Chapter 1

Introduction

1.1 Cancer

Cancer is a family of over 100 diseases and it can originate from most cell types and in nearly all organs (Stratton et al., 2009). In 2012 cancer was responsible for 8.2 million deaths worldwide, while 14.1 million new cases were diagnosed (Ferlay et al., 2015). In the UK, it is estimated that 1 in 2 people will be diagnosed with the disease at some point during their lives (Ahmad et al., 2015). Decades of research have made tangible improvements to patient outcome, in large part, due to the introduction of effective treatment strategies (Narod et al., 2015). For example, breast and prostate cancer incidence rates in the UK have tripled from 1972 to 2013, but the number of mortalities attributed to these cancer types has decreased over the same time frame [1].

1.2 Hallmarks

The body of accumulated research can be summarised into six hallmarks (Hanahan and Weinberg, 2011), of which a brief summary is provided here.

(**1**) Tumours acquire and maintain **chronic proliferation** by activating signalling pathways. Somatic mutations that activate genes involved in growth-promoting and controlling pathways deregulate growth signalling and allow cells to take control over its own destiny. Alternatively, somatic mutations can disrupt negative-feedback mechanisms of cell proliferation.

(**2**) Tumours can become **adverse to growth-suppressors**, either through losses of gene copies or through deactivating somatic mutations (Klein, 1987). Tumour suppressor genes

[1] https://visual.ons.gov.uk/40-years-of-cancer

as *TP53* and *RB1* play a pivotal role in pathways that control the decisions whether cells proliferate, or activate senescence (a state where viable cells no longer proliferate) and apoptosis (programmed cell death).

(**3**) Apoptosis is thought to be a protective measure against cancer. Tumour cells **avoid apoptosis** to stay in a proliferative state, which allows a tumour to continue to grow. The most common circumvention strategy is by deactivating *TP53* that acts as a sensor and can trigger apoptosis. Alternatively, apoptosis can be circumvented by altering the expression of its regulators.

(**4**) Tumour cells acquire the ability to **infinitely replicate**. Cells escape senescence and cell death, which allows tumours to grow to large sizes. The caps that protect the end of chromosomes (telomeres) are thought to play an important role in this process. The length of a cells' telomeres corresponds to how many generations of offspring it can produce and erode over time, triggering cell death when they have sufficiently shortened. Tumour cells circumvent this process by extending telomeric DNA to prohibit the consequences of their erosion. Germline variants and somatic mutations are implicated in telomere lengthening (Robles-Espinoza et al., 2014).

(**5**) As the tumour grows bigger, it becomes more difficult to provide cells with the required oxygen and nutrients. To prevent cells from starvation tumours can acquire new blood vessels (**angiogenesis**) that allow for transportation of these requirements and for removal of waste.

(**6**) Tumour cells acquire the ability to **invade and metastasise** to distant sites. Evidence exists that the process of metastatic dissemination can occur late during tumour evolution (late dissemination) (Yachida et al., 2010; Yates et al., 2017). But there is also evidence to suggest that metastatic ability can arise early in tumour development (early dissemination) (Coghlin and Murray, 2010).

1.3 Genetics

By observing the brief description of cancer hallmarks above it becomes apparent that somatic mutation and genome instability play key roles in tumourigenesis.

1.3.1 Early observations suggest a role for the genome

Work in the late 19th and early 20th century reported on chromosomal alterations in cells of various human cancers (Boveri, 2008; von Hansemann, 1890). Von Hansemann theorised that the abnormal chromosome counts he observed were due to cellular defects and that the alterations could lead to the development of cancer (von Hansemann, 1890). Theodore Boveri revived the idea in 1914 when he applied observations made in sea urchins to cancers. He hypothesised that tumours may originate from a single cell, that tumour cell populations are genetically unstable and that acquired aneuploidy is passed on to progenitor cells (Boveri, 2008). 16 years later Winge (1930) built on these theories and proposed that consecutive chromosomal alterations could lead to disease progression.

1.3.2 Confirmation of a role in cancer

But it was only after DNA was identified as the molecule that conveys inheritance (Avery et al., 1944; Franklin and Gosling, 1953; Watson and Crick, 1953; Wilkins et al., 1953) that true confirmation came that genetics plays a key role in cancer development: with the discovery of the Philadelphia chromosome (Nowell and Hungerford, 1960). The Philadelphia chromosome is a translocation between chromosomes 9 and 22 in chronic myelogenous leukemia that creates a fusion-gene between *BCR* and *ABL1* (Rowley, 1973). Its finding lead to more chromosome count abnormalities being reported in patients with advanced disease (Sandberg, 1966).

But in the early 1980s it was thought more likely that cancer development was caused by transposon activity, then by changes in the genetic sequence of genes (Cairns, 1981). It wasn't until 1982 that the first single sequence change was shown to be the activating event of an oncogene when Reddy et al. (1982) showed that a single G>T substitution in *HRAS* is enough to activate its oncogenic potential.

1.4 Evolution

Confirmation of a role for alterations in the genome lead to theories about how many mutations would be required for a tumour to arise.

1.4.1 Estimates of the number of mutations to form a cancer

In 1953 Nordling combined observations reported in several papers from the 1940s into a theory of a cancer-inducing mechanism. He reasoned that if cells were left a sufficiently long time, a genetic mutation would occur and if the cell with a mutation produces daughter cells, then it could acquire more mutations. If the mutation speeds up propagation of a cell, then this process would occur more quickly. By examining the age distribution of cancer patients from various countries he observed that cancer mortality increased "by a certain power (to the sixth) of age" and thus postulated that six mutations were required for tumour development (Nordling, 1953).

In 1971 Knudson compared the ages of patients with sporadic and familial retinoblasoma and observed that patients with the familial predisposition were much younger (Knudson, 1971). He speculated that familial retinoblastoma patients were already born with one mutation and therefore required just a single mutation, while patients without the predisposition required two, hence the difference in ages between the groups. The gene in question was later shown to be *RB1* (Murphree and Benedict, 1984).

1.4.2 An integrated theory of tumour evolution

For some time, the thought prevailed that tumour cell populations were underpinned by one or multiple stem cells that drive the growth of each population (stemlines) (Roberts and Trevan, 1960). But mounting evidence suggested that tumours could arise from a single cell due to the similarity in karyotypes observed between cells of the same tumour (Ford and Clarke, 1963; Hauschka, 1961; Levan and Biesele, 1958; Makino, 1957) and that tumours progress as tumour-cell populations acquire additional mutations in a process termed clonal evolution (Adam et al., 1970; de Grouchy et al., 1966; Foulds, 1957). Furthermore, it was observed that neoplasms could give rise to malignant growths (Morson, 1974).

It was Peter Nowell (1976) who combined all these observations into a single theory of tumour evolution. He proposed the following model: tumour initiation occurs when a normal cell acquires a selective growth advantage, which allows its offspring to become neoplastic. The cells proliferate and due to ongoing chromosomal and genetic instability they generate mutant daughter cells. Nearly all the introduced mutations are eliminated due to a lack of selective advantage, but a cell that acquires a mutation that does convey a selective advantage becomes the precursor for a new subpopulation.

1.5 Drivers

Nowell suggested that tumours evolve through a process of clonal expansion, where expansions are initiated by the selective advantage gained through a driver mutation and that process could eventually lead to metastasis and resistance to therapy (Nowell, 1976). Cairns (1975) suggested that these driver mutations may be introduced as errors through cell renewal programmes, solidifying that the process of carcinogenisis is an internal process.

1.5.1 Oncogenes

By the mid-1980s there was evidence to support Nowell's theory as 40 oncogenes (genes for which their oncogenenic function is activated through a single mutation, i.e. are dominant) had been identified (Weinberg, 1985) and there was evidence that oncogenes could be activated through point mutations, amplification or deletions and translocations (Nowell, 1986). Meanwhile, Pegoraro et al. (1984) suggested that the successive activation of two oncogenes could explain the aggressive clinical behaviour of ALL cases (first a t(14,18) translocation (Tsujimoto et al., 1984), followed by a t(8,14) translocation (Dalla-Favera et al., 1982) that fuse both *BCL2* and *MYC* to the immune heavy IGH region on chromosome 14), underpinning the theory of a decade earlier.

1.5.2 Tumour suppressor genes

The existence of tumour suppressor genes (TSGs), which require both copies to be deactivated (i.e. are recessive), was confirmed shortly after. *RB1* was the first to be identified (Morson, 1974). A number of other genomic regions were already suspected (Klein, 1987) and were subsequently shown to contain TSGs, partly due to families with a germline predisposition (Knudson, 1993), and included some of the most frequently mutated cancer genes: *VHL* (Seizinger et al., 1988; Tory et al., 1989), *TP53* (Nigro et al., 1989), *WT1* (Haber et al., 1990), *APC* (Nishisho et al., 1991) and *BRCA2* (Wooster et al., 1995, 1994), among others.

1.5.3 Drivers and passenger mutations

We now know that tumours can acquire 1000s of mutations during their life time (Pleasance et al., 2009). Not all of these mutations convey a selective advantage or disadvantage to the cell in which they occur, which leads to the notion of *driver* and *passenger* mutations (Stratton et al., 2009). Tumours often contain many more passenger mutations than drivers and the passengers are thought not to contribute to cancer development (a thought that is not entirely uncontested (Supek et al., 2014)), however, they have provided useful to study the process of tumour evolution, as will be explained further below.

1.6 High throughput technology

The advent of high throughput technology to perform genome wide screening for genomic alterations has proven to be a rich medium on which to measure somatic alterations. Somatic alterations can be found by performing the same experiment on a tumour sample and a normal sample from the same donor. The normal sample is often taken from a blood sample, but sometimes from adjacent tissue. By calling events in the matched normal against a reference sample one obtains those that constitute the 'germline' of the donor, which can then be subtracted from those found in the tumour to obtain somatic events (Pleasance et al., 2009). All the technology briefly described below operate on pooled DNA from many individual cells. Which means the somatic mutations measured must be carried by a large proportion (but not all, as will be covered later) of the cells that are prepared for the high throughput procedure. Mutations that are available at the level of a single cell (or shared between small proportions of cells) are not measured.

1.6.1 Array based technology

The first high throughput technology was comparative genomic hybridization (CGH) (Kallioniemi et al., 1992), for which the array development could be readily used to detect copy number alterations down to 100kb in cancers (Pinkel and Albertson, 2005). Soon afterwards SNP arrays arrived on the scene which had the advantage that they could detect regions of loss of heterozygosity (LOH) (Pfeifer et al., 2007; Schaaf et al., 2011). The CGH platform could only detect the total amount of DNA available (logR), SNP arrays also include the b-allele frequency (BAF) measure that accounts the availability of the two alleles at the SNP location. Heterozygous SNPs could therefore be used to quantify allele specific copy number.

1.6.2 Sequencing technology

But it wasn't until massively parallel sequencing technology arrived that the full compendium of somatic mutations could be measured (Margulies et al., 2005; Shendure et al., 2005). As was demonstrated by Pleasance et al. (2009), and is detailed further below, sequencing of exomes first provided access to all protein coding regions of the genome at base-pair resolution, while genome sequencing also yielded mutations in intergenic regions, highly detailed copy number and structural variation.

1.6.3 The emergence of sequencing consortia

The availability of these high throughput technologies, coupled with a drop in price, saw the emergence of large cancer sequencing consortia in the American The Cancer Genome Atlas (TCGA) and later the International Cancer Genome Consortium (ICGC). Both consortia aimed to paint a complete picture across cancer types by collecting large numbers of samples for exome (TCGA, although genomes were also sequenced) and whole genome

(ICGC) sequencing. TCGA systematically collected DNA, RNA, methylation and clinical data from over 10,000 cancer patients with the aim to improve diagnosis through a better understanding of landscape of somatic alterations in cancers. ICGC aims to further increase our understanding by coordinating the sequencing of 25,000 whole cancer genomes with the participation of individual (national) projects that contribute particular cancer types.

The availability of these data sets allows researchers to paint an ever increasingly detailed picture of what cancer genomes look like.

1.7 Copy number

The role of aneuploidy in cancer development has been long since known. When high throughput technology became available to systematically measure aneuploidy across the genome it was directly applied and provided further insight into the extent and the patterns by which the cancer genome is altered.

1.7.1 Confirming classic knowledge

Pollack et al. (2002) reported that patterns of copy number alterations (CNAs) across 44 primary breast cancers and 10 breast cancer derived cell lines correspond well with what was known from cytogenetic studies. This study also included micro-array based expression profiling, which showed that CNAs can lead to big changes in gene expression. The authors reported that a 2-fold change in copy number was associated with a 1.5-fold change in expression and that the majority of highly amplified genes are highly to moderately high expressed.

Expression arrays had already shown that the expression profiles of breast cancers cluster in subtypes (Perou et al., 2000). Bergamaschi et al. (2006) then showed that copy number alterations in breast cancers are linked to these new subtypes. Basal-like tumours were associated with more gains and losses, while luminal-B tumours showed more high amplifications. High level amplifications were associated with genes that could be drug targets (Chin et al., 2006). These findings highlighted that breast cancer subtypes have distinct copy number profiles that contain clues about the underlying differences in biological process that shaped the cancer.

1.7.2 Pan-cancer overview of CNAs

SNP arrays were quickly shown to also detect regions of copy neutral LOH (Nannya et al., 2005; Zhao et al., 2004) and therefore to provide a more complete picture of copy number alterations. The first landscape paper about CNAs used SNP arrays and reported on profiles from 3,131 samples across 26 types of cancers (Beroukhim et al., 2010). The study revealed that across cancer types copy number profiles have several characteristics in common. The size distribution of copy number events appears bimodal: one mode represents arm level events, the other focal events and the frequency at which focal events are observed is inverse proportional to their size. Many tumours contain focal deletions in known tumour suppressor genes, while the focal gains amplify known oncogenes, further strengthening the link between CNAs and oncogenic function.

When the TCGA project was devised, it was set up such that SNP arrays were collected for every tumour to perform copy number analysis. Zack et al. (2013) paint the emerging picture across 4,934 cancers and report that 37% of cancers have a whole genome duplication and that tumours with a duplication contained more CNAs. The authors speculate that the bimodal CNA size distribution could be due to different mechanisms by which CNAs are acquired. They observe that both focal and arm level events are larger if one end of the event contained a telomere. Finally, they report that recurrent copy number events that do not affect a known cancer gene. Some regions also contain significantly mutated genes suggesting these regions may play an important role in tumour development.

1.8 Massively parallel sequencing of cancer genomes

The advent of massively parallel sequencing brought with it a new era in which the whole cancer genome could be interrogated for single base substitutions, as well as larger scale copy number alterations and rearrangements. A first large scale screening of all genes in the RAS–RAF–MEK–ERK–MAP kinase pathway in the early 2000s had already shown the potential of such approaches with the identification of BRAF as a cancer gene in melanoma (Davies et al., 2002) and non-small cell lung cancer (Brose et al., 2002). And a screen of all protein kinases in 25 breast cancers had already revealed that some tumours contain no mutations in these genes, whilst some contained numerous mutations, suggesting the existence of a mutator phenotype (Stephens et al., 2005).

1.8.1 Early findings from exome sequencing

Due to initial technical limitations, early sequencing studies focussed on the coding regions of the genome, which means single nucleotide variants (SNVs) and short insertions and deletions (indels) could be detected in about 3% of the genome. The early exome sequencing studies nonetheless immediately revealed interesting insights. Wood et al. (2007) reported that only a handful of genes across 11 breast and 11 colorectal tumours were commonly mutated, but that many other genes were mutated at low frequency. This finding was corroborated by Ding et al. (2008) when they reported that 26 out of 623 sequenced genes were significantly mutated across 188 lung adenocarcinomas. And the pilot of TCGA project contained the exome sequences of 206 glioblastoma cases which revealed an unexpectedly high number of mutations in *PIK3R1* (TCGA Network, 2008), which was later confirmed to be a glioblastoma driver (Weber et al., 2011).

1.8.2 The full compendium of somatic alterations

The first full catalogues of somatic mutations across the whole genome, and including copy number and structural variations, arrived a little later. Pleasance et al. (2009) sequenced a cell line that is derived from a metastatic melanoma case. The sample yielded 33,345 SNVs, 66 indels, 37 rearrangements and the copy number analysis yielded several highly amplified and several homozygously deleted genes. Similar findings were reported on the whole genome sequence of a cell line derived from a bone metastasis of a small-cell lung cancer patient; 22,910 SNVs (of which the majority in intergenic regions), 65 indels, 58 rearrangements and a range of copy number alterations (Pleasance et al., 2010).

1.8.3 Whole genome sequencing reveals the extent of somatic alterations in cancer genomes

Numerous sequencing studies have since found small numbers of recurrently mutated genes in relatively small data sets (Ellis et al., 2012; Fujimoto et al., 2012; Puente et al., 2011; Waddell et al., 2015; Wang et al., 2014). Larger sequencing studies have yielded larger numbers of frequently mutated genes, but often a handful are shared among many samples and the remaining genes are found mutated in only a few cases (Dulak et al., 2013; Nik-Zainal et al., 2016; Stephens et al., 2012). The combined studies have given us a good idea of the mutation rates in human cancers and have shown that mutation rate correlates with DNA replication time (Lawrence et al., 2013).

Whole genome sequencing lead to the discovery of recurrent mutations in the TERT promotor that are important for tumour development in a number of cancer types (Fujimoto et al., 2012; Horn et al., 2013; Huang et al., 2013; Vinagre et al., 2013) and have shown evidence of L1-retrotransposon activity in over half the tumours evaluated (Tubio et al., 2014). Studies have reported on mutational patterns that correlate with subtypes (Ellis et al., 2012; Puente et al., 2011; Waddell et al., 2015) and treatment outcome (Puente et al., 2011; Wang et al., 2014), including chemotherapy resistance (Patch et al., 2015).

1.9 Mutational processes

With the full catalogue of mutations now detectable it became possible to investigate the processes by which these mutations are generated. Pleasance et al. (2009) observed that a large proportion of mutations found in their melanoma case are C>T/G>A substitutions as a result of exposure to UV light. The lung cancer case reported in Pleasance et al. (2010) showed multiple signs of a smoking signature; for example, the mutation type distribution showed a close correspondence to that observed in mutations within *TP53* in small cell lung cancer cases obtained from literature, and the mutations appeared more often in unmethylated CpG dinucleotides, which confirmed earlier knowledge about smoking associated carcinogens.

1.9.1 Automated extraction of mutational signatures

With more genomes sequenced it became possible to automatically extract signatures that correspond to the mutational processes operative on cancer genomes. Nik-Zainal et al. (2012b) reported on five signatures found across 21 breast cancers and observed that cancers with a *BRCA1* or *BRCA2* mutations clustered together, suggesting the mutations are generated by double-strand break-repair mechanisms. It also contained the first mention of localised hypermutation known as kataegis, which appeared with a particular mutational spectrum that suggested the mutations may be due to the APOBEC family of deaminases.

Characterisation of the mutational processes of over 7,000 exomes and genomes revealed evidence of at least 21 mutational signatures (Alexandrov et al., 2013). Most cancer genomes contain evidence of activity of more than one process, with some genomes containing signs of activity of six signatures and many different combinations of signatures were observed to be jointly active.

1.9.2 Linking signatures to mutational processes

By analysing the samples in which certain signatures were detected it became possible, for some signatures, to suggest the processes by which they were generated. Signatures 1A/1B were strongly correlated with the age of diagnosis, and based on the accumulation of prior evidence it was suggested these mutations might be due to spontaneous deamination (Alexandrov et al., 2015), signature 4 corresponded to previous knowledge about the mutation types generated by tobacco smoke and was predominantly found in the cancer genomes from smokers and signature 7 conformed to prior knowledge about UV induced mutagenesis, suggesting a role of UV light exposure. Alexandrov et al. (2015) further increased the number of signatures to 30 after analysing 10,250 cancer genomes.

These studies suggest that the underlying biological processes that generate the driver and passenger mutations by which cancers evolve are varied and complex.

1.10 Heterogeneity

The ongoing activity of mutational processes in every tumour cell means that no two tumour cells are genetically the same and that tumours are therefore heterogeneous. Driver mutations allow cells to proliferate quicker and expand into a subpopulation of cells. If mutations in these subpopulations can be measured, then one could use these mutations to assess the extent of genetic heterogeneity in the tumour.

1.10.1 Detecting heterogeneity from sequencing data

A pilot project revealed that high-level heterogeneity can be measured through sequencing data. Campbell et al. (2008) sequenced the IGH locus of 22 CLL cases and showed that subclonal populations of tumour cells could be detected through massively parallel sequencing. The IGH locus was chosen in particular because CLL patients show signs of hypermutation within this region, and due to 264 base-pair long reads, it was possible to arrange the SNVs in haplotypes and to arrange the haplotypes into phylogenetic trees. The results indicated that tumours are heterogeneous and that intra-tumour heterogeneity can be detected from sequencing data.

1.10.2 Cancer type specific studies highlight evolutionary properties

Nik-Zainal et al. (2012a) were the first to show that subclones can be detected through bulk whole genome sequencing and that the uncovered evidence could be compiled into the individual life history of a cancer. The authors developed algorithms to detect subclonal copy number, construct haplotypes from nearby SNVs and devised theory that can be used to construct the evolutionary trajectory of a tumour.

The authors reported that each of the 21 tumours in the data set contain a dominant subclone and detect large scale subclonal CNAs in nearly every case. Timing of gains by means of SNVs on one and two chromosome copies (Greenman et al., 2012) revealed the evolutionary patterns that have given rise to each tumour and suggested that breast cancers of the same subtype may evolve similarly. Inspection of the base substitution types showed that mutational signature activity can change between clonal and subclonal mutations.

Since then, various articles that focused on a single cancer type report vast differences in heterogeneity between patients, in which known genes are mutated early in one case, but late in another (Yates et al., 2015) and that some tumours can show evidence of rapid evolution, while other tumours in the same cohort show a stable balance between subclones (Schuh et al., 2012).

The application of treatment can introduce a phase of rapid tumour evolution (Landau et al., 2013, 2015), in which mutations in known drivers are observed to be subclonal (Gerlinger et al., 2014; Landau et al., 2013). Mechanisms of resistance can be acquired in parallel in different lesions (Gerlinger et al., 2014; Gundem et al., 2015), subclones can persist through treatment (Schuh et al., 2012) and the existence of a subclonal driver mutation can be an independent risk factor for disease progression (Landau et al., 2013).

A primary tumour can contain observable signs of metastatic and treatment resistance potential before onset (Yates et al., 2015) and in some cases can contain patterns that predict the evolutionary progression (Landau et al., 2015). Mutational processes can differ between clones and subclones through spatially (de Bruin et al., 2014) and temporally (Bolli et al., 2014) separated samples from the same cancer. Gundem et al. (2015) reported metastasis-to-metastasis seeding in a number of lethal metastatic prostate cancers and Cooper et al. (2015) observed clonal expansions in morphologically normal cells in multifocal prostate tumours.

Two recent in-depth studies of ITH suggest that early tumour development is consistently driven by point mutations, while later evolution contains more CNAs in both small cell lung and colorectal cancers (Jamal-Hanjani et al., 2017; Mamlouk et al., 2017). Jamal-Hanjani et al. (2017) further observe that genome doublings and ongoing genetic instability are associated with ITH and could result in parallel evolution of CNAs. Mamlouk et al. (2017) report on a 3D reconstruction of a single cancer revealed that point mutations in the *APC* and *TP53* genes were evenly distributed throughout the cancer, but gene copy numbers appeared highly variable.

1.10.3 Pan-cancer studies reveal widespread ITH across cancer types

These separate studies hint that intra-tumour heterogeneity is widespread and that tumours of the same cancer type can differ greatly. McGranahan and Swanton (2015a) analysed somatic mutations across 2,694 exome-sequenced tumours representing 9 cancer types from TCGA and found that protein altering mutations in known cancer genes that are possibly actionable in the clinic are typically clonal, but can also be observed subclonal. Analysis of mutational signatures suggested a link between subclonal driver mutations and APOBEC-related mutagenesis.

Andor et al. (2016) performed subclonal reconstruction on 1,165 exome-sequenced tumours from TCGA and report that 86% of tumours across 12 cancer types contain at least one subclone. The authors report that subclones can contain driver mutations and that subclone size correlates with treatment outcome.

These studies show that much can be learned about tumour evolution and heterogeneity through massively parallel sequencing data.

1.11 Subclonal inference

The studies named in the previous section are possible due to the development of two types of methods: Callers for somatic copy number and subclonal architectures. A subclonal inference method, in general, first estimates the proportion of tumour cells that carry each mutation (this is known as cancer cell fraction, CCF). Mutations carried by only a subset of tumour cells can be used as a marker of the existence of the subpopulation, as these mutations will appear with similar CCF values. The raw CCF values are therefore clustered to infer subclones.

1.11.1 Clustering of mutations

Figure 1.1 illustrates how this can be done: (A) During cancer evolution, a tumour acquires driver mutations (marked with a plus sign) that can initiate clonal expansions. (B) Over time, a number of these clonal expansions can occur, resulting in the increase of subpopulations of cells harbouring distinct sets of mutations. Tumour samples typically consist of a mixture of tumour cells with mutations (solid lines) and normal cells without mutations (dashed lines).

(C) Some mutations are carried by all tumour cells (marked with a square), whereas others are present in a subset of cells (triangle and circle). Using allele frequencies of mutations obtained from sequencing data and accounting for copy number aberrations, an estimate of the fraction of tumour cells carrying each mutation can be obtained. A set of mutations can then be used as a marker for a population of cells, allowing estimation of the fraction of tumour cells of the corresponding subclone. Clustering algorithms can be applied to obtain the cancer cell fractions (CCFs) of each subclone. (D and E) The relationship between subclones can be visualized as a tree. (D) Some methods perform this clustering in fraction-of-tumour-cells space, and (E) others in the space of fraction of all cells.

1.11.2 Statistical and computational strategies for subclonal reconstruction

Many subclonal inference methods are based on a Dirichlet Process (effectively a distribution of statistical distributions with properties to estimate the composition), including PyClone (Roth et al., 2014), PhyloSub (Jiao et al., 2014) and PhyloWGS (Deshwar et al., 2015). These methods require Markov chain Monte Carlo (MCMC) during their estimation process, which is computationally heavy.

Alternatively, one can model the data as a mixture of distributions and use variational Bayesian methods to estimate the composition (SciClone, which requires specification of the number of clusters (Miller et al., 2014)). CloneHD is based on a hidden Markov model and can couple SNV and CNA data to perform subclonal reconstruction (Fischer et al., 2014).

The method that I have worked on, DPClust, is also based on a Dirichlet Process. The method is explained in Chapter 2 and is shown to be amongst the best performing methods in a comparison in Chapter 6.

1.12 Copy number calling

To calculate CCF values for SNVs, one must take into account copy number alterations (this will be explained in Chapter 2). Copy number calling is therefore an important part of the subclonal reconstruction pipeline.

Copy number calling consists of two major components: estimating the sample purity (the proportion of tumour cells in the sample) and ploidy (the average number of chromosome copies per tumour cell), and obtaining copy number states for each genomic segment. Callers predominantly rely on the logR and BAF measures. The logR is a quantification of the amount of DNA that is available (i.e. total copy number). The BAF of SNPs that are heterozygous in the germline of the sample donor (identified from the matched normal sample) can be used to quantify the contributions of the maternal and paternal allele to the total copy number.

Copy number callers can call subclonal copy number (Carter et al., 2012; Fischer et al., 2014; Kleinheinz et al., 2017; Nik-Zainal et al., 2012a), of which the Battenberg algorithm (Nik-Zainal et al., 2012a) is presented in the next chapter. Calling subclonal copy number requires very precise BAF estimates as small deviations from a clonal state are used to detect alterations. Many of these methods therefore perform haplotype reconstruction to order SNPs correctly and improve the accuracy of the BAF estimate.

The principles outlined above are implemented in the Battenberg algorithm (Nik-Zainal et al., 2012a). Other BAF-based methods apply similar metrics to detect deviation from

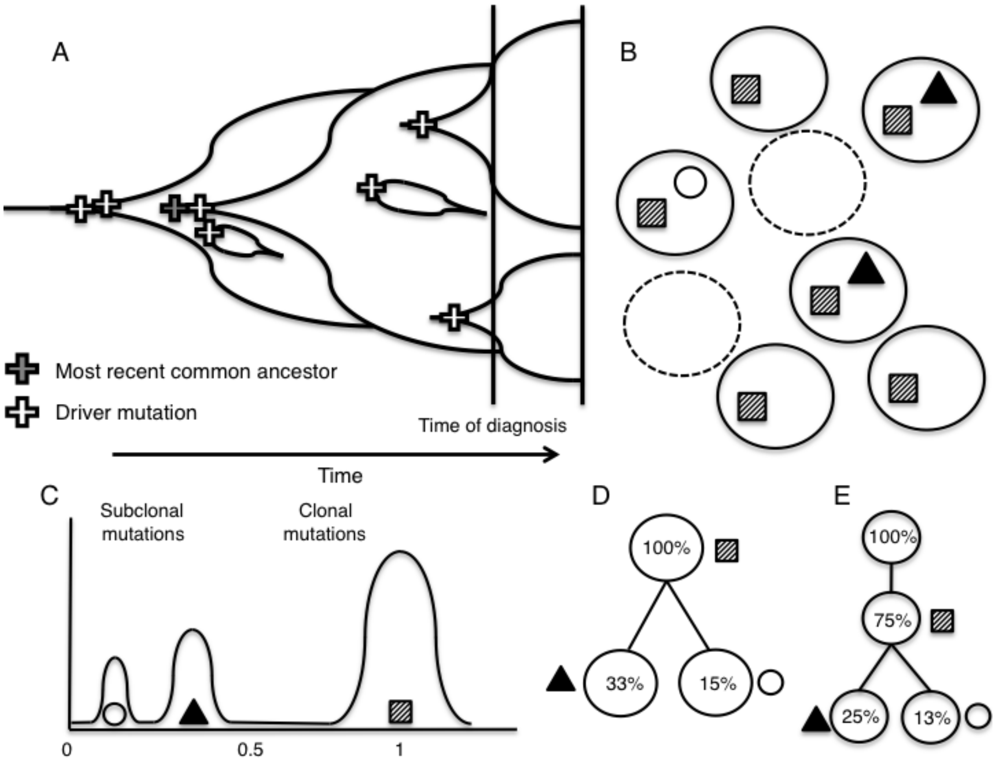

Fig. 1.1 (A) During their lifetime, tumours acquire mutations, of which drivers can lead to clonal expansions. (B) At any time, a tumour consists of multiple populations of cells, tumour cells (circles with continuous line) and infiltrating normal cells (dashed circles). As mutations are acquired gradually, some mutations will be carried by all tumour cells (marked by a square), whilst other mutations are only available in a subset of cells (marked by a triangle and circle). These subclonal mutations serve as a marker of the presence of subclonal cellular populations when they are measured via massively parallel sequencing. (C) By adjusting the measured allele frequency of each mutation for local copy number alterations and the tumour purity one can estimate the fraction of tumour cells that carry each mutation, of which a density is shown in this panel. The clonal mutations marked by a square in panel (B) will appear at approximately 1 (100% of tumour cells), while subclonal mutations appear at values smaller than 1. Mutations can be clustered in this fraction of tumour cell space to estimate the presence of subclonal populations. (D and E) The relationship between obtained mutation clusters can be visualized as a tree, in CCF space (D) or cellular prevalence (CP) space (E).

clonal copy number. There are two different approaches to establish these values: event-based or population-based. Event-based callers, such as the Battenberg algorithm, aim to establish these values for each segment individually (Carter et al., 2012; Nik-Zainal et al., 2012a), while population-based callers aim to explain as many segments as possible with a single subclonal fraction (Fischer et al., 2014; Ha et al., 2014).

It is also possible to estimate total copy number from read depth alone by binning reads across the genome and comparing the relative differences between bins with a matched normal sample. The advantage of methods such as Battenberg that rely heavily on BAF values is that allele frequencies are less affected by various biases that affect read depth (such as wave bias related to GC content and/or replication timing (Diskin et al., 2008; Koren et al., 2012)), as these biases affect both alleles equally and will therefore be cancelled out in the BAF calculation.

1.13 Tumour micro-environment

A tumour consists of a mixture of cancer and non-cancer cells, and with recent high throughput measurements show that the mixture of cell types (the tumour micro-environment, TME) plays an active role in shaping the tumour from neoplasm to advanced disease (Hanahan and Coussens, 2012).

1.13.1 Carcinoma-associated fibroblasts

Fibroblasts can be permanently activated to support a growing tumour, where the cancer can be thought of as a wound that does not heal (Wang et al., 2017). Typically, fibroblasts are deactivated when a tissue lesion is repaired, however, when fibroblasts remain active (known as carcinoma-associated fibroblasts (CAFs), or myofibroblasts) they can impact a growing tumour. CAFs alter the extra cellular matrix (ECM), communicate with epithelial, endothelial and immune cells by secreting growth factors (Kalluri and Zeisberg, 2006) and can induce epithelial-mesenchymal transition (EMT) (Erez et al., 2010), enhance vascularisation and promote inflammation (Orimo et al., 2005).

1.13.2 Tumour-associated macrophages

Macrophages can be recruited into a tumour supporting role, promoting angiogenesis, cell migration, tumour cell intravasation and metastasis (Condeelis and Pollard, 2006). These tumour-associated macrophages (TAMs) are typically abundant and are thought to contribute to tumour evolution from neoplasia to invasive disease (Qian and Pollard, 2010). TAMs are relevant for treatment choices and response: their abundance is associated with poor prognosis (Bingle et al., 2002) and they are sensitive to checkpoint blockade immunotherapies (Mantovani et al., 2017).

1.13.3 Tumour-infiltrating lymphocytes

Many types T and B cells can be found within the TME (T cells) and at the tumour margin and adjacent lymph nodes (B cells) (Balkwill et al., 2012). Both T and B cells can have a positive or negative effect on prognosis: for example, CD4+ T cells that produce cytokines interleukin-2 (IL-2) and interferon gamma (IFN-) are associated with good prognosis, but CD4+ cells that produce IL-4, IL-5 and IL-13 are thought to promote tumour growth (Fridman et al., 2012). Infiltrating B cells are generally thought to exhibit a positive effect on tumour prognosis (Wouters and Nelson, 2018), sharp contrast exists however: B cells are a survival benefit for HER2-positive and triple negative breast cancer, but have an adverse effect on HER2-negative breast cancers (Denkert et al., 2018). T cells are a major target for immunotherapy by blocking cytotoxic T lymphocyte–associated protein 4 (CTLA-4) or programmed cell death 1 (PD-1) expression, however treatment leads to resistance in approximately one-in-three patients (Ribas and Wolchok, 2018; Sharma et al., 2017), leading to calls for combining targeted and immune-based therapies (Gotwals et al., 2017).

1.13.4 Tumour-associated neutrophils

Neutrophils play an important role in tumour initiation, growth, proliferation, angiogenesis, suppression of antitumour immunity (Coffelt et al., 2016) and metastasis establishment (Wculek and Malanchi, 2015), and can exert pro- and anti-tumour functions (Galdiero et al., 2013). A high neutrophil count has been shown to correlate with poor prognosis (Coffelt et al., 2016), while a decline in neutrophils-to-lymphocytes ratio has been associated with improved outcomes (Templeton et al., 2016).

1.13.5 Other cell types

The TME is host to a number of additional cell types that influence evasion of immune destruction (NK cells), angiogenesis (myeloid-derived suppressor, dendritic and vascular endothelial cells), cell death resistance (adipocytes) and invasion and metastasis (pericytes) (Balkwill et al., 2012; Hanahan and Coussens, 2012; Joyce and Fearon, 2015).

1.13.6 Immune evasion

As a tumour grows, somatic mutations in tumour cells may introduce newly formed antigens that could trigger a response from the immune system via immune cells present in the TME (Schumacher and Schreiber, 2015). Tumours have been reported with evidence of, and have shown signs of negative selection against neoantigens: through point mutations (Rizvi et al., 2015; Robbins et al., 2013), copy number loss (McGranahan et al., 2017) and promotor hypermethylation (Rosenthal et al., 2019).

By separating these events into clonal and subclonal it is possible to observe selection against neoantigens. Clonal analysis of neoantigens in lung and skin cancers suggested that tumours with a high clonal neoantigen burden and low ITH have a longer disease-free survival (McGranahan et al., 2016). And a recent study suggests that the immune microenvironment actively shapes evolution of lung cancers (Rosenthal et al., 2019): Untreated tumours with low tumour infiltrating lymphocytes (TIL) showed signs of earlier immune editing or copy number loss of antigens that were previously carried by all tumour cells, while tumours with high TIL contained evidence of continued editing and repression of neoantigens.

These findings highlight the importance of the tumour micro-environment for patient care, as tumours treated with immunotherapy showed a better response when a high clonal neoantigen burden was observed (McGranahan et al., 2016).

1.14 Clinical implications of heterogeneity

The realisation that a tumour is an ecosystem with its own unique properties has led to the idea of prescribing treatment specifically based on a tumour's characteristics, also known as targeted therapy (Sawyers, 2004). These prescription of a targeted therapy based on genetic profiling of the tumour has been shown to improve prognosis, for example for patients with difficult to treat metastatic lung adenocarcinoma (Kris et al., 2014) and unresectable metastatic gastrointestinal stromal tumours expressing *KIT* (Blanke et al., 2008).

A higher amount of heterogeneity is associated with poorer prognosis (Brioli et al., 2014; Gerlinger et al., 2012; Jamal-Hanjani et al., 2017; Marusyk et al., 2012; Turner and Reis-Filho, 2012). For example, Jamal-Hanjani et al. (2017) reported a 4.9 hazard ratio for recurrence or death for patients with a high rate of subclonal CNAs, compared to those with a low rate. However, current targeted therapy approaches do not take into account whether the targeted event is clonal or subclonal (McGranahan and Swanton, 2015b), leaving considerable room for improvement, as a therapy targeting a subclonal mutation will not target all tumour cells.

Despite the successful application of targeted therapies, tumours can quickly develop resistance (McGranahan and Swanton, 2015b; Misale et al., 2014; Russo et al., 2016), which typically occurs within 1-2 years (Dagogo-Jack and Shaw, 2018). Mechanisms via which resistance can arise include pre-existing or *de novo* mutations (Gainor et al., 2016; Jr et al., 2012; Kwak et al., 2015; Sequist et al., 2011; Wagle et al., 2011), switching to alternative pathways (Zhang et al., 2012) or change in cell lineage (Sequist et al., 2011). In light of the ease at which resistance occurs, there are several efforts to develop combination- or serial-therapies with the aim to overcome resistance to a single drug (Bozic et al., 2013; Duncan et al., 2012; Sharma and Allison, 2015; Szerlip et al., 2012).

It is currently unclear how often mechanisms of resistance are already present in low proportions of cells within heterogeneous tumours. A recent study reported a comparison between a single sample biopsy of a primary tumour and tumour DNA extracted from a blood sample (also known as circulating tumour DNA (Mattos-Arruda et al., 2013) or cell-free tumour DNA (ctDNA)) at the same time-point (Parikh et al., 2019). The study consisted of 42 cases of gastrointestinal adenocarcinoma which were enrolled in a targeted therapy programme and showed signs of disease progression. The authors found that 76% of the cases showed evidence of at least one active treatment resistance mechanism in obtained ctDNA at disease progression, while multiple resistance mechanisms were identified in 17 cases (40% of all patients) (Parikh et al., 2019). These findings require confirmation in a larger cohort spanning more tissue and cancer types, however it suggests understanding intra-tumour heterogeneity is crucial to understand treatment effectiveness and is key to developing successful targeted therapies.

1.15 Summary

From this brief review of the relevant literature it becomes apparent that intra-tumour heterogeneity is an important component of tumour evolution, with clinical implications. Throughout the life-time of a tumour, mutational processes generate mutations throughout the genomes of cancer cells. By chance such a process can generate a driver mutation that initiates a clonal expansion, also increasing the cellular frequency of the passenger mutations that occurred in the cell with the new driver. Concurrently, the micro-environment co-evolves and allows the tumour to expand. Massively parallel sequencing allows for detection of the somatic mutations and of copy number alterations in tumour cells, and therefore provides access to the life history of a tumour. Careful curation of subclonal architectures and life histories across cancers can shed light on the pan-cancer landscape of ITH and on general characteristics by which tumours develop.

To this end, in the next chapter I will provide an in-depth description of the algorithms that I have maintained and developed further during my Ph.D. One algorithm for estimating somatic copy number alterations (Battenberg, first used in Nik-Zainal et al. (2012a)) and one for inferring the subclonal architecture of a cancer (DPClust, first used in Bolli et al. (2014)). Chapter 3 contains an extensive validation of the methods on simulated data, while in Chapter 4 I will explain a thorough QC procedure for copy number and subclonal architecture calls.

In Chapter 5 I apply the methods to a single tumour to illustrate what can be learned about the life history of a cancer from its genome. Chapter 6 contains further computational methods for a pan-cancer analysis of ITH, while in Chapter 7 I describe the results of applying those methods to 2,778 cancer genomes. Finally, Chapter 8 contains the discussion.