Mechanisms of cancer evolution and drivers of tumour heterogeneity

Nicholas Louis McGranahan

University College London

and

Cancer Research UK London Research Institute

PhD Supervisor: Charles Swanton

A thesis submitted for the degree of

Doctor of Philosophy

University College London

September 2015

Declaration

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

Abstract

Cancer drug resistance is almost inevitable in the majority of patients with advanced metastatic tumours. Intra-tumour heterogeneity, facilitating rapid tumour evolution, is a main cause of resistance to cancer therapies.

In this thesis, I explore how cancer genome sequencing data can shed light intra-tumour heterogeneity and the processes shaping cancer genome evolution over space and time. Multi-region and single-sample sequencing data was harnessed to temporally and clonally dissect mutations across 10 major cancer-types. The existence of branched tumour evolution and widespread heterogeneity was demonstrated. Although mutations in known cancer genes typically occurred early in cancer evolution, subclonal 'actionable' mutations, including *BRAF* (V600E), *IDH1* (R132H), *PIK3CA* (E545K), and *EGFR* (L858R), were also identified.

Temporal dissection of mutational signatures revealed that APOBEC-mediated-mutagenesis is frequently a late event in cancer evolution and plays a key role in the acquisition of subclonal driver mutations.

Copy number analysis suggested that genome doubling is prevalent across tumour types and that it frequently occurs early in tumour evolution in colorectal cancer. A cancer cell-line system was used to demonstrate that rare cells that survive genome-doubling display increased tolerance to chromosome aberrations and a genome-doubling event was found to be independently predictive of reduced relapse-free survival in two independent cohorts.

Finally, the clinical impact of intra-tumour heterogeneity was explored in the context of cancer neo-antigens and immune-modulation. The number of clonal neo-antigens was associated with survival outcome in lung adenocarcinoma patients and T cells reactive to clonal neo-antigens were identified. Sensitivity to anti-PD-1 therapy was dependent on neo-antigen clonal burden and intra-tumour heterogeneity. Thus, immunotherapeutic strategies targeting clonal neo-antigens in combination with checkpoint-blockade may provide a tractable approach to tackling lung adenocarcinomas.

This thesis demonstrates how analyses of genomic data can shed light on the biology and clinical relevance of cancer evolution and intra-tumour heterogeneity.

Acknowledgements

I would like to thank Dr Charlie Swanton for welcoming me into his lab, first to start a Masters project, and then to begin a PhD. This thesis would not have been possible without his endless supply of encouragement, ideas and, above all, infectious enthusiasm.

I would also like to thank the entire Translational Cancer Therapeutics Laboratory, both past and present, for their help, collaboration and friendship. Special thanks must go to Gareth Wilson for his assistance and advice during proofreading this thesis.

I am also indebted to those whom I have been fortunate enough to be able to collaborate with. In particular, I would like to thank Sally Dewhurst, Elza de Bruin, Nirupa Murugaesu, Rachel Rosenthal and Sergio Quezada.

Within the LRI and UCL, I would like to acknowledge my thesis committee, Professors Nicholas Luscombe, Paul Bates and Karen Page, the bioinformatics service lab, and my fellow students.

Last, but not least, I would like to thank my friends and family for all their patience and encouragement. In particular, I would like to thank Josephine for her support, counselling and perspective throughout the journey.

Table of Contents

Table of figures

List of tables

Abbreviations

ALL – acute lymphoblastic leukaemia

APOBEC – apoliprotein B mRNA editing enzyme catalytic polypeptide-like

BLCA – bladder adenocarcinoma

BRCA – breast carcinoma

CCF – cancer cell fraction

CGH – comparative genome hybridisation

CIN – chromosomal instability

COAD – colon adenocarcinoma

DrGAP – driver genes and pathway

ER-positive – oestrogen receptor positive

ER-negative – oestrogen receptor negative

ESCA – oesophageal cancer

FFPE – formalin fixed paraffin embedded

FM-bias – functional impact bias

FISH – fluorescence *in situ* hybridisation

GBM – glioblastoma mulitforme

GD – genome doubled

nGD – non genome doubled

GII – genome instability index

HNSC – head and neck squamous cell carcinoma

HR – hazard ratio

INDEL – small insertion and/or deletion

KIRC – clear cell renal cell carcinoma

LOH – loss of heterozygosity

LUAD – lung adenocarcinoma

LUSC – lung squamous cell carcinoma

MATH – mutant allele tumour heterogeneity

MDS – myelodysplastic syndrome

Mb – mega base

MSI – microsatellite instability

MuSiC – mutational significance in cancer

NMF – non-negative matrix factorization

NSCLC – non small-cell lung cancer

OS – overall survival

RFS – relapse free survival

SKCM – skin cutaneous melanoma

SNP – single-nucleotide polymorphism

SNV – single nucleotide variant

TILs – tumour infiltrating lymphocytes

TKI – tyrosine kinase therapy

UV – ultraviolet light

wGII – weighted genome instability index

WES – whole exome sequencing

WGS – whole genome sequencing

Chapter 1. Overview of the literature

1.1 Introduction

Cancer represents a disease of the genome and epigenome, and tumours have been found to harbour widely varying numbers and types of somatic aberrations. While the vast majority of these mutations are likely to be passenger events that do not provide any selective benefit to the cancer cell, a subset will represent cancer driver events, conferring a selective advantage to the tumour (Stratton et al., 2009)

Emerging evidence – much of which has accumulated since the start of this PhD - suggests that not every somatic mutation, whether a driver or a passenger, will necessarily be found in every cancer cell within the tumour population. While the types and distribution of somatic mutations across the tumour genome can be used to decipher the mutational processes that have been active during its evolutionary history, the degree of heterogeneity in space and over time can reveal a tumour's life history.

In this chapter, I explore our current understanding of the processes shaping the cancer genome and place these in the context of intra-tumour heterogeneity and cancer genome evolution. In addition, I outline some of the clinical implications of intra-tumour heterogeneity.

The work presented in this chapter builds upon reviews I have published (as first author, or joint first author) concerning heterogeneity and tumour evolution (see Appendix 1, and (McGranahan and Swanton, 2015, McGranahan et al., 2012, Burrell et al., 2013b, Gerlinger et al., 2014b)).

1.2 Cancer as an evolutionary system

Darwinian evolution by natural selection can be summarised by the requirement of three key tenets:

1. Phenotypic variation – individuals within a population exhibit different phenotypes, e.g. different physiologies, behaviours and morphologies

2. Differential fitness – phenotypes are associated with different fitness as measured by rates of survival and reproduction

3. Heredity – fitness is heritable and can be passed from parents to offspring

Although originally framed in relation to the evolution of individual organisms within a population, these principles of evolution can be viewed as a powerful predictive system for changes at multiple levels of biological organisation (Lewontin, 1970).

An evolutionary synthesis of cancer progression, whereby the cancer cell is the unit of selection, was first outlined in 1976 by Peter Nowell (1976). Nowell posited that cancers arise from a single founder cell and tumour progression occurs through a process of clonal evolution, involving genetic instability and the sequential acquisition and selection of more aggressive subclones (Nowell, 1976).

While our models of tumour evolution have become more nuanced since the original postulations of Nowell (1976), the notion of cancer as an evolutionary disease involving clonal selection is now supported by an overwhelming body of evidence (Greaves, 2015, Greaves and Maley, 2012, Yates and Campbell, 2012). This evidence is underpinned by a number of key findings. First, selection of specific somatic events has been observed across tumour types (Stratton et al., 2009, Lawrence et al., 2014). Second, clonal populations of cancer cells have been demonstrated both through copy number and next generation sequencing studies (Greaves and Maley, 2012, Greaves, 2015). Third, exogenous and endogenous genomic instability processes, providing the fuel for somatic evolution, have been documented across a wide variety of cancers (Alexandrov et al., 2013a).

1.2.1 Types of cancer-causing mutations: drivers and passengers

In keeping with the evolutionary synthesis of cancer, somatic genomic aberrations in cancer cells – ranging from single nucleotide substitutions to chromosomal aberrations – can broadly be classified as either drivers or passengers, reflecting their relative contribution to the fitness of cancer cells (Stratton et al., 2009). Driver aberrations increase cellular fitness, and thus are thought to play a significant role in promoting tumourigenesis and tumour progression, while passenger mutations confer a negligible benefit to the cancer cell (Stratton et al., 2009). Throughout this thesis, the terms cancer gene and driver gene will be used interchangeably, while recognising the fact that it is necessarily a somatic driver event that confers driver or cancer status to a gene.

This binary classification of somatic aberrations as either drivers or passengers is underpinned by the multi-stage model of cancer development, which posits the existence of a number of rate-limiting steps in tumour progression – possibly equating to a number of key driver events – that punctuate cancer development (Armitage and Doll, 1954, Foulds, 1958, Hahn and Weinberg, 2002). On the basis of age-incidence statistics, it has been suggested that the number of rate limiting steps in common solid tumours is approximately 5-7 (Miller, 1980), and such estimates are supported by recent next-generation sequencing efforts (Vogelstein et al., 2013, Kandoth et al., 2013, Tamborero et al., 2013). For example, a study of 3,053 sequenced tumour exomes from 12 major cancer types, including lung, breast and colorectal cancers, found that the vast majority of tumours (93%) harboured at least one non-synonymous mutation in a cancer gene, and the average number of somatic aberrations in cancer genes per tumour was generally low, with most tumours harbouring between 2-6 (Kandoth et al., 2013).

Given the extensive number of tumours that have already been sequenced (>10,000) it is likely that the most common cancer genes have already been identified. Indeed, it has been estimated that cancer genes mutated in >20% of tumours are approaching saturation (Lawrence et al., 2014, Vogelstein et al., 2013). However, genes mutated at frequencies of <20% are still thought to be rising, particularly those at <5% frequency (Lawrence et al., 2014). Conceivably, the tail of cancer genes, present at lower frequencies across the cancer population, may contain a large number of cancer genes that confer relatively minor fitness effects, aptly named 'mini-drivers' (Leedham and Tomlinson, 2012).

It is also worth considering that the selective advantage of a mutation is context dependent - it is intrinsically linked to the environment it enters, both in terms of the genomic and epigenetic landscape, as well as the external 'microenvironment'. As such, it is theoretically possible that a passenger mutation may become elevated to driver status later in tumour evolution, or, alternatively, the selective advantage of a particular mutation may eventually become negligible, demoting it to passenger status. The importance of context is illustrated by the fact that mutations to certain cancer genes are more prevalent in certain cancer types compared to others, for example, while mutations to *VHL* occur in the majority of clear cell renal cell carcinomas (Cancer Genome Atlas Research, 2013), they are infrequent in ovarian cancers, which almost all (96%) harbour mutations to *TP53* (Cancer Genome Atlas Research, 2011). Furthermore, particular combinations of somatic mutations co-occur more than expected by chance, such as mutations in *TP53* and *IDH1* in GBM while other combinations of somatic mutations, such as mutations in *TP53* and *CDH1* in breast cancer, are generally mutually exclusive (Kandoth et al., 2013).

1.2.2 Approaches to identifying cancer genes

The search for cancer genes remains an active area of research and many different methods have been proposed (Ding et al., 2014, Raphael et al., 2014). Broadly, three distinct, yet complementary, approaches can be distinguished, based on their underlying strategies: identifying cancer genes based on recurrence and evidence of positive selection; predicting functional impact of somatic variants; and pathway or network analysis.

1.2.2.1 *Identifying drivers based on recurrence and or evidence of positive selection*

Most methods used to identify cancer genes are underpinned by the notion that these genes are likely to harbour a higher burden of coding variants than expected by chance. Thus, a key requirement of such methods is to determine the background mutation rate and what constitutes a higher mutation burden than that expected by chance (Figure 1:1).

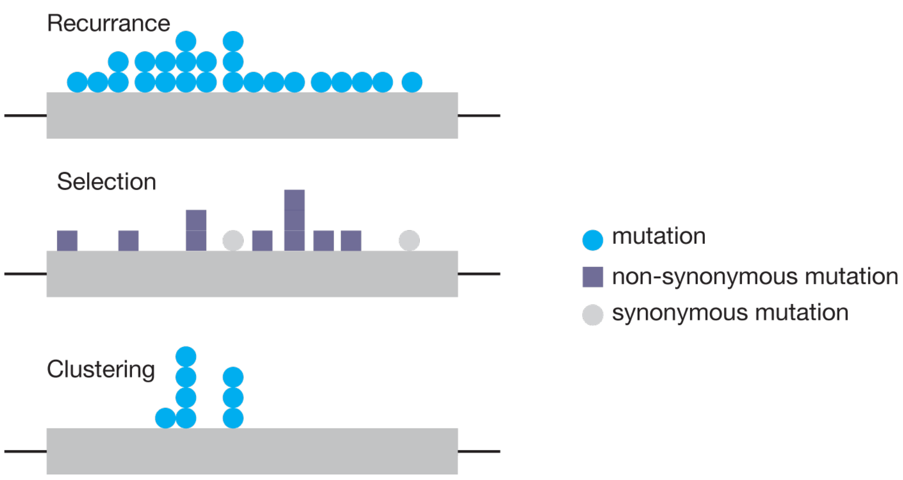


Figure 1:1 Identifying cancer genes from mutational catalogues
Cancer genes can be identified based on: exhibiting a mutation rate higher than expected; showing evidence of selection whereby a higher proportion of non-synonymous or functional mutations are observed compared to expected; or, based on mutation clustering, indicating particular amino-acid residues are being targeted.

In evolutionary biology, a common approach for identifying signatures of selection involves comparing the ratio of non-synonymous to synonymous mutations, under the assumption that synonymous mutations are selectively neutral and can be used to approximate the background mutation rate (Hurst, 2009). Such methods have been adapted to the search for cancer genes (Greenman et al., 2006, Gonzalez-Perez and Lopez-Bigas, 2012, Hua et al., 2013). Greenman et al. (2006) developed a hypothesis test in a Poisson regression framework to explore evidence of selection in genomes, accounting for differences in the mutation spectrum, and this method has since been extended and used to identify cancer genes in a catalogue of over 7,651 cancer genomes (Murtaza et al., 2013). Similar approaches are also adopted in DrGAP (Hua et al., 2013) and FM-bias (Gonzalez-Perez and Lopez-Bigas, 2012), where the main evidence for driver mutations is a signal of positive selection based on the ratio of damaging to benign mutations. An advantage of these approaches is that, in theory, they automatically adjust for gene length and genomic features, which may influence mutation rate. A drawback, however, is that they generally require large mutational catalogues or tumours with high mutation rates to accurately estimate the background mutation rate.

Other approaches attempt to use the genomic features that may influence the background mutation rate as part of a model to identify cancer genes. MuSiC for instance, takes into account many of the covariates that affect the background mutation rate, such as gene expression, to estimate per-gene and per-patient mutation rate and identify genes mutated at a high rate (Dees et al., 2012). MutSigCV also estimates gene-specific background mutation rates from multiple genomic features, including replication timing and expression of genes, as well as coverage and the rate of non-synonymous mutations (Lawrence et al., 2014). These two methods have been used extensively in large scale sequencing studies (Lawrence et al., 2014, Kandoth et al., 2013, Cancer Genome Atlas, 2012, Govindan et al., 2012, Cancer Genome Atlas Research, 2014).

An alternative, complementary, approach has been to consider the types of mutations and whether these cluster at specific sites within the gene sequence. Vogelstein proposed the 20/20 rule, postulating that a gene could be classified as a driver if either at least 20% of its missense mutations are located at a particular residue, or if at least 20% of the mutations are inactivating (for example, nonsense, frame-shift, splice site or stop codon mutations) (Vogelstein et al., 2013). A similar notion underpins OncodriverCLUST, which identifies genes harbouring a significant clustering of mutations as compared to a background model using synonymous mutations (Tamborero et al., 2013). A clustering component is also incorporated into the analysis of MutSigCV.

In general, the identification of cancer genes based on copy number analysis is also based on the notion that segments of the genome harbouring driver events are likely to be altered at higher frequencies relative to genomic segments containing only passengers (Weir et al., 2007). Recent approaches, such as GISTIC2 (Mermel et al., 2011) and JISTIC (Sanchez-Garcia et al., 2010), also incorporate sophisticated background models to enable the statistical significance of overlapping copy number segments of variables lengths to be estimated.

1.2.2.2 Predicting functional impact of somatic variants

A number of tools have been developed that seek to determine the functional impact of somatic variants, thereby providing clues as to which may be involved in cancer development. Determining which amino acid substitutions are likely to affect protein function can, for example, be assessed by presuming that important amino acids in a sequence are more likely to be conserved in the protein family than less well conserved amino acids, and SIFT (sorting intolerant from tolerant) (Ng and Henikoff, 2003), PolyPhen (Adzhubei et al., 2010), MutationAssessor (Reva et al., 2011) and Condel (Gonzalez-Perez and Lopez-Bigas, 2011) are commonly used approaches. Other tools, such as ActiveDriver (Tamborero et al., 2013), have been developed to predict effects that are related to protein aggregation, protein stability and alterations of residues targeted by post-translational modification.

1.2.2.3 Pathway or network analysis

Somatic mutations can also be considered together, enabling an examination of whether collections of genes linked to particular pathways are disrupted.

A straightforward approach is to determine whether a catalogue of coding mutations occurs more frequently in a particular gene set than expected by chance, as determined by a Fisher’s exact test (Ding et al., 2014). Indeed, such an approach is incorporated into the MuSic analysis suite (Dees et al., 2012). However, while overcoming some of the limitations of exploring individual genes, such methods ignore the fact that pathways do not act in isolation and can only ever be as useful and informative as the gene sets they use, thus restricting identification of novel combinations of genes. Other approaches therefore either rely on curated pathways to identify genes involved in cancer development or progression, or identify pathways *de novo* from the data itself (Ding et al., 2014).

Notably, with the exception of a recent study by Sakoparnig and colleagues (2015), approaches to identify cancer genes generally do not make use of evolutionary timing or heterogeneity to disentangle drivers from passengers. As such, in general, current methods are likely biased to detect clonally dominant recurrent driver events, as these events will necessarily be present in all tumour cells and will likely exhibit a stronger signal of selection. Estimates for the number of cancer genes operating in an advanced tumour may rise considerably when one considers the possibility that each subclonal population in a tumour may harbour a driver mutation (Gerlinger et al., 2014a).

1.3 Heterogeneity within tumours

1.3.1 A brief history of heterogeneity research

Within tumours, diversity can occur across multiple different biological scales, from genomic and epigenetic differences between single cells to macroscopic diversity and morphological differences between tumour regions. The biological scales at which heterogeneity can occur to a large extent reflect the history of research into intra-tumour heterogeneity itself, from macroscopic to single nucleotide, from past to present.

Macroscopic intra-tumour heterogeneity has likely been evident ever since tumours were first excised thousands of years ago (Mukherjee, 2011). Many solid tumours, for example, often display distinct regions of growth, necrosis and hypoxia. As early as the 1800s pathologists such as Rudolf Virchow, with the aid of microscopes, documented morphological diversity between tumour cells within individual tumours (reviewed in (Brown and Fee, 2006)). Today, in recognition of its clinical importance, the pathologist's grading system of breast tumours incorporates an element of heterogeneity (Russnes et al., 2011).

The notion of clonal heterogeneity and the possibility of distinct tumour subclones within a single tumour gained prominence in the 1980s and 1990s as tools were developed to investigate intra-tumour heterogeneity at the chromosomal level. Chromosome G-banding and karyotyping methods have permitted detailed analysis of the chromosomal complements of cancer cells, revealing extensive clonal heterogeneity and multiple distinct subclones in many cancer types. For example, an in-depth study of multiple regions (one from each tumour quadrant) from 10 breast carcinomas revealed shared clonal chromosomal abnormalities in the majority of samples (89%), as well as private abnormalities, not detected in all tumour regions, in seven out of the 10 carcinomas (Teixeira et al., 1996). Similarly, a study in pancreatic cancer found nineteen of 25 pancreatic carcinomas harboured at least two detectable subclones, with one tumour exhibiting a total of 58 clones (Gorunova et al., 1998).

Fluorescence *in situ* hybridisation (FISH) techniques have been used extensively to explore heterogeneity of specific genomic loci (Zojer et al., 1998, Maley et al., 2006), and, more recently, comparative genomic hybridization (CGH) microarrays have been used to dissect the extent to which copy number events may diverge in distinct tumour subpopulations (Navin et al., 2010, Benetkiewicz et al., 2006, Torres et al., 2007b). An analysis of chromosome 22 array-CGH profiling of multiple tumour regions in 15 breast tumours found four out of 15 exhibited evidence of intra-tumour heterogeneity for this chromosome (Benetkiewicz et al., 2006), while in depth array-CGH profiling across the tumour genome of 34 tumour samples from 9 breast carcinomas revealed extensive intra-tumour heterogeneity and a median of 17.4% of genetic alterations identified as heterogeneous across the cohort (Torres et al., 2007b). Sector-ploidy-profiling has also been used to dissect the clonal heterogeneity of high-grade ductal breast carcinomas (Navin et al., 2010). In this case two classes of tumour were observed: monoclonal, where only one single major clonal population was detected within the tumour, and polyclonal, in which multiple subclonal populations were observed (Navin et al., 2010).

Similarly, in GBM, multi-region copy number profiling revealed extensive heterogeneity, with different GBM subtypes identified within the same tumour (Sottoriva et al., 2013).

1.3.1.1 Next generation sequencing to explore heterogeneity and clonal architecture

The advent of next-generation sequencing has permitted analysis of heterogeneity at the single-nucleotide level across the genome and revolutionised the ability to dissect clonal heterogeneity in tumours, both from multi-region sequencing and single sample analysis. Notably, the majority of studies adopting next generation sequencing analysis, and development of associated methods, were undertaken after the start of the work for this thesis (i.e. post 2011).

One of the first studies to use next-generation sequencing to dissect the clonal architecture of a tumour explored whether mutations identified in the metastasis of a lobular breast tumour were also present in the primary tumour, which arose 9 years earlier (Shah et al., 2009). Consistent with previous work demonstrating heterogeneity at the karyotypic level (Torres et al., 2007b), only five of 32 mutations identified in the metastasis were also present at likely clonally dominant levels within the primary tumour, with 19 not detected and six present at allele frequencies indicative of subclonal prevalence. Similarly, an analysis of seven prostate cancer tumours with paired metastatic samples revealed that on average 36% (range 17-52%) of mutations identified were present in one or more of the metastatic samples exampled, but not in the primary region (Yachida et al., 2010).

In contrast to these data, which demonstrate a high degree of divergence between metastasis and primary tumour, one report documented that 48 out of 50 mutations were shared between a primary basal-like breast cancer and a brain metastasis (Ding et al., 2010). Similarly, in colorectal adenocarcinoma cancer, sanger-sequencing revealed that only a minority of mutations (~3%) identified in the metastasis were not present in the primary tumour (Jones et al., 2008).

While these studies shed light on the extent to which metastatic samples differed from their matched primary, the extent of heterogeneity within the primary tumour was not fully considered. To further elucidate the extent of heterogeneity in solid tumours, a study from our lab in KIRC (clear cell renal cell carcinoma) performed whole exome sequencing, SNP chromosome aberration analysis, and DNA ploidy profiling to multiple spatially separated biopsy samples and their associated metastatic sites in four patients (Gerlinger et al., 2012). Of all non-synonymous somatic mutations identified through multi-region sequencing, 63 to 69% of variants were heterogeneous. In every tumour region, inactivation of the tumour suppressor gene *VHL* was identified as a ubiquitous event, while mutations in other genes, such as *SET2D* and *BAP1,* were often subclonal.

The analysis of KIRC has since been extended to 10 tumours, revealing heterogeneity in each case (Gerlinger et al., 2014a), and multi-region sequencing has also been performed in relatively small cohorts of patients in high-grade serous ovarian (Bashashati et al., 2013), bladder (Nordentoft et al., 2014), colorectal (Sottoriva et al., 2015), glioma (Johnson et al., 2014, Suzuki et al., 2015), glioblastoma (Favero et al., 2015b), prostate (Cooper et al., 2015, Gundem et al., 2015), and, more recently, breast tumours (Yates et al., 2015). In every case heterogeneity has been observed, with subclonal driver events present in only a subset of cancer cells. However, given that each study has performed different types of analysis, with different numbers of regions sequenced, different tumour stage, different depth etc., it is difficult to directly compare the extent of heterogeneity across tumour types, and further studies are required to identify patterns of heterogeneity across and within cancer types.

1.3.1.2 Analysis of single tumour samples to reveal heterogeneity

Deep analysis of single biopsies has also been used to shed light on the clonal architecture of tumours and extent of intra-tumour heterogeneity.

Exome sequencing and deep re-sequencing of somatic variants in 54 triple negative breast cancer tumours, coupled with a clustering process (see Computational tools to dissect clonal architecture, below), enabled estimates of the number of subclones in single breast tumour biopsies to be made (Shah et al., 2012). The majority of tumours were found to harbour multiple subclones, with known driver mutations in cancer genes such as *TP53* tending to occur in the dominant clone (Shah et al., 2012). Applying a similar clustering method, but leveraging data from whole genome sequencing, a contemporaneous study found that for the majority of 21 breast cancer tumours, the number of subclonal mutations exceeded the number of clonal mutations (Nik-Zainal et al., 2012b). However, in each case a dominant clone, representing 50-95% of tumour cells, was detected (Nik-Zainal et al., 2012b). Heterogeneity analysis from exome sequencing has also been performed in chronic lymphocytic leukaemia (Landau et al., 2013), and multiple myeloma (Lohr et al., 2014, Bolli et al., 2014), where most tumours were found to be composed of multiple subclones. Whole genome sequencing of single biopsies from non small-cell lung cancer (NSCLC) uncovered evidence for subclonal mutations in ten out of 17 cases (Govindan et al., 2012).

Taken together, these studies demonstrate that heterogeneity has been observed across a wide variety of cancers, with both clonal and subclonal driver mutations. However, the majority of studies considering heterogeneity in detail have been limited to a small number of patients, or only investigated heterogeneity based on one sample from each tumour, thereby potentially underestimating the true extent of diversity within tumours.

1.3.2 Computational tools to dissect clonal architecture

Data available from next-generation sequencing experiments are particularly suited to statistical analysis and clonal dissection given that they represent a random sample of DNA molecules, and by extension cancer cell genomes, within a given tumour cell population. As such, the advent of next-generation sequencing has seen a surge in computational tools to explore the clonal architecture of tumours.

The fraction of reads reporting a point mutation is dependent upon the copy number at that locus, the level of tumour purity (or normal cell contamination) and finally, the cancer cell fraction, describing the fraction of cancer cells that harbour the mutation (Figure 1:2). The majority of tools to dissect clonal architecture rely on the relationship between these variables (see equation 1), to estimate whether mutations are likely clonal or subclonal.

$$\text{VAF} = p\text{*CCF} / (\text{CPNnorm*}(1\text{-}p) + p\text{*CPNmut}) \qquad \text{(equation 1)}$$

Where VAF=variant allele frequency; *p*= tumour purity; CCF = cancer cell fraction; CPNnorm = local copy number in normal genome; CPNmut = local copy number in tumour genome.

A first step in dissecting the clonal architecture of a tumour involves estimating its genomic copy number profile and also its purity. Both of these variables can be obtained from methods to estimate copy number across the genome, such as ASCAT (Van Loo et al., 2010), ABSOLUTE (Carter et al., 2012), OncoSNP (Yau et al., 2010), PICNIC (Greenman et al., 2010) or Sequenza (Favero et al., 2015a). These methods utilize mathematical frameworks to model the observed copy number array data as an amalgamation of measurements from a population of different cell types present at different proportions: tumours cells that contain an *unknown* amount of DNA, as well as an *unknown* proportion of normal cells, which have a *known* amount of DNA per cell. While the system of equations is undetermined, only a few combinations of purity and ploidy result in biologically meaningful solutions. For instance, ploidy cannot be negative or infinitely large and, in the absence of subclonal events, copy numbers must be positive integers.

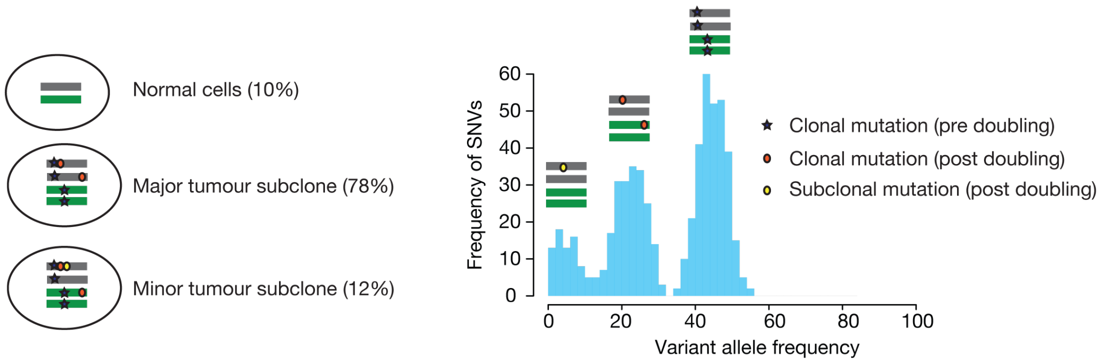


Figure 1:2 Estimating cancer cell fraction

The cancer cell fraction of each mutation reflects both the local copy number and the variant allele frequency.

Once the local copy number and purity of a sample has been determined, the cancer cell fraction and mutational multiplicity, describing the number of chromosome copies a mutation has, can be estimated from equation 1. A simple approach to assess whether a given mutation is subclonal is to assume the mutation reflects a binomial distribution and to calculate the probability that the observed variant allele frequency differs from what would be expected given a clonal mutation (Carter et al., 2012, Stephens et al., 2012).

More sophisticated methods utilize the fact that multiple mutations with similar variant allele frequencies may correspond to a clonal or subclonal cluster of mutations. For example, PyClone (Shah et al., 2012, Roth et al., 2014) pioneered the integration of variant allele frequencies with allele specific copy number and purity estimates in order to define the subclonal composition of individual biopsies. The method uses a Bayesian Dirichlet clustering process to jointly group deeply sequenced (>100x) mutations, and infers posterior density estimates over the cancer cell fraction (CCF) for each mutation. Indeed, modelling the number of subclones as coming from a Dirichlet process does not require knowledge of the number of subclones *a priori*, thus allowing both mutations to be assigned to clusters and the number of clusters to be inferred as part of the model (Roth et al., 2014). A limitation of PyClone, however, is that it assumes all copy number events are clonal, and deep sequencing is required (Roth et al., 2014). The method adopted by Nik-Zainal (2012b) leverages data from whole genome sequencing to circumvent the need for deep sequencing and allows mutations to reside on subclonal copy numbers. By contrast, SciClone (Miller et al., 2014) - which also applies a Bayesian clustering method - focuses exclusively on single-nucleotide variants (SNVs) in copy-number neutral, loss of heterozygosity (LOH)-free portions of the genome. Although this feature of SciClone circumvents issues associated with clonal and subclonal copy number aberrations, it means not every mutation can be associated with an SNV cluster. All these methods have also been extended to allow multiple samples over space or time to be included in subclonal clustering and this may considerably improve the accuracy of subclonal reconstruction.

Importantly, in the absence of multi-region data, it may be impossible to accurately de-convolve the subclonal structure in a tumour. For example, if two clonal populations have similar cancer cell fractions in one tumour region, they will appear as one clone. Analysis of another tumour region may enable their separation. Alternatively, further information regarding the clonal composition of tumours can be inferred based on the mutual exclusivity or co-occurrence of mutations in cancer cells, either from single cell sequencing or from ‘phasing’. Phasing involves determining whether mutations co-occur or are mutually exclusive, allowing different subclones to be delineated; if two mutations never occur together they are likely to represent distinct subclones, whereas if two mutations can be phased in the same cancer cell they are necessarily present in the same lineage (Nik-Zainal et al., 2012b). However, such approaches are currently limited to analysis of mutations in regions of hyper-mutations or high mutation burden.

The majority of computational tools focus on dissecting the heterogeneity of SNVs. More recently, tools have also been developed to explore heterogeneity at the copy number level (Ha et al., 2014, Oesper et al., 2014). For example, THetA (Tumour Heterogeneity Analysis) is an algorithm that seeks to estimate tumour, purity and clonal and subclonal copy numbers directly from DNA sequencing data (Oesper et al., 2014), while TITAN uses a hidden Markov model (HMM) framework to estimate clonal and subclonal clusters of copy numbers (Ha et al., 2014).

1.4 Types and patterns of tumour evolution

1.4.1 Branched and linear tumour evolution

Traditionally, tumourigenesis has been thought to proceed in a linear fashion, involving the sequential accumulation of mutations, accompanied by successive clonal sweeps where ancestral clones are replaced by fitter progeny. However, evidence of intra-tumour heterogeneity has led to a more complex view of tumour evolution (Greaves and Maley, 2012).

While linear evolution can give rise to heterogeneity if, for example, a daughter clone fails to outcompete its ancestral clones, the presence of multiple subclones with distinct sets of mutations is incompatible with linear progression (Figure 1:3). Indeed, distinct subclonal populations are indicative of branched tumour evolution and can be represented as phylogenetic trees, reminiscent of Darwin’s iconic speciation tree (Figure 1:3). In this case, mutations present in all cancer cells represent the trunk of the tumour phylogenetic tree, while distinct subclones reflect branches (Figure 1:3).

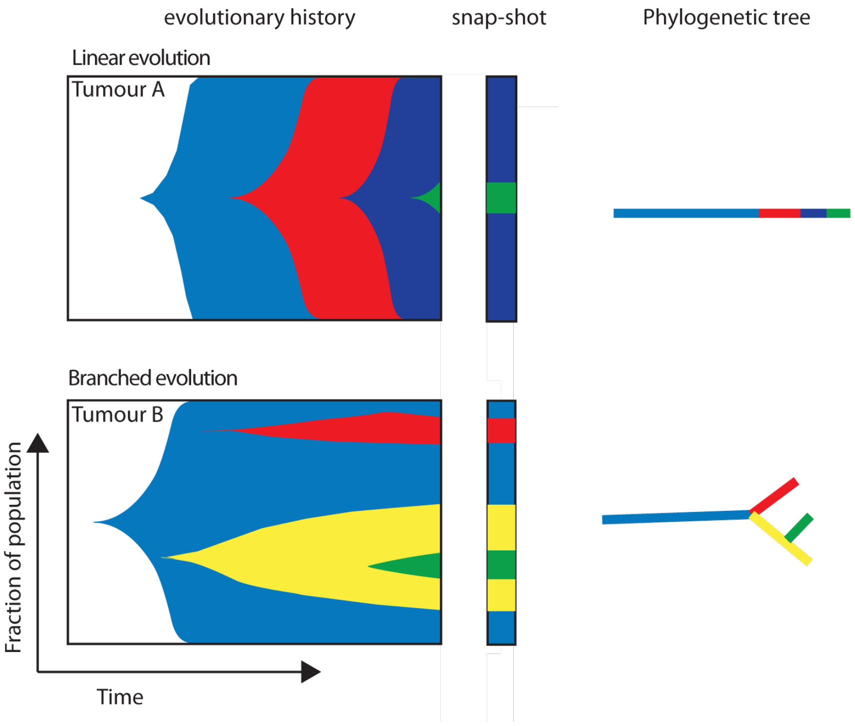


Figure 1:3 Branched versus linear evolution

Heterogeneity can be observed during linear evolution; however there will always be one subclone that harbours all the mutations in the tumour. In this case the green subclone carries all mutations that have occurred during tumour evolution. By contrast, during branched evolution distinct subclones evolve simultaneously, and at least two clones exist that harbour different private mutations.

Notably, the subclonal structure identified by sequencing a single tumour sample at one point in time is almost always compatible with linear evolution, as without information on mutual exclusivity of mutations in tumour populations (from multi-region sequencing, single cell or phasing information) it is always possible to nest subclones within each other, like Russian dolls. It is perhaps not surprising; therefore, that evidence for branched evolution has largely accumulated as a result of multi-region, serial sampling or single cell analysis.

One of the first studies to unequivocally demonstrate the presence of branched tumour evolution used FISH to explore driver copy number differences in single cells of 30 *ETV6-RUNX1* positive acute lymphoblastic leukaemias (ALL). Twenty-four malignancies were found to exhibit clonal architectures only compatible with branched evolution, while six cases followed a linear progression model (Anderson et al., 2011).

In keeping with these findings, multi-region and single-cell sequencing of tumours has also predominantly uncovered branched tumour evolution across different cancer types, including KIRC (Gerlinger et al., 2012, Gerlinger et al., 2014a), ovarian (Bashashati et al., 2013), breast (Yates et al., 2015, Wang et al., 2014), bladder (Nordentoft et al., 2014), glioma (Johnson et al., 2014, Suzuki et al., 2015), colorectal (Sottoriva et al., 2015) and prostate (Cooper et al., 2015, Gundem et al., 2015).

Despite accumulating studies documenting branched evolution, it is worth bearing in mind that the majority of studies exploring intra-tumour heterogeneity have generally been limited to a handful of tumours studied at depth. Thus the prevalence and extent of intra-tumour heterogeneity and branched tumour evolution in many cancers remains unclear. The observations that some tumour types can develop in either a linear or branched manner (Anderson et al., 2011, Navin et al., 2010) demonstrates that branched evolution is not necessarily a prerequisite for tumour development and, moreover, intra-tumour heterogeneity may not be a universal phenomenon.

Notably, the underlying reasons for branched or linear evolution remains unclear in almost all cases. In colorectal cancer it has been proposed intra-tumour heterogeneity results from clonal sweeps at early stages of evolution, followed by neutral evolution (Sottoriva et al., 2015). Conceivably, a greater understanding of patterns in tumour evolution may shed light on the mechanisms underpinning branched tumour evolution and the mechanisms driving intra-tumour heterogeneity across cancers.

1.4.2 Recurrent patterns in tumour evolution

A long-standing and contentious debate in evolutionary biology concerns whether macro-evolutionary rules and trends can be deciphered, and the veracity of Gould's famous assertion that if the tape of life were rewound and played again a different evolutionary outcome would result (Gould, 2000). Examination of rules and patterns in evolution is an emerging theme in cancer research and has important clinical implications.

A seminal study in colorectal cancer used the frequency of somatic events across independent colorectal tumours at different stages of tumour development to infer their likely temporal order (Fearon and Vogelstein, 1990). According to this model, there are two routes resulting in colorectal cancer, one through inactivation of APC, and the other through mis-match repair deficiency. Methods that rely on (and assume) common patterns across tumours have since been extended and developed into more sophisticated mathematical models (for a review, see (Beerenwinkel et al., 2014)).

The timing of mutations can also be dissected based on their copy number. For example, if a mutation precedes a regional duplication, its copy number will be doubled. Conversely, if a mutation occurs after the duplication it will be present at only one copy (Figure 1:4). Application of such a procedure revealed mutations in *TP53* mutations are early events in cutaneous squamous cell carcinomas (Durinck et al., 2011). In breast cancer, mutation data was used to time copy number events suggesting that large-scale chromosomal gains did not arise across the genome until at least 15-20% of somatic point mutations had accumulated, but subsequently were an on-going process (Nik-Zainal et al., 2012b).

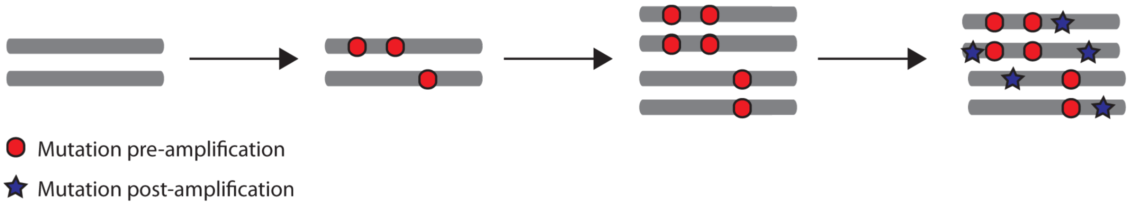


Figure 1:4 Using copy number to time mutations
Mutations occurring in regions of copy number gain can be timed based on their mutational multiplicity. Mutations occurring prior to an amplification or doubling event will be present at multiple copes (red circles), whereas those occurring after amplification will only be present at one copy (blue stars).

Finally, heterogeneity itself can also illuminate the temporal sequence of somatic events in cancer. Clonal mutations, occurring on the trunk of a tumour's phylogenetic tree, are by definition early events, whereas subclonal events, occurring on the branches, reflect later events (summarised in Table 1-1).

Multi-region sequencing of 10 KIRCs revealed that mutations in *VHL*, together with loss of chromosome 3p, appear always to occur as early events in this cancer type (Gerlinger et al., 2014a). Conversely, mutations in *TP53*, *SETD2*, *BAP1*, *PTEN* and *KDM5C* were only ever found to be subclonal, suggesting these are often later events in KIRC evolution. In other cancer types, however, *TP53* mutations have been found generally to be early events (Shah et al., 2012, Nik-Zainal et al., 2012b, Bashashati et al., 2013, Weaver et al., 2014, Yachida et al., 2010). Indeed, mutations in *TP53* were found to be the only somatic events that could predict progression from Barrett's oesophagus to oesophageal adenocarcinoma (Weaver et al., 2014), while in breast cancer and ovarian cancer, mutations in *TP53* were found to occur as clonal events (Shah et al., 2012, Bashashati et al., 2013). In prostate cancer and type one low-grade glioma tumours, mutations to *TP53* were often found to be subclonal (Hong et al., 2015, Suzuki et al., 2015). In combination, this implies that in many cancers mutations in *TP53* may be one of the founder events during tumour evolution, while in other cancers it may play a role in maintenance and progression, occurring after the emergence of the founding cancer clone.

Analysis of paired primary tumour and metastasis in colorectal cancer consistently identified driver mutations in *KRAS*, *NRAS* and *BRAF* in both contexts, suggesting these somatic events are generally early and are maintained through the metastatic dissemination process in this cancer type (Brannon et al., 2014). Conversely, in myelodysplastic syndrome (MDS), clonal analysis of single biopsies revealed that mutations in *NRAS* were often a later event while mutations in genes involved in splicing, such as U2AF1, were often the earliest (Papaemmanuil et al., 2013).

Table 1-1 Summary of truncal and branched driver events across cancer types

Tumour type	Trunk Drivers[1]	Branch Drivers	References
Acute myeloid leukemia (AML)	*DNMT3A*, *TET2*, t(15;17),t(8;21), t(16;16), inv(16)	*WT1*, *KRAS*, *NRAS*, *KIT*	(Welch, 2014)
Breast	*TP53*, *PIK3CA*	*BRCA2*	{Martins,2012; Nik-Zainal ,2012.; Shah, 2012; Yates, 2015)
Chronic lymphocytic leukaemia (CLL)	*MYD88*	*SF3B1*, *TP53*	(Landau et al., 2013)
Colorectal[2]	*KRAS*, *NRAS*, *BRAF*	*TP53*, *PIK3CA*	(Brannon et al., 2014, Vakiani et al., 2012)
Ewing Sarcoma	*EWSR1–ETS* fusion	*STAG2*	{Tirode, 2014)
Follicular lymphoma	BCl2-IGH (14;18), *MLL2*, *CREBBP*, *EZH2*	*MYD88*, *TNFAIP3*, *MYC*, *TP53*	(Okosun et al., 2014)
Glioma	*IDH1*	*SMARCA4*, *BRAF*, *TP53*, *ATRX*	(Johnson et al., 2014)
Myelodysplastic syndrome (MDS)	*SF3B1*, *SRSF2*, *U2AF1*, *DNMT3A*	*NRAS*	(Papaemmanuil et al., 2013)
Melanoma	*BRAF*	*NRAS*, *MEK1*	(Van Allen et al., 2014)
Myeloma	*IgH* rearrangement	*KRAS*, *NRAS*, *BRAF*, *FAM46C*	(Bolli et al., 2014, Lohr et al., 2014, Melchor et al., 2014)
Non-small cell lung cancer (NSCLC)	*EGFR, KRAS*, *TP53*	*HGF*	(Govindan et al., 2012)
Oesophageal adenocarcinoma	*TP53*, *SMAD4*	*MYO18B*, *TRIM58*, *CNTNAP5*, *ABCB1*, *PCDH9*, *UNC13C*, *SEMA5A* , *CCDC102B*	(Weaver et al., 2014)
Ovarian	*TP53*	*PIK3CA*, *CTNNB1*, *NF1*	(Bashashati et al., 2013, Schwarz et al., 2015)
Prostate	*ERG* rearrangements, 21q22 deletion, *NKX3-1* deletion *FOXP1*, *SPOP*	*CDKN1B*, *AR* amplification	(Baca et al., 2013, Haffner et al., 2013, Gundem et al., 2015)
Pancreatic	*KRAS*, *CDKN2A*, *TP53*, *SMAD4*	*OVCH1*	(Yachida and Iacobuzio-Donahue, 2013)
Renal	*VHL*, *PBRM1**, 3p LOH	*SETD2*, *BAP1*, *KDM5C, MTOR*, *TSC1*, *TSC2*, *TP53*	(Gerlinger et al., 2014a, Gerlinger et al., 2012)

[1] Genes with * have also been found to be subclonal in multi-region samples. LOH refers to Loss of Heterozygosity. [2] Comparative sequencing analysis was used between matched primary and metastatic colorectal lesions to define potential branched status.

In multiple myeloma (MM) known driver events, such as *BRAF* mutations, were found to be clonal in some tumours and subclonal in others, suggesting these alterations can contribute to either tumour initiation or maintenance and progression (Lohr et al., 2014, Bolli et al., 2014). Similarly, in chronic lymphocytic leukaemia (CLL), mutations in *TP53* were identified as subclonal in approximately 50% of tumours, and only mutations to *MYD88* were almost always present in the founding cancer clones when identified (Landau et al., 2013).

Taken together these data suggest rules for tumour evolution can be deciphered but are often specific within tumour types and cannot necessarily be applied across different cancers.

1.4.3 Epistatic interactions inform temporal order

Epistatic interactions, whereby the action of a gene is influenced by the genetic background, may play a key role in dictating the order in which mutations are acquired. For example, in the presence of P53, loss of *BRCA* results in acute cell-cycle arrest, thus it is likely that *TP53* mutations usually occur before BRCA loss-of-function (Ashworth et al., 2011). This is supported by observations that in breast cancer, loss of wild-type *BRCA1* often occurs after loss of *TP53* (Martins et al., 2012). Moreover the evolutionary trajectory of breast tumours reflects their subtype, with the majority of luminal tumours displaying early mutations to *TP53*, while loss of *PTEN* is observed as the first event in basal-like subtypes

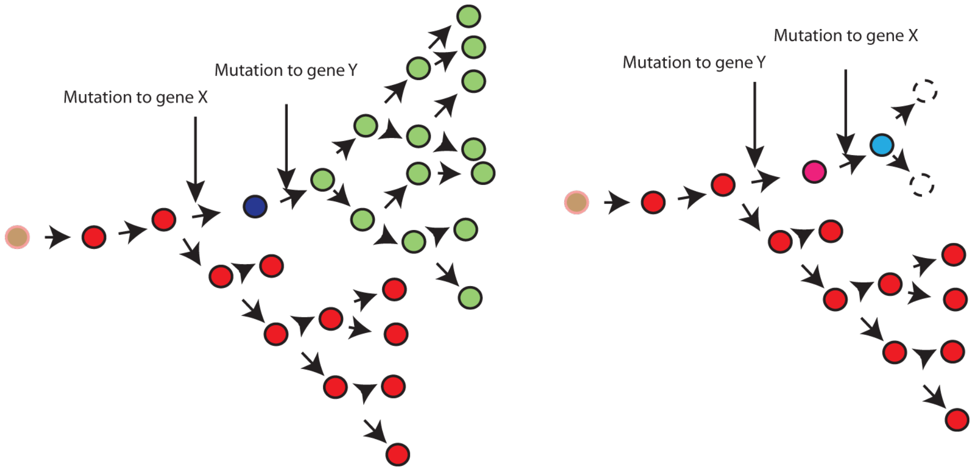


Figure 1:5 Epistatic interactions influence order of mutation acquisition
The temporal order of mutation acquisition may be influenced by epistatic interactions; for example, if gene X is mutated before gene Y, this results in a subclonal expansion, whereas if gene Y is mutated before gene X, this results in cell death.

In myelodysplastic syndrome (MDS), it was found the type of early driver mutations could play a role in dictating the future evolutionary trajectories of the disease course (Papaemmanuil et al., 2013), suggesting that certain somatic events may lead to a form of genetic canalization, in which a tumour is forced down a particular evolutionary path.

Evidence from myelo-proliferative neoplasms has reinforced this notion and further demonstrated the importance of gene order influencing the biology and outcome of tumour development (Ortmann et al., 2015). Sequencing of 246 neoplasms harbouring mutations in *JAK2*^{V617F} found that 24 also carried a mutation in *TET2.* Clonal analysis in these neoplasms revealed that twelve were characterized by early *TET2* mutations, followed by *JAK2* mutations (*TET2*-first), while in the remaining twelve neoplasms, mutations in *JAK2* were acquired first (*JAK2*-first). Intriguingly, patients harbouring *JAK2*-first mutations were significantly younger at disease diagnosis than their *TET2*-first counterparts and, moreover, the *JAK2*-first patients were characterized by an increased propensity for developing polycythaemia-vera, a condition in which red blood cells are overproduced. Moreover, in *JAK2*-first neoplasms the *JAK2* mutations increased proliferation of stem and progenitor cells, while a *JAK2* mutation did not have the same impact in neoplasms that first acquired a *TET2* mutation (Ortmann et al., 2015).

1.4.4 Parallel evolution

In evolutionary biology, parallel evolution is defined as the development of similar traits in related but distinct species, descending from the same ancestor. In tumour evolution, observations of distinct subclones converging upon disruption or deregulation of identical sets of genes or signalling pathways provide evidence for parallel tumour evolution (Figure 1:6). Parallel evolution may reflect epistatic interactions and suggest the presence of potentially recurrent patterns in tumour evolution.

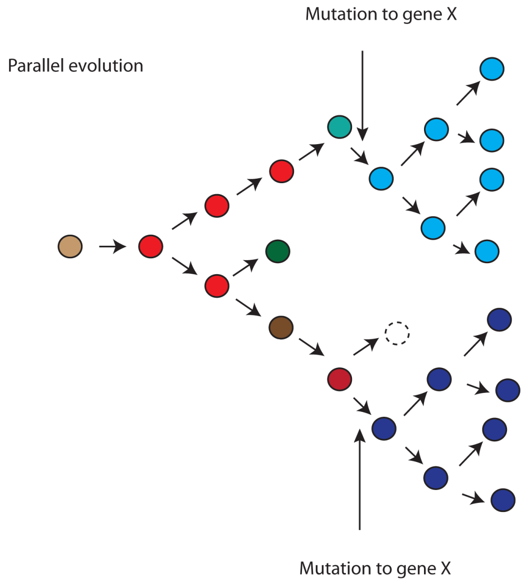


Figure 1:6 Parallel tumour evolution

Parallel evolution can occur during tumour development, in which, for example, a particular gene is mutated independently in different subclones.

In metastatic pancreatic cancer, distinct metastatic sites have been found to harbour multiple independent out of frame deletions of exon 6 of *PARK2* (Campbell et al., 2010), suggesting convergence upon inactivation of *PARK2,* while in acute lymphoblastic leukaemia (ALL), deletions in *ETV6*, *PAX5* and *CDKN2A* were found to occur independently in distinct subclones from the same tumour (Anderson et al., 2011). Our lab found evidence of parallel evolution in six out of ten KIRC tumours, with distinct somatic events in different tumour regions affecting the same gene (e.g. *SETD2*, *KDM5C*), pathway (*PIK3CA*, *PTEN*, *mTOR*) or protein complex (*PBRM1*, *ARID1A* and *SMARCA4*) (Gerlinger et al., 2014a). Moreover, an analysis of four tumours occurring in the kidneys of a patient with Von Hippel Lindau syndrome uncovered independent clonal origins in different tumours, with distinct chromosome 3p LOH events, resulting in bi-allelic inactivation of VHL. However, despite distinct tertiary driver events in every tumour, convergence for functional activation of the mTOR pathway was observed in all four tumours (Fisher et al., 2014).

In keeping with these studies, in glioma, multiple distinct somatic mutations in *TP53* and *ATRX* have been observed (Johnson et al., 2014, Suzuki et al., 2015), while in glioblastoma, a single cell sequencing approach, leveraging data from bulk genomic sequencing, documented parallel evolution with additional somatic alterations to *EGFR* following its amplification (Francis et al., 2014). Finally, in myeloma, independent subclones within the same tumour driving RAS/MAPK pathway activation through distinct RAS mutations have been observed (Melchor et al., 2014).

Ultimately, evidence for parallel evolution and patterns in the temporal acquisition of mutations emphasizes the existence of constraints to tumour development. However, the extent to which these constraints can be predicted *a priori* remains unclear.

1.5 Cancer as an ecosystem

The evolutionary synthesis of cancer can be extended to viewing cancer as an evolving ecosystem. Thus, cancer cells interact with the plethora of different cells in their microenvironment and can compete and cooperate for the resources available in order to proliferate and survive (Tabassum and Polyak, 2015).

1.5.1 Cooperation and competition between tumour subclones

The interactions between different species have been explored extensively in ecology and evolutionary biology. More recently, the concepts developed to study ecosystems have been applied to investigate cancers, in recognition of the fact that reductionist view of cancer is too simplistic and subclones may act in a mutualistic or a competing fashion (Tabassum and Polyak, 2015).

Co-operation, and potential mutualism, between tumour subclones has been documented in both mouse Wnt-driven mammary tumours (Cleary et al., 2014), and in *Drosophila melanogaster* where distinct tumour clones harbouring RASV12 and scribbled loss of function somatic aberrations have been found to cooperate to induce JNK signalling and activation of growth promoting cytokines (Wu et al., 2010). Likewise, co-operation has been observed in a glioblastoma tumour, where a low frequency subclone bearing an EGFRvIII mutation contributed to the growth of the dominant subclone through paracrine mechanisms (Inda et al., 2010). Consistent with this, evidence from a zebra-fish melanoma xenograft suggested subpopulations with limited invasive potential acting in isolation could co-invade in a symbiotic fashion (Chapman et al., 2014).

In a mouse model of small cell lung cancer, evidence was found of crosstalk between two histopathologically distinct populations of neuroendocrine and mesenchymal cells, sharing the same genetic origin (Calbo et al., 2011). Intriguingly, the neuroendocrine cells only acquired metastatic potential when the two cellular populations were engrafted together (Calbo et al., 2011). Similarly, in colorectal cancer it was found that low frequency KRAS mutant subclones – that are resistant to cetuximab – can support the survival of KRAS wild type, drug sensitive, subclones through the paracrine release of TGF beta and amphiregulin (Hobor et al., 2014).

Potentially clonal cooperation may apply to many aspects of tumour growth and progression, as well as emergence of resistance phenotypes. However, it is worth noting that tumour populations with a reliance on clonal cooperation must maintain a dynamic equilibrium and disruption of this may lead to tumour collapse through clonal interference, an example being if a non-cell-autonomous driver subclone is outcompeted by a subclone with higher proliferative potential that cannot survive independently (Marusyk et al., 2014).

Taken together, these data suggest tumours represent a complex dynamic ecological system where heterogeneity is not only a substrate for evolution, but can also promote or even be a requirement for continued tumour development and progression.

1.5.2 Cancer and the immune system

The view of cancer as an ecosystem also encourages an appreciation of the microenvironment of tumours and the interaction between cancer cells and the immune system.

Cancers can only arise if they can avoid predation by the immune system, and successfully achieve immune escape (Schumacher and Schreiber, 2015). Just as organisms have evolved a wide variety of mechanisms to avoid predation, tumour cell escape can also occur through a multitude of different methods (reviewed in (Schreiber et al., 2011)) which are not necessarily mutually exclusive. For instance, a tumour cell may acquire alterations resulting in reduced immune recognition (such as loss of tumour antigens) or alterations leading to enhanced resistance to the cytotoxic effects of the immunity may be acquired (for example, through activation of anti-apoptotic mechanisms) (Schreiber et al., 2011). In addition, escape may be achieved by creating an immunosuppressive state within the tumour microenvironment by effectively modulating the immune system (Quezada and Peggs, 2013). For example, stimulation of regulatory T cells, such as $CD4^+$ T cells, can result in the inhibition of tumour specific T lymphocytes through, for example, expression of the negative co-stimulatory molecules CTLA-4, PD-1, and PD-L1 (Quezada et al., 2011).

1.5.2.1 Neo-antigens and immune escape

Although a large proportion of all somatic variants may ultimately be silent, occurring in genes that are transcriptionally dormant or not yielding altered amino acid sequences, a subset of all somatic variants will result in altered protein products that are expressed. If processed in a specific way, peptides, harbouring the mutated allele, may encounter HLA molecules present within a cell. Depending on the binding affinity of the peptide to the HLA molecule, some of these may be presented to a patient's T-cell compartment. Given that peptides resulting from mutated amino acid sequences are necessarily different from their wild-type counterparts, these ‘neo-epitopes’ have the potential to give rise to an immune response, and thus may be considered ‘neo-antigens’.

The clinical importance of the neo-antigenic repertoire contained within a tumour will depend on the inflammatory environment and, moreover, whether immune-regulatory checkpoints are permissive for T cell function (Quezada and Peggs, 2013). A recent study, using a B16 murine melanoma model, identified tumour specific HLA-binding neo-antigens and found that cytotoxic T lymphocyte responses were induced in mice immunized with the mutated peptide but not the wild-type equivalent, and disease could be controlled therapeutically and prophylactically (Castle et al., 2012). An analysis of five cancer types found tumours with predicted neo-antigens exhibited an improved prognosis compared to tumours without neo-antigens (Brown et al., 2014). However, given that only a limited number of cancers were investigated, it was not possible to reliably assess the applicability of these findings within individual cancer types, and moreover, whether the number of neo-antigens a tumour harbours is clinically significant. More recently, a pan-cancer found the number of neo-antigens correlated with cytolytic activity – as measured by expression of granzyme A (GZMA) and perforin (PRF1) – and neo-antigen load was lower than expected in colorectal cancer, indicative of immune-editing (Rooney et al., 2015).

The relevance of modulating the immune system is underscored by the interest in and success of trials that attempt to remove the immunological brakes that block the induction of anti-tumour responses, for example through inhibition of CTLA-4 or PDL-1 (Quezada and Peggs, 2013, Soria et al., 2015). Moreover, in support of the importance of immune modulation and neo-antigens, in both melanoma and NSLCC a relationship between number of predicted neo-antigens and progression free survival in the context of immune blockade has been observed (Snyder et al., 2014, Rizvi et al., 2015).

1.6 Genome instability processes

The substrate for tumour evolution is heritable somatic aberrations. These aberrations can occur at multiple genomic levels, from single nucleotide substitutions through to whole doubling events, and may be genetic or epigenetic in origin. Thus, a key component of understanding cancer evolution involves understanding the processes that give rise to somatic events, and their evolutionary importance.

The prevalence of the different possible types of mutations identified within a cancer genome can shed light on the mutational processes that have been active during a tumour's evolution. Indeed, the characteristic mutations associated with a particular genome instability process can be considered a "mutational signature", reflecting the imprint of the type of DNA damage that has occurred.

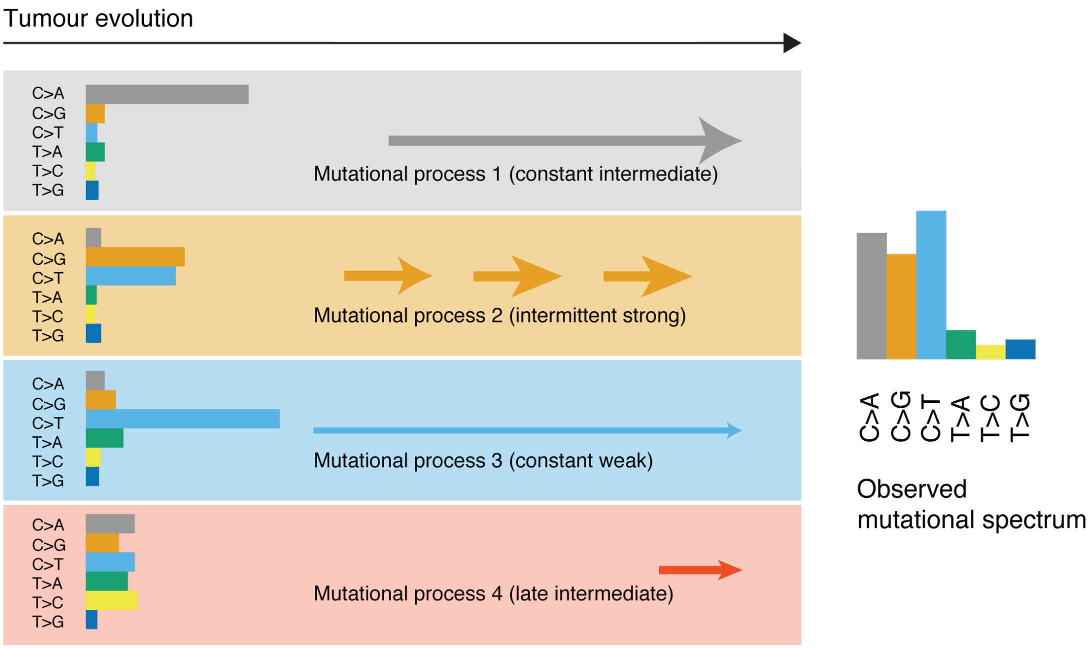

Tumour evolution
C>A
C>G
C>T
T>A
T>C
T>G
Mutational process 1 (constant intermediate)
Mutational process 2 (intermittent strong)
Mutational process 3 (constant weak)
Mutational process 4 (late intermediate)
Observed
mutational spectrum

Figure 1:7 Mutational signatures in cancer evolution

Multiple different mutational processes give rise to the mutational signature observed in the cancer genome. These processes may differ in their strength and temporal decomposition.

1.6.1 Genome instability processes at the single nucleotide level

While there are twelve possible single-base changes (C>A; G>T; C>G; G>C; C>T; G>A; T>A; A>T; T>C; A>G; T>G; A>C), these effectively reduce to six types (C:G>A:T; C:G>G:C; C:G>T:A; T:A>A:T; T:A> C:G; T:A>G:C) as, given the double stranded nature of DNA, it is impossible to distinguish on which strand the mutation occurs. The nomenclature used to describe mutations varies considerably. Within this thesis, mutations will be described from the context of the purine base that is mutated, for example, a C:G>A:T mutation will be described as a C>A transversion, despite the fact that this necessarily also defines a G>T transversion.

Historically, the exploration of mutational spectra has been restricted to analysis of targeted sequencing of established cancer genes (DeMarini et al., 2001, Giglia-Mari and Sarasin, 2003, Schwemmle and Pfeifer, 2000). Using the catalogue of mutation documented in the tumour suppressor gene *TP53,* clear differences in the mutational spectra between lung carcinomas and other cancers were observed, with C>A transversions more prevalent in tumours from smokers (Schwemmle and Pfeifer, 2000). Furthermore, even at codons that are common hotspots across a variety of cancer types, a preponderance of C>A transversions in lung cancers of smokers was identified compared to other tumours and never-smokers. The C>A transversions derived from smokers' tumours have been found to exhibit a strong transcriptional strand bias with fewer G>T transversions on the transcribed than the non-transcribed strand, likely reflecting the past activity of transcription-coupled nucleotide excision repair on bulky adducts of guanine caused by tobacco carcinogens (Hainaut et al., 2001). Ultraviolet (UV) light associated damage has been shown to induce C>T and CC>TT transitions in melanoma genomes, and also exhibited transcriptional strand bias, with fewer C>T mutations on the transcribed compared to the non-transcribed strand, probably due to the action of transcription-coupled repair on impaired pyrimidines (Pfeifer et al., 2005).

While informative, a limitation of these single gene studies is that they rely on mutations in driver genes, making it difficult to disentangle the effects of selection with the mutational signatures operating. Moreover, aggregate signals from thousands of samples are required, meaning only strong exposures or dominant repair processes operating across the majority of tumours can be accurately deciphered.

1.6.1.1 Whole-genome and whole-exome analysis of mutational catalogues

Analyses of mutational catalogues obtained from whole-genome sequencing of tumours with high mutation rates can provide a detailed picture of the processes operating. For example, two studies involving sequencing of a malignant melanoma and a single lung cancer were some of the first to illustrate the power of this approach (Pleasance et al., 2010a; Pleasance et al., 2010b). These two studies identified the characteristic mutational spectra of ultraviolet light and tobacco carcinogens respectively within single tumours. Since these seminal studies, further analysis of a large series of sequenced lung tumours has revealed that smokers on average exhibit a 10-fold increase in the burden of somatic mutations in their cancer genomes compared to never-smokers (Govindan et al., 2012; Imielinski et al., 2012). Moreover, consistent with previous findings and experimental evidence, this elevation is mainly due to the increase in the number of C>A transversions.

In depth examination of the mutational spectra within cancers has also highlighted key differences in mutational processes within different tumours. For example, examination of hyper-mutated endometrial and colorectal tumours has revealed preponderance of C>A and C>G mutations at TpCpT sites specifically in tumours harbouring mutations in *POLE-E* (Palles et al., 2013), demonstrating somatic aberrations can result in the preponderance of a particular mutational signature.

More recently, algorithms have been developed to quantify the number and contributions of mutational signatures operating within cancers at the single-nucleotide level (Alexandrov et al., 2013b, Fischer et al., 2013). These approaches assume recurrent processes operate across different cancers, and make use of mathematical frameworks to de-convolve the different processes operating within each cancer. Application of NMF (non-negative matrix factorization) and model selection to over 30 cancer types, represented by more than 7000 tumours, identified 20 distinct mutational signatures (Alexandrov et al., 2013a).

In the majority of cancer samples analysed, at least two mutational processes were identified, consistent with an elevated mutation rate in most cancers (Alexandrov et al., 2013a). The most widespread mutational signature, identified in 25 cancer types, was characterized by C>T transitions at CpG sites, probably reflecting deamination of 5-methylcytosines at CpG sites. This signature correlated with patient age (Alexandrov et al., 2013a), consistent with a large proportion of these mutations having been acquired prior to tumourigenesis.

Another pervasive mutational signature, identified in 15 cancer types, was characterized by C>T and C>G mutations at TpC sites. Orthogonal analysis, adopting a simpler method that used the prevalence of mutations with this motif compared to what would be expected given random mutagenesis, also identified the presence of this signature across multiple cancer types (Burns et al., 2013, Roberts et al., 2013). This signature has been linked to the family of apoliprotein B mRNA editing enzyme catalytic polypeptide-like (APOBEC) enzymes, involved in cytosine deamination. Cell line studies have demonstrated that APOBEC1, APOBEC3A, and APOBEC3B are capable of mutating DNA by the deamination of cytosine flanked by a 5’ thymine and thus result in C>T mutations at TpCpN tri-nucleotides (Harris et al., 2002, Hultquist et al., 2011, Suspene et al., 2011, Taylor et al., 2013), consistent with their role in facilitating tumour mutagenesis. Furthermore, it has been shown that the activation of enzymes APOBEC3A and APOBEC3B in yeast can also lead to in C>G at TpCpN tri-nucleotides (Taylor et al., 2013). This mutational pattern was attributed to replication over an abasic site, formed when an APOBEC deaminated cytosine is excised by uracil-DNA glycosylase, which is catalysed by REV1 (Taylor et al., 2013).

Certain mutational signatures were also identified exclusively in specific cancer types. For instance, in oesophageal carcinomas (including both adeno and squamous cell-carcinomas), a signature characterized by T>G and T>C mutations, particularly at CpTpT sites, was identified, which has been linked to gastric acid exposure (Dulak et al., 2013, Weaver et al., 2014). Similarly, the previously characterized smoking signature, involving C>A transversions, was also identified in lung and head-and-neck cancers.

In combination these studies demonstrate that many of the mutational processes underlying cancer genome evolution are beginning to be elucidated. Importantly, however, the underlying processes responsible for many of the observed signatures still remain unclear, and in general their temporal decomposition is unknown.

1.6.2 Therapy as an exogenous mutational process

Therapy may also act as an exogenous source of genome instability (Hunter et al., 2006, Johnson et al., 2014, Cahill et al., 2007, Ding et al., 2012). Examination of the clonal evolution of primary and relapsed acute myeloid leukaemia (AML) found an increase in transversion mutations, and in particular C>A mutations, following cytotoxic therapy at relapse (46%) compared with mutations prior to therapy (30.7%) (Ding et al., 2012). Consistent with this, in *Caenorhabditis elegans*, cisplatin treatment has been found to lead to a striking, dose-dependent, increase in base substitutions - predominantly C>A transversions - as well as an elevated rate of dinucleotide substitutions, indels, and structural variants (Meier et al., 2014). However, the mutational context of C>A transversions was not explored in the analysis of Ding (2012).

Temozolomide, an alkylating agent, has also been found to leave a somatic imprint in the cancer genome in the form of an elevated rate of C>T transitions - primarily at CpC and CpT sites (Alexandrov et al., 2013a, Johnson et al., 2014). Analysis of primary gliomas revealed that recurrent tumours exhibited evidence of hyper-mutation, and mutations that could be linked to temozolomide included those in *RB1* and *CDKN2A* (Johnson et al., 2014).

Thus, in these examples, conceivably therapy was not simply acting as an exogenous source of mutations, but also as a selection barrier, influencing the evolutionary trajectory of a tumour and its progression to a more aggressive phase. Consistent with this notion, in NSCLC, chemotherapy has been found to be associated with reduced prevalence of mutations in EGFR (Bai et al., 2012), and treatment of colorectal cancer clones with oxaliplatin resulted in the outgrowth of previously dormant, resting clones (Kreso et al., 2013).

1.6.3 Genome instability processes at copy number level

Genome instability processes and mutational signatures can also be deciphered through copy number analysis. Chromosomal instability (CIN), a driving force of intercellular genetic heterogeneity, occurs at high frequencies across many different cancer types and describes an increased rate of change in chromosome number or structure, resulting in a range of karyotypic abnormalities, including whole-chromosome and segmental aneuploidies, as well as translocations, inversions and deletions.

A study from our lab found the weighted genome instability index (wGII) – a measure capturing the extent of the genome that differs from ploidy – can discriminate between chromosomally unstable and stable colorectal cancer cell-lines (Burrell et al., 2013a). A signature of CIN can also be deciphered from gene expression data. Indeed, the CIN70 signature, which encompasses a set of 70 genes that correlate with an aberrant transcriptional profile (Carter et al., 2006), has been found to correlate with structural complexity (defined from SNP microarrays) and numerical CIN (as assessed by DNA image cytometry) (Birkbak et al., 2011).

Chromosomal instability likely arises through multiple mechanisms including both pre-mitotic and mitotic defects (Bakhoum et al., 2009, Crasta et al., 2012, Burrell et al., 2013a, Janssen et al., 2011, Giam and Rancati, 2015), and is often categorized as either numerical CIN or structural CIN. Numerical CIN occurs following the missegregation of whole chromosomes at mitosis. Conversely, structural CIN results in the disordered integrity of parts of chromosomes. While traditionally considered in isolation, recent studies have revealed that both types of chromosomal instability can be mechanistically related and reciprocally causative.

Gross changes in chromosome structure arising from erroneous DNA repair render them susceptible to missegregation (Burrell et al., 2013a, Pampalona et al., 2010), and conversely, whole chromosome missegregation due to mitotic defects including merotelic attachments, alterations of the spindle assembly checkpoint (SAC) and centrosome number and positioning defects (Bakhoum et al., 2009, Silkworth et al., 2009, Sotillo et al., 2007), expose chromosomes to mechanical stress, DNA damage and erroneous repair, engendering structural CIN (Crasta et al., 2012, Janssen et al., 2011).

The mechanisms underpinning CIN may also leave specific genomic imprints in the genome, which can be deciphered from sequencing or SNP6.0 analysis. For example, a key source of DNA damage is homologous recombination (HR) deficiency, which has been shown to result in a specific copy number profile involving loss of heterozygosity (Birkbak et al., 2012, Popova et al., 2012, Abkevich et al., 2012). As such, the telomeric imbalance score, assessing the number of sub-chromosomal regions with allelic imbalance extending to the telomeres is thought to be indicative of impaired DNA repair (Birkbak et al., 2012). The clinical importance of this HR signature is underscored by the observations that it predicts cisplatin sensitivity *in vitro* and response to preoperative cisplatin treatment in patients with triple-negative breast cancer (Birkbak et al., 2012).

Another likely source of much of the DNA damage in tumours is DNA replication stress. DNA replication stress is also thought to lead to a specific copy number profile involving enrichment of chromosome breakpoints and loss of heterozygosity in cancer genomes at fragile sites (regions of the genome that are particularly susceptible to replication stress) (Dereli-Oz et al., 2011). In keeping with the importance of replication stress, in general, copy number aberrations are enriched around fragile sites, in large genes and often display loss of heterozygosity.

These studies demonstrate the presence of a number of different mutational signatures at the copy number level and highlight the importance of chromosomal instability. However, copy number data has not been subject to the same level of analysis as SNV data. Moreover, from a biological perspective, it is important to recognise that cellular adaptations are likely to be required for cancer cells to tolerate the gross genome remodelling associated with CIN (Mcclelland et al., 2009, Gordon et al., 2012), and without such tolerance these signatures would not be observed in cancer genomes. Thus, for example, a signature of chromosomal instability may not only indicate the presence of karyotypic instability, but also that the cancer cell can tolerate it.

One of the most frequently studied tolerance mechanisms is disruption to *TP53*, the so-called 'guardian of the genome' (Lane, 1992). Indeed, it has been found p53 can limit the proliferation of cells with an abnormal chromosome complement and it is up-regulated in response to an unstable karyotype (Giam and Rancati, 2015). Thus, disruption of p53 may facilitate chromosomal instability, permitting cancer genome evolution at the chromosomal level. However, it is likely additional tolerance mechanisms, beyond *TP53* disruption, remain to be elucidated.

1.6.4 Tetraploidy and whole genome doubling

Whole genome doubling events have been crucial in the development of many different complex traits during organismal evolution, and a number of tetraploidisation events have been thought to occur early in vertebrate evolution (Huminiecki and Conant, 2012). Moreover, genome-doubling events are thought to facilitate the rise of novel gene functions, allowing functional divergence of duplicated genes (Huminiecki and Conant, 2012, Adams and Wendel, 2005).

Whole genome doublings have been documented to occur frequently across a range of cancers and can be estimated from allele specific copy number data (Carter et al., 2012, Zack et al., 2013). Across common epithelial tumours, including colorectal, breast, lung, ovarian and oesophageal, it was estimated that genome doubling occurs in approximately 50% of tumours. In this case tumours were classified as genome doubled if over 50% of the genome exhibited a major allele copy number of at least two.

It has been posited that a genome-doubling event may be important as an intermediate stage leading to aneuploidy and CIN in cancers (Shackney et al., 1989, Galipeau et al., 1996). In a seminal study comparing tetraploid and diploid cancer cells, it was found that tetraploid cells formed tumours in p53-null nude mice, exhibiting both numerical and structural chromosome aberrations, while their diploid counterparts did not (Fujiwara et al., 2005). Moreover, tetraploid cells generated through cell fusion events resulted in karyotypically diverse tumours in mice (Duelli et al., 2007, Nguyen et al., 2009). Likewise, the continual passage of tetraploid mouse ovarian cells results in aneuploid progeny (Lv et al., 2012).

Tetraploid cells have also been observed following telomere crisis, and it has been shown that tumours that are initiated by tetraploid cells exhibit more complex aneuploidy karyotypes *in vivo* (Davoli and de Lange, 2012). More recently, mathematical modelling of tumour progression in neuroblastomas, combined with karyotypic analysis of tumours suggested that an early genome-doubling event was the most parsimonious explanation for the chromosomal complement observed in neuroblastoma tumours (Lundberg et al., 2013).

Taken together, these studies suggest genome doubling may be a frequent event across cancers. However, the relationship between genome doubling and chromosomal instability remains unclear, and the timing of genome doubling across cancers is unknown.

1.7 Clinical implications of intra-tumour heterogeneity

1.7.1 Tumour sampling bias

Given that therapeutic decision-making is frequently based on archival tumour material or, in cases where patients present with advanced disease, a single metastatic sample, tumour sampling bias may present a significant clinical challenge (Swanton, 2012). Indeed, if different regions of the same tumour are genetically distinct, one sample will not be representative of the tumour and, by extension, if biomarkers are not clonal, sampling bias may confound clinical decision-making.

In support of difficulties associated with validating clinical biomarkers, potentially due to sampling bias, it has been documented that while over 150,000 biomarkers have been proposed in the literature, approximately only 100 of these are used in clinical practice (Poste, 2011). Moreover, evidence for multiple distinct prognostic biomarkers within single tumours can be found in KIRC, where seven out of ten tumours were found to harbour distinct gene expression signatures, one associated with good prognosis, and the other associated with poor prognosis (Gulati et al., 2014). Likewise, an analysis of 11 glioblastomas revealed that most patients harboured multiple different subtypes within the same tumour (Sottoriva et al., 2013).

Sampling bias may also lead to underestimates for the prevalence of particular mutations in cancers. Notably, in KIRC, an analysis in our lab found that single sample analysis suggests the prevalence of mutations in *TP53* is only 10%, while multi-region analysis, with more complete sequencing of cancer cells, identified mutations in *TP53* in approximately 40% of tumours (Gerlinger et al., 2014a).

1.7.2 Heterogeneity and outcome

1.7.2.1 Clinical relevance of chromosomal instability

It has long been established that chromosomal instability – resulting in cell-to-cell genetic heterogeneity – is associated with poor prognosis across a wide range of cancers (see Table 1). For instance, multiple studies in NSCLC using FISH and gene expression signatures to measure CIN have found an association between CIN status and overall survival, independent of conventional risk factors, including tumour stage (Carter et al., 2006, Mettu et al., 2010, Choi et al., 2009). In breast cancer, CIN has been linked to poor prognosis and also is enriched in more aggressive subgroups of breast cancer, such as ER-negative and triple-negative.

Furthermore, evidence has suggested that this form of genomic instability is associated with multi-drug resistance in colorectal cancer (Lee et al., 2011a).

The potential benefits of chromosomal instability, and intra-cellular genetic heterogeneity, must be balanced against its potentially deleterious consequences. An abnormal chromosome number in yeast and murine systems has been demonstrated to be deleterious to normal cells, reducing proliferation (Torres et al., 2007a, Williams et al., 2008). However, other studies have suggested chromosomal instability may redress imbalances in the stoichiometry of protein complexes which can result from an abnormal chromosome complement and CIN may effectively allow proliferation genes to be hardwired to the genome (Ozery-Flato et al., 2011). It has thus been proposed there is an optimal level of CIN in tumours, beyond which it becomes unfavourable (Swanton, 2012, Cahill et al., 1999). The notion of extreme CIN being deleterious for tumour development may be analogous to 'mutational meltdown' in bacteria, or error catastrophe in viruses (McGranahan et al., 2012).

Consistent with excessive levels of CIN having adverse consequences for tumour progression, excessive CIN, induced by inactivation of spindle assembly checkpoint components, leads to excessive aneuploidy and cell death in human cancer cells, and multipolar cell divisions generate non-viable cells that are highly aneuploid (Giam and Rancati, 2015). Similarly, mice with reduced levels of CENP-E – involved in spindle elongation – develop tumours with CIN. However, if CIN is increased through depletion of CENP-E in tumours that already had a pre-existing level of aneuploidy, the depletion can have tumour-suppressive abilities (Sotillo et al., 2007).

Table 1-2 Clinical significance of chromosomal instability

Cancer Type	Method of measuring CIN	CIN associated with	Reference
Lung cancer (NSCLC)	FISH (n=63)	Poor prognosis (OS & DFS)	(Choi et al., 2009)
	FISH (n=47)	Poor prognosis (OS)	(Yoo et al., 2010)
	FISH (n=50)	Poor prognosis (OS)	(Nakamura et al., 2003)
	12-gene signature (n=647)	Poor prognosis (OS)	(Mettu et al., 2010)
	CIN70 signature (n=62)	Poor clinical outcome	(Carter et al., 2006)
Breast cancer	SSI (n=890)	Poor prognosis (OS)	(Kronenwett et al., 2004)
	SNP (n=313)	Poor prognosis (MFS)	(Smid et al., 2011)
	12-gene signature (n=469)	Poor prognosis (DFS & RFS)	(Habermann et al., 2009)
	CIN70 signature (n=1866)	Poor clinical outcome	(Carter et al., 2006)
	FISH (n=31)	Lymph-node metastasis and ER negativity.	(Takami et al., 2001)
Myelodysplastic syndrome	FISH (n=65)	Poor prognosis (DFS)	(Heilig et al., 2010)
Endocrine pancreatic tumours	CGH (n=62)	Metastasis	(Jonkers et al., 2005)
Colon cancer	12 gene genomic instability signature (n=92)	Recurrence of colon cancer	(Mettu et al., 2010)
	Flow cytometry/ image cytometry (n = 10 126)	Poor prognosis.	(Walther et al., 2008)
Ovarian cancer	12-gene genomic instability signature (n=124)	Poor prognosis (RFS).	(Mettu et al., 2010)
Endometrial cancer	SNP (n=31)	Poor prognosis (OS).	(Murayama-Hosokawa et al., 2010)
Synovial sarcoma	CGH (n=22)	Poor prognosis (OS)	(Nakagawa et al., 2006)
Oral cancer (SCCs)	FISH (n=77)	Poor prognosis (OS & DFS)	(Sato et al., 2010)
	FISH (n=20)	(Loco) regional tumour outgrowth	(Bergshoeff et al., 2008)
Diffuse Large B-cell Lymphoma	Anaphase segregation errors(n=54)	Poor prognosis (RFS)	(Bakhoum et al., 2011)
Cervical cancer	CIN70 signature (n=79)	Para-aortic nodal relapse	(How et al., 2015)

Abbreviations: NSCLC, non-small cell lung cancer; SCC, squamous cell carcinoma; FISH, fluorescence *in situ* hybridization; SSI, stem line scatter index; CGH, comparative genome hybridisation; SNP, single-nucleotide polymorphisms, OS, overall survival; DFS, disease-free survival; MFS, metastasis-free survival; RFS, relapse free survival

Evidence substantiating this hypothesis has been found when further dissecting the relationship between CIN and cancer outcome. Specifically, patients who harbour tumours with extreme CIN, defined as tumours in the upper quartile of CIN70 expression, exhibit significantly better prognosis, in terms of recurrence-free or distance metastasis-free survival, compared to patients harbouring tumours in the third CIN70 expression quartile (Birkbak et al., 2011). This association has been observed in ER negative breast, ovarian, gastric, breast and non-small cell lung cancers, but not in ER positive breast cancers (Birkbak et al., 2011).

1.7.2.2 Clonal heterogeneity and outcome

One of the first studies to specifically evaluate the clinical relevance of clonal heterogeneity adapted diversity measures from ecology and evolution to explore whether these could predict progression to adenocarcinoma in the premalignant condition Barrett's oesophagus (Maley et al., 2006). Using the Shannon diversity index to capture both the number and abundance of clones, it was found that using multiple different types of somatic alterations, clonal diversity was predictive of progression to cancer in a large cohort of patients (Maley et al., 2006). These results have since been confirmed in an independent cohort of tumours, encompassing 239 patients (Merlo et al., 2010).

More recently, in CLL, it has been found that the presence of subclonal drivers is associated with a shorter time to retreatment or death (Landau et al., 2013) while in head-and-neck cancer, a measurement capturing the clonal diversity – termed mutant allele tumour heterogeneity (MATH) – was found to correlate with poor prognosis (Mroz and Rocco, 2013). However, in MDS, the number of driver events was the key determinant of outcome, regardless of their clonal status; i.e. the presence of a driver was more critical than whether it was subclonal or clonal (Papaemmanuil et al., 2013).

Clonal heterogeneity may also impact upon the efficacy of therapeutic treatments, through the presence of subclones harbouring resistance mutations, which may be barely detectable at diagnosis. In NSCLC with activating mutations in *EGFR* presence of subclonal gatekeeper T790M resistance mutations are associated with shorter progression free survival (Maheswaran et al., 2008, Su et al., 2012). In a small cohort of high grade serous ovarian cancers (n=14) treated with platinum-based chemotherapy a copy number based clonal heterogeneity index was found to have predictive value for survival after chemotherapy treatment (Schwarz et al., 2015).

Evidence in colorectal cancer (Diaz et al., 2012), as well as melanoma (Van Allen et al., 2014, Shi et al., 2014) suggests that multiple resistance events can occur independently in the same tumour. For instance, following BRAF inhibitor therapy in *BRAF* V600 mutant melanoma, resistance mutations in both *NRAS* and *MEK1* were identified in one tumour while another patient presented with a tumour harbouring two distinct *NRAS* mutations (Van Allen et al., 2014). It has also been demonstrated that resistance to BRAF inhibitors can occur through both MAPK pathway dependent and PI3K-AKT dependent mechanisms in the same tumour simultaneously (Shi et al., 2014). Likewise, in one patient with colorectal cancer, through longitudinal tracking of cell free tumour DNA, four distinct *KRAS* mutations, that increased in frequency during the acquisition of resistance to panitumumab therapy (targeting EGFR), were detected (Diaz et al., 2012).

1.8 Summary

In this chapter, I have provided a literature review of studies exploring cancer evolution and heterogeneity in cancer. I have situated this by considering the mutational processes that mould the cancer genome at both the single nucleotide and whole genome level.

Taken together, this work demonstrates that clonal diversity and branched tumour evolution is an emerging theme across studies. However, a pan-cancer analysis exploring the extent of diversity across tumours in a unified way is lacking. Additionally, the extent to which a single biopsy may underestimate the extent of heterogeneity remains unresolved. Moreover, while mutational processes at both the single nucleotide and copy number level have been deciphered, their temporal decomposition is unresolved. Finally, while studies have started to shed light on the clinical importance of neo-antigens – especially in the context of immune modulation – the impact of intra-tumour heterogeneity on this relationship has not been investigated.

In the subsequent chapters of this thesis, I will build upon the work outlined here to explore the extent of heterogeneity in human cancers, focussing in depth on

NSCLC and oesophageal adenocarcinoma, and how this information can be used to understand tumour evolution. I will perform a pan-cancer analysis of somatic intra-tumour heterogeneity, exploring the extent to which tumours harbour subclonal mutations in driver genes, and, moreover, whether temporal dissection can assist in the identification of cancer genes. I will explore the incidence and timing of genome doubling across cancers, and systematically investigate the effects of a whole genome-doubling event on chromosomal instability and tumour evolution in colorectal cancers. Temporal dissection of mutations will be further used to shed light on the timing of mutational processes in cancer, revealing which likely occur as early events, and which represent later processes. Finally, I will investigate the clinical implications of heterogeneity in the context of immune modulation.

Chapter 2. Experimental procedures

2.1 Introduction

While the heterogeneous nature of tumour populations may confound treatment success, it may also serve a role in shedding light on tumour evolution. As discussed in the previous chapter, an understanding of the extent of diversity within individual tumours can both be used to temporally dissect mutations, and, moreover, reveal the type of evolution that has occurred. Thus, an understanding of how to dissect the extent of heterogeneity and the evolutionary history of tumours is crucial.

In this chapter, I outline the bioinformatics and experimental methods adopted throughout the thesis. First, I briefly describe the main datasets used throughout the proceeding chapters. Second, I explain how heterogeneity is determined and measured using both single and multi-region samples. I also provide details regarding the analysis of mutational signatures and their temporal decomposition, and in addition, I explore how copy number data can be used to infer genome doubling. Finally, I provide brief details regarding the analysis of neo-antigens and the wet-lab experimental procedures performed.

2.2 Data used in this thesis

2.2.1 Multi-region NSCLC data

The multi-region NSCLC tumours for sequencing were obtained from six patients diagnosed with NSCLC who underwent definitive surgical resection prior to receiving any form of adjuvant therapy, such as chemotherapy or radiotherapy. These samples were collected from University College London Hospital, London (UCLHRTB 10/H1306/42) and informed consent had been obtained. Details regarding patient characteristics can be found in Table 3-1.

2.2.2 Multi-region oesophageal adenocarcinoma data

Pre-treatment tumour regions were obtained endoscopically from a single tumour mass and post-chemotherapy tumour regions were obtained from the surgical tumour resection. Patients were treated with neo-adjuvant combination chemotherapy and did not receive any concurrent radiation treatment. Details regarding patient clinical characteristics can be found in Table 3-2

2.2.3 TCGA data

TCGA data was obtained from the TCGA data portal, or, alternatively from the broad maf dashboard (https://confluence.broadinstitute.org/display/GDAC/MAF+Dashboard). Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at http://cancergenome.nih.gov/.

2.2.4 Additional lung single sample data

To explore the mutational spectrum of lung cancer genomes, data was obtained from the supplementary tables of Imielinksi et al. (2012).

2.2.5 Validation data for Genome Doubling

Validation cohort data was kindly provided by Dr Oliver Sieber, and colleagues at the Victorian Cancer Biobank (Australia). Patients were recruited from the Royal Melbourne Hospital, Western Hospital Footscray and St Vincent's Hospital Sydney in Australia. The study was ethics approved, and patients gave informed consent.

2.2.6 NSCLCs treated with pembrolizumab

Samples obtained from Rizvi and colleagues (2015) reflected a patient cohort of stage IV NSCLC, and detailed clinical characteristics of this cohort are provided in Table 8-11.

2.3 Bioinformatics analyses

All bioinformatics analysis was carried out in the R statistical environment, unless otherwise specified.

2.3.1 SNV calling for TCGA tumours

For TCGA samples, SNV calling was performed by individual sequencing centres. To ensure consistency, further filtering was also applied. Specifically, every SNV was filtered to ensure that each variant had at least five tumour reads and coverage of ≥30. Only single nucleotide variants were used for analysis (insertions and deletions were excluded due lower validation rates).

2.3.2 SNV calling from multi-region NSCLC WES and WGS data

The pipeline for multi-region SNV calling was developed by Richard Mitter and Gareth Wilson. For multi-region NSCLC tumours, raw paired end reads in FastQ format generated by the Illumina pipeline (WES or WGS) were aligned to the full hg19 genomic assembly (including unknown contigs), using bwa 0-5.9 (Li and Durbin, 2009) with a seed length of 72bp for data sequenced on the GAII and 100bp for data sequenced on the HiSeq. Up to 3 or 4 mismatches were allowed per read for the GAII or HiSeq respectively, and all other settings were left as default. Picard tools v1.8 was used to both merge samples from the same patient region and to remove duplicate reads (http://picard.sourceforge.net) prior to determining sequence coverage (Table 8-1).

Variant calling was performed between tumour and matched germ-line using the “somatic” tool from VarScan2 v2.3.3 (Koboldt et al., 2012). SAMtools mpileup output from combined tumour and normal samples generated by skipping bases with a phred score of <20 or reads with a mapping-quality <2 was used as input for VarScan2. SAMtools BAQ computation was disabled and the coefficient for downgrading mapping quality was set to 50. VarScan2 somatic was run with default settings, except for the following: minimum coverage was set to 10 for germline and 6 for tumour regions, the minimum variant frequency for calling a heterozygote was set to 0.01 and tumour purity was set to 0.5. Calls were subsequently filtered for false positives using Varscan2's fpfilter.pl script, run with the settings outlined by Ding (2012). Additionally, variants were only accepted if present in ≥ 5% of reads in at least one tumour region and present with ≤2 reads in germ-line and ≥ 2 reads in at least one tumour region.

Small insertions and deletions (indels) were identified using pindel version 0.2.4 (Ye et al., 2009) in paired tumour-normal mode. Variants were annotated using both ANNOVAR (Wang et al., 2010) and dbNSFP (Liu et al., 2011). Each variant identified as non-silent was manually reviewed by Elza de Bruin, using Integrated Genomics Viewers (IGV) (Robinson et al., 2011).

Variants not subjected to ultra-deep orthogonal validation were further filtered using an in-house filter, VarSLR, developed by Max Salm, which models strand-bias, mapping-quality, base-quality and position-in-read in a stepwise logistic regression framework. Any substitution identified in dbSNP Build 132 was removed. For the WGS data, variants detected in repetitive regions (RepeatMasker, USCS genomicSuperDups tracks) or blacklisted by the Encode Mappability were also removed from downstream analysis.

2.3.3 SNV calling from multi-region oesophageal adenocarcinoma tumours and NSCLC tumours from Rizvi et al. (2015)

SNV calling for multi-region oesophageal adenocarcinoma and NSCLC tumours obtained from Rizvi et al. (2015) was carried out as described above, with a number of modifications. Specifically, bwa mem (bwa-0.7.7) (Li and Durbin, 2009) was used for genome alignment. Picard tools v1.107 was used to clean sort and merge files from the same patient region and to remove duplicate reads (http://broadinstitute.github.io/picard). Quality control metrics were obtained using a combination of picard tools (1.107), GATK (2.8.1) and FastQC (0.10.1) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

For those variants not subjected to Ion Torrent validation, further filtering was applied whereby variants were only accepted if present in ≥ 5 reads and ≥ 5% variant allele frequency (VAF) in at least one tumour region with germ-line VAF ≤ 1%. If a variant was found to meet these criteria in a single region, then the VAF threshold was reduced to ≥ 1% in order to detect low frequency variants. VarSLR was not applied to filter mutations.

All variants were annotated using ANNOVAR (Wang et al., 2010). Variants identified as non-silent were manually reviewed using Integrated Genomics Viewers (IGV) (Robinson et al., 2011).

2.3.4 Copy number analysis from multi-region NSCLC

Copy number processing was performed by Max Salm. For exome data, relative copy number was estimated using VarScan2 (v2.2.11) (Koboldt et al., 2012) with default parameters, excluding the sex chromosomes and low mapability regions (ENCODE 'DAC blacklisted' regions) and adjusting for GC-content. To identify genomic segments of constant copy number, logR values were quantile normalized, winsorized using the Median Absolute Deviation, and jointly segmented at the patient level (gamma = 1000). Integer copy numbers were calculated from relative copy numbers using ABSOLUTE (v1.0.6) (Carter et al., 2012). Minimum/maximum ploidy was set to within +/-0.5 of the prior ploidy estimate, calculated from the sample's FACS-based DNA-index. The top 5 ABSOLUTE models (ranked by log-likelihood) were retrieved for each exome, and a set of inter-sample models was identified that minimized the total pairwise distance derived from the segments' expected modal copy-number, whilst maximizing the model’s posterior log likelihood. Final model solutions were manually reviewed.

For multi-region WGS regions ASCAT was used for allele specific copy number estimation (Van Loo et al., 2010). Subclonal copy number analysis was performed using the Battenberg algorithm and was used as the input for the mutation clustering (Nik-Zainal et al., 2012b).

2.3.5 Copy number analysis from multi-region oesophageal adenocarcinoma

Processed sample exome SNP and copy number data from paired tumour-normal was generated using VarScan2 as described above. Output from VarScan was processed using the Sequenza R package 2.1.1 (Favero et al., 2015a) to provide segmented copy number data as well as purity and ploidy estimates. The following settings were used: breaks.method = 'full', gamma = 40, kmin = 5, gamma.pcf = 200, kmin.pcf = 200.

Manual verification was performed on the automatically selected models for ploidy and cellularity, and for 6 cases the model fitting was re-run using the second most optimum solution returned by Sequenza.

2.3.6 Copy number analysis from NSCLC exomes from Rizvi et al (2015)

Processed sample exome SNP and copy number data from paired tumour-normal was generated using VarScan2 as described above. Output from VarScan was processed using the ASCAT v2 (Van Loo et al., 2010) to provide segmented copy number data as well as purity and ploidy estimates The following setting was altered from its default value: Threshold for setting ACF to 1 was adjusted from 0.2 to 0.15 and the package was run with gamma setting of 1.

2.3.7 Estimating allele specific integer copy numbers in TCGA samples

Affymetrix SNP6 data from paired tumour-normal samples were normalized and pre-processed to obtain logR and B-allele fractions (BAF) using the Aroma Affymetrix CRMAv2 algorithm (in the allele-specific setting) (Bengtsson et al., 2009). The BAF was subsequently further adjusted using the CalMaTe and TumourBoost algorithms (Ortiz-Estevez et al., 2012, Bengtsson et al., 2010). Tumour copy number aberrations, ploidy and normal cell contamination was then determined using either ASCAT (Van Loo et al., 2010) or OncoSNP (Yau et al., 2010) with normal samples as references and hg19 coordinates. Sex chromosomes were excluded from all analysis. Samples that failed ASCAT processing due to poor model fit were discarded from all subsequent analysis.

Cell-lines were also analysed with Affymetrix SNP6.0 arrays. In this case, PICNIC (Greenman et al., 2010) was used for both normalization and integer copy number estimation.

Validation cohort for COAD analysis was performed on Illumina 610 Quad arrays. LogRs and BAFs were obtained using GenomeStudio V2011.1 and Genotyping Module V1.9.4. For QC, samples with moving standard deviation >0.28 were discarded. Integer copy numbers were estimated using OncoSNP (Yau et al., 2010).

Ploidy was estimated for each tumour sample by summing the weighted median integer copy for each chromosome and dividing by number of chromosomes analysed (n=22). The number of chromosomes in each sample was estimated by summing the modal copy numbers from the segmented copy number profile of each chromosome. Each segment was weighted according to the number of base pairs it covered. Copy number segments of loss and gain were defined relative to ploidy. wGII was calculated as in (Burrell et al., 2013a), by calculating the proportion of each chromosome that different from the genome median ploidy, and calculating the average across chromosomes.

2.3.8 Estimating the cancer cell fraction of TCGA mutations

As discussed in the Introduction, the variant allele frequency of a mutation is dependent on the copy number at the mutated site in the cancer cells, the copy number at the mutated site in the normal cells, the number of copies of the mutation itself, the purity of the cancer sample (i.e. the stromal contamination) and, finally, the fraction of cancer cells that harbour the mutation. If the variant allele frequency (obtained from VarScan2), the purity and local copy number (estimates from ASCAT) are known, the cancer cell fraction can be estimated. Specifically, the relationship between copy-number, purity and variant allele frequency can be described by the following formula:

Expected VAF (CCF) = $\mathbf{p}$*CCF / CPN_{norm} (1-$\mathbf{p}$) + $\mathbf{p}$*CPN_{mut}.

Where CPN_{mut} corresponds to the local copy number of the tumour, **p** is the tumour purity and CPN_{norm} is the local copy number of the matched normal sample. For a given mutation with 'a' alternative reads, and a depth of 'N', the probability of a given CCF can be estimated using a binomial distribution P(CCF) = binom(a|N, VAF(CCF)). CCF values can then be calculated over a uniform grid of 100 CCF values (0.01,1) and subsequently normalized to obtain a posterior distribution over the CCF. The 95% confidence intervals of the cancer cell fraction can be obtained from this distribution. Given that sex chromosomes were excluded from this analysis CPN_{norm} was assumed to be 2.

Similarly, the mutation copy number, n_{mut}, (the number of chromosomal alleles harbouring the mutation) can be calculated as follows:

n_{mut} = (VAF/$\mathbf{p}$)*(($\mathbf{p}$*CNt)+CNn*(1-$\mathbf{p}$))

To ensure the number of subclonal mutations was not overestimated, a conservative method for subclonal classification was utilized for TCGA samples. Any mutations was classified as clonal (present in all tumour cells sequenced) if the upper band of the 95% cancer cell fraction confidence interval was ≥1 and subclonal otherwise.

2.3.9 Estimating the cancer cell fraction of multi-region NSCLC

For multi-region NSCLC, the cancer cell fraction was estimated using a multi-dimensional Dirichlet process. This process was implemented by David Wedge.

In brief, subclonal clusters were identified using a Markov Chain Monte Carlo (MCMC) implementation of the Dirichlet process. Based on the MCMC assignment of mutations to clusters, the most likely configuration was obtained using a stepwise, greedy expectation-maximization (EM) algorithm. The best set of clusters was chosen using the Bayesian information criterion. Clusters containing less than 1% of mutations identified in a tumour were excluded from further analysis.

2.3.10 Estimating the cancer cell fraction of multi-region oesophageal adenocarcinoma and NSCLC from Rizvi et al. (2015)

For each mutation, reference read counts were first corrected for copy number and purity using a modified version of the methods described above. In brief, the mutation copy number, n_{mut}, was first calculated for each mutation using the formula described above. The expected copy number, n_{chr} was then determined using maximum likelihood. Notably, the expected copy number could take the form of a subclonal copy number. An estimate of the cancer cell fraction was obtained by dividing n_{mut} by $n_{chr.}$ This therefore yielded cancer cell fraction estimates corrected for clonal and subclonal copy number and purity

Next, cancer cell fraction estimates were grouped using the PyClone Dirichlet process clustering (Roth et al., 2014). Mutations were clustered based on modified total read depth as well as the observed alternative read depth. The modified read-depth was set such that the alternative read depth divided by the total read depth was equal to the cancer cell fraction calculated. Given that copy number and purity had already been corrected, integer copy numbers were set to 1 and purity to 1, allowing clustering to simply group clonal and subclonal mutations. PyClone was run with 10,000 iterations and a burn-in of 1000, and default parameters.

2.3.11 Identification of SNV heterogeneity driven by copy number alterations

For multi-region NSCLC mutation heterogeneity that could be driven by copy number heterogeneity was evaluated by visual inspection of copy number profiles. However, for oesophageal adenocarcinoma tumours a more systematic approach was devised.

For each tumour any SNV residing in a genomic segment of copy number heterogeneity across multiple tumour regions was identified, with minor and major copy number aberrations considered separately. For each chromosome, mutations were grouped into non-contiguous genomic segments with consistent copy number states within tumour regions and within SNV clusters defined above.

In order to restrict the analysis to mutations lost in at least one tumour region, the median CCF value of each SNV group was then determined, and only SNV groups where the median CCF value was <=0.25 in at least one tumour region were considered. Whether copy number loss coincided with lower CCF levels was assessed using a one-sided Wilcoxon test or, if more than two copy number states were present across tumour regions, a one-sided Cochrane-Armitage trend test. To ensure the lower CCF value was driven by copy number and not simply which tumour region harboured the mutation, a regression analysis was also implemented, including both copy number and region in the model. In total, across all oesophageal tumour regions, 100 mutations were filtered as being driven by copy number change.

2.3.12 Phylogenetic tree analysis of NSCLC multi-region tumours

The phylogenetic relationships between tumours subclones were inferred using Maximum Parsimony as implemented in the MEGA5 package (Tamura et al., 2011). For each tumour, every identified mutation (both non-silent and silent) was used as input. In cases where subclonal analysis of given tumour region revealed considerable subclonal diversity, this region was divided into a major and a minor subclone based on the mutation clusters described above. In addition, in cases where there was evidence for ubiquitous mutations becoming heterogeneous due to regional copy number losses, mutations were considered ubiquitous prior to tree reconstruction.

Maximum parsimony trees were inferred using the max-mini branch-and-bound algorithm calculating branch lengths with the average pathway method. The germline sample was designated as the outgroup, and, in the event of multiple optimal topologies, a consensus tree was constructed by collapsing branches reproduced in less than 50% of the trees.

2.3.13 Phylogenetic tree analysis of oesophageal adenocarcinoma multi-region tumours

Two types of phylogenetic trees were constructed for oesophageal adenocarcinoma tumours, region-trees and subclones trees. Region trees were constructed above using a Maximum Parsimony approach, treating each region as a separate species.

For subclonal cluster trees, the pigeonhole principle was used (Nik-Zainal et al., 2012a). That is, in each region if two subclonal clusters have a combined cancer cell fraction greater than 100% the smaller subclonal cluster must be nested within the larger. For each region, each possible subclonal structure was delineated, and this was compared across tumour regions to give a final subclonal relationship based on all tumour regions. Subclones that were incompatible with the phylogenetic relationship were removed. To illustrate this process, consider the following four clusters identified in three tumour regions. Cluster 1 is present in 100% of cells in all tumour regions. Cluster 2 is present in 90% of cells in region R1 and R2, and present in region 20% of cells in region R3. Conversely, cluster 3 is present in 80% of cells in region R3, but not present regions R1 and R2. Cluster 4 is present in 70% of cells in region R1 and R2 and 15% of cells in R3. Given that a tumour region cannot contain more than 100% of cells, and cluster 1 is present at 100% in all tumour regions, this cluster must represent the trunk node, with all other clusters contained within it. Cluster 2 and cluster 3 when combined never exceed 100% and so may represent separate branches of the phylogenetic tree. Cluster 4 must be contained within cluster 3 as their combined sum exceeds 100% in regions R1 and R2, yielding a complete phylogenetic relationship.

2.3.14 Classification of known driver mutations in multi-region NSCLC

In order to identity putative driver mutations, all genes harbouring non-silent mutations were compared with lists of potential driver genes in NSCLC, containing all genes identified as frequently mutated by large-scale lung cancer sequencing studies (Ding et al., 2008, Govindan et al., 2012, Imielinski et al., 2012, The Cancer Genome Atlas Research Network, 2012) or large-scale pan-cancer analyses (Lawrence et al., 2014, Davoli et al., 2013) using $q<0.05$ as cut-off, or present in the COSMIC cancer gene census (downloaded February 2013). For genes with a non-silent mutation identified in the tumours by multi-region sequencing that were present in one of these lists, whether the amino acid substitution had been previously identified in COSMIC was explored. In addition, when available, functional prediction scores (SIFT, Polyphen2, LRT and MutationTaster) were used and a mutation was scored as 'deleterious' when at least two out of the four predictors classified the mutation as deleterious. Finally, all non-silent mutations were grouped into four distinct categories. Category 1, 'high confidence driver mutations', contained all disrupting mutations (nonsense, frameshift, splicing or 'deleterious' missense) in tumour suppressor genes or activating amino acid substitutions in oncogenes as described in the literature. Category 2, 'putative driver mutations', on the other hand, represented all other amino acid substitutions located at the same position or up to 5 amino acids away from a substitution present in COSMIC. Category 3, 'low confidence driver mutations', contained all other non-silent mutations in genes that were present in the lists of cancer-related genes described above. Finally, category 4, 'unknown significance', contained the remaining non-silent mutations.

2.3.15 Identification and classification of driver mutations in oesophageal adenocarcinoma

Driver mutations in oesophageal adenocarcinoma were identified as above, with the substitution of lung specific genes with those identified in previous oesophageal sequencing studies (Dulak et al., 2013).

2.3.16 Classification of known driver mutations in TCGA tumours

Driver mutations were classified as non-silent mutations residing in genes representing the intersection of genes identified by the recent Lawrence analysis (Lawrence et al., 2014) and those residing in the cancer gene census (Futreal et al., 2004).

2.3.17 Identification of novel cancer genes using temporal analysis

To identify novel cancer genes, for each cancer type, MutSigCV was applied to all mutations as well as to ‘early’, ‘late’, clonal and subclonal mutations separately. A q-value threshold of 0.05 was adopted. For each run of MutSigCV only samples with at least 5 mutations within a given category were used and only genes with at least 10 non-silent mutations were considered. MutSigCV was implemented using default settings, using hg19 coordinates and the covariates and coverage tables provided (assuming full exome coverage).

2.3.18 Intra-tumour heterogeneity index in NSCLC and oesophageal adenocarcinoma

An intra-tumour heterogeneity index was calculated for each tumour. The index was calculated by firstly calculating for each tumour the proportion of heterogeneous mutations, relative to total number of mutations, for each possible pairwise comparison of tumour regions. The index was then determined by assessing the mean value across each pairwise comparison.

2.3.19 Permutation testing to assess clonal enrichment

In order to assess whether a specific gene had an enrichment or depletion of clonal/subclonal mutations a permutation test was devised. Consider a cancer gene with 50 non-silent mutations across 500 separate samples. Of these 50 mutations, 35 are clonal and 15 subclonal. To assess whether this gene exhibits an enrichment or depletion in clonal mutation 50 non-silent mutations are randomly sampled from 500 samples 10,000 times to obtaining a background distribution representing the expected proportion of clonal/subclonal mutations. A p-value can then be obtained by comparing the observed proportion of clonal/subclonal mutations with this background distribution.

2.3.20 Temporal dissection of mutations of multi-region tumours

For each multi-region sequenced tumours, each mutation was classified as 'early' or 'late' based on whether it was located on the trunk or branch of the phylogenetic tree, with all truncal mutations classified as 'early' and any branch mutation as 'late'. Mutation copy number estimates calculated as described above were used to time mutations relative to genome doubling. Only regions with at least two copies of the major allele were used, whilst regions where the minor allele was equal to one were excluded. Any mutations that had an estimated mutation copy number >1 was considered to have occurred before doubling, and all copy number 1 mutations as after doubling. Chi-square tests were used to compare the mutation spectra of the six mutations types (C>A, C>G, C>T, T>A, T>C, T>G). To compare the relative frequency of specific mutation types a two-sided Fisher's exact test was used. APOBEC enrichment was assessed as described below.

2.3.21 Temporal dissection of mutations in TCGA samples

For each single-region TCGA sample, it was not possible in these to construct phylogenetic trees as described above. Therefore mutations were classified as early or late based on their clonal status and when possible mutations were timed relative to copy number events.

In brief, for timing mutations relative to copy number events, analysis was restricted to mutations occurring in regions with at least two copies of the major allele. For any such region, mutations at mutation copy number >1 were classified as 'before event' and any mutations with a mutation copy number of 1 were classified as 'after event'. Combining this with cancer cell fraction estimates (see above), all clonal mutations that were not classified as 'after event' were aggregated as 'early', whilst all subclonal or 'after event' mutations were aggregated as 'late'.

2.3.22 Estimating smoking strand bias

Strand bias was calculated based on annotating each C>A mutation as to whether it fell on the transcribed or untranscribed strand in the UCSC hg19 gene track (available from http://genome.ucsc.edu/cgi-bin/hgTables). TCGA Samples, grouped according to histology and smoking status, were considered in aggregate. Two-sided Fisher's exact tests were used to determine significance.

2.3.23 Detecting an APOBEC mutation pattern

To detect an APOBEC mutation pattern the methods outlined by Roberts et al (2013) were utilized. In brief, the enrichment E_{TCW} relating to the strength of mutagenesis at the TCW motif across the genome was calculated as follows:

$$E_{TCT} = \frac{mutations_{TCW} \times context_{CorG}}{mutations_{CorG} \times context_{TCW}}$$

where $\text{mutations}_{\text{TCW}}$ is the number of mutated cytosines (and guanines) falling in a TCW (or WGA) motif, $\text{mutations}_{\text{C (or G)}}$ is the total number of mutated cytosines (or guanines), $\text{context}_{\text{TCW}}$ is the total number of TCW (or WGA) motifs within a 41-nucleotides region centred on the mutated cytosines (and guanines) and $\text{context}_{\text{C (or G)}}$ is the total number of cytosines (or guanines) within the 41-nucleotides region centred on the mutated cytosines (or guanines). Only specific base substitutions were included (TCW to TTW or TGW, WGA to WAA or WCA, C to T or G, and G to A or C).

Over-representation of APOBEC signature mutations in each sample was determined using a two-sided Fisher's exact test comparing the ratio of the number of cytosine-to-thymine or cytosine-to-guanine substitutions and guanine-to-adenine or guanine-to-cytosine substitutions that occurred in and out of the APOBEC target motif (TCW or WGA) to an analogous ratio for all cytosines and guanines that reside inside and outside of the TCW or WGA motif within 41-nucleotide region centred on the mutation cytosine (and guanine). P-values were corrected using

Benjamin-Hochberg multiple testing correction, and a significance threshold of q<0.05. For each sample, APOBEC mutation enrichment was determined for all mutations, 'early' mutations and 'late' mutations separately.

2.3.24 Detecting a platinum mutation pattern

As discussed in the Introduction, previous work (Meier et al., 2014) has demonstrated the propensity for platinum exposure to lead to C>A transversion mutations within a CpC context. To detect a platinum mutation pattern the methods outlined for APOPEC enrichment, described above, were adapted. The enrichment E_{CpC} relating to the strength of mutagenesis at the CpC motif across the genome was calculated as follows:

$$E = (\text{mutations}_{CpC} * \text{context}_{C}) / (\text{mutations}_{C} * \text{context}_{CpC})$$

where mutations_{CpC} is the number of mutated cytosines (and guanines) falling in a CpC (or GpG) dinucleotide, mutations_{C} (or G) is the total number of mutated cytosines (or guanines), context_{CpC} is the total number of CpC (or GpG) dinucleotides within a 41-base region centred on the mutated cytosines (and guanines) and context_{C} (or G) is the total number of cytosines (or guanines) within the 41 base region centred on the mutated cytosines (or guanines). A two-sided Fisher's exact test was used to determine if an overrepresentation of platinum signature mutations was present in each sample. The test compared the ratio of the number of cytosine-to-adenine and guanine-to-thymine substitutions that occurred in and out of the CpC (GpG) platinum target dinucleotide to an analogous ratio for all cytosines and guanines that reside inside and outside of the CpC or GpG dinucleotide within 41 base region centred on the mutation cytosine (and guanine), representing the genomic background. P-values were adjusted using Benjamin-Hochberg multiple test correction. A two-tailed Fisher's exact test was performed to compare for platinum enrichment between the pre- and post-chemotherapy samples.

2.3.25 Mutational signature analysis in TCGA tumours

For each cancer type, mutational signature analysis was implemented separately on early and late mutations, as well as on all mutations. To ensure sufficient statistical power, mutations classified as ‘early’ or ‘late’ rather than simply clonal and subclonal were used.

Hierarchical clustering was used to ensure the same signatures were identified in early, late mutations and all mutations. Each signature identified was then hierarchically clustered together with the signatures identified in Alexandrov et al. (2013a) and visually inspected to ensure a good fit. Tumours were classified as harbouring a mutational signature if at least 25% of mutations or over 100 mutations were found to belong to a given signature (as in (Alexandrov et al., 2013a)).

Finally, to compare the prevalence of mutational signatures in early and late mutations a paired Wilcoxon test was used, comparing, for each tumour, the proportion of mutations that corresponded to a given signature in their early and late mutations. Only samples that harboured at least 10 early and 10 late mutations were considered. A small number of samples with outlier mutation signatures were identified and discarded so as to avoid biasing mutational signature interpretation (e.g. one HNSC sample with a prevalent UV-signature).

2.3.26 Mutational signature analysis in TCGA oesophageal tumours

The deconstructSigs package for R, which determines the contributions from a given set of mutational signatures to an individual tumour sample, was used to explore mutational signatures in TCGA oesophageal tumours. As above, for each tumour sample, the package was applied separately to all, early and late mutations. Comparisons between signatures were also performed as described above.

2.3.27 Genome doubling algorithm

A modified version of a published algorithm (Carter et al., 2012) was used to estimate the prevalence of genome doubling. Each sample, *s*, was represented as an aberration profile of major and minor allele copy numbers at chromosome arm resolution. From this profile, *Ns*, the total number of aberrations (relative to diploid) and *Ps*, the probabilities of loss/gain for each allele at each chromosome arm were calculated. 10,000 simulations were then run for each sample *s*. In each simulation, *Ns* sequential aberrations, based on *Ps*, were applied to a diploid profile. A p-value for 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. Different p-value thresholds were used depending on the ploidy of the tumour. For samples with ploidy ≤3, a p-value threshold of 0.001 was used. A p-value ≤ 0.05 was used if ploidy=4, and all samples where ploidy≥5 were classified as genome-doubled.

2.3.28 Estimating timing of genome Doubling

Each genome-doubled sample was represented as an array of genotype proportions reflecting copy numbers ranging from zero to eight. Sixteen possible genotypes were discriminated: zero copies, A|B (1 copy); AA|BB and AB (2 copies); AAA|BBB and AAB|ABB (3 copies)…. (where A and B represent the two parental alleles).

To explore the timing of genome doubling, only losses to two copies (AA, BB, AB) were used as these can either reflect losses before (AA or BB) or after genome doubling (AB). Samples with a higher proportion AB compared to AA/BB were classified as having genome doubled before the majority of losses, whereas those where AA/BB > AB were classified as having genome doubled after the majority of losses

2.3.29 Significance of correlation between wGII and copy number loss

Based on the observed probability for loss, given by the percentage of genome that was identified as lost relative to the ploidy in that sample, an aberration state (loss or no loss) was calculated for each sample separately. A point-biserial correlation between aberration state and wGII was then calculated across samples. This process was repeated 10,000 times and a p-value was obtained for each gene by counting the percentage of simulations showing a greater correlation coefficient than that observed for that gene.

2.3.30 Shannon diversity index in genome-doubled and non-genome doubled colonies

Based on data from clonal FISH data from chromosome 2 and 8, the Shannon diversity index was calculated using the following formula:

$$\mathrm{SD} = -\sum_{i=1}^{R} p_i \ln p_i$$

where p_i is the proportion of colonies with a mode i. The Shannon diversity index quantifies the uncertainty in predicting the chromosome complement from a colony that is taken at random from the population. For example, if the colony modes of chromosome 2 and 8 are always diploid the SD will be 0. Conversely, if the colony modes differ greatly, the SD will be a larger number.

2.3.31 HLA typing of patient samples

The 4-digit HLA type of each TCGA patient was determined using POLYSOLVER (POLYmorphic loci reSOLVER)(Shukla et al., In press). Patients L011 and L012 were serotyped and simultaneously genotyped using Optitype (Szolek et al., 2014), which produced concordant results.

2.3.32 Identification of putative neo-antigens

Identified non-silent mutations were used to generate a comprehensive list of peptides 9-11 amino acids in length with the mutated amino acid represented in each possible position. The binding affinity of every mutant peptide and its corresponding wild-type peptide to the patient's germline HLA alleles was predicted using netMHCpan-2.8 (Hoof et al., Nielsen et al., 2007). Candidate neo-antigens were identified as those with a predicted binding strength of < 500 nM.

2.3.33 Survival analyses

Survival curves were plotted according to the Kaplan Meier method. Log-rank test statistics were used to assess significance for univariate analysis. Cox proportional hazards regression models were conducted for multivariate survival analysis (R package *survival*).

2.4 Experimental work

While I carried out the bioinformatics work described in this thesis, unless otherwise indicated, Elza de Bruin, Sally Dewhurst and Andrew Furness carried out the experimental work.

2.4.1 Cell culture

Cell culture was performed by Sally Dewhurst. HCT-116 was obtained from the Cancer Research UK cell services facility. STR DNA fingerprinting for HCT-116 was carried out in October 2010. The HCT-116_MLH1.3 was STR fingerprinted in February 2010 and was a gift of Françoise Praz, with similar results to HCT-116 except at vWA. Cell lines were maintained at 37°C in 5% CO_2 in Dulbeco's Modified Eagle medium (DMEM) media (1X), High Glucose with L-Glutamine (Gibco, Invitrogen), with 10% foetal bovine serum (FBS) and 1xPenStrep (Sigma). During long-term culture, clones were passaged approximately once a week, and clones were always split at the same dilution. All passage numbers described within the thesis are within <4 passages of the numbers represented in figures.

CellTiter-Blue® (Promega) assays were performed following the manufacturers instructions. Cloning efficiency was estimated using the Poisson distribution as follows: Efficiency = (-100) * ln(#of wells with no colony / total number of wells)%. Under the assumptions that the expected proportion of colonies accurately approximates the true efficiency and that the observed proportion of failures approximates the probability of zero growth in a single well (as described in (Leight and Sugden, 2001)).

2.4.2 Fluorescence Activated Cell Sorting (FACS)

FACS analysis was carried out by Sally Dewhurst. Cells were stained with 10μg/ml Hoescht 33342 (Sigma) for 1hr at 37°C. A MofLo (Beckman Coulter) cell sorter was used, with a 100mW JDS Uniphase XCyte pulsed UV laser, emitting at 355nm. Single cells were sorted into 96 well plates containing 20% FBS media. Genome doubled clones were expanded before assessment of DNA content by flow cytometry using either Propidium Iodide (Sigma) with RNase (Life Technologies), or Hoescht 33342, after fixation in 70% Ethanol.

2.4.3 Clonal FISH, metaphase spreads & immunofluorescence of genome doubled and non genome doubled clones

FISH was performed probing for chromosomes 2 (CEP2 D271, SO), and 8 (CEP8, D872, SGn, both Abbott Molecular probes) and was carried out as previously described (Burrell et al., 2013a). Slides were either scored semi-automatically using the Ariol system (Leica Microsystems). After a low magnitude scan (5x) discrete colonies were selected and scanned at a higher magnification (40x). Nine stacks of 0.7μm were taken through each colony. Analysis was performed utilizing the automated SPOT assay before manual curation. An Olympus DeltaVision RT microscope (Applied Precision, LLC) equipped with a Coolsnap HQ camera with an Olympus ×100, ×60 or x40 1.4 numerical apertures UPlanSApo oil immersion objective was used to score slides. As previously described (Burrell et al., 2013a), metaphase spreads were prepared and probed with an all-human centromere probe (Posiedon)]. Finally, immunofluorescence and segregation error classification were performed as previously described (Burrell et al., 2013a).

2.4.4 H2B-mRFP transfection and live-cell imaging analysis

H2B-mRFP cells: cells were transfected with pH2B-mRFP (gift from A. Straube) using Fugene 6.0 (Promega), and selected in 1 mg ml^{-1} G418 (Life technologies) before flow-sorting for mRFP expression. Cells were maintained in 500µg ml^{-1} G418. Cells were imaged in an 8 well imaging chamber (LabTek). Fourteen-micrometre *z*-stacks were recorded using an Olympus ×40 1.3 numerical aperture UPlanSApo oil immersion objective using a DeltaVision microscope in a 5% CO_2 atmosphere at 37°C. Stills were captured every 3mins for 6hrs and every 15mins thereafter, for approx. 60 hours. Analysis was carried out with Softworx Explorer (Applied Precision, LLC). Daughter cells were scored as arrested if they failed to undergo division within 48 hours of the first division. Multipolar divisions were excluded from analysis.

2.4.5 APOBEC3B expression analysis

Total RNA was isolated from tumour regions of which fresh frozen material was available (L001, L002, L003, L004 and L011), using the AllPrep DNA/RNA kit from Qiagen according to the manufacturer's instruction, and used to synthesize cDNA. The cDNA was then amplified using *APOBEC3B* Taqman Assay (Applied Biosystems) on a 7500 FAST Real Time PCR machine (Applied Biosystems). Taqman Assays for the housekeeping gene *TBP*. *APOBEC3B* expression was normalised towards *TBP*, and the fold-change in expression was determined against the expression in the adjacent normal lung.

2.4.6 MHC multimer generation and combinatorial encoding-flow cytometry analysis

MHC-multimers holding the predicted neo-antigen were produced in-house at Technical University of Denmark, in the laboratory of Sine R. Hadrup. HLA molecules matching the HLA-expression of L011 (HLA-A1101, A2402, and B3501) and L012 (HLA-A1101, A2402, and B0702) were refolded with a UV-sensitive peptide, and exchanged to peptides of interest following UV exposure (Toebes et al., Bakker et al., Frosig et al., Chang et al.). Briefly, HLA complexes loaded with UV-sensitive peptide were subjected to 366-nm UV light (CAMAG) for one hour at 4°C in the presence of candidate neo-antigen peptide in a 384-well plate. Peptide-MHC multimers were generated using a total of 9 different fluorescent streptavidin (SA) conjugates: PE, APC, PE-Cy7, PE-CF594, Brilliant Violet (BV)421, BV510, BV605, BV650, Brilliant Ultraviolet (BUV)395 (BioLegend). MHC-multimers were generated with two different streptavidin-conjugates for each peptide-specificity to allow a combinatorial encoding of each antigen responsive T cells, enabling analyses for reactivity against up to 36 different peptides in parallel (Hadrup et al., Andersen et al.).

2.4.7 Identification of neo-antigen-reactive CD8$^+$ T cells

MHC-multimer analysis was performed on in-vitro expanded CD8$^+$ T lymphocytes isolated from lung tumour regions and adjacent normal lung tissue. 290 and 355 putative epitopes (with predicted HLA binding affinity <500nM, including multiple potential peptide variations from the same missense mutation) were synthesized and used to screen expanded L011 and L012 tumour infiltrating lymphocytes (TILs) respectively. For staining of expanded CD8$^+$ T lymphocytes, samples were thawed, treated with DNAse for 10 min, washed and stained with MHC multimer panels for 15 min at 37°C. Subsequently, cells were stained with LIVE/DEAD® Fixable Near-IR Dead Cell Stain Kit for 633 or 635 nm excitation (Invitrogen, Life Technologies), CD8-PerCP (Invitrogen, Life Technologies) and FITC coupled antibodies to a panel of CD4, CD14, CD16, CD19 (all from BD Pharmingen) and CD40 (AbD Serotec) for an additional 20 min at 4°C. Data acquisition was performed on an LSR II flow cytometer (Becton Dickinson) with FACSDiva 6 software. Cut-off values for the definition of positive responses were ≥0.005% of total CD8+ cells and ≥10 event

Chapter 3. Evidence and extent of intra-tumour heterogeneity

3.1 Introduction

Accumulating evidence suggests intra-tumour heterogeneity may be widespread across human cancers. The extent of this diversity within tumours has important clinical implications. For instance, in colorectal cancer, subclonal mutations in the oncogene *RAS* have been shown to precipitate resistance to cetuximab (Misale et al., 2012), while in NSCLC, subclonal *EGFR* T790M mutations are associated with resistance to *EGFR*-TKI (tyrosine kinase therapy) (Kobayashi et al., 2005). Further complicating the issue, the use of targeted therapy against a subclonal driver mutation, present in a subset of cancer cells within a tumour`, may lead to stimulation of wild-type subclones which lack the actionable alteration (Lohr et al., 2014).

Targeting clonally dominant somatic events, present in all tumour cells, or adopting combinatorial targeted therapy approaches, may therefore be necessary for optimal tumour control (Yap et al., 2012a). However, although the clonal status of driver mutations has received attention in certain cancers (Papaemmanuil et al., 2013, Bolli et al., 2014, Lohr et al., 2014, Shah et al., 2012, Nik-Zainal et al., 2012b, Landau et al., 2013, Gerlinger et al., 2012, Gerlinger et al., 2014a), a broad understanding of the heterogeneity of mutations in cancer genes, and deciphering their clonal and subclonal frequencies, is lacking.

More generally, an understanding of the patterns of cancer evolution may begin to shed light on whether rules dictating progression of cancers can be discerned. Such rules may inform clinical practice, and reveal novel avenues for drug discovery and clinical trial design. Relatedly, there is a need to understand the impact of therapy itself on tumour evolution and the extent of diversity within tumours.

In this chapter I explore the degree of intra-tumour heterogeneity across ten major cancer types. I make use of next-generation sequencing data from both multi-region sampling and single tumour sample data to quantify intra-tumour heterogeneity and tumour evolution. I investigate the extent to which both driver and passenger mutations are clonal or subclonal across cancers, focussing on heterogeneity at the single-nucleotide level. In addition, heterogeneity within tumours is used to time the acquisition of driver events in cancer evolution, revealing which events may be crucial for tumour transformation, whilst also shedding light on others which may play important roles in tumour progression and metastasis. Finally, I investigate the impact of neo-adjuvant platinum chemotherapy on tumour evolution and intra-tumour heterogeneity in oesophageal adenocarcinoma.

The data presented in this chapter of the thesis largely forms sections of three separate publications (McGranahan et al., 2015, de Bruin et al., 2014, Murugaesu et al., 2015), which can be found in Appendix 2.

3.2 Results

3.2.1 Multi-region sequencing data to explore heterogeneity in NSCLC

To determine the degree of heterogeneity in NSCLC, 27 regions from tumours derived from six patients were subject to whole-exome or whole genome sequencing, as well as matched germ-line blood or normal tissue. On average, 4 tumour regions were sequenced per patient (range 3-6), to mean coverage depths of 107X for exome and 96X for whole genome (Table 8-1). Likewise, matched germ-line blood was sequenced to mean coverage depth of 98X for exome and 38X for whole genome.

Table 3-1 Clinical characteristics of multi-region NSCLC

Patient ID	Age (years)	Gender	Histology	Lymph node(s)/ location	Stage (I-IV)	Regions sequenced	Smoking status (pack-years[*])
L003	84	F	LUAD	2/Station 4	IIIB	R2 (RLL), R4 (RUL), LN	never-smoker
L008	75	M	LUAD	2/Hilar	IIIA	R1 (RUL), R3 (RML), LN	ex-smoker (25)
L001[†]	59	F	LUAD	3/Hilar	IIA	R1-R5, LN	ex-smoker (10)
L004[‡]	73	M	Undiff. NSCLC	none	IIB	R1-R4	current smoker (50)
L011	49	F	LUAD	none	IB	R1-R3	current smoker (45)
L002	78	M	LUAD/ LUSC	2/Station 5	IIIA	R1-R4	current smoker (>50)

Abbreviations: LLL, lower lobe; LUL, left upper lobe; RLL, right lower lobe; RUL, right upper lobe; RML, right middle lobe; R, region; LN, lymph node; Undiff, undifferentiated. * a pack-year is defined as the number of packs of cigarettes smoked per day multiplied by the number of years the person has smoked. [†]L001 presented a synchronous MEN1 syndrome-associated tumour, classified as separate tumour based on histological morphology, biochemical profile and octreotide scan imaging. [‡]L004 presented a synchronous oesophageal adenocarcinoma, classified as separate primary tumours based on histological morphology and immunohistochemistry marker profile.

Patients represented a range of ages (49-84) and included a never-smoker (L003), as well as two former smokers (L008 and L001) and three current smokers (L004, L011, and L002) (Table 3-1). Four tumours were classified as lung adenocarcinoma (LUAD), one tumour classified as an undifferentiated NSCLC, and a final tumour as a mix of both LUAD and lung squamous cell carcinoma (LUSC) (Figure 3:1).

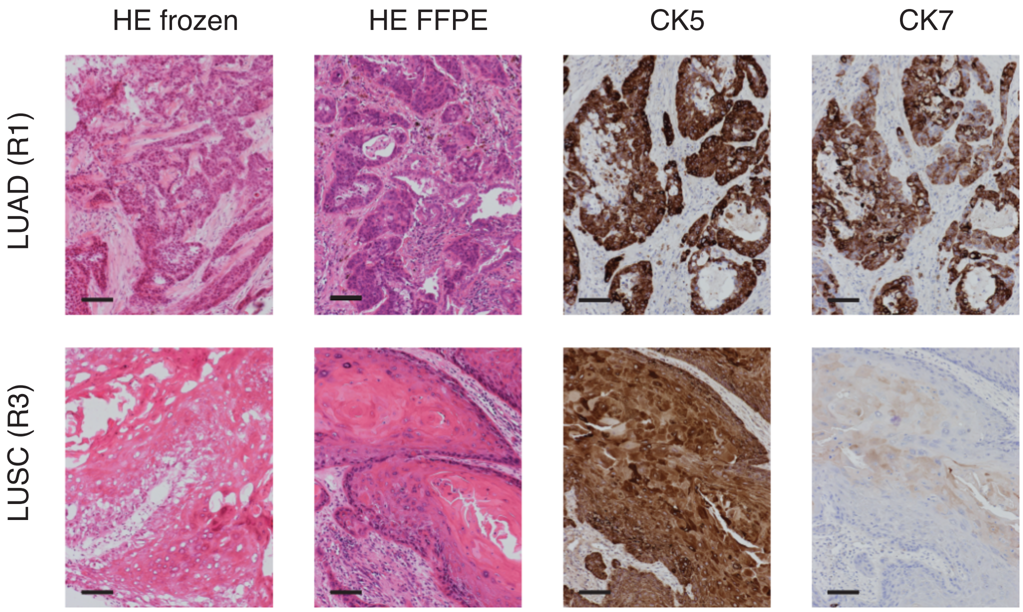


Figure 3:1 L002 Histology

L002 histopathology staining. Sections from LUAD (upper) or LUSC (lower) regions are shown. Left section is taken from frozen material adjacent to material used for DNA extraction for sequencing. The remaining sections are taken from FFPE representing LUAD or LUSC tumour regions and stained with hematoxylin and eosin or with an antibody detecting CK5 or CK7, as indicated.

Mutation calling was performed on individual tumour regions using VarScan2, as outlined in the Experimental procedures section. To ensure heterogeneity was not overestimated, mutation calling leveraged data from all sequenced regions within a tumour. Thus, if a mutation was identified in any tumour region its presence was assessed in all other tumour regions. In total, 44,882 mutations were identified across the 6 tumours, of which 1,758 were non-silent. An average of 293 (range 117- 476) non-silent mutations were identified per tumour, and 229 (range 42-403) per tumour region and, on average, 20% (range 3-87%) of each region's mutations were only identified as a result of leveraging multi-region sequencing analysis (Figure 3:2). Use of leveraged mutation calling was most prominent in tumour region R4 in L002 - while only 31 non-silent mutations passed calling thresholds within that specific tumour region, multi-region sequencing was used to confirm the likely presence of an additional 212 non-silent mutations. These data highlight the benefits of utilizing a multi-region approach to more fully capture the catalogue of mutations within a tumour.

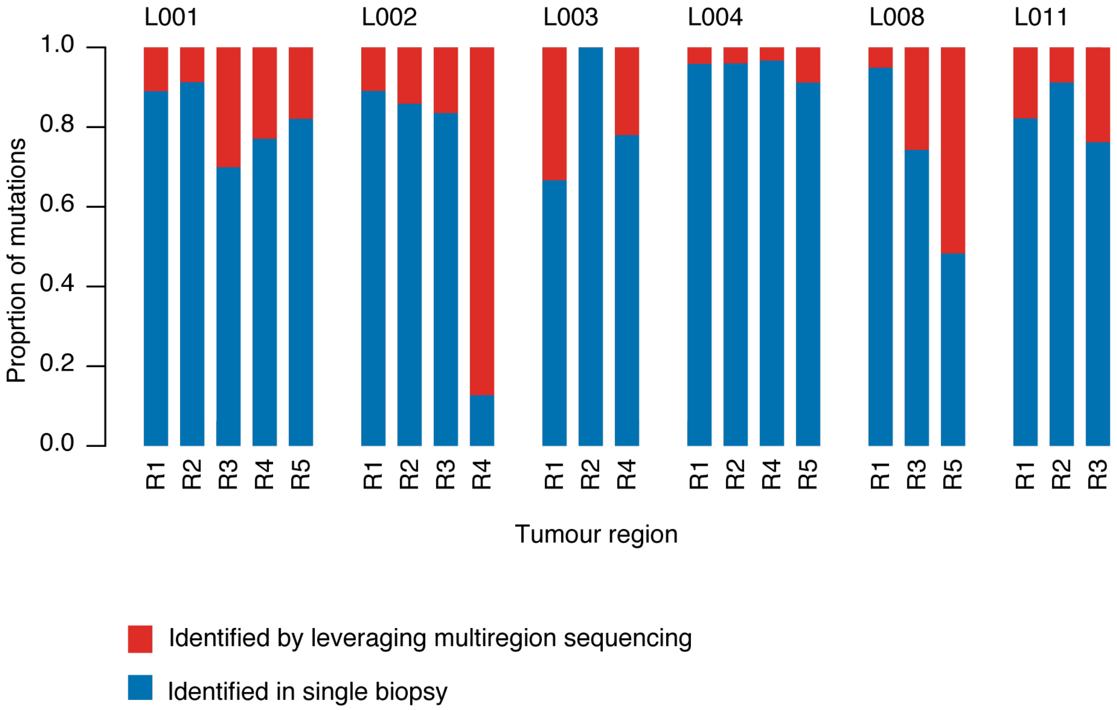


Figure 3:2 Mutations leveraged by multi-region sequencing
Mutations identified as present within tumour regions are split into the proportion identified by leveraging multi-region sequencing (red) and the proportion identified regardless of multi-region sequencing.

3.2.2 NSCLC exhibits extensive heterogeneity at SNV level

To obtain a measure of heterogeneity, independent of the number of tumour regions sequenced, a heterogeneity score was devised, normalizing for the number of regions sequenced. In brief, for a given tumour, each tumour region was compared to each and every other tumour region and the mean proportion of heterogeneous non-silent SNVs between pairs of tumour regions was used to give an indication of the extent of heterogeneity at the non-silent mutation level.

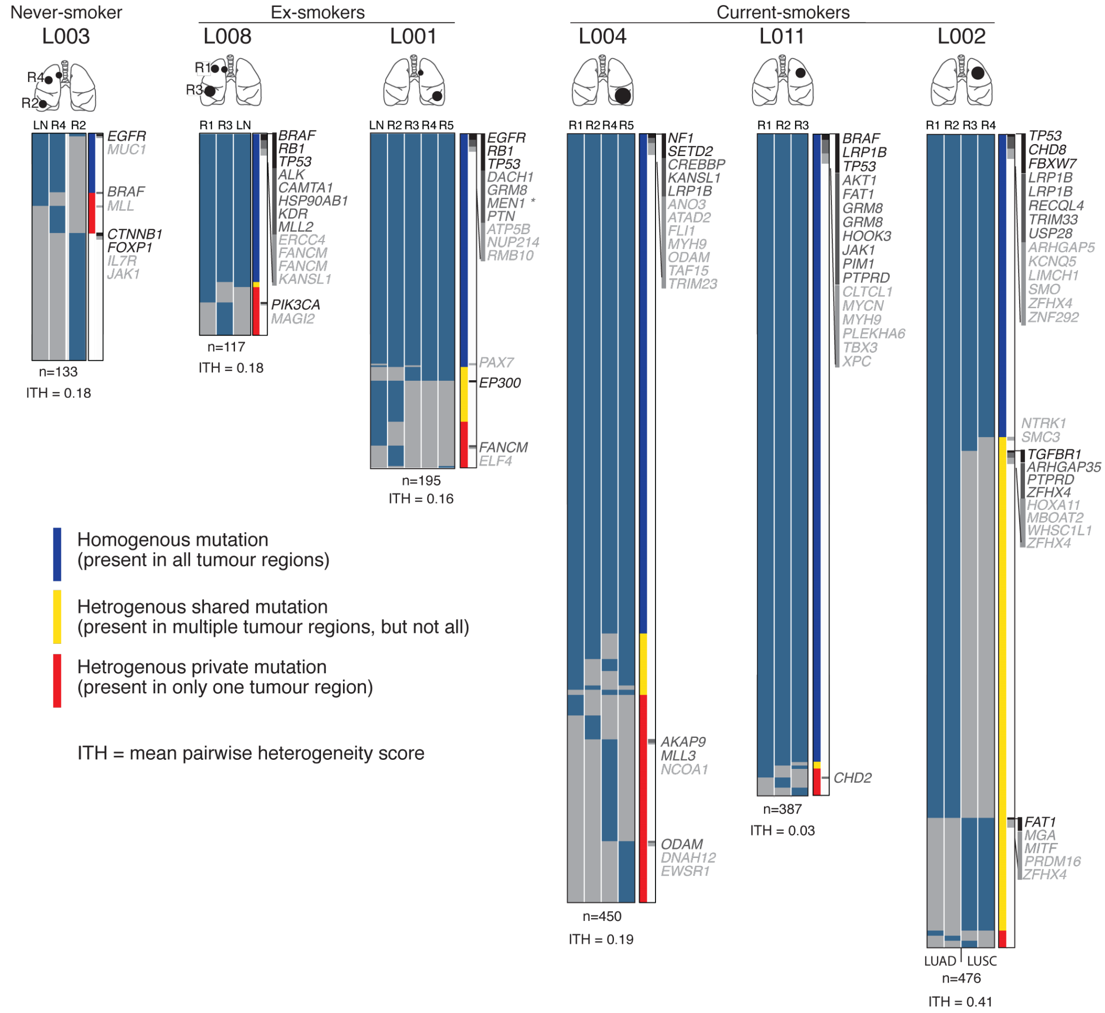


Figure 3:3 NSCLC Multi-region heatmap

Heatmap displaying non-silent mutations identified in NSCLC multi-region sequencing data. Tumours are grouped according to smoking status. The extent of heterogeneity and number of mutations is indicated below each heatmap. Category 1-3 driver mutations are shown from black to light grey.

Intra-tumour heterogeneity at the SNV level was evident in all six NSCLC tumours (Figure 3:3), with a median of 30% of non-silent mutations not present in all regions sampled within an individual tumour (range 4-63%), and a median heterogeneity score of 18% (range 3-41%).

The extent of the heterogeneity varied considerably between tumours derived from different patients. The adeno-squamous cell carcinoma tumour, L002, displayed extensive intra-tumour heterogeneity, with 63% of non-silent mutations identified as heterogeneous and not present in all tumour regions, and a heterogeneity score of

41%. Although heterogeneous mutations showed regional separation, concordant with LUAD (regions R1 and R2) or LUSC (regions R3 and R4) histopathologies, a shared clonal origin was evident, with 37% of the non-silent mutations identified as present in all tumour regions sequenced (Figure 3:3). By contrast, L011 exhibited minimal heterogeneity at the SNV level, with only 4% of mutations identified as heterogeneous and a heterogeneity score of 3%.

Both patient L003 and patient L008 presented with two tumours in separate lobes of the lung (Figure 3:3). Multi-region sequencing was therefore used to decipher the clonal origins of these tumours. In L008, 74% of non-silent mutations were ubiquitously identified in both tumour regions derived from the middle lobe and the upper lobe of the lung, suggesting a shared clonal origin between the two tumours. Thus, for L008 one of the tumours likely represents a metastasis of the other. Conversely, in L003 only a single mutation (EGFR-L858R) was identified as shared between the tumours from the upper and lower lobes of the lung. Given that EGFR-L858R is a highly recurrent mutation, occurring in approximately 43% of *EGFR* mutated lung tumours (Mitsudomi and Yatabe, 2010), and no other mutations (silent or non-silent) were identified as shared, it is likely that the tumours were of independent clonal origin, converging upon identical *EGFR* activating mutations.

3.2.3 Single sample analysis to explore heterogeneity across 9 major cancer types

To further explore the level of heterogeneity at the SNV level across a range of cancer types, data from the TCGA was used. In total 2,694 tumour samples, representing 9 major cancer types, bladder (BLCA, n=130), breast (BRCA, n=903), colon adenocarcinoma (COAD, n=190), head-and neck squamous cell carcinoma (HNSC, n=250), glioblastoma mulitforme (GBM, n=289), kidney clear cell renal cell carcinoma (KIRC, n=303), lung adenocarcinoma (LUAD, n=257), lung squamous cell carcinoma (LUSC, n=131), and skin cutaneous melanoma (SKCM, n=241), were investigated.

For TCGA tumours, mutations were called by individual sequencing centres and further filtering was applied to ensure consistency within and across tumour types (see Experimental procedures). A total of 516,672 somatic mutations were used for downstream analysis, consisting of 326,918 missense, 145,009 silent, 24,236 nonsense, 9867 RNA [5' un-translated region (UTR) or 3'UTR], 9828 splice-site, 341 nonstop and 473 translational start-site mutations.

Sequencing coverage at mutated sites varied considerably both within and between tumour types, with a median coverage of 94.5X across the pan-cancer cohort (range 73-125). Notably, the read coverage across the genome will directly impact upon the ability to detect low frequency mutations. It is estimated that at a sequencing depth of 70, the probability of detecting mutations at a variant allele frequency (VAF) of 0.1 is 0.98; however, the probability is reduced to 0.69 when attempting to detect a variant with a VAF of 0.05 (Cibulskis et al., 2013). Thus, in this analysis, the number of low-frequency subclonal mutations may have been underestimated.

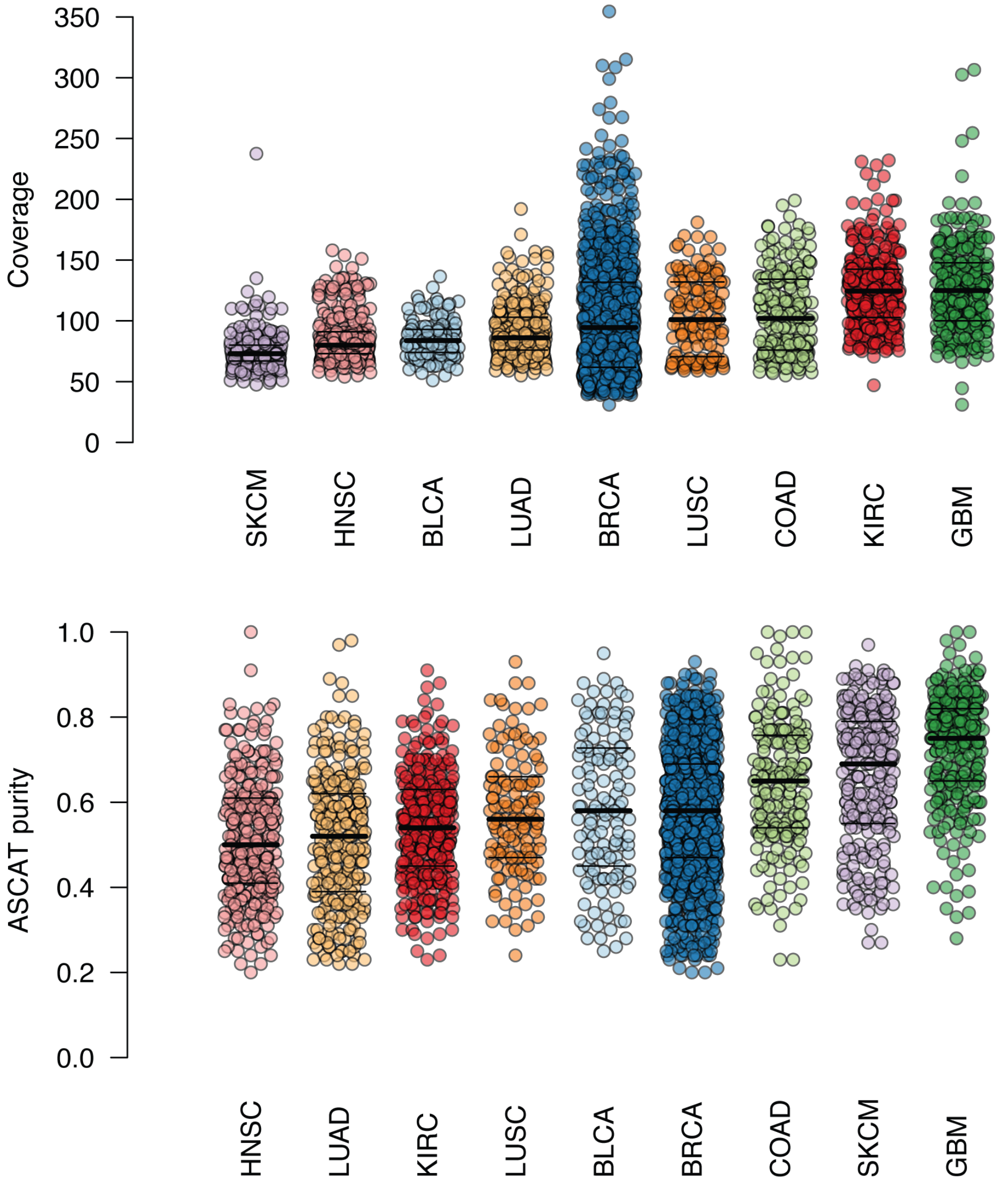


Figure 3:4 Purity and coverage in TCGA tumours

A) Tumour coverage at mutated bases across different tumours. For each cancer type, each dot represents one TCGA tumour. B) ASCAT purity estimates across different tumours.

Relatedly, the purity of the sample will also directly impact upon the ability to identify mutations, with low purity samples necessarily exhibiting mutations with lower VAF than their higher purity counterparts when matched for copy number. Encouragingly, the vast majority of TCGA samples (>90%) exhibited purity above 35%, as estimated by ASCAT (Figure 3:4 B).

3.2.4 Deciphering heterogeneity and timing of SNVs within single tumour samples

Next, the clonal status and timing of mutations was estimated by integrating variant allele frequencies with copy number and purity estimates (see Experimental procedures). The cancer cell fraction, describing the proportion of cancer cells harbouring a mutation, was calculated and mutations were classified as clonal if the 95% confidence interval of the cancer cell fraction overlapped with 1. The subclonal fraction, describing the proportion of mutations classified as subclonal within a tumour, varied considerably within and across tumour types (Figure 3:5).

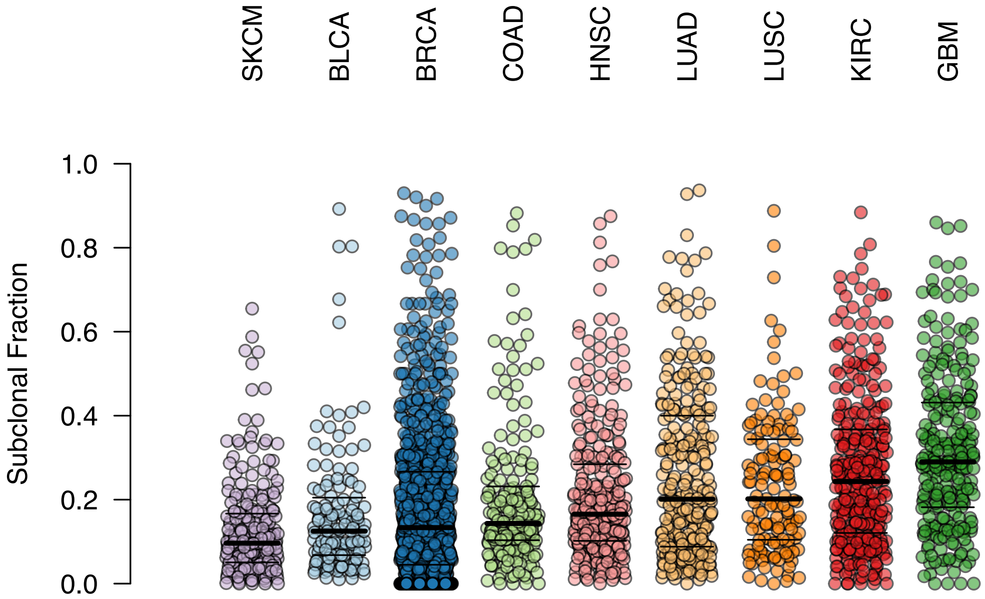


Figure 3:5 Subclonal fractions across and within tumour types
For each tumour type, each dot represents one tumour, and the proportion of the entire mutational burden that is classified as subclonal is depicted. Notably, the subclonal fraction varies considerably both between and within tumour types.

To determine whether the observed differences in subclonal fraction was primarily driven by differences in tumour purity, ASCAT purity was directly compared to the subclonal fraction in all tumours. Reassuringly, the subclonal fraction was not significantly different between high purity (purity>0.8) and medium purity (purity 0.6-0.8) samples. Nevertheless, a lower subclonal fraction was observed for tumours with a purity lower than 40%, confirming that low purity levels likely hamper the ability to decipher clonal architecture, leading to underestimates of intra-tumour heterogeneity.

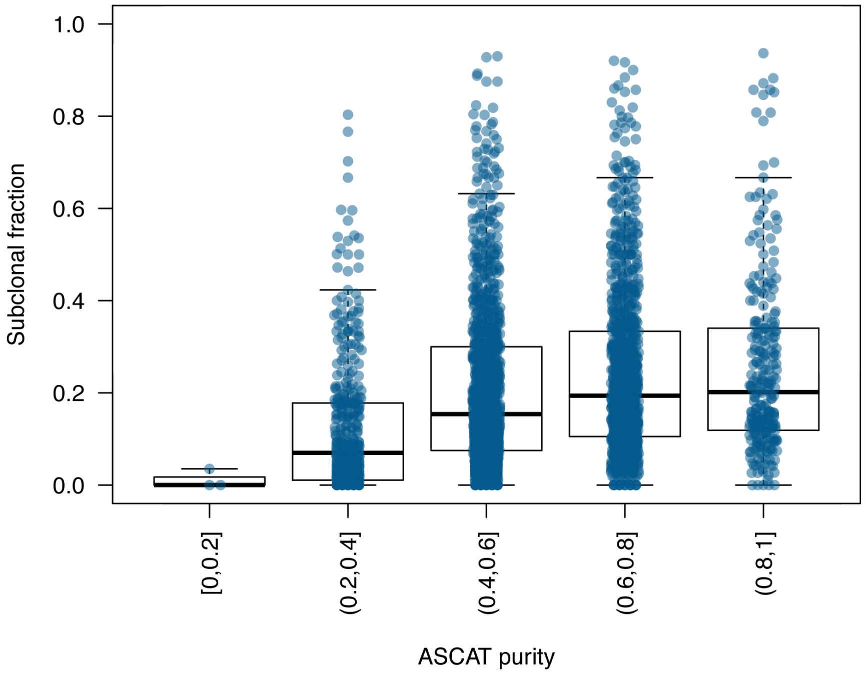


Figure 3:6 Tumour purity in relation to subclonal fractions

For the entire pan-cancer cohort, the subclonal fraction is depicted in relation to different ASCAT purity levels. Each blue dot represents one TCGA tumour. As can be seen, tumours with lower purity exhibit a lower subclonal fraction.

Although it is difficult to directly compare different cancer types, given differences in tumour coverage and tumour purity as well as mutational calling strategies, it is interesting to note that SKCM was identified as the tumour type with, on average, the lowest subclonal fraction. Given that most SKCM samples in TCGA represent metastatic samples, these data potentially indicate a bottlenecking event occurring during metastatic dissemination in this tumour type.

3.2.5 Deciphering clonal heterogeneity of SNVs within multi-region tumour samples

For multi-region sequencing samples, the clonal estimation was further refined using a Dirichlet clustering process extended into multiple dimensions (see Experimental procedures). This analysis was undertaken in collaboration with David Wedge at the Sanger Institute. As outlined in the Introduction, a Dirichlet clustering process can be implemented in order to cluster mutations within similar cancer cell fraction estimates. Mutation clusters likely represent subclones that are present, or have been present, during tumour evolution.

For most pair-wise comparisons between tumour regions, a cluster at (1,1) was detected. Such mutation clusters represent mutations present in 100% of tumour cells within at least two tumour regions. Mutations present at (1,1) for each and every region comparison represent fully clonal mutations, identified as present in 100% of tumour cells sequenced. These mutations likely occurred prior to the emergence of the most recent common ancestor. Conversely, mutations present at (0,1) or (1,0) reflect mutations giving the illusion of clonality. That is, such mutation clusters appear clonal within one tumour region but are not identified within another tumour region. For example, in L002, over 10,000 mutations were identified as clonal in region R1, but not present in region R3.

As in TCGA data, subclonal mutations were also identified within single biopsies. For example, all private mutations within region R3 in L001 were classified as present in less than 100% of cancer cells within that tumour region, and in region R1 of L008, both a regionally clonal and regionally subclonal mutation cluster was identified, harbouring private mutations only identified in region R3.

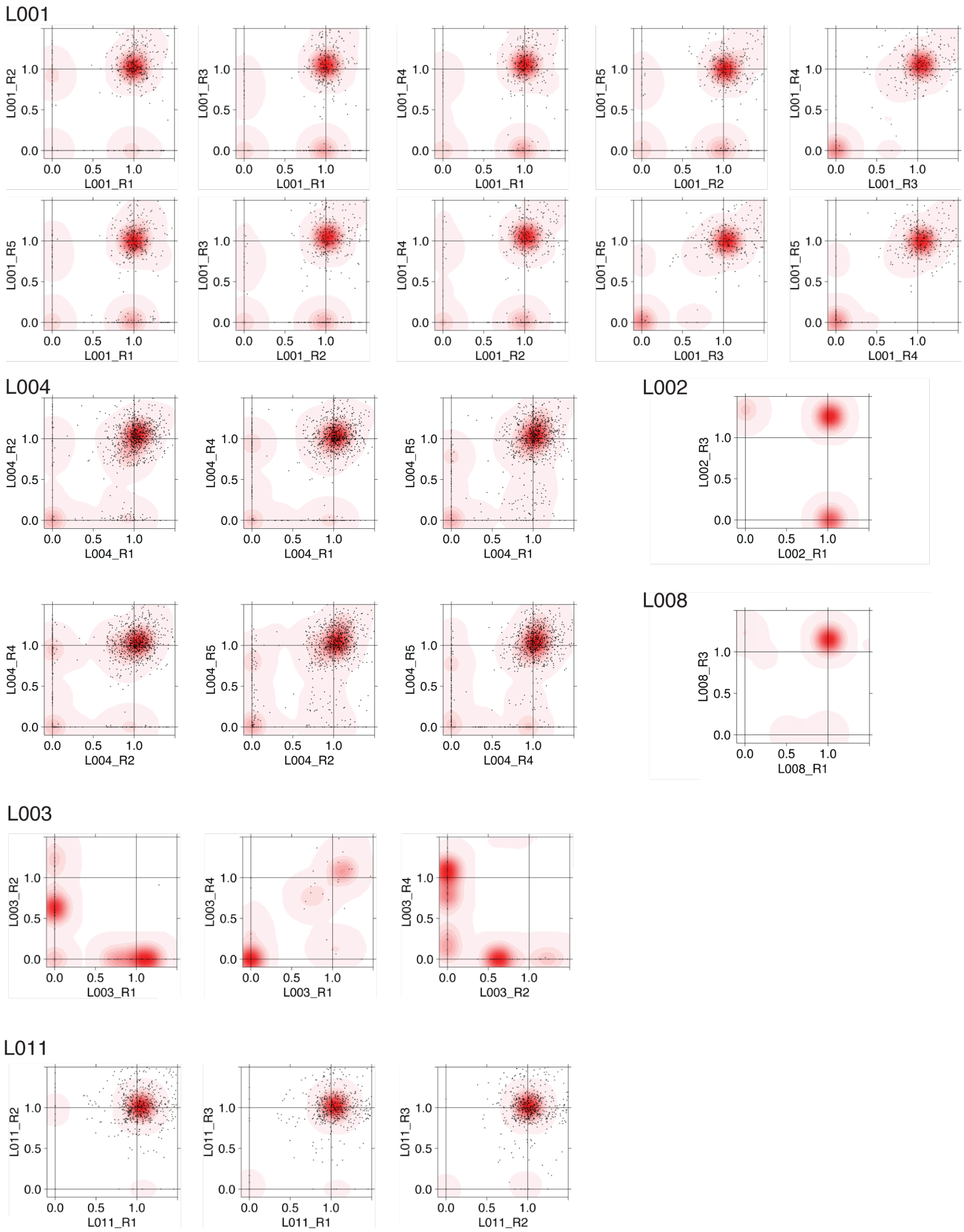


Figure 3:7 2D Dirichlet plots for multi-region NSCLC tumours

2D Dirichlet plots showing mutation clusters based on cancer cell fraction estimates. Clusters at (1,1) represent mutations present in 100% of cancer cells within at least two tumour regions. Conversely, mutation clusters at (1,0) or (0,1) represent mutations that give an illusion of clonality, appearing fully clonal within one tumour region, but not detectable within another tumour region.

3.2.6 Deciphering timing of SNVs and phylogenetic relationships of multi-region tumour samples

Next, for multi-region sequencing data, cancer cell fraction estimates for individual tumour regions were used to construct phylogenetic trees, to yield insights into the relationships between tumour regions and evolutionary histories of individual tumours. Phylogenetic trees were constructed using the maximum parsimony approach (see Experimental procedures).

When constructing phylogenetic trees it is important to recognise that subclonal mutations, present in only a subset of cancer cells, are not necessarily later events, but may reflect copy number driven heterogeneity. For example, in L008 a segment of the genome on chromosome 7q was lost specifically in region R3 but not region R1. As a result, any mutations present on this chromosomal segment will only be found in region R1, despite the fact that many of these mutations likely occurred prior to the emergence of the common ancestor between region R1 and region R3.

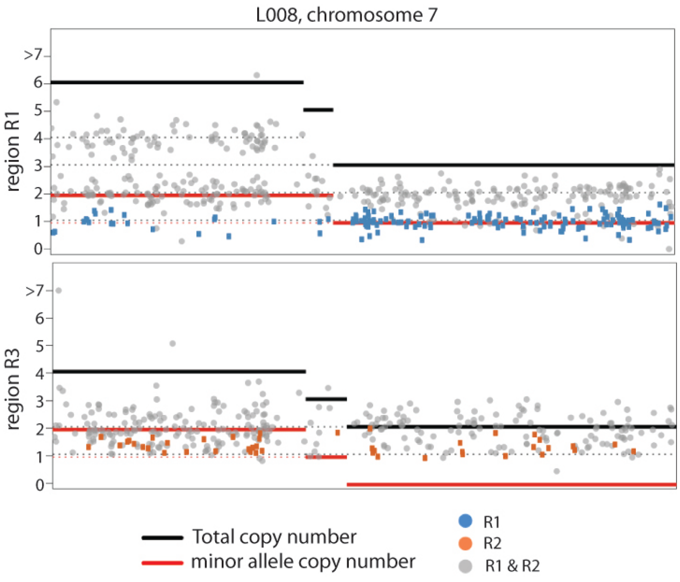


Figure 3:8 Copy number driven heterogeneity in L008

A genomic segment on chromosome 7q was lost specifically in region R3, but not in region R1. Loss of this segment likely creates intra-tumour heterogeneity as mutations which were previously homogenous become heterogeneous, only to be found in one tumour region.

In addition, certain tumour regions were found to harbour multiple distinct subclones, such that the region tree, where each region is considered as a separate species, did not match the clonal dissection tree. For instance, in L004 region R5 harboured three distinct mutations clusters (Figure 3:7). One cluster was present at 100% cancer cell fraction, thereby representing the ancestral clone. Another cluster was present at a cancer cell fraction of 80%, and was unique to region R5, thereby representing a leaf-node on the tree. The final, smallest cluster, present at 20%, contained mutations also present in regions R1, thereby suggesting a minor subclone within region R5.

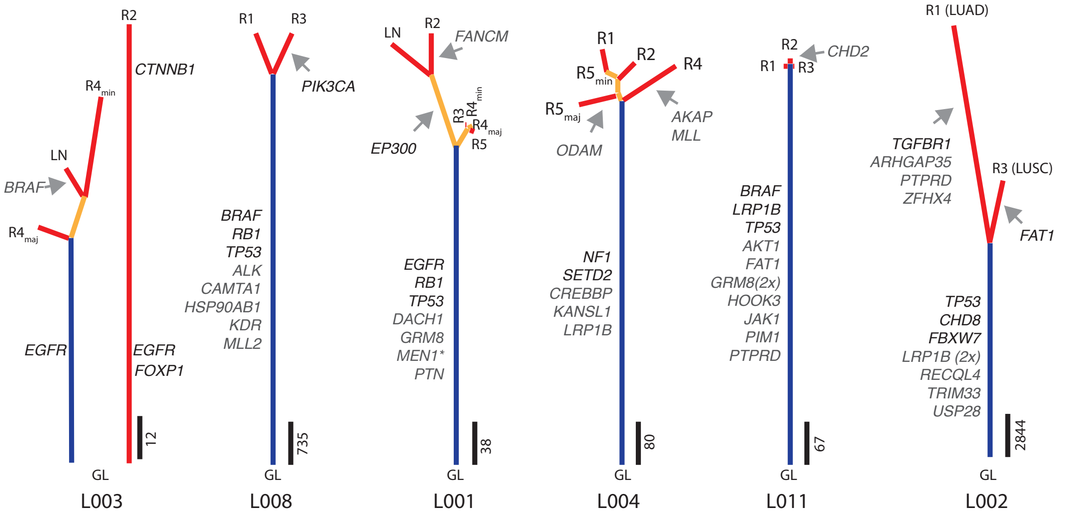


Figure 3:9 Phylogenetic trees of multi-region NSCLC tumours

Phylogenetic trees generated by a maximum parsimony approach based on the distribution of all detected mutations. Scale is indicated for each sample next to the trunk with the number of mutations (silent and non-silent) indicated. GL indicates germ-line. Trees of L002 and L008 are based on WGS data, and the remaining ones on WES data. Categories 1 and 2 driver mutations (black and grey) are indicated next to the trunk or with an arrow pointing to the branches where they were acquired (see Experimental procedures).

For each NSCLC, branched tumour evolution was observed, with distinct subclones observed in separate tumour regions. It is also notable that, in general, very few mutations that were subclonal in multiple regions were identified, suggesting that NSCLCs often harbour geographically distinct, regionally dominant, subclones, with little subclonal mixing.

3.2.7 Heterogeneity of SNVs in tumour suppressors and oncogenes

To explore the regional heterogeneity of candidate tumour driver mutations in multi-region NSCLC, non-silent mutations were classified into four categories, based on the confidence that it was a driver mutation (Table 8-2, see Experimental Procedures). Categories 1-3 contained all non-silent mutations that were present in genes that have previously been implicated as tumour driver mutations, with category 1 representing ‘high confidence’, category 2 ‘probable’ and category 3 ‘potential’ driver mutations. Conversely, category 4 represented all other non-silent mutations. Further details regarding mutation classification can be found in the Experimental procedures section.

Within the multi-region sequenced samples a median of three ‘high confidence’ (category 1), six ‘putative’ (category 2) and six ‘low confidence’ (category 3) driver mutations were identified in each tumour (Table 8-2). Every tumour showed evidence for regional heterogeneity in driver mutations (Figure 3:9). Moreover, many of the heterogeneous driver mutations gave an illusion of clonality in one tumour region. For example, mutations in *TGFBR1* and *EP300* were clonally dominant in a subset of tumour regions but not detectable in other regions from the same tumour (Table 8-2). These data further imply that subclonal estimates based on single samples likely underestimate the true extent of heterogeneity.

To investigate heterogeneity and clonal status of candidate driver mutations in the pan-cancer cohort, non-silent mutations were classified as occurring in established cancer genes based on the recent pan-cancer analysis by Lawrence et al. (2014), combined with the manually curated COSMIC gene census (Futreal et al., 2004).

Consistent with multi-region results in NSCLC, non-silent subclonal mutations in known cancer genes were identified in every cancer type (Figure 3:10).

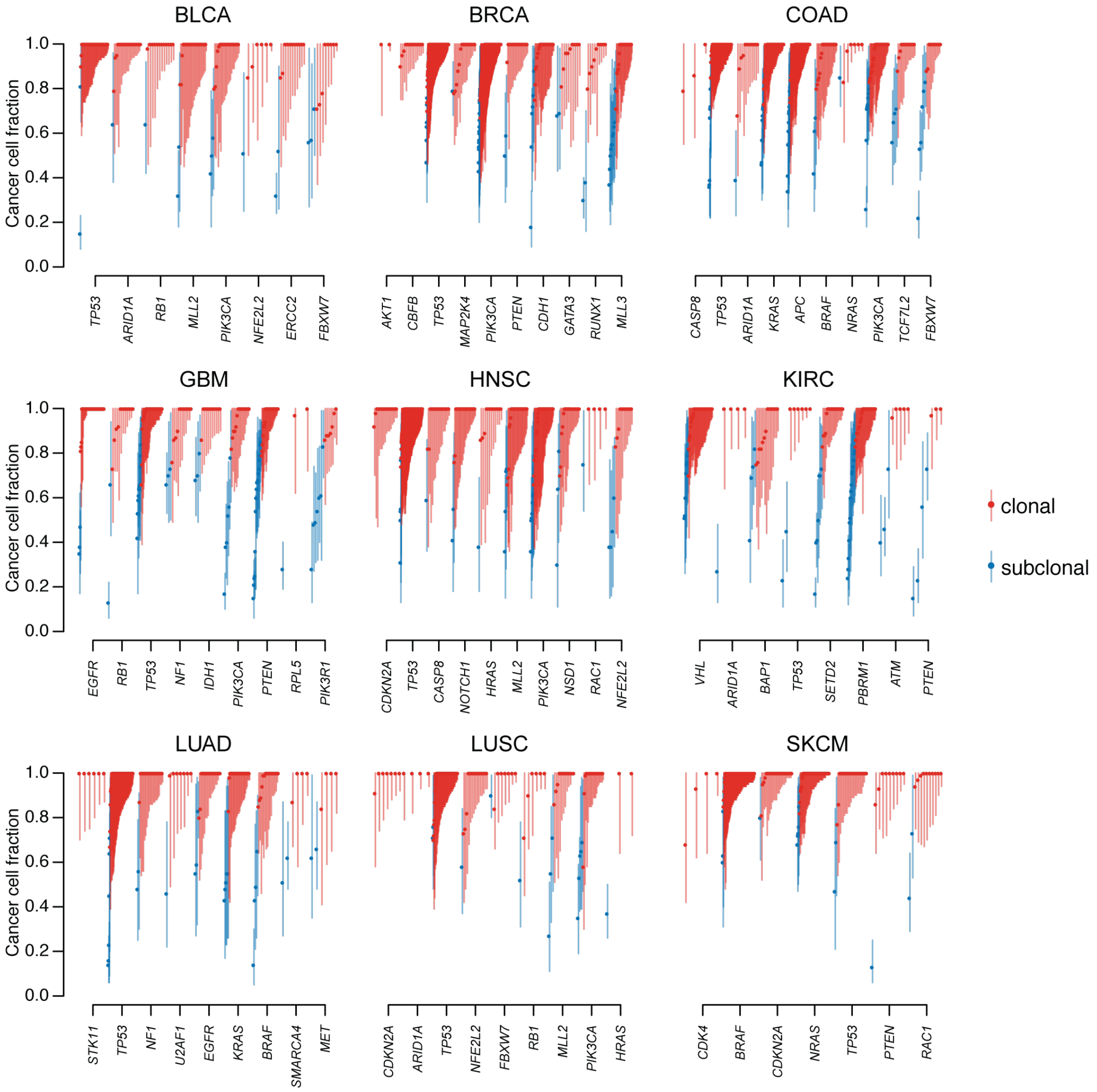


Figure 3:10 Pan-cancer cancer cell fraction of cancer genes

Pan-cancer cancer cell fraction estimates for established cancer related genes. For each tumour type, clonal (red) and subclonal (blue) mutations in established cancer genes are shown. Each dot represents one mutation, with 95% confidence interval of the cancer cell fraction indicated. Mutations are classified as clonal if the 95% confidence interval overlaps with 1.

3.2.8 Established driver mutations typically occur early in tumour evolution

To explore whether certain genes had a tendency to be clonal or subclonal, beyond what might be expected by chance, permutation tests were performed. In KIRC, over 50% of non-silent mutations in *PTEN* were found to be subclonal (significantly more than the background rate, Figure 3:11 *P*=0.0116). These results are consistent with observations from a smaller cohort of KIRC samples subject to multi-region sequencing (Gerlinger et al., 2014a), and support the role that inactivation of *PTEN* may play later in tumour evolution in this cancer type. In BRCA, mutations in *CBFB* were exclusively clonal, as were mutations in *CDKN2A* in LUSC and HNSC, and a similar trend was observed for mutations in *ARID1A* in BLCA.

Within the multi-region NSCLC cohort, four tumours (L002, L011, L001, and L008) were found to harbour mutations in *TP53* and, in each case, these were identified as fully clonal, occurring in every cancer cell sequenced. Consistently, in TCGA LUAD and LUSC tumours and in the pan-cancer dataset as a whole, somatic mutations in *TP53* were significantly more often clonal than the background rate (Figure 3:11, *P*<0.0001). Conversely, mutations in *PIK3R1* and *MLL3* were frequently subclonal across cancer types, such that a higher proportion of subclonal mutations was observed in these genes compared to the background rate (Figure 3:11).

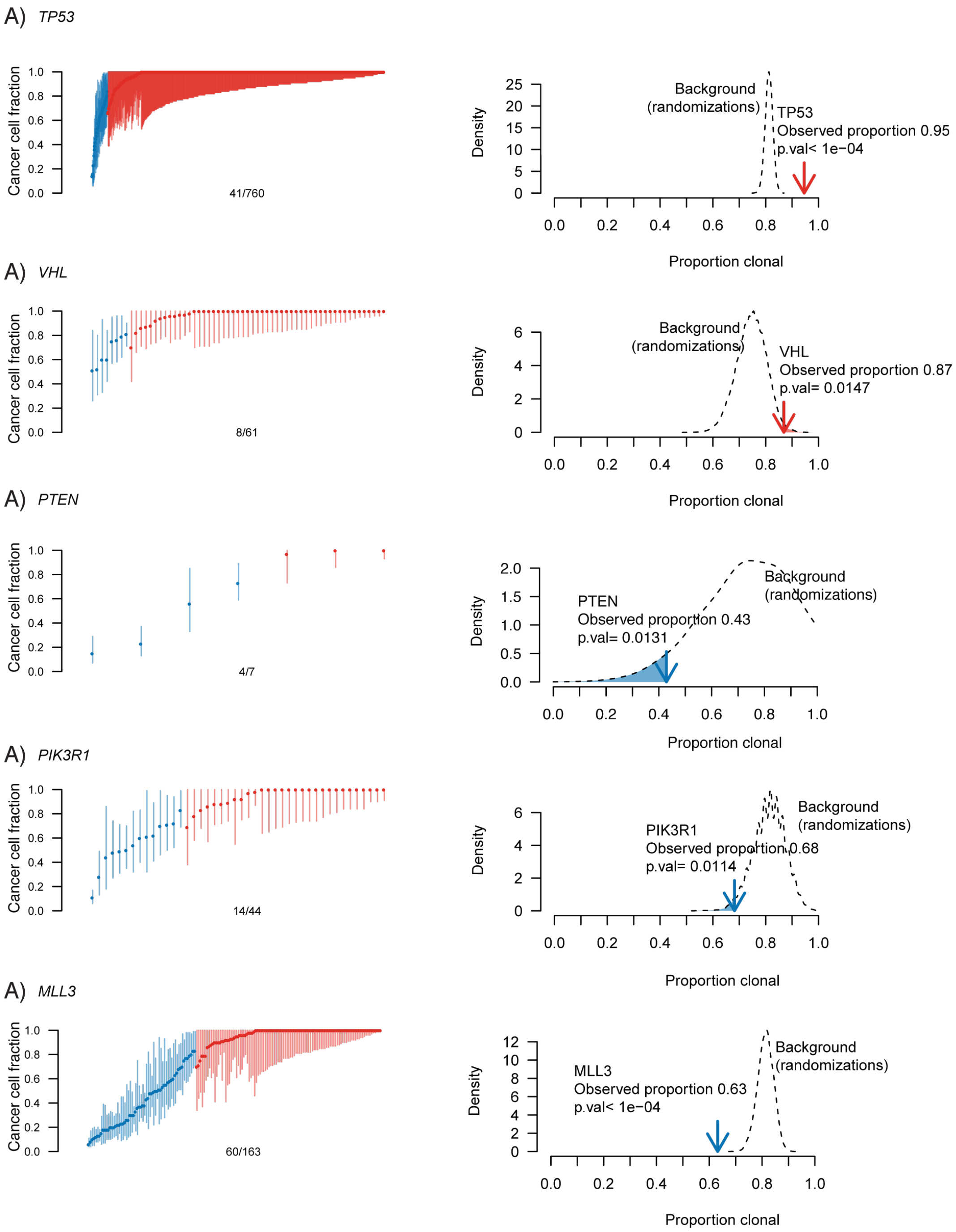


Figure 3:11 Clonal status significance using permutations

In the left panel, for each gene the proportion of mutations that are clonal (red) and subclonal (blue) are depicted. The significance is shown in the right panel, comparing the observed proportion of clonal mutations with the background distribution obtained from 10,000 randomizations, assuming the same mutation number. As can be seen, mutations in *TP53* (across the pan-cancer cohort) and *VHL* (in KIRC tumours) are significantly more often clonal compared to background, whereas mutation in *PTEN* in KIRC and *PIK3R1* as well as *MLL3* in the pan-cancer cohort show a significant tendency to be subclonal.

Next, to determine whether mutations in established cancer genes had a tendency to be clonal or subclonal, all non-silent mutation in cancer genes were grouped together and compared to all non-silent mutations in non-cancer genes. In multi-region NSCLC tumours, driver mutations were significantly more often truncal compared to non-driver mutations (*P*=0.016, Fisher's exact test), indicating that established driver mutations tend to occur early in tumour evolution. This observation was confirmed across the pan-cancer cohort, where a clear tendency for mutations in driver genes to be clonal compared to mutations in non-cancer genes was observed (Figure 3:12). In every cancer type, with the exception of KIRC, mutations in driver genes (considered in aggregate) were enriched to a statistically significant degree for clonal mutations, compared to mutations in non-driver genes. In KIRC, *VHL* was the only cancer gene that had a significantly higher proportion of clonal mutations than the background rate representing all non-silent mutations, consistent with multi-region sequencing results in KIRC (Gerlinger et al., 2014a) and supporting the approach to distinguish clonal from subclonal somatic events used in this thesis (*P* = 0.0147, Figure 3:11). These results remained robust to multiple different cut-offs to define clonal/subclonal mutations.

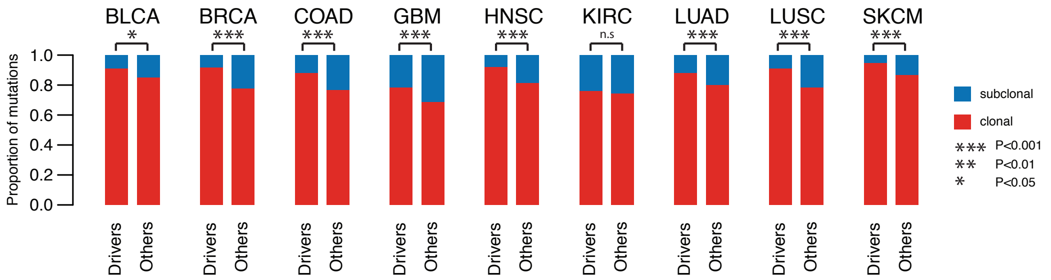


Figure 3:12 Clonal versus subclonal mutations in cancer driver genes
Bar-plots showing the proportion of clonal versus subclonal non-silent mutations in driver genes and other genes across 9 cancer types.

Taken together, these results suggest that many known cancer genes have a tendency to be clonal within single samples and corroborate the 'driver' capabilities of these genes. These data also suggest that mutations in known driver genes frequently occur before subclonal diversification, and are less likely to be subject to sampling bias compared to non-driver genes. However, although a tumour may harbour a founder driver aberration, such as mutation to *TP53*, branched subclonal expansions can result in additional driver events, such as a mutation to *PIK3R1*, dominating only one region of the tumour, or a subset of tumour cells within a region.

3.2.9 Recurrent patters in tumour evolution

Across the pan-cancer cohort in a minority of cases multiple subclonal mutations occurring in the same cancer gene or in separate cancer genes whose function might be expected to be redundant were identified.

For example, in KIRC-B0-5399, two distinct mutations in *SETD2*, one occurring in 40% of cancer cells and the other in 19%, were observed. These observations are consistent with evidence for parallel tumour evolution, whereby the same genetic pathway, gene or protein complex is disrupted independently in distinct tumour subpopulations within one tumour (Gerlinger et al., 2014a). Likewise, in GBM-28-2513, two subclonal mutations in genes involved in the PI3K signalling axis were identified, with one mutation in *PIK3R1* present in 54% of cancer cells and a mutation in *PTEN* present in only 36% of cancer cells.

In further support of parallel evolution, in general when multiple non-silent mutations were identified in the same cancer gene within one tumour sample, these mutations exhibited a significantly lower cancer cell fraction compared to mutations in cancer genes occurring only once in a cancer sample (Figure 3:13, $P<0.001$, Wilcoxon rank sum test).

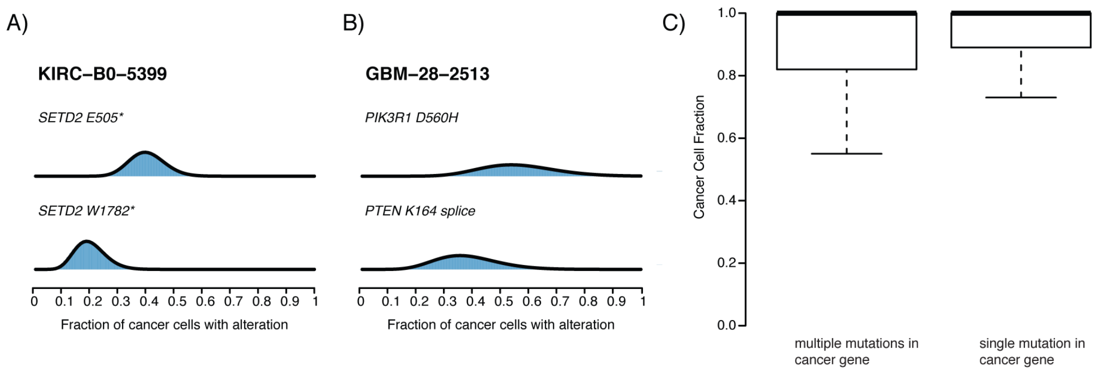


Figure 3:13 Examples of parallel evolution

A-B) Probability distributions over the cancer cell fraction for individual mutations are shown for specific tumours. In both cases, the cancer cell fractions are consistent with mutations occurring in independent tumour subclones. C) When multiple non-silent mutations were identified in the same cancer gene within one tumour sample these mutations exhibited a significantly lower cancer cell fractions compared to mutations in cancer genes occurring only once in a cancer sample (*P*=7.151e-07, Wilcoxon rank sum test).

3.2.10 Identification of novel putative subclonal and clonal cancer genes using temporal and clonal dissection

Given the observation that many clonal mutations likely occur before tumourigenesis and the fact that many branches of NSCLC phylogenetic trees did not harbour known driver genes, it seems reasonable to postulate that it might be possible to identify drivers of subclonal expansions by focusing exclusively on late or subclonal mutations (see Experimental procedures).

The MutSigCV algorithm (Lawrence et al., 2014), a statistical analysis that takes into account nucleotide context, gene expression, replication timing, and the somatic background mutation rate to identify cancer genes, was therefore applied to temporally and clonally dissected mutations in the pan-cancer cohort. Given that MutSigCV requires a large number of tumour samples in order to identify putative cancer genes, multi-region NSCLC samples were not utilized for this part of the analysis.

In total, 32 late putative driver genes were identified across the 9 cancer types (q<0.05, Table 8-3). Of these, 12 would have been missed in at least one cancer type without temporally dissecting mutations. Temporal dissection was required to uncover the cancer gene *PIK3CA* in HNSC, suggesting that mutations in this gene may often lead to, or be permissive for, subclonal expansions in this cancer type.

A number of putative cancer genes that were identified can be linked to tumour development, maintenance, and progression. A notable example is cell adhesion gene *CTNNA2*, catenin (cadherin-associated protein) alpha 2, identified in LUAD. *CTNNA2* has previously been implicated in laryngeal cancer as a tumour suppressor, and its inactivation in HNSC cells is associated with migration and invasion advantages, consistent with it playing a role at later stages of tumour development (Fanjul-Fernandez et al., 2013). In LUAD, mutations in *NRXN3* were identified as a putative subclonal driver event. A polymorphic site of this gene (rs10146997) has been associated with higher risk of breast cancer development (Kusinska et al., 2012), and low expression of *NRXN3* is associated with reduced overall-survival in lung cancer (Figure 3:14).

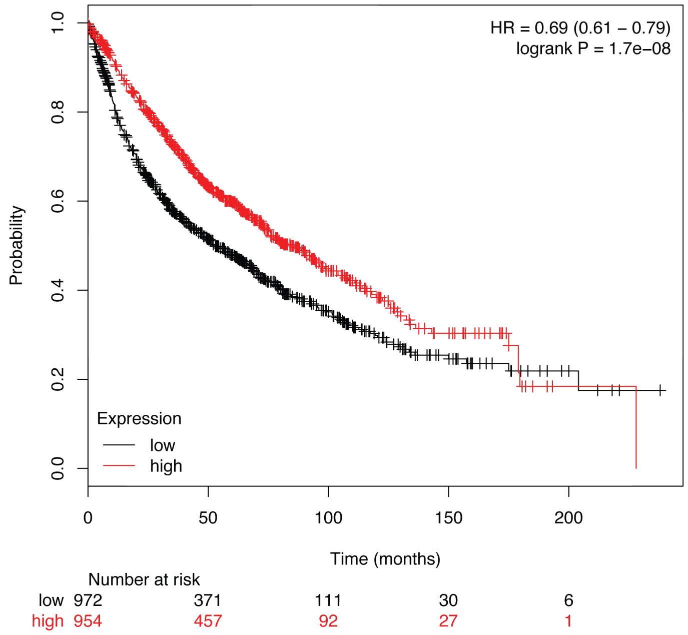


Figure 3:14 Overall survival and expression of NRXN3.

High expression of NRXN3 in NSCLC is associated with significantly improved prognosis compared to low expression. High and low expression groups were defined relative to median expression level. Survival plot was generated using KM plotter (www.kmplot.com).

Another putative cancer gene, identified in COAD, was *ATXN1*. The ATXN1 protein family plays an important role in transcriptional control of extracellular matrix remodelling, and mutations in *ATXN1* have been putatively linked to cancer metastasis, consistent with its occurrence as a later event in COAD (Lee et al., 2011b). In LUSC, *LRP1B*, which encodes a member of the low-density lipoprotein (LDL) receptor gene family, was identified as a putative cancer gene. It has been suggested that *LRP1B* acts as a tumour suppressor gene (Varela et al., 2010), and in KIRC its depletion results in increased anchorage-independent growth, cell migration, and invasion *in vitro* (Ni et al., 2013).

By temporally separating mutations, driver genes that may be missed by the inclusion of late or subclonal mutations were also identified. For example, when focusing exclusively on clonal mutations it was possible to identify *BRAF* in COAD, consistent with BRAF playing an important role early in tumour evolution and the clonal nature of this event in published studies (Brannon et al., 2014). Likewise, in LUSC, *BRAF*, *KRAS,* and *EGFR* were only identified as drivers in this cancer type by focusing on early mutations.

3.2.11 Mutations in genes with clinical relevance can be subclonal

Mutations in genes and gene pathways for which therapeutics have been developed or are in development (Table 8-4), have received much attention in oncology given their appeal as targets for personalized medicine. For instance, the presence of mutations that activate the PI3K/AKT/mTOR pathway and contribute to carcinogenesis has engendered much interest in inhibitors of this signalling axis (Janku et al., 2011).

It is therefore notable that, with the exception of *CDKN2B* and *CDKN1B*, for every gene that has been linked with a targeted therapy approach, a subclonal mutation was identified in at least one tumour in the pan-cancer cohort (Figure 3:15). Over 10% of all non-silent *PIK3CA* mutations and over 20% of all non-silent mutations in *PTEN* were found to be subclonal and over 15% of mutations in genes in the PI3K/AKT/mTOR pathway overall were identified as subclonal (Figure 3:15 B). In

GBM, the use of IDH-targeted therapies has been proposed for tumours with *IDH1* or *IDH2* mutations (Rohle et al., 2013), and yet over 20% of *IDH1* mutations in GBM were identified as subclonal. On the other hand, all *IDH1* mutations detected in SKCM were clonal. In KIRC, mTOR inhibition is a common therapeutic option; however, over 30% of mutations in *mTOR* were subclonal within this disease type within single tumour samples.

Clear differences between different cancer pathways were also observed; significantly fewer subclonal mutations in genes associated with cyclin dependent kinases (*CDKN1A*, *CDKN1B*, *CDKN2B*, *RB1*, *CDKN2A*, *CDK6,* and *CDK5*) and the RAS-MEK pathway were observed compared to mutations in the AKT/mTOR/PI3K pathway (P<0.05, Fisher's exact test).

Finally, in order to restrict the analysis to well-characterized mutations, some of which are considered 'actionable', the database of curated mutations (DoCM, docm.genome.wustl.edu) was used. Only mutations occurring in at least three tumour samples within the cohort were considered. For the majority of these mutations, both clonal and subclonal mutations were identified in at least one cancer type (Figure 3:16). Of particular clinical relevance, tumour samples with subclonal mutations in known sites with therapeutic relevance such as *NRAS* (Q61K, Q61R, Q61L)*; BRAF* (V600E); *KRAS* (G12C, G12D, G12V); *PIK3CA* (E542K, 545K, H1047R); *IDH1* (R132H), as well as many subclonal loss-of-function mutations in tumour suppressor genes such as *PTEN* (Figure 3:17) were observed.

Identified subclonal driver mutations often occurred in tumours where clonal mutations in established cancer genes were also present (Figure 3:17; mean 45% across cancers [range 15-67%]). For example, in patient HNSC-CV-7177, a *PIK3CA* (E545K) mutation in the highly conserved helical domain was estimated to be present in only 36% of cancer cells, whereas a mutation in *TP53* in the same tumour was found to be present in all cancer cells. Similarly, patient SKCM−ER−A2NE exhibited a clonal mutation in *NRAS* (Q61K), whereas a *PTEN* mutation (Y178*) was present only in 13% of cells.

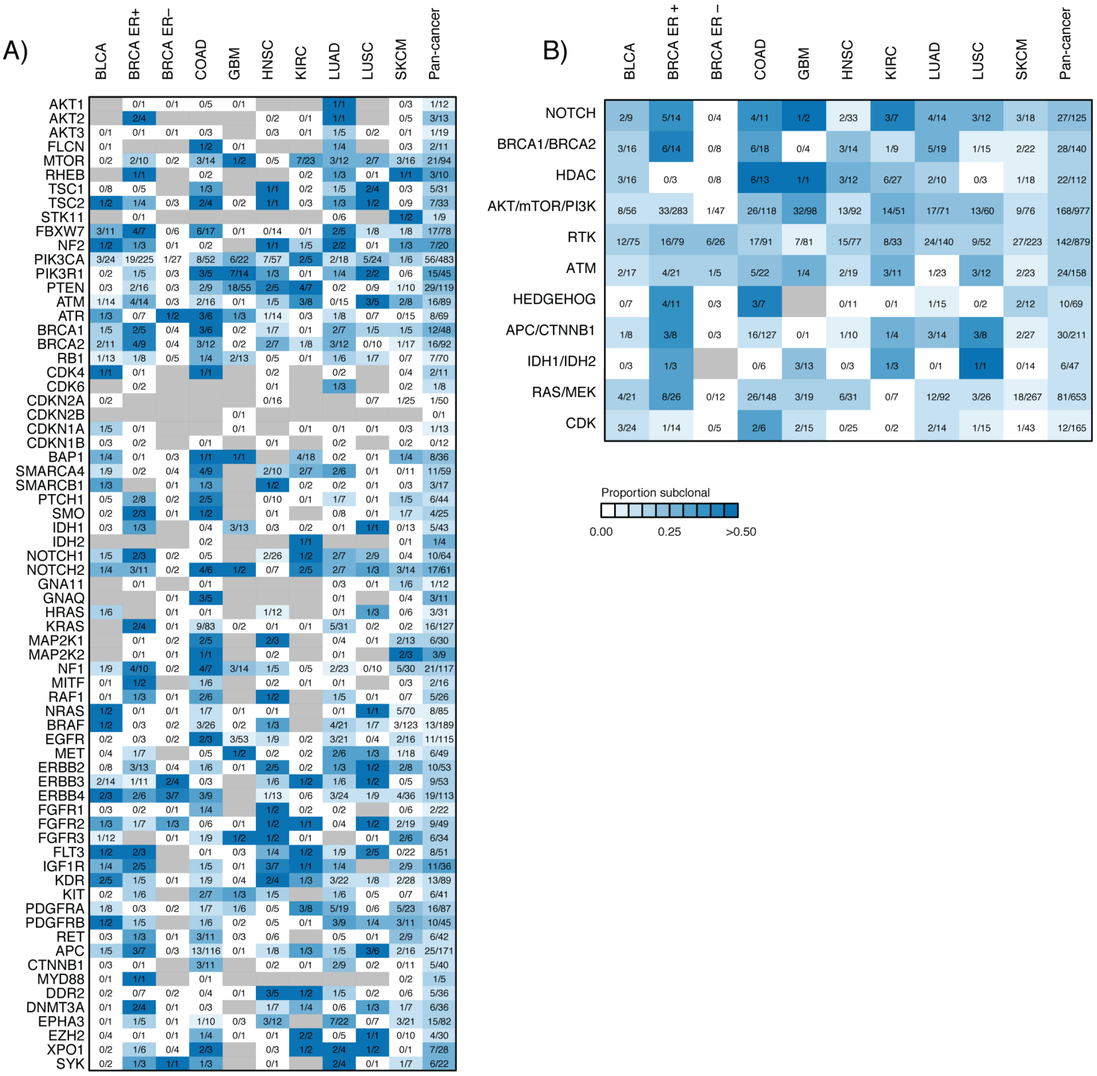
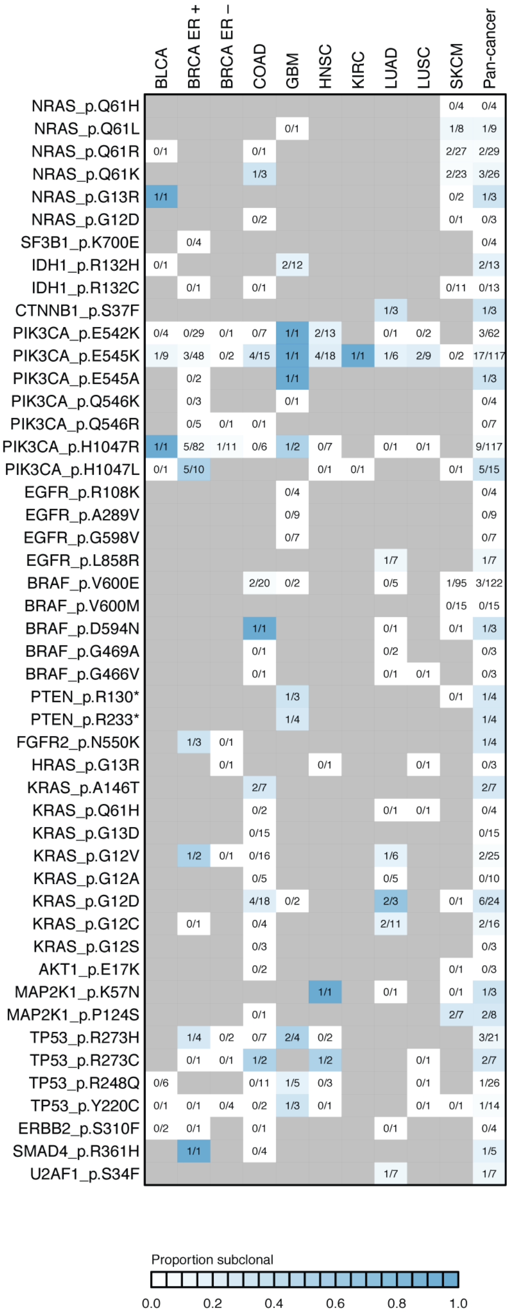


Figure 3:15 Heterogeneity of mutations in genes linked with therapies

(A) Heatmap showing the proportion of non-silent mutations that are subclonal for each potentially actionable gene across nine cancer types. For each mutation, the number of subclonal mutations identified is indicated for each cancer type and the combined pan-cancer data set. Grey indicates the absence of a mutation. (B) The clonality of actionable pathways is depicted for each cancer type and the combined pan-cancer data set. Pathways are ordered according to sub-clonality in the pan-cancer data set, with pathways that have a higher proportion of subclonal mutations at the top of the heatmap. Genes related to CDKs (cyclin-dependent kinases) have very few subclonal mutations. RTK, receptor tyrosine kinases. For details of all the genes within each pathway, see Table 8-4.



	BLCA	BRCA ER +	BRCA ER –	COAD	GBM	HNSC	KIRC	LUAD	LUSC	SKCM	Pan-cancer
NRAS_p.Q61H										0/4	0/4
NRAS_p.Q61L					0/1					1/8	1/9
NRAS_p.Q61R	0/1			0/1						2/27	2/29
NRAS_p.Q61K				1/3						2/23	3/26
NRAS_p.G13R	1/1									0/2	1/3
NRAS_p.G12D				0/2						0/1	0/3
SF3B1_p.K700E		0/4									0/4
IDH1_p.R132H	0/1				2/12						2/13
IDH1_p.R132C		0/1		0/1						0/11	0/13
CTNNB1_p.S37F								1/3			1/3
PIK3CA_p.E542K	0/4	0/29	0/1	0/7	1/1	2/13		0/1	0/2		3/62
PIK3CA_p.E545K	1/9	3/48	0/2	4/15	1/1	4/18	1/1	1/6	2/9	0/2	17/117
PIK3CA_p.E545A		0/2			1/1						1/3
PIK3CA_p.Q546K		0/3			0/1						0/4
PIK3CA_p.Q546R		0/5	0/1	0/1							0/7
PIK3CA_p.H1047R	1/1	5/82	1/11	0/6	1/2	0/7		0/1	0/1		9/117
PIK3CA_p.H1047L	0/1	5/10				0/1	0/1			0/1	5/15
EGFR_p.R108K					0/4						0/4
EGFR_p.A289V					0/9						0/9
EGFR_p.G598V					0/7						0/7
EGFR_p.L858R								1/7			1/7
BRAF_p.V600E				2/20	0/2			0/5		1/95	3/122
BRAF_p.V600M										0/15	0/15
BRAF_p.D594N				1/1				0/1		0/1	1/3
BRAF_p.G469A				0/1				0/2			0/3
BRAF_p.G466V				0/1				0/1	0/1		0/3
PTEN_p.R130*					1/3					0/1	1/4
PTEN_p.R233*					1/4						1/4
FGFR2_p.N550K		1/3	0/1								1/4
HRAS_p.G13R			0/1			0/1			0/1		0/3
KRAS_p.A146T				2/7							2/7
KRAS_p.Q61H				0/2				0/1	0/1		0/4
KRAS_p.G13D				0/15							0/15
KRAS_p.G12V		1/2	0/1	0/16				1/6			2/25
KRAS_p.G12A				0/5				0/5			0/10
KRAS_p.G12D				4/18	0/2			2/3		0/1	6/24
KRAS_p.G12C		0/1		0/4				2/11			2/16
KRAS_p.G12S				0/3							0/3
AKT1_p.E17K				0/2						0/1	0/3
MAP2K1_p.K57N						1/1		0/1		0/1	1/3
MAP2K1_p.P124S				0/1						2/7	2/8
TP53_p.R273H		1/4	0/2	0/7	2/4	0/2					3/21
TP53_p.R273C		0/1	0/1	1/2		1/2			0/1		2/7
TP53_p.R248Q	0/6			0/11	1/5	0/3			0/1		1/26
TP53_p.Y220C	0/1	0/1	0/4	0/2	1/3	0/1			0/1	0/1	1/14
ERBB2_p.S310F	0/2	0/1		0/1				0/1			0/4
SMAD4_p.R361H		1/1		0/4							1/5
U2AF1_p.S34F								1/7			1/7

Figure 3:16 Clonal and subclonal actionable mutations

For each known ‘actionable’ site (http://docm.genome.wustl.edu/), the number of subclonal (numerator) and total mutations (denominator) is listed for each cancer type. Grey indicates absence of a mutation.

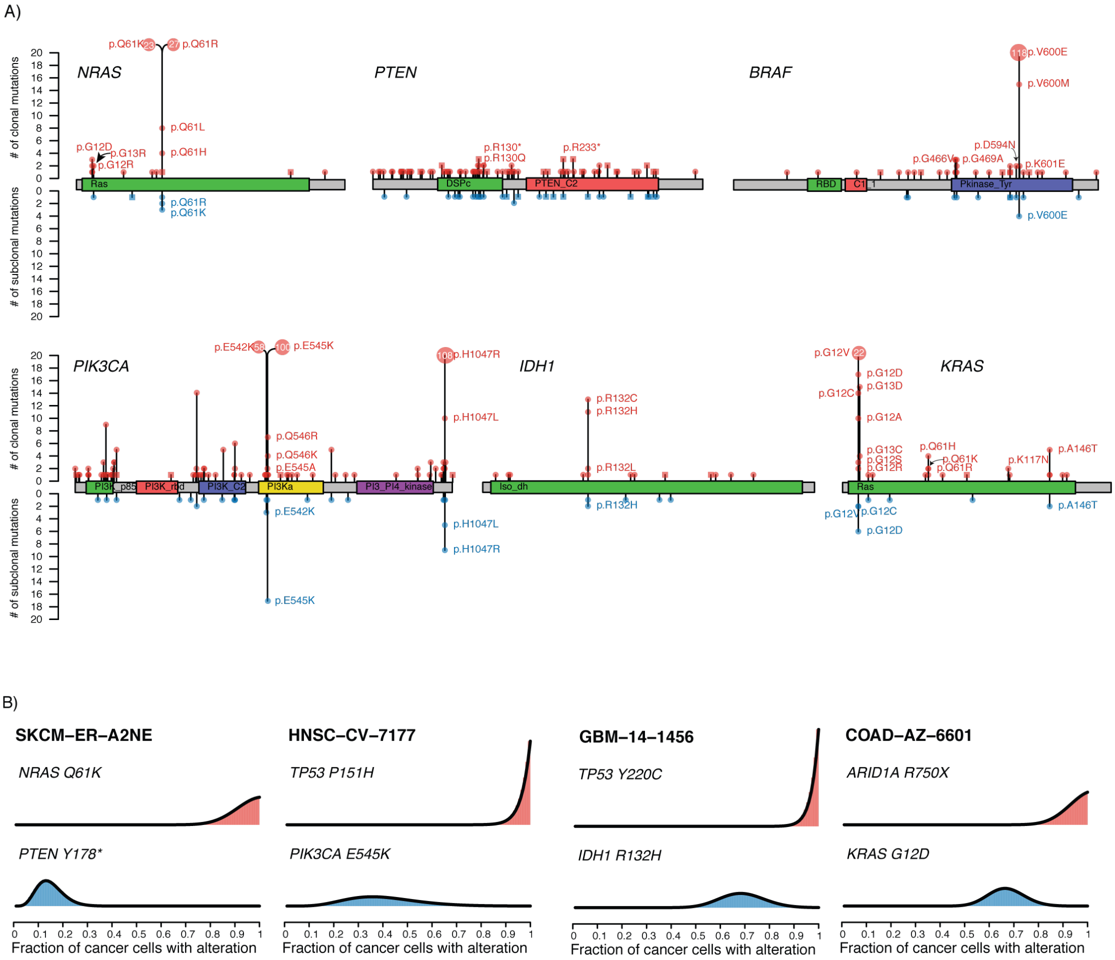


Figure 3:17 Clonal and subclonal actionable mutations across cancers

A) For specific cancer genes, lollipop plots are shown depicting the frequency of clonal and subclonal mutations across the coding sequence. B) Probability distributions over the cancer cell fraction for individual mutations are shown for specific tumours.

3.2.12 Multi-region sequencing in oesophageal adenocarcinoma

Work in this chapter has thus far has focussed on predominantly primary tumours that have not received any treatment. In order to gain a deeper understanding of the clinical implication of intra-tumour heterogeneity, and, further, explore the impact of treatment on tumour evolution and tumour diversity, multi-region sequencing data from eight oesophageal adenocarcinomas pre and post-exposure to platinum chemotherapy was obtained. The clinical characteristics of these 8 patients are shown in Table 3-2, and included patients with stage IIB to IIIC disease.

Table 3-2 Clinical characteristics of oesophageal multi-region tumours

Sample ID	Gender	Pre-op TNM Staging	Pre-op Stage	Mandard Score	Path TNM Stage	Post-op Stage	Response	Outcome
OG005	M	T2N1	2B	5	pT3N2	3B	Upstaged	Poor
OG015	M	T3N1	3A	5	pT4aN3	3C	Upstaged	Poor
OG001	M	T3N0	3A	5	pT3N0	2B	Same	Intermediate
OG006	M	T1N0	1B	3	pT1N1	2B	Upstaged	Intermediate
OG014	M	T3N1	3A	4	pT1bN1	2B	Down-staged	Intermediate
OG003	M	T3N1	3A	3	pT3N1	3A	Same	Good
OG017	M	T3N1	3A	3	pT2N1	2B	Down-staged	Good
OG009	M	T4N2	3C	2	pT1bN1	2B	Down-staged	Good

In total 40 tumour regions were sequenced at a median 90-fold coverage (range 56-191), with matched germ-line samples also sequenced to a median depth of 90X (range 42-132). As in NSCLC, mutation calling leveraged data from all tumour regions. The total number of mutations identified if only a single tumour region had been samples was significant lower than the number of mutations identified using multi-region sequencing (P<0.01, Wilcoxon test).

3.2.13 Intra-tumour heterogeneity in oesophageal adenocarcinoma

To determine the evolution and heterogeneity of oesophageal adenocarcinoma, mutations were first classified as either ubiquitously detected, if detected in all tumour regions, or heterogeneous, if identified in only a subset of tumour regions (Figure 3:18). Mutations were further grouped according to whether they were identified pre-chemotherapy or post-chemotherapy treatment.

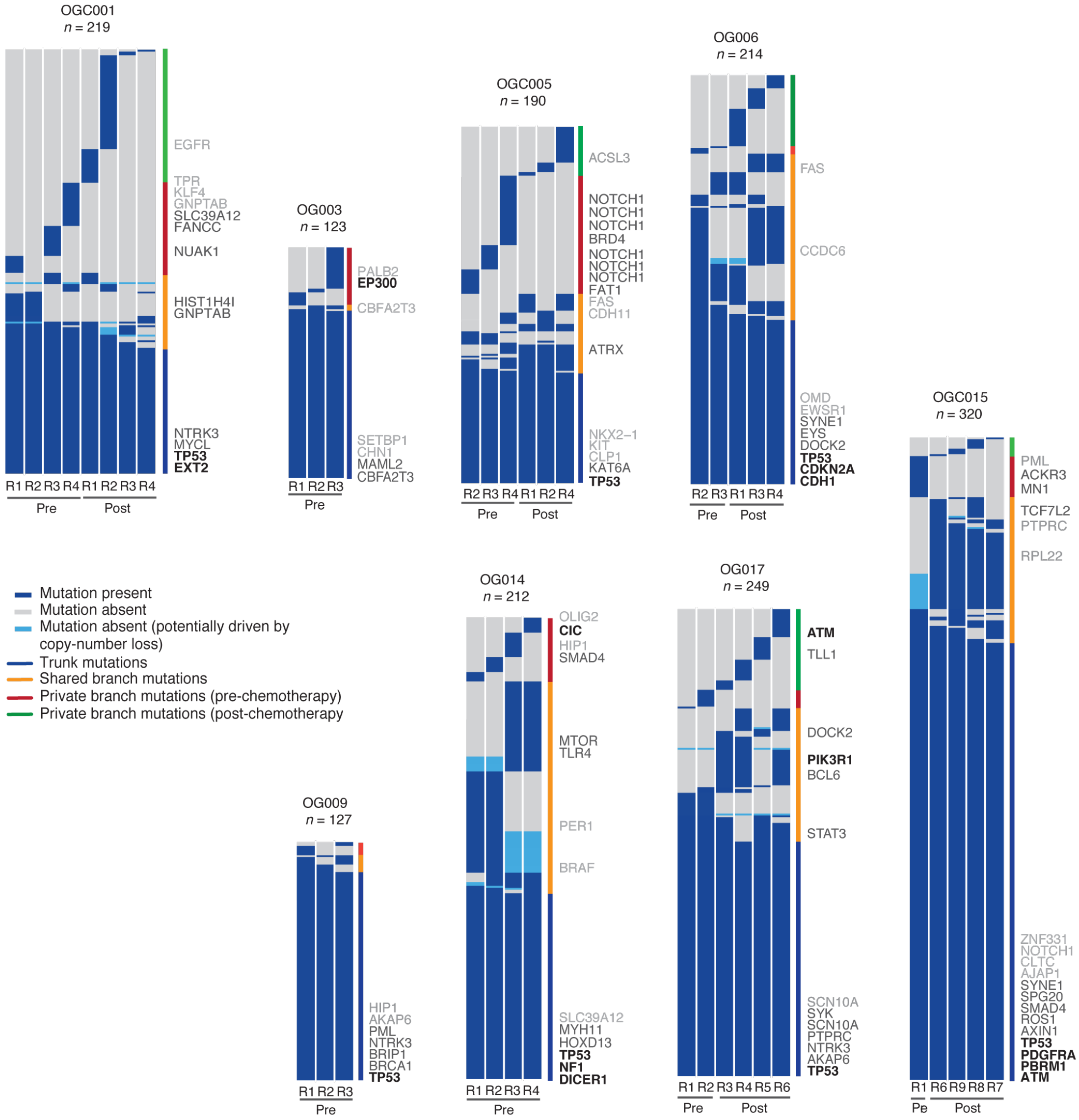


Figure 3:18 Intra-tumour heterogeneity of oesophageal adenocarcinoma

Heatmaps depicting the extent of intra-tumour heterogeneity of non-silent mutations in multi-region sequenced oesophageal adenocarcinomas. Pre- and post- chemotherapy tumour regions are indicated. Category 1-3 mutations in cancer genes are indicated, with black, bold, genes category 1, dark grey category 2 and light grey category 3.

All primary tumours exhibited intra-tumour heterogeneity, indicative of branched tumour evolution (Figure 3:18). A median of 55.63% of non-silent mutations were classified as heterogeneous (range 12.6% to 69.4%). The degree of intra-tumour heterogeneity was assessed in each tumour by the use of the heterogeneity index described above. In this small cohort, a strong correlation between the heterogeneity index and response to neo-adjuvant chemotherapy was observed (Figure 3:19), suggesting a simple heterogeneity score may predict response to therapy in oesophageal adenocarcinoma.

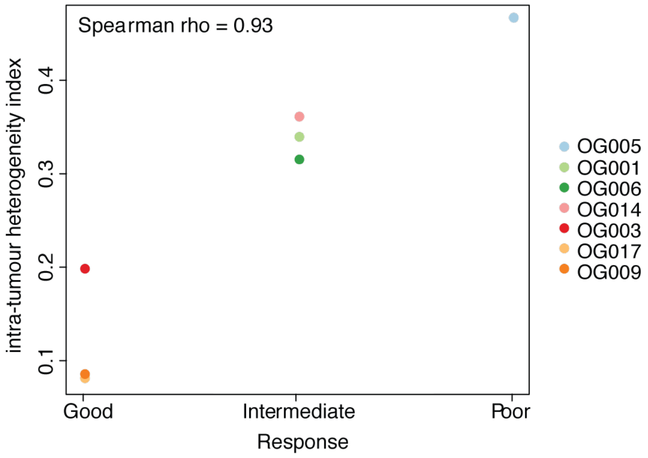
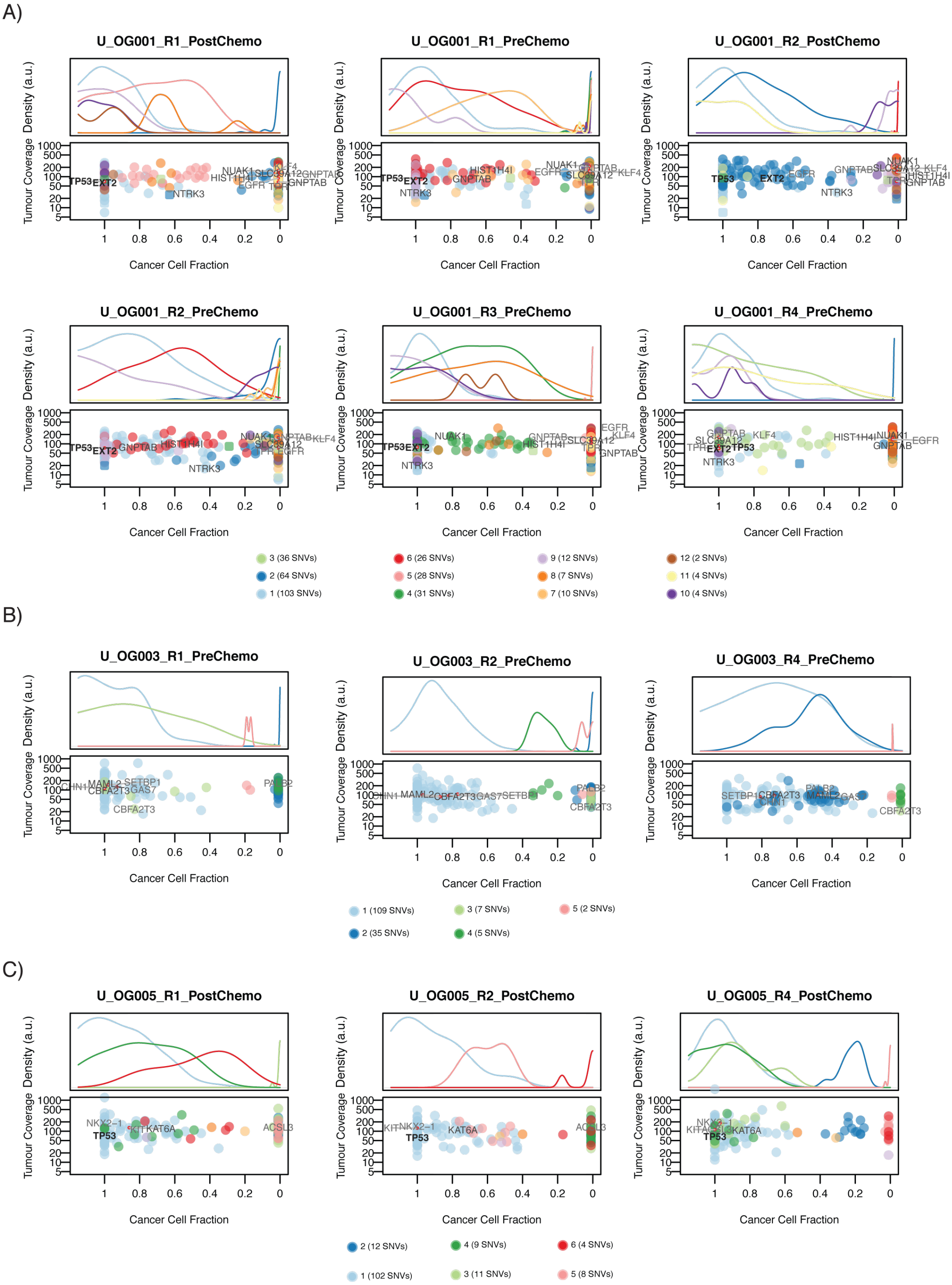


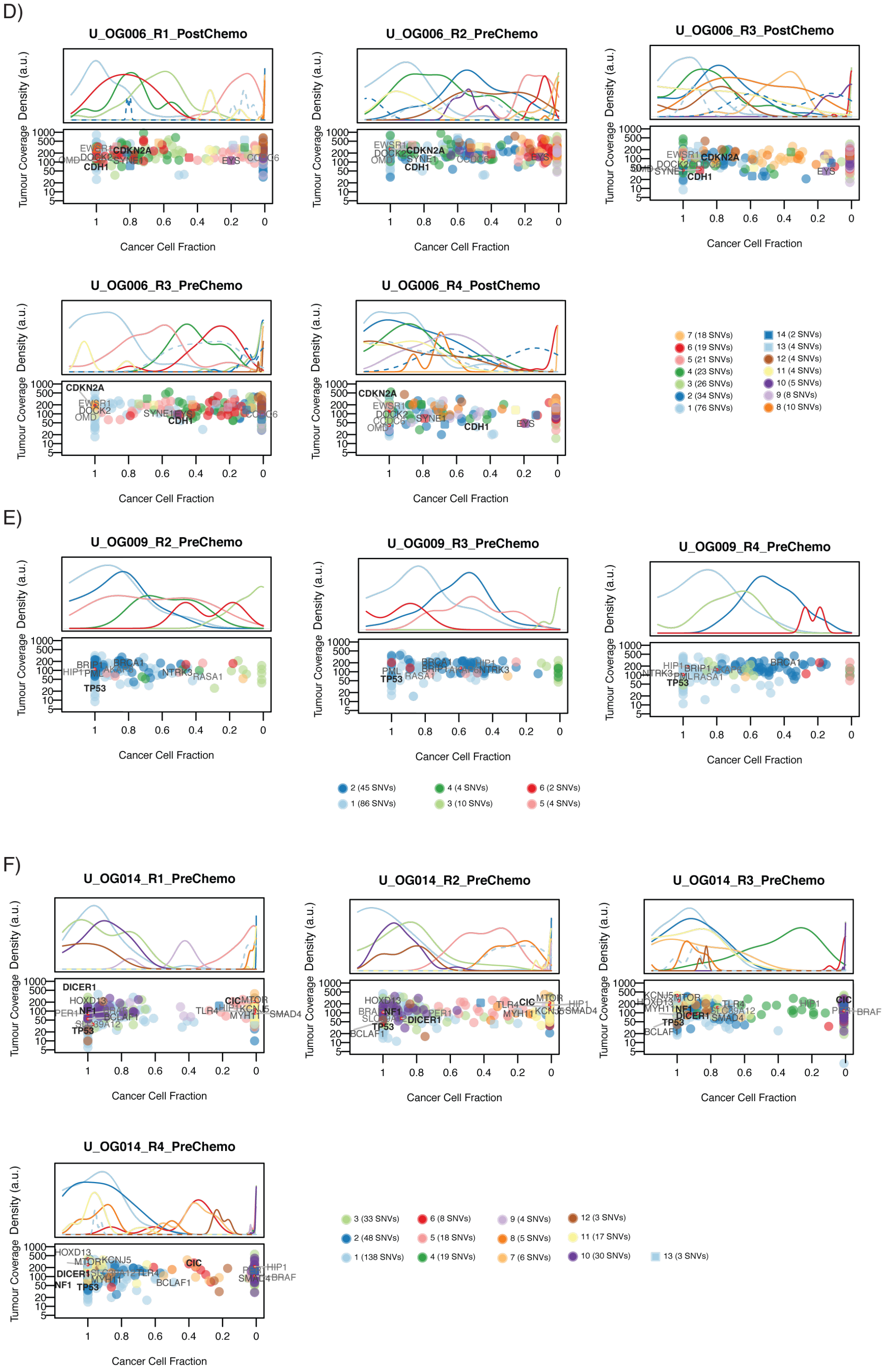
Figure 3:19 Intra-tumour heterogeneity and response in oesophageal adenocarcinoma

Relationship of intra-tumour heterogeneity and response to neo-adjuvant chemotherapy treatment. Spearman rho is indicated.

3.2.14 Clonal architecture pre and post chemotherapy treatment

Next, to further resolve the clonal architecture of mutations both pre and post-chemotherapy treatment, the cancer cell fraction of each mutation was determined and mutations were clustered using a multi-dimensional Dirichlet process (Figure 3:20). As in NSCLC, the cancer cell fraction was determined by integrating copy number and purity estimates.


A)
U_OG001_R1_PostChemo
U_OG001_R1_PreChemo
U_OG001_R2_PostChemo
U_OG001_R2_PreChemo
U_OG001_R3_PreChemo
U_OG001_R4_PreChemo
Density (a.u.)
Tumour Coverage
1000
500
200
100
50
20
10
5
1
0.8
0.6
0.4
0.2
0
Cancer Cell Fraction
TP53
EXT2
NTRK3
NUAK1
HIST1H4I
SLC39A12
GNPTAB
KLF4
EGFR
TPR
3 (36 SNVs)
2 (64 SNVs)
1 (103 SNVs)
6 (26 SNVs)
5 (28 SNVs)
4 (31 SNVs)
9 (12 SNVs)
8 (7 SNVs)
7 (10 SNVs)
12 (2 SNVs)
11 (4 SNVs)
10 (4 SNVs)
B)
U_OG003_R1_PreChemo
U_OG003_R2_PreChemo
U_OG003_R4_PreChemo
MAML2
SETBP1
CBFA2T3
GAS7
PALB2
NUAK1
1 (109 SNVs)
2 (35 SNVs)
3 (7 SNVs)
4 (5 SNVs)
5 (2 SNVs)
C)
U_OG005_R1_PostChemo
U_OG005_R2_PostChemo
U_OG005_R4_PostChemo
NKX2−1
TP53
KIT
KAT6A
ACSL3
2 (12 SNVs)
1 (102 SNVs)
4 (9 SNVs)
3 (11 SNVs)
6 (4 SNVs)
5 (8 SNVs)


D)
U_OG006_R1_PostChemo
U_OG006_R2_PreChemo
U_OG006_R3_PostChemo
U_OG006_R3_PreChemo
U_OG006_R4_PostChemo
Density (a.u.)
Tumour Coverage
1000 500 200 100 50 20 10 5
Cancer Cell Fraction
1 0.8 0.6 0.4 0.2 0
CDKN2A
CDH1
EWSR1
DOCK2
OMD
SYNE1
EYS
CCDC6
7 (18 SNVs)
6 (19 SNVs)
5 (21 SNVs)
4 (23 SNVs)
3 (26 SNVs)
2 (34 SNVs)
1 (76 SNVs)
14 (2 SNVs)
13 (4 SNVs)
12 (4 SNVs)
11 (4 SNVs)
10 (5 SNVs)
9 (8 SNVs)
8 (10 SNVs)
E)
U_OG009_R2_PreChemo
U_OG009_R3_PreChemo
U_OG009_R4_PreChemo
TP53
BRIP1
BRCA1
HIP1
PML
NTRK3
RASA1
2 (45 SNVs)
1 (86 SNVs)
4 (4 SNVs)
3 (10 SNVs)
6 (2 SNVs)
5 (4 SNVs)
F)
U_OG014_R1_PreChemo
U_OG014_R2_PreChemo
U_OG014_R3_PreChemo
U_OG014_R4_PreChemo
DICER1
HOXD13
NF1
TP53
PER1
SLC39A12
BCLAF1
CIC
MTOR
TLR4
HIP1
KCNJ5
MYH11
SMAD4
BRAF
3 (33 SNVs)
2 (48 SNVs)
1 (138 SNVs)
6 (8 SNVs)
5 (18 SNVs)
4 (19 SNVs)
9 (4 SNVs)
8 (5 SNVs)
7 (6 SNVs)
12 (3 SNVs)
11 (17 SNVs)
10 (30 SNVs)
13 (3 SNVs)

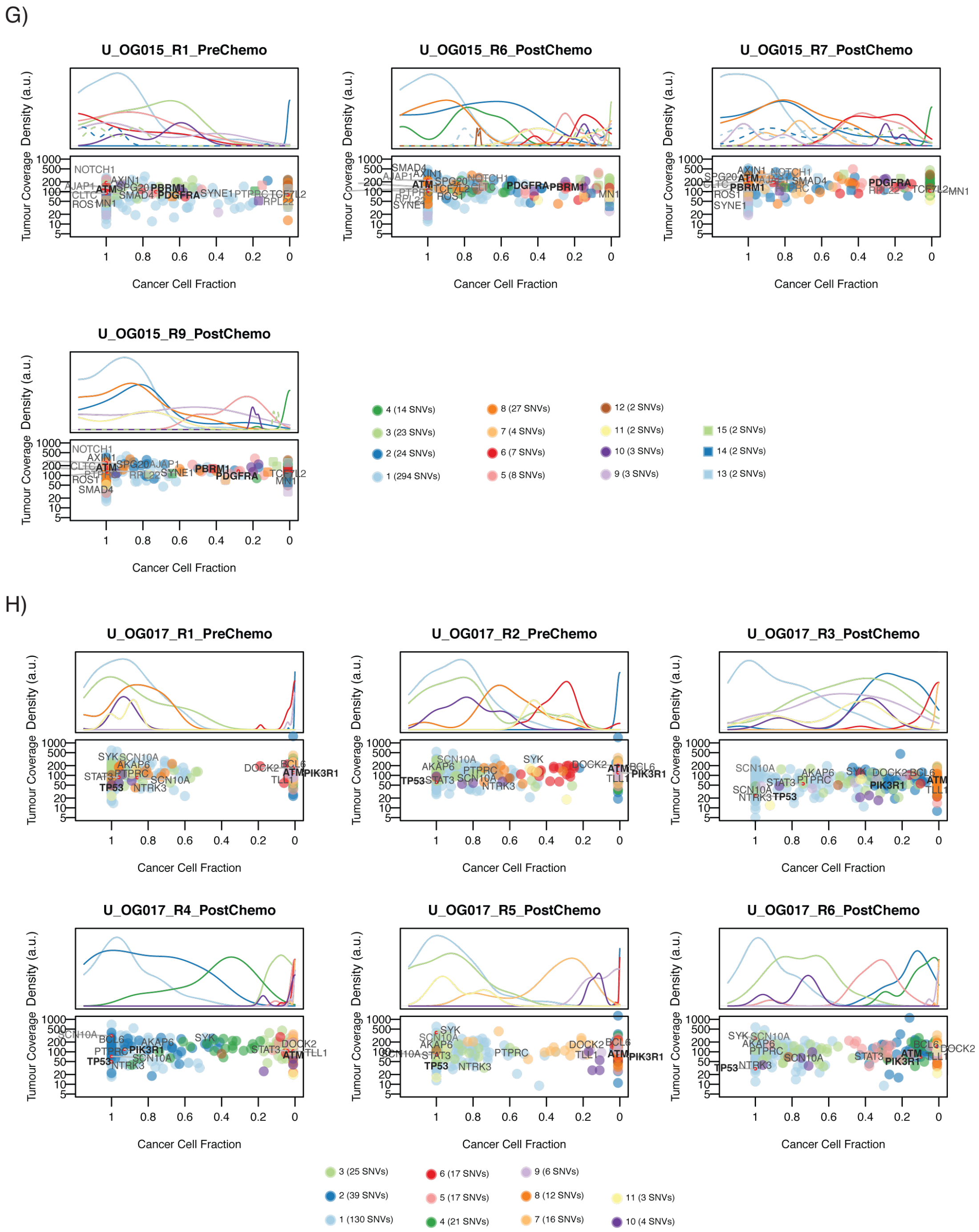


Figure 3:20 Clonal Architecture of oesophageal adenocarcinoma

For each tumour region within each tumour, the clonal composition is shown, with the cancer cell fraction of each mutation indicated. Mutation clusters, as determined by a multidimensional Dirichlet clustering process are depicted. Category 1-3 mutations in cancer genes are indicated.

Analysis of the subclonal architecture of oesophageal adenocarcinoma revealed a multitude of subclones in each tumour, exceeding the number of regions sequenced in every case (Figure 3:20). The number of subclones identified was directly related to the number of tumour regions sequenced, highlighting the necessity of multi-region to resolve the clonal architecture of tumours. Indeed, many mutation clusters identified exhibited near-identical cancer cell fractions within the same tumour region, rendering their de-convolution intractable without multi-region sequencing.

A fully clonal cluster was identified in every tumour, reflecting the mutations present in the most recent common ancestor, and the founding cancer clone. On average, this mutation cluster represented 51% of mutations in each tumour (range 28-69%). In general, the majority of mutation clusters present prior to adjuvant chemotherapy were found to be present in 100% of tumour cells within at least one tumour region, while sometimes absent or present in only subset of tumour cells within other tumour regions (Figure 3:21). Such clusters may reflect a clonal sweep within one tumour region. Mutation clusters identified exclusively after treatment with platinum chemotherapy were often subclonal, occurring in less than 100% of tumour cells within every tumour region, suggesting platinum chemotherapy may foster additional subclonal diversity. Further samples will be needed to confirm this observation, and it is difficult to assess fully without a control group that has also been serially sampled.

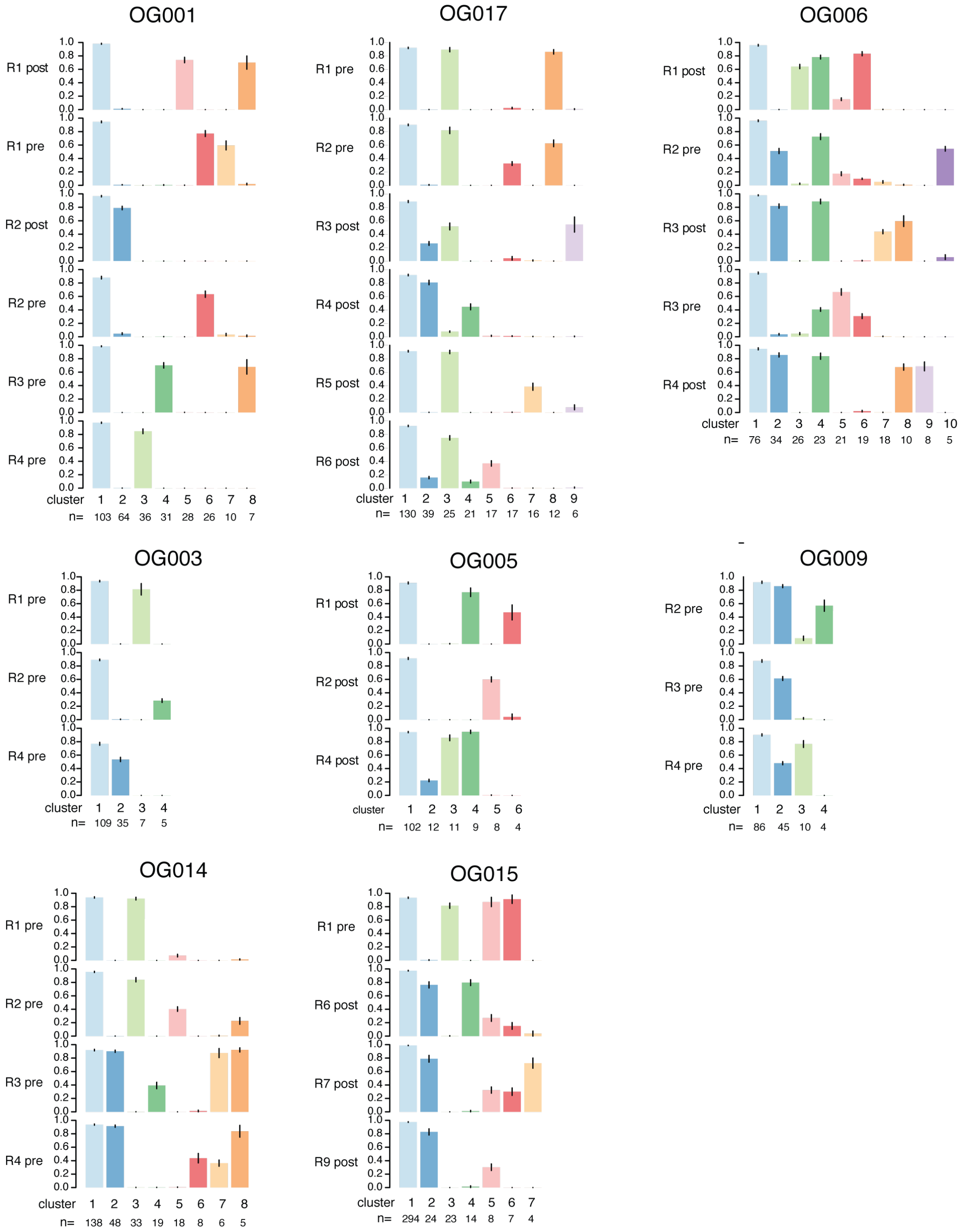


Figure 3:21 Clonal and subclonal clusters in tumour regions

For each tumour region, the median cancer cell fraction of all mutations within each cluster identified is shown as a barplot. Number of mutations within each cluster is shown under plot. Clusters were heterogeneity may be driven by copy number loss were removed from analysis.

3.2.15 Phylogenetic trees for oesophageal adenocarcinoma

In order to understand the evolutionary relationship between tumour subclones, and to further explore how chemotherapy may influence the subclonal architecture and tumour diversity, phylogenetic relationships between clonal and subclonal mutation clusters were investigated.

First, for each mutation cluster identified, the possibility of copy number driven heterogeneity was determined (see Experimental procedures). In total the heterogeneity of 100 mutations was identified as potentially driven by copy number diversity. For instance, in OG015 a chromosomal segment harbouring 16 mutations was identified as deleted in tumour region R1, but present in all other tumour regions.

Next, two types of phylogenetic trees were constructed, region trees, where each region was treated as a separate species and subclonal architecture trees, where the subclonal architecture of each tumour region was taken into account. To determine the subclonal evolutionary history of each tumour, the pigeonhole principle was adopted, such that any two branches of the evolutionary tree when summed could not exceed 100% of tumour cells and ancestral clones were necessarily at a higher or equal cancer cell fraction compared to their descendants (see Experimental procedures).

Every oesophageal adenocarcinoma exhibited branched tumour evolution. In two of eight tumours, parallel evolution was observed. In OG005, multiple distinct mutations in *NOTCH1* were observed on separated tumour branches and in OG001 distinct mutations in *GNPTAB* were identified in separate tumour regions.

Two distinct modes of tumour evolution were observed following chemotherapy treatment. In two tumours (OG005 and OG015) post-chemotherapy regions clustered separately as a monophyletic group (indicating a shared common ancestor). However, in the remaining tumours, paraphyly was observed whereby pre- and post- chemotherapy regions clustered together. Interestingly, in this small cohort, tumours OG005 and OG015, which had a poor response to chemotherapy (Table 3-2), exhibited monophyly with clustering of post-chemotherapy tumour regions from a shared common ancestor, consistent with a common genomic event conferring resistance to therapy in these tumours.

In six of eight subclonal architecture trees, at least one tumour region was found to harbour subclones from different nodes of the phylogenetic tree, highlighting the fact that these tumours exhibit a complex subclonal structure, and subclones can be shared between different tumour regions of oesophageal adenocarcinoma tumours (Figure 3:23).

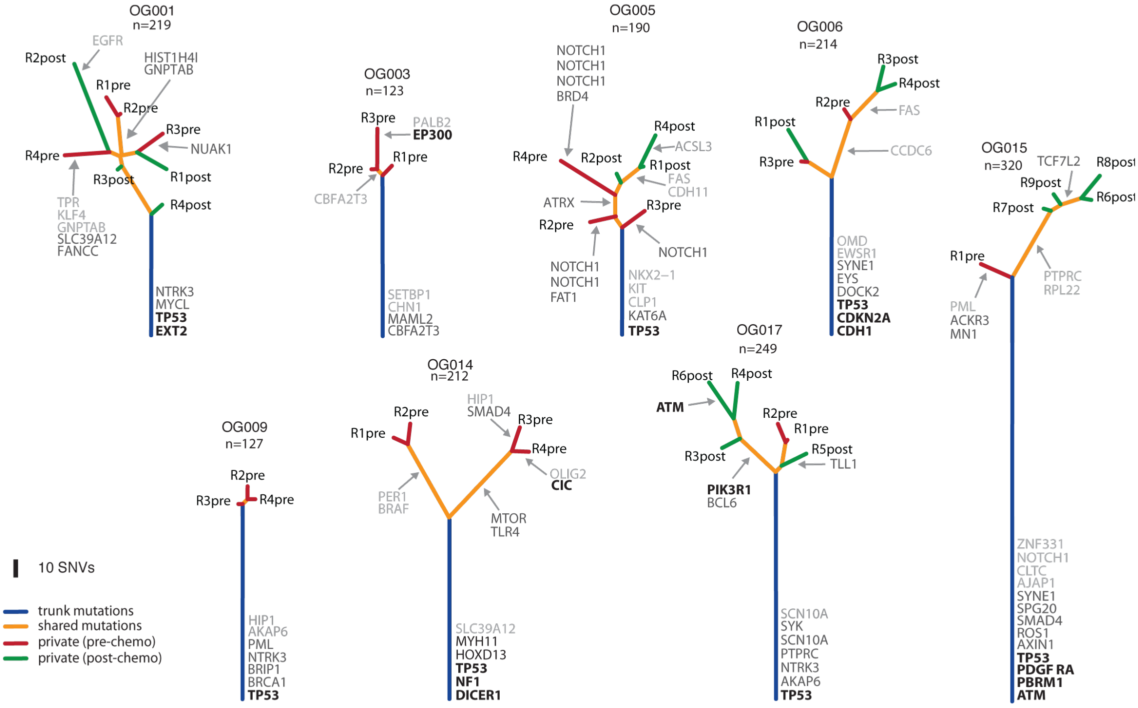


Figure 3:22 Oesophageal tumour region phylogenetic trees

Phylogenetic trees generated by a maximum parsimony approach based on the distribution of all detected mutations. Trunk and branch lengths are proportional to the number of mutations acquired. Categories 1 -3 driver mutations are indicated next to the trunk or with an arrow pointing to the branches where they were acquired

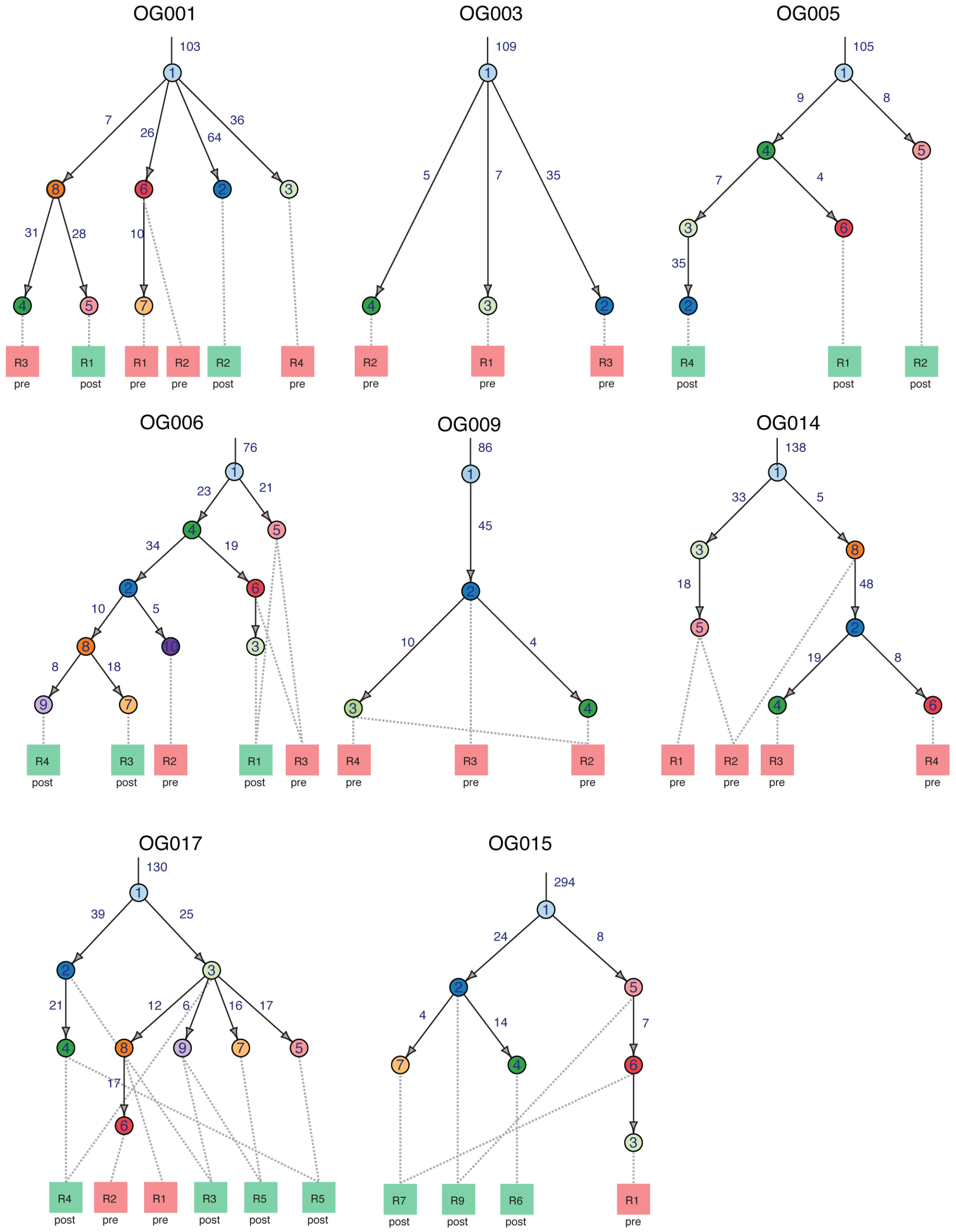


Figure 3:23 Subclonal phylogeny in oesophageal adenocarcinoma

Phylogenetic trees based on subclones. Relationships between subclones are computed based on pigeonhole principle. Tumour regions are depicted beneath tree, with lines indicating presence of subclone within tumour region. Number of SNVs is indicated.

3.3 Conclusions

The results of this chapter demonstrate the presence of considerable intra-tumour across cancer types, and provide further evidence for branched tumour evolution, involving competing or cooperating subclones. Indeed, in both NSCLC and oesophageal adenocarcinoma, using multi-region sequencing, branched tumour evolution was observed.

Subclonal driver mutations including known canonical hotspot mutations, such as *IDH1* (R132H) and *PIK3CA* (E545K) were identified in every cancer type investigated. Further, almost every gene linked with a targeted therapy was found to harbour subclonal mutations in at least one tumour within the cohort. Moreover, certain pathways were found to be more likely to harbour subclonal mutations than others, and evidence for parallel evolution was identified, suggesting constraints to tumour evolution that may one day be therapeutically exploitable.

Analysis of paired pre- and post- neo-adjuvant platinum chemotherapy oesophageal adenocarcinoma tumours suggested intra-tumour heterogeneity might facilitate development of resistance to therapy. However, additional samples are required to confirm this observation. In addition, mutations detected post-chemotherapy tended to be subclonal in all tumour regions.

Nevertheless, despite extensive heterogeneity, the results presented here also demonstrate that mutations in established driver genes have a tendency to be clonal compared to mutations in non-driver genes, suggesting that these mutational events may often be required as early events in tumourigenesis and might represent suitable candidates for cancer screening approaches (Martinez et al., 2013). The enrichment for clonal mutations in established driver genes is likely related to the fact that current methods for detecting driver genes are underpowered to detect subclonal drivers of tumour subclade expansions, present at low variant allele frequencies or spatially or temporally separated within tumours. Consistent with this, by focusing on later mutations, it was possible to identify cancer genes that may be responsible for subclonal clade expansions. Indeed, genes such as *ATXN1* and *CTNNA2* have been linked to tumour metastasis and cell migration (Lee et al., 2011b).

Importantly, comparisons of multi-region and single sample techniques to decipher heterogeneity revealed that sampling bias, which can create an illusion of clonal dominance, can render single sample analysis inaccurate and potentially misleading. However, further work is needed to characterize the optimum number of tumour samples required to accurately de-convolve the subclonal structure of a tumour. Moreover, current methods for deciphering intra-tumour heterogeneity are not without limitations and are heavily influenced by purity and sequencing depth. Future studies, such as TRACERX in NSCLC (Jamal-Hanjani et al., 2014), employing deep-multi-region sequencing on large cohorts of tumour may be required to gain a more complete picture of the genomic and clonal landscape of tumours.

In summary, the evidence presented here highlights the need to understand the subclonal composition of tumours. Moreover, the results demonstrate that known driver mutations not only play a role in tumour initiation, but also likely influence tumour behaviour after tumour branching within distinct subclones. The next two chapters further explore the mechanisms generating this intra-tumour heterogeneity.

Chapter 4. Genome doubling in cancer evolution

4.1 Introduction

Chromosomal instability contributes to intra-tumour heterogeneity by creating a genomically diverse pool of tumour cells upon which selection can act. Consistent with providing a substrate for tumour evolution, CIN is associated with both poor prognosis and intrinsic multi-drug resistance across a wide range of cancers (McGranahan et al., 2012, Lee et al., 2011a, Duesberg et al., 2001).

Another common feature of tumour cells is an abnormal chromosomal content. Polyploid cells have been observed in multiple cancer types, and genome doubling events have been proposed as an intermediate state en route to chromosomal instability (Carter et al., 2012, Davoli and de Lange, 2011, Shackney et al., 1989). However, the relationship between genome doubling and CIN, and the effect of a whole genome-doubling event on cancer genome evolution remains unclear.

In this chapter, I explore the timing and impact of a whole genome-doubling event on genomic instability and cancer genome evolution. I make use of SNP6.0 copy number and mutation data from 10 major cancer types. In addition an isogenic colorectal cell-line system is used to explore the acute affects of a whole genome-doubling event. Finally, I investigate the clinical implications of a genome doubling in colorectal cancer.

The majority of the work presented in this chapter forms the basis of a publication in Cancer Discovery (see Appendix 2, Dewhurst and McGranahan et al. 2014). I undertook all the bioinformatics work presented in this chapter, while experimental work was performed by Sally Dewhurst.

4.2 Results

4.2.1 Exploring chromosomal instability using copy number data

As discussed in the introduction to this thesis, CIN represents a dynamic state in which cells gain or lose whole chromosomes, or parts of chromosomes, at an elevated rate, creating intercellular genetic heterogeneity. Allele specific copy number data can shed light on the average copy number state of a population of tumour cells, revealing the level of aneuploidy in a tumour population. However, given that the copy number state of a tumour genome does not necessarily reflect the underlying rate of copy number loss/gain, it is necessary to infer proxies of chromosomal instability from copy number data.

A simple method to capture a gross level of instability, first proposed by Chin et al. (2007), involves measuring the extent of the genome that is subject to copy number change. This measure, termed genome instability index (GII), can broadly distinguish unstable from stable genomes. However, a drawback of this method is that it ignores the ploidy of a sample, and, moreover, is biased by the fact that larger chromosomes will necessarily contribute more to the score than smaller chromosomes.

The wGII, the weighted genome integrity index, builds directly upon the GII by measuring gains and losses relative to the ploidy of a simple, and further, is weighted to ensure each chromosome contributes equally to the score, and thereby losses or gains affecting longer chromosomes will not bias the score (Burrell et al., 2013a).

4.2.2 Genome doubling can be inferred from allele specific copy number

A tetraploid genome may result from successive gains of individual chromosomes and chromosome arms, or from a genome-doubling event. In order to distinguish between the two and identify genome-doubling events, allele specific copy-number information can be used. Starting from a diploid genome, a genome doubling will initially result in both major and minor alleles (i.e. the two homologous chromosome segments at each locus) being even numbers. Following the genome-doubling event, specific chromosomes, or chromosome segments, may be lost, but the major allele would be expected to have a copy number ≥2 throughout the majority of the genome. Thus, the extent to which a tumour sample's genome exhibits a major allele >=2 across multiple chromosome arms can be used to assess the probability of genome doubling (see Experimental procedures for full details).

4.2.3 A relationship between ploidy and genomic complexity in COAD

To explore the relationship between ploidy, CIN and genome doubling across a range of cancer types, including COAD, allele specific copy number was used.

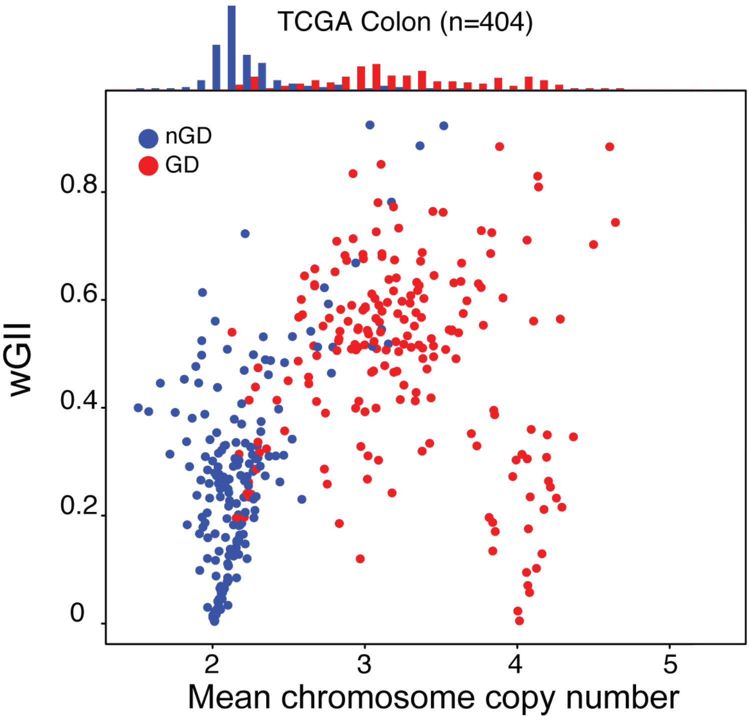


Figure 4:1 Genome doubling and chromosomal instability

Relationship between weighted mean chromosome copy number and wGII. Each circle represents one TCGA COAD tumour sample. Red depicts genome-doubled (GD) samples; blue non–genome-doubled (nGD) samples (see Experimental Procedures). A histogram of weighted mean chromosome copy number for GD (red) and nGD (blue) is shown on top.

In COAD tumour samples, significantly higher wGIIs were found in polyploid (ploidy ≥3) compared to diploid tumours (*P*<0.0001, Student's t-test, Figure 4:1), although some stable (wGII <0.2) tetraploid tumours were also observed. Furthermore, significantly higher wGIIs were observed in tumours classified as genome-doubled compared to non genome-doubled (*P*<0.0001, Student's t-test, Figure 4:1), suggesting a potential relationship between genome doubling and genomic complexity.

The majority of tumours with a triploid karyotype appeared to have undergone a genome-doubling event (105/110 tumours), consistent with a genome-doubling event occurring en route to further genomic change (Carter et al., 2012, Davoli and de Lange, 2011, Shackney et al., 1989). However, near-diploid tumours with high wGIIs and no evidence of genome doubling were also observed within the cohort, suggesting that there exist multiple routes to an unstable genome in COAD tumour evolution. A small number of relatively stable genome-doubled tetraploid tumours were also observed; potentially indicating genome doubling may be an early event in COAD evolution.

To explore whether these results were consistent across cancer types, both the prevalence of genome doubling and its association with the wGII was explored in a further 9 cancer types (Figure 4:2). The prevalence of genome doubling was found to vary considerably across cancers, with almost 80% of oesophageal carcinomas exhibiting evidence for genome doubling, in contrast to only 23% of GBM tumours. In all cancer types, wGIIs were found to be significantly higher in genome-doubled compared to non-genome-doubled tumours (*P*<0.001,all cancer types, Figure 4:2).

Taken together, these results suggest genome doubling and chromosomal instability may be linked in a range of cancer types.

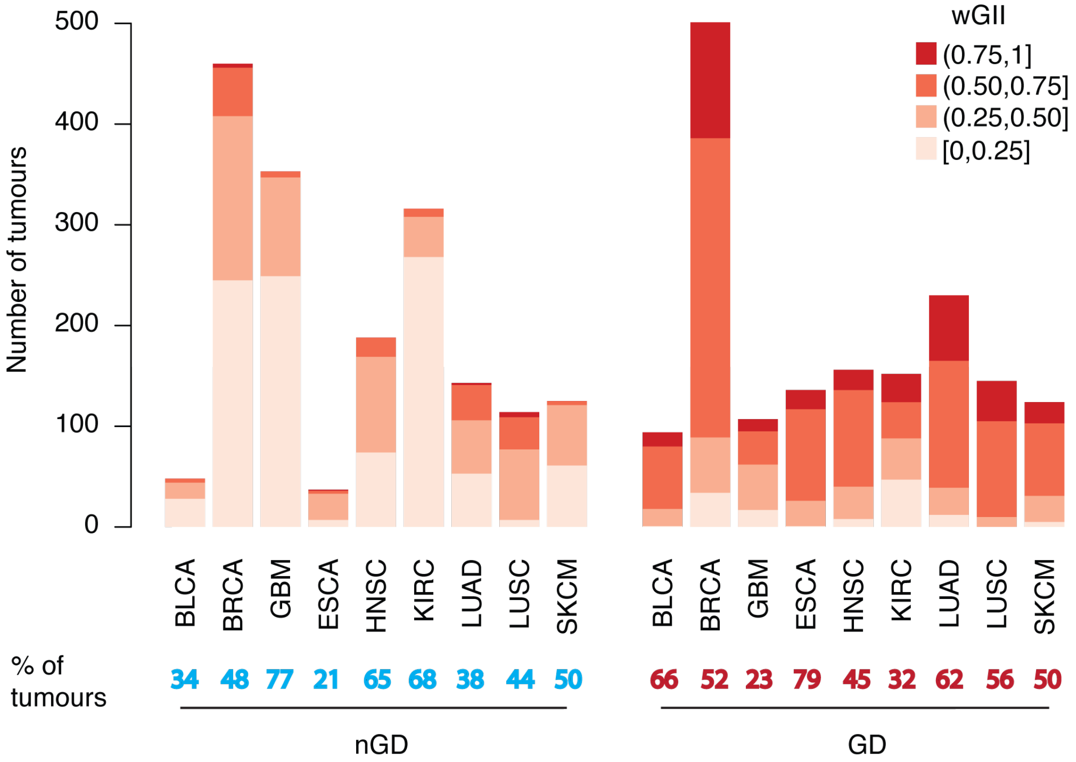


Figure 4:2 Genome doubling across cancer types

wGIIs for genome-doubled (GD) and non-genome-doubled (nGD) tumours is shown for 9 different cancer types. In every case, wGIIs are significantly higher in GD tumours compared to nGD tumours.

4.2.4 Genome doubling is an early event in the majority of COADs

To determine the timing of genome doubling in colorectal cancer evolution, the relative abundance of different copy number states can be used. Copy number losses occurring on the background of a diploid genome will necessarily result in loss of heterozygosity (LOH), as either the maternal or paternal allele will be lost. This LOH will leave a permanent footprint in the genome, persisting after a genome-doubling event and will be indelibly imprinted for the remainder of the life history of the tumour (Figure 4:3-2 A i). By contrast, losses occurring after genome doubling are less likely to exhibit LOH (Figure 4:3-2 A ii). The types of losses in a genome-doubled sample may thus shed light on the relative timing of a genome-doubling event in cancer evolution.

In the majority of genome-doubled TCGA COAD tumours, genome doubling likely occurred as a relatively early event, prior to the majority of copy number losses (Fig. 1B). In almost two thirds (130/196) of genome-doubled tumour, a higher fraction of the genome was found to be present at heterozygous diploid compared to homozygous diploid. Further, in 20% of these samples, the proportion of the genome exhibiting LOH was less than is typical for stable tumours exhibiting microsatellite instability (MSI).

These data not only suggest that genome doubling is frequently an early event in colorectal tumours, but importantly, also imply that CIN often occurs after genome doubling *in vivo*.

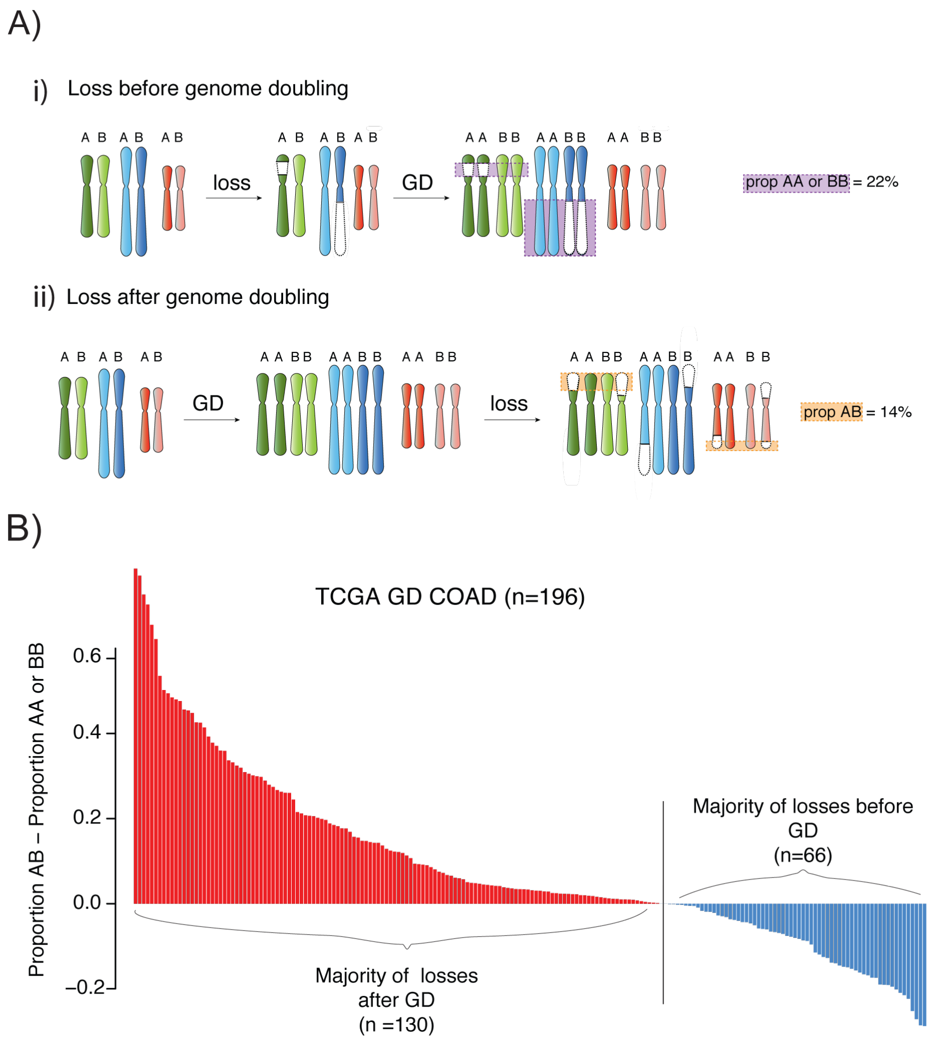


Figure 4:3 Timing genome doubling

A) i) Copy number losses that occur on the background of a diploid genome, before genome doubling, will result in LOH. In tumours that harboured CIN before doubling, the majority of losses will be unbalanced, involving LOH. Unbalanced losses to two copies (AA or BB) are depicted with a purple box with purple dotted lines. ii), in tumours where genome doubling was an early event, before the onset of CIN, the majority of losses to two copies will be balanced without LOH. Balanced losses to two copies (AB) are depicted with an orange box with orange dotted lines around them. B) Timing of genome doubling estimated using copy number and LOH profiles. Each bar represents one genome-doubled TCGA tumour and its height corresponds to the proportion of AB − proportion AA or BB copy number states. Tumour genomes in which the majority of losses to two copies are likely to have occurred after genome doubling are shown in red (n = 130; proportion AB > proportion AA or BB), and those where the majority of losses are likely to have occurred before doubling are shown in blue (n = 66; proportion AB < proportion AA or BB).

4.2.5 Timing of genome doubling varies across cancer types

Applying the same reasoning to a further nine cancer types revealed that the timing of genome doubling is variable across cancer types. Interestingly, in BRCA, ESCA, LUSC and SKCM, there was a highly significant trend for the majority of deletions in genome-doubled tumours having occurred before a genome doubling, suggesting genome doubling may be a later event in these cancer types, occurring after the onset of CIN (Figure 4:4, *P*<0.001, paired Wilcoxon test). Conversely, in KIRC, like COAD, there was a significant trend for genome doubling occurring as a relatively early event (Figure 4:4).

Taken together, these data suggest divergent patterns of genome doubling both within and across tumour types.

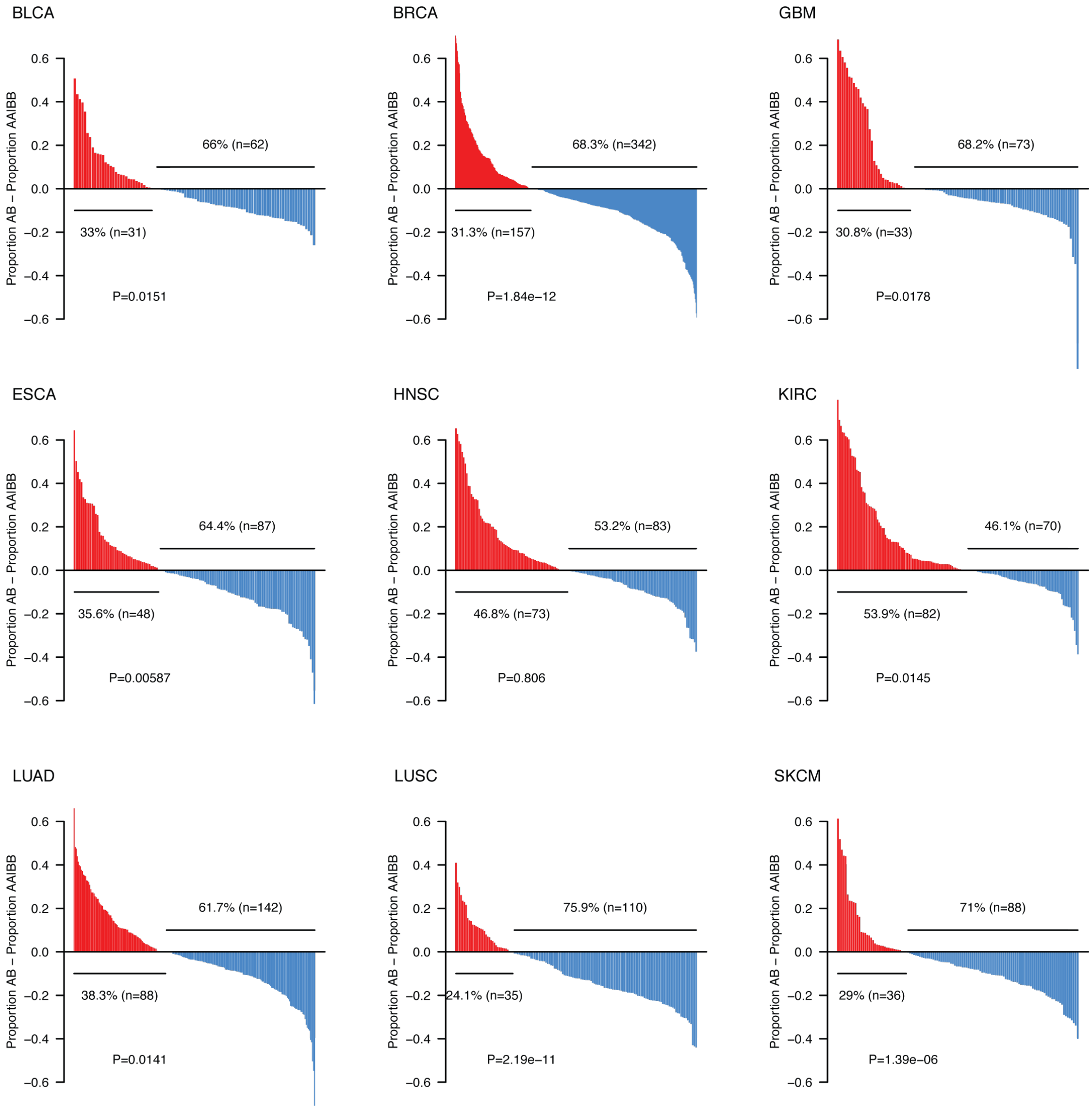


Figure 4:4 Timing of genome doubling across cancer types

Timing of genome doubling across 9 cancer types. In each cancer type, the proportion of the genome that shows balanced losses to diploid (AB) is compared to unbalanced (BB or AA). P-values are indicated, reflecting the results of a paired Wilcoxon test.

4.2.6 Genome doubling often occurs prior to the last clonal sweep at SNV level

Using multi-region sequencing data from oesophageal and NSCLC (Table 3-1 and Table 3-2), it was possible to gain further resolution regarding the timing of genome doubling. Genome doubling events were observed in all 8 oesophageal adenocarcinomas, in keeping with its high prevalence in TCGA data. In every case, the genome-doubling event was shared in all tumour regions, indicating it likely occurred on the trunk of the phylogenetic tree, prior to last clonal sweep and before the emergence of the most recent common ancestor. In addition, extensive LOH was observed in every tumour, suggesting the doubling may have occurred after the onset of chromosomal instability. These data corroborate the findings discussed above using TCGA copy number data, indicating genome doubling in oesophageal cancers often occurs after the onset of CIN.

In the multi-region NSCLC cohort, four out of six tumours displayed evidence for whole-genome–doubling events. In three tumours (L001, L004, and L008), the genome-doubling event was identified in every tumour region, indicating it occurred on the trunk of the tumours' evolutionary tree. In each of these cases, the majority of truncal mutations (84-88%) were present at ploidy >=2, consistent with a large mutational burden preceding genome doubling. Further, in these three tumours, all truncal driver mutations were found to exhibit mutation copy numbers >=2, suggesting these were present prior to doubling.

These data are consistent with genome doubling being a relatively late event in NSCLC. In further support of this, in L002 the majority of heterogeneous mutations were found to be present at ploidy >=2, indicative of two independent genome-doubling events: one in the LUAD region, region R1, and one in the LUSC region, region R3 (Figure 4:5). Additional NSCLC multi-region sequencing data will be required to decipher the frequency of multiple independent genome doubling events during tumour evolution in this cancer type.

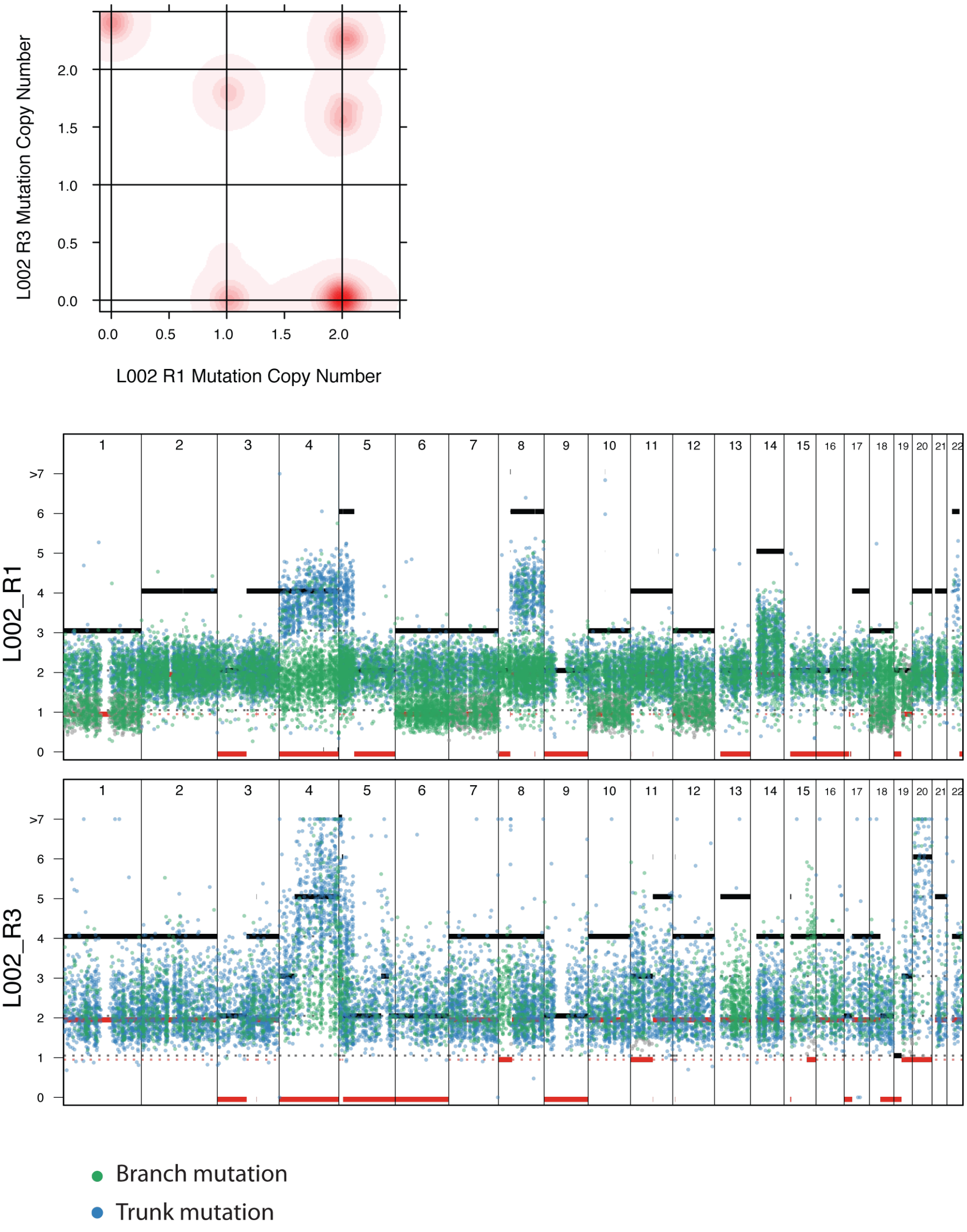


Figure 4:5 Genome doubling in L002

2D-dirichlet process plot (upper) using mutation copy numbers; increasing intensity of red indicates the location of a high posterior probability of a cluster. Private mutations clustered at a copy number 2 are observed in both regions R1 and R3, suggesting two independent genome-doubling events. Copy number profiles of L002 regions R1 and R3 (lower). Trunk mutations are shown in blue, whilst private mutations are depicted in green. Total copy number is depicted as a black line, with minor allele as a red line.

In combination, these data suggest genome doubling is a common event across human cancers and that polyploid tumours often arise from a genome-doubling event, rather than through sequential gains of individual chromosomes. While, in COAD and KIRC tumours genome doubling appears to be a relative early event, potentially preceding CIN, in the majority of other cancer types, genome doubling appears to occur after the onset of chromosomal instability.

4.2.7 Enrichment for mutations in *TP53* in genome-doubled tumours

Given the established associated between disruption of the tumour suppressor gene *TP53* and chromosomal instability (Giam and Rancati, 2015), the relationship between genome doubling and mutations in *TP53* was next explored.

Across the pan-cancer cohort, a highly significant enrichment of *TP53* mutations was observed in genome-doubled compared to non genome-doubled tumours (P<0.0001, Fisher's exact test). Further, almost invariably (>90% cases), these mutations occurred prior to doubling, significantly more than the background rate, consistent with TP53 playing an important role in tolerance of genome doubling (Figure 4:6).

Nevertheless, despite this strong association, a substantial majority of genome-doubled tumours (831/1270) did not exhibit somatic mutation disruption to *TP53*, suggesting there are likely multiple routes to tolerance of genome doubling.

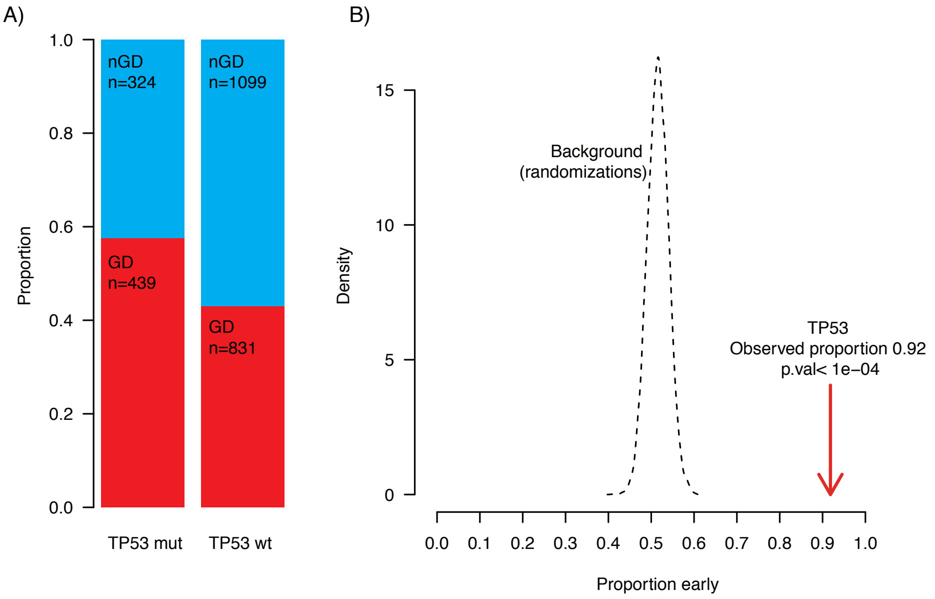


Figure 4:6 *TP53* and genome doubling

A**)** Association between mutations in *TP53* and prevalence of genome doubling. B) Timing of mutations in *TP53* compared to background randomizations. 92% of mutations in *TP53* are classified as early, significantly more than expected given random mutations.

4.2.8 An isogenic cell line system to study the effects of genome doubling

To explore the impact of genome doubling on genome stability in colorectal cancer, an isogenic, p53 wild-type cell line system was used. The experimental work described in this section of the thesis was undertaken by Sally M. Dewhurst.

HCT-116, a diploid micro-satellite unstable COAD cell line, was found to have a small sub-population (<2%) with >4N DNA content (Figure 4:7 A), reflecting a genome-doubled population. Single tetraploid (genome-doubled) and diploid (non genome-doubled) cells were isolated by flow cytometry. The cloning efficiency of genome-doubled cells was lower than non-diploid cells, suggesting a genome doubling event is poorly tolerated in HCT-116 cells under standard culture conditions (2N= 63%, 4N= 6%, Figure 4:7 B),

One diploid clone (nGD 8) and two rare surviving genome-doubled, tetraploid, clones (GD 3, GD 4, Figure 4:7 C), were expanded in culture. It was also possible to isolate a second generation of genome-doubled clones (GD-13, GD-16, GD-17, GD-35) as well as two diploid, non-genome-doubled, clones (nGD-14, nGD-25) from the diploid clone nGD 8 (Figure 4:7 D). Thus, all second-generation clones shared a common ancestor (nGD 8), and had arisen spontaneously within the time frame of the experiment. The ancestry of genome-doubled and non-doubled clones can be seen in Figure 4:7 E.

To verify that any observed differences between genome-doubled and non-genome-doubled clones were not impacted upon by MSI status, an MLH1 competent clone, HCT-116_MLH, was also investigated. In this clone, a wild-type *MLH1* allele was reincorporated into the genome, thus effectively rendering the clone micro-satellite stable. Both genome-doubled (MLH-GD11 and MLH-GD16) and non-genome-doubled (MLH-nGD8) clones were isolated from this MLH competent cell-line (Figure 4:7 E).

All genome-doubled clones were found to have a seemingly functional p53 response to DNA damage, and no mutations were found in the coding regions of *TP53* or *CDKN1A* (p21), suggesting that in these clones tolerance to genome doubling cannot be accounted for by direct disruption of p53.

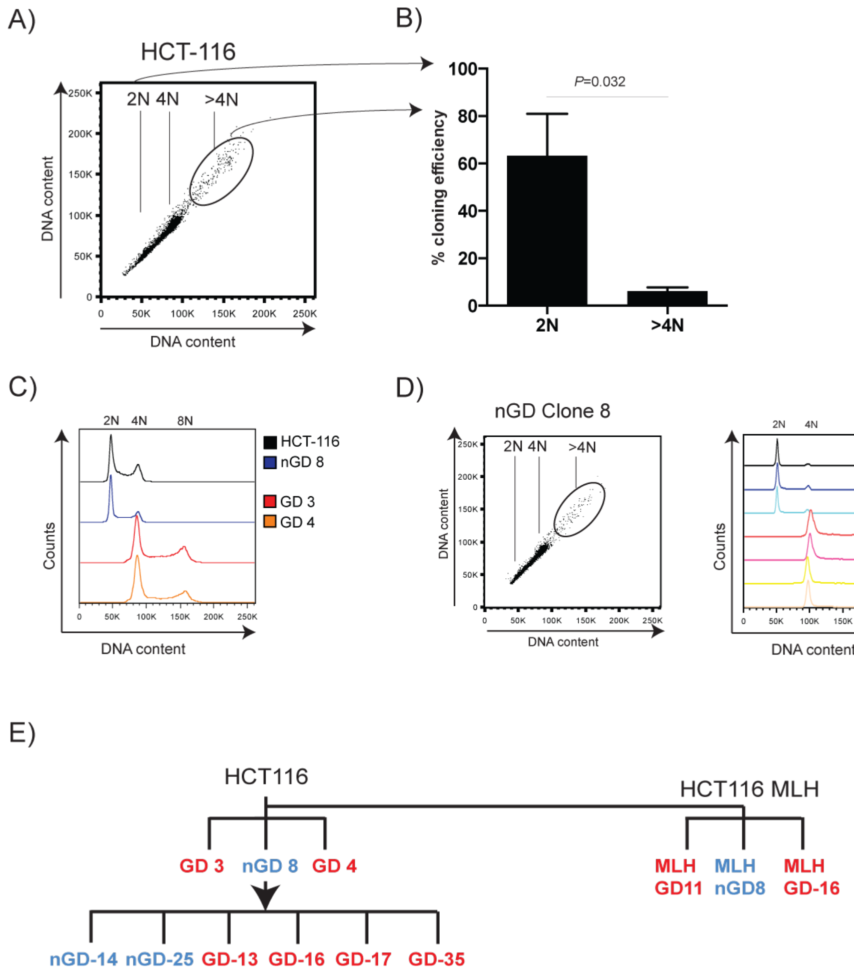


Figure 4:7 Isogenic genome doubling cell-line system

A**)** Flow cytometry shows a **>**4N population in the MSI colon cancer cell line HCT-116, indicative of a genome-doubling event. B) Cloning efficiency of 2N and **>**4N cells is shown with mean and standard error of mean (three experiments). The cloning efficiency of the >4N, genome-doubled, population was significantly lower than cloning efficiency of the diploid, non genome-doubled, population (*P* **=** 0.032, Student's t-test). C) After single-cell sorting, DNA content was assessed by flow cytometry with Hoescht staining. Two genome-doubled clones (GD 3 and GD 4), one non genome-doubled, diploid, clone (nGD 8), as well as HCT-116 are depicted. D) Flow cytometry of the non genome-doubled clone nGD 8 also shows a small **>**4N subpopulation. Two further non-genome-doubled, diploid, clones (nGD-14 and nGD-25) and four genome-doubled clones (GD-13, GD-16, GD-17, and GD-35) were isolated from nGD 8, and their DNA content, by flow cytometry with Hoescht staining. E) A family tree depicting the relationships between genome-doubled and non genome-doubled clones. GD= genome doubled; nGD= non genome doubled.

4.2.9 Chromosomal instability in genome-doubled clones

To explore the level of chromosomal instability in both genome-doubled and non-doubled clones data from experiments conducted by Sally Dewhurst were analysed.

A measure of numerical instability can be obtained by assessing the percentage of cells deviating from the modal chromosome number within individual cell colonies, using clonal fluorescence *in situ* hybridization (FISH). Within a given colony, a high degree of deviation from the mode is indicative of cell-to-cell variation in chromosome number, and the presence of chromosomal instability during the clonal expansion. By contrast, colonies exhibiting little deviation from the mode, and low cell-to-cell variation, suggest limited chromosomal instability during clonal expansions.

Colonies from genome-doubled clones displayed significantly higher cell-to-cell variation than non-genome-doubled colonies (Figure 4:8 A: passage 5 non genome-doubled mean=7% (0-23%); genome-doubled mean=28% (5-57%), *P*<0.0001; passage 50 non genome-doubled mean=13% (3-34%); genome-doubled mean=33%, (7-68%), *P*<0.0001, Student's t-test). The results remained consistent in genome-doubled clones derived from a microsatellite competent clone of HCT-116.

To further assess the level and extent of chromosomal instability, segregation errors occurring during mitosis were assessed on both a per-cell and a per-chromosome basis using immunofluorescence. Specific segregation errors scored included lagging centric chromosomes, anaphase bridges, lagging acentric chromosome fragments, and joint lagging dicentric chromosomes.

Genome doubled cells exhibited a significantly higher proportion of segregation errors compared to non-genome-doubled cells across three different passages (Figure 4:8 B). However, the segregation error rate on a per chromosome basis was not found to be significantly different between genome-doubled and non-genome-doubled clones. These data suggest the elevated segregation error rate in genome-doubled clones may simply result from the greater number of chromosomes that these cells harbour. In further support of this, the profile of segregation errors in genome-doubled and non-genome-doubled clones was not significantly different.

The prevalence of structural chromosome aberrations in all clones was scored from metaphase chromosome spreads hybridised to an all-centromere fluorescent probe (Figure 4:8 C). Specifically, the presence of di-centric chromosomes, a-centric chromosomes as well as double strand breaks was scored. The number of structural abnormalities per cell was found to be significantly higher in genome-doubled compared to non genome-doubled cells (Figure 4:8 C, non genome-doubled mean: 0.39 (0.26-0.58) abnormalities; genome-doubled mean: 0.93 (0.60-1.62) abnormalities, all passage *P*<0.05, Student’s T test). However, in keeping with the result from scoring segregation errors, on a per chromosome basis the number of structural abnormalities was not significantly different between genome-doubled and non-genome-doubled clones (Figure 4:8, abnormalities per chromosome: diploid mean=0.0088 (0.0058-0.0130); tetraploid mean=0.0108 (0.0073-0.0180), *P*=0.1093, Student’s t-test).

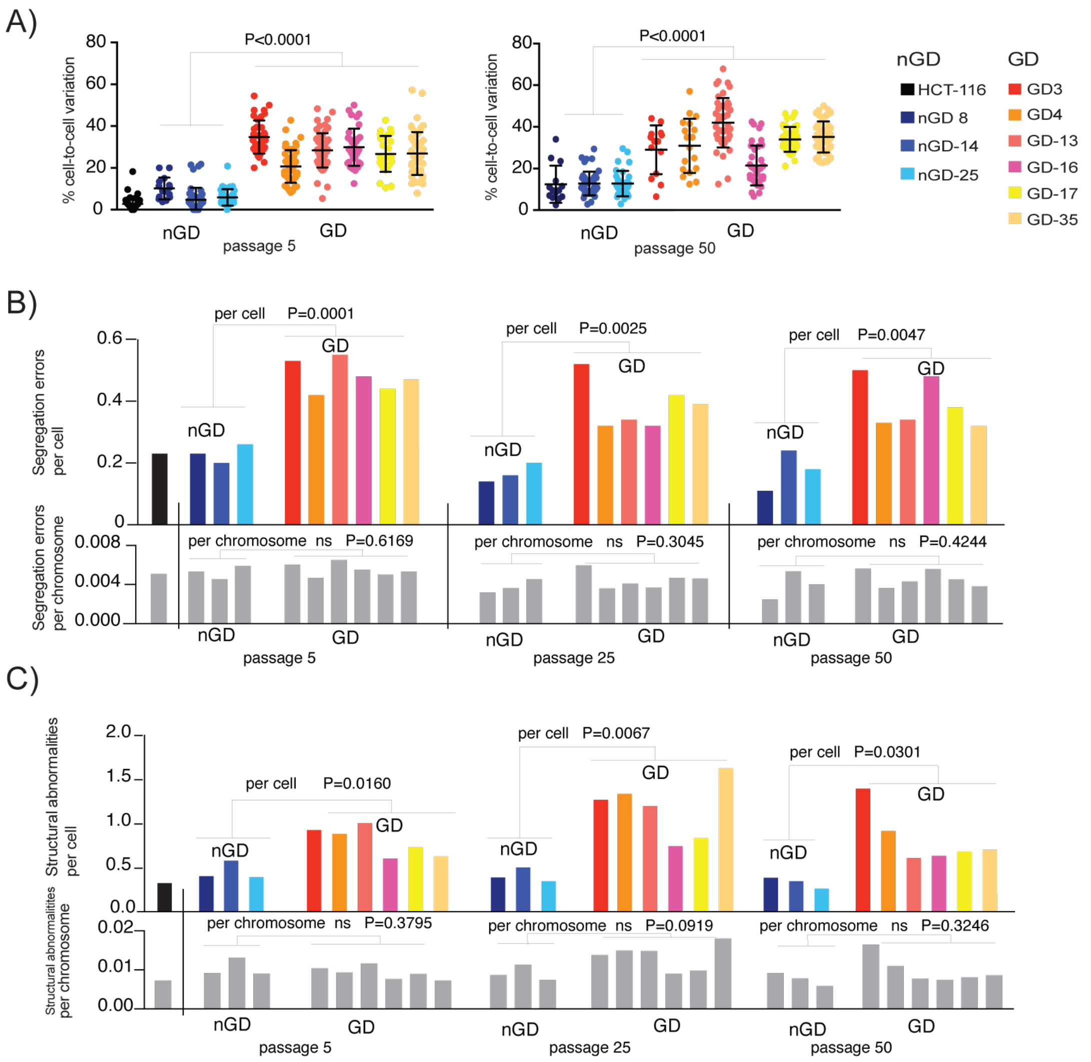


Figure 4:8 Chromosomal instability in genome-doubled clones

A) Cell-to-cell variation in chromosome number. The average percentage deviation of chromosome 2 and 8 is shown for each clone at passages 5 and 50. Each point represents one colony. Median number of cells: at passage 5 = 2,479, passage 50 = 2,105. B**)** Chromosome segregation errors on a per-cell and per-chromosome basis. Errors on a per-cell basis are shown on top graph and on a per-chromosome basis on bottom. C) Structural abnormalities on a per-cell and per-chromosome basis. Abnormalities on a per cell basis are shown on the top, and on per chromosome are shown below. Median number of spreads scored at each passage: passage 5 = 25, passage 25 = 29, passage 50 = 27, and HCT-116 = 37; ns, not significant. GD= genome doubled; nGD non-genome doubled

Finally, to determine whether chromosomal instability was present in genome-doubled clones prior to genome doubling, all clones were subject to SNP6.0 analysis, and the proportion of the genome showing LOH was determined (Figure 4:9). As discussed above, any losses occurring prior to a genome-doubling event will necessarily result in LOH. Thus, if chromosomal instability were present prior to doubling one would expect to see genomic scars in the genome in the form of LOH. However, early passage genome-doubled clones showed limited to no LOH beyond that harboured by the diploid clones (diploid, non genome-doubled, mean: 6.17% (5.96–6.36%) of genome; genome-doubled mean 6.46% (6.16-6.79%), *P*=0.181, Kolmogorov-Smirnov test). These data are in agreement with findings in TCGA COAD tumours, suggesting genome doubling often occurs prior to CIN in this cancer type.

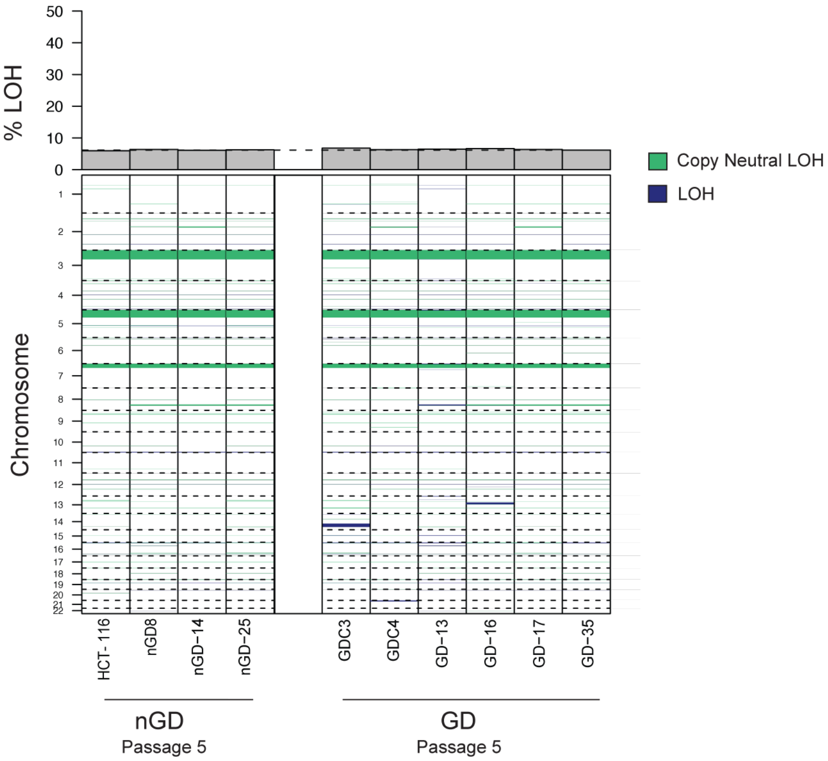


Figure 4:9 LOH in GD and nGD clones

SNP6.0 data illustrating extent of LOH in genome-doubled and non genome double early passage clones. LOH is depicted in purple. Notably, little difference in genomic regions of LOH is seen between genome-doubled and non genome-doubled clones. The barplot above displays the proportion of the genome exhibiting LOH. GD= genome-doubled, nGD = non genome-doubled.

4.2.10 Tolerance of chromosomal instability observed in genome-doubled clones

A chromosomal instability phenotype requires not only a high rate of errors, but also a tolerance of errors, enabling propagation of aneuploidy (Thompson and Compton, 2010). As such, a high rate of both cell-to-cell variation and segregation errors in genome-doubled clones does not equate with tolerance to chromosomal instability. To explore whether genome-doubled clones exhibited enhanced tolerance to chromosomal instability, SNP6.0, clonal FISH experiments and live-cell imaging was performed.

Given that SNP6.0 data provides a population measure of chromosomal content, any aberrations detected by SNP6.0 are suggestive of not only tolerance of errors, but also their propagation. Thus, to assess tolerance and propagation of chromosomal instability, all clones were subject to SNP6.0 analysis at multiple different passages over 18 months in culture.

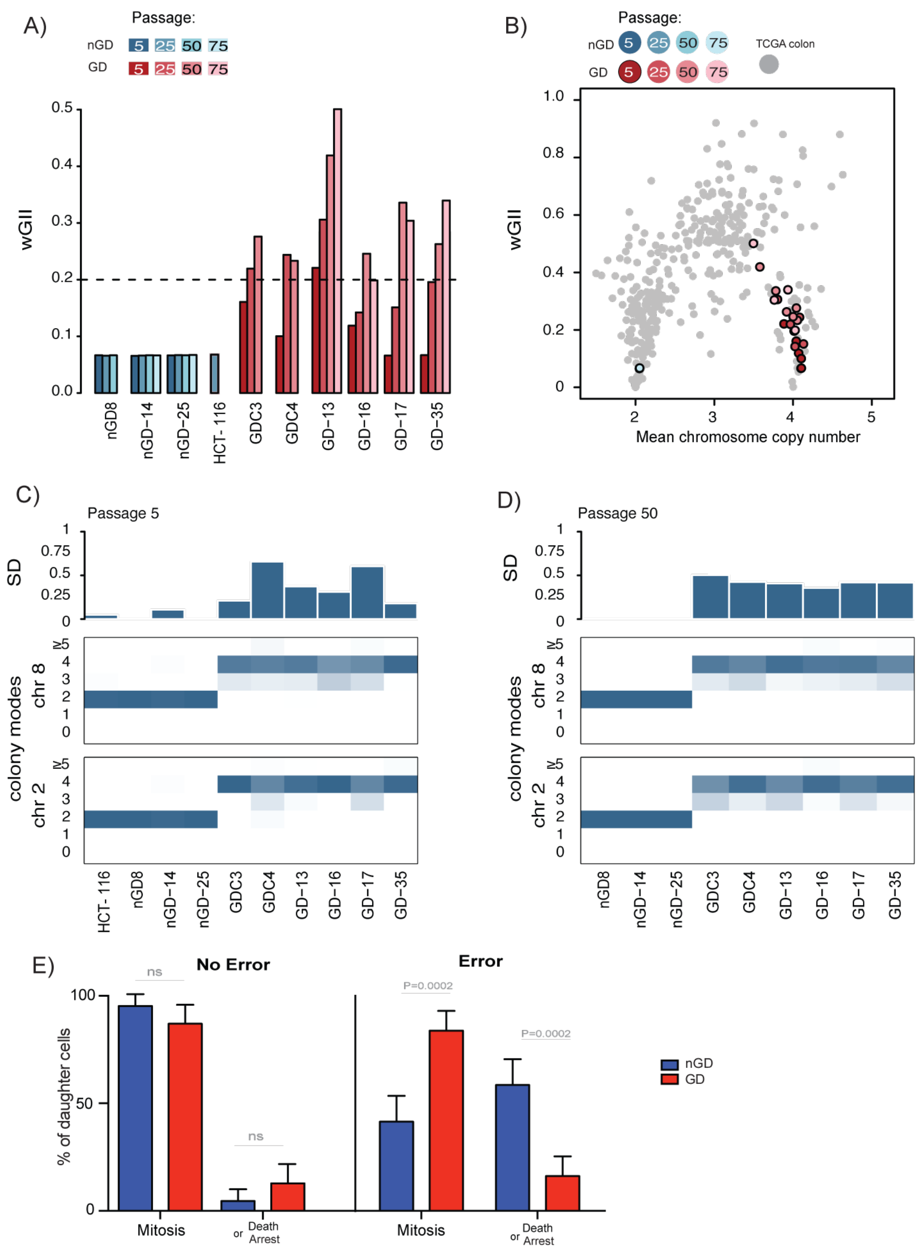


Figure 4:10 Tolerance to CIN in GD clones

A) wGIIs for nGD and GD clones from passage 5 to 75. Dashed line indicates wGII = 0.2, a threshold separating MSI and CIN cell lines (Burrell et al., 2013a). GD clones are depicted in red, nGD in blue. B) wGIIs for GD and nGD clones in relation to COAD tumours. Each grey dot represents one TCGA COAD tumour while red dots and blue dots represents GD and nGD clones respectively. C-D) Frequency of different colony modes from clonal FISH data is shown from all clones at passage 5 (C) and at passage 50 (D). Shannon diversity (SD) index is show above each colony. Median number of colonies scored: passage 5 = 44, passage 50 = 39. E) Live-cell imaging of H2B-mRFP–expressing cell reveals different daughter cell fates after segregation errors. GD= genome doubled; nGD non-genome doubled.

Early-passage non-genome-doubled clones exhibited wGII scores similar to stable COAD tumours. Furthermore, non-genome-doubled clones remained stable over time, suggesting either a lack of tolerance of segregation errors, or, alternatively, no selection for an altered karyotype. Conversely, in genome-doubled clones, wGII significantly increased from passage 5 through to passage 75 clones (One way ANOVA test for passage 5, 25 and 50, GD *P*=0.0002, nGD *P*=0.5907,Figure 4:10 A). By the late passage (approximately 18 months), genome-doubled clones had wGIIs similar to those of genomic complex polyploidy COAD tumours in the TCGA data set (Figure 4:10 B). Changes in genome-doubled clones over time in culture are therefore in keeping with a model of genome doubling occurring as a precursor state to a complex triploid karyotype, commonly observed in COAD.

To confirm the increased tolerance to chromosomal segregation errors observed in genome-doubled clones, clonal FISH data was further utilized. The presence of multiple colonies with distinct modal populations may be indicative of colony-to-colony variation and tolerance of CIN. The extent of this variation can be quantified by the Shannon diversity index. This index, frequently used in ecology, integrates both the number and abundance of colonies with differing modes. In non-genome-doubled, diploid clones, only a single aneuploid colony was observed in HCT-116 (1.7% of colonies) and one tetraploid colony was observed in a non genome-doubled clone, consistent with the observation that genome doubling is a rare and poorly tolerated event in diploid cells. Overall, genome-doubled clones exhibited significantly greater Shannon diversity indices than their non-doubled counterparts, both at passage 5 and passage 50 (Figure 4:10, *P*<0.01, Wilcox test). HCT-116_MLH1 genome-doubled clones also exhibited similar colony-to-colony variation in modal chromosome number.

Finally, live-cell imaging was used to track the fate of both genome-doubled and non-doubled cells following a segregation error (Figure 4:10 D). In non-genome-doubled clones, after a cell underwent a segregation error its daughter cells frequently died or underwent cell-cycle arrest. However, the majority of non genome-doubled cells that underwent a normal division (i.e. no detectable segregation errors) continued through a subsequent mitosis. By contrast, in genome-doubled clones, after a cell exhibited a segregation error its daughter cells frequently continued through to a normal mitosis in the subsequent cell cycle. These data suggest that lack of aneuploidy karyotypes observed in non-genome-doubled clones reflects a lack of tolerance of segregation errors and not a lack of selection for aneuploidy progeny.

4.2.11 Genome doubled clones evolve specific chromosome losses

Next, to explore whether convergent losses or gains were observed in genome-doubled clones, copy number profiles were assessed over long-term culture.

Chromosomal aberrations present in parental HCT-116 were observed in all analysed clones, confirming the ability of SNP6.0 data to identify clonal copy number aberrations, persisting through time and evolution of these clones. For example, chromosome 8q was identified as amplified in all early passage clones (Figure 4:11).

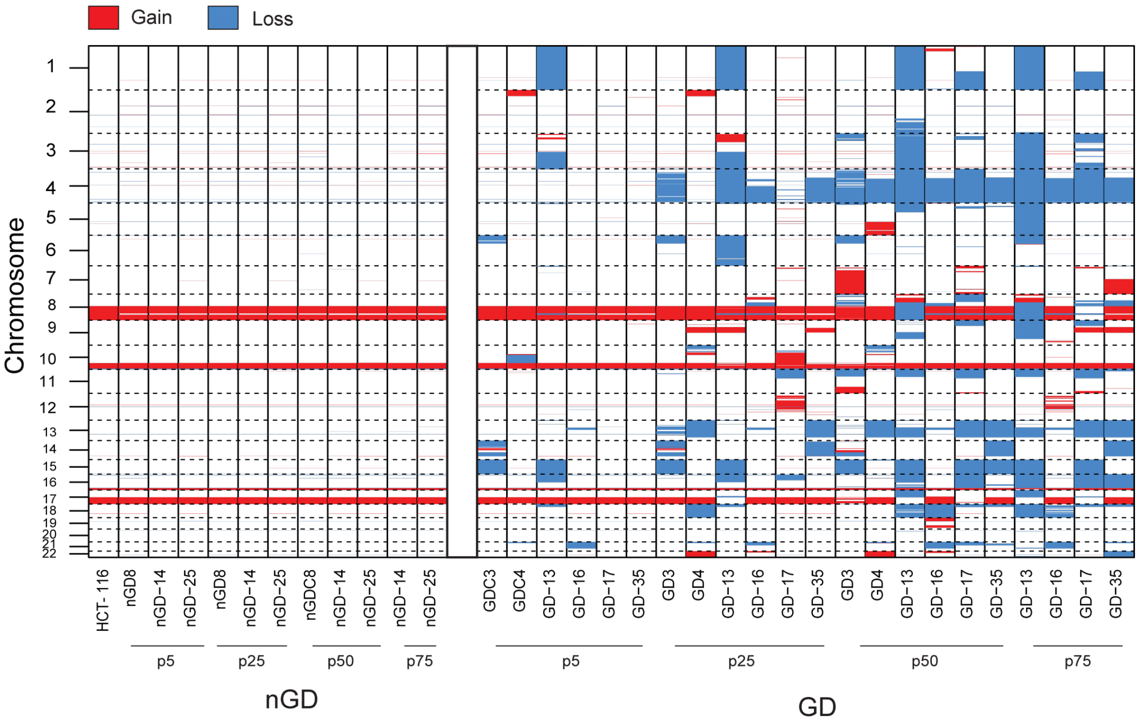


Figure 4:11 Specific losses in GD clones

Genome-wide copy number losses and gains for all clones at passage 5, 25, and 50 (and passage 75 for nGD-14, nGD-25, GD-13, GD-16, GD-17, and GD-35). Blue sections represent copy number losses and red sections represent copy number gains (relative to ploidy).

As expected, no novel chromosomal gains or losses were observed exclusively in non-doubled clones. Likewise, in early passage genome-doubled clones no novel losses or gains were observed in every clone, indicating no single detectable genomic change at the copy number level can account for tolerance to a whole genome-doubling event. However, a non-contiguous segment of chromosome 4q containing 362 genes was lost to three copies in every genome-doubled clone by passage 50. This chromosomal alteration occurs after genome doubling as it does not display LOH and not occur before passage 25. This is indicative of selection for loss of this chromosomal segment during prolonged culture in every clone. A similar pattern of chromosome losses was also observed in HCT-116 MLH1 genome-doubled clones, suggesting loss of this segment of the genome is independent of MIN status (Figure 4:12).

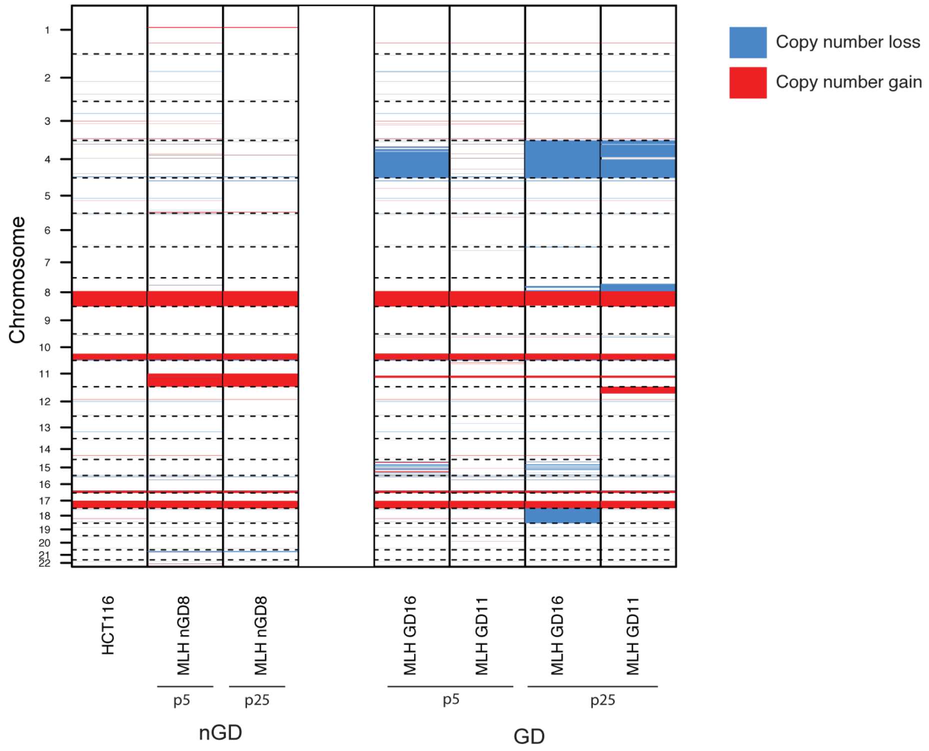


Figure 4:12 Specific losses in MLH GD and nGD clones

Copy number losses and gains for one MLH1 diploid and two tetraploid clones compared to parental HCT-116. Clone and passage number is shown on x-axis. Y-axis shows chromosomal position. Blue represents loss and red represents gain relative to ploidy.

4.2.12 Losses in genome-doubled clones recapitulate losses in unstable COAD tumours

Next, to explore whether loss chromosome of chromosome 4q was specific to the cell-line system or whether it recapitulates data observed *in vivo*, copy number data from COAD TCGA data was used.

Loss of genes identified on chromosome 4q in the cell-line system was found to significantly correlate with increasing wGII (Figure 4:13 A) in COAD tumours. The relationship remained significant when controlling for the increased likelihood of chromosomal loss in unstable tumours (Figure 4:13 B).

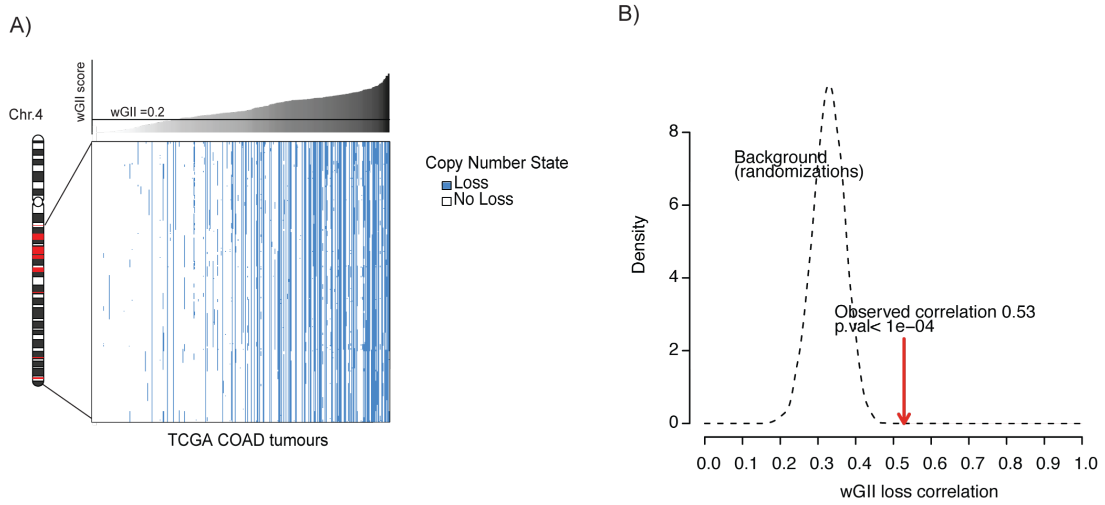


Figure 4:13 Loss in genome-doubled clones mirror losses in COAD tumours

A) Correlation between loss in COAD tumours and wGII. Notably, loss of genes on chromosomes 4q highly correlates with an unstable genome. B) Simulation to control for increased likelihood of chromosomal loss in unstable tumours. Correlation remained significant when taking into account the increased number of losses observed in tumours with high wGIIs.

These data suggest the genome doubling cell-line system mirrors the evolution of genome-doubled COAD tumours in terms of large-scale copy number events.

4.2.13 Genome doubling is associated with poor prognosis in CRC tumours

Given the timing of genome doubling in COAD tumours, and moreover, the established relationship between genomic instability and poor clinical outcome across cancer types (McGranahan et al., 2012), the association between genome doubling and clinical outcome was explored.

Clinical data, including information regarding both overall and relapse-free survival, was obtained for 150 stage 1-3 COAD patients from TCGA. In addition, a validation cohort, with relapse free survival, consisting of 189 stage 2-3 COAD patients was identified. Given that 80% of colorectal cancer recurrences occur within the first 2 years, and the paucity of data beyond this time point for TCGA tumours, survival data was censored at 2 years.

A genome-doubling event, as predicted using allele specific copy number, was found to be significantly associated with relapse in both the TCGA and the validation cohort (*P*=0.019 Discovery; *P*=0.0022, Validation, log-rank test) (Figure 4:14). When extending outcome to beyond 2 years, genome doubling remained a significant predictor of relapse in the validation cohort (*P*=0.00081, log-rank test), but not in the smaller TCGA cohort (*P*=0.099, log-rank test).

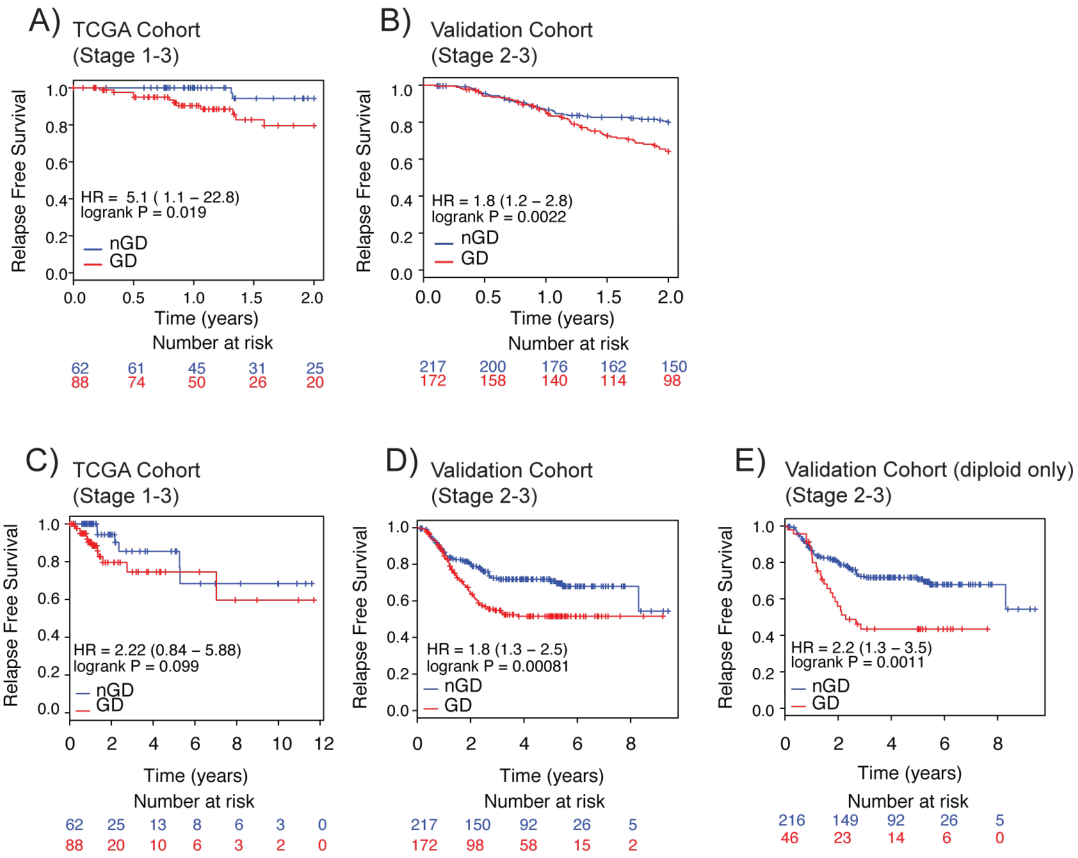


Figure 4:14 Genome doubling and relapse free survival in COAD.

Relationship between genome doubling and relapse free survival in A) TCGA data cohort (stage 1-3), censored at 2 years. B) Validation cohort (stage 2-3), censored at 2 years. C) TCGA data cohort (stage 1-3) not censored. D) Validation cohort (stage 2-3) not censored. E) Validation cohort (stage 2-3), not censored, diploid tumours only. In each case, the Kaplan Meier plots are shown, and p-values correspond to the log-rank test.

To determine whether genome doubling was predictive beyond classical predictors of survival in colorectal cancer a multivariate survival analysis was performed using a cox proportional hazard model, including tumour stage, age, and MSI status in the model. Genome doubling remained significant in multi-variate analysis in both cohorts (TCGA, *P*=0.045, Table 8-5; Validation, *P*= 0.028, Table 8-7). In the larger validation cohort, genome doubling was significant when restricting to just diploid tumours (Figure 4:14, *P*=0.001, log-rank test), and also when including polyploidy (ploidy ≥ 3) in a multivariate analysis (*P*=0.0209, Table 8-9).

A genome-doubling event may therefore provide prognostic relevance with a greater sensitivity than aneuploidy to detect high-risk tumours. Indeed, while the wGII was found to be significantly associated with relapse in a univariate analysis using TCGA data, it was not a significant predictor using the validation cohort. Moreover, wGII was not significant in multi-variate analysis in either cohort.

Furthermore genome doubling was significant in predicting overall 5-year survival when restricting to just early stage 1-2 tumours (available for the TCGA cohort only, Table 8-5).

Finally, the relationship between chromosomal content, genomic complexity and tumour stage was considered specifically within genome-doubled tumours. Consistent with a model where genome doubling is an early event in COAD tumours, accelerating cancer genome evolution, tetraploid genome-doubled tumours were found to be genomically more stable and exhibited lower tumour stage than sub-tetraploid genome-doubled (ploidy <4) samples (Figure 4:15).

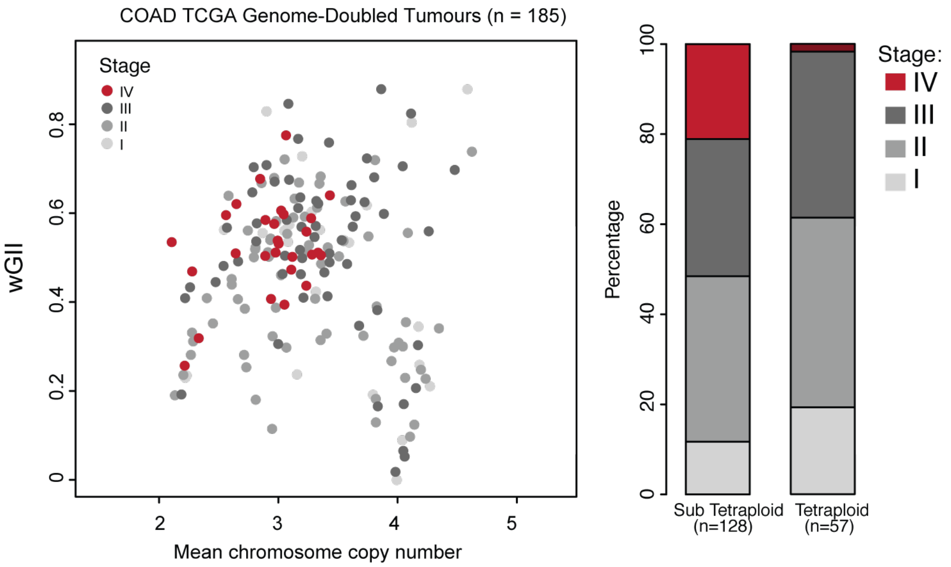


Figure 4:15 Tumour stage in genome doubled tumours

Each circle represents one genome-doubled tumour. The stacked bar-chart shows the proportion of different tumour stages for tetraploid and sub-tetraploid samples. P = 0.0062, Cochrane– Armitage test for trend.

4.3 Conclusions

Analysis of allele specific copy number and mutation data revealed the presence of genome doublings in every cancer type, and an association between genome doubling and genomic complexity. However, both the prevalence and timing of doublings were found to vary across cancer types. In COAD tumours, genome doubling appears to be an early event, often occurring prior to detectable chromosomal instability. Conversely, in other tumour types, genome doubling often appears to occur after the onset of detectable CIN.

In keeping with previous findings, a significant association between genome doubling and disruption to the tumour suppressor gene *TP53* was evident across the pan-cancer cohort. Nevertheless, despite this association, a substantial majority of genome-doubled tumours were found to be *TP53* wild-type, suggesting there are multiple routes to genome doubling and chromosomal instability, some of which may be independent of *TP53* disruption.

CIN is a complex phenotype, requiring not only segregation errors to be made, but also tolerance of the genomic remodelling from one cell division to the next. Indeed, increasing segregation error rates in diploid MIN COAD cells does not lead to propagation of chromosomally unstable progeny (Thompson and Compton, 2010). Through long-term culture of naturally occurring, rare surviving tetraploid clones, the evolution of genomic complexity specifically was explored in *TP53* wild-type colorectal cancer. A year post genome doubling, the HCT-116 genome was considerably altered relative to its genomically stable diploid progenitors, suggesting that tolerance of genome doubling facilitates rapid genome evolution.

Conceivably, a genome-doubling event may provide a permissive genetic background for selection of high-risk genomic copy number aberrations over time. Consistent with the impact of genome doubling upon genome instability, a tetraploidisation event was an independent predictor of reduced relapse-free survival time in COAD from the TCGA and in a larger validation cohort in both univariate and multivariate analyses. These data support studies that have linked aneuploidy with disease outcome in COAD (Araujo et al., 2007), and genome doubling with poor prognosis in ovarian cancers (Carter et al., 2012). Genome doubling may forecast the onset and tolerance of elevated CIN, which has previously been shown to be associated with both poor prognosis and intrinsic drug resistance (McGranahan et al., 2012, Lee et al., 2011a).

Current methods for detecting genome doubling are not without limitation. Predicting genome doubling relies on accurate copy number calling, which in itself is influenced by tumour purity. Future methods to define genome doubling and its timing may be improved by integrating SNV and copy number data.

Nevertheless, taken together, these results suggest genome doubling may play a key role in driving diversity at the copy number-level. In the next chapter of this thesis I will consider how mutational processes at the SNV-level may contribute to intra-tumour heterogeneity and their dynamics during tumour evolution.

Chapter 5. Dynamics of mutational processes

5.1 Introduction

The mutational processes that have been active during the evolutionary life history of a tumour leave scars in the cancer genome in the form of somatic events, such as SNVs and copy number aberrations. A particular mutational process may result in a preponderance of a particular type of somatic event. UV irradiation damage, for instance, results in both C>T single nucleotide substitutions as well as CC>TT di-nucleotide substitutions (Pleasance et al., 2010a). Similarly, chemicals contained in tobacco are thought to lead to an elevated number of C>A transversions (Govindan et al., 2012, Pleasance et al., 2009). As such, quantifying the types and abundance of different types of somatic events can reveal the processes that have been active during the evolutionary history of a tumour.

Recently, mathematical frameworks have been developed to formally quantify the number and contributions of mutational signatures operating within cancers at the single-nucleotide level (Alexandrov et al., 2013b, Fischer et al., 2013). Application of NMF (non-negative matrix factorization) and model selection to more than 7000 tumours from over 30 cancer types identified 20 distinct mutational signatures (Alexandrov et al., 2013a).

However, although the mutational signatures present during tumour evolution are beginning to be characterized, little is known about their temporal decomposition. Understanding the dynamics of mutational signatures can reveal the importance of different mutational processes at different stages of a tumour's life history. Moreover, these data may reveal whether patterns for tumour evolution can be deciphered. From a clinical perspective, this will inform whether mutational signatures can act as biomarkers and, further, whether certain processes can be considered as potential therapeutic targets.

In this chapter, I determine the presence and abundance of mutational processes and also explore their temporal decomposition in cancer genomes. I focus in-depth on NSCLC and oesophageal adenocarcinoma tumours, exploring the impact of both endogenous and exogenous mutational processes, and whether certain processes fuel intra-tumour heterogeneity and the acquisition of subclonal driver events.

The work presented in this chapter forms sections of three first or joint first-author publications, (McGranahan et al., 2015, de Bruin et al., 2014, Murugaesu et al., 2015), all of which can be found in Appendix 2.

5.2 Results

5.2.1 A signature of tobacco smoke in lung cancer genomes

To shed light on the mutational processes operating in lung cancer genomes, the prevalence of the six types of base substitutions (C>A; C>G; C>T; T>A; T>C; T>G) was explored in the cohort of six multi-region NSCLC tumours (Table 3-1). The tumours exhibited marked differences in both their mutational load and their mutational spectra (Figure 5:1). L002, L004 and L011 were characterized by a high mutation burden, with 476, 450 and 387 non-silent mutations identified in these tumours respectively. By contrast, in L001, L008 and the two tumours from L003, only 195, 117, 76 and 87 non-silent mutations were respectively identified. The dichotomy between high and low mutation burden directly reflect the smoking status of patients (Table 3-1). L002, L004 and L011 were all current smokers, while L001 and L008 were reformed smokers and L003 was a never-smoker.

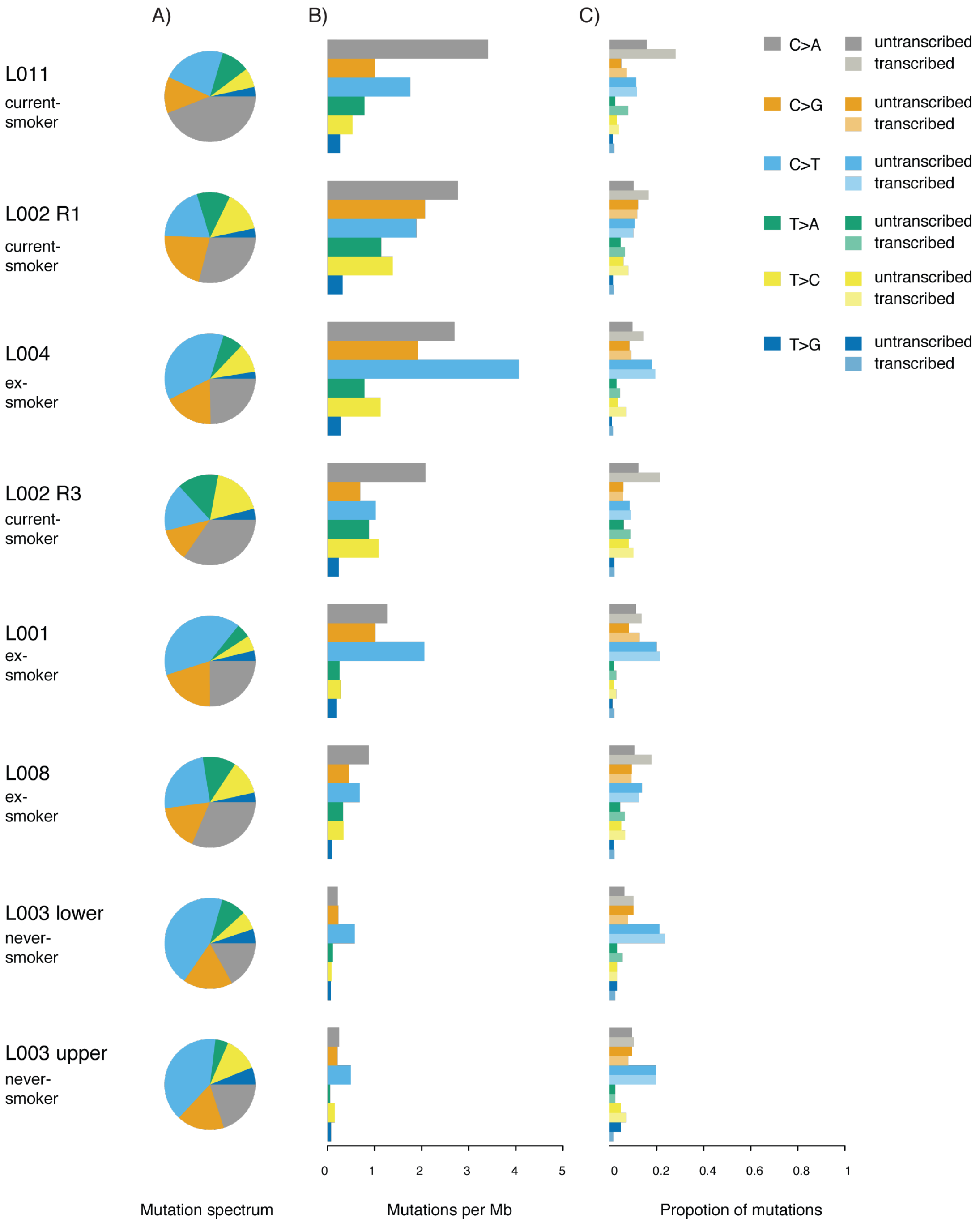


Figure 5:1 Mutational spectra and rate of multi-region NSCLC tumours

A) Relative abundance of the six mutation types is shown for each tumour. B) Number of mutations per mega-base (Mb) with adequate coverage for the six mutation types. C) For mutations occurring in transcribed regions of the genome, the proportion of each mutation type that was on the transcribed versus untranscribed strand is shown. All tumours are ordered with decreasing C>A transversion rate, from top to bottom.

The smoking status dichotomy also extended to difference in mutational spectra. L011, one of the heaviest smokers (45 pack-years – an average of 1.5 packets of cigarettes per day for 30 years), exhibited the greatest proportion of C>A transversions and highest rate of C>A transversions per megabase sequenced, with over 40% of base substitutions occurring as C>A transversion mutations, and a C>A transversion rate of over 3 substitutions per sequenced megabase (Figure 5:1). Conversely, the two tumours in L003, a never-smoker, were characterized by predominantly by C>G and C>T mutations, and also exhibited a low mutation rate; indeed, the total mutation rate of all nucleotide substitutions was lower than that for L011 C>A transversions alone.

To gain a deeper understanding of the types of base substitutions occurring in this cohort of tumours, mutations occurring in transcribed regions of the genome were considered further (Figure 5:1 B). While it is not possible to determine on which strand a mutation occurred, it is possible to assess whether a bias exists. Thus, it is possible to assess, for example, if C>T mutations preferentially occur on the transcribed strand relative to the non-transcribed strand or vice versa.

A significant strand bias was observed specifically for C>A transversion mutations for every tumour (*P*<0.01, chi-square test), with the exception of L003 and L001. As discussed, while L003 was a never-smoker, L001 was a former smoker, who stopped smoking 25 years ago (Table 3-1). The bias observed was characterized by fewer C>A transversions on the non-transcribed compared to the transcribed strand. This is indicative of transcription coupled repair, and specifically nucleotide excision repair, which is thought to remove nucleotides with bulky adducts from the transcribed strands of genes. The presence of an elevated number of C>A transversions on the transcribed strand (at the expense of a reduced number of G>T transversions) suggest that bulky adduct damage to guanine may be the underlying cause of the observed base substitution. These data are consistent with findings of Pleasance and colleagues (2009).

5.2.2 Smoking induced mutations are more likely to give rise to non-silent mutations than C>T transitions

The degenerate nature of the genetic code results in a subset of base substitutions being synonymous, such that, despite a change in DNA nucleotide, no change in amino acid is observed. For example, both triplets AGC and AGT code for the amino acid serine, and, as such, a C>T transition in the 3rd codon position of ApGpC will be silent at the amino acid level.

Given the different mutational spectra observed in different cancer genomes, and, by extension, the different mutational processes that have been active, it is worth considering that each mutation type, and associated mutational process, may be associated with a different likelihood of yielding a non-synonymous mutation.

Comparing the ratio of C>T transition (the most prevalent nucleotide substitution in never-smokers) to C>A transversion mutations (the most prevalent in smokers), and the proportion of base substitutions that lead to amino acid substitutions reveals that a C>A transversion is more likely to result in non-synonymous mutations than a C>T transition (Figure 5:2). This suggests tobacco induced mutations may not only increase the mutation rate as a whole, but will also increases the proportion of the total mutational load that results in sequence changes at the amino acid level.

An analysis of mutations at cytosine sites within the 3rd codon position, the position that is most likely to be degenerate, revealed a clear difference between heavy smokers and never-smokers, using data obtained from Imielinski et al. (2012) (Figure 5:2). Specifically, for heavy smokers, on average, over 30% of mutations occurring at the 3rd codon position result in non-synonymous mutations, yielding a change in coding sequence. Conversely, in never-smokers, on average, only 11% of mutations at the 3rd codon position result in a non-synonymous mutation. Thus, mutations at the 3rd codon position are, on average, 2.72 times more likely to be non-synonymous in smokers than never-smokers.

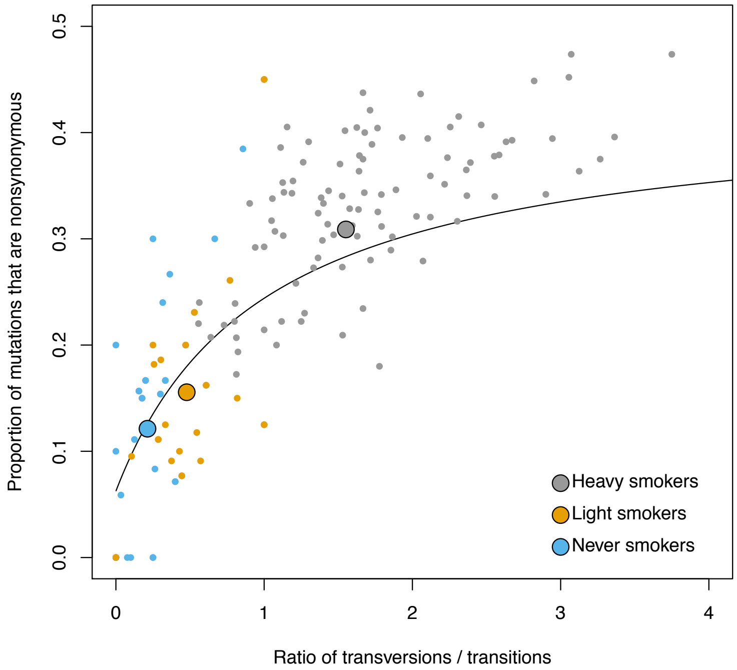


Figure 5:2 Likelihood of non-synonymous coding mutations

The relationship between the ratio of C>A transversions compared to C>T transitions is shown for each tumour within the Imielenski cohort (2012). The line represents the expected relationship between the two given the genetic code. Each dot represents one tumour, with the colour indicating smoking status. Large circles represent the median value for each category. In each case, mutations in the 3rd codon position are considered. Light smokers are defined by less than ten pack years of tobacco use), and heavy smokers more than ten pack years

The effect of tobacco induced C>A transversions can also be considered in the context of specific DNA sequences or cancer genes. Across the coding sequence of the tumour suppressor gene *TP53* over 80% of all possible C>A transversion mutations result in a non-synonymous amino acid substitution. By contrast, only 63% of all possible C>T transition mutations are non-synonymous at the amino acid level. These data highlight the need to consider both the nucleotide composition and mutational spectra of genomes when attempting to identify cancer genes.

It is also interesting to note that the majority of tumours exhibited an elevated rate of non-synonymous mutation acquisition compared to what would be expected by chance given the genetic code (Figure 5:2). This suggests positive selection may be pervasive during lung cancer genome evolution.

5.2.3 Temporal dissection of mutation spectra in NSCLC

Mutational signatures may not only vary between different tumours, but also over time, reflecting the dynamics of mutational processes during the evolutionary history of a tumour. The temporal dynamics of these processes can be deciphered by comparing the spectrum of point mutations acquired on the trunk of a tumour's phylogenetic tree with those acquired on the branches.

Early (trunk) mutations likely reflect processes involved prior to and during tumour initiation and early development. Conversely, late (branch) mutations reveal mutational processes shaping the genome during tumour maintenance and progression, and may shed light on mechanisms driving diversity within tumours.

Every multi-region NSCLC tumour, with the exception of L003, had sufficient numbers of branch mutations to permit temporal analysis. Given that L002 represents two histologically distinct subtypes, each histological subtype was treated separately, allowing comparisons of LUAD (L002_R1) and LUSC (L002_R3) histologies within the same tumour.

Statistically significant shifts in the mutation spectra were observed between early and late mutations in all five tested NSCLC tumours (Figure 5:3; *P*<0.05 all cases, chi-square test). As might be expected, the two ex-smokers both exhibited a statistically significant decrease in the proportion of C>A transversions, indicating a relative decrease in the mutational burden attributable to smoking in late mutations (Figure 5:3; *P*<0.05 all cases). Curiously, a significant decrease in C>A transversions was also observed for current smokers (Figure 5:3; *P*<0.05 all cases). Conceivably, this suggests either the rate of smoking induced mutations may be decreasing, despite continuous exposure to tobacco smoke, or, alternatively, an additional mutational process may dominate later in tumour evolution.

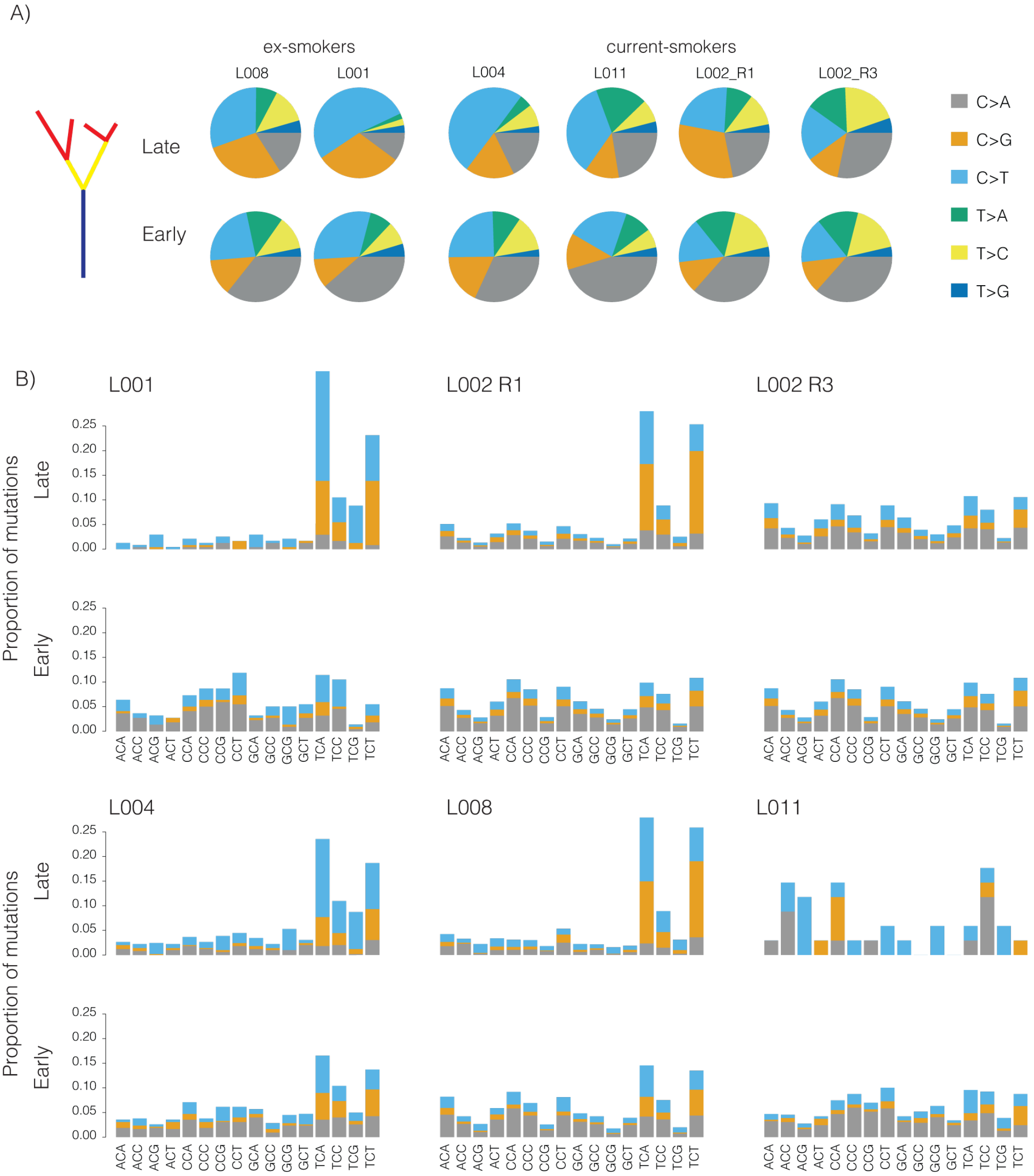


Figure 5:3 Temporal dissection of multi-region NSCLC

A) Fraction of early mutations (trunk) and late mutations (branch) accounted for by each of the six mutation types in all multi-region NSCLC. B) Tri-nucleotide context of three mutation types (C>A; C>G; C>T). For each tumour, early (trunk) mutations are depicted on bottom, with late (branch) mutations depicted on top.

5.2.4 Temporal heterogeneity of APOBEC mutations in NSCLC

In every NSCLC tumour, the decreasing proportion of C>A mutations later in evolution was accompanied by an increase in C>T and C>G mutations. In order to gain further resolution regarding the mutational processes active during the evolution of NSCLC tumours, the tri-nucleotide context of each mutation was considered. Thus, the nucleotides both 5' and 3' to the mutated base were taken into account, yielding a total of 96 possible mutation types, and, when focussing on substitutions from cytosine to any other base, a total of 48 nucleotide substitutions (Figure 5:3 B).

While no particular tri-nucleotide context preference was observed for C>A transversions, C>G and C>T mutations occurring on the branches of NSCLC tumours exhibited a clear preference towards occurring at TpCpW sites (where W represents A or T) (Figure 5:3 B). This particular mutational motif is similar to that of APOBEC3B and APOBEC3A, and may therefore be indicative of APOBEC mediated mutagenesis (Roberts et al., 2013, Nik-Zainal et al., 2012a).

To formally assess the significance of an APOBEC mutagenesis pattern within each sample, the frequency of mutations occurring with an APOBEC context (TpCpW to TpTpW or TpGpW) was compared to that expected with random mutagenesis. Enrichment for APOBEC associated mutations was calculated relative to the frequency of the APOBEC mutation motif within the genomic region surrounding mutated bases (20 nucleotides on either side). This approach was chosen as the APOBEC enzyme is thought to scan a limited area of single stranded DNA (Roberts et al., 2013).

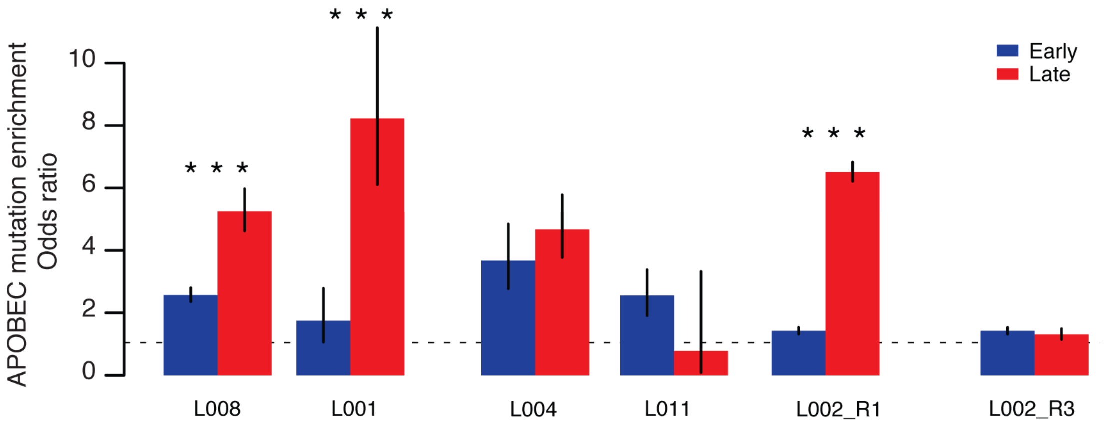


Figure 5:4 Temporal dissection of APOBEC enrichment

APOBEC mutation enrichment odds ratio for early (trunk, blue bars) and late (branch, red bars) mutations for multi-region NSCLC samples. The APOBEC signature encompasses C>T and C>G mutations in a TpC context. . The 95% confidence intervals for Fisher's exact test are indicated. Stars indicate *P*<0.001 of comparison between early and late APOBEC enrichment using Fisher's exact test.

APOBEC mutation enrichment was identified in five tumours, L008, L001, L004, L011, and L002 R1. In four of these five tumours APOBEC enrichment was found to be higher on the branches compared to the trunk (L008, L001, L004, and L002 R1), and the difference was statistically significant in three cases (L008, L001, and L002 R1). On average, in L001, L004, L002 R1 and L008 36% of non-silent branch mutations were found to occur in an APOBEC mutation context (range 29-41%), significantly more than compared to non-silent trunk mutations (range 7-16%). Notably, L011, the only tumour that did not show any evidence for branched APOBEC enrichment, was also the least heterogeneous tumour.

To confirm whether APOBEC3B was responsible for the observed pattern of mutagenesis, APOBEC3B mRNA expression was also assessed by qPCR in 5 tumours (L001; L002; L003; L004 and L011). This experiment was undertaken by Elza de Bruin. In every tumour, APOBEC3B expression was enriched compared to matched normal, consistent with APOBEC3B being responsible for the observed signature.

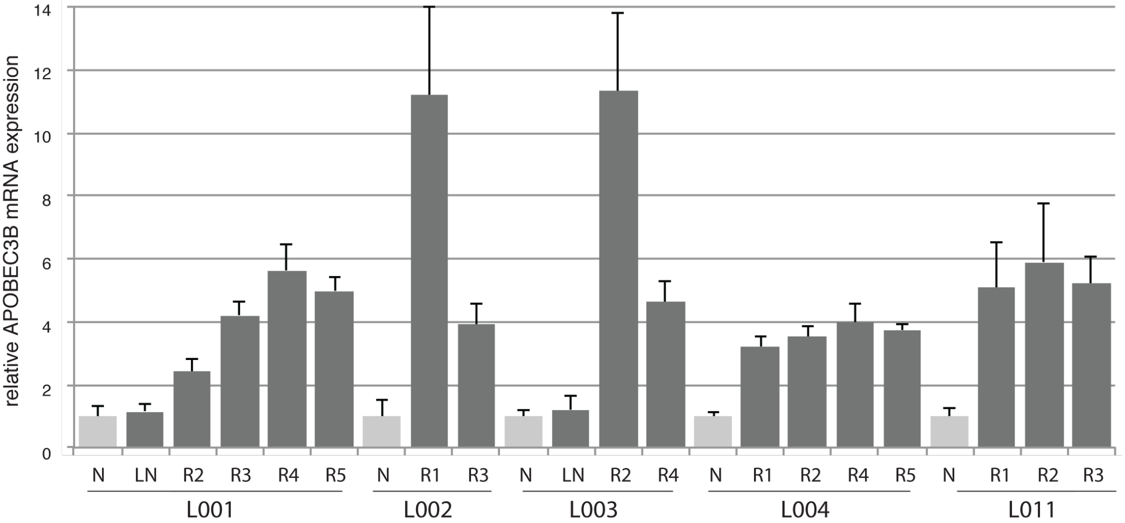


Figure 5:5 APOBEC3B mRNA expression

Barplot showing APOBEC3B mRNA expression in tumour regions relative to the adjacent normal lung for each tumour, using *TBP* mRNA expression for normalization.

5.2.5 Spatial heterogeneity of APOBEC mediated mutagenesis

Next, to explore whether APOBEC mediated mutagenesis may not only be temporally heterogeneous, but also spatially heterogeneous, tumours with multi-region WGS data were used.

Clear differences in mutational spectra and APOBEC enrichment was observed between the two tumour regions in L002, derived from a current smoker. While the adenocarcinoma region, region R2, showed a highly significant enrichment of APOBEC mutagenesis (*P*<0.001), with over 5378 mutations out of 15285 mutations classified as C>T and C>G mutations occurring at TpC sites, no enrichment was observed in the squamous cell carcinoma region, region R3. The predominant mutation type in region R3 remained C>A transversions, consistent with smoking still playing a role in moulding the tumour genome. With regards to mutational load, region R1 exhibited 3.6 times more mutations than regions R3, suggesting APOBEC mediated mutagenesis results in a highly elevated mutation rate. mRNA expression analyses confirmed a higher *APOBEC3B* expression in L002 region R1 compared to R3 (Figure 5:5).

Spatial heterogeneity in mutational spectra was also observed for L008. While the tumour regions from both the middle and upper lobes of the lung exhibited mutational signatures consistent with APOBEC mediated mutagenesis, this was more pronounced in the tumour from the middle lobe, region R3. In this tumour, no significant difference in mutational load was observed between the two tumour regions, however, region R3 exhibited a low purity level which may have led to many mutations occurring at variant allele frequencies below the level of detection.

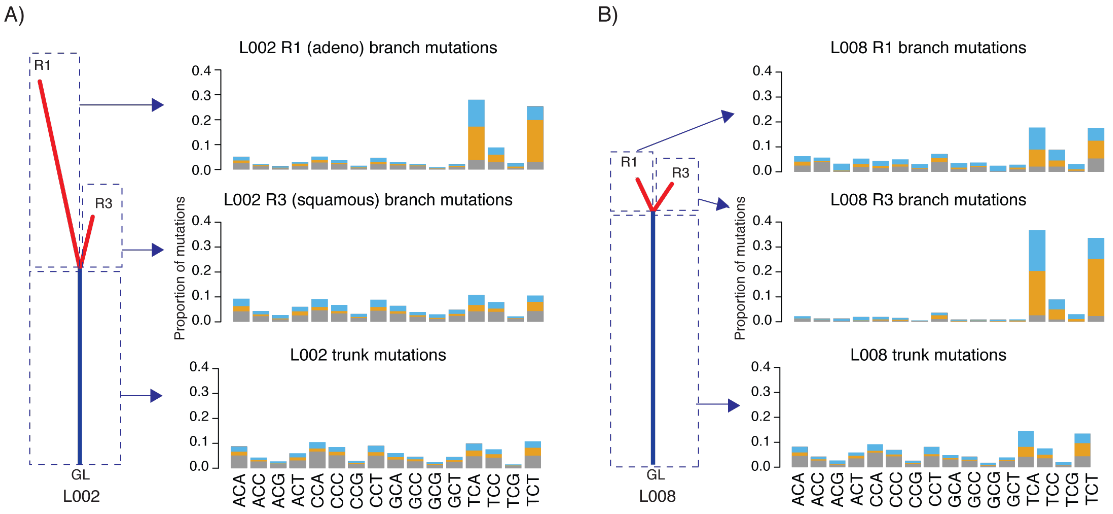


Figure 5:6 Spatial heterogeneity of APOBEC mediated mutagenesis

Three mutation types (C>A; C>G and C>T) at all 16 possible tri-nucleotide contexts for L002 (A) and L008 (B). For both samples, trunk mutations as well as branch mutations from two regions are depicted. Notably, in both cases one region shows a clear preponderance of C>T and C>G mutations at TpC sites, indicative of APOBEC mediated mutagenesis.

5.2.6 Mutational signatures identified in the pan-cancer cohort

To validate the findings in NSCLC in a larger cohort and further explore the dynamics of mutational processes across a range of different cancer types, data from the TCGA was used. For this analysis, breast cancer was divided into two cohorts, oestrogen receptor positive and oestrogen receptor negative. In total, 10 cancer types were considered, BLCA, ER-positive BRCA, ER-negative BRCA, COAD, GBM, HNSC, KIRC, LUAD, LUSC and SKCM.

To explore the mutational signatures present in this large cohort of tumours, mathematical tools, principally non-negative matrix factorization (NMF) and model selection, were applied separately to each cancer type. In brief, this approach considers the tri-nucleotide of the catalogue of mutations contained within each tumour genome and attempts to quantify both the underlying signatures responsible for the patterns of mutations observed and their abundance within individual tumours (Alexandrov et al., 2013b).

In total, across the 10 cancer types, 10 robust mutational signatures were identified (Figure 5:7, Experimental procedures). Each signature was clustered with the set of mutational signatures previously identified by Alexandrov et al. (2013a) to assess whether the signatures identified in this analysis corresponded to known signatures from a larger analysis. Clustering analysis suggested every signature identified could be linked to an established mutational signature (Figure 5:7).

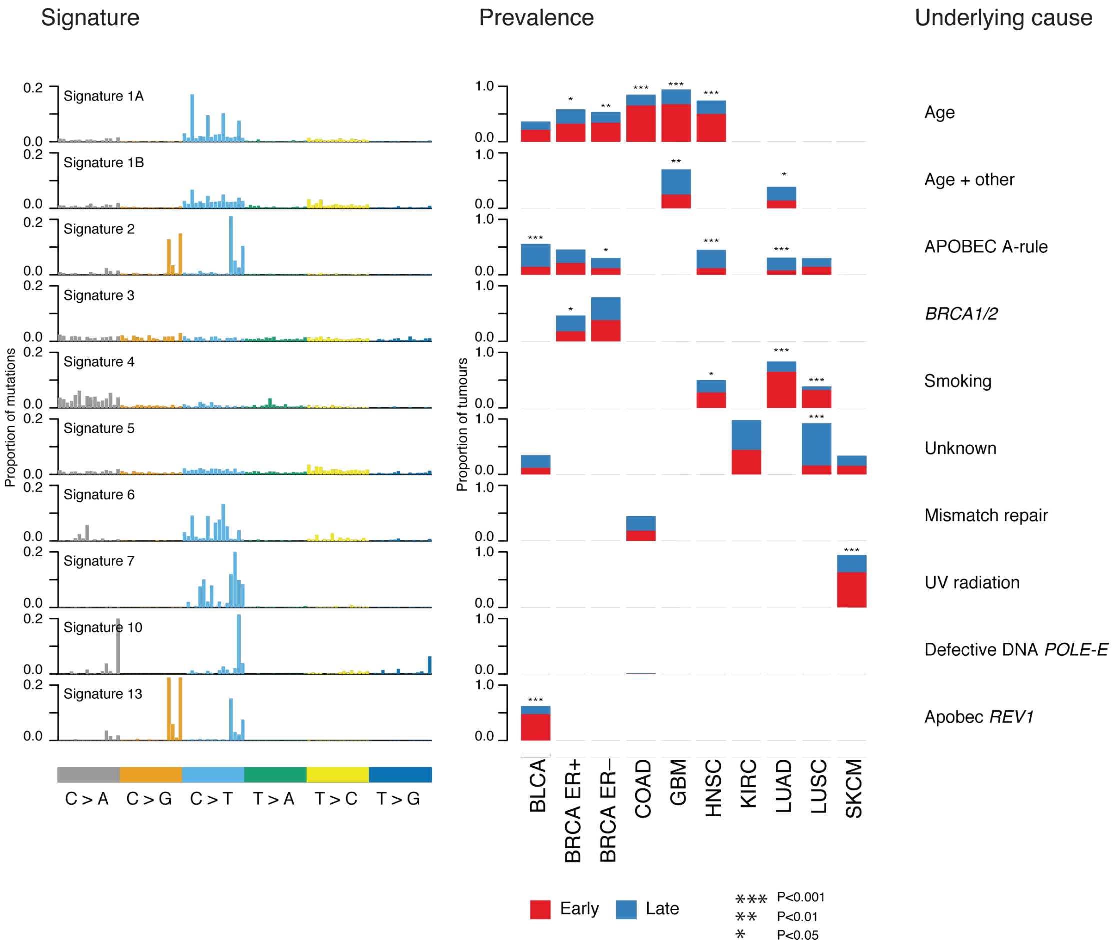


Figure 5:7 Prevalence and temporal dissection of mutational signatures

Mutational signatures identified across 9 cancer types and their temporal dissection. Signature names correspond to those identified by Alexandrov. For each signature present in a cancer type, the proportion of tumours where the signature is prominent early (red), compared to more prominent late (blue) is shown.

The two most prominent signatures identified across the pan-cancer cohort were Signature 1A and Signature 2. Signature 1A is characterized by C>T mutations at CpG sites and has been attributed to spontaneous deamination of methylated cytosine. This signature is thought to reflect cell-turnover, and has been shown to correlate with patient age at diagnosis (Alexandrov et al., 2013a).

Signature 2, on the other hand, is dominated by C>T and C>G mutations at TpC sites and is thought to relate to APOBEC3B mediated mutagenesis. To validate the similarity between Signature 2 and the APOBEC enrichment signature defined above, the two signatures were compared using all TCGA LUAD samples. Reassuringly, Signature 2 was found to significantly correlate with the APOBEC signature (*P*<2.2e-16, R^2 = 0.74, Figure 5:8), suggesting the two independent methods to define APOBEC mediated mutagenesis are concordant.

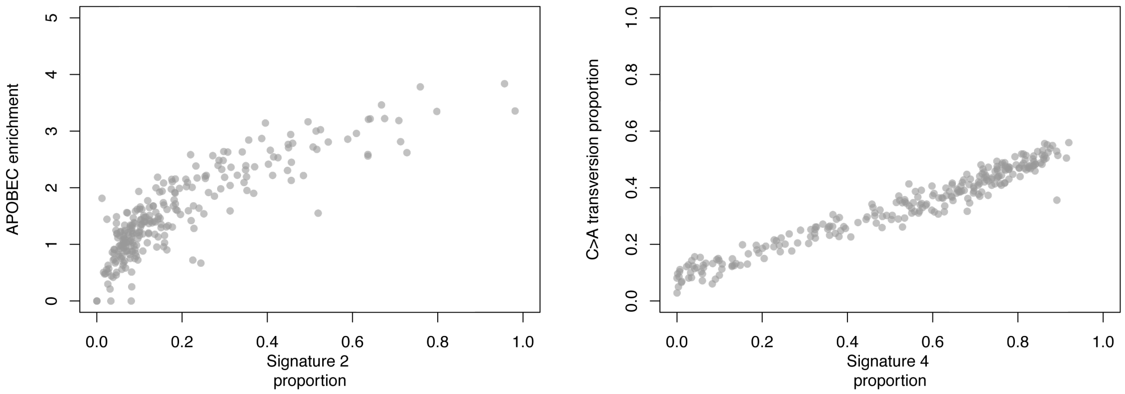


Figure 5:8 Comparisons of algorithms to infer mutational processes

Left panel shows comparison between two independent approaches to identify APOBEC mediated mutagenesis patterns. The fold enrichment of APOBEC-signature mutations was determined for all TCGA LUAD tumours and compared to the proportion of mutations identified as Signature 2 using NMF and model selection. High correlation, as determined by Pearson correlation, indicates concordance between the two approaches. Right panel shows comparison between the proportion of C>A transversions within each cancer tumour and the proportion of Signature 4 mutations using NMF and model selection. High correlation, as determined by Pearson correlation, indicates these independent methods are highly concordant.

In addition, as expected, a signature reflecting exposure to tobacco carcinogens was also identified in the cohort. Consistent with multi-region sequencing analysis, Signature 4, characterized by a preponderance of C>A transversions was identified in both LUAD and LUSC tumours, and this signature was also identified in HNSC tumours. Further, the proportion of mutations classified as Signature 4 was found to highly correlate with the proportion C>A transversions in LUAD tumours (R^2 =0.95, *P*<2.2e-16, Figure 5:8). These data suggest considering the proportion of C>A transversion mutations is equivalent to determining the proportion of Signature 4 mutations.

Certain signatures were also specific to given cancer types. For example, every SKCM tumour was characterized by Signature 7 mutations, which is thought to relate to exposure to UV radiation, and is characterized by a preponderance of C>T transition mutations, especially at TpCpC sites (Alexandrov et al., 2013a). Signature 6, on the other hand, was specific to COAD tumours and is thought to reflect micro-satellite instability (Alexandrov et al., 2013a). A subset of COAD tumours were also characterized by a highly specific mutational profile, involving C>A transversion mutations at TpCpT sites and C>T transitions at TpCpG sites. The mutational process responsible for this signature is thought to be mutations in *POLE-E*, resulting in a lack of recognition and removal of mis-paired nucleotides by the exonucelase activity of DNA polymerase-ε (Alexandrov et al., 2013a)

In both oestrogen positive and oestrogen negative breast cancers, a signature (Signature 3) thought to reflect defects in *BRCA1* and *BRCA2*, genes involved in DNA double strand break repair by homologous recombination, was identified.

Mutational signatures where the underlying aetiology remains unclear were also identified. In five cancer types, Signature 5 was identified. Given its relatively flat profile, with no particular tri-nucleotide context prominent, it is difficult to ascertain the mutational process responsible for this signature.

5.2.7 Temporal dissection of mutations reveals dynamics of mutational processes

Next, temporal dissection of mutations was used to shed light on the dynamics of the mutational processes during tumour evolution and to explore whether findings using multi-region sequencing in NSCLC could be validated. Mutations were classified as ‘early’ or ‘late’ based on both their cancer cell fraction and their mutation copy number in given tumours. It was reasoned that clonal mutations likely represent relatively early events in tumour evolution, occurring before or at the time of the most recent clonal sweep, whereas subclonal mutations represent later events. In the case of genome doubling or amplification events, the timing of mutations was further refined; a mutation occurring before doubling would be expected to be present at multiple copies, whereas a mutation occurring after a doubling would only be present at one copy. Thus, mutations were defined as ‘early’ if they were clonal and did not occur after amplification or a genome-doubling event. Conversely, ‘late’ mutations were defined as those that were subclonal or occurring after genome doubling or amplification events (Figure 5:9).

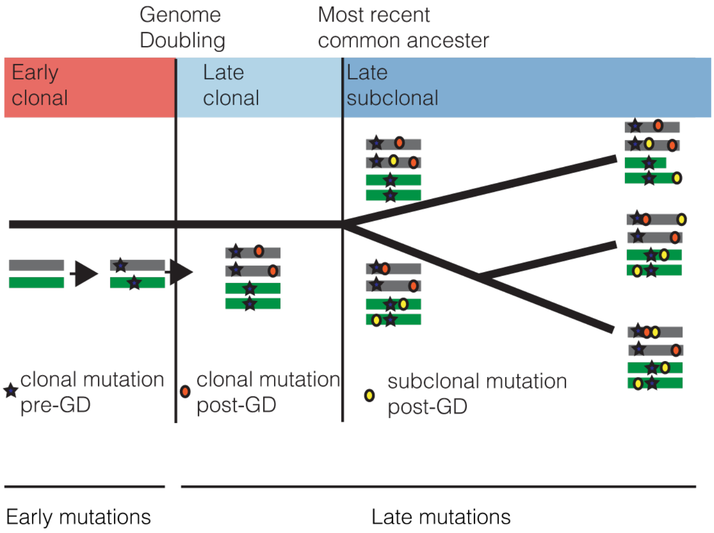


Figure 5:9 Timing mutations using copy number and clonal status
Mutations can be timed by using both copy number and clonal status of mutations. Clonal mutations occurring prior to genome doubling of amplification events are likely ‘early’ mutations, whereas mutations occurring after genome doubling, or only in a subset of cancer cells are likely ‘late’ events.

In four cancer types, Signature 1A was significantly more prevalent in early compared to late mutations (BRCA ER-positive, *P*=0.0359;BRCA ER-negative, *P* =0.000629; COAD, *P*=7.59e-15; GBM, *P*=3.55e-10; and HNSC, *P*=2.26e-09, Figure 5:7). These data suggest that a large proportion of early mutations represent non-cancer-specific mutational processes, such as aging, and additional cancer-specific mutational processes may be required for tumour development.

Consistent with additional cancer-specific mutational processes increasing in prevalence later in tumour evolution, and in keeping with temporal dissection of multi-region NSCLC, Signature 2, which, as discussed, has been linked with up-regulation of APOBEC cytosine deaminases, was found to increase over time in the majority of tumour types in which it was detected. A highly significant increase in Signature 2 was observed for later mutations in LUAD tumours (*P*=3.33e-06), and a similar pattern was observed for HNSC (*P*=5.57e-07) and BLCA

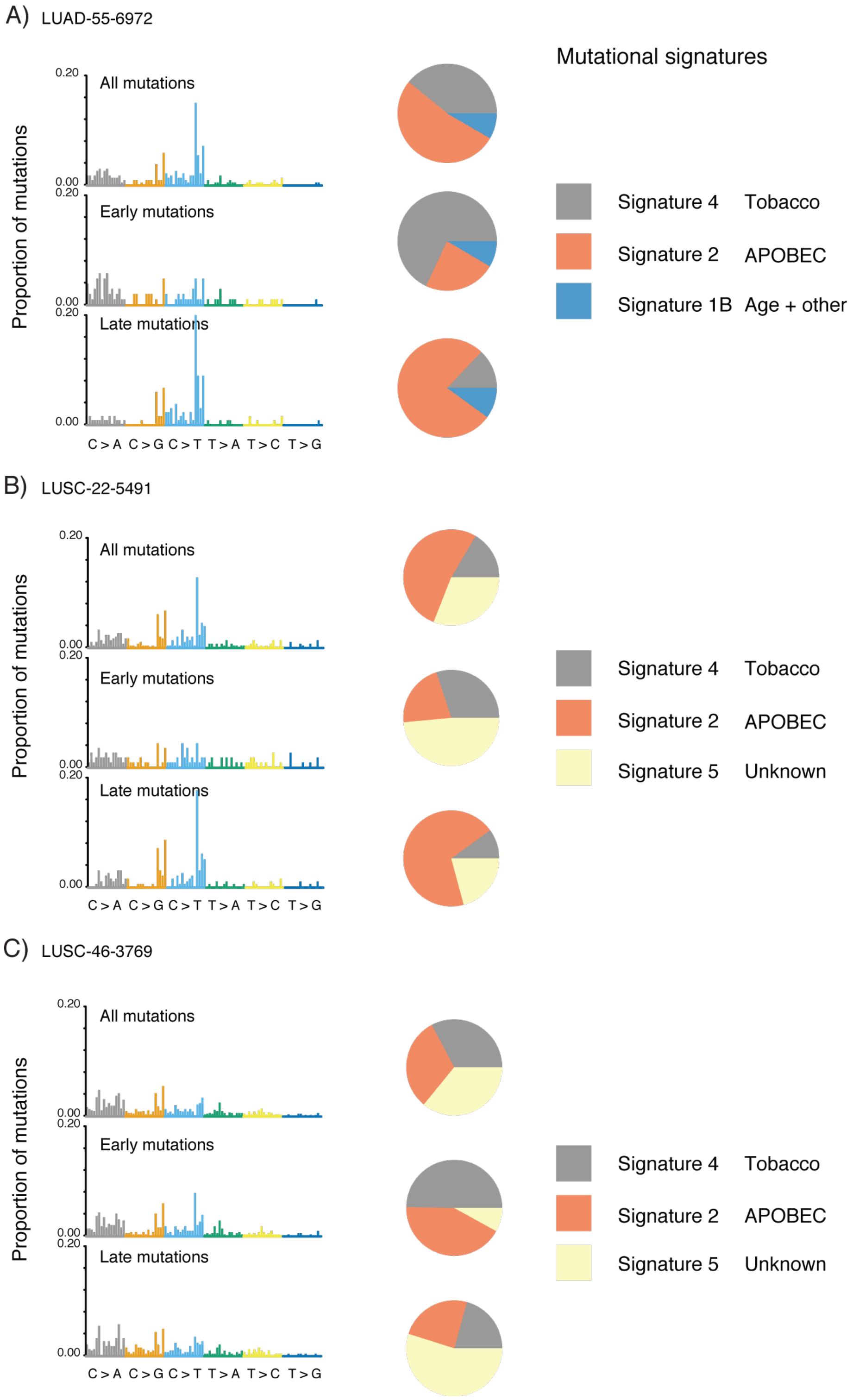


Figure 5:10 Examples of temporal dissection of TCGA LUAD and LUSC tumours

For three tumours A) LUAD-55-6972; B) LUSC-22-5491; C) LUSC-46-3769 mutational spectra, showing proportion of mutations at each of the 96 tri-nucleotide contexts is shown for all mutations, early mutations and late mutations. In addition, mutational signatures for different mutational sets are shown.

(*P*=4.39e-06). LUAD-55-6972 provided a clear example of a TCGA LUAD tumour exhibiting an increase APOBEC signature over time (Figure 5:10 A): while less than 25% of early mutations in this tumour were classified as Signature 2, over 70% of late mutations were found to be potentially APOBEC-related.

In LUSC tumours, although APOBEC was frequently detected, no clear pattern in temporal dissection was observed - with some tumours exhibiting APOBEC as a late pattern, and others as early. For example, in LUSC-22-5491, APOBEC mediated mutagenesis was a later event, with 69% of late mutations classified as belonging to Signature 2, compared to only 21% of early mutations (Figure 5:10 B). LUSC-46-3679, by comparison, exhibited strong APOBEC enrichment in its early mutations, with less APOBEC enrichment detected in late mutations (Figure 5:10 C).

Interestingly, in BRCA a tendency for Signature 2 to increase in frequency in later mutations was observed in ER-negative BRCA (*P*=0.0126), but not in ER-positive breast cancers (*P*=0.597), highlighting the differences between these two subtypes of breast cancer. Signature 13, also linked to APOBEC cytosine deaminases, was found to be significantly more prevalent in early compared to late mutations in BLCA (*P*=6.67e-07). Signature 13 has been postulated to be associated with APOBEC coupled with excess activity of DNA repair protein *REV1*. These data suggest Signatures 2 and 13 are distinct mutational processes that can be separated temporally in BLCA (Figure 5:12).

In terms of exogenous mutational processes, Signature 4, associated with smoking-induced mutations, was significantly more prevalent in early compared to late mutations in LUAD (*P*=4.56e-17) and LUSC (*P*=8.1e-07), and a similar pattern was observed in HNSC (*P*=0.0338). For example, while over 50% of early mutations in LUAD-55-6972 were classified as Signature 4, after the whole genome-doubling event, only 10% of late mutations were linked to tobacco exposure (Figure 5:10 A). These data act as a validation for the multi-region sequencing findings described above and likely reflect the fact that despite the mutagenic effects of tobacco smoke, additional mutational processes occur during tumour development.

Furthermore, in keeping with the small cohort of multi-region NSCLC tumours, the temporal decrease in C>A mutations, associated with Signature 4, was observed both for smokers and never-smokers, suggesting its proportional decline does not simply represent an absence of tobacco induced mutations (Figure 5:11).

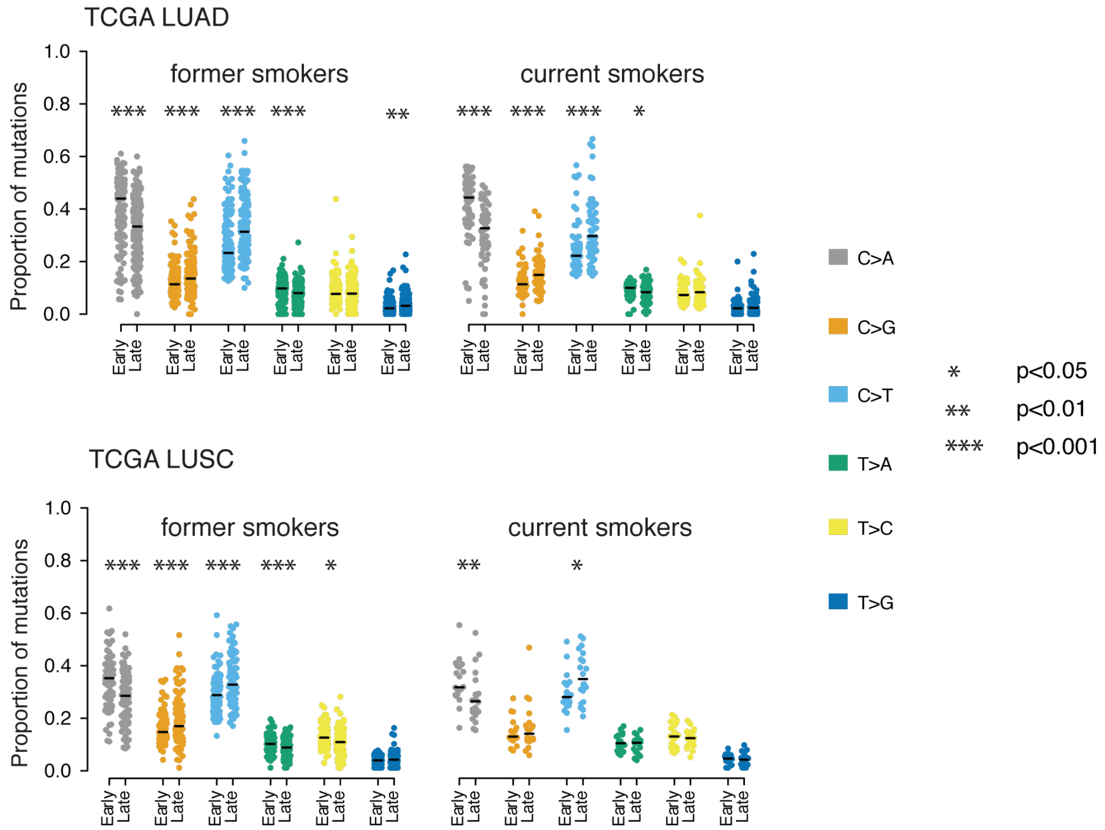


Figure 5:11 Temporal dissection of LUAD and LUSC TCGA tumours

Proportion of early and late mutations that correspond to the six different mutation types is shown for TCGA LUAD and LUSC tumours, further grouped according to smoking status. Significance is shown for a paired t-test for each mutation type.

To shed further light on the impact of smoking in tumour development, and further verify whether tobacco smoke still induces mutations later in tumour evolution, mutations occurring in transcribed regions of the genome were considered. Both LUAD and LUSC ex-smokers revealed a statistically significant decrease in the strand bias in late compared to early C>A transversions (LUAD, *P*=0.00354; LUSC, *P*=0.046), consistent with smoking having left a footprint on these genomes but no longer being active. By contrast, no statistically significant difference in strand bias was observed between early and late mutations in current-smokers in either cancer type (LUAD, *P*=0.23; LUSC, *P*=0.220). The data indicate that in current-smokers, smoking likely still impacts upon the cancer genome.

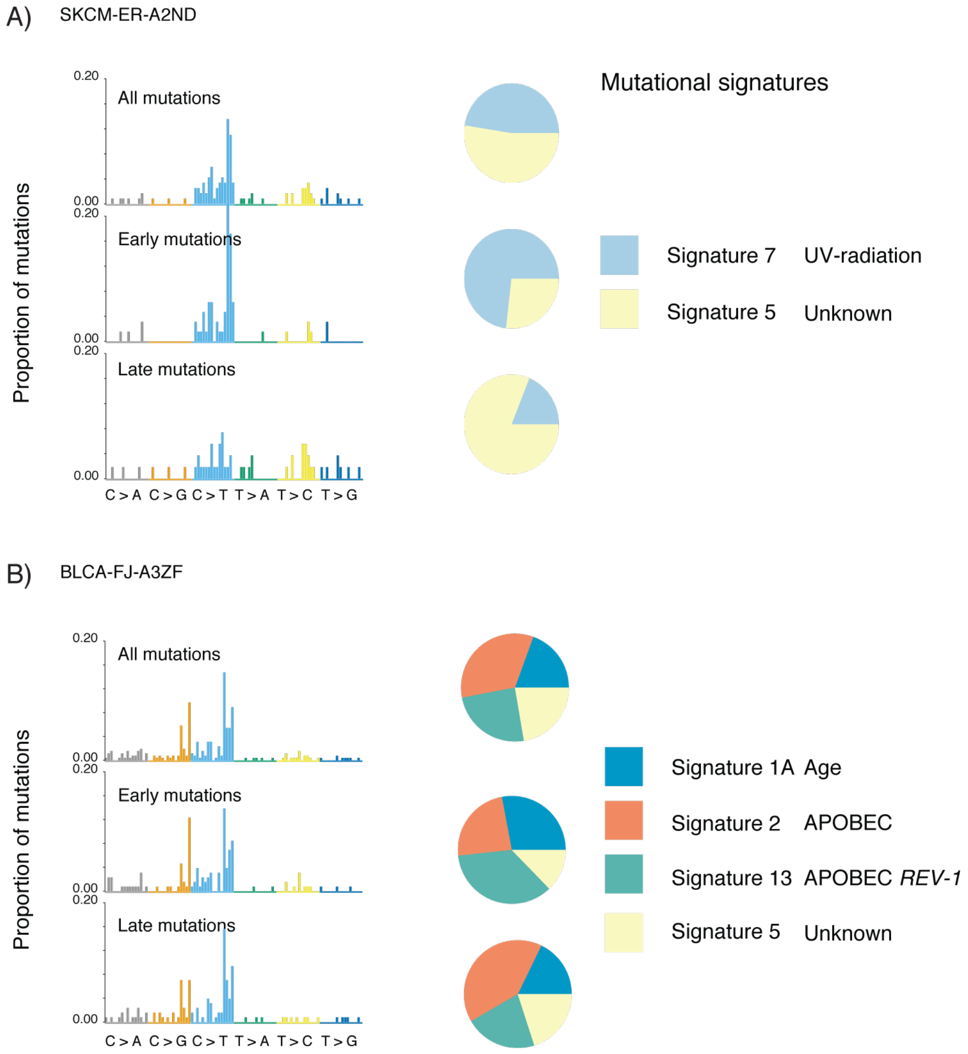


Figure 5:12 Mutational signatures in a SKCM and a BLCA tumour

For two tumours A) SKCM-ER-A2ND; B) BLCA-FJ-A32F mutational spectra, showing proportion of mutations at each of the 96 tri-nucleotide contexts is shown for all mutations, early mutations and late mutations. In addition, mutational signatures for different mutational sets are shown.

Similar to Signature 4, Signature 7, related to mutations caused by exposure to UV, was elevated in early compared to late mutations in SKCM tumours (*P*=5.67e-11;Figure 5:7; Figure 5:12 A). In SKCM, it is also worth noting that most samples were derived from metastatic sites, thus many of the later mutations may have been acquired when tumour cells were no longer exposed to UV light.

5.2.8 APOBEC mediated mutagenesis is not merely a transient event

It is important to note that, when detected in tumours, Signature 2 was most frequently identified in both early and late mutations, albeit often to a lesser degree in early mutations. This suggest that although APOBEC may dominate later in tumour evolution, its activity is not transient and it does not represent an historical relic within the tumour genome, active at only one point in time during the disease course (Roberts and Gordenin, 2014, Swanton et al., 2015).

5.2.9 Deciphering dynamics of mutational processes in oesophageal cancer using multi-region sequencing

To further explore the temporal dynamics of mutational processes in cancer genome evolution, multi-region sequencing data from oesophageal adenocarcinoma was used (Table 3-2). As in multi-region NSCLC, the mutational spectra of early (trunk) and late (branch) mutations were compared for each tumour. In every case, significant shifts in the mutational spectra over time were observed. Notably, in 6 out of 8 cases, a trend for a decrease in the proportion of T>G mutations was observed.

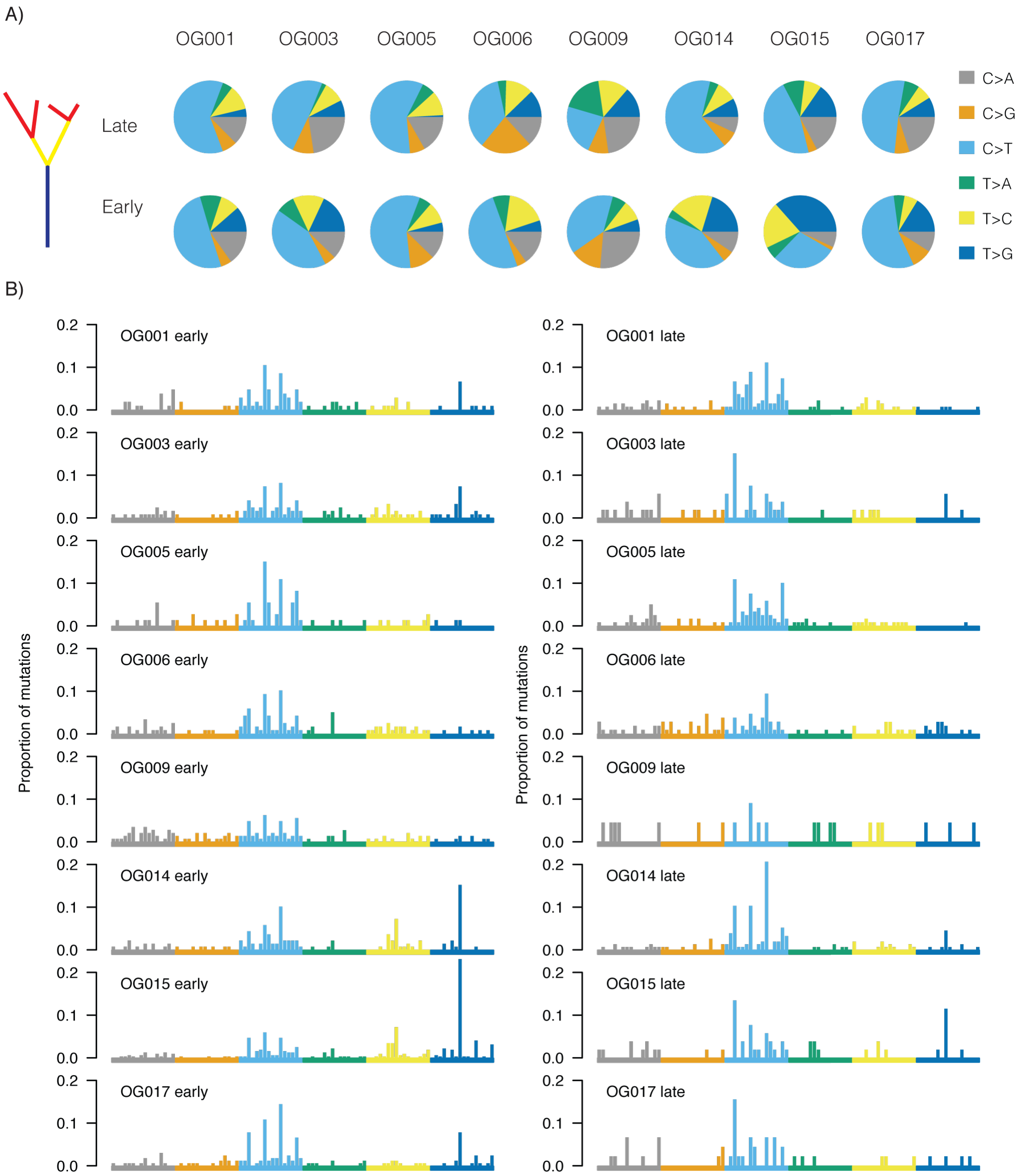


Figure 5:13 Temporal dissection of mutations in oesophageal adenocarcinoma

A) Proportions of the six base substitutions is shown for all eight oesophageal adenocarcinoma tumours for both early and late base substitutions. B) Proportions of base substitutions within each tri-nucleotide context, for early and late mutations, are shown. Notably, five of the early oesophageal adenocarcinomas are characterized by T>G mutations occurring specifically within a CpTpT context.

In the majority of cases, trunk mutations were characterized by a strong enrichment for T>G transversion mutations occurring with a CpTpT context (Figure 5:13). When focussing on late mutations, however, a relative decrease in the proportion of T>G mutations occurring within a CpTpT context was observed, accompanied by a relative increase in C>T transition mutations at CpG sites. To formally assess the enrichment for T>G transversion mutations and T>C transition mutations occurring within a CpTpT context, a test was devised in order to compare the prevalence of these mutations compared to random mutagenesis. Significant enrichment for this mutational signature was identified in five out of eight tumours, and in general, it was found to decrease over time (Figure 5:14).

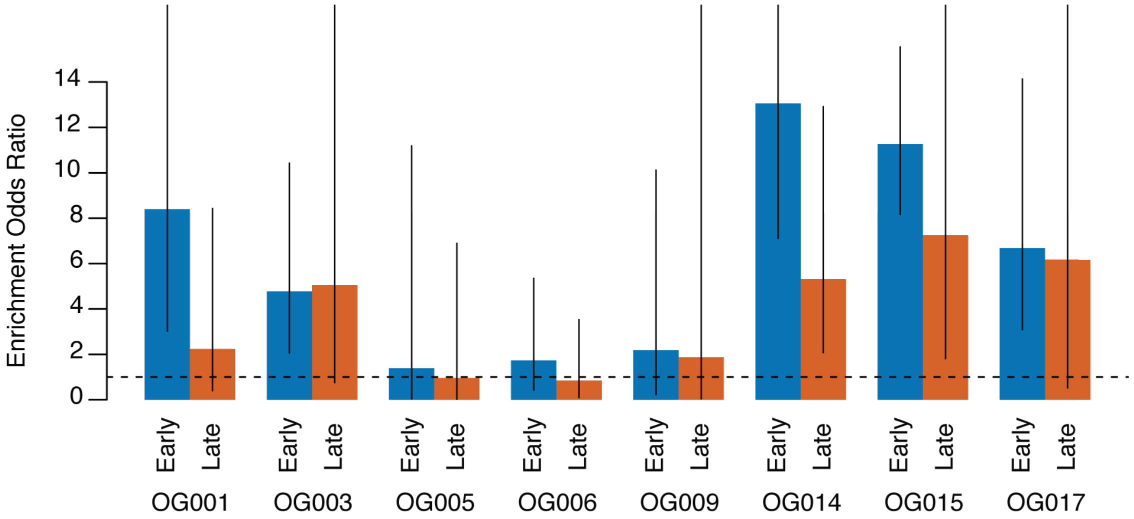


Figure 5:14 Enrichment of T>G and T>C mutations at CpTpT sites

Mutation enrichment odds ratio for early (trunk, blue bars) and late (branch, red bars) mutations for multi-region NSCLC samples. The 95% confidence intervals for Fisher's exact test are indicated.

While T>G mutations occurring within a CpTpT context have been linked to acid reflux (Dulak et al., 2013, Weaver et al., 2014, Nones et al., 2014), C>T transition mutations at CpG sites, as previously discussed, likely correspond to spontaneous deamination of methylated cytosines (Yates and Campbell, 2012). Thus, these data suggest that founder genomic events occur within an environmental context, possibly attributable to exposure to gastric acid, but this is not the major process contributing to mutational diversity in the growing neoplasm.

5.2.10 Differences between oesophageal adeno and squamous cell carcinoma

To validate the temporal dynamics of observed mutational processes in oesophageal adenocarcinoma and compare the differences between oesophageal adenocarcinoma and oesophageal squamous cell carcinoma, data from TCGA was used. For this analysis, mutations signatures were deciphered using the R package deconstructSigs, developed by Rachel Rosenthal.

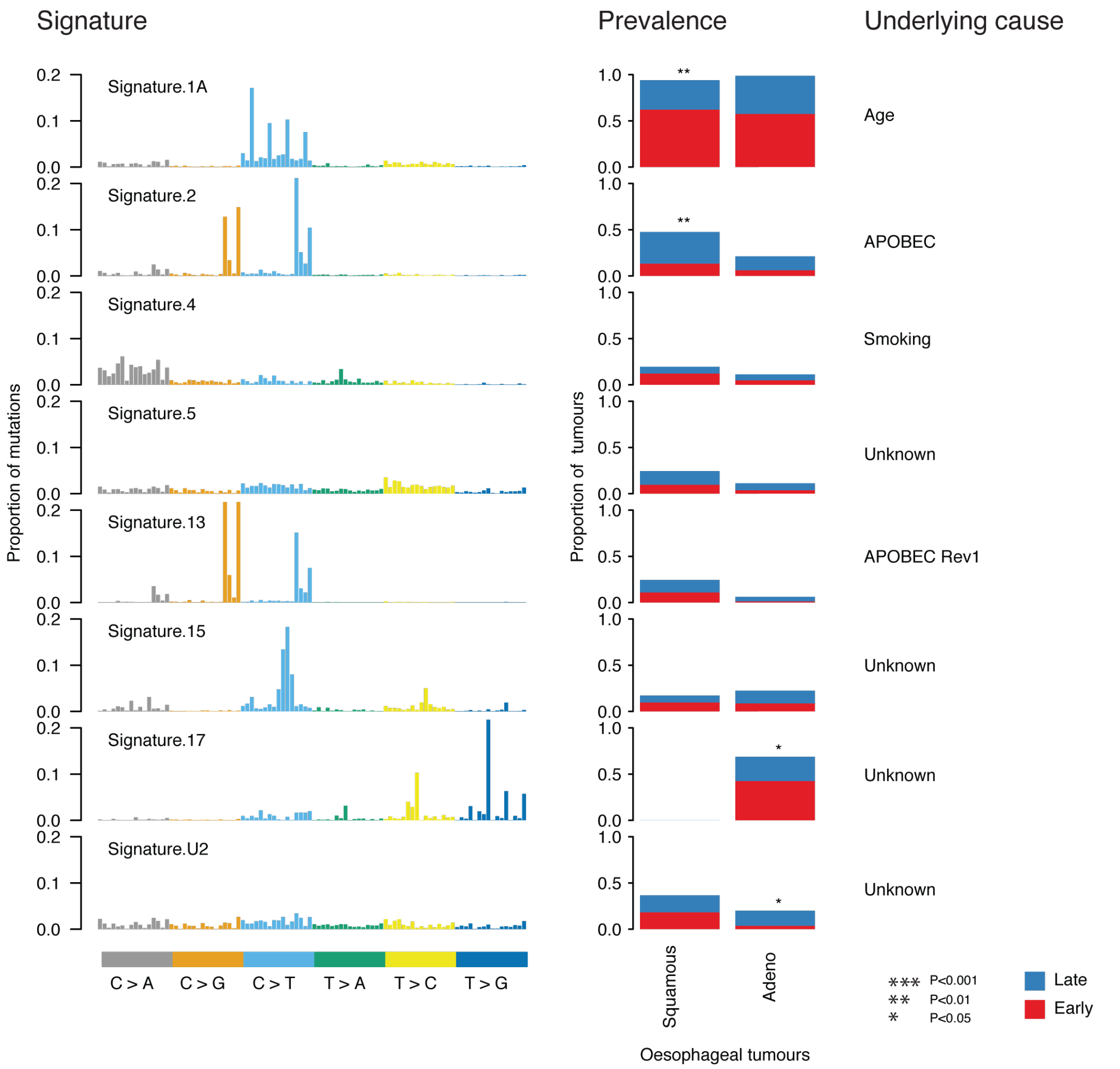


Figure 5:15 Temporal dissection of mutational signatures in oesophageal squamous and adenocarcinoma tumours

Mutational signatures identified across oesophageal adeno and squamous cell carcinoma and their temporal dissection. Signature names correspond to those identified by Alexandrov (2013a). For each signature present in a cancer type, the proportion of tumours where the signature is prominent early (red), compared to more prominent late (blue) is shown.

Signature 17, characterized predominantly by T>G transversion mutations occurring with a CpTpT context, was identified in 60 out of 82 patients with oesophageal adenocarcinoma. Consistent with multi-region sequencing data, in the majority of patients in which this signature was identified, it was found to decrease over time. A notable example was ESCA-2H-A9GR, a tumour that exhibited a genome-doubling event. A marked decrease in the proportion of mutations corresponding to Signature 17 was observed after the genome doubling, accompanied by an increase in the proportion of mutations classified as Signature 1A. However, a small number of patients were also identified in which the reverse was observed. For instance, ESCA-2H-A9GR, exhibited a trend for Signature 17 to increase at the expense of Signature 1A (Figure 5:16 B).

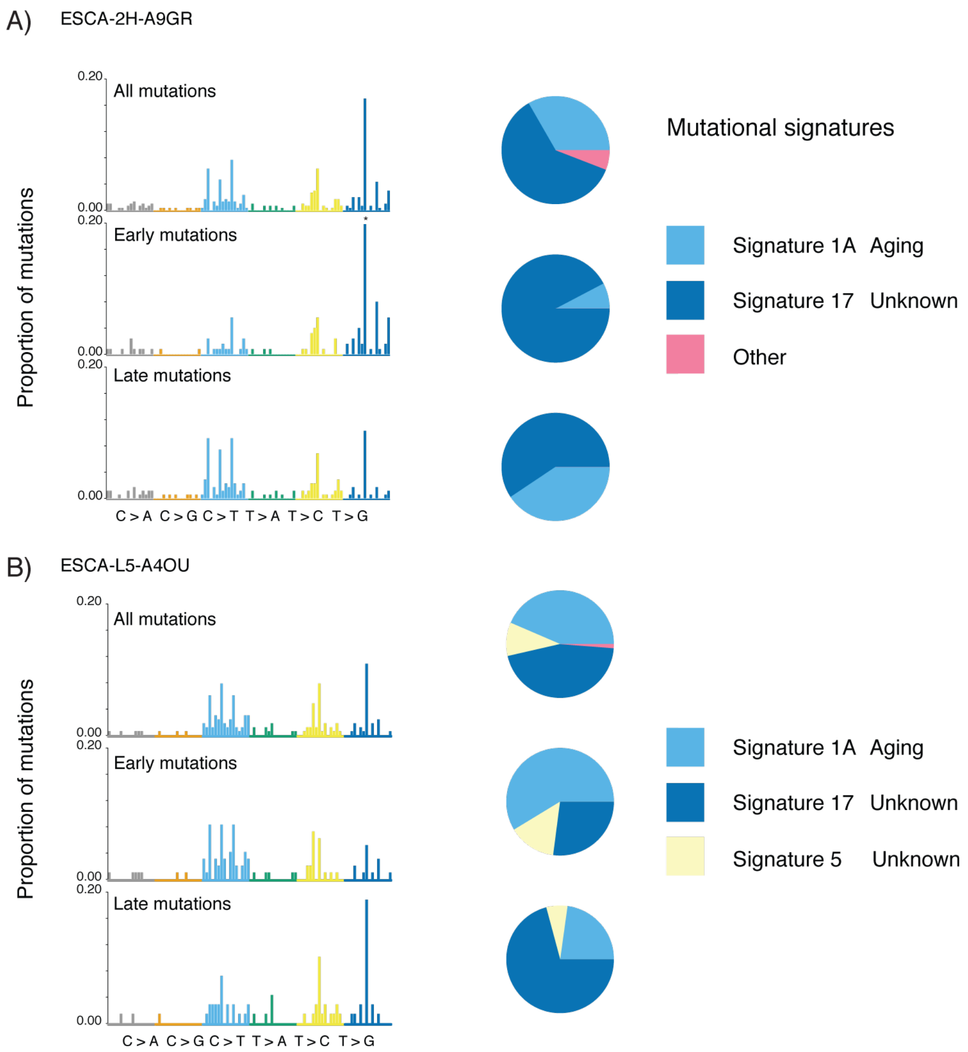


Figure 5:16 Example of temporal shift in mutational signature in oesophageal adenocarcinoma using TCGA

A, B) Left panel shows mutational spectra of all, early and late mutations in an oesophageal adeno (A) and squamous cell carcinoma (B). Proportion of different mutational signatures is shown in right panel.

Signature 17 was not detected in oesophageal squamous cell carcinoma tumours, highlighting the differences between these two types of tumour. Another notable differences between the two cancer types was that a significantly higher proportion of oesophageal squamous cell carcinoma tumours were found to exhibit APOBEC mediated mutagenesis (*P*=0.011, Fisher's Exact test). Similar to LUAD, HNSC, and oestrogen negative BRCA, the proportion of APOBEC-related mutations was found to be significantly elevated in later compared to early mutations in oesophageal squamous cell carcinomas (*P*=0.004), and a similar trend was observed for oesophageal adenocarcinoma (*P*=0.09).

5.2.11 A signature of platinum therapy in oesophageal adenocarcinoma

Therapy may at as an exogenous source of mutations, leaving scars in the cancer genome, which are subject to selection and neutral evolution (Ding et al., 2010, Johnson et al., 2014). To assess whether platinum therapy directly influenced cancer evolution by providing additional mutagenic fuel, the mutational spectra and clonal composition of oesophageal adenocarcinoma tumours were investigated following treatment with chemotherapy.

When considered in aggregate, statistically significant shifts in the mutational spectra were observed between pre and post-chemotherapy somatic mutations. Specifically, a significant decrease in the proportion of C>T transition mutations was observed, accompanied by an increase in C>A transversion mutations. Furthermore, within post-chemotherapy somatic mutations, C>A transversion mutations were found to preferentially occur within a CpC context, more than would be expected given random mutagenesis. This mutational signature was also identified in *C.elegans*, following platinum treatment (Meier et al., 2014).

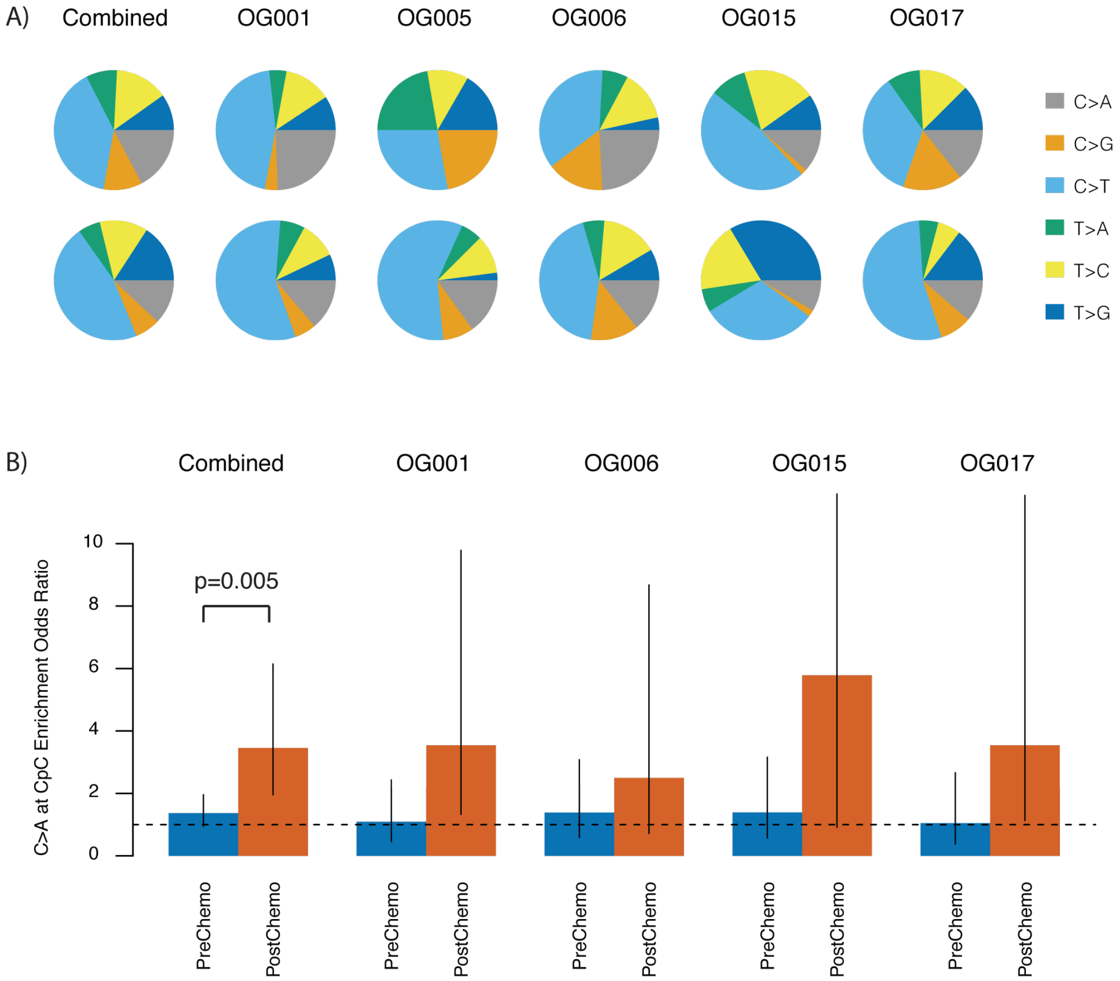


Figure 5:17 Mutational spectra pre and post-platinum therapy

A) Pie-chart showing the fraction of pre-chemotherapy mutations and post-chemotherapy mutations accounted for by each of the six mutation types in all M-seq samples. B) Barplot showing platinum signature enrichment odds ratio for pre-chemotherapy (blue bars) and post-chemotherapy (red bars) mutations. The platinum signature encompasses C>A in CpC context (25). 95% confidence intervals for Fisher's exact test are indicated.

These data are indicative of mutations being directly attributable to exposure to platinum therapy. On average, of the mutations identified after treatment with platinum therapy, 10% were C>A transversions occurring within a CpC context.

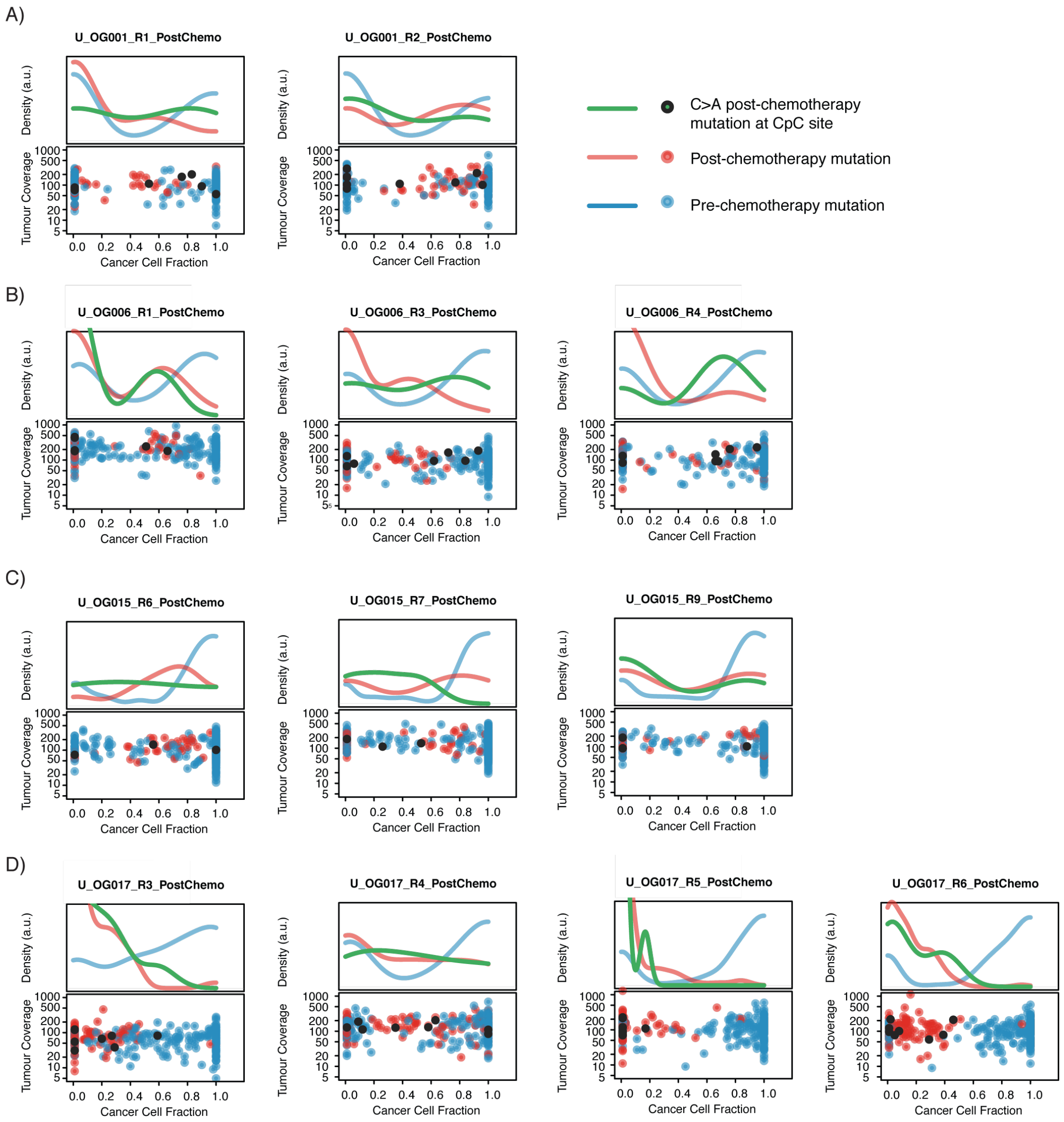


Figure 5:18 Clonal architecture of tumour regions following platinum therapy

For four patients with C>A transversion mutations in post-chemotherapy regions, the cancer cell fraction of mutations pre and post chemotherapy are shown. C>A mutations occurring at CpC sites are highlighted. Pre-chemotherapy mutations are more likely to be clonal compared to post-chemotherapy mutations.

Almost all of these transversions were found to be present at cancer cell fractions below 100%(28/32) in all tumour regions within a given tumour (Figure 5:18). Indeed, as a whole, post-chemotherapy mutations were very rarely clonal in any tumour region, with less than 3% identified as having a CCF of 100% in any tumour region. By contrast, over 50% of mutations identified prior to adjuvant chemotherapy were clonal in at least one tumour region.

Taken together, these data suggest a lack of a full clonal sweep following adjuvant chemotherapy and minimal evolutionary bottlenecking. Moreover, it appears chemotherapy may foster subclonal diversity, although further tracking of tumours over time will be required to fully elucidate whether the size of clones is driven by selection.

5.2.12 Mutational processes fuel the acquisition of somatic events in cancer genes

Next, the extent to which mutational processes can explain the temporal acquisition of non-silent mutations in known cancer genes was explored across cancer types. In both LUAD and LUSC TCGA tumours, over 30% of clonal non-silent mutations in cancer genes were found to be C>A transversions, consistent with smoking providing the mutagenic fuel for tumourigenesis. For example, over 40% of clonal mutations in *TP53* were C>A transversions. In keeping with these results, in multi-region NSCLC L011, a clonal C>A *TP53* mutation was detected in all tumour regions. In BLCA, by contrast, over 40% of clonal mutations in cancer genes were found to occur in an APOBEC context, consistent with APOBEC-mediated mutagenesis shaping the early evolutionary trajectory of many BLCA tumours. For example, BLCA-GU-A42R harboured two clonal APOBEC mutations in *ARID1A*.

Focusing on subclonal mutations in known cancer driver genes, APOBEC represented a dominant mutational process in at least five cancer types: BLCA, BRCA, HNSC, LUAD, and LUSC. In LUAD, LUSC, and HNSC samples showing evidence of APOBEC-mediated mutagenesis, on average only 21% (range 19-24%) of clonal mutations in cancer driver genes occurred in an APOBEC context, but over 45% (range 35-59%) of subclonal mutations in cancer genes could be explained by APOBEC-mediated mutagenesis. Strikingly, in these cancers, over 90% of subclonal mutations in *PIK3CA* occurred in an APOBEC context, subclonal APOBEC mutations were identified in multiple other cancer driver genes, including

PTEN, EGFR, and *TP53*. Likewise, in multi-region NSCLC, subclonal mutations in driver genes *PIK3CA, EP300, TGFBR1* and *AKAP9* were identified, and, in APOBEC-associated BLCA, over 45% of subclonal mutations in driver genes occurred in an APOBEC context. These data suggest that APOBEC cytosine deaminases may play a key role in driving subclonal diversification in these cancer types.

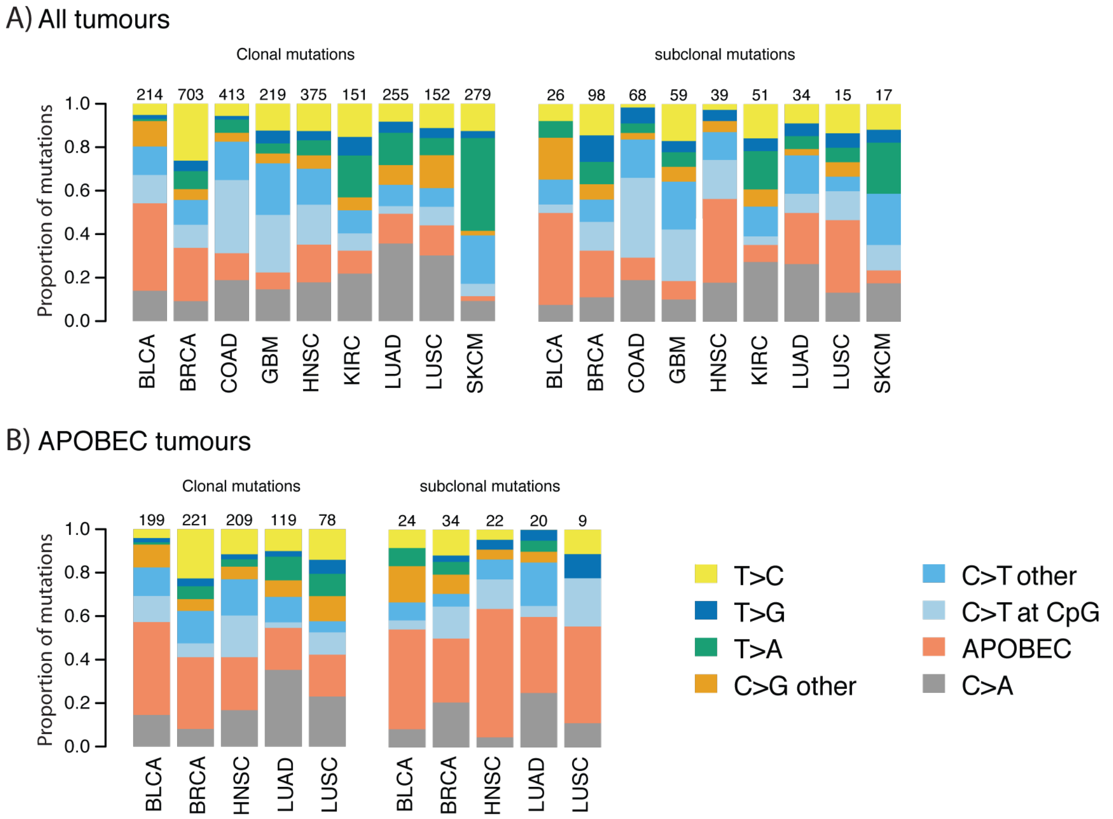


Figure 5:19 Mutational spectra of mutations in driver genes.

A) Proportion of 8 distinct mutation types is shown for clonal and subclonal mutations within cancer genes in each tumour type. APOBEC refers to C>T or C>G mutations at TpCp(CorT) sites. B) Proportion of mutation types for samples showing a significant enrichment of APOBEC mediated mutagenesis. Notably, in HNSC, LUAD, and LUSC there is a clear enrichment for APOBEC mutations in subclonal cancer driver genes compared to clonal mutations.

5.2.13 A model of the evolutionary history of NSCLC

Finally, to gain a deeper understanding of the dynamics of early stage NSCLC evolution, two tumours with high coverage whole-genome sequencing were considered in detail (Figure 5:20). The mutational processes shaping the cancer genome, both before and after the emergence of the most-recent common ancestor, were placed on the phylogenetic trees of theses tumours.

In L002, a current-smoker, tobacco carcinogens clearly played a significant role early in tumour development, with over 39% of the truncal mutations being C>A transversion mutations. Prior to the emergence of the most recent common ancestor, mutations in multiple driver genes were acquired, including mutations in *TP53* and *CHD8*.

After diversification, into a LUAD subclone and a LUSC subclone, APOBEC mutational processes were elevated specifically in the LUAD region. Despite continual exposure to tobacco carcinogens, on the LUAD branch only 22% of mutations were C>A transversions and driver mutations in *TGFBR1* and *PTPRD* were found to occur within an APOBEC context. These data are consistent with APOBEC providing the mutagenic fuel for subclonal expansions. Furthermore, the LUAD branch of the phylogenetic tree harboured almost 10,000 more mutations than the LUSC branch. In the LUSC region 28% of mutations were C>A transversions, suggesting smoking remained a key genomic instability process shaping the genome on this branch of the phylogenetic tree. Genome doubling events were detected in both tumour regions, in both cases occurring towards the end of the branches.

In L008, an ex-smoker who stopped over 20 years ago (Table 1), further temporal resolution was gained by exploring the mutations before and after the truncal genome-doubling event. Notably, every truncal driver mutation, including mutations in *TP53* and *BRAF*, likely occurred before the genome-doubling event. A signature of tobacco smoke, characterized by over 30% of mutations being C>A transversions, was identified both before and after the genome-doubling event.

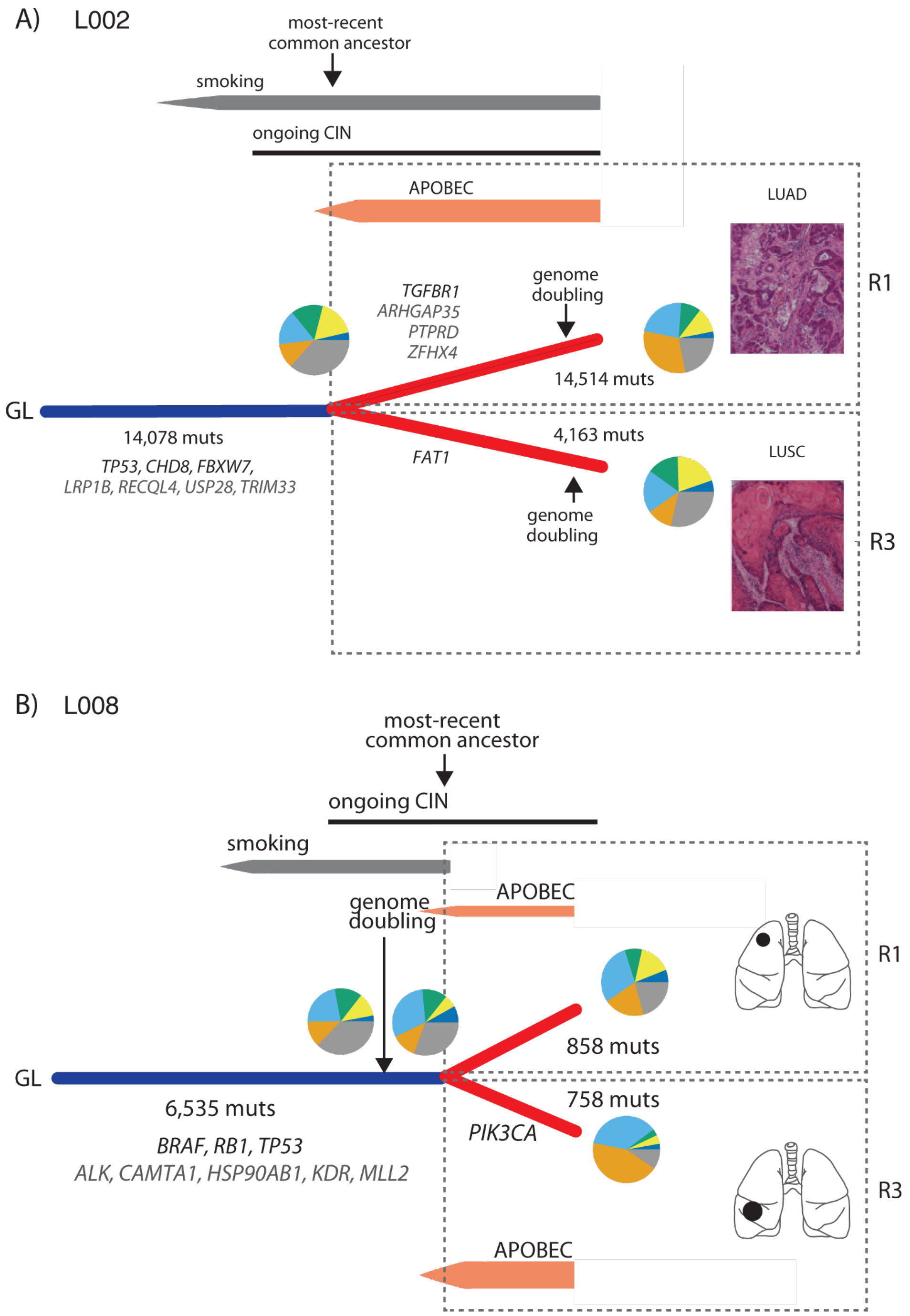


Figure 5:20 A model of the evolutionary history of NSCLC.

Evolutionary histories of tumours from patients L002 (A) and L008 (B) are depicted. Genomic instability processes defining NSCLC evolution have been placed on their phylogenetic trees. In each case, the timing of genome-doubling events is indicated with an arrow. CIN, chromosomal instability; muts, mutations.

After the emergence of the most recent common ancestor, only 21% and 9% of mutations in the upper lobe, region R1 and middle lobe, region R3, were C>A transversions. These data imply genome doubling occurred within a smoking carcinogenic context, and, by extension therefore, over 20 years ago. In support of a prolonged tumour latency period, following genome doubling prior to clinical detection in NSCLC, the genome-doubling event in L001 also appeared to occur within a smoking context, over 20 years prior to surgery.

5.3 Conclusions

Analysis of the catalogues of somatic mutations from both multi-region and single samples tumour genomes has yielded several insights into the underlying mutational processes that have sculpted the cancer genomes over space and time.

In most cancers while known mutagenic processes, such spontaneous deamination of methylated cytosines and smoking induced C>A transversions, appear to dominate the early life history of a tumour, other processes, such as APOBEC mediated mutagenesis, frequently dominate later, driving subclonal diversification. Indeed, APOBEC mediated mutagenesis could be linked the acquisition of multiple subclonal driver events in HNSC, LUAD, LUSC, BLCA, and BRCA. Thus these data fit with a model of APOBEC mediated mutagenesis providing a substrate for tumour evolution, which may lead to intra-tumour heterogeneity

Moreover, these findings are indicative of an elevated mutation rate in most cancers, characterized by multiple mutational processes that specifically arise after or during tumour formation, and support the notion of a mutator phenotype defining tumour evolution (Loeb, 2001).

Analysis of pre- and post-chemotherapy oesophageal adenocarcinoma tumours identified a platinum signature following therapy, highlighting how therapy itself can directly impact upon the evolutionary trajectory of tumours. Two distinct modes of tumour evolution were deciphered following therapy, with both paraphyly and monophyly observed (Figure 5:21). Prior to chemotherapy a decrease in mutations potentially attributable to gastric acid exposure was observed, at the expense of an increase in C>T mutations at CpG sites, potentially indicating increased cell-turnover and proliferation after tumour formation (Figure 5:21), consistent with previous findings in KIRC (Gerlinger et al., 2014a).

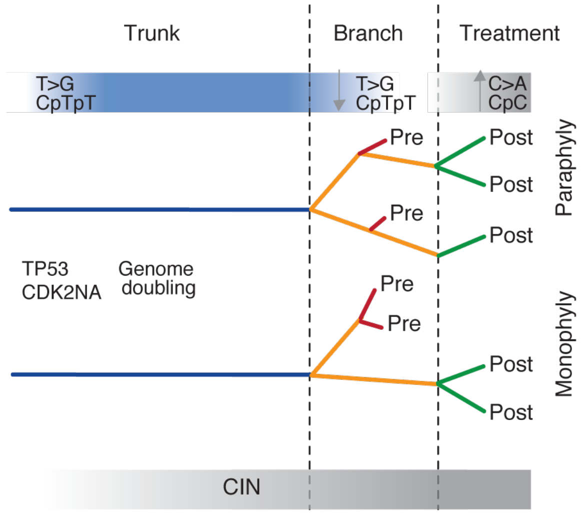


Figure 5:21 A model of tumour progression in oesophageal adenocarcinoma
Following treatment, two distinct modes of tumour evolution were observed, paraphyly and monophyly.

While extremely useful in deconstructing the mutational processes present in large catalogues of mutation from hundreds of cancers, it is also worth considering some of the limitations of NMF and model selection to identify the mutational signatures. Most notably, the process makes the assumption that every mutational signature is present to some degree in all samples, which can mean signatures with little effect may be missed, or dominant signatures within specific samples may be over estimated across the cohort. In the future it will be worth exploring novel computational tools that reduce parameter space, and moreover tools that search for signatures that allow SNVs, copy number and epigenetic space to be evaluated.

Understanding how the mutations derived from the vast array of mutational processes can be exploited in the clinical setting and the clinical significance of intra-tumour heterogeneity revealed in this thesis will be explored further in the next chapter.

Chapter 6. Neo-antigens and intra-tumour heterogeneity

6.1 Introduction

In the previous chapters of this thesis I have investigated the extent of diversity within tumours and explored the processes driving cancer genome evolution and intra-tumour heterogeneity and also considered their timing. In this chapter, I explore how cancer genome sequencing data can be used to inform immunotherapy approaches and the clinical relevance of intra-tumour heterogeneity.

Given the dynamic and heterogeneous nature of tumour evolution, an optimal therapy may need to be highly personal, customized based on the specific mutational processes and somatic events present in a patient’s tumour, ideally targeting multiple clonal events in each tumour. As demonstrated in Chapter 3, targeted therapy approaches may be hampered by the fact that many actionable mutations are frequently only present in a subset of cancer cells within a tumour.

Recent studies have shed light on the immunotherapeutic potential of immune-blockade and tumour neo-antigens. Studies in both mice and humans have demonstrated T-cell responses elicited towards neo-antigens (Castle et al., 2012, Rizvi et al., 2015, Linnemann et al., 2015), and a recent study suggested the presence of tumour neo-antigens might lead to improved overall survival (Brown et al., 2014). Relatedly, modulation of the immune system through blockade of the immune inhibitory receptors PD-1, PD-L1 or CTLA-4 has been shown to produce significant clinical benefits against a variety of cancers, including NSCLC (Jamal-Hanjani et al., 2013, Sharma and Allison, 2015, Soria et al., 2015). However, further work is needed to fully understand the repertoire of tumour neo-antigens in tumours, and an understanding of neo-antigens in the context of intra-tumour heterogeneity is lacking.

In this chapter, I develop a pipeline to decipher the neo-antigen repertoire in tumours. The extent to which neo-antigens may be heterogeneous, and the clinical relevance of this to checkpoint inhibitor response, is explored in NSCLC.

The work presented in this chapter of the thesis is currently in preparation for submission. Some of the bioinformatics work presented here was conducted in collaboration with Rachel Rosenthal, while the experimental work was performed by Andrew Furness and collaborators at the Danish Technical University.

6.2 Results

6.2.1 A pipeline to identify putative neo-antigens

Each mutation that results in an altered protein product has the potential to give rise to a tumour neo-antigen, thereby eliciting an immune response. In order to identify putative neo-antigens in NSCLC a pipeline was devised. In brief, the pipeline incorporates the following steps:

1. Determine patient's HLA class I alleles using Optitype (this part of the pipeline was developed by Rachel Rosenthal)

2. Identify amino acid substitutions associated with each coding non-synonymous mutation

3. Create mutant peptide sequencing of 9-11 amino acids in length, including every possible position for altered amino acid to sit. Also create equivalent wild-type peptides.

4. Determine binding of each peptide to patient's HLA-class I alleles using netMHCpan (Hoof et al., 2009). netMHCpan relies on a neural network-based learning approach to predict binding affinity of peptides to class I MHC molecules

5. Evaluate binding scores of mutant and wild-type peptides to obtain a filtered list of putative neo-antigens

6.2.2 Considerable variation in number of putative neo-antigens across LUAD tumours

To explore the presence of putative neo-antigens across a large cohort of tumours, data from the TCGA was used. Application of the pipeline to 124 LUAD and 124 LUSC TCGA tumours revealed the presence of putative neo-antigens in the majority of tumours (Figure 6:1). Consistent with previous reports, each LUAD tumour harboured on average 122 putative neo-antigens, while each LUSC tumour an average of 142 (Rajasagi et al., 2014). Notably, a considerable range in neo-antigens per tumour was observed, especially in LUAD tumours (Figure 6:1A), with a median absolute deviation of 82 and interquartile range of 112.

As might be expected, the number of predicted neo-antigens was significantly correlated with the mutational burden of the tumour (*P*< 2.2e-16, r= 0.95).

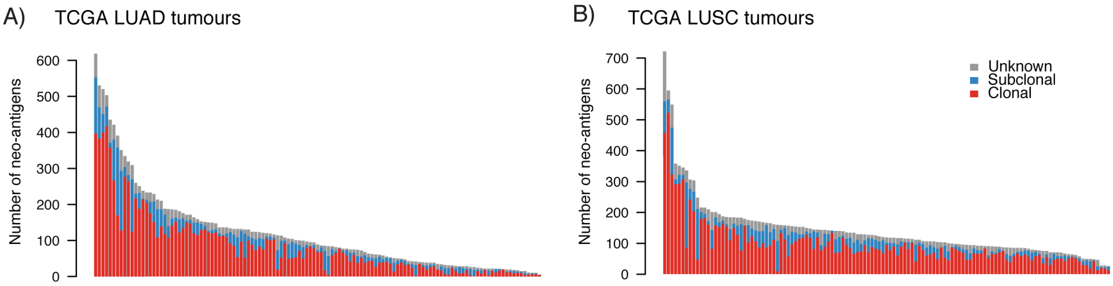


Figure 6:1 Putative neo-antigens in TCGA LUAD and LUSC tumours
Number of putative neo-antigens identified in (A) TCGA LUAD and (B) TCGA LUSC tumours is depicted. Each bar represents one tumour, and neo-antigens are labelled as deriving from clonal, subclonal or mutations of unknown clonal status.

6.2.3 Relationship between survival and number of putative neo-antigens in LUAD tumours

Conceivably, the presence of a large number of tumour neo-antigens may impede the ability of a tumour to successfully evade the immune system through immune escape (Schumacher and Schreiber, 2015). Accordingly, one would predict a high number of tumour neo-antigens might be associated with improved patient prognosis. To assess this hypothesis, matched clinical data for TCGA samples was obtained.

In support of the clinical importance of neo-antigens (Brown et al., 2014), in the LUAD cohort, a high neo-antigen load (defined as the upper quartile of the number of putative neo-antigens in the cohort) was associated with significantly longer overall survival times compared to tumours in the remaining three quartiles (Figure 6:2, *P*=0.011, log-rank test). For tumours in the LUSC cohort, which exhibited a narrower range of putative neo-antigens (with a median absolute deviation of 50 and interquartile range of 71), a statistically significant association between overall survival and neo-antigen load was not observed (Figure 6:2, *P*=0.81, log-rank test).

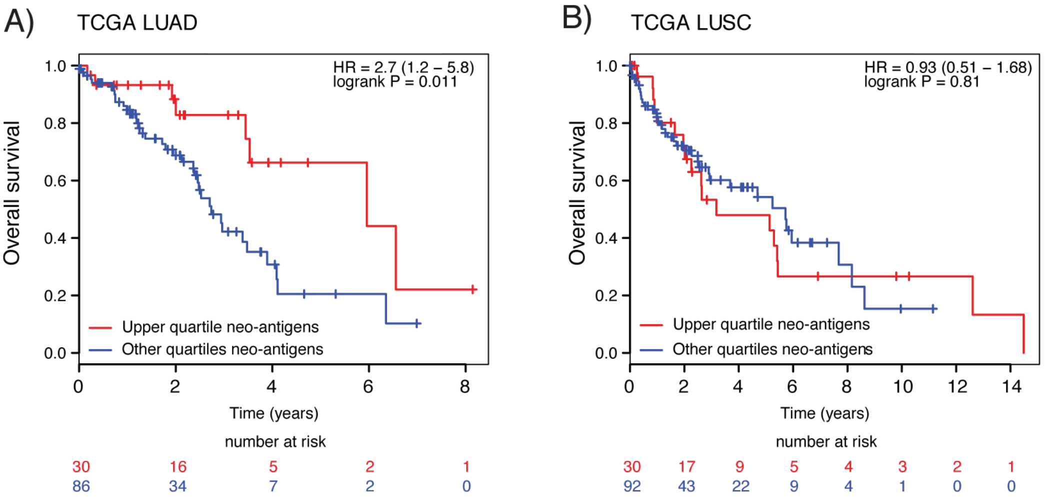


Figure 6:2 Relationship between survival and neo-antigenic load

Overall survival for (A) TCGA LUAD tumours and (B) TCGA LUSC tumours comparing tumours with a high neo-antigenic load (upper quartile) versus all other quartiles. For TCGA LUAD, tumours with a high neo-antigens load have significantly improved survival (*P*=0.011, log-rank test). For TCGA LUSC, no relationship between neo-antigenic load and survival was observed (*P*=0.81, log-rank test).

6.2.4 Relationship between survival and number of clonal putative neo-antigens in LUAD tumours

Next, for TCGA LUAD tumours, the clonal status of each putative neo-antigen was determined and the relationship between the number of clonal neo-antigens and overall survival was considered (Figure 6:3).

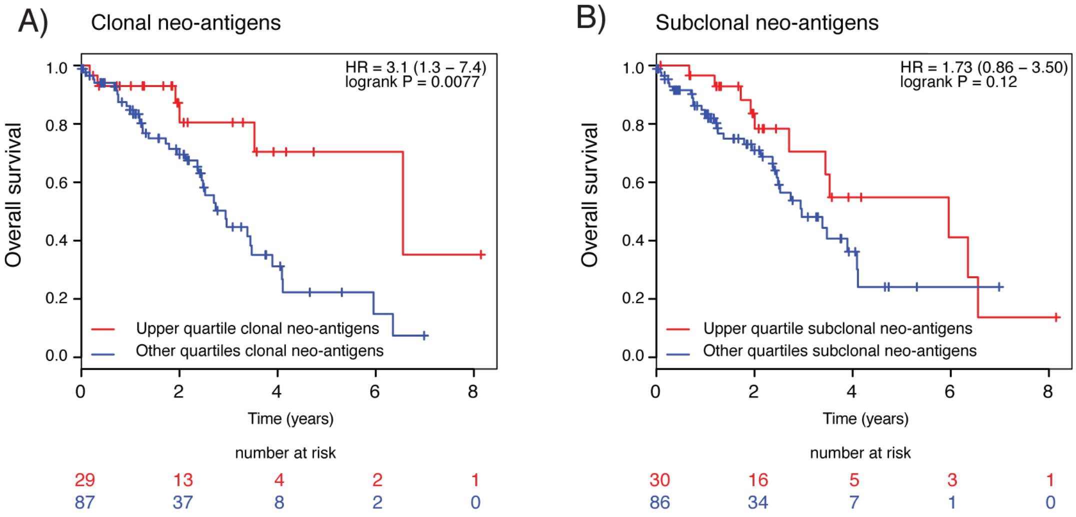


Figure 6:3 Clonality and number of epitopes in relation to overall survival in NSCLC

A) Overall survival for TCGA LUAD tumours with a high clonal neo-antigenic load (upper quartile) versus lower clonal neo-antigenic load (all other tumours). B) Overall survival for tumours with a high subclonal neo-antigenic load (upper quartile) versus lower subclonal neo-antigenic load (all other tumours).

LUAD tumours harbouring a high number of predicted clonal neo-antigens (defined as the upper quartile of the cohort) were associated with significantly longer overall survival compared to all other tumours in the cohort (Figure 6:3, *P*=0.0077, log-rank test). Conversely, the number of predicted subclonal neo-antigens was not significantly associated with overall survival (Figure 6:3, *P*=0.12, log-rank test). The relationship between clonal neo-antigens and overall-survival also remained significant in a multivariate analysis, including stage and age in the model (*P*=0.00389, Table 6-1)

Table 6-1 Clonal neo-antigen survival analysis

Variable	Univariate HR (95%)	p-val (log-rank)	Multivariate HR (95%)	p-val (Wald)
Clonal neo-antigens (other vs. upper quartile)	3.08 (1.29-7.36)	0.0077	3.96 (1.55-10.08)	0.0039
Age (continuous)	1.01 (0.98-1.04)	0.67	1.01 (0.98-1.04)	0.51
Stage II (vs. I)	1.77 (0.78-4.01)	0.17	1.74 (0.75-4.04)	0.20
Stage III (vs. I)	4.03 (2.04-7.99)	6.38e-05	5.141 (2.50-10.59)	8.97e-06
Stage IV (vs. I)	2.64 (0.60-11.60)	0.20	2.49 (0.54 - 11.30)	0.24

While a significant association was also observed between high clonal SNV load and poor prognosis (*P*=0.018, log-rank-test, HR=2.0 [1.1 − 3.7]), this association was less evident compared to clonal neo-antigenic burden, suggesting the estimated neo-antigenic burden is not simply a proxy for SNV load.

Finally, tumours with a high burden of clonal neo-antigens were found to be significantly more homogenous compared to other tumours in the cohort (Figure 6:4, P=0.005, Student's t-test).

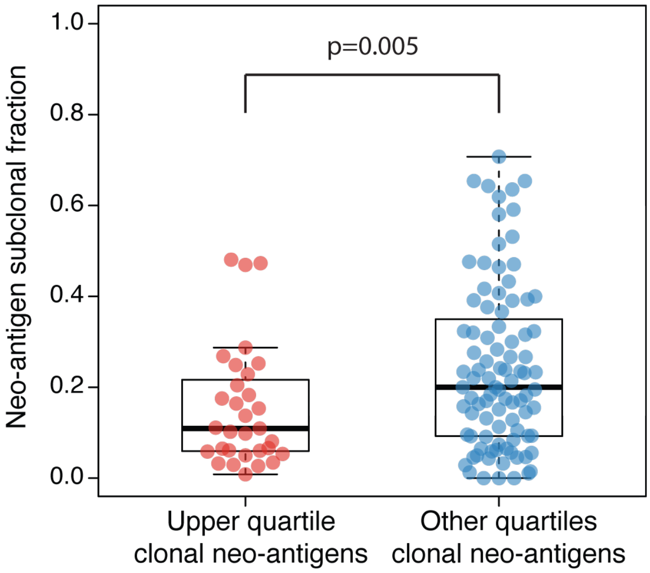


Figure 6:4 Relationship between number of clonal neo-antigens and neo-antigen subclonal fraction in TCGA LUAD tumours

Boxplot showing difference in neo-antigen subclonal fraction between tumours harbouring a high number of clonal neo-antigens (upper quartile) and those harbouring intermediate to intermediate to low numbers of clonal neo-antigens (remaining quartiles). A significant difference is observed between the two groups, *P*=0.005, Student's T-test.

Taken together, these data suggest the number and clonal status of neo-antigens may impact upon the likelihood of immune escape, and thereby patient prognosis. However, additional samples are needed to reliably assess the importance of clonal neo-antigens in relation to survival. Moreover, given the myriad of potential mechanisms leading to immune escape, and the many checkpoints of the immune system, it is likely that many tumours with clonal neo-antigens still eventually evade the immune system.

6.2.5 Identification of putative neo-antigens in multi-region tumours

Next, to address whether CD8$^+$ T cells reactive to clonal neo-antigens could be identified in primary NSCLC tumours, the neo-antigen pipeline was implemented on two multi-region sequenced tumours, L011 and L012 (Figure 6:5).

While both tumours were derived from female heavy smokers (>40 pack-years), their mutation burden and the extent of their intra-tumour heterogeneity was considerably different (Figure 6:5 A). L011, an adenocarcinoma, exhibited a relatively homogenous primary tumour and metastatic dissemination to the brain (M1-M4), likely originating from tumour region R3 (Figure 6:5 A). A total of 313 neo-antigens were predicted from the mutations contained within the primary tumour of L011, 88% of which were clonal, identified in every region of the primary tumour (Figure 6:5 B). Conversely, L012, a squamous cell carcinoma, exhibited a relatively low mutation burden and extensive heterogeneity, with 75% of the predicted neo-antigens deriving from subclonal mutations, only present in a subset of cancer cells (Figure 6:5 A,C).

6.2.5.1 Confirmation of predicted neo-antigens

To assess whether predicted neo-antigens resulted in a T-cell response, MHC-multimers were loaded with predicted neo-antigens and screened for CD8$^+$ T cells expanded from different tumour regions and adjacent normal lung tissue. In total, 290 and 355 putative neo-antigens (with predicted HLA binding affinity <500nM, including multiple potential peptide variations from the same missense mutation) were synthesized and used to screen expanded L011 and L012 tumour infiltrating lymphocytes (TILs) respectively. This experimental work was performed by Andrew Furness.

In L011, CD8$^+$ T cells reactive to an HLA-B3501-multimer derived from mutant *MTFR2*^{D326Y} (FAFQE***Y***DSF), a clonal mutation with high predicted HLA binding to HLA-B3501 in wild type (10nM) and mutant (22nM) forms (Figure 6:5 B), were identified in all tumour regions (2.8-4.4%) and at lower frequency in normal regions (0.1%) (Figure 6:5 D). No CD8$^+$ T cells reactive to the wild type multimer with sequence FAFQE**D**DSF were identified. Likewise, no responses were found against multimers with overlapping mutant peptides AFQE**Y**DSFEK and KFAFQE**Y**DSF, suggesting CD8$^+$ T cells were highly specific.

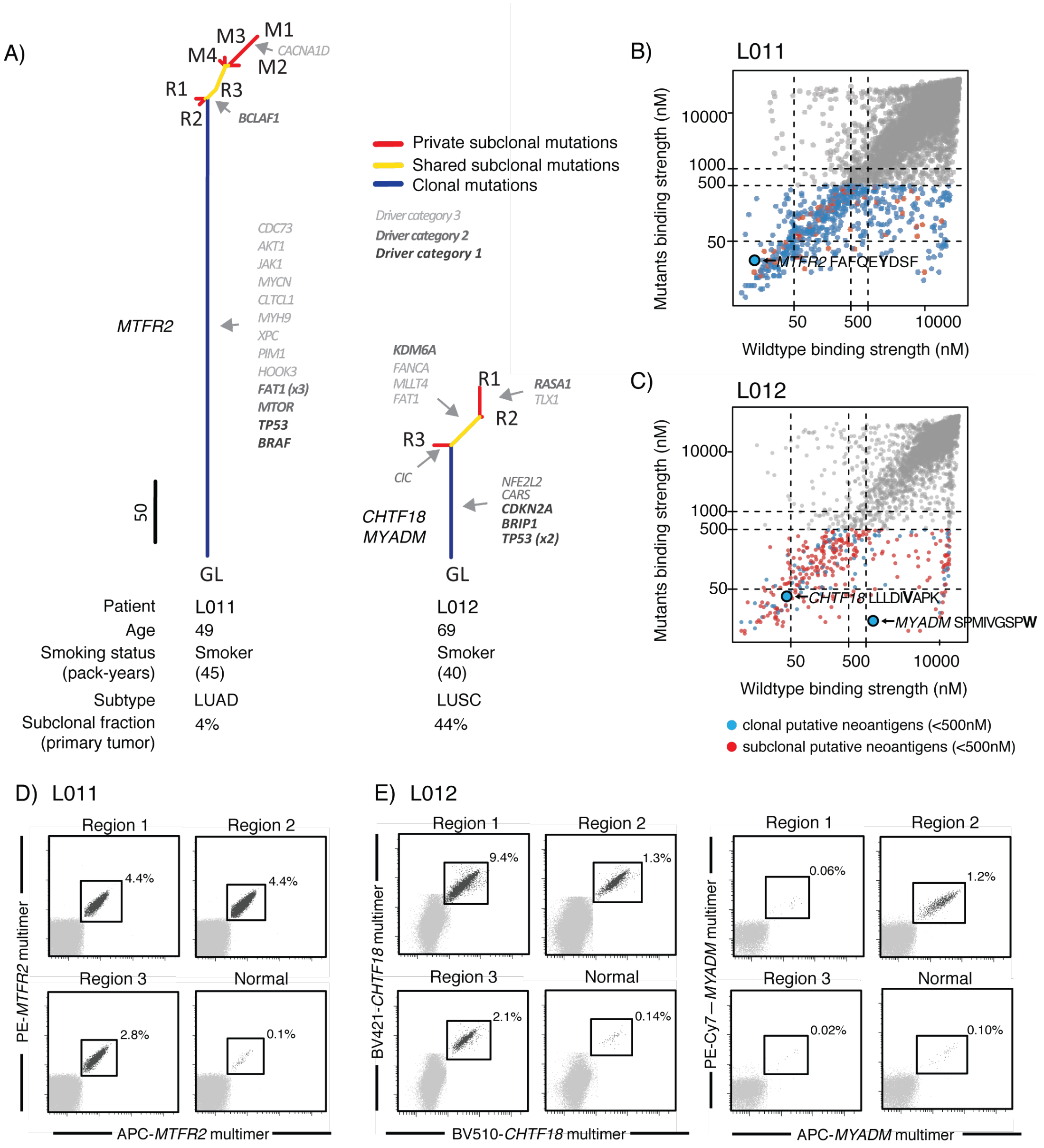


Figure 6:5 Phylogenetic trees and neo-antigens in L011 and L012

A) Phylogenetic trees for L011 and L012, with branch lengths proportional to number of non-silent mutations. B) Putative neo-antigens predicted for all missense mutations in L011. The MTFR2^{D326Y} neo-antigen (FAFQE**Y**DSF) is highlighted. C) Putative neo-antigens predicted for all missense mutations in L012. The CHTF18 L769V neo-antigen (LLLDI**V**APK) and MYADM^{R30W} neo-antigen (SPMIVGSP**W**) are indicated. D, E) MHC-multimer analysis of in-vitro expanded CD8^{+} T lymphocytes deriving from three tumour regions and normal tissues for L011 (D) and L012 (E). In both cases, frequency of CD3^{+}CD8^{+} T lymphocytes reactive to mutant peptides is indicated.

In L012, CD8$^+$ T cells reactive to an HLA-A1101-multimer derived from mutant *CHTF18*^{L769V} (LLDI***V***APK) and an HLA-B0702-multimer derived from *MYADM*^{R30W} (SPMIVGSP***W***) were identified in all tumour regions and at lower frequencies in normal tissue (Figure 6:5 E). Both were clonal mutations, *CHTF18* with high-predicted binding to HLA-A1101 (<50nM) in mutant and wild type forms, and *MYADM* with lower predicted binding in wild type to HLA-B0702 (>1000nM) compared to mutant form (<50nM) (Figure 6:5 C). An HLA-A1101-multimer derived from the wild type sequence LLLDI**L**APK (41 nM), for *CHTF18* was not found to illicit a T-cell response. However, an HLA-B0702-multimer derived from the wild type sequence SPMIVGSP**R** (1329 nM) for *MYADM* was found to provoke a T-cell response. Notably, as the wild type peptide had a very weak predicting peptide to HLA binding score (>1000 nM), it is predicted that the actual tetramer may be very unstable, and therefore rarely, if ever, seen by the immune system. Thus the perceived T cell response may reflect T-cells primed towards the mutant peptide.

6.2.6 Clonal dissection of tumour exomes from patients treated with pembrolizumab

Blockade of immune inhibitory receptors, such as PD-1, PD-L1 and CTLA-4, has been shown to be associated with significantly improved response rates in NSCLC (Topalian et al., 2012, Garon et al., 2015, Soria et al., 2015). Relatedly, a higher neo-antigen burden – predicted by a pipeline similar to that described above – has been shown to be associated with improved objective response, durable clinical benefit, and progression free survival in NSCLC patients treated with pemrolizumab, an antibody targeting PD-1 (Rizvi et al., 2015).

To explore whether the clonal status of putative neo-antigens impacts upon sensitivity to PD-1 blockade, exome sequencing data from a recent study in which two independent NSCLC cohorts were treated with pembrolizumab was obtained (Rizvi et al., 2015). In total, sequencing data from 32 NSCLC tumours was acquired, 16 of which were originally designated as a discovery cohort, and 18 of which were considered a validation cohort. The clinical characteristics of all patients can be found in Table 8-11, in Appendix 3.

The clonal architecture of each tumour was dissected by estimating the cancer cell fraction of each mutation and clustering these using a Dirichlet clustering process (see Experimental procedures). (For two tumours, ZA6965 and GR0134, reliable copy number, mutation and purity estimations could not be extracted, rendering clonal architecture analysis intractable.) Both the mutational burden and the degree of intra-tumour heterogeneity, as measured by the proportion of subclonal mutations, varied considerably for tumours across the two tumour cohorts (Figure 6:6).

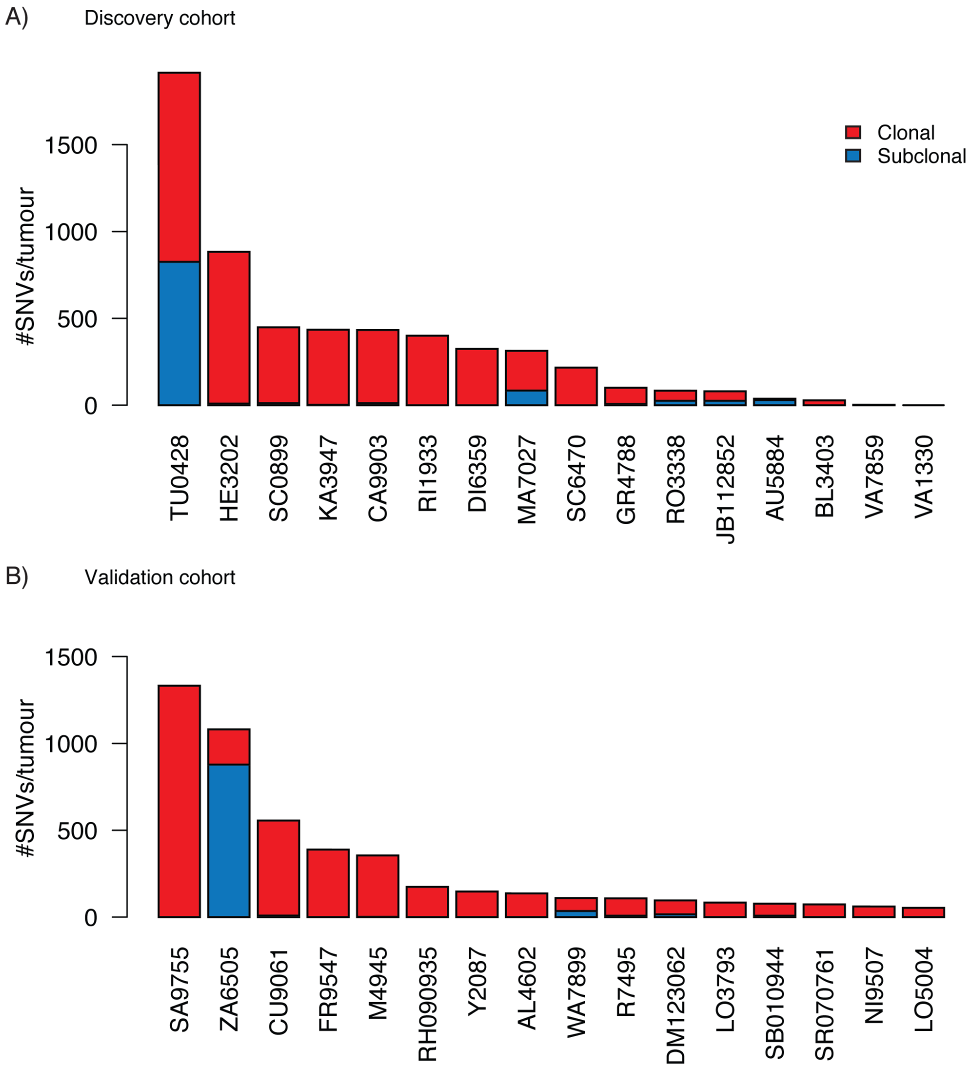


Figure 6:6 Mutation burden and subclonal fraction in discovery and validation cohort.

Total number of SNVs identified for each sequenced tumour, with number of clonal (red) and subclonal (blue) displayed in the barplot.

6.2.7 Relationship between number of clonal neo-antigens and sensitivity to PD-1 blockade in NSCLC

To determine the number of neo-antigens within the discovery and validation cohort, the neo-antigen prediction pipeline described above was applied to all sequenced tumours.

As previously reported (Rizvi et al., 2015), neo-antigen burden was related to the clinical efficacy of pembrolizumab in the discovery and validation cohort, with a high neo-antigen repertoire associated with improved outcome (Figure 6:7). This relationship was also contingent upon the clonal architecture of each tumour (Figure 6:7 A-H). In the discovery cohort, every tumour exhibiting durable clinical benefit (DCB, defined as in (Rizvi et al., 2015) as partial response or stable disease lasting > 6 months) harboured a high clonal neo-antigen burden (defined as above or equal to the median number of clonal neo-antigens in the discovery cohort, 91) coupled with a neo-antigen subclonal fraction lower than 5% (Figure 6:7). Conversely, every tumour exhibiting a non-durable benefit (NDB) exhibited either a low clonal neo-antigen repertoire (<91) or a high neo-antigen subclonal fraction (>5%). Indeed, the only tumour with a high neo-antigenic load and a non-durable response, TU0428, was an extremely heterogeneous tumour, with 301 out of 902 somatic mutations classified as subclonal. Thus, in the discovery cohort, combining both neo-antigen repertoire and neo-antigen heterogeneity was able to predict sensitivity to pembrolizumab, better than either measure alone (Figure 6:7 A-C).

Similarly, in the validation cohort, five of six tumours with a high clonal neo-antigen burden (defined as greater than or equal to the median of the validation cohort, 69) and low subclonal neo-antigen fraction (<5%) were associated with DCB (Figure 6:7). Conversely, eight out of ten tumours with low clonal neo-antigen burdens or high neo-antigen heterogeneity were associated with NDB. For example, SA9755, a homogenous tumour with a high clonal neo-antigenic burden, responded well to treatment, exhibiting an on-going durable clinical response at the time of data lock (Figure 6:8 A). By contrast, despite a large neo-antigen burden, ZA6505 exhibited a non-durable clinical response, relapsing after 2 months (Figure 6:8 A). ZA6505 was the one of most heterogeneous tumour within the cohort, with over 80% of mutations classified as present in only a subset of tumour cells within this metastatic liver sample. Moreover, consistent with many of the mutations occurring after the emergence of the founding cancer clone and metastatic dissemination to the liver, two mutational signatures that have been identified in liver tumours, Signature 16 and Signature 12, were identified within ZA6505 (Figure 6:8 B).

In summary, when the extent of neo-antigen heterogeneity and the clonal neo-antigen burden were considered together, clinical benefit could be predicted in almost all cases (Figure 6:7). Notably, two of the three outliers were LUSC tumours (CU9061 and SB010944) and consistent with survival analysis from TCGA data, when the four LUSC tumours in both cohorts were removed, an even clearer relationship was observed. Moreover, a greater PD-L1 expression was observed in tumours harbouring a large clonal neo-antigen burden and low neo-antigen heterogeneity compared to those with a low neo-antigen load or high neo-antigen heterogeneity (*P*=0.0017, chi-square test; Figure 6:9).

These results also remained consistent when considering all mutations rather than class-I restricted putative neo-antigens, supporting the notion that unidentified MHC class II-restricted neo-antigens may also play a significant role in immune reactivity (Linnemann et al., 2015) and the need for refinement of neo-antigen prediction algorithms (Schumacher and Schreiber, 2015).

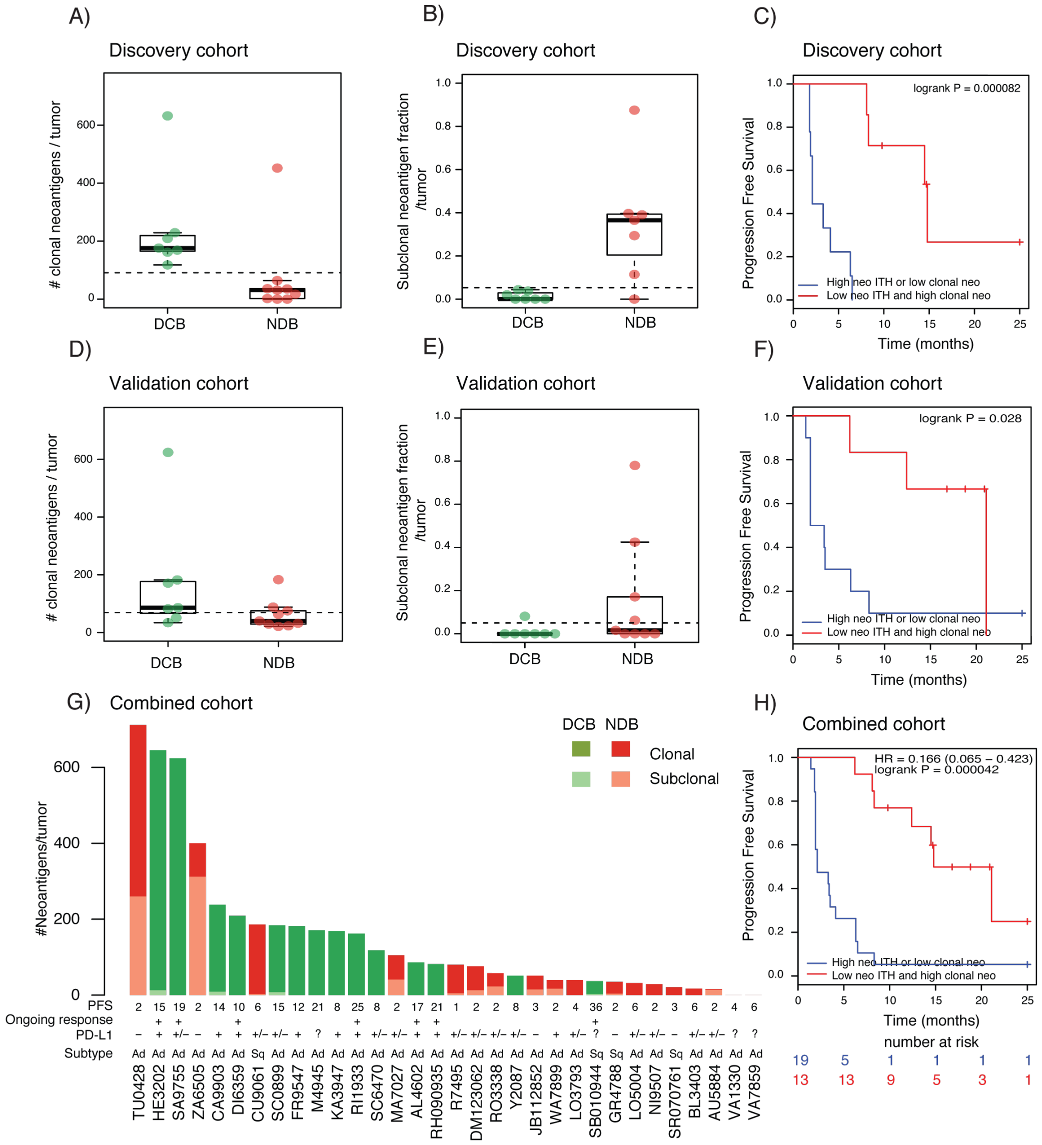
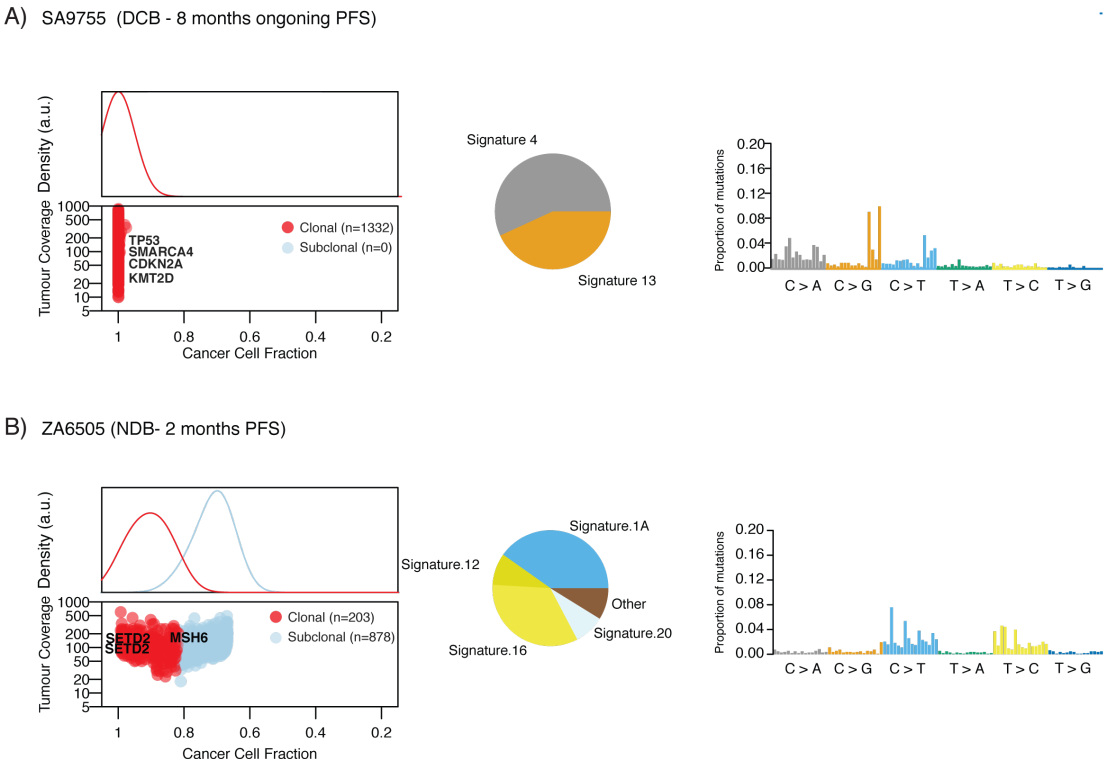


Figure 6:7 Relationships of clonal neo-antigens and intra-tumour heterogeneity with clinical benefit of anti-PD-1 therapy

For discovery (A-C) and validation cohort (D-F), number of clonal neo-antigens and fraction of subclonal neo-antigens is shown for patients with a durable clinical benefit (DCB), or non-durable benefit (NDB). Progression free survival in tumours with a higher number of neo-antigens and low subclonal fraction compared to those with a lower number of neo-antigens or high subclonal fraction is shown for discovery (C) and validation (F) cohorts. G) Clonal architecture for each sequenced tumour. PFS are reported under bar-plot and those with on-going progression-free survival are labelled with +. PD-L1 is indicated below bar-plot: Strong (+) 50% membranous staining; Weak (+/-), 1-49% membranous staining; Negative (-),<1% membranous staining; Unknown (?). (H) Progression free survival in combined tumour cohort comparing tumours with a higher number of neo-antigens and low subclonal fraction with those with a lower number of neo-antigens or high subclonal fraction.



A) SA9755 (DCB - 8 months ongoing PFS)

B) ZA6505 (NDB- 2 months PFS)

Figure 6:8 Clonal architecture and mutational signatures in two tumours

For two tumours the clonal architecture and mutational signatures identified using the deconstructSigs mutational signatures package are depicted. Notably, ZA6505 harbours mutational signatures that are generally found in liver tumours, suggesting a large proportion of mutations occurred after metastatic dissemination to the liver. Names of mutational signatures correspond to those used by Alexandrov et al. (2013a).

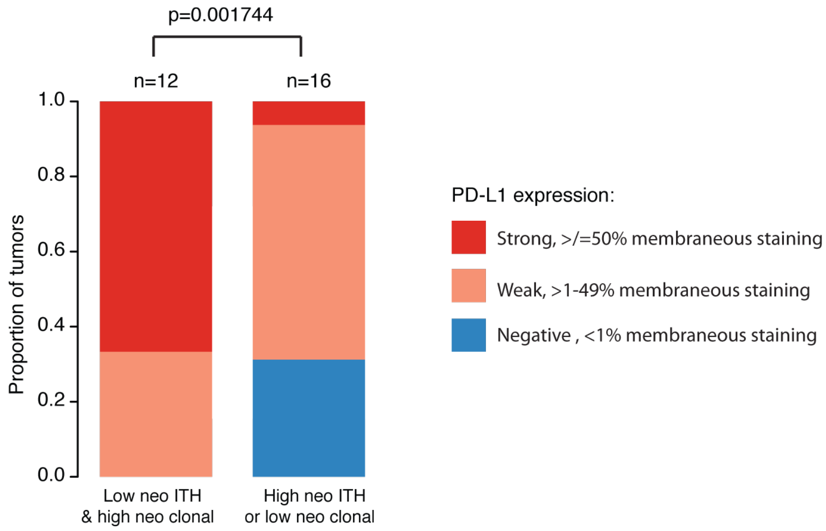


Figure 6:9 PD-L1 expression for two groups of tumours
PD-L1 exhibits significantly stronger expression in tumours harbouring a high clonal neo-antigen burden and a low subclonal neo-antigen fraction compared to tumours harbouring a low clonal neo-antigen burden or high subclonal neo-antigen fraction.

6.2.8 A clonal neo-antigen specific CD8+ T cell response reflects tumour regression

Finally, given that candidate neo-antigens had previously been examined in a tumour (CA9903) with exceptional response to pembrolizumab within this cohort (Rizvi et al., 2015), the clonal status of mutations and candidate neo-antigens were further considered, specifically for this tumour.

Previous analysis demonstrated a CD8+ T cell response against a neo-antigen resulting from a *HERC1* P3278S mutation (ASNA***S***SAAK). Subclonal architecture analysis revealed that this mutation was likely present in all tumour cells within the sequenced tumour and exhibited mutation copy number of 1.73, suggesting it was likely clonal within the tumour population and present at multiple copies (Figure 6:10)

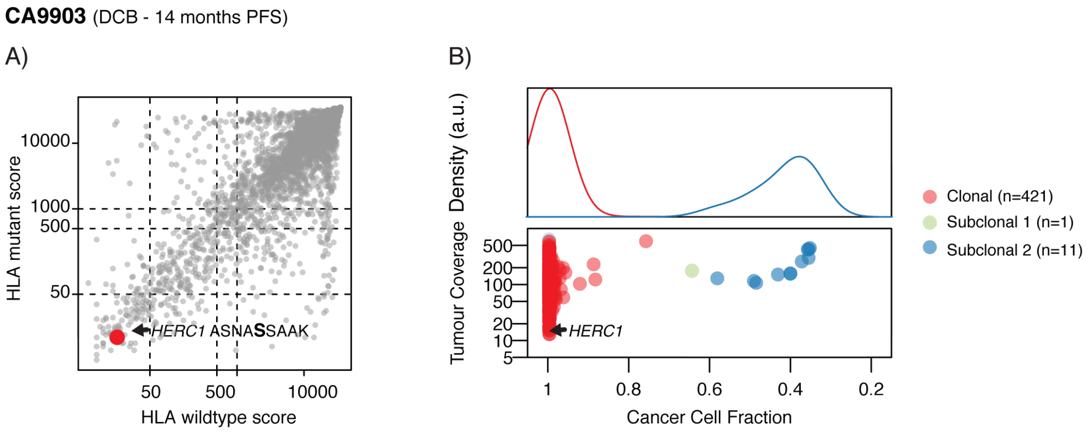


Figure 6:10 Clonal status of neo-antigen in CA9903

A) Putative neo-antigens predicted for all non-synonymous mutations in CA9903. The HERC1 P3278S neo-antigen (ASNA**S**SAAK) is highlighted. Each dot represents one peptide. Binding scores IC50 (nM) is indicated, a lower binding score indicates stronger binding affinity. B) Clonal architecture of CA9903 tumour sample with HERC1 mutation highlighted. A small subclonal population, representing 3% of the mutational burden, is detected.

6.3 Conclusions

Exploration of neo-antigens in LUAD and LUSC tumours revealed that the majority of these tumours harbour putative neo-antigens, however there was considerable range across cancers. Consistent with the clinical importance of neo-antigens (Schumacher and Schreiber, 2015, Brown et al., 2014, Rooney et al., 2015), LUAD tumours with a high neo-antigen burden were associated with significantly longer overall survival times. Clonal neo-antigens eliciting a T-cell response could be identified in multi-region sequenced NSCLC tumours.

Importantly, the relationship between neo-antigens and survival was also found to be more pronounced when considering clonal neo-antigens. Likewise, the established clinical benefit associated with immune modulation through anti-PD-1 (Rizvi et al., 2015) was found to be related to both the number of clonal neo-antigens and the neo-antigen subclonal fraction

These data suggest that clonal and subclonal neo-antigens might not drive equally effective anti-tumour immunity. Low intra-tumour heterogeneity and high clonal neo-antigen burden might result in higher antigen dosage compared to tumours with low clonal neo-antigen burden and high heterogeneity. Moreover, extensive heterogeneity might foster generation of T cells favouring the recognition of mutations not expressed by all tumour cells, limiting tumour control.

Targeting clonal mutations, shared by all tumour cells might hold promise to address the challenges of intra-tumour heterogeneity outlined in the preceding chapters.

Chapter 7. Discussion

7.1 Introduction

Precision medicine will ultimately require not only a catalogue of driver mutations and mutational processes, but also an understanding of their spatial and temporal dynamics during a tumour's evolution. More specifically, the drivers of intra-tumour heterogeneity, which provides a substrate for tumour evolution, will need to be understood in order to identify the key causes of tumour adaptation and resistance to cancer therapies (Greaves, 2015).

In the course of this thesis, detailed genomic analysis of tumours from ten cancer types has yielded insights into the clonal status of somatic events in cancer, as well as the temporal dynamics of mutational processes. Analysis of allele specific tumour copy number data and a colorectal cancer cell-line system has shed light on the importance of genome doubling in cancer genome evolution. Finally, the clinical relevance of tumour heterogeneity in NSCLC has been explored in the context of cancer neo-antigens and immune modulation.

In this chapter, the biological insights engendered by this analysis are explored and considered within the context of the current literature. Potential future directions and the clinical implications of these findings are also discussed.

7.2 Heterogeneity is widespread across human cancers

Consistent with emerging data from multiple studies undertaken since the start of this PhD (Gerlinger et al., 2012, Gerlinger et al., 2014a, Shah et al., 2012, Yates et al., 2015, Greaves, 2015, Landau et al., 2013, Lohr et al., 2014, Alexandrov et al., 2013a, Suzuki et al., 2015), the results of this thesis highlight that intra-tumour heterogeneity is widespread across cancers.

For the first time, multi-region whole-exome and/or whole genome sequencing was applied to both NSCLC and oesophageal adenocarcinoma primary tumours

(Chapter 3). In every tumour sequenced, branched evolution was observed, with distinct subclonal populations spatially separated within the primary tumour.

The heterogeneity observed in LUAD, with a median of 30% of non-silent mutations classified as heterogeneous (Figure 3:3), is consistent with a recent study, which also performed multi-region sequencing in this cancer type, and identified, on average, 24% of mutations as heterogeneous (Zhang et al., 2014). Zhang and colleagues (2014) also found that the number of trunk mutations increased upon deep sequencing of mutations. Arguably, however, this superficially intriguing finding is the result of their definition of a trunk mutation, which is both confusing and flawed. From a phylogenetic perspective, placing a mutation on the trunk of a phylogenetic tree implies it occurred prior to the emergence of the most recent common ancestor. Thus, a trunk mutation should be clonal, or at least exhibit evidence for having been clonal at one point in time. According to Zhang and colleagues (2014) a mutation identified as subclonal in all tumour regions can be truncal, which is counterintuitive and does not fit within a logical evolutionary framework. As such, the results of their deep sequencing analysis does not suggest there is an illusion of heterogeneity that is resolved upon deeper sequencing; rather, deep sequencing can further elucidate the subclonal structure of a tumour, identifying branched mutations present in only a small subset of tumour cells that may be missed at lower sequencing depths.

Moreover, it is worth considering that although multi-region sampling permits spatial heterogeneity to be evaluated, it still does not equate to comprehensive sampling of the tumour. Indeed, in the analysis of multi-region NSCLC tumours, on average <5% of the tumour was sequenced, thus it is likely that the extent of heterogeneity in these tumours exceeds that estimated here. Further research will be required to explore the optimum number of biopsies and optimum depth required to adequately capture the extent of heterogeneity in NSCLC tumours. In the future, it will also be important to explore the extent and importance of heterogeneity at the transcriptomics and epi-genomics levels.

The extent of heterogeneity observed can also be put into context by a comparison with tumours from other cancer types where multi-region sequencing has been performed. While the proportion of heterogeneous mutation was greater in stage IV KIRC compared to the early stage NSCLC tumours explored here (76% vs. 30%) (Gerlinger et al., 2012, Gerlinger et al., 2014a), the actual number of subclonal mutations was, on average, slightly greater in NSCLC with 98 (range, 19-300) non-silent mutations identified as subclonal per tumour in NSCLC, compared to 57 (range, 21-122) in KIRC. These data likely reflect differences in mutational processes operating in these tumour types over time (see below). In oesophageal adenocarcinoma, a median of 56% of non-silent mutations were classified as heterogeneous (range 12.6% to 69.4%), and the average non-silent mutation burden was 206.75 (range 127-320), greater than KIRC (85.6, range 54-140) yet smaller than NSCLC (280.5, range 58-476). Unfortunately, it is difficult to compare to ovarian cancer given the considerable inconsistencies in supplementary tables associated with this study (Bashashati et al., 2013).

7.2.1 Illusion of clonality and clonal architecture of tumours

The results presented here also highlight the potential limitations of using a single sample to identify the driver events of a tumour and to characterize its subclonal architecture.

In both lung cancer and oesophageal adenocarcinoma a large proportion of mutations analysed in a single tumour region context were found to exhibit an illusion of clonality (Chapter 3). That is, mutations appeared fully clonal in one tumour region while they were entirely absent or subclonal in other tumour regions. For instance, in L002, over 10,000 mutations from region R1 were identified as clonal, but not identified as present in region R3. Similarly, in oesophageal adenocarcinoma, driver mutations exhibiting an illusion of clonality were identified in all but one of eight tumours. These findings extend the observations of Gerlinger (Gerlinger et al., 2014a), highlighting that multi-region sequencing may be required to gain a more complete picture of the genomic landscape of tumours. More recently, similar findings have also been documented in breast cancer (Yates et al., 2015), where seven of twelve tumours subject to multi-region sequencing were found to harbour driver mutations that could erroneously be classified as clonal, based on a single sample.

From a clinical perspective, these results emphasize how intra-tumour heterogeneity can compromise the utility of a single biopsy to inform treatment strategies, and suggest targeting truncal drivers may be necessary (Yap et al., 2012b). For instance, L008 presented with an activating *BRAF* (G469A) mutation (Davies et al., 2002) in all tumour regions and an activating *PIK3CA* (E542K) mutation (Kang et al., 2005) only in region R3 (Figure 3:3). Thus, a biopsy taken from region R3 might suggest treatment with an inhibitor of the PI3K/mTOR signalling axis and combination therapy. Conversely, a single biopsy from any other region would suggest treatment with a *BRAF* inhibitor, for which the tumour cells from region R3 might be resistant as a result of the activating *PIK3CA* mutation (Shi et al., 2014).

The clinical implications of heterogeneity are also underlined by the analysis of oesophageal adenocarcinoma, where a strong correlation between the heterogeneity index and response to neo-adjuvant platinum chemotherapy was observed (Figure 3:19), suggesting a simple heterogeneity score may predict response to chemotherapy. However, given the limited number of cases in the cohort (n=8), there is a need to confirm these results in an appropriately powered prospective cohort study. Prospective studies such as TRACERX (Jamal-Hanjani et al., 2014) will shed light on the clinical relevance of intra-tumour heterogeneity in NSCLC.

7.2.2 Parallel evolution

In keeping with accumulating evidence across cancers (Johnson et al., 2014, Gerlinger et al., 2012, Gerlinger et al., 2014a, Yates et al., 2015, Suzuki et al., 2015), in oesophageal adenocarcinoma and using single samples across the pan-cancer cohort it was also possible to identify evidence for parallel evolution of subclones, with distinct somatic events affecting the same gene or pathway appearing to occur in distinct branches of the tumour phylogenetic tree. These data likely reflect constraints to tumours' evolution that conceivably might be therapeutically exploitable in the future. As additional data accumulates it will be interesting to explore whether further patterns across cancer types can be deciphered.

7.2.3 Driver genes can be subclonal

Most drug development programs employing next-generation sequencing as a stratification tool do not consider the clonal or subclonal frequencies of a driver alteration, simply their presence or absence (Swanton, 2012). Using both multi-region sequencing and single sample sequencing data, it was possible to assess the extent to which driver mutations are subclonal, present in only a subset of cancer cells, or clonal, present in all tumour cells (Chapter 3).

Across the pan-cancer cohort, and multi-region sequencing data, the presence of considerable intra-tumour heterogeneity in driver events, including known canonical hotspot mutations such as *IDH1* (R132H) and *PIK3CA* (E545K), was identified. Indeed, almost every gene linked with a target therapy was found to harbour a subclonal mutation in at least one tumour. Intriguingly, in the pan-cancer analysis, genes involved in the PI3K-AKT-mTOR pathway were found to harbour a higher proportion of subclonal mutations compared to genes associated with cyclins or cyclin-dependent kinases or the RAS-MEK pathway (Figure 3:15). Conceivably, this may shed light on the limited success of drugs targeting the PI3K signalling axis (Khan et al., 2013).

Importantly, targeted therapy applied to cancer cells lacking the targeted mutation could also have a paradoxical stimulatory effect, resulting in increased growth of wild-type subclones (Lohr et al., 2014). The results therefore suggest the need to stratify targeted therapy response according to the proportion of tumour cells in which the driver is identified and, moreover, indicate that certain pathways may be more actionable than others. Clinical trials such as the DARWIN program (Deciphering Anti-tumour Response With INtra-tumour heterogeneity) may shed light on whether targeting a clonally dominant driver event results in improved progression free survival outcomes relative to targeting the same driver event when it is present subclonally.

7.2.4 Mutations in known cancer genes tend to occur early

Nevertheless, despite evidence for subclonal mutations in cancer genes, it is important to emphasize that the majority of known driver mutations were found to be clonal, present in all tumour cells, using both multi-region and single sample analysis. In oesophageal adenocarcinoma, when focussing on mutations in known cancer genes there was a trend for mutations to be clonal, and, likewise, in NSCLC, there was a significant enrichment for mutations in cancer genes to be clonal. In the pan-cancer cohort, non-silent mutations in known cancer across every cancer type harboured a significantly higher proportion of clonal mutations compared to non-silent mutations in other genes, with the exception of KIRC (Figure 3:12), where *VHL* was the only cancer gene with a clear enrichment of clonal mutations.

These observations extend previous findings across cancer types (Shah et al., 2012, Weaver et al., 2014, Yachida et al., 2010). For instance, Shah (Shah et al., 2012) and Nik-Zainal (Nik-Zainal et al., 2012b) found that non-silent mutations in *TP53* in breast cancer exhibited a tendency to be clonal, while evidence from Zhang and colleagues (Zhang et al., 2014) suggests known driver mutations are predominantly clonal in NSCLC. In KIRC, the only cancer driver found to harbour a significant enrichment of clonal mutations was *VHL*, consistent with previous findings (Gerlinger et al., 2012, Gerlinger et al., 2014a).

7.2.5 New cancer genes can be identified based on subclonal analysis

The tendency for known canonical driver mutations to occur early in tumour evolution may also reflect the fact that current methods of defining driver mutations are biased to detect clonal drivers, present in all cancer cells. Consistent with this, for the first time, by focusing on later mutations, it was possible to detect novel putative cancer genes (Table 8-3). These genes may play roles in tumour maintenance and progression. Indeed, some of the cancer genes that were identified, such as *ATXN1* and *CTNNA2*, have been linked to tumour metastasis and cell migration (Lee et al., 2011b). Although follow-up studies will be required to validate the functional impact of these genes, these results suggest that the catalogue of cancer genes is far from complete, consistent with saturation studies carried out by Lawrence and colleagues (2014). Future studies, involving multi-region sequencing in hundreds of tumours, may reveal the extent to which selection or neutral evolution is the dominant process driving tumour evolution after subclonal diversification.

7.3 Mutational processes vary over time

Temporal dissection of mutations also permitted insights into the dynamics of mutational processes during the disease course (Chapter 5).

7.3.1 An aging signature dominates early tumour evolution in most cancer types

A mutational signature dominated by C>T transitions at CpG sites was found to dominate early mutations in four cancer types, COAD, GBM, HNSC and BRCA (Figure 5:7). This signature has been linked to aging and likely reflects spontaneous deamination of methylated cytosines (Yates and Campbell, 2012). The results therefore suggest a large proportion of mutations in cancer genomes precede tumour formation, consistent with mutational signature analysis of 20 breast cancers (Nik-Zainal et al., 2012a) and mathematical modelling studies (Tomasetti et al., 2013).

In oesophageal primary tumours, however, a similar ‘aging’ signature was found to increase in prevalence on the branches of phylogenetic trees (Figure 5:13). In this case, this likely reflects the fact that founder genomic events are caused by a distinct, oesophageal adenocarcinoma signature (T>G at CpTpT sites), potentially attributable to gastric acid exposure (Dulak et al., 2013, Weaver et al., 2014, Nones et al., 2014), which reduces in prevalence after tumour formation. The increase in

C>T mutations at CpG sites could also reflect more rapid cell turnover induced by the onset of chromosomal instability.

7.3.2 Exogenous mutational processes decrease over time

Concordant with previous studies (Govindan et al., 2012, Imielinski et al., 2012, Lee et al., 2010, Pleasance et al., 2010b) a smoking signature, characterized by C>A transversion mutations, was identified in LUAD and LUSC tumours, both from single sample and multi-region data (Figure 5:1,Figure 5:7). Temporal dissection of mutations revealed that the smoking induced mutations are likely the principal cause of many early clonal mutations, yet their relative contribution to the mutational load decreases during the evolution of tumours, both in current smokers and ex-smokers. These data suggest that even in the face of the continued mutagenic insults of tobacco carcinogens, an additional mutational process may often be required for tumour progression and maintenance, supporting the notion of a mutator phenotype in cancer (Loeb, 2001).

7.3.3 APOBEC mediated mutagenesis occurs later in tumour evolution

In multi-region NSCLC tumours, the subclonal mutational spectrum was characterized by C>T and C>G mutations, particularly at TpC sites (Figure 5:3). This mutational signature has been linked to cytosine deamination by APOBEC enzymes (Nik-Zainal et al., 2012a, Roberts et al., 2013), and suggests APOBEC mediated mutagenesis may become a pervasive mutational force later in tumour evolution. In keeping with these results, in the pan-cancer analysis, mutations within an APOBEC motif were found to increase in prevalence during the disease course (Figure 5:7), in LUAD as well as HNSC and ER-negative BRCA. APOBEC-mediated mutagenesis could also be linked to the acquisition of subclonal driver events (Figure 5:19), highlighting how mutational processes can alter the evolutionary trajectory of tumours consistent with findings in HNSC (Henderson et al., 2014).

Clearly an understanding of the underlying mechanisms that trigger APOBEC mediated mutagenesis is of paramount importance and an important avenue for future work. A recent paper highlighted the need to consider both APOBEC3A and APOBEC3B as potential culprits (Chan et al., 2015).

7.3.4 A platinum signature may contribute to intra-tumour heterogeneity

Multi-region sequencing of oesophageal adenocarcinoma tumour samples obtained before and after adjuvant platinum chemotherapy provided insights into how therapy can shape the genomic landscape of tumours (Chapter 5).

Combined analysis of post-chemotherapy mutations from five tumours revealed an enrichment of C>A mutations at CpC sites (Figure 5:17), a signature that has been linked to exposure to platinum chemotherapy in model organisms (Meier et al., 2014). In two tumours that exhibited a poor response to chemotherapy, tumour regions collected after therapy clustered together on the phylogenetic tree as a monophyletic group, potentially indicating a common somatic event conferring resistance. Conversely, the three other tumours, associated with good or intermediate response, exhibited paraphyly, whereby post-chemotherapy regions did not form a distinct branch on the evolutionary tree. In all tumours, mutations identified after treatment were often subclonal (Figure 5:18). While it is difficult to discern how many of these mutations can be directly linked to chemotherapy, these data highlight that a greater understanding of the impact of chemotherapy on the evolutionary trajectory of tumours is required and emphasize the need to identify patients that are unlikely to benefit from chemotherapy (Devarakonda and Govindan, 2015).

7.3.5 Genome doubling may represent an evolutionary leap

In four of six multi-region NSCLC tumours, genome-doubling events were observed (Chapter 4). In one tumour there was evidence that the doubling event occurred after a *TP53* mutation, but within a smoking signature context, 20 years prior to clinical detection (de Bruin et al., 2014). This indicates a prolonged tumour latency period after genome doubling and before clinical detection in NSCLC and supports recent approaches for screening in this cancer type (Peled and Ilouze, 2015).

In oesophageal adenocarcinoma tumours, clonal disruption to *TP53* and early genome-doubling events were identified in every tumour, suggesting mutations to this gene and tetraploidisation events may serve as useful clinical biomarkers. These findings corroborate observations that mutations to *TP53* occur early in oesophageal adenocarcinomas (Weaver et al., 2014), and support the results from a previous longitudinal case-control study that found chromosomal instability, genome doubling and genome diversity in samples obtained from patients with Barrett's oesophagus could predict progression to oesophageal adenocarcinoma (Li et al., 2014).

Conceivably, large-scale genomic changes such as genome doubling may often be required for cells to make the leap from a benign to a malignant phase and therefore may serve as useful markers for clinical risk prediction in other cancer types. Indeed, in support of the notion that genome doubling may facilitate accelerated cancer genome evolution, using a colorectal cancer cell-line system it was demonstrated that rare cells that survive genome-doubling display an increased tolerance to chromosome aberrations compared to their diploid counterparts (Figure 4:10). Moreover, genome doubling was found to be associated elevated genomic complexity and with reduced relapse-free survival times in two independent cohorts in this cancer type (Figure 4:14). In keeping with these results, a recent study found multi-drug resistance is promoted upon tetraploidisation in human cells (Kuznetsova et al., 2015), and modelling of the dynamics of tumour heterogeneity suggested a genome-doubling event provides the most parsimonious explanation for the frequency of triploid tumours across cancer types (Laughney et al., 2015). Moreover, a recent pan-cancer study also found a relationship between genome doubling an elevated genomic complexity (Zack et al., 2013).

Taken together, these results suggest a genome-doubling event could represent a macro-evolutionary leap in tumours, which can precipitate and sustain extensive chromosomal rearrangements. Indeed, a recent study in yeast found that compared with haploids and diploids, tetraploids undergo significantly faster adaptation (Selmecki et al., 2015). In the future, it will be important to investigate mechanisms leading to the emergence of genome doubled cells in tumours, such as, for example, cellular stress, cytokinesis failure or telomere shortening (Davoli and de Lange, 2011), and, moreover, the reason why its timing appears to vary across tumours.

7.4 Clonal cancer neo-antigens elicit T cell immunoreactivity and anti-PD-1 response

As illustrated in this thesis, NSCLCs often harbour a considerable mutation burden, a large proportion of which can be attributed to smoking related carcinogens. This mutational burden may be associated with a substantial neo-antigen repertoire, which may be clinically relevant. Indeed, a link between high tumour burden and efficacy of immune checkpoint blockade has been demonstrated in NSCLC (Rizvi et al., 2015), and the presence of neo-antigens has been associated with improved prognosis in a meta-analysis of cancer types (Brown et al., 2014).

In support of the importance of neo-antigens, patients with LUAD tumours harbouring a high burden of predicted neo-antigens were found to have significantly improved overall survival compared to patients carrying tumours with intermediate or low numbers of predicted neo-antigens (Figure 6:2). This relationship was more pronounced when considering specifically clonal neo-antigens, present in all tumour cells (Figure 6:3). These observations therefore extend previous observations (Brown et al., 2014, Rooney et al., 2015), suggesting not only the presence of neo-antigens but also that their number and clonal status may impact upon the likelihood of effective immune surveillance. Moreover, T cells reactive to clonal neo-antigens were identified from two multi-region sequenced NSCLC tumours, validating the ability of the neo-antigen pipeline to predict bona-fide neo-antigens (Figure 6:5).

Analysis of the clonal architecture of NSCLC tumours treated with pembrolizumab revealed that progression free survival benefit was dependent on both the number of neo-antigens and their clonal status, with tumours harbouring a high clonal neo-antigen burden coupled with low neo-antigen intra-tumour heterogeneity exhibiting durable complete responses (Figure 6:7). This relationship may reflect the fact that high intra-tumour heterogeneity results in a lower antigen dosage and higher T-cell clonal competition compared to tumours harbouring a high clonal neo-antigen burden (Kedl et al., 2000).

Larger perspective studies using multi-region sequencing to evaluate intra-tumour heterogeneity may be required to further validate these observations, and, in the future, it will be important to explore the impact of intra-tumour heterogeneity and immune modulation in early-stage primary tumours (Soria et al., 2015). Data from DARWIN II will shed further light on the relationship between intra-tumour neo-antigen heterogeneity and neo-antigen burden and progression free survival in patients treated with a monoclonal antibody targeting anti-PD-L1. It also remains to be elucidated why this relationship is observed particularly in LUAD tumours and not LUSC tumours and whether the same relationship holds for tumour types such as melanoma. Finally, it will also be important to consider whether filtering neo-antigens based on those that can also be identified through mass-spectrometry data can refine and improve the analysis considerably.

Nevertheless, taken together, these results suggest the extensive clonal mutational repertoire present in smoking-associated LUAD might render this disease vulnerable to vaccination or T cell therapies, targeting multiple clonal neo-epitopes, in combination with checkpoint blockade.

7.5 Conclusions

The data presented in this thesis suggest diversity within tumours is widespread across tumour types and represents a significant clinical challenge. While temporal and clonal dissection of mutations revealed subclonal mutations in established cancer genes, in general these genes show a significant enrichment of clonal mutations, present in all tumour cells.

Dissection of mutational signatures suggests exogenous processes, including tobacco carcinogens, dominate the early life history of tumours, while endogenous processes gone awry, such as APOBEC mediated mutagenesis, become pervasive mutational forces later in tumour evolution, contributing to the acquisition of subclonal driver events. Platinum chemotherapy can also leave scars in the tumour genome, potentially leading to elevated tumour diversity. Analyses of tumour copy number data shed light on the importance of genome doubling in cancer genome evolution, conceivably acting as an evolutionary leap, facilitating propagation of chromosomal instability.

Finally, the clinical relevance of intra-tumour heterogeneity was demonstrated in the context of immune modulation and neo-antigens. The number of clonal neo-antigens was found to correlate with lung adenocarcinoma survival outcome and T cells reactive to clonal neo-antigens were identified. Sensitivity to anti-PD-1 therapy was found to be dependent on neo-antigen clonal burden and the extent of intra-tumour heterogeneity. Thus, immunotherapeutic strategies targeting clonal neo-antigens, present in all tumour cells, in combination with checkpoint-blockade, may provide a tractable approach to tackling lung adenocarcinomas.

Chapter 8. Appendix

8.1 Appendix 1 – List of peer-reviewed papers and reviews published during the production of this thesis

PRIMARY RESEARCH ARTICLES:

1. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution
   N McGranahan, F Fever, EC de Bruin, NJ Baraka, Z Szallasi, C Swanton
   Science Translational Medicine, 2015.

2. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution
   EC de Bruin, **N McGranahan***, R Mitter, M Salm, DC Wedge, L Yates, ...
   Science 346 (6206), 251-256, 2014
   ** Joint first author*

3. Tolerance of Whole-Genome Doubling Propagates Chromosomal Instability and Accelerates Cancer Genome Evolution
   SM Dewhurst, **N McGranahan***, RA Burrell, AJ Rowan, E Grönroos, ...
   Cancer discovery, 2014
   ** Joint first author*

4. Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy
   N Murugaesu, GA Wilson, NJ Birkbak, T Watkins, **N McGranahan***, ...
   Cancer discovery, 2015
   **Joint first author*

5. Glioblastoma adaptation traced through decline of an IDH1 clonal driver and macroevolution of a double minute chromosome
   F Favero, **N McGranahan***, M Salm, NJ Birkbak...
   Annals of Oncology, 2015
   ** Joint first author*

6. A breast cancer meta-analysis of two expression measures of chromosomal instability reveals a relationship with younger age at diagnosis and high risk histopathological variables.
   D Endesfelder, **N McGranahan***, NJ Birkbak, Z Szallasi, M Kschischo, ...
   Oncotarget 2 (7), 529-537, 2011
   ** Joint first author*

7. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing
M Gerlinger, S Horswell, J Larkin, AJ Rowan, MP Salm, I Varela, R Fisher, **N McGranahan**
Nature genetics, 2014

8. Computational optimisation of targeted DNA sequencing for cancer detection
P Martinez, **N McGranahan**, NJ Birkbak, M Gerlinger, C Swanton
Scientific reports 3, 2013

9. Parallel evolution of tumour subclones mimics diversity between tumours
P Martinez, NJ Birkbak, M Gerlinger, **N McGranahan**, RA Burrell, ...
The Journal of pathology 230 (4), 356-364, 2013

10. Chromosomal Instability Selects Gene Copy-Number Variants Encoding Core Regulators of Proliferation in ER+ Breast Cancer
D Endesfelder, RA Burrell, N Kanu, **N McGranahan**, M Howell, PJ Parker, ...
Cancer research 74 (17), 4853-4863

11. Ultra-deep T cell receptor sequencing reveals the complexity and intratumour heterogeneity of T cell clones in renal cell carcinomas
M Gerlinger, SA Quezada, KS Peggs, AJS Furness, R Fisher, T Marafioti, H Vishyesh, H Shende, **N McGranahan,…**
The Journal of pathology *231* (4), 424-432, 2013

12. Development of synchronous VHL syndrome tumors reveals contingencies and constraints to tumor evolution
R Fisher, S Horswell, A Rowan, MP Salm, EC de Bruin, S Gulati, **N McGranahan**...
Genome biology 15 (8), 433, 2014

13. SETD2 loss-of-function promotes renal cancer branched evolution through replication stress and impaired DNA repair
N Kanu, E Grönroos, P Martinez, RA Burrell, XY Goh, J Bartkova...**N McGranahan…**
Oncogene, 2015

REVIEW ARTICLES

14. Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution
N McGranahan, C Swanton
Cancer cell 27 (1), 15-26, 2014

15. The causes and consequences of genetic heterogeneity in cancer evolution
RA Burrell, **N McGranahan***, J Bartek, C Swanton
Nature 501 (7467), 338-345, 2013
**Joint first author*

16. Cancer: Evolution Within a Lifetime
M Gerlinger, N **McGranahan***, SM Dewhurst, RA Burrell, I Tomlinson, ...
Annual review of genetics 48, 215-236, 2014
**Joint first author*

17. Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine
C Hiley, EC de Bruin, **N McGranahan***, C Swanton
Genome Biol 15, 453
**Joint first author*

18. Cancer chromosomal instability: therapeutic and diagnostic challenges
N McGranahan, RA Burrell, D Endesfelder, MR Novelli, C Swanton
EMBO reports 13 (6), 528-538, 2012

19. APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity
C Swanton, **N McGranahan**, GJ Starrett, RS Harris
Cancer discovery 5 (7), 704-712

20. Tracking genomic cancer evolution for precision medicine: the lung TRACERx study
M Jamal-Hanjani, A Hackshaw, Y Ngai, J Shaw, C Dive, S Quezada, ..., **N McGranahan**, C. Swanton
PLoS biology 12 (7), 2014

21. Inferring mutational timing and reconstructing tumour evolutionary histories
S Turajlic, **N McGranahan**, C Swanton
Biochimica et Biophysica Acta (BBA)-Reviews on Cancer 1855 (2), 2015

8.2 Appendix 2 – Selection of primary research papers