A Genomic Study of the Evolution of Paediatric Solid Tumours

George David Cresswell

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

The Francis Crick Institute

University College London

February 26, 2018

ii

I, George David Cresswell, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Paediatric cancers evolve very differently to adult cancers: they develop over months, not decades, and are exposed to few exogenous mutagens. Despite major strides in our understanding of adult cancers, little is known about the evolutionary trajectories of paediatric solid cancers. This thesis focuses on the commonest paediatric malignancies of the kidney and liver to determine their genetic past.

Constructing the evolutionary history of a cancer requires estimating the number of cells possessing each mutation in a tumour and determining their ordering. In the first part of the thesis, I compare copy number alterations (CNAs) in multiple tumour regions to construct the histories of 20 Wilms' tumours (WTs). I uncover patterns of CNA occurrence, describe their role in the different stages of tumorigenesis, and relate phylogeny to clinically important features.

In the second part, I develop a method to estimate cellular proportion of CNAs by modelling heterozygous single nucleotide polymorphisms from array data as a Gaussian mixture. By determining the composition of spatially separate and locally mixing clones, I perform a comprehensive reconstruction of the evolutionary histories of >70 WTs and 11 hepatoblastomas.

Supporting my analysis using targeted sequencing of ~180 genes, I show that convergent evolution of CNAs and single nucleotide variation highlights context-dependent selection. I discover subtype-specific features of WT evolution, including a consistent trajectory of stromal WT evolution. Finally, I show CNA instability

as a potential biomarker in WT, suggesting that burden of CNAs may identify high-risk patients.

In summary, this thesis describes the diversity of evolution in two paediatric solid cancers and novel approaches to studying these diseases. By uncovering evidence for how these cancers evolve over time, I gain insights into the sequential genetic changes that form the histories of these cancers and demonstrate the clinical relevance both in terms of sub-type diagnosis and prognosis.

Impact Statement

The importance and complexity of cancer evolution for modern medicine has been demonstrated profoundly in the last decade. Despite the fact that cancer researchers have accepted that tumours evolve through an evolutionary process since the mid-20th century, contemporary genome-wide assays are now enabling scientists to model tumour evolution in increasingly advanced ways. Evolution as a system is adept at allowing for the proliferation of a species in harsh conditions exemplified by extremophiles in nature. Evolution in cancer is integral in causing metastasis, relapse and therefore death. The implications of tumour evolution are pervasive amongst all cancer types including childhood cancers.

This thesis explores cancer evolution in the setting of cancers that occur early in life. Cancer is a major cause of death in childhood, however survival rates have increased dramatically since the advent of chemotherapy. Unfortunately, childhood cancer survivors subsequently face a decreased quality of life into adulthood and reduced life expectancy due to the long-term effects of cytotoxic therapies such as chemotherapy and radiotherapy.

Clinical trials that search for genetic biomarkers for patient stratification may provide an opportunity to select patients for optimised therapeutic programmes to balance survival and long-term life quality. However, clinical trials should also be viewed in the context of tumour evolution. Tumour evolution produces diversity and diversity affects biomarker detection. By understanding tumour evolution, future trials can be designed to take into account the likelihood of detecting genetic

biomarkers. Conversely, alternative sampling strategies may be utilised to overcome this challenge, for example by assaying tumour DNA present in the blood instead of sampling the tumour directly.

In this thesis I explore the idea of using traits of tumour evolution itself to stratify patients. Redefining our idea of a genetic biomarker, from a single mutation, to the relationships between mutations may prove to be a powerful improvement for our perception of cancer prognosis. Studying evolution in cancer remains expensive and challenging; findings in this thesis may help to intelligently design large tumour evolution studies.

Understanding tumour evolution can also aid decision making in the context of routine clinical problems. Are synchronous tumours that occur in separate organs related or independent? What drives strikingly different phenotypes within the same tumour mass? Medical practitioners are faced with these issues on a regular basis. Work in this thesis seeks to communicate with clinicians ways in which evolution can help answer these questions to shape the perception of a case in light of the underlying biology.

This research contributes to the growing repertoire of tumour evolution studies. Together many groups across the world are adding detail to the concept of tumour evolution. These findings often show why cancer is prone to resistance and is such a difficult disease to treat. By supplementing this school of thought this thesis should contribute to the education of the general public. The more aware that cancer patients are of how cancers grow and survive, the closer the relationship will be between patient and doctor, as well as the public and the researcher.

Preface

This thesis describes research I conducted during my PhD studentship at the Francis Crick Institute (formerly CRUK London Research Institute), London, UK, from September 2013 to September 2017 under the supervision of Prof. Nicholas Luscombe.

Work presented in Chapters 3 and 5 was conducted in collaboration with Prof. Kathy Pritchard-Jones' group at the UCL Great Ormond Street Institute of Child Health. The project was co-led by William Mifsud. Generally, I performed all analysis of the data presented in this thesis, however I generated no data. Data in these chapters was generated by John Apps, Tasnim Chagtai, Mariana Maschietto, Øystein E. Olsen and Richard Williams. Single nucleotide polymorphism array data presented was generated by UCL Genomics. Targeted sequencing data generation was performed by Reem Al-Saadi, Taryn Treger and the UCL Pathogen Genomics Unit. Circulating tumour DNA extraction and digital droplet polymerase chain reaction experiments were performed by Taryn Treger. Work presented in Chapter 6 was performed in collaboration with Prof. John Anderson and Tessa Kasia.

Research presented in Chapter 3 has been published in the following article: George D. Cresswell, John R. Apps, Tasnim Chagtai, Borbala Mifsud, Christopher C. Bentley, Mariana Maschietto, Sergey D. Popov, Mark E. Weeks, Øystein E. Olsen, Neil J. Sebire, Kathy Pritchard-Jones, Nicholas M. Luscombe, Richard D. Williams, William Mifsud. Intra-Tumor Genetic Heterogeneity in Wilms Tumor: Clonal Evolution and Clinical Implications. EBioMedicine 9, 120–129 (2016).

Acknowledgements

"Over thinking, over analysing separates the body from the mind."

Lateralus, Tool

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Cancer in Early Life

Cancer is a disease in which phenotypically normal cells in the body transform by acquiring the ability to proliferate uncontrollably, overcoming normal biological barriers of unlimited cell growth and forming malignant neoplasms. In order for a cancer to acquire the hallmarks of malignancy, key genes in a cell of origin must acquire mutations. The disease is a major cause of death across the world and one in two people born after 1960 will be diagnosed with cancer during their lifetime [1].

Cancer can arise in any organ of the body, however cancer incidence per tissue is underpinned by the innate proliferation rate of the cell, its relative exposure to mutagens, as well as other underlying biological and genetic factors [2]. Some of the most common tissues for cancer development include lung, breast, prostate and colon [1]. Lung cancer is a common example of a tumour type linked to environmental exposures as it is strongly associated with tobacco smoking which leads to long-term mutagenic insults to the genetic material of cells [3].

The ability of smokers to live for decades cancer-free demonstrates the ability of the body to prevent the emergence of cancer through DNA damage repair, cell signalling, controlled cell death and immune surveillance [4, 5]. In this regard, the delay in the development of cancer over time illustrates why it is considered a

disease of age as well as a disease of genetic aberration. Cancer incidence increases dramatically with age and in the United Kingdom approximately half of cancers are diagnosed in individuals older than 70 years [1].

Despite this, cancer can also present in childhood. Cancer contributes significantly to childhood mortality, causing 1 in 5 deaths in 1–14 year olds, and is increasing in incidence [6, 7]. Genetic predispositions are an important factor for the development of cancer in the early stages of life, but only explain a subset of childhood cancers [8]. In the 1960s childhood cancer survival rates were much lower compared to modern standards, as more than 50% of patients did not survive past 5 years [9]. Since then, through the rational design of clinical trials centred around the use of chemotherapy and radiotherapy in children, survival rates have achieved ~80%. Unfortunately, this success has come at a cost. The use of chemotherapy and radiotherapy has severe long-term implications when used in young children, leading to a reduced quality of life and an increased chance of death decades later. For example, the use of anthracyclines is known to cause cardiotoxicity that can start in adulthood [10]. Additionally, more than 10% of survivors of childhood cancer treated with anthracyclines in combination with radiotherapy experience heart failure before the age of 45 [11]. These effects, in addition to the contribution of other cytotoxic anti-cancer agents, contribute to the fact that 18% of individuals that were diagnosed with a childhood cancer in the 1970s/1980s and achieved five-year survival, died within the subsequent quarter of a century.

Childhood cancers highlight a separate spectrum of affected tissues types compared with adult cancer. These are broadly split into 12 categories of both liquid and solid tumours [7, 12]. The 11 defined categories are leukaemia, lymphoma, central nervous system tumours, sympathetic nervous system tumours, retinoblastoma, renal tumours, hepatic tumours, bone tumours, soft tissue sarcomas, germ cell and gonadal tumours and, epithelial tumours and melanoma. Most adult cancers are carcinomas as most tissues with high proliferation rates and carcinogen

exposure in adults are epithelial cells. Contrasting, in children many other histological types are common, perhaps due to many childhood cancers being caused by aberrations in developmental processes that are largely concluded in adults [12, 13].

Cancer incidence rates in 0–4 year olds are higher than in later childhood years until 14 years old [1]. As a proportion of the total incidence, renal and hepatic tumours contribute more to the 0–4 year old age bracket than older age brackets, meaning they are significant contributors to cancer in childhood in early years [7]. This thesis focuses on studying the two most common tumour types in the kidney and liver, Wilms' tumours (nephroblastomas) and hepatoblastomas, as representative of solid childhood tumours.

1.2   Wilms' tumour biology

Approximately 90% of all paediatric renal tumours are Wilms' tumours (WTs). These tumours make up ~6% of all childhood tumours and 1 in every 100,000 child under 15 years of age will develop a WT annually (95–140 cases per year in the UK). The median age of diagnosis is ~3 years of age [14]. WTs are thought to develop as a result of aberrant mesenchymal-to-epithelial transition during the differentiation of the ureteric bud to form the mature kidney [15]. Interestingly, the opposite process, epithelial-to-mesenchymal transition often occurs in adult carcinomas as a pathway towards tumour metastasis [16].

The ureteric bud, which invades the metanephric blastema, triggers the differentiation of stromal and epithelial cells to form the nephron structure of the mature kidney (Fig. 1.1A) [18]. When this differentiation is blocked or disrupted in WT development, these cell types are produced aberrantly [15]. This is thought to lead to the presentation of the 'classic' triphasic histology seen in WTs. Here, WTs are often seen to form tumours with variable compositions of the three cell types related to ureteric bud induced differentiation – blastema, stroma and epithelia (Fig. 1.1B–C). The cellular composition of a WT may provide clues of the natural his-
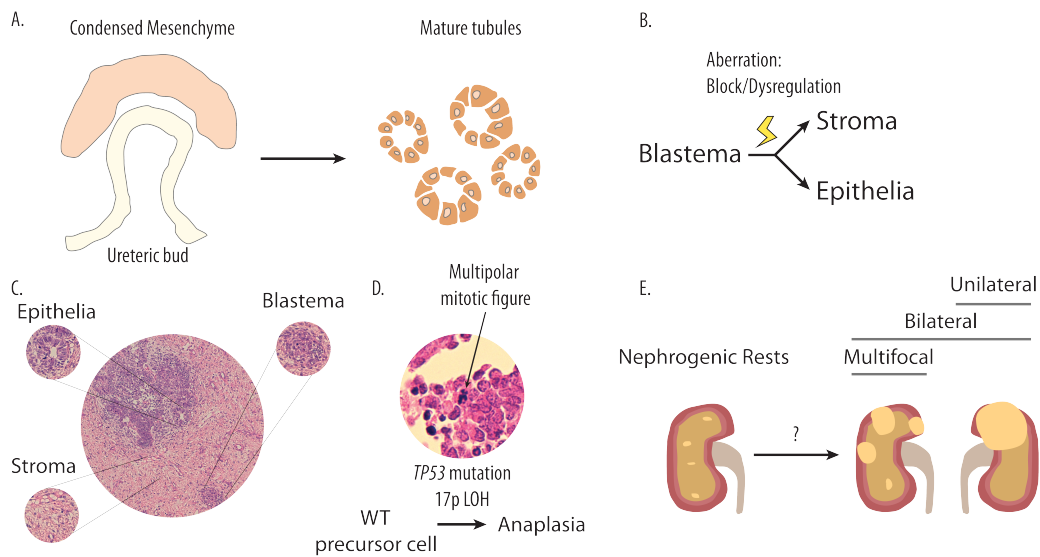
Figure 1.1: Understanding how a WT develops and its links to clinical presentation are vital to understanding how the cancer evolves. In normal kidney development, the ureteric bud invades the metanephric blastema to begin the process of differentiation that will form the mature tubules of the kidney (A – inspired by [17]). Normal kidney development involves the differentiation of blastema into stromal and epithelial components. A block/dysregulation of this differentiation process is linked to WT formation (B). This is thought to cause the development of triphasic WT histology involving variable components of blastema, stroma and epithelia. (C) represents the histology of a mixed type WT in which no single cell-type dominates (10X magnification). The three components of triphasic WT histology (blastema, epithelia and stroma) are enlarged in the image. An additional important feature of WT histology is the development of anaplasia which is associated with *TP53* mutation and 17p LOH (D). This image is 40X magnification and shows a multipolar mitotic figure alongside hyperchromatic nuclei. Finally, small nodules with the appearance of immature kidney, known as nephrogenic rests, are thought to be the precursor lesion to WT. When WTs present at diagnosis they can be unilateral, multifocal or bilateral (E).

tory of the tumour. Another histological feature also present in a subset of WTs is anaplasia, in which chaotic mitotic figures are observed often showing multiple poles in addition to cells with hyperchromatic nuclei (Fig. 1.1D) [19].

The precursor lesion of WT is thought to be nephrogenic rests (Fig. 1.1E). These are often observed as small nodules of immature kidney which are benign. These rests are often associated with genetic syndromes such as Beckwith-Wiedemann

syndrome that is associated with WT and hepatoblastoma development [20, 21].

According to International Society of Paediatric Oncology (SIOP) guidelines in Europe, risk stratification for WT is largely based on histological presentation [22]. WTs are categorised according to their variable cellular composition. If any given cell type compromises >66% of the tumour, the tumour is classified according to that cell type (e.g. blastemal, stromal or epithelial). If no one single cell type dominates the tumour, the tumour is classified as being 'mixed' type. Additionally the tumour is assessed for anaplasia. If anaplasia is observed in multiple loci the tumour is considered 'diffuse anaplastic' regardless of the triphasic composition. If anaplasia is observed in only a single location, it is considered 'focal anaplastic'. Tumours with >66% dead cells – necrosis – are considered to be regressive type tumours and if the tumour is >99% necrotic the tumour is considered a necrotic type tumour.

All patients that are treated according to SIOP guidelines undergo pre-surgery chemotherapy prior to removal of the tumour, often performed as a nephrectomy. The typing of the tumour post-nephrectomy determines further treatment. Diffuse anaplastic and blastemal type tumours are considered high risk due to their worse survival, due to their association with genomic instability and chemotherapy-resistant undifferentiated cells, respectively. Only necrotic type tumours are considered to be low risk. All other types are considered to be intermediate risk. Higher risk types receive extended post-surgery chemotherapy, whilst low risk tumours receive no post-surgery chemotherapy [22].

Approximately 15% of WTs patients will relapse within two years. More than half of WT cases that relapse are not high risk cases, clearly showing that there is a need for increased treatment in some cases [14]. However, WT patients are not exempt from the negative long term effects of a chemotherapy and/or radiotherapy programme of treatment in early life. Furthermore, half of high risk cases do not relapse but are treated with anthracycline which is linked to heart tissue damage

and heart failure decades later in life [11].

An additional form of high risk is associated with tumours which present primary lesions in both kidneys upon diagnosis – bilateral tumours. WTs can also be multifocal within a single kidney (Fig. 1.1E). Bilateral WTs are associated with syndromic disorders which can predispose to WT formation [23]. These malformation syndromes, in which patients present with constitutional molecular defects, are present in up to ~17% of WT cases, despite only around 1–2% of WT cases being familial, indicating that defects are acquired post-fertilisation [24, 25].

1.3   Hepatoblastoma tumour biology

Hepatoblastoma is the most common hepatic tumour in children, however it is very rare as only 1–1.5 children per 1,000,000 are diagnosed with it annually (10–25 cases per year in the UK) [26, 27]. It is described as an 'embryonal' tumour and despite its aetiology remaining relatively unknown, it is thought to arise from aberrant liver development [28]. Support for this concept comes from its histological presentation as cells in hepatoblastoma resemble immature liver cells in the different stages of liver development [29].

The histology of hepatoblastomas is very heterogeneous both within a tumour and between tumours. Hepatoblastomas show variable amounts of epithelial and stromal components. The epithelial components show variability within themselves as these components can present as embryonal or fetal epithelia. A mixed presentation of both epithelia and stromal is most common in hepatoblastoma (~44%). Purely fetal epithelial type compromise 31% of cases and purely embryonal make up 19% of cases [28].

Tumour treatment largely depends upon the Pretreatment Extent of Disease (PRETEXT) protocol in which risk depends on how many sections of the liver the tumour has invaded [30]. Survival rates have doubled since the use of chemother-

apy in hepatoblastomas to ~80% [31]. The incidence of hepatoblastoma is also reportedly increasing [26]. Treatment of high risk hepatoblastoma involves the use of an anthracycline which damages heart tissue [11, 28]. Hepatoblastoma is also associated with syndromes such as Beckwith-Wiedermann syndrome and familial adenomatous polyposis and syndromic patients make up ~15% of all hepatoblastoma patients [28, 31].

1.4   Genetic mutation in childhood cancer

1.4.1   Small sequence mutation in childhood cancers

Small sequence mutations are a type of genetic alteration acquired in carcinogenesis that can provide a strong phenotypic change to drive cancer development and are one of the best characterised types of mutation that are present in cancer. Generally speaking small sequence mutation are made up of three different types of alterations (1) point mutations, in which single bases are mutated to an alternate base, (2) insertions, in which short sequence are acquired at a locus in the genome and (3) deletions in which short sequences are lost. The latter two types of mutations are generally grouped together as 'indels'.

If these mutations occur in the exon of a gene they can potentially cause changes to the codon sequence that allow for transfer RNAs possessing alternate amino acids to be incorporated during protein synthesis and the protein sequence of the mutated gene to be permanently changed on a single allele. In the context of point mutations, changes to the genetic sequence that affect the amino acid sequence are considered non-synonymous mutations. These come in two forms, missense mutations which alter a single amino acid in a protein sequence and nonsense mutations which cause a premature stop codon to be incorporated in the protein sequence, causing the early termination of the synthesis of the protein.

Missense mutations can cause a gain-of-function by altering the chemical structure of key protein motifs. One of the most common examples of this in cancer is

the alteration of a valine to a glutamic acid (V600E) in the protein product of the gene *BRAF*. BRAF propagates growth signals in the cell and this amino acid change allows the protein to become constitutively active which leads to increased proliferation rate and is associated with skin and large intestine malignancies [32]. Genes like *BRAF* which often gain function when mutated in cancer are known as 'oncogenes'. Other mutations may cause a loss-of-function – the removal of a protein's function or the removal of a stable version of the protein. This may be caused by a missense mutation and it is nearly always the consequence of a nonsense mutation. Genes which often exhibit loss-of-function mutations are considered tumour suppressors. An example of a tumour suppressor gene is *RB1* that encodes for a protein product that controls cell cycle progression [33]. Single nucleotide point mutations that do not cause a change in amino acid sequence, due to codon redundancy, are known as synonymous or silent mutations, often leaving the protein unaffected.

Indels always alter the protein coding sequence if they are present in the gene exon. If the size of the insertion/deletion is not divisible by three, the mutation causes a frameshift. Here for the entire length of the gene, downstream of the mutation, the amino acid sequence may be altered. Frameshifts often lead to a loss-of-function, as the new protein sequence is unlikely to be biologically functional. If the size of the indel is divisible by three the mutation causes an 'in-frame' insertion/deletion. Here a number of amino acids are either removed or inserted in an amino acid sequence, whilst the majority of the sequence remains unaffected.

In WT, recurrent small sequence mutations are known to affect several genes. These include *WT1*, a gene that encodes for a transcription factor that controls cell proliferation and is involved in kidney development [15]. Germline *WT1* mutations are associated with predisposition to WT development and *WT1* is also frequently mutated somatically. Mutations in *AMER1* and *CTNNB1* are also common in WT. These three genes were the earliest known recurrently mutated genes and account for approximately one third of all WTs [34]. *AMER1*, also known as *WTX*, ap-

pears to be involved in WNT signalling, potentially promoting the degradation of β-catenin [35]. β-catenin, encoded by *CTNNB1*, transmits signalling induced by the WNT ligands to the nucleus of the cell to alter cellular transcription and is associated with several oncogenic processes such as growth and differentiation. The protein is often mutated at sites of phosphorylatable amino acids (e.g. serines and threonine) to prevent phosphorylation which would normally target the protein for degradation, therefore allowing WNT signalling to propagate [36]. The tumour suppressor p53, that is encoded by *TP53* and conducts the response of the cell to stresses such as DNA damage, is mutated in at least half of diffuse anaplastic WTs [37, 38, 39].

Large exome sequencing studies of WT have highlighted less recurrent mutations in micro RNA (miRNA) processing pathways and SIX1/SIX2 signalling genes, which are important developmental regulators [40, 41]. Genes such as *DICER1*, *DGCR8* and *DROSHA* all play key roles in the maturation of miRNAs [42]. *SIX1* and *SIX2* are frequently mutated at the hotspot Q177R in the homeobox domain in blastemal type WTs which alters its DNA-binding potential [40]. Recurrent mutations have also been highlighted in *MLLT1* in the YEATS domain. The protein product of *MLLT1*, ENL, is involved in transcriptional elongation in early development [43]. *MYCN*, an oncogene discovered in neuroblastoma, has also found to be mutated in WTs [40, 44].

Largely owing to its rarity, very few recurrent mutations have been discovered in hepatoblastomas. The clearest recurrent mutation in this cancer is in *CTNNB1*. Other sequencing projects have highlighted other potential candidates such as *NFE2L2* which implicates the activation of the NFE2L2-KEAP1 pathway, which protects cells against oxidative damage, in hepatoblastoma tumorigenesis [29, 45].

1.4.2 Copy number alterations and childhood cancers

Most cells in the human body are diploid and possess two copies of each gene in the human genome, one copy from each parent. Normal cells are capable of replicating

this diploid genome accurately to then segregate the chromosomes evenly into two daughter cells to allow for cell replication and proliferation through mitosis. A key hallmark of cancer is the loss of genomic integrity. This leads to the genomes of cancer cells becoming unstable, possibly owing to defective DNA repair machinery or chromosome segregation in combination with evasion of regulated cell death [5, 46]. In this situation, replication of the genome can provide an increased risk of improper segregation of chromosomes to daughter cells that may be propagated and tolerated in the progeny. The somatic inheritance of additional or fewer copies in daughter cells changes the copy number of sections of the genome therefore also alters the number of copies of each gene in the affected region. Altering gene copy number can affect gene expression levels in a positively correlated manner [47].

Copy number states of genomic regions may increase or decrease and are referred to as gains and losses respectively. Copy number gains are linked to increased expression and by contrast copy number losses are associated with reduced gene expression. As copy number changes can affect either the maternal or paternal chromosome these events can be specific to particular alleles, which has implications to both expression levels of transcripts and the expression of protein variations which may arise as a consequence of allele-specific constitutional or somatic nucleotide variants. A deletion of the remaining wild-type copy of a gene to produce loss of heterozygosity (LOH) is often observed following a somatic mutation as a 'second hit' to the gene locus and may be required to produce an oncogenic effect [48].

Copy number alterations (CNAs) can be of any size, affecting both whole chromosomes and chromosome arms, but also being focal, affecting a small group of genes. In large chromosome CNAs it can be difficult to pinpoint the important genes which are being altered by the variation. It is often only possible to hypothesise as to the importance of the large CNA if it affects a region which contains a gene of known importance. An example of this may be the loss of the p-arm of chromosome 17 that contains a wild-type copy of *TP53* in a situation in which

the other copy is mutated. Focal CNAs are easier to interpret biologically in terms of their consequences as they affect much fewer genes. Focal changes are also more likely to be tolerated in extreme copy number states such as complete loss (a loss of both alleles) or amplification to high copy number states. Amplifications of the *MYC* gene family, transcription factors that regulate cell proliferation, often produce high copy number states in several cancer types for example, *MYCN* in neuroblastoma, owing to their large oncogenic potential [44, 49].

Chromosomal structure can also be affected in a manner in which the total copy number state remains constant but one allele is lost and another is effectively duplicated. This is considered copy number neutral LOH (CNNLOH) and leads to the same copy number but a LOH. This is also called uniparental disomy (UPD), especially in contexts highlighting parental-specific allele expression. The overall effect of CNNLOH/UPD is often a subtle allele-specific alteration in expression levels leaving genes with no allele-specific expression unaltered.

The underlying mechanisms of genomic instability in cancer are broad and may act in many combinations. These include unrepaired double stranded breaks in DNA and chromosome missegregation among others. The process of mitotic recombination, in which chromosomes recombine in a somatic mitotic setting is thought to be the major mechanism for the formation of CNNLOH/UPD [50]. Additionally, breakage-fusion-bridge cycling, a process in which breakage followed by fusion can lead to a cyclic process which generates an increasingly complex chromosome structure, is considered a contributing factor to the generation of amplifications [51].

Somatic CNAs are informative of cancer biology and patient prognosis and there are several examples of informative CNAs in WT. Chromosome 11p UPD is a common event in WT. This event will almost always affect 11p15.5, the telomeric cytoband of the p-arm. *IGF2*, an important gene for regulating fetal growth, maps to this cytoband. This gene is imprinted and is therefore expressed in an allele-

specific manner. Here, the paternal chromosome expresses *IGF2* whilst the materal copy of the gene is repressed through DNA methylation. In WT, UPD of 11p15.5 often leads to the loss of the materal copy and the duplication of the paternal allele, leading to *IGF2* overexpression [52, 53]. As *WT1* is also present on chromosome 11p (11p13), a UPD event can often alter both *IGF2* and *WT1*, therefore having implications for both genes and demonstrating the context of mutation events that have horizontal relationships across the genome. Additionally, *MYCN* amplification in WT demonstrates the importance of *MYCN* as an oncogene in a subset of WTs [54]. Lastly, as mentioned previously, *TP53* mutations are common in diffuse anaplastic WTs. *TP53* is located on the p-arm of chromosome 17 and LOH of this chromosome arm is often indicative of the presence of a mutation in *TP53* as LOH removes the remaining wild-type copy of the gene [38].

CNAs are also prognostic in WTs. The observation of a combination of chromosome 1p loss and chromosome 16q loss leads to a WT patient being considered high-risk in the USA [55]. Chromosome 1q gain is also currently being trialled as a prognostic marker in WT and may prove to be of greater use to prognosis as it is more common owing to the fact it is one of the most prevalent CNAs in cancer [56, 57, 58, 59]. Finally, chromosome 4q and 14q losses are associated with high risk diffuse anaplastic WTs, indicating they may play an important role in this phenotype [54].

In hepatoblastomas, gains are frequently present in chromosome 1q, 2, 8 and 20 [60]. Additionally, chromosome losses are often observed in the p-arm of chromosome 1. Recurrent copy number gain of the cytoband 20q13.2 has been associated with poor prognosis in these tumours [61].

1.4.3 Ploidy states and childhood cancers

It is possible for cancer cells to possess genomes with higher or lower ploidy states than diploid (2n). The most commonly accepted mechanism for higher ploidy states to be achieved is through whole genome duplication (WGD). Here, through mis-

regulation of mitosis – perhaps through nondisjunction occurring during anaphase – one daughter cell may inherit a completely duplicated copy of the entire genome [62]. High ploidy states may also be achieved by multiple successive chromosomal duplications.

High ploidy states, such as tetraploidy (4n), are often observed in cancer, particularly in tumours that often present with high genomic instability [63]. A high ploidy state may allow cancer cells to achieve a more subtle set of CNA relationships and expression patterns. This may provide an evolutionary advantage to the tumour. In a related manner high ploidy states may also be more tolerant of subsequent genomic instability. Interestingly, polyploidy is a potential mechanism of rapid evolution in speciation [64].

WGDs have been timed as having occurred decades prior to the presentation of cancer at diagnosis. The timing of these events by analysing mutational types and their frequencies in cancer samples demonstrate that adult cancers have long life histories [65, 66]. However, despite being cancers with relatively short life histories, paediatric cancers such as WT and hepatoblastoma have also been shown to display high ploidy states such as tetraploidy [67, 68, 69].

1.4.4 Chromothripsis and childhood cancers

Another mechanism of genomic instability in cancer is the occurrence of a phenomenon known as 'chromothripsis'. Here, often within a relatively short region of a chromosome, many different copy number states are observed. These are thought to be produced in a single catastrophic event that causes the shattering of a region of the chromosome. This shattered region often involves the telomeric end of a chromosome indicating the shattering affects the end of a chromosome arm. This shattered chromosome is then aberrantly reassembled, propagating a chromosome with a chaotic new structure to daughter cells. Chromothripsis is a mechanism that causes extreme changes to the cancer genome in a single event [70].

It has been shown that *TP53* mutation is associated with chromothripsis, suggesting a role for p53 in protecting against chromothripsis events being propagated to daughter cells [71]. Chromothripsis has been observed in several paediatric tumours [72]. In WTs chromothripsis is strongly associated with both high ploidy states and *TP53* mutation. Additionally these tumours have poor outcomes and were associated with anaplasia [40]. Clearly, chromothripsis events are associated with high-risk WTs. Overall, extreme events such as chromothripsis can be seen as symptoms of genomic instability.

1.4.5 Burden of mutations across cancer types

Across the many different types of cancer, the burden of genomic mutation varies due to several factors including, but not limited to, age at diagnosis, environmental exposure and underlying DNA repair defects, which may be acquired somatically or inherited. The process of ageing and damaging effects of environmental mutagens leave characteristic 'mutational signatures' on the genomes of a cancer, allowing researchers to deconvolute the mutational processes that a cancer is subject to during its development [74]. The gradual accumulation over time of these mutational processes can burden the genome of a cancer with large numbers of somatic mutations.

It is vital to recognise that paediatric tumours, particularly those that mostly occur in children younger than 4 years of age, are not exposed to regular mutagens such as ultra-violet (UV) light or tobacco smoke to the same extent as adults. Additionally the ageing process is extremely nascent. In this regard the context of genomic mutation is largely different in these cancers. This is demonstrated by the fact that most paediatric cancers have far fewer SNVs in exome-wide sequencing studies (range of 2–13.5 median mutations, Fig. 1.2). This is in vast contrast to many adult cancers which can have 10-100 fold more non-synonymous point mutations and therefore have far more plastic genomes [29, 41, 73]. The almost complete removal of the effects of age and common carcinogens in paediatric cancers places them in a completely alternative context to the majority of adult cancers.
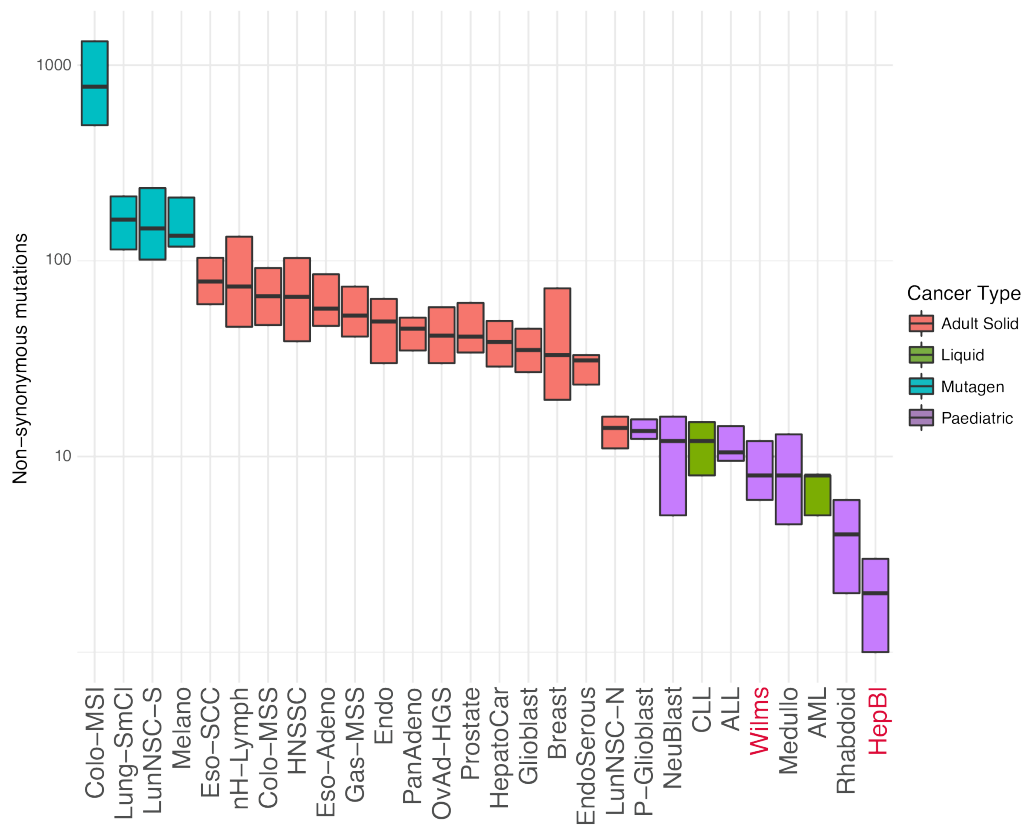
Figure 1.2: The number of non-synonymous point mutations varies in each tumour type. Box plots for each cancer type represent the median number of non-synonymous point mutations in studies of different cancer subtypes and the size of the box represents the interquartile range (IQR) of the distribution of point mutations reported. Generally cancer types with a mutagenic pheno-type (turquoise) such as colorectal cancers with microsatellite instability (Colo-MSI), small cell lung cancer (Lung-SmCl), non-small cell lung cancers in smokers (LunNSC-S) and melanomas (Melano) have a very large number of mutations (>100 non-synonymous mutations). Solid adult tumours (red) exhibit between 30-100 non-synonymous mutations across the exome. Paediatric cancers generally have fewer non-synonymous point mutations (purple) similar to liquid cancers (green). WTs (Wilms) and hepatoblastomas (HepBl) have a low median number of non-synonymous point mutations (8 and 2 respectively). Data is derived from Vogelstein et al, 2013 [73] and supplemented with data from Walz et al, 2015 [41] and Eichenmüller et al, 2014 [29].

1.5 Clonal evolution in cancer

1.5.1 The principles of clonal evolution

Cancers develop through a process of clonal evolution. Here a cell of origin begins the process of carcinogenesis by acquiring a somatic mutation which gives it a selective advantage over the surrounding normal cells. If this cell survives, the cell
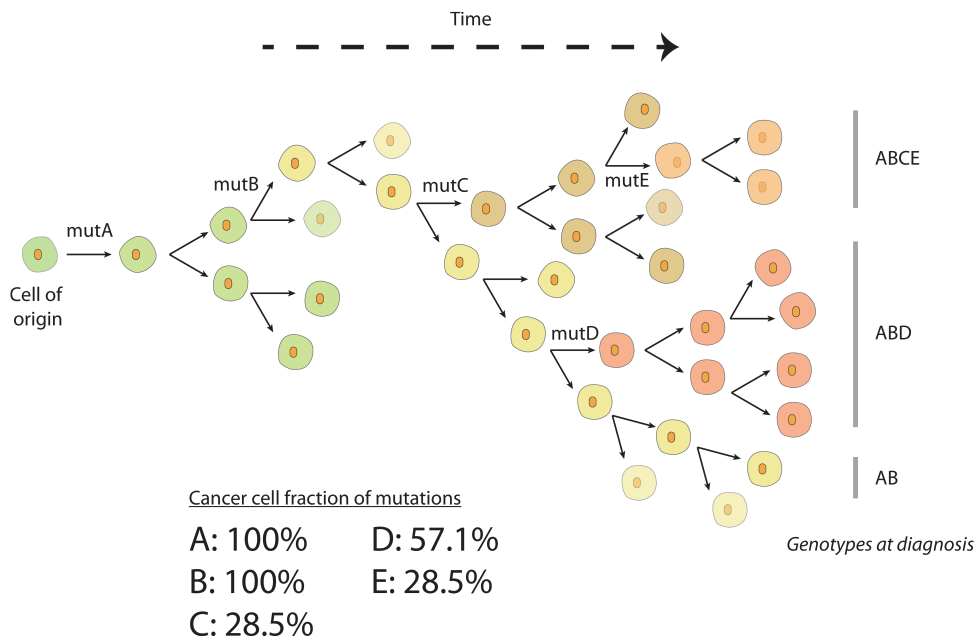
Figure 1.3: Cancer evolution as first conceptualised by Peter Nowell [75] dictates that cancers successively accumulate mutations in a process that can be branched and lead to the accumulation of genetically distinct subclones. A graphical simplification of the process is demonstrated and shows the accumulation of five mutations (mutA–E) across a branching population of cells from a cell of origin. The evolutionary history stipulates that three clones are present at diagnosis (AB, ABD and ABCE) and the abundance of each of the mutations is linked to the abundance of the clones and their ancestry.

will replicate to produce a population of cells sharing this advantageous mutation. This population is known as a clonal population as each cell in the population is in effect a clone of the cell of origin in that they all share this common genetic trait (Fig. 1.3) [75].

This process can continue to repeat itself as the clonal population can acquire additional mutations, which increase the fitness of its progeny further, potentially allowing the new clonal population to proliferate at a higher rate than the previous clonal population. This process is essentially branched, as cells replicate to produce two daughter cells that have an opportunity to produce progeny. Key mutations that provide a selective advantage are known as 'driver mutations' as they are thought to accelerate the process of tumorigenesis forward. However, other mutations are also

acquired that are largely considered to have little effect on the tumour but persist in the tumour population. These mutations are known as 'passenger' mutations as they 'hitchhike' a ride on the back of the effect of driver mutations and are far more frequent than driver mutations as the vast majority of possible mutations in the human genome will not produce an oncogenic effect. Passenger mutations that are acquired prior to the first driver mutation, particularly in cancers exposed to a life-time supply of carcinogens, for example UV light in melanomas, are then present in all cancer cells and contribute to the genetic background of the cancer [76, 77]. As the majority of mutations are thought to have little effect on fitness, and are therefore not negatively selected against, passenger mutations nearly always outnumber driver mutations and are a tribute to the tolerance of the cancer cell to mutational burden.

The concept of clonal evolution dictates that the process of mutation, cell survival and selection proceeds until a cancer has acquired the necessary hallmarks to become malignant [5]. As cell division is a branched process and mutation may occur in practically every cell division, genetic diversity – also commonly referred to as heterogeneity – is an inherent property of a tumour.

1.5.2 Mutability of the cancer genome is an essential property of evolution

Mutability is an essential part of all evolutionary processes and the clonal evolution of cancer is no exception. These mutations may be small sequence mutations such as point mutations and indels or large chromosomal alterations as is the case with CNAs. Additionally they can be random permeations in epigenetic patterning which are passed onto daughter cells. Regardless of the mutation type, all human cells have a baseline level of mutation as an inherent risk of cell division, ergo all cells which still divide have a chance of making the first step towards transformation [78]. However this chance is greatly increased when the tissue is exposed to carcinogens. The continued mutability of the cancer genome during clonal evolution is essential for the cancer to explore the fitness landscape and to achieve phenotype

alterations which are potentially advantageous.

Many mutations may not have large effects on fitness individually but may have a cumulative effect on cell fitness. Mutability also must exist in a 'Goldilocks' zone as, if mutability is too high, the effect on fitness may become negative. Considering that clonal evolution in cancer is analogous to asexual reproduction in microbes, cancer cells are at risk of 'Muller's ratchet' in which deleterious mutations become irreversible [78]. This is potentially reflected in the fact that some cancers with extreme genomic instability have slightly better survival rates than those with a more intermediate range of genomic instability [79].

1.5.3 Drift in the context of cancer evolution

It is also important to recognise that cancer cells are subject to genetic drift. A vital point in this concept is that the acquisition of a mutation occurs in just a single cell. At this point with a population of one, the cell is in a precarious position for achieving survival. Even if the cell possesses an advantageous mutation, the stochastic process of drift, in which all cells have a chance of not surviving, can affect the ability of a clonal population to become fixed. Small populations of cells are at an increased risk of being eliminated due to drift than larger populations, therefore the effect of drift is at its most important in the early stages of mutational acquisition [80, 81].

1.5.4 Selection in cancer evolution

Mutation and drift are largely stochastic processes, although there is evidence to suggest that mutation is non-random, at least in terms of the areas of the genome that it affects [82]. However, clearly the major deterministic force in clonal evolution is selection. Once the clonal population has become robust to drift, it is selection that ultimately determines the success of a clonal population. Selection can be both positive or negative and is clearly highly dependent on context. Clonal populations may remain a subset of the total population (a 'subclone') or grow to dominate the entire tumour. Negative selection means that in certain contexts,

mutations are unlikely to survive in the cell population, or have reduced survival rates compared to the rest of the population. Despite the fact that each mutation appears in a different context, for some mutations selection may be predictable. For instance, the identification of common driver mutations in different cancer types dictates that these mutations are likely to be selected for. However, we ultimately do not know how many 'trials' a driver mutation in each gene has had in the cancer population before it becomes established [78, 80].

An important factor when considering the forces of evolution in a cancer population is the size of the population. In small population sizes, the effect of selection on the make up of the total population is likely to be higher as the new clonal population will constitute to a larger fraction of the cancer population. However, as the tumour grows to a large size new populations, even advantageous mutations that would be under strong selection in a small population may find it more difficult to establish themselves and may explain the presence of a subclonal driver mutations [83]. Furthermore, as the population of cells increases and the number of cell divisions that have occurred goes up, the chance that the mutable cancer genome has mutated at any given base at least once in the population increases, meaning the probability of an important driver mutation being acquired in at least one cell increases in larger tumours [78].

As the cancer cell population increases, so does the size of the tumour itself and the tumour begins to occupy more space. As the tumour grows in size the importance of spatial location increases and the context dependence of selection related to where the cell sits becomes more important. Cell selection may be influenced by its interaction with neighbouring normal tissue or may also be affected by its viability in the interior of the tumour, that may produce increasingly hypoxic conditions [84].

Essentially the ecosystem in which tumour cells find themselves is highly important for determining cell survival [85]. Cells, like animals or other units of

selection, must compete for resources with other cancer cells and normal cells. The demand of the cancer cell for resources may also be much greater than that of the surrounding normal tissue as cancer cells have a higher metabolic rate [5]. Therefore a population of cancer cells must overcome several environmental barriers before they are fully capable of establishing large malignant populations. These barriers could be key to the long periods of latency observed in tumour development in adults [65, 66, 85].

Perhaps the most extreme example of the importance of the environment in selection in cancer is the establishment of a metastasis. Here a cancer cell or a group of cancer cells must establish a new population in a completely new context [78]. The drastic change in environmental context is likely to select only for a particularly viable clone from the primary tumour, even though the cells from the primary tumour have already established most/all of the hallmarks of cancer. This selection in the new environment is likely to mould the metastasis into a unique population, distinct from the primary tumour. The required divergence of the metastasis from the primary tumour is likely to be closely related to tumour type and metastatic site [86].

Selection pressures clearly must exist in tumour populations, however it may be the case that at diagnosis tumour are not longer under active selection. 'Neutral evolution' is a term used to describe the existence of a population of cancer cells that is not under active selection following a series of driver events and may appropriately describe a large proportion of tumours at diagnosis [87]. It may be possible that neutral evolution must be rejected in order to determine the presence of active selection in a tumour population.

1.5.5   The temporal nature of cancer evolution

It is also important to recognise that the process of evolution occurs through time and many factors that shape the evolving tumour are temporal. For instance, the mutability of the genomes of the cancer cells is changeable and over time the

average mutation rate may increase, fuelling tumour evolution, as intrinsic repair mechanisms break down or environmental exposure increases damage. Perhaps one of the most dramatic changes in selection pressure for a cancer is the insult derived from treatment, such as chemotherapy. Here therapy aims to alter the environment of the tumour in order to increase negative selection of tumour cells and cause cell death. However, it is also common that within the large tumour population there is a clone that is resistant to therapy and becomes positively selected for as part of a process known as 'competitive release' [88]. In this way, genetic diversity becomes essential for the population to survive large shifts in selection pressure.

The genetic context of tumour evolution also changes temporally. The order in which mutations are capable of establishing themselves may be dictated by epistatic prerequisites. Here a mutation may only cause a positively selected phenotypic change if it follows from another mutation which provides a phenotype fertile for the development of this sequential mutation [78, 88].

Furthermore as the cancer evolves and subclones become established, selection will shift depending on the competition between the clones as they compete for space and resources. It is this competition that may reduce a clone to a low undetectable population level, that may be released upon a therapeutic shift in the selection environment. Alternatively, clones may cooperate, developing a symbiotic-like relationship in which clones depend upon each other [89, 90]. Overall, clones are not existing independently but make up a system which interacts with itself and is malleable.

1.5.6 General patterns of clonal evolution

Overall the cumulation of these factors in the framework of this system may provide a predictable pattern of mutation acquisition and clonal success. However, there is much debate about the nature of these changing steps. Early work in cancer progression from cross-sectional studies suggested tumours evolve largely as a linear set of transitions [91]. These may take the form of successive clonal sweeps in
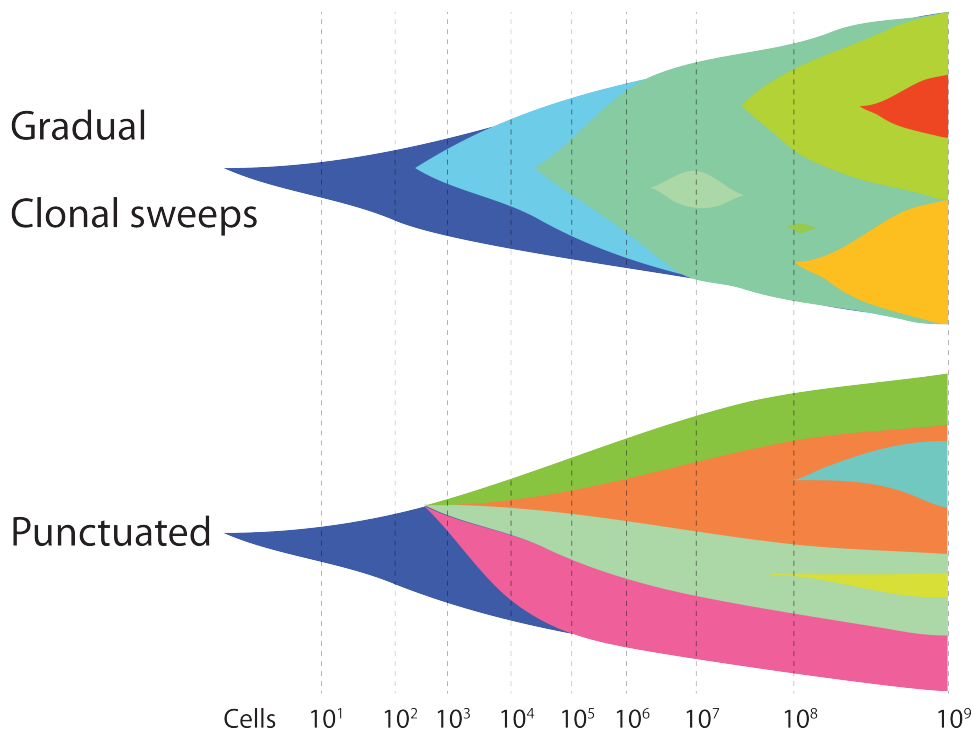
Figure 1.4: There are conflicting theories as to how clonal evolution occurs over time in tumours. These can be generalised into two groups, gradual evolution in which clones emerge and then grow to dominate the tumour in gradual successive clonal sweeps, and punctuated evolution in which large shifts in clonal make-up occur in dramatic bursts which subsequently undergo few adjustments in latter stages. This is represented in this in figure which shows clones as colours and the width of the graphic is related to tumour size across its evolutionary history.

which new clones with increased fitness advantage grow to dominate the tumour. Perhaps these steps may equate to a gradual process in which the tumour develops traits in a piecemeal-like fashion [88].

Alternatively evolution may occur in a more punctuated pattern in which cells acquire various additional traits and then in a short period of time dramatically increase in size or diversify in terms of population composition, causing expansions in size or shifts in fitness to occur in bursts. Punctuated evolution may even show large explosions in size in which several clones all dramatically increase in size largely at the same time (Fig. 1.4). This is known as the 'big bang' model

of evolution [83]. Here distant, spatially separate clones in the diagnosed tumour largely developed at the same time point as opposed to spatially separate clones in a tumour that developed through selective sweeps in which distant clones may represent observable ancestors [92].

Overall patterns of evolution are likely to be determined by the multiple layers of factors that may only be generalisable depending on the cell of origin and the cancer subtype [93]. Evolution is a repeated process in cancer but no two contexts are identical nor are the stochastic processes driving them, which ultimately leads to each cancer being unique.

1.6 Cancer from an evolutionary perspective

1.6.1 Tumour evolution and cancer treatment

Cancer evolution underpins tumour growth and phenotypic presentation in every tumour. All cancer biology can be put in the context of tumour evolution as each cancer hallmark must be selected for [93]. Understanding tumour evolution allows for important genetic events to be reconstructed across real time and for the dynamics of tumour growth prior to diagnosis to be elucidated by utilising principles derived from evolutionary biology [66, 83]. Through studying large cohorts of patients the patterns of driver mutation ordering can also be resolved [66, 94, 95]. Determining mutational ordering in tumour types reveal the genetic events that are important in the earliest stages of tumour growth and are likely to be clonal in a diagnosed tumour, in addition to revealing mutations that are more likely to be acquired late in the tumour's evolution and be present in only a fraction of the tumour population.

Determining if targetable genetic events are more likely to be clonal mutations present in all tumour cells can allow for targeted treatments to be selected. If therapies target mutations often present in only a fraction of the tumour cells, a proportion of the population will be unaffected and likely be ancestral to the clones that present at relapse [96]. Studies that quantify the frequency of subclonality in

important genetic alterations relevant for potential treatments, for example human leukocyte antigen LOH in lung cancers that may be candidates for immunotherapy, provide rationale for selecting patients for treatment [97]. Furthermore, recognising that clonal diversity in 3D space in solid tumour tissue may prevent the detection of important biomarkers has highlighted the advantages of assaying tumour DNA in circulation. Tumour DNA extracted from the blood may represent a mixture of DNA from all clones present in a tumour and may allow for better detection of biomarkers across a patient cohort [98].

Assessing the relatedness of tumour samples across space and time also reveals how the population of tumour cells alters in the establishment of metastases or evasion of therapy. For instance, metastases may develop through a linear progression of changes, in parallel, or may be derived from a dormant population of tumour cells [99]. In each model of metastasis the relatedness of the metastatic clones to the primary tumour determines the likelihood that a single therapy will eradicate the cancer at all tumour sites. Assaying tumour divergence between timepoints in addition to between tumour sites also reveals if rare populations of cells are capable of replacing the original tumour population following therapy, revealing the potential of a cancer population to be completely restructured across the course of treatment [100].

Modelling clonal growth dynamics over time, as well as the common ordering of mutations, may allow for predictions to be generated of the ability of the tumour, or a precursor cell, to continue to grow, or allow for the capability of a tumour to adapt to changes in selection pressure (e.g. during treatment) to be predicted [101, 102]. Approaches that model evolutionary dynamics may outperform genetic biomarker assays as they will assess tumour population fitness directly.

Nevertheless, a recognition of intratumour genetic heterogeneity should alter the perspective of the ability of single sample studies to detect biomarkers, for instance

when assessing chromosome 1q gain as a biomarker in WT, as these studies are likely to be affected by clonal diversity [103, 104]. Understanding the relatedness of clones in separate tumour sites, masses, nodules or areas presenting with phenotypic differences, allows us to address how similar different tumour cell populations are and if they should be treated as a single homogeneous mass.

1.6.2 Clonal evolution in paedatric tumours

Studies assessing the clonal evolution of paediatric cancer by directly assaying multiple tissue samples predate contemporary studies. At the turn of the 21st century groups studying neuroblastoma assessed multiple samples from primary tumour, metastasis and recurrence to assess how tumour cytogenetics altered over the course of the disease [105, 106]. More recently, assaying primary-relapse pairs in neuroblastoma has implicated mutations in the growth-promoting RAS-MAPK pathway in clones capable of causing recurrence [107].

Research in medulloblastoma metastasis and recurrence samples has demonstrated that these devastating lesions can be formed by clones that are rare in the primary tumour. Here clones identified in metastatic and recurrent tumour samples were strongly divergent from the primary tumour [108, 109]. Interestingly multiple tumour samples taken from primary medulloblastomas showed that the tumour is transcriptomically homogeneous, yet by the same approach somatic mutations were strongly heterogeneous. Therefore demonstrating that somatic mutation may be pressured into a form of parallel evolution to achieve a uniform expression phenotype [110].

In the aggressive paediatric brain tumour, diffuse intrinsic pontine glioma (now included in the World Health Organization 2016 entity of 'diffuse midline glioma' [111]), multiple tumour samples revealed that histone 3 K27M mutations, that can induce cell proliferation, occur early in tumour evolution. Contrastingly mutations in the PI3K pathway, that similarly propagates growth factor signals, were found to occur later on. The genetic history of the tumour shows that most of the spreading

of the cancer occurs before its diversification but that main drivers, unlike in medul-loblastoma, are clonal [112].

CNAs have been timed in single sample studies in both rhabdomyosarcoma and adrenocortical tumours. Interestingly, in both tumours chromosome 11p15.5 LOH was considered an early event. In adrenocortical tumours chromosome 17p LOH was also observed as an early event which may be related to the association of the tumour with germline *TP53* mutations [113]. In rhabdomyosarcoma mutations in the RAS pathway were also found to occur early [114].

In WTs within-sample mixing of clones detected in single nucleotide polymor-phism (SNP) arrays has been considered to be prognostic. Additionally multiple tumour pieces have shown genetic diversity [115]. Many key aberrations in WT have also been observed to be heterogeneous and likely late events in some tumours, such as *AMER1* mutation [116], *TP53* deactivation [117] and *MYCN* upregulation [118].

1.7   Assessing clonal evolution in cancer

Through understanding the mechanism of clonal evolution and abstracting its impli-cations for tumour growth, we can conceptualise the course of tumour progression. However, it is ultimately through assaying the genomic composition of a tumour and inferring evolutionary relationships through these assessments that we are capa-ble of adding detail to this picture. Owing largely to the advent of modern genomic assays, there has been an explosion in research into assaying the products of tumour evolution.

Analysis of the total population however is difficult. In a tumour with a diame-ter of approximately 2cm, there are likely to be several billion tumour cells [119]. Therefore, to assay a tumour's evolutionary history multiple approaches have been taken to decipher the details of such a large and diverse population (Fig. 1.5A).
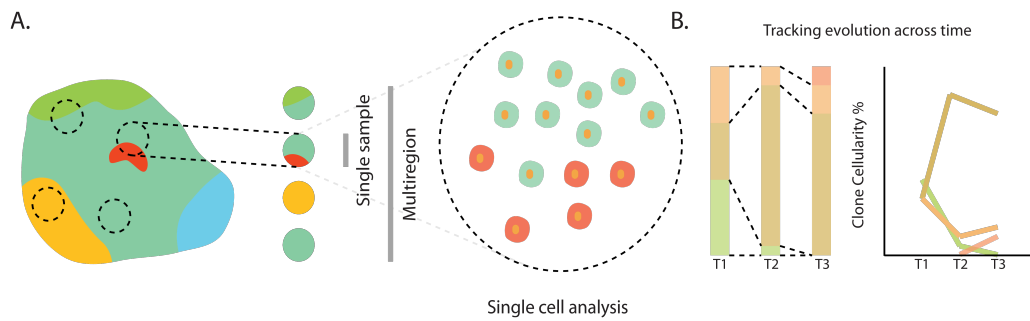
Figure 1.5: Assaying a resected primary tumour to determine its evolutionary history is performed in many ways. It can be performed on single samples which contain a mixed population, by biopsying many samples that contain spatially separated clones or by extracting single cells. Each method provides a unique perspective on evolutionary history although no single method is likely to allow the assessment of all clones. In (A) this is represented by the 'blue' clone which evades detection by any methodology in the example. Lastly liquid tumours and ctDNA can be traced across time by liquid biopsies at multiple time points (T1–T3) (B). Here clones can be determined by resolving the changes in cancer cell fraction across time. This approach can be used to determine the expansion, regression, eradication or the emergence of clones across a given time-frame.

1.7.1 Single bulk tumour samples

The majority of tumour tissue samples taken for research from patient derived tumours are single biopsies. This may relate to the fact that many tumour assays have been based on the assumption, or acceptance, that tumours are genetically homogeneous enough that single samples are representative of the total mass.

However, the sample of tissue taken still remains a sample of the overall tumour and contains a mixed population of cells. By assessing the contribution of each mutation to the total cancer cell population, the structure of the population can be inferred. For example, mutations present in all cells can be considered 'clonal', that is to say that they occurred early in the evolution of the cancer and each cell has subsequently cloned this mutation [120, 121]. By identifying a set of clonal mutations it is possible to consider these mutations as having come before other non-clonal mutations. Non-clonal mutations, also called 'subclonal', are only present in a fraction of the population and are assumed to have occurred after the clonal mutations. Subclonal mutations can then be ordered amongst themselves. It is assumed that

during the expansion of a subclone in the population that the basket of mutations that it carries will be present with the same/similar estimated contribution to the cancer cell population (cancer cell fraction). Therefore subclonal mutation clusters can be identified by being calculated as groups of mutations with similar cancer cell fractions. This problem can be modelled as an *n*-component mixture that can be solved with a hierarchical Bayesian Dirichlet process, for example, that may be used to estimate the unknown number of subclones, their proportion in the tumour and their mutational burden [120].

The cancer cell fractions of each subclone may then consequently be used to order their appearance in the tumour's evolution. Subclones with a lower cancer cell fraction may be considered to appear within the population of a subclone with a higher cancer cell fraction in a linear fashion. Branched evolution may also be shown to exist, only if the cancer cell fractions of potentially independent subclones (those on separate branches) do not total more than 100% of the tumour population. These rules were succinctly described by Nik-Zainal and colleagues and are relevant for a range of single sample evolutionary studies [120].

1.7.2   Multiregion tumour profiling

One method for overcoming the complexities of inferring population structure within a single sample, where all estimations of mutation cancer cell fraction are decoupled from their underlying subclonal association, is to take advantage of the fact that separate clones are likely to be present in spatially separate areas as physical mixing of clones can be restricted in solid tumours (although this is not the case in all tumours, such as in the Big Bang model of tumour growth [83]). This can be achieved by assessing multiple pieces of tissue separately, which can also allow for a wider range of clones to be sampled.

The mutational profiles of each sample can therefore be assessed and samples that share mutations are subsequently known to likely be related. For instance, the presence of a mutation in all biopsy samples taken suggests the mutation is

clonal, whereas a subset of samples sharing a mutation may be related at a later point of divergence from other samples. Multiregion sequencing has shown great strength in assessing a snapshot of the multiple clones present in separate tumour sites highlighting, for example, the presence of genes presenting unique mutations in separate tumour regions (convergent evolution) [65, 122, 123]. Additionally, microdissection can also be performed to analyse specific cells across the tumour that may aid phylogenetic reconstruction, for example glands in colorectal tumours [83].

1.7.3 Temporal assessment of tumour evolution

The previous two approaches to tumour sampling often involve sampling at only a single given timepoint. It is sometimes possible to detect mixing in the cell population and to resolve subclones by tracking the contributions of mutations to the total population across a time-course (Fig. 1.5B). Mutations belonging to the same clone will produce the same pattern of cancer cell fraction changes over time. This method is limited in solid tumours, often to comparing between primary tumours and relapse samples [124, 125]. However, this is far more feasible in liquid cancers in which the cancer can be monitored at regular intervals across time [126, 127]. This can allow for the tracking of clonal progression across long periods of time [128]. The prospect of assaying circulating tumour DNA (ctDNA) released from primary solid tumours and potentially their metastases into the bloodstream may allow for solid tumours to be monitored in a similar fashion to a liquid cancer [129, 130].

1.7.4 Single tumour cell assays

The previous assessments of tumour clonality all involve assaying a mixed population of cells. However, it is clear that the single cell itself is the unit of selection in cancer evolution. Assessing single cells individually was one of the first ways in which genetic diversity in cancer was observed [131].

By assaying genome-wide mutations in single cells there is no requirement to infer subclones as each subclone is being measured directly [132, 133]. Single cell

assays however are limited by the numbers of cells which can be assessed and by the limitations of noise that comes with attempting to assay directly only a few molecules of DNA for each base pair.

1.7.5 Combining approaches

These different approaches may be combined to great effect. For instance, the principles of single sample bulk assessments can be applied to each assessed tumour piece in a multi-sample assay to potentially greatly increase the resolution of phylogenetic reconstruction across multiple samples and also avoid erroneous identification of convergent evolution [94, 134, 135]. Additionally, assaying multiple regions in a primary tumour can allow for subclones to be resolved and tracked longitudinally in ctDNA [98]. Furthermore, to overcome sampling bias, single cell assays can be combined with bulk assays. Using a set of single cells to compare with bulk samples can help resolve subclones that present in the bulk sample with less reliance on previously described assumptions [136, 137].

1.8 Assaying the genome

To assess genome-wide mutation in an evolving cancer, a researcher must utilise a genomic assay. There are two common techniques to approaching this in cancer research (1) second generation / next generation sequencing (NGS) which assays genomic DNA by sequencing millions of short DNA fragments and (2) SNP arrays in which genome-wide allele-specific CNAs can be assessed.

1.8.1 Next generation sequencing

Until the mid-2000s Sanger sequencing was the predominant method of resolving the base sequence of a given section of the genome. However, in the latter half of the decade several companies had developed methods of achieving massively parallel DNA sequencing, termed NGS. The key to this technology was to separate individual DNA fragments and to clone them in spatially separate compartments or regions. These separate fragment clones could then be sequenced in parallel to sequence millions of fragments simultaneously [138, 139].

This method of sequencing is often referred to as 'massively parallel' and allows for many short sequences to be obtained from the same sample of mixed DNA fragments. Sequences are commonly short ~35–150bp due to a decrease in base quality as sequencing progresses, often due to 'dephasing' in which base sequencing in the clone cluster goes out of sync [138, 139].

NGS can be performed on any DNA sequence that is of a large enough length and allows for whole genome sequencing (WGS) to be possible in pragmatic time frames. However, NGS is also often used to sequence a subset of the genome that is of interest to a particular study [140], for example, the exome or a panel of cancer genes. Additionally, NGS can be used to sequence copy DNA reverse transcribed from RNA (RNA-seq). Sequencing only a subset of the genome allows for the target region to be sequenced more times (a higher coverage) or for more samples to be sequenced for the same sequencing cost. NGS can also be used to sequence low quantities of DNA, for example, DNA derived from single cells [132].

NGS can detect all types of sequencing alterations and as the probability of a region being sequenced is proportional to its copy number, sequencing can also be used to identify CNAs.

1.8.2 Next generation sequencing data analysis

The primary output of a NGS run are large numbers of reads – small sequences of a determined length – often stored as FASTQ files which also possess information regarding the quality of the base call in each base of the reads. To interpret these large numbers of small sequences (often 1×10^6–1×10^8) these reads need to be mapped back to the genomic location from which they were derived. This computational challenge is called alignment. Common software developed for this challenge are BWA [141] and bowtie2 [142]. If enough reads have been produced across the sequenced DNA multiple reads will map to overlapping locations forming pile-ups. The number of overlapping reads at any given genomic location is called coverage.

It is in this aligned form that sequence variation can be assessed. It is important to note that there is an inherent trade-off between alignment and biological sequence variation. Read aligners require large section of the read to match perfectly to the reference genome to achieve accurate alignment. However, sequence variants will not align perfectly to the genome. Most aligners are robust to this however as they must take into account mismatches generated by false base calls in addition to biological variation, nevertheless alignment and variant calling are closely related [141, 142].

Perhaps the sequence variation that highlights this problem to the least extent is the single nucleotide variation (SNV). However identifying a SNV remains non-trivial and there are several software developed to achieve this including MuTect [143], VarScan2 [144] and Freebayes [145]. Small insertions and deletions are more complex to identify as they alter larger sections of the genome and cause alignment to become increasingly difficult. Software such as MuTect2 [146] approach this problem using local realignment. These algorithms report candidate variants as alternative bases to the reference base(s) and also detail the numbers of reads supporting the variant compared to the reference nucleotide(s).

1.8.3 Single nucleotide polymorphism arrays

Developed originally to genotype human DNA, SNP arrays have become a common method for assessing mutations in cancer, principally CNAs. By probing for SNPs in the genome the array can report on total copy number level of the SNP as well as the allelic balance between the two bases being probed. SNP arrays are based on nucleotide biochemistry that dictates that complementary base pairs will preferentially hybridise. The genome of the sample is fragmented and denatured to become single stranded. These genomic DNA fragments are then hybridised to complementary probe sequences on an array that are designed to match regions surrounding SNPs with a high minor allele frequency in the population that will bind to fragments and allow for allele specific output. An allele-specific fluorescence-based signal intensity is then reported for each probed SNP and ultimately the signal is proportional

Illumina Single-Base Extension
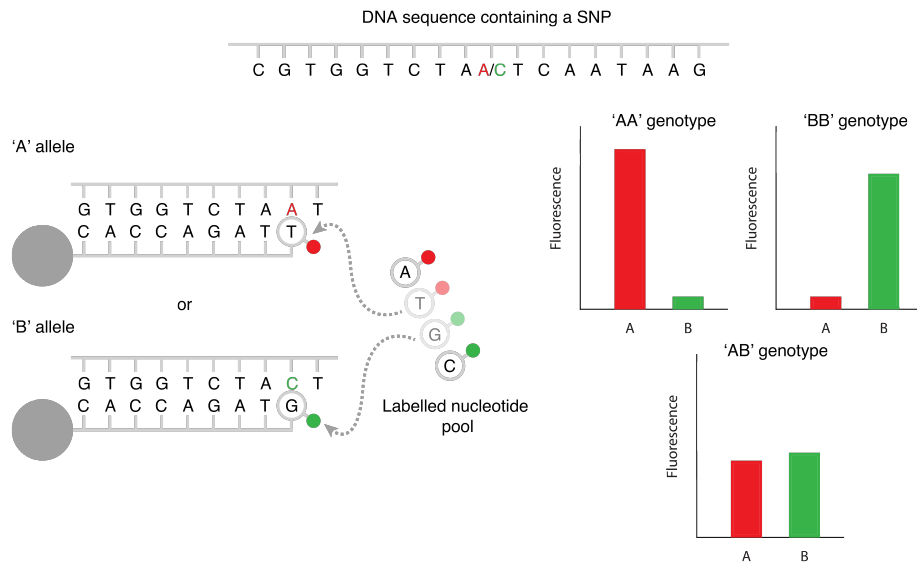
Figure 1.6: Illumina® SNP arrays utilise single-base extension technology. Here sample DNA containing a known SNP hybridises with a complementary sequence adjacent to the location of the known SNP attached to a bead. This leaves the SNP position in the DNA as the first base in the overhang of the hybridisation. Single-base extension is then performed in which the fluorescent-labelled nucleotide complementary to the SNP locus is incorporated. This fluorescence is then read as output for both the arbitrarily assigned 'A' allele and 'B' allele. This figure is adapted from LaFramboise, 2009 [147].

to the amount of sample DNA that is hybridised. This yields an estimation of total DNA content for each SNP in the sample and can be used for both genotyping and studying the somatic CNAs in a cancer genome. The major manufacturers of SNP arrays are Affymetrix® and Illumina®, each with unique chemistry and algorithms to process the raw data, however the output is broadly equivalent [147]. Figure 1.6 demonstrates single-base extension utilised by Illumina® SNP arrays. Here, fluorescent-labelled nucleotides representing the polymorphism are incorporated as a base extension following the hybridisation of the sample DNA sequence to an adjacent complementary sequence. The genotype of the SNP can be interpreted from the relative proportions of the allele-specific fluorescence based signal.

33

1.8.4 Single nucleotide polymorphism data analysis

Output for SNP arrays produced by either Illumina® or Affymetrix® can be interpreted using two types of data, Log R ratio (LRR) and B-allele frequency (BAF) (although they are an Illumina® concept). LRR is defined as the $\log_2$ of the fluorescence observed over fluorescence expected for both alleles at a single position. In this respect the LRR represents the total copy number state at the position of the SNP. The two alleles of the SNP being probed can be arbitrarily assigned as either *A* or *B*. BAF represents the ratio between the signal from these two alleles and therefore represents allelic imbalance by calculating the proportion of the signal generated by the *B* allele as a fraction of the total signal. Homozygous SNPs in an *AA* or *BB* state therefore produce BAF values of 0 and 1 respectively, whereas heterozygous SNPs will produce values in between 0 and 1 [148]. LRR and BAF can also be generated from NGS data using the same principles, where allele-specific coverage represents allelic abundance.

Analysis of this output involves identifying regions of genomically adjacent SNPs with a similar signal that may represent a particular copy number state. This is often performed by a process called segmentation and algorithms such as circular binary segmentation [149] are used to perform this. Once the genome has been segmented into regions of similar/the same copy number state, the average LRR and BAF values of these segments can then be used to model the underlying integer values. To do this in cancer, aspects such as tumour purity and ploidy must be first estimated when interpreting the copy number profile of the cancer sample. Popular software for calculating allele-specific cancer copy number profiles include ASCAT [150] and ABSOLUTE [151], among others, and take advantage of simultaneous modelling of both purity and ploidy.

1.8.5 Cellular prevalence of copy number alterations

As with somatic short sequence mutations, it is also possible to estimate the cancer cell fractions of CNAs. Like all mutations in cancer, CNAs may also be present in only a fraction of the cancer cell population and be carried by a non-dominant

subclone. By estimating the cell cancer fraction of CNAs we can also infer the clonal structure of single tumour sample assays utilising the same principles as for short sequence mutations.

As described by Nik-Zainal and colleagues, to achieve an accurate estimation of CNA cancer cell fraction, it is necessary to have a precise estimation of major allele frequency for each CNA to determine allelic imbalance. To achieve this it is necessary to estimate the BAF distribution that is produced by the major allele [120]. BAF data produced by a cancer sample is comprised of four different distributions. Two distributions tightly bound to 0 and 1 that represent pure homozygous states (AA and BB, respectively) that are produced by both tumour cells and contaminating normal cells if the SNPs producing the homozygous signal are also homozygous in the germline. SNPs that are homozygous in the germline are unimportative as their signal does not change in the event of a CNA and therefore by identifying germline homozygous SNPs, for example in a tissue sample known to be healthy, these distributions can be removed from the data.

Additionally, there are two distributions produced by the major and minor alleles at heterozygous SNPs. The major allele is defined as the chromosomal locus with the greater number of copies and in allelic balance both the major and minor alleles have the same copy number. In segments of the genome where there is allelic balance in both the tumour and normal cells, i.e. a normal diploid state, these two distributions are equal with a mean of 0.5. However, in a somatic allelic imbalance, these distributions separate according to the ratio of the major allele to the minor allele and the percentage of normal cell contamination (tumour purity of the sample). The means of both distributions are equidistant from 0.5 and the major allele distribution has a mean BAF greater than the minor allele distribution.

A simple, but insufficient, method of estimating the major allele distribution is to subtract all germline heterozygous BAF values less than 0.5 from 1, essentially
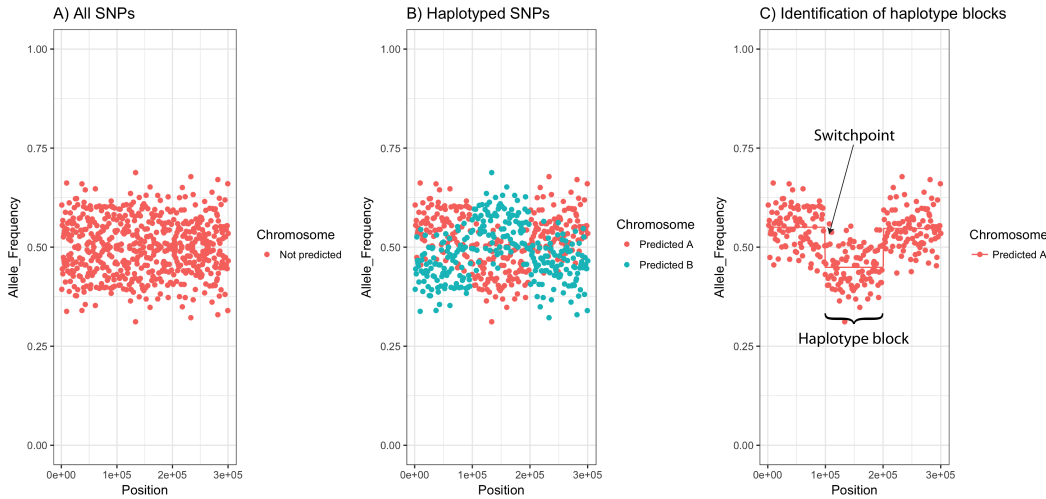
Figure 1.7: Performing a haplotype prediction of a cancer patient can allow for the identification of the major allele distribution. Panel A shows the allele frequencies of variants in heterozygous SNPs across a hypothetical chromosome in which there is an allelic imbalance. The separation of the major and minor allele distributions is not clear as the tails of the distributions overlap. Panel B shows the product of a hypothetical haplotype prediction, the two chromosomes are termed 'Predicted A' and 'Predicted B'. When taking only a single chromosome prediction ('Predicted A', panel C), three 'haplotype blocks' are observed produced by two 'switchpoints' in which the linkage of adjacent SNPs was falsely predicted. Two of these haplotype blocks are part of the major allele distribution (mean >0.5) and the middle block is part of the minor allele distribution (mean <0.5). The allele frequencies in the minor allele haplotype block are subtracted from 1 to generate an estimation of the major allele distribution.

mirroring the BAF distributions at 0.5 [152]. This sufficiently estimates the major allele distribution in pure CNAs with a large imbalances (e.g. hemizygous losses) if no BAF values in the lower tail of the major allele distribution are less than 0.5. However, in impure allelic imbalances and in allelic balance, a significant proportion of the lower tail of the major allele distribution is less than 0.5 causing the mirroring transformation to skew the estimated major allele distribution and its mean value. To overcome this, and to accurately estimate the major allele distribution, Nik-Zainal and colleagues used haplotype 'phasing' [120].

Here, Nik-Zainal and colleagues take advantage of the fact that genetic linkage between variants in SNPs on chromosomes is predictable, as the closer two variants are to each other on a chromosome, the less likely they are to be separated during

chromosomal crossover. Using reference information from a haplotyped dataset, such as the 1000 Genomes Project [153], a computational prediction of the physical linkage of each variant in the heterozygous SNPs of a patient sample can be performed [154]. This generates a prediction of the haplotype representing each chromosome, based on the genotype of patient. Consequently, the measured allele frequencies of each variant in a haplotype can be identified. It is unlikely that a haplotype prediction is accurate for the length of a whole chromosome, however a sufficient prediction will be accurate for a consecutive series of SNP loci in blocks. These blocks will then belong interchangeably to the major and minor alleles, separated by an inaccurate prediction of linkage between two SNP loci (Figure 1.7). A series of correct linkage predictions are known as 'haplotype blocks' and the inaccurate predictions between two SNP loci that separate them are known as 'switchpoints'. A lower frequency of inaccurate linkage prediction between two adjacent SNPs increases the size of the 'haplotype blocks' and reduces the number of 'switchpoints'. To estimate the major allele distribution the 'haplotype blocks' are identified through segmentation and the allele frequency of each SNP in the minor allele 'haplotype blocks' (a mean allele frequency <0.5) are subtracted from 1. This produces an estimation of the major allele distribution and retains the lower tail in regions of allelic balance and impure allelic imbalance.

This approach forms the basis of a tool known as 'Battenberg' in reference to the patterning of the haplotype blocks. The accurate major allele distribution estimation can then be used to model each CNA loci as a mixture of multiple states and allows for the cancer cell fraction to be estimated [120].

1.9   Inferring phylogeny in cancer data

A key output for the study of tumour evolution is the inference of phylogeny between the observed clones. There are many approaches to understanding cancer evolution, many of which have roots in approaches designed for speciation phylogenetics. However, clonal evolution presents a unique context compared to speciation
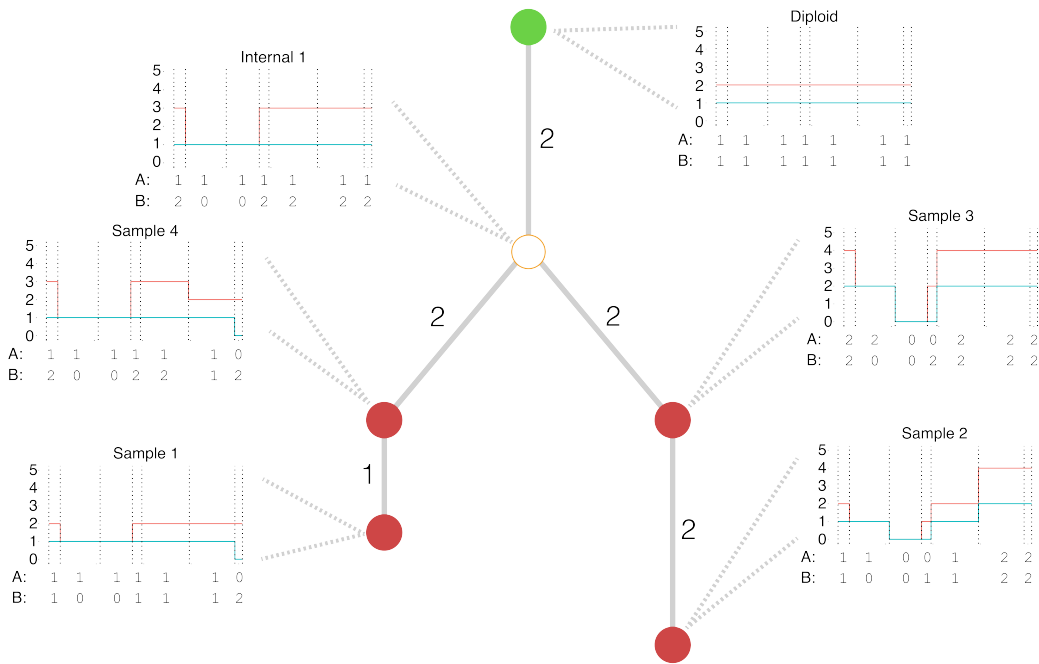
Figure 1.8: The phylogenetic history of a tumour can be inferred from the copy number profiles of multiple tumour pieces. This figure represents the inferred phylogeny produced from four unique copy number profiles in a single chromosome by MEDICC. The inferred phylogeny is rooted by a normal diploid sample (green). By determining minimum evolution and the copy number state changes between these four samples, an ancestral copy number state is reconstructed (Internal 1) as well as the order of copy number changes that may have produced the observed states (tumour samples – red). Performing this allows the relationships between the tumour samples and their history to be determined. Each profile shows total copy number state (red) and copy number of the *A* allele (turquoise). Integer copy numbers for each allele are displayed under the profile and the number of events between related profiles is displayed next to each branch. This figure is an adaption of Figure 1 in Schwarz et al, 2014 [155]

due to factors such as purely asexual reproduction, large population sizes with large amounts of diversity and perhaps most importantly, rapid mutation rates including dramatic structural rearrangements [156].

The approach for understanding clonal evolution also depends on the type of data being collected and the method of choice for sampling the cell population. The most commonly measured type of variant for evolution is SNV and CNA, although other variations can be used such as gene expression or DNA methylation [156, 157].

There are many ways in which relationships between clones can be inferred, for example, maximum parsimony [65, 158], minimum evolution [159, 160] and neighbour joining [132, 133]. SNVs are often assumed to only occur once in tumour evolution (infinite sites assumption) as well as being irreversible. These assumptions are likely not true for all mutations in the tumour history but these events are apparently so rare they are assumed to have negligible effect on modelling evolution [157]. These assumptions allow for the interpretation of SNV fraction in the cancer cell population to be an indirect measure of the cancer cell fraction of the subclone of which it has arisen from.

Some tumour phylogeny algorithms use only SNVs to infer phylogeny. These approaches either ignore or avoid local copy number states and their effects on variant allele frequency (VAF) [161, 162]. SNV cancer cell fraction should be corrected for local copy number state when that state is aberrant and there are several algorithms that approach solving this problem in order to infer phylogenetic relationships [163, 164, 165].

Evolution can also be traced using CNAs alone, however CNAs have a more complex relationship across clones compared with SNVs. This is due to multiple differences between the nature of CNAs and SNVs. For instance, a CNA of a given set of breakpoints can alter in copy number to multiple levels, for instance a chromosome arm can exist with a total copy number state of 3 in one clone and 4 in another whilst originating from a copy number state of 2. Therefore the ability of a CNA to change in total copy number whilst retaining identical breakpoints must be accounted for. Furthermore, a tumour can possess independent CNAs in which the genomic location of the breakpoint boundaries can be different but the lost or gained regions overlap in different clones. Therefore, distinct CNAs can affect the same genomic location. CNAs are also allele-specific and one allele in one clone may be affected by a similar CNA as its opposite allele is in another clone (mirrored subclonal allelic imbalance, MSAI) [94]. Additionally, once an allele has

been deleted in an evolutionary trajectory, it cannot be reversed. Lastly, large whole genome changes such as WGD can affect the CNA evolution in dramatic fashion [156, 157].

Some algorithms such as TuMult [166] and MEDICC [155] have looked specifically at CNA evolution. MEDICC solves the problem of CNA evolution by taking CNA profiles across multiple samples and 'phasing' them by predicting the assignment of CNA states in the maternal and paternal alleles by calculating the minimum pairwise event distance between the multiple samples. MEDICC then constructs the phylogenetic tree using minimum evolution, a parsimony-like approach to a distance matrix between the multiple CNA profiles (Fig. 1.8). This method allows for CNA phylogenetics to be compared across many different platform types.

The field of tumour phylogenetic inference is nascent despite many novel approaches to solving the problems presented by clonal evolution being proposed. Ultimately the appropriate strategy for choosing algorithmic approach depends on data type and collection. No single solution has been found to suit all situations in the field yet. Indeed, it is through the current cyclic process of collecting genomic data, inferring evolution, drawing conclusions and adjusting modelling of clonal evolution, that the field of tumour phylogenetics will develop.

1.10   Aims of the thesis

This thesis aims to explore the role of tumour evolution in the context of paediatric solid tumours, focusing primarily on a large cohort of WTs and hepatoblastomas. Assessing tumour evolution in a large of cohort of solid paediatric tumours remains largely unexplored, preventing the illumination of the underlying dynamics of the genetic development of these tumours. I hypothesise that by assessing multiple tissue samples from the same primary tumour and utilising genome-wide assays to explore their mutational landscape, I will be able to identify multiple, unique cancer clones. Moreover, if multiple clones within the primary tumour of a patient

are resolved, the phylogenetic history of each individual tumour may be inferred by comparing mutations that are both common and unique to the clones. I aim to achieve this by utilising phylogenetic methods such as minimum evolution.

I will analyse multiple CNA profiles generated from SNP arrays across several tumour tissue samples, as well as targeted next generation sequencing of important genes in a subset of cases in the same DNA extractions. I aim to develop a bioinformatics approach to compare between different mutational profiles from the same patient, particularly of CNA profiles. Furthermore, I also aim to estimate the cancer cell fraction of each CNA/mutation to identify mixed clones within a single tissue sample. I hypothesise that this will allow for further timing of CNAs/mutations in the phylogenetic history of the tumour.

By determining the evolutionary histories of a large number of tumours of the same type, I will be capable of identifying common patterns of evolution, not only specific to the tumour type, but also in tumour subtypes and risk stratifications. Furthermore, I aim to compare my results to common clinical questions and I hypothesise that understanding tumour evolution in a single patient may allow us to increase our understanding of many clinical aspects, including but not limited to: differential treatment response between areas of the tumour, the presentation of histological heterogeneity within multiple tissue samples, and the relatedness of separate tumour masses (e.g. bilateral tumours). Inter- and intra-tumour phenotypic differences are unexplained from a phylogenetic standpoint, therefore I aim to use a phylogenetic understanding to support a rationale for differences in phenotypic presentation.

Moreover, I hypothesise that common molecular biomarkers, such as 1q gain in WT, can potentially occur as late events that have variable detection within a single tumour. I aim to evaluate intratumour heterogeneity of CNA biomarkers in order to inform future studies of the ability to detect biomarkers and to place old studies in

41

this new context. Furthermore, I aim to use my results to evaluate aspects of tumour evolution that may provide novel prognostic value for stratifying patients with solid paediatric tumours.

Also, I aim to identify common selection pressures in WT and hepatoblastoma. By assessing if CNAs/mutations occur early or late in tumour evolution I hypothesise that I will be able to discuss the selection for, and importance of, particular CNAs/mutations at the early or late stages of tumorigenesis. I aim to identify convergent evolution, if it occurs in the data, as an example of both repeated evolution within a tumour and potentially strong selection pressure for a CNA/mutation. Finally, I hypothesise that by identifying common selection pressures in a large cohort of cancers that occur early in life, I will be able to describe tumour evolution in a unique context and will highlight new patterns of tumour evolution rarely observed in adult cancers.

Chapter 2

Methods

This chapter will outline the methods used throughout the thesis. The relationship between each section in this chapter and the corresponding chapter later in the thesis is referred to in the text or made explicit by referring to the datasets separately by their shorthand names. The corresponding datasets are as follows:

- **WT20** – An initial series of twenty multi-sampled WTs presented in Chapter 3. The reanalysis of this dataset is presented in Chapter 4.

- **MicMa** – MicMa highlights breast carcinoma data derived from the Oslo Micrometastasis project previously presented as a series of 112 Illumina® Human-1 109K BeadChip SNP arrays by van Loo and colleagues [150] and a series of six patients analysed using Affymetrix® Genome-Wide Human SNP 6.0 arrays by Demeulemeester and colleagues [136]. This data is presented in Chapter 4.

- **PKC66** – A series of sixty-six paediatric kidney cancers (~94% WTs) presented in Chapter 5.

- **Hep11** – Eleven hepatoblastomas cases presented in Chapter 6.

Methods detailed in Sections 2.1.1, 2.2, 2.3 and 2.11 were performed in collaboration with the laboratory of Professor Kathy Pritchard-Jones at UCL Great Ormond Street Institute of Child Health.

2.1 Sample collection

2.1.1 Sample collection and histological analyses in the WT20 and PKC66 series

We obtained multiple tumour samples from WT nephrectomy/nephron-sparing surgery specimens from patients enrolled on the SIOP WT2001 trial [167], the subsequent Improving Population Outcomes for Renal Tumours of Childhood (IM-PORT) study or whose parents had consented for additional tissue to be used in research as part of the UK Children's Cancer and Leukaemia Group (CCLG) tissue bank. In a subset of cases, we also studied material obtained at diagnostic needle core biopsy. In addition, in one case, tissue from a relapse biopsy was also used. The research was approved by a national research ethics committee. Patients received preoperative chemotherapy as per the SIOP WT 2001 trial protocol or according to national clinical guidelines based on this trial.

WTs were classified for diagnostic purposes as previously described [22]. Each research tissue sample was divided into two pieces. One piece was formalin-fixed and paraffin-embedded and from this at least one histological section was prepared, stained with haematoxylin and eosin (H&E stain) and reviewed by one pathologist (Dr. William Mifsud). Standard histological methods, as per standard operating procedures in place at the Department of Histopathology at Great Ormond Street Hospital, were used. The matching piece was flash-frozen in liquid nitrogen. Only samples with more than 50% viable tumour (the remainder consisting of necrotic tumour/post-chemotherapy change) on histological section were used in this study. DNA was extracted from frozen tissue samples using standard techniques, and from adjacent non-tumorous kidney tissue and/or peripheral blood lymphocytes where they were available.

2.1.2 Sample collection in the Hep11 series

This was performed in collaboration with Professor John Anderson's laboratory at the UCL Great Ormond Street Institute of Child Health. Frozen tumour samples

and where available peripheral blood lymphocytes were obtained by Professor Anderson's laboratory from the CCLG tissue bank. They were assessed and DNA was extracted by Dr. Tessa Kasia using standard methods in Professor Anderson's laboratory.

2.2 Magnetic resonance tumour imaging

In one case (Case 8) in Chapter 3, different regions within the same tumour were identified prospectively as distinct nodules in the same overall tumour mass on T1- and T2-weighted magnetic resonance (MR) imaging, and matched on comparison of pre- and post-chemotherapy images. Apparent diffusion coefficient (ADC) was calculated by one observer as previously described [168]. Radiological images were assessed by Dr. Øystein E. Olsen at Great Ormond Street Hospital.

2.3 Molecular analyses

2.3.1 Illumina® SNP microarrays

Illumina® HumanCytoSNP-12 v2.1 microarrays (~300,000 probes) and Illumina® CoreExome-24 SNP arrays (~500,000 probes) were hybridised with 250ng DNA per sample according to the manufacturer's instructions. Log R ratio ($\log_2$[observed intensity/reference intensity], LRR) and B-allele frequency (BAF) were calculated using the Illumina® GenomeStudio software for each array using default settings. SNP array data in these array platforms was generated by UCL Genomics.

2.3.2 Methylation-specific multiplex ligation-dependent probe amplification

Methylation-specific multiplex ligation-dependent probe amplification (MS-MLPA) for 11p15 was carried out as previously described [169], using the Salsa MS-MLPA BWS/RSS ME030-C3 probemix (MRC-Holland®) and data visualised in Coffalyser.NET (MRC-Holland®), performed and analysed in the Pritchard-Jones lab.

2.3.3 Design of targeted bait capture

A selection of genes was chosen to be pulled down for DNA sequencing using NGS to assay a subset of the WT genome. In total 181 genes were pulled down for sequencing analysis and are listed in Table B.6. These were used to create a bait capture design using Agilent® SureSelect in the SureDesign suite. These genes were chosen based on known WT genes from the literature, results from an exome study by Wegert and colleagues [40] (taking genes mutated multiple times in the study) and recommendations from experienced WT biologists within the group of Professor Pritchard-Jones. The gene list was curated to ensure a potential role in WT biology. Maximum tiling density was used to pull down the exons of this gene list (5X). In total the capture design probes for 0.881 MB of genes.

Additionally, a region of 125 bp surrounding high minor allele frequency SNPs were also probed in this bait capture design to assess a subset of CNAs, determined by the 'OneSeq' backbone provided by Agilent®. Targeted regions were chromosome 1 (1p and 1q), 4q, 11p, 14q, 16q and 22 and cytobands containing the genes *MYCN* and *TP53* (2p24.3 and 17p13.1, respectively). In whole chromosomes or chromosome arms I took every one in ten SNPs from the Agilent® OneSeq backbone for inclusion into the bait capture, for the single cytobands, I included every other SNP for panel inclusion (50%). Tiling density of the CNA regions was 1X. This CNA data was not included in this study but is referred to briefly in Chapter 5.

2.3.4 Targeted next generation sequencing of tumour samples

Library preparation, targeted capture and sequencing was carried out at the UCL Pathogen Genomics Unit. NGS was performed on an Illumina® NextSeq machine. For each sample paired-end reads of 151 base pairs were generated. Sequencing data was provided as FASTQ files.

2.4 Multiregion copy number alteration phylogenetic analysis in SNP arrays

2.4.1 Correction of genomic waviness in LRR data

LRR genomic waves [170] were detected in normal tissue samples and corrected from all arrays in an array specific manner for Illumina® HumanCytoSNP-12 v2.1 and Illumina® CoreExome-24 v1.0/v1.1 SNP arrays. A normal sample cohort for each array type was selected. For each autosomal SNP array probe, the mean LRR across all arrays in the normal tissue cohort was calculated, generating a mean normal profile. For each autosome a moving average LRR profile was generated from the mean normal profile using a window size of 10 probes. The moving average profile from the normal samples represents the genomic waves of each autosome in the array. Genomic wave detection was applied separately for male and female X chromosome LRRs. To centre the male X chromosome signal around zero, the profile was subtracted by the median value. The contribution of the genomic waves to genome-wide LRR variation in each array is determined by subtracting the genomic waves (G) from corresponding array probe LRRs (S) using the coefficient (α) which minimises LRR population variation across all genomic probes to generate normalised LRR (N),

$$argmin\{var(S_\mathrm{i} - \alpha G_\mathrm{i})\} = \alpha_{\min},$$

$$S_\mathrm{i} - \alpha_{\min} G_\mathrm{i} = N_\mathrm{i}.$$

LRR values were corrected for genomic waves in each array for all autosome and X chromosome probes if included in the study.

Methods described in the following sections, 2.4.2–2.4.6, were developed for and specifically used in Chapter 3.

2.4.2 Identifying regions of copy number changes per case using CGHcall-CGHregions

The LRRs from each WT20 array were segmented and copy number states were determined using the CGHcall R package [171] in Bioconductor [172] using the 'sdundo' option for the 'undo.splits' parameter for segmentation (undo.SD = 5, clen = 13, relSDlong = 8.33). Called copy number segment boundaries were smoothed and segments were condensed into genomic regions with corresponding copy number state in each region per case using the 'CGHregions' R package [173] in Bioconductor ($c = 0$). Genomic regions that contained less than 100 probes were removed. Genomic regions were retained if the number of base pairs per array probe was within the median probe density plus/minus double the interquartile range of all genomic regions of all cases analysed in WT20 ($10^{3.5}$ to $10^{4.4}$ base pairs per probe in the WT20 series, i.e. median(all regions) $\pm$ 2*IQR*[all regions]) to filter genomic regions with extreme tiling densities.

2.4.3 Tumour-specific mirrored B-allele frequency profile

In order to generate a tumour specific mirrored BAF (mBAF) profile the estimation of aberrant cell fraction (A) generated by ASCAT [150] (min. 0.8, max. 0.95) was used to define a threshold (T) above which probe mBAF values are considered non-informative homozygous SNPs and removed,

$$1/(1 + (1 - A)) = T$$

The remaining mBAF values (B) that are greater than 0.56 (used as a threshold of allelic imbalance in [152]) are scaled between 0.5 and the non-informative homozygous SNP threshold (T) to give a tumour specific mBAF value (M),

$$((0.5B) - 0.5)/(T - 0.5) + 0.5 = M$$

Mean mBAF was calculated for each copy number loss (copy number = 1) and gain (copy number = 3) state genomic region and for each region, calls were reversed to normal if the mean mBAF was not greater than 0.58 and 0.64 for gain and loss

states respectively. This was on an empirical cut off to remove the distribution of normal mBAF values (~0.5) when allelic imbalance was expected.

2.4.4 Copy number neutral loss of heterozygosity detection

As 'CGHcall' uses LRR data to generate copy number segments, copy number neutral loss of heterozygosity events, that do not alter LRR, are not detected. Therefore I developed an algorithm to identify LOH in genomic regions of a total copy number of 2. Here I counted the number of probes with a tumour-specific mBAF $<= 0.66$ in windows of 100 SNP probes, sliding by 10 probes per chromosome. Windows with 1% or fewer probes with mBAF $<= 0.66$ were considered to not be in the 'AB state' of normal allelic balance. Consecutive windows of allelic imbalance were merged for each array and boundaries across different arrays from the same tumour were smoothed to make these regions comparable. These newly identified regions were then incorporated into the genomic regions generated from the LRR data, if (1) an allelic imbalance region was called within a genome region with a copy number state of 2 (as described earlier) and (2) the allelic imbalance region contained equal to or more than 1000 SNP probes (more than 90 consecutive windows called as allelic imbalance).

2.4.5 Calculation of major and minor copy numbers

Once the genomic regions of copy number states for a case were identified, I calculated major (A) and minor (B) allele copy numbers for each region using the total copy number (T – as called by 'CGHcall') and the mean scaled mBAF (M) for the probes within each genomic region. I used the following equation to determine major and minor allele copy numbers;

$$round(TM) = A$$

$$T - A = B$$

2.4.6 Phylogenetic analysis of tumour samples

The major and minor allele copy numbers per chromosome per array were used as input for MEDICC [155]. MEDICC was used to infer clonal evolution of samples in each case. Normal tissue samples were used to root phylogenetic trees.

2.5 MicMa breast carcinoma array datasets

The dataset of 110 MicMa breast carcinomas analysed in Chapter 4 was generated using Illumina® Human-1 109K BeadChip SNP arrays and published by van Loo and colleagues [150]. The original dataset contains 112 samples however in Chapter 4 only 110 are presented as two samples failed to provide a purity and ploidy solution in ASCAT.

The five MicMa breast carcinoma samples also presented in Chapter 4 were analysed using Affymetrix® Genome-Wide Human SNP 6.0 arrays and taken from a previous study by Demeulemeester and colleagues [136].

2.6 Copy number alteration cancer cell fraction phylogenetic analysis

2.6.1 Phasing of SNP array data to predict major allele distribution

Phasing was performed in two array types, Affymetrix® Genome-Wide Human SNP 6.0 arrays and Illumina® HumanCytoSNP-12 v2.1, as described by Nik-Zainal and colleagues [120]. This initially involves converting CEL files to LRR and BAF using the PennCNV-Affy pipeline (available at penncnv.openbioinformatics.org/, accessed September 12, 2017) excluding the generation of a canonical genotype clustering file (step 1.3) for the Affymetrix® Genome-Wide Human SNP 6.0 arrays. Phasing was then performed on both array types using impute2 [154] and the phased 1000 Genome Project reference genomes (v3, 2012) [153] using heterozygous SNPs determined by the corresponding normal sample. Haplotype blocks

were identified by segmenting a single haplotype prediction using piecewise constant fitting (PCF) segmentation as developed in the ASCAT package (gamma = 1, kmin = 3) [150]. BAF values in haplotype block segments with a mean BAF less than 0.5 are transformed to represent the alternate allele by subtracting from 1 (1 – BAF). This produces a major allele distribution prediction which is then subsequently segmented using PCF with stricter segmentation settings (gamma = 10) to identify allelic imbalances and therefore CNAs.

2.6.2 Mixture modelling approach to predicting major allele distribution

This method is described in Chapter 4 and uses two component Gaussian mixture modelling to predict the major allele distribution, in this Section I will detail supplementary information relevant to the implementation of this method. This approach was applied to four separate datasets in this study as presented in Chapters 4, 5 and 6. These datasets are referred to as MicMa 109K (Chapter 4), WT20 (Chapters 3, 4), PKC66 (Chapter 5) and Hep11 (Chapter 6). Genomic waviness was corrected for as outlined in Section 2.4.1 except for the MicMa 109K dataset which was normalised using the ASCAT methodology [150] (github.com/Crick-CancerGenomics/ascat, accessed September 12, 2017). Only autosomes were analysed using this approach in these datasets as male X chromosomes do not possess heterozygous SNPs. LRRs from each array were normalised to the median and outliers were smoothed according to the CGHcall `normalize()` function [171]. Heterozygous probes were identified either by selecting probes in a matched normal sample with BAF values between $0.3 - 0.7$ or through manual selection in cases without a good quality matched normal sample. The heterozygous probe BAF values are transformed to mBAF using the equation,

$$|BAF - 0.5| + 0.5 = mBAF$$

The mBAF for each array is then segmented using PCF (kmin = 3). For WT and hepatoblastoma datasets that were multi-sampled, segmentation gamma was chosen

51

to be 4.72 as this value is able to identify a pure [2+1] state of 44 SNPs 50% of the time given a BAF *sd* = 0.033 (*sd*, standard deviation). In the MicMa dataset, gamma was reduced to 1 to identify more segments in the single sample study (Table 2.1). A two component Gaussian mixture model is then fitted to the BAFs of each segment using expectation maximisation (EM) and the `normalmixEM()` function in the mixtools R package version 1.0.4 [174]. The starting value for the component means is the mean mBAF and 1 – mean mBAF. The value of *sd* converged upon for each segment is recorded across the array. The values of *sd* are smoothed using kernel density smoothing using the basic R function `density()` with default settings for bandwidth (BW) calculation. The *sd* that produces the highest peak of this density is taken as the representative *sd* of the array and a range either side of the array *sd* is chosen for parameter testing as listed in Table 2.1. The comparison of the phasing and mixture modelling approach in five tumour samples analysed on Affymetrix® Genome-Wide Human SNP 6.0 arrays in Chapter 4 used a distance of $\pm0.01sd$.

Segmentation of the mBAF is then performed twice, on the separate chromosomal arms (split version), in addition to the whole chromosome. These two versions of segmentation are then compared as outlined in Figure 2.1. Firstly the distance of each mBAF value to the mean mBAF of its segment is measured for each probe and summed. The summed distance for the two versions are compared and if the sum distance of the split version minus the sum distance of the original version is greater than 0.58, the original version is retained. This prevents missing segments by splitting at the centromere. If the split version is taken, the distribution of BAFs in the segments which lie immediately adjacent of the centromere are tested to be significantly different using the Kolmogorov-Smirnov test. If the distributions are significantly different ($p<0.05$) the centromere split is retained as the final segmentation output, if not, the original version is kept.

In each segment the BAF distribution is then tested for a normal BAF distribution (mean = 0.5), using the expected normal distribution given the array BAF *sd* and
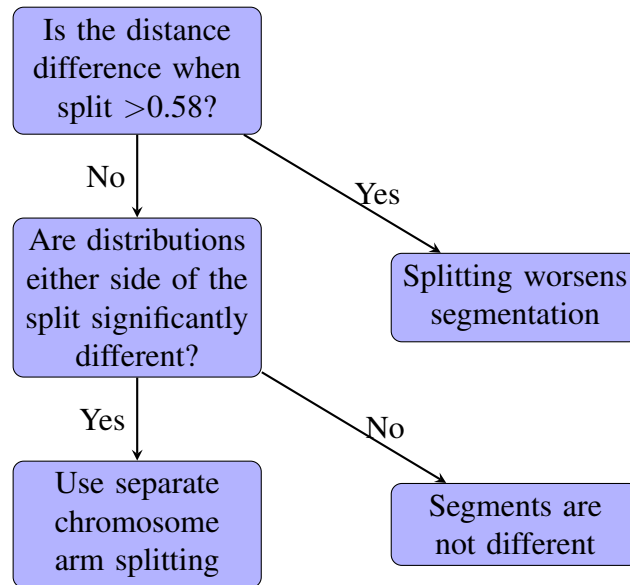
Figure 2.1: The decision tree used for choosing to segment chromosomes as a whole or split by the centromere (segment chromosome arms separately). Firstly, the difference between each mBAF value in the segment and the mean mBAF of the segment is calculated for each probe and summed for both the split and the not split version. If the difference between, the sum of the differences for the split version and the sum of the differences for the non-split version, is greater than 0.58 then the split version is considered to worsen segmentation as the difference between the raw data and segments is too large. If not the raw BAF distributions in the segments immediately neighbouring the centromere in the p arm and the q arm are tested to be different using a Kolmogorov-Smirnov test. If the segments are considered different ($p < 0.05$) then the split at the centromere is taken and the chromosome arms are segmented separately.

using a Kolmogorov-Smirnov test. If the test shows significant difference ($p<0.05$) then the segment proceeds to be modelled as two mixtures. Values for distance from 0.5 (distribution means) are used in the range of $0 - 0.5$ in intervals of 0.01 in combination with the range of *sd*s with a given number of equal intervals (20) to perform a 'global' grid search. The combination that produces the largest log-likelihood is the chosen as the 'global' solution. Next, a 'local' search takes place using the values in the grid immediately neighbouring the 'global' solution using 10 intervals between the *sd* values and 20 intervals for the values of the means of the distributions. The combination of parameters producing the largest log-likelihood in the 'local' search is taken as the final solution for determining the means and *sd* of the major/minor allele distributions.

2.6.3 Predicting copy number state mixing and calculating cancer cell fraction

The median BAF and the mean LRR of each segment are then inputted into the AS-CAT workflow to determine the purity, ρ and ploidy, ψ, of the sample as well as the array gamma (a multiplier that represents the compression effect in the SNP array as described by van Loo and colleagues), γ, using the ASCAT equations that model LRR and BAF [150]. In short this involves calculating major and minor allele copy number states across a grid search of ρ (0.1 – 1) and ψ (1 – 5.5) values. These copy numbers are then measured genome-wide for distance from integer states. I altered this workflow by also performing this across a range of γ values and selecting the grid search with closest solution to an integer state, in order to also select for the most appropriate γ. The grid is then searched for local optima and the combination of ρ and ψ is reported for each local minima. I then selected for the solution within a range of accepted ψ and ρ values that produced the highest ρ, apart from in the MicMa 109K dataset. These parameters were adjusted in particular cases which showed likely tetraploidy. In general diploid solutions were taken unless segment produced combinations of values impossible in a diploid state (e.g. allelic balance in the BAF but a loss-like negative LRR value – a [1+1], diploid, state in a [2+2], tetraploid, genome). The chosen ρ (purity), ψ (ploidy) and γ are then used to model copy number state mixing.

The modelling of mixed copy number states was performed according to the method of Nik-Zainal and colleagues which is an extension of the ASCAT equations and detailed in 'Extended Experimental Procedures' in the following citation [120]. This algorithm is known as 'Battenberg' and was developed as part of a pipeline which utilises phasing (https://github.com/cancerit/cgpBattenberg, accessed September 12, 2017). In short, the closest integer state to the major and minor alleles copy numbers calculated for each segment according to the ASCAT equations. The theoretical mean BAF value of this integer state is then calculated and is tested if it is representative of the distribution of major allele frequency values (two-sided

Approach / Parameter	MicMa 109K	WT20	PKC66	Hep11
Genomic Wave Correction	ASCAT meth.	Sec. 2.4.1	Sec. 2.4.1	Sec. 2.4.1
PCF gamma	1	4.72	4.72	4.72
sd range	$\pm 0.01sd$	$\pm 1/3sd$	$\pm 1/3sd$	$\pm 1/3sd$
γ range	0.43–0.63	0.25–0.5	0.25–0.5	0.25–0.5
ρ range	0.2–1	0.5–1	0.5–1	0.5–1
ψ range	1.6–5.5	1.6–3.4	1.6–3.4	1.6–3.4
Highest ρ solution?	No	Yes	Yes	Yes

Table 2.1: Parameters and approaches taken for datasets analysed by mixture modelling.

t-test, $p>=0.05$). If the theoretical clonal BAF of the segment is considered to be representative the segment is considered to be in the closest integer state in 100% of cells. If the theoretical clonal BAF is not representative of the major allele distribution ($p<0.05$) then two state mixtures of copy numbers are modelled according to the Battenberg methodology [120]. Here, I choose to ensure the first two state solution is between a normal state [1+1] and the closest aberrant state, then remaining combinations as determined by the Battenberg methodology are tested. For each subclonal segment, a series of two state copy number combinations are reported as cancer cell fraction mixtures, the first combination of [1+1] and the closest aberrant state is then taken as the combination with first priority.

2.6.4 Exceptions for higher than diploid ploidy states

Two cases clearly harboured tetraploid genomes, these were IMPORTs 108 and 170 in the PKC66 dataset. Here a larger range of ψ was tested in order to allow for a tetraploid solution ($1.6 - 5.5$) than listed in Table 2.1. To settle on a tetraploid solution in the R2 sample of IMPORT 108, the range of tested ρ values was also expanded ($0.2 - 1$) in this case and R2 of IMPORT 108 reached a tetraploid solution with a purity of 49%. Also for both profiles in IMPORT 108 the best fitting local minima solution was taken, not the solution with the highest ρ.

Owing to the fact that the background in tetraploid cases is no longer a [1+1] state, the two state mixture of [1+1] and an aberrant state is no longer taken as the first option and the the ordering of states is taken entirely according to the

Battenberg methodology.

2.6.5 Clustering copy number alternations to identify subclones

Once the cancer cell fraction of each copy number change has been calculated, each individual array is then analysed to identify clusters of CNAs that may have formed from the same clonal expansion and therefore represent the same clone. I performed this utilising kernel density estimation. Firstly, segments with a mixed copy number state of two aberrations are assumed to have acquired the aberration with a copy number state closer to [1+1] earlier and the state is considered to have been previously clonal and considered to have 100% purity.

Next Gaussian density smoothing is performed using the basic `density()` function in R using a fixed BW of 0.08. Troughs in the kernel density estimation are identified and used as boundaries between groups of CNAs. CNAs within these boundaries are then grouped together and considered to be part of the same 'clone' and the peak density of each of these clones is taken as the cancer cell fraction of the clone.

The first option for a mixed modelling is a combination of a normal state [1+1] and the nearest aberrant state. This can produce estimated cancer cell fractions of the aberrant state that exceed 100%. If an identified 'clone' has a cancer cell fraction greater than 110% the 'second' option is considered for all CNAs belonging in it (first Battenberg option that is not a combination with [1+1]). As this will be a mixture of two 'aberrant' states, the state closest to [1+1] is considered to have once been clonal and the cancer cell fraction is then considered to be 100%. The identification of clones using kernel density smoothing is then repeated and all clone cancer cell fractions will be <110%.

Finally, for each inferred clone duplicated CNA locations are searched for and the CNA that is furthest in distance from the normal state [1+1] and therefore has the lowest cancer cell fraction within the clone is retained whereas the other state

of the segment is removed from the clone. An example of this may be a mixture of [2+1] and [2+2] in which the former cancer cell fraction is 10% and the latter cancer cell fraction is 90%. As [2+1] is closer in distance to [1+1] it is considered to have been the precursor CNA to the [2+2] state and therefore its cancer cell fraction is considered to be 100%. However as the cancer cell fraction of the [2+2] state is 90%, this change may be grouped together with the clonal cluster of changes that are ~100%. Both CNAs cannot co-exist in the same clone and therefore the CNA closer to the normal state is removed.

2.6.6   Inferring phylogenetic relationships in multiple tumour regions with multiple clones

Once clones have been identified within each tumour array across multiple samples, phylogeny is inferred. For an initial interpretation of phylogeny the clones within each sample are considered to be 'nested' even if they can, through virtue of the 'pigeon hole principle', be on separate branches in cases with 3 or more subclones in a single sample. However, the 'pigeon hole principle' was only considered later in the interpretation of the phylogenetic tree. The profiles of nested clones was then condensed into regions of CNAs using CGHregions ($c = 0$) [173]. Two separate inputs were then used for running MEDICC [155], all profiles including 'clonal' and 'subclonal' and only the 'clonal' profiles (CNAs in each sample with the highest cancer cell fraction).

This was performed as MEDICC treats each copy number profile used as input entirely separate and is, in the purest sense, only suitable for interpreting copy number profiles taken as averages from separate samples, as presented in Chapter 3. To this effect it is unable to understand the predetermined temporal relationship between two clones identified in a single sample, i.e. that the clone with the higher cancer cell fraction must have arisen prior to the clone with the lower cancer cell fraction if this clone is considered to be 'nested' inside the higher cancer cell fraction clone. Therefore each phylogenetic tree was produced by required manual assessment of both approaches, using all identified CNA profiles (all 'subclones')

and the phylogenetic tree produced by only the 'clonal' CNA profiles of each sample (the cluster with the highest cancer cell fraction). The MEDICC phylogenetic tree could then be assessed for misordering of nested clones when all CNA profiles were used as input and the MEDICC tree produced using only the CNA profiles generated from the highest cancer cell fraction cluster per sample could be interpreted as a possible 'backbone' of clonal phylogenetic changes and the nested 'subclonal' profile could be mapped on as extension of this backbone.

However, further considerations were also taken when manually assessing the output of MEDICC. These include:

- CNAs affecting largely the same region with clearly different breakpoint boundaries were mostly interpreted as being separate events affecting the same region as opposed to the shortening or lengthening of a pre-existing CNA.

- Breakpoints that were very close but inaccurately identified due to tumour sample impurity were manually defined to be the same if another sample clearly showed a purer version of the same CNA when inspecting profiles.

- Some CNA mixture combinations were re-interpreted if such a change was supported by another sample. Often if the CNA mixture was considered to change the total copy number state but did not alter LRR, it was re-interpreted as a mixture of 2+0 and 1+1, this was often the case for the mis-identification of 11p LOH as a mixture of 2+1 and 1+1 or 2+0 and 2+1.

- If CNAs across a chromosome overlapped they were considered for the possibility of being a mixture of two overlapping CNA states in separate clones.

- The detection of MSAI (a CNA affecting a different allele) was used to separate CNAs as being separate events, even if the breakpoint boundary was equal.

Some sets of CNAs were too challenging to manually assess, such as regions of

58

possible chromothripsis and MEDICC was used to interpret these changes. In some cases WGD makes interpreting CNA evolution difficult (e.g. IMPORTs 108 and 170), and for these cases the interpretation of CNA evolution was carried out by using the results of MEDICC using only the highest cancer cell fraction clones in each sample and then mapping the additional within-sample clones on to the tree.

2.7 Identifying mirrored subclonal allelic imbalance

MSAI was identified in each case by utilising the regions produced following the detection of within-sample clones. Here, for the BAF values of each region of CNA boundaries across multiple samples, k-means clustering was performed using the basic function `kmeans()` in R (centers = 2). A region showing allelic imbalance is one in which the centres of the clusters of BAF values within the region of a single sample show a difference of 0.1, if two samples show allelic imbalances in opposite directions, a region of MSAI is called.

2.8 Copy number alteration evolution features as predictors for event-free survival

Follow up information, in the form of event-free survival, was available for 54 of the 66 cases in PKC66 in Chapter 5. In order to quantify mutational burden across tumour evolution, the number of CNA events was counted for each CNA-only phylogenetic tree split by laterality (bilateral tumours were counted separately). The number of events that was considered representative for the case as a whole was not the sum of the events across the bilateral tumours but the unilateral tumour with the maximum number of events. Subclonality was also calculated for each lateral tumour, as the number of non-clonal CNA events as a percentage of the total number of CNA events, taking the maximum value of the two tumours in bilateral cases to be representative of the case. If two separate tumours were suspected, subclonality was considered to be 100%. Tumours with a subclonality of 100% were removed from the predictive assessment of subclonality.

2.8.1 Receiver operating characteristic curve analysis

Receiver operating characteristic (ROC) curve analysis was performed using the pROC R package [175] using the event-free survival status as the response and the number of events or the percentage of subclonal CNAs as the predictor. The area under the curve was calculated using the `auc()` function in the pROC package. The threshold to split the cases by the predictor was calculated by measuring the part of the ROC curve which is furthest from the function that represents a random predictor ($y = -x + 1$).

2.8.2 Survival analysis

The Kaplan-Meier curves and log-rank test were calculated and performed using the survival R package using the `survdiff()` and `survfit()` functions [176].

2.9 Sequencing data bioinformatic processing and analysis

2.9.1 Removing adapter sequences

FASTQ files for each sample were firstly processed by trimming Illumina® TruSeq Universal Adapter sequences from paired end reads. This was performed using cutadapt [177].

2.9.2 Alignment to the human genome and further processing

Trimmed, paired end FASTQ files were then aligned to the human genome (GRCh37.p13) using bowtie2 [142]. Bowtie2 was run using the settings `--sensitive --score-min L,0,-0.12`. This translates to roughly allowing for one mismatch in every 50bp. Samtools, using the settings `-huF 4` – which outputs only mapped reads – was used to pipe into sambamba which was used to sort the sam file output of bowtie2 and compress the file into a bam format [178]. Sambamba is available online (github.com/lomereiter/sambamba, accessed September 12, 2017). Duplicated reads were then marked using MarkDuplicates which is part of the picard tools software suite (broadinstitute.github.io/picard, accessed September 12, 2017).

2.9.3 Calling point mutations

For each case point mutations were called in each individual tumour sample using the normal tissue sample in each case as a control. Mutations were called using MuTect version 1.1.7 [143] using the dbsnp138 release as reference for the option `--dbsnp` and only performing calling on the targeted genes according to the bait capture design pull-down regions in the option `--intervals`. Failed calls that use the MuTect judgment call 'nearby_gap_events' and 'alt_allele_in_normal' alone were reversed.

2.9.4 Calling insertions and deletions

Indels were called in each tumour sample using the case normal sample as a control and using MuTect2 in the Genome Analysis Tool Kit version 3.6 [146]. The dbsnp138 release as reference for the option `--dbsnp`. MuTect2 was used to analyse the targeted pull down regions of the genes *CTNNB1*, *WT1*, *AMER1*, *DICER1*, *DGCR8*, *TP53*, *DROSHA* and *MLLT1*. Failed calls that use the MuTect2 judgement call 'str_contraction' and 'alt_allele_in_normal' alone were reversed.

2.9.5 Filtering mutations

Point mutations and indels were then only kept if the VAF in the tumour sample was 10X greater than in the normal sample and that there were at least 10 reads reporting the mutation in the tumour sample. Each mutation called in the samples across the whole tumour were then compared in all samples to see if they were present and the VAF of the mutation in all samples is also reported. Mutations were annotated using Oncotator [179].

2.10 Supplementing point mutations into copy number based phylogenies

To incorporate sequence mutations into copy number phylogeny the VAF of the sequence mutation is compared to the local copy number state in order to estimate the cancer cell fraction of the mutation. The prediction of the cancer cell fraction of mutant cells as a proportion of the tumour is performed using the following equa-

tions. As stated in a recent paper by Jiang and colleagues [164], VAF represents mutant cancer cell fraction when a CNA is pure in the following relationship:

$$V = (B*R)/(2*(1-R)+A*R)*M$$

Here and throughout the rest of this section V represents the VAF, R is the proportion of the sample cells that are tumorous, M is the proportion of the cancer cells that possess the mutation, A is the total copy number state and B is the number of alleles within the CNA that possess the mutation.

This equation can be rearranged to make M the subject:

$$M = V*((A-2)*R+2)/(B*R)$$

If the copy number alteration precedes the SNV/indel but is impure in the tumour sample as a whole and is only mixed with a diploid state, the following adjustment of this equation can be made in which the local CNA state is modelled as if it is pure. This is achieved by calculating the proportion of the total sample that contains the CNA:

$$r = R*C$$

Here r is the purity of the CNA as a proportion of the total sample and C is the cancer cell fraction of the CNA as a proportion of the tumour cells, as reported by Battenberg. r can then replace R in the equation to calculate the percentage of the mutant cells, m, as a proportion of the cells carrying the CNA:

$$m = V*((A-2)*r+2)/(B*r)$$

The cancer cell fraction of the mutant cells as a proportion of the tumour cells (M) is then calculated by adjusting m by the cancer cell fraction of the CNA as a proportion of the tumour cells (C):

$$M = m * C$$

If using this method, m, is greater than 1, the mutation is likely to be in all cells possessing the CNA plus additional cells (providing B is correct), therefore it is likely that the SNV/indel precedes the CNA. To assess the mutant cancer cell fraction (M) of a SNV/indel that is in all cells possessing the CNA as well as a fraction of diploid cells in a heterozygous state, the following equation is used:

$$M = (r * ((A * V) - B - (2 * V) + 1) + (2 * V))/R$$

The cancer cell fraction of each mutation is then used to recalculate subclonal grouping in the sample according to Section 2.6.5. The clones are then remapped onto the CNA phylogenetic tree to assess relationships between the clones.

This was performed on IMPORT 54 and 143 in Chapter 5. VAF was derived from output from MuTect for point mutations, however due to the fact MuTect2 filters uninformative reads which alters the VAF compared to the raw data, I derived the VAF for indels from the raw bam file as produced by bowtie2 alignment. Tumour purity was taken as produced by ASCAT in the Battenberg workflow, apart from in R2 of IMPORT 54 where the tumour purity was adjusted to 1 as the purity of the sequenced mutations was much higher than the copy number profile purity (0.59). Number of mutated alleles was assumed to be 1 except for in copy number states in which an alleles was duplicated, then it was assumed that the duplicated allele was the mutated allele in these two cases (B = 2).

2.11 Assessing circulating tumour DNA using digital droplet PCR

2.11.1 Isolating ctDNA and performing ddPCR

Isolation of ctDNA/utDNA (urinary tumour DNA) and the digital droplet polymerase chain reaction (ddPCR) assay was performed by Dr. Taryn Treger in the

laboratory of Professor Kathy Pritchard-Jones. Briefly, extraction of ctDNA from 200μL of plasma, serum and urine was carried out with the Circulating Nucleic Acid Kit (Qiagen®) as per protocol.

For mutation detection in four cases with *TP53* mutations, assays with sequence-specific primers and TaqMan-based probes for mutant and wild type alleles were designed (PrimePCR ddPCR Mutation Assay Bio-Rad for *TP53*®). Twenty microlitres of ddPCR mixture were partitioned into droplets using the QX100 Droplet Generator (Bio-Rad®). Droplets were read on a QX100 Droplet Reader (Bio-Rad®). Three technical replicates were used for each sample. QuantaSoft software (version 1.3.2.0, Bio-Rad®) was used to process the droplet data.

2.11.2 ddPCR data processing

The concentration of positive alleles ([A]) was calculated by counting the number of positive (P) and negative (N) droplets and applying the following equation:

$$[A] = -\ln(1 - (P/P + N))$$

The concentration of the mutant allele ([A_M]) is then calculated via the positive and negative droplets of the mutant probe and the concentration of the wild-type allele ([A_W]) is calculated from the droplet results produced by the wild type probe. The mutant allele frequency (MAF) is calculated as a ratio of the calculated concentrations for each sample at a given time period.

$$MAF = [A_M]/[A_M] + [A_W]$$

2.12 Analysing miRNA expression array data

miRNA expression in WT was analysed using a series of miRNA microarrays (Agilent® Unrestricted Human miRNA V16.0 Microarray) analysed by Ludwig and colleagues [40, 180]. This data was downloaded from GEO database (GSE57370). The miRNAs in the C19MC cluster and cases by histological subtype were subset

from these arrays and hierarchical clustering was performed using the `Heatmap()` function in the R package ComplexHeatmap [181].

Chapter 3

Multiple Wilms' tumour regions reveal genetic heterogeneity

3.1 Introduction

In order to investigate whether important CNA biomarkers in WT are heterogeneous and to place this heterogeneity in the context of evolution, myself and colleagues began a study in which multiple pieces of tumour tissue were taken from a series of 20 WT patients at Great Ormond Street Hospital, London (WT20). WT histological subtypes were not preselected in these patients in order to ensure subtype-specific biases were not introduced (for instance anaplastic WTs are more likely to have chromosome 17p LOH events [38]). The key to this investigation was our attempt to sample as many pieces of tumour as was feasible. This resulted in a range of sampling from 2–6 tumour pieces per patient and allowed for a thorough assessment of CNA heterogeneity within individual patients and across a cohort. Additionally, the larger the number of tumour samples taken, the more detailed the reconstruction of the evolution of the tumour could be resolved through tracking CNAs, as each sample would represent a state in the phylogenetic tree. As WTs can present as bilateral tumours, sampling multiple tumour pieces in WT patients should be thought of in two ways; firstly in terms of the number of samples taken per patient and secondly in terms of the number of samples from each kidney, as bilateral tumours present as separate masses in each kidney (as represented in Table 3.1).

Number of samples	By patient	By kidney
1	0	2
2	3	7
3	11	11
4	1	1
5	3	1
6	2	2

Table 3.1: Number of samples taken in the WT20 dataset per patient (n=20) and per kidney (n=24).

There were multiple aims to this investigation, firstly we assessed how heterogeneous important WT CNAs are. The heterogeneity of CNAs, such as, 1p–, 1q+, 11p LOH, 16q– and 17p LOH, within WT tissue has strong implications on the feasibility of these CNAs as biomarkers that can stratify WT patients for treatment appropriate to their risk of relapse/death, especially in the context of single tumour piece profiling. Therefore our investigation intended to inform those who may design future clinical trials using these CNA biomarkers of the ability they would have to detect the presence of a CNA from a single tissue sample and, retrospectively, may highlight that results from previous investigations may require reinterpretation.

Secondly, it is possible to use CNAs from multiple regions to understand the phylogenetic relationships between different sites in the tumour and to understand how the genome has been altered across time (specifically in terms of ploidy of chromosomes and subchromosomal regions). We aimed to reconstruct the tumour phylogeny in order to gain insight into the evolution of a solid paediatric tumour and the underlying biology influencing this process, as well as clinical presentation in the context of evolution. We were also particularly interested in understanding whether different tumour masses in different kidneys had any phylogenetic relationship, at least in terms of CNAs. We hypothesised that this would help answer the question of whether bilateral WTs masses were related and separated early in embryogenesis, arose as a consequence of metastasis, or whether they were completely separate tumours. Lastly, we compared all phylogenetic analysis to all available

clinical information, including per tumour region histology and available imaging data to associate tumour evolution to defining clinical features.

To assess CNAs in DNA extracted from these tumour pieces we used Illumina® CytoSNP12v2.1 arrays to assess both total copy number and allelic imbalance in each piece to profile somatic CNAs, as well as a matched normal kidney sample and/or a peripheral blood leukocytes sample in all bar one case to assess germline CNAs. My specific role was to generate phylogenetic trees using the processed fluorescence data from these SNP arrays (LRR and BAF) as well as to assess the presentation of the CNAs in general in each patient and to associate the phylogeny of each tumour to available clinical data (tumour imagining, sample site maps and per piece histology). My assessment of tumour phylogenetics using copy number profiles per tumour piece was inspired by work done by Schwarz and colleagues and their phylogenetic algorithm MEDICC which uses minimum evolution of CNAs to infer a cancer's evolution [155].

This chapter was performed in collaboration with the laboratory of Professor Kathy Pritchard-Jones, who generated the presented data. Further details are given in Chapter 2.

3.2 Genetic diversity in Wilms' tumour

I studied high resolution SNP array data from 70 distinct tumour samples from 24 tumours in 20 patients (mean 3.5 samples/case, range 2–6 samples), with matched DNA from non-tumorous kidney and/or peripheral blood leukocytes in 19 cases. Five patients (Cases 9, 10, 16, 17, 20) had bilateral WT, and we obtained samples from both tumours in four of them; in Case 9, the contralateral tumour had been removed prior to the start of the study and was not available for analysis. Patient characteristics and samples are summarised in Table 3.2. To achieve our objectives, I developed a pipeline that processed the raw data from genome-wide allele-specific CNA and LOH events using SNP arrays hybridised with genomic

| Case	Gender	Age	Stage	Subtype	Normal sample
1	F	10	1	S	NK, PBL
2	M	18	2	M	-
3	M	32	3	M	PBL
4	F	58	2	B	NK
5	M	48	2	R	NK
6	M	19	3	M	NK
7	F	10	2	M	NK
8	M	34	2	M	NK
9	F	32	1	M	NK
10	M	36	1 (R), 2 (L)	M (R), M (L)	NK (L)
11	F	20	1	R	NK
12	F	6	1	M	NK
13	F	27	3	M	NK
14	M	12	1	M	NK
15	M	14	2	M	NK
16	M	28	3 (R), 3 (L)	R (R), R (L)	NK (R), PBL
17	F	5	2 (R), 1 (L)	M (R), M (L)	NK (R)
18	F	50	3	DA	NK
19	F	42	1	M	NK
20	M	29	1 (R), 1 (L)	M (R), S (L)	NK (R)

Table 3.2: Clinical information of twenty patient cohort. Patient age is recorded as age in months at the time of nephrectomy and therefore removal of the primary tumour(s). Stage and subtype refer to histology and are noted separately for right (R) and left (L) kidneys. Histological subtypes in the study include; Mixed (M), Blastemal (B), Stromal (S), Regressive (R) and Diffuse Anaplastic (DA). Normal refers to the normal tissue sample taken (if at all) and whether it is normal kidney (NK) or peripheral blood lymphocytes (PBL). In the case of bilateral WTs, the side in which the NK sample was taken is noted.

DNA from each tumour region by first segmenting the LRR signal, which is representative of total copy number, and subsequently the BAF values, representative of allelic (im)balance. This data was used to call allele-specific copy number states per tumour region. The copy number states were compared across these regions and differences between states were considered 'events', changes that must have been acquired as the tumour evolved. Segments were then subject to several stages of quality control to ensure they were reliable (number of probes representing the segment, being present in a region of the genome with unusual SNP coverage [SNP bp^{-1}], see Section 2.4.2). LRR and BAF were segmented separately in order to include all possible events such as CNNLOH that could not be detected in the LRR
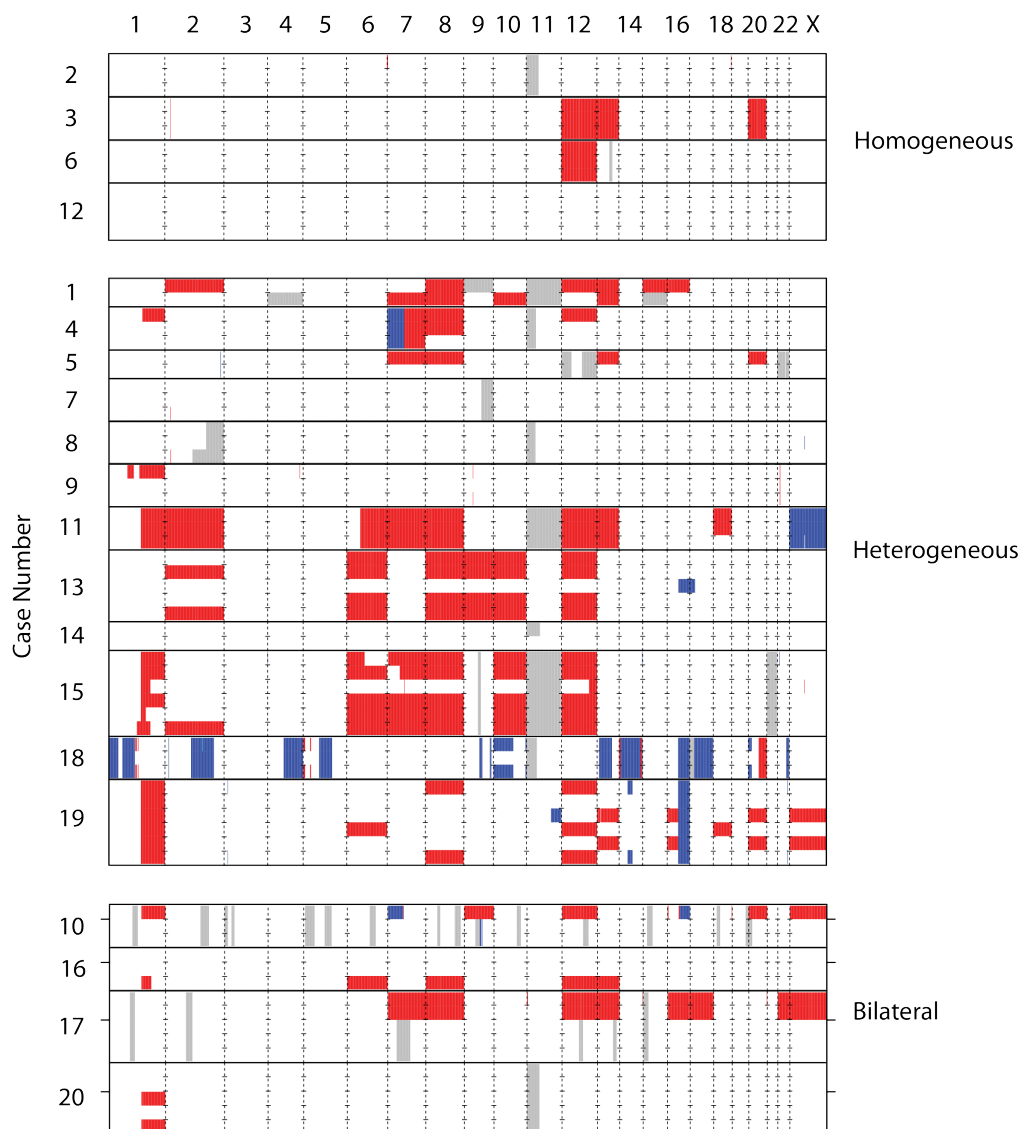
Figure 3.1: Copy number alterations in the WT20 series. Chromosomes are displayed in order horizontally. Patients are grouped vertically into three groups, those with genetically homogeneous unilateral tumours, heterogeneous unilateral tumours (according to the presence of private CNA events in the MEDICC output) and bilateral tumours. These patients are labelled by case number and separated by a bold line. Tumour samples from each patients are separated as rows with hatching at borders between chromosomes. White represents an expected copy number (2 for autosomes and female X chromosomes, 1 for male X chromosomes). Red represents copy number gain and blue represents copy number loss. Grey represents CNNLOH (2+0 state). Contralateral tumour samples in the bilateral cases are separated by a bold tick between the samples.

alone (a 2+0 state produces normal LRR in a diploid genome). In summary, this procedure condenses the genome into regions of copy number changes and pro-
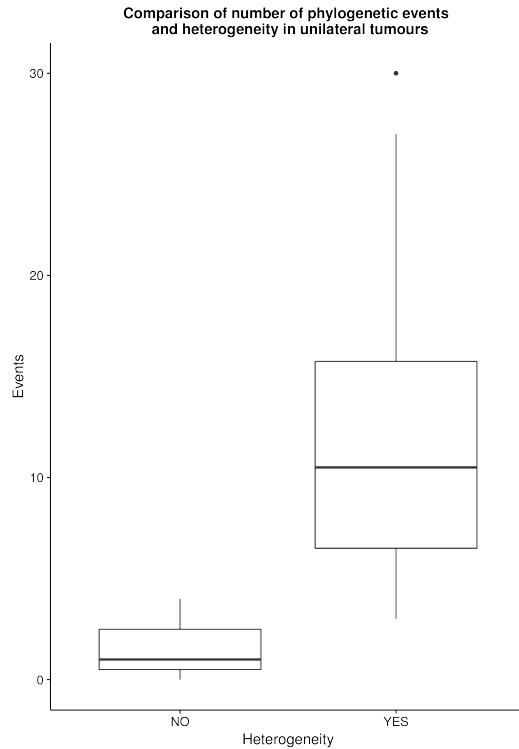
Figure 3.2: A comparison of the number of phylogenetic events as determined by MEDICC in 13 cases and their heterogeneity status (unilateral tumours only). Events is defined as the number of changes required to explain all CNA state differences.

duces a major and minor allele copy number per region for each sample (including the normal sample), by comparing total copy number to the major allele frequency.

Figure 3.1 is a graphical representation of all CNA and CNNLOH events across the 70 tumour samples. We detected most known recurrent WT CNA/LOH, including those associated with poor outcome [55, 56, 57, 58]. Surprisingly, 1q+ was heterogeneous in four of seven (57%) multi-sampled tumours with this change (see Section 3.5). In general, we found remarkable diversity in the extent of intratumour CNA and CNNLOH heterogeneity, ranging from cases with unique CNA/CNNLOH events in all or most samples (such as Cases 1 and 15) to tumours exhibiting no CNA/CNNLOH heterogeneity across all samples. The latter either lacked somatic CNAs/CNNLOH (Case 12 – these samples were confirmed to be tumour through histological assessment) or showed a consistent pattern of somatic CNAs/CNNLOH across all samples (Cases 2, 3, 6), where the single dominant clone in each tu-

mour showed few (0–4) CNAs/CNNLOH events, hinting at a relationship between intratumoural genetic diversity and mutational burden (Figure 3.2). The four patients with these homogeneous (and unilateral) tumours were not statistically significantly younger than the other twelve patients with heterogeneous unilateral tumours (Welch two sample one-tailed t-test, t = -1.52, p = 0.08), suggesting that heterogeneity does not arise purely as a consequence of later age at diagnosis.

3.3   Evolutionary Patterns in Wilms' Tumour

In order to infer the evolutionary history of tumours, I generated phylogenetic trees depicting the relationships between the multiple tumour samples. My pipeline required a minimum of four samples, as dictated by the minimum requirements of MEDICC, and this condition was satisfied in seventeen cases. In thirteen cases, all tumour samples were from the same kidney (unilateral). We obtained a 'flat' phylogenetic tree, where all samples contain the same clone, in Cases 3, 6 and 12. Cases 7 and 9 gave a linear tree, in which clones are derived from 'ancestral' clones sharing all prior evolutionary events. Branched evolution, in which more than one clone contains unique events compared with a common ancestor, was observed in the remaining seven unilateral tumours (1, 4, 8, 11, 13, 15, 18, 19).

Branched evolution is particularly well demonstrated by Cases 15 and 19, with marked branching that began early in the clonal evolution of the tumour. In Case 19 (Figure 3.3), the most recent common ancestor (MRCA) clone observed in sample R4 shows 1q+ and 16q−. All other tumour samples display extra CNAs: R3 and R2 share four additional events (+13, 16p+, +20, +X) and cluster separately from R5, R1 and R6. R2 contains a clone with two additional events (11q− and 13q13.1−) to R3. Interestingly, 13q13.1 loss removes the specific chromosomal region encompassing *BRCA2*, after previous gain of the entire chromosome 13, indicating that the increased expression of the tumour suppressor *BRCA2* as a consequence of +13 was compensated for with this specific loss. R5, R1 and R6 share a +12 event that explains the distinct branching of the samples. However, further branching is seen
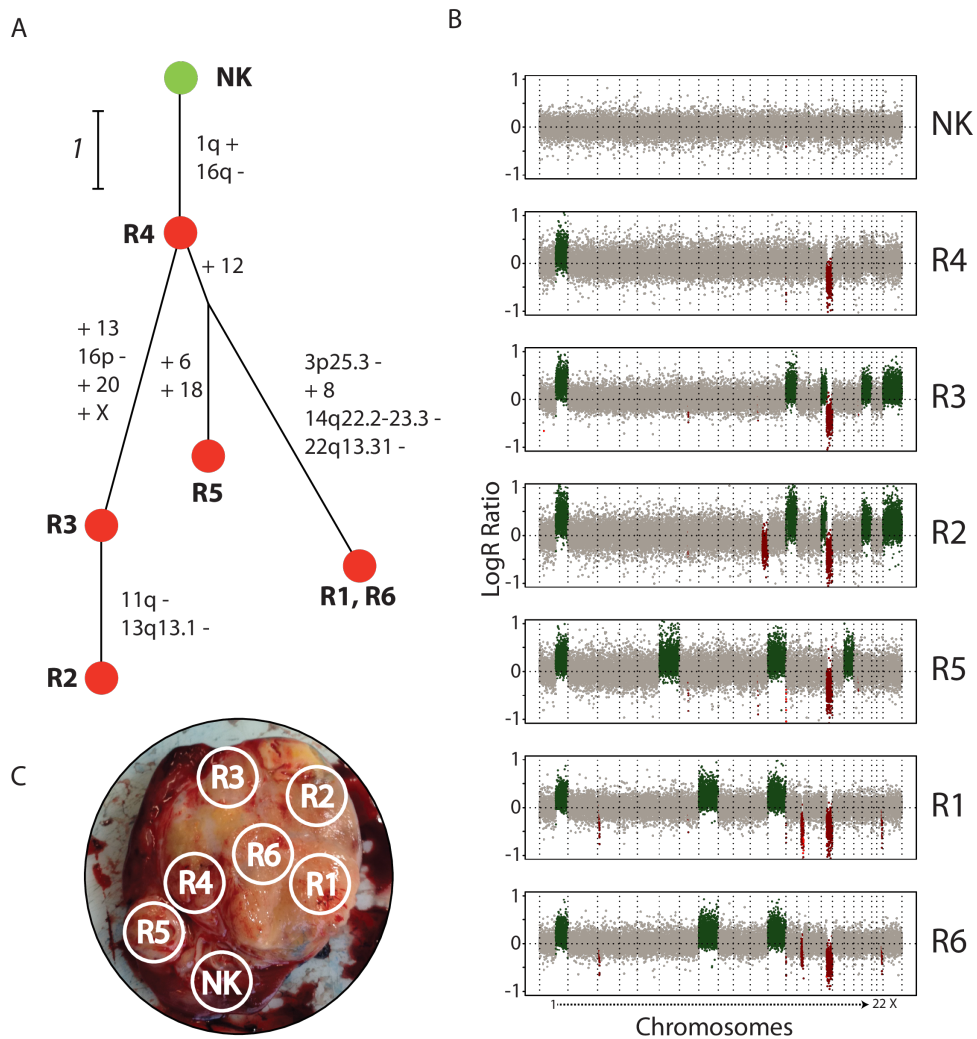
Figure 3.3: Branched evolution in Case 19. A) Phylogenetic analysis of the multiple tumour regions reveals acquisition of CNAs across the evolution of the tumour. The green node represents the NK, germline copy number profile and red nodes represent six tumour regions (R1–6). The edges represent the events explaining the difference between the states with the length being proportional to number of events and the events being annotated next to the edges. B) Genome-wide plots of LRR representing total copy number in the NK and tumour regions. Point colour represents copy number call with grey representing normal, green representing copy number gain and red representing copy number loss. C) Image of the tumour from which the regions were taking with the white circles representing the mapped positions of these samples.

in these samples as R5 contains two unique events (+6, +18), and R1 and R6 both exclusively share four events (+8, 3p25.3–, 14q22.2–23.3–, 22q13.31–). The clone present in R1 and R6 acquired three focal losses late in its evolution that encompass

74

known cancer genes such as *VHL*, a tumour suppressor gene commonly mutated in clear cell renal cell carcinoma (ccRCC) [182], and *SIX1*, a gene that was recently reported to be recurrently mutated in WT [40, 41] (3p25.3– and 14q22.2–23.3–, respectively). Chromosome 22q13.31 contains the gene *WNT7B*, which indicates that this CNA may affect WNT signalling, a commonly dysregulated pathway in WT [183]. Case 15 is even more diversified, with unique events in five of six samples.

3.4 Chromosome 11p15 UPD is an Early Event in WT tumorigenesis

I observed 11p15 CNNLOH in 9 of 20 cases (Figure 3.1). In six of these cases, LOH also involved the entire 11p13 region, which contains the gene *WT1*, and in Case 18, it involved part of the 11p13 region, including the *WT1* locus. In Cases 4 and 8, 11p LOH did not involve 11p13. In all but Case 14, 11p15 CNN LOH was detected in all tumour samples, indicating that it occurred in a common ancestor of all observed clones in these tumours, and is thus a consistent early event in WT tumorigenesis. The observation that 11p15 CNNLOH was a truncal event in the evolution of these WTs was confirmed by phylogenetic analysis. In Case 14, 11p15 CNNLOH was called in just one of the two samples taken from this tumour, where it was the only CNA observed (Figure 3.1). Visual inspection of the chromosome 11 BAF plots for this case shows that, despite not being identified in one of the samples through calling allelic imbalance by detecting sustained absence of balanced SNPs, the event does appear to be present in a low fraction of cells, indicating a detection limit in my approach (Figure 3.4). Therefore, Case 14 does not contradict the evidence from other cases that 11p15 CNNLOH is a consistently early event in WT tumorigenesis.

Since chromosome 11p15 contains a cluster of imprinted genes, including the *IGF2* oncogene, we performed MS-MLPA on the 11p15.5 locus in all cases with 11p CNNLOH, and in all tested samples found UPD, as indicated by hypermethylation of the H19 differentially methylated region (DMR) (consistent with overexpression
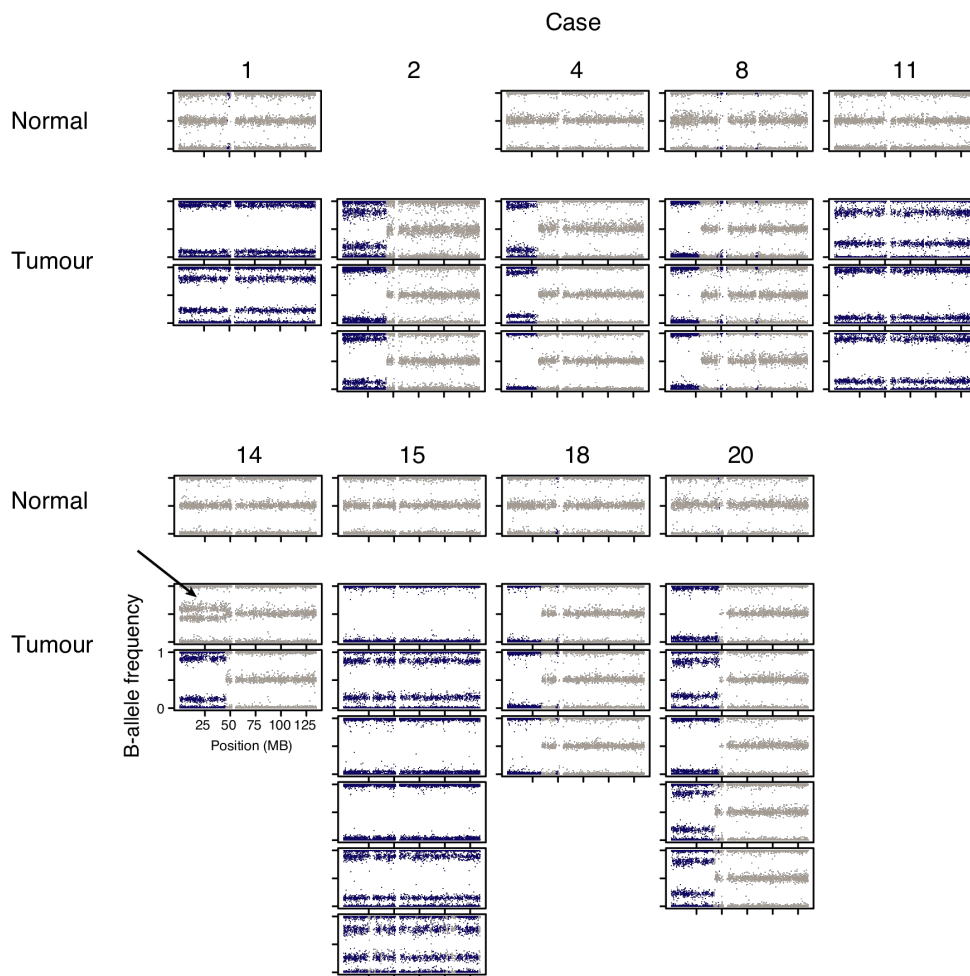
Figure 3.4: Chromosome 11p LOH is an early event in Wilms' tumorigenesis. Chromosome 11 wide BAF plots for nine cases in which chromosome 11p LOH occurs show the event to be present in all tumour samples yet absence in the normal sample (unproven in Case 2 due to a missing normal samples). The x-axis represents position along the chromosome in MB. Points coloured in blue represent SNPs called as being in a state of allelic imbalance whereas grey points represent SNPs called to be normal. The arrow highlights a chromosome 11p LOH event in a tumour sample of Case 14 that was not called as being in imbalance due to being present in a low number of cells. For reference, the centromere in chromosome 11 is present at 53.7 MB.

of *IGF2*) and hypomethylation of the KvDMR, (i.e. a paternal pattern) [169, 184]. DNA from adjacent normal kidney was available for eight of the nine cases with 11p15 CNNLOH, and in all eight cases, there was neither 11p CNNLOH nor abnormal methylation, indicating that the CNNLOH and methylation abnormalities were somatic events in the tumour cells. This indicates that methylation changes caused

by CNNLOH in 11p15.5 leading to a paternal methylation pattern are crucial steps in the early stages of tumorigenesis in a subset of WTs.

In contrast, MS-MLPA on the cases without somatic 11p15 CNNLOH showed five cases (Cases 6, 10, 13, 17, 19) with hypermethylation of the H19 DMR only, with normal methylation of KvDMR. In these cases, the abnormal methylation pattern was homogeneous across all tumour regions and was also present in adjacent histologically normal kidney. Case 6 had known hemihypertrophy, but the other cases had not been diagnosed with Beckwith-Wiedemann syndrome, hemihypertrophy, abnormal growth, macroglossia, hypoglycemia or other tumours. This suggests that the hypermethylation of the H19 DMR may be a mosaic event in these patients that precedes tumour formation. Hypermethylation of the H19 DMR, and therefore *IGF2* overexpression, was an early event in 14 of the 20 patients investigated, suggesting a crucial founding role in WT formation. This finding also builds on the previous observations of somatic 11p CNNLOH in WT precursor lesions (nephrogenic rests) [185].

3.5 Gain of Chromosome 1q Shows Variable Timing

We observed 1q+ in eight patients (40%, Figure 3.5). For seven of these we had multisampled the tumours in which we detected the change. 1q+ is present in all the tumour samples in three of these cases (Cases 11, 15, 19). In Cases 11 and 19 the same 1q+ is in all samples. In Case 15 there is gain affecting all or part of chromosome 1q in all tumour samples (n = 6); there are several unique CNAs that affect this chromosome arm, thus displaying intra-tumour genetic heterogeneity (ITGH) in the extent of 1q+ itself, indicating that the event occurred several times. In the remaining four cases, 1q+ is present in one of three (Cases 4 and 9), one of two (Case 16, right kidney tumour) and two of three (Case 20, right kidney tumour) samples. This indicates that, if we had single sampled our tumours, we would have probably only been certain of identifying 1q+ in three of seven tumours; the probability of finding 1q+ in all seven tumours with single sampling is only 0.037.
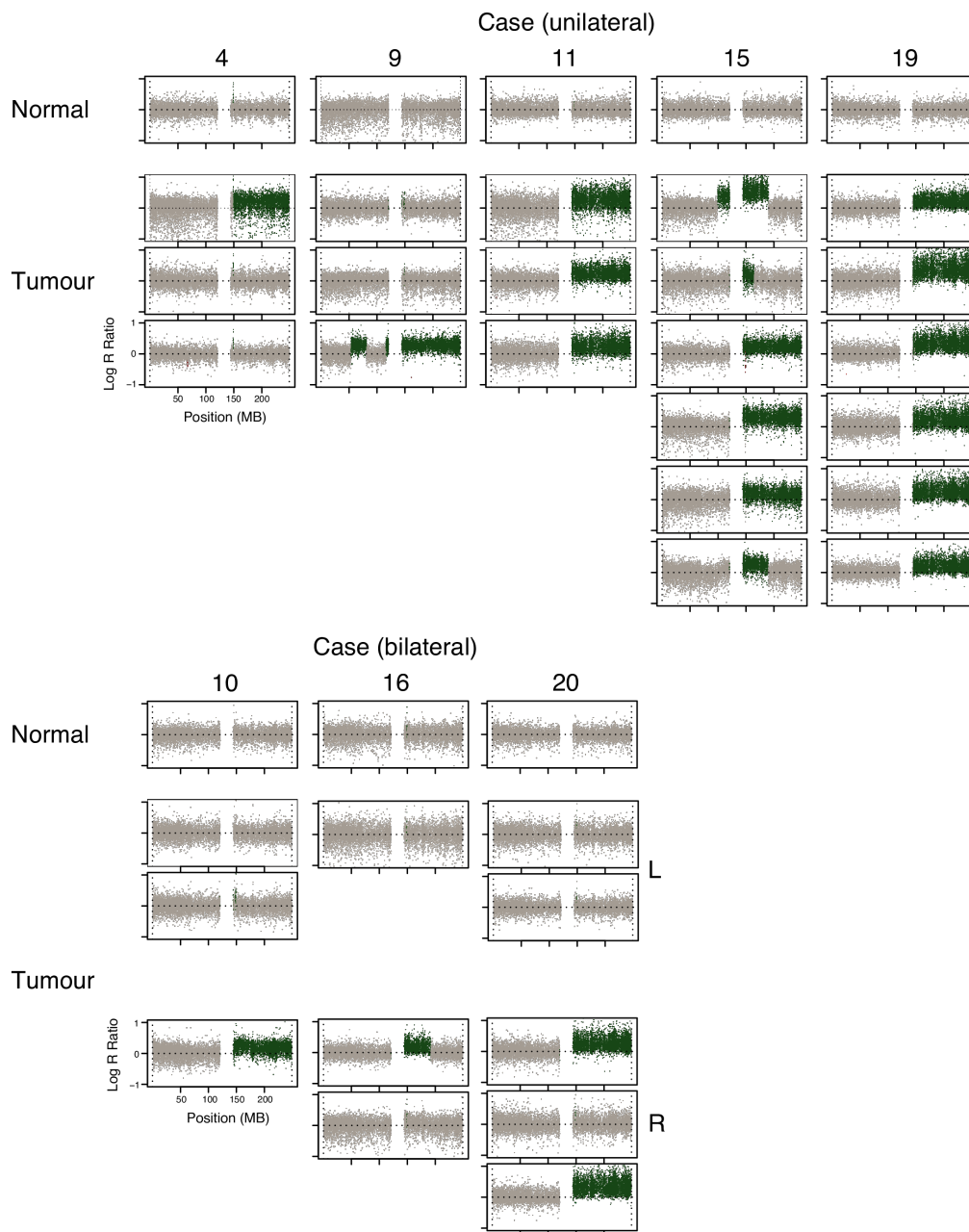
Figure 3.5: Heterogeneity of chromosome 1q gain in WT. Chromosome 1 wide LRR plots for the eight patients in which we observe 1q+, split into the five patients which were sampled from a single tumour and three patients where samples were taken from two bilateral tumours, separating out the individual left (L) and right (R) tumour samples. The x-axis represents position along the chromosome in MB. Points which are grey represent SNPs called as having a normal copy number, green points represent copy number gain and red points represent copy number loss. Normal sample chromosome 1 LRR is displayed for reference. The centromere in chromosome 1 is present at 125 MB.

Indeed, the average probability of obtaining a negative single sample per tumour in our set of seven multi-sampled tumours with 1q+ is 0.31 (95% confidence interval is 0.08–0.54), despite 1q+ being present in all samples obtained from three of the seven positive tumours. The cumulative binomial distribution with p = 0.69 indicates that at least three samples per tumour must be studied in order to ensure that greater than 95% of tumours with 1q+ are detected, assuming a uniform probability of detecting the change in each tumour.

In the four tumours which display 1q+ ITGH, 1q+ shows no preference in evolutionary timing. In Cases 9 and 16 chromosome 1q+ timing is indistinguishable from other copy number changes present in the tumour because samples from these tumours represent either the only clone with tumour-specific CNAs or have a copy number profile that is indistinguishable from matched germline DNA. In Cases 4 and 20 1q+ follows other CNAs (e.g. UPD of chromosome 11p in both cases) indicating that 1q+ evolved late. In contrast, in Cases 11 and 19, where 1q+ is homogeneous with identical breakpoint boundaries across the samples, it is a truncal event that occurred early in tumour evolution.

3.6   Rarer Biomarkers

Other WT biomarkers are difficult to assess without larger patient cohorts. Nevertheless, I observed 16q– in four tumours, three of which are multi-sampled, one of which was a single sampled right kidney mass in Case 10 (Figure 3.1). Of the multi-sampled tumours, Cases 18 and 19 show homogeneous 16q–, accompanied in Case 18 by homogeneous 1p– and 4q–. In Case 13, 16q– is apparently heterogeneous, but the change is present in the one sample that is from a tumour that probably originated independently in the same kidney (see Section 3.8); therefore, we have no evidence to support 16q– as a heterogeneous event. Two cases contain a LOH event that affects the *TP53* locus: homogeneous 17p CNNLOH in Case 18 and apparently heterogeneous loss of *TP53* in Case 13 occurring in the single sample from the tumour with probable independent origin, again indicating a lack of evidence
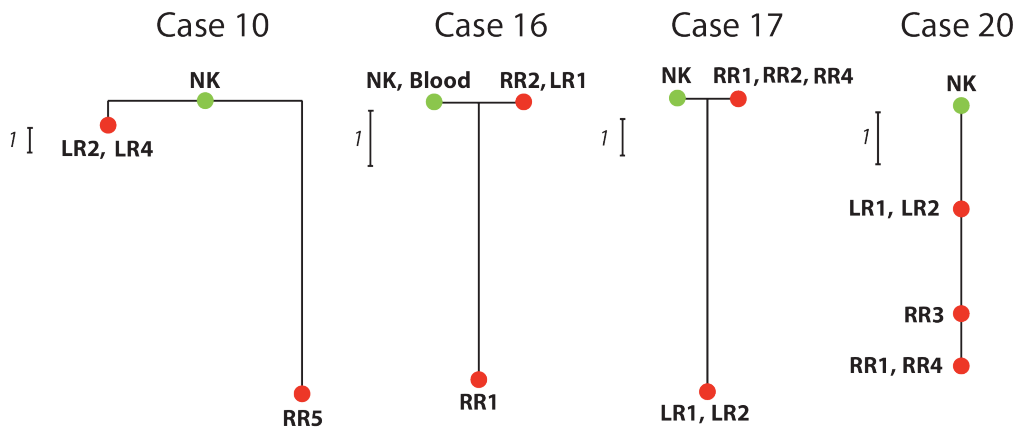
Figure 3.6: Phylogenetics of the four bilateral WT cases as interpreted by MEDICC. Germline copy number profiles are represented as green nodes (blood and NK) and tumour regions are represented as red nodes. Region numbers are preceded by L (left) or R (right) to state the laterality of the tumour sample (e.g. RR1, right region 1). Edges are not weighted in the horizontal axis, this is used to separate nodes of the same copy number profile. Edges are weighted by phylogenetic events (copy number changes) vertically and the scale for each case is displayed using a key at the left of the tree.

for heterogeneous *TP53* loss. Case 18 showed diffuse anaplasia on histological examination, whereas Case 13 showed neither diffuse nor focal anaplasia or nuclear unrest, instead the tumour region with 17p LOH was dominated by epithelial structures. Focal *MYCN* gain is heterogeneous in Cases 7 and 8 and homogeneous in Cases 3 and 16. *MYCN* gain is germline in Case 16 (previously reported [118]).

3.7 Bilateral tumours

Five cases in the dataset are bilateral WTs, and in four of these both sides were sampled (Cases 10, 16, 17 and 20, Figure 3.6). Overall, bilateral tumours appear genetically distinct and probably arose independently. Cases 10 and 17 provide the clearest examples of the striking differences in copy number profiles between contralateral tumours. In Case 10 the only tumour-specific CNAs in the two samples taken from the left tumour are two focal deletions in chromosome 9, whereas the sample from the right tumour shows 1q+, +9, +12, +20, +X, 7p– and 16q–. In Case 17, the left tumour of Case 17 shows +7, +8, +12, +13, +16, +17, +22 and +X, while the right tumour shows no tumour-specific CNAs. In both these cases,
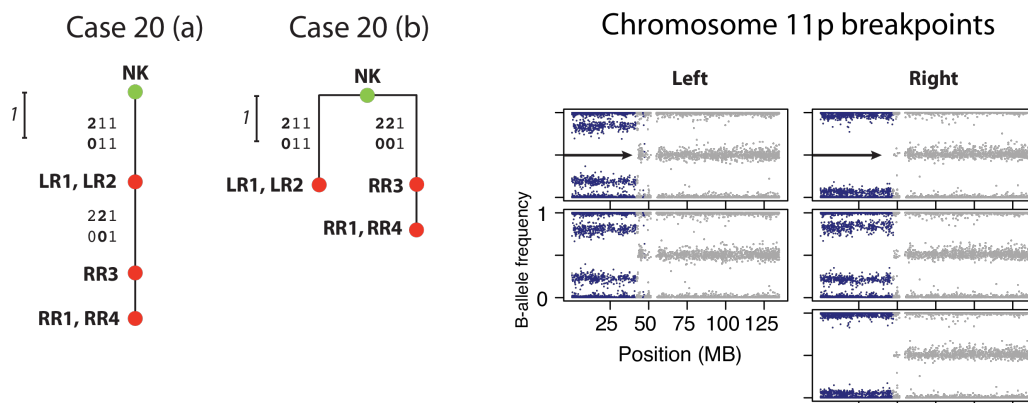
Figure 3.7: Multiple interpretations that are equally parsimonious of the phylogenetic events in Case 20. Two interpretations of the same tumour regions of Case 20 are displayed on the left (Case 20 [a] and Case 20 [b]). The green node represents the germline copy number profile (NK) and the red nodes represent the tumour regions, each tumour region is preceded by L (left) / R (right) to indicate tumour laterality of the tumour region. Edges are weighted vertically by phylogenetic events. The grid of numbers to the left of the edges represents the changes that occur to the regions representing chromosome 11 in these two interpretations, where numbers in **bold** have changed from the preceding state (all normal in this case). Chromosome 11 BAF plots of the tumour regions taken from the left and right tumour are displayed on the right of the figure. The x-axis displays chromosome position in MB. Points coloured in blue represent SNPs which are called to be in state of allelic imbalance, grey points represent normal allelic balance regions. The arrows highlight the group of SNPs which have differential allelic imbalance across the two regions, highlighting the different breakpoints of this chromosome 11p LOH across the left and right tumours.

the two contralateral tumours share no tumour-specific CNAs and are not heterogeneous within the individual kidney. Case 16 only showed CNAs in one sample of two taken from the right tumour, thus making the right tumour heterogeneous. The left tumour sample of Case 16 did not have any CNAs.

Case 20 was heterogeneous within the right tumour, with 11p CNNLOH in all three samples but 1q+ in only two samples. The left tumour also contained 11p CNNLOH, but with a different breakpoint boundary on the centromeric side, indicating this may have been a separate unrelated event in the left tumour (Figure 3.7). The UPD event is allele specific in the same way in both tumours (loss of maternal, gain of paternal). However, we cannot formally exclude the possibility that
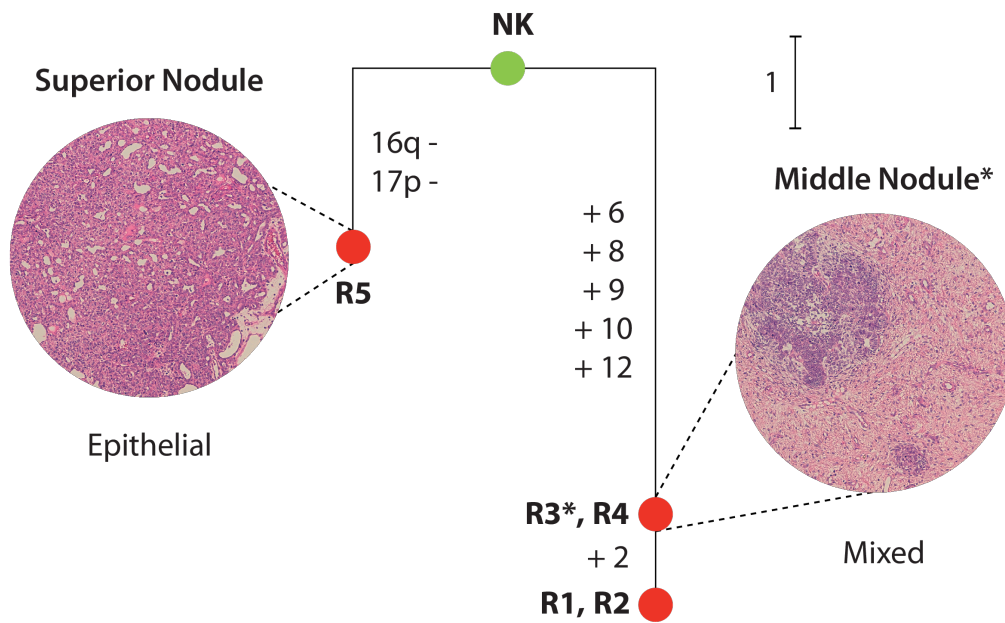
Figure 3.8: Evidence for two tumours in a single mass in Case 13. Phylogenetic analysis reveals two genetically distinct WTs with tumour regions taken from the superior nodule (R5, red) and middle nodule (R1–4, red) as their only ancestral state is the copy number profile of the normal kidney (NK, green). Specific CNA are noted adjacent to the edges of the tree. Histology of the tumour nodules is displayed at 10X magnification for R5 (superior nodule) and R3 (middle nodule). The histology of the middle nodule is triphasic, with blastemal, epithelial and stromal elements (i.e. mixed), whereas the superior nodule, represented by R5, is markedly different and composed exclusively of more mature epithelial elements.

the tumours in the two kidneys developed from a single clone with the shorter 11p CNNLOH, showing no further evolution in the left tumour and extension of 11p CNNLOH in the right kidney. The chosen phylogenetic tree algorithm MEDICC outputs this latter interpretation, however a phylogeny in which the tumours evolved separately is equally parsimonious to this solution.

3.8 Evidence for Two Separate Tumours in the Same Mass

Case 13 presented as one tumour mass consisting of a larger middle nodule, presenting mixed blastemal, epithelial and stromal elements, contiguous with a superior nodule composed of relatively well-differentiated epithelial structures, thus show-

ing two contrasting WT histological phenotypes in the same apparent tumour. Four samples were obtained from the middle nodule and one from the superior nodule. The samples from the middle nodule were heterogeneous due to the presence or absence of +2, with homogeneous +6, +8, +9, +10, +12 as clonal CNAs. The sample from the superior nodule contained 16q– and 17p– and none of the CNAs in the middle nodule. Thus, the two nodules showed completely independent branching from germline DNA, suggesting that they arose independently, based on copy number and histological data, despite appearing as one contiguous tumour (Figure 3.8). Phylogenetic reconstruction was necessary for interpreting the clonality of the 16q– and 17p– events (see Section 3.6), showing that heterogeneity must be interpreted in the context of evolution. By applying this approach we have uncovered evidence of independent origins of two synchronous WT, not only in bilateral cases, but also within the same kidney containing an intratumoural nodule with divergent histology.

3.9 Relation to Treatment Response

Case 8 also presented with a large intrarenal mass contiguous with a smaller nodule (Figure 3.9A). Following neoadjuvant chemotherapy, however, the smaller nodule showed a greater shrinkage (78% vs. 42%, Figure 3.9B) and increased ADC (0.71 to 1.71 vs. 0.73 to 1.32, Figure 3.9C) than the main tumour mass, indicating a better response to chemotherapy due to greater diffusion of water molecules in the tissue and a disintegration of tissue structure [168, 186]. Histologically, all areas of the tumour showed the same, mixed histology of blastema, epithelia and stroma.
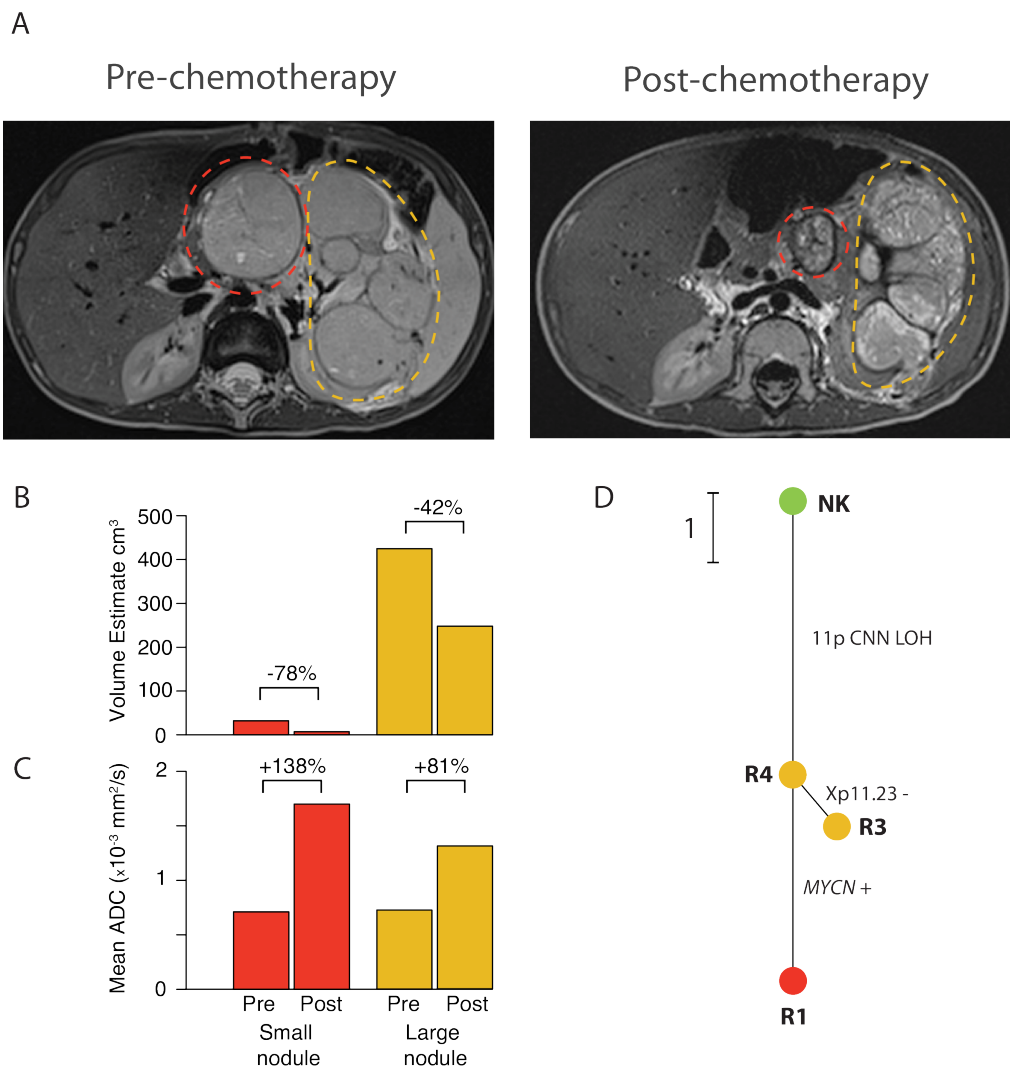
Figure 3.9: Tumour phylogenetics and chemotherapy response in Case 8. (A) Pre- and post-chemotherapy T2-weighted MR images show the two masses indicated by red (small nodule) and orange (large nodule) dashed lines. (B) Estimated tumour volumes pre- and post-treatment shows that the small nodule (red) shrank by 78% and the large nodule shrank by 42% (orange). (C) The small nodule (red) showed a greater post-chemotherapy gain in the mean assessed diffusion coefficient (ADC) than the large nodule (orange), indicating a better response to chemotherapy. (D) A phylogenetic tree shows the genetic relationships between samples from the large nodule (R3–4, orange) and the small nodule (R1, red). The normal kidney sample is represented by a green node (NK). The small nodule is related to the rest of the tumour, but has evolved additional changes, including focal gain of *MYCN*.

We obtained one sample from the smaller nodule (R1) and two from the larger mass (R3, R4); these samples display ITGH and their phylogeny shows that R4 represents a clonal ancestor of the other two samples. R3, also from the larger mass, evolved an additional Xp11.23–, while the sample taken from the smaller nodule (R1) developed 2q CNNLOH and a copy number gain of the oncogene *MYCN*. As *MYCN* amplification is known to result in increased cell proliferation, it is possible that this nodule specific CNA causes this smaller nodule to become more sensitive to neoadjuvant chemotherapy.

3.10   Concluding Remarks

In this chapter I presented results of my initial approach for automatically reconstructing tumour phylogeny in an initial pilot series of 20 WT cases, surveying CNA heterogeneity and evolution patterns across our cohort. I was able to demonstrate that in a subset of patients, even in tumours sampled up to six times, unique clones can be detected in nearly every sample of a WT. Patients such as Case 19 exemplify the large genetic diversity present in a single WT. This chapter also demonstrates that, not only can this diversity be detected in multiple samples, but can also be utilised to infer the ordering of CNA acquisitions as the multiple WT clones develop through branched evolution. Six tumour samples represents the greatest number of WT samples assessed in an evolutionary study, yet demonstrates that in these WT patients the limit to gaining additional information, in respect to genetic diversity, remains undetectable, suggesting more unique clones may be identified if sampling was increased further.

This analysis highlighted key issues produced by CNA biomarker heterogeneity that are relevant for clinical trials, as demonstrated by 1q+. Two recent studies assaying 1q+ as a prognostic biomarker have both shown the prevalence of the CNA to be 28% in their independent cohorts, based on just a single tissue piece per tumour [104, 103]. Interestingly, if the probability of obtaining a negative sample in a tumour that contains at least one sample with a 1q+ is 0.31 (as proposed in Section

3.5), single sampling our cohort of 20 patients would produce a 1q+ prevalence of ~28%. Exemplifying that despite our low number of patients, our cohort may be representative of 1q+ prevalence in WT and therefore also representative of 1q+ heterogeneity, suggesting that ~17% of WTs considered 1q+ negative in these studies may be 1q+ positive. It is important to recognise that despite closely matching 1q+ prevalence, these results are only based on 7 multisampled WT with variable sampling (2–6 samples). Furthermore, a uniform probability of detecting a negative sample in each 1q+ positive tumour is not likely to be a true representation of detection probability. It is likely that a subset of patients possess a clonal 1q gain and others posses subclonal 1q gain. This suggests that there are multiple modes of detection probability based on the evolutionary histories of patients in the cohort. The probability of detecting a negative sample in a patient with a truly clonal 1q gain is 0, yet the probability of detecting a negative sample in a tumour with subclonal 1q+ is dependent on the timing of the event per tumour. Larger studies with consistent sampling strategies are required to understand the true distribution of 1q+ detection.

This study showed for the first time in the context of tumour evolution, in a large number of patients, that 11p15 LOH is an early mutation event in the development of WT. This finding strings together many other indications that 11p15 LOH is an early event, such as the common presence of 11p15 LOH in a precursor lesion of WTs, nephrogenic rests, and the fact that 11p15 LOH is detected in a large portion of WT patients (suggesting detection is not limited by subclonality) [185, 187]. This is further supported by the observation of 11p15 LOH as an early event in other paediatric solid tumours [113, 114].

Additionally, I have highlighted that WT patients have the capacity to develop multiple tumours, with no evidence for evolutionary relationships between bilateral tumours as well as being able to suggest the presence of two tumours in the same tumour mass in Case 13. Case 13 highlights that multicentric WTs may be under-recognised and not treated appropriately despite this being only a single case

in the study. Additionally, the genetic independence of bilateral WTs may prove to be a defining feature of these synchronous tumours, although only four patients with bilateral WT were assayed from both tumours in this study. Assessing genetic independence in these patients is complicated by the absence of CNAs in samples taken from Cases 16 and 17. As samples with no somatic CNAs are equivalent to the root of the phylogenetic tree, they can be interpreted as being a common ancestor to all/none of the samples, making them uninformative to tumour sample relatedness.

Finally, by comparing tumour phylogenetics to data such as tumour imaging and physical characteristics (i.e. volume and ADC), I have been able to associate tumour phylogeny with clinical features such as response to treatment. Although just a single case, a heterogeneous *MYCN* gain may be associated with increased treatment response in the subclones possessing the alteration in Case 8. This study demonstrates that associating differential, intratumour treatment response with genetic diversity is possible and may allow us to understand why some regions of a tumour respond better to therapy than others.

In this chapter I have presented a detailed picture of WT evolution. However, there are several limitations to this study. Firstly, the study assumes the copy number profiles determined for each tumour sample are not subject to the mixing of clones. Additionally, the methodology in this chapter firstly segments the LRR and then incorporates the CNAs only detectable in the BAF (CNNLOH/UPD) as a second, latter step. A singular approach that is able to detect CNNLOH/UPD alongside other CNAs that alter LRR would allow for a more concise approach to CNA detection. The specific detection of CNNLOH/LOH developed for this chapter (detailed in Section 2.4.4) was unable to detect an impure 11p15 UPD event in Case 14 as shown in Figure 3.4. An improved approach to CNA detection should be able to detect impure changes. Finally, although 20 cases is a large enough cohort to describe WTs generally, it is clear that different WT subtypes may display different

patterns of clonal evolution. For example Case 18, the only diffuse anaplastic case in this series, largely displays copy number losses despite the majority of cases in this study being dominated by copy number gains. A larger cohort of patients would increase the amount of evidence for subtype specific evolutionary patterns.

These findings describe the spectrum of WT evolution, from homogeneous tumours dominated by single clones, to genetically diverse tumours, and helped me to conceptualise the process of evolution in WT. This study highlights key characteristics of the evolution of this solid paediatric cancer. Additionally, this work provided a context for meeting the challenge of understanding CNA evolution across multiple tumour SNP array samples.

Chapter 4

Detecting and Reconstructing Clonal Mixtures in Allele-Specific Copy Number Profiles

4.1 Introduction

Assessing multiple regions of a tumour is a powerful approach to determine the evolutionary history of a cancer. It is clear that by assessing tissue from several sites across a tumour, a much broader assessment of tumour genetics is achieved and a much larger collection of genomic variations are identified, compared with assaying just a single sample. However, it is also apparent that each sample still represents a large mixture of tumour cells that produce heterogeneity and the fundamental problem that is presented when assaying clonal architecture using single samples still affects multiregion studies. Ignoring the fact each sample from a multiregion assay represents a mixture of cells may reduce the information gained from these studies, as well as bias the results to describe genetic variations across the tumour as opposed to within local tumour regions that make up a single tumour sample. Therefore, to gain a complete insight into tumour evolution we must assay genetic variation both within a tumour sample and across several tumour samples.

To address the heterogeneous nature of a single tumour sample in the context

of CNAs we must assess the underlying mixed population of tumour cells. To achieve this it is important to estimate the fraction of tumour cells – the cancer cell fraction – harbouring each CNA individually and remove the limitation of assuming a CNA profile is homogeneous across all genomes in a single sample population. As noted by Nik-Zainal and colleagues [120], this requires an accurate estimation of the BAF for each segment by determining the separation between the major and minor allele distributions. An approach to achieve this must balance the detection of true separation with ensuring normal segments are identified correctly in which these distributions are identical.

Nik-Zainal and colleagues approached this problem by directly predicting the physical chromosomal linkage of SNPs in each patient based on their genotype. This was achieved by phasing the genotype using an imputation algorithm called impute2 [154] and by comparing this computationally predicted phasing to the BAF distributions derived from the patient to identify the major allele distribution. This has been performed using data from high-density Affymetrix® SNP6 arrays as well as in whole-genome sequencing experiments [120, 136]. However, it is unclear if this approach can be performed on inexpensive SNP arrays with lower tiling densities and therefore a reduced information content on the genotype of the patient.

Here I developed an alternative approach to contrast phasing by identifying the separation in major and minor allele distributions using a two component Gaussian mixture model. This chapter will focus on comparing the performance of this mixture model approach on high-density Affymetrix® SNP6 arrays to the 'gold standard' of phasing the genotype. I will then present results of using this approach on a low density array type that probes 9X fewer SNPs to identify possible subclonality in a dataset of breast carcinoma patients. Finally this chapter will focus on comparing mixture modelling to phasing in the dataset presented in Chapter 3 to assess which approach is more suited to the 'medium' density Illumina® CytoSNP arrays. The advantages gained from understanding within sample subclonality in

SNP array name	Array size (SNPs)	Density
Affymetrix® Genome-Wide Human 6.0	901,095	High
Illumina® HumanCytoSNP-12 v2.1	288,852	Medium
Illumina® Human-1 109K BeadChip	109,302	Low

Table 4.1: This chapter will focus on data from three different arrays, one produced by Affymetrix® and two produced by Illumina®. For each array the number of SNPs probed is noted (array size) and the relative density is also presented as will be referred to in this investigation. The number of SNPs represents tiling across chromosomes 1 to 22 and X.

multi-region datasets such as WT20 will be expanded upon. The various array types described in this Chapter are described in Table 4.1.

Previously published datasets were kindly made available to me by Dr. Peter van Loo and his lab to perform these comparisons [136, 150].

4.2 Mixture modelling allows for the detection of allele-specific BAF distributions

In order to generate accurate estimates of copy number cancer cell fraction, it is important to estimate BAF accurately as this value is representative of copy number state and the proportion of cells in which the state is present. For states of allelic imbalance with high purity this task is often trivial and can be performed by calculating the mean mBAF:

$$|BAF - 0.5| + 0.5 = mBAF$$

Calculating the mBAF only produces accurate mean major allele BAF values if the lower tail of the major allele distribution does not produce values in the range of values representative of the minor allele distribution $(0 - 0.5)$. However if an allelic imbalance is present in a low number of cells (a low cancer cell fraction or if the tumour itself is impure) the lower tail of the major allele distribution and the upper tail of the minor allele distributions may overlap and the mean mBAF is no longer representative of the major allele BAF. Furthermore, calculating mBAF for

each segment prevents the accurate detection of normal allelic balance in which the major and minor allele distributions are identical.

One approach to this problem developed by Nik-Zainal and colleagues [120] is to phase the genotype of the patient. Here, using a genotype imputation algorithm such as impute2 [154] and a dataset such as the 1000 Genomes Project as a reference panel representative of the human population [153], a prediction is made of the physical linkage between bases in each SNP that is, which base of the SNP lies on the maternal chromosome and which base is present on the paternal chromosome. Once the prediction has been made, if the phasing is completely accurate across the genome, the major and minor alleles would have been identified. However, due to the large number of SNPs probed in each chromosome and the variation in the human population, phasing the entire chromosome perfectly is practically impossible. Instead, phasing may accurately predict physical linkage correctly for a short series of adjacent SNPs across the chromosome. These are referred to as 'haplotype blocks'. Taking the BAF values of SNPs that are predicted to lie on one physical chromosome, haplotype blocks can be identified through segmentation even if these haplotype blocks are located interchangeably on the major and minor alleles. The larger the haplotype blocks, the easier it is to identify these correctly using segmentation. Haplotype blocks with a mean BAF <0.5 are considered to belong to the minor allele and all SNPs within this block are transformed by subtracting from 1 $(1 - \text{BAF})$ to generate a predicted major allele distribution. This technique serves to preserve the lower tail of the major allele distribution and allows for accurate BAF values to be generated.

If phasing accurately detects haplotype blocks, it demonstrates an ability to detect very small cancer cell fraction CNAs, however it has only been shown to be effective thus far in whole genome sequencing and high density SNP arrays [120, 136]. It may not be possible to phase lower density SNP arrays accurately due to the reduced information content of the array (fewer probed SNPs) meaning less

data is available for genotype prediction. To address this problem and to provide an alternative to phasing, I developed an approach which makes use of a two component Gaussian mixture model to identify major and minor allele distributions. This approach contains several advantages over phasing, as mixture modelling:

- Requires much less computational time.

- Does not require a pre-phased reference dataset (e.g. 1000 Genomes Project) which may not contain enough reference information relevant to the genotype being phased and can incorporate ethnic group bias if the reference dataset is not representative.

- Does not require accurate identification of haplotype blocks. If haplotype blocks are not identified accurately by segmentation, minor allele SNPs will be determined to be present in the major allele distribution regardless of the purity of the allelic imbalance, distorting mean major allele distribution.

The mixture modelling workflow used to detect the major allele distribution is outlined in Figure 4.1. The initial objective of the workflow is to identify segments which are affected by the same allelic imbalance. It is only possible to model a segment as a two component Gaussian mixture if a single CNA is affecting it. To achieve this the mBAF of each chromosome is segmented using PCF (Fig. 4.1A). In order to conduct a parameter search using a grid search, an appropriate range of parameters must be determined. The range of means that are possible in the BAF values of the allelic distributions of a CNA segment are bounded by their distance from 0.5 ($0 - 0.5$). However, the range of standard deviation (*sd*) of the distributions are unbounded. Therefore to find an appropriate range of *sd* to test in a grid search a two component Gaussian mixture model is fitted to the BAFs of each segment across the genome using EM. The model is bounded by the following constraints which reduce parameter space and are appropriate for the underlying data type:

- The means of the two fitted distributions, *m*, must be equidistant from 0.5.

- The *sd* of each of the two distributions fitted must be equal.

93

EM then searches this parameter space to optimise log-likelihood across a limited of number of iterations (Fig. 4.1B). Once this is performed on each segment the *sd* converged upon is recorded for each segment and a representative *sd* is chosen as well as a range either side of this *sd* (see Section 2.6.2).
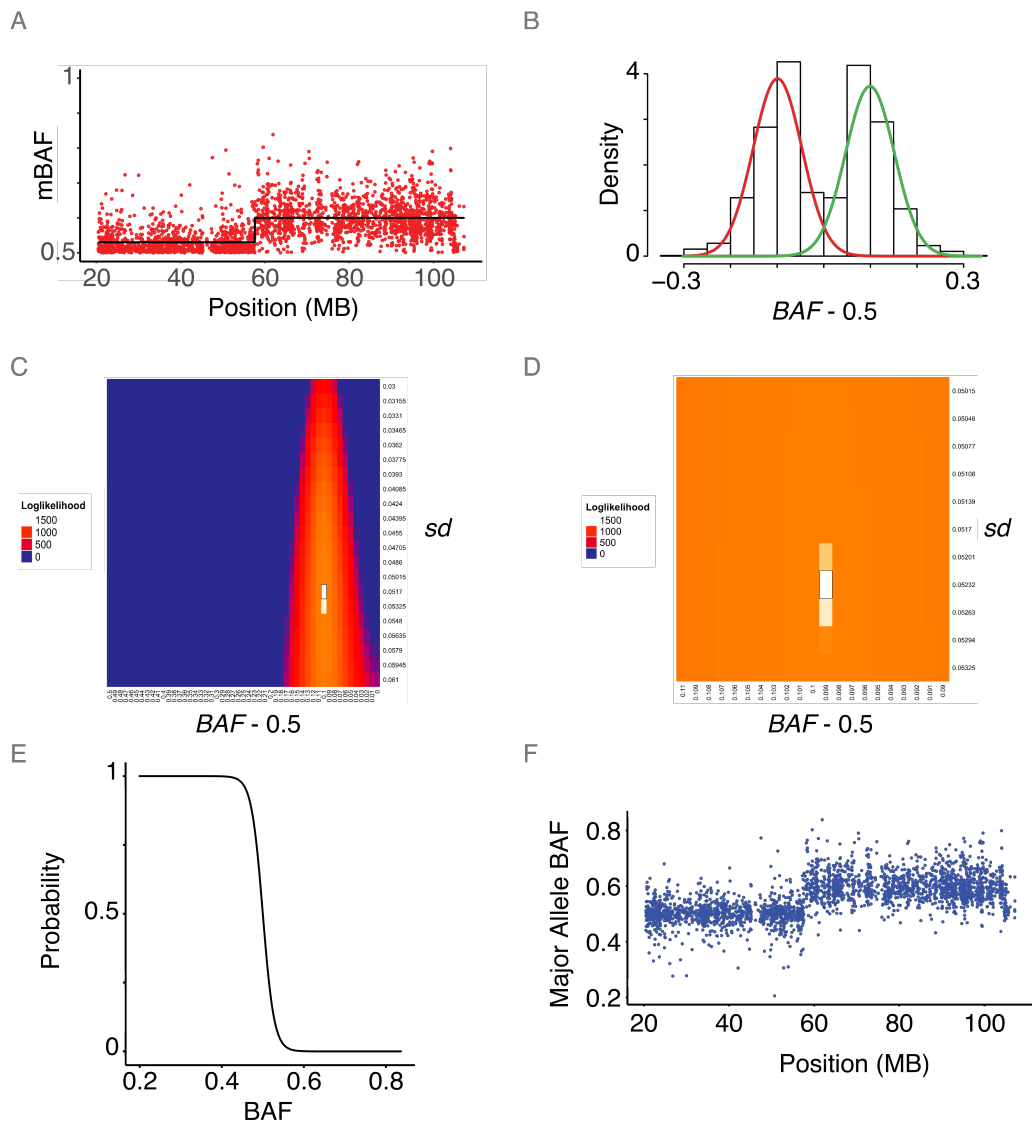
Figure 4.1: Using a mixture model of two Gaussian distributions allows for the detection of the major allele distribution. The beginning of this workflow requires the identification of a segment in which only a single CNA state is present. To achieve this the BAF across each chromosome is mirrored on 0.5 (mBAF) and segmented (A). Taking the non-reflected BAF within each segment, a two component Gaussian mixture model is fitted to the data using EM (B). The mixture model is constrained by the means being equidistant from a balanced BAF and the *sd* of each distribution being equal. Based on the fitting of the model across all segments, a representative range of *sd* is chosen. This range is then used to conduct a 'global' grid search including all possible BAF means (C) for each segment followed by a 'local' search (D) of a range of parameters surrounding the parameter combination with the largest log-likelihood of the global search.

(continued)

95

Figure 4.1: The mean and *sd* of the two Gaussian distribution are used generate a ratio of the densities of the minor allele distribution (mean <0.5) over the sum of the two densities, this ratio is then interpreted as the probability of a SNP being present in the minor allele distribution given its BAF (E). A representative major allele distribution is then generated by reassigning each BAF value as 1 – BAF, with the reassignment based on the minor allele distribution probability (F). This approach can replace phasing in order to estimate CNA cancer cell fraction using the Battenberg workflow.

By defining a bounded parameter range of *sd* to search, and a range of intervals between the bounds of distances from 0.5 appropriate for the distribution means (0 – 0.5), a grid search can be performed. Here, a log-likelihood is calculated for each combination of distance from 0.5 for the means of the distributions and *sd* of those distributions and the highest log-likelihood is chosen for the given data (Fig. 4.1C). A grid search overcomes the random seeding of EM and its ability to converge on local optima. Conducting a grid search also ensures all extreme values are tested (such as the means of both distributions being equal, i.e. distance = 0), yet is computationally efficient. The dimensionality of the grid search is greatly reduced by introduced constraints on the *sd* and means of the distributions. Following this first 'global' grid search, a 'local' grid search is implemented in a more refined range of values surrounding the chosen solution to the global grid search to improve the chosen solution (Fig. 4.1D).

Once two distributions have been fitted to the BAFs in the segment a 'minor allele ratio' is calculated using the densities of these two normal distributions given the chosen mean and *sd*. The distribution with a mean <0.5 is considered to be the minor allele distribution and the distribution with a mean >0.5 is defined as the major allele distribution. The ratio is calculated for a given BAF value as the density of minor allele distribution divided by the sum of the densities of the minor and major allele densities. This ratio is then considered to be representative of a probability that a SNP with a given BAF belongs to the minor allele distribution (Fig. 4.1E). Each SNP in the segment is subsequently mirrored at 0.5 (1 – BAF) according to the probability that it is in the minor allele distribution. This then

transforms the BAFs across the segments into a representation of the major allele distribution (Fig. 4.1F).

4.3 Comparison to phasing of breast carcinoma high-density SNP arrays

In order to test the mixture modelling approach I obtained five MicMa breast carcinoma samples in which the major allele distribution was estimated using phasing and the 1000 Genomes Project dataset as a reference [136]. These cancer samples were run on high density Affymetrix® Genome-Wide Human SNP 6.0 arrays. For each segment identified by segmenting the estimation of the major allele distribution by phasing this dataset, I applied the mixture modelling workflow from the point of estimating the *sd* of the distributions (Fig. 4.1B). I then compared the estimated major allele BAF obtained by the mixture model to the mean of the segment identified following phasing (Fig. 4.2).

Overall the concordance between the two approaches is very good. The lowest R^2 across the five arrays was 0.97. Generally phasing has a tendency to produce a higher estimation of BAF closer to the normal state (0.5). This is most obvious in MicMa 044 where there are higher BAF values close to 0.5 for a cluster of segments in the phasing estimation. This may indicate the greater sensitivity of the phasing approach to very low cancer cell fraction allelic imbalances. This effect however is not found in higher BAF estimations.

In the opposite manner segments of a higher cancer cell fraction tend to have a slightly higher estimated major allele BAF in the mixture model approach. This may be due to the fact that mixture modelling is not affected by accurate segmentation of haplotype blocks which if not identified properly can underestimate the mean major allele BAF. This effect increases as the purity of the allelic imbalance rises.

97
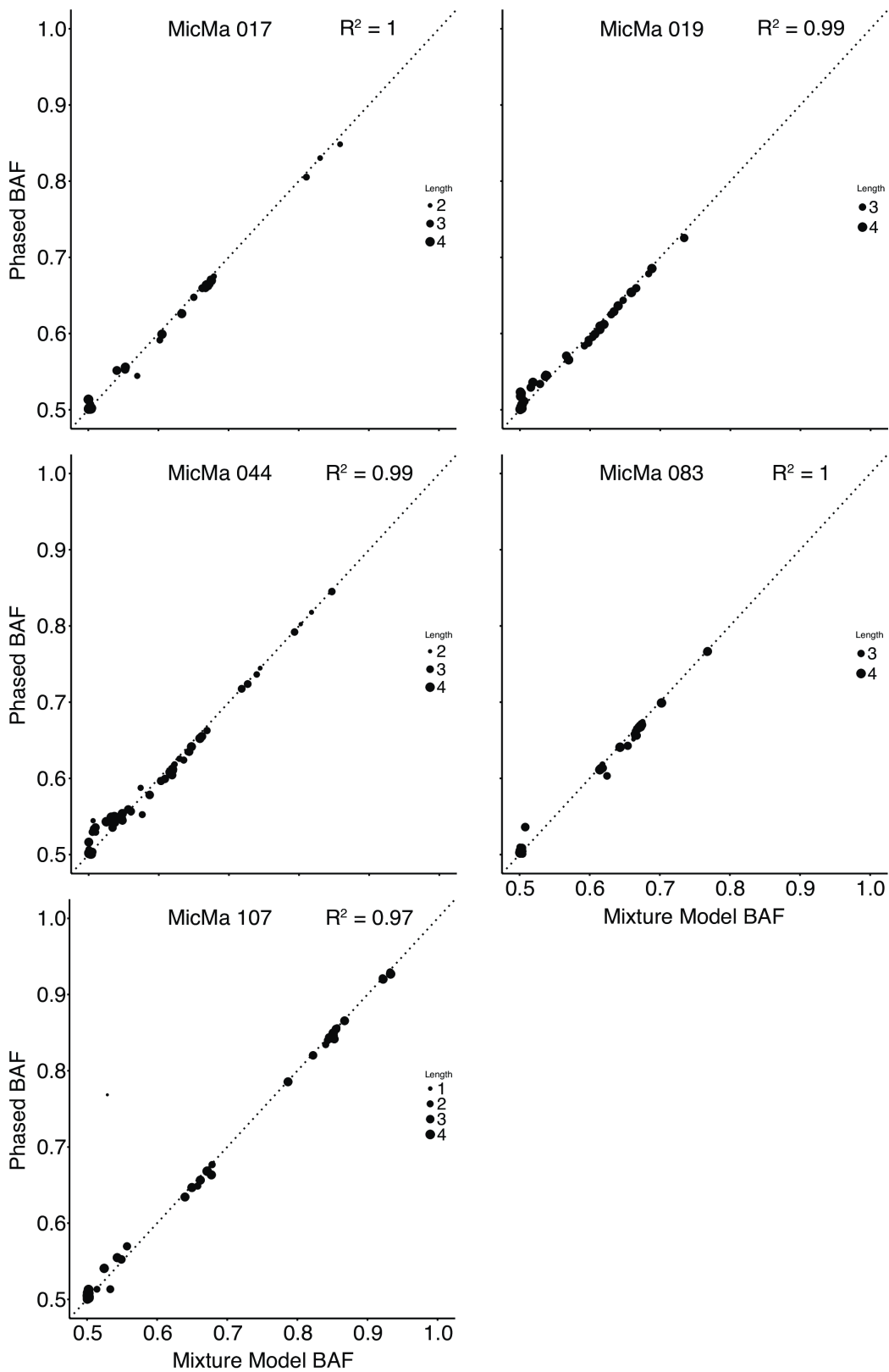
Figure 4.2: A comparison of the predicted major allele distribution as predicted by the approach outlined in Fig. 4.1 and produced by phasing using impute2 and the

(continued)

98

Figure 4.2: 1000 Genomes Project dataset as a reference. The data is derived from five Affymetrix SNP6 arrays which assayed samples taken from the MicMa breast carcinoma study. Segments produced by the phasing approach were then used as input for the mixture model approach (replacing step A in Fig. 4.1) to allow for a direct comparison of predicted major allele BAF. The points in the scatter plot represent each segment and the size of the points are scaled to the length of the segment, as detailed in the legend of each plot (value represents the exponent of base 10 of the numbers of SNPs). The dotted line represents an identical prediction and the R^2 is also displayed.

Figure 4.3 shows two chromosomes, 12 and 17, taken from MicMa 107. The figure shows the estimated BAFs of the major and minor alleles using the phasing approach (red) and the mixture modelling approach (blue). Firstly, it is clear both phasing and mixture modelling correctly detect the absence of a CNA in chromosome 12. However in both chromosome 12 and 17 the estimations of BAF derived from the two methods, in both the major and minor allele, can differ. In chromosome 12 the CNA is estimated to be purer by the phasing approach, however for the CNAs in chromosome 17 with higher BAFs, the mixture modelling estimates BAF to be higher, representative of the pattern show in Figure 4.2. This may display the ability of phasing to be more sensitive to low cancer cell fraction CNA but the ability of mixture modelling to estimate BAF of purer allelic imbalances better.

4.4 Analysis of genome-wide low density breast carcinoma SNP arrays

After testing the mixture modelling approach on high density Affymetrix® SNP6 arrays, I tested the approach on lower density SNP arrays, to gain insight into the ability of the workflow to estimate major allele distributions in arrays not conventionally used for subclonality analysis. Therefore I analysed 110 MicMa breast carcinoma samples that were run on Illumina® Human-1 109K BeadChip SNP arrays [150]. As noted in Table 4.1, these arrays probe for ~88% fewer SNPs than Affymetrix® Genome-Wide Human SNP6 arrays.
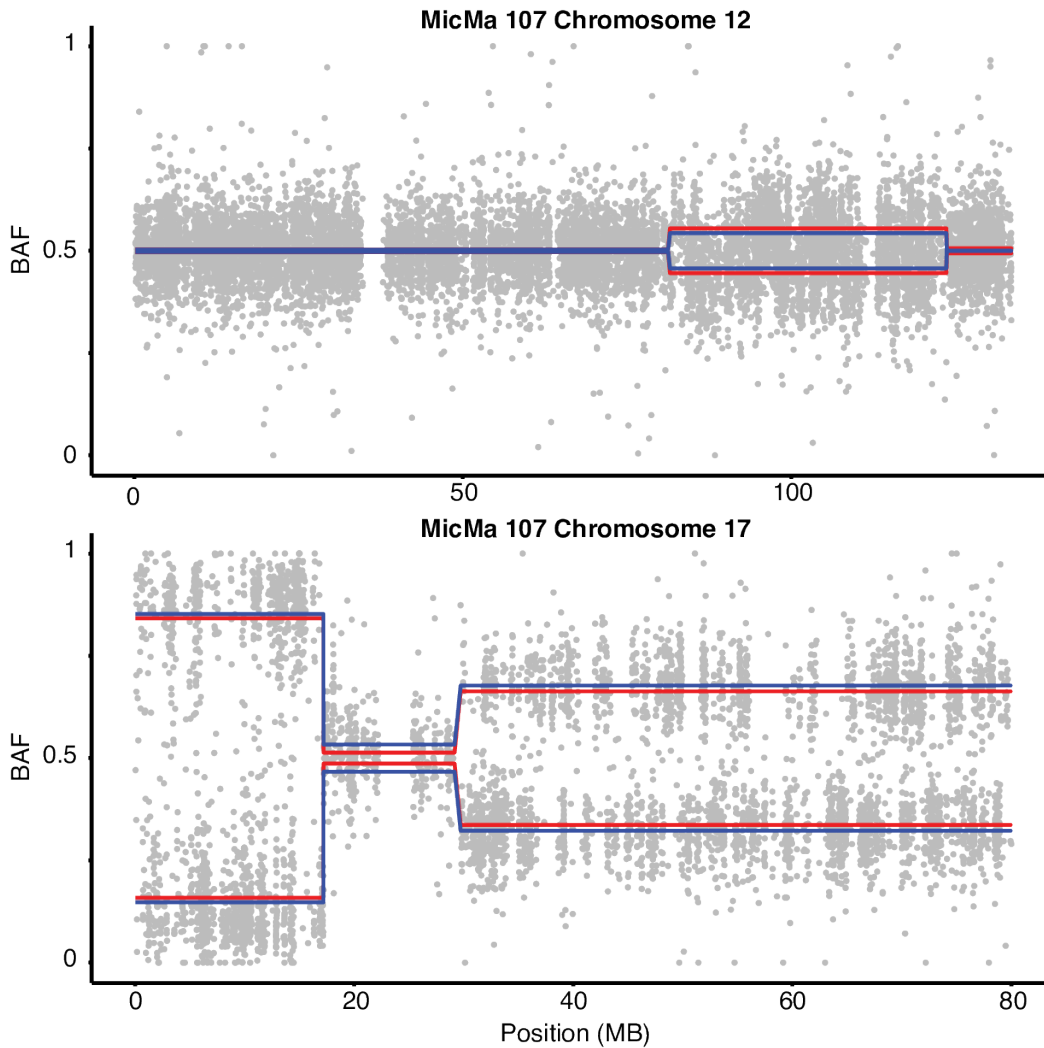
Figure 4.3: A comparison of the estimations of major and minor allele frequencies across chromosome 12 and 17 in MicMa 107. Blue lines represent the results from the mixture model and red lines represent the results from the phasing-based approach. Chromosome 12 shows a CNA surrounded by normal allelic balance. In this region the phasing BAF estimation is higher than the mixture model (0.555 and 0.543, respectively). In chromosome 17 of MicMa 107, the predicted BAF by phasing is reduced compared to mixture modelling. The regions of normal allelic balance in chromosome 12 show phasing and mixture modelling identifying the normal state.

As shown in Figure 4.4, the mixture modelling approach is able to detect subclonal CNAs in these low density SNP arrays. The upper panel of Fig. 4.4 shows the output of cancer cell fraction estimates according to the Battenberg workflow for chromosome 1 of a breast carcinoma sample in which mixture modelling has replaced phasing [120]. This chromosome possesses two CNAs, a whole arm single
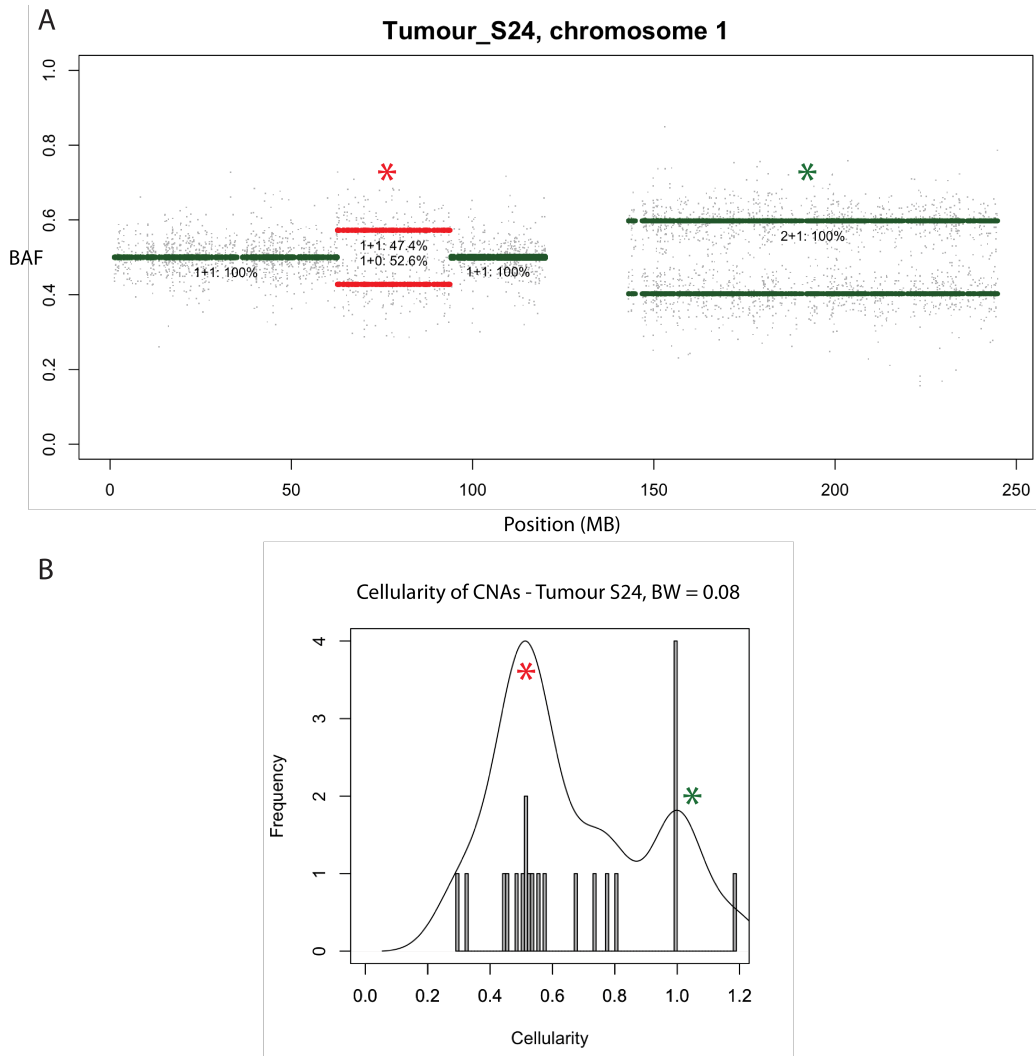
Figure 4.4: Using mixture modelling on low-density Illumina® Human-1 109K BeadChip SNP array from a breast carcinoma sample to determine CNA cancer cell fraction. In the upper panel (A) the BAF and segment copy number calls are shown along with their cancer cell fractions. This tumour sample contains a chromosome 1q gain [2+1] in all cells but a copy number loss [1+0] affecting 1p31.3–p22.1 (62.7–93.8 MB) that is determined to be present in 52.6% cells. The lower panel (B) represents the cancer cell fractions of all CNAs called in the sample as a mixture of an aberrant and diploid state [1+1] across all chromosomes. Here two groups of cancer cell fractions are observed, the clonal CNAs which are fixed at 100% (green star) and a distribution of subclonal CNAs distributed around 0.5 (red star). The line in the lower panel represents the kernel density estimation of the distribution using a BW of 0.08.

gain of chromosome 1q [2+1] and a single loss of 1p31.3–p22.1. The chromosome 1q gain is estimated to be in 100% of cells whereas the loss in 1p31.3–p22.1 is considered to be subclonal (red) in 52.6% of cells. The mixture modelling estimates a

normal allelic balance in the rest the chromosome.

The lower panel in Figure 4.4 shows the distribution of the cancer cell fractions of all the CNAs identified in the same tumour sample across the genome. The whole arm gain of chromosome 1q is clonal, and along with other CNAs, forms part of the 'clonal' distribution (green star). Other CNAs are considered to have cancer cell fractions less than 100% and are potentially subclonal. The loss in 1p31.3–p22.1 is in 52.6% of cells and forms part of the distribution of 'subclonal' CNAs (red star). By identifying troughs in the kernel density estimate, it is possible to separate CNAs into groups of similar cancer cell fractions. One CNA is estimated to have a cancer cell fraction >100% because the closest integer copy number state in this CNA is modelled as a mixture with a normal copy number state [1+1] but gives an impossible result (>100%) based on the purity and ploidy of the sample. This may be due to the purity or ploidy of the sample being inaccurate or that the CNA is a mixture of two aberrant states.

By running the mixture modelling and Battenberg workflow on these 110 MicMa breast carcinoma samples I could, for the first time, analyse subclonality in these low density arrays. A genome-wide summary of this analysis is shown in Figure 4.5. The upper panel (Fig. 4.5A) displays the number of CNA alterations identified across the genome in this dataset as a fraction of the total number of samples analysed (n = 110). Overall the regions with some of the largest number of CNAs included chromosome 1q, 8, 11q, 16q and 17p (>60% of samples).

The lower panel in the figure shows the percentage of subclonal mutations called as a fraction of the total CNA count. The genomic location with the highest enrichment of subclonal CNAs was in chromosome 1p. The cytoband affected by the most subclonal CNAs was mapped as being 1p22.2 where CNAs affecting this cytoband were subclonal in >70% of CNAs. Perhaps more interestingly some whole chromosomes had a tendency not to be subclonal. Chromosomes 3, 17 and

21 showed a depletion in subclonal CNAs indicating these chromosomes are often clonal and may play important early roles in the evolution of this cancer type.
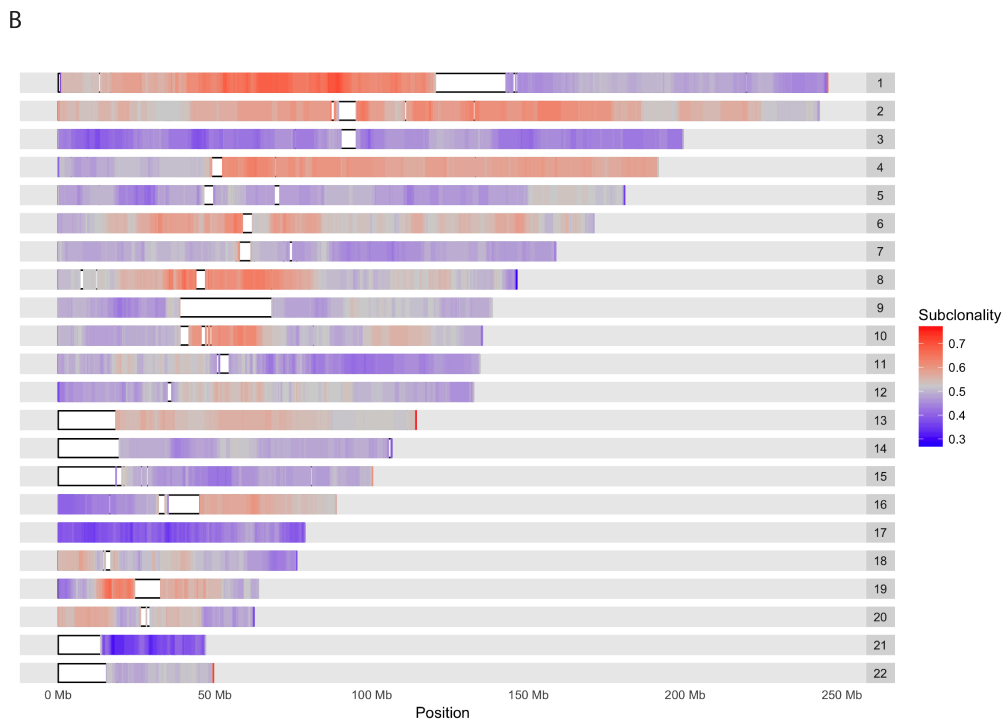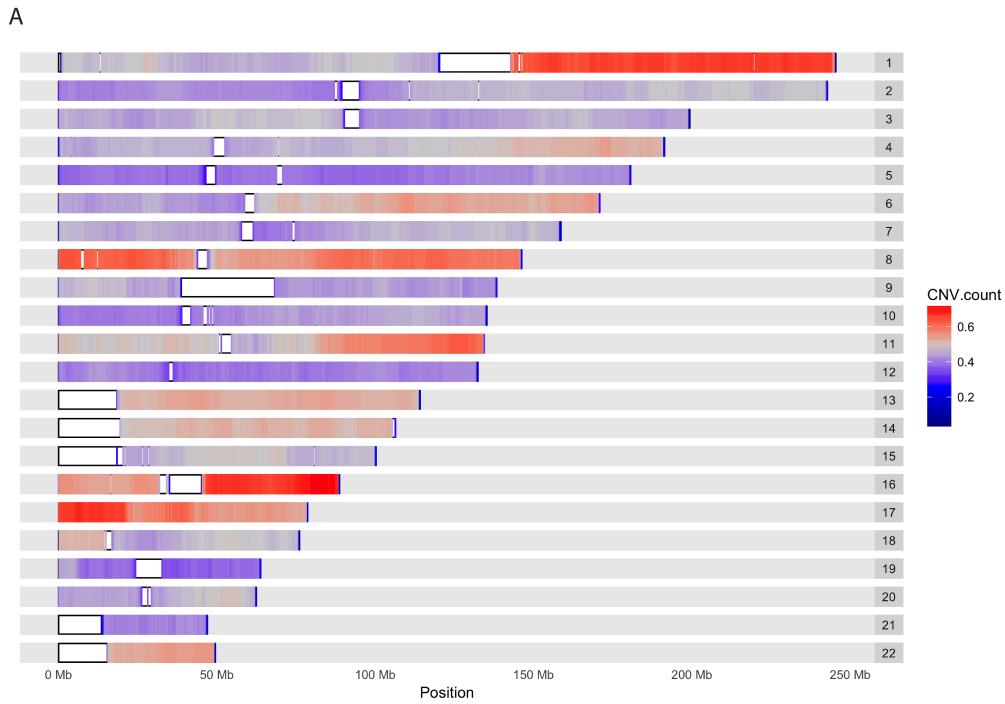
Figure 4.5: By taking a large set of low density SNP arrays run on breast carcinoma samples and using mixture modelling to identify major allele distributions, it is possible to gain insight into subclonality of CNAs across the set, despite the relatively low number of SNPs probed. Panel A displays the percentage of samples containing a CNA in a given genomic location. Generally this shows

(continued)
</center>

Figure 4.5: chromosome 1q, 8, 11q, 16q and 17p to be most affected by CNAs. The second panel (B) shows the percentage of CNAs that are subclonal. Patterns of subclonality are less clear genome-wide, however the q-end region of the p-arm of chromosome 1 appears to harbour a large number of subclonal CNAs. CNAs across in chromosomes 3, 17 and 21 are more commonly clonal.

4.5 Comparison to phasing of medium density SNP arrays in the WT20 series

To explore the power of mixture modelling for detecting subclonality further, I applied the approach to the WT20 series of Wilms' tumours presented in Chapter 3. The SNP arrays analysed in the WT20 series are 'medium' density Illumina® HumanCytoSNP-12 v2.1 arrays which have only ~30% the number of SNPs that are probed for in high density Affymetrix® SNP6 arrays. I compared the ability to estimate the major allele distribution using mixture modelling as well as phasing using impute2 and the 1000 Genome Project reference genomes. Generally the phasing approach under-performed as it produced poor haplotype block predictions that led to an underestimation of the major allele BAF and erroneous segmentation. The mixture modelling approach was not affected by these two general problems.

The disadvantages of phasing medium density SNP arrays are presented in Figure 4.6. Here, Figure 4.6A shows data from chromosome 2 in a tumour sample in Case 15. The left graph shows the estimated major allele distribution produced by the phasing approach. Here it is clear that there are several SNPs with BAFs less than 0.5 that are derived from the minor allele distribution. These remain present as the predicted haplotype blocks were not sufficiently accurate enough to be correctly identified through segmentation. The right panel of Figure 4.6A then shows that the mean BAF (red) of the phasing result is reduced compared with the mixture modelling result (blue) and is inaccurate, likely leading to lower cancer cell fraction estimates.
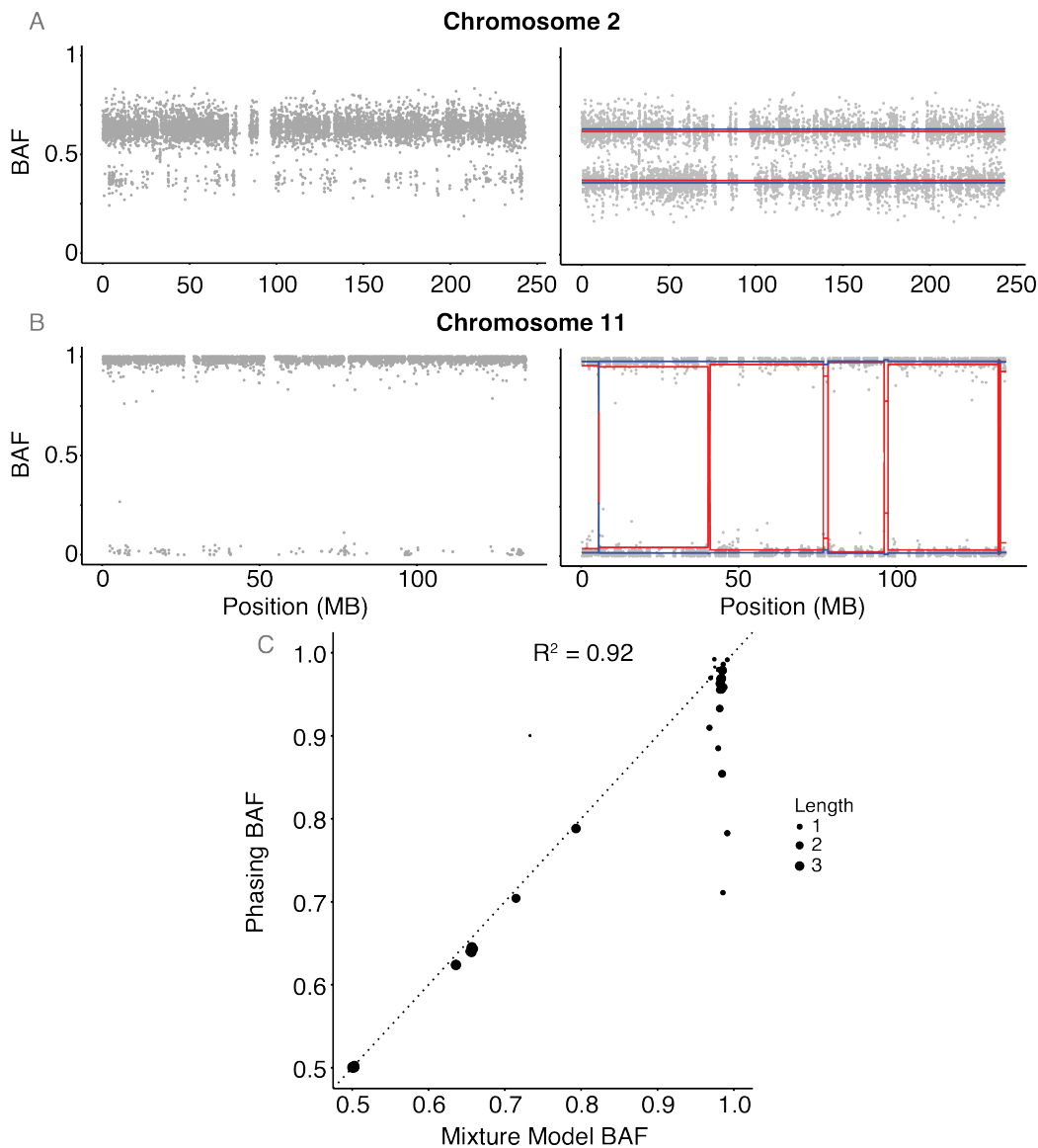
Figure 4.6: Both phasing and mixture modelling were used to estimate major allele BAF in the WT20 dataset presented in Chapter 3 that was performed on medium density SNP arrays. A copy number gain in chromosome 2 in Case 15 shows SNPs of the minor allele distribution present in the estimation of major allele distribution by phasing (A). This reduces the mean BAF produced by the phasing approach (red line) compared to the mean BAF estimated by mixture modelling (blue line). Poor phasing in chromosome 11 and the purity of the allelic imbalance in this sample causes erroneous segmentation (red line) due to segmentation detecting minor allele SNPs (B). This effect is shown in the comparison scatter plot (C) as a series of segments with high major allele frequency produced by the mixture model but variable BAF produced by phasing.

Figure 4.6B shows chromosome 11 from the same tumour sample. Again BAF values far less than 0.5 (~0) are present in the estimation of the major allele distribution in chromosome 11 due to poor haplotype block identification. However in this instance the segmentation performed on the major allele distribution estimated by phasing, to identify CNAs, calls SNPs in the minor allele distribution as separate segments from the major allele distribution. These segments are clearly problematic for accurate identification of CNAs in this chromosome. The mixture modelling approach is not subject to this issue.

Additionally, Figure 4.6C shows a comparison of the major allele BAFs of segments produced by the phasing approach compared to the mixture model approach. The clearest evidence of erroneous segmentation is the presence of many segments which have high (~1) BAF values as determined by mixture modelling, but have highly variable BAF values according to the phasing approach, likely due to poor haplotype block identification. Finally, it is also evident that generally speaking, BAF value estimations are slightly higher for the mixture modelling approach compared with the phasing approach.

4.6 Reanalysis of the WT20 study

Reanalysing the WT20 dataset using mixture modelling to estimate the major allele distribution and estimating CNA cancer cell fraction using Battenberg allows for the re-evaluation of inferred phylogeny in those cases. Figure 4.7 shows the reinterpretation of Figure 3.3 from Chapter 3. The first improvement that understanding subclonality produces is the identification of more observable states in the tumour. In Figure 4.7A there are five observed tumour states across six samples, however in 4.7B the detection of subclonal architecture means seven states are observed, adding more detail to our picture of tumour evolution. Secondly, through allowing for the detection of subclonal CNAs, more CNAs in total are detected, as CNA calling does not require CNAs to match to a single profile. This is seen in R4 where four additional CNAs are observed (17p-, 17p+, +20 and +21), clearly detailing a
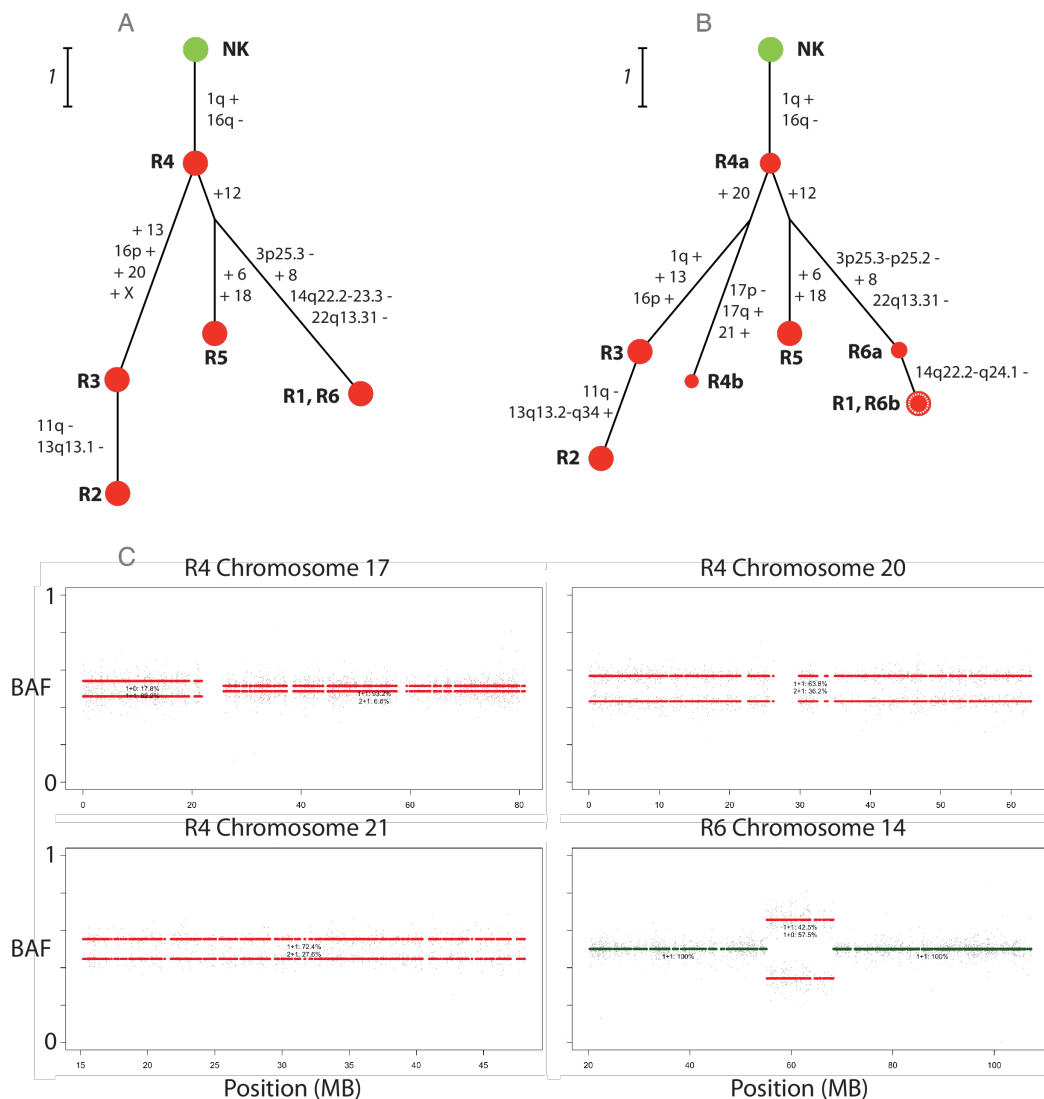
Figure 4.7: Detecting subclonal CNAs in Case 19 from Chapter 3 allows for a more detailed picture of WT evolution. Panel A is taken from Fig. 3.3 and panel B includes subclonal detection in phylogenetic reconstruction. Here the mixture modelling and Battenberg workflow was applied to the array samples of Case 19 and five CNAs in four chromosomes across two samples were detected as subclonal. Four of the subclonal CNAs were present in R4. These included 17p-, 17q+, +20 and +21. The chromosomes with detected subclonal CNAs in R4 and R6 are shown in panel C.

more accurate detection of the range of CNAs in the tumour.

The detection of both more CNAs and estimation of their cancer cell fraction allows for the ordering of CNAs to be further clarified in the inferred phylogeny. For

instance, the gain of chromosome 20 present in R4 allows us to interpret this event as occurring first in the successive CNAs that are acquired by R2 and R3, which also contain a chromosome 20 gain. Additionally, the detection of 14q22.2–q24.1 loss as being subclonal in R6 informs us that in the acquisition of CNAs unique to R1 and R6, this loss of 14q22.2–q24.1 occurred last. Discrepancies observed between Figure 4.7A and 4.7B are explained by the differences between this mixture modelling-Battenberg approach and the methodology of Chapter 3. These include:

- The removal of the X chromosome from analysis in Figure 4.7B due to the lack of heterozygous SNPs in male patients.

- Segmentation is performed on the mBAF of heterozygous SNPs in Figure 4.7B and not primarily the LRR as in Figure 4.7A. This explains boundary differences seen in the CNAs of chromosome 3 and chromosome 14, as well as the different interpretation of chromosome 13.

- The increased sensitivity to higher gain states in Figure 4.7B and therefore the additional gain of chromosome 1q in R2 and R3 [3+1] by using the ASCAT equations which gives unlimited copy number states as opposed to CGHcall which is limited to a total copy number state of 4.

Figure 4.7C shows the five subclonal CNA calls in the four chromosomes across R4 and R6. Cancer cell fraction estimations range from 6.8% to 57.5%.

4.7 Concluding Remarks

To gain full insight into the evolution of a tumour, it is clear that subclonality within single tumour samples must be accounted for. By estimating CNA subclonality, it is possible to observe more states in the clonal architecture of a tumour. This increases the amount of information used for constructing phylogeny as well as providing important constraints for establishing the progress of genetic change in a tumour's evolution. Through detecting that a CNA is subclonal, we make clear that it is highly likely to have occurred after a clonal CNA (or more accurately, a CNA present in 100% of cells in a sample). Elucidating subclonality in one sample can

then provide insight to the ordering of genetic events in other samples which may be hidden by changes being present in all cells, as exemplified by the subclonal +20 in R4 that is clonal in R3, allowing it to be timed as occurring prior to the other clonal CNAs in R3 that are unique to the sample.

Sensitive and accurate detection of allelic distributions across the cancer genome is vital for understanding CNA cancer cell fraction. Whereas it is established that phasing provides a powerful tool for achieving this, the approach I developed using mixture modelling is also capable of achieving the same objective. By providing an alternative to phasing, mixture modelling opens up the option of assessing subclonality in a range of situations in which phasing a genome effectively is not possible. These situations may include difficult to phase genotypes that are not represented accurately by a reference genotype dataset. This situation may arise if phased genotypes are not available for specific ethnicities or if a genotype is uncommon in the general population, such as those arising from consanguineous parentage, that may occur in childhood malignancies [188].

Mixture modelling is capable of detecting the allelic distributions in arrays with lower tiling density such as Illumina® HumanCytoSNP-12 v2.1 arrays and Illumina® Human-1 109K BeadChip SNP arrays and may also perform well on targeted sequencing experiments such as exome sequencing. In general the ability of mixture modelling to perform on lower tiling density arrays may 'unlock' many large legacy genomic studies previously thought to be unsuitable for subclonality analysis. As shown in the WT20 Illumina® HumanCytoSNP-12 v2.1 array series, phasing an array with approximately 300,000 SNPs does not appear to produce accurate haplotype block predictions. Yet, in this chapter CNA subclonality was assessed in the MicMa breast carcinoma dataset analysed on Illumina® Human-1 109K BeadChip SNP arrays using mixture modelling despite their low tiling density (~100,000 probes) [150].

It should be noted that phasing may be more robust to noise in the BAF data as compared to mixture modelling. Fitting a mixture model to the BAF data will be affected by noise in the BAF distributions in the SNP array. However, phasing is performed on the genotype information that, assuming the identification of homozygous and heterozygous SNPs is accurate, is performed on the genotype of the patient directly and is not affected by this type of noise initially. However, the subsequent identification of haplotype blocks requires accurate segmentation and this will be compromised by poor quality BAF data.

Phasing a patient's genotype is a powerful approach to assessing CNA cancer cell fraction and accurately estimating BAF. However, mixture modelling provides several advantages that makes it a viable alternative to phasing. Whereas mixture modelling may not be more preferable to phasing in all situations, the fact that it is computed quickly and at low cost makes it a good source of comparison, at a minimum, for all studies of CNA subclonality.

Chapter 5

Comprehensive Evolutionary Reconstruction of Paediatric Kidney Cancer

5.1 Introduction

Extending my investigation into paediatric solid tumour evolution, I analysed a larger more diverse cohort of patients diagnosed with paediatric kidney cancer (PKC) in which cancerous tissue was both numerously and longitudinally sampled, where possible. To expand even further our knowledge of these tumours and to display the advantages of detecting clones both within tumour samples as well as across tumour samples, the aim was to assay copy number variations using an accurate estimation of copy number cancer cell fraction from data derived from Illumina CoreExome-24 SNP arrays. Additionally, we assayed mutations in multiple regions using a WT-specific DNA pull down panel to capture the exons of genes implicated in WT development using deep sequencing in a subset of cases.

By utilising methods presented in Chapter 4 I aimed to infer mixing of clones in all array samples individually, to construct a more accurate picture of PKC evolution by further elucidating the individual stepped changes that CNAs undergo when evolving. Overall, as well as being able to provide an interpretation of tu-

mour evolution by comparing between multiple samples as presented in Chapter 3, I hypothesised that being able to detect these mixed clones in single samples would allow me to order copy number changes further when comparing to multiple tissue sites. The detection of mixing clones would identify more directly observed states to be used in phylogeny construction and the increased and accurate detection of low cancer cell fraction changes, as well as the recognition that these low cancer cell fraction changes do not require to be explained in a single 'average' profile alongside higher cancer cell fraction/clonal CNAs, would also provide a larger number of detected events across the multiple tumour pieces that may also help explain ordering of changes between samples as well as identifying changes that may have previously gone undetected. Performing this across multiple samples would synergise two major approaches in detecting diversity produced by evolution in the field, the reconstruction of clones from bulk data and the derivation of phylogenetic history from comparing between multiple sites that possess spatially separated clones.

As presented in Chapter 3, copy number changes occur frequently in WTs and are informative for deducing the order of events that occurred during the evolution of the cancer. However, like the vast majority of cancers, WTs are known to also possess point mutations and small insertions/deletions. Until this stage we had not addressed these mutations from the view point of heterogeneity and evolution. We chose to address this because, unlike many CNAs – and especially in large CNAs like those that often appear in WT (1p–, 1q+, 16q- etc.) – the functional consequences of small sequence mutations are predictable as we can determine the amino acid changes that affect the translated protein from the mutated gene. In this regard, point mutations and indels are perhaps more informative as individual mutations than CNAs from a functional perspective. Additionally, these point mutations are likely to alter the phenotype therefore relative fitness of clones. However, as explored in Chapter 1, WTs possess relatively few non-synonymous mutations comparatively to the majority of adult cancers. Bearing this in mind, and whilst aiming

to perform an assay of small sequence mutations in multiple regions of a sizeable cohort of cases with relatively high coverage to ensure accurate VAF estimations, we chose to design a WT-specific targeted sequencing panel that we would apply to the PKCs in our cohort. We felt this approach to be both the most cost effective as well as the best way to 'hedge our bets' of finding the majority of mutations in each paedatric cancer genome. In addition to assaying the spatial presence of SNVs and indels in the tumour, the panel would also serve as a prototype for a diagnostic assay.

I designed the WT-specific targeted sequencing panel to include both coding and non-coding genes implicated in WT development, specifically those which have previous evidence of mutation, as well as eight regions of the genome that often exhibit potentially prognostic CNAs in WTs. In total I included 167 coding genes and 14 non-coding genes, this included any gene that showed a level of recurrence in WTs (mutated more than once in a recent exome study [40]). Further details can be found in Section 2.3.3.

Alongside the aim to reconstruct the evolutionary trajectories of PKCs more accurately using improved processing of CNA data and assaying SNVs and indels, we also took advantage of the tissue collection routinely being performed as part of the IMPORT study. This UK-wide study enrolled all consenting PKC patients in the UK across multiple centres from ~2013 and stored these samples at Great Ormond Street Hospital, London. The study is of PKCs generally and therefore is representative of the spectrum of PKCs, ergo the majority of the cases (~90%) are WTs [14]. As part of this study patients are sampled in various ways longitudinally. Firstly, a multiple sampling approach takes place at nephrectomy if possible, additionally tumour samples may also be available from a diagnostic needle biopsy, providing a tissue sample from before chemotherapy and therefore prior to nephrectomy, furthermore tumour samples taken from metastases as well as relapses (therefore a later timepoint) are also taken in this study. In addition to tumour sampling, liquid samples were also available as both blood and urine samples are stored from these

patients at several timepoints (diagnosis, mid-point of chemotherapy, pre-operation, post-operation, end of treatment). These samples were taken to assess ctDNA and utDNA.

In total I analysed tissue samples from 66 patients (PKC66) that were enrolled as part of IMPORT to determine ordering of CNAs in the life history of these PKCs. This included 10 patients with pre-chemotherapy biopsies taken at diagnosis. Of the PKC66 patients, multiple tumour regions from 30 cases were assayed using a WT-specific sequencing panel.

I hypothesised that using a larger cohort could potentially allow us to assess a large enough number of WT histological subtypes to understand evolution in specific phenotypes, particularly in those subtypes that have differential risk stratifications. Additionally, by utilising a large enough dataset, I aimed to identify patterns of diversification and mutational ordering to paint a generalised picture of evolution in these solid paediatric tumours. By utilising methods to detect clonal mixing in single samples, I was also able to present within sample mixing across multiple regions, to display a comparison of 'micro'- and 'macro'-heterogeneity (as described by [115]), as I gained an overall picture of both of these approaches to studying intratumour genetic diversity in a large number of patients.

We also wished to compare clonality of specific mutations to their levels in ctDNA, specifically of *TP53* in diffuse anaplastic type WT to understand if levels of *TP53* in the blood reflected primary tumour clonality. Additionally, I hypothesised that assessing general characteristics of WT phylogeny may be informative for distinguishing between high and intermediate risk WTs and may additionally be prognostic. For the majority of cases we had sample specific histological assessment as well as images of the primary tumours in a subset, I aimed to utilise this data to map tumour phylogeny back to spatial position and phenotype information.

My role in this part of the study was to create an analysis structure to call allele-specific copy number changes and to calculate the cancer cell fraction of each of these copy number changes. I then aimed to infer subclones from this data by clustering copy number cancer cell fractions which may have arose as part of the same clonal expansion. Once assigning copy number variations to mixed clones within single samples, I compared these events to the multiple samples of the tumour to recreate the phylogenetic tree for the case, informed by MEDICC. I also processed the raw sequencing data from the WT-specific panel to call mutations in these samples and to use these to aid phylogenetic reconstruction. My role was then to compare my findings to all available clinical data and to highlight general patterns in the cohort.

This chapter was performed in collaboration with the laboratory of Professor Kathy Pritchard-Jones. In general, none of the data presented was generated by myself, however I performed all presented analysis. Further details are presented in Chapter 2.

5.2 Subclonal reconstruction in sixty-six paediatric kidney cancers

The PKC66 set of patients were selected in an unbiased manner from cases that produced DNA which could be assayed. The majority of the patients were diagnosed with WT (62/66). The non-WT cases included diagnoses of diffuse hyperplastic perilobar nephroblastomatosis (DHPN), diffuse nephroblastomatosis, metanephric adenofibroma and rell cell carcinoma. The commonest subtype of WT was 'mixed' when considering unilateral cases alone and together with bilateral tumours (Table 5.1). Interestingly, diffuse anaplastic WTs were the second most representative subtype in unilateral tumours, however when including the individual tumours from bilateral cases, stromal was the second most representative subtype, highlighting an enrichment of stromal type tumours in the bilateral cases. Approximately 31% of unilateral tumours were a high risk subtype (blastemal and diffuse anaplastic), ~4%

Tumour Type	Subtype	Uni. tumours	Uni. & Bi.
	Blastemal	6	6
	Diffuse Anaplastic	9	9
	Epithelial	5	9
Wilms' tumour	Focal Anaplastic	1	1
	Mixed	12	22
	Necrotic	2	2
	Regressive	6	7
	Stromal	7	13
DHPN	-	0	1
Diffuse Nephroblastomatosis	-	0	1
Metanephric Adenofibroma	-	1	1
Renal Cell Carcinoma	-	1	1

Table 5.1: A summary of the tumour types in the cohort of 66 paediatric kidney cancer cases. The majority of these cases were WT (~96%) and the individual subtypes of WT are also noted, firstly for only the unilateral tumours (Uni. tumours), and secondly including the subtypes of each individual side from the 16 bilateral cases (Uni & Bi.).

were the low risk necrotic subtype and the majority (~65%) were intermediate type WTs.

In the PKC66 cohort 14 cases were bilateral WTs. Table 5.2 displays these cases labelled by their IMPORT IDs (CCLG-897 is not strictly an IMPORT case as the patient was not eligible due to not being a UK resident but was multi-sampled and assayed in line with the IMPORT study and is therefore included). The table summarises the diagnosed subtypes of the tumours from each kidney in these cases as well as the number of tumour samples taken from each side. In total ~25% of the WTs cases were bilateral cases, this higher than expected (~8%) [14]. Due to findings in Chapter 3 indicating that bilateral tumour appear not to be related, for general conclusions of the patterns of phylogeny I consider these tumours to have arose separately and only highlight the bilateral nature of these cases if it is in a relevant context.

Across the PKC66 series multiple tumour samples were taken for the majority of cases (~80%). The mean number of samples taken was 3.2 tumour samples and

IMPORT ID	Left Subtype	# Samples	Right Subtype	# Samples
4	Mixed	1	Mixed	2
7	Epithelial	2	Mixed	4
9	Mixed	2	Mixed	3
16	Stromal	2	Mixed	4
21	Epithelial	2	Regressive	3
30	Epithelial	2	Epithelial	3
47	Stromal	1	Stromal	3
61	Mixed	1	NA	0
66	NA	0	Mixed	2
78	Stromal	6	NA	0
147	NA	0	Mixed	4
153	Mixed	1	NA	0
171	Stromal	4	NA	0
CCLG-897	NA	0	Stromal	6

Table 5.2: A summary of the 14 bilateral WT cases in the PKC66 series stating the IM-PORT ID, the subtype and the number of tumour samples taken for each tumour by laterality.

two non-bilateral cases were sampled up to seven samples. Table 5.3 shows the spectrum of sampling in the PKC66 cohort, including per case and per kidney to take into account the bilateral cases. The majority of tumour samples are frozen nephrectomy (FN) samples taken from the primary tumour when it is removed after surgery, however 11 of these tumour samples are pre-chemotherapy biopsies (FBs) in 10 cases, one case contains a lung metastasis sample taken at relapse (FR). Eight cases (~12%) did not have a matching normal sample (blood/normal kidney) available which produced a high quality SNP array, therefore for these cases heterozygous SNPs had to be identified manually (IMPORTs 25, 27, 52, 85, 88, 108, 111 and 146). Six cases in the PKC66 series overlap with the WT20 series (Cases 15–20 were reanalysed as IMPORT 3, 4, 7, 8, 12 and 16 respectively).

5.3 Subtype-specific evolutionary observations

Following phylogeny reconstruction for the PKC66 cases, my aim was to assess the evolutionary trajectories of these cases as an 'ecosystem', studying the landscape as a whole by grouping by subtypes and comparing between them. I split the

Number of Samples	By Patient	By Kidney
1	13	15
2	19	25
3	8	10
4	7	9
5	9	6
6	7	6
7	3	2

Table 5.3: A summary of the number of tumour samples per case and per kidney (treating each tumour in bilateral cases as separate unilateral tumours). The number of tumour samples includes FNs, FBs and FRs.

cohort by the diagnosed tumour type to search for patterns that may be specific for each subtype. Overall, patterns were not completely consistent within each subtype, however some generalisable conclusions were apparent.

5.3.1 Necrotic tumours

Necrotic tumours represent the only low risk histological subtype in WTs. With only two cases in the cohort I was limited in conclusions that I could draw from these cases (Fig. 5.1). Both tumours were only sampled once at primary nephrectomy and the only multisampled case including biopsy (32) had no events detected and the second case with only a single sample at nephrectomy contained six clonal copy number variations (133). Sampling in these tumours is limited by tissue quality.

5.3.2 Mixed tumours

WTs with a histology comprising of variable amounts of blastemal, epithelial and stromal cells with no single cell type being completely predominant ($>66\%$) represented most of the unilateral tumours (12 cases – Fig. 5.2 and Fig. 5.3). Generally speaking there was a range of events and diversification in the mixed type tumours. Three cases had no copy number events and the most prevalent cell type was blastema in two cases (71, 120) and epithelial in one (52). Four other cases had only a few events (1–7), apparently enriched for 11p LOH changes and stroma (the predominant cell type in three cases) relative to other cell types (5, 38, 39, 64). Contrastingly, some cases presented with large diversity and clear branched evolu-
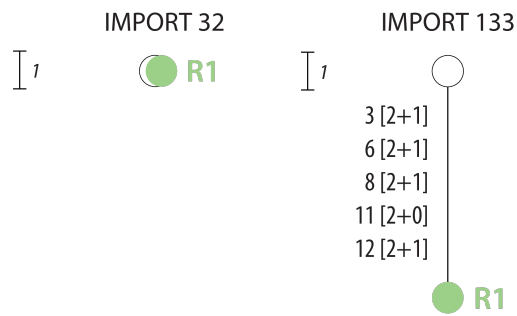
Figure 5.1: Necrotic WTs. Two cases in our series were diagnosed as necrotic type WT, neither was multi-sampled at nephrectomy, one case shows no CNAs (32) and one case (133) shows several different changes. Generally there are too few cases and samples to draw conclusions from these necrotic tumours, in IMPORT 133 no CNAs were called as subclonal. White circles represent the normal state and coloured circles represent clones in each tumour sample.

tion displaying the presence of non-clonal CNAs after relatively few clonal events (3, 12, 85). This striking explosion in diversification is markedly different from 'quieter' mixed tumours. IMPORT 85 was considered a 'borderline' blastemal case (65% blastemal content) and displayed MSAI in four chromosomes (explored in Section 5.6.4). The most dominant cell type in IMPORT 3 and 12 was stroma (60% and 55% respectively). IMPORT 108 was a tetraploid case (ploidy of 3.78 – R1, and 3.94 – R2) likely due to a WGD event and consequently showed many events plus branching. IMPORT 24 was only single sampled but did present several events and some non-clonal CNAs such as a 1q+.

5.3.3 Stromal tumours

Seven cases in the PKC66 cohort were unilateral stromal tumours (Fig. 5.4). Of the several WT subtypes in the cohort, stromal tumours displayed the most consistent evolutionary pattern. Six of the seven cases (23, 26, 29, 37, 143, 146) presented with an 11p LOH event, with various breakpoints, that always affected both 11p15 and therefore *IGF2* as well as 11p13, the *WT1* locus, and this change was always clonal consistent with findings in Chapter 3. Five of the cases have aberrations affecting chromosome 3 and in four cases this aberration affects the p-arm (26, 29,

IMPORT 3

‖ 1

6 [2+1]
8 [2+1]
10 [2+1]
11 [2+0]
12 [2+1]
21 [2+0]

R3b *11 R3a
*10 *9 *6 *1 R2a R5a
*12 *7 *2 R5b
R4a *4 *3
2 [2+1] R6a *8 R2b
6p25.3-p13.2 [2+0] R6b
6q12-q16.3 [2+1]
7p22.3-p14.1 [2+1] *5
R4b R1

*1 7 [2+1]   *5 1p21.3-p12 [3+1]
*2 1q31.1-q44 [2+1]   1q21.1-q31.1 [4+1]
*3 1q21.1-q31.1 [2+1]   2 [2+1]
*4 1q21.1-q23.3 [2+1] *6 1q [2+1]
                      *7 7p12.1-q36.3 [2+1]
                      *8 7p22.3-p12.1 [2+1]
                      *9 7p14.1-q36.3 [2+1]
                      *10 6q16.3-q27 [2+2]
                      *11 7p22.3-p14.1 [2+1]
                      *12 6q12-ter [1+1]

IMPORT 5

‖ 1

6q21 [1+0]
11p [2+0]
R1a
12p13.33-p11.23 [2+0]   12p13.33-p11.21 [2+0]
R1b   R1c

IMPORT 12

‖ 1

16q [1+0]
R4a *1   12 [2+1]
20 [2+1]   1q [2+1]
1q [3+1]   3p25.3-25.2 [1+0]
13 [2+1] *4   *5   8 [2+1]
16p [2+1]   22q13.31 [1+0]
R3 *2   R6a
R2a   R5 R1 *6 R6b
*3 R4b
R2b

*1 1q [2+1]   *5 1q [2+1]
*2 13q13.2-q34 [3+1]   6 [2+1]
*3 11q21-25 [1+0]   18 [2+1]
*4 17p [1+0]   *6 14q22.2-q24.1 [1+0]
17q [2+1]
21 [2+1]

IMPORT 24

‖ 1

1p [1+0]
4q [1+0]
11 [1+0]
14q12-q32.33 [1+0]
16 [1+0]
21q21.2-q22.3 [1+0]
R1a
1q [2+1]
14q11.2-q12 [2+1]
21q11.2-q21.2 [1+0]
R1b

IMPORT 38

‖ 1

11p [2+0]
R1, R2

IMPORT 39

11p14.1-p13 [1+0] ‖ 1
R1, R2

Figure 5.2: Six of the twelve mixed type Wilms' tumours in the series. Mixed tumours show a variety of evolutionary patterns. IMPORTs 3, 5 and 12 show branched evolution and IMPORTs 3 and 12 have a large number of events. Contrastingly IMPORTs 24, 38 and 39 show a linear trajectory of evolution and IMPORTs 38 and 39 only have aberrations affecting loci in the 11p arm. Areas of the circles represent the contribution of each clone to the sample in which it was detected.

143, 169). In three cases the aberration affects the *CTNNB1* locus (26, 29, 143) and in two cases it is detected as non-clonal (26, 143). Stromal WTs have previously been associated with 11p LOH events, *WT1* and *CTNNB1* mutation [189] and these findings could hint at a model for stromal WT development in which 11p LOH occurs early in the development of a stromal WT followed by a *CTNNB1* mutation
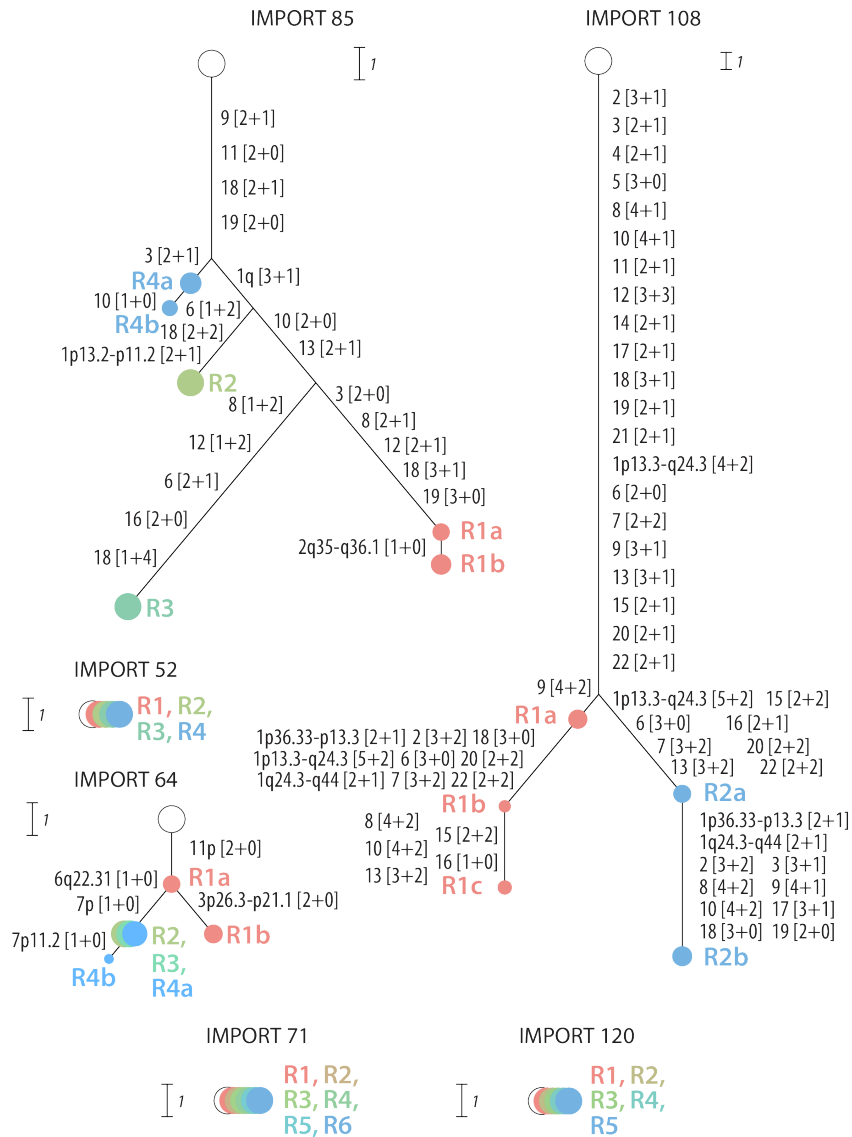
Figure 5.3: The remaining six mixed type Wilms' tumours of the twelve in the series show the same variety in evolutionary patterns. IMPORT 52, 71 and 120 have no copy number events yet IMPORTs 85 and 108 have many (IMPORT 108 has a possible WGD). IMPORTs 64, 85 and 108 display branched evolution.

and subsequent CNA of the *CTNNB1* locus. As presented in Section 5.9, three unilateral stromal WT case in the PKC66 series were sequenced (143, 146 and 169). Both IMPORTs 143 and 146 fit the previously stated model, however IMPORT 169 did not show *WT1* or *CTNNB1* mutation. Curiously, the CNA affecting the p arm in IMPORT 169 did not affect the *CTNNB1* locus (3p22.1), however it did affect a region immediately neighbouring the location (3p21.31–14.3). As IMPORT 169

123
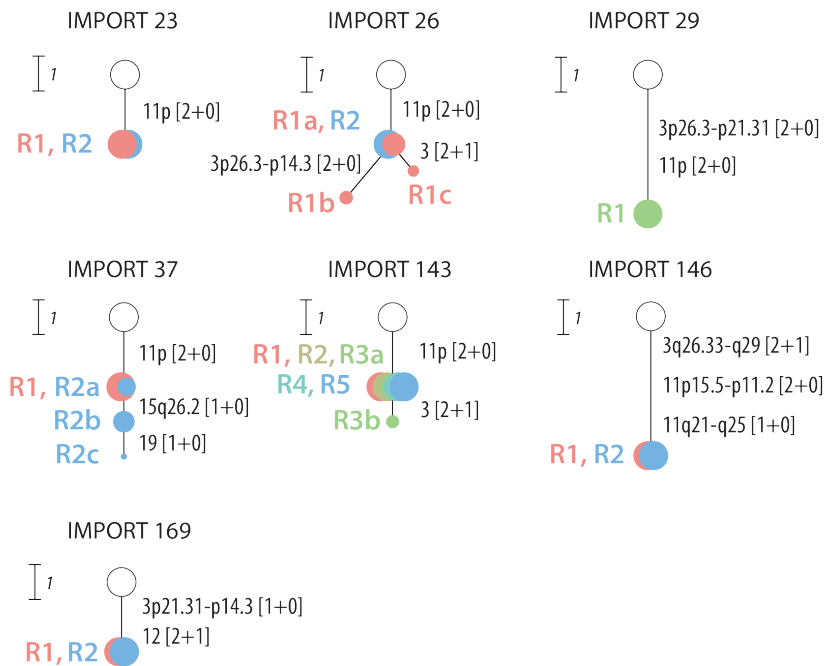
Figure 5.4: Stromal WTs. Seven tumours in the series were unilateral stromal tumours. Generally these tumours showed few events (2-5 CNA events), however there was a clear evidence of CNA preference in these cases. In all but one case 11p LOH occurred and was clonal. In 5 cases (26, 29, 143, 146, 169) chromosome 3 aberrations were also present, affecting the p arm of the chromosome in four of these cases. In two cases these changes occur after the 11p LOH event and there was no evidence for chromosome 3 aberrations occurring before 11p LOH.

lacks an 11p LOH event, a *WT1* or a *CTNNB1* mutation, it may represent an alternative stromal tumour phenotype.

5.3.4 Epithelial tumours

There were five unilateral WT cases which were diagnosed as epithelial type (36, 75, 88, 98, 166). These tumours had relatively few events (Fig. 5.5). The case with the most events (166) presented with three clonal CNAs implicated in WT (1p–, 1q+ and 4q LOH) as well as a subclonal +2. The only single sampled case (98) possessed a –22. The remaining three cases (36, 75, 88) all presented with a clonal 19q LOH event as the only CNA, this is further explored in Section 5.7.
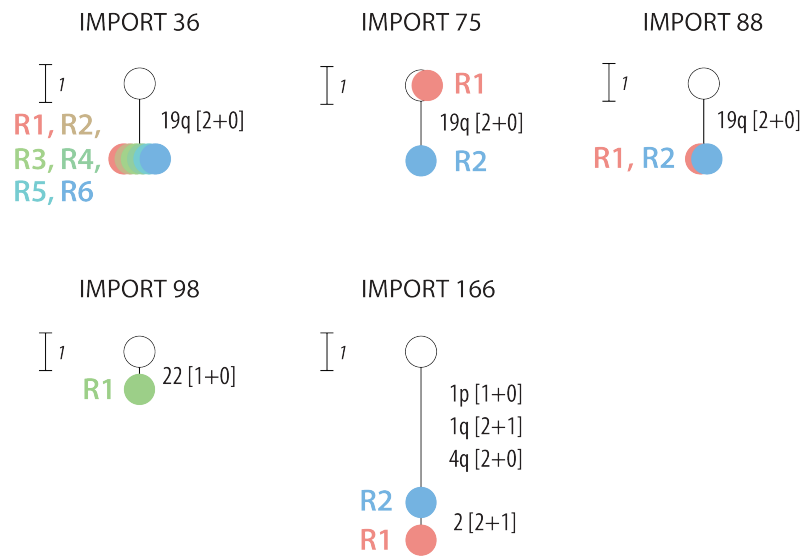
Figure 5.5: Epithelial WTs. There were five cases of unilateral epithelial type WTs in the series. They contained relatively few events (1–5). Four of the five tumours only contained a single CNA and in three cases this was a 19q LOH event (36, 75, 88). Only one case was not multi-sampled (98). The case with the most changes was IMPORT 166.

5.3.5 Regressive tumours

Regressive type WTs are defined by the presence of >66% necrosis histologically, which indicates response to chemotherapy. Due to this, sampling was limited in the six unilateral cases which presented with regressive WTs in the PKC66 set. Half of these cases were not multisampled (22, 27, 28) and of these only one was sampled at the point of nephroctomy (22) (Fig. 5.6). This case did present with 25 events including a non-clonal +8 and –22. The other two cases were sampled at the point of diagnosis and relapse respectively (28, 27). The relapse sample presented with three typical WT CNAs – 1q+, 14q LOH and 16q– – indicating that these CNAs may provide resistance to treatment in this case. Of the multisampled cases one case presented with five events, all of which were clonal (55). The remaining cases presented with no or few events (91, 145, respectively). Overall, there was no obvious pattern in this subtype.
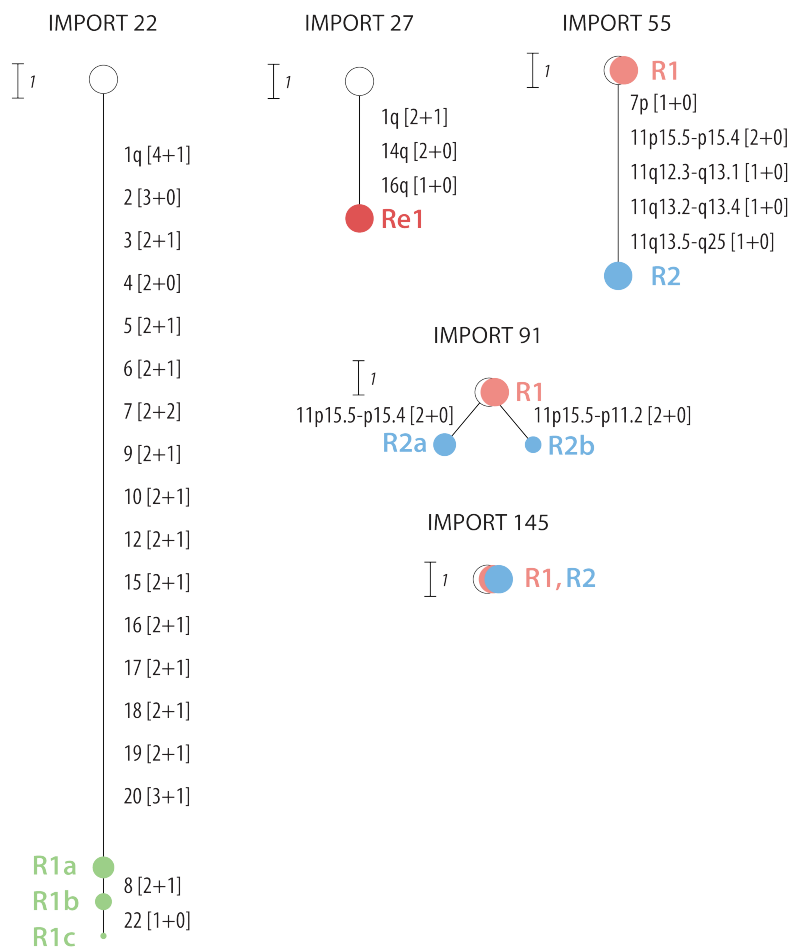
Figure 5.6: Regressive WTs. Six cases were unilateral regressive type WTs. One case (28) was only sampled once with a biopsy sample and is not included in this figure. One case only contains a relapse sample (27) noted as 'Re1'. Of the cases with initial nephrectomy samples (22, 55, 91, 145), only one is not multi-sampled (22). Generally, the patterns of evolution are varied. IMPORT 22 has many events (25), yet all other cases have 6 or less and IMPORT 145 shows no events. There is within sample clonal mixing detected IMPORT 22, meaning more than one clone is observed despite being single sampled. IMPORT 91 shows evidence for the mixing of two clones sharing no copy number changes.

5.3.6 Focal anaplastic

Focal anaplasia was the rarest WT diagnosis in the PKC cohort and I only analysed one case with a single sample (Fig. 5.7). The case (43) displayed 10 clonal copy number gain events, including a 1q++ [3+1]. This case may not be representative of focal anaplastic tumours.

126
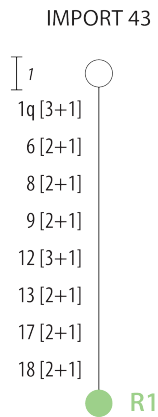
Figure 5.7: Focal Anaplastic WTs. There was just a single case of focal anaplastic WT in the series which showed 10 events, all of which were copy number gains. There was only a single sample taken in this case which did not have detectable clonal mixing.

5.3.7 Diffuse anaplastic tumours

Diffuse anaplastic tumours are chiefly characterised by high numbers of copy number events and clear branched evolution (Fig. 5.8 and 5.9). IMPORTs 8, 54, 59, 106 and 170 show apparent diversification involving many CNAs. Interestingly, one region (R3) in IMPORT 54 appears to show no shared CNAs with the other samples, which will be explored further in Section 5.10 (Fig. 5.23). IMPORT 170 has a tetraploid genome, reaffirming the concept that diffuse anaplastic evolution can alter the entire genome (ploidy of 3.72 – R1, and 3.65 – R2). IMPORTs 10 and 233 also display large number of changes, however the phylogeny is mostly linear. In IMPORT 233 the only single CNA (+8) is unique to R2 whereas R1 has additional events. In stark contrast to other cases, two cases (20, 83) show relatively few CNAs (7 and 4 events respectively) and little diversity, indicating that these aspects are not completely archetypal in these tumours. However, I cannot rule out that the most diverse clones were simply not sampled in these cases as other, diverse diffuse anaplastic tumours, display living ancestors with fewer events compared to the most evolved in the tumour (10, 106). Three diffuse anaplastic cases (10, 83, 106) had clonal 19q LOH. All diffuse anaplastic tumours bar one case (83) displayed 17p LOH involving the *TP53* locus and this change was determined to be not clonal in 4 cases (8, 10, 106 and 233) and apparently in an independent tumour in IMPORT 54.

127
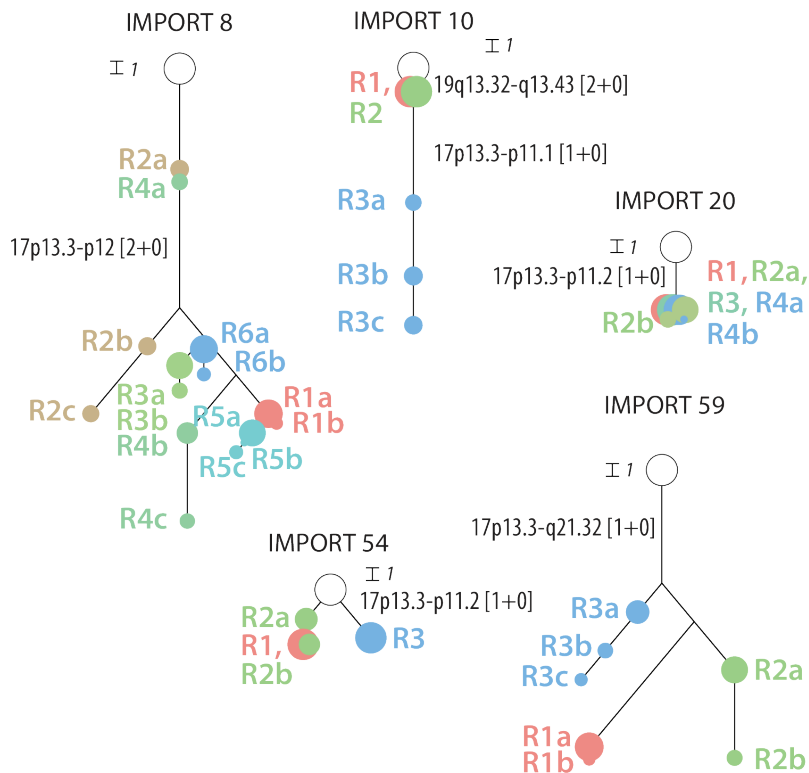
Figure 5.8: Diffuse anaplastic WTs have the potential to display the highest number of copy number events seen in WTs. In this figure and Fig 5.9, the length of the branches is one-third of that in other figures in this section (as displayed by the scale bar next to each phylogenetic tree). Many cases of diffuse anaplastic Wilms' tumours show large amounts of branching and many events (IMPORTs 8 and 59). However, diffuse anaplastic WT can also be largely linear such as in IMPORT 10, despite having many CNA events. Furthermore, large numbers of CNAs is not a prerequisite for the development of a diffuse anaplastic Wilms' tumours as displayed by the low numbers of events seen in IMPORTs 20 and 54. Only clonal 19q LOH events and CNAs affecting the *TP53* locus are displayed.
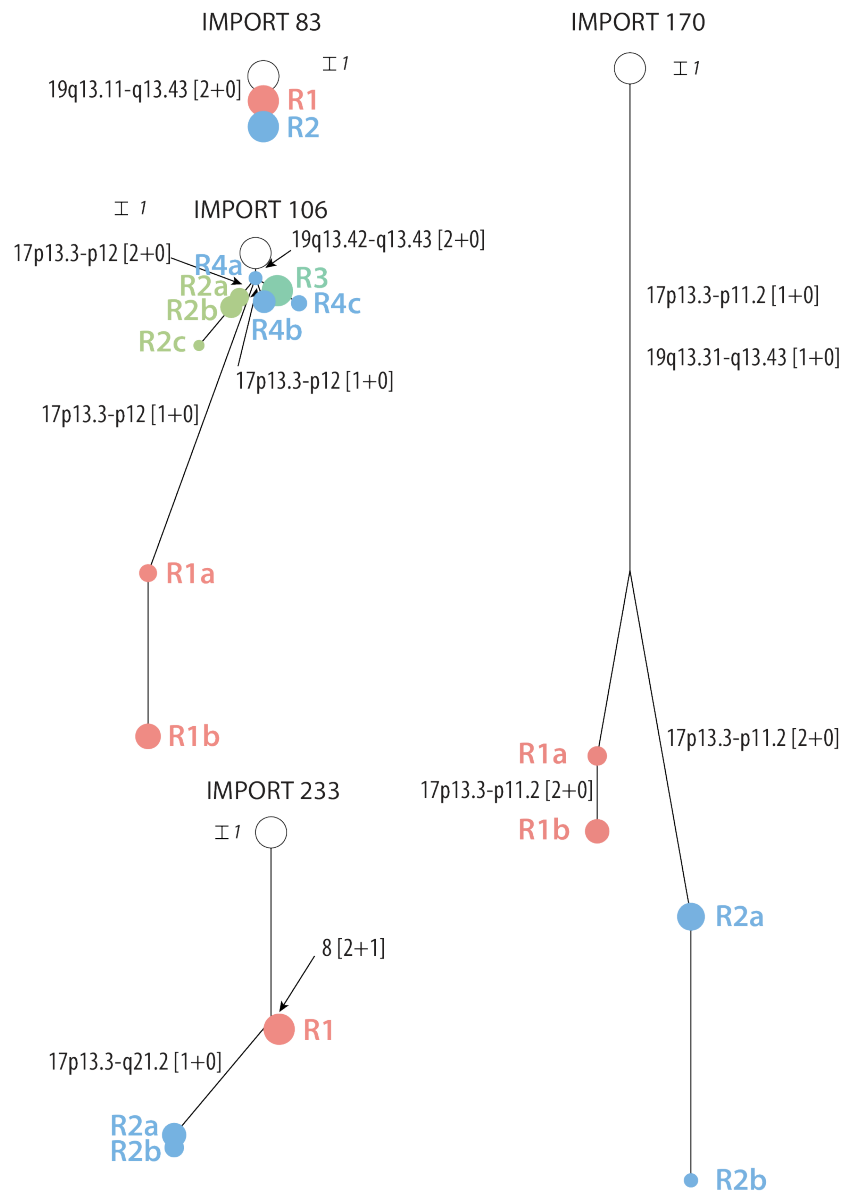
Figure 5.9: In this figure and Fig 5.8 the length of the branches is one-third of that in other figures in this section (as displayed by the scale bar next to each phylogenetic tree). Many cases of diffuse anaplastic WTs show large amounts of branching and many events (IMPORTs 106 and 170). IMPORT 170 shows likely WGD event. However, diffuse anaplastic Wilms' tumour can also be largely linear such as in IMPORT 233, despite having many CNA events. Furthermore, large numbers of CNAs is not a prerequisite for the development of a diffuse anaplastic WTs as displayed by the low numbers of events seen in IMPORT 83. Only clonal 19q LOH events and CNAs affecting the *TP53* locus are displayed.

5.3.8   Blastemal tumours

Blastemal type WTs present with a large range of diversification (Fig. 5.10). Four of the six unilateral blastemal WTs were largely homogeneous (25, 58, 80, 99). These tumours were mostly dominated by a single clone, although one region in IMPORT 58 represented an ancestral clone to the clone which dominates the other samples. Unfortunately one case (25) was not multisampled, limiting our detection of diversity, however it presented with several changes, all of which were clonal. The remaining two cases (82, 192) show large diversification and many events. The six samples in IMPORT 192 all contained at least one unique clone. IMPORT 82 showed large diversity in the two samples taken. Indicating in this small set that generally blastemal tumours are capable showing vast heterogeneity but may also be dominated by a single clone, presumably produced by a selective sweep. Generally these tumours exhibited large numbers of events (Fig. 5.22).
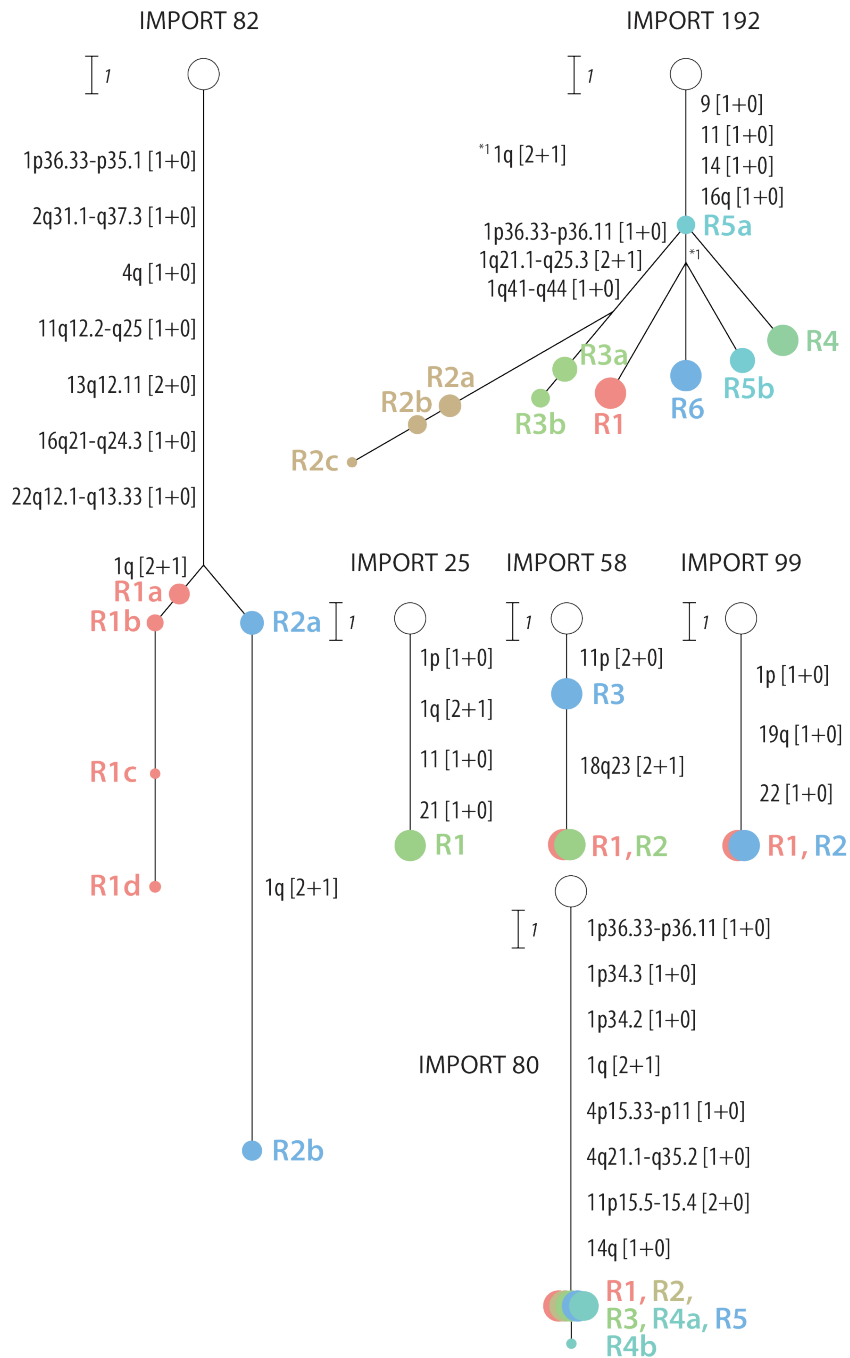
Figure 5.10: Blastemal type Wilms' tumours show a range of evolutionary patterns similar to those seen in mixed type Wilms' tumours. Some cases have many events and branching (IMPORTs 82 and 192). Yet they may also show a similar process of evolution showing little heterogeneity (IMPORT 25, 58, 99 and 80). Only a subset of CNAs are displayed.

131

5.3.9   Bilateral WT cases

Fourteen cases in the PKC66 series were bilateral WTs. These tumours are difficult to group and summarise as individual subtypes as a single patient may present with a different tumour subtype in each kidney. Overall, as in Chapter 3, there was no evidence to suggest that bilateral tumours were genetically related, although there was evidence of convergent evolution as explored in Section 3.7. I chose to summarise these tumours individually, referring back to subtype specific observation in unilateral tumour subtypes.

In seven of the bilateral cases, samples were taken from both the right and the left kidney (4, 7, 9, 16, 21, 30, 47) (Fig. 5.11). In four of these seven cases the histological subtype on both sides was the same (4, 9, 30, 47), for the remainder of cases each side presented with a different histological subtype.

IMPORT 4 presented with two mixed subtype tumours and no shared CNAs. All events were clonal in both tumours expect for a subclonal 1q+ event. Both tumours in IMPORT 47 only exhibited an 11p LOH event, but with different breakpoint boundaries for each side. These tumours were both the stromal subtype and the presence of a clonal 11p LOH event is in line with our observations in the unilateral cases. The left and right tumours of IMPORT 30 were both the epithelial subtype. Again, in line with observation in unilateral epithelial tumours, both sides had only a clonal 19q LOH event with no shared boundaries between the lateral sides and within the left tumour there was also no shared breakpoint between the 19q LOH events in the two samples taken, this is explored further in Section 5.7. IMPORT 9 was mixed type in both sides. The right side contained no CNAs and the CNAs presenting in the left side were not shared, indicating the presence of two independent tumours in the left kidney of IMPORT 9.
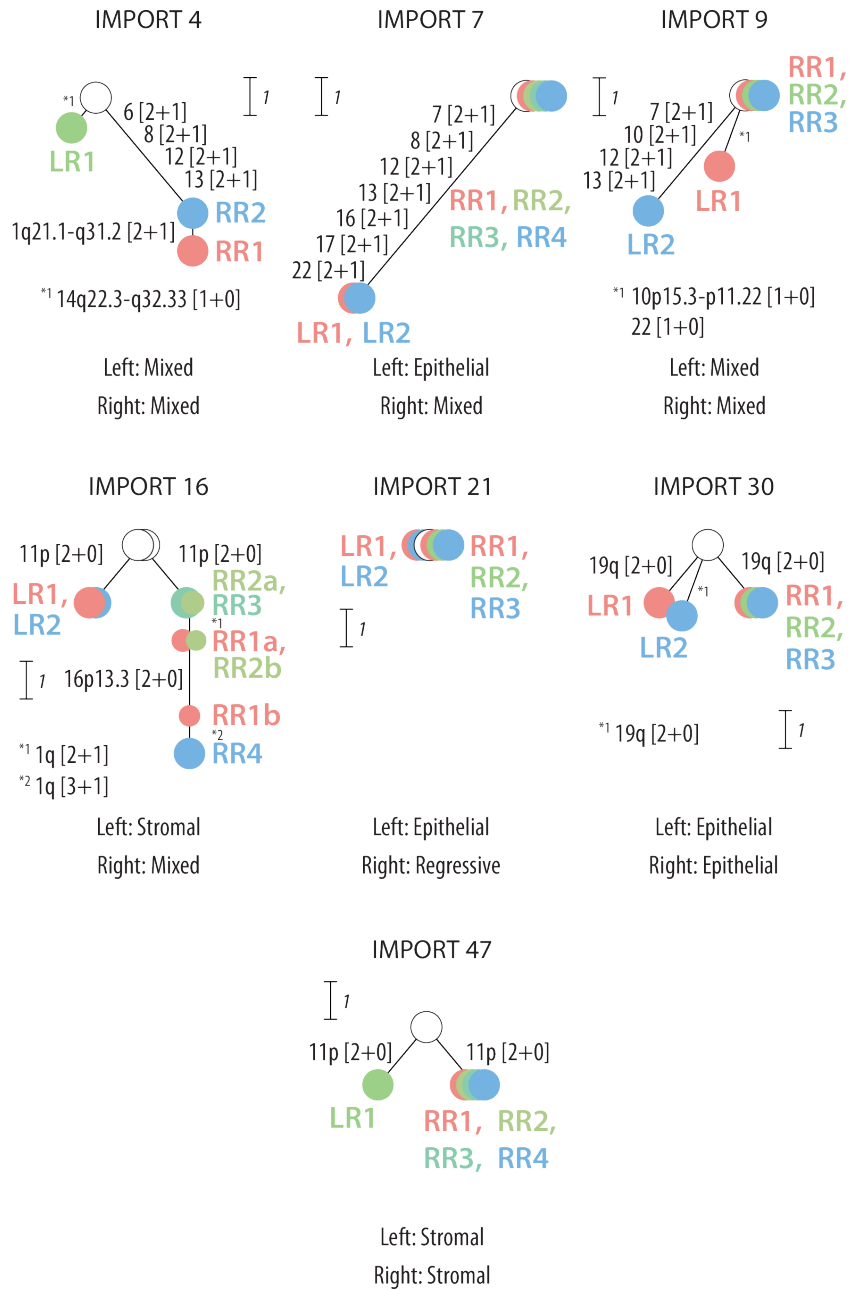
Figure 5.11: Bilateral WTs with both sides sampled. Seven cases of the fourteen bilateral WT cases in our series were sampled from both sides. There was no evidence that bilateral tumours were related to each other (no CNAs with shared break points between the tumours). In four cases the subtypes were identical in each kidney (4, 9, 30, 47). In IMPORT 30 and 47 the same type of CNA was acquired in each kidney also, indicating convergent evolution of two independent tumours within a single patient.
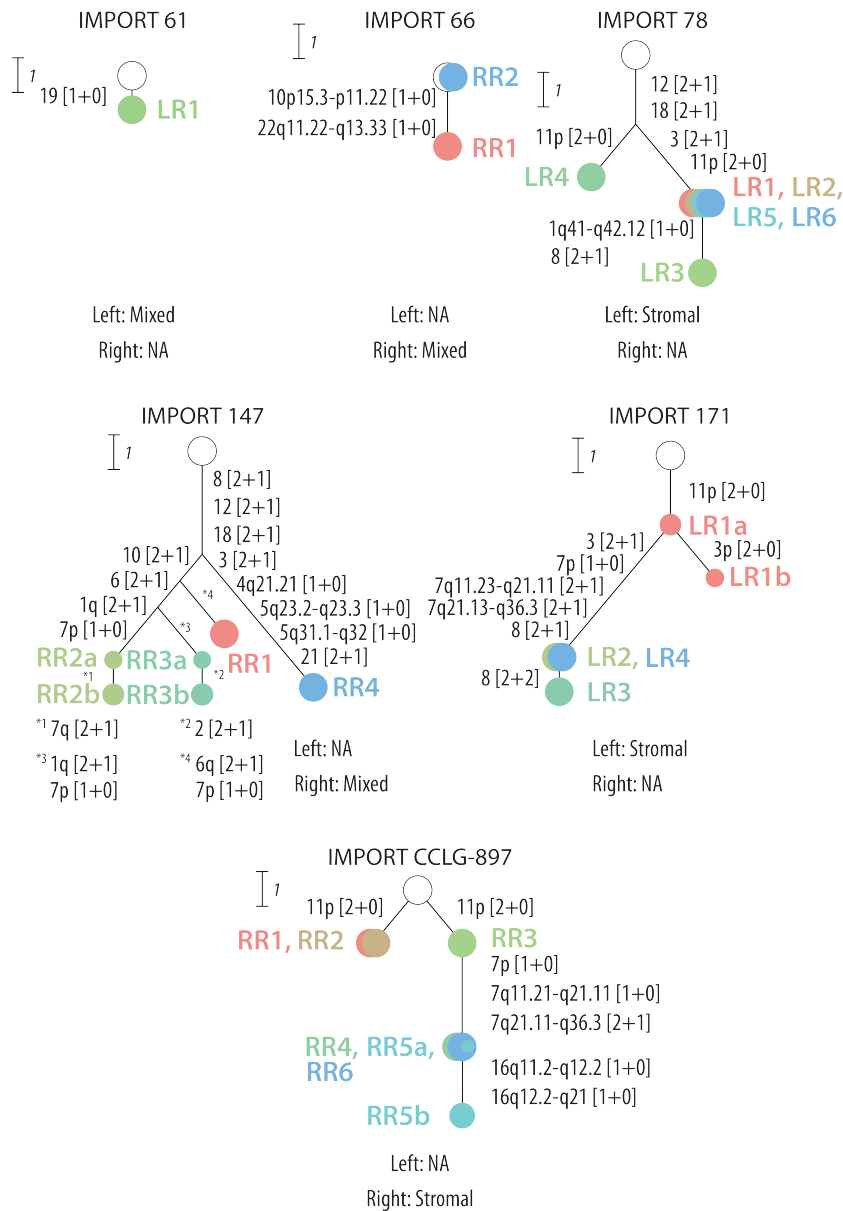
Figure 5.12: Bilateral WTs with only one side sampled. In total there were seven cases that were bilateral WTs that were only sampled from a single side. Four of these were mixed type tumours (61, 66, 147, 153) and three stromal tumours (78, 171, CCLG-897). IMPORT 153 was only sampled at diagnosis with a biopsy and is not included in the figure. Generally the patterns of evolution are similar to unilateral tumours of the same type. The mixed type tumours show a range of events and degree of branching and stromal tumours show 11p LOH events often followed by chromosome 3 events (78, 171).

Three cases had different histological subtypes on each side (7, 16, 21). IMPORT 21 presented with a left-sided epithelial tumour and a regressive right-sided tumour. Neither side displayed CNAs. IMPORT 7 also presented with a left-sided epithelial tumour but the tumour presented with 7 whole chromosome gains. The right-sided tumour was mixed subtype and did not exhibit any CNAs. IMPORT 16 was stromal type in the left tumour and mixed type in the right. Both sides presented with 11p LOH of different breakpoint boundaries. The right, mixed type tumour, acquired additional CNAs.

The remaining seven bilateral cases were only sampled on a single side (61, 66, 78, 147, 153, 171, CCLG-897) (Fig. 5.12). Two of these cases (61, 153) were not multisampled and one case only was sampled at biopsy. Both these tumours were mixed type and only IMPORT 61 showed a single CNA (–19). Two more of the remaining cases were also mixed type tumours (66, 147). IMPORT 66 was sampled twice and one sample presented with no CNAs and the other with only two. Alternatively, IMPORT 147 displayed many events and the majority of these were non-clonal (83.3%). These two cases displayed a range of events and diversification similar to those seen in the unilateral mixed WTs.

Lastly, the remaining three bilateral cases sampled only on a single side were all stromal type (78, 171, CCLG-897). All three cases contained 11p LOH events and in two cases (78, CCLG-897) exhibited breakpoint boundary differences. The 11p LOH events affected both the *IGF2* and *WT1* locus in all but one convergent 11p LOH event (LR4 of IMPORT 78 that only affected the *IGF2* locus – 11p15.5-p15.4). Two cases also presented with 3p aberrations (78, 171), with no evidence to suggest that these changes occurred before the 11p events – showing that stromal tumours in bilateral cases also appear to be restricted to the model of tumour development displayed in the unilateral cases. The presence of convergent evolution in each of these cases further supports this, as explored in Section 5.6. Each stromal case displayed 9–10 CNA events and diversification.

5.3.10 non-WT cases

Four cases were not WTs and all cases were multisampled when also including FB samples (Fig. 5.13). Three of four tumours displayed no CNAs (DHPN, Diffuse Nephroblastomatosis and Metanephric Adenofibroma). The only non-WT displaying CNAs was a case of renal cell carcinoma. This case was striking as it presented with four tumour pieces with seemingly unrelated clones in each region (although no region presented with more than 2 CNAs), this is further explored in Section 5.6.
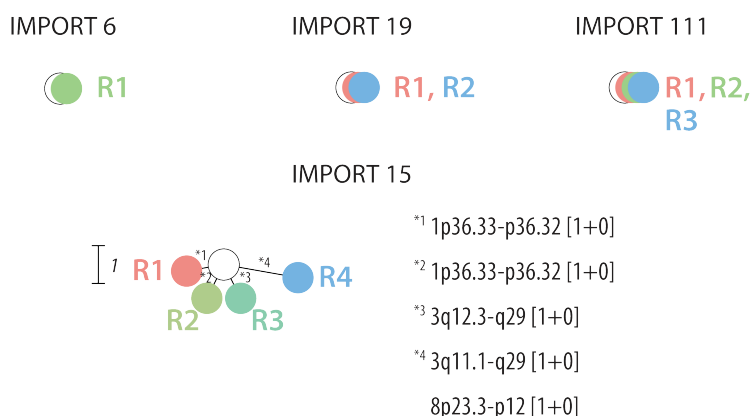
Figure 5.13: non-WT paediatric kidney cancers. Four cases were not WTs. These diagnoses included diffuse nephroblastomatosis (6), renal cell carcinoma (15), metanephric adenofibroma (19) and diffuse hyperplastic perilobar nephroblastomatosis (111). The only case showing CNAs was IMPORT 15 which had five events. In the four samples from this case, no CNAs were shared (the 1p36.33–p36.32 event affects different alleles between R1 and R2), indicating the possible presence of unrelated tumours. IMPORT 15 was diagnosed in a patient aged ~14.5 years old, much older than the majority of patients in the series.

5.4 Mapping clonal evolution against spatial location

For nine cases we had photographed the primary tumour post-nephrectomy and annotated the sampled regions (8, 12, 16, 20, 54, 78, 147, 171, 192), allowing me to compare the locations of the regions taken for analysis with the inferred phylogeny of the cases (Fig. 5.14). Here I will present the mapping of the clones in six cases (8, 16, 78, 147, 171, 192). Generally these images allowed us to examine the trajectory of evolution against spatial position which could represent clonal movement during tumorigenesis.

136

In IMPORT 8 the ordering of clones in evolutionary history matches the physical location of the tumour regions (Fig. 5.14A). Regions R2, R3 and R6 are present at the bottom of the image and represents clones that are clustered together phylogenetically. Regions R1 and R4 present with clones further evolved and are present in the middle of the tumour. Some of the most evolved clones are present in R5 at the top of the tumour image. This generally indicates that as the clones gain more changes as they are moving towards the top of the image. Interestingly however, R2 (bottom) and R4 (middle) contain some living ancestors of these latter clones.

IMPORT 16 is a bilateral tumour and therefore we have two images, one from the right and left kidney (Fig. 5.14B). The left kidney presents a small tumour and the two samples contain the same 11p LOH event. In the right kidney the two regions with the most ancestral clones are in the middle (RR2 and RR3) and RR2 contains a subclonal 1q+ that is present in the other regions. More evolved clones are present in the regions at the periphery (RR1 and RR4) and RR4 contains the most evolved clones.

A single clone dominates the majority of the tumour in IMPORT 78 (LR1, LR2, LR5, LR6) (Fig. 5.14C). However, there are two unique clones in LR3 and LR4 (left side of the image). Interestingly LR3 contains a clone that has acquired additional changes compared to the dominant clone, whereas LR4 contains a clone that uniquely branched from an ancestor of the dominant clone (convergent 11p LOH).
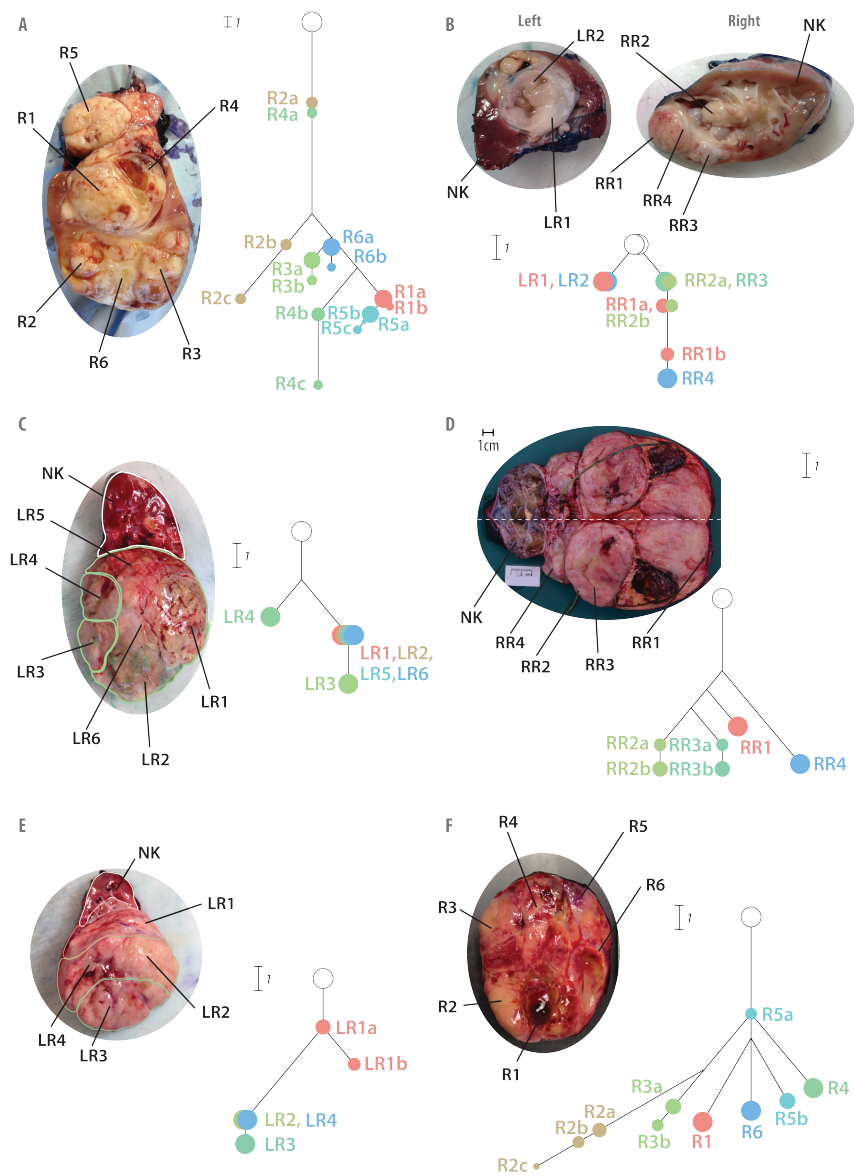
Figure 5.14: By comparing the inferred phylogenetic trees of tumours to spatial location of samples, it is possible to hypothesise the the genetic history of the tumour and the movement of the clones in 3 dimensional space. IMPORT 8 (A) shows the locations of the six tumour regions compared to the phylogenetic relationship, showing that, for instance the lower part of the image may represent less developed clones. IMPORT 16 (B) displays the phylogenetic events present in regions taken from a bilateral WT. IMPORT 78 (C) displays the dominance of a single clone present in four regions (LR1, LR2, LR5 and LR6), but also describes the location of other genetically different samples. IMPORT 147 (D) shows a tumour in which the halving of the tumour is not complete and the dashed white line shows the axis at which the cut is made.

(continued)

138

Figure 5.14: This image has a scale and shows that despite being ~1cm apart RR2 and RR3 show branching and interestingly, convergent evolution. The image also shows that RR4 is divergent from RR1–3. IMPORT 171 (E) displays a relationship between distance from the normal kidney and acquisition of CNAs, where LR3 contains the most mutated clone and is furthest from the normal kidney. IMPORT 192 (F) shows a complex 'flowering' pattern in which generally speaking, only R2 and R3 show a higher degree of relatedness (shared events) relative to the other samples.

The tumour in the image for IMPORT 147 shows a tumour not fully sliced in half, therefore tissue that is reflected across the white dashed line represents tissue which was originally in contact (Fig. 5.14D). The tumour has a branched evolutionary pattern. RR4 shows an early divergence and was sampled from a location closest to the normal kidney. RR1 then shows the next point of divergence and incidentally is sampled from a point furthest from the normal kidney tissue. RR2 and RR3 which are also closely genetically related are also located close spatially (~1cm). Incredibly, this there is clear convergent between these clones in such a small area, indicating that local environment may influence selection (Section 5.6).

The most convincing example of spatial location being related to evolutionary history is present in IMPORT 171 (Fig. 5.14E). Here, as clones gain more events they move further away from the normal kidney. The only clonal change in LR1 is 11p LOH. However, LR2 and LR4 develop an additional five changes and are further from the normal kidney. The most evolved clone is present in LR3, the furthest region from the normal, which has an additional +8.

IMPORT 192 presents a more complicated picture of spatial evolution (Fig. 5.14F). All clones in the case branch from a living common ancestor present in R5 in a 'flowering'-like pattern. From this clone, one clone branches to form the only clone in R4. Three clones are clustered phylogenetically as they share a 1q+ event but also develop further unique changes; these include R5b, R6 and R1. R1 exhibits the most events following the branching point after the shared 1q+. Furthermore, related clones present in R2 and R3 share several unique changes and branch from

the ancestral clone of R5a. These are also spatially separated from the other regions, encompassing most of the tissue on the left side of the tumour image. The MRCA present in R5 may represent a highly plastic clone from which the explosion of diversity then spreads across the tumour.

5.5   Pre-chemotherapy biopsies reveal pre-treatment clones

By assaying needle-biopsy tumour samples taken at diagnosis we were able to gain some insight into the nature of clones present in the sample prior to treatment (chemotherapy) and tumour removal. In total there were eleven biopsy samples taken from 10 patients (~15% patients).

Five of the eleven biopsies had no CNAs. In two cases the biopsy sample was the only tumour sample available (28, 153), indicating that in these tumours there were either no CNAs – a possible scenario in WT, or that these samples were potentially not tumour. Two diagnosis biopsies were taken from tumours displaying no CNAs in the nephrectomy sample also, therefore displaying a CNA profile equal to the removed tumour post-chemotherapy (6, 7).

IMPORT 10 was the only case in which two pre-chemotherapy biopsies were taken. One sample contained no CNAs, a rare occurrence in a case diagnosed as diffuse anaplastic, whereas the other piece contained a large number of genetic events comparable with the most evolved clone in the tumour sampled at nephrectomy. This biopsy also appears to potentially be a separate tumour, not sharing the early clonal events seen in the samples taken at nephrectomy. However, if it is a separate tumour, it shows convergent evolution of 2p42.2–p11.1 and 17p loss (Fig. 5.15).
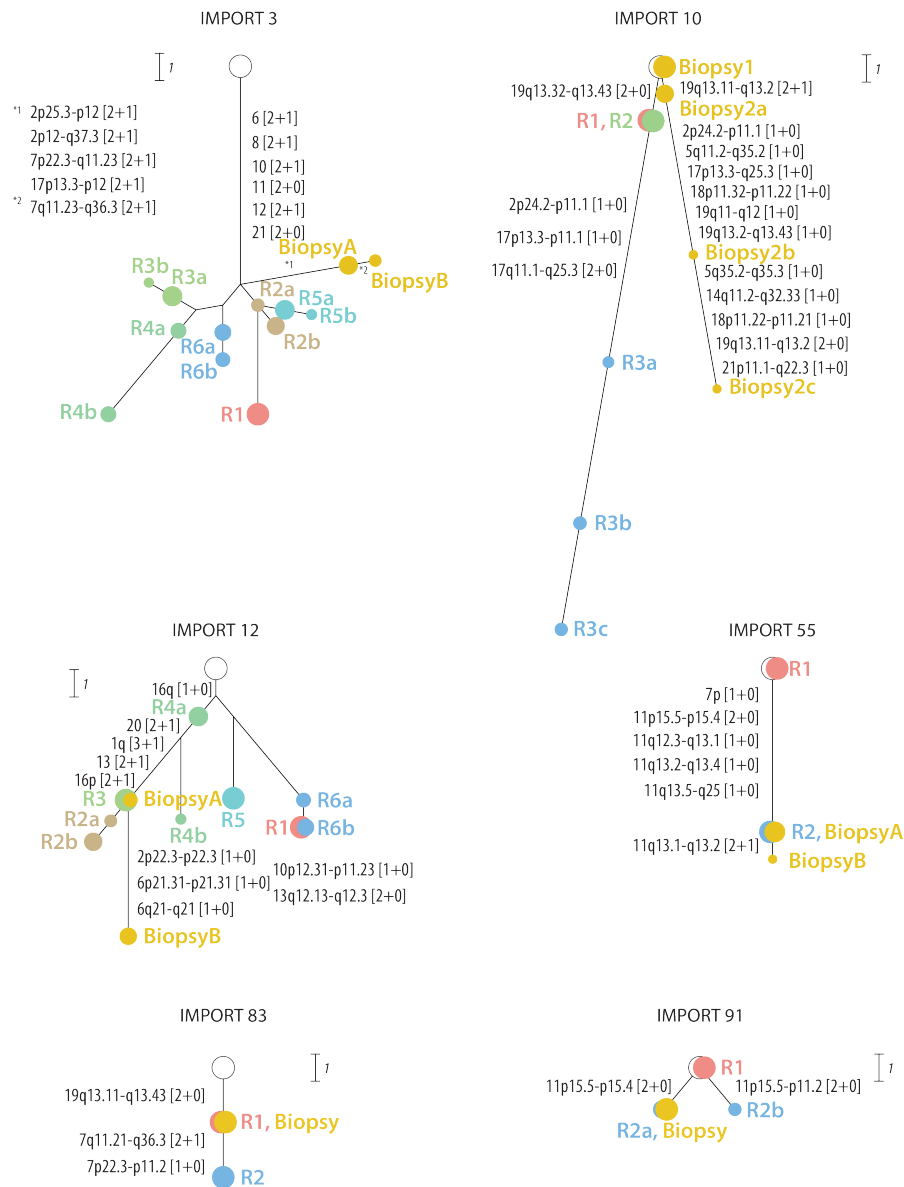
Figure 5.15: Six cases had biopsy samples that had CNAs which were comparable with the primary nephrectomy samples. Four of six of the cases contained clones in the biopsy which were identified in the primary nephrectomy (IMPORTs 12, 55, 83 and 91) and in these samples clones with additional CNAs were also sometimes present, for example, the subclone in the biopsy of IMPORT 12 (BiopsyB). In IMPORT 3 the biopsy sample shares clonal CNAs with the rest of the primary nephrectomy samples but is divergent from the other samples. In IMPORT 10 Biopsy2 is potentially a product of an independent tumour with convergent evolution of events in chromosome 2 and 17, however this is not the only possible explanation of the relationship of these events.

141

There were also several CNAs in the FBs of IMPORT 3 and 55. The FB sample in IMPORT 3 represented a clone that is similarly evolved compared to most FN samples and branches from the sample MRCA as the other FN clones. In IMPORT 55 only one of the FNs contains CNAs (R2), however it shares all of these changes with the CNAs present in the highest cancer cell fraction clone of the FB (BiopsyA), suggesting the FN and the FB represent the same clone.

In three more cases the biopsy sample contains clones identical to clones in a sample taken from the removed tumour. In IMPORT 12 the clonal CNAs in the FB sample are identical to the clonal changes R3, allowing us to map this pre-surgery biopsy spatially onto the removed tumour. This FB sample also contains several additional non-clonal copy number changes indicating the presence of a sub-population potentially not present post-treatment.

In IMPORT 83 the FB sample contained the same CNAs as one of the two FN samples (19q LOH) and represented an ancestral clone to the other further evolved region, proving that this ancestral clone was present before treatment. In IMPORT 91 the FB contains an 11p LOH event that appears to be mixed together with another 11p LOH event in the tissue sample from the post-treatment primary tumour.

5.6 Convergent evolution and mirrored subclonal allelic imbalance of copy number changes

Convergent evolution can be thought of as separate units of evolution acquiring genotypic/phenotypic traits in parallel due to selective pressure strongly favouring those traits. In the context of cancer, this refers to separate clones acquiring the same/similar mutation, in this study a CNA that affects the same genomic region. In the context of PKC, parallel evolution may occur between separate tumours in separate kidneys, separate tumours in the same kidney/mass and parallel evolution of related clones in the same tumour.

5.6.1 Convergent evolution between bilateral tumours

Three bilateral cases show clear convergent evolution between tumours in separate kidneys. Two cases show separate acquisition of 11p LOH in the separate sides (16, 47, Fig. 5.11) developing tumours dominated by stroma (three stromal subtype tumours and a mixed tumour with 60% stroma – IMPORT 16 right tumour). These tumours show clearly that 11p LOH is an important initiating event, particularly in tumours dominated by stroma. IMPORT 30 develops two bilateral epithelial tumours. The only event in each sample (2 regions from the left, 3 regions from the right) was a 19q LOH event (Fig. 5.16A). These events have three separate breakpoints indicating three separate events. All regions taken from the right share a breakpoint but it is unique from the left breakpoints. The two regions from the left have separate breakpoints. This indicates that 19q LOH may have occurred three times producing potentially three separate tumours – two in the left, one in the right. Clearly there is strong selective pressure in this individual to develop a 19q LOH event that then produces an epithelial WT.

5.6.2 Convergent evolution between independent tumours in the same mass

In addition to the possible presence of two tumours in the left kidney of IMPORT 30 that show convergent evolution. There were six other examples of convergence in two independent tumours in the same mass – similar to the phenomenon described in Section 3.8. The clearest evidence for separate tumours which are convergent based on copy number alterations are in IMPORT 54 and IMPORT CCLG-897. In IMPORT 54 there are no shared CNAs between R5 and the other two samples R1 and R2 with identical breakpoints (Fig. 5.23). However, there are CNAs affecting chromosome 16q and 15 in R5 that are different from those breakpoints present in R1 and R2, indicating clear convergence for these alterations. These convergent changes affect chromosome 15q14–q21.2 and 16q21–q24.3 in all samples. This case will be further explored in Section 5.10. IMPORT CCLG-897 is a bilateral case, but samples were only taken from a single side. All samples contain a 11p

LOH however, two regions R1 and R2 only contain an 11p LOH event, with different breakpoint boundaries to the 11p LOH event in R3–R6, indicating that R1 and R2 could represent a separate tumour from the same kidney and both tumour have converged onto an 11p LOH event that affects 11p15 and the *WT1* locus (Fig. 5.12).

IMPORT 15, a renal cell carcinoma, was sampled four times. Interestingly, each sample contains unique CNAs, hinting at the possible development of four separate tumours and two of the CNAs are convergent, yet definitely separate events. These events affect chromosome 3q and 1p36.33–36.32. These changes are certainly convergent as they affect separate alleles, a phenomenon termed 'mirrored subclonal allelic imbalance' and explored in Section 5.6.4. The regions affected by a 3q loss are R3 and R4, in both regions chromosome 3q12.3–q29 is affected. Interestingly, the losses in R1 and R2 both have identical breakpoints, affecting exactly 1p36.33–36.32, yet different alleles are lost in each region. This may have been two separate initiating events or potentially could have been a single event that caused two alterations with identical boundaries (Fig. 5.13).

The only FN sample taken from IMPORT 91 appears to potentially contain a mixture of two clones, each with a separate 11p LOH event, indicating the mixture of two tumours in one region but a convergence of 11p LOH. The biopsy sample for this case contained only one of the two 11p LOH events (Fig. 5.15).

The two tumour regions from the left tumour of IMPORT 9 share no CNA events. In both regions a CNA event affects chromosome 10p15.3–p11.22, one is a loss of this specific region and the other is a whole chromosome 10 gain, these changes also affect alternate alleles (Fig. 5.11).

IMPORTs 9, 30, and CCLG-897 are bilateral cases that also appear to have developed two tumours of independent origin in one kidney in addition to the tumour developed in the other (3 tumours in total). This clearly indicates that these patients
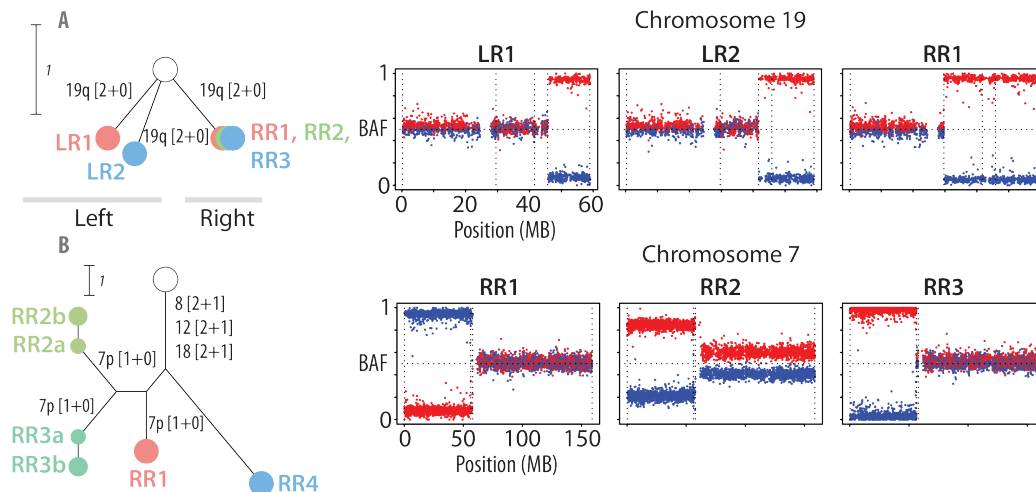
Figure 5.16: Evidence for convergent evolution in WTs. IMPORT 30 (A) represents a bilateral tumour in which an epithelial type tumour has developed in both kidneys. The case also represents convergent evolution both within potentially two tumours (left) and between tumours of a different laterality. The only CNA in this case is a 19q LOH event however, across the five tumour samples taken, three different breakpoints of the chromosome 19q LOH are recorded, one affecting all samples taken from the right (RR1–3) and a unique breakpoint in both samples taken from the left tumour. IMPORT 147 represents the presence of convergent evolution within a single tumour (B). Samples RR1–3 all contain unique chromosome 7 loss events. The losses in RR2 and RR3 have different breakpoints and despite the events in RR1 and RR2 affecting the same breakpoint (insofar as is detectable on the SNP array), the event affects different alleles in both samples (as represented by the colour of the SNPs).

have a predisposition for WT development.

5.6.3   Convergent evolution in the same tumour

Parallel evolution is also possible within the evolution of a single tumour when clones acquire similar changes in parallel. There were several examples of this in the PKC66 cohort. IMPORT 12 shows strong convergence of 1q gain (Fig. 5.2). The change originates three times in the evolution of this tumour, affecting both alleles. This indicates that there is high selection pressure for 1q gain in IMPORT 12. The breakpoint boundaries are also always identical according to the detection of the Illumina® HumanCoreExome array, as each changes affects the whole arm.

IMPORT 147 also shows striking convergent evolution (Fig. 5.16B). Similarly to IMPORT 12, IMPORT 147 shows convergent evolution of 1q gain, again this

change affects separate alleles in R2 and R3. Interestingly, there is also convergent evolution of 7p loss, an event that occurs three times in IMPORT 147, differentiated by the fact these changes affect both alleles and possess two different breakpoints. Fascinatingly parallel evolution of 7p loss and 1q gain can be observed in R2 and R3, two regions that are ~1cm apart in a tumour with a length of ~15cm, indicating that this particular tumour environment may be selecting for these mutations (Fig. 5.14D).

Chromosome 11p LOH evolved in parallel in IMPORT 78, potentially indicating a late selection for the initiation of stromal tumour evolution (Fig. 5.12). I also saw parallel evolution of 3p CNAs in two other stromal cases following an 11p LOH, indicating the convergent acquisition of changes required to achieve stromal tumour development (26 – Fig. 5.4, 171 – Fig. 5.12).

Chromosome 1q+ also appears to be convergent in IMPORTs 3, 82 and 192. Additionally, IMPORT 106 potentially shows convergent evolution of 17p LOH in three separate clones.

5.6.4   Mirrored subclonal allelic imbalance

In several cases parallel evolution was confirmed by observing mirrored subclonal allelic imbalance (MSAI), that is defined by the same region of the genome being altered on separate alleles during evolution [94]. One of the most distinguished examples of MSAI being present in tumour evolution was in IMPORT 85 (Fig. 5.3). In this case MSAI affected whole chromosome arm gains in four chromosomes (6, 8, 12, 18). Displaying clear examples of parallel evolution, yet these clones are related to each other based on sharing allele-specific 11 and 19 LOH and 9 gain.

In total 12 cases are affected by MSAI, five mixed-type tumours, two blastemal type, four diffuse anaplastic and one renal cell carcinoma. In half of these cases MSAI affects chromosome 1 (five 1q changes and one 1p change). This is three or more times more frequent than any other chromosome indicating the *non*-allele
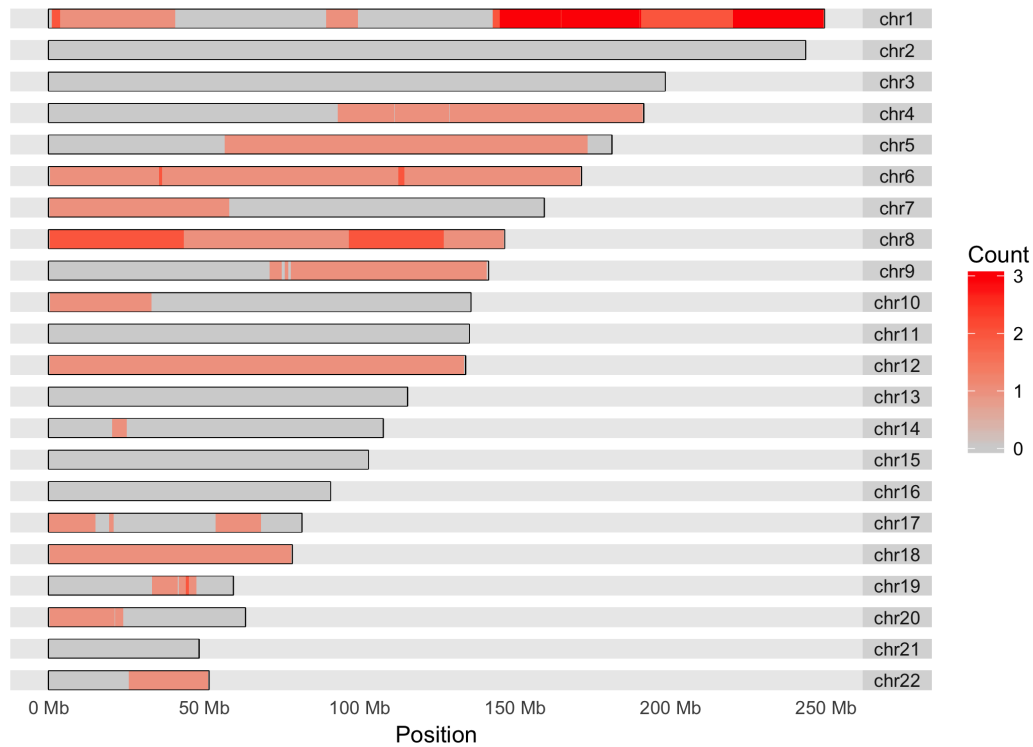
Figure 5.17: MSAI was detected in 12 cases in the PKC66 series. It was generally distributed across the genome. The most common location for MSAI was in chromosome 1q. Six of the twelve cases with detected MSAI contained the presence of MSAI in chromosome 1. Chromosome 11 is not affected by any MSAI events.

specific selection pressure for chromosome 1q changes (Fig. 5.17). Interestingly, despite being often observed multiple times in parallel in the same case, MSAI never affects chromosome 11p LOH, unveiling the allele-specific selection for this change.

5.6.5 Possible evidence for translocations in multiregion SNP array data

By assaying MSAI in SNP arrays from multiple tumour pieces I uncovered evidence for a possible translocation in R5 of IMPORT 192 (Fig. 5.18). Here I have successfully phased the alleles across the multiple regions by identifying SNPs which form the major and minor allele BAF distributions in well separated CNAs in chromosome 1q in IMPORT 192. For CNAs with clear breakpoints, observing the distri-

147
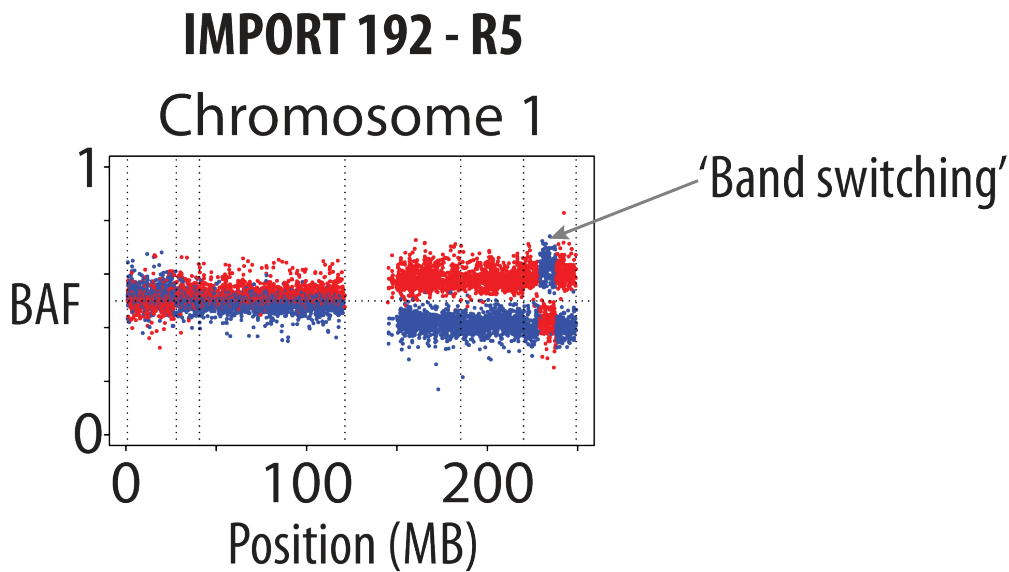
Figure 5.18: Possible evidence for a translocation within the q–arm of a sample (R5) taken from IMPORT 192. Here 'band switching' is observed in which SNPs which have been phased together in other samples have switched in a section of the chromosome arm. This is not detectable by segmentation indicating that a possible translocation may have occurred, which may explain its identical BAF to its surrounding regions. Points in blue and red represent the heterozygous SNPs of the two alleles.

bution of SNPs that are often the minor allele distribution but form the major allele distribution in a subset of samples allowed me to identify MSAI. However, in R5 in IMPORT 192, despite the CNA identified being a whole chromosome 1q gain, a section of the major allele distribution is comprised of SNPs usually found in the minor allele distribution. The fact that this band of SNPs matches the surrounding allele distributions appears to indicate that the switching could have been produced before the CNA by a process similar to mitotic recombination. This event was not taken into account phylogenetically as it did not dominate the entire CNA.

The observation of MSAI in losses that affect identical breakpoint in IMPORT 15 (1p36.33–36.32) in R1 and R2 may have been produced by a similar single event to a translocation or mitotic recombination. Without additional validation I was unable to confirm these events to be evidence of a translocation.
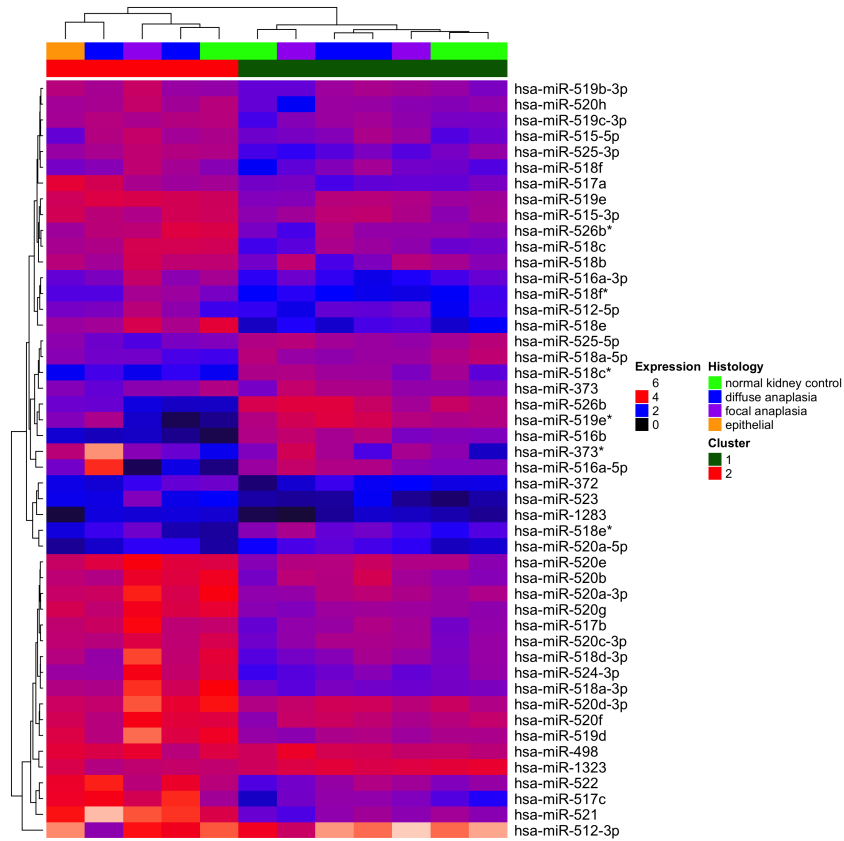
Figure 5.19: miRNA array data generated by Ludwig and colleagues [180] shows that C19MC miRNAs are overexpressed in a subset of epithelial and anaplastic WTs. The heatmap represents expression levels of miRNAs present at the C19MC cluster clustered using hierarchical clustering in both the columns (samples) and the rows (miRNAs). The dendrograms representing the clustering of the rows and columns are displayed alongside the data. Samples were grouped into two clusters as displayed by the bar above the heatmap (red and dark green). The histological subtype of each samples is also displayed above the heatmap. Diffuse anaplastic (blue), focal anaplastic (purple) and epithelial (orange) subtypes were included for analysis. Additionally, normal kidney samples were also included (green). Cluster 1 (dark green) appears to represent the normal expression pattern of the C19MC miRNAs as the majority of the normal samples belong to the group (3/4), cluster 2 (red) shows increase expression levels in the several of the C19MC miRNAs and potentially represents an aberrant C19MC expression pattern.

5.7 Chromosome 19q uniparental disomy is an early event in a subset of Wilms' tumours

I observed LOH affecting the telomeric end of chromosome 19q in 15 cases (10, 30, 36, 37, 59, 61, 75, 82, 83, 85, 88, 99, 106, 108, 170). In 11 cases the LOH

was clonal (subclonal in 37, 59, 82, 108). Of the cases in which the chromosome 19q LOH event was clonal, eight of these events were UPD [2+0], the other three were losses (61, 99, 170). Interestingly, of the eight cases in which clonal UPD of 19q was observed, four were epithelial type WT (30, 36, 75, 88), three were diffuse anaplastic type (10, 83, 106) and one was a mixed histology case (85). Additionally, two of these three diffuse anaplastic cases were dominated by epithelial cells (83 – 70%, 106 – 80%).

Remarkably, in all four epithelial tumours a clonal 19q UPD event is the only observed change (IMPORT 30 developed three tumours each only with a chromosome 19q UPD event). Indicating that, from a purely CNA perspective, 19q UPD is sufficient to form an epithelial WT. Additionally, in all three diffuse anaplastic tumours with a clonal 19q UPD event there is at least one FN sample in each in which the aberration is the only CNA observed (10, 83) or the only observed CNA which is completely clonal (106). Potentially these diffuse anaplastic cases may be further evolved epithelial tumours.

By examining the 19q UPD events in each tumour in which they are clonal, the genomic location which is affected can be mapped. In all eight cases the UPD event affects the entirety of chromosome 19q13.43 and in seven cases it also affects all of chromosome 19q13.42. The case in which the UPD event doesn't affect all of chromosome 19q13.42 is the case which lacked a single sample in which chromosome 19q UPD was the only observed event (106). All chromosome 19q13.42–43 clonal UPD events are not affected by MSAI, indicating that the event is allele-specific.

These observations implicate loss of imprinting (LOI) as a potential mechanism for oncogenic effect produced by UPD in this region. I identified the imprinted primate-specific microRNA gene cluster – C19MC – in chromosome 19q13.42 as a candidate for producing the oncogenic effect when expression is aberrant due to UPD. This cluster is aberrantly expressed in other paediatric cancers such as embryonal tumours with multilayered rosettes (a classification that includes tumours
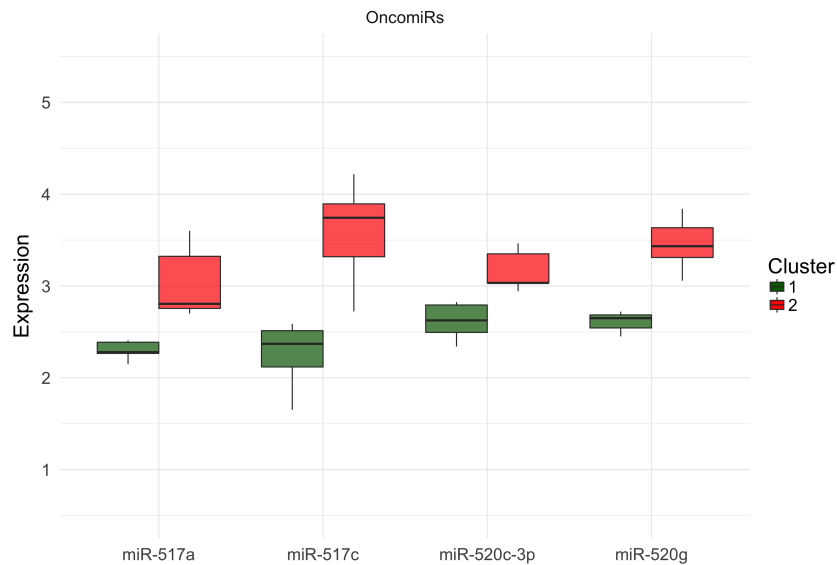
Figure 5.20: C19MC oncomiRs show increased expression in the group of patients with potentially aberrant C19MC expression (cluster 2). Clusters are equal to those that appear in Figure 5.19. OncomiRs are defined by studies detailed by Girardot and colleagues [190]. Data is downloaded from [180].

formerly known as aggressive primitive neuroectodermal brain tumours) [111, 191].

To investigate if this cluster of miRNAs was aberrantly expressed in WTs, I downloaded a dataset of miRNA expression arrays published by Ludwig and colleagues [180]. Here I selected the miRNAs belonging to the C19MC cluster [190]. As 19q UPD appears to preferentially affect epithelial and diffuse anaplastic tumours in the PKC66 dataset, I performed hierarchical clustering specifically on epithelial and anaplasic subtypes, as well as normal sample controls taken from WT patients (Fig. 5.19). The result showed two clear clusters, one with the majority of C19MC miRNA showing high expression levels (cluster 2). Interestingly, one of the four normal kidney samples was grouped in cluster 2, indicating that the miRNA C19MC LOI event may occur in apparently normal kidney in WT, similarly to *IGF2* LOI [192].

In addition to clustering samples based on expression levels across the C19MC miRNAs, I also compared the expression levels of individual oncogenic miRNAs
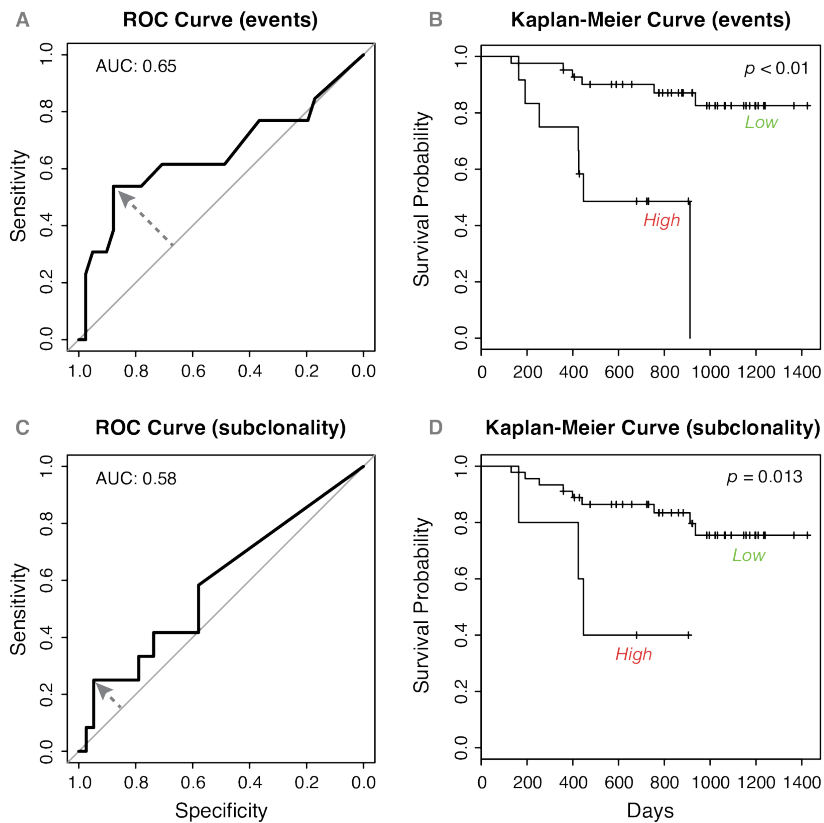
Figure 5.21: Genomic instability, measured as total number of CNA events in tumour evolution, and percentage of CNA which are subclonal, appear to be prognostic for event-free survival. To investigate this both the number of events per tumour and fraction of subclonal events were used to perform ROC analysis (A, C), here the number of events has a greater area under the curve (AUC) than fraction subclonal (0.65 v 0.58) indicating that events is more predictive of event-free survival than subclonality. Based on the the ROC analysis the threshold that produced the value furthest from the line representing a random predictor ($x = 1 - y$) in ROC space was taken (as shown by the grey dashed arrows). This value was 11 events, for number of events, and 81.4% CNAs subclonal, for subclonality. Performing a log-rank test on the Kaplan-Meier curves for each predictor (B, D) showed both the number of events ($p = 1.03 \times 10^{-4}$) and the fraction of subclonal CNAs ($p = 0.013$) to be predictive of event-free survival.

(onco-miRs) which form part of the C19MC cluster (miR-517a, miR-517c, miR-520c-3p, miR-520g) [190]. For each of these onco-miRs the expression level was higher in cluster 2 in the subset of epithelial and anaplastic tumours (Fig. 5.20).

5.8   Copy number evolution and patient follow-up

As we collected follow up data for the PKC66 cases I was able to test if general features of the evolution of these cancers were prognostic. We obtained follow up data for 54 of the PKC66 cases. The median follow up time for cases was ~2.48 years (0.45 – 3.9 years) for patients that did not have observed death/relapse (n = 41). I assessed both total number of CNA events across the series as well as percentage of CNAs subclonal, as possible predictors of relapse/death. For both predictors I generated a ROC curve to test for the relationship between specificity and sensitivity of these predictors for predicting event-free survival. Additionally, the ROC curve allowed me to choose a threshold for each predictor that seeks to maximise sensitivity and specificity.

Using the number of events for predicting relapse/death showed predictive value (AUC = 0.65, Fig. 5.21A). The threshold that produced a sensitivity/specificity with the greatest distance from what is expected randomly was 11 events. Using this value to differentiate between patients in a category of high/low number of events significantly predicted event-free survival ($p = 1.03 \times 10^{-4}$, log-rank test, Fig. 5.21B). This indicates that genomic instability is predictive of event-free survival in PKCs. Performing the same tests on percentage of subclonal CNAs required removing tumours which could possibly be multiple tumours in a single mass as subclonality is difficult to assess in these cases.

Performing a ROC curve on CNA subclonality showed this predictor to be less predictive (AUC = 0.58, Fig. 5.21C). A threshold of 81.4% CNAs subclonal was determined to provide the best balance between sensitivity and specificity. Using this threshold as a classifier predicted event-free survival ($p = 0.013$, log-rank test, Fig. 5.21D). This indicated that a high percentage of subclonal CNA events is predictive of event-free survival. However, only five cases had a subclonality greater than 81.4% and all of these cases also had a number of events greater than 11 (18–67). The threshold chosen by ROC curve analysis selects for a highly specific
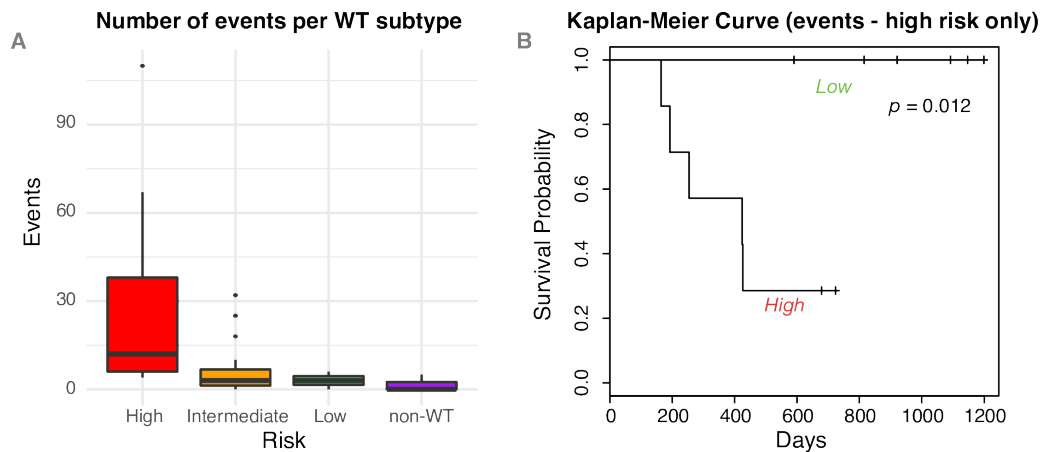
Figure 5.22: The number of CNA events per histological WT risk type is variable. Low risk WTs and other PKCs have very few CNA events over the course of their evolutionary history (low risk median number of events = 3, n = 2). In stark contrast, high risk WTs show many events, with a median number of events of 12 and a range from 4–110 (n = 13). Intermediate risk WTs generally have much fewer events than high risk WTs (median = 3, n = 42), however there are outlier cases that have more than 15 CNA events (A). When using the threshold for determining a high/low number of events (11) from Fig. 5.21 and performing a log-rank test on the Kaplan-Meier curves produced for the high risk cases considered to have either a high or low number of CNA events, the survival differences are considered to be significant ($p = 0.012$, n = 13).

threshold (0.95) but not a sensitive threshold (0.25). Taking a threshold of 4% CNAs subclonal for example (sensitivity = 0.58, specificity = 0.58) is not predictive of event-free survival ($p = 0.26$, log-rank test).

For bilateral cases the higher value for the individual tumours for both predictors was taken for analysis, as I felt this would better represent the level of risk in the individual case.

5.8.1 Number of CNA events predicts event-free survival in the high-risk subtype

To explore further the role of the number of CNAs events in WTs and its relationship to prognosis, I assessed the numbers of CNA events across different risk subtypes (in cases with follow-up data whilst treating the bilateral tumours as unilateral tumours). Generally, I observed that high risk tumours have a larger number of CNA events (median number of events = 12, range 4–110). The median numbers

Number of Samples	Cases
2	5
3	0
4	9
5	8
6	8

Table 5.4: A summary of the number of tumour samples per case in the 30 PKC66 WT cases sequenced.

of CNA events for intermediate and low risk were both 3 events and the median number of CNA events for non-WT PKCs was 0. The range of events however was greater for intermediate type WTs (0–32) and five tumours, shown as outliers in Fig. 5.22A, have more than 15 CNA events.

I then chose to test if number of CNA events could be prognostic within high risk WTs themselves. Taking the threshold for high/low risk patients from the dataset as a whole (number of CNA events = 11), which is similar to the median number of events in high risk cases (12), I performed a log-rank test on the Kaplan-Meier curves of the high risk cases, separating cases as high/low instability based on if they possessed >11 CNA events. Number of evolutionary CNA events in these cases appeared to be prognostic of relapse/death ($p = 0.012$, log-rank test, Fig. 5.22B). This suggests that the number of CNA events can be used to distinguish between high-risk cases that are likely to relapse or not, adding further power to risk stratification (n = 13).

5.9 Integration of sequence mutation evolution in Wilms' tumour

We used our WT-specific targeted sequencing panel to sequence tumour pieces taken from 30 WT cases in the PKC66 series. The mean extent of sampling was ~4.47 (range 2–6), at least one normal sample was also sequenced for each case to allow for somatic variant calling.

Number of SNVs	Cases
0	8
1	11
2	5
3	3
4	2
5	0
6	0
7	1

Table 5.5: A summary of the number of SNVs identified across the 30 PKC66 WT cases sequenced with the WT-specific bait capture design.

I identified a mean of ~1.5 somatic point mutations per case across all sequenced samples (0–7 mutations). The number of point mutations identified did not correlate with number of samples sequenced (Spearman's rho = -0.007). Five genes showed point mutations in multiple cases. *CTNNB1* was mutated in five stromal cases, a mixed type case and blastemal case (n = 7). *TP53* was mutated in six cases, all of which were diffuse anaplastic cases (n = 6). *DROSHA*, *AMER1* and *RYR2* were mutated in two cases each.

Across the series 41 unique SNVs were identified. Missense mutations made up 25 of the point mutations, four were nonsense mutations, four silent, six were intronic, one affected a splice site and one was in an intergenic region.

Three of the six *TP53* point mutations appeared clonal (8, 59, 170), and only two of the seven *CTNNB1* mutated cases had apparent clonal *CTNNB1* mutation (146, 171). Non-synonymous point mutations also appeared clonal in *DROSHA* (54 and 78), *WT1* (64) and *TRIM37* (80).

Additionally, I analysed eight genes for insertions and deletions (indels). These genes were *AMER1*, *CTNNB1*, *DGCR8*, *DICER1*, *DROSHA*, *MLLT1*, *TP53* and *WT1*. Across the 30 cases, 7 cases displayed an indel in one of these genes. No case possessed an indel in more than one of these genes. These seven cases contained five unique mutations, surprisingly three of these mutations were in frame

deletions whereas one was a frame shift deletion and one was a frame shift insertion.

Three cases possessed a *CTNNB1* deletion (7, 64, 78) and strikingly each time this deletion was identical – S45del. This in-frame deletion perfectly deletes the 45th codon of *CTNNB1* which encodes for serine (CTT). This mutation was clonal in the right tumour of IMPORT 7 (RR1 – RR4), but was only present in a single sample in IMPORT 64 (R1) and IMPORT 78 (LR2).

Two cases contained a *WT1* indel (146, 143), one insertion (146) and one deletion (143) that both cause frameshifts (L302fs and Q26fs, respectively). Both *WT1* indels appeared clonal in the samples sequenced. Deletions were discovered in *TP53* (54) and *DICER1* (4) also, both caused in frame deletions (E339del and I1744del, respectively) and were only discovered in a single sample in each case (R3 and LR1, respectively).

Mutation cancer cell fraction was calculated based on allele frequency, tumour purity and local copy number state and was used to cluster alongside copy number variations to identify subclones and infer phylogeny alongside copy number data in two cases, IMPORT 54 and 143.

5.10 Diversification of Wilms' tumour evolution

Performing WT-specific targeted sequencing revealed further complexity to WT evolution and allowed us to further explain relationships between tumour samples. Diversification was most striking in cases with heterogeneous *TP53* mutation. Four cases appeared to display heterogeneous *TP53* and all cases were diagnosed as diffuse anaplastic (10, 54, 106, 233).

The phylogeny inferred from the CNAs in IMPORT 54 suggest that there may be two tumours present as tumour section R3 shares no CNAs with R1 and R2
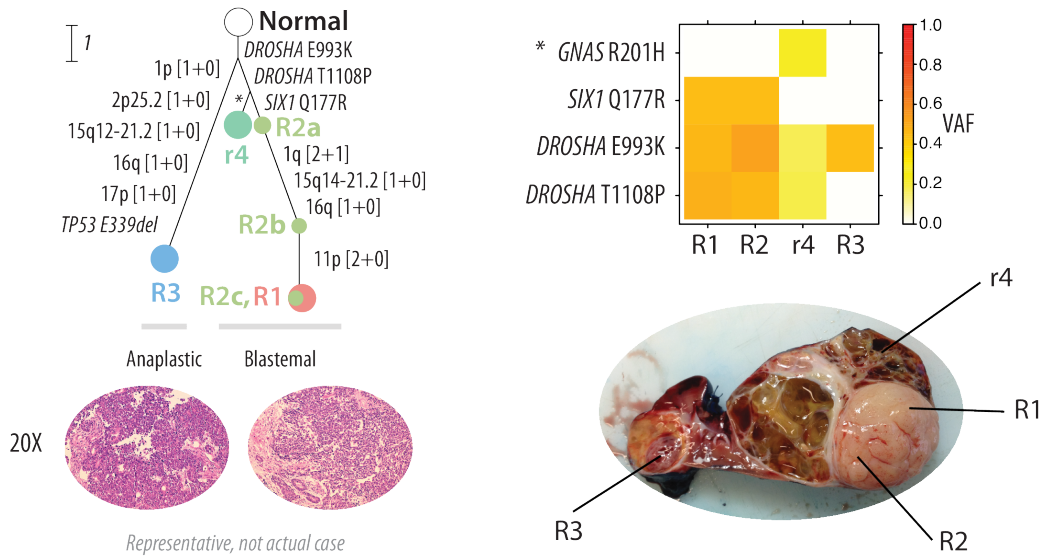
Figure 5.23: Strong diversification in a WT case (54). Tumour regions R1–2, r4 appear to follow a separate evolutionary trajectory (top left) to R3 following the acquisition of a clonal *DROSHA* E993K mutation. R1–2, r4 acquire an additional *DROSHA* mutation (T1108P) and R1, R2 further diversify acquiring a *SIX1* Q177R mutation and CNAs. R1 and R2 showed blastemal histology. R3 developed unique CNAs (including 17p loss) and a *TP53* E339del mutation. R3 showed diffuse anaplasia and led to the case being diagnosed as such. The heatmap (top right) shows the VAF of each point mutation across the regions supporting this phylogeny. The photograph of the tumour (bottom right) clearly shows that the region are located in distinct sites. The most evolved clones (R1 and R2c) form part of what appears to be an outgrowth, perhaps due to increased the proliferation of these clones. R3 is spatially separate from R1–2, r4.

(although all regions have losses affecting 16q and 15q14–21.2, the losses in R3 have different boundaries to R1 and R2, as described in Section 5.6). However, sequencing reveals that all regions possess a *DROSHA* E993K mutation, indicating that these regions are indeed related and that they are not completely independent tumours. Therefore *DROSHA* E993K mutation is the only detected clonal mutation in this case. R3 uniquely contains a *TP53* small deletion (E339del). Regions R1 and R2 share two mutations uniquely between them. A second *DROSHA* mutation (T1108P) and a *SIX1* Q177R mutation. *SIX1* Q177R mutations have previously been reported in blastemal type WTs [40]. Phenotypically R1 and R2 were dominated by blastema whereas R3 showed diffuse anaplasia (leading to the diagnosis). The tumour therefore shows the phenotype of two high risk WT subtypes in a single

158

mass. This phenotypic disparity is clearly a result of large mutational divergence following the *DROSHA* E993K mutation – which may have been a field effect causing mutation.

A fourth region (r4) was also sequenced in IMPORT 54 but unfortunately did not produce a SNP array with adequate quality for analysis. This region only shared the clonal *DROSHA* E993K mutation with all regions and the *DROSHA* T1108P mutation with the R1 and R2 but uniquely possessed a *GNAS* R201H mutation, a known driver mutation [193]. Manual assessment of the SNP array and assessment of the CNA probes included as part of the target sequencing design suggest the region contains no CNAs. This would place *DROSHA* T1108P mutation earlier than all CNAs in R1 and R2 and would indicate an ancestral subclone to R1 and R2 uniquely possessing *GNAS* R201H is present in the tumour. This case represents the large potential for diversity in WT with three of the four regions showing drastically different yet related genotypes.

IMPORT 233 also contains a heterogeneous *TP53* mutation. This tumour shows a strikingly more linear divergence (Fig. 5.9). In the two tumour samples assayed, R2 uniquely contains a *TP53* H179L mutation and possesses several additional CNA events whilst sharing all but one of the CNA events with R1, therefore displaying a linear fashion of evolution. It appears that the *TP53* mutation may cause the expansion of a genomically unstable clone in R2. R1 uniquely contains a +8 and a *MLLT1* N115K mutation. Indicating that R1 may have further evolved to develop a unique phenotype.

IMPORT 10 displayed heterogeneous *TP53* mutation in the two samples which were sequenced (R2 and R3). Similarly to IMPORT 233, this mutation appears to be associated with a linear progression in CNAs (Fig. 5.8). R2 only contained a 19q UPD event and did not possess a *TP53* mutation, however, R3, which contained several additional CNAs, possessed a E285K mutation in *TP53*.

Strikingly, in IMPORT 106, there were two unique point mutations in *TP53* identified in separate samples (R2 – R342*, R4 – G245S) but not in the other two samples sequenced (R1 and R3). This case showed extensive branching in the CNA following a 19q UPD event and the development of unique *TP53* events may explain this (Fig. 5.9). Surprisingly, R1, which showed the largest number of CNA events was not called as having a point mutation or an indel in *TP53* according to my analysis, however, an insertion in *TP53* was called in this sample by muTect2 but removed due to my custom filter based on low coverage as reported by muTect2 ($<$10 reads reporting mutation – five reads reported a P82fs insertion with 1 read reporting the reference). If this insertion is real it is likely the cause of such chromosomal chaos. This would mean *TP53* mutation was converged on three times in this case.

Striking pheno-/genotypic diversification was also observed in IMPORT 147, here one sample, RR4, exclusively contains a N387K *CTNNB1* mutation and the tissue piece is composed of 95% stroma. The other tumour pieces RR1–RR3 share a R358* *AMER1* mutation, present in only 2% of reads in RR4. Tumour pieces RR1–RR3 possess a mixed-type histology in which no one cell-type dominates, ergo the tumour is diagnosed as mixed-type. These tumours are related by shared whole chromosomal gains of 8, 12 and 18, yet show divergence both genetically and phenotypically (Fig. 5.12). Interestingly RR4 also contains a +3, suggesting these cells are evolving down a pathway common to stromal tumours in this series.

5.11 Heterogeneity and convergent evolution of *CTNNB1* mutation

In addition to observing convergent evolution of CNA in several cases and of *TP53* in IMPORT 106, I also observed convergent evolution of sequence mutation in *CTNNB1* in three stromal tumours (47, 78, 143). Five regions of IMPORT 143 were sequenced (R1–R5). The stromal WT contains a clonal 11p UPD event, in line with almost all stromal tumours in the PKC66 series. *CTNNB1* mutation was
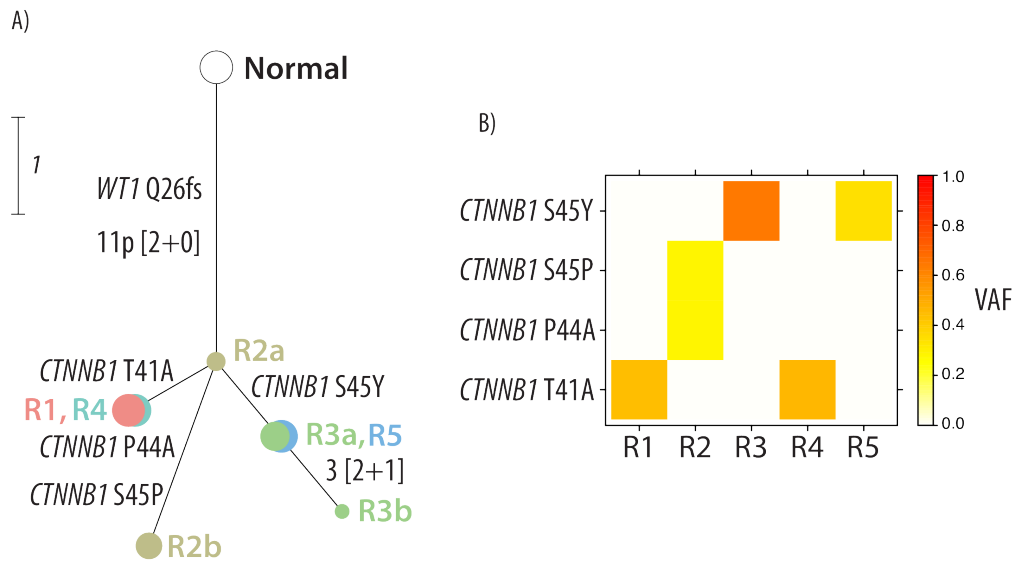
Figure 5.24: IMPORT 143 shows convergent evolution of *CTNNB1* mutations. Here following a truncating *WT1* frameshift mutation and a 11p UPD event, four *CTNNB1* have been acquired, as observed in five separate tumour pieces (A). The heatmap (B) shows the recorded VAF of the *CTNNB1* mutations.

observed in all regions of this tumour, however strikingly two regions (R1 and R4) possess a T41A mutation exclusively, two regions (R3 and R5) contain a S45Y mutation, also exclusively, and R2 contained two additional mutations affecting the same allele (P44A and S45P – reported on the same read). This indicates that following the 11p LOH event there is a very strong selection pressure for a *CTNNB1* mutation and as a result, three mutated alleles emerge in parallel. Three of four of these mutations are known to be functional as they mutate phosphorylatable sites in *CTNNB1* which allow it to appropriately respond to degradation signals. The mutation of a proline to an alanine in R2 is also likely to cause a structural alteration. Interestingly, the VAF (0.65) for the S45Y mutation in R3 does not fit the copy number profile of the sample closely (there is a chromosome 3 gain [2+1] but it is subclonal and in ~20% of cells). This VAF is could potentially be explained by this mutation occurring twice in this sample.

Curiously all incidences of a P44A *CTNNB1* reported in a WT in COSMIC (can-

161

cer.sanger.ac.uk/cosmic) are reported in samples that also contain a S45P mutation [194, 195, 196, 197]. This possibly indicates a common mechanism of mutation that causes two point mutations in a distance of two codons or a high selection for this combination of mutations when the same allele is affected.

IMPORT 47 also exhibited convergent evolution of *CTNNB1*. The only tumour sample taken from the left kidney in this bilateral case showed no *CTNNB1* mutation (LR1). Three of the four samples taken from the right kidney contained a K335I mutation (RR2–RR4), whereas RR1 contained a S45P mutation (Fig. 5.11).

IMPORT 78 displayed convergent evolution of *CTNNB1* also as LR1, LR5 and LR6 all contained a S45Y mutation and LR2 contains a S45del in frame deletion. All regions sequenced in this case (LR1–LR3, LR5, LR6) contain a clonal E1147K *DROSHA* mutation and also all contain clonal gains of chromosome 12 and 18 (Fig. 5.12).

Tumour region R3 in IMPORT 192 uniquely possessed a S45C mutation in *CTNNB1*. IMPORT 192 is a blastemal WT case.

5.12 Possible evidence for mosaicism in Wilms' tumour

Deep targeted sequencing of Wilms' tumour normal tissue samples can reveal mutations present in a low number of cells and therefore could identify possible mosaicisms in the individual that predispose to a tumour. Two cases reported several reads, that were mutant in the tumour, in their respective normal samples that became clonal in the tumour evolution. In the normal kidney sample taken from IMPORT 54 there are 7/571 reads reporting the clonal E993K mutation in *DROSHA*, yet there is not a single read reporting the other non-clonal point mutations discovered in *DROSHA*, *SIX1* and *GNAS* in the normal kidney sample. Potentially 2.45% (1.19% – 5.017%, binomial confidence interval, Wilson method) of the cells in this

normal kidney sample may already carry the E993K *DROSHA* mutation. This may explain the striking diversification seen after this mutation in this tumour as these tumours may have developed from two separate cells in a mosaic kidney.

Additionally, 6/709 reads in the normal sample of IMPORT 64 report the clonal E128* mutation in *WT1*. Here potentially 1.69% (0.78% – 3.67%, binomial confidence interval, Wilson method) of cells may carry this mutation in the normal sample and these cells may have given rise to the tumour. Interestingly, the normal sample in this case from derived from the blood, indicating that the mosaicism may be across the whole individual.

5.13 Comparing *TP53* heterogeneity status to circulating tumour DNA

To assess the possibility of assaying ctDNA in WT, we conducted a study in which we selected four diffuse anaplastic cases in the PKC66 series which we had sequenced (8, 59, 170, 233) and knew to contain a *TP53* mutation, a mutation associated with this high risk subtype. The mutations in *TP53* were R248Q (8), R337C (59), R273C (170) and H179L (233).

A ddPCR assay was designed to measure patient-specific mutant *TP53* from circulating serum, plasma and urine samples across several time-points, including at diagnosis, at the two week point of the four week chemotherapy programme prior to surgery, prior to removing the tumour (pre-operation) and after the operation. To assess ctDNA specifically, a DNA extraction protocol to specifically select for ctDNA was used when extracting DNA from these liquid samples.

The clearest results and most abundant observations of mutant DNA were seen in the plasma and serum derived from IMPORT 59. This case was sampled at diagnosis (Dx), mid-chemotherapy and post-operation. At diagnosis the mutant allele

163
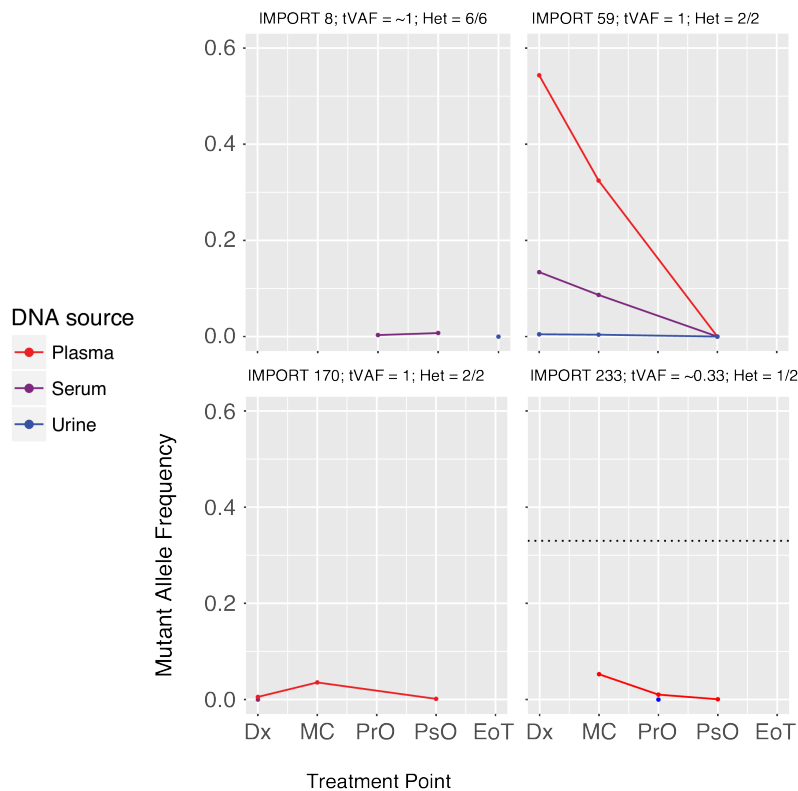
Figure 5.25: Mutant *TP53* can be detected in ctDNA but not in utDNA in diffuse anaplastic WT cases. Mutant *TP53* was detected in blood plasma (red) in all three cases in which is was measured (59, 170, 233) displaying the highest MAFs prior to nephrectomy, at diagnosis (Dx) in IMPORT 59 and mid-chemotherapy (MC) in IMPORTs 170 and 233. Mutant *TP53* was also detected in serum (purple) samples taken from IMPORT 59. Urine (blue) samples appeared to show no detectable mutant *TP53*. In the title for each panel the VAF of the bulk tumour is estimated (tVAF) as calculated based on all tumour samples sequenced and arrayed per case for comparison and the number of tumour samples containing mutant *TP53* as a ratio of the total numbers of samples sequenced is noted (Het). IMPORT 233 (bottom right) was the only case that showed heterogeneous *TP53* and the dotted horizontal line displays the tVAF for this case for comparison. Additional timepoints compared to those already mentioned include pre-operation (PrO), post-operation (PsO) and end of treatment (EoT).

frequency (MAF) of *TP53* R337C in the plasma was 0.54, meaning over half of the alleles detected in the ctDNA of this patient at this time point were mutant *TP53*, the corresponding value was 0.13 in the serum. During chemotherapy these values declined to 0.32 in the plasma and 0.09 in serum and post-operation *TP53* levels were negligible, reflecting the removal of the source of the mutant DNA and the short half-life of ctDNA in this example. The VAF in the bulk tumour sample was 1.

IMPORT 170 and 233 show much lower MAF in plasma samples compared to IMPORT 59. For both cases the timepoint at which the MAF was highest was mid-chemotherapy, possibly due to increased cell death and the release of tumour DNA into the bloodstream. However, the MAF values were significantly lower than IMPORT 59 at 0.04 and 0.05 for IMPORT 170 and 233 respectively. Similarly to IMPORT 59 the post-operation timepoint had negligible MAF in the plasma in 170 and 233, probably due to tumour removal. Interestingly, IMPORT 233 had a heterogeneous *TP53* mutation as detected in the two samples sequenced, if all circulating DNA was produced by these assayed primary tumour cells the expected MAF would only be 0.33 (2 wild type alleles in R1, 1 mutant allele in R2). Ergo, relatively speaking, tumour DNA is contributing more to the circulating DNA in IMPORT 233 than IMPORT 170 by virtue of the fact the mutation is heterogeneous in IMPORT 233, despite the MAF in both cases being similar.

IMPORT 8 showed little signal and actually displayed a small increase post-operation compared to pre-operation, potentially the tumour may not have been entirely removed. Overall, urine showed little/no signal.

5.14 Concluding Remarks

In this chapter I have presented a wide ranging assessment of paediatric kidney cancer evolution. By exploring the genomes of multiple tumour pieces of these tissues using two major genome-wide assays in cancer, genome-wide SNP array genotyping and NGS, I integrated the approaches to gain a holistic assessment of PKC clonal evolution. These findings are fundamentally important for under-standing how PKC and indeed other solid paediatric tumour evolve, how events in their evolution may lead to characteristic phenotypes and how understanding these cancers in the context of their evolution is essential for understanding and further exploring the importance of particular genomic mutations.

Furthermore, these discoveries are clearly significant for future clinical practice. By assaying the genomic make-up of multiple tumour pieces it becomes apparent that PKCs and in particular WTs have great potential for evolutionary diversification and further still seemingly show great potential for developing multiple possibly independent tumours. We have also assessed the detection of mutant *TP53* in ctDNA over time and compared detection to the expected VAF based on the characterisation of the tumour as a heterogeneous mixture of clones – assessing the use of a minimally invasive assay for the detection of an important mutation in the context of a heterogeneous tumour.

By attempting to take an unbiased selection of PKC patients we sampled a large set of patients and then subdivided into different histological groups and assessed the similarities and differences between and within different histological groups. Furthermore, I assessed if patterns within these subgroups may be indicative of different journeys down the same path of selection.

The most striking example of a consistent evolutionary pattern was the observation of a progression model in the stromal subtype of WT. All cases bar one unilateral stromal tumour (6/7) displayed a clonal 11p LOH event. All stromal tumours in the bilateral cases also showed 11p LOH, although in two cases it was not clonal, it emerged multiple times through convergent evolution. Many of these 11p LOH events were then followed by 3p changes, which may be possibly due to the acquisition of a *CTNNB1* mutation, as supported by sequencing evidence. This indicates that the progression of a stromal WT tumour begins with a 11p15 LOH event followed by a *CTNNB1* and chromosome 3p CNA, highly likely to be mutant allele specific. This path of mutation acquisition is elegantly displayed in the 13 stromal tumours in the PKC66 series.

Another model of tumour evolution appears to connect epithelial type WTs and high risk diffuse anaplastic WTs as both these subtypes appear to present with

clonal chromosome 19q UPD in a subset of cases, drawing an evolutionary link between two WT subtypes. The importance of 19q UPD in epithelial tumours is dramatically supported by the convergence onto this event in IMPORT 30 in which three potentially separate epithelial tumours have developed following the acquisition of chromosome 19q UPD, indicating how understanding repeated evolution enables us to understand the importance of this change as an initial CNA in epithelial WTs. When chromosome 19q UPD was acquired in epithelial tumours it was always the only event and of the diffuse anaplastic tumours in which 19q UPD was clonal, there were samples in which it is the only clonal change observed, suggesting it occurred first. It is possible that 19q UPD is a priming event for diffuse anaplasia and that the epithelial tumours, in which 19q UPD is the only event, may have been removed and diagnosed prior to developing diffuse anaplasia. In the diffuse anaplastic cases with 19q UPD, a clone may have developed a second hit to develop an anaplasic phenotype and lose genomic stability, this second hit may be a *TP53* mutation and indeed IMPORT 10 and 106 have chromosome 17p LOH events in the samples with the highest mutational burden. Furthermore, two of these three diffuse anaplastic cases would have been diagnosed as epithelial subtype if they did not show diffuse anaplasia (over 66% epithelial cells in 83 and 106), further supporting the idea they may have initially been genomically stable epithelial tumours. Not only do these results link two WT subtypes but highlight a potential progression model from one subtype to another. These observations however remain limited by the number of cases in which they were observed (4 epithelial cases, 3 diffuse anaplastic cases). This link may only be relevant for half of epithelial-type tumours (4/8 cases) and a third of diffuse anaplastic tumours (3/9 cases) but is only identifiable by conducting a study on a large cohort of unselected patients. Future studies may seek to confirm this link by specifically focusing on this quesiton by generating larger cohorts of epithelial and diffuse anaplastic type tumours.

This chapter demonstrates that evolutionary diversification continues to be a com-

mon theme in multisampled WT studies. Branched evolution is demonstrated most clearly in mixed, blastemal and diffuse anaplastic type histologies and can be visualised by comparing these evolutionary patterns to the images mapping the origin of these tumour samples. Although caution should be taken when assuming the directionality of clonal movement in these images as they are merely a snapshot of tumour evolution. Diversification can develop early in WT evolution as demonstrated in IMPORT 54 in Section 5.10. This case supports the observation of a single mass presenting with tissue samples that do not have CNAs with equal breakpoints as is presented in Section 3.8, but exemplifies that highly divergent clones may still share ancestral point mutations and highlights that NGS is vital for determining true genetic independence. This chapter also presents an additional 4 cases of bilateral WT in which both tumours were sampled and supports the observation in Chapter 3 that bilateral tumours are genetically independent.

WTs are highly diverse in their clinical presentation and cellular composition and are also diverse in their phylogenetic patterning. The nature of genetic diversity in WT risks undermining the success of single biomarker assays, however a study with more consistent sampling strategy would be more appropriate for elucidating this risk. Overall, the variegated nature of the PKC66 series and WTs generally makes summarising findings for translating into clinical practice challenging. However, it is clear that the ability to acquire mutational burden across the tumour's evolutionary history is associated with poor prognosis. An assay that assesses multiregion CNA events could potentially be clinically useful as I have shown that the number of these events can identify patients that are likely to relapse. This test is capable of identifying high-risk patients irrespective of underlying histology. This is somewhat unsurprising as a high burden of CNAs is typically displayed by high-risk diffuse anaplastic cases but I also observe high CNA burden a subset of the high-risk blastemal cases. Perhaps more strikingly, this test was able to distinguish between patients that relapse earlier within high-risk patients alone. Further studies may seek to consistently multisample WTs and assay multiregion CNA burden

to validate this finding. Myself and colleagues have also shown that one potential biomarker of high CNA burden, mutant *TP53*, is detectable longitudinally in patient blood and that DNA released by the tumour can contribute to the majority of the circulating alleles in a patient prior to nephrectomy (e.g. IMPORT 59). A ctDNA assay could be utilised to pre-screen patients prior to surgery to assess the risk of the patient presenting with a genomically unstable tumour. Despite the association of mutant *TP53* with genomic instability, IMPORT 54 was not categorised as having high CNA burden despite possessing mutant *TP53*. However, this case is an example of how early divergence can obscure the detection of multiregion CNA burden as the area of the tumour displaying diffuse anaplasia was essentially only single sampled.

Overall, this chapter attempts to convey the dynamism and diversity of sixty-six PKCs and exhibit the selective pressures that they are under, whilst constantly interacting with important clinical questions.

Chapter 6

Copy Number Evolution in Hepatoblastoma

6.1 Introduction

In addition to studying WTs as a model of solid paediatric tumour evolution, I also analysed a dataset consisting of a small blue round cell tumour of the liver, hepatoblastoma. This tumour is rarer than WT, with an incidence of 1.2–1.5 cases/million population/year [31]. We were only able to gather tumour samples from eleven patients. Multiple tumour pieces were taken from a resection sample where possible (7/11 cases, 63.6%). The range of tumour sampling was 1–4. Tumour samples were also derived in the form of pre-chemotherapy biopsies in 4/11 cases. Constitutional DNA as peripheral blood lymphocytes was only available in 7/11 cases. Heterozygous SNPs were determined manually in those cases without normal samples. In Patient 1, a patient with confirmed Beckwith-Wiedemann syndrome, the non-tumorous liver was also sampled multiple times (L1–L4).

The primary aim of this chapter is to describe for the first time the process of copy number evolution in this rare solid paediatric tumour, despite a relatively small number of cases and samples. I aimed to determine the clonal composition of the tumours in each patient across tumour samples and within each sample in order to generate a detailed picture of chromosomal evolution in these cases as possible. To

Number of samples	Cases (resection-only)	Cases (inc. biopsy)
0	2	0
1	2	4
2	3	2
3	2	2
4	2	3

Table 6.1: Number of samples taken in the study per patient taking into account numbers of samples derived from primary tumour resection and with the addition of pre-chemotherapy biopsies.

achieve this, I used analytical methods outlined in Chapter 4. Each sample in this series was analysed on a Illumina® HumanCoreExome array to generate LRR and BAF values across the ~500,000 SNPs on the array. Once the cancer cell fraction contribution of each copy number alteration was determined and clones inferred, phylogeny was determined using MEDICC and inspected manually for a recreation of the orderings supported by inferred subclonal composition.

This chapter was carried out as a collaboration with the laboratory of Professor John Anderson who collected the samples and generated the SNP array data. Further details are described in Chapter 2.

6.2 Evidence for two independent tumours in a hepatoblastoma case

One case (Patient 3) in the eleven patients analysed shows evidence for two independent tumours in a single hepatoblastoma (Fig. 6.1). Here, R1–R2 contain no CNAs with identical breakpoint boundaries to the CNAs observed in R3, indicating that this patient potentially presented with a mass containing two independent tumours. R1–R2a contain losses of chromosome 1p, 11 and 18q and gains of 1q and 20q. R3 contains many focal single and double losses in the q-arm of chromosome 4 (4q34.1–q35.2), a focal UPD event affecting the *CDKN1C* locus (chr11:2749015–2929552, hg19), a gain of 2q22.3–q33.3 and a gain of the majority of chromosome 1q (1q21.3–q44). However, this gain in 1q does not share the same break points as the gain in R1–R2, indicating it is potentially unrelated to the gain of the whole arm
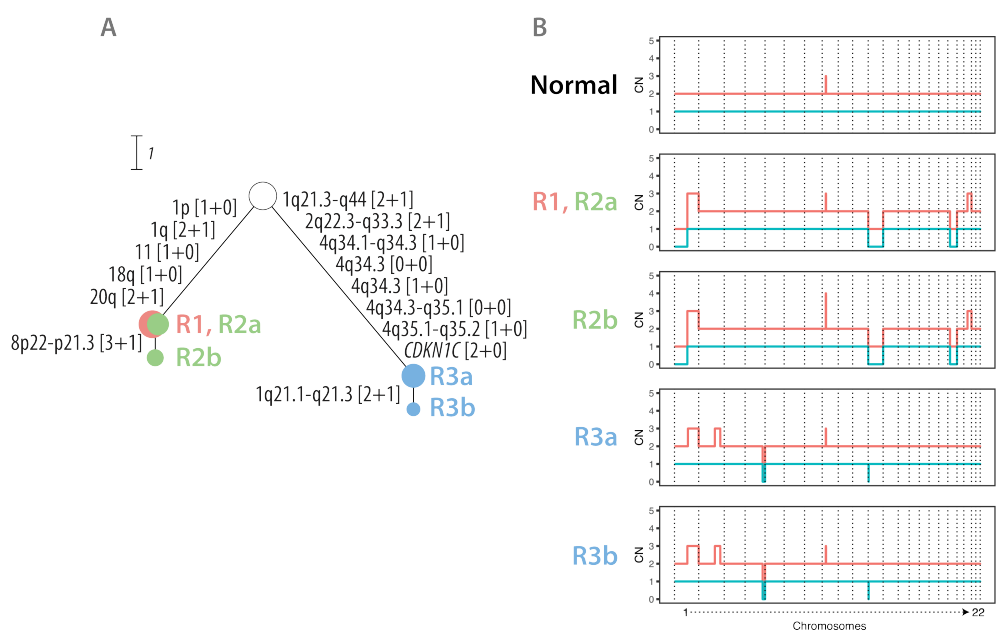
Figure 6.1: Evidence for the development of two independent hepatoblastomas in Patient 3. An interpretation of the phylogeny for this patient is displayed of the left of this figure that shows that no copy number alterations in R1–R2 and R3 share CNAs with equal breakpoint boundaries. The right of the figure show the copy number profiles inferred for all clones across the regions. Both the total copy number (red) and the minor allele copy number (blue) are displayed.

in R1–R2. A gain of a region of chromosome 1q (1q21.1–1q21.3) is also present in 25% of tumour cells in R3.

It is possible that R1–R2 and R3 are related to one another via a whole gain of chromosome 1q and then a loss back to a normal 1+1 state in 75% of cells of 1q21.1–1q21.3 in R3. This possibility is actually more parsimonious. However, chromosome 1q gain was the commonest CNA in this series, indicating its propensity to occur in hepatoblastomas. Furthermore, as presented in section 6.3, there is no other evidence for such branched evolution in this series. Overall, it is not possible to prove definitively that these samples are completely independent.

The constitutional DNA sample taken from this patient contains a gain of 8p22–p21.3 (chr8:18574874–19379421, hg19) which is duplicated to a 3+1 state in ~20%

of cells in R2. Interestingly, the mean BAF for this gain in the constitutional DNA is 0.65 which would indicate the gain is in 85.7% of cells and not 100%. The gain is reported by 71 SNPs on the array and by bootstrapping the BAF data of this gain through random replacement, calculating the mean BAF through 1,000,000 iterations, the 2.5th and 97.5th percentiles suggest the cancer cell fraction of the gain may be between 80.5% and 90.9% of cells in the sample.

This suggests that the gain of 8p22–p21.3 is potentially somatic and occurs as a mosaicism in the individual, causing a possible field effect for the formation of multiple hepatoblastomas.

6.3 No evidence for branched evolution in hepato-blastoma

Seven hepatoblastomas in our series were multisampled. Taking the interpretation of Patient 3 as a mass consisting of two independent tumours, all seven tumours showed no evidence of branched evolution and all phylogenetic relationships were of a living ancestor to a more evolved clone (Fig. 6.2).

Linear evolution of these hepatoblastoma cases is further supported by the identification of subclonal copy number changes in Patients 2, 4, 5 and 9 that become clonal in other regions. In Patient 2, ~21% of tumour cells in R1 contain nine CNAs that contribute to the clonal CNA profile of R2, therefore displaying linear evolution but also confirming that these CNAs belong to the same clone. In Patient 4 a chromosome 6p gain was present in 43.1% of tumour cells in R1 and is found to be clonal in R2. R3 and R4 in this case only contain the ancestral 5q21.1–q31.1 loss that is clonal in R1. Patient 5 contained a biopsy sample (B) which contained a clonal 13q34 gain, that was also clonal in R3. Two chromosome gains of 2 and 20 are found to be subclonal (~15% of tumour cells) in the biopsy sample that became clonal in the most evolved tumour clone present in R2, the subclone present in R3 (~60% of tumour cells) also contains a chromosome 17 gain which is found to be
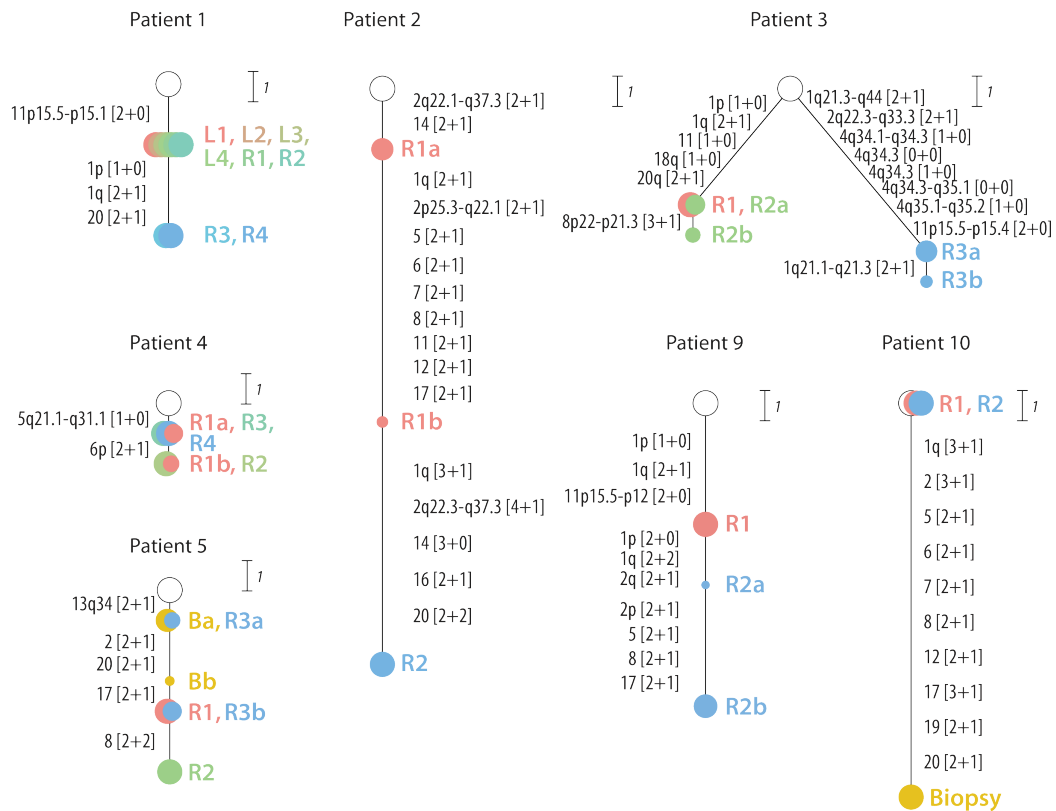
Figure 6.2: There is no evidence for branched evolution in hepatoblastoma. Seven cases in our series were multisampled. Taking the interpretation discussed in section 6.2 that the tumour taken from Patient 3 contains two independent tumours, all multisampled hepatoblastomas showed linear evolution. In Patients 2, 4, 5 and 9, this is further supported by the subclonal deconstruction of copy number changes that are subclonal in one region and become clonal in a further evolved region. Multisampled liver is denoted as L1–4 in Patient 1.

clonal in R2, which incidentally does not show any clonal mixing. Unfortunately there is no normal sample taken from Patient 5 and we are unable to determine if the 13q34 gain is present in normal tissue, however the BAF of this gain do not indicate it is in all cells in any of the tumour samples taken, indicating normal cells without this change must exist. Finally, subclonal gains of 2p, 5, 8 and 17 are present in ~88.7% of tumour cells in R2 of Patient 9, therefore according to the pigeon hole principle (these four CNA gain all have cancer cell fractions ~88.7% – there must be cells which contain all these changes), this subclone supports the continuation of linear evolution in this case.

175

Patient 1 shows an 11p LOH event in six samples, including in all the multiple liver samples (L1–4, R1–2), which then further develops in a linear fashion three additional CNA events in the remaining two samples (R3–R4). Furthermore, it is striking that the primary tumour samples taken at surgical resection in Patient 10 display no CNAs yet the pre-surgery biopsy sample contains several copy number gains. This may potentially owe to the elimination of the clone discovered in the biopsy over a treatment course.

The presence of linear evolution in multisampled tumours meant that orderings implied by subclonal inference were supported by MEDICC, except in Patient 3. MEDICC described branched evolution in Patient 9 due to a difference of two probes in the breakpoint separation between chromosome 1p and 1q in R1 and R2. This was manually reinterpreted to be the same breakpoint boundary upon inspection and linear evolution was taken as the solution.

6.4 Single sampled hepatoblastomas

There are four cases in this hepatoblastoma series that were not multisampled (Fig. 6.3). Two of these cases were only sampled by biopsy (Patient 6 and 11) and in this biopsy in both cases, only a 11p UPD event was described. Chromosome 11p UPD is associated with Beckwith-Wiedemann syndrome and patients with Beckwith-Wiedemann syndrome have an increased incidence of hepatoblastoma [31, 198], in Patient 6 the constitutional DNA also shows this aberration. In Patient 11 the 11p UPD event was reinterpreted as a mixture of 2+0 and 1+1 from a mixture of 2+1 and 2+0 after manual inspection.

Patients 7 and 8 showed several CNA events despite being only sampled once. Patient 7 contained many loss events in a region of chromosome 4q (4q32.2–q35.2), in addition to focal CNAs in 2q24.2–q24.3 and 15q15.1 (a gain and a loss respectively). Patient 8 contained the most events in the series (26 events) including a focal UPD event encompassing the locus of *BMP4* (chr14:54198782– 54862710,

176
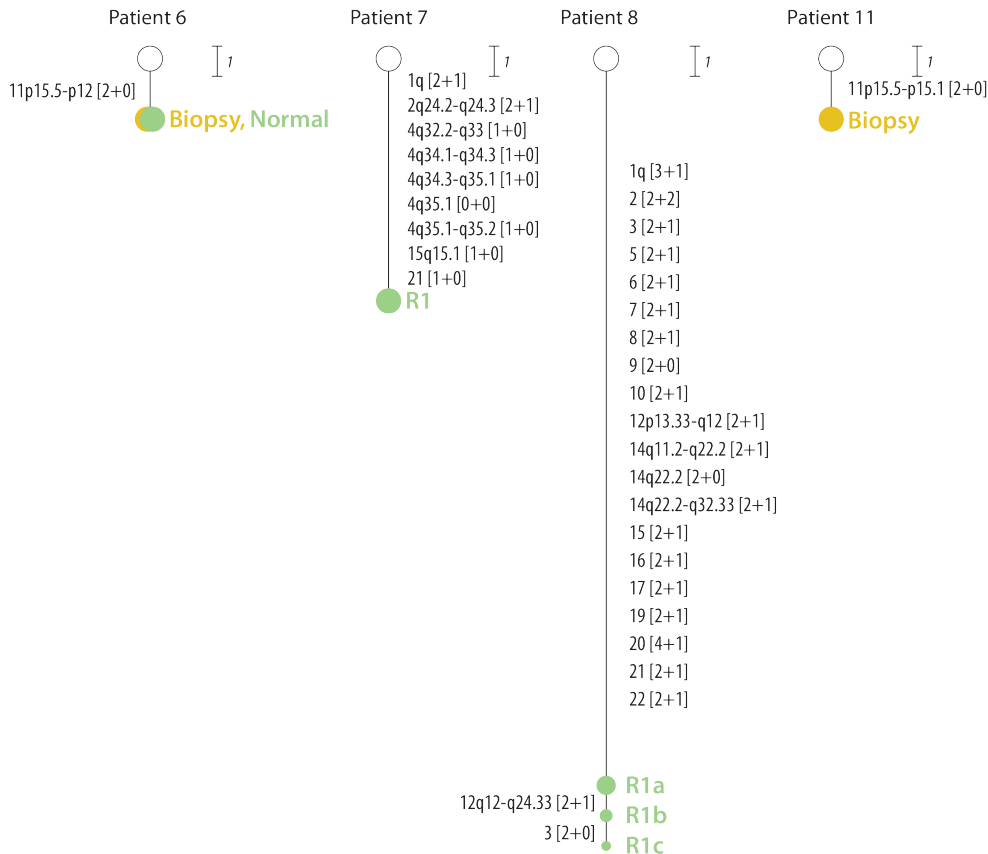
Figure 6.3: Four hepatoblastoma cases in the series were only sampled once and in two cases (6 and 11), this was in the form of a pre-chemotherapy biopsy. Both of these biopsies contained a chromosome 11p UPD event, and the constitutional DNA sample in Patient 6 also contains this aberration. Patients 7 and 8 show several CNAs. Patient 8 also shows clonal mixing as aberrations of 12q and chromosome 3 are determined to be subclonal.

hg19). This case was the only single sampled case which contained subclonal CNAs (chromosome 12q12–q24.33 gain and chromosome 3 UPD).

6.5 Concluding Remarks

Despite this series being limited in terms of both sampling and number, I was able to describe for the first time the nature of copy number evolution in this rare paediatric cancer of the liver. All eleven cases displayed CNAs and only the primary tumour samples taken at surgical resection in Patient 10 showed no CNAs despite the fact the pre-surgery biopsy sample did. Overall, the commonest CNA seen in these tumours was chromosome 1q gain, which was seen in seven cases (1, 2, 3, 7, 8, 9

and 10). Gain of chromosome 20 was also in six cases (1, 2, 3, 5, 8 and 10), UPD of 11p in 5 cases (1, 3, 6, 9 and 11) and chromosome 1p loss in 3 cases (1, 3 and 9). Interestingly, large sets of focal losses were present in the 4q34.1–q35.2 region of chromosome 4 in Patients 3 and 7. These close groupings of focal losses may be explained by catastrophic events such as chromothripsis, however it would only be possible to prove this with additional data such as fluorescence in situ hybridization (FISH) or DNA sequencing.

Similar to a case study presented in Section 3.8, of a WT presenting as a single mass with no shared CNA events with identical breakpoint boundaries between different sets of tumour pieces samples, it appears that Patient 3 in this series of hepatoblastoma may also present as a single mass in which tumour samples R1–R2 and R3 appear to share no CNAs. This would provide evidence that hepatoblastomas, like nephroblastomas, may be capable of forming multiple tumours in a single mass. The origin of a heightened propensity to develop multiple tumours in a single mass may have origins in somatic CNAs developing early in human development and being present as a mosaicism in individual. In Patient 3 this is supported by the fact that the BAF of a copy number gain of 8p22–p21.3 does not support the gain being present in all cells in the individual. This finding sheds light on the ability of hepatoblastomas to develop in a single patient and may force treatment strategies to take into account phenotypic heterogeneity potentially produced by multiple tumours. It is also important to note that the presence of a 1q+ in all samples in Patient 3 – despite each gain possessing separate breakpoint boundaries – makes it difficult to exclude the possibility of a MRCA existing that possessed only a 1q+.

One of the most distinctive features of this study appears to be the lack of branched evolution seen across this series of eleven hepatoblastomas. Despite the limited size of the series and number of tumour samples taken per case, it is still striking that no branched evolution is clearly present. This may indicate that the evolution and

178

development of hepatoblastomas may occur across a linear series of clonal sweeps. This may present an opening for understanding the genetic changes that lead to the development of hepatoblastoma by determining the linear events that produce these tumours through prediction and allow for further studies to identify the driver(s) of a linear trajectory of evolution. Determining the order of CNAs in hepatoblastoma may in turn provide therapeutic opportunity for treating hepatoblastomas by understanding how 'evolved' a tumour is and what the subsequent genetic changes are likely to be. This study is unable to present an order of CNA events as a progression model for hepatoblastoma as the study was limited by both patient number (n=11) and sampling (4 patients with only a single tumour sample).

Chapter 7

Discussion

This thesis presents a broad and detailed assessment of paediatric solid tumour evolution. Taking advantage of multiple tumour samples to gain a high resolution picture of paediatric cancer evolution with a particular focus on CNA evolution.

Chapter 3 highlights the beginning of my investigation in a set of 20 WTs. Here my approach involves comparing 'average' CNA profiles across multiple tumour samples. This study highlighted important bioinformatics challenges regarding segmenting the genome across multiple tumour samples. Steps were taken to ensure potential false positive calls were not incorporated across the genome. The approach also showed the limitation of segmenting on the LRR as common CNNLOH/UPD events were unaccounted for by this initial segmentation that led me to develop a downstream analysis to incorporate regions of genomic imbalance.

Nevertheless the work conducted in Chapter 3 proved vital for understanding the initial aspects of WT evolution and the bioinformatic challenges that must be addressed to unveil them. Chapter 3 reveals that WTs can be broadly homogeneous with only few CNAs or can be highly branched with many CNAs. This simple observation of the broad range of evolutionary trajectories was not obvious prior to this study and is an important finding unto itself. By focusing on these patterns I unravelled important aspects of WT genetics that were largely consistent with the expansion of the investigation in Chapter 5. Important findings included biomarker

heterogeneity (particularly of chromosome 1q+), describing the founding CNA events in WT development such as 11p15.5 UPD, and resolving the differential genetic composition of tumour nodules that are important for understanding tumour origin as well as differential treatment response.

Chapter 4 focuses on two major advancements from methodologies proposed in Chapter 3, (1) the detection of a single major allele distribution, essential for accurate segmentation and allelic imbalance modelling and (2) the detection of non-clonal CNAs. A mixture modelling approach proves to be effective at separating the major and minor allele distributions allowing for segments to be modelled using the Battenberg workflow [120]. These improvements allow for more CNAs to be identified as well as more observable clone states to be used for phylogenetic reconstruction as demonstrated in Figure 4.7. These findings lead to more sensitive CNA detection and allowed for a more accurate phylogenetic profile to be drawn. Overall this methodology proved vital for uncovering the details of evolution presented in Chapter 5 and Chapter 6.

Detailing the evolution of 66 PKCs of which 30 were subject to targeted sequencing, Chapter 5 represents one of the most extensive assessments of WT genetics in the field. By carefully dissecting tumour evolution, taking into account histological subtypes as well as tumour laterality, I find consistent observations that are difficult to highlight in smaller studies. Here, consistent patterns of evolution are determined and parallels are drawn between different tumour types. For instance, 19q UPD may provide a link between epithelial and diffuse anaplastic tumours. Parallel evolution also provides a framework for understanding selection in WTs. Finally, I show that tumour phylogenies can be translated into clinically relevant variables. Chapter 6 then demonstrates the use of the approach to extract a maximal impression of tumour evolution in a rare cancer and shows that it evolves mostly linearly as well as being able to form potentially unrelated masses.

Overall this study has highlighted concepts and challenges that are not only important for solid paediatric tumours but also for the study of tumour evolution in general. From their origin and their development to their eventual clinical presentation, studying evolution allows for the nature of individual cancers to be illuminated and their commonalities to be explored. Putting the spotlight on the concept of tumour evolution uncovers the dynamics of cancers and allows us to perceive the disease in new ways.

7.1 Patterns of paediatric solid tumour evolution are diverse

WTs produce a spectrum of different phenotypes, often characterised by a range of different histologies. This thesis shows that WTs are also genetically diverse and this translates to its evolutionary history. There are several examples of the extremes of genomic instability as well as diversification in WTs, from cases with no CNAs to cases with many changes and high ploidy states. Additionally, patterns of evolution can be linear or branched irrespective of mutational burden.

Despite this variation between tumours being complex, there are still patterns which can be mapped onto tumour subtypes. Broadly speaking, stromal, epithelial, regressive, necrotic and focal anaplastic tumour evolutionary patterns were not frequently branched. In regressive and necrotic tumours this may be due to the lack of multiple sampling because of limitations in DNA extraction due to tissue quality and in the focal anaplastic subtype because only one case was assayed.

However, there was sufficient multi-sampled cases taken from stromal and epithelial tumours to be confident that this lack of branching was convincing. These tumours are often evolving linearly and have few CNAs. Interestingly, patterns in these CNAs are strikingly consistent. Most unilateral stromal tumours contained a clonal 11p UPD event and the majority of tumours contained a chromosome 3 aberration, which appears subclonal in some cases. CNAs affecting these two

chromosomes therefore composed the majority of CNAs observed in stromal type tumours. Stromal histologies were also exclusively affected by convergent evolution of *CTNNB1* mutation, indicating *CTNNB1* mutation as a late event in stromal tumour evolution that is under strong selection. All cases with somatic *WT1* mutation the mutation appeared clonal and occurred in tumours that also had *CTNNB1* mutation, in which two cases had heterogeneous *CTNNB1* mutation. One of these cases however was a mixed type tumour. Clearly *WT1* and *CTNNB1* mutation are associated as described previously [199]. I hypothesise that stromal tumours undergo a predictable pattern of evolution. A typical evolution of a stromal tumour likely involves *WT1* mutation followed by a 11p UPD event that involves 11p13 (therefore the *WT1* locus). This then leads to strong selection for *CTNNB1* mutation as well as chromosome 3 gains or CNNLOH/UPD events (no chromosome 3 losses were observed in stromal WT that had a 11p UPD event). This linear progression is similar to Vogelstein's model of colorectal cancer evolution [91] and indeed could be used to test how 'evolved' a given stromal tumour is. For example, detection of *CTNNB1* mutation may be informative of a late stage stromal tumour.

It is unclear why selection for *CTNNB1* mutation is only strong following the 11p UPD event despite *WT1* mutation preceding the CNA in some cases. Potentially *IGF2* overexpression caused by CNNLOH/UPD of both 11p15.5 and 11p13 could play a role in triggering *CTNNB1* mutation selection in the genetic context of a *WT1* mutated tumour. This progression model has already been implicated in mouse models developed by Huff and colleagues. Here, *Wt1* ablation and *Igf2* upregulation were shown to be capable of forming WTs in mice and more recent experiments have also tested *Wt1* ablation in combination with stabilised *Ctnnb1* in nephron and stromal progenitors of the kidney. Interestingly, double knock out of *Wt1* and *Ctnnb1* stabilisation showed the highest potential for generating WTs in mice in nephron progenitors [200, 201]. The tumorigenicity of *Wt1* double knock out, *Igf2* overexpression and *Ctnnb1* stabilisation was not tested. Additionally, there is evidence of many *CTNNB1* mutations in a single WT case suggesting tumours

of independent origin or parallel evolution [202]. This model of evolution is potentially also occurring in stromal components of other WTs for instance in mixed type tumours. RR4 in IMPORT 147, which is a mixed type tumour, uniquely possesses a chromosome 3 gain and a *CTNNB1* and unlike the other samples which showed mixed histology RR4 showed 95% stromal cells. Mixed type tumours could be conglomerates of different subtypes evolving in parallel, as appears plausible given the observations of diversification in Chapters 3 and 5. Overall stromal tumour evolution is the most predictable sequence of events in any WT subtype.

Epithelial tumours appear to also progress largely linearly with few events. However, like 11p UPD in stromal tumours, clonal 19q UPD is common – indeed, 3/5 unilateral epithelial tumours contain only a 19q UPD event and the only case with bilateral epithelial tumours showed convergent evolution of 19q UPD. Three diffuse anaplastic tumours (IMPORTs 10, 83, 106) also appeared to possess clones with only 19q UPD present whereas many other CNAs were present in other clones across the tumour samples. We sequenced two of these cases and found heterogeneous *TP53* mutation in both (IMPORTs 10, 106). This possibly indicates that 19q UPD may act as a founding event for late *TP53* mutation causing heterogeneous CNA instability and diffuse anaplasia. This could also potentially implicate epithelial tumours as having a higher risk of developing anaplasia. Furthermore, evidence for an early role of 19q UPD in Wilms' tumorigenesis, in a similar fashion to 11p15 UPD, suggests that the subtle effects of UPD may be particularly advantageous in the early stages of WT development.

Additionally, hepatoblastomas appear less diverse than WTs in terms of their evolutionary trajectories. In Chapter 6 they are seen to always evolve linearly with an association with clonal 11p UPD which can often be present in normal tissue. Although they display a range of a genomic instability similar to WTs.

Mixed, blastemal and diffuse anaplastic type WTs are most likely to develop

branched evolution and genomic instability. However, even in diffuse anaplastic tumours, which are associated with unstable genomes, there are tumours with little branching and very few events (e.g. IMPORTs 20, 54, 83). However, it may be possible that in these tumours, regions that contained *TP53* mutation and an increase in CNA events were not sampled for study (although a E339del *TP53* mutation was found in IMPORT 54, it seemingly does not cause extensive CNA instability, suggesting not all *TP53* mutations lead to CNA instability). Clearly further work is required to elucidate the pattern of development of these tumours, particularly in blastemal and mixed type tumours.

7.2   WTs and hepatoblastomas can develop separate tumours in parallel

Evidence in Chapters 3 and 5 shows that bilateral tumours are likely to independent masses and such be treated as such. Additionally Chapters 3, 5 and 6 show examples of tumours which appear independent but present in the same organ. Clearly both WTs and hepatoblastomas can form tumours which are highly divergent. This is not exclusive to solid paediatric tumours as it has been suggested that independent tumours appear in other adult cancers also, for instance breast and lung cancers [65, 100].

However, whether these tumours are completely independent, especially in the case of two highly divergent tumours in the same mass, is cast into doubt by observations in IMPORT 54 (Fig. 5.23). The CNAs in the tumours samples taken from the potentially independent tumours are not shared (when taking the interpretation that separate CNA breakpoint boundaries are separate events). However, when sequencing the samples a single *DROSHA* E993K mutation was shared between the candidate independent tumours, implying they did not exist independently. Not all cases with potentially independent tumours were assayed by sequencing, for example in Case 13 from Chapter 3, and if targeted sequencing was extended to other cases, it may have identified a shared ancestral mutation between samples

also making them divergent but not independent.

WGS would likely be the most comprehensive method of assaying these tumours to separate the difference between early divergence and independence. Indeed, such an assay may be applicable to bilateral tumours also. There is evidence in IMPORT 54 that the *DROSHA* E993K mutation is present in ~2.45% of cells in the normal sample, implying that the mutation may be mosaic and potentially causing a field effect from which the blastemal and diffuse anaplastic type tumours arose from by exploring different fitness landscapes. Despite this possible mosaicism being on the edge of detection, even for deep sequencing, ~2.45% of cells in a tissue sample likely translates to 1–10 million cells, each of which can prime subsequent steps in oncogenesis due to their apparent tumorigenic phenotype. CNAs may also be mosaic as suggested in Patient 3 in Chapter 6, for example.

An interesting point of cancer aetiology is the question of at what stage should we consider tumours unrelated as all cells in the body derive from the fertilised egg and ultimately have a shared lineage. If at each cell division there is a possibility of mutation, it is then possible to trace this lineage in a method similar to studying cancer evolution [203, 204]. All divergent tumours including bilateral tumours may be affected by a mutation early during cell division in embryogenesis. Bilateral tumours may simply be characterised by this mutation occurring early enough that it is passed to both kidneys. This may be caused by poorly characterised WT mutations that cause only mild oncogenic potential that may explain the latency of bilateral tumour developing, leaving normal kidney development largely unaffected. Additionally, there is evidence that CNAs have the potential to occur very early on in gestation [205].

Regardless of the specifics of cancer aetiology, it is practically important for clinicians to recognise that tumours can be formed by masses that are divergent enough genetically that they should treated as independent cancers. Furthermore, only

187

genetically assaying multiple, phenotypically different samples will identify this phenomena.

7.3 WTs and hepatoblastomas support gradual evolution

In Chapter 1 I outlined that there are two broadly contrasting theories of the general pattern of cancer evolution, (1) gradual evolution, which involves successive accumulation of driver mutations and clonal sweeps that establish a clone carrying the new driver mutation as the dominant clone and (2) punctuated evolution in which evolutionary tempo increases in short bursts causing wide ranging diversification which is then followed by a period of expansion but no large changes in clonal architecture.

The common detection of observable ancestors and linear evolution appears to suggest that WTs and hepatoblastomas may tend to develop as part of the gradual model. Here observable ancestors represent 'older' clones that have not been eliminated from the cancer population and this are thought to be a consequence of a gradual model of evolution [92]. However, the detection of *TP53* mutation in tumour regions with large numbers of CNAs, in diffuse anaplastic cases, may suggest the late acquisition of genomic instability that has not dominated the tumour. Here *TP53* mutation may lead to a punctuated burst of genetic diversity that is localised spatially. Additionally, some heavily branched tumours may also fit into the burst of diversity associated with the punctuated model of tumour evolution, especially when the relatedness of the samples is weak (e.g. IMPORT 3, IMPORT 106, IMPORT 192).

Mapping phylogenetic trees to tumour images can given an impression of the growth of clones. These images remind us that tumour clones are spatially constrained in 3D spaces (Fig. 5.14). Tumours such as IMPORT 16 (Fig. Fig. 5.14B) and IMPORT 171 (Fig. Fig. 5.14E) seemingly support a gradual model of evolution

as observable clones are still present and are positionally ordered along the tissue in a manner that suggests expansion of clones away from the normal kidney. Perhaps in these cases older clones are not outcompeted by new clones as the tumour still has space to expand into, therefore allowing for the accommodation of both the 'new' and the 'old' clone. Other more branched phylogenies may represent rapid expansion of multiple clones into several direction which may have occurred in a punctuated manner.

These assumptions of the temporal dynamics of clonal evolution must be treated with caution however as evolution is not being directly measured in real time. Assays of primary tumour resections are snapshots of a point in evolution that can be interpreted in the perspective of a model, but the process has not been directly observed. Direct observation are only possible in temporal studies of tumour evolution. This thesis presents two types of temporal evolution assessment, (1) pre-treatment, pre-resection biopsy samples in comparison with resection samples and (2) ctDNA assessment. Conclusions derived from pre-treatment biopsies are difficult to draw when only a single biopsy is taken, like in this thesis. This is due to the fact that it is impossible to know if clones identified in the pre-treatment biopsy are also present in the primary tumour at resection if they are not observed in the tumour samples taken, therefore temporal context is difficult to pinpoint. Multiple biopsies may provide a more extensive phylogeny of the cancer population prior to treatment for comparison with the resection samples but it is practically difficult to perform large numbers of invasive biopsies. In this thesis pre-treatment biopsies were used to show the existence of clones prior to resection and to physically map the location from which the biopsy was taken from as a proof of the power of phylogeny mapping (IMPORT 12).

Assaying ctDNA is more likely to be informative of temporal clonal dynamics. In this thesis I showed that *TP53* allele abundance in serum and plasma is variable across the course of treatment. In this study this likely translates to tumour burden

and not clonal dynamics. However, work by Jamal–Hanjani and colleagues has shown that subclone dynamics and emergence are traceable across time in ctDNA [94]. This assessment would be possible in WTs as Section 5.13 shows that ctDNA provides detectable signal of an important mutation when it is both clonal and sub-clonal (IMPORT 233).

Measuring time in a snapshot assessment of tumour evolution may be possible by inferring time using mutational signatures associated with age. Mutational signatures as defined by Alexandrov and colleagues showed that the 1A/B signatures strongly correlated with age positively. Signatures 1A/B are associated with spon-taneous deamination of 5'-methylcytosine which is considered a sporadic process across the lifetime of an individual and could be associated with clonal and sub-clonal mutations to infer real time [66, 74]. The applicability of this assessment to very young cancers is underappreciated.

Overall, despite gradual and punctuated evolution representing two competing ideas in cancer evolutionary theory, it is highly likely that both of these models are relevant. Gradual and punctuated evolution likely depends on context and may even co-exist between tumours of the same type or even within the same tumour.

7.4 Convergence is a common feature of paediatric solid tumour evolution

There are several examples of convergent evolution in Chapters 3, 5 and 6 of this thesis. Parallel evolution is seen between tumours that appear independent, includ-ing bilateral tumours as well as tumours which appear independent in the same organ, and during the evolution of a single tumour. Parallel evolution is highly in-formative as it provides evidence for which mutations are likely being selected for at different stages of tumour evolution. Often parallel evolution shows researchers how mutations are selected for at later stages of tumour evolution, for example convergent evolution of *SETD2* in renal carcinomas [122]. However, the examples

of convergent evolution in independent tumours also reveals important mutations in the early stages of tumour development, for example 11p UPD in WT.

Parallel evolution acts like a repeated experiment in the sense that there are several cells of a clone with similar genomes and when parallel evolution occurs, multiple cells have navigated the same pathway through the fitness landscape they are presented with. Parallel evolution is an example of how selection acts on the whole population and that different clones with the same mutation can establish themselves before a parallel clone can dominate the tumour.

In this study parallel evolution is mostly seen in CNAs and mutations which are of known importance in WT development, confirming their relevance to the disease. Examples of parallel evolution between independent tumours appear to often involve CNAs that are important for the early stages of WT development such as 11p and 19q UPD. Examples of parallel evolution within an individual tumour is perhaps evidence for the importance of particular mutations later in the tumour's life history. These mutations include 1q+, 7p–, 17p LOH and mutation of *TP53* and *CTNNB1*.

The mechanism by which some of these mutations drive WT evolution are already characterised, for example *IGF2* overexpression and *WT1* homozygosity explains selection for 11p UPD, the evolutionary benefits of genomic instability is the likely cause of selection of *TP53* mutation and 17p LOH, and WNT signalling dysregulation leads to selection of *CTNNB1* mutation. However, the phenotypic alteration leading to selection of 19q UPD, 1q+ and 7p– remain unresolved but are clearly important in WT development, although 19q UPD may be selected for due to C19MC dysregulation. Furthermore, by examining MSAI in CNAs it is possible to estimate if these changes are allele specific in nature, for example 11p UPD did not show MSAI in this thesis yet 1q+ was not allele specific and regularly displayed MSAI.

Mapping convergent evolution to tumour images can reveal the local vicinity in which parallel evolution emerges. For example, IMPORT 147 (Fig. 5.14D) shows parallel evolution of both 1q+ and 7p– in a space of ~1cm, indicating a potentially strong local selection pressure for these changes in a relatively small region of the tumour. Local parallel evolution may represent microenvironment selection pressures specific to a tumour region. The tumour's dimensions are 16.5cm x 14.5cm x 7cm, indicating just how much diversity may be being missed in such a large mass.

Parallel evolution presents several examples in which epistasis may be occurring. In 3 stromal cases in the PKC66 series I observed parallel evolution of *CTNNB1* mutation. Considering this followed an 11p UPD in each case and in one case a *WT1* mutation, this could be an example of the epistatic relationship between 11p UPD and *CTNNB1*, in which selection for *CTNNB1* mutation is greatly increased following a 11p UPD event, leading to extensive convergent evolution. Similarly, the presence of 2 (possibly 3) *TP53* mutations in IMPORT 106 following a single 19q UPD event, as well as the presence of 19q UPD as the initiating event in several other diffuse anaplastic tumours, may indicate epistasis between 19q UPD and *TP53* mutation.

It is worth recognising that, as illustrated by Alves and colleagues [135], convergent evolution can only be accurately recognised if the mixing of cancer cell population within a single sample is addressed. Sample trees, as produced in Chapter 3, can incorrectly show parallel evolution if a clone is present as a fraction of multiple samples. I only describe convergent evolution in Chapter 5 in which my methods address the possibility of CNA profiles existing as admixtures in each tumour sample. The vast majority of examples of convergent evolution of CNAs in this series are also supported by clear differences in breakpoint boundaries or via the observation of MSAI. Nearly all described observations of convergent evolution derived from small sequence mutations in the targeted sequencing data involve

unique mutations affecting the same gene. The correct reasoning presented by Alves and colleagues is in reference to incorrectly called convergent evolution of mutations that are identical.

Interestingly Lipinski and colleagues suggest that parallel evolution is an inherent property of large tumours which have a low mutation rate [78]. As WTs contain both of these properties, it may be that parallel evolution is a predictable property of WT development. It is likely that even with up to seven samples per tumour we may not be able to detect all clones that are undergoing parallel evolution, particularly if it can occur in a small region of the tumour as in IMPORT 147. Furthermore, IMPORT 143 which developed 4 (possibly 5) *CTNNB1* mutations challenges the concept that genomic instability is required in order to 'hit' driver mutations frequently, as in this case only a clonal *WT1* mutation was also present as well as just two CNAs. This may support evidence that some regions of the genome are more vulnerable to mutations than others. Although to judge this one must take into the numbers of cell divisions that have occurred prior to diagnosis as well as the mutation rate of the tumour.

7.5   Inferring cancer evolution remains challenging

Inferring cancer evolution is a conceptually and computationally difficult problem. Deconvoluting mutational heterogeneity in both small sequence mutations and CNAs remains very challenging. This thesis highlights several of these challenges.

Using multiple tumour regions is a powerful method of finding the clonal architecture of a heterogeneous tumour as the spatial separation of clones provides key information of the clonal belonging of mutations. However, as each region is analysed separately, it demands that analytical methods are capable of performing well across each tumour region to detect mutations across each site. Furthermore, there is a cumulative effect in the unwanted detection of false calls as these calls will be accumulated across all respective samples. This is true for detection in both

small sequencing mutations and CNAs, however CNAs are further complicated by achieving accurate breakpoint boundary detection. Most of these potential artefacts are affected by altering tumour sample purity as well as issues relating to varying qualities of the generated data per region (variable coverage in NGS experiments and variable noise levels in SNP arrays).

I address these challenges in this thesis by, for example (1) removing regions which were small or had unusual tiling density from my analysis in Chapter 3 to remove spurious CNA calls that remodel phylogeny, (2) using CGHregions to smooth boundaries between segments across multiple samples as well as using centromere splitting to test for improvements in segmentation and (3) by searching for evidence of all mutations called in any sample across all samples in data from targeted sequencing in Chapter 5. Unfortunately low tumour purities can lead to highly divergent CNA segments and in some examples I had to manually curate segments which were erroneous in order to achieve accurate phylogenies in Chapters 5 and 6. False detection as well as incorrectly segmented regions can dramatically alter inferred phylogenies.

Once each segment has been identified, work in Chapters 4, 5 and 6 utilised the Battenberg workflow to test for possible combinations of CNAs that may produce the LRR and BAF observed in a particular segment. Firstly, this requires very accurate estimations of LRR and BAF in order to model small percentage changes in CNA composition, which is addressed in Chapter 4. Most importantly however, solving a two-state mixture of CNAs for a segment produces infinite solutions. For example, a 2+0 state in 100% of cells may also be explained by a 50:50 mixture of 3+0 and 1+0. Additionally, MSAI within a single sample can make its detection very difficult as, for example, a 50:50 mixture of two 2+1 gains displaying MSAI would appear as a 2+2 state with a 50% purity. Solving >2 states makes this problem increasingly complex and yet is possible in a cancer cell population. Deciding on the appropriate mixture often derives from choosing the solution with the closest

distance from the diploid state. In this thesis I always initially test combinations of 1+1 and the nearest aberrant state (normal + aberrant). This allowed for 2+0 and 1+1 mixtures to be tested, which are known mixtures in WTs and hepatoblastoma. Modelling a mixture with 1+1 was not appropriate in cases with probable WGD.

Inferring clonal architecture in a sample from these CNA mixtures also requires assumptions. Here I assume CNAs that are in all cells arose first in tumour evolution, therefore implying a phylogenetic relationship when reconstructing the clones in a single sample. I also assumed that CNAs arise from a 1+1 state and that 1+1 is always representative of an ancestor despite the fact that CNAs could be reversed from an aberrant state to 1+1. In mixtures of two CNAs I assumed that the nearest state to diploid occurred prior to the other CNA in the mixture. These assumptions are likely often true but are not rules of cancer evolution.

Furthermore, subsequent clustering of CNAs is based on the assumption that CNAs of a similar cancer cell fraction arise from the same clonal expansion. This is very likely to be true in clusters >50% cancer cell fraction as these likely co-exist due to the pigeon hole principle. However cancer cell fraction clusters in less than 50% of cells cannot be applied to the pigeon hole principle. Mixing two branched clones equally would yield two groups of cellularies of ~50% that by the principle of 'similar cancer cell fraction equals same clone' would lead to these clones being grouped together. As cancer cell fractions decrease, the space for the number of potentially, branched clones producing similar cancer cell fractions increases exponentially and, for example, at ~10% cancer cell fraction there are 10 subclonal 'compartments' in which mutations may exist in various combinations. In this thesis I use the assumption that mutations of a similar cancer cell fraction belong to the same clone and that clones have evolved linearly within the same sample unless mutations in other samples conflict with this interpretation. For almost all cases there was no conflict. One interpretation of CNA states that necessitates branched evolution within a sample however is the possibility of CNAs overlapping as these

states cannot co-exist.

To my knowledge there is no approach to solving the problem of CNA evolution that addresses all of the following problems:

1. Interpreting within sample phylogenetic relationships of CNA cancer cell fraction in the context of multiple region datasets.

2. Interpreting different breakpoint boundaries in different regions as being likely to have been produced by separate events.

3. Interpreting the possibility that CNAs in different clones with different breakpoints may co-exist in the same sample.

4. Understanding that different alleles may have been affected by similar CNAs as is the case with MSAI.

5. Understanding that a WGD is a potential single event state change but one that is likely to be infrequent.

6. Understanding that 2+0 states probably arise as a consequence of mitotic recombination and are therefore a single event.

7. Identifying chromothripsis and calling it as a single chaotic event.

8. Interpreting CNA data in a platform naïve manner. The majority of algorithms that provide an interpretation of CNA evolution are designed for NGS data input.

Overall CNA evolution is a highly complex problem that requires further study and this thesis provides evidence of the challenges that require addressing. The problem of CNA evolution produces multiple solutions and therefore requires many assumptions to settle on an answer. To solve CNA phylogenies in Chapters 4, 5 and 6 I used manual curation of multiple interpretations of the data by using Battenberg and MEDICC. This approach is not sustainable for larger cohorts and more genomically unstable tumours. Indeed, I relied on MEDICC to interpret complex tetraploidy and

chromothripsis.

Additionally, assumptions are made when interpreting sequence mutation, including the belonging of a mutation to a CNA or the infinite sites assumption that each alteration only occurs once, which may possibly be incorrect in R3 of IMPORT 143.

New algorithms, such as Canopy and SPRUCE, signal continued development in the field [164, 165]. However, despite the plethora of algorithmic development in cancer evolution, each approach still requires extensive testing and manual assessment as well as conceptual critique. Different cancer types present different contexts and problems and paediatric solid tumours are an example of that. Future holistic approaches may seek to develop algorithms applying principles to CNA evolution that are synthesised by mathematical modelling and evolutionary theory but also that account for cancer biology and mechanisms of genomic instability.

Single cell approaches may be the future of studying CNA evolution over bulk assays as no inference of CNA/clonal mixing is required. However it may be several years before data quality issues are resolved for consistently accurate profiling and for enough cells to be sequenced across representative regions of the tumour. Furthermore, the prospect of the erroneous sequencing of multiple cells makes detecting WGDs difficult [136]. These assays may require 10s of 1000s of cells to be assayed to capture the full repertoire of CNA evolution which will likely make studies of large tumour cohorts an expensive, labour intensive task without technological advancements.

7.6 Understanding cancer evolution is vital for clinical biomarkers

At the same time as addressing key fundamental questions of paediatric solid tumour evolution, this thesis relates findings to outstanding clinical questions. One of the primary aims of the investigation in Chapter 3 was to assess CNA biomarker

heterogeneity in WT, particularly of chromosome 1q gain. This CNA is attractive as a biomarker in WT as it is common and shows value for prognostication [56, 57, 58]. However, as results in Section 3.5 show (as well as results across tumours in Chapter 5), chromosome 1q gain is frequently heterogeneous. Other important biomarkers appear to be clonal more frequently, such as 11p15 UPD and 16q loss, the former CNA was proposed as the most prognostic CNA in WT in a recent systematic literature review [187]. Discovering CNAs with the potential to be frequently clonal may circumvent the detection challenges associated with heterogeneous biomarkers.

Understanding evolution however is vital for interpreting heterogeneity. If Case 13 wasn't interpreted as being composed of two independent masses in Chapter 3, 16q loss and 17p loss would have been interpreted as heterogeneous. Furthermore, reanalysis of Case 19 in Chapter 5 as IMPORT 12 shows that despite chromosome 1q gain presenting as homogeneous in Chapter 3, recognition of MSAI and a more sensitive detection of higher copy number states dictates that chromosome 1q gain likely occurred four times in the tumour and is not a clonal event despite appearing homogeneous.

Due to the fact no single CNA biomarker appears to satisfactorily explain WT outcomes, I hypothesised that mutational burden, interpreted as the number of CNA event occurring across the evolution of a multiple sampled WT, may be prognostic. Indeed, results presented in Section 5.8 suggest that the number of CNA alterations occurring during the evolution of a WT is prognostic and that if a tumour possesses more than 11 CNA events, it should be considered high-risk. Interestingly, this appears to also be prognostic within currently defined high-risk patients, potentially indicating that high-risk patients with low CNA burden could be candidates for treatment reduction or that high-risk patients with high CNA burden require longer treatment regimes. To assess this biomarker, multiple tumour samples must be assessed and tumour phylogenetics must be inferred. Sampling biases may cause

the detection of clones of varying genomic instability and poor sampling may miss clones with high numbers of CNA events such as R1 in IMPORT 106, that are caused by heterogeneous *TP53* mutation. A simplified, single test, that would not require sophisticated phylogenetics to assay mutational burden in a clinical trial may involve sampling a fixed number of tumour regions and mixing DNA into a single sample assay to 'count' the number of CNAs regardless of their phylogenetic relationship. This could however still lead to an underestimate of CNA burden, for example, if MSAI is present across the tumour regions. Biomarkers such as chromosome 1q gain, that are common but weakly prognostic and heterogeneous, may simply be a proxy measure of CNA burden. CNA burden may be prognostic as it represents the progress of evolution in a given tumour and may be particular useful as it is agnostic of tumour histology. Unfortunately, as event-free survival in WT is high, larger numbers of cases are required to be confident of a prognostic value.

Mengelbier and colleagues have proposed that detecting heterogeneity – which they termed 'microdiversity' – in single tumour samples is prognostic in WTs [115]. Across multiple samples in the PKC66 study, I did not find subclonality, as defined as percentage of CNA event that are not clonal, to be a useful biomarker, as despite it showing statistical significance, it was outperformed by mutational burden. This is supported by the fact that the threshold generated by ROC curve analysis only separated five cases as having 'high' subclonality ($>81.4\%$ CNAs subclonal), each of which showed more than 11 events. Furthermore, my own analysis of clonal mixing in single samples revealed that the detection of clonal mixing is also subject to sampling bias. Many tumours in the PKC66 series possessed samples both with and without 'microdiversity' indicating that detecting heterogeneity is itself subject to issues of heterogeneity. I hypothesise that understanding the phylogenetic history is more important than understanding its by-product, heterogeneity.

Understanding evolution may overcome barriers set by histological classification.

For example, the diffuse anaplastic phenotype could be replaced by a subtype that presents with high CNA instability and *TP53* mutation or a 11p UPD–*CTNNB1* mutation subtype could replace stromal type tumours. Introducing a molecular basis of subtyping may overcome unsatisfactory histological classifications such as 'mixed' type which is highly diverse both molecularly and phenotypically.

The molecular relationships between subtypes also requires further expansion. Future investigations primed by this research may seek to determine in a larger cohort of epithelial tumours if 19q UPD is prognostic of event-free survival and if relapses in these cases tend to present as diffuse anaplastic tumours considering the association between epithelial and diffuse anaplastic tumours and clonal 19q UPD in this study. Furthermore, regressive type WT interestingly contribute to 44% of all metastasising WTs yet is a tumour defined by response to therapy (>66% necrotic cells) [14]. This could potentially be due to the fact that diversity within this type is under recognised and some regressive type WTs may possess aggressive clones. Further studies may seek to assay metastasising regressive WTs to identify metastatic clones that may expand in size due to 'competitive release', such a study may require prudent microdissection of viable tissue in order to assay DNA.

7.7   Final thoughts

The principles of evolution are redefining how the field perceives cancer biology and treatment. This work contributes to a refinement of our outlook of solid paediatric tumour evolution with a particular focus on WTs. Liquid biopsies and advanced phylogenetic tumour profiling should advance the clinic in new unprecedented ways. Clinical approaches that take into account the principles of tumour evolution should be used to fine-tune methods of drug development and biomarker identification as well as patient care. As we increase our understanding of tumour evolution we must seek to determine ways to make a phylogenetic understanding of a cancer translatable to a health-care scenario. This may require phylogenetic algorithms to take a form that is understandable to non-experts and for researchers

to determine easy to interpret output variables that describe tumour evolution.

Research into WTs and hepatoblastomas will be advanced by assaying more cases and in WTs specifically, by building larger cohorts of individual subtypes. One possible route of analysis in non-anaplastic WTs may be to disregard the blastemal, epithelial, stromal and mixed subtypes and to study general principles across the spectrum of triphasic histology as a whole. This may unravel the complexity of the mixed WT subtype.

Additionally, tumour sampling should also be increased to gain a more detailed picture of 3D tumour composition in future studies. This is limited by tumour size and tissue quality, however large tumours could be potentially sampled 10 or more times, effectively tripling the number of samples taken on average in this study. In smaller tumours, sampling smaller tumour regions may aid the generation of a higher number of samples. These higher sampling strategies would require expert tissue dissection and therefore a good working relationship with clinical colleagues, as higher sampling increases labour time and the need for streamlined documentation. Larger studies may seek to produce a device for routine sampling that may cut a tumour into regular shaped cubes designed for easy labelling and documentation of the location of cubed tissue. Such a device may also allow for 3D computational reconstruction of the tumour clones following a genomic assay as demonstrated by Mamlouk and colleagues [206]. Similarly future phylogenetic algorithms may seek to consider 3D spatial constraints of tumour region locations when generating phylogenetic trees and resolving clones in studies with a large number of multiple samples.

Further work may also integrate methylomics and transcriptomics in the study of WT and hepatoblastoma evolution. These additional data types may be particularly informative in cases with few somatic mutations or CNAs. However, the heterogeneous cellular composition of WTs will produce cell type specific expres-

sion patterns that would have to be delineated from patterns produced specifically by evolutionary diversity. Tumours with no obvious driver mutations may also be subject to WGS in order to determine potential intergenic mutations.

One case study in this thesis in Chapter 3 related tumour phylogeny to specific response of nodules in a WT to chemotherapy. Performing an integration of tumour imaging, sampling and phylogenetic analysis in a larger cohort of patients would produce an invaluable study for determining the genetic rationale for treatment response. As mentioned previously, this would require a close working relationship between cancer geneticists and radiologists, as well as the creation of databases that are practical for both sets of experts.

This thesis begins with the observation of the simple features of genetic heterogeneity driven by tumour evolution in WT, but extends to the complex phylogenetic patterning of WT subtypes. Despite assaying over 90 cases of WT and hepatoblastoma, it appears that the repertoire of evolutionary patterns in solid paediatric tumours is only now being gradually unravelled.

Appendix A

List of abbreviations

Apparent diffusion coefficient	ADC
APC membrane recruitment protein 1	AMER1
Area under the curve	AUC
B-allele frequency	BAF
B-Raf proto-oncogene	BRAF
Band-width	BW
UK Children's Cancer and Leukaemia Group	CCLG
Copy number aberrations	CNA
Copy number neutral loss of heterozygosity	CNNLOH
Circulating tumour deoxyribonucleic acid	ctDNA
Catenin Beta 1	CTNNB1
Digital droplet polymerase chain reaction	ddPCR
DiGeorge syndrome critical region gene 8	DGCR8
Diffuse hyperplastic perilobar nephroblastomatosis	DHPN
Differentially methylated region	DMR
Deoxyribonucleic acid	DNA
Diagnosis	Dx
Expectation maximisation	EM
Eleven-nineteen-leukaemia protein	ENL
Frozen biopsy	FB
Fluorescence in-situ hybridization	FISH
Frozen nephrectomy	FN
Frozen relapse	FR
Set of 11 hepatoblastoma cases	Hep11
Insulin-like growth factor 2	IGF2
Improving population outcomes in renal tumours	IMPORT
Insertion and/or deletion	indel
Interquartile range	IQR
Intra-tumour genetic heterogeneity	ITGH
Kelch-like ECH associated protein 1	KEAP1
Loss of heterozygosity	LOH
Loss of imprinting	LOI
Log R ratio	LRR
Mutant allele frequency	MAF
Mitogen-activated protein kinase	MAPK
Megabases	MB
Mirrored B-Allele Frequency	mBAF
Minimum event distance for intra-tumour copy-number comparisons	MEDICC

Micrometastasis project breast carcinoma samples	MicMa
Micro ribonucleic acid	miRNA
Myeloid/lymphoid or mixed-lineage leukemia translocated to, 1	MLLT1
Most recent common ancestor	MRCA
Magnetic resonance	MR
Mirrored subclonal allelic imbalance	MSAI
Methylation-specific multiplex ligation-dependent probe amplification	MS-MLPA
Myc proto-oncogene	MYC
Neuroblastoma MYC	MYCN
Nuclear Factor, Erythroid 2 Like 2	NFE2L2
Next generation sequencing	NGS
Normal kidney	NK
Oncogenic micro ribonucleic acid	onco-miR
Piecewise Constant Fitting	PCF
Phosphatidylinositol-4,5-bisphosphate 3-kinase	PI3K
Paediatric kidney cancer	PKC
Set of sixty-six paediatric kidney cancer patients	PKC66
Pretreatment extent of disease	PRETEXT
Retinoblastoma transcriptional corepressor 1	RB1
Ribonucleic acid	RNA
Receiver operating characteristic	ROC
Standard deviation	sd
International Society of Paediatric Oncology	SIOP
Sine oculis homeobox homolog	SIX
Single nucleotide polymorphism	SNP
Single nucleotide variation	SNV
Tumour Protein P53	TP53
Uniparental disomy	UPD
Urinary tumour deoxyribonucleic acid	utDNA
Ultra-violet	UV
Variant allele frequency	VAF
Whole genome duplication	WGD
Whole genome sequencing	WGS
Wilms' tumour	WT
Wilms tumour on the X	WTX
Wilms' tumour 1	WT1
Initial Wilms' tumour series of twenty patients	WT20

Appendix B

Data produced by targeted sequencing analysis in PKC66

Case	Region	Gene	Mut	R	V	A	B	C	M
54	R1	DROSHA	T1108P	0.84	0.49	2	1	1.00	1.16
54	R1	DROSHA	E993K	0.84	0.47	2	1	1.00	1.12
54	R1	SIX1	Q177R	0.84	0.45	2	1	1.00	1.06
54	R2	DROSHA	T1108P	1.00*	0.47	2	1	1.00	0.95
54	R2	DROSHA	E993K	1.00*	0.53	2	1	1.00	1.06
54	R2	SIX1	Q177R	1.00*	0.45	2	1	1.00	0.89
54	R3	DROSHA	E993K	0.92	0.45	2	1	1.00	0.98
54	R3	TP53	E339del	0.92	0.83	1	1	1.00	0.98
54	r4	DROSHA	T1108P	1.00	0.19	2	1	1.00	0.39
54	r4	DROSHA	E993K	1.00	0.18	2	1	1.00	0.36
54	r4	GNAS	R201H	1.00	0.23	2	1	1.00	0.46
143	R1	CTNNB1	T41A	0.97	0.45	2	1	1.00	0.92
143	R1	WT1	Q26fs	0.97	0.86	2	2	1.00	0.89
143	R2	CTNNB1	P44A	0.89	0.29	2	1	1.00	0.65
143	R2	CTNNB1	S45P	0.89	0.29	2	1	1.00	0.66
143	R2	WT1	Q26fs	0.89	0.77	2	2	1.00	0.87
143	R3	CTNNB1	S45Y	0.86	0.65	3	2	0.20	1.43
143	R3	WT1	Q26fs	0.86	0.70	2	2	0.97	0.81
143	R4	CTNNB1	T41A	0.91	0.47	2	1	1.00	1.03
143	R4	WT1	Q26fs	0.91	0.71	2	2	1.00	0.78*
143	R5	CTNNB1	S45Y	1.00	0.35	2	1	1.00	0.70
143	R5	WT1	Q26fs	1.00	0.51	2	2	0.67	0.51

Table B.1: A table of the variables used for calculating mutation cellularity. *R* is the tumour purity as determined by ASCAT. In R2 of IMPORT 54 the purity is manually changed to 1 due to the sequencing mutations being much purer than the copy number changes (0.59). The cellularity of the *WT1* Q26fs mutation was called lower than the clonal cluster, but is considered clonal due to the weight of evidence in R1–3 and R5. VAFs for *TP53* E339del and *WT1* Q26fs were derived from the raw data, whereas VAFs for the other mutations was taken from MuTect.

IMPORT	Region	Gene	Contig	Mut	Ref	Alt	VAF
3	R3	AMER1	chrX:63412095	G>A	9	161	0.95
3	R4	AMER1	chrX:63412095	G>A	13	114	0.90
3	R5	AMER1	chrX:63412095	G>A	143	89	0.38
3	R6	AMER1	chrX:63412095	G>A	173	123	0.42
8	R1	PIK3C3	chr18:39575890	C>T	440	24	0.05
8	R3	PIK3C3	chr18:39575890	C>T	364	11	0.03
8	R4	PIK3C3	chr18:39575890	C>T	16	3	0.16
8	R5	PIK3C3	chr18:39575890	C>T	442	8	0.02
8	R2	RYR1	chr19:38964226	G>A	491	92	0.16
8	R4	RYR1	chr19:38964226	G>A	84	12	0.12
8	R1	TP53	chr17:7577538	C>T	30	179	0.86
8	R2	TP53	chr17:7577538	C>T	45	67	0.60
8	R3	TP53	chr17:7577538	C>T	35	144	0.80
8	R4	TP53	chr17:7577538	C>T	12	45	0.79
8	R5	TP53	chr17:7577538	C>T	21	142	0.87
8	R6	TP53	chr17:7577538	C>T	5	87	0.95
10	R3	TP53	chr17:7577085	C>T	4	157	0.98
12	R3	GLI3	chr7:42063065	A>C	269	10	0.04
15	R2		chr11:2021676	G>A	207	8	0.04
15	R3		chr11:2021676	G>A	335	18	0.05
15	R4	ABL2	chr1:179112169	C>T	250	46	0.16
15	R2	ALK	chr2:29443901	G>T	182	4	0.02
15	R3	ALK	chr2:29443900	G>C	129	13	0.09
15	R3	ALK	chr2:29443901	G>T	132	12	0.08
15	R3	REST	chr4:57785960	C>T	379	35	0.08
15	R4	RYR2	chr1:237670117	A>G	184	18	0.09
15	R4	USP9X	chrX:41026835	G>A	72	42	0.37
36	R3	CUX1	chr7:101892227	C>A	387	10	0.03
36	R4	CUX1	chr7:101892227	C>A	193	5	0.03
36	R6	CUX1	chr7:101892227	C>A	469	3	0.01
36	R1	NBPF10	chr1:145325997	A>G	319	15	0.04
36	R2	NBPF10	chr1:145325997	A>G	143	28	0.16
36	R3	NBPF10	chr1:145325997	A>G	72	16	0.18
36	R4	NBPF10	chr1:145325997	A>G	68	14	0.17
36	R5	NBPF10	chr1:145325997	A>G	92	23	0.20
36	R6	NBPF10	chr1:145325997	A>G	122	15	0.11
36	R4	SALL3	chr18:76755106	C>A	251	10	0.04
36	R6	SALL3	chr18:76755106	C>A	480	6	0.01

continued...

IMPORT	Region	Gene	Contig	Mut	Ref	Alt	VAF
47	RR1	CTNNB1	chr3:41266136	T>C	206	148	0.42
47	RR2	CTNNB1	chr3:41268766	A>T	130	97	0.43
47	RR3	CTNNB1	chr3:41268766	A>T	194	142	0.42
47	RR4	CTNNB1	chr3:41268766	A>T	123	99	0.45
54	r4	DROSHA	chr5:31422991	T>G	149	36	0.19
54	r4	DROSHA	chr5:31435937	C>T	113	25	0.18
54	R1	DROSHA	chr5:31422991	T>G	89	84	0.49
54	R1	DROSHA	chr5:31435937	C>T	70	62	0.47
54	R2	DROSHA	chr5:31422991	T>G	48	43	0.47
54	R2	DROSHA	chr5:31435937	C>T	42	47	0.53
54	R3	DROSHA	chr5:31435937	C>T	221	180	0.45
54	r4	GNAS	chr20:57484421	G>A	206	62	0.23
54	R1	SIX1	chr14:61115378	T>C	143	115	0.45
54	R2	SIX1	chr14:61115378	T>C	136	110	0.45
59	R1	TP53	chr17:7574018	G>A	16	66	0.80
59	R3	TP53	chr17:7574018	G>A	21	176	0.89
64	R1	WT1	chr11:32421574	C>A	18	208	0.92
64	R2	WT1	chr11:32421574	C>A	8	404	0.98
64	R3	WT1	chr11:32421574	C>A	22	222	0.91
64	R4	WT1	chr11:32421574	C>A	23	445	0.95
71	R3	CTR9	chr11:10790026	C>T	424	73	0.15
78	R1	CTNNB1	chr3:41266137	C>A	454	150	0.25
78	R5	CTNNB1	chr3:41266137	C>A	235	97	0.29
78	R6	CTNNB1	chr3:41266137	C>A	310	108	0.26
78	R1	DROSHA	chr5:31421465	C>T	289	185	0.39
78	R2	DROSHA	chr5:31421465	C>T	235	131	0.36
78	R3	DROSHA	chr5:31421465	C>T	295	205	0.41
78	R5	DROSHA	chr5:31421465	C>T	102	86	0.46
78	R6	DROSHA	chr5:31421465	C>T	214	146	0.41

continued...

IMPORT	Region	Gene	Contig	Mut	Ref	Alt	VAF
80	R1	HRAS	chr11:532749	C>A	428	4	0.01
80	R4	HRAS	chr11:532749	C>A	620	6	0.01
80	R5	HRAS	chr11:532749	C>A	211	11	0.05
80	R1	TRIM37	chr17:57181701	G>A	205	168	0.45
80	R2	TRIM37	chr17:57181701	G>A	403	213	0.35
80	R3	TRIM37	chr17:57181701	G>A	302	206	0.41
80	R4	TRIM37	chr17:57181701	G>A	207	137	0.40
80	R5	TRIM37	chr17:57181701	G>A	208	139	0.40
106	R2	TP53	chr17:7574003	G>A	8	143	0.95
106	R4	TP53	chr17:7577548	C>T	60	21	0.26
120	R3	RYR2	chr1:237863520	A>G	37	11	0.23
143	R1	CTNNB1	chr3:41266124	A>G	237	191	0.45
143	R2	CTNNB1	chr3:41266133	C>G	238	97	0.29
143	R2	CTNNB1	chr3:41266136	T>C	235	98	0.29
143	R3	CTNNB1	chr3:41266137	C>A	123	225	0.65
143	R4	CTNNB1	chr3:41266124	A>G	137	121	0.47
143	R5	CTNNB1	chr3:41266137	C>A	243	130	0.35
146		CTNNB1	chr3:41266137	C>T	279	179	0.39
146	R1	CTNNB1	chr3:41266137	C>T	309	154	0.33
147		AMER1	chrX:63412095	G>A	149	100	0.40
147	RR1	AMER1	chrX:63412095	G>A	240	197	0.45
147	RR2	AMER1	chrX:63412095	G>A	167	128	0.43
147	RR3	AMER1	chrX:63412095	G>A	172	158	0.48
147	RR4	AMER1	chrX:63412095	G>A	320	8	0.02
147	RR4	CTNNB1	chr3:41274911	T>A	409	173	0.30
170	R1	TP53	chr17:7577121	G>A	13	166	0.93
170	R2	TP53	chr17:7577121	G>A	11	147	0.93
171	LR1	CTNNB1	chr3:41266137	C>G	134	266	0.66
171	LR2	CTNNB1	chr3:41266137	C>G	198	397	0.67
171	LR3	CTNNB1	chr3:41266137	C>G	133	243	0.65
171	LR4	CTNNB1	chr3:41266137	C>G	216	376	0.64
192	R3	CTNNB1	chr3:41266137	C>G	178	124	0.41
233	R1	BCL9	chr1:147084752	C>A	310	3	0.01
233	R2	BCL9	chr1:147084752	C>A	560	61	0.10
233	R1	MLLT1	chr19:6230656	G>C	178	49	0.22
233	R2	MLLT1	chr19:6230656	G>C	250	6	0.02
233	R1	TP53	chr17:7578394	T>A	220	6	0.03
233	R2	TP53	chr17:7578394	T>A	29	153	0.84

Table B.2: This table details all point mutations called in the different regions of the 30 cases sequenced. Ref and Alt shows reads reported by MuTect supporting each allele. Contig reports the location according to hg19.

IMPORT	Region	Gene	Contig	Mut	Ref	Alt	VAF*
4	LR1	DICER1	chr14:95560356	AGAT>A	653	11	0.02
7	RR1	CTNNB1	chr3:41266133	CCTT>C	102	13	0.11
7	RR2	CTNNB1	chr3:41266133	CCTT>C	313	31	0.09
7	RR3	CTNNB1	chr3:41266133	CCTT>C	91	12	0.12
7	RR4	CTNNB1	chr3:41266133	CCTT>C	392	65	0.14
54	R3	TP53	chr17:7574009	TCTC>T	77	107	0.58
64	R1	CTNNB1	chr3:41266133	CCTT>C	205	84	0.29
78	R2	CTNNB1	chr3:41266133	CCTT>C	439	14	0.03
143	R1	WT1	chr11:32450099	TGC>T	12	8	0.40
143	R2	WT1	chr11:32450099	TGC>T	18	9	0.33
143	R3	WT1	chr11:32450099	TGC>T	22	15	0.41
143	R4	WT1	chr11:32450099	TGC>T	31	12	0.28
143	R5	WT1	chr11:32450099	TGC>T	49	6	0.11
146		WT1	chr11:32410609	G>GC	90	92	0.51
146	R1	WT1	chr11:32410609	G>GC	94	61	0.39

Table B.3: This table details all indels called in the different regions of the 30 cases sequenced. Ref and Alt shows reads reported by MuTect2 supporting each allele. Contig reports the location according to hg19. VAF are not used for phylogenetic reconstruction due to filtering of uninformative reads by MuTect2.

Gene	Chr	Start	Ref	Alt	Change	Type
ABL2	1	179112169	C	T		Intron
ALK	2	29443900	G	C		Intron
ALK	2	29443901	G	T		Intron
AMER1	X	63412095	G	A	p.R358*	Nonsense
BCL9	1	147084752	C	A	p.L42M	Missense
CTNNB1	3	41266124	A	G	p.T41A	Missense
CTNNB1	3	41266133	C	G	p.P44A	Missense
CTNNB1	3	41266134	CTT	-	p.S45del	In_Frame
CTNNB1	3	41266136	T	C	p.S45P	Missense
CTNNB1	3	41266137	C	A	p.S45Y	Missense
CTNNB1	3	41266137	C	G	p.S45C	Missense
CTNNB1	3	41266137	C	T	p.S45F	Missense
CTNNB1	3	41268766	A	T	p.K335I	Missense
CTNNB1	3	41274911	T	A	p.N387K	Missense
CTR9	11	10790026	C	T	p.S699S	Silent
CUX1	7	101892227	C	A	p.P1475T	Missense
DICER1	14	95560357	GAT	-	p.I1744del	In_Frame
DROSHA	5	31421465	C	T	p.E1147K	Missense
DROSHA	5	31422991	T	G	p.T1108P	Missense
DROSHA	5	31435937	C	T	p.E993K	Missense
GLI3	7	42063065	A	C		Splice_Site
GNAS	20	57484421	G	A	p.R201H	Missense
HRAS	11	532749	C	A	p.E153*	Nonsense

continued...

Gene	Chr	Start	Ref	Alt	Change	Type
MLLT1	19	6230656	G	C	p.N115K	Missense
NBPF10	1	145325997	A	G	p.Q1290Q	Silent
PIK3C3	18	39575890	C	T	p.R275W	Missense
REST	4	57785960	C	T	p.R302R	Silent
RYR1	19	38964226	G	A	p.A1325A	Silent
RYR2	1	237670117	A	G		Intron
RYR2	1	237863520	A	G		Intron
SALL3	18	76755106	C	A	p.P1039T	Missense
SIX1	14	61115378	T	C	p.Q177R	Missense
TP53	17	7574003	G	A	p.R342*	Nonsense
TP53	17	7574010	CTC	-	p.E339del	In_Frame
TP53	17	7574018	G	A	p.R337C	Missense
TP53	17	7577085	C	T	p.E285K	Missense
TP53	17	7577121	G	A	p.R273C	Missense
TP53	17	7577538	C	T	p.R248Q	Missense
TP53	17	7577548	C	T	p.G245S	Missense
TP53	17	7578394	T	A	p.H179L	Missense
TRIM37	17	57181701	G	A	p.R26C	Missense
Unknown	11	2021676	G	A		IGR
USP9X	X	41026835	G	A		Intron
WT1	11	32410609	-	C	p.L302fs	Frame_Shift
WT1	11	32421574	C	A	p.E128*	Nonsense
WT1	11	32450100	GC	-	p.Q26fs	Frame_Shift

Table B.4: This table details types of mutations called in this study as annotated by Oncotator (http://portals.broadinstitute.org/oncotator/).

IMPORT	Array.Region	Type	ICH_number
3		FK	10627
3	R1	FN	10629
3	R2	FN	10630
3	R3	FN	10632
3	R4	FN	10633
3	R5	FN	10634
3	R6	FN	10635
4	RR1	FN	10888
4	RR2	FN	10890
4		FK	10892
4	LR1	FN	10894
4		FN	12311
7	RR1	FN	10687
7	RR2	FN	10689
7	RR3	FN	10691
7	RR4	FN	10693
7		FK	10695
7	LR1	FN	10697
7	LR2	FN	10699
8	R1	FN	10673
8	R2	FN	10675
8	R3	FN	10677
8	R4	FN	10679
8	R5	FN	10681
8	R6	FN	10683
8		FK	10685
9		FN	11039
9	LR1	FN	11040
9	LR2	FN	11042
9	RR1	FN	11585
9	RR2	FN	11586
9	RR3	FN	11587
9		FK	11588
10	R2	FN	11214
10	R3	FN	11215
10		CF	11206
12	R1	FN	10998
12	R2	FN	10999
12	R3	FN	11000
12	R4	FN	11001
12	R5	FN	11002
12	R6	FN	11003
12		FK	11004

continued...

IMPORT	Array.Region	Type	ICH_number
15		FK	11052
15	R1	FN	11053
15	R2	FN	11054
15	R3	FN	11055
15	R4	FN	11056
16	RR1	FN	10900
16	RR4	FN	10903
16		FK	10905
16	LR1	FN	11143
16	LR2	FN	11144
16		FK	11146
21	RR2	FN	11186
21	RR3	FN	11188
21	LR1	FN	11466
21		FK	11467
21	LR2	FN	11468
30	RR1	FN	11495
30	RR2	FN	11496
30	RR3	FN	11497
30	LR1	FN	11501
30	LR2	FN	11502
30		FK	11503
36	R1	FN	11170
36	R2	FN	11172
36	R3	FN	11174
36	R4	FN	11176
36	R5	FN	11178
36	R6	FN	11180
36		FK	11182
47		CF	11484
47	RR1	FN	11487
47	RR2	FN	11488
47	RR3	FN	11489
47	RR4	FN	11490
47	LR1	FN	11641
54	R1	FN	11409
54	R2	FN	11410
54	r4	FN	11411
54	R3	FN	11413
54		FK	11415
59		CF	11716
59	R3	FN	12180
59		CF	11706
59	R1	FN	11734

continued...

IMPORT	Array.Region	Type	ICH_number
64	R1	FN	11550
64	R2	FN	11551
64	R3	FN	11552
64	R4	FN	11553
64		CF	11864
71	R1	FN	11562
71	R2	FN	11563
71	R3	FN	11564
71	R4	FN	11565
71	R5	FN	11566
71	R6	FN	11567
71		FK	11568
78	LR1	FN	11940
78	LR2	FN	11941
78	LR3	FN	11942
78	LR5	FN	11944
78	LR6	FN	11945
78		FK	11946
80	R1	FN	11851
80	R2	FN	11852
80	R3	FN	11853
80	R4	FN	11854
80	R5	FN	11855
80		FK	11856
85		FN	12219
85	R1	FN	12220
85	R2	FN	12221
85	R3	FN	12222
85	R4	FN	12223
85		CF	12239
106	R1	FN	12203
106	R2	FN	12204
106	R4	FN	12206
106		FK	12202
106	R3	FN	12205
106		CF	13167
120	R1	FN	12129
120	R2	FN	12130
120	R3	FN	12131
120	R4	FN	12132
120	R5	FN	12133
120		FK	12134

continued...

IMPORT	Array.Region	Type	ICH_number
143	R1	FN	12389
143	R2	FN	12390
143	R3	FN	12391
143	R4	FN	12392
143	R5	FN	12393
143		FK	12394
146		FN	12718
146	R1	FN	12720
146		FK	12721
147	RR1	FN	12819
147		FN	12820
147	RR3	FN	12821
147	RR4	FN	12822
147		FK	12823
147	RR2	FN	12824
169	R1	FN	12608
169	R2	FN	12609
169		FN	12610
169		FN	12611
169		CF	12653
170	R1	FN	13328
170	R2	FN	13329
170		CF	13372
171	LR1	FN	12553
171	LR2	FN	12554
171	LR3	FN	12555
171	LR4	FN	12556
171		FK	12557
192	R1	FN	13191
192	R2	FN	13192
192	R3	FN	13193
192	R4	FN	13194
192	R5	FN	13195
192	R6	FN	13196
192		FK	13197
233		CF	14228
233	R1	FN	14239
233	R2	FN	14240

Table B.5: This table details samples sequenced in the 30 cases in PKC66. FN = frozen nephrectomy, FK = frozen kidney and CF = cell fraction (normal tissue from blood). ICH number is an internal sample ID.

ABL2	CUX1	HOXD11	MSH6	SETD5
AGO1	DACT1	HOXD12	MTOR	SH3BP1
AGO2	DCLRE1B	HOXD13	MYCN	SIX1
AGO3	DDX1	HOXD3	NBPF10	SIX2
AGO4	DGCR10	HOXD4	NBPF4	SIX4
AKT2	DGCR6	HOXD8	NBPF6	SIX6
ALK	DGCR8	HOXD9	NIPBL	SMARCA4
AMER1	DGCR9	HRAS	NOTCH1	SMC1A
AR	DICER1	IGF1	NSL1	SMC3
ARID1A	DIS3	IGF2	NTRK2	SMURF1
ARID1B	DIS3L2	IRS2	PAX5	SOX1
ARID3B	DNAJA2	JAK2	PAX6	SOX2
ATRX	DROSHA	KIF7	PHF2	SOX6
BCL11A	ERG	KMT2C	PHLDA1	TAF1
BCL9	EYA1	LIN28A	PIK3C3	TARBP2
BLM	EYS	LIN28B	PKHD1	TBX22
BMP5	FANCG	MAP3K2	PLCH2	TFE3
BRAF	FBXO11	MAPK10	POU3F1	TGFB2
BRCA2	FBXW7	MAPK8	POU4F1	TOP3B
BUB1B	FGF19	MAX	POU6F2	TP53
CDC73	FGFR4	MDM4	PPARA	TRAF2
CDH1	GAREM	MED12	PREX2	TRIM37
CDH4	GAS6	MIRLET7A1	PRODH	U2AF1
CDK20	GATA3	MIRLET7A2	PTF1A	USP9X
CDK6	GATA6	MIRLET7A3	PTPN14	VHL
CDK8	GDF7	MIRLET7B	RAB22A	WBSCR17
CDKN1C	GLI3	MIRLET7C	RAI14	WHSC1
CDKN2B	GNAS	MIRLET7D	REG4	WNT7B
CEP57	GPC3	MIRLET7E	RERE	WT1
CHD1L	GREM2	MIRLET7F1	REST	XPO5
CHD3	H19	MIRLET7F2	RMI2	YEATS2
CHD4	HACE1	MIRLET7G	RUNX2	ZFHX3
CREBBP	HDAC4	MIRLET7I	RYR1	ZFPM2
CRLF1	HFM1	MIS18A	RYR2	
CSNK2A2	HIF1A	MLLT1	SALL1	
CTNNB1	HOXD1	MLLT3	SALL3	
CTR9	HOXD10	MSH5	SETD2	

Table B.6: Genes included on the WT-specific targeted sequencing panel designed using Agilent SureSelect.

Appendix C

CNA manual adjustment documentation in PKC66

Table C.1:

IMPORT	BP difference	Segmentation Correction	Overlapping CNAs	MSAI	CNA remodelled	Curated MEDICC	Manual inference	Tetraploid
3			x	x	x	x	x	
5								
8	x							
10				x				
12				x				
15				x				
16	x				x	x		
26			x		x	x		
30						x		
47						x		
54	x				x	x		
55		x			x	x		
58		x			x	x		
59	x			x	x	x		
64					x	x		
78	x	x		x	x	x		
82		x		x	x	x		
85				x	x	x	x	
91			x	x	x	x		
106		x		x	x	x	x	
108					x	x		x
146					x	x		
147	x			x	x	x		
169		x			x	x		
170				*		x		x
171	x				x	x		
192						x	x	
CCLG-897	x			x		x		

Table C.1: Overview of manual adjustments to PKC66 CNA trees. BP difference represents cases in which a breakpoint difference is taken into account. Segmentation correction shows cases that had manual segmentation alterations. Overlapping CNAs cases had CNAs remodelled as overlapping CNAs in separate clones. MSAI lists cases that contained MSAI in which MSAI was accounted for in the tree. It was not possible to account for MSAI in IMPORT 170 (*) due to tree complexity. CNA remodelled lists cases in which another mixing option was taken for one or more CNAs. Manual inference details cases inferred manually, other cases are considered curated versions of MEDICC output. Tetraploid highlights cases with tetraploid genomes. More details regarding decision making can be found at: www.dropbox.com/sh/95nh8ssic0jx5b6/AAC_Q1U3jeOaoFd-aWbBZc7wa?dl=0

Bibliography

[1] Cancer Research UK. http://www.cancerresearchuk.org/health-professional/cancer-statistics. Accessed: 2017-07-11.

[2] Cristian Tomasetti, Lu Li, and Bert Vogelstein. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 355(6331):1330–1334, Mar 2017.

[3] Gerd P Pfeifer. Environmental exposures and mutational patterns of cancer genomes. *Genome Med.*, 2(8):54, Aug 2010.

[4] D Hanahan and RA Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000.

[5] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, Mar 2011.

[6] Cancer Research UK. http://www.cancerresearchuk.org/health-professional/cancer-statistics/childrens-cancers/mortality. Accessed: 2017-08-17.

[7] Eva Steliarova-Foucher, Murielle Colombet, Lynn A G Ries, Florencia Moreno, Anastasia Dolya, Freddie Bray, Peter Hesseling, Hee Young Shin, Charles A Stiller, and the IICC-3 contributors. International incidence of childhood cancer 2001–10: a population-based registry study. *The Lancet Oncology*, 18(6):719–731, Jun 2017.

[8] Tim Ripperger, Stefan S. Bielack, Arndt Borkhardt, Ines B. Brecht, Birgit Burkhardt, Gabriele Calaminus, Klaus-Michael Debatin, Hedwig Deubzer,

Uta Dirksen, Cornelia Eckert, Angelika Eggert, Miriam Erlacher, Gudrun Fleischhack, Michael C. Frühwald, Astrid Gnekow, Gudrun Goehring, Norbert Graf, Helmut Hanenberg, Julia Hauer, Barbara Hero, Simone Hettmer, Katja von Hoff, Martin Horstmann, Juliane Hoyer, Thomas Illig, Peter Kaatsch, Roland Kappler, Kornelius Kerl, Thomas Klingebiel, Udo Kontny, Uwe Kordes, Dieter Körholz, Ewa Koscielniak, Christof M. Kramm, Michaela Kuhlen, Andreas E. Kulozik, Britta Lamottke, Ivo Leuschner, Dietmar R. Lohmann, Andrea Meinhardt, Markus Metzler, Lüder H. Meyer, Olga Moser, Michaela Nathrath, Charlotte M. Niemeyer, Rainer Nustede, Kristian W. Pajtler, Claudia Paret, Mareike Rasche, Dirk Reinhardt, Olaf Rieß, Alexandra Russo, Stefan Rutkowski, Brigitte Schlegelberger, Dominik Schneider, Reinhard Schneppenheim, Martin Schrappe, Christopher Schroeder, Dietrich von Schweinitz, Thorsten Simon, Monika Sparber-Sauer, Claudia Spix, Martin Stanulla, Doris Steinemann, Brigitte Strahm, Petra Temming, Kathrin Thomay, Andre O. von Bueren, Peter Vorwerk, Olaf Witt, Marcin Wlodarski, Willy Wössmann, Martin Zenker, Stefanie Zimmermann, Stefan M. Pfister, and Christian P. Kratz. Childhood cancer predisposition syndromes-a concise review and recommendations by the cancer predisposition working group of the society for pediatric oncology and hematology. *American Journal of Medical Genetics Part A*, 173(4):1017–1037, Feb 2017.

[9] Gregory T. Armstrong, Yan Chen, Yutaka Yasui, Wendy Leisenring, Todd M. Gibson, Ann C. Mertens, Marilyn Stovall, Kevin C. Oeffinger, Smita Bhatia, Kevin R. Krull, Paul C. Nathan, Joseph P. Neglia, Daniel M. Green, Melissa M. Hudson, and Leslie L. Robison. Reduction in Late Mortality among 5-Year Survivors of Childhood Cancer. *New England Journal of Medicine*, 374(9):833–842, Mar 2016.

[10] Danielle Harake, Vivian I Franco, Jacqueline M Henkel, Tracie L Miller, and Steven E Lipshultz. Cardiotoxicity in childhood cancer survivors: strategies for prevention and management. *Future Cardiology*, 8(4):647–670, Jul 2012.

[11] Gregory T. Armstrong, Kevin C. Oeffinger, Yan Chen, Toana Kawashima, Yutaka Yasui, Wendy Leisenring, Marilyn Stovall, Eric J. Chow, Charles A. Sklar, Daniel A. Mulrooney, Ann C. Mertens, William Border, Jean-Bernard Durand, Leslie L. Robison, and Lillian R. Meacham. Modifiable risk factors and major cardiac events among adult survivors of childhood cancer. *Journal of Clinical Oncology*, 31(29):3673–3680, Oct 2013.

[12] Charles A Stiller. Epidemiology and genetics of childhood cancer. *Oncogene*, 23(38):6429–6444, Aug 2004.

[13] Paul J. Scotting, David A. Walker, and Giorgio Perilongo. Opinion: Childhood solid tumours: a developmental disorder. *Nature Reviews Cancer*, 5(6):481–488, May 2005.

[14] J Brok, TD Treger, SL Gooskens, den Heuvel-Eibrink MM van, and K Pritchard-Jones. Biology and treatment of renal tumours in childhood. *Eur J Cancer*, 68:179–195, Oct 2016.

[15] Vicki Huff. Wilms' tumours: about tumour suppressor genes an oncogene and a chameleon gene. *Nature Reviews Cancer*, 11(2):111–121, Jan 2011.

[16] Raghu Kalluri and Robert A. Weinberg. The basics of epithelial-mesenchymal transition. *Journal of Clinical Investigation*, 119(6):1420–1428, Jun 2009.

[17] Rachel Shukrun, Naomi Pode-Shakked, Oren Pleniceanu, Dorit Omer, Einav Vax, Eyal Peer, Sara Pri-Chen, Jasmine Jacob, Qianghua Hu, Orit Harari-Steinberg, Vicki Huff, and Benjamin Dekel. Wilms' Tumor Blastemal Stem Cells Dedifferentiate to Propagate the Tumor Bulk. *Stem Cell Reports*, 3(1):24–33, Jul 2014.

[18] Seppo Vainio and Yanfeng Lin. Organogenesis: Coordinating early kidney development: lessons from gene targeting. *Nature Reviews Genetics*, 3(7):533–543, Jul 2002.

[19] J. B. Beckwith and N. F. Palmer. Histopathology and prognosis of Wilms tumor: Results from the first national Wilms' tumor study. *Cancer*, 41(5):1937–1948, May 1978.

[20] J. Bruce Beckwith. Precursor lesions of Wilms tumor: Clinical and biological implications. *Medical and Pediatric Oncology*, 21(3):158–168, 1993.

[21] M Steenman, A Westerveld, and M Mannens. Genetics of beckwith-wiedemann syndrome-associated tumors: common genetic pathways. *Genes Chromosomes Cancer*, 28:1–13, May 2000.

[22] Gordan M. Vujanić, Bengt Sandstedt, Dieter Harms, Anna Kelsey, Ivo Leuschner, and Jan de Kraker. Revised International Society of Paediatric Oncology (SIOP) working classification of renal tumors of childhood. *Medical and Pediatric Oncology*, 38(2):79–82, Jan 2002.

[23] Brigitte Royer-Pokora, Manfred Beier, Markus Henzler, Rita Alam, Valrie Schumacher, Angela Weirich, and Vicki Huff. Twenty-four new cases of WT1 germline mutations and review of the literature: Genotype/phenotype correlations for Wilms tumor development. *American Journal of Medical Genetics*, 127A(3):249–257, 2004.

[24] Johannes Hans M. Merks, Huib N. Caron, and Raoul C.M. Hennekam. High incidence of malformation syndromes in a series of 1,073 children with cancer. *American Journal of Medical Genetics Part A*, 134A(2):132–143, 2005.

[25] Norman Breslow, Jane Olson, Jami Moksness, J. Bruce Beckwith, and Paul Grundy. Familial Wilms' tumor: A descriptive study. *Medical and Pediatric Oncology*, 27(5):398–403, Nov 1996.

[26] Logan G. Spector and Jill Birch. The Epidemiology of Hepatoblastoma. *Pediatric Blood & Cancer*, 59(5):776–779, Jun 2012.

[27] C. E. Herzog. Childhood Cancers: Hepatoblastoma. *The Oncologist*, 5(6):445–453, Dec 2000.

[28] Talita Ferreira Marques Aguiar, Thaise Nayane Carneiro, Cecilia Maria Lima da Costa, Carla Rosenberg, Isabela Werneck da Cunha, and Ana Cristina Victorino Krepischi. The genetic and epigenetic landscapes of hepatoblastomas. *Applied Cancer Research*, 37(1):20, May 2017.

[29] Melanie Eichenmüller, Franziska Trippel, Michaela Kreuder, Alexander Beck, Thomas Schwarzmayr, Beate Häberle, Stefano Cairo, Ivo Leuschner, Dietrich von Schweinitz, Tim M. Strom, and Roland Kappler. The genomic landscape of hepatoblastoma and their progenies with HCC-like features. *Journal of Hepatology*, 61(6):1312–1320, Dec 2014.

[30] Rebecka L Meyers, Rudolf Maibach, Eiso Hiyama, Beate Häberle, Mark Krailo, Arun Rangaswami, Daniel C Aronson, Marcio H Malogolowkin, Giorgio Perilongo, Dietrich von Schweinitz, Marc Ansari, Dolores Lopez-Terrada, Yukichi Tanaka, Rita Alaggio, Ivo Leuschner, Tomoro Hishiki, Irene Schmid, Kenichiro Watanabe, Kenichi Yoshimura, Yurong Feng, Eugenia Rinaldi, Davide Saraceno, Marisa Derosa, and Piotr Czauderna. Risk-stratified staging in paediatric hepatoblastoma: a unified analysis from the children's hepatic tumors international collaboration. *The Lancet Oncology*, 18(1):122–131, Jan 2017.

[31] Piotr Czauderna, Dolores Lopez-Terrada, Eiso Hiyama, Beate Häberle, Marcio H. Malogolowkin, and Rebecka L. Meyers. Hepatoblastoma state of the art. *Current Opinion in Pediatrics*, 26(1):19–28, Feb 2014.

[32] Helen Davies, Graham R. Bignell, Charles Cox, Philip Stephens, Sarah Edkins, Sheila Clegg, Jon Teague, Hayley Woffendin, Mathew J. Garnett, William Bottomley, Neil Davis, Ed Dicks, Rebecca Ewing, Yvonne Floyd, Kristian Gray, Sarah Hall, Rachel Hawes, Jaime Hughes, Vivian Kosmidou, Andrew Menzies, Catherine Mould, Adrian Parker, Claire Stevens, Stephen Watt, Steven Hooper, Rebecca Wilson, Hiran Jayatilake, Barry A. Gusterson, Colin Cooper, Janet Shipley, Darren Hargrave, Katherine Pritchard-Jones, Norman Maitland, Georgia Chenevix-Trench, Gregory J. Riggins, Darell D.

Bigner, Giuseppe Palmieri, Antonio Cossu, Adrienne Flanagan, Andrew Nicholson, Judy W. C. Ho, Suet Y. Leung, Siu T. Yuen, Barbara L. Weber, Hilliard F. Seigler, Timothy L. Darrow, Hugh Paterson, Richard Marais, Christopher J. Marshall, Richard Wooster, Michael R. Stratton, and P. Andrew Futreal. Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949–954, Jun 2002.

[33] Robert A Weinberg. The retinoblastoma protein and cell cycle control. *Cell*, 81(3):323–330, May 1995.

[34] E. Cristy Ruteshouser, Stephen M. Robinson, and Vicki Huff. Wilms tumor genetics: Mutations in WT1,WTX and CTNNB1 account for only about one-third of tumors. *Genes Chromosomes and Cancer*, 47(6):461–470, 2008.

[35] M. B. Major, N. D. Camp, J. D. Berndt, X. Yi, S. J. Goldenberg, C. Hubbert, T. L. Biechele, A.-C. Gingras, N. Zheng, M. J. MacCoss, S. Angers, and R. T. Moon. Wilms Tumor Suppressor WTX Negatively Regulates WNT/β-Catenin Signaling. *Science*, 316(5827):1043–1046, May 2007.

[36] Jamie N. Anastas and Randall T. Moon. WNT signalling pathways as therapeutic targets in cancer. *Nature Reviews Cancer*, 13(1):11–26, Dec 2012.

[37] Arnold J Levine. p53 the Cellular Gatekeeper for Growth and Division. *Cell*, 88(3):323–331, Feb 1997.

[38] Nabeel Bardeesy, David Falkoff, Mary-Jane Petruzzi, Norma Nowak, Bernhard Zabel, Mohammed Adam, Maria C. Aguiar, Paul Grundy, Tom Shows, and Jerry Pelletier. Anaplastic Wilms' tumour a subtype displaying poor prognosis, harbours p53 gene mutations. *Nature Genetics*, 7(1):91–97, May 1994.

[39] Mariana Maschietto, Richard D. Williams, Tasnim Chagtai, Sergey D. Popov, Neil J. Sebire, Gordan Vujanic, Elizabeth Perlman, James R. Anderson, Paul

Grundy, Jeffrey S. Dome, and Kathy Pritchard-Jones. TP53 Mutational Status Is a Potential Marker for Risk Stratification in Wilms Tumour with Diffuse Anaplasia. *PLoS ONE*, 9(10):e109924, Oct 2014.

[40] Jenny Wegert, Naveed Ishaque, Romina Vardapour, Christina Geörg, Zuguang Gu, Matthias Bieg, Barbara Ziegler, Sabrina Bausenwein, Nasenien Nourkami, Nicole Ludwig, Andreas Keller, Clemens Grimm, Susanne Kneitz, Richard D. Williams, Tas Chagtai, Kathy Pritchard-Jones, Peter van Sluis, Richard Volckmann, Jan Koster, Rogier Versteeg, Tomas Acha, Maureen J. O'Sullivan, Peter K. Bode, Felix Niggli, Godelieve A. Tytgat, Harm van Tinteren, Marry M. van den Heuvel-Eibrink, Eckart Meese, Christian Vokuhl, Ivo Leuschner, Norbert Graf, Roland Eils, Stefan M. Pfister, Marcel Kool, and Manfred Gessler. Mutations in the SIX1/2 Pathway and the DROSHA/DGCR8 miRNA Microprocessor Complex Underlie High-Risk Blastemal Type Wilms Tumors. *Cancer Cell*, 27(2):298–311, Feb 2015.

[41] Amy L. Walz, Ariadne Ooms, Samantha Gadd, Daniela S. Gerhard, Malcolm A. Smith, Jaime M. Guidry Auvil, Daoud Meerzaman, Qing-Rong Chen, Chih Hao Hsu, Chunhua Yan, Cu Nguyen, Ying Hu, Reanne Bowlby, Denise Brooks, Yussanne Ma, Andrew J. Mungall, Richard A. Moore, Jacqueline Schein, Marco A. Marra, Vicki Huff, Jeffrey S. Dome, Yueh-Yun Chi, Charles G. Mullighan, Jing Ma, David A. Wheeler, Oliver A. Hampton, Nadereh Jafari, Nicole Ross, Julie M. Gastier-Foster, and Elizabeth J. Perlman. Recurrent DGCR8, DROSHA and SIX Homeodomain Mutations in Favorable Histology Wilms Tumors. *Cancer Cell*, 27(3):426, Mar 2015.

[42] Shuibin Lin and Richard I. Gregory. MicroRNA biogenesis pathways in cancer. *Nature Reviews Cancer*, 15(6):321–333, May 2015.

[43] Elizabeth J. Perlman, Samantha Gadd, Stefan T. Arold, Anand Radhakrishnan, Daniela S. Gerhard, Lawrence Jennings, Vicki Huff, Jaime M. Guidry Auvil, Tanja M. Davidsen, Jeffrey S. Dome, Daoud Meerzaman, Chih Hao Hsu, Cu Nguyen, James Anderson, Yussanne Ma, Andrew J. Mungall,

Richard A. Moore, Marco A. Marra, Charles G. Mullighan, Jing Ma, David A. Wheeler, Oliver A. Hampton, Julie M. Gastier-Foster, Nicole Ross, and Malcolm A. Smith. MLLT1 YEATS domain mutations in clinically distinctive favourable histology wilms tumours. *Nature Communications*, 6:10013, Dec 2015.

[44] G. Brodeur, R. Seeger, M Schwab, H. Varmus, and J. Bishop. Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science*, 224(4653):1121–1124, Jun 1984.

[45] Keiko Taguchi, Hozumi Motohashi, and Masayuki Yamamoto. Molecular mechanisms of the Keap1-Nrf2 pathway in stress response and cancer evolution. *Genes to Cells*, 16(2):123–140, Jan 2011.

[46] Prasad V. Jallepalli and Christoph Lengauer. Chromosome segregation and cancer: cutting through the mystery. *Nature Reviews Cancer*, 1(2):109–117, Nov 2001.

[47] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavare, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis. Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. *Science*, 315(5813):848–853, Feb 2007.

[48] Georgina L. Ryland, Maria A. Doyle, David Goode, Samantha E. Boyle, David Y.H. Choong, Simone M. Rowley, Jason Li, David DL Bowtell, Richard W. Tothill, Ian G. Campbell, and Kylie L. Gorringe. Loss of heterozygosity: what is it good for? *BMC Medical Genomics*, 8(1):45, Aug 2015.

[49] Gerard I. Evan, Andrew H. Wyllie, Christopher S. Gilbert, Trevor D. Littlewood, Hartmut Land, Mary Brooks, Catherine M. Waters, Linda Z. Penn, and David C. Hancock. Induction of apoptosis in fibroblasts by c-myc protein. *Cell*, 69(1):119–128, Apr 1992.

[50] M. S. Esposito. Evidence that spontaneous mitotic recombination occurs at the two-strand stage. *Proceedings of the National Academy of Sciences*, 75(9):4436–4440, Sep 1978.

[51] P. J. Hastings, James R. Lupski, Susan M. Rosenberg, and Grzegorz Ira. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564, Aug 2009.

[52] J. Scott, J. Cowell, M. E. Robertson, L. M. Priestley, R. Wadey, B. Hopkins, J. Pritchard, G. I. Bell, L. B. Rall, C. F. Graham, and T. J. Knott. Insulin-like growth factor-II gene expression in Wilms' tumour and embryonic tissues. *Nature*, 317(6034):260–262, Sep 1985.

[53] Osamu Ogawa, Michael R. Eccles, Jenny Szeto, Leslie A. McNoe, Kankatsu Yun, Marion A. Maw, Peter J. Smith, and Anthony E. Reeve. Relaxation of insulin-like growth factor II gene imprinting implicated in Wilms' tumour. *Nature*, 362(6422):749–751, Apr 1993.

[54] Richard D. Williams, Reem Al-Saadi, Rachael Natrajan, Alan Mackay, Tasnim Chagtai, Suzanne Little, Sandra N. Hing, Kerry Fenwick, Alan Ashworth, Paul Grundy, James R. Anderson, Jeffrey S. Dome, Elizabeth J. Perlman, Chris Jones, and Kathy Pritchard-Jones. Molecular profiling reveals frequent gain of MYCN and anaplasia-specific loss of 4q and 14q in wilms tumor. *Genes Chromosomes and Cancer*, 50(12):982–995, Aug 2011.

[55] Paul E. Grundy, Norman E. Breslow, Sierra Li, Elizabeth Perlman, J. Bruce Beckwith, Michael L. Ritchey, Robert C. Shamberger, Gerald M. Haase, Giulio J. D'Angio, Milton Donaldson, Max J. Coppes, Marcio Malogolowkin, Patricia Shearer, Patrick R.M. Thomas, Roger Macklis, Gail Tomlinson, Vicki Huff, and Daniel M. Green. Loss of Heterozygosity for Chromosomes 1p and 16q Is an Adverse Prognostic Factor in Favorable-Histology Wilms Tumor: A Report From the National Wilms Tumor Study Group. *Journal of Clinical Oncology*, 23(29):7312–7321, Aug 2005.

[56] R Natrajan, RD Williams, SN Hing, A Mackay, JS Reis-Filho, K Fenwick, M Iravani, H Valgeirsson, A Grigoriadis, CF Langford, O Dovey, SG Gregory, BL Weber, A Ashworth, PE Grundy, K Pritchard-Jones, and C Jones. Array CGH profiling of favourable histology Wilms tumours reveals novel gains and losses associated with relapse. *The Journal of Pathology*, 210(1):49–58, 2006.

[57] Eric J. Gratias, Lawrence J. Jennings, James R. Anderson, Jeffrey S. Dome, Paul Grundy, and Elizabeth J. Perlman. Gain of 1q is associated with inferior event-free and overall survival in patients with favorable histology Wilms tumor: A report from the Children's Oncology Group. *Cancer*, 119(21):3887–3894, Aug 2013.

[58] H. Segers, M. M. van den Heuvel-Eibrink, R. D. Williams, H. van Tinteren, G. Vujanic, R. Pieters, K. Pritchard-Jones, and N. Bown. Gain of 1q is a marker of poor prognosis in Wilms' tumors. *Genes Chromosomes and Cancer*, 52(11):1065–1074, Sep 2013.

[59] Teresa Davoli, Hajime Uno, Eric C. Wooten, and Stephen J. Elledge. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science*, 355(6322):eaaf8399, Jan 2017.

[60] Gail E. Tomlinson and Roland Kappler. Genetics and epigenetics of hepatoblastoma. *Pediatric Blood & Cancer*, 59(5):785–792, Jul 2012.

[61] Pavel Sumazin, Yidong Chen, Lisa R. Treviño, Stephen F. Sarabia, Oliver A. Hampton, Kayuri Patel, Toni-Ann Mistretta, Barry Zorman, Patrick Thompson, Andras Heczey, Sarah Comerford, David A. Wheeler, Murali Chintagumpala, Rebecka Meyers, Dinesh Rakheja, Milton J. Finegold, Gail Tomlinson, D. Williams Parsons, and Dolores López-Terrada. Genomic analysis of hepatoblastoma identifies distinct molecular and prognostic subgroups. *Hepatology*, 65(1):104–121, Nov 2016.

[62] Qinghua Shi and Randall W. King. Chromosome nondisjunction yields tetraploid rather than aneuploid cells in human cell lines. *Nature*, 437(4061):1038–1042, Oct 2005.

[63] Travis I Zack, Steven E Schumacher, Scott L Carter, Andrew D Cherniack, Gordon Saksena, Barbara Tabak, Michael S Lawrence, Cheng-Zhong Zhang, Jeremiah Wala, Craig H Mermel, Carrie Sougnez, Stacey B Gabriel, Bryan Hernandez, Hui Shen, Peter W Laird, Gad Getz, Matthew Meyerson, and Rameen Beroukhim. Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45(10):1134–1140, Sep 2013.

[64] A Madlung. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity*, 110(2):99–104, Nov 2012.

[65] EC de Bruin, N McGranahan, R Mitter, M Salm, DC Wedge, L Yates, M Jamal-Hanjani, S Shafi, N Murugaesu, AJ Rowan, E Grönroos, MA Muhammad, S Horswell, M Gerlinger, I Varela, D Jones, J Marshall, T Voet, P van Loo, DM Rassl, RC Rintoul, SM Janes, SM Lee, M Forster, T Ahmad, D Lawrence, M Falzon, A Capitanio, TT Harkins, CC Lee, W Tom, E Teefe, SC Chen, S Begum, A Rabinowitz, B Phillimore, B Spencer-Dene, G Stamp, Z Szallasi, N Matthews, A Stewart, P Campbell, and C Swanton. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, 346:251–6, Oct 2014.

[66] Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez, Thomas J. Mitchell, Yulia Rubanova, Pavana Anur, Daniel Rosebrock, Kaixan Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kortine Kleinheinz, Ignacio Vazquez-Garcia, Kerstin Haase, Subhajit Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G. Livitz, Marek Cmero, Jonas Demeulemeester, Steve Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Paul C. Boutros, David D. Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhim, S. Cenk Sahinalp, Yuan Ji, Martin Peifer, Florian Markowetz, Ville Musto-

nen, Ke Yuan, Wenyi Wang, Quaid D. Morris, Paul T. Spellman, David C. Wedge, and Peter Van Loo. The evolutionary history of 2,658 cancers. *bioRxiv*, Jul 2017.

[67] Fang Chen, Zhong-Chu Li, Ren-Quan Ge, Guo-Hua Liu, Wei Cai, and Zhong-De Zhang. The measurement of DNA content in Wilms' tumor and its clinical significance. *Journal of Pediatric Surgery*, 29(4):548–550, Apr 1994.

[68] Kenneth W. Gow and James J. Murphy. Cytogenetic and histologic findings in Wilms' tumor. *Journal of Pediatric Surgery*, 37(6):823–827, Jun 2002.

[69] Z. Storchova and C. Kuffer. The consequences of tetraploidy and aneuploidy. *Journal of Cell Science*, 121(23):3859–3866, Nov 2008.

[70] Philip J. Stephens, Chris D. Greenman, Beiyuan Fu, Fengtang Yang, Graham R. Bignell, Laura J. Mudie, Erin D. Pleasance, King Wai Lau, David Beare, Lucy A. Stebbings, Stuart McLaren, Meng-Lay Lin, David J. McBride, Ignacio Varela, Serena Nik-Zainal, Catherine Leroy, Mingming Jia, Andrew Menzies, Adam P. Butler, Jon W. Teague, Michael A. Quail, John Burton, Harold Swerdlow, Nigel P. Carter, Laura A. Morsberger, Christine Iacobuzio-Donahue, George A. Follows, Anthony R. Green, Adrienne M. Flanagan, Michael R. Stratton, P. Andrew Futreal, and Peter J. Campbell. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell*, 144(1):27–40, Jan 2011.

[71] Tobias Rausch, David T.W. Jones, Marc Zapatka, Adrian M. Stütz, Thomas Zichner, Joachim Weischenfeldt, Natalie Jäger, Marc Remke, David Shih, Paul A. Northcott, Elke Pfaff, Jelena Tica, Qi Wang, Luca Massimi, Hendrik Witt, Sebastian Bender, Sabrina Pleier, Huriye Cin, Cynthia Hawkins, Christian Beck, Andreas von Deimling, Volkmar Hans, Benedikt Brors, Roland Eils, Wolfram Scheurlen, Jonathon Blake, Vladimir Benes, Andreas E. Kulozik, Olaf Witt, Dianna Martin, Cindy Zhang, Rinnat Porat,

Diana M. Merino, Jonathan Wasserman, Nada Jabado, Adam Fontebasso, Lars Bullinger, Frank G. Rücker, Konstanze Döhner, Hartmut Döhner, Jan Koster, Jan J. Molenaar, Rogier Versteeg, Marcel Kool, Uri Tabori, David Malkin, Andrey Korshunov, Michael D. Taylor, Peter Lichter, Stefan M. Pfister, and Jan O. Korbel. Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. *Cell*, 148(1-2):59–71, Jan 2012.

[72] X Chen, A Pappo, and M A Dyer. Pediatric solid tumor genomics and developmental pliancy. *Oncogene*, 34(41):5207–5215, Feb 2015.

[73] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer Genome Landscapes. *Science*, 339(6127):1546–1558, Mar 2013.

[74] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinsk, Natalie Jäger, David T. W. Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R. Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C. Munshi, Hiromi Nakamura, Paul A. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N. J. Tutt, Rafael Valdés-Mas, Marit M. van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R. Yates, Jessica Zucman-Rossi, P. Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M. Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M. Pfister, Peter J. Campbell, and Michael R. Stratton.

Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, Aug 2013.

[75] P. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, Oct 1976.

[76] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, Apr 2009.

[77] K. Harbst, M. Lauss, H. Cirenajwis, K. Isaksson, F. Rosengren, T. Torngren, A. Kvist, M. C. Johansson, J. Vallon-Christersson, B. Baldetorp, A. Borg, H. Olsson, C. Ingvar, A. Carneiro, and G. Jönsson. Multiregion Whole-Exome Sequencing Uncovers the Genetic Evolution and Mutational Heterogeneity of Early-Stage Metastatic Melanoma. *Cancer Research*, 76(16):4765–4774, May 2016.

[78] Kamil A. Lipinski, Louise J. Barber, Matthew N. Davies, Matthew Ashenden, Andrea Sottoriva, and Marco Gerlinger. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in Cancer*, 2(1):49–63, Jan 2016.

[79] N. J. Birkbak, A. C. Eklund, Q. Li, S. E. McClelland, D. Endesfelder, P. Tan, I. B. Tan, A. L. Richardson, Z. Szallasi, and C. Swanton. Paradoxical Relationship between Chromosomal Instability and Survival Outcome in Cancer. *Cancer Research*, 71(10):3447–3452, Jan 2011.

[80] Andrea Sottoriva, Chris P Barnes, and Trevor A Graham. Catch my drift? Making sense of genomic intra-tumour heterogeneity. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1867(2):95–100, Apr 2017.

[81] Sasha F. Levy, Jamie R. Blundell, Sandeep Venkataram, Dmitri A. Petrov, Daniel S. Fisher, and Gavin Sherlock. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, 519(7542):181–186, Feb 2015.

[82] Iñigo Martincorena, Aswin S. N. Seshasayee, and Nicholas M. Luscombe. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature*, 485(7396):95–98, Apr 2012.

[83] Andrea Sottoriva, Haeyoun Kang, Zhicheng Ma, Trevor A Graham, Matthew P Salomon, Junsong Zhao, Paul Marjoram, Kimberly Siegmund, Michael F Press, Darryl Shibata, and Christina Curtis. A Big Bang model of human colorectal tumor growth. *Nature Genetics*, 47(3):209–216, Feb 2015.

[84] Douglas Hanahan and Lisa M. Coussens. Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. *Cancer Cell*, 21(3):309–322, Mar 2012.

[85] Mel Greaves and Carlo C. Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, Jan 2012.

[86] S. Turajlic and C. Swanton. Metastasis as an evolutionary process. *Science*, 352(6282):169–175, Apr 2016.

[87] Marc J Williams, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48(3):238–244, Jan 2016.

[88] Nicholas McGranahan and Charles Swanton. Clonal Heterogeneity and Tumor Evolution: Past Present, and the Future. *Cell*, 168(4):613–628, Feb 2017.

[89] Allison S. Cleary, Travis L. Leonard, Shelley A. Gestl, and Edward J. Gunther. Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. *Nature*, 508(7494):113–117, Apr 2014.

[90] Andriy Marusyk, Doris P. Tabassum, Philipp M. Altrock, Vanessa Almendro, Franziska Michor, and Kornelia Polyak. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature*, 514(7520):54–58, Jul 2014.

[91] Eric R. Fearon and Bert Vogelstein. A genetic model for colorectal tumori-genesis. *Cell*, 61(5):759–767, Jun 1990.

[92] William CH Cross, Trevor A Graham, and Nicholas A Wright. New paradigms in clonal evolution: punctuated equilibrium in cancer. *The Journal of Pathology*, 240(2):126–136, Aug 2016.

[93] M. Greaves. Evolutionary Determinants of Cancer. *Cancer Discovery*, 5(8):806–820, Jul 2015.

[94] Mariam Jamal-Hanjani, Gareth A. Wilson, Nicholas McGranahan, Nicolai J. Birkbak, Thomas B.K. Watkins, Selvaraju Veeriah, Seema Shafi, Diana H. Johnson, Richard Mitter, Rachel Rosenthal, Max Salm, Stuart Horswell, Mickael Escudero, Nik Matthews, Andrew Rowan, Tim Chambers, David A. Moore, Samra Turajlic, Hang Xu, Siow-Ming Lee, Martin D. Forster, Tanya Ahmad, Crispin T. Hiley, Christopher Abbosh, Mary Falzon, Elaine Borg, Teresa Marafioti, David Lawrence, Martin Hayward, Shyam Kolvekar, Niko-laos Panagiotopoulos, Sam M. Janes, Ricky Thakrar, Asia Ahmed, Fiona Blackhall, Yvonne Summers, Rajesh Shah, Leena Joseph, Anne M. Quinn, Phil A. Crosbie, Babu Naidu, Gary Middleton, Gerald Langman, Simon Trot-ter, Marianne Nicolson, Hardy Remmen, Keith Kerr, Mahendran Chetty, Les-ley Gomersall, Dean A. Fennell, Apostolos Nakas, Sridhar Rathinam, Girija Anand, Sajid Khan, Peter Russell, Veni Ezhil, Babikir Ismail, Melanie Irvin-Sellers, Vineet Prakash, Jason F. Lester, Malgorzata Kornaszewska, Richard Attanoos, Haydn Adams, Helen Davies, Stefan Dentro, Philippe Taniere, Brendan O'Sullivan, Helen L. Lowe, John A. Hartley, Natasha Iles, Har-riet Bell, Yenting Ngai, Jacqui A. Shaw, Javier Herrero, Zoltan Szallasi, Roland F. Schwarz, Aengus Stewart, Sergio A. Quezada, John Le Quesne, Peter Van Loo, Caroline Dive, Allan Hackshaw, and Charles Swanton. Track-ing the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, 376(22):2109–2121, Jun 2017.

[95] Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Rebeca Sanz-Pamplona, Luca De Sano, Giancarlo Mauri, Victor Moreno, Marco Antoniotti, and Bud Mishra. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences*, 113(28):E4025–E4034, Jun 2016.

[96] Michael W. Schmitt, Lawrence A. Loeb, and Jesse J. Salk. The influence of subclonal resistance mutations on targeted cancer therapy. *Nature Reviews Clinical Oncology*, 13(6):335–347, Oct 2015.

[97] Nicholas McGranahan, Rachel Rosenthal, Crispin T. Hiley, Andrew J. Rowan, Thomas B.K. Watkins, Gareth A. Wilson, Nicolai J. Birkbak, Selvaraju Veeriah, Peter Van Loo, Javier Herrero, Charles Swanton, Charles Swanton, Mariam Jamal-Hanjani, Selvaraju Veeriah, Seema Shafi, Justyna Czyzewska-Khan, Diana Johnson, Joanne Laycock, Leticia Bosshard-Carter, Rachel Rosenthal, Pat Gorman, Robert E. Hynds, Gareth Wilson, Nicolai J. Birkbak, Thomas B.K. Watkins, Nicholas McGranahan, Stuart Horswell, Richard Mitter, Mickael Escudero, Aengus Stewart, Peter Van Loo, Andrew Rowan, Hang Xu, Samra Turajlic, Crispin Hiley, Christopher Abbosh, Jacki Goldman, Richard Kevin Stone, Tamara Denner, Nik Matthews, Greg Elgar, Sophia Ward, Marta Costa, Sharmin Begum, Ben Phillimore, Tim Chambers, Emma Nye, Sofia Graca, Maise Al Bakir, Kroopa Joshi, Andrew Furness, Assma Ben Aissa, Yien Ning Sophia Wong, Andy Georgiou, Sergio Quezada, John A. Hartley, Helen L. Lowe, Javier Herrero, David Lawrence, Martin Hayward, Nikolaos Panagiotopoulos, Shyam Kolvekar, Mary Falzon, Elaine Borg, Teresa Marafioti, Celia Simeon, Gemma Hector, Amy Smith, Marie Aranda, Marco Novelli, Dahmane Oukrif, Sam M. Janes, Ricky Thakrar, Martin Forster, Tanya Ahmad, Siow Ming Lee, Dionysis Papadatos-Pastos, Dawn Carnell, Ruheena Mendes, Jeremy George, Neal Navani, Asia Ahmed, Magali Taylor, Junaid Choudhary, Yvonne Summers, Raffaele Califano, Paul Taylor, Rajesh Shah, Piotr Krysiak, Kendadai Rammohan, Eustace Fontaine, Richard Booton, Matthew Evison, Phil Crosbie,

Stuart Moss, Faiza Idries, Leena Joseph, Paul Bishop, Anshuman Chaturved, Anne Marie Quinn, Helen Doran, Angela Leek, Phil Harrison, Katrina Moore, Rachael Waddington, Juliette Novasio, Fiona Blackhall, Jane Rogan, Elaine Smith, Caroline Dive, Jonathan Tugwood, Ged Brady, Dominic G. Rothwell, Francesca Chemi, Jackie Pierce, Sakshi Gulati, Babu Naidu, Gerald Langman, Simon Trotter, Mary Bellamy, Hollie Bancroft, Amy Kerr, Salma Kadiri, Joanne Webb, Gary Middleton, Madava Djearaman, Dean Fennell, Jacqui A. Shaw, John Le Quesne, David Moore, Apostolos Nakas, Sridhar Rathinam, William Monteiro, Hilary Marshall, Louise Nelson, Jonathan Bennett, Joan Riley, Lindsay Primrose, Luke Martinson, Girija Anand, Sajid Khan, Anita Amadi, Marianne Nicolson, Keith Kerr, Shirley Palmer, Hardy Remmen, Joy Miller, Keith Buchan, Mahendran Chetty, Lesley Gomersall, Jason Lester, Alison Edwards, Fiona Morgan, Haydn Adams, Helen Davies, Malgorzata Kornaszewska, Richard Attanoos, Sara Lock, Azmina Verjee, Mairead MacKenzie, Maggie Wilcox, Harriet Bell, Allan Hackshaw, Yenting Ngai, Sean Smith, Nicole Gower, Christian Ottensmeier, Serena Chee, Benjamin Johnson, Aiman Alzetani, Emily Shaw, Eric Lim, Paulo De Sousa, Monica Tavares Barbosa, Alex Bowman, Simon Jordan, Alexandra Rice, Hilgardt Raubenheimer, Chiara Proli, Maria Elena Cufari, John Carlo Ronquillo, Angela Kwayie, Harshil Bhayani, Morag Hamilton, Yusura Bakar, Natalie Mensah, Lyn Ambrose, Anand Devaraj, Silviu Buderi, Jonathan Finch, Leire Azcarate, Hema Chavan, Sophie Green, Hillaria Mashinga, Andrew G. Nicholson, Kelvin Lau, Michael Sheaff, Peter Schmid, John Conibear, Veni Ezhil, Babikir Ismail, Melanie Irvin-sellers, Vineet Prakash, Peter Russell, Teresa Light, Tracey Horey, Sarah Danson, Jonathan Bury, John Edwards, Jennifer Hill, Sue Matthews, Yota Kitsanta, Kim Suvarna, Patricia Fisher, Allah Dino Keerio, Michael Shackcloth, John Gosney, Pieter Postmus, Sarah Feeney, Julius Asante-Siaw, Hugo J.W.L. Aerts, Stefan Dentro, and Christophe Dessimoz. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*, 171(6):1259–1271.e11,

Nov 2017.

[98] Christopher Abbosh, Nicolai J. Birkbak, Gareth A. Wilson, Mariam Jamal-Hanjani, Tudor Constantin, Raheleh Salari, John Le Quesne, David A. Moore, Selvaraju Veeriah, Rachel Rosenthal, Teresa Marafioti, Eser Kirk-izlar, Thomas B. K. Watkins, Nicholas McGranahan, Sophia Ward, Luke Martinson, Joan Riley, Francesco Fraioli, Maise Al Bakir, Eva Grönroos, Francisco Zambrana, Raymondo Endozo, Wenya Linda Bi, Fiona M. Fennessy, Nicole Sponer, Diana Johnson, Joanne Laycock, Seema Shafi, Justyna Czyzewska-Khan, Andrew Rowan, Tim Chambers, Nik Matthews, Samra Turajlic, Crispin Hiley, Siow Ming Lee, Martin D. Forster, Tanya Ahmad, Mary Falzon, Elaine Borg, David Lawrence, Martin Hayward, Shyam Kolvekar, Nikolaos Panagiotopoulos, Sam M. Janes, Ricky Thakrar, Asia Ahmed, Fiona Blackhall, Yvonne Summers, Dina Hafez, Ashwini Naik, Apratim Ganguly, Stephanie Kareht, Rajesh Shah, Leena Joseph, Anne Marie Quinn, Phil A. Crosbie, Babu Naidu, Gary Middleton, Gerald Langman, Simon Trotter, Marianne Nicolson, Hardy Remmen, Keith Kerr, Mahendran Chetty, Lesley Gomersall, Dean A. Fennell, Apostolos Nakas, Sridhar Rathinam, Girija Anand, Sajid Khan, Peter Russell, Veni Ezhil, Babikir Ismail, Melanie Irvin-Sellers, Vineet Prakash, Jason F. Lester, Malgorzata Kornaszewska, Richard Attanoos, Haydn Adams, Helen Davies, Dahmane Oukrif, Ayse U. Akarca, John A. Hartley, Helen L. Lowe, Sara Lock, Natasha Iles, Harriet Bell, Yenting Ngai, Greg Elgar, Zoltan Szallasi, Roland F. Schwarz, Javier Herrero, Aengus Stewart, Sergio A. Quezada, Karl S. Peggs, Peter Van Loo, Caroline Dive, C. Jimmy Lin, Matthew Rabinowitz, Hugo J. W. L. Aerts, Allan Hackshaw, Jacqui A. Shaw, Bernhard G. Zimmermann, The TRACERx consortium, The PEACE consortium, and Charles Swanton. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature*, 545(7655):446–451, Apr 2017.

[99] Kamila Naxerova and Rakesh K. Jain. Using tumour phylogenetics to identify the roots of metastasis in humans. *Nature Reviews Clinical Oncology*,

12(5):258–272, Jan 2015.

[100] Christopher A. Miller, Yevgeniy Gindin, Charles Lu, Obi L Griffith, Malachi Griffith, Dong Shen, Jeremy Hoog, Tiandao Li, David E. Larson, Mark Watson, Sherri R Davies, Kelly Hunt, Vera J. Suman, Jacqueline Snider, Thomas Walsh, Graham A. Colditz, Katherine DeSchryver, Richard K. Wilson, Elaine R. Mardis, and Matthew J. Ellis. Aromatase inhibition remodels the clonal architecture of estrogen-receptor-positive breast cancers. *Nature Communications*, 7:12498, Aug 2016.

[101] Peter Eirew, Adi Steif, Jaswinder Khattra, Gavin Ha, Damian Yap, Hossein Farahani, Karen Gelmon, Stephen Chia, Colin Mar, Adrian Wan, Emma Laks, Justina Biele, Karey Shumansky, Jamie Rosner, Andrew McPherson, Cydney Nielsen, Andrew J. L. Roth, Calvin Lefebvre, Ali Bashashati, Camila de Souza, Celia Siu, Radhouane Aniba, Jazmine Brimhall, Arusha Oloumi, Tomo Osako, Alejandra Bruna, Jose L. Sandoval, Teresa Algara, Wendy Greenwood, Kaston Leung, Hongwei Cheng, Hui Xue, Yuzhuo Wang, Dong Lin, Andrew J. Mungall, Richard Moore, Yongjun Zhao, Julie Lorette, Long Nguyen, David Huntsman, Connie J. Eaves, Carl Hansen, Marco A. Marra, Carlos Caldas, Sohrab P. Shah, and Samuel Aparicio. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426, Nov 2014.

[102] Pierre Martinez, Margriet R. Timmer, Chiu T. Lau, Silvia Calpe, Maria del Carmen Sancho-Serra, Danielle Straub, Ann-Marie Baker, Sybren L. Meijer, Fiebo J. W. ten Kate, Rosalie C. Mallant-Hent, Anton H. J. Naber, Arnoud H. A. M. van Oijen, Lubbertus C. Baak, Pieter Scholten, Clarisse J. M. Bhmer, Paul Fockens, Jacques J. G. H. M. Bergman, Carlo C. Maley, Trevor A. Graham, and Kausilia K Krishnadath. Dynamic clonal equilibrium and pre-determined cancer risk in Barrett's oesophagus. *Nature Communications*, 7:12158, Aug 2016.

[103] Eric J. Gratias, Jeffrey S. Dome, Lawrence J. Jennings, Yueh-Yun Chi, Jing Tian, James Anderson, Paul Grundy, Elizabeth A. Mullen, James I. Geller, Conrad V. Fernandez, and Elizabeth J. Perlman. Association of Chromosome 1q Gain With Inferior Survival in Favorable-Histology Wilms Tumor: A Report From the Children's Oncology Group. *Journal of Clinical Oncology*, 34(26):3189–3194, Sep 2016.

[104] Tasnim Chagtai, Christina Zill, Linda Dainese, Jenny Wegert, Suvi Savola, Sergey Popov, William Mifsud, Gordan Vujanić, Neil Sebire, Yves Le Bouc, Peter F. Ambros, Leo Kager, Maureen J. O'Sullivan, Annick Blaise, Christophe Bergeron, Linda Holmquist Mengelbier, David Gisselsson, Marcel Kool, Godelieve A.M. Tytgat, Marry M. van den Heuvel-Eibrink, Norbert Graf, Harm van Tinteren, Aurore Coulomb, Manfred Gessler, Richard Dafydd Williams, and Kathy Pritchard-Jones. Gain of 1q As a Prognostic Biomarker in Wilms Tumors (WTs) Treated With Preoperative Chemotherapy in the International Society of Paediatric Oncology (SIOP) WT 2001 Trial: A SIOP Renal Tumours Biology Consortium Study. *Journal of Clinical Oncology*, 34(26):3195–3203, Sep 2016.

[105] Takahiro Gotoh, Hiroyuki Sugihara, Takafumi Matsumura, Kanade Katsura, Tetsuro Takamatsu, and Tadashi Sawada. Human neuroblastoma demonstrating clonal evolution in vivo. *Genes Chromosomes and Cancer*, 22(1):42–49, May 1998.

[106] J Mora, NKV Cheung, and WL Gerald. Genetic heterogeneity and clonal evolution in neuroblastoma. *British Journal of Cancer*, 85(2):182–189, Jul 2001.

[107] Thomas F Eleveld, Derek A Oldridge, Virginie Bernard, Jan Koster, Leo Colmet Daage, Sharon J Diskin, Linda Schild, Nadia Bessoltane Bentahar, Angela Bellini, Mathieu Chicard, Eve Lapouble, Valérie Combaret, Patricia Legoix-Né, Jean Michon, Trevor J Pugh, Lori S Hart, JulieAnn Rader, Edward F Attiyeh, Jun S Wei, Shile Zhang, Arlene Naranjo, Julie M Gastier-

Foster, Michael D Hogarty, Shahab Asgharzadeh, Malcolm A Smith, Jaime M Guidry Auvil, Thomas B K Watkins, Danny A Zwijnenburg, Marli E Ebus, Peter van Sluis, Anne Hakkert, Esther van Wezel, C Ellen van der Schoot, Ellen M Westerhout, Johannes H Schulte, Godelieve A Tytgat, M Emmy M Dolman, Isabelle Janoueix-Lerosey, Daniela S Gerhard, Huib N Caron, Olivier Delattre, Javed Khan, Rogier Versteeg, Gudrun Schleiermacher, Jan J Molenaar, and John M Maris. Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. *Nature Genetics*, 47(8):864–871, Jun 2015.

[108] Xiaochong Wu, Paul A. Northcott, Adrian Dubuc, Adam J. Dupuy, David J. H. Shih, Hendrik Witt, Sidney Croul, Eric Bouffet, Daniel W. Fults, Charles G. Eberhart, Livia Garzia, Timothy Van Meter, David Zagzag, Nada Jabado, Jeremy Schwartzentruber, Jacek Majewski, Todd E. Scheetz, Stefan M. Pfister, Andrey Korshunov, Xiao-Nan Li, Stephen W. Scherer, Yoon-Jae Cho, Keiko Akagi, Tobey J. MacDonald, Jan Koster, Martin G. McCabe, Aaron L. Sarver, V. Peter Collins, William A. Weiss, David A. Largaespada, Lara S. Collier, and Michael D. Taylor. Clonal selection drives genetic divergence of metastatic medulloblastoma. *Nature*, 482(7386):529–533, Feb 2012.

[109] A. Sorana Morrissy, Livia Garzia, David J. H. Shih, Scott Zuyderduyn, Xi Huang, Patryk Skowron, Marc Remke, Florence M. G. Cavalli, Vijay Ramaswamy, Patricia E. Lindsay, Salomeh Jelveh, Laura K. Donovan, Xin Wang, Betty Luu, Kory Zayne, Yisu Li, Chelsea Mayoh, Nina Thiessen, Eloi Mercier, Karen L. Mungall, Yusanne Ma, Kane Tse, Thomas Zeng, Karey Shumansky, Andrew J. L. Roth, Sohrab Shah, Hamza Farooq, Noriyuki Kijima, Borja L. Holgado, John J. Y. Lee, Stuart Matan-Lithwick, Jessica Liu, Stephen C. Mack, Alex Manno, K. A. Michealraj, Carolina Nor, John Peacock, Lei Qin, Juri Reimand, Adi Rolider, Yuan Y. Thompson, Xiaochong Wu, Trevor Pugh, Adrian Ally, Mikhail Bilenky, Yaron S. N. Butterfield, Rebecca Carlsen, Young Cheng, Eric Chuah, Richard D. Corbett, Noreen

Dhalla, An He, Darlene Lee, Haiyan I. Li, William Long, Michael Mayo, Patrick Plettner, Jenny Q. Qian, Jacqueline E. Schein, Angela Tam, Tina Wong, Inanc Birol, Yongjun Zhao, Claudia C. Faria, José Pimentel, Sofia Nunes, Tarek Shalaby, Michael Grotzer, Ian F. Pollack, Ronald L. Hamilton, Xiao-Nan Li, Anne E. Bendel, Daniel W. Fults, Andrew W. Walter, Toshihiro Kumabe, Teiji Tominaga, V. Peter Collins, Yoon-Jae Cho, Caitlin Hoffman, David Lyden, Jeffrey H. Wisoff, James H. Garvin, Duncan S. Stearns, Luca Massimi, Ulrich Schüller, Jaroslav Sterba, Karel Zitterbart, Stephanie Puget, Olivier Ayrault, Sandra E. Dunn, Daniela P. C. Tirapelli, Carlos G. Carlotti, Helen Wheeler, Andrew R. Hallahan, Wendy Ingram, Tobey J. MacDonald, Jeffrey J. Olson, Erwin G. Van Meir, Ji-Yeoun Lee, Kyu-Chang Wang, Seung-Ki Kim, Byung-Kyu Cho, Torsten Pietsch, Gudrun Fleischhack, Stephan Tippelt, Young Shin Ra, Simon Bailey, Janet C. Lindsey, Steven C. Clifford, Charles G. Eberhart, Michael K. Cooper, Roger J. Packer, Maura Massimino, Maria Luisa Garre, Ute Bartels, Uri Tabori, Cynthia E. Hawkins, Peter Dirks, Eric Bouffet, James T. Rutka, Robert J. Wechsler-Reya, William A. Weiss, Lara S. Collier, Adam J. Dupuy, Andrey Korshunov, David T. W. Jones, Marcel Kool, Paul A. Northcott, Stefan M. Pfister, David A. Largaespada, Andrew J. Mungall, Richard A. Moore, Nada Jabado, Gary D. Bader, Steven J. M. Jones, David Malkin, Marco A. Marra, and Michael D. Taylor. Divergent clonal selection dominates medulloblastoma at recurrence. *Nature*, 529(7586):351–357, Jan 2016.

[110] A Sorana Morrissy, Florence M G Cavalli, Marc Remke, Vijay Ramaswamy, David J H Shih, Borja L Holgado, Hamza Farooq, Laura K Donovan, Livia Garzia, Sameer Agnihotri, Erin N Kiehna, Eloi Mercier, Chelsea Mayoh, Simon Papillon-Cavanagh, Hamid Nikbakht, Tenzin Gayden, Jonathon Torchia, Daniel Picard, Diana M Merino, Maria Vladoiu, Betty Luu, Xiaochong Wu, Craig Daniels, Stuart Horswell, Yuan Yao Thompson, Volker Hovestadt, Paul A Northcott, David T W Jones, John Peacock, Xin Wang, Stephen C Mack, Jüri Reimand, Steffen Albrecht, Adam M Fontebasso,

Nina Thiessen, Yisu Li, Jacqueline E Schein, Darlene Lee, Rebecca Carlsen, Michael Mayo, Kane Tse, Angela Tam, Noreen Dhalla, Adrian Ally, Eric Chuah, Young Cheng, Patrick Plettner, Haiyan I Li, Richard D Corbett, Tina Wong, William Long, James Loukides, Pawel Buczkowicz, Cynthia E Hawkins, Uri Tabori, Brian R Rood, John S Myseros, Roger J Packer, Andrey Korshunov, Peter Lichter, Marcel Kool, Stefan M Pfister, Ulrich Schüller, Peter Dirks, Annie Huang, Eric Bouffet, James T Rutka, Gary D Bader, Charles Swanton, Yusanne Ma, Richard A Moore, Andrew J Mungall, Jacek Majewski, Steven J M Jones, Sunit Das, David Malkin, Nada Jabado, Marco A Marra, and Michael D Taylor. Spatial heterogeneity in medulloblastoma. *Nature Genetics*, 49(5):780–788, Apr 2017.

[111] David N. Louis, Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, Hiroko Ohgaki, Otmar D. Wiestler, Paul Kleihues, and David W. Ellison. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathologica*, 131(6):803–820, May 2016.

[112] Hamid Nikbakht, Eshini Panditharatna, Leonie G. Mikael, Rui Li, Tenzin Gayden, Matthew Osmond, Cheng-Ying Ho, Madhuri Kambhampati, Eugene I. Hwang, Damien Faury, Alan Siu, Simon Papillon-Cavanagh, Denise Bechet, Keith L. Ligon, Benjamin Ellezam, Wendy J. Ingram, Caedyn Stinson, Andrew S. Moore, Katherine E. Warren, Jason Karamchandani, Roger J. Packer, Nada Jabado, Jacek Majewski, and Javad Nazarian. Spatial and temporal homogeneity of driver mutations in diffuse intrinsic pontine glioma. *Nature Communications*, 7:11185, Apr 2016.

[113] Emilia M. Pinto, Xiang Chen, John Easton, David Finkelstein, Zhifa Liu, Stanley Pounds, Carlos Rodriguez-Galindo, Troy C. Lund, Elaine R. Mardis, Richard K. Wilson, Kristy Boggs, Donald Yergeau, Jinjun Cheng, Heather L. Mulder, Jayanthi Manne, Jesse Jenkins, Maria J. Mastellaro, Bonald C. Figueiredo, Michael A. Dyer, Alberto Pappo, Jinghui Zhang, James R.

Downing, Raul C. Ribeiro, and Gerard P. Zambetti. Genomic landscape of paediatric adrenocortical tumours. *Nature Communications*, 6:6302, Mar 2015.

[114] Li Chen, Jack F. Shern, Jun S. Wei, Marielle E. Yohe, Young K. Song, Laura Hurd, Hongling Liao, Daniel Catchpoole, Stephen X. Skapek, Frederic G. Barr, Douglas S. Hawkins, and Javed Khan. Clonality and Evolutionary History of Rhabdomyosarcoma. *PLOS Genetics*, 11(3):e1005075, Mar 2015.

[115] Linda Holmquist Mengelbier, Jenny Karlsson, David Lindgren, Anders Valind, Henrik Lilljebjörn, Caroline Jansson, Daniel Bexell, Noémie Braekeveldt, Adam Ameur, Tord Jonson, Hanna Göransson Kultima, Anders Isaksson, Jurate Asmundsson, Rogier Versteeg, Marianne Rissler, Thoas Fioretos, Bengt Sandstedt, Anna Börjesson, Torbjörn Backman, Niklas Pal, Ingrid Øra, Markus Mayrhofer, and David Gisselsson. Intratumoral genome diversity parallels progression and predicts outcome in pediatric cancer. *Nature Communications*, 6:6125, Jan 2015.

[116] Jenny Wegert, Stefanie Wittmann, Ivo Leuschner, Eva Geissinger, Norbert Graf, and Manfred Gessler. WTX inactivation is a frequent but late event in Wilms tumors without apparent clinical impact. *Genes, Chromosomes and Cancer*, 48(12):1102–1111, Dec 2009.

[117] Sergey D. Popov, Gordan M. Vujanic, Neil J. Sebire, Tasnim Chagtai, Richard Williams, Sucheta Vaidya, and Kathy Pritchard-Jones. Bilateral Wilms Tumor with TP53 -Related Anaplasia. *Pediatric and Developmental Pathology*, 16(3):217–223, May 2013.

[118] Richard D. Williams, Tasnim Chagtai, Marisa Alcaide-German, John Apps, Jenny Wegert, Sergey Popov, Gordan Vujanic, Harm van Tinteren, Marry M. van den Heuvel-Eibrink, Marcel Kool, Jan de Kraker, David Gisselsson, Norbert Graf, Manfred Gessler, and Kathy Pritchard-Jones. Multiple mecha-

nisms of MYCN dysregulation in Wilms tumour. *Oncotarget*, 6(9):7232–7243, Apr 2015.

[119] Ugo Del Monte. Does the cell number 10^9 still really fit one gram of tumor tissue? *Cell Cycle*, 8(3):505–506, Feb 2009.

[120] Serena Nik-Zainal, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, Adam Shlien, Susanna L. Cooke, Jonathan Hinton, Andrew Menzies, Lucy A. Stebbings, Catherine Leroy, Mingming Jia, Richard Rance, Laura J. Mudie, Stephen J. Gamble, Philip J. Stephens, Stuart McLaren, Patrick S. Tarpey, Elli Papaemmanuil, Helen R. Davies, Ignacio Varela, David J. McBride, Graham R. Bignell, Kenric Leung, Adam P. Butler, Jon W. Teague, Sancha Martin, Goran Jönsson, Odette Mariani, Sandrine Boyault, Penelope Miron, Aquila Fatima, Anita Langerød, Samuel A.J.R. Aparicio, Andrew Tutt, Anieta M. Sieuwerts, Åke Borg, Gilles Thomas, Anne Vincent Salomon, Andrea L. Richardson, Anne-Lise Børresen-Dale, P. Andrew Futreal, Michael R. Stratton, and Peter J. Campbell. The Life History of 21 Breast Cancers. *Cell*, 149(5):994–1007, May 2012.

[121] Sohrab P. Shah, Andrew Roth, Rodrigo Goya, Arusha Oloumi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, Jiarui Ding, Kane Tse, Gholamreza Haffari, Ali Bashashati, Leah M. Prentice, Jaswinder Khattra, Angela Burleigh, Damian Yap, Virginie Bernard, Andrew McPherson, Karey Shumansky, Anamaria Crisan, Ryan Giuliany, Alireza Heravi-Moussavi, Jamie Rosner, Daniel Lai, Inanc Birol, Richard Varhol, Angela Tam, Noreen Dhalla, Thomas Zeng, Kevin Ma, Simon K. Chan, Malachi Griffith, Annie Moradian, S.-W. Grace Cheng, Gregg B. Morin, Peter Watson, Karen Gelmon, Stephen Chia, Suet-Feung Chin, Christina Curtis, Oscar M. Rueda, Paul D. Pharoah, Sambasivarao Damaraju, John Mackey, Kelly Hoon, Timothy Harkins, Vasisht Tadigotla, Mahvash Sigaroudinia, Philippe Gascard, Thea

Tlsty, Joseph F. Costello, Irmtraud M. Meyer, Connie J. Eaves, Wyeth W. Wasserman, Steven Jones, David Huntsman, Martin Hirst, Carlos Caldas, Marco A. Marra, and Samuel Aparicio. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–399, Apr 2012.

[122] Marco Gerlinger, Andrew J. Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, Ignacio Varela, Benjamin Phillimore, Sharmin Begum, Neil Q. McDonald, Adam Butler, David Jones, Keiran Raine, Calli Latimer, Claudio R. Santos, Mahrokh Nohadani, Aron C. Eklund, Bradley Spencer-Dene, Graham Clark, Lisa Pickering, Gordon Stamp, Martin Gore, Zoltan Szallasi, Julian Downward, P. Andrew Futreal, and Charles Swanton. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine*, 366(10):883–892, Mar 2012.

[123] Lucy R Yates, Moritz Gerstung, Stian Knappskog, Christine Desmedt, Gunes Gundem, Peter Van Loo, Turid Aas, Ludmil B Alexandrov, Denis Larsimont, Helen Davies, Yilong Li, Young Seok Ju, Manasa Ramakrishna, Hans Kristian Haugland, Peer Kaare Lilleng, Serena Nik-Zainal, Stuart McLaren, Adam Butler, Sancha Martin, Dominic Glodzik, Andrew Menzies, Keiran Raine, Jonathan Hinton, David Jones, Laura J Mudie, Bing Jiang, Delphine Vincent, April Greene-Colozzi, Pierre-Yves Adnet, Aquila Fatima, Marion Maetens, Michail Ignatiadis, Michael R Stratton, Christos Sotiriou, Andrea L Richardson, Per Eystein Lønning, David C Wedge, and Peter J Campbell. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature Medicine*, 21(7):751–759, Jun 2015.

[124] Sohrab P. Shah, Ryan D. Morin, Jaswinder Khattra, Leah Prentice, Trevor Pugh, Angela Burleigh, Allen Delaney, Karen Gelmon, Ryan Guliany, Janine Senz, Christian Steidl, Robert A. Holt, Steven Jones, Mark Sun,

Gillian Leung, Richard Moore, Tesa Severson, Greg A. Taylor, Andrew E. Teschendorff, Kane Tse, Gulisa Turashvili, Richard Varhol, René L. Warren, Peter Watson, Yongjun Zhao, Carlos Caldas, David Huntsman, Martin Hirst, Marco A. Marra, and Samuel Aparicio. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 461(7265):809–813, Oct 2009.

[125] Niccolo Bolli, Hervé Avet-Loiseau, David C. Wedge, Peter Van Loo, Ludmil B. Alexandrov, Inigo Martincorena, Kevin J. Dawson, Francesco Iorio, Serena Nik-Zainal, Graham R. Bignell, Jonathan W. Hinton, Yilong Li, Jose M.C. Tubio, Stuart McLaren, Sarah O' Meara, Adam P. Butler, Jon W. Teague, Laura Mudie, Elizabeth Anderson, Naim Rashid, Yu-Tzu Tai, Masood A. Shammas, Adam S. Sperling, Mariateresa Fulciniti, Paul G. Richardson, Giovanni Parmigiani, Florence Magrangeas, Stephane Minvielle, Philippe Moreau, Michel Attal, Thierry Facon, P Andrew Futreal, Kenneth C. Anderson, Peter J. Campbell, and Nikhil C. Munshi. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature Communications*, 5:2997, Jan 2014.

[126] Matthew J. Walter, Dong Shen, Li Ding, Jin Shao, Daniel C. Koboldt, Ken Chen, David E. Larson, Michael D. McLellan, David Dooling, Rachel Abbott, Robert Fulton, Vincent Magrini, Heather Schmidt, Joelle Kalicki-Veizer, Michelle O'Laughlin, Xian Fan, Marcus Grillot, Sarah Witowski, Sharon Heath, John L. Frater, William Eades, Michael Tomasson, Peter Westervelt, John F. DiPersio, Daniel C. Link, Elaine R. Mardis, Timothy J. Ley, Richard K. Wilson, and Timothy A. Graubert. Clonal architecture of secondary acute myeloid leukemia. *New England Journal of Medicine*, 366(12):1090–1098, Mar 2012.

[127] A. Schuh, J. Becq, S. Humphray, A. Alexa, A. Burns, R. Clifford, S. M. Feller, R. Grocock, S. Henderson, I. Khrebtukova, Z. Kingsbury, S. Luo, D. McBride, L. Murray, T. Menju, A. Timbs, M. Ross, J. Taylor, and

D. Bentley. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, 120(20):4191–4196, Aug 2012.

[128] T H Kim, K Yoshida, Y K Kim, M S Tyndel, H J Park, S H Choi, J-S Ahn, S-H Jung, D-H Yang, J-J Lee, H J Kim, G Kong, S Ogawa, Z Zhang, H J Kim, and D D H Kim. Clonal dynamics in a single AML case tracked for 9 years reveals the complexity of leukemia progression. *Leukemia*, 30(2):295–302, Oct 2015.

[129] Muhammed Murtaza, Sarah-Jane Dawson, Dana W. Y. Tsui, Davina Gale, Tim Forshew, Anna M. Piskorz, Christine Parkinson, Suet-Feung Chin, Zoya Kingsbury, Alvin S. C. Wong, Francesco Marass, Sean Humphray, James Hadfield, David Bentley, Tan Min Chin, James D. Brenton, Carlos Caldas, and Nitzan Rosenfeld. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature*, 497(7447):108–112, Apr 2013.

[130] Sarah-Jane Dawson, Dana W.Y. Tsui, Muhammed Murtaza, Heather Biggs, Oscar M. Rueda, Suet-Feung Chin, Mark J. Dunning, Davina Gale, Tim Forshew, Betania Mahler-Araujo, Sabrina Rajan, Sean Humphray, Jennifer Becq, David Halsall, Matthew Wallis, David Bentley, Carlos Caldas, and Nitzan Rosenfeld. Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer. *New England Journal of Medicine*, 368(13):1199–1209, Mar 2013.

[131] Ö. Winge. Zytologische untersuchungen über die natur maligner tumoren. *Zeitschrift für Zellforschung und Mikroskopische Anatomie*, 10(4):683–735, Dec 1930.

[132] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W. Richard McCombie,

James Hicks, and Michael Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, Mar 2011.

[133] Xun Xu, Yong Hou, Xuyang Yin, Li Bao, Aifa Tang, Luting Song, Fuqiang Li, Shirley Tsang, Kui Wu, Hanjie Wu, Weiming He, Liang Zeng, Manjie Xing, Renhua Wu, Hui Jiang, Xiao Liu, Dandan Cao, Guangwu Guo, Xueda Hu, Yaoting Gui, Zesong Li, Wenyue Xie, Xiaojuan Sun, Min Shi, Zhiming Cai, Bin Wang, Meiming Zhong, Jingxiang Li, Zuhong Lu, Ning Gu, Xiuqing Zhang, Laurie Goodman, Lars Bolund, Jian Wang, Huanming Yang, Karsten Kristiansen, Michael Dean, Yingrui Li, and Jun Wang. Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor. *Cell*, 148(5):886–895, Mar 2012.

[134] Gunes Gundem, Peter Van Loo, Barbara Kremeyer, Ludmil B. Alexandrov, Jose M. C. Tubio, Elli Papaemmanuil, Daniel S. Brewer, Heini M. L. Kallio, Gunilla Högnäs, Matti Annala, Kati Kivinummi, Victoria Goody, Calli Latimer, Sarah O'Meara, Kevin J. Dawson, William Isaacs, Michael R. Emmert-Buck, Matti Nykter, Christopher Foster, Zsofia Kote-Jarai, Douglas Easton, Hayley C. Whitaker, David E. Neal, Colin S. Cooper, Rosalind A. Eeles, Tapio Visakorpi, Peter J. Campbell, Ultan McDermott, David C. Wedge, and G. Steven Bova. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, Apr 2015.

[135] João M. Alves, Tamara Prieto, and David Posada. Multiregional Tumor Trees Are Not Phylogenies. *Trends in Cancer*, 3(8):546–550, Aug 2017.

[136] Jonas Demeulemeester, Parveen Kumar, Elen K. Møller, Silje Nord, David C. Wedge, April Peterson, Randi R. Mathiesen, Renathe Fjelldal, Masoud Zamani Esteki, Koen Theunis, Elia Fernandez Gallardo, A. Jason Grundstad, Elin Borgen, Lars O. Baumbusch, Anne-Lise Børresen-Dale, Kevin P. White, Vessela N. Kristensen, Peter Van Loo, Thierry Voet, and Bjørn Naume. Tracing the origin of disseminated tumor cells in breast cancer using single-cell sequencing. *Genome Biology*, 17(1):250, Dec 2016.

252

[137] Lili Wang, Jean Fan, Joshua M. Francis, George Georghiou, Sarah Hergert, Shuqiang Li, Rutendo Gambe, Chensheng W. Zhou, Chunxiao Yang, Sheng Xiao, Paola Dal Cin, Michaela Bowden, Dylan Kotliar, Sachet A. Shukla, Jennifer R. Brown, Donna Neuberg, Dario R. Alessi, Cheng-Zhong Zhang, Peter V. Kharchenko, Kenneth J. Livak, and Catherine J. Wu. Integrated single-cell genetic and transcriptional analysis suggests novel drivers of chronic lymphocytic leukemia. *Genome Research*, 27(8):1300–1311, Jul 2017.

[138] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, Oct 2008.

[139] K. V. Voelkerding, S. A. Dames, and J. D. Durtschi. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, 55(4):641–658, Feb 2009.

[140] Lira Mamanova, Alison J Coffey, Carol E Scott, Iwanka Kozarewa, Emily H Turner, Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J Turner. Target-enrichment strategies for next-generation sequencing. *Nature Methods*, 7(2):111–118, Feb 2010.

[141] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, May 2009.

[142] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, Mar 2012.

[143] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219, Feb 2013.

[144] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. VarScan 2: Somatic

mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, Feb 2012.

[145] E. Garrison and G. Marth. Haplotype-based variant detection from short-read sequencing. *ArXiv E-Prints*, 1207.3907, 2012.

[146] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, Jul 2010.

[147] T. LaFramboise. Single nucleotide polymorphism arrays: a decade of biological computational and technological advances. *Nucleic Acids Research*, 37(13):4181–4193, Jul 2009.

[148] Johan Staaf, Johan Vallon-Christersson, David Lindgren, Gunnar Juliusson, Richard Rosenquist, Mattias Höglund, Åke Borg, and Markus Ringnér. Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. *BMC Bioinformatics*, 9(1):409, 2008.

[149] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, Oct 2004.

[150] P. Van Loo, S. H. Nordgard, O. C. Lingjaerde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A.-L. Borresen-Dale, and V. N. Kristensen. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, Sep 2010.

[151] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S

Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421, Apr 2012.

[152] Johan Staaf, David Lindgren, Johan Vallon-Christersson, Anders Isaksson, Hanna Goransson, Gunnar Juliusson, Richard Rosenquist, Mattias Hoglund, Ake Borg, and Markus Ringner. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biology*, 9(9):R136, 2008.

[153] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 2010.

[154] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, 5(6):e1000529, Jun 2009.

[155] Roland F. Schwarz, Anne Trinh, Botond Sipos, James D. Brenton, Nick Goldman, and Florian Markowetz. Phylogenetic Quantification of Intra-tumour Heterogeneity. *PLoS Computational Biology*, 10(4):e1003535, Apr 2014.

[156] Russell Schwartz and Alejandro A. Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229, Feb 2017.

[157] N. Beerenwinkel, R. F. Schwarz, M. Gerstung, and F. Markowetz. Cancer Evolution: Mathematical Models and Computational Inference. *Systematic Biology*, 64(1):e1–e25, Oct 2014.

[158] J. Zhang, J. Fujimoto, J. Zhang, D. C. Wedge, X. Song, J. Zhang, S. Seth, C.-W. Chow, Y. Cao, C. Gumbs, K. A. Gold, N. Kalhor, L. Little, H. Mahadeshwar, C. Moran, A. Protopopov, H. Sun, J. Tang, X. Wu, Y. Ye, W. N. William, J. J. Lee, J. V. Heymach, W. K. Hong, S. Swisher, I. I. Wistuba,

and P. A. Futreal. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science*, 346(6206):256–259, Oct 2014.

[159] David Brocks, Yassen Assenov, Sarah Minner, Olga Bogatyrova, Ronald Simon, Christina Koop, Christopher Oakes, Manuela Zucknick, Daniel Bernhard Lipka, Joachim Weischenfeldt, Lars Feuerbach, Richard Cowper-Sal·lari, Mathieu Lupien, Benedikt Brors, Jan Korbel, Thorsten Schlomm, Amos Tanay, Guido Sauter, Clarissa Gerhäuser, and Christoph Plass. Intratumor DNA Methylation Heterogeneity Reflects Clonal Evolution in Aggressive Prostate Cancer. *Cell Reports*, 8(3):798–806, Aug 2014.

[160] Roland F. Schwarz, Charlotte K. Y. Ng, Susanna L. Cooke, Scott Newman, Jillian Temple, Anna M. Piskorz, Davina Gale, Karen Sayal, Muhammed Murtaza, Peter J. Baldwin, Nitzan Rosenfeld, Helena M. Earl, Evis Sala, Mercedes Jimenez-Linan, Christine A. Parkinson, Florian Markowetz, and James D. Brenton. Spatial and Temporal Heterogeneity in High-Grade Serous Ovarian Cancer: A Phylogenetic Analysis. *PLOS Medicine*, 12(2):e1001789, Feb 2015.

[161] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15(1):35, 2014.

[162] Habil Zare, Junfeng Wang, Alex Hu, Kris Weber, Josh Smith, Debbie Nickerson, ChaoZhong Song, Daniela Witten, C. Anthony Blau, and William Stafford Noble. Inferring Clonal Composition from Multiple Sections of a Breast Cancer. *PLoS Computational Biology*, 10(7):e1003703, Jul 2014.

[163] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: Reconstructing subclonal composition

and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, 2015.

[164] Yuchao Jiang, Yu Qiu, Andy J. Minn, and Nancy R. Zhang. Assessing intra-tumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences*, 113(37):E5528–E5537, Aug 2016.

[165] Mohammed El-Kebir, Gryte Satas, Layla Oesper, and Benjamin J. Raphael. Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. *Cell Systems*, 3(1):43–53, Jul 2016.

[166] Eric Letouzé, Yves Allory, Marc A Bollet, François Radvanyi, and Frédéric Guyon. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biology*, 11(7):R76, 2010.

[167] Kathy Pritchard-Jones, Christophe Bergeron, Beatriz de Camargo, Marry M van den Heuvel-Eibrink, Tomas Acha, Jan Godzinski, Foppe Oldenburger, Liliane Boccon-Gibod, Ivo Leuschner, Gordan Vujanic, Bengt Sandstedt, Jan de Kraker, Harm van Tinteren, and Norbert Graf. Omission of doxorubicin from the treatment of stage II–III intermediate-risk Wilms' tumour (SIOP WT 2001): an open-label non-inferiority, randomised controlled trial. *The Lancet*, 386(9999):1156–1164, Sep 2015.

[168] Kirsteen McDonald, Neil J. Sebire, John Anderson, and Øystein E. Olsen. Patterns of shift in ADC distributions in abdominal tumours during chemotherapy—feasibility study. *Pediatric Radiology*, 41(1):99–106, Jul 2010.

[169] R H Scott, J Douglas, L Baskcomb, A O Nygren, J M Birch, T R Cole, V Cormier-Daire, D M Eastwood, S Garcia-Minaur, P Lupunzina, K Tatton-Brown, J Bliek, E R Maher, and N Rahman. Methylation-specific multiplex

ligation-dependent probe amplification (MS-MLPA) robustly detects and distinguishes 11p15 abnormalities associated with overgrowth and growth retardation. *Journal of Medical Genetics*, 45(2):106–113, Sep 2007.

[170] Sharon J. Diskin, Mingyao Li, Cuiping Hou, Shuzhang Yang, Joseph Glessner, Hakon Hakonarson, Maja Bucan, John M. Maris, and Kai Wang. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research*, 36(19):e126–e126, Sep 2008.

[171] Mark A. van de Wiel, Kyung In Kim, Sjoerd J. Vosse, Wessel N. van Wieringen, Saskia M. Wilting, and Bauke Ylstra. CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, 23(7):892–894, Jan 2007.

[172] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean YH Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

[173] MA van de Wiel and WN Wieringen. Cghregions: dimension reduction for array cgh data with minimal information loss. *Cancer Inform*, 3:55–63, Feb 2007.

[174] Tatiana Benaglia, Didier Chauveau, David R. Hunter, and Derek Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009.

[175] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. pROC: an open-source package for r and s+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1):77, 2011.

[176] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. version 2.41.

[177] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, May 2011.

[178] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Jun 2009.

[179] Alex H. Ramos, Lee Lichtenstein, Manaswi Gupta, Michael S. Lawrence, Trevor J. Pugh, Gordon Saksena, Matthew Meyerson, and Gad Getz. Oncotator: Cancer Variant Annotation Tool. *Human Mutation*, 36(4):E2423–E2429, Mar 2015.

[180] Nicole Ludwig, Tamara Werner, Christina Backes, Patrick Trampert, Manfred Gessler, Andreas Keller, Hans-Peter Lenhof, Norbert Graf, and Eckart Meese. Combining miRNA and mRNA Expression Profiles in Wilms Tumor Subtypes. *International Journal of Molecular Sciences*, 17(4):475, Mar 2016.

[181] Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, Sep 2016.

[182] J.R. Gnarra, K. Tory, Y. Weng, L. Schmidt, M.H. Wei, H. Li, F. Latif, S. Liu, F. Chen, F.-M. Duh, I. Lubensky, D.R. Duan, C. Florence, R. Pozzatti, M. M. Walther, N.H. Bander, H.B. Grossman, H. Brauch, S. Pomer, J.D. Brooks, W.B. Isaacs, M.I. Lerman, B. Zbar, and W.M. Linehan. Mutations of the VHL tumour suppressor gene in renal carcinoma. *Nature Genetics*, 7(1):85–90, May 1994.

[183] Chi-Ming Li, Connie E. Kim, Adam A. Margolin, Meirong Guo, Jimmy Zhu, Jacqueline M. Mason, Terrence W. Hensle, Vundavalli V.V.S. Murty, Paul E.

Grundy, Eric R. Fearon, Vivette D'Agati, Jonathan D. Licht, and Benjamin Tycko. CTNNB1 Mutations and Overexpression of Wnt/β-Catenin Target Genes in WT1-Mutant Wilms' Tumors. *The American Journal of Pathology*, 165(6):1943–1953, Dec 2004.

[184] N. J. Smilinich, C. D. Day, G. V. Fitzpatrick, G. M. Caldwell, A. C. Lossie, P. R. Cooper, A. C. Smallwood, J. A. Joyce, P. N. Schofield, W. Reik, R. D. Nicholls, R. Weksberg, D. J. Driscoll, E. R. Maher, T. B. Shows, and M. J. Higgins. A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. *Proceedings of the National Academy of Sciences*, 96(14):8064–8069, Jul 1999.

[185] Adrian K. Charles, Keith W. Brown, and P. Jeremy Berry. Microdissecting the Genetic Events in Nephrogenic Rests and Wilms' Tumor Development. *The American Journal of Pathology*, 153(3):991–1000, Sep 1998.

[186] Patrick W. Hales, Øystein E. Olsen, Neil J. Sebire, Kathy Pritchard-Jones, and Chris A. Clark. A multi-Gaussian model for apparent diffusion coefficient histogram analysis of Wilms' tumour subtype and response to chemotherapy. *NMR in Biomedicine*, 28(8):948–957, Jun 2015.

[187] Eugene B. Cone, Stewart S. Dalton, Megan Van Noord, Elizabeth T. Tracy, Henry E. Rice, and Jonathan C. Routh. Biomarkers for Wilms Tumor: A Systematic Review. *The Journal of Urology*, 196(5):1530–1535, Nov 2016.

[188] T Révész, C Mpofu, C Oyejide, and A Bener. Socioeconomic factors in the families of children with lymphoid malignancy in the UAE. *Leukemia*, 11(4):588–593, Apr 1997.

[189] D. Perotti, P. Hohenstein, I. Bongarzone, M. Maschietto, M. Weeks, P. Radice, and K. Pritchard-Jones. Is Wilms Tumor a Candidate Neoplasia for Treatment with WNT/β-Catenin Pathway Modulators?–A Report from

the Renal Tumors Biology-Driven Drug Development Workshop. *Molecular Cancer Therapeutics*, 12(12):2619–2627, Nov 2013.

[190] Michael Girardot, Jérôme Cavaillé, and Robert Feil. Small regulatory RNAs controlled by genomic imprinting and their contribution to human disease. *Epigenetics*, 7(12):1341–1348, Dec 2012.

[191] Meihua Li, Kyle F. Lee, Yuntao Lu, Ian Clarke, David Shih, Charles Eberhart, V. Peter Collins, Tim Van Meter, Daniel Picard, Limei Zhou, Paul C. Boutros, Piergiorgio Modena, Muh-Lii Liang, Steve W. Scherer, Eric Bouffet, James T. Rutka, Scott L. Pomeroy, Ching C. Lau, Michael D. Taylor, Amar Gajjar, Peter B. Dirks, Cynthia E. Hawkins, and Annie Huang. Frequent Amplification of a chr19q13.41 MicroRNA Polycistron in Aggressive Primitive Neuroectodermal Brain Tumors. *Cancer Cell*, 16(6):533–546, Dec 2009.

[192] Wan-Heng Wang, Jian-Xiong Duan, Thanh H. Vu, and Andrew R. Hoffman. Increased Expression of the Insulin-like Growth Factor-II Gene in Wilms' Tumor Is Not Dependent on Loss of Genomic Imprinting or Loss of Heterozygosity. *Journal of Biological Chemistry*, 271(44):27863–27870, Nov 1996.

[193] Eliana Amato, Marco dal Molin, Andrea Mafficini, Jun Yu, Giuseppe Malleo, Borislav Rusev, Matteo Fassan, Davide Antonello, Yoshihiko Sadakari, Paola Castelli, Giuseppe Zamboni, Anirban Maitra, Roberto Salvia, Ralph H Hruban, Claudio Bassi, Paola Capelli, Rita T Lawlor, Michael Goggins, and Aldo Scarpa. Targeted next-generation sequencing of cancer genes dissects the molecular profiles of intraductal papillary neoplasms of the pancreas. *The Journal of Pathology*, 233(3):217–227, Jun 2014.

[194] Ryuji Fukuzawa, Rosemary W. Heathcott, Makoto Sano, Ian M. Morison, Kankatsu Yun, and Anthony E. Reeve. Myogenesis in Wilms Tumors Is Associated with Mutations of the WT1 Gene and Activation of Bcl-2 and the

Wnt Signaling Pathway. *Pediatric and Developmental Pathology*, 7(2):125–137, Apr 2004.

[195] Y Satoh, H Nakadate, T Nakagawachi, K Higashimoto, K Joh, Z Masaki, J Uozumi, Y Kaneko, T Mukai, and H Soejima. Genetic and epigenetic alterations on the short arm of chromosome 11 are involved in a majority of sporadic Wilms' tumours. *British Journal of Cancer*, 95(4):541–547, Aug 2006.

[196] R. D. Williams, R. Al-Saadi, T. Chagtai, S. Popov, B. Messahel, N. Sebire, M. Gessler, J. Wegert, N. Graf, I. Leuschner, M. Hubank, C. Jones, G. Vujanic, and K. Pritchard-Jones. Subtype-Specific FBXW7 Mutation and MYCN Copy Number Gain in Wilms' Tumor. *Clinical Cancer Research*, 16(7):2036–2045, Mar 2010.

[197] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague, M. R. Stratton, U. McDermott, and P. J. Campbell. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(D1):D805–D811, Oct 2014.

[198] Valeria Romanelli, Heloisa N M Meneses, Luis Fernández, Victor Martínez-Glez, Ricardo Gracia-Bouthelier, Mario F Fraga, Encarna Guillén, Julián Nevado, Esther Gean, Loreto Martorell, Victoria Esteban Marfil, Sixto García-Miñaur, and Pablo Lapunzina. Beckwith–Wiedemann syndrome and uniparental disomy 11p: fine mapping of the recombination breakpoints and evaluation of several techniques. *European Journal of Human Genetics*, 19(4):416–421, Jan 2011.

[199] S Maiti, R Alam, CI Amos, and V Huff. Frequent association of beta-catenin and wt1 mutations in wilms tumors. *Cancer Research*, 60:6288–92, Nov 2000.

[200] Qianghua Hu, Fei Gao, Weihua Tian, E. Cristy Ruteshouser, Yaqing Wang, Alexander Lazar, John Stewart, Louise C. Strong, Richard R. Behringer, and Vicki Huff. Wt1 ablation and Igf2 upregulation in mice result in Wilms tumors with elevated ERK1/2 phosphorylation. *Journal of Clinical Investigation*, 121(1):174–183, Jan 2011.

[201] Le Huang, Sharada Mokkapati, Qianghua Hu, E. Cristy Ruteshouser, M. John Hicks, and Vicki Huff. Nephron Progenitor But Not Stromal Progenitor Cells Give Rise to Wilms Tumors in Mouse Models with β-Catenin Activation or Wt1 Ablation and Igf2 Upregulation. *Neoplasia*, 18(2):71–81, Feb 2016.

[202] C. Uschkereit, N. Perez, C. de Torres, M. Kuff, J. Mora, and B. Royer-Pokora. Different CTNNB1 mutations as molecular genetic proof for the independent origin of four Wilms tumours in a patient with a novel germ line WT1 mutation. *Journal of Medical Genetics*, 44(6):393–396, Feb 2007.

[203] Sam Behjati, Meritxell Huch, Ruben van Boxtel, Wouter Karthaus, David C. Wedge, Asif U. Tamuri, Iñigo Martincorena, Mia Petljak, Ludmil B. Alexandrov, Gunes Gundem, Patrick S. Tarpey, Sophie Roerink, Joyce Blokker, Mark Maddison, Laura Mudie, Ben Robinson, Serena Nik-Zainal, Peter Campbell, Nick Goldman, Marc van de Wetering, Edwin Cuppen, Hans Clevers, and Michael R. Stratton. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, 513(7518):422–425, Jun 2014.

[204] Young Seok Ju, Inigo Martincorena, Moritz Gerstung, Mia Petljak, Ludmil B. Alexandrov, Raheleh Rahbari, David C. Wedge, Helen R. Davies, Manasa Ramakrishna, Anthony Fullam, Sancha Martin, Christopher Alder, Nikita Patel, Steve Gamble, Sarah O'Meara, Dilip D. Giri, Torril Sauer, Sarah E. Pinder, Colin A. Purdie, Åke Borg, Henk Stunnenberg, Marc van de Vijver, Benita K. T. Tan, Carlos Caldas, Andrew Tutt, Naoto T. Ueno, Laura J. van 't Veer, John W. M. Martens, Christos Sotiriou, Stian

Knappskog, Paul N. Span, Sunil R. Lakhani, Jórunn Erla Eyfjörd, Anne-Lise Børresen-Dale, Andrea Richardson, Alastair M. Thompson, Alain Viari, Matthew E. Hurles, Serena Nik-Zainal, Peter J. Campbell, and Michael R. Stratton. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*, 543(7647):714–718, Mar 2017.

[205] Rajiv C. McCoy. Mosaicism in Preimplantation Human Embryos: When Chromosomal Abnormalities Are the Norm. *Trends in Genetics*, 33(7):448–463, Jul 2017.

[206] Soulafa Mamlouk, Liam Harold Childs, Daniela Aust, Daniel Heim, Friederike Melching, Cristiano Oliveira, Thomas Wolf, Pawel Durek, Dirk Schumacher, Hendrik Bläker, Moritz von Winterfeld, Bastian Gastl, Kerstin Möhr, Andrea Menne, Silke Zeugner, Torben Redmer, Dido Lenze, Sascha Tierling, Markus Möbs, Wilko Weichert, Gunnar Folprecht, Eric Blanc, Dieter Beule, Reinhold Schäfer, Markus Morkel, Frederick Klauschen, Ulf Leser, and Christine Sers. DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer. *Nature Communications*, 8:14093, Jan 2017.