The role of chromosomal instability and parallel evolution in cancer

Thomas Benjamin Kingdon Watkins

PhD supervisor: Charles Swanton

A dissertation submitted for the degree of

Doctor of Philosophy

University College London

December 2018

Declaration

I, Thomas Watkins, confirm all 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 thesis.

December 19, 2018

Thomas Watkins

Abstract

Although chromosomal instability (CIN) is recognised as an initiating process in cancer, the extent and relevance of ongoing somatic copy number alterations (SCNAs) that result from it later in tumour development is unclear. In this thesis I describe a comprehensive analysis, including 1421 tumour samples (394 patients; 22 tumour types), to evaluate ongoing CIN and SCNAs in tumour evolution and show that intratumor heterogeneity mediated through chromosomal instability is associated with an increased risk of recurrence or death in non-small cell lung cancer (NSCLC), a finding that supports the potential value of CIN as a prognostic predictor.

I also uncover pervasive SCNA intratumour heterogeneity across cancers, with recurrent clonal and subclonal events identified and found to demonstrate enrichment for cancer genes. I develop novel techniques for obtaining a phasing of heterozygous SNPs from multi-region next generation sequencing data and apply them to observe recurrent parallel evolutionary events converging upon disruption to the same genes in distinct subclones within 146 individual tumours.

The most prevalent recurrent parallel loss event involved chromosome 14, including *HIF1A* and *HIF1B*. In addition, chromosome 5p, including *TERT*, was recurrently gained and subject to parallel evolution in 7 tumour types. Tumour type-specific constraints to early tumour development were identified in the form of obligatory clonal LOH, including LOH of 3p in clear cell renal cell carcinoma, lung squamous cell carcinoma (LUSC) and triple-negative breast cancer and LOH of 17p in LUSC, colorectal adenocarcinoma, triple negative and HER2+ breast cancer. Whole-genome doubling (WGD) was generally an early event in tumour evolution, associated with an increased acquisition of both clonal and subclonal SCNAs. For instance, *CCNE1* amplifications, which occurred exclusively in WGD tumours, were subclonal in 45% of these cases, suggesting this event may be selected following a WGD event. Mathematical modelling of subclonal SCNA evolution demonstrated that models that incorporate ongoing selection with respect to SCNAs significantly outperform evolutionary neutral models, particularly in the context of WGD.

This thesis highlights the importance of ongoing CIN and recurrent subclonal chromosomal alterations in tumour evolution, reveals parallel evolution of SCNAs, and sheds light on the dynamics and order of events that influence metastasis.

Impact statement

Cancer is one of the main causes of mortality and morbidity in the world today and an important factor that contributes to poor outcome in cancer is intratumour heterogeneity. This heterogeneity provides the variation of somatic genetic alterations upon which the force of natural selection may act causing the expansion of subclonal populations of cancer cells that carry a subset of alterations increasing their evolutionary fitness. The source of these somatic alterations is genetic instability, a subtype of which is the process of chromosomal instability (CIN) that allows a state of aneuploidy to occur through changes in chromosome copy number and structural variation through changes in chromosome architecture. The characterisation of the chromosomal complement of cancer cells using cytological techniques has contributed greatly to early theories of cancer evolution but, in the era of next-generation sequencing, has taken a lesser role compared to contributions of the newly described wealth of sequence-level variants and their heterogeneity. My research has focused on methods to quantify chromosomal copy number heterogeneity from multi-region next generation sequencing data and assess its impact and role in both early and late cancer evolution. Parallel evolution of sequence-level variants in separate subclones of the same patient's disease is thought to demonstrate selection and likely constraints on later tumour evolution. A better understanding of chromosome copy number changes in cancer evolution and identification of similar instances of copy number based instances of parallel evolution may help identify patient subgroups and allow a choice of treatment interventions in a patient specific manner.

Acknowledgements

First and foremost I would like to thank my supervisor Charlie Swanton for his unwavering support throughout my time in the lab. He has been a constant source of insight, enthusiasm and exhortations to focus!

I am very grateful to have had the opportunity to work on the TRACERx project with a fantastic team from whom I have learnt so much. I would particularly like to thank Nicholas McGranahan, Nicolai Birkbak, Gareth Wilson and Emilia Lim who have not only taught me almost everything I know about cancer genomics but have also made the last four years in the lab so much fun. In addition, I am lucky to have been able to discuss phylogenetics, somatic copy number calling and phasing with Christophe Dessimoz, Nick Luscombe and Roland Schwarz who have all guided me at critical junctures.

I would also like to thank my friends and family for their encouragement, especially my father whose calm advice has always proved helpful. Finally, I would like to thank Nelly whose wit, cheerfulness and genuine interest in my work have been invaluable throughout my PhD.

Equations

(Equations 1) Sequence-based mutational cancer cell calculation.
(Equations 2) Calculating continuous values of “raw” copy number for each BAF locus.
(Equations 3) Ploidy and purity dependent copy number thresholds.

Abbreviations

AI	AI
ALL	acute lymphoblastic leukaemia
ASCN	allele-specific copy number
BAF	B-allele frequency
BE	Barrett's oesophagus
BLCA	bladder adenocarcinoma
BRCA	breast carcinoma
BRCA ER+	oestrogen receptor positive breast carcinoma
BRCA HER+	human epidermal growth factor receptor 2 positive breast carcinoma
BRCA TN	triple-negative breast carcinoma
CCF	cancer cell fraction
CGH	comparative genomic hybridisation
CLL	chronic lymphocytic leukaemia
COAD	colon adenocarcinoma
ESCA	oesophageal adenocarcinoma
FISH	fluorescence *in situ* hybridization
GBM	glioblastoma multiforme
GBMLGG	glioma
GD	genome doubled
HGSOC	high-grade serous ovarian cancer
HD	homozygous deletion

HLA	human leukocyte antigen
HNSC	head and neck squamous cell carcinoma
ISM	infinite sites model
ITH	intratumour heterogeneity
INDEL	small insertion and/or deletion
LogR	log-ratio
KIRP	papillary renal cell carcinoma
KICH	chromophobe renal cell carcinoma
MB	megabase
MRCA	most recent common ancestor
GB	gigabase
LOH	loss-of-heterozygosity
LOR	log odds ratio
LUAD	lung adenocarcinoma
LUSC	lung squamous cell carcinoma
MEDICC	minimum event distance for intra-tumour copy number comparisons
MHC	major histocompatibility complex
MRP	multi-region phasing
MSI	microsatellite instability
nGD	non-genome doubled
NGS	next-generation sequencing
NSCLC	non-small cell lung cancer
OS	overall survival

OVCA	ovarian serous cystadenocarcinoma
PCAWG	pan-cancer analysis of whole genomes
PCR	polymerase chain reaction
PRAD	prostate adenocarcinoma
PFS	progression free survival
RFS	relapse free survival
SAC	spindle assembly checkpoint
SARC	sarcoma
SCS	single cell sequencing
SDC	salivary duct carcinoma
sGD	subclonal genome doubling
SKCM	skin cutaneous melanoma
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
TCGA	the cancer genome atlas
TCN	total copy number
THCA	thyroid carcinoma
TSG-OG	tumour suppressor – oncogene score
TRACERx	tracking cancer evolution through therapy (Rx)
WES	whole exome sequencing
WGD	whole genome doubling
WGA	whole genome amplification
WGS	whole genome sequencing

Chapter 1 Introduction

Cancer is now known to be a somatic evolutionary process, where populations of cells descended from a single common ancestor acquire mutations conferring beneficial phenotypic traits that allow them to clonally expand and outcompete less fit neighbouring cells.

These mutations are somatic variants: genetic alterations that occur in any of the cells of the body, apart from the germ cells. The length of nucleotide sequences affected by these somatic variants in cancer ranges enormously, from one nucleotide to entire chromosomes.

The categories these variants are organized into include single-nucleotide variants (SNVs) that may be silent, cause a missense or nonsense amino acid substitution, or affect the normal splicing of mRNA. Alternatively, one or more nucleotides may be involved in duplications, deletions, insertions, or complex patterns with a nucleotide(s) deletion combined with a nucleotide(s) insertion known as an INDELs.

Structural variants (SVs) which result from chromosomal instability and genomic rearrangements, include inversions, duplications, deletions, or translocations, and are often found in tumour DNA. Many SVs result in changes in the number of copies of sections of the genome and genes within them. These somatic copy number changes (SCNAs) may also occur without a structural variant if, often through whole genome doubling (WGD) or chromosomal missegregation events, a whole chromosome is gained or one of its homologous copies lost. Additionally, SVs may cause gene fusions and result in proteins with oncogenic properties, such as the *EML4-ALK* fusion that is seen in non-small cell lung cancer (NSCLC).

This knowledge has been gained through enormous advances in technology and study design over the last 50 years which have allowed the life history of a tumour to be reconstructed in greater and greater detail from the somatic mutations present within its genome, as well as revealing the great variety of mutational processes that may shape a cancer's development.

Since the inception and completion of the Human Genome Project, high-throughput technologies to investigate the cancer genome at greater and greater resolution have been developed. Each has then been applied as broadly as possible to large cohorts of single-region samples from many patients' disease in order to identify genes that,

when somatically altered, might “drive” a cancer’s growth and provide a treatment opportunity.

High-throughput Sanger sequencing revealed novel therapeutic targets such as the *BRAF* gene. Expression arrays allowed genetic classification of subtypes within breast cancer and other tumour types(Perou et al., 2000) while array comparative genomic hybridisation and then later the construction of the HapMap resource and development of high density SNP arrays allowed unprecedented characterization of SCNAs. Finally, the ability to interrogate the entire exome and even whole genome through next generation sequencing (NGS) has allowed comprehensive efforts to catalogue all mutations in a wide variety of adult and paediatric cancers by international consortia that have, together, sequenced tens of thousands of single region samples.

However in the same period smaller studies which either took multiple regions from the same tumour or focused on novel computational methods applied to single regions have demonstrated intratumour heterogeneity (ITH) of mutations and branched evolution: the presence of multiple divergent subpopulations of cancer cells, or subclones, within a single patient’s disease. In the main, these studies, whilst cataloguing chromosomal copy number changes and incorporating them into their analyses, have focused on the ITH and clonality of sequence level variants such as SNVs.

1.1.1 Chromosomal Instability and aneuploidy

The term aneuploidy describes the state of having an atypical, or non-euploid, number of chromosomes within a cell. This definition is often extended to distinguish the presence of so-called “numerical” aneuploidy that refers to abnormal numbers of intact chromosomes and “structural” aneuploidy to describe the presence of abnormal chromosomes themselves. For the purposes of this thesis the term aneuploidy will encompass both numerical and structural abnormalities unless otherwise specified.

In contrast to aneuploidy, chromosomal instability refers to process of ongoing errors in chromosomal segregation during anaphase that occur through consecutive mitoses. This distinction between the “state” measurement that is aneuploidy and the “rate” measurement of CIN is important as, though the two phenomena often occur together in cancer, there are examples of aneuploidies that are stable that exist without CIN such as in AML (Paulsson and Johansson, 2007).

The state of aneuploidy is relatively simple to assess from a wide variety of experimental techniques including: single-region bulk sample SNP array or next generation sequencing, single sample fluorescence in situ hybridization, or single sample karyotyping. These methods have been variously estimated aneuploidy to be present in 60-80% or approximately 90% of solid human tumours (Weaver and Cleveland, 2006) and over 50% of haematopoietic cancers (Mitelman, 2015). However, the presence of ongoing CIN in cancers assayed using these methods cannot be conclusively shown and has left debate in the field as to whether cancer development is predominantly characterized by early punctuated macro-evolutionary change followed by copy number-stasis (Gao et al., 2016), or continual karyotypic change.

Investigations that are powered to identify the ongoing rate of chromosome missegregation either directly observe CIN in cell anaphases or infer CIN by examining chromosomal differences between multiple samples of the same experiment or patient's disease that are spatially or temporally separated. The logical extension of this multi-sample or multi-region approach are single cell approaches that should be able to definitively identify chromosomal heterogeneity; however, currently there are limitations with this approach that will be covered in section 1.3.19.

1.1.2 CIN and early theories of cancer evolution

The earliest work that hinted at a role for genetic alterations in cancer was performed by von Hansemann in the late 19th century. His systematic descriptions of cytology in papers published from 1890 to 1906 noted abnormalities in tumour tissue during cell division and theorized that abnormal mitotic spindle formation resulted in the unequal distribution of chromosomes in tumour cells (von Hansemann, 1890). Hansemann's ideas were expanded upon by Theodor Boveri who observed aberrant mitoses while following cell division in the of a sea urchin egg and noted how they lead to abnormalities in the number of chromosomes of the resulting daughter cells. These findings lead Boveri to propose his "Chromosome Theory of Cancer" which suggested both that cancerous growth occurs in cells that acquire abnormal numbers of chromosomes from "asymmetrical mitosis" and that a cancer originates from single ancestral cell (Boveri, 1914).

Otto von Winge provided additional cytological evidence in support of Boveri's theory in two studies from 1927 and 1930. The first investigated the Crown-gall tumour that may affect beet and tobacco that is caused by the Bacterium tumefaciens, he noted

that the vast majority of cells had more than the normal complement of chromosomes. In the second study, he induced tumours in 80 mice with coal tar and noted that cells within the same tumour were heterogeneous, and contained 35-138 chromosomes (the normal mouse karyotype contains 40) (Winge, 1927, Winge, 1930). These findings lead Winge to further propose that multiple chromosomal alterations over generations could lead to disease progression.

However, it was only after DNA was demonstrated to be the molecule by which genetic information is encoded that it was demonstrated beyond doubt that genetic changes were involved in carcinogenesis with discovery of the Philadelphia chromosome (Nowell, 1960) and its link to chronic myelogenous leukaemia (CML). However, theoretical work had begun to provide a mutational theory of cancer. In 1953 Nordling described how cancer mortality increased "by a certain power (to the sixth) of age" and proposed a requirement of six mutations for tumour development (Nordling, 1953). Armitage and Doll also expanded upon this idea, proposing a multi-stage theory of cancer based on interpreting cancer incidence as a function of age (Armitage and Doll, 1954).

In 1971 Knudson compared patients with sporadic and familial retinoblastoma. He noted that patients with familial disease were far younger and thus must already have a germline mutation whereas the patients with the non-familial form of the disease would need two mutations and that this would explain the difference in onset between the two groups. This formed the basis of his famous “two-hit hypothesis” (Knudson, 1971). Genetic linkage studies following up on this theoretical basis examined chromosomal abnormalities in families affected by hereditary retinoblastoma and centred upon the genomic region of 13q14 (Yunis and Ramsay, 1978). Evidence for loss of heterozygosity (LOH) covering this genomic area in tumour cases from familial disease was then described (Cavenee et al., 1983, Sparkes et al., 1983, Murphree and Benedict, 1984, Cavenee et al., 1985), strongly supporting the two-hit hypothesis by suggesting that an abnormal chromosome was being inherited from one parent and the wild type chromosome from the other was then lost in tumour cells.

Many other groups had also theorised that all cells in a patient's disease must derive from a common ancestral cell based on how similar the karyotypes of individual cells from a tumour were (Makino, 1957, Hauschka, 1961, Ford and Clarke, 1963) and still others had observed that neoplasms could give rise to malignant tumours (Morson, 1974). Additionally, in agreement with Winge's earlier suggestion others proposed that tumours might progress as populations of tumour cells gained additional

mutations, which at this time were understood to be chromosomal alterations, by a process termed clonal evolution. This fit well with observations that sequences of karyotypes could be observed to replace each other in patients' disease as it progressed (Foulds, 1957, de Grouchy et al., 1966, Adam et al., 1970).

In 1976 Peter Nowell combined all of these observations into a unified description of tumour evolution where the cancer is subject to selection (Nowell, 1976). In his successive clonal evolution model, he posited that tumours arise from a single normal ancestral cell that acquires a growth advantage over its neighbours and forms a clonal expansion. As this clonal population proliferates it acquires mutations through chromosomal and genetic instability, selection then acts upon this genetic variability to successively select more aggressive and proliferative subclones.

1.1.3 The discovery of sequence level variants and their role in cancer

Inherent in Nowell's model of tumour evolution was the idea of a "driver" mutation that provided an advantage to cellular growth. However, while the role of chromosomal aberrations in certain tumour types were well established, particularly in haematological disease in which it was relatively easier to obtain metaphase spreads and interrogate them using Giemsa staining, the extent, role and methods of generation of mutations in the genetic sequence itself were unknown. Presciently, Cairns suggested that driver mutations might be the result of mistakes made in cell intrinsic renewal processes (Cairns, 1975). However, even he in 1981, though accepting that localized lesions in DNA could be carcinogenic, suggested that transposon activity was the more likely the source of the sequential mutations leading to human cancer (Cairns, 1981).

The first direct demonstration of that single change sequence could be an activating or so called "oncogenic" event (genes whose oncogenic effect is produced by only a single mutational "hit") was when Reddy et al. 1982 and Capon in 1983 demonstrated a single G to T transversion in the *HRAS* gene was capable of transforming cells and present in tumour samples (Reddy et al., 1982, Capon et al., 1983). Small variations in genetic sequence including in-frame deletions, and somatic single nucleotide variants (SNVs) were also shown to inactivate the *RB1* tumour suppressor by a multitude of studies (Lee et al., 1987, Friend et al., 1987, Fung et al., 1987, Horowitz et al., 1989, Jones et al., 1990, Shew et al., 1990).

It became clear that a plethora of genes were affected by likely sequence level mutations. By the mid 1980s at least 40 oncogenes had been described (Weinberg, 1985) and increasing numbers of tumour suppressors identified such as *VHL* (Tory et al., 1989), *TP53* (Nigro et al., 1989)*, APC* (Nishisho et al., 1991) amongst others.

The sequencing of human genome sequence was completed in draft form in 2001. In the process of accomplishing this enormous feat the price of Sanger sequencing fell 100-fold. This allowed targeted resequencing of many genes thought to potentially be involved in cancer. Such screens identified somatic SNVs of *BRAF* as a driver in melanoma (Davies et al., 2002) and of *EGFR* in NSCLC(Lynch et al., 2004).

However, the real paradigm shift came with the advent of massively parallel or next generation sequencing (NGS) that became commercially available from 2005 onwards, which involved and allowed use of short read shotgun sequencing. Initially, due to technical limitations, studies focused on assaying the entire exome (protein coding region of the genome) and were able to reveal SNVs and short insertions and deletions in the genetic sequence.

These early studies revealed that some genes seemed to be recurrently affected by mutations while others seem to be affected far less (Wood et al., 2007, Ding et al., 2008). This pattern was again confirmed by the first pilot study of The Cancer Genome Atlas (TCGA) which examined the exomes of over 200 glioblastoma (GBM) cases. The gene *PIK3R1* was identified as potential driver due the number of cases affected by SNVs in comparison the relative paucity of mutations in other genes. In the following years the TCGA and other groups continued to release studies of an enormous number of tumour types analysed with more and more advanced algorithms detailing the range of SNVs across tumour types. These studies continued to identify so-called driver genes whilst relegating most SNVs observed in each patient's disease to the status of "passenger" mutations thought to have little effect on fitness of the cancer cells.

1.2 Next generation sequencing reveals intratumour heterogeneity

In the early 1800s, German physiologist Johannes Muller applied microscopy to human tumour samples. Until this point all knowledge of tumours had been gathered from observations made with the naked eye. Muller and his student, the famous German pathologist Rudolf Virchow, characterised and distinguished cancer

subtypes and noted heterogeneity within single tumours (Balkwill and Mantovani, 2001, Hajdu, 2004).

The first application of NGS techniques to cancer that revealed ITH was when subclonal populations of tumour cells were demonstrated through the diversity of sequence variation in the Ig heavy chain locus in 22 cases of chronic lymphocytic leukaemia (CLL). This locus was chosen as it was known that CLL patients demonstrate hypermutation in this region and given the depth of the sequencing and length of the reads used, it was possible to create and assemble phased tumour haplotypes that determined which sequence variants were likely to be present within the same subclonal set of cells. Phylogenetic relationships between these tumour haplotypes could be constructed, and demonstrated the subclonal nature of the patients' disease (Campbell et al., 2008).

1.2.1 Bioinformatics methods applied to single and multi-region studies reveal SNV ITH

The first group to demonstrate SNV ITH from single region samples were Nik-Zainal et. al 2012 through the development and application of a major bioinformatics advance. The group subjected 21 BRCA tumours to whole genome sequencing (WGS) which drastically increased the number SNVs detected. This enabled the clonal architecture of the tumour to be determined by clustering the SNVs by their prevalence within the tumour, measured as their cancer cell fraction (CCF), into the likely subpopulations (or subclones) of cancer cells present that carried them. Later, similar versions of this Dirichlet clustering approach were also implemented to cluster deeply sequenced SNVs revealed by targeted sequencing and was first applied to a single region BRCA TN cases (Shah et al., 2012) and NSCLC (Govindan et al., 2012). A stand-alone bioinformatics tool called PyClone that required less stringent coverage working with depths (> 100x) commonly found in WES was released in 2014 (Roth et al., 2014) though other groups had by this stage also begun applying CCF based clustering techniques to WES data from CML (Landau et al., 2013) and multiple myeloma (Lohr et al., 2014, Bolli et al., 2014).

Earlier in 2012, another approach to identifying ITH was demonstrated using WES: multi-region NGS. Gerlinger and colleagues showed that by taking multiple spatially separate regions from the same KIRC tumour and sequencing them independently they could demonstrate heterogeneous SNVs. This approach revealed ubiquitous SNVs present in all tumour regions, SNVs present in a subset of regions and finally

private SNVs present only in a single region (Gerlinger et al., 2012). The presence of different SNVs in distinct regions demonstrated the presence of sub-clonal populations of cancer cells.

One advantage of this approach of considering SNVs in a binary manner as either present or absent in different regions compared to the clustering of their CCFs is that it requires far fewer SNVs. Multi-region studies adopting this approach were able to characterise ITH in tumour types with a lower mutation rate or using sequencing modalities such as WES or targeted panel sequencing that examine far less of the genome than WGS. This was particularly important before the advent of bioinformatics tools such as PyClone designed perform clustering on high depth WES data (Roth et al., 2014).

Multi-region study designs are also able to reveal ITH that cannot be described from single regions alone. Somatic alterations present in one region of a patient's tumour but completely absent from another would either be entirely missed by a single region approach or potentially considered to be clonal, present in all cancer cells of a patient's disease depending on which region was assayed. Such events have been said to demonstrate the "illusion of clonality" (IOC) and emphasise the importance the multi-region approach (Jamal-Hanjani et al., 2017).

1.2.2 Modes of evolution revealed by SNV ITH

In Nowell's 1976 theory, tumourigenesis was thought to be linear, proceeding in success of clonal sweeps as more and more aggressive subclones developed. However, the more recent observations from NGS single and multi-region studies already described showing multiple subclones with distinct sets of SNVs present in primary tumours do not fit with a linear progression model but rather that the "divergent cancer clone genotypes and phenotypes correspond to allopatric speciation in separate natural habitats" (Greaves and Maley, 2012).

1.2.3 Branched Evolution revealed by SNVs

It is important to note that without information on the mutual exclusivity of SNVs, the variants described by any single-region sampling approach could be explained by linear evolution: unless a subset of mutations are mutually exclusive it would always be theoretically possible for one clone to be a less (or at most equally) prevalent descendant of another and containing all its ancestors' SNVs. The now predominant

view of “branched” tumour evolution with multiple diverging subclones present in the same patient’s disease has been supported by three sources of mutual exclusivity:

1) Phasing of SNVs: phasing is generally defined as the process of assigning alleles to maternal or paternal chromosomes. However, SNVs can also be phased relative to one another. If using NGS, any two SNVs can be shown to occur in the same NGS read (which must originate from a single contiguous DNA fragment) they must therefore have occurred within the same cell and thus lineage. If the two SNVs are never present on the same read then they are likely to be mutually exclusive, never having been present in the same cell and demonstrate distinct, branched tumour evolution. However, this approach is limited to cases in which SNVs are close enough to one another to be covered by the same NGS read and has therefore been applied in WGS (Nik-Zainal et al., 2012) that by assaying the entire genome has a greater chance to detect these rare cases. Another application is in tumour types or areas of the genome known to be affected by hypermutation (Campbell et al., 2008).

2) Single cell sequencing: numerous studies have in which SNVs have been profiled by single cell DNA sequencing and found to form distinct lineages in (Navin et al., 2011, Hou et al., 2012, Li et al., 2012b, Yu et al., 2014, Gawad et al., 2014, Wang et al., 2014b, Gao et al., 2016).

3) Multi-region sequencing revealing distinct sets of SNVs: these studies demonstrating ITH of SNVs have consistently identified SNVs present in all regions and others only in mutually exclusive subsets of regions assayed across cancer types.

The multi-region NGS approach has now been applied to relatively small cohorts in many cancer types and with differing study designs. Tumour types examined with this method and found to demonstrate branched evolution include: BLCA (Thomsen et al., 2016, Lamy et al., 2016), ESCA (Murugaesu et al., 2015, Findlay et al., 2016), neuroblastoma (Eleveld et al., 2015), GBMLGG (Suzuki et al., 2015), SKCM (Sanborn et al., 2015, Harbst et al., 2016, Reuben et al., 2017), COAD (Kim et al., 2015, Sottoriva et al., 2015, Uchi et al., 2016), HNSC (Hedberg et al., 2016), STAD (Lim et al., 2016) BRCA (Snyder et al., 2014, Savas et al., 2016), UCEC (Gibson et al., 2016), OVCA (Bashashati et al., 2013), NSCLC (Zhang et al., 2014, de Bruin et al., 2014).

A review of the multi-region sequencing literature showed that, in spite of caveats including differences between the number of regions assayed, primary and metastatic sites, stage and sequencing depth these studies have revealed mutational heterogeneity and branched evolution appears to occur across cancer types (McGranahan and Swanton, 2017). However, the degree of SNV ITH can be highly variable, between 0 and more than 8,000 thousand coding mutations were found to be heterogeneous between primary and metastatic sites in GBMLGG (Johnson et al., 2014).

1.2.4 SNV ITH and inference of subclonality allows the construction of tumour phylogenies

The ITH discovered by both single-region and multi-region methods can be used to reconstruct the life history of a tumour using phylogenetic methods (Beerenwinkel et al., 2014). The resulting phylogenetic trees may then be used to time events in cancer evolution: those events that are clonal (thought to be present in all cancer cells) are early and form the trunk of the tree whilst those that are subclonal (thought to be present in only a subset of cancer cells) are thought to occur later in the tumour's evolution and appear in the branches.

Initial approaches applied WES multi-region sequencing involved treating the presence or absence of SNVs as a binary variable (essentially ignoring their VAF and CCF) on a region by region basis and the use of maximum parsimony approaches to create the phylogenetic trees (Gerlinger et al., 2012, Gerlinger et al., 2014, Murugaesu et al., 2015, Harbst et al., 2016). However, this approach could only highlight differences between regions and could not assign the SNVs themselves to individual subclones, indeed because of this some groups argue they are not true phylogenies at all (Alves et al., 2017).

While there are many tools designed to create phylogenies based on subclones derived from CCF clustering, they all rely on two principles: the infinite sites assumption (ISA) (Kimura, 1969) and the pigeonhole principle. The ISA implies that there is no loss of SNVs and that each mutation occurs at a novel site. The pigeonhole principle, in its simplest case states that if the total of the CCFs of two SNVs is greater than 100%, at least one cell must have contained both SNVs, and because the same mutation cannot be gained twice independently by the ISA, one subclone must be the ancestor of the other.

A large number of phylogeny building algorithms using these two principles have been designed to arrange subclones into phylogenies (Strino et al., 2013, Hajirasouliha et al., 2014, Jiao et al., 2014, El-Kebir et al., 2015, Deshwar et al., 2015, Malikic et al., 2015, Popic et al., 2015, Yuan et al., 2015). This approach has also extended and applied to multi-region WGS or high depth exome sequencing, combining the ability to identify subclones within each region with the mutual exclusivity information from multi-region sequencing to identify subclones that could be present across multiple regions or only make up a small fraction of the cancer cells in a single region (Yates et al., 2015, Gundem et al., 2015, McPherson et al., 2016, Jamal-Hanjani et al., 2017, Yates et al., 2017). Our group's work in analysing data from 100 patients' NSCLC disease also took this hybrid approach, utilizing high mutation rate of NSCLC and the high depth 400x exome sequencing to allow the application of a modified version of PyClone (Jamal-Hanjani et al., 2017).

1.2.5 Timing of SNVs

The inference of clonality and the construction of these phylogenetic or pseudo-phylogenetic trees have allowed the relative timing of events in tumour evolution to be established. These life histories of individual tumours can be combined to identify evolutionary subtypes (Turajlic et al., 2018b) and characterize particular events as potentially being important to overcome either early or late constraints on the cancer's progression. The relative timing of events to one another is also important to establish given that it is now known that the order in which mutations occur may affect the result of the selective pressures that a tumour undergoes, to restrict the evolutionary possibilities open to it (Papaemmanuil et al., 2013) as well as influence the clinical presentation of disease and its response to therapy (Ortmann et al., 2015).

Analysis of the single region data available for nine cancer types from the TCGA using PyClone found that potentially clinically actionable non-synonymous SNVs were in large part clonal, but, could also be observed to occur subclonally (McGranahan et al., 2015). Multi-region studies have also revealed that certain driver mutations are almost always clonal in certain cancer types but often subclonal in others. One example of this cancer type specific timing of a driver event is TP53 mutation. This has been found to be clonal in the majority of cancer types including NSCLC (Zhang et al., 2014, de Bruin et al., 2014) and OVCA (Bashashati et al., 2013) but in CLL and KIRC was found to often occur subclonally (Landau et al., 2013, Gerlinger et al., 2014). A review of other driver mutations classified as early or late

by phylogenetic means from a multitude of smaller multi-region studies already mentioned in this thesis as well as others can be found here (Turajlic et al., 2015).

1.2.6 Parallel evolution revealed by SNVs

Multi-region NGS studies, though relatively few in number compared to those investigating only single regions have repeatedly demonstrated instances of the same gene being affected by SNVs in distinct regions of a patient's disease. As these mutations cannot be present within the same tumour subclones they must have occurred independently. This has been described using terminology from evolutionary biology as parallel (or convergent) evolution where similar traits develop in related but distinct species each descended from a common ancestor. These parallel events may demonstrate constraints on later tumour evolution as their statistically improbable occurrence suggests selection pressures for particular cellular pathways to be perturbed. It is also important to note that identifying instances of parallel evolution of somatic events from single tumour regions is often infeasible as some evidence of the events' mutual exclusivity in subclonal populations is required to demonstrate that one is not simply the descendant of the other.

These parallel SNV events have been found to occur in primary tumours, with parallel SNVs found in *KDM5C* and *PTEN* as well as multiple inactivating mutations in *SETD2* in a multi-region study of clear cell renal cell carcinoma (KIRC) (Gerlinger et al., 2012). Parallel evolution was subsequently observed to affect *PIK3CA* in separate follow up study of the same tumour type (Gerlinger et al., 2014). Other tumour types documented to have been affected by sequence level parallel evolution in multi-region or in samples from multiple timepoints include breast cancer (Brastianos et al., 2015, Juric et al., 2015, Spoerke et al., 2016), colorectal cancer (Misale et al., 2012, Brannon et al., 2014), glioma (Suzuki et al., 2015), melanoma (Shi et al., 2014), multiple myeloma (Bolli et al., 2014), oesophageal adenocarcinoma (Murugaesu et al., 2015) and are comprehensively reviewed in (Venkatesan et al., 2017).

1.2.7 Evolutionary and clinical Importance of ITH revealed by SNVs

While branched evolution inferred from SNV ITH has now been described across a large proportion of cancer types, the exact cause remains the subject of heated debate. Indeed, work first performed in colorectal cancer (Sottoriva et al., 2015, Uchi et al., 2016) and subsequently pan-cancer (Williams et al., 2016) has suggested subclonal evolutionary dynamics might be best characterized by neutral growth,

whilst other groups have argued it is likely the result of positive selection (Tarabichi et al., 2018, Balaparya and De, 2018). However, parallel evolutionary events, converging upon the same biological endpoint (Jamal-Hanjani et al., 2017), and subclonal selective sweeps resulting in an illusion of clonality (Sun et al., 2017, de Bruin et al., 2014) are strong indicators of positive selection, but their detection from single samples remains challenging. Despite this fact, all pan-cancer studies of ITH to date have relied on single region analyses.

Andor and colleagues examined twelve cancer types from the TCGA and reported identifying at least two subclones with a CCF of >10% in 86% of cases, and that the size of the observed subclones correlated with survival (Andor et al., 2016); this supported an earlier single tumour type study identifying association between a measure of ITH and survival in HNSCC (Mroz and Rocco, 2013). More recently the PCAWG consortium has reported a pan-cancer ITH analysis of single region WGS data from 36 cancer types and, while not commenting on any potential relationship between ITH and survival, they have argued that 95.1% of cases demonstrated evidence of recent subclonal expansions as well as showing clear signs of positive selection in both clonal and subclonal SNVs (Dentro et al., 2018).

Regardless of whether ITH and subclonality in primary untreated disease indicates neutral or positive selection in cancer evolution, its clinical relevance is in no doubt as it provides the genetic variation that selection may act upon leading to resistance to therapy. Treatment often engenders rapid tumour evolution (Landau et al., 2013, Landau et al., 2015) and the presence of a subclonal driver mutation alone has been observed to be an independent risk factor for progression (Landau et al., 2013). A pan-cancer study of SNV clonality revealed that while driver genes typically occurred early in cancer evolution, subclonal "actionable" mutations that might affect efficacy of targeted therapy approaches were identified in all 9 tumour types investigated (McGranahan et al., 2015). Finally, there are numerous examples of genomic changes in response to both chemotherapeutics (Ding et al., 2012, Schuh et al., 2012, Johnson et al., 2014, Murugaesu et al., 2015), and therapies targeted to particular genomic alterations (Diaz et al., 2012, Shah et al., 2012, Misale et al., 2014, Shi et al., 2014, Bettegowda et al., 2014). Such targeted therapies have even been documented to result in parallel evolution of resistance mutations (Juric et al., 2015) suggesting that such parallel events are the result of selective pressures. ITH has also been shown to influence the efficacy of immunotherapy with a relationship

described between clonal neoantigen burden and overall survival in in primary LUAD cases (McGranahan et al., 2016).

ITH has also been linked to progression with studies of Barrett's oesophagus, a precursor for ESCA, linked higher subclonal diversity to an increased risk of progression to malignancy (Maley et al., 2006, Merlo et al., 2010). In a later extension of this finding Ross-Innes and colleagues found that by the time of diagnosis the proportion of mutations shared between the BE and ESCA may be under 20% (Ross-Innes et al., 2015). This subclonal diversity is not only limited to progression from a pre-cancerous state to malignancy but can also be observed in untreated disease such as been observed in longitudinal sequencing data from CLL which demonstrated changes in selection pressures and clonal dynamics over time (Nadeu et al., 2016).

1.3 Detection of somatic copy number alterations

1.3.1 What is an SCNA?

Throughout this thesis the term somatic copy number alterations (SCNA) is used to refer to somatic changes in the number of copies of a DNA sequence that arise during the process of cancer development. Therefore, an SCNA is a sequence of DNA that is found at different copy numbers in an individual's germline DNA compared to the DNA present in a subpopulation of somatic cells. The term copy number variant (CNV) is sometimes incorrectly used to describe SCNAs, when it more properly describes normal copy number variation of sections of the genome present in the germline DNA of multiple individuals' genomes.

1.3.2 History of SCNA detection

Aneuploidy and karyotypic changes were the first genetic alterations described in cancer. These are generated through chromosomal instability (CIN) that occurs at high frequencies across many cancer types. CIN describes an increased rate in variation in chromosome number or structure, the result of which are karyotypic abnormalities including whole chromosome and segmental aneuploidies as well as translocations, inversions and deletions.

These events were first described through cytology where light microscopy is used to visualize chromosomes. However, without any treatment, structural details of chromosomes were found to be difficult to detect using a light microscopy. Indeed,

the correct number of human chromosomes was mistakenly thought to be 48 as described by Theophilus Painter for over thirty years until 1956 when Hin Tijo and Albert Levan reported the correct number was 46 (Painter, 1923, Tjio and Levan, 1956). The most important aspect of this finding was not that the correct number was established but rather that now with the technique of colchicine to arrest the dividing cells in metaphase and hypotonic shock to spread cells' nuclear contents, it was possible to distinguish the different chromosomes with a degree of confidence. In 1960 Peter Nowell discovered that phyohaemaglutinin simulated white blood cells to divide (Nowell, 1960), greatly advancing the field of cytogenetics. In the same year, he and Peter Hungerford described the Philadelphia chromosome which was the first specific chromosomal change to be associated with a human cancer (Nowell and Hungerford, 1960). Torbjorn Caspersson's group introduced the technique of Giemsa chromosome banding (or G-banding) that allowed perfect discrimination between chromosomes and most importantly allowed identification of chromosomes involved in structural rearrangements (Caspersson et al., 1970). It was with G-banding that Janet Rowley demonstrated that the Philadelphia chromosome was in fact a translocation between chromosomes 9 and 22 and not a simple deletion of chromosome 22 (Rowley, 1973).

The first probes to identify single genes in cells were designed well before the completion of the human genome project. Mouse human somatic cell hybrids which lost human chromosomes preferentially allowed human gene mapping by correlating the gene expression and chromosomal patterns present (Harris and Watkins, Weiss and Green, 1967). Finally, when gene isolation became possible, labelled probes were hybridized to human chromosome metaphase spreads and were able to mark a single gene (Morton et al., 1984) using fluorescent in situ hybridisation (FISH). Fluorescent "painting" of entire chromosomes (Lichter et al., 1988) and finally multicolour fluorescent in situ hybridization (M-FISH) (Speicher et al., 1996) and spectral karyotyping (SKY) (Schrock et al., 1996) allowed every chromosome to be visualised with a distinct colour.

The first technology allowing the high-throughput assessment of SCNAs was comparative genomic hybridization (CGH) (Kallioniemi et al., 1992). While this technology allowed the detection of alterations in total copy number through log ratio (LogR), the $\log_2$ ratio of fluorescence between the normal and test DNA using oligonucleotide probes, which could be as short as 100KB (Pinkel and Albertson, 2005), it could not detect allelic imbalance or loss of heterozygosity. The first relatively

high-throughput method to assess AI or LOH, was to examine a limited number of microsatellite markers with PCR fragment size analysis. However, this technique could not be used to assay the entire genome (Zenklusen et al., 1994, Negrini et al., 1995).

Array CGH was then used in combination with expression microarrays to demonstrate that SCNAs influenced the expression levels of the genes present within them. Pollack and colleagues showed that a 2-fold change in DNA copy number corresponded to a 1.5-fold change in expression (Pollack et al., 2002). The next advance in high-throughput SCNA detection were SNP arrays that combined measures of B-allele frequency of heterozygous SNPs with LogR measurement which allowed detection of total copy number as well as AI and LOH were then introduced (Pfeifer et al., 2007).

The first pan-cancer study of SCNAs utilized SNP arrays to report on over 3000 single region bulk samples from 26 cancer types (Beroukhim et al., 2010) and was followed by a TCGA consortium effort that also used SNP arrays to characterize SNCAs in 4934 single region bulk samples (Zack et al., 2013). These studies catalogued the variety of SCNAs across cancer types, demonstrating LOH, copy number gains and losses as well as high level amplifications and homozygous deletions across cancer types for the first time.

During this period, a plethora of bioinformatics tools to call SCNAs from SNP array data were published. However, NGS technologies soon became available (Metzker, 2010) and though initially focused on the detection of sequence level variants, NGS was also used to detect SCNAs through LogR measurements derived from the relative depths of NGS reads in tumour versus normal samples and BAF and coverage of heterozygous SNPs (Zare et al., 2017). In this thesis LogR is used to describe the comparison of signal intensity between oligonucleotide probes in array CGH, allele-specific oligonucleotide probes in SNP arrays and the relative sequencing depths observed between normal (or “germline” DNA) versus tumour DNA in the context of NGS.

WES and some larger targeted sequencing panels have similar capabilities to SNP arrays although they most often have fewer “probe” equivalents than larger SNP arrays such as Affymetrix SNP6. SNP arrays, WES and targeted sequencing, though able to quantify SCNAs, cannot accurately describe the majority of structural variants as the most SVs occur in intergenic regions. However, in an advance to relative to

SNP arrays, they are able to identify SVs in exonic regions that result in fusion proteins such as EML-ALK4.

Some groups adapted and applied tools originally designed for SNP array SCNA calling e.g. ascatNGS (Zheng et al., 2016) but many WES specific SCNA/CNV callers were developed such as CNVer (Medvedev et al., 2010), ExomeCNV (Sathirapongsasuti et al., 2011), exomeCopy (Love et al., 2011), CONTRA (Li et al., 2012a) , CoNIFER (Krumm et al., 2012), ExomeDepth (Plagnol et al., 2012), VarScan 2 (Koboldt et al., 2012), XHMM (Fromer et al., 2012), ngCGH (Gartner et al., 2012), EXCAVATOR (Magi et al., 2013), CANOES (Backenroth et al., 2014), PatternCNV (Wang et al., 2014a), Control-FREEC (Boeva et al., 2012), cn.MOPS (Klambauer et al., 2012), CODEX (Jiang et al., 2015), Sequenza (Favero et al., 2015) and FACETs (Shen and Seshan, 2016) Including some which utilized off-target reads, NGS reads that fall outside the targeted genomic areas in exome or targeted sequencing, including cnvOffSeq (Bellos and Coin, 2014), copyWriteR (Kuilman et al., 2015) and cnvkit (Talevich et al., 2016).

As the price of NGS decreased, WGS became a viable technology with which to examine tumours. A novel set of algorithms were required to infer copy number from WGS which, covering the whole genome, provided a much greater quantity of data compared to WES which covers only 1-2% of the genome. Unlike arrayCGH, SNP arrays, targeted sequencing and WES, WGS is able to identify the exact breakpoints of SVs in both exonic and intergenic regions. This whole genome coverage also allowed both greater statistical power in SCNA inference and, as has been used in a subset of copy number calling tools, the phasing of heterozygous SNPs using sets of reference haplotypes and statistical imputation. This phasing allows greater accuracy in subclonal copy number calling by producing very accurate BAF estimation (Nik-Zainal et al., 2012). Copy number calling tools specifically designed for WGS that utilize phasing include Battenberg (Nik-Zainal et al., 2012) and ReMixT (McPherson et al., 2017) which also incorporates SVs called from WGS data. Other copy number calling tools designed for WGS that do not use phasing include TITAN (Ha et al., 2014) and ACEseq (Kleinheinz et al., 2017) which like ReMixT also incorporates SV calling.

1.3.3 SCNAs and structural variants

Structural variants and SCNAs are inherently connected, with unbalanced SVs producing SCNAs, and loss or gain of “neochromosomes” created by SVs joining

sections of normally separate chromosomes causing SCNAs appearing to affect sections of the chromosomes from which the genetic material they carry originates. However, the two categories of somatic event are not synonymous, as not all SVs will result in SCNAs: inversions and balanced translocations will create no change in overall somatic copy number. Similarly, not all SCNAs will be linked to an SV: whole chromosome missegregations will cause the copy number of intact chromosomes in the genome to be altered with no associated structural variants.

1.3.4 SCNA terminology

A normal human cell is diploid with a complement of 46 chromosomes, that is, it has two copies of each of the 22 autosomes (non-sex chromosomes) and either two X chromosomes in females or a single X and a single Y chromosome in the case of males. Each pair of chromosomes consists of two homologous chromosomes: one maternal and one paternal in origin. These two homologous chromosomes will differ from the reference genome in genetic sequence with each having its own single nucleotide polymorphisms (SNPs) among other classes of genetic variants such as CNVs and polymorphic microsatellites (a tract of repetitive DNA in which certain patterns from 1 to 10s of base pairs are repeated between, on average, 5 and 50 times (Gulcher, 2012)). In a single individual, SNPs that are present on both the paternal and maternal homologous chromosomes are referred to as homozygous while those present on only one homologous chromosome are termed heterozygous SNPs. These heterozygous SNPs are can be picked up by some assays and can be used to assess the which of the two parental chromosomes are affected by an SCNA. All cells discussed in the rest of this thesis are human and can be assumed to have, or originate from, a cell with the normal diploid complement of 46 chromosomes.

However, cancer cells are often aneuploid, having an abnormal number of chromosomes. As already discussed, this aneuploidy is very often not simply variations in the numbers of whole chromosomes but may also include gross structural alterations of parts of chromosomes. The field of cytogenetics has a codified and detailed nomenclature for describing these copy number changes and structural chromosomal alterations in single cells: the international system for human cytogenetic nomenclature (ISCN) that has undergone many revisions since it was first introduced in 1960 (Robinson, 1960). However, this naming system is not applied to cancer samples that are assayed by non-cytological methods and a new and uncodified lexicon of terms has been used to describe SCNAs. Indeed, the language employed and level of detail in which SCNAs are described is heavily dependent on

both the assay used to assess copy number changes and the bioinformatics approaches applied.

Nonetheless, in the post-NGS era of cancer genomics there are some general terms that are broadly used. In the study of genetics, the ploidy of a cell refers to the number of complete sets of chromosomes it possesses, with the states of monoploidy, diploidy, triploidy and tetraploidy referring to 1, 2, 3 and 4 complete sets of chromosomes. In cancer genomics, the term has been adapted: if a diploid state is considered to have two copies of the haploid genome, an aneuploid cell with uneven numbers of copies of chromosomes or sections of chromosomes across the genome can be thought to represent a non-integer factor of the haploid genome and therefore a non-integer ploidy. Indeed, in cancer genomics the inferred ploidy of a cancer cell is often calculated as the mean copy number across the genome (normally weighted by chromosome size), resulting in non-integer ploidies. The allele specific copy number estimates across the genome produced by the copy number calling algorithms such as ASCAT (Van Loo et al., 2010) used in this thesis are termed the copy number profiles (CNP). Perhaps confusingly these cells, while recognized as being aneuploid and likely affected by both SCNAs and SVs and thus likely to have non-integer ploidies, are often referred to by the closest integer ploidy obtained by rounding, so cancer cells with ploidies of 2.6 and 3.4 may be considered “triploid”.

In addition, in cancer genomics somatic copy number estimation, the genome is considered in terms of sections termed “segments”. These segments refer to a stretch of contiguous genome in the genomic architecture of a normal diploid cell and consist of three values: the chromosome as well as the start and positions the segment occupies of it. The format that these are generally referred to in and will be referred to in this thesis is in the order of “chromosome”, “start” and “end”, for example: chr1:100-200. It is important to note that unless combined with structural variant calls determined by bioinformatics tools for this purpose, the genomic positions these segments refer to may not be correct; the genetic material to which they refer could have been rearranged and be part of neochromosomes bearing little relationship to the original normal diploid genome. There are five attributes generally used to describe the copy number state of any section or “segment” of the genome:

1) Total copy number (TCN)
   The total number of copies of the genome present at a genomic position with no reference to whether they are the same or different homologous

chromosomes. In a normal diploid cell this will be always be two, unless referring to the sex chromosomes in a male.

2) Allele-specific copy number (ASCN)

The number of copies of the genome present at a genomic position from each of the two homologous chromosomes originally present in the last normal diploid ancestral cell. These allele specific copy number measures are determined by exploiting changes in the ratio of heterozygous genetic variants from the normal 1:1 caused by SCNAs affecting one of the homologous chromosomes. These two separate counts are often referred to as the “major” or “A allele” and “minor” or “B allele” copy number as, unless trio sequencing (sequencing of the patient, their mother and father) has been performed, which homologous chromosome is paternal or maternal cannot be determined.

If, due to an SCNA, there is more of part of the genome from one homologous chromosome than the other at one position in the genome, the more common chromosome is referred to as the “major” allele and the less common as the “minor” allele. If the two homologous chromosomes occur at an equal frequency then one is arbitrarily referred to as the major and the other the minor.

In a diploid normal cell, when referring to autosomes, the ratio of heterozygous SNPs to reference allele will be 1:1 and the allele specific copy number will always be 1:1, that is 1 major copy and 1 minor copy.

3) Whether AI (AI) is present

Allelic imbalance refers to when, at a genomic position, there is more of one homologous chromosome present than the other e.g. there may be two copies of the paternal chromosome present and only one of the maternal.

In a normal diploid cell this should not occur on at any genomic locus in the autosomes as there should be one homologous chromosome of maternal origin and one of paternal origin. This is, in somatic copy number terminology, an allelically balanced state. Allelically balanced states may co-occur with SCNAs e.g. there may be two copies of genetic material from the maternal chromosome and two copies of the paternal chromosome at a genomic

position giving a total copy number state of 4 and allele specific copy number state of 2:2. However, in a normal diploid cell, in the context of autosomes and the X chromosome in females, no AI will be present.

4) Whether loss of heterozygosity (LOH) is present
   LOH is a subset of AI when only genetic material from a single parental chromosome is present with the other reduced to zero copies.

5) Whether a homozygous deletion (HD) is present
   A homozygous deletion is when zero copies of either the paternal or maternal genetic material are present at a locus in the genome, giving it a total copy number state of zero and an allele specific copy number state of 0:0. This is an allelically balanced state.

1.3.5 SCNA amplitude: gains, amplifications, losses and deletions

SCNAs were first described using cytogenetic techniques, and in the corresponding literature a copy number gain generally referred to an increase of one copy or a small number of copies of a section of DNA that would typically span a large genomic region. An amplification referred to a smaller or more focal section of DNA that reached much higher copy numbers. A similar distinction was applied to “losses” and “deletions”, with losses referring to large-scale events and deletions to smaller focal events (Beroukhim et al., 2010).

These widely used if inexact general terms were used to discuss patterns observed over many cases, but also co-existed with a very specific codified system for describing cytogenetic findings in single cells: the international system for human cytogenetic nomenclature (ISCN) that has undergone many revisions since it was first introduced in 1960 (Robinson, 1960).

1.3.6 Gains and losses in the context of bulk samples

When describing copy number in single cells, all copy number states must necessarily be integers, for example a gain of 1 copy of chromosome 2 in a single normal diploid cell would lead to there being 3 copies of chromosome 2 in a now aneuploid cell. These integer states are often described as absolute (Carter et al., 2012) and, ideally, in order to characterise the cancer genome, it would be possible

to infer these integer states for all cancer cells present in any tumour sample analysed.

However, while cytogenetic techniques examine single cells, nearly all high-throughput methods of quantifying SCNAs such as arrayCGH, SNP arrays, and various forms of NGS (WES, WGS and targeted panel sequencing) all analyse bulk tumour samples. The nature of these bulk samples, which are homogenized to produce DNA for analysis, makes inference of the integer copy number states of individual cells impossible. Even estimating the average integer copy number state (or ploidy) of cancer cells within such a bulk sample is technically challenging for three reasons:

1) A bulk sample is admixture of unknown proportions of normal diploid cells and cancer cells.
2) The ploidy of the cancer cells themselves is unknown and will be the result of CIN causing both large and small-scale chromosomal alterations as well as potential whole genome doubling events.
3) There may be heterogeneous subclonal populations of cancer cells that are of different aneuploid states (see section 1.3.9).

Due to these difficulties, the early analysis of bulk tumour samples with high-throughput methods was limited to calling relative gain and loss compared either to matched normal samples or a modality based internal reference. In addition, those modalities, such as SNP array and NGS, that assayed heterozygous SNPs while able to highlight areas of AI are unable to identify LOH as the relative gain and loss classifications cannot indicate whether there is a total absence of one parental allele in a section of the genome, only that there is an imbalance in the copy number of the two parental alleles.

Alternatives that could infer absolute copy number without the need to surmount these challenges of bulk sample interpretation include the aforementioned cytogenetics approaches or single cell sequencing. However, while these techniques are rapidly improving they have not been found to sufficiently high-throughput to allow the investigation of the tens of thousands of tumour samples that have been analysed by international consortia.

1.3.7 The importance of absolute integer copy number estimation

The proportion of a bulk sample consisting of cancer cells is often termed its purity and the mean copy number state of the cancer cells present in the bulk sample the ploidy. Relative gains or losses caused by SCNAs that are quantified in terms of the ploidy of a normal diploid sample cannot not give integer copy number estimates if either the overall ploidy of the cancer cells differs from diploid or the purity is not correctly estimated.

Purity also has an important effect on identifying gains and losses in a bulk tumour sample: if the sample is of low purity, then the effects of gain or loss events will be subtler in the LogR or equivalent measure than if it were higher purity. Therefore, it will be more difficult to observe a gain or loss event in lower purity sample whichever method of detection is employed. This effect is accentuated if the same relative thresholds based on a diploid normal sample are used for low and high purity tumour samples and the results compared. Absolute integer copy number estimates reduce the effect of this issue as, in order to generate them, purity must be taken into account, making the absolute integer results from tumour samples of differing purities comparable.

Another important class of SCNA event in tumour evolution besides gains and losses is LOH, an integral aspect of Knudson's two hit hypothesis. The AI events revealed by high-throughput methods that assay heterozygous SNPs can only be further subclassified as LOH by incorporating purity estimates to give absolute integer allele-specific copy number calls. These absolute integer calls would show whether a section of the genome has zero copies of one parental allele demonstrating LOH or, despite there being AI, there are still one or more copies remaining of the less prevalent parental allele demonstrating the absence of LOH but presence of AI.

Therefore, without absolute integer copy number estimation, purity may confound analyses attempting to identify common gains or losses across tumour samples and without allele specific absolute integer copy number estimation LOH cannot be reliably identified. In addition, without an estimation of the ploidy of the cancer cells within a tumour sample, integer copy number estimates cannot be made. Therefore, in order to obtain absolute integer copy number estimates both purity and ploidy must be known or estimated.

1.3.8 Inferring purity and ploidy from bulk samples

Methods of estimating purity from tumour samples include the visual assessment of the proportion of tumour cells by pathologist or computational estimation from histological slides. An orthogonal approach is to use laser capture microdissection to specifically enrich for tumour cells in sample before it is assayed, potentially removing the need for the inference of purity. Tumour ploidy can be estimated in a non-computational manner using cytological methods or FACS ploidy estimation of material from the same area of the tumour that will be used to provide the bulk sample to be analysed. Estimates of purity and ploidy from any of these sources may be combined to produce absolute integer copy number.

However, the most widely adopted solution is to use same computational method to jointly infer allele-specific absolute integer copy number, purity and ploidy from a bulk of single region using copy number calling algorithms. Published algorithms taking this approach include those designed for analysis of SNP array data such as GAP (Popova et al., 2009), PICNIC (Greenman et al., 2010), ASCAT (Van Loo et al., 2010) and ABSOLUTE (Carter et al., 2012) and similar approaches that apply the same principles to NGS data including Sequenza (Favero et al., 2015) and FACETS (Shen and Seshan, 2016). These methods all employ two stages: a segmentation step and a modelling step.

The segmentation step examines many datapoints of two measures across the genome: the first indicative of total copy number change in tumour sample relative to a normal sample known as the log-ratio or LogR and a second measure examining the proportions of heterozygous SNPs known as B-allele frequency of BAF. These data are used to determine where an SCNA begins and ends, or at least maps to, in the normal genome to produce the segments in which SCNAs are described in cancer genomics. These copy number calling algorithms take a number of approaches to this including: circular binary segmentation (CBS), Hidden Markov models (HMM), piecewise constant fitting algorithms (PCF) and allele-specific piecewise constant fitting (ASPCF), but all produce segments across the genome that reflect changes in BAF and LogR induced by SCNAs. The proportion of the genome that is affected by SCNAs varies greatly by cancer type and so, in some tumour samples, much of the genome will not differ from diploid normal sample. The schematic Figure 1-1 demonstrates the theoretical values of LogR and BAF of heterozygous SNP datapoints on a chromosome unaffected by SCNAs.

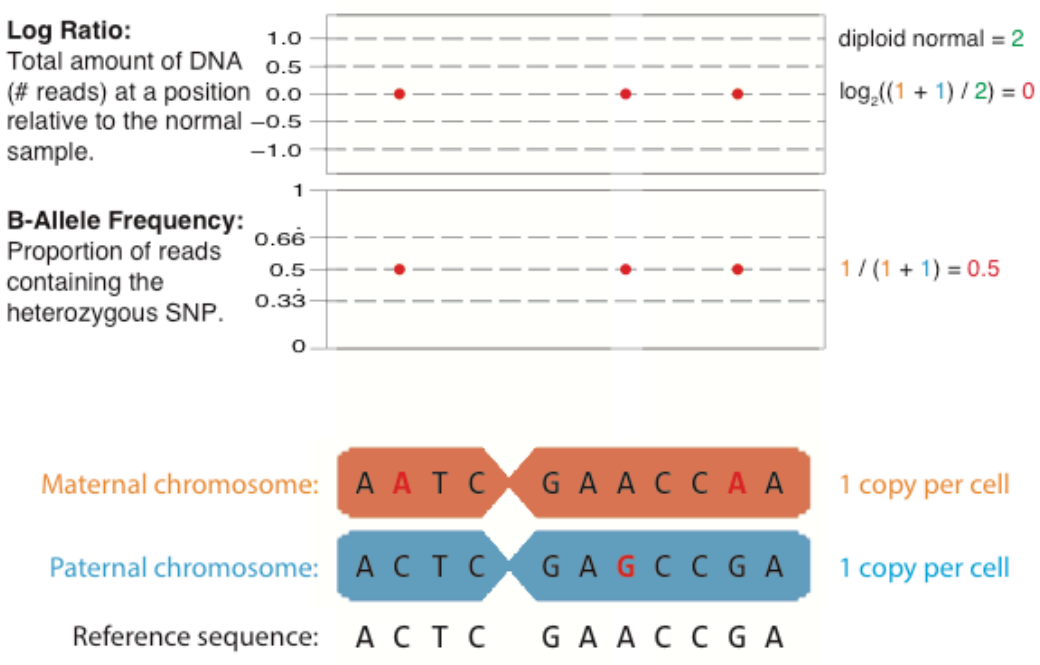

Figure 1-1: LogR and BAF in a normal sample
A normal diploid complement of chromosomes produces a LogR value of 0 and a BAF of 0.5

Any deviation in copy number in the tumour from the diploid value of 2 in the normal sample will result in a change in the value of in the LogR. In addition, any SCNA or combination of SCNAs that result in material from one homologous chromosome being present at greater copy number than the other will cause the BAF of the heterozygous SNPs present on the more common homologous chromosome (termed major allele or copy number) to increase, and the BAF of those on the less common chromosome (termed minor allele, or copy number) to decrease. An example of how a single copy gain would affect theoretical values of BAF and LogR is shown in Figure 2-1.

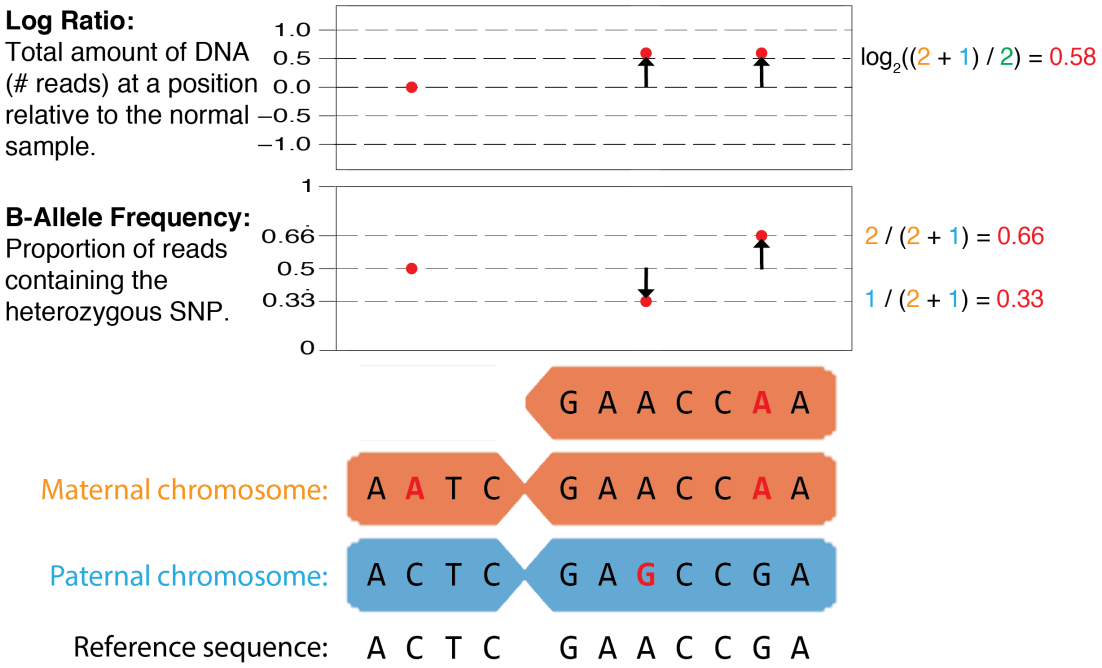

Figure 1-2: LogR and BAF in a tumour region
A gain in the q-arm of the chromosome causes an increase in LogR to 0.5 and deviation of BAF into two distributions centred around 0.66 and 0.33 depending upon which chromosome the heterozygous SNP' is present upon which is its phasing.

Once all segments are inferred across the genome, summarised values of the BAF and LogR data within their boundaries are generated and used as input for the modelling step that attempts to infer absolute integer copy number, ploidy and purity. This modelling step varies between copy number calling algorithms with some able to incorporate other sources of information such as SNV frequencies or known karyotypes from previous cytogenetic studies. However, all these algorithms attempt, by considering all combinations of purity and ploidy, to choose the optimal solution that estimates the allele specific copy number for each segment to be closest to non-negative absolute integer values. The modelling step will produce multiple possible solutions and most copy number calling algorithms require manual curation for error checking and selection of the most probable solution. Figure 1-3 shows a schematic of allele specific copy number segmentation output, termed the copy number profile (CNP), for a chromosome affected a by a single copy gain with the simulated tumour sample demonstrating 100% purity.

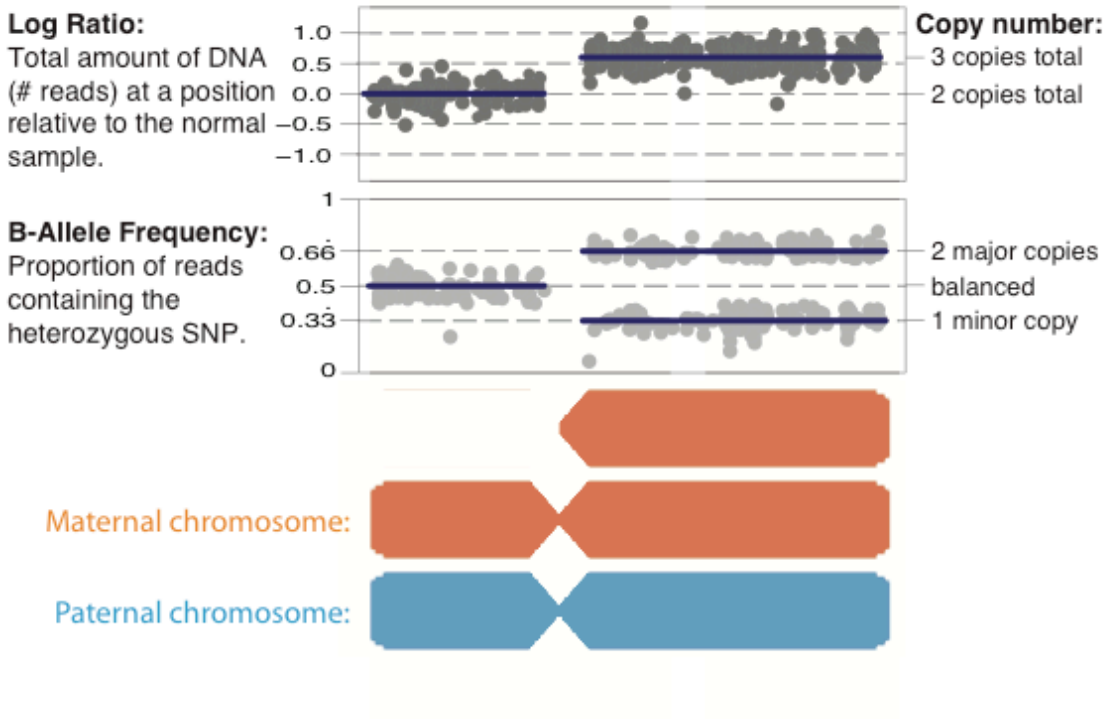

Figure 1-3: Copy number calling in a tumour region
Allele specific copy number calling tools use the combination of LogR and BAF to call total and allele specific copy number, an example of which matching the previous figure is shown here.

1.3.9 Non-integer SCNA estimates and subclonality in single samples

These SCNA calling algorithms, by optimising to the ploidy and purity solutions that produce CNPs with results closest to absolute integer copy numbers, assume there is a clonal, or at least dominant subclone, in which all cells share the same SCNAs. This assumption is necessary given the lack of resolution available due to homogenization of the DNA of the cells from a bulk sample in order to gain the advantages of identifying LOH as well as gains and losses through removing the confounding factor of purity. However, subclonal populations of cancer cells with differing copy number states are known to occur within a single bulk sample taken from a tumour. Subclones within the same tumour region with differing copy number states at the same genetic locus may lead to non-integer copy number states that appear even with the “optimal” purity and ploidy solutions that aim to minimize their occurrence across the genome.

In order to be detectable using bulk sequencing from single regions, subclonal populations with differing copy number states must be sufficiently prevalent and, in combination, produce a non-integer copy number estimate that appears to be

significantly different to integer copy number states with the chosen purity and ploidy solutions, even accounting for measurement noise. However, a subset of possible combinations of subclonal populations with different copy number states will produce data that is congruent with an integer copy number state. A simple example of this would be if there were two subclonal cancer cell populations each representing 50% of the entire population, one with a total copy number of 3 and the other with a total copy number of 5 at the same locus. These two subclonal populations would fit a solution consistent with single clonal population of cancer cells with a total copy number state of 4 at this locus.

Theoretically, these combinations of subclonal populations with differing copy number states could mask gains and losses in the initial modelling step of copy number calling algorithms. If gains and losses of the same loci were to occur within the same sample in separate subclonal populations there could appear to be no overall gain or loss in the inferred copy number states. In addition, two subclonal populations that both contained an allelically imbalanced gain at the same locus, but each from different homologous chromosomes, could appear to be a single clonal population demonstrating a balanced gain event (see Figure 1-4). Masking of copy number gains and losses through extensive subclonal heterogeneity is difficult to rule out from bulk analysis alone. However, prior to high-throughput technologies a huge number of cytogenetic studies with resolution at the single cell level were performed and identified recurrent gains and losses present in all cells of a patients' disease with little evidence recorded of co-occurring gains and losses of the same locus (Mitelman, 2015).

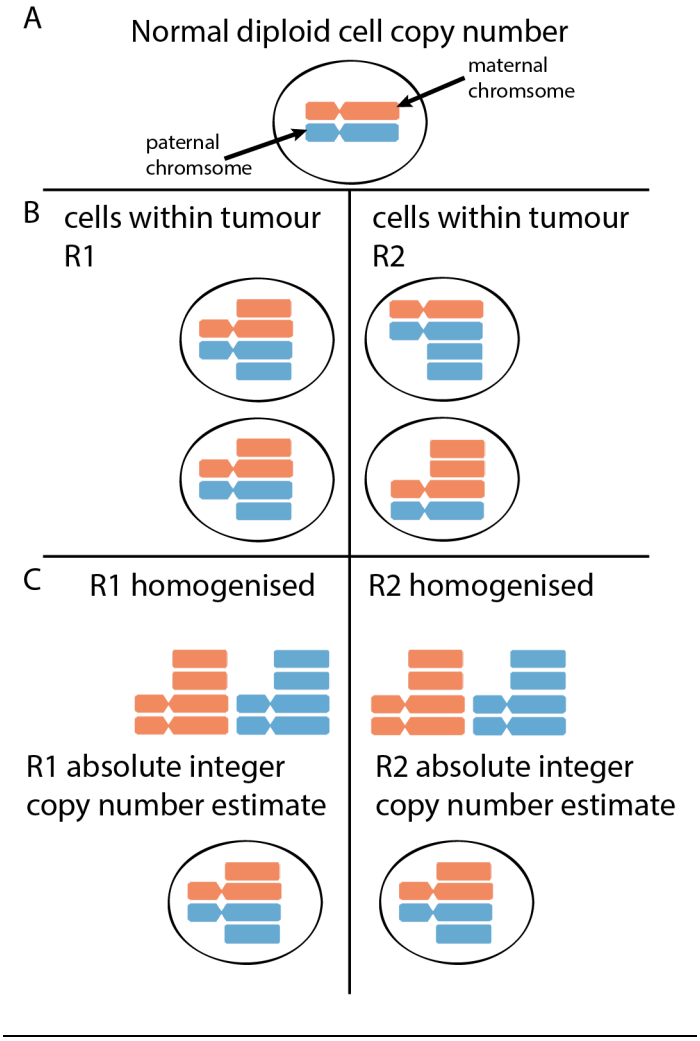

Figure 1-4: Bulk sequencing may mask ASCN heterogeneity
A) shows a schematic representation of a maternal (orange) and paternal (blue) chromosomes in a diploid cell. B) Shows cells from two theoretical regions of the same tumour. R1 contains only cells demonstrating a balanced gain of the q arm of the chromosome, one copy from each homologous chromosome while R2 contains two separate subclones each demonstrating allelically imbalanced gains of the q arm from different homologous chromosomes. C) shows the total DNA present when each region is homogenized. D) shows the likely solution that would be produced by a copy number calling algorithm given the input derived from the homogenised DNA from both regions.

However, these same cytogenetic studies cannot distinguish homologous chromosomes and thus cannot give estimates of allele-specific copy number at the single cell level. Indeed, there are currently no widely employed NGS or experimental methods able to describe allele specific copy number at the single cell level (see section 1.3.19) and therefore the degree of allele-specific copy number heterogeneity within single samples and whether any “balanced” SCNA gain and loss events described by copy number calling algorithms represent “averages” of heterogeneous events remains an unanswered question.

Despite these caveats many copy number calling algorithms attempt to identify and characterise subclonal copy number from single bulk samples. While some initial approaches to subclonal SCNA calling did not use BAF values (Oesper et al., 2013), many subclonal copy number calling tools utilize BAF estimates to detect slight deviations from integer copy number states. Subclonal copy number calling algorithms follow two general approaches using BAF values: attempting to inferring subclonal SCNAs in each copy number segment separately (Carter et al., 2012, Nik-Zainal et al., 2012) or attempting to cluster potentially subclonal segment values in attempt to pool information and infer a more accurate subclonal fraction common to multiple events (Fischer et al., 2014, Shen and Seshan, 2016).

1.3.10 Haplotype reconstruction and phasing

A BAF based approach to subclonal copy number calling has the advantage that BAF values are less likely to be affected by biases such as replication timing or GC content related wave biases that will affect read depth and LogR in NGS. As both the reference and variant alleles are affected by these biases to a near identical degree they will be factored out in the calculation of the BAF (Diskin et al., 2008, Koren et al., 2012, Kleinheinz et al., 2017). Further improvements in the accuracy of BAF values can be achieved through haplotype reconstruction and phasing of the heterozygous SNPs.

Phasing is defined as assigning heterozygous germline SNPs to one of the two homologous chromosomes present present at their locus in a diploid cell. In the context of subclonal copy number calling this is achieved by using a reference set of haplotypes that have been generated from a large population study and imputing the matching haplotypes. This imputation uses the homozygous and heterozygous SNPs present identified in a patients' NGS data to infer which two haplotypes are likely to be present in each so-called "haploblock" using a tool such as Impute2 (Howie et al., 2009) or SHAPEIT (Delaneau et al., 2013). Using the reference haplotypes derived from the 1000 genomes project leads to a haploblock size of approximately 300kb separated by recombination hotspots which, in areas of AI, cause the chequered pattern in BAF when plotted by genomic position along a chromosome that gives the Battenberg algorithm its name.

Subclonal copy calling tools designed for single regions that utilise this phasing approach include ABSOLUTE (Carter et al., 2012), Battenberg (Nik-Zainal et al., 2012) and ACEseq (Kleinheinz et al., 2017). Phasing improves copy number

estimation from BAFs in two ways: it allows identification of previously unobservable AI and improves the accuracy of the quantification of previously observed but subtle of AI. With current technology, haplotype reconstruction and phasing is limited to WGS and in some cases SNP arrays as these provide sufficient genotype data across the genome to allow imputation of haplotypes, while NGS methods such as WES that sample a lower proportion of the genome provide insufficient resolution of the genotype.

SCNAs often result in AI when genetic material from one homologous chromosome is present at a greater copy number (termed the major copy number or allele) and the other is present at a lower copy number (termed the minor copy number or allele). In areas of the genome which do not exhibit AI there is a unimodal distribution of BAFs that should be centred around 0.5. In areas of AI the BAFs from heterozygous SNPs present on the major allele and those on the minor allele form a bimodal distribution with the centre of the upper and lower peaks determined by the relative proportions of the two homologous chromosomes.

However, the difference between the upper and lower peaks of BAF in an area of AI is made more or less distinct by a number of factors: the read depth (in the case of NGS), the magnitude of difference in copy number between the major and minor, the proportion of cancer cell population carrying the SCNA(s) causing the AI and the proportion of normal balanced diploid cells in the sample assayed. When the difference between the two peaks of BAF is clear it is easy for SCNA calling algorithms to separate the two distributions as coming from different haplotypes and detect AI.

However, as the two peaks of BAFs in an area of AI approach 0.5, which may occur in cases where an SCNA is subclonal and present in a low proportion of the cancer cells of the tumour region sequenced, the two distributions increasingly overlap. Phasing enables these overlapping distributions to be easily distinguished as each BAF data point is assigned to one of two haplotypes. The BAFs assigned to these two haplotypes can then be tested to see if their distributions significantly differ revealing AI that would be missed using segmentation methods on unphased data.

The phasing also improves the quantification of AI that is detected in unphased data. This is because in areas of AI in which the BAF from heterozygous SNPs on different homologous chromosomes overlaps but is still detectable when unphased, some

values will be mis-assigned to the incorrect allele leading to inaccuracies that are removed when their assignation is determined by phasing.

1.3.11 Differing information content of SCNAs and SNVs from bulk samples

SNVs when described using NGS have certain, often obvious, properties that make their use in analyses of ITH and construction of tumour phylogenies from bulk samples more tractable than SCNAs. In addition, SNVs, even when described by different variations of NGS such as WGS, WES or targeted panel sequencing maintain the same properties whereas the sequencing modality greatly affects the detail with which SCNAs can be described. In the following section I will describe compare the inherent properties of SNVs and SCNAs when examined in single bulk samples.

1.3.12 Size and location

By definition, an SNV has a size of 1 base pair and a definitive position in the genome. SCNAs cannot be so easily described and vary in size from whole chromosomes and arm level events to smaller “focal” events that may even fall just outside the category of an INDEL (approx. 1-50bp).

It is only possible to resolve the exact positions of a subset of SCNAs through the use of WGS to identify their corresponding SV breakpoints. However, few algorithms that combine SV and SCNA estimation are present in the literature with only ACEseq (Kleinheinz et al., 2017), CONSERTING (Chen et al., 2015) and ReMiXT (McPherson et al., 2017) attempting to do so. Even with these computational approaches and WGS it will not be possible to resolve the start and end positions of SCNAs that begin or end in areas of the genome that are not amenable to NGS such as centromeric, telomeric or other highly repetitive regions.

Indeed, a central issue for most SCNA callers, even those designed for WGS, is segmentation: how to determine where an SCNA begins and ends in the genome. SCNA calling algorithms that rely on LogR and BAF measures (for SNP array, WES, targeted approaches) and have taken a number of approaches to this including circular binary segmentation (CBS), Hidden Markov models (HMM), piecewise constant fitting algorithms (PCF) and allele-specific piecewise constant fitting (ASPCF). At best, the output from these methods will offer narrow estimate of a potential “breakpoint” between copy number segments and not an exact position.

1.3.13 Identity, persistence and the infinite sites hypothesis

In cancer genomics SNVs are assumed to adhere to Kimura's infinite sites model (ISM) (Kimura, 1969) and occur only once in the life history of a tumour as the statistical likelihood of the same locus being mutated multiple times in the same location is thought to be vanishingly small. In accordance with the model SNVs are also thought not to back mutate to the reference base. These properties combined with an exact size and location give each SNV a unique identity in the life history of a tumour when assayed with any form of NGS and allow their use in phylogeny building.

SCNAs do not conform to the infinite sites model. Unlike SNVs, in most cases SCNAs cannot be said to demonstrate a comparable unique identity. Most SCNAs described in the cancer genomics literature do not have well defined location in terms of an accurate start and end position in the genome. Even SCNAs described with WGS will often lack accurate start and end positions if either is in the centromere, telomere or another section of the genome non-amenable to WGS. SCNAs analysed with any other high-throughput technique will nearly always lack accurate start and end positions and therefore lack a unique identity.

The varying sizes of SCNAs also contribute to another difference compared with SNVs: multiple SCNA events are known to overlap, compounding the issue of attempting to assign them a unique identity from start and end positions that would allow phylogeny building. In addition, overlapping of SCNAs may allow back mutation to occur. For example, a gain of single copy of a whole chromosome in an originally diploid cell, itself an event without uniquely identifying characteristics such as a detectable exact start or end position, could occur in a tumour cell and be followed by a subsequent loss of the same whole chromosome in descendant cell thus returning the chromosome to a normal copy number.

Indeed, given the possibility of larger and smaller overlapping SCNAs, many SCNAs are possibly the result of compounded separate genome altering events rather than single changes themselves. In this case, by allowing SCNA back-mutation at loci across the genome, only complete loss of genetic material from a cell could be considered to be a permanent alteration to the cancer cell's genome. For example, LOH refers to the complete loss of the genetic material from one of the two homologous chromosomes, therefore once a locus demonstrates LOH in a cancer cell all subsequent generations will also be affected. The same can be said to be true

for homozygous deletion though the area of the genome affected is highly likely to be an order of magnitude lower than that affected by LOH due to the lethal nature of entirely losing both homologous chromosomes material for large sections of the genome.

1.3.14 Measures of cellular prevalence within a sample

The small and defined size and location of SNVs allows them to be detected with comparative ease by NGS using relatively few sequencing reads, dependent on sequencing depth and modality. This is in contrast to the much larger quantity of sequencing reads required to reliably detect even small SCNAs. SNVs when detected by NGS are nearly always accompanied by a variant allele frequency or VAF. Given that relatively few NGS reads are required to detect the presence of an SNV, the proportion of reads that contain it compared to those that do not imparts useful information regarding the prevalence of the event in the sequencing performed. In contrast, instances of an SCNAs are not contained within a single read and require summary measures of a multitude of sequencing reads coupled with algorithmic quantification to observe. There is no easy counterpart for an SNV's VAF available from NGS for SCNAs and thus no simple proxy for cellular prevalence.

This makes the inference of subclonal SCNAs from single bulk samples more difficult than the inference of subclonal SNVs. Indeed, most copy number calling algorithms do not call subclonal SCNAs unless designed for WGS and these algorithms then often employ haplotype reconstruction and phasing to attempt to further resolve these events in bulk samples. Algorithms that do attempt to assess subclonal SCNAs in data types other than WGS will place restrictions on the proportion of the genome that can be found subject to subclonal SCNAs in order to keep the problem of purity and ploidy inference tractable. Furthermore, tools that attempt to assay subclonal SCNAs from bulk samples are limited in their sensitivity, only being able to reliably detect SCNAs (even in the case of large allelically imbalanced events with phasing) at much higher subclonal proportions than those at which it is possible to detect subclonal SNVs. This is even without addressing the issue that distinct subclonal populations with differing SCNAs may actively mask each other's events when the two populations are homogenized in the processing steps of NGS analyses.

1.3.15 SCNAs and their influence upon SNV ITH

When a sequence level mutation occurs, a single instance of it must initially be present on one of the chromosomal copies of the genome present in the cell. However, as cancers may gain or lose sections of chromosomes through chromosomal instability, the number of copies of a mutation present in the later generation of the same cell may change: either increasing, remaining at a single copy or being completely lost. To counteract this, one early approach even attempted to avoid the influence of SCNAs on SNVs by only assessing SNVs in areas of the genome thought not to demonstrate SCNAs (Miller et al., 2014).

When a mutation is detected using NGS it will appear in a certain a certain proportion of reads covering its the genomic position with the other reads detecting the wild type allele; this is quantified as its variant allele frequency (VAF). One of the major approaches to identifying heterogeneity within NGS from a bulk sample of a tumour region is to apply clustering approaches to the SNVs by the proportion of the cancer cells in which they are present, this is quantified as their cancer cell fraction or CCF and is calculated using the equation below:

$$VAF = \frac{\rho \times CCF}{CPNnorm \times (1 - \rho) + \rho \times CPNmut}$$

(Equations 1) Sequence-based mutational cancer cell calculation
Where VAF = variant allele frequency; ρ = tumour purity; CCF = cancer cell fraction; CPNnorm = local copy number in normal genome; CPNmut = local copy number in tumour genome.

Therefore, in a single tumour region assayed with NGS of a bulk sample the proportion of reads on which an SNV can be observed is dependent on three factors: the purity (or proportion of the bulk sample that consists of cancer cells), the total copy number of genome at that genomic position and finally the number of these copies that carry the SNV.

In addition, the purity estimate that is a term in this equation is also (most commonly) the result of the same somatic copy number calling algorithm that determines the local total copy number estimate. Therefore, much of the literature on ITH, subclonality and branched evolution is at least partially dependent on copy number estimates from bioinformatics tools that were originally designed for single region analysis using SNP arrays.

SCNAs may therefore affect both true SNV heterogeneity as well as the accuracy of its quantification. True heterogeneity would result from loss of chromosomal material in which an SNV was present in an ancestral clone in one tumour subclone but not another. The accuracy of SCNA calling and the accompanying purity estimates may also significantly affect the CCF calculations and thus resulting subclonal populations identified when the CCFs are clustered. It has also been argued that these potential losses of SNVs through SCNAs must be taken into account in phylogeny building approaches as, if they occur, they will violate aspects of the ISM that specify back mutation (McPherson et al., 2016).

1.3.16 SCNAs detection from bulk samples summary

The description of SCNAs from bulk samples provides less detail, dependent on the modality used to investigate and the SCNA calling algorithm used, and less likely to accurately reveal subclonal events than the analysis of SNVs from a bulk sample. It is likely for these reasons that the majority of recent work on establishing levels of ITH has focused on SNVs, especially given the large single region bulk datasets available. However, strikingly the inference of SCNAs and accurate purity and ploidy from single bulk tumour regions may have a significant impact upon CCF calculations and subsequent determination of the heterogeneity and subclonality of the SNVs present in a sample. Therefore, SCNAs have influence upon even studies in which SNVs, using their adherence to the ISM, have been used to generate tumour phylogenies. As SCNAs most often do not have unique identities, may overlap or back mutate and their prevalence within a single bulk tumour sample is difficult to estimate, relatively few efforts have been made to assess their heterogeneity or construct phylogenies from them. In fact, due to these difficulties, many efforts at quantification of their heterogeneity employ non-bulk NGS approaches. Those studies that do attempt to assess their heterogeneity using NGS or other bulk sampling methods must take multiple samples, as while single region samples assess overall SCNA burden within a single region they have, depending on the modality used, a limited ability to detect SCNA heterogeneity and subclonality.

1.3.17 Non-bulk sample approaches to SCNA detection

An ideal method to quantify SCNAs would be able to characterise the SCNAs and SVs of a single cell with allele-specific and breakpoint information. Unfortunately, no such assay exists. The limitations of bulk sequencing assays such as SNP arrays and various NGS modalities have been covered in the previous section, this section

describes the multitude of single cell sequencing and low-throughput non-NGS methods have been applied to detect SCNAs in cancer, each with its own particular advantages and disadvantages.

1.3.18 Levels of resolution of non-bulk sample SNP array or NGS methods of SCNA description

Fluorescence-activated cell sorting (FACS) ploidy estimates can be used to identify subpopulations of cells with differing ploidies within the same sample, to compare ploidies across patients, and to compare multiple samples from the same patient. The work of Gerlinger et al. utilised multi-region ploidy profiling using FACs in KIRC to show a diploid profile in most of the regions assayed but also revealed both diploid and tetraploid regions within the same patient (Gerlinger et al., 2012). While FACs ploidy may reveal a spread of ploidies within a single sample that SNP arrays or bulk NGS approaches might struggle to identify, it is limited to describing the overall ploidy of populations of individual cells and cannot describe any of the individual SCNAs present within them. In addition, diploid or close to diploid tumour cells will be indistinguishable from normal tissue diploid cells that will contaminate the sample. FACs ploidy estimates have been used as validation for copy number calling tools such as ASCAT (Van Loo et al., 2010).

FISH approaches which mark small areas of the genome can be used to assay the copy number of genomic loci within single cells. However, these approaches are often difficult and involve challenging laboratory work and image analysis that is required to produce estimates of the copy number changes. Most FISH studies can only examine loci on a few chromosomes in each cell. This results in the rate of aneuploidy all chromosomes being extrapolated from the few that are studied which may cause over or under estimation in the total rate of aneuploidy. Indeed it is recognized that there is an inherent error rate in FISH measurements (Bartlett et al., 2008), these could occur through failure of the probe to hybridise, or non-specific binding. This potential error rate has been suggested as one cause of a current controversy regarding whether neurons in healthy tissues and Alzheimer's disease may demonstrate chromosome copy number changes. Multiple groups have reported relatively high levels of abnormalities in brain tissue using FISH but recent investigations using single cell sequencing have failed to recapitulate the results, reviewed in (Ferronika et al., 2017).

In contrast to FISH, Spectral Karyotyping (SKY) and M-FISH can examine all chromosomes within a cell at once. However, this method requires cells in metaphase (unlike FISH which can be performed on cells in interphase) and thus access to cultures of dividing cells. SKY is also thought to be likely to over emphasise chromosome loss due to chromosomes being washed away from the slide onto which they are dropped which has been suggested as one potential reason for the high rates of hypoploid cells observed in some studies of neurons performed with this technique (Rehen et al., 2001, van den Bos et al., 2018). A limitation of FISH, SKY and M-FISH is that they are allele insensitive. There is no way to distinguish the two homologous chromosomes and as such while total copy number ITH might be observed with these techniques, LOH or heterogeneous AI between single cells would be entirely undetected.

Before the advent of SNP arrays and NGS, the only way to assess AI and potential LOH was through the use of microsatellite markers with PCR fragment size analysis (see section 1.3.2) and this is again limited to bulk samples that are homogenised in the analysis process. Indeed, at the time of writing, no technique has been applied to cancer samples that can reveal allele-specific copy number changes at the level of single cells. One technique that could be theoretically be used to identify such heterogeneity is specific to allele fluorescence in situ hybridisation (STARFISH) that was used to visually identify SNVs in the *PIK3CA* gene at the single cell level (Janiszewska et al., 2015). However, this technique has not been used to attempt to identify normal heterozygous SNPs in cancer cell context to investigate potential AI.

1.3.19 Single cell sequencing methods to detect SCNAs

Development of single cell sequencing (SCS) protocols has recently allowed the observation of all chromosomes at the level of non-dividing single cells. There are many SCS protocols currently available with which libraries are constructed from individual nuclei or cells and their exact details are beyond scope of this thesis (see (Gawad et al., 2016)). However, most SCS technologies depend on whole genome amplification (WGA) as part of their protocol (van den Bos et al., 2018). While WGA is required in order to create enough coverage to allow for SNV calling, it is problematic for SCNA calling as it introduces significant PCR amplification biases which can obscure SCNAs. To avoid this issue, some protocols have been created that do not use WGA, however this makes them unsuitable for SNV calling (van den Bos et al., 2016, Zahn et al., 2017) but the low genomic coverage this results in can be used to assess total copy number changes within single cells. Therefore, while the

SCS methods are likely an improvement on FISH and SKY as they allow all chromosomes of non-dividing single cells to assessed for total copy number changes, the results they produce are still limited by being allele insensitive as both FISH and SKY are. In addition, because of the low coverage these SCS SCNA techniques provide, they cannot provide resolution of structural variants due to a lack of coverage of breakpoints at base pair resolution.

SCS approaches have been applied to BRCA (Navin et al., 2011, Wang et al., 2014b, Baslan et al., 2015, Gao et al., 2016) and shown multiple subclones within single patients' tumours when cells are clustered by SCNA gains and losses. These findings have been argued to show that BRCA tumours experience a short burst of aneuploid rearrangements followed by stable clonal expansion, demonstrating short punctuated bursts in the early stages of tumour evolution rather than gradual accumulation of SCNA events (Gao et al., 2016). A separate group working on analysing SCNAs from single cells in FFPE BRCA ER+ samples revealed that ductal carcinomas in situ show intratumour genetic heterogeneity at diagnosis and that these lesions may progress to invasive breast cancer through a variety of evolutionary processes, showing either "parallel progression" of multiple SCNA subclones or "bottle-necking" progression of single SCNA subclone (Martelotto et al., 2017). These results show common total copy number heterogeneity in BRCA samples though the degree of ITH estimated differs between studies. SCS approaches have also demonstrated SCNA heterogeneity in a SCLC patient (Ferronika et al., 2017) with a high level of ITH in cells in the primary tumour, lymph node and adrenal metastases, but a much lower level of ITH with a distinct pattern of SCNAs in a liver metastasis. Finally a range in levels of SCNA heterogeneity was found both human B-cell ALL and murine models of the disease (Bakker et al., 2016).

A study of single cells, isolated from six non-metastatic BRCA patients, that were thought to be disseminated tumour cells traced their origins to either the dominant tumour clone, primary tumour subclones, or subclones in an axillary lymph node sample suggesting a heterogeneous population of tumour cells with differing SCNAs that separated from the primary tumour comparatively late in tumour evolution (Demeulemeester et al., 2016). Other cancer types in which, like Demeulemeester et al., circulating tumour cells (CTCs) have been analysed for SCNAs include PRAD, in which the analysis of a single patient's disease revealed two subclones with differing SCNAs (Dago et al., 2014), and SCLC in which chemotherapy sensitive and

refractory disease could be distinguished based on the SCNAs present within the CTCs (Carter et al., 2017).

It is important to note that SCS approaches, SCNA quantification and phylogeny building from these SCS SCNA analyses were performed in highly study specific manner. Few if any bioinformatics tools to analyse SCNAs in SCS are used outside the groups in which they are created. This contrasts to single cell SNV calling studies for which multiple phylogeny construction algorithms have been published including Bitphylogeny (Yuan et al., 2015), OncoNem (Ross and Markowetz, 2016) and SCITE (Jahn et al., 2016).

1.3.20 Non-bulk sample approaches summary

While approaches with greater resolution than bulk sample analyses exist for detecting SCNAs within tumours at the single cell level, they are limited in a number of ways. All single cell DNA sequencing, FISH, M-FISH and SKY approaches to date used for copy number analysis are not able to reveal AI or LOH and only describe total copy number changes. Nonetheless, SCS approaches have revealed widespread total copy number heterogeneity across a multitude of cancer types and have, on occasion, been interpreted to imply punctuated rather than gradual accumulation of SCNAs in tumour evolution. However, with the emergence of single cell RNAseq it is possible that AI, LOH and potentially allele-specific copy number may be estimated from single cells. This is because there will be many transcripts per cell that will carry heterozygous SNPs that provide allele-specific information rather than the few copies of DNA per cell and this provides greater coverage and reduces the risk of allelic drop out faced by SCS DNA sequencing approaches. Indeed, there is already one tool that purports to already be able to achieve LOH detection in single cells from single cell RNA sequencing data is HoneyBADGER (Fan et al., 2018).

1.3.21 Multi-region bulk sample copy number calling

In summary, aside from relatively unproven technologies that could be applied to produce allele-specific copy number estimation from single cells including STARFISH (Janiszewska et al., 2015) and HoneyBADGER (Fan et al., 2018), the only current technology that is able to detect allele-specific copy number heterogeneity within a tumour are multi-region bulk assays including SNP array and NGS approaches. Single-region bulk studies can only identify one homologous chromosome as being

the most prevalent at a genomic position within a region whereas multi-region bulk approaches could potentially identify SCNAs originating from different homologous chromosomes being the most prevalent at the same genomic position in different regions from the same patient's disease. It is using this property of multi-region bulk sequencing that I designed and implemented a bioinformatics tool to analyse multi-region bulk data for evidence of parallel evolution that will be shown in later chapters.

1.4 The origins of CIN and SCNAs

An enormous amount of effort and computational and experimental innovation has been expended in the invention of methods to detect SCNAs. This has been coupled with efforts to understand the mechanism underpinning the generation of these chromosomal alterations and though this thesis is primarily concerned with the timing of SCNA events, quantification of SCNA heterogeneity and CIN mediated parallel evolution, I will provide brief introduction to the processes underpinning CIN in this section.

1.4.1 Causes of CIN

Numerical CIN and aneuploidy originates from missegregation of whole chromosomes during anaphase. However, structural aneuploidy and CIN may also be caused by segregation errors such as chromatin bridges, ultrafine DNA bridges, and acentric fragments that fail to establish direct attachments to spindle microtubules. While both types of CIN originate from different pre-mitotic and mitotic defects they may occur together and are often interrelated (Burrell et al., 2013b).

The pre-mitotic causes CIN of include replication stress (Burrell et al., 2013a) while at least a partial explanation for the ubiquity of CIN across cancer types likely reflects the number of signaling pathways whose deregulation has been implicated in influencing chromosome segregation during anaphase (Thompson et al., 2010, Orr and Compton, 2013).

A detailed review of the mitotic causes of CIN is beyond the scope of thesis but these are known to include problems with sister chromatid cohesion (Solomon et al., 2011), centrosome replication (Ganem et al., 2009), spindle assembly checkpoint (SAC) defects (Giam and Rancati, 2015), or defects in microtubule attachments to chromosomes (Ertych et al., 2014, Godek et al., 2015). Indeed, the presence of an extra chromosomes themselves causing an aneuploid state has been found to lead to genomic instability (Passerini et al., 2016).

Other cell intrinsic sources of chromosome copy-number alterations in cancer include homologous recombination defects that are known to be associated with telomeric AI that forms so-called “genomic scars” (Birkbak et al., 2012) and telomere dysfunction that causes breakage-fusion-bridge cycles that may result in amplifications and dicentric chromosomes (Maciejowski et al., 2015). Finally, many cancer therapies used may disturb chromosome segregation during anaphase (Bakhoum et al., 2014, Lee et al., 2016).

Other consequences of chromosome missegregation, apart from aneuploidy and relatively large structural alterations, are focal and complex rearrangements called chromothripsis, the formation of extrachromosomal DNA and double-minute chromosomes that are often causes of high level copy-number amplifications (Stephens et al., 2011, Zhang et al., 2015, Ly et al., 2017, Turner et al., 2017)

Aneuploidy is known to often lead to the activation of p53, the protein product of the most commonly altered driver gene in cancer *TP53* (Bailey et al., 2018), and the proliferation of aneuploid cells is known often be limited by p53 dependent mechanism (Thompson and Compton, 2010) but how this activation occurs in response to aneuploidy remains somewhat controversial. It has variously been proposed to occur through the elevated levels of DNA damage caused by increased levels of reactive oxygen species (Li et al., 2010), the action of the p38 stress kinase (Thompson and Compton, 2010) or chromosome breakage that activate p53 through the action of *ATM* (Janssen et al., 2011). Interestingly, recent work has found that some whole-chromosome aneuploidies are propagated in p53 proficient cells, suggesting that aneuploidy does not always lead to activation of p53. However the same group found that survival and proliferation of cells demonstrating structural aneuploidies, involving loss or gain of parts of chromosomes, seemed limited to p53 deficient cells (Soto et al., 2017).

1.4.2 Whole genome doubling

Cells with multiple sets of the normal diploid complement of chromosomes are classed as polyploid. While whole organism polyploidy is unknown in current mammalian species (Svartman et al., 2005), it has been suggested as being likely to have occurred in early vertebrate evolution and is relatively common in other genera including various existing plant species (Otto and Whitton, 2000). However, in mammals there are examples of specialized cells that exhibit polploidy such as placental giant trophoblast cells (Lee et al., 2009), megakaryocytes (Ravid et al.,

2002) and hepatocytes (Toyoda et al., 2005). The form of polyploidy that these cell types most often show is tetraploidy in which they have twice the normal diploid chromosomal complement which occurs through whole genome doubling (WGD). WGD has also been observed to occur in many different tissue types in response to stress: telomere dysfunction and oxidative stress in hepatocytes (Gorla et al., 2001, Lazzerini Denchi et al., 2006), in response to inflammation in bronchial tissue (Lothschutz et al., 2002) and post-myocardial infarction in heart tissue (Meckert et al., 2005).

WGD can therefore be viewed in as both an evolutionary mechanism thought to allow new gene function by allowing repurposing of duplicated genes (Adams and Wendel, 2005) and as a potential stress response in normal tissue.

Following on from the observations of Winge in the 1930s that large numbers of chromosomes, often more than twice the normal diploid number, could be observed in mouse tumours induced by tar, other groups began to describe evidence for polyploidisation in cancer. WGD was repeatedly reported in cell culture of rodent cell lines (Barski and Cassingena, 1963, Jackson et al., 1970) and when transforming normal fibroblasts with the oncogenic DNA virus SV40 (Cooper and Black, 1963, Moorhead and Saksela, 1965).

These observations from cell culture also matched cytological observations in human cancers that demonstrated a bimodal distribution in chromosome number in cancer types where metaphase spreads had been observed. This bimodal distribution had a peak at a near diploid state as well as second peak with a triploid to tetraploid state (Storchova and Kuffer, 2008, Mitelman). It was proposed that this second, approximately triploid peak was the result of a single WGD event that created a tetraploid state rather multiple chromosomal missegregation errors resulting in increases in chromosome number (Storchova and Kuffer, Davoli and de Lange, 2011, Laughney et al., 2015).

In genomics and SNP array studies of human tumours WGD has been identified in single region pan-cancer studies occurring at different rates in distinct cancer types (Carter et al., 2012, Zack et al., 2013). In addition, WGD has been found to have a number of associations with genomic events including mutation of *TP53* and amplification *CCND1* and *CCNE1* amongst others (Zack et al., 2013, Bielski et al., 2018). However, the exact contribution of WGD to CIN and tumour evolution remains unclear. In 2014 Dewhurst and colleagues isolated WGD tetraploid clones from a

diploid culture of HCT116 colorectal cancer cell line. Long term cultures of isogenic WGD and non-WGD HCT116 clones revealed that cells which survived a WGD event demonstrated an increased tolerance for CIN. This increased tolerance allowed the WGD clones to acquire SCNAs that recapitulated chromosomal changes in genomically complex colorectal tumours, moving from a tetraploid to triploid state over time (Dewhurst et al., 2014).

Multi-region studies from our group in NSCLC (Jamal-Hanjani et al., 2017) and ESCA cohort (Murugaesu et al., 2015) identified WGD but also observed that WGD could occur subclonally, occurring only in a subset of regions assayed, a distinction that could not have been made using the single region approach.

1.5 Classification of SCNAs

There are many assays which may detect SCNAs with differing levels of resolution including cytogenetics, SCS, SNP arrays and NGS of bulk samples approaches to quantifying SCNAs. However, the by far largest and most detailed source of information available for interrogating the SCNA landscape of tumours are datasets of single-region bulk samples assayed with either SNP arrays or NGS.

An enormous amount of effort has been expended to create algorithmic approaches to detect LogR, BAF and SCNAs from these single-region bulk tumour samples with tens if not hundreds of algorithms published to analyse SNP arrays, WES, WGS and targeted panel sequencing. However, this plethora of methods and modalities has perhaps held back the development of standardized approaches to categorise, analyse and assess the importance of the processed SCNA “calls” resulting from them.

In this section I will review previously employed approaches and specific methods that have been developed to attempt to mine the processed SCNA calls from these large datasets. Specifically, I will review attempts to classify SCNAs by size, identify the SCNA equivalent of SNV “driver” events and efforts to find mutational signatures from SCNA data.

1.5.1 Arm and focal SCNAs

SCNAs in cancer are generally split into two categories: large-scale and focal events. Large-scale events include whole chromosome and so-called arm level copy number changes. Across cancer types these large-scale whole chromosome and arm level

affect more of the genome than any other somatic genetic alteration (Beroukhim et al., 2010, Zack et al., 2013, Mitelman, 2015) and arm level events were found to more common than whole chromosome SCNAs in 12 out of 16 tumour types examined (Beroukhim et al., 2010).

Arm level events have been defined as SCNAs affecting ≥ 98% (Beroukhim et al., 2010, Mermel et al., 2011, Zack et al., 2013) or ≥ 80% (Taylor et al., 2018) of a chromosome arm. Focal events have been variously defined in the literature as < 98% of a chromosome arm (Mermel et al., 2011, Zack et al., 2013), covering < 50% of a chromosome arm (Beroukhim et al., 2010) or covering < 25% of a chromosome arm (Koboldt et al., 2012). Focal events have also been defined using strict size cut-offs, for example of 1MB (Bignell et al., 2010) or 3 MB (Bierkens et al., 2013).

1.5.2 Identifying recurrent SCNA events and assessing their importance

Many algorithms have been created to attempt to identify separate driver sequence-level variants from “passenger” events and similar efforts have been undertaken to identify driver SCNA events in tumour evolution.

These efforts face two challenges. The first is that that somatic alterations are randomly acquired each cell division and only some of these will be driver events that promote the development of the cancer improving its fitness. Somatic alterations that have a neutral or even a slight deleterious effect on the tumour’s fitness may also become fixed in the tumour if they occur within a tumour subclone that also acquires beneficial driver alterations (McGranahan and Swanton, 2017). In this respect, the issue of identifying SCNA drivers is similar to that of identifying sequence-level alterations.

However, the second is an issue that is unique to SCNAs: that SCNAs may be small or large-scale events effecting from only a handful to thousands of genes in whole chromosome or chromosome arm level events. However, the improvement in cancer cell fitness that these events produce might be caused only by the alteration of a tiny subset of all the genes altered by an SCNA.

Therefore identifying the exact means by which an SCNA might be advantageous to fitness requires additional analyses compared to SNVs which, by only affecting a single gene, are easier to investigate. The first step in most efforts to identify SCNA drivers is to analyse large datasets of SCNAs from multiple patients’ disease, exploiting the idea that areas of the genome containing driver events will be altered

at higher frequencies than areas only containing passenger events. Algorithms attempting to achieve this include Significance Testing for Aberrant Copy number (STAC) (Diskin et al., 2006), Genomic Identification of Significant Targets in Cancer or GISTIC (Beroukhim et al., 2007). Subsequent efforts such as JISTIC (Sanchez-Garcia et al., 2010) and GISTIC2 (Mermel et al., 2011) aimed to address the issue of identifying smaller important areas of the genome from potential masking by large-scale events and to take into account that chromosome arm-level events had been found to be more common than larger or smaller SCNA events (Leach et al., 2004, Beroukhim et al., 2010, Zack et al., 2013). Still other approaches such as WIFA (Hur and Lee, 2011) and ADMIRE (van Dyk et al., 2013) attempted to at least partially address of effect of the size of SCNAs and potential masking of smaller SCNA events by, unlike the algorithms previously mentioned, avoiding the use of processed segmented results in their analyses and focusing instead on LogR datapoints as input.

However, the majority of these approaches predate the wide-spread use of using SCNA generating allele-specific copy number from SNP array data and the widespread use of NGS data to infer SCNA events. Indeed, since the publication of GISTIC2 relatively few additional algorithms for integrating SCNA data across multiple tumours have been published despite the obvious differences between NGS data covering large contiguous sections of the genome and SNP arrays examining only a small set of probes. A recent effort that has attempted to address differences has been WIFA-seq (Notta et al., 2016), an algorithm designed to be applied to single-region WGS data. Nonetheless GISTIC2, as the current gold-standard analysis tool has been applied to a pan-cancer cohort of single region datasets (Zack et al., 2013) and as periodically updated online databases of results (TCGA, 2017) which provides estimates of significantly recurrent arm-level SCNAs and separately focal SCNAs that are used throughout this thesis as reference dataset.

A relatively tractable subset of SCNAs are those that are both the smallest events (encompassing the smallest number of genes) and demonstrate the greatest magnitude of copy number change: focal high-level increases in copy number termed amplifications and homozygous deletions where zero copies of a small section the genome remain. In 2010, Santarius and colleagues described a framework for identifying driver genes by SCNA amplifications through the use of other means of validation including expression assays, experimental work and animal modelling and showed a large number of genes exert an oncogenic effect when overexpressed

(Santarius et al., 2010). More recently a census of homozygous deletions made possible by application novel statistical models for their detection from SNP arrays has been used to identify both established tumour suppressor genes and novel candidates (Cheng et al., 2017).

However, aneuploidy affects a greater proportion of the genome than other somatic alterations (Zack et al., 2013, Mitelman) and arm-level SCNA events occur more frequently than smaller focal SCNAs: The most recurrent focal SCNAs have been reported to occur at a frequency below 15% whilst the most frequent arm-level alterations have been found occur in 30% of tumours in single-region bulk sample pan-cancer study (Beroukhim et al., 2010). The fitness effect of a large-scale SCNA such as an arm-level event has often been thought to be due to the overexpression of an oncogene or loss of a tumour suppressor. However, due to the size of these events it has been difficult to determine single candidates that might explain the frequency of their recurrence (Liu et al., 2016). Nonetheless groups have taken computational approaches to attempt to identify the particular genes that might be targeted by these large-scale events to increase cancer cell fitness and perhaps negatively affect patient survival (Roy et al., 2016). However, a growing body of evidence suggests that effects of large-scale SCNAs may result from the perturbation of a combination of genes (Xue et al., 2012, Bonney et al., 2015), with experimental work now modelling large-scale chromosomal losses and examining their effects on cell fitness.

Indeed, early work using microsatellite markers to perform high-resolution allelotyping of chromosome 3p in NSCLC and preneoplastic bronchial epithelium revealed three regions of frequent breakpoints suggesting a combination of genes might responsible for the frequent appearance of this 3p LOH in NSCLC (Wistuba et al., 2000). More recent SCNA modelling has identified functional effects of whole arm deletions including that 8p LOH in BRCA cell lines result in changes in lipid metabolism that triggers invasiveness and confers tumour growth under stress conditions due to increased autophagy (Cai et al., 2016) and that loss of syntenic areas of human 17p in mouse models of leukaemia that do not affect the *TP53* gene that is normally assumed to be the target of 17p losses produce more aggressive disease (Liu et al., 2016). Other recent work has shown that long term passage of NSCLC cell lines engineered to have loss of chromosome 3p results in selection of advantageous alterations (including gaining an additional copy of chromosome 3p)

or expression changes that allow cells with a deletion of 3p to overcome the initial negative growth effects of aneuploidy a model system (Taylor et al., 2018).

However, in spite of these new experimental approaches, the exact manner in which most arm-level events may contribute to increases in cancer cell fitness remains unknown. For this reason, additional analysis and experimentation is required to distinguish whether particular “driver” genes or combinations of expression changes in genes present on these arms have advantageous functional effects. Indeed, while recurrence is assumed to indicate the likelihood of SCNAs influencing cell fitness, recent results suggesting that different chromosomes may be affected by missegregation at differing rates (Worrall et al., 2018) suggest further work may be needed to establish a null distribution to assess the relative recurrence of SCNAs against. Furthermore, while many single-region studies have assessed how commonly focal and arm level gains and losses occur (Beroukhim et al., 2010, Zack et al., 2013, Dentro et al., 2018, Taylor et al.), little is known about the clonality of these changes within tumours. Due to the limitations of bulk single-region assays, the only means to assess the clonality of these SCNA events is through multi-region bulk approaches or SCS.

1.5.3 SCNA mutational signatures

Mutational signatures have been used to dissect the different processes causing somatic alterations at the sequence level. These mutational signatures were defined by taking into account the type of substitution occurring in the mutated base as well as identity of the nucleotide preceding and following it. In total this created six classes of base substitution which, when combined with the identities of the two bases either side genrate 96 possible mutational contexts or signatures. The first application of this technique to a large sequencing cohort containing thousands of samples was performed by Alexandrov and Nik-Zainal et al. who used this framework coupled with non-negative matrix factorization (NMF) to extract a set of signatures that give an indication of which processes are active in shaping the cancer genome (Nik-Zainal et al., 2012, Alexandrov et al., 2013). Studies following this approach performed on single region bulk samples have shown a range of signatures whose distribution differs between cancer types. Signatures that have been identified to date have been attributed to UV light damage, smoking, the action of the APOBEC family of cytidine deaminases and defects in DNA mismatch repair amongst other mechanisms (Alexandrov et al., 2013).

Attempts to quantify and investigate of SCNAs for patterns which might be indicative of mutational processes began in earnest with the availability of large cohorts of SNP array data from single-region bulk samples. It was found that these data could be used to identify SCNAs that were associated with homologous recombination deficiency (HRD) which is often caused by loss of *BRCA1* or *BRCA2* function. Three measures or signatures of this phenomenon were created: the large-scale state transitions score (Popova et al., 2012), the telomeric imbalance score (HRD-TAI) (Birkbak et al., 2012) and loss of heterozygosity profiles (HRD-LOH) (Abkevich et al., 2012). These three measures were later combined into a single score (Telli et al., 2016) and more recently the HRD-LOH score has become part of a combined SCNA, SNV and INDEL machine learning based predictor of HRD designed for WGS called HRDetect (Davies et al., 2017).

A more general approach using NMF aimed to identify SCNA signatures active in high-grade serous ovarian cancer tumour evolution rather than find SCNA events indicative of single phenotype such as HRD. Using low-pass allele-insensitive WGS and NMF, 6 signatures were identified and each subsequently found to associate with combinations of sequence level alterations, sequence level mutational signatures, structural variants and even survival (Dentro et al., 2018). This approach, by linking patterns of SCNAs to previously unconnected aspects of the cancer genome landscape, may reveal novel mechanisms of generation and effects of SCNAs. In addition, as the low-pass WGS NGS modality used in the study is allele insensitive, additional signatures and signals of biological processes might be discovered by extending the approach to allele-specific copy number data. Finally, this approach could be extended to multi-region data which has previously revealed that different sequence-level mutational signatures are active at different points in tumour evolution with APOBEC associated mutations appearing to occur late in NSCLC evolution (de Bruin et al., 2014).

1.6 The role of CIN and SCNAs in tumour evolution

Efforts to identify which SCNAs, at both focal and arm level, might be advantageous to cancer cells have focused on identifying frequently recurrent events from large single-region cohort studies and then attempting to validate these through experimental work. However, other computational and experimental approaches have attempted to assess the role of the level of CIN or identify evidence of selection in patterns of SCNAs found within and across cancer types during tumour evolution.

1.6.1 The relationship between CIN and tumorigenicity

CIN and aneuploidy are prevalent across different cancer types and their levels when elevated are often associated with poor prognosis. However, there are many conflicting results as to defining their role in tumour evolution.

Models of aneuploidy that attempt to address aneuploidy's role in tumour development have included single whole chromosome trisomies and monosomies in yeast where they have been found to induce proteasomal stress and cause a lower rate of proliferation (Torres et al., Sheltzer et al.). Robertsonian translocations in mouse cells have found that aneuploidy decreases cellular fitness and elicits traits that are shared between different aneuploid cells (Williams et al.). In addition, cells that had chromosomes added by through microcell cellular transfer were found to show that whole chromosome aneuploidy leads to cellular senescence and decreased proliferative and transformative abilities (Stingele et al., 2012, Sheltzer et al., 2017). However, increased chromosome missegregation in primary tumours has been found to be associated with increased risk of distant metastasis and relapse BRCA and large B cell lymphoma. Furthermore, mouse models of lung cancer that were manipulated to produce increase rates of CIN resulted in increased tumour recurrence when the effect of a driver gene was removed (Sotillo et al., 2010).

1.6.2 Evidence of selection in tumours

While experimental assessment of the role of CIN and aneuploidy in tumorigenicity has often produced mixed results, bioinformatics based investigations of large datasets have often found evidence of selection for increased levels of CIN and aneuploidy tumours. Increased levels of SCNAs have been found to correlate with proliferation pathways (Carter et al., 2006, Davoli et al., 2017, Buccitelli et al., 2017) as well as reflecting metabolic selection pressures (Graham et al., 2017).

Pan-cancer assessment of aneuploidy patterns in terms of the distribution and inferred effect of oncogenes and tumour suppressors defined from sequence-level variants have suggested that cumulative haploinsufficiencies (losses to single copies) and triplosensitivities drive SCNAs that sculpt the genome to maximise proliferative potential (Davoli et al., 2013). The previous year a separate paper also suggested that recurrent patterns "hemizygous" (losses to single copies) might optimize proliferative potential (Solimini et al., 2012). These results fit with the observation that karyotype evolution in trisomic cell lines was able to mitigate the initially negative

consequences of an aneuploidy caused by the microcell cellular transfer addition of a chromosome (Sheltzer et al., 2017). Overexpression of different oncogenes in the same cell line have also been found to influence the landscape of breakpoints and structural variants present (Miron et al., 2015) perhaps suggesting different SCNAs being selected for in an oncogene specific manner.

Another line of evidence springs from the observation that cancer types are very often characterised by specific patterns of chromosomal arm-level alterations (Ried et al., 2012), including LUSC (Sundaresan et al., 1992, Hung, 1995), ESCA, and BLCA (Hoadley et al., 2014) as well as in studies on a pan-cancer level (Taylor et al., 2018). If these chromosomal events have no fitness effect it seems unlikely that they would display such distinct patterns apparently specific to their tissue of origin. Indeed, even in the pre-malignant state, SCNAs have been shown to contribute to driving early clonal expansions in the peripheral blood (Loh et al., 2018).

1.6.3 SCNA convergent and parallel evolution

Parallel evolution of SNVs affecting the same gene within a tumour but in different subclones has been posited as evidence of selection (see section 1.2.6). However, the first described case of parallel evolution in cancer was the result of SCNAs. Observations of parallel deletions of *ETV6*, *CDKN2A* and *PAX5* were made using the single cell assaying technique of fluorescence in situ hybridization (FISH) in childhood acute lymphoblastic leukaemia (Anderson et al., 2011).

The theory of clonal succession advanced by Nowell in 1976 was proposed without knowledge of sequence level variants and was based upon observations of karyotypic changes and mathematical modelling. Indeed, observations of karyotypic ITH of have been made numerous times before the completion of the human genome project. For instance, heterogeneity in amplifications of oncogenes and deletions of tumour suppressors within the same tumour had previously been observed (Szollosi et al., 1995, Di Vinci et al., 1999, Farabegoli et al., 2001). Indeed, the one of the major advances on theory clonal succession, that branched diverging evolution of subclones that persist and develop within a tumour was demonstrated to occur via copy number losses. Anderson and colleagues used fluorescence in situ hybridisation (FISH) to describe persistent heterogeneity of SCNAs in single cells from 30 cases of *ETV6-RUNX1* positive acute lymphoblastic leukaemias, 80% of which showed evidence of branched evolution with the other 20% being compatible with linear model (Anderson et al., 2011). However, it should be noted that due to the

limitations of the technique used to assay, only a few driver genes could be assayed and, if more up to date methods were used, these other cases might also have exhibited branched evolution.

A subtype of branched evolution that has been argued to demonstrate selection is parallel evolution, where the same gene is affected by distinct SNVs in different subclones. As was discussed in the previous section these parallel SNV events have been identified in a multitude of cancer types. Anderson and colleagues as well as demonstrating branched evolution also showed parallel losses of *ETV6* and *PAX5* in distinct subclones they identified using FISH. However, because of the inherent difficulties in characterising SCNAs and the lack of resolution of these events provided by WES and SNP array it is only recently that examples of SCNA mediated parallel evolution have been revealed using multi-region WGS. Parallel evolution of focal amplifications was identified centred upon *FGFR2* in a BRCA case (Yates et al., 2015) and a PRAD case (Gundem et al., 2015). The presence of these independent SCNAs in the small subset of studies powered to detect them suggested a potentially larger role for SCNAs in later tumour evolution than previously appreciated.

One of the first multi-region studies of SCNA evolution analysed 48 regions from eight KIRC cases and compared SCNAs in regions from the same tumour with single-region tumour biopsies from a separate TCGA cohort. Unsupervised hierarchical clustering of TCGA and multi-region KIRC cases showed segregation of samples from the same tumour into unrelated clusters. This resulted 25% of multi-region samples appearing more similar to unrelated cases than to any other region originating from the same case (Martinez et al., 2013). The authors suggest that this indicates convergent evolution with distinct subclones within a single case closely mirroring completely different cases in a separate cohort suggesting a finite number of viable evolutionary paths. In this case the pattern of evolution is classified as convergent rather than parallel as the populations of cancer cells independently exhibiting the same SCNA events are from distinct rather than the same ancestor.

1.6.4 SCNA heterogeneity in cancer

A large number of groups have identified the presence SCNA heterogeneity from multi-region studies in different cancer types. Some examples include multi-region FACS ploidy profiling and AI characterisation from SNP arrays that revealed ITH of both ploidy and SCNAs between tumour regions of single cases in addition to the better-known finding of SNV heterogeneity and branched evolution in clear cell renal

cell carcinoma (KIRC) (Gerlinger et al., 2012). The following year a multi-region SNP array study in GBM examined samples from 12 cases and demonstrated ITH as well as identifying *EGFR* and *CDKN2A* as early SCNAs while *PDGFRA* and *PTEN* were identified as later SCNA events (Sottoriva et al., 2013). Subsequent multi-region studies using SNP arrays to examine high-grade serous ovarian cancer (HGSOC) (Schwarz et al., 2015), multi-region SNP arrays in COAD (Sottoriva et al., 2015), multi-region WES in ESCA (Murugaesu et al., 2015), multi-region WGS in PRAD (Yates et al., 2017), multi-region WGS in BRCA (Yates et al., 2015), multi-region WGS in HGSOC (McPherson et al., 2016), multi-region WES NSCLC (Jamal-Hanjani et al., 2017) and targeted panel NGS analyses of KIRC (Turajlic et al., 2018b) have all identified the presence of SCNA heterogeneity. However, across all these studies a range of SCNA detection algorithms, thresholds and definitions have been applied resulting in results which are difficult to compare and do not provide pan-cancer multi-region comparison of different cancer types' levels of SCNA heterogeneity or whether any SCNA events are recurrently clonal or subclonal within or across tumour types.

While pan-cancer studies of SNV ITH and total SCNA burden have been performed, only single pan-cancer attempt at quantifying SCNA ITH and subclonality has been made using single-region bulk datasets (Dentro et al., 2018) and due to its single-region nature can only capture a subset of potentially heterogeneous events (see section 1.3.21)

1.6.5 SCNA branched evolution and phylogeny building

Relatively few of the multi-region studies listed in the previous section attempted to generate phylogenies to explore the branched evolution of SCNAs that was revealed by the presence of SCNA ITH across tumour regions from the same patient's disease. One of the reasons for this is likely the differing inherent properties of SNVs and SCNAs outlined in sections 1.3.11 through 1.3.14 and SCNAs' violation of the ISM (Kimura, 1969) which make SNVs a more tractable unit with which to infer an evolutionary history. Two tools that were applied to multi-region SCNA estimates to generate phylogenies include TuMult (Letouze et al., 2010) which has been applied to a study examining COAD (Sottoriva et al., 2013) and MEDICC (Beerenwinkel et al., 2014) which, amongst other cancer types and studies, has been applied to high-grade serous ovarian cancer (Schwarz et al., 2015), COAD (Sottoriva et al., 2015) and BRCA (Brown et al., 2017).

TuMult makes no attempt to address the non-adherence of SCNAs to the ISM (Kimura, 1969) whereas MEDICC (Minimum Event Distance for Intra-tumour Copy-number Comparisons) attempts to account for the unique properties of SCNAs derived from bulk samples. To do this MEDICC models a phasing of SCNA events and maintains two distinct haplotypes for each chromosome, allows SCNA events to overlap, and implements a stepwise model of transitions between valid SCNA states on each chromosome independently (Beerenwinkel et al., 2014). The measure "minimum-event distance" between copy number profiles generated through its stepwise transitions between valid SCNA states allows both the generation of SCNA based phylogenies and the statistical assessment of them.

Although MEDICC takes into account many of the issues inherent in SCNAs inferred from bulk data, its current implementation does leave others unaddressed including how to incorporate WGD into an SCNA phylogeny and how to identify and deal with genomic regions affected by extrachromosomal genetic material that may not behave in the manner as larger SCNA events resulting from whole-chromosome or arm-level SCNA events.

1.7 Clinical importance of CIN and SCNAs in cancer

1.7.1 SCNA burden and heterogeneity are linked to progression and survival

Measures of aneuploidy used as a surrogate for CIN are known to correlate with poor prognosis across cancer types (McGranahan et al., 2012). Pre-NGS studies in NSCLC that demonstrated such an association included assessment of CIN using FISH for c-myc, chromosome 6, *EGFR*, and chromosome 5 (Choi et al., 2009), a 12-gene expression signature (Mettu et al., 2010) and the CIN70 gene signature (Carter et al., 2006) all demonstrating an association between poor OS and inferred CIN status independent of tumour stage.

More recently multiple pan-cancer single region studies of SCNA burden across multiple cohorts have been found to be significantly associated with survival. An analysis of TCGA data found SCNA burden to be associated with DFS and OS in BRCA, UCEC, KIRC and COAD as validated these findings in an independent pan-cancer cohort analysed with a target NGS panel (Hieronymus et al., 2018). A separate study, that did use some of the same data, analysed 17,879 single-region tumour samples and found that tumour SCNA burden was prognostic for overall

survival in multiple cancer types and that focal SCNAs tended to confer a worse prognosis than “broad” SCNAs (Smith and Sheltzer, 2018).

In addition, tumour types that demonstrate high rates of WGD and aneuploidy are often those that have higher rates of treatment resistance, metastasis, and poor OS, and these include BRCA TN, PAAD, LUAD, LUSC, THCA, PRAD, SARC, GBM, ESCA, and microsatellite stable COAD (Carter et al., 2012, Bielski et al., 2018, Taylor et al., 2018). WGD events have also been found to be independently predictive of poor relapse-free survival in early-stage COAD disease in two independent cohorts in multivariate analyses (Dewhurst et al., 2014).

Multi-region studies that assessed measures capturing SCNA ITH found these to be significantly associated with worse PFS and OS in HGSOC (Schwarz et al., 2015) and that elevated SCNA heterogeneity was significantly associated with an increased risk of recurrence or death in NSCLC but that overall SCNA burden was not (Jamal-Hanjani et al., 2017). In HGSOC a subset of copy-number signatures potentially linked to distinct biological processes were found to be associated with survival (Macintyre et al., 2018) while evolutionary subtypes partially defined by SCNA burden and heterogeneity were found to influence progression in KIRC (Turajlic et al., 2018b, Turajlic et al., 2018a). CIN was also found to drive metastasis through a cytosolic DNA response with a significantly higher SCNA burden found in metastatic samples and suppression of CIN resulting in delayed metastasis in highly aneuploid tumour models (Bakhoum et al.).

1.7.2 CIN, therapy and resistance

CIN generates aneuploidy and SCNA heterogeneity that serves as a fuel for natural selection enhancing tumour fitness and providing a means for immune evasion, metastasis and drug resistance (Pavelka et al., 2010, Chen et al., 2012, Potapova et al., 2013, Laughney et al., 2015, Davoli et al., 2017). This ITH is known to frustrate targeted therapies against amplification of oncogenes such as in treatment-resistant BRCA HER2+ (Rye et al., 2018). However, aneuploidy also defines groups of tumours that are responsive to particular therapies, such as low-grade gliomas with 1p/19q co-deletions that have been shown to be responsive to specific chemoradiotherapy regimens (Cairncross et al., 2013) or provides trackable biomarkers and actionable targets such as chromosome 1q21.3 amplification for breast cancer recurrence (Goh et al., 2017).

Much like the experimental models assessing the impact of CIN on tumorigenicity, the influence of CIN on therapy is a mixed picture with CIN and complex aneuploidies correlating with resistance to antineoplastic agents, in tumour cell lines and the clinic (Carter et al., 2006, Swanton et al., 2009, Bakhoum et al., 2011). In contrast, excessive levels of CIN have also been reported to indicate enhanced sensitivity to cytotoxic therapies such as cisplatin in OVCA, COAD and BRCA (Swanton et al., 2009, Roylance et al., 2011, Zaki et al., 2014, Jamal-Hanjani et al., 2015).

More recently CIN has been shown to be responsible for a mechanism of immune escape where it facilitates LOH of the region of the genome containing the human leukocyte antigen (HLA) genes in 40% of NSCLCs. This HLA LOH is associated with a high subclonal neoantigen burden (McGranahan et al., 2017) and the burden of clonal neoantigens has previously been shown to be significantly associated with OS in NSCLC (McGranahan et al., 2016).

Chapter 2 Data and Methods

2.1 Data

2.1.1 TRACERx NSCLC multi-region sequencing data

The TRACERx 100 cohort comprises of the first 100 patients prospectively analyzed by the lung TRACERx main study (https://clinicaltrials.gov/ct2/show/NCT01888601, approved by an independent Research Ethics Committee, 13/LO/1546) and mirrors the prospective 100 patient cohort described in (Jamal-Hanjani et al., 2017).

All surgically resected tumour samples were macroscopically reviewed by a pathologist. Spatially separated tumour regions, documented by photography, were collected and snap frozen in liquid nitrogen for subsequent DNA extraction. At least two regions from each tumour, separated by at least 3mm, were collected. The samples were taken so as to maximize tumour purity (areas that were obviously necrotic, fibrotic, or hemorrhagic were avoided) and to reflect the observed macroscopic morphological heterogeneity of the tumour. Peripheral blood was also obtained at time of surgery.

2.1.2 Other exome sequencing data

All exome data was requested and downloaded as FASTQ files from either the EGA, dbGaP or transferred directly from the original group who produced it.

2.1.1 SNP array data

All SNP array data was downloaded from GEO as processed BAF and LogR values or transferred directly from the group who produced it as processed BAF or LogR calls and analysed using with ASCAT (v2.3) (Van Loo et al., 2010).

2.1.2 WGS data

All WGS data was transferred directly from the group who produced it as processed 1000 Genome Projects positions calls as described in (Nik-Zainal et al., 2012).

2.2 Methods

2.2.1 Whole genome sequencing

Somatic copy number alteration data was derived from WGS using ASCAT (Van Loo et al., 2010). Within-sample clonal and subclonal copy number changes in WGS data using the Battenberg algorithm as previously described (Nik-Zainal et al., 2012) and was also used to challenge and confirm copy number and ploidy estimates derived from ASCAT. Battenberg phases germline SNPs within NGS data through the use of Impute2 (Howie et al., 2009) that utilises a panel of polymorphic SNPs.

2.2.2 Whole exome sequencing

All exome data was analysed using the main study version of the TRACERx pipeline designed by Gareth Wilson and Richard Mitter. Full details of the pipelines can be found in (de Bruin et al., 2014) and (Jamal-Hanjani et al., 2017).

2.2.3 Multi-region somatic alteration calling

All somatic alteration calling was performed using the main study version of the TRACERx pipeline designed by Gareth Wilson and Richard Mitter. Full details of the pipelines can be found in (de Bruin et al., 2014) and (Jamal-Hanjani et al., 2017)

As described in greater detail in (Jamal-Hanjani et al., 2017) raw paired-end sequencing reads were aligned to hg19 and somatic variants were identified using VarScan2 somatic (v2.3.6) and extracted using VarScan2 process Somatic (Koboldt et al., 2012). Additionally, MuTect (v1.1.4) was used to identify SNVs (Cibulskis et al., 2013). All variants were annotated using COSMIC (v75) and Annovar (Wang et al., 2010).

2.2.4 Copy number analysis

Varscan2 copynumber was run to generate copy number data from paired tumor-normal samples, which produced per-region logR values, that were subsequently GC corrected (Koboldt et al., 2012). Homozygous and heterozygous single nucleotide polymorphisms (SNPs) were identified from the germline sample using Platypus (v0.8.1) (Rimmer et al., 2014). The B-allele frequency (BAF) of each SNP was calculated as the proportion of the reads at that position that contained the variant base.

The logR and BAF values were used with ASCAT (v2.3) (Van Loo et al., 2010) and Sequenza (Favero et al., 2015) in order to generate segmented allele-specific copy number data, purity, and ploidy estimates.

2.2.5 mPHASE multi-region phasing and mirrored subclonal allelic imblance detection

Performed as described in Chapter 3.

2.2.6 TRACERx SNV timing and phylogenetic tree construction

As described in greater detail in (Jamal-Hanjani et al., 2017), to estimate whether mutations were clonal or subclonal, a modified version of PyClone was used (Roth et al., 2014). For each mutation, an observed CCF (obsCCF) and a phylogenetic CCF (phyloCCF), which took into consideration any subclonal copy number events potentially altering the CCF, was calculated. Mutations were clustered using PyClone Dirichlet process clustering.

Phylogenetic trees for the TRACERx study were constructed using CITUP (v0.1.0) (Malikic et al., 2015), which takes as input mutation clusters and their mean cancer cell fraction with additional manual input for six tumours.

2.2.7 The isolation and passaging of diploid and tetraploid HCT116 clones

As described in further detail in Dewhurst et al. (Dewhurst et al., 2014) the ploidy of all clones was checked using flow cytometry for DNA content. Andrew Rowan carried out the isolation and passaging of both diploid clones from the 2N peak, and tetraploid clones selected from the >4N peak. Tetraploid and diploid clones were then passaged weekly through the seeding of 150,000 cells into fresh media in a 10cm dish.

2.2.8 HCT116 SNP array and analysis

HCT116 analysed with SNP arrays from described in this thesis were subject to Affymetrix SNP6.0 arrays by Aros(Applied Biotechnology). Normalization and integer copy number estimation were carried out using PICNIC (Greenman et al., 2010).

2.2.9 Inference of WGD from copy number profiles

WGD estimation was performed as previously described(Jamal-Hanjani et al., 2017). Briefly, each sample was represented as a copy number profile of major and minor allele copy numbers at the level of each chromosome arm. Using these arm level summaries, the total alterations (relative to diploid) and the probabilities of loss/gain for each allele at each chromosome arm were calculated. 10,000 simulations were then run for each sample. In each simulation, a number of sequential aberrations, based on the probabilities of loss/gain for each allele already calculated, were applied to a diploid profile. A p-value for whole genome doubling was obtained by counting the percentage of simulations where the proportion of chromosome arms with a major allele copy number ≥2 was higher than that observed in the sample.

2.2.10 Hierarchical clustering of clonal and subclonal SCNA events

Two patient-by-cytoband matrices were created, one consisting only of clonal events, and the other only of subclonal events. Gain events were indicated with “1”, while loss/LOH events were indicated with “-1”. Hierarchical clustering was performed based on Euclidean distances and using Complete hclust clustering.

2.2.11 Allele specific expression

To obtain allele-specific expression (ASE) profiles we ran phASER (Castel et al., 2016) on the TRACERx LUAD and LUSC samples for which we had RNA-seq data for (Rosenthal, under review). For each sample, ASE was determined for each heterozygous SNP that was reported from Platypus analysis of the corresponding exome data. SNPs that had at least 8 reads that were mapped to it were considered for analysis, and ASE was reported as (number of reads representing alternate allele/total number of reads mapping to SNP) For each gene, ASE was determined by taking the average ASE of all corresponding SNPs. Phase groups for SNPs were determined based on exome analysis.

Allelic imbalance in expression (AIE) of each SNP was determined by a binomial test of ASE value; with a significance threshold of $P<0.05$. Allelic imbalance in expression intra-tumour heterogeneity (AIE ITH) is determined for each gene, where AIE ITH is declared when some but not all regions of a tumour have AIE.

2.2.12 Assessing the significance of clonal and subclonal SCNA classed as gains and losses

To assess the significance of SCNAs, for each tumor type, the proportion of the genome subject to SCNA either gain or loss, and further subdivided into total, clonal and subclonal categories of each is determined. This value was assumed to reflect the probability of that category of SCNA event in each tumor. Based on this probability, an aberration state (SCNA category present or absent) for each sample was generated and the proportion of samples exhibiting this chosen category of SCNA event was determined. This process was repeated 10,000 times to obtain a background distribution reflecting the likelihood of observing an SCNA of the chosen category given the probability of that SCNA category occurring in each sample. A p-value reflecting the likelihood of observing the level of the SCNA category under consideration was determined by counting the percentage of simulations showing a higher proportion of the genome affected by the SCNA category under consideration than that observed.

2.2.13 Testing for within-sample subclonal LOH

Two one tailed t-tests were used to assess whether the raw allele specific copy number of the minor allele in suspected region was both significantly greater than a raw allele-specific copy number threshold of 0.2 and significantly less than a raw allele-specific copy number threshold of 0.8. The segment's raw minor allele copy number was then tested against all other segments with the same rounded value of integer minor allele copy number using a t-test to ensure any significant result was not simply the result of an incorrect ploidy or purity estimate.

2.2.14 Generating arm-level summaries for mathematical modelling

The average copy number of each chromosome arm in all samples was determined by taking the average copy number of each segment weighted by the proportion of the chromosome arm it occupied.

2.2.15 Description of Markov chain model that incorporates arm level events

Dr Sergi Elizalde and Saumel Bakhoum adapted a Markov chain which has been previously described (Elizalde et al., 2018) that keeps track of the distribution of the

number of copies of a given chromosome arm. Below is an example for a given chromosome arm (e.g. 1p).

States of the Markov chain are triples of (a,f,tf), where

a = number of attached copies of 1p (either part of chromosome 1 or of a neo-chromosome that contains 1p),

f = number of free copies of 1p,

tf = total number of any free arms in the cell; this number is positive if there are free p arms, and negative if there are free q arms.

There is an additional state corresponding to dead cells. Cells are considered dead if a+f, which is the total number of copies of 1p, goes below 1 or above N.

Each step of the Markov chain corresponds to one generation. The transition probabilities are computed from the following scenarios. At each step, each cell dies spontaneously with probability computed from its fitness score, as previously described (Elizalde et al., 2018) this time incorporating arm-specific scores rather than whole chromosome scores. The contribution of a given chromosome arm (e.g. 1p) to the survival probability of the cell is exp(d*score1p*(a+f)).

Assuming the cell survives, it undergoes whole genome duplication with probability pGD, otherwise it divides and the values a,f,tf are updated as follows:

- Each copy of a whole chromosome or neo-chromosome containing 1p breaks at the centromere with a probability psplit.

- Each copy of a whole chromosome or neo-chromosome containing 1p missegregates with probability pmisseg.

- Each free arm 1p and 1q missegregates with probability 0.5 as these broken arms cannot form proper attachments to the mitotic spindle.

- The value of tf is updated by estimating chromosome breakages and free arm missegregations in the whole cell.

- Any remaining free arms 1p fuse with any existing free q arms in the cell. In particular, after these fusions, the resulting state (a,f,tf) cannot have f>0 and tf<0 simultaneously, since in that case the free p arms would fuse with the free q arms.

Starting in state (A,0,0) and running the Markov chain for g steps, we obtain the probability distribution on the states of the Markov chain after g generations when the founder cell has A copies of arm 1p. The probability that a random cell in the colony after g generations has B total copies of arm 1p is then equal to the sum of the probabilities of states (a,f,tf) with a+f=B, divided by the sum of the probabilities of all states (a,f,tf) with 1<=a+f<=N, which correspond to live cells. Deviance scores were computed to take into account the square of the differences in copy number between the average predicted karyotype and the actual sample karyotype and normalized to the neutral evolution model.

Model Parameters

These parameters were based on the previously developed model (Elizalde et al., 2018) pGD, pmisseg, and g (below) were empirically derived to minimize the deviance between the predicted and actual values. Robustness analysis indicated that our primary conclusion is robust over a very wide range of genome duplication rate, number of generations, and chromosome missegregation rates.

Maximum number of allowed copies of any given arm: N=8

Missegregation rate: pmisseg=0.00422

Probability of a whole chromosome splitting into 2 arms at a given cell division: psplit = 0.4*pmisseg

Missegregation probability for free broken arms: 0.5

Parameter used to translate chromosome scores into survival probabilities: d=0.00039047, as computed previously (Elizalde et al., 2018) and (Laughney et al., 2015).

Arm TSG-OG scores: as derived from (Davoli et al., 2013)

Probability of whole genome duplication at each cell division: pGD=0 in nGD cases, pGD=0.006 in GD and scGD cases.

Number of generations: g=100 for nGD and GD data, g=300 for scGD.

2.2.16 Fragment analysis validation of MSAI results

MSAI in patient T_KTX169 was validated using a polymorphic Sequence-Tagged Site (STR) marker D14S306 located on chromosome 14. 20ng of patient germline and tumor region DNA was amplified using the PCR. The PCR was comprised of 35 cycles of denaturing at 95C for 45 seconds, followed by an annealing temperature of 55C for 45 seconds and then a PCR extension at 720C for 45 seconds. PCR products were separated on the ABI 3730xl DNA analyzer. Fragment length and area under the curve of each allele was determined using the Applied Biosystems software GeneMapper v5. When two separate alleles were identified for a particular marker, the fragments could be analyzed for AI using the formula $(A_{tumor}/B_{tumor})/(A_{normal}/B_{normal})$. The output of this formula was defined as the normalized allelic ratio.

2.2.17 SCNA phylogeny construction

Phylogenetic trees and ancestral genomes were reconstructed from phased allele-specific copy-number profiles using MEDICC v1.0 (Beerenwinkel et al., 2014). Tests for star topology were carried out using MEDICCquant (Beerenwinkel et al., 2014). Extraction of individual events and other intermediate processing was done in R using packages APE (Paradis et al., 2004) and igraph (Nepusz, 2006).

Chapter 3 mPHASE: quantifying allele specific and total copy number heterogeneity

3.1 Introduction

Efforts to quantify SCNA heterogeneity from bulk sequencing of multiple regions from the same patient's disease have, until recently, relied on processed calls from SCNA calling algorithms designed to analyze only a single bulk sample from a tumour region. The processed output, normally integer or continuous values of TCN or ASCN, from these multiple separate runs of the SCNA caller on different regions are then combined and compared across regions. This processed output is further categorized by thresholds based on TCN that often include: deep loss or homozygous deletions, loss, gains and amplifications though the exact thresholds may vary between studies and tools (Mermel et al., 2011). Generally, heterogeneity of ASCN is only quantified in terms of LOH, if it is examined, though there are some notable exceptions (Schwarz et al., 2015). These processed data are then compared across regions to quantify SCNA heterogeneity as well as, on some occasions, to infer SCNA phylogenies.

However, the approach of running the SCNA calling algorithm on each tumour region in isolation from all others from the same patient's disease will miss the opportunity to leverage information across regions. This may result in a less accurate estimation of the SCNAs present in each sample and erroneous classification by the various thresholding approaches often used. A more accurate picture of the SCNAs present in each tumour region from a multi-region case would, by definition, allow an equal or better-quality quantification of SCNA heterogeneity.

To date, efforts to call SCNAs jointly from multiple regions of a patient's disease at the same time have been limited, perhaps because of the relative paucity of multi-region studies compared to those using single region data. However, two approaches which attempt aspects of joint multi-region SCNA inference from bulk samples include ReMixT (McPherson et al., 2017) and, at the time of writing still a preprint released before peer review, asmultipcf (Ross et al., 2017).

Structural variants and SCNAs are inherently connected, with unbalanced SVs producing SCNAs, and loss or gain of "neochromosomes" affected by SVs causing SCNAs appearing to affect sections of the chromosomes from which the genetic

material they carry originates. Indeed, the traditional approach of considering and displaying SCNAs in the order of the position of the DNA affected in the normal genome is, to an extent, imperfect given the likelihood that the genetic material will have been affected by structural variants and rearranged. However, given the limitations of most data used to call SCNAs: arrayCGH, SNP arrays, targeted sequencing and WES which are unable or have very limited ability to detect actual breakpoints resulting from SVs, this is understandable. However, with WGS that covers intergenic regions, breakpoints can be identified at base-pair resolution which often allows the determination of the exact starts and ends of corresponding SCNAs.

ReMixT is an SCNA calling algorithm designed for bulk WGS and can run on single or multiple regions. It aims not only to call SCNAs but deconvolute clonal and subclonal populations present within tumour regions. In order to achieve this, it uses the exact breakpoints caused by SVs (that WGS has the ability to detect) alongside equivalent measures of traditional LogR and BAF to obtain clone- and allele-specific segment copy numbers, and clone-specific breakpoint copy number. ReMixT, through at least partially defining SCNA boundaries by actual SV breakpoints identified at the nucleotide level, removes much ambiguity regarding the start and end positions of SCNAs and by including SVs found in different regions leverages information across all of those given to it as input and provides a joint segmentation present across all regions.

A different approach to improving segmentation is taken by asmultipcf, which is designed to remedy a frequent segmentation issue encountered in multi-region studies using tools designed for single regions. When these SCNA callers are run on multi-region data the inferred breakpoints of SCNAs (that are may be identical in origin) differ between regions due to experimental noise (Schwarz et al., 2015). This is especially problematic when the regions are interrogated with another modality apart from WGS (or with WGS when not examined for SV breakpoint calls) and is likely to have a greater effect with technologies that offer a lower resolution of SCNAs e.g. WES compared to SNP arrays. The idea executed by asmultipcf, is to ensure a joint segmentation is present in all regions examined unless there is significant evidence that the breakpoints differ.

While both algorithms attempt to improve the accuracy of SCNA breakpoints across multiple regions, ReMixT takes an additional step to reducing potentially false copy number heterogeneity: keeping track of alleles across tumour regions. Like Battenberg and some versions of ABSOLUTE it uses haplotype imputation with

SHAPEIT2 (Delaneau et al., 2013) and reference set from the 1000 genomes reference set. As discussed previously this allows very accurate BAF quantification that improves subclonal copy number calling within a tumour region. Unlike previous mutli-region studies using imputation phasing (Yates et al., 2015, Gundem et al., 2015), ReMiXT takes this further by keeping track of the phased alleles obtained across multiple regions. This is an important advance as distinguishing alleles as either major (more copies) or minor (fewer copies) is not insufficient when comparing multiple regions. A copy number profile of 2:1 (2 major and 1 minor) in the first region might be different from a copy number profile of 2:1 in a second region if a different allele has been gained.

In this chapter I will present novel methods that acquire and maintain a phasing of alleles across multiple regions of the same patient's disease as well as address multi-region segmentation issues to quantify SCNA ITH. I will also show methods I implemented and applied earlier in my studies which follow previous approaches of analysing single regions in isolation and comparing their processed output published in two co-first author papers (Murugaesu et al., 2015, Jamal-Hanjani et al., 2017).

The multi-region phasing aspect of this chapter has been published as a part of a co-first author publication (Jamal-Hanjani et al., 2017) and its extension to multi-region copy number calling and SCNA heterogeneity quantification is part of a recently submitted paper.

3.2 Overview of and applications of multi-region phasing (MRP)

3.2.1.1 Imputation phasing

SCNAs often result in AI when genetic material from one homologous chromosome is present at a greater copy number (the major copy number or allele) and the other is present at a lower copy number (the minor copy number or allele). The allele frequencies from heterozygous SNPs where the non-reference base is present on the homologous chromosome which contributes to the major copy number are higher while those that are present on homologous chromosome that contributes to the minor are lower forming two different distributions.

The difference between observed distributions of allele frequencies are made more or less distinct by a number of factors: the read depth, the difference in copy number between the major and minor and the proportion of cancer cells carrying the SCNA present in the sample sequenced. When these differences are clear it is easy for

SCNA calling algorithms to separate the two distributions as coming from different haplotypes and detect AI.

However, as the BAFs approach 0.5, which may occur in cases where an SCNA is subclonal and present in a low proportion of the cancer cells of the tumour region sequenced, the two distributions increasingly overlap. Multiple groups have addressed this issue by assigning the non-reference bases present in the sequencing data to distinct haplotypes using tools such as such as SHAPEIT2 (Delaneau et al., 2014) to statistically infer which haplotypes are present from known reference sets (e.g. 1000 Genome Project Consortium). This allows these overlapping BAF distributions to be distinguished once more and subclonal SCNA events detected.

3.2.1.2 SCNA heterogeneity revealed only by multi-region approaches

In addition to the within-region forms of subclonal SCNAs that can be characterized using single phased tumour regions, two forms of between-region subclonal SCNA can only be detected in multi-region data: illusion of clonality and mirrored subclonal allelic imbalance (MSAI).

SCNAs that demonstrate illusion of clonality are present in one region but are the absent in another, if only the region in which the SCNA is present was assayed it might be thought to be present in all regions of the tumour and thus give an illusion of clonality. These illusion of clonality events can be picked up by approaches designed for single regions applied to multi-region data. However, MSAI detection requires tools designed specifically for multi-region data.

3.2.1.3 Multi-region phasing allows MSAI detection

MSAI refers to the pattern of AI observed when two regions from the same patient's disease both demonstrate AI but a different homologous chromosome is responsible for the major copy number in each. This results in a "mirrored" or inverse appearance of BAF profiles if the identity of the SNPs with their BAF in the upper peak and lower peaks of the bimodal distribution of one region chosen as the reference sample is maintained across the two regions. This is because the separation of the two peaks of the BAF distribution due to AI resulting from SCNAs in one region effectively phases the variant bases on each homologous chromosome. The two peaks of the BAF distribution correspond to two long-range phased haplotypes that may extend the entire length of each homologous chromosome, depending on the genomic extent of the SCNA. Furthermore, the validity of this phasing can be compared by applying

it to another region of the tumour that demonstrates AI, if the BAF from the SNPs now thought to be on the same chromosome cluster together again then the phasing is likely correct.

An instance of MSAI may occur through the presence of different combinations of SCNA events in the tumour regions from the same patient's disease that are considered. The only constant is that parental allele that results in the highest copy number in areas of AI differs between at least two regions. This may occur if two different tumour subclones lose the same section of the genome but from different paternal alleles: a parallel loss event (Figure 3-1).

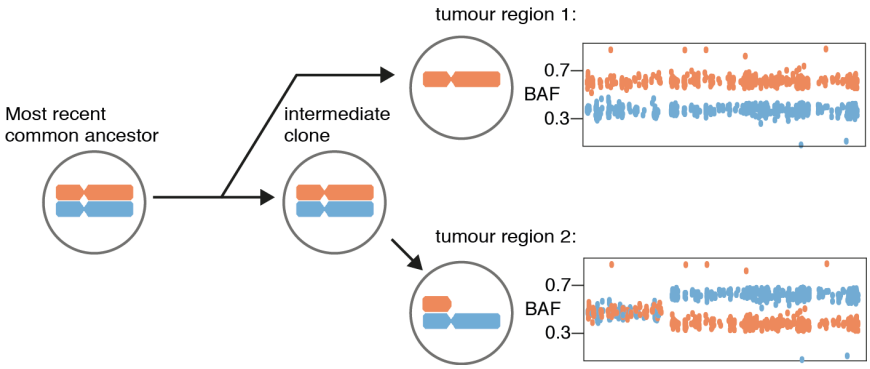

Figure 3-1: **MSAI from parallel losses**
A schematic of how MSAI might arise through parallel single copy losses of different parental alleles in distinct tumour regions with similar levels of tumour purity. The presence of normal diploid cells with a balanced copy number state in each region prevent the BAF of heterozygous SNPs in these areas of LOH approaching 1 for the remaining allele and 0 for the lost allele.

A similar parallel manner in which MSAI might occur is if two separate tumour subclones gain the same section of the genome, but independently and from different paternal alleles (Figure 3-2). These parallel loss and parallel gain events that can be revealed through MSAI detection are instances of parallel evolution as the same genetic event is occurring in independent tumour subclones each descended from a single ancestral tumour cell. However, it must be noted that these parallel gain and loss events are not the only manner in which MSAI can occur, a gain of a section of genetic material in one tumour region from one parental allele and a loss of an overlapping section of genetic material in another tumour region from the same parental allele will also result in MSAI and would not demonstrate parallel evolution. This is because though the same section of the genome is affected, the events occurring independently are not the same (a gain in one region, and a loss in another) and therefore these events do not demonstrate parallel evolution. However, all MSAI

events detected represent additional SCNA heterogeneity that would not be detected if the tumour multi-region

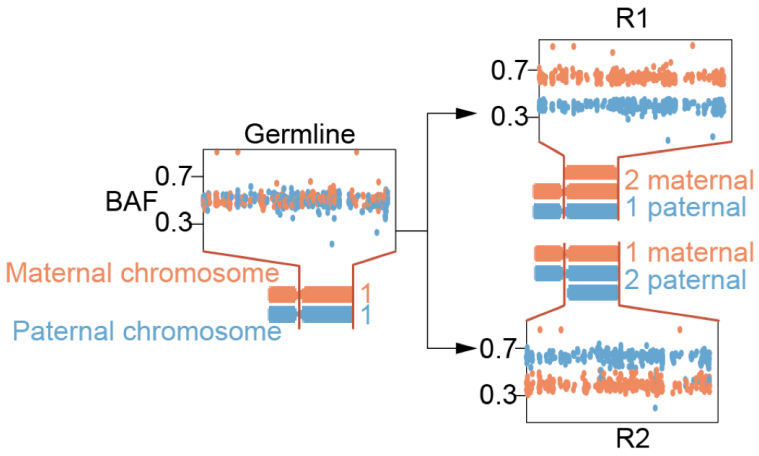

Figure 3-2: MSAI from parallel gains
A schematic of how MSAI might arise through parallel single copy gains of distinct parental alleles in distinct tumour regions. In this case high levels of tumour purity in both regions with few normal diploid cells in with a balanced copy number state result in the BAF of heterozygous SNPs in these areas of AI approaching 0.66 for the gained allele and 0.33 for the lost allele.

Equally important to note is that while some parallel gain or parallel loss events in different tumour regions will result from SCNAs in affecting different parental alleles and be detectable through multi-region phasing, many will not. As such, assuming the allele itself has no positive or negative fitness effect, only 50% of parallel events will result in MSAI.

3.2.1.4 Multi-region phasing allows calling of accurate phased copy number across multiple tumour regions

This multi-region phasing which highlights instances of MSAI can also be used to re-estimate allele-specific copy number in order to take the switch in which homologous chromosome makes up the major copy number in different regions into account. This keeps the alleles consistent across tumour regions much like ReMiXT and therefore detects heterogeneity that might otherwise have been missed.

The use of MRP also allows haplotypes to be used to identify AI when the upper and lower peaks of the BAF distributions begin to approach 0.5 due either to within-sample subclonal SCNAs or low purity affecting within-sample clonal SCNAs. With further work, similar within-sample subclonal SCNA calling approaches to Battenberg (Nik-Zainal et al., 2012) could potentially be implemented. However, in its current implementation the application of MRP allows within-sample clonal SCNAs resulting

in AI to be picked up in low purity samples. Examining low purity samples with a similar level of detail (the same haplotyping information) as those with high purity will reduce the instances of false heterogeneity being introduced to due to differences in the ease of detection of SCNAs between regions if analysed in isolation.

3.2.1.5 Overview of multi-region phasing in mPHASE: MSAI detection and phased SCNA calling and re-segmentation

While MSAI is often simple to detect visually by applying a multi-region phasing determined from the AI from one tumour region to all others, it is more complicated to optimally detect and quantify in an algorithmic manner. The optimum reference region to detect one instance of MSAI on one chromosome between two tumour regions may not be the ideal reference for another MSAI instance on a separate chromosome. Indeed, the optimum reference region to apply to all other regions may vary within chromosomes.

In order to take this into account I designed a set of scripts that are now part of a larger unpublished bioinformatics tool called mPHASE. Briefly, these produce a consensus segmentation across tumour regions, identify the optimum region to act as the reference for each segment, attempt to detect MSAI and then re-estimates SCNAs in all regions from the patient's disease using the multi-region phasing

The next sections will outline the input required by mPHASE for these tasks that are reliant on the MRP and exact process in which task performed.

3.2.1.6 Input to mPHASE multi-region phasing and MSAI detection

The input requires four separate sets of data:

1) The ASCAT or equivalent copy number segmentation for all tumour regions to be examined.
2) The purity (proportion of cancer cells present in tumour region) estimate and the ploidy (the mean total copy number across all chromosomes) estimate for each sample being investigated.
3) The LogR and BAF values of all data points passed to ASCAT or an alternative) tool used to infer the purity, ploidy and the SCNA segmentation for each sample.
4) Whether the samples being examined were sequenced using SNP array or NGS and if SNP array was used the correct value of gamma (a parameter

that accounts for technological differences and refers to the compaction of LogR profiles) must be provided.

3.2.1.7 SCNA segmentation input format

ASCAT or ASCAT format integer copy number estimates and segmentation data with the following columns:

sample: a unique identifier for each tumour sample from the patient being investigated. There must be a minimum of two for MSAI detection or multi-region phasing to occur.

chromosome: an integer representation of autosomes 1 to 22 with 23 for the X chromosome and 24 for the Y chromosome.

start position: an integer representation of the start of the genomic segment called by ASCAT or an equivalent.

end position: an integer representation of the end of the genomic segment called by ASCAT or an equivalent copy number calling tool.

nMinor: this is the integer copy number estimate of the least common allele at this position produced by ASCAT or an equivalent.

nMajor: this is the integer copy number estimate of the most common allele at this position produced by ASCAT or an equivalent. This column will always be greater than or equal to the nMinor column.

3.2.1.8 Sample summary information input format

An input file with the following columns:

sample: a unique identifier for each tumour sample from the patient being investigated. There must be a minimum of two for MSAI detection or multi-region phasing to occur.

ACF: the aberrant cell fraction or purity, the proportion of cancer cells thought to make up the bulk sequencing that the copy number segmentation for each sample was inferred from produced by ASCAT or an alternative.

Ploidy: the mean total copy number weighted for chromosome size across the all chromosomes in each sample as called by ASCAT or an alternative.

3.2.1.9 Sample BAF and LogR input format

An input file with the following columns:

sample: a unique identifier for each tumour sample from the patient being investigated. There must be a minimum of two for MSAI detection or multi-region phasing to occur.

chromosome: an integer representation of autosomes 1 to 22 with 23 for the X chromosome and 24 for the Y chromosome.

position: an integer representation of the genomic position of the data point on its particular chromosome.

LogR: This should be the same LogR data passed to ASCAT or an equivalent copy number caller used to generate the SCNA segmentation for the sample. In the case of NGS, this will be the GC corrected LogR values derived from sequencing coverage comparisons between the tumour and normal samples. In the case of SNP arrays this will be the final output LogR after all processing steps have been followed.

BAF: This should be the same BAF data passed to ASCAT or an equivalent copy number caller used to generate the SCNA segmentation for the sample.

3.2.1.10 Generating a consensus segmentation profile from multiple samples

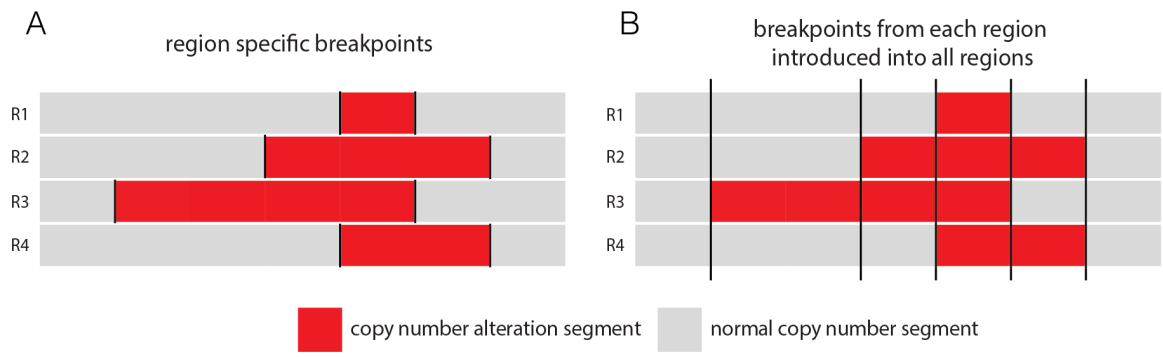

Figure 3-3:Creation of consensus segmentation
Schematic showing the process of creating a consensus segmentation from all regions subjected to copy number analysis.

Breakpoints from each individual tumour sample's SCNA segmentation profile are combined to create a single patient specific consensus segmentation profile. This profile was then applied to all tumour samples while maintaining their original integer allele specific copy number calls at all genomic positions. This allows the data in the genomic regions behind all SCNAs to be compared across regions even if they were initially not called as present in subset of regions.

This approach is vulnerable to the issues that asmultipcf (Ross et al., 2017) seeks to address, that identical SCNAs may have differing breakpoints in different regions due to experimental noise and may produce a multitude of small segments that are difficult to compare between regions.

3.2.1.11 Selecting the optimum reference region for multi-region phasing

An optimal region to act as the reference region will produce an accurate phasing if it has a clearly bimodal distribution with clear a higher and lower peak of BAFs representing the variant alleles present on each homologous chromosome.

In order to be confident of obtaining a high-quality phasing several there are several conditions that must be met before a segment can be considered for multi-region phasing.

1) Only regions that are called by ASCAT (or the equivalent SCNA caller used) can be considered as a reference region as having demonstrating AI, defined as the integer nMinor not equal to integer nMajor in the corresponding segment call for the region from the segmentation input.
2) There must be a minimum of 5 heterozygous SNPs within the segment for it to be considered.

The potential remaining regions are then ranked first by the p-value from Hartigan's dip test statistic for multimodality (Hartigan and Hartigan, 1985) implemented in R as package "diptest" (Maechler, 2015) and then a measure of mean absolute deviation (MAD) of the BAF in that the segment from 0.5 for each region.

3.2.1.12 Choice of samples and segments in which to detect MSAI

Two approaches were taken to describe MSAI. The first, more conservative approach applied in Jamal-Hanjani et al. was to only examine regions for MSAI in which the segmentation input indicated AI was present, defined as minor copy number not equaling major.

A second more explorative approach was also implemented in which all tumour regions, providing one region from the patient's disease met the criteria to be selected as the reference region were tested for MSAI.

3.2.1.13 Log truncated normal probability density

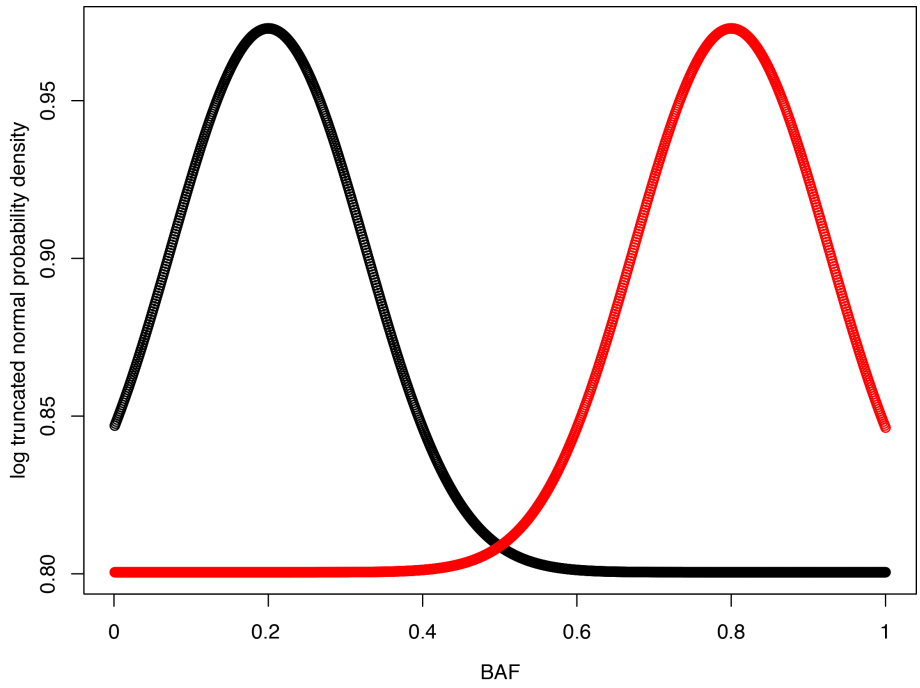

Figure 3-4: Log truncated normal probability density values for upper and lower BAF distributions
Plot showing the log truncated normal probability distribution values returned for BAFs between 1 and 0, black are the values for the LOWER input and red are the values for the UPPER input.

Figure 3-4 is a plot demonstrating the log truncated normal distributions used to determine the likelihood of a BAF value belonging to the or lower or upper peak in a bimodal BAF distribution in a region of AI. The distributions are truncated at 0 and 1. Each BAF value is assigned a log probability density value when passed to the function with a chosen sigma, and μ (which is termed LOWER or UPPER for the two different distributions in the next section referring to whether it is meant to model the lower or upper peak).

First we define our truncated normal distribution:

ϕ is defined as the unit normal distribution $N(0,1)$

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\,\xi^2\right)$$

Φ is defined as the cumulative unit normal distribution.

$$f(x;\mu,\sigma,a,b) = \frac{1}{\sigma}\ \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

And f = 0 otherwise.

For a ≤ x ≤ b the cumulative distribution of *f* is:

$$F(x;\mu,\sigma,a,b) = \frac{\Phi\left(\frac{x-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

And *F* = 0 for x < a , *F* = 1 for x > b

LTNPD (BAF, μ, σ, a, b) = log($f(\mathrm{BAF}, \mu, \sigma, a, b)$)

In the following section a = 0 and b = 1 and are therefore omitted from the argument list to LTNPD. Sigma is also unchanged in all equations and therefore omitted from the argument list to LTNPD.

For example, a call to LTNPD would appear like so:

LTNPD (BAF, μ)

Where BAF is a B-allele frequency value between 0 and 1 and μ is defined as either UPPER or LOWER.

3.2.1.14 Mirrored subclonal AI detection using multi-region phasing

There are now sets of:

N consensus segments present in all samples, ***S*** heterozygous SNPs present in all samples, and ***K*** samples.

MSAI detection is performed on a single segment and sample basis.

For each SNP ***s*** in sample ***k,*** the log truncated normal probability density of belonging to the upper peak of the bimodal BAF distribution in a segment of AI is calculated using the corresponding BAF value from SNP **s** in sample **k**, written BAF(s, k), as input:

U (s, k) = LTNPD(BAF(s, k), UPPER)

or lower peak of the bimodal BAF distribution in a segment with AI is calculated:

L(s, k) = LTNPD (BAF(s, k), LOWER)

Each segment N is phased from one of the K samples which is selected as its reference sample through the heuristic outlined in section 3.2.11.1 This reference sample is referred to as k_ref and the sample being examined for MSAI is referred to as k_test.

We then enumerate the probabilities of each SNP belonging to the upper and lower peaks of the bimodal AI BAF distribution in the k_test:

U(s, k_test) and L(s, k_test)

and in the reference sample:

U(s, k_ref) and L(s, k_ref)

The probability that each SNP is present in the same peak: upper in both or lower in both, in the bimodal BAF distributions of k_ref and k_test is calculated:

P_same(s) = U(s, k_test)U(s, k_ref) + L(s, k_test)L(s, k_test)

The probability that each SNP is present in a different peak: upper in one and lower in the other or vice versa in the bimodal BAF distributions of k_ref and k_test is also calculated:

P_diff(s) = U(s, k_test)L(s, k_ref) + L(s, k_test)U(s, k_ref)

NB it should be noted that P_diff(s, k) = 1 - P_same(s, k)

The log odds ratio of the SNP demonstrating a switch in which peak it is most likely to belong to in the bimodal BAF AI distributions of from k_test and k_ref is then calculated:

LOR(s, k) = log (P_diff(s,k) / P_same(s,k)

All SNPs within a segment will now have a LOR which is examined using a one tailed t-test:

MSAI(k) is the t-test one-sided p-value testing that the mean of (LOR(1,k), …, LOR(S,k)) is greater than 0.

In NGS data, it is known that there is often a significant bias toward higher mapping rates for the reference allele in comparison to those of the variant allele at heterozygous loci (Degner et al., 2009). Therefore, due to the possibility of some BAF values being unreliable due to local reference bias as well as the potential for other heterogeneous systematic errors at different heterozygous loci we considered more robust to assess the distribution of LORs rather than summing them.

3.2.1.15 Examples of multi-region phasing and MSAI detection

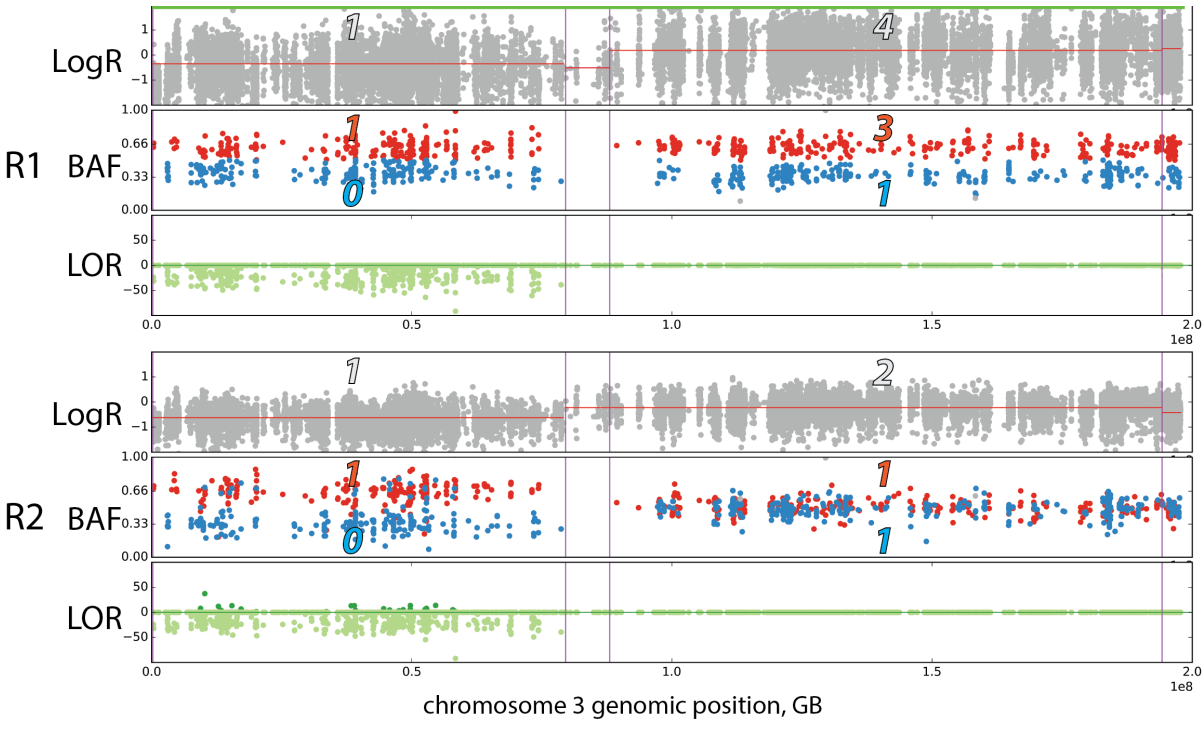

Figure 3-5: MSAI detection in the absence of MSAI
Example of MSAI detection in a pair of samples that do not demonstrate MSAI.

Figure 3-5 shows chromosome 3 from two regions of the same tumour analysed for MSAI where none is present. Each region has its own subplot with three tracks for LogR, phased BAF, and the LOR.

These tracks are separated into the 4 consensus segments present in both regions by vertical purple lines. In this case R1 was chosen as the reference for phasing in all consensus segments as indicated by the green line at the top of the R1 LogR track. The total copy number is shown by the grey integers at the top of each LogR track matching the grey colour in which all LogR is plotted for both heterozygous and 1000 Genomes loci.

The BAF tracks for region one and two maintain the same colouring for each heterozygous SNP representing the phasing obtained from the reference region, in this case R1. Orange indicates the BAF of the heterozygous SNPs that were found most likely to belong to the upper peak of the bimodal distribution in the chosen reference sample and blue represents if the heterozygous SNP was more likely to be present in the lower peak of the chosen reference sample.

The integers at the top and bottom of the BAF track in both regions' subplots indicate the allele specific copy number called for that segment by the original SCNA calling software that was provided as input to mPHASE.

The LOR is plotted in the third track for each region in light green, in the first consensus segment we can see that in R1 all the LORs are below 0 and in R2 only a small proportion of these values become positive and plotted in dark green. This segment does not demonstrate MSAI and the one tailed t-test performed on the LORs is not significant.

The next three consensus segments are not examined for MSAI as the first has no heterozygous SNPs, and the for the next two only R1 demonstrates AI with a copy number of 3:1 while R2 has 1:1 meaning MSAI cannot be detected.

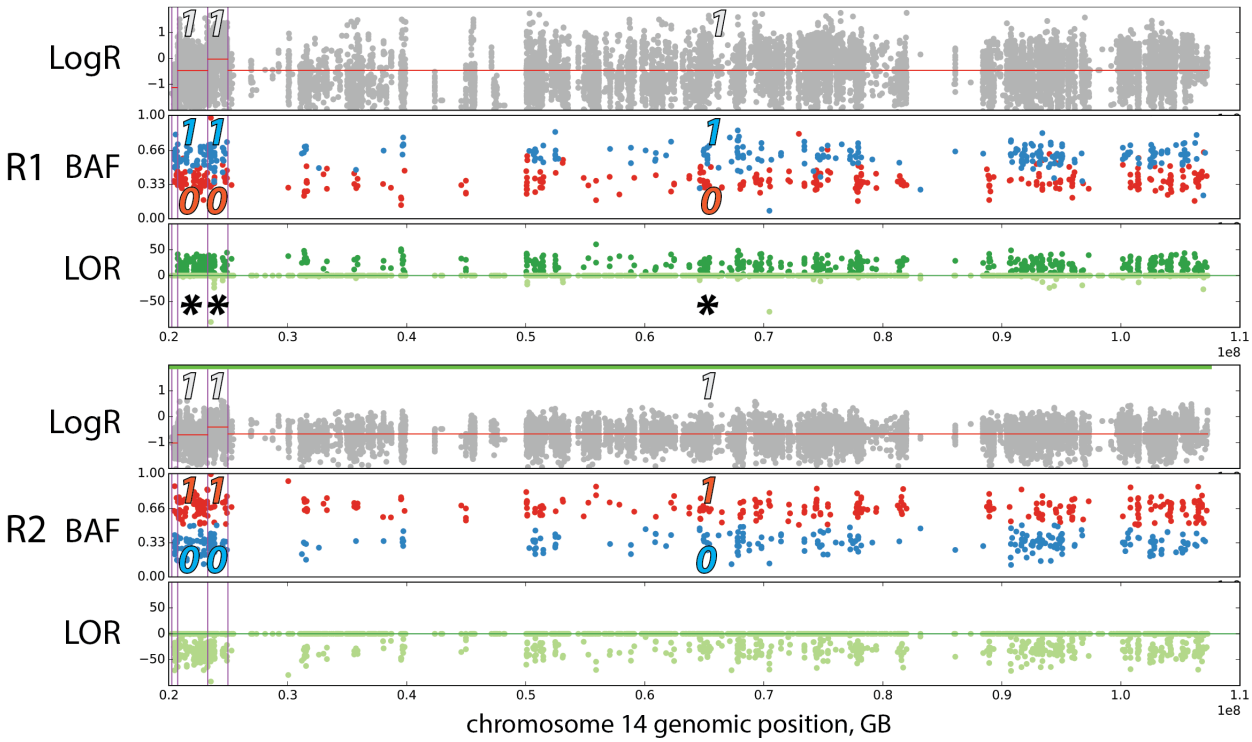

Figure 3-6: MSAI detection in the presence of MSAI
Example of MSAI detection in a pair of samples that demonstrate MSAI.

Figure 3-6 shows a successful case of MSAI detection in chromosome 14 from the same two tumour regions as Figure 3-5. In this case R2 has been chosen as the reference sample as shown by the green line at the top of the LogR track. For this chromosome, all consensus segments were analysed for MSAI as all exhibited AI with a copy number state of 1:0. As before, the phasing is of BAF values is maintained from the reference region (R2) and shown by the colouring of the plotted BAF values. There is a clear reversal in the identity of SNPs present in the upper peak of the bimodal distribution, in R1 the “blue” SNPs’ BAF values are in the upper peak while in R2 the “orange” SNPs’ BAF values are in the lower peak.

The LOR plot for R2 shows all heterozygous SNPs’ values are negative as they are in the reference, but in contrast to Figure 3-5, in R1 the majority of SNPs’ LOR values are positive and plotted in dark green. This shift in LOR distribution is picked by the one tailed t-test and all three consensus segments are called as significant (indicated by the asterisks) indicating the occurrence of MSAI in each.

3.2.1.16 Non-NGS validation of MSAI detection

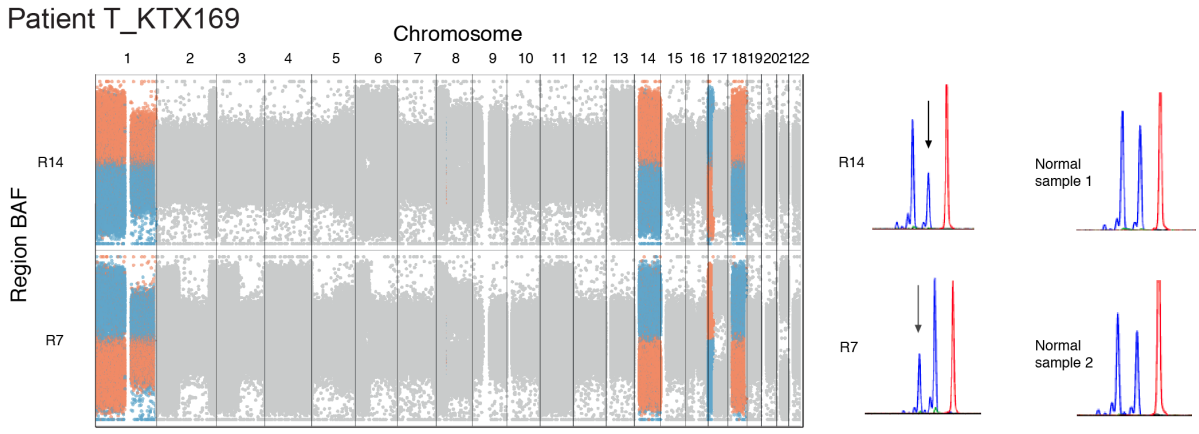

Figure 3-7: **Polymorphic microsatellite marker validation of MSAI**
BAF profiles from patient T_KTX169 - regions T_KTX169_R14 and T_KTX169_R7 using polymorphic microsatellite marker D3S2408 PCR fragment analysis profiling.

In order to verify the presence of MSAI using a method that was not dependent on NGS, a validation experiment was performed using a polymorphic microsatellite marker. The peaks indicating the two different sizes of dinucleotide repeat D3S2408 on chromosome 3, that must be present on distinct homologous chromosomes, are shown by the two blue peaks of the dinucleotide plot and the reversal in the highest peak in R7 compard to R14, that indicates the most common homologous chromosome, reflects the pattern indicated by the MSAI detection.

3.2.1.17 Revealing subtle AI using multi-region phasing

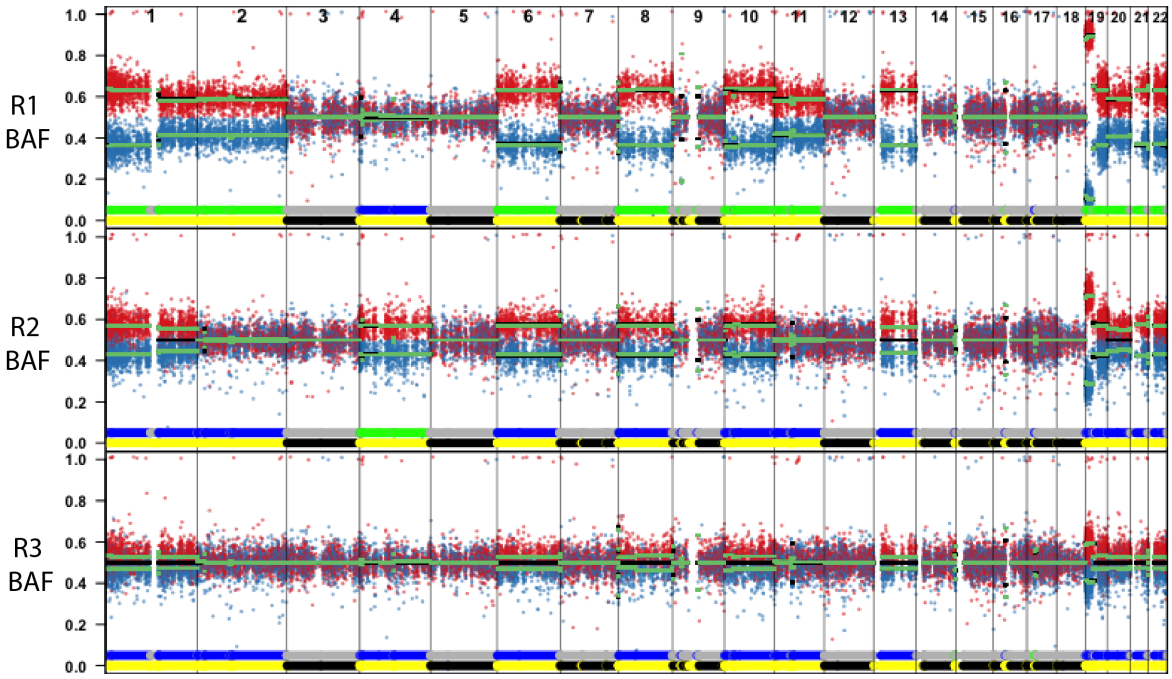

Figure 3-8: **Rescuing low purity samples using multi-region phasing**
Tumour regions of different purities with multi-region phasing applied

Multi-region phasing, as well as being used to test for MSAI in different regions of the same tumour, can also be used to reveal previously missed or undetectable AI in samples from the same patient. Figure 3-8 shows BAF profiles for chromosomes 1 to 22 from three regions of the same tumour.

These three regions have very different purity estimates from ASCAT: R1 is the purest with 72%, R2 is the next most pure with 31% and R3 has an extremely low purity of 13% of cells in the sample estimated to be cancer cells. These different purities affect the degree of bi-modality in BAF distributions in areas of AI.

At the bottom of each region's BAF plot are two coloured lines. The topmost line describes the multi-region phasing in that sample. If the line is grey it indicates that the there was no AI in any region in the input to the algorithm for that segment and therefore phasing was not possible. A green or blue line indicates that phasing was possible as there was at least one region which, in the input to the algorithm, demonstrated AI. A green line indicates that this region was chosen as the reference sample for the multi-region phasing and blue line indicates that region was not chosen as a reference.

The next line indicates whether MSAI was called in any region examined using the multi-region phasing obtained from the reference region. Yellow indicates AI was present in at least one region and it was possible to test for MSAI in the others but no MSAI was identified, red indicates that MSAI is present in at least one region. Black indicates that as no AI was present in the SCNA segmentation provided as input, no multi-region phasing could be obtained and therefore no MSAI detection was possible.

In this data, we can see that chromosomes 1, 2, 4, 6, 8, 10, 11, 14, 19, 20, 21 and 22 demonstrate clear AI in R1 which is the highest purity region. This is shown by the clear separation of the BAF into two separate distributions, the means of which are indicated by the darker green line plotted over them. Whether the original segmentation algorithm (in this case allele-specific piecewise constant fitting (ASPCF) from ASCAT) detected AI is indicated by a black line that is also plotted over the mean of the BAF distribution above and below 0.5. In cases such as R1 where the BAF is clearly separate these two lines may be overlaid with the green on top. However, if the black line is plotted at 0.5 and visible but green lines are plotted above and below this indicates that mPHASE is able to detect AI that was missed by the ASPCF segmentation performed by ASCAT. An example of detection of previously missed allelic can be seen in chromosome 13 from R2 in figure 3-8.

It should be noted that while the ASPCF segmentation performed by ASCAT may not detect some instances of AI, the final copy number segmentation inferred may still report AI in the form of unequal major and minor allele specific copy number. This is because ASCAT is able to deduce the presence of AI by inferring an odd integer total copy number through the LogR values present in the segment. Segments with an odd total copy number, as determined by the most likely purity and ploidy for that tumour region, must have inferred allele specific copy number solutions that are unequal and therefore suggest AI. However, this method of inference of AI strongly relies upon ASCAT selecting an accurate purity and ploidy and is a potentially error prone substitute for detecting the allelic balance through BAF segmentation alone.

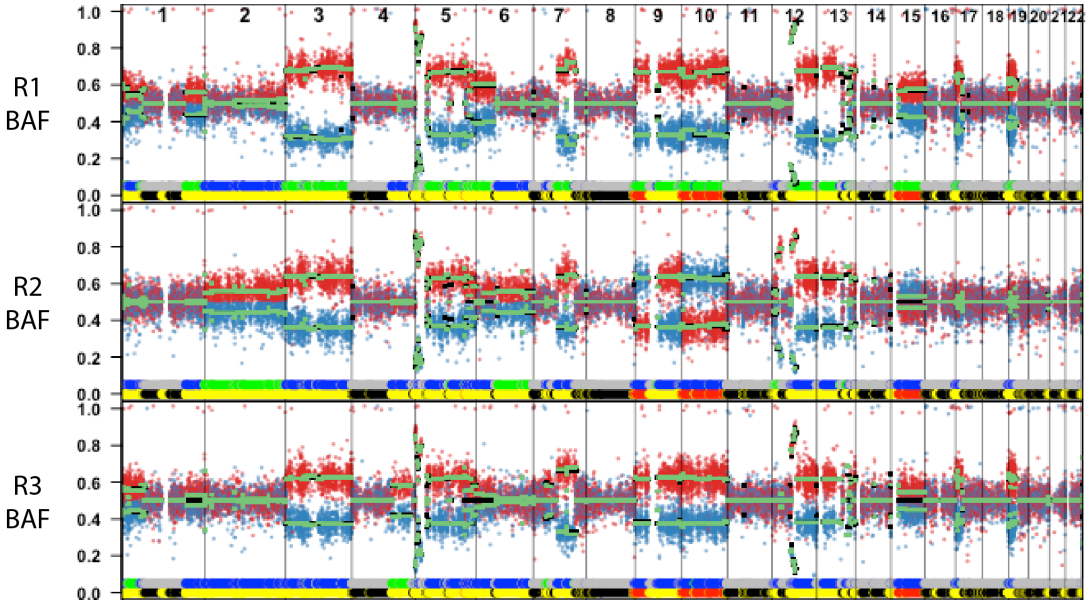

Figure 3-9: Multi-region phasing reveals MSAI affecting multiple chromosomes
Tumour regions demonstrating MSAI and AI revealed by multi-region phasing

The previous example shown in Figure 3-8 showed data from a relatively homogeneous tumour with few apparent between-sample subclonal SCNA events leading to region specific AI: chromosome 2 in R1 and chromosome 4 in R2. However, Figure 3-9 shows three separate regions from another tumour all of similar purity, that demonstrate both region specific AI events and MSAI. R1 and R3 have very similar patterns of AI while R2 diverges from them both. R2 possesses AI over the entirety of chr2 and chr3 while R2 and R3 have only partial areas of these chromosomes affected. R2 also demonstrates MSAI of 9p, chr10 and chr5 relative to R1 and R3.

Though the instances of MSAI detected on 9p and chr10 in Figure 3-8 were detected on the background of clearly distinguishable AI in all regions without needing additional resolution from multi-region phasing, the detection of the MSAI event on chromosome 15 is reliant upon it. The ASPCF segmentation run by ASCAT could only detect AI in R1, as shown by the black line plotted at 0.5 over the entirety of chr15 in R2 and R3. It is only because of application of multi-region phasing that AI can be observed and the MSAI event detected.

3.2.1.18 Estimation of allele-specific copy number using multi-region phasing

The vast majority of studies in which allele specific copy number is determined from bulk sequencing use a bioinformatics tool that jointly estimates the cancer cell fraction of the bulk sample (its purity) and the average ploidy. In contrast, mPHASE takes a hybrid approach taking both purity and ploidy estimates as input.

This allows the calculation of a pair of "raw" or continuous copy number values for each heterozygous locus that correspond to copy number estimates of genetic material from the two homologous chromosomes at that position, an "A" allele copy number and a "B-allele" copy number. SCNA estimation tools designed for single bulk samples may arbitrarily assign these and then use the higher value as the "A-allele" or major allele and the lower as the "B-allele" or minor allele. However, as mPHASE maintains a phasing of each heterozygous SNP across samples, these continuous copy number values incorporate this phasing allowing the "A-allele" to be part of the minor copy number and "B-allele" the major copy in sections of the genome affected by MSAI.

The equations that calculate the raw copy number values are the same as are used in ASCAT and use ρ = tumour purity and ψ = tumour ploidy from the values that are given to mPHASE as input for each tumour region. At each SNP site, an estimate of the phased A-allele and B-allele copy number is obtained with equations below with the corresponding LogR value for the SNP provided as input to mPHASE.

Gamma (γ) is a parameter that accounts for technological differences and refers to the compaction of LogR profiles (the expected decrease in case of deletion in a diploid sample of 100% of aberrant cells); 1 in ideal case such as for NGS, or 0.55 in Illumina 109K arrays).

$$cpnA = \frac{\rho - 1 + 2^{\frac{log_2[R]}{\gamma}}(1 - BAF)(2(1-\rho) + \rho\psi)}{\rho}$$

$$cpnB = \frac{\rho - 1 + 2^{\frac{log_2[R]}{\gamma}} BAF(2(1-\rho) + \rho\psi)}{\rho}$$

(Equations 2) Calculating continuous values of "raw" copy number for each BAF locus.

The phasing is then applied to keep the “A” and “B” copy number in line with the phasing of the BAF values obtained from the previous step.

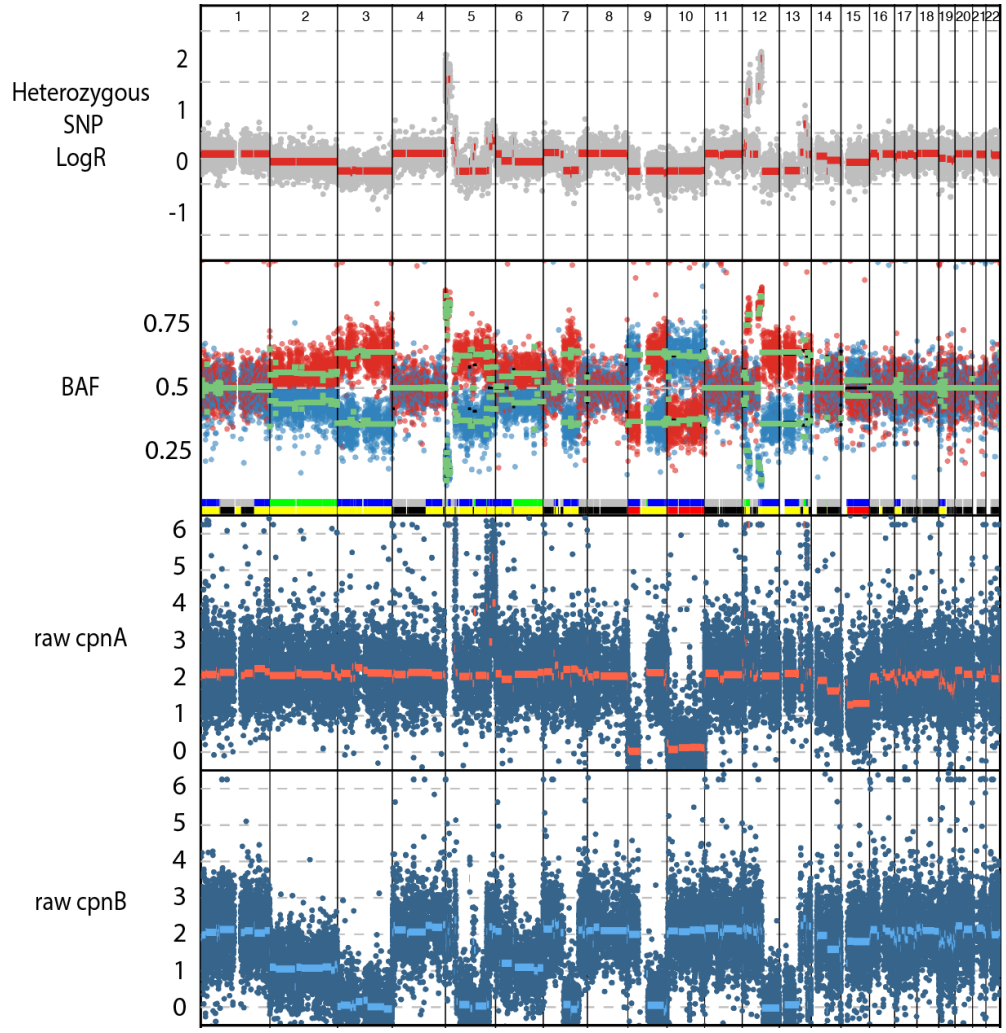

Figure 3-10: Phased estimation of “raw” allele-specific copy number
Phased estimation of non-integer “raw” allele-specific copy number using the multi-region phasing.

Figure 3-10 shows R2 from the previous figure (Figure 3-9) with the multi-region phasing still applied. The top track shows the BAF values used as input to the equations to determine phased cpnA and cpnB. The instances of MSAI previously discussed can still be observed to be on chromosomes 9, 10 and 15. The two tracks below show the continuous (or “raw”) allele specific copy number values calculated using corresponding the LogR and BAF for each heterozygous SNP locus as well as the purity and ploidy provided as input. In the raw cpnA track the median raw copy

number of each segment is indicated by an orange line and blue line is plotted in the cpnB track to indicate its median raw copy number.

Areas of AI in the BAF track can be seen to produce raw cpnA values that are higher and raw cpnB values that are lower, an example of this can be seen on chromosomes 2 and 3 where the differing degrees of BAF separation produce a raw cpnB values close to 1 on chromosome 2 and 0 on chromosome 3. However, in areas of MSAI the situation is reversed, with the raw cpnB values being higher than those of raw cpnA values; an example of this can be observed on 9p where an instance of MSAI can be observed in the BAF track and the mean cpnA values have a median value close to zero whereas the raw cpnB values are have median value close to 2. This pattern is repeated in the other chromosomes (10 and 15) demonstrating MSAI in this patient's disease.

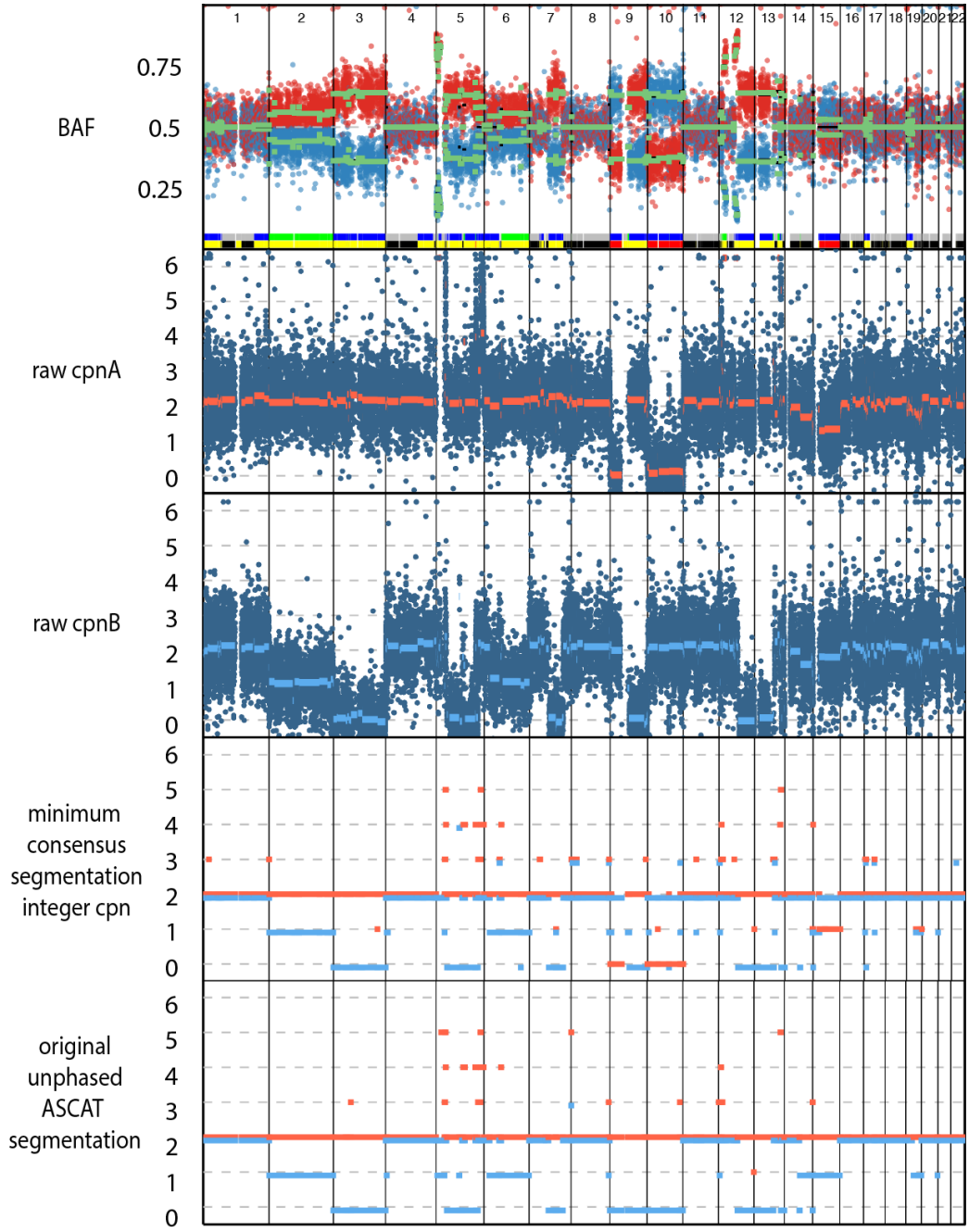

Figure 3-11: Phased integer copy number estimates
Phased estimation of absolute integer allele-specific copy number using the multi-region phasing

The final absolute integer copy number calls that ASCAT produces as output result from the rounding of the median values of raw copy number estimates within each segment. The original allele-specific copy number values for all segments across the genome from the region from the patient's disease examined in Figure 3-9 are shown in Figure 3-10 and are displayed in the last track "original unphased ASCAT segmentation". The final SCNA output from mPHASE is performed in the same manner except for taking into account the multi-region phasing derived in earlier

steps. The corresponding phased results for the same patient and region are displayed in the track labelled “minimum consensus segmentation integer cpn”.

The phased minimum consistent segmentation has more segments than the original region-specific segmentation as breakpoints have been introduced from the two other regions analysed in the mPHASE run, some of these may be seen as additional small segments at different copy number compared to the segments present in the original ASCAT segmentation. However, apart from these additional small segments, the main difference between the original ASCAT and the mPHASE profiles is that the three MSAI events detected from the BAF profiles (using data from all three regions analysed) are now present in the final segmentation output. These can be seen on chromosomes 9, 10 and 15 which now have blue (signifying cpnB values) lines at a higher copy number than the orange lines (signifying cpnA values) when the opposite arrangement is present in the original ASCAT segmentation in the track below.

The instances of MSAI detected on chromosome 9 and 10 are particularly important for the correct assessment of heterogeneity. This is because the absolute integer copy number states associated with them indicate that both are LOH events, with one allele entirely lost from the cancer cells present in that tumour region. The other two regions examined from this patient’s disease also demonstrate LOH but, as revealed by MSAI detection (Figure 3-10), from a different allele. Without utilising the multi-region phasing described earlier in this chapter these instances LOH present in these three tumour regions would have been misclassified as the same clonal and likely early event in tumour evolution. Their reclassification as parallel and independent late subclonal LOH events raises the question of how many events from single regions might be subclonal had multiple regions been taken and multi-region phasing applied.

In conclusion, multi-region phasing improves the quality of SCNA heterogeneity estimation in two ways. First, by preventing independent events being misclassified as having the same origin by identifying their occurrence on different homologous chromosomes and second, by preventing the introduction of false heterogeneity caused by purity differences between samples as shown in Figure 3-8.

3.3 Quantifying total copy number relative to ploidy heterogeneity using mPHASE

Leveraging BAF information across samples from the same patient’s disease is clearly advantageous quantifying allele-specific heterogeneity through the

identification of MSAI and rescue of SCNA detection in low purity samples. However, LogR information can also be shared between regions in a manner that may rescue detection of SCNAs in low purity samples, particularly when compared to previous methods of generating SCNA heterogeneity estimates from processed copy number calling data where multi-region data has been analysed with copy number calling algorithms designed for single regions.

Studies analysing SCNA heterogeneity have taken the output for copy number calling tools designed for single regions and compared either the absolute integer total copy number calls or non-integer "raw" total copy number calls to thresholds dependent on the ploidy of the same region (Murugaesu et al., 2015, Jamal-Hanjani et al., 2017). However, in order to make comparisons across regions from the same patient's disease, a minimum consistent segmentation must be created in similar to fashion to that outlined in section 3.2.10 by introducing the inferred segment breakpoints from all tumour regions into each tumour region's segmentation. While this produces identical start and end positions for all segments in each tumour region that allows comparison across all regions assayed in a patient's disease, it also causes the values of total and allele-specific copy number assigned to individual segments to often be, at least partially, dependent on values of LogR that are no longer contained within the boundaries of the new segment. This is because in the absence of reanalysis incorporating the new starts and end positions introduced, the copy number values from the original "parent" segment that is split into "daughter" segments by the process of minimum consistent segmentation are simply passed on to the "daughter" segments despite occupying only a fraction of the original area of the genome covered by the "parent" segment. To avoid this, mPHASE generates a consensus segmentation from each region's copy number profile and then assesses SCNA heterogeneity by examining the raw LogR data within these new consensus segments' genomic boundaries.

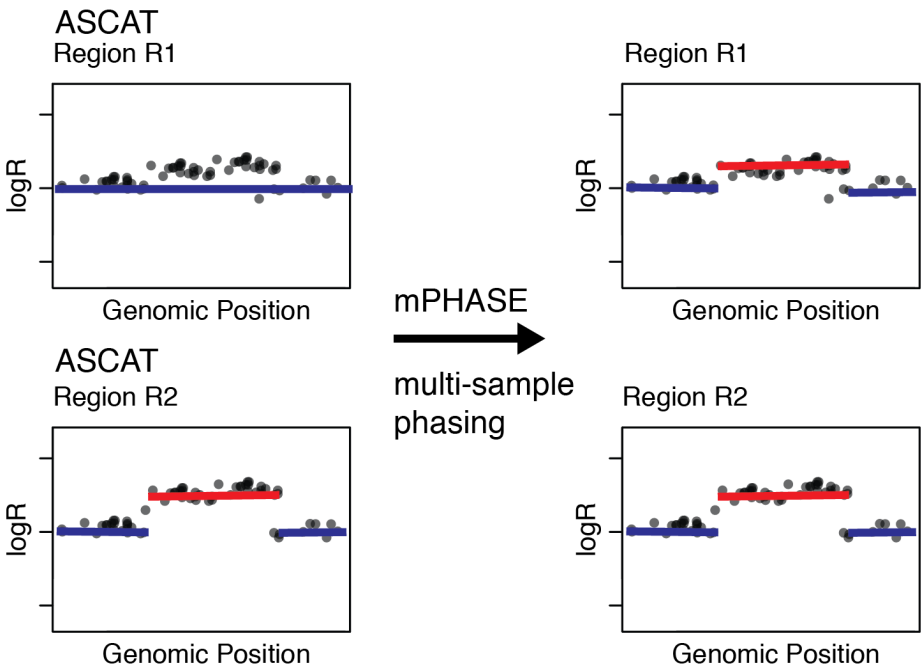

Figure 3-12: Consensus segmentation assessment of total copy number heterogeneity
Each dot represents a LogR value with a line indicating the median value of LogR, colour of the line indicates whether an SCNA relative to ploidy is found to be present (red) or absent (blue).

The LogR values within the boundaries of each consensus segment in each region can then be compared against a threshold adjusted by the corresponding purity and overall ploidy of the tumour region. This improves the comparison of copy number across regions in two ways: firstly, it ensures that only like for like information is compared between regions as only data from the same loci are compared and secondly it specifically tests for the presence of an SCNA found in one region in all others by re-examining all regions within the start and end positions of an SCNA identified any region. A schematic demonstrating both of these advantages can be seen in Figure 3-12, here we can see that in the ASCAT segmentation of R2 a small segment highlighted in red has been determined to demonstrate an SCNA. However, in R1 within the same genomic region as the SCNA segment in R2 there is a small increase in the values of LogR that alone were not sufficient to cause the creation of a new segment in the original ASCAT segmentation. However, with the creation of minimum consistent segmentation and interrogation of the LogR values through the application of mPHASE, the small localised increase in LogR values in R1 in the area of the SCNA originally detected in R2 is now found to be above the threshold for identifying an SCNA. By specifically interrogating the corresponding raw LogR data

of the loci within each and every SCNA identified across regions, mPHASE aims to avoid inferring false heterogeneity.

Four thresholds are used by mPHASE to produce 5 possible copy number states relative to ploidy: amplification, gain, neutral, loss and deep loss. The highest of these thresholds is amplification with deep loss being the lowest, neutral is the state between the loss and gain thresholds. The LogR values in each segment with a sufficient number of loci within it (the default is set to 5 but is adjustable the user) in all tumour regions are then tested against the LogR thresholds set out in Equations 3.

$$amp\_LogR\_threshold = \ log_2(\frac{4}{2})$$

$$gain\_LogR\_threshold = \ log_2(\frac{2.5}{2})$$

$$loss\_LogR\_threshold = \ log_2(\frac{1.5}{2})$$

$$deep\_loss\_LogR\_threshold = \ log_2(\frac{1.5}{2})$$

$$exp\_LogR = log_2(\ \frac{2 \times (1 - \ \rho) \ + \rho \ \times \ \psi \ \times \ 2^{threshold}}{2 \times (1 - \rho) + (\rho \ \times \ \psi)}\ )$$

(Equations 3) Ploidy and purity dependent copy number thresholds where ρ represents tumour purity and ψ represents tumour ploidy.

To generate the respective thresholds in terms of a value that raw LogR values may be compared to, a corresponding expected LogR value for amplification, gain, loss and deep loss is calculated. This is achieved by incorporating by purity and ploidy as well as an expected “raw” or continuous estimate of total copy number that would meet the definition of an amplification, gain, loss or deep loss in a diploid cell in the final equation show in *Equations 3.* The LogR values within a segment are then compared to each of these thresholds using a one-tailed t-test, ensuring they are greater than the threshold in when amplifications and gains are examined and lower when losses and deep losses are examined. A segment which is found to reach significance for an amplification will also achieve significance for a gain. Similarly, a

segment that is found to be significant for a deep loss will also be found to be significant for loss. In these cases, the more an amplification or deep loss copy number event will be reported. The resulting calls in for a single segment in each region of a patient's disease can then be compared and clonal events occurring in all regions of a patient's disease identified with other events defined as subclonal.

3.4 Integrating AI, LOH and SCNAs relative to ploidy in SCNA heterogeneity estimates

The previous sections have shown that mPHASE is able to use multi-region phasing to infer independent events and quantify allele-specific copy number heterogeneity as well as total copy number heterogeneity across multiple samples from a patient's disease taking into account the corresponding purity and ploidy estimates. However, in order to report SCNA heterogeneity accurately mPHASE reports the clonality of gains and amplifications separately to losses and deep losses. This is because it is possible, though uncommon to have to have the same genomic area lost in one region of a patient's disease and gained in another region of the same patient's disease.

An additional complication is caused when attempting to integrate the allele-specific SCNA calling information with the relative to ploidy classifications of amplification, gain, neutral, loss and deep loss. That is because every segment of the genome will have multiple different levels of classification that must be integrated. The first is the actual absolute integer copy number of the segment and whether this is different or the same between regions, in the case of mPHASE heterogeneity of actual absolute integer copy number though recorded is not reported in the final definition of whether an event is subclonal. As well as absolute integer copy number call each segment will have a classification of whether it is allelically balanced, allelically imbalanced or in an extension of AI demonstrating LOH. Any of these AI states can coexist with total copy number amplifications, gains, neutral, loss or deep loss states depending on the ploidy of the cancer cells examined, see Figure 3-13 for schematic.

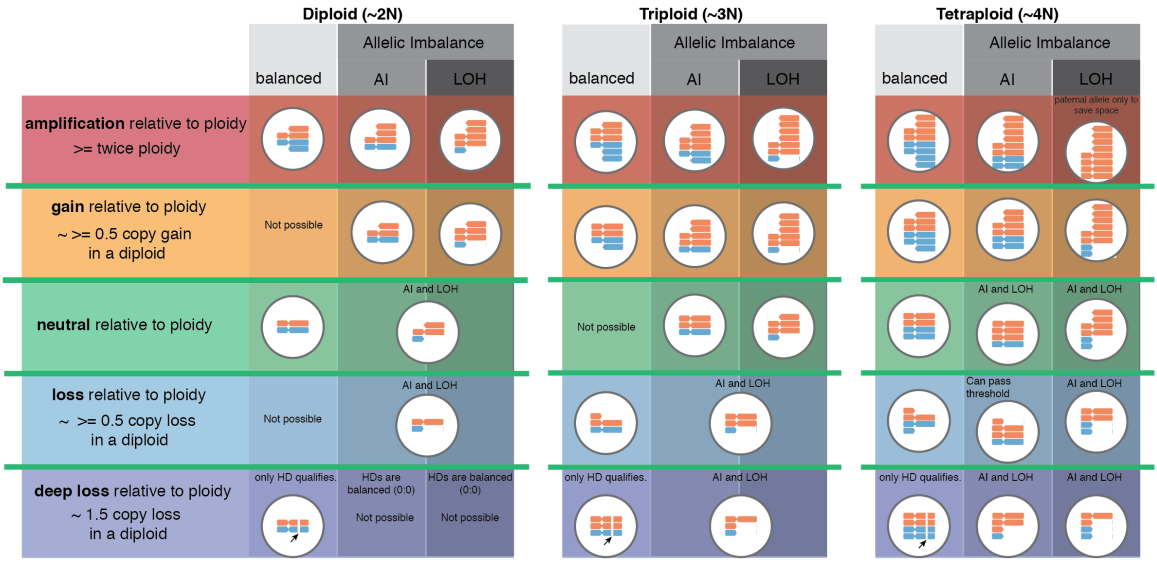

Figure 3-13: Schematic of MPhase thresholds at different ploidies
MPhase calls categories of copy number change relative to ploidy, this schematic shows how each of those can or cannot be associated with a balanced allelic state, AI and/or LOH in diploid, triploid or tetraploid cases. HD stands for homozygous deletion. Each cell shows only the chromosomal complement from one of the 22 autosomes with q-arm affected an SCNA if any is present. The two distinct homologous chromosomes are indicated by differing colours of orange and blue.

While mPHASE reports a measure of whether AI is present or absent across all regions only LOH is factored into its final definitions of the clonality of SCNA. LOH may co-exist with any of the copy number states relative to ploidy, e.g. it is possible to have a neutral copy number state relative to ploidy and LOH in all regions. In another more extreme example that is also possible, LOH may co-exist with an amplification in all regions. LOH signifies that a complete loss of genetic material from one homologous chromosome must have occurred at some point in the life history of the tumour. In contrast, AI events that do not demonstrate LOH may have occurred through gain of material from a diploid state or, from a loss or gain of genetic material from a tetraploid state after genome doubling. Therefore, given the ambiguity of the origins of such non-LOH AI and the importance of LOH events in cancer evolution in acting as the second-hit for deleterious sequence-based variants (see Chapter 1) LOH events are considered as “loss” events regardless of the copy number state relative to ploidy co-occurring with it. This results in copy number segments that may be classified as both “gain” and “loss” as a combination of these changes must have occurred in evolution of the tumour affecting the genomic area covered by the copy number segment.

The final definitions of clonality for a single segment across multiple regions of the same patient's disease integrating both LOH and copy number relative to ploidy are as follows and are binary variables indicating the presence or absence of that particular classification:

Clonal amplification: present if every region analysed from the patient's disease has a relative to ploidy classification of amplification. The LOH status in all regions is ignored.

Subclonal amplification: present if at least one, but not all, region of a patient's disease has a relative-to-ploidy classification of amplification. The relative-to-ploidy copy number status of regions not demonstrating an amplification may be ignored. The LOH status of all regions is ignored.

Clonal gain: present if every region analysed from the patient's disease has a relative to ploidy classification of gain or amplification. The LOH status all regions is ignored.

Subclonal gain: present if one or more but not all regions analysed from the patient's disease has a relative to ploidy classification of gain or amplification. The LOH status all regions is ignored.

Clonal loss: present if either all regions demonstrate a loss/deep loss relative to ploidy **or** all regions demonstrate LOH. A region may demonstrate both LOH and a loss relative ploidy and LOH and still count towards either of these definitions. However, a clonal loss may **not** consist of a combination of regions where only a subset demonstrates LOH.

Subclonal loss: present if at least one or more but not all regions demonstrate loss/deep loss relative to ploidy **or** at least one or more but not all regions demonstrate LOH. Cannot demonstrate a subclonal loss if a clonal loss has already been called for the segment.

Clonal deep loss: present if every region analysed has a relative to ploidy classification of a deep loss. LOH status in all regions is ignored.

Subclonal deep loss: present if one or more but not all regions analysed from the patient's disease has a relative to ploidy classification of deep loss.

Again, it should be noted that more than one of these classifications may apply to the same segment e.g. the same segment could be both a loss and a gain if it demonstrates gain relative to ploidy but also LOH in all regions. However, while a single segment under these definitions may be classified as both a gain and a loss it cannot not simultaneously be a both a subclonal and clonal gain or a subclonal and clonal loss.

As mPHASE requires as input purity and ploidy as well as LogR and BAF, it can be applied to any multi-region data where these values can be generated for each region. Therefore, it can be applied to multi-region cases that have been analysed with an SCNA calling algorithm designed to analyse single region bulk samples with a modality that provides sufficient numbers of heterozygous SNPs to allow allele-specific absolute integer copy number calling. Such modalities include some larger targeted panel NGS platforms (Shen and Seshan, 2016), WES, SNP arrays and WGS. If assessing heterogeneity within cases and comparing the results across a larger cohort of different patients' disease that has been assayed with the same modality all heterogeneity classifications may be used. However, this is with the caveat that if a modality is not normally used to assess a particular SCNA event e.g. WES and homozygous deletions, then the corresponding less discriminative classification should be used e.g. loss in place of deep loss. In addition, when analysing a cohort of patients' disease examined with mixed assays that provide differing levels of SCNA resolution e.g. WES and WGS (as in Chapters 5 and 6), it may be advisable to use the less discriminating heterogeneity classifications.

3.5 Conclusions

In this chapter I have outlined and demonstrated methods, grouped together into a bioinformatics tool termed mPHASE, to leverage multi-region data to improve the quantification of both allele-specific and total copy number. At the time of writing and to my knowledge, there is no equivalent multi-region copy number calling algorithm that specifically aims to characterise SCNA heterogeneity across multiple bulk samples of regions of the same tumour into summary calls of between-sample clonality incorporating both total and allele-specific copy number measures while taking into account tumour ploidy and purity. An application of the methods described to a pan-cancer cohort will be shown in Chapters 5 and 6.

Multi-region phasing provides an opportunity to identify hitherto undescribed heterogeneity and identify parallel evolution of independent events that originate from

different homologous chromosomes. However, it is not the first form of phasing employed in somatic copy number calling. Statistical imputation phasing based on sets of reference haplotypes has been used with WGS analysis of single bulk regions (and in the case of ReMiXt to multiple samples of a patient's disese) to allow the extremely accurate quantification in changes BAF . This accurate BAF characterisation is then used by many SCNA calling algorithms to identify non-integer copy number changes within single tumour regions to attempt to infer within-sample subclonal copy number calls (see Chapter 1). Multi-region phasing could also potentially be applied to address within-sample subclonality as well as the between-sample subclonality the current implementation of mPHASE described in this Chapter uses it for. Indeed, it may have certain advantages when compared to phasing by imputation:

1) It does not require WGS data and can be applied to WGS, SNP array, WES or even targeted panel sequencing depending on the number of heterozygous SNPs assayed.
2) Phasing by imputation is limited to relatively small haplotype blocks of approximately 300kb (though some methods have recently increased this with new reference panels (Loh et al., 2018)) in which to attempt to detect AI. In contrast, multi-region phasing is only limited by the size of the SCNA that caused AI in the tumour region chosen to be the reference sample. The relatively small haplotype blocks used in statistical phasing are also separated by recombination hotspots which lead to "switching" of the blocks and which cause the so-called "Battenberg" pattern which requires an extra segmentation step to identify and rectify (Nik-Zainal et al., 2012). This limits how many BAF loci can be grouped together to identify subclonal AI and is complication that would not be encountered using multi-region phasing.
3) Haplotyping by imputation is computationally onerous, though this is improving with new software releases (Delaneau et al., 2013, Howie et al., 2009) while, in comparison multi-region phasing, if choosing identities for SNPs based on a single region of AI, is trivial.
4) By not relying on a set of reference haplotypes it can be applied to patients from understudied ethnic groups for whom reference sets of haplotypes may not be available.

The long-range haplotypes acquired by multi-region phasing from bulk samples could also have applications outside of SCNA calling from multi-region bulk context and

could potentially be applied to detection of SCNAs in circulating tumour DNA, single cell sequencing approaches and in examining separate tumours from patients with germline cancer predisposition syndromes for evidence of convergent evolution.

Chapter 4 Investigating SCNAs in ESCA, NSCLC and KIRC

4.1 Introduction

In this chapter I will present the results of my evolving efforts to quantify SCNA ITH over the first three years of my PhD. This will, in the order in which they were performed, cover analyses of multi-region data to quantify SCNA ITH on ESCA published as co-first author in (Murugaesu et al., 2015), work on multi-region NSCLC data from the TRACERx 100 study published as co-first author (Jamal-Hanjani et al., 2017) and work on KIRC as part of the TRACERx renal study investigating KIRC published as a co-author (Turajlic et al., 2018b).

In the subsection of this chapter devoted to each cancer type, I describe the methods used for the characterization of its SCNA heterogeneity that produced the corresponding published results and the opportunities for improvement upon them I identified and have subsequently implemented. In all three cancer types I document extensive SCNA heterogeneity and categorise SCNAs as clonal or subclonal giving insights into their timing in tumour evolution. I also reveal multiple instances of parallel evolution of SCNA events in both NSCLC and KIRC for the first time through the creation and application of a novel multi-region phasing tool to reveal them.

4.2 Tracking the genomic evolution of oesophageal adenocarcinoma through neoadjuvant chemotherapy

4.2.1.1 Data and SCNA ITH quantification methods applied

In this study 40 regions from 8 patients (EAC001, EAC003, EAC005, EAC006, EAC007, EAC014 and EAC017) were subjected to WES and analysed for both sequence-level variants and SCNAs. This was the first multi-region NGS study of ESCA and allowed investigation into the clonal architecture and evolution of each tumour, allowing the assessment of both spatial and temporal heterogeneity. This study also allowed ITH pre-and post-treatment to be assessed as some regions were sampled at a second timepoint after chemotherapy had been administered.

I assessed the ITH and timing of SCNAs present within each patient's disease. This is also the first multi-region dataset of any kind that I had the opportunity to analyse

during my PhD and therefore the approach I took was heavily based on those of previous attempts in the literature (Sottoriva et al., 2013). Following the general approach outlined in Chapter 3, I did not perform any systematic multi-region phasing in order to keep track of alleles across multiple regions taken from a single patient's disease. Furthermore, each tumour region was subject to SCNA calling using Sequenza (Favero et al., 2015) in isolation from all other regions from the same patient. The only point at which information was shared between regions from the same patient's disease during SCNA calling was at the manual curation step, when all regions' independent potential purity and ploidy estimates from Sequenza were examined when choosing each region's final purity and ploidy solutions.

In addition, while a consensus SCNA segmentation in which inferred breakpoints for all regions analysed from a single patient's disease were introduced to all others was produced, no assessment of the raw data present within each consensus SCNA segment across all regions was undertaken to quantify heterogeneity solely within the boundaries of the consensus segment. Instead, when an original SCNA "parent" segment from a single tumour region is split into multiple "child" consensus segments by the introduction of breakpoints from other regions the same values (e.g. average LogR, total copy number and allele specific copy number) from the "parent" segment are inherited in the "child" segments. Therefore, each consensus segment may not accurately reflect the smaller subset of data within its genomic position but rather the larger segment in originated from in the SCNA calling for that region pre-consensus segmentation. However, despite these limitations in the view of my later work and development of the mPHASE multi-region SCNA caller, this study revealed that ESCA is tumour type characterized by early CIN leading to amplifications which encompass clinically targetable oncogenes that persist thought chemotherapy, potentially suggesting a future therapeutic rationale.

4.2.1.2 CIN and focal amplifications are early drivers

Oesophageal cancer has the eighth highest incidence in the world (Ferlay et al., 2015) and is associated with poor clinical outcome, usually presenting at a late stage as well as being relatively resistant to chemotherapy with five-year survival rates of 15% (Smyth et al., 2017). In addition to only modest improvements in survival over the last few decades the incidence of ESCA has increased significantly in the last twenty years (Pohl and Welch, 2005). Previous WES studies have identified sequence based drivers in ESCA (Dulak et al., 2013) and in its preinvasive forms

(Weaver et al., 2014). In some of these studies and others single bulk regions were also analysed with SNP arrays and focal SCNAs that contain kinases such as *ERBB2*, *EGFR*, *FGFR1*, *FGFR2*, and *MET* that are therapeutically targetable have been described (Dulak et al., 2012, Dulak et al., 2013, Agrawal et al., 2012). However, given the limitations of bulk assays for determining timing of SCNAs, their relative position in the life history of ESCA was unknown.

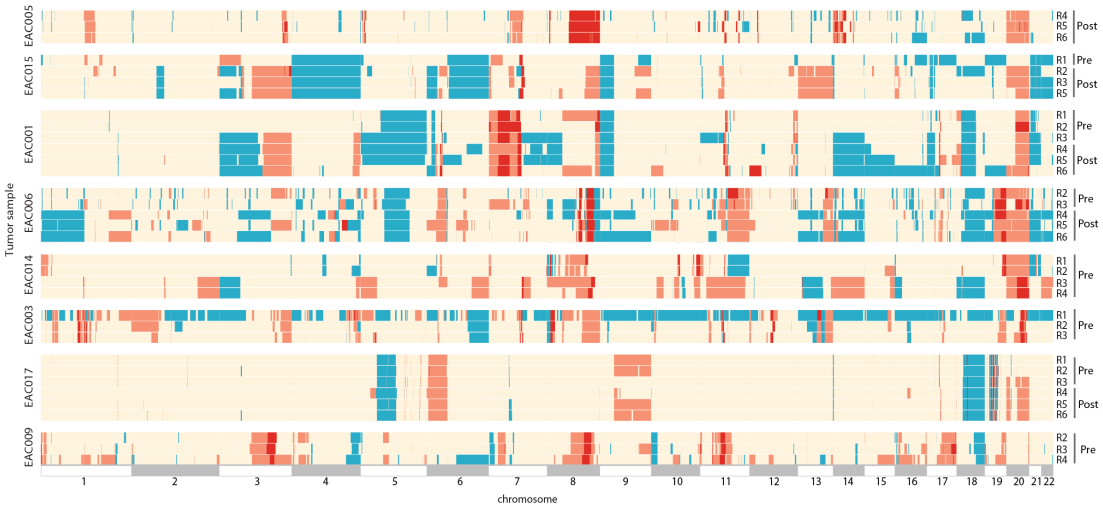

Figure 4-1: Visualising losses, gains and amplifications in multi-region ESCA data
Heat map of all regions from each of 8 multi-region exome sequenced ESCA cases analysed. Amplifications are indicated by red, gains by orange and losses by blue. Areas of the genome without any SCNAs reaching the threshold for any of these categories are represented with pale yellow.

SCNA analysis performed on all regions from each patient showed that all the cases examined showed extensive evidence of SCNAs (Figure 4-1). To compare the levels of CIN in this cohort a weighted genomic instability index (wGII) score, a single state measurement used as proxy for the process of CIN, was calculated for each tumour region. All regions analysed demonstrated a high wGII score (median 0.53, range 0.31-0.66) which was similar to that observed in samples from the TCGA (Figure 4-2). In addition, all regions showed evidence clonal *TP53* disruption either through sequence-level variants, as seen in 7 of 8 cases, or through the amplification of *MDM2* as observed in one case (EAC003), all tumour regions also showed evidence of WGD using the algorithm outlined in Chapter 2. These observations were in keeping with previous studies that had described early inactivation of *TP53* and subsequent WGD and CIN as an early event in ESCA evolution (Weaver et al., 2014, Nones et al., 2014). Chromothripsis events were observed in two cases (EAC003 on chromosome 1 and EAC017 on chromosome 19, Figure 4-1) and were found to be

present in all regions assayed from these patients suggesting that these events were also likely to have occurred early in tumour evolution.

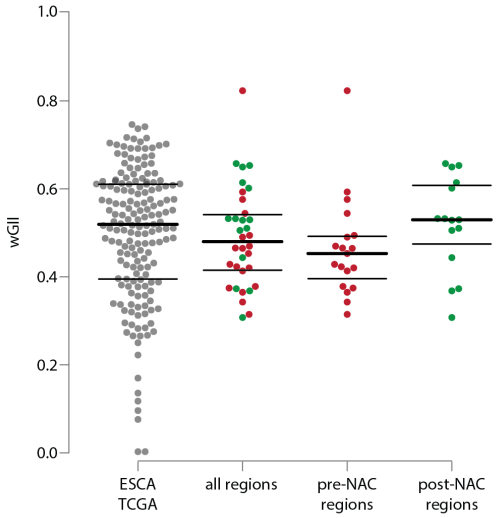

Figure 4-2: Multi-region WES SCNA burden comparison with TCGA
Comparison of weighted genomic instability index (wGII) values between single ESCA region SCNA profiles from the TCGA, all regions analysed for SCNAs in this study and the same regions divided into those take before and after neoadjuvant chemotherapy (NAC). A red dot indicates a pre-NAC region and a green indicated a post-NAC region. No significant differences were found between any of the four sets of regions.

SCNA heterogeneity was then explored across the cohort of 8 tumours. Each segment was assigned a value of SCNA "gain" that was equivalent to ≥1 copy relative to ploidy, SCNA "loss" that was equivalent to ≤1 copy number relative to ploidy or SCNA amplification that had a copy number of ≥ twice the ploidy of the region. These segments were then used to explore whether the landscape of SCNAs present in the multi-region cohort were similar to those identified in previous analyses performed on ESCA SNP array data by the TCGA (TCGA, 2015) using GISTIC2 (Mermel et al., 2011). The multi-region ESCA cohort demonstrated 15 chromosomal regions that had been previously been shown to be recurrently amplified by the TCGA to be amplified in at least one region from the 8 patients (Figure 4-3). It was notable that while the majority of recurrent amplifications identified by the TCGA when investigated in our cohort appeared to occur in all regions of a tumour, some including amplifications encompassing *KRAS* and *ERBB2* were present only in a subset of tumour regions suggesting that they occurred later in tumour evolution.

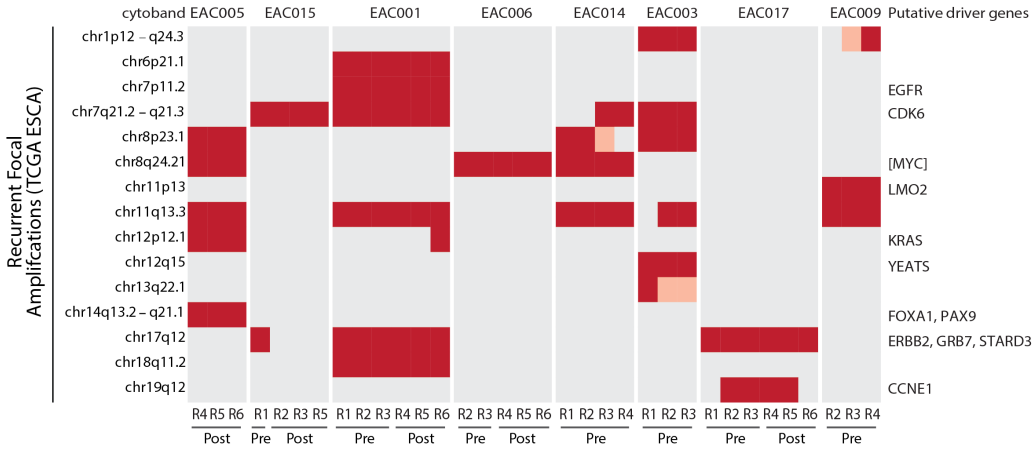

Figure 4-3: TCGA GISTIC2 focal amplification regions characterised in multi-region WES
Heat map of recurrent focal amplifications from TCGA GISTIC2 analysis (TCGA, 2015) investigated in the ESCA multi-region cohort. Shows the distribution of potential tumour driver copy number amplifications and deletions for each tumour region based on recurrently amplified and deleted chromosomal segments identified from TCGA oesophageal cancer (ESCA) data. For each region, amplification was determined as ≥2x ploidy and copy number loss was determined as ≤1 copy number, relative to ploidy.

However, when recurrent regions of loss in the TCGA were examined in our cohort using the same procedure (Figure 4-4) 47 recurrently lost areas of the genome were identified to also be lost in at least one region of the 8 patients' disease examined. Indeed, there appeared to be a much less consistent pattern compared to amplifications with many loss events occurring only in a subset of tumour regions from a patient's disease. In addition, no significantly recurrent loss event that appeared in more than one patient's disease was always ubiquitously present. Even the most significant recurrent loss from TCGA analyses (9p21.3 covering CDKN2A) was observed to occur in only a subset of regions in patient EAC006.

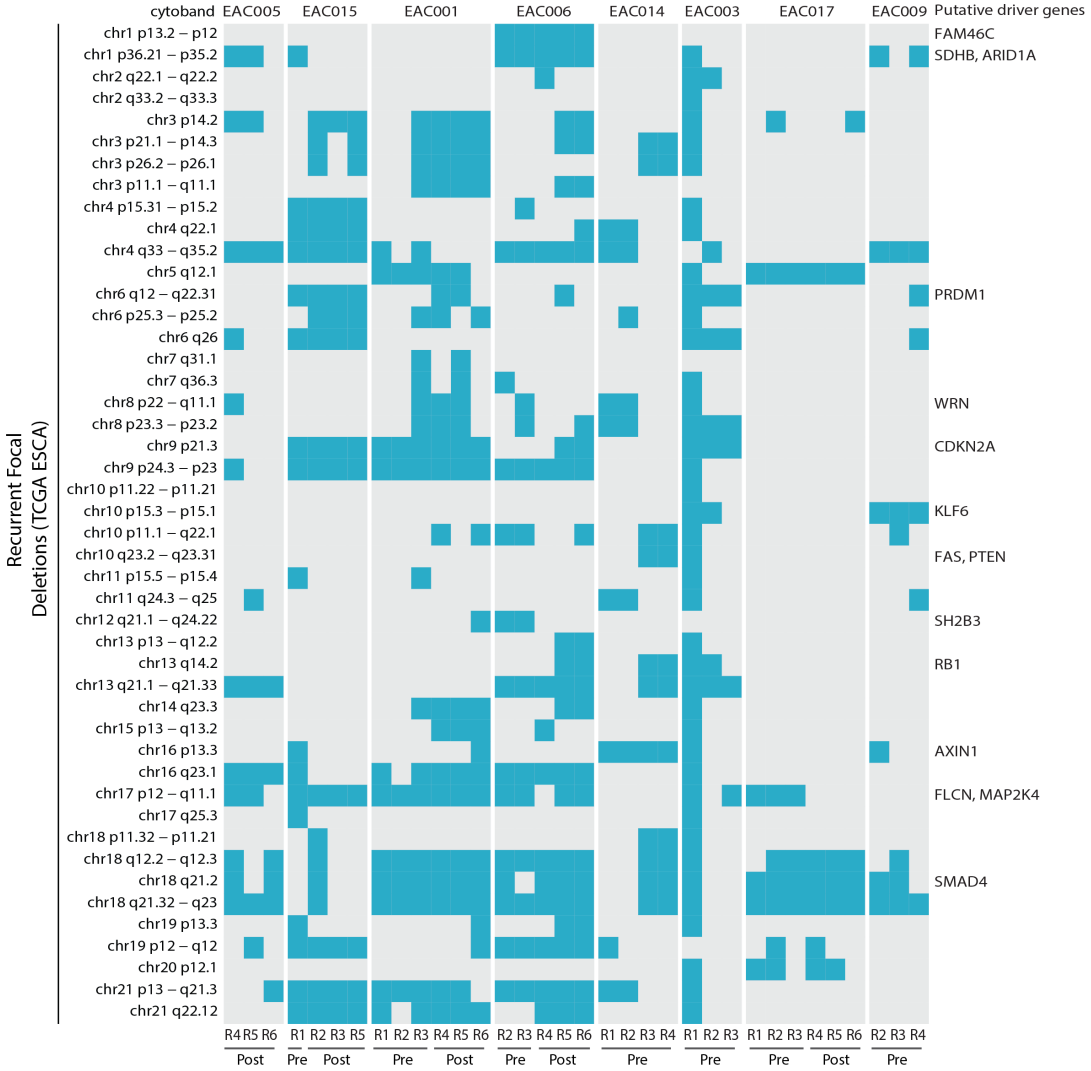

Figure 4-4: TCGA GISTIC2 focal loss regions characterised in multi-region WES
Heat map of recurrent focal losses from TCGA GISTIC2 analysis (TCGA, 2015) investigated in the ESCA multi-region cohort. Shows the distribution of potential tumour driver copy number amplifications and deletions for each tumour region based on recurrently deleted chromosomal segments identified from TCGA oesophageal cancer (ESCA) data. For each region, copy number loss was determined as ≤1 copy number, relative to ploidy.

The three categories of SCNA (amplification, gain and loss) were then examined to see if they demonstrated a different pattern of clonality and each segment as assigned a value of heterogeneous or ubiquitous across all regions. All segments, not just those previously identified as significantly recurrent in TCGA analyses were then examined. On average, only 24% of gains and 13% of loss segments were found to be ubiquitous, whereas 43% of all segments that were classified as amplified were found to be ubiquitous, these differences between the categories were found to be significant (Fishers exact, $p<0.0001$, Figure 4-5).

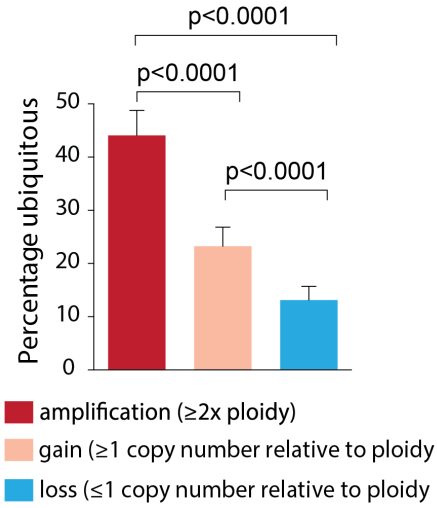

Figure 4-5: **Differing rates heterogeneity of amplifications, gains and losses**
Histogram showing the percentage of ubiquitous copy number amplifications, gains and losses (+SEM) for all tumours.

SCNA amplifications were also observed to be present in all regions both pre and post-treatment (neoadjuvant chemotherapy) and were often centred on specific oncogenes, such as *CCND1*, *EGFR*, *ERBB2*, *KRAS* and *MYC* (Figure 4-6). These ubiquitous SCNA amplifications pre and post-therapy often contained genes that have previously been identified as potentially targetable copy number alterations in the target database . However, it should be noted that some of these same targetable

genes were present in heterogeneous amplifications dephasing the potential importance of considering clonality of alterations in future treatment stratagems.

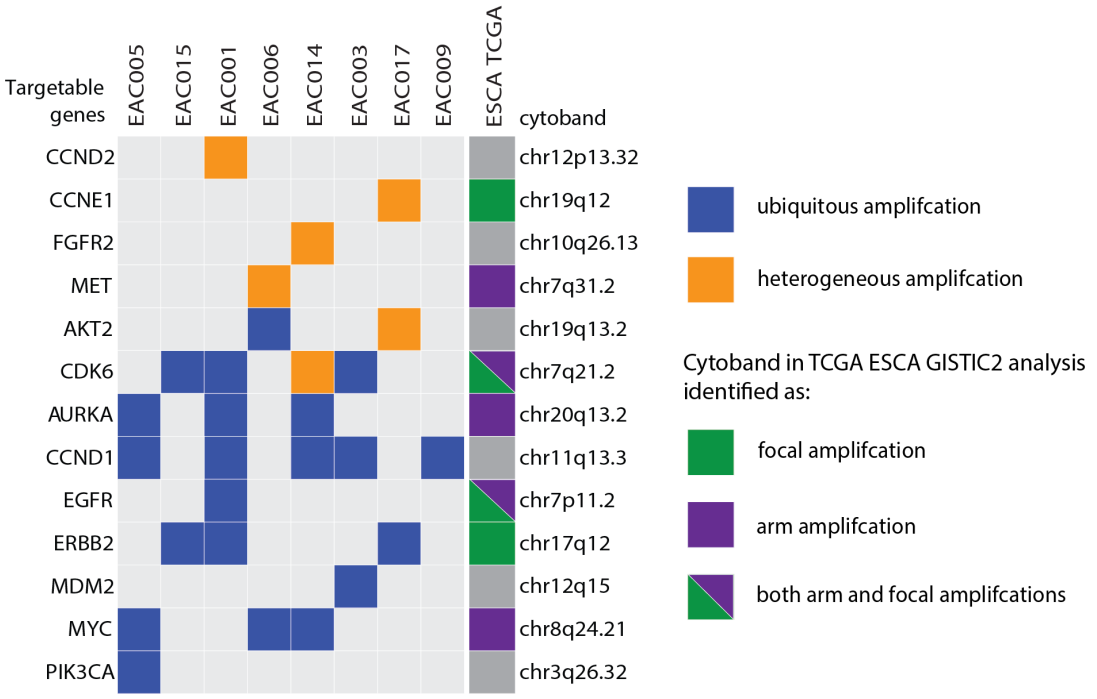

Figure 4-6: **Heterogeneity of SCNAs affecting clinically targetable genes**
Heat map showing amplifications identified within the cohort of tumours containing a targetable oncogene as identified by the TARGET dataset and whether these occur as ubiquitous (blue) or heterogeneous (orange) amplifications and if they occur as recurrent focal or arm level amplifications based on TCGA ESCA data (TCGA, 2015)

4.2.1.3 Summary and opportunities for methodological improvements

While this study did not identify instances of SCNA based parallel evolution, in the phylogenies based on sequence level variants (produced by Dr Gareth Wilson) it can be seen that multiple independent SNVs in *NOTCH1* occurred in EAC005. This parallel evolution of SNVs was in keeping with previous results from our group that had identified parallel SNVs in KIRC (Gerlinger et al., 2012, Gerlinger et al., 2014). The discovery of SNV parallel evolution and absence of SCNA based parallel evolution is perhaps not surprising given the lack of systematic attempts to identify SCNAs in which the most prevalent homologous chromosome differs between regions to demonstrate MSAI coupled with the fact that in WES analysis actual sequence level breakpoints cannot be reliably identified and act as unique identifiers for SCNAs. Indeed, this entire cohort is reanalysed in Chapters 5 and 6 and is shown to demonstrate large amounts of MSAI, with patient EAC001 showing evidence of

MSAI in 14 of its 22 autosomes. However, in this publication an ad-hoc approach was taken to examine the SCNA amplifications ubiquitous in all regions of patients' disease that had also been identified as recurrent amplifications by the TCGA (TCGA, 2015). Using this method none were found to demonstrate MSAI that might have suggested these amplifications might have occurred later in tumour evolution. Therefore, the paper's main CIN related conclusions of early LOH affecting TP53, early WGD and early SCNA amplifications in ESCA tumour evolution (Figure 4-7) are largely unaffected.

One other issue this study highlighted was difficulty in setting SCNA thresholds for calling "gain", "loss" and "amplification" using integer copy number estimates and continuous estimates of ploidy from Sequenza (Favero et al., 2015). Region 1 from EAC003 can be seen to demonstrate SCNA losses over much of its genome (see Figures 4-1 and 4-4), this is a function of the non-integer ploidy estimate produced by Sequenza coupled with the integer total copy number estimates of the region's segments. Together, these result in a threshold that considers much of the genome at a loss relative to ploidy. At the time of writing Sequenza v2.1.2 cannot produce continuous estimates of each SCNA segment's copy number that might better reflect the raw data, especially when calculating thresholds using continuous ploidy estimates. It should be noted that even with the exclusion of SCNA segments from EAC003 R1 the ordering of SCNA amplifications and gains being significantly more likely to be present in all regions of patient's disease remained significant.

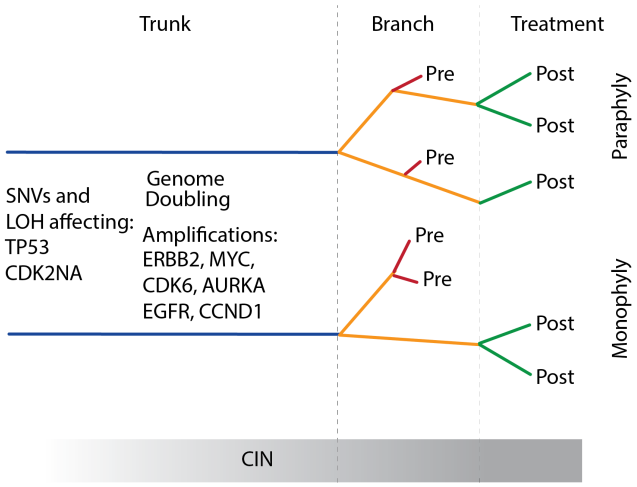

Figure 4-7: A model of tumour progression in ESCA
Inferring the evolutionary trajectories of EAC tumours through NAC by multi-region WES, identifying early and late mutational processes.

4.3 Tracking the Evolution of Non–Small-Cell Lung Cancer

4.3.1.1 Data and SCNA ITH quantification methods applied

This study analysed NSCLC tumour samples collected from 100 patients recruited by the TRACERx multi-centre prospective cohort study. These included 61 cases of LUAD, 32 cases of LUSC and 7 cases of other NSCLC subtypes. In total 327 tumour regions and 100 matched germline (blood) samples were subjected to high depth (~400x) WES.

In order to accurately call amplifications, gains and losses relative to ploidy we aimed to avoid introducing potential errors from the rounding of raw copy number values to produce integer copy number estimates. Therefore, instead of using Sequenza (Favero et al., 2015) as in Murugaesu et al. I, with supervision of Dr Nicolai Birkbak, chose to employ a modified version of ASCAT (Van Loo et al., 2010). This version of ASCAT provides continuous "raw" non-integer values of major and minor copy number derived directly from the measures of LogR, BAF, purity and ploidy that ASCAT uses to generate its integer copy number estimates (discussed in more detail in Chapter 3). As a single bulk region may contain multiple subclonal populations with different integer total copy number states, continuous values of copy number may more accurately reflect the mixed cancer cell population present. Therefore, the rounding involved in selecting a single integer total copy number state to represent an entire region may mask copy number gains relative to the average ploidy of cancer cells in the region. Total raw copy number was used in comparison to the copy number thresholds outlined in Chapter 3 in place of integer values as was performed in Murugaesu et al. Consensus segmentation was performed in the manner as Murugaesu et al. with each region's "child" segments caused by the introduction of all inferred breakpoints inheriting the values associated with their region specific "parent" segment.

These continuous values of raw allele specific copy number also allow potentially more accurate CCF calculation from SNV variant allele frequencies. Again, by removing the potential for rounding errors in estimating integer copy number, subclonal populations may be more accurately represented. However, it should be noted that the subclonal deconvolution of the TRACERx 100 regions using these raw copy number values was performed by Dr Nicholas McGranahan and Dr Nicolai

Birkbak. This deconvolution and subsequent phylogeny construction is used to generate the trees shown in Figure 4-13.

In this study I wrote applied the first instance of multi-region phasing was applied in this paper in order to detect MSAI with the approach outlined in Chapter 3 as a post SCNA calling step. Multi-region phasing was not used to improve copy number estimates as outlined in Chapter 3.

4.3.2 Multi-region phasing reveals pervasive CIN and MSAI in NSCLC

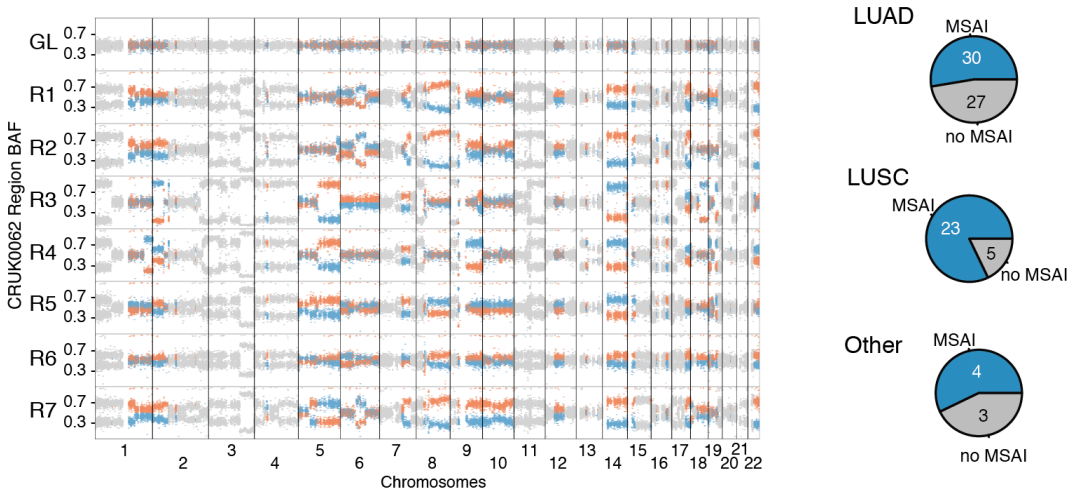

Figure 4-8: Prevalence of MSAI within a single tumour and incidence in cohort
Panel A) shows the BAF profile across the genome of a tumour sample obtained from Patient CRUK0062. Areas of BAF in regions (including tumour regions R1 through R7 and a germline [GL] reference region) that have mirrored subclonal allelic imbalance are highlighted in blue or orange. Events that showed mirrored subclonal allelic imbalance were identified in more than 40% of the genome.

Using multi-region phasing that selected optimum reference regions revealed that large proportions of the genome in some patients demonstrated MSAI. CRUK0062 in (Figure 4-8) was found to have 43% of its genome exhibiting MSAI in at least one region. None of this additional SCNA heterogeneity could be described using an SCNA caller designed for single bulk region analysis on region independently and in isolation. This large proportion of MSAI events indicate an enormous degree of ongoing chromosomal instability in relatively early stage NSCLC and argues against early punctuated macro-evolutionary change followed by copy number stasis (Gao et al., 2016). These MSAI events were common in the cohort and across histology subtypes, MSAI was observed in 62% of the 92 tumours in which there was sufficient quality of copy number calls in multiple regions. Of these 57 patients, 30 were LUAD, 23 LUSC and 4 were other NSCLC subtypes.

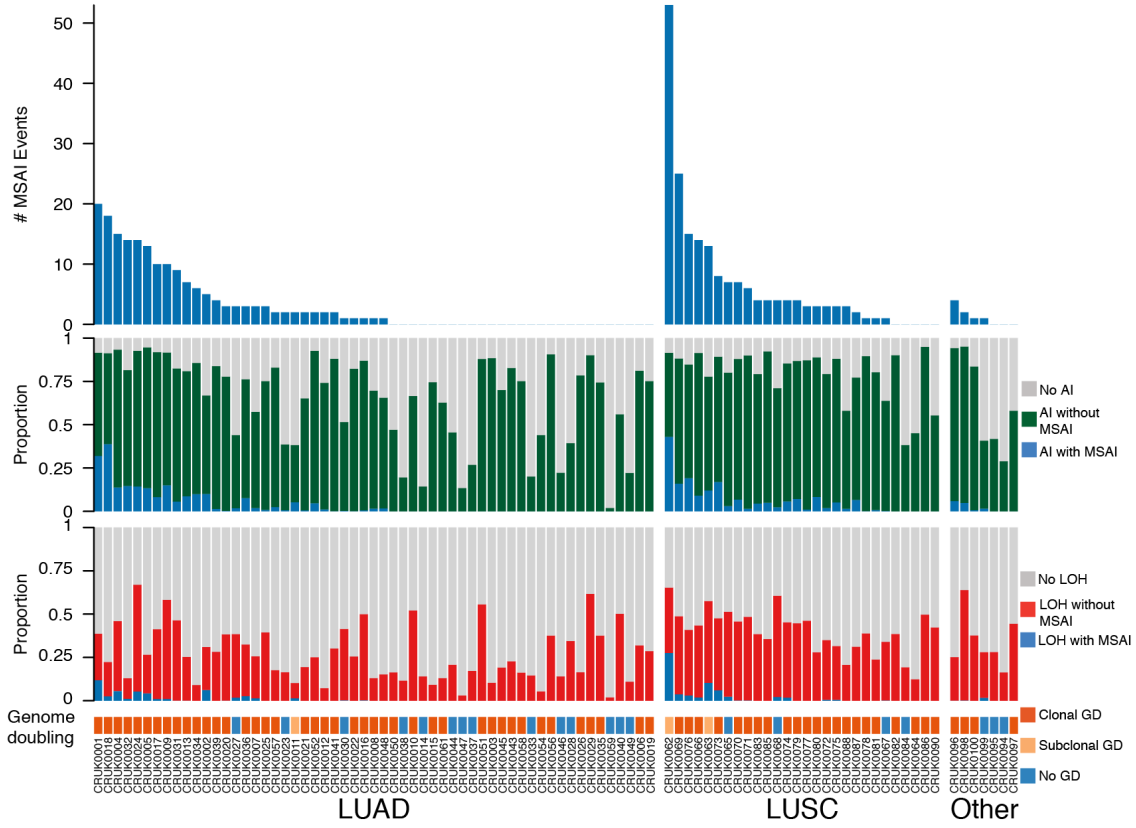

Figure 4-9: Prevalence of AI and MSAI in the TRACERx lung cohort
Allelic imbalance and mirrored subclonal allelic imbalance in each patient examined. Plot illustrating the extent of AI across the TRACERx cohort. The top panel shows the number of mirrored subclonal AI events detected in each tumour. The panels below highlight the proportion of each tumour displaying AI and loss of heterozygosity. The proportion displaying mirrored subclonal AI is shown (blue). AI: Allelic imbalance.

In the entire cohort we observed 375 MSAI events that varied from whole chromosome to focal events and affected from less than 1% to 43% of affected tumour genomes (Figure 4-9). Previous work from our group and others has demonstrated that WGD events may predict prognosis are associated with CIN (Fujiwara et al., 2005, Carter et al., 2012, Dewhurst et al., 2014). In this cohort 76% of tumours were found to demonstrate genome doubling and were clonal apart from three cases in which the WGD was found to be subclonal CRUK011, CRUK0062 and CRUK0063 and MSAI events (P=0.004, by Fisher's exact test).

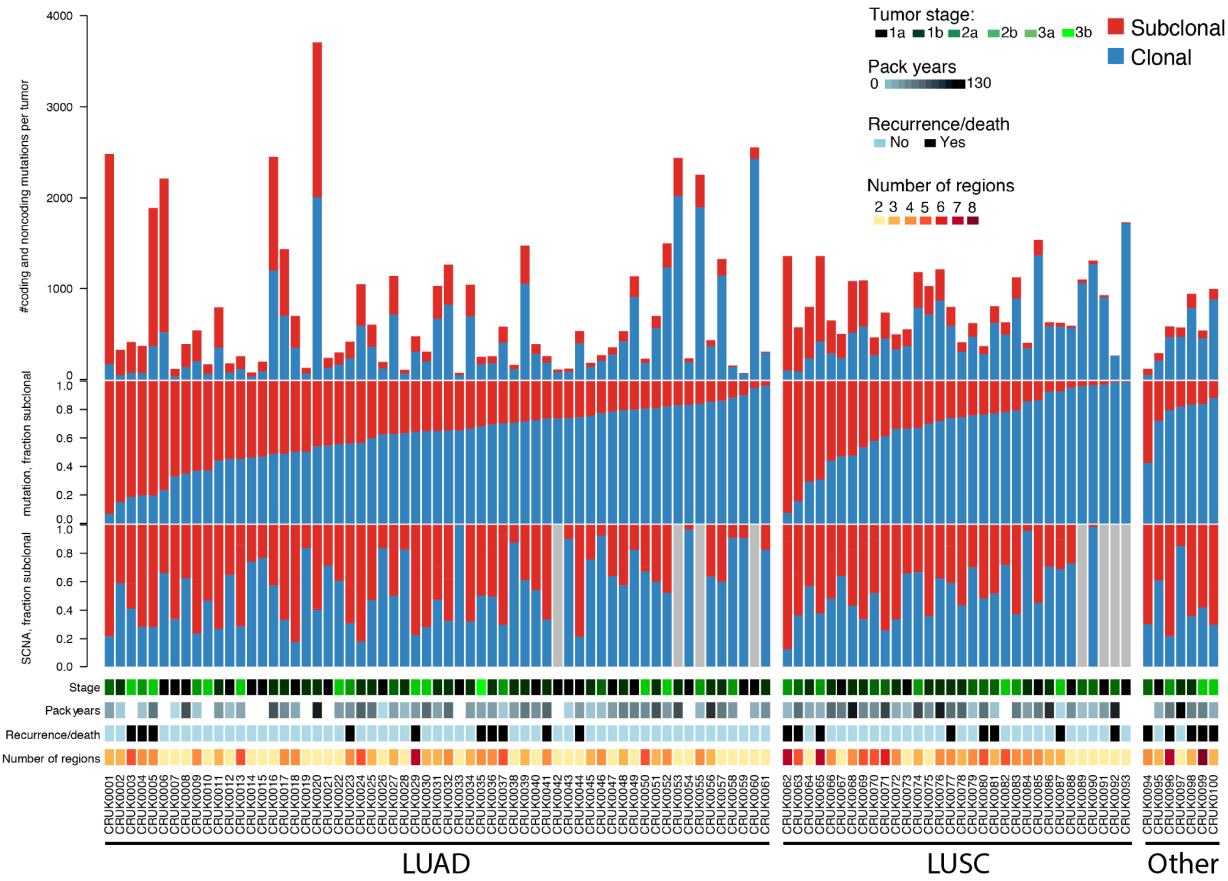

Figure 4-10: SCNA and mutation heterogeneity in each patient
Kaplan-Meier curve showing the recurrence-free survival versus the overall burden of SCNAs as defined by the mean proportion of the genome deviating from the tumour ploidy status defined as wGII across tumour regions from the same patient, divided in to high and low based on the media.

Sequence-level mutation calling and ITH analysis was, in the main, performed by Dr Gareth Wilson, Dr Nicolai Birkbak and Dr Nicholas McGranahan while I focused on the characterization of SCNA heterogeneity. To determine genome-wide estimates of SCNA ITH all segments were split into areas of gain or loss. As in Murugaesu et al. a consensus segmentation was generated by combining all inferred breakpoints from each region's copy number profile. The continuous raw total copy number estimate for each consensus segment in each region was then divided by that region's ploidy estimate and log2 transformed. SCNA gain was defined as log2(2.5 / 2) and SCNA loss as log2(1.5 / 2). Any segment defined as gain or loss that overlapped across all regions was defined as clonal and all other segments of copy number aberrations as subclonal. Within each tumour, I then summarized the percent of the genome subjected to SCNA in any region (total SCNA), the percent of the genome subjected to clonal SCNA (SCNA shared by all tumour regions), and the percent of the genome subjected to subclonal SCNA (SCNA found in some but not

all tumour regions). The proportion of subclonal SCNAs was then defined as the percent of the genome subjected to subclonal SCNA divided by the percent of the genome subjected to SCNA in any region (total SCNA).

Extensive ITH was observed across the cohort with a median of 30% (range, 0.5 to 93) of sequence level variants being classified as subclonal. SCNA ITH was also observed with a median of 48% (range 0.3 to 88) of SCNAs classified as subclonal. These findings of mutational and SCNA heterogeneity suggest that mutational and chromosomal genomic instability processes are continuing in later tumour evolution (Figure 4-10). The difference in levels of ITH between tumours was considerable, with the number of subclonal sequence level variants in a single case being between 2 and 2310. Similarly, the percentage of the genome affected by SCNA while the proportion of the genome affected by subclonal SCNAs ranging from 0.06 to 81%. WGD events were found to be significantly associated with subclonal SCNAs

If only a single region were sequenced rather than multiple regions from the same patient's disease, 76% of subclonal sequence level mutations could have been classified to be clonal across the cohort. However, given the lack of available SCNA callers able to detect subclonal SCNAs within tumour regions not assayed with WGS, 100% of SCNAs called from a single region would have had to be considered as clonal or of indeterminate clonality.

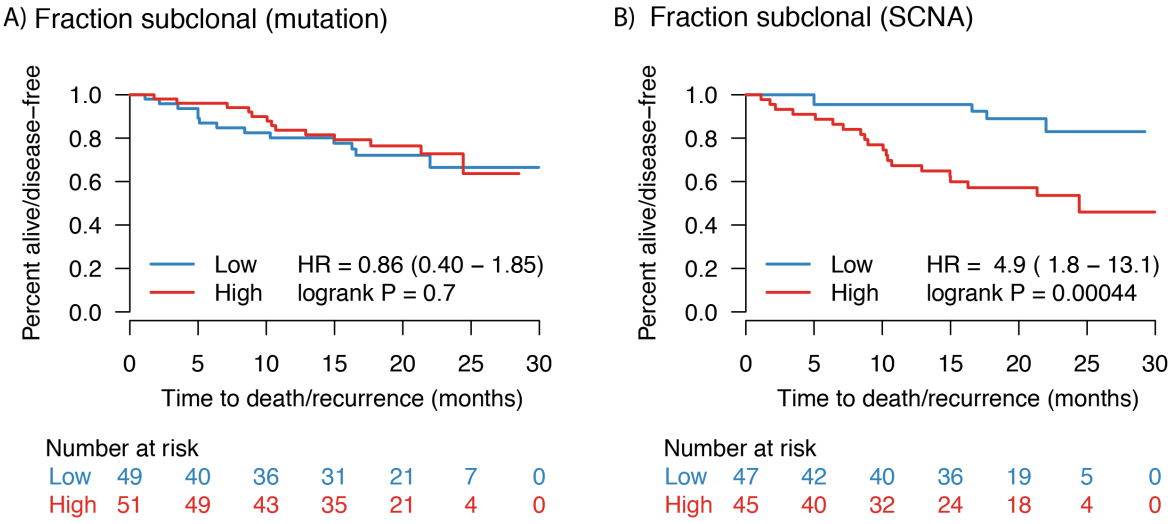

Figure 4-11: Survival analyses investigating sequence-level and SCNA ITH
There was no significant association between the proportion of subclonal sequence level mutations and relapse-free survival (P=0.70), but patients who had tumours with a high proportion of subclonal SCNAs were at a significantly higher risk for recurrence or death than those with a low proportion ($P=4.4x10^{-4}$).

As TRACERx is clinical trial (ClinicalTrials.gov number NCT01888601) survival analyses were performed by the trial statistician Prof Alan Hackshaw. However, I

contributed to the results used in these analyses. No significant association between the proportion of subclonal sequence level variants (median in cohort, 30%) and relapse-free survival was identified. However, though this analysis is preliminary given aim to recruit 842 patients in total to the TRACERx study, a significant association was observed between patients who had tumours with a high proportion of subclonal SCNAs (≥48%, the median of the cohort) were at a higher risk of recurrence or death than those patients whose disease demonstrated a low proportion (hazard ratio, 4.9; 95% confidence interval, 1.8 to 13.1, ($P=4.4x10^{-4}$, Fig 4-11). Interestingly, a measure of total SCNA burden that does not capture heterogeneity between tumour regions was also examined for association with relapse-free survival with no significant relationship found (Figure 4-12).

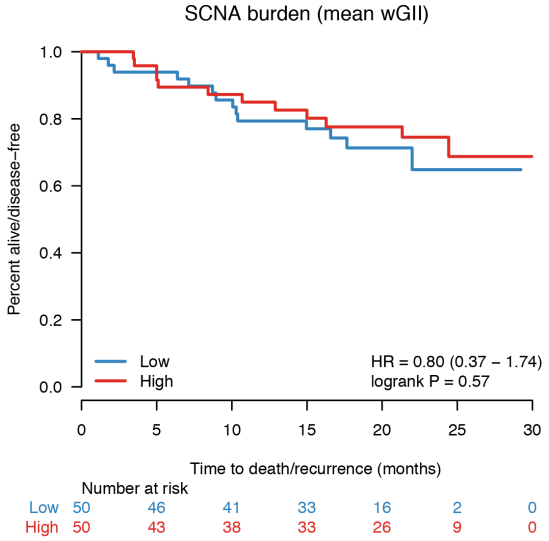

Figure 4-12: Survival analysis based on overall SCNA burden.
Kaplan-Meier curve showing the recurrence-free survival versus the overall burden of SCNAs as defined by the mean proportion of the genome deviating from the tumour ploidy status defined as wGII across tumour regions from the same patient, divided in to high and low based on the media.

4.3.3 Multi-region phasing reveals parallel evolution in NSCLC

Parallel evolution, where somatic events in distinct subclones of a patient's disease converge on the same gene, protein complex or pathway, are thought to reveal evidence of ongoing selection in tumours and potentially reveal evolutionary constraints which may reveal therapeutic targets. Our group's previous multi-region studies had revealed sequenced based mutational parallel evolution in both KIRC (Gerlinger et al., 2012, Gerlinger et al., 2014) and ESCA (Murugaesu et al., 2015). However, despite assaying 327 tumour regions from 100 patients no SNV based

parallel evolution was identified in the TRACERx100 cohort. In contrast, SCNA mediated parallel evolution was revealed through the detection of focal amplification events that demonstrate MSAI.

These focal SCNA amplifications (see Chapter 2 for definitions used) could be shown to originate from distinct homologous chromosomes in different tumour regions and therefore have independent origins. Five of these focal amplifications were described in five tumours and affected known cancer genes including *CDK4*, *MUC1*, *CHD8*, *NKX2-1* (Figure 4-13).

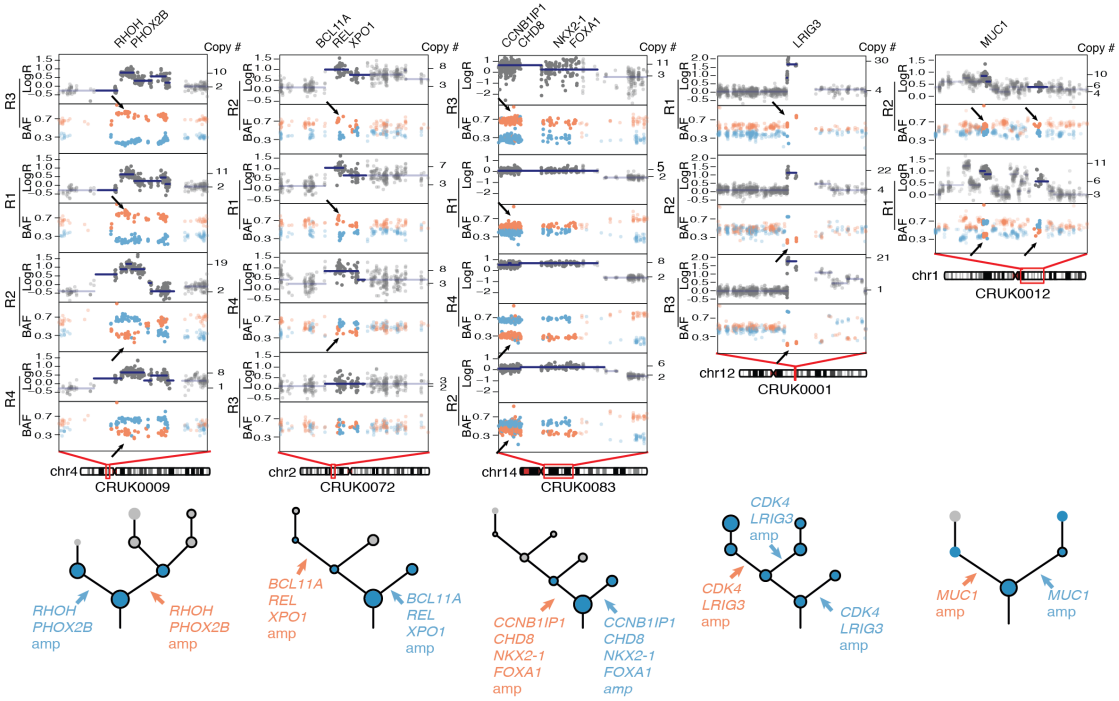

Figure 4-13: MSAI reveals parallel focal amplifications
Shows phylogenetic trees that indicate parallel evolution of driver amplifications detected through the observation of mirrored subclonal AI (arrows). Subclones that are coloured blue carry a cancer driver event, and those that are coloured grey carry no driver event; black outlining of the circles indicates that the subclone appears to be clonal in at least one tumour region.

These focal SCNA amplification events, as well as being present in different regions of the tumour, also mapped to different branches of the subclonal phylogenies constructed from CCF clusters identified by PyClone (Roth et al., 2014). This mapping was produced by assigning the parallel MSAI amplifications to the SNV CCF cluster(s) that had the greatest contribution from the regions in which the amplification events were identified. This further emphasizes that these events occurred in distinct subclones within a patient's disease.

These parallel MSAI events were not limited to amplification events but also affected arm level gain and loss events. Chromosomal arm gain and loss were called by

requiring at least one region to show at least 98% of the arm's extent as either gain or loss by the threshold relative to ploidy previously described in Chapter 2. Clonal arm gain or loss was then called if the same chromosomal arm showed at least 75% gain or loss across all remaining regions, otherwise a subclonal arm level event was called.

-

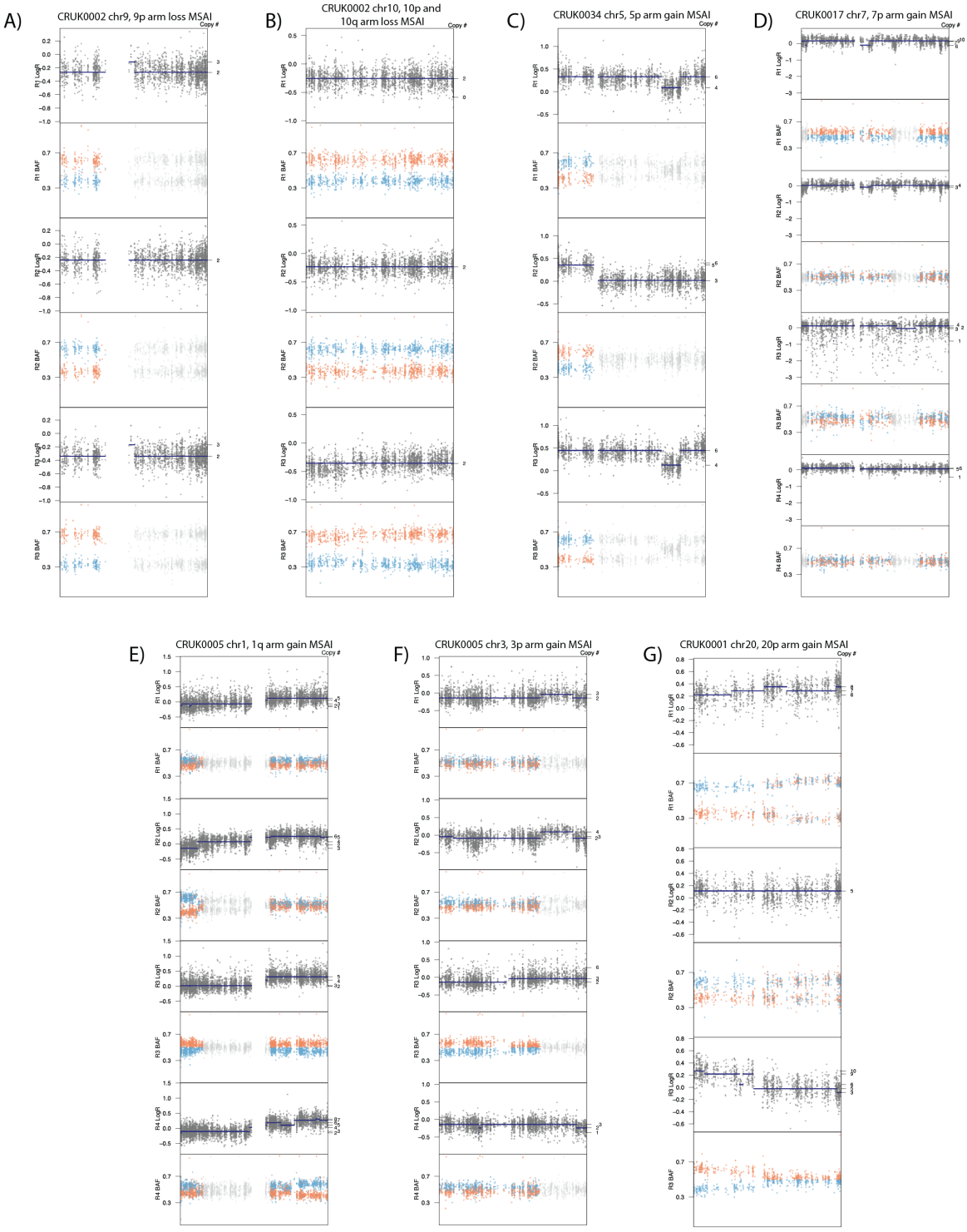

Figure 4-14: Arm-level MSAI events in LUAD
Plots A – G illustrate LogR and BAF of heterozygous SNPs across all tumour regions from patients in which whole chromosomes on which arm-level gains demonstrating MSAI were observed. Copy number is indicated on the right-hand side of each plot. Only heterozygous SNPs in areas involved in MSAI, including the arm event itself are coloured orange and blue.

Parallel MSAI gain or loss events were observed at the arm level in 13 tumours (6 LUSC, 5 LUAD, 2 other NSCLC histological subtypes) see Figure 4-14 and 4-15 for LUAD and LUSC events, other histological subtypes not shown. The majority of the MSAI parallel arm level gains (10 out of 11 events) and MSAI parallel arm losses (6 of 8 events) have previously been classified as significantly gained or lost in NSCLC (Campbell et al., 2016, TCGA, 2017).

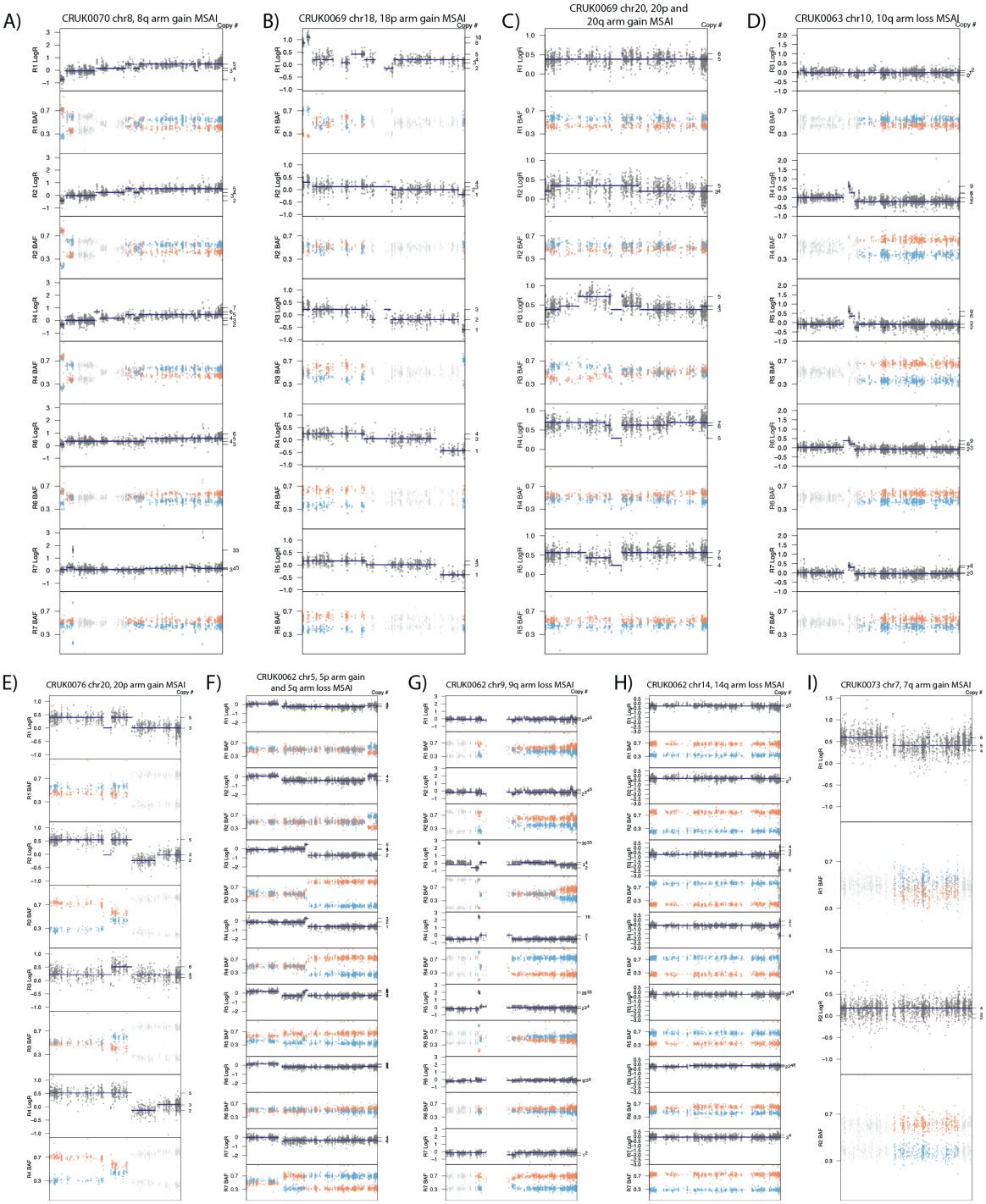

Figure 4-15: Arm-level MSAI events in LUSC
Plots A – I illustrate LogR and BAF of heterozygous SNPs across all tumour regions from patients in which whole chromosomes on which arm-level gains demonstrating MSAI were observed. Copy number is indicated on the right-hand side of each plot. Only heterozygous SNPs in areas involved in MSAI, including the arm event itself are coloured orange and blue. Copy number is indicated on the right-hand side of the plot.

4.3.3.1 Summary

ITH is thought to provide a substrate for tumour evolution and drug resistance (Greaves, 2015). In this study, my work revealed pervasive ongoing CIN and SCNA

ITH with the use of multi-region phasing that enabled the detection of allele-specific copy number heterogeneity through MSAI that is not possible to detect from single region WES alone.

These MSAI events could also be used to detect independent focal amplification and arm level gains and losses affecting the same sections of the genome. Again, these events would not be detectable without multi-region sequencing or the multi-region phasing. Together, this newly revealed SCNA heterogeneity and parallel evolution suggests an important role for CIN in later tumour evolution. This is especially true if one considers that independent SCNA events occurring in a single tumour's development may not always occur on different homologous chromosomes. It seems quite possible that an SCNA may arise from the paternal chromosome in one region of the tumour and, entirely separately in another region again from the paternal chromosome.

These independent SCNAs would be undetectable using multi-region phasing. Even WGS might not be able to distinguish independent events whose breakpoints end in telomeric or centromeric regions. Therefore, while the scale of the heterogeneity revealed by this study is novel, an unknown proportion of SCNAs considered clonal using the techniques available to us might in reality also be subclonal.

This study is preliminary with only the first 100 of a total of 842 patients investigated. However, the association between the proportion of SCNAs affecting a patient's disease being subclonal and relapse-free survival emphasizes the need to investigate biological effects of SCNA heterogeneity and create accurate bioinformatics tools to quantify it.

4.3.3.2 Opportunities for methodological improvements

When applied to this cohort, multi-region phasing and MSAI detection revealed many SCNA events to be of independent origin and be present in separate subclonal populations of the cancer. However, it also revealed that the SCNA caller used (ASCAT) failed to sometimes failed to detect AI in a region that was clearly present when the phasing obtained. This occurred when, in a region of AI, the two peaks of the BAF distribution resulting from heterozygous SNPs on the different homologous chromosomes were difficult to clearly distinguish. This could be due a low purity tumour region resulting in the AI signal being drowned out by proportion of NGS reads

coming from normal diploid cells. One other possibility is that there are SCNAs occurring in subclonal populations of cancer cells within the tumour region that cause only relatively small separations in the means of the BAF resulting from different homologous chromosomes.

These observations of AI observed with multi-region phasing but missed by ASCAT and leading to incorrect allele specific copy number calls lead me to implement SCNA calling using multi-region phasing in Chapter 3.

4.4 Deterministic Evolutionary Trajectories Influence Primary Tumour Growth: TRACERx Renal

4.4.1.1 Introduction

Previous work from our group has reported sequence based mutational and SCNA ITH in multi-region studies of KIRC (Gerlinger et al., 2014, Martinez et al., 2013). These studies demonstrated that a single region analysed alone may miss important somatic events or mistakenly consider them as clonal due to their "illusion of clonality". This lack of resolution may hold back our efforts to understand tumour evolution. Until the publication of the three TRACERx renal papers in which the following work features (Turajlic et al., 2018b, Turajlic et al., 2018a, Mitchell et al., 2018), large-scale attempts (with greater than ten patient's disease examined) to classify KIRC have depended on single region analyses focusing on sequence level variants (Hakimi et al., 2013, Kapur et al., 2013, Sato et al., 2013) or gene expression and methylation (TCGA_Consortium, 2013, Chen et al., 2016).

These efforts to conduct a molecular characterization of KIRC from single regions have revealed frequent inactivation of the *VHL* tumour suppressor gene, alterations in histone-modifying genes (Dalgliesh et al., 2010), SWI/SNF complex (Varela et al., 2011) and the PI3K/AKT/mTOR pathway (Scelo et al., 2014). Significant arm level or focal losses have been observed on chromosomes 1p, 3p, 4q, 6q, 8p, 9p and 14q while significant focal and arm level gains have been described on 1q, 2q, 5q, 7q, 8q, 12p and 20q (Beroukhim et al., 2009, TCGA_Consortium, 2013).

TRACERx renal is a prospective longitudinal cohort study (study number: NCT03226886) with a protocol-specified endpoint of examining the association of ITH with disease stage and clinical outcomes through multi-region genomic profiling of primary tumours. The study aims to recruit 320 patients in total but the results

shown in this chapter come from the interim findings investigating patterns of ITH, clonal evolution in the first 101 patients with a diagnosis of non-familial KIRC. All patients in this study had multi-region targeted panel sequencing.

Dr Hang Xu, Dr Kevin Litchfield and Mr Stuart Horswell performed SCNA and sequence-level mutation characterization of the NGS targeted panel data. Amongst many findings they demonstrated extensive parallel evolution of sequence-based mutations with 13% of untreated primary tumours demonstrating at least one instance. Furthermore *SETD2*, *BAP1*, and *PTEN* showed significant enrichment for a parallel evolution with some tumours exhibiting large numbers of parallel sequence-based mutations, for example one tumour exhibited 10 distinct *SETD2* mutations in different regions (Turajlic et al., 2018b).

However, the targeted panel sequencing was found not to cover sufficient numbers of heterozygous SNPs in all patients to allow allele specific copy number calling and only total copy number estimation was performed through the use of copy number caller that exploits off-target reads CNVkit (Talevich et al., 2016). Therefore, multi-region phasing and MSAI detection, which require allele-specific copy number estimates, could not be performed on these data. In addition, as the targeted NGS panel used only assayed around 110 genes (depending on the version used) it delivered unequal coverage of different chromosome arms making any potential ability to detect MSAI non-uniform over the genome. Therefore, I focused on investigating the subset of 41 cases that, as well as being investigated with targeted panel sequencing, were subjected to either multi-region WES or WGS.

4.4.2 Multi-region phasing reveals parallel evolution in KIRC

Somatic copy number calling was performed on each region separately by Dr Thomas Mitchell at the Sanger Institute using ASCAT (Van Loo et al., 2010) for WES samples and Battenberg for WGS (Nik-Zainal et al., 2012). Of the 41 tumours examined 15 were found to demonstrate at least on instance of MSAI and in total 23 MSAI events were observed (Fig 4-16).

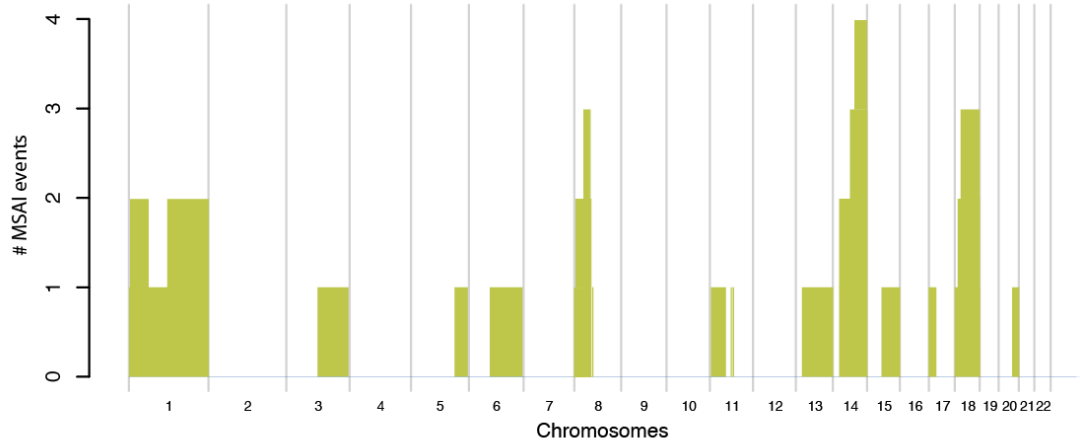

Figure 4-16: MSAI prevalence in the TRACERx renal WES and WGS samples
Across-genome plot showing the relative size and genomic position of all MSAI events identified in the 41 KIRC tumours.

These MSAI events were then further subdivided into gains and losses relative to ploidy. In this study WES copy number calling was performed using a version of ASCAT that could not output continuous estimates of copy number, therefore integer copy number estimates were used for both the WES data and WGS analysed with Battenberg. The total copy number of each segment was then compared to the rounded mean ploidy estimate (derived from either Battenberg and ASCAT) of its respective tumour region and was considered a gain if higher, a loss if lower and neutral if there was no difference. These classifications were then combined with MSAI detection to identify parallel evolution of losses or gains within the same patient's disease coming from different homologous chromosomes.

In total 8 of the 15 tumours in which MSAI was described also demonstrated parallel evolution of SCNA gains and losses with a total of 14 events. However only two resulted from parallel independent gains from different homologous chromosomes relative to ploidy: an arm-level gain of 1q and a focal gain event on chromosome 11. In contrast, parallel loss MSAI events were much more common with 12 separate instances of losses relative to ploidy or LOH coming identified across the cohort. The majority of these were large-scale events with only two focal losses observed on chromosome 8 (Figure 4-17). Of the 10 large-scale parallel losses 9 (all apart from 1p) occurred in chromosome arms that had previously been classified as significantly lost by TCGA analysis of a large single region cohort.

The 9 remaining MSAI events not classified as instances of parallel evolution resulted from areas of AI where (in at least one region affected by MSAI) the total copy number

was equal to the rounded mean ploidy of that tumour sample, e.g. an area of AI with 2 copies of the major allele and 1 of the minor allele in a tumour region where the mean rounded ploidy is 3. Therefore at least one of the SCNA events underlying the MSAI in these cases are classed as “neutral” rather than a gain or loss.

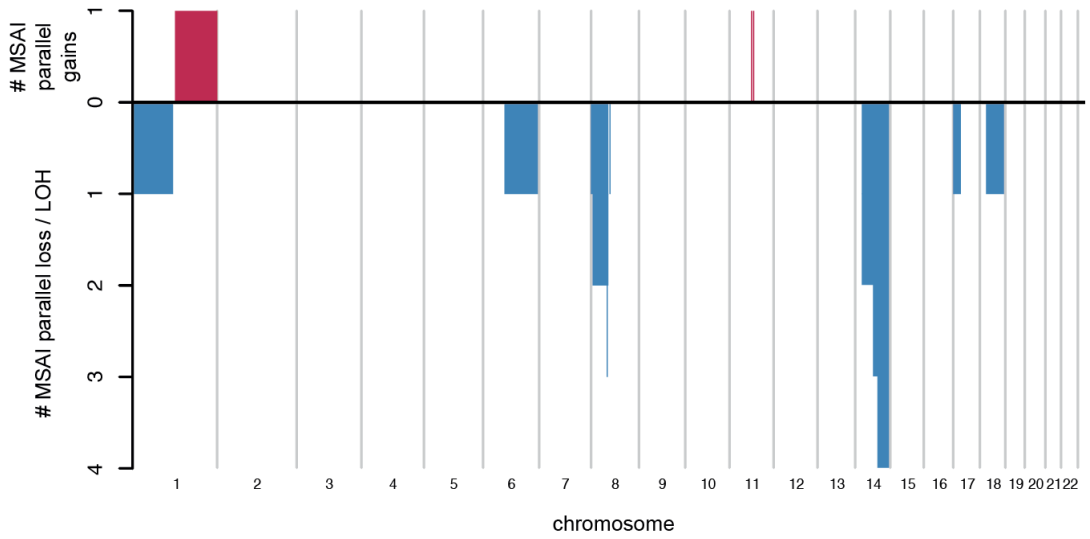

Figure 4-17: Across-genome plot showing parallel gain and parallel loss events
Count, size and genomic position of parallel gain and loss events revealed by copy number analysis and MSAI detection in the TRACERx renal cohort.

The most common MSAI parallel loss events were large-scale and observed on 14q in four cases (Figure 4-18) and 8p in two cases. Previous analysis of a large cohort of single region KIRC samples performed by the TCGA using GISTIC2 have identified these two arms as being both significant loss events and the second and third most common arm loss events occurring in KIRC (TCGA, 2017). All instances of parallel MSAI 14q loss encompassed the KIRC tumour suppressor *HIF1A* locus (Shen et al., 2011). This previous evidence of significant recurrence of these losses combined with the novel description of their parallel evolution within suggests they may have important role the development of KIRC.

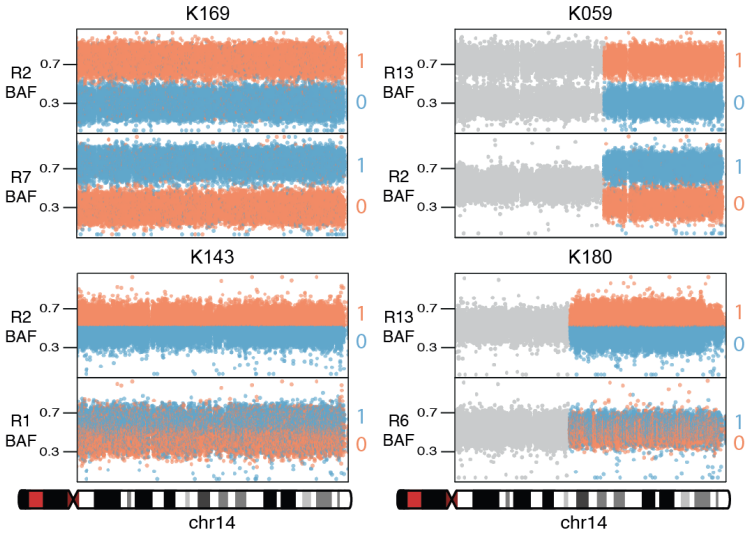

Figure 4-18: Chromosome 14 parallel loss and LOH in WGS samples
BAF profiles of two regions from each of the four cases demonstrating parallel losses of chromosome 14 revealed by MSAI detection. Allele specific copy number estimates are on the right-hand side of the plot and match the colour of the BAF from homologous chromosome they refer to.

4.4.2.1 Summary

Unfortunately, the choice of targeted NGS using a gene panel of limited size (~110 genes depending on the version used (Turajlic et al., 2018b)) that assayed different chromosomal arms to greatly different degrees meant that MSAI detection could not be reliably applied to the majority of the 1,206 regions from the 101 patients assayed in this study. However, in the subset of 41 tumours that were also interrogated using multi-region WGS or WES 15 demonstrated MSAI and 8 of these were found to demonstrate parallel evolution of chromosome arm-level gains and losses.

It is interesting to note that the targeted panel NGS that allowed description of sequence-level variants (but only total copy number not allele-specific or MSAI) found parallel evolution of sequence level mutations in 13% of cases whereas the WES and WGS demonstrated parallel MSAI gains or losses in 20% (8 of 41 cases). This is despite a greater average number of regions per tumour being analysed with targeted NGS than the WES and WGS used to identify MSAI. Targeted NGS was applied to 1206 tumour regions from 101 patients (median 7 [range, 3 – 75]) whereas WES was performed on 107 regions from 17 tumours and WGS on 81 regions from 27 patients (note there was an overlap of 7 patients that was assayed both WES and WGS).

Furthermore, as many of the MSAI events were the result of independent parallel losses occurring in chromosome arms previously identified as significantly recurrent

losses in TCGA analyses of KIRC, this suggests that these parallel events may be selected due to providing some selective advantage to the subclones developing them.

4.4.2.2 Opportunities for methodological improvements

My work in this study integrated copy number calls made using two different bioinformatics tools (ASCAT and Battenberg) applied to two different NGS modalities: WES and WGS. Battenberg produced continuous estimates of copy number for all tumour regions examined while the version of ASCAT applied by Dr Thomas Mitchell did not. This made finding a set of thresholds to define copy number gains and losses to apply to both the WES and WGS data difficult and I took the decision to rely on the integer copy number calls made by both tools. These integer calls will involve rounding of continuous estimates of raw copy number that may introduce rounding errors that mask events in the underlying raw data.

The necessity of using these integer calls to allow comparisons highlighted the need for integrated copy number calling tools that are able to analyse different data types as well as quantify SCNA heterogeneity in a standardized manner. Indeed as, discussed in Chapter 3, few SCNA calling tools are designed for multi-region data and still fewer bioinformatics tools designed to quantify SCNA heterogeneity across multiple regions.

Multi-region phasing was applied to both the WES and WGS in order to detect MSAI. However, the initial analysis of WGS using Battenberg to produce copy number estimates applied statistical phasing using 1000 Genomes reference haplotypes. This statistical phasing is intended to allow detection of subclonal events by distinguishing the two BAF distributions resulting from the different homologous chromosomes in areas of AI. However, when comparing the copy number calls and phased BAF data between regions from the same patient, it became clear that the phasing of the same heterozygous SNPs was not consistent across regions and it was necessary to use unphased BAF with the multi-region phasing technique in order to detect MSAI. This finding further emphasized the need for integrated analysis of multiple regions from the same patient when available.

4.5 Conclusions

Multi-region bulk NGS data analysis allows a far greater resolution of SCNAs than single regions alone and in this chapter I have demonstrated that SCNAs may be present in one region and absent in another region from the same patient's disease. These SCNAs might simply be considered "clonal" from analyses of single region data and could be considered to demonstrate "illusion of clonality". I have also shown that AI in multi-region data may be used to create a multi-region phasing that is consistent across regions from the same patient's disease. This phasing allows the detection of MSAI, a type of subclonal SCNA that cannot be detected using single region data.

The detection of MSAI revealed hitherto underappreciated SCNA heterogeneity in NSCLC and KIRC and suggests that a greater degree of CIN is active later in tumour evolution than has been previously appreciated. My work quantifying this heterogeneity, that could only have been assessed using multi-region data, as part of the TRACERx lung prospective cohort study is among the first to investigate SCNA heterogeneity in relation to outcome in cancer. The significant association between the proportion of SCNAs present within a patient's disease that are subclonal and relapse free-survival suggests that NSCLC patients with early-stage tumours with high SCNA heterogeneity may be a high-risk group who could benefit from early therapeutic intervention during follow up. Interestingly, the proportion of sequence-level variants found to be subclonal were not found to be significantly associated with relapse-free progression, potentially suggesting a functional impact of SCNA heterogeneity distinct to sequence-level heterogeneity, though this result is preliminary and could also reflect measures that simply sample general ITH with different levels of resolution. However, the lack of significant association between a measure of total SCNA burden that does not capture their subclonality and relapse-free survival emphasizes the importance of further efforts to examine SCNA heterogeneity using methods that are able to assess it rather than simply continuing with single region bulk analyses.

Sequence-based parallel evolution has previously been described in many cancer types while SCNA based parallel evolution has been described in fewer. However, this may be a function of the sensitivity of the assays available for describing SCNA events (see Chapter 1). Previously, only breakpoint analyses through multi-region WGS or FISH in very specific experimental conditions have been powered to detect such events (see Chapter 1). Multi-region phasing offers the opportunity to detect at

least a subset (those intendent events originating from the same homologous chromosome will still be undetectable) of parallel events without the use of high complex experimental approaches or expensive and time consuming to analyse WGS. The surprisingly high rate of MSAI and the parallel SCNA events it reveals observed in multi-region analyses of both KIRC and LUAD suggest that both the degree, ongoing nature and influence on cancer cell fitness of CIN might be underappreciated across cancer types.

Chapter 5 Pan-cancer SCNA heterogeneity reveals pervasive chromosomal instability in tumour evolution

5.1 Introduction

Understanding the genomic alterations underpinning cancer development and evolution is a fundamental aim of biomedical research. Large-scale sequencing efforts have revealed many of the high-frequency driver alterations associated with cancers (Lawrence et al., 2013, Bailey et al., 2018), and also elucidated some of the key mutational processes shaping the cancer genome (Alexandrov et al., 2013). Recent work has also highlighted the prevalence and importance of ITH across human cancers, contributing to treatment failure and resistance (Schwarz et al., 2015, Greaves, 2015).

A major contributor to cancer genome evolution and ITN is CIN, involving an elevated rate of acquisition of SCNAs such as WGD, segmental gains or losses, high-level amplifications and LOH events (Zack et al., 2013). Bulk genomics analysis of single tumour regions has defined the state of the cancer genome at the time of sampling (Carter et al., 2012, Zack et al., 2013) and revealed significant deviations from a normal diploid karyotype across tumour types. However, single region sequencing cannot capture ongoing CIN, a dynamic measure defined as the ongoing rate of change of chromosome number or structure. In order to gain a detailed understanding of the dynamics of CIN either spatial (multiple tumour regions or single cell analysis) or temporal (multiple tumour regions over time) sequencing strategies are required (Jamal-Hanjani et al., 2015).

Figures 5-13, 5-17 through 5-31 were produced by Dr Emilia Lim, though I performed the analysis of the data plotted. Isolation and passaging of the HCT116 cell line clones (displayed in Figure 5-8) was performed by Dr Andrew Rowan, Dr Sally Dewhurst and initial SCNA analysis (not including MSAI detection) of the SNP arrays used to analyse the SCNA landscape of the HCT116 cell lines was performed by Dr Nicholas McGranahan.

5.2 Detailed characterization of SCNA heterogeneity across cancer types

I derived allele-specific copy number profiles in 1421 multi-region cancer samples from 394 patients from 25 cancer subtypes (on average, 4 cancer samples per patient [range 2-16], Figure 5-1) using standardised pipelines (see Chapter 2) and the mPHASE algorithm (see Chapter 3). Out of the 1421 samples, 1019 were from primary tumour samples, 363 were of metastatic origin, 32 were obtained post-treatment, and 7 were obtained at relapse. Tumour types examined with samples from 10 or more patients included:

Tumour type	Tumour type abbreviation	Number of tumours	Number of samples	Publications
Lung adenocarcinoma	LUAD	84	221	Brastianos et al.(Brastianos et al., 2015) Jamal-Hanjani et al.(Jamal-Hanjani et al., 2017)
Clear cell renal cell carcinoma	KIRC	54	245	Gerlinger et al.(Gerlinger et al.) Martinez et al.(Martinez et al., 2013) Gerlinger et al.(Gerlinger et al., 2014) Brastianos et al.(Brastianos et al., 2015) Mitchell et al.(Mitchell et al., 2018) Turajlic et al.(Turajlic et al., 2018b)
Lung squamous cell carcinoma	LUSC	31	107	Brastianos et al.(Brastianos et al., 2015) Jamal-Hanjani et al.(Jamal-Hanjani et al., 2017)
Melanoma	SKCM	30	115	Sanborn et al.(Sanborn et al., 2015) Brastianos et al. (Brastianos et al., 2015) Harbst et al.(Harbst et al., 2016) Reuben et al.(Reuben et al., 2017)
Endometrial carcinoma	UCEC	27	62	Gibson et al.(Gibson et al., 2016)

Bladder urothelial carcinoma	BLCA	26	66	Thomsen et al.(Thomsen et al., 2016) Lamy et al.(Lamy et al., 2016)
Oesophageal adenocarcinoma	ESCA	22	65	Murugaesu et al.(Murugaesu et al., 2015) Brastianos et al.(Brastianos et al., 2015) Findlay et al.(Findlay et al., 2016)
Oestrogen receptor-positive breast cancer	BRCA ER+	19	91	Yates et al.(Yates et al., 2015) Brastianos et al.(Brastianos et al., 2015) Savas et al.(Savas et al., 2016) Yates et al.(Yates et al., 2017) Messaoudene et al. Unpublished.
HER2+ breast cancer	BRCA HER2+	18	75	Brastianos et al.(Brastianos et al., 2015) Yates et al.(Yates et al., 2015) Yates et al.(Davies et al., 2017) Messaoudene et al. Unpublished.
Triple-negative breast cancer	BRCA TN	17	47	Brastianos et al.(Brastianos et al., 2015) Savas et al.(Savas et al., 2016) Yates et al.(Yates et al., 2015) Yates et al.(Yates et al., 2017)
Glioma	GBM/LGG	12	45	Suzuki et al.(Suzuki et al., 2015)
Prostate adenocarcinoma	PRAD	10	50	Gundem et al.(Gundem et al., 2015)
Colorectal adenocarcinoma	COAD	13	103	Sottoriva et al.(Sottoriva et al., 2015) Brastianos et al.(Brastianos et al., 2015)

Other tumour types with fewer than 10 patients each contributed a further 129 tumour regions from 31 tumours.

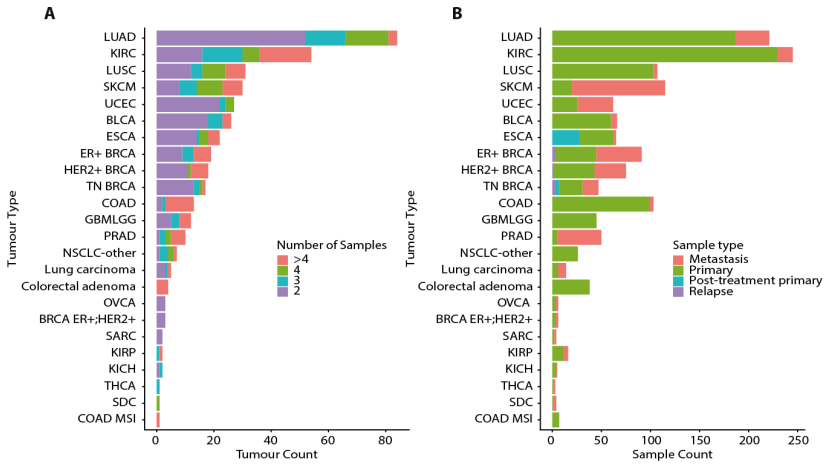

Figure 5-1: Pan-cancer cohort outline
A) Histogram showing the number of tumours present in the each tumour type and coloured according to the number of samples analysed from the tumour. B) Histogram showing the total number of tumour samples present in each tumour tyope and coloured by sample type.

To assess the degree of CIN, dynamics of SCNA acquisition, and extent of SCNA intra-tumour heterogeneity, I applied a novel multi-region phasing approach that I developed (Chapter 3) followed by tree inference and ancestral reconstruction using established methods (Beerenwinkel et al., 2014) updated to incorporate MSAI detection. The approach leverages the multi-region data for long-range phasing of heterozygous germline variants and their associated SCNAs. This phasing information allows for the sensitive detection of subclonal SCNAs both within and between regions frequently missed by single-region studies, and enables rescuing of samples with low tumour cell content (Chapter 3).

For each tumour region from each patient I derived two distinct readouts: (i) continuous total copy number relative to ploidy, which enables stratification of SCNA events into gains, amplifications, losses and deep-losses (Chapter 3), and (ii) haplotype-specific copy number, for detection of allelic imbalance (AI), loss of heterozygosity (LOH) and MSAI (Jamal-Hanjani et al., 2017). MSAI reflects at least two distinct subclones of a tumour harbouring independent SCNA events affecting overlapping genomic loci on distinct homologous chromosomes, that may indicate parallel evolution. Haplotype-specific SCNA profiles were used for tree inference and reconstruction of the most recent common ancestor (MRCA) using established methods (Beerenwinkel et al., 2014). Importantly, unlike structural rearrangements,

homozygous deletions and focal events, these readouts were available regardless of the sequencing approach adopted.

5.3 Distinct patterns of CIN characterise cancer types

I first set out to quantify the prevalence of subclonal SCNAs across the cohort as indicators of ongoing CIN after the emergence of the MRCA. In 99% of tumours (389/394), multi-region analysis revealed at least one subclonal SCNA, with a median of 16% of the genome found to be subject to subclonal SCNAs across cancer types. Considerable variation was observed between and within cancer types with regards to the total percentage of genome subject to SCNAs as well as subclonal and clonal SCNAs (Figure 5-2).

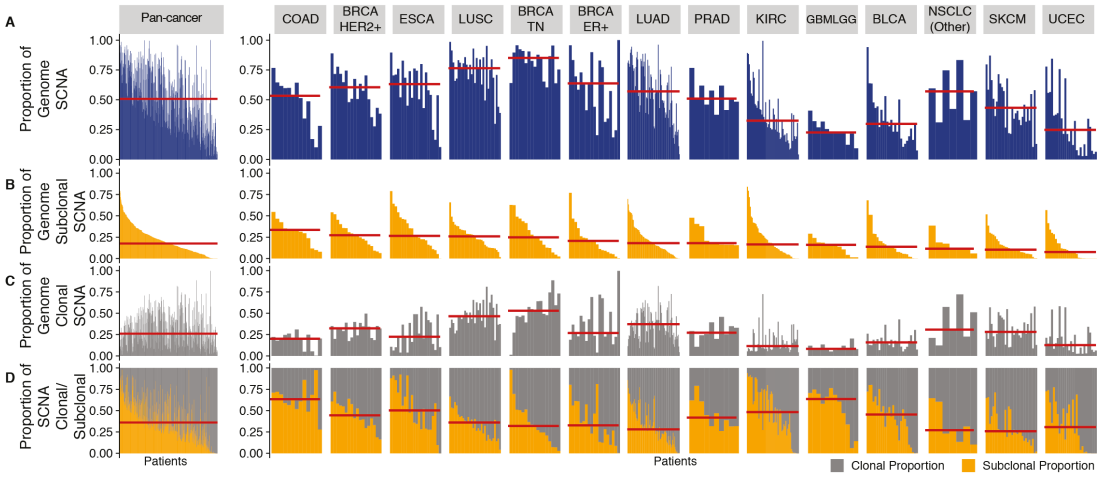

Figure 5-2: Overview of SCNA burden and heterogeneity in the pan-cancer cohort
A) For each patient, the percentage of the genome that is affected by SCNAs is indicated. Cancer types and patients are ordered based on the median percentage of subclonal SCNAs -- this order is maintained throughout the figure.
B - C) Barplots indicating the percentage of the genome affected by subclonal (B) and clonal (C) SCNAs. D) The proportion of SCNAs that are subclonal and clonal is displayed.

For instance, despite comparable total percentage of genome affected by SCNA between LUAD and BRCA HER2+ (54% vs. 57% genome. P=0.7, t-test), LUAD exhibited a significantly larger proportion of the genome subject to clonal SCNAs (72%, P=0.0055, t-test, Figure S4), while BRCA HER2+ was characterized by a higher frequency of subclonal SCNAs (53%) across the genome. Thus, while overall SCNA profiles from single samples would suggest similar CIN patterns for these two cancer types, multi-region analysis reveals distinct levels of both clonal and subclonal SCNAs.

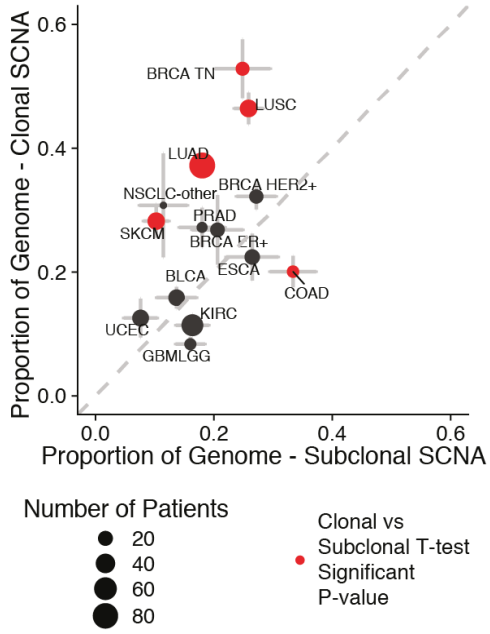

Figure 5-3: Comparing clonal and subclonal SCNA burden within cancer types
Scatterplot comparing the proportion of the genome affected by clonal and subclonal SCNAs. The median value for each cancer type is indicated. The size of the dots reflects the number of patients present in each cancer type. Red dots indicate cancer types with significant differences in the proportion of the genome affected by clonal subconal SCNAs ($p < 0.05$, paired t-test)

More generally, we observed that certain cancer types were characterized by a high proportion of the genome affected by clonal SCNAs, including LUAD (34%), LUSC (44%), and BRCA TN (52%) compared to a median of 22% across the entire cohort (Figure 5-2). Indeed, in these cancer types and SKCM, a larger proportion of SCNAs were found to be clonal compared to subclonal ($P<0.05$, paired t-test, Figure 5-3). By contrast, in KIRC and LGG/GBM, the proportion of the genome subject to clonal SCNAs was frequently low (median, <15%), with only one or two arm-level SCNA events occurring clonally, and the remaining SCNA occurring subclonally (median, >20%). For instance, in the KIRC case, R_GLX001, only 5% of the genome is impacted by clonal SCNAs (LOH of 3p, including *VHL*, *SETD2*, *BAP1*, *PBRM1*), while 84% of the genome was affected by subclonal SCNAs. Thus, in certain cancer types (including KIRC and LGG/GBM) the total SCNAs acquired generally mirrors the number of subclonal SCNAs while in other cancer types a more complex relationship exists with evidence of SCNAs driven by CIN fixed early in the MRCA and ongoing CIN after subclonal diversification.

The application of the mPHASE algorithm (Chapter 3) to the data analysed in this chapter produces estimates of gains and losses relative to ploidy and allele-specific copy number. These allele-specific copy number estimates indicate areas of AI (AI)

and, a sub-category of AI, loss of heterozygosity (LOH) which, when combined with the multi-region nature of the data analysed, produce estimates of clonal and subclonal AI and LOH. These measures will often, but not always, co-occur with events that are classed as gains and losses (Figure 5-4). This multi-region assessment of AI allows the detection of mirrored subclonal allelic imblance (MSAI) in which the major allele (or most prevalent homologous chromosomes' genetic material) differs between two regions of the same patient's disease. This indicates the presence of 2 or more subclones of which a subset may demonstrate parallel events. Parallel events are observed if the same event occurs at a chromosome locus in multiple regions of the patient's disease but originates from a different allele. These measures of between-region subclonal LOH and MSAI could not be detected in single region studies. In some patients' disease MSAI affected large proportions of the genome. Figure 5-4 shows three regions from a C_STX003, a COAD patient originally from (Sottoriva et al., 2015), demonstrating large-scale MSAI events across the genome affecting a total of 37% of the genome.

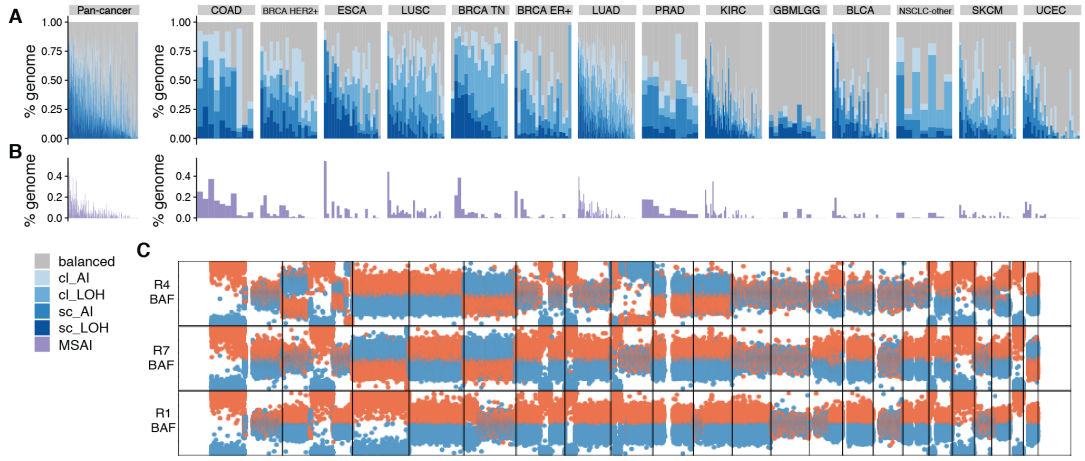

Figure 5-4: Overview of clonal and subclonal AI across cancer types
A) For each patient, the percentage of the genome that is affected by clonal and subclonal AI as well as clonal and subclonal LOH (a subset of AI) is indicated. Cancer types and patients are ordered based on the median percentage of subclonal total copy number SCNAs.
B) For each patient the proportion of the genome affected MSAI is shown C) BAF profiles from three regions of COAD tumour C_STX003 (originally from (Sottoriva et al., 2015)) demonstrating 37% of the genome affected by MSAI events.

Clonal and subclonal SCNAs were observed to occur in cancer type specific patterns with some clonal events shared between cancer types. For example Loss of chromosome 3p was a highly clonal event in BRCA TN, KIRC, LUSC, NSCLC-other

while gain of 5p appeared to be a subclonal event in BRCA ER+, BRCA TN, COAD, ESCA, GBMLGG and SKCM (Figure 5-5).

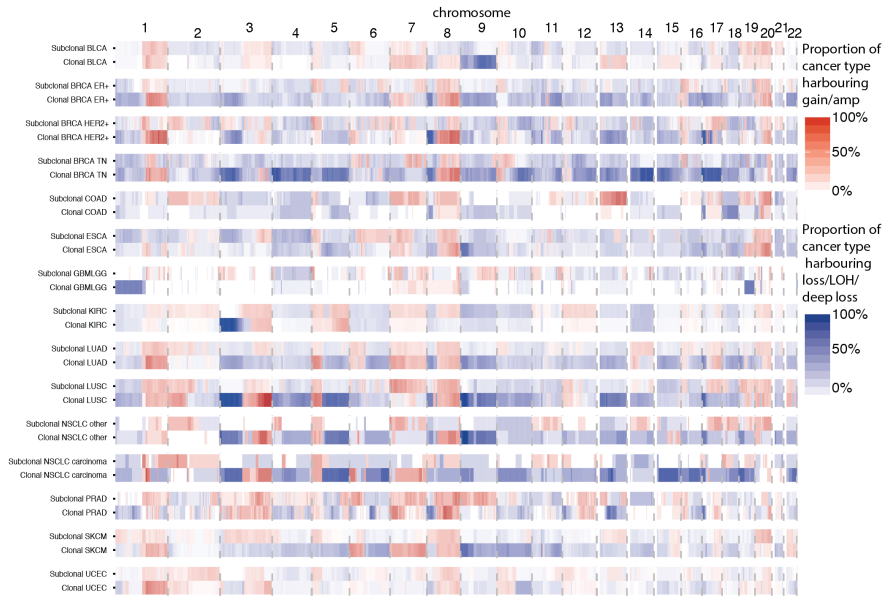

Figure 5-5: Across genome clonal and subclonal gain and loss events by cancer type
Each cancer type is represented by two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event.

In addition, our multi-sample cohort permitted high-resolution clonality analysis of WGD. In total, 46% of cancers examined were classified as clonally genome-doubled (GD), 7% as subclonally genome-doubled (sGD), and 46% as non-genome-doubled (nGD). Thus, in most cases WGD is a clonal event, consistent with studies looking to time the occurrence of WGD in single samples (Dentro et al., 2018). The tumour types with the highest proportion of subclonal WGD were MSS COAD (20%), KIRC (17%), SKCM (10%) and PRAD (10%) (Figure 5-6A).

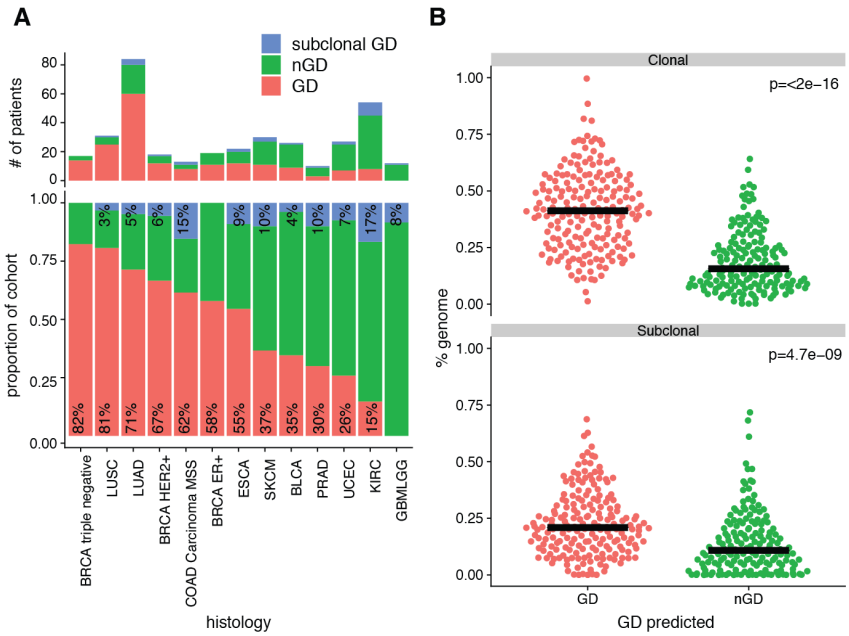

Figure 5-6: Multi-region whole genome doubling detection and SCNA clonality
A) Barplots indicating the number and proportion of cases of each histology type that are whole-genome doubled (GD). Subclonal GD cases (indicated in blue) are those with GD in some but not all regions of the tumour. B) Beeswarm plots comparing the proportion of the genome affected by clonal or subclonal SCNAs in GD and no GD (nGD) cases.

To determine the impact a WGD event may have on cancer genome evolution, we explored differences in SCNA acquisition between GD and nGD tumours. GD tumours showed significantly increased levels of clonal gains and losses ($p<0.01$, t-test), as well as subclonal gains and losses ($p<0.01$, t-test), suggesting that WGD may be permissive for ongoing CIN (Figure 5-6B).

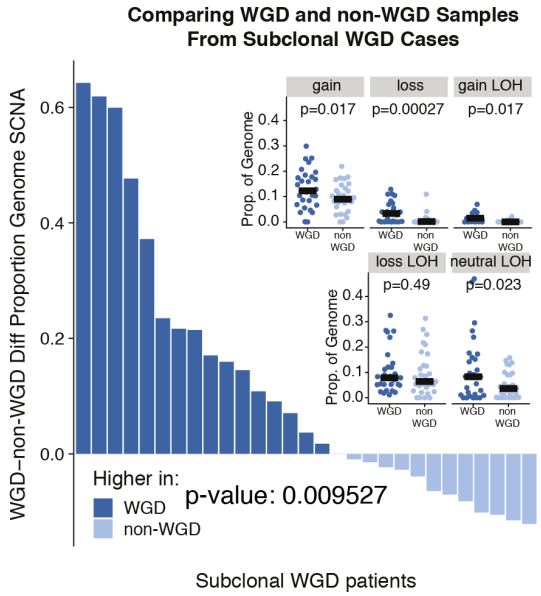

Figure 5-7: Investigating SCNA burden in subclonal GD cases
Comparing the proportion of the genome affected by clonal or subclonal SCNAs in matched GD and nGD samples from subclonal GD cases. Bars indicate for each subclonal GD case the difference between the median percentage of the genome affected by SCNAs in GD and nGD samples. The inset beeswarm plots compare the percentage of the genome affected by different types of SCNAs in GD and nGD samples.

Consistently, in sGD tumours we observed a higher frequency of SCNA events in the subclones affected by WGD compared to their nGD sister clones (p= 0.0095, paired t-test, Figure 5-7), thus removing the germline background and prior clonal somatic alterations as potential confounding.

I also noted that in some sGD cases MSAI events occurred mainly in the regions of the tumour subject to GD regions, an example of which is shown in Figure 5-8A in KIRC patient's disease whose original identifier is EV001 from (Gerlinger et al., 2012). This lead me to explore the relationship between MSAI events and GD by examining the proportion of the genome effected in GD versus nGD cases. There was a higher proportion of the genome was subject to MSAI in those with clonal GD compared to nGD cases (p=0.000006, paired t-test, Figure 5-8B). Next, to attempt to observe this relationship in an experimental system I analysed data from a previously (Dewhurst et al., 2014) that examined GD and nGD sister clones from the same ancestral clone of an HCT116 colorectal cancer cell line(Figure 5-8C). These diploid and tetraploid clones were passaged and subjected to SNP array and SCNA analysis at regular intervals of 20 - 25 passages. Previous analysis of these data had observed anincrease in SCNA events in the tetraploid clones with increased passage number mainly driven by losses which resulted in decreasing ploidy over time. I reanalysed

these SCNA calls and revealed multiple instances of MSAI affecting chromosomes 3, 4, 9, 11 and 16 that increased in prevalence over time (Figure 5-8D).

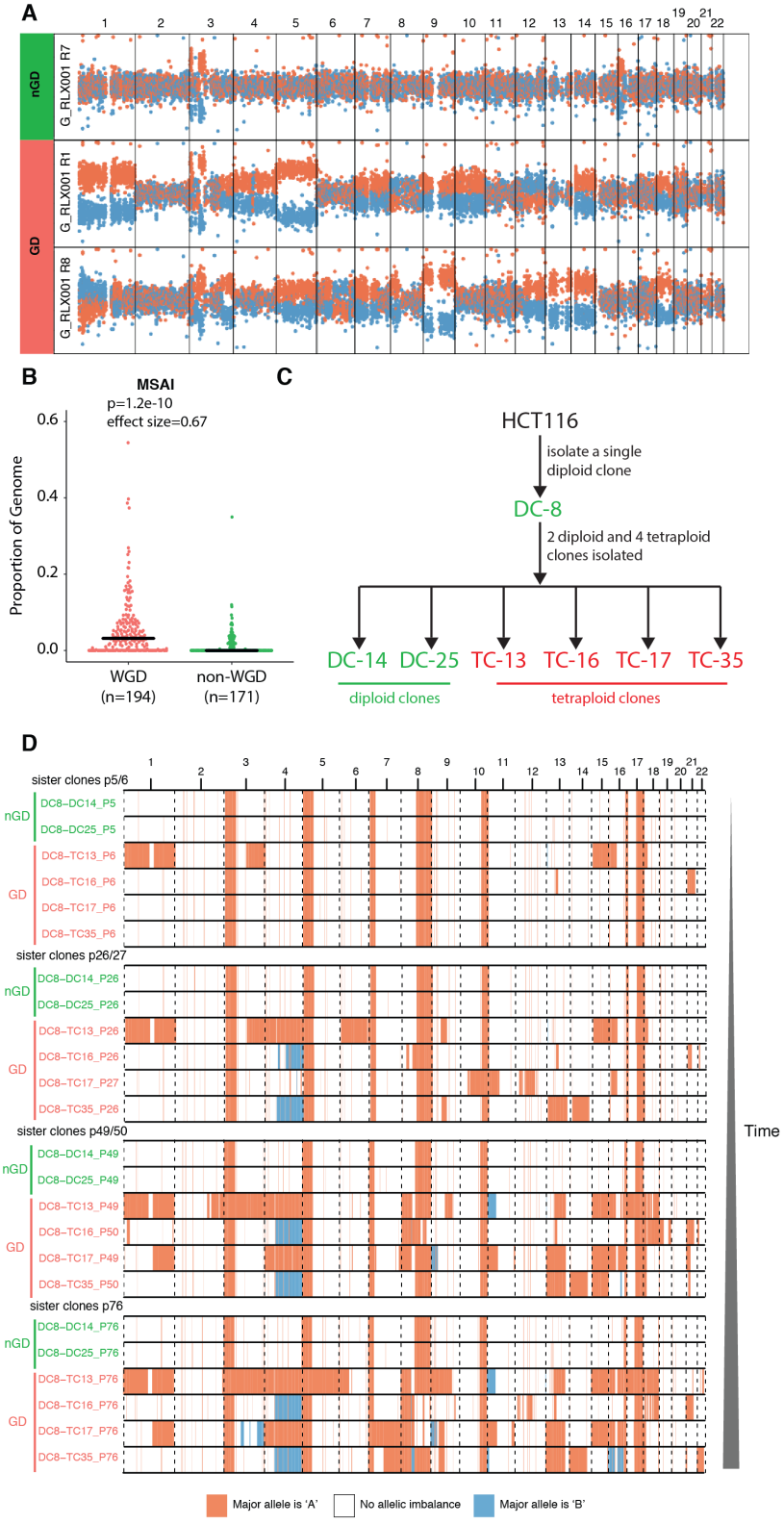

Figure 5-8: Investigating the relationship between MSAI and GD
A) shows the BAF profiles across the entire genome from three regions of a KIRC tumour G_RLX001 (EV001 from (Gerlinger et al., 2012)). The first track is from a nGD and demonstrates little allelic imbalance, the second two

tracks are from two GD regions and demonstrate AI and MSAI. B) Beeswarm plot comparing the proportion of the genome affected by MSAI in GD and non-GD cases. C) Schematic showing the derivation of genome doubled and non-genome doubled HCT116 cell line clones. D) Across genome plots of the major allele in areas of AI for the nGD and GD clones at four increasing passage intervals, MSAI can be observed between the HCT116 clones where the major allele at the same genomic position differs between clones in areas demonstrating AI.

SCNAs that were present in the ancestral HCT116 clone were observed in all diploid tetraploid clones that examined. No novel SCNA loss events were common to the first passage of the tetraploid HCT116 clones examined. However, an area of 4q containing 362 genes went from four to three copies in all in all four tetraploid clones (necessarily by independent events) by passage 49/50 (~1 year). This loss of chromosome 4q occurs after genome doubling, as it did not occur before passage 25 in any HCT116 tetraploid clone. The ubiquitous loss of the 4q arm is consistent with potential selection for loss of this region during prolonged culture and has also been previously found to be a significantly lost arm-level loss event by TCGA GISTIC2 analyses of COAD cases (TCGA, 2017).

Taken together, these data provide substantial evidence that ongoing subclonal CIN is pervasive during the evolution of many tumours, but requires adequate sampling of the tumour mass to be detected. We find WGD to generally be an early event in tumour evolution, associated with an increased acquisition of both clonal and subclonal SCNAs. The diversity generated by this pervasive ongoing CIN may act as a substrate for Darwinian selection.

5.4 Early tumour evolution is defined by clonal LOH followed by WGD

To investigate whether order and repeated patterns of SCNAs exist in tumour evolution, we assessed whether recurrent SCNAs could be identified. Using a permutation-based assessment of gains and losses independently, we considered the proportion of the genome subject to clonal/subclonal SCNAs within each tumour region and tested which SCNAs were present more frequently than expected by chance ($p<0.05$, 10,000 permutations, see Chapter 2).

Within each cancer type, distinct clonal and subclonal significantly recurrent events were identified, suggesting the acquisition of SCNAs is non-random during tumour evolution. Focusing first on clonal events, we then considered whether our identified significant events were enriched for cancer genes identified independently through

mutation-based statistical analysis (Bailey et al., 2018). We observed a tendency for significantly recurrent gains to be enriched for oncogenes (P <0.01 in 4/12 cancer types), while losses were enriched for tumour suppressor genes (TSGs) (P<0.01 in 12/14 cancer types) (Figure 5-9A).

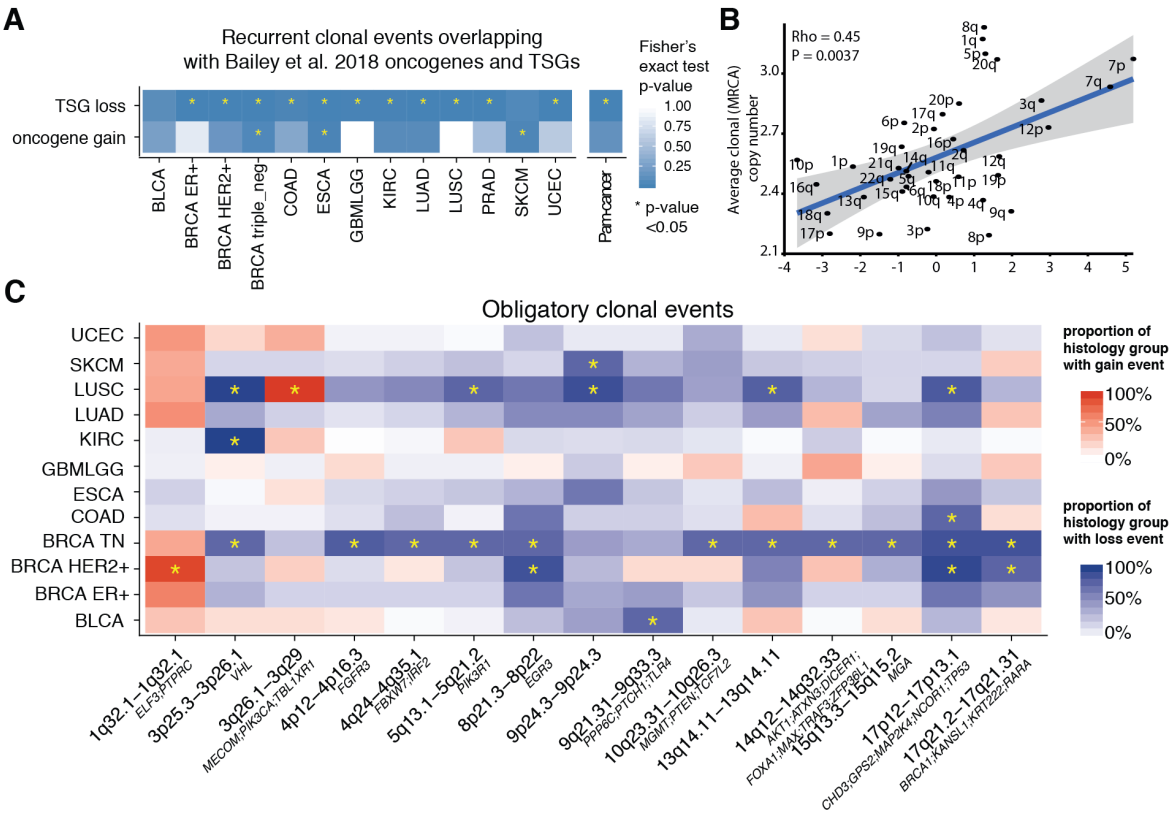

Figure 5-9: Investigating recurrent clonal SCNAs in the pan-cancer cohort
A) Heat map indicating where groups of recurrent clonal events significantly overlap with known TSG and oncogenes. B) Scatter plot depicting a positive correlation between TSG-Oncogene scores and average clonal copy number of each chromosomal arm. C) Heatmaps indicating obligatory clonal SCNA events for each histology type. Red shading indicates that a gain is predominant, while blue shading indicates that loss/LOH is predominant. A yellow asterisk indicates that the event is significantly obligatory clonal in that cancer type.

It was recently revealed that the potency and distribution of oncogenes and tumour suppressor genes (referred to as the TSG-OG score) influence cancer karyotypes and lead to amplification of chromosomes with a net oncogenic effect (Davoli et al., 2013). To evaluate this relationship between oncogenicity and copy-number status in our cohort, we considered clonal arm-level events present in the most recent common ancestor (MRCA) inferred using MEDICC (Beerenwinkel et al., 2014) of each patient and found a positive correlation between the net changes in the copy number of chromosome arms and their TSG-OG score (Figure 5-9B).In 7 out of 13 cancer types we observed evidence for obligatory clonal SCNAs, occurring at frequencies >75%

in their respective cancer types. 13/15 obligatory clonal events involved LOH, while only 2 were gains/amplifications (Figure 5-9C).

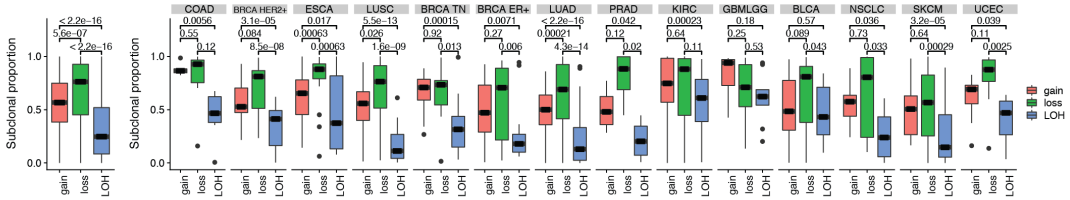

Figure 5-10: Investigating the clonality of gain, loss and LOH events.
Boxplots are shown for pan-cancer and each cancer type with ≥ 10 separate cases. For each SCNA type (gain/amp, loss/deep loss, LOH), the proportion of subclonal SCNAs is indicated, where each dot represents a patient.

Consistent with this, losses involving LOH were significantly more likely to be clonal than gains (Figure 5-10), potentially reflecting that LOH is necessarily permanent; once an allele is lost it cannot be regained. The tendency for clonal LOH was particularly pronounced in GD tumours, consistent with loss and LOH of tumour suppressor genes prior to GD, followed by further gains and losses (Figure 5-11).

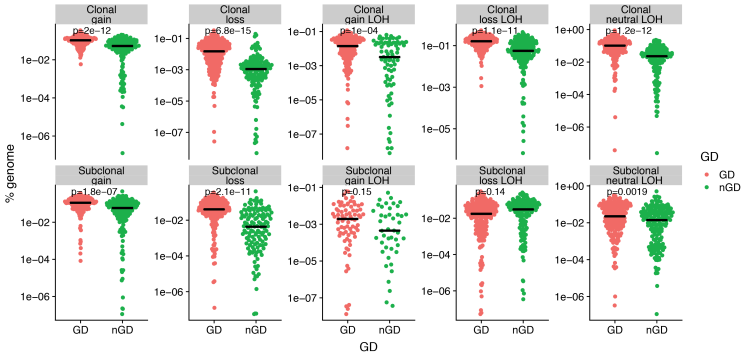

Figure 5-11: Clonal and subclonal SCNAs in relation to genome doubling
For each SCNA type, the proportion of the genome affected by clonal or subclonal alterations is indicated for each genome doubled (GD) and non-genome doubled (nGD) case. With the exception of subclonal loss LOH, SCNAs tend to affect a greater proportion of the genome in GD cases than in nGD cases.

SCNAs have been noted to affect more of the genome than any other somatic genetic alteration (Beroukhim et al., 2010, Zack et al., 2013). Extending this observation, it is interesting to note that obligatory clonal SCNA events generally occurred at frequencies that exceed most clonal point mutations in cancer genes (Figure 5-12).

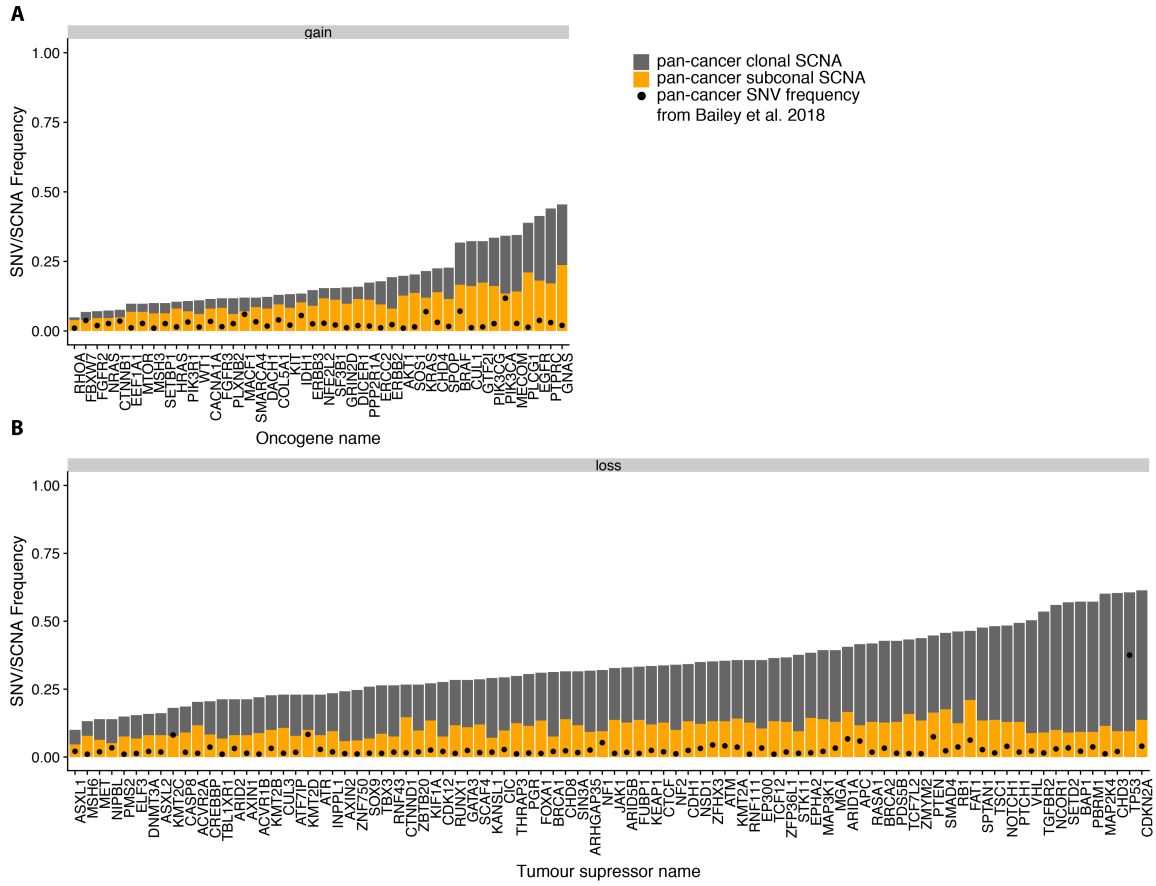

Figure 5-12: SCNA events are generally more frequent than point mutations in cancer genes
A) Shows oncogenes found to be significant in a previous analysis (Bailey et al., 2018) and their frequency of alteration by sequence-level variants in the pan-cancer dataset from that publication shown by the black points. The stacked bar chart shows the frequency of clonal and subclonal gain/amplification SCNAs affecting the same oncogenes relative to ploidy in the pan-cancer dataset examined in this thesis, clonal in grey and subclonal in yellow B) Shows tumour suppressor genes and their frequency of alteration by sequence-level variants in the same dataset as panel A. Clonal and subclonal loss/LOH/deep loss SCNA events are in the pan-cancer cohort examined in this thesis are shown in the stacked bar chart as grey and yellow respectively.

For example in LUSC, clonal mutations in TP53 occur in >50% of tumours, whereas clonal loss/LOH to 3p (including *SETD2*), 9p (including *CDKN2A*) and 17p (including *TP53*) as well as clonal gains of 3q (including SOX2, PIK3CA) were present in over 90% of the cohort and were identified in over 90% of cases as clonal (Figure 5-13A). Likewise, 9q loss/LOH in BLCA was present in greater than >75% of the cohort, whereas the previously reported highest-frequency clonal somatic point mutation is in *TP53*, occurring in 51% of patients (Figure 5-13B).

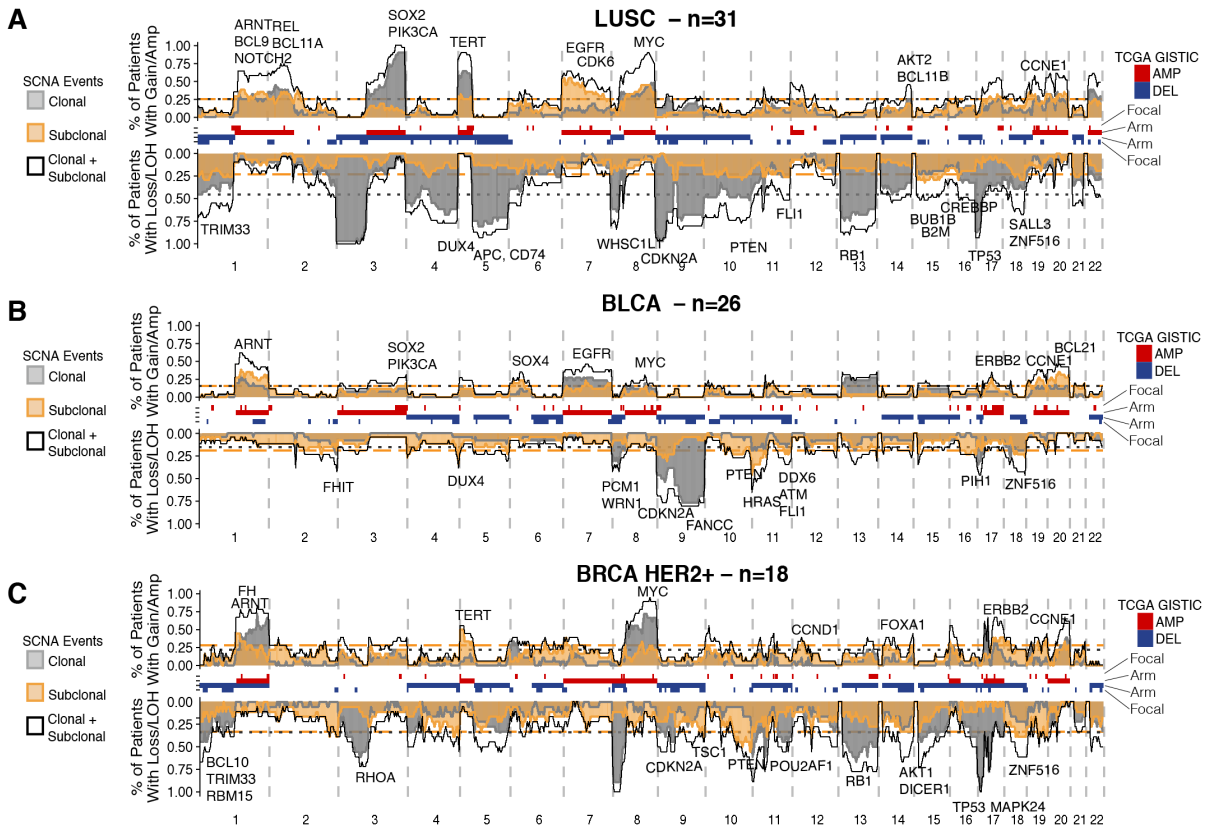

Figure 5-13: Across-genome by cytoband plots show the SCNA frequency and clonality
The cancer types represented are A) LUSC, B) BLCA and C) BRCA HER2+. The frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards. The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events. Events have been enumerated according to chromosome cytoband. The dotted grey and orange lines represent significance thresholds for clonal and subclonal events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type.

In BRCA HER2+, loss/LOH of 8p and 17p (including TP53) were present in more than 75% of tumours and were almost exclusively clonal (Figure 5-13C).Importantly, 17p was subject to copy-number neutral LOH, while 8p losses were associated with only 1 remaining copy. In regions of 17p LOH, the two remaining copies frequently harbour TP53 mutations (the classical ‘double hit’), suggesting a point mutation followed by LOH and subsequent WGD as the most parsimonious succession of events. 8p LOH does not contain any known tumour-suppressor genes and its loss has been linked to a chromosome-dosage effect, influencing lipid metabolism and metastatic potential in BRCA HER2+ (Cai et al., 2016). Indeed, in 8/13 clonally genome doubled BRCA HER2+ cases, 8p was present only at a single copy and was therefore likely to have undergone either two separate loss events: one before and one after WGD, or three loss events following WGD.

Intrigued by the observed interplay between LOH events and WGD, we also investigated enrichment of specific SCNA events in GD tumours (Zack et al., 2013, Cheng et al., 2017, Bielski et al., 2018). Consistent with previous studies, we identified LOH at the loci of *TP53* (17p13.1), *RB1* (13q14.2) and gains at the loci of *CSMD1* (8p23.2), *ABL 1*(9q34.12) and *CCNE1* as significantly enriched in GD compared to nGD tumours (Table 5-1).

Event	Estimate	Std_Error	z_value	p_value	significance	previously identified by
(Intercept)	-2.11698	0.53121	-3.985	6.74E-05	***	
histologyLUAD	1.64155	0.54485	3.013	0.00259	**	
8p23.2_CSMD1-8p21.2_PPP2R2A_LOH	0.72302	0.26618	2.716	0.0066	**	Zack et al. 2013
histologyCOAD	1.71325	0.73308	2.337	0.01944	*	
17p13.1_TP53-17p12_MAP2K4_LOH	0.71951	0.3101	2.32	0.02033	*	Bielski et al. 2018
19q12_CCNE1_amp	2.45028	1.09426	2.239	0.02514	*	Bielski et al. 2018
9q34.12_ABL1_LOH	0.61424	0.29147	2.107	0.03508	*	Bielski et al. 2018
13q14.2_RB1_LOH	0.59484	0.29507	2.016	0.04381	*	Bielski et al. 2018
histologyLUSC	1.42172	0.70909	2.005	0.04496	*	
14q21.1_FOXA1_LOH	-0.45594	0.31344	-1.455	0.14576	ns	Bielski et al. 2018
9p21.3_CDKN2A_LOH	0.4415	0.30358	1.454	0.14586	ns	Zack et al. 2013
histologyBRCA	0.8585	0.61943	1.386	0.16576	ns	
histologySKCM	0.77831	0.62422	1.247	0.21245	ns	
1p12_NOTCH2_LOH	-0.25937	0.32282	-0.803	0.4217	ns	Bielski et al. 2018
11q13.3_CCND1_amp	0.38529	0.50131	0.769	0.44215	ns	Bielski et al. 2018
histologyESCA	0.52827	0.72107	0.733	0.46379	ns	
histologyNSCLC-other	0.52955	0.78328	0.676	0.499	ns	
17q11.2_NF1_LOH	0.22179	0.34954	0.635	0.52575	ns	Zack et al. 2013
histologyGBMLGG	-0.74806	1.19214	-0.627	0.53033	ns	
10q23.31_PTEN_LOH	0.17053	0.28148	0.606	0.54461	ns	Zack et al. 2013
3p21.1_BAP1-3p24_TGFBR2.1_LOH	0.21249	0.36011	0.59	0.55516	ns	Bielski et al. 2018
histologyOther	0.51233	0.88466	0.579	0.56251	ns	
17q23.2_BRIP_amp	0.41231	0.76878	0.536	0.59174	ns	Bielski et al. 2018
histologyUCEC	0.37012	0.70009	0.529	0.59703	ns	
20p12.2_PAK7_LOH	-0.12462	0.34113	-0.365	0.71487	ns	Bielski et al. 2018
20q11.21_BCL2L1_amp	-0.43558	1.19408	-0.365	0.71527	ns	Bielski et al. 2018
histologyPRCA	0.15048	0.85744	0.175	0.86069	ns	
6p21.2_CDKN1A-6q21_PRDM1-6q26_PARK2_LOH	0.05966	0.34981	0.171	0.86459	ns	Bielski et al. 2018
histologyKIRC	0.10252	0.63788	0.161	0.87232	ns	
8q24.21_MYC_amp	-0.02538	0.35051	-0.072	0.94229	ns	Zack et al. 2013

Table 5-1: Investigating known associations with genome-doubling in a pan-cancer multi-region cohort
Full results of a multi-variable regression model incorporating of previously described associations with WGD in the pan-cancer multi-region cohort.

None of these events were found to be mutually exclusive with one another. While LOH at the *TP53* (17p13.1) locus was found to be clonal in 90% of GD cases in which it was observed, *CCNE1* amplifications, which occurred exclusively in GD tumours, were subclonal in 45% of these cases. This relatively low clonality suggests *CCNE1* amplifications may be selected for after WGD, rather than promoting WGD. Similarly, subclonal LOH occurred at the loci of *RB1*(13q14.2) in 24%, *CSMD1*(8p23.2) in 20%, and *ABL1*(9q34.12) in 20% of WGD cases in which they were identified (Figure 5-14).

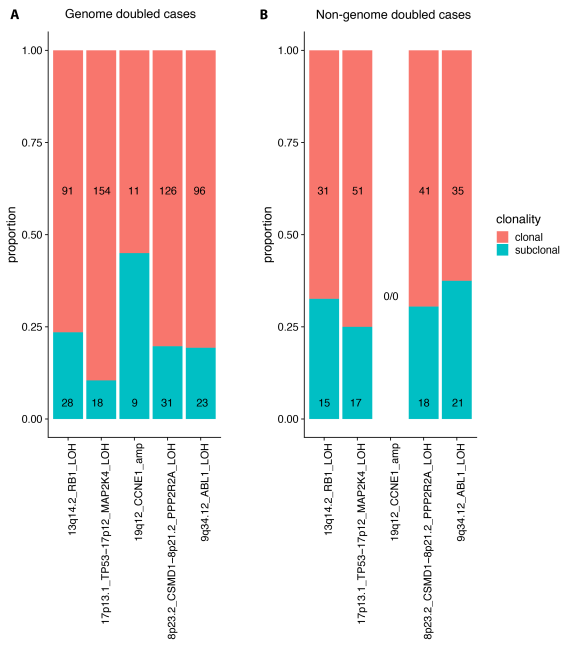

Figure 5-14: Clonality of significant associations with genome-doubling
A) Shows the clonality and exact counts in genome-doubled cases of events that have been previously described as associated with WGD in other studies and found significant in the pan-cancer multi-region cohort investigated. B) Shows the same events as in A but in non-genome doubled cases.

In UCEC, two major divergent paths of clonal SCNA acquisition were observed (Figure 5-15) 1q gain (including *H3F3A*, *PTPRC*) occurred in 52% (12/23) of patients and was clonal in 75% of cases. 1q gain was mutually exclusive with 17p LOH (including TP53), which occurred in 30% of patients and was clonal in 85% of cases. 71% of the cases with 17p LOH had undergone GD, while none of the cases with 1q gain had undergone GD, suggesting that early evolutionary paths in UCEC constrain the acquisition of future WGD events. Notably, a significant difference in the proportion of the genome subject to subclonal SCNAs was observed between patients harbouring 17p LOH and those without ($P<0.01$, t-test). In LGG/GBM we also observed two divergent evolutionary paths, with 25% (3/12) patients harbouring a clonal 17p LOH events which was mutually exclusive with 19p and 1q loss which co-occurred and were found in 67% (8/12) of patients. Jointly, these data suggest that *TP53* LOH typically occurs prior to WGD, in line with its role in enabling tolerance for chromosome copy number alterations (Dewhurst et al., 2014), often followed by the selection of other subclonal losses or amplifications.

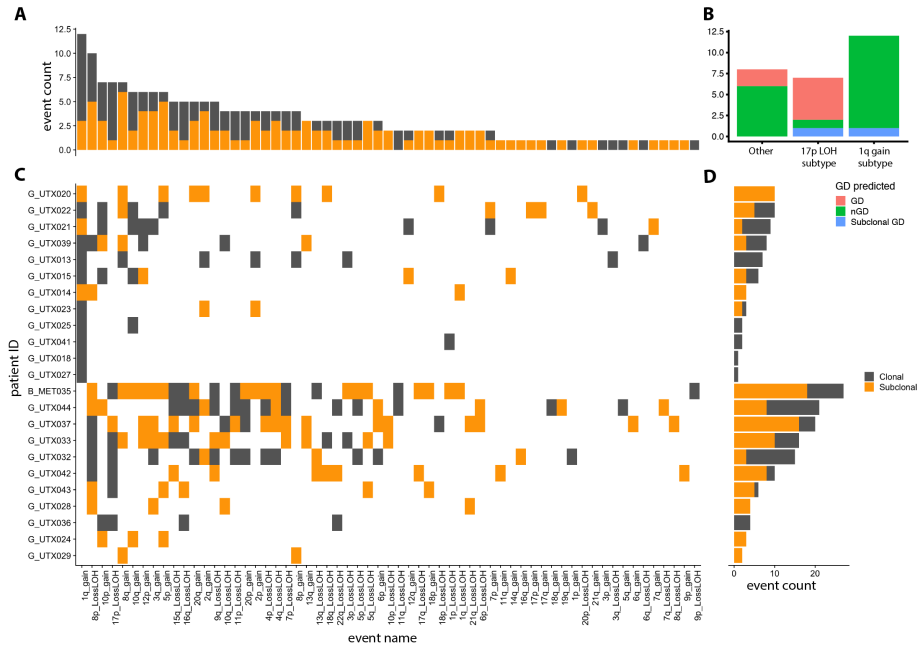

Figure 5-15: Divergent paths of SCNA acquisition in UCEC
Oncoprint showing events present in UCEC cases. A) shows total count of each arm level event (the events are enumerated at the bottom of the panel C) across the entire cohort. B) shows the number of cases found to be GD, sGD or nGD depending on whether they demonstrate 17p LOH, 1q gain or neither. C) Shows the arm level events present in each individual case. D) Shows the total number of arm level events observed in each tumour.

Figure 5-5 displayed summaries of the clonal and subclonal SCNA events across the genome in different cancer types. However, for completeness, I have included more detailed individual plots illustrating the clonal and subclonal SCNA prevalence with the cases of that cancer type across the genome, how these events compare to previous studies investigating single regions, MSAI and plots investigating clonal and subclonal LOH. These are displayed for all cancer types contributing greater than 10 cases to the cohort investigated in this chapter and constitute plots 5-16 to 5-29. Areas of the genome that reach the significance threshold for their respective SCNA category (split by gain and loss as well as clonal, subclonal and combined, see Chapter 2) are shown to cross a line representing the respective threshold. In order to discuss the SCNA landscape of each cancer type I make a distinction between small-scale peaks of SCNAs whose significant area occupies less than half of a chromosome arm and large-scale peaks of SCNAs that occupy greater than half of a chromosome arm. These classifications are not equivalent to arm and focal events as all SCNAs of any size from each from individual patient's disease are combined to produce a cohort level dataset that is then investigated.

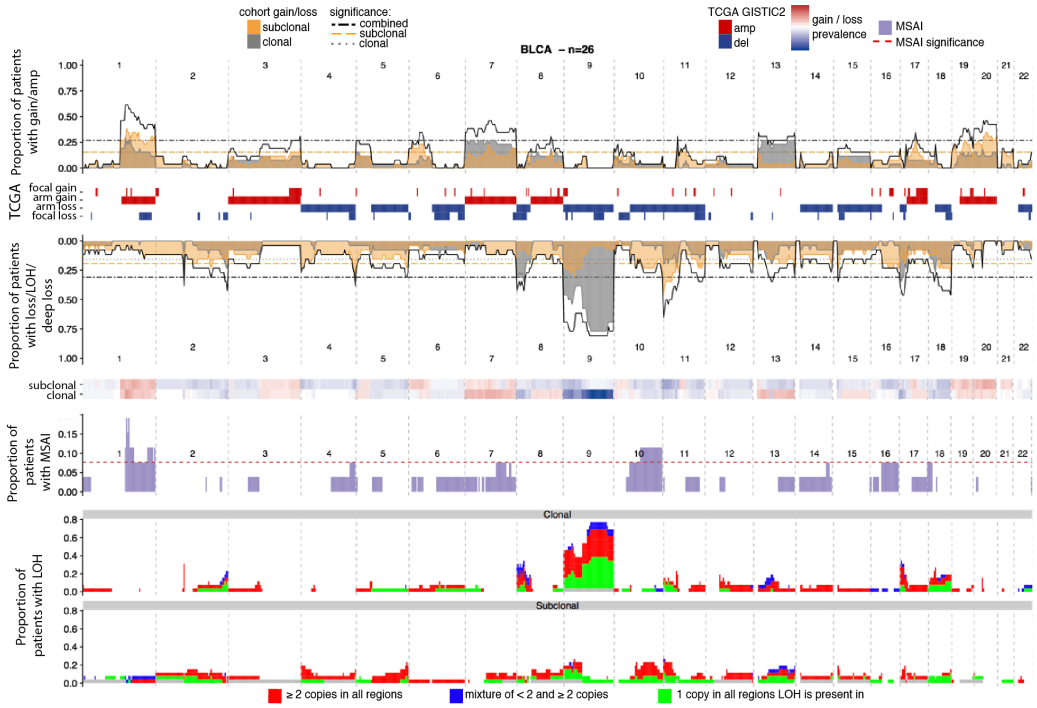

Figure 5-16: BLCA across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The SCNA landscape of the 26 multi-region BLCA cases examined (Figure 5-16) is dominated by large-scale clonal losses on 9q present in over 75% of the cohort. These losses are in the vast majority of cases also LOH events with a mixture of overall loss and copy neutral LOH. Significant large-scale clonal gains include areas of chromosomes 1q, 7p, 7q, 13q with small-scale significant subclonal gains observed on 11q and 17q. Other clonal loss events reaching significance include large-scale areas of 8p and 17p and 18q. There are few significant areas of significant subclonal loss with small-scale areas of chromosomes 4q and 10q and large-scale areas of 11q, 16q and 18q.

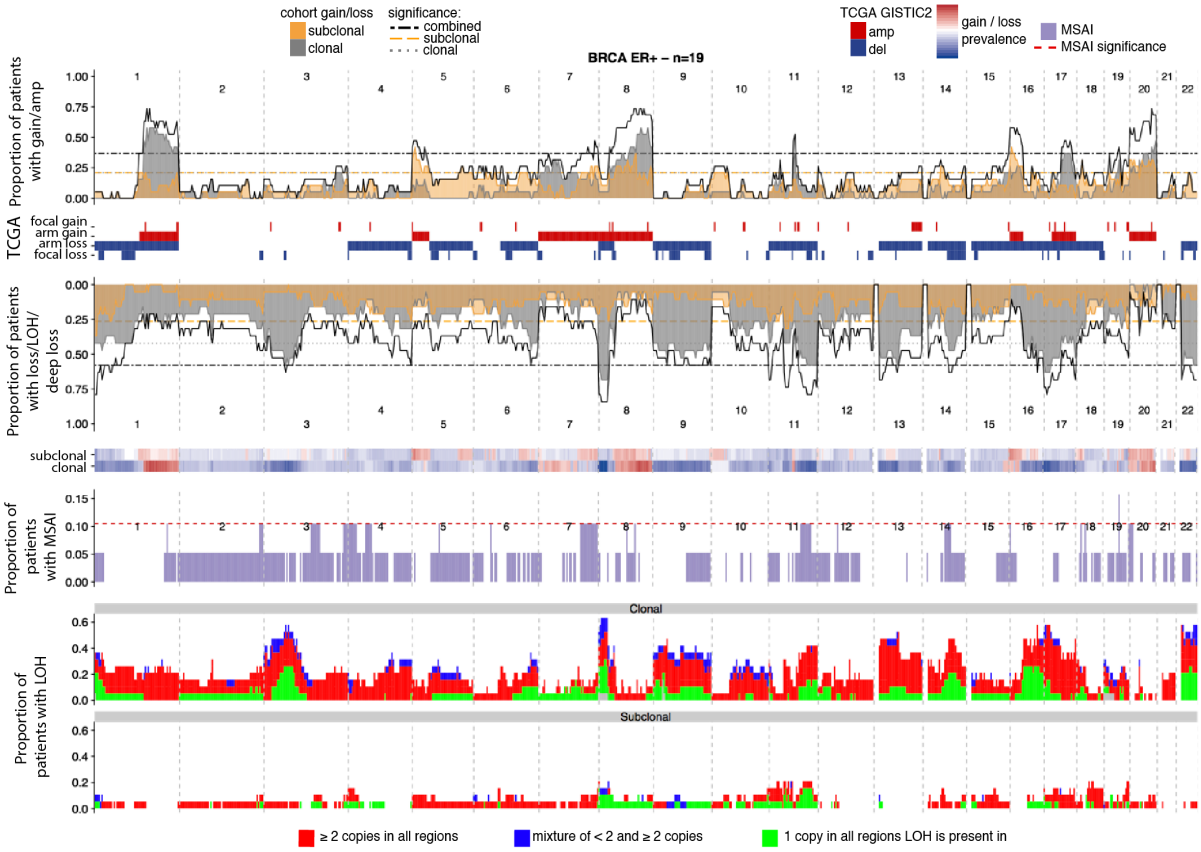

Figure 5-17: BRCA ER+ across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.
.

The SCNA landscape of the 19 multi-region BRCA ER+ cases (Figure 5-17) demonstrates significant large-scale clonal gains of areas of 1q, 8q, 17q and 20q. Significant small-scale areas of clonal gain were observed on 11q, 16p and 19q. Significant large-scale subclonal gains were found on 5p and 16p and significant small-scale losses were found on only 1p and 14q.

Significant large-scale clonal losses were observed on a large number of chromosome arms: 3p, 8p, 9p, 9q, 11q, 13q 16q, 17p, 22p and 22q while only a few arms demonstrated significant small-scale clonal losses with 1p and 14q affected. Significant subclonal losses were limited to only small-scale subclonal losses on 1p, 11q, and 19q.

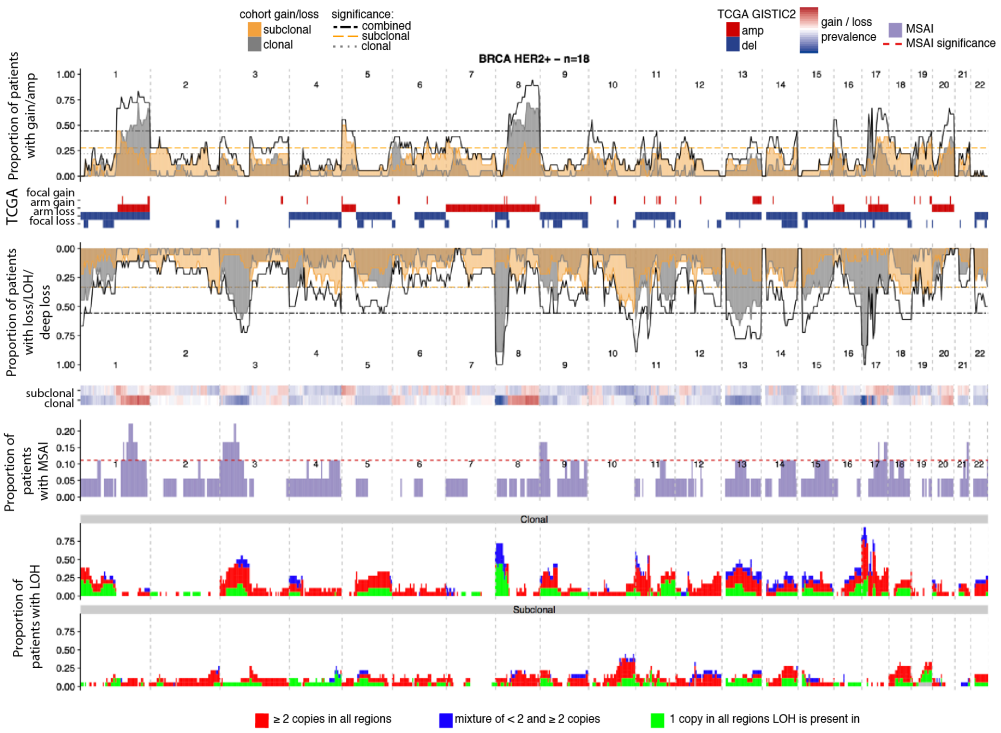

Figure 5-18: BRCA HER2+ across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The SCNA landscape of the 18 multi-region BRCA HER2+ cases (Figure 5-18) examined demonstrated significant large-scale clonal gains of areas of 1q, 6p, 8q with the only significant area of small-scale clonal gain was found on 20q. Like BRCA ER+ but unlike BRCA TN one of the significant large-scale subclonal gains was observed to be on 5p, with 19p and 19q the only others affected. Significant small-scale subclonal gains affected 1q, 2p, 3q, 6q, 7p and 14q.

Significant large-scale clonal losses affected 3p, 8p, 13q and 17p while small-scale closes affected 9p and 15p. Large-scale subclonal losses were observed on 10q and 18q while small-scale subclonal losses were found on 2q and 4q.

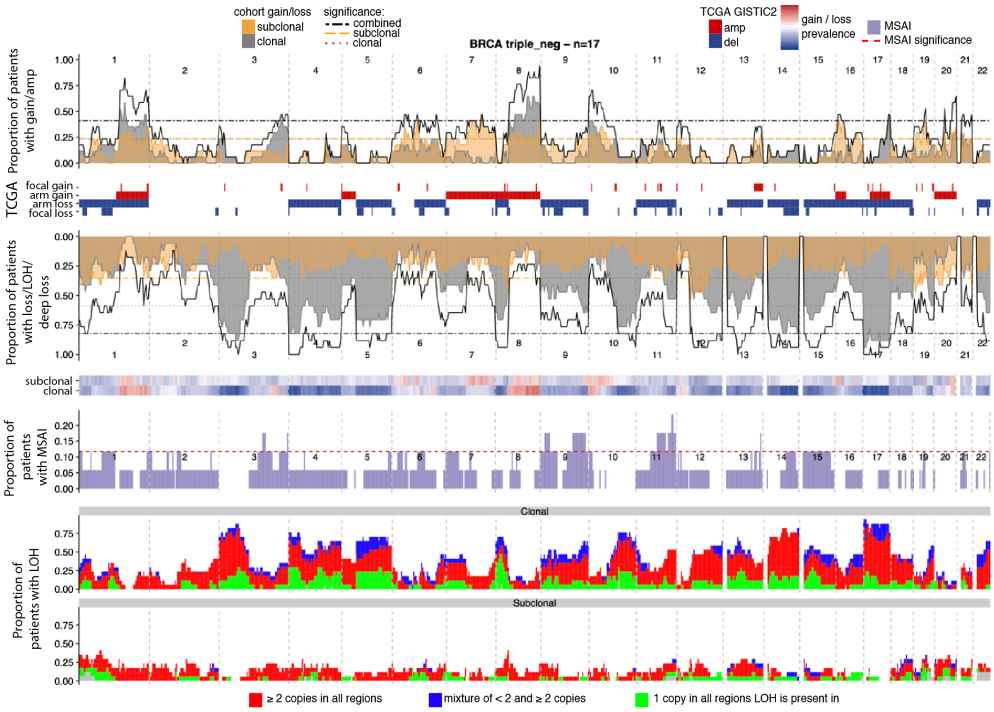

Figure 5-19: BRCA TN across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type

cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The 17 cases of BRCA TN (Figure 5-19) demonstrated significant large-scale clonal gains affecting 1q, 3q, 8q and 10p with a single small-scale clonal gain on 17q. Large-scale subclonal gains affected 7q, 10p, 16p and 19p with small-scale subclonal gains on 1q, 6p, 17q.

BRCA TN was the cancer type most affected by LOH and large-scale clonal losses affected 3p, 4p, 4q, 5q, 8p, 10q, 15q, 17p and 17q with small-scale clonal losses: affecting 9q, 12q, 13q and 22q. Significant large scale subclonal losses were found on 1p, 12p, 19p and small-scale losses observed on 8p and 20p.

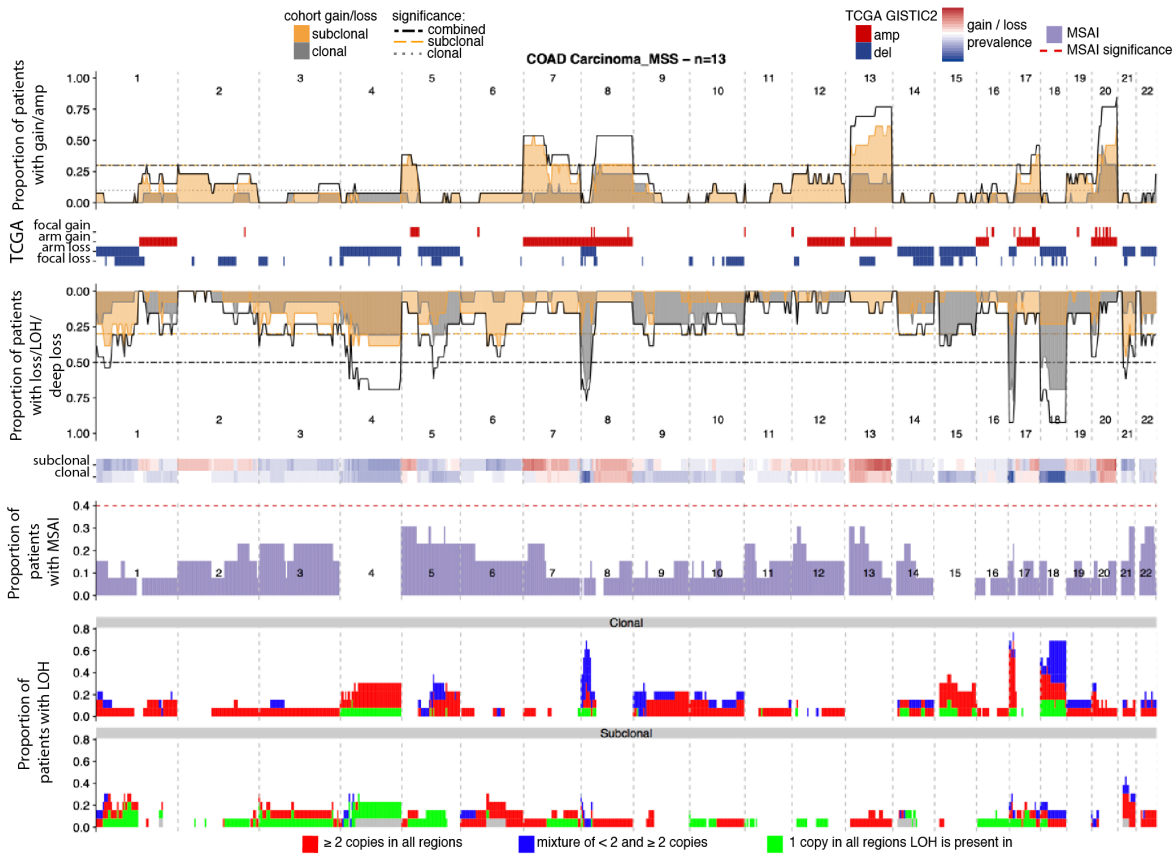

Figure 5-20: COAD across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the

TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The SCNA landscape the 13 COAD cases (Figure 5-20) demonstrate significant large-scale clonal gains on 8q, 13q and 20q with no significant small-scale clonal gains identified. Significant large-scale subclonal gains were found on 5p, 7p, 13q and 20q with the only small-scale subclonal gain on 17q.

Significant large-scale clonal losses on 4q, 8p, 17p and 18q were observed as well as a single small-scale clonal loss on 15q. Large-scale subclonal losses were observed on 4q and 21q with only 6q affected by a significant small-scale subclonal loss. It is interesting to note that the 4q large-scale significant subclonal loss affects the same area of the genome observed to undergo ubiquitous loss in the HCT116 genome-doubled cell lines clones when passaged over time. However, unlike the HCT116 cell line clones, in this relatively small cohort no MSAI of 4q was observed.

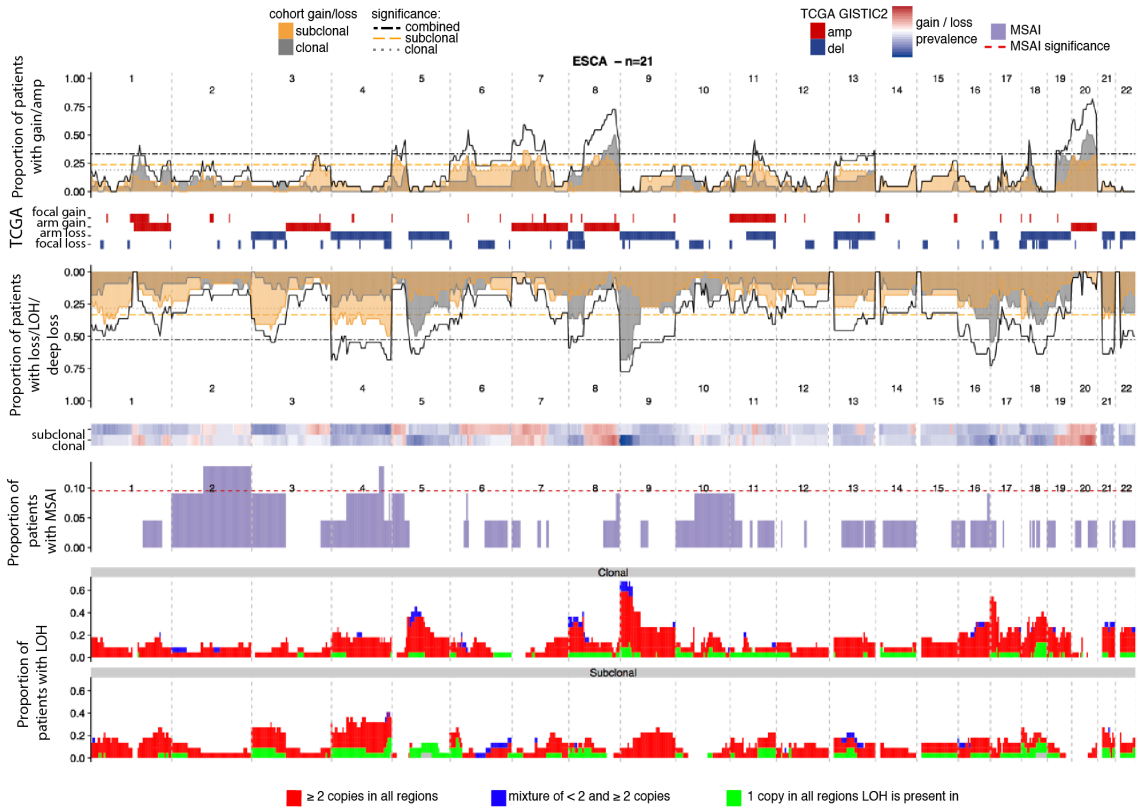

Figure 5-21: ESCA across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The SCNA landscape of the 21 ESCA cases analysed (Figure 5-21) in this chapter demonstrated significant large-scale clonal gains of 8q, 19q, 20p, 20q and small-scale clonal gains of 1q and 18q. Similarly to a number of other cancer types examined, a significant large-scale subclonal gain was found on 5p which, in this cancer type, is the only large-scale subclonal gain. Significant small scale subclonal gains were observed on 3q, 6q and 11q.

Significant large-scale clonal losses affected 5q, 8p, 9p, 16q, 17p, 18q, 20p and 20q with small-scale clonal losses observed on 1q, 7p and 18p. Significant large-scale

subclonal losses were observed on 3p, 4p, 4q and 21q with small-scale losses found on 1p and 18p.

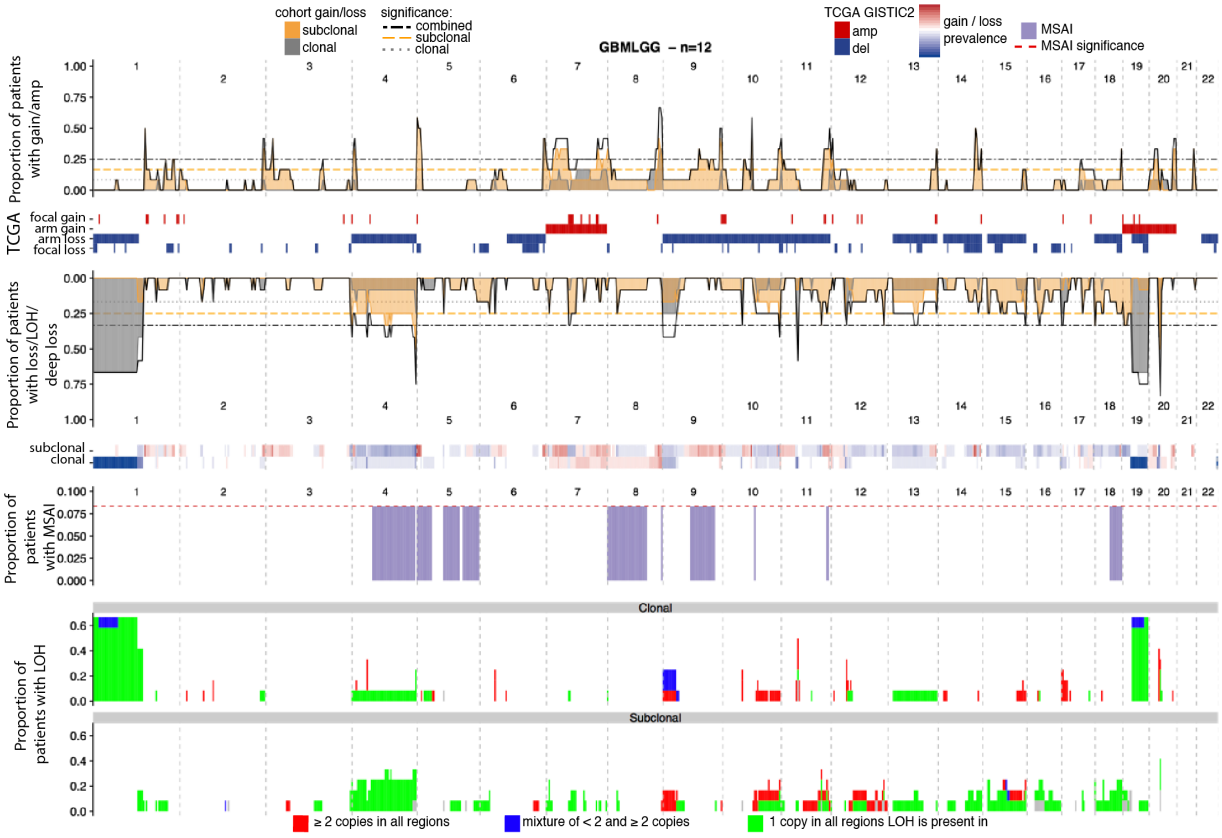

Figure 5-22: GBMLGG across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The SCNA landscape of the 12 cases of GBMLGG analysed in this chapter (Figure 5-22) have a relatively sparse copy number landscape with relatively few SCNAs. There were no significant large-scale clonal gain events were limited to a single small-scale clonal gain on 7q. Subclonal gains were more common with significant large-

scale subclonal gains observed on 5p and 7p and small-scale subclonal gains seen on 3p, 4p, 9q, 10q, 11q, 14q and 20q.

There were more significant clonal losses than gains in GBMLGG with significant large-scale clonal losses on 1p and 19q and small-scale clonal loss on 17p. Large scale subclonal losses were found in 4p, 4q and small-scale losses were observed on 16p.

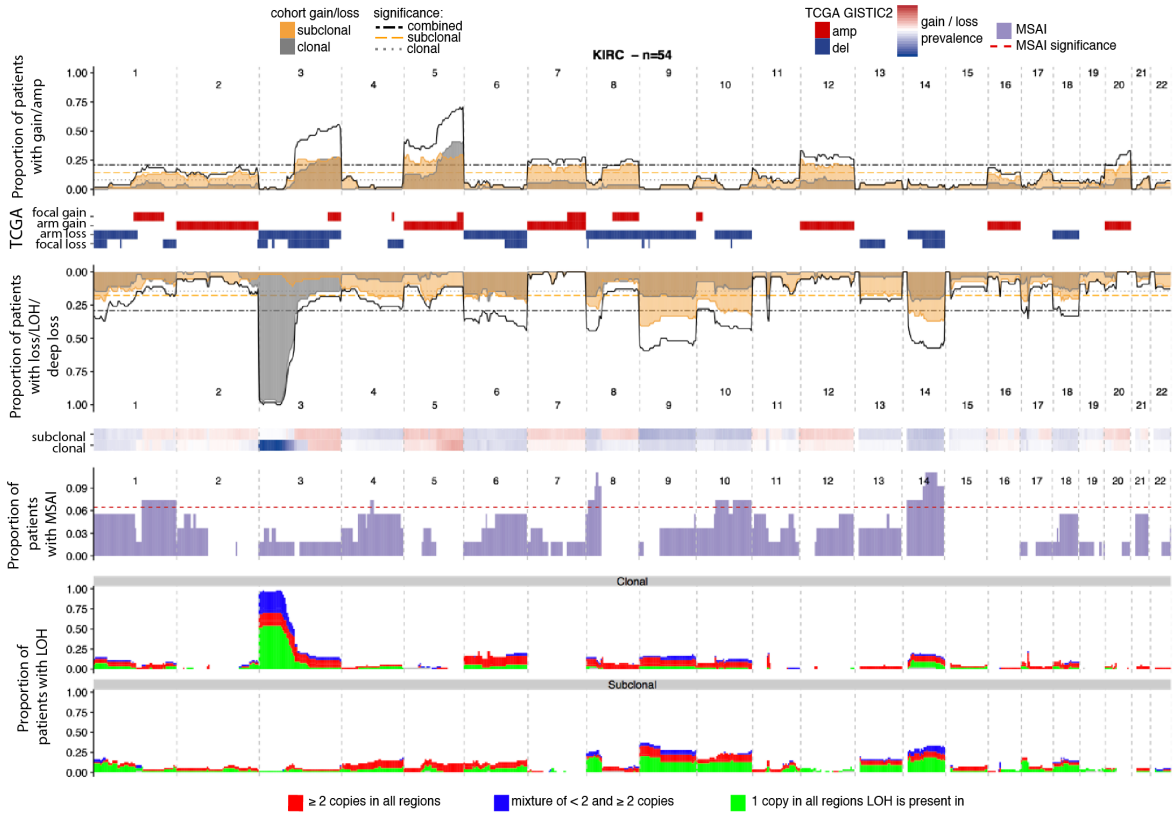

Figure 5-23: KIRC across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The SCNA landscape of KIRC (Figure 5-23) is dominated by large-scale events. No significant small-scale clonal gains were observed but with large-scale clonal gains observed on 3q, 5p, 5q and 20q. Significant large-scale subclonal gains were observed on 3q, 5p, 5q, 7p, 7q, 8q, 12p, 12q, 20p and 20q with small-scale subclonal gains observed on 11p and 17q.

Very few significant small-scale loss events were observed with only significant small-scale subclonal losses of 1p and 5q. Significant large-scale clonal losses were found on 3p, 6p, 6q, 8p, 9p, 9q and 14q. Significant large-scale subclonal losses were found on 4q, 6p, 6q, 8p, 9p, 9q, 10p, 10q, 14q, 17p, 18q. Significant peaks of MSAI could be found on 8p and 14q, these events coincided with parallel losses suggesting there might be sufficient selective pressure not only to result in subclonal events being observed across cancer types but also multiple loss events within a single patient's disease.

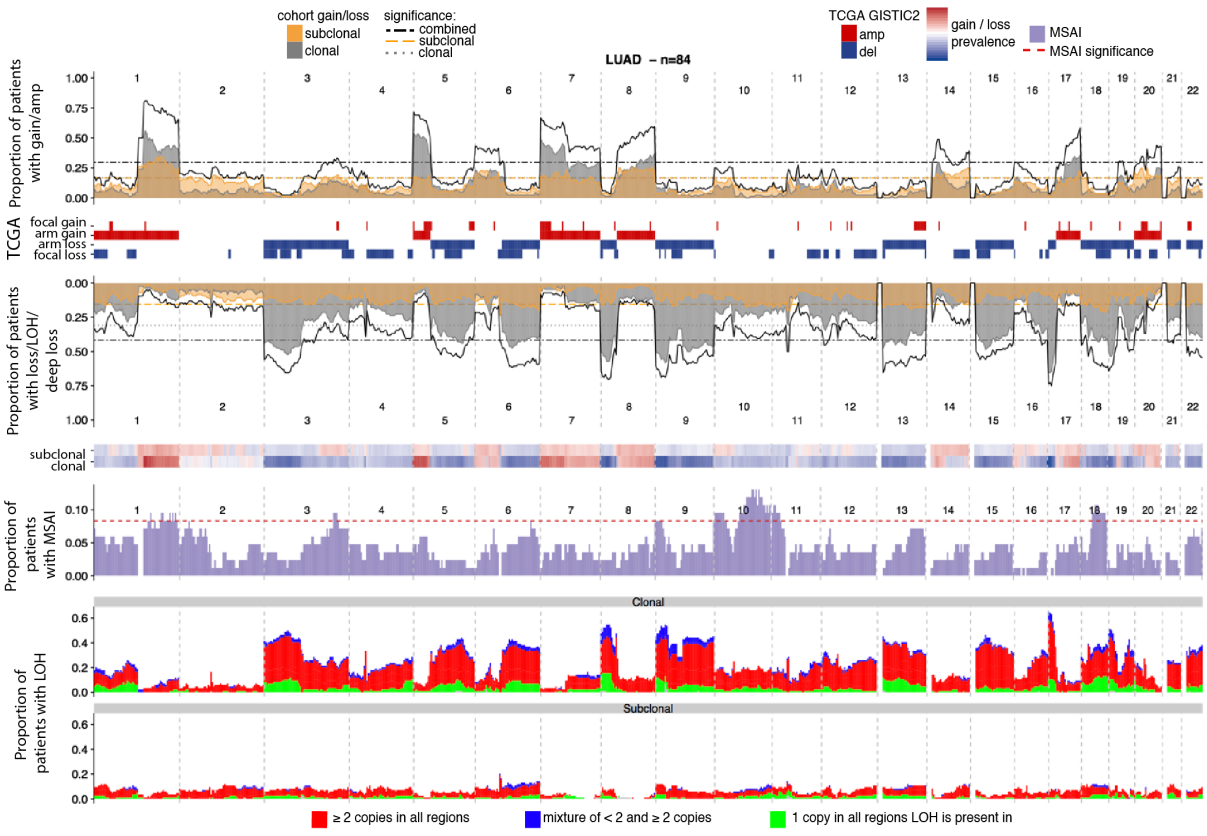

Figure 5-24: LUAD across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all

(clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The SCNA landscape of the 84 multi-region LUAD cases analysed in this chapter (Figure 5-24) is not dominated by any single event unlike many of the cancer types examined, the most common event observed is the significant large-scale loss of 17p that is present in close to 75% of the cohort though there are many other large-scale losses and gains occurring in greater than 50% of the cohort. Significant large-scale clonal gains observed include 1q, 5p, 7p, 7q, 8q, 17q and 20q with small-scale clonal gains seen on 14q and 19p. Large-scale subclonal gains were observed on 1q, 5p, 6p, 7p, 8q, 14q and 20p with small-scale subclonal gains found on 2q and 17q.

Large-scale clonal losses were observed across much of the genome including 3p, 5q, 6q, 8p, 9p, 9q, 13q, 15q, 16q, 17p, 18q and 19q with only a single small-scale clonal loss found on 12p. Large-scale subclonal losses were observed on 6q, 10q, 13q and 18q with small-scale subclonal losses on 21q and 22q.

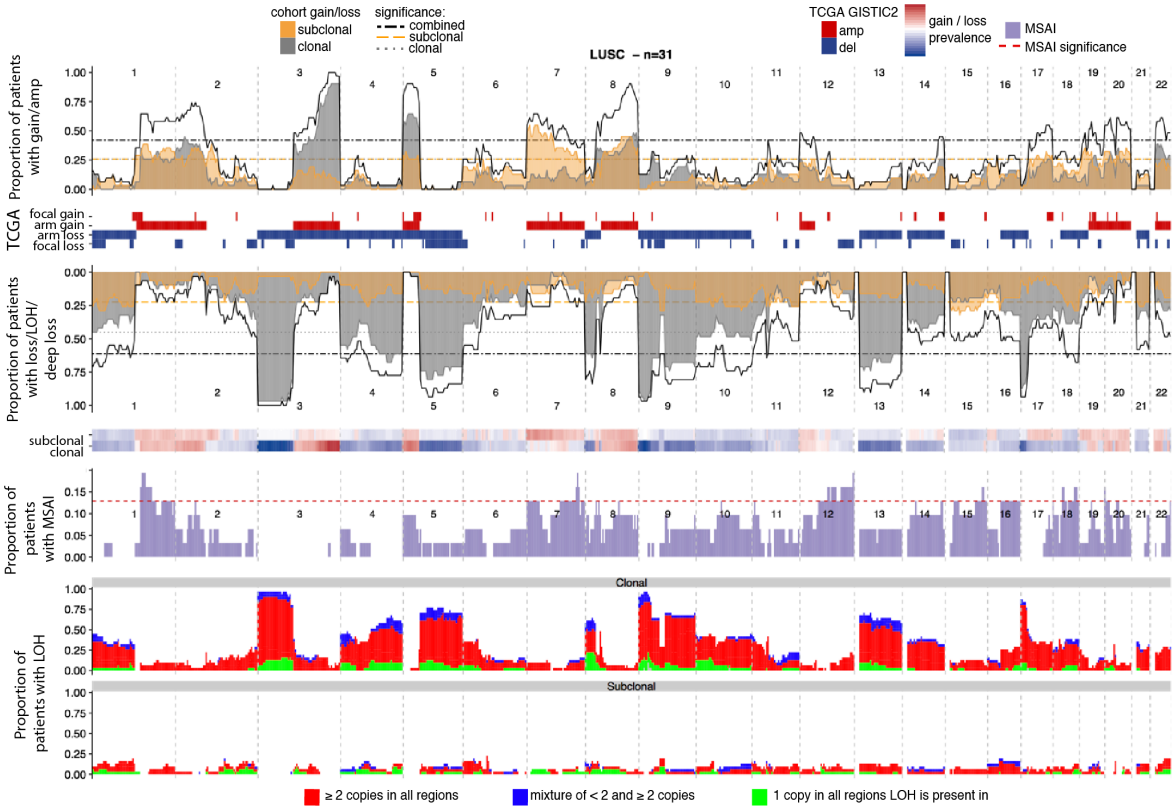

Figure 5-25: LUSC across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The SCNA landscape of the 31 LUSC cases analysed for this chapter (Figure 5-25) revealed four obligatory clonal (occurring in >75% of the cohort examined) events with large-scale losses affecting 3p, 9p and 17p and a large-scale clonal gain affecting 3q, the most of any cancer type examined. These obligatory loss events, given they are always accompanied by clonal LOH, are likely to have occurred before genome doubling. Other large-scale clonal gains were observed on 1q, 2p, 5p, 8q, 22q with small-scale clonal gains seen on 19q and 20q. Significant large-scale subclonal gains were present on 1q, 2p, 5p, 7p, 7q, 8q, 19p and small-scale gains on

17q and 19q. Interestingly, when combining both subclonal and clonal events affecting 5p and 8q, their combined totals are present in greater than 75% of cases suggesting an important role in this cancer type's evolution.

Significant large-scale clonal losses apart from the three obligatory clonal events of 3p, 9p and 17p include 4p, 4q, 5q, 8p, 9q, 10p, 10q and 13q with no small-scale clonal loss events detected. Large-scale subclonal losses included 1p, 4q, 9q, 10q, 15q, 18q and 21q with only a single subclonal loss affecting 11q.

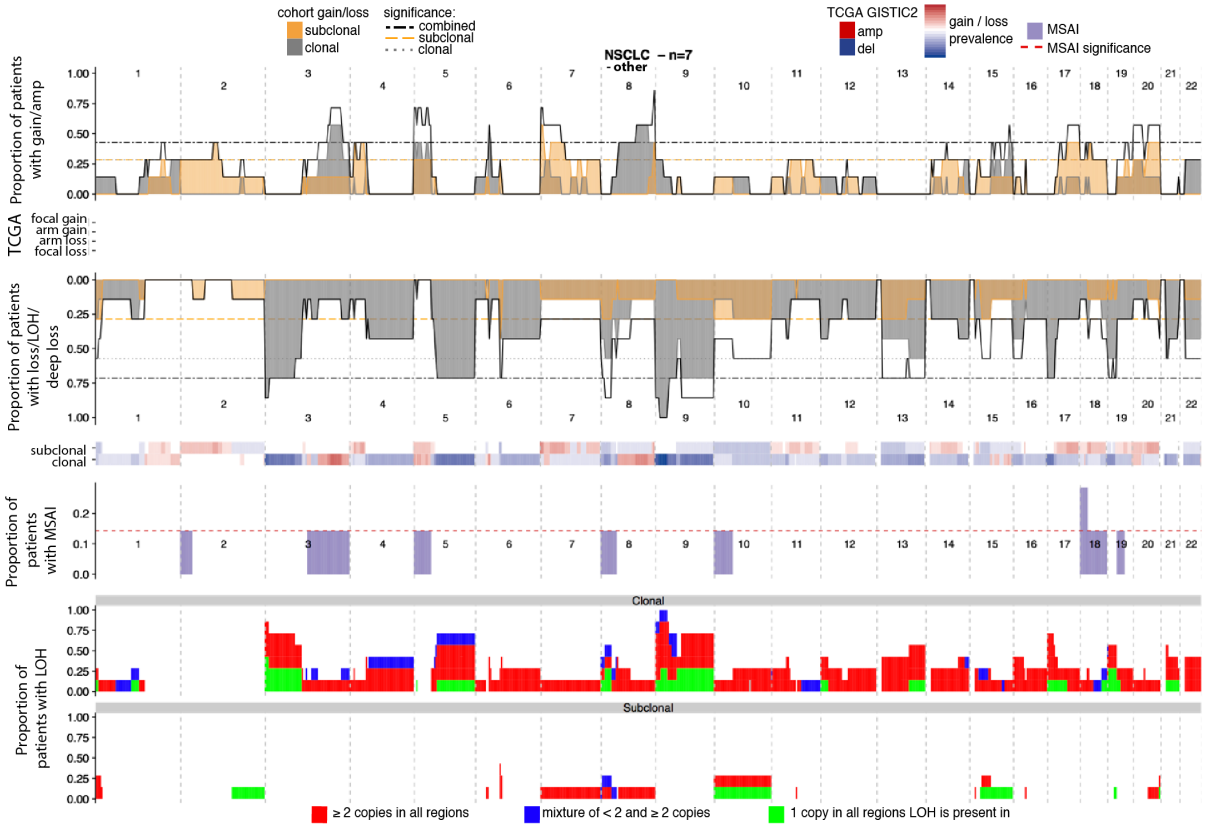

Figure 5-26: NSCLC-other across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with

significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The SCNA landscape of NSCLC-other (Figure 5-26) is quite similar to that of LUSC (Figure 5-25), sharing the same three large-scale clonal losses of: 3p, 9p and 17p. Significant large-scale clonal gains affected 3q, 8q and 15q with no small-scale significant losses found. Large-scale subclonal gains included 7p, 17q and 20q with small-scale subclonal gains affecting 2p, 4p and 18p.

Apart from the already mentioned obligatory clonal losses of 3p, 9p and 17p other large-scale clonal losses described included 5q and 9q with no small-scale clonal losses found, large-scale subclonal losses or small-scale subclonal losses described, perhaps as a function of the small number of cases present of this cancer type.

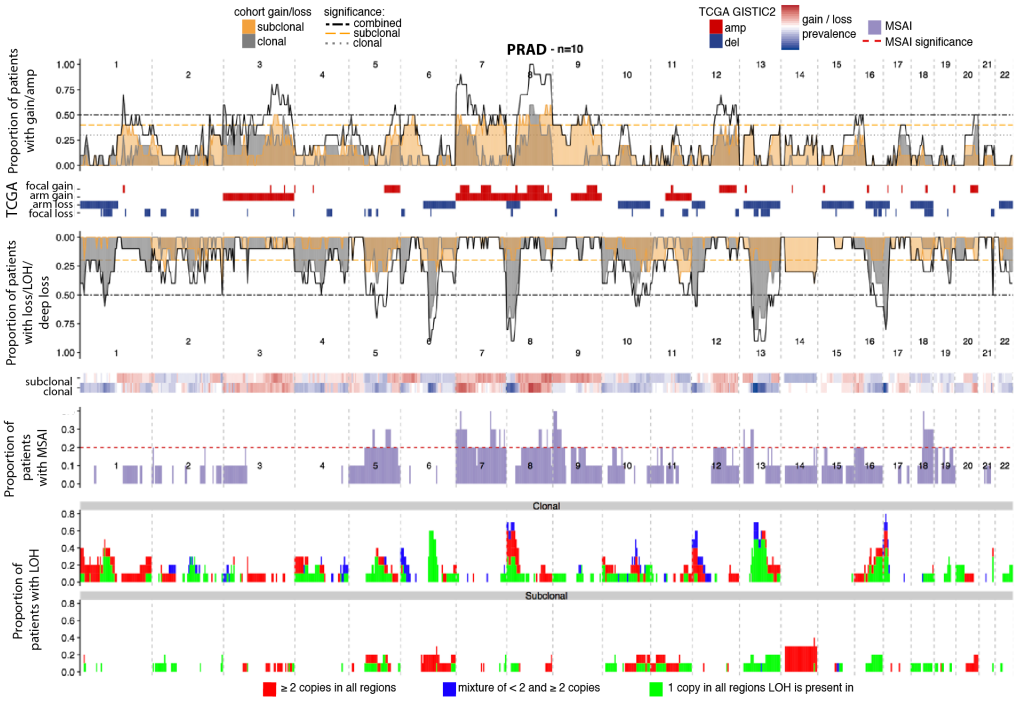

Figure 5-27: PRAD across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for

every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The SCNA landscape of PRAD (Figure 5-27) demonstrate large-scale clonal gains on 7p and 20q, small-scale clonal gains were found upon 3p and 8q. Significant large-scale subclonal gains include 8q and small-scale gains were found on 3q, 6p and 9q.

Large-scale clonal losses were observed on 8p, 13q and 17p with small-scale clonal losses seen on 1p, 6p, 10p, 12p. In terms of subclonal events, large-scale subclonal losses were found on 11q and 14q and small-scale losses observed on 5q, 2p and 10q.

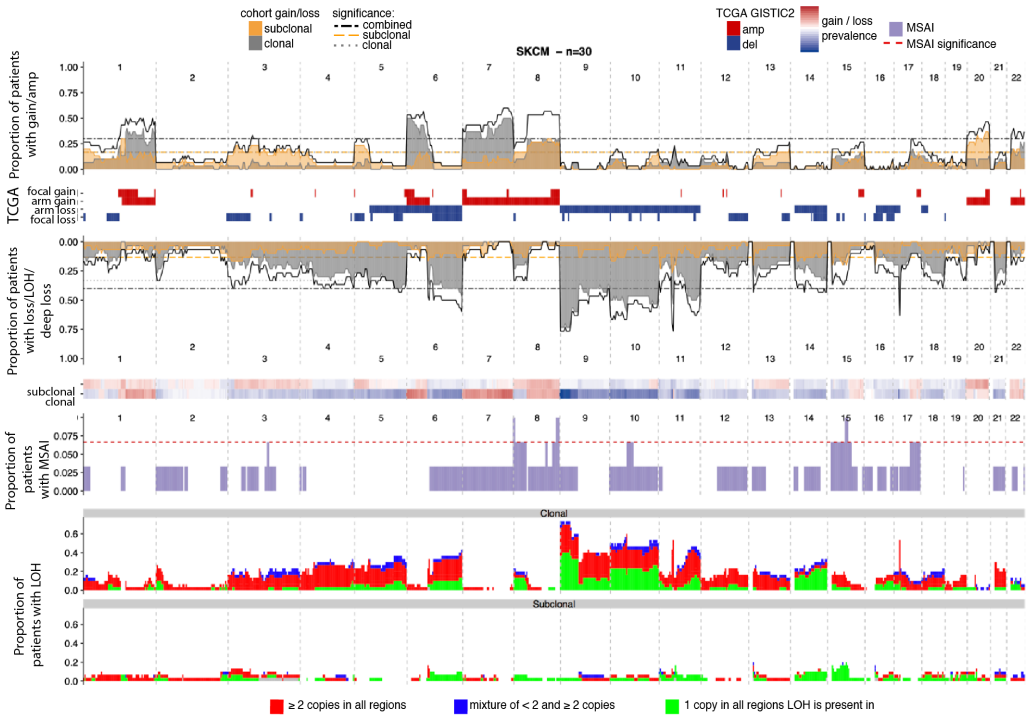

Figure 5-28: SKCM across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all (clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for

every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.
.

The 30 multi-region SKCM cases analysed in this chapter (Figure 5-28) demonstrated significant large-scale clonal gains on 1q, 6p, 7p, 7q and 8q with no small-scale clonal gains observed. Much like many of the cancer types analysed in this study SKCM demonstrated a large-scale subclonal gains of 5p and 8q. Small-scale gains of 3p and 3q were also observed.

Large-scale losses observed were on 6q, 9p, 9q, 10p, 10q and 11q with a single small-scale clonal loss on 5q. Large-scale subclonal losses were found on 3p and 19p and small-scale subclonal losses on 11p, 14q and 15q.

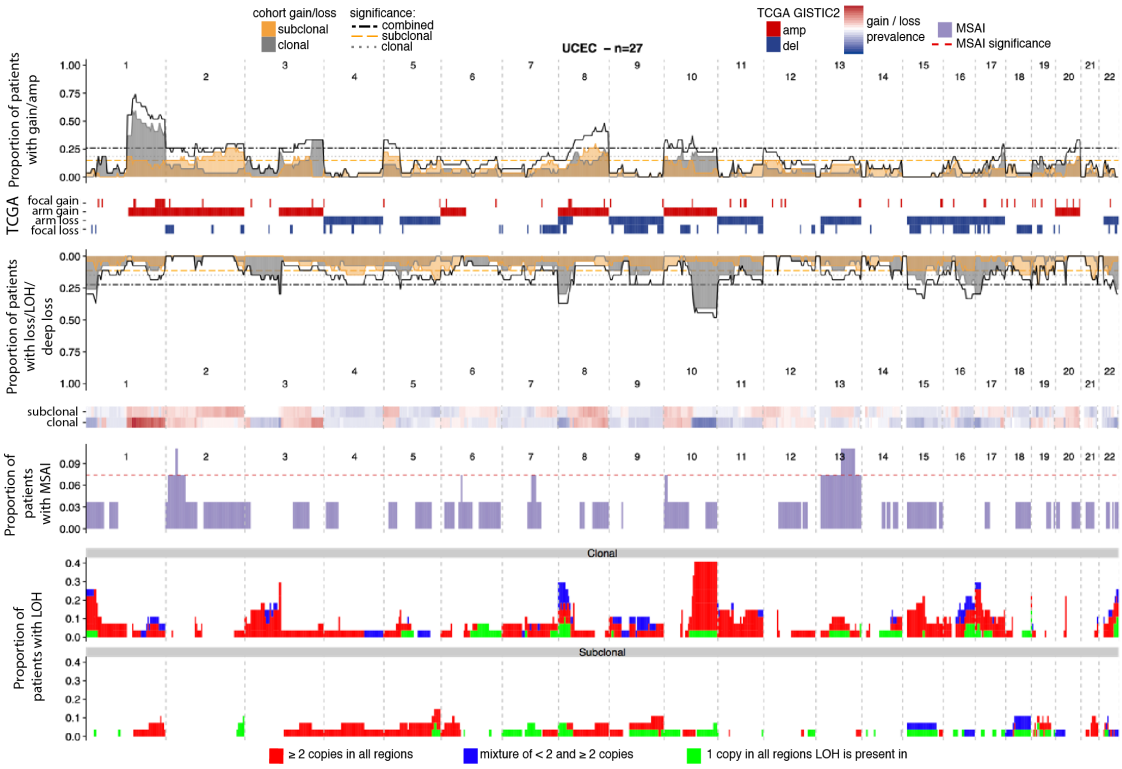

Figure 5-29: UCEC across-genome plot showing the SCNA frequency and clonality
The first track shows the frequency of amplification/gain events pointing upwards, and the frequency of loss/LOH/deep loss events pointing downwards, the most common event type present at each cytoband is plotted (it should be noted that LOH at the same position as a gain in a tumour region will count towards both the loss and gain totals). The solid grey line and grey filled bars indicate the frequency of clonal events, the solid yellow line and yellow filled bars indicate the frequency of subclonal events, while the solid black line indicates the frequency of all

(clonal and subclonal) events The dotted grey, yellow and black lines represent significance thresholds for clonal, subclonal and all events, respectively -- bars and lines that are found to exceed these lines represent cytobands that harbour significantly recurrent SCNA events. The red/blue tracks in between the bar and line plots indicate amplifications (amp) and deletions (del) that were previously reported to be significant in GISTIC2 analyses of the TCGA cohort for that cancer type. In the next panel down there are two across genome heatmap tracks that for every cytoband demonstrate the most common event type present in the cohort examined. I.e. if gain relative to ploidy was the most common event at a cytoband for a cancer type this is plotted in red with the intensity of colour reflecting the proportion of the cohort from this cancer type that contains a gain event while for losses the colour is blue. The next panel shows the proportion of the cancer type cohort affected by MSAI events at each cytoband with significance threshold indicated by a dotted grey line. The final two tracks show the proportion of the cancer type cohort affected by clonal and subclonal LOH respectively. The colour indicates whether the LOH is single copy in all regions (very likely a copy number loss relative to ploidy) in green, 2 or more copies in all regions in red (very likely a copy neutral event relative to ploidy) and blue if a mixture between < 2 copies in at least one region as well as ≥ 2 in least one other region.

The 27 cases of UCEC analysed (Figure 5-29) demonstrated large-scale clonal gains on 1q and 10q and small-scale clonal gains on 3q, 8q and 17q with the most prevalent event in being the 1q clonal gain occurring approximately half the cohort when clonal and subclonal gains are combined. Large-scale subclonal gains were found on 2p, 5p and 8q with small-scale subclonal gains found on 10p and 20p.

Significant large-scale clonal losses were found on 1p, 3p, 8p, 10q, 17p and 22q with small-scale clonal losses on 15q, 19p, 19q and 21q. Large-scale subclonal losses were found on 4q and 18q with small-scale subclonal losses on 5q and 7q.

5.5 Conclusions

Previous studies have used single region SNP array and sequencing approaches to quantify aneuploidy through total SCNA burden. The multi-region data and novel approach of utilising multi-region phasing I have implemented has, for the first time, described SCNA heterogeneity across a pan-cancer cohort in a manner that is able to detect MSAI and SCNAs that demonstrate an illusion of clonality in single regions. This study revealed pervasive SCNA heterogeneity across cancer types, strongly suggesting a role for ongoing CIN throughout cancer evolution regardless of cancer type. This argues against the acquisition of SCNAs through early punctuated bursts that some have proposed from SCS (Gao et al., 2016). Interestingly, this single cell study was performed in BRCA TN with technology unable to detect LOH or AI. In the cohort examined in this chapter BRCA TN was one tumour type that demonstrated relatively high levels of AI and LOH. In addition, BRCA TN also demonstrated MSAI and other forms of subclonal AI. For these reasons, it is possible that allele insensitive techniques such as SCS might underestimate the SCNA burden and heterogeneity present in this cancer type.

Though pervasive SCNA heterogeneity was observed across cancer types, each cancer-type demonstrated evidence of order in copy number evolution over time. Furthermore, some events were found to be common across cancer types such as clonal 3p loss in KIRC, NSCLC (both LUSC and LUAD) and breast cancer subtypes. In addition, some significantly recurrent clonal loss events occurred at very high prevalence within cancer types, defined as obligatory early events, occurring in at least 75% of cases within a cancer type. These obligatory clonal loss events, given their co-occurrence with LOH present in all regions of each patient's tumour, are likely to have occurred before WGD. WGD was associated with increased clonal and subclonal SCNA burden including MSAI, an observation validated by analysis of previous experimental work on genome-doubled and non-genome doubled sister colorectal cell line clones.

The work presented in this chapter suggests that cancer evolution is characterized by ongoing CIN but early clonal LOH events, specific to cancer type, which often occur before WGD and that WGD events are subsequently permissive for an increased level of CIN, aneuploidy and SCNA heterogeneity.

Chapter 6 Positive selection influences the subclonal SCNA landscape

6.1 Introduction

Multi-region sequencing strategies may also shed light on debates regarding the presence of positive selection, mode of tumour evolution and the extent to which order exists during SCNA evolution (Sun et al., 2017). For instance, whether cancer development is predominantly characterized by early punctuated macro-evolutionary change followed by copy number-stasis (Gao et al., 2016), or continual karyotypic change is unclear and requires an exploration of both the state and rate of these karyotypic changes. Likewise, echoing debates in evolutionary theory, recent work has questioned whether subclonal evolutionary dynamics are best characterized by neutral growth or positive selection (Williams et al., 2016, Williams et al., 2018, Tarabichi et al., 2018, Balaparya and De, 2018, Heide et al., 2018, Cannataro and Townsend, 2018, Niida et al., 2018). Parallel evolutionary events, converging upon the same biological endpoint (Jamal-Hanjani et al., 2017, Turajlic et al., 2018b), or subclonal selective sweeps, often resulting in an illusion of clonality, where subclonal events appear clonal within an individual tumour region (Sun et al., 2017, de Bruin et al., 2014), are strong indicators of such positive selection. However the detection of such events from single samples remains challenging. As such, there is a need for large-scale, pan-cancer multi-regional analyses to systematically assess the extent, timing and importance of CIN and ongoing SCNA heterogeneity and selection across different cancer types.

To address this, I assembled 1421 multi-region tumour samples from 394 patients across 22 tumour types profiled by WGS, WES or SNP arrays. Using novel computational approaches for long-range phasing of germline variants and somatic evolutionary events, I reveal the extent of subclonal recurrent events in these tumours and detect parallel evolutionary events at an unprecedented scale and prevalence across tumour types.

Furthermore, mathematical modelling shown in Figure 6-9 of subclonal SCNA evolution performed by my collaborators Dr. Sergi Elizalde and Dr. Samuel Bakhoum, shows that a model incorporating ongoing selection with respect to SCNAs significantly outperforms evolutionary neutral models, particularly in the context of WGD when measured against the observed data in this cohort. In addition, while I

performed the analyses to produce the data shown in Figures 6-1 and 6-2 the figures themselves were generated by Dr Emilia Lim and the SCNA based MEDICC tree building shown in Figure 6-5 was performed by Marina Petkovic, though the figure design was mine. RNAseq processing was performed by Dr Emilia Lim and Dr Rachel Rosenthal.

6.2 Recurrent subclonal copy number events across tumour types

Subclonal recurrent SCNAs were identified in every tumour type (with ≥ 10 cases examined) and revealed both common evolutionary trajectories later in tumour evolution both between tumour types and within tumour subtypes (Figure 6-1). In LUSC, gains to 7p (encompassing *EGFR*) were more than three times more likely to occur at a subclonal rather than clonal level (13% clonal vs. 42% subclonal), while in LUAD gains to 7p were only significantly recurrent at the clonal level (51% clonal vs. 12% subclonal). Across the cohort, 7p gain was identified as a significantly recurrent clonal gain event in BLCA, BRCA ER+, ESCA, LUAD and SKCM, while it was identified as significantly recurrent subclonal gain in COAD, LGG/GBM, KIRC, LUSC, and PRAD. Similarly, gains to 7q (including *CDK6*) were identified as significantly recurrent subclonal events in BRCA TN, COAD, KIRC, LUSC, ESCA and as significantly recurrent clonal events in BLCA, ESCA, LGG/GBM, LUSC and SKCM. With regards to significantly recurrent losses in LUSC, only chromosome 15 exhibited a greater preponderance for subclonal compared to clonal SCNAs. In this case, given the large size of the genomic segments subject to loss it is difficult to identify whether specific genes, such as B2M and BUB1B were targeted. In BRCA HER2+, loss events on 10q (including *ARID5B*, *PTEN* and *TCF7L2*), 18q (including *SMAD4*) and 19q (including *ARHGAP35*, *CIC*, *KMT2B*) showed significant recurrence on the

subclonal, but not the clonal level.

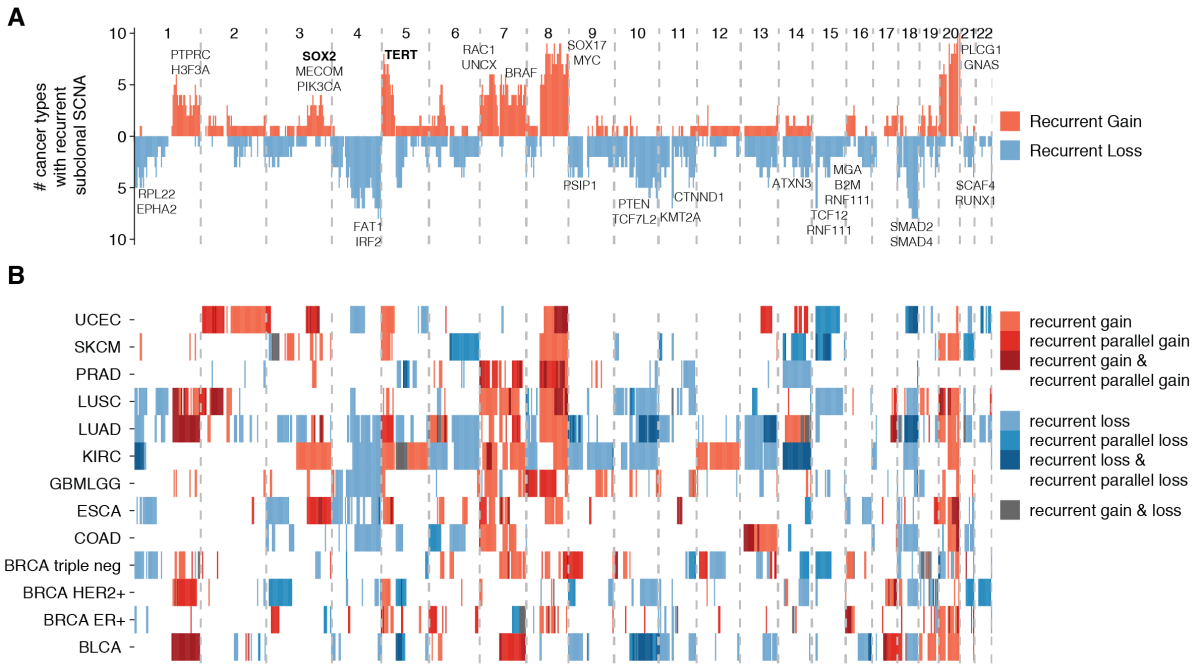

Figure 6-1: Recurrent subclonal SCNAs across tumour types
A) Across-genome plot indicating the number of tumour types that have a recurrent subclonal SCNA peak of either gain (red, facing up), loss (blue, facing down) affecting a cytoband. Only the 10 tumour types with recurrent subclonal and parallel events are included B) Heat map indicating cytobands that have recurrent subclonal or parallel events. Recurrent gains are indicated in shades of red, while recurrent losses are indicated in shades of blue.

Overall, the most common subclonal gain events found in the pan-cancer cohort were 5p (*TERT*, significantly recurrent in 7 tumour types), 8q (*MYC*, significantly recurrent in 8 tumour types), 20p (TASP1, significantly recurrent in 5 tumour types) and 20q (*GNAS*, *PLCG1*, significantly recurrent in 11 tumour types) (Figure 6-1). The most common subclonal loss events were found on 1p (*ARID1A*, *EPHA2*, *FUBP1*, *JAK1*, *THRAP3*, significantly recurrent in 6 tumour types), 10q (*GATA3*, significantly recurrent in LUSC, SKCM), 11p (*CARS*, significantly recurrent in BLCA, COAD, LUAD and LUSC) and 18q (*SMAD4*, significantly recurrent in 8 tumour types) (Figure 6-1). Notably, these events were significantly enriched in previously defined GISTIC regions (Figure 6-2), emphasising their potential functional relevance.

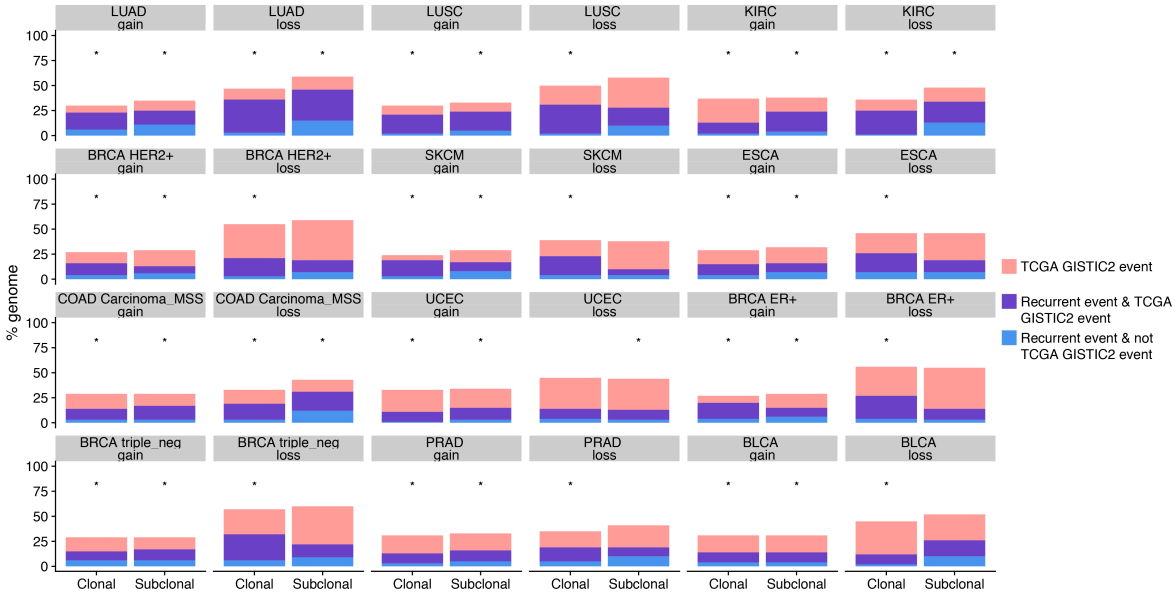

Figure 6-2: Overlap between TCGA GISTIC2 analyses and recurrent SCNAs
Overlap between TCGA GISTIC2 analyses (arm and focal combined) and recurrent clonal and subclonal events identified in the pan-cancer multi-region cohort by tumour type. Significant overlap measured by Fisher's exact test is indicated by an asterisk.

Another previously described SCNA event in cancer involves losing the ability to present neoantigens through human leukocyte antigen (HLA) LOH, which may facilitate immune evasion, and has recently been shown to occur as a predominantly subclonal event in early stage non-small cell lung cancers. Here we confirm these results for NSCLC but also demonstrate that it is a predominantly subclonal event in all tumour types examined, including KIRC, BRCA, BLCA, NSCLC-other, UCEC, and ESCA (Figure 6-3).

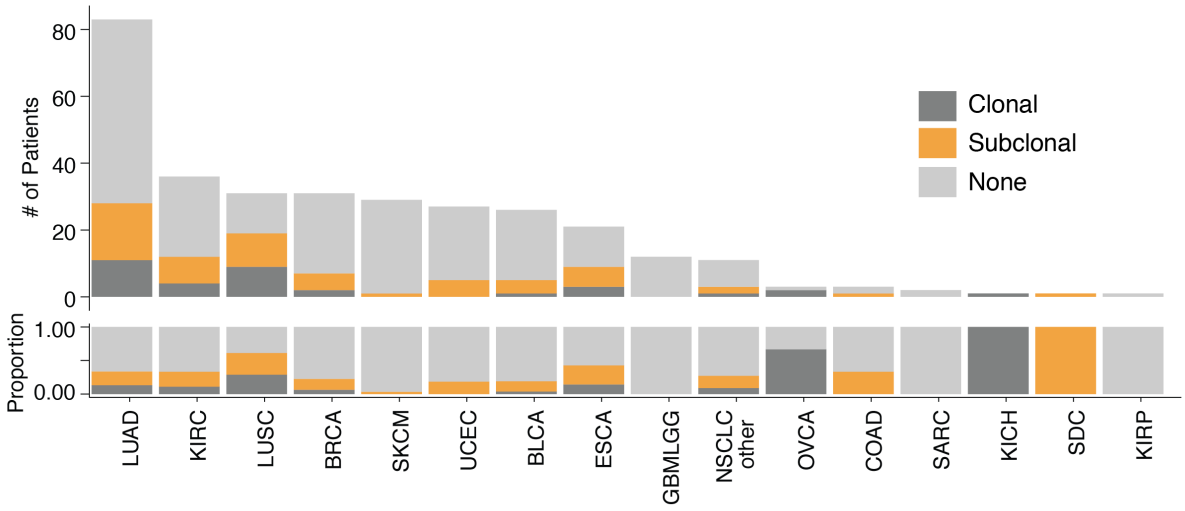

Figure 6-3: HLA LOH examined in WES cases
Histogram depicting the clonality HLA LOH detected in WES cases examined across tumour types.

Interestingly, in SKCM, a tumor type characterised by a high mutational burden (Vogelstein et al., 2013), the lowest rate of HLA-LOH of all tumour types examined was observed. Indeed, the HLA locus on 6p was encompassed by the most common gain event in SKCM, harbouring the putative oncogene *NEDD9* (Kim et al., 2006)(Figure 5-30), suggesting profound constraints to HLA copy loss in this tumour type and that other mechanisms of immune evasion may be more prevalent.

To further evaluate whether these significantly recurrent subclonal gain and loss events were potentially consistent with positive selection, I examined whether any of the affected regions showed signs of parallel evolution. Parallel evolution is characterised by multiple distinct subclones in the same tumour showing independent SCNA disruption encompassing the same gene. Parallel evolution of SCNAs is traditionally difficult to detect, but can be revealed using the multi-region phasing technique developed in this thesis, if two SCNA events affecting the same region occur on different parental haplotypes in different subclones (Jamal-Hanjani et al.), chapter 3). Because these events require both haplotypes to be affected independently, they likely only reflect a fraction of parallel evolutionary SCNA events in any given tumour since we cannot detect parallel events involving the same alleles. Despite this, we observed that the majority of detected parallel events co-occur with the most common subclonal gain events, supporting evidence for selection in these subclonal dynamics (Figure 6-4).

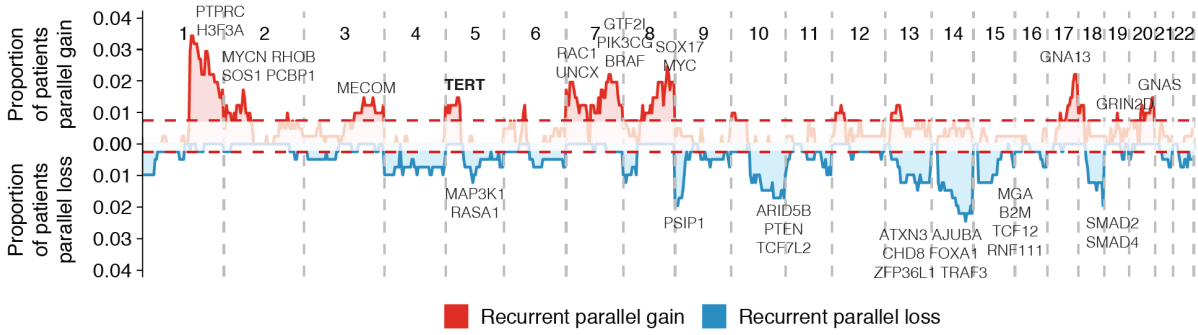

Figure 6-4: Recurrent parallel gains and parallel losses across cancer types
A) Across-genome plot indicating the proportion of patients that have a significant parallel SCNA peak of either gain (red, facing up), loss (blue, facing down) affecting a cytoband. The dashed line indicates the significance threshold.

The most prevalent parallel event across the cohort, with 24 parallel events observed (12 cases: 5 LUAD, 3 BLCA, 2 LUSC, 1 KIRC and 1 BRCA HER2+) extended from 1q23.1 to 1q44 including *ARNT* (also known as HIF1-Beta). The *ARNT* locus has previously been described as a significantly recurrent focal gain event by TCGA GISTIC analyses in BLCA, LUSC and LUAD (Figure 6-1). Another parallel gain locus that matches significantly recurrent subclonal gains found across tumour types harbours *TERT* (Figure 6-4, 6-5) on 5p. Contributing tumour types include LUAD and LUSC, in which it is also a known GISTIC arm event. In total, 12 parallel gain events involving 5p were observed (6 cases: 3 LUAD, 1 LUSC, 1 COAD, 1 BRCA ER+) (Figure 6-4).

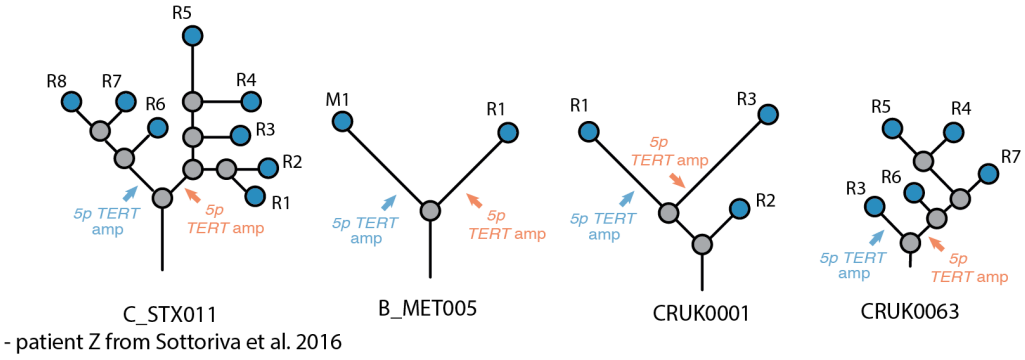

Figure 6-5: Parallel gain *TERT* events map to different branches of copy number phylogenies
MEDICC trees derived from SCNA profiles highlighting 4 of the 6 examples of parallel MSAI 5p (*TERT*) gain/amplifications. Blue nodes represent terminal codes of the trees; these represent the copy number profiles of each region that we have obtained. The blue text and arrows represent one allele; while the orange text and arrows represent the other allele. Both alleles have been impacted by independent 5p amplifications.

The most prevalent significantly recurrent parallel loss/LOH event was found on chromosome 14, with 14 parallel events found (3 KIRC, 1 BRCA, 1 LUSC, 1 ESCA, and 1 SKCM cases) (Figure 6-4). Loss of chromosome 14 harboring tumour suppressor HIF1A has been previously identified as having clinically prognostic value in KIRC (Monzon et al., 2011). It results in a decrease in expression of *HIF1A* (Shen et al., 2011) which is part of the hypoxia-inducible pathway perturbed by the pathognomonic alterations in *VHL* observed in KIRC. Therefore, parallel evolution for alterations in gene dosage later in tumour evolution at both *HIF1A* and *HIF1B* loci were identified across tumour types.

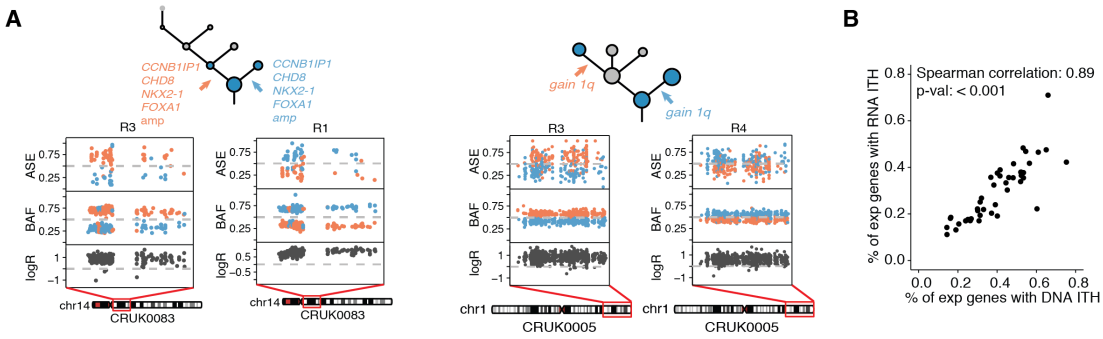

Figure 6-6: Allele-specific expression reflects MSAI events
A) Shows the log ratio, B-allele frequencies and allele specific expression of an MSAI parallel amplification observed in patients CRUK0083 and CRUK0005, whose SNV based subclone tree from (Jamal-Hanjani et al., 2017)is also displayed with annotation of where the independent MSAI events occur highlighted with orange and blue arrows. B) Scatter plot showing that the percentage of expressed genes with allele-specific DNA ITH correlates with the percentage of expressed genes with allele-specific RNA ITH.

We next examined whether parallel events lead to similar gene expression dosage effects from different homologous chromosomes. For this, we used available in-house matched RNA-seq data from 48 tumour regions from the TRACERx-100 cohort (Rosenthal et al., Under Review)) and quantified allele-specific expression (ASE) at heterozygous SNPs (see Chapter 2). Across these 48 regions, ASE profiles tracked genetic MSAI dosage with a correlation of 0.89 (Figure 6-6B, Spearman correlation, p-val < 0.001). For example, in LUSC patient CRUK0083, we previously identified parallel amplification of *NKX2-1*, *CHD8* and *FOXA1* (Jamal-Hanjani et al., 2017). The ASE profiles clearly mirror the changing amplification state of the haplotypes between subclones (Figure 3D), further supporting the potential phenotypic relevance of this

MSAI event. Consistently, in LUAD patient CRUK0005, a previously identified 1q MSAI gain was also reflected in the ASE profiles (Figure 6-6A).

6.3 Gene enrichment and modelling approaches

While evidence of recurrence and parallel evolution are consistent with positive selection, non-random mis-segregation, negative and neutral evolution may also play a role in shaping the subclonal SCNA landscape. To evaluate the extent to which positive selection shapes the subclonal SCNA landscape, I sought to address whether SCNAs characterised by either significant recurrence or parallel evolution were enriched for cancer genes identified through orthogonal mutation-based analysis. Intriguingly, while a significant enrichment for recurrent subclonal losses to harbour TSGs identified through mutation-based analysis in 5/13 tumour types ($P<0.01$) was observed, I did not find any significant enrichment for subclonal recurrent gains to harbour oncogenes. However, when significant SCNA events identified through parallel evolution analysis were considered, I observed a significant enrichment for oncogenes in gains in 5/13 tumour types, and TSGs in losses for 4/13 tumour types (Figure 6-7A). These data are consistent with parallel evolution providing greater resolution to detect positive selection and either suggest subclonal recurrent gain events do not necessarily reflect positive selection or, alternatively, oncogenes subject to subclonal gains may be distinct from those oncogenes subject to statistical mutation significance. Indeed, the latter may be the case for oncogenes *TERT* and *MYC*. These data suggest that the observed parallel evolutionary events are not merely a result of the increased frequency of the underlying subclonal SCNAs, but that instead selection is driving the increased enrichment of functional genes in these regions.

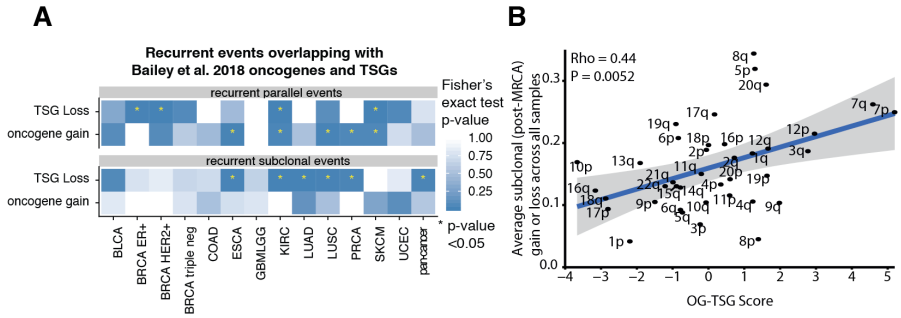

Figure 6-7: Gene enrichment evidence of subclonal SCNAs being subject to selection
A) A summary plot showing for each histology group the proportion of cases with recurrent subclonal and parallel events. A yellow asterisk indicates significance. B) Scatter plot showing the positive correlation between TSG-Oncogene score and average arm-level subclonal gain or loss occurring in each tumour's evolution since the MRCA. C) Beeswarm plot showing, for each patient, the deviance of the non-neutral model from the neutral (Control) state. Majority of deviance scores were found below 1, indicating that the non-neutral model tended to outperform the neutral model in all three scenarios (GD, nGD and subclonal GD); significance tested using chi-squared test.

Next, I sought to determine whether, at the cohort level, subclonal chromosome and arm-level copy number changes are likely to be selected for during tumour evolution. Extending on the analysis of the relationship between TSG-OG score (Davoli et al., 2013) and clonal events in the previous chapter (Figure 5-9A). I compared the net changes in the subclonal copy number of chromosome arms and their TSG-OG score. A significant positive correlation was observed, suggesting ongoing optimization of subclonal fitness through CIN (Figure 6-7B). Notably, when the relationship between average subclonal chromosome copy number changes and chromosome size (instead of the TSG-OG score) was investigated, no significant relationship between chromosome size and subclonal copy number changes were observed (Figure 6-8). Thus, the global landscape of SCNAs across tumours likely reflects the impact of positive selection operating both before and after the MRCA, clonally and subclonally.

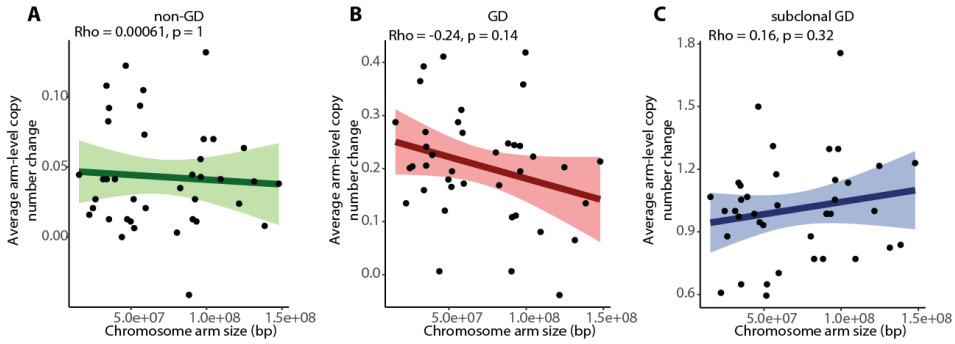

Figure 6-8: Impact of chromosome size on average arm-level copy number change

Scatter plots showing the lack of strong correlation between chromosome size and average arm-level subclonal gain or loss occurring in each tumour's evolution since the MRCA in A) nGD cases, B) GD cases and C) subclonal GD cases.

To explore positive selection in more depth within individual tumours, I obtained an inferred copy number profile of each patient's most recent common ancestor (MRCA) using an updated version of the MEDICC copy number phylogeny construction tool that is able to take MSAI into account (Beerenwinkel et al., 2014). I then summarised the mean copy number of each chromosome arm into a single integer value for the MRCA and tumour regions used to infer it for each patient's disease. These MRCA and region arm copy number summaries were then used with a modified version of a previously published mathematical model previously developed that simulates karyotype evolution during clonal expansion (Laughney et al., 2015, Elizalde et al., 2018). The revised model probabilistically predicts population karyotypes over time, allowing for whole-chromosome missegregation as well as chromosome arm breakage-fusion cycles (see Methods). The modifications to and running of the model were performed by Dr Sergi Elizalde.

For each patient, the MRCA (containing only clonal SCNAs) was considered as the starting state and the predictions of the model were compared to the arm-level copy number summaries of individual tumour regions (that will contain both clonal and subclonal SCNAs). The degree of deviation in terms of chromosome arm copy number changes between prediction and the tumour regions is measured and defined as the "relative deviance score". To ask whether chromosome copy number alterations are subject to ongoing selection rather than a neutral evolution model, we compared the model's output when chromosome-specific TSG-OG scores were either included, excluded or randomly permuted. In the first scenario, arm-level copy number profiles with higher oncogenic propensity were favoured (selection model), whereas the second treated all chromosomes equally (neutral model) and the third randomly permuted the TSG-OG scores across the chromosome arms (scrambled model). In all groups of cases split by genome doubling status, the selection model outperformed both the neutral and scrambled model when compared with patient region data as evidenced by significantly reduced relative deviance scores. The weighted model outperformed the neutral or scrambled models in 64% of GD, 59% of subclonal GD and 54% non-GD tumours) (Figure 6-9).

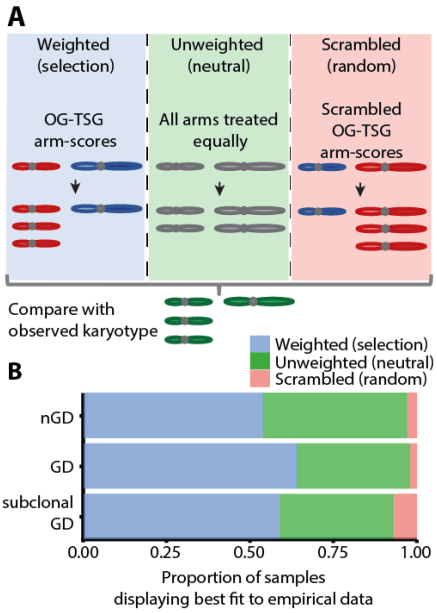

Figure 6-9: Neutral and non-neutral Markov chain modelling of chromosome missegregation
A)The three conditions under which karyotype evolution was modelled: chromosome arms with OG–TSG scores included (weighted model); chromosome arms were treated equally (neutral model); OG–TSG scores were randomly permuted (scrambled model). B) For each context (GD, n = 194 tumours; nGD, n = 171 tumours; and subclonal GD, n = 29 tumours), the percentage of tumours for which each model condition best recapitulates the empirically observed data is shown.

6.4 Late-emerging subclones frequently seed metastases

To determine whether the acquisition of SCNAs is associated with metastases, I considered the percentage of the genome affected by SCNAs in both primary and metastatic samples. Consistent with previous efforts (Bakhoum et al., 2018), I observed a greater percentage of the genome affected by SCNAs in metastatic tumour regions (n=178 patients) compared with primary tumour regions (n=366 patients) (Figure 6-10A, t-test, p-value = 5.3e-07). This trend was also observed when comparing matched primary and metastatic samples (n=152 patients; p-value=0.02246, Figure 6-10B). However, no significant increase in ploidy was observed between matched primary tumour and metastatic samples in the cohort as a whole, or in any individual tumour type., in contrast to previous findings (Priestley et al., 2018). This lack of increased ploidy between matched primary and metastasis samples may be due to the histology of the available metastatic samples, over a third of which are from LUAD and the various breast cancer subtypes, histologies with high rates of WGD in the primary tumour (Zack et al., 2013, Bielski et al., 2018) and which are associated with GD in our own analyses (Table 5-1). Overall, LOH displayed the greatest increase from primary to metastatic samples (p-value=0.014, test=t-test) (Figure 6-10C).

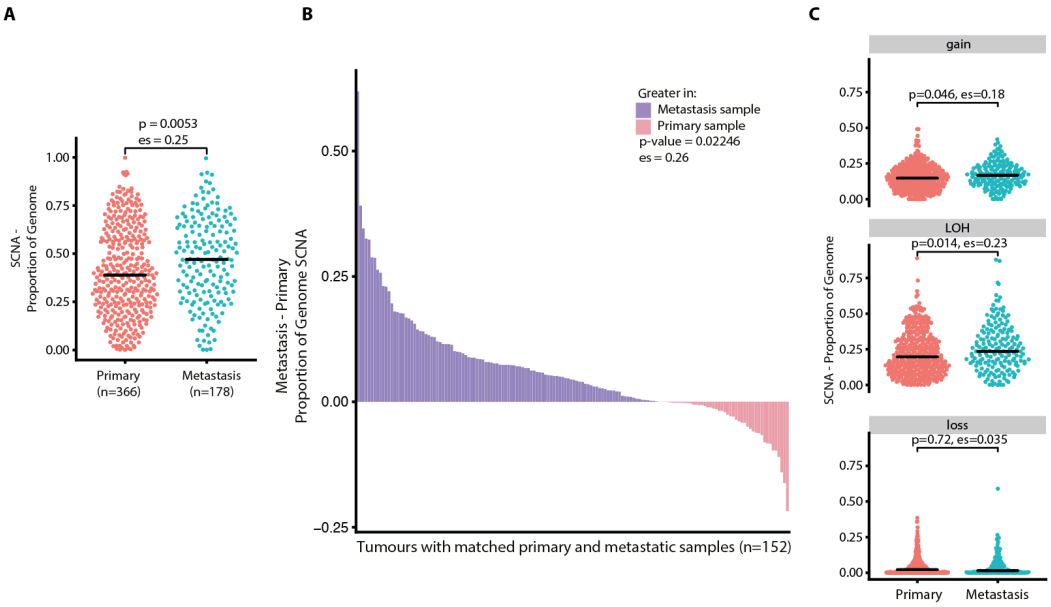

Figure 6-10: Increased SCNA burden in metastatic samples
A) Beeswarm plot demonstrating an increase in the percentage of the genome affected by SCNAs in metastatic samples when compared to primary samples. B) For each patient with matched primary and metastatic samples (n=152), the difference in the percentage of the genome affected by SCNAs between metastatic and primary samples is indicated. For a majority of patients, there is a greater percentage of the genome affected by SCNAs in metastatic samples when compared with primary samples (purple bars). C) Beeswarm plots showing the percentage of the genome affected by gain, LOH and losses in metastatic samples compared to primary samples. es, effect size.

To evaluate the importance of the recurrent clonal and subclonal SCNA events (Figure 6-11), I determined their relative prevalence in primary tumours compared to non-matched metastases. In this case, recurrent clonal SCNAs, defined from all tumour regions in the cohort, were found to be enriched in metastatic samples and may reflect SCNA events which generally occur early in tumour evolution, yet provide cancer cells with a predisposition towards metastasis. Conversely, recurrent subclonal SCNAs defined in the same way and enriched in metastatic samples may potentially reflect events occurring later which specifically enhance metastatic potential of specific subclones.

To quantify whether any specific subclonal recurrent SCNAs were enriched in metastasis compared to the primary tumour, an enrichment score at the cytoband-level was calculated, using the background rate of SCNA acquisition between primary and metastatic samples. In this case, we restricted our analysis to the two largest tumour types in our cohort (>50 tumour regions) LUAD and KIRC. Consistent with previous analyses of multi-region targeted NGS (Turajlic et al., 2018b) but in a new cohort, LOH of 9p21 (including *CDKN2A*) as a SCNA significantly associated with metastasis in KIRC was identified (60% of metastatic samples, compared to only 35%

of primary tumour regions). In LUAD, a number of SCNAs which were enriched in metastatic samples were identified, including gains of *TRAF3* and losses of *APC*, *RASA1*, *KMT2B*, *ARID1A*, *MTOR*, *RPL22*, *EPHA2*, and *CARS* (Figure 6-11A). Intriguingly, mutations in *ARID1A* have previously been linked to metastatic spread in one LUAD tumour (Abbosh et al., 2017), consistent with alterations to this gene potentially engendering increased metastatic potential.

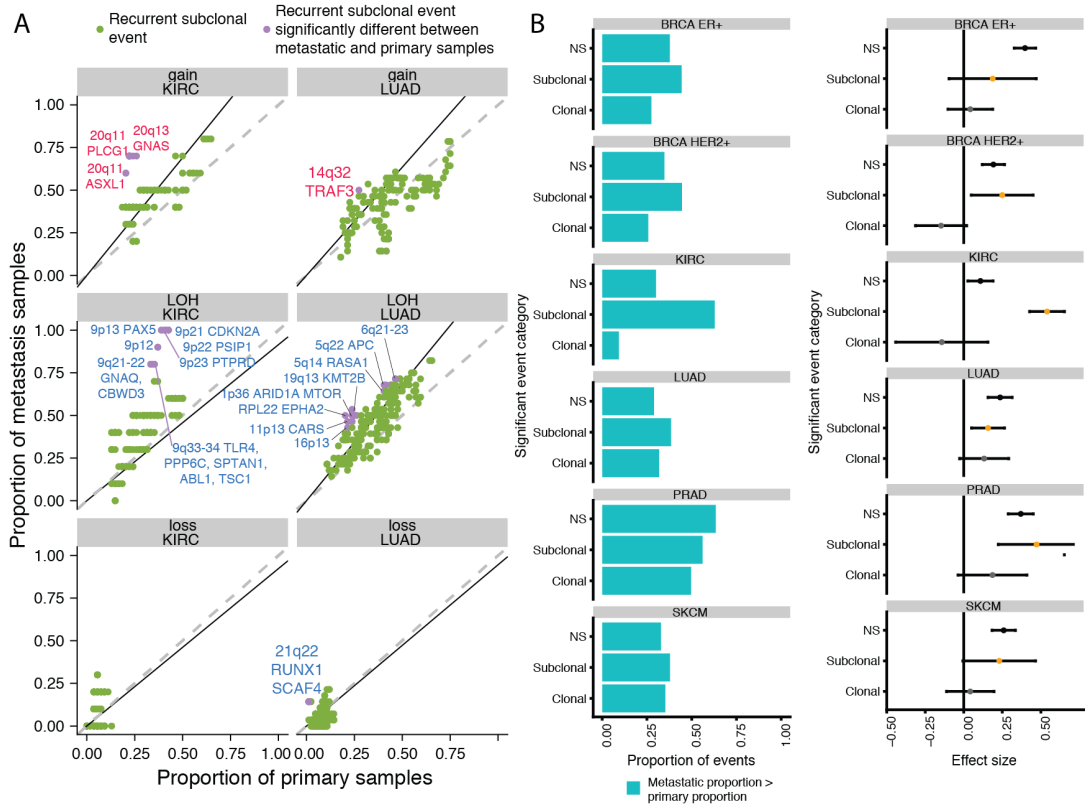

Figure 6-11: Enrichment of recurrent subclonal SCNA events in metastatic samples
A) Scatter plots comparing the prevalence of recurrent subclonal events in primary and metastatic samples. Cytobands with significantly different prevalence in primary and metastatic samples are labeled: gains in red and losses in blue. B)The proportion of different SCNA events (Significant Clonal, Significant Subclonal and Non-Significant (NS)) for each tumour type with ≥ 10 metastatic samples that are more commonly found in metastatic samples than primary samples. In the right panel, the effect size (odds ratio) of the different SCNA events being more prevalent in metastatic samples than primary samples is shown.

In 5/6 tumour types (with >= 5 metastatic cases) a significantly higher prevalence of subclonally recurrent SCNA events (considered in aggregate) was observed in in metastatic samples compared to non-matched primary tumours (Figure 6-11B), consistent with these events potentially being important in metastatic transition. Moreover, the recurrent subclonal SCNAs were generally enriched to a greater degree than recurrent clonal SCNA (6/6 tumour types) and non-significantly recurrent

events (5/6 tumour types) (Figure 6-11B). These data suggest tumours frequently acquire their metastatic potential later in tumour evolution.

To shed further light on patterns of metastatic spread and their clonal composition, intra-metastatic SCNA heterogeneity and intra-tumour SCNA heterogeneity in the primary tumour were compared, utilising tumours with multiple regions sampled from the metastatic site or multiple primary tumour regions. In our cohort of tumour types, SCNAs were found to be more clonal in metastatic samples than primary samples (Figure 6-12), consistent with an evolutionary bottleneck and a shared common ancestor between subclones identified at metastatic sites.

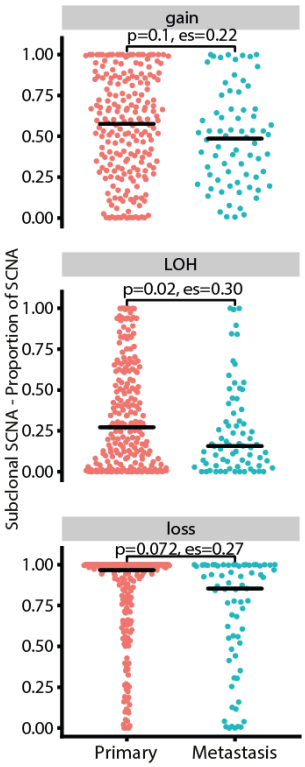

Figure 6-12: Clonality of SCNA categories in primary and metastatic samples
A) Heat map indicating where groups of recurrent clonal events significantly overlap with known TSG and oncogenes. B) Scatter plot depicting a positive correlation between TSG-Oncogene scores and average clonal copy number of each chromosomal arm. C) Heatmaps indicating obligatory clonal SCNA events for each histology type. Red shading indicates that a gain is predominant, while blue shading indicates that loss/LOH is predominant. A yellow star indicates that the event is significantly obligatory clonal in that histology type.

If metastatic subclones share a common ancestor, it follows that a minor subclone which seeds a metastatic site may be present within the matched primary tumour. To explore this possibility, I harnessed the multi-region haplotype phasing approach to identify minor subclones within individual primary tumour regions. For example, in one LUAD tumour, B_MET019 (Figure 6-13), in approximately 44% of cancer cells

sequenced in the primary tumour, there was evidence for an LOH event affecting 10q, which included cancer genes *PTEN*, *BUB3*, *SMC3* and *NFKB2*. However, in the matched tumour region from the brain metastasis, 10q LOH was observed in 100% of cancer cells, consistent with this minor subclone in the primary tumour seeding the metastatic site.

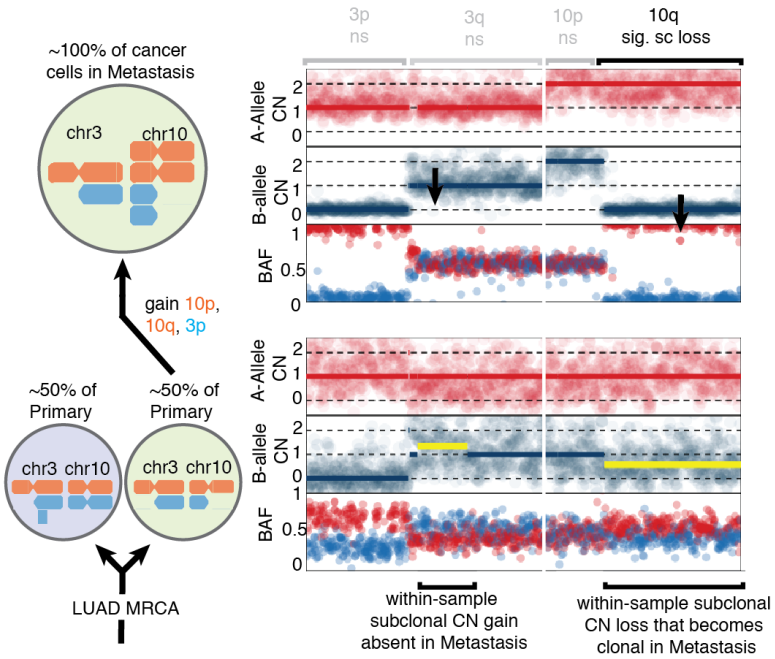

Figure 6-13: B_MET019 minor SCNA subclone in primary seeds a brain metastasis
Schematic on the left shows the likely copy number states of the two subclones (shown in blue and green) present in the primary sample and that the metastasis is likely seeded from the green subclone in the primary. The panel to the right shows the raw (continuous) values of the allele-specific copy number and BAF profiles for each a primary and metastasis region from the patient's disease. Continuous or "raw" allele specific copy number is shown as red and blue dots, with integer copy number estimates in solid lines of the same colour. Subclonal copy number estimates are shown with a yellow line when present. Events found as significant in recurrence analyses for the tumour type examined are highlighted in black at top of the copy number plot for the metastasis region with those that were found to be non-significant shown in grey.

Similarly, in a BRCA HER2+ case (B_MET066), in the primary tumour there was evidence for a within-sample subclonal LOH event affecting chromosome 9q (including cancer genes *TSC1*, *SMC2*, *SMC5* and *TRAF2*) that was clonal within the metastasis (Figure 6-13).

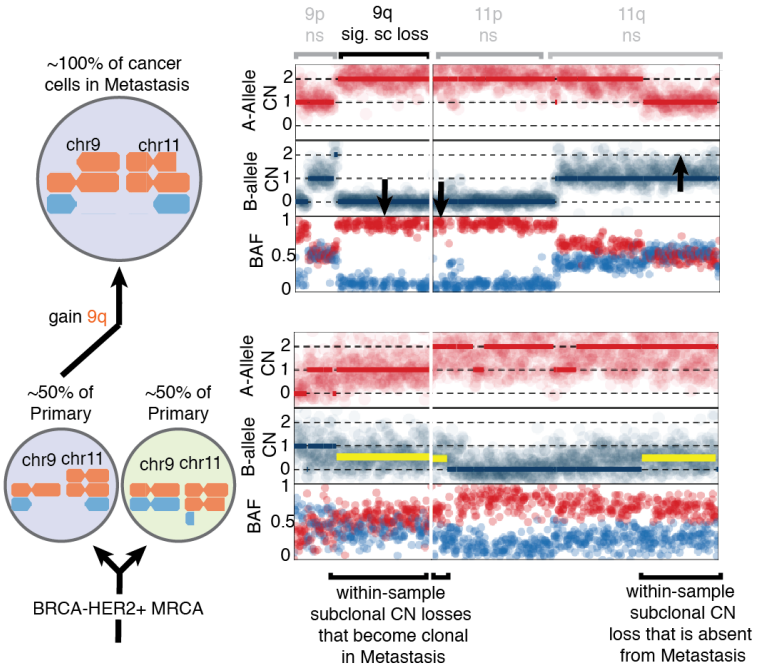

Figure 6-14: B_MET066 minor SCNA subclone in primary seeds a brain metastasis
Schematic on the left shows the likely copy number states of the two subclones (shown in blue and green) present in the primary sample and that the metastasis is likely seeded from the blue subclone in the primary. The panel to the right shows the raw (continuous) values of the allele-specific copy number and BAF profiles for each a primary and metastasis region from the patient's disease. Continuous or "raw" allele specific copy number is shown as red and blue dots, with integer copy number estimates in solid lines of the same colour. Subclonal copy number estimates are shown with a yellow line when present. Events found as significant in recurrence analyses for the tumour type examined are highlighted in black at top of the copy number plot for the metastasis region with those that were found to be non-significant shown in grey.

In another case (B_MET034, a salivary duct carcinoma) a WGD event that appeared clonal in all 3 metastatic regions but was absent in the primary tumour was observed. Interestingly only the metastatic WGD samples had a *CCNE1* amplification which appeared completely absent from the primary, an event significantly associated with WGD in our cohort and others (Bielski et al., 2018).

We also observed evidence for a parallel evolutionary trajectory between the primary and metastatic tumour regions. For example, in the primary tumour region of KIRC - B_MET080 - there was evidence for mono-allelic loss on chromosome 14q containing *HIF1A*, a significantly recurrent SCNA identified through our subclonal analysis and a known risk event in KIRC (Monzon et al., 2011, Shen et al., 2011) (Figure 6-15). In the metastatic sample, however, LOH at 14q was also observed but affected the opposite parental haplotype (MSAI event), indicating parallel evolution converging upon LOH at 14q through independent events in primary and metastatic subclones.

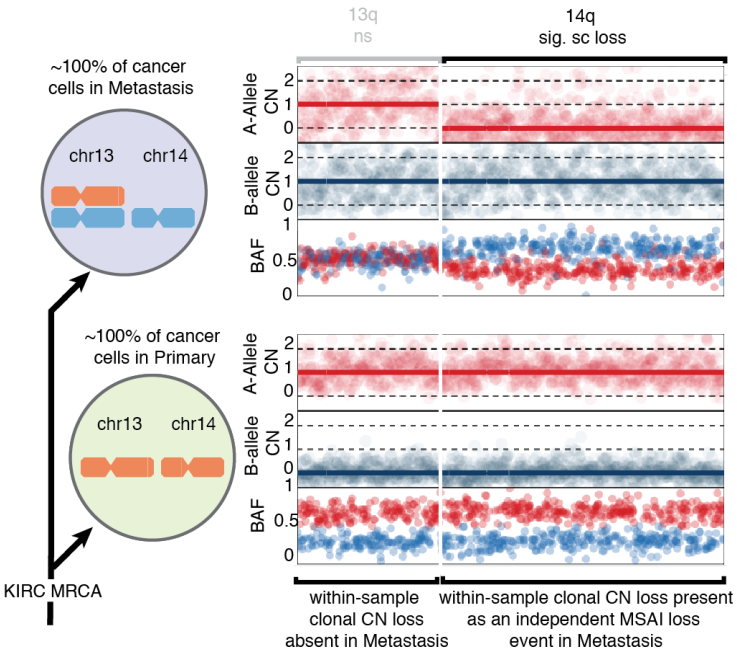

Figure 6-15: B_MET080 SCNA subclone undetectable in primary seeds a brain metastasis
Schematic on the left shows the on likely copy number state of the single detectable subclone present in the primary sample(shown in green) in this case the clone present in the metastasis likely originated form a subclone not detectable in the primary sample. The panel to the right shows the raw (continuous) values of the allele-specific copy number and BAF profiles for each a primary and mestatasis region from the patient's disease. Continuous or "raw" allele specific copy number is shown as red and blue dots, with integer copy number estimates in solid lines of the same colour. Subclonal copy number estimates are shown with a yellow line when present. Events found as significant in recurrence analyses for the tumour type examined are highlighted in black at top of the copy number plot for the metastasis region with those that were found to be non-significant shown in grey.

6.5 Conclusions

Pan-cancer investigations of SCNAs in from single regions have been performed on SNP arrays (Zack et al., 2013) and more recently on WGS (Dentro et al., 2018). However, these studies either entirely ignored SCNA subclonality as in the case of Zack et al. or were limited to the detection of non-integer SCNAs (with the caveats that entails, see Chapter 1) in single regions and were unable to detect mirrored subclonal allelic imbalance or any form of SCNA demonstrating an illusion of clonality as the in case of Dentro et al. The investigation of subclonal SCNAs in this chapter is the first pan-cancer exploration of subclonal SCNAs without such limitations. The identification of recurrent subclonal events is therefore a novel finding potentially explained by later selection pressures acting upon ongoing copy number alteration acquisition generated by CIN. This is supported by the differences in distribution across the genome of significant subclonal SCNAs between tumour types, suggesting that different SCNAs may be advantageous later in tumour evolution according to histology. The significant overlap of these events with previously

described significant GISTIC2 events from the TCGA analyses also suggests that these events may be having some fitness effect. Additional orthogonal analyses of gene enrichment in subclonal, and particularly parallel SCNA events as well as mathematical modelling of subclonal SCNA changes lend further credence to the idea of subclonal selection. Finally, that these subclonal SCNA changes appear to be enriched in metastatic disease, itself a bottle-necking event, suggests that these later emerging subclones often seed metastases.

Chapter 7 Discussion

The extent of SCNA intra-tumour heterogeneity, mode of tumour evolution and nature of evolutionary dynamics all have implications for predicting treatment failure and guiding treatment decisions (Greaves, 2015). My analyses of SCNAs in the TRACERx prospective cohort study contributed to the finding that intratumor heterogeneity mediated through chromosome instability was associated with an increased risk of recurrence or death in NSCLC, which supports the potential value of chromosome instability as a prognostic predictor. However, despite this demonstration and others of the clinical relevance of CIN, a systematic pan-cancer analysis of the prevalence and importance of ongoing SCNA acquisition in tumour evolution is lacking.

In this thesis I describe creation and application of a novel bioinformatics approach named mPHASE utilising multi-region phasing to quantify SCNA heterogeneity and identify instances parallel evolution through the detection of mirrored subclonal allelic imbalance from multi-region data. I applied this method and integrated data from a diverse array of multi-region studies across 22 tumour types, to interrogate the temporal landscape of SCNA acquisition and the propensity for clonal and subclonal SCNA events to be non-random and potentially subject to selection.

Extensive subclonal SCNA acquisition was observed across the cohort, with up to 84% of the genome subject to subclonal SCNA. Our results therefore argue against a simple punctuated model of tumour evolution, with limited SCNA acquisition following the emergence of the MRCA (Gao et al., 2016). However, consistent with saltationary leaps in tumour evolution, WGD was observed as a landmark event in many tumours development and found to be associated with increased levels of both clonal and subclonal SCNA acquisition. LOH events affecting tumour suppressor genes, including *TP53* and *RB1* were found to precede WGD, while gains to oncogenes frequently followed the duplication event and, as such, were more likely to be subclonal. Consistent with this, obligatory clonal SCNA were almost exclusively LOH events. Moreover, echoing studies exploring driver mutations in NSCLC (Jamal-Hanjani et al., 2017) and breast cancer (Yates et al., 2017) and pan-cancer analysis (Dentro et al., 2018), while clonal events were found to be subtype specific, subclonal events were more likely to converge across tumour types.

Taken as a whole, the landscape of SCNAs was found to be highly non-random throughout tumour evolution, with distinct as well as overlapping recurrent clonal and subclonal SCNA identified across tumour types. This reflects the impact of four related forces: mutation generation (CIN), genetic-drift, as well as negative and positive selection. To tease apart the relative contribution of these different forces and evaluate the extent to which the acquisition of SCNA is constrained, three largely orthogonal approaches were considered : i) overlap with known tumour suppressor genes (TSGs) and oncogenes identified through mutation analysis and GISTIC, ii) phasing-based detection of parallel evolution and iii) mathematical modelling of different modes of cancer evolution.

Consistent with recent findings focusing on mutation data (Martincorena et al., 2017, Dentro et al., 2018), my results strongly support the notion that positive selection plays a key role in shaping tumour development. Recurrent events were found to significantly overlap with previously identified SCNA drivers through TCGA GISTIC analyses, as well as cancer genes identified through mutation-based statistical approaches. This overlap was particularly pronounced for clonal and subclonal loss events, consistent with the Knudson two-hit hypothesis (Knudson, 1971). While recurrent subclonal gains were not found to significantly overlap with known oncogenes, we found an enrichment of oncogenes in gains identified through parallel evolution analysis. These data suggest constraints to tumour evolution which may limit the number of evolutionary avenues that a tumour explores. Notable genes affected by subclonal SCNA included *TERT*, which exhibited recurrent as well as parallel subclonal gains, and, *ARID1A*, which was found to be subject to both recurrent and parallel losses, as well as enrichment in LUAD metastatic sites.

Finally, the analysis of the data I have produced coupled with the mathematical modelling using this data performed by collaborators indicates the karyotypic evolution of a tumour can be better recapitulated using a modelling approach which incorporates positive selection, as compared to solely neutral dynamics. As the understanding of the intrinsic propensity for different chromosomes to mis-segregate (Worrall et al., 2018), as well as the extent to which chromosome alterations may be deleterious to the cancer cell improves, it will be possible to refine the parameters of this model and improve our ability to detect novel SCNA drivers.

In conclusion, the findings in thesis highlight the importance of ongoing SCNA acquisition as a key substrate of positive selection in cancer evolution, and begin to

shed light on the complex dynamics of contingencies and constraints that shape tumour evolutionary trajectories.

References

TARGET Database [Online]. Available: https://software.broadinstitute.org/cancer/cga/target [Accessed 2015].

ABBOSH, C., BIRKBAK, N. J., WILSON, G. A., JAMAL-HANJANI, M., CONSTANTIN, T., SALARI, R., LE QUESNE, J., MOORE, D. A., VEERIAH, S., ROSENTHAL, R., MARAFIOTI, T., KIRKIZLAR, E., WATKINS, T. B. K., MCGRANAHAN, N., WARD, S., MARTINSON, L., RILEY, J., FRAIOLI, F., AL BAKIR, M., GRONROOS, E., ZAMBRANA, F., ENDOZO, R., BI, W. L., FENNESSY, F. M., SPONER, N., JOHNSON, D., LAYCOCK, J., SHAFI, S., CZYZEWSKA-KHAN, J., ROWAN, A., CHAMBERS, T., MATTHEWS, N., TURAJLIC, S., HILEY, C., LEE, S. M., FORSTER, M. D., AHMAD, T., FALZON, M., BORG, E., LAWRENCE, D., HAYWARD, M., KOLVEKAR, S., PANAGIOTOPOULOS, N., JANES, S. M., THAKRAR, R., AHMED, A., BLACKHALL, F., SUMMERS, Y., HAFEZ, D., NAIK, A., GANGULY, A., KAREHT, S., SHAH, R., JOSEPH, L., MARIE QUINN, A., CROSBIE, P. A., NAIDU, B., MIDDLETON, G., LANGMAN, G., TROTTER, S., NICOLSON, M., REMMEN, H., KERR, K., CHETTY, M., GOMERSALL, L., FENNELL, D. A., NAKAS, A., RATHINAM, S., ANAND, G., KHAN, S., RUSSELL, P., EZHIL, V., ISMAIL, B., IRVIN-SELLERS, M., PRAKASH, V., LESTER, J. F., KORNASZEWSKA, M., ATTANOOS, R., ADAMS, H., DAVIES, H., OUKRIF, D., AKARCA, A. U., HARTLEY, J. A., LOWE, H. L., LOCK, S., ILES, N., BELL, H., NGAI, Y., ELGAR, G., SZALLASI, Z., SCHWARZ, R. F., HERRERO, J., STEWART, A., QUEZADA, S. A., PEGGS, K. S., VAN LOO, P., DIVE, C., LIN, C. J., RABINOWITZ, M., AERTS, H., et al. 2017. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature,* 545**,** 446-451.

ABKEVICH, V., TIMMS, K. M., HENNESSY, B. T., POTTER, J., CAREY, M. S., MEYER, L. A., SMITH-MCCUNE, K., BROADDUS, R., LU, K. H., CHEN, J., TRAN, T. V., WILLIAMS, D., ILIEV, D., JAMMULAPATI, S., FITZGERALD, L. M., KRIVAK, T., DELOIA, J. A., GUTIN, A., MILLS, G. B. & LANCHBURY, J. S. 2012. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br J Cancer,* 107**,** 1776-82.

ADAM, M., THORBURN, M. J., GIBBS, W. N., BROOKS, S. E. & HANCHARD, B. 1970. Clonal evolution in two patients with autoimmune disease and lymphoreticular neoplasia. *Br J Cancer,* 24**,** 266-76.

ADAMS, K. L. & WENDEL, J. F. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol,* 8**,** 135-41.

AGRAWAL, N., JIAO, Y., BETTEGOWDA, C., HUTFLESS, S. M., WANG, Y., DAVID, S., CHENG, Y., TWADDELL, W. S., LATT, N. L., SHIN, E. J., WANG, L.-D., WANG, L., YANG, W., VELCULESCU, V. E., VOGELSTEIN, B., PAPADOPOULOS, N., KINZLER, K. W. & MELTZER, S. J. 2012. Comparative Genomic Analysis of Esophageal

Adenocarcinoma and Squamous Cell Carcinoma. *Cancer Discovery,* 2**,** 899-905.

ALEXANDROV, L. B., NIK-ZAINAL, S., WEDGE, D. C., APARICIO, S. A., BEHJATI, S., BIANKIN, A. V., BIGNELL, G. R., BOLLI, N., BORG, A., BORRESEN-DALE, A. L., BOYAULT, S., BURKHARDT, B., BUTLER, A. P., CALDAS, C., DAVIES, H. R., DESMEDT, C., EILS, R., EYFJORD, J. E., FOEKENS, J. A., GREAVES, M., HOSODA, F., HUTTER, B., ILICIC, T., IMBEAUD, S., IMIELINSKI, M., JAGER, N., JONES, D. T., JONES, D., KNAPPSKOG, S., KOOL, M., LAKHANI, S. R., LOPEZ-OTIN, C., MARTIN, S., MUNSHI, N. C., NAKAMURA, H., NORTHCOTT, P. A., PAJIC, M., PAPAEMMANUIL, E., PARADISO, A., PEARSON, J. V., PUENTE, X. S., RAINE, K., RAMAKRISHNA, M., RICHARDSON, A. L., RICHTER, J., ROSENSTIEL, P., SCHLESNER, M., SCHUMACHER, T. N., SPAN, P. N., TEAGUE, J. W., TOTOKI, Y., TUTT, A. N., VALDES-MAS, R., VAN BUUREN, M. M., VAN 'T VEER, L., VINCENT-SALOMON, A., WADDELL, N., YATES, L. R., AUSTRALIAN PANCREATIC CANCER GENOME, I., CONSORTIUM, I. B. C., CONSORTIUM, I. M.-S., PEDBRAIN, I., ZUCMAN-ROSSI, J., FUTREAL, P. A., MCDERMOTT, U., LICHTER, P., MEYERSON, M., GRIMMOND, S. M., SIEBERT, R., CAMPO, E., SHIBATA, T., PFISTER, S. M., CAMPBELL, P. J. & STRATTON, M. R. 2013. Signatures of mutational processes in human cancer. *Nature,* 500**,** 415-21.

ALVES, J. M., PRIETO, T. & POSADA, D. 2017. Multiregional Tumor Trees Are Not Phylogenies. *Trends Cancer,* 3**,** 546-550.

ANDERSON, K., LUTZ, C., VAN DELFT, F. W., BATEMAN, C. M., GUO, Y., COLMAN, S. M., KEMPSKI, H., MOORMAN, A. V., TITLEY, I., SWANSBURY, J., KEARNEY, L., ENVER, T. & GREAVES, M. 2011. Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature,* 469**,** 356-61.

ANDOR, N., GRAHAM, T. A., JANSEN, M., XIA, L. C., AKTIPIS, C. A., PETRITSCH, C., JI, H. P. & MALEY, C. C. 2016. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med,* 22**,** 105-13.

ARMITAGE, P. & DOLL, R. 1954. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer,* 8**,** 1-12.

BACKENROTH, D., HOMSY, J., MURILLO, L. R., GLESSNER, J., LIN, E., BRUECKNER, M., LIFTON, R., GOLDMUNTZ, E., CHUNG, W. K. & SHEN, Y. 2014. CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res,* 42**,** e97.

BAILEY, M. H., TOKHEIM, C., PORTA-PARDO, E., SENGUPTA, S., BERTRAND, D., WEERASINGHE, A., COLAPRICO, A., WENDL, M. C., KIM, J., REARDON, B., NG, P. K., JEONG, K. J., CAO, S., WANG, Z., GAO, J., GAO, Q., WANG, F., LIU, E. M., MULARONI, L., RUBIO-PEREZ, C., NAGARAJAN, N., CORTES-CIRIANO, I., ZHOU, D. C., LIANG, W. W., HESS, J. M., YELLAPANTULA, V. D., TAMBORERO, D., GONZALEZ-PEREZ, A., SUPHAVILAI, C., KO, J. Y., KHURANA, E., PARK, P. J., VAN ALLEN, E. M., LIANG, H., GROUP, M. C. W., CANCER GENOME ATLAS RESEARCH, N., LAWRENCE, M. S., GODZIK, A., LOPEZ-BIGAS, N., STUART, J., WHEELER, D., GETZ,

G., CHEN, K., LAZAR, A. J., MILLS, G. B., KARCHIN, R. & DING, L. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell,* 173**,** 371-385 e18.

BAKHOUM, S. F., DANILOVA, O. V., KAUR, P., LEVY, N. B. & COMPTON, D. A. 2011. Chromosomal instability substantiates poor prognosis in patients with diffuse large B-cell lymphoma. *Clin Cancer Res,* 17**,** 7704-11.

BAKHOUM, S. F., KABECHE, L., MURNANE, J. P., ZAKI, B. I. & COMPTON, D. A. 2014. DNA-damage response during mitosis induces whole-chromosome missegregation. *Cancer Discov,* 4**,** 1281-9.

BAKHOUM, S. F., NGO, B., LAUGHNEY, A. M., CAVALLO, J. A., MURPHY, C. J., LY, P., SHAH, P., SRIRAM, R. K., WATKINS, T. B. K., TAUNK, N. K., DURAN, M., PAULI, C., SHAW, C., CHADALAVADA, K., RAJASEKHAR, V. K., GENOVESE, G., VENKATESAN, S., BIRKBAK, N. J., MCGRANAHAN, N., LUNDQUIST, M., LAPLANT, Q., HEALEY, J. H., ELEMENTO, O., CHUNG, C. H., LEE, N. Y., IMIELENSKI, M., NANJANGUD, G., PE'ER, D., CLEVELAND, D. W., POWELL, S. N., LAMMERDING, J., SWANTON, C. & CANTLEY, L. C. 2018. Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature,* 553**,** 467-472.

BAKKER, B., TAUDT, A., BELDERBOS, M. E., PORUBSKY, D., SPIERINGS, D. C., DE JONG, T. V., HALSEMA, N., KAZEMIER, H. G., HOEKSTRA-WAKKER, K., BRADLEY, A., DE BONT, E. S., VAN DEN BERG, A., GURYEV, V., LANSDORP, P. M., COLOME-TATCHE, M. & FOIJER, F. 2016. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol,* 17**,** 115.

BALAPARYA, A. & DE, S. 2018. Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data. *Nat Genet,* 50**,** 1626-1628.

BALKWILL, F. & MANTOVANI, A. 2001. Inflammation and cancer: back to Virchow? *Lancet,* 357**,** 539-45.

BARSKI, G. & CASSINGENA, R. 1963. Malignant transformation in vitro of cells from C57BL mouse normal pulmonary tissue. *J Natl Cancer Inst,* 30**,** 865-83.

BARTLETT, J. M., CAMPBELL, F. M. & MALLON, E. A. 2008. Determination of HER2 amplification by in situ hybridization: when should chromosome 17 also be determined? *Am J Clin Pathol,* 130**,** 920-6.

BASHASHATI, A., HA, G., TONE, A., DING, J., PRENTICE, L. M., ROTH, A., ROSNER, J., SHUMANSKY, K., KALLOGER, S., SENZ, J., YANG, W., MCCONECHY, M., MELNYK, N., ANGLESIO, M., LUK, M. T., TSE, K., ZENG, T., MOORE, R., ZHAO, Y., MARRA, M. A., GILKS, B., YIP, S., HUNTSMAN, D. G., MCALPINE, J. N. & SHAH, S. P. 2013. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J Pathol,* 231**,** 21-34.

BASLAN, T., KENDALL, J., WARD, B., COX, H., LEOTTA, A., RODGERS, L., RIGGS, M., D'ITALIA, S., SUN, G., YONG, M., MISKIMEN, K., GILMORE, H., SABOROWSKI, M., DIMITROVA, N., KRASNITZ, A., HARRIS, L., WIGLER, M. & HICKS, J. 2015. Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. *Genome Res,* 25**,** 714-24.

BEERENWINKEL, N., SCHWARZ, R. F., GERSTUNG, M. & MARKOWETZ, F. 2014. Cancer evolution: mathematical models and computational inference. *Syst Biol*.

BELLOS, E. & COIN, L. J. 2014. cnvOffSeq: detecting intergenic copy number variation using off-target exome sequencing data. *Bioinformatics,* 30**,** i639-45.

BEROUKHIM, R., BRUNET, J. P., DI NAPOLI, A., MERTZ, K. D., SEELEY, A., PIRES, M. M., LINHART, D., WORRELL, R. A., MOCH, H., RUBIN, M. A., SELLERS, W. R., MEYERSON, M., LINEHAN, W. M., KAELIN, W. G., JR. & SIGNORETTI, S. 2009. Patterns of gene expression and copy-number alterations in von-hippel lindau disease-associated and sporadic clear cell carcinoma of the kidney. *Cancer Res,* 69**,** 4674-81.

BEROUKHIM, R., GETZ, G., NGHIEMPHU, L., BARRETINA, J., HSUEH, T., LINHART, D., VIVANCO, I., LEE, J. C., HUANG, J. H., ALEXANDER, S., DU, J., KAU, T., THOMAS, R. K., SHAH, K., SOTO, H., PERNER, S., PRENSNER, J., DEBIASI, R. M., DEMICHELIS, F., HATTON, C., RUBIN, M. A., GARRAWAY, L. A., NELSON, S. F., LIAU, L., MISCHEL, P. S., CLOUGHESY, T. F., MEYERSON, M., GOLUB, T. A., LANDER, E. S., MELLINGHOFF, I. K. & SELLERS, W. R. 2007. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences,* 104**,** 20007-20012.

BEROUKHIM, R., MERMEL, C. H., PORTER, D., WEI, G., RAYCHAUDHURI, S., DONOVAN, J., BARRETINA, J., BOEHM, J. S., DOBSON, J., URASHIMA, M., MC HENRY, K. T., PINCHBACK, R. M., LIGON, A. H., CHO, Y.-J., HAERY, L., GREULICH, H., REICH, M., WINCKLER, W., LAWRENCE, M. S., WEIR, B. A., TANAKA, K. E., CHIANG, D. Y., BASS, A. J., LOO, A., HOFFMAN, C., PRENSNER, J., LIEFELD, T., GAO, Q., YECIES, D., SIGNORETTI, S., MAHER, E., KAYE, F. J., SASAKI, H., TEPPER, J. E., FLETCHER, J. A., TABERNERO, J., BASELGA, J., TSAO, M.-S., DEMICHELIS, F., RUBIN, M. A., JANNE, P. A., DALY, M. J., NUCERA, C., LEVINE, R. L., EBERT, B. L., GABRIEL, S., RUSTGI, A. K., ANTONESCU, C. R., LADANYI, M., LETAI, A., GARRAWAY, L. A., LODA, M., BEER, D. G., TRUE, L. D., OKAMOTO, A., POMEROY, S. L., SINGER, S., GOLUB, T. R., LANDER, E. S., GETZ, G., SELLERS, W. R. & MEYERSON, M. 2010. The landscape of somatic copy-number alteration across human cancers. *Nature,* 463**,** 899-905.

BETTEGOWDA, C., SAUSEN, M., LEARY, R. J., KINDE, I., WANG, Y., AGRAWAL, N., BARTLETT, B. R., WANG, H., LUBER, B., ALANI, R. M., ANTONARAKIS, E. S., AZAD, N. S., BARDELLI, A., BREM, H., CAMERON, J. L., LEE, C. C., FECHER, L. A., GALLIA, G. L., GIBBS, P., LE, D., GIUNTOLI, R. L., GOGGINS, M., HOGARTY, M. D., HOLDHOFF, M., HONG, S. M., JIAO, Y., JUHL, H. H., KIM, J. J., SIRAVEGNA, G., LAHERU, D. A., LAURICELLA, C., LIM, M., LIPSON, E. J., MARIE, S. K., NETTO, G. J., OLINER, K. S., OLIVI, A., OLSSON, L., RIGGINS, G. J., SARTORE-BIANCHI, A., SCHMIDT, K., SHIH L, M., OBA-SHINJO, S. M., SIENA, S., THEODORESCU, D., TIE, J., HARKINS, T. T., VERONESE, S., WANG, T. L., WEINGART, J. D., WOLFGANG, C. L., WOOD, L. D., XING, D., HRUBAN, R. H., WU, J.,

ALLEN, P. J., SCHMIDT, C. M., CHOTI, M. A., VELCULESCU, V. E., KINZLER, K. W., VOGELSTEIN, B., PAPADOPOULOS, N. & DIAZ, L. A., JR. 2014. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med,* 6**,** 224ra24.

BIELSKI, C. M., ZEHIR, A., PENSON, A. V., DONOGHUE, M. T. A., CHATILA, W., ARMENIA, J., CHANG, M. T., SCHRAM, A. M., JONSSON, P., BANDLAMUDI, C., RAZAVI, P., IYER, G., ROBSON, M. E., STADLER, Z. K., SCHULTZ, N., BASELGA, J., SOLIT, D. B., HYMAN, D. M., BERGER, M. F. & TAYLOR, B. S. 2018. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat Genet,* 50**,** 1189-1195.

BIERKENS, M., KRIJGSMAN, O., WILTING, S. M., BOSCH, L., JASPERS, A., MEIJER, G. A., MEIJER, C. J., SNIJDERS, P. J., YLSTRA, B. & STEENBERGEN, R. D. 2013. Focal aberrations indicate EYA2 and hsa-miR-375 as oncogene and tumor suppressor in cervical carcinogenesis. *Genes Chromosomes Cancer,* 52**,** 56-68.

BIGNELL, G. R., GREENMAN, C. D., DAVIES, H., BUTLER, A. P., EDKINS, S., ANDREWS, J. M., BUCK, G., CHEN, L., BEARE, D., LATIMER, C., WIDAA, S., HINTON, J., FAHEY, C., FU, B., SWAMY, S., DALGLIESH, G. L., TEH, B. T., DELOUKAS, P., YANG, F., CAMPBELL, P. J., FUTREAL, P. A. & STRATTON, M. R. 2010. Signatures of mutation and selection in the cancer genome. *Nature,* 463**,** 893-898.

BIRKBAK, N. J., WANG, Z. C., KIM, J. Y., EKLUND, A. C., LI, Q., TIAN, R., BOWMAN-COLIN, C., LI, Y., GREENE-COLOZZI, A., IGLEHART, J. D., TUNG, N., RYAN, P. D., GARBER, J. E., SILVER, D. P., SZALLASI, Z. & RICHARDSON, A. L. 2012. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov,* 2**,** 366-75.

BOEVA, V., POPOVA, T., BLEAKLEY, K., CHICHE, P., CAPPO, J., SCHLEIERMACHER, G., JANOUEIX-LEROSEY, I., DELATTRE, O. & BARILLOT, E. 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics,* 28**,** 423-5.

BOLLI, N., AVET-LOISEAU, H., WEDGE, D. C., VAN LOO, P., ALEXANDROV, L. B., MARTINCORENA, I., DAWSON, K. J., IORIO, F., NIK-ZAINAL, S., BIGNELL, G. R., HINTON, J. W., LI, Y., TUBIO, J. M., MCLAREN, S., S, O. M., BUTLER, A. P., TEAGUE, J. W., MUDIE, L., ANDERSON, E., RASHID, N., TAI, Y. T., SHAMMAS, M. A., SPERLING, A. S., FULCINITI, M., RICHARDSON, P. G., PARMIGIANI, G., MAGRANGEAS, F., MINVIELLE, S., MOREAU, P., ATTAL, M., FACON, T., FUTREAL, P. A., ANDERSON, K. C., CAMPBELL, P. J. & MUNSHI, N. C. 2014. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun,* 5**,** 2997.

BONNEY, M. E., MORIYA, H. & AMON, A. 2015. Aneuploid proliferation defects in yeast are not driven by copy number changes of a few dosage-sensitive genes. *Genes Dev,* 29**,** 898-903.

BOVERI, T. 1914. Zur Frage der Entstehung Maligner Tumoren.

BRANNON, A. R., VAKIANI, E., SYLVESTER, B. E., SCOTT, S. N., MCDERMOTT, G., SHAH, R. H., KANIA, K., VIALE, A., OSCHWALD, D. M., VACIC, V., EMDE, A. K., CERCEK, A., YAEGER, R., KEMENY,

N. E., SALTZ, L. B., SHIA, J., D'ANGELICA, M. I., WEISER, M. R., SOLIT, D. B. & BERGER, M. F. 2014. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome Biol,* 15**,** 454.

BRASTIANOS, P. K., CARTER, S. L., SANTAGATA, S., CAHILL, D. P., TAYLOR-WEINER, A., JONES, R. T., VAN ALLEN, E. M., LAWRENCE, M. S., HOROWITZ, P. M., CIBULSKIS, K., LIGON, K. L., TABERNERO, J., SEOANE, J., MARTINEZ-SAEZ, E., CURRY, W. T., DUNN, I. F., PAEK, S. H., PARK, S. H., MCKENNA, A., CHEVALIER, A., ROSENBERG, M., BARKER, F. G., 2ND, GILL, C. M., VAN HUMMELEN, P., THORNER, A. R., JOHNSON, B. E., HOANG, M. P., CHOUEIRI, T. K., SIGNORETTI, S., SOUGNEZ, C., RABIN, M. S., LIN, N. U., WINER, E. P., STEMMER-RACHAMIMOV, A., MEYERSON, M., GARRAWAY, L., GABRIEL, S., LANDER, E. S., BEROUKHIM, R., BATCHELOR, T. T., BASELGA, J., LOUIS, D. N., GETZ, G. & HAHN, W. C. 2015. Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. *Cancer Discov,* 5**,** 1164-1177.

BROWN, D., SMEETS, D., SZEKELY, B., LARSIMONT, D., SZASZ, A. M., ADNET, P. Y., ROTHE, F., ROUAS, G., NAGY, Z. I., FARAGO, Z., TOKES, A. M., DANK, M., SZENTMARTONI, G., UDVARHELYI, N., ZOPPOLI, G., PUSZTAI, L., PICCART, M., KULKA, J., LAMBRECHTS, D., SOTIRIOU, C. & DESMEDT, C. 2017. Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nat Commun,* 8**,** 14944.

BUCCITELLI, C., SALGUEIRO, L., ROWALD, K., SOTILLO, R., MARDIN, B. R. & KORBEL, J. O. 2017. Pan-cancer analysis distinguishes transcriptional changes of aneuploidy from proliferation. *Genome Res,* 27**,** 501-511.

BURRELL, R. A., MCCLELLAND, S. E., ENDESFELDER, D., GROTH, P., WELLER, M. C., SHAIKH, N., DOMINGO, E., KANU, N., DEWHURST, S. M., GRONROOS, E., CHEW, S. K., ROWAN, A. J., SCHENK, A., SHEFFER, M., HOWELL, M., KSCHISCHO, M., BEHRENS, A., HELLEDAY, T., BARTEK, J., TOMLINSON, I. P. & SWANTON, C. 2013a. Replication stress links structural and numerical cancer chromosomal instability. *Nature,* 494**,** 492-496.

BURRELL, R. A., MCGRANAHAN, N., BARTEK, J. & SWANTON, C. 2013b. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature,* 501**,** 338-45.

CAI, Y., CROWTHER, J., PASTOR, T., ABBASI ASBAGH, L., BAIETTI, M. F., DE TROYER, M., VAZQUEZ, I., TALEBI, A., RENZI, F., DEHAIRS, J., SWINNEN, J. V. & SABLINA, A. A. 2016. Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism. *Cancer Cell,* 29**,** 751-766.

CAIRNCROSS, G., WANG, M., SHAW, E., JENKINS, R., BRACHMAN, D., BUCKNER, J., FINK, K., SOUHAMI, L., LAPERRIERE, N., CURRAN, W. & MEHTA, M. 2013. Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: long-term results of RTOG 9402. *J Clin Oncol,* 31**,** 337-43.

CAIRNS, J. 1975. Mutation selection and the natural history of cancer. *Nature,* 255**,** 197-200.

CAIRNS, J. 1981. The origin of human cancers. *Nature,* 289**,** 353-7.

CAMPBELL, J. D., ALEXANDROV, A., KIM, J., WALA, J., BERGER, A. H., PEDAMALLU, C. S., SHUKLA, S. A., GUO, G., BROOKS, A. N., MURRAY, B. A., IMIELINSKI, M., HU, X., LING, S., AKBANI, R., ROSENBERG, M., CIBULSKIS, C., RAMACHANDRAN, A., COLLISSON, E. A., KWIATKOWSKI, D. J., LAWRENCE, M. S., WEINSTEIN, J. N., VERHAAK, R. G., WU, C. J., HAMMERMAN, P. S., CHERNIACK, A. D., GETZ, G., CANCER GENOME ATLAS RESEARCH, N., ARTYOMOV, M. N., SCHREIBER, R., GOVINDAN, R. & MEYERSON, M. 2016. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat Genet,* 48**,** 607-16.

CAMPBELL, P. J., PLEASANCE, E. D., STEPHENS, P. J., DICKS, E., RANCE, R., GOODHEAD, I., FOLLOWS, G. A., GREEN, A. R., FUTREAL, P. A. & STRATTON, M. R. 2008. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci U S A,* 105**,** 13081-6.

CANNATARO, V. L. & TOWNSEND, J. P. 2018. Neutral Theory and the Somatic Evolution of Cancer. *Mol Biol Evol,* 35**,** 1308-1315.

CAPON, D. J., SEEBURG, P. H., MCGRATH, J. P., HAYFLICK, J. S., EDMAN, U., LEVINSON, A. D. & GOEDDEL, D. V. 1983. Activation of Ki-ras2 gene in human colon and lung carcinomas by two different point mutations. *Nature,* 304**,** 507-13.

CARTER, L., ROTHWELL, D. G., MESQUITA, B., SMOWTON, C., LEONG, H. S., FERNANDEZ-GUTIERREZ, F., LI, Y., BURT, D. J., ANTONELLO, J., MORROW, C. J., HODGKINSON, C. L., MORRIS, K., PRIEST, L., CARTER, M., MILLER, C., HUGHES, A., BLACKHALL, F., DIVE, C. & BRADY, G. 2017. Molecular analysis of circulating tumor cells identifies distinct copy-number profiles in patients with chemosensitive and chemorefractory small-cell lung cancer. *Nat Med,* 23**,** 114-119.

CARTER, S. L., CIBULSKIS, K., HELMAN, E., MCKENNA, A., SHEN, H., ZACK, T., LAIRD, P. W., ONOFRIO, R. C., WINCKLER, W., WEIR, B. A., BEROUKHIM, R., PELLMAN, D., LEVINE, D. A., LANDER, E. S., MEYERSON, M. & GETZ, G. 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology,* 30**,** 413-21.

CARTER, S. L., EKLUND, A. C., KOHANE, I. S., HARRIS, L. N. & SZALLASI, Z. 2006. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat Genet,* 38**,** 1043-8.

CASPERSSON, T., ZECH, L. & JOHANSSON, C. 1970. Differential binding of alkylating fluorochromes in human chromosomes. *Exp Cell Res,* 60**,** 315-9.

CASTEL, S. E., MOHAMMADI, P., CHUNG, W. K., SHEN, Y. & LAPPALAINEN, T. 2016. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat Commun,* 7**,** 12817.

CAVENEE, W. K., DRYJA, T. P., PHILLIPS, R. A., BENEDICT, W. F., GODBOUT, R., GALLIE, B. L., MURPHREE, A. L., STRONG, L. C. & WHITE, R. L. 1983. Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature,* 305**,** 779-84.

CAVENEE, W. K., HANSEN, M. F., NORDENSKJOLD, M., KOCK, E., MAUMENEE, I., SQUIRE, J. A., PHILLIPS, R. A. & GALLIE, B. L. 1985. Genetic origin of mutations predisposing to retinoblastoma. *Science,* 228**,** 501-3.

CHEN, F., ZHANG, Y., SENBABAOGLU, Y., CIRIELLO, G., YANG, L., REZNIK, E., SHUCH, B., MICEVIC, G., DE VELASCO, G., SHINBROT, E., NOBLE, M. S., LU, Y., COVINGTON, K. R., XI, L., DRUMMOND, J. A., MUZNY, D., KANG, H., LEE, J., TAMBOLI, P., REUTER, V., SHELLEY, C. S., KAIPPARETTU, B. A., BOTTARO, D. P., GODWIN, A. K., GIBBS, R. A., GETZ, G., KUCHERLAPATI, R., PARK, P. J., SANDER, C., HENSKE, E. P., ZHOU, J. H., KWIATKOWSKI, D. J., HO, T. H., CHOUEIRI, T. K., HSIEH, J. J., AKBANI, R., MILLS, G. B., HAKIMI, A. A., WHEELER, D. A. & CREIGHTON, C. J. 2016. Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma. *Cell Rep,* 14**,** 2476-89.

CHEN, X., GUPTA, P., WANG, J., NAKITANDWE, J., ROBERTS, K., DALTON, J. D., PARKER, M., PATEL, S., HOLMFELDT, L., PAYNE, D., EASTON, J., MA, J., RUSCH, M., WU, G., PATEL, A., BAKER, S. J., DYER, M. A., SHURTLEFF, S., ESPY, S., POUNDS, S., DOWNING, J. R., ELLISON, D. W., MULLIGHAN, C. G. & ZHANG, J. 2015. CONSERTING: integrating copy-number analysis with structural-variation detection. *Nat Methods,* 12**,** 527-30.

CHEN, Z. Y., ZHONG, W. Z., ZHANG, X. C., SU, J., YANG, X. N., CHEN, Z. H., YANG, J. J., ZHOU, Q., YAN, H. H., AN, S. J., CHEN, H. J., JIANG, B. Y., MOK, T. S. & WU, Y. L. 2012. EGFR mutation heterogeneity and the mixed response to EGFR tyrosine kinase inhibitors of lung adenocarcinomas. *Oncologist,* 17**,** 978-85.

CHENG, J., DEMEULEMEESTER, J., WEDGE, D. C., VOLLAN, H. K. M., PITT, J. J., RUSSNES, H. G., PANDEY, B. P., NILSEN, G., NORD, S., BIGNELL, G. R., WHITE, K. P., BORRESEN-DALE, A. L., CAMPBELL, P. J., KRISTENSEN, V. N., STRATTON, M. R., LINGJAERDE, O. C., MOREAU, Y. & VAN LOO, P. 2017. Pan-cancer analysis of homozygous deletions in primary tumours uncovers rare tumour suppressors. *Nat Commun,* 8**,** 1221.

CHOI, C. M., SEO, K. W., JANG, S. J., OH, Y. M., SHIM, T. S., KIM, W. S., LEE, D. S. & LEE, S. D. 2009. Chromosomal instability is a risk factor for poor prognosis of adenocarcinoma of the lung: Fluorescence in situ hybridization analysis of paraffin-embedded tissue from Korean patients. *Lung Cancer,* 64**,** 66-70.

CIBULSKIS, K., LAWRENCE, M. S., CARTER, S. L., SIVACHENKO, A., JAFFE, D., SOUGNEZ, C., GABRIEL, S., MEYERSON, M., LANDER, E. S. & GETZ, G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol,* 31**,** 213-9.

COOPER, H. L. & BLACK, P. H. 1963. Cytogenetic studies of hamster kidney cell cultures transformed by the simian vacuolating virus (SV40). *J Natl Cancer Inst,* 30**,** 1015-43.

DAGO, A. E., STEPANSKY, A., CARLSSON, A., LUTTGEN, M., KENDALL, J., BASLAN, T., KOLATKAR, A., WIGLER, M., BETHEL, K., GROSS, M. E., HICKS, J. & KUHN, P. 2014. Rapid phenotypic and genomic change in response to therapeutic pressure in prostate cancer inferred by high content analysis of single circulating tumor cells. *PLoS One,* 9**,** e101777.

DALGLIESH, G. L., FURGE, K., GREENMAN, C., CHEN, L., BIGNELL, G., BUTLER, A., DAVIES, H., EDKINS, S., HARDY, C., LATIMER, C., TEAGUE, J., ANDREWS, J., BARTHORPE, S., BEARE, D., BUCK, G., CAMPBELL, P. J., FORBES, S., JIA, M., JONES, D., KNOTT, H., KOK, C. Y., LAU, K. W., LEROY, C., LIN, M. L., MCBRIDE, D. J., MADDISON, M., MAGUIRE, S., MCLAY, K., MENZIES, A., MIRONENKO, T., MULDERRIG, L., MUDIE, L., O'MEARA, S., PLEASANCE, E., RAJASINGHAM, A., SHEPHERD, R., SMITH, R., STEBBINGS, L., STEPHENS, P., TANG, G., TARPEY, P. S., TURRELL, K., DYKEMA, K. J., KHOO, S. K., PETILLO, D., WONDERGEM, B., ANEMA, J., KAHNOSKI, R. J., TEH, B. T., STRATTON, M. R. & FUTREAL, P. A. 2010. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature,* 463**,** 360-3.

DAVIES, H., BIGNELL, G. R., COX, C., STEPHENS, P., EDKINS, S., CLEGG, S., TEAGUE, J., WOFFENDIN, H., GARNETT, M. J., BOTTOMLEY, W., DAVIS, N., DICKS, E., EWING, R., FLOYD, Y., GRAY, K., HALL, S., HAWES, R., HUGHES, J., KOSMIDOU, V., MENZIES, A., MOULD, C., PARKER, A., STEVENS, C., WATT, S., HOOPER, S., WILSON, R., JAYATILAKE, H., GUSTERSON, B. A., COOPER, C., SHIPLEY, J., HARGRAVE, D., PRITCHARD-JONES, K., MAITLAND, N., CHENEVIX-TRENCH, G., RIGGINS, G. J., BIGNER, D. D., PALMIERI, G., COSSU, A., FLANAGAN, A., NICHOLSON, A., HO, J. W., LEUNG, S. Y., YUEN, S. T., WEBER, B. L., SEIGLER, H. F., DARROW, T. L., PATERSON, H., MARAIS, R., MARSHALL, C. J., WOOSTER, R., STRATTON, M. R. & FUTREAL, P. A. 2002. Mutations of the BRAF gene in human cancer. *Nature,* 417**,** 949-54.

DAVIES, H., GLODZIK, D., MORGANELLA, S., YATES, L. R., STAAF, J., ZOU, X., RAMAKRISHNA, M., MARTIN, S., BOYAULT, S., SIEUWERTS, A. M., SIMPSON, P. T., KING, T. A., RAINE, K., EYFJORD, J. E., KONG, G., BORG, A., BIRNEY, E., STUNNENBERG, H. G., VAN DE VIJVER, M. J., BORRESEN-DALE, A. L., MARTENS, J. W., SPAN, P. N., LAKHANI, S. R., VINCENT-SALOMON, A., SOTIRIOU, C., TUTT, A., THOMPSON, A. M., VAN LAERE, S., RICHARDSON, A. L., VIARI, A., CAMPBELL, P. J., STRATTON, M. R. & NIK-ZAINAL, S. 2017. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med,* 23**,** 517-525.

DAVOLI, T. & DE LANGE, T. 2011. The causes and consequences of polyploidy in normal development and cancer. *Annu Rev Cell Dev Biol,* 27**,** 585-610.

DAVOLI, T., UNO, H., WOOTEN, E. C. & ELLEDGE, S. J. 2017. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science,* 355.

DAVOLI, T., XU, A. W., MENGWASSER, K. E., SACK, L. M., YOON, J. C., PARK, P. J. & ELLEDGE, S. J. 2013. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell,* 155**,** 948-62.

DE BRUIN, E. C., MCGRANAHAN, N., MITTER, R., SALM, M., WEDGE, D. C., YATES, L., JAMAL-HANJANI, M., SHAFI, S., MURUGAESU, N., ROWAN, A. J., GRONROOS, E., MUHAMMAD, M. A., HORSWELL, S., GERLINGER, M., VARELA, I., JONES, D., MARSHALL, J., VOET, T., VAN LOO, P., RASSL, D. M., RINTOUL, R. C., JANES, S. M., LEE, S. M., FORSTER, M., AHMAD, T., LAWRENCE, D., FALZON, M., CAPITANIO, A., HARKINS, T. T., LEE, C. C., TOM, W., TEEFE, E., CHEN, S. C., BEGUM, S., RABINOWITZ, A., PHILLIMORE, B., SPENCER-DENE, B., STAMP, G., SZALLASI, Z., MATTHEWS, N., STEWART, A., CAMPBELL, P. & SWANTON, C. 2014. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science,* 346**,** 251-6.

DE GROUCHY, J., DE NAVA, C., CANTU, J. M., BILSKI-PASQUIER, G. & BOUSSER, J. 1966. Models for clonal evolutions: a study of chronic myelogenous leukemia. *Am J Hum Genet,* 18**,** 485-503.

DEGNER, J. F., MARIONI, J. C., PAI, A. A., PICKRELL, J. K., NKADORI, E., GILAD, Y. & PRITCHARD, J. K. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics,* 25**,** 3207-12.

DELANEAU, O., MARCHINI, J., GENOMES PROJECT, C. & GENOMES PROJECT, C. 2014. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun,* 5**,** 3934.

DELANEAU, O., ZAGURY, J. F. & MARCHINI, J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods,* 10**,** 5-6.

DEMEULEMEESTER, J., KUMAR, P., MOLLER, E. K., NORD, S., WEDGE, D. C., PETERSON, A., MATHIESEN, R. R., FJELLDAL, R., ZAMANI ESTEKI, M., THEUNIS, K., FERNANDEZ GALLARDO, E., GRUNDSTAD, A. J., BORGEN, E., BAUMBUSCH, L. O., BORRESEN-DALE, A. L., WHITE, K. P., KRISTENSEN, V. N., VAN LOO, P., VOET, T. & NAUME, B. 2016. Tracing the origin of disseminated tumor cells in breast cancer using single-cell sequencing. *Genome Biol,* 17**,** 250.

DENTRO, S. C., LESHCHINER, I., HAASE, K., TARABICHI, M., WINTERSINGER, J., DESHWAR, A. G., YU, K., RUBANOVA, Y., MCINTYRE, G., VAZQUEZ-GARCIA, I., KLEINHEINZ, K., LIVITZ, D. G., MALIKIC, S., DONMEZ, N., SENGUPTA, S., DEMEULEMEESTER, J., ANUR, P., JOLLY, C., CMERO, M., ROSEBROCK, D., SCHUMACHER, S., FAN, Y., FITTALL, M., DREWS, R. M., YAO, X., LEE, J., SCHLESNER, M., ADAMS, D. J., GETZ, G., BOUTROS, P. C., IMIELINSKI, M., BEROUKHIM, R., SAHINALP, S. C., JI, Y., PEIFER, M., MARTINCORENA, I., MARKOWETZ, F., MUSTONEN, V., YUAN, K., GERSTUNG, M.,

SPELLMAN, P. T., WANG, W., MORRIS, Q. D., WEDGE, D. C. & VAN LOO, P. 2018. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. *bioRxiv*.

DESHWAR, A. G., VEMBU, S., YUNG, C. K., JANG, G. H., STEIN, L. & MORRIS, Q. 2015. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol,* 16**,** 35.

DEWHURST, S. M., MCGRANAHAN, N., BURRELL, R. A., ROWAN, A. J., GRONROOS, E., ENDESFELDER, D., JOSHI, T., MOURADOV, D., GIBBS, P., WARD, R. L., HAWKINS, N. J., SZALLASI, Z., SIEBER, O. M. & SWANTON, C. 2014. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov,* 4**,** 175-185.

DI VINCI, A., INFUSINI, E., PEVERI, C., SCIUTTO, A., ORECCHIA, R., GEIDO, E., MONACO, R. & GIARETTI, W. 1999. Intratumor heterogeneity of chromosome 1, 7, 17, and 18 aneusomies obtained by FISH and association with flow cytometric DNA index in human colorectal adenocarcinomas. *Cytometry,* 35**,** 369-75.

DIAZ, L. A., JR., WILLIAMS, R. T., WU, J., KINDE, I., HECHT, J. R., BERLIN, J., ALLEN, B., BOZIC, I., REITER, J. G., NOWAK, M. A., KINZLER, K. W., OLINER, K. S. & VOGELSTEIN, B. 2012. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature,* 486**,** 537-40.

DING, L., GETZ, G., WHEELER, D. A., MARDIS, E. R., MCLELLAN, M. D., CIBULSKIS, K., SOUGNEZ, C., GREULICH, H., MUZNY, D. M., MORGAN, M. B., FULTON, L., FULTON, R. S., ZHANG, Q., WENDL, M. C., LAWRENCE, M. S., LARSON, D. E., CHEN, K., DOOLING, D. J., SABO, A., HAWES, A. C., SHEN, H., JHANGIANI, S. N., LEWIS, L. R., HALL, O., ZHU, Y., MATHEW, T., REN, Y., YAO, J., SCHERER, S. E., CLERC, K., METCALF, G. A., NG, B., MILOSAVLJEVIC, A., GONZALEZ-GARAY, M. L., OSBORNE, J. R., MEYER, R., SHI, X., TANG, Y., KOBOLDT, D. C., LIN, L., ABBOTT, R., MINER, T. L., POHL, C., FEWELL, G., HAIPEK, C., SCHMIDT, H., DUNFORD-SHORE, B. H., KRAJA, A., CROSBY, S. D., SAWYER, C. S., VICKERY, T., SANDER, S., ROBINSON, J., WINCKLER, W., BALDWIN, J., CHIRIEAC, L. R., DUTT, A., FENNELL, T., HANNA, M., JOHNSON, B. E., ONOFRIO, R. C., THOMAS, R. K., TONON, G., WEIR, B. A., ZHAO, X., ZIAUGRA, L., ZODY, M. C., GIORDANO, T., ORRINGER, M. B., ROTH, J. A., SPITZ, M. R., WISTUBA, II, OZENBERGER, B., GOOD, P. J., CHANG, A. C., BEER, D. G., WATSON, M. A., LADANYI, M., BRODERICK, S., YOSHIZAWA, A., TRAVIS, W. D., PAO, W., PROVINCE, M. A., WEINSTOCK, G. M., VARMUS, H. E., GABRIEL, S. B., LANDER, E. S., GIBBS, R. A., MEYERSON, M. & WILSON, R. K. 2008. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature,* 455**,** 1069-75.

DING, L., LEY, T. J., LARSON, D. E., MILLER, C. A., KOBOLDT, D. C., WELCH, J. S., RITCHEY, J. K., YOUNG, M. A., LAMPRECHT, T., MCLELLAN, M. D., MCMICHAEL, J. F., WALLIS, J. W., LU, C., SHEN, D., HARRIS, C. C., DOOLING, D. J., FULTON, R. S., FULTON, L. L., CHEN, K., SCHMIDT, H., KALICKI-VEIZER, J., MAGRINI, V. J.,

COOK, L., MCGRATH, S. D., VICKERY, T. L., WENDL, M. C., HEATH, S., WATSON, M. A., LINK, D. C., TOMASSON, M. H., SHANNON, W. D., PAYTON, J. E., KULKARNI, S., WESTERVELT, P., WALTER, M. J., GRAUBERT, T. A., MARDIS, E. R., WILSON, R. K. & DIPERSIO, J. F. 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature,* 481**,** 506-10.

DISKIN, S. J., ECK, T., GRESHOCK, J., MOSSE, Y. P., NAYLOR, T., STOECKERT, C. J., JR., WEBER, B. L., MARIS, J. M. & GRANT, G. R. 2006. STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res,* 16**,** 1149-58.

DISKIN, S. J., LI, M., HOU, C., YANG, S., GLESSNER, J., HAKONARSON, H., BUCAN, M., MARIS, J. M. & WANG, K. 2008. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res,* 36**,** e126.

DULAK, A. M., SCHUMACHER, S. E., VAN LIESHOUT, J., IMAMURA, Y., FOX, C., SHIM, B., RAMOS, A. H., SAKSENA, G., BACA, S. C., BASELGA, J., TABERNERO, J., BARRETINA, J., ENZINGER, P. C., CORSO, G., ROVIELLO, F., LIN, L., BANDLA, S., LUKETICH, J. D., PENNATHUR, A., MEYERSON, M., OGINO, S., SHIVDASANI, R. A., BEER, D. G., GODFREY, T. E., BEROUKHIM, R. & BASS, A. J. 2012. Gastrointestinal Adenocarcinomas of the Esophagus, Stomach, and Colon Exhibit Distinct Patterns of Genome Instability and Oncogenesis. *Cancer Research,* 72**,** 4383-4393.

DULAK, A. M., STOJANOV, P., PENG, S., LAWRENCE, M. S., FOX, C., STEWART, C., BANDLA, S., IMAMURA, Y., SCHUMACHER, S. E., SHEFLER, E., MCKENNA, A., CARTER, S. L., CIBULSKIS, K., SIVACHENKO, A., SAKSENA, G., VOET, D., RAMOS, A. H., AUCLAIR, D., THOMPSON, K., SOUGNEZ, C., ONOFRIO, R. C., GUIDUCCI, C., BEROUKHIM, R., ZHOU, Z., LIN, L., LIN, J., REDDY, R., CHANG, A., LANDRENAU, R., PENNATHUR, A., OGINO, S., LUKETICH, J. D., GOLUB, T. R., GABRIEL, S. B., LANDER, E. S., BEER, D. G., GODFREY, T. E., GETZ, G. & BASS, A. J. 2013. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet,* 45**,** 478-86.

EL-KEBIR, M., OESPER, L., ACHESON-FIELD, H. & RAPHAEL, B. J. 2015. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics,* 31**,** i62-70.

ELEVELD, T. F., OLDRIDGE, D. A., BERNARD, V., KOSTER, J., COLMET DAAGE, L., DISKIN, S. J., SCHILD, L., BENTAHAR, N. B., BELLINI, A., CHICARD, M., LAPOUBLE, E., COMBARET, V., LEGOIX-NE, P., MICHON, J., PUGH, T. J., HART, L. S., RADER, J., ATTIYEH, E. F., WEI, J. S., ZHANG, S., NARANJO, A., GASTIER-FOSTER, J. M., HOGARTY, M. D., ASGHARZADEH, S., SMITH, M. A., GUIDRY AUVIL, J. M., WATKINS, T. B., ZWIJNENBURG, D. A., EBUS, M. E., VAN SLUIS, P., HAKKERT, A., VAN WEZEL, E., VAN DER SCHOOT, C. E., WESTERHOUT, E. M., SCHULTE, J. H., TYTGAT, G. A., DOLMAN, M. E., JANOUEIX-LEROSEY, I., GERHARD, D. S., CARON, H. N., DELATTRE, O., KHAN, J., VERSTEEG, R.,

SCHLEIERMACHER, G., MOLENAAR, J. J. & MARIS, J. M. 2015. Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. *Nat Genet,* 47**,** 864-71.
ELIZALDE, S., LAUGHNEY, A. M. & BAKHOUM, S. F. 2018. A Markov chain for numerical chromosomal instability in clonally expanding populations. *PLoS Comput Biol,* 14**,** e1006447.
ERTYCH, N., STOLZ, A., STENZINGER, A., WEICHERT, W., KAULFUSS, S., BURFEIND, P., AIGNER, A., WORDEMAN, L. & BASTIANS, H. 2014. Increased microtubule assembly rates influence chromosomal instability in colorectal cancer cells. *Nat Cell Biol,* 16**,** 779-91.
FAN, J., LEE, H. O., LEE, S., RYU, D. E., LEE, S., XUE, C., KIM, S. J., KIM, K., BARKAS, N., PARK, P. J., PARK, W. Y. & KHARCHENKO, P. V. 2018. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res,* 28**,** 1217-1227.
FARABEGOLI, F., SANTINI, D., CECCARELLI, C., TAFFURELLI, M., MARRANO, D. & BALDINI, N. 2001. Clone heterogeneity in diploid and aneuploid breast carcinomas as detected by FISH. *Cytometry,* 46**,** 50-56.
FAVERO, F., JOSHI, T., MARQUARD, A. M., BIRKBAK, N. J., KRZYSTANEK, M., LI, Q., SZALLASI, Z. & EKLUND, A. C. 2015. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol,* 26**,** 64-70.
FERLAY, J., SOERJOMATARAM, I., DIKSHIT, R., ESER, S., MATHERS, C., REBELO, M., PARKIN, D. M., FORMAN, D. & BRAY, F. 2015. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer,* 136**,** E359-86.
FERRONIKA, P., VAN DEN BOS, H., TAUDT, A., SPIERINGS, D. C. J., SABER, A., HILTERMANN, T. J. N., KOK, K., PORUBSKY, D., VAN DER WEKKEN, A. J., TIMENS, W., FOIJER, F., COLOME-TATCHE, M., GROEN, H. J. M., LANSDORP, P. M. & VAN DEN BERG, A. 2017. Copy number alterations assessed at the single-cell level revealed mono- and polyclonal seeding patterns of distant metastasis in a small-cell lung cancer patient. *Ann Oncol,* 28**,** 1668-1670.
FINDLAY, J. M., CASTRO-GINER, F., MAKINO, S., RAYNER, E., KARTSONAKI, C., CROSS, W., KOVAC, M., ULAHANNAN, D., PALLES, C., GILLIES, R. S., MACGREGOR, T. P., CHURCH, D., MAYNARD, N. D., BUFFA, F., CAZIER, J. B., GRAHAM, T. A., WANG, L. M., SHARMA, R. A., MIDDLETON, M. & TOMLINSON, I. 2016. Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. *Nat Commun,* 7**,** 11111.
FISCHER, A., VAZQUEZ-GARCIA, I., ILLINGWORTH, C. J. & MUSTONEN, V. 2014. High-definition reconstruction of clonal composition in cancer. *Cell Rep,* 7**,** 1740-52.
FORD, C. E. & CLARKE, C. M. 1963. Cytogenetic Evidence of Clonal Proliferation in Primary Reticular Neoplasms. *Proc Can Cancer Conf,* 5**,** 129-46.
FOULDS, L. 1957. Tumor Progression: Guest Editorial. *Cancer Research,* 17**,** 355-356.

FRIEND, S. H., HOROWITZ, J. M., GERBER, M. R., WANG, X. F., BOGENMANN, E., LI, F. P. & WEINBERG, R. A. 1987. Deletions of a DNA sequence in retinoblastomas and mesenchymal tumors: organization of the sequence and its encoded protein. *Proc Natl Acad Sci U S A,* 84**,** 9059-63.

FROMER, M., MORAN, J. L., CHAMBERT, K., BANKS, E., BERGEN, S. E., RUDERFER, D. M., HANDSAKER, R. E., MCCARROLL, S. A., O'DONOVAN, M. C., OWEN, M. J., KIROV, G., SULLIVAN, P. F., HULTMAN, C. M., SKLAR, P. & PURCELL, S. M. 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet,* 91**,** 597-607.

FUJIWARA, T., BANDI, M., NITTA, M., IVANOVA, E. V., BRONSON, R. T. & PELLMAN, D. 2005. Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. *Nature,* 437**,** 1043-7.

FUNG, Y. K., MURPHREE, A. L., T'ANG, A., QIAN, J., HINRICHS, S. H. & BENEDICT, W. F. 1987. Structural evidence for the authenticity of the human retinoblastoma gene. *Science,* 236**,** 1657-61.

GANEM, N. J., GODINHO, S. A. & PELLMAN, D. 2009. A mechanism linking extra centrosomes to chromosomal instability. *Nature,* 460**,** 278-282.

GAO, R., DAVIS, A., MCDONALD, T. O., SEI, E., SHI, X., WANG, Y., TSAI, P. C., CASASENT, A., WATERS, J., ZHANG, H., MERIC-BERNSTAM, F., MICHOR, F. & NAVIN, N. E. 2016. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet,* 48**,** 1119-30.

GARTNER, J. J., DAVIS, S., WEI, X., LIN, J. C., TRIVEDI, N. S., TEER, J. K., PROGRAM, N. C. S., MELTZER, P. S., ROSENBERG, S. A. & SAMUELS, Y. 2012. Comparative exome sequencing of metastatic lesions provides insights into the mutational progression of melanoma. *BMC Genomics,* 13**,** 505.

GAWAD, C., KOH, W. & QUAKE, S. R. 2014. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc Natl Acad Sci U S A,* 111**,** 17947-52.

GAWAD, C., KOH, W. & QUAKE, S. R. 2016. Single-cell genome sequencing: current state of the science. *Nat Rev Genet,* 17**,** 175-88.

GERLINGER, M., HORSWELL, S., LARKIN, J., ROWAN, A. J., SALM, M. P., VARELA, I., FISHER, R., MCGRANAHAN, N., MATTHEWS, N., SANTOS, C. R., MARTINEZ, P., PHILLIMORE, B., BEGUM, S., RABINOWITZ, A., SPENCER-DENE, B., GULATI, S., BATES, P. A., STAMP, G., PICKERING, L., GORE, M., NICOL, D. L., HAZELL, S., FUTREAL, P. A., STEWART, A. & SWANTON, C. 2014. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet,* 46**,** 225-33.

GERLINGER, M., ROWAN, A. J., HORSWELL, S., LARKIN, J., ENDESFELDER, D., GRONROOS, E., MARTINEZ, P., MATTHEWS, N., STEWART, A., TARPEY, P., VARELA, I., PHILLIMORE, B., BEGUM, S., MCDONALD, N. Q., BUTLER, A., JONES, D., RAINE, K., LATIMER, C., SANTOS, C. R., NOHADANI, M., EKLUND, A. C., SPENCER-DENE, B., CLARK, G., PICKERING, L., STAMP, G., GORE, M., SZALLASI, Z., DOWNWARD, J., FUTREAL, P. A. &

SWANTON, C. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med,* 366**,** 883-92.
GIAM, M. & RANCATI, G. 2015. Aneuploidy and chromosomal instability in cancer: a jackpot to chaos. *Cell Div,* 10**,** 3.
GIBSON, W. J., HOIVIK, E. A., HALLE, M. K., TAYLOR-WEINER, A., CHERNIACK, A. D., BERG, A., HOLST, F., ZACK, T. I., WERNER, H. M., STABY, K. M., ROSENBERG, M., STEFANSSON, I. M., KUSONMANO, K., CHEVALIER, A., MAULAND, K. K., TROVIK, J., KRAKSTAD, C., GIANNAKIS, M., HODIS, E., WOIE, K., BJORGE, L., VINTERMYR, O. K., WALA, J. A., LAWRENCE, M. S., GETZ, G., CARTER, S. L., BEROUKHIM, R. & SALVESEN, H. B. 2016. The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. *Nat Genet,* 48**,** 848-55.
GODEK, K. M., KABECHE, L. & COMPTON, D. A. 2015. Regulation of kinetochore-microtubule attachments through homeostatic control during mitosis. *Nat Rev Mol Cell Biol,* 16**,** 57-64.
GOH, J. Y., FENG, M., WANG, W., OGUZ, G., YATIM, S. M. J. M., LEE, P. L., BAO, Y., LIM, T. H., WANG, P., TAM, W. L., KODAHL, A. R., LYNG, M. B., SARMA, S., LIN, S. Y., LEZHAVA, A., YAP, Y. S., LIM, A. S. T., HOON, D. S. B., DITZEL, H. J., LEE, S. C., TAN, E. Y. & YU, Q. 2017. Chromosome 1q21.3 amplification is a trackable biomarker and actionable target for breast cancer recurrence.
GORLA, G. R., MALHI, H. & GUPTA, S. 2001. Polyploidy associated with oxidative injury attenuates proliferative potential of cells. *J Cell Sci,* 114**,** 2943-51.
GOVINDAN, R., DING, L., GRIFFITH, M., SUBRAMANIAN, J., DEES, N. D., KANCHI, K. L., MAHER, C. A., FULTON, R., FULTON, L., WALLIS, J., CHEN, K., WALKER, J., MCDONALD, S., BOSE, R., ORNITZ, D., XIONG, D., YOU, M., DOOLING, D. J., WATSON, M., MARDIS, E. R. & WILSON, R. K. 2012. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell,* 150**,** 1121-34.
GRAHAM, N. A., MINASYAN, A., LOMOVA, A., CASS, A., BALANIS, N. G., FRIEDMAN, M., CHAN, S., ZHAO, S., DELGADO, A., GO, J., BECK, L., HURTZ, C., NG, C., QIAO, R., TEN HOEVE, J., PALASKAS, N., WU, H., MUSCHEN, M., MULTANI, A. S., PORT, E., LARSON, S. M., SCHULTZ, N., BRAAS, D., CHRISTOFK, H. R., MELLINGHOFF, I. K. & GRAEBER, T. G. 2017. Recurrent patterns of DNA copy number alterations in tumors reflect metabolic selection pressures. *Mol Syst Biol,* 13**,** 914.
GREAVES, M. 2015. Evolutionary Determinants of Cancer. *Cancer Discov,* 5**,** 806-20.
GREAVES, M. & MALEY, C. C. 2012. Clonal evolution in cancer. *Nature,* 481**,** 306-13.
GREENMAN, C. D., BIGNELL, G., BUTLER, A., EDKINS, S., HINTON, J., BEARE, D., SWAMY, S., SANTARIUS, T., CHEN, L., WIDAA, S., FUTREAL, P. A. & STRATTON, M. R. 2010. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics,* 11**,** 164-75.
GULCHER, J. 2012. Microsatellite markers for linkage and association studies. *Cold Spring Harb Protoc,* 2012**,** 425-32.

GUNDEM, G., VAN LOO, P., KREMEYER, B., ALEXANDROV, L. B., TUBIO, J. M. C., PAPAEMMANUIL, E., BREWER, D. S., KALLIO, H. M. L., HOGNAS, G., ANNALA, M., KIVINUMMI, K., GOODY, V., LATIMER, C., O'MEARA, S., DAWSON, K. J., ISAACS, W., EMMERT-BUCK, M. R., NYKTER, M., FOSTER, C., KOTE-JARAI, Z., EASTON, D., WHITAKER, H. C., GROUP, I. P., NEAL, D. E., COOPER, C. S., EELES, R. A., VISAKORPI, T., CAMPBELL, P. J., MCDERMOTT, U., WEDGE, D. C. & BOVA, G. S. 2015. The evolutionary history of lethal metastatic prostate cancer. *Nature,* 520**,** 353-357.

HA, G., ROTH, A., KHATTRA, J., HO, J., YAP, D., PRENTICE, L. M., MELNYK, N., MCPHERSON, A., BASHASHATI, A., LAKS, E., BIELE, J., DING, J., LE, A., ROSNER, J., SHUMANSKY, K., MARRA, M. A., GILKS, C. B., HUNTSMAN, D. G., MCALPINE, J. N., APARICIO, S. & SHAH, S. P. 2014. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res,* 24**,** 1881-93.

HAJDU, S. I. 2004. A note from history: the first tumor pathologist. *Ann Clin Lab Sci,* 34**,** 355-6.

HAJIRASOULIHA, I., MAHMOODY, A. & RAPHAEL, B. J. 2014. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics,* 30**,** i78-86.

HAKIMI, A. A., OSTROVNAYA, I., REVA, B., SCHULTZ, N., CHEN, Y. B., GONEN, M., LIU, H., TAKEDA, S., VOSS, M. H., TICKOO, S. K., REUTER, V. E., RUSSO, P., CHENG, E. H., SANDER, C., MOTZER, R. J., HSIEH, J. J. & CC, R. C. C. C. G. A. R. N. I. 2013. Adverse outcomes in clear cell renal cell carcinoma with mutations of 3p21 epigenetic regulators BAP1 and SETD2: a report by MSKCC and the KIRC TCGA research network. *Clin Cancer Res,* 19**,** 3259-67.

HARBST, K., LAUSS, M., CIRENAJWIS, H., ISAKSSON, K., ROSENGREN, F., TORNGREN, T., KVIST, A., JOHANSSON, M. C., VALLON-CHRISTERSSON, J., BALDETORP, B., BORG, A., OLSSON, H., INGVAR, C., CARNEIRO, A. & JONSSON, G. 2016. Multiregion Whole-Exome Sequencing Uncovers the Genetic Evolution and Mutational Heterogeneity of Early-Stage Metastatic Melanoma. *Cancer Res,* 76**,** 4765-74.

HARRIS, H. & WATKINS, J. F. 1965. Hybrid Cells Derived from Mouse and Man: Artificial Heterokaryons of Mammalian Cells from Different Species. *Nature,* 205**,** 640-6.

HARTIGAN, J. A. & HARTIGAN, P. M. 1985. The Dip Test of Unimodality. *The Annals of Statistics,* 13**,** 70-84.

HAUSCHKA, T. S. 1961. The chromosomes in ontogeny and oncogeny. *Cancer Res,* 21**,** 957-74.

HEDBERG, M. L., GOH, G., CHIOSEA, S. I., BAUMAN, J. E., FREILINO, M. L., ZENG, Y., WANG, L., DIERGAARDE, B. B., GOODING, W. E., LUI, V. W., HERBST, R. S., LIFTON, R. P. & GRANDIS, J. R. 2016. Genetic landscape of metastatic and recurrent head and neck squamous cell carcinoma. *J Clin Invest,* 126**,** 1606.

HEIDE, T., ZAPATA, L., WILLIAMS, M. J., WERNER, B., CARAVAGNA, G., BARNES, C. P., GRAHAM, T. A. & SOTTORIVA, A. 2018. Reply to 'Neutral tumor evolution?'. *Nat Genet,* 50**,** 1633-1637.

HIERONYMUS, H., MURALI, R., TIN, A., YADAV, K., ABIDA, W., MOLLER, H., BERNEY, D., SCHER, H., CARVER, B., SCARDINO, P., SCHULTZ, N., TAYLOR, B., VICKERS, A., CUZICK, J. & SAWYERS, C. L. 2018. Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. *Elife,* 7.

HOADLEY, K. A., YAU, C., WOLF, D. M., CHERNIACK, A. D., TAMBORERO, D., NG, S., LEISERSON, M. D. M., NIU, B., MCLELLAN, M. D., UZUNANGELOV, V., ZHANG, J., KANDOTH, C., AKBANI, R., SHEN, H., OMBERG, L., CHU, A., MARGOLIN, A. A., VAN'T VEER, L. J., LOPEZ-BIGAS, N., LAIRD, P. W., RAPHAEL, B. J., DING, L., ROBERTSON, A. G., BYERS, L. A., MILLS, G. B., WEINSTEIN, J. N., VAN WAES, C., CHEN, Z., COLLISSON, E. A., CANCER GENOME ATLAS RESEARCH, N., BENZ, C. C., PEROU, C. M. & STUART, J. M. 2014. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell,* 158**,** 929-944.

HOROWITZ, J. M., YANDELL, D. W., PARK, S. H., CANNING, S., WHYTE, P., BUCHKOVICH, K., HARLOW, E., WEINBERG, R. A. & DRYJA, T. P. 1989. Point mutational inactivation of the retinoblastoma antioncogene. *Science,* 243**,** 937-40.

HOU, Y., SONG, L., ZHU, P., ZHANG, B., TAO, Y., XU, X., LI, F., WU, K., LIANG, J., SHAO, D., WU, H., YE, X., YE, C., WU, R., JIAN, M., CHEN, Y., XIE, W., ZHANG, R., CHEN, L., LIU, X., YAO, X., ZHENG, H., YU, C., LI, Q., GONG, Z., MAO, M., YANG, X., YANG, L., LI, J., WANG, W., LU, Z., GU, N., LAURIE, G., BOLUND, L., KRISTIANSEN, K., WANG, J., YANG, H., LI, Y., ZHANG, X. & WANG, J. 2012. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell,* 148**,** 873-85.

HOWIE, B. N., DONNELLY, P. & MARCHINI, J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet,* 5**,** e1000529.

HUNG, J. 1995. Allele-specific chromosome 3p deletions occur at an early stage in the pathogenesis of lung carcinoma. *JAMA: The Journal of the American Medical Association,* 273**,** 558-563.

HUR, Y. & LEE, H. 2011. Wavelet-based identification of DNA focal genomic aberrations from single nucleotide polymorphism arrays. *BMC Bioinformatics,* 12**,** 146.

JACKSON, J. L., SANFORD, K. K. & DUNN, T. B. 1970. Neoplastic conversion and chromosomal characteristics of rat embryo cells in vitro. *J Natl Cancer Inst,* 45**,** 11-23.

JAHN, K., KUIPERS, J. & BEERENWINKEL, N. 2016. Tree inference for single-cell data. *Genome Biol,* 17**,** 86.

JAMAL-HANJANI, M., A'HERN, R., BIRKBAK, N. J., GORMAN, P., GRONROOS, E., NGANG, S., NICOLA, P., RAHMAN, L., THANOPOULOU, E., KELLY, G., ELLIS, P., BARRETT-LEE, P., JOHNSTON, S. R., BLISS, J., ROYLANCE, R. & SWANTON, C. 2015. Extreme chromosomal instability forecasts improved outcome in ER-negative breast cancer: a prospective validation cohort study from the TACT trial. *Ann Oncol,* 26**,** 1340-6.

JAMAL-HANJANI, M., WILSON, G. A., MCGRANAHAN, N., BIRKBAK, N. J., WATKINS, T. B. K., VEERIAH, S., SHAFI, S., JOHNSON, D. H.,

MITTER, R., ROSENTHAL, R., SALM, M., HORSWELL, S., ESCUDERO, M., MATTHEWS, N., ROWAN, A., CHAMBERS, T., MOORE, D. A., TURAJLIC, S., XU, H., LEE, S. M., FORSTER, M. D., AHMAD, T., HILEY, C. T., ABBOSH, C., FALZON, M., BORG, E., MARAFIOTI, T., LAWRENCE, D., HAYWARD, M., KOLVEKAR, S., PANAGIOTOPOULOS, N., JANES, S. M., THAKRAR, R., AHMED, A., BLACKHALL, F., SUMMERS, Y., SHAH, R., JOSEPH, L., QUINN, A. M., CROSBIE, P. A., NAIDU, B., MIDDLETON, G., LANGMAN, G., TROTTER, S., NICOLSON, M., REMMEN, H., KERR, K., CHETTY, M., GOMERSALL, L., FENNELL, D. A., NAKAS, A., RATHINAM, S., ANAND, G., KHAN, S., RUSSELL, P., EZHIL, V., ISMAIL, B., IRVIN-SELLERS, M., PRAKASH, V., LESTER, J. F., KORNASZEWSKA, M., ATTANOOS, R., ADAMS, H., DAVIES, H., DENTRO, S., TANIERE, P., O'SULLIVAN, B., LOWE, H. L., HARTLEY, J. A., ILES, N., BELL, H., NGAI, Y., SHAW, J. A., HERRERO, J., SZALLASI, Z., SCHWARZ, R. F., STEWART, A., QUEZADA, S. A., LE QUESNE, J., VAN LOO, P., DIVE, C., HACKSHAW, A., SWANTON, C. & CONSORTIUM, T. R. 2017. Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med,* 376**,** 2109-2121.

JANISZEWSKA, M., LIU, L., ALMENDRO, V., KUANG, Y., PAWELETZ, C., SAKR, R. A., WEIGELT, B., HANKER, A. B., CHANDARLAPATY, S., KING, T. A., REIS-FILHO, J. S., ARTEAGA, C. L., PARK, S. Y., MICHOR, F. & POLYAK, K. 2015. In situ single-cell analysis identifies heterogeneity for PIK3CA mutation and HER2 amplification in HER2-positive breast cancer. *Nat Genet,* 47**,** 1212-9.

JANSSEN, A., VAN DER BURG, M., SZUHAI, K., KOPS, G. J. & MEDEMA, R. H. 2011. Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations. *Science,* 333**,** 1895-8.

JIANG, Y., OLDRIDGE, D. A., DISKIN, S. J. & ZHANG, N. R. 2015. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res,* 43**,** e39.

JIAO, W., VEMBU, S., DESHWAR, A. G., STEIN, L. & MORRIS, Q. 2014. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics,* 15**,** 35.

JOHNSON, B. E., MAZOR, T., HONG, C., BARNES, M., AIHARA, K., MCLEAN, C. Y., FOUSE, S. D., YAMAMOTO, S., UEDA, H., TATSUNO, K., ASTHANA, S., JALBERT, L. E., NELSON, S. J., BOLLEN, A. W., GUSTAFSON, W. C., CHARRON, E., WEISS, W. A., SMIRNOV, I. V., SONG, J. S., OLSHEN, A. B., CHA, S., ZHAO, Y., MOORE, R. A., MUNGALL, A. J., JONES, S. J., HIRST, M., MARRA, M. A., SAITO, N., ABURATANI, H., MUKASA, A., BERGER, M. S., CHANG, S. M., TAYLOR, B. S. & COSTELLO, J. F. 2014. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science,* 343**,** 189-93.

JONES, R. D., KERR, D. J., HARNETT, A. N., RANKIN, E. M., RAY, S. & KAYE, S. B. 1990. A pilot study of quinidine and epirubicin in the treatment of advanced breast cancer. *Br J Cancer,* 62**,** 133-5.

JURIC, D., CASTEL, P., GRIFFITH, M., GRIFFITH, O. L., WON, H. H., ELLIS, H., EBBESEN, S. H., AINSCOUGH, B. J., RAMU, A., IYER, G., SHAH,

R. H., HUYNH, T., MINO-KENUDSON, M., SGROI, D., ISAKOFF, S., THABET, A., ELAMINE, L., SOLIT, D. B., LOWE, S. W., QUADT, C., PETERS, M., DERTI, A., SCHEGEL, R., HUANG, A., MARDIS, E. R., BERGER, M. F., BASELGA, J. & SCALTRITI, M. 2015. Convergent loss of PTEN leads to clinical resistance to a PI(3)Kalpha inhibitor. *Nature,* 518**,** 240-4.

KALLIONIEMI, A., KALLIONIEMI, O. P., SUDAR, D., RUTOVITZ, D., GRAY, J. W., WALDMAN, F. & PINKEL, D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science,* 258**,** 818-21.

KAPUR, P., PENA-LLOPIS, S., CHRISTIE, A., ZHREBKER, L., PAVIA-JIMENEZ, A., RATHMELL, W. K., XIE, X. J. & BRUGAROLAS, J. 2013. Effects on survival of BAP1 and PBRM1 mutations in sporadic clear-cell renal-cell carcinoma: a retrospective analysis with independent validation. *Lancet Oncol,* 14**,** 159-167.

KIM, M., GANS, J. D., NOGUEIRA, C., WANG, A., PAIK, J. H., FENG, B., BRENNAN, C., HAHN, W. C., CORDON-CARDO, C., WAGNER, S. N., FLOTTE, T. J., DUNCAN, L. M., GRANTER, S. R. & CHIN, L. 2006. Comparative oncogenomics identifies NEDD9 as a melanoma metastasis gene. *Cell,* 125**,** 1269-81.

KIM, T. M., JUNG, S. H., AN, C. H., LEE, S. H., BAEK, I. P., KIM, M. S., PARK, S. W., RHEE, J. K., LEE, S. H. & CHUNG, Y. J. 2015. Subclonal Genomic Architectures of Primary and Metastatic Colorectal Cancer Based on Intratumoral Genetic Heterogeneity. *Clin Cancer Res,* 21**,** 4461-72.

KIMURA, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics,* 61**,** 893-903.

KLAMBAUER, G., SCHWARZBAUER, K., MAYR, A., CLEVERT, D. A., MITTERECKER, A., BODENHOFER, U. & HOCHREITER, S. 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res,* 40**,** e69.

KLEINHEINZ, K., BLUDAU, I., HUEBSCHMANN, D., HEINOLD, M., KENSCHE, P., GU, Z., LOPEZ, C., HUMMEL, M., KLAPPER, W., MOELLER, P., VATER, I., WAGENER, R., BRORS, B., SIEBERT, R., EILS, R. & SCHLESNER, M. 2017. ACEseq - allele specific copy number estimation from whole genome sequencing. *bioRxiv*.

KNUDSON, A. G., JR. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A,* 68**,** 820-3.

KOBOLDT, D. C., ZHANG, Q., LARSON, D. E., SHEN, D., MCLELLAN, M. D., LIN, L., MILLER, C. A., MARDIS, E. R., DING, L. & WILSON, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res,* 22**,** 568-76.

KOREN, A., POLAK, P., NEMESH, J., MICHAELSON, J. J., SEBAT, J., SUNYAEV, S. R. & MCCARROLL, S. A. 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet,* 91**,** 1033-40.

KRUMM, N., SUDMANT, P. H., KO, A., O'ROAK, B. J., MALIG, M., COE, B. P., PROJECT, N. E. S., QUINLAN, A. R., NICKERSON, D. A. &

EICHLER, E. E. 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Res,* 22**,** 1525-32.

KUILMAN, T., VELDS, A., KEMPER, K., RANZANI, M., BOMBARDELLI, L., HOOGSTRAAT, M., NEVEDOMSKAYA, E., XU, G., DE RUITER, J., LOLKEMA, M. P., YLSTRA, B., JONKERS, J., ROTTENBERG, S., WESSELS, L. F., ADAMS, D. J., PEEPER, D. S. & KRIJGSMAN, O. 2015. CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol,* 16**,** 49.

LAMY, P., NORDENTOFT, I., BIRKENKAMP-DEMTRODER, K., THOMSEN, M. B., VILLESEN, P., VANG, S., HEDEGAARD, J., BORRE, M., JENSEN, J. B., HOYER, S., PEDERSEN, J. S., ORNTOFT, T. F. & DYRSKJOT, L. 2016. Paired Exome Analysis Reveals Clonal Evolution and Potential Therapeutic Targets in Urothelial Carcinoma. *Cancer Res,* 76**,** 5894-5906.

LANDAU, D. A., CARTER, S. L., STOJANOV, P., MCKENNA, A., STEVENSON, K., LAWRENCE, M. S., SOUGNEZ, C., STEWART, C., SIVACHENKO, A., WANG, L., WAN, Y., ZHANG, W., SHUKLA, S. A., VARTANOV, A., FERNANDES, S. M., SAKSENA, G., CIBULSKIS, K., TESAR, B., GABRIEL, S., HACOHEN, N., MEYERSON, M., LANDER, E. S., NEUBERG, D., BROWN, J. R., GETZ, G. & WU, C. J. 2013. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell,* 152**,** 714-26.

LANDAU, D. A., TAUSCH, E., TAYLOR-WEINER, A. N., STEWART, C., REITER, J. G., BAHLO, J., KLUTH, S., BOZIC, I., LAWRENCE, M., BOTTCHER, S., CARTER, S. L., CIBULSKIS, K., MERTENS, D., SOUGNEZ, C. L., ROSENBERG, M., HESS, J. M., EDELMANN, J., KLESS, S., KNEBA, M., RITGEN, M., FINK, A., FISCHER, K., GABRIEL, S., LANDER, E. S., NOWAK, M. A., DOHNER, H., HALLEK, M., NEUBERG, D., GETZ, G., STILGENBAUER, S. & WU, C. J. 2015. Mutations driving CLL and their evolution in progression and relapse. *Nature,* 526**,** 525-30.

LAUGHNEY, A. M., ELIZALDE, S., GENOVESE, G. & BAKHOUM, S. F. 2015. Dynamics of Tumor Heterogeneity Derived from Clonal Karyotypic Evolution. *Cell Rep,* 12**,** 809-20.

LAWRENCE, M. S., STOJANOV, P., POLAK, P., KRYUKOV, G. V., CIBULSKIS, K., SIVACHENKO, A., CARTER, S. L., STEWART, C., MERMEL, C. H., ROBERTS, S. A., KIEZUN, A., HAMMERMAN, P. S., MCKENNA, A., DRIER, Y., ZOU, L., RAMOS, A. H., PUGH, T. J., STRANSKY, N., HELMAN, E., KIM, J., SOUGNEZ, C., AMBROGIO, L., NICKERSON, E., SHEFLER, E., CORTES, M. L., AUCLAIR, D., SAKSENA, G., VOET, D., NOBLE, M., DICARA, D., LIN, P., LICHTENSTEIN, L., HEIMAN, D. I., FENNELL, T., IMIELINSKI, M., HERNANDEZ, B., HODIS, E., BACA, S., DULAK, A. M., LOHR, J., LANDAU, D. A., WU, C. J., MELENDEZ-ZAJGLA, J., HIDALGO-MIRANDA, A., KOREN, A., MCCARROLL, S. A., MORA, J., CROMPTON, B., ONOFRIO, R., PARKIN, M., WINCKLER, W., ARDLIE, K., GABRIEL, S. B., ROBERTS, C. W. M., BIEGEL, J. A., STEGMAIER, K., BASS, A. J., GARRAWAY, L. A., MEYERSON, M., GOLUB, T. R., GORDENIN, D. A., SUNYAEV, S., LANDER, E. S. &

GETZ, G. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature,* 499**,** 214-218.
LAZZERINI DENCHI, E., CELLI, G. & DE LANGE, T. 2006. Hepatocytes with extensive telomere deprotection and fusion remain viable and regenerate liver mass through endoreduplication. *Genes Dev,* 20**,** 2648-53.
LEACH, N. T., REHDER, C., JENSEN, K., HOLT, S. & JACKSON-COOK, C. 2004. Human chromosomes with shorter telomeres and large heterochromatin regions have a higher frequency of acquired somatic cell aneuploidy. *Mech Ageing Dev,* 125**,** 563-73.
LEE, H. O., DAVIDSON, J. M. & DURONIO, R. J. 2009. Endoreplication: polyploidy with purpose. *Genes Dev,* 23**,** 2461-77.
LEE, H. S., LEE, N. C., KOUPRINA, N., KIM, J. H., KAGANSKY, A., BATES, S., TREPEL, J. B., POMMIER, Y., SACKETT, D. & LARIONOV, V. 2016. Effects of Anticancer Drugs on Chromosome Instability and New Clinical Implications for Tumor-Suppressing Therapies. *Cancer Res,* 76**,** 902-11.
LEE, W. H., BOOKSTEIN, R., WHEATLEY, W., BENEDICT, W. F. & LEE, E. Y. 1987. A null allele of esterase D is a marker for genetic events in retinoblastoma formation. *Hum Genet,* 76**,** 33-6.
LETOUZE, E., ALLORY, Y., BOLLET, M. A., RADVANYI, F. & GUYON, F. 2010. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biol,* 11**,** R76.
LI, J., LUPAT, R., AMARASINGHE, K. C., THOMPSON, E. R., DOYLE, M. A., RYLAND, G. L., TOTHILL, R. W., HALGAMUGE, S. K., CAMPBELL, I. G. & GORRINGE, K. L. 2012a. CONTRA: copy number analysis for targeted resequencing. *Bioinformatics,* 28**,** 1307-13.
LI, M., FANG, X., BAKER, D. J., GUO, L., GAO, X., WEI, Z., HAN, S., VAN DEURSEN, J. M. & ZHANG, P. 2010. The ATM-p53 pathway suppresses aneuploidy-induced tumorigenesis. *Proc Natl Acad Sci U S A,* 107**,** 14188-93.
LI, Y., XU, X., SONG, L., HOU, Y., LI, Z., TSANG, S., LI, F., IM, K. M., WU, K., WU, H., YE, X., LI, G., WANG, L., ZHANG, B., LIANG, J., XIE, W., WU, R., JIANG, H., LIU, X., YU, C., ZHENG, H., JIAN, M., NIE, L., WAN, L., SHI, M., SUN, X., TANG, A., GUO, G., GUI, Y., CAI, Z., LI, J., WANG, W., LU, Z., ZHANG, X., BOLUND, L., KRISTIANSEN, K., WANG, J., YANG, H., DEAN, M. & WANG, J. 2012b. Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *Gigascience,* 1**,** 12.
LICHTER, P., CREMER, T., TANG, C. J., WATKINS, P. C., MANUELIDIS, L. & WARD, D. C. 1988. Rapid detection of human chromosome 21 aberrations by in situ hybridization. *Proc Natl Acad Sci U S A,* 85**,** 9664-8.
LIM, C. H., CHO, Y. K., KIM, S. W., CHOI, M. G., RHEE, J. K., CHUNG, Y. J., LEE, S. H. & KIM, T. M. 2016. The chronological sequence of somatic mutations in early gastric carcinogenesis inferred from multiregion sequencing of gastric adenomas. *Oncotarget,* 7**,** 39758-39767.
LIU, Y., CHEN, C., XU, Z., SCUOPPO, C., RILLAHAN, C. D., GAO, J., SPITZER, B., BOSBACH, B., KASTENHUBER, E. R., BASLAN, T.,

ACKERMANN, S., CHENG, L., WANG, Q., NIU, T., SCHULTZ, N., LEVINE, R. L., MILLS, A. A. & LOWE, S. W. 2016. Deletions linked to TP53 loss drive cancer through p53-independent mechanisms. *Nature,* 531**,** 471-475.

LOH, P. R., GENOVESE, G., HANDSAKER, R. E., FINUCANE, H. K., RESHEF, Y. A., PALAMARA, P. F., BIRMANN, B. M., TALKOWSKI, M. E., BAKHOUM, S. F., MCCARROLL, S. A. & PRICE, A. L. 2018. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature,* 559**,** 350-355.

LOHR, J. G., STOJANOV, P., CARTER, S. L., CRUZ-GORDILLO, P., LAWRENCE, M. S., AUCLAIR, D., SOUGNEZ, C., KNOECHEL, B., GOULD, J., SAKSENA, G., CIBULSKIS, K., MCKENNA, A., CHAPMAN, M. A., STRAUSSMAN, R., LEVY, J., PERKINS, L. M., KEATS, J. J., SCHUMACHER, S. E., ROSENBERG, M., MULTIPLE MYELOMA RESEARCH, C., GETZ, G. & GOLUB, T. R. 2014. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell,* 25**,** 91-101.

LOTHSCHUTZ, D., JENNEWEIN, M., PAHL, S., LAUSBERG, H. F., EICHLER, A., MUTSCHLER, W., HANSELMANN, R. G. & OBERRINGER, M. 2002. Polyploidization and centrosome hyperamplification in inflammatory bronchi. *Inflamm Res,* 51**,** 416-22.

LOVE, M. I., MYSICKOVA, A., SUN, R., KALSCHEUER, V., VINGRON, M. & HAAS, S. A. 2011. Modeling read counts for CNV detection in exome sequencing data. *Stat Appl Genet Mol Biol,* 10.

LY, P., TEITZ, L. S., KIM, D. H., SHOSHANI, O., SKALETSKY, H., FACHINETTI, D., PAGE, D. C. & CLEVELAND, D. W. 2017. Selective Y centromere inactivation triggers chromosome shattering in micronuclei and repair by non-homologous end joining. *Nat Cell Biol,* 19**,** 68-75.

LYNCH, T. J., BELL, D. W., SORDELLA, R., GURUBHAGAVATULA, S., OKIMOTO, R. A., BRANNIGAN, B. W., HARRIS, P. L., HASERLAT, S. M., SUPKO, J. G., HALUSKA, F. G., LOUIS, D. N., CHRISTIANI, D. C., SETTLEMAN, J. & HABER, D. A. 2004. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med,* 350**,** 2129-39.

MACIEJOWSKI, J., LI, Y., BOSCO, N., CAMPBELL, P. J. & DE LANGE, T. 2015. Chromothripsis and Kataegis Induced by Telomere Crisis. *Cell,* 163**,** 1641-54.

MACINTYRE, G., GORANOVA, T. E., DE SILVA, D., ENNIS, D., PISKORZ, A. M., ELDRIDGE, M., SIE, D., LEWSLEY, L. A., HANIF, A., WILSON, C., DOWSON, S., GLASSPOOL, R. M., LOCKLEY, M., BROCKBANK, E., MONTES, A., WALTHER, A., SUNDAR, S., EDMONDSON, R., HALL, G. D., CLAMP, A., GOURLEY, C., HALL, M., FOTOPOULOU, C., GABRA, H., PAUL, J., SUPERNAT, A., MILLAN, D., HOYLE, A., BRYSON, G., NOURSE, C., MINCARELLI, L., SANCHEZ, L. N., YLSTRA, B., JIMENEZ-LINAN, M., MOORE, L., HOFMANN, O., MARKOWETZ, F., MCNEISH, I. A. & BRENTON, J. D. 2018. Copy number signatures and mutational processes in ovarian carcinoma. *Nat Genet,* 50**,** 1262-1270.

MAECHLER, M. 2015. Diptest: Hartigan's dip test statistic for unimodality—corrected. *R package version 0.75-7. See https://cran.r-project.org/package=diptest*.
MAGI, A., TATTINI, L., CIFOLA, I., D'AURIZIO, R., BENELLI, M., MANGANO, E., BATTAGLIA, C., BONORA, E., KURG, A., SERI, M., MAGINI, P., GIUSTI, B., ROMEO, G., PIPPUCCI, T., DE BELLIS, G., ABBATE, R. & GENSINI, G. F. 2013. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol,* 14**,** R120.
MAKINO, S. 1957. The Chromosome Cytology of the Ascites Tumors of Rats, with Special

Reference to the Concept of the Stemline Cell. *International Review of Cytology,* 6**,** 25–84.
MALEY, C. C., GALIPEAU, P. C., FINLEY, J. C., WONGSURAWAT, V. J., LI, X., SANCHEZ, C. A., PAULSON, T. G., BLOUNT, P. L., RISQUES, R. A., RABINOVITCH, P. S. & REID, B. J. 2006. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet,* 38**,** 468-73.
MALIKIC, S., MCPHERSON, A. W., DONMEZ, N. & SAHINALP, C. S. 2015. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics,* 31**,** 1349-56.
MARTELOTTO, L. G., BASLAN, T., KENDALL, J., GEYER, F. C., BURKE, K. A., SPRAGGON, L., PISCUOGLIO, S., CHADALAVADA, K., NANJANGUD, G., NG, C. K., MOODY, P., D'ITALIA, S., RODGERS, L., COX, H., DA CRUZ PAULA, A., STEPANSKY, A., SCHIZAS, M., WEN, H. Y., KING, T. A., NORTON, L., WEIGELT, B., HICKS, J. B. & REIS-FILHO, J. S. 2017. Whole-genome single-cell copy number profiling from formalin-fixed paraffin-embedded samples. *Nat Med,* 23**,** 376-385.
MARTINCORENA, I., RAINE, K. M., GERSTUNG, M., DAWSON, K. J., HAASE, K., VAN LOO, P., DAVIES, H., STRATTON, M. R. & CAMPBELL, P. J. 2017. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*.
MARTINEZ, P., BIRKBAK, N. J., GERLINGER, M., MCGRANAHAN, N., BURRELL, R. A., ROWAN, A. J., JOSHI, T., FISHER, R., LARKIN, J., SZALLASI, Z. & SWANTON, C. 2013. Parallel evolution of tumour subclones mimics diversity between tumours. *J Pathol,* 230**,** 356-64.
MCGRANAHAN, N., BURRELL, R. A., ENDESFELDER, D., NOVELLI, M. R. & SWANTON, C. 2012. Cancer chromosomal instability: therapeutic and diagnostic challenges. 'Exploring aneuploidy: the significance of chromosomal imbalance' review series. *EMBO Rep*.
MCGRANAHAN, N., FAVERO, F., DE BRUIN, E. C., BIRKBAK, N. J., SZALLASI, Z. & SWANTON, C. 2015. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci Transl Med,* 7**,** 283ra54.
MCGRANAHAN, N., FURNESS, A. J., ROSENTHAL, R., RAMSKOV, S., LYNGAA, R., SAINI, S. K., JAMAL-HANJANI, M., WILSON, G. A., BIRKBAK, N. J., HILEY, C. T., WATKINS, T. B., SHAFI, S., MURUGAESU, N., MITTER, R., AKARCA, A. U., LINARES, J., MARAFIOTI, T., HENRY, J. Y., VAN ALLEN, E. M., MIAO, D.,

SCHILLING, B., SCHADENDORF, D., GARRAWAY, L. A., MAKAROV, V., RIZVI, N. A., SNYDER, A., HELLMANN, M. D., MERGHOUB, T., WOLCHOK, J. D., SHUKLA, S. A., WU, C. J., PEGGS, K. S., CHAN, T. A., HADRUP, S. R., QUEZADA, S. A. & SWANTON, C. 2016. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science,* 351**,** 1463-9.

MCGRANAHAN, N., ROSENTHAL, R., HILEY, C. T., ROWAN, A. J., WATKINS, T. B. K., WILSON, G. A., BIRKBAK, N. J., VEERIAH, S., VAN LOO, P., HERRERO, J., SWANTON, C. & CONSORTIUM, T. R. 2017. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*.

MCGRANAHAN, N. & SWANTON, C. 2017. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell,* 168**,** 613-628.

MCPHERSON, A., ROTH, A., LAKS, E., MASUD, T., BASHASHATI, A., ZHANG, A. W., HA, G., BIELE, J., YAP, D., WAN, A., PRENTICE, L. M., KHATTRA, J., SMITH, M. A., NIELSEN, C. B., MULLALY, S. C., KALLOGER, S., KARNEZIS, A., SHUMANSKY, K., SIU, C., ROSNER, J., CHAN, H. L., HO, J., MELNYK, N., SENZ, J., YANG, W., MOORE, R., MUNGALL, A. J., MARRA, M. A., BOUCHARD-COTE, A., GILKS, C. B., HUNTSMAN, D. G., MCALPINE, J. N., APARICIO, S. & SHAH, S. P. 2016. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat Genet,* 48**,** 758-67.

MCPHERSON, A. W., ROTH, A., HA, G., CHAUVE, C., STEIF, A., DE SOUZA, C. P. E., EIREW, P., BOUCHARD-COTE, A., APARICIO, S., SAHINALP, S. C. & SHAH, S. P. 2017. ReMixT: clone-specific genomic structure estimation in cancer. *Genome Biol,* 18**,** 140.

MECKERT, P. C., RIVELLO, H. G., VIGLIANO, C., GONZALEZ, P., FAVALORO, R. & LAGUENS, R. 2005. Endomitosis and polyploidization of myocardial cells in the periphery of human acute myocardial infarction. *Cardiovasc Res,* 67**,** 116-23.

MEDVEDEV, P., FIUME, M., DZAMBA, M., SMITH, T. & BRUDNO, M. 2010. Detecting copy number variation with mated short reads. *Genome Res,* 20**,** 1613-22.

MERLO, L. M., SHAH, N. A., LI, X., BLOUNT, P. L., VAUGHAN, T. L., REID, B. J. & MALEY, C. C. 2010. A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. *Cancer Prev Res (Phila),* 3**,** 1388-97.

MERMEL, C. H., SCHUMACHER, S. E., HILL, B., MEYERSON, M. L., BEROUKHIM, R. & GETZ, G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol,* 12**,** R41.

METTU, R. K., WAN, Y. W., HABERMANN, J. K., RIED, T. & GUO, N. L. 2010. A 12-gene genomic instability signature predicts clinical outcomes in multiple cancer types. *Int J Biol Markers,* 25**,** 219-28.

METZKER, M. L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet,* 11**,** 31-46.

MILLER, C. A., WHITE, B. S., DEES, N. D., GRIFFITH, M., WELCH, J. S., GRIFFITH, O. L., VIJ, R., TOMASSON, M. H., GRAUBERT, T. A., WALTER, M. J., ELLIS, M. J., SCHIERDING, W., DIPERSIO, J. F., LEY, T. J., MARDIS, E. R., WILSON, R. K. & DING, L. 2014. SciClone:

inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol,* 10**,** e1003665.

MIRON, K., GOLAN-LEV, T., DVIR, R., BEN-DAVID, E. & KEREM, B. 2015. Oncogenes create a unique landscape of fragile sites. *Nat Commun,* 6**,** 7094.

MISALE, S., DI NICOLANTONIO, F., SARTORE-BIANCHI, A., SIENA, S. & BARDELLI, A. 2014. Resistance to anti-EGFR therapy in colorectal cancer: from heterogeneity to convergent evolution. *Cancer Discov,* 4**,** 1269-80.

MISALE, S., YAEGER, R., HOBOR, S., SCALA, E., JANAKIRAMAN, M., LISKA, D., VALTORTA, E., SCHIAVO, R., BUSCARINO, M., SIRAVEGNA, G., BENCARDINO, K., CERCEK, A., CHEN, C. T., VERONESE, S., ZANON, C., SARTORE-BIANCHI, A., GAMBACORTA, M., GALLICCHIO, M., VAKIANI, E., BOSCARO, V., MEDICO, E., WEISER, M., SIENA, S., DI NICOLANTONIO, F., SOLIT, D. & BARDELLI, A. 2012. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature,* 486**,** 532-6.

MITCHELL, T. J., TURAJLIC, S., ROWAN, A., NICOL, D., FARMERY, J. H. R., O'BRIEN, T., MARTINCORENA, I., TARPEY, P., ANGELOPOULOS, N., YATES, L. R., BUTLER, A. P., RAINE, K., STEWART, G. D., CHALLACOMBE, B., FERNANDO, A., LOPEZ, J. I., HAZELL, S., CHANDRA, A., CHOWDHURY, S., RUDMAN, S., SOULTATI, A., STAMP, G., FOTIADIS, N., PICKERING, L., AU, L., SPAIN, L., LYNCH, J., STARES, M., TEAGUE, J., MAURA, F., WEDGE, D. C., HORSWELL, S., CHAMBERS, T., LITCHFIELD, K., XU, H., STEWART, A., ELAIDI, R., OUDARD, S., MCGRANAHAN, N., CSABAI, I., GORE, M., FUTREAL, P. A., LARKIN, J., LYNCH, A. G., SZALLASI, Z., SWANTON, C., CAMPBELL, P. J. & CONSORTIUM, T. R. R. 2018. Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. *Cell,* 173**,** 611-623 e17.

MITELMAN, F. 2015. Mitelman database of chromosome aberrations and gene fusions in cancer. http://cgap.nci.nih.gov/Chromosomes/Mitelman.

MONZON, F. A., ALVAREZ, K., PETERSON, L., TRUONG, L., AMATO, R. J., HERNANDEZ-MCCLAIN, J., TANNIR, N., PARWANI, A. V. & JONASCH, E. 2011. Chromosome 14q loss defines a molecular subtype of clear-cell renal cell carcinoma associated with poor prognosis. *Mod Pathol,* 24**,** 1470-9.

MOORHEAD, P. S. & SAKSELA, E. 1965. The sequence of chromosome aberrations during SV 40 transformation of a human diploid cell strain. *Hereditas,* 52**,** 271-84.

MORSON, B. C. 1974. Evolution of cancer of the colon and rectum. *Cancer,* 34**,** 845–849.

MORTON, C. C., KIRSCH, I. R., TAUB, R., ORKIN, S. H. & BROWN, J. A. 1984. Localization of the beta-globin gene by chromosomal in situ hybridization. *Am J Hum Genet,* 36**,** 576-85.

MROZ, E. A. & ROCCO, J. W. 2013. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncol,* 49**,** 211-5.

MURPHREE, A. L. & BENEDICT, W. F. 1984. Retinoblastoma: clues to human oncogenesis. *Science,* 223**,** 1028-33.

MURUGAESU, N., WILSON, G. A., BIRKBAK, N. J., WATKINS, T., MCGRANAHAN, N., KUMAR, S., ABBASSI-GHADI, N., SALM, M., MITTER, R., HORSWELL, S., ROWAN, A., PHILLIMORE, B., BIGGS, J., BEGUM, S., MATTHEWS, N., HOCHHAUSER, D., HANNA, G. B. & SWANTON, C. 2015. Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discov,* 5**,** 821-831.

NADEU, F., DELGADO, J., ROYO, C., BAUMANN, T., STANKOVIC, T., PINYOL, M., JARES, P., NAVARRO, A., MARTIN-GARCIA, D., BEA, S., SALAVERRIA, I., OLDREIVE, C., AYMERICH, M., SUAREZ-CISNEROS, H., ROZMAN, M., VILLAMOR, N., COLOMER, D., LOPEZ-GUILLERMO, A., GONZALEZ, M., ALCOCEBA, M., TEROL, M. J., COLADO, E., PUENTE, X. S., LOPEZ-OTIN, C., ENJUANES, A. & CAMPO, E. 2016. Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. *Blood,* 127**,** 2122-30.

NAVIN, N., KENDALL, J., TROGE, J., ANDREWS, P., RODGERS, L., MCINDOO, J., COOK, K., STEPANSKY, A., LEVY, D., ESPOSITO, D., MUTHUSWAMY, L., KRASNITZ, A., MCCOMBIE, W. R., HICKS, J. & WIGLER, M. 2011. Tumour evolution inferred by single-cell sequencing. *Nature,* 472**,** 90-4.

NEGRINI, M., RASIO, D., HAMPTON, G. M., SABBIONI, S., RATTAN, S., CARTER, S. L., ROSENBERG, A. L., SCHWARTZ, G. F., SHILOH, Y., CAVENEE, W. K. & ET AL. 1995. Definition and refinement of chromosome 11 regions of loss of heterozygosity in breast cancer: identification of a new region at 11q23.3. *Cancer Res,* 55**,** 3003-7.

NEPUSZ, G. C. A. T. 2006. The igraph software package for complex network research. *InterJournal,* Complex Systems**,** 1695.

NIGRO, J. M., BAKER, S. J., PREISINGER, A. C., JESSUP, J. M., HOSTETTER, R., CLEARY, K., BIGNER, S. H., DAVIDSON, N., BAYLIN, S., DEVILEE, P. & ET AL. 1989. Mutations in the p53 gene occur in diverse human tumour types. *Nature,* 342**,** 705-8.

NIIDA, A., IWASAKI, W. M. & INNAN, H. 2018. Neutral Theory in Cancer Cell Population Genetics. *Mol Biol Evol,* 35**,** 1316-1321.

NIK-ZAINAL, S., VAN LOO, P., WEDGE, D. C., ALEXANDROV, L. B., GREENMAN, C. D., LAU, K. W., RAINE, K., JONES, D., MARSHALL, J., RAMAKRISHNA, M., SHLIEN, A., COOKE, S. L., HINTON, J., MENZIES, A., STEBBINGS, L. A., LEROY, C., JIA, M., RANCE, R., MUDIE, L. J., GAMBLE, S. J., STEPHENS, P. J., MCLAREN, S., TARPEY, P. S., PAPAEMMANUIL, E., DAVIES, H. R., VARELA, I., MCBRIDE, D. J., BIGNELL, G. R., LEUNG, K., BUTLER, A. P., TEAGUE, J. W., MARTIN, S., JONSSON, G., MARIANI, O., BOYAULT, S., MIRON, P., FATIMA, A., LANGEROD, A., APARICIO, S. A., TUTT, A., SIEUWERTS, A. M., BORG, A., THOMAS, G., SALOMON, A. V., RICHARDSON, A. L., BORRESEN-DALE, A. L., FUTREAL, P. A., STRATTON, M. R., CAMPBELL, P. J. & BREAST CANCER WORKING GROUP OF THE INTERNATIONAL CANCER GENOME, C. 2012. The life history of 21 breast cancers. *Cell,* 149**,** 994-1007.

NISHISHO, I., NAKAMURA, Y., MIYOSHI, Y., MIKI, Y., ANDO, H., HORII, A., KOYAMA, K., UTSUNOMIYA, J., BABA, S. & HEDGE, P. 1991. Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science,* 253**,** 665-9.

NONES, K., WADDELL, N., WAYTE, N., PATCH, A.-M., BAILEY, P., NEWELL, F., HOLMES, O., FINK, J. L., QUINN, M. C. J., TANG, Y. H., LAMPE, G., QUEK, K., LOFFLER, K. A., MANNING, S., IDRISOGLU, S., MILLER, D., XU, Q., WADDELL, N., WILSON, P. J., BRUXNER, T. J. C., CHRIST, A. N., HARLIWONG, I., NOURSE, C., NOURBAKHSH, E., ANDERSON, M., KAZAKOFF, S., LEONARD, C., WOOD, S., SIMPSON, P. T., REID, L. E., KRAUSE, L., HUSSEY, D. J., WATSON, D. I., LORD, R. V., NANCARROW, D., PHILLIPS, W. A., GOTLEY, D., SMITHERS, B. M., WHITEMAN, D. C., HAYWARD, N. K., CAMPBELL, P. J., PEARSON, J. V., GRIMMOND, S. M. & BARBOUR, A. P. 2014. Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nature Communications,* 5.

NORDLING, C. O. 1953. A new theory on cancer-inducing mechanism. *Br J Cancer,* 7**,** 68-72.

NOTTA, F., CHAN-SENG-YUE, M., LEMIRE, M., LI, Y., WILSON, G. W., CONNOR, A. A., DENROCHE, R. E., LIANG, S. B., BROWN, A. M., KIM, J. C., WANG, T., SIMPSON, J. T., BECK, T., BORGIDA, A., BUCHNER, N., CHADWICK, D., HAFEZI-BAKHTIARI, S., DICK, J. E., HEISLER, L., HOLLINGSWORTH, M. A., IBRAHIMOV, E., JANG, G. H., JOHNS, J., JORGENSEN, L. G., LAW, C., LUDKOVSKI, O., LUNGU, I., NG, K., PASTERNACK, D., PETERSEN, G. M., SHLUSH, L. I., TIMMS, L., TSAO, M. S., WILSON, J. M., YUNG, C. K., ZOGOPOULOS, G., BARTLETT, J. M., ALEXANDROV, L. B., REAL, F. X., CLEARY, S. P., ROEHRL, M. H., MCPHERSON, J. D., STEIN, L. D., HUDSON, T. J., CAMPBELL, P. J. & GALLINGER, S. 2016. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature,* 538**,** 378-382.

NOWELL, P. C. 1960. Phytohemagglutinin: an initiator of mitosis in cultures of normal human leukocytes. *Cancer Res,* 20**,** 462-6.

NOWELL, P. C. 1976. The clonal evolution of tumor cell populations. *Science,* 194**,** 23-8.

NOWELL, P. C. & HUNGERFORD, D. A. 1960. A minute chromosome in human chronic granulocytic leukemia *Science,* 132.

OESPER, L., MAHMOODY, A. & RAPHAEL, B. J. 2013. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol,* 14**,** R80.

ORR, B. & COMPTON, D. A. 2013. A double-edged sword: how oncogenes and tumor suppressor genes can contribute to chromosomal instability. *Front Oncol,* 3**,** 164.

ORTMANN, C. A., KENT, D. G., NANGALIA, J., SILBER, Y., WEDGE, D. C., GRINFELD, J., BAXTER, E. J., MASSIE, C. E., PAPAEMMANUIL, E., MENON, S., GODFREY, A. L., DIMITROPOULOU, D., GUGLIELMELLI, P., BELLOSILLO, B., BESSES, C., DÖHNER, K., HARRISON, C. N., VASSILIOU, G. S., VANNUCCHI, A., CAMPBELL, P. J. & GREEN, A. R. 2015. Effect of Mutation Order on

Myeloproliferative Neoplasms. *New England Journal of Medicine,* 372**,** 601-612.
OTTO, S. P. & WHITTON, J. 2000. Polyploid incidence and evolution. *Annu Rev Genet,* 34**,** 401-437.
PAINTER, T. S. 1923. Studies in mammalian spermatogenesis. II. The spermatogenesis of man. *Journal of Experimental Zoology,* 37**,** 291-336.
PAPAEMMANUIL, E., GERSTUNG, M., MALCOVATI, L., TAURO, S., GUNDEM, G., VAN LOO, P., YOON, C. J., ELLIS, P., WEDGE, D. C., PELLAGATTI, A., SHLIEN, A., GROVES, M. J., FORBES, S. A., RAINE, K., HINTON, J., MUDIE, L. J., MCLAREN, S., HARDY, C., LATIMER, C., DELLA PORTA, M. G., O'MEARA, S., AMBAGLIO, I., GALLI, A., BUTLER, A. P., WALLDIN, G., TEAGUE, J. W., QUEK, L., STERNBERG, A., GAMBACORTI-PASSERINI, C., CROSS, N. C., GREEN, A. R., BOULTWOOD, J., VYAS, P., HELLSTROM-LINDBERG, E., BOWEN, D., CAZZOLA, M., STRATTON, M. R., CAMPBELL, P. J. & CHRONIC MYELOID DISORDERS WORKING GROUP OF THE INTERNATIONAL CANCER GENOME, C. 2013. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood,* 122**,** 3616-27; quiz 3699.
PARADIS, E., CLAUDE, J. & STRIMMER, K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics,* 20**,** 289-90.
PASSERINI, V., OZERI-GALAI, E., DE PAGTER, M. S., DONNELLY, N., SCHMALBROCK, S., KLOOSTERMAN, W. P., KEREM, B. & STORCHOVA, Z. 2016. The presence of extra chromosomes leads to genomic instability. *Nat Commun,* 7**,** 10754.
PAULSSON, K. & JOHANSSON, B. 2007. Trisomy 8 as the sole chromosomal aberration in acute myeloid leukemia and myelodysplastic syndromes. *Pathol Biol (Paris),* 55**,** 37-48.
PAVELKA, N., RANCATI, G. & LI, R. 2010. Dr Jekyll and Mr Hyde: role of aneuploidy in cellular adaptation and cancer. *Curr Opin Cell Biol,* 22**,** 1-7.
PEROU, C. M., SORLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H., AKSLEN, L. A., FLUGE, O., PERGAMENSCHIKOV, A., WILLIAMS, C., ZHU, S. X., LONNING, P. E., BORRESEN-DALE, A. L., BROWN, P. O. & BOTSTEIN, D. 2000. Molecular portraits of human breast tumours. *Nature,* 406**,** 747-52.
PFEIFER, D., PANTIC, M., SKATULLA, I., RAWLUK, J., KREUTZ, C., MARTENS, U. M., FISCH, P., TIMMER, J. & VEELKEN, H. 2007. Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. *Blood,* 109**,** 1202-10.
PINKEL, D. & ALBERTSON, D. G. 2005. Array comparative genomic hybridization and its applications in cancer. *Nat Genet,* 37 Suppl**,** S11-7.
PLAGNOL, V., CURTIS, J., EPSTEIN, M., MOK, K. Y., STEBBINGS, E., GRIGORIADOU, S., WOOD, N. W., HAMBLETON, S., BURNS, S. O., THRASHER, A. J., KUMARARATNE, D., DOFFINGER, R. & NEJENTSEV, S. 2012. A robust model for read count data in exome

sequencing experiments and implications for copy number variant calling. *Bioinformatics,* 28**,** 2747-54.
POHL, H. & WELCH, H. G. 2005. The Role of Overdiagnosis and Reclassification in the Marked Increase of Esophageal Adenocarcinoma Incidence. *JNCI Journal of the National Cancer Institute,* 97**,** 142-146.
POLLACK, J. R., SORLIE, T., PEROU, C. M., REES, C. A., JEFFREY, S. S., LONNING, P. E., TIBSHIRANI, R., BOTSTEIN, D., BORRESEN-DALE, A. L. & BROWN, P. O. 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A,* 99**,** 12963-8.
POPIC, V., SALARI, R., HAJIRASOULIHA, I., KASHEF-HAGHIGHI, D., WEST, R. B. & BATZOGLOU, S. 2015. Fast and scalable inference of multi-sample cancer lineages. *Genome Biol,* 16**,** 91.
POPOVA, T., MANIE, E., RIEUNIER, G., CAUX-MONCOUTIER, V., TIRAPO, C., DUBOIS, T., DELATTRE, O., SIGAL-ZAFRANI, B., BOLLET, M., LONGY, M., HOUDAYER, C., SASTRE-GARAU, X., VINCENT-SALOMON, A., STOPPA-LYONNET, D. & STERN, M. H. 2012. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res,* 72**,** 5454-62.
POPOVA, T., MANIÉ, E., STOPPA-LYONNET, D., RIGAILL, G., BARILLOT, E. & STERN, M. 2009. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biology,* 10**,** R128.
POTAPOVA, T. A., ZHU, J. & LI, R. 2013. Aneuploidy and chromosomal instability: a vicious cycle driving cellular evolution and cancer genome chaos. *Cancer Metastasis Rev,* 32**,** 377-89.
PRIESTLEY, P., BABER, J., LOLKEMA, M., STEEGHS, N., DE BRUIJN, E., DUYVESTEYN, K., HAIDARI, S., VAN HOECK, A., ONSTENK, W., ROEPMAN, P., SHALE, C., VODA, M., BLOEMENDAL, H., TJAN-HEIJNEN, V., VAN HERPEN, C., LABOTS, M., WITTEVEEN, P., SMIT, E., SLEIJFER, S., VOEST, E. & CUPPEN, E. 2018. Pan-cancer whole genome analyses of metastatic solid tumors. *bioRxiv*.
RAVID, K., LU, J., ZIMMET, J. M. & JONES, M. R. 2002. Roads to polyploidy: the megakaryocyte example. *J Cell Physiol,* 190**,** 7-20.
REDDY, E. P., REYNOLDS, R. K., SANTOS, E. & BARBACID, M. 1982. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature,* 300**,** 149-52.
REHEN, S. K., MCCONNELL, M. J., KAUSHAL, D., KINGSBURY, M. A., YANG, A. H. & CHUN, J. 2001. Chromosomal variation in neurons of the developing and adult mammalian nervous system. *Proc Natl Acad Sci U S A,* 98**,** 13361-6.
REUBEN, A., SPENCER, C. N., PRIETO, P. A., GOPALAKRISHNAN, V., REDDY, S. M., MILLER, J. P., MAO, X., DE MACEDO, M. P., CHEN, J., SONG, X., JIANG, H., CHEN, P. L., BEIRD, H. C., GARBER, H. R., ROH, W., WANI, K., CHEN, E., HAYMAKER, C., FORGET, M. A., LITTLE, L. D., GUMBS, C., THORNTON, R. L., HUDGENS, C. W., CHEN, W. S., AUSTIN-BRENEMAN, J., SLOANE, R. S., NEZI, L.,

COGDILL, A. P., BERNATCHEZ, C., ROSZIK, J., HWU, P., WOODMAN, S. E., CHIN, L., TAWBI, H., DAVIES, M. A., GERSHENWALD, J. E., AMARIA, R. N., GLITZA, I. C., DIAB, A., PATEL, S. P., HU, J., LEE, J. E., GRIMM, E. A., TETZLAFF, M. T., LAZAR, A. J., WISTUBA, II, CLISE-DWYER, K., CARTER, B. W., ZHANG, J., FUTREAL, P. A., SHARMA, P., ALLISON, J. P., COOPER, Z. A. & WARGO, J. A. 2017. Genomic and immune heterogeneity are associated with differential responses to therapy in melanoma. *NPJ Genom Med,* 2.

RIED, T., HU, Y., DIFILIPPANTONIO, M. J., GHADIMI, B. M., GRADE, M. & CAMPS, J. 2012. The consequences of chromosomal aneuploidy on the transcriptome of cancer cells. *Biochim Biophys Acta,* 1819**,** 784-93.

RIMMER, A., PHAN, H., MATHIESON, I., IQBAL, Z., TWIGG, S. R., CONSORTIUM, W. G. S., WILKIE, A. O., MCVEAN, G. & LUNTER, G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet,* 46**,** 912-8.

ROBINSON, A. 1960. A proposed standard system of nomenclature of human mitotic chromosomes. *JAMA,* 174**,** 159-62.

ROSS, E. M., HAASE, K., VAN LOO, P. & MARKOWETZ, F. 2017. Allele-specific multi-sample copy number segmentation. *bioRxiv***,** 166017.

ROSS, E. M. & MARKOWETZ, F. 2016. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol,* 17**,** 69.

ROSS-INNES, C. S., BECQ, J., WARREN, A., CHEETHAM, R. K., NORTHEN, H., O'DONOVAN, M., MALHOTRA, S., DI PIETRO, M., IVAKHNO, S., HE, M., WEAVER, J. M. J., LYNCH, A. G., KINGSBURY, Z., ROSS, M., HUMPHRAY, S., BENTLEY, D., FITZGERALD, R. C., FOR THE OESOPHAGEAL CANCER, C., MOLECULAR STRATIFICATION STUDY, G., HAYES, S. J., ANG, Y., WELCH, I., PRESTON, S., OAKES, S., SAVE, V., SKIPWORTH, R., TUCKER, O., DAVIES, J., CRICHTON, C., SCHUSTERREITER, C., UNDERWOOD, T., NOBLE, F., STACEY, B., KELLY, J., BYRNE, J., HAYDON, A., SHARLAND, D., OWSLEY, J., BARR, H., LAGERGREN, J., GOSSAGE, J., DAVIES, A., MASON, R., CHANG, F., ZYLSTRA, J., SANDERS, G., WHEATLEY, T., BERRISFORD, R., BRACEY, T., HARDEN, C., BUNTING, D., ROQUES, T., NOBES, J., LOO, S., LEWIS, M., CHEONG, E., PRIEST, O., PARSONS, S. L., SOOMRO, I., KAYE, P., SAUNDERS, J., PANG, V., WELCH, N. T., CATTON, J. A., DUFFY, J. P., RAGUNATH, K., LOVAT, L., HAIDRY, R., MIAH, H., KERR, S., ENEH, V., BUTAWAN, R., ROQUES, T., LEWIS, M., CHEONG, E., KUMAR, B., IGALI, L., WALTON, S., DANN, A., SAFRANEK, P., HINDMARSH, A., SUDJENDRAN, V., SCOTT, M., CLUROE, A., MIREMADI, A., MAHLER-ARAUJO, B., NUTZINGER, B., PETERS, C., ABDULLAHI, Z., CRAWTE, J., MACRAE, S., NOORANI, A., ELLIOTT, R. F., BOWER, L., EDWARDS, P., TAVARE, S., ELDRIDGE, M., BORNSCHEIN, J., SECRIER, M., YANG, T.-P., et al. 2015. Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nature Genetics,* 47**,** 1038.

ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-COTE, A. & SHAH, S. P. 2014. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods,* 11**,** 396-8.

ROWLEY, J. D. 1973. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature,* 243**,** 290-3.

ROY, D. M., WALSH, L. A., DESRICHARD, A., HUSE, J. T., WU, W., GAO, J., BOSE, P., LEE, W. & CHAN, T. A. 2016. Integrated Genomics for Pinpointing Survival Loci within Arm-Level Somatic Copy Number Alterations. *Cancer Cell,* 29**,** 737-750.

ROYLANCE, R. R., ENDESFELDER, D., GORMAN, P., BURRELL, R., SANDER, J., TOMLINSON, I., HANBY, A. M., SPEIRS, V., RICHARDSON, A., BIRKBAK, N., DOWNWARD, J., KSCHISCHO, M., SZALLASI, Z. & SWANTON, C. 2011. Relationship of Extreme Chromosomal Instability with Long-term Survival in a Retrospective Analysis of Primary Breast Cancer. *Cancer Epidemiol Biomarkers Prev,* in press.

RYE, I. H., TRINH, A., SAETERSDAL, A. B., NEBDAL, D., LINGJAERDE, O. C., ALMENDRO, V., POLYAK, K., BORRESEN-DALE, A. L., HELLAND, A., MARKOWETZ, F. & RUSSNES, H. G. 2018. Intratumor heterogeneity defines treatment-resistant HER2+ breast tumors. *Mol Oncol,* 12**,** 1838-1855.

SANBORN, J. Z., CHUNG, J., PURDOM, E., WANG, N. J., KAKAVAND, H., WILMOTT, J. S., BUTLER, T., THOMPSON, J. F., MANN, G. J., HAYDU, L. E., SAW, R. P., BUSAM, K. J., LO, R. S., COLLISSON, E. A., HUR, J. S., SPELLMAN, P. T., CLEAVER, J. E., GRAY, J. W., HUH, N., MURALI, R., SCOLYER, R. A., BASTIAN, B. C. & CHO, R. J. 2015. Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proc Natl Acad Sci U S A,* 112**,** 10995-1000.

SANCHEZ-GARCIA, F., AKAVIA, U. D., MOZES, E. & PE'ER, D. 2010. JISTIC: identification of significant targets in cancer. *BMC Bioinformatics,* 11**,** 189.

SANTARIUS, T., SHIPLEY, J., BREWER, D., STRATTON, M. R. & COOPER, C. S. 2010. A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer,* 10**,** 59-64.

SATHIRAPONGSASUTI, J. F., LEE, H., HORST, B. A., BRUNNER, G., COCHRAN, A. J., BINDER, S., QUACKENBUSH, J. & NELSON, S. F. 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics,* 27**,** 2648-54.

SATO, Y., YOSHIZATO, T., SHIRAISHI, Y., MAEKAWA, S., OKUNO, Y., KAMURA, T., SHIMAMURA, T., SATO-OTSUBO, A., NAGAE, G., SUZUKI, H., NAGATA, Y., YOSHIDA, K., KON, A., SUZUKI, Y., CHIBA, K., TANAKA, H., NIIDA, A., FUJIMOTO, A., TSUNODA, T., MORIKAWA, T., MAEDA, D., KUME, H., SUGANO, S., FUKAYAMA, M., ABURATANI, H., SANADA, M., MIYANO, S., HOMMA, Y. & OGAWA, S. 2013. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet,* 45**,** 860-7.

SAVAS, P., TEO, Z. L., LEFEVRE, C., FLENSBURG, C., CARAMIA, F., ALSOP, K., MANSOUR, M., FRANCIS, P. A., THORNE, H. A., SILVA,

M. J., KANU, N., DIETZEN, M., ROWAN, A., KSCHISCHO, M., FOX, S., BOWTELL, D. D., DAWSON, S. J., SPEED, T. P., SWANTON, C. & LOI, S. 2016. The Subclonal Architecture of Metastatic Breast Cancer: Results from a Prospective Community-Based Rapid Autopsy Program "CASCADE". *PLoS Med,* 13**,** e1002204.

SCELO, G., RIAZALHOSSEINI, Y., GREGER, L., LETOURNEAU, L., GONZALEZ-PORTA, M., WOZNIAK, M. B., BOURGEY, M., HARNDEN, P., EGEVAD, L., JACKSON, S. M., KARIMZADEH, M., ARSENEAULT, M., LEPAGE, P., HOW-KIT, A., DAUNAY, A., RENAULT, V., BLANCHE, H., TUBACHER, E., SEHMOUN, J., VIKSNA, J., CELMS, E., OPMANIS, M., ZARINS, A., VASUDEV, N. S., SEYWRIGHT, M., ABEDI-ARDEKANI, B., CARREIRA, C., SELBY, P. J., CARTLEDGE, J. J., BYRNES, G., ZAVADIL, J., SU, J., HOLCATOVA, I., BRISUDA, A., ZARIDZE, D., MOUKERIA, A., FORETOVA, L., NAVRATILOVA, M., MATES, D., JINGA, V., ARTEMOV, A., NEDOLUZHKO, A., MAZUR, A., RASTORGUEV, S., BOULYGINA, E., HEATH, S., GUT, M., BIHOREAU, M. T., LECHNER, D., FOGLIO, M., GUT, I. G., SKRYABIN, K., PROKHORTCHOUK, E., CAMBON-THOMSEN, A., RUNG, J., BOURQUE, G., BRENNAN, P., TOST, J., BANKS, R. E., BRAZMA, A. & LATHROP, G. M. 2014. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat Commun,* 5**,** 5135.

SCHROCK, E., DU MANOIR, S., VELDMAN, T., SCHOELL, B., WIENBERG, J., FERGUSON-SMITH, M. A., NING, Y., LEDBETTER, D. H., BAR-AM, I., SOENKSEN, D., GARINI, Y. & RIED, T. 1996. Multicolor spectral karyotyping of human chromosomes. *Science,* 273**,** 494-7.

SCHUH, A., BECQ, J., HUMPHRAY, S., ALEXA, A., BURNS, A., CLIFFORD, R., FELLER, S. M., GROCOCK, R., HENDERSON, S., KHREBTUKOVA, I., KINGSBURY, Z., LUO, S., MCBRIDE, D., MURRAY, L., MENJU, T., TIMBS, A., ROSS, M., TAYLOR, J. & BENTLEY, D. 2012. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood,* 120**,** 4191-6.

SCHWARZ, R. F., NG, C. K., COOKE, S. L., NEWMAN, S., TEMPLE, J., PISKORZ, A. M., GALE, D., SAYAL, K., MURTAZA, M., BALDWIN, P. J., ROSENFELD, N., EARL, H. M., SALA, E., JIMENEZ-LINAN, M., PARKINSON, C. A., MARKOWETZ, F. & BRENTON, J. D. 2015. Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLoS Med,* 12**,** e1001789.

SHAH, S. P., ROTH, A., GOYA, R., OLOUMI, A., HA, G., ZHAO, Y., TURASHVILI, G., DING, J., TSE, K., HAFFARI, G., BASHASHATI, A., PRENTICE, L. M., KHATTRA, J., BURLEIGH, A., YAP, D., BERNARD, V., MCPHERSON, A., SHUMANSKY, K., CRISAN, A., GIULIANY, R., HERAVI-MOUSSAVI, A., ROSNER, J., LAI, D., BIROL, I., VARHOL, R., TAM, A., DHALLA, N., ZENG, T., MA, K., CHAN, S. K., GRIFFITH, M., MORADIAN, A., CHENG, S. W., MORIN, G. B., WATSON, P., GELMON, K., CHIA, S., CHIN, S. F., CURTIS, C., RUEDA, O. M., PHAROAH, P. D., DAMARAJU, S., MACKEY, J., HOON, K., HARKINS, T., TADIGOTLA, V., SIGAROUDINIA, M., GASCARD, P., TLSTY, T., COSTELLO, J. F., MEYER, I. M., EAVES, C. J.,

WASSERMAN, W. W., JONES, S., HUNTSMAN, D., HIRST, M., CALDAS, C., MARRA, M. A. & APARICIO, S. 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature,* 486**,** 395-9.

SHELTZER, J. M., BLANK, H. M., PFAU, S. J., TANGE, Y., GEORGE, B. M., HUMPTON, T. J., BRITO, I. L., HIRAOKA, Y., NIWA, O. & AMON, A. 2011. Aneuploidy drives genomic instability in yeast. *Science,* 333**,** 1026-30.

SHELTZER, J. M., KO, J. H., REPLOGLE, J. M., HABIBE BURGOS, N. C., CHUNG, E. S., MEEHL, C. M., SAYLES, N. M., PASSERINI, V., STORCHOVA, Z. & AMON, A. 2017. Single-chromosome Gains Commonly Function as Tumor Suppressors. *Cancer Cell,* 31**,** 240-255.

SHEN, C., BEROUKHIM, R., SCHUMACHER, S. E., ZHOU, J., CHANG, M., SIGNORETTI, S. & KAELIN, W. G., JR. 2011. Genetic and functional studies implicate HIF1alpha as a 14q kidney cancer suppressor gene. *Cancer Discov,* 1**,** 222-35.

SHEN, R. & SESHAN, V. E. 2016. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res,* 44**,** e131.

SHEW, J. Y., LIN, B. T., CHEN, P. L., TSENG, B. Y., YANG-FENG, T. L. & LEE, W. H. 1990. C-terminal truncation of the retinoblastoma gene product leads to functional inactivation. *Proc Natl Acad Sci U S A,* 87**,** 6-10.

SHI, H., HUGO, W., KONG, X., HONG, A., KOYA, R. C., MORICEAU, G., CHODON, T., GUO, R., JOHNSON, D. B., DAHLMAN, K. B., KELLEY, M. C., KEFFORD, R. F., CHMIELOWSKI, B., GLASPY, J. A., SOSMAN, J. A., VAN BAREN, N., LONG, G. V., RIBAS, A. & LO, R. S. 2014. Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. *Cancer Discov,* 4**,** 80-93.

SMITH, J. C. & SHELTZER, J. M. 2018. Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. *Elife,* 7.

SMYTH, E. C., LAGERGREN, J., FITZGERALD, R. C., LORDICK, F., SHAH, M. A., LAGERGREN, P. & CUNNINGHAM, D. 2017. Oesophageal cancer. *Nat Rev Dis Primers,* 3**,** 17048.

SNYDER, A., MAKAROV, V., MERGHOUB, T., YUAN, J., ZARETSKY, J. M., DESRICHARD, A., WALSH, L. A., POSTOW, M. A., WONG, P., HO, T. S., HOLLMANN, T. J., BRUGGEMAN, C., KANNAN, K., LI, Y., ELIPENAHLI, C., LIU, C., HARBISON, C. T., WANG, L., RIBAS, A., WOLCHOK, J. D. & CHAN, T. A. 2014. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med,* 371**,** 2189-99.

SOLIMINI, N. L., XU, Q., MERMEL, C. H., LIANG, A. C., SCHLABACH, M. R., LUO, J., BURROWS, A. E., ANSELMO, A. N., BREDEMEYER, A. L., LI, M. Z., BEROUKHIM, R., MEYERSON, M. & ELLEDGE, S. J. 2012. Recurrent hemizygous deletions in cancers may optimize proliferative potential. *Science,* 337**,** 104-9.

SOLOMON, D. A., KIM, T., DIAZ-MARTINEZ, L. A., FAIR, J., ELKAHLOUN, A. G., HARRIS, B. T., TORETSKY, J. A., ROSENBERG, S. A., SHUKLA, N., LADANYI, M., SAMUELS, Y., JAMES, C. D., YU, H., KIM,

J. S. & WALDMAN, T. 2011. Mutational inactivation of STAG2 causes aneuploidy in human cancer. *Science,* 333**,** 1039-43.
SOTILLO, R., SCHVARTZMAN, J. M., SOCCI, N. D. & BENEZRA, R. 2010. Mad2-induced chromosome instability leads to lung tumour relapse after oncogene withdrawal. *Nature,* 464**,** 436-40.
SOTO, M., RAAIJMAKERS, J. A., BAKKER, B., SPIERINGS, D. C. J., LANSDORP, P. M., FOIJER, F. & MEDEMA, R. H. 2017. p53 Prohibits Propagation of Chromosome Segregation Errors that Produce Structural Aneuploidies. *Cell Rep,* 19**,** 2423-2431.
SOTTORIVA, A., KANG, H., MA, Z., GRAHAM, T. A., SALOMON, M. P., ZHAO, J., MARJORAM, P., SIEGMUND, K., PRESS, M. F., SHIBATA, D. & CURTIS, C. 2015. A Big Bang model of human colorectal tumor growth. *Nat Genet,* 47**,** 209-16.
SOTTORIVA, A., SPITERI, I., PICCIRILLO, S. G., TOULOUMIS, A., COLLINS, V. P., MARIONI, J. C., CURTIS, C., WATTS, C. & TAVARE, S. 2013. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A,* 110**,** 4009-14.
SPARKES, R. S., MURPHREE, A. L., LINGUA, R. W., SPARKES, M. C., FIELD, L. L., FUNDERBURK, S. J. & BENEDICT, W. F. 1983. Gene for hereditary retinoblastoma assigned to human chromosome 13 by linkage to esterase D. *Science,* 219**,** 971-3.
SPEICHER, M. R., GWYN BALLARD, S. & WARD, D. C. 1996. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nat Genet,* 12**,** 368-75.
SPOERKE, J. M., GENDREAU, S., WALTER, K., QIU, J., WILSON, T. R., SAVAGE, H., AIMI, J., DERYNCK, M. K., CHEN, M., CHAN, I. T., AMLER, L. C., HAMPTON, G. M., JOHNSTON, S., KROP, I., SCHMID, P. & LACKNER, M. R. 2016. Heterogeneity and clinical significance of ESR1 mutations in ER-positive metastatic breast cancer patients receiving fulvestrant. *Nat Commun,* 7**,** 11579.
STEPHENS, P. J., GREENMAN, C. D., FU, B., YANG, F., BIGNELL, G. R., MUDIE, L. J., PLEASANCE, E. D., LAU, K. W., BEARE, D., STEBBINGS, L. A., MCLAREN, S., LIN, M. L., MCBRIDE, D. J., VARELA, I., NIK-ZAINAL, S., LEROY, C., JIA, M., MENZIES, A., BUTLER, A. P., TEAGUE, J. W., QUAIL, M. A., BURTON, J., SWERDLOW, H., CARTER, N. P., MORSBERGER, L. A., IACOBUZIO-DONAHUE, C., FOLLOWS, G. A., GREEN, A. R., FLANAGAN, A. M., STRATTON, M. R., FUTREAL, P. A. & CAMPBELL, P. J. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell,* 144**,** 27-40.
STINGELE, S., STOEHR, G., PEPLOWSKA, K., COX, J., MANN, M. & STORCHOVA, Z. 2012. Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol Syst Biol,* 8**,** 608.
STORCHOVA, Z. & KUFFER, C. 2008. The consequences of tetraploidy and aneuploidy. *J Cell Sci,* 121**,** 3859-66.

STRINO, F., PARISI, F., MICSINAI, M. & KLUGER, Y. 2013. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res,* 41**,** e165.

SUN, R., HU, Z., SOTTORIVA, A., GRAHAM, T. A., HARPAK, A., MA, Z., FISCHER, J. M., SHIBATA, D. & CURTIS, C. 2017. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat Genet,* 49**,** 1015-1024.

SUNDARESAN, V., GANLY, P., HASLETON, P., RUDD, R., SINHA, G., BLEEHEN, N. M. & RABBITTS, P. 1992. p53 and chromosome 3 abnormalities, characteristic of malignant lung tumours, are detectable in preinvasive lesions of the bronchus. *Oncogene,* 7**,** 1989-97.

SUZUKI, H., AOKI, K., CHIBA, K., SATO, Y., SHIOZAWA, Y., SHIRAISHI, Y., SHIMAMURA, T., NIIDA, A., MOTOMURA, K., OHKA, F., YAMAMOTO, T., TANAHASHI, K., RANJIT, M., WAKABAYASHI, T., YOSHIZATO, T., KATAOKA, K., YOSHIDA, K., NAGATA, Y., SATO-OTSUBO, A., TANAKA, H., SANADA, M., KONDO, Y., NAKAMURA, H., MIZOGUCHI, M., ABE, T., MURAGAKI, Y., WATANABE, R., ITO, I., MIYANO, S., NATSUME, A. & OGAWA, S. 2015. Mutational landscape and clonal architecture in grade II and III gliomas. *Nat Genet,* 47**,** 458-68.

SVARTMAN, M., STONE, G. & STANYON, R. 2005. Molecular cytogenetics discards polyploidy in mammals. *Genomics,* 85**,** 425-30.

SWANTON, C., NICKE, B., SCHUETT, M., EKLUND, A. C., NG, C., LI, Q., HARDCASTLE, T., LEE, A., ROY, R., EAST, P., KSCHISCHO, M., ENDESFELDER, D., WYLIE, P., KIM, S. N., CHEN, J. G., HOWELL, M., RIED, T., HABERMANN, J. K., AUER, G., BRENTON, J. D., SZALLASI, Z. & DOWNWARD, J. 2009. Chromosomal instability determines taxane response. *Proc Natl Acad Sci U S A,* 106**,** 8671-6.

SZOLLOSI, J., BALAZS, M., FEUERSTEIN, B. G., BENZ, C. C. & WALDMAN, F. M. 1995. ERBB-2 (HER2/neu) gene copy number, p185HER-2 overexpression, and intratumor heterogeneity in human breast cancer. *Cancer Res,* 55**,** 5400-7.

TALEVICH, E., SHAIN, A. H., BOTTON, T. & BASTIAN, B. C. 2016. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol,* 12**,** e1004873.

TARABICHI, M., MARTINCORENA, I., GERSTUNG, M., LEROI, A. M., MARKOWETZ, F., EVOLUTION, P., HETEROGENEITY WORKING, G., SPELLMAN, P. T., MORRIS, Q. D., LINGJAERDE, O. C., WEDGE, D. C. & VAN LOO, P. 2018. Neutral tumor evolution? *Nat Genet,* 50**,** 1630-1633.

TAYLOR, A. M., SHIH, J., HA, G., GAO, G. F., ZHANG, X., BERGER, A. C., SCHUMACHER, S. E., WANG, C., HU, H., LIU, J., LAZAR, A. J., CANCER GENOME ATLAS RESEARCH, N., CHERNIACK, A. D., BEROUKHIM, R. & MEYERSON, M. 2018. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell,* 33**,** 676-689 e3.

TCGA, B. I. T. G. D. A. C. 2015. SNP6 Copy number analysis (GISTIC2) (2014) [Internet]. *Broad Institute of MIT and Harvard*.

TCGA, B. I. T. G. D. A. C. 2017. SNP6 Copy number analysis (2016) (GISTIC2). *Broad Institute of MIT and Harvard*.

TCGA_CONSORTIUM, C. G. A. R. N. 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature,* 499**,** 43-9.

TELLI, M. L., TIMMS, K. M., REID, J., HENNESSY, B., MILLS, G. B., JENSEN, K. C., SZALLASI, Z., BARRY, W. T., WINER, E. P., TUNG, N. M., ISAKOFF, S. J., RYAN, P. D., GREENE-COLOZZI, A., GUTIN, A., SANGALE, Z., ILIEV, D., NEFF, C., ABKEVICH, V., JONES, J. T., LANCHBURY, J. S., HARTMAN, A. R., GARBER, J. E., FORD, J. M., SILVER, D. P. & RICHARDSON, A. L. 2016. Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer. *Clin Cancer Res,* 22**,** 3764-73.

THOMPSON, S. L., BAKHOUM, S. F. & COMPTON, D. A. 2010. Mechanisms of chromosomal instability. *Curr Biol,* 20**,** R285-95.

THOMPSON, S. L. & COMPTON, D. A. 2010. Proliferation of aneuploid human cells is limited by a p53-dependent mechanism. *J Cell Biol,* 188**,** 369-81.

THOMSEN, M. B., NORDENTOFT, I., LAMY, P., HOYER, S., VANG, S., HEDEGAARD, J., BORRE, M., JENSEN, J. B., ORNTOFT, T. F. & DYRSKJOT, L. 2016. Spatial and temporal clonal evolution during development of metastatic urothelial carcinoma. *Mol Oncol,* 10**,** 1450-1460.

TJIO, J. H. & LEVAN, A. 1956. THE CHROMOSOME NUMBER OF MAN. *Hereditas,* 42**,** 1-6.

TORRES, E. M., SOKOLSKY, T., TUCKER, C. M., CHAN, L. Y., BOSELLI, M., DUNHAM, M. J. & AMON, A. 2007. Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science,* 317**,** 916-24.

TORY, K., BRAUCH, H., LINEHAN, M., BARBA, D., OLDFIELD, E., FILLING-KATZ, M., SEIZINGER, B., NAKAMURA, Y., WHITE, R., MARSHALL, F. F. & ET AL. 1989. Specific genetic change in tumors associated with von Hippel-Lindau disease. *J Natl Cancer Inst,* 81**,** 1097-101.

TOYODA, H., BREGERIE, O., VALLET, A., NALPAS, B., PIVERT, G., BRECHOT, C. & DESDOUETS, C. 2005. Changes to hepatocyte ploidy and binuclearity profiles during human chronic viral hepatitis. *Gut,* 54**,** 297-302.

TURAJLIC, S., MCGRANAHAN, N. & SWANTON, C. 2015. Inferring mutational timing and reconstructing tumour evolutionary histories. *Biochim Biophys Acta,* 1855**,** 264-75.

TURAJLIC, S., XU, H., LITCHFIELD, K., ROWAN, A., CHAMBERS, T., LOPEZ, J. I., NICOL, D., O'BRIEN, T., LARKIN, J., HORSWELL, S., STARES, M., AU, L., JAMAL-HANJANI, M., CHALLACOMBE, B., CHANDRA, A., HAZELL, S., EICHLER-JONSSON, C., SOULTATI, A., CHOWDHURY, S., RUDMAN, S., LYNCH, J., FERNANDO, A., STAMP, G., NYE, E., JABBAR, F., SPAIN, L., LALL, S., GUARCH, R., FALZON, M., PROCTOR, I., PICKERING, L., GORE, M., WATKINS, T. B. K., WARD, S., STEWART, A., DINATALE, R., BECERRA, M. F., REZNIK, E., HSIEH, J. J., RICHMOND, T. A., MAYHEW, G. F., HILL, S. M., MCNALLY, C. D., JONES, C., ROSENBAUM, H., STANISLAW, S., BURGESS, D. L., ALEXANDER, N. R., SWANTON, C., PEACE & CONSORTIUM, T. R. R. 2018a. Tracking Cancer Evolution Reveals

Constrained Routes to Metastases: TRACERx Renal. *Cell,* 173**,** 581-594 e12.

TURAJLIC, S., XU, H., LITCHFIELD, K., ROWAN, A., HORSWELL, S., CHAMBERS, T., O'BRIEN, T., LOPEZ, J. I., WATKINS, T. B. K., NICOL, D., STARES, M., CHALLACOMBE, B., HAZELL, S., CHANDRA, A., MITCHELL, T. J., AU, L., EICHLER-JONSSON, C., JABBAR, F., SOULTATI, A., CHOWDHURY, S., RUDMAN, S., LYNCH, J., FERNANDO, A., STAMP, G., NYE, E., STEWART, A., XING, W., SMITH, J. C., ESCUDERO, M., HUFFMAN, A., MATTHEWS, N., ELGAR, G., PHILLIMORE, B., COSTA, M., BEGUM, S., WARD, S., SALM, M., BOEING, S., FISHER, R., SPAIN, L., NAVAS, C., GRONROOS, E., HOBOR, S., SHARMA, S., AURANGZEB, I., LALL, S., POLSON, A., VARIA, M., HORSFIELD, C., FOTIADIS, N., PICKERING, L., SCHWARZ, R. F., SILVA, B., HERRERO, J., LUSCOMBE, N. M., JAMAL-HANJANI, M., ROSENTHAL, R., BIRKBAK, N. J., WILSON, G. A., PIPEK, O., RIBLI, D., KRZYSTANEK, M., CSABAI, I., SZALLASI, Z., GORE, M., MCGRANAHAN, N., VAN LOO, P., CAMPBELL, P., LARKIN, J., SWANTON, C. & CONSORTIUM, T. R. R. 2018b. Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell,* 173**,** 595-610 e11.

TURNER, K. M., DESHPANDE, V., BEYTER, D., KOGA, T., RUSERT, J., LEE, C., LI, B., ARDEN, K., REN, B., NATHANSON, D. A., KORNBLUM, H. I., TAYLOR, M. D., KAUSHAL, S., CAVENEE, W. K., WECHSLER-REYA, R., FURNARI, F. B., VANDENBERG, S. R., RAO, P. N., WAHL, G. M., BAFNA, V. & MISCHEL, P. S. 2017. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature,* 543**,** 122-125.

UCHI, R., TAKAHASHI, Y., NIIDA, A., SHIMAMURA, T., HIRATA, H., SUGIMACHI, K., SAWADA, G., IWAYA, T., KURASHIGE, J., SHINDEN, Y., IGUCHI, T., EGUCHI, H., CHIBA, K., SHIRAISHI, Y., NAGAE, G., YOSHIDA, K., NAGATA, Y., HAENO, H., YAMAMOTO, H., ISHII, H., DOKI, Y., IINUMA, H., SASAKI, S., NAGAYAMA, S., YAMADA, K., YACHIDA, S., KATO, M., SHIBATA, T., OKI, E., SAEKI, H., SHIRABE, K., ODA, Y., MAEHARA, Y., KOMUNE, S., MORI, M., SUZUKI, Y., YAMAMOTO, K., ABURATANI, H., OGAWA, S., MIYANO, S. & MIMORI, K. 2016. Integrated Multiregional Analysis Proposing a New Model of Colorectal Cancer Evolution. *PLoS Genet,* 12**,** e1005778.

VAN DEN BOS, H., BAKKER, B., SPIERINGS, D. C. J., LANSDORP, P. M. & FOIJER, F. 2018. Single-cell sequencing to quantify genomic integrity in cancer. *Int J Biochem Cell Biol,* 94**,** 146-150.

VAN DEN BOS, H., SPIERINGS, D. C., TAUDT, A. S., BAKKER, B., PORUBSKY, D., FALCONER, E., NOVOA, C., HALSEMA, N., KAZEMIER, H. G., HOEKSTRA-WAKKER, K., GURYEV, V., DEN DUNNEN, W. F., FOIJER, F., TATCHE, M. C., BODDEKE, H. W. & LANSDORP, P. M. 2016. Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and Alzheimer's disease neurons. *Genome Biol,* 17**,** 116.

VAN DYK, E., REINDERS, M. J. & WESSELS, L. F. 2013. A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control. *Nucleic Acids Res,* 41**,** e100.

VAN LOO, P., NORDGARD, S. H., LINGJAERDE, O. C., RUSSNES, H. G., RYE, I. H., SUN, W., WEIGMAN, V. J., MARYNEN, P., ZETTERBERG, A., NAUME, B., PEROU, C. M., BORRESEN-DALE, A. L. & KRISTENSEN, V. N. 2010. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A,* 107**,** 16910-5.

VARELA, I., TARPEY, P., RAINE, K., HUANG, D., ONG, C. K., STEPHENS, P., DAVIES, H., JONES, D., LIN, M. L., TEAGUE, J., BIGNELL, G., BUTLER, A., CHO, J., DALGLIESH, G. L., GALAPPATHTHIGE, D., GREENMAN, C., HARDY, C., JIA, M., LATIMER, C., LAU, K. W., MARSHALL, J., MCLAREN, S., MENZIES, A., MUDIE, L., STEBBINGS, L., LARGAESPADA, D. A., WESSELS, L. F., RICHARD, S., KAHNOSKI, R. J., ANEMA, J., TUVESON, D. A., PEREZ-MANCERA, P. A., MUSTONEN, V., FISCHER, A., ADAMS, D. J., RUST, A., CHAN-ON, W., SUBIMERB, C., DYKEMA, K., FURGE, K., CAMPBELL, P. J., TEH, B. T., STRATTON, M. R. & FUTREAL, P. A. 2011. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature,* 469**,** 539-42.

VENKATESAN, S., BIRKBAK, N. J. & SWANTON, C. 2017. Constraints in cancer evolution. *Biochem Soc Trans,* 45**,** 1-13.

VOGELSTEIN, B., PAPADOPOULOS, N., VELCULESCU, V. E., ZHOU, S., DIAZ, L. A., JR. & KINZLER, K. W. 2013. Cancer genome landscapes. *Science,* 339**,** 1546-58.

VON HANSEMANN, D. 1890. Ueber asymmetrische Zelltheilung in epithel Krebsen und deren

biologische Bedeutung. *Virchow's Archiv für pathologische Anatomie und Physiologie und für klinische Medicin***,** 299-326.

WANG, C., EVANS, J. M., BHAGWATE, A. V., PRODDUTURI, N., SARANGI, V., MIDDHA, M., SICOTTE, H., VEDELL, P. T., HART, S. N., OLIVER, G. R., KOCHER, J. P., MAURER, M. J., NOVAK, A. J., SLAGER, S. L., CERHAN, J. R. & ASMANN, Y. W. 2014a. PatternCNV: a versatile tool for detecting copy number changes from exome sequencing data. *Bioinformatics,* 30**,** 2678-80.

WANG, K., LI, M. & HAKONARSON, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res,* 38**,** e164.

WANG, Y., WATERS, J., LEUNG, M. L., UNRUH, A., ROH, W., SHI, X., CHEN, K., SCHEET, P., VATTATHIL, S., LIANG, H., MULTANI, A., ZHANG, H., ZHAO, R., MICHOR, F., MERIC-BERNSTAM, F. & NAVIN, N. E. 2014b. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature,* 512**,** 155-60.

WEAVER, B. A. & CLEVELAND, D. W. 2006. Does aneuploidy cause cancer? *Curr Opin Cell Biol,* 18**,** 658-67.

WEAVER, J. M., ROSS-INNES, C. S., SHANNON, N., LYNCH, A. G., FORSHEW, T., BARBERA, M., MURTAZA, M., ONG, C. A., LAO-SIRIEIX, P., DUNNING, M. J., SMITH, L., SMITH, M. L., ANDERSON, C. L., CARVALHO, B., O'DONOVAN, M., UNDERWOOD, T. J., MAY,

A. P., GREHAN, N., HARDWICK, R., DAVIES, J., OLOUMI, A., APARICIO, S., CALDAS, C., ELDRIDGE, M. D., EDWARDS, P. A., ROSENFELD, N., TAVARE, S., FITZGERALD, R. C. & CONSORTIUM, O. 2014. Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat Genet,* 46**,** 837-43.

WEINBERG, R. A. 1985. The action of oncogenes in the cytoplasm and nucleus. *Science,* 230**,** 770-6.

WEISS, M. C. & GREEN, H. 1967. Human-mouse hybrid cell lines containing partial complements of human chromosomes and functioning human genes. *Proc Natl Acad Sci U S A,* 58**,** 1104-11.

WILLIAMS, B. R., PRABHU, V. R., HUNTER, K. E., GLAZIER, C. M., WHITTAKER, C. A., HOUSMAN, D. E. & AMON, A. 2008. Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. *Science,* 322**,** 703-9.

WILLIAMS, M. J., WERNER, B., BARNES, C. P., GRAHAM, T. A. & SOTTORIVA, A. 2016. Identification of neutral tumor evolution across cancer types. *Nat Genet,* 48**,** 238-244.

WILLIAMS, M. J., WERNER, B., HEIDE, T., BARNES, C. P., GRAHAM, T. A. & SOTTORIVA, A. 2018. Reply to 'Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data'. *Nat Genet,* 50**,** 1628-1630.

WINGE, O. 1927. Zytologische Untersuchungen ueber die Natur maligner Tumoren.1. "Crown gall" der Zuckerrilbe. *Ztschr. f. Zellforsch. und mikr. Anat.***,** 397-423.

WINGE, O. 1930. Zytologische Untersuchungen uber die Natur maligner tumoren. II. Tecrkarzinome bei Mausen. Z. Zellforsch. *Z. Zellforsch. Mikrosk. Anat.,***,** 683–735.

WISTUBA, II, BEHRENS, C., VIRMANI, A. K., MELE, G., MILCHGRUB, S., GIRARD, L., FONDON, J. W., 3RD, GARNER, H. R., MCKAY, B., LATIF, F., LERMAN, M. I., LAM, S., GAZDAR, A. F. & MINNA, J. D. 2000. High resolution chromosome 3p allelotyping of human lung cancer and preneoplastic/preinvasive bronchial epithelium reveals multiple, discontinuous sites of 3p allele loss and three regions of frequent breakpoints. *Cancer Res,* 60**,** 1949-60.

WOOD, L. D., PARSONS, D. W., JONES, S., LIN, J., SJOBLOM, T., LEARY, R. J., SHEN, D., BOCA, S. M., BARBER, T., PTAK, J., SILLIMAN, N., SZABO, S., DEZSO, Z., USTYANKSKY, V., NIKOLSKAYA, T., NIKOLSKY, Y., KARCHIN, R., WILSON, P. A., KAMINKER, J. S., ZHANG, Z., CROSHAW, R., WILLIS, J., DAWSON, D., SHIPITSIN, M., WILLSON, J. K., SUKUMAR, S., POLYAK, K., PARK, B. H., PETHIYAGODA, C. L., PANT, P. V., BALLINGER, D. G., SPARKS, A. B., HARTIGAN, J., SMITH, D. R., SUH, E., PAPADOPOULOS, N., BUCKHAULTS, P., MARKOWITZ, S. D., PARMIGIANI, G., KINZLER, K. W., VELCULESCU, V. E. & VOGELSTEIN, B. 2007. The genomic landscapes of human breast and colorectal cancers. *Science,* 318**,** 1108-13.

WORRALL, J. T., TAMURA, N., MAZZAGATTI, A., SHAIKH, N., VAN LINGEN, T., BAKKER, B., SPIERINGS, D. C. J., VLADIMIROU, E., FOIJER, F. & MCCLELLAND, S. E. 2018. Non-random Mis-segregation of Human Chromosomes. *Cell Rep,* 23**,** 3366-3380.

XUE, W., KITZING, T., ROESSLER, S., ZUBER, J., KRASNITZ, A., SCHULTZ, N., REVILL, K., WEISSMUELLER, S., RAPPAPORT, A. R., SIMON, J., ZHANG, J., LUO, W., HICKS, J., ZENDER, L., WANG, X. W., POWERS, S., WIGLER, M. & LOWE, S. W. 2012. A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. *Proc Natl Acad Sci U S A,* 109**,** 8212-7.

YATES, L. R., GERSTUNG, M., KNAPPSKOG, S., DESMEDT, C., GUNDEM, G., VAN LOO, P., AAS, T., ALEXANDROV, L. B., LARSIMONT, D., DAVIES, H., LI, Y., JU, Y. S., RAMAKRISHNA, M., HAUGLAND, H. K., LILLENG, P. K., NIK-ZAINAL, S., MCLAREN, S., BUTLER, A., MARTIN, S., GLODZIK, D., MENZIES, A., RAINE, K., HINTON, J., JONES, D., MUDIE, L. J., JIANG, B., VINCENT, D., GREENE-COLOZZI, A., ADNET, P. Y., FATIMA, A., MAETENS, M., IGNATIADIS, M., STRATTON, M. R., SOTIRIOU, C., RICHARDSON, A. L., LONNING, P. E., WEDGE, D. C. & CAMPBELL, P. J. 2015. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med,* 21**,** 751-9.

YATES, L. R., KNAPPSKOG, S., WEDGE, D., FARMERY, J. H. R., GONZALEZ, S., MARTINCORENA, I., ALEXANDROV, L. B., VAN LOO, P., HAUGLAND, H. K., LILLENG, P. K., GUNDEM, G., GERSTUNG, M., PAPPAEMMANUIL, E., GAZINSKA, P., BHOSLE, S. G., JONES, D., RAINE, K., MUDIE, L., LATIMER, C., SAWYER, E., DESMEDT, C., SOTIRIOU, C., STRATTON, M. R., SIEUWERTS, A. M., LYNCH, A. G., MARTENS, J. W., RICHARDSON, A. L., TUTT, A., LONNING, P. E. & CAMPBELL, P. J. 2017. Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell,* 32**,** 169-184 e7.

YU, C., YU, J., YAO, X., WU, W. K., LU, Y., TANG, S., LI, X., BAO, L., LI, X., HOU, Y., WU, R., JIAN, M., CHEN, R., ZHANG, F., XU, L., FAN, F., HE, J., LIANG, Q., WANG, H., HU, X., HE, M., ZHANG, X., ZHENG, H., LI, Q., WU, H., CHEN, Y., YANG, X., ZHU, S., XU, X., YANG, H., WANG, J., ZHANG, X., SUNG, J. J., LI, Y. & WANG, J. 2014. Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. *Cell Res,* 24**,** 701-12.

YUAN, K., SAKOPARNIG, T., MARKOWETZ, F. & BEERENWINKEL, N. 2015. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol,* 16**,** 36.

YUNIS, J. J. & RAMSAY, N. 1978. Retinoblastoma and subband deletion of chromosome 13. *Am J Dis Child,* 132**,** 161-3.

ZACK, T. I., SCHUMACHER, S. E., CARTER, S. L., CHERNIACK, A. D., SAKSENA, G., TABAK, B., LAWRENCE, M. S., ZHANG, C. Z., WALA, J., MERMEL, C. H., SOUGNEZ, C., GABRIEL, S. B., HERNANDEZ, B., SHEN, H., LAIRD, P. W., GETZ, G., MEYERSON, M. & BEROUKHIM, R. 2013. Pan-cancer patterns of somatic copy number alteration. *Nat Genet,* 45**,** 1134-1140.

ZAHN, H., STEIF, A., LAKS, E., EIREW, P., VANINSBERGHE, M., SHAH, S. P., APARICIO, S. & HANSEN, C. L. 2017. Scalable whole-genome single-cell library preparation without preamplification. *Nat Methods,* 14**,** 167-173.

ZAKI, B. I., SURIAWINATA, A. A., EASTMAN, A. R., GARNER, K. M. & BAKHOUM, S. F. 2014. Chromosomal instability portends superior

response of rectal adenocarcinoma to chemoradiation therapy. *Cancer,* 120**,** 1733-42.

ZARE, F., DOW, M., MONTELEONE, N., HOSNY, A. & NABAVI, S. 2017. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics,* 18**,** 286.

ZENKLUSEN, J. C., BIECHE, I., LIDEREAU, R. & CONTI, C. J. 1994. (C-A)n microsatellite repeat D7S522 is the most commonly deleted region in human primary breast cancer. *Proceedings of the National Academy of Sciences,* 91**,** 12155-12158.

ZHANG, C. Z., SPEKTOR, A., CORNILS, H., FRANCIS, J. M., JACKSON, E. K., LIU, S., MEYERSON, M. & PELLMAN, D. 2015. Chromothripsis from DNA damage in micronuclei. *Nature,* 522**,** 179-84.

ZHANG, J., FUJIMOTO, J., ZHANG, J., WEDGE, D. C., SONG, X., ZHANG, J., SETH, S., CHOW, C. W., CAO, Y., GUMBS, C., GOLD, K. A., KALHOR, N., LITTLE, L., MAHADESHWAR, H., MORAN, C., PROTOPOPOV, A., SUN, H., TANG, J., WU, X., YE, Y., WILLIAM, W. N., LEE, J. J., HEYMACH, J. V., HONG, W. K., SWISHER, S., WISTUBA, II & FUTREAL, P. A. 2014. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science,* 346**,** 256-9.

ZHENG, S., CHERNIACK, A. D., DEWAL, N., MOFFITT, R. A., DANILOVA, L., MURRAY, B. A., LERARIO, A. M., ELSE, T., KNIJNENBURG, T. A., CIRIELLO, G., KIM, S., ASSIE, G., MOROZOVA, O., AKBANI, R., SHIH, J., HOADLEY, K. A., CHOUEIRI, T. K., WALDMANN, J., METE, O., ROBERTSON, A. G., WU, H. T., RAPHAEL, B. J., SHAO, L., MEYERSON, M., DEMEURE, M. J., BEUSCHLEIN, F., GILL, A. J., SIDHU, S. B., ALMEIDA, M. Q., FRAGOSO, M., COPE, L. M., KEBEBEW, E., HABRA, M. A., WHITSETT, T. G., BUSSEY, K. J., RAINEY, W. E., ASA, S. L., BERTHERAT, J., FASSNACHT, M., WHEELER, D. A., CANCER GENOME ATLAS RESEARCH, N., HAMMER, G. D., GIORDANO, T. J. & VERHAAK, R. G. W. 2016. Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. *Cancer Cell,* 30**,** 363.