The Intra-Tumour Heterogeneity Landscape of Human Cancers

Stefan Christiaan Dentro

Christ's College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

December 2017

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee

Stefan Christiaan Dentro
December 2017

Abstract

The Intra-Tumour Heterogeneity Landscape of Human Cancers

Stefan Christiaan Dentro

Tumours accumulate many somatic mutations in their lifetime. Some of these mutations, drivers, convey a selective advantage and can induce clonal expansions. Incomplete clonal expansions give rise to intra-tumour heterogeneity. Somatic mutations can be measured through massively parallel sequencing, where mutations that are supporting incomplete expansions will appear as subclonal. These mutations can be used as a marker of the existence of the expansion and allow for a window into the clonal and subclonal architecture of the tumour at diagnosis.

During my Ph.D. I have developed computational methods to infer intra-tumour heterogeneity from massively parallel sequencing data and applied these to the 2,778 tumour whole genome sequences in the International Cancer Genome Consortium Pan-Cancer Analysis of Whole Genomes initiative to paint the pan-cancer landscape of intra-tumour heterogeneity.

I will first introduce the methods; a method to call somatic copy number alterations (Battenberg) and a method to infer subclones from single nucleotide variants (DPClust). Both are extensively validated on simulated and on real data, and I describe a rigorous quality control procedure. The methods are then applied to a single sample to showcase what can be learned about the life history of a cancer, before introducing additional computational methods for a pan-cancer study of heterogeneity. Finally, I describe the findings.

I find that nearly all cancers, for which there is sufficient power, contain at least one subclone (96.7% of 1,801 primary tumours). The subclones contain driver mutations that are under positive selection, and known cancer genes contain subclonal driver mutations in low proportions. 9.5% of tumours contain only subclonal drivers that are clinically actionable, suggesting that heterogeneity could inform treatment choices. Finally, the analysis reveals that activity of smoking and UV-light associated mutational signatures goes down as the tumour evolves, while activity of the APOBEC associated signatures goes up.

Acknowledgements

I would like to thank my supervisors, Peter Van Loo, David Wedge and David Adams. I’m very grateful to have had the opportunity to work with you and on this very exciting project! Thank you very much for your guidance, thoughts, ideas and your patience. It has been an enormous pleasure to learn from you and to work with you! Thank you for the great discussions and for continuously pushing me to be a better scientist in a friendly and positive environment.

During the PCAWG project I’ve been in the fortunate position to closely collaborate with members of the Van Loo lab. Thank you Maxime Tarabichi, Kerstin Haase, Clem Jolly, Jonas Demeulemeester and Matt Fittall for the fantastic collaboration and the many in-depth discussions on tumour evolution and heterogeneity. I’d like to also thank the Van Loo lab for being my ’home’ for the last few years. Annelien, Clem, Jonas, Kerstin, Lilly, Matt, Maxime and Peter: it’s been an enormous pleasure!

Most of the work in this thesis was done as part of the international PCAWG collaboration, I’d like to thank everyone of the PCAWG collaborators and in particular Quaid Morris, Moritz Gerstung, Jeff Wintersinger, Ignaty Leshchiner, Amit Deshwar, Yulia Rubanova and Peter Campbell.

I would like to extend my thanks to Jason J. Pitt for the fruitful collaboration on the West-African breast cancer study and to members of the Adams lab, the Wedge lab and of the Cancer Genome Project for stimulating discussions and continuous helpful feedback.

This thesis could not have happened without the love and support from my parents and sister. Thank you for supporting me and believing in me!

Finally, I would like to thank the Wellcome Trust for generously funding this Ph.D.

Preface

During my Ph.D. I have been in the very fortunate position to heavily collaborate with colleagues close by and far away. Nearly all of the work reported in this thesis was performed as part of an international collaboration project to jointly analyse the cancer whole genome sequencing samples that are part of the International Cancer Genome Consortium (ICGC) The Pan-Cancer Analysis of Whole Genomes (PCAWG) initiative. As this data set is referenced throughout this thesis I have included below a high level description of the project and data set.

Nearly all of the work reported in this thesis is performed in collaboration with, or building on top of the work done by others. Throughout this thesis I therefore systematically refer to work that is solely to my credit with "I", and work that was done by others (with my involvement) as "we". There is also the occasional reference to "a collaborator", where work was done that I had no involvement with.

To help the story line I have spread the introduction across Chapters 1 and 2. Chapter 1 contains a brief review of the relevant literature, but descriptions of algorithms which were already in advanced development when I started are described in Chapter 2. I have worked on and with these algorithms throughout my Ph.D. and have made numerous improvements and adjustments (of which improvements on computational resource requirements are not reported as they are not of direct scientific interest). I felt it would make this thesis more easily readable when the algorithms and updates are described in one chapter.

This setup was chosen to paint a comprehensive story that can hopefully be understood from this thesis alone.

The ICGC Pan-Cancer Analysis of Whole Genomes initiative

The International Cancer Genome Consortium (ICGC) was created to coordinate cancer genome sequencing projects spanning 50 different types of cancer, with the aim to sequence over 25,000 cancer genomes (ICGC Consortium, 2010). ICGC is organised as a series of projects based in countries spread all over the world that focus on analysis of a single cancer type. Over 17,000 cancers have now been sequenced, of which the majority through whole exome sequencing.

The Pan-Cancer Analysis of Whole Genomes (PCAWG) project was launched to comprehensively characterise those samples for which whole genome sequencing (WGS) is available as a single data set (Campbell et al., 2017). The advantage of focussing on samples for which WGS is available is that the full genome can be interrogated, including the full array of single nucleotide variants (SNV), indels (short insertions and deletions) and structural variants (SVs).

The project consists of 16 working groups with each their own distinct theme. I have been a member of the working group that focusses on tumour evolution and heterogeneity, which is a collaboration of about 60 scientists representing 12 different laboratories.

The tumours that are part of ICGC PCAWG had to meet a series of criteria to be included: a minimal set of clinical annotations should be available, both tumour and normal samples have to be paired-end sequenced from an Illumina machine to a coverage of at least 30x and 25x respectively (Campbell et al., 2017). The data set consists of primarily treatment-naive primary tumours and nearly all matched normals are generated from blood samples.

Data from 2,834 donors was selected to be included in ICGC PCAWG, of which data from 2,658 donors passed quality control procedures (Whalley et al., 2017). The analysed data set consists of 2,778 tumours, of which 2,605 are primary tumours and 173 from a metastasis or relapse case, and spreads 39 histologically distinct types of cancer. Each sequencing sample was processed using a standardised set of primary analysis pipelines, that includes alignment of the sequencing reads and variant calling and filtering from pipelines provided by the Sanger, Broad and EMBL/DKFZ (Yung et al., 2017). These pipelines were extended by one additional SNV and one additional indel caller to further increase the reliable detection of low allele frequency variants (Campbell et al., 2017). Clinical data was systematically collected and standardised (Campbell et al., 2017).

Output from the three primary variant calling pipelines was combined into a high quality set of somatic consensus SNVs (Campbell et al., 2017), indels (Campbell et al., 2017), SVs (Campbell et al., 2017) and copy number alterations (CNAs) (Dentro et al., 2017, manuscript in preparation), of which the latter is described in this thesis. The optimal strategy to find consensus SNVs and indels was found by first running 19 different variant callers across a selected set of 64 tumours. 250.000 calls were selected for validation through deep targeted capture sequencing such that every combination of variant callers is represented and were stratified by allele frequency, after which consensus strategy that maximises precision and recall was then developed to generate the final PCAWG calls (Campbell et al., 2017).

These steps have created the largest set of whole genome cancer sequences to date, spreading a broad range of cancer types. The data set is uniformly processed and the variant calling pipelines have been extensively validated. It therefore provides a unique opportunity for a high quality, in-depth study of tumour heterogeneity.

Table of contents

Chapter 1

Introduction

1.1 Cancer

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

1.2 Hallmarks

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

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

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

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

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

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

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

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

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

1.3 Genetics

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

1.3.1 Early observations suggest a role for the genome

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

1.3.2 Confirmation of a role in cancer

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

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

1.4 Evolution

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

1.4.1 Estimates of the number of mutations to form a cancer

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

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

1.4.2 An integrated theory of tumour evolution

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

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

1.5 Drivers

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

1.5.1 Oncogenes

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

1.5.2 Tumour suppressor genes

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

1.5.3 Drivers and passenger mutations

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

1.6 High throughput technology

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

1.6.1 Array based technology

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

1.6.2 Sequencing technology

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

1.6.3 The emergence of sequencing consortia

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

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

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

1.7 Copy number

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

1.7.1 Confirming classic knowledge

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

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

1.7.2 Pan-cancer overview of CNAs

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

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

1.8 Massively parallel sequencing of cancer genomes

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

1.8.1 Early findings from exome sequencing

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

1.8.2 The full compendium of somatic alterations

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

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

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

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

1.9 Mutational processes

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

1.9.1 Automated extraction of mutational signatures

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

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

1.9.2 Linking signatures to mutational processes

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

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

1.10 Heterogeneity

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

1.10.1 Detecting heterogeneity from sequencing data

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

1.10.2 Cancer type specific studies highlight evolutionary properties

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

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

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

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

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

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

1.10.3 Pan-cancer studies reveal widespread ITH across cancer types

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

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

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

1.11 Subclonal inference

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

1.11.1 Clustering of mutations

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

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

1.11.2 Statistical and computational strategies for subclonal reconstruction

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

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

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

1.12 Copy number calling

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

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

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

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

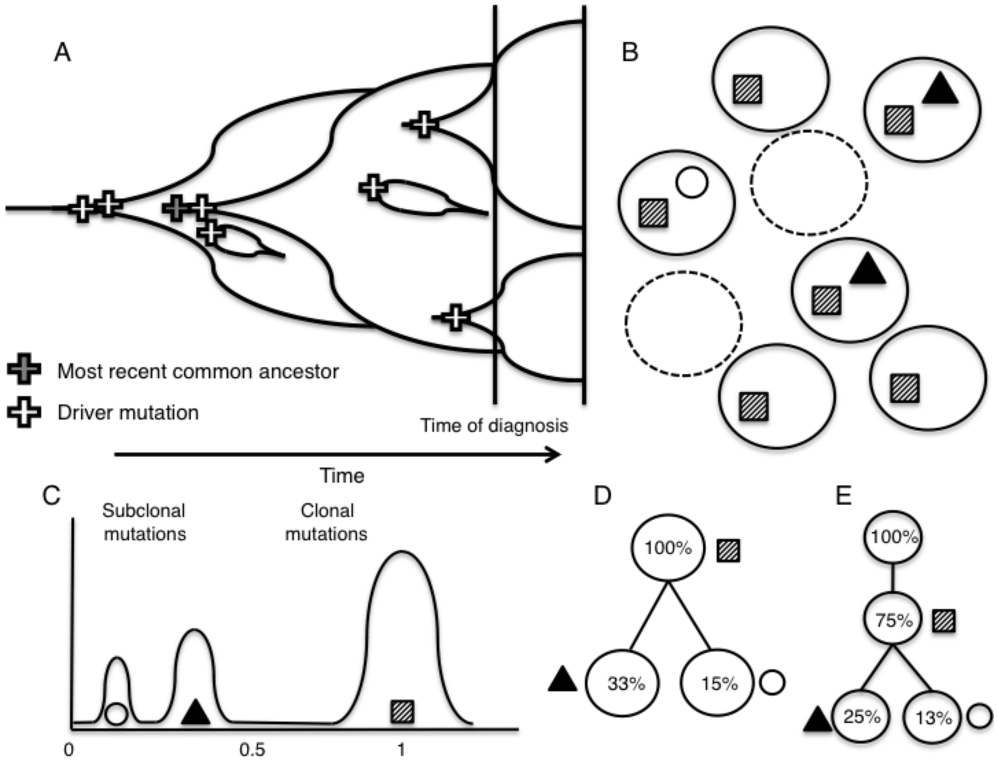


Fig. 1.1 (A) During their lifetime, tumours acquire mutations, of which drivers can lead to clonal expansions. (B) At any time, a tumour consists of multiple populations of cells, tumour cells (circles with continuous line) and infiltrating normal cells (dashed circles). As mutations are acquired gradually,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

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

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

1.13 Tumour micro-environment

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

1.13.1 Carcinoma-associated fibroblasts

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

1.13.2 Tumour-associated macrophages

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

1.13.3 Tumour-infiltrating lymphocytes

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

1.13.4 Tumour-associated neutrophils

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

1.13.5 Other cell types

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

1.13.6 Immune evasion

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

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

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

1.14 Clinical implications of heterogeneity

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

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

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

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

1.15 Summary

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

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

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

Chapter 2

Methods

2.1 Principles of subclonal reconstruction

Reconstruction of the subclonal architecture of a tumour involves three main components: Estimating copy number, adjusting SNV VAFs for copy number alterations to obtain CCF values and inferring the subclonal architecture from the CCF data. This section contains a description of all methods that I use for subclonal inference. These methods form the basis of all results reported on in this thesis, sometimes as part of a much larger procedure as is detailed in Chapter 6. This chapter also contains a description of avenues that have been explored, but were deemed not an improvement. An earlier version of the text in this section has appeared in Dentro et al. (2017).

2.2 The Battenberg algorithm

Battenberg was originally developed to study the unique PD4120 sample and was briefly described in the supplement of Nik-Zainal et al. (2012a). Since then it has been adapted and extended to run with whole genome sequencing and SNP 6.0 data from 1000s of genomes and has become a standard part of the cancer genome analysis pipelines at the Sanger. This section contains a complete description of the whole genome sequencing pipeline and algorithm. In brief: Battenberg uses the 1000 Genomes SNP locations with B-allele frequency (BAF) and relative amounts of DNA (logR) as input from either whole genome sequencing or SNP 6.0 arrays. Heterozygous SNPs are identified from the matched normal sample, after which the SNPs are phased into haplotype blocks to obtain accurate BAF values. Battenberg then performs segmentation, finds an initial purity and ploidy combination before fitting a global copy number profile. Finally, it identifies segments for which the underlying BAF cannot be explained by clonal copy number and it will fit subclonal copy number as a mixture of two separate major and minor allele states.

2.2.1 Pre-processing

Battenberg starts by reading in allele counts for all 1000 genomes SNPs, which are directly obtained from the tumour and matched normal BAM. SNPs are removed from the pool if they appear on the list of unreliable SNPs (identified in a panel of 200 normal genome sequences) or when they are covered by fewer than 10 reads in the normal or 1 read in the tumour. The normal is used to identify SNPs that are heterozygous in the germline of the patient and therefore requires that the normal is from the same individual as the tumour. All SNPs then go into haplotype reconstruction, after which the germline heteroygous SNPs are used for segmentation and fitting.

2.2.2 Reconstructing haplotype blocks

Battenberg primarily uses allelic imbalances to estimate copy number. To observe these imbalances, it is helpful to look at the B-allele frequency (BAF) of a germline heterozygous SNP. For sequencing data the BAF can be calculated as:

$$BAF_i = \frac{r_{B,i}}{r_{A,i} + r_{B,i}} \tag{2.1}$$

where $r_{A,i}$ and $r_{B,i}$ represent the total reads reporting allele A and B respectively. Alternatively, the BAF can be expressed as a function of the number of chromosome copies of allele A and B (n_A and n_B respectively):

$$BAF_i = \frac{n_{B,i}}{n_{A,i} + n_{B,i}} \tag{2.2}$$

A germline heterozygous SNP will have a BAF of approximately 0.5 in the absence of any copy number changes. Deviations from 0.5 therefore can be used to detect somatic aberrations. As tumours are often admixed with normal cells, establishing the copy number state of an aberration based on the deviation of BAF requires estimating the fraction of tumour cells in the sample (the tumour purity). The number of chromosome copies in the formula above should therefore be split into a contribution of ρ tumour cells and (1-ρ) normal cells:

$$BAF_i = \frac{\rho n_{B,t} + (1-\rho)n_{B,n}}{\rho(n_{A,t} + n_{B,t}) + (1-\rho)(n_{A,n} + n_{B,n})} \tag{2.3}$$

where ρ represents the tumour purity, $n_{A,t}$ and $n_{B,t}$ the number of chromosome copies in tumour cells and $n_{A,n}$ and $n_{B,n}$ the number of chromosome copies in normal cells. Several methods have been developed to co-estimate clonal copy number states and tumour purity based on these allele-specific signals (Carter et al., 2012; Ha et al., 2014; Van Loo et al., 2010).

Tumours that exhibit much clonal genomic instability will show deviation of the BAF for large proportions of the genome. In such tumours, the BAF values show clear levels corresponding to different clonal states, which translates into more usable signal for methods that co-estimate copy number states and tumour purity. However, genomes that show large amounts of subclonal genomic instability will show a range of different BAF values and will be more difficult to fit.

Fig. 2.1 shows allele frequency values for a number of example cases that are affected by copy number changes and different normal cell admixtures. Panel A shows a region with no copy number alterations in a tumour that has no normal cell infiltration. One expects both alleles to be present in equal proportions, resulting in allele frequencies of 0.5. Panel B shows a region with a clonal gain. The bands representing allele A and B are clearly separated, with allele A representing two thirds of the total chromosome copies and allele B one third. Panel C contains a similar gain, but in a sample with 75% tumour purity, resulting in a smaller difference between the bands. Panel D shows the gain, again with 75% tumour cells, but now the coverage is reduced from 100X (as in panels A, B and C) to 40X. The bands appear to be overlapping as lowering the depth increases the noise and widens the bands. Panel E shows an example where the gain is subclonal in 60% of tumour cells resulting in further overlap of both bands. And finally panel F shows a subclonal loss in 40% of tumour cells.

Fig. 2.1 illustrates that the allele frequencies of individual SNPs are subject to statistical variation and this noise increases with lower coverage. Combining SNPs into haplotype blocks through phasing can mitigate this effect (Carter et al., 2012; Nik-Zainal et al., 2012b). Through haplotype phasing, information can be combined across multiple SNPs within a region of copy number change, by matching alleles across SNPs. For example, for SNP i, allele A may correspond to the maternal allele, while for SNP $i+1$, allele B may correspond to the maternal allele. If these are combined appropriately, smaller deviations of the BAF from the normal state can be detected, and higher precision copy number changes, including subclonal copy number changes, can be inferred.

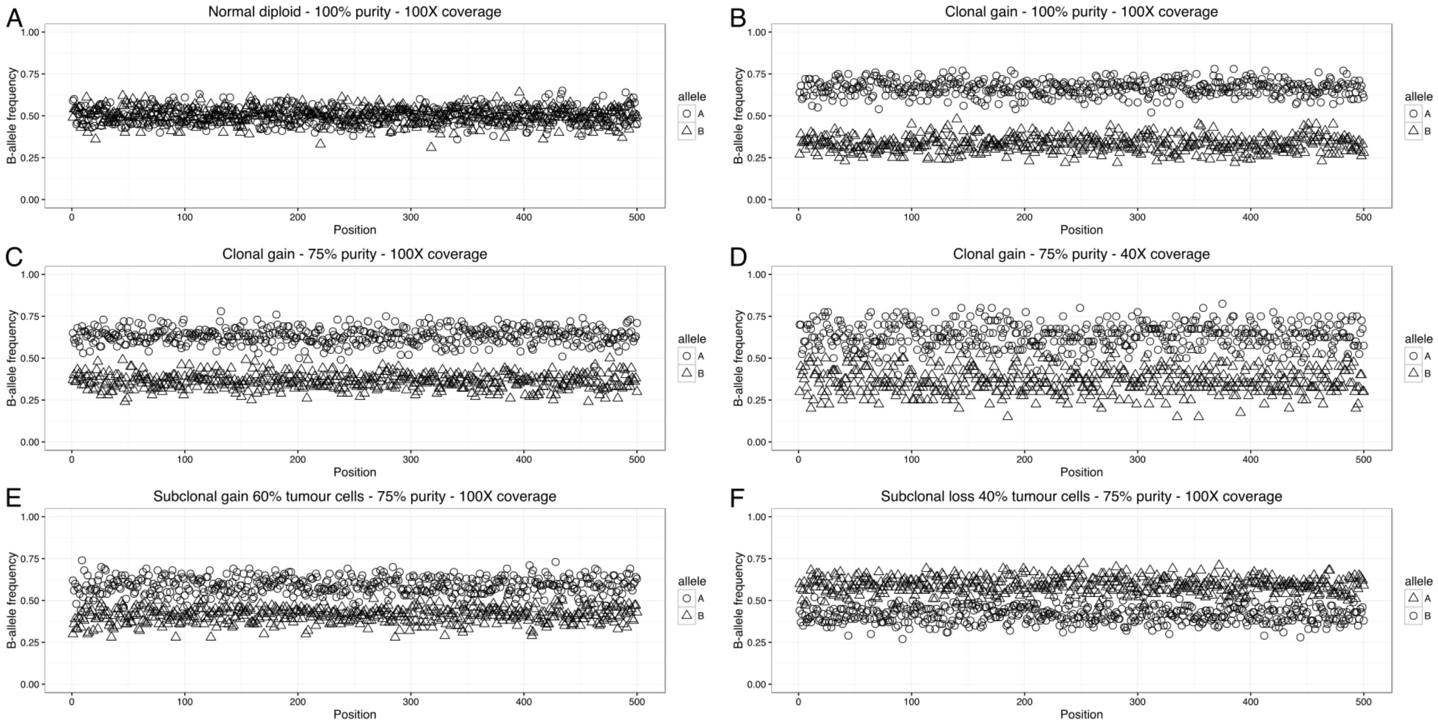


Fig. 2.1 The BAF is used to determine whether subclonal copy number exists, often in low fractions of overall cells within the sequencing sample. This figure illustrates that for alterations in low proportions of cells one should perform haplotype reconstruction to reduce noise. (A) The BAF centres around 0.5 (50% of sequenced alleles contain the SNP, as it is heterozygous in a background of two chromosome copies) when no copy number alterations have occurred and the circles (allele A) and triangles (allele B) appear exactly on top of each other. (B) When an alteration occurs, the BAF changes showing two clear ‘bands‘ (at 100% purity and 100X coverage). In this scenario an unbiased estimate of the BAF can be obtained from either of the bands separately. (C) When the tumour purity drops to 75% the two bands appear closer to each other and closer to 0.5, but at 100X they still visually separate. (D) When the gain occurs in a purity of 75%, but the tumour is sequenced at 40X instead of 100X the bands both widen considerably and start to merge. By calculating the BAF from one of the bands as both bands now contain both circles and triangles. (E) So far, only clonal gains have been simulated. This panel shows a subclonal gain in 60% of tumour cells in a tumour with 75% purity sequenced at 100X. Compared to panel C, the bands are much closer together and more difficult to separate. Panel F shows that the bands are equally difficult to separate form a loss in 40% of tumour cells occurs.

2.2.3 Fitting a global copy number profile

Two main components must be taken into account when fitting a copy number profile: Infiltration of normal cells (tumour purity) and tumour cell aneuploidy (tumour ploidy). Battenberg takes a similar approach as ASCAT (Van Loo et al., 2010) by considering a range of purity and ploidy combinations to pick a solution. After a combination is established, each segment is then assigned allele specific copy number states.

The grid search procedure is performed twice, first with large steps to find an initial optimum and then with small steps to refine the solution. The grid search procedure takes a range of purity (ρ) and ploidy (ψ_t) values and calculates the proportion of the genome fit with clonal copy number with each combination, through the steps described in the next section. It then picks the ρ and ψ_t pair that maximises the proportion of the genome with clonally altered copy number.

Finally, the copy number states of both alleles of a segment s are established through:

$$n_{A,s} = \frac{\rho - 1 - (1 - b_s)2^{l_s}(2(1-\rho) + \rho\psi_t)}{\rho} \tag{2.4}$$

$$n_{B,s} = \frac{\rho - 1 + b_s 2^{l_s}(2(1-\rho) + \rho\psi_t)}{\rho} \tag{2.5}$$

where $n_{A,s}$ is the copy number call for allele A of segment s, b_s and l_s are the BAF and logR of the segment and ψ_t is the average ploidy of all tumour cells in the sequencing sample.

2.2.4 Testing whether a segment is clonal

After fitting clonal major and minor allele copy number states, we can test whether these states accurately explain the observed BAF. If the BAF is not well explained by the best clonal states, then the segment is subclonal. This section explains the details of the test, the next section explains how the test is applied. The obtained n_A and n_B (through eqs. 2.4 and 2.5) can be non-integer values and therefore have to be rounded to obtain clonal copy number states. This can be achieved by rounding either allele up or down, yielding four possible options (explained further in the next section). For each option the expected BAF, given rounded alleles $\widehat{n}_{A,s}$ and $\widehat{n}_{B,s}$, is calculated using:

$$\widehat{b}_s = \frac{1-\rho-\rho\widehat{n}_{A,s}}{2+2\rho+\rho(\widehat{n}_{A,s}+\widehat{n}_{B,s})} \tag{2.6}$$

A choice is made between the four options by taking the combination of alleles that minimises the distance between the observed BAF b_s and the expected BAF $\widehat{b}_s$.

Finally, the $\widehat{b}_s$ value corresponding to the chosen allele combination is tested against the observed BAF through a t-test and accepted as clonal if the p-value is not significant _using 0.05 as the significance cutoff.

2.2.5 Fitting subclonal copy number

Once exact allele frequencies of segments have been calculated and a clonal copy number profile has been fit, subclonal copy number changes can be detected. As a first step, for each segment, one can determine whether the BAF value of this segment can be explained by a clonal copy number change (as detailed in the previous section). Deviation of the observed exact allele frequency from the theoretical allele frequency can be used to identify a segment having a subclonal copy number state, i.e. a combination of two or more populations of tumour cells with different copy number states, in addition to a population of normal cells.

When such a segment is fit with a clonal copy number state, the multiple subclonal states are combined into a single (integer) representation. For example, if the real copy number state of the segment is 2+1 (2 copies of one parental allele and 1 copy of the other allele) in 80% and 1+1 in 20% of tumour cells (i.e. on average 1.8+1), its clonal fit will likely be 2+1 in 100% of tumour cells (1.8+1 rounded up). The observed allele frequency will therefore deviate from the frequency expected under the clonal copy number fit, allowing us to infer that the segment cannot be explained by a clonal copy number state.

The type of subclonal copy number depends on the different copy number states at the locus and their respective fractions of tumour cells. This problem has multiple solutions, as there can be any number of subclones with distinct subclonal copy number states. However, for any given segment, the most parsimonious assumption is that there are only two distinct copy number states, and that those copy number states differ at most by one chromosome copy (i.e. are separated by only one copy number event). Battenberg therefore assumes two distinct major and minor allele states, which are separated by one copy number event.

Under this assumption, given allele-specific copy number values n_A and n_B (integer if clonal, non-integer if subclonal), there are four options for the theoretical clonal allele frequency $\widehat{h}_f$ (assuming diploid copy number in the normal cell population):

Allele A and B are both rounded down:

$$\widehat{h_f} = \frac{\rho \lfloor n_B \rfloor + 1 - \rho}{\rho(\lfloor n_A \rfloor + \lfloor n_B \rfloor) + 2(1-\rho)} \tag{2.7}$$

Allele A is rounded down and B is rounded up:

$$\widehat{h_f} = \frac{\rho \lceil n_B \rceil + 1 - \rho}{\rho(\lfloor n_A \rfloor + \lceil n_B \rceil) + 2(1-\rho)} \tag{2.8}$$

Allele A is rounded up and B is rounded down:

$$\widehat{h_f} = \frac{\rho \lfloor n_B \rfloor + 1 - \rho}{\rho(\lceil n_A \rceil + \lfloor n_B \rfloor) + 2(1-\rho)} \tag{2.9}$$

Allele A and B are both rounded up:

$$\widehat{h_f} = \frac{\rho \lceil n_B \rceil + 1 - \rho}{\rho(\lceil n_A \rceil + \lceil n_B \rceil) + 2(1-\rho)} \tag{2.10}$$

Subclonal segments can be identified by testing the observed allele frequency h_f against the theoretical $\widehat{h_f}$ of all four scenarios and accepting a segment as subclonal if the observed h_f is significantly different from $\widehat{h_f}$ in all. If the segment is deemed to be subclonal we choose one of the above four scenarios as the most likely explanation of how subclonal copy number was rounded into clonal. The scenario that explains the observed h_f best is picked, providing two combinations of major and minor allele copy number states.

Finally, having obtained the states, we estimate the proportions of tumour cells that contain each of the two major and minor allele combinations. Formally, if a fraction of tumour cells τ shows copy number state $n_{A,1}$ + $n_{B,1}$ and a fraction of tumour cells 1-τ shows copy number state $n_{A,2}$ + $n_{B,2}$, τ can be calculated as:

$$\tau = \frac{1 - \rho + \rho n_{B,2} + 2h_f(1-\rho) - h_f\rho(n_{A,2} + n_{B,2})}{h_f\rho(n_{A,1} + n_{B,1}) - h_f\rho(n_{A,2} + n_{B,2}) - \rho n_{B,1} + \rho n_{B,2}} \tag{2.11}$$

2.2.6 Extensions to segmentation

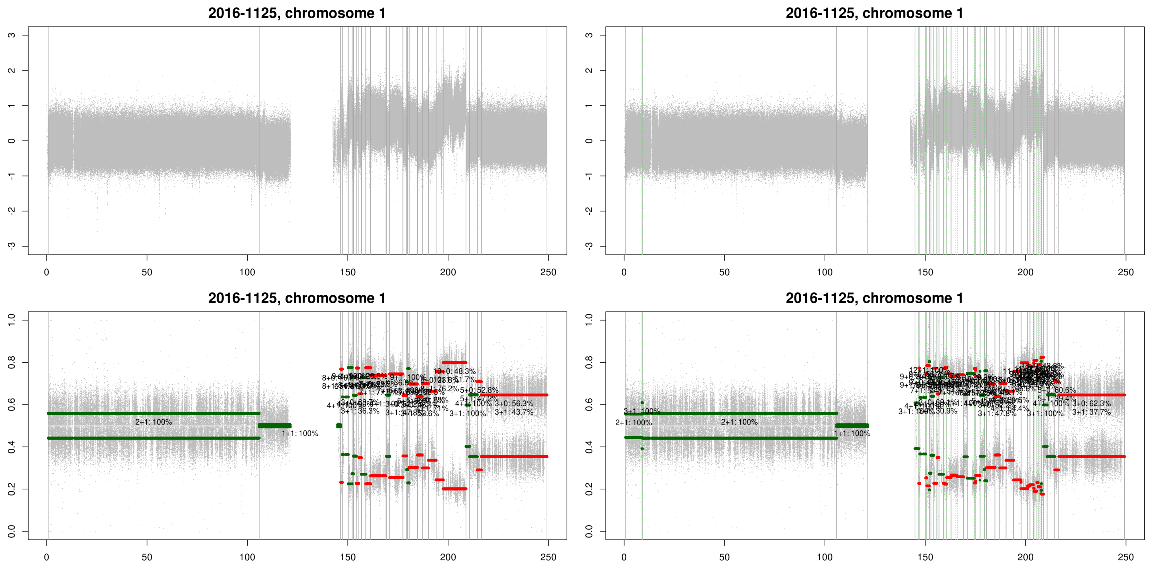
Segmentation of the phased BAF data is performed by piecewise constant fitting (PCF) in Battenberg. PCF models the data as a step-function to explain the observed data by a number of discrete copy number segments as described in (Nilsen et al., 2012). PCF is provided with BAF data for heterozygous SNPs and requires two parameters: the penalty for starting a new segment and a minimum segment length defined by the number of supporting SNPs. That means a new segment always starts with a heterozygous SNP and the startpoint may not be precise as the parameters require sufficient evidence of a step in the BAF signal before a new breakpoint is added. Finally, Battenberg does not use the logR for segmentation, which means a region in which both alleles are gained are difficult to detect as the BAF does not change.

I have therefore added the option to incorporate pre-defined breakpoints into the segmentation procedure (see Fig. 2.2 for an example). This allows for inclusion of breakpoints with base-pair resolution from SV calling. The approach starts with pre-segmenting the genome with the supplied breakpoints. It assumes the breakpoints are clean and therefore performs no further filtering. Then PCF is performed in each pre-segment to detect further breakpoints not covered by a SV, such as a chromosome arm event. However, not every structural variant constitutes a copy number change (inversions for example) and the SVs can therefore lead to spurious segments. A segment merging step is therefore added that formally tests the BAF and logR of each adjacent pair of segments through a t-test and merges the pair if the BAF and logR are not significantly different or when the major and minor allele of both segments have the same clonal values. An exception is made for segments between which there is a gap of 3Mb or larger. The assumption is made that there is either missing data or a centromere between the segments and as there is no data we make no call.

2.2.7 GC content correction

Coverage of sequencing data can be affected by artefacts that manifest themselves as a wave pattern across the genome (Diskin et al., 2008). These artefacts are correlated with local GC content and can be corrected for by a regression approach (Benjamini and Speed, 2012; Diskin et al., 2008). I observed that a substantial set of tumours reported on in this thesis are affected by this problem. Fitting an initial copy number profile was impossible as it yielded whole chromosome homozygous deletions where the profile looked generally correct for other chromosomes (Fig. 2.3, with details of chromosome 8 in Fig. 2.4). These deletions would be surprising given that about 10% of genes are thought to be essential for cell function (Wang et al., 2015), which makes it likely that every chromosome contains at least one gene required for cell survival. I have therefore implemented an approach for Battenberg that corrects the relative tumour coverage (logR) for wave patterns.

Similarly to the method implemented in ASCAT, the GC content correction function considers each SNP given in the input as the centre-point of a window. The GC content for window-sizes varying from 25kb to 10Mb have been pre-calculated. Similarly to ASCAT, we consider two window sizes to correct for high and low frequency waves. After calculating correlations the data with the GC content of the logR data we select a window $<$ 1Mb



(a) Without SV breakpoints

(b) With SV breakpoints

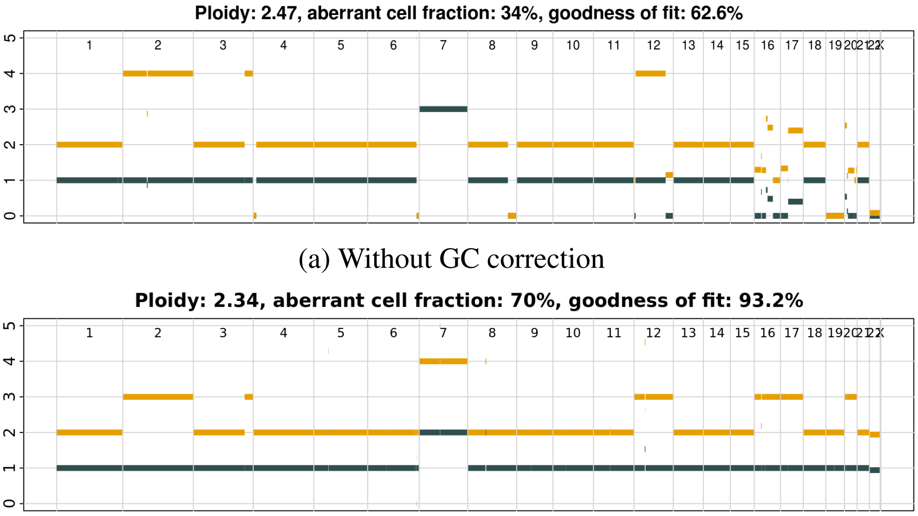
Fig. 2.2 The above figures show the segmented data with a copy number fit on a chromosome that consists of many small segments. The top plot in both figures contains the GC corrected raw logR data (grey dots) with the segment boundaries overlayed (vertical lines). The bottom plot contains the BAF with fit segments overlayed (green represents clonal copy number, while red represents subclonal). **a**) The fit without inclusion of SV breakpoints misses a series of consecutive breakpoints around 200Mb. **b**) After inclusion of the SV breakpoints (green vertical lines) Battenberg is able to call all visible segments on this complex chromosome.

(denoted as $w < 1$) and one $>= 1$Mb ($w >= 1$) and perform regression on a model that allows for both a linear and a non-linear effect of GC content:

$$l = G_{w<1} + G_{w>=1} + G^2_{w<1} + G^2_{w>=1} \tag{2.12}$$

where G is the precalculated GC content data. The residuals (expected logR) are then taken as the corrected logR value and saved for use further down the pipeline.

This approach corrects for the majority of the wave effect and has allowed a substantial number of tumours to be included in the analysis described further into this thesis. It does however not completely remove the artefacts (see Fig. 2.4b), which suggest that there are additional factors that have not yet been accounted for.



(a) Without GC correction

(b) With GC correction

Fig. 2.3 Whole copy number profile for sample SA514993, with in orange the total copy number and in dark grey the minor allele. (**a**) A copy number profile with homozygous deletions on chromosomes 4, 6, 8, 19 and 22. (**b**) The homozygous deletions disappear after correction for GC content. The purity estimate also increases, which reduces the gains on chromosomes 2 and 12 by one copy and on chromosome 7 by three copies.

2.3 Subclonal architecture inference with DPClust

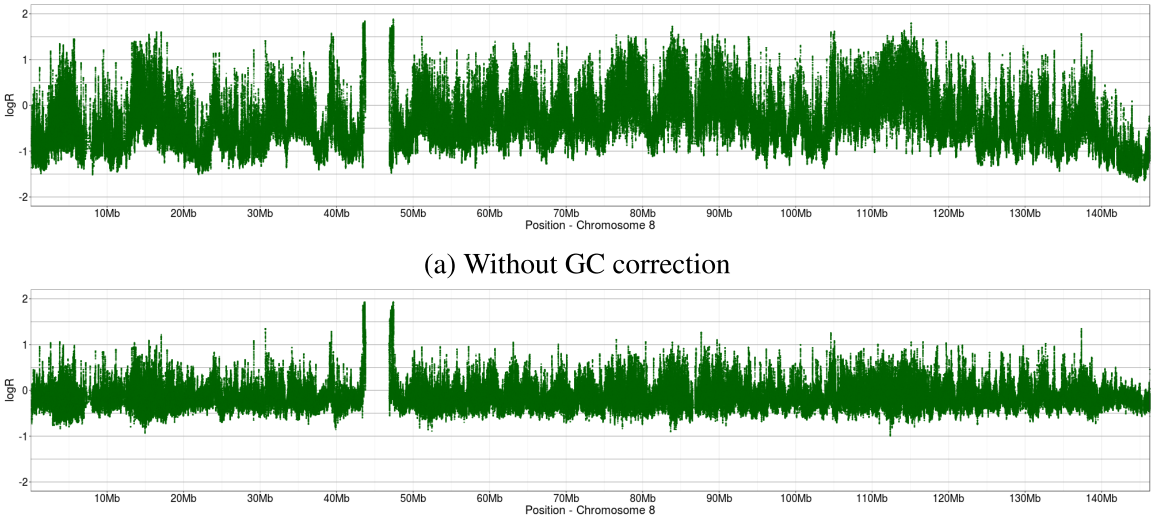
A subclone is a population of tumour cells that carry a unique subset of mutations (SNVs, indels or copy number). These mutations will appear in a similar fraction of tumour cells in the sequenced sample and can therefore be used as a marker of the population. By clustering the mutations, one can infer the existence of a subpopulation and therefore the subclonal architecture contained within the sequencing sample.

For such an approach to work one must assume that mutations occur only once during the life time of the tumour, which is referred to as the *infinite sites assumption* (Jiao et al., 2014). For SNVs and indels that assumption holds true in general given the size of the human genome, but for copy number alterations there is accumulating evidence that the same locus can be mutated on multiple occasions (Jamal-Hanjani et al., 2017).

This section describes the approach implemented in the DPClust software package.

2.3.1 Estimating cancer cell fractions

To infer the subclonal architecture of a tumour one must first obtain an estimate of the fraction of tumour cells (cancer cell fraction, CCF) for each mutation, which can be inferred from VAFs of SNVs. Massively parallel sequencing results in short reads, which can then be



(a) Without GC correction

(b) With GC correction

Fig. 2.4 LogR data of chromosome 8 from sample SA514993. Smoothing was performed by applying a running median with a window-size of 101 SNPs to make the signal more visible at this scale. (**a**) Raw logR before GC correction shows a long wave pattern with a varying frequency. The homozygous deletion visible in Fig. 2.3a is situated at about 140Mb where the logR is clearly sloping downwards. (**b**) The big steps in logR are removed after correcting for GC content. The sloping at around 140Mb is reduced dramatically, now stopping Battenberg from calling a homozygous deletion (Fig. 2.3b). A light wave pattern is still visible, suggesting further improvements can be made.

aligned to a reference genome, followed by SNV calling. Both the variant and reference alleles of an SNV are supported by a number of reads, r_{mut} and r_{ref} respectively. The VAF of SNV i, f_i, can straightforwardly be calculated as:

$$f_i = \frac{r_{mut,i}}{r_{mut,i} + r_{ref,i}} \tag{2.13}$$

However, mutation clustering to identify (sub)clonal populations cannot be performed directly using VAFs, as copy number changes impact allele frequencies. Fig. 2.5 shows four SNVs in a sample that consists of 80% tumour cells and 20% normal cells. SNV 1 is clonal and occurs in a region with a normal diploid copy number state. This mutation is therefore carried by approximately half the reads that represent tumour DNA. SNV 2 is subclonal and also occurs in a region of normal diploid copy number. As both copy number and normal cell contamination are equal for both SNV 1 and 2, their allele frequencies are directly comparable and proportional to the fraction of tumour cells by which they are carried. SNV 3 falls into an area that was subclonally lost. As the subclonal loss has occurred on the other allele, this SNV's VAF is increased compared to SNV 1. SNV 4 is clonal, falls into an

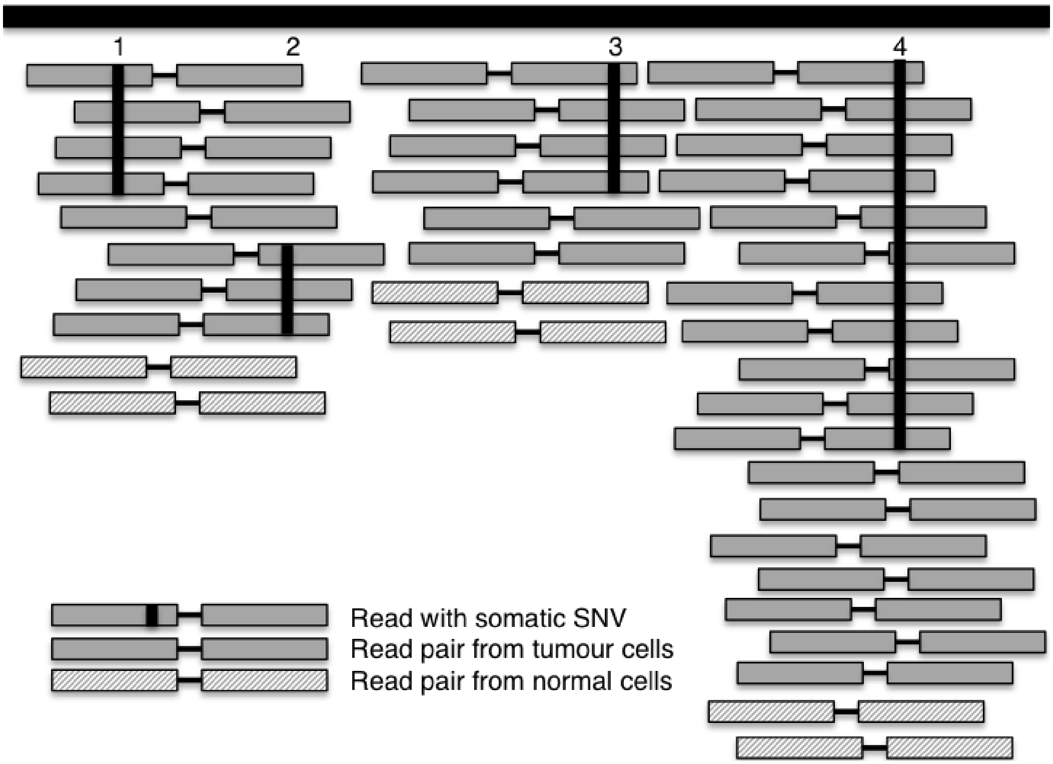


Fig. 2.5 Allele frequencies of SNVs must be transformed to Cancer Cell Fractions, accounting for copy number changes, before they can be clustered to identify subclonal populations. This illustration shows 4 SNVs in different (sub)clonal populations and in regions with different copy number states, to illustrate this principle. SNVs 1 and 2 are clonal and subclonal respectively and appear in a non-aberrated copy number state. SNV 3 coincides with a subclonal deletion, with the SNV falling on the retained allele (i.e. the other allele is subclonally deleted). SNV 4 has occurred before a gain and is therefore carried by two chromosome copies. Even though SNV 1, 3 and 4 are clonal, their allele frequencies differ due to copy number alterations.

area that is clonally gained and is on the gained allele. Its VAF is therefore higher than that of SNV 1. If these SNVs were clustered in VAF space, SNVs 3 and 4 would be mistaken for evidence of additional mutation clusters, while they in fact belong to the clonal cluster.

This example illustrates that the copy number state of an SNV, also called its multiplicity, is key to understanding VAF distributions of mutations. Estimating the multiplicity of an SNV is challenging, as it requires establishing the copy number state of a single base. Copy number callers often estimate copy number states for large stretches of DNA, which might not accurately represent the copy number state exactly at the base of the SNV. To assist with resolving this issue, it is helpful to consider the product of mutation multiplicity m_i of a mutation i and its cancer cell fraction CCF_i:

$$u_i = CCF_i m_i \tag{2.14}$$

Let us consider the properties of u_i. A clonal SNV will have a CCF of 1.0 (i.e. 100% of tumour cells) and in each cell the number of chromosome copies, m_i is an integer. It follows from the above equation that for clonal mutations $u_i \geq 1$. A subclonal mutation has a CCF less than 1.0 (for example 0.4, or 40% of tumour cells) and can only be carried by a single chromosome copy (unless also affected by a subclonal CNA), therefore m_i = 1. It follows that $u_i < 1$ for subclonal mutations. We can use these observations to obtain m_i from u_i:

$$m_i = \begin{cases} |u_i|, & \text{if } u_i \geq 1 \\ 1, & \text{if } u_i < 1 \end{cases} \tag{2.15}$$

Furthermore, u_i can be written as a function of the fraction of tumour cells ρ with a total number of chromosome copies in tumour cells at locus i, $n_{tot,t,i}$, and a fraction of normal cells 1-ρ with a total number of chromosome copies in normal cells at locus i, $n_{tot,n,i}$:

$$u_i = f_i \frac{1}{\rho}[\rho n_{tot,t,i} + (1-\rho)n_{tot,n,i}] \tag{2.16}$$

In the formula above, ρ and $n_{tot,t,i}$ can be obtained through copy number analysis, f_i can be calculated from r_{mut} and r_{ref} using Eq. 2.13, and the $n_{tot,n,i}$ values are considered known (typically 2). This equation therefore provides us with a way to calculate u_i and by extension to obtain the multiplicity of the SNV.

SNV 1 in Fig. 2.5 for example is clonal and has 4 reads reporting the variant and 6 reporting the reference allele. The purity is 0.8 (80% of total cells are tumour cells) and the total copy number of both the tumour and normal cells is 2. Its u_i therefore becomes:

$$\frac{4}{4+6} \times \frac{1}{0.8} \times [0.8 \times 2 + 0.2 \times 2] = 1.000 \tag{2.17}$$

Which translates into a CCF of 1.0 via Eq. 2.15. While for SNV 4 it yields:

$$\frac{11}{11+9} \times \frac{1}{0.8} \times [0.8 \times 3 + 0.2 \times 2] = 1.925 \tag{2.18}$$

Which also translates into a CCF of 1. SNV 4 illustrates that u_i must be rounded to obtain the multiplicity of a clonal SNV. It differs slightly from the expected value 2 because of variability in the number of reads due to limited sequencing depth. A similar mutation with 12 variant reads out of 20 would lead to an estimate of 2.100.

The accuracy of the multiplicity estimate in practice depends on the accuracy of the VAF and local copy number. Slight deviation in the VAF due to read sampling can result in minor deviation of the multiplicity estimates, as illustrated in the example above. Incorrect copy number profiles may also result in large errors if, for example, the CNA profile has been called diploid instead of tetraploid. Ambiguity in estimating whole genome duplications is a difficult problem in copy number analysis. If a copy number profile is erroneously called as diploid then SNVs carried by two chromosome copies will be estimated to have a multiplicity of 1, while SNVs on 1 chromosome copy will become subclonal as they appear to be on 0.5 copies (e.g. exactly half of tumour cells). The CCF space will therefore show an SNV cluster at exactly 0.5, while the copy number profile may also contain subclonal CNAs at exactly 50% of tumour cells. The uncertainty may be mitigated through the application of a key assumption: a CNA profile is thought to be in its normal state (diploid) unless substantial evidence of a whole genome duplication is available (i.e. the most parsimonious diploid state is assumed unless there is evidence otherwise). However in rare cases, when whole genome duplications occur late and are not followed by other copy number alterations, they leave no traces in the data and it is mathematically impossible to infer from the data available that they occurred.

We now have obtained a series of formulas to calculate CCF from a VAF and copy number profile. First, we obtain u_i through Eq. 2.16 and then calculate the multiplicity and CCF using Eqs. 2.15 and 2.14 respectively.

Finally, we adjust the multiplicity to address SNVs that may appear subclonal due to a subclonal deletion. In these cases it is unknown whether the SNV occurred first and was then deleted in a fraction of cells, or the SNV occurred after the deletion. It is important to account for such subclonal deletions (e.g. by appropriately adjusting multiplicity estimates), and ensure that these subclonal deletions do not result in the inference of spurious subclonal populations.

2.3.2 Filtering

Not all mutations that are provided as input are clustered. Mutations for which there is no copy number are removed because it is not possible to estimate their CCF value. Mutations in regions with fewer than 4 reads total coverage are also removed. Mutations in regions identified with localised somatic hypermutation (*kataegis*) are also filtered out. Short read alignment in regions with kataegis can be difficult because of the many reads carrying one or multiple variant alleles and the fact that kataegis is often observed close to a SV breakpoint. These mutations are removed to reduce the opportunity for a spurious cluster to be inferred.

2.3.3 Algorithm

DPClust clusters SNVs with a similar CCF, derived from VAF values as described in the last section. However, the VAF of a SNV - and therefore also its CCF - can be a relatively coarse measure and is a function of local sequencing depth, which should be taken into account when clustering SNVs. For example, if the SNV falls in a region of diploid copy number with a depth of 20 reads in a sample with 50% tumour cells, its CCF changes by 0.2 when a variant read is added or removed (e.g. 3 mutant reads correspond to a CCF of 0.6, while 4 mutant reads correspond to a CCF of 0.8). If the same SNV is sequenced to 80X depth, one additional variant read would change the CCF by only 0.05. Tumours are often sequenced at 30X average coverage or higher, but this coverage is not constant across the genome. Due to this discrete sampling of mutant and non-mutant reads, and the variability of the sequencing depth, CCF estimates of mutations from specific (sub)clones will show a distribution of values. For example, clonal mutations will display a range of CCF values around 1.0 (Fig. 1.1C).

A suitable error model can account for this variability. The number of variant reads can be seen as the number of successes of n independent coin tosses, where n is the total read depth. The number of successes (variant reads) can therefore be modelled through a binomial distribution with r_i the number of reads reporting the variant at location i, $r_{tot,i}$ the total depth at location i and $r_{tot,i}$ the probability of observing a mutant read:

$$r_i \sim \text{Bin}(r_{tot,i}, p_i) \tag{2.19}$$

Both r_i and $r_{tot,i}$ are observed in the data. p_i can be considered the product of two factors: the proportion of reads one expects to see if the mutation is fully clonal, ζ_i, and the true fraction of tumour cells carrying the mutation π_i:

$$p_i = \zeta_i \pi_i \tag{2.20}$$

ζ_i can be calculated from the tumour purity and the copy number state of the locus, as detailed above. Take for example a clonal SNV in a balanced diploid copy number region in a sequencing sample consisting of 80% tumour cells. The SNV is heterozygous and therefore expected to be carried by half the reads that represent tumour DNA. The expected proportion of reads is therefore 0.5 * 0.8, i.e. 0.4. If the region has three copies and the SNV is carried by two copies, one expects two thirds of the reads representing tumour DNA to be carrying the variant allele, making the expected fraction 2 * 0.8 / (3 * 0.8 + 2 * 0.2), i.e. 0.57.

The key estimate in subclonal reconstruction is the true fraction of tumour cells that are carrying mutation i, π_i. Many methods (Deshwar et al., 2015; Jiao et al., 2014; Landau et al., 2013; Roth et al., 2014) use a Dirichlet Process, which models subclonal fractions as:

$$\pi_i \sim \mathrm{DP}(\alpha P_0) \tag{2.21}$$

where $DP(P_0)$ is a Dirichlet Process with a given probability distribution P_0 and a dispersion parameter α. A realisation of a Dirichlet Process (DP) can be seen as a distribution over a (possibly) infinite sample space, or alternatively as a sampling from an unknown number of unknown distributions (Dunson, 2010). This approach allows for co-estimating both the number of contributing distributions K (the number of cellular populations) and their properties (fraction of tumour cells and number of mutations they contain). The observed sampling represents of the (possibly) infinite number of distributions and can be used to estimate K (i.e. cellular populations) through the stick-breaking representation (Sethuraman, 1994). Stick-breaking implies that the real probability distribution P can be expressed as follows:

$$P = \sum_{h=1}^{\infty} \omega_h \pi_{\theta_h}\ , \qquad \theta_h \sim P_0 \tag{2.22}$$

where π_{θ_h} is a location in CCF space and ω_h represents the probability weight of cluster h

$$\omega_h = V_h \prod_{l<h} (1 - V_h) \tag{2.23}$$

with

$$V_h \sim \mathrm{Beta}(1, \alpha) \tag{2.24}$$

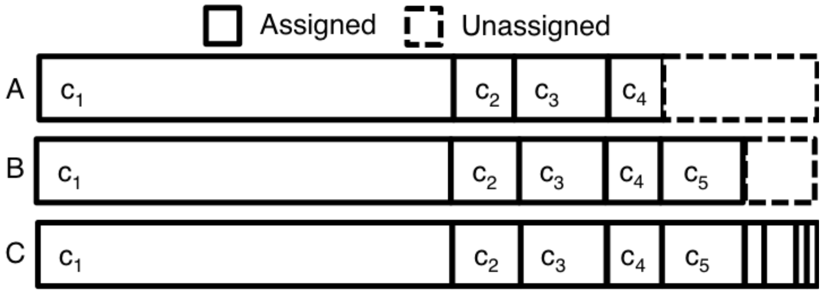


Fig. 2.6 The stick-breaking property of the Dirichlet Process is used to estimate the number of mutation clusters in the data. For each mutation, a stick of arbitrary length is broken into randomly sized bits that represent a cluster. At point A, breaks have been introduced, corresponding to clusters c1-c4. B shows the stick after introducing break 5, while C shows the completed stick-breaking procedure. The size of each broken part represents the weight associated with a cluster and influences the mutation assignments, where a high weight makes it more likely that a mutation is assigned to that cluster. These weights are updated after probabilities for each cluster have been obtained for each mutation, eventually converging on a solution.

The V_h represent parts of a unit length stick that are iteratively broken off from the remaining stick. The V_h get increasingly smaller as more parts are broken off, providing a discrete representation of an infinite space.

Fig. 2.6 symbolizes the stick at various iterations of the stick-breaking procedure. Fig. 2.6A and 2.6B show the stick after 4 and 5 breaks respectively, while Fig. 2.6C shows it after completion. Each substick represents a fraction of the total weight (number of SNVs) of a cluster and can be assigned a CCF through resampling using the assigned SNVs. Then for each SNV and for each substick, a likelihood can be calculated representing the probability that that SNV is generated by that substick, taking the characteristics of the SNV, the stick location and its associated weight into account. After assigning all SNVs, the weights are updated such that they reflect the overall likelihood across SNVs.

The DP models an appropriate number of clusters because the assigned SNVs (influenced by the cluster weight) are used to resample the cluster CCF and the weight represents the fraction of total SNVs assigned to the cluster. By repeating this process over many iterations, the weight and SNV assignments will accumulate in certain locations that correspond to the estimated clusters. Therefore, the DP has the advantage that the number of clusters does not have to be specified *a priori*, making it ideally suited to this problem.

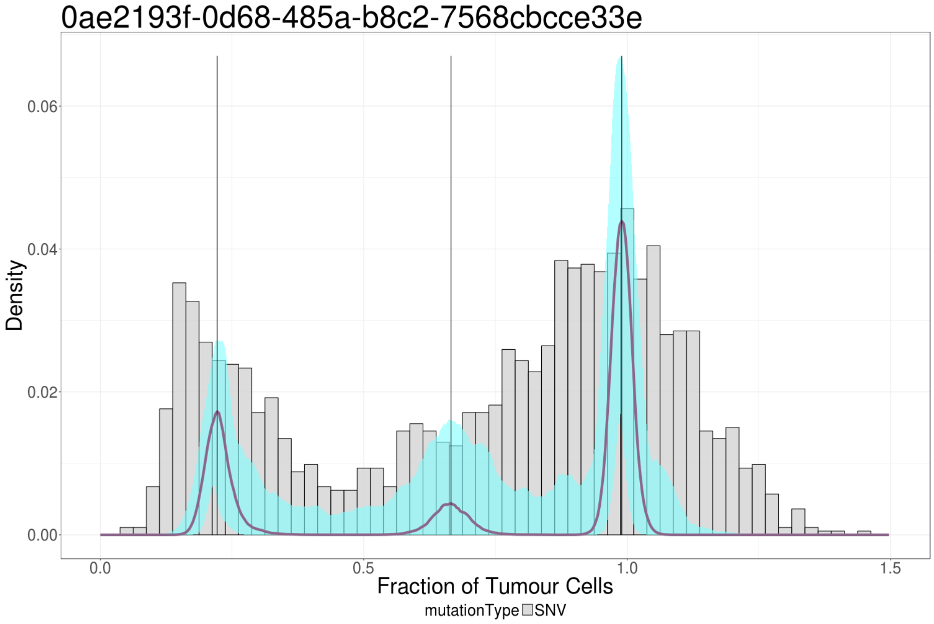


Fig. 2.7 Main output figure from a DPClust run. The grey histogram represents the input SNV data. In front of the histogram is a density line in purple with a turqoise 95% confidence interval. The density line is built up by carefully recording where each SNV is assigned throughout the MCMC iterations. The number of clusters is obtained by obtaining all peaks in the density (vertical black lines). To assign mutations to clusters, first the local minimum density between each pair of cluster locations is obtained. Mutation assignment probabilities are then obtained by going back to the MCMC iterations to record how often each mutation would have been assigned to the final clusters if those were the clusters available at that iteration. The mutation is finally assigned to the cluster with the highest number of assigments.

2.3.4 Post-processing

After completing the MCMC iterations we aim to obtain three estimates: (1) An estimate of the finite number of distributions (cell populations), K, that are present in the input data, (2) the proportion of tumour cells that each population consists of (CCF_k) and (3) likelihoods of each SNV belonging to each population. The number of cell populations K is determined by finding peaks in the posterior weight density Fig. 2.7. In each iteration j the stick-breaking procedure assigns a weight $\omega_{k,j}$ to each cluster that represents its size and the cluster has a $CCF_{k,j}$. Over many iterations weight accumulates in the CCF space, where a large amount of weight corresponds to a high likelihood of the existence of a mutation cluster. We then obtain an estimate of the number of clusters K (cell populations) by obtaining all local maxima in the weight density.

With the K clusters and their locations (CCF_k) established, SNVs can be assigned to clusters. We first establish the CCF area covered by each $k \in K$ by finding the CCF location between each pair of neighbouring clusters that corresponds to the minimum density. The minimum density on either side of a cluster represents its upper and lower CCF bound. Probabilities of a mutation belonging to a cluster are then established by accounting how often a SNV would have been assigned to each k throughout the MCMC iterations. Finally, small clusters smaller than 30 SNVs are removed.

2.3.5 Extension to multi-sample cases

Obtaining multiple samples from the same donor allows for extraction of more detailed subclonal reconstructions. These datasets can consist of multiple tumours taken from different sites (e.g. multiple primary sites, primary and metastasis), multiple samples from the same tumour or multiple samples from the same cancer that represent different time points (e.g. primary and relapse).

Multiple sampling strategies provide a series of advantages. Consider a tumour that has two subclones that each comprise 20% of tumour cells. A single sample analysis will not be able to separate the two groups of mutations as both occur in 20% of tumour cells. But if in another sample the cellular prevalence of the two subclones does vary, one can separate the two groups of mutations. In addition, having multiple samples may help resolve tree topologies. In single sample cases it is often not possible to resolve phylogeny, as more rare subclones may be placed in multiple positions in the tree. By applying the pigeonhole principle across the samples for each subclone, one can often rule out various configurations where a subclone may fit in multiple places in one sample, but not the other. Finally, with multiple sampling strategies, mutations with low allele fractions in one sample can be confirmed (or detected) in another sample where they have higher allele fractions due to higher tumour purity or higher CCF.

Approaches based on a DP can be extended into multiple dimensions (Bolli et al., 2014). The read counts across samples can be modelled as independent draws from n Binomial distributions.

$$\begin{aligned} r_{i,1} &\sim \text{Bin}(r_{tot,i,1}, p_{i,1}) \\ r_{i,n} &\sim \text{Bin}(r_{tot,i,n}, p_{i,n}) \end{aligned} \tag{2.25}$$

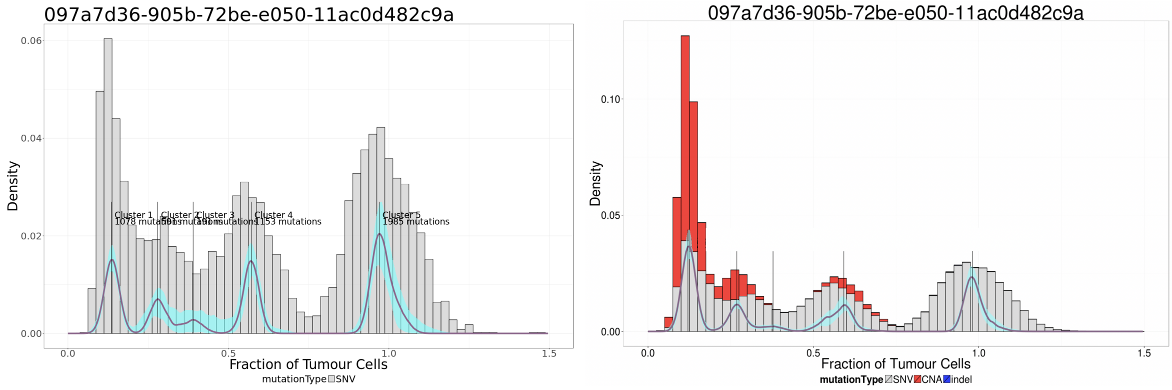
The stick-breaking procedure is performed across the samples where a cluster has a single weight (representing the number of mutations), but a separate location in each of the samples.

Posteriors are obtained across samples by calculating the total probability for each mutation for each cluster under consideration. Finally, the DP can be used to jointly perform clustering and infer phylogenetic relationships between the clusters by interleaving two stick-breaking procedures (Ghahramani et al., 2010).

Several methods for single sample analysis, including PyClone (Roth et al., 2014), SciClone (Miller et al., 2014) and CloneHD (Fischer et al., 2014), can be used to analyse multiple samples. Furthermore, automated tree inference has been implemented in PhyloSub (Jiao et al., 2014) and extended to include SNVs in copy number aberrant regions in PhyloWGS (Deshwar et al., 2015).

2.3.6 Co-clustering of indels and CNAs

Up until now CNAs have only been used to adjust the allele frequency of point mutations. CNAs can also be used to identify cellular populations. The Battenberg algorithm estimates CCF values for each subclonal alteration and it is therefore possible in principle to reconstruct the subclonal architecture through CNAs only, or jointly with SNVs. However, unlike SNVs, there are often far fewer subclonal CNAs measured, which leads to a sparser CCF space and therefore to a reconstruction with less detail. Jointly clustering SNVs and CNAs is preferred



(a) SNV only reconstruction

(b) SNV, indel and CNA combined reconstruction

Fig. 2.8 Subclonal reconstruction on tumour SA6164 (also known as PD4120 and 097a7d36-905b-72be-e050-11ac0d482c9a) using (**a**) only SNVs and (**b**) SNVs, indels and CNAs. There are relatively few indels (blue bars) measured in this tumour, but those available are automatically assigned to mutation clusters. The addition of CNAs (red bars) has a more profound effect, but it does not alter the inferred subclonal architecture substantially. The CNAs provide additional support for clusters 1, 2 and 4 (counted from the left edge of the figure).

as the SNVs will anchor the cluster locations, while CNAs are then assigned to their most likely cluster.

To include CNAs in the clustering process they must be encoded with properties that the DPClust algorithm can understand. The CNA is therefore encoded as an artificial SNV, termed pseudo-SNV. But with a single pseudo-SNV representation it is not immediately clear how many reads should support the pseudo-variant and pseudo-wild-type alleles. A very high coverage could represent a large CNA event, but it would create an artificially high amount of confidence in the VAF, while low read counts do not reflect the size of the CNA events accurately. It is also not directly clear how to balance the evidence between SNVs and CNAs such that one does not dominate the other.

To resolve this issue I encode the CNAs as groups of pseudo-SNVs. First the mutation rate of the tumour is calculated using the measured SNVs. Each CNA covers a certain area of the genome and the equivalent number of mutations that a stretch of DNA would contain given the mutation rate is calculated. Each pseudo-SNV is then assigned a number of mutant and wild-type reads such that the CCF of the SNV corresponds to the CCF of the subclonal CNA.

To mimick read sampling variability the total number of reads are drawn from a Poisson distribution that takes as input the exact depth and the mutant reads are drawn from a binomial that takes the inexact depth and the exact probability of success mandated by the CCF of the CNA. By introducing read sampling variability we transform the pseudo-SNVs into an independent estimate of the CCF of the CNA. The exact total depth is set to either the median depth of all measured SNVs or, if the CNAs cannot be represented by pseudo-SNVs due to insufficient reads per chromosome copy, by 90 reads.

The Battenberg algorithm also provides a measure of confidence in the CCF of each subclonal CNA in the form of a standard deviation on the CCF obtained through bootstrapping. The tighter the standard deviation, the more confident we are in the accuracy of the CCF estimate. The binomial can be used to take this certainty into account by increasing or decreasing the number of trials undertaken. If the number of trials is lower the number of successes given the same probability of success will be more coarse, giving rise to a wider distribution. The total depth is therefore scaled down by the amount of uncertainty, which is represented by the standard deviation. As the standard deviation for the most certain cases is close to 0 we add 1 to it before scaling down the total depth. Finally, the copy number status of each pseudo-SNV is irrelevant and is set to 1 chromosome copy out of 2.

It is important to balance the evidence obtained from SNVs and CNAs such that one does not dominate the other. I have implemented the balancing using the following observation: tumours often have more SNVs than CNAs and each subclonal SNV or CNA is an independent measure of the CCF of a subpopulation of cells. With more samples the estimate of the sampled value becomes more accurate, which gives SNVs an advantage. CNAs however stretch much larger regions of the genome. The evidence is therefore balanced such that the CNAs can provide support for an (extra) cluster, but not dominate the CCF space. For this reason clonal CNAs are represented by a single pseudo-SNV and assigned to the cluster to which the pseudo-SNV is assigned afterwards. Fig. 2.8 shows an example run on the PD4120 tumour that was first described in Nik-Zainal et al. (2012a).

Co-clustering of indels is performed by including the indels as pseudo-SNVs into the input to DPClust. CCF estimates are obtained from the number of reads carrying the variant and wild-type using the procedure described for SNVs. That approach assumes the VAF estimates of the indels are recalibrated by local assembly. Due to alignment difficulties around indels the raw VAF values are often an underestimate. By assembling the local sequence and local realignment of the reads a less biased VAF estimate can be obtained that is useful for subclonal architecture inference.

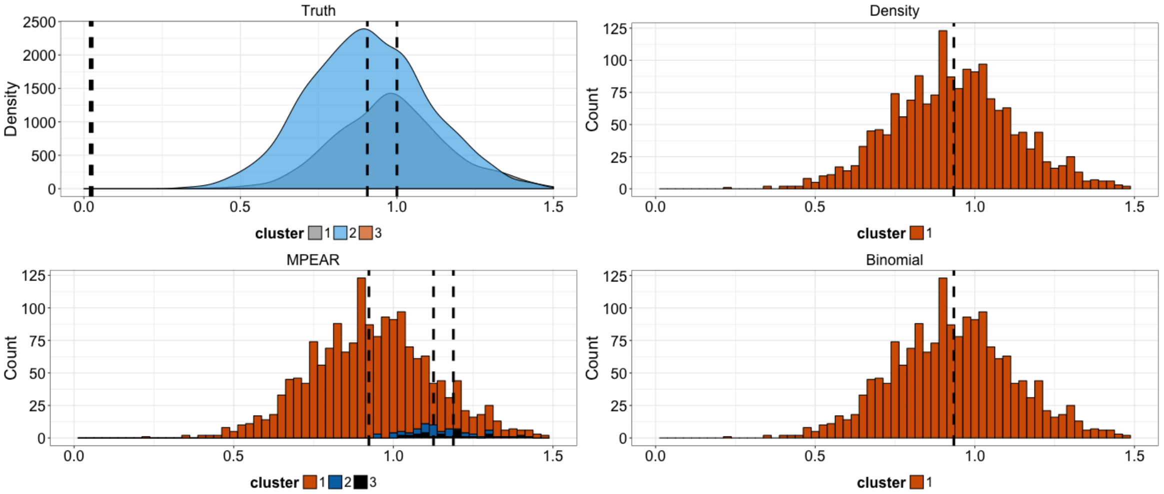


Fig. 2.9 The MPEAR cluster finding approach often finds many small mutation clusters. In this randomly generated example the truth (top left) contains three clusters: A small clone in grey (behind the blue density), a large subclone in blue and a large subclone in orange that falls below the detection limit given this tumours' purity, ploidy and coverage combination. The default density (top right) and binomial assigment (bottom right) approaches find a single cluster in between the major subclone and the clone, effectively merging the two clusters. The size of the clone and its close proximity to the subclone makes it impossible to disentangle the two clusters. MPEAR (bottom left) returns two small additional superclonal clusters in an incorrect position and therefore often requires an additional merging step, more often than the default density approach.

2.3.7 Alternative post-processing steps

In search for increased sensitivity to real clusters I have implemented alternative strategies for obtaining the number of clusters and their contents from the MCMC chain and developed additional procedures for assigning mutations to clusters. The current assignment approach is prone to find small clusters that need to be filtered from the output. It is not easy to come up with a list of criteria that capture these clusters without removing real results. The current implementation of the filtering step removes all clusters below 30 mutations. Often these clusters appear at the end of the data histogram, in the far tail of a large mutation distribution. As the MCMC chain progresses it places a cluster where the large mutation distribution belongs, but depending on its exact placement it leaves the need to explain the far tail with an extra cluster in some iterations. This process is part of the mixing required by a clustering method and it allows the chain to find evidence for extra clusters, but it has the side effect of yielding spurious small clusters. I have therefore attempted to find alternative methods for obtaining clusters that do not have this property. However, none of these new strategies yielded an improvement in performance from evaluation on real and simulated data and have therefore not been used in production.

A new method for obtaining clusters is using hierarchical clustering of mutations followed by a cut of the tree using the MPEAR (maximal posterior expected Rand index) criterion

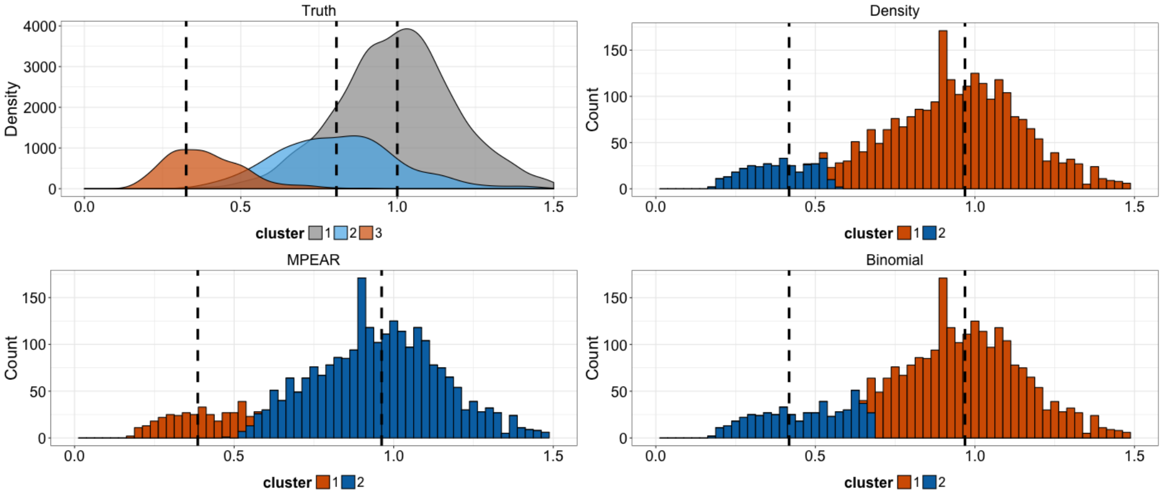


Fig. 2.10 The approach that assigns mutations using the most likely cluster based on the cluster that yields the maximum binomial probability often has the effect of assigning a mutation to its closest cluster. In this example there are three mutation clusters (top left) and all approaches find only two. Both the default density and MPEAR approaches underestimate the size of the subclone slightly (top right and bottom left), while the binomial approach estimates it to be nearly twice the actual size (bottom right).

(Fritsch and Ickstadt, 2009), also used by PyClone (Roth et al., 2014) and BitPhylogeny (Yuan et al., 2015). For this approach I first build a mutation similarity matrix through co-assignment probabilities. Each cell contains the probability that a pair of mutations belong to the same cluster. This matrix is build from the MCMC chain by counting how often the pair is assigned to the same cluster and dividing by the total number of iterations, after excluding the burn-in. After performing hierarchical clustering the MPEAR criterion is applied to k cuts of the tree, with $k < (\lceil \text{mutations}/8 \rceil)$. The cut that yields the maximum score is chosen as the optimum solution. This approach however yields more spurious clusters, it often splits clear existing clusters found by the DPClust default approach into multiple (Fig. 2.9), and the co-assignment matrix cannot easily be constructed for large numbers of mutations.

I have also experimented with an alternative mutation assignment approach. The DPClust default approach is to calculate likelihoods of a mutation belonging to a cluster by counting how often the mutation would have been assigned to that cluster if it had been available in each MCMC iteration. That tends to yield very high probabilities of one cluster, which may not reflect the uncertainty correctly. I therefore wondered if calculating the binomial likelihood would provide a more accurate reflection:

$$\ell_{i,c} = r_{mut,i} \log \mathbf{E}(f_{i,c}) + r_{ref,i} \log(1 - \mathbf{E}(f_{ref,c})) \tag{2.26}$$

Equation 2.26 contains the total number of reads supporting the variant and reference alleles ($r_{mut,i}$ and $r_{ref,i}$) and the expected allele frequency ($\mathbf{E}(f_{ref,c}$) if the mutation belongs to cluster c, calculated using Eq. 2.16. The binomial likelihood however effectively works as assigning the mutation to its closest cluster and therefore tends to overestimate the size of small clusters (Fig. 2.10). It is also a point estimate and does not take into account the cluster size, which the default DPClust assignment approach does. The mutation assignment approach used by Gerstung et al. (2017) calculates beta-binomial probabilities with the inclusion of the cluster size and may be an interesting option in the future.

2.3.8 A downsampling strategy

Clustering a large number of mutations can take a very long time with MCMC based approaches. DPClust uses Gibbs sampling, which means it has to execute a routine for all mutations in every iteration. To improve on runtime and resource usage I have implemented a downsampling strategy that samples mutations and is capable of assigning the mutations not used for clustering afterwards. The routine performs uniform sampling of a specified number of mutations. Large clusters therefore have a higher chance of being sampled from over small clusters, keeping their relative sizes intact. For every mutation not used during clustering I find the mutation with the most similar allele frequency (referred to as its *mate*) that is clustered. By using the allele frequency the selection process is biased towards finding a mate in a similar copy number configuration. After clustering the mutation is assigned to the same cluster as its mate.

I have considered alternative strategies. Selecting copy number segments and using only the mutations in those genomic regions for clustering, but that does not leave fine-grained control over the number of sampled mutations. A biased sampling approach was also considered. It operated by first creating bins across the CCF space and then sampling equally from each bin. That approach changes the shape of the cluster distributions, which detriments the ability to correctly identify clusters. The idea was to perform the biased sampling a number of times and then combine the results from multiple MCMC runs. But preference was given to the unbiased selection due to its simplicity.

Downsampling initially started with 5,000 mutations, which affects nearly half of the tumours reported in this thesis. Later the number of sampled mutations was scaled up to 50,000 after various performance improvements had been implemented, which only affects 134 tumours reported on in this thesis.

2.4 Automated post-hoc tree building

For practical applications it is useful to have an overview of the possible trees that can be built from a given subclonal reconstruction. Nearly all data that I've worked with consists of single sample cases where the tree is difficult to derive, often multiple options are possible and multiple, disjoint, low CCF clusters cannot be disentangled. But for multi-sample cases it is informative and the tree represents the evolutionary story that links the multiple samples together.

I have therefore developed a procedure that builds all possible trees using the DPClust output, which operates regardless of the number of samples. First it classifies each pair of mutation clusters into categories that denote the possible pair-wise relationships. Then the classification is used to iterate over all possible trees, which are provided as a tree structured figure.

2.4.1 Cluster-pair classification

Clusters a and b can have the following relationships: (1) The CCF of a can be strictly greater than b, (2) it can be greater or equal than b, (3) it can be equal, (4) smaller or equal, (5) strictly smaller or (6) it can be unknown. Pairs of clusters are classified into these categories by first establishing the support for each cluster from the MCMC iterations and then sampling pairs of mutations to establish per category.

The classification procedure starts by recording a mutation preferences matrix after mutations are assigned to clusters (Fig. 2.11a and b). This matrix contains a row for each mutation and a column for each cluster and cell (i,j) contains the proportion of MCMC iterations mutation i would have been assigned to cluster j if the final clusters were available.

The approach then iterates over all cluster pairs (Fig. 2.11c). When considering clusters a and b we first sample 1000 mutations from a and b separately to create 1000 mutation pairs. The sampling is performed with replacement to reduce the effect of the different sizes of clusters a and b. Probabilities are calculated by, for each mutation pair (k,l), obtaining how often mutation k is assigned to a lower CCF than mutation l and then aggregating the counts across pairs. The same procedure holds for the greater-than and equals relationships.

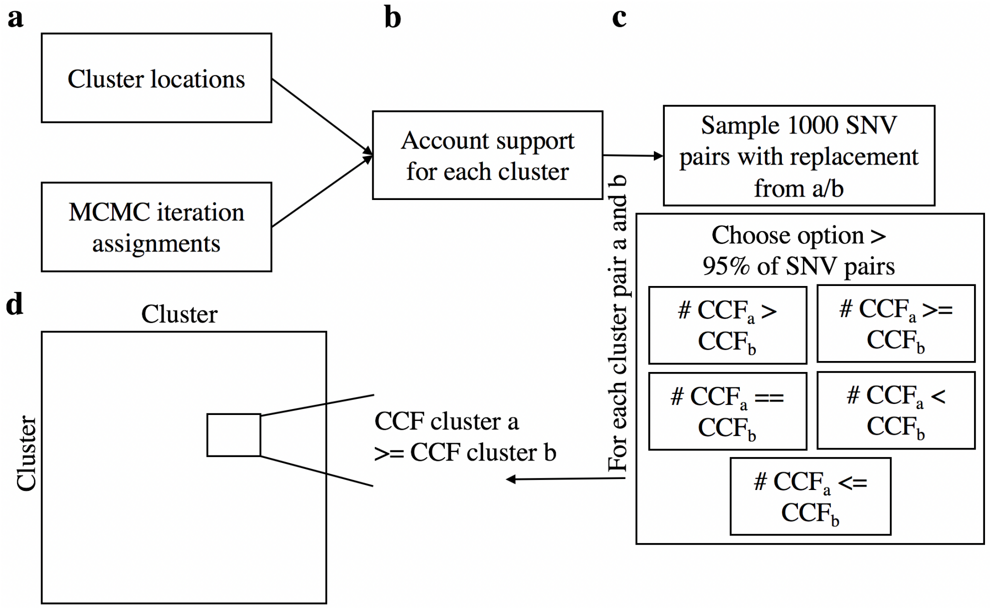

a
b
c
Cluster locations
MCMC iteration assignments
Account support for each cluster
Sample 1000 SNV pairs with replacement from a/b
For each cluster pair a and b
Choose option > 95% of SNV pairs
# CCF_a > CCF_b
# CCF_a >= CCF_b
# CCF_a == CCF_b
# CCF_a < CCF_b
# CCF_a <= CCF_b
d
Cluster
Cluster
CCF cluster a >= CCF cluster b

Fig. 2.11 Before trees are constructed all pairs of clusters are classified into pre-defined relationships. (**a**) The procedure starts with the cluster locations and the mutation assignments during MCMC. (**b**) For each mutation it is recorded how often it would have been assigned to each cluster during the MCMC iterations if that the final cluster locations had been available, yielding a probability per cluster per mutation. (**c**) Then for each pair of clusters 1000 mutation pairs are sampled with replacement and it is counted how often the pair are assigned to the same cluster or to a different cluster, providing support for five different scenarios. (**d**) Finally, the scenario that yields support from greater than 95% of sampled SNV pairs is chosen as the final classification. If no scenario yields a 95% support the pair of clusters is classified as *unknown*.

Having obtained a probability that clusters (a,b) have a greater-than, lesser-than or equal CCF we can classify the pair into a category with a threshold at 0.95 (Fig. 2.11d). If a pair does not pass the threshold for any category, or for multiple categories, it is assigned the label unknown.

2.4.2 Tree building

The tree building process begins with creating a full inventory of all possible edges by obtaining all possible parents for each mutation cluster. The trees are then built in two phases: In the first phase all clusters that fit into a single location are placed on the tree, starting with the cluster that has the highest CCF. The pigeonhole principle is not enforced in this phase, so the phase is followed by a screening that yields an error if a the combined CCF of daughter nodes exceeds the CCF of the parent.

Then in the second phase, all clusters that fit in multiple places are considered. For each cluster, we iterate over all the possible edges involving that cluster from the inventory and over all trees obtained so far. Clusters are added to the tree in a greedy fashion on first-come first-serve basis. The pigeonhole principle is strictly enforced during this process. Some clusters may therefore not fit on the tree, which results in warnings which point to clusters that cannot coexist and warrant further investigation. If a cluster can fit in multiple places, then new trees are recorded for each configuration. This process yields a list of possible trees after all iterations are complete.

Single sample cases do not yield any warnings, because it is always possible to construct a linear tree. Multi-sample cases are more complicated however. In such cases there are two possible options to be considered: (1) The data is not clean enough and an artefact cluster is prohibiting the tree building and (2) the number of whole genome duplications is not correctly accounted for and clonal mutations have become subclonal. The output of the tree builder is useful for automated checking for violations and it will point to the clusters that are problematic.

Chapter 3

Validation of methods

3.1 Introduction

In the previous chapter I have introduced the Battenberg algorithm for calling (subclonal) copy number from whole genome sequencing data, an approach to estimate the CCF values for SNVs and indels, and the DPClust method to infer the subclonal architecture from the distribution of CCF values measured in a single cancer. In this chapter, I will focus on validating the performance of these methods *in silico*. To this end, I have developed a subclonal architecture simulator called SimClone and metrics are introduced to evaluate the performance of a subclonal architecture caller. These metrics have a theoretical lower bound of performance, but a realistic upper bound does not exist (apart from the worst possible score). To set a realistic upper bound for a subclonal architecture caller, I introduce a series of simple, naive methods (termed RandomClone) that produce random subclonal architectures. An edited version of the text and figures describing SimClone, it's simulated data set and RandomClone will appear in the supplement of the PCAWG consensus subclonal architecture calling paper (Yu et al. 2017, manuscript in preparation). The simulated data set will be further used in Chapter 6 to compare the performance of subclonal architecture callers. Figure 3.6 is created by Maxime Tarabichi and is used with permission.

3.2 Simulating subclonality with SimClone

3.2.1 Introduction

SimClone was developed to evaluate the performance of DPClust. It can be used to generate subclonal architectures with underlying data that can test specific scenarios, or to build a set of random samples that can be used to evaluate overall performance. For it to be a true evaluation it is important for the simulator to generate problems that are as realistic as possible. I have therefore aimed to build SimClone such that it can take high level characteristics of real data as input.

A typical workflow goes as follows: (1) A subclonal architecture is generated in the form of a phylogenetic tree with subclones and their locations. Relationships between nodes (mutation clusters) are created, where each node has a parent and sits at a particular level in the tree (the level of a node is determined as the minimum number of steps required to "walk" from the root of the tree to the node). (2) Each node on the tree is assigned a number of mutations. (3) Then a genome wide copy number profile is simulated, after which (4) mutations are simulated separately for each node, which requires the user to also specify a tumour purity value that is used for all mutation clusters. The user can provide input for steps 1, 2 and 3 to have full control over the solution to be simulated.

3.2.2 Assumptions

Both the mutation and wild type alleles are supported by a number of reads. I assume that the distribution of the number of mutation-supporting-reads takes on the shape of a binomial distribution. To model variation on the total number of reads covering the locus where the mutation has occurred, I assume that the depth can be modelled as a Poisson distribution. The mutation is carried by a number of chromosome copies (multiplicity). The shape of this distribution is partly determined by the copy number profile, that bounds the possible multiplicity states, and by cancer type specific development, i.e. if gains occur late there will be many SNVs on multiple copies, while if gains occur early there will be few. I model multiplicity through a Poisson, and learn the parameter that determines the shape of the distribution from real data. Finally, as a simplification, subclonal mutations cannot be carried by more or less than 1 chromosome copy.

3.2.3 Simulating a tree

A tree consists of nodes and edges, and each node has a parent that is either another node or the root. The tree simulation step generates a tree independently of other sample characteristics such as coverage, cluster sizes, etc. The procedure is provided with a number of nodes to place on the tree and the number of tries allowed to place each node.

The procedure starts by placing a root node, that represents mutations that are clonal. Then, iteratively, new nodes are placed until the required number is reached. Before a new node can be inserted SimClone first selects the new nodes parent, that resides in the tree at a level and in a branch. The level is selected by a draw from a uniform distribution that covers all levels below the root in the current tree and a node at that level is selected to determine the parent with uniform probability.

For the new node to fit on the tree it must be assigned a CCF value such that the total CCF at the level of insertion does not exceed the CCF of the parent. The possible CCF space is therefore constrained. A further constraint can be placed (this is a user setting) in requiring that the new node position must be at least a minimum CCF away from its parent because clusters that are too close cannot be separated during clustering.

A CCF is then selected for the node between the max possible value and 0 as the CCF of the new node and the node is placed if it doesn't violate any of the constraints. The addition is retried with a new sampling of parameters if no suitable location is found, but the insertion is aborted if no location can be found after a specified maximum number of tries, resulting in one fewer node on the tree. This particular scenario is more likely when large numbers of clusters are requested as placed nodes will constrain the allowed space for new nodes. The user can then opt to either work with fewer nodes, rerun the procedure for an alternative tree or manually add extra nodes afterwards. Alternatively, a function is provided where a custom tree can be created.

3.2.4 Determining cluster sizes

The cluster size determination procedure takes the minimum and maximum total number of mutations in the tumour as input and an optional proportion of those mutations that are clonal. The total number of mutations is drawn from a uniform distribution between the minimum and maximum. I then determine cluster sizes by applying a stick breaking procedure where iteratively a randomly sized chunk is broken of the remaining stick. Each chunk then represents the proportion of total mutations that belong to a cluster. If a minimum proportion of clonal mutations is specified, then the first chunk will be constrained to be at least that specified size.

3.2.5 Simulating mutations

With node locations and sizes determined or provided as input SimClone now simulates the mutations per node. Further input is required in a copy number profile (the copy number simulation procedure is explained in the next section), a tumour purity value, coverage and a multiplicity λ parameter (also explained in the next section). These parameters are sample specific and clusters are therefore simulated independently per sample. Mutations are generated by calculating the expected number of reads reporting the mutation and wild-type alleles. But the multiplicity must first be determined before those can be calculated.

The multiplicity (m_m) is drawn from a Poisson distribution with the provided λ parameter as input.

$$m_m \sim \text{Pois}(\lambda) \tag{3.1}$$

The mutations are randomly assigned to a copy number segment in the provided profile. If the multiplicity is not possible given the major and minor allele of the selected segment I adjust it to the copy number of the major allele.

Then the number of reads per chromosome copy for the tumour (c_t) and normal (c_n) cells are calculated from the total coverage (C), tumour purity (ρ) and tumour ploidy (ψ_t). The total copy number of the normal cells is assumed to be 2:

$$c_t = C\frac{\rho}{\rho\psi_t + 2(1-\rho)} \tag{3.2}$$

$$c_n = C\frac{1-\rho}{\rho\psi_t + 2(1-\rho)} \tag{3.3}$$

Then expected number of mutant alleles r_m is determined by the multiplicity of the mutation, the mutations fraction of tumour cells (f) and the number of reads per tumour chromosome copy (c_t):

$$\mathbf{E}(r_m) = m_m f c_t \tag{3.4}$$

The expected number of wild type alleles r_w consists of three components: (1) Reads from normal cells (can be zero when the sample is pure and does not contain normal cells), (2) reads from whole chromosome copies from tumour cells that are not carrying the mutation (can also be zero when the copy number is 1+0) and (3) if the mutation is subclonal, an additional number of reads from cells that are not part of the subclone that carries the mutation:

$$\mathbf{E}(r_w) = 2c_n + m_w c_t + m_m(1-f)c_t \tag{3.5}$$

The total number of observed reads are then drawn from a Poisson distribution:

$$r_d \sim \mathrm{Pois}(\mathbf{E}(r_m) + \mathbf{E}(r_w)) \tag{3.6}$$

And the final mutant and wild type alleles are determined by a draw from a binomial distribution:

$$r_m \sim \mathrm{Bin}(r_d, \frac{\mathbf{E}(r_m)}{\mathbf{E}(r_m) + \mathbf{E}(r_w)}) \tag{3.7}$$

3.2.6 Extension to simulating multi-sample cases

The above procedure is already extended to simulate multi-sample cases. During my Ph.D. I have mostly worked with single-sample cases and this thesis contains results on that type of data only. I have therefore opted not to include any multi-sample simulations and validations.

The tree building step can be provided with an additional parameter that specifies the number of samples a multi-sample case should contain. It then simulates mutation clusters with (potentially different) CCF values in all the samples. However, it does require one single node that is clonal in all, but that node does not need to contain any mutations (i.e. to simulate multi-focal tumours). The procedure to simulate mutations for each cluster can take a multi-sample tree as input and it then simulates the mutations belonging to that cluster with CCF values for all the requested samples.

It is currently not possible to simulate multi-sample copy number profiles. One could use the same copy number profile or run the copy number simulator a number of times on the same input data. The method can be adapted in the future to simulate copy number profiles for multi-sample cases where the samples share a number of common alterations.

3.2.7 Simulating copy number

A copy number profile consists of segments and a certain number of copies are available for every segment. SimClone simulates copy number in three steps: (1) it selects a segmentation from a catalogue, (2) then it models the total copy number profile and (3) it breaks down the total copy number into allele specific contributions. Fig. 3.1 shows an example of a real copy number profile (top) and a simulation inspired by it (bottom). The described approach is simulating clonal copy number only.

A segmentation is selected either randomly from a catalogue of real segmentations, or can be chosen as containing only whole chromosome or whole chromosome arm segments. The total copy number is then modelled through a Poisson distribution, where the λ parameter is learned from a real copy number profile. Before learning the λ parameter I first subtract 1 from the total copy number and after drawing from the learned distribution I add one to the simulated total copy number because the Poisson distribution often draws many zeroes. This means SimClone does not simulate homozygous deletions.

$$n_{tot} \sim \text{Pois}(\lambda) \tag{3.8}$$

Total copy number is drawn from the learned distribution and assigned to randomly selected copy number segments, until the fraction of the genome covered by total copy number distribution looks similar to that of the real tumour (Fig. 3.2 shows the total copy number distributions of the real and simulated profiles shown in Fig. 3.1). Often there is a minor discrepancy between the distribution from real and simulated data inspired by the real sample due to the random assignment of total copy number to segments. The real distribution can often only be obtained by recreating the real profile exactly, which is what SimClone aims to avoid.

The total copy number is then broken down into separate contributions from two alleles to obtain allele specific copy number by using the multiplicity distribution of the SNVs from the real sample. Multiplicity values are drawn from a Poisson distribution with its λ parameter

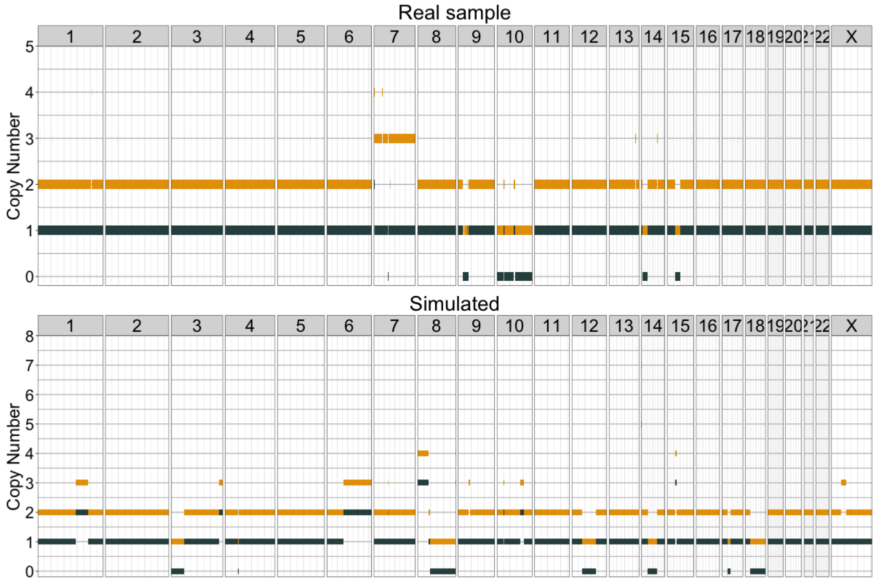


Fig. 3.1 Genome wide overview of a real copy number profile (top) and a simulation that is inspired by the real profile (bottom).

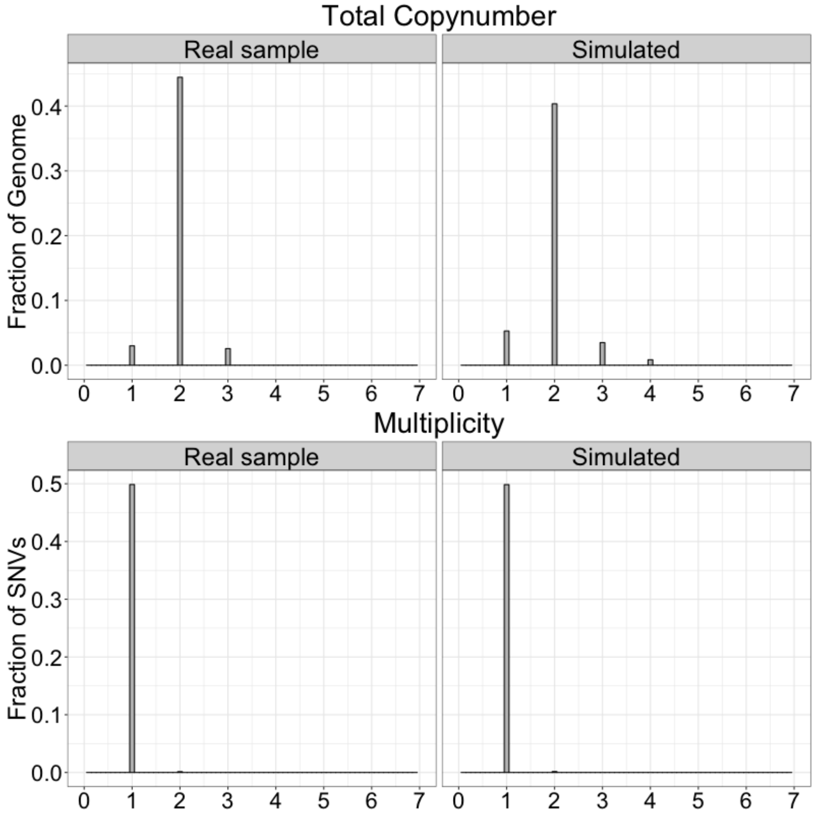


Fig. 3.2 Comparison of real distributions (left) with distributions of simulated data. The figures on the top row show that the copy number states distribution of the simulated data follows that of the real data, but there is a noticeable discrepancy. The algorithm aims to approximate the total copy number states distribution in the real tumour as closely as possible by iterating over the observed copy number states and assigning the state to a randomly selected segment until similar proportions of the genome are covered. But due to the variability in segment lengths it is not always possible to exactly match the real distribution. The multiplicity distribution (bottom) closely resembles that of the real sample.

learned from the real data (Fig. 3.2). A multiplicity value is drawn for each SNV and SNVs are assigned to segments (this assignment is purely for establishing copy number) where the total copy number is greater than or equal to the multiplicity of the SNV. The maximum multiplicity, m_i, assigned to segment i is then used to determine the one allele, $n_{A,i}$.

$$n_{A,i} = \max(m_i) \tag{3.9}$$

The other allele, $n_{B,i}$, is then determined by subtracting the copy number of $n_{A,i}$ from the total copy number $n_{tot,i}$:

$$n_{B,i} = n_{tot,i} - n_{A,i} \tag{3.10}$$

Finally, the major allele for each segment is established as the allele with the highest copy number state, the minor allele is the allele with the lowest copy number state.

This procedure ensures that if many SNVs have a high multiplicity state, then many segments will be created with an allele specific copy number configuration that can support them. If, for example, the multiplicity and total copy number distributions are very similar, then the profile will have many major alleles that closely follow the multiplicity, leading to a profile with much loss of heterozygosity (Fig. 3.3). By randomly distributing copy number states across the genome it becomes possible to create very difficult and truly chaotic profiles to test a methods' limits (Fig. 3.4)

The procedure however doesn't restrict particular copy number states to particular chromosomal areas. That means the actual copy number profile will most likely not resemble the profile used as inspiration. But it should be covered by the same allele specific copy number state combinations in similar proportions, if the same segmentation is used.

A different choice in segmentation can cause a bigger discrepancy between the real and simulated distributions of total copy number and multiplicity. Fig. 3.5 shows three simulations inspired by the same real profile, but with segmentations that are restricted to whole chromosome arms (middle) or whole chromosomes (bottom). The real tumour consists for large parts of normal copy number with a few large and small scale alterations. When restricting segments to whole chromosomes, many segments are too large for a reasonable approximation of the proportion of the genome covered by total copy number. SimClone therefore tends to pick small segments, that reside on the small chromosomes.

In the future it could be interesting to experiment with an additional catalogue of cancer type specific common events to create more biologically accurate copy number profiles. However, currently the aim is to simulate the effect of copy number on the VAF of SNVs, for which it does not matter where on the genome the alterations are placed.

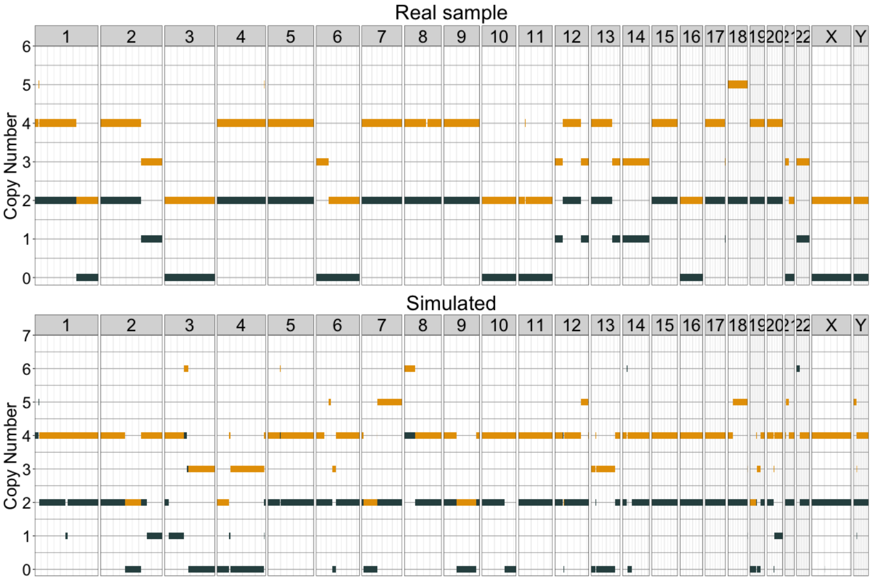


Fig. 3.3 A copy number profile simulation that is inspired by a real tumour with loss of heterozygosity and a whole genome duplication.

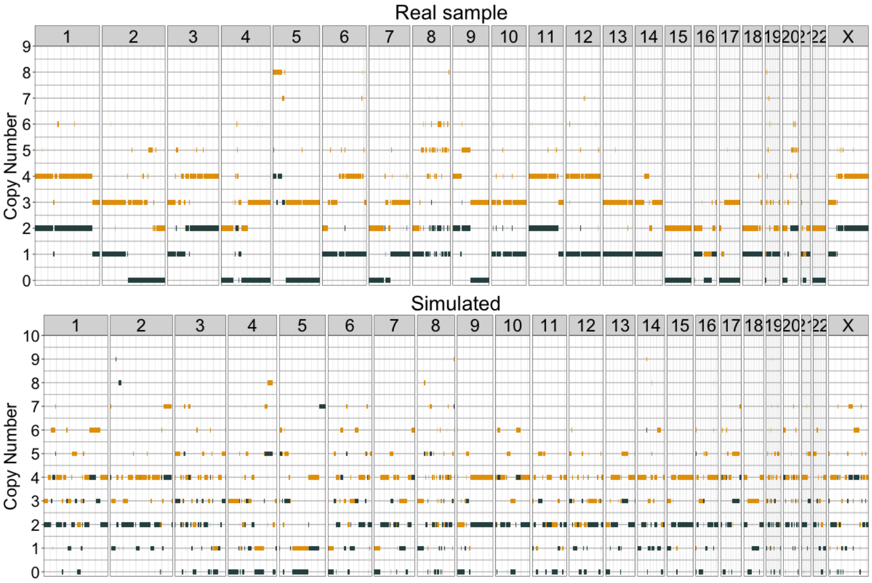
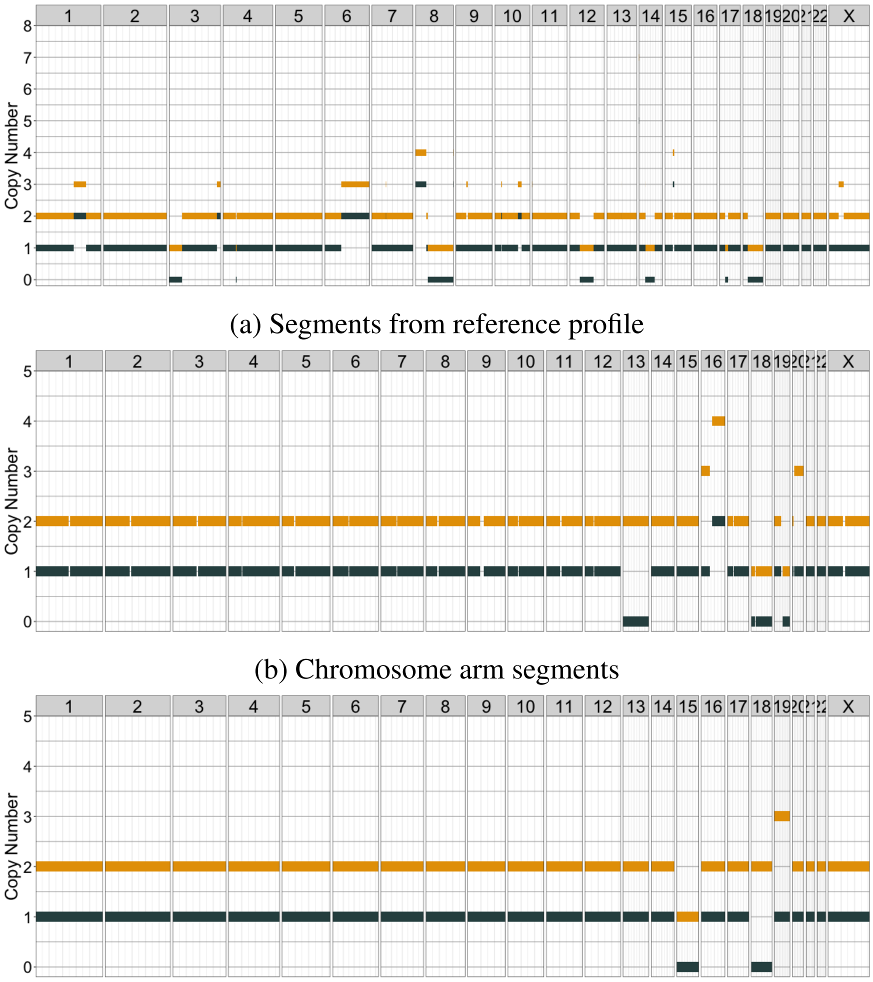


Fig. 3.4 Example of a simulation based on a very fragmented and messy real copy number profile. The random assignment of copy number to segments creates a chaotic simulated profile.



(a) Segments from reference profile

(b) Chromosome arm segments

(c) Whole chromosome segments

Fig. 3.5 Simulations using the same real sample as inspiration, but with different segmentations: (a) The segmentation of the real sample (b) segmentation where each segment is a full chromosome arm and (c) segmentation where each segment is a whole chromosome. The simulator aims to approximate the proportion of the genome covered by certain copy number states. The real tumour (shown in figure 3.1) contains a few large scale and a few small scale alterations, but consists mostly of normal copy number. Due to the large segment sizes in (b) and (c) it therefore tends to place alterations in the smaller segments that reside on smaller chromosomes.

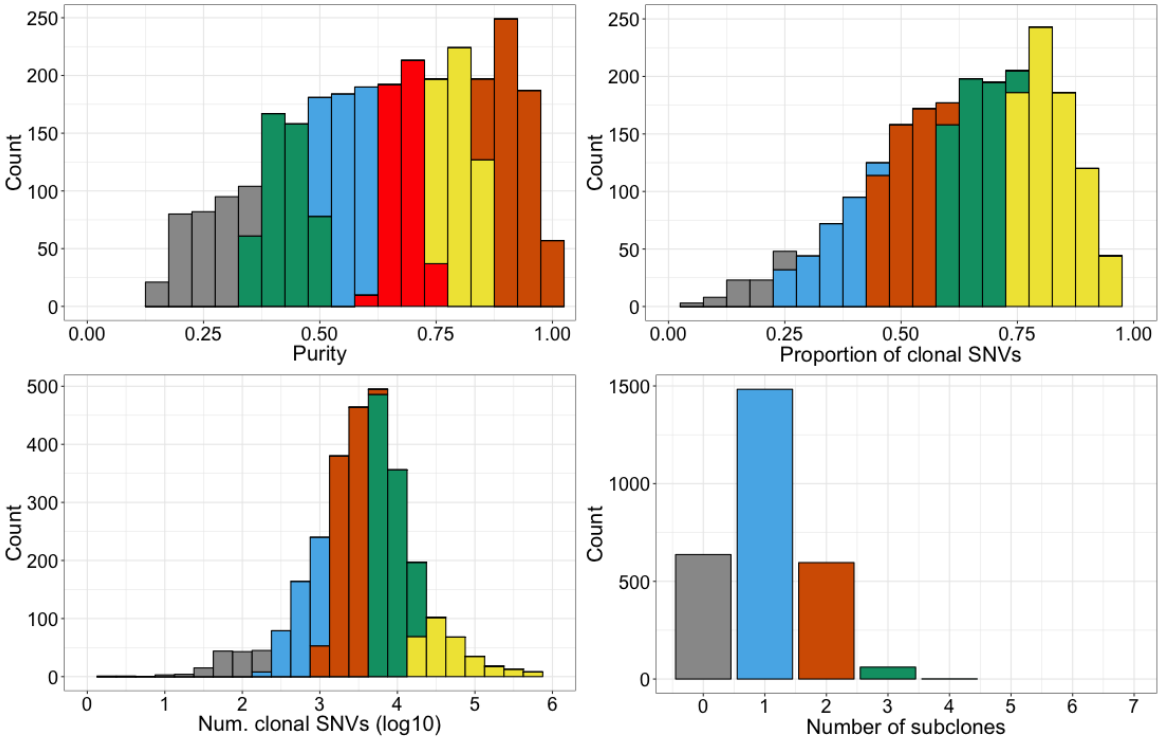


Fig. 3.6 The simulated data set was created as a grid with four axis. Each axis represents a type of measurement that can be obtained from real data. This figure shows the histogram of these four measurements from the PCAWG data and the colours represent bins along each grid axis. A simulated tumour falls somewhere on the grid, which amounts to a combination of 4 bins (one on each axis). The parameters for this sample are then generated by sampling a single value from each of the 4 bins.

3.3 SimClone1000, a validation data set for PCAWG

SimClone was used to simulate 1000 tumours with the aim to evaluate the performance of subclonal architecure callers within PCAWG. The data set consists of 700 unique subclonal architecture simulations and 300 cases where the exact subclonal architecture was simulated a second time on a copy number profile without any alterations. The 300 paired cases allow us to investigate whether subclonal architecture callers perform better without having to adjust for copy number alterations. We created a grid with four key parameters by which tumours vary when considering their subclonal architectures: Purity, the fraction of clonal SNVs, the number of clonal SNVs and the number of subclones. Of the 1000 tumours 36 yielded too few mutations (less than 20) or did not complete the simulation process due to time constraints. The data set therefore contains 964 tumours.

The axis of the grid were determined based on the distribution of each of these values in the PCAWG data set, shown in Fig. 3.6. By applying k-means clustering we obtained 6 purity clusters, 5 clusters for the number of clonal mutations (with one cluster fixed at 10^5 to represent hypermutators) and 5 clusters for the fraction of clonal mutations (one cluster was fixed at 0.995 to represent a typical hypermutator). The grid axis for the number of subclones was determined by creating 4 classes corresponding to 0, 1, 2 and 3+ subclones, where the 3+ category contains tumours with 3 to 7 subclones. A single tumour is then assigned a purity drawn from a bin on the purity axis, a number of clonal mutations from a bin on the clonal mutations axis, etc. This results in combinations of real parameters, but they may not have been observed as a combination. The 6-by-5-by-5-by-4 grid yields 600 unique combinations of parameters, which we extended by sampling another 100 combinations to reach 700.

Copy number profiles were chosen at random from the PCAWG data set. Once a real purity is selected for a particular simulation we also assigned it the copy number profile of the real tumour. We simulated 300 tumours a second time without copy number alterations and therefore only allowed tumours with at least 10% of their genome altered to be included in the grid to not inflate the number of quiet diploid tumours. From the 700 simulations we selected 300 at random for another simulation with normal diploid copy number. A change in ploidy affects the number of reads per chromosome copy, when purity and coverage remain equal, potentially altering the CCF space of the simulation substantially and making the envisaged comparison difficult.

In the regular simulation we, for example, do not have sufficient power to simulate SNVs at a CCF below 0.3. This means that a the distributions of mutations belonging to a subclone at 0.4 CCF will be truncated as some of its mutations cannot be represented by a number of supporting reads greater than 0. When the number of reads per chromosome copy is increased we gain power to simulate subclonal mutations, resulting in a lower CCF cutoff point. That means we can simulate more mutations of the 0.4 CCF subclone, making it potentially easier to correctly identify it by subclonal architecture callers, and rendering a comparison between the diploid and non-diploid simutations uninformative.

We therefore opted to adjust the purity of the non-diploid copy number profile (ρ_n) to correct for the reads per chromosome copy shift when changing the ploidy from the real sample (ψ_n) to create a purity for the diploid simulation ρ_d:

$$\rho_d = \rho_n \frac{2}{\psi_n} \tag{3.11}$$

Subclone positions and sizes were determined as described in the previous section about SimClone. And coverage was fixed to the PCAWG average of 48.46621.

3.4 Metrics to evaluate a subclonal reconstruction

To aid the large scale performance evaluation of tumours on simulated data we developed three metrics around two key descriptions of a subclonal architecture: Clusters (location and size) and assignments of mutations to clusters. The metrics compare a provided subclonal architecture against a known truth. It's also possible to use these metrics to measure how similar a pair of solutions are, which is used later in this thesis to compare performance of subclonal architecture callers.

The overall subclonal architecture can be roughly described by the number of subclones (π) and the proportion of clonal mutations (θ). Eqs. 3.12 and 3.13 capture the absolute difference between solutions k and l. For both metrics, the lower the score, the better.

$$\frac{|\pi_k - \pi_l|}{\pi_k + \pi_l} \tag{3.12}$$

$$\frac{2|\theta_k - \theta_l|}{\theta_k + \theta_l} \tag{3.13}$$

Comparing the cluster locations is more complicated, because the solutions to be compared may not contain the same number of clusters. Instead, we use the mutation assignments. Each mutation is hard-assigned to a cluster, and each cluster has a location. For each mutation i we compare the CP of the assigned cluster between solutions k and l. A small distance across all mutations reflects a good concordance between the solutions. Eq. 3.14 calculates the average difference in CP ($\varphi_{i,k}$ is the cellular prevalence assigned to mutation i by method k), where a lower value is better. The score is divided by the tumour purity (ρ) to correct for purity differences between tumours.

$$\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\varphi_{i,k} - \varphi_{i,l})^2}}{\rho} \tag{3.14}$$

3.5 A lower bound generated by RandomClone

3.5.1 Introduction

The metrics described above can be used to assess the performance of a subclonal architecture caller, where low scores mean a method is performing well. Often methods will not get the perfect solution and therefore their scores will deviate from the perfect score. What is not clear however is when performance can be described as poor. I reasoned that a caller should be able to outperform a simple random method. Running a random method on the same data as the caller would then provide an upper bound of what can be considered reasonable performance. To this end I have developed three simple methods that generate random subclonal reconstructions and one method that returns only clonal tumours.

3.5.2 RC – Stick breaking

The stick breaking method starts with drawing a random number between 0 and 6 to determine the number of clusters. It then orders the mutations by their CCF and iteratively breaks a randomly sized chunk of the ordered mutations. Each of these chunks represents a mutation cluster and its location is obtained by taking the mean CCF of the mutations in the chunk. Mutations are automatically assigned by belonging to a particular chunk. The advantage of this approach is that it is more likely to place a cluster where there are large real clusters, but it does also tend to place multiple clusters within large real clusters.

3.5.3 RC – Informed

The downside of the stick breaking approach described above is that it performs a single series of breaks and returns that as a solution. I wondered whether the method could be improved by selecting the best solution from a series of random models. The informed method runs the stick breaking implementation described above 100 times. Contrary to the above approach, the informed method records the size and locations of clusters, but does not record the assignments. It runs the MutationTimer approach (used in Gerstung et al. (2017) to assign mutations), which models each mutation cluster as a beta-binomial and takes into account the size of the cluster. MutationTimer then calculates the proportion of mutations that are poorly explained (i.e. fall in the outermost 5% of the beta-binomial distributions). This proportion is calculated for all 100 runs, after which the run that yields the lowest value is selected as the returned subclonal architecture.

3.5.4 RC – Uniform

The uniform approach is an alternative to stick breaking. It starts with drawing a random number between 0 and 6 to determine the number of clusters to be found. Then that number of draws are made from a uniform distribution that has as minimum the lowest 5% of the mutation CCF space and as maximum the highest 5%. That means cluster locations are drawn from within the CCF space that is occupied by mutations. Mutations are then assigned to clusters by calculating the binomial likelihood per mutation and cluster, and assign to the most likely cluster. This approach does not utilise the shape of the CCF space and therefore usually places clusters in unexpected locations. As it is nearly always outperformed by the stick breaking approach I omit results from this method.

3.5.5 RC – Single cluster

This approach places a single cluster to explain the data. It can obtain the single cluster by taking the mean mutation CCF, or it can be forced to place a clonal cluster at a CCF of 1. All mutations are assigned to the single cluster.

3.6 Validation of multiplicity calls

DPClust takes as input a fixed multiplicity value for every mutation. That value is obtained during pre-processing using the equations from the previous chapter. Incorrect multiplicity values could artificially alter the CCF space, which DPClust may explain through extra mutation clusters. I therefore used the over 42 million mutations from the 964 simulations to validate the performance of this pre-processing step. Table 3.1 contains the proportion of mutations with a correct multiplicity for four different splits of the data and shows that over 99% of mutations are assigned the correct multiplicity.

The subset of mutations that are gained (i.e. have a multiplicity greater than one) has a lower success rate at just over 93%. For most of these mutations there is not much ambiguity about their multiplicity, but some will fall between two multiplicity states. The addition of binomial noise to the reads supporting the mutation and reference alleles can cause the mutation to shift away from the correct multiplicity. The DPClust pre-processing considers the evidence provided and assigns the most likely multiplicity, which, due to the noise, can be incorrect.

3.7 Assesment of a subclonal architecture through resimulations

In addition to the metrics described earlier I have developed a measurement that aims to capture how well a subclonal architecture explains the raw CCF space it is provided as input. The idea is that if the subclonal architecture called by a method is the true architecture, then their corresponding CCF spaces should be very similar. A distance metric can then be used to calculate the difference, where a small deviation would serve as a good score because the called subclonal architecture describes the observed input data very well.

Evaluation of multiplicity calls across 964 simulations

Type	Total	1st Qu.	Median	Mean	3rd Qu.	S.D.
All	42,536,567	0.9915	0.9992	0.9917	1.0000	0.0186
Gained	2,125,495	0.9195	0.9808	0.9336	0.9987	0.1071
CNA	11,869,768	0.9800	0.9946	0.9840	1.0000	0.0277
CNA & Gained	1,756,994	0.9121	0.9762	0.9311	0.9973	0.1014

Table 3.1 Multiplicity values are compared between the truth and the DPClust calls. Proportions of mutations correct (1st quartile, median, mean and 3rd quartile) are shown for four different splits of the data: All mutations, mutations on more than one chromosome copy (Gained), mutations in regions of aberrant copy number (CNA) and mutations that are in a region of aberrant copy number and are gained (CNA & Gained). The table shows that over 99% of mutations are assigned a correct multiplicity value. Most mistakes are made with gained mutations. This pertains to mutations that fall exactly between two multiplicity values and the binomial noise pushes the mutation away from the correct multiplicity. Even in that scenario over 93% of mutations are assigned the correct value.

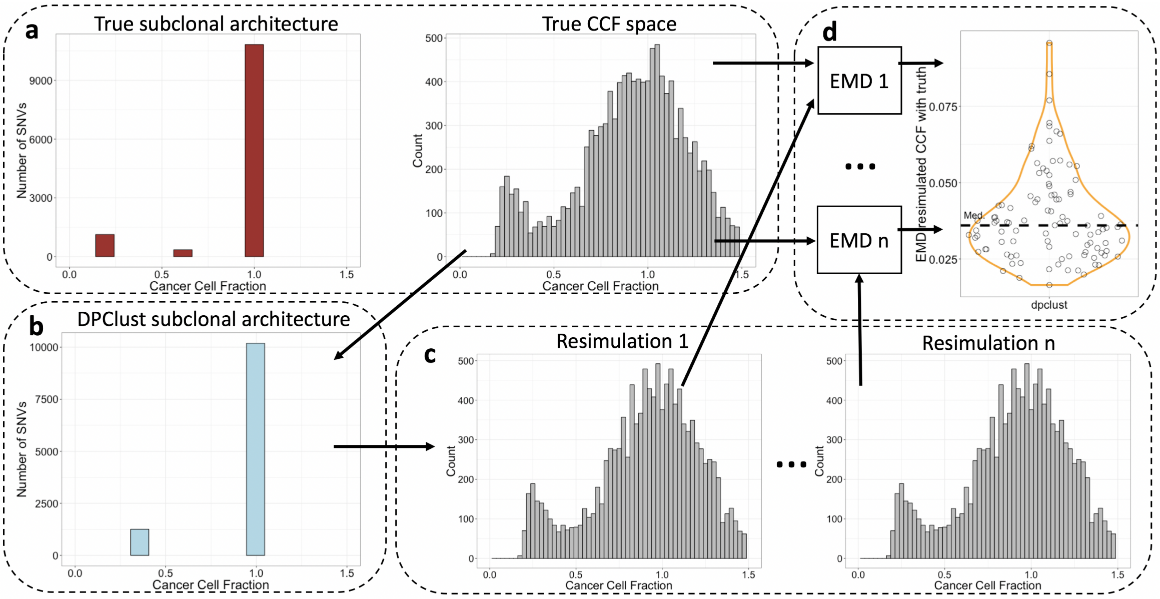


Fig. 3.7 (**a**) SimClone generates a true subclonal architecture and its associated CCF space. (**b**) The mutations that make up the true CCF space are provided as input to DPClust, which returns a subclonal reconstruction. (**c**) That reconstruction serves as input to SimClone for resimulations, which returns a number of CCF spaces. (**d**) The EMD is calculated between each resimulation and the true CCF space that measures how well the DPClust subclonal reconstruction is explaining the true CCF space. The EMD distribution is summarised by the median, resulting in a single similarity value per sample. This similarity value can be compared across methods and across samples.

The approach is illustrated in Fig. 3.7. For a single tumour SimClone returns (among other things) a subclonal architecture with cluster locations and sizes, and the simulated mutations form a CCF space. DPClust takes the CCF space as input and returns a subclonal architecture. That obtained architecture is then fed back into SimClone a 100 times (this process is referred to as *resimulation*), yielding 100 CCF spaces. The earth movers distance (EMD) is then calculcated between each resimulated CCF space and the true CCF space. A summary of the resulting histogram can then be used as a metric of performance.

When provided with a subclonal architecture SimClone simulates mutations with binomial noise on the number of supporting reads and Poisson noise on the coverage. It is therefore expected that a pair of resimulations also differ, hence 100 resimulations are performed for every called subclonal architecture by DPClust, the RandomClone methods and for the truth. The distribution of EMDs obtained from resimulating the truth can then be compared with the EMDs from the DPClust and RandomClone solutions.

Figure 3.8 shows the EMD distributions for a selected simulated tumour for the truth, DPClust and the three RandomClone methods, while table 3.2 contains the raw results and scores. It shows that the stick variant of RandomClone comes closest to the expected number of subclones and DPClust does best on the fraction of clonal mutations. The RMSE scores are very close, suggesting the CCFs of the clusters to which mutations are assigned in general are close to the expected value. The median EMD (dashed horizontal line) tells us that DPClust explains the CCF space best, followed by the informed RandomClone variant.

To summarise the EMDs (for a whole data set comparison in the next section), relative to the variation obtained from resimulating the truth, I then calculate the following score:

$$1 - \frac{1}{n}\sum_{i=1}^{n} I(e_t, e_m) \tag{3.15}$$

This score is obtained by sampling n pairs of values with replacement from the truth and from one of the methods' EMD distributions and determining whether the EMD of the method (e_m) is greater than the EMD from the truth (e_t). The index function in eq. 3.15 returns a 1 when e_m is greater than e_t. n is set to 1000.

If the method has returned a subclonal architecture that explains the CCF space very well, then a score of 0.5 is returned. A value greater than 0.5 means the method has not perfectly described the true CCF space, with a higher value meaning a bigger discrepancy. Finally, a lower value means the method is better at explaining the true CCF space than resimulations of the truth do (i.e. the method is overfitting on the input data).

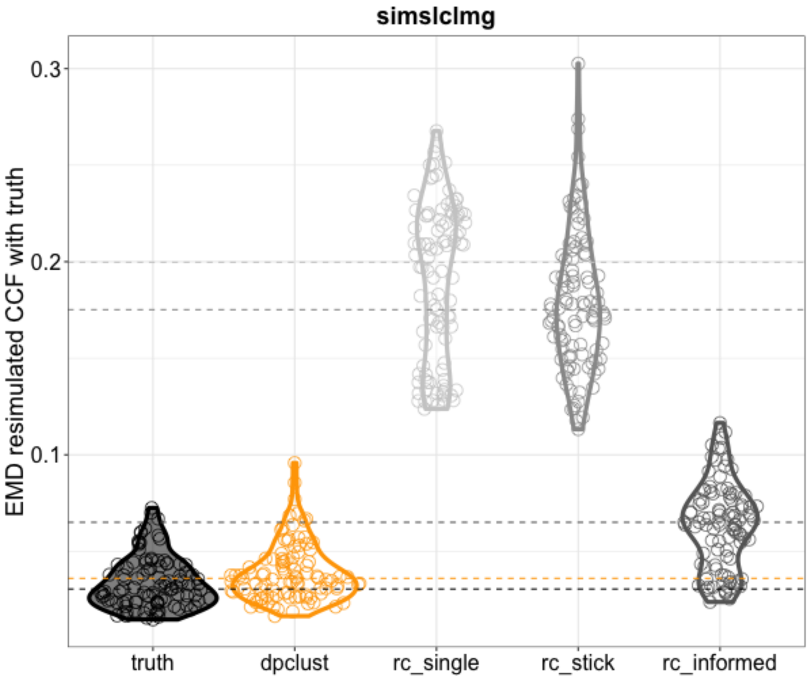


Fig. 3.8 Earth movers distance (EMD) of the CCF space between the truth and resimulations of solutions found by DPClust and the randomclone methods. Resimulating the truth provides a lower bound of what is obtainable, while the random methods provide an upper bound. The similar median (dashed lines) EMD that of the truth suggests DPClust has found a solution that explains the true CCF space quite well.

Evaluation of performance on sample simslclmg

Method	Calls		Scores			
	Num. subcl.	Frac. clonal	Num. subcl.	Frac. clonal	RMSE	EMD
Truth	4	0.196				0.030
DPClust	2	0.235	0.491	0.197	0.007	0.036
RC single	0	0.000	0.999	1.000	0.011	0.199
RC stick	3	0.000	0.221	0.931	0.006	0.175
RC informed	1	0.312	0.822	0.592	0.008	0.065

Table 3.2 A comparison of the scores on simulated tumour simslclmg reveals that the scores capture different characteristics of the reported solutions. RandomClone-stick, for example, comes closest to the true number of subclones and therefore receives the best score in that category. However, it assigns very few mutations to the clonal cluster and therefore attains a poor fraction of clonal mutations score.

3.8 Validation of DPClust

Figure 3.9 shows the outcome for DPClust and the RandomClone methods for the three PCAWG scores and the resimulation metric on the SimClone1000 data set. For the three

PCAWG metrics a lower score means better, with zero being perfect. DPClust comfortably outperforms the three random methods on number of subclones, fraction of clonal mutations and mutation assignments. The resimulation score is expected to be 0.5 when DPClust finds a perfect solution. A value higher than 0.5 represents a drop in performance, while values below 0.5 can be considered overfitting. DPClust also outperforms the random methods on the resimulation score. DPClust is evaluated on the three PCAWG scores against 10 other subclonal architecture callers in Chapter 6.

The scores indicate that DPClust performs well. However, it does not always find the exact true subclonal architecture. Figure 3.10 is an attempt to explore where the differences lie. In nearly half the tumours DPClust calls the correct number of subclones and in those cases the proportion of clonal mutations is close to the truth, indicating cluster locations have been called in roughly the correct locations. For the other cases DPClust nearly always undercalls the number of subclones. Where undercalling occurs, DPClust often calls a larger proportion of mutations clonal. This suggests it merges a nearby subclone into the clone.

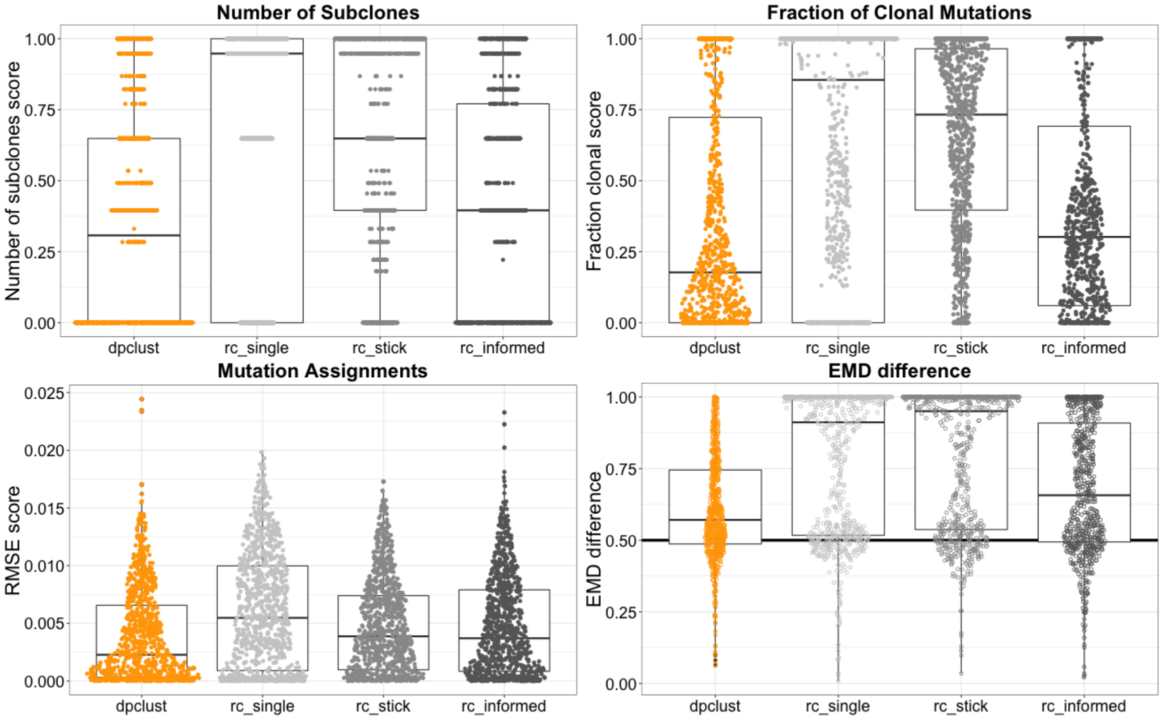


Fig. 3.9 General overview of the four scores obtained on the SimClone1000 comparing DPClust against the the truth and performance by a random method. The best score is 0 in the first three metrics and 0.5 in the fourth. All four scores show that DPClust easily outperforms the random callers and does well on calling the fraction of clonal mutations (top right), the mutation assignments (bottom right) and explaining the true CCF space (bottom left). The discrepancy in the number of called subclones is further explored in figures 3.10 and 3.11.

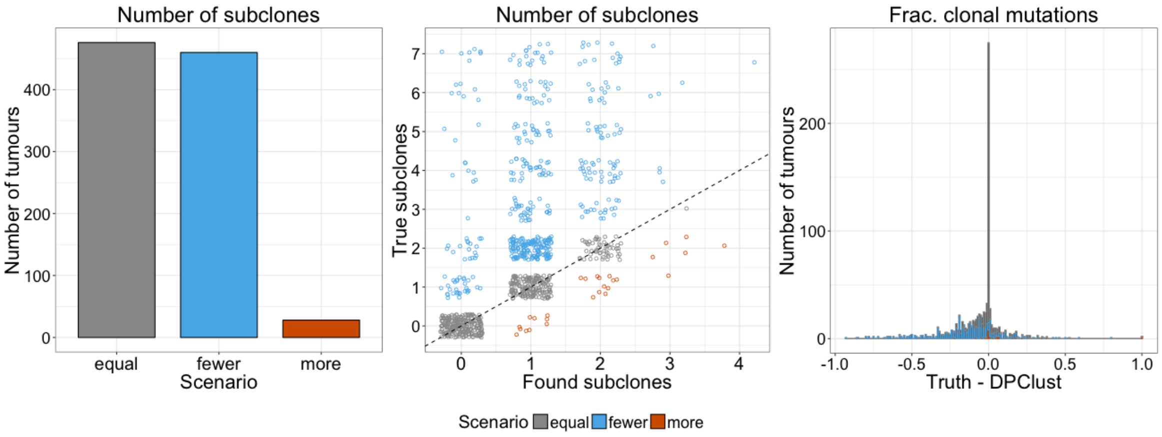


Fig. 3.10 The trend across the 964 simulated tumours is that DPClust calls the correct number of subclones in nearly half of the cases, in the other half it nearly always undercalls by one or more subclones (left). Undercalling occurs in two major scenarios: Cases where in reality there are two subclones, but DPClust calls a single subclone and cases where the true number of subclones is 3 or more (middle). This affects estimates of the fraction of clonal mutations, where in cases where DPClust undercalls it often overestimates the fraction of clonal mutations (right). Combined these results suggest DPClust can be considered conservative in its statements about the amount of subclonality found in a data set. An explanation where the number of subclones discrepancy comes from is explored in Fig. 3.11.

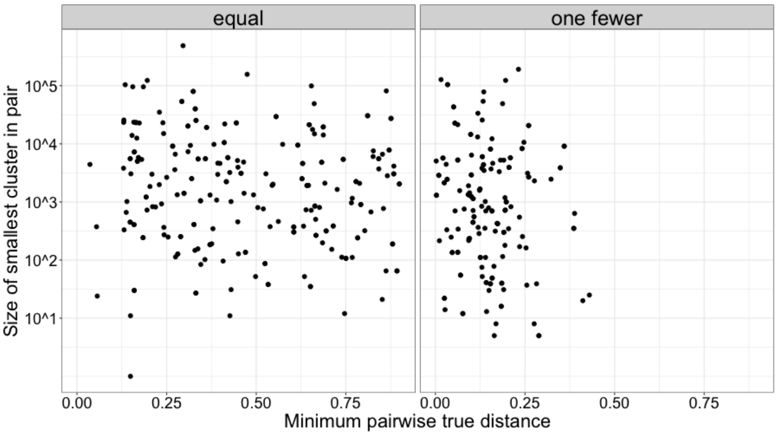


Fig. 3.11 The results in Fig. 3.10 suggest that subclones are merged in about half of the cases within the SimClone1000 data set. This figure compares the distance between the closest pair of clusters (x-axis) and the size of the smallest subclone within that pair (y-axis) for cases where DPClust finds the correct number of subclones (left, clonal tumours omitted) and where it finds a single subclone where two are expected. The data show that merging of clusters occurs when a pair of clusters are within 0.25 CCF of each other, regardless of their size.

Figure 3.11 shows this phenomenon, with on the x-axis the CCF difference of the two closest mutation clusters and on the y-axis the number of mutations that belong to the smallest cluster of the selected pair of clusters. When comparing these data between cases where DPClust finds the correct number of subclones (clonal tumours are excluded) and cases where it finds a one subclone where two are expected, it is clearly visible that in the latter category merging occurs frequently when the distance between a pair of clusters goes below 0.25 CCF. The size of the cluster (and the size difference, data not shown) plays little to no role in the separability of clusters.

DPClust assumes each of these clusters is its own statistical distribution influenced by binomial noise. In this merging scenario a pair of clusters are significantly overlapping and DPClust cannot find sufficient evidence of there being two clusters. It therefore takes the conservative approach and calls one fewer subclones, protecting against overstating the amount of heterogeneity.

These analysis explain that discrepancy between the truth and the DPClust solutions can be explained by clusters that are too close to separate. It shows that tumour heterogeneity estimates based on DPClust are reliable, yet a conservative underestimate of the true amount of heterogeneity.

3.9 Validation of assignments of gained mutations

A gained mutation (i.e. a mutation on more than 1 one chromosome copy) is assumed to be clonal by the described procedure to establish a mutation's multiplicity in the previous chapter. It is the maximum parsimony explanation when, for example, a mutation appears to be reliably carried by two chromosome copies, and the local copy number allows for this to occur (there are two copies of at least one of the two alleles). In such a scenario one would expect DPClust to assign this mutation to the clonal cluster, but DPClust is not constrained and can assign a gained mutation to a subclone. I therefore set out to investigate how often this occurs.

Across the SimClone1000 data set an average 10% of mutations are gained (Fig. 3.12, top and middle). Nearly all these gained mutations are correctly assigned, with an average of 94% of mutations assigned to the clone (Fig. 3.12, bottom). However, in some samples this percentage is much higher, in some cases nearly all gained mutations are incorrectly assigned. Furthermore, samples in which a very high proportion of gained mutations are assigned incorrectly contain a large number of mutations overall.

Investigation revealed that in many cases clearly clonal mutations (i.e. mutations with a CCF near or greater than 1) were assigned to a subclone (sample sim01bxzd is provided

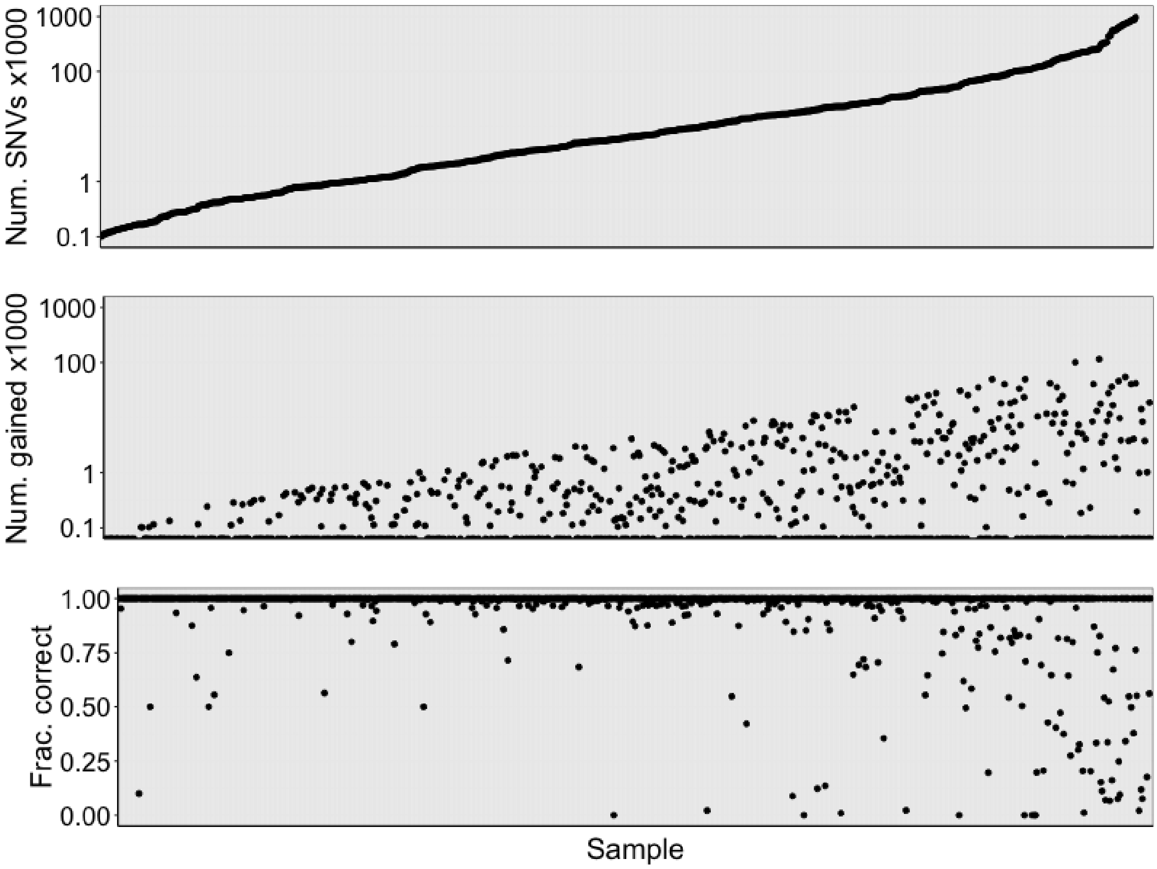


Fig. 3.12 The figure shows the total number of mutations (x1,000) at the top, the number of gained mutations (x1,000) in the middle and the fraction of those gained mutations that are correctly assigned to the clone at the bottom. A high proportion of incorrectly assigned mutations is concentrated in the samples with a high number of mutations, which lead to the exploration of a number of DPClust runs with different parameters.

as an example in Fig. 3.13), often forming a clear cluster around the location of the clone. Combined with the observation that samples with poor performance often contain many mutations, I postulated that these represent cases where the MCMC chain had not yet converged. At each MCMC iteration a mutation is assigned to the most likely cluster and after all iterations are complete these assignments are amalgamated into the most likely call. If the chain has not yet converged it is still in a state of flux, where mutation assignments can be volatile, leading to an increased likelihood of assignment to an incorrect cluster.

I therefore performed additional runs on a number of samples that show a poor assignment performance. I increased the total number of iterations and also altered the number of iterations that are discarded as burn-in, leading to the following combinations: 1,250 iterations with 250 burn-in (default), 2,000 iterations with 1,000 burn-in, and to runs of 5,000 with 4000 burn-in and 10,000 with 9,000 burn-in. The proportion of incorrectly assigned mutations for sample sim01bxzd (Fig. 3.13) directly decreases from 16% to 0.4% when increasing to 2,000 iterations and 1,000 discarded as burn-in. This behaviour is consistent across nearly all selected samples (Fig. 3.14), highlighting that the number of iterations the MCMC chain is run for by default is too short and should be increased to at least 2,000 and 1,000 iterations should be used for burn-in.

More iterations did not have the desired effect on all samples. Figure 3.15 shows one such cases, sample sim6zrlr0 as an example. It contains a vast subclone, that is truncated on one side, while it engulfs the clonal cluster on the other side (top row). In such an extreme case it becomes difficult to separate a pair of clusters, which causes a good number of clonal mutations to be assigned to the subclone (middle and bottom). in a scenario where clusters are within each others space it will be difficult to reliably assign mutations. A further adjustment should be made that prohibits gained mutations to be assigned a subclone.

3.10 Validation of Battenberg

The SimClone1000 data set cannot be used to assess the performance of Battenberg, because SimClone simulates just the final copy number profile and not the underlying data. For this validation I therefore introduce a separate data set that contains manually created subclonal architectures and copy number profiles that have subsequently been embedded into a BAM file using BAMsurgeon (Ewing et al., 2015). These tumours have been created for the Somatic Mutation Calling - heterogeneity (SMC-het) project that aimed to evaluate performance of subclonal reconstruction algorithms, where Battenberg copy number profiles were provided to all participants. The organisers of SMC-het used the high coverage NA12878 (Zook et al., 2016) and the previously sequenced parents of NA12878 (NA12891 and NA12892) as part of the 1000 Genomes project (1000 Genomes Consortium, 2012) to obtain the maternal and paternal genome that make up the genome of NA12878, which allows for creation of haplotype correct copy number profiles (details will be available in Salcedo et al. (2017, manuscript in preparation)).

In total 50 tumours were designed by the SMC-het team, inspired by real tumours reported in literature and from PCAWG. The copy number profiles were limited to whole chromosome alterations, as this is a requirement for BAMsurgeon, and subsequently 50 BAMs were generated. For the purposes of this validation I have excluded 8 cases that were designed as subclonal architecture corner cases as these all have the exact same copy number profile.

Figure 3.16 shows a comparison of the expected and measured raw data (BAF and logR, top panels), total estimated copy number and the cancer cell fraction estimates of subclonal

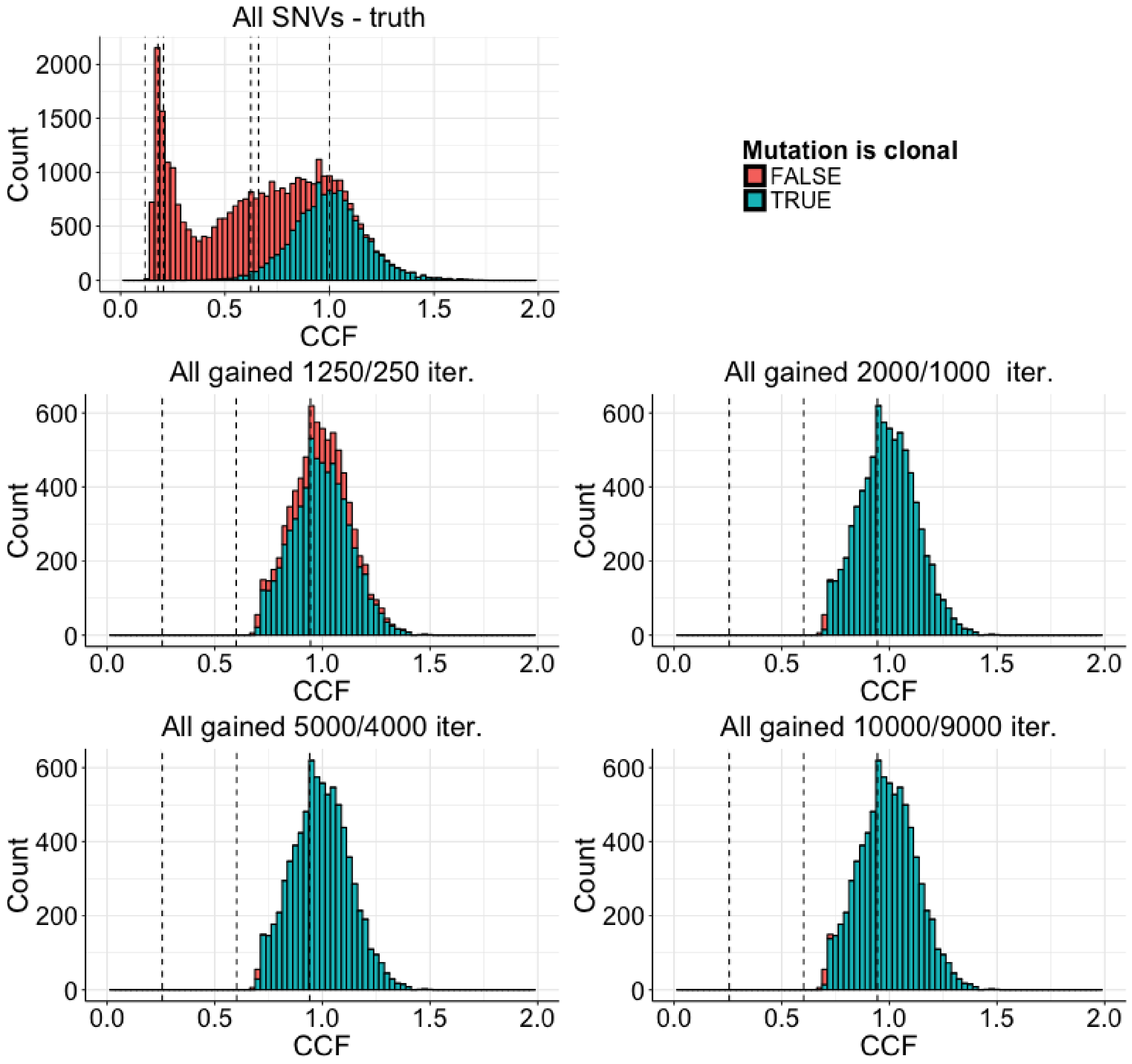


Fig. 3.13 Overview of the truth (top) and the output of multiple DPClust runs (middle and bottom rows) with varying numbers of MCMC iterations and burn-in. The top row shows the full true CCF space where bars are coloured to represent clonal and subclonal mutations, the black dashed lines represent the true cluster locations. The middle and bottom rows show all gained mutations that are clonal and offers a breakdown of these mutations into whether they are correctly assigned to the clone in green or incorrectly assigned to the subclone in red, while the dashed lines represent the found cluster locations. The figure shows that increasing the parameters to 2,000 iterations and 1,000 burn-in yields a reduction in the number of incorrectly assigned mutations.

copy number segments, where each dot represents a segment. Both BAF and logR show very high R^2 values, with only a few segments just of the diagonal. This means that the

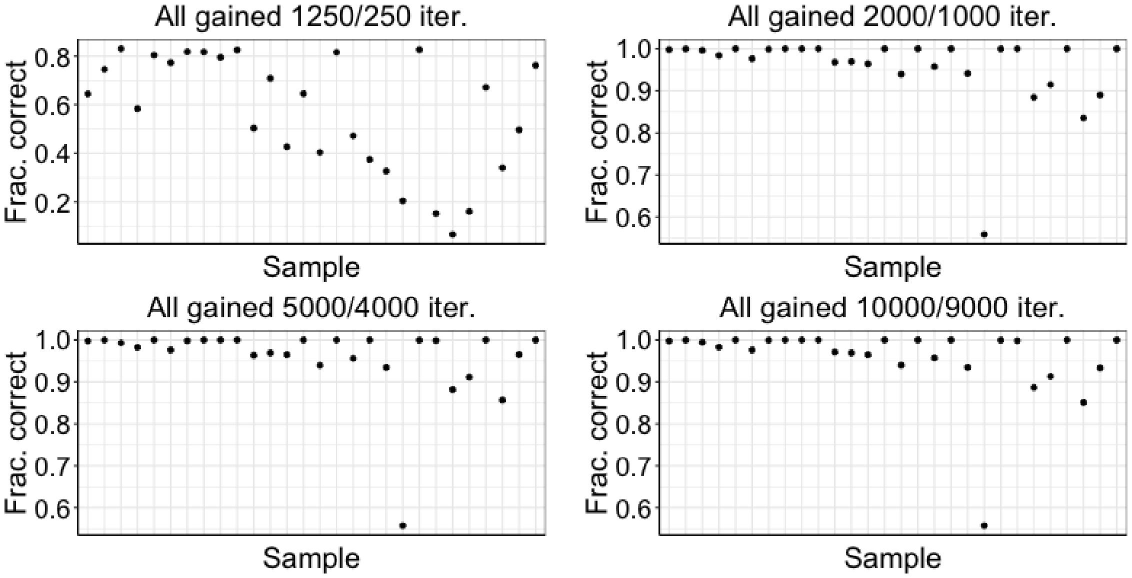


Fig. 3.14 The fraction of incorrectly assigned gained mutations across a number of selected samples showing poor performance. Three additional DPClust runs were performed beyond the original (1,250 iterations and 250 burn-in): 2,000 iterations and 1,000 burn-in, 5,000 iterations and 4,000 burn-in and 10,000 iterations with 9,000 burn-in. The results show that increasing the number of iterations yields considerable improvement, therefore the number of iterations should be increased.

Battenberg phasing and logR creation and correction steps are performing well and are able to adequately recreate the data that goes into copy number calling.

The total copy number estimates for these segments also show a very high correlation with the expected values (bottom left in Fig. 3.16), albeit slightly lower than the correlations obtained on the BAF and logR. One major source of discrepancy is caused by Battenberg calling clonal copy number, where subclonal copy number was expected. This occurs because Battenberg performs a t-test on the BAF and calls subclonal copy number only when the observed BAF is significantly different from the expected BAF of a clonal copy number state. A comparison of the CCFs (bottom right in Fig. 3.16) of the subclonal segments reveals that this affects a low proportion of all subclonal segments and most of these have low expected CCF values, which means the BAF is very similar to that of the closest clonal state.

A comparison of the CCF values reveals a strong correspondence between the observed and expected output (bottom right in Fig. 3.16). There is however a larger discrepancy than observed on the other three measures. In part this is caused by the aforementioned segments that are fit with clonal copy number. The data appears on a slightly discrepant diagonal, where the observed CCF is consistently higher than the expected. This effect is most likely explained by deviations in the purity estimate. The correlation between observed

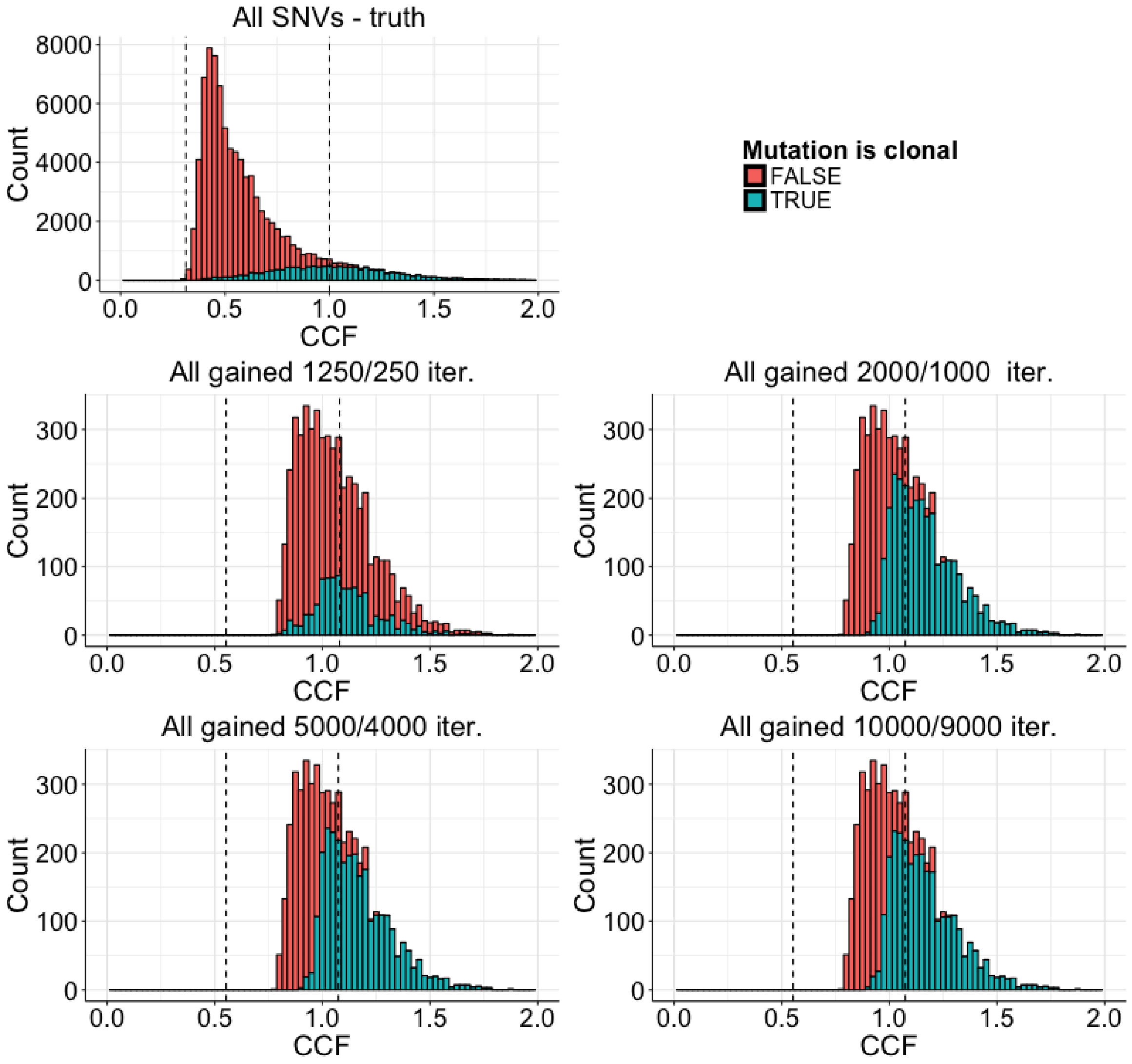


Fig. 3.15 Overview of a sample for which increasing the DPClust parameters had no effect on the number of incorrectly assigned mutations. This particular sample contains a very large and very broad subclone that contains the clone in its tail (top). In this scenario the mutations within the two clusters are split incorrectly, leading to a consistent number of gained mutations assigned to the subclone. A post-hoc step to reassign these mutations would resolve the issue, as there always is considerable uncertainty for mutations in between two subclones.

and expected CCF values is still strong however, but this result shows the CCF estimates of subclonal copy number are under the influence of some variation.

Finally, the bottom panels of Fig. 3.16 suggest that the Battenberg purity estimate may play a role in whether estimates of total copy number and CCF deviate from the diagonal. Fig. 3.17 contains the purity estimates for all samples and shows that Battenberg systematically

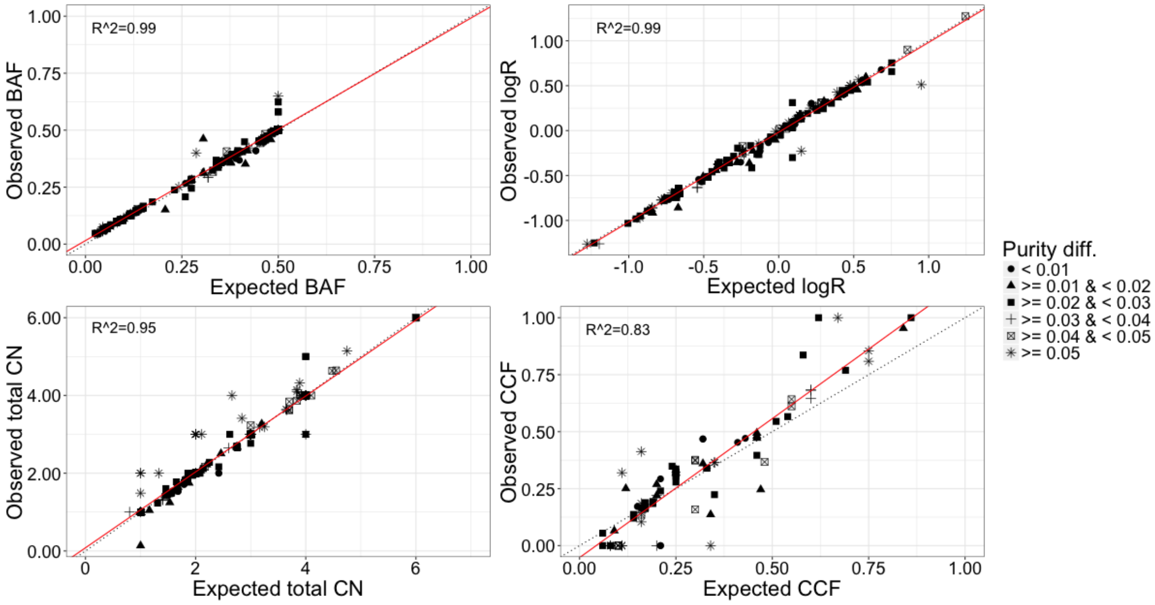


Fig. 3.16 The correlations between observed and expected BAF and logR per segment (each dot represents a segment) are very high, suggesting the Battenberg pipeline does well at obtaining the raw data required for copy number fitting (top panels). These correct raw values result in high correlations on the total copy number estimates per segment (bottom left) and strong correlations on the CCFs of subclonal copy number segments (bottom right). One source of discrepancy are cases where Battenberg calls clonal copy number, where subclonal was expected, due to the BAF not being significantly different from the closest clonal state. The slight bias in the bottom right panel is most likely explained by the deviations observed in purity estimates.

underestimates the purity (left panel). However, when the simulated BAF is replaced by the true BAF (while keeping the simulated logR), Battenberg calls the correct purity (right panel). This shows that the fitting algorithm works correctly and the small deviations of the BAF (the simulated BAF is on average 0.003 lower than it should be) are responsible for the purity discrepancy.

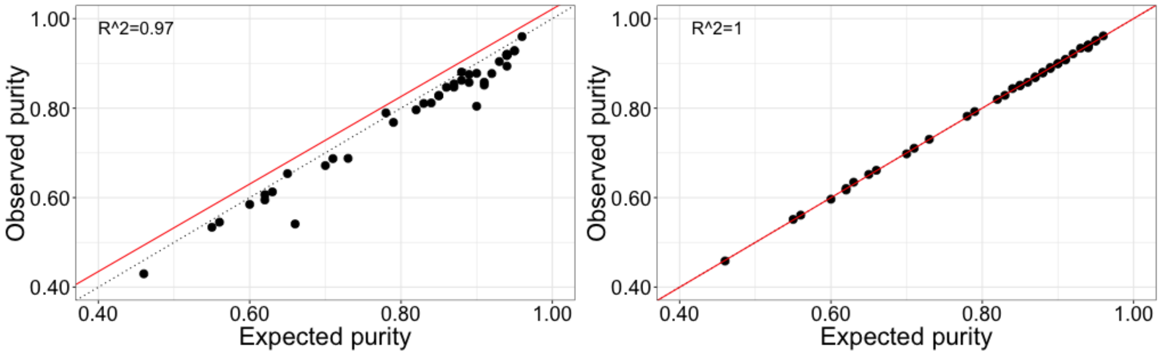


Fig. 3.17 A comparison of Battenberg purity calls (y-axis) from the simulated data with the true purity (x-axis) shows that Battenberg systematic underestimates (left). A run of the Battenberg fitting using the true BAF and simulated logR however shows that the fitting algorithm works as expected and yields the almost exact purity values (right). This means the very small deviations of the simulated BAF are responsible for the offset and shows how sensitive Battenberg is to correct BAF data.

Chapter 4

A rigorous quality control procedure

4.1 Introduction

During my projects I have developed a rigorous quality control (QC) procedure for copy number profiles. In the previous chapter there was always a ground-truth available for the simulated cases, but for real data that is not available. Furthermore, my experience of working with Battenberg and other somatic copy number callers across 1000s of cases has revealed that whole genome duplication calling and handling various types of noise on the input data are difficult problems. For the ICGC pan-cancer project we therefore developed a consensus copy number calling approach (described later in this thesis) that is robust against outlier calls. However, the effect can be mitigated for a single caller by a quality control review and refitting procedure.

The procedure consists of a series of QC failure criteria, which can be assessed using a series of figures and are described further below. Once a profile fails QC it must go through a refit procedure that either involves an automatic or a manual refit (see Section 4.3). The profile then either passes QC or it will fail again, resulting in another refit. Most profiles that require this procedure pass after one refit, but for some samples it is impossible to find a fit that does not violate any of the criteria highlighted below. In such a scenario the QC violation could be unexpected interesting biology and requires further investigation. Examples of such violations can be multi-focal tumours or cases with a pre-malignant lesion.

The QC procedure is based on the expectation that a cancer sample contains the clone (the most recent common ancestor) that contained SNVs and, depending on the cancer type, CNAs that are shared by all cancer cells and are therefore clonal. The expectation is that the sequencing data shows those clonal mutations by means of a clonal mutation cluster and a large proportion of clonal CNAs. Furthermore, clonal SNVs are carried by a number of chromosome copies, which distribution should roughly follow the proportion of the genome covered by different copy number states. Those three expectations link the SNV and CNA data together and they should fit as a trio as they are different views of the same cancer.

In an ideal world, every copy number profile is backed by independent validation. However, for the data described in this thesis there is very little validation data available. The samples described also provide the most heterogeneous data set that Battenberg has seen to date, with samples sequenced on different platforms and protocols, at different time points, to different specifications and as part of different projects. The diversity of these data required curation to obtain information about data quality and method performance.

A series of metrics have been developed that aim to capture the QC metrics described below. However, at the time of writing there is no substantial analysis on those metrics available. With the ICGC pan-cancer consensus copy number profiles available however, it should now in theory be possible to create a set of metrics that capture every QC failure.

4.2 Quality control metrics

The quality of a subclonal architecture is addressed by inspection of the copy number profile, the subclonal reconstruction and the estimated copy number states from the SNV data. In general, most samples pass after the first fit, but the success rate can be variable depending on the type of cancer (biology) or the sequencing project (data or biology).

I use Fig. 4.1 for initial assessment of the criteria highlighted below. In general one expects the copy number profile to be without any of the *fail* criteria, for the copy number estimate of SNVs to show peaks at integers (i.e. if there are 2 copies of a particular allele, then in general it is expected that some mutations are present on two copies) and the mutation clustering should normally show a peak at 1 for the clonal cluster.

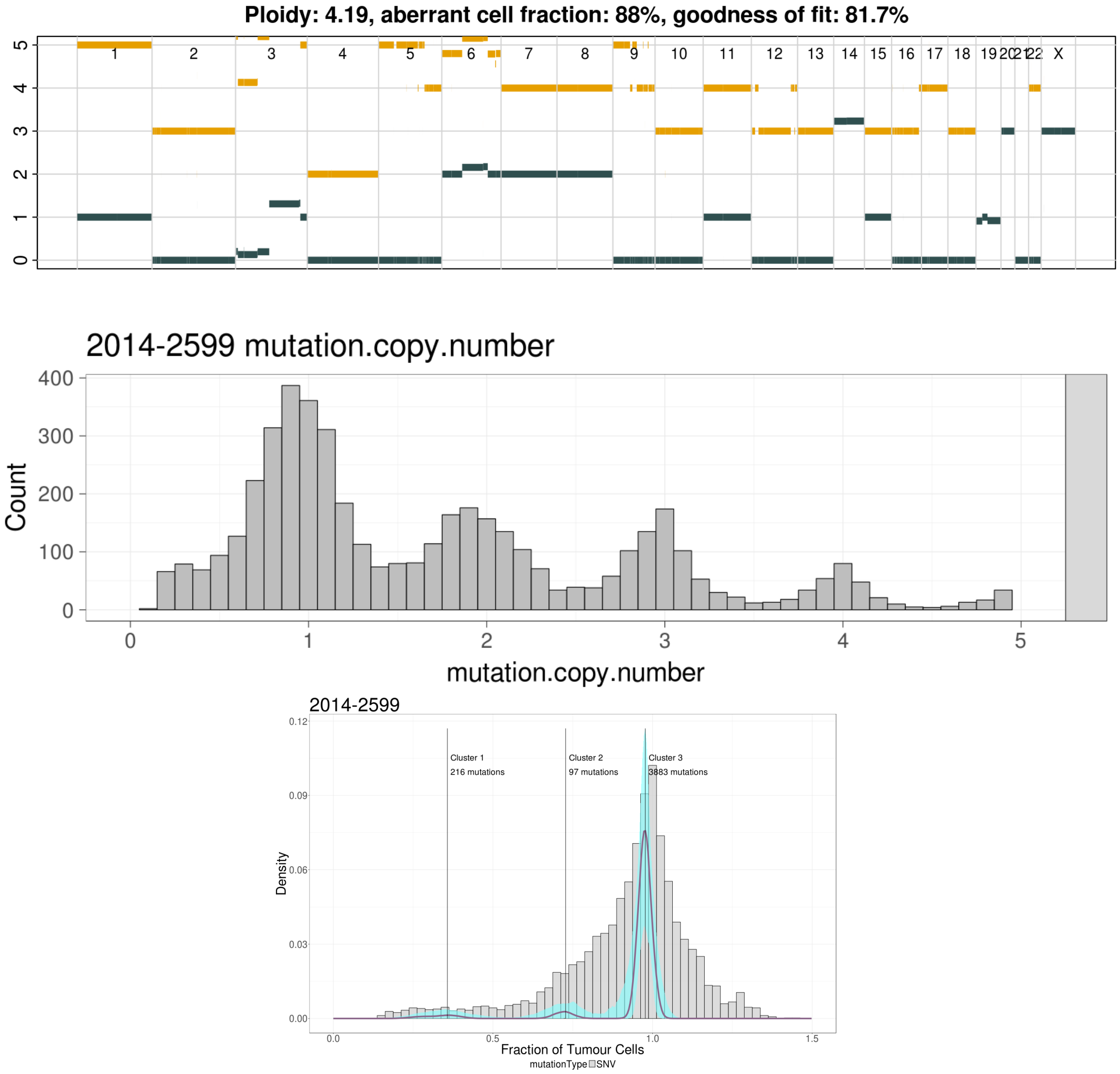


Fig. 4.1 The main quality control figure used to assess the copy number fit and subclonal reconstruction initially. The top figure contains the copy number profile with the major allele in orange and the minor allele in dark grey. The middle figure shows the copy number estimate of the SNV data (which is calculated through Eq. 2.16). The bottom figure shows the subclonal architecture for this tumour with the mutation data as a histogram in the background and the clustering result in the foreground as a purple line with a turquoise confidence interval. The vertical lines represent found cluster locations.

4.2.1 Large homozygous deletions

Large homozygous deletions are an instant QC fail. As previously discussed, it would be unexpected if a whole chromosome was lost completely. But it is not directly clear where the cutoff lies for a homozygous deletion to be believable. To be conservative I flag homozygous or subclonal homozygous deletions of 10Mb or greater, which means they are clearly visible in the copy number figure. In some cases the homozygous deletion should then be accepted as real, after closer inspection.

The case shown in Fig. 4.2 contains two subclonal homozygous deletions, of which one covers the majority of chromosome 18. This example also shows that multiple QC metrics can be triggered as it also contains a large number of chromosomes with subclonal states that would become clonal by doubling. This profile therefore also triggers a failure for the metric described in section 4.2.2. The mutation copy number and CCF space do not trigger any failures. With a purity of 17% one expects the clonal peak to appear somewhat shifted because with the relatively low coverage a sizeable number of clonal mutations fall below the detection limit. The shift is therefore the result of the *winner's curse*, which is addressed in Chapter 6. The solution for this sample is to add a whole genome duplication, which makes the homozygous deletions become a mixture of 1+0 and 2+0 (Fig. 4.3).

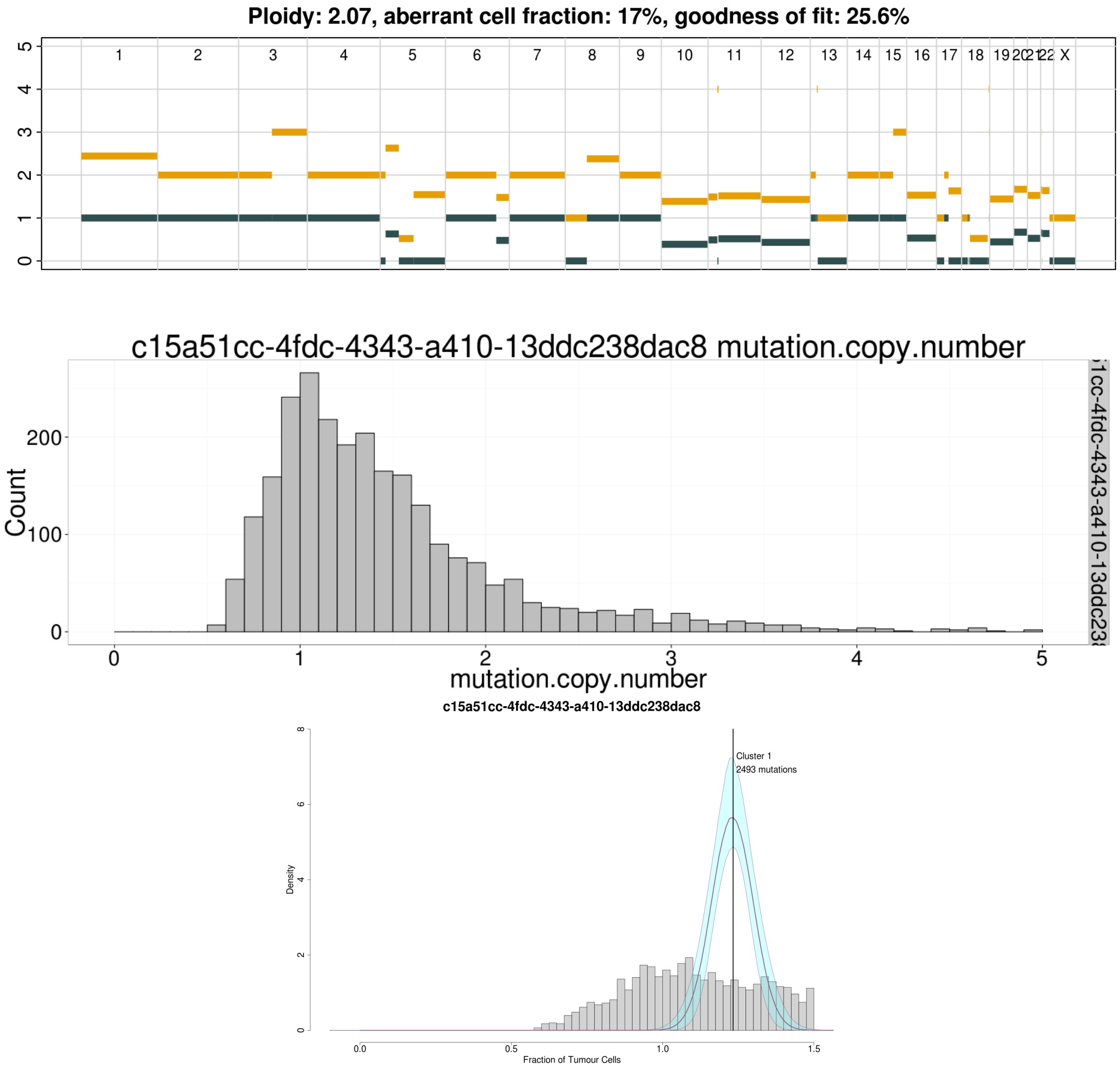


Fig. 4.2 QC failure case because of large (subclonal) homozygous deletions on chromosomes 5 and 18. This copy number profile also contains a number of subclonal copy number segments near 50% of tumour cells (detailed in Section 4.2.2).

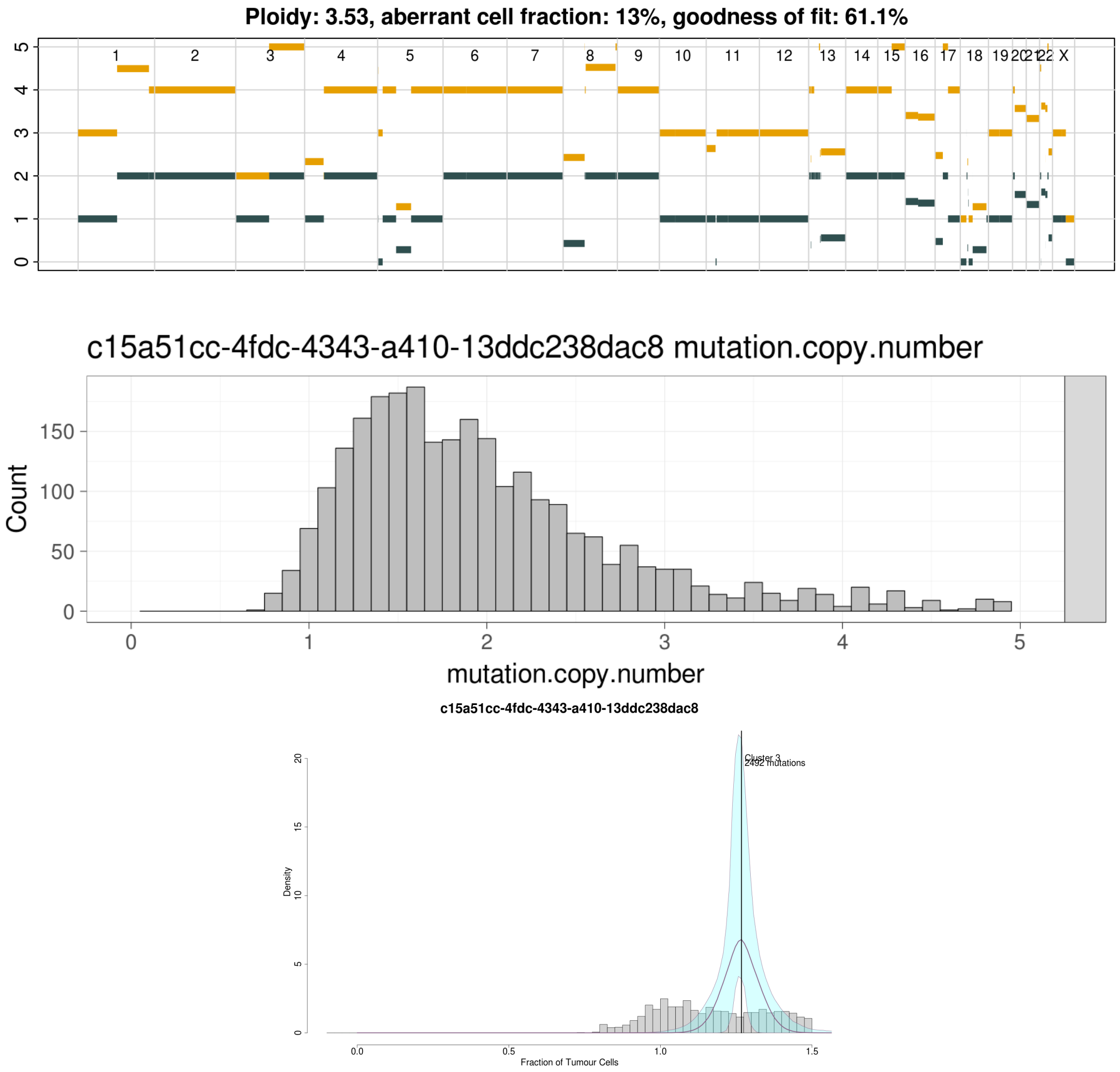


Fig. 4.3 The sublonal homozygous deletions on chromosomes 5 and 18 are resolved by adding a whole genome duplication (normal diploid 1+1 regions become 2+2). The subclonal segments near 50% tumour cells on chromosomes 1, 10, 12 and 19 become clonal. This tumour is of very low purity (13% of the sequencing sample contains tumour cells, according to this copy number fit), which in this case means a number of clonal SNVs fall below the detection limit. In the DPClust output figure (bottom) we therefore identify the clonal cluster shifted to a CCF higher than one (the shifting due to what is referred to as the *winner's curse*, which is briefly addressed in Chapter 6). This is a characteristic of the data and cannot be reverted by adjusting the copy number profile.

4.2.2 50% subclone in copy number

Beyond the subclonal homozygous deletions, Fig. 4.2 contains segments covering nearly the whole of a number of chromosomes that appear right in between two clonal states. The profile would initially fail and closer inspection of the detailed copy number figures that Battenberg produces reveals that the segments on chromosomes 11 and 16 have estimates close to 50% of tumour cells (Fig. 4.4). A whole chromosome arm or multiple large segments on different chromosomes is enough to trigger a fail. The next step is to refit the profile by doubling one of the clonal alterations (in this case for example the large segment on chromosome 13). A refit is accepted if the identified subclonal segments become clonal and no other failure criteria are triggered.

4.2.3 Empty odd numbered copy number state

A whole genome duplication can always be added to a copy number profile, and it produces an equally likely explanation of the data. But when a duplication too many is added it sometimes leaves a copy number state empty. In Fig. 4.5 there is no segment that takes on copy number state 1. Furthermore, in the mutation copy number figure there is no clear peak at 1 either. That suggests that either the whole genome duplication was the last event that became clonal, or that the duplication was added erroneously. Without further evidence it is not possible to distinguish between the former and latter, but the profile without a duplication provides a simpler explanation of the observed data and is considered a maximum parsimony explanation. The approach taken with samples reported in this thesis is that a whole genome duplication must be supported by clear evidence, preferably from the copy number and SNV data. The solution for this sample is therefore to halve the ploidy, as there is no information to support the duplication (Fig. 4.6).

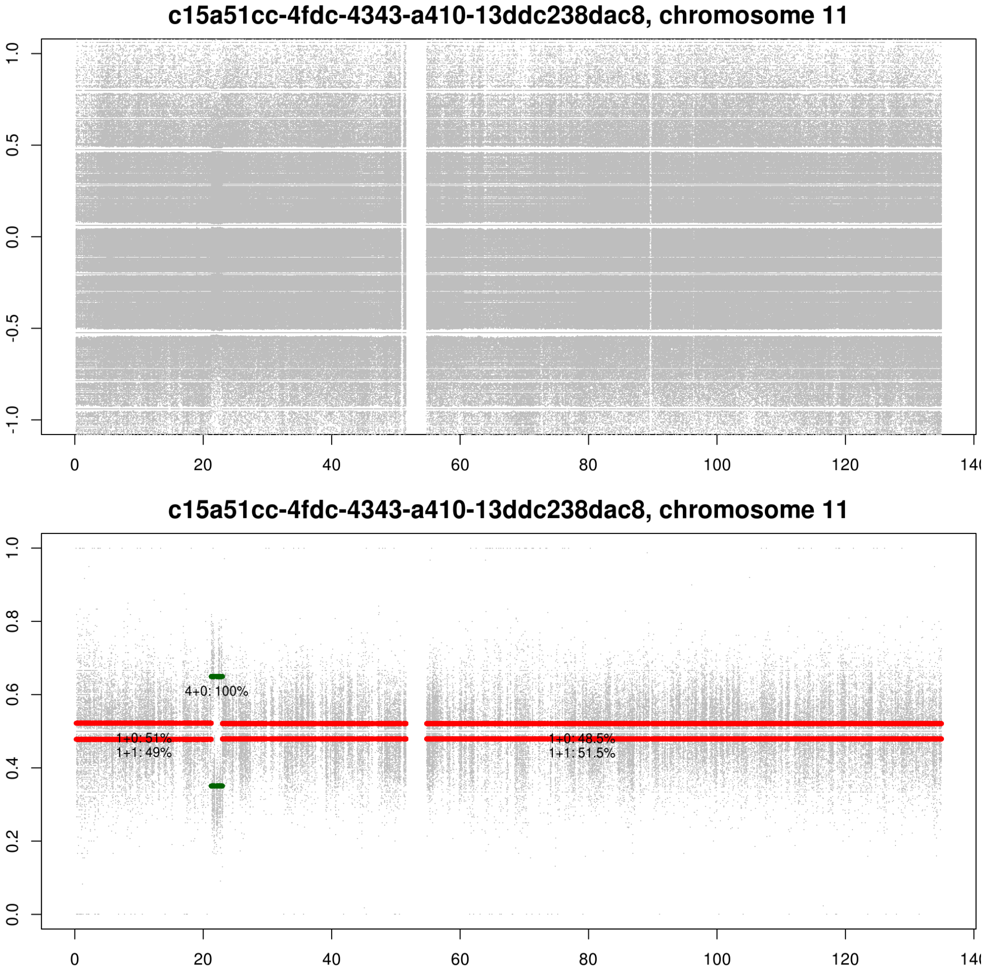


Fig. 4.4 Detailed copy number fit of chromosome 11. The top figure shows the relative copy number (logR), which is not informative for these purposes. The bottom figure contains the BAF and the copy number fit where subclonal copy number is plot by a red line. The subclonal segments on this chromosome are fit with CCF values close to 50% of tumour cells. This QC fail can be resolved by adding a whole genome duplication to this copy number profile.

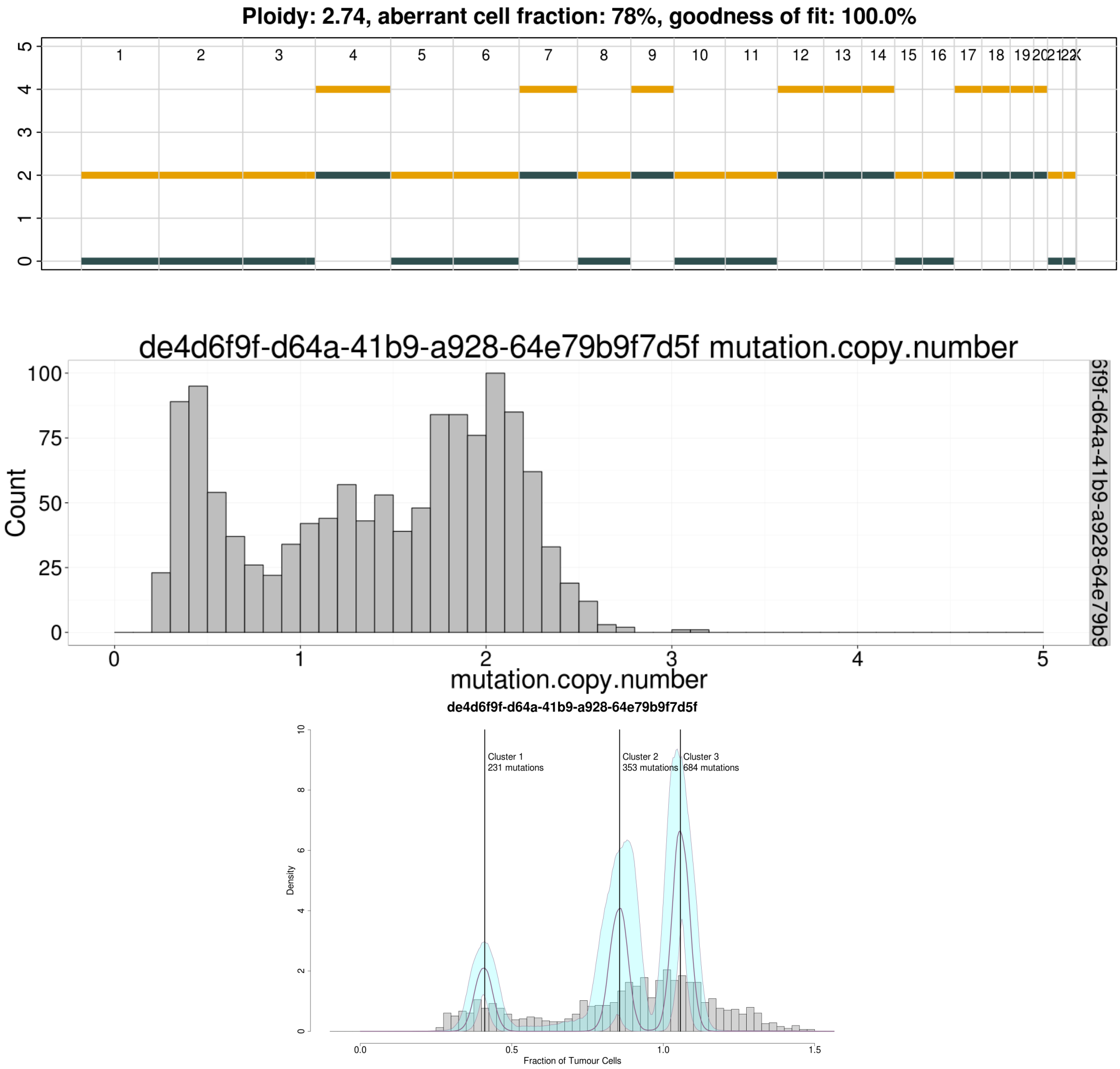


Fig. 4.5 This case fails QC because there are no segments fit with copy number state 1. In this scenario a whole genome duplication has been added that has not yielded an increase in the proportion of clonal copy number.

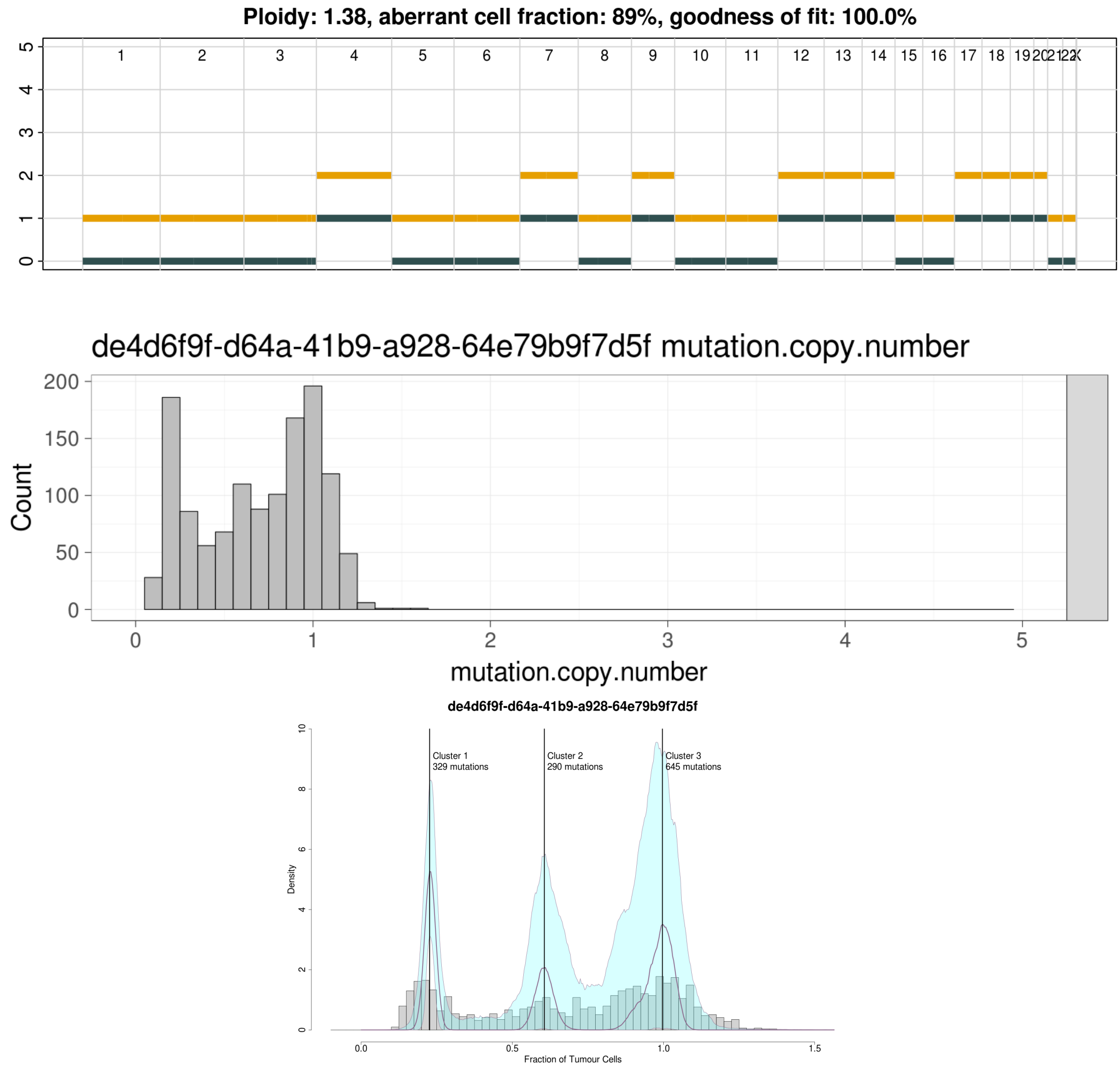


Fig. 4.6 The whole genome duplication is removed by refitting chromosome 1 with a 1+0 copy number state. This adjustment creates a clear mutation copy number peak at 1. Note that in the bottom figure Cluster 2 is not called at 50% of tumour cells and therefore does not fail that criteria (see Section 4.2.6).

4.2.4 No clonal copy number alteration

In some cases Battenberg does not find a solution with a single clonal copy number alteration. Battenberg requires at least one clonal alteration to estimate the purity, hence in cases like the one shown in Fig. 4.7 the purity estimate is incorrect. If there are no alterations in the profile then an alternative source must be used to estimate purity (for example from clonal SNVs). But in this case Battenberg has not been able to fit the segments on chromosomes 7 or 8 with a clonal copy number state, and in this scenario, the purity estimate is too high. This has affected the mutation copy number and CCF spaces by shifting the clonal peak to the left. There are two possible solutions for cases like this: force Battenberg to fit a selected segment with a particular copy number state, or obtain a purity estimate by other means (from SNVs in normal diploid 1+1 regions for example). The former approach focusses on the belief that there must be at least one clonal CNA, the latter on the belief that there might not be a single CNA.

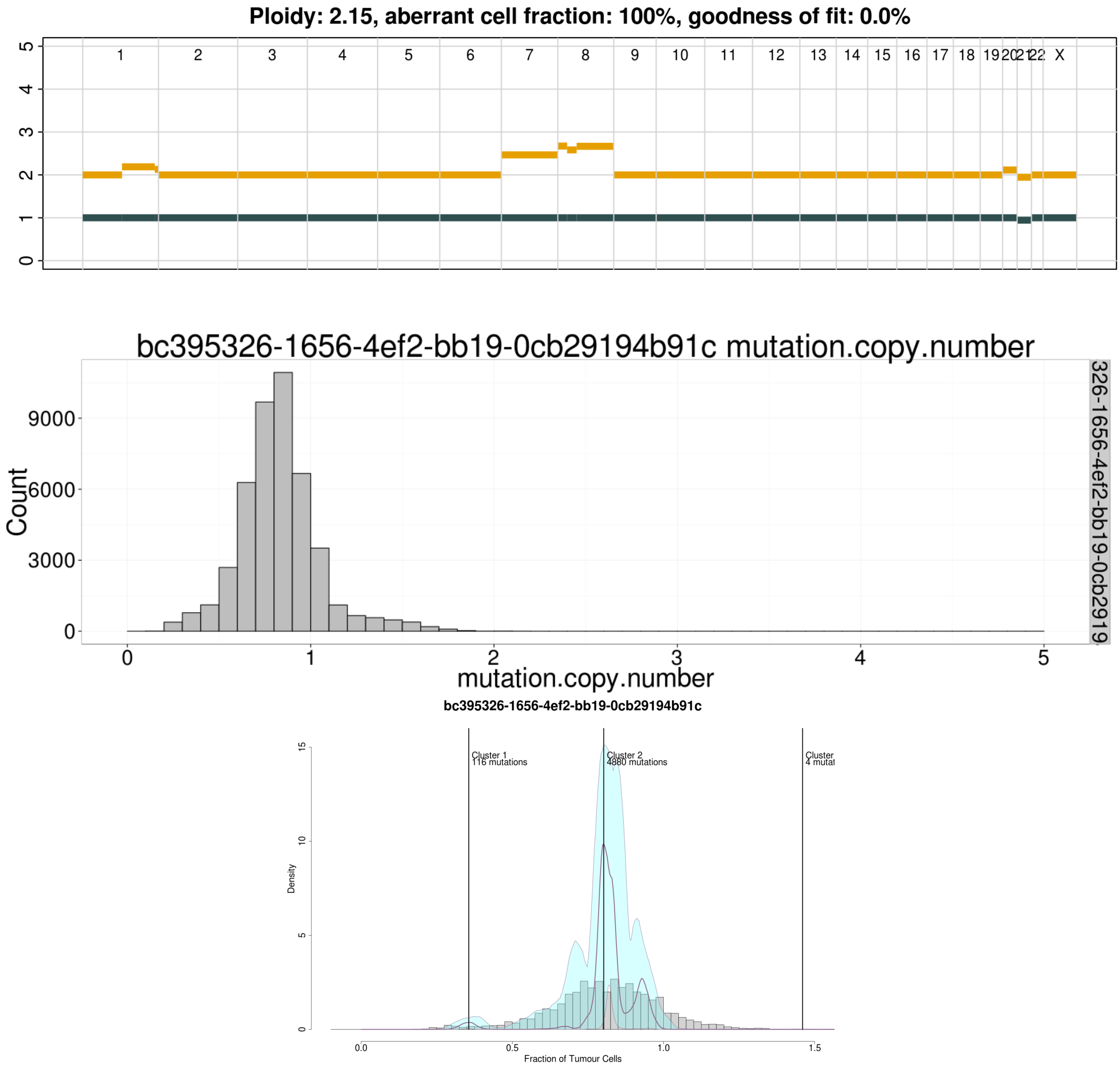


Fig. 4.7 Battenberg has not fit a segment with a clonal alteration. That means the purity estimate is most likely incorrect, supported by the shifted peaks in mutation copy number and CCF space.

4.2.5 Shifted clonal mutation cluster

The example highlighted in Fig. 4.7 would require attention due to its shifted clonal peak. In some cases a clonal peak can be shifted towards the right. That shift is caused by the winner's curse (see Chapter 6) where a number of clonal SNVs fell below the detection limit and that has caused the weight of the clonal distribution to shift. However, in cases where the clonal cluster is either seemingly fully sampled or the cluster is shifted to the left, the shift could be an indication that the purity estimate is incorrect. That can have different reasons for different profiles. For example, in Fig. 4.7 it is due to no segments being fit with a clonal alteration, but it can also be the effect of the wrong segment fit with a clonal state. In Fig. 4.7 fitting the altered segment on chromosome 1 as 2+1 will yield a very different purity and ploidy then fitting the chromosome 7 as 2+1. In cases of a shifted clonal cluster and no other QC violations the solution is often to look for an alternative clonal segment until the shift is resolved.

4.2.6 Mutation cluster at 50% of tumour cells

A clear mutation cluster at 50% of tumour cells can be a QC violation. But by chance one can observe a tumour with a subclone that takes up exactly half of the tumour. It is therefore not always clear whether a sample should pass or fail QC. In the example depicted in Fig. 4.8 there is a clear SNV cluster at 0.5 visible and the copy number profile also violates the subclone at 50% QC criterion (Section 4.2.2). Furthermore, DPClust finds an SNV cluster at a CCF of 1, but from the histogram it is not clearly there. The fact that Battenberg fits this profile with only chromosome 1q as a clonal alteration is suspicious as it would appear that chromosomes 6, 7, 8p, 8q, 11 and 18 could become clonal by 'stretching' the profile (i.e. 8p is 1+0, 8q is 5+1). That solution possibly shifts the 50% mutation cluster to become clonal. In such a scenario, a solution that yields a large proportion of the alterations as clonal is preferable as it one of the foundations upon which Battenberg is based. For the purpose of describing heterogeneity it would also lead to a conservative estimate if the clonal solution is incorrect. Fig. 4.9 shows that this provides a coherent fit that does not violate any of the criteria listed in this Chapter.

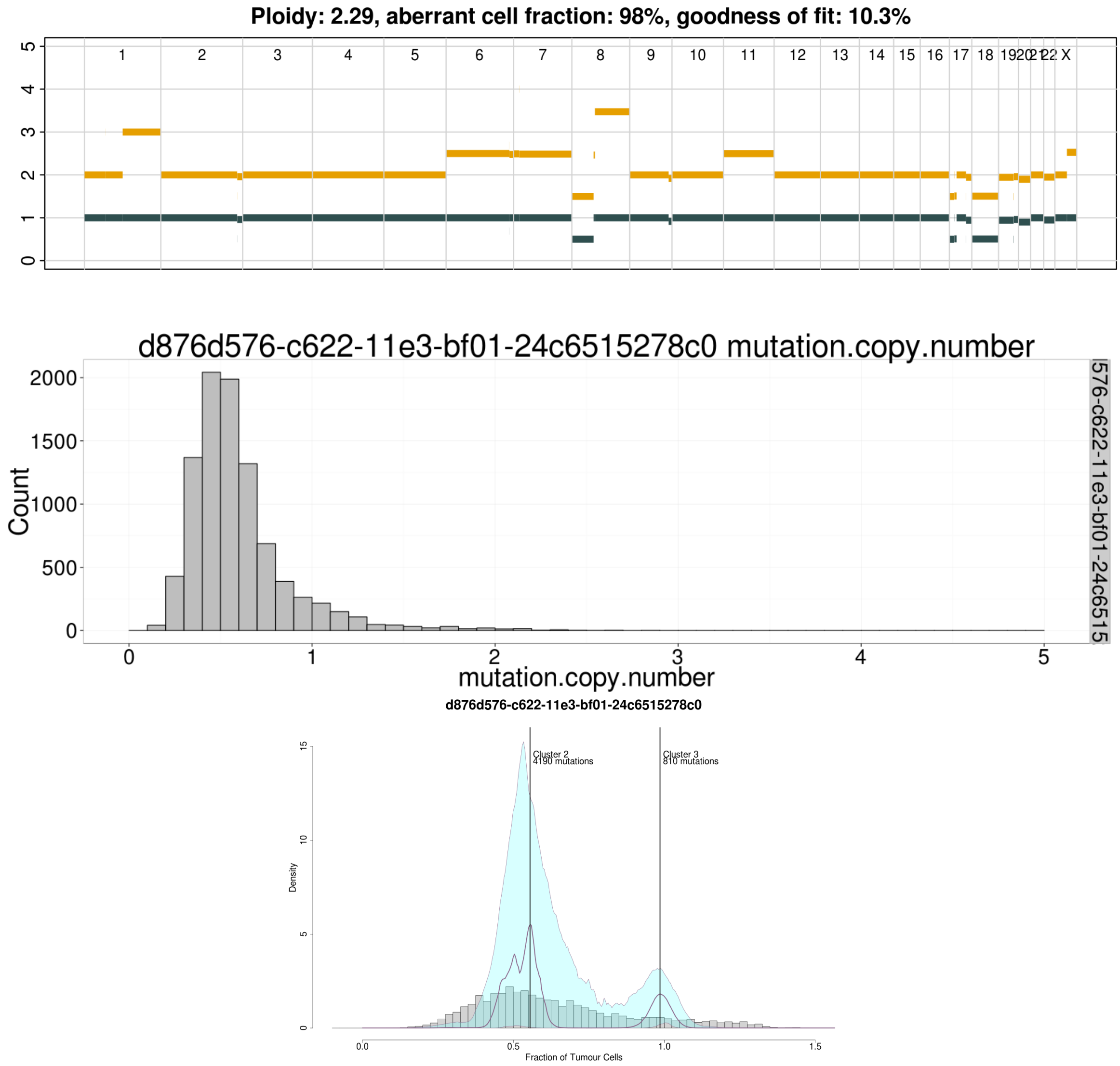


Fig. 4.8 Example of a case with a clear large mutation cluster at a CCF of about 0.5. In this particular scenario there are other QC violations, most notably the copy number segments in between two clonal states. The mutation cluster at 0.5 could be a sign that the ploidy needs doubling, in this case one could also try to choose another clonal segment over chromosome 1q.

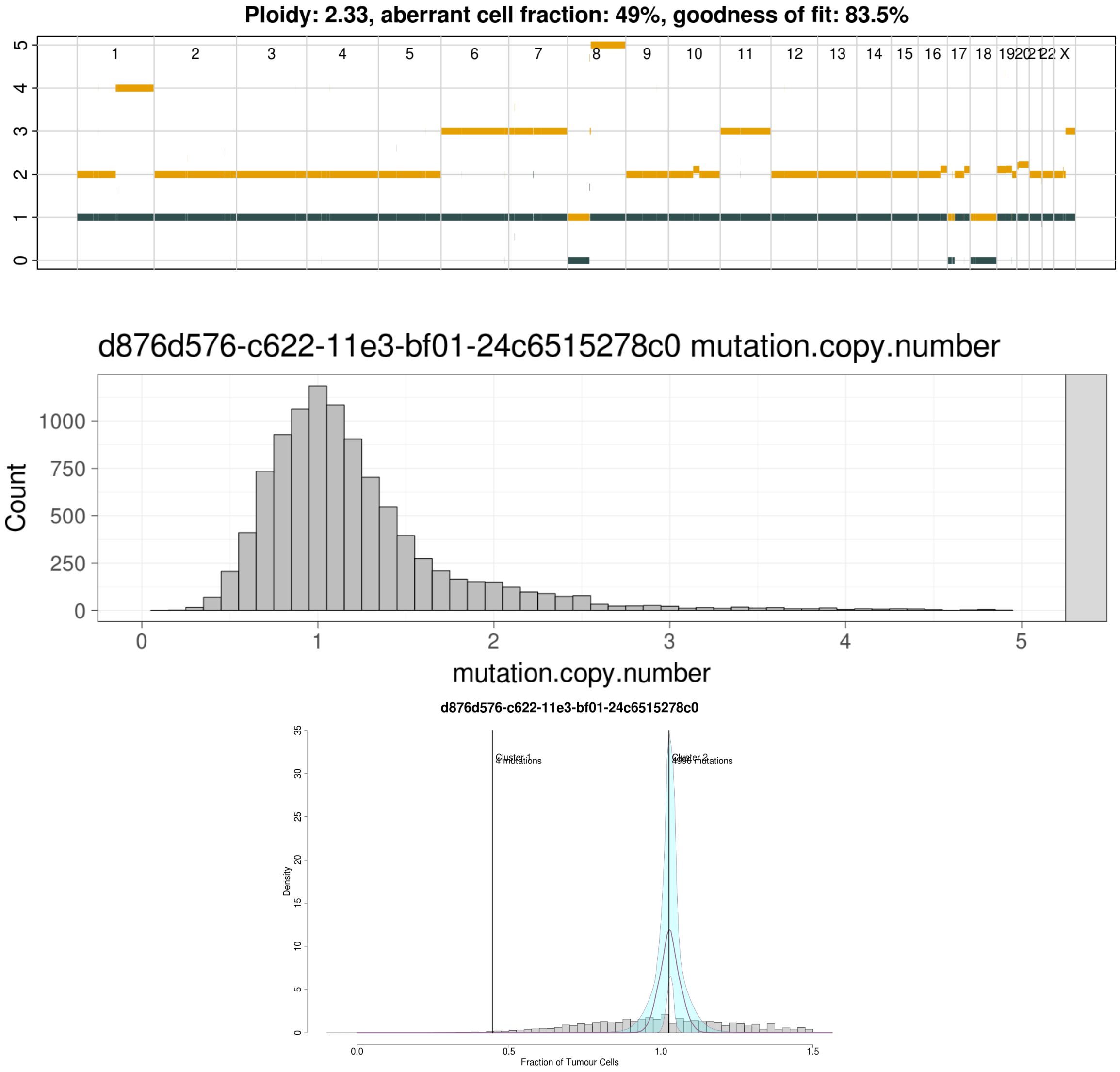


Fig. 4.9 The refit copy number profile does not contain any QC violations. The mutation cluster at 50% of tumour cells is removed, as are the copy number segments exactly in between two clonal states.

4.2.7 Empty mutation copy number state

In some cases the addition of an incorrect whole genome duplication or of an extra copy to some alleles can yield an empty mutation copy number state. Fig. 4.10 shows no clear peak at a mutation copy number state of 1, suggesting there are very few SNVs that are clonal and carried by a single chromosome copy. In this case the peak at mutation copy number 3 contains mutations on chromosome segments that are 3+0, whilst the peak at mutation copy number 2 contains mutations on the balanced 2+2 chromosomes. This particular example also contains an empty copy number state (see Section 4.2.3).

In this scenario the raw data can be equally likely explained by subtracting copies from every segment that does not have a copy number count of 0. One could refit with for example chromosome 2 as 2+0 or 1+0. This compresses the profile and will adjust the mutations with copy number states 2 and 3 to 1 and 2. That leads to a maximum parsimony explanation of the data as the additional duplication in the current profile does not allow for much more of the alterations to be explained as clonal and hence the data can be explained without the duplication. Fig. 4.11 shows the refit with chromosome 2 fit as 1+0, which yields a clean mutation copy number and DPClust figure and no QC criteria are violated.

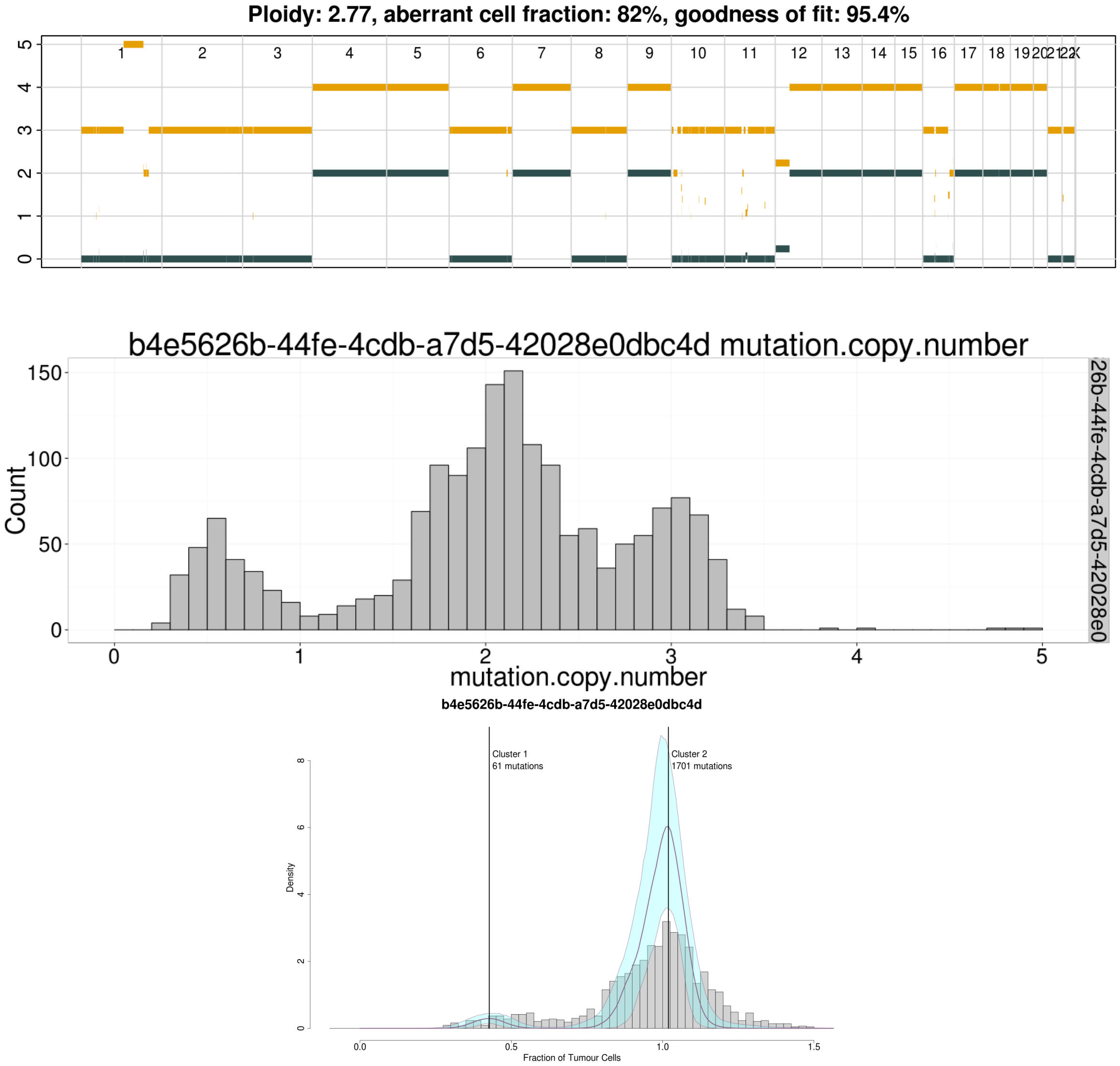


Fig. 4.10 An empty mutation copy number state can be an indication that additional chromosome copies have been added that do not help in explaining the largest possible proportion of the alterations as clonal. In this case there is no peak at mutation copy number 1, and the copy number profile also contains an empty state at 1.

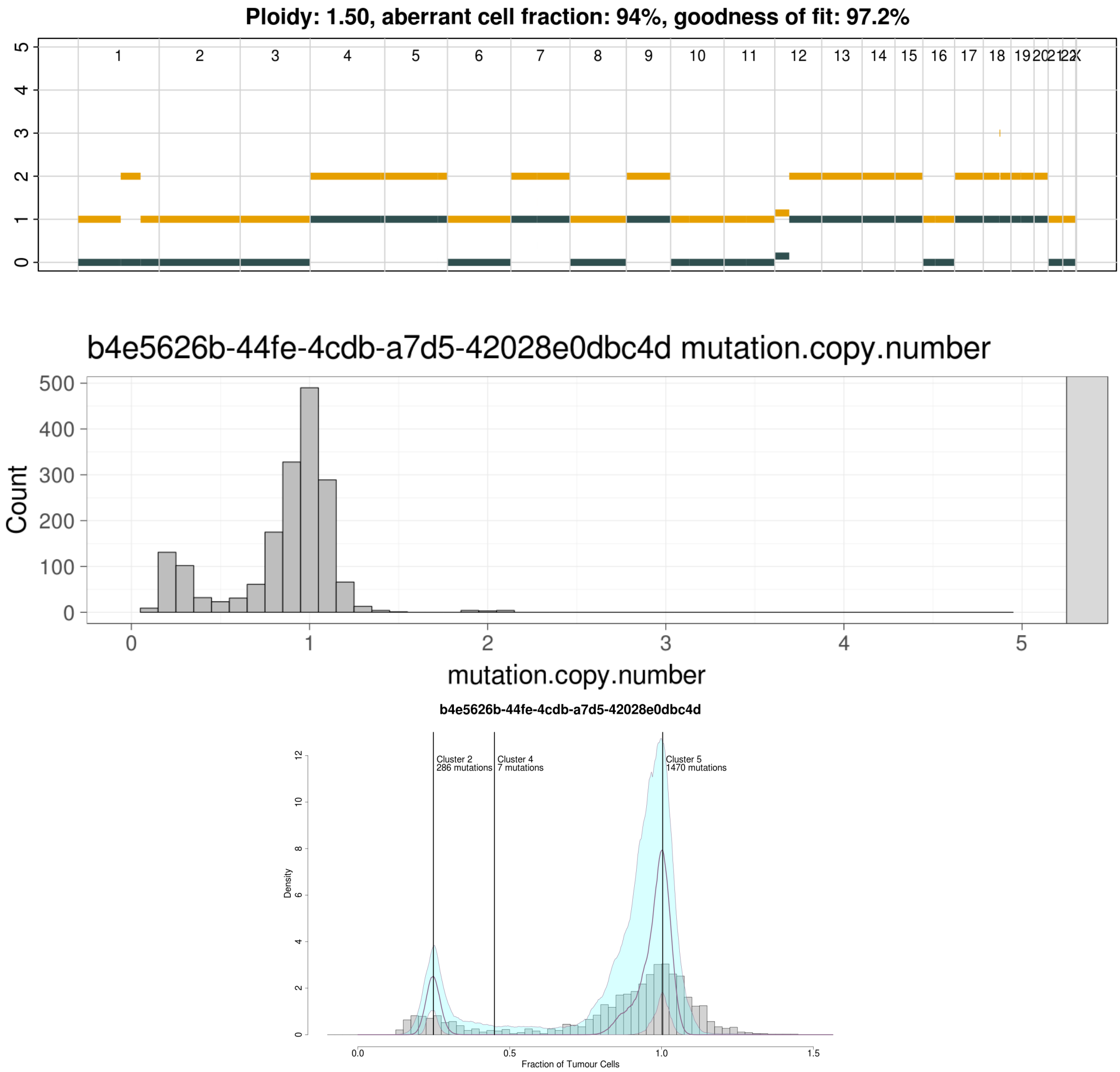


Fig. 4.11 A refit with chromosome 2 as 1+0 yields a clean CNA profile, mutation copy number figure and DPClust result.

4.3 Resolving a quality control failure

A QC fail can be resolved in multiple ways. An automatic procedure for Battenberg exists that obtains a purity estimate from SNVs, but leaves the ploidy unchanged. There is also a manual approach where one can request Battenberg to fit a copy number profile with a certain segment with a particular combination of major and minor allele states.

4.3.1 Automatic correction

An automatic correction is currently possible for cases where the purity should be estimated from another source because there are no clear clonal copy number alterations. A purity estimate can be obtained from SNVs in balanced copy number regions (preferably 1+1) where the VAF of clonal SNVs can be directly used. An estimate can be obtained by running DPClust in VAF space and to take the location of the SNV cluster closest to 1 and multiply it by 2 for a purity estimate. If SNVs in 2+2 regions are taken, the estimate should be adjusted appropriately. This approach does not make any adjustments to the ploidy and is therefore best suited for types of cancer with very quiet copy number profiles.

4.3.2 Manual correction

Manual correction is possible for copy number profiles that contain clonal alterations. One can hypothesize that a particular segment should have a particular combination of major and minor allele states, for example, chromosome 8p 1+0.

The ASCAT equations can be rewritten to obtain Eqs. 4.1 and 4.2 that convert the hypothesis into a suggested ρ and ψ parameter combination that Battenberg takes in when it fits a profile. Battenberg then skips the first fitting step to obtain an initial global fit, and it will start with the local optimisation to find the best solution.

In Eqs. 4.1 and 4.2, $n_{A,i}$, $n_{B,i}$ are the major and minor allele copy number states suggested for segment i and b_i and l_i are the BAF and logR respectively of segment i.

$$\rho = \frac{2b_i - 1}{2b_i - b_i(n_{A,i} + n_{B,i}) - 1 + n_{A,i}} \tag{4.1}$$

$$\psi = \frac{\rho(n_{A,i} + n_{B,i}) + 2 - 2\rho}{2^{l_i}} \tag{4.2}$$

After a refit suggestion, Battenberg produces a new profile, which is followed by a DPClust run and a new QC procedure.

4.4 Inventory of metric triggers in the PCAWG data set

I have attempted to incorporate the above metrics into a series of automated checks, through which a profile can be automatically flagged as either pass or fail. And applying these metrics to Battenberg profiles where no refitting has taken place can reveal how often these scenarios occur. I have therefore performed a rerun across all samples in the data set without refit suggestions of the copy number fitting pipeline, with the Battenberg version (2.2.5) that was used for PCAWG.

There are two main reasons why an additional run was required and these numbers could not be extracted from notes taken during the Battenberg PCAWG QC, or by simply comparing Battenberg with the PCAWG consensus. First, Battenberg has received a range of upgrades over the course of PCAWG. Most notably GC correlated wave correction and the inclusion of SV breakpoints. Both these additions were essential to increase performance on this heterogeneous data set. Second, a comparison against the PCAWG consensus profiles (detailed in section 6.2) would be imperfect as the PCAWG consensus consists of only clonal copy number states. This means that an unknown percentage of copy number segments is represented with a slightly different fit than the data suggests, which affects the calculated ploidy. A rerun of the exact same version of Battenberg with and without refitting is not affected by these downsides.

The comparison of two Battenberg runs reveals that refitting causes a discrepancy in either purity or ploidy in 15.2% of 2,748 samples for which output of both runs was available (Fig. 4.12). Nearly half of these are caused by the lack of a clonal copy number alteration (Table 4.1), while the other metrics trigger either between 20-30% or 10% or fewer cases. A total of 14 samples did not trigger any of the metrics. Manual inspection of the profiles revealed that in 7 cases the refit profile may be incorrect, as it triggered one or more metrics. The other 7 are the result of a bug that has been fixed at the time of writing, but was still prevalent in Battenberg version 2.2.5. This bug caused a slight discrepancy in the stored ploidy value, which could occasionally push subclonal gains into losses or vice versa.

It is often a single metric that causes the bulk of the triggers in a cancer type. Many cancer types with often quiet copy number profiles (pilo-astrocytoma, thyroid adenocarcinoma, benign bone cancers and AML) therefore often trigger the "No clonal CNA" metric. In this scenario one can use SNVs to estimate the purity and, without further evidence of a whole genome duplication available, set the ploidy to equal 2. It is for this scenario that

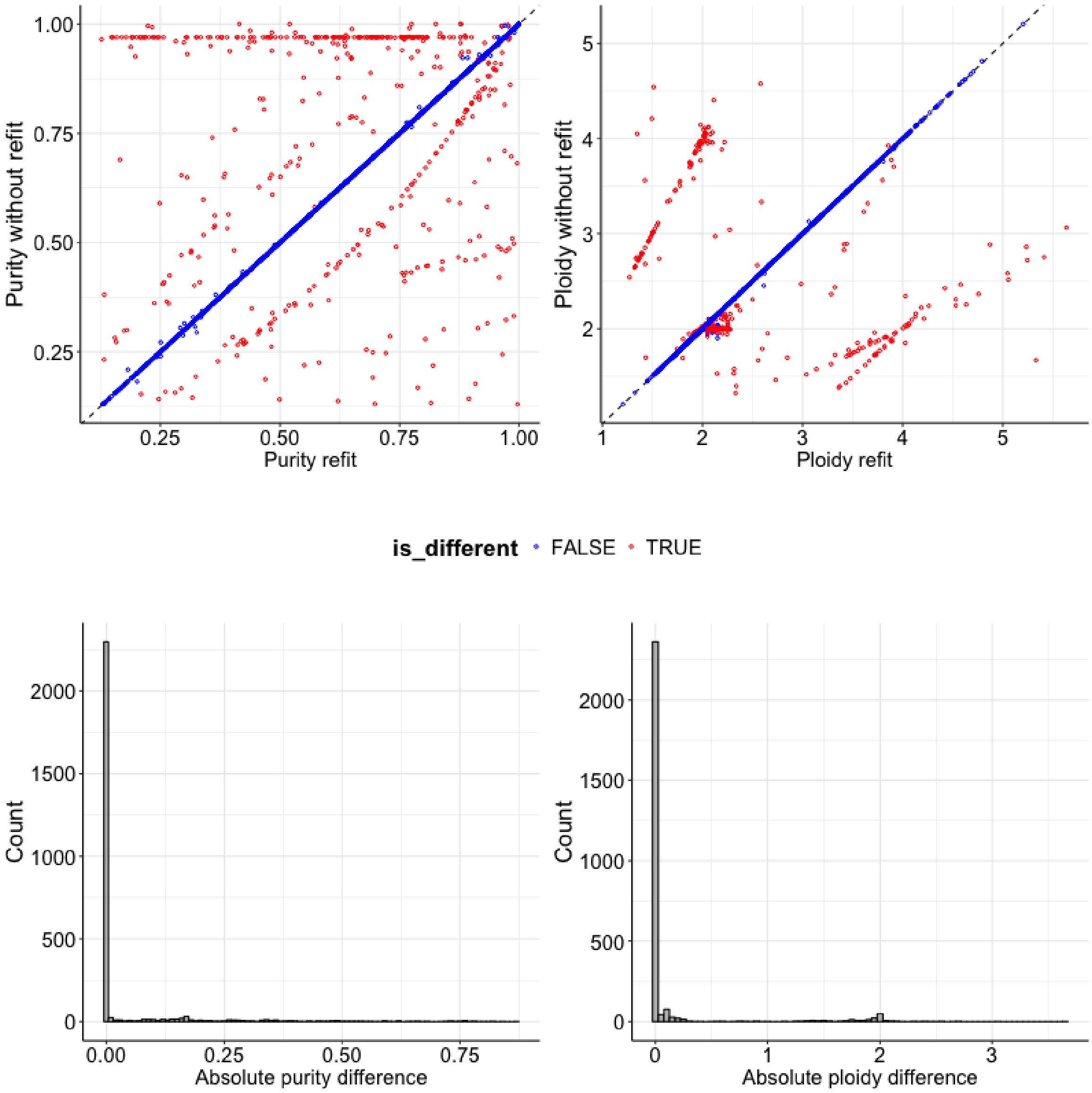


Fig. 4.12 Comparison of purity and ploidy values generated by Battenberg with and without refitting. The top figures compare the purity (left) and ploidy (right) where a discordant sample is coloured red. The bottom figures show a distribution of the difference between purity (left) and ploidy (right) between the two runs.

the automatic refitting pipeline was developed (see section 4.3.1), which should be a future extension of the Battenberg pipeline.

Pancreatic endocrine cancers often trigger scenario C (empty copy number state). This may also reflect the underlying biology. Pancreatic endocrine tumours often show very little subclonal copy number alterations, often contain whole chromosome LOH and (to a lesser extent) whole chromosome gains and relatively frequent whole genome duplications.

Metric	Num. cases	Frac. different	Frac. total
A. No clonal CNA	207	49.2%	7.4%
B. CNA subclone at 50%	108	25.77%	3.8%
C. Empty CN state	98	23.3%	3.5%
D. Shifted clone	88	20.9%	3.2%
E. SNV cluster at 50%	43	10.2%	1.5%
F. Large hom del	32	7.6%	1.1%

Table 4.1 Overview of QC metric triggers between Battenberg with and without refitting. Almost half the cases with a discrepancy in either purity or ploidy contain copy number profiles without a clonal CNA. Between 20-30% of cases trigger a CNA subclone at near 50% of tumour cells, an empty copy number state or a shifted clonal cluster. Around 10% of cases contain an SNV cluster near 50% of tumour cells or a large homozygous deletion.

Histology	A	B	C	D	E	F	Samples	Samples diff.	Frac. diff.
CNS-PiloAstro	64	2	3	2	1	2	88	69	0.78
Prost-AdenoCA	44	15	5	6	5	2	284	65	0.22
Liver-HCC	7	19	5	8	7	4	326	34	0.10
CNS-Medullo	4	8	12	12	4	6	139	31	0.22
Panc-Endocrine	0	2	26	1	2	2	85	27	0.31
Thy-AdenoCA	20	0	4	1	1	0	48	23	0.47
Lymph-BNHL	8	5	5	8	0	1	106	19	0.17
Kidney-RCC.clearcell	4	2	6	6	1	1	111	14	0.12
Lymph-CLL	8	2	7	4	2	0	90	14	0.15
Kidney-ChRCC	2	1	6	0	0	1	45	11	0.24
Bone-Benign	9	0	1	0	0	0	16	10	0.62
Myeloid-AML	7	2	0	0	0	0	16	7	0.43
Kidney-RCC.papillary	3	2	3	3	1	2	33	6	0.18
Bone-Osteosarc	2	3	1	0	2	0	38	6	0.15
Myeloid-MPN	4	0	1	1	1	0	45	5	0.11

Table 4.2 The number of samples triggering the six metrics, split per cancer type: A=No clonal CNA, B=CNA subclone at 50%, C=Empty CN state, D=Shifted clone, E=SNV cluster at 50%, F=Large hom del.

This means that for the Battenberg metric (the proportion of the genome that is fit with a clonal state) is often extremely similar between a profile with and without a whole genome duplication.

Furthermore, it is possible that some pancreatic cancers indeed contain an empty copy number state. Such a scenario can occur if the last copy number alteration to occur is a whole genome doubling. Sample SA570847, shown in Fig. 4.13, contains many SNVs on 1 and many SNVs on 2 chromosome copies. During the PCAWG expert panel review (see

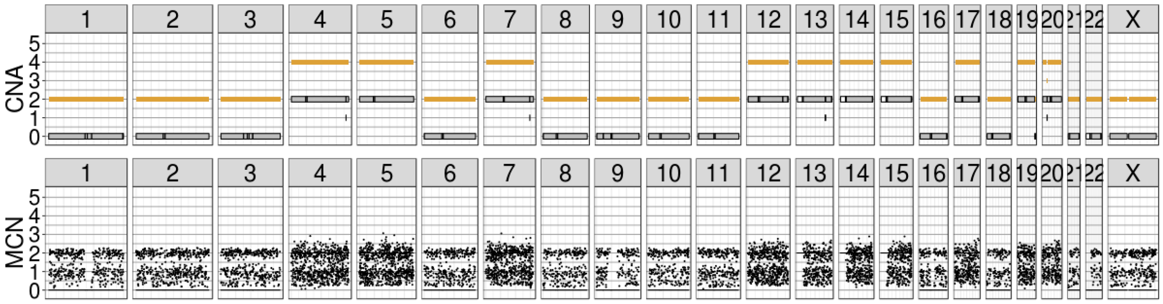


Fig. 4.13 The copy number profile of PCAWG tumour SA570847 (top), with the total copy number in orange and the minor allele in grey. The bottom figure shows mutation copy number (MCN, the raw estimated number of chromosome copies an SNV is carried by) for all SNVs detected in this tumour. SA570847 clearly shows a large number of SNVs on 1 and on 2 chromosome copies, justifying the addition of a whole genome doubling to the copy number profile, even though it leaves almost no segment at 1 chromosome copy.

section 6.2.6) we have occasionally allowed an empty copy number state based on convincing evidence of a genome doubling.

This example suggests that, even though these metrics capture the essence of what a manual QC captures, there can be exceptions due to specific characteristics of a particular type of cancer. It also highlights that a combination of metrics can sometimes lead to convincing evidence that contradicts a single metric. This is a sign that a combination of metrics could be a fruitful approach. It is unclear however, how to adequately weight multiple metrics against each other. A machine learning approach may be able to learn weights between the metrics by using the metrics from the PCAWG data set. This may be an interesting direction to explore in the future in order to further improve the metrics system.

Chapter 5

The subclonal architecture and life history of a single cancer

5.1 Introduction

In the previous chapters I have described methods to call copy number, infer the subclonal architecture of a tumour and quality control the results. In this chapter I explore what the inferred architecture reveals about the events that took place during a tumour's development and what can be learned about a single cancer from these data. The aim of this chapter is to visibly show what the previously introduced methods do, without a deep understanding of their internal workings.

To this end I have selected a single breast cancer sample of which the copy number profile and subclonal architecture are very clear. The sample however originates from a project that is not yet complete and therefore not described in this thesis. This project comprises the whole genome sequencing of breast cancers from patients in Nigeria with the aim to explore the tumours' subclonal architecture and life history, the background of this project is briefly described in the next section. For the purposes of this chapter I present the selected tumour as 'a cancer' and therefore do not consider its additional unique features, such as the germline context, occurrence rates of breast cancer subtypes and differences in (access to) healthcare when compared to tumours from patients in the Western world.

The sequencing read alignment and variant calling work described in section 5.3 was performed by Jason Pitt.

5.2 Background

Over the last few decades, enormous progress has been made in treatment of breast cancer. In the UK, between 1988 and 2013 mortality has dropped from 60 per 100,000 to below 40, even though incidence rate has gone up from 120 to 170 per 100,000 [1]. However, a big discrepancy remains between women of different ethnic backgrounds. American women of African American ancestry consistently showing lower treatment success, even though survival rates show a similar improvement obtained for American women of European descent (Servick, 2014).

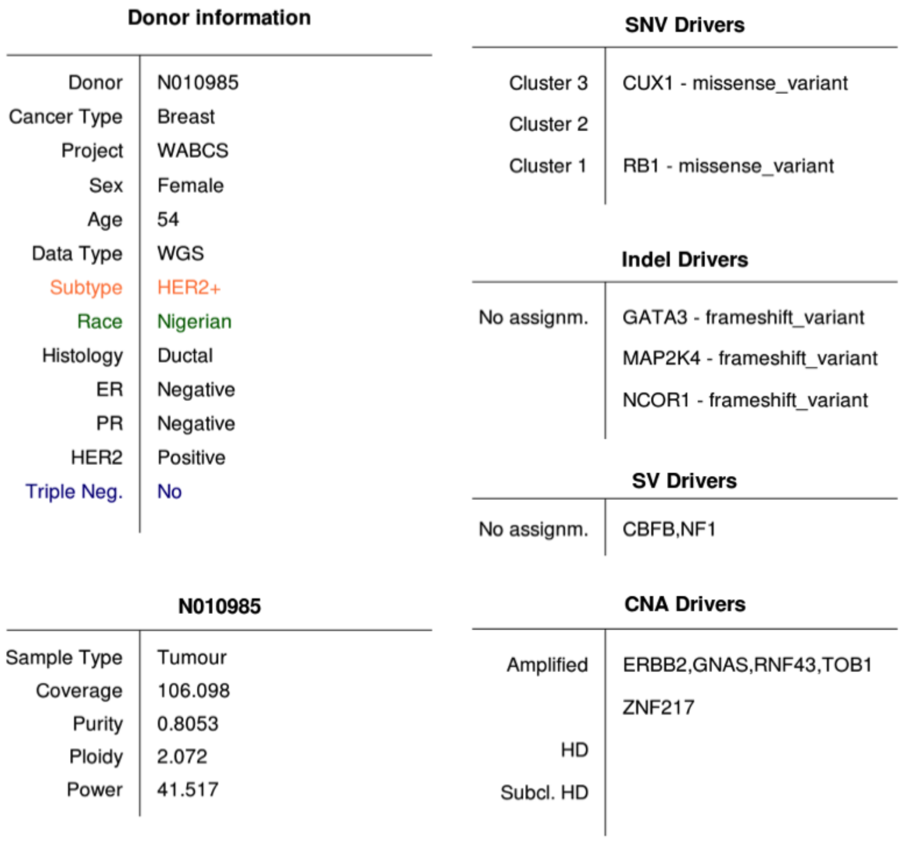
The reasons behind this disparity are thought to be a complex interplay between socio-economic and tumour biology differences (Daly and Olopade, 2015). American women of African ancestry are less likely to be diagnosed with breast cancer, however tumours are diagnosed at an earlier age and at higher tumour stage compared to American women of European descent (Iqbal et al., 2015). Breast tumours are more often of the triple negative subtype (Ray and Polite, Feb) and there is a higher prevalence of *BRCA1* and *BRCA2* germline carriers among African women (Fackenthal et al., 2012). Meanwhile, several social boundaries (Jones et al., 2014) and differences in patterns of referral have been described (Daly and Olopade, 2015), including that African American women with a family history of breast cancer are less likely to undergo genetic counselling (Armstrong et al., 2005).

The West African Breast Cancer Study (WABCS) was set up to further investigate the tumour biology and genetics of breast cancers from western Africa and is aimed to provide a comprehensive overview by sourcing and sequencing tumours (WXS or WGS of DNA and RNA-seq) from West Africa. The study consists of various projects focussing on predisposing germline loci, a landscape of somatic alterations and unveiling patterns tumour evolution. I am part of the tumour evolution project where we aim to describe the life history of breast cancers from Africa and investigate whether there are different patterns of evolution, when compared to those obtained from women in North America. At the point of writing, the study is in progress, with no finalised results. The sample described in this chapter is one of 98 whole genome sequenced tumours that are part of the study and was specifically picked for its clear copy number profile and subclonal architecture to aid the purpose of this chapter.

5.3 Methods

Sample N010985 was resected from a 54 year old patient in Nigeria (Fig. 5.1). Six needle biopsy samples were taken, of which one was prepared for whole genome sequencing. The

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



Donor information

Donor	N010985
Cancer Type	Breast
Project	WABCS
Sex	Female
Age	54
Data Type	WGS
Subtype	HER2+
Race	Nigerian
Histology	Ductal
ER	Negative
PR	Negative
HER2	Positive
Triple Neg.	No

N010985

Sample Type	Tumour
Coverage	106.098
Purity	0.8053
Ploidy	2.072
Power	41.517

SNV Drivers

Cluster 3	CUX1 - missense_variant
Cluster 2	
Cluster 1	RB1 - missense_variant

Indel Drivers

No assignm.	GATA3 - frameshift_variant MAP2K4 - frameshift_variant NCOR1 - frameshift_variant

SV Drivers

No assignm.	CBFB,NF1

CNA Drivers

Amplified	ERBB2,GNAS,RNF43,TOB1 ZNF217
HD	
Subcl. HD	

Fig. 5.1 General annotations of breast cancer case N010985 (left column) and identified potential drivers (right column). The top left table contains information about the donor, including age, ethnicity, project (WABCS stands for West African Breast Cancer Study), the type of sequencing and the inferred ER/PR/HER2 status. The bottom left table shows statistics about the tumour sample: coverage, purity and ploidy. It also shows the number of reads per chromosome copy (labelled as power), which determines the power to detect subclones (see Chapter 6 for a description of the metric).

tumour biopsy and a blood sample from the same patient were sequenced on an Illumina X10 machine to a coverage of 100x and 30x respectively. Histology of the tumour was examined by pathologists in Nigeria and at the University of Chicago, after which it was classified as a ductal carcinoma. Accompanying RNA-seq data was used to infer that the tumour ER-negative and HER2-positive.

After passing initial sequencing quality control metrics the obtained reads were aligned to the GRCh37 reference genome using BWA (Li and Durbin, 2009), after which SNV calling was performed using Strelka (Saunders et al., 2012) and Mutect (Cibulskis et al., 2013), indel calling using Strelka and SVs were obtained by applying Delly (Rausch et al., 2012) and Lumpy (Layer et al., 2014). To obtain reliable SNV and SV calls the results from the two methods were intersected and filtered by an unmatched normal panel. For indels only the filtering by panel was applied.

5.4 Subclonal architecture

The sequencing yielded 18,813 SNVs, 382 indels and 335 SVs. A copy number profile was fit using the Battenberg algorithm, which yields a relatively quiet profile with a ploidy just over 2 and a purity of 81% (Fig. 5.2a). The tumour consists of a clone with an estimated 9,794 SNVs and two subclones with 2,792 and 5,530 SNVs (Fig. 5.2b). At the time of writing the VAF adjustment pipeline for indels is not ready, hence indels are not assigned to mutation clusters.

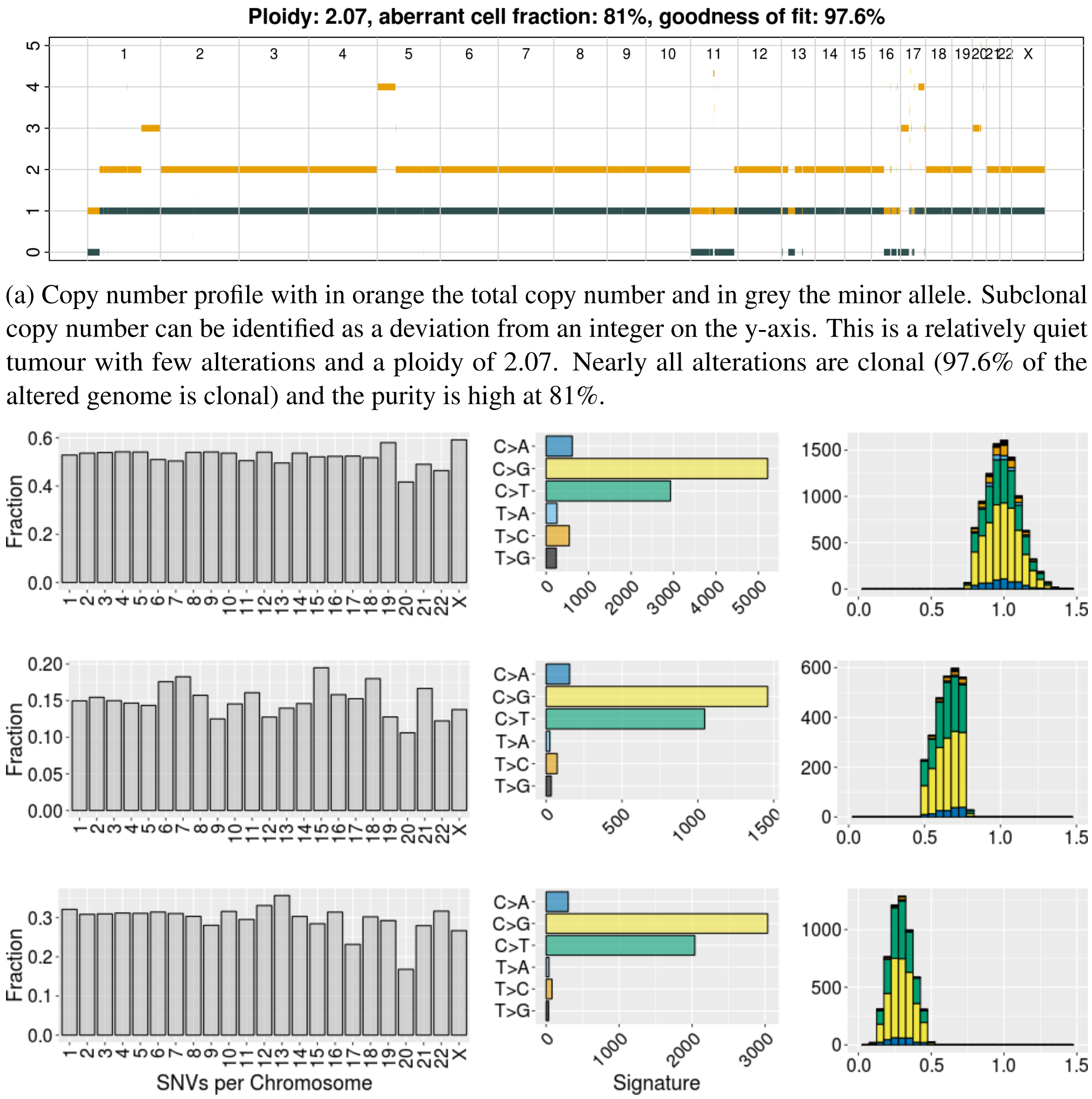
The mutations (SNVs, indels, SVs, amplifications and homozygous deletions) were intersected with a putative list of 149 genes thought to be involved in breast cancer development (Fig. 5.1). This list consists of genes taken as the top hits reported in (Nik-Zainal et al., 2016) and all genes in which a driver was found in a breast cancer in the ICGC pan-cancer dataset (Sabarinathan et al., 2017).

This analysis yields a clonal missense variant in *CUX1*, which is carried by 1 chromosome copy in a balanced copy number region and a subclonal missense variant in *RB1*, which falls in a region of clonal LOH where only a single copy of the locus is available. It's unclear whether the RB1 mutation deactivates the remaining copy of *RB1*.

Analysis of indels yields deletions of 11 bases in *MAP2K4* and 14 in *NCOR1* in a region of LOH where there are 3 copies of one allele and an insertion of a single base into *GATA3* in a region of balanced copy number. Raw CCF values for these variants are 1.00, 1.07 and 1.05 respectively, all three are therefore most likely clonal. The *MAP2K4* deletion is reported by 104 out of 124 reads and the *NCOR1* deletion by 132 out of 159, which suggests the mutations are clonal and carried by multiple chromosome copies.

Amplifications and homozygous deletions were obtained from the copy number data by selecting focal segments (< 1Mb). A segment is classified as an amplification when the total copy number exceeds 2*ploidy+1 and as a homozygous deletion when both alleles have been lost (clonally or subclonally). This results in three genes on chromosome 17 (*ERBB2*, *TOB1* and *RNF43*) and two on chromosome 20 (*ZNF217* and *GNAS*) being classified as amplified. The *ERBB2* is also known as *HER2* and is a primary driver of this tumour.

A copy number or SV breakpoint is found within the *CBFB* and *NF1* genes (Fig. 5.3). *CBFB* contains a copy number breakpoint where the first 3 exons are deleted. No other disrupting event is found, which suggests one copy of *CBFB* remains intact. *NF1* contains multiple breakpoints, which results in gaps in the local copy number as segments in Bat-

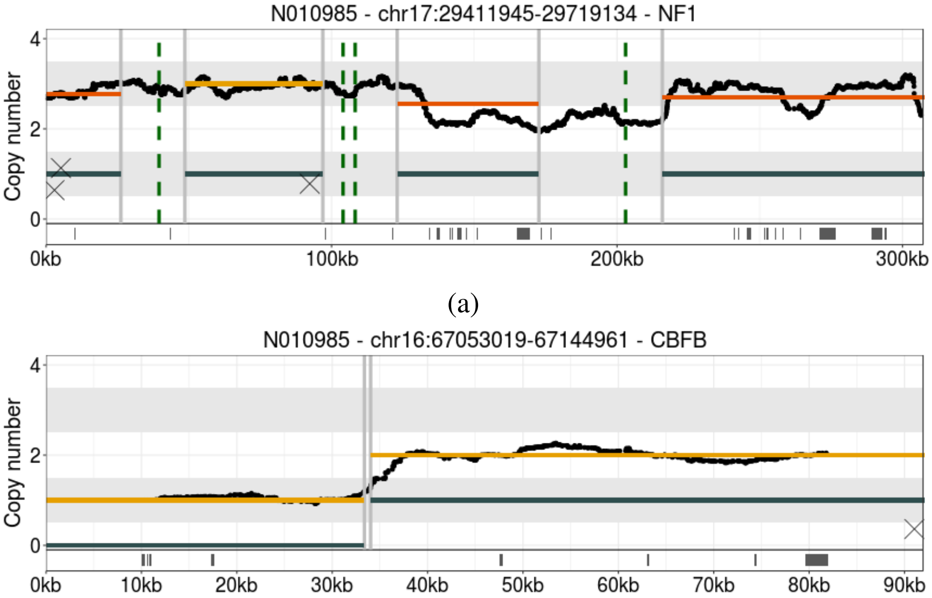


(a) Copy number profile with in orange the total copy number and in grey the minor allele. Subclonal copy number can be identified as a deviation from an integer on the y-axis. This is a relatively quiet tumour with few alterations and a ploidy of 2.07. Nearly all alterations are clonal (97.6% of the altered genome is clonal) and the purity is high at 81%.

(b) Summary of the subclonal architecture with a row for each mutation cluster identified. The left column shows the number of SNVs per chromosome, the middle column counts for each of the six possible base substitutions and the right column the raw CCF values of the SNVs assigned to the cluster. This tumour consists of three mutation clusters, a clone and two subclones. All three contain a high number of C>G and C>T mutation types.

Fig. 5.2 Subclonal architecture and copy number profile of N010985.

tenberg start and end at a germline heterozygous SNP. The copy number fit suggests there are three copies of *NF1*, of which one contains a deletion of exons 6-16. The SVs suggest the whole gene up to 200kb was duplicated, and both regions marked with subclonal copy number are supported by deletion calls. Regardless, given the copy number, there is at least



(a)

(b)

Fig. 5.3 Detailed figures showing the mutations measured in *NF1* and *CBFB*. The figures contain the copy number profile in orange and grey (total copy number and minor allele), raw total copy number calculated from the coverage in the background in black, SNVs as X-es (grey when non-coding, black when coding, there are no coding mutations found in either gene), copy number breakpoints as grey vertical lines and SV breakpoints as green dashed lines. Below the mutations is a track that shows the exons of the default transcript from Ensembl.

one working copy of *NF1* remaining as no other disrupting events have been found. *NF1* is unlikely to be a driver of this tumour as it is a tumour suppressor gene (Cichowski and Jacks, 2001). One copy of *CBFB* remains in tact, and without evidence of a fusion with *MYH11* or deactivation of *RUNX1* this gene is also unlikely to be a driver (Banerji et al., 2012).

5.5 Mutational Signatures

Mutational signature analysis was restricted to nine selected signatures that have been called de novo by Jason Pitt on a large set of breast cancer exomes from Nigerian patients, also part of the WABCS project (Pitt et al., 2018). The signatures found have been matched against the COSMIC signatures to determine the labels (Forbes et al., 2017). I subsequently quantified the activity of each of the nine signatures using the MutationalPatterns R package (Blokzijl et al., 2018).

The signatures reveal strong APOBEC activity in the clone and both subclones (Fig. 5.4). There is a larger relative contribution of the C>T APOBEC signature in both subclones when compared to the clone, which may be an indication that the C>T signature has a later onset in this tumour or that the activity rate of the two APOBEC signatures varies. The other seven signatures do not contribute substantially and their detected presence in low proportions could be noise.

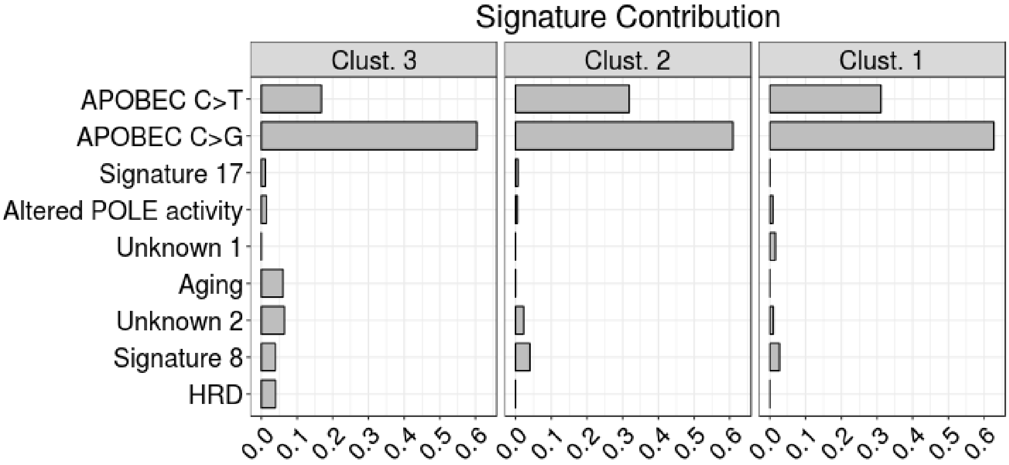


Fig. 5.4 Mutational signature analysis of SNVs assigned to the three mutation clusters reveals steady APOBEC C>G and C>T signature activity.

Kataegis events in N010985

Region	Chromosome	Size (bp)	Num SNVs	C>A	C>G	C>T	T>G
1	2	12356	23	3	10	10	0
2	17	1063	11	2	0	7	2
3	19	6079	13	1	6	6	0
4	20	10666	35	3	22	10	0

Table 5.1 Four regions containing kataegis have been identified in N010985. All four regions contain a large proportion of C>G and C>T mutations associated with APOBEC activity. Regions 1 and 3 contain an equal number of C>Gs and C>Ts, while regions 2 and 4 show an imbalance between the two types of substitutions. No T>A and T>C substitutions have been identified in any of the regions.

5.6 Kataegis

APOBEC activity is associated with local hypermutation, known as kataegis (Nik-Zainal et al., 2012b). Regions containing kataegis are obtained by first segmenting the intermutational

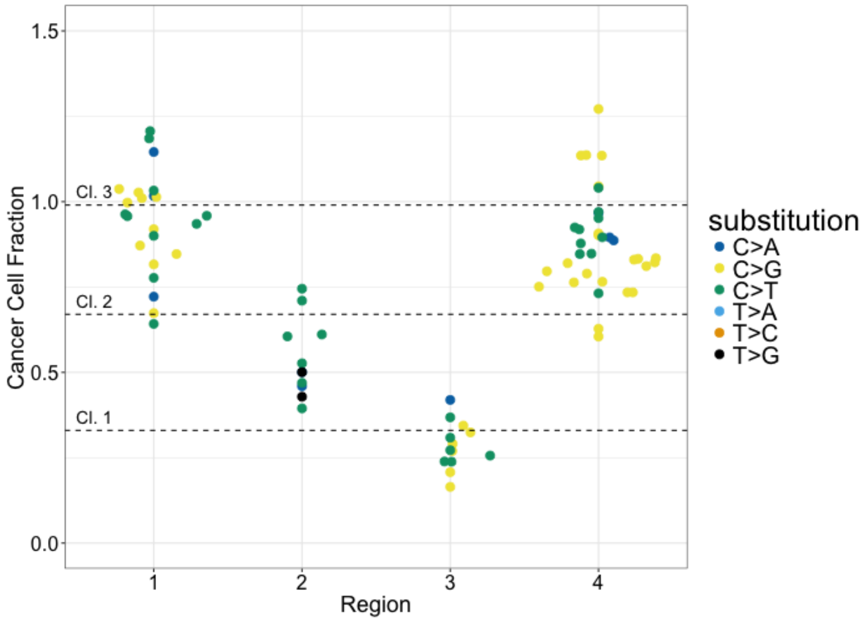


Fig. 5.5 Overview of four kataegis regions found in N010985. Regions 2 and 3 appear to be subclonal, while regions 1 and 4 may consist of multiple kataegis events. All four regions predominantly contain SNVs in APOBEC C>G and C>T contexts.

distance using (i.e. grouping mutations in stretches of similar distance) and subsequently selecting regions with a consistent short distance. The distance threshold is set depending on the mutation rate of the tumour: it must be below an average of 100 base-pair in tumours with over 90,000 SNVs, 250 in tumours with between 50,000 and 90,000 SNVs, 500 when between 10,000 and 50,000 and the threshold is set to 1,000 base-pair below 10,000 SNVs.

I identify four regions are with local hypermutation in this tumour (Table 5.1). All four regions contain SNVs that can be explained as the result of APOBEC signature activity. Two regions show an imbalance between the number of C>G and C>T substitutions, with region 2 containing no C>Gs and region 4 containing more than double the number of C>Gs. This suggest that both APOBEC signatures can independently generate kataegis events and that some of the regions identified may be a combination of multiple events.

SNVs in kataegis regions are routinely excluded from subclonal architecture inference. The localised hypermutation causes reads to contain multiple variants, which may impact the read alignment quality and result in more variable VAFs. However, analysis of the raw CCF estimates of the SNVs in the four regions suggests regions 2 and 3 contain subclonal kataegis events (Fig. 5.5). There appears to be separation between C>G and C>T SNVs in region 4, with the C>G SNVs possibly belonging to cluster 2.

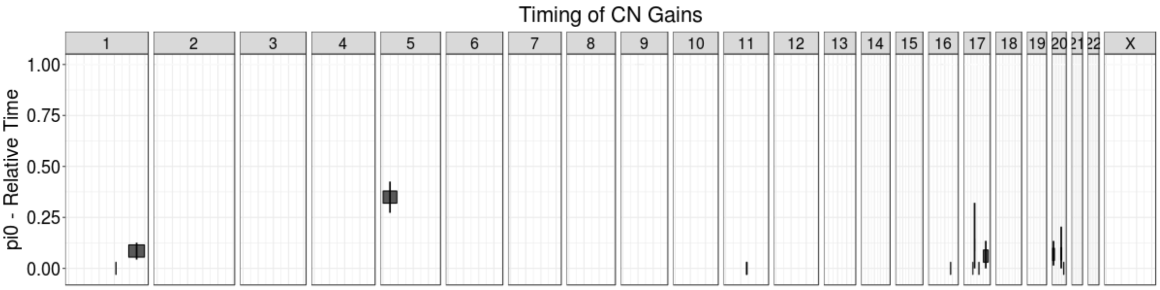


Fig. 5.6 Timing of gains analysis showing all gains with at least 10 SNVs. The y-axis represents the relative ordering, with a low value meaning the gain occurred relatively early according to the mutation data on the segment. The thin vertical lines are confidence intervals.

5.7 The life history of N010985

Analysis of the timing of gains (Fig. 5.6) reveals that gains on chromosome 17, 20 and 1q are early, followed by a gain of chromsome arm 5p. This analysis utilises the ratio of SNVs that are on multiple copies and on a single chromosome copy, where a low ratio indicates the gain is early. Timing of gains was performed using cancerTiming (Purdom et al., 2013), with segments restricted to those with 10 or more SNVs.

Previously I have split events measured in this tumour into clonal and subclonal. The timing of gains analysis allows for splitting clonal events into clonal early, late or undefined. Meanwhile, potential driver mutations can also be classified by taking into account the multiplicity.

The tumour's life history can now be compiled from the accumulated evidence (Fig. 5.7). It starts with deletions in *NCOR1* and *MAP2K4*, which are subsequently gained. The loss of 17p is most likely also early as it deletes the remaining intact copy of both genes, but the loss cannot be timed. Gains of chromosomes 1q and 20 are also early. These events appear in the first 150 measured SNVs. Mutational signature analysis suggests both APOBEC mutational signatures are already active.

Then follows a range of events that cannot be accurately timed. This phase contains an SNV in *CUX1* and an insertion into *GATA3*, in both cases on one of two available copies, and a loss of one copy of *RB1*. It also contains amplifications of *ERBB2* and other genes on chromsome 17p and 20q and losses of 1p, 11 and 16q. This period represents a large proportion of the tumour's life history consisting of 9,647 SNVs. APOBEC signatures remain constantly active. Multiple kataegis events with an APOBEC context are observed.

Finally, subclonally there is the second deactivating event for *RB1*, which suggests *RB1* is the driver of a subclonal expansion. Also observed are further gains of segments on chromosomes 11 and 17 and multiple kataegis events associated with APOBEC activity.

The subclonal architecture leaves two possible tree representations, a branching and a linear tree (Fig. 5.8). Phasing of SNVs did not yield a mutually exclusive pair that could have ruled out the linear tree. It is therefore not possible to resolve the tree topology.

This life history can be put in perspective by comparing it to the combined life history of all breast cancers (which is available as Appendix B, as it is part of the supplementary figures of Gerstung et al. (2017)). N010985 does not have a whole genome duplication. PCAWG tumours without a genome doubling event typically contain early gains, which is also observed in N010985 (Appendix B Fig. A).

The overall breast cancer life history (Appendix B Fig. B) shows that loss of 17p and 13q (*RB1*) are indeed most likely early. The gain of 1q and driver mutations in *GATA3* are typically early, but later than losses of 17p and 13q. Mutational signature analysis of the PCAWG breast cancers suggests that APOBEC activity is highly variable between cancers, with some tumours showing high early or late activity, while in others APOBEC activity remains constant (signatures 2+13 in Appendix B Fig. D/E).

These findings highlight what a subclonal reconstruction can tell about the life history of a single cancer.

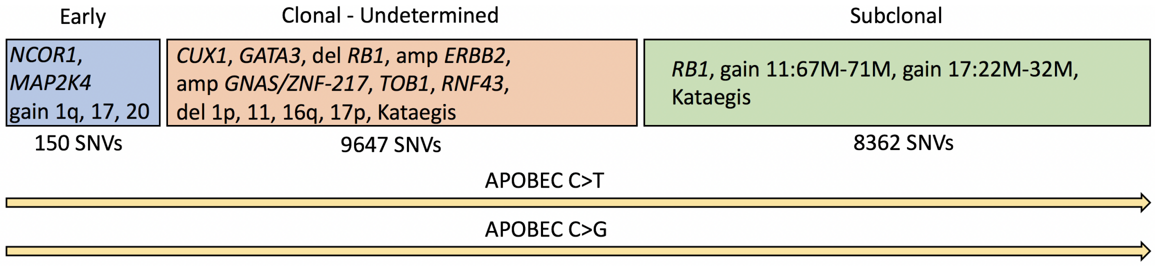

Early
Clonal - Undetermined
Subclonal
NCOR1, MAP2K4 gain 1q, 17, 20
CUX1, GATA3, del RB1, amp ERBB2, amp GNAS/ZNF-217, TOB1, RNF43, del 1p, 11, 16q, 17p, Kataegis
RB1, gain 11:67M-71M, gain 17:22M-32M, Kataegis
150 SNVs
9647 SNVs
8362 SNVs
APOBEC C>T
APOBEC C>G

Fig. 5.7 The compiled life history for N010985. Early drivers are *NCOR1* and *MAP2K4* (blue square). A range of events cannot be timed and could be early or late (red square). An *RB1* mutation and two gains are subclonal (green square). APOBEC mutational signatures are active throughout the life history of this tumour.

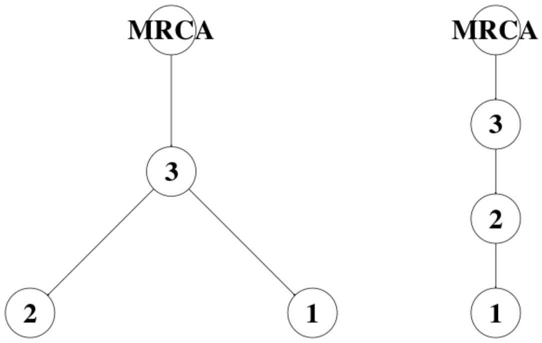


Fig. 5.8 Possible trees reconstructed from the subclonal architecture. The numbers on each node refer to the cluster number, cluster 3 is the clone. The cluster locations provide the option of either a linear or a branching tree. Mutation phasing information did not provide evidence to rule out one of the scenarios.

Chapter 6

Methods for a pan-cancer study of tumour heterogeneity

6.1 Introduction

In the previous chapter I described what can be learned about a single cancer through the application of subclonal architecture and life history methods. In this chapter I introduce methods to scale the analysis up to many tumours. These methods have been applied to tumours in the International Cancer Genome Consortium (ICGC) Pan-cancer Analysis of Whole Genomes (PCAWG) project. Results obtained from the application of these methods are described in the next chapter.

The work described in this chapter is the result of a long standing collaboration that has occupied my whole Ph.D. This chapter therefore contains that is not solely mine; however this additional work is essential to make this into a complete chapter. My main contribution is the procedure that combines copy number profiles from six different methods into a robust profile. It was also my responsibility to deliver the profiles of all PCAWG tumours. The text describing the consensus breakpoints is based on text by Jeff Wintersinger for the Dentro et al. manuscript. Jeff developed the consensus breakpoints component of the consensus copy number workflow.

I have also helped lead the development of the consensus subclonal architecture procedure that combines eleven subclonal architectures into a consensus. I was involved in calibration of the eleven callers by extensive comparisons on real and simulated data, was involved in the development of a simulation data set to validate the approaches and delivered the PCAWG-wide release of the results. The brief methods described in the section about consensus subclonal architecture procedure below however contain methods developed by Kaixian Yu, Maxime Tarabichi and Amit Deschwar, which included here to create a complete story.

The chapter covers a range of methods and contains text that will serve as a basis for methods descriptions in different manuscripts, including Dentro et al. (2017, manuscript in preparation) and Yu et al. (2017, manuscript in preparation). Fig. 6.1 is inspired by a figure made by Jeff Wintersinger for Dentro et al.

6.2 Consensus copy number

ICGC PCAWG relied on a consensus strategy for SNVs, SVs, and indels. Calls made separately by algorithms that are based on different principles were understood to be high-confidence predictions. For copy number calls, we relied on a similar consensus approach, which combined results from six individual copy number callers: ABSOLUTE (Carter et al., 2012), ACEseq (Kleinheinz et al., 2017), Battenberg (Nik-Zainal et al., 2012b), CloneHD (Fischer et al., 2014), JaBbA (manuscript in preparation) and Sclust (manuscript in preparation).

Each copy number caller uses a two-step process, first segmenting the genome into regions assumed to have a constant copy number status, then determining the clonal and subclonal copy number states of each segment. Disagreement amongst copy number callers arises primarily from two factors: differences in genome segmentation, and uncertainty concerning whether a whole-genome duplication (WGD) occurred. Thus, our consensus strategy resolved both factors for each sample, allowing us to determine a consensus copy number state for much of the genome across samples.

6.2.1 Assumptions behind different copy number callers

Copy number callers differ in their implementation choices and underlying assumptions, which contribute to differences in their output (Table 6.1). The copy number callers used in this project come in two different flavours: *Event based*, that fit copy number per segment (ABSOLUTE, Aceseq, Battenberg, and Sclust), and *state based*, that aim to explain the observed data by the least number of copy number states (cloneHD). The former group are more flexible to fit different copy number states, but in principle more sensitive to noise, while the latter group is generally more conservative as it aims to minimise the number of different copy number states.

Methods also utilise different approaches perform the fitting itself. Some callers first fit total copy number to the coverage ratio data and then break that into allele specific calls

(Sclust), others perform a grid-search across a range of purity and ploidy values to jointly fit allele frequencies of heterozygous SNPs and coverage data (ABSOLUTE, Aceseq and Battenberg) or train hidden markov models separately to each type of data (cloneHD). The order of events, and how much trust is put in the allele frequency or coverage data determines how sensitive the method is to noise.

Noise levels, however, will be different between methods due to differing processing steps. Some methods perform phasing of heterozygous SNPs to reduce noise on allele frequency data (ABSOLUTE, Aceseq and Battenberg), some count reads in 1kb bins across the genome to obtain a smoothed out coverage track (ABSOLUTE, Aceseq, cloneHD and Sclust) or use coverage at single SNP positions (Battenberg). Some methods correct coverage data to remove potential wave artifacts for GC content and replication timing (ABSOLUTE and Aceseq), just GC content (Battenberg and cloneHD) or not at all (Sclust). Noise therefore does not only affect methods differently due to the fitting choices, noise itself will be different due to processing choices.

Finally, approaches differ in how subclonal copy number is considered to transform a problem of potentially millions of subclonal copy number profiles per tumour sample into a tractable problem for which a solution can be found. To do so, assumptions are made on the number of copy number states per segment (2 or 3 for ABSOLUTE, Battenberg and Sclust) and how much the separate states can differ (1 copy for Battenberg and Sclust). For the JaBbA caller there currently is neither code nor a manuscript available currently and it is therefore omitted from this comparison.

Copy number methods implementation choices and assumptions

Name	ABSOLUTE	Aceseq	Battenberg	cloneHD	Sclust
Event based	X	X	X		X
State based				X	
Allele counts for heterozygous SNPs	X	X	X	X	X
Binned read counts logR	X	X		X	X
logR from SNP positions			X		
Phasing of SNPs	X	X	X		
Replication timing correction of logR	X	X			
GC content correction of logR	X	X	X	X	
Assume GC artifact same in tumour and normal				X	
Assume raw data shape*				X	
Maximises genome with clonal copy number states	X	X	X	X	X
Purity/ploidy grid search fitting	X	X	X		
Hidden Markov model fitting				X	
Step-wise fitting					X
Estimates subclonal CNA	X	X	X	X	X
Fits subclonal CNA	X		X	X	X
Number of subclonal states allowed	3		2	many	2
Max. differences between subclonal states			1		1

Table 6.1 Copy number callers differ by implementation choices and assumptions, which contribute to differences between callers. The table lists, from top to bottom, basic strategy for calling alterations, how raw data is obtained, how the raw data is adjusted, approaches to fitting a copy number profile and assumptions related to fitting subclonal copy number. * = cloneHD explicitly assumes the coverage data takes on a shape of a (overdispersed) Poisson distribution and allele frequencies the shape of a (overdispersed) binomial.

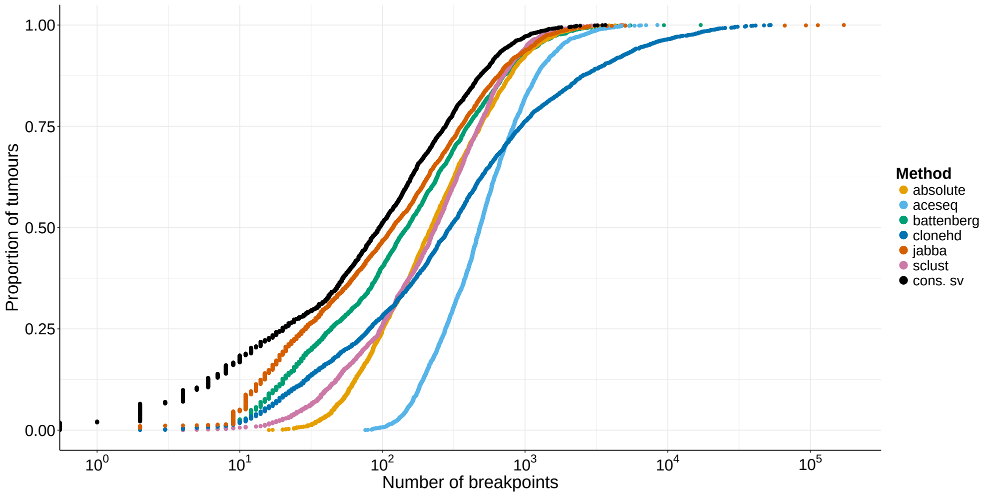


Fig. 6.1 Number of breakpoints for each of the methods used to create the consensus breakpoints (the JaBbA calls are plot for reference) and the consensus structural variants (black). cloneHD and ACEseq call more breakpoints than the other methods, hence their characterisation as *liberal* methods. This figure is inspired by one made by Jeff Wintersinger.

6.2.2 Determining consensus segment breakpoints

Copy number callers segment a sample's genome into regions assumed to have constant copy number. Each segment is bounded by a breakpoint at either end, where breakpoints correspond to a change in copy number. The collection of segments is then used to infer purity and ploidy and fit to copy number states.

We observed substantial disagreement in segmentation between the different algorithms and aimed to develop a consensus set of breakpoints, which the six callers subsequently used to call copy number.

Jeff Wintersingers consensus strategy aims to maximise "true positive" breakpoints at the potential cost of increasing "false negatives". Orthogonal evidence of copy number breakpoints from structural variants was used to quantify the "true positive" and "false negative" rate of our consensus approach. When fitting the copy number profile callers are allowed to merge adjacent segments, therefore the cost of introducing spurious breakpoints was less than that of missing breakpoints.

Copy number methods differed substantially in the number of breakpoints they defined (Fig. 6.1), with some methods calling an order-of-magnitude more breakpoints than others. Broadly speaking, these can be broken into two classes: *liberal* methods (ACEseq and cloneHD) called, on average, a great many more breakpoints than *conservative* methods (ABSOLUTE, Battenberg, JaBbA, and Sclust).

Copy number methods determine breakpoints based on data derived from the sequencing output. Methods use the BAF, logR or coverage for such purposes, sometimes in combination. These three views have their advantages and disadvantages, as was explained in the sections about Battenberg earlier in this thesis, which is a source of the observed differences in segmentation between methods Furthermore, methods differ in how BAF, logR or coverage is obtained from the sequencing data. LogR, for example, can be obtained through windows placed across the genome (overlapping or non-overlapping) or through a set of predefined single base genomic locations, while GC content can be corrected for using different approaches (Benjamini and Speed, 2012; Diskin et al., 2008).

Finally, methods can call the same segment with different breakpoints. Here the implementation matters: the exact location of a breakpoint can correspond to the edge of a window, to a known SNP or to a measured SV breakpoint. That means the called breakpoints for the same segment can be ambiguous, especially in regions with many small segments.

The algorithm that was developed for determining consensus breakpoints draws on the insight that regions between adjacent segments can be used to quantify a method's uncertainty in the exact location of the breakpoint. The segmentation released by each method consists of a set of regions defined by the genomic loci S_i and E_i, with the interval (S_i, E_i) representing a region of constant copy number. On a given chromosome, however, the region (E_{i-1}, S_i) has undefined copy number—the segmentation method inferred that CN status changed at some point within this interval, but cannot pinpoint the location.

The algorithm uses the space between segments and a fixed window size to create leeway on calls from the individual copy number methods and then looks for overlaps between methods to define consensus breakpoints. The algorithm consists of six steps, which are executed for each chromosome separately:

1. For each copy number segmentation method M, take each reported segment (S_i, E_i), and generate an interval spanning the end point of the current segment and the start point of the next, $(E_i - \delta, S_{i+1} + \delta)$. This interval indicates the belief of M that a breakpoint lies somewhere in this interval, permitting the breakpoint to move δ bases upstream or downstream beyond the reported boundaries. Here, we set δ = 50 kb, which we selected after manually comparing the breakpoints generated by a range of δ values to the underlying signal in the data. δ = 50 kb achieved a reasonable balance between false-positive consensus breakpoints (when δ was too large) and false-negative consensus breakpoints (when δ was too small).

2. Compute the intersection of intervals between the methods. Scanning from the start of the chromosome, find the first intersection I_s supported by the threshold methods T.

We defined T to be any combination of at least three of the six copy number methods, or any combination of two of the *conservative* methods (i.e., ABSOLUTE, Battenberg, JaBbA, and Sclust). This avoided calling consensus breakpoints supported by only the two *liberal* methods (ACEseq and cloneHD).

3. For a given intersection I_s: select all reported breakpoints falling within I_s. Score each breakpoint according to the size of the associated gap $G_i = \text{rank}(S_{i+1} - E_i)$, where G_i corresponds to the rank in the empirical cumulative distribution of all gaps generated by the given method. Thus, if a method assigns a large gap between two segments relative to the other segments it generates, its uncertainty in breakpoint placement is understood to be relatively large; conversely, a relatively small gap indicates high certainty. The consensus breakpoint of the intersection is then the breakpoint with the smallest G_i. In the case that two breakpoints in the intersection have the same G_i (which occurs, e.g., because both E_i and $Si+1$ fell in the intersection), arbitrarily prefer end locus to start locus and record this as the single consensus breakpoint. Otherwise, in the rare case that no input start and end loci fall in the intersection, report the upstream-most end of the intersection as the consensus breakpoint. Such cases arise when only the δ bases padding each input segment intersect, meaning that the intersection as a whole is relatively small, and that either end of the intersection can be taken as a reasonable representation of a breakpoint's position.

4. Remove all intervals that contributed to the intersection. Return to step 2. Repeat until no intersections passing the threshold remain on the chromosome.

5. Add PCAWG consensus SVs to the consensus breakpoint set. To do so, find all consensus breakpoints within 100 kb of a consensus SV. Replace the consensus BP with the consensus SV, as the SV presumably represents the same mutational event, but with greater precision concerning position. For any SVs lacking a consensus BP within 100 kb, add the SV as an additional consensus breakpoint.

6. Add breakpoints at centromeres and telomeres as necessary, as copy number status cannot be called across these boundaries. Use the chromosome lengths and centromere start and end locations reported in the hg19 human reference genome. If any centromere start or end lacks a consensus breakpoint within 1 Mb, add an additional consensus BP at that location; if a consensus breakpoint occurs within the centromere, move it to the start or end of the centromere, according to whichever point is closer. Likewise, if no breakpoint occurs within 1 Mb of a chromosome start or end position (representing telomere locations), add an additional breakpoint at the chromosome start or end.

The output of this approach is a list of breakpoints per chromosome. Each pair of adjacent breakpoints corresponds to a consensus segment, for which the six methods produced copy number calls. The next step is to combine those calls into a consensus profile.

6.2.3 Constructing consensus copy number

The consensus copy number profile should contain a call for every consensus segment, if there are enough calls. To do so I first identified 6 ways of extracting agreement between the CNA callers on a single segment (summarised in table 6.2):

(a) All methods agree on a clonal copy number call (both major and minor alleles).

(b) A single method disagrees on the copy number state of a single segment, leaving the call from this method out creates agreement.

(c) A single method disagrees on the ploidy of a sample, leaving the profile out creates agreement.

(d) The strict majority of available methods agree on clonal copy number.

(e) Complete or leave-one-out agreement is achieved by rounding subclonal copy number.

(f) Majority vote is achieved after rounding subclonal copy number.

For each sample, every segment goes through the list starting at a, until agreement is reached. On average, that obtains consensus on 90% of the genome in 86% of samples after reaching level f (Fig. 6.2). The segments that remain without a consensus call go through a second approach that is designed to find a call from a single method to be selected into the consensus profile.

To select a call I first calculate, for every CNA method, what proportion of the consensus profile it agrees with after reaching level f. This allows ranking of the methods, where an excluded profile (due to disagreement on the ploidy) is not included (see filtering below). The following additional levels were then devised:

(g) Take the call from the best method. If there is consensus for the copy number state of one of the alleles we require the best method to agree with it (see rounding below).

(h) Take the call from another method, iterating from the best to the worst performing method.

Consensus copy number levels

Star	Level	Description
3	a	Complete clonal agreement
	b	Clonal agreement of n-1 methods
	c	Clonal agreement excluding ploidy outliers
2	d	Strict majority vote on clonal copy number
	e	Complete agreement after rounding subclonal copy number
	f	Strict majority vote after rounding subclonal copy number
1	g	Best method, one allele with consensus
	h	Best method, no consensus on either allele
	i	No ploidy consensus from panel of experts

Table 6.2 Each consensus copy number segment is assigned a star quality and a confidence level. The level is based on how the consensus is obtained and can be used as a measure of the amount of confidence in the call. The star assignment is aimed to capture the quality of the copy number call in a broad scale that can be understood without details of how the consensus was established.

A special level was added to distinguish between samples where the expert panel did not reach consensus on the ploidy of a sample during a review of all the profiles and raw data for that sample. These copy number profiles were assigned copy number states through the procedure detailed above and each segment received assignment of the level corresponding to how consensus was obtained. But segments were later re-marked as level *i* to denote the extra uncertainty about the assigned copy number states.

I then devised a star rating system that denotes the amount of confidence in each of the calls. Levels *a*, *b* and *c* are the most strict and require all-but-one methods to agree at the least. These segments are therefore assigned *3 stars*. Segments for which a majority of the methods agree on either clonal or rounded clonal copy number are assigned *2 stars* (levels *d*, *e*, *f*). The remaining levels (*g*, *h*, *i*) receive *1 star* to denote the lowest confidence.

6.2.4 Rounding subclonal copy number

Subclonal copy number is reported in three different ways across the 6 methods. ABSOLUTE reports up to 3 different copy number states per segment, of which 1 is termed the ancestral state. Battenberg and Sclust report subclonal copy number as a mixture of two states, while ACEseq returns a single non-integer state (i.e. a mixture). Both cloneHD and JaBbA provided clonal calls only.

Rounded profiles were obtained in the following way:

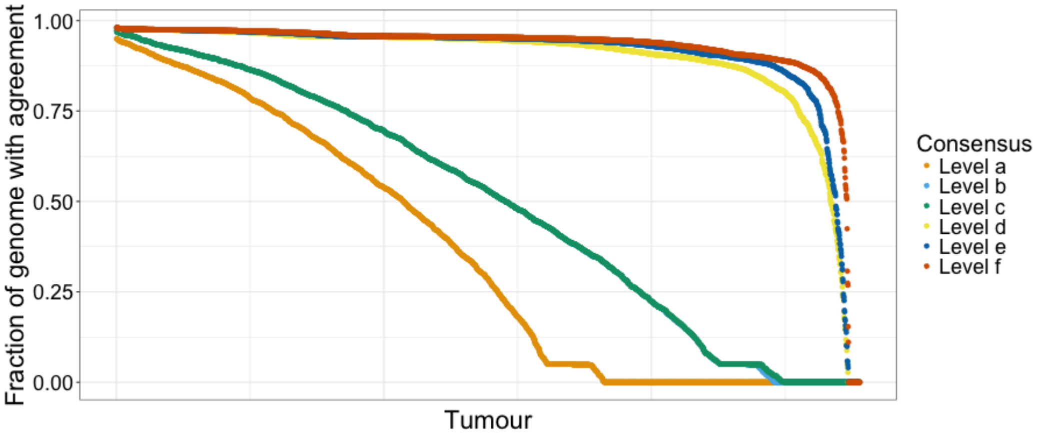


Fig. 6.2 The fraction of the genome for which a consensus can be created increases as more levels are added. Levels correspond to how consensus was obtained. In general levels a-c are the highest confidence, levels d-f medium and levels g-i low (not show in this figure). Agreement on over 90% of the genome is reached for 2406 out of 2778 samples (86%).

- ABSOLUTE: 6 ways, corresponding to rounding both alleles up and down of the ancestral state, the highest CCF state and the lowest CCF state.
- Battenberg and Sclust: 4 ways, the highest CCF state and the lowest CCF state.
- ACEseq: 4 ways, rounding both alleles up and down.

To create a consensus call for a segment I first obtain an inventory of the available copy number states across all roundings and the clonal calls from cloneHD and JaBbA are included. If there is a major/minor allele combination that satisfies the minimum number of methods criterion (either leave-one-out or majority vote) we select that state as the consensus.

If no agreement is reached I attempt to establish consensus by voting for the major and minor allele separately. An allele is accepted if it passes the minimum number of methods threshold. In some cases this leads to consensus on one of the alleles. The state of that allele is saved and fed into levels *g* and *h*, where a call is selected where one of the alleles agrees with the established consensus allele.

6.2.5 Chromosomes X and Y

Fewer methods report on X and Y chromosomes:

- X: ABSOLUTE, Battenberg (females), ACEseq, cloneHD and JaBbA
- Y: ABSOLUTE, ACEseq and JaBbA

The required number of methods to agree for the separate levels are adjusted accordingly.

6.2.6 Panel of experts review

For a range of samples the copy number callers did not unanimously agree on the ploidy. An initial computational analysis developed by Jeff Wintersinger revealed up to 361 profiles where affected. Jeff's approach was developed to automatically adjust a copy number profile in two ways: Halve the ploidy (the effect of removing a whole genome duplication), or subtracting 1 copy from each allele (the effect of removing a normal genome from the profile). A sample name was saved if an adjustment yielded a larger agreement between the methods. An additional 315 tumours did not reach over 20% agreement in a first run of the consensus approach, these tumours were also reviewed.

The samples have been put through a panel of experts review procedure. Initially to understand where the discrepancy lay between the profiles, and later to resolve the differences. The discrepancy cases can be grouped into two categories: Erroneous addition or omission of a genome doubling, or a method specific error scenario. We opted to use the manual approach after initial inspection of 100 samples because fixing the method specific scenarios would have set the process back for months, while there was only a short timeline possible to make the consensus copy number calls available PCAWG-wide.

The expert panel consisted of three core and five alternating members and sat down for four afternoon sessions. Each member prepared a figure per sample with all possibly interesting information. A central figure was used to feed the discussion contained: Copy number profiles from all methods and raw BAF, copy ratio (logR) and multiplicity values from ABSOLUTE. My personal figures contained the Battenberg profile, DPClust reconstruction, multiplicity (Fig. 6.3) and copy ratio (Fig. 6.4); an assembly of figures shown in the QC chapter of this thesis.

During the review a sample was marked as *WGD* or *no_WGD* to reflect a high or low ploidy solution the panel agreed upon. A sample was only marked on unanimous agreement amongst the panel and a maximum of roughly two minutes was maintained to discuss a sample. A sample was marked *unkown* if no agreement could be obtained within the set time. Over time, we observed that methods often show similar behaviour in particular scenarios, which made it easy to determine the disagreeing method as an outlier.

For example, Aceseq showed difficulty calling a copy number profiles with a low purity, which has since been improved. It will not only call a much higher purity, it also shifts the copy number profile up leaving a profile without losses and with empty copy number states. Figures 6.3 and 6.4 show an example, lung squamous cell carcinoma SA305293. The copy number profiles plot it showed Aceseq as calling a higher ploidy by adding an extra copy to every allele on every chromosome. The ABSOLUTE allele specific copy ratio plot clearly showed four separate states, suggesting a genome doubling. Adding a copy to every allele on chromsome 17 would leave one copy of either allele active, and it would remove all other LOH (the Battenberg QC figure shows LOH at chromosome 17p, where *TP53* resides (Fig. 6.3). Finally, the Battenberg raw data figure shows that this is a low purity tumour (separation of purple and blue lines in the bottom plot) and it confirms that, for example, chromosome 4 has the lowest coverage (top) and lowest minor allele frequency (bottom), suggesting that chromosome 4 holds the profile up. There was little discussion about this case, it is clearly whole genome doubled and should have LOH on chromosomes 4 and 17p.

Another example are cases with very heavy, wave-like, coverage artifacts where some methods experienced difficulties. Figures 6.5 and 6.6 show an example, liver cancer SA529774. cloneHD calls a large number of segments that have been fit with an additional gain, creating a much more fragmented genome as compared to the other methods. The Battenberg profile (Fig. 6.5) shows a clean fit (bottom), with a clear clonal mutation cluster (top) and mutation copy number states at integer values (middle), while the raw coverage log ratio plot (Fig. 6.6, top) confirms very noisy signal. Here the panel decided that one method gained a better fit by following the noise, whilst there is no clear evidence of additional gains, which could have lead to the calling of a whole genome doubling.

The above two examples occurred commonly enough for the panel to recognise the scenario and quickly resolve the discrepancy. In some cases however, the panel could not agree within the time limit we set for discussing a sample (which was set to 2 minutes). Figures 6.7 and 6.8 show such an example, pancreatic adenocarcinoma SA533746. ABSOLUTE and cloneHD disagree with Aceseq, Battenberg and Sclust about whether there has been a whole genome doubling. The discussion focussed on whether the allele ratio plot showed 4 distinct states and whether the subclonal segments on chromosome 4 could be fit with a clonal state when the ploidy was doubled (bottom plot Fig. 6.7). I maintain that neither of these segments would become clonal when doubling the ploidy and see no evidence of a clear mutation cluster centered around 50% of tumour cells (top and middle plots Fig. 6.7). However, within the group there was considerable doubt. We did not reach consensus within the time limits and therefore marked it as *unknown*.

Overall, the panel did not manage to agree on the WGD status of 38 cases. All segments of these genomes have been re-marked with confidence level *i* accordingly.

6.2.7 Determining consensus purity

To obtain a consensus purity I extended the calls from the 6 CNA methods with calls from a number of SNV based approaches: CliP, CTPsingle, PhyloWGS, cloneHD (on SNVs) and Ccube. Outlier calls are first removed for CNA and SNV methods separately (see filtering

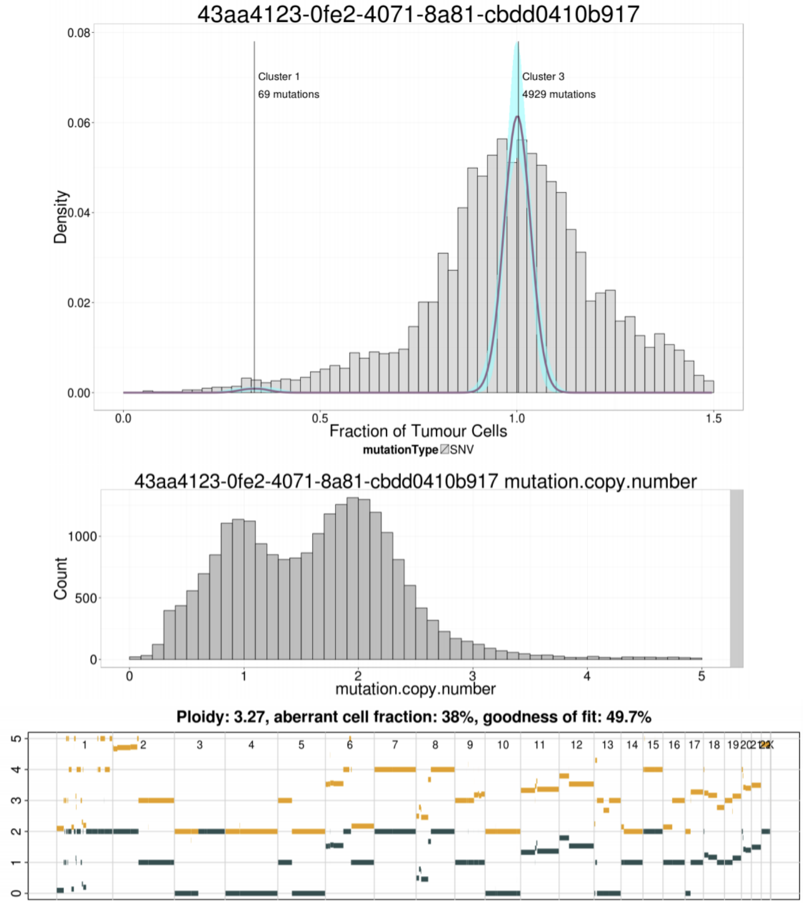


Fig. 6.3 The Battenberg QC figure for sample SA305293. It contains the mutation clustering result (top) with a histogram containing the raw CCF values of SNVs in the background and a density through where weight occurs in the foreground, SNV cluster locations are marked. The middle plot shows a histogram of mutation copy number, the raw estimate of the number of chromosomes an SNV is thought to be carried by. The bottom figure contains the Battenberg copy number profile with total copy number in orange and the minor allele in grey. This sample does not violate any of the QC metrics, and the three views of the same tumour correspond well (i.e. a clear clonal SNV cluster, SNV peaks at integer mutation copy number states 1 and 2 and a lot of allele specific copy number states at 2 that explain the ratio of SNVs on 1 and 2 chromosome copies.)

below). For each sample I establish a density over the combined data. Analogous to taking the mode we select the call that is closest to the highest peak in the density as the consensus.

There is a larger discrepancy in purity calls from CNA methods on samples with few copy number alterations. I therefore calculate the density over the calls from SNV based methods only for samples where less than 8% of the genome is altered by CNAs.

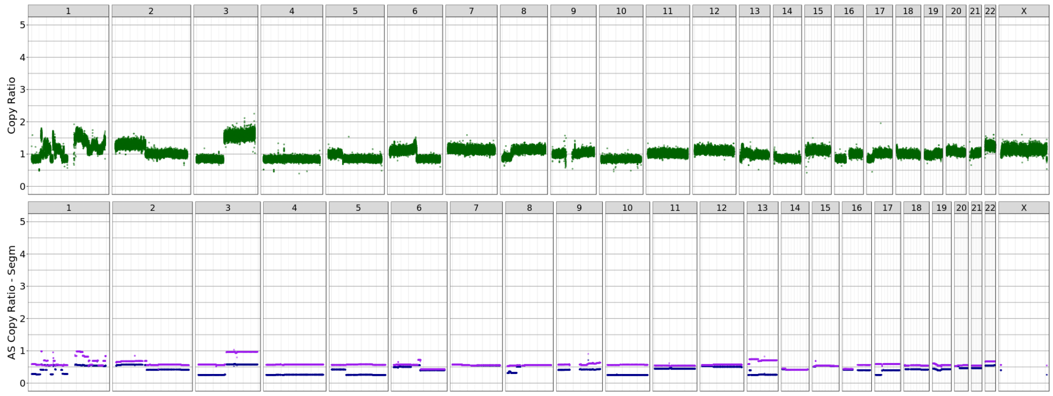


Fig. 6.4 The raw data that Battenberg produces for sample SA305293. The top figure shows the copy ratio (the R in logR), the bottom figure shows the allele specific copy ratio (R multiplied by the BAF). The top figure shows that the coverage of this sample is clean (SNP dots fall in clear straight green lines) and it shows which genomic regions correspond to the lowest relative coverage in this sample, which can be used to identify which segments are of the lowest total copy number. The bottom figure contains information on how the total copy number is relatively split into major (purple) and minor allele (blue). Focussing on chromosome 4, the figure shows it has the lowest relative coverage and the BAF of the alleles is split. That means this segment is a candidate for a 1+0 or 2+0 fit. Such reasoning helps to read a copy number profile from the raw data.

Finally, the median absolute deviation of the purity calls on a sample is calculated to capture the amount of agreement between the methods and is used as a measure of confidence.

6.2.8 Filtering

After the expert panel review of ploidy-uncertain cases, a rough reference ploidy can be obtained for almost all samples. The methods either all agreed on large portions of the genome and therefore, by extension, on the ploidy or certain ploidy calls were overruled by the expert panel.

With the accepted ploidies in hand it became possible to calculate a rough reference ploidy that serves to overrule calls from individual CNA callers. This is necessary because copy number callers can return a different ploidy in different runs and I required a way to automatically accept or reject a ploidy call. A method is allowed to deviate from the reference ploidy by a relative amount to allow for larger discrepancies on higher ploidy calls. I set the threshold at 0.25 times the reference ploidy. If a profile differed by more than this threshold, it was automatically overruled and excluded from the procedure for both consensus copy number and purity creation.

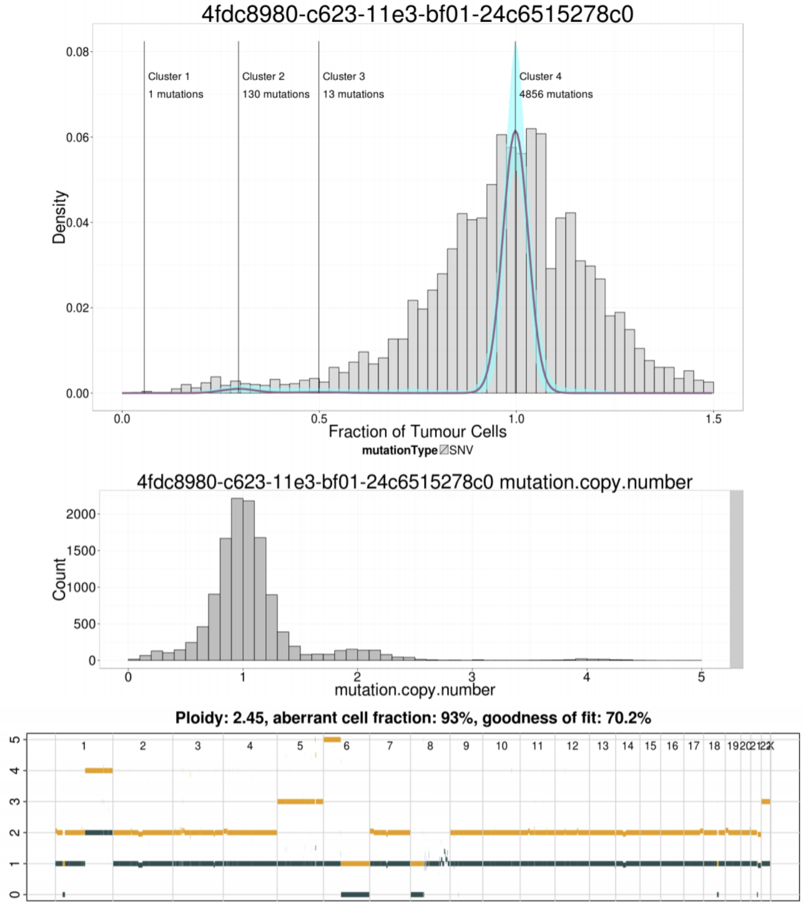


Fig. 6.5 The DPClust and Battenberg results show a clean CCF (top) and mutation copy number (middle) space which corresponds to a large clonal peak, consisting of mostly SNVs on one chromosome copy and a much smaller number of SNVs on two copies. The copy number profile (bottom) shows a relatively clean fit, with quite a few small segments that are fit with a near clonal copy number state. It shows Battenberg accounts for the noise in this sample.

Filtering on the purity calls was performed to remove outliers. A purity call was filtered out if it differed from all other non-ploidy-overruled purity calls by more than 0.2. This method was applied separately to SNV based purity values.

Finally, I also excluded calls on complex regions chromosomes 13p, 14p, 15p and 21p from some methods as they consistently appeared as losses across the entire data set.

6.3 Consensus subclonal architecture

With consensus copy number profiles established we sought to construct consensus subclonal architectures. A similar philosophy to the consensus copy number is applied: combine

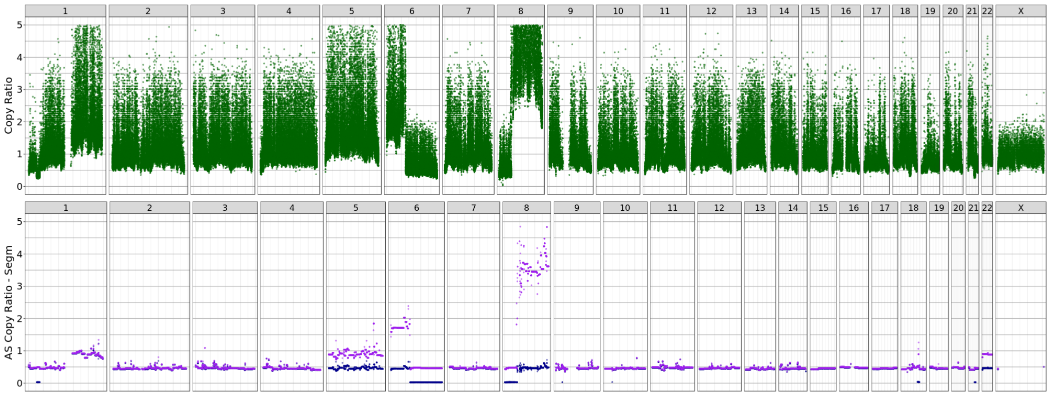


Fig. 6.6 The raw Battenberg data show what is causing problems when fitting a copy number profile for sample SA529774. The relative coverage ratio (top) is very messy and affects the allele ratio space by adding noise to the dots (bottom). This effect is most likely due to either the tumour or the normal containing strong coverage bias across the genome.

the output from multiple methods, which allows for recovery from a mistake by a single method. We have developed three orthogonal approaches that combine output from eleven individual callers (Bayclone (Sengupta et al., 2015), Ccube (manuscript in preparation), CliP (manuscript in preparation), CloneHD (Fischer et al., 2014), CTPsingle (Donmez et al., 2016), DPClust (Nik-Zainal et al., 2012b), Phylogic (Landau et al., 2013), PhyloWGS (Deshwar et al., 2015), PyClone (Roth et al., 2014), Sclust (Cun et al., 2018) and SVclone (Cmero et al., 2017)) into a consensus. We show through simulated data that the three approaches are equivalent and are consistently ranked amongst the top performing methods.

6.3.1 Three consensus approaches

Weighted Median (WeMe, by Amit Deshwar) - takes the cluster location and sizes reported by the individual methods and combines the output by minimising the earth movers distance (EMD) to the median clustering. Where the median clustering is defined as the clustering minimises its EMD to all input clusterings. To constrain the number of clusters it then performs a grid search over cluster location and size parameters to fit the median number of clusters obtained from the provided input.

Cluster ID Consensus Clustering (CICC, by Maxime Tarabichi) - uses groups of SNVs that are consistently assigned to the same cluster across methods. It first creates a vector for each mutation with as contents the cluster to which the eleven methods have assigned the mutation. Then, for each pair of vectors, a distance is calculated, resulting in a distance matrix. Hierarchical clustering is performed to cluster the mutations. The resulting tree is then cut to the median number of clusters that the eleven methods reported, rounding

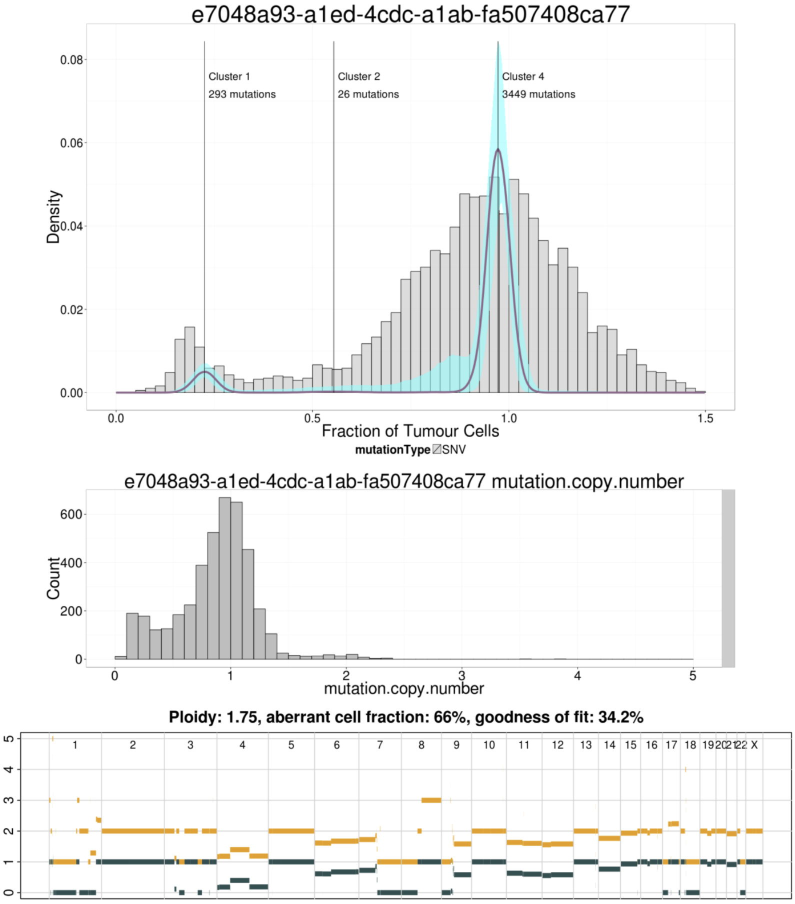


Fig. 6.7 During the review it was suggested that segments on chromosome 4 could be a separate clonal state when sample SA533746 is fit with a whole genome doubling. In that case one would expect these segments to be fit with subclonal copy number in exactly 50% of tumour cells if the profile is fit without a doubling. This figure shows that subclonal segments on chromosome 4 are not exactly within two clonal copy number states, which supports the theory that this sample has not had a whole genome duplication.

up when that number is not an integer, which results in a number of consensus clusters with their mutation assignments. Cluster locations are then determined by first calculating a consensus CCF estimate through the equations 2.14 and 2.16 in chapter 2, and then per cluster taking the median CCF of the SNVs assigned to the cluster.

Sparse Clustering for Subclonal Reconstruction (CSR, by Kaixian Yu) - starts from a mutation-to-mutation co-clustering matrix, in which cell (i, j) contains the probability that mutation i belongs to the same cluster as mutation j. The input matrix M is deconvolved using dictionary learning into a dictionary matrix D and a sparse code matrix A. A contains a sparse representation of the structure in M by its most essential components, which makes the mutations better separable. k-means clustering is then applied to A with k set to the

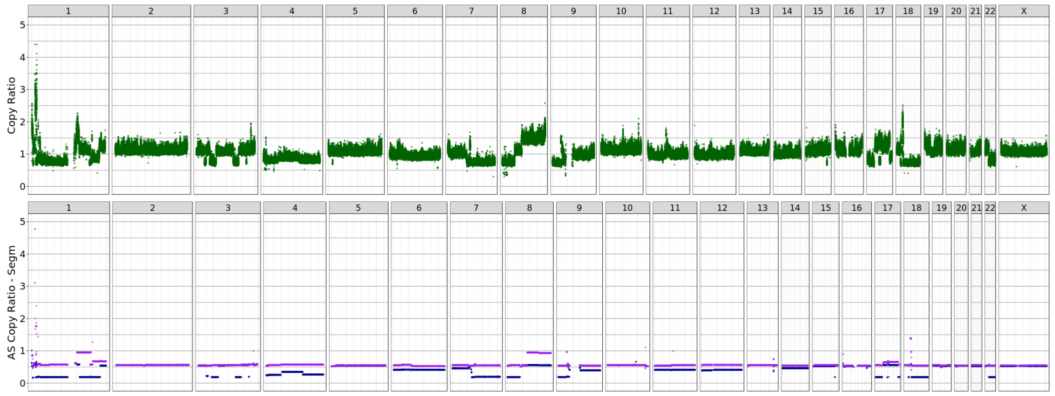


Fig. 6.8 The raw Battenberg data for sample SA533746 shows that the coverage ratio is clean (top).

median number of clusters called by the eleven methods. Cluster locations are obtained by first calculating the median cellular prevalence (CP) of each SNV across the eleven methods and then taking the average per cluster.

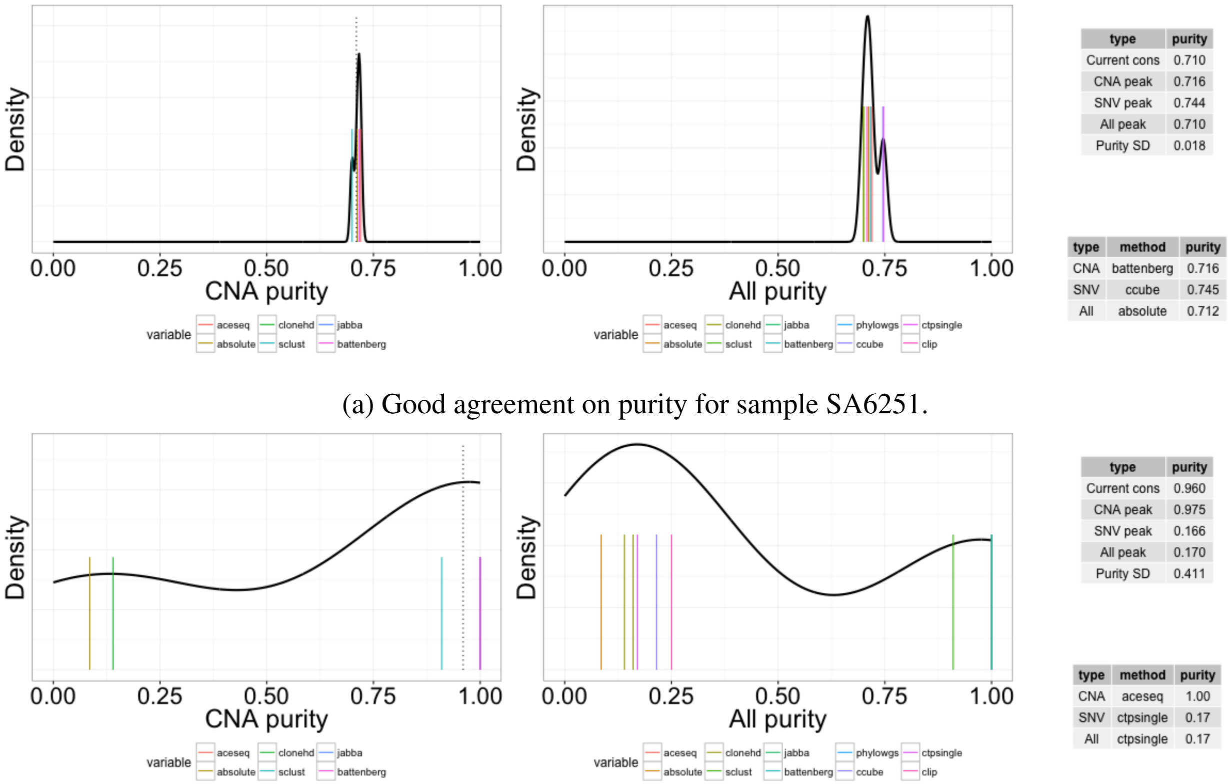
6.3.2 Performance comparison

To asses the performance of the consensus methods we compare the three consensus approaches with the eleven input methods and three methods that produce random solutions, that do not serve as input to the consensus, on the simulated data that was introduced in section 3.3 as part of the validation chapter.

In the validation chapter I also introduced three metrics that can be used to compare the results of a method to the truth or to another method: number of subclones, fraction of clonal mutations and the root-mean-squared-error (RMSE) between mutation assignments. The three metrics can be combined into a single measure by calculating the rank sum across the metrics for each method. The rank sums are normalised for whether the ranking is increasing or decreasing, resulting sum values between a best case 3/17 and worst case 3*17 as there are 17 methods.

Figure 6.10 shows the rank sums of all samples across the 17 methods when they are compared to the truth, with a black bar indicating the median rank of the results of that method and the red bar denoting the mean. A dashed line is drawn that corresponds to the lowest median rank across the methods.

In general there are a number of methods that show a very similar performance. Phylogic, DPClust, CCube, PyClone, PhyloWGS and all three consensus methods have similar median ranks, with cloneHD and SVclone and CTPsingle not far off. The first group is followed by a second group that includes Sclust, CliP, BayClone and the informed method from



type	purity
Current cons	0.710
CNA peak	0.716
SNV peak	0.744
All peak	0.710
Purity SD	0.018

type	method	purity
CNA	battenberg	0.716
SNV	ccube	0.745
All	absolute	0.712

(a) Good agreement on purity for sample SA6251.

type	purity
Current cons	0.960
CNA peak	0.975
SNV peak	0.166
All peak	0.170
Purity SD	0.411

type	method	purity
CNA	aceseq	1.00
SNV	ctpsingle	0.17
All	ctpsingle	0.17

(b) Methods disagree on the purity for sample SA528952.

Fig. 6.9 Consensus purity establishment for two example cases. Each figure contains the purity calls from the CNA methods on the left and all methods on the right. The top table on the far right contains information about where the peaks in density are for CNA methods only, SNV methods only and all methods. The bottom contains information about which method agrees best with the consensus. The dashed line in the CNA purity figure contains the median purity from CNA methods (labelled as current consensus in the table). In scenario (**a**) there is a good agreement between the methods and a good agreement between CNA and SNV purities. This represents most tumours. I therefore opted to establish a consensus based on the overall density peak, which amounts to the mode across all CNA and SNV methods. The purity value that is closest to the peak location is then chosen as the consensus. Scenario (**b**) can occur when a copy number profile contains no clonal alterations. Not all CNA methods are capable of handling these cases, which results in disagreement and possibly an incorrect call. Inclusion of the SNV data leads shows that SNV methods agree with the lower purity value. To remove the uncertainty that CNA methods introduce in this scenario I establish consensus by evaluating the SNV methods only.

RandomClone. Finally, the third group contains the remaining two RandomClone methods that perform considerably worse on this dataset.

These results show that the three consensus approaches have a performance comparable to the best individual methods. It also shows that all eleven methods comfortably outperform a

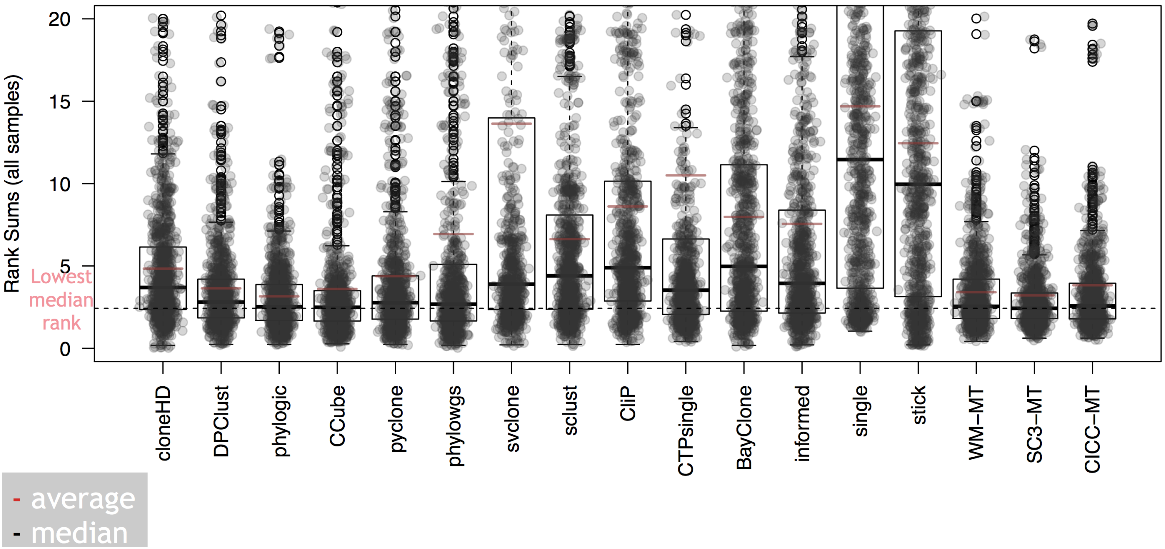


Fig. 6.10 Ranksum comparison of subclonal reconstructions across the methods. The three consensus approaches systematically perform comparable to the best individual method. There appear to be two groups of individual callers: Those that perform comparable to the consensus or are close to it, and those that perform similar to the RandomClone informed method. All methods comfortably outperform simple random approaches. These findings show that the consensus are not sensitive to an outlier solution or a poorly performing caller.

simple random approach (RandomClone stick) and assigning all mutations to a single cluster (RandomClone single). However, not all methods outperform a slightly more sophisticated random method (RandomClone informed) and these methods have been included into the consensus approaches. That the consensus methods show a comparable performance to the best methods shows that the consensus is invariant to the inclusion of a poor solution when it is constructed.

Figure 6.11 contains a pairwise comparison between the 17 methods. Each square contains a matrix with a comparison between a pair of methods (i, j) that is sorted by number of subclones (columns) and number of reads per chromosome copy (rows). A cell in the matrix is coloured blue if the *column* method has a higher ranking than the *row* method on a sample, it's coloured red if the *row* method performs better, while it's white if there is not much difference. These figures allow for exploration of performance in relation to increasing number of subclones and increasing number of reads per chromosome copy.

Similar to fig. 6.10, there appear to be three groups of methods, but their members are slightly different. Phylogic, DPClust, CCube, PyClone and PhyWGS form a block of white or lightly coloured squares in relation to each other, meaning these methods correlate very well. The second group is formed of SVclone, Sclust, CliP, CTPsingle, BayClone and the informed RandomClone method. The third group contains the final two RandomClone methods that contain nearly completely blue squares on their rows indicating they are nearly always outperformed by the callers.

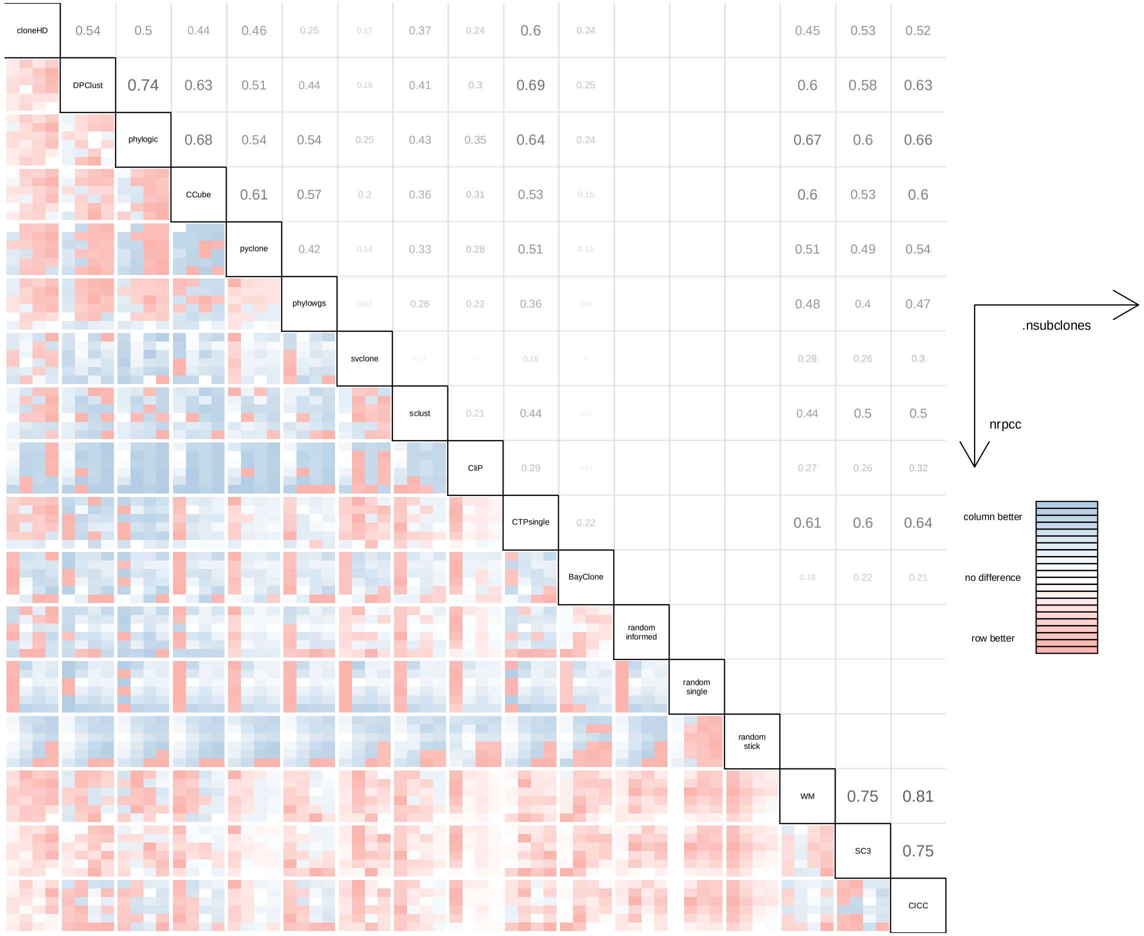


Fig. 6.11 Pairwise comparison of methods performance on the simulation data set. The comparison shows that the three consensus methods (bottom right) agree very well with each other, even though the individual methods only sporadically correlate well. The consensus methods also systematically perform better than any individual caller (including the random methods). Although in some scenarios an individual method can perform slightly better.

The consensus methods again show a very similar performance to each other and all three perform comparable to the methods in the first group. It does however appear that CICC performs better than CSR on low numbers of subclones. The squares comparing WM to CSR and CICC appear more pale, indicating that their solutions are very similar in general.

Combined, these findings show that a consensus approach is robust against an outlier solution. And because WM appears to correspond best to both other consensus methods we opted to use that for further analysis of which results are reported in the next chapter.

6.4 Purity, ploidy and sequencing coverage determine ability to detect subclones

Subclonal reconstruction depends on the ability to call subclonal SNVs in a sequenced tumour. The number of reads required to call a SNV depends on the properties of the SNV caller, and on the sequencing error rate distribution. As a rough rule of thumb, three mutant reads are typically required to detect an SNV, and mutations present in small fractions of tumour cells may be missed. The coverage at which the tumour was sequenced, the admixture of tumour and normal cells in the sequencing sample and the total amount of DNA from each tumour cell all contribute to the ability to detect clonal and subclonal mutations. The following formula combines these three factors into a power metric

$$p_s = c_s \frac{\rho}{\rho \psi_t + (1-\rho)\psi_n} \tag{6.1}$$

Here, c_s is the sequencing coverage of the tumour sample, ρ is the tumour purity, and ψ_t and ψ_n are the ploidy of the tumour and normal cells respectively (the amount of genomic material per cell, expressed in number of haploid genome copies). p_s is equivalent to the number of reads per chromosome copy and represents the expected number of reads reporting a clonal SNV. If, for example, p_s equals 10 and an SNV can be detected when there are three mutant reads, then (as an approximation) mutations present a subclone taking up 30% of tumour cells can be detected.

Figure 6.12 shows that this theoretical bound roughly corresponds to what is possible in real data. Each dot represents a tumour in the ICGC PCAWG data set. The left hand plot shows that the ability to detect subclones goes up as the number of reads per chromosome copy increases, with tumours without subclones, with 1, 2 and more subclones showing clearly visible ’bands’. The right hand plot shows the minimum CCF of the detected mutation clusters, plot against the number of reads per chromosome copy. The dashed line represents

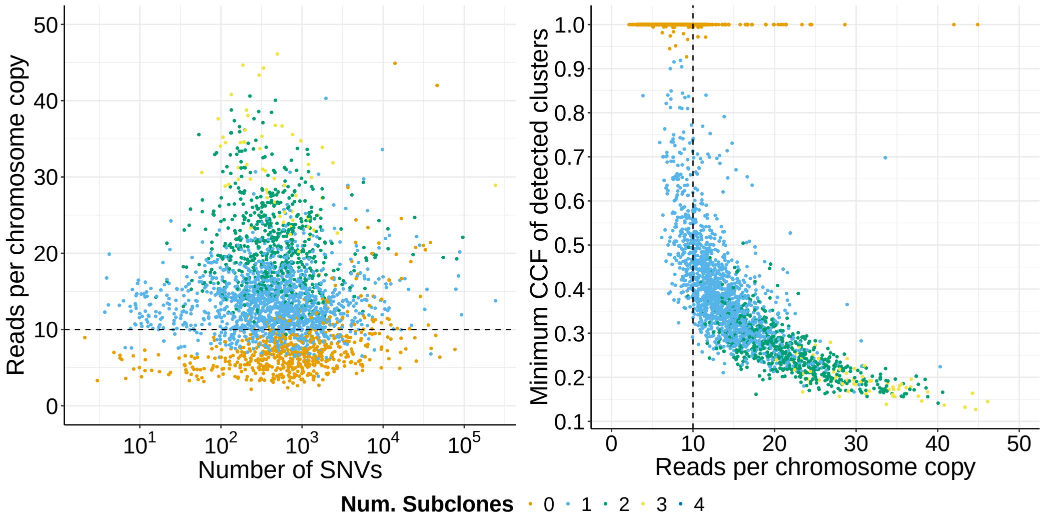


Fig. 6.12 The number of reads per chromosome copy, calculated by combining tumour purity, ploidy and sequencing coverage, determines the power to detect subclones (left). 10 reads per chromosome copy allows for detection of a subclone at 30% of tumour cells.

our theoretical bound of 10. The figure indicates that at 10 reads per chromosome copy we almost have the power to detect a subclone at 30% of tumour cells.

6.5 Correcting for the winner's curse

Figure 6.13 shows the CCF space for the same subclonal architecture and copy number profile, simulated four times with different coverage values. As coverage, and therefore the number of reads per chromosome copy, increases the light green vertical lines move closer to the black vertical lines, which indicates that the mean CCF of mutations visible in the plot moves closer to the true CCF of the clusters from which they were generated. This shifting of the weight of the clusters is caused by the winner's curse due to the clusters being represented by the mutations that by chance made it over the threshold of minimum number of supporting reads required.

As the weight of the clusters is shifted, subclonal reconstruction algorithms will also infer a shifted cluster location (if the clusters can be disentangled at all, see the top left plot in Fig 6.13). To obtain the true cluster locations and their sizes Amit Deshwar and Ignaty Leshchiner developed approaches to correct for this winner's curse effect. One approach simulates additional mutations and iteratively adjusts the cluster location depending on how much the cluster location changes. The process converges when the true cluster location has been obtained, with a corresponding size estimate from the simulated mutations. A second

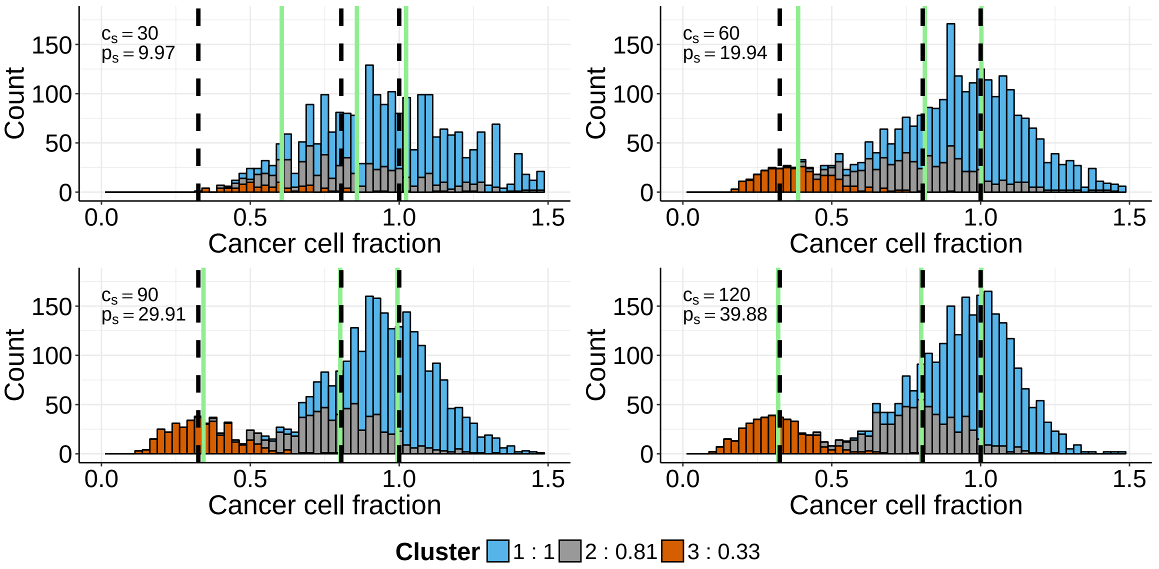


Fig. 6.13 The number of reads per chromosome copy determines the power to detect subclones. This figure shows the same tumour simulated four times with the same purity and ploidy, with a range of coverage values: 30x, 60x, 90x and 120x. The dashed black lines represent the true cluster locations, while the light green lines represent the mean CCF per cluster of mutations shown. The histogram clearly shows the effect of increasing the reads per chromosome copy: the left hand tail extends further towards 0. As the power goes up the three clusters are more fully represented, resulting in the light green bars (mean CCF of mutations present) moving towards the true cluster locations (black dashed lines). This shifting of the weight of the clusters is called the winner's curse as clusters are only represented by the mutations that by chance are supported by enough reads to be called.

approach uses moment matching to match the observed distribution to a library of available shapes and picks the shape that best corresponds to the observed CCF distribution. In the next chapter we correct the ICGC PCAWG data set for the winner's curse effect by taking the average adjustment between the two methods.

Chapter 7

A pan-cancer overview of tumour heterogeneity

7.1 Introduction

In the previous chapters I have introduced methods to analyse copy number and the subclonal architecture of cancers. Individual methods were introduced as well as approaches to construct a consensus copy number profile and a consensus subclonal architecture. These approaches were applied to the 2,778 cancer genomes contained within the ICGC PCAWG project (a description of how the consensus subclonal architecture was obtained can be found in the next section). This chapter describes the pan-cancer landscape of intra-tumour heterogeneity and evolution, as it emerges from the consensus results. The chapter, for the most part, covers the results that will be reported in Dentro et al. (manuscript in preparation) and also contains a high-level overview of results described in Gerstung et al. (2017), which is attached to this thesis as Appendix A. The results reported in this chapter are the culmination of my Ph.D. and are the result of a process in which I have been deeply involved over the last 3.5 years. These results also represent the outcome of a long standing collaboration between members of the PCAWG Evolution and Heterogeneity working group, without whom this project could not have succeeded. The figures in this chapter will appear in Dentro et al. Figs. 7.1 and 7.2 have been created by Kerstin Haase and Figs. 7.7 and 7.8 are by Maxime Tarabichi. All figures are used with permission. Fig. 7.4 is inspired by the driver figure made by Ignaty Leshchiner for the Dentro et al. manuscript.

7.2 Methods

We set out to obtain a robust consensus subclonal architecture for every tumour based on the consensus approaches introduced previously. We first applied the consensus copy number procedure to combine profiles from the six different copy number callers (ABSOLUTE, Carter et al. (2012); ACEseq, Kleinheinz et al. (2017); Battenberg, Nik-Zainal et al. (2012a); cloneHD, Fischer et al. (2014), Sclust Cun et al. (2018) and JaBbA (manuscript in preparation)) into a robust, high confidence consensus. Every copy number profile consists of a series of segments with each an assigned confidence level. Not every segment in every genome is of high confidence and an incorrect copy number call for a single segment could cause a subclonal architecture method to call a spurious mutation cluster, as the CCF values of mutations on that copy number segment would be incorrectly calculated from the VAF. I therefore created a subset of high confidence segments by ordering the segments of each tumour by their confidence level and select segments from the top until at least 75% of the genome was covered. The 11 subclonal architecture callers were restricted to use copy number and SNVs in the selected regions only.

The 11 subclonal architecture callers (BayClone-C, Sengupta et al. (2015); cloneHD, Fischer et al. (2014); CTPSingle, Donmez et al. (2016); DPClust, Bolli et al. (2014), Phylogic, Landau et al. (2013); PhyloWGS, Deshwar et al. (2015); PyClone, Roth et al. (2014); SVclone, Cmero et al. (2017), Ccube (manuscript in preparation), CliP (manuscript in preparation) and Sclust, Cun et al. (2018)) produced three key features to describe every tumour in the data set: The number of mutation clusters identified, properties of those clusters (the estimated number of mutations and the proportion of tumour cells that each cluster represents) and mutation assignments (either probabilistic or hard assignments). The three consensus subclonal architecture procedures were applied to produce a consensus set of mutation clusters described by a location (proportion of tumour cells estimate) and size (number of SNVs that the cluster contains).

I then applied the MutationTimer pipeline (Gerstung et al., 2017) to assign all available consensus SNVs (including the SNVs that were previously excluded when selecting highly confident copy number segments) and all indels and SVs for which allele frequencies were available. MutationTimer assumes each mutation cluster can be modelled by a beta-binomial and calculates probabilities for each mutation belonging to each cluster whilst also taking into account the size of the mutation clusters. It produced the final consensus subclonal architecture with the aforementioned key features, while also performing timing of mutations relative to gains to classify mutations in *clonal early*, *clonal not specified*, *clonal late* and *subclonal*. MutationTimer does this by evaluating the multiplicity state of a mutation and the copy number of the segment on which the mutation resides. If the mutation is on a gained

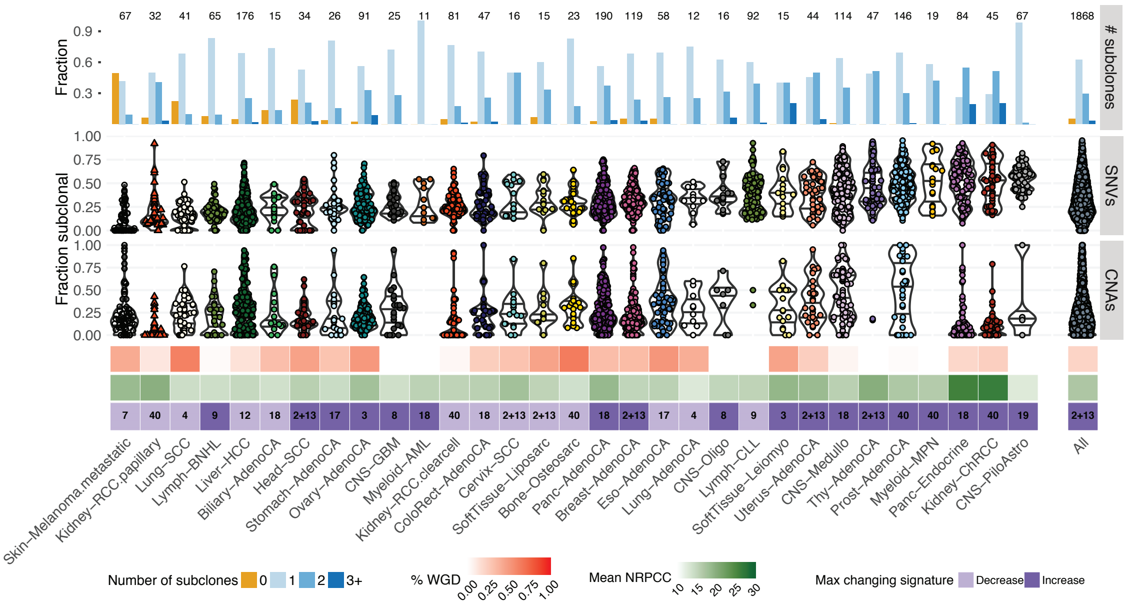


Fig. 7.1 A pan-cancer overview of intra-tumour heterogeneity. The top bar shows the total number of tumours per cancer type and the proportion of tumours where we identify zero (orange), one, two or three+ subclones (shades of blue). Below are the proportion of subclonal SNVs and CNAs, where the CNAs represent tumours with at least 5 whole chromosome arm CNA events. Whole genome duplication rate is show in boxes coloured red (high duplication rate) to white (low) and average reads per chromosome copy (introduced in section 6.4) is shown in shades of green. The most changing mutational signature is shown in purple, where the number refers to the COSMIC signature.

chromosome and has a multiplicity greater than one it is *clonal early*, if the mutation is on a gained chromosome and has a multiplicity equal to one it is *clonal late*, if the mutation falls in a region that has normal diploid copy number or is a loss it will classify clonal mutations as *clonal not specified* and if the cluster with the highest assignment probabality is subclonal, then the mutation is assigned *subclonal*.

The results reported in this chapter are from the WeMe consensus method, but due to the high similarity between the consensus subclonal reconstruction methods we could have chosen CSR or CICC and have shown the same basic results (Yu et al. 2017, manuscript in preparation).

7.3 Nearly all primary tumours contain detectable subclones

Figure 7.1 shows the pan-cancer overview of intra-tumour heterogeneity (ITH) that our analysis reveals. The figure contains all tumours with a number of reads per chromosome copy of 10 or more (which allows us to find a subclone at 30% of tumour cells or higher) to exclude tumours where not enough subclonal signal is obtained due to a combination of purity, ploidy and sequencing coverage. We selected all primary tumours and reduced multi-sample cases to their preferred tumour (a label that is provided by the PCAWG consortium). As there are only 2 primary melanoma tumours available in PCAWG, we instead included melanoma metastasis. The remaining metastasis and relapse tumours are discussed in the next section. The figure shows fractions of subclonal SNVs and CNAs. The fraction of subclonal CNAs indicates the number of arm-level subclonal CNAs over the total number of arm-level CNAs per tumour. A tumour is only included in the CNA plot if it's profile contains at least 5 arm-level events in total. Cancer types are sorted by the median proportion of subclonal SNVs.

The overview across 36 histologically distinct cancer types reveals that 96.7% of the 1,801 primary tumours contain at least one subclone. Patterns of ITH differ markedly between types of cancer: Prostate, uterus and esophageal adenocarcinomas show high proportions of both subclonal SNVs and CNAs. Kidney chromophobe and pancreatic endocrine tumours also show high proportions of subclonal SNVs, but differ from the previous group by containing few subclonal CNAs. On the other hand, hepatocellular carcinomas and squamous cell carcinomas of head-and-neck and lung contain low proportions of subclonal SNVs, but high proportions of subclonal CNAs. Finally, in osteosarcomas we find a high proportion of subclonal CNAs and varying degrees of subclonal SNVs. These findings suggest that tumour types exhibit their own, distinct, evolutionary narratives.

7.4 Metastatic melanomas are often clonal

In stark contrast to the high proportion of primary tumours with at least one subclone, we observe that over half of metastatic melanomas are clonal (Fig. 7.1). A comparison to metastasis of other cancer types available in this data set suggests that this might be a unique property (Fig. 7.2, left), although, the other cancer types are represented by a low number of cases and the observation would need to be verified in a larger cohort. There are only two breast and ten prostate metastasis cases available (the prostate tumours are from the multi-sample study Gundem et al. (2015)).

In contrast, melanoma relapse tumours are as heterogeneous as relapse cases from other types of cancer (Fig. 7.2, right). These findings may highlight properties of metastatic melanomas, for example that these metastasis belong to a group of rapidly developing melanoma tumours (Liu et al., 2006).

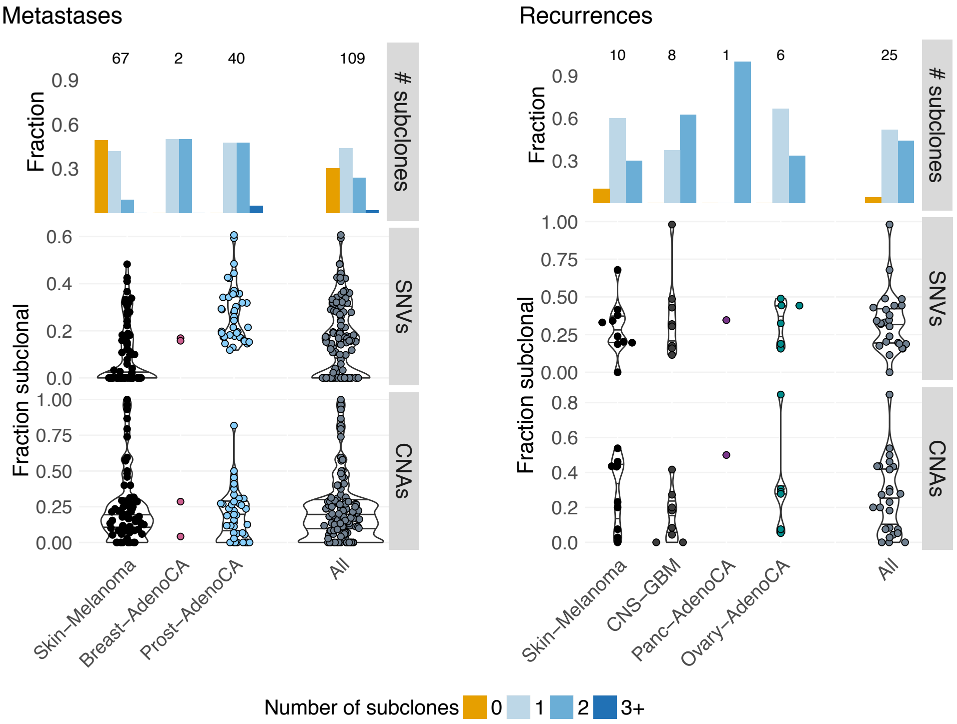


Fig. 7.2 The proportions of subclones, fraction of subclonal SNVs and CNAs found in metastasis and relapse tumours.

7.5 Subclonal driver mutations in known cancer genes

I used the catalogue of driver mutations identified in the PCAWG cohort by Sabarinathan et al. (2017) to obtain an overview of clonal and subclonal drivers in the PCAWG cohort. A number of filters have been applied to obtain the pan-cancer picture of subclonal drivers (Fig. 7.3). I excluded CNA drivers, as they are not assigned in our consensus subclonal architectures, and also excluded SV assignments, as their CCF values tend to be of quite variable quality. Furthermore, tumours with low reads per chromosome copy have been removed, multi-sample cases have been reduced to their PCAWG preferred sample and relapse and metastasis cases have been filtered out (apart from melanomas). After filtering there are 4,152 identified drivers remaining, spanning 362 different genes. 1,423 of 1,865 (76%) tumours contain an identified driver, but only 24% of tumours and 20% of subclones contain at least one subclonal driver.

Figure 7.4 provides a pan-cancer overview of the top 30 genes with subclonal drivers, identified by summing their probability of being subclonal. Each square is sized depending on the proportion of tumours of that cancer type containing a driver in that gene and the

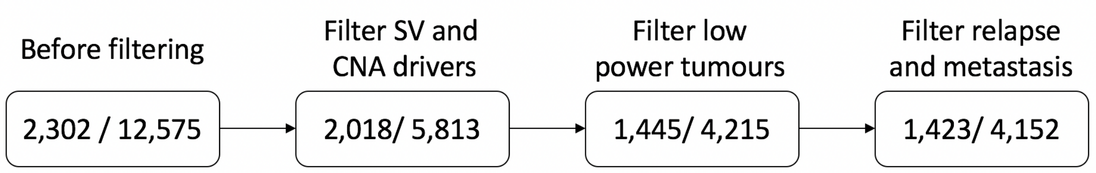
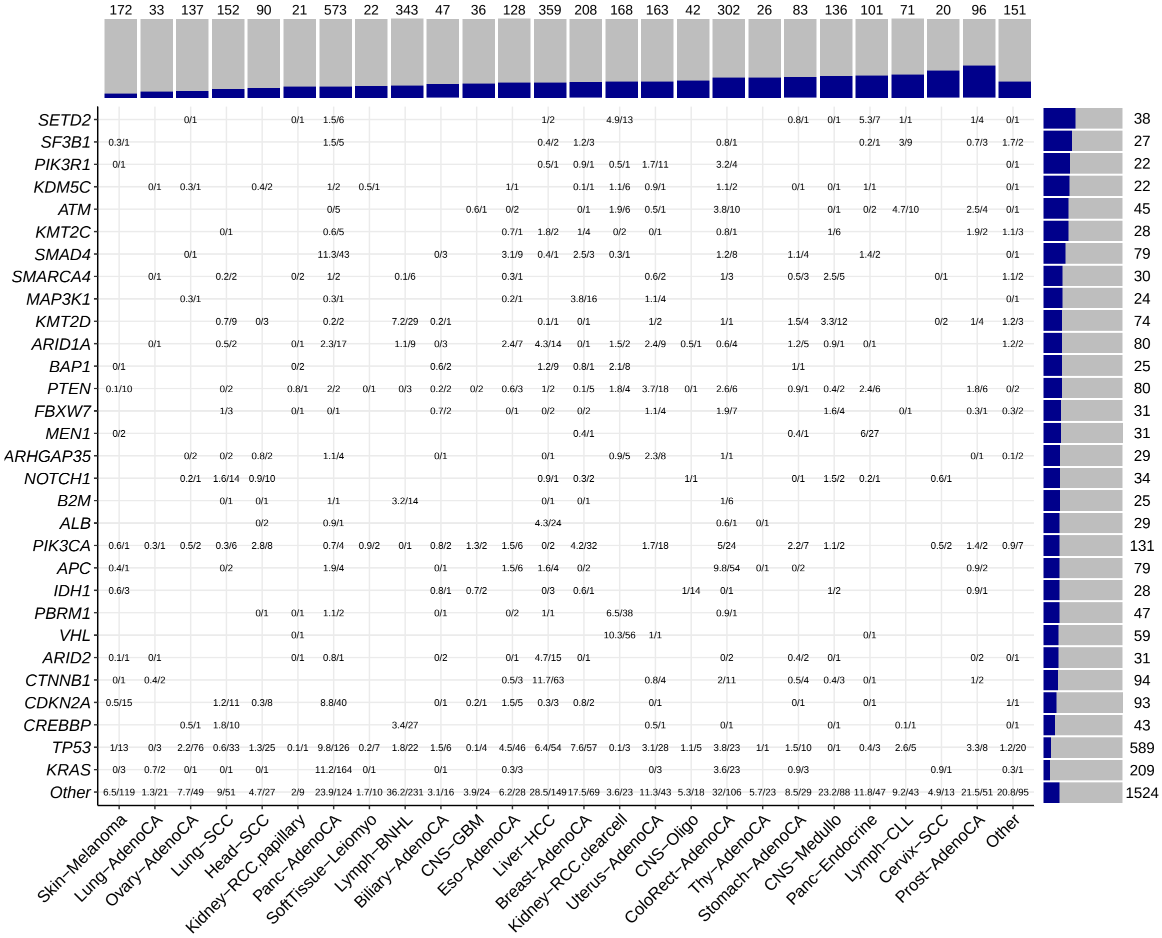


Fig. 7.3 Flow diagram showing the filters applied to obtain the driver overview. Each square shows two values, separated by a divider: First the number of unique tumours that are left in the data set after a filter has been applied and second the number of driver contained within those samples.

square is coloured depending on the proportion of tumours in which the driver is subclonal (again identified by summing probabilities of being clonal and subclonal), where darkblue means high proportion subclonal and lightblue means a high proportion clonal. Figure 7.5 shows the same data that is provided in each square in Fig. 7.4, but instead of showing the proportion of subclonal drivers, they are provided as counts to show the total number of tumours that support each square. Bars have been greyed out when they represent fewer than six tumours.

TP53, *TERT* and *VHL* show large, light squares for a number of cancer types, which means that these drivers are often activated early and are thus frequently observed as clonal. The figure does not contain any large, dark squares, which means that no gene is primarily identified as a subclonal driver in large numbers of tumours. The presence of small dark squares shows that driver mutations in known cancer genes can appear late during tumour evolution, but only in small proportions of tumours. This suggests that early drivers may be constrained to a select number of genes, while the spectrum of drivers becomes more diverse as tumours evolve further.



	Skin–Melanoma	Lung–AdenoCA	Ovary–AdenoCA	Lung–SCC	Head–SCC	Kidney–RCC.papillary	Panc–AdenoCA	SoftTissue–Leiomyo	Lymph–BNHL	Biliary–AdenoCA	CNS–GBM	Eso–AdenoCA	Liver–HCC	Breast–AdenoCA	Kidney–RCC.clearcell	Uterus–AdenoCA	CNS–Oligo	ColoRect–AdenoCA	Thy–AdenoCA	Stomach–AdenoCA	CNS–Medullo	Panc–Endocrine	Lymph–CLL	Cervix–SCC	Prost–AdenoCA	Other	
	172	33	137	152	90	21	573	22	343	47	36	128	359	208	168	163	42	302	26	83	136	101	71	20	96	151	
SETD2			0/1			0/1	1.5/6						1/2		4.9/13					0.8/1	0/1	5.3/7	1/1		1/4	0/1	38
SF3B1	0.3/1						1.5/5						0.4/2	1.2/3				0.8/1				0.2/1	3/9		0.7/3	1.7/2	27
PIK3R1	0/1												0.5/1	0.9/1	0.5/1	1.7/11		3.2/4								0/1	22
KDM5C		0/1	0.3/1		0.4/2		1/2	0.5/1				1/1		0.1/1	1.1/6	0.9/1		1.1/2		0/1	0/1	1/1				0/1	22
ATM							0/5				0.6/1	0/2		0/1	1.9/6	0.5/1		3.8/10			0/1	0/2	4.7/10		2.5/4	0/1	45
KMT2C				0/1			0.6/5					0.7/1	1.8/2	1/4	0/2	0/1		0.8/1			1/6				1.9/2	1.1/3	28
SMAD4			0/1				11.3/43			0/3		3.1/9	0.4/1	2.5/3	0.3/1			1.2/8		1.1/4		1.4/2				0/1	79
SMARCA4		0/1		0.2/2		0/2	1/2		0.1/6			0.3/1				0.6/2		1/3		0.5/3	2.5/5			0/1		1.1/2	30
MAP3K1			0.3/1				0.3/1					0.2/1		3.8/16		1.1/4										0/1	24
KMT2D				0.7/9	0/3		0.2/2		7.2/29	0.2/1			0.1/1	0/1		1/2		1/1		1.5/4	3.3/12			0/2	1/4	1.2/3	74
ARID1A		0/1		0.5/2		0/1	2.3/17		1.1/9	0/3		2.4/7	4.3/14	0/1	1.5/2	2.4/9	0.5/1	0.6/4		1.2/5	0.9/1	0/1				1.2/2	80
BAP1	0/1					0/2				0.6/2			1.2/9	0.8/1	2.1/8					1/1							25
PTEN	0.1/10			0/2		0.8/1	2/2	0/1	0/3	0.2/2	0/2	0.6/3	1/2	0.1/5	1.8/4	3.7/18	0/1	2.6/6		0.9/1	0.4/2	2.4/6			1.8/6	0/2	80
FBXW7				1/3		0/1	0/1			0.7/2		0/1	0/2	0/2		1.1/4		1.9/7			1.6/4		0/1		0.3/1	0.3/2	31
MEN1	0/2													0.4/1						0.4/1		6/27					31
ARHGAP35			0/2	0/2	0.8/2		1.1/4			0/1			0/1		0.9/5	2.3/8		1/1							0/1	0.1/2	29
NOTCH1			0.2/1	1.6/14	0.9/10								0.9/1	0.3/2			1/1			0/1	1.5/2	0.2/1		0.6/1			34
B2M				0/1	0/1		1/1		3.2/14				0/1	0/1				1/6									25
ALB					0/2		0.9/1						4.3/24					0.6/1	0/1								29
PIK3CA	0.6/1	0.3/1	0.5/2	0.3/6	2.8/8		0.7/4	0.9/2	0/1	0.8/2	1.3/2	1.5/6	0/2	4.2/32		1.7/18		5/24		2.2/7	1.1/2			0.5/2	1.4/2	0.9/7	131
APC	0.4/1			0/2			1.9/4			0/1		1.5/6	1.6/4	0/2				9.8/54	0/1	0/2					0.9/2		79
IDH1	0.6/3									0.8/1	0.7/2		0/3	0.6/1			1/14	0/1			1/2				0.9/1		28
PBRM1					0/1	0/1	1.1/2			0/1		0/2	1/1		6.5/38			0.9/1									47
VHL						0/1									10.3/56	1/1						0/1					59
ARID2	0.1/1	0/1				0/1	0.8/1			0/2		0/1	4.7/15	0/1				0/2		0.4/2	0/1				0/2	0/1	31
CTNNB1	0/1	0.4/2										0.5/3	11.7/63			0.8/4		2/11		0.5/4	0.4/3	0/1			1/2		94
CDKN2A	0.5/15			1.2/11	0.3/8		8.8/40			0/1	0.2/1	1.5/5	0.3/3	0.8/2		0/1				0/1		0/1				1/1	93
CREBBP			0.5/1	1.8/10					3.4/27							0.5/1		0/1			0/1		0.1/1			0/1	43
TP53	1/13	0/3	2.2/76	0.6/33	1.3/25	0.1/1	9.8/126	0.2/7	1.8/22	1.5/6	0.1/4	4.5/46	6.4/54	7.6/57	0.1/3	3.1/28	1.1/5	3.8/23	1/1	1.5/10	0/1	0.4/3	2.6/5		3.3/8	1.2/20	589
KRAS	0/3	0.7/2	0/1	0/1	0/1		11.2/164	0/1		0/1		0.3/3				0/3		3.6/23		0.9/3				0.9/1		0.3/1	209
Other	6.5/119	1.3/21	7.7/49	9/51	4.7/27	2/9	23.9/124	1.7/10	36.2/231	3.1/16	3.9/24	6.2/28	28.5/149	17.5/69	3.6/23	11.3/43	5.3/18	32/106	5.7/23	8.5/29	23.2/88	11.8/47	9.2/43	4.9/13	21.5/51	20.8/95	1524

Fig. 7.4 A pan-cancer overview of subclonal drivers. Each cell contains two values, separated by a divider: The sum of the subclonal probabilities and the total number of drivers identified in the gene and the cancer type. Bars at the edges show the proportions of tumours with clonal (grey) and subclonal (dark blue) drivers for cancer types (top) and genes (side). The top 30 genes are shown, obtained by summing the subclonal probability of all drivers per gene.

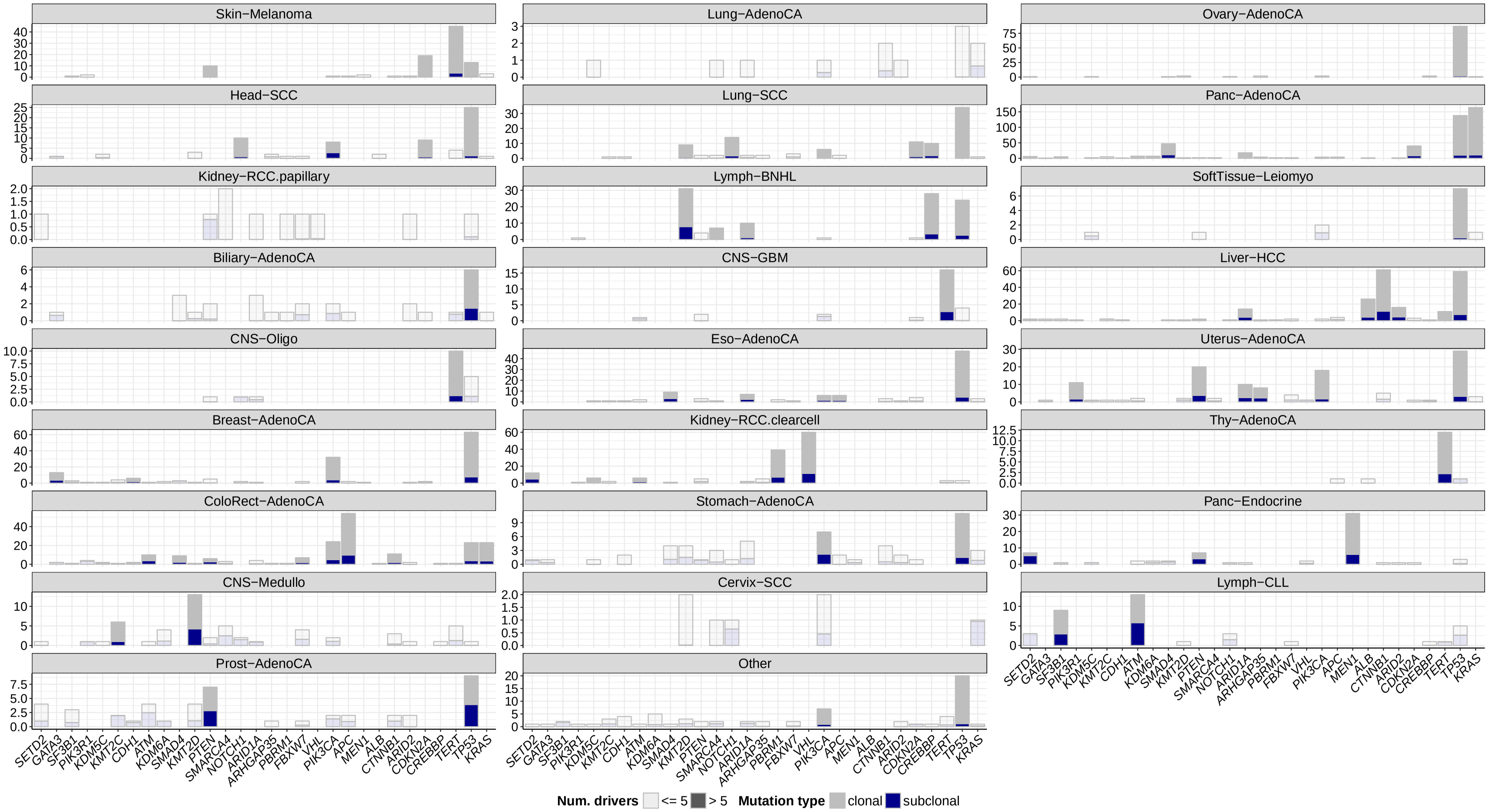


Fig. 7.5 Counts of samples per cancer type, per top 30 drivers shown in Fig. 7.4. Bars are greyed out when fewer than 5 tumours are found for a cancer type and gene combination.

Figure 7.5 shows that several genes are subclonal in high proportions of relatively few tumours: *SETD2* in pancreatic endocrine tumours, *ATM* in CLL cases and *PTEN* and *TP53* in prostate cancers. Other genes are at the top of the list of subclonal driver genes pan-cancer wide, but are supported by low numbers across cancer types: *GATA3*, *KMT5C* and *CDH1*.

These findings suggest that a large proportion of late drivers have yet to be identified. They show that known driver genes can be active late during tumour evolution, which may suggest that a particular environment is required for a driver to be effective.

If is likely however that the picture painted in this section is just the tip of the iceberg. Mutations identified as clonal in the sequencing sample may in fact not be carried by all tumour cells. In the scenario of small biopsies one may be painting a picture of the small biopsy only and therefore overestimate the number of mutations that are clonal, and what is determined to be subclonal to consist of relatively minor subclones. When a large portion of the tumour is obtained for sequencing one can be more confident about whether it represents the whole tumour, however subclones will represent major cellular populations at this scale. How the sequencing samples were obtained exactly has not been recorded as part of the PCAWG effort and it is therefore unclear how this affects the painted picture.

It is important to note however that what is subclonal at any point within the tumour, is in fact subclonal in the whole tumour. The painted picture therefore represents a conservative estimate of the amount of subclonality, regardless of the sampling strategy. Finally, our findings have been partially reported elsewhere, with *SETD2* often appearing as subclonal in a detailed study of kidney cancers (Turajlic et al., 2018) and driver mutations in chromatin remodellers shown to correlate with later transitions between tumour progression states in melanomas (Shain et al., 2018).

7.6 14% of mutations are undetected

We applied our methods to correct for the *winner's curse* (which were introduced in section 6.5) and estimate the number of mutations that are unaccounted for given a tumour's subclonal architecture. Figure 7.6 shows the correction applied for cluster position (left) and cluster size (right). Clonal clusters (shown in grey) often are corrected very little, highlighting that typically (nearly) all clonal mutations are detectable above 30X sequencing coverage. Subclonal clusters however can be corrected extensively with many mutations falling below the detection limit, in line with the observations on simulated data in the previous chapter. The average correction across all tumours (i.e. the difference between total detected mutations and total estimated mutations) is 14%, suggesting that 14% of mutations have been missed because they fell below the detection limit.

7.7 Clear signs of positive selection in subclonal mutations

We observe clear signs of positive selection within clonal and subclonal mutations and for missense, nonsense and splice-site SNVs (Fig. 7.7). Inspection of driver mutations reveals that the detected subclones contain driver mutations in known cancer genes (Fig. 7.4).

We next looked for signs of positive selection in both the clonal and subclonal mutations by analysing the ratio of non-synonymous and synonymous mutations, an approach often referred to as dN/dS. A ratio larger than 1 is considered a sign of positive selection, a

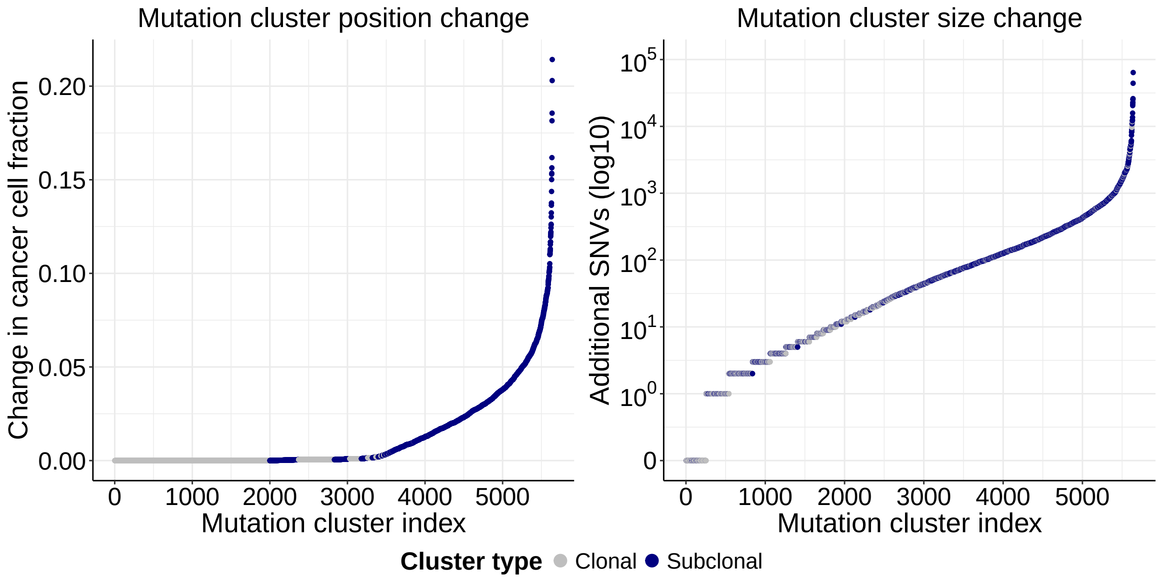


Fig. 7.6 The correction applied to cluster location (left) and cluster size (right). The average correction across all tumours is 14%. Clonal clusters are often adjusted very little, while subclonal clusters can be adjusted considerably, in line with observations on simulations in the previous chapter.

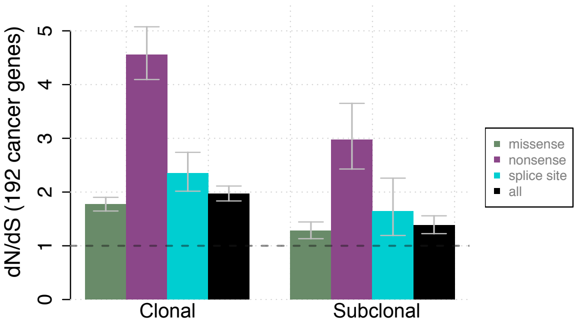


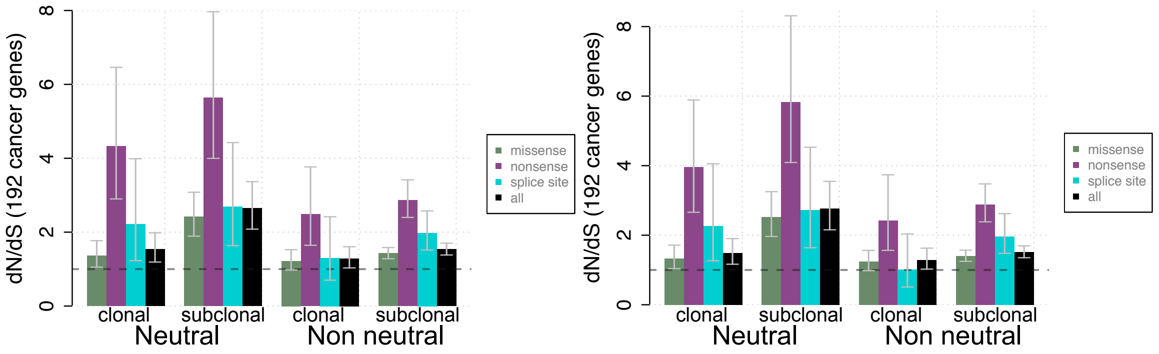
Fig. 7.7 dN/dS values for clonal and subclonal SNVs across all primary tumours as described by Martincorena et al. Values for missense, nonsense and all mutations are shown, along with the 95% percentage intervals. Positive selection is observed in all mutation classes.

ratio below 1 represents negative selection, while a ratio of 1 can mean either no selection (neutral) or an equal amount of positive and negative selection. dN/dS ratios have been used extensively in the field of evolutionary biology and have recently been adapted to study selection in cancer genomes (Greenman et al., 2006; Martincorena et al., 2016). We used the approach published in Martincorena et al. (2017) that models tri-nucleotide contexts and considers additional non-synonymous mutations beyond missense mutations, such as nonsense and splice-site mutations and indels, and has been shown to accurately recapitulate existing knowledge about cancer drivers (Martincorena et al., 2017). The analysis was performed on the 192 cancer genes in the COSMIC Cancer Gene Census v80 (Futreal et al., 2004) to provide a conservative estimate of positive selection.

Recently there has been discussion in the field of tumour evolution about whether positive selection may no longer be present and that further evolution of these tumours occurs due to genetic drift (Sottoriva et al., 2015; Williams et al., 2016). Williams et al. (2016) recently proposed a test that can be used on bulk whole genome sequencing data to identify tumours for which this is the case. The test is based on the principle that the further one zooms into a neutrally evolving tumour, the number of subclones and mutations increases at an exponential rate. If these neutral subclones are captured in a sequencing sample, one therefore expects the number of mutations to increase at an exponential rate as the VAF distribution goes to zero. Williams et al. (2016) propose to test the mutation VAF space of a sequencing sample against an exponential curve (which I'll denote as a "1/f" tail) and a high correlation between the VAF tail and the "1/f" tail would indicate the tumour is evolving neutrally. Williams et al. (2016) recommend a correlation of over 0.98 indicates neutral evolution.

The test and results are however not without controversy (Noorbakhsh and Chuang, 2017; Tarabichi et al., 2017). We therefore applied this principle to the PCAWG data set to identify neutrally evolving tumours and did so separately for all mutations and for mutations in unaltered copy number regions (one copy of the maternal and paternal alleles). A tumour was called neutral when the cumulative VAF space yielded a correlation of over 0.98 with a "1/f" tail. This test identified 557 tumours as neutrally evolving (531 on all mutations and 499 tumours when considering only mutations in normal copy number). The tumours that are identified as neutrally evolving have significantly higher reads per chromosome copies (p-value $8.74*10^{-90}$, Mann-Whitney-U test), which may suggest the number identified is an underestimate, as many tumours in the PCAWG dataset it is not possible to identify sufficient subclonal mutations.

We next applied the dN/dS pipeline to the mutations in these tumours, which identified positive selection in both clonal and subclonal mutations in neutral and non-neutral tumours (Fig. 7.8). Tumours identified as neutrally evolving also contain subclonal driver mutations



(a) Mutations in CNA regions

(b) Mutations in normal copy number

Fig. 7.8 Tumours were classified into neutral and non-neutral according to the rationale described by Williams et al. (2016) dN/dS values for clonal and subclonal SNVs were derived separately across all primary tumours in those two groups, as described by Martincorena et al. (2017). Values for missense, nonsense and all mutations are shown, along with the 95% percentage intervals. The same figure is shown after grouping neutral and non-neutral tumours based on SNVs in the diploid genome only.

in known cancer genes. 345 of the 557 identified tumours contain at least one identified driver. The tumours contain a total of 893 driver mutations, of which 114 have a probability > 0.95 of being subclonal. We find subclonal driver mutations in *TP53* and *PTEN* (6 each), *SETD2* (4), *ATM*, *FBXW7*, *KIT*, *NF1*, *SF3B1* and *TGFBR2* (3), 16 genes with 2 subclonal drivers and 48 with 1.

These findings show that tumours identified by the "1/f"-tail test contain subclones under positive selection and that the identified clonal expansions contain driver mutations in known cancer genes.

7.8 Subclonal clinically actionable events

I considered driver mutations (SNVs and indels) in genes and pathways for which drugs are either developed or in development to look specifically for tumours with subclonal targetable driver mutations (as predicted by Cancer Genome Interpreter (Tamborero et al., 2018)). A patient with a targetable driver mutation could in the near future be prescribed a targeted therapy, but the therapy is inherently flawed if the targetable mutation is not shared by all tumour cells. In this analysis we excluded all metastasis and relapse tumours, except melanomas. For multi-sample cases we only considered the PCAWG provided preferred sample for each donor.

Our consensus subclonal architecture approach produces probabilistic cluster assignments for each mutation and identifies a mutation cluster as clonal (the clonal cluster has CCF

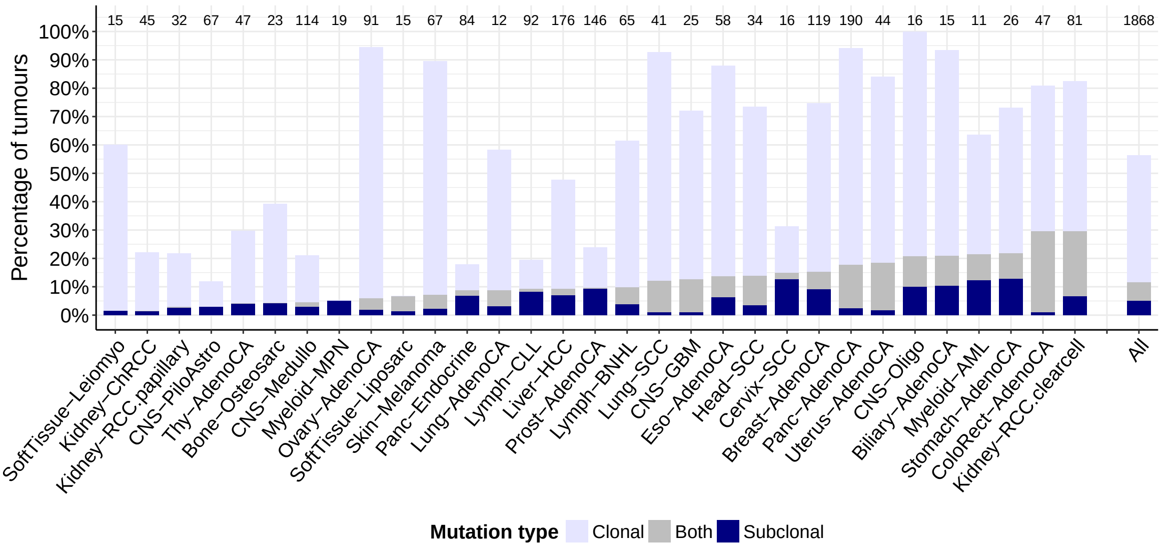


Fig. 7.9 Clinically actionable driver mutations were surveyed across the cohort (Sabarinathan et al., 2017) and assigned a probability of being clonal or subclonal. Per cancer type probabilities are combined to provide a fraction of tumours that contain only clonal actionable drivers, only subclonal actionable drivers or both (see supplementary methods). On average 11.7% of tumours contain at least one subclonal actionable driver, while in 5.1% of tumours we found that all actionable drivers are subclonal. Cancer types show markedly different proportions, ranging from 4.3% of thyroid cancers with at least one subclonal actionable driver to 29.7% of kidney clear cell carcinoma cases.

of 1, while a subclonal cluster has a CCF < 1). Through the consensus I can establish the probability whether a mutation is clonal or subclonal. The procedure is as follows: For each sample, I establish the probability that all actionable mutations are clonal, all actionable mutations are subclonal and the probability of observing at least one pair of clonal and subclonal targetable events.

The probability (p) of observing all n actionable mutations as clonal is:

$$\prod_{i=1}^{n} p_{i,clonal} \tag{7.1}$$

The probability of (p) of observing all n actionable mutations as subclonal is:

$$\prod_{i=1}^{n} p_{i,subclonal} \tag{7.2}$$

Then the probability of observing at least one pair of actionable mutations where one is clonal and one is subclonal is:

$$1 - \left(\prod_{i=1}^{n} p_{i,clonal} + \prod_{i=1}^{n} p_{i,subclonal} \right) \tag{7.3}$$

The three probabilities were summed to create the three classes per type of cancer: *Clonal*, *Subclonal* and *Both*.

Through this analysis I find that 11.7% of tumours have an identified subclonal driver mutation that is clinically actionable (Fig. 7.9). In 5.1% of tumours, I find targetable driver mutations only in subclones and 6.6% of tumours contain both a clonal and a subclonal target. These estimates are likely a lower bound as tumours are only represented by a single sample, which likely shows (depending on how the samples were obtained) either local heterogeneity or large subclones, and mutations that appear to be clonal in one area of the tumour may in fact be subclonal overall when they are not present in another region.

These findings suggest it is important to consider the clonal status of a targetable mutation before treatment is started. Prescribing a drug that targets a mutation not carried by all tumour cells is certain to be ineffective. Meanwhile, in 6.6% both a clonal and a subclonal targetable mutation is found. In this scenario, clonality assessment would highlight the clonal mutation as the best candidate.

However, for clonality analysis to be truly informative in clinical application, one must be highly confident that a mutation that appears clonal is indeed carried by all tumour cells. Ultimately, it may prove impossible to truly establish a mutation is carried by every tumour cell as it would require assessing every tumour cell. Clonality assessment strategies may therefore be limited to identifying subclonal targetable mutations (a mutation that is subclonal in one region of the tumour is subclonal overall) that would provide ineffective treatment options.

7.9 Evidence of additional heterogeneity

Several studies have shown (Jamal-Hanjani et al., 2017; Sun et al., 2017) that multi-region sequencing is better powered to detect subclones, compared to single-region sequencing approaches. We reasoned that some of the subclones that cannot be reliably disentangled on single-region sequencing may leave a trace that can be detected in a single sample. Mutation clusters may be merged during a single-region based subclonal reconstruction when multiple subclones appear at a similar CCF. We therefore explored two aspects that could be informative about the number of additional subclones within a sequencing sample: Subclonal

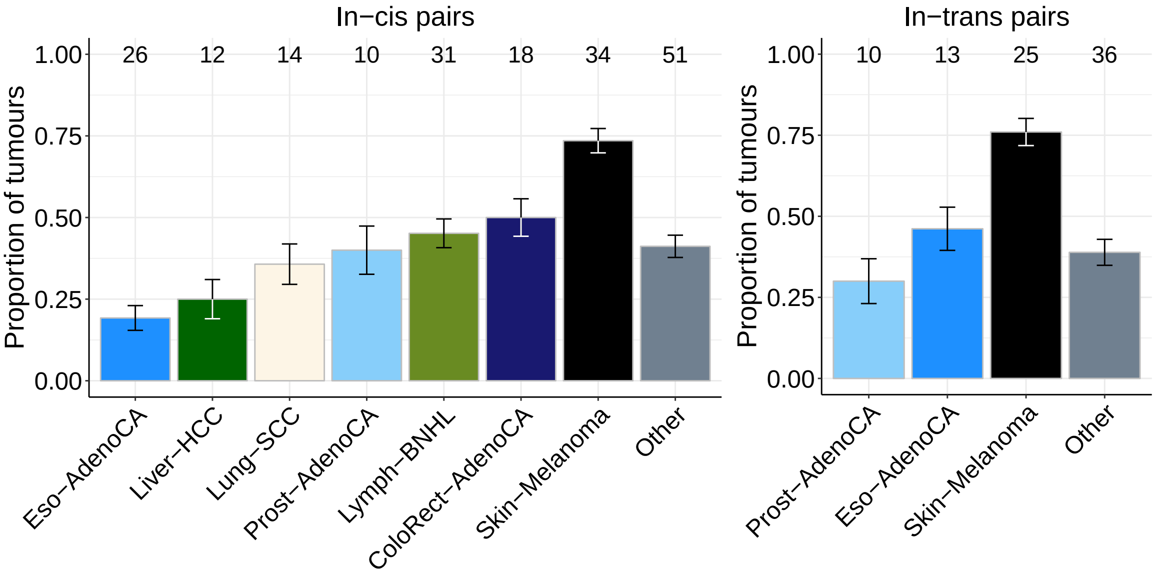


Fig. 7.10 Fraction of tumours where phased mutations provide evidence of additional heterogeneity for tumours where the mutations are *in-cis* (left, co-linear) or *in-trans* (right, branching) Error bars represent the binomial standard deviation of the total number of tumours for each type of cancer and the associated ratios.

mutational signature changes that are not close to a boundary between mutation clusters and evidence of mutations assigned to the same subclone that cannot have occurred in the same cell.

Within our working group, Yulia Rubanova has developed an approach (termed Tracksig) to estimate changes in mutational signature activity in approximate-time ordered mutations. Tracksig bins mutations and orders the bins by pseudo-time (mutations on two chromosome copies have occurred before mutations on the same segment carried by one chromosome copy, subclonal mutations occur after clonal mutations, etc) and subsequently detects in which pseudo-time bin mutational signature exposures change (Rubanova et al. 2017, manuscript in preparation). Yulia has applied her algorithm to the PCAWG data, which reveals that 37.4% of tumours had a signature exposure change of at least 10%, while 30.1% of signature changes correspond to a boundary between the mutations from a clone / subclone and 39.7% represent boundaries between subclones. We further find that an average of about 0.5 changes per sample are not within a subclone boundary, which suggests that additional subclones are measured by the sequencing data but have not been detected.

Within our working group Amit Deshwar has looked into pairs of mutations that cannot have occurred in the same cell, building on data that I generated. I generated counts for mutation pairs that fall within 700bp that could be spanned by a single read pair. For these mutations, it is possible to determine whether both mutations fall on the same chromosome copy (i.e. are phased) by examining the read pairs that cover both. A mutually exclusive pair of mutations (mutations are *in-trans*) that cannot have occurred in a single cell is measured as a pair of mutations (a,b) without a read-pair that report both variant alleles, while some reads report the variant allele of a and the reference allele of b and vice-versa, and the pair fall in a genomic region where only a single chromosome copy is available. In contrast, mutations where some read-pairs report both variant alleles and some pairs report only one (mutations are *in-cis*) the mutations represent clusters in an ancestral relationship. When considering phased mutation pairs Amit finds that in 44% (86 of 196) of tumours there is evidence of mutually exclusive mutations (Fig. 7.10).

These findings highlight that the found amounts of ITH, as is reported in Fig. 7.1, are an lower bound for the amount of ITH available in the sequencing samples.

7.10 Cancer types follow individual evolutionary narratives

We next characterised the evolutionary histories of the 2,658 tumour cases in the PCAWG dataset, described in full in Gerstung et al. (2017) and attached to this thesis as Appendix A. This section describes results that are the culmination of work by Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner and Santiago Gonzalez. In brief: three different mutation timing analysis were performed. (1) A classification of mutations into clonal early (gained mutations, on more than 1 chromosome copy), clonal late (mutations on a gained chromosome, but on 1 copy only), clonal unspecified (mutations in non-gained copy number regions) or subclonal.

(2) Timing of copy number gains was performed by accounting for multiplicity states (for example, a high ratio of multiplicity two mutations on a gained chromosome suggests the gain was late (Fig. A1.3)) and (3) CNAs were timed relatively against each other by league model analysis (these models pitch every pair of detected CNAs against each other, and like a sports league, build a league table out of all the matchups). We also overlayed mutational signature activity, and through the use of clock-like signatures we convert timing analysis into real time estimates.

We find that driver mutations predominantly occur early and are therefore observed as clonal (Fig. 7.11a and b). Drivers in *TP53* and *KRAS*, for example, are 5-9x more likely to occur early than late clonal. For *TP53* this effect is independent of tumour type (Fig. 7.11c). In general, the diversity in driver genes increases as tumour evolution progresses: 50% of all early driver mutations are found in 12 genes, while late clonal and subclonal drivers occur across 39 and 36 genes respectively (Fig. 7.11d). These findings suggest that early drivers occur within a specific set of genes, while late drivers are more diverse, giving rise to the "long tail" of driver genes active in low proportions of tumours.

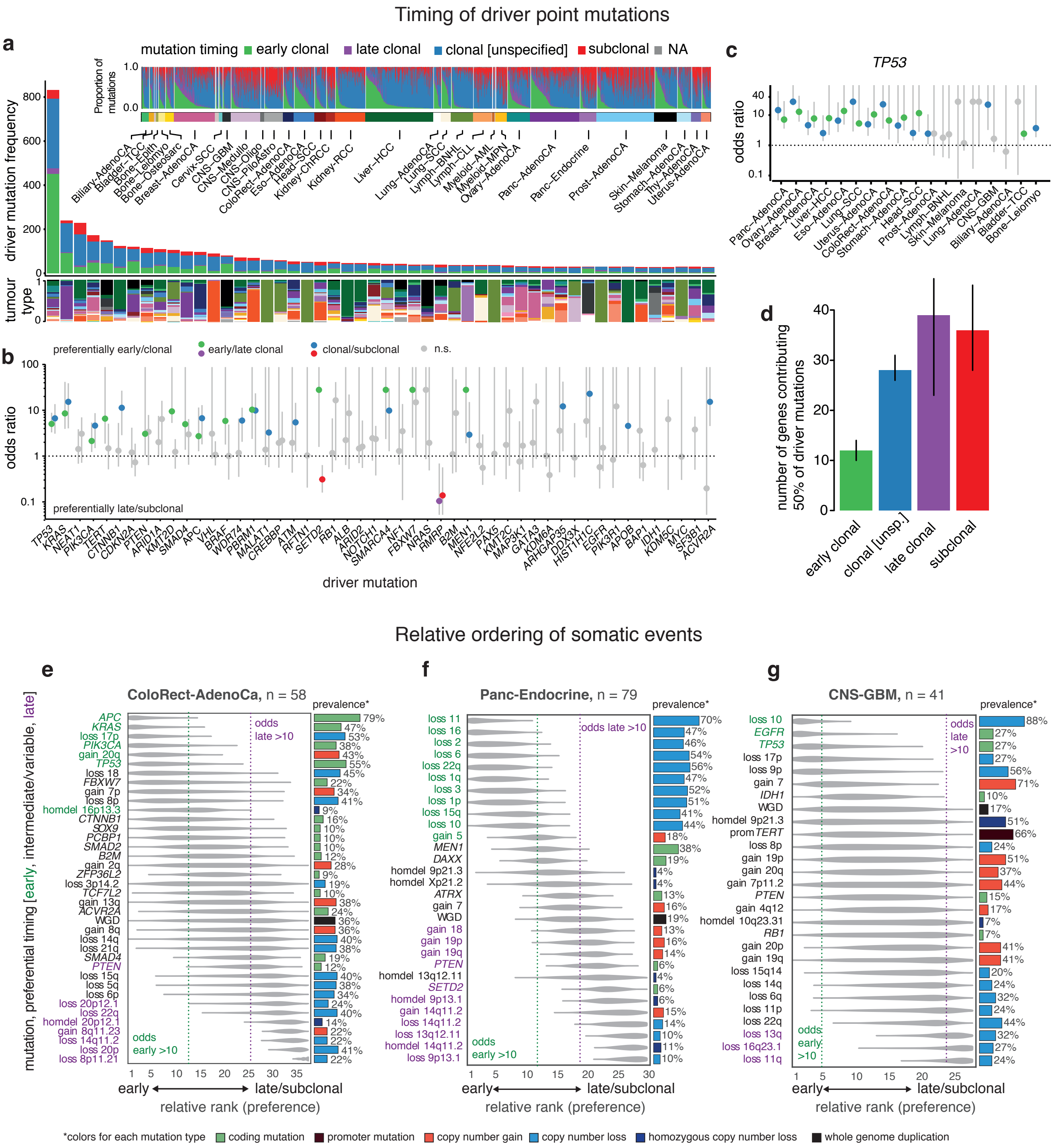
Our relative timing of events (Fig. 7.11e) reveals that in colorectal adenocarcinoma *APC* has the highest odds of occurring early, followed by *KRAS*, loss of 17p, *TP53*, loss of 8p, gain of 8q, in concordance with the progression model proposed by Fearon and Vogelstein (1990). For pancreatic endocrine cancers the relative timing suggests that in these tumours losses are frequently early, followed by driver mutations in *MEN1* and *DAXX* and a whole genome duplication (Fig. 7.11f). In glioblastoma cases we find that loss of chromosome 10 and driver mutations in *TP53* and *EGFR* are typically early, preceding early gains of chromosomes 7, 19 and 20 (Fig. 7.11g).

The timing of gains reveals that, pan-cancer wide, copy number gains typically occur during the second half of tumour evolution. But cancer types show marked differences (Fig. A2): Glioblastoma tumours show consistent early gains of chromosomes 7, 19 and 20, while medulloblastomas contain early gains of 17q. Gains are typically early in glioblastoma, medulloblastoma and pancreatic neuroendocrine cancers, late in squamous cell lung cancers and melanomas, while they occur during broad periods in other cancers.

Analysis of mutational process activity early, clonal and late reveals that signature activity typically changes by less than 30%, which indicates that signature activity is relatively constant during tumour evolution (Fig. 7.12a). Life style associated signatures, such as 4 (smoking associated) in lung adenocarcinoma, 7 (UV-light) in melanoma and 12 (aetiology unknown) in liver cancers, typically decrease in activity late, while signatures 2 and 13 typically increase in activity (Fig. A4b). The clock-like mutational signature 1 was used to infer real time estimates: whole genome duplications typically appear 2-11 years before diagnosis (Fig. 7.12b), while the most recent common ancestor appears six months to six years before we observe the tumour (Fig. 7.12c).

These analysis combined confirm previous knowledge about the distribution of driver genes and the classic progression model of colorectal adenocarcinoma, and reveal that cancer types follow distinct evolutionary patterns.

Figure 3: Timing of driver mutations and relative ordering of somatic events



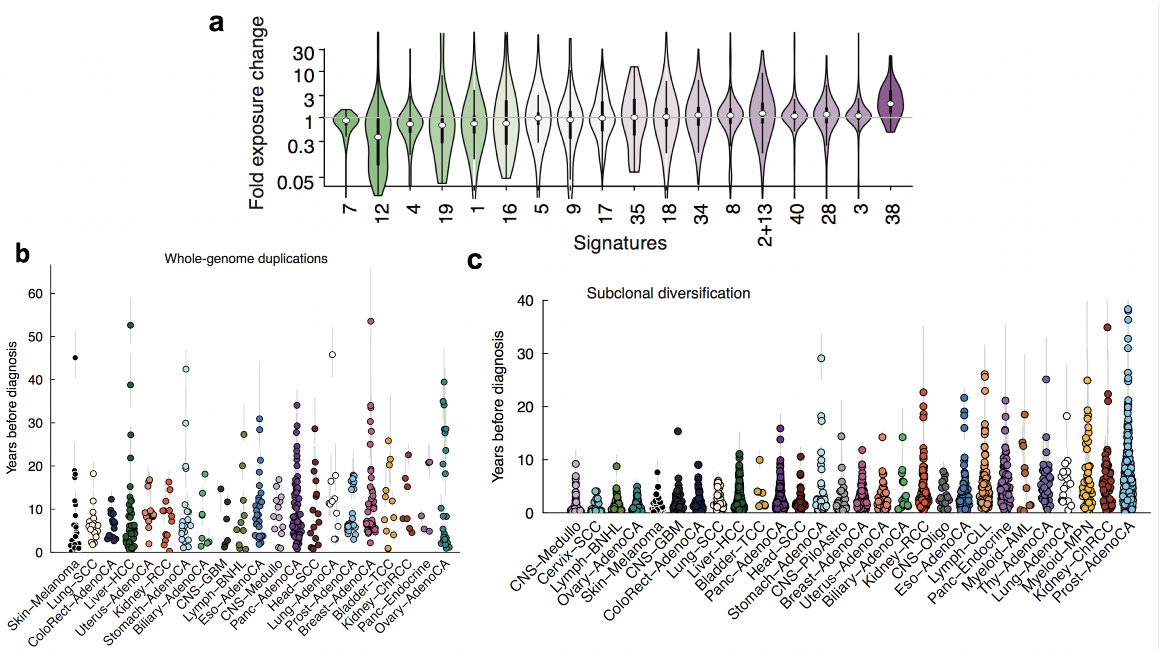


Fig. 7.12 This figure is a combination of various figures from Gerstung et al. (2017). (a) Fold changes in signature exposures between early and late clonal stages for all tumours. Each violin shows the distribution of exposure changes across tumour types in one signature. Signatures are sorted by the ratio of tumours with a positive signature change. (b) Time of occurrence of whole genome duplications in individual patients, split by tumour type, based on CpG>TpG mutations and patient age. Results are shown for a 5x acceleration of the mutation rate. (c) Timing of subclonal diversification using CpG>TpG mutations in individual patients.

Chapter 8

Discussion

8.1 Overall summary

In this thesis I have introduced computational methods to unravel the life history of a tumour from massively parallel sequencing data and applied them to the 2,778 cancer genomes in the ICGC PCAWG project. To aid with the subclonal reconstruction of tumours I introduced algorithms to estimate subclonal copy number, multiplicity and cancer cell fraction values for SNVs and indels and to infer the subclonal architecture.

The methods have been extensively validated using different simulated data sets and cross-compared to those from other labs. The validation shows that, even though DPClust is one of the best performing methods on the simulated data within the PCAWG consortium and that performance is consistent with the results on real data, it is not perfect. The observation that methods are not perfect lead to the development of consensus procedures for both copy number and subclonal architectures. I showed that the output of the consensus is more consistent across methods, and for the subclonal architectures that corresponded well to the patterns observed on simulated data.

The copy number consensus procedure first combines breakpoints from six CNA calling methods with SV calls into a single, complete consensus. By synchronising the segmentations across the six callers we achieve six profiles that are complete and directly comparable. There is substantial complete agreement between the callers, but agreement on the near full genome is only achieved through a strict majority vote. This highlights that there is still substantial disagreement between some of the methods, which has already resulted in method improvements. Consensus subclonal architectures were obtained across 11 individual subclonal inference methods. Three consensus approaches were developed based on different representations of the input data and validation showed that they are consistently comparable to the best individual caller, while assigning all SNVs, indels and SVs for which allele frequencies were available.

When these methods are applied to the PCAWG data set we find that nearly all tumours contain at least one subclone, that subclones contain driver mutations in known cancer genes and those drivers are under positive selection. These findings show that tumours are still evolving at the point of diagnosis. Furthermore, the PCAWG data set covers 36 histologically distinct cancer types and donors cover a wide range of ages, while recruitment has occurred in Europe, North America and Asia, resulting in a diverse population that covers many germline genetic backgrounds. The fact that we observe subclones in nearly all tumours across this range of variables may suggest that tumours are in a continuous process of clonal expansion.

Recent work suggests that tumours could evolve neutrally through genetic drift. We applied the concepts from Williams et al. (2016) and found signs of positive selection in subclonal mutations from tumours identified as evolving neutrally and these tumours can contain driver mutations for which we are highly confident that they are subclonal. However, dN/dS ratios were used to estimate the amount of positive selection, and dN/dS analysis pools mutations across samples. It is therefore possible that a subset of tumours is indeed no longer under the influence of positive selection, or that the VAF tail that is used to detect neutrality contains subclones that are under positive selection and those that don't. But that subset is sufficiently small to not affect the dN/dS ratios obtained on the pooled mutations.

Analysis of clinically actionable driver mutations reveals that the detected clones can be informative in the clinic. 11% of tumours contain at least one subclonal actionable event, which means a prescribed treatment is inherently flawed as it would not target all tumour cells. However, over half these tumours also contain a clonal actionable event. That may provide a route to apply more effective treatment as clones could be targeted separately.

Evaluation of mutational signatures shows that activity of life-style associated signatures decreases during tumour evolution, although this signal could also be explained by the increase of a combination of other signatures. Meanwhile, APOBEC activity typically increases. Finally, cancer timelines were created by combining the evolutionary histories of tumours within a cancer type. The analysis confirms classic knowledge and suggests that cancer types follow distinct patterns of tumour evolution.

8.2 Future directions

8.2.1 A more in-depth view of intra-tumour heterogeneity

The single sample whole genome sequencing data presented in this thesis ultimately provides only a high level view of intra-tumour heterogeneity. Work by others based on multiple samples from the same cancer (Jamal-Hanjani et al., 2017) shows that there is more high level heterogeneity, and our analysis about additional heterogeneity corroborates that. It's possible that the view obtained by these bulk sequencing approaches is just the tip of the iceberg and that tumours consist of many 100s to 1000s of clones. Bigger and deeper sequencing studies will reveal the true extend of intra-tumour heterogeneity.

Larger sequencing studies, that cover more cases of the same cancer type, such as the Pan Prostate Genomics Consortium, are needed to extend our knowledge of late drivers. The work in this thesis suggests that tumours become more diverse as they evolve, with a larger set of genes acting as late drivers. A complete overview of late drivers may lead to new treatment options and it may shed light on drivers that are rarely early.

A currently mostly unexplored angle in these kinds of heterogeneity studies is expression data, which could be overlayed onto the subclonal architectures. Subclonal inference is focussed on establishing the genotype of evolution. What effect these additional mutations contained within the cells in the subclone have on the expression profile remains unclear. Using matched RNAseq data it should be possible to observe expressed transcripts with a subclonal mutation. This kind of analysis will however not provide a complete overview of the expression profile of a subclone. Single cell sequencing technology, especially sequencing of the genome and transcriptome from the same cell, does give access to that information (Macaulay et al., 2015).

Single cell technology also provides access to the lower levels of heterogeneity that are not visible with current bulk sequencing approaches. One could paint the 3D landscape of tumour heterogeneity by carefully sampling cells from the tumour environment (Mamlouk et al., 2017). However, single cell genome and transcriptome technology is still too expensive and cannot easily be scaled up to the numbers required to paint a comprehensive picture, while the genomic data is hampered by factors such as allele dropout and sequencing errors (Van Loo and Voet, 2014).

Higher sample counts per tumour can not only lengthen the branches of the evolutionary trees, it may also provide a more fine grained picture of the tree trunk. SNVs that appear as clonal in one tumour region, may in fact appear subclonal in another. A more fine grained trunk will provide a better view of the very early events that have given rise to the tumour and provides a clearer sequence of events. A clearer picture can then lead to cancer type evolutionary histories with smaller confidence intervals.

8.2.2 Tumour evolution

Recent work on normal tissue from healthy individuals reported many small clones at an in-depth view of epithelial tissue and showed complex clonal dynamics, containing clones with driver mutations in well known cancer genes (Martincorena et al., 2016). Given these observations, and assuming they can be extrapolated to other tissues, it is surprising that the cancer evolutionary timelines reported in this thesis suggest that cancers develop over decades. The process of clonal expansion may appear similarly in healthy tissues and in tumours. More detailed experiments are required to understand the dynamics of the environment in which the tumour grows and why the malignant growth can escape while many early clones cannot. Ultimately, we cannot easily observe the micro-environment *in vivo* when a driver mutation has its selective advantage. Normal tissue observational experiments will be required to better understand the dynamics of clones that provides the breeding ground for malignant lesions.

More detailed experimental data will also facilitate those who work on mathematical models of tumour evolution. So far this field has been held back by the lack of a bridge between models of early evolution and the tumours observed in the clinic that represent a much later evolutionary phase. Projects are needed where model development is provided with measurements of input variables, and longitudinal follow-up of clonal dynamics of the populations that provided the input variables could provide the ground truth for intermediate predictions made by developed models (for example, clone sizes and distributions at various time-points). Scenarios should include the introduction of driver mutations, through CRISPR for example, or population bottlenecks as created by the application of drugs. A combination of these two should lead to closing of the gap between those that study tumour evolution top-down (as is reported in this thesis) and those that study it bottom-up.

8.2.3 Towards clinical application

The findings reported in this thesis reveal that tumour evolution continues up to (shortly before) diagnosis. This has profound clinical implications as it suggests every tumour can in principle become resistant to treatment (Holohan et al., 2013), which can be unlocked via a single mutation (Nazarian et al., 2010; Zaretsky et al., 2016). In cancer types with a high mutation burden it is therefore not implausible that every conceivable somatic mutation is available. The resistance mechanism may therefore already be available and will be selected for via treatment application. It is therefore important to assess how ubiquitous the resistance mechanism has to be before treatment becomes inevitably unsuccessful. If a large subpopulation of cells with the resistance mechanism is required (as opposed to a single cell with a single mutation), then a routine screening could be facilitated to rule out certain types of treatment. Large scale collection and deep profiling of pre- and post-resistance cancers is required, along with a more complete list of resistance mechanism markers, is required to see whether heterogeneity profiling can be a useful clinical application.

The findings also suggest that broad timelines exist across cancers of a certain type. It has recently been suggested ctDNA provides a complete picture of the major clonal populations within a tumour, and a blood test could therefore in principle be used to not only detect a tumour, but to also assess and follow its progression (Abbosh et al., 2017). A non-invasive blood test could help detect cancers more easily and potentially earlier. If one can classify a tumour as to be on a evolutionary ‘pathway‘, then it may be possible to predict how it will develop further and choose treatments that slow tumour progression by closing possible evolutionary routes. However, even though evidence is beginning to emerge that overall timelines may exist within cancer types (Fearon and Vogelstein, 1990; Gerstung et al., 2017; Makohon-Moore et al., 2018), it is currently not clear whether individual tumours truly follow such a pathway, or whether the observed order of driver mutations is one out-of-many ways to malignancy. A more detailed picture is required of evolutionary paths that tumours take, but it is equally important to assess which other evolutionary paths arise during a tumours' life time and are outcompeted.

The emergence of ctDNA based tumour tracking potentially allows for much earlier diagnosis via routine testing. However, it will increasingly emphasise that better understanding of the difference between normal and malignant tissue evolution is required. Driver mutations in *TP53* in morphologically normal epithelial have been described and are suggesting that just the acquisition of these driver mutations is not enough for a cancer to arise (Martincorena et al., 2018). A more comprehensive overview of normal somatic evolution is required to demarcate the crossover point to malignancy more clearly, and to show that somatic mutations alone (as measured via ctDNA) can clearly differentiate between the two states.

8.2.4 Methods

Within the PCAWG consortium there has been a great emphasis on consensus strategies that build confident calls by combining evidence from multiple methods. These ensemble methods perform well, however they come at considerable computational cost. Our copy number consensus procedure required six methods to run across the full data set to obtain a complete set of breakpoints, followed by another full run to obtain calls on the consensus segmentation. It is interesting that the review sessions, where copy number was discussed in great detail, has already yielded improvements to methods and has sparked additional development.

There is not only a need for more accurate methods, there is also a need for public benchmarking data sets where performance can be validated and to aid the development process. The extensive comparison of subclonal architecture callers that was performed internally and lead to the development of three consensus procedures also lead to improvements to individual methods. Efforts like the DREAM somatic variant calling - heterogeneity project are an important part of this, as are simulators like BAMsurgeon (Ewing et al., 2015) and SimClone.

More real data with multiple samples of the same tumour are also required to aid further benchmarking. As is discussed in this thesis, multi-region approaches can be more powerful to detect intra-tumour heterogeneity, as it allows for a larger area of the tumour to be sampled and mutation clusters with a similar CCF in one sample can be more easily separated if their CCF differs in another sample. The richer view that is obtained via multi-region sequencing can be used for validation purposes by applying subclonal architecture callers on just a single sample. So far however, comparatively little multi-region whole genome sequencing data has been published. The PCAWG project did include cases where multiple samples (primary-primary, primary-metastasis, primary-relapse) were available. However, these cases span few cancer types, each cancer type is represented by low numbers of cases and the cases are further split by multi-focality and whether they were obtained before or after start of treatment. A large cohort, with multi-region whole genome sequencing is required to help further validate performance of subclonal architecture callers.

My analysis of performance of DPClust on the SimClone1000 data set highlights limits of what can be detected. Subclones that are within 0.25 CCF of each other stand a good chance of being merged. Other methods may have a lower threshold for disentangling subclones, but there is always a limit. The subclone distance limit however is much less likely to affect cases with multi-region sequencing where a pair of mutation clusters may appear in similar proportions in one sample, but different in another. It is important to understand these limits when interpreting the output of subclonal architecture callers. Simulations will help to discover the limits and could ultimately be used to understand their impact on interpretation.

There is considerable scope to develop subclonal inference methods that use SNVs, indels, CNAs and SVs. There should be more signal to detect subclones by considering evidence across all types of mutations. In this thesis I have presented an approach that includes CNAs as pseudo-SNVs, based on CCF values estimated by Battenberg. This approach however does not take into account the characteristics of the underlying data upon which the copy number estimates are based. A combined approach should take those into account, should have a suitable error model for the CNA CCFs and should include a step that detects whether the assumptions required to estimate CCF values for CNAs are violated.

References

1000 Genomes Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Abbosh, C., Birkbak, N. J., Wilson, G. A., Jamal-Hanjani, M., Constantin, T., Salari, R., Quesne, J. L., Moore, D. A., Veeriah, S., Rosenthal, R., Marafioti, T., Kirkizlar, E., Watkins, T. B. K., McGranahan, N., Ward, S., Martinson, L., Riley, J., Fraioli, F., Bakir, M. A., GrÖnroos, E., Zambrana, F., Endozo, R., Bi, W. L., Fennessy, F. M., Sponer, N., Johnson, D., Laycock, J., Shafi, S., Czyzewska-Khan, J., Rowan, A., Chambers, T., Matthews, N., Turajlic, S., Hiley, C., Lee, S. M., Forster, M. D., Ahmad, T., Falzon, M., Borg, E., Lawrence, D., Hayward, M., Kolvekar, S., Panagiotopoulos, N., Janes, S. M., Thakrar, R., Ahmed, A., Blackhall, F., Summers, Y., Hafez, D., Naik, A., Ganguly, A., Kareht, S., Shah, R., Joseph, L., Quinn, A. M., Crosbie, P., Naidu, B., Middleton, G., Langman, G., Trotter, S., Nicolson, M., Remmen, H., Kerr, K., Chetty, M., Gomersall, L., Fennell, D. A., Nakas, A., Rathinam, S., Anand, G., Khan, S., Russell, P., Ezhil, V., Ismail, B., Irvin-sellers, M., Prakash, V., Lester, J. F., Kornaszewska, M., Attanoos, R., Adams, H., Davies, H., Oukrif, D., Akarca, A. U., Hartley, J. A., Lowe, H. L., Lock, S., Iles, N., Bell, H., Ngai, Y., Elgar, G., Szallasi, Z., Schwarz, R. F., Herrero, J., Stewart, A., Quezada, S. A., Van Loo, P., Dive, C., Lin, C. J., Rabinowitz, M., Aerts, H. J., Hackshaw, A., Shaw, J. A., Zimmermann, B. G., the TRACERx Consortium, the PEACE Consortium, and Swanton, C. (2017). Phylogenetic ctDNA analysis depicts early stage lung cancer evolution. *Nature*, advance online publication.

Adam, M., Thorburn, M. J., Gibbs, W. N., Brooks, S. E. H., and Hanchard, B. (1970). Clonal Evolution in Two Patients with Autoimmune Disease and Lymphoreticular Neoplasia. *British Journal of Cancer*, 24(2):266–276.

Ahmad, A. S., Ormiston-Smith, N., and Sasieni, P. D. (2015). Trends in the lifetime risk of developing cancer in Great Britain: Comparison of risk for those born from 1930 to 1960. *British Journal of Cancer*, 112(5):943.

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., and Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47(12):1402.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013). Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*, 3(1):246–259.

Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., Ji, H. P., and Maley, C. C. (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, 22(1):105–113.

Armstrong, K., Micco, E., Carney, A., Stopfer, J., and Putt, M. (2005). Racial Differences in the Use of BRCA1/2 Testing Among Women With a Family History of Breast or Ovarian Cancer. *JAMA*, 293(14):1729–1736.

Avery, O. T., Macleod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxycibonucleic acis fraction isolated from pneumococcus type III. *The Journal of Experimental Medicine*, 79(2):137–158.

Balkwill, F. R., Capasso, M., and Hagemann, T. (2012). The tumor microenvironment at a glance. *Journal of Cell Science*, 125(23):5591–5596.

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K. K., Carter, S. L., Frederick, A. M., Lawrence, M. S., Sivachenko, A. Y., Sougnez, C., Zou, L., Cortes, M. L., Fernandez-Lopez, J. C., Peng, S., Ardlie, K. G., Auclair, D., Bautista-Piña, V., Duke, F., Francis, J., Jung, J., Maffuz-Aziz, A., Onofrio, R. C., Parkin, M., Pho, N. H., Quintanar-Jurado, V., Ramos, A. H., Rebollar-Vega, R., Rodriguez-Cuevas, S., Romero-Cordoba, S. L., Schumacher, S. E., Stransky, N., Thompson, K. M., Uribe-Figueroa, L., Baselga, J., Beroukhim, R., Polyak, K., Sgroi, D. C., Richardson, A. L., Jimenez-Sanchez, G., Lander, E. S., Gabriel, S. B., Garraway, L. A., Golub, T. R., Melendez-Zajgla, J., Toker, A., Getz, G., Hidalgo-Miranda, A., and Meyerson, M. (2012). Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 486(7403):405–409.

Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10).

Bergamaschi, A., Kim, Y. H., Wang, P., Sørlie, T., Hernandez-Boussard, T., Lonning, P. E., Tibshirani, R., Børresen-Dale, A.-L., and Pollack, J. R. (2006). Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes and Cancer*, 45(11):1033–1040.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., Mc Henry, K. T., Pinchback, R. M., Ligon, A. H., Cho, Y.-J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M. S., Weir, B. A., Tanaka, K. E., Chiang, D. Y., Bass, A. J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F. J., Sasaki, H., Tepper, J. E., Fletcher, J. A., Tabernero, J., Baselga, J., Tsao, M.-S., Demichelis, F., Rubin, M. A., Janne, P. A., Daly, M. J., Nucera, C., Levine, R. L., Ebert, B. L., Gabriel, S., Rustgi, A. K., Antonescu, C. R., Ladanyi, M., Letai, A., Garraway, L. A., Loda, M., Beer, D. G., True, L. D., Okamoto, A., Pomeroy, S. L., Singer, S., Golub, T. R., Lander, E. S., Getz, G., Sellers, W. R., and Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905.

Bingle, L., Brown, N. J., and Lewis, C. E. (2002). The role of tumour-associated macrophages in tumour progression: Implications for new anticancer therapies. *The Journal of Pathology*, 196(3):254–265.

Blanke, C. D., Demetri, G. D., von Mehren, M., Heinrich, M. C., Eisenberg, B., Fletcher, J. A., Corless, C. L., Fletcher, C. D., Roberts, P. J., Heinz, D., Wehre, E., Nikolova, Z., and Joensuu, H. (2008). Long-Term Results From a Randomized Phase II Trial of Standard- Versus Higher-Dose Imatinib Mesylate for Patients With Unresectable or Metastatic Gastrointestinal Stromal Tumors Expressing KIT. *Journal of Clinical Oncology*, 26(4):620–625.

Blokzijl, F., Janssen, R., van Boxtel, R., and Cuppen, E. (2018). MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. *Genome Medicine*, 10(1):33.

Bolli, N., Avet-Loiseau, H., Wedge, D. C., Van Loo, P., Alexandrov, L. B., Martincorena, I., Dawson, K. J., Iorio, F., Nik-Zainal, S., Bignell, G. R., Hinton, J. W., Li, Y., Tubio, J. M. C., McLaren, S., O' Meara, S., Butler, A. P., Teague, J. W., Mudie, L., Anderson, E., Rashid, N., Tai, Y.-T., Shammas, M. A., Sperling, A. S., Fulciniti, M., Richardson, P. G., Parmigiani, G., Magrangeas, F., Minvielle, S., Moreau, P., Attal, M., Facon, T., Futreal, P. A., Anderson, K. C., Campbell, P. J., and Munshi, N. C. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature Communications*, 5.

Boveri, T. (2008). Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *Journal of Cell Science*, 121 Suppl 1:1–84.

Bozic, I., Reiter, J. G., Allen, B., Antal, T., Chatterjee, K., Shah, P., Moon, Y. S., Yaqubie, A., Kelly, N., Le, D. T., Lipson, E. J., Chapman, P. B., Diaz, Jr, L. A., Vogelstein, B., and Nowak, M. A. (2013). Evolutionary dynamics of cancer in response to targeted combination therapy. *eLife*, 2:e00747.

Brioli, A., Melchor, L., Cavo, M., and Morgan, G. J. (2014). The impact of intra-clonal heterogeneity on the treatment of multiple myeloma. *British Journal of Haematology*, 165(4):441–454.

Brose, M. S., Volpe, P., Feldman, M., Kumar, M., Rishi, I., Gerrero, R., Einhorn, E., Herlyn, M., Minna, J., Nicholson, A., Roth, J. A., Albelda, S. M., Davies, H., Cox, C., Brignell, G., Stephens, P., Futreal, P. A., Wooster, R., Stratton, M. R., and Weber, B. L. (2002). BRAF and RAS Mutations in Human Lung Cancer and Melanoma. *Cancer Research*, 62(23):6997–7000.

Cairns, J. (1975). Mutation selection and the natural history of cancer. *Nature*, 255(5505):197–200.

Cairns, J. (1981). The origin of human cancers. *Nature*, 289(5796):353.

Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O., Stein, L. D., and -ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Net (2017). Pan-cancer analysis of whole genomes. *bioRxiv*, page 162784.

Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., Follows, G. A., Green, A. R., Futreal, P. A., and Stratton, M. R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proceedings of the National Academy of Sciences*, 105(35):13081–13086.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., and Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B. M., Esserman, L., Albertson, D. G., Waldman, F. M., and Gray, J. W. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213.

Cichowski, K. and Jacks, T. (2001). NF1 Tumor Suppressor Gene Function. *Cell*, 104(4):593–604.

Cmero, M., Ong, C. S., Yuan, K., Schröder, J., Mo, K., Group, P. E. a. H. W., Corcoran, N. M., Papenfuss, A. T., Hovens, C. M., Markowetz, F., and Macintyre, G. (2017). SVclone: Inferring structural variant cancer cell fraction. *bioRxiv*, page 172486.

Coffelt, S. B., Wellenstein, M. D., and de Visser, K. E. (2016). Neutrophils in cancer: Neutral no more. *Nature Reviews Cancer*, 16(7):431–446.

Coghlin, C. and Murray, G. I. (2010). Current and emerging concepts in tumour metastasis. *The Journal of Pathology*, 222(1):1–15.

Condeelis, J. and Pollard, J. W. (2006). Macrophages: Obligate Partners for Tumor Cell Migration, Invasion, and Metastasis. *Cell*, 124(2):263–266.

Cooper, C. S., Eeles, R., Wedge, D. C., Van Loo, P., Gundem, G., Alexandrov, L. B., Kremeyer, B., Butler, A., Lynch, A. G., Camacho, N., Massie, C. E., Kay, J., Luxton, H. J., Edwards, S., Kote-Jarai, Z., Dennis, N., Merson, S., Leongamornlert, D., Zamora, J., Corbishley, C., Thomas, S., Nik-Zainal, S., Ramakrishna, M., O'Meara, S., Matthews, L., Clark, J., Hurst, R., Mithen, R., Bristow, R. G., Boutros, P. C., Fraser, M., Cooke, S., Raine, K., Jones, D., Menzies, A., Stebbings, L., Hinton, J., Teague, J., McLaren, S., Mudie, L., Hardy, C., Anderson, E., Joseph, O., Goody, V., Robinson, B., Maddison, M., Gamble, S., Greenman, C., Berney, D., Hazell, S., Livni, N., the ICGC Prostate Group, Fisher, C., Ogden, C., Kumar, P., Thompson, A., Woodhouse, C., Nicol, D., Mayer, E., Dudderidge, T., Shah, N. C., Gnanapragasam, V., Voet, T., Campbell, P., Futreal, A., Easton, D., Warren, A. Y., Foster, C. S., Stratton, M. R., Whitaker, H. C., McDermott, U., Brewer, D. S., and Neal, D. E. (2015). Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nature Genetics*, 47(4):367–372.

Cun, Y., Yang, T.-P., Achter, V., Lang, U., and Peifer, M. (2018). Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. *Nature Protocols*, 13(6):1488–1501.

Dagogo-Jack, I. and Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81–94.

Dalla-Favera, R., Bregni, M., Erikson, J., Patterson, D., Gallo, R. C., and Croce, C. M. (1982). Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. *Proceedings of the National Academy of Sciences*, 79(24):7824–7827.

Daly, B. and Olopade, O. I. (2015). A perfect storm: How tumor biology, genomics, and health care delivery patterns collide to create a racial survival disparity in breast cancer and proposed interventions for change. *CA: A Cancer Journal for Clinicians*, 65(3):221–238.

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J., Bottomley, W., Davis, N., Dicks, E., Ewing, R., Floyd, Y., Gray, K., Hall, S., Hawes, R., Hughes, J., Kosmidou, V., Menzies, A., Mould, C., Parker, A., Stevens, C., Watt, S., Hooper, S., Wilson, R., Jayatilake, H., Gusterson, B. A., Cooper, C., Shipley, J., Hargrave, D., Pritchard-Jones, K., Maitland, N., Chenevix-Trench, G., Riggins, G. J., Bigner, D. D., Palmieri, G., Cossu, A., Flanagan, A., Nicholson, A., Ho, J. W. C., Leung, S. Y., Yuen, S. T., Weber, B. L., Seigler, H. F., Darrow, T. L., Paterson, H., Marais, R., Marshall, C. J., Wooster, R., Stratton, M. R., and Futreal, P. A. (2002). Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949.

de Bruin, E. C., McGranahan, N., Mitter, R., Salm, M., Wedge, D. C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A. J., Grönroos, E., Muhammad, M. A., Horswell, S., Gerlinger, M., Varela, I., Jones, D., Marshall, J., Voet, T., Loo, P. V., Rassl, D. M., Rintoul, R. C., Janes, S. M., Lee, S.-M., Forster, M., Ahmad, T., Lawrence, D., Falzon, M., Capitanio, A., Harkins, T. T., Lee, C. C., Tom, W., Teefe, E., Chen, S.-C., Begum, S., Rabinowitz, A., Phillimore, B., Spencer-Dene, B., Stamp, G., Szallasi, Z., Matthews, N., Stewart, A., Campbell, P., and Swanton, C. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, 346(6206):251–256.

de Grouchy, J., de Nava, C., Cantu, J. M., Bilski-Pasquier, G., and Bousser, J. (1966). Models for clonal evolutions: A study of chronic myelogenous leukemia. *American Journal of Human Genetics*, 18(5):485–503.

Denkert, C., von Minckwitz, G., Darb-Esfahani, S., Lederer, B., Heppner, B. I., Weber, K. E., Budczies, J., Huober, J., Klauschen, F., Furlanetto, J., Schmitt, W. D., Blohmer, J.-U., Karn, T., Pfitzner, B. M., Kümmel, S., Engels, K., Schneeweiss, A., Hartmann, A., Noske, A., Fasching, P. A., Jackisch, C., van Mackelenbergh, M., Sinn, P., Schem, C., Hanusch, C., Untch, M., and Loibl, S. (2018). Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: A pooled analysis of 3771 patients treated with neoadjuvant therapy. *The Lancet Oncology*, 19(1):40–50.

Dentro, S. C., Wedge, D. C., and Van Loo, P. (2017). Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harbor Perspectives in Medicine*.

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., and Morris, Q. (2015). PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16:35.

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B., Fulton, L., Fulton, R. S., Zhang, Q., Wendl, M. C., Lawrence, M. S., Larson, D. E., Chen, K., Dooling, D. J., Sabo, A., Hawes, A. C., Shen, H., Jhangiani, S. N., Lewis, L. R., Hall, O., Zhu, Y., Mathew, T., Ren, Y., Yao, J., Scherer, S. E., Clerc, K., Metcalf, G. A., Ng, B., Milosavljevic, A., Gonzalez-Garay, M. L., Osborne, J. R., Meyer, R., Shi, X., Tang, Y., Koboldt, D. C., Lin, L., Abbott, R., Miner, T. L., Pohl, C., Fewell, G., Haipek, C., Schmidt, H., Dunford-Shore, B. H., Kraja, A., Crosby, S. D., Sawyer, C. S., Vickery, T., Sander, S., Robinson, J., Winckler, W., Baldwin, J., Chirieac, L. R., Dutt, A., Fennell, T., Hanna, M., Johnson, B. E., Onofrio, R. C., Thomas, R. K., Tonon, G., Weir, B. A., Zhao, X., Ziaugra, L., Zody, M. C., Giordano, T., Orringer, M. B., Roth, J. A., Spitz, M. R., Wistuba, I. I., Ozenberger, B., Good, P. J., Chang, A. C., Beer, D. G., Watson, M. A., Ladanyi, M., Broderick, S., Yoshizawa, A., Travis, W. D., Pao, W., Province, M. A., Weinstock, G. M., Varmus, H. E., Gabriel, S. B., Lander, E. S., Gibbs, R. A., Meyerson, M., and Wilson, R. K. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216):1069.

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M., and Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research*, 36(19):e126–e126.

Donmez, N., Malikic, S., Wyatt, A. W., Gleave, M. E., Collins, C. C., and Sahinalp, S. C. (2016). Clonality Inference from Single Tumor Samples Using Low Coverage Sequence Data. In *Research in Computational Molecular Biology*, Lecture Notes in Computer Science, pages 83–94. Springer, Cham.

Dulak, A. M., Stojanov, P., Peng, S., Lawrence, M. S., Fox, C., Stewart, C., Bandla, S., Imamura, Y., Schumacher, S. E., Shefler, E., McKenna, A., Carter, S. L., Cibulskis, K., Sivachenko, A., Saksena, G., Voet, D., Ramos, A. H., Auclair, D., Thompson, K., Sougnez, C., Onofrio, R. C., Guiducci, C., Beroukhim, R., Zhou, Z., Lin, L., Lin, J., Reddy, R., Chang, A., Landrenau, R., Pennathur, A., Ogino, S., Luketich, J. D., Golub, T. R., Gabriel, S. B., Lander, E. S., Beer, D. G., Godfrey, T. E., Getz, G., and Bass, A. J. (2013). Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature Genetics*, 45(5):478.

Duncan, J. S., Whittle, M. C., Nakamura, K., Abell, A. N., Midland, A. A., Zawistowski, J. S., Johnson, N. L., Granger, D. A., Jordan, N. V., Darr, D. B., Usary, J., Kuan, P.-F., Smalley, D. M., Major, B., He, X., Hoadley, K. A., Zhou, B., Sharpless, N. E., Perou, C. M., Kim, W. Y., Gomez, S. M., Chen, X., Jin, J., Frye, S. V., Earp, H. S., Graves, L. M., and Johnson, G. L. (2012). Dynamic Reprogramming of the Kinome in Response to Targeted MEK Inhibition in Triple-Negative Breast Cancer. *Cell*, 149(2):307–321.

Dunson, D. B. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Ellis, M. J., Ding, L., Shen, D., Luo, J., Suman, V. J., Wallis, J. W., Tine, B. A. V., Hoog, J., Goiffon, R. J., Goldstein, T. C., Ng, S., Lin, L., Crowder, R., Snider, J., Ballman, K., Weber, J., Chen, K., Koboldt, D. C., Kandoth, C., Schierding, W. S., McMichael, J. F., Miller, C. A., Lu, C., Harris, C. C., McLellan, M. D., Wendl, M. C., DeSchryver, K., Allred, D. C., Esserman, L., Unzeitig, G., Margenthaler, J., Babiera, G. V., Marcom, P. K., Guenther, J. M., Leitch, M., Hunt, K., Olson, J., Tao, Y., Maher, C. A., Fulton, L. L., Fulton, R. S., Harrison, M., Oberkfell, B., Du, F., Demeter, R., Vickery, T. L., Elhammali, A., Piwnica-Worms, H., McDonald, S., Watson, M., Dooling, D. J., Ota, D., Chang, L.-W., Bose, R., Ley, T. J., Piwnica-Worms, D., Stuart, J. M., Wilson, R. K., and Mardis, E. R. (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*, 486(7403):353.

Erez, N., Truitt, M., Olson, P., and Hanahan, D. (2010). Cancer-Associated Fibroblasts Are Activated in Incipient Neoplasia to Orchestrate Tumor-Promoting Inflammation in an NF-κB-Dependent Manner. *Cancer Cell*, 17(2):135–147.

Ewing, A. D., Houlahan, K. E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T. N., Bare, J. C., P'ng, C., Waggott, D., Sabelnykova, V. Y., Participants, I.-T. D. S. M. C. C., Xi, L., Dewal, N., Fan, Y., Wang, W., Wheeler, D., Wilm, A., Ting, G. H., Li, C., Bertrand, D., Nagarajan, N., Chen, Q.-R., Hsu, C.-H., Hu, Y., Yan, C., Kibbe, W., Meerzaman, D., Cibulskis, K., Rosenberg, M., Bergelson, L., Kiezun, A., Radenbaugh, A., Sertier, A.-S., Ferrari, A., Tonton, L., Bhutani, K., Hansen, N. F., Wang, D., Song, L., Lai, Z., Liao, Y., Shi, W., Carbonell-Caballero, J., Dopazo, J., Lau, C. C. K., Guinney, J., Kellen, M. R., Norman, T. C., Haussler, D., Friend, S. H., Stolovitzky, G., Margolin, A. A., Stuart, J. M., and Boutros, P. C. (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods*, 12(7):623.

Fackenthal, J. D., Zhang, J., Zhang, B., Zheng, Y., Hagos, F., Burrill, D. R., Niu, Q., Huo, D., Sveen, W. E., Ogundiran, T., Adebamowo, C., Odetunde, A., Falusi, A. G., and Olopade, O. I. (2012). High prevalence of BRCA1 and BRCA2 mutations in unselected Nigerian breast cancer patients. *International Journal of Cancer*, 131(5):1114–1123.

Fearon, E. R. and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, 136(5):E359–E386.

Fischer, A., Vázquez-García, I., Illingworth, C. J. R., and Mustonen, V. (2014). High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Reports*, 7(5):1740–1752.

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., Kok, C. Y., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T., and Campbell, P. J. (2017). COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783.

Ford, C. and Clarke, C. (1963). Cytogenetic evidence of clonal proliferation in primary reticular neoplasms. *Proceedings. Canadian Cancer Conference*, 5:129–146.

Foulds, L. (1957). Tumor Progression: Guest Editorial. *Cancer Research*, 17(5):355–356.

Franklin, R. E. and Gosling, R. G. (1953). Molecular Configuration in Sodium Thymonucleate. *Nature*, 171(4356):740.

Fridman, W. H., Pagès, F., Sautès-Fridman, C., and Galon, J. (2012). The immune contexture in human tumours: Impact on clinical outcome. *Nature Reviews Cancer*, 12(4):298–306.

Fritsch, A. and Ickstadt, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2):367–391.

Fujimoto, A., Totoki, Y., Abe, T., Boroevich, K. A., Hosoda, F., Nguyen, H. H., Aoki, M., Hosono, N., Kubo, M., Miya, F., Arai, Y., Takahashi, H., Shirakihara, T., Nagasaki, M., Shibuya, T., Nakano, K., Watanabe-Makino, K., Tanaka, H., Nakamura, H., Kusuda, J., Ojima, H., Shimada, K., Okusaka, T., Ueno, M., Shigekawa, Y., Kawakami, Y., Arihiro, K., Ohdan, H., Gotoh, K., Ishikawa, O., Ariizumi, S.-i., Yamamoto, M., Yamada, T., Chayama, K., Kosuge, T., Yamaue, H., Kamatani, N., Miyano, S., Nakagama, H., Nakamura, Y., Tsunoda, T., Shibata, T., and Nakagawa, H. (2012). Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nature Genetics*, 44(7):760.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183.

Gainor, J. F., Dardaei, L., Yoda, S., Friboulet, L., Leshchiner, I., Katayama, R., Dagogo-Jack, I., Gadgeel, S., Schultz, K., Singh, M., Chin, E., Parks, M., Lee, D., DiCecca, R. H., Lockerman, E., Huynh, T., Logan, J., Ritterhouse, L. L., Le, L. P., Muniappan, A., Digumarthy, S., Channick, C., Keyes, C., Getz, G., Dias-Santagata, D., Heist, R. S., Lennerz, J., Sequist, L. V., Benes, C. H., Iafrate, A. J., Mino-Kenudson, M., Engelman, J. A., and Shaw, A. T. (2016). Molecular Mechanisms of Resistance to First- and Second-Generation ALK Inhibitors in ALK-Rearranged Lung Cancer. *Cancer Discovery*, 6(10):1118–1133.

Galdiero, M. R., Bonavita, E., Barajon, I., Garlanda, C., Mantovani, A., and Jaillon, S. (2013). Tumor associated macrophages and neutrophils in cancer. *Immunobiology*, 218(11):1402–1410.

Gerlinger, M., Horswell, S., Larkin, J., Rowan, A. J., Salm, M. P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C. R., Martinez, P., Phillimore, B., Begum, S., Rabinowitz, A., Spencer-Dene, B., Gulati, S., Bates, P. A., Stamp, G., Pickering, L., Gore, M., Nicol, D. L., Hazell, S., Futreal, P. A., Stewart, A., and Swanton, C. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics*, 46(3):225–233.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A., and Swanton, C. (2012). Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine*, 366(10):883–892.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., Mitchell, T. J., Rubanova, Y., Anur, P., Rosebrock, D., Yu, K., Tarabichi, M., Deshwar, A., Wintersinger, J., Kleinheinz, K., Vazquez-Garcia, I., Haase, K., Sengupta, S., Macintyre, G., Malikic, S., Donmez, N., Livitz, D. G., Cmero, M., Demeulemeester, J., Schumacher, S., Fan, Y., Yao, X., Lee, J., Schlesner, M., Boutros, P. C., Bowtell, D. D., Zhu, H., Getz, G., Imielinski, M., Beroukhim, R., Sahinalp, S. C., Ji, Y., Peifer, M., Markowetz, F., Mustonen, V., Yuan, K., Wang, W., Morris, Q. D., Spellman, P. T., Wedge, D. C., Van Loo, P., PCAWG Evolution and Heterogeneity Working Group, and PCAWG Network (2017). The evolutionary history of 2,658 cancers. *bioRxiv*, page 161562.

Ghahramani, Z., Jordan, M. I., and Adams, R. P. (2010). Tree-Structured Stick Breaking for Hierarchical Data. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 19–27. Curran Associates, Inc.

Gotwals, P., Cameron, S., Cipolletta, D., Cremasco, V., Crystal, A., Hewes, B., Mueller, B., Quaratino, S., Sabatos-Peyton, C., Petruzzelli, L., Engelman, J. A., and Dranoff, G. (2017). Prospects for combining targeted and conventional cancer therapy with immunotherapy. *Nature Reviews Cancer*, 17(5):286–301.

Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R., and Easton, D. F. (2006). Statistical Analysis of Pathogenicity of Somatic Mutations in Cancer. *Genetics*, 173(4):2187–2198.

Greenman, C. D., Pleasance, E. D., Newman, S., Yang, F., Fu, B., Nik-Zainal, S., Jones, D., Lau, K. W., Carter, N., Edwards, P. A. W., Futreal, P. A., Stratton, M. R., and Campbell, P. J. (2012). Estimation of rearrangement phylogeny for cancer genomes. *Genome Research*, 22(2):346–361.

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L. B., Tubio, J. M. C., Papaemmanuil, E., Brewer, D. S., Kallio, H. M. L., Högnäs, G., Annala, M., Kivinummi, K., Goody, V., Latimer, C., O'Meara, S., Dawson, K. J., Isaacs, W., Emmert-Buck, M. R., Nykter, M., Foster, C., Kote-Jarai, Z., Easton, D., Whitaker, H. C., ICGC Prostate UK Group, Neal, D. E., Cooper, C. S., Eeles, R. A., Visakorpi, T., Campbell, P. J., McDermott, U., Wedge, D. C., and Bova, G. S. (2015). The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357.

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Gilks, C. B., Huntsman, D. G., McAlpine, J. N., Aparicio, S., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*, 24(11):1881–1893.

Haber, D. A., Buckler, A. J., Glaser, T., Call, K. M., Pelletier, J., Sohn, R. L., Douglass, E. C., and Housman, D. E. (1990). An internal deletion within an 11p13 zinc finger gene contributes to the development of Wilms' tumor. *Cell*, 61(7):1257–1269.

Hanahan, D. and Coussens, L. M. (2012). Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. *Cancer Cell*, 21(3):309–322.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674.

Hauschka, T. S. (1961). The Chromosomes in Ontogeny and Oncogeny. *Cancer Research*, 21(8):957.

Holohan, C., Van Schaeybroeck, S., Longley, D. B., and Johnston, P. G. (2013). Cancer drug resistance: An evolving paradigm. *Nature Reviews Cancer*, 13(10):714–726.

Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., Schadendorf, D., and Kumar, R. (2013). TERT Promoter Mutations in Familial and Sporadic Melanoma. *Science*, 339(6122):959–961.

Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., and Garraway, L. A. (2013). Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science*, 339(6122):957–959.

ICGC Consortium (2010). International network of cancer genome projects. *Nature*, 464(7291):993–998.

Iqbal, J., Ginsburg, O., Rochon, P. A., Sun, P., and Narod, S. A. (2015). Differences in Breast Cancer Stage at Diagnosis and Cancer-Specific Survival by Race and Ethnicity in the United States. *JAMA*, 313(2):165–173.

Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B., Veeriah, S., Shafi, S., Johnson, D. H., Mitter, R., Rosenthal, R., Salm, M., Horswell, S., Escudero, M., Matthews, N., Rowan, A., Chambers, T., Moore, D. A., Turajlic, S., Xu, H., Lee, S.-M., Forster, M. D., Ahmad, T., Hiley, C. T., Abbosh, C., Falzon, M., Borg, E., Marafioti, T., Lawrence, D., Hayward, M., Kolvekar, S., Panagiotopoulos, N., Janes, S. M., Thakrar, R., Ahmed, A., Blackhall, F., Summers, Y., Shah, R., Joseph, L., Quinn, A. M., Crosbie, P. A., Naidu, B., Middleton, G., Langman, G., Trotter, S., Nicolson, M., Remmen, H., Kerr, K., Chetty, M., Gomersall, L., Fennell, D. A., Nakas, A., Rathinam, S., Anand, G., Khan, S., Russell, P., Ezhil, V., Ismail, B., Irvin-Sellers, M., Prakash, V., Lester, J. F., Kornaszewska, M., Attanoos, R., Adams, H., Davies, H., Dentro, S., Taniere, P., O'Sullivan, B., Lowe, H. L., Hartley, J. A., Iles, N., Bell, H., Ngai, Y., Shaw, J. A., Herrero, J., Szallasi, Z., Schwarz, R. F., Stewart, A., Quezada, S. A., Le Quesne, J., Van Loo, P., Dive, C., Hackshaw, A., and Swanton, C. (2017). Tracking the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine*.

Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., and Morris, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15(1):35.

Jones, C. E., Maben, J., Jack, R. H., Davies, E. A., Forbes, L. J., Lucas, G., and Ream, E. (2014). A systematic review of barriers to early presentation and diagnosis with breast cancer among black women. *BMJ Open*, 4(2):e004076.

Joyce, J. A. and Fearon, D. T. (2015). T cell exclusion, immune privilege, and the tumor microenvironment. *Science*, 348(6230):74–80.

Jr, L. A. D., Williams, R. T., Wu, J., Kinde, I., Hecht, J. R., Berlin, J., Allen, B., Bozic, I., Reiter, J. G., Nowak, M. A., Kinzler, K. W., Oliner, K. S., and Vogelstein, B. (2012). The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*, 486(7404):537–540.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (New York, N.Y.)*, 258(5083):818–821.

Kalluri, R. and Zeisberg, M. (2006). Fibroblasts in cancer. *Nature Reviews Cancer*, 6(5):392–401.

Klein, G. (1987). The Approaching Era of the Tumor Suppressor Genes. *Science*, 238(4833):1539–1545.

Kleinheinz, K., Bludau, I., Huebschmann, D., Heinold, M., Kensche, P., Gu, Z., Lopez, C., Hummel, M., Klapper, W., Moeller, P., Vater, I., Wagener, R., Project, I. M.-S., Brors, B., Siebert, R., Eils, R., and Schlesner, M. (2017). ACEseq - allele specific copy number estimation from whole genome sequencing. *bioRxiv*, page 210807.

Knudson, A. G. (1971). Mutation and cancer: Statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 68(4):820–823.

Knudson, A. G. (1993). Antioncogenes and human cancer. *Proceedings of the National Academy of Sciences*, 90(23):10914–10921.

Koren, A., Polak, P., Nemesh, J., Michaelson, J. J., Sebat, J., Sunyaev, S. R., and McCarroll, S. A. (2012). Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. *The American Journal of Human Genetics*, 91(6):1033–1040.

Kris, M. G., Johnson, B. E., Berry, L. D., Kwiatkowski, D. J., Iafrate, A. J., Wistuba, I. I., Varella-Garcia, M., Franklin, W. A., Aronson, S. L., Su, P.-F., Shyr, Y., Camidge, D. R., Sequist, L. V., Glisson, B. S., Khuri, F. R., Garon, E. B., Pao, W., Rudin, C., Schiller, J., Haura, E. B., Socinski, M., Shirai, K., Chen, H., Giaccone, G., Ladanyi, M., Kugler, K., Minna, J. D., and Bunn, P. A. (2014). Using Multiplexed Assays of Oncogenic Drivers in Lung Cancers to Select Targeted Drugs. *JAMA*, 311(19):1998–2006.

Kwak, E. L., Ahronian, L. G., Siravegna, G., Mussolin, B., Borger, D. R., Godfrey, J. T., Jessop, N. A., Clark, J. W., Blaszkowsky, L. S., Ryan, D. P., Lennerz, J. K., Iafrate, A. J., Bardelli, A., Hong, T. S., and Corcoran, R. B. (2015). Molecular Heterogeneity and Receptor Coamplification Drive Resistance to Targeted Therapy in MET-Amplified Esophagogastric Cancer. *Cancer Discovery*, 5(12):1271–1281.

Landau, D. A., Carter, S. L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M. S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., Wan, Y., Zhang, W., Shukla, S. A., Vartanov, A., Fernandes, S. M., Saksena, G., Cibulskis, K., Tesar, B., Gabriel, S., Hacohen, N., Meyerson, M., Lander, E. S., Neuberg, D., Brown, J. R., Getz, G., and Wu, C. J. (2013). Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell*, 152(4):714–726.

Landau, D. A., Tausch, E., Taylor-Weiner, A. N., Stewart, C., Reiter, J. G., Bahlo, J., Kluth, S., Bozic, I., Lawrence, M., Böttcher, S., Carter, S. L., Cibulskis, K., Mertens, D., Sougnez, C. L., Rosenberg, M., Hess, J. M., Edelmann, J., Kless, S., Kneba, M., Ritgen, M., Fink, A., Fischer, K., Gabriel, S., Lander, E. S., Nowak, M. A., Döhner, H., Hallek, M., Neuberg, D., Getz, G., Stilgenbauer, S., and Wu, C. J. (2015). Mutations driving CLL and their evolution in progression and relapse. *Nature*, 526(7574):525–530.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol*, 9(8):e1003118.

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):R84.

Levan, A. and Biesele, J. J. (1958). Role of Chromosomes in Cancerogenesis, as Studied in Serial Tissue Culture of Mammalian Cells. *Annals of the New York Academy of Sciences*, 71(6):1022–1053.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Liu, W., Dowling, J. P., Murray, W. K., McArthur, G. A., Thompson, J. F., Wolfe, R., and Kelly, J. W. (2006). Rate of Growth in Melanomas: Characteristics and Associations of Rapidly Growing Melanomas. *Archives of Dermatology*, 142(12):1551–1558.

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., der Aa, N. V., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., and Voet, T. (2015). G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12(6):519.

Makino, S. (1957). The Chromosome Cytology of the Ascites Tumors of Rats, with Special Reference to the Concept of the Stemline Cell. In Bourne, G. H. and Danielli, J. F., editors, *International Review of Cytology*, volume 6, pages 25–84. Academic Press.

Makohon-Moore, A. P., Matsukuma, K., Zhang, M., Reiter, J. G., Gerold, J. M., Jiao, Y., Sikkema, L., Attiyeh, M. A., Yachida, S., Sandone, C., Hruban, R. H., Klimstra, D. S., Papadopoulos, N., Nowak, M. A., Kinzler, K. W., Vogelstein, B., and Iacobuzio-Donahue, C. A. (2018). Precancerous neoplastic cells can move through the pancreatic ductal system. *Nature*, 561(7722):201–205.

Mamlouk, S., Childs, L. H., Aust, D., Heim, D., Melching, F., Oliveira, C., Wolf, T., Durek, P., Schumacher, D., Bläker, H., von Winterfeld, M., Gastl, B., Möhr, K., Menne, A., Zeugner, S., Redmer, T., Lenze, D., Tierling, S., Möbs, M., Weichert, W., Folprecht, G., Blanc, E., Beule, D., Schäfer, R., Morkel, M., Klauschen, F., Leser, U., and Sers, C. (2017). DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer. *Nature Communications*, 8.

Mantovani, A., Marchesi, F., Malesci, A., Laghi, L., and Allavena, P. (2017). Tumour-associated macrophages as treatment targets in oncology. *Nature Reviews Clinical Oncology*, 14(7):399–416.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

Martincorena, I., Fowler, J. C., Wabik, A., Lawson, A. R. J., Abascal, F., Hall, M. W. J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M. R., Fitzgerald, R. C., Handford, P. A., Campbell, P. J., Saeb-Parsy, K., and Jones, P. H. (2018). Somatic mutant clones colonize the human esophagus with age. *Science*, page eaau3879.

Martincorena, I., Jones, P. H., and Campbell, P. J. (2016). Constrained positive selection on cancer mutations in normal skin. *Proceedings of the National Academy of Sciences*, 113(9):E1128–E1129.

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Loo, P. V., Davies, H., Stratton, M. R., and Campbell, P. J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, 171(5):1029–1041.e21.

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: A looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334.

Mattos-Arruda, L. D., Cortes, J., Santarpia, L., Vivancos, A., Tabernero, J., Reis-Filho, J. S., and Seoane, J. (2013). Circulating tumour cells and cell-free DNA as tools for managing breast cancer. *Nature Reviews Clinical Oncology*, 10(7):377–389.

McGranahan, N., Furness, A. J. S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., Jamal-Hanjani, M., Wilson, G. A., Birkbak, N. J., Hiley, C. T., Watkins, T. B. K., Shafi, S., Murugaesu, N., Mitter, R., Akarca, A. U., Linares, J., Marafioti, T., Henry, J. Y., Allen, E. M. V., Miao, D., Schilling, B., Schadendorf, D., Garraway, L. A., Makarov, V., Rizvi, N. A., Snyder, A., Hellmann, M. D., Merghoub, T., Wolchok, J. D., Shukla, S. A., Wu, C. J., Peggs, K. S., Chan, T. A., Hadrup, S. R., Quezada, S. A., and Swanton, C. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, page aaf1490.

McGranahan, N., Rosenthal, R., Hiley, C. T., Rowan, A. J., Watkins, T. B. K., Wilson, G. A., Birkbak, N. J., Veeriah, S., Van Loo, P., Herrero, J., and Swanton, C. (2017). Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*, 171(6):1259–1271.e11.

McGranahan, N. and Swanton, C. (2015a). Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell*, 27(1):15–26.

McGranahan, N. and Swanton, C. (2015b). Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell*, 27(1):15–26.

Miller, C. A., White, B. S., Dees, N. D., Griffith, M., Welch, J. S., Griffith, O. L., Vij, R., Tomasson, M. H., Graubert, T. A., Walter, M. J., Ellis, M. J., Schierding, W., DiPersio, J. F., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. (2014). SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Comput Biol*, 10(8):e1003665.

Misale, S., Nicolantonio, F. D., Sartore-Bianchi, A., Siena, S., and Bardelli, A. (2014). Resistance to Anti-EGFR Therapy in Colorectal Cancer: From Heterogeneity to Convergent Evolution. *Cancer Discovery*, 4(11):1269–1280.

Morson, B. C. (1974). Evolution of cancer of the colon and rectum. *Cancer*, 34(S3):845–849.

Murphree, A. L. and Benedict, W. F. (1984). Retinoblastoma: Clues to human oncogenesis. *Science*, 223(4640):1028–1033.

Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D. K., Kennedy, G. C., and Ogawa, S. (2005). A Robust Algorithm for Copy Number Detection Using High-Density Oligonucleotide Single Nucleotide Polymorphism Genotyping Arrays. *Cancer Research*, 65(14):6071–6079.

Narod, S. A., Iqbal, J., and Miller, A. B. (2015). Why have breast cancer mortality rates declined? *Journal of Cancer Policy*, 5(Supplement C):8–17.

Nazarian, R., Shi, H., Wang, Q., Kong, X., Koya, R. C., Lee, H., Chen, Z., Lee, M.-K., Attar, N., Sazegar, H., Chodon, T., Nelson, S. F., McArthur, G., Sosman, J. A., Ribas, A., and Lo, R. S. (2010). Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature*, 468(7326):973–977.

Nigro, J. M., Baker, S. J., Preisinger, A. C., Jessup, J. M., Hosteller, R., Cleary, K., Signer, S. H., Davidson, N., Baylin, S., Devilee, P., Glover, T., Collins, F. S., Weslon, A., Modali, R., Harris, C. C., and Vogelstein, B. (1989). Mutations in the p53 gene occur in diverse human tumour types. *Nature*, 342(6250):705.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K. W., Mudie, L. J., Varela, I., McBride, D. J., Bignell, G. R., Cooke, S. L., Shlien, A., Gamble, J., Whitmore, I., Maddison, M., Tarpey, P. S., Davies, H. R., Papaemmanuil, E., Stephens, P. J., McLaren, S., Butler, A. P., Teague, J. W., Jönsson, G., Garber, J. E., Silver, D., Miron, P., Fatima, A., Boyault, S., Langerød, A., Tutt, A., Martens, J. W., Aparicio, S. A., Borg, Å., Salomon, A. V., Thomas, G., Børresen-Dale, A.-L., Richardson, A. L., Neuberger, M. S., Futreal, P. A., Campbell, P. J., and Stratton, M. R. (2012a). Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, 149(5):979–993.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjærde, O. C., Langerød, A., Ringnér, M., Ahn, S.-M., Boyault, S., Brock, J. E., Broeks, A., Butler, A., Desmedt, C., Dirix, L., Dronov, S., Fatima, A., Foekens, J. A., Gerstung, M., Hooijer, G. K. J., Jang, S. J., Jones, D. R., Kim, H.-Y., King, T. A., Krishnamurthy, S., Lee, H. J., Lee, J.-Y., Li, Y., McLaren, S., Menzies, A., Mustonen, V., O'Meara, S., Pauporté, I., Pivot, X., Purdie, C. A., Raine, K., Ramakrishnan, K., Rodríguez-González, F. G., Romieu, G., Sieuwerts, A. M., Simpson, P. T., Shepherd, R., Stebbings, L., Stefansson, O. A., Teague, J., Tommasi, S., Treilleux, I., Van den Eynden, G. G., Vermeulen, P., Vincent-Salomon, A., Yates, L., Caldas, C., van't Veer, L., Tutt, A., Knappskog, S., Tan, B. K. T., Jonkers, J., Borg, A., Ueno, N. T., Sotiriou, C., Viari, A., Futreal, P. A., Campbell, P. J., Span, P. N., Van Laere, S., Lakhani, S. R., Eyfjord, J. E., Thompson, A. M., Birney, E., Stunnenberg, H. G., van de Vijver, M. J., Martens, J. W. M., Børresen-Dale, A.-L., Richardson, A. L., Kong, G., Thomas, G., and Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., Gamble, S. J., Stephens, P. J., McLaren, S., Tarpey, P. S., Papaemmanuil, E., Davies, H. R., Varela, I., McBride, D. J., Bignell, G. R., Leung, K., Butler, A. P., Teague, J. W., Martin, S., Jönsson, G., Mariani, O., Boyault, S., Miron, P., Fatima, A., Langerød, A., Aparicio, S. A. J. R., Tutt, A., Sieuwerts, A. M., Borg, A., Thomas, G., Salomon, A. V., Richardson, A. L., Børresen-Dale, A.-L., Futreal, P. A., Stratton, M. R., Campbell, P. J., and Breast Cancer Working Group of the International Cancer Genome Consortium (2012b). The Life History of 21 Breast Cancers. *Cell*, 149(5):994–1007.

Nilsen, G., Liestøl, K., Van Loo, P., Moen Vollan, H. K., Eide, M. B., Rueda, O. M., Chin, S.-F., Russell, R., Baumbusch, L. O., Caldas, C., Børresen-Dale, A.-L., and Lingjærde, O. C. (2012). Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics*, 13:591.

Nishisho, I., Nakamura, Y., Miyoshi, Y., Miki, Y., Ando, H., Horii, A., Koyama, K., Utsunomiya, J., Baba, S., and Hedge, P. (1991). Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science (New York, N.Y.)*, 253(5020):665–669.

Noorbakhsh, J. and Chuang, J. H. (2017). Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. *Nature Genetics*, 49(9):1288–1289.

Nordling, C. O. (1953). A New Theory on the Cancer-inducing Mechanism. *British Journal of Cancer*, 7(1):68–72.

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

Nowell, P. C. (1986). Mechanisms of tumor progression. *Cancer Research*, 46(5):2203–2207.

Nowell, P. C. and Hungerford, D. (1960). A minute chromosome in human chronic granulocytic leukemia.

Orimo, A., Gupta, P. B., Sgroi, D. C., Arenzana-Seisdedos, F., Delaunay, T., Naeem, R., Carey, V. J., Richardson, A. L., and Weinberg, R. A. (2005). Stromal Fibroblasts Present in Invasive Human Breast Carcinomas Promote Tumor Growth and Angiogenesis through Elevated SDF-1/CXCL12 Secretion. *Cell*, 121(3):335–348.

Parikh, A. R., Leshchiner, I., Elagina, L., Goyal, L., Levovitz, C., Siravegna, G., Livitz, D., Rhrissorrakrai, K., Martin, E. E., Seventer, E. E. V., Hanna, M., Slowik, K., Utro, F., Pinto, C. J., Wong, A., Danysh, B. P., de la Cruz, F. F., Fetter, I. J., Nadres, B., Shahzade, H. A., Allen, J. N., Blaszkowsky, L. S., Clark, J. W., Giantonio, B., Murphy, J. E., Nipp, R. D., Roeland, E., Ryan, D. P., Weekes, C. D., Kwak, E. L., Faris, J. E., Wo, J. Y., Aguet, F., Dey-Guha, I., Hazar-Rethinam, M., Dias-Santagata, D., Ting, D. T., Zhu, A. X., Hong, T. S., Golub, T. R., Iafrate, A. J., Adalsteinsson, V. A., Bardelli, A., Parida, L., Juric, D., Getz, G., and Corcoran, R. B. (2019). Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. *Nature Medicine*, 25(9):1415–1421.

Patch, A.-M., Christie, E. L., Etemadmoghadam, D., Garsed, D. W., George, J., Fereday, S., Nones, K., Cowin, P., Alsop, K., Bailey, P. J., Kassahn, K. S., Newell, F., Quinn, M. C. J., Kazakoff, S., Quek, K., Wilhelm-Benartzi, C., Curry, E., Leong, H. S., The Australian Ovarian Cancer Study Group, Hamilton, A., Mileshkin, L., Au-Yeung, G., Kennedy, C., Hung, J., Chiew, Y.-E., Harnett, P., Friedlander, M., Quinn, M., Pyman, J., Cordner, S., O'Brien, P., Leditschke, J., Young, G., Strachan, K., Waring, P., Azar, W., Mitchell, C., Traficante, N., Hendley, J., Thorne, H., Shackleton, M., Miller, D. K., Arnau, G. M., Tothill, R. W., Holloway, T. P., Semple, T., Harliwong, I., Nourse, C., Nourbakhsh, E., Manning, S., Idrisoglu, S., Bruxner, T. J. C., Christ, A. N., Poudel, B., Holmes, O., Anderson, M., Leonard, C., Lonie, A., Hall, N., Wood, S., Taylor, D. F., Xu, Q., Fink, J. L., Waddell, N., Drapkin, R., Stronach, E., Gabra, H., Brown, R., Jewell, A., Nagaraj, S. H., Markham, E., Wilson, P. J., Ellul, J., McNally, O., Doyle, M. A., Vedururu, R., Stewart, C., Lengyel, E., Pearson, J. V., Waddell, N., deFazio, A., Grimmond, S. M., and Bowtell, D. D. L. (2015). Whole–genome characterization of chemoresistant ovarian cancer. *Nature*, 521(7553):489–494.

Pegoraro, L., Palumbo, A., Erikson, J., Falda, M., Giovanazzo, B., Emanuel, B. S., Rovera, G., Nowell, P. C., and Croce, C. M. (1984). A 14;18 and an 8;14 chromosome translocation in a cell line derived from an acute B-cell leukemia. *Proceedings of the National Academy of Sciences*, 81(22):7166–7170.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A.-L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747.

Pfeifer, D., Pantic, M., Skatulla, I., Rawluk, J., Kreutz, C., Martens, U. M., Fisch, P., Timmer, J., and Veelken, H. (2007). Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. *Blood*, 109(3):1202–1210.

Pinkel, D. and Albertson, D. G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, 37(6s):S11.

Pitt, J. J., Riester, M., Zheng, Y., Yoshimatsu, T. F., Sanni, A., Oluwasola, O., Veloso, A., Labrot, E., Wang, S., Odetunde, A., Ademola, A., Okedere, B., Mahan, S., Leary, R., Macomber, M., Ajani, M., Johnson, R. S., Fitzgerald, D., Grundstad, A. J., Tuteja, J. H., Khramtsova, G., Zhang, J., Sveen, E., Hwang, B., Clayton, W., Nkwodimmah, C., Famooto, B., Obasi, E., Aderoju, V., Oludara, M., Omodele, F., Akinyele, O., Adeoye, A., Ogundiran, T., Babalola, C., MacIsaac, K., Popoola, A., Morrissey, M. P., Chen, L. S., Wang, J., Olopade, C. O., Falusi, A. G., Winckler, W., Haase, K., Loo, P. V., Obafunwa, J., Papoutsakis, D., Ojengbede, O., Weber, B., Ibrahim, N., White, K. P., Huo, D., Olopade, O. I., and Barretina, J. (2018). Characterization of Nigerian breast cancer reveals prevalent homologous recombination deficiency and aggressive molecular features. *Nature Communications*, 9(1):1–12.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordóñez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., Kokko-Gonzales, P. I., Gormley, N. A., Grocock, R. J., Haudenschild, C. D., Hims, M. M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L. J., Ning, Z., Royce, T., Schulz-Trieglaff, O. B., Spiridou, A., Stebbings, L. A., Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M. T., Campbell, P. J., Bentley, D. R., Futreal, P. A., and Stratton, M. R. (2009). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196.

Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M.-L., Beare, D., Lau, K. W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H. R., Ordoñez, G. R., Mudie, L. J., Latimer, C., Edkins, S., Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J. W., Mangion, J., Sun, Y. A., McLaughlin, S. F., Peckham, H. E., Tsung, E. F., Costa, G. L., Lee, C. C., Minna, J. D., Gazdar, A., Birney, E., Rhodes, M. D., McKernan, K. J., Stratton, M. R., Futreal, P. A., and Campbell, P. J. (2010). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278):184.

Pollack, J. R., Sørlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Børresen-Dale, A.-L., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, 99(20):12963–12968.

Puente, X. S., Pinyol, M., Quesada, V., Conde, L., Ordóñez, G. R., Villamor, N., Escaramis, G., Jares, P., Beà, S., González-Díaz, M., Bassaganyas, L., Baumann, T., Juan, M., López-Guerra, M., Colomer, D., Tubío, J. M. C., López, C., Navarro, A., Tornador, C., Aymerich, M., Rozman, M., Hernández, J. M., Puente, D. A., Freije, J. M. P., Velasco, G., Gutiérrez-Fernández, A., Costa, D., Carrió, A., Guijarro, S., Enjuanes, A., Hernández, L., Yagüe, J., Nicolás, P., Romeo-Casabona, C. M., Himmelbauer, H., Castillo, E., Dohm, J. C., de Sanjosé, S., Piris, M. A., de Alava, E., Miguel, J. S., Royo, R., Gelpí, J. L., Torrents, D., Orozco, M., Pisano, D. G., Valencia, A., Guigó, R., Bayés, M., Heath, S., Gut, M., Klatt, P., Marshall, J., Raine, K., Stebbings, L. A., Futreal, P. A., Stratton, M. R., Campbell, P. J., Gut, I., López-Guillermo, A., Estivill, X., Montserrat, E., López-Otín, C., and Campo, E. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, 475(7354):101.

Purdom, E., Ho, C., Grasso, C. S., Quist, M. J., Cho, R. J., and Spellman, P. (2013). Methods and challenges in timing chromosomal abnormalities within cancer samples. *Bioinformatics*, 29(24):3113–3120.

Qian, B.-Z. and Pollard, J. W. (2010). Macrophage Diversity Enhances Tumor Progression and Metastasis. *Cell*, 141(1):39–51.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339.

Ray, M. and Polite, B. N. (2010 Jan-Feb). Triple-negative breast cancers: A view from 10,000 feet. *Cancer Journal (Sudbury, Mass.)*, 16(1):17–22.

Reddy, E. P., Reynolds, R. K., Santos, E., and Barbacid, M. (1982). A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature*, 300(5888):149–152.

Ribas, A. and Wolchok, J. D. (2018). Cancer immunotherapy using checkpoint blockade. *Science*, 359(6382):1350–1355.

Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., Lee, W., Yuan, J., Wong, P., Ho, T. S., Miller, M. L., Rekhtman, N., Moreira, A. L., Ibrahim, F., Bruggeman, C., Gasmi, B., Zappasodi, R., Maeda, Y., Sander, C., Garon, E. B., Merghoub, T., Wolchok, J. D., Schumacher, T. N., and Chan, T. A. (2015). Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science*, 348(6230):124–128.

Robbins, P. F., Lu, Y.-C., El-Gamil, M., Li, Y. F., Gross, C., Gartner, J., Lin, J. C., Teer, J. K., Cliften, P., Tycksen, E., Samuels, Y., and Rosenberg, S. A. (2013). Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nature Medicine*, 19(6):747–752.

Roberts, D. C. and Trevan, D. J. (1960). Cell Division in an Ascites Tumour in vitro. *British Journal of Cancer*, 14(4):716–723.

Robles-Espinoza, C. D., Harland, M., Ramsay, A. J., Aoude, L. G., Quesada, V., Ding, Z., Pooley, K. A., Pritchard, A. L., Tiffen, J. C., Petljak, M., Palmer, J. M., Symmons, J., Johansson, P., Stark, M. S., Gartside, M. G., Snowden, H., Montgomery, G. W., Martin, N. G., Liu, J. Z., Choi, J., Makowski, M., Brown, K. M., Dunning, A. M., Keane, T. M., López-Otín, C., Gruis, N. A., Hayward, N. K., Bishop, D. T., Newton-Bishop, J. A., and Adams, D. J. (2014). POT1 loss-of-function variants predispose to familial melanoma. *Nature Genetics*, 46(5):478.

Rosenthal, R., Cadieux, E. L., Salgado, R., Bakir, M. A., Moore, D. A., Hiley, C. T., Lund, T., Tanić, M., Reading, J. L., Joshi, K., Henry, J. Y., Ghorani, E., Wilson, G. A., Birkbak, N. J., Jamal-Hanjani, M., Veeriah, S., Szallasi, Z., Loi, S., Hellmann, M. D., Feber, A., Chain, B., Herrero, J., Quezada, S. A., Demeulemeester, J., Loo, P. V., Beck, S., McGranahan, N., and Swanton, C. (2019). Neoantigen-directed immune escape in lung cancer evolution. *Nature*, 567(7749):479–485.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2014). PyClone: Statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398.

Rowley, J. D. (1973). A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature*, 243(5405):290.

Russo, M., Siravegna, G., Blaszkowsky, L. S., Corti, G., Crisafulli, G., Ahronian, L. G., Mussolin, B., Kwak, E. L., Buscarino, M., Lazzari, L., Valtorta, E., Truini, M., Jessop, N. A., Robinson, H. E., Hong, T. S., Mino-Kenudson, M., Nicolantonio, F. D., Thabet, A., Sartore-Bianchi, A., Siena, S., Iafrate, A. J., Bardelli, A., and Corcoran, R. B. (2016). Tumor Heterogeneity and Lesion-Specific Response to Targeted Therapy in Colorectal Cancer. *Cancer Discovery*, 6(2):147–153.

Sabarinathan, R., Pich, O., Martincorena, I., Rubio-Perez, C., Juul, M., Wala, J., Schumacher, S., Shapira, O., Sidiropoulos, N., Waszak, S., Tamborero, D., Mularoni, L., Rheinbay, E., Hornshoj, H., Deu-Pons, J., Muinos, F., Bertl, J., Guo, Q., Weischenfeldt, J., Korbel,

J. O., Getz, G., Campbell, P. J., Pedersen, J. S., Beroukhim, R., Perez-Gonzalez, A., Lopez-Bigas, N., PCAWG Drivers and Functional Interpretation Group, and ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Net (2017). The whole-genome panorama of cancer drivers. *bioRxiv*, page 190330.

Sandberg, A. A. (1966). The Chromosomes and Causation of Human Cancer and Leukemia. *Cancer Research*, 26(9 Part 1):2064–2081.

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817.

Sawyers, C. (2004). Targeted cancer therapy. *Nature*, 432(7015):294–297.

Schaaf, C. P., Wiszniewska, J., and Beaudet, A. L. (2011). Copy Number and SNP Arrays in Clinical Diagnostics. *Annual Review of Genomics and Human Genetics*, 12(1):25–51.

Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., Feller, S. M., Grocock, R., Henderson, S., Khrebtukova, I., Kingsbury, Z., Luo, S., McBride, D., Murray, L., Menju, T., Timbs, A., Ross, M., Taylor, J., and Bentley, D. (2012). Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, 120(20):4191–4196.

Schumacher, T. N. and Schreiber, R. D. (2015). Neoantigens in cancer immunotherapy. *Science*, 348(6230):69–74.

Seizinger, B. R., Rouleau, G. A., Ozelius, L. J., Lane, A. H., Farmer, G. E., Lamiell, J. M., Haines, J., Yuen, J. W. M., Collins, D., Majoor-Krakauer, D., Bonner, T., Mathew, C., Rubenstein, A., Halperin, J., McConkie-Rosell, A., Green, J. S., Trofatter, J. A., Ponder, B. A., Eierman, L., Bowmer, M. I., Schimke, R., Oostra, B., Aronin, N., Smith, D. I., Drabkin, H., Waziri, M. H., Hobbs, W. J., Martuza, R. L., Conneally, P. M., Hsia, Y. E., and Gusella, J. F. (1988). Von Hippel-Lindau disease maps to the region of chromosome 3 associated with renal cell carcinoma. *Nature*, 332(6161):268.

Sengupta, S., Wang, J., Lee, J., Müller, P., Gulukota, K., Banerjee, A., and Ji, Y. (2015). Bayclone: Bayesian nonparametric inference of tumor subclones using NGS data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 467–478.

Sequist, L. V., Waltman, B. A., Dias-Santagata, D., Digumarthy, S., Turke, A. B., Fidias, P., Bergethon, K., Shaw, A. T., Gettinger, S., Cosper, A. K., Akhavanfard, S., Heist, R. S., Temel, J., Christensen, J. G., Wain, J. C., Lynch, T. J., Vernovsky, K., Mark, E. J., Lanuti, M., Iafrate, A. J., Mino-Kenudson, M., and Engelman, J. A. (2011). Genotypic and Histological Evolution of Lung Cancers Acquiring Resistance to EGFR Inhibitors. *Science Translational Medicine*, 3(75):75ra26–75ra26.

Servick, K. (2014). Breast Cancer: A World of Differences | Science. http://science.sciencemag.org/content/343/6178/1452.

Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4(2):639–650.

Shain, A. H., Joseph, N. M., Yu, R., Benhamida, J., Liu, S., Prow, T., Ruben, B., North, J., Pincus, L., Yeh, I., Judson, R., and Bastian, B. C. (2018). Genomic and Transcriptomic Analysis Reveals Incremental Disruption of Key Signaling Pathways during Melanoma Evolution. *Cancer Cell*, 34(1):45–55.e4.

Sharma, P. and Allison, J. P. (2015). Immune Checkpoint Targeting in Cancer Therapy: Toward Combination Strategies with Curative Potential. *Cell*, 161(2):205–214.

Sharma, P., Hu-Lieskovan, S., Wargo, J. A., and Ribas, A. (2017). Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. *Cell*, 168(4):707–723.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)*, 309(5741):1728–1732.

Sottoriva, A., Curtis, C., Shibata, D., Kang, H., Zhao, J., Siegmund, K., Salomon, M. P., Press, M. F., Marjoram, P., Graham, T. A., and Ma, Z. (2015). A Big Bang model of human colorectal tumor growth. *Nature Genetics*, 47(3):209.

Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., Hunter, C., Bignell, G., Teague, J., Smith, R., Stevens, C., O'Meara, S., Parker, A., Tarpey, P., Avis, T., Barthorpe, A., Brackenbury, L., Buck, G., Butler, A., Clements, J., Cole, J., Dicks, E., Edwards, K., Forbes, S., Gorton, M., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jones, D., Kosmidou, V., Laman, R., Lugg, R., Menzies, A., Perry, J., Petty, R., Raine, K., Shepherd, R., Small, A., Solomon, H., Stephens, Y., Tofts, C., Varian, J., Webb, A., West, S., Widaa, S., Yates, A., Brasseur, F., Cooper, C. S., Flanagan, A. M., Green, A., Knowles, M., Leung, S. Y., Looijenga, L. H. J., Malkowicz, B., Pierotti, M. A., Teh, B., Yuen, S. T., Nicholson, A. G., Lakhani, S., Easton, D. F., Weber, B. L., Stratton, M. R., Futreal, P. A., and Wooster, R. (2005). A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nature Genetics*, 37(6):590.

Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G. R., Yates, L. R., Papaemmanuil, E., Beare, D., Butler, A., Cheverton, A., Gamble, J., Hinton, J., Jia, M., Jayakumar, A., Jones, D., Latimer, C., Lau, K. W., McLaren, S., McBride, D. J., Menzies, A., Mudie, L., Raine, K., Rad, R., Chapman, M. S., Teague, J., Easton, D., Langerød, A., Consortium (OSBREAC), T. O. B. C., Karesen, R., Schlichting, E., Naume, B., Sauer, T., Ottestad, L., Lee, M. T. M., Shen, C.-Y., Tee, B. T. K., Huimin, B. W., Broeks, A., Vargas, A. C., Turashvili, G., Martens, J., Fatima, A., Miron, P., Chin, S.-F., Thomas, G., Boyault, S., Mariani, O., Lakhani, S. R., van de Vijver, M., Veer, L. v. t., Foekens, J., Desmedt, C., Sotiriou, C., Tutt, A., Caldas, C., Reis-Filho, J. S., Aparicio, S. A. J. R., Salomon, A. V., Børresen-Dale, A.-L., Richardson, A. L., Campbell, P. J., Futreal, P. A., and Stratton, M. R. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400.

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239):719–724.

Sun, R., Hu, Z., Sottoriva, A., Graham, T. A., Harpak, A., Ma, Z., Fischer, J. M., Shibata, D., and Curtis, C. (2017). Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nature Genetics*, 49(7):1015–1024.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. *Cell*, 156(6):1324–1335.

Szerlip, N. J., Pedraza, A., Chakravarty, D., Azim, M., McGuire, J., Fang, Y., Ozawa, T., Holland, E. C., Huse, J. T., Jhanwar, S., Leversha, M. A., Mikkelsen, T., and Brennan, C. W. (2012). Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response. *Proceedings of the National Academy of Sciences*, 109(8):3041–3046.

Tamborero, D., Rubio-Perez, C., Deu-Pons, J., Schroeder, M. P., Vivancos, A., Rovira, A., Tusquets, I., Albanell, J., Rodon, J., Tabernero, J., de Torres, C., Dienstmann, R., Gonzalez-Perez, A., and Lopez-Bigas, N. (2018). Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, 10(1):25.

Tarabichi, M., Martincorena, I., Gerstung, M., Markowetz, F., Spellman, P. T., Morris, Q. D., Lingjaerde, O. C., Wedge, D. C., Loo, P. V., and -PCAWG Evolution and Heterogeneity Working Group (2017). Neutral tumor evolution? *bioRxiv*, page 158006.

TCGA Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061.

Templeton, A. J., Knox, J. J., Lin, X., Simantov, R., Xie, W., Lawrence, N., Broom, R., Fay, A. P., Rini, B., Donskov, F., Bjarnason, G. A., Smoragiewicz, M., Kollmannsberger, C., Kanesvaran, R., Alimohamed, N., Hermanns, T., Wells, J. C., Amir, E., Choueiri, T. K., and Heng, D. Y. C. (2016). Change in Neutrophil-to-lymphocyte Ratio in Response to Targeted Therapy for Metastatic Renal Cell Carcinoma as a Prognosticator and Biomarker of Efficacy. *European Urology*, 70(2):358–364.

Tory, K., Brauch, H., Linehan, M., Barba, D., Oldfield, E., Katz, M. F., Seizinger, B., Nakamura, Y., White, R., Marshall, F. F., Lerman, M. I., and Zbar, B. (1989). Specific Genetic Change in Tumors Associated With von Hippel-Lindau Disease. *JNCI: Journal of the National Cancer Institute*, 81(14):1097–1101.

Tsujimoto, Y., Finger, L. R., Yunis, J., Nowell, P. C., and Croce, C. M. (1984). Cloning of the chromosome breakpoint of neoplastic B cells with the t(14;18) chromosome translocation. *Science (New York, N.Y.)*, 226(4678):1097–1099.

Tubio, J. M. C., Li, Y., Ju, Y. S., Martincorena, I., Cooke, S. L., Tojo, M., Gundem, G., Pipinikas, C. P., Zamora, J., Raine, K., Menzies, A., Roman-Garcia, P., Fullam, A., Gerstung, M., Shlien, A., Tarpey, P. S., Papaemmanuil, E., Knappskog, S., Loo, P. V., Ramakrishna, M., Davies, H. R., Marshall, J., Wedge, D. C., Teague, J. W., Butler, A. P., Nik-Zainal, S., Alexandrov, L., Behjati, S., Yates, L. R., Bolli, N., Mudie, L., Hardy, C., Martin, S., McLaren, S., O'Meara, S., Anderson, E., Maddison, M., Gamble, S., Group, I. B. C., Group, I. B. C., Group, I. P. C., Foster, C., Warren, A. Y., Whitaker, H., Brewer, D., Eeles, R., Cooper, C., Neal, D., Lynch, A. G., Visakorpi, T., Isaacs, W. B., van't Veer, L., Caldas, C., Desmedt, C., Sotiriou, C., Aparicio, S., Foekens, J. A., Eyfjörd, J. E., Lakhani, S. R., Thomas, G., Myklebost, O., Span, P. N., Børresen-Dale, A.-L., Richardson, A. L., de Vijver, M. V., Vincent-Salomon, A., den Eynden, G. G. V., Flanagan, A. M., Futreal, P. A., Janes, S. M., Bova, G. S., Stratton, M. R., McDermott, U., and Campbell, P. J. (2014). Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science*, 345(6196):1251343.

Turajlic, S., Xu, H., Litchfield, K., Rowan, A., Chambers, T., Lopez, J. I., Nicol, D., O'Brien, T., Larkin, J., Horswell, S., Stares, M., Au, L., Jamal-Hanjani, M., Challacombe, B., Chandra, A., Hazell, S., Eichler-Jonsson, C., Soultati, A., Chowdhury, S., Rudman, S., Lynch, J., Fernando, A., Stamp, G., Nye, E., Jabbar, F., Spain, L., Lall, S., Guarch, R., Falzon, M., Proctor, I., Pickering, L., Gore, M., Watkins, T. B. K., Ward, S., Stewart, A., DiNatale, R., Becerra, M. F., Reznik, E., Hsieh, J. J., Richmond, T. A., Mayhew, G. F., Hill, S. M., McNally, C. D., Jones, C., Rosenbaum, H., Stanislaw, S., Burgess, D. L., Alexander, N. R., and Swanton, C. (2018). Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. *Cell*, 173(3):581–594.e12.

Turner, N. C. and Reis-Filho, J. S. (2012). Genetic heterogeneity and cancer drug resistance. *The Lancet Oncology*, 13(4):e178–e185.

Van Loo, P., Nordgard, S. H., Lingjærde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Børresen-Dale, A.-L., and Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915.

Van Loo, P. and Voet, T. (2014). Single cell analysis of cancer genomes. *Current Opinion in Genetics & Development*, 24:82–91.

Vinagre, J., Almeida, A., Pópulo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L., Melo, M., da Rocha, A. G., Preto, A., Castro, P., Castro, L., Pardal, F., Lopes, J. M., Santos, L. L., Reis, R. M., Cameselle-Teijeiro, J., Sobrinho-Simões, M., Lima, J., Máximo, V., and Soares, P. (2013). Frequency of TERT promoter mutations in human cancers. *Nature Communications*, 4:2185.

von Hansemann, D. (1890). Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung. *Virchow's Arch Path Anat*, (119).

Waddell, N., Pajic, M., Patch, A.-M., Chang, D. K., Kassahn, K. S., Bailey, P., Johns, A. L., Miller, D., Nones, K., Quek, K., Quinn, M. C. J., Robertson, A. J., Fadlullah, M. Z. H., Bruxner, T. J. C., Christ, A. N., Harliwong, I., Idrisoglu, S., Manning, S., Nourse, C., Nourbakhsh, E., Wani, S., Wilson, P. J., Markham, E., Cloonan, N., Anderson, M. J., Fink, J. L., Holmes, O., Kazakoff, S. H., Leonard, C., Newell, F., Poudel, B., Song, S., Taylor, D., Waddell, N., Wood, S., Xu, Q., Wu, J., Pinese, M., Cowley, M. J., Lee, H. C., Jones, M. D., Nagrial, A. M., Humphris, J., Chantrill, L. A., Chin, V., Steinmann, A. M., Mawson, A., Humphrey, E. S., Colvin, E. K., Chou, A., Scarlett, C. J., Pinho, A. V., Giry-Laterriere, M., Rooman, I., Samra, J. S., Kench, J. G., Pettitt, J. A., Merrett, N. D., Toon, C., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Jamieson, N. B., Graham, J. S., Niclou, S. P., Bjerkvig, R., Grützmann, R., Aust, D., Hruban, R. H., Maitra, A., Iacobuzio-Donahue, C. A., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Corbo, V., Bassi, C., Falconi, M., Zamboni, G., Tortora, G., Tempero, M. A., Gill, A. J., Eshleman, J. R., Pilarsky, C., Scarpa, A., Musgrove, E. A., Pearson, J. V., Biankin, A. V., and Grimmond, S. M. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, 518(7540):495–501.

Wagle, N., Emery, C., Berger, M. F., Davis, M. J., Sawyer, A., Pochanard, P., Kehoe, S. M., Johannessen, C. M., MacConaill, L. E., Hahn, W. C., Meyerson, M., and Garraway, L. A. (2011). Dissecting Therapeutic Resistance to RAF Inhibition in Melanoma by Tumor Genomic Profiling. *Journal of Clinical Oncology*, 29(22):3085–3096.

Wang, K., Yuen, S. T., Xu, J., Lee, S. P., Yan, H. H. N., Shi, S. T., Siu, H. C., Deng, S., Chu, K. M., Law, S., Chan, K. H., Chan, A. S. Y., Tsui, W. Y., Ho, S. L., Chan, A. K. W., Man, J. L. K., Foglizzo, V., Ng, M. K., Chan, A. S., Ching, Y. P., Cheng, G. H. W., Xie, T., Fernandez, J., Li, V. S. W., Clevers, H., Rejto, P. A., Mao, M., and Leung, S. Y. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature Genetics*, 46(6):573.

Wang, M., Zhao, J., Zhang, L., Wei, F., Lian, Y., Wu, Y., Gong, Z., Zhang, S., Zhou, J., Cao, K., Li, X., Xiong, W., Li, G., Zeng, Z., and Guo, C. (2017). Role of tumor microenvironment in tumorigenesis. *Journal of Cancer*, 8(5):761–773.

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., and Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science (New York, N.Y.)*, 350(6264):1096–1101.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Wculek, S. K. and Malanchi, I. (2015). Neutrophils support lung colonization of metastasis-initiating breast cancer cells. *Nature*, 528(7582):413–417.

Weber, G. L., Parat, M.-O., Binder, Z. A., Gallia, G. L., and Riggins, G. J. (2011). Abrogation of PIK3CA or PIK3R1 reduces proliferation, migration, and invasion in glioblastoma multiforme cells. *Oncotarget*, 2(11):833–849.

Weinberg, R. A. (1985). The action of oncogenes in the cytoplasm and nucleus. *Science (New York, N.Y.)*, 230(4727):770–776.

Whalley, J. P., Buchhalter, I., Rheinbay, E., Raine, K. M., Kleinheinz, K., Stobbe, M. D., Werner, J., Beltran, S., Gut, M., Huebschmann, D., Hutter, B., Livitz, D., Perry, M., Rosenberg, M., Saksena, G., Trotta, J.-R., Eils, R., Korbel, J., Gerhard, D. S., Campbell, P., Getz, G., Schlesner, M., Gut, I. G., PCAWG-Tech, Pcawg-Qc, and Pcawg Network (2017). Framework For Quality Assessment Of Whole Genome, Cancer Sequences. *bioRxiv*, page 140921.

Wilkins, M. H. F., Stokes, A. R., and Wilson, H. R. (1953). Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature*, 171(4356):738.

Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A., and Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48(3):238.

Winge, O. (1930). Zytologische Untersuchungen uber die Natur maligner tumoren. II. Tecrkarzinome bei Mausen. *Z. Zellforsch. Mikrosk. Anat.*, 10:683–735.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K. V., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V. K., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E., and Vogelstein, B. (2007). The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science*, 318(5853):1108–1113.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., Micklem, G., Barfoot, R., Hamoudi, R., Patel, S., Rices, C., Biggs, P., Hashim, Y., Smith, A., Connor, F., Arason, A., Gudmundsson, J., Ficenec, D., Kelsell, D., Ford, D., Tonin, P., Bishop, D. T., Spurr, N. K., Ponder, B. A. J., Eeles, R., Peto, J., Devilee, P., Cornelisse, C., Lynch, H., Narod, S., Lenoir, G., Egilsson, V., Barkadottir, R. B., Easton, D. F., Bentley, D. R., Futreal, P. A., Ashworth, A., and Stratton, M. R. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature*, 378(6559):789.

Wooster, R., Neuhausen, S. L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., and Averill, D. (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science (New York, N.Y.)*, 265(5181):2088–2090.

Wouters, M. C. A. and Nelson, B. H. (2018). Prognostic Significance of Tumor-Infiltrating B Cells and Plasma Cells in Human Cancer. *Clinical Cancer Research*, 24(24):6125–6135.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., and Iacobuzio-Donahue, C. A. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, 467(7319):1114.

Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L. B., Larsimont, D., Davies, H., Li, Y., Ju, Y. S., Ramakrishna, M., Haugland, H. K., Lilleng, P. K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D., Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, L. J., Jiang, B., Vincent, D., Greene-Colozzi, A., Adnet, P.-Y., Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M. R., Sotiriou, C., Richardson, A. L., Lønning, P. E., Wedge, D. C., and Campbell, P. J. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature Medicine*, 21(7):751–759.

Yates, L. R., Knappskog, S., Wedge, D., Farmery, J. H. R., Gonzalez, S., Martincorena, I., Alexandrov, L. B., Loo, P. V., Haugland, H. K., Lilleng, P. K., Gundem, G., Gerstung, M., Pappaemmanuil, E., Gazinska, P., Bhosle, S. G., Jones, D., Raine, K., Mudie, L., Latimer, C., Sawyer, E., Desmedt, C., Sotiriou, C., Stratton, M. R., Sieuwerts, A. M., Lynch, A. G., Martens, J. W., Richardson, A. L., Tutt, A., Lønning, P. E., and Campbell, P. J. (2017). Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell*, 32(2):169–184.e7.

Yuan, K., Sakoparnig, T., Markowetz, F., and Beerenwinkel, N. (2015). BitPhylogeny: A probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biology*, 16:36.

Yung, C. K., O’Connor, B. D., Yakneen, S., Zhang, J., Ellrott, K., Kleinheinz, K., Miyoshi, N., Raine, K. M., Royo, R., Saksena, G. B., Schlesner, M., Shorser, S. I., Vazquez, M., Weischenfeldt, J., Yuen, D., Butler, A. P., Davis-Dusenbery, B. N., Eils, R., Ferretti, V., Grossman, R. L., Harismendy, O., Kim, Y., Nakagawa, H., Newhouse, S. J., Torrents, D., Stein, L. D., and PCAWG Technical Working Group (2017). Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments. *bioRxiv*, page 161638.

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhang, C.-Z., Wala, J., Mermel, C. H., Sougnez, C., Gabriel, S. B., Hernandez, B., Shen, H., Laird, P. W., Getz, G., Meyerson, M., and Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45(10):1134–1140.

Zaretsky, J. M., Garcia-Diaz, A., Shin, D. S., Escuin-Ordinas, H., Hugo, W., Hu-Lieskovan, S., Torrejon, D. Y., Abril-Rodriguez, G., Sandoval, S., Barthly, L., Saco, J., Homet Moreno, B., Mezzadra, R., Chmielowski, B., Ruchalski, K., Shintaku, I. P., Sanchez, P. J., Puig-Saus, C., Cherry, G., Seja, E., Kong, X., Pang, J., Berent-Maoz, B., Comin-Anduix, B., Graeber, T. G., Tumeh, P. C., Schumacher, T. N., Lo, R. S., and Ribas, A. (2016). Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *New England Journal of Medicine*, 375(9):819–829.

Zhang, Z., Lee, J. C., Lin, L., Olivas, V., Au, V., LaFramboise, T., Abdel-Rahman, M., Wang, X., Levine, A. D., Rho, J. K., Choi, Y. J., Choi, C.-M., Kim, S.-W., Jang, S. J., Park, Y. S., Kim, W. S., Lee, D. H., Lee, J.-S., Miller, V. A., Arcila, M., Ladanyi, M., Moonsamy, P., Sawyers, C., Boggon, T. J., Ma, P. C., Costa, C., Taron, M., Rosell, R., Halmos, B., and Bivona, T. G. (2012). Activation of the AXL kinase causes resistance to EGFR-targeted therapy in lung cancer. *Nature Genetics*, 44(8):852–860.

Zhao, X., Li, C., Paez, J. G., Chin, K., Jänne, P. A., Chen, T.-H., Girard, L., Minna, J., Christiani, D., Leo, C., Gray, J. W., Sellers, W. R., and Meyerson, M. (2004). An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. *Cancer Research*, 64(9):3060–3071.

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., Henaff, E., McIntyre, A. B. R., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., Saghbini, M., Dzakula, Z., Hastie, A., Cao, H., Deikus, G., Schadt, E., Sebra, R., Bashir, A., Truty, R. M., Chang, C. C., Gulbahce, N., Zhao, K., Ghosh, S., Hyland, F., Fu, Y., Chaisson, M., Xiao, C., Trow, J., Sherry, S. T., Zaranek, A. W., Ball, M., Bobe, J., Estep, P., Church, G. M., Marks, P., Kyriazopoulou-Panagiotopoulou, S., Zheng, G. X. Y., Schnall-Levin, M., Ordonez, H. S., Mudivarti, P. A., Giorda, K., Sheng, Y., Rypdal, K. B., and Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, 3.

List of figures

List of tables

Glossary

APOBEC apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like

ASCAT allele-specific copy number analysis of tumors

BAF B-allele frequency

CAF Carcinoma-associated fibroblasts

CCF Cancer cell fraction - The fraction of tumour cells that carry a particular mutation, or the fraction of tumour cells in the sequencing sample that is represented by a mutation cluster

CGH Comparative genomic hybridization

Clonal Is used to refer to mutations carried by all tumour cells. A cluster of these mutations is referred to as clone

CNA Copy number alteration - Somatic copy number change

CNLOH Copy neutral loss of heterozygosity

COSMIC Catalogue of somatic mutations in cancer

CP Cellular proportion/prevalence - The proportion of cells in the sequencing sample that a mutation is carried by or is represented by a mutation cluster. CP is CCF multiplied by the tumour purity: CP=CCF*ρ

ctDNA Cell free tumour DNA

dbSNP Single Nucleotide Polymorphism Database

DNA Deoxyribonucleic acid

Driver mutation A somatic mutation that is thought to convey any selective advantage

ECM Extra cellular matrix

EMD Earth movers distance

EMT Epithelial-mesenchymal transition

FDR False discovery rate

GC content The proportion of bases within a range of the types C or G

Haplotype phasing The process of estimating which SNP alleles appear on the same chromosome within an individual organism

Haplotype A set of SNPs on the same chromosome

hg19 Human reference genome build 19

HMM Hidden Markov model

ICGC International Cancer Genome Consortium

IGH locus Immunoglobulin heavy locus

Indel Somatic short insertion or deletion

Infinite sites assumption The assumption that mutations occur only once during the life time of the tumour

ITH Intra-tumour heterogeneity

K-S Kolmogorov-Smirnov test

Kataegis Localised somatic hypermutation

logR Quantification of the amount of DNA available for a certain locus. In sequencing data, logR represents the ratio of the coverage of the tumour over that in the matched normal

LOH Loss of heterozygosity

MCMC Markov chain Monte Carlo

MPEAR Maximal posterior expected Rand index

Multiplicity The number of chromosome copies that carry a somatic mutation

Passenger mutation A somatic mutation that is thought to not convey any selective advantage

PCAWG Pan-Cancer Analysis of Whole Genomes

PCF Piecewise constant fitting

Ploidy The average number of chromosome copies of a cell

Purity The proportion of tumour cells available in the data

QC Quality control

RMSE Root mean squared error

RNA Ribonucleic acid

SMC-het Somatic Mutation Calling heterogeneity - Challenge to benchmark subclonal reconstruction methods

SNP Single nucleotide polymorphism - A germline single base difference from the reference genome

SNV Single nucleotide variant - A somatic single base substitution

Subclonal reconstruction Estimation of the number of subclonal cell populations within a tumour sequencing sample, the number of mutations in each population and the size of each population (fraction of tumour cells)

Subclonal Is used to refer to mutations carried by a subset of tumour cells. A cluster of these mutations is referred to as a subclone

SV Somatic structural variant

TAM Tumour-associated macrophages

TCGA The Cancer Genome Project

TIL Tumour-infiltrating lymphocytes

TME Tumour micro-environment

TSG Tumour suppressor gene

UV Ultraviolet

VAF Variant allele frequency

WABCS West African Breast Cancer Study

WGS Whole genome sequencing

WXS Whole exome sequencing

Appendix A

The evolutionary history of 2,658 cancers

This thesis describes my Ph.D. work that was undertaken for the ICGC PCAWG project. The working group that I am part of has produced two papers, at the point of writing, on both of which I am a shared first author. The bulk of my work however has focussed on the pan-cancer description of intra-tumour heterogeneity. I have participated in the evolutionary history of 2,658 cancers story to a lesser extend, where my role was to deliver the right input data required for the evolutionary history analysis. I have therefore opted to attach the manuscript of the evolutionary history paper in this appendix and include a brief overview of the results in Chapter 7.

The evolutionary history of 2,658 cancers

Moritz Gerstung[1,2,#,*], Clemency Jolly[3,#], Ignaty Leshchiner[4,#], Stefan C. Dentro[2,3,5,#], Santiago Gonzalez[1], Thomas J. Mitchell[2,6], Yulia Rubanova[7], Pavana Anur[8], Daniel Rosebrock[4], Kaixian Yu[9], Maxime Tarabichi[3], Amit Deshwar[7], Jeff Wintersinger[7], Kortine Kleinheinz[10,11], Ignacio Vázquez-García[2,6], Kerstin Haase[3], Subhajit Sengupta[12], Geoff Macintyre[13], Salem Malikic[14], Nilgun Donmez[14], Dimitri G. Livitz[4], Marek Cmero[15], Jonas Demeulemeester[3,16], Steven Schumacher[4], Yu Fan[9], Xiaotong Yao[17,18], Juhee Lee[19], Matthias Schlesner[10], Paul C. Boutros[7,20], David D. Bowtell[21,22], Hongtu Zhu[9], Gad Getz[4], Marcin Imielinski[17,18], Rameen Beroukhim[4], S. Cenk Sahinalp[23], Yuan Ji[12,24], Martin Peifer[25], Florian Markowetz[13], Ville Mustonen[26], Ke Yuan[13,27], Wenyi Wang[9], Quaid D. Morris[7], Paul T. Spellman[8,#], David C. Wedge[5,#], Peter Van Loo[3,16,#,*], on behalf of the PCAWG Evolution and Heterogeneity Working Group[28] and the PCAWG network.

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom; [2]Wellcome Trust Sanger Institute, Cambridge, United Kingdom; [3]The Francis Crick Institute, London, United Kingdom; [4]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [5]Big Data Institute, University of Oxford, Oxford, United Kingdom; [6]University of Cambridge, Cambridge, United Kingdom; [7]University of Toronto, Toronto, Canada; [8]Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA; [9]The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [10]German Cancer Research Center (DKFZ), Heidelberg, Germany; [11]Heidelberg University, Heidelberg, Germany; [12]NorthShore University HealthSystem, Evanston, IL, USA; [13]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom; [14]Simon Fraser University, Vancouver, Canada; [15]University of Melbourne, Melbourne, Australia; [16]Department of Human Genetics, University of Leuven, Leuven, Belgium; [17]Weill Cornell Medicine, New York, NY, USA; [18]New York Genome Center, New York, NY, USA; [19]University of California Santa Cruz, Santa Cruz, CA, USA; [20]Ontario Institute for Cancer Research, Toronto, Canada; [21]Peter MacCallum Cancer Centre,

Melbourne, VIC 3052, Australia; [22]Garvan Institute of Medical Research, Sydney, NSW 2010, Australia; [23]Indiana University, Bloomington, IN, USA; [24]The University of Chicago, Chicago, IL, USA; [25]University of Cologne, Cologne, Germany; [26]University of Helsinki, Helsinki, Finland; [27]University of Glasgow, Glasgow G12 8RZ, United Kingdom.

[#]These authors contributed equally.

[*]To whom correspondence may be addressed:

Moritz Gerstung, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, CB10 1SD, United Kingdom. Tel: +44 (0) 1223 49 4636, email: Moritz.Gerstung@ebi.ac.uk.

Peter Van Loo, The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, United Kingdom. Tel: +44 (0) 20 3796 1719, e-mail: Peter.VanLoo@crick.ac.uk.

[28]A list of members of the PCAWG Evolution and Heterogeneity Working Group can be found at the end of the manuscript.

Summary

Cancer develops through a process of somatic evolution. Here, we reconstruct the evolutionary history of 2,778 tumour samples from 2,658 donors spanning 39 cancer types. Characteristic copy number gains, such as trisomy 7 in glioblastoma or isochromosome 17q in medulloblastoma, are found amongst the earliest events in tumour evolution. The early phases of oncogenesis are driven by point mutations in a restricted set of cancer genes, often including biallelic inactivation of tumour suppressors. By contrast, increased genomic instability, a more than three-fold diversification of driver genes, and an acceleration of mutational processes are features of later stages. Clock-like mutations yield estimates for whole genome duplications and subclonal diversification in chronological time. Our results suggest that driver mutations often precede diagnosis by many years, and in some cases decades. Taken together, these data reveal common and divergent trajectories of cancer evolution, pivotal for understanding tumour biology and guiding early cancer detection.

Introduction

Cancer arises through natural selection: initiated by mutations in a single cell, the accumulation of subsequent aberrations and the effects of selection over time result in the clonal expansions of cells, ultimately leading to the formation of a genomically aberrant tumour[1]. This model has been underpinned by genetic studies, starting with classical work on retinoblastoma[2] and the sequence of *APC*, *KRAS* and *TP53* mutations during colorectal adenoma to adenocarcinoma progression[3]. Establishing a particular order of mutations during the somatic evolution of cancers systematically across cancer types, however, has proven to be complicated due to small sample sizes and the stochastic nature of evolution between individuals.

Deep sequencing of bulk tumour samples makes it possible to examine the evolutionary history of individual tumours, based on the catalogue of somatic mutations they have accumulated[4]. Many studies have reconstructed the phylogenetic relationships between tumour samples and metastases from individual patients[5-8], corroborating the clonal evolution model. From single samples, the timing of chromosomal gains can be estimated using point mutations within duplicated regions[9,10]. In addition, the relative ordering of events within a tumour type can be determined by aggregating pairwise timing estimates of genomic changes (for example clonal *vs*. subclonal) across many samples using preference models[11,12]. While these approaches provide insights into tumour development, they have only been applied to a limited number of cancers.

Here, we use the Pan-Cancer Analysis of Whole Genomes (PCAWG)[13] dataset, as part of the International Cancer Genome Consortium (ICGC)[14] and The Cancer Genome Atlas (TCGA)[15] to characterise the evolutionary history of 2,778 cancers from 2,658 unique donors across 39 cancer types. We determine the order and timing of mutations in cancer development to delineate the patterns of chromosomal evolution across and within different cancer types. We then define broad periods of tumour evolution and examine how drivers and mutational signatures vary between these stages. Finally, using CpG>TpG mutations, we convert timing estimates into approximate real time, and create typical timelines of tumour evolution.

Results

Reconstructing the life history of single tumours

A cancer cell's genome is the cumulative result of the somatic aberrations that have arisen during its evolutionary past, and part of this history can be reconstructed from deep whole genome sequencing data (**Fig. 1a**)[4]. Initially, each point mutation occurs on a single chromosome in a single cell. If that chromosomal locus is subsequently duplicated, the point mutation will be co-amplified with the gained allele, which can be detected in deep sequencing data. Likewise, mutations found in a subset of tumour cells have not swept through the population, and must have occurred after most recent common ancestor (MRCA) of the tumour cells in the sequenced sample.

Mapping point mutations to the proportion of cells and chromosomes enables us to define three categories, which we term *early clonal*, *late clonal* and *subclonal,* each associated with broad epochs of tumour evolution (**Fig. 1a**). *Clonal* mutations have occurred before the occurrence of the MRCA and are common to all cancer cells. These can often be further subdivided as either *early clonal* if they occurred before copy number gains, or *late clonal* otherwise. Additionally, *subclonal* mutations are only observed in a fraction of cancer cells. Importantly, the number of early (and late) clonal mutations provides information about the timing of the underlying copy number segment. For example, there would be few, if any, coamplified early clonal mutations if the gain had occurred right after fertilisation (**Fig. 1a** and **Online Methods**)[9].

These analyses are illustrated in **Fig. 1b**. As expected, the frequency of somatic point mutations cluster tightly around the values imposed by the purity of the sample, local copy number configuration and identified subclones. As the sample pictured has undergone whole genome duplication (WGD), the mutation time estimates of all copy number segments scatter narrowly around a single time-point, independently of the exact copy number state, confirming that WGD is a single catastrophic event.

Timing patterns of copy number gains

To systematically explore the timing of copy number gains pan-cancer, we applied mutational timing analysis to all 2,778 samples from 2,658 distinct donors across the PCAWG dataset (see **Supplementary Methods**). We find that chromosomal gains are typically acquired during the second half of clonal evolution (median value 0.76, IQR = 0.43-0.94), with systematic differences between tumour types (**Fig. 2a, Supplementary Fig. 1**). In glioblastoma, medulloblastoma and pancreatic neuroendocrine cancers, a substantial fraction of gains occurs early in mutational time. Conversely, in squamous cell lung cancers and melanomas, gains arise towards the end of the mutational time scale. Most tumour types, including breast, ovarian and colorectal cancer, show relatively broad periods of chromosomal instability, rather than staggered events throughout clonal evolution.

There are, however, certain tumour types with consistently early gains of specific chromosomal regions. Most pronounced is glioblastoma, where single copy gains of chromosomes 7, 19 and/or 20 are present in 90% of tumours (**Fig. 2a-b**). Strikingly, these gains are consistently timed within the first 10% of clonal mutational time. Similarly, the duplications leading to isochromome 17q in medulloblastoma are timed exceptionally early. Although less pronounced, gains of chromosome 18 in B-cell non-Hodgkin lymphoma, as well as gains of the q arm of chromosome 5 in clear cell renal cell carcinoma, often have a distinctively early timing within the first 50% of mutational time.

We observed that co-occurring gains in the same tumour often appear to occur at a similar time, pointing towards punctuated bursts of copy number gains involving the majority of gained segments (**Fig. 2c**). While this is expected in tumours with WGD (**Fig. 1b**), it may seem surprising to observe synchronous gains (defined as more than 80% of gained segments in a single event) in near-diploid tumours. Still, synchronous gains are frequent, occurring in a striking 58% (469/814) of informative near-diploid tumours, 61% more frequently than expected by chance ($p < 0.01$, permutation test; **Fig. 2d**). These data indicate that tumour evolution is often driven in short bursts involving multiple chromosomes, confirming earlier observations in breast cancer[16].

Timing of mutations in driver genes

As outlined above, point mutations can be qualitatively assigned to different time categories, allowing the timing of driver mutations (**Fig. 1a, 3a**). Using a panel of 453 cancer driver genes[17], we find that the timing distribution of pathogenic mutations in the 50 most common drivers is predominantly clonal, and often early clonal (**Fig. 3a-b**). For example, *TP53* and *KRAS* are 5-9x more likely to be mutated in the early than in the late clonal stage. For *TP53*, this trend is independent of tumour type (**Fig. 3c**). Mutations in *PIK3CA* are 4x more frequently clonal than subclonal, while non-coding changes near the *TERT* gene are 8x more frequently early clonal than expected. In contrast, *SETD2* mutations are frequently subclonal, in agreement with previous reports[5]. Mutations in the non-coding RNA *RMRP* appear to be frequently late and subclonal.

Overall, common driver mutations predominantly occur early during tumour evolution. To understand how the entire landscape of all 453 driver genes changes over time, we calculated how the number of driver mutations relates to the number of driver genes in each of the evolutionary stages. This reveals an increasing diversity of driver genes mutated at later stages of tumour development: 50% of all early clonal driver mutations are found in only 12 different genes, whereas the corresponding proportion of late and subclonal mutations occur in approximately 39 and 36 different genes, respectively, a more than 3-fold increase (**Fig. 3d**). These results are consistent with previous findings in non-small-cell lung cancers[18], and suggests that, across cancer types, the very early carcinogenic events occur in a constrained set of common drivers, while a more diverse array of drivers is involved in late tumour development.

Relative timing of somatic driver events

Next, we sought to better understand the sequence and timing of events during tumour evolution by integrating the timing of driver point mutations and recurrent copy number changes across cancer samples. We calculated an overall probabilistic ranking of lesions, detailing whether each lesion occurs preferentially early or late during tumour evolution, by aggregating order relations between pairs of lesions from individual samples within each cancer type (**Supplementary Methods,** section 3.2, **Supplementary Fig. 2**).

In colorectal adenocarcinoma, for example, we find *APC* mutations to have the highest odds of occurring early, followed by *KRAS,* loss of 17p and *TP53,* and *SMAD4* (**Fig. 3e**). Whole-genome duplications have an intermediate ranking, indicating a variable timing, while many chromosomal gains and losses are typically late. These results are in agreement with the classical progression of *APC-KRAS-TP53* proposed by Vogelstein and Fearon[3], but add considerable detail.

In other cancer types, the sequence of events in cancer progression has not previously been studied in as much detail as colorectal cancer. For example, in pancreatic neuroendocrine cancers, we find that many chromosomal losses, including those of chromosomes 2, 6, 11 and 16, occur early, followed by driver mutations in *MEN1* and *DAXX* (**Fig. 3f**). WGD events occur late, after many of these tumours have reached a pseudo-haploid state due to wide-spread chromosomal losses. In glioblastoma, we find that loss of chromosome 10 and driver mutations in *TP53* and *EGFR* are very early, often preceding early gains of chromosomes 7, 19 and 20 (as described above) (**Fig. 3g**). *TERT* promoter mutations tend to occur at early to intermediate time points, while other driver mutations and copy number changes tend to be later events.

Across cancer types, we typically find *TP53* mutations early, as well as losses of chromosome 17 (**Supplementary Fig. 1**). WGD events usually have an intermediate ranking and the majority of copy number changes occur after WGD. We also find that losses typically precede gains, and consistent with the results above, we find that common drivers typically occur earlier than rare drivers.

Timing of mutational signatures

Mutagenic processes acting on the tumour genome often leave characteristic signatures of their activity[19,20]. In order to quantify how these processes change over time, we estimated the intensity of active signatures within each sample, across the qualitative epochs of tumour evolution (early clonal, late clonal and subclonal). The changes in proportion of mutations associated with a given signature in each of these epochs provide a measure of the dynamics of relative signature activity (**Fig. 4**, **Supplementary Fig. 3**).

Overall, we find that signature activities typically change during clonal evolution by less than 30% (median fold change 0.98, IQR [0.70-1.36]), indicating that mutational processes act at a rather constant rate during tumour progression. This is in contrast with the variation of signatures across patients, which varies 10 to 100-fold. There are, however, particular signatures that show consistent trends over time, both pan-cancer and within certain tumour types (**Fig. 4**). For example, the relative activity of the mutational signature associated with DNA damage caused by tobacco smoking (signature 4) decreases at least 1.2-fold in 70% of cancers where it is active clonally, consistent with previous reports in lung adenocarcinoma[21,22].

Other signatures, including UV light (signature 7) in melanoma (40% of samples with clonally active signature), and signature 12, of unknown aetiology, in liver cancer (83% of samples) show a similar ≥1.2-fold decrease in activity towards the later stages of clonal evolution (**Fig. 4**). We also observe that some signatures increase in late clonal evolution, most notably signatures 2 and 13, which are associated with the activity of APOBEC enzymes and increase by more than 1.2-fold in 58% of samples that have this signature. Similarly, the signature associated with *BRCA* mutations and defective double strand break repair (signature 3) increases in late clonal evolution in 35% of the samples where it is active. Similar trends also hold between clonal and subclonal phases of tumour evolution (**Supplementary Fig. 3**).

Chronological time estimates of whole genome duplications and subclonal diversification

Any changes in the mutation rate of cancers influence timing estimates made from mutational data. Due to increased proliferation and in some cases acquired hypermutation, one would generally expect an increase in the mutation rate (per year) in cancer, yet some mutational processes appear more variable than others.

The above analysis of signature changes revealed that the relative contribution of signature 1 usually decreases as other mutational processes become more active (**Fig. 4**). Mutational signature 1, characterised by CpG>TpG mutations, is a promising candidate for a clock-like process, as it is ubiquitously active in all tissues and has been described as correlating with age in normal tissues[23,24] and multiple tumour types[25]. The latter implies not only that it is fairly constant in a given cell lineage, but also that it varies little across patients. For the purpose of timing mutations in chronological time, only the former property is required, as the age at diagnosis provides a reference by which relative timing estimates are scaled.

The acceleration of overall mutation rate and CpG>TpG rate can be directly estimated from sequencing data of matched primary and relapse samples from the same donor by comparing the rates of mutations that have accumulated between fertilisation and primary diagnosis to those accumulated between diagnosis and relapse. Suitable samples are publicly available for ovarian cancer[26], breast cancer[27] and acute myeloid leukaemia[28]. While for all point mutations, the median acceleration ranges between 3.3 for AML and 11.7 for ovarian cancer, CpG>TpG mutations display lower values and less variability (ranging from 2.8 to 6.7; **Fig. 5a**). To some extent this acceleration may be driven by treatment, but we may use it as a conservative reference for other tumour types.

Accounting for the acceleration above, we inferred the chronological time of whole-genome duplications based on CpG>TpG mutations (**Supplementary Methods**, section 5; **Fig. 5b**). While the typical timing of WGD is about one decade before diagnosis (assuming a 5x CpG>TpG mutation acceleration), we observe substantial variability among samples of a given tumour type, with many cases dating back more than two decades. Ovarian adenocarcinoma shows very early occurrences of WGD with approximately half of the samples having WGD more than two decades before diagnosis (**Fig. 5b**). A similar phenomenon is seen for breast adenocarcinoma. Without any acceleration, the estimated median occurrence of WGD would be 15-25yrs for the majority of cancer types; this value decreases with greater values of CpG>TpG acceleration (**Fig. 5c**).

We used a similar approach to calculate the timing of the emergence of the MRCA, and therefore the onset of subclonal diversification. The typical timing is considerably closer to diagnosis although, interestingly, there are also cases dating back more than ten years before diagnosis (**Fig. 5d**). We note, however, that timing the occurrence of the MRCA is more difficult, as it is not always possible to calculate the phylogenetic relationship between subclones. The MRCA may date back longer if subclones arise sequentially.

While the exact timing of individual samples remains challenging due to low mutation numbers and unknown mutation rates for individual tumours, on average, a picture emerges where across tumour types, the median MRCA ranges between six months and six years before diagnosis, while WGD typically occurs 2-11 years before diagnosis (**Fig. 5e**). These findings dovetail with epidemiological observations: cancer generally arises past the age of 50[29], and the typical latency between carcinogen exposure and cancer detection, most notable in tobacco-associated cancers, is several years to multiple decades[30]. Furthermore the progression of most known precancerous lesions to carcinomas occurs usually over multiple years, if not decades[31-38]. The data presented here corroborate that these time scales hold also in cases without detectable premalignant conditions, raising hopes that these tumours could also be detected in precancerous stages.

Discussion

Taken together, these analyses begin to build an overall picture of tumour development. Across cancer types, early tumour development is characterised by mutations in a handful of canonical driver genes, and biallelic inactivation of tumour suppressor genes, such as *TP53*. Copy number gains during this time are relatively infrequent in many tumour types, but can be distinctive in others. Throughout the later stages of tumour evolution, increased genetic instability, a greater diversity of drivers, and an acceleration of mutational processes shape the final subclonal diversification.

Our combined approaches allow us to draw timelines of tumour development over different cancer types (**Fig. 6**; **Supplementary Fig. 1**). We see that many years before a tumour is diagnosed, endogenous and exogenous mutational processes have resulted in key driver mutations and chromosomal instability. An intriguing finding is that large somatic events, such as WGD, can occur decades before the appearance and diagnosis of a tumour. Thus, the process of tumour development may span an entire lifetime.

Our findings raise the possibility of early detection, if cells carrying early mutations can be detected and distinguished from cells not progressing further. The discovery of distinctive, early mutations in certain tumour types, such as gains of chromosomes 7, losses of chromosome 10, and EGFR mutations in glioblastoma, and isochromosome 17q in medulloblastoma, begin to unveil possible candidate lesions.

Individual tumour types show characteristic sets of evolutionary trajectories, reflecting differences in the underlying biology of tumorigenesis (**Fig. 6**; **Supplementary Fig. 1**). Where applicable, these trajectories agree with previous studies of genomic aberrations acquired at different stages of tumour progression (e.g. in colorectal cancer[3]). Unlike most other cancers, high grade serous ovarian adenocarcinomas typically acquire chromosomal gains within the first half of clonal evolution (**Fig. 6d**). Our findings are consistent with these tumours being the most genomically unstable of all solid cancers[39], and with their high frequency of *TP53* and homologous recombination repair defects[40]. Both across and within cancer types, these typical evolutionary trajectories and their correlations with clinical features may provide an opportunity to develop prognostic markers and more effective therapies.

Our findings provide insight into the process of selection acting on tumours throughout their development. The genetic canalization in early tumour development, and increased diversity of driver mutations later in tumour evolution, is striking. It suggests a strong epistasis of fitness effects constraining evolution initially to a small set of mutational events that are able to initiate neoplastic transformation. Over time, as tumours evolve, the small- and large-scale somatic changes they subsequently accumulate propel them towards increasingly specialised developmental paths driven by individually rare, atypical driver mutations.

In summary, we present the first pan-cancer analysis of the evolutionary history of tumours. The timelines we derive from this analysis show that in a wide range of cancer types, tumour evolution often follows a typical pattern. This can begin decades before diagnosis, thus providing a window for early diagnosis and clinical intervention.

References

1 Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23--28, doi:10.1126/science.959840 (1976).

2 Knudson, A. G. J. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820--823 (1971).

3 Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759--767, doi:10.1016/0092-8674(90)90186-I (1990).

4 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007, doi:10.1016/j.cell.2012.04.023 (2012).

5 Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* **366**, 883-892, doi:10.1056/NEJMoa1113205 (2012).

6 Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353-357, doi:10.1038/nature14347 (2015).

7 Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med* **21**, 751-759, doi:10.1038/nm.3886 (2015).

8 Brastianos, P. K. *et al.* Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. *Cancer Discov* **5**, 1164-1177, doi:10.1158/2159-8290.CD-15-0369 (2015).

9 Durinck, S. *et al.* Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov* **1**, 137-143, doi:10.1158/2159-8290.CD-11-0028 (2011).

10 Purdom, E. *et al.* Methods and challenges in timing chromosomal abnormalities within cancer samples. *Bioinformatics* **29**, 3113-3120, doi:10.1093/bioinformatics/btt546 (2013).

11 Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616-3627, doi:10.1182/blood-2013-08-518886 (2013).

12 Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression

and relapse. *Nature* **526**, 525-530, doi:10.1038/nature15395 (2015).

13 Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *bioRxiv* (2017).

14 Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391-2405, doi:10.1182/blood-2016-03-643544 (2016).

15 McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068, doi:10.1038/nature07385 (2008).

16 Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet* **48**, 1119-1130, doi:10.1038/ng.3641 (2016).

17 PCAWG working group 2-5-9-14 (Analysis of mutations). *Manuscript in preparation* (2017).

18 Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med* **376**, 2109-2121, doi:10.1056/NEJMoa1616288 (2017).

19 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

20 PCAWG working group 7 (Mutation signatures and processes). *Manuscript in preparation* (2017).

21 McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci Transl Med* **7**, 283ra254, doi:10.1126/scitranslmed.aaa1408 (2015).

22 Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* **17**, 31, doi:10.1186/s13059-016-0893-4 (2016).

23 Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260-264, doi:10.1038/nature19768 (2016).

24 Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264-278, doi:10.1016/j.cell.2012.06.023 (2012).

25 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet*, doi:10.1038/ng.3441 (2015).

26 Patch, A.-M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489-494, doi:10.1038/nature14410 (2015).

27 Yates, L. R. & others. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* **in press** (2017).

28 Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, doi:10.1038/nature10738 (2012).

29 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer Statistics, 2017. *CA Cancer J Clin* **67**, 7-30, doi:10.3322/caac.21387 (2017).

30 Thun, M. J., Henley, S. J. & Calle, E. E. Tobacco use and cancer: an epidemiologic perspective for geneticists. *Oncogene* **21**, 7307-7325, doi:10.1038/sj.onc.1205807 (2002).

31 Bostwick, D. G. & Qian, J. High-grade prostatic intraepithelial neoplasia. *Mod Pathol* **17**, 360-379, doi:10.1038/modpathol.3800053 (2004).

32 Brenner, H. *et al.* Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840,149 screening colonoscopies. *Gut* **56**, 1585-1589, doi:10.1136/gut.2007.122739 (2007).

33 Gazdar, A. F. & Brambilla, E. Preneoplasia of lung cancer. *Cancer Biomark* **9**, 385-396, doi:10.3233/CBM-2011-0166 (2010).

34 Sanders, M. E., Schuyler, P. A., Dupont, W. D. & Page, D. L. The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. *Cancer* **103**, 2481-2484, doi:10.1002/cncr.21069 (2005).

35 Schlecht, N. F. *et al.* Human papillomavirus infection and time to progression and regression of cervical intraepithelial neoplasia. *J Natl Cancer Inst* **95**, 1336-1343 (2003).

36 Whitson, M. J. & Falk, G. W. Predictors of Progression to High-Grade Dysplasia or Adenocarcinoma in Barrett's Esophagus. *Gastroenterol Clin*

North Am **44**, 299-315, doi:10.1016/j.gtc.2015.02.005 (2015).

37 Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med* **20**, 1472-1478, doi:10.1038/nm.3733 (2014).

38 Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**, 2477-2487, doi:10.1056/NEJMoa1409405 (2014).

39 Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45**, 1127-1133, doi:10.1038/ng.2762 (2013).

40 Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).

Figure Legends

Figure 1. Principles of timing mutations

(**a**) Principles of timing mutations based on deep whole genome sequencing. According to the clonal evolution model of cancer, tumour cells evolve in multiple selective sweeps. During some of these sweeps, copy number gains are acquired, which can be used for timing analyses (green and purple epochs). Mutations acquired after the last clonal expansion are present in distinct subclonal populations (red epoch). The number of sequencing reads reporting point mutations can be used to discriminate variants as early or late clonal (green/purple) in cases of specific copy number gains, as well as clonal (blue) or subclonal (red) in cases without (right). The distribution of the number of early and late clonal mutations carries information about the timing of the copy number gains with the exact relation depending on the resulting copy number configuration (bottom). (**b**) Example case illustrating the annotation of point mutations based on the variant allele frequency (VAF, top) and copy number configuration (middle), each shown as a function of genomic coordinate (x-axis). The resulting timing estimates for each copy number segment are shown at the bottom, indicating that all segments were gained at a similar time (whole genome duplication).

Figure 2. Pan-cancer timing patterns of arm-level gains

(**a**) Overview of timing arm-level copy number gains across different cancer types. Depicted are the smoothened histograms (y-axes; scale bar 5% recurrence) of the timing estimates of large gains at decile resolution (x-axes), split by tumour type and chromosome on which gains are detected. (**b**) Heatmaps representing timing estimates of gains on different chromosome arms (x-axis) for individual samples (y-axis) for selected tumour types. (**c**) Two near-diploid example cases illustrating synchronous gains with a single peak in amplification activity (top) and asynchronous gains with multiple amplification periods (bottom). (**d**) Distribution of synchronous and asynchronous gain patterns across samples, split by whole genome duplication status (left). Uninformative samples carry too few or too small gains to be timed accurately. Systematic permutation tests reveal a 61% enrichment of synchronous gains in near-diploid samples (right).

Figure 3. Timing of driver mutations and relative ordering of somatic events

(**a-d**) **Timing of driver point mutations.** (**a**) Top: distribution of point mutations over different mutations periods in 2,583 samples from unique donors. Middle: timing distribution of driver point mutations in the 50 most recurrent lesions. Bottom: distribution of driver mutations across cancer types; colour as defined in the inset. (**b**) Relative timing of the 50 most recurrent driver lesions, calculated as the odds ratio of early versus late clonal driver mutations versus background (green, purple) or clonal versus subclonal (blue, red). Odds ratios overlapping 1 in less than 5% of bootstrap samples are considered significant and have been coloured. (**c**) Relative timing of *TP53* mutations across cancer types, coloured as in (**b**). (**d**) Estimated number of unique lesions (genes) contributing 50% of all driver mutations in different timing epochs. Error bars denote the range between 0 and 1 pseudocounts. (**e**, **f**, **g**) **Relative ordering of somatic events.** Preferential ordering diagrams of somatic copy number events and driver point mutations within tumour types, for (**e**) colorectal adenocarcinoma, (**f**) pancreatic neuro-endocrine cancer and (**g**) glioblastoma. Probability distributions show the uncertainty of timing for specific events in the cohort. Events with odds above 10 (either earlier or later) are highlighted.

Figure 4. Timing of signatures

(**a**) Fold changes in signature exposures between early and late clonal stages for all tumours. Each violin shows the distribution of exposure changes across tumour types in one signature. Signatures are sorted by the ratio of tumours with a positive signature change. (**b**) Fold changes in signature exposures in individual tumours (early *vs*. late clonal). Within cancer types, tumours are ordered according to hierarchical clustering. White indicates inactive signatures.

Figure 5. Real-time estimation of mutational landmarks

(**a**) Mutation rate acceleration inferred from paired samples. CpG>TpG mutations (right) display a lower acceleration rate compared to all point mutations (left). (**b**)

Time of occurrence of whole genome duplications in individual patients, split by tumour type, based on CpG>TpG mutations and patient age. Results are shown for a 5x acceleration of the mutation rate. (**c**) Median time of WGD occurrence per cancer type, as a function of CpG>TpG acceleration. (**d**) Timing of subclonal diversification using CpG>TpG mutations in individual patients. (**e**) Comparison of inferred median occurrence of WGD and subclonal diversification.

Figure 6. Cancer timelines

Typical timelines of tumour development, for (**a**) glioblastoma, (**b**) colorectal adenocarcinoma, (**c**) squamous cell lung cancer, (**d**) ovarian adenocarcinoma, and (**e**) pancreatic adenocarcinoma. Each timeline represents the length of time, in years, between the fertilised egg and the median age of diagnosis per cancer type. Point estimates for major events, such as WGD and the emergence of the MRCA are used to define early, intermediate, late and subclonal stages of tumour evolution approximately in chronological time. Driver mutations and copy number aberrations are shown in each stage according to their preferential timing, as defined by relative ordering. Mutational signatures that fluctuate during tumour evolution, either considerably (median change +/- 20%), or consistently (75% samples change in the same direction) are annotated as well.

Methods

Timing of gains

We used three related approaches to calculate the timing of copy number gains (see **Supplementary Methods**, section 1). In brief, the common feature is that the expected variant allele frequency of a mutation is related to the underlying number of alleles carrying a mutation according to the formula

$$E[X] = n\, m\, f\, /\, [N\, (1-\rho) + C\, \rho]$$

Here X is the number of reads, n denotes the coverage of the locus, the mutation copy number m is the number of alleles carrying the mutation (which is usually inferred), f is the frequency of the clone carrying the given mutation ($f = 1$ for clonal mutations). N is the normal copy number (2 on autosomes, 1 or 2 for chromosome X and 0 or 1 for chromosome Y), C the total copy number of the tumour and ρ the purity of the sample.

The number of mutations at each allelic copy number then informs about the time when the gain has occurred. The basic formulae for timing each gain are, depending on the copy number configuration:

Copy number 2+1: $T = 3\, n_2\, /\, (2n_2 + n_1)$

Copy number 2+2: $T = 2\, n_2\, /\, (2n_2 + n_1)$

Copy number 2+0: $T = 2\, n_2\, /\, (2n_2 + n_1)$

Here 2+1 refers to major and minor copy number of 2 and 1, respectively. Methods differ slightly in how the number of mutations present on each allele are calculated and how uncertainty is handled.

Timing of mutations

The mutation copy number m and the clonal frequency f is calculated according to the principles indicated above. Details can be found in **Supplementary Methods**, section 1.2. Mutations with $f = 1$ are denotes as clonal, and mutations with $f < 1$ as *subclonal*. Mutations with $f = 1$ and $m > 1$ are denoted as *early clonal* (coamplified). In cases with $f = 1$, $m = 1$ and $C > 2$, mutations were annotated as *late clonal*, if the minor copy number was 0, otherwise *clonal* [*unspecified*] (**Supplementary Methods**, section 1.2.)

Timing of driver mutations

A catalogue of driver point mutations was provided by the PCAWG Drivers and Functional Interpretation Group[17]. The timing category was calculated as above. From the four timing categories, odds ratios of early/late clonal and clonal (early, late or unspecified clonal)/subclonal were calculated for driver mutations against the distribution of all other mutations in the samples with each particular driver. The background distribution of these odds ratios was assessed with 1000 bootstraps (**Supplementary Methods**, section 3.1.)

Integrative timing

For each pairs of driver point mutations and recurrent copy number variants it was established what the ordering of the given pair was (earlier, later or unspecified). The information underlying this decision was derived from the timing of each driver point mutation, as well as from the timing status of clonal and subclonal copy number segments. These tables were aggregated across all samples and a sports statistics model was employed to calculate the overall ranking of driver mutations. A full description is given in **Supplementary Methods**, section 3.2.

Timing of mutational signatures

Mutational trinucleotide substitution signatures, as defined by the PCAWG Mutational Signatures Working Group[20], were refit to samples with observed signature activity, after splitting point mutations into either of the 4 timing categories. Time-resolved exposures were calculated using non-negative linear least squares. Full details are given in **Supplementary Methods**, section 4.

Real-time estimation of copy number gains

For tumours with multiple time points, the set of mutations shared between diagnosis and relapse (n_D) and those specific to the relapse (n_R) was calculated. The rate acceleration was calculated as $a = n_R / n_D \times t_D / t_R$. This analysis was performed separately for all substitutions and for CpG>TpG mutations.

The correction for transforming an estimate of a copy number gain in mutation time into chronological time depends not only on the rate acceleration, but also on the time at which this acceleration occurred. As this is generally unknown, we performed Monte Carlo simulations of rate accelerations spanning an interval of 0.66 to 1.0 of relative time and averaged the results. Subclonal mutations were assumed to occur at full acceleration. The proportion of subclonal mutations was divided by the number of identified subclones, thus conservatively assuming branching evolution. Full details are given in **Supplementary Methods**, section 5.

Supplementary Figure Legends

Supplementary Figure 1. Summary of all results obtained per cancer type

(**a**) Clustered heatmaps of mutational timing estimates for gained segments, per patient. Colours as indicated in main text: green represents early clonal events, purple represents late clonal. (**b**) Relative ordering of copy-number events and driver mutations across all samples per cancer type. (**c**) Distribution of mutations across early clonal, late clonal and subclonal stages, for the most common driver genes per cancer type. A maximum of 10 driver genes are shown. (**d**) Clustered mutational signature fold changes between early clonal and late clonal stages, per patient. Green and purple indicate, respectively, a signature decrease and increase in late clonal from early clonal mutations. Inactive signatures are coloured white. (**e**) As in (**d**) but for clonal *vs*. subclonal stages. Blue indicates a signature decrease and red an increase in subclonal from clonal mutations. (**f**) Typical timeline of tumour development, per cancer type.

Supplementary Figure 2. Correlation between league model and Bradley-Terry model order of events.

The two approaches for determining the order of recurrent somatic mutations and copy number events are compared directly for each tumour type. We show how the order derived from the league model compares to that derived from the Bradley-Terry model, quantified by Spearman's rank correlation coefficient.

Supplementary Figure 3. Timing of signatures

(**a**) Fold changes in signature exposures between clonal and subclonal stages for all tumours. Each violin shows the distribution of exposure changes across tumour types in one signature. Signatures are sorted by the ratio of tumours with a positive signature change. (**b**) Fold changes in signature exposures in individual tumours (clonal *vs*. subclonal). Within cancer types, tumours are ordered according to hierarchical clustering. White indicates inactive signatures.

Author contributions

MG, CJ, IL, SG, PA, DR, DGL, PTS and PVL performed timing of point mutations and copy number gains. SG and MG performed qualitative timing of driver point mutations. IL, TJM, DR, DGL, DCW and GG performed relative timing of somatic driver events and implemented integrative models. CJ, YR, PVL and QDM performed timing of mutational signatures. MG performed real-time estimation of whole-genome duplication and subclonal diversification. CJ, MG, IL, YR, DR and PVL constructed cancer timelines. MG, CJ, IL, SCD, SG, TJM, YR, PA, JD, PCB, DDB, VM, QDM, PTS, DCW and PVL interpreted the results. SCD, IL, JW, AD, IVG, KeY, GM, MP, SM, ND, KaY, SSe, KH, MT, JD, DGL, DR, JL, MC, SCS, YJ, FM, VM, HZ, WW, QDM, DCW and PVL performed subclonal architecture analysis. SCD, IL, KK, VM, MP, XY, DGL, SSc, RB, MI, MS, DCW and PVL performed copy number analysis. JW, SCD, IL, KH, DGL, KK, DR, DCW, QDM and PVL derived a consensus of copy number analysis results. KaY, MT, AD, SCD, IL, DCW, MG, PVL, QDM and WW derived a consensus of subclonal architecture results. YF and WW contributed to subclonal mutation calls. PTS, DCW and PVL coordinated the study. MG, CJ, PTS, YR, IL, QDM, DCW and PVL wrote the manuscript.

Acknowledgements

This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202), and the Wellcome Trust (FC001202). MT and JD are postdoctoral fellows supported by the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie Grant Agreement No. 747852-SIOMICS and 703594-DECODE). JD is a postdoctoral fellow of the FWO. FM, GM and KeY would like to acknowledge the support of the University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited. GM, KeY and FM were funded by CRUK core grants C14303/A17197 and A19274. SSe and YJ are supported by NIH R01 CA132897. HZ is supported by grant NIMH086633 and an endowed Bao-Shan Jing Professorship in Diagnostic Imaging. WW is supported by the U.S. National

Cancer Institute (1R01 CA183793 and P30 CA016672). PTS was supported by U24CA210957 and 1U24CA143799. DCW is funded by the Li Ka Shing foundation. PVL is a Winton Group Leader in recognition of the Winton Charitable Foundation’s support towards the establishment of The Francis Crick Institute.

Members of the PCAWG Evolution and Heterogeneity Working Group

Stefan C. Dentro[1,2,3,*], Ignaty Leshchiner[4,*], Moritz Gerstung[5,*], Clemency Jolly[1,*], Kerstin Haase[1,*], Jeff Wintersinger[6,*], Pavana Anur[7], Rameen Beroukhim[4], Paul C. Boutros[6,8], David D. Bowtell[9,10], Peter J. Campbell[2], Elizabeth L. Christie[9], Marek Cmero[11], Yupeng Cun[12], Kevin Dawson[2], Jonas Demeulemeester[1,13], Amit Deshwar[6], Nilgun Donmez[14], Roland Eils[15,16], Yu Fan[17], Matthew Fittall[1], Dale W. Garsed[9], Gad Getz[4], Santiago Gonzalez[5], Gavin Ha[4], Marcin Imielinski[18,19], Yuan Ji[20,21], Kortine Kleinheinz[15,16], Juhee Lee[22], Henry Lee-Six[2], Dimitri G. Livitz[4], Geoff Macintyre[23], Salem Malikic[14], Florian Markowetz[23], Inigo Martincorena[2], Thomas J. Mitchell[2,24], Ville Mustonen[25], Layla Oesper[26], Martin Peifer[12], Myron Peto[7], Benjamin J. Raphael[27], Daniel Rosebrock[4], Yulia Rubanova[6], S. Cenk Sahinalp[28], Adriana Salcedo[8], Matthias Schlesner[15], Steve Schumacher[4], Subhajit Sengupta[20], Lincoln D. Stein[8], Maxime Tarabichi[1], Ignacio Vázquez-García[2,24], Shankar Vembu[6], Wenyi Wang[17], David A. Wheeler[29], Tsun-Po Yang[12], Xiaotong Yao[18,19], Fouad Yousif[8], Kaixian Yu[17], Ke Yuan[23,30], Hongtu Zhu[17], Quaid D. Morris[6,#], Paul T. Spellman[7,#], David C. Wedge[3,#], Peter Van Loo[1,13,#]

[1]The Francis Crick Institute, London NW1 1AT, United Kingdom; [2]Wellcome Trust Sanger Institute, Cambridge CB10 1SA, United Kingdom; [3]Big Data Institute, University of Oxford, Oxford OX3 7LF, United Kingdom; [4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [5]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Cambridge CB10 1SD, United Kingdom; [6]University of Toronto, Toronto, ON M5S 3E1, Canada; [7]Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR 97231, USA; [8]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada; [9]Peter MacCallum Cancer Centre, Melbourne, VIC 3052, Australia; [10]Garvan Institute of Medical Research, Sydney, NSW 2010, Australia; [11]University of Melbourne, Melbourne, VIC 3010, Australia; [12]University of Cologne, 50931 Cologne, Germany; [13]Department of Human Genetics, University of Leuven, B-3000 Leuven, Belgium; [14]Simon Fraser University, Burnaby, BC V5A1S6, Canada; [15]German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany; [16]Heidelberg University, 69120 Heidelberg, Germany; [17]The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; [18]Weill Cornell Medicine, New York, NY 10065, USA; [19]New York Genome Center, New York, NY 10013, USA; [20]NorthShore University HealthSystem, Evanston, IL 60201, USA; [21]The

University of Chicago, Chicago, IL 60637, USA; [22]University of California Santa Cruz, Santa Cruz, CA 95064, USA; [23]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, United Kingdom; [24]University of Cambridge, Cambridge CB2 0QQ, United Kingdom; [25]University of Helsinki, 00014 Helsinki, Finland; [26]Carleton College, Northfield, MN 55057, USA; [27]Princeton University, Princeton, NJ 08540, USA; [28]Indiana University, Bloomington, IN 47405, USA; [29]Baylor College of Medicine, Houston, TX 77030, USA; [30]University of Glasgow, Glasgow G12 8RZ, United Kingdom.

*: These authors contributed equally

[#]: These authors jointly directed the work

Figure 1. Principles of timing mutations

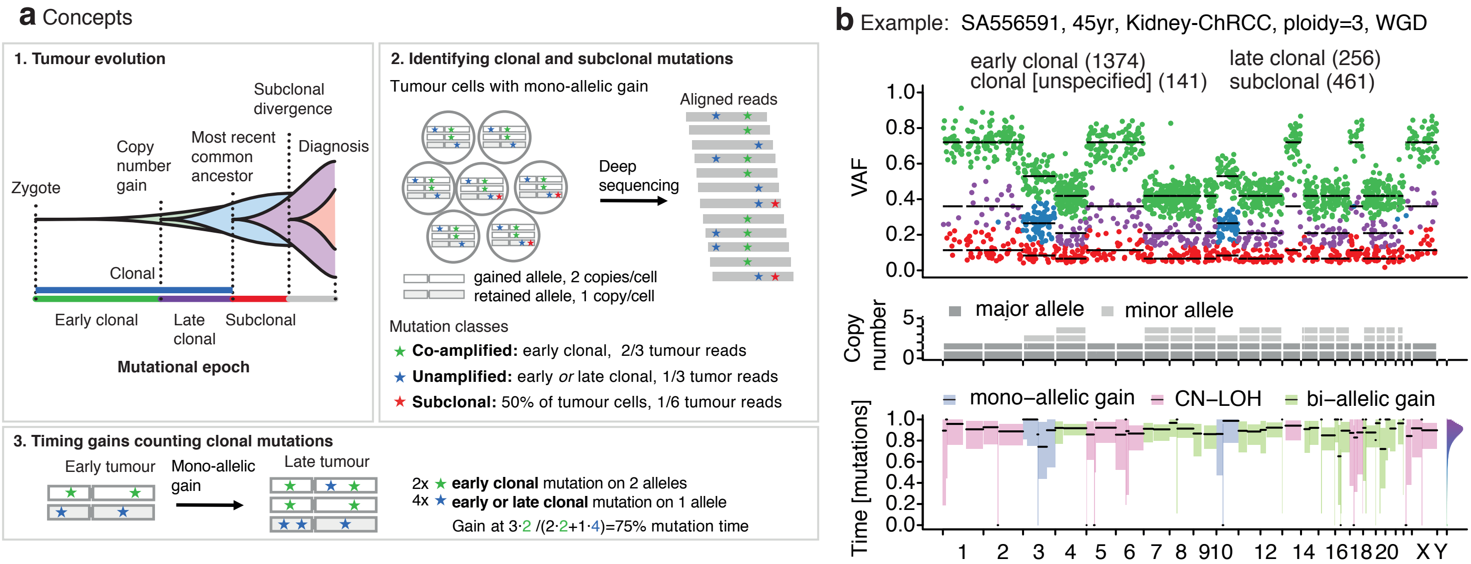


Figure 2: Pan-cancer timing patterns of arm-level gains

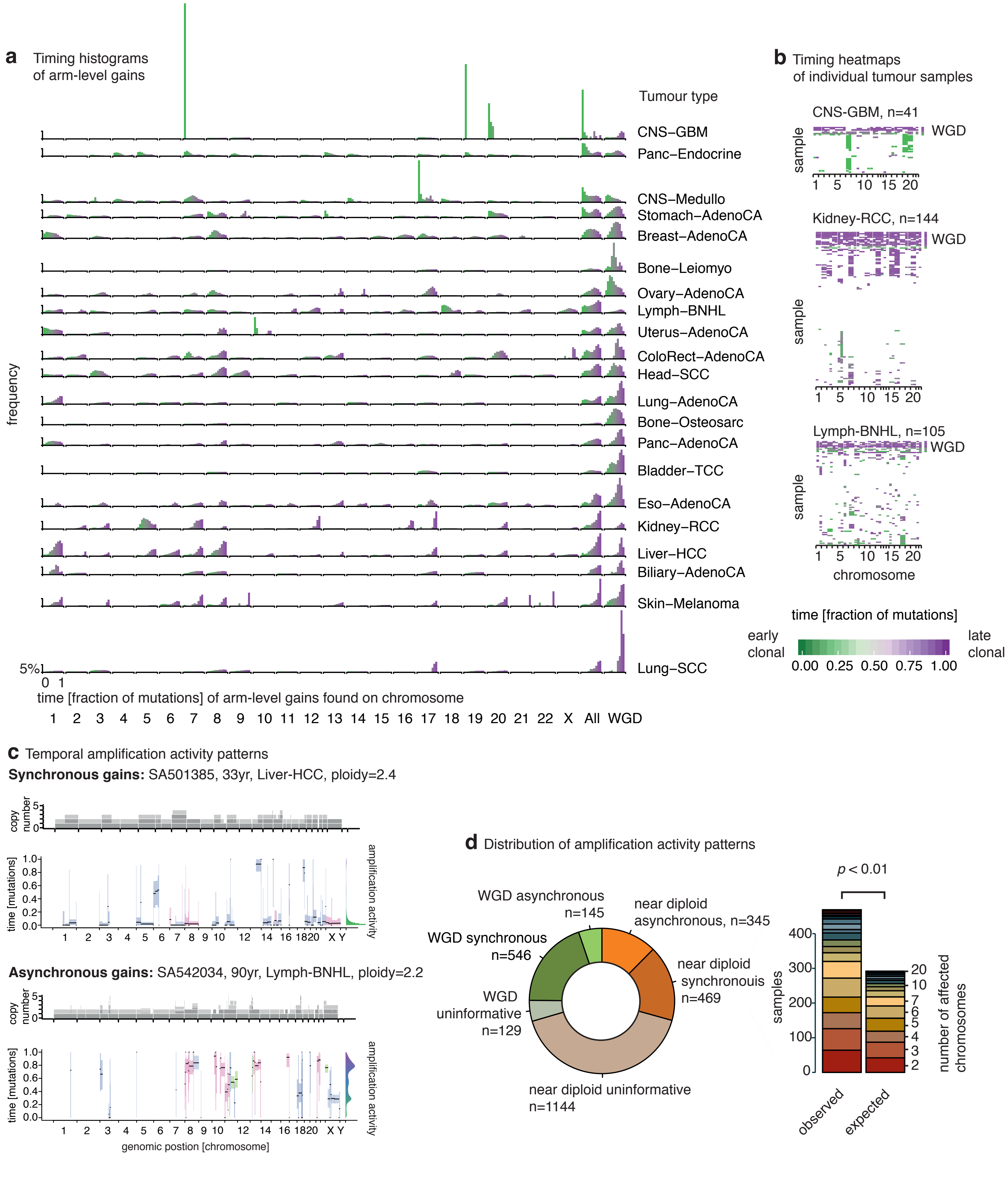


Figure 3: Timing of driver mutations and relative ordering of somatic events

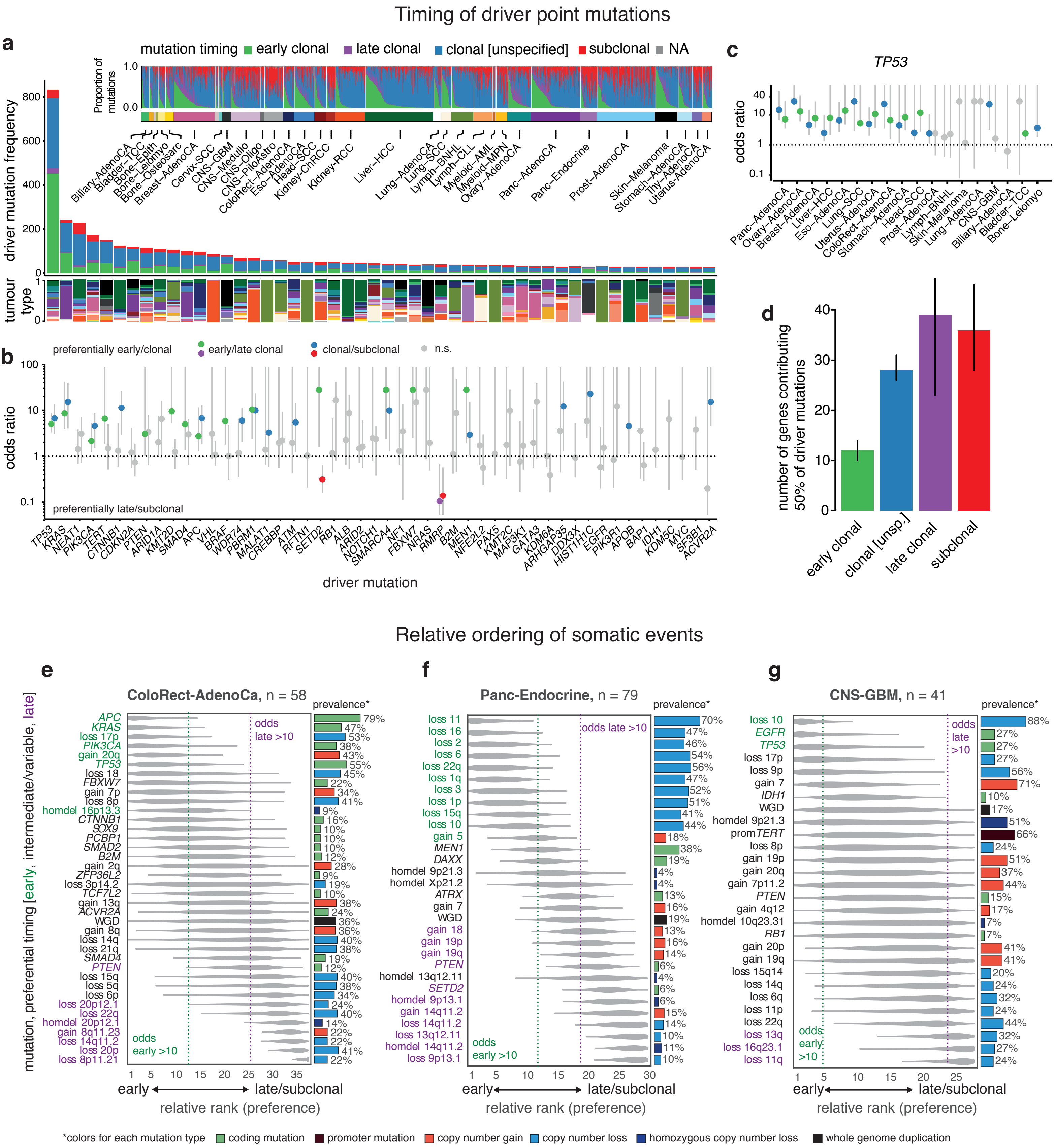


Figure 4: Evolution of signatures: early clonal vs late clonal

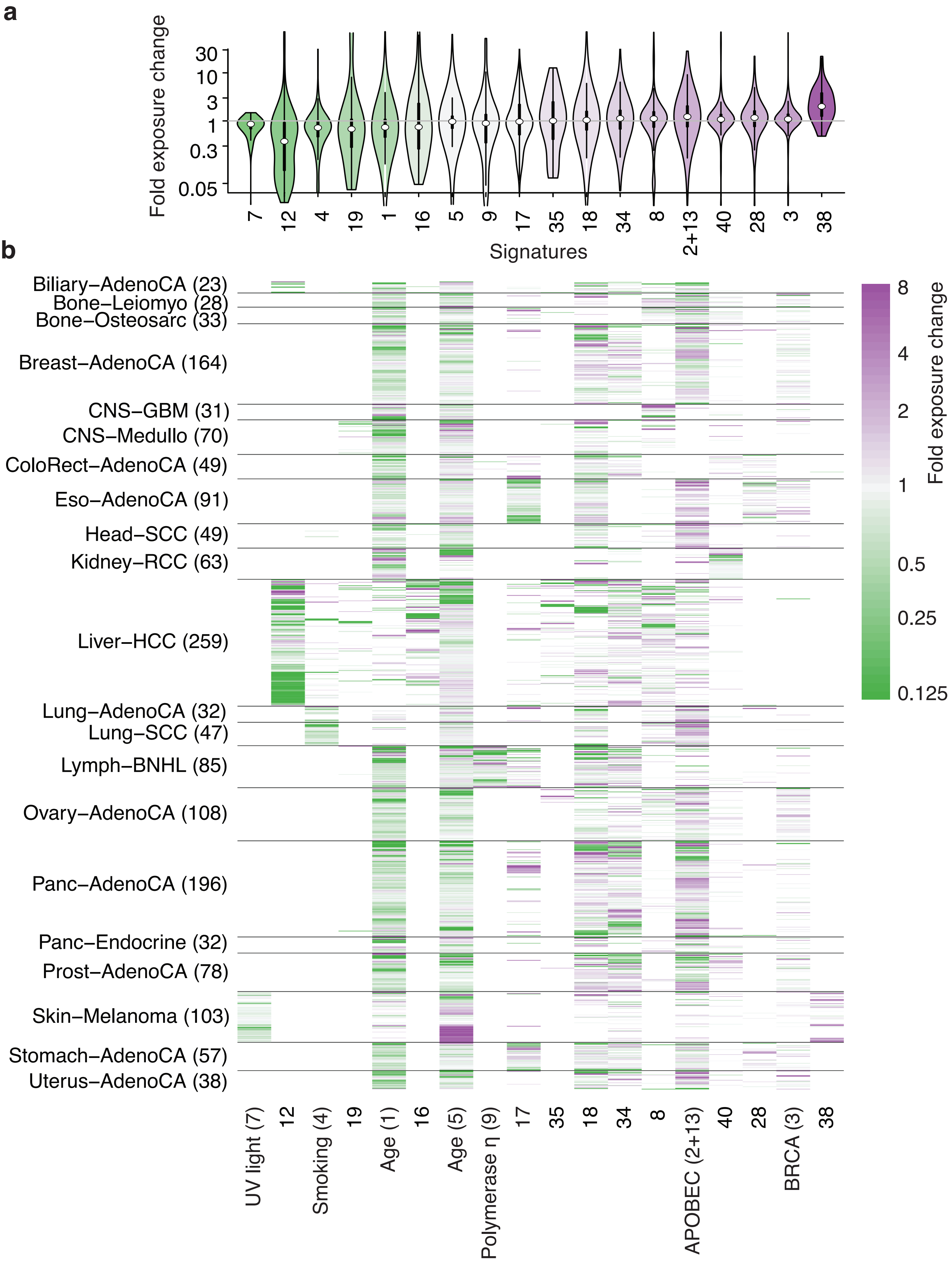


Figure 5: Real-time estimates

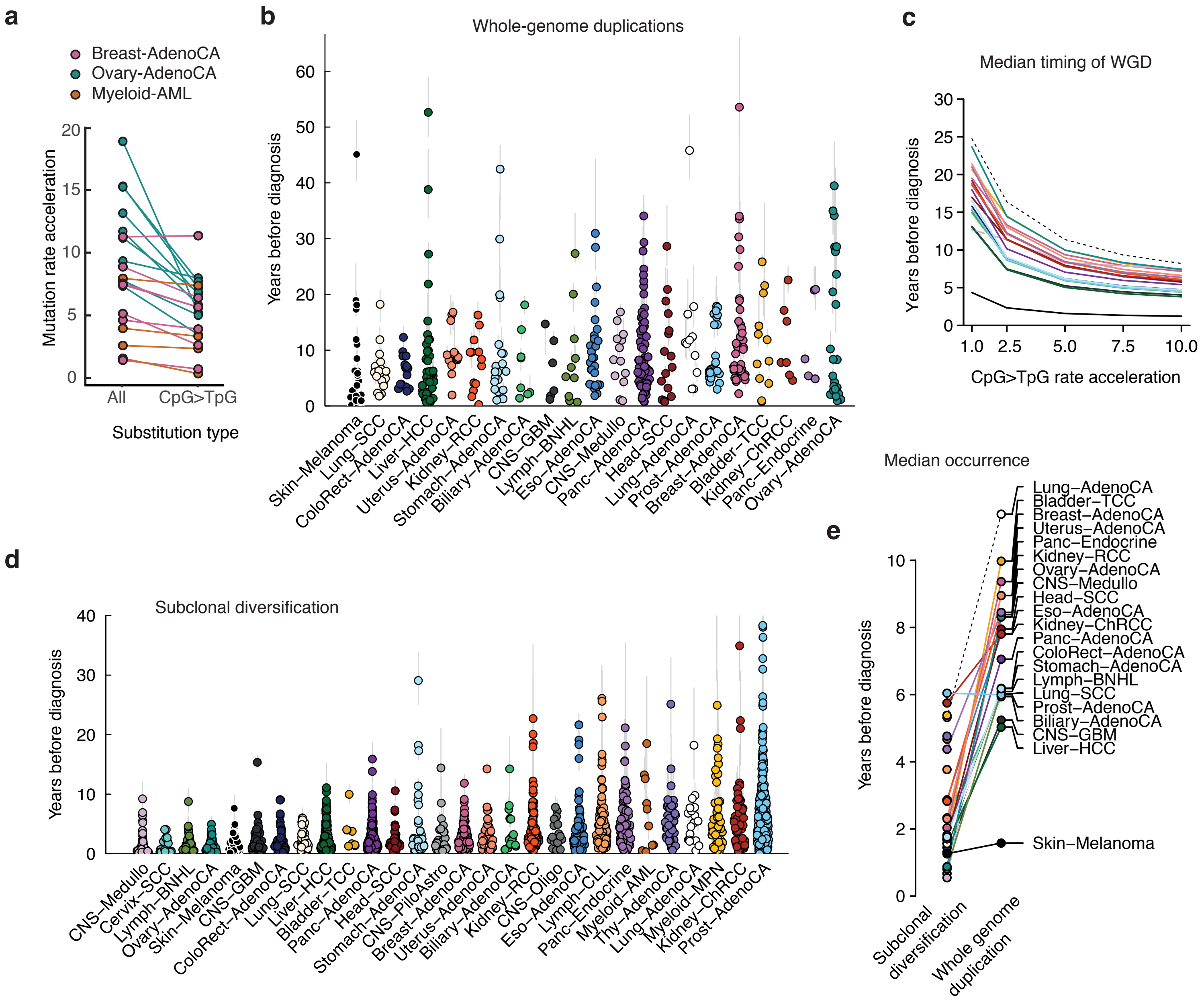
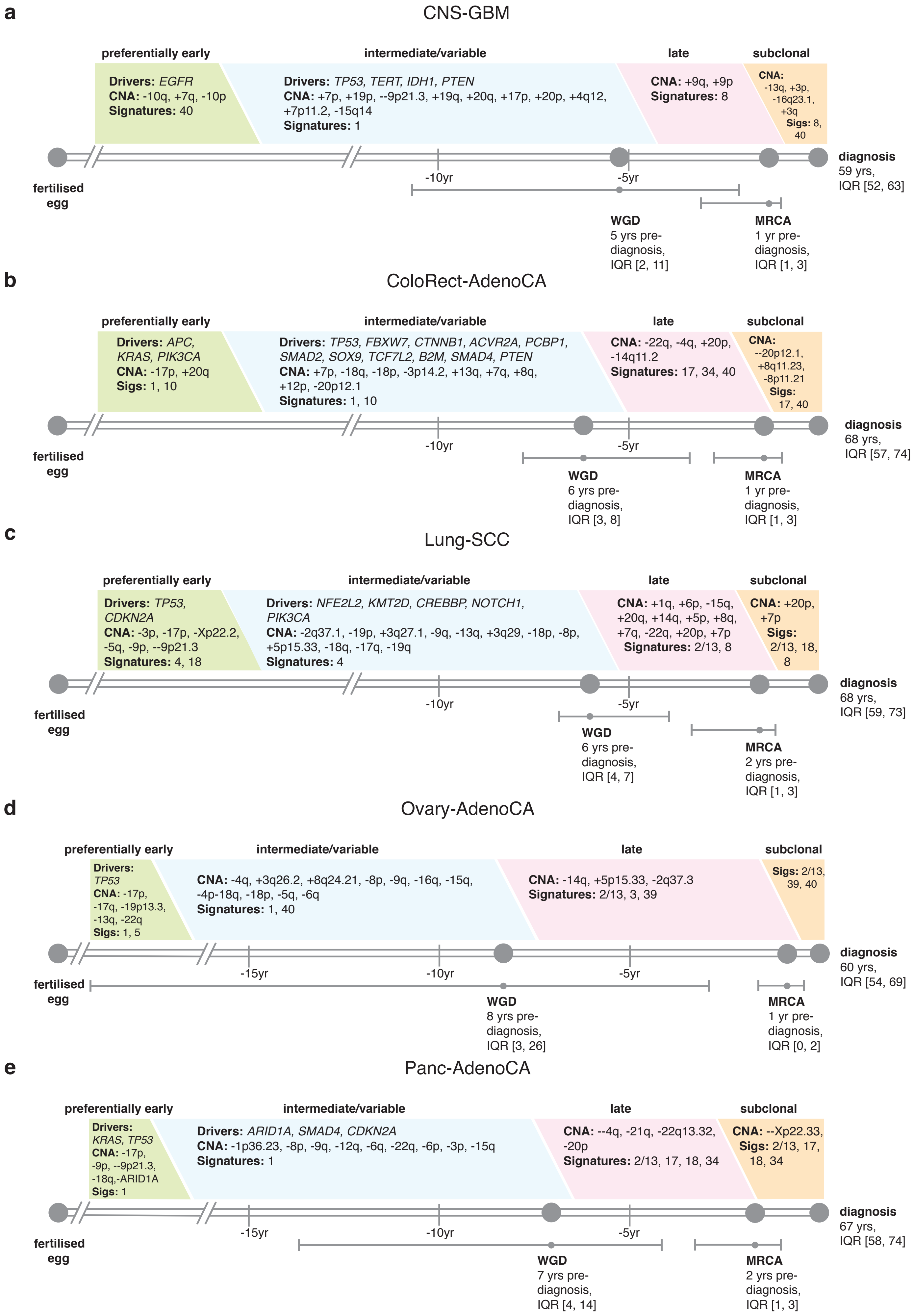


Figure 6: Oncogenic timelines


a
CNS-GBM
preferentially early
Drivers: EGFR
CNA: -10q, +7q, -10p
Signatures: 40
intermediate/variable
Drivers: TP53, TERT, IDH1, PTEN
CNA: +7p, +19p, --9p21.3, +19q, +20q, +17p, +20p, +4q12, +7p11.2, -15q14
Signatures: 1
late
CNA: +9q, +9p
Signatures: 8
subclonal
CNA: -13q, +3p, -16q23.1, +3q
Sigs: 8, 40
fertilised egg
-10yr
-5yr
diagnosis
59 yrs, IQR [52, 63]
WGD
5 yrs pre-diagnosis, IQR [2, 11]
MRCA
1 yr pre-diagnosis, IQR [1, 3]
b
ColoRect-AdenoCA
preferentially early
Drivers: APC, KRAS, PIK3CA
CNA: -17p, +20q
Sigs: 1, 10
intermediate/variable
Drivers: TP53, FBXW7, CTNNB1, ACVR2A, PCBP1, SMAD2, SOX9, TCF7L2, B2M, SMAD4, PTEN
CNA: +7p, -18q, -18p, -3p14.2, +13q, +7q, +8q, +12p, -20p12.1
Signatures: 1, 10
late
CNA: -22q, -4q, +20p, -14q11.2
Signatures: 17, 34, 40
subclonal
CNA: --20p12.1, +8q11.23, -8p11.21
Sigs: 17, 40
fertilised egg
-10yr
-5yr
diagnosis
68 yrs, IQR [57, 74]
WGD
6 yrs pre-diagnosis, IQR [3, 8]
MRCA
1 yr pre-diagnosis, IQR [1, 3]
c
Lung-SCC
preferentially early
Drivers: TP53, CDKN2A
CNA: -3p, -17p, -Xp22.2, -5q, -9p, --9p21.3
Signatures: 4, 18
intermediate/variable
Drivers: NFE2L2, KMT2D, CREBBP, NOTCH1, PIK3CA
CNA: -2q37.1, -19p, +3q27.1, -9q, -13q, +3q29, -18p, -8p, +5p15.33, -18q, -17q, -19q
Signatures: 4
late
CNA: +1q, +6p, -15q, +20q, +14q, +5p, +8q, +7q, -22q, +20p, +7p
Signatures: 2/13, 8
subclonal
CNA: +20p, +7p
Sigs: 2/13, 18, 8
fertilised egg
-10yr
-5yr
diagnosis
68 yrs, IQR [59, 73]
WGD
6 yrs pre-diagnosis, IQR [4, 7]
MRCA
2 yrs pre-diagnosis, IQR [1, 3]
d
Ovary-AdenoCA
preferentially early
Drivers: TP53
CNA: -17p, -17q, -19p13.3, -13q, -22q
Sigs: 1, 5
intermediate/variable
CNA: -4q, +3q26.2, +8q24.21, -8p, -9q, -16q, -15q, -4p-18q, -18p, -5q, -6q
Signatures: 1, 40
late
CNA: -14q, +5p15.33, -2q37.3
Signatures: 2/13, 3, 39
subclonal
Sigs: 2/13, 39, 40
fertilised egg
-15yr
-10yr
-5yr
diagnosis
60 yrs, IQR [54, 69]
WGD
8 yrs pre-diagnosis, IQR [3, 26]
MRCA
1 yr pre-diagnosis, IQR [0, 2]
e
Panc-AdenoCA
preferentially early
Drivers: KRAS, TP53
CNA: -17p, -9p, --9p21.3, -18q,-ARID1A
Sigs: 1
intermediate/variable
Drivers: ARID1A, SMAD4, CDKN2A
CNA: -1p36.23, -8p, -9q, -12q, -6q, -22q, -6p, -3p, -15q
Signatures: 1
late
CNA: --4q, -21q, -22q13.32, -20p
Signatures: 2/13, 17, 18, 34
subclonal
CNA: --Xp22.33,
Sigs: 2/13, 17, 18, 34
fertilised egg
-15yr
-10yr
-5yr
diagnosis
67 yrs, IQR [58, 74]
WGD
7 yrs pre-diagnosis, IQR [4, 14]
MRCA
2 yrs pre-diagnosis, IQR [1, 3]

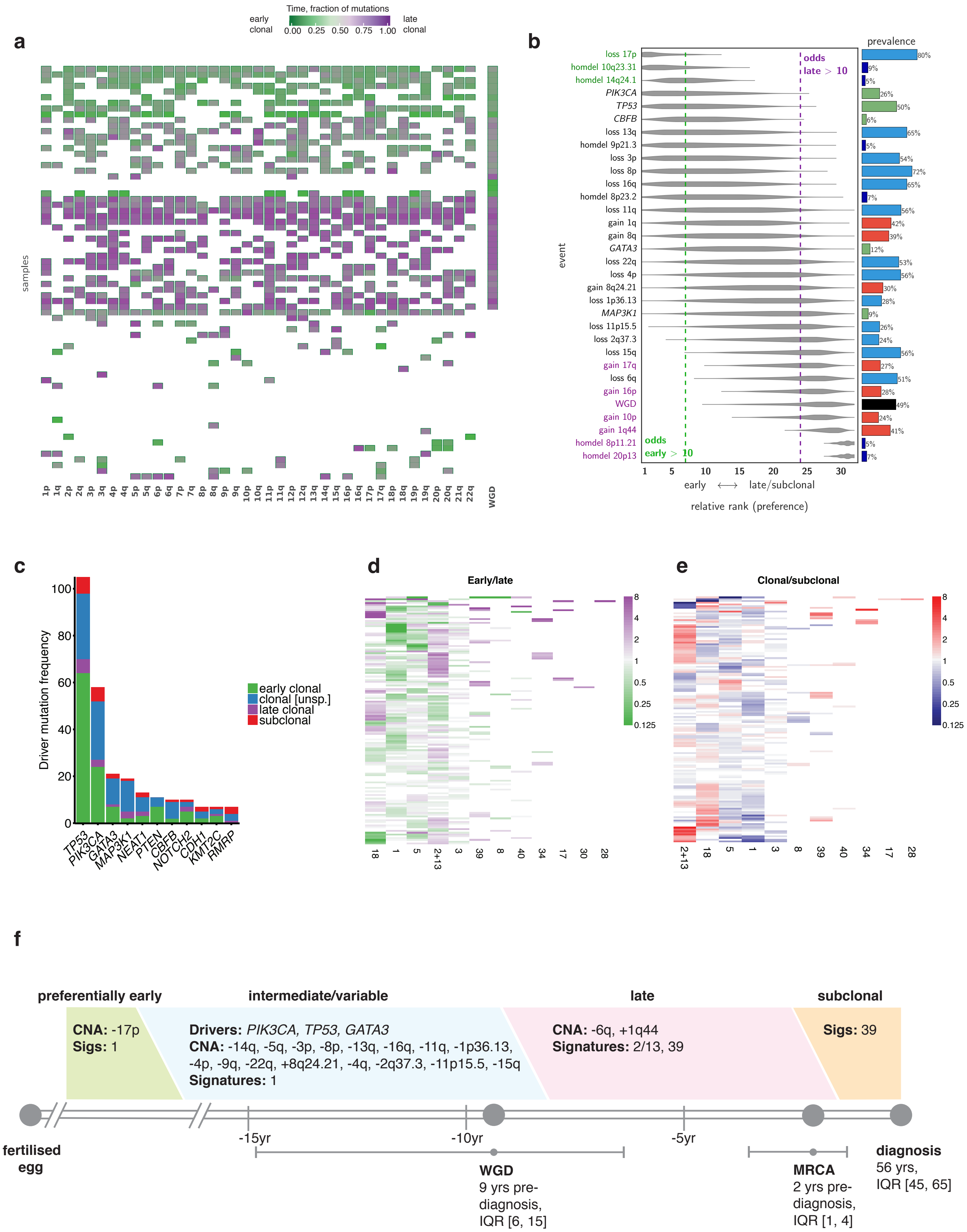
Appendix B

The evolutionary history of breast adenocarcinoma

This appendix contains the evolutionary history of breast adenocarcinoma, which was constructed as part of Gerstung et al. (2017). It is referenced in Chapter 5.

The figure on the next page shows the summary of all results obtained per cancer type (a) Clustered heatmaps of mutational timing estimates for gained segments, per patient. Colours as indicated in main text: green represents early clonal events, purple represents late clonal. (b) Relative ordering of copy-number events and driver mutations across all samples per cancer type. (c) Distribution of mutations across early clonal, late clonal and subclonal stages, for the most common driver genes per cancer type. A maximum of 10 driver genes are shown. (d) Clustered mutational signature fold changes between early clonal and late clonal stages, per patient. Green and purple indicate, respectively, a signature decrease and increase in late clonal from early clonal mutations. Inactive signatures are coloured white. (e) As in (d) but for clonal vs. subclonal stages. Blue indicates a signature decrease and red an increase in subclonal from clonal mutations. (f) Typical timeline of tumour development, per cancer type.

Supplementary Fig. 1. Breast-AdenoCA



Appendix C

Published works

Named authorships

Wedge, D. C.*, Gundem, G.*, Mitchell, T.*, Martincorena, I., Ghori, M., Zamora, J., Butler, A., Whitaker, H., Kote-Jarai, Z., Alexandrov, L. B., Van Loo, P., Massie, C. E., **Dentro, S.**, Warren, A. Y., Verrill, C., Berney, D., Dennis, N., Merson, S., Hawkins, S., Howat, W., Yu, Y., Lambert, A., Kay, J., Kremeyer, B., Karaszi, K., Luxton, H., Camacho, N., Marsden, L., Edwards, S., Matthews, L., Bo, V., Leongamornlert, D., McLaren, S., Ng, A., Yu, Y., Zhang, H., Dadaev, T., Thomas, S., Easton, D. F., Ahmed, M., Bancroft, E., Fisher, C., Livni, N., Nicol, D., Tavaré, S., Gill, P., Greenman, C., Khoo, V., Van As, N., Kumar, P., Ogden, C., Cahill, D., Thompson, A., Mayer, E., Rowe, E., Dudderidge, T., Gnanapragasam, V., Shah, N. C., Raine, K., Jones, D., Menzies, A., Stebbings, L., Teague, J., Hazell, S., de Bono, J., Attard, G., Isaacs, W., Visakorpi, T., Fraser, M., Boutros, P. C., Bristow, R. G., Workman, P., Sander, C., The TCGA consortium, Hamdy, F. C., Futreal, A., McDermott, U., Al-Lazikani, B., Lynch, A .G., Bova, G. S., Foster, C. S., Brewer, D., Neal, D., Cooper, C. S., and Eeles, R. A. Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. Nature Genetics - Manuscript accepted.

Linch, M.*, Goh, G.*, Hiley, C., Shanmugabavan, Y., McGranahan, N., Rowan, A., Wong, Y. N. S., King, H., Furness, A., Freeman, A., Linares, J., Akarca, A., Herrero, J., Rosenthal, R., Harder, N., Schmidt, G., Wilson, G. A., Birkbak, N. J., Mitter, R., **Dentro, S.**, Cathcart, P., Arya, M., Johnston, E., Scott, R., Hung, M., Emberton, M., Attard, G., Szallasi, Z., Punwani, S., Quezada, S. A., Marafioti, T., Gerlinger, M., Ahmed, H. U., and Swanton, C. (2017). Intratumoural evolutionary landscape of high-risk prostate cancer: The PROGENY study of genomic and immune parameters. Annals of Oncology, 28(10):2472–2480.

Jamal-Hanjani, M.*, Wilson, G. A.*, McGranahan, N.*, Birkbak, N. J.*, Watkins, T. B.*, Veeriah, S., Shafi, S., Johnson, D. H., Mitter, R., Rosenthal, R., Salm, M., Horswell, S., Escudero, M., Matthews, N., Rowan, A., Chambers, T., Moore, D. A., Turajlic, S., Xu, H., Lee, S.-M., Forster, M. D., Ahmad, T., Hiley, C. T., Abbosh, C., Falzon, M., Borg, E., Marafioti, T., Lawrence, D., Hayward, M., Kolvekar, S., Panagiotopoulos, N., Janes, S. M., Thakrar, R., Ahmed, A., Blackhall, F., Summers, Y., Shah, R., Joseph, L., Quinn, A. M., Crosbie, P. A., Naidu, B., Middleton, G., Langman, G., Trotter, S., Nicolson, M., Remmen, H., Kerr, K., Chetty, M., Gomersall, L., Fennell, D. A., Nakas, A., Rathinam, S., Anand, G., Khan, S., Russell, P., Ezhil, V., Ismail, B., Irvin-Sellers, M., Prakash, V., Lester, J. F., Kornaszewska, M., Attanoos, R., Adams, H., Davies, H., **Dentro, S.**, Taniere, P., O'Sullivan, B., Lowe, H. L., Hartley, J. A., Iles, N., Bell, H., Ngai, Y., Shaw, J. A., Herrero, J., Szallasi, Z., Schwarz, R. F., Stewart, A., Quezada, S. A., Le Quesne, J., Van Loo, P., Dive, C., Hackshaw, A., and Swanton, C. (2017). Tracking the Evolution of Non–Small-Cell Lung Cancer. New England Journal of Medicine.

Dentro, S. C., Wedge, D. C., and Van Loo, P. (2017). Principles of Reconstructing the Subclonal Architecture of Cancers. Cold Spring Harbor Perspectives in Medicine.

Rabbie, R., Rashid, M., Arance, A. M., Sanchez, M., Tell-Marti, G., Potrony, M., Conill, C., van Doorn, R., **Dentro, S.**, Gruis, N. A., Corrie, P., Iyer, V., Robles-Espinoza, C. D., Puig-Butille, J. A., Puig, S., and Adams, D. J. (2017). Genomic analysis and clinical management of adolescent cutaneous melanoma. Pigment Cell & Melanoma Research, 30(3):307–316.

Sandhu, V., Wedge, D. C., Lothe, I. M. B., Labori, K. J., **Dentro, S. C.**, Buanes, T., Skrede, M. L., Dalsgaard, A. M., Munthe, E., Myklebost, O., Lingjærde, O. C., Børresen-Dale, A.-L., Ikdahl, T., Van Loo, P., Nord, S., and Kure, E. H. (2016). The Genomic Landscape of Pancreatic and Periampullary Adenocarcinoma. Cancer Research.

Manuscripts in preparation

Gerstung, M.*, Jolly, C.*, Leshchiner, I.*, **Dentro, S. C.***, Gonzalez, S., Mitchell, T. J., Rubanova, Y., Anur, P., Rosebrock, D., Yu, K., Tarabichi, M., Deshwar, A., Wintersinger, J., Klein- heinz, K., Vazquez-Garcia, I., Haase, K., Sengupta, S., Macintyre, G., Malikic, S., Donmez, N., Livitz, D. G., Cmero, M., Demeulemeester, J., Schumacher, S., Fan, Y., Yao, X., Lee, J., Schlesner, M., Boutros, P. C., Bowtell, D. D., Zhu, H., Getz, G., Imielinski, M., Beroukhim, R., Sahinalp, S. C., Ji, Y., Peifer, M., Markowetz, F., Mustonen, V., Yuan, K., Wang, W., Morris, Q. D., Spellman, P. T., Wedge, D. C., Van Loo, P., PCAWG Evolution and Heterogeneity Working Group, and PCAWG Network (2017). The evolutionary history of 2,658 cancers. bioRxiv, page 161562.

Dentro, S. C.*, Ignaty Leshchiner, I.*, Haase, K.*, Wintersinger, J.*, Deshwar, A. G.*, Tarabichi, M.*, Rubanova, Y., Yu, K., Vázquez-García, I., Macintyre, G., Kleinheinz, K., Livitz, D. G., Malikic, S., Donmez, N., Sengupta, S., Demeulemeester, J., Anur, P., Jolly, C., Cmero, M., Rosebrock, D., Schumacher, S., Fan, Y., Fittall, M., Yao, X., Lee, J., Schlesner, M., Zhu, H., Adams, D. J., Getz, G., Boutros, P., Imielinski, M., Beroukhim, R., Sahinalp, C. S., Ji, Y., Peifer, M., Martincorena, I., Markowetz, F., Mustonen, V., Yuan, K., Gerstung, M., Wang, W., Spellman, P. T., Morris, Q., Wedge, D. C., Van Loo, P., on behalf of the PCAWG Evolution and Heterogeneity Working Group and the PCAWG network. Pervasive intra-tumour heterogeneity and subclonal selection across cancer types. Manuscript in preparation.

Consortium authorships

Abbosh, C.*, Birkbak, N. J.,* Wilson, G. A.*, Jamal-Hanjani, M.,* Constantin, T.*, Salari, R.*, Quesne, J. L., Moore, D. A., Veeriah, S., Rosenthal, R., Marafioti, T., Kirkizlar, E., Watkins, T. B. K., McGranahan, N., Ward, S., Martinson, L., Riley, J., Fraioli, F., Bakir, M. A., GrOnroos, E., Zambrana, F., Endozo, R., Bi, W. L., Fennessy, F. M., Sponer, N., Johnson, D., Laycock, J., Shafi, S., Czyzewska-Khan, J., Rowan, A., Chambers, T., Matthews, N., Turajlic, S., Hiley, C., Lee, S. M., Forster, M. D., Ahmad, T., Falzon, M., Borg, E., Lawrence, D., Hayward, M., Kolvekar, S., Panagiotopoulos, N., Janes, S. M., Thakrar, R., Ahmed, A., Blackhall, F., Summers, Y., Hafez, D., Naik, A., Ganguly, A., Kareht, S., Shah, R., Joseph, L., Quinn, A. M., Crosbie, P., Naidu, B., Middleton, G., Langman, G., Trotter, S., Nicolson, M., Remmen, H., Kerr, K., Chetty, M., Gomersall, L., Fennell, D. A., Nakas, A., Rathinam, S., Anand, G., Khan, S., Russell, P., Ezhil, V., Ismail, B., Irvin-sellers, M., Prakash, V., Lester, J. F., Kornaszewska, M., Attanoos, R., Adams, H., Davies, H., Oukrif, D., Akarca, A. U., Hartley, J. A., Lowe, H. L., Lock, S., Iles, N., Bell, H., Ngai, Y., Elgar, G., Szallasi, Z., Schwarz, R. F., Herrero, J., Stewart, A., Quezada, S. A., Van Loo, P., Dive, C., Lin, C. J., Rabinowitz, M., Aerts, H. J., Hackshaw, A., Shaw, J. A., Zimmermann, B. G., the TRACERx Consortium, the PEACE Consortium, and Swanton, C. (2017). Phylogenetic ctDNA analysis depicts early stage lung cancer evolution. Nature.

McGranahan, N., Rosenthal, R., Hiley, C. T., Rowan, A. J., Watkins, T. B. K., Wilson, G. A., Birkbak, N. J., Veeriah, S., Van Loo, P., Herrero, J., and Swanton, C. (2017). Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. Cell, 171(6):1259–1271.