Chapter 1

Introduction to the cancer genome

The journey of an individual human genome begins with its formation in the fertilised egg—a chance meeting between maternal and paternal chromosomes in a totally unique combination, never to be repeated. After the normal germline genome is first established in the zygote, it faces the immediate prospect of copying itself into two daughter cells as faithfully as possible. Indeed, in each mitotic cell division through embryogenesis, infancy, and adulthood, a volley of biochemical activity operates to replicate and disseminate the six gigabases of inherited genome many millions of times over. Inevitably, occasional errors in DNA repair, replication, and segregation accrue with each cell division, and somatic genomes gradually diverge from their common ancestor in the zygote. A subset of these somatic mutations confer a selective advantage to the cell lineage, sometimes culminating in pathological unchecked cell growth broadly classified as cancer. With advances in whole genome DNA sequencing technology, somatic mutation in cancer samples can now be identified at base-pair resolution on any scale from single base substitution to rearrangement of kilobases, megabases, and whole chromosomes. In this thesis I analyse somatic rearrangement observed in more than 2500 cancer genomes from common cancer types all over the human body. The diverse structural patterns which emerge are testament to the complex bio-molecular challenges a genome may encounter in the course of its somatic evolution. By charting the landscape of possible genome configurations in the soma, we begin to understand the repertoire of genetic manoeuvres available to a cancer, and can better appreciate the underlying reasons for cancer's heterogeneous clinical presentation.

1.1 The somatic genome in mitosis and cancer

Throughout the mitotic cell cycle, the information content and structural integrity of the nuclear genome must be preserved and carefully promulgated to maintain regulated programs of cell behaviour and function. To this end, breaks or lesions in the DNA are repaired where possible, or may trigger cell death. The DNA content is replicated in S phase to produce sister chromatid pairs, which then condense and separate into opposite daughter cells during M phase. Errors in the dynamic orchestration of genome state generate mutations which transmit through the descendent cell lineage. Such genome alterations include single nucleotide variants (SNV), small insertions or deletions (indels), and a diverse range of larger structural variation (SV). Although most mutations have negligible fitness effects, some may confer a selective advantage driving clonal expansion into oncogenesis. (Stratton et al., 2009; Martincorena and Campbell, 2015; Tubbs and Nussenzweig, 2017)

1.1.1 DNA damage response

DNA lesions arise from endogenous and exogenous sources, including UV radiation, ionizing radiation, reactive oxygen species, chemical mutagens, and the inherent instability of biochemical molecules in a reactive environment. Different lesion types signal specialised DNA damage response pathways. For example, abasic sites and spontaneous deamination of 5-methylcytosine are repaired by base excision repair; pyrimidine dimers and bulky adducts by nucleotide excision repair; and incorrect DNA base-pairing by mismatch repair. Double-stranded DNA breaks may signal a variety of repair pathways, including non-homologous end-joining and homologous recombination, discussed further in Section 1.4. If the DNA injury is beyond repair, then the p53 pathway may trigger senescence or apoptosis to remove the cell from the population. When a DNA lesion is replicated without repair, or repaired incorrectly, mutations fix into the cell lineage. Some cancers have loss-of-function mutations in the genes controlling DNA repair, and develop a hypermutator phenotype as a result of compromised repair capacity. (Jackson and Bartek, 2009; Helleday et al., 2014; Tubbs and Nussenzweig, 2017)

1.1.2 DNA replication

DNA replication begins at many thousands of licensed origins[a], which 'fire' at different time points during S phase to recruit the replisome complex at two bi-directional replication forks. The replisome includes: helicase for separating parental duplex DNA into single stranded templates; topoisomerase for cutting the DNA backbone to release super-coil tension ahead of the fork and precatenane[b] structures behind the fork; polymerases for synthesising new DNA strands; and the DNA clamp PCNA for tethering the polymerase to the template strand. Different DNA polymerases have specialised roles in priming DNA synthesis and elongating nascent DNA along the leading and lagging strands[c]. The polymerases completing the bulk of replication have an inbuilt proof-reading domain and an estimated error rate of 10^{-7} mismatches per base. In contrast, the specialised translesion polymerases for replicating past DNA damage have lower fidelity, and are prone to incorporating small indels and SNVs. (Loeb and Monnat, 2008; Branzei and Foiani, 2010; Gaillard et al., 2015)

In addition to the small mutations caused by polymerase error, DNA replication can also generate larger structural variation through aberrant origin licensing, topoisomerase errors, and replication fork stalling and collapse. For example, inefficient origin licensing leads to incomplete replication and breaks in late-replicating regions, whereas unscheduled origin firing can lead to re-replication and fork collisions. Fork progression is also impeded by nucleotide pool depletion or physical obstacles such as DNA lesions or breaks, non-B DNA structures, or transcription bubbles. S phase checkpoint pathways respond to stalled forks and try to complete replication via translesion polymerases, template switching to the sister chromatid, or licensing of dormant origins. Failure to do so gives rise to double-stranded DNA breaks and subsequent error-prone repair. (Branzei and Foiani, 2010; Gaillard et al., 2015; Cortez, 2015)

[a] A 'licensed' replication origin is bound by helicases and the origin recognition complex during G1 phase, in preparation for active replication 'firing' during S phase.

[b] A precatenane is formed by sister DNA duplexes intertwining after synthesis.

[c] As DNA polymerases add new nucleotides to the free 3' hydroxyl group on the sugar-phosphate backbone, synthesis must proceed in a 5' to 3' direction. At the replication fork, leading strand synthesis is able to proceed continuously as it travels in the same direction as the opening fork. On the lagging strand, DNA is synthesised in discontinuous fragments building away from the replication fork and later joined through ligation.

1.1.3 Chromosome segregation

During interphase, the nuclear DNA spreads out to occupy large chromosomal territories with looping domain structures to regulate gene expression (Gibcus and Dekker, 2013). In preparation for mitotic cell division, the nuclear membrane breaks down as the chromosomes condense into their compact form, with sister chromatids initially still linked together via cohesin complexes (prophase). To achieve equal chromosome segregation, each chromatid in a sister pair must attach to kinetochore microtubules emanating from opposite spindle poles (metaphase). As the mitotic checkpoint proteins decay to signal successful spindle attachments, the cohesin disbands and sister chromatids are pulled to opposite poles (anaphase). In the final stages of telophase and cytokinesis, nuclear membranes reform around the two separated DNA masses, and the cellular membrane cleaves the cytoplasm to produce two daughter cells with equal chromosome content. (Hirano, 2015; Funk et al., 2016)

Errors in mitotic division can change the overall ploidy, and even be a root cause of DNA breaks and rearrangement. If cytokinesis fails to divide the replicated DNA into separate daughter cells, then the doubled genome content can persist in tetraploid state. If successful cytokinesis follows uneven chromosome segregation, then the abnormal chromosome count can persist in aneuploid state. Causes of chromosome missegregation include: mitotic checkpoint failure permitting premature entry into anaphase; cohesin defects causing sister chromatids to prematurely decouple or remain linked during anaphase; and aberrant kinetochore attachments (syntelic or merotelic) or centromere content (dicentric or acentric). These errors may pull both sister chromatids into the same daughter cell, or may result in DNA being caught between poles (either an entire lagging chromosome, or a smaller DNA section caught in a 'bridge'). Lagging or bridge DNA can be a substrate for large-scale rearrangement, as discussed further in Section 1.4. (Orr et al., 2015; Funk et al., 2016)

1.1.4 Genome and chromosome instability

In some cancers, the normal programs of DNA repair, replication, and mitotic segregation become so disordered that the cells develop persistent genomic and/or chromosomal instability. Genomic instability (GIN) refers to the continual generation of structural rearrangements *within* chromosomes, whereas chromosomal instability (CIN) refers specifically to unstable aneuploidy and a consistently high rate of chromosome missegregation.

Both instability phenotypes are associated with ongoing replication stress as a result of excessive DNA damage, excessive oncogenic transcriptional programs, or loss-of-function mutations in relevant genes (Burrell et al., 2013; Macheret and Halazonetis, 2015). Under these stress conditions, slow, stalled, or collapsed replication forks give rise to SVs and missegregating acentric or dicentric chromosomes. CIN is also possible in a competent replication background with compromised mitotic function. Although high rates of CIN are associated with cell death and tumour suppression, low rates of CIN are thought to be weakly tumour promoting, and provide a gradually diversifying genetic pool to facilitate adaptation. In the Mitelman cytogenetic database, 44% of solid tumours and 14% of blood cancers show evidence of CIN, while a further 42% (solid) and 58% (blood) have stable aneuploidies. (Zasadil et al., 2013; Funk et al., 2016)

1.1.5 Somatic mutations give rise to cancer

The hallmark properties of cancer include: sustained proliferative signalling and replicative immortality; evasion of growth suppression and cell death; and acquisition of invasive and metastatic abilities. These abnormal cellular properties are acquired via driver genome alterations, and thus somatic genome instability and mutation are considered an 'enabling' cancer hallmark (Hanahan and Weinberg, 2011).

Oncogenesis requires a small accumulation of driver events, with between two and ten currently identifiable in most cancer genomes (Vogelstein et al., 2013; Tomasetti et al., 2015; Martincorena et al., 2017; Sabarinathan et al., 2017). In general, oncogenes promoting cell growth are up-regulated by gain-of-function mutations, and tumour suppressor genes providing normal control and repair functions are down-regulated by loss-of-function mutations. Although most driver mutations are acquired in the soma, some may be inherited in the germline and increase the lifetime cancer risk (for example, *BRCA1* and *BRCA2* polymorphisms). Active mutagenic processes also generate a vast number of 'passenger' somatic alterations with no fitness benefit, thus confounding the search for genuine drivers in cancer genome sequencing studies (Pon and Marra, 2015).

1.2 Cancer genome sequencing projects

In a prescient opinion piece, Dulbecco (1986) predicted that an undertaking to sequence the human genome would yield invaluable insight into cancer biology. Despite being a stretch of blue-sky thinking at the time, his initial vision—to interrogate any gene of interest with probes designed off the reference—has long since been surpassed. The advent of affordable high-throughput DNA sequencing technologies ushered in a new field of cancer genomics research, with the first samples sequenced in their entirety by Ley et al. (2008) and Pleasance et al. (2010). Following this success, large collaborations within the International Cancer Genome Consortium (ICGC), The Cancer Genome Atlas (TCGA), and other local projects, set out to systematically catalogue genetic mutations in most common cancer types (International Cancer Genome Consortium et al., 2010; Cancer Genome Atlas Research Network et al., 2013; Wheeler and Wang, 2013). To date, research publications have summarised the genome landscape in dozens of patient cohorts, from the earliest reports characterising hundreds of *exomes* in ovarian and colorectal cancer (Cancer Genome Atlas Research Network, 2011; Cancer Genome Atlas Research Network, 2012) to more recent work analysing hundreds of *whole genomes* in breast cancer (Nik-Zainal et al., 2016) and medulloblastoma (Northcott et al., 2017), to cite just a few examples.

1.2.1 Study design

The classical study design for a cancer genome project is to sequence the bulk DNA of matched cancer–normal samples from a cohort of donors with the same or similar disease pathology (Mwenifumbo and Marra, 2013). Matching each cancer sample with normal DNA from the same individual[d] is critical for distinguishing somatic mutations specific to the cancer lineage from germline polymorphisms present in all tissues of the body.

To date, the vast majority of cancer genome projects have used the Illumina DNA sequencing platform. This technology sequences the last 100–150 bases of billions of DNA fragments by detecting the stepwise addition of fluorescently-labelled, reversibly-terminating nucleotides (Reuter et al., 2015). Sophisticated bioinformatics pipelines map these short reads (usually paired ends from a

[d]Normal DNA is usually taken from blood, or nearby non-cancerous tissue surgically extracted at the same time as the tumour. For blood cancers, the normal sample must be taken from an isolate of non-cancerous cell type/s (or another tissue if available).

fragment < 1 kb long) to their most likely origin in the reference genome, and identify variants which differ from the reference sequence.

So far, TCGA studies have primarily focussed on whole exome capture sequencing (WES), limiting high resolution findings to protein-coding regions covering less than 2% of the total genome. Studies by the ICGC and other groups are now turning to the more expensive whole genome sequencing (WGS) methods, which allow variation to be detected in non-coding regions and in the form of structural rearrangement. In addition to DNA sequencing, most cancer genome projects include complementary assays such as SNP arrays to detect copy number variation (CNV) and RNA-seq to quantify gene expression levels. (Mwenifumbo and Marra, 2013)

Moving beyond this traditional template of bulk DNA sequencing in matched cancer–normal pairs, other approaches to cancer genome interrogation include multi-sample, multi-region, and single-cell designs, combined with a burgeoning variety of new long-read and single-molecule sequencing technologies.

1.2.2 Insight from somatic SNVs

As a core output of both WES and WGS data with relatively simple properties to identify and analyse, the SNV has been the most intensively studied class of somatic genome alteration in the modern sequencing era. Analysis of somatic SNVs has yielded substantial insight into their underlying generative mechanisms (Alexandrov et al., 2013b; Helleday et al., 2014) and functional implications as driver events within genes (Kandoth et al., 2013; Lawrence et al., 2014) and, to a lesser extent, non-coding regions (Khurana et al., 2016). Patterns of SNV allele fraction have shed light on the sub-clonal phylogenetic evolution of tumours, and the relationships between primary and metastatic sites (Macintyre et al., 2016a; Schwartz and Schäffer, 2017). In concert with other -omics assays, SNV data has also been instrumental in describing the molecular subtypes of different cancer histologies (Hoadley et al., 2014; Bailey et al., 2016). Comprehensive studies of structural variation have been slower to emerge, partly because of the paucity of WGS relative to exome data, and partly because the complexity and variety of rearrangement events pose considerable analytical challenges. Section 1.4 outlines our current understanding of the somatic rearrangement landscape in human cancer.

1.2.3 Clinical translation

Efforts to characterise cancer genomes are motivated partly by the insight into molecular biology, and partly by the promise of clinical translation and improved patient outcomes. Findings from cancer genome studies are already proving their clinical worth, with at least eleven genetic alterations specifically targeted by FDA-approved therapies in ten different cancer types (as of early 2017), and dozens more genes on track for targeted drug development (Hyman et al., 2017). As diagnosis moves to incorporate molecular and genetic markers, new 'basket' clinical trials are beginning to test therapies by gene target in addition—or even in preference—to histology and tissue of origin. For example, drugs approved to target BRAF V600 mutations in melanoma may be used to treat other cancers with the same driver mutation (Hyman et al., 2015). Beyond precision therapies, detailed genomic profiling also improves prognostic accuracy (Ng et al., 2016; Gerstung et al., 2017), and has led to novel technologies for personalised medicine such as relapse monitoring of circulating cell-free DNA (Wan et al., 2017; Siravegna et al., 2017). Personalised, genome-driven oncology may soon be a routine addition to patient care, with the Genomics England initiative currently in progress to sequence whole genomes of 25,000 cancer patients in a clinical setting (Peplow, 2016; Genomics England, 2017).

1.3 Discovering rearrangements in the cancer genome

In addition to SNVs and small indels, somatic genomes also develop larger structural variants wherein kilobases, megabases, or whole chromosomes are deleted, amplified, or otherwise rearranged from the germline state. In this thesis I use the terms genome rearrangement and structural variation (SV) interchangeably. With the first deluge of cancer sequencing data over 2010–2015, publication of SNV analyses far outpaced those on SVs. However, long before high-throughput DNA sequencing and the focus on point mutations, cancer genomes were historically described in terms of large cytogenetic aberrations. As the cancer genomics field matures and the task of gleaning new insight from SNVs becomes harder, the time is right to refocus attention on somatic rearrangements, capitalising on the improved power and resolution afforded by WGS technology.

1.3.1 History of SV discovery in cancer

Advances in biotechnology have revealed several types of genome rearrangement. In the late 19th and early 20th centuries, David Paul von Hansemann and Theodor Boveri proposed the first chromosomal theories on the origins of cancer after observing abnormal chromosome content and asymmetric mitoses in tumour cells (contributions reviewed by Bignold et al. (2006)). As cytogenetic techniques improved, researchers visualised whole chromosome gains and losses (Spriggs et al., 1962), double minutes (Cox et al., 1965), translocations (Rowley, 1973), breakage-fusion-bridge cycles (Gisselsson et al., 2000), and megabase-scale deletions, insertions, and inversions (Sandberg, 1991).

One of the earliest successes from the cytogenetic era was the characterisation of the chr9;chr22 translocation causing the *BCR–ABL* oncogenic fusion gene in chronic myeloid leukaemia (Rowley, 1973) (Nowell (2007) recounts the history of its discovery). With the consequent development of targeted tyrosine kinase inhibitors, the life expectancy of CML patients is now comparable to the general population (Bower et al., 2016).

Moving beyond cytogenetic visualisation of M-phase chromosomes, the detection resolution for copy number alterations (CNA) was refined to a sub-megabase scale with the development of aCGH (Pinkel et al., 1998) and SNP arrays (Zhao et al., 2004; Bignell et al., 2004). CN array methods quantify the degree of copy loss or gain along the reference genome to a resolution of several kilobases, and are still commonly used to supplement WES studies (Zack et al., 2013). However, array technology cannot pinpoint the underlying events actually causing copy number change, and are powerless to detect copy-neutral rearrangement (with the exception of loss-of-heterozygosity (LOH) detectable by SNP array).

1.3.2 Somatic SVs in WGS data

Whole genome sequencing allows all rearrangement classes at any size[e] to be identified at base-pair breakpoint resolution (Korbel et al., 2007; Campbell et al., 2008). In addition to the many chromosome abnormalities identified in the cytogenetic era, sequencing data has revealed novel rearrangement patterns including chromothripsis (Stephens et al., 2011), chromoplexy (Berger et al., 2011; Baca et al., 2013), and chromoanasynthesis (Liu et al., 2011), described further in Section 1.4.2.

[e]SV detection below ~ 1 kb is poor if the read-group orientation is normal (deletion-type).

First generation SV calling algorithms (reviewed by Liu et al. (2015)) use reference-mapped paired-end reads to find groups of split[f] and/or discordantly mapping[g] read pairs which demarcate breakpoint junction positions. The range of possible SV detection methods continues to expand, with more than 20 published algorithms for short-read WGS data available as of late 2017. Some of the more recent contributions concentrate on:

- incorporating depth of coverage (copy number) (for example, SV-Bay finds likely breakpoints under a Bayesian model linking discordant read positions with concordant read depth (Iakovishina et al., 2016); COSMOS prioritises SV calls using strand-specific coverage (Yamagata et al., 2016));
- local assembly around purported breakpoints (for example, novoBreak assembles reads containing the same cancer-specific k-mers (Chong et al., 2016); SvABA assembles abnormal reads mapping to the same reference loci (Wala et al., 2017b)); and
- different ways of comparing matched cancer-normal samples to account for the germline SV background (for example, SMUFIN performs reference-free raw read comparison (Moncunill et al., 2014); PSSV estimates the joint probability of specific hidden genotype states (Chen et al., 2016)).

Regardless of the method, all algorithms are bound by the intrinsic limitation of read lengths being shorter than some repeat sequences, and have low power to detect SVs in ambiguous regions around telomeres and centromeres.

The core output from a standard SV caller is a set of breakpoint junctions (BPJ), each identifying two reference positions juxtaposed in a specified orientation. In addition, the nucleotide sequence detail can detect microhomology or small non-templated base insertions at each junction. WGS data also facilitates genome-wide CN estimation by segmentation of normalised read depth (reviewed by Liu et al. (2013)).

Ideally, a bioinformatics pipeline would also classify SV events by their broader structural context to distinguish simple events of one or two BPJ from medium complexity events of ∼3–9 BPJ or highly complex clusters of ∼10–1000 BPJ. So far, systematic SV classification in cancer WGS data has been largely confined to the basic orientation pattern of individual junctions (Yang et al., 2013; Zhuang and Weng, 2015; Alaei-Mahabadi et al., 2016). Some studies have

[f] A split read has a portion mapping to the reference location, with the remaining portion soft-clipped.

[g] A discordantly mapping read pair has non-standard orientations and/or a mapping distance inconsistent with the library insert size.

augmented this with one or two additional caveats by copy number, cluster separation, and/or broad classification of chromothriptic patterns (Patch et al., 2015; Nik-Zainal et al., 2016; Fraser et al., 2017).

1.4 Patterns of structural variation

The breadth of rearrangement observed in cancer sequencing data reflects the diverse range of DNA alteration that is not only possible, but evidently both consistent with and beneficial to cellular survival, even to the point of continuous pathological growth. Somatic SV catalogues are a window into the dynamics of genome upkeep, and hint at where and when different structural changes arise, whether in specific genome loci, cell types, genotype background, stage of tumour evolution and so on. However, the underlying mechanisms actually generating these rearrangements are not always obvious, and we rely on characteristic fingerprints such as microhomology and copy number profile to implicate known and undiscovered pathways of DNA damage and repair.

1.4.1 Mechanisms of repair and rearrangement at a DNA break

Genome rearrangements are generated by a variety of mechanisms, with many details still unknown. In general, they form during repair of double-strand breaks (DSB) caused by DNA damage, replication fork collapse, telomere attrition, or enzymatic activity. Free DNA ends are substrates for several possible processes, including resection, annealing, ligation, strand invasion, polymerisation, and telomere capture (Kasparek and Humphrey, 2011). DNA repair pathways employ these steps in varying combinations to secure ongoing genome integrity, even at the expense of some local rearrangement.

DSB repair pathways fall into two broad camps: 'break and ligate' mechanisms where two free DNA ends are pasted together; and 'template and replicate' mechanisms where one free end is extended through DNA polymerisation against a template sequence. For detailed reviews, see Willis et al. (2015), Ceccaldi et al. (2016), and Rodgers and McVey (2016).

In brief, the classic 'break and ligate' pathway of non-homologous end-joining (NHEJ) operates throughout the cell cycle (especially in G0/G1) to ligate blunt DNA ends. An alternative mechanism termed microhomology-mediated

end-joining (MMEJ) ligates slightly resected DNA ends[h] with a few bases of overlapping microhomology (MH). If heavily resected DNA ends share long (> 20 bp) homology, then single-stranded annealing (SSA) can stabilise their connection in new duplex DNA, and ligate the backbones after 3' flap digestion and DNA synthesis to fill in the gaps.

The classic 'template and replicate' pathway of homologous recombination (HR) operates during S and G2 phases of the cell cycle, and starts with strand invasion of a 3' single strand overhang to a template sequence with shared homology—preferably finding the sister chromatid for exact sequence preservation. Following strand invasion, DNA synthesis extends the nascent strand along the template, leaving the other strand displaced in a 'D-loop'. Somatic cells primarily resolve HR with synthesis-dependent strand annealing, in which the nascent strand is free to anneal to homologous sequence as it detaches from the template, and ideally finds its duplex partner on the opposing side of the original DSB to mediate error-free repair. An alternative form termed break-induced replication (BIR) continues synthesis of the invading strand in a migrating D-loop for many kilobases, proceeding until the D-loop destabilises or encounters the next obstacle (e.g. replication fork, transcription bubble, chromosome end). The stretch of newly synthesised single stranded DNA trailing from the D-loop is vulnerable to mutation, and is a probable substrate for APOBEC-mediated kataegis clusters[i]. In contrast to the established BIR model which relies on RAD51 homology search to initiate strand invasion, a RAD51-independent pathway termed microhomology-mediated break-induced replication (MMBIR)[j] appears to act in similar fashion, with the relaxed requirement of short MH between the invading and template strands. Indeed, the low-fidelity action of translesion polymerases may even facilitate MMBIR strand invasion in the absence of any pre-existing MH (Sakofsky et al., 2015).

DNA break repair mechanisms have a propensity to generate rearrangement structures through ligation of non-contiguous sequences, or inappropriate template choice and template switching. For example, stalled replication forks may trigger tandem duplication, either by end-joining of staggered breaks in two sister chromatids or re-replication bubble (break and ligate), or by strand invasion to the sister behind the original break locus (template and replicate)

[h] Enzymatic resection at the DSB leaves 3' overhanging single stranded DNA.

[i] Kataegis is a dense hypermutation cluster of ∼5–100 SNV. APOBEC is a family of cytidine deaminases which act on single stranded nucleic acid, with an important role in mutational disarmament of invading viral sequence.

[j] The MMBIR mechanism is also described in the literature as fork-stalling and template switching (FOSTES, Lee et al. (2007)).

(Finn and Li, 2013; Costantino et al., 2014; Willis et al., 2015). Likewise, deletions and translocations may be caused by aberrant end-joining of two DSB positions, or by strand invasion to a distant locus (Roukos and Misteli, 2014; Sakofsky and Malkova, 2017).

The repercussions of structural DNA repair and remodelling extend well beyond one or two break positions, and occasional bursts of genomic upheaval generate complex SV spanning tens or hundreds of breakpoint junctions.

1.4.2 Complex rearrangements

SV clusters arise from special cases of DNA breakage, and are not typically the mere overlap of simple events independently acquired.

Stephens et al. (2011) first described chromothripsis, characterised by dozens of BPJ shuffled together over one or more reference chromosomes with an oscillating copy number profile (Korbel and Campbell, 2013). This complex configuration results from a catastrophic shattering event, such as befalls lagging DNA caught in a micronucleus (Zhang et al., 2015) or chromatin bridge (Maciejowski et al., 2015) after aberrant mitosis. Subsequent ligation of a random combination of disjoint fragments generates a highly disordered derivative chromosome, with several fragments lost altogether.

Another 'break and ligate' pattern termed chromoplexy was first described in prostate cancer as a largely copy-neutral cycle of reciprocal exchange at multiple loci (Berger et al., 2011; Baca et al., 2013). The observed balancing of translocation partners across many chromosomes is hypothesised to result from correlated DSBs in spatio-temporal proximity, presumably mediated by androgen receptor activity in prostate.

Extrachromosomal DNA fragments generated by chromothripsis-type shattering events (or other means) often circularise to form double minutes (DM). These acentric DNA circles are free to segregate asymmetrically during mitosis, and are an efficient vehicle for oncogene amplification. DM copies can also re-integrate into the linear chromosome complement, forming intrachromosomal amplicon structures (also known as homogeneously staining regions). Internal DM composition may combine non-templated sequence insertions with small and large segments from several reference chromosomes, evolving through multiple rounds of integration and recombination. (Sanborn et al., 2013; L'Abbate et al., 2014; Vogt et al., 2014; Turner et al., 2017)

A different route to intrachromosomal sequence amplification is through successive breakage-fusion-bridge (BFB) cycles. In the classic model proposed by McClintock (1941), fusion of two atelomeric sister chromatids forms a dicentric chromosome which gets pulled apart during anaphase, passing a foldback SV (one-sided inversion) to one daughter cell. If multiple cell divisions repeat this cycle before the derivative is stabilised via telomere acquisition, then BFB imparts a characteristic foldback SV cluster with a step-like CN profile (Kinsella and Bafna, 2012; Greenman et al., 2016).

Break and ligate events—such as BFB, DM formation and chromothripsis—sometimes overlap to generate highly convoluted derivatives with little resemblance to their germline chromosome antecedents (Garsed et al., 2014; Li et al., 2014; Notta et al., 2016). Presumably, the inherent instability of some aberrant structures means that one large rearrangement may beget another, thus accounting for the prevalence of complex overlap observed in several cancers.

Replication mechanisms also generate complex SV via serial template switching, with distinctive patterns of copy gain, MH enrichment, and small, locally-templated insertions in the junctions between more distal BPJ (Lee et al., 2007; Zhang et al., 2009). These events have primarily been described in germline developmental disorders, and range from medium complexity SV like the duplication–inverted triplication–duplication (Carvalho et al., 2011), to high complexity events involving five or more BPJ termed chromoanasynthesis (Liu et al., 2011), possibly triggered by interstrand crosslinks or other persistent DNA lesions (Meier et al., 2014). Experimental studies support a MMBIR mechanism (Sakofsky et al., 2015; Hartlerode et al., 2016) with low-fidelity polymerases also generating nearby SNVs and indels (Carvalho et al., 2013).

1.4.3 Prevalence and distribution across the genome

The character and extent of somatic rearrangement is highly variable, depending on the fidelity of replication, rate of DNA breakage, choice of repair pathway, and subsequent effectiveness of that repair. WGS data indicate that most cancer samples have tens to hundreds of detectable BPJ, with the burden varying by an order of magnitude both across and within cancer types, from highly rearranged breast and ovarian genomes, to relatively stable genomes in kidney and thyroid cancer (Yang et al., 2013; Alaei-Mahabadi et al., 2016). Some cancers present with a strong tandem duplicator phenotype, especially those breast and ovarian cancers with both *BRCA1* and *TP53* mutations (Menghi

et al., 2016). Moreover, bone and soft-tissue cancers are particularly prone to chromothripsis (Stephens et al., 2011; Cai et al., 2014), while prostate cancer is notable for the prevalence of chromoplexy (Baca et al., 2013). The observation that most somatic BPJ have no or micro (1–5 bp) junction homology suggests that NHEJ, MMEJ, and MMBIR are the major pathways to cancer rearrangement, while non-allelic HR is largely confined to germline disorders (Drier et al., 2013; Malhotra et al., 2013; Yang et al., 2013; Carvalho and Lupski, 2016).

The variable forces of DNA breakage and repair not only dictate the *number* of BPJ per sample, but also their *location* in the genome. In B cells, deliberate enzymatic DSB generation renders immune loci particularly prone to translocation, often contributing to oncogenic fusions (Vaandrager et al., 2000; Alt et al., 2013). In prostate, androgen receptor signalling leads to topoisomerase DSBs in specific regulatory locations, often triggering the *TMPRSS2–ERG* fusion driver (Lin et al., 2009; Haffner et al., 2010). Retrotransposons are another source of recurrent SV, with particular L1 hotspots generating dozens of somatic insertion/transduction events in some cancers (Lee et al., 2012; Tubio et al., 2014; Helman et al., 2014). Common fragile sites are recurrent foci of deletion in many cancer types, associated with low density of replication forks, late replication time, large genes, and active transcription (Ozeri-Galai et al., 2012; Sarni and Kerem, 2016; Glover et al., 2017). Aside from these rearrangement hotspots, BPJ also correlate more generally with: spatial proximity inside the nucleus (Fudenberg et al., 2011; Hakim et al., 2012; Zhang et al., 2012); replication timing[k] (De and Michor, 2011; Pedersen and De, 2013); simple repeats (Bacolla et al., 2016); chromatin modifications (Black et al., 2013; Burman et al., 2015); and show sample-specific association patterns (Drier et al., 2013).

1.5 Functional consequences of rearrangement

Rearrangement landscapes observed in clinically-detectable cancer samples reflect the distribution of events at generation, moulded by selection on the functional consequences. Events which substantially reduce cell fitness are subject to purifying selection, and are not typically observed. Conventional theories posit that most somatic mutations are passenger events with negligible fitness effect, and that only a handful of positively-selected drivers are responsible for clonal expansion of the cancer lineage. A high passenger-to-driver ratio is well substantiated for point mutations (Tomasetti et al., 2015; Martincorena et al.,

[k]Replication timing tends to be late for copy loss, and early for both copy gain and LOH.

2017; Sabarinathan et al., 2017), and presumably extends to most SV classes as well. As a probable exception to this general paradigm, those complex SV events that restructure hundreds of megabases effect such a drastic departure from the normal diploid genome that passenger status seems unlikely.

Rearrangements drive the cancer phenotype through various means, including production of oncogenic fusion genes, amplification of oncogenes, deletion or disruption of tumour suppressors, and repurposing of regulatory regions. These alterations play a major role in cancer development, with COSMIC curating 73% of 547 census cancer genes as being affected by translocation or CNA (v71, (Forbes et al., 2015)). Even with the additional insight provided by RNA-seq data, it remains extremely challenging to distinguish the driver SVs from the passengers, and to discern which of the many changes to genes and/or regulatory elements meaningfully contribute to oncogenesis.

1.5.1 Fusion genes

Some rearrangements create fusion genes by placing one gene (or part thereof) downstream of a different promoter region (with or without the 5' end of the promoter's native open reading frame). Fusion genes drive cancer by placing a proto-oncogene under the control of a highly active promoter, or by the translation of a chimeric protein product with novel oncogenic properties. (Mertens et al., 2015)

Any SV class is capable of generating a fusion gene via the juxtaposition of non-contiguous sequences. For example, the *BCR–ABL* fusion driving chronic myeloid leukaemia is generated by translocation (Salesse and Verfaillie, 2002; Nowell, 2007); *KIAA1549–BRAF* in pilocytic astrocytoma is generated by tandem duplication (Jones et al., 2008); whereas *TMPRSS2–ERG* in prostate cancer is fused through deletion or chromoplexy (St John et al., 2012; Baca et al., 2013).

1.5.2 Gene dosage

A gene's transcriptional output is roughly correlated with its copy number in the genome (Fehrmann et al., 2015), and thus SV events generating regions of copy gain or loss may drive cancer by oncogene over-expression or tumour suppressor haploinsufficiency or two-hit loss. Roughly 40 peak regions of recurrent CNA span a known cancer gene (Beroukhim et al., 2010; Zack et al., 2013), such as

the *MYC* oncogene amplified in 13–17% of all breast and ovarian cancers and the *CDKN2A* tumour suppressor lost in 33% of brain cancers[l].

Regions of copy alteration often span multiple genes, and may drive cancer through the combined fitness effect of their synchronous dosage change. In the maximal case, whole chromosome or arm-level aneuploidy simultaneously alters the copy level for hundreds of genes. Some arms are strongly biased towards gain (e.g. 7p, 8q, 20q) or loss (e.g. 9p, 13q, 17p), reflecting the uneven distribution of tumour promoting or suppressing regions (Beroukhim et al., 2010; Kim et al., 2013; Davoli et al., 2013). Considering a smaller scale of several megabases, Liu et al. (2016) reported that the selective advantage of *TP53* tumour suppressor loss is boosted by co-deletion of neighbouring genes. Likewise, Hagerstrand et al. (2013) described the joint amplification in 3q26 of two oncogenes promoting cell growth and invasion. Beyond the single-copy gains proffered by aneuploidy or tandem duplication, the most efficient route to high-magnitude amplification is via extrachromosomal DMs, frequently boosting oncogenes like *MYC* and *EGFR* to CN levels above ten (Turner et al., 2017).

Amplifying enhancer[m] dosage is another route to oncogene over-expression, without necessarily changing the copy level of the gene itself (Zhang et al., 2016; Glodzik et al., 2017).

1.5.3 Altered regulation

Interphase chromosomes are organised in a looping architecture of topologically associating domains (TAD) which divide the linear sequence into self-interacting blocks (typically hundreds of kilobases) with coordinated gene expression and replication timing. TADs are physically separated from their neighbours by insulating boundary regions held together by CTCF and cohesin. Within a TAD, DNA looping allows enhancer elements to recruit transcription factors for genes up to a megabase away. DNA looping also ensures that enhancers are typically restricted from accessing and regulating genes in any separate TAD. Although TAD boundaries are conserved across cell types (and even species), the TADs themselves are dynamic units, localising in either active or repressive nuclear compartments to regulate tissue-specific gene expression programs. (Pombo and Dillon, 2015; Ruiz-Velasco and Zaugg, 2017)

[l]CNA statistics from the COSMIC database (Forbes et al., 2015); other cancer types not specified are also commonly affected by CNA at *MYC* and *CDKN2A*.

[m]Enhancer elements are *cis*-acting regulatory regions which recruit transcription factors to promote expression of genes brought in to proximity by DNA looping.

Mouse models show that SV events which duplicate or delete TAD boundaries result in merged or neo-TAD structures. Such alterations place genes in a novel regulatory context, drastically changing their expression levels with potentially serious phenotypic consequences. (Lupiáñez et al., 2015; Franke et al., 2016)

Chromatin topology remodelling and ectopic enhancer activity has also been observed in cancer, with the capacity to activate oncogenes and down-regulate tumour suppressors (Valton and Dekker, 2016). Early findings highlighted recurrent 'enhancer-hijacking' rearrangements up-regulating *EVI1* (alias *MECOM*) in acute myeloid leukaemia (Gröschel et al., 2014) and *GFI1A/B* in medulloblastoma (Northcott et al., 2014). Weischenfeldt et al. (2017) surveyed over 7000 cancer samples to find more than a dozen oncogenes likely to be activated in this manner, including the *IGF2* gene recurrently involved in a boundary-spanning tandem duplication in colorectal cancer. This simple SV event generates a neo-TAD structure linking *IGF2* with an active super-enhancer from the neighbouring region (usually insulated from each other by the boundary), causing an oncogene expression increase of more than 250-fold.

Given the immense influence of enhancer contact on gene regulation, TAD-disrupting SV events can drastically affect genes as far as a megabase from the breakpoint, irrespective of any fusion or dosage changes. The ability of rearrangements to transmute the chromatin organisation domains so faithfully preserved across tissues and species is now emerging as an under-appreciated pathway to the cancer phenotype.

1.6 Overview of this work

In this thesis I analyse somatic genome rearrangements within 2559 samples from the ICGC Pan-Cancer Analysis of Whole Genomes dataset, focussing on structural classes and properties (Chapter 2), the genome-wide distribution pattern (Chapter 3), co-occurrence signatures of underlying process (Chapter 4), and complex SV intractable to simple classification (Chapter 5).