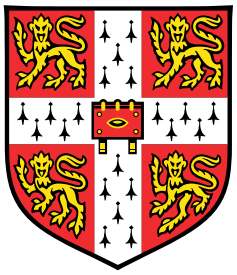
Reconstructing Chromothriptic Chromosomes in Oesophageal Adenocarcinomas



Jannat Ijaz

Wellcome Sanger Institute
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

Churchill College

September 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the Biological Sciences Degree Committee.

Jannat Ijaz
September 2021

Reconstructing Chromothriptic Chromosomes in Oesophageal Adenocarcinomas

Jannat Ijaz

Abstract

The epigenetic landscape is regulated by a myriad of factors. This regulation ranges from functional compartmentalisation of genomic sequences into topologically associating domains, to chromosome looping, to short-range promoter-enhancer interactions. The underlying genome sequence contributes to this regulation, likely at a variety of scales, however the extent of this contribution is not fully understood. Chromothripsis is a localised catastrophic genome shattering event that can be used to study how the underlying genomic sequence affects this higher order structuring. Since chromothripsis tends to affect only one of the two alleles, in every cell a direct comparison can be made between the wild-type chromosome and the chromothriptic chromosome. The wild-type chromosome represents the genome sequence and structure before reshuffling and the chromothriptic derivative chromosome can be used to query the direct effects of this reshuffling.

Chromothripsis has been seen in up to 32% of cases of oesophageal adenocarcinomas. Therefore, patient-derived oesophageal adenocarcinoma organoids with evidence of chromothripsis restricted to one allele were used to better understand how the genome is regulated. Complex regions of structural variation between alleles in cancer genomes coupled with subclonal variants means haplotype-aware *de novo* assemblies are essential for contiguous cancer genome assemblies. Our method takes haplotype blocks and assigns PacBio circular consensus sequencing reads to the appropriate allele using B-allele frequencies of single nucleotide polymorphisms and presence of structural variants. The chromosomes are then assembled separately and scaffolded using Hi-C reads, which we also haplotype resolve. This produces contiguous assemblies, even on chromosomes with over 900 structural rearrangements compared to the reference genome. This methodology has been used to reconstruct chromothriptic derivative chromosomes and the associated wild-type chromosomes in five organoid models, as well as other chromosomes with complex rearrangements. All types

of structural variant have been reconstructed, other than tandem duplications which are collapsed by current assembly tools.

With these cancer-specific reference assemblies, the epigenome of the chromothriptic and wild-type chromosomes can be profiled. Hi-C chromosome capture has been used to study topologically associated domains; ATAC-seq to study chromatin accessibility; ChIP-seq to identify CTCF binding and histone modifications (H3K27me3, H3K4me3, H3K27ac) and Iso-seq to phase long read transcripts to their respective chromosomes. There are widespread differences between the chromothriptic and wild-type chromosomes for each epigenetic mark. This indicates that the shattering of the chromosome has dramatic consequences for gene regulation, far beyond what we see when comparing two wild-type alleles of the same chromosome. It highlights that, while underlying genome sequence has a fundamental role in gene regulation, the epigenetic context of that sequence also has a profound impact. The work done to assemble these chromosomes allows for unprecedented insight into the regulatory impact of structural variation.

Acknowledgements

I would first like to thank my supervisor, Dr Peter Campbell, for proposing a project that has challenged and inspired me every single day over the last 3 years. Your patience and guidance as I learnt to code and become a computational biologist was unwavering and your constant scientific curiosity and excitement when seeing data is infectious and kept me motivated, even when nothing was working.

I would also like to thank many people in the Cancer, Ageing and Somatic Mutation department in the Wellcome Sanger Institute. While this thanks extends to everyone in the department, there are some without whom this thesis would not have materialised. Thanks to Dr Tim Coorens for being my mentor when I started, for keeping my spirits up, for all the crosswords and for being my sound board whenever I needed it. To Dr Tim Butler for the never-ending support, not just for me but for the entire lab. To Dr Andrew Lawson for an unrivalled attention to detail, for proofing the final version of this thesis and for always telling me where the anagram indicator is. To Dr Heather Machado for being an inspiration to us all, not just as a phenomenal scientist but also for the incredible kindness and warmth you cannot help but give out. To Dr Alex Cagan, I am constantly in awe of you for your ability to do it all: scientist, artist and entrepreneur. To Luke Harvey for always making me laugh, even when all my jobs were crashing on the farm and my genomes were in pieces. To Dr Pantelis Nicola for being my own personal doctor and your willingness to answer all my obscure medical questions. To Dr Daniel Leongamornlert, for the countless pep talks and advice in both the peaks and troughs of this project. To Dr Sam Behjati for an unending enthusiasm and for always wanting to only be involved in the most exciting science. To Keiran Raine, Andrew Menzies, Dr Kathryn Beal, Adam Butler and Jon Teague for helping me establish the computational pipelines I needed and always answering my "but why?" questions. And of course, the lab support team, the IT support team, the core facilities and the sequencing team.

I would also like to thank my parents for always letting me believe I could do everything I wanted to. For putting me first, no matter what. I owe you everything. I want to thank my uncle, Nico Uitenbroek, for everything he has done for me, we may be in different countries but I would not be here without you. I would also like to thank James Miller for his constant

support throughout this entire thesis writing process, for reminding me to hyphenate and for never failing to make me smile no matter how badly the writing was going.

I would also like to thank the Wellcome Trust for funding the study and the patients for donating the samples, this work would not have been possible without them. And thank you, the reader, for reading this thesis. I hope it is as fun to read as it was to write.

Table of contents

List of figures **xiii**

List of tables **xv**

Nomenclature **xvii**

1 Introduction **1**

1.1 What drives gene expression? 1

1.2 Genome assembly 3

1.2.1 Assembling the human genome 3

1.2.2 Diversity of the reference genome 4

1.2.3 Current sequencing technologies 4

1.2.4 Genome assembly methods 6

1.2.5 Haplotype-aware assemblies 8

1.3 Rearrangements in cancer genomes 8

1.3.1 Mutations in cancer genomes 9

1.3.2 Chromothripsis 10

1.3.3 Other complex rearrangements 14

1.4 Epigenetic alteration 15

1.4.1 Mutations and gene expression 17

1.4.2 Mutations and chromatin accessibility 19

1.4.3 Mutations and topologically associating domains 20

1.5 Chromothripsis as a model 22

2 Materials and Methods **23**

2.1 Organoid culture 23

2.2 Sequencing and associated protocols 23

2.2.1 Genomic sequencing 23

2.2.2 Epigenomic sequencing 24

2.3 Sequence alignment . . . 25
2.4 Genomic variant calling . . . 25
2.5 *De novo* haplotype resolution . . . 26
2.6 Genomic assembly methods . . . 26
2.7 Assembly-based haplotype resolution . . . 27
2.8 Calling topologically associating domains . . . 28
2.9 Peak calling in ChIP-seq and ATAC-seq data . . . 28
2.10 Identifying differential transcript expressions . . . 28
2.11 Gene essentiality . . . 29
2.12 Integrated analysis . . . 29

3 Generating a cancer-specific reference genome 31
3.1 Chapter highlights . . . 31
3.2 Introduction . . . 31
3.3 Selection of organoids . . . 33
3.4 Growth of organoids . . . 37
3.5 Karyotyping . . . 38
3.6 Haplotype-unaware assemblies . . . 40
3.7 Haplotype-resolved *de novo* assemblies . . . 41
3.7.1 Calling and phasing structural variants . . . 41
3.7.2 Phasing reads . . . 42
3.7.3 *De novo* assemblies . . . 47
3.8 Scaffolding using Hi-C reads . . . 49
3.9 Haplotype-resolved structural variant calling . . . 53
3.10 Final assembly statistics . . . 56
3.10.1 Chromothripsis in other samples . . . 56
3.10.2 Non-chromothriptic chromosome assemblies in all organoids . . . 62
3.10.3 Assembling all types of structural variants . . . 64
3.10.4 Assembling other complex rearrangements . . . 67
3.10.5 Unexpected assembly difficulties . . . 70
3.11 Discussion . . . 70

4 Structural variants and the epigenome 75
4.1 Chapter highlights . . . 75
4.2 Introduction . . . 75
4.3 Haplotype resolution of epigenetic and expression data . . . 76
4.4 Differential expression and essentiality . . . 78

4.4.1 Comparison of transcripts on haplotypes . . . 78
4.4.2 Classification of reads . . . 79
4.4.3 Differential expression between the two haplotypes . . . 79
4.4.4 Relationship between SVs and differential transcript expression . . . 82
4.4.5 Essential genes . . . 85
4.5 Differential protein binding and accessibility . . . 86
4.5.1 Comparison of histone modifications on haplotypes . . . 86
4.5.2 Summary of marks . . . 86
4.5.3 Differential histone modifications and accessibility . . . 88
4.5.4 Relationship between SVs and differential histone modifications and chromatin accessibility . . . 91
4.6 Topologically associating domains . . . 94
4.6.1 Comparison of long-range interactions . . . 94
4.7 Integrative analysis of differential epigenomic patterns . . . 99
4.7.1 Altered histone marks . . . 99
4.7.2 Altered chromatin interactions . . . 101
4.7.3 Fusion gene formation . . . 103
4.7.4 Unexplained alterations . . . 103
4.7.5 Overview . . . 105
4.8 Discussion . . . 107

5 Conclusions and outlook 111
5.1 Conclusions . . . 111
5.2 Future work . . . 112
5.2.1 Cancer genome assembly . . . 112
5.2.2 Epigenetics . . . 114
5.3 Final remarks . . . 116

References 117

Appendix A Amplified Genes in chromothripsis 133

List of figures

1.1 Chromothripsis mechanism 11

3.1 Sequence ambiguity from increased copy number segments 33
3.2 Chromothriptic regions in each sample 35
3.3 BAF of chromothriptic regions in each sample 37
3.4 Organoid cell line images 38
3.5 Cell line heterogeneity 39
3.6 Haplotype switching 44
3.7 Haplotype switching schematic 45
3.8 Dot plot alignments of CCS assemblies 48
3.9 Steps in Hi-C scaffolding 49
3.10 Dot plot alignments of scaffolds 51
3.11 Main steps in assembly method 52
3.12 xmatchview alignments for WTSI-OESO_103 53
3.13 Chromosome 6 WTSI-OESO_103 summary statistics 54
3.14 Reconstructing chromothripsis in other samples 57
3.15 WTSI-OESO_152 chromosome 9 contig alignment to WTSI-OESO_117 . 59
3.16 Reconstructed small duplications 61
3.17 Summary assembly statistics for non-chromothriptic chromosomes 63
3.18 Assembly fragmentation 64
3.19 Reconstructing inversions and deletions 65
3.20 Reconstructing tandem duplications 66
3.21 Reconstructing complex rearrangements 68
3.22 Assembly limitations 71

4.1 Transcript summary 80
4.2 Differential transcript analysis 83
4.3 SVs nearest to differential genes 84

4.4 Summary of peaks called by MACS2 87
4.5 Overall histone modifications on chromosomes across samples 89
4.6 Differential protein binding and histone modifications 92
4.7 SVs nearest to differential peaks 93
4.8 Structural variant altered TADs 95
4.9 Conserved TADs 97
4.10 Structural variation can lead to altered TADs 98
4.11 Differential gene expression from altered surrounding histone activity states 100
4.12 Differential gene expression from altered TAD structures 102
4.13 Differential gene expression and fusion genes 104
4.14 Unexplained differential gene expression 106

List of tables

2.1 Probability assignment 28

3.1 Evidence of chromothripsis 36
3.2 Randomness of DNA joins in chromothripsis 36
3.3 Clinical organoid data 36
3.4 Genome sequencing coverage 38
3.5 Haplotype-unaware assembly 41
3.6 WhatsHap phasing 43
3.7 Structural variant density 46
3.8 Haplotype-aware initial wild-type assemblies 47
3.9 Haplotype-aware initial chromothriptic assemblies 47
3.10 Final wild-type assembly and scaffolding metrics 50
3.11 Final chromothriptic assembly and scaffolding metrics 50
3.12 SVs spanning genomic features 55

4.1 Epigenetic sequencing coverage 77
4.2 Read assignment groups 77
4.3 Differential transcripts in other samples 81
4.4 SVs nearest to differential genes 84
4.5 Differential ChIP-seq and ATAC-seq peaks in other samples 90
4.6 SVs nearest to differential peaks 93

Nomenclature

Acronyms / Abbreviations

ATAC-seq	Assay for Transposase-Accessible Chromatin using Sequencing
BAF	B-Allele Frequency
BME2	Basement Membrane Extract, type 2
bp	Base Pairs
BrdU	Bromodeoxyuridine
BWA	Burrows-Wheeler algorithm
CAUS	Chromosome Assignment Using Synteny
CaVEMan	Cancer Variants Through Expectation Maximization
CCS	Circular Consensus Sequencing
CGaP	Cellular Genotyping and Phenotyping
ChIP-seq	Chromatin Immunoprecipitation Sequencing
chr	Chromosome
CLR	Continuous Long Read
COSMIC	Catalogue of Somatic Mutations in Cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTCF	CCCTC-Binding Factor
CT	Chromothriptic

DLS	Direct Label and Stain
EZH2	Enhancer of Zeste Homologue 2
GRCh38	Genome Reference Consortium Human Build 38
H3K27ac	Histone H3 protein Lysine (K) 27 acetylation
H3K27me3	Histone H3 protein Lysine (K) 27 Tri-methylation
H3K4me3	Histone H3 protein Lysine (K) 4 Tri-methylation
Iso-seq	Full-Length Isoform Sequencing
kb	Kilo Base pairs
LOH	Loss of Heterozygosity
Mb	Mega Base pairs
MNase	Micrococcal Nuclease
MSPC	Multiple Sample Peak Calling
NGMLR	coNvex Gap-cost alignMents for Long Reads
PacBio	Pacific Biosciences
PCR	Polymerase Chain Reaction
RIN	RNA Integrity Number
SMRT	Single-Molecule Real-Time
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SV	Structural Variant
TAD	Topologically Associating Domain
TCGA	The Cancer Genome Atlas
VAF	Variant Allele Fraction
WT	Wild-Type

Chapter 1

Introduction

1.1 What drives gene expression?

The human body is composed of a huge number of cell types, which differ in cellular structure and function. It is staggering to believe that, for example, the underlying genome in a normal skin cell is the same as that of a normal liver cell. In order for these cells to be so dramatically different, it is essential that there is tight regulation on gene expression. Rather than directly altering the genetic code, this diversity in cell function is enabled by highly regulated, heritable and reversible changes in chromatin, chemical modification of bases and functional compartmentalisation of the genome into discrete regions. This is known as epigenetic modification. The term epigenetics was coined in 1942 as changes in phenotype without changes in genotype (Waddington, 1942a,b). The term epigenetics is now used more broadly to encompass the study of chromatin biology and modifications, which is the definition for epigenetics used in this thesis.

While epigenetic modifications are fundamental in regulation of gene expression and establishment of cell fate, underlying genome sequence is a primary driver of gene expression. Human chromosome 21 is broadly syntenic to mouse chromosome 16, with a few additional regions of human chromosome 21 found on mouse chromosomes 10 and 17 (Shinohara *et al.*, 2001). When a human chromosome 21 is placed in a transgenic mouse, fundamental questions can be asked about the role of primary genome sequence on higher-order structuring. If the transcription factors bind to the human chromosome in a mouse cell at the same sites as a human chromosome in a human cell, this suggests it is the underlying genome sequence that dictates this binding and therefore gene expression. If there are multiple binding sites on the human chromosome in a mouse cell that are not present on a human chromosome in a human cell, then other factors must be regulating this binding. A study investigating this question (Wilson *et al.*, 2008) found that a human chromosome 21 in a mouse has

transcription factor binding profiles and active histone modifications (H3K4me3 marks) that were almost identical to a human chromosome in a human cell. This is particularly notable given the propensity of transcription factors to bind to noncanonical binding sites. It suggests neither the cellular microenvironment nor differences in remodelling complexes between human and mouse cells influence the placement of active chromatin marks or transcription (Wilson *et al.*, 2008).

However, there is also undoubtedly an influence from chromatin organisation in order for cells with the same genome to function so differently. The degree of this influence and its relationship to primary genome sequence is currently poorly understood and the aim of this thesis is to gain further insight into this influence. This sequence-function relationship can be studied in many ways. Since the evolution of functional sequences is slower than the evolution of non-functional sequences, comparing DNA sequences in different species can identify conserved and functional regions of the genome (Frazer *et al.*, 2003). When there is an evolutionary difference between species in one of these conserved functional regions, the differences that result from this mutation can be studied. However this is complicated by the variation in neutral evolution rates in different regions of the genome and so systematic genome-wide investigations are difficult (Waterston and Pachter, 2002). Alternatively, germline mutations and polymorphisms can be used to study the epigenome. For example, in prostate cancer germline polymorphisms have been used to study alterations in the methylation profile and subsequently the effect on chromatin structure and gene expression (Houlahan *et al.*, 2019). However, germline variation is strongly influenced by positive and negative selection and emerges over evolutionary time scales. Therefore, it can be difficult to isolate sequence effects from other non-local effects. Another method utilises genome editing technologies, such as clustered regularly interspaced short palindromic repeats-Cas9 (CRISPR-Cas9), to activate or repress enhancer function (Li *et al.*, 2020a) or to use saturation mutagenesis to target specific functional regions (Kircher *et al.*, 2019). However, both methods are time-consuming and low-throughput, allowing only certain changes to the genome to be studied. Finally cancers, particularly those that have undergone catastrophic genome reshuffling events, often exhibit large-scale alteration in the genome sequence of one allele. Therefore, cancer genomes are useful tools to understand this regulation and this is the method that will be used to elucidate the sequence-function relationship in this thesis. Combining chromatin accessibility, histone modification and functional compartmentalisation of the genome into topologically associating domains will allow elucidation of how the 3D structure affects gene expression. While each regulatory layer will be informative in and of itself, a view encompassing multiple layers will allow

insight into mechanisms which remain elusive when only looking at one layer. This in-depth analysis has, to my knowledge, not been done before.

This introduction provides an overview of the historical perspective and recent advances in genomic and epigenomic sequencing technologies, rearrangement in cancer genomes and epigenetic alteration.

1.2 Genome assembly

1.2.1 Assembling the human genome

Sequencing the entire human genome was first suggested in 1985 (Sinsheimer, 1989), which was met with enthusiasm from research councils throughout the world. By 1990, the Human Genome Project was launched with genome centres in the UK, the US, France, Japan and later also Germany and China. In the UK, the Sanger Centre, directed by John Sulston, played a pivotal role. The initial goal of the project was to produce a draft genome; which would provide a thorough, if incomplete, representation of the human genome (Lander *et al.*, 2001).

The first draft chromosomes to be finished were chromosomes 21 (Hattori *et al.*, 2000) and 22 (Dunham *et al.*, 1999) as they are the smallest human chromosomes. This was done using hierarchical shotgun sequencing. The first draft of the human genome sequence was released in 2001, containing roughly 94% of the human genome (Lander *et al.*, 2001; Venter *et al.*, 2001). The "complete genome" was finished in 2003. It was composed of 2.85 billion bases, with 99% of euchromatic sequence being resolved but still containing 341 gaps (International Human Genome Sequencing Consortium, 2004). It was the first vertebrate genome to have been sequenced, 25 times larger than any other genome that had been sequenced until then and eight times larger than any genome assembled by that point.

Over the following years, sequencing technologies and assembly methods continued to improve, allowing for the initial draft genome to be routinely updated. In 2020, the first telomere-to-telomere assembly was produced (Miga *et al.*, 2020). It was an assembly of chromosome X which primarily used ultra-long-read Nanopore sequencing of the complete hydatidiform mole CHM13, an effectively haploid cell line. It led to a gap-less assembly. Chromosome 8 was the next chromosome with a full telomere-to-telomere assembly to be released (Logsdon *et al.*, 2021) and the remaining chromosomes are hoped to also have telomere-to-telomere assemblies soon.

1.2.2 Diversity of the reference genome

The human reference genome is based on the genetic sequences of many donors. However, a single European donor constitutes 70% of the sequence. Since this individual was found to have a high risk of diabetes, they did not represent a "healthy" sample (Chen and Butte, 2011). Furthermore, a reference based largely on an individual of European ethnicity does not accurately represent differences seen in other ethnic groups. These differences include sequence variation and linkage disequilibrium. Sequencing efforts have begun to catalogue these differences. Examples include the International HapMap project (Gibbs *et al.*, 2003), the African Genome Variation project (Gurdasani *et al.*, 2015), the Genome of the Netherlands project (Francioli *et al.*, 2014), the GenomeAsia 100K project (Wang *et al.*, 2008) and the Simons Genome Diversity project (Mallick *et al.*, 2016).

Efforts are now being made to incorporate alternative loci into newer reference genome builds, including the current human reference genome (GRCh38) (Schneider *et al.*, 2017). These projects have also highlighted that there are missing sequences when comparing the GRCh38 reference genome to the assembled genomes of other ethnicities. For example, the pan-assembly of African genomes found that the current reference genome lacks roughly 10% of sequence found in these genomes, much of which falls within protein coding regions (Sherman *et al.*, 2019). These ethnicity-specific genomes will be vital to accurately determine disease-causing variants, because they are less likely to include erroneous calls arising from differences in the underlying genome sequence. In cancer genome analysis, this will mean that determining somatic verses germline variants will not be complicated by reference verses germline variants.

1.2.3 Current sequencing technologies

The human genome is highly complex with many regions of repetitive elements and structural variation, meaning both sequencing and assembly of those regions is problematic. This often leads to sequencing and assembly gaps (Treangen and Salzberg, 2012). Short read sequencing reads are typically 150 bp and therefore often do not cover the entire repeat region. This means they are often unable to resolve highly complex, repetitive regions unambiguously. Long-read sequencing methods produce reads which are kilobases in length meaning many of these regions can be covered by a single read and therefore unambiguously resolved. There are three main types of long-read sequencing: single-molecule real-time (SMRT) sequencing, nanopore sequencing and synthetic long-read sequencing.

PacBio sequencing uses SMRT sequencing. It relies on immobilising the polymerase to the bottom of a well on a flow cell and allowing the DNA strand to move through the poly-

merase. Each molecule is then visualised using a laser and camera. The fluorescently-labelled base that is incorporated by the polymerase is then recorded by monitoring the duration and colour of light emitted (Eid *et al.*, 2009). Continuous long read (CLR) sequencing will pass each DNA molecule through the polymerase once and will have a high error rate (8–15%) (Logsdon *et al.*, 2020) with read lengths ranging from 250 bp to 50 kb (Amarasinghe *et al.*, 2020). Circular consensus sequencing (CCS) will use a circular template in order to sequence each DNA molecule multiple times and subsequently generate a consensus sequence. These reads become highly accurate (99.8%) as errors in each pass of the template can be corrected by consensus (Wenger *et al.*, 2019). In order to do this, a read length size selection is needed to ensure that the polymerase can sequence each read multiple times. This is often 13.5 kb (Wenger *et al.*, 2019), however larger CCS reads are now possible when using a higher-processivity polymerase.

Oxford Nanopore sequencing directly detects which base is being incorporated into the DNA. The DNA molecule moves through a biological pore and the current through the pore is measured. The sequence of DNA can then be inferred from the changes in current as different bases will cause different fluctuations based on the run of nucleotides already in the pore. This even allows measurement of base modifications, such as methylation (Clarke *et al.*, 2009). The Nanopore sequencing reads are much longer than PacBio sequencing reads. They can be up to 2.3 Mb and read length is limited by length of input DNA rather than the processivity of a DNA polymerase (Amarasinghe *et al.*, 2020). However, this comes at a cost and these reads have a high error rate (~14%), much higher than the CCS reads (Sahlin and Medvedev, 2021).

Synthetic long-read sequencing generates long reads using barcoded short-read sequencing reads. They segregate reads into wells or emulsions so that very few DNA molecules are present in each segregation. These reads are then uniquely labelled and sequenced on a short-read platform. Short-reads which share barcodes are derived from the same long-read and therefore these reads can be assembled together (Goodwin *et al.*, 2016). There are two main types of synthetic long-read sequencing: Illumina long-read sequencing and 10X Genomics linked-reads. Illumina long-read sequencing separates reads in a microtitre plate, whereas 10X Genomics linked-reads uses a microfluidic device to separate reads into droplets. The 10X Genomics linked-reads have gaps in coverage of each individual long read, however this can be overcome by sequencing multiple long reads of the same genomic region (Goodwin *et al.*, 2016).

It is also important to note that optical mapping can be used to determine long sequences of DNA. Optical mapping uses enzymes which label DNA at specific sequence motifs and this generates DNA fingerprints (Schwartz *et al.*, 1993). The average length of a optical map

molecule is 255 kb, considerably longer than the current average read length of long-read sequencing reads (Yuan *et al.*, 2020). Until 2012, optical mapping was low-throughput (Yuan *et al.*, 2020), however Bionano optical mapping uses a massively parallel platform to generate these maps. This makes optical mapping a more feasible method for large-scale genome projects (Lam *et al.*, 2012). The direct label and stain (DLS) technology has resulted in chromosome-scale maps being produced (Lam *et al.*, 2012). The resolution of Bionano optical mapping is much lower than long-read sequencing technologies and, although improvements are constantly being made, the 500 bp resolution is incomparable to the base pair resolution of long-read sequencing technologies.

Long-read sequencing alone is sufficient for many biological research applications. However, for some applications, such as genome assembly, hybrid approaches can improve results by combining long-read and short-read sequencing data (Mahmoud *et al.*, 2019; Shi *et al.*, 2019). These hybrid approaches allows the strengths and benefits of different technologies to complement each other. Short reads and CCS reads have a high accuracy whereas CLR reads, ultra-long nanopore reads and optical mapping are less accurate but provide longer-range sequencing context.

Hybrid approaches often also combine long reads with Hi-C chromosome capture. Hi-C is an extension of low-throughput 3C approaches (Dekker *et al.*, 2002) and it measures the frequency of interaction between two loci. Regions which are in close spatial proximity are crosslinked. DNA is fragmented using a restriction enzyme and these interacting regions are then ligated together and subsequently sequenced. The initial goal of Hi-C was to study 3D organisation of the genome (Lieberman-Aiden *et al.*, 2009), but it has also been used to successfully improve genome assemblies. This is because the genome is partitioned into functional compartments and therefore physical distance correlates very strongly with the number of interactions between two regions.

The lower cost of sequencing has meant that long-read sequencing is becoming more feasible as a routine sequencing method (Lightbody *et al.*, 2019). Read lengths are increasing (Goodwin *et al.*, 2015) and computation methods are improving (Gavrielatos *et al.*, 2021). Together these factors mean that sequencing an entire genome is becoming somewhat trivial and more comprehensive studies into fundamental biology and disease are possible.

1.2.4 Genome assembly methods

One use of long-read sequencing is genome assembly. The goal is to take sequencing reads and construct the sequence of longer regions of the genome by inferring how the reads overlap one another. From this a consensus of continuous sequences, known as a contig, is generated. Genome assembly methods can be reference-based, meaning they derive from a pre-existing

reference genome, or *de novo*, meaning they are generated using only information contained in the sequencing reads themselves.

Reference-based assemblies utilise the similarity of reference and target assemblies in order to use the reference as a scaffold upon which to place the target contigs or reads. (Bao *et al.*, 2014; Pop *et al.*, 2004). Reference-based assemblies can first align reads to a reference genome and then generate a consensus based on known sequence and deviation from that sequence (Vezzi *et al.*, 2011). Alternatively, an initial *de novo* assembly can be produced and then these contigs can be aligned against the reference to ascertain order and scaffold these contigs together. Using reference-based methods also allows positions of genes to be easily ascertained as reference genomes are often highly annotated (Bao *et al.*, 2014). These methods work well for generating genome assemblies for closely related species and for generating patient-specific assemblies if genomes do not greatly diverge from the reference. The more divergent the target genome is from the reference genome; the more mistakes are generated using reference-based methods (Card *et al.*, 2014). Furthermore, any errors in the reference genome will introduce mistakes into the assembly (Ekblom and Wolf, 2014).

Conversely, *de novo* genome assemblers do not use a reference. There are three main methods of *de novo* genome assembly (Liao *et al.*, 2019): overlap consensus, de Bruijn and string graph. Both overlap consensus and string graph assemblers generate data structures, known as graphs, bases on full reads whereas de Bruijn graph assemblers split reads into a sequence of bases that are of a length designated k, a k-mer. The larger the initial k-mer or read, the greater the number of repeat regions covered. However, this will lead to a greater number of shorter unconnected sequences and a more fragmented assembly (Luo *et al.*, 2015). Overlap consensus assembly was the primary method for assembling the human genome initially. Wtdbg2 (Ruan and Li, 2020) and Canu (Koren *et al.*, 2017) are assemblers which use overlap consensus graphs. SPAdes (Bankevich *et al.*, 2012) is a de Bruijn graph-based assembler. Hifiasm (Cheng *et al.*, 2021) and FALCON (Chin *et al.*, 2016) are assemblers which use string graphs.

There are many general challenges associated with genome assembly. Firstly, many long reads have high sequencing error rates which can increase the complexity of the graphs. De Bruijn graphs are particularly affected by this and the resultant graphs tend to be unsatisfactory. CCS reads, which have a much lower sequencing error rate, or error correction methods can be used to overcome this. Secondly, there is often sequencing bias, which can lead to uneven coverage of the genome. Thirdly, repeat regions are distributed throughout the human genome (Cordaux and Batzer, 2009), with non-random repeats constituting 50% of the genome (Kazazian, 2004). These elements and the subsequent uneven sequencing depth in these regions lead to gaps, breaks or errors in the genome assembly. However if

the length of the repeat is shorter than the read length, or if each repeated element contains occasional unique base changes, these repeats can be unambiguously resolved. Finally, it is computationally intensive to assemble large genomes (Liao *et al.*, 2019).

Once the initial contigs have been formed, their order and orientation relative to one another is deduced and they are scaffolded together, with Ns denoting regions between two scaffolded contigs. Misassemblies can also be corrected using inconsistencies in read depths. This leads to a more contiguous assembly and is a key step in the assembly process. Hi-C and 10X reads are often used to scaffold assemblies together as they contain long-range contacts while still being highly accurate (Luo *et al.*, 2021a).

1.2.5 Haplotype-aware assemblies

The human reference genome represents both alleles of each chromosome with a single reference, and so is known as a squashed assembly. The first diploid human genome was produced using Sanger sequencing (Levy *et al.*, 2007). Haplotype-aware assembly methods fall into two main groups. Alignment-based methods align reads to a reference and assign single nucleotide polymorphisms (SNPs) to a specific haplotype to segregate reads. This requires a high-quality reference genome as the phased SNPs are then placed in the reference genome to produce modified reference sequence patches. *De novo*-based methods incorporate haplotypes into the initial assembly graph (Luo *et al.*, 2021b). These do not rely on a reference genome and often an initial squashed assembly is produced which is subsequently split into haplotypes. Many genome assemblies being produced are now haplotype-resolved (Cheng *et al.*, 2021; Porubsky *et al.*, 2021; Xu *et al.*, 2021; Yen *et al.*, 2020). This is particularly useful for cancer genomes where mutations are allele-specific, meaning haplotype-resolved cancer genome assemblies will allow mutations to be phased and allele-specific differences to be queried (Adey *et al.*, 2013). Since library preparation and sequencing artefacts will not phase with SNPs, haplotype-resolved assemblies may also assist accurate variant calling.

1.3 Rearrangements in cancer genomes

The first whole genome sequence from a cancer was published in 2008 (Ley *et al.*, 2008). This genome was used to discover somatic mutations, both single nucleotide variants (SNVs) and small insertion and deletion events (indels), in coding sequences of an acute myeloid leukaemia. The first fully analysed cancer genomes analysing SNVs, indels and structural variants (SVs) in both coding and non-coding regions occurred in a malignant melanoma and

a small-cell lung cancer cell line (Pleasance *et al.*, 2010a,b). All three studies fundamentally relied on the presence of a reference genome to analyse these cancer genomes.

Today, the sequencing of cancer genomes is widespread and many cancers have been comprehensively characterised (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) giving insight into frequently altered regions of the genome. The Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census reports somatically acquired mutated genes (Sondka *et al.*, 2018) and with every genome sequenced more information is gained.

1.3.1 Mutations in cancer genomes

Somatic cells accumulate spontaneous mutations throughout their lifetime. Most of these mutations are selectively neutral providing no phenotypic advantage or disadvantage. These are known as passenger mutations. Others are beneficial and will be positively selected, driving the progression of the cell towards a disease phenotype or contributing to aging. The final set are harmful to the growth and function of the cell and will be negatively selected. This process of mutation accumulation and selection leads to the Darwinian evolution of a cell and overtime leads to the accumulation of mutations. It is also important to note that mutations that are harmful or beneficial at the cellular level may have a different effect at the organismal level.

The Darwinian evolution of cells requires natural genetic variation within a population. In the context of cancer, this manifests as somatic mutations in different cells within a population. It also requires that the genetic variation changes the phenotype of the cells, rather than just being passenger variation. Finally, there must be competition for resources among the cells in the population. In the context of cancer, this is density-dependent growth suppression signals from in-built and host microenvironment surveillance as well as nutrient availability. Together, this Darwinian evolution leads to cells with specific mutations having a proliferative advantage and outcompeting cells without these mutations to become the dominant clone. This mutation accumulation is constant, meaning there is constant clonal evolution of cellular populations.

The most common model of cancer development is this Darwinian evolution and progressive accumulation of mutations over time which inactivate tumour suppressor genes and activate oncogenes (Stratton *et al.*, 2009). For example, in oesophageal adenocarcinoma there is a progressive accumulation of mutations in many genes including *TP53*, *SMAD4*, *ARID1A* and *CDKN2A*. While mutations in *ARID1A* and *CDKN2A* occur at all points in cancer development, mutations in *TP53* occur as the abnormal cells progress to Barrett's oesophagus with high-grade dysplasia and mutations in *SMAD4* occur after progression to

early invasive oesophageal adenocarcinoma (Weaver *et al.*, 2014). These mutations can act as markers of cancer progression stage. They provide a selective advantage to the cell, allowing it to develop a more tumourigenic phenotype.

Mutations in cancer genomes can occur through SNVs, indels and larger SVs. While SNVs and indels can alter a few bases, structural variation amplifies, deletes and reorders regions of the genome at a scale of 100s of bases through to whole chromosomes. A structural variant can be balanced if both sides of the double-stranded break are subsequently joined to another region of DNA. Balanced structural variants remain connected to a centromere and protected by telomeres. SVs can also be unbalanced if only one side of the double-stranded break is joined to another region of DNA. This manifests as a change in the copy number of the segment that is not joined to DNA (Li *et al.*, 2020b). Structural variants can be grouped based on number of breakpoints (simple or complex) and on whether the mechanism of action is cut-and-paste or copy-and-paste. The cut-and-paste mechanism causes rearrangements where there is loss or reshuffling of genomic regions from incorrect ligation of DNA ends (Li *et al.*, 2020b). This includes deletions, inversions, translocations, chromoplexy (Baca *et al.*, 2013) and chromothripsis (Stephens *et al.*, 2011). The copy-and-paste mechanism produces rearrangements where there are genomic templates inserted into sequences of DNA leading to associated copy number changes (Li *et al.*, 2020b). This includes tandem duplications, local-distant clusters, local n-jumps, cycles of templated insertions and breakage-fusion-bridge-cycles (Li *et al.*, 2014).

All types of structural rearrangements have the potential to have functional impact on gene expression, including the creation of chimeric proteins, inactivation of genes through loss, overexpression of genes through amplification and alteration of gene regulation from enhancer hijacking or placement into a different regulatory environment. Assembling the genomes of cancers which have undergone large-scale rearrangements will allow better understanding of this functional impact and the relationship between primary sequence alteration and higher-order regulation.

1.3.2 Chromothripsis

Chromothripsis is a catastrophic genome shattering event whereby a localised region of the genome is fragmented into 10s to 100s of pieces and religated in a seemingly random order and orientation (Stephens *et al.*, 2011) (Figure 1.1). Many cells will not be able to survive this catastrophic event. The rare cells that do survive will have heavily rearranged genomes. Many of these rearrangement events will be selectively neutral, passenger mutations but a handful of them will provide a proliferative advantage, be under positive selection and aid the cell on the tumourigenic progression (Hanahan and Weinberg, 2011).

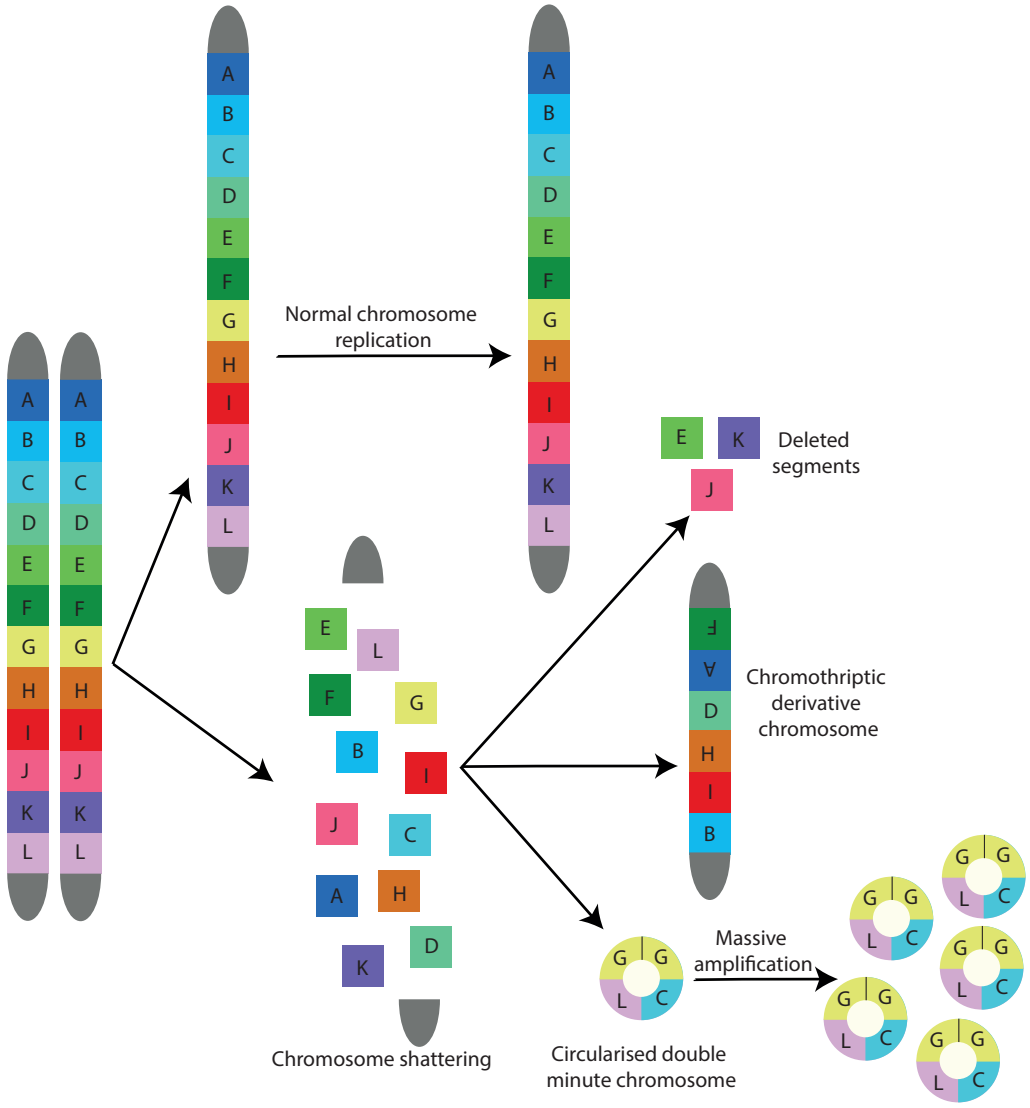


Fig. 1.1 Schematic representation of chromothripsis. Chromothripsis tends to only affect one of the two alleles. One of the alleles undergoes normal DNA replication. The other is fragmented. Some of the segments are retained and ligated together in a random order in both the original and inverted orientations. This forms the derivative chromosome. Of the segments that are not incorporated into the derivative chromosome, some are lost completely and others can form circularised double minute chromosomes. Sequences in these double minutes can become duplicated and the double minutes may be subsequently highly amplified. After the chromothriptic event, there is a wild-type chromosome and a highly rearranged derivative chromosome.

Chromothripsis is characterised by four main hallmarks. Firstly, the rearrangements are highly localised to a small region of the genome and within that region breakpoints are clustered and juxtapose regions which are linearly far apart (Stephens *et al.*, 2011). This

is in contrast to the progressive mutational mechanism whereby rearrangements are found throughout the genome or, if localised, genomic segments in that region become highly amplified (Campbell *et al.*, 2008; Zhang *et al.*, 2009). Secondly, the copy number profile oscillates between usually two, occasionally three, different copy number states depending on whether a genomic segment is lost or retained (Stephens *et al.*, 2011). Conversely in a progressive mutational mechanism, any number of copy number states can occur depending on where each sequential break occurs (Zack *et al.*, 2013). Thirdly, heterozygosity is maintained in the regions with a copy number greater than one, even when there are juxtaposed tandem duplication orientation and deletion orientation rearrangements (Stephens *et al.*, 2011). This differs from a progressive mechanism where a deletion occurring before a duplication event would lead to loss of heterozygosity. Finally, the regions with a copy number of one result from a series of complex rearrangements rather than just simple deletions (Stephens *et al.*, 2011). Chromothripsis tends to only affect one parental chromosome. In 40% of cases, only one chromosome is affected (Cortés-Ciriano *et al.*, 2020) but it can sometimes affect multiple chromosomes if multiple chromosomes are segregated and shattered together or if a translocation has occurred prior to the chromothriptic event. This is particularly common in osteosarcomas where at least five chromosomes are affected in 61% of chromothripsis cases (Cortés-Ciriano *et al.*, 2020).

There are multiple proposed mechanisms for chromothripsis (Forment *et al.*, 2012; Holland and Cleveland, 2012; Jones and Jallepalli, 2012; Kloosterman and Cuppen, 2013). During interphase, chromosomes are unpacked with many regions of long exposed DNA (Misteli, 2007). When DNA is unpacked like this, the clustering of breakpoints seen in chromothripsis is unlikely to occur and consequently it is unlikely that chromothripsis occurs in interphase. *In vitro* data suggest that a mitotic catastrophe may be the origin of chromothripsis, with micronucleus formation being a key mechanism to physically isolate chromosomes so that localized rearrangement can occur (Crasta *et al.*, 2012; Kato and Sandberg, 1968; Zhang *et al.*, 2013, 2015a). Micronuclei contain whole or parts of chromosomes and form when acentric, dicentric or lagging chromosomes do not divide accurately between the two daughter nuclei during anaphase (Cimini, 2008). The microenvironment within the micronucleus is different to that of the primary nucleus. There are defective DNA damage responses which do not induce cell cycle arrest and DNA is replicated asynchronously with the primary nucleus (Crasta *et al.*, 2012). This means that in some micronuclei, premature chromosome condensation occurs before the DNA has been fully replicated (Johnson and Rao, 1970). This leads to extensive genome shattering. Repair mechanisms within the micronucleus attempt to religate the shattered fragments of DNA and form a derivative chromosome (Figure 1.1). Some chromosomal fragments may not be incorporated into the

derivative chromosome and these are lost. Other fragments may be ligated together but not incorporated into the derivative chromosome, instead they may form extrachromosomal circular DNA fragments called double minute chromosomes. These double minutes can become subsequently amplified (Stephens *et al.*, 2011). While these micronuclei can persist for several cell cycles, they are occasionally reincorporated into the main nucleus (Crasta *et al.*, 2012), particularly if the derivative chromosome has retained a centromere (Huang *et al.*, 2012).

Chromothripsis is a common occurrence in cancer, with some cancer types exhibiting chromothriptic events at a frequency greater than 50% (Cortés-Ciriano *et al.*, 2020). For example, 60–65% of pancreatic cancer (Notta *et al.*, 2016) and 32% of oesophageal adenocarcinoma have evidence of chromothripsis (Nones *et al.*, 2014). It has been shown to cause initiating driver mutations, such as 3p loss and 5q gain in clear cell renal cell carcinoma (Mitchell *et al.*, 2018). There is huge variation in the number of breakpoints involved in the chromothriptic event (Cortés-Ciriano *et al.*, 2020). The rearrangement events inactivate tumour suppressor genes and activate oncogenes which progresses the cell toward the tumourigenic phenotype. The regions that are frequently lost and gained in chromothripsis are the same regions that are lost and gained in cells of the same tumour type without chromothripsis, suggesting that alterations gained from chromothripsis provide a selective advantage to cells (Voronina *et al.*, 2020).

Chromothriptic rearrangements can alter genes in multiple ways (Stephens *et al.*, 2011). Tumour suppressor genes on the chromothriptic allele may not be incorporated into the derivative chromosome. However, in order for the tumour suppressor to be fully inactive, the other allele also needs to be inactivated, by mechanisms including point mutation or loss. For example, chromothriptic rearrangement on one allele and mutation of the other allele can deactivate CDKN2A in chordomas (Stephens *et al.*, 2011) or SMAD4 in pancreatic cancers (Notta *et al.*, 2016). A by-product of chromothripsis can be the formation of double minute chromosomes (Shoshani *et al.*, 2021). These double minute chromosomes may contain oncogenes that provide a selective advantage and may get amplified in subsequent cell divisions. For example, MYC can be amplified up to 200-fold in double minutes that were formed during a chromothriptic event (Stephens *et al.*, 2011). Another example is the *MYCN*, *MET* and *GLI2* amplifications in Sonic-Hedgehog subtype medulloblastoma which occur in double minute chromosomes produced as a result of the chromothriptic event (Rausch *et al.*, 2012). Furthermore, chromothripsis can cause juxtaposition of two genes in the same orientation with an open reading frame, leading to fusion transcripts. 44.8% of 2,493 gene fusions identified from RNA-seq data for 455 The Cancer Genome Atlas (TCGA) tumours have at least one gene partner in a chromothripsis region (Cortés-Ciriano *et al.*, 2020).

61% of breakpoints in high-confidence chromothripsis regions occur in intronic/exonic regions (Cortés-Ciriano *et al.*, 2020), but there is no apparent tissue-specific recurrence of gene targeting. Canonical fusions transcripts are also often not seen on chromothriptic chromosomes, so many of these fusion events may not be driving the cancer (Stephens *et al.*, 2011). Chromothripsis can also lead to the expression of different gene isoforms through truncation of genes (Kahles *et al.*, 2018). It can move a gene out of its normal regulatory context causing aberrant expression. There are many underlying mechanisms for this epigenetic dysregulation and these are discussed in the next section.

Chromothriptic chromosomes have previously been reconstructed (Garsed *et al.*, 2014). In one study, a chromothriptic event caused the formation of a neochromosome, with two to four copies per cell. The structure of the chromosome was extremely complex with a highly amplified core region and variation between copies of neochromosomes outside this core. This study determined the size and structure of these chromosomes, including presence of recurrent SNVs, indels and SVs. Further functional characterisation was also done, involving gene expression, promoter methylation and fusion genes but higher-order chromatin conformation was not studied. To date, this was the most comprehensive characterisation of a chromothriptic chromosome.

1.3.3 Other complex rearrangements

Since chromothripsis causes extensive rearrangement of primary genome sequence, it is a useful model to investigate the relationship between DNA sequence and higher-order chromosomal structuring. However, other complex rearrangements also alter the primary sequence and therefore provide additional insights into this relationship.

Extensive rearrangement occurs after repair of co-occurring double-strand DNA breaks, in a process known as chromoplexy. These chains of balanced interchromosomal and intrachromosomal rearrangements occur over a few events. While it was first described in prostate cancers (Baca *et al.*, 2013), this chromosome shuffling process has been reported in multiple cancer types and frequently rearranges disease-specific genes (Demeulemeester *et al.*, 2018). Unlike chromothripsis, chromoplexy breakpoints are unclustered and involve multiple chromosomes. There are also often fewer breakpoints than chromothriptic rearrangements and chromoplexy can occur multiple times in tumour development (Shen, 2013).

Breakage-fusion-bridge cycles (McClintock, 1938) are an active mutational mechanism in some cancers (Gisselsson *et al.*, 2001). In anaphase, fused sister chromatids are pulled apart asymmetrically meaning that one daughter cell will inherit a chromosome with a terminal deletion and the other cell will inherit a chromosome with a terminal inverted duplication. Both cells inherit chromosomes with exposed DNA ends and this process repeats until a

telomere is regained. The final result is a chromosome with many regions with increased copy number (Kinsella and Bafna, 2012).

Some cancer chromosomes have clusters of two to ten structural variants that are restricted to a local region (Li *et al.*, 2020b). There are many examples of these types of rearrangements in cancer genomes, including in liver hepatocellular carcinoma and oesophageal adenocarcinoma. An intuitive example is a translocation followed by a deletion of some of the adjacent DNA with an insertion of another region of DNA at the break site. This inserted sequence may be local to the breakpoint suggesting a copy-and-paste mechanism or distal to the initial breakpoint suggesting a cut-and-paste mechanism (Li *et al.*, 2020b).

Other examples include duplication–inverted triplication–duplication structures (Carvalho *et al.*, 2011), duplicated regions surrounding inverted segments and copy number loss regions with duplicated and inverted segments in the same region (Li *et al.*, 2020b). These regions of copy number gain are thought to be generated by a single process acting with a copy-and-paste mechanism. If these all occur locally then the process is a local n-jump but if the copied segments are found throughout the genome, then cycles or chains of insertions are occurring.

Complex rearrangements can also be classified based on copy number of the rearrangment junctions. This partitions rearrangements into rigma, pyrgo or tyfonas (Hadi *et al.*, 2020). Rigma rearrangements result from the accumulation of deletion events within genomic regions that contain large and late-replicating genes. Pyrgo rearrangments result from a series of nested duplication events, often involving super-enhancer regions. Finally, tyfonas are highly-amplified segments that result from the self-assembly of multiple sequences which are joined and duplicated over multiple sequential cell cycles (Hadi *et al.*, 2020).

1.4 Epigenetic alteration

The genome is hierarchically organised at a variety of scales (Serizay and Ahringer, 2018). Methyl groups can be added to cytosines to form 5-methylcytosine by DNA methyltransferases. They can be removed by DNA demethylases. The addition of a methyl group can inhibit transcription factor binding or can cause proteins that contribute to gene silencing to be recruited. DNA methylation most commonly occurs on cytosines in CpG sites. When these CpG sites are found in isolation, they are often methylated (Bird *et al.*, 1985). However they can also exist in regions of DNA with high CpG density, known as CpG islands, and in this context, they are often unmethylated. CpG islands often contain promoters (Saxonov *et al.*, 2006) and the methylation state of the cytosines in the promoter-containing CpG islands are vital in regulation of gene expression.

In order for a gene to be expressed, RNA polymerase needs to be recruited to the transcriptional start site. Promoters are located very close to the start site and facilitate this binding. Simply the presence of a promoter is sufficient to recruit the RNA polymerase complex but transcription is often weak without interactions from distal regulatory elements (Shlyueva *et al.*, 2014). Juxtaposition with regulatory elements, such as enhancers, occurs through DNA folding and looping. The enhancers contain binding sites for transcription factors and can facilitate transcriptional activation irrespective of the distance and orientation between the enhancer and promoter (Banerji *et al.*, 1981). Enhancers can also recruit histone modifying enzymes or chromatin remodelling complexes to increase chromatin accessibility around target genes and allow proteins to bind; a process essential for gene activation in eukaryotes (Clapier and Cairns, 2009). Enhancers themselves often have high levels of histone displacement (Mito *et al.*, 2007) and are often enriched for active histone marks, including H3K4me1, H3K4me2 and H3K27ac (Heintzman *et al.*, 2009). Histone marks are denoted by which histone protein is affected (H number), then by which amino acid residue that becomes altered and this is commonly lysine (K number), and then by the type and number of modifications present, for example me3 represents three methylation groups.

Cluster of hyper-active enhancers that work cooperatively to control gene networks are known as super-enhancers. They are frequently bound by master transcription factors and Mediator leading to the stable formation of the transcription pre-initiation complex and subsequent RNA polymerase II regulation (Allen and Taatjes, 2015). Super-enhancers have higher levels of active chromatin marks and are more frequently bound by chromatin remodelers when compared to enhancers (Khan *et al.*, 2018). Together, the high activity and large size of super-enhancers leads to the formation of distinct regulatory units where the overall activity of the super-enhancer is greater than the sum of the contained enhancers.

The fundamental unit of chromatin is a nucleosome which is composed of 145-147 bp of DNA is wrapped around a histone octamer (Luger *et al.*, 1997). These nucleosomes are connected via linker DNA and further organised into chromatin fibres. These fibres can be densely compacted to form condensed chromatin (Tremethick, 2007). The amino acids on the nucleosomes can be modified post-translationally to alter this level of compaction (Bannister and Kouzarides, 2011), from euchromatin to heterochromatin. Histones can be post-translationally modified by at least 80 covalent alterations including methylation, phosphorylation, acetylation, and ubiquitylation (Kouzarides, 2007; Zhang *et al.*, 2015b).

Nucleosome positioning in relation to genomic sequence is also important in gene regulation as nucleosomes sequester DNA from other DNA binding proteins. In the absence of other factors, the underlying DNA sequence also determines the propensity for histones to bind DNA. This affinity ranges over three orders of magnitude (Thåström *et al.*, 1999) and is

based on the entire 145-147 bp sequence binding to the histone octamer, rather than just a few base pairs like most DNA-protein interaction. There is often depletion of nucleosomes at regulator sequences, such as promoters and enhancers, so that protein binding can occur (Struhl and Segal, 2013). The overall positioning and compaction of nucleosomes, which are themselves determined by many factors, determines the overall DNA accessibility within a region. The more accessible a region, the more transcriptionally active it is (Henikoff, 2008).

At a larger scale, DNA is organised into topologically associating domains (TADs) which allow regions of the genome which are far apart in linear space to interact in 3D space. These interactions functionally compartmentalise large portions of the chromosome and interactions within domains are preferential to interactions across domains. These interactions occur through chromatin looping (Szabo *et al.*, 2019). The boundaries of these sites are often occupied by CCCTC-binding factor (CTCF) which acts as an insulator between adjacent TADs (Lupiáñez *et al.*, 2015). CTCF is also important as both a mediator of chromatin looping (Splinter *et al.*, 2006) and in preventing aberrant spreading of opposing chromatin marks into adjacent TADs, for example spreading of active marks into a repressive TAD (Cuddapah *et al.*, 2009).

Finally, interphase chromosomes themselves exist in distinct regions of the nucleus known as chromosome territories (Cremer and Cremer, 2010). Territories containing gene-rich chromosomes, for example chromosome 19, are less compact than territories containing gene-poor chromosomes, for example chromosome 18 (Croft *et al.*, 1999). However the chromosomes in these territories are highly dynamic in their compaction and this is essential for the genes within these chromosomes to be carefully regulated.

It is this hierarchical structuring that causes different genes to be expressed in different cells and allows a single fertilised egg to give rise to all cell types in the human body. Gene regulation is likely influenced at all scales; however, this influence is not fully understood and there are likely to be overlapping influences acting on every gene. Mutations, such as those seen in cancer genomes, can alter this epigenetic regulation at all scales (Jones and Baylin, 2002).

1.4.1 Mutations and gene expression

The mutation rate of non-coding regions of the genome is almost double the mutation rate of the coding genome, however the role of these mutations is poorly understood (Weinhold *et al.*, 2014). While single base substitutions and small indels rarely cause non-coding driver mutations (Rheinbay *et al.*, 2020), structural variants are a more common mechanism of generating regulatory dysfunction relevant to cancer. Structural rearrangements can directly affect gene expression by juxtaposing two transcripts in frame to produce a fusion gene or by

deleting or translocating part of a gene, thereby silencing it. However, mutations can also occur directly in promoters leading to altered gene expression. For example, while mutations in the coding region of telomerase are rare, the *TERT* promoter regulates telomerase activity and is mutated in more than 70% of melanomas (Huang *et al.*, 2013). This causes increased telomerase expression, often by the formation of a novel binding motif for ETS family transcription factors.

Structural variation can also occur in enhancers or lead to enhancer hijacking (Karnuta and Scacheri, 2018; Northcott *et al.*, 2014). Since genes are often under the control of multiple enhancers, structural variation involving enhancers may affect multiple genes (Hobert, 2010). Deletions can cause gene silencing by the removal of active enhancers or can lead to gene activation by deleting intervening regions of DNA and bringing an enhancer closer to a promoter (Lupiáñez *et al.*, 2015). Insertions can cause novel binding sites for transcription factors or disrupt binding of transcription factors to enhancers (Laurell *et al.*, 2012). Inversions and translocations can move enhancers away from their cognate promoter and allow them to regulate other promoters (Benko *et al.*, 2009; Lettice *et al.*, 2011; Northcott *et al.*, 2014; Watson *et al.*, 2016). Duplications can upregulate or downregulate transcription by creating novel enhancer sequences (Hyon *et al.*, 2015).

While many of these structural variants affecting enhancers lead to developmental disease, they are also commonplace in cancer genomes. In Burkitt's lymphoma, there is often a translocation that places *MYC* under the control of immunoglobulin heavy chain enhancers leading to upregulation of *MYC* (Schmitz *et al.*, 2014). Copy number gains of enhancers surrounding *KLF5*, *USP12*, *PARD6B* and *MYC* are seen in head and neck squamous cell carcinoma, colorectal carcinoma, liver hepatocellular carcinoma and lung adenocarcinoma. This leads to upregulation of those key oncogenes (Zhang *et al.*, 2016). In leukaemia, a common translocation conferring poor prognosis places a GATA2 transcription factor enhancer upstream of the *EVI1* proto-oncogene. Some tumours also acquire clusters of highly active super-enhancers around key oncogenes, for example near *MYC* in multiple myeloma (Lovén *et al.*, 2013).

Much remains unknown about the function of enhancers and there is no consensus on their hallmark properties (Kolovos *et al.*, 2012). The high mutation rates, extreme structural variation and regions of aneuploidy found in cancers, particularly those that have undergone chromothriptic events, will commonly lead to alterations affecting regulatory regions. Therefore, using cancers as models will allow further insight into the mechanisms underlying these regulatory elements and how removal from their normal context affects the gene expression landscape.

1.4.2 Mutations and chromatin accessibility

Histone modifications alter the accessibility of chromatin. The complement of all histone modifications is thought to form an epigenetic code which determines the structure and function of chromatin regions (Jenuwein and Allis, 2001). There are often alterations in chromatin accessibility and histone modifications seen during tumour development and progression. These alterations are caused by mutations in epigenetic readers, writers and erasers and chromatin remodeling complexes. Mutations in histone genes can also promote cancer development (Lu *et al.*, 2016) and epigenetic regulatory proteins are frequently mutated in cancers, highlighting the importance of epigenetic dysregulation as a central factor in tumour initiation and development.

There are many examples of mutations in chromatin remodelers. While point mutations can activate or inactivate these remodelers, structural variants can juxtapose them under the control of other regulatory elements. Many cancers, including breast, bladder and prostate, overexpress *EZH2* (Kleer *et al.*, 2003; Varambally *et al.*, 2002; Weikert *et al.*, 2005). Since EZH2 is a histone methyltransferase, this overexpression likely results in elevated levels of H3K27me3. However, in some myeloid malignancies, *EZH2* is inactivated (Ernst *et al.*, 2010; Nikoloski *et al.*, 2010). This dichotomy is striking but may be explained the need of histone modifications to be carefully regulated. Perturbing the levels of H3K27me3 in either direction can lead to aberrant expression of genes and subsequent cancer progression (Bannister and Kouzarides, 2011). KDM6A demethylates H3K27me3. Since some cancer have mutations in *KDM6A* (Van Haaften *et al.*, 2009), there is further support for the idea that H3K27me3 levels need to be carefully regulated.

Histone acetyltransferases and histone deacetylases are also overexpressed in some cancers (Halkidou *et al.*, 2004; Song *et al.*, 2005; Yang, 2004). In acute myeloid leukemia, the most common translocation is t(8;21). This juxtaposes DNA binding domains of RUNX1 to ETO which leads to transcriptional repression of RUNX1 targets, by mechanisms such as recruitment of histone deacetylase complexes by ETO (Zhang *et al.*, 2004). This leads to alteration in the overall chromatin accessibility of these cells. Notably, this change in chromatin accessibility is dependent on the other mutations within the genome. The other mutations in the genome determine which transcription factor binding sites have enrichment of accessibility peaks, suggesting the epigenetic landscape is dictated by the sum of multiple mutations (Chin *et al.*, 2020). Knockdown of *RUNX1-ETO* or upregulation of *C/EBP*α, a target of RUNX1-ETO, leads to alterations in chromatin accessibility. It causes formation of peaks containing binding sites for C/EBPα, which are repressed by RUNX1-ETO (Loke *et al.*, 2018).

Nucleosome positioning can also be a contributing factor to cancer development. For example, nucleosomes can be remodelled in cancers so that there is nucleosome occupancy at transcriptional start sites (Lin *et al.*, 2007). Once nucleosomes occupy the transcriptional start site, there can be further recruitment of polycomb repressive complexes and DNA methyltransferases which leads to permanent silencing of the gene. This has been shown in leukaemia at PML-RARA target genes via H3K27me3 repression (Morey *et al.*, 2008). Alterations in nucleosome positioning can also occur through mutations in the SWI-SNF complex which is estimated to be mutated in more than 20% of cancers (Kadoch and Crabtree, 2015). For example, mutations in the BAF47 subunit can lead to inactivation of p16 and p21 in rhabdoid tumours (Chai *et al.*, 2005).

Since alterations in the histone modifications and accessibility are inherently reversible, therapeutic interventions targeting chromatin remodelers are an appealing approach for cancer management. For example, histone deacetylase inhibitors are effective at preventing tumour growth and triggering differentiation (Sharma *et al.*, 2010). However, there is still a lot unknown about the dynamic landscape of histone modifications and what regulates which sites are modified. It is likely that different cancers will be more or less reliant on specific epigenetic modifications and what governs this is currently poorly understood. The fact that some combinations of epigenetic therapies are less effective than the therapies in isolation highlights the need for a better understanding of the mechanistic effects of epigenetic modifiers (Prebet *et al.*, 2014). In order to understand what is happening in disease contexts, a better understanding of the relationship between genome sequence and regulation of these marks is needed. Cancer genomes provide an opportunity to better understand this relationship as both alleles may exist with different histone modifications due to allele-specific mutations.

1.4.3 Mutations and topologically associating domains

Topologically associating domains (TADs) are largely conserved across cell types (Dekker and Heard, 2015). These domains encompass all the regulatory elements acting on the genes within the domain (De Laat and Duboule, 2013). Within these TADs, gene expression and histone modifications are correlated. Therefore, alteration of domain boundaries by structural variants can lead to altered gene regulation and subsequently altered gene expression (Lupiáñez *et al.*, 2015).

Within TADs, structural variants may lead to disruption of CTCF-CTCF chromatin loops (Hnisz *et al.*, 2016). Across TADs, structural variants may lead to disruption of TAD boundaries and if the boundary is between two transcriptionally different domains, this can lead to fused repressed–active domains and upregulation of gene expression in the repressed

domain (Dowen *et al.*, 2014). This is by the formation of a neo-TAD (Franke *et al.*, 2016). For example, a deletion affecting TAD boundaries between an active and repressed domain in lymphoma led to a 37-fold upregulation of *WNT4* which was previously in the repressed domain. This upregulation was not seen in samples that did not contain the boundary deletion (Akdemir *et al.*, 2020a). The alteration of TAD boundaries can lead to enhancer hijacking which causes dysregulation of neighbouring genes (Franke *et al.*, 2016; Lieberman-Aiden *et al.*, 2009). It is important to note that this effect is not universal and only 14% of the boundary affecting deletions between active and repressed domains led to alteration of gene expression.

Deletions tend to be within TADs and duplications tend to span different TADs. (Akdemir *et al.*, 2020a). In different cancer types, specific chromosomes tend to have more structural variants affecting TAD boundaries. For example, in oesophageal adenocarcinomas, chromosome 17 tends to have the most boundary affecting structural variants (Akdemir *et al.*, 2020a). This may be due to underlying pathogenesis driving each cancer type.

Domains containing active histone marks are highly correlated with enrichment in open chromatin peaks whereas domains containing repressive histone marks are enriched with DNA methylation (Akdemir *et al.*, 2020b). In both normal and cancer genomes, these domain marks can be used in isolation to predict somatic mutational burden, with the most dramatic change at the boundaries between two transcriptionally different domains rather than transcriptionally similar domains (Akdemir *et al.*, 2020b). The effect is domain-wide and not restricted to the genic regions suggesting that higher-order chromosome structuring is an influencing factor in somatic mutation rate.

The distinct mutational processes active in TADs are also different for those that have active marks and those that have inactive marks. Mutations caused by exogenous DNA damage agents, for example from UV light, tend to accumulate more in inactive domains. Conversely, deficiencies in DNA damage repair pathways, such as base excision repair or nucleotide excision repair, tend to accumulate more in active domains. Highly clustered regions of mutations, so called kategis-like loci, contain C > T and C > G substitutions which occur at TpCpN trinucleotides in single stranded DNA (Nik-Zainal *et al.*, 2012; Roberts *et al.*, 2012). These loci are enriched in transcriptionally active domains suggesting that the 3D structure could affect the persistence of single stranded DNA, which then is susceptible to hypermutation.

Since it is clear that chromatin configuration has a defining role in where specific mutational processes are active, it raises questions about how fixed that is. Large-scale structural variants can cause alterations in higher-order structures leading to altered gene regulation, histone marks and the formation of neo-TADs. It is conceivable that these structural rear-

rangements then alter the mutational processes acting on these regions. Many questions arise from this. A comprehensive investigation into the effect that large-scale rearrangements have on these epigenetic marks when compared to the marks seen before the rearrangement event will begin to elucidate some of these regulatory mechanisms.

1.5 Chromothripsis as a model

While chromothripsis produces extensive genomic rearrangements in a single catastrophic event, many of these rearrangements will not contribute to tumour initiation or progression and therefore are labelled passenger mutations (Stratton *et al.*, 2009). Identifying which rearrangements are tumour promoting and which are passengers is difficult (Pon and Marra, 2015), particularly in cases of chromothripsis where there is a high density of breakpoints in a localised region.

However, the high density of rearrangements, regardless of whether they are driver or passenger mutations, provides a unique opportunity to query more fundamental biological questions. The rearrangements generated by chromothripsis provide an interesting model whereby we can study how altering the primary genome structure affects higher-order structures. Since chromothripsis tends to affect only one of the two alleles, in every cell there is a rearranged allele and an internal control: the wild-type chromosome representing the sequence, structure and epigenetic landscape before reshuffling and the chromothriptic derivative chromosome representing the changes in landscape due to the reshuffling. This system will allow investigations into which epigenetic patterns are conserved and which are lost when comparing the wild-type and chromothriptic derivative chromosomes. These observations will also allow investigations into how these epigenetic alterations and rearrangements can impact gene expression. In order to do this, an unprecedented number of sequencing technologies will be needed to generate highly accurate chromothriptic and wild-type assemblies and query the effects of taking regions of the genome out of their normal context.

Chapter 2

Materials and Methods

2.1 Organoid culture

Five oesophageal adenocarcinoma organoid cell lines, from four different patients, were derived by the Cellular Genotyping and Phenotyping (CGaP) team at the Wellcome Sanger Institute using a previously described method (Li *et al.*, 2018). Lines were WTSI-OESO_103, WTSI-OESO_117, WTSI-OESO_143, WTSI-OESO_148 and WTSI-OESO_152. Two of the lines, WTSI-OESO_117 and WTSI-OESO_152, were derived from the same patient pre-chemotherapy and post-chemotherapy, respectively. All patients were male, two of which were Caucasian and two had unknown ethnicities. Patient were 61, 62, 68 and 82 years old. While cells in organoid culture accumulate single nucleotide variants at a low rate over long culture periods, they are mutationally stable at the structural level (Huch *et al.*, 2015). Consequently, to limit the number of culture-associated changes, cells were cultured for either a maximum of 30 weeks or below 30 passages from the point of banking, whichever occurred first. Organoids were passaged every 7-14 days, depending on the specific growth rate of the organoid.

2.2 Sequencing and associated protocols

2.2.1 Genomic sequencing

Multiple sequencing platforms were used. DNA extraction and sequencing was carried out by the core facilities at the Wellcome Sanger Institute. For HiSeq X Ten sequencing, the Qiagen DNA Maxi Kit protocol (Qiagen, Germantown, MD USA) was used to extract DNA. Short insert 500 bp genomic libraries were constructed, flow cells were prepared and clusters were generated using standard Illumina no-PCR library protocols (Illumina Cambridge Ltd,

Little Chesterford, UK). The Illumina HiSeq2500 platform was used to generate 150 bp paired-end whole genome sequencing with 32-43x coverage. Matched normal blood cells were also sequenced using this methodology.

For the 1M PacBio continuous long read (CLR) sequencing, 10X Genomics linked-read sequencing and BioNano optical mapping, high molecular weight DNA was extracted using the BioNano plug-based preparation protocol (BioNano, San Diego, CA USA). Libraries were sequenced on the PacBio Sequel, 10X Chromium and the BioNano Saphyr platforms, respectively. DNA for the 8M CLR and Circular Consensus Sequencing (CCS) PacBio sequencing and Oxford Nanopore sequencing was extracted using a Qiagen MagAttract extraction protocol and then sequenced on the PacBio Sequel and MinION, respectively. The CCS reads were size selected at 12 kb. Coverage for PacBio CCS and CLR reads ranged from 46-72x and 35-101x, respectively. Coverage for linked-reads ranged from 30-35x.

Karyotypic data was also generated for all cell lines by the Karyotyping facility at the Wellcome Sanger Institute. Cells were arrested in metaphase using 100ng/ml colcemid for three hours and 20 metaphase cells were karyotyped per sample.

2.2.2 Epigenomic sequencing

The Hi-C libraries were produced by the Research and Development team at the Wellcome Sanger Institute using the Dovetail Genomics Hi-C library preparation kit (Dovetail Genomics, Scotts Valley, CA USA) and the crosslinking for cell lines protocol. DpnII was used to fragment the DNA. Flow cells were prepared and libraries were sequenced on the HiSeq X Ten using recommended protocols. This was done by core facilities at the Wellcome Sanger Institute. Total coverage generated was 115-121x.

For ChIP-seq, DNA was crosslinked using the Diagenode (Denville, NJ, US) cell fixation protocol. Formaldehyde and subsequently glycine were added directly to organiods in basement membrane extract, type 2 (BME2). Organoids were then dissociated using TrypLE and stored in fixation buffer. Chromatin prepration, ChIP optimisation and shearing were then performed by Diagenode. Chromatin was prepared using the iDeal ChIP-seq kit for Transcription Factors protocol and immunoprecitation was done using antibodies against H3K4me3, H3K27me3, H3K27ac and CTCF. Libraries were sequenced on the Illumina NovaSeq 6000 to generate 100 bp paired-end sequencing reads. Each mark had 7-9x coverage.

For ATAC-seq, the Fast-ATAC protocol was followed (Corces *et al.*, 2016). Cells were incubated with Tn5 transposase, which fragments DNA in open chromatin regions and adds adapters simultaneously. Cells were subsequently lysed using digitonin. Library preparation was done using qPCR-only Nextera Dual indexing, flow cells were prepared and DNA was

sequenced on a HiSeq V4 to produce 75 bp paired-end sequencing reads. The coverage was 6-7x.

For Iso-Seq, RNA was extracted by core facilities at the Wellcome Sanger Institute using the Qiagen AllPrep DNA/RNA/miRNA Universal Kit (Qiagen, Germantown, MD USA). The extracted RNA had a RIN $\geq$ 7. An Iso-seq library was constructed from the extracted RNA using the Iso-Seq Express Template Preparation for Sequel and Sequel II Systems protocol. These libraries were sequenced by core facilities at the Wellcome Sanger Institute on the Sequel II with one SMRT cell per sample leading to 7-12x coverage.

2.3 Sequence alignment

GRCh38 was used as the reference genome. The HiSeq X Ten reads, ATAC-seq reads, ChIP-seq reads and the Hi-C reads were aligned using the Burrows-Wheeler algorithm (BWA mem) (v0.7.17-r1188) (Li and Durbin, 2009). The ChIP-seq adapters were removed using cutadapt (Martin, 2011) and the ATAC-seq adapters were removed using skewer (Jiang *et al.*, 2014). The Hi-C reads were aligned as single-ended reads. The HiSeq X Ten reads, ATAC-seq reads and ChIP-seq reads were aligned as paired-end reads. The PacBio CCS, CLR and Iso-seq reads were aligned using NGMLR (v0.2.7) (Sedlazeck *et al.*, 2018), after CCS consensus sequences were generated for the CCS and Iso-seq reads using ccs (v5.0.0). Primers and polyA tails in the Iso-seq reads were removed using lima (v2.0.1) and isoseq3 refine (v5.0.0), respectively, as part of the PacBio SMRTAnalysis software. The linked-reads were aligned using LongRanger align (v2.2.2). The BioNano reads were aligned using the BioNano software, RefAligner. Comparison of sequences to the reference was done using BLAT version 36x2 (Kent, 2002).

2.4 Genomic variant calling

In the short reads, single nucleotide variants (SNVs) were called using CaVEMan (v1.11.1) (Jones *et al.*, 2016) and filtered on PASS variants, ASMD $\geq$ 140 and CLPM equal to 0. Indels were called using Pindel (v2.2.5) (Ye *et al.*, 2009). Structural variants (SVs) were called using BRASS (v6.2.0) (Nik-Zainal *et al.*, 2016) and validated by the BRASS implementation of local assembly. In the HiSeq X Ten matched normal blood cell sequencing, germline variants were called using Strelka2 (Kim *et al.*, 2018) and SVs were called using both BRASS (v6.2.0) and GRIDSS (v2.11.1) (Cameron *et al.*, 2017).

In the PacBio CCS and CLR reads, SVs were called using Sniffles (v1.0.9) (Sedlazeck *et al.*, 2018). The final haplotype-resolved call set was filtered to remove germline variants

present on both haplotypes, variants that were in the normal blood cells called using BRASS (Nik-Zainal *et al.*, 2016) or GRIDSS (unmatched) (Cameron *et al.*, 2017) and variants called in repeat regions or decoy sequences as these regions are more difficult to map to and therefore calls are less confident. SNVs and indels were called using DeepVariant (Poplin *et al.*, 2018).

The SVs in the BioNano reads were called using the internal BioNano pipeline.

2.5 *De novo* haplotype resolution

WhatsHap (Patterson *et al.*, 2015) was used to phase germline heterozygous variants and generate phase blocks on the chromosomes using only CCS reads. Heterozygous variants in were called in matched normal blood and were filtered to remove regions of low complexity so only the most confident single nucleotide variant calls were used.

On the chromothriptic chromosome in WTSI-OESO_103, an initial consensus set of haplotype-unresolved SVs were generated from the SVs called by Sniffles in the CCS and CLR reads. These SVs were validated in the raw CCS and raw CLR reads. SVs without support were removed, while those that were not called by Sniffles but were found and validated in the raw reads were rescued. These SVs were used to assign phase blocks to the wild-type and chromothriptic chromosomes, with the expectation that the chromothriptic chromosome would have more structural variants than the wild-type chromosome.

The chromothriptic regions in the other samples had lower density of SVs than WTSI-OESO_103 and therefore the structural variants were less informative for phasing. Instead, variant allele fractions (VAFs) of heterozygous single nucleotide polymorphisms (SNPs) and local read depths were generated for each phase block and were used to assign blocks to the appropriate allele. VAFs and read depth were also used to complement the SV-based phasing for WTSI-OESO_103.

Together, structural variants, VAFs and regions of loss of heterozygosity (LOH) present on either allele allowed phase blocks with informative features to be confidently assigned. Where there were no informative variants, phase blocks were randomly assigned.

2.6 Genomic assembly methods

To determine the best assembly tool for these cancer genomes, four tools were trialled: Canu (v2.0) (Koren *et al.*, 2017), Falcon (v0.0.8) (Chin *et al.*, 2016), wtdbg2 (v3.3) (Ruan and Li, 2020) and hifiasm (v0.7) (Cheng *et al.*, 2021). Assemblies were produced by Dr Zemin Ning. Separate assemblies were produced for each allele using haplotype-resolved reads to

generate an initial assembly for the chromothriptic and wild-type chromosomes. In most cases, hifiasm was used. In cases where assemblies from hifiasm were insufficient, wtdbg2 was used. The initial assemblies were then scaffolded using 3D-DNA (v180922) (Dudchenko *et al.*, 2017). This was done by Dr Edward Harry. The scaffolds were then further corrected by chromosome assembly using synteny (caus3D) (v1.0) which corrects scaffolding errors generated by 3D-DNA using the original CCS assembly. This was done by Dr Zemin Ning. These error-corrected assemblies were then used as the basis for subsequent analysis.

2.7 Assembly-based haplotype resolution

Hi-C reads, ATAC-seq reads, ChIP-seq reads and Iso-seq reads were all haplotype resolved using alignment-based methods and the custom assembly generated for each chromosome. The reads were aligned to the assemblies using the appropriate algorithm described above. All reads were aligned to both the chromothriptic assembly and the wild-type assembly. The most appropriate mapping for every read was determined. Ordered by priority, criteria used for assigning reads was:

1. Presence of a SNP that was assigned to either assembly using WhatsHap
2. Read mapping to a region of LOH in either assembly
3. Mapping quality for the read or read pair is higher on either assembly
4. Insert size between read pairs differs from the expected size

Regions without SNPs and SVs where sequences were identical in both assemblies would provide identical mapping scores and therefore when identical mapping was seen, reads were randomly assigned.

For the Hi-C reads, reads were assigned to insert size groups. If the insert size group was the same for both alignments and none of the above criteria were met, then the reads were randomly assigned. If the insert size groups were different, a weighted probability was used to determine which chromosome the reads were most likely to be derived from. If either insert size was over 75 Mb, reads were assigned to the chromosome where the alignment was not over 75 Mb. Groups and probabilities can be found in table 2.1 and were based on the underlying insert size distribution of the Hi-C read pairs.

For Iso-seq reads, read alignments allowed for splicing of transcripts.

Table 2.1 Probability assignment

Insert size (Mb)	Rounded probability (%)
0-0.001	54.0
0.001-1	34.0
1-3	4.4
3-5.5	1.4
5.5+	6.4

2.8 Calling topologically associating domains

Hi-C read interactions and topologically associating domains (TADs) were called using Juicer (Durand *et al.*, 2016). The restriction enzyme used to make Hi-C libraries was DpnII and therefore the interactions were normalised around DpnII restriction binding sites. TADs were called at the following resolutions: 2.5 Mb, 1 Mb, 500 kb, 250 kb, 100 kb, 50 kb, 25 kb, 10 kb and 5 kb. Hi-C matrices were visualised using JBrowse2. Both reads assigned based on the criteria described above and randomly assigned reads were used in order to ensure that the overall TAD structures could be seen.

2.9 Peak calling in ChIP-seq and ATAC-seq data

ATAC-seq and ChIP-seq peaks were called using MACS2 (Zhang *et al.*, 2008), with biological repeats peaks being merged using MSPC (Jalili *et al.*, 2015) where appropriate. Broad peak calling was used for H3K27me3 reads and narrow peak calling was used for all other ChIP-seq marks and for ATAC-seq reads. Differential binding was called using DiffBind (Ross-Innes *et al.*, 2012). In order to do this, reads were mapped back to GRCh38 using BWA mem (Li and Durbin, 2009). Only reads mapping to regions of the genome that are present in both assemblies and reads which were explicitly assigned were included. Regions only present in one assembly will give a false positive signal for differential binding since signal will be due to absence of the region in the other assembly and not due to true differential binding. Reads which are randomly assigned may give a false negative signal and can mask real differential binding in a region.

2.10 Identifying differential transcript expressions

Raw Iso-seq reads were processed using IsoSeq v3 (https://github.com/PacificBiosciences/IsoSeq). To identify differences in transcripts between the two chromosomes, only reads that

could be explicitly assigned to a chromosome in regions which were present in both assemblies were used. cDNA_Cupcake (v24.3.0) (https://github.com/Magdoll/cDNA_Cupcake) was used to collapse transcripts and SQANTI (https://github.com/ConesaLab/SQANTI) was used to classify the transcripts using the GENCODE version 37 (Ensembl 103) annotation file for GRCh38. DESeq2 (Love *et al.*, 2014) was used to determine which transcripts were differentially expressed (Wald test, q value < 0.05). Reads were mapped back to GRCh38 using NGMLR (Sedlazeck *et al.*, 2018) and, again, only reads in regions present in both assemblies that were explicitly assigned were used. Gene set enrichment analysis was done using gprofiler2 (Kolberg *et al.*, 2020) using only the assignable portions of the chromosome.

2.11 Gene essentiality

Cells were first transduced with lentiviral vector carrying Cas9. Those that were stably expressing Cas9 were subsequently transduced with a single guide RNA (sgRNA) library using a previously described method (Iorio *et al.*, 2018). Each gene had on average 5 guides. The guides were based on genes from the GRCh37 reference genome and essential genes were identified using CRISPRcleanR (https://github.com/francescojm/CRISPRcleanR).

2.12 Integrated analysis

To visualise multiple data types concurrently, JBrowse 2 was used (https://jbrowse.org/jb2/).

Chapter 3

Generating a cancer-specific reference genome

3.1 Chapter highlights

This chapter describes a methodology for reconstructing cancer genomes. These genomes vary in the level of complexity of structural variation, from normal chromosomes to simple rearrangements to chromothripsis events. The main messages from this chapter are:

1. Complex regions of structural variation between alleles in cancer genomes coupled with subclonal variants means haplotype-aware *de novo* assemblies are essential for contiguous cancer genome assemblies.

2. Phase blocks are assigned to a specific haplotype using B-allele frequencies (BAFs) of single nucleotide polymorphism (SNPs) and presence of structural variants (SVs).

3. Contiguous assemblies are produced even on chromosomes with over 900 structural rearrangements compared to the reference genome.

4. All types of structural variation are reconstructed, other than tandem duplications which are collapsed by current assembly tools.

3.2 Introduction

The human reference genome is an invaluable tool for studying human biology and disease. However in some diseases, such as cancer, the genomes of the cells vary substantially from the human reference genome. In cases where variation is small, using the reference genome

will allow differences to be studied in the greater genomic and epigenomic context. However in cases with complex variation, such as extensive chromothripsis, information about newly formed genomic and epigenomic structures is much more difficult to untangle. To study newly-formed structures in these cases, it is necessary to reconstruct the exact sequence of a cancer genome rather than comparing variation to a known reference. That is, to produce a *de novo* genome assembly.

Genome assembly becomes more difficult in complex and repetitive regions and these regions are abundant in the human genome. Cancer genomes often also have subclonal variation and copy number alterations and this makes assembling cancer genomes increasingly difficult. Heterogeneous regions of subclonality mean that only the dominant clone is reconstructed, unless multiple assemblies are generated. Regions of copy number variation make determining the exact genome sequence problematic. When the copy number oscillates between 1 and 2 with only one allele containing the rearrangements, an unambigious underlying sequence solution can be determined (Figure 3.1A). This is because there is a single solution which will connect all breakpoints once. However when there are multiple copies of segments, the same copy number profile can result from many different underlying genomic sequences (Figure 3.1B). The true solution can be found if a single read spans all regions of copy number alteration, if there are occasional unique bases in the amplified regions or by combining information in multiple reads. However, there may still be ambiguity in the final sequence and often this manifests as the final genome assembly being more fragmented.

This chapter aims to address these problems and develop a methodology that allows cancer-specific genome assemblies to be generated for cancers which have undergone extensive genome reshuffling. While the main objective was to generate assemblies in regions of complex variation, the entire cancer genome was assembled. As such, the method is robust to variable levels of rearrangement: from normal chromosomes to simple rearrangements through to chromothripsis.

Assemblies for five oesophageal adenocarcinoma organoids were generated, each with different genomic features that provide different biological insights but also presented a range of technical problems. As a basis for these assemblies, I used multiple long-read sequencing technologies including PacBio sequencing and Hi-C chromosome capture. To my knowledge, this is the first time cancer-specific genome assemblies have been generated for chromosomes which have undergone extensive rearrangement. These assemblies will be used in subsequent chapters to gain insight into the role of the primary genome sequence in higher-order structuring of the genome. Regions around rearrangements may have altered histone modifications, accessibility and topologically associating domains (TADs). This in turn may lead to alterations in gene expression and subsequently the cancer phenotype.

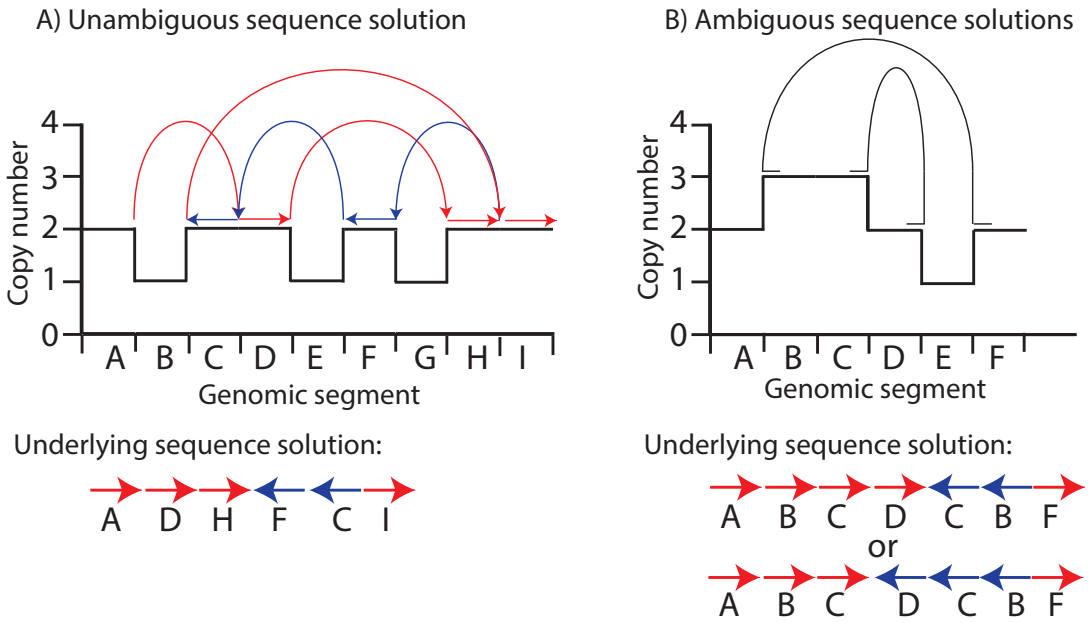


Fig. 3.1 Schematic representations of theoretical copy number plots. Only one allele is affected by the rearrangements so the other allele will be wild-type and copy number one in all segments. Curved lines represent read pair evidence for regions that are joined. Red arrows represent forward orientation joining, blue arrows represent reverse orientation joining and black lines are ambiguous. For the black lines, the location of the read pairs are shown by a horizontal line to denote which segments are joined. A) Unambiguous sequence solutions occur for rearrangements where there has been no sequence amplification. There is only one path that will encompass all segments just once. B) An ambiguous sequence solution occurs due to the regions of increased copy number. Two possible solutions are present and it is impossible to determine which is the true solution without further information.

In order to understand the impact of the rearrangements, cancer-specific references will be imperative.

3.3 Selection of organoids

Short-read Illumina HiSeq X Ten sequencing from 98 cell lines were screened in order to select organoid models with evidence of chromothripsis. These organoids were derived, cultured and sequenced by the Cellular Generation and Phenotyping (CGaP) facility at the Wellcome Sanger Institute using the protocol described previously (Li *et al.*, 2018). They spanned three cancer types: oesophageal adenocarcinomas, colorectal adenocarcinomas and colorectal liver metastases. Of these types, the oesophageal adenocarcinoma organoids had the highest incidence of chromothripsis. Therefore, five organoids derived

from oesophageal adenocarcinomas with evidence of chromothripsis were selected for this research: WTSI-OESO_103, WTSI-OESO_117, WTSI-OESO_143, WTSI-OESO_148 and WTSI-OESO_152.

These organoid models fit the previously defined criteria for chromothripsis (Korbel and Campbell, 2013) (Table 3.1 shows where evidence can be found). The main chromothriptic chromosomes in each sample can be seen in Figure 3.2. The chromothriptic chromosomes in WTSI-OESO_103 and WTSI-OESO_152 show oscillations over copy number states 1 and 2, characteristic of chromothripsis. On chromosomes 9 in WTSI-OESO_117, there are oscillations over copy number states 2 and 3 because a whole chromosome duplication of the wild-type allele had occurred. However the other two samples exhibit some regions which do not oscillate over just two copy numbers. On chromosome 18 of WTSI-OESO_143, there is oscillation over mainly two copy number states with a few small regions of amplification. Chromothripsis has been shown to be a major driver of double minute formation which leads to massive amplifications of small regions of DNA (Shoshani *et al.*, 2021). These double minutes may persist as extra-chromosomal segments or have been shown to reincorporate into chromosomes (Von Hoff *et al.*, 1990). The amplified segments on chromosome 18 of WTSI-OESO_143 are likely double minutes produced by the chromothriptic event. The genes in these regions are listed in Appendix A. On chromosome 1 in WTSI-OESO_148, the copy number oscillates mainly over three states as the wild-type chromosome has a large duplication and the chromothriptic chromosome has regions of amplifications that are likely to be from breakage-fusion-bridge prior to chromothripsis. The genes in these amplified regions can also be seen in Appendix A.

The organoids were selected by calculating the frequency of the minor allele relative to the minor and major allele combined (B-allele frequency, BAF) at sites of single nucleotide polymorphisms (SNPs). This allowed identification of copy number states along a chromosome to query whether chromothripsis was restricted to only one allele (Figure 3.3), another key chromothripsis characteristic. If the chromothripsis was present on both alleles, there would be regions of the chromosome which were homozygously deleted leading to no SNPs and no coverage in that area. This was not seen in the BAFs of the selected samples and there are no copy number 0 regions (Figure 3.2). Loss and retention of hetrozygozity (Figure 3.2), breakpoint clustering (Figure 3.2), randomness of DNA joins (Table 3.2) and randomness of fragment order (Figure 3.2) were seen in all organoid lines. Therefore, the criteria for chromothripsis was met. The organoids were derived from male patients aged 61 to 82 with varying levels of prior treatment (Table 3.3).

WTSI-OESO_103 was selected as the pilot sample to test initial methods on. The high density of structural rearrangements present in the chromothriptic region and the oscillation

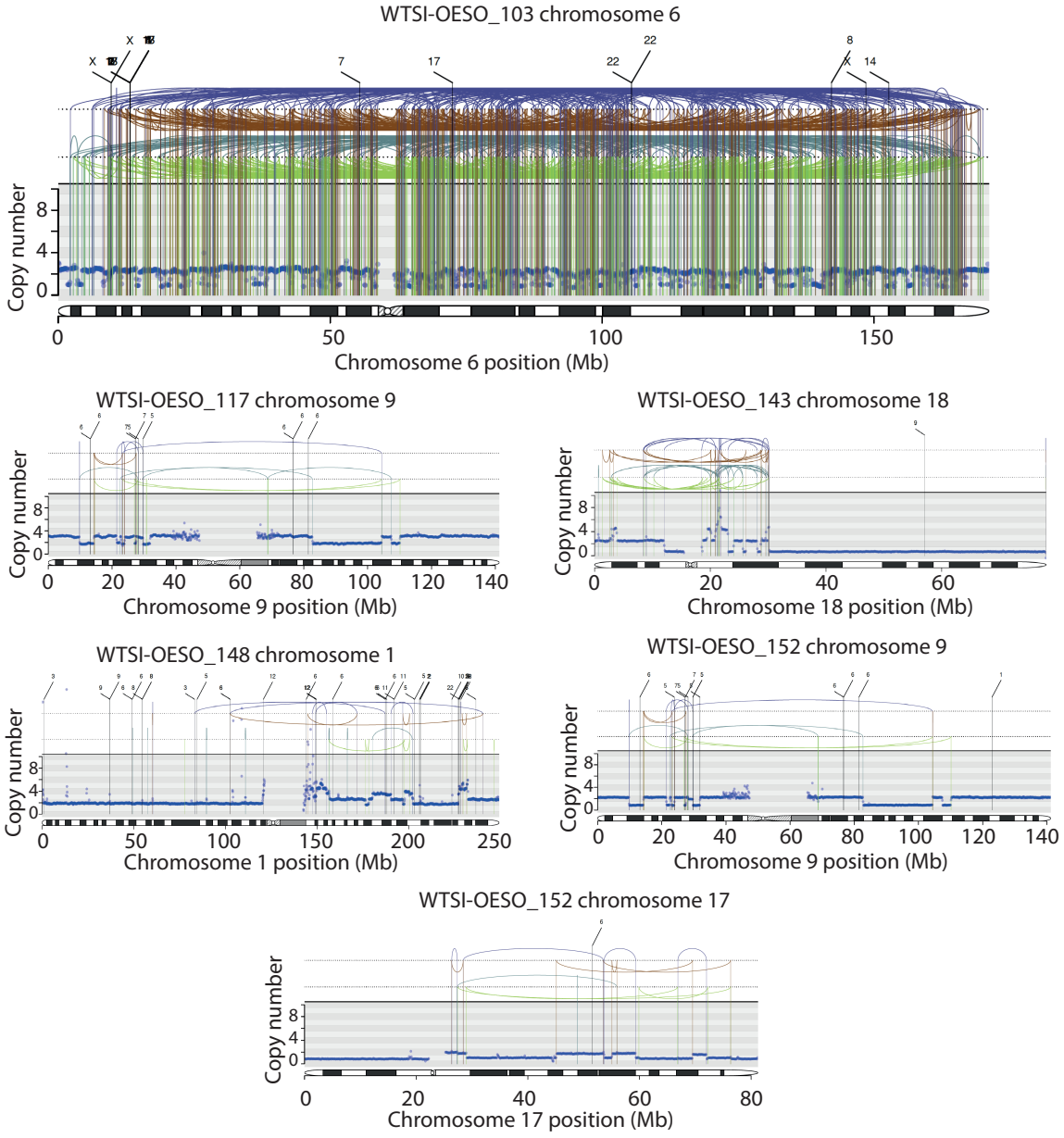


Fig. 3.2 Chromothriptic regions in each sample. Horizontal blue dots denote copy number. Vertical lines denote breakpoints: black represents translocations, green represents deletions, teal represents tandem-duplications, dark blue represents tail-to-tail inversions and brown represents head-to-head inversions. WTSI-OESO_152 has two chromothriptic chromosomes and all other samples have one chromothriptic chromosome. Copy number and variant calls were derived from Illumina X Ten sequencing. Key features of chromothripsis are displayed in these plots: breakpoints in these samples show clustering, copy number oscillations characteristic of chromothripsis are present and there is randomness in DNA fragment order.

Table 3.1 Evidence of chromothripsis

Criteria	Evidence
Breakpoints clustering	Fig 3.2
Oscillation of copy number states	Fig 3.2
Loss and retention of hetrozygozity	Fig 3.3
Rearrangements affecting one haplotype	Fig 3.2 & 3.3
Randomness of DNA joins	Table 3.2
Randomness of fragment order	Fig 3.2

Table 3.2 Randomness of DNA joins in chromothripsis

	Orientation of read pairs			
	- -	- +	+ -	+ +
WTSI-OESO_103 chr6	204	222	216	241
WTSI-OESO_117 chr9	4	3	5	5
WTSI-OESO_143 chr18	20	14	12	13
WTSI-OESO_148 chr1	18	16	19	12
WTSI-OESO_152 chr9	3	4	5	4
WTSI-OESO_152 chr17	2	4	4	3

Table 3.3 Clinical organoid data

Tumour Tissue	WTSI-OESO_103	WTSI-OESO_117	WTSI-OESO_143	WTSI-OESO_148	WTSI-OESO_152
Age	82	61	62	68	61
Cytotoxic Chemo	No	Yes	Yes	Yes	Yes
Ethnic Category	White British	Not known	Not known	White British	Not known
Gender	Male	Male	Male	Male	Male
Lines Therapy	0	0	2	2	2
Primary tumour	Yes	Yes	Yes	Yes	Yes
M Pathology	M0	M0	M0	M0	M0
N Pathology	N3	N2	N0	N0	N2
T Pathology	T3	T3	T3	T2	T3
Prior Drug	NA	Unknown	Unknown	Capecitabine	Capecitabine
Prior Drug	NA	Unknown	Unknown	Cisplatin	Dexamethasone
Prior Drug	NA	Unknown	Unknown	Epirubicin	NA
Therapy Outcome	NA	Unknown	Progressive disease	Partial Response	Stable disease
Sample collection	Surgical	Laproscopy	Surgical	Surgical	Surgical
Smoking Status	Unknown	Non-smoker	Non-smoker	Ex smoker	Non-smoker
TNM Stage	IIIC	IIIB	IIA	IB	IIIB

over mainly two copy number states in this region meant that reads spanning multiple SVs could be used to confidently join shattered segments together. This enabled the reconstruction of segments harbouring complex structural variation in the order of megabases (Mb) in length. This sample also contained other chromosomes with complex rearrangements as well as regions with a lower density of SVs. Therefore, a variety of reconstruction methods could be

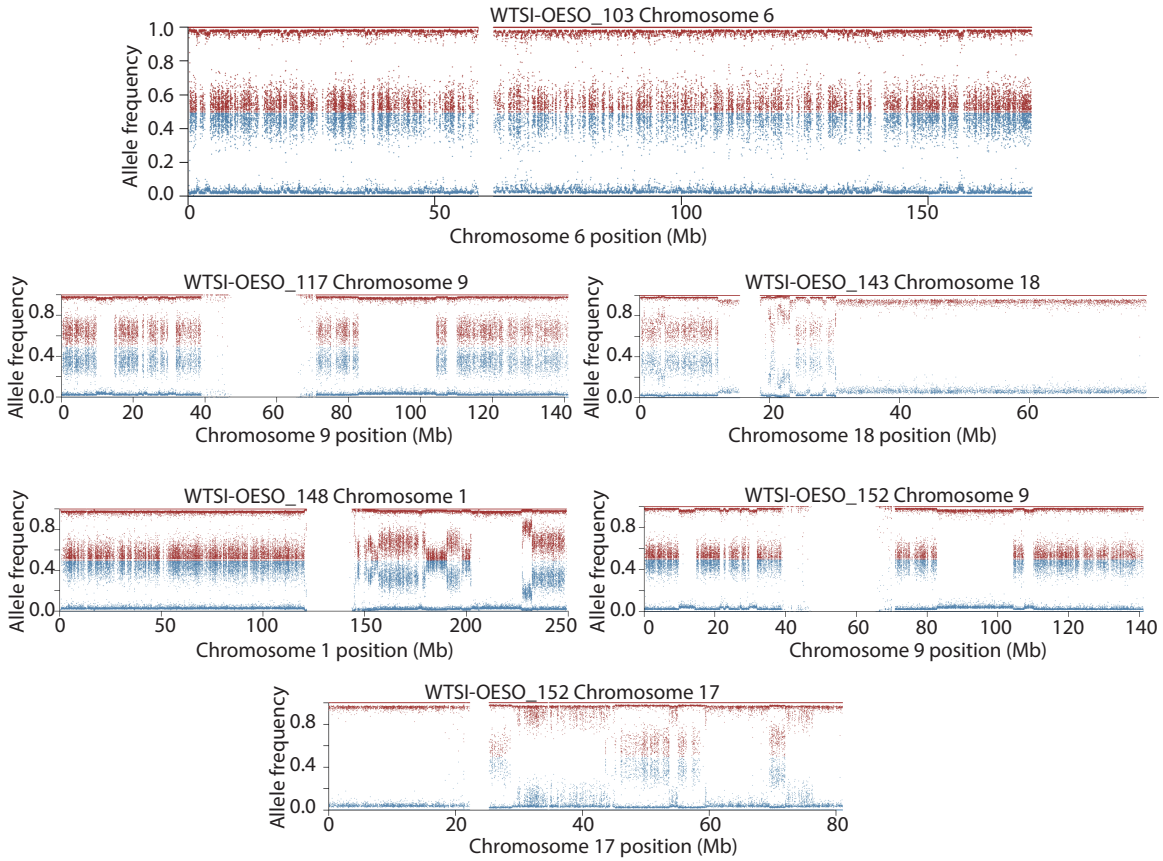
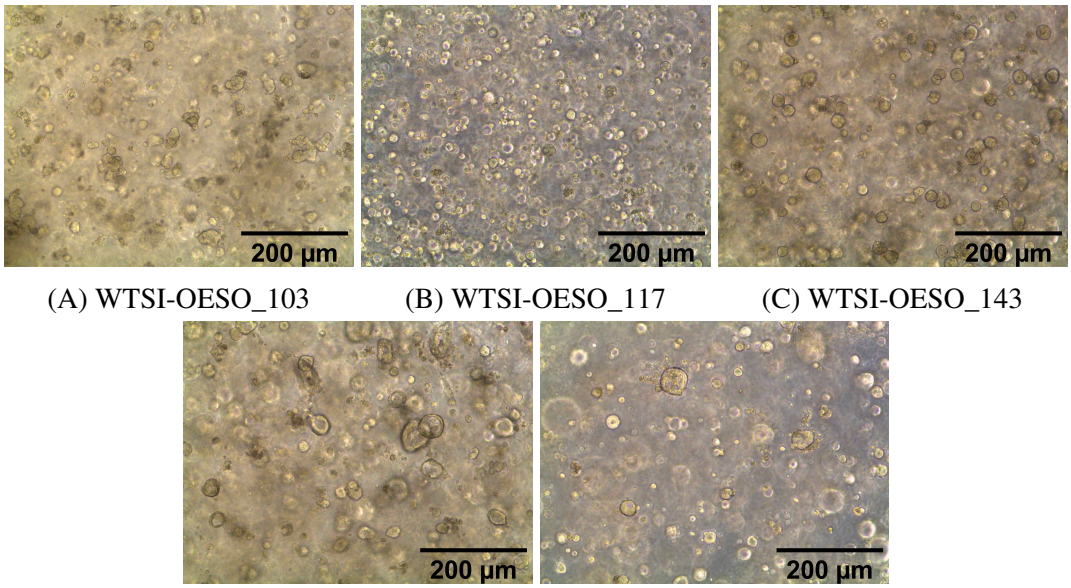


Fig. 3.3 BAF of chromothriptic region in each sample derived from Illumina X Ten sequencing. Red dots represent one allele and blue dots represent the other. Alleles are not phased so adjacent red and blue dots may be on opposite alleles. All samples show regions of loss and retention of heterozygosity

trialled on this sample. This would allow identification of the best reconstruction method for chromosomes with a variety of structural variation types, including chromothriptic, complex, simple and no rearrangements.

3.4 Growth of organoids

In all organoid lines, cells formed spheroid structures. However, the morphology varied between organoid lines (Figure 3.4). WTSI-OESO_117 and WTSI-OESO_143 formed small spheroid organoids. WTSI-OESO_148 formed larger organoids. The organoids formed by WTSI-OESO_103 and WTSI-OESO_152 were much less uniform and often exhibited protrusions.



(A) WTSI-OESO_103 (B) WTSI-OESO_117 (C) WTSI-OESO_143

(D) WTSI-OESO_148 (E) WTSI-OESO_152

Fig. 3.4 Organoid cell lines growing in Basement Membrane Matrix, type 2 (BME2) imaged in a widefield microscope at 10x magnification. Different samples have variable spheroid structures.

These organoids were sequenced on multiple sequencing platforms at varying coverage (Table 3.4). They also underwent optical mapping and karyotyping.

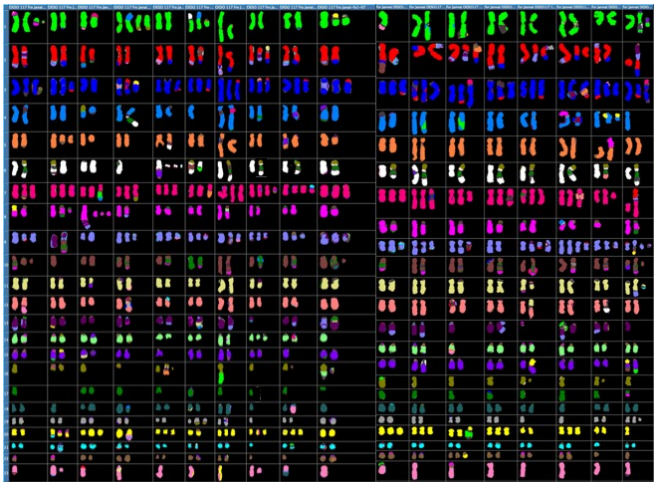
Table 3.4 Genome sequencing coverage

	Median sequencing coverage (x)				
Sequencing	WTSI-OESO_103	WTSI-OESO_117	WTSI-OESO_143	WTSI-OESO_148	WTSI-OESO_152
X Ten	38.6	32.0	40.4	42.8	41.6
PacBio CCS	45.5	72.3	71.7	71.7	54.3
PacBio CLR	41.8	68.6	101.0	46.4	35.6
Linked-reads	32.6	30.3	35.2	31.0	32.4
Hi-C	120.8	117.8	114.9	115.4	118.8

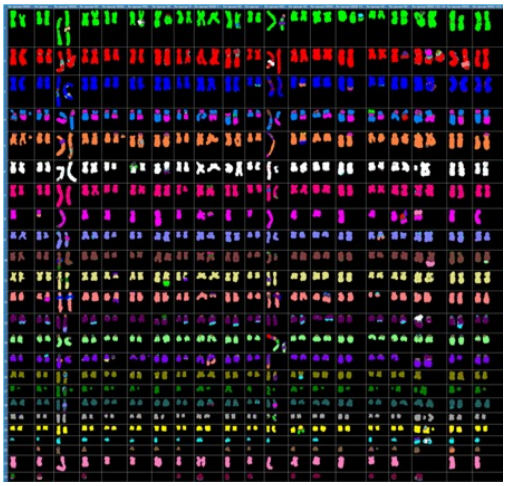
3.5 Karyotyping

Karyotypic data revealed that the organoids were highly variable in the total number of chromosomes and that subclonal rearrangements were present in all samples (Figure 3.5). The most common sources of variation were whole-chromosome duplication, whole-genome duplication and small fragments of chromosomes being duplicated. Often, rearrangements can be seen in multiple cells, even if they are not completely clonal. WTSI-OESO_117

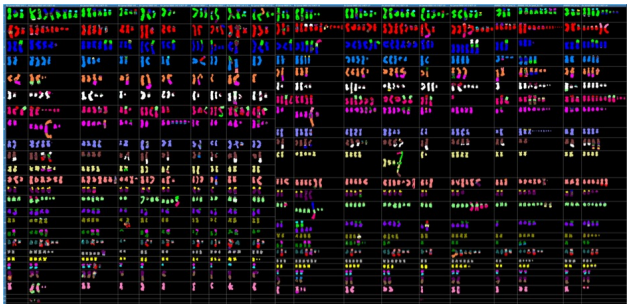
(A) WTSI-OESO_103



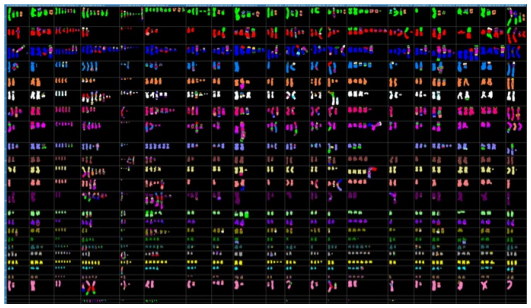
(B) WTSI-OESO_117



(C) WTSI-OESO_152



(D) WTSI-OESO_143



(E) WTSI-OESO_148

Fig. 3.5 Heterogeneity shown in karyotypic data. Each row is a separate chromosome, ordered by chromosome number. Each column is a different cell. 20 cells per sample underwent karyotyping. Different chromosomes are stained different colours so chromosomes with multiple colours show translocations.

and WTSI-OESO_152 had near-diploid genomes with fewer whole-chromosome and whole-genome duplications than WTSI-OESO_103, WTSI-OESO_143 and WTSI-OESO_148.

3.6 Haplotype-unaware assemblies

To determine the best assembly method for chromothriptic regions, chromosome 6 from WTSI-OESO_103, the chromosome with the highest density of SVs, was used as a basis for exploratory analysis. Reads were mapped to the GRCh38 reference genome and those mapping to chromosome 6 were extracted. All following assembly statistics described in this section are from WTSI-OESO_103, unless otherwise stated.

Initially, both reference-based assemblies and *de novo* assemblies were generated. For the reference-based assemblies, PacBio continuous long read (CLR) sequencing reads were aligned to the GRCh38 reference genome, chromosome 6 reads were selected, non-primary hit fragments were removed and gap5 (Bonfield and Whitwham, 2010) was used to generate a multiple sequence alignment database. This was done by Dr Zemin Ning. The consensus sequence was calculated based on these alignments. The generated assembly had 280 gaps. However, a reference-based approach may not be optimal for the reconstruction of chromothriptic chromosomes. Gaps may be caused by the chromothriptic chromosome no longer resembling the reference chromosome 6 due to the extensive structural variation. Instead, chromothriptic chromosomes contain large sequences identical to GRCh38 chromosome 6 joined to sequences located 100s-1000s of bases away in GRCh38 chromosome 6. These regions then become gaps in the assembly. Reads may also be incorrectly joined together based on reference position rather than the underlying sequence of the sample. To circumvent this problem, *de novo* assemblies can be produced.

Chromosome 6 aligned CLR reads were assembled using wtdbg2 (Ruan and Li, 2020) by Dr Zemin Ning. While wtdbg2 produced a more representative assembly than the reference-based methods, the assembly was still fragmented into 2,895 segments, known as contigs (Table 3.5). This high fragmentation is likely due to the highly convergent and divergent nature of the two alleles of chromosome 6 in WTSI-OESO_103. Many regions on the two alleles are similar for 100s of bases to Mb which results in an assembly with high confidence in that region. However, at a breakpoint where a structural variant occurred on one allele, the downstream sequences become divergent. Current assemblers have been built to expect small changes between the two alleles (Lischer and Shimizu, 2017), such as SNPs or sequencing errors. They do not explicitly account for large structural changes within a sample, particularly at the density seen in this chromosome. This causes the assembler to

stop assembling the contig as it does not know which sequence to progress down. Since these regions occur so frequently, the resulting assembly is highly fragmented.

In order to circumvent these problems, a haplotype-resolved approach was needed.

Table 3.5 Haplotype-unaware assembly

Assembly metric	*de novo* Assembly
Sum (bp)	189,680,514
Total contigs	2,895
Average length	65,520.04
Largest contig (bp)	1,195,726
N50 (bp)	229,295
L50	239
N90 (bp)	25,358
L90	1,219
Gaps	0

Assembly metrics: L90 is the smallest number of contigs that constitute 90% of the total assembly size. N90 is sequence length of the shortest contig at which 90% of the total assembly size is reached. L50 and N50 have the same definitions as L90 and N90 but for 50% of the genome or assembly.

3.7 Haplotype-resolved *de novo* assemblies

In order to use haplotype-based approaches, sequencing reads needed to be haplotype-resolved so that a *de novo* assembly can be produced for each parental chromosome separately. This will produce contiguous assemblies, continuing through regions where the two parental haplotypes diverge.

3.7.1 Calling and phasing structural variants

Resolution of haplotypes in chromothriptic chromosomes with a high SV density requires comprehensive identification of unphased SVs present on the chromosome. For WTSI-OESO_103, this was chromosome 6. All reads were aligned to the reference GRCh38 using two long read aligners: NGMLR (Sedlazeck *et al.*, 2018) and minimap2 (Li, 2018). When comparing these alignments, reads were mapped in similar positions with variation due to differences in gap opening and extension penalties. SVs in the NGMLR-aligned reads and minimap2-aligned reads were called using Sniffles (Sedlazeck *et al.*, 2018) and PBSV (Wenger *et al.*, 2017), respectively. Variants called were broadly similar. NGMLR and

Sniffles were selected as the final aligner and SV caller based on the ability of Sniffles to output read name information, which PBSV was not able to do at the time. In total, Sniffles called 12,168 SVs in PacBio circular consensus sequencing (CCS) reads, 1,497 of which were on chromosome 6. These include both germline and somatic SVs and both groups are informative for generating the initial assembly. These SVs appear in at least 10 reads and on average in 25 reads and therefore, they were used as the initial call set.

These SVs were further validated using raw read data as follows. For every read in both the CCS and CLR reads, all alignments and therefore all SVs present in the read were identified. Then for each SV in the initial call set, the adjacent SVs were identified. For each read supporting an SV, there could be no adjacent SV, an adjacent SV present in the initial SV call set or an adjacent SV that was not identified. SVs which were not initially identified by Sniffles were then added to the SV list. From this the reads overlapping each SV were used to determine if, for the region of interest, at least 80% of supporting reads joined the same two segments. In some cases, a consensus could not be generated as a single break point was joined to multiple different regions. This may be the result of subclonality. Alternatively, it may due to local complexity making the exact breakpoint ambiguous. This may be particularly relevant in CLR reads due to the high base error rate. Regardless of the cause, these SVs had low confidence and were subsequently removed. This built a consensus SV call set from both the CCS and CLR reads, producing a list of 1,201 partially phased SVs. The initial list contained 1,497 SVs, 1,146 of which were validated, 351 of which were removed and 55 new SVs were added. Since the base accuracy in the CCS reads was high, highly accurate SV calls were generated from the CCS reads. The CLR reads gave further long-distance phasing and validation.

While the inbuilt phasing of Sniffles only phases variants which are found explicitly on a single read, this method allows partial phasing of SVs across reads. However, the CCS reads underwent a size selection at 12 kb meaning that phasing stopped in regions where the distance between two SVs was greater than the maximum read length in that region, often around 12 kb. This distance could be extended using the CLR reads spanning two SVs. A phase block is the length of a region of the genome where each heterozygous variant in that region can be assigned to an allele. The phase block ends when a variant occurs which cannot be confidently phased with the previous variant. The longest phase block generated using germline and somatic SVs was 30 kb.

3.7.2 Phasing reads

Phasing based solely on read-spanning SVs is insufficient to produce highly contiguous assemblies. In order to expand phase blocks, SNPs can also be incorporated. WhatsHap

(Patterson *et al.*, 2015) was used to phase SNPs. It is able to deal with variable sequencing error rates seen in different data types. Read depth was down-sampled to 15 x to retain the reads with the greatest number of SNPs while still remaining computationally feasible.

One of the main benefits of WhatsHap is its ability to incorporate multiple sequencing data types into the phasing. Thus, depending on which sources of data are used, the length and accuracy of the phasing varies (Table 3.6). The most intuitive metric of phasing success is the total number of phase blocks in which the genome, or chromosome in this case, is partitioned. The smaller the total number of blocks, the more contiguous the assembly and the better the phasing. Often the length of the longest phase block is also used as a metric of phasing success (Choi *et al.*, 2018). According to these metrics, the more data provided to WhatsHap, the better the phasing becomes. As expected, longer read lengths allow for more extensive phasing. Therefore, longer phase blocks were seen when using CLR reads rather than CCS reads. Interestingly, using linked-reads alone led to highly fragmented phase blocks despite the long-distance information inherently in these reads. Since the matched normal blood only underwent Illumina X Ten sequencing, the haplotype blocks in this sample were highly fragmented and not informative for further phasing.

Table 3.6 WhatsHap phasing

		Largest block variants	
	Total Number of blocks	Counts	Percentage (%)
CCS	1279	2092	1.4
CCS with no simple repeats	1264	2052	1.5
CLR	202	5206	3.5
Linked-reads	30512	885	0.8
CCS+CLR	205	5206	3.5
CLR,CCS+XTEN	928	49703	31.6
CCS,CLR,XTEN+linked-reads	863	119559	31.6
CCS,CLR,XTEN,linked-reads+Hi-C	134	125863	79.7
Blood XTEN	28408	775	0.6

WhatsHap randomly assigns the separated phase blocks to haplotype 0 and haplotype 1. For both haplotypes, the number of SVs, both germline and somatic, and the number of regions of LOH were determined. Phase blocks were assigned to the chromothriptic or wild-type alleles based on whether there was an increased number of SVs in one phase block over the other when compared to the average frequency of SVs in the rest of the genome. The haplotype with a greater number of SVs was assigned to the chromothriptic chromosome. Large scale deletions are also expected to occur more on the chromothriptic than the wild-type chromosome, providing a second method of determining which chromosome the haplotype

block was derived from if SV density was uninformative. Some phase blocks contained no SVs or large scale deletions and these were randomly assigned. Some regions were not phased by WhatsHap and hence reads in these regions were also randomly assigned.

Upon closer inspection, using multiple data types led to phase switching errors. Usually the easiest way to accurately identify phase switching is to use trio data. However, maternal and paternal samples were not available so this was not possible for this study. The chromothripsis in these samples is restricted to one parental allele, as evidenced by BAFs. This allowed identification of switch errors as, if phasing was correct, most of the SVs would be present on one of the two alleles. However this was not the case when combining multiple data types in order to phase reads. When CCS, CLR, XTEN, linked-reads and Hi-C reads were all used, multiple SVs were seen on both alleles within each phase block (Figure 3.6A and schematic in Figure 3.7A). This is due to phase switching within a haplotype block.

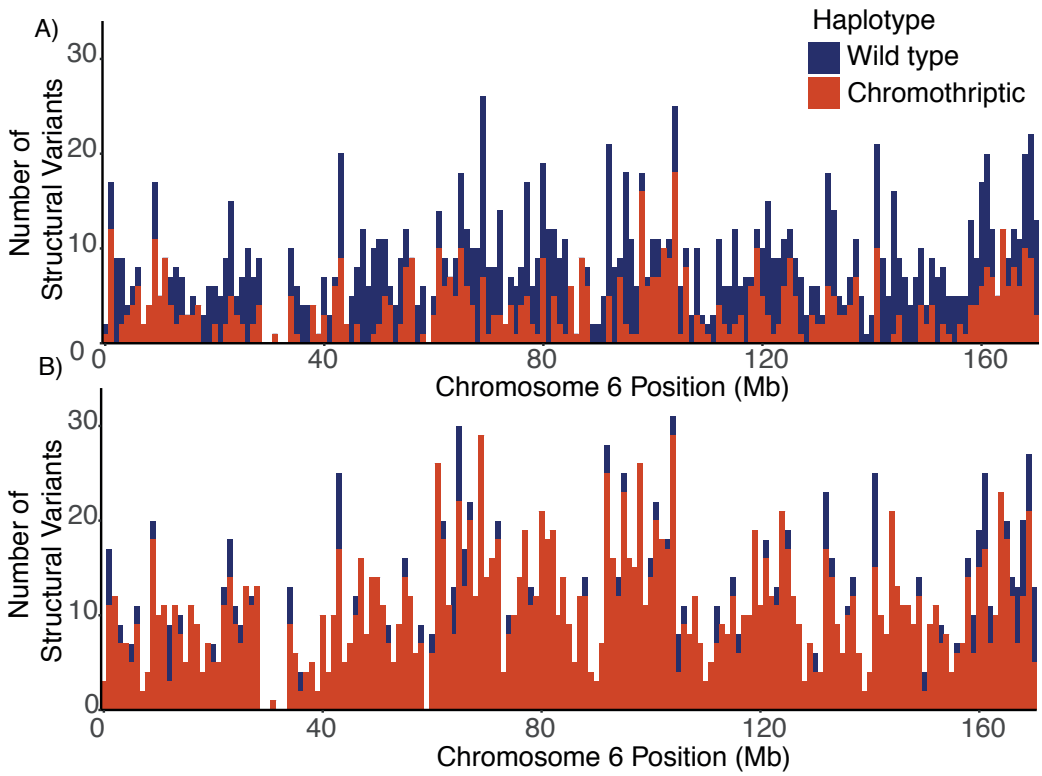


Fig. 3.6 Number of SVs seen on each haplotype, binned into 1 Mb bins. A) Initial SV counts using WhatsHap on CCS, CLR, linked-reads, XTEN and Hi-C reads has many SVs on both haplotypes. B) Final SV counts using only CCS data has SVs mainly on the chromothriptic chromosome. Minor differences in counts between plots are due to differences in which regions are assigned using different data types. SVs which are present on both alleles have been removed. When using multiple data types, SVs are erroneously assigned to both alleles and will be erroneously filtered out as germline homozygous SVs in later steps.

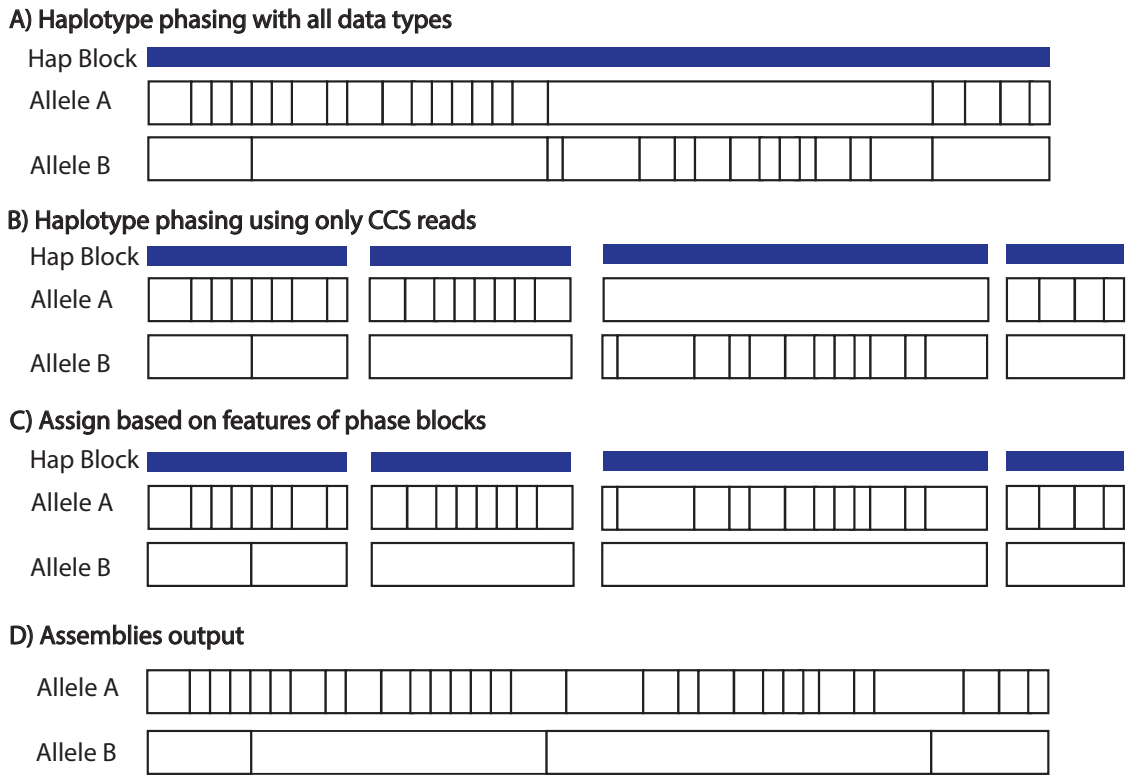


Fig. 3.7 Haplotype switching schematic. Haplotype switching can be identified using the distribution of SVs across haplotype blocks. A) When using multiple data types, both haplotypes can contain regions of high SV density followed by low SV density. B) When only using CCS reads, phase blocks are shorter but SVs are often restricted to one allele. C) SVs density can be used to group the haplotype blocks which derive from the chromothriptic allele together (Allele A). D) Reads from the high SV blocks and low SV blocks are assembled separately. Blue bars are haplotype blocks from whatshap. Lines within A and B demarcate SV boundaries. Other features can also be used to assign phase blocks.

Excluding simple repeat regions improved the phasing slightly. When excluding these regions and phasing only using CCS reads, the total number of phase blocks was 1,264 with the largest phase block containing 2,052 variants (1.5% of accessible variants). This is more fragmented when compared to the number of phase blocks when using multiple data types. However when examining these phase blocks, it was apparent that the SVs were restricted to one haplotype (Figure 3.6B) and schematic in Figure 3.7B). It is difficult to determine where exactly the switching occurs and as it may not be at an SV boundary. Consequently it is preferential to have smaller phase blocks which contain fewer switch errors than trying to resolve switch errors within reads. Other features, as described above, can then be used to assign allele A or allele B to the chromothriptic chromosome (schematic using SVs to

assign blocks in Figure 3.7C). Using this method and phasing only using CCS reads led to a reduction of switch errors and most SVs on one haplotype. Therefore, these blocks were used to generate subsequent genome assemblies (schematic in Figure 3.7D).

This methodology works well when there is a very high density of SVs on the chromothriptic chromosome. For WTSI-OESO_103, the chromothriptic phase blocks needed more than 3.16 times the number of germline and somatic SVs when compared to the wild-type phase blocks in order to be confidently assigned to the chromothriptic chromosome. This is because chromosome 6 had, on average, 3.16 times more structural variants than the average number of structural variants seen on other chromosomes. However the density of structural variants is much lower on the chromothriptic regions of other samples (Table 3.7) and on non-chromothriptic chromosomes.

Table 3.7 Structural variant density

	SV density per Mb	
Sample	Chromothriptic chromosome	Rest of Genome
WTSI-OESO_117 chr9	3.86	3.90
WTSI-OESO_103 chr6	6.77	2.14
WTSI-OESO_143 chr18	4.14	3.64
WTSI-OESO_148 chr1	3.79	3.65
WTSI-OESO_152 chr9	3.36	3.47
WTSI-OESO_152 chr17	3.94	3.47

For these chromosomes, VAFs and read depth were used in order to assign phase blocks to the appropriate allele. The copy number of the chromothriptic chromosomes was determined and VAFs of phase blocks generated by WhatsHap were assigned accordingly. If the average copy number of the chromothriptic chromosome was higher than the wild-type, then the higher VAF haplotype blocks were assigned to the chromothriptic chromosome. If the copy number of the wild-type chromosome was higher than the chromothriptic chromosome, then the higher VAF haplotype block was assigned to the wild-type chromosome, unless read depth in that region suggested that the higher VAF haplotype belongs on the chromothriptic chromosome. Read depth was also used in regions of complexity in order to best assign the VAFs to the most probable chromosome. In regions of LOH, the remaining allele was assigned to the wild-type chromosome. The overall copy number of the chromosome was determined using read depth in the LOH regions relative to the non-LOH regions. If the average copy number of the chromothriptic chromosome and the wild-type were equal, blocks were randomly assigned.

This method was also used for normal chromosomes and chromosomes with simple rearrangements. It is most effective when the two alleles have an asymmetric copy number, for example two copies of one allele and only one copy of the other. However, it still performed well in symmetric copy number regions with and without structural variation between the two alleles.

3.7.3 *De novo* assemblies

Table 3.8 Haplotype-aware initial wild-type assemblies

Assembly metric	wtdbg2	Canu	Falcon	hifiasm
Sum (bp)	147409805	154034761	14153210	148544969
Total contigs	420	619	1787	360
Average length (bp)	350975.73	248844.53	7920.10	412624.91
Largest contig (bp)	10454899	16896470	405002	10457240
N50 (bp)	1838477	1304051	10419	1475599
L50	20	23	468	20
N90 (bp)	171054	133883	4319	204185
L90	132	163	1307	123
Gaps	0	0	0	0

Table 3.9 Haplotype-aware initial chromothriptic assemblies

Assembly metric	wtdbg2	Canu	Falcon	hifiasm
Sum (bp)	120761587	128869853	10449542	122500881
Total contigs	695	801	1313	528
Average length (bp)	173757.68	160886.21	7958.52	232009.24
Largest contig (bp)	2822132	2827000	92443	2829437
N50 (bp)	447088	624957	11131	699270
L50	73	58	377	53
N90 (bp)	84331	64026	4379	99130
L90	298	291	976	229
Gaps	0	0	0	0

Assembly metrics: L90 is the smallest number of contigs that constitute 90% of the total assembly size. N90 is sequence length of the shortest contig at which 90% of the total assembly size is reached. L50 and N50 have the same definitions as L90 and N90 but for 50% of the genome and assembly.

A separate *de novo* assembly was produced for each allele by Dr Zemin Ning in order to produce more contiguous assemblies. This reduced some of the fragmentation seen in regions where the sequence of the two alleles diverged, i.e. at regions of structural variation.

Initially four assembly tools were used: wtdbg2 (Ruan and Li, 2020), Canu (Koren *et al.*, 2017), Falcon (Chin *et al.*, 2016) and hifiasm (Cheng *et al.*, 2021). Hifiasm produced the fewest number of contigs with the longest average length for both the chromothriptic and wild-type chromosomes (Table 3.8 and Table 3.9). Wtdbg2 produced the second fewest number of contigs and the second longest average length for both the chromothriptic and wild-type chromosomes. Canu outperformed hifiasm, wtdbg2 and Falcon in terms of bases covered in both the wild-type and chromothriptic assemblies, however these assemblies were more fragmented than hifiasm and wtdbg2 assemblies. Despite having haplotype-resolved reads, the Falcon assemblies were still highly fragmented for both haplotypes.

In order to decide which assembly provided the most accurate representation of the cancer genome, the wild-type assemblies were compared to chromosome 6 in the GRCh38 reference

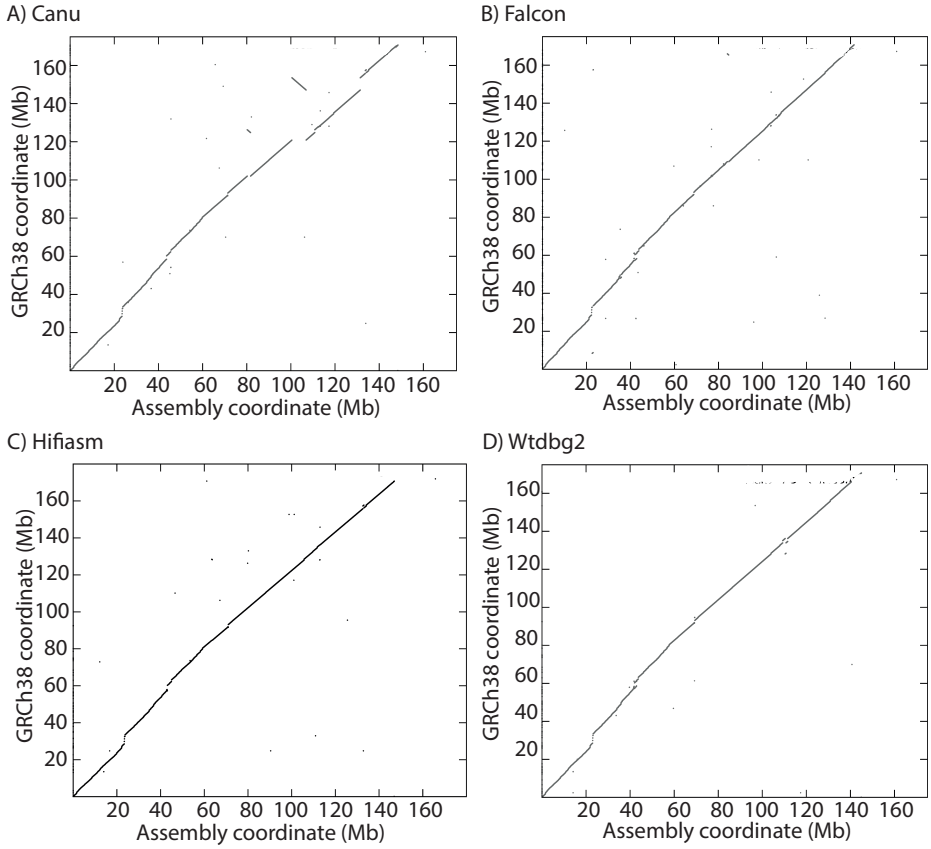


Fig. 3.8 Dot plot alignments of the wild-type assemblies (x-axis) against the reference GRCh38 chromosome 6 (y-axis). A sequence identical to the reference GRCh38 genome would appear as a 45° diagonal line.

genome (Figure 3.8). These wild-type assemblies should be grossly normal with very few rearrangements present. Therefore, they should be broadly consistent with the reference genome at the structural level. The SVs that are present should also be small events. The Falcon, hifiasm and the wtdbg2 assemblies all closely resembled the reference chromosome 6, shown in Figure 3.8 by the limited deviation of the assembly from the reference chromosome. The Canu assembly exhibited deviation from the reference genome between 100 Mb and 140 Mb (Figure 3.8), and this deviation is not see in the other assemblies, suggesting it is an assembly error produced by Canu.

3.8 Scaffolding using Hi-C reads

Since the wtdbg2 and hifiasm assemblies were the most contiguous with the longest average lengths, these assemblies were both scaffolded. Scaffolding was done by Dr Edward Harry using 3D-DNA (Dudchenko *et al.*, 2017) and haplotype-resolved Hi-C reads. 3D-DNA has three main steps. First misjoined contigs in the assemblies are identified and split based on inconsistencies in the signal seen in the Hi-C read alignments (Figure 3.9A,B). This results in contigs which have a continuous Hi-C signal. Next, these segments are used for iterative scaffolding based on Hi-C reads mapping to the ends of different contigs, with multiple rounds possible in order to generate the most likely solution (Figure 3.9C). Finally, evidence for overlapping sequences are queried in segments which have a similar Hi-C signal. If there is sufficient evidence, these segments are merged. This produces a scaffolded assembly (Figure 3.9D).

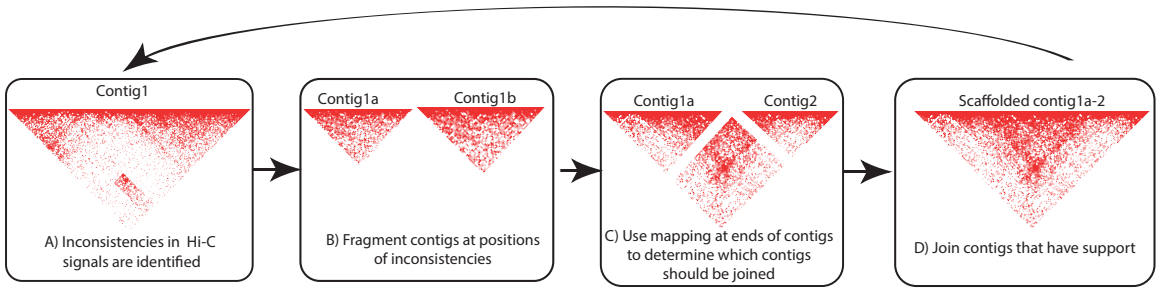


Fig. 3.9 Scaffolding using Hi-C reads follows 4 main steps. 1) Interactions that are occurring at high frequency outside of normal TAD structures are identified as discrepancies. 2) Regions with discrepancies are fragmented. 3) Interactions at the ends of all contigs are compared to look for support. 4) Contig with supported interactions are joined.

Before Hi-C reads can be used for scaffolding, they need to be haplotype-resolved. This was done in a reference-aware manner. All reads were aligned to the chromothriptic assembly

and all reads were aligned to the wild-type assembly. The most appropriate mapping for every read was determined as described in Chapter 2 and the Hi-C reads were separated into wild-type and chromothriptic reads. These reads were used by 3D-DNA to scaffold the initial CCS assemblies and many gaps were closed. In the wtdbg2 assemblies, 389 and 575 gaps were closed on the wild-type and chromothriptic assemblies respectively. In the hifiasm assemblies, 362 and 500 gaps were closed on the wild-type and chromothriptic assemblies respectively.

3D-DNA uses Hi-C coverage profiles to identify assembly errors by fragmenting the assembly at regions which have inconsistent Hi-C coverage. However this can lead to over-fragmentation of assemblies since regions may have inconsistent coverage due to difficulties

Table 3.10 Final wild-type assembly and scaffolding metrics

Assembly metric	CCS assembly	Scaffolds	Final assembly
Sum (bp)	148544969	148725969	148718969
Total contigs	360	54	12
Average length (bp)	412624.91	2754184.61	12393247.42
Largest contig (bp)	10457240	99230238	99925482
N50 (bp)	1475599	99230238	99925482
L50	20	1	1
N90 (bp)	204185	7746498	8732116
L90	123	5	4
Gaps	0	362	348

Table 3.11 Final chromothriptic assembly and scaffolding metrics

Assembly metric	CCS assembly	Scaffolds	Final assembly
Sum (bp)	122500881	122750881	122739881
Total contigs	528	176	50
Average length (bp)	232009.24	697448.19	2454797.62
Largest contig (bp)	2829437	20124411	20584246
N50 (bp)	699270	12049381	12151762
L50	53	5	5
N90 (bp)	99130	3663039	4094508
L90	229	12	11
Gaps	0	500	478

Assembly metrics: L90 is the smallest number of contigs that constitute 90% of the total assembly size. N90 is sequence length of the shortest contig at which 90% of the total assembly size is reached. L50 and N50 have the same definitions as L90 and N90 but for 50% of the genome.

in mapping to that region. 150 bp Hi-C reads are more affected by complex or repetitive regions that are difficult to map to when compared to the 12 kb CCS reads. In order to overcome this, Chromosome Assignment Using Synteny (CAUS) was used by Dr Zemin Ning. It rejoins regions which were broken by 3D-DNA but neither end was subsequently joined to anything else. This led to the final assemblies (Tables 3.10 and 3.11). In the wtdbg2-based final assembly, a total of 41 and 153 contigs were produced for the wild-type and chromothriptic assemblies, respectively. The hifiasm-based final assemblies were more contiguous than the wtdbg2-based final assemblies with a total of 12 and 50 contigs for the wild-type and chromothriptic assemblies, respectively. The hifiasm-based final assemblies also had a longer average contig length and a longer largest contig when compared to the wtdbg2-based final assemblies. This was true for both the wild-type and chromothriptic assemblies.

However, while contiguity is an important metric, it is also important that the assemblies resemble the underlying cancer genome sequence. In order to assess which assembly was a more accurate representation, the wild-type assemblies were compared to chromosome 6 in the GRCh38 reference genome (Figure 3.10). Since the wild-type chromosome should have very few SVs and those that are present should be small events, the assembly should closely resemble the reference chromosome. This was true for the hifiasm-based final assembly (Figure 3.10A), however there appeared to be some assembly errors around the

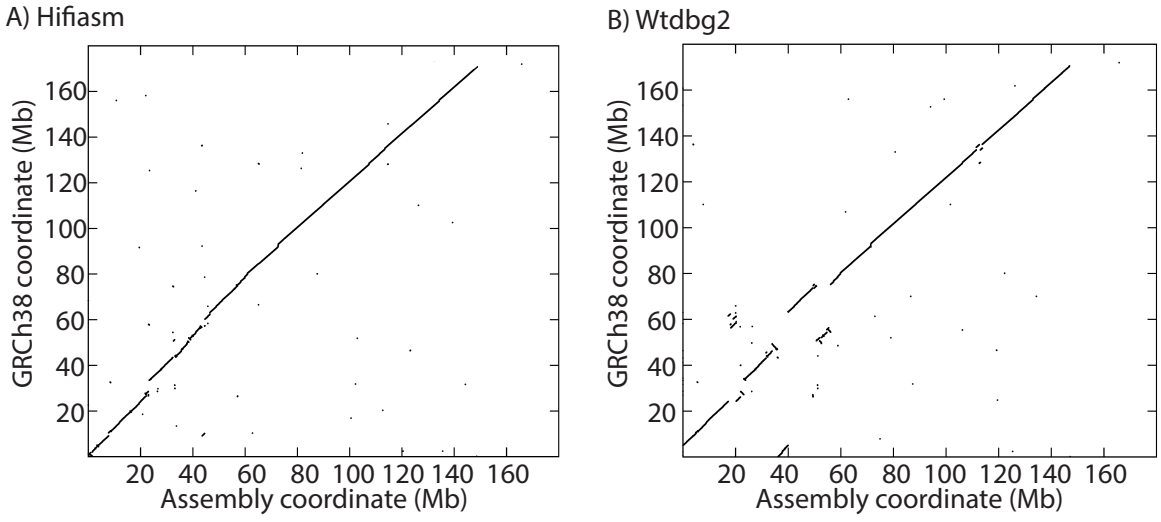


Fig. 3.10 Dot plot alignments of the wild-type assemblies (x-axis) against the reference GRCh38 chromosome 6 (y-axis). The wtdbg2 assembly has regions in the assembly which do not align to the same region of the reference GRCh38 assembly between 25 Mb and 80 Mb.

HLA locus and the centromere in the wtdbg2-based assembly. This, coupled with the increased contiguity, meant that the hifiasm-based assemblies were selected as the final assemblies for chromosome 6 in this sample. All other chromosomes in this sample and other samples were also assembled using hifiasm as the initial assembler.

This haplotype-oriented *de novo* assembly method led to contiguous cancer genome assemblies, even in highly rearranged chromosomes with, for example, over 900 rearrangements on a single allele. An overview of the main steps in the final assembly method can be seen in Figure 3.11. For WTSI-OESO_103, the wild-type assembly was slightly more contiguous than the chromothriptic assembly, with a total contig number of 12 rather than 50 (Tables 3.10 and 3.11). L90 is the smallest number of contigs that constitute 90% of the total assembly size. The smaller the L90, the more contiguous the assemblies are. The wild-type assembly also had a smaller L90 of 4 when compared with 11 on the chromothriptic assembly and a longer average contig length of 12,393,247.42 when compared with an average of 2,454,797.62 on the chromothriptic assembly.

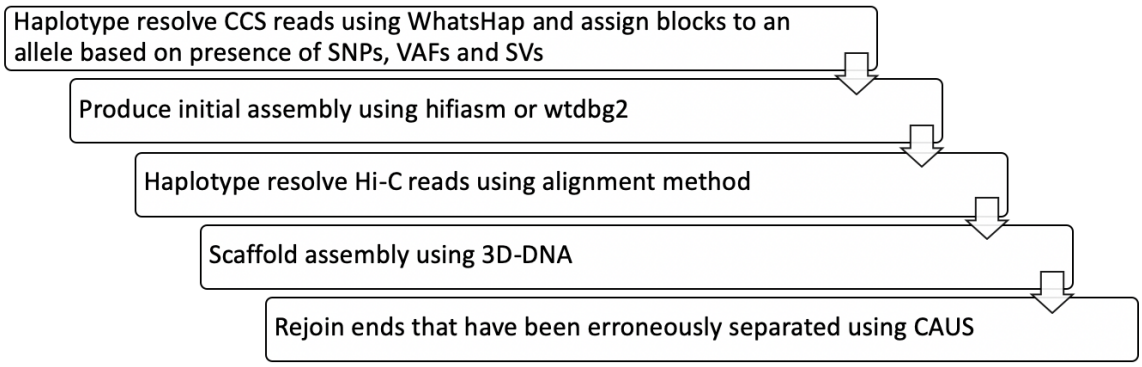


Fig. 3.11 Contigous haplotype-resolved cancer-specific genome assemblies were generated using 5 main steps.

These assemblies recapitulate the differences between the two alleles of chromosome 6 (Figure 3.12). As expected, the wild-type contigs are highly contiguous with large regions identical to the reference GRCh38 genome if allowing for SNPs and indels. Conversely, the chromothriptic contigs are derived from regions found throughout the reference GRCh38 genome, consistent with the breakage and repair caused by the chromothriptic event. This suggests the generated assemblies are highly representative of the specific sequence seen in these organoids and will be suitable for identifying differences between haplotypes in regions of interest.

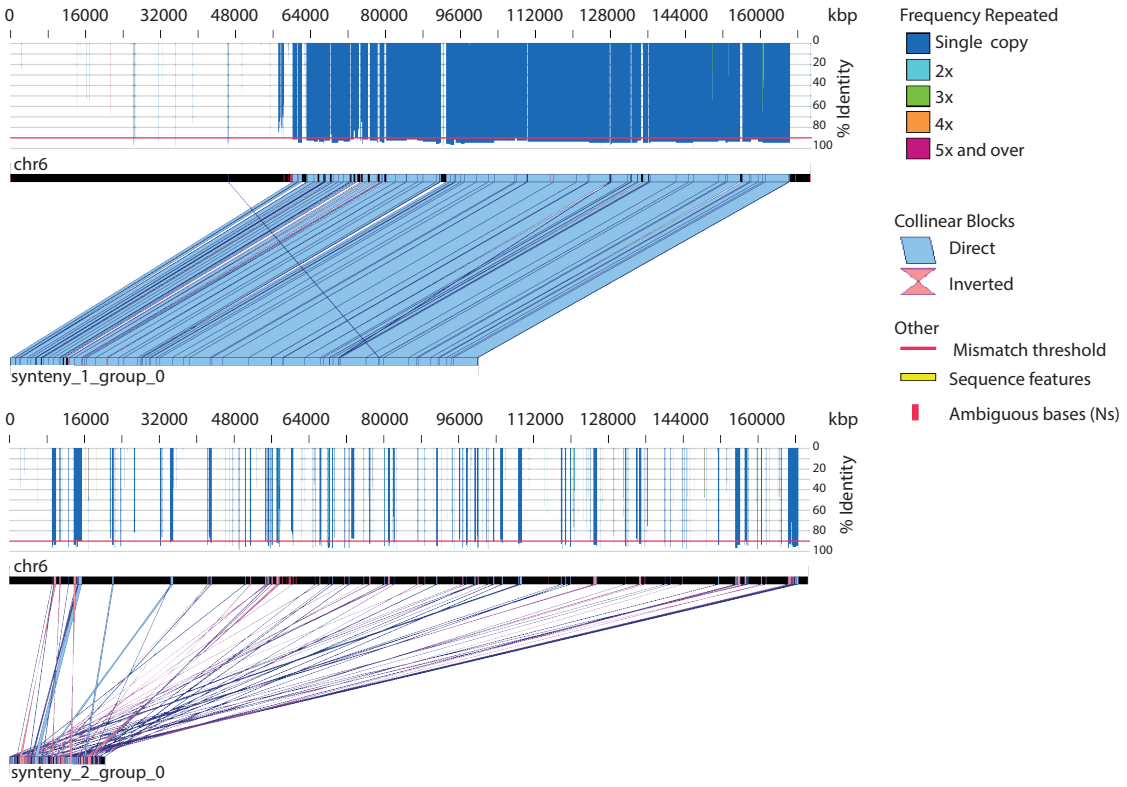


Fig. 3.12 Alignment plots produced using xmatchview for largest contigs from the chromothriptic and wild-type assembly. Black line represents chromosome 6 from GRCh38 and the line below represents the contig. Blue regions are direct repeats and pink regions are inverted repeats. An identity threshold was set at 90%.

3.9 Haplotype-resolved structural variant calling

Having a confident set of SVs is important in both characterising the extent of the rearrangement in the chromothriptic regions as well as in assessing how changes in the linear genome sequence affect higher-order structures. For the purpose of genome assembly, both germline and somatic SVs were used as both were needed to reconstruct the underlying sequence. However germline SVs would not be informative in studying how chromothriptic SV leads to alterations in higher order structures. As such, germline SVs needed to be filtered. The initial SV list used for genome assembly was called from all CCS reads and therefore was not haplotype-resolved. In order to study how the allele-specific expression and chromatin configurations alter in relation to SVs, SVs needed to be haplotype-resolved.

Haplotype-resolved SVs were called using Sniffles on haplotype-resolved PacBio CCS reads aligned to the GRCh38 reference genome. Since matched normal CCS reads were not available, germline variants could not be filtered out based on a haplotype-resolved normal

genome. Instead, germline homozygous variants were filtered out if they were called on both alleles. Further germline variants were filtered out using variants identified in Illumina X Ten short reads. Both BRASS and unmatched GRIDSS calls were used to identify variants in the matched normal blood sample. Furthermore, SVs mapping to decoy sequences or repeat regions as annotated by UCSC repeat masker were removed as mapping in these regions would be difficult and may lead to erroneous SV calls. Short insertions (between 250 bp and 350 bp in length) were also aligned to the reference genome using BLAT (Kent, 2002). If they did not align to chromosome 6, they were filtered out as these were likely to be Alu elements not present in the GRCh38 reference or retrotransposition reactivation events,

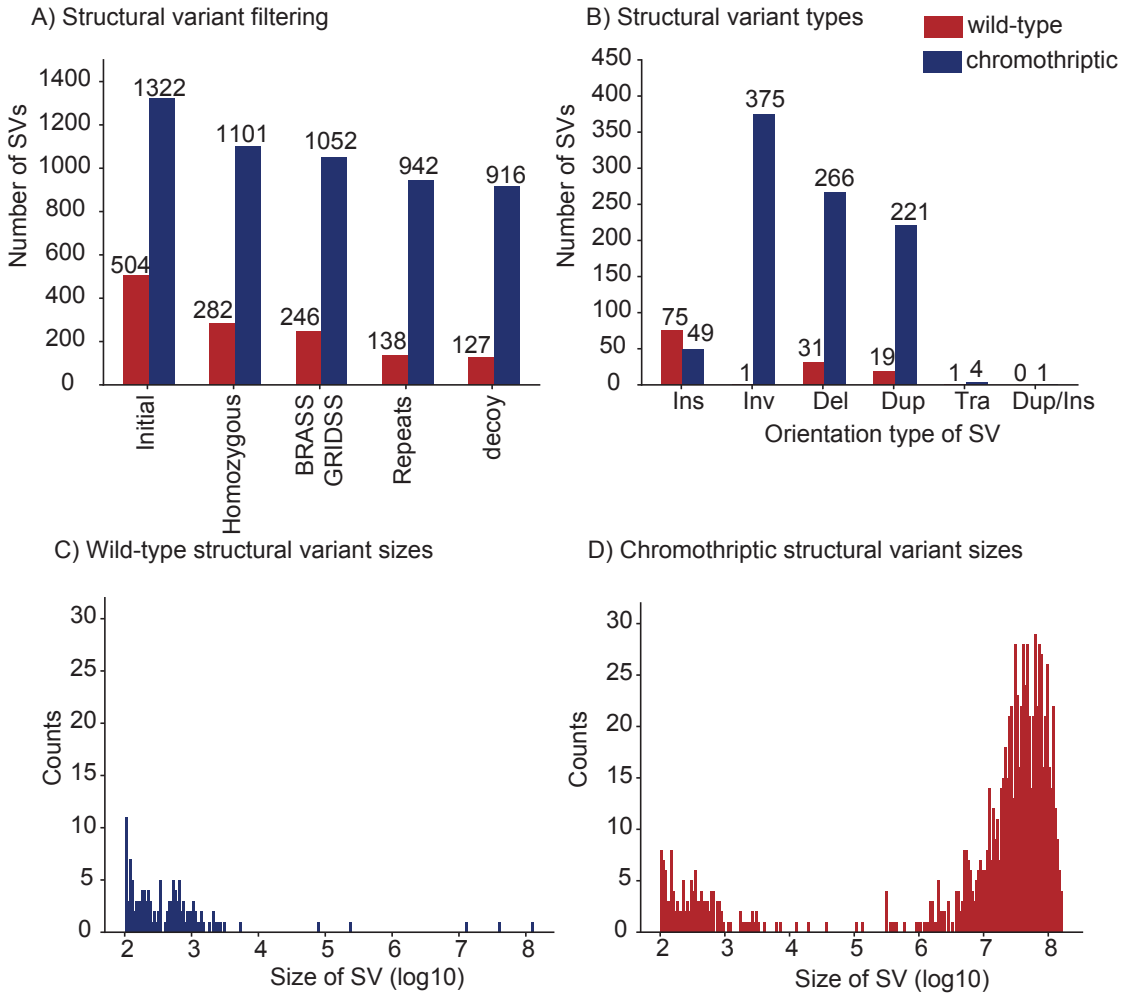


Fig. 3.13 Summary SV statistics for WTSI-OESO_103. A) SV filtering to remove germline SVs and SVs in repeat regions which are likely to be erroneous calls. B) Orientation of SV breakpoints. C) and D) SV size distribution for wild-type and chromothriptic chromosomes, respectively.

which are highly prevalent in this sample. The number of variants filtered at each stage can be seen in Figure 3.13A. It is important to note that, while this filtering method led to the removal of many germline SVs, some true germline SVs will evade filtering particularly since match normal CCS reads were not available. However erroneously removing SVs as germline, will cause the loss of true somatic SVs and therefore a balance is needed. The wild-type assembly contained 127 somatic SVs when compared to the GRCh38 reference genome and the chromothriptic assembly contained 916 somatic SVs. Unless otherwise stated, the SVs which are used in Chapter 4 are haplotype-resolved and filtered to remove germline variants.

On the chromothriptic allele, there is a similar number of deletion and duplication orientation SVs as inversion orientation SVs (Figure 3.13B). There is a different distribution of structural variant sizes on the chromothriptic and wild-type alleles (Figure 3.13C and Figure 3.13D). On the wild-type allele, most of the SVs (81.1%) are small events under 1,000 bp in size. Conversely, on the chromothriptic allele, there is a bimodal distribution of SV sizes. There is a peak of events under 1,000 bp which accounts for 113 SVs (12.3%). This is comparable to the total number of SVs seen on the wild-type allele. However, there is also a peak seen above 1 Mb which accounts for 770 SVs (84.1%). These SVs are generated by the chromothriptic event which brings regions of the genome that are far apart in the reference GRCh38 genome close together. These breakpoints span repeat and regulatory regions. On the chromothriptic chromosome, the distribution of these events can be compared to how frequently 1,000,000 randomly simulated breakpoints span the same repeat and regulatory regions. There are many breakpoints that span repeat regions and the distribution of these repeat-spanning breakpoints are different to those that were randomly simulated (Mann–Whitney U test, p-value = 2.8×10^{-14}) (Table 3.12). Breakpoints also span regulatory regions and when comparing to randomly simulated breakpoints, distal enhancer spanning breakpoints and CpG island spanning breakpoints have different distributions to

Table 3.12 SVs spanning genomic features

	SVs spanning features (%)		
Genomic features	Real breakpoints	Simulated breakpoints	difference (p-value)
Simple repeats	50.3	49.7	2.8×10^{-14}
Proximal enhancers	0.55	1.13	0.17
Distal enhancers	5.08	6.74	4.9×10^{-2}
Promoter	0.38	0.30	0.24
CTCF	0.49	0.48	0.47
CpG islands	1.75	0.56	3.8×10^{-2}

simulated one (Mann–Whitney U test, p-vaules = $4.9x10^{-2}$ and $3.8x10^{-2}$, respectively) (Table 3.12). However, for proximal enhancers, promoters and CTCF binding sites, the distributions of these breakpoints are not different to what would occur by chance (Mann–Whitney U test, p-vaules > 0.05) (Table 3.12).

3.10 Final assembly statistics

3.10.1 Chromothripsis in other samples

So far this chapter has only discussed the chromosome 6 assembly in WTSI-OESO_103. On this chromosome, there is an extremely high density of structural variants throughout the entire chromosome. However, it is imperative that this methodology works when reconstructing less extreme chromothripsis as well. This was investigated using the chromothripsis present in the other organoids. Many of the chromothriptic variants were successfully reconstructed (Figure 3.14).

WTSI-OESO_117 has a single chromothriptic region on chromosome 9 which spans the entire chromosome (Figure 3.14A). The overall copy number of the chromosome is 3, with 2 wild-type copies and a single chromothriptic copy. This asymmetric copy number profile makes haplotype resolution much easier than if a symmetric copy number profile was present. This is because, even in regions with no other discernible structural features, reads can be assigned based on the VAFs of the SNPs in those regions. This results in fewer blocks which are randomly assigned because they have no informative features, and subsequently a more contiguous assembly due to this better assignment. This is evident as the wild-type assembly is macroscopically highly similar to the reference GRCh38 genome (Figure 3.14C), as expected. This wild-type assembly is also highly contiguous with 4 total contigs, one of which forms the majority of the p-arm and another forms the majority of the q-arm. The chromothriptic assembly is also highly contiguous, with 12 total contigs, and an L90 of 5. Many deletions occur as a result of the chromothripsis and these can be seen in the assembled chromosome along with long stretches which are highly similar to the reference GRCh38 chromosome 9 (Figure 3.14B). For example, the deleted region between 80.0 Mb and 101.4 Mb is missing in the chromothriptic assembly but present in the wild-type assembly.

Interestingly, WTSI-OESO_117 and WTSI-OESO_152 are organoids derived from the same patient. WTSI-OESO_117 was derived from a sample obtained pre-chemotherapy and the WTSI-OESO_152 was derived post-chemotherapy. Somewhat unsurprisingly, the chromothriptic rearrangements on chromosome 9 are strikingly similar in these two organoids (Figure 3.14A,D) with only the addition of a few further rearrangements, such as a transloca-

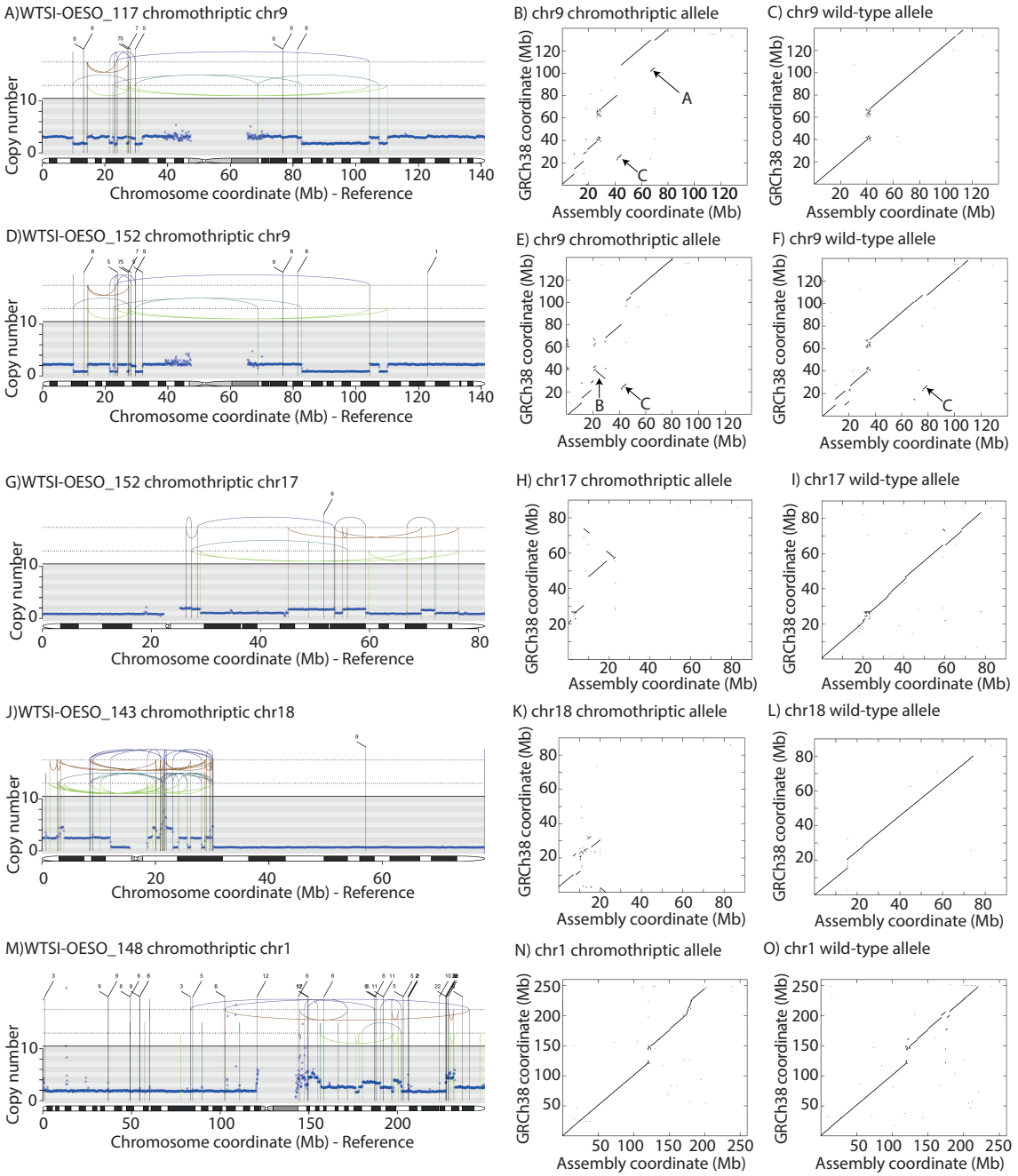


Fig. 3.14 Chromothriptic regions in organoids. Rearrangements plots are derived from Illumina X Ten sequencing as previously described. Dot plots show final assembly alignments to the reference GRCh38 genome. A-C) Chromosome 9 in WTSI-OESO_117. D-F) Chromosome 9 in WTSI-OESO_152. G-I) Chromosome 17 in WTSI-OESO_152. J-L) Chromosome 18 in WTSI-OESO_143. M-O) Chromosome 1 in WTSI-OESO_148. Arrow A highlights a rearrangement in WTSI-OESO_117 but not in WTSI-OESO_152. Arrow B highlights an inversion in WTSI-OESO_152 but not in WTSI-OESO_117. Arrow C in B,E+F highlight a rearrangement in the chromothriptic chromosome in WTSI-OESO_117 but both assemblies in WTSI-OESO_152.

tion to chromosome 1 in WTSI-OESO_152. This provides us with an opportunity to assess the reproducibility of the assembly methods to assemble the same set of structural variants in different organoids.

For WTSI-OESO_152, the overall copy number was two. One of the wild-type alleles has been lost when compared to the pre-chemotherapy sample. This will make the haplotype resolution more difficult in regions where there are no SNPS, no structural features or only copy number neutral structural features as haplotype blocks will need to be randomly assigned. Despite this, the assembled chromothriptic regions are broadly similar (Figures 3.14B,E). However, there are some notable differences. There is a rearrangement which has been reconstructed in the WTSI-OESO_117 assembly which brings a region at 130 Mb to 100 Mb (Figure 3.14B arrow A) and this is not seen in the WTSI-OESO_152 assembly. This rearrangement is not called by SV callers in the sequencing reads, suggesting that this is an assembly error.

There is also an inversion present in the reconstructed chromothriptic chromosome 9 of WTSI-OESO_152 (Figure 3.14E arrow B) which is not seen in the WTSI-OESO_117 assembly. It is possible that this is a true structural variation present between WTSI-OESO_117 and WTSI-OESO_152 and that the chemotherapy selected for a subclone which did not have the rearrangement. The alternative is that the assemblers have mis-assembled some regions of the chromosome.

In order to determine which was the case and to better visualise the similarity between the two assemblies, they can be aligned against each other. In the dotplots in Figure 3.14, the segments are aligned and oriented in relation to the reference genome. If the contig spans both an inverted and non-inverted region, this will be shown. However if the entire inversion is contained within a single contig and does not span any non-inverted sequence, the dot plot will show it as not inverted as it will orient the inversion relative to the reference genome. By aligning these contigs against each other, orientation is taken into account relative to the assemblies rather than the reference GRCh38 genome. This was done for the large segment that is inverted when the two chromothriptic chromosomes are compared (Figure 3.15). When aligning the WTSI-OESO_152 contig containing the inverted sequence to WTSI-OESO_117, it is evident that part of the contig is inverted (pink segments in Figure 3.15) and the other part is in the same orientation (blue segments in Figure 3.15). By doing this, it is clear that the assemblies contain a segment that is inverted relative to each other.

To ascertain if the rearrangement was a real SV or an assembly error, SVs called in the sequencing reads can be informative. There are rearrangements called in both assemblies at these breakpoints, but their orientation is inconsistent with an inversion. This suggests this

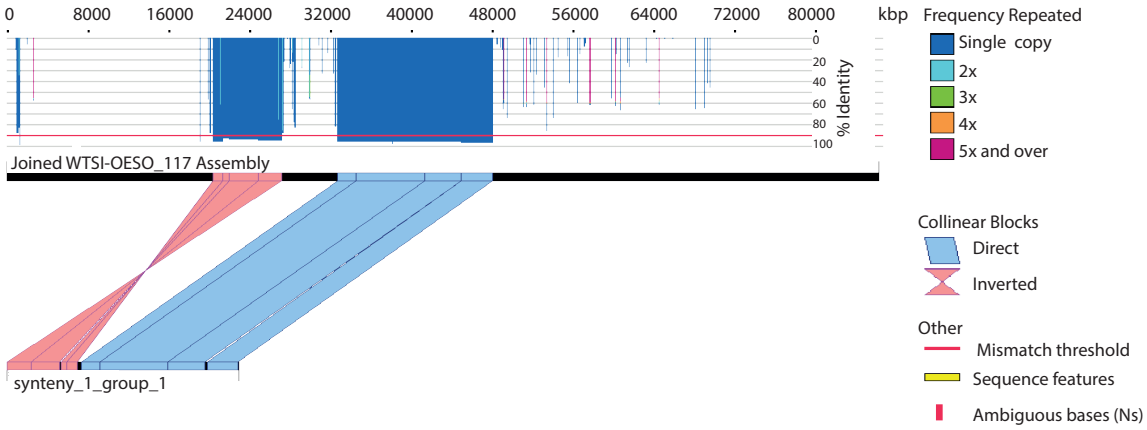


Fig. 3.15 Alignment plots produced using xmatchview for the contig spanning the inverted region on chromosome 9 of WTSI-OESO_152 (bottom line) against the entire chromosome 9 assembly for the chromothriptic chromosome 9 of WTSI-OESO_117 joined by GRCh38 reference genome order (Black line). Blue regions are direct repeats and pink regions are inverted repeats. An identity threshold was set at 90%.

inversion is likely to be an assembly error rather than a rearrangement gained in the time between the two samples.

The wild-type allele assembly of chromosome 9 in WTSI-OESO_117 is very similar to the reference GRCh38 assembly (Figure 3.14C). Interestingly, there is more variation seen in the wild-type assembly of chromosome 9 in WTSI-OESO_152 (Figure 3.14F). Again, these rearrangements may be real SVs or misassemblies attributable to better phasing of haplotypes in WTSI-OESO_117. In particular, there are some rearranged segments present in both haplotypes of chromosome 9 WTSI-OESO_152. For example, there is a segment which has been placed not at the reference location in both the chromothriptic and wild-type assemblies of chromosome 9 (arrow C in Figure 3.14E,F). This suggests this is a germline variant or a region that has not been haplotype-resolved correctly in WTSI-OES_152 and therefore led to an assembly error. In WTSI-OESO_117, this rearrangement only appears in the chromothriptic allele suggesting the latter may be the case.

In WTSI-OESO_152, there is a second chromothriptic region on chromosome 17 (Figure 3.14G) which is not present in WTSI-OESO_117. Chromosome 17 in WTSI-OESO_117 (not shown) has one haplotype that looks macroscopically very similar to the reference GRCh38 genome and the other is similar from 26.6 Mb to 78.7 Mb but outside of these regions sequence has been lost. This loss of sequence may be important for chromothripsis that is seen later in the tumour progression, i.e in WTSI-OESO_152 but not WTSI-OESO_117. There are two copies of the normal chromosome and one copy of the shortened one in WTSI-OESO_117. Interestingly, in WTSI-OESO_152 there has been a loss of one of the

normal chromosomes of chromosome 17, similar to what has happened on chromosome 9. Again haplotype resolution of chromosome 17 in WTSI-OESO_152 will be more difficult than for WTSI-OESO_117.

The chromothripsis on chromosome 17 in WTSI-OESO_152 is relatively easy to reconstruct based on the rearrangements seen in Figure 3.14G. The assembled chromosome for WTSI-OESO_152 recapitulates many of these features. Large deletions can be seen below 20.9 Mb, between 30.7 Mb and 46.9 Mb, between 61.2 Mb and 71.5 Mb and above 74.3 Mb. The deletion between 55.7 Mb and 56.9 Mb seen in Figure 3.14G is not visible at this Mb scale visualisation in Figure 3.14H but this region is absent in the assembly of the chromothriptic allele. Importantly, we are also able to reconstruct the inversions. Both BRASS and Sniffles called inversions at 55.6 Mb to 61.2 Mb and 68.8 Mb to 74.0 Mb and these have been accurately reconstructed in the chromothriptic allele. The final chromothriptic assembly has 7 total contigs and an L90 of just 3 contigs. The wild-type assembly is less contiguous but assembled more of the chromosome rather than just the regions between 26.6 Mb and 78.7 Mb. The wild-type assembly consists of 20 contigs in total, with an L90 of 8.

In WTSI-OESO_143, chromothripsis occurred on chromosome 18. This chromothripsis led to a complete loss of sequence above 32.7 Mb of reference GRCh38 chromosome 18 (Figure 3.14J). The copy number of the chromothriptic chromosome is 2 and the copy number of the wild-type chromosome is 1. Therefore, the whole-chromosome duplication of the chromothriptic allele must have occurred after the chromothriptic event. This asymmetric copy number means that haplotype resolution worked very well on this chromosome. While the overall copy number of the chromothriptic chromosome is 2, there are also regions of further local amplification throughout the chromosome. When these regions are small, the assembler is able to reconstruct them. For example, there is a duplication of sequences found between 3.27 Mb and 3.86 Mb and many small duplications of sequences between 21.9 Mb and 25.3 Mb (Figure 3.14J). These amplifications are also present in the assembly (Figure 3.14K and Figure 3.16 arrows) as different regions of the assembly align to the same region of the reference GRCh38 reference.

The chromothriptic assembly is highly contiguous with 6 total contigs and an L90 of just 2. The wild-type assembly is less contiguous. Like with chromosome 17 in WTSI-OESO_152, the total sequence of the wild-type assembly is greater than the chromothriptic assembly, since the latter lost a large part of chromosome 18. The wild-type assembly is composed of 10 contigs, with an L90 of just 6. It is macroscopically very similar to the reference GRCh38 chromosome 18 (Figure 3.14L).

The chromothripsis in WTSI-OESO_148 affects chromosome 1q (Figure 3.14M). Chromosome 1p in both the chromothriptic and wild-type chromosomes are relatively normal

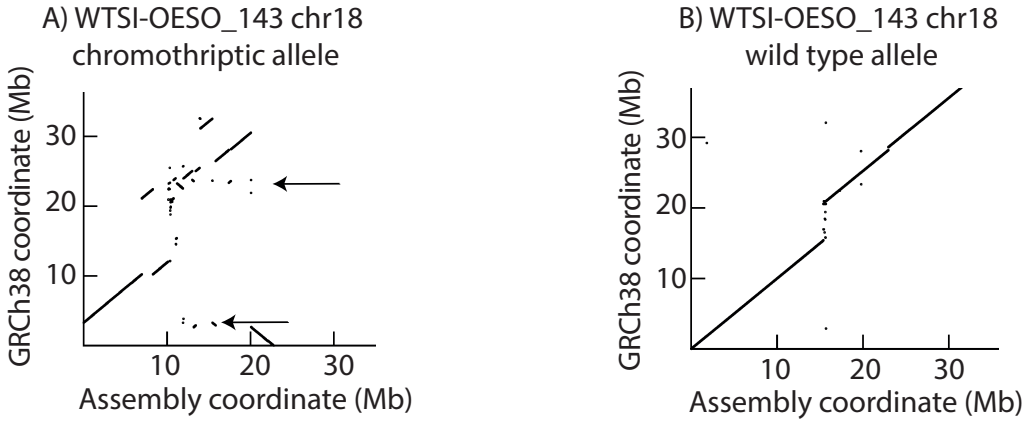


Fig. 3.16 Chromothriptic chromosome in WTSI-OESO_143. Dot plots show alignments to the reference GRCh38 genome chr18:0-35,000,000. A) Chromothriptic allele and B) wild-type allele. Arrows in A highlight regions of reconstructed duplications present on chromothriptic allele. These duplications are not present in the wild-type allele.

and both assemblies have reconstructed this (Figures 3.14N,O). The chromothripic q-arm was difficult to assemble due to many regions of amplification. Unlike on the chromothriptic chromosome 18 in WTSI-OESO_143 where the duplicated regions are relatively small, the duplicated regions in chromosome 1 in WTSI-OESO_148 are large. Most assembly tools are built to assemble relatively normal chromosomes. Therefore, regions of large tandem duplication, without reads which cover the entire length of the tandem duplication, are often collapsed. As a consequence, these large tandem duplications, such as the one between 161.7 Mb and 175.7 Mb (Figure 3.14M), have not been recapitulated to the correct copy number and only appear once in the assembly (Figure 3.14N). These duplicated regions may have important functional and phenotype effects on the cell. Therefore, it is important to account for the increased copy number when making conclusions about higher-order structuring.

Despite not being able to reconstruct the large amplifications, the deletions on this allele have been recapitulated accurately. For example, the region from 176.6 Mb to 178.1 Mb, that is deleted on the chromothriptic chromosome, is not present in the chromothriptic assembly (Figure 3.14N). There also appears to be some rearrangements in the wild-type assembly (Figure 3.14O). The wild-type assembly is more contiguous than the chromothriptic assembly as the assembly is formed of 16 contigs, with L90 of just 4 contigs. Conversely the chromothriptic assembly consists of 66 contigs, with an L90 of 6.

The haplotype resolution and subsequent assembly method has successfully reconstructed six incidences of chromothripsis. It is effective in different densities of structural variation and

with different SV types. The main limitation lies in the assemblers collapsing regions where there are large duplications and therefore these regions only appear once in the assembly.

3.10.2 Non-chromothriptic chromosome assemblies in all organoids

The non-chromothriptic chromosomes in all samples were also reconstructed using the methodology described above. It was successful at the varying levels of rearrangement, from two alleles with very little difference at the structural level to simple rearrangements to regions of complex variation, such as breakage-fusion-bridges. In most cases, hifiasm performed well as the initial CCS read assembler and 3D-DNA performed well as the scaffolder.

In order to summarise the assemblies and measure assembly contiguity, L90 and N90 were used. In the assemblies for the non-chromothriptic chromosomes, the mean L90 was 4 and the median L90 was 2. Of the assembled chromosomes, 76.5% had an L90 of less than 5 contigs and 95.9% had an L90 of less than 10 contigs (Figure 3.17A). This speaks to the high quality of the generated assemblies as they are composed of large contiguous sequences.

Noticeably, chromosome 16 had a larger L90 compared to many of the other chromosomes. This can be attributed to the lack of scaffolding in three of the five samples on chromosome 16. Scaffolding caused fragmentation of chromosome 16 in WTSI-OESO_103, WTSI-OESO_117 and WTSI-OESO_152 meaning the assembly post-scaffolding was less representative of the allele than the assembly pre-scaffolding. For example, in WTSI-OESO_117, both haplotypes are relatively normal so should be very similar to the GRCh38 reference chromosome at the macroscopic level. The initial haplotype 1 and 2 assemblies are in 36 and 45 total contigs respectively, with 81,438,785 and 85,532,930 total base pairs respectively. However, post-scaffolding, the haplotype 1 and 2 assemblies were in 6 and 9 contigs respectively with 81,453,785 and 37,761,532 total base pairs respectively. While there are fewer total contigs, the scaffolding introduced an assembly error on haplotype 1 (Figure 3.18A vs 3.18C) and fragmented haplotype 2 so only 44.1% of the total initial sequence remained (Figure 3.18B vs 3.18D). It is unclear why this occurred. However the CCS assemblies are better representations of the chromosomes and therefore these will be used as the final assemblies for these chromosomes. Similar changes between initial assemblies and scaffolded assemblies can be seen on chromosome 16 in WTSI-OESO_103 and WTSI-OESO_152. Therefore, the CCS assemblies, rather than the scaffolded assemblies, were used for these chromosomes as well.

It is also important to note that, while the assemblies are fragmented by scaffolding, the chromosome 16 CCS assemblies often contained more of the centromere than present in the

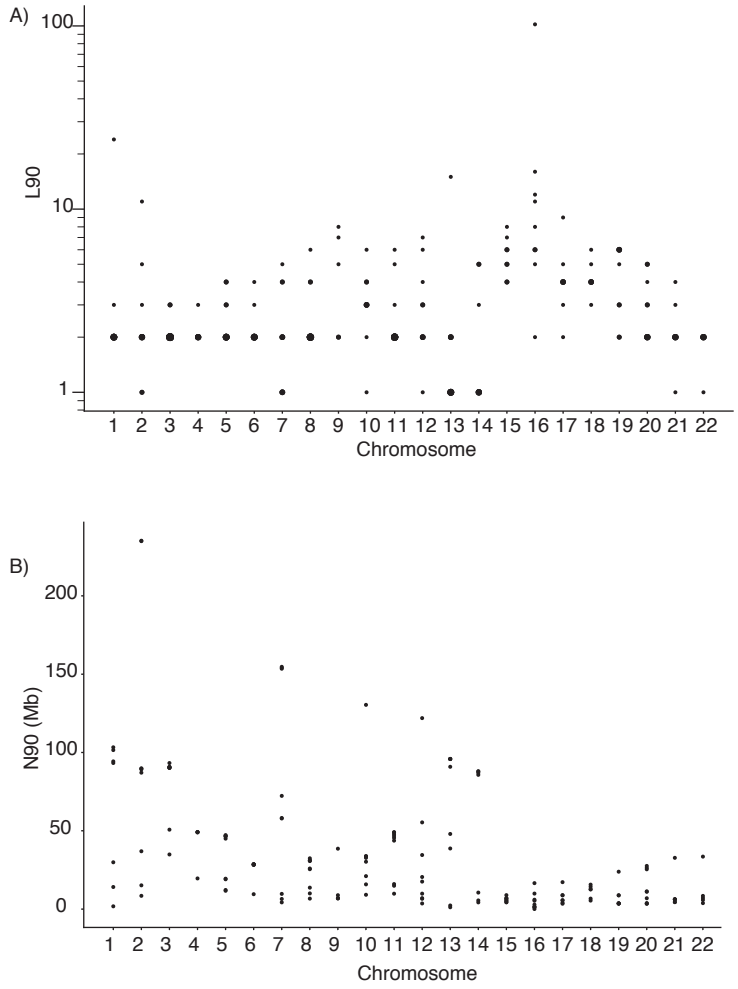


Fig. 3.17 Summary SV statistics for non-chromothriptic chromosomes. A) Assembly L90 split by chromosome. The size of the dot is scaled relative to the number of data points. Due to a highly fragmented chromosome 16 assembly with an L90 of 102, L90 is visualised on a log scale. B) Assembly N90 split by chromosome.

GRCh38 reference (Figure 3.18A,B). This suggests that these assemblies better reconstructed the centromeric sequence than the reference chromosome 16.

The other important metric for assessing assembly success is N90. If the contigs are ordered by size in descending order, N90 is sequence length of the shortest contig at which 90% of the total assembly size is reached. It is an important metric as the assemblies will be used to look at higher order structuring of chromosomes. Therefore, large contigs are required to reconstitute long-range interactions. In the non-chromothriptic assemblies, the mean N90 is 35.5 Mb and the median N90 is 18.3 Mb. Of the assemblies, 36 had an N90 greater than 50 MB and 9 had a N90 greater than 100 Mb.

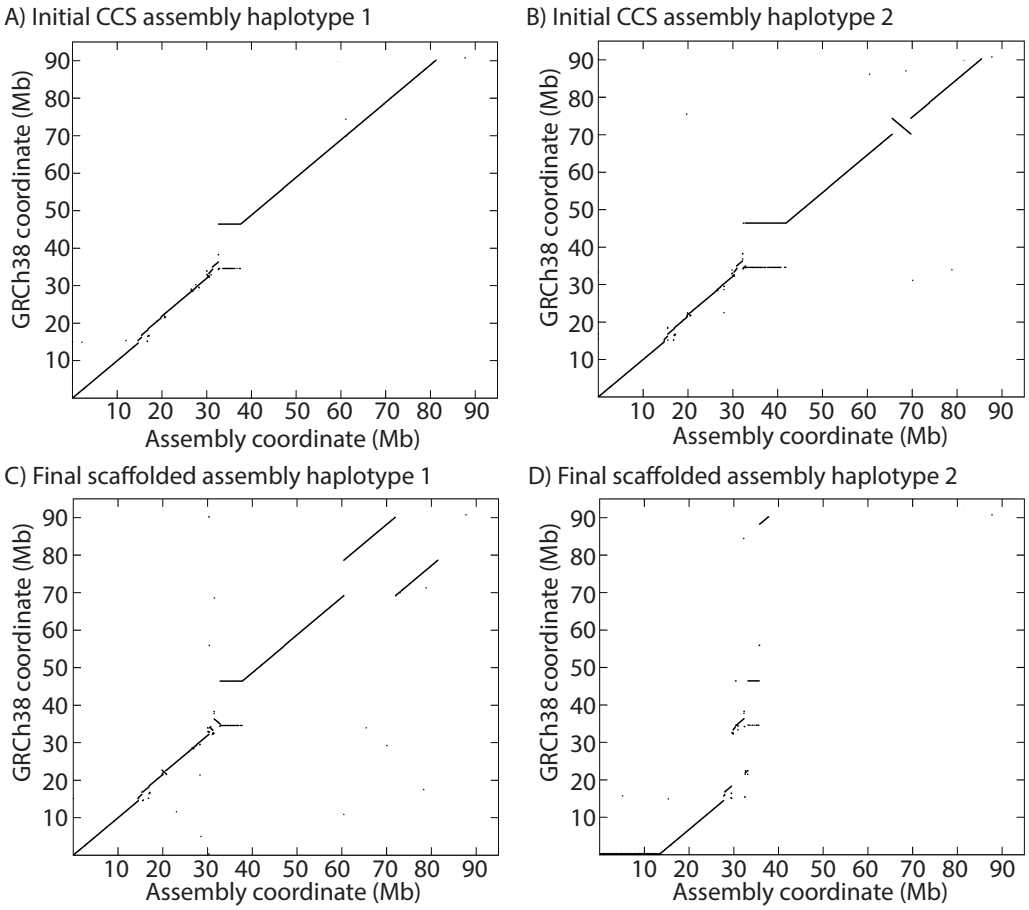


Fig. 3.18 Chromosome 16 in WTSI-OESO_117. A) and B) Inital CCS assembly dotplots of haplotype 1 and 2, respectively. C) and D) Scaffolded assembly dotplots of haplotype 1 and 2, respectively. Scaffolding fragmented the initial CCS assemblies making them less representative of the underlying genome sequence.

3.10.3 Assembling all types of structural variants

Many genome assemblers are built to assemble relatively normal chromosomes. Even haplotype-aware assemblers are built to expect only minor differences between two alleles, and not the extensive variation that is seen in cancer genomes. In cases of chromothripsis, there is extreme variation between the two alleles as well as relative to the reference genome. I have previously discussed how some types of structural variation have been reconstructed in chromothriptic regions. However, it is important to characterise whether it was possible to recapitulate all types of structural variation. This is most apparent in chromosomes with only a few rearrangements.

The simplest type of SVs to reconstruct is a deletion. True clonal deletions will be easy to assemble as there will be no reads in the deleted region and there will be reads spanning

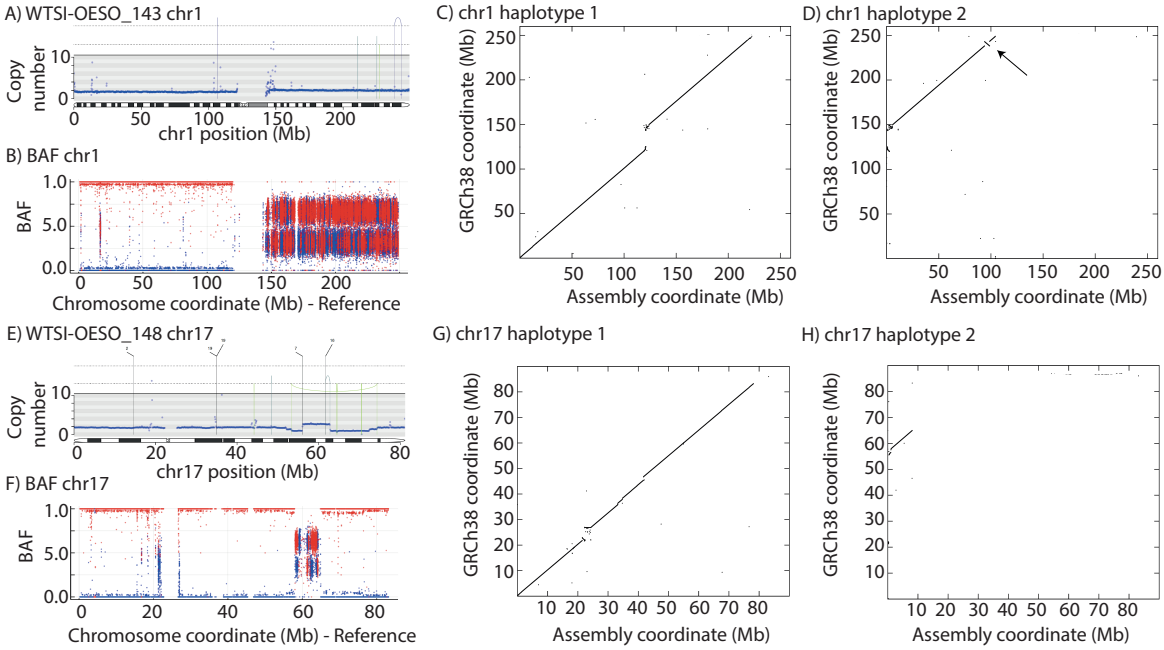


Fig. 3.19 Structural variants can be reconstructed using this assembly method. Panels A and E show rearrangement plots derived from Illumina X Ten sequencing as previously described. Panels B and F show unphased BAF of the chromosome of interest derived from Illumina X Ten sequencing as previously described. Panels C,D,G and H show dot plots of assemblies of reconstructed haplotypes relative to the reference GRCh38 genome. Arrow highlights an inversion.

both ends of the deletion. Deletions can be detected using BAFs and changes in read depths. They can span whole chromosome arms, as observed in chromosome 1 in WTSI-OESO_143 (Figure 3.19A,B). The final assemblies reflect this successfully as one allele is similar to the reference genome (Figure 3.19C) and the other allele begins after the centromere (Figure 3.19D).

Another example of a deletion can be seen on chromosome 17 in WTSI-OESO_148. While one allele is relatively normal when compared to the GRCh38 reference genome, the second allele contains only a segment of chromosome 17 from 58.0Mb to 65.0Mb (Figure 3.19E,F). There are two copies of the smaller allele and one copy of the normal allele. The assemblies generated recapitulate the deletion, with one relatively normal assembly (Figure 3.19G) and one assembly without the deleted sequences (Figure 3.19H). Smaller deletions can also be reconstructed and this has been shown in the chromothriptic regions described previously (Figure 3.14B,E,H,K,N).

Inversions can be difficult to detect in short-read sequencing as they are copy number neutral. However, they have been reconstructed in these cancer assemblies. An example of

this is on chromosome 1 in WTSI-OESO_143. There is a heterozygous inversion of sequence from 238 Mb to 243 Mb (Figure 3.19A) present on haplotype 2, which can be seen in the assembly (Figure 3.19D arrow).

Assembly-based methods may also pick up inversions that were missed in the short-read sequencing and long-read sequencing. An example of this can be seen on chromosome 8 in WTSI-OESO_148. From the short-read sequencing, large inversions had not been detected before 20 Mb (Figure 3.20A) and they have also not be detected by long-read variant callers (not shown). However, in both the haplotype 1 (Figure 3.20C) and haplotype 2 (Figure 3.20D) assemblies, inversions can be seen. It is difficult to ascertain whether these are real inversions or assembly errors. While in this case it is likely to be an assembly error as there is no evidence in SV calls, these inversions being real cannot be completely ruled out as inversions are difficult to detect. Further validation using higher sequencing depth would be needed in order to ascertain this.

Tandem duplications are very difficult to reconstruct. In cases where these tandem duplications are very small and there are reads which span both ends of the duplication, it is possible to reconstruct these. This has been seen in a chromothriptic chromosome

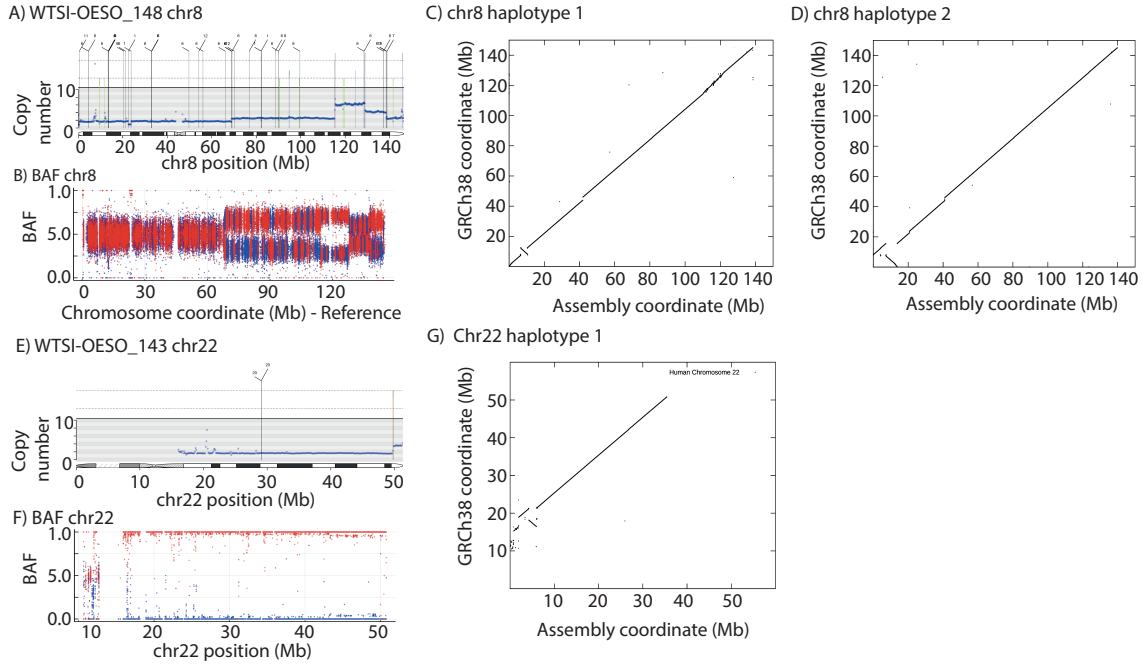


Fig. 3.20 Some structural variants can not be reconstructed using this assembly method. Panels A and E show rearrangement plots derived from Illumina X Ten sequencing as previously described. Panels B and F show unphased BAF of the chromosome of interest derived from Illumina X Ten sequencing as previously described. Panels C,D and G show dot plots of assemblies of reconstructed haplotypes relative to the reference GRCh38 genome

(Figures 3.14K and 3.16). However, when these tandem duplications are larger than the largest read length in that area, assembly algorithms tend to collapse these regions. An example of this can be seen in WTSI-OESO_148 on chromosome 8 (Figure 3.20A,B). In this particular example, both alleles should have amplified regions which is evident from read depth data (not shown). Haplotype 2 is simpler and has a duplication of DNA above 68 Mb until the end of the chromosome. Haplotype 1 remains copy number 1 until 113.8 Mb. There is then a duplication of 113.8 Mb to 127.8 Mb to copy number 4. This is an important amplification because the *MYC* gene lies within this segment and the increased copies of the *MYC* gene may have important functional ramifications for this tumour (Schaub *et al.*, 2018). The copy number then drops to 3 for the sequence between 127.8 Mb and 137.3 Mb. The remainder of this allele is then copy number 1 until the end of the chromosome. Disappointingly, these copy number changes were also missed by the short-read SV callers (not called as rearrangements in Figure 3.20A). However, it is somewhat unsurprising that these amplifications have not been reconstructed because the duplicated segments are far greater than the average read length of the CCS reads (Yahav and Privman, 2019).

Notably this inability to reconstruct tandem duplications is not caused by incorrect haplotype resolution, as duplications in a haploid chromosome are also not reconstructed. An example of this can be seen on chromosome 22 in WTSI-OESO_143 (Figure 3.20E,F). Cells in this organoid model have undergone copy number neutral LOH of chromosome 22, some of which subsequently lose one of the copies. There is a duplication of sequence above 49.4 Mb (Figure 3.20E), however this is not reconstructed in the assembly (Figure 3.20G).

3.10.4 Assembling other complex rearrangements

Chromothripsis is just one example of a complex rearrangement. Cancer genomes often have many different types of complex rearrangements and it is important to accurately recapitulate a variety of them. Here examples of other complex rearrangements present in the organoids are discussed.

In WTSI-OESO_103, chromosome 1 and chromosome 9 have undergone a chromothripsis-like rearrangement event that has inserted regions of chromosome 9 into chromosome 1 (Figure 3.21A,C). Subsequently, for the haplotype where rearrangements between chromosomes 1 and 9 have occurred, these chromosomes have been assembled together to give the most contiguous assembly (Figure 3.21B). The haplotypes without these rearrangements have been assembled separately (Figures 3.21D,E). The start of the p-arm of chromosome 1 has intermingled with the end of the q-arm of chromosome 9, therefore the rearranged assembly is mainly chromosome 1 joined to chromosome 1 and chromosome 9 joined to chromosome 9 with a few regions spanning both chromosomes. The deletions seen on the

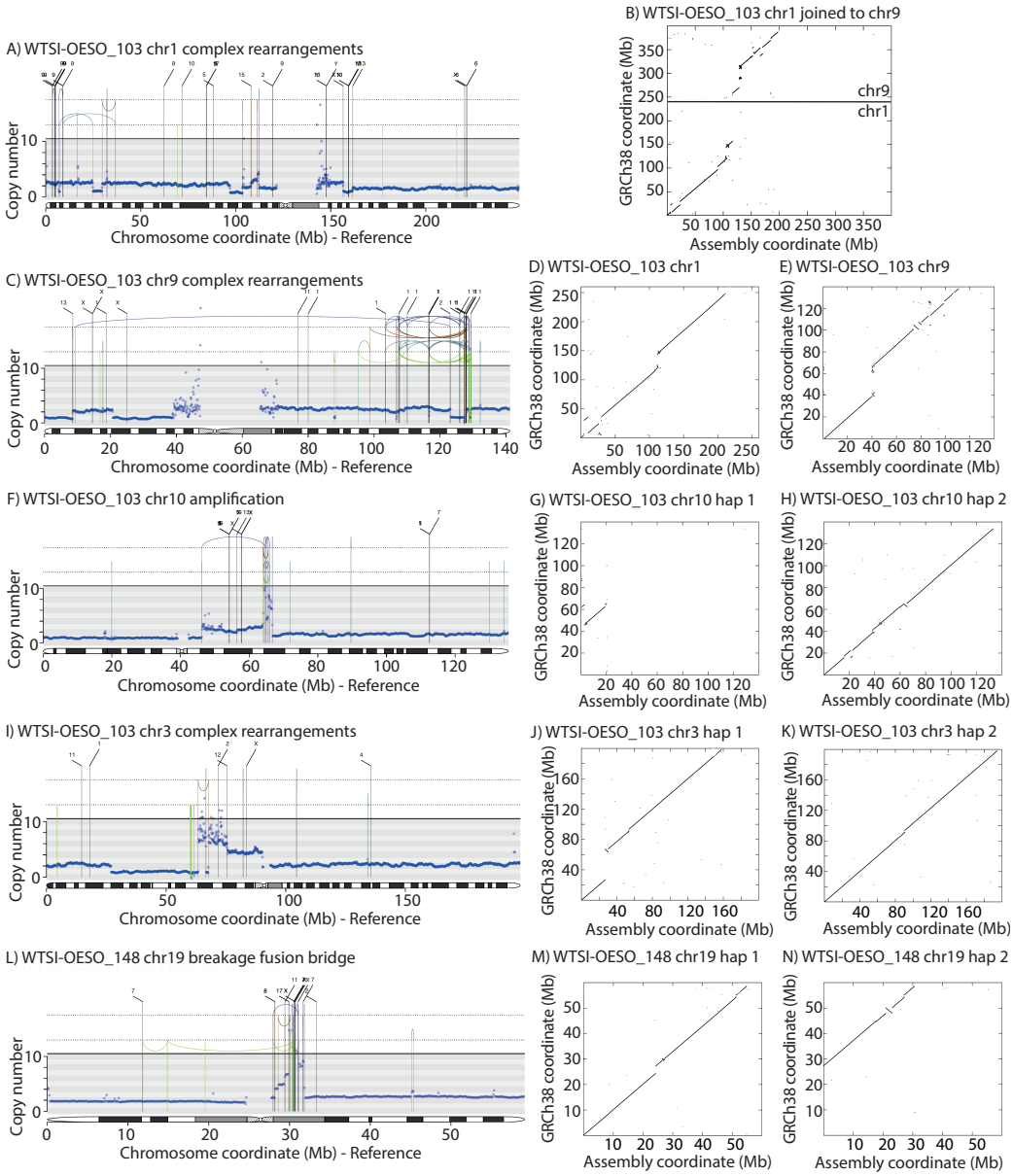


Fig. 3.21 Other types of complex rearrangements have been reconstructed. Plots A,C,F,I,L show rearrangement plots derived from Illumina X Ten sequencing as previously described. B,D,E,G,H,J,K,M,N show dot plot alignments of final reconstructed assemblies relative to the reference GRCh38 chromosome. For panel B, the GRCh38 coordinate is chromosome 1 followed directly by chromosome 9 so 0-249 Mb represents chromosome 1 and 249-387 Mb represents chromosome 9.

rearranged haplotype have been reconstructed in the assembly. For example, above 156.9 Mb on chromosome 1 has been completely lost in the rearranged assembly (Figure 3.21B) but not the wild-type assembly (Figure 3.21D). When examining the rearranged chromosome 9, deletions have also been recapitulated here. For example, sequence between 21.2 Mb to 39.0 Mb is not present in the rearranged chromosome. This can be seen in the dot plot (Figure 3.21B) as the region 279.1 Mb to 288.0 Mb is missing.

These assemblies are more fragmented than the assemblies for equivalent chromosomes in other samples which have not undergone rearrangement or chromothripsis-like events (Figure 3.17). The chromosome 1 L90, if excluding WTSI-OESO_103, is always either 2 or 3. Whereas the L90 for the non-rearranged haplotype is 24. The non-rearranged chromosome 9 L90 is similar to the L90 seen in the other samples. All chromosome 9 L90 values are under 10 contigs and the non-rearranged chromosome 9 L90 in WTSI-OESO_103 is 7. The rearranged assembly containing contigs from both chromosome 1 and 9 has an L90 value of 24.

On chromosome 10 of WTSI-OESO_103, one haplotype has an amplification rearrangement as well as LOH of all regions of chromosome 10 outside of 45.6 Mb to 65.5 Mb. The assembly generated for this haplotype reconstructs many key features. There has been an amplification of the regions of sequence between 62.4 Mb and 65.0 Mb and when these duplications remain in tandem, they are collapsed and not seen in the assembly. However, when these duplications have been moved to other parts of the chromosome, they have been reconstructed. This can be seen in Figure 3.21G. On haplotype 1, there is only a single copy of DNA from reference regions 45.6 Mb to 62.4 Mb. The sequence between 62.4 Mb and 65.0 Mb has been duplicated multiple times. There has also been a rearrangement of some of the duplicated sequence to the start of the chromosome, making this region appear twice in the assembly. Noticeably, the non-rearranged haplotype (Figure 3.21H) also has some rearrangements, which may be misassemblies or real SVs.

Chromosome 3 in WTSI-OESO_103 has very large amplifications over multiple copy number steps as well as regions of large deletions adjacent to the amplifications. Again, since these amplifications are larger than the longest CCS read, these regions have been collapsed in the assembly. On the rearranged haplotype, the LOH region between 27.0 Mb and 63.3 Mb has not been reconstructed but the amplified regions from 63.3 Mb until the centromere has been completely collapsed (Figure 3.21J). The wild-type chromosome looks macroscopically similar to the reference GRCh38 chromosome 3 (Figure 3.21K).

A breakage-fusion-bridge cycle has occurred on chromosome 3 in WTSI-OESO_148. This is apparent from the characteristic copy number profile (Figure 3.21L) representative of large sequences of DNA folding back (Greenman *et al.*, 2016a). There are two copies of

the wild-type chromosome and a single copy of the breakage fusion bridge chromosome. There has been loss of the p-arm in the breakage-fusion-bridge haplotype and several cycles of breakage-fusion-bridges affecting 27.4 Mb to 31.3 Mb. However, these have not been reconstructed in the assembly and have been collapsed into one copy (Figure 3.21N). These amplifications may be functionally important and the inability to reconstruct them means that the impact on higher order structuring cannot be fully understood.

3.10.5 Unexpected assembly difficulties

Despite this methodology working well on most chromosomes, some chromosomes proved difficult to assemble. In WTSI-OESO_148, haplotype1 of chromosome 16 underwent a complex rearrangement producing multiple copy number changes within a 18.7 Mb region of the chromosome (Figure 3.22A). However, hifiasm was unable to produce a sufficient assembly for this haplotype (Figure 3.22B,C) as it was unable to assemble the region above 74.4 Mb. The initial assembly produced by hifiasm had spotty coverage of this region (Figure 3.22B) and many of these contigs were further removed by the "purge duplicates" step of hifiasm (Figure 3.22C). It also was unable to produce a contiguous assembly for the complex region which contained many regions of duplication between 15.9 Mb and 34.6 Mb. The reason for this poor assembly is unclear. However, there was coverage from initial input CCS reads in that region (Figure 3.22E). When the same reads were assembled by wtdbg2, the resultant assembly had a greater contiguity and was more representative of the rearrangements present (Figure 3.22D). This was the only incidence of wtdbg2 being used as the initial assembler instead of hifiasm. In all other cases, hifiasm worked well.

3.11 Discussion

This chapter presents a methodology to generate haplotype-resolved assemblies of cancer genomes, which often contain highly rearranged chromosomes. Methods used to assemble chromosomes in normal cells often generate an initial squashed assembly and then generate two haplotypes based on this initial assembly (Porubsky *et al.*, 2021). In cancer genomes, the initial squashed assembly generated is highly fragmented due to many macroscopic differences between alleles. Subsequently, haplotype resolution of all reads is needed before the initial assembly is produced. Separate assemblies of both alleles then get generated.

One inherent problem of assembling cancer genomes is subclonality of the cancer sample. In this study, a heterogenous population of tumour organoids were grown and subsequently sequenced on multiple sequencing platforms. For the long-read sequencing, DNA was

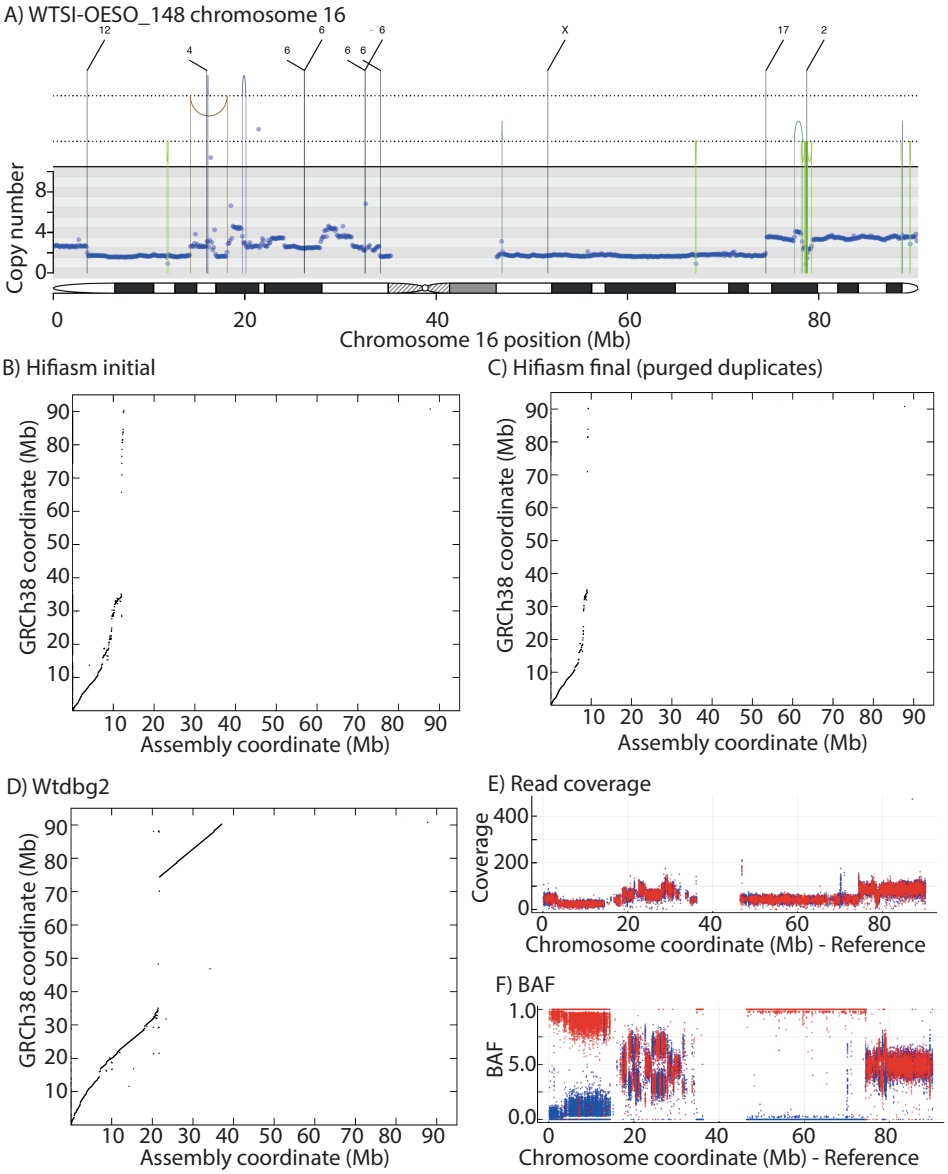


Fig. 3.22 Chromosome 16 in WTSI-OESO_148 assembly using hifiasm and wtdbg2. A) Rearrangement plot of chromosome 16 derived from Illumina X Ten sequencing as previously described. Notably, many copy number variants are not called by BRASS. B) Initial contigs produced for the rearranged haplotype after haplotype resolution and assembly by hifiasm. C) Final contigs produced by hifiasm after the purge duplicates step. D) Contigs produced after assembly using wtdbg2. E) Read coverage of all reads after haplotype resolution. F) Unphased B-allele frequency of chromosome 16 derived from Illumina X Ten sequencing as previously described. LOH from 0 to 15.9Mb is subclonal. LOH from 34.6 Mb to 74.4 Mb is clonal.

extracted from a population of cells at a single time point and this DNA was used for all types of long-read sequencing. Because of this, different assembly solutions from different sequencing technologies cannot be due to differences in the clonal structure of the cancer organoids.

However, while this mitigated different assembly solutions, the problem of subclonality is still present. The underlying chromothriptic events were clonal, as evidenced by the B-allele frequencies. Since these regions are the main regions of interest, they could be reconstructed successfully. However more contiguous assemblies may have been generated if the tumour samples were completely clonal. This is because fragmentation happens at regions where there are multiple paths that the sequence may follow (Du and Liang, 2019), which can be caused by subclonal rearrangements. When this occurs, the assembly graph does not adequately resolve the different paths. In certain regions, subclonal rearrangements may lead to altered read depths, which results in a more fragmented assembly. This is particularly relevant in the scaffolding step. If a region was known to be affected by a subclonal rearrangement, then these reads could be removed from the initial assembly. However, it is difficult to accurately identify all subclonal rearrangements and erroneously removing reads will lead to inaccurate and fragmented assemblies as well.

Single-cell derived organoids can be grown in order to ensure there are no subclonal rearrangements in the initial tumour sample (Roerink *et al.*, 2018). Since each organoid is itself clonal, a single organoid can be picked and then grown into a homogenous population. However, this in itself is a highly inefficient process as organoids have an optimum seeding density at which they grow most successfully (Xie and Wu, 2016). This varies depending on cell type but is much greater than the number of cells from a single organoid. While single-cell derived organoids would have removed some of the technical aspects of the project, this cloning process would have caused the data generation stage of this PhD to be much longer. It is also unclear what the extent of the improvement in assembly contiguity would be with clonal populations. While undeniably there would be some improvement, fragmentation of assemblies also occurs in highly complex or repetitive regions (Du and Liang, 2019) which would still be present in a clonal organoid sample. It is important to note that these organoids tend to be very stable in culture so are unlikely to gain culture-associated structural changes (Huch *et al.*, 2015). However these culture-associated structural changes cannot be completely ruled out, particularly during large expansions of the organoids. Furthermore, tumour samples obtained from patients are often subclonal and it may not always be possible to derive organoids from single cells. Therefore, the ability of this methodology to generate contiguous assemblies even with subclonal samples, means that it is robust and can readily be used on patient-derived samples.

Assembling highly rearranged genomes provides an opportunity to test the benefits and limitations of current assembly methods. Four different genome assemblers were trialled and there was substantial variation in the quality of the assembly produced. Before this it was possible to ascertain the best combination of data to use for haplotype assignment. WhatsHap is a widely used phaser which can use multiple data types in order to phase a genome. Highly accurate CCS reads will provide short length haplotype blocks with highly accurate phasing. The addition of longer read lengths, for example using CLR or linked-read reads, or long-range interactions, for example from Hi-C reads, will connect phase blocks that are further away. Therefore, long phase blocks can be generated and the accuracy of these phase blocks can be validated using trio data (Patterson *et al.*, 2015).

While there was no trio data which could be used for ascertaining the accuracy of phase blocks for the samples in this study, chromothriptic chromosomes have a high density of structural variants which are restricted to only one parental allele. Therefore, there is an inherent truth set. It was evident that, for cancer genomes, increasing the number of data types used did not necessarily improve phasing and in fact led to switch errors. It is unclear why this may be the case. However, only using highly accurate CCS reads and then assigning confident phase blocks based on other features, such as VAFs and SV density, seemed to mitigate the phase switching seen and provides an alternative methodology for phasing.

The assemblies generated by this method adequately reflect many features of cancer genomes. However, these assemblies sometimes poorly represent duplicated regions. When these regions are small, they may be accurately incorporated into the assembly. However, when these regions are large, the regions become collapsed and only occur once in the assembly. This is problematic because copy number variants spanning genes can be associated with differential gene expression (Shao *et al.*, 2019) and therefore may be important drivers of cancer development. One way around this would be to generate an initial assembly where tandem duplications are collapsed and then resolve the collapsed regions. Tools have been developed to do this, for example SDA (Vollger *et al.*, 2019) which uses local read depth in order to identify collapsed duplications and then performs a local assembly using Canu (Koren *et al.*, 2017). This may be able to rescue collapsed regions.

As read lengths for long-read sequencing platforms are increasing, more contiguous assemblies could be generated. This methodology is heavily reliant on CCS sequencing and its associated low error rate. When using just CLR reads, the base error rate is too high leading to an increase in phasing errors. This is evident when looking at haplotype switching from just CLR reads. While longer reads would lead to more contiguous assemblies, these assemblies would contain many errors as phasing would be less accurate. Currently the high

error rate of both Nanopore sequencing and CLR sequencing means that it is imperative that CCS reads are used for complex cancer genome assemblies.

This chapter presents the first attempt to build haplotype-resolved cancer-specific genome assemblies for genomes with extensive, complex chromosomal rearrangement. Key structural features, specific to each cancer, have been reconstructed as well as regions of complex rearrangement, such as chromothripsis. While the assemblies presented in this chapter are imperfect, they are highly representative of the underlying genome sequence and will allow attribution of phenotypic data to specific alleles. These assemblies will be used as a basis to investigate allele-specific differences in gene expression, chromatin accessibility, histone modifications and topologically associating domains in the subsequent chapter.

Chapter 4

Structural variants and the epigenome

4.1 Chapter highlights

This chapter describes the differences in the epigenome and transcriptome between two haplotypes within a sample. The main messages from this chapter are:

1. There is differential transcription, chromatin accessibility and protein binding between chromothriptic and wild-type chromosomes
2. Differences between chromothriptic and wild-type chromosomes are greater than differences between two wild-type chromosomes in some cases of chromothripsis
3. Structural variants cause alterations in TAD structures, differentially active and repressed regions and differential gene expression
4. The mechanisms by which structural variants alter gene expression and epigenetic landscape are complex and often traverse many epigenetic layers.

4.2 Introduction

The primary genome sequence plays a major role in the regulation of the epigenetic landscape in somatic cells (Wilson *et al.*, 2008). This regulation likely operates over a range of genomic scales, from the nuclear positioning of chromosomes into chromosome domains; through megabase-scale chromosome looping determined by topological-associating domains (TADs); to local regulation through enhancer-promoter juxtaposition. The DNA sequence informs this regulation, probably at all scales, but studies relating primary sequence to

chromosome organisation are hampered by the strong within-species similarity of genome sequence and the high between-species variability of chromosome configuration.

Chromothripsis provides an interesting model whereby the relationship between underlying genome sequence and genome regulation can be interrogated. A direct comparison can be made between haplotypes in the same cell; one that has undergone the catastrophic rearrangement event and one that is representative of the chromosomal organisation before this occurred. The genome organisation of both haplotypes can be investigated across many epigenetic levels, from nucleosome modifications and accessibility to topologically associating domains, and the overall effect on gene expression can be queried.

This chapter aims to use haplotype-resolved chromothriptic and wild-type assemblies to gain further insight into the biological consequences of variation in primary genome sequence. Initially, the chromothriptic chromosome 6 in the oesophageal adenocarcinoma organoid WTSI-OESO_103 will be used to answer these questions, after which a broader overview of findings using other samples will be presented. Iso-seq will be used to query differential expression of full-length transcripts between haplotypes and this will be discussed in section 4.4. ChIP-seq and ATAC-seq will be used to investigate changes in histone modifications and accessibility and this will be the focus of section 4.5. Hi-C will be used to reconstruct higher order structures and chromatin contacts and this will be explored in section 4.6. Finally, these data types will be integrated in section 4.7 to generate an overview of some example regions where specific differences between the two haplotypes can be seen.

4.3 Haplotype resolution of epigenetic and expression data

To query haplotype-specific differences, haplotype-resolved data from numerous sequencing modalities needed to be generated (Table 4.1): Iso-seq reads on the PacBio Sequel platform, Hi-C reads on the Illumina HiSeq X Ten platform, ChIP-seq reads on the Illumina Novaseq 6000 platform and ATAC-seq reads on the Illumina HiSeq V4 platform. Duplicate sequencing was performed for all techniques, other than Hi-C, for all five organoids. Haplotype resolution was done by aligning all reads to the chromothriptic assembly and all reads to the wild-type assembly and determining which had the higher mapping score, as described in Chapter 2. This ultimately separated reads into either wild-type or chromothriptic reads.

Some stretches of DNA sequence were exactly identical in the chromothriptic and wild-type haplotypes: these sequences contained no heterozygous SNPs and no structural variation to distinguish the two parental haplotypes. Because of this, assigning reads to the correct

Table 4.1 Epigenetic sequencing coverage

	Average sequencing coverage (x)				
	WTSI-OESO_103	WTSI-OESO_117	WTSI-OESO_143	WTSI-OESO_148	WTSI-OESO_152
Iso-seq	7.9	12.2	8.2	7.2	6.5
ChIP-seq	8.2	8.3	6.8	9.0	7.3
ATAC-seq	6.4	7.0	6.7	5.5	6.3
Hi-C	120.8	117.8	114.9	115.4	118.8

haplotype in these regions using the alignment method outlined above was impossible. For completeness, particularly for the visualisation of TAD structures, reads mapping to these regions were randomly assigned between the parental haplotypes. However, in subsequent analyses of differential regions between haplotypes, only reads that are explicitly assigned, and not those randomly assigned, were used. This strict filtering ensured that there was a high degree of specificity at the cost of slightly lowered sensitivity.

The number of reads randomly assigned varied depending on the underlying data type. This is evident in chromosome 6 of WTSI-OESO_103 (Table 4.2): 68.9% of ATAC-seq reads, 63.8% of ChIP-seq reads, 56.7% of Hi-C reads and 30.1% of Iso-seq reads were randomly assigned. The rest of the reads were explicitly assigned. These varying proportions were due to differences in the underlying sequencing methodology. ChIP-seq reads were 100 bp in length and ATAC-seq reads were 75 bp. These short read lengths meant that there was still a large proportion of the chromosome that did not have an SV or SNP in the region the read covers. ChIP-seq and ATAC-seq reads are aligned as paired-end reads with a fixed distance between read pairs. Since they are paired reads, only one read needs to have a SNP or SV present for both reads in a read pair to be assigned. However, there is still a large proportion of reads that need to be randomly assigned.

Comparatively, more Hi-C reads were assigned since they were 150 bp in length. Hi-C reads are mapped as single-ended reads to allow for variable mapping distances between read pairs caused by long-range interactions. These variable mapping distances mean it is difficult to use mapping distance to distinguish between normal interactions and interactions

Table 4.2 Read assignment groups

	Percentage of reads in each group		
	chromothriptic	wild_type	randomly assigned
ATAC-seq	11.4	19.7	68.9
ChIP-seq	15.0	21.2	63.8
Hi-C	16.2	27.1	56.7
Iso-seq	34.2	35.7	30.1

affected by haplotype-specific SVs. When the difference in mapping distance between the chromothriptic and wild-type chromosome is small, it is impossible to distinguish real 3D interactions from an increased distance caused by a haplotype-specific SV. However, since regions which are close in linear sequence are more likely to interact than regions that are far apart in linear sequence, some read pairs can be distinguished using a weighted probability of likelihood that reads are interacting at specific distances. Together, this distance feature and the longer read lengths meant that a larger proportion of Hi-C reads were assigned when compared to ChIP-seq and ATAC-seq reads.

In contrast, Iso-seq reads are full-length transcripts and therefore have much longer read lengths. This increased read length means that reads are more likely to cover an informative SNP or SV and be assigned based on it. The long lengths caused Iso-seq reads to have the smallest proportion of randomly assigned reads.

4.4 Differential expression and essentiality

Structural variation can have many effects on gene expression. It can directly affect expression, for example by a rearrangement occurring within a gene body. This can lead to the silencing of the gene, a different, often truncated, isoform of the gene being transcribed or a novel fusion gene. Examples of these effects are seen in many cancer genomes (Alaei-Mahabadi *et al.*, 2016) and have been discussed in detail in the Introduction. This section aims to investigate and quantify the extent of this dysregulation using the chromothriptic and wild-type haplotypes.

4.4.1 Comparison of transcripts on haplotypes

Iso-seq transcripts were assigned to the appropriate haplotype based on the criteria described in Chapter 2, using reads which passed initial mapping quality filters. For chromosome 6 in WTSI-OESO_103, 34.2% reads were explicitly assigned to the chromothriptic haplotype and 35.7% were explicitly assigned to the wild-type haplotype. The remaining 30.1% had no informative features and therefore had to be randomly assigned. Reads were then further filtered to remove regions of LOH, representing regions of the reference genome present on one haplotype but absent from the other. In these regions it would be unclear as to whether the differences seen would be due to true biological differences, say a true deletion, or due to a region which was not assembled in one of the assemblies, for example due to complexity in that region. This approach ensured any differences seen in the transcriptome of these haplotypes were specific to differences in the regions the reads originated from. The

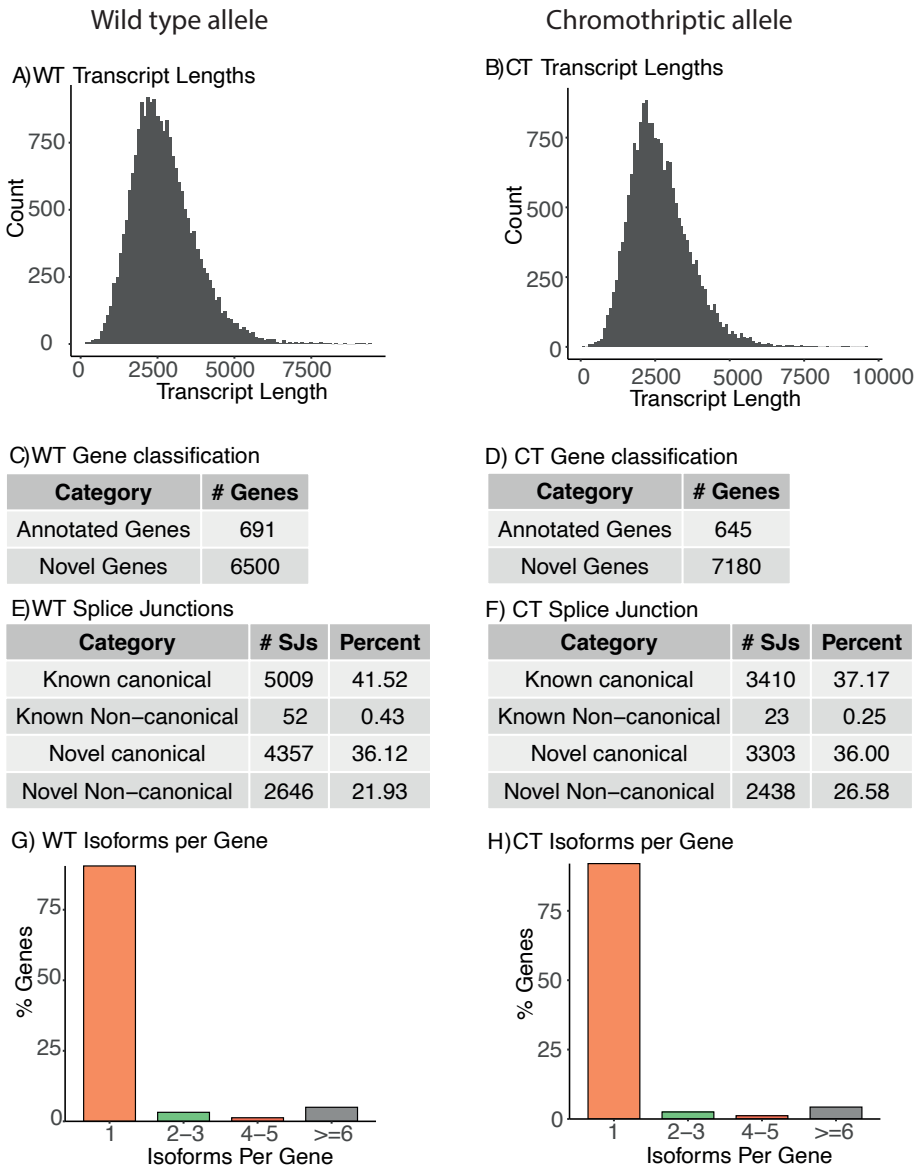
total non-LOH sequence in the genome assemblies of chromosome 6 in WTSI-OESO_103 represented 62.8% of the reference GRCh38 chromosome 6 sequence. These haplotype-resolved explicitly assigned reads which did not map to regions of LOH were then remapped to the GRCh38 reference genome. This was necessary to compare the transcriptomes to each other.

4.4.2 Classification of reads

The distribution of read lengths was similar on the wild-type and chromothriptic haplotypes with a peak around 2,500 bp in length (Figure 4.1A and 4.1B respectively). Transcripts from wild-type and chromothriptic haplotypes were annotated using SQANTI. For chromosome 6 in WTSI-OESO_103, the chromothriptic haplotype expressed more genes than the wild-type haplotype: 7,191 total genes expressed on the wild-type haplotype and 7,825 expressed on the chromothriptic haplotype. Despite this, the wild-type haplotype had 2,328 more unique isoforms expressed when compared to the chromothriptic: 18,455 unique isoforms were expressed on the chromothriptic haplotype to 20,783 unique isoforms expressed from the wild-type haplotype. Many genes were annotated as novel genes in both the wild-type and chromothriptic chromosomes (Figure 4.1C and 4.1D respectively). There were also fewer splice sites in the chromothriptic genes (9,174) when compared to the wild-type genes (12,064). However the relative proportion of whether these splice sites were known or novel and canonical or non-canonical was similar on both haplotypes (Figure 4.1E and 4.1F). The majority of genes on both haplotypes only had one splice isoform expressed (Figure 4.1G and 4.1H)

4.4.3 Differential expression between the two haplotypes

Differential transcripts were identified using DESeq2 (Love *et al.*, 2014). In WTSI-OESO_103, chromosome 6 had evidence of chromothripsis and 910 genes were identified as expressed on at least one haplotype. Of those, 10.7% (97 genes) were differentially expressed: 50 genes were expressed at significantly lower levels on the chromothriptic haplotype than wild-type haplotype; versus 47 genes at higher levels on the chromothriptic than wild-type (Wald test, q-value < 0.05). Notably, 8 differential genes on the wild-type chromosome and 17 differential genes on the chromothriptic chromosomes also contain point mutations. While it is important to note these mutations as potential causes of differential expression, it is difficult to determine whether these mutations or other factors are driving the differences in gene expression. A gene set enrichment analysis of these genes can be informative to identify whether genes which show differential expression are enriched in specific biological path-



C)WT Gene classification

Category	# Genes
Annotated Genes	691
Novel Genes	6500

D) CT Gene classification

Category	# Genes
Annotated Genes	645
Novel Genes	7180

E)WT Splice Junctions

Category	# SJs	Percent
Known canonical	5009	41.52
Known Non−canonical	52	0.43
Novel canonical	4357	36.12
Novel Non−canonical	2646	21.93

F) CT Splice Junction

Category	# SJs	Percent
Known canonical	3410	37.17
Known Non−canonical	23	0.25
Novel canonical	3303	36.00
Novel Non−canonical	2438	26.58

Fig. 4.1 Summary of underlying iso-seq data. Transcripts length distribution (A and B), number of annotated genes (C and D) and percentage of splice junctions in each splice category (E and F) and isoforms per gene (G and H) are shown for wild-type (left) and chromothriptic (right) chromosomes.

ways. Since not all regions can be confidently assigned, in it not suitable to look at all genes on chromosome 6. Instead this analysis was restricted to genes in regions on chromosome 6 that could be assigned. Within these regions, there was no significant enrichment of the differential genes in any biological pathway (hypergeometric test, q-value < 0.05).

In WTSI-OESO_117, WTSI-OESO_143 and WTSI-OESO_152, chromosome 6 was largely unaffected by structural rearrangements and therefore these chromosomes can be used

to determine the baseline differential expression between two haplotypes of chromosome 6. Of genes that were identified as expressed on at least one haplotype, 1.9%, 1.8% and 1.0% of chromosome 6 genes were differentially expressed between the two parental copies of the chromosome for WTSI-OESO_117, WTSI-OESO_143 and WTSI-OESO_152 respectively (Table 4.3). This suggests the chromothriptic event caused more differential transcripts than expected to be present between two wild-type copies of chromosome 6 (Fisher's Exact test, p-value < 10^{-12} for all comparisons of chromosome 6 in WTSI-OESO_103 to chromosome 6 in other samples).

Chromosome 6 in WTSI-OESO_103 had a very high density of structural variants, much higher than that seen in the chromothriptic chromosomes in other samples. The percentage of genes that were differentially expressed on chromothriptic chromosomes in other samples was lower than that seen on chromosome 6 in WTSI-OESO_103 (Table 4.3 percentages in bold). However, when looking at a subset of other normal chromosomes in the same sample, in WTSI-OESO_143 and WTSI-OESO_148 the chromothriptic chromosome had a higher percentage of differential transcripts than these normal chromosomes (Table 4.3 bold versus not bold). While it is important to note that the percentage of differential transcripts varied across samples, this higher percentage of differential transcripts seen on the chromothriptic chromosomes in these samples suggests that chromothripsis may be responsible for perturbing normal gene expression. Interestingly, WTSI-OESO_117 and WTSI-OESO_152 do not appear to follow this trend. WTSI-OESO_117 chromosome 17, which does not have evidence of chromothripsis, has the highest level of differential expression. WTSI-OESO_152, which is derived from the same donor as WTSI-OESO_117 but after chemotherapy rather than before, had a second chromothriptic event that fragments chromosome 17. The high level of differential gene expression on chromosome 17 in WTSI-OESO_117 before this chromothriptic event occurs is intriguing and may indicate an abnormal chromosome 17 state, relative to the haplotype that does not undergo chromothripsis, prior to the genome shattering

Table 4.3 Differential transcripts in other samples

	Percentage of total genes which are differentially expressed (%)				
chr	WTSI-OESO_103	WTSI-OESO_117	WTSI-OESO_143	WTSI-OESO_148	WTSI-OESO_152
chr1	R	0.96	1.7	**8.1**	3.5
chr6	**10.7**	1.9	1.8	LOH	1.0
chr9	R	**1.5**	2.8	3.1	**0.72**
chr17	0.92	4.3	3.1	6.0	**2.3**
chr18	R	1.1	**5.1**	LOH	0.31

R = Rearranged. Chromosomes 1, 9 and 18 in WTSI-OESO_103 are highly rearranged so do not represent normal chromosomes and have been excluded from this analysis.

event. It is important to note that in WTSI-OESO_117, one allele of chromosome 17 has been incorporated into chromosome 6. It is unclear whether this difference between alleles is resultant of the fusion into chromosome 6 or whether it actually foreshadows chromothripsis.

When looking across samples, chromosome 1 and chromosome 18 had a higher percentage of differential transcripts when chromothripsis was present (Table 4.3 numbers in bold versus not bold). This was not true for chromosomes 9 and 17; the chromothriptic chromosome had a lower percentage of differential transcripts when compared to some of the normal chromosomes. It is important to note that different samples have different baseline levels of differential gene expression between two haplotypes and that this makes comparing differences directly across samples difficult. For example, chromosomes 9 and 17 in WTSI-OESO_148 had more differential transcripts than the same chromosomes in other samples, even against samples which contained chromothripsis, but this was less than the differences on the chromothriptic and wild-type chromosome 1 in WTSI-OESO_148, the chromothriptic chromosome. These differences may also be chromosome specific or specific to the chromothriptic breakpoints in each samples. It highlights the difficulty in comparing across samples and a much larger sample size is needed to make a statistically robust conclusion about whether the differential gene expression in chromothriptic regions is greater or less than the difference between two normal chromosomes.

4.4.4 Relationship between SVs and differential transcript expression

The top 40 most differentially expressed genes on chromosome 6 in WTSI-OESO_103 can be seen in Figure 4.2A and include *TMEM30A*, *ROS1*, *ADGRG6* and *AKAP12*. There is high concordance between biological repeats. For some genes, such as *ADGRG6* the differential expression occurs because the gene has been fragmented by the chromothriptic event and is no longer functional even though the gene sequence is still present in the chromosome. This leads to the gene only being expressed on the wild-type chromosome.

To determine whether SVs have a direct effect on gene expression, the distance of differentially expressed transcripts to the nearest SV on either the chromothriptic or wild-type haplotype can be measured. From the Iso-seq data, a difference in expression between the two parental chromosomes is observed, but it is impossible to know whether it is the expression on the wild-type or chromothriptic haplotype that has been affected. Thus, there are two categories of differential expression that can be identified: “higher on chromothriptic / lower on wild-type” and “lower on chromothriptic / higher on wild-type”. For any gene in these categories, the distance from that gene to the nearest SV was estimated on the chromothriptic haplotype, and the distance to the nearest SV on the wild-type haplotype was estimated. The distances for differentially expressed genes against equivalent distances

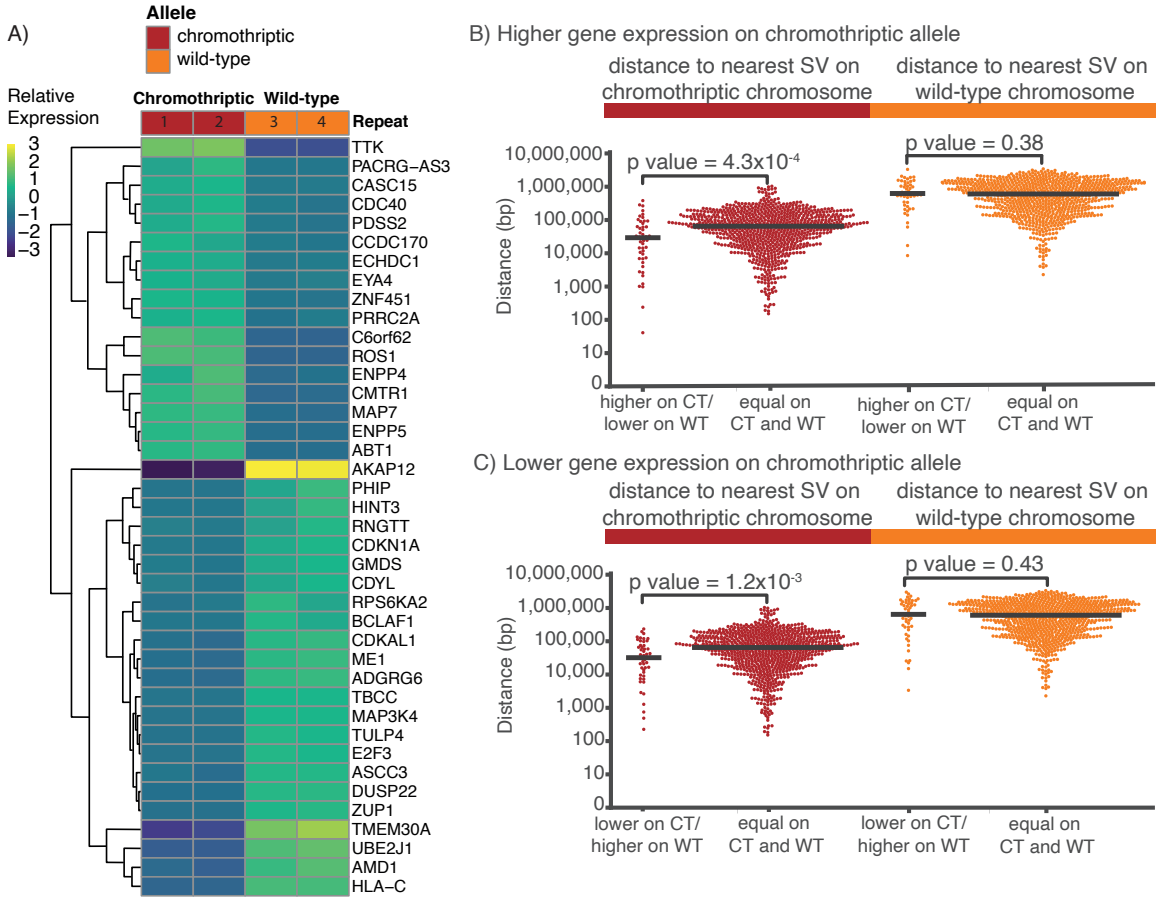


Fig. 4.2 Differential transcript analysis. A) Heatmap of 40 most differentially expressed genes on the wild-type and chromothriptic chromosome 6 in WTSI-OESO_103. B) Distance of differential genes with higher expression on the chromothriptic haplotype to nearest SV relative to the genes with equal expression on both haplotypes. Chromothripsis causes gain of gene expression. C) Distance of differential genes with lower expression on the chromothriptic haplotype to nearest SV relative to the genes with equal expression on both haplotypes. Chromothripsis causes loss of gene expression. Median lines shown in B and C.

for the control set of genes that did not show evidence of differential expression were then compared.

If SVs generated by chromothripsis are causing increased expression, genes which are more highly expressed on the chromothriptic chromosome relative to the wild-type chromosome would be closer to SVs than genes with equal or non-differential expression. Similarly, if SVs generated by chromothripsis are causing loss of expression, genes which have lower expression on the chromothriptic haplotype would be closer to SVs than genes with equal or non-differential expression. This was, in fact, the case for both upregulated

(Wilcoxon rank-sum test, p-value = 4.3×10^{-4}) and downregulated (Wilcoxon rank-sum test, p-value = 1.2×10^{-3}) genes on chromosome 6 in WTSI-OESO_103 (Figure 4.2 B,C).

On the chromothriptic haplotype, the median distance of genes that had higher expression on the chromothriptic haplotype to SVs was similar to the median distance of genes that have lower expression on the chromothriptic haplotype, 29,363 bp and 32,074 bp respectively. They are both much closer to SVs than the median distance of non-differentially expressed genes to SVs, 64,917 bp. Interestingly, inversions tend to be close to genes that have higher expression on the chromothriptic chromosome whereas insertions tend to be close to genes which have lower expression on the chromothriptic chromosomes. Duplications and deletions occur at a similar frequency near genes with higher and lower gene expression on the chromothriptic chromosome (Table 4.4). Furthermore, there is a difference between the SV size distribution of the closest SV to genes which have higher expression and genes that

Table 4.4 SVs nearest to differential genes

SV type	Higher expression on CT	Lower expression on CT
Deletion	13	8
Duplication	8	10
Insertion	2	30
Inversion	24	2

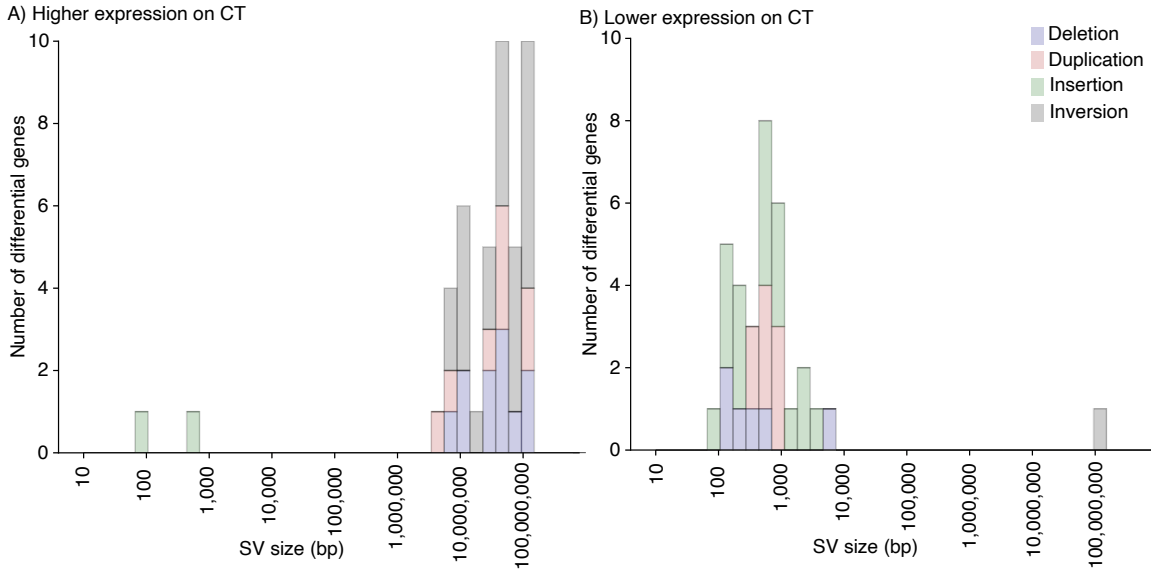


Fig. 4.3 SV size of the nearest SV to genes with differential expression. A) Genes that have higher expression on the chromothriptic chromosome. B) Genes that have lower expression on the chromothriptic chromosome.

have lower expression on the chromothriptic allele (Mann Whittney-U, p-value=$1.0x10^{-13}$). The nearest SV to genes which have higher expression on the chromothriptic chromosome tend to be large SVs whereas the nearest SV to genes which have lower expression on the chromothriptic chromosome tend to be smaller SVs. SV types do not separate based on size (Figure 4.3).

Importantly, the distance relationship was not true for SV distance to genes which were upregulated or downregulated on the wild-type haplotype (Wilcoxon rank-sum test, p-values = 0.43 and 0.38, respectively) where SV density is much lower. On the wild-type haplotype, genes that had higher expression on the wild-type haplotype, genes that had lower expression on the wild-type haplotype and non-differential genes had only a modest difference in the median distance to SVs, 642,272 bp, 625,459 bp and 604,683 bp respectively.

In the other samples, the SV density was much lower and therefore comparing structural variant distance of differential transcripts on chromothriptic and wild-type haplotypes is less informative.

4.4.5 Essential genes

There are core sets of genes whose function is essential to all cell types; there are additional sets of genes that may be essential to particular cell types; and there is likely a third set of genes which are essential in the context of malignant transformation. Among this third set are, for example, genes that could be unmasked by chromothripsis. An example of this could be by upregulation of a gene which is placed under the control of a highly active enhancer after a chromothriptic event.

Investigation of essential genes may allow identification of genes which are important mechanistically to the cell. This can be done using a CRISPR–Cas9 fitness screen. A gene is targeted by a guide RNA which leads to loss-of-function of the gene. The effect of this loss-of-function on cell survival can then be quantified. Of the 927 genes screened on chromosome 6 in WTSI-OESO_103, 143 genes were identified as essential. WTSI-OESO_143 and WTSI-OESO_152 also underwent CRISPR–Cas9 fitness screens and a similar number of essential genes were identified, 136 and 195 genes respectively.

These genes are essential on chromosome 6, however many of them will be essential to oesophageal adenocarcinoma generally and not essential due to the specific mutations present in the organoid. In order to identify genes that may be essential as a result of the chromothriptic event, genes which are essential in the other oesophageal adenocarcinomas were removed from the list of essential genes. This left 41 genes on chromosome 6 in WTSI-OESO_103. Of those, 7 genes were both essential and differentially expressed: *AKAP12*, *ASCC3*, *ROS1*, *CYB5R4*, *NKAIN2*, *MTFR2* and *KIAA1586*. These genes point to regions

which may have been disrupted by chromothripsis and provide potential candidates for further investigation. *AKAP12* will be revisited later in the chapter.

4.5 Differential protein binding and accessibility

Structural variation can also affect histone modifications leading to differential accessibility and subsequently differential protein binding and gene expression. Evidence of this is present in cancer genomes (Zhao and Shilatifard, 2019) and the complement of mutations has also been shown to affect which proteins are differentially bound, with different mutations leading to the same protein binding to different sites. This has been discussed in detail in the Introduction and this section aims to investigate the extent of histone modification and chromatin accessibility dysregulation seen when comparing chromothriptic and wild-type haplotypes.

4.5.1 Comparison of histone modifications on haplotypes

Cells were profiled for two active histone marks, H3K27ac and H3K4me3, and one repressive histone mark, H3K27me3. They were also profiled for CTCF, a TAD boundary marker. Overall chromatin accessibility was studied using ATAC-seq. Reads which passed initial mapping quality filters were assigned to the appropriate haplotype using criteria described in Chapter 2. For chromosome 6 in WTSI-OESO_103, 31.1% and 36.2% of ATAC-seq and ChIP-seq reads, respectively, were explicitly assigned to either the chromothriptic or wild-type haplotype and only these reads were used in the subsequent analysis. These explicitly assigned reads were further filtered to remove regions of LOH. These filtered reads were then remapped to the GRCh38 reference genome to compare histone marks and accessibility between the two haplotypes.

4.5.2 Summary of marks

Peaks were called in ChIP-seq and ATAC-seq reads using MACS2 (Zhang *et al.*, 2008). These peaks represent sites which are bound by a protein profiled using ChIP-seq or regions of open chromatin profiled by ATAC-seq. 2,704 different peak sites were called on chromosome 6 in WTSI-OESO_103, with overlapping peaks merged into a single peak site by MACS2. Total number of peaks called on wild-type and chromothriptic chromosomes were broadly similar, particularly when only using assigned reads in non-LOH regions (Figure 4.4). The total number of peaks called using assigned ATAC-seq reads in non-LOH regions vary between biological repeats more than that seen between biological repeats of ChIP-seq peaks. While

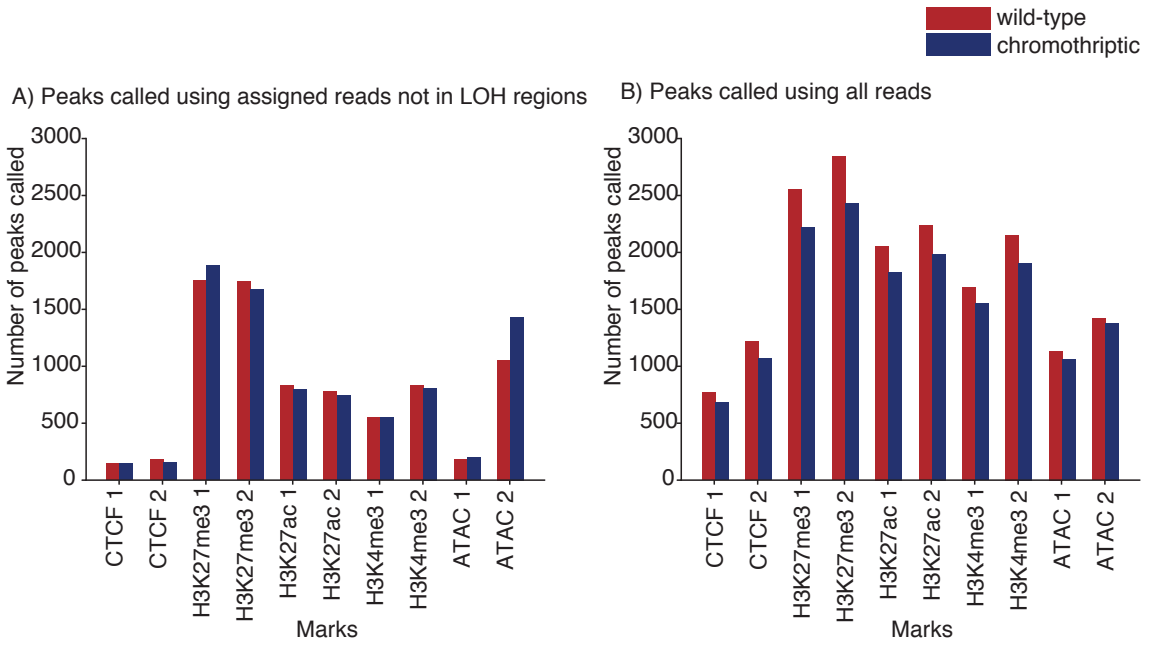


Fig. 4.4 Total number of peaks called by MACS for each ChIP-seq mark and for ATAC-seq peaks using A) only assigned reads not in regions of LOH and B) all reads. Two biological repeats were taken and are annotated by 1 and 2.

the difference seen when calling peaks using all reads is not as great, this difference may be representative of the dynamic nature of chromatin accessibility or indicate less technical reproducibility of the assay.

Some regions of the genome contained only active marks, other regions contained only repressive marks and the remaining regions contained both active and repressive marks at varying ratios. Each haplotype was segregated into 10 kb bins and the relative activity state was calculated. Activity state was determined using only the H3K4me3 and H3K27ac marks (active marks) and the H3K27me3 mark (repressive mark). To determine whether a region was more active or repressed the overall counts of active marks (A) verses repressed marks (R) were used. A sliding window of 30 kb was used to generate the overall score in a 10 kb bin. For each 10 kb bin (i), the counts were determined (B_i) and a sliding window was used to determine overall activity state (AS_i):

$$B_{\mathrm{i}} = A_i - R_i \tag{4.1}$$

$$AS_i = \frac{B_{\mathrm{i\text{-}1}} + B_{\mathrm{i}} + B_{\mathrm{i+1}}}{3} \tag{4.2}$$

When using assigned reads in non-LOH regions on chromosome 6 in WTSI-OESO_103, there were broadly similar total number of active versus repressed marks: 5,534 active marks and 5,156 repressive marks. Both the chromothriptic and the wild-type chromosomes had more active than repressive marks. However, the chromothriptic chromosome had fewer overall active marks (an excess of 66 active peaks) when compared to the wild-type chromosome (an excess of 313 active peaks). Transcriptionally active promoters tend to have high levels of H3K4me3, active cis-regulatory elements have high levels of H3K27ac and repressed genes have a high density of H3K27me3 (Gates *et al.*, 2017). Therefore, these marks provide a good indication of modifications in functionally important regions and an overview of the overall chromatin state. However, it is important to note that more histone marks would be needed to ascertain that the chromothriptic and wild-type chromosomes have similar relative proportions of active and repressive marks, with the chromothriptic haplotype being slightly less active. However, when considering only H3K27me3, H3K4me3 and H3K27ac histone marks on chromosome 6 in WTSI-OESO_103, this seems to be the case.

To determine whether this pattern was seen in other chromothriptic chromosomes and the normal difference in overall histone activity marks, chromosomes in other samples were investigated. The overall accessibility varied on different chromosomes and in different samples (Figure 4.5). There was no consistent accessibility pattern emerging in the subset of chromosomes investigated in these samples. Chromosomes in different samples may have broadly similar overall marks between haplotypes regardless of presence or absence of chromothripsis, for example chromosome 18. Other chromosomes have highly variable overall marks, for example chromosome 6. In WTSI-OESO_143 on chromosome 6, one haplotype has a large surplus of repressive marks (2201) and the other has a surplus of active marks (577). Conversely, in WTSI-OESO_152, both haplotypes of chromosome 6 are more repressive with very little difference between the total number of overall marks. Other chromosomes have very little difference between haplotypes of the same chromosomes in most samples, but some sample show large haplotype specific differences, for example chromosome 17 and chromosome 1 in WTSI-OESO_143. This highlights the high variability in the overall histone marks between chromosomes in different samples and also suggests that this may be sample-specific rather than chromosome-specific. Again, this may be indicative of the highly dynamic nature of chromatin accessibility.

4.5.3 Differential histone modifications and accessibility

Differential peaks were then called using all ChIP-seq marks as well as ATAC-seq reads. This was done using DiffBind (Ross-Innes *et al.*, 2012). In WTSI-OESO_103, chromosome 6 has

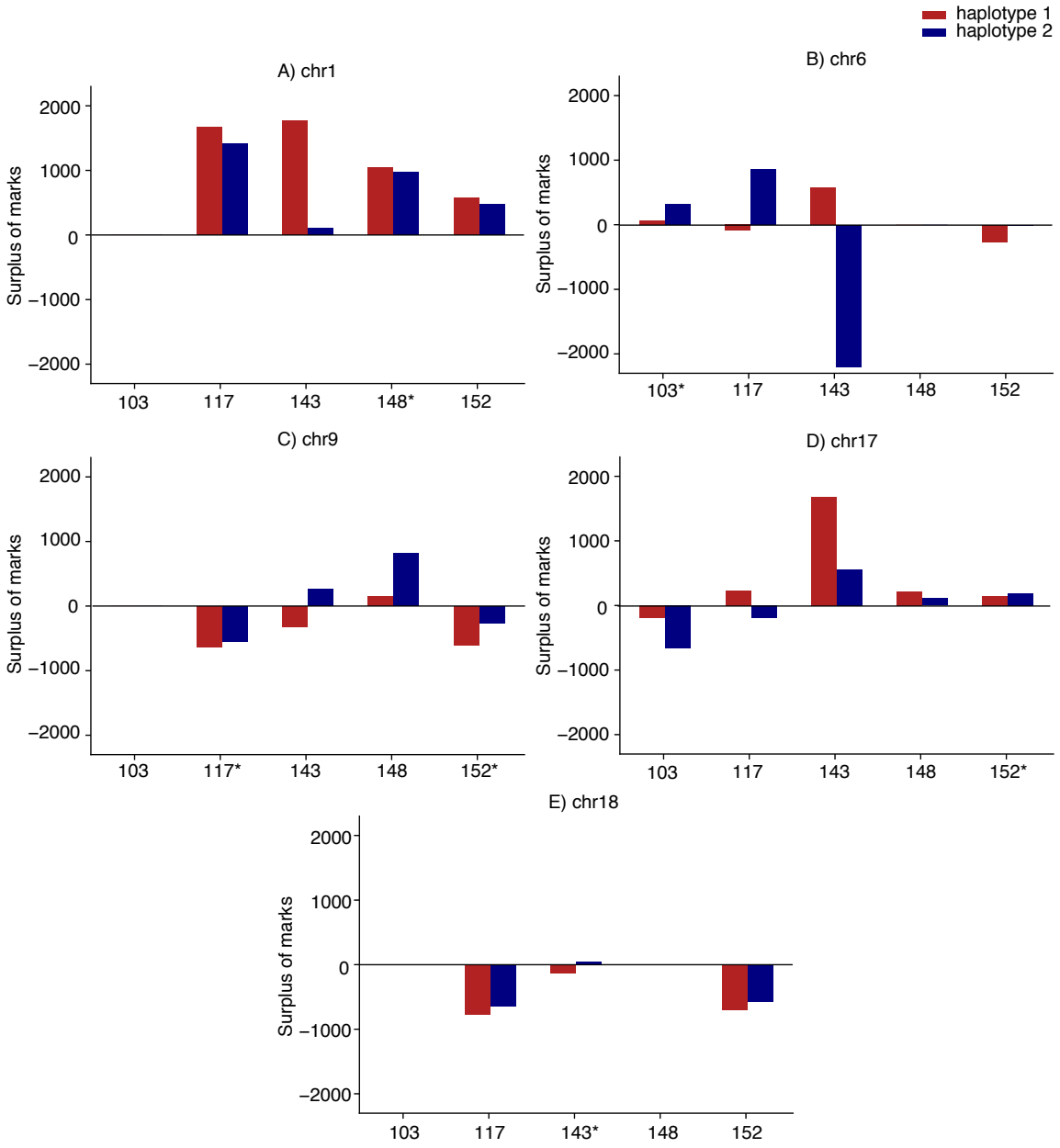


Fig. 4.5 The overall accessibility of chromatin calculated using assigned reads in non-LOH regions for H3K27me3, H3K4me3 and H3K27ac marks on chromosomes 1 (A), 6 (B), 9 (C), 17 (D) and 18 (E). A net negative number represents overall more repressive marks. A net positive number represents overall more active marks. Samples marked with asterisks indicate chromothripsis on haplotype 1. Chromosomes 1, 9 and 18 in WTSI-OESO_103 are highly rearranged so do not represent normal chromosomes and have been excluded from this analysis. WTSI-OESO_148 exhibits LOH on chromosomes 6 and 18 so no haplotype-specific data is present.

evidence of chromothripsis and 13.4% (361 total) of peaks were differential between the two haplotypes. 197 peaks were stronger on the chromothriptic haplotype and 164 peaks were stronger on the wild-type haplotype. In other samples, chromosome 6 is largely unaffected by rearrangements and therefore these chromosomes were used to determine the baseline number of differential ChIP-seq and ATAC-seq peaks between two haplotypes of chromosome 6. In WTSI-OESO_117, WTSI-OESO_143 and WTSI-OESO_152, 5.8%, 2.5% and 2.7% of chromosome 6 peaks were differential between the two haplotypes (Table 4.5). This suggests the chromothriptic event caused more differential histone modifications and accessibility than expected to be present between two normal haplotypes of chromosome 6 (Fisher's Exact test, p-value < 10^{-16} for all comparisons of chromosome 6 in WTSI-OESO_103 to chromosome 6 in other samples).

Table 4.5 Differential ChIP-seq and ATAC-seq peaks in other samples

	Percentage of total peaks which are differential (%)				
chr	WTSI-OESO_103	WTSI-OESO_117	WTSI-OESO_143	WTSI-OESO_148	WTSI-OESO_152
chr1	R	5.7	16.8	**38.5**	5.2
chr6	**13.4**	5.8	2.5	LOH	2.7
chr9	R	**11.9**	7.5	10.7	**8.1**
chr17	9.4	38.2	17.1	27.2	**12**
chr18	R	2.7	**16.3**	LOH	2.4

R = Rearranged. Chromosomes 1, 9 and 18 in WTSI-OESO_103 are highly rearranged so do not represent normal chromosomes and have been excluded from this analysis.

The chromothripsis present on chromosomes in other samples can be used to identify whether this is a feature specific to the high density of SVs and the particular biology of this sample, or whether it is seen as a general consequence of chromothripsis. In WTSI-OESO_148 and WTSI-OESO_152, the chromothriptic chromosomes exhibited more differential peaks between the chromothriptic haplotype and the wild-type haplotype than the normal chromosomes queried in that sample. Interestingly in WTSI-OESO_117, chromosome 17 exhibits the highest percentage of differential peaks of the chromosomes studied. While this chromosome is not chromothriptic, it becomes chromothriptic later in the tumour progression post-chemotherapy, as evidenced by chromothripsis present in WTSI-OESO_152. This high level of differential chromatin accessibility and histone modifications, alongside the high level of differential gene expression, provides further evidence of an altered chromatin state. The chromothriptic chromosome 9 in WTSI-OESO_117 has a high percentage of differential peaks, higher than the normal chromosomes if ignoring chromosome 17. In WTSI-OESO_143, the chromothriptic chromosome has a high percentage of differential peaks but it is similar to other chromosomes in sample. Strikingly, all the chromothriptic

chromosomes other than chromothriptic chromosome 17, had a higher level of differential peaks than seen on the same chromosome in other samples. The highest percentage of differential peaks on chromosome 17 was present in WTSI-OESO_117 and, again, this may be due to an altered chromosome state prior to chromothripsis occurring in this sample or as a consequence of the fusion of one haplotype into chromosome 6. It is also important to note that chromosome 17 in WTSI-OESO_143 and WTSI-OESO_148 had a higher percentage of differential peaks when compared to the percentage seen between chromothriptic and wild-type chromosomes in WTSI-OESO_152.

It is clear that there is huge variation across samples and between chromosomes in terms of haplotype-specific overall histone modifications (Figure 4.5). However, it is intriguing that the percentage of differential sites is consistently high when chromothripsis is present. It is unclear whether this is due to the underlying biology of the chromothriptic chromosomes, whether this occurs normally or whether this is a sporadic event that happened to have occurred on the chromothriptic chromosomes. It is clear that this difference is not as simple as one chromosome becoming overall more repressed or more active and it points to there being specific changes at sites on each haplotype. Further investigations are needed with a larger sample size in order to elucidate the mechanisms behind these observed differences.

4.5.4 Relationship between SVs and differential histone modifications and chromatin accessibility

The differentially bound sites on the chromothriptic and wild-type haplotype of chromosome 6 in WTSI-OESO_103 can be see in Figure 4.6A. Sites of differential binding cluster based on haplotype; chromothriptic marks cluster together as do wild-type marks. There is also clustering based on type of mark profiled, for example there are clusters of active histone marks and repressive histone marks specific to either the wild-type or the chromothriptic haplotype. For some of these haplotype-specific sites, SVs will directly alter binding sites, thereby preventing binding. However for many of these marks, the binding site will not be directly interrupted by an SV. Instead, the reason for the differences between haplotypes is less immediately obvious but may be attributed to a nearby SV disrupting the chromatin state.

To determine whether SVs are having a direct effect on differential accessibility and histone modifications between the two haplotypes, the distance of called peaks to SVs can be measured. From the ChIP-seq and ATAC-seq reads, a difference in protein binding and accessibility between the two parental chromosomes was observed. However, as with the Iso-seq reads, it is impossible to identify the affected allele. Therefore, there are two

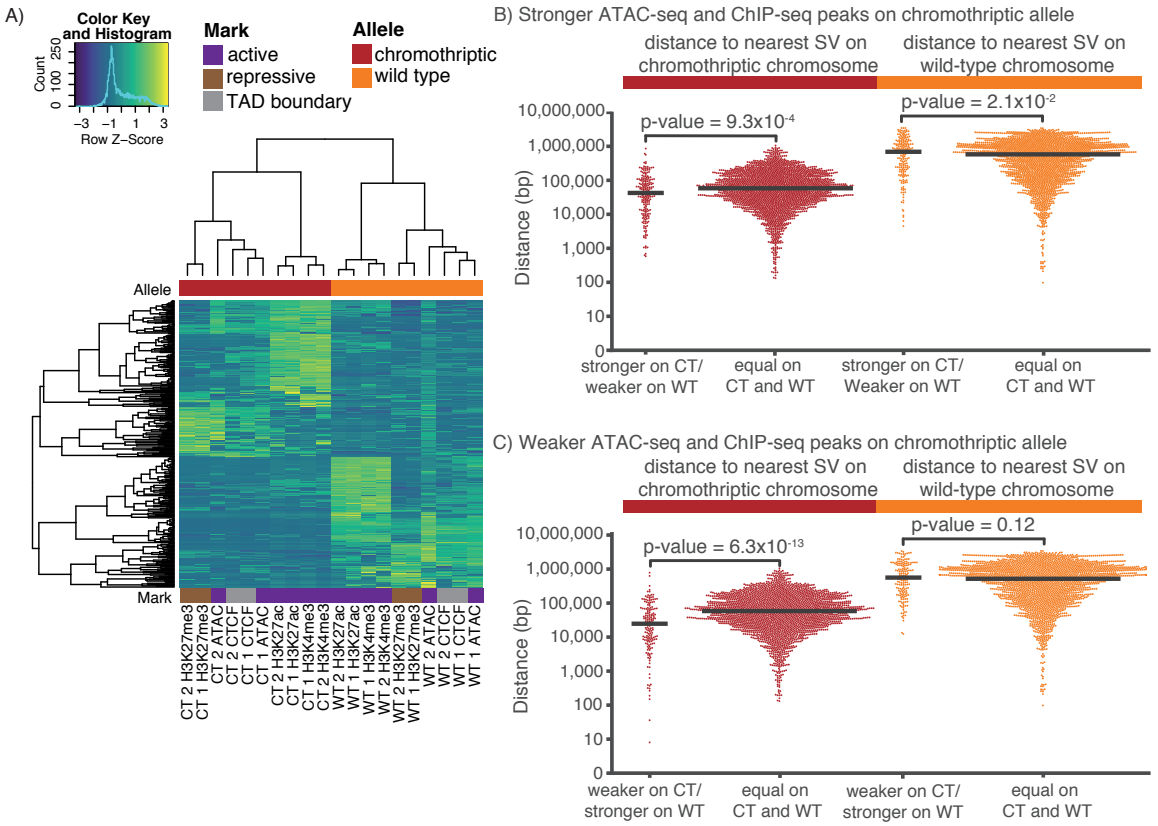


Fig. 4.6 Differential protein binding and histone modifications. A) Heatmap of differentially bound and open chromatin sites on the wild-type and chromothriptic chromosome 6 in WTSI-OESO_103. B) Distance of peaks which are stronger on the chromothriptic chromosome to the nearest SV, showing gain of binding. C) Distance of peaks which are weaker on the chromothriptic chromosome to the nearest SV, showing loss of binding.

categories to describe accessibility and protein binding peaks for this data: "stronger on chromothriptic/ weaker on wild-type" and "weaker on chromothriptic/ stronger on wild-type". Stronger binding in relation to accessibility from ATAC-seq represents a more open chromatin configuration. The distance of the chromatin peaks in these categories to SVs was determined on both the wild-type and chromothriptic alleles.

As with the Iso-seq reads, peaks which are stronger and those which are weaker on the chromothriptic haplotype were closer to SVs than non-differential peaks (Wilcoxon rank-sum test, p-value = 9.3×10^{-4} and p-value = 6.3×10^{-13}, respectively). This suggests that these SVs are causing both the formation of new and loss of old histone modifications as well as alterations in overall accessibility (Figure 4.6B,C). The median distance of peaks that are more weakly bound to SVs on the chromothriptic chromosome was smaller than the distance of peaks that are more strongly bound to SVs, 24,713 bp and 42,523 bp respectively. They

were both closer to SVs than the median distance of non-differentially bound sites to SVs, 58,207 bp.

Furthermore, like with differentially expressed genes, inversions tend to be close to peaks that have stronger binding on the chromothriptic chromosome whereas insertions tend to be close to genes which have weaker binding on the chromothriptic chromosome. More deletions occur near peaks which have stronger binding than weaker binding on the chromothriptic chromosome (Table 4.6). There is a difference between the size distributions of the nearest SV to differential peaks between peaks which are stronger and peaks which are weaker on the chromothriptic chromosome (Mann Whittney-U, p-value=7.6×10^{-34}). The nearest SV to peaks which have weaker binding on the chromothriptic chromosome tend to be small SVs. Conversely, there is a bimodal distribution of SV sizes near peaks which have

Table 4.6 SVs nearest to differential peaks

SV type	Stronger binding on CT	Weaker binding on CT
Deletion	65	32
Duplication	36	22
Insertion	9	99
Inversion	87	11

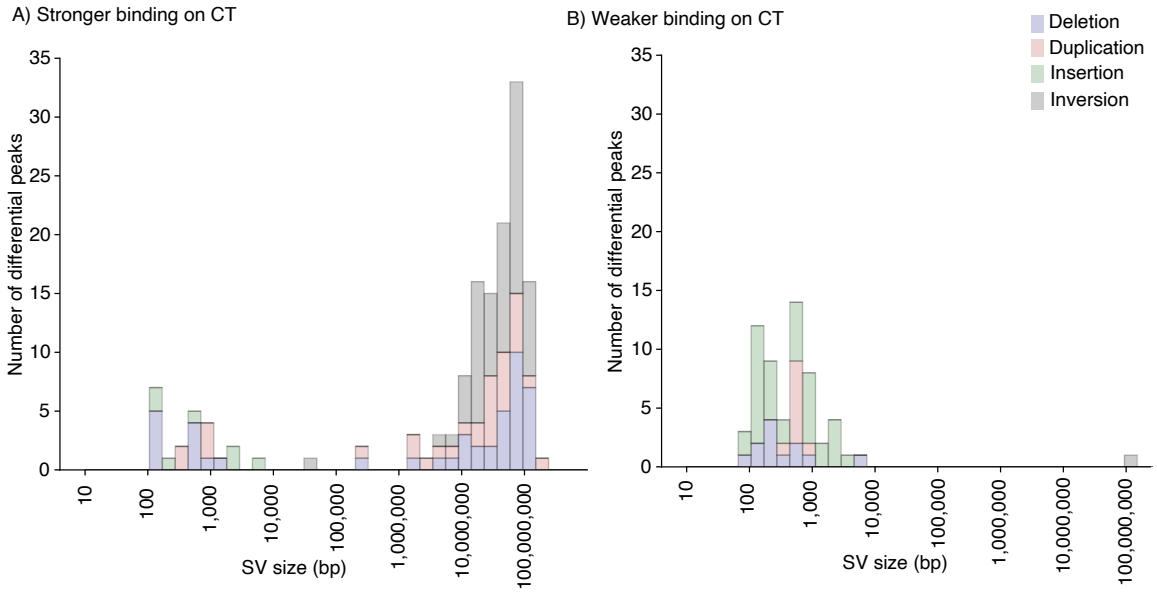


Fig. 4.7 SV size of the nearest SV to peaks with differential binding. A) Peaks with stronger binding on the chromothriptic chromosome. B) Peaks that have weaker binding on the chromothriptic chromosome.

stronger binding, with a larger peak from large SV sizes. SV types do not separate by size (Figure 4.7).

Furthermore, as with the gene expression data, the distance relationship is not true for peaks which are stronger on the wild-type haplotype (Wilcoxon rank-sum test, p-value = 0.12) and is a much weaker relationship for peaks which are weaker on the wild-type haplotype (Wilcoxon rank-sum test, p-value = 2.1×10^{-2}) where SV density is lower (Figure 4.6B,C). The median distance for sites which are more strongly bound on the wild-type haplotype, sites which have weaker binding and non-differentially bound sites have only a small difference in distance to SVs, 561,399 bp, 690,787 bp and 580,004 bp respectively.

Since the SV density was much lower in other samples, this analysis was not informative in other samples.

4.6 Topologically associating domains

Structural variation is known to affect large scale chromatin organisation by disrupting topologically associating domains (TADs). In developmental diseases, there has been evidence of neo-TADs forming which place regions of the chromosome under the control of different regulatory elements (Melo *et al.*, 2020). This has been discussed in detail in the Introduction and this section aims to investigate the extent of topologically associating domain alteration using the chromothriptic and wild-type haplotypes.

4.6.1 Comparison of long-range interactions

When re-mapping haplotype-resolved Hi-C reads to the GRCh38 reference genome, the usual TAD structures have clearly been altered by structural variation. This is evident in WTSI-OESO_103. When looking at an 800 kb window of chr6:116,860,000-117,660,000 on the reference genome, the wild-type chromosome has no SVs. Conversely, the chromothriptic chromosome has 11 rearrangements in this region (Figure 4.8 SV position on GRCh38 tracks). On the wild-type haplotype, there are obvious regions of high interaction and the formation of TAD structures (Arrow A on wild-type track of Figure 4.8). However, on the chromothriptic haplotype this region of high interactivity is lost. This is because the two regions which are interacting are in relatively close proximity in the wild-type haplotype, 520,196 bp apart (B and C in Figure 4.8). However, on the chromothriptic haplotype they have been moved to different parts of the chromosome, 2,139,896 bp apart. Therefore, they are no longer close in linear space or 3D space and do not interact. When viewed from the perspective of the reference genome, positions of SVs on the chromothriptic chromosome

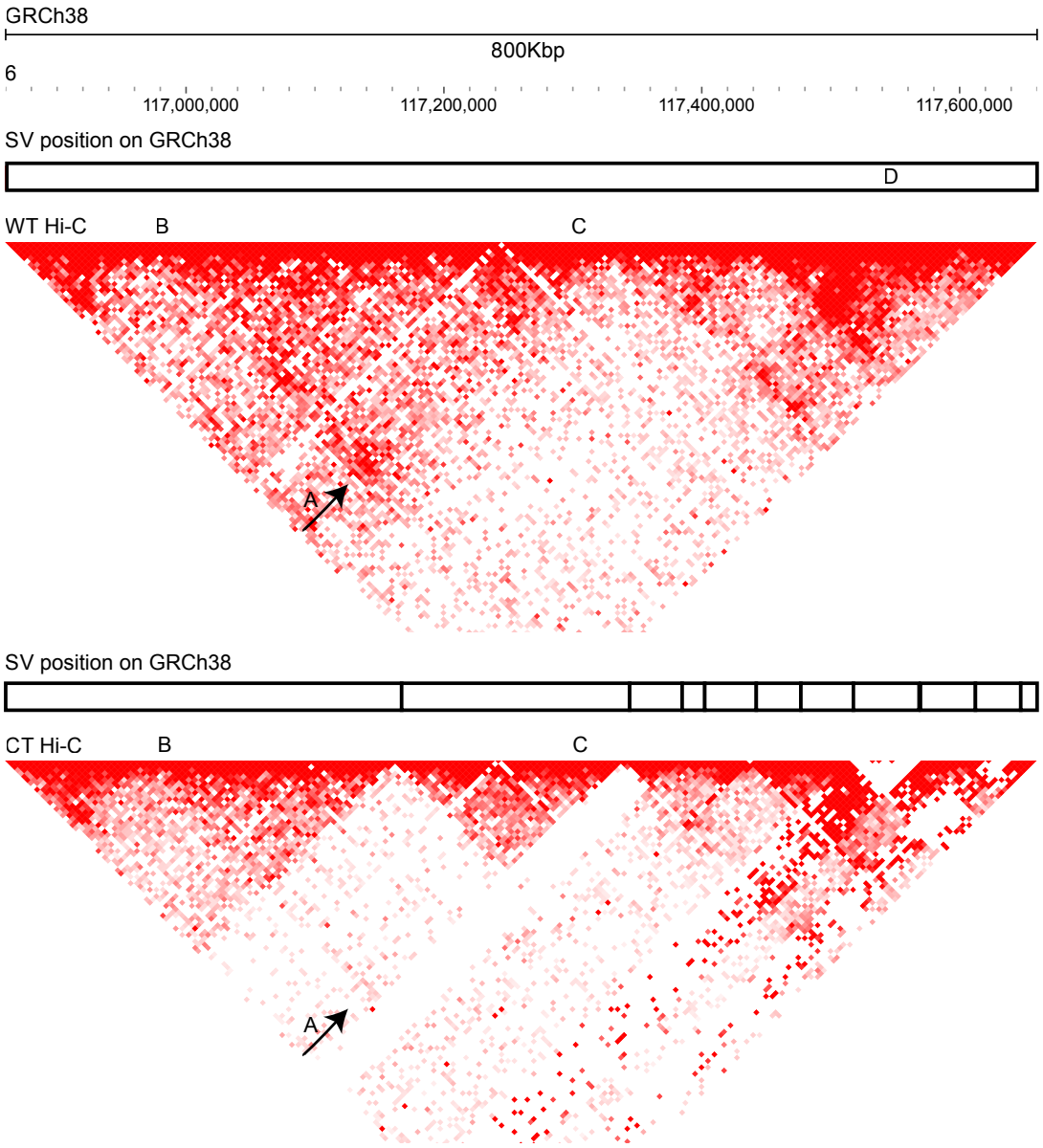


Fig. 4.8 TADs are disrupted by SVs. This plot shows WTSI-OESO_103 reads aligned to the GRCh38 reference genome at chr6:116,860,000-117,660,000. Every pixel is an interaction. The more red a region appears, the greater the number of contacts between those two regions. The SV postion on GRCh38 track show regions which are contiguous with the reference genome and no SVs present. A block is contiguous within itself but is not found next to the adjacent block in the assemblies. Arrow A shows an interaction that is present on the wild-type haplotype and missing in the chromothriptic haplotype. B and C are two regions that are adjacent in the wild-type chromosome and so interact but in the chromothriptic chromosome they have been moved away from each other and no longer interact WT = wild-type, CT = chromothriptic

usually demarcate boundaries of interaction, since two segments that might be adjacent on the reference chromosome will have been shuffled to very distant regions on the derivative chromosome.

Using the reference GRCh38 genome highlights that the original TAD structures can be disrupted by structural variation but this does not give an indication of what new structures, if any, form and how much they differ from the original structures. Custom assemblies, which have the haplotype-specific sequences, can be used to do this. When reads are mapped to custom assembles, equivalent regions on the two haplotypes can be identified and compared. These regions will vary in small events such as indels and SNPs but macroscopically will be very similar. In regions when there are no SVs, TAD structures on both haplotypes are the same with the same regions of high interaction (annotated by arrows in Figure 4.9). This is somewhat unsurprising as these regions are identical, if ignoring SNPs and indels.

In regions that are altered by SVs, some TAD structures become altered and some remain the same. On the wild-type chromosome, chr6:149,813,037-150,391,726 is completely contiguous. However, on the chromothriptic chromosome, it has been shattered by the chromothriptic event. This is depicted in Figure 4.10. Some regions have been lost (R3) and others have been retained (R1, R2, R4). This is an interesting area to query how these TAD structures have changed. R4 encompasses almost an entire TAD structure and R4 is retained in the chromothriptic chromosome. The interactions in R4 are very similar in the chromothriptic and wild-type chromosomes. In particular, the regions of high interaction between sequence A and sequence B are retained. Arguably the TAD structure in the chromothriptic chromosome is more defined, in particular, interactions of sequences in R4 with sequence B. Interestingly, R2 does not contain a full TAD structure and therefore the SV has disrupted the original TAD. The TAD in R2 on the chromothriptic chromosomes is smaller than the one seen on the wild-type chromosome with a different boundary. However, there are still interactions between sequence D and sequences in R2 so this compartmentalisation has been preserved. R1 also does not contain a full TAD structure. The chromothriptic event has caused the region upstream of chr6:149,813,037 to be placed adjacent to other regions of the genome. Consequently, this original TAD structure is no longer found on the chromothriptic chromosome. Instead, another TAD structure has formed incorporating the segment found upstream on the derivative chromosome (chr6:31,195,351-31,235,729 and Figure 4.10 arrow H). This example highlights the dynamic nature of chromatin contacts and the ability of new contact to form as well as old contains to persist when sequence context change.

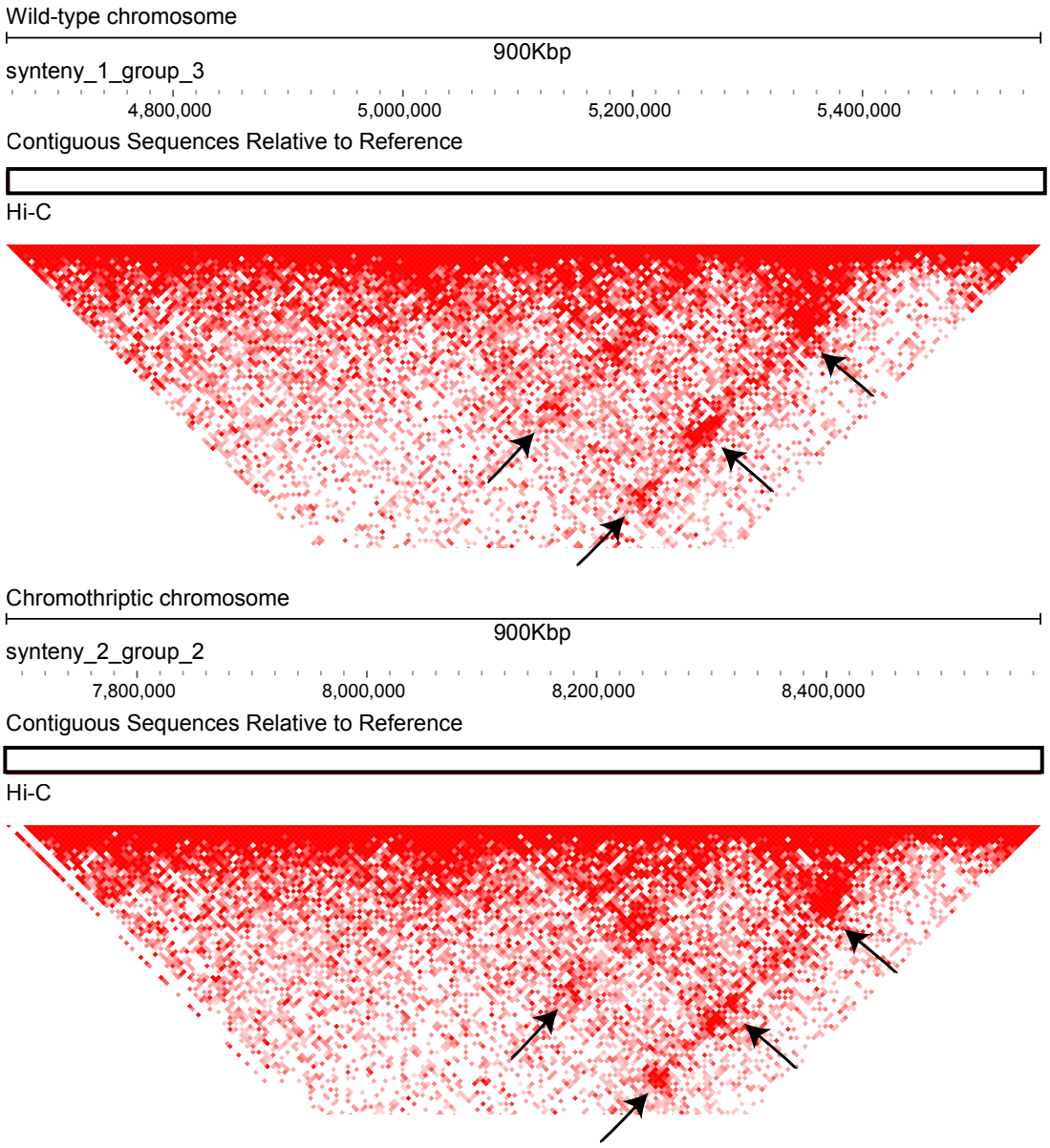


Fig. 4.9 Some TADs are conserved in both chromothriptic and wild-type haplotypes. This plot show regions in the custom assemblies of WTSI-OESO_103 that are equivalent to chr6:38,533,390-39,433,390. Every pixel is an interaction. The more red a region appears, the greater the number of contacts between those two regions. In regions where there have been no SVs, interactions are the same on wild-type and chromothriptic chromosomes. Regions of increased interactions are highlighted by arrows.

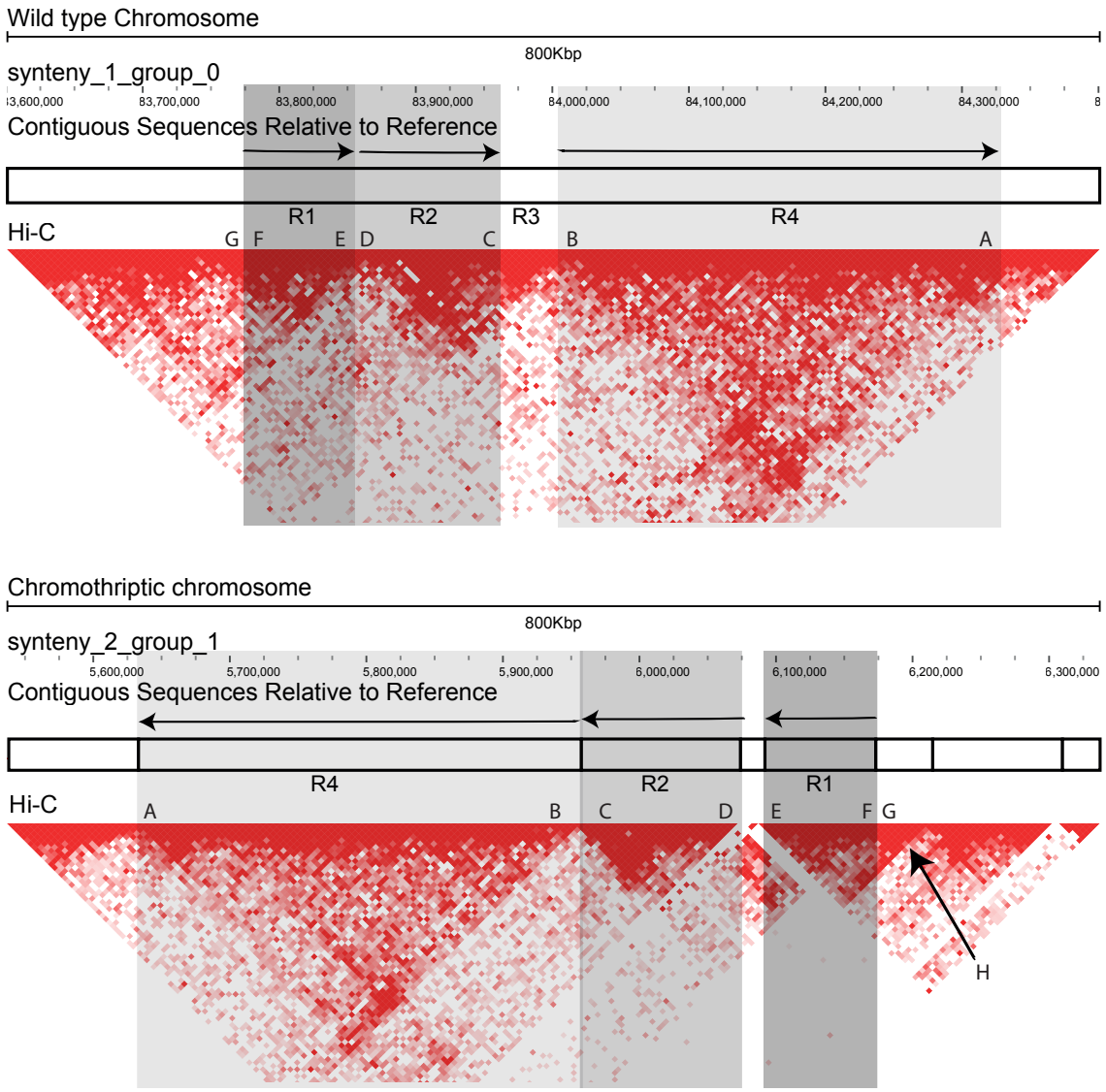


Fig. 4.10 These plots show an 800 kb region of the chromosome 6 custom assemblies of WTSI-OESO_103. The sequence is uninterrupted in the wild-type chromosome (top) but on the chromothriptic chromosome (bottom) the resultant sequence is the product of 8 SVs. The regions in grey are identical sequences in the two chromosomes if ignoring indels and SNPs and are the same as the sequence chr6:149,813,037-150,391,726 in the GRCh38 reference genome. Blocks in the contiguous sequences track are contiguous sequences found in the reference genome. A block is contiguous but is not found in the reference GRCh38 adjacent to the next block. The arrows denote orientation of the regions relative to the reference GRCh38 genome. Each pixel in the Hi-C track is an interaction. The more red a region is, the more interactions are occurring between the two sequences. While some TADs are conserved, there are also noticeable differences between the TAD structures on the chromosomes.

4.7 Integrative analysis of differential epigenomic patterns

Looking at expression, differential histone modifications or TADs in isolation will not be able to elucidate many mechanisms of dysregulation caused by structural variation. While the reason behind altered expression of a gene directly affected by SVs may be apparent, genes that are near SVs may have altered expression for a variety of reasons. For example, juxtaposing two regions which are in opposite accessibility states may lead to one of the regions changing accessibility state through altered histone modifications and subsequently a new chromatin configuration. Alternatively, these regions may remain in opposing accessibility states. The two regions may also form new interactions and these new interactions may be long-range and influence gene expression. Conversely the regions may remain functionally isolated from each other if they are encompassed in different TADs. It is likely that multiple mechanisms underlie the dysregulation of expression of a single gene. For example, when one of the juxtaposed regions has an altered accessibility state, it may result in alterations in chromatin contacts as regions that were inaccessible become accessible.

In order to gain a more complete understanding of the extent and mechanisms underlying gene expression dysregulation, different data types must be studied concurrently. This will allow understanding of mechanisms which span multiple levels of epigenetic regulation. Here, four such examples are discussed.

4.7.1 Altered histone marks

Altered gene expression can result from alterations in the overall histone activity state in a region. An example of this occurs in a 187 kb sequence which can be found in the reference GRCh38 genome at chr6:151,209,484-151,396,541 (Figure 4.11 grey boxes). This region is not affected by SVs on the wild-type chromosome but on the chromothriptic chromosome it has been taken out of its usual context and placed between regions of DNA which it is not usually in contact with (Figure 4.11 contiguous sequences track). This has led to *AKAP12* being expressed 570-fold higher on the wild-type chromosome than on the chromothriptic chromosome (Wald test, q-value = 2.2×10^{-113}) (Figure 4.11 dark grey box). Interestingly, the adjacent gene, ZBTB2, is not differentially expressed between the two alleles (Wald test, q-value = 0.98). Mechanistic understanding of these changes can be gained by examining the overall histone modifications surrounding the 187 kb segment. On the wild-type chromosome, the surrounding histone marks are mainly active marks (H3K27ac and H3K4me3) rather than

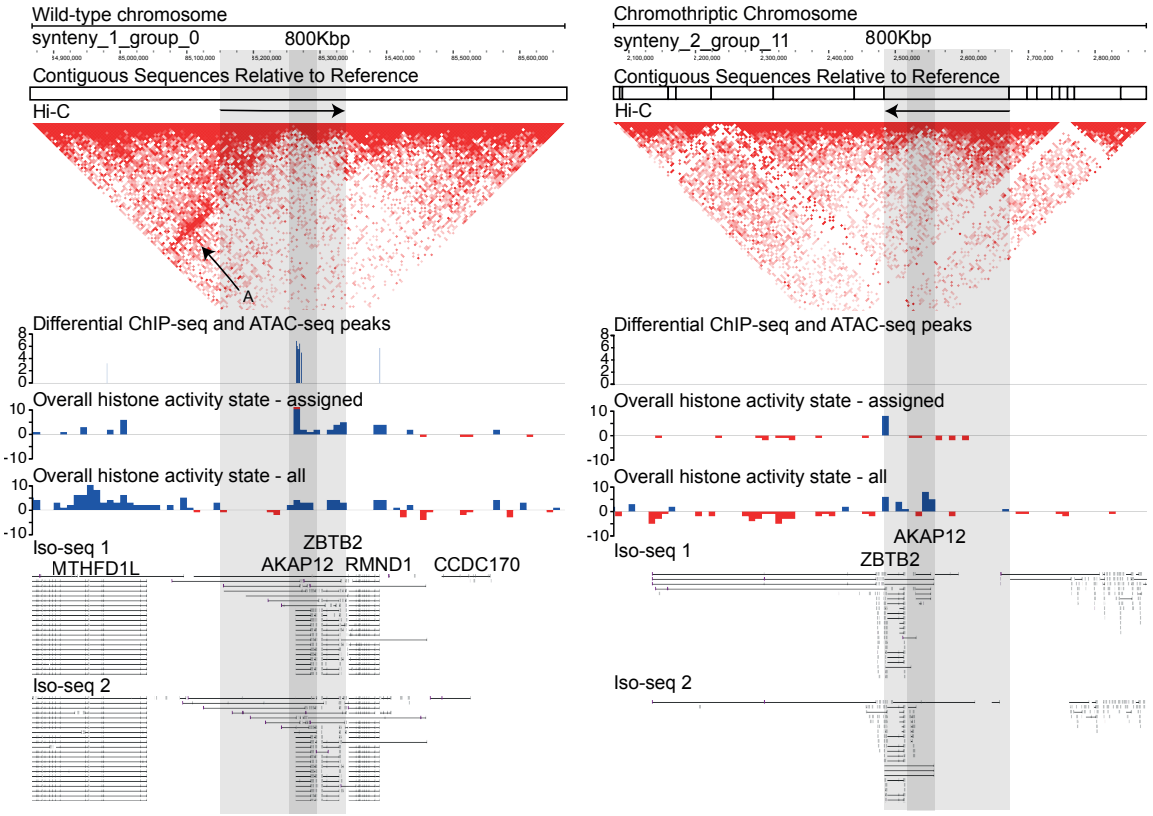


Fig. 4.11 These plots show an 800 kb region of the custom assemblies. Wild-type assembly is on the left and chromothriptic assembly is on the right. The regions in grey are identical sequences in the two chromosomes if ignoring indels and SNPs and are the same as the sequence chr6:151,209,484-151,396,541 in the GRCh38 reference genome. This region contains *AKAP12* and *ZBTB2*. The dark grey region highlights the *AKAP12* gene. Blocks in the contiguous sequences track are contiguous sequences found in the reference genome. A block is contiguous but is not found in the reference GRCh38 genome adjacent to the next block. The arrows denote orientation of the regions relative to the reference GRCh38 genome. Each pixel in the Hi-C track is an interaction so the more red a region is, the more interactions are occurring between the two sequences. Differential ChIP-seq and ATAC-seq peaks were called using DiffBind, and scale measures relative peak score. The peak is shown on the chromosome where that mark is upregulated. There are multiple differentially active sites on the wild-type chromosomes and no differential sites on the chromothriptic chromosomes in this region. One histone activity state track represents only peaks that can be explicitly assigned, and the other includes all peaks using both randomly and explicitly assigned peaks. Blue peaks are more active regions and red peaks are more repressed regions. Iso-seq tracks represent two independent biological repeats. Only a subset of reads are shown. Black lines show splicing and grey boxes represent exons

repressive marks (H3K27me3). This results in a more active region as a whole and expression of the gene. Conversely on the chromothriptic chromosome, the *AKAP12* gene has been placed into a more repressive environment, despite the histone marks directly surrounding the *AKAP12* gene remaining active. There are also differentially active histone marks which are specific to the wild-type chromosomes but lost on the chromothriptic chromosome. These marks may be essential to full expression of *AKAP12*. The region containing the differentially active marks is interacting with the upstream region on the wild-type chromosome (Figure 4.11 Hi-C track arrow A), however this interaction is lost when this sequence is placed in a different context on the chromothriptic chromosome. New interactions occur with adjacent sequences upstream of this region on the chromothriptic chromosome – these interactions link the *AKAP12* gene body with a generally repressed segment of DNA and this may also contribute to the reduced *AKAP12* expression on the chromothriptic chromosome. Loss of these specific differential interactions may be why *AKAP12* expression is altered but *ZBTB2* expression is unaffected. Together, this example shows how altered gene expression can be caused by alteration in the histone modifications of surrounding regions and altered contacts with adjacent sequences.

4.7.2 Altered chromatin interactions

Altered gene expression can occur in the absence of a change in overall histone activity in a region. An example of this occurring is present in a 485 kb sequence which can be found in the reference GRCh38 genome at chr6:124,941,256-125,426,117 (Figure 4.12 grey box). SVs occur around this segment in the chromothriptic but not the wild-type chromosome. This 485 kb sequence, including the *TPD52L1* gene, is removed from its normal context and joined to other segments of chromosome 6. *TPD52L1* is upregulated on the wild-type chromosome 19.2-fold relative to the chromothriptic chromosome (Wald test, q-value = $2.3\text{x}10^{-2}$). However, unlike the *AKAP12* gene, this upregulation cannot be explained by the region changing overall histone activity state as both chromosomes have similar surrounding histone marks (Figure 4.12 overall histone activity state tracks). There are also no specific differential histone marks between the two chromosomes (Figure 4.12 differential ChIP-seq and ATAC-seq peaks track). However, insight may be gained by looking at chromosomal contacts and alterations in TAD structures (Figure 4.12 Hi-C track). On both chromosomes, the 111 kb region containing *TPD52L1* (Figure 4.12 dark grey box) lies within a TAD structure. There are many interactions between the start of the *TPD52L1* gene with upstream sequences on the wild-type chromosome and these interactions may be important in regulation of gene expression (Figure 4.12 Hi-C track arrow A). However, on the chromothriptic chromosome, these interactions are lost as the sequences are no longer in

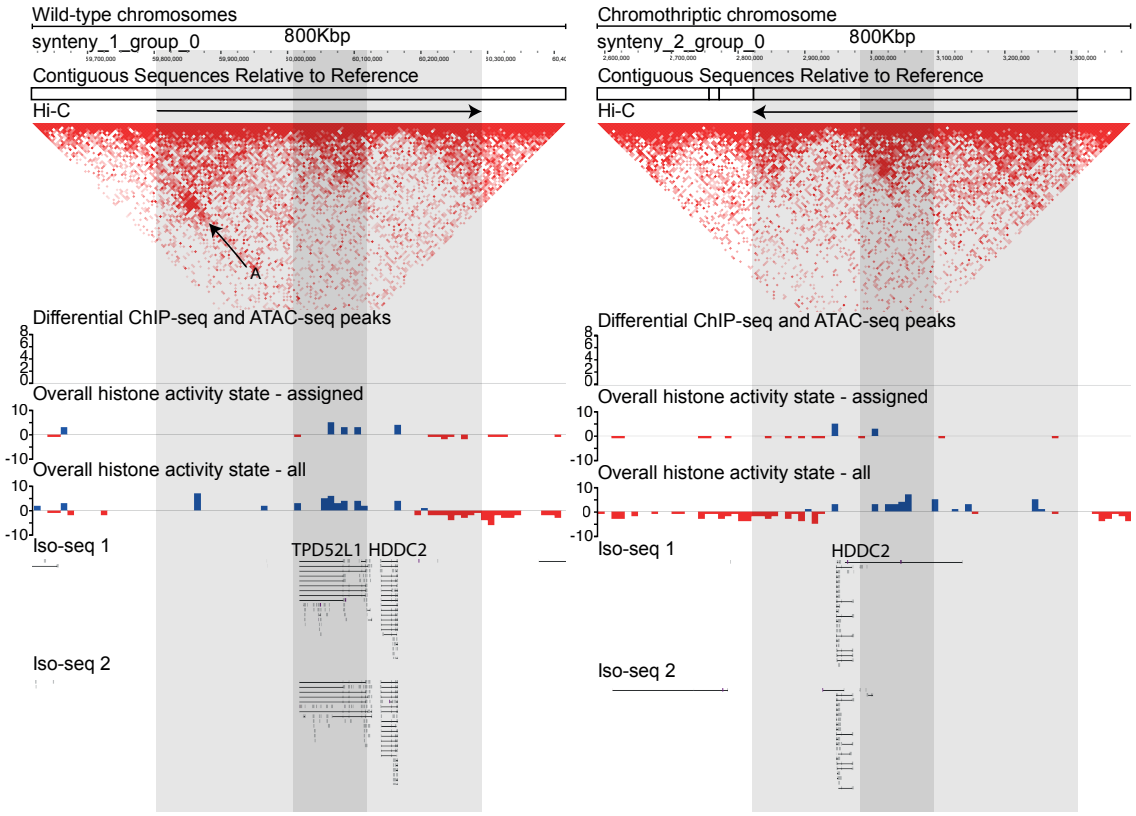


Fig. 4.12 These plots show an 800 kb region of the custom assemblies. Wild-type assembly is on the left and chromothriptic assembly is on the right. The regions in grey are identical sequences in the two chromosomes if ignoring indels and SNPs and are the same as the sequence chr6:124,941,256-125,426,117 in the GRCh38 reference genome. The dark grey region highlights the *TPD52L1* gene. Blocks in the contiguous sequences track are contiguous sequences found in the reference genome. A block is contiguous but is not found in the reference GRCh38 genome adjacent to the next block. The arrows denote orientation of the regions relative to the reference GRCh38 genome. Each pixel in the Hi-C track is an interaction so the more red a region is, the more interactions are occurring between the two sequences. Differential ChIP-seq and ATAC-seq peaks were called using DiffBind, and scale measures relative peak score. One histone activity state track represents only peaks that can be explicitly assigned, and the other includes all peaks using both randomly and explicitly assigned peaks. Blue peaks are more active regions and red peaks are more repressed regions. Iso-seq tracks represent two independent biological repeats with expression of *TPD52L1* only on the wild-type chromosomes. Only a subset of reads are shown. Black lines show splicing and grey boxes represent exons. Arrow A highlights regions of high interaction.

close spatial proximity (Figure 4.12 Hi-C track). Segregation of *TPD52L1* from its usual contacts and loss of local regulation may be the cause of this reduced expression of *TPD52L1* on the chromothriptic chromosome.

4.7.3 Fusion gene formation

Fusion genes can also occur as a result of structural variation. An example of this occurs in a 282 kb sequence which can be found in the reference GRCh38 genome at chr6:75,536,872-75,819,360 (Figure 4.13 grey box, note opposite orientations). This region contains both the *SENP6* gene and the *MYO6* gene. There is no difference in expression of *SENP6* but *MYO6* is downregulated 3.9-fold on the chromothriptic chromosome (Wald test, q-value = 5.5×10^{-5}). On the chromothriptic chromosome, a structural variant has fragmented the *MYO6* gene, however there is still low expression of the first half of transcript. Like the region containing the *TPD52L1* gene, the activity state of the regions containing the *SENP6* and *MYO6* genes are similar on the chromothriptic and wild-type chromosomes (Figure 4.13 overall histone activity state tracks, regions are in opposite orientations) and there are no differential ChIP-seq and ATAC-seq peaks within or surrounding these genes. However, there are alterations in the chromatin contacts. On the wild-type chromosome, the 282 kb region forms a TAD and very few interactions occur outside of this region. However, on the chromothriptic chromosome, a new larger TAD forms with interaction upstream of the *MYO6* gene. Some of these new interactions manifest as novel fusion genes. There is a fusion of *SENP* and *MYO6* and both genes lie within a TAD (Figure 4.13 read A). There is also a fusion between *MYO6*, *ASCC3* and *TAB2* as *ASCC3* and *TAB2* have been placed upstream of *MYO6* by the chromothriptic event (Figure 4.13 read B). This produces a complex fusion gene. The segment of *MYO6* begins at the start of the gene. The segment of *ASCC3* incorporated into this transcript contains only one exon as well as some intronic sequence. The segment of *TAB2* begins within an intron. *ASCC3* and *TAB2* are in the same orientation as the gene reference but in opposite orientation relative to *MYO6*. Regions of high interaction between the MYO6-ASCC-TAB2 fusion regions can been seen in Hi-C contact matrices (Figure 4.13 Hi-C contact matrix and arrows C and D).

4.7.4 Unexplained alterations

In some cases, the cause of the alterations in gene expression is unclear. An example of this occurs in a 48 kb sequence which can be found in the reference GRCh38 region chr6:37,433,221-37,481,508 and encompasses the *CMTR1* gene (Figure 4.14 dark grey box). The chromothriptic event has shattered the sequence surrounding the 48 kb segment

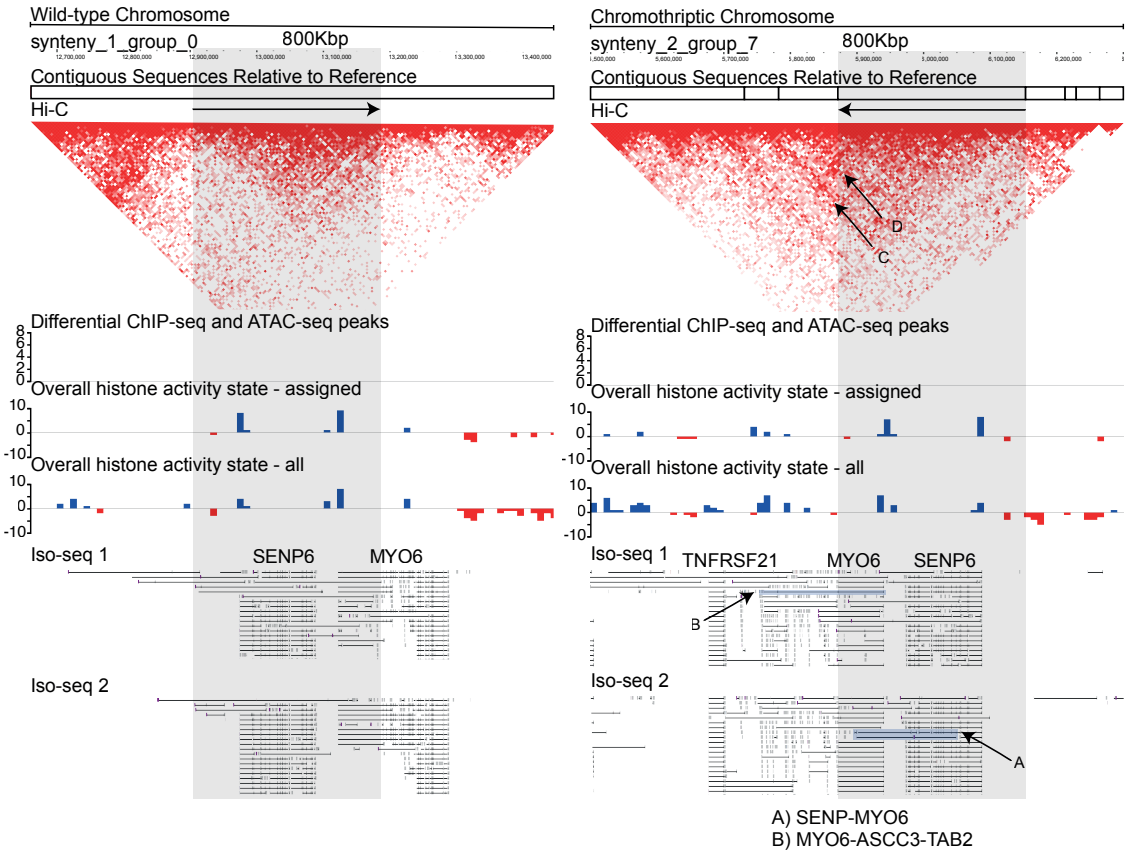


Fig. 4.13 These plots show an 800 kb region of the custom assemblies. Wild-type assembly is on the left and chromothriptic assembly is on the right. The regions in grey are identical sequences in the two chromosomes if ignoring indels and SNPs and are the same as the sequence chr6:75,536,872-75,819,360 in the GRCh38 reference genome. Blocks in the contiguous sequences track are contiguous sequences found in the reference genome. A block is contiguous but is not found in the reference GRCh38 genome adjacent to the next block. The arrows denote orientation of the regions relative to the reference GRCh38 genome. Each pixel in the Hi-C track is an interaction so the more red a region is, the more interactions are occurring between the two sequences. Differential ChIP-seq and ATAC-seq peaks were called using DiffBind, and scale measures relative peak score. One histone activity state track represents only peaks that can be explicitly assigned, and the other includes all peaks using both randomly and explicitly assigned peaks. Blue peaks are more active regions and red peaks are more repressed regions. Iso-seq tracks represent two independent biological repeats and show expression of many genes including *MYO6* and *SENP6* on both chromosomes. Only a subset of reads are shown. Black lines show splicing and grey boxes represent exons. A *SENP6-MYO6* fusion transcript is highlighted by read A and a *MYO6-ASCC3-TAB2* fusion transcript is highlighted by read B.

leading to multiple rearrangements. *CMTR1* is expressed on the chromothriptic chromosome 47.9-fold higher than on the wild-type chromosome (Wald test, q-value = $3.9x10^{-19}$). There is no differential expression of neighbouring gene *RNF8*. The overall histone marks over both genes are similar on both the chromothriptic and wild-type allele (Figure 4.14 histone activity state track). The TAD structure encompassing these genes also have similar overall interactions (Figure 4.14 Hi-C track arrows), although the adjacent TADs have different histone activity states. There are also no immediately discernible interaction which could explain the release of inhibition of *CMTR1* on the chromothriptic allele. The over-expression of *CMTR1* with no change to *RNF8* highlights that there is very careful control of gene expression which allows for only specific genes in a region to be affected. In both this organoid line and the others studied, *CMTR1* is an essential gene, as defined by the CRISPR knockout studies. This suggest that this gene is important mechanistically to the cancer but the exact mechanism remain elusive.

The answer to why *CMTR1* is only being expressed on the chromothriptic chromosome may lie in regional context. There are subtle differences between interactions occurring within the TAD structure encompassing *CMTR1* on the wild-type and chromothriptic chromosome. The interactions on the chromothriptic chromosome are less pronounced than those on the wild-type and some of these may be important in hindrance of *CMTR1* expression on the wild-type chromosome. If these interactions are in fact weaker on the chromothriptic chromosome, this weaken level of interaction may not be enough to release inhibition of expression. However, looking at the resolution of expression, accessibility and topologically associating domains, it is impossible to say for sure is there is release of inhibition or simply an artefact of the inherently noisy chromatin interactions. Further epigenetic layers will be needed to fully elucidate the mechanism occurring on *CMTR1*.

4.7.5 Overview

Together these examples highlight an array of ways in which structural variation can lead to alterations in gene expression by changes in chromatin accessibility, histone modifications and TAD structures. In some cases, differential gene expression could be explained by SVs directly fragmenting a gene. However, in other cases, such as the examples outlined above, the underlying mechanisms could not be understood without a multi-omic approach. In all examples outlined above, the cause of differential expression could not be attributed to rearrangements directly affecting genes - instead, the rearrangements caused alterations in higher order structures which changed the overall chromatin landscape. It is important to note that while we are using a multi-omic approach, there will undoubtedly be information in other epigenetic layers not studied here. Inclusion of more epigenetic layers will help

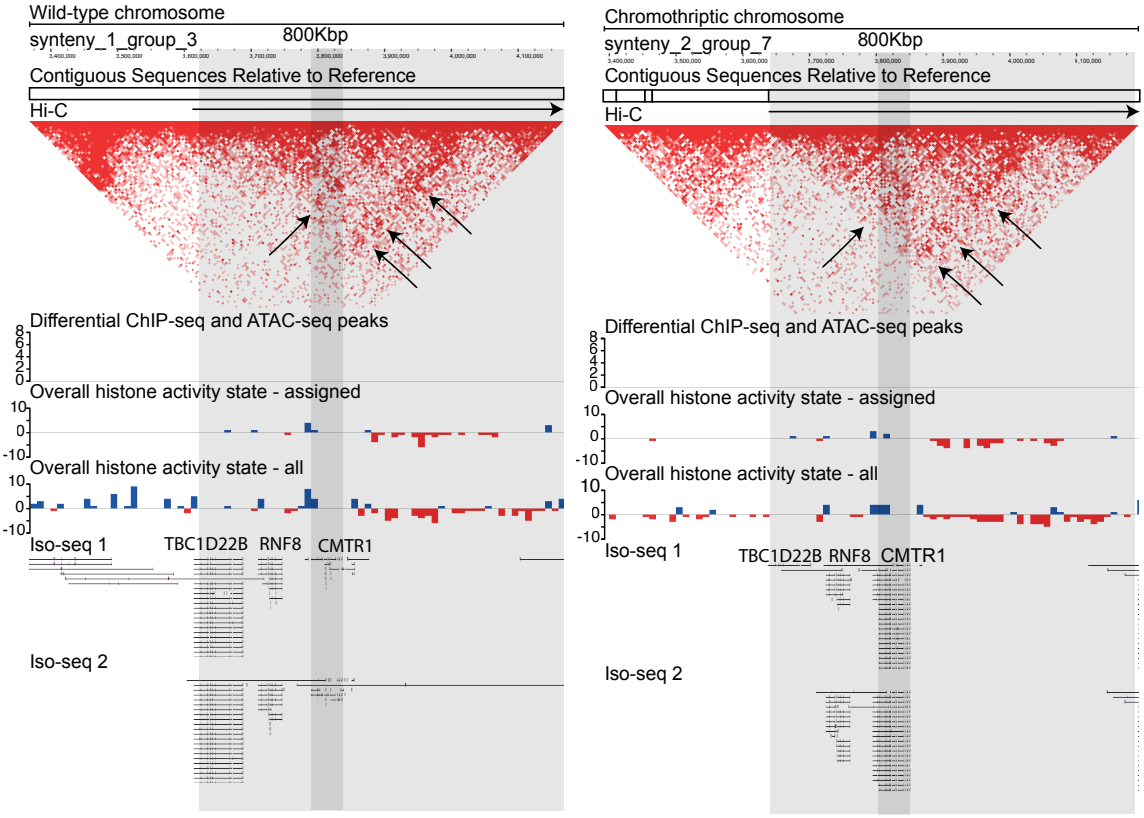


Fig. 4.14 These plots show an 800 kb region of the custom assemblies. Wild-type assembly is on the left and chromothriptic assembly is on the right. The regions in grey are identical sequences in the two chromosomes if ignoring indels and SNPs and are the same as the sequence chr6:37,267,542-37,828,872 in the GRCh38 reference genome. This region contains *CMTR1* and *RNF8*. The dark grey region highlights the *CMTR1* gene. Blocks in the contiguous sequences track are contiguous sequences found in the reference genome. A block is contiguous but is not found in the reference GRCh38 genome adjacent to the next block. The arrows denote orientation of the regions relative to the reference GRCh38 genome. Each pixel in the Hi-C track is an interaction so the more red a region is, the more interactions are occurring between the two sequences. Differential ChIP-seq and ATAC-seq peaks were called using DiffBind, and scale measures relative peak score. One histone activity state track represents only peaks that can be explicitly assigned, and the other includes all peaks using both randomly and explicitly assigned peaks. Blue peaks are more active regions and red peaks are more repressed regions. Iso-seq tracks represent two independent biological repeats. Only a subset of reads are shown. Black lines show splicing and grey boxes represent exons.

elucidate further mechanisms, such as the mechanism behind upregulation of *CMTR1* on the chromothriptic chromosome.

4.8 Discussion

Using cancer-specific haplotype-resolved genome assemblies has been pivotal in elucidating the underlying mechanisms surrounding genome dysregulation caused by chromothriptic structural variation. Haplotype-specific alterations caused by chromothripsis can be mechanistically important to the biology of the cancer. Here an attempt has not been made to understand specific mechanisms underlying oesophageal adenocarcinoma biology but instead these cancers have been used as a model to understand how alteration in primary genome structure affects higher order chromatin biology. This chapter has demonstrated that SVs are causing differential changes in gene expression, histone modifications, chromatin accessibility and topologically associating domains. This may be due to SVs directly occurring within genes or at sites of chromatin marks or through long-range effects and altered compartmentalisation of chromatin. Unsurprisingly, the overall message is complex as structural variation can impact all epigenetic layers and the overall effect on gene expression is dependent on a multitude of factors. This chapter has also highlighted a few examples where there are specific differences between haplotypes. In these cases, there is an obvious need for combining multiple epigenetic layers in order to understand why haplotype specific differences occur (Figures 4.11, 4.12, 4.13 and 4.14). This highlights the need for future studies investigating complex rearrangements to be multi-omic and transverse multiple epigenetic layers.

Five oesophageal adenocarcinoma organoids have been studied in this chapter and the methodology presented here may be the basis for future studies. Some intriguing findings have been presented, such as higher percentage of differential ChIP-seq and ATAC-seq peaks between chromothriptic and wild-type haplotypes and the unclear trend of percentage of differential transcripts on different chromothriptic chromosomes. While these initial findings are interesting starting points, larger sample sizes and more chromosomes need to be studied to determine whether these findings are common features of large-scale chromosome reorganisation or whether sporadic differences between haplotypes are commonplace. These future studies may also incorporate other epigenetic factors that have not been studied in this thesis. These factors will be both affected by structural rearrangements and important mechanistically in understanding the overall effect of structural variation on higher order structuring. Obvious examples include methylation, chromatin looping and other histone marks. By including more samples and more epigenetic layers, it will be possible to get closer to understanding the complexity of gene regulation by higher order structures.

It is also interesting to consider the findings from this chapter in the context of the finding from a study of the effect of rearrangement on gene expression and genome topology in Drosophila (Ghavi-Helm *et al.*, 2019). These Drosophila chromosomes exhibit altered TAD structure with 76% of the genome being affected by rearrangement. Similar to the alterations seen in the TAD encompassing *AKAP12*, the alterations in the Drosophila appear to alter only a subset of genes within that TAD rather than globally altering all genes within a TAD. Furthermore, in Drosophila, differential genes have twice as many differential promoter contacts but the majority of altered promoter contacts do not affect gene expression suggesting an uncoupling of genome topology and gene expression (Ghavi-Helm *et al.*, 2019). Difference in specific mechanisms, such as the lack of activation of genes in reshuffled TADs, may be due to biological differences Drosophila and humans. Alternatively, chromothriptic chromosome undergo massive selection as rearrangements which are too detrimental are lost and this level of selection was not present in the study in Drosophila. Both studies conclude that genome regulation is highly complex and rearrangement of primary genome sequence does not globally alter all genes in a specific region, instead a few genes are altered while the majority are unaffected. Both studies also conclude that specific changes are caused by SVs, even if there are neighbouring genes which remain unaffected. In this chapter mechanisms behind certain alterations have been elucidated, however neither this chapter nor the study in Drosophila have systematically elucidated why this may be the case genome wide.

While the cancer genomes in this thesis have been comprehensively characterised, it is important to note that this characterisation was restricted to only a subsection of the genome. The proportion of the genome characterised was dependent on a variety of factors including chromosome SNP density within a particular sample, density of SVs and variants which cause differences in read mapping and specific properties of the data being phased. SNP and variant density are fixed within a sample. Read length of the underlying data had a considerable impact of the proportion of the genome being phased. This is apparent as the proportion of Iso-seq reads phased was greater than the proportion of ChIP-seq, ATAC-seq or Hi-C reads phased. This read length effect is seen even over small increases in read length. 7.1% more Hi-C reads were phased than ChIP-seq reads. Furthermore 5.1% more ChIP-seq reads were phased than ATAC-seq reads. This highlights that more of the genome is accessible with longer read lengths.

While the study has been restricted to only certain regions of the genome, this does not take away from the insights gained in the chapter. Genome regulation is complex and subtle changes in one epigenetic layer can have profound impacts on cellular phenotype which may not necessarily be immediately intuitive. Structural variation provides an opportunity to query the role of primary sequence in regulation of phenotype and its relationship with

epigenetic regulation. The methods developed and used in this chapter can be used in future studies to further elucidate gene expression regulation and the dysregulation seen in disease. It is a versatile approach for haplotype resolution of a variety of sequencing data types which is tuned for features of the underlying data.

Chapter 5

Conclusions and outlook

5.1 Conclusions

This study is the first that I am aware of to have reconstructed highly rearranged chromosomes in a haplotype-aware manner and subsequently used these assemblies to query allele-specific changes in gene expression and higher-order structuring. A haplotype-aware approach has been vital in the reconstruction of these highly rearranged chromosomes.

Chromosomes with simple rearrangements may be reconstructed by generating a single assembly for both alleles, known as a collapsed assembly, and then separating the two haplotypes. However, chromosomes with extreme variation between alleles exhibit large regions of identical sequences as well as large regions of divergent sequences. At regions where an identical sequence in the two chromosomes become divergent, a haplotype-unaware approach was not able to reconstruct both sequences and this subsequently led to a break in the assembly. Haplotype-aware methods circumvented this problem by segregating reads and reducing, though not totally eliminating, the number of instances in which fragmentation occurred. However, segregation of reads required both long and highly accurate reads. The advent of CCS PacBio sequencing meant that this was possible, albeit non-trivial. Reconstruction of cancer samples has the added complexity of subclonal variation, complex structural variation and copy number amplifications. However, despite these complexities, many key features of cancer genomes can be reconstructed using the methodology presented in Chapter 3. The method is also robust to varying degrees of complexity ranging from normal alleles unaffected by large scale chromosomal rearrangement to simple rearrangements to complex reshuffling.

While the genome assemblies presented in this thesis are in no way perfect, they allowed investigation of the effect of altering primary genome sequence on higher-order chromatin structure. Specifically, I could use the assemblies to assess how the 3D genome structure

and associated gene expression change when a sequence is taken out of its normal regulatory context and is put into a different regulatory environment. The answer to that is far from simple. While primary sequence is a key driver of gene expression, slight variations in primary sequence can cause altered higher-order structuring and this can be used to explain the altered phenotype. This occurs at the level of chromatin contacts, histone modifications and chromatin accessibility and there is often interplay between these levels. It is clear that structural variation is having a direct effect on this alteration of higher-order structure, as evidenced by altered expression and histone modifications on the chromothriptic allele being closer to structural variation than marks which have not been altered. The equivalent statement is not true for wild-type alleles where structural variant density is much lower.

Through the combination of haplotype-resolved assembly and integrating across multiple sequencing platforms, several concrete examples of the chromothripsis-induced alteration of the primary genome sequence and subsequent altered gene expression have been identified. This appears to occur via several different mechanisms including loss and gain of histone modifications, altered overall accessibility and newly formed chromatin contacts. Although many questions remain, this approach has allowed for an unprecedented mechanistic insight into the relationship between primary genome sequence and higher-order structure and would not have been possible without haplotype resolution or the mutli-platform approach.

5.2 Future work

Remarkable insight into genome regulation has been gained from this study. However, it has also uncovered many avenues for future investigation. The remainder of this chapter will describe future work needed to better understand the regulation between primary genome sequence and higher-order structure.

5.2.1 Cancer genome assembly

Genome sequencing has become commonplace and is used in the study of cancer genomics to query the presence and effect of mutations. Cancer-specific genome assemblies can be used to reconstruct complex genomes and better understand mechanisms of dysregulation that arise from complex rearrangement. This may not be necessary for simple rearrangements or even cancers where the mutational burden is low.

The methodology presented here is robust to a wide range of structural rearrangement. However, it is restricted by density of variants which can be used to assign reads to one chromosome over another. The longer the read length, the more likely the read is able to

cover an informative variant and the better the chromosome reconstructions will be. However, in order to accurately assign reads, the base accuracy of the reads must also be high. This means the method relies heavily on PacBio CCS reads over PacBio CLR or Oxford Nanopore reads, where the error rate is much higher, despite the longer read lengths. As PacBio CCS technologies improve and a higher processivity polymerase is used, CCS reads will become longer and lead to more contiguous assemblies using the method presented here. However, it will be a long time, if at all, before these highly accurate reads reach the length of Oxford Nanopore reads. This means a single telomere-to-telomere haplotype-resolved assembly will be difficult to achieve. Despite this, information can still be gleaned from these imperfect assemblies and lead to a better understanding of fundamental genome regulation as well as cancer biology.

There are methods which allow phasing through an entire chromosome. An example is Strand-seq, a single cell sequencing method which allows full chromosome phasing by only sequencing a single allele (Sanders *et al.*, 2017). This method is particularly interesting as it uses a micrococcal nuclease (MNase) digestion. MNase can be used to study chromatin accessibility as it measures nucleosome occupancy (Mieczkowski *et al.*, 2016). Since Strand-seq allows phasing across the entire chromosome and the nucleosome occupancy information is inherently in the data, this raises the possibility of phasing the nucleosome occupancy for the entire chromosome and not just the regions which have informative SNPs or variants. There is also potential for the Strand-seq protocol to be adapted to allow for other sequencing techniques, for example a proximity ligation step to investigate chromatin contacts. However, despite the interesting potential, these techniques are yet to be developed.

The other limitation of the methodology presented in this thesis is the difficulty in resolving regions of increased copy number. These regions are ambiguous, with the number of sequence solutions increasing more than exponentially with each amplified segment (Greenman *et al.*, 2016b). Haplotype-resolved reads are input into genome assemblers and while hifiasm was the preferred assembler for this thesis, any assembler could be used as and when they develop and improve. Currently, assemblers are unable to resolve regions of the genome that are duplicated unless the entire duplication is encompassed in a single read. This is because without anchoring points that inform the assembler that the two duplicated segments are unique, they will be used to create one copy of the segment. Most assemblers were designed to assemble relatively normal genomes where duplications are rare events. There are tools which can be used to expand collapsed duplications and if generating cancer-specific genome assemblies become more common, assemblers specifically designed to encounter these regions may be developed. Until then, these regions may remain somewhat elusive.

Despite these limitations and opportunities for future work, this haplotype-resolved assembly method is a malleable system and the assemblies produced will improve as assembly algorithms and underlying read lengths and quality improve. In theory, the principles underlying the methodology can also be used to generate assemblies for polyploid chromosomes, if a polyploid phaser is used to initially assign haplotypes to reads.

5.2.2 Epigenetics

This thesis has begun to relate primary genome sequence and higher-order structures of chromatin at the resolution of gene expression, histone modifications, chromatin accessibility and chromatin contacts. It has highlighted some examples of dysregulation and began to elucidate the underlying mechanisms. While the impact of some rearrangements can be explained by directly fragmenting genes, for others there is clear dysregulation but the reason cannot be elucidated by simply examining the primary genome sequence. Instead, dysregulation lies in the higher-order chromatin structure. Higher-throughput methods of systematically querying regions of dysregulation will be needed to efficiently query multiple samples. An indication of dysregulation (here, differential gene expression was used) is needed to find candidate regions where there may be an altered state. However, investigation of these regions was a manual process and when scaling this type of study to multiple samples, an automated approach will need to be developed.

A major caveat of this study is the ability to only access a region of the genome with an informative variant nearby, either structural or single nucleotide. While Strand-seq offers an alternative method of phasing through entire chromosomes and methods may be developed to incorporate phasing of other epigenetic marks, longer reads could also be used to access more of the chromosome. Some regions will remain inaccessible as variant density is low, however increasing read length of ChIP-seq, ATAC-seq and Hi-C reads will allow more reads to be phased. Evidence for this can be seen when examining the improvement in phasing from 75 bp to 150 bp in the ATAC-seq to Hi-C reads. While longer read lengths are not necessary for uniquely assigning reads to a location in the genome, the likelihood of traversing an informative variant will increase with longer read lengths and even only marginal increases in phasing will lead to further insight about allele-specific regulation and how this becomes dysregulated in an abnormal context.

Arguably, the most difficult epigenetic mark to analyse in this study was chromatin conformation from Hi-C reads. Hi-C contacts are inherently noisy due to the extremely transient nature of the interactions. Often functional compartmentalisation of regions can be seen by eye but computationally calling these regions is non-trivial. The nested and hierarchical nature of TADs further complicates this. Many methods have been developed to

approach this problem, however there is huge variability in the sizes and numbers of TADs called using different callers (Zufferey *et al.*, 2018). Cancer samples are further complicated by subclonal variation which can alter TAD structures or cause novel interactions which are present in only a subset of the cells. Interactions from both the major clone and any subclones will be present in the Hi-C data. These interactions will be incorporated when calling TAD structures and therefore may result in a TAD not being called, despite it being clearly visible by eye. This made it impossible to robustly identify differential TADs in a quantitative manner in this study and interpretations were based on manual inspection. TAD callers specifically designed to deal with cancer samples or using samples that are completely clonal will be imperative in order to extend the analysis to include a systematic study of how TAD structures are altered by alterations in primary sequence.

Finally, copy number variants have cropped up multiple times in this thesis. The impact on genome assembly has already been discussed, however it is important to note that they also affect differential analyses. For ChIP-seq and ATAC-seq analysis, calling differential peaks uses untreated DNA which does not undergo an immunoprecipitation step in order to account for baseline number of reads in a specific area. This means that regions which are prone to mapping artefacts or regions of increased copy number are accounted for and not erroneously called as sites of increased protein binding or increased accessibility. This is not the case when calling differential transcripts. Here the underlying copy number is not factored into the calling of differentially expressed transcripts. An increase in transcript quantity due to increase in copy number is indistinguishable from an increased transcript quantity from a gene being placed under the control of a more active regulatory mechanism. In the examples shown in the previous chapter, the regions were copy number one on both alleles. However, this was manually identified after calling differential transcripts to determine that the cause was not due to increased copies of the region and instead was due to altered regulatory networks. Depending on the underlying question, the development of algorithms specifically designed to detect differential transcripts in cancer samples with a copy number correction will be essential.

The transient nature of epigenetic marks means studying the 3D genome structure is much more difficult than studying the primary genome sequence. A mutation, single nucleotide or structural, is either present or absent, with the added complexity of interpreting where a structural variant starts and ends. Comparatively, protein binding can be transient. Chromatin contacts are dynamic with interactions constantly forming and reforming within defined TADs to bring, for example, an enhancer in close physical proximity to a promoter to allow gene expression. All levels of regulation work together to allow cells to adapt to a particular

cell state or exogenous stress. Understanding this regulation is far from trivial but is important in order to understand the dysregulation seen in disease.

5.3 Final remarks

This thesis has capitalised on the unique model of chromothripsis occurring in cancer cells to study how alterations in primary genome sequence can alter higher-order structuring. This dysregulation traversed many epigenetic layers, undoubtedly more than just the layers studied here. The novelty of this system lies in the presence of both a wild-type and rearranged chromosome in a single cell. While this required haplotype resolution of a variety of data types and assembly of both alleles, it is unique in the sense that any inter-individual differences or differences caused by unique cell stages or exogenous stresses affecting different cells are removed.

These chromothriptic chromosomes have regions of extreme variation and using chromosome 6 in WTSI-OESO_103 as a pilot study has allowed identification of what is possible using this method. However, it has also painted a highly complex picture. Many of the rearrangements are passenger events, having no real functional impacts even if new structures are formed. For example, a neo-TAD may be formed encompassing multiple fragments found in different regions of the chromosomes. While these regions may now interact, they may not contain regulatory domains or the regulatory domains may remain unchanged and not interact with genes in the neo-TAD. Identification of which regions are causative of the alteration in higher-order structure and which are simply passenger events is difficult, particularly when the total number of alterations is very high. Future studies may be more directed at rearrangements affecting specific regions and therefore completed in a system with fewer overall structural variants. Alternatively, another interesting study based on the work presented in this thesis would be to see how these regulatory domains change over time. This may be in development where this higher-order regulation is highly dynamic or in a system where rearrangements are expected over time.

To conclude, this thesis has presented a methodology for how highly rearranged cancer genomes can be reconstructed in a haplotype-aware manner and how the generated assemblies can be used to query fundamental biological questions. Much work is needed to gain a greater understanding of how the primary genome sequence relates to higher-order structuring. Undoubtedly more will be learnt over the coming years and it is an exciting time to study cancer genomics and the 3D genome in these abnormal cells.

References

Adey, A., Burton, J. N., Kitzman, J. O., *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, 500(7461):207–211, 2013.

Akdemir, K. C., Le, V. T., Chandran, S., *et al.* Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nature Genetics*, 52(3):294–305, 2020a.

Akdemir, K. C., Le, V. T., Kim, J. M., *et al.* Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nature Genetics*, 52(11):1178–1188, 2020b.

Alaei-Mahabadi, B., Bhadury, J., Karlsson, J. W., Nilsson, J. A., and Larsson, E. Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers. *Proceedings of the National Academy of Sciences*, 113(48): 13768–13773, 2016.

Allen, B. L. and Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nature Reviews Molecular Cell Biology*, 16(3):155–166, 2015.

Amarasinghe, S. L., Su, S., Dong, X., *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1):1–16, 2020.

Baca, S. C., Prandi, D., Lawrence, M. S., *et al.* Punctuated evolution of prostate cancer genomes. *Cell*, 153(3):666–677, 2013.

Banerji, J., Rusconi, S., and Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2):299–308, 1981.

Bankevich, A., Nurk, S., Antipov, D., *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5): 455–477, 2012.

Bannister, A. J. and Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, 2011.

Bao, E., Jiang, T., and Girke, T. AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references. *Bioinformatics*, 30(12):319–328, 2014.

Benko, S., Fantes, J. A., Amiel, J., *et al.* Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nature Genetics*, 41(3):359–364, 2009.

Bird, A., Taggart, M., Frommer, M., Miller, O. J., and Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1):91–99, 1985.

Bonfield, J. K. and Whitwham, A. Gap5—editing the billion fragment sequence assembly. *Bioinformatics*, 26(14):1699–1703, 2010.

Cameron, D. L., Schröder, J., Penington, J. S., *et al.* GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Research*, 27 (12):2050–2060, 2017.

Campbell, P. J., Stephens, P. J., Pleasance, E. D., *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics*, 40(6):722–729, 2008.

Card, D. C., Schield, D. R., Reyes-Velasco, J., *et al.* Two low coverage bird genomes and a comparison of reference-guided versus de novo genome assemblies. *PLoS One*, 9(9): e106649, 2014.

Carvalho, C. M., Ramocki, M. B., Pehlivan, D., *et al.* Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nature Genetics*, 43(11):1074–1081, 2011.

Chai, J., Charboneau, A. L., Betz, B. L., and Weissman, B. E. Loss of the hSNF5 gene concomitantly inactivates p21CIP/WAF1 and p16INK4a activity associated with replicative senescence in A204 rhabdoid tumor cells. *Cancer Research*, 65(22):10192–10198, 2005.

Chen, R. and Butte, A. J. The reference human genome demonstrates high risk of type 1 diabetes and other disorders. In *Biocomputing 2011*, pages 231–242. World Scientific, 2011.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2):170–175, 2021.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13(12):1050–1054, 2016.

Chin, P. S., Assi, S. A., Ptasinska, A., *et al.* RUNX1/ETO and mutant KIT both contribute to programming the transcriptional and chromatin landscape in t (8; 21) acute myeloid leukemia. *Experimental Hematology*, 92:62–74, 2020.

Choi, Y., Chan, A. P., Kirkness, E., Telenti, A., and Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLoS Genetics*, 14(4):e1007308, 2018.

Cimini, D. Merotelic kinetochore orientation, aneuploidy, and cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1786(1):32–40, 2008.

Clapier, C. R. and Cairns, B. R. The biology of chromatin remodeling complexes. *Annual Review of Biochemistry*, 78:273–304, 2009.

Clarke, J., Wu, H.-C., Jayasinghe, L., *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 4(4):265–270, 2009.

Corces, M. R., Buenrostro, J. D., Wu, B., *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, 48 (10):1193–1203, 2016.

Cordaux, R. and Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10):691–703, 2009.

Cortés-Ciriano, I., Lee, J. J.-K., Xi, R., *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature Genetics*, 52(3):331–341, 2020.

Crasta, K., Ganem, N. J., Dagher, R., *et al.* DNA breaks and chromosome pulverization from errors in mitosis. *Nature*, 482(7383):53–58, 2012.

Cremer, T. and Cremer, M. Chromosome territories. *Cold Spring Harbor Perspectives in Biology*, 2(3):a003889, 2010.

Croft, J. A., Bridger, J. M., Boyle, S., *et al.* Differences in the localization and morphology of chromosomes in the human nucleus. *Journal of Cell Biology*, 145(6):1119–1131, 1999.

Cuddapah, S., Jothi, R., Schones, D. E., *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Research*, 19(1):24–32, 2009.

De Laat, W. and Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, 502(7472):499–506, 2013.

Dekker, J. and Heard, E. Structural and functional diversity of Topologically Associating Domains. *FEBS Letters*, 589(20):2877–2884, 2015.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, 2002.

Demeulemeester, J., Tarabichi, M., Fittall, M., *et al.* 4 Patterns of clustered mutational processes: Pan-Cancer analysis of chromothripsis, chromoplexy and kataegis. *ESMO Open*, 3:A2, 2018.

Dowen, J. M., Fan, Z. P., Hnisz, D., *et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, 159(2):374–387, 2014.

Du, H. and Liang, C. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *Nature communications*, 10(1):1–10, 2019.

Dudchenko, O., Batra, S. S., Omer, A. D., *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95, 2017.

Dunham, I., Hunt, A., Collins, J., *et al.* The DNA sequence of human chromosome 22. *Nature*, 402(6761):489–495, 1999.

Durand, N. C., Shamim, M. S., Machol, I., *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*, 3(1):95–98, 2016.

Eid, J., Fehr, A., Gray, J., *et al.* Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

Ekblom, R. and Wolf, J. B. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, 7(9):1026–1042, 2014.

Ernst, T., Chase, A. J., Score, J., *et al.* Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nature Genetics*, 42(8):722–726, 2010.

Forment, J. V., Kaidi, A., and Jackson, S. P. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nature Reviews Cancer*, 12(10):663–670, 2012.

Francioli, L. C., Menelaou, A., Pulit, S. L., *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 46(8):818–825, 2014.

Franke, M., Ibrahim, D. M., Andrey, G., *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624):265–269, 2016.

Frazer, K. A., Elnitski, L., Church, D. M., Dubchak, I., and Hardison, R. C. Cross-species sequence comparisons: a review of methods and available resources. *Genome Research*, 13(1):1–12, 2003.

Garsed, D. W., Marshall, O. J., Corbin, V. D., *et al.* The architecture and evolution of cancer neochromosomes. *Cancer Cell*, 26(5):653–667, 2014.

Gates, L. A., Foulds, C. E., and O'Malley, B. W. Histone marks in the 'driver's seat': functional roles in steering the transcription cycle. *Trends in Biochemical Sciences*, 42 (12):977–989, 2017.

Gavrielatos, M., Kyriakidis, K., Spandidos, D. A., and Michalopoulos, I. Benchmarking of next and third generation sequencing technologies and their associated algorithms for de novo genome assembly. *Molecular Medicine Reports*, 23(4):1–1, 2021.

Ghavi-Helm, Y., Jankowski, A., Meiers, S., *et al.* Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature genetics*, 51(8): 1272–1282, 2019.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., *et al.* The international HapMap project. *Nature*, 426:789–796, 2003.

Gisselsson, D., Jonson, T., Petersén, Å., *et al.* Telomere dysfunction triggers extensive DNA fragmentation and evolution of complex chromosome abnormalities in human malignant tumors. *Proceedings of the National Academy of Sciences*, 98(22):12683–12688, 2001.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., *et al.* Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*, 25 (11):1750–1756, 2015.

Goodwin, S., McPherson, J. D., and McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016.

Greenman, C., Cooke, S., Marshall, J., Stratton, M., and Campbell, P. Modeling the evolution space of breakage fusion bridge cycles with a stochastic folding process. *Journal of Mathematical Biology*, 72(1-2):47–86, 2016a.

Greenman, C., Cooke, S., Marshall, J., Stratton, M., and Campbell, P. Modeling the evolution space of breakage fusion bridge cycles with a stochastic folding process. *Journal of Mathematical Biology*, 72(1-2):47–86, 2016b.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., *et al.* The African genome variation project shapes medical genetics in Africa. *Nature*, 517(7534):327–332, 2015.

Hadi, K., Yao, X., Behr, J. M., *et al.* Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell*, 183(1):197–210, 2020.

Halkidou, K., Gaughan, L., Cook, S., *et al.* Upregulation and nuclear recruitment of HDAC1 in hormone refractory prostate cancer. *The Prostate*, 59(2):177–189, 2004.

Hanahan, D. and Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell*, 144(5): 646–674, 2011.

Hattori, M., Fujiyama, A., Taylor, T., *et al.* The DNA sequence of human chromosome 21. *Nature*, 405(6784):311–319, 2000.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009.

Henikoff, S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nature Reviews Genetics*, 9(1):15–26, 2008.

Hnisz, D., Day, D. S., and Young, R. A. Insulated neighborhoods: structural and functional units of mammalian gene control. *Cell*, 167(5):1188–1200, 2016.

Hobert, O. Gene regulation: enhancers stepping out of the shadow. *Current Biology*, 20(17): R697–R699, 2010.

Holland, A. J. and Cleveland, D. W. Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. *Nature Medicine*, 18 (11):1630–1638, 2012.

Houlahan, K. E., Shiah, Y.-J., Gusev, A., *et al.* Genome-wide germline correlates of the epigenetic landscape of prostate cancer. *Nature Medicine*, 25(10):1615–1626, 2019.

Huang, F. W., Hodis, E., Xu, M. J., *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science*, 339(6122):957–959, 2013.

Huang, Y., Jiang, L., Yi, Q., *et al.* Lagging chromosomes entrapped in micronuclei are not 'lost' by cells. *Cell Research*, 22(5):932–935, 2012.

Huch, M., Gehart, H., van Boxtel, R., *et al.* Long-term culture of genome-stable bipotent stem cells from adult human liver. *Cell*, 160(1):299–312, 2015.

Hyon, C., Chantot-Bastaraud, S., Harbuz, R., *et al.* Refining the regulatory region upstream of SOX9 associated with 46, XX testicular disorders of Sex Development (DSD). *American Journal of Medical Genetics Part A*, 167(8):1851–1858, 2015.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, 2020.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, 2004.

Iorio, F., Behan, F. M., Gonçalves, E., *et al.* Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. *BMC Genomics*, 19(1):1–16, 2018.

Jalili, V., Matteucci, M., Masseroli, M., and Morelli, M. J. Using combined evidence from replicates to evaluate ChIP-seq peaks. *Bioinformatics*, 31(17):2761–2769, 2015.

Jenuwein, T. and Allis, C. D. Translating the histone code. *Science*, 293(5532):1074–1080, 2001.

Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 15(1):1–12, 2014.

Johnson, R. and Rao, P. Mammalian cell fusion: induction of premature chromosome condensation in interphase nuclei. *Nature*, 226(5247):717–722, 1970.

Jones, D., Raine, K. M., Davies, H., *et al.* cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Current Protocols in Bioinformatics*, pages 15–10, 2016.

Jones, M. J. and Jallepalli, P. V. Chromothripsis: chromosomes in crisis. *Developmental Cell*, 23(5):908–917, 2012.

Jones, P. A. and Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, 3(6):415–428, 2002.

Kadoch, C. and Crabtree, G. R. Mammalian SWI/SNF chromatin remodeling complexes and cancer: Mechanistic insights gained from human genomics. *Science Advances*, 1(5): e1500447, 2015.

Kahles, A., Lehmann, K.-V., Toussaint, N. C., *et al.* Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell*, 34(2):211–224, 2018.

Karnuta, J. M. and Scacheri, P. C. Enhancers: bridging the gap between gene control and human disease. *Human Molecular Genetics*, 27(R2):R219–R227, 2018.

Kato, H. and Sandberg, A. A. Chromosome pulverization in human cells with micronuclei. *Journal of the National Cancer Institute*, 40(1):165–179, 1968.

Kazazian, H. H. Mobile elements: drivers of genome evolution. *Science*, 303(5664): 1626–1632, 2004.

Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, 2002.

Khan, A., Mathelier, A., and Zhang, X. Super-enhancers are transcriptionally more active and cell type-specific than stretch enhancers. *Epigenetics*, 13(9):910–922, 2018.

Kim, S., Scheffler, K., Halpern, A. L., *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, 15(8):591–594, 2018.

Kinsella, M. and Bafna, V. Combinatorics of the breakage-fusion-bridge mechanism. *Journal of Computational Biology*, 19(6):662–678, 2012.

Kircher, M., Xiong, C., Martin, B., *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature Communications*, 10(1):1–15, 2019.

Kleer, C. G., Cao, Q., Varambally, S., *et al.* EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proceedings of the National Academy of Sciences*, 100(20):11606–11611, 2003.

Kloosterman, W. P. and Cuppen, E. Chromothripsis in congenital disorders and cancer: similarities and differences. *Current Opinion in Cell Biology*, 25(3):341–348, 2013.

Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., and Peterson, H. gprofiler2–an R package for gene list functional enrichment analysis and namespace conversion toolset g: Profiler. *F1000Research*, 9, 2020.

Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R., and Papantonis, A. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics & Chromatin*, 5 (1):1–8, 2012.

Korbel, J. O. and Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell*, 152(6):1226–1236, 2013.

Koren, S., Walenz, B. P., Berlin, K., *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, 2017.

Kouzarides, T. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

Lam, E. T., Hastie, A., Lin, C., *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology*, 30(8):771–776, 2012.

Lander, E. S., Linton, L. M., Birren, B., *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

Laurell, T., VanderMeer, J. E., Wenger, A. M., *et al.* A novel 13 base pair insertion in the sonic hedgehog ZRS limb enhancer (ZRS/LMBR1) causes preaxial polydactyly with triphalangeal thumb. *Human Mutation*, 33(7):1063–1066, 2012.

Lettice, L. A., Daniels, S., Sweeney, E., *et al.* Enhancer-adoption as a mechanism of human developmental disease. *Human Mutation*, 32(12):1492–1499, 2011.

Levy, S., Sutton, G., Ng, P. C., *et al.* The diploid genome sequence of an individual human. *PLoS Biology*, 5(10):e254, 2007.

Ley, T. J., Mardis, E. R., Ding, L., *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218):66–72, 2008.

Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18): 3094–3100, 2018.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754–1760, 2009.

Li, K., Liu, Y., Cao, H., *et al.* Interrogation of enhancer function by enhancer-targeting CRISPR epigenetic editing. *Nature Communications*, 11(1):1–16, 2020a.

Li, X., Francies, H. E., Secrier, M., *et al.* Organoid cultures recapitulate esophageal adenocarcinoma heterogeneity providing a model for clonality studies and precision therapeutics. *Nature Communications*, 9(1):2983–2996, 2018.

Li, Y., Schwab, C., Ryan, S. L., *et al.* Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature*, 508(7494):98–102, 2014.

Li, Y., Roberts, N. D., Wala, J. A., *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature*, 578(7793):112–121, 2020b.

Liao, X., Li, M., Zou, Y., *et al.* Current challenges and solutions of de novo assembly. *Quantitative Biology*, 7(2):90–109, 2019.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326 (5950):289–293, 2009.

Lightbody, G., Haberland, V., Browne, F., *et al.* Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Briefings in Bioinformatics*, 20(5):1795–1811, 2019.

Lin, J. C., Jeong, S., Liang, G., *et al.* Role of nucleosomal occupancy in the epigenetic silencing of the MLH1 CpG island. *Cancer Cell*, 12(5):432–444, 2007.

Lischer, H. E. and Shimizu, K. K. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, 18(1):1–12, 2017.

Logsdon, G. A., Vollger, M. R., and Eichler, E. E. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10):597–614, 2020.

Logsdon, G. A., Vollger, M. R., Hsieh, P., *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature*, 593(7857):101–107, 2021.

Loke, J., Chin, P. S., Keane, P., *et al.* C/EBPα overrides epigenetic reprogramming by oncogenic transcription factors in acute myeloid leukemia. *Blood Advances*, 2(3):271–284, 2018.

Love, M. I., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014.

Lovén, J., Hoke, H. A., Lin, C. Y., *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2):320–334, 2013.

Lu, C., Jain, S. U., Hoelper, D., *et al.* Histone H3k36 mutations promote sarcomagenesis through altered histone methylation landscape. *Science*, 352(6287):844–849, 2016.

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, 1997.

Luo, J., Wang, J., Li, W., *et al.* EPGA2: memory-efficient de novo assembler. *Bioinformatics*, 31(24):3988–3990, 2015.

Luo, J., Wei, Y., Lyu, M., *et al.* A comprehensive review of scaffolding methods in genome assembly. *Briefings in Bioinformatics*, 22(5):33, 2021a.

Luo, X., Kang, X., and Schönhuth, A. phasebook: haplotype-aware de novo assembly of diploid genomes from long reads. *bioRxiv*, 2021b.

Lupiáñez, D. G., Kraft, K., Heinrich, V., *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

Mahmoud, M., Zywicki, M., Twardowski, T., and Karlowski, W. M. Efficiency of PacBio long read correction by 2nd generation Illumina sequencing. *Genomics*, 111(1):43–49, 2019.

Mallick, S., Li, H., Lipson, M., *et al.* The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016.

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.

McClintock, B. The production of homozygous deficient tissues with mutant characteristics by means of the aberrant mitotic behavior of ring-shaped chromosomes. *Genetics*, 23(4): 315, 1938.

Melo, U. S., Schöpflin, R., Acuna-Hidalgo, R., *et al.* Hi-C identifies complex genomic rearrangements and TAD-shuffling in developmental diseases. *The American Journal of Human Genetics*, 106(6):872–884, 2020.

Mieczkowski, J., Cook, A., Bowman, S. K., *et al.* MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nature Communications*, 7(1):1–11, 2016.

Miga, K. H., Koren, S., Rhie, A., *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585(7823):79–84, 2020.

Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell*, 128(4): 787–800, 2007.

Mitchell, T. J., Turajlic, S., Rowan, A., *et al.* Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx renal. *Cell*, 173(3):611–623, 2018.

Mito, Y., Henikoff, J. G., and Henikoff, S. Histone replacement marks the boundaries of cis-regulatory domains. *Science*, 315(5817):1408–1411, 2007.

Morey, L., Brenner, C., Fazi, F., *et al.* MBD3, a component of the NuRD complex, facilitates chromatin alteration and deposition of epigenetic marks. *Molecular and Cellular Biology*, 28(19):5912–5923, 2008.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, 2012.

Nik-Zainal, S., Davies, H., Staaf, J., *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47, 2016.

Nikoloski, G., Langemeijer, S. M., Kuiper, R. P., *et al.* Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes. *Nature Genetics*, 42(8): 665–667, 2010.

Nones, K., Waddell, N., Wayte, N., *et al.* Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nature Communications*, 5(1):1–9, 2014.

Northcott, P. A., Lee, C., Zichner, T., *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*, 511(7510):428–434, 2014.

Notta, F., Chan-Seng-Yue, M., Lemire, M., *et al.* A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature*, 538(7625):378–382, 2016.

Patterson, M., Marschall, T., Pisanti, N., *et al.* WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology*, 22(6):498–509, 2015.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196, 2010a.

Pleasance, E. D., Stephens, P. J., O'Meara, S., *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278):184–190, 2010b.

Pon, J. R. and Marra, M. A. Driver and passenger mutations in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 10:25–50, 2015.

Pop, M., Phillippy, A., Delcher, A. L., and Salzberg, S. L. Comparative genome assembly. *Briefings in Bioinformatics*, 5(3):237–248, 2004.

Poplin, R., Chang, P.-C., Alexander, D., *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, 2018.

Porubsky, D., Ebert, P., Audano, P. A., *et al.* Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology*, 39(3):302–308, 2021.

Prebet, T., Sun, Z., Figueroa, M. E., *et al.* Prolonged administration of azacitidine with or without entinostat for myelodysplastic syndrome and acute myeloid leukemia with myelodysplasia-related changes: results of the US Leukemia Intergroup trial E1905. *Journal of Clinical Oncology*, 32(12):1242, 2014.

Rausch, T., Jones, D. T., Zapatka, M., *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, 148(1-2):59–71, 2012.

Rheinbay, E., Nielsen, M. M., Abascal, F., *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, 578(7793):102–111, 2020.

Roberts, S. A., Sterling, J., Thompson, C., *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Molecular cell*, 46(4): 424–435, 2012.

Roerink, S. F., Sasaki, N., Lee-Six, H., *et al.* Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature*, 556(7702):457–462, 2018.

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381):389–393, 2012.

Ruan, J. and Li, H. Fast and accurate long-read assembly with wtdbg2. *Nature methods*, 17 (2):155–158, 2020.

Sahlin, K. and Medvedev, P. Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nature Communications*, 12(1):1–13, 2021.

Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C., and Lansdorp, P. M. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nature Protocols*, 12(6):1151–1176, 2017.

Saxonov, S., Berg, P., and Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, 103(5):1412–1417, 2006.

Schaub, F. X., Dhankani, V., Berger, A. C., *et al.* Pan-cancer alterations of the MYC oncogene and its proximal network across the cancer genome atlas. *Cell Systems*, 6(3):282–300, 2018.

Schmitz, R., Ceribelli, M., Pittaluga, S., Wright, G., and Staudt, L. M. Oncogenic mechanisms in Burkitt lymphoma. *Cold Spring Harbor Perspectives in Medicine*, 4(2):a014282, 2014.

Schneider, V. A., Graves-Lindsay, T., Howe, K., *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, 2017.

Schwartz, D. C., Li, X., Hernandez, L. I., *et al.* Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science*, 262(5130):110–114, 1993.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15:461–468, 2018.

Serizay, J. and Ahringer, J. Genome organization at different scales: nature, formation and function. *Current Opinion in Cell Biology*, 52:145–153, 2018.

Shao, X., Lv, N., Liao, J., *et al.* Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Medical genetics*, 20(1):1–14, 2019.

Sharma, S., Kelly, T. K., and Jones, P. A. Epigenetics in cancer. *Carcinogenesis*, 31(1): 27–36, 2010.

Shen, M. M. Chromoplexy: a new category of complex rearrangements in the cancer genome. *Cancer Cell*, 23(5):567–569, 2013.

Sherman, R. M., Forman, J., Antonescu, V., *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, 51(1):30–35, 2019.

Shi, J., Ma, X., Zhang, J., *et al.* Chromosome conformation capture resolved near complete genome assembly of broomcorn millet. *Nature Communications*, 10(1):1–9, 2019.

Shinohara, T., Tomizuka, K., Miyabara, S., *et al.* Mice containing a human chromosome 21 model behavioral impairment and cardiac anomalies of Down's syndrome. *Human Molecular Genetics*, 10(11):1163–1175, 2001.

Shlyueva, D., Stampfel, G., and Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286, 2014.

Shoshani, O., Brunner, S. F., Yaeger, R., *et al.* Chromothripsis drives the evolution of gene amplification in cancer. *Nature*, 591(7848):137–141, 2021.

Sinsheimer, R. L. The santa cruz workshop — May 1985. *Genomics*, 5(4):954–956, 1989.

Sondka, Z., Bamford, S., Cole, C. G., *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, 2018.

Song, J., Noh, J. H., Lee, J. H., *et al.* Increased expression of histone deacetylase 2 is found in human gastric cancer. *APMIS*, 113(4):264–268, 2005.

Splinter, E., Heath, H., Kooren, J., *et al.* CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus. *Genes & Development*, 20(17): 2349–2354, 2006.

Stephens, P. J., Greenman, C. D., Fu, B., *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40, 2011.

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

Struhl, K. and Segal, E. Determinants of nucleosome positioning. *Nature Structural & Molecular Biology*, 20(3):267–273, 2013.

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

Thåström, A., Lowary, P., Widlund, H., *et al.* Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *Journal of Molecular Biology*, 288(2):213–229, 1999.

Treangen, T. J. and Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, 2012.

Tremethick, D. J. Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell*, 128(4): 651–654, 2007.

Van Haaften, G., Dalgliesh, G. L., Davies, H., *et al.* Somatic mutations of the histone H3k27 demethylase gene UTX in human cancer. *Nature Genetics*, 41(5):521–523, 2009.

Varambally, S., Dhanasekaran, S. M., Zhou, M., *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, 419(6907):624–629, 2002.

Venter, J. C., Adams, M. D., Myers, E. W., *et al.* The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

Vezzi, F., Cattonaro, F., and Policriti, A. e-RGA: enhanced reference guided assembly of complex genomes. *EMBnet. Journal*, 17(1):46–54, 2011.

Vollger, M. R., Dishuck, P. C., Sorensen, M., *et al.* Long-read sequence and assembly of segmental duplications. *Nature Methods*, 16(1):88–94, 2019.

Von Hoff, D., Forseth, B., Clare, C., *et al.* Double minutes arise from circular extrachromosomal DNA intermediates which integrate into chromosomal sites in human HL-60 leukemia cells. *The Journal of Clinical Investigation*, 85(6):1887–1895, 1990.

Voronina, N., Wong, J. K., Hübschmann, D., *et al.* The landscape of chromothripsis across adult cancer types. *Nature Communications*, 11(1):1–13, 2020.

Waddington, C. H. Canalization of development and the inheritance of acquired characters. *Nature*, 150(3811):563–565, 1942a.

Waddington, C. H. The epigenotype. *Endeavour*, 1:18–20, 1942b.

Wang, J., Wang, W., Li, R., *et al.* The diploid genome sequence of an Asian individual. *Nature*, 456(7218):60–65, 2008.

Waterston, R. H. and Pachter, L. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

Watson, C. M., Crinnion, L. A., Harrison, S. M., *et al.* A Chromosome 7 pericentric inversion defined at single-nucleotide resolution using diagnostic whole genome sequencing in a patient with hand-foot-genital syndrome. *PloS One*, 11(6):e0157075, 2016.

Weaver, J. M., Ross-Innes, C. S., Shannon, N., *et al.* Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nature Genetics*, 46(8):837–843, 2014.

Weikert, S., Christoph, F., Köllermann, J., *et al.* Expression levels of the EZH2 polycomb transcriptional repressor correlate with aggressiveness and invasive potential of bladder carcinomas. *International Journal of Molecular Medicine*, 16(2):349–353, 2005.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genetics*, 46(11):1160–1165, 2014.

Wenger, A., Hickey, L., Chin, J., and Korlach, J. Structural variant detection with low-coverage PacBio sequencing. *Nature*, 517(7536):608–611, 2017.

Wenger, A. M., Peluso, P., Rowell, W. J., *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, 2019.

Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., *et al.* Species-specific transcription in mice carrying human chromosome 21. *Science*, 322(5900):434–438, 2008.

Xie, B.-Y. and Wu, A.-W. Organoid culture of isolated cells from patient-derived tissues with colorectal cancer. *Chinese Medical Journal*, 129(20):2469, 2016.

Xu, M., Guo, L., Du, X., *et al.* Accurate haplotype-resolved assembly reveals the origin of structural variants for human trios. *Bioinformatics*, 37(15):2095–2102, 2021.

Yahav, T. and Privman, E. A comparative analysis of methods for de novo assembly of hymenopteran genomes using either haploid or diploid samples. *Scientific Reports*, 9(1): 1–10, 2019.

Yang, X.-J. The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases. *Nucleic Acids Research*, 32(3):959–976, 2004.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, 2009.

Yen, E. C., McCarthy, S. A., Galarza, J. A., *et al.* A haplotype-resolved, de novo genome assembly for the wood tiger moth (Arctia plantaginis) through trio binning. *GigaScience*, 9(8):88, 2020.

Yuan, Y., Chung, C. Y.-L., and Chan, T.-F. Advances in optical mapping for genomic research. *Computational and Structural Biotechnology Journal*, 2020.

Zack, T. I., Schumacher, S. E., Carter, S. L., *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45(10):1134–1140, 2013.

Zhang, C.-Z., Leibowitz, M. L., and Pellman, D. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes & Development*, 27(23): 2513–2530, 2013.

Zhang, C.-Z., Spektor, A., Cornils, H., *et al.* Chromothripsis from DNA damage in micronuclei. *Nature*, 522(7555):179–184, 2015a.

Zhang, F., Carvalho, C. M., and Lupski, J. R. Complex human chromosomal and genomic rearrangements. *Trends in Genetics*, 25(7):298–307, 2009.

Zhang, J., Kalkum, M., Yamamura, S., Chait, B. T., and Roeder, R. G. E protein silencing by the leukemogenic AML1-ETO fusion protein. *Science*, 305(5688):1286–1289, 2004.

Zhang, T., Cooper, S., and Brockdorff, N. The interplay of histone modifications–writers that read. *EMBO Reports*, 16(11):1467–1481, 2015b.

Zhang, X., Choi, P. S., Francis, J. M., *et al.* Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nature Genetics*, 48(2):176–182, 2016.

Zhang, Y., Liu, T., Meyer, C. A., *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):1–9, 2008.

Zhao, Z. and Shilatifard, A. Epigenetic modifications of histones in cancer. *Genome Biology*, 20(1):1–16, 2019.

Zufferey, M., Tavernari, D., Oricchio, E., and Ciriello, G. Comparison of computational methods for the identification of topologically associating domains. *Genome Biology*, 19 (1):1–18, 2018.

Appendix A

Amplified Genes in chromothripsis

Two chromothriptic chromosomes studied in this thesis have large regions of amplification over multiple copy number states. These regions contain genes which may have altered gene expression. The genes in regions of amplification on these chromothriptic chromosomes are listed below.

WTSI-OESO_143 chromosome 18

1. *TYMSOS*
2. *TYMS*
3. *ENOSF1*
4. *EMILIN2*
5. *LPIN2*
6. *CHORDC1P4*
7. *SNRPCP4*
8. *MYL12B*
9. *LINC01895*
10. *RPL31P59*
11. *IGLJCOR18*
12. *RPL21P127*
13. *TGIF1*
14. *BOD1P1*
15. *GAPLINC*
16. *DLGAP1*
17. *RN7SL39P*
18. *DLGAP1-AS1*
19. *DLGAP1-AS2*
20. *RNU6-120P*
21. *EXOGP1*
22. *GREB1L-DT*
23. *GREB1L*
24. *ESCO1*
25. *SNRPD1*
26. *ABHD3*
27. *MIR320C1*
28. *MIB1*
29. *RN7SL233P*
30. *MIR133A1HG*
31. *MIR133A1*
32. *MIR1-2*
33. *RNU6-1038P*
34. *RNA5SP451*
35. *LINC01900*
36. *RNU6ATAC20P*

37. *GATA6-AS1*
38. *GATA6*
39. *RNU6-702P*
40. *CTAGE1*
41. *ATP7BP1*
42. *RPS4XP18*
43. *RNU6-1032P*
44. *RBBP8*
45. *UBE2CP2*
46. *MIR4741*
47. *RN7SL745P*
48. *CABLES1*
49. *TMEM241*
50. *RIOK3*
51. *RMC1*
52. *NPC1*
53. *ANKRD29*
54. *RPS10P27*
55. *LAMA3*
56. *RPL23AP77*
57. *TTC39C*
58. *TTC39C-AS1*
59. *RNU5A-6P*
60. *CABYR*
61. *OSBPL1A*
62. *RNA5SP452*
63. *RNU6-435P*
64. *MIR320C2*
65. *IMPACT*
66. *HRH4*
67. *EIF4A3P1*
68. *LINC01915*
69. *RAC1P1*
70. *PPIAP57*
71. *LINC01894*
72. *WBP2P1*
73. *ZNF521*
74. *KCTD1*
75. *AQP4-AS1*
76. *AQP4*
77. *CDH2*

WTSI-OESO_148 chromosome 1

1. *SEC22B*
2. *NBPF8*
3. *PFN1P2*
4. *PDE4DIPP2*
5. *NOTCH2NLR*
6. *NBPF26*
7. *RNVU1-19*
8. *PPIAL4A*
9. *LINC00623*
10. *RNVU1-4*
11. *PDE4DIPP4*
12. *FCGR1B*
13. *H2BP1*
14. *H3P4*
15. *RPL22P6*
16. *FAM72B*
17. *SRGAP2C*
18. *SRGAP2-AS1*
19. *LINC02798*
20. *MTIF2P1*
21. *EMBP1*
22. *LINC01691*
23. *NBPF17P*
24. *PFN1P12*

25. *PPIAL4F*
26. *LINC01632*
27. *RNA5SP59*
28. *SRGAP2B*
29. *FAM72D*
30. *LINC01145*
31. *RNU1-153P*
32. *PPIAL4D*
33. *RNVU1-14*
34. *NBPF20*
35. *PFN1P3*
36. *RNU1-154P*
37. *RNVU1-31*
38. *PDE4DIPP5*
39. *NBPF25P*
40. *GPR89A*
41. *PDZK1*
42. *CD160*
43. *RNF115*
44. *POLR3C*
45. *NUDT17*
46. *PIAS3*
47. *MIR6736*
48. *ANKRD35*
49. *ITGA10*
50. *PEX11B*
51. *GNRHR2*
52. *RBM8A*
53. *LIX1L-AS1*
54. *LIX1L*
55. *ANKRD34A*
56. *POLR3GL*
57. *TXNIP*
58. *HJV*
59. *RNVU1-6*
60. *LINC01719*
61. *NBPF10*
62. *NOTCH2NLA*
63. *RNU6-1071P*
64. *NUDT4P2*
65. *SEC22B4P*
66. *PPIAL4H*
67. *RNVU1-29*
68. *RNVU1-25*
69. *HYDIN2*
70. *RNA5SP536*
71. *NBPF12*
72. *PFN1P8*
73. *RNU1-151P*
74. *SSBL4P*
75. *RNVU1-8*
76. *NBPF13P*
77. *PRKAB2*
78. *PDIA3P1*
79. *FMO5*
80. *CCT8P1*
81. *RPL7AP15*
82. *CHD1L*
83. *LINC00624*
84. *OR13Z1P*
85. *OR13Z2P*
86. *OR13Z3P*
87. *BCL9*
88. *ACP6*
89. *RN7SL261P*
90. *GJA5*
91. *GJA8*
92. *GPR89B*
93. *PDZK1P1*
94. *LINC02804*
95. *RNU1-129P*
96. *RNVU1-7*
97. *PDE4DIPP1*

98. *NBPF11*
99. *PFN1P4*
100. *ABHD17AP1*
101. *LINC02805*
102. *RNA5SP57*
103. *RNVU1-22*
104. *LINC01731*
105. *LINC01138*
106. *LINC02806*
107. *MIR5087*
108. *RNVU1-21*
109. *RNVU1-1*
110. *RNU1-155P*
111. *MIR6077*
112. *RNVU1-2*
113. *PDE4DIPP6*
114. *RNVU1-3*
115. *PPIAL4G*
116. *RNVU1-27*
117. *NBPF14*
118. *NOTCH2NLB*
119. *RNU6-1171P*
120. *NUDT4B*
121. *SEC22B3P*
122. *PDE4DIP*
123. *RN7SKP88*
124. *RNU2-38P*
125. *NBPF9*
126. *RNVU1-24*
127. *SEC22B2P*
128. *NOTCH2NLC*
129. *NBPF19*
130. *PPIAL4C*
131. *LINC00869*
132. *RNVU1-30*
133. *FAM91A2P*
134. *PDE4DIPP7*
135. *RNU1-68P*
136. *FCGR1A*
137. *H2BC18*
138. *H3C13*
139. *H4C14*
140. *H3C14*
141. *H2AC18*
142. *H2BC19P*
143. *H2BC20P*
144. *H2AC19*
145. *H3C15*
146. *H4C15*
147. *H2BC21*
148. *H2AC20*
149. *H2AC21*
150. *BOLA1*
151. *SV2A*
152. *SF3B4*
153. *MTMR11*
154. *OTUD7B*
155. *VPS45*
156. *PLEKHO1*
157. *RN7SL480P*
158. *ANP32E*
159. *RNU2-17P*
160. *CA14*
161. *APH1A*
162. *C1orf54*
163. *CIART*
164. *MRPS21*
165. *PRPF3*
166. *RPRD2*
167. *TARS2*
168. *MIR6878*
169. *ECM1*
170. *FALEC*

171. *ADAMTSL4-AS2*
172. *ADAMTSL4*
173. *MIR4257*
174. *ADAMTSL4-AS1*
175. *MCL1*
176. *RN7SL473P*
177. *RN7SL600P*
178. *ENSA*
179. *GOLPH3L*
180. *HORMAD1*
181. *RNU6-1042P*
182. *CTSS*
183. *CTSK*
184. *UBE2D3P3*
185. *ARNT*
186. *RNU6-1309P*
187. *RPS27AP6*
188. *CTXND2*
189. *CYCSP51*
190. *SETDB1*
191. *CERS2*
192. *ANXA9*
193. *MINDY1*
194. *PRUNE1*
195. *RNU6-884P*
196. *BNIPL*
197. *C1orf56*
198. *CDC42SE1*
199. *MLLT11*
200. *GABPB2*
201. *RPS29P29*
202. *SEMA6C*
203. *TNFAIP8L2*
204. *SCNM1*
205. *LYSMD1*
206. *TMOD4*
207. *VPS72*
208. *PIP5K1A*
209. *PSMD4*
210. *ZNF687-AS1*
211. *ZNF687*
212. *PI4KB*
213. *RN7SL444P*
214. *RFX5*
215. *RFX5-AS1*
216. *SELENBP1*
217. *PSMB4*
218. *POGZ*
219. *RNY4P25*
220. *CGN*
221. *TUFT1*
222. *MIR554*
223. *SNX27*
224. *RNU6-1062P*
225. *CELF3*
226. *RIIAD1*
227. *RNU6-662P*
228. *MRPL9*
229. *OAZ3*
230. *TDRKH*
231. *TDRKH-AS1*
232. *LINGO4*
233. *RORC*
234. *C2CD4D*
235. *C2CD4D-AS1*
236. *THEM5*
237. *THEM4*
238. *KRT8P28*
239. *S100A10*
240. *NBPF18P*
241. *S100A11*
242. *SPTLC1P4*
243. *TCHHL1*

244. *TCHH*
245. *PUDPP2*
246. *RPTN*
247. *FLG-AS1*
248. *HRNR*
249. *FLG*
250. *FLG2*
251. *HMGN3P1*
252. *CRNN*
253. *LCE5A*
254. *CRCT1*
255. *LCE3E*
256. *LCE3D*
257. *LCE3C*
258. *LCE3B*
259. *LCE3A*
260. *LCEP4*
261. *LCEP3*
262. *LCE2D*
263. *LCE2C*
264. *LCE2B*
265. *LCE2A*
266. *LCE4A*
267. *C1orf68*
268. *LCEP2*
269. *LCEP1*
270. *KPRP*
271. *LCE1F*
272. *LCE1E*
273. *LCE1D*
274. *LCE1C*
275. *LCE1B*
276. *LCE1A*
277. *LCE6A*
278. *SMCP*
279. *IVL*
280. *LINC01527*
281. *SPRR5*
282. *SPRR4*
283. *SPRR1A*
284. *SPRR3*
285. *SPRR1B*
286. *SPRR2D*
287. *SPRR2A*
288. *SPRR2B*
289. *SPRR2E*
290. *SPRR2F*
291. *SPRR2C*
292. *SPRR2G*
293. *LELP1*
294. *PRR9*
295. *LORICRIN*
296. *PGLYRP3*
297. *PGLYRP4*
298. *RNU6-160P*
299. *S100A9*
300. *S100A12*
301. *LAPTM4BP1*
302. *S100A8*
303. *S100A15A*
304. *S100A7A*
305. *S100A7P1*
306. *S100A7L2*
307. *S100A7*
308. *RN7SL44P*
309. *S100A6*
310. *S100A5*
311. *S100A4*
312. *S100A3*
313. *S100A2*
314. *S100A16*
315. *S100A14*
316. *S100A13*

317. *S100A1*
318. *CHTOP*
319. *SNAPIN*
320. *ILF2*
321. *NPR1*
322. *MIR8083*
323. *RN7SL372P*
324. *GEMIN2P1*
325. *INTS3*
326. *SLC27A3*
327. *GATAD2B*
328. *DENND4B*
329. *CRTC2*
330. *SLC39A1*
331. *MIR6737*
332. *CREB3L4*
333. *JTB*
334. *RAB13*
335. *RPS27*
336. *NUP210L*
337. *RNU6-179P*
338. *RPS7P2*
339. *MIR5698*
340. *TPM3*
341. *RN7SL431P*
342. *MIR190B*
343. *C1orf189*
344. *C1orf43*
345. *UBAP2L*
346. *SNORA58B*
347. *HAX1*
348. *RNU6-239P*
349. *RNU6-121P*
350. *AQP10*
351. *ATP8B2*
352. *RNU7-57P*
353. *RPSAP17*
354. *IL6R-AS1*
355. *IL6R*
356. *PSMD8P1*
357. *SHE*
358. *TDRD10*
359. *UBE2Q1*
360. *UBE2Q1-AS1*
361. *CHRNB2*
362. *ADAR*
363. *KCNN3*
364. *PMVK*
365. *PBXIP1*
366. *PYGO2*
367. *SHC1*
368. *CKS1B*
369. *MIR4258*
370. *FLAD1*
371. *LENEP*
372. *ZBTB7B*
373. *DCST2*
374. *DCST1*
375. *DCST1-AS1*
376. *ADAM15*
377. *EFNA4*
378. *EFNA3*
379. *CDC42BPA*
380. *LINC01641*
381. *NUCKS1P1*
382. *BTF3P9*
383. *TUBB8P10*
384. *TUBB8P9*
385. *RNA5SP77*
386. *ZNF678*
387. *FAM133FP*
388. *ZNF847P*
389. *SNAP47*

390. *JMJD4*
391. *SNAP47-AS1*
392. *PRSS38*
393. *WNT9A*
394. *MIR5008*
395. *CICP26*
396. *SEPTIN14P17*
397. *WNT3A*
398. *LINC02809*
399. *ARF1*
400. *MIR3620*
401. *C1orf35*
402. *MRPL55*
403. *CIAO2AP2*
404. *GUK1*
405. *GJC2*
406. *IBA57-DT*
407. *IBA57*
408. *OBSCN-AS1*
409. *OBSCN*
410. *TRIM11*
411. *MIR6742*
412. *TRIM17*
413. *H3-4*
414. *H2AW*
415. *RPL23AP15*
416. *H2BU1*
417. *MIR4666A*
418. *H2BU2P*
419. *RNF187*
420. *BTNL10*
421. *RNA5SP19*
422. *RNA5SP162*
423. *RNA5S1*
424. *RNA5S2*
425. *RNA5S3*
426. *RNA5S4*
427. *RNA5S5*
428. *RNA5S6*
429. *RNA5S7*
430. *RNA5S8*
431. *RNA5S9*
432. *RNA5S10*
433. *RNA5S11*
434. *RNA5S12*
435. *RNA5S13*
436. *RNA5S14*
437. *RNA5S15*
438. *RNA5S16*
439. *RNA5S17*
440. *RNA5SP18*
441. *DUSP5P1*
442. *FTH1P2*
443. *RHOU*
444. *LINC02815*
445. *ISCA1P2*
446. *LINC02814*
447. *TMEM78*
448. *RAB4A*
449. *CCSAP*
450. *RNU6-180P*
451. *RN7SKP276*
452. *ACTA1*
453. *NUP133*
454. *ABCB10*
455. *RNU4-21P*
456. *RNA5SP78*
457. *HMGN2P19*
458. *TAF5L*
459. *URB2*
460. *HMGB1P26*
461. *LINC01682*
462. *LINC01736*

463. *GALNT2*
464. *PGBD5*
465. *LINC01737*
466. *COG2*
467. *AGT*
468. *CAPN9*
469. *RNA5SP79*
470. *C1orf198*
471. *RN7SL837P*
472. *TTC13*
473. *ARV1*
474. *FAM89A*
475. *MIR1182*
476. *TRIM67*
477. *TRIM67-AS1*
478. *C1orf131*
479. *GNPAT*
480. *RNA5SP80*
481. *EXOC8*
482. *SPRTN*
483. *EGLN1*
484. *SNRPD2P2*
485. *TSNAX*
486. *TSNAX-DISC1*
487. *LINC00582*
488. *DISC1*
489. *RNU5A-5P*
490. *DISC1-IT1*