Methods and practice of detecting selection in human cancers

Marc J Williams

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Centre of Mathematics and Physics in the Life Sciences and Experimental Biology

University College London

February 18, 2019

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

Abstract

Cancer development and progression is an evolutionary process, understanding these evolutionary dynamics is important for treatment and diagnosis as how a cancer evolves determines its future prognosis. This thesis focuses on elucidating selective evolutionary pressures in cancers and somatic tissues using population genetics models and cancer genomics data.

First a model for the expected diversity in the absence of selection was developed. This neutral model of evolution predicts that under neutrality the frequency of subclonal mutations is expected to follow a power law distribution. Surprisingly more than 30% of cancer across multiple cohorts fitted this model.

The next part of the thesis develops models to explore the effects of selection given these should be observable as deviations from the neutral prediction. For this I developed two approaches. The first approach investigated selection at the level of individual samples and showed that a characteristic pattern of clusters of mutations is observed in deep sequencing experiments. Using a mathematical model, information encoded within these clusters can be used to measure the relative fitness of subclones and the time they emerge during tumour evolution. With this I observed strikingly high fitness advantages for subclones of above 20%. The second approach enables measuring recurrent patterns of selection in cohorts of sequenced cancers using dN/dS, the ratio of non-synonymous to synonymous mutations, a method originally developed for molecular species evolution. This approach demonstrates how selection coefficients can be extracted by combining measurements of dN/dS with the size of mutational lineages. With this approach selection coefficients were again observed to be strikingly high.

Finally I looked at population dynamics in normal colonic tissue given that many mutations accumulate in physiologically normal tissue. I found that the current view of stem cell dynamics was unable to explain sequencing data from individual colonic crypts. Some new models were proposed that introduce a longer time scale evolution that suppresses the accumulation of mutations which appear consistent with the data.

Impact Statement

In this thesis I have developed mathematical methods to measure evolutionary and population dynamics across different scales in human cancers and tissues. The principal results concerned unravelling evolutionary dynamics using genetics and mathematical modelling. Cancer development and progression is now widely recognised as an evolutionary system. Relapse and resistance to treatment can be viewed as evolutionary events and thus a better understanding of the evolutionary rules that govern cancer progression will likely have clinical impact in the future.

In terms of direct impact from this thesis, the results concerning neutral evolution in cancer, presented in Chapter 3 have somewhat surprisingly (at least to me!) stimulated quite a lot of debate in the field. This is perhaps due to selection being assumed to be pervasive and that the role of neutral evolution having been largely ignored. This has brought a new way of analysing cancer genomic data into the field. I hope that this work shows the importance of analysing data from cancers with this in mind.

Chapters 4 and 5 concentrated on quantifying selection in cancer. I showed a theoretical example where measuring selection coefficients allow for predicting future evolutionary trajectories. While such an approach would need to be validated, this potentially allows for patient specific approaches for rationalising treatment strategies and prognostication.

Acknowledgements

First of all I would like to thank Trevor Graham and Chris Barnes for supervising me for this PhD. I am grateful to you both for always giving me the freedom to explore my own ideas and could not have wished for better supervisors. Trevor, thank you for ideas, insights and thoughts which shaped much of this PhD. And Chris thank you for your rigorous statistical and mathematical input which improved much of this work immensely. Most of all thank you both for your confidence in me as a scientist.

A special thanks to Andrea Sottoriva and Ben Werner whose enthusiasm and insight were invaluable and always welcome. My PhD would not have been anywhere near as complete and more importantly fun without your input.

Thanks to the other members of the Evolution and Cancer lab, particularly Will Cross for your patience while teaching me the dark arts of bioinformatics and for your expert knowledge of the local selection of pubs! I am also indebted to Ibrahim al-Bakir, Laura Gay, Annie Baker and Chris Kimberley for their work in the lab that generated some of the data I used in the thesis. Thanks also to the other members of the Computational Systems and Synthetic Biology lab and my Complex cohort with whom I had many interesting discussions and much fun over the past few years.

Thanks to my friends and family. In particular thanks to my wonderful parents, Aled and Sue. Diolch i chdi, Dad am rannu dy frwdrydedd mewn gwyddionaeth a mathemateg er gwaethaf dy anallu yn y pwnc! And thanks to you, Mum for nurturing my mathematically inclined mind which is no doubt more yours than Dad's. Lastly a special thanks to Bethan for being a wonderful support and for always getting me to see any difficulties in a better and more positive light. I feel

immensely fortunate to have met you near the beginning of this journey and to have had you by my side throughout. I would not have done this nearly as well without your love and support these past few years. Diolch.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancers originate from somatic cells in the human body that have accumulated genetic alterations. These mutations modify the phenotype of the cells, which allows them to escape the precise homeostatic regulation of cells under normal conditions in the body. Viewed though the lens of evolutionary biology the transformation of normal cells into cancerous ones is evolution in action. Cancer initiation, progression, treatment resistance and subsequent disease relapse can and perhaps *should* all be viewed as evolutionary events. Indeed the oft quoted *Nothing in biology makes sense except in the light of evolution* comes to mind, and applies to cancer biology equally well (Dobzhansky, 1983).

While survival rates in the UK have doubled in the previous 40 years, cancer still exerts a considerable burden on society. In the UK more than 150,000 people die of the disease each year and 350,000 new cases are reported annually (*Cancer Research UK*). This is despite the considerable efforts of successive governments, scientists and funding bodies over the previous 50 years. Understanding how evolutionary forces shape cancer progression is likely to be key in new strategies to combat the disease (Greaves & Maley, 2012). For example, recognising that drug treatment introduces a selective pressure has recently led to the idea that long term control of the disease may be achieved by managing the evolution of resistance rather than attempting to completely eradicate the disease (Enriquez-Navas *et al.,* 2016). With this in mind, this thesis is principally concerned with evolution in cancer and somatic tissues, and how to measure and quantify its contribution.

I will argue that evolution is best described in terms of mathematical equations. The combination of a mathematical framework with sequencing data from human cancers allows for inferring and quantifying aspects of the evolutionary process *in vivo* from a single time point. The remainder of this chapter will discuss relevant concepts and questions related to this approach.

1.1 Cancer evolution

The current evolutionary perspective on cancer was first described by Nowell in the 1970's (Nowell, 1976). In this paradigm, physiologically normal cells acquire a growth advantage over their neighbours and clonally expand. Following this initiating event, subsequent alterations may be acquired, inducing further clonal expansions and increasing fitness. This process ultimately leads to a malignant tumour. Viewed from this angle the evolutionary dynamics can be divided into 2 stages; firstly, how the transformation of a normal cell to a neoplastic cell proceeds and secondly, how the evolution within growing neoplasms progresses. From a patient perspective, understanding the first stage is key to the early detection and possible prevention of the disease. If we understand how the homeostatic regulation is hijacked by cancer, it may provide clues as to how to spot early signs and possible avenues for early intervention. The second stage is more relevant to later stage cancer, where key questions remain largely unanswered. Such as, how resistance to treatment emerges and how primary tumours disseminate to other sites within the body. Both questions are key to controlling the progression of the disease. Broadly, this thesis will discuss the second stage of evolution: the evolution within tumours in Chapters 3, 4 and 5. Chapter 6 discusses the evolutionary dynamics in the pre-transformation stage and will explore approaches to measure stem cell dynamics during normal homeostasis.

The advent of high throughput genomics and its application to cancer has validated cancer progression as an evolutionary system. Applications of genomics in cancer has revealed that complex genetic architectures are a feature of all cancer types (Stratton, 2011). Cancers have been shown to harbour many thousands of

point mutations (Lawrence *et al.,* 2014), and in many cases, to be highly aneuploid (Gordon *et al.,* 2012). Furthermore, many studies have now demonstrated that within tumours, genetic diversity is pervasive (Gay *et al.,* 2016). Diversity is the *fuel* on which Darwinian selection acts, hence understanding how this diversity originates and develops is a key question. Furthermore, genetic diversity been shown to have prognostic value in some cancer types (Andor *et al.,* 2016; Martinez *et al.,* 2016). This intra-tumour heterogeneity is evident at all genomic length scales, from single nucleotide alterations through to whole chromosome losses and gains. Moreover, as technologies have improved, genomic resolution has increased and the degree of diversity uncovered continues to accelerate to the point that every cell in a tumour is likely to be genetically distinct from all others, as has been demonstrated in single cell sequencing studies (Wang *et al.,* 2014b).

Cancers grow via bifurcating cell division where DNA in the cell is copied and then passed on to daughter cells. While this process is relatively robust, it is far from perfect and there remains some non-negligible probability that an error is made. Due to somatic cells reproducing asexually, any errors will then be passed onto the daughter cells and all subsequent descendants. Given a conservative estimate of the base pair mutation rate of 10^{-9} and the size of the genome being $3 * 10^9$ it is likely that every cell division will introduce new mutations. Furthermore, tumours are often subject to genomic instability, either through increased point mutation rates, due to defective DNA repair processes (Campbell *et al.,* 2017) or chromosomal instability, which can result in heterogeneous copy number states across the tumour. Billions of cell divisions coupled with imperfect DNA copying makes intra-tumour heterogeneity inevitable.

1.2 Components of the evolutionary system

Like every evolutionary system, clonal evolution in cancer is shaped by the fundamental evolutionary forces: (deterministic) selection, (stochastic) mutation and (stochastic) genetic drift. Mutation and drift are by nature stochastic processes as they depend on the chance acquisition of a mutation or in the case of drift, ran-

dom birth and death events. Selection, meanwhile, is deterministic. Once a lineage overcomes the genetic drift barrier (given when the strength of selection is of the order of the inverse population size $\sim 1/N$), the expansion of the lineage becomes predictable (Ewens, 2012).

The growing field of cancer evolution interrogates the relative and combined contributions of these evolutionary components. Large sequencing studies such as TCGA have uncovered many recurrent so-called driver mutations across cancer types (mutations that lead to a positively selected phenotype, and so expansion of the clone of cells carrying the driver mutation). These types of analysis particularly highlight the importance clonal selection in cancer development . The mutation rate itself has also received considerable attention. The mutation burden varies considerably across cancers, suggesting large differences in the underlying mutation rate between individual tumours and tumour types (Lawrence *et al.,* 2013). The realisation that different mutational processes (a combined term for the interrelated processes of mutagenesis and defective DNA repair), such as damage from UV light or defective mismatch repair, each leave distinctive (e.g. non-random) patterns of mutation across the genome has been instructive in mapping genetic mutations to underlying biological process (Alexandrov *et al.,* 2013). On the other hand, the role of stochastic drift in shaping tumour evolution has largely been neglected.

The field of population genetics provides a quantitative framework with which to study evolution (Lynch, 2007), and has proved useful in organismal evolution. It is perhaps the area of biology that has made the most use of mathematical theories (Queller, 2017), and for good reason given the difficulty in conducting experiments over evolutionary timescale. Population genetics can also be applied and is well suited to applications in the study of cancer evolution (Hu *et al.,* 2017). Cancer evolution suffers from the same problem as organismal evolution in that the evolutionary process cannot be observed across time. It can however be inferred from genetic diversity at the time of observation. Population genetics provides a principled way to approach this problem. The cancer genomics field has now accumulated a wealth of data (which continues apace) with which to apply these approaches. Fur-

thermore, recent technological advances in measuring biological parameters and the development of sophisticated experimental systems with which to track evolution provides exciting opportunities to produce quantitative measurements of cancer evolution across space and time.

1.2.1 Classifying tumour evolution

Many studies attempt to classify the evolution of cancers into distinct modes. Frequently discussed modes include neutral, punctuated, branching and linear (Davis *et al.,* 2017). The conventional view of cancer evolution is that it proceeds in a linear fashion where successively fitter mutant arise and sweep to fixation, replacing less fit lineages. Neutral evolution on the other hand is a description of what happens in the absence of selection where all cell lineages have equal fitness and grow at the same rate (I'll return to this in Chapter 3). The term effectively neutral is also used, in this paradigm subclonal variants do not contribute substantially to the clonal architecture even if they are under selection (Sun *et al.,* 2017). Branched evolution describes the scenario where multiple subclones with selective growth advantages co-occur within the tumour (Gerlinger *et al.,* 2014). A punctuated event can be thought of as a catastrophic event which induces a radical change in phenotype followed by strong selection for that phenotype (Baca *et al.,* 2013) Possible examples of punctuated events in cancer progression included chromothripsis (Korbel & Campbell, 2013) where many rearrangements of the genome occur simultaneously and kateagis, localised hypermutation resulting in many single base pair changes (Nik-Zainal *et al.,* 2012a).

Classifying cancers in this way is an illusion in many respects. A single cancer may go through distinct periods of evolution that may be described by all these modes. How the evolutionary process appears a single time point will depend on how and when the tumour is sampled as well as the resolution of the assay that is used. For example, if a tumour is sampled right after a clone has swept, then the evolution would appear ?linear?, but sampling at a time just before the fixation event would appear ?branched? (Figure 1.1). Furthermore, how the samples are taken in space could also lead to the appearance of linear (if only the swept clone

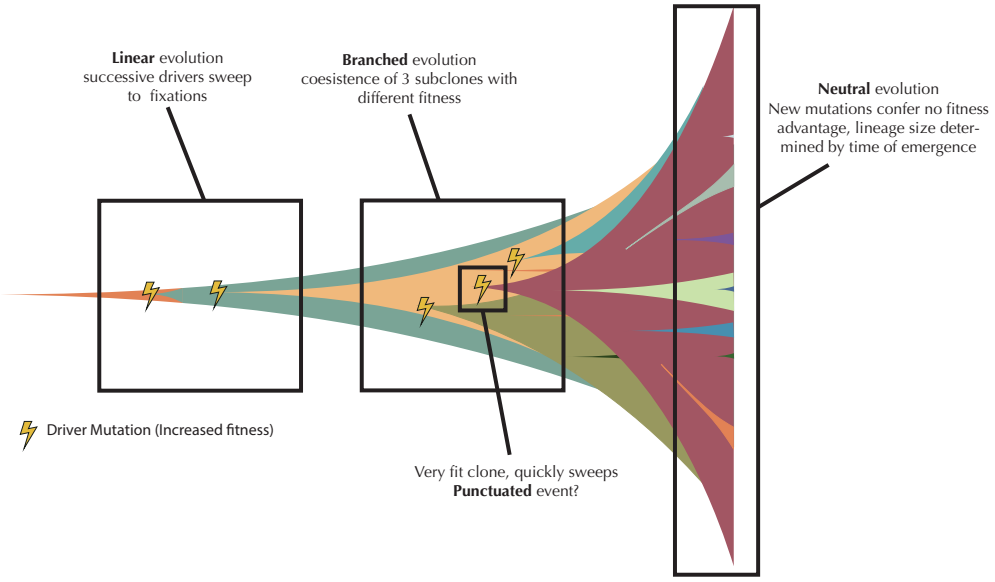

Figure 1.1: *This schematic shows how it's possible that many modes of evolution may occur during the expansion of a single tumour*

is sampled), branched (if the sweeping clone and residual tumour population is sampled) or neutral evolution (if only the clone, or residual population, is sampled) (Figure 1.1). Spatially biased sampling and limited genetic resolution can also mean some clones are missed and others overrepresented in the samples.

A more informative approach is perhaps to think about the contribution of the different evolutionary forces and how these may impact the future evolutionary trajectory. Specifically, this would mean measuring the mutation rate (separately for different types of mutations), the distribution of fitness effects of these mutations (potentially also taking into the account the current microenvironment) and elucidating the relative importance of stochastic effects (e.g. the prevalence of drift). Here we can think of the evolutionary forces and their relative contribution as inputs into the cancer ecosystem, the output is what we ultimately observe in the clinic or laboratory - a tumour, see Figure 1.2. Understanding the relative contribution of these inputs and how they relate to the output is in most cases unknown. For example, we might like to know what are the different inputs that appear to be causative for metastasis, aggressive tumours or resistance to treatment.

This introductory chapter will discuss how genomics can be used to infer and

quantify evolution in cancer. In particular I will focus on the concept of lineage tracing and how this together with quantitive models provides a powerful way to study and measure evolutionary processes. I will also briefly mention other ways genomics can be used to infer evolution in cancer, such as with the particular properties of mutations. First of all however, I will discuss in brief the different evolutionary forces and their peculiarities in cancer, and what measurements may be useful in the goal laid out in this thesis of quantifying evolutionary forces in tumours.

1.2.2 Selection

Selection from a population genetics perspective is the increase in frequency of a particular genotype due to increased fitness. Classically this is defined as more offspring per capita per generation. In tumours, fitness can be intuitively understood as the net growth rate of lineages relative to other lineages. Despite this simple framing, from a mechanistic point of view, selection can come in many different flavours and in cancer is likely to be variable in time and space; a genotype that is selected for in one environment, say in a pre-malignant lesion may not have a fitness advantage for metastasizing cells. Treatment radically changes the selective pressures in the tumour micro-environment. This means the fitness landscape is radically altered, this will often lead to the emergence of resistant sub-populations whose genotype may have conveyed no differential fitness pre-treatment. The eco-

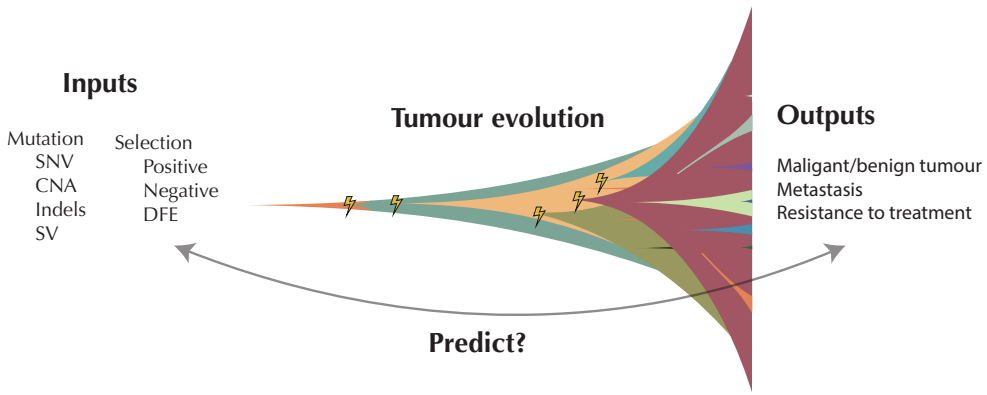

Figure 1.2: *The evolutionary forces can be thought of inputs into the progression of tumours. We ultimately observe different kinds of tumours that may be malignant, metastasise and be resistant to treatment. If we can characterise the inputs (evolutionary forces) we may be able to predict the outputs.*

logical context within which mutations arise and how it changes over time is thus likely to be important (Scott & Marusyk, 2017). That is the fitness landscape in cancers is almost certainly a dynamic entity.

Positive selection, when subclones within a tumour grow more rapidly than others is perhaps the dominant mode of selection during tumour initiation. Many of the so called hallmarks of cancer result in increased proliferation or the ability of cells to evade the homeostatic regulation of physiologically normal tissues (Hanahan & Weinberg, 2011). Sequencing of large cohorts has revealed recurrent mutations in certain genes suggesting large fitness effects. Other *driver* mutations are rarer, perhaps due to weaker selection and cancer requiring many small effect drivers in the absence of strong drivers. An alternative mode of selection that may be important in cancer is negative or purifying selection whereby subclones have a negative fitness and are more likely to be lost from the population. This may be particularly important in immunotherapies for example, where one hypothesis for their effectiveness in some cancer types is that the increased mutational load in some cancers result in increased presentation of neoantigens on the cell surface which can be detected by the immune system (McGranahan *et al.,* 2016). Thus cells with high numbers of neoantigens may be negatively selected.

Another curious property of selection in cancer is that because tumours are growing populations (at least for a significant proportion of their life history), the effects of selection are reduced (Korolev *et al.,* 2012). Thus, subclones that have fitness advantages may never reach an appreciable frequency to effect the evolutionary trajectory of tumours. In this case the dominant clone is what is important for defining the biology of the tumour as selection may not be strong enough (given the short timescales) to allow a new subclone to replace the dominant one. Related to this point a pertinent question is firstly what types and strength of selection are we able to observe? And how can we measure these effects. Some further interesting questions related to selection in cancer are what is the distribution of fitness effects of driver mutations? How frequent are punctuated events? How does ecological context change selection pressure and to what degree? Is negative selection strong

enough to mould the cancer genome? The answer to these questions have thus far remained elusive.

1.2.3 Mutation

Mutation too comes in different flavours. The typical cancer genome is modified in radically different ways. The most straightforward to identify are point mutations, single base pair changes that can alter the protein coding region and render it non-functional (eg tumour suppressive mutations) or alter its function (eg oncogenic mutations). Slightly bigger changes such as deletions and insertions (collectively called indels) can induce similar effects. Larger structural variation across the genome are also common across cancer, including whole genome doubling, chromosomal loss or gains and translocations (Beroukhim *et al.,* 2010). Relatively little is known about the rates of these distinct processes. It appears likely that the point mutation rate is elevated in somatic tissue compared to germline (Seshadri *et al.,* 1987), but the degree with which it is elevated is unknown. Many recent studies have shown that distinct processes can contribute different types of base pair changes. For example the process of ageing has a distinct *mutational signature* characterised by an abundance of $C > T$ mutations (Alexandrov *et al.,* 2015).

Meanwhile it is unknown whether chromosomal genomic instability is continuous throughout some tumours lifetime or comes in bursts. Recent analyses suggest that genome doubling is a common feature of cancer genome evolution ($\sim$ 30% of cases) and is a driver of copy number instability (Bielski *et al.,* 2018). Furthermore some tumours are found to be hypermutated (Campbell *et al.,* 2017) due to inactivating mutations in DNA repair pathways. Measuring the mutation rates of the different types of mutations (passengers, drivers, structural variation, copy number alterations) and how the various kinds of genomic instability modify the baseline rate is a current gap in our understanding. There is also some speculation that the average fitness of cancer cells may decrease over time due to the constant accrual of slightly deleterious mutations (McFarland *et al.,* 2013).

1.2.4 Neutral drift

Another important aspect is the evolutionary dynamics in the absence of selection, ie neutral drift processes. Understanding these processes is useful as it enables quantifying the degree of diversity we would expect to see in a tumour if all cells have the same fitness. Furthermore neutrality provides the natural null model for molecular evolution (Wu *et al.,* 2016). This is important as it enables distinguishing selection from any variation that would be expected when there is no selection.

The peculiarities of cancer growth complicate how drift is manifested in cancer. Given that tumours are growing populations modifications to classical models of drift may be needed and are being developed (Chen *et al.,* 2017b). Furthermore, the expected frequencies of new mutations entering the population is also affected by the growth of tumours. For example mutations that appear early during tumour growth will be at a higher frequency that those that appear late when the population is large. High death rates can also lead to greater variability (Bozic *et al.,* 2016), but the death rate of cells particularly after transformation when the tumour is small (where drift effects are likely to dominate) is unknown. Further non-Darwinian variability may arise from spatial phenomenon such as gene-surfing where mutations acquired on the expanding front of a population rise in frequency as has been observed in bacteria (Fusco *et al.,* 2016; Kostadinov *et al.,* 2016).

1.3 Measuring cancer evolution

Despite the many unknowns and challenges raised thus far, quantifying these evolutionary forces is becoming a more realistic goal. Recent advances in high throughput assays enable precise measurements of biological parameters to be made. While sophisticated experimental systems are also being developed to interrogate evolutionary forces in cancer. Taken together this provides exciting opportunities to produce quantitative measurements of cancer evolution in both model systems and *in vivo* across space and time (McPherson *et al.,* 2017). Some of these approaches will form the bulk of the rest of this introductory chapter and motivate the approaches taken in the remaining chapters of this thesis.

Attempting to unravel the complexity of the genomic landscape of cancer that has been revealed in recent years may at first appear rather daunting. Intra tumour heterogeneity by its very nature does provide opportunities to study cancer evolution. Due to the continual acquisition of genetic abnormalities and the unavoidable intra-tumour genetic heterogeneity as the tumour progresses, the cancer genome, or more accurately the differences between genomes of single cells in the cancer hold a record of its evolutionary history. Each new mutation acquired by a single cell in a cancer will be passed on to its daughter cells. These mutations thus record ancestral relationships between cells, and can be thought of as scars which label different sub-populations. In summary the genome of every cancer cell is an imperfect copy of another cancer cell that existed some time in the past. This simple observation, that heterogeneity emerges from cell divisions coupled with mutations therefore allows for inferring the past history of a cancer, or indeed any somatic tissue. This circumvents one of the biggest issues facing the study of cancer as an evolutionary system; the inability to follow cancers *in vivo* unperturbed over time, due to clinical necessity and ethical issues.

Broadly this type of approach - using a label of some kind, in this case mutations to track populations of cell - can be called lineage tracing and comes in many different flavours. As will be shown in this thesis, lineage tracing coupled with theoretical models provides a powerful method with which to study evolution and population dynamics, and is the basis of the methodological approach taken. The concept of lineage tracing also has been used extensively in model systems, particularly in developmental biology and stem cell biology to track the progeny of particular cell types of interest (Blanpain & Simons, 2013).

1.3.1 Lineage tracing

Lineage tracing is a powerful method to study evolution in populations of cells. Defining a lineage as a group of cells who all share some common ancestor, the idea is to follow or trace these lineages over space and/or time. To apply this principle, a label of some kind needs to be induced or acquired in a single cell which is then passed onto all of its progeny, and thus label that lineage. In experimen-

tal systems, these labels can be introduced artificially, using fluorescent markers or DNA barcodes for example. Labelling of cell population with fluorescent reports has been used in mouse models of tumour growth for example, identifying stem cell populations in squamous skin tumours (Driessens *et al.,* 2012) and the clonal dynamics required for the formation of skin tumours (Sánchez-Danés *et al.,* 2016). Lineage tracing principles can also be applied *in vivo* in human tissues by using mutations as naturally occurring labels. The labelling of different populations within a tumour with mutations lie at the heart of phylogenetic principles applied to cancer for example.

1.3.2 Naturally occurring lineage markers

An early example of using lineage tracing in the context of cancer was to study whether tumours were of single cell origin or not. The first genetic defect that was found to be associated with cancer progression is the so-called Philadelphia chromosome. Patients suffering from chronic myeloid leukaemia were found to have an abnormally large chromosome 21. Later studies showed that this was due to the translocation of genetic material between chromosome 21 and chromosome 9, resulting in a fusion gene called BCR-ABL1. This fusion gene results in the uncontrolled proliferation of myeloid cells. This genetic defect was found to be pervasive throughout all malignant cells in the tumour. Furthermore it was found that it was always the same allele that was elongated, providing strong evidence for the single cell of origin hypothesis. Other early studies used X-chromosome inactivation as a clonal lineage marker. In females, only one of the two X-chromosome are genetically active in somatic cells, the choice of which X-chromosome is active is random but once decided upon early during embryogenesis is fixed and passed on to all progeny. Female somatic tissue is therefore a mosaic of cells with either the maternal or paternal X-chromosome being genetically active. A corollary of the single cell of origin hypothesis would be that in tumours all cells would have the same X-chromosome active. Across tumours from multiple different tissues sites this was indeed found to be the case (Fialkow, 1979; Fialkow, 1974).

These early studies ultimately led the foundations of what become the clonal

evolution theory of cancer, detailed in the seminal paper by Peter Nowell in 1976 (who was the first person to observe the Philadelphia chromosome) (Nowell, 1976). These approaches could not however, probe the evolutionary dynamics within the tumour, as at the resolution of these approaches, all cells look the same. Molecular biology techniques developed later, do however enable researchers to distinguish different lineages via divergent genomic events within tumours.

1.3.3 Multi-region sequencing

These early studies looked for conserved genetic markers across space as evidence of single cell origin, differences in the genetics across space also hold valuable information. As I have already discussed, due to the imperfect DNA copying machinery in cells, tumour cells inevitably harbour distinct mutations that make them genetically different from other tumour cells. A relatively simple experiment is then to measure these differences. One pioneering study from Shibata and colleagues did exactly this by looking at non-coding microsatellite loci in spatially distinct regions within microsatellite unstable colorectal tumours. Such tumours are deficient in their mismatch repair (MMR) machinery which results in a hypermutator phenotype. Using the divergence (genetic distance between samples) as a summary statistic between these spatially distinct regions and theoretical models of cancer growth they were able to estimate tumour ages in terms of cell divisions and the time at which the cancers diverged from their pre-curser lesions (adenomas) (Tsao *et al.,* 1999; Tsao *et al.,* 2000). To conduct their inference Tsao *et al.* developed a computational model of tumour growth with micro-satellite instability and then could compare the simulations with the observed data to infer the ages of adenomas vs cancers. Perhaps surprisingly, the adenomas and carcinomas were found to be of similar ages. These studies were perhaps the first so called multi-region sequencing study.

Another study taking a similar approach demonstrated that colorectal adenomas were often of polyclonal origin (Thirlwell *et al.,* 2010). This study revealed multiple subclones within the same adenoma harboured different mutations in the tumour suppressor *APC*. Further phylogenetic analysis of a limited number of mark-

ers implicated in the progression of colon cancers (including mutations in *KRAS* and *TP53* and loss of heterozygosity on chromosomes 5, 11 and 18) also showed the coexistence of different subclones within the same adenoma.

Since these early studies, multi-region sequencing studies have become common place, particularly in the current era of next generation sequencing technologies. These technologies allow simultaneously identifying mutations across the whole genome of the tumour with whole genome sequencing, or selectively targeting a region of the genome such as with whole exome sequencing. Multi-region sequencing consists of taking "bulk" samples containing millions of cells and sequencing the aggregate genome of this bulk sample. Similarities and differences in the mutations in these bulk samples can then be used for evolutionary inferences. Studies using multi-region sequencing often use phylogenetics to interrogate the evolutionary relationships between samples. In recent years many phylogenetic methods have been developed to deal with the peculiarities of cancer evolution such as the different types of mutations (SNVs vs CNAs), hypermutability and high heterogeneity (Schwartz & Schäffer, 2017).

One prominent study which used this approach was Gerlinger *et al.,* 2012, here they used whole exome sequencing of different tumour regions to profile clear cell renal carcinomas, finding a large degree of intra-tumour heterogeneity. Later studies from the same group using a approach showed evidence of convergent evolution with distinct putative driver mutations in *SETD2* found on different branches of the phylogenetic tree (Gerlinger *et al.,* 2014). Further multi region sequencing studies have shown intra-tumour heterogeneity is pervasive across cancer types including lung (de Bruin *et al.,* 2014; Zhang *et al.,* 2014), breast (Yates *et al.,* 2015), lymphoma (Okosun *et al.,* 2014), brain (Sottoriva *et al.,* 2013a) and colon (Sottoriva *et al.,* 2015) amongst others.

Multi-region sequencing has proved useful in elucidating the timing that mutations are acquired. For example driver mutations are often found to be truncal on the phylogenetic tree, that is found ubiquitously across all sampled regions. This suggests that many of the important driver events are acquired early relative to the

time patients present with symptoms of their disease. This is particularly true in some cancer types such as colon and lung (Sottoriva *et al.,* 2015; Zhang *et al.,* 2014), while kidney cancers for example often appear to have subclonal driver mutations (Gerlinger *et al.,* 2014). Multiregion sequencing together with phylogenetic analysis has also been useful in determining how evolution is influenced by environmental factors (de Bruin *et al.,* 2014) and elucidating the seeding patterns of metastasis (El-Kebir *et al.,* 2018; Mcpherson *et al.,* 2016). Studies such as the TRACERx clinical trial are currently underway to determine the effects of intra tumour heterogeneity on patient prognosis using multi region sequencing assays in multiple cancer types (Jamal-Hanjani *et al.,* 2017).

In general however, the evolutionary dynamics that produce most of the observed ITH remain uncharacterised. In particular, how to accurately construct the phylogenetic relationships between tumours sites and how best to interpret such data remain uncertain (Schwartz & Schäffer, 2017). Issues include sampling bias, where samples may not be taken uniformly across the tumour mass and may be confined to sub regions. This can give the illusion of longer or shorter branch lengths. Another issue is that the typical limited sampling (4-5 samples per tumour) can result in misclassifying truncal mutations (Werner *et al.,* 2017). Also bulk tumour samples potentially consist of multiple subclones and therefore ideally the phylogenetic relationships should be constructed based on the deconvolved clonal structure (Alves *et al.,* 2017). Deconvolving bulk tumour samples into its respective subclones remains technically challenging however (Sun *et al.,* 2017). Typically these methods integrate copy number changes and mutation frequencies to calculate the cellular prevalence (or cancer cell fraction - CCF) of mutations within the tumour and then cluster mutations with similar cellular prevalence into distinct groups. The logic is that these are subclones within the tumour, however this may not always be the case. Correctly inferring CCF before clustering is also challenging, in particular inferring the relative timing of a point mutation vs a copy number gain or loss is prone to error which will result in incorrect calculation of the CCF. Some kind of clustering methodology is typically used to group mutations into clusters, a

popular method being Bayesian Dirichilet Process clustering (Dunson, 2009). The accuracy of these methods particularly with low depth sequencing is uncertain as robust benchmarking on data where the ground truth is known is lacking.

1.3.4 Single cell sequencing

Recent advances in single cell sequencing resolve some of the issues of multi-region sequencing type studies, but introduce others. Single cell sequencing potentially provides unparalleled resolution of tumour genetic diversity, identifying mutations present in individual cells that would most likely be missed by conventional bulk sequencing. Single cell sequencing thus provides opportunities for fine grained analysis of cancer evolution. It also has the benefit that resolving the clonal structure is not required as each cell is a pure sample by its very nature. This type of approach does however come with its own set of problems. In particular the degree of technical noise is higher than with other sequencing approaches and issues of sampling bias remain given that perhaps only a hundred cells of a tumour comprised of billions are typically sampled.

Technical issues in single cell sequencing technology arise from the naturally low quantity of DNA extracted from single cells meaning whole genome amplification is generally required to generate sufficient DNA for sequencing. This additional step introduces technical artefacts such as non-uniform coverage and allele dropout (Davis & Navin, 2016). SNVs are also difficult to accurately detect due to the high technical error rates. This means true positives are often difficult to distinguish from sequencing errors (which is of the order 1%) (Roth *et al.,* 2016). For this reason copy number profiling is generally preferred. This however is also not without issues due to the aforementioned technical issues with whole genome amplification. Sophisticated single cell specific algorithms have been developed for analysing these data. Recent technical advances showed that single cell sequencing without whole genome amplification is possible (Zahn *et al.,* 2017) and that by pooling single cells together a "virtual" bulk could be generated which makes single nucleotide variant calling more robust (Salehi *et al.,* 2017). Further advances in this area are likely to provide exquisite fine grained data with which to conduct

evolutionary analyses.

The small number of large-scale studies using single cell sequencing have already revealed interesting aspects of the evolutionary process. Gao *et al.,* 2016 used single cell sequencing to look at aneuploidy in triple negative breast cancer, this technology allowed them to look at copy number alterations at very fine grained resolution. Interestingly, copy number alterations appeared spatially and therefore temporally stable, suggesting that large scale copy number changes are perhaps rare events during tumour evolution.

Another interesting study that employed single cell sequencing investigated the temporal dynamics of cancer evolution using a patient derived xenograft model (Eirew *et al.,* 2015). Interestingly, they found that minor clones ($<5\%$) often come to dominate the tumour population. Suggesting that some clones acquire large fitness advantages, likely necessary to induce such large expansions.

1.3.5 Deep sequencing

An orthogonal approach to using sequencing from spatially distinct regions is to leverage the information from deep sequencing of bulk tumour samples. This approach can be used to simultaneously measure mutations that are at different cellular proportions within the tumour. Given that bulk samples are composed of many millions of cells, deep sequencing effectively sequences an aggregate genome of these cells. While a particular mutation cannot be assigned to a particular cell in the tumour or indeed the co-occurrence of mutations within the same lineage is unknown, this approach does measure the frequency of cells with a particular mutation. In other words, deep sequencing can measure the size of lineages within the population. While copy number aberrations and low tumour purity can confound these measurements methods exist to correct for these issues. Cancer cell fraction (CCF) transformations are often used for this purpose. Alternatively using mutations in diploid regions of the tumour provides a straightforward mapping between lineage size and mutation frequency. Lineage size is the crucial piece of information that is revealed by deep sequencing studies that enables the population dynamics to be inferred. Furthermore in asexual evolution such as cancer cell populations where

there is no recombination, mutations hitchhike and a set of mutations at a particularly frequency potentially allow identification of sub populations of cells (Fay & Wu, 2000; Gillespie, 2000; Nik-Zainal *et al.,* 2012b).

A useful way to summarise the information from a deep sequencing experiment (of a bulk sample) is by plotting a histogram of the mutation frequencies (or lineage sizes). This is commonly referred to as the variant allele frequency (VAF) distribution. In population genetics this distribution is known as the "site frequency spectrum", and there is a considerable body of work devoted to exploiting it to measure evolutionary dynamics (Ronen *et al.,* 2013; Keinan & Clark, 2012). This type of data is readily available, in particular due to large scale cancer sequencing projects such as TCGA (the Cancer Genome Atlas) and ICGC (International Cancer Genome Consortium) many thousands of tumours have been deep sequenced. For this reason this type of data will be the primary focus of this thesis, a more thorough discussion of this type of data and the methods associated with its analysis are presented in Chapter 2.

1.3.6 Lineage tracing in stem cell biology

Lineage tracing has also been used in experimental systems to measure evolution and population dynamics. The field of stem cell biology in particular has embraced the use of experimental lineage tracing as it provides a robust way to interrogate the stemness of different cell types. A common definition of a stem cell is a cell that is long lived and that can give rise to multiple cell types. Following the progeny of single (suspected) stem cell allows for identifying the different cell types that can be derived from this labelled cell. Thus determining the potency of particular cells as well as their potential for self-renewal. This type of approach is now deemed the gold standard in stem cell identification (Wright, 2012). Typically, these experimental systems use cre-recombinase to induce fluorescent reporters that can be followed over time. By targeting different cell-type specific promoters, reporters in different sub-populations of cells can be induced and the stem-like capabilities of these cells investigated (Blanpain & Simons, 2013). Not only do these approaches allow for identification of stem cells, together with statistical modelling they can

be used to investigate the population dynamics of stem cell self renewal. This has uncovered that neutral competition of stem cells through stochastic loss and replacement as a universal phenomenon across tissues (Klein & Simons, 2011). This has been demonstrated through the observation that the distribution of clone sizes exhibits a property called "scaling" where the shape of the distribution is preserved over time. Where the "scale" is defined by the average clone size. This scaling property only arises in populations of equipotent stem cells undergoing stochastic loss replacement and not alternate models such as populations of slow cycling cells undergoing asymmetric division. That is, asymmetry of cell fate is maintained at a population level rather than an individual cell level.

An archetypal system with which to study stem cells is the colonic crypt. The colon is comprised of small finger like protrusions into the epithelia where cells undergo constant turnover, so much so that the entire epithelia is replaced over the course of a week (Vermeulen & Snippert, 2014). Numerous studies in mice using fluorescent based lineage tracing have demonstrated that stem cells reside in the bottom of the crypt and undergo stochastic loss and replacement with their neighbours resulting in a neutral drift process (Ritsma *et al.,* 2014; Lopez-Garcia *et al.,* 2010). This has also been demonstrated in humans using mitochondrial DNA as a lineage marker (Baker *et al.,* 2014). It has also been shown that genetic alterations can disrupt the stem cell dynamics. Using genetically engineered mice, Vermeulen *et al.,* 2013 quantified the selective advantage of mutations introduced into stem cells and found mutant *KRAS* and *APC* stem cells had between a 2-4 increased probability of fixation in mouse intestinal crypts over what would be expected from a neutral process. This process is thought to be the first step in the progression of cancer in the colon. With this in mind, in the final chapter of this thesis I use sequencing of individual crypts in an attempt to quantify stem cell dynamics in humans using naturally occurring single nucleotide changes as labels to track the dynamics.

1.3.7 High throughput experimental lineage tracing

A major issue with experimental lineage tracing that rely on fluorescent reporters for studying these processes is that they suffer from a lack of resolution, a limited number of lineages can be followed at any one time. To circumvent such issues, high throughput lineage tracing protocols have been developed via the use of multiplexed genomic barcodes. These barcodes can be inserted via viral transfection into the genomes of single cells and provide a unique tag for each cell. Many millions of clones can be traced simultaneously with this approach via sequencing pools of barcoded cells (Bhang *et al.,* 2015). Barcode libraries are constructed such that each transfected cell caries a unique label which can be used to measure its size. Just as normal somatic mutations and deep sequencing of tumour cells measures the size of lineages within a tumour, deep sequencing of a pool of barcoded cells will also measure the lineage size. However as only a few base pairs of the genome is needed to be sequenced (the barcodes), high depth coverage of the barcode is employed which results in very high resolution tracking of individual lineages.

Levy *et al.,* 2015 used this approach to measure lineage sizes in serial passages of yeast cells. Using this data, together with theoretical population genetics they were able to quantify the time fitter lineages emerged and the distribution of fitness effects. This was done by identifying lineages that increased in size faster than could be expected from stochastic neutral drift. Similar experimental strategies have recently been applied to cancer model systems. For example the fitness of 11 tumour suppressor pathways was measured using CRISPR-Cas9 genome editing to introduce mutations followed by barcoding to measure tumour size in mouse models of human cancer (Rogers *et al.,* 2017). Using this approach, Rogers *et al.* found that mutations in *SETD2* and *LKB1* had the largest fitness effect and resulted in the largest tumours. More complex experimental strategies are likely to provide further insight, for example barcoding potentially allows for tracking the size of competing lineages over time, which could be used to measure the relative fitness of subclones. Quantitative theoretical approaches are also likely to be key to leveraging the power of these experiments, just as the VAF distribution reported by traditional sequenc-

ing approaches can be used to infer the population level dynamics, the lineage size distribution from barcoding experiments can be similarly insightful. Indeed this approach was recently used by Lan *et al.,* 2017 in barcoded glioblastoma models in mice. They showed that intra-tumour heterogeneity in glioblastoma could be explained by stochastic fate of cells in a stem cell hierarchy, while treatment resistant clones could be identified via deviations from this model.

1.4 Mutational identity and measuring evolution

The previous section discussed how the concept of lineage tracing can be used to trace and map evolution through space and time, for these purposes mutations are merely labels which track lineages. Mutations themselves can also elucidate aspects of evolutionary processes in cancer.

1.4.1 Driver mutations

This type of approach is perhaps best exemplified by large cancer sequencing studies such as the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). The focus with these projects was that recurrent so called driver mutations could be identified by sequencing large cohorts of cancers. Driver mutations are those mutations which are thought to drive the progression of the disease. There is great interest in identifying these alterations as they provide targets for the development of therapies, and for stratifying patients based on their mutational profile. From an evolutionary perspective mutations that are frequently mutated can be seen as evidence of convergent evolution. Such mutations may have therefore have strong fitness effects on cells. Unfortunately these studies have demonstrated that there are very few highly recurrent mutations and a long tail of rare driver mutations. For example only a handful of mutations are found at appreciable frequencies across all cancer types, Kandoth *et al.,* 2013 for example only found mutations in *TP53* and *PIK3CA* to be above 10% across cancer types. By looking at the occurrence of mutations across specific cancer types the results are improved, for example mutations in *APC* are found in 80% if colorectal cancers and VHL in 50% of renal carcinomas. Overall however these results demonstrated

a large degree of heterogeneity in cancer drivers. Furthermore little is known about epistatic interactions and how multiple driver mutations affect the fitness of cells.

Classifying mutations as drivers is also problematic and poses many challenges. For example merely looking at the frequency of hits across large cohorts neglects important cofounders, such as the size of genes. The genomic context of certain genes is also important, chromatin structure for example and in particular regions of open chromatin have been shown to have higher mutation rates (Schuster-Böckler & Lehner, 2012). Genes with high expression also tend to have more mutations. Methods such as MutSigCV have therefore been developed to correct for the variability of mutation rates across the genome and across tissues (Lawrence *et al.,* 2013). Such methods apply corrections based on the whole cohort and so variability at a patient level may be masked.

Genomic changes other than point mutations also undoubtedly drive the progression of the disease. Indeed, the first driver mutation to be identified was the *BCR-ABL* translocation in chronic myeloid leukaemia (Sandberg & Hossfeld, 1970). In general however, identifying structural variation that drives the disease is even more technically challenging because many cancers display genomic instability. Distinguishing changes that may modify fitness from mutations arising from genomic instability is challenging. One class of structural variation that has sound biological significance is loss of heterozygosity (LOH), coupled with point mutations in tumour suppressors. For example LOH in chromosome 5 is commonly observed in colon cancer coupled with a mutation in the *APC* gene. *APC* is a tumour suppressor gene commonly found inactivated in colon cancer. One common route to inactivation appears to be a mutation in one allele followed by LOH which removes the remaining functional allele (Pino & Chung, 2010). Genome doubling is another structural change commonly observed, it is thought that genome doubling enables cells to tolerate genomic instability rather than providing a differential fitness effect in of itself.

1.4.2 Mutational signatures

Another aspect that has also become prominent in recent years is the context in which mutations arise, both their environmental context and genomic context. In a seminal study Alexandrov *et al.,* 2013 found that mutations occurring in specific mutational contexts could be assigned to different signatures. Some of these signatures were found across cancer types and are likely age related where as others were found to be cancer specific. Examples include tobacco specific mutational signatures found in lung cancers (Alexandrov *et al.,* 2016) or mutations due to defects in DNA repair pathways. For example $C > A$ mutations particularly when flanked by a C and A base (ie $CCA > CAA$) is enriched in tobacco smoker. Other signatures exhibit clock like properties (Alexandrov *et al.,* 2015) making them ideal for the use in evolutionary inferences. Many signatures remain of unknown biological origin however, suggesting the presence of as yet unknown mutagens.

Mutational signatures potentially provide a window into past exposures which may be useful for designing preventative strategies. There is also potential to shed light on the relative contribution of mutation vs selection. Certain mutational processes may predispose people to certain driver mutations. While other driver mutations may be less likely but have higher fitness. Indeed, a mathematical treatise of this kind of analysis showed that *BRAF* v600E a common driver in many cancer types is unlikely to occur compared to other mutations but is highly selected (Temko *et al.,* 2018).

1.4.3 dN/dS

dN/dS is an alternate method that can be used to infer selection. Originally developed for comparative genomics in species evolution, dN/dS quantifies if there are more protein changing mutations than would be expected by chance. This is achieved by looking at the ratio of normalized non-synonymous mutations to synonymous mutations, where synonymous mutations are assumed to be neutral and thus provide a baseline rate. A ratio of 1 is expected if all mutations were neutral, and ratios of <1 and >1 if there are abundance of negatively or positively

selected variants. In organismal population genetics a $dN/dS < 1$ is typically observed, meaning the dominant form of selection observed at the genomic coding level is negative or purifying selection. Interestingly, in cancer many studies report dN/dS close to 1 (Martincorena *et al.,* 2017; Wu *et al.,* 2016), suggesting the absence of negative selection and that most mutations are neutral. Robustly measuring dN/dS in cancer genomes however is challenging, mutations in cancer genomes exhibit strong context dependence exemplified by mutational signatures which needs to be corrected for as does other confounders such as chromatin accessibility and transcription factor rates which are known to influence mutations rates. Two recent studies, Martincorena *et al.,* 2017 and Weghorn & Sunyaev, 2017 attempt to account many of these confounding factors. Both studies showed strong selection for many common driver mutations. Negative selection on the other hand appeared much harder to observe. However one factor that has received little attention is how population dynamics effects evolutionary inferences gained from dN/dS, and in particular how and when dN/dS could be used in inferring selection within growing tumours. In classical dN/dS, only mutations that are fixed within lineages are used to measure selection pressures. In cancer however many mutations are subclonal and intra tumour heterogeneity is widespread which may confound these measurements. In Chapter 5 I explore some of these issues using a population genetics based theory together with dN/dS. This enables a mapping between the selection coefficient and dN/dS which has thus far been lacking in the application of dN/dS in cancer.

1.5 What is a clone?

Since the clonal evolution model of cancer has gained traction Nowell, 1976, talk of clones has become widespread, unfortunately many different meanings are used. As the concept of a clone will be used heavily throughout this thesis and to avoid any confusion I will spend a brief moment discussing how I will define a clone in the context of cancer evolution throughout.

The Oxford Dictionary provides the following definition of a clone as is used

in Biology *"An organism or cell, or group of organisms or cells, produced asexually from one ancestor or stock, to which they are genetically identical"*. While it is true that cancer cells produce asexually and are from a single ancestor, they are far from being genetically identical. As already discussed cancers are very diverse populations of cells, potentially every cell is genetically distinct (Wang *et al.,* 2014b), so using this definition is not particularly useful. Such a definition would result in a situation where there are millions of clones in a tumour. Perhaps a more useful definition, and the one that I will employ here is that a clone is a group of cells that share some phenotypic trait that make them functionally distinct from all the other cells in the tumour but can also be labelled a clone via some shared genetic alterations. As this thesis takes a population genetics perspective to the evolution in cancer, this phenotypic trait should have some effect on the population dynamics such as a higher birth or death rate or on increased mutation rate. I will not in general attempt to infer or describe the mechanisms that may alter the population dynamics of a clone, but simply talk about its effects on the overall population dynamics. For example a cell in a cancer may acquire some phenotypic trait that changes its metabolism meaning it can grow faster, from a population genetic perspective this clone can then be defined by an increased birth rate without regard to what caused the increased birth rate. This is the approach I will take. I will also commonly employ the term subclone to refer to clones that occupy a subfraction of the tumour. For example, if we are looking at a sample of cancer cells, they will all have a common ancestor some time in the past, any functionally distinct cell that initiates a clonal expansions I will then refer to as a subclone of the ancestral clone. With these definitions in mind, Chapter 3 investigates the population dynamics when there is only a single initiating clone, while Chapter 4 looks at the dynamics when subclones arise within this ancestral clone.

One final additional comment is that sometimes it will be useful to refer to a cells ancestry in terms of its relationship to other cells, particularly according to what mutations they share. Rather than use the term clone here, I will use lineage, any 2 cells that share some genetic alteration are part of the same lineage. Cells can

be part of multiple lineages but only one clone. Cancers therefore will have millions of lineages but a limited number of (sub)clones.

I will attempt to keep to the above definition throughout the thesis, however in keeping with the literature I will sometimes refer to the clone size distribution in Chapter 6.

1.6 Summary of thesis

In this introductory chapter I have discussed approaches to measure evolution and population dynamics with a particular emphasis on how genomics and theoretical models can be used for this. I will use deep sequencing of tissue samples as a read-out of lineage size and by combining this data with theoretical models inspired from population dynamics and stochastic processes will show how these measurements encode the evolutionary history cancers. The next chapter will discuss in more detail some of the methodological and technical approaches used for these purposes.

Chapter 2

Technical background and methods

The previous chapter laid the foundations and motivations of this thesis, which is principally concerned with deciphering patterns of evolution and population dynamics in human cancers and somatic tissues. This chapter discusses some of the technical background and the methods used. To begin, I will discuss how high throughput-sequencing data is generated and analysed, and the challenges involved with the analysis. I will also discuss how this data can be effectively summarised and used for evolutionary analyses. Then I will discuss mathematical and computational methods that can be used to simulate evolutionary processes, and how these have been applied to cancer. The final part of the chapter will discuss how methods from Bayesian statistics can be used to integrate data with mathematical models and together, can be used to extract mechanistic insight into the underlying evolutionary processes that drive evolution in cancer.

2.1 Bioinformatics

In this section I will discuss the methods used to identify somatic changes in cancer genomes, small changes such as single nucleotide variants (SNVs) and insertions and deletions (indels) and larger structural variants. The data used in this thesis is primarily deep sequencing of bulk tumour samples, where a piece of tumour tissue is taken from the cancer, DNA is extracted and then fragmented into short reads for next generation sequencing. By comparing sequencing data from the tumour with a matched control sample from the same patient (either blood or physiologically

normal tissue) it is possible to identify somatic mutations present exclusively in the tumour sample, Figure 2.1 shows a screenshot from the integrated genome viewer (IGV) which is used to visualise this type of data (Robinson *et al.,* 2011). Here the grey bars are sequencing reads and it can be seen that in a proportion of these reads, one base has been mutated from a C to a G when compared to the reference. The main methodological challenge in analysing this type of data is confidently identifying these changes, and distinguishing them from sequencing errors, mapping errors and germline polymorphisms.

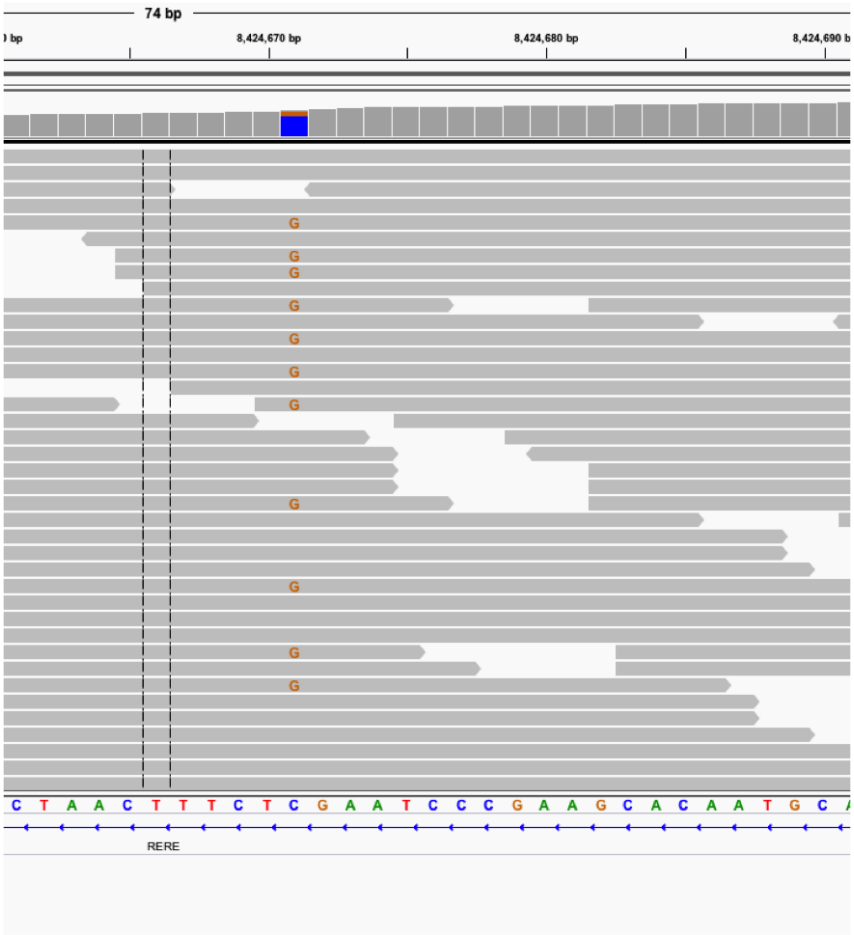

Figure 2.1: *Screenshot from the integrated genomic viewer of a suspected mutation in a lung cancer sample. In a proportion of the reads (grey bars) the reference base C has been mutated to a G.*

2.1.1 Sequencing preprocessing steps

The sequence data comes in the form of a fastq file which contains the sequences for each read with associated quality scores for each base in the read. The first step is to map these reads to the reference human genome (human genome reference 19 was used throughout unless stated otherwise) and produce a binary alignment map file (BAM), the Burrows Wheeler Aligner (BWA) was used throughout for this purpose and is one of the most popular tools for this step (Li, 2013). This produces a sequence alignment map (SAM) file that contains the sequence reads and associated genomic coordinates for each of the reads, as well as quality scores for the mapping. With this SAM file, we then need to sort the file and convert it to a binary representation, SAM files can be very large, so this compression reduces the file size, this binary compressed SAM is called a BAM file. During library preparation, it is often necessary to perform some amplification of the fragmented DNA so that there is sufficient DNA for sequencing. This PCR based amplification step can lead by chance, to the same DNA molecule being sequenced multiple times, as these duplicate reads are generated during the library preparation step and are hence technical artefacts, these duplicate reads are removed. I used the Picard MarkDuplicates tool for this purpose (Van der Auwera *et al.,* 2013). The amount of duplicate reads depends on the complexity (ie the information entropy) of the pre-amplified DNA pool, starting with small amounts of DNA often results in higher amount of duplicate reads. Following this duplicate marking the BAM file is then ready to be used for somatic mutation calling. A summary of this pipeline is shown in Figure 2.2.

2.1.2 Somatic variant calling

Having processed the sequencing data via the steps described above, the next task is identify somatic changes that are exclusive to the tumour. This is technically challenging for numerous reasons. To illustrate the challenges it is instructive to compare this to the similar problem of identifying germline polymorphisms from high throughput sequencing data. In this case, a number of assumptions can be made which makes distinguishing true mutations from sequencing noise easier.

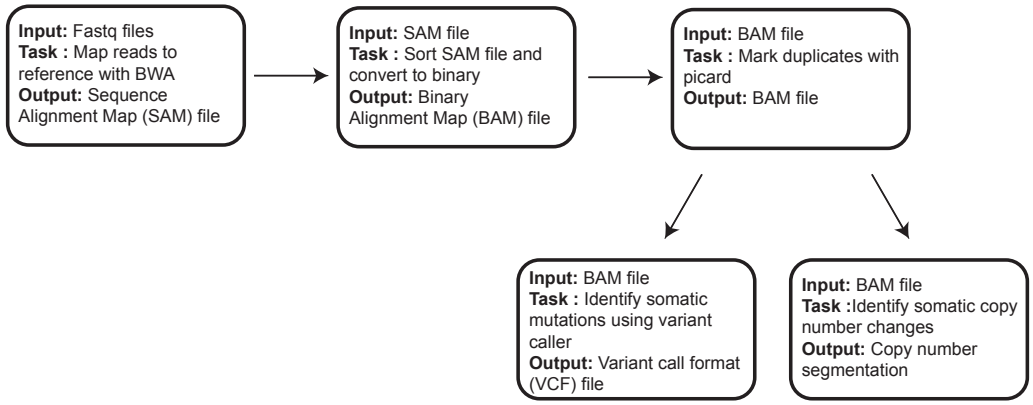

Figure 2.2: *Overview of sequencing analysis pipeline*

Due to humans having diploid chromosomes, heterozygous mutations would be expected to have variant allele frequency 50% and homozygous mutations variant allele frequency 100%. Mutations at low frequency can therefore be discounted as sequencing noise, this is not so in the discovery of somatic mutations in cancer for 3 principal reasons. i) Genetic heterogeneity in cancer means that mutations that are present in a small proportion of cells means there exist *bone fide* true mutations with low read counts, ii) Copy number alterations results in frequency of mutations being distorted and iii) tumour samples are rarely pure tumour tissue and will contain stromal tissue as well as immune cells that can distort naive expectations of the frequency of mutations.

Numerous tools have been developed to overcome these issues to confidently identify somatic mutations in cancer samples. Among the most popular are Varscan2 (Koboldt *et al.,* 2012) and Mutect (Cibulskis *et al.,* 2013). Mutect was the primary tool used in this thesis, due to it being designed specifically to identify low allelic frequency mutations with high confidence. Furthermore many independent tests have consistently shown Mutect to be among the best performing somatic variant caller in terms of sensitivity and specificity, and it can be considered the current gold standard in terms of variant calling (Wang *et al.,* 2013; Kim *et al.,* 2014; Goode *et al.,* 2013; Griffith *et al.,* 2015). Varscan 2 was used for variant calling of a whole genome gastric cancer dataset in chapter 3.This was due to the BAM files obtained from the original study being incompatible with Mutect. Mutect requires the

sequencing reads to be processed in a specific manner to comply with the Genome Analysis Toolkit best practices developed by the Broad institute, unfortunately this was not the case for this data set. Varscan 2 uses a Fisher's Exact Test to compare candidate variants in the tumour and normal control sample while Mutect uses a Bayesian classifier to identify variants. Both these tools are capable of identifying small insertions or deletions as well as single nucleotide variants.

2.1.3 Somatic copy number calling

Acquisition of single nucleotide variants and indels is only one aspect of the genetic changes observed in cancer. Another equally important observation is that cancers often have larger structural changes in their genomes. Copies of whole chromosomes are often observed to have been lost or gained (Beroukhim *et al.,* 2010), and even doubling of the whole genome is thought to be common in some cancer types (Carter *et al.,* 2012). A common approach to detect these is to use SNP arrays, where DNA molecules hybridise to probes for common human polymorphisms and induce a fluorescent signal. The strength of the fluorescence between normal samples and tumour samples can be used to detect gains or losses of genetic material, while comparing the strength of signal of polymorphisms at the same locus can be used to determine which of the alleles has been lost or gained.

Recently, studies have moved away from SNP arrays to use next generation sequencing directly. Similar principles can be applied to this type of data, where the difference in coverage between normal and tumour samples can help identify losses or gains and utilising the frequency of germline SNPs reported by NGS at different loci can help identify which of the alleles has been gained or lost. As in the case of identifying SNVs, contamination from normal tissue can make this challenging, as can the sparsity of coverage in targeted assays such as whole exome sequencing. One tool that attempts to circumvent these issues is Sequenza. Sequenza uses a Bayesian hierarchical model to simultaneously estimate allele specific copy number and tumour sample purity, from either whole exome sequencing or whole genome sequencing (Favero *et al.,* 2015). This method shows high concordance with results from SNP array assays, and was used to analyse data in all results chapters. Fig-

ure 2.3 shows the copy number profile of two gastric cancer samples showing the depth of coverage (logR ratio) and SNP frequencies (B-Allele Frequencies). The top panel shows a cancer with a highly aberrant genome with multiple copy number alterations across different chromosome, while the bottom shows a relatively stable cancer with the only observable change a gain on chromosome 8.

2.1.4 Summarising high throughput sequencing data

The above methods and tools will generate mutation calls (both point mutations and structural changes) across the whole genome or exome depending on the sequencing strategy. To make use of this data for population or evolutionary dynamic purposes requires summarising this data in some way. Fortunately this is quite straightforward as it is possible to leverage the information on the size of mutational lineages that is naturally reported in deep sequencing assays to produce a summary of the size of cell lineages in a population.

In a deep sequencing experiment each mutation found by the variant calling algorithm will have an associated variant allele frequency, VAF:

$$VAF_i = \frac{N_i}{N_T} \tag{2.1}$$

Where N_i is the number of reads with a mutation (ie a C>T) at a particular locus i and N_T is the total number of reads at that locus. If we think of mutations as labels of cell lineages, then the VAF of a mutation is related to the number of cells carrying the label in the population of cells that were sequenced, ie the lineage size. For example, in a diploid genome the lineage size is simply 2 times the VAF of a mutation. Copy number changes and low purity can however make inferring the correct lineage size from the VAF of a mutation cumbersome. In particular, if a region of the genome is amplified inferring the true lineage size from the VAF requires knowing if the mutation occurred before or after the amplification which in general will be unknown. These transformed VAFs which measure the lineage size are normally referred to in the literature as as the cancer cell fraction (CCF). Due to issues with making this transformation, for the most part I relied on using mutations that fell in

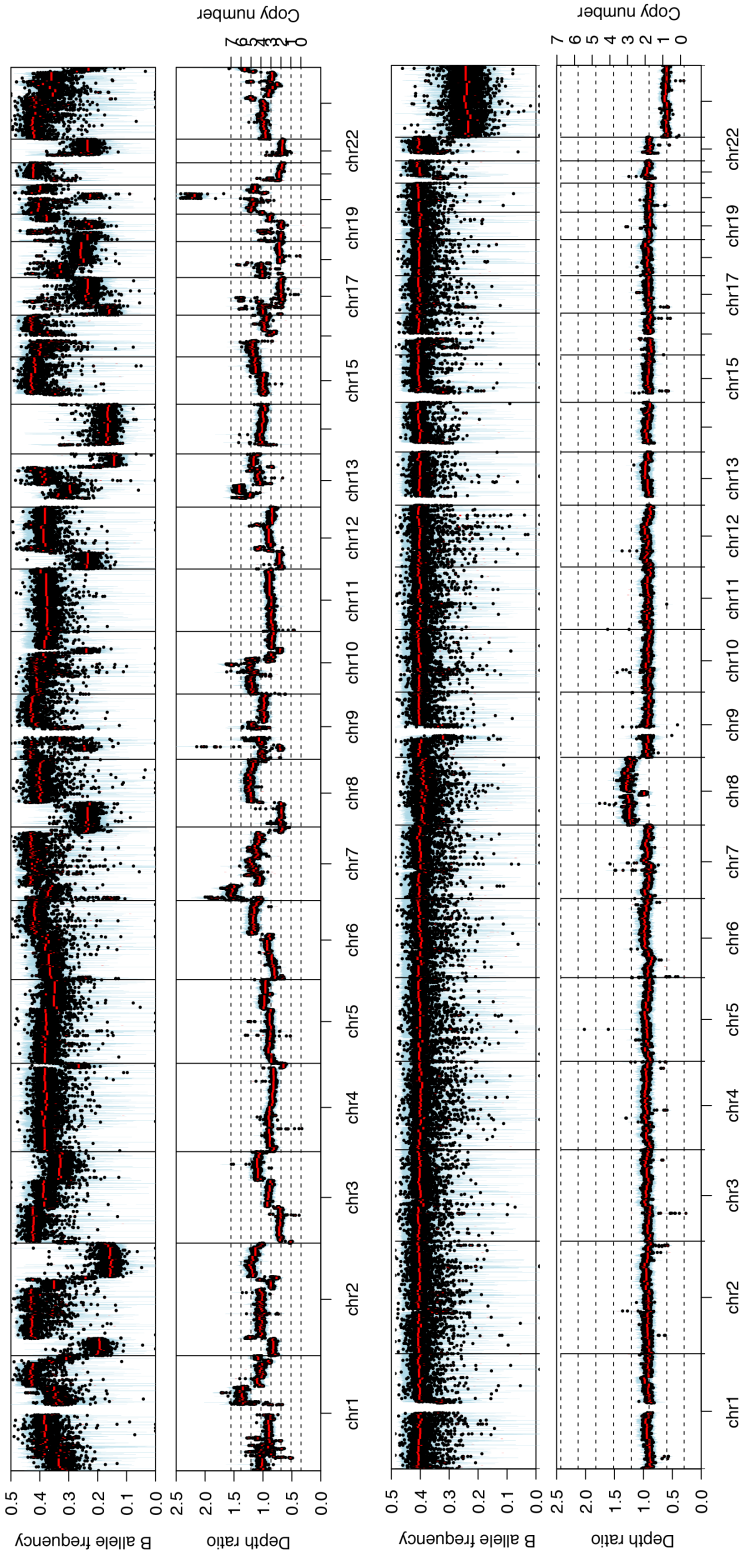

Figure 2.3: *Top panel shows B-Allele frequency and depth ratio for a highly abnormal genome, while the bottom panel shows the B-Allele frequency and depth ratio for a relatively stable cancer genome.*

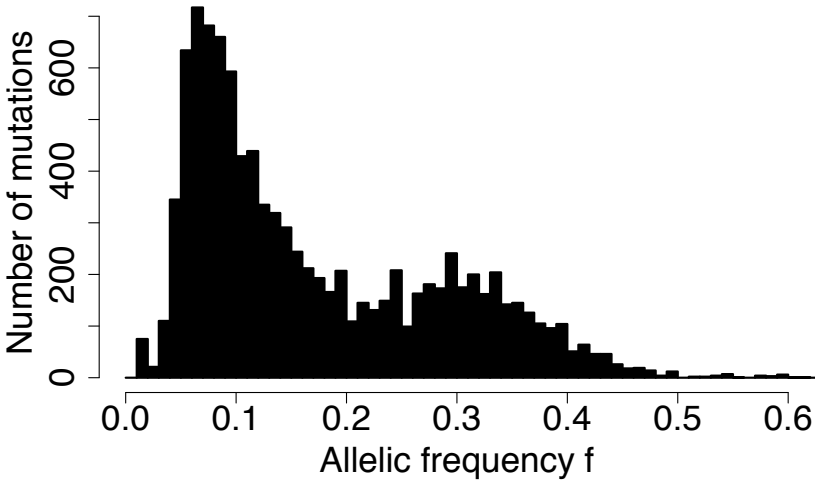

Figure 2.4: *Histogram of variant allele frequencies*

diploid regions of the genome where the relationship between VAF and lineage size is straightforward. Having acquired point mutations and their corresponding VAFs from a cancer, a useful way to summarise the data is to generate a histogram of the frequencies, that is count the number of mutations at a particular frequency. This is equivalent to the site frequency spectrum in population genetics. An example of this from a whole genome sequenced gastric cancer is shown in Figure 2.4. As will become apparent over the remainder of this thesis this distribution holds a surprisingly large amount of information on the population dynamics of the population of cells under question. This information can be unravelled by using mathematical models of what this distribution would be expected to look like under different circumstances. In the next section I will discuss approaches to model evolution and how they have been applied to cancer.

2.2 Modelling tumour evolution

The analysis of high-throughput genomics in cancer has predominantly relied on statistical methods to extract meaningful insight. This has proved fruitful in many areas, for example such studies have revealed the complexity of cancer genomes (Vogelstein *et al.,* 2013), studies of large cohorts has enabled the identification of driver mutations across many cancer types (Lawrence *et al.,* 2014), and statistical algorithms have elucidated that cancers often contain complex clonal architectures

(Nik-Zainal *et al.,* 2012a; Roth *et al.,* 2014). Despite its success, this type of approach fails to provide mechanistic insight into how and why these changes occur. For example some clinically relevant questions that were touched upon in the previous chapter might be i) how does a cancer genome change over time? ii) what is the mutation rate in cancers? iii) at what rate do subclones evolve? iv) do any of the above correlate with patient survival? Given that the data available is often a sample from a single time point. a purely statistical model is insufficient to answer these kind of questions and rather requires integration of the data with some (dynamical) model.

The use of mathematical modelling has a long history in cancer research. Early examples include the multistage theory of cancer progression that attempted to explain cancer age incidence curves through the use of mathematical models of cancer progression (Armitage & Doll, 1957; Knudson, 1971). These models suggested that a number of *hits* was necessary for a somatic cell in the body to initiate a cancer.

The use of mathematical modelling has become more popular in recent years with the advent of high-throughput data technologies and the drive to understand cancer as an evolutionary process (Altrock *et al.,* 2015; Beerenwinkel *et al.,* 2015). Taken together this has opened up the door for cancer researchers to use population genetics theory, a mathematical theory of evolution which models the evolutionary process in terms of changes in gene frequencies (Ewens, 2012), which has in turn led to this type of analyses becoming common place in many cancer genomic studies (Schwartz & Schäffer, 2017).

The simplest population genetic model is the Wright-Fisher model, in this model generations are discrete, the population size stays constant and is well-mixed. New generations are constructed by sampling from the previous one based on some offspring probability distribution. In the absence of selection this process explains the change in gene frequencies purely due to neutral drift, and quantities of interest such as fixation times and probabilities of extinction are readily calculable. A similar model is the Moran model, but rather than being discrete, generations are overlapping so that at each time point a random cell is chosen to give birth and a

random cell dies. The Wright-Fisher process has been applied to cancer evolution with extensions to include mutation, selection and multiple cell types (S Datta *et al.,* 2013). Beerenwinkel *et al.,* 2007 for example used a Wright-Fisher model to estimate the waiting time to cancer using genetic data from colorectal cancers.

For modelling growing populations, stochastic branching processes have commonly been employed. In these types of models, individuals give birth or die according to some rates, if the birth rate is greater than the death rate then the population will grow exponentially. Extensions to this simple model can be made to include mutations (Griffiths & Pakes, 1988; Champagnat *et al.,* 2012) or multiple types where differential fitness can be modelled via different birth and death rates (Antal & Krapivsky, 2011). Some applications of these types of models in the context of cancer include modelling the emergence of resistance (Iwasa *et al.,* 2006), the speed of selective sweeps (Durrett & Schweinsberg, 2004), the expected degree of heterogeneity (Durrett *et al.,* 2011) and the shape of genealogies in an expanding cancer cell population (Durrett *et al.,* 2015). Bozic *et al.,* 2010 used a branching process model with genetic data to estimate the selection coefficient, estimating a small selection coefficient of the order 0.004 per driver. This model was also used to assess the relative numbers of driver mutations to passenger mutations. These types of models that use branching processes are closely related to the Luria-Delbrück distribution, originally used to demonstrate that resistance to a particular bacterial phage in bacterial colonies was pre-existing in the population rather than in response to the introduction of the phage (Luria & Delbrück, 1943). The Luria-Delbrück distribution describes mutation accumulation in growing populations, it has received considerable attention from mathematicians since it was originally developed (Zheng, 1999). I will use some of these developments and extensions of the original Luria-Delbrück model in Chapter 5.

Often more complex models are desired where no analytical solution is available. For example one might want to include spatial effects such as migration of the cell population (Waclaw *et al.,* 2015), the effects of the tumour microenvironment (Anderson *et al.,* 2006) or a hierarchical organisation of the tumour

tissue (Poleszczuk & Enderling, 2014). In these cases where no analytical solution is available, it is necessary to simulate the evolutionary process. Simulation also allows recapitulating some of the experimental details or noise in the experimental setup, which can be important when wanting to make inferences from the model using data (Sottoriva *et al.,* 2015; Sottoriva *et al.,* 2017; Sievers *et al.,* 2016). The challenge then is how best to efficiently and accurately simulate such models. Many simulation based approaches to cancer modelling use one of the above approaches as a basis to model the birth and death of cells but add additional complexity depending on the question. For example the underlying process maybe a branching process with birth and death rates but the process evolves on a lattice to include spatial effects.

In Chapters 3, 4 and 5 I took this type of approach and use branching type processes to model the expansion of a tumour but increased the complexity by including mutation accumulation, differential fitness and elements of the data generation procedure. In Chapter 3 I use a discrete generation branching processes where birth and death were specified by an offspring probability distribution. In Chapter 4 due to the need to have multiple types and have more flexibility in expressing fitness values a continuous time branching process is used. In the case of a discrete time model as used in Chapter 2, simulation is straightforward as all that is required is drawing random numbers according to the specified offspring probability distribution for each cell at each generation.

In the case of a continuous time branching process, the simulation method needs more careful consideration. Broadly, methods to simulate these type of processes where events are governed by rate parameters are called Kinetic Monte Carlo methods. Of these methods, the most common approach is known as the Gillespie algorithm (Gillespie, 1977), which is a rejection-free Kinetic Monte Carlo. Alternatively one can use rejection Kinetic Monte Carlo methods (Schulze, 2008). In the context of a birth-death process with birth rate b and death rate d, the two algorithms are described in *Algorithms* 1 and 2.

Often the rejection-free KMC is preferred as there are no redundant time steps

Algorithm 1: Rejection kinetic monte carlo

input : birth and death rates b, d
output: Population size after time t

start with one individual, N = 1;
set r_0 such that it is $\geq b+d$;

while $t < t_{end}$ **do**
    randomly sample an individual ;
    individual gives birth with probability b/r_0, dies with probability d/r_0;
    get uniform random number $u' \in (0,1]$;
    update time $t = t + \Delta t$, where $\Delta t = (Nr_0)^{-1}\log(1/u')$

Algorithm 2: Rejection-free kinetic monte carlo (Gillespie)

input : birth and death rates b, d
output: Population size after time t

start with one individual, N = 1;
rates q_n are birth and death rates b, d;

while $t < t_{end}$ **do**
    calculate sum Q and partial sum Q_n of rates $Q = \sum_{n=1}^{N} q_n$;
    get uniform random number $r \in (0,Q]$;
    find event n that satisfies $Q_{n-1} \leq r < Q_n$;
    select event n ;
    choose random cell which undergoes event n ;
    get uniform random number $u' \in (0,1]$;
    update time $t = t + \Delta t$, where $\Delta t = Q^{-1}\log(1/u')$

where no event occurs. In some applications however the rejection KMC method is more computationally efficient. This can be the case if the rejection rate is low in which case the rejection KMC avoids the (relatively) costly searching and summation steps in the rejection-free KMC. Simulating a basic birth-death process, I found the rejection KMC to be approximately twice as fast, see Figure 2.5A. To confirm the accuracy of the simulation method it is possible to derive the exact solution of the probability distribution. Given the probability of any cell giving birth in the interval $(t, t+\delta t)$ as $b\delta t$, and the probability of dying is $d\delta t$ the differential

difference equation for the birth-death process is as follows:

$$\begin{aligned} \frac{dp_0(t)}{dt} &= dp_1(t) \\ \frac{dp_n(t)}{dt} &= b(n-1)p_{n-1}(t) - (b+d)np_n(t) + b(n+1)p_{n+1}(t) \ , \ (n \geq 1) \end{aligned} \tag{2.2}$$

The solution to this equation can be found in Bailey, 1964:

$$\begin{aligned} p_0 &= \alpha \\ p_n &= (1-\alpha)(1-\beta)\beta^{n-1} \ , \ (n \geq 1) \end{aligned} \tag{2.3}$$

where

$$\begin{aligned} \alpha &= \frac{d(e^{(b-d)t}-1)}{be^{(b-d)t}-d} \\ \beta &= \frac{b(e^{(b-d)t}-1)}{be^{(b-d)t}-d} \end{aligned} \tag{2.4}$$

Figure 2.5B shows the result of 1000 simulations together with the theoretical expectation from equation (2.3) showing that the simulation and theoretical expectation are in good agreement.

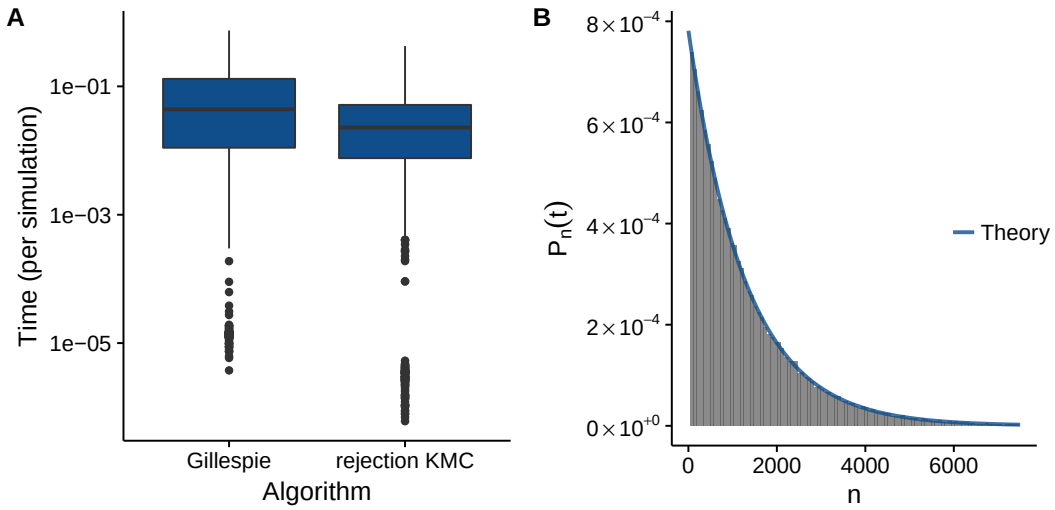

Figure 2.5: *Left plot shows the distribution of simulation times for the Gillespie algorithm and the rejection KMC algorithm. On average the rejection KMC algorithm was found to be approximately twice as fast. Simulation parameters were* $b = \log(2)$, $d = log(2)/2$ *and distributions shows the results for 100 simulations. Right hand side show that the probability distribution described by equation* (2.3) *agrees well with the simulation*

2.3 Statistical inference

As should already be apparent from Section 2.1, the type of data this thesis is principally concerned with can appear very complex. High throughput sequencing of tumour samples shows cancer genomes have mutations in varying proportions, with large regions of the genome potentially lost or amplified. On top of this complexity the data suffers from additional sources of noise such as contamination from normal tissue and lack of resolution due to limited read depth. Furthermore as discussed above, due to the complexity of the mechanistic models needed to describe these kind of data, there is often no analytical solution available and they can only be simulated. To be able to correctly infer the parameters from a model we need some way to integrate data and a model. Fortunately, Bayesian statistics provides an answer, in particular Approximate Bayesian Computation (ABC) methods allow the full power of Bayesian statistics to be applied to simulation based models. We'll first discuss some preliminaries of Bayesian statistics, before introducing the basic ABC algorithm and some of its extensions that were used later in this thesis.

2.3.1 Bayesian inference

Statistical inference is the process by which we can make quantitive conclusions from data. For example we might want to draw some generalisable conclusions from some sample of a population, or quantify how well some data we have collected fits a particular scientific model. Statistical inference comes in two principal flavours, Bayesian and frequentist. From a philosophical point of view, these two approaches differ in their interpretation of probability, to a Bayesian, probability is interpreted as a quantification of uncertainty. Frequentist inference on the other hand, interprets probability strictly as the frequency of an event over a large number of trials, assuming that the frequency will converge to the true probability as the number of trials increases. In practice these philosophical differences mean that frequentists treat the parameter of interest as fixed and the data as varying, while a Bayesian treats the data as fixed and parameters as random variables and attempts to quantify the uncertainty in parameter values based on the data. There are many

arguments for using one approach over the other which I will not go into, but I will briefly discuss the merits of Bayesian statistics for analysing large complex datasets, and the algorithms that have been developed over the last 20-30 years that have made Bayesian inference a popular choice for many applications.

The quantity of interest in Bayesian methods is the posterior distribution, $p(\theta|D)$ which quantifies the uncertainty in the inference of a particular parameter or parameter set θ, given the data D. The posterior distribution is obtained by combining any prior beliefs we may have about a particular parameter with how well that parameter value explains the observed data. This is formally expressed via Bayes rule:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{2.5}$$

where θ is a vector of parameters for our model, $p(\theta)$ is the prior distribution, ie our prior beliefs on θ, and $p(D|\theta)$ is the likelihood function. $p(D)$ is the model evidence or marginal likelihood. With the likelihood and prior distribution specified, any question can then be answered by constructing the posterior distribution, $p(\theta|D)$. The simplicity in going from assumptions to conclusion makes Bayesian inference an attractive methodology. For example a Bayesian probability interval has the common sense interpretation of having a high probability of containing the true parameter of interest, while a frequentist confidence interval should be interpreted in terms of repeated sampling, a 95% confidence interval means that 95% of repeated sampling steps should contain the true value (Gelman *et al.,* 2014).

Despite this apparent simplicity and ease of interpretation two challenges remain. First how best to choose the prior distribution, and second calculating the posterior distribution via equation (2.5) often involves high dimensional integrals and in most cases can only be solved numerically. In the past 30 years these problems have been overcome with the development of Markov Chain Monte Carlo (MCMC) algorithms to sample from the posterior and the trend to use uninformative priors, which make Bayesian inference less subjective (Hastie *et al.,* 2016), a common criticism from frequentist schools of thought.

MCMC algorithms work by constructing a Markov Chain whose equilibrium

distribution is the probability distribution of interest. Sampling from the equilibrium distribution (usually via a random walk) is then equivalent to sampling from the probability distribution. One of the earliest and most widely used algorithms that implements this is the Metropolis-Hastings algorithm (Metropolis *et al.,* 1953; Hastings, 1970). This algorithm contains two steps, a proposal step and a rejection step. The proposal step perturbs the current state, while the rejection step rejects any proposals that stray too far from regions of high probability. Although adequate for many applications, these type of random walk Metropolis algorithms tend to scale poorly for high dimensional probability distributions. Recently, to overcome this problem of inefficient sampling, Hamiltonian Monte Carlo (HMC) methods have become popular and are implemented in probabilistic programming tools such as *Stan* (Carpenter *et al.,* 2017). These algorithms construct a guided random walk by exploiting information on the geometry of the probability distribution so that proposals are made that follow contours of regions of high probability mass. This is done by calculating the gradient of the field defined by the probability distribution, and using formulations from classical physics known as Hamiltonian dynamics to explore the probability distribution efficiently (Betancourt, 2017; Mackay, 2016). Being able to efficiently calculate posterior distributions for complex hierarchical probability models is one of the principal reasons Bayesian inference has gained considerable in interest in an age where large datasets are commonplace, and many parameters are required. In Chapter 4 I used MCMC to infer the degree of overdispersion our sequencing data by modelling it as a Beta-Binomial process. I also used MCMC in Chapter 4 to implement Dirichilet process clustering, a popular clustering approach used to cluster mutation frequencies and uncover the clonal structure of tumours (Roth *et al.,* 2014; Nik-Zainal *et al.,* 2012b).

2.3.2 Approximate Bayesian Computation

Despite its wide applicability, Bayesian inference using MCMC algorithms is not suitable for all problems. For complex models involving multiple parameters or based on a simulation, analytical forms for the likelihood are intractable or unavailable, such models can however often be simulated efficiently on a computer. A class

of methods called Approximate Bayesian Computation (ABC) has been developed to tackle this problem where calculation of the likelihood is not available. These methods rely on comparison of simulated data with real data via some distance measure and allow the full power of Bayesian inference to be applied to simulation based models as used in this thesis in Chapter 4.

ABC methods were first developed for applications in the field population genetics. Tavaré *et al.,* 1997 was the first application of using ABC to approximate the posterior distribution. In this article ABC was used to infer the most recent common ancestor from samples of a population under different demographic models. ABC methods were further developed for applications in population genetics (Beaumont *et al.,* 2002) and are now used in a wide range of applications in fields as diverse as ecology and physics (Sunnåker *et al.,* 2013; Lintusaari *et al.,* 2016). The flexibility in being able to construct models of arbitrary complexity and parameterise them accurately makes ABC an attractive method for high throughput biological data in particular. Recently, some studies have used genomic data from cancer together with ABC to look at stem cell organisation (Sottoriva *et al.,* 2013b), mutation accumulation (Zhao *et al.,* 2017) and colon cancer growth dynamics (Sottoriva *et al.,* 2015; Sievers *et al.,* 2016).

2.3.3 ABC rejection

The simplest ABC algorithm is the rejection algorithm (Pritchard *et al.,* 1999; Tavaré *et al.,* 1997) which compares simulated data D^* with parameters θ and the target data D, and if they match sufficiently well accepts the parameters θ:

S1 Sample θ^* from $p(\theta)$

S2 Simulate a dataset D^* from model $M(D|\theta^*)$

S3 If $d(D^*, D) \leq \varepsilon$, accept θ^*, otherwise reject

S4 Return to S1

As $\varepsilon \to 0$, the estimates of θ will converge to the true posterior. In most cases, rather than using the full data the distance is calculated based on a set of summary

statistics.

In many applications we may have a number of competing models and would like to infer the most probable model. For this we can turn to Bayesian model selection. If m_0 and m_1 are two models, we would like to choose which model provides the best support for the data, Bayes factors, the ratio of posterior odds to prior odds of the two models provides a way to quantitively test which of these models has the greater support. The Bayes factor in favour of m_0 over m_1 is defined as:

$$B_{01} = \frac{P(m_1|D)/P(m_2|D)}{P(m_1)/P(m_2)} \tag{2.6}$$

Where $P(m_n)$ is the prior probability of model n and $P(m_n|D)$ is the posterior probability.

Incorporating model selection into the ABC framework is relatively straightforward as we can effectively treat the model as an additional parameter in the inference scheme, where each model m_n will have a corresponding model specific parameter vector θ_n. The ABC rejection with model selection then becomes (Grelaud *et al.,* 2009):

S1 Sample m^* from $p(m)$

S2 Sample θ^* from $p(\theta|m^*)$

S3 Simulate a dataset D^* from model $M(D|\theta^*, m^*)$

S4 If $d(D^*, D) \leq \varepsilon$, accept (m^*, θ^*), otherwise reject

S5 Return to S1

It has been shown that use of ABC to calculate Bayes factors can result in inconsistent results due to the loss of information from using summary statistics (Robert *et al.,* 2011), these issues however can be overcome if the full data is used rather than summary statistics (Barnes *et al.,* 2012).

The downside of the ABC rejection algorithm is that the acceptance rate is generally low, requiring a large amount of datasets to be simulated. A number of

extensions of the basic ABC rejection approach exist such as ABC MCMC (Marjoram *et al.,* 2003), which uses a Metropolis Hastings step to sample more efficiently from the posterior. Another approach is to use sequential importance sampling (Del Moral *et al.,* 2006) to propagate a set of parameter vectors through a sequence of ever decreasing tolerances (ε) until it is small enough to provide an accurate estimate of the posterior distribution. This algorithm, called Approximate Bayesian Computation Sequential Monte Carlo (ABC SMC) also provides increased efficiency over the basic ABC rejection and overcomes issues in ABC MCMC where the sampler can get stuck in regions of low probability for extended periods of time. Additionally the ABC SMC algorithm can be extended to perform Bayesian model selection (Toni *et al.,* 2009; Toni & Stumpf, 2010). For these reasons I implemented an ABC SMC algorithm which was used in Chapter 4.

2.3.4 ABC SMC

In ABC SMC, parameter vectors, particles (m_n, θ_n) are sampled from the prior distribution and then propagated through a series of distributions with decreasing tolerances, ε_i, until $\varepsilon_i = \varepsilon_T$ the target tolerance. We therefore gradually evolve toward the target posterior distribution $p(\theta|d(D^*,D) \leq \varepsilon_T)$ as ε_i decreases. The ABC SMC model selection algorithm is as follows (Toni & Stumpf, 2010):

S1 Set the population indiciator to $t = 1$

S2 Set the particle indiciator $i = 1$

S3 If $t = 1$, sample (m^{**}, θ^{**}) from the prior distribution $P(m, \theta)$
if $t > 1$, sample m^* from $P_{t-1}(m^*)$ and then perturb according to $m^{**} \sim KM_t(m|m^*)$. Sample θ^* from previous populations with weights w_{t-1} and perturb parameter vector according to $\theta^{**} \sim KP_{t,m^{**}}(\theta|\theta^*)$

S4 If $P(m^{**}, \theta^{**}) = 0$, return to **S3**

S5 Simulate data D^* for model m^{**} and parameters θ^{**}, then calculate $d(D^*,D)$, if $d(D^*,D) > \varepsilon_t$ go to **S3**

S6 Set $(m_t^i, \theta_t^i) = (m^{**}, \theta^{**})$ and calculate the weight of the particle w_t. If $i < N$ set $i = i + 1$ and go to **S3**

S7 Normalize the particle weights and calculate the marginal model probabilities, $P_t(m_t = m) = \sum_{i,m_t^i=m} w_t^i(m_t^i, \theta_t^i)$

S8 Calculate the perturbation kernels and next tolerance value ε_t, if $\varepsilon_t > \varepsilon_T$, set $t = t + 1$ and go to **S3**.

The particle weights are calculated as follows:

$$w_t^i(m_t^i, \theta_t^i) = \begin{cases} 1, & \text{if } t = 1 \\ \frac{P(m_t^i, \theta_t^i)}{S}, & \text{if } t > 1 \end{cases} \tag{2.7}$$

where S is:

$$S = \sum_{j=1}^{M} P_{t-1}(m_{t-1}^j) KM_t(m_t^i, m_{t-1}^j) \times \sum_{k, m_{t-1} - m_t^i} \frac{w_{t-1}^k KP_{t,m_t}^i(\theta_t^i | \theta_{t-1}^k)}{P_{t-1}(m_{t-1} = m_t^i)} \tag{2.8}$$

Here KM is the model perturbation kernel and KP is the parameter perturbation kernel. Particles that have been sampled from the previous distribution are denoted by a single asterisk, the perturbed particles are denoted with a double asterisk. To implement the ABC SMC algorithm one needs to choose the perturbation kernels, KM and KP. For the model perturbation kernel a simple approach is to assign probabilities at the beginning for models to stay the same after perturbation, for example:

$$KM_t(m|m^*) = \begin{cases} \alpha, & \text{if } m = m^* \\ \beta, & \text{if } m \neq m^* \end{cases} \tag{2.9}$$

where α and β are number between 0 and 1 and $\alpha + \beta = 1$. For the particle perturbation kernel the simplest approach is to use the uniform distribution with limits determined from the range of parameter values from the previous population (Filippi *et al.,* 2013), for parameter k, $KP_t(k|k^*) = U(k_i - \sigma, k_i + \sigma)$, where σ is given

by:

$$\sigma = \frac{1}{2}(max(k)_{t-1} - min(k)_{t-1}) \tag{2.10}$$

Other perturbation kernels are also possible such as using the normal distribution or multivariate normal when it is known a priori that parameters may be correlated (Filippi *et al.,* 2013). Finally, the ABC SMC algorithm requires choosing the tolerance schedule, or alternatively implementing an adaptive tolerance schedule where for example the tolerance is taken as the αth quantile of the distances of the previous population.

2.3.5 Algorithm performance and accuracy

To illustrate the accuracy of the ABC approach and to demonstrate the increased efficiency of the ABC SMC algorithm over the basic ABC rejection sampler I will introduce a basic example where we can both estimate the underlying parameters and perform Bayesian model selection. We will use an example (originally from (Grelaud *et al.,* 2009)) that takes advantage of conjugate priors so that calculation of the Bayes factor is analytically tractable, in this way comparison between the true Bayes factor and calculation of the Bayes factor via ABC is possible (Didelot *et al.,* 2011).

I'll consider 2 models, in the first model (M_1), observations are distributed according to $Poission(\lambda)$ while in the second (M_2) they are assumed to be $Geometric(\mu)$. For these models the likelihoods are given by

$$p(x|\theta_1, M_1) \propto exp\left(\sum_{j=1}^{n} x_j\theta_1 - \sum_{j=1}^{n} \log x_j!\right) \tag{2.11}$$

$$p(x|\theta_2, M_2) \propto exp\left(\sum_{j=1}^{n} x_j\theta_2\right) \tag{2.12}$$

where $\theta_1 = log(\lambda)$ and $\theta_2 = log(1-\mu)$. I'll assign equal probabilities to both models and assign an $Exponential(1)$ prior to λ in M_1 and a $Uniform(0,1)$ prior to μ in M_2. As these priors are conjugate to the likelihoods we can calculate the

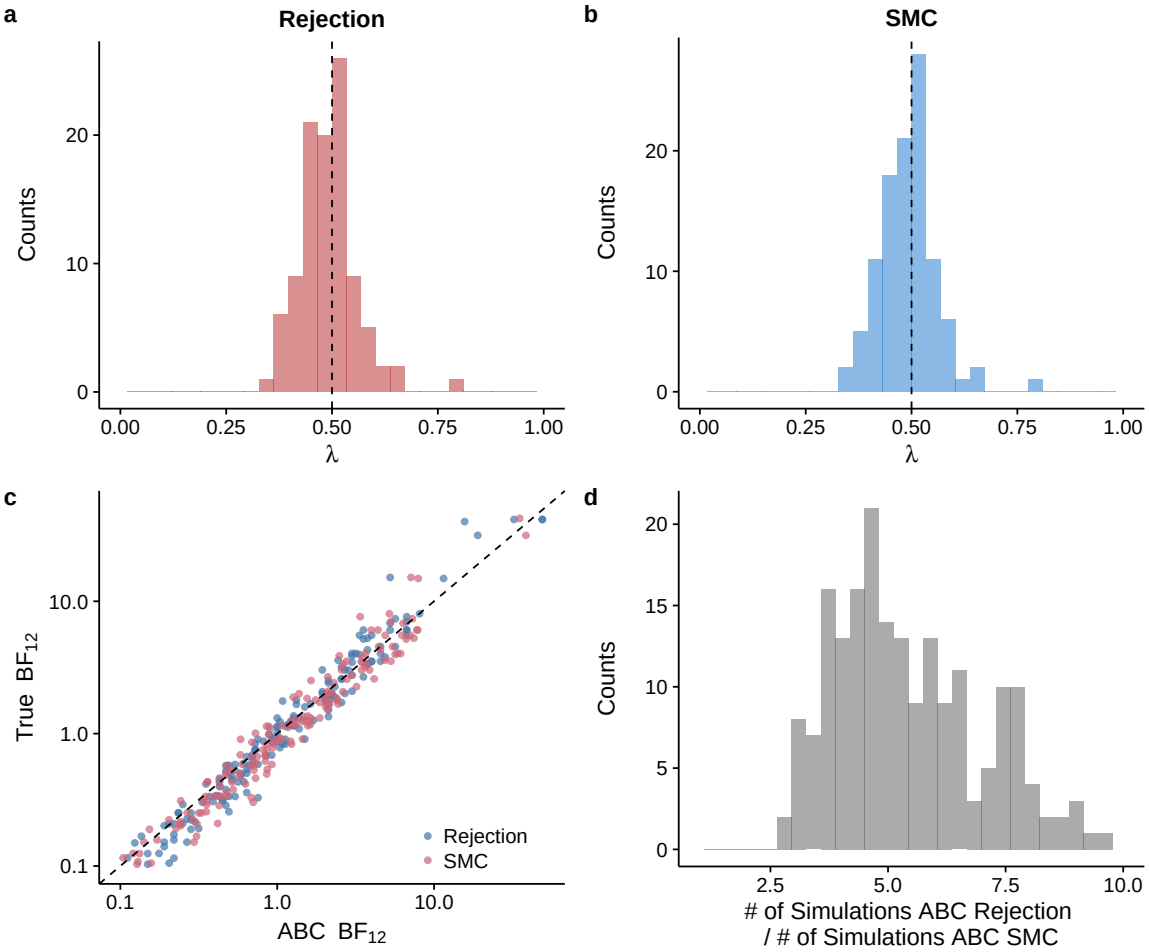

Figure 2.6: ***a***, ***b*** *Shows that the ABC rejction algorithm and the ABC SMC algorithm both accurately identify the true parameter across 50 simulations, dashed line is the true* $\lambda = 0.5$ *from model 1.* ***c*** *Both methods correctly calculate the Bayes Factors (dashed line is* $x = y$*) while the ABC SMC algorithm shows increased efficiency over the ABC rejection algorithm* ***d***

marginal probabilities of the models:

$$p(x|M_1) = \frac{s_1!}{exp(t_1) \times (n+1)^{s_1+1}} \tag{2.13}$$

$$p(x|M_2) = \frac{n!s_1!}{(n+s_1+1)!} \tag{2.14}$$

It is often convenient to summarise data with 1 or 2 numbers, when the posterior distribution of a parameter of interest depend on the data only through one of these summary statistics then these statistics are said to be sufficient, that is they provide the maximal amount of information. For the above models the sufficient statistics

(s_1, t_1) are given by:

$$s_1 = \sum_j x_j \tag{2.15}$$

$$t_1 = \sum_j \log x_j! \tag{2.16}$$

I will use these sufficient statistics as my summary statistics in the ABC, and so the distance will be calculated based on these.

With the models set up it is now possible to assess how accurately the two ABC algorithms introduced above infer the correct parameters and Bayes factors. I generated 100 datasets from M_1 with $\lambda = 0.5$ and applied both algorithms to these datasets. Figure 2.6c shows a good agreement between the true Bayes factor and approximate Bayes factor from ABC, while both algorithms also inferred the correct model parameter, Figures 2.6a and 2.6b. For this example, the ABC SMC algorithm exhibits an average 5 fold increase in efficiency see Figure 2.6d.

2.4 Software

All the bioinformatic analysis was done on a linux based high performance computer, with the pipelines implemented in bash scripts. Any subsequent analysis was done in the R statistical programming language, including plotting which was done in base R or ggplot2 (R Core Team, 2016; Wickham, 2009). Simulations were written in the Julia technical programming language ("Julia: A fresh approach to numerical computing"), a relatively new programming language that has been designed specifically for scientific technical applications. Julia uses *just in time* compilation to achieve speeds comparable to statically typed languages such as *C* and *Fortran* by inferring function types and aggressively specialising code based on the type inference. The ABC algorithms described above where also implemented in Julia and is available as a package at `https://github.com/marcjwilliams1/ApproxBayes.jl`.

Chapter 3

Identification of neutral tumour evolution across cancer types

3.1 Introduction

Despite much progress in understanding how selection and mutation shape the cancer genome, the role of neutral processes has largely been neglected. Given that the vast majority of point mutations are thought to be passengers and the relative paucity of putative driver mutations per cancer (Lawrence *et al.,* 2014), it is conceivable that tumours undergo periods of stable growth. Neutral evolution provides the null model for intra-tumour genetic diversity, and provides the necessary theoretical framework to identify selection via deviations from this neutral model.

Recent studies that have formally tested this assumption in colorectal cancer and its precursor lesions have surprisingly shown no evidence of strong selection. Siegmund *et al.,* 2009 showed using methylation pattern diversity that the molecular age of glands (clonal units within colon tumours) from different regions of the same tumours were similar, consistent with a single expansion. Similar observations have been made in adenomas, a precursor lesion to cancers of the colorectum (Humphries *et al.,* 2013). More recently, the *Big Bang* model of clonal evolution was proposed and validated using genomic data from colon cancer (Sottoriva *et al.,* 2015). This *Big Bang* model of clonal evolution, posits that once a cell has acquired the genetic alterations necessary for malignancy the resultant clonal expansion is effectively

neutral, with the size of lineages within the expansion determined by the time they appear rather than stringent strong selection. Sottoriva *et al.* found evidence for this model of tumour evolution by observing mutations that were not pervasive (present in all cancer cells) could be found on opposite sides of the tumour. A computational model showed that these mutations must have appeared early when the cancer was small suggesting that these cancers resulted from a single clonal expansion with a lack of strong selection. Studies in mice have also suggested that intra-tumour heterogeneity can be explained by stochastic processes due to a proliferative hierarchy rather than functional differences between sub populations of cells (Driessens *et al.,* 2012).

In this chapter I will develop a mathematical model of neutral tumour evolution that can be applied to widely available sequencing data to test whether a neutral model of cancer evolution is consistent with other cancer types.

3.2 Neutral tumour evolution

To test whether a neutral model of cancer evolution can plausibly explain the genetic variation we observe in cancer cell populations I'll take inspiration from population genetics theory and apply a model that describes the theoretical expectation of a neutral model to widely available cancer genomic data such as The Cancer Genome Atlas (TCGA). To summarise and recap briefly some of the key points from Chapter 2, deep sequencing of cancer cell populations measures the size of cell lineages within the population and this data can be summarised via a histogram of these lineage sizes. Developing a model of the expected distribution of lineage sizes will can then be used to test whether these models fit the observed data and extract features of the cancer cell population dynamics from data taken at a single time point. The model I will now develop is a simple neutral growth model where all cells proliferate at the same rate and accumulate mutations. I will derive the expected distribution of lineage sizes and then apply this to large cohorts of both whole genome sequencing data and whole exome sequencing data.

3.2.1 Mathematical model of neutral tumour evolution

Beginning at time $t = 0$ with a single transformed cancer cell, the number of tumour cells at time t will be $N(t)$. Given a growth rate b, a mutation rate μ per chromosome set and the ploidy of the tumour (number of chromosomes per cell), π, the differential equation describing the expected number of mutations at some time t, $M(t)$, can be written as

$$\frac{dM(t)}{dt} = \pi b \mu N(t). \tag{3.1}$$

Solving this equation requires integrating over some growth function $N(t)$

$$M(t) = \pi b \mu \int_{t_0}^{t} N(t). \tag{3.2}$$

The simplest model of growth is exponential growth which is given by,

$$N(t) = e^{b\beta t}. \tag{3.3}$$

Where β denotes the fraction of cells producing 2 viable offspring, it is only these divisions that contribute to the growth of the population. Substituting this expressions into equation (3.2) we arrive at

$$M(t) = \frac{\mu \pi}{\beta} \left(e^{b\beta t} - e^{b\beta t_0} \right). \tag{3.4}$$

This equation is analogous to the Luria-Delbruck distribution, originally developed to describe the accumulation of mutations in bacteria over time (Zheng, 1999). This equation is however of little use for modelling cancer evolution as it is dependent on time and obtaining time resolved sequencing data from human cancers is challenging for obvious reasons. What we do know however is that the frequency of a mutation, f will be inversely proportional to the population size (or more accurately the number of chromosome sets in the population) when it arose in the population

$$f = \frac{1}{\pi N(t)} = \frac{1}{\pi e^{b\beta t}}. \tag{3.5}$$

Crucially in the absence of selection or significant genetic drift the frequency will remain (approximately) constant as the population grows. For example if a tumour is comprised of 100 cells and 1 cell acquires a mutation its frequency will be $1/100$, if the tumour then doubles, the tumour will have 200 cells, with 2 of those cells carrying the mutation giving a frequency of $2/200 = 1/100$. Frequency and time are therefore (approximately) equivalent under neutral growth dynamics. With this we can finally substitute equation (3.5) into equation (3.4) with which we arrive at

$$M(f) = \frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{max}}\right), \tag{3.6}$$

where $f_{max} = e^{-b\beta t_0}$. Given that the VAF distribution from high throughput sequencing data measures the frequency of mutations, equation (3.6) provides a means to first test whether individual cancers are consistent with a neutral exponential expansion and secondly to measure the effective mutation rate $\hat{\mu}$ $(= \mu/\beta)$, (the mutation rate scaled by the death rate) on a sample by sample basis. To summarise, this model predicts that cumulative number of mutations with frequency greater than f, $M(f)$ should be linear in $1/f$. With this it is then possible to fit a simple linear model to the data and measure the goodness of fit. This result converges to results obtained in similar models in the population genetics and stochastic processes literature (Durrett, 2013a; Griffiths & Tavaré, 1998). The probability density of mutations in the VAF distribution (ie in non cumulative space) has also been derived and follows a $1/f^2$ dependence (Keller & Antal, 2015; Kessler & Levine, 2013; Ohtsuki & Innan, 2017; Nicholson & Antal, 2016), I will return to this in Chapter 5.

3.2.2 Stochastic simulations

Before proceeding to apply equation (3.6) to real sequencing data I first confirm its prediction via a stochastic simulation of tumour growth and investigate how selection causes deviations from the predicted distribution. The simulation scheme used here is a simplified version of the one presented in the next chapter so I will discuss it briefly here.

The approach taken was to generate synthetic data sets that capture the characteristics of NGS data, and by exploring various evolutionary histories of tumour growth verify the predictions of the model. NGS data is plagued by various sources of noise, these include tumour sampling, contamination of the sample with non-tumour cells, limited sequencing depth and difficulties in mutation calling, particularly at low frequencies. The model I implemented models tumour growth using a branching type process in discrete generations and then generates synthetic datasets from the model output using a empirically motivated sampling procedure.

The simulation begins with a single "transformed" cancer cell that gives rise to the malignancy. Cells then die and proliferate at each generation by sampling from an an offspring probability distribution The offspring probability distribution can be written as $P = (p_0, p_1, p_2)$, where p_n is the probability of having n offspring. Therefore assuming exponential growth the population at time t will be given by

$$N(t) = X^t = e^{ln(X)t}, \tag{3.7}$$

where X is the average number of offspring per cell (expectation of P) and t is in units of generations. $X = 2$ is the case where there is no cell death and every division produces 2 viable offspring. At each division, cells acquire mutations at a rate μ and it is assumed every mutation is unique (infinite sites approximation) (Ewens, 2012). The number of mutations acquired by a newborn cell at division is a random number drawn from a Poisson distribution. I record the evolutionary history of the population by recording the parent of each newborn cell, this allows reconstructing the entire history of the tumour and calculate the variant allele frequencies of all mutations in the population. Selection can be incorporated into the model by introducing populations of cells with a different offspring probability distribution, populations of cells with positive selection will for example have on average a larger number of offspring per generation than wildtype cells ($E[P_{mutant}] > E[P_{wildtype}]$. Finally, the output of the simulation - cells with associated mutations - undergoes a process of Binomial sampling to produce synthetic data that mimic the characteristic noise associated with sequencing.

Given a simulated data set we can then fit the analytical model and asses the goodness of fit of the model and extract the effective mutation rate. To fit the analytical model to the synthetic data I used the same methodology that I use later for fitting the real sequencing data. I fitted the model using subclonal mutations in the frequency range ($f_{min} = 0.12, f_{max} = 0.24$). The lower limit was chosen to mitigate against the resolution limit of moderate depth (50-100X) sequencing data which was empirically observed to be $0.05 - 0.1$. The upper limit was chosen to ensure that only subclonal mutations were interrogated, for a diploid genome we would expect to observe mutations at a frequency 0.5 while in a tetraploid genome these clonal mutations would be expected to be at frequency 0.25, therefore mutations < 0.25 would be expected to be subclonal. Exploiting the constraint on the intercept of the linear model given by equation (3.6) the model $y = m\left(x - \frac{1}{f_{max}}\right) + 0$ can be fitted using ordinary least squares, where $y = M(f)$, $x = 1/f$ and the value extracted from the fit being the effective mutation, which is known *a priori* for the simulations. The R^2, coefficient of determination statistic was used to assess the goodness of fit, where values closer to 1 are indicative of good fits. The ability to recover the true value for the mutation rate was also assessed.

3.2.3 Simulation results

3.2.3.1 Neutral tumour growth

Figure 3.1 shows an example synthetic dataset generated from the model, qualitative comparisons between real data shown in Figure 2.4 from the previous chapter and this synthetic data confirm that the model is able to generate synthetic datasets that have similar characteristics to real NGS data. Applying the fitting methodology described above, Figure 3.1B shows that under neutrality, the simulation validates the prediction of equations (3.6) of a linear relationship between $M(f)$ and $1/f$.

To assess the robustness of the inferences from fitting equation (3.6), a cohort of 10,000 synthetic tumours was generated and the ability to recover the input mutation rate and the R^2 values was assed. Analysis of these simulations confirmed that the model is robust to the noise introduced from sequencing. On average we recover the input mutation rate to within average error of 1% (Figure 3.2) and the

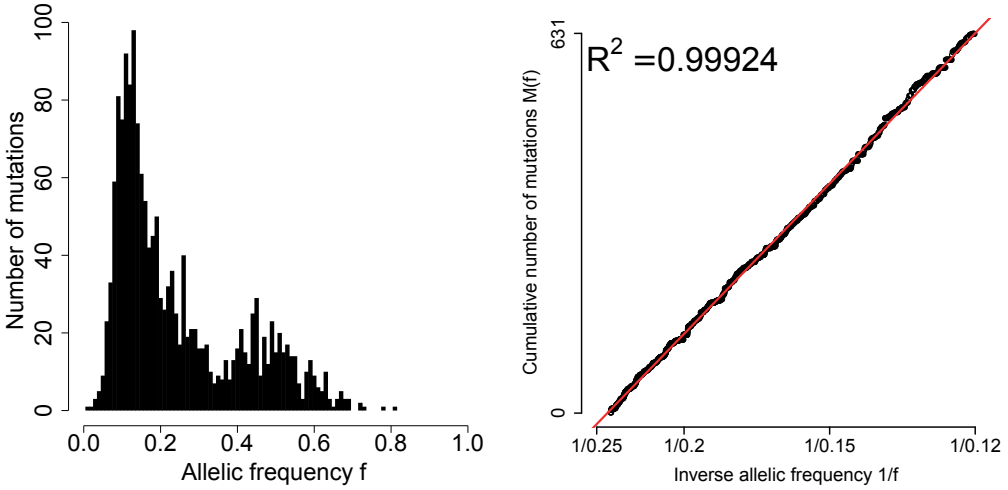

Figure 3.1: ***A*** *We were able to produce realistic synthetic NGS data using a stochastic simulation of tumour growth that accounts for neutral accumulation of mutations in the tumour as well as the different sources of noise of sequencing (sampling, sequencing depth and normal contamination).* ***B*** *The prediction of the analytical model on the cumulative distribution of subclonal allelic frequencies agrees with the stochastic simulation.*

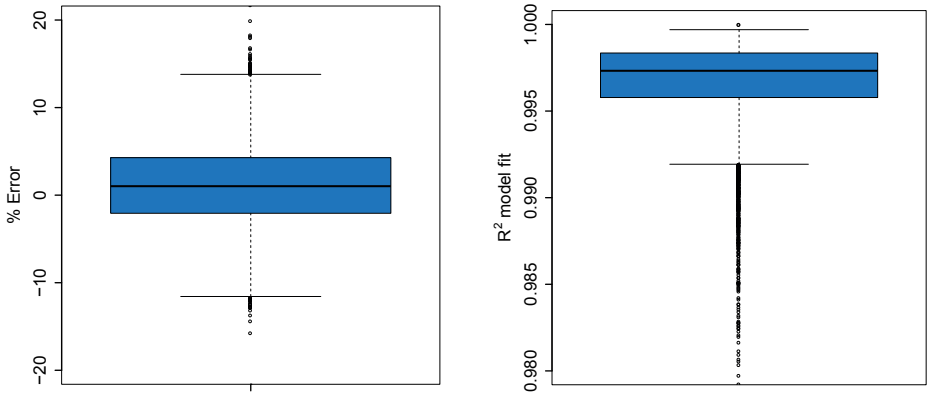

Figure 3.2: *Over 10,000 simulations, the interquartile range of the percentage error in the estimates of the mutation rate is <5%, demonstrating the ability of the analytic model to accurately estimate tumour growth parameters from NGS data. The R^2 values of the fits are consistently high over 10,000 simulations. Unless otherwise stated the input parameters for the simulation and subsequent sampling were $\mu = 100$ mutations/cell division, $b = ln(2)$, detection limit = 10%, normal contamination = 0%, depth = 100X and number of clonal mutation = 200.*

R^2 of the model fit is consistently high (> 0.98), see Figure 3.2.

3.2.4 Effect of selection on the allelic frequency distribution

To confirm that NGS data that follows the allelic frequency distribution of equation (3.6) is indeed dominated by neutral tumour evolution we would predict that including selection in the model results in synthetic data that deviates from (3.6). Introducing a second fitter population early during tumour growth causes an overrepresentation of variants at high frequency compared to what we would expect from a model of neutral tumour growth. This is evident in Figure 3.3A with the appearance of an intermediate peak between the clonal peak and the $1/f$ tail. This causes the cumulative distribution to deviate from the linear relationship predicted by neutral growth. An overrepresentation of variants at high frequency, as compared to what we would expect from our null model is caused by the selection of the fitter subclones. These variants that cluster at higher frequency than would be expected from a neutral model will primarily be passenger mutations that hitchhike to higher frequency on the back of the fitter subclone (Fay & Wu, 2000; Gillespie, 2000). These high frequency clusters are consistent with previous studies that have shown that selected subclones produce distinct clusters in the VAF distribution (Nik-Zainal *et al.,* 2012b).

Simulations also demonstrate that a change in mutation rate early during tumour growth results in deviations from the predicted distribution, see Figures 3.3C & 3.3D.

3.3 Neutral evolution across cancer types

Having validated that neutral evolution should be observable and not obscured by sequencing noise in synthetic datasets, next I applied the model to large publicly available datasets.

3.3.1 Data & Data Processing

A large cohort of gastric cancers sequenced to high depth across the whole genome from a study by Wang *et al.,* 2014a was acquired, which was used to test the prediction of the neutral null model. We also tested the model on data from the TCGA and Sottoriva *et al.,* 2015.

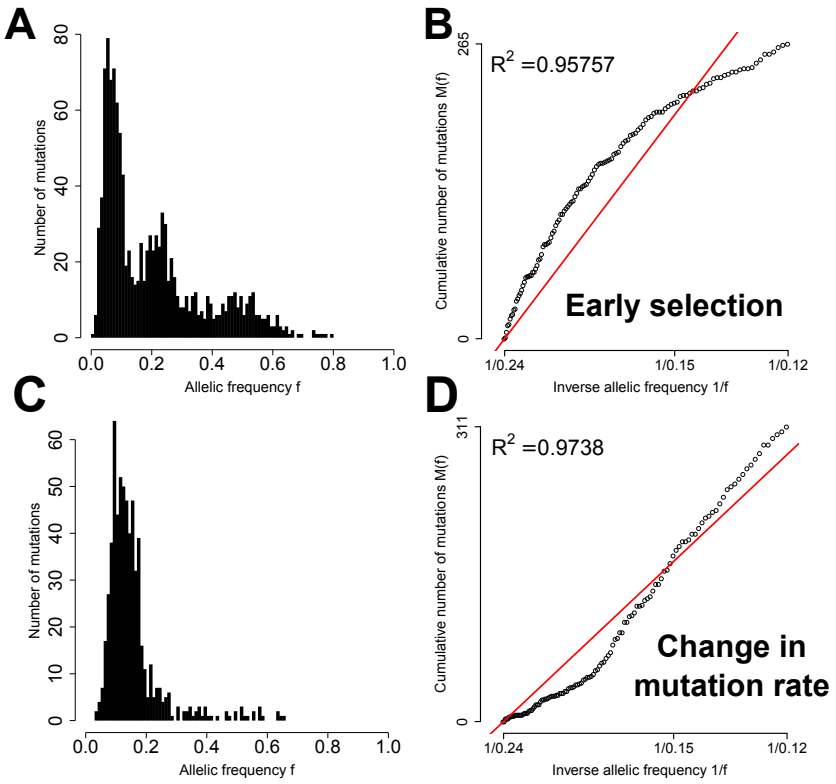

Figure 3.3: *By introducing a second population with a large 65% fitness advantage* ($P_{ln(1.2)} = (p_0 = 0, p_1 = 0.8, p_2 = 0.2)$, $Q_{ln(1.98)} = (q_0 = 0, q_1 = 0.02, q_2 = 0.98)$) *when the tumour is comprised of 80 cells we see a second peak at VAF∼ 0.2 (**A**) and a bend in the cumulative distribution plot (**B**). A new phenotypically distinct clone introduced with a 10-fold higher mutations rate (20 per division to 200 per division) also produces a deviation from neutrality (**C** & **D**).*

The gastric cancer data was acquired in the BAM format which then required point mutation calling as well as copy number calling. Point mutations were called using the VarScan2 software package (Koboldt *et al.,* 2012) and then annotated using ANNOVAR (Wang *et al.,* 2010). Using the output from Varscan, mutations were filtered out if the depth of coverage was below 10X in either the tumour and normal sample and fewer than 3 reads reported the variant in the tumour sample. The Sequenza software package was used to produce allele specific copy number segmentations across the whole genomes (Favero *et al.,* 2015).

A pan-cancer cohort from the TCGA and 2 colon cancer cohorts (one from TCGA and one from Sottoriva *et al.*) were examined. As these cohorts consisted of

exome sequencing data there were fewer mutations called. For this reason an additional filtering criteria requiring that there must be at least 12 subclonal mutations in the range of interest (f_{min},f_{max}) for a sample to be included in the analysis was included. Additionally, across all cohorts only samples with tumour purity $\geq 70\%$ were included in the analysis (as determined by inspection from a pathologist and reported in the TCGA metadata), this is because low tumour purity can confound the results resulting in clonal mutations being present in the integration range.

3.3.2 Gastric cancer results

An allelic frequency distribution as measured by NGS whole-genome sequencing of a gastric cancer is shown in Figure 3.4A (this is the same data presented previously in Figure 2.4. This exhibits all the characteristics predicted from the analytical model and the stochastic simulation of tumour growth. Transforming this data and plotting as the cumulative distribution, $M(f)$ shows that subclonal mutations in this tumour follow the distribution predicted by equation (3.6), Figure 3.4B. The high goodness of fit measure R^2 indicating that the growth dynamics of this tumour was dominated by neutral evolutionary dynamics.

Examining all 78 samples from the gastric cancer cohort shows that a large proportion are dominated by neutral evolutionary dynamics. We classify tumours as being dominated by neutral evolutionary dynamics if the model fit produces a R^2 value > 0.98. Stratifying according to micro-satellite stability, 57/68 (76.9%) MSS (micro-satellite stable) cancers were classified as neutral compared to 3/10 (30%) for MSI cases (micro-satellite unstable). We also stratified according to coding and non-coding regions, due to a smaller number of mutations in coding regions fewer samples were classified as neutral, although those that were, were also classified as neutral when considering mutations across the whole genome Figure 3.4C.

The model also allows us to estimate the effective mutation rate of subclonal mutations for those samples that grow as neutral clonal expansions ($R^2 > 0.98$). Cancers were stratified into micro satellite stable (MSS) and micro-satellite unstable (MSI). Micro satellite unstable cancers have defects in their mismatch repair machinery which result in elevated mutational loads. Mutation rate estimates for

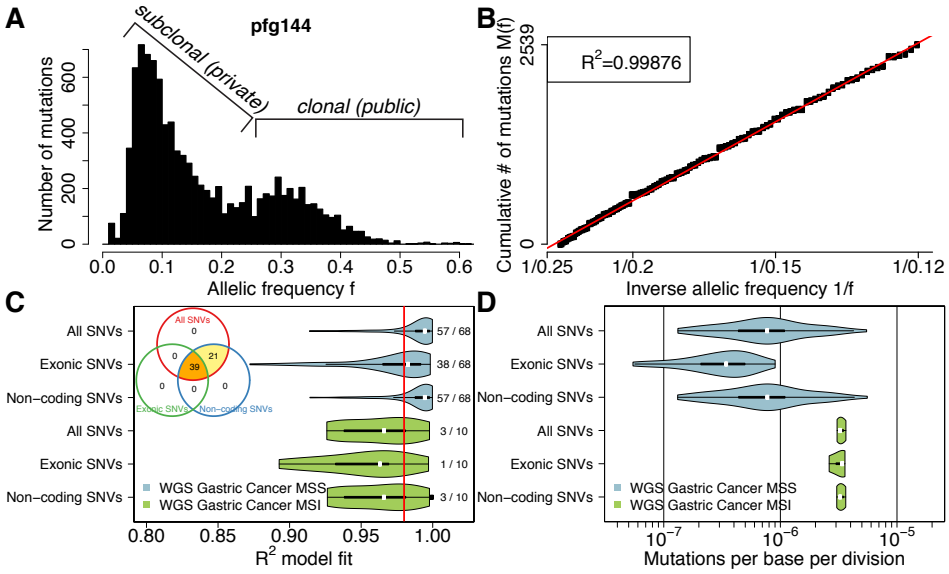

Figure 3.4: *Variant allele frequency **A** and corresponding cumulative distribution **B** for sample pfg144. **C** R^2 model fit across the whole-genome gastric cancer cohort separated into MSS and MSI cancers as well as exonic and non-coding SNVs. Exonic SNVs are mutations found in the exome (coding region of the genome). 60/78 samples (76.9%) of samples were consistent with neutral evolution ($R^2 >$ 0.98). **D** Mutation rates across the gastric cohort for the 60 samples that are consistent with neutral evolution. MSI cases exhibit a 4-fold higher mutation rate.*

MSI cancers were over 4-fold higher than MSS cases, consistent with this known biological mechanism (Figure 3.4D).

3.3.3 Colon cancer cohort

Given that colon cancer was the first cancer type where neutral evolution was observed and explained the observed intra-tumour heterogeneity, next I examined 2 colon cancer cohorts, one from TCGA and one from Sottoriva *et al.,* 2015. Figures 3.5 A & B show an example VAF distribution and corresponding cumulative distribution, which again shows the characteristics predicted by the neutral evolution model. As in the gastric cancer cohort a large proportion of tumours fit the neutral model well, 4/7 in the Sottoriva *et al.,* 2015 cohort, 31/82 in CIN (copy number unstable) TCGA colon cancers and 3/19 MSI positive TCGA colon cancers, see Figure 3.5C. Again, as would be expected MSI colon cancers had higher mutation rates, Figure 3.5D.

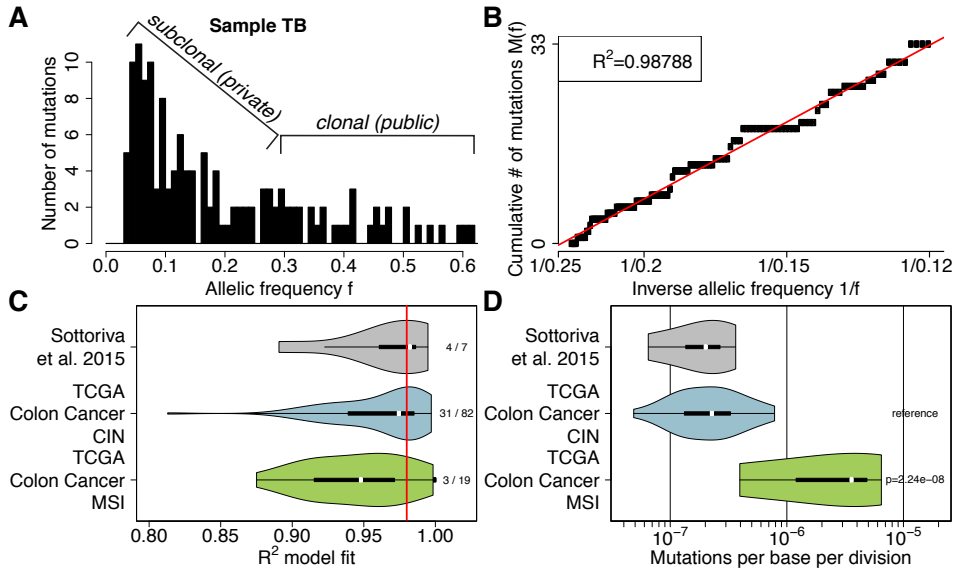

Figure 3.5: *Variant allele frequency **A** and corresponding cumulative distribution **B** for sample TB from Sottoriva* et al., *2015. **C** R^2 model fits across the 2 cohorts with TCGA stratified into MSI and CIN cancers. **D** Higher effective mutation rates were observed in the MSI cancers as would be expected.*

3.3.4 Pan-cancer cohort results

The final data-set included 819 exome sequenced samples from 14 tumour types, again from the TCGA consortium (Figure 3.6). Again a high proportion of samples fitted the neutral evolution model (259/819, 31.6%). Interestingly, some cancers types exhibited a consistently high model fit ($R^2 > 0.98$), while others showed a poorer fit suggesting that in some cancer types neutral evolution is more prominent. Good model fits were seen in stomach, lung, bladder and colon consistent with multi-region sequencing studies of these cancer types that find most putative driver mutations to be truncal (Sottoriva *et al.,* 2015; Zhang *et al.,* 2014). Meanwhile tumour types that were predominantly non-neutral include renal, pancreatic and thyroid. Multi-region sequencing of renal cancers has previously shown distinct putative drivers affecting the same pathway in different regions from the same tumour (Gerlinger *et al.,* 2014) consistent with the emergence of selected subclones.

Analysis of the mutation rates shows high mutation rates in lung cancers and melanoma consistent with known carcinogens driving these diseases and producing elevated point mutation rates. The lowest mutation rates were seen in low grade

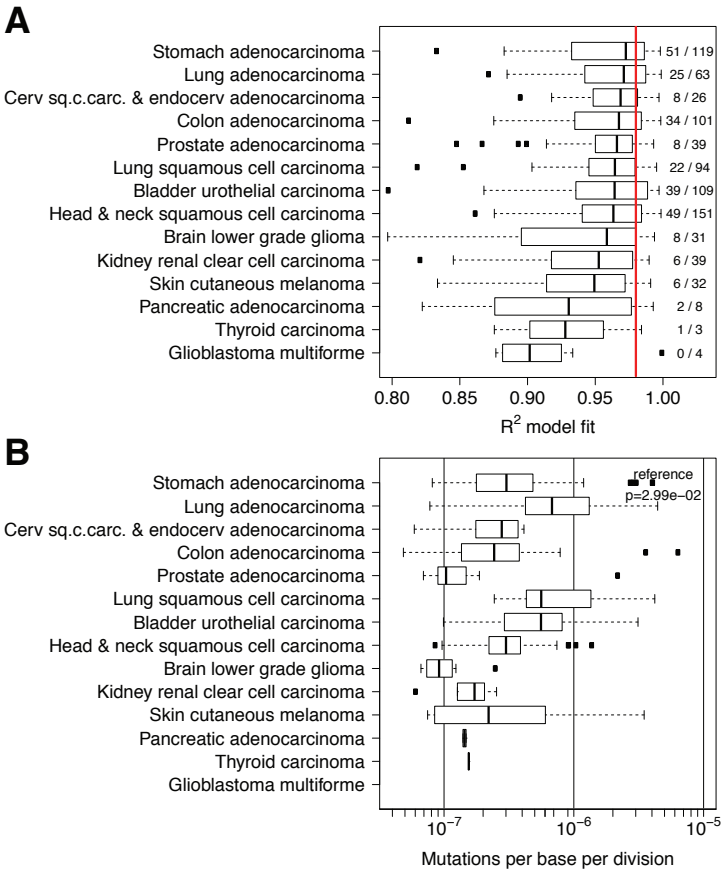

Figure 3.6: ***A*** *259/819 (31.6%) of cases from a cohort of 14 different types were consistent with neutral evolution ($R^2 > 0.98$). Stomach and lung adenocarcinoma had the highest proportion of cases that were consistent with our model, while few cases were consistent with neutral evolution in glioblastoma, thyroid and pancreatic cancer. Lung adenocarcinoma and lung squamous cell exhbited the highest mutation rates -* ***B****.*

glioma and in prostate.

3.3.5 Validation

3.3.5.1 Copy number changes

To discount the possibility that copy number changes could influence the identification of neutral evolution, the gastric cancer dataset was re-analysed following the removal of any mutations falling in non-diploid regions. This reduced the size of the cohort to 46 due to some samples having < 12 mutations remaining after filtering for copy number changes. Nevertheless, a similar proportion (74%) of cases were classified as neutral, see Figure 3.8

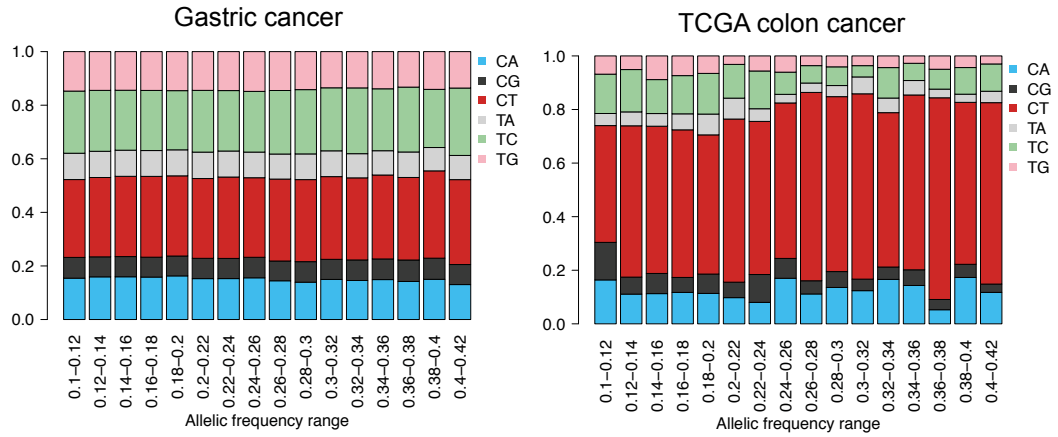

Figure 3.7: *Proportion of mutations in each mutation channel as a function their frequency in the gastric cancer cohort and the pan cancer cohort.*

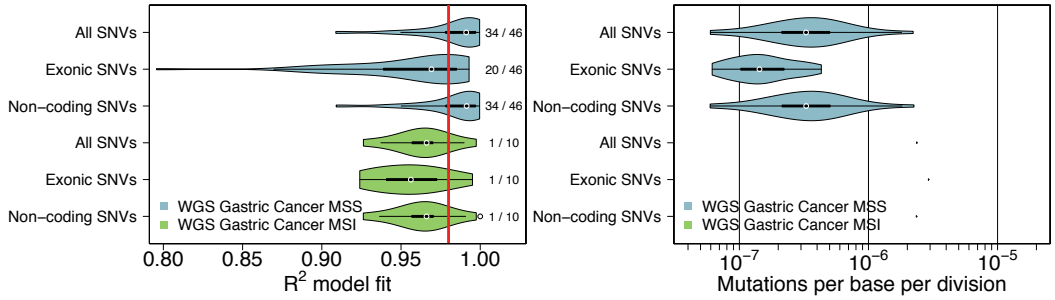

Figure 3.8: *Analysis of the proportion of gastric cancers called as neutral and their mutation rates for mutations in diploid regions only, mitigating any effects of copy number alterations.*

3.3.5.2 Sequencing errors

To discount the possibility that mutation calling errors, particularly at low frequencies could effect the inference we confirmed that the proportion of mutation types across the frequency range was consistent. Reasoning that if sequencing or PCR errors were prominent at the low frequency range then the proportion of particular types of mutations would be different from those observed at high frequency. No difference in the gastric or pan cancer cohort was observed, see Figure 3.7.

3.4 Discussion

The results in this chapter have demonstrated that cancer genome sequencing is often dominated by a signature of neutral growth, characterised by an abundance of

low frequency mutations, where the expectation from a theoretical model is that we find evermore mutations at ever lower frequencies. This was found to be true across cancer types and was robust to different cohorts and sequencing strategies. This null model of tumour evolution provides a means to quantitively test whether intra-tumour heterogeneity routinely observed can be explained by a simple neutral model rather than perhaps more complicated models involving selection of sub populations of cells.

The model described here also provides a means to estimate the mutation rate *in vivo* in human cancers, a measurement that has previously relied on *in vitro* experiments (Araten *et al.,* 2005; Rouhani *et al.,* 2016), or assuming growth rates and ages of tumours (Bozic *et al.,* 2010). In the approach presented here the mutation rate is naturally encoded in the distribution reported in the frequency spectrum of sequencing data. Further work is needed however to decouple the effect of cell death from the true per division rate. Being able to both identify neutrally evolving tumours, and measure their mutation rates provides fundamental insight into the evolutionary process, and has potentially utility in estimating the number of cells harbouring treatment resistant mutations or calculating the expected diversity of the whole tumour population. Much theoretical work has been done in this area (Iwasa *et al.,* 2006), but calculations often rely on assuming rates and evolutionary dynamics, these results show that evolutionary dynamics and their rates can be quantified on a patient by patient basis using routinely available sequencing data. Recently neutral evolution has been demonstrated to be predictive of clinical outcome in myeloma, demonstrating the potential clinical utility of evolutionary analyses of cancer genomes (Johnson *et al.,* 2017).

Although a large proportion of tumours were consistent with a neutral model of tumour evolution, a large proportion were not (70% in the pan-cancer cohort). This suggests that other evolutionary dynamics may be in play in these cases, such as the presence of functionally distinct sub-clones which lead to deviations from the null model as in Figure 3.3. In the next chapter I will explore in detail this scenario, and show how clusters of mutations that are characteristic of this type of dynamics can

elucidate further the evolutionary dynamics in these non-neutral tumours. It may also be the case that the quality of the data may be restrictive in some cases, exome sequencing in particular only reports a small subset of the mutations present in the cancer genome, when the model was applied to a whole genome cohort a larger proportion of cases were shown to be consistent with a neutral model. Other factors may also contribute to miss-classification such as low tumour content or misidentified copy number changes, the latter in particular can be challenging in exome sequencing data and estimating cellularity often shows large discordance between pathologist estimated values and values estimated from bioinformatic measures. In the next chapter I will also devote further effort to mitigate these effects.

3.5 Acknowledgements

This project was conducted in collaboration with Trevor Graham and Chris Barnes (my PhD supervisors) and Andrea Sottoriva and Benjamin Werner. We jointly developed the mathematical model I describe here. I analysed the gastric cancer cohort to call mutations, mutations for the pan-cancer cohort were kindly provided by Noemi Andor, and Andrea Sottoriva called mutations in the colon cancer cohort. I then fitted the model to these mutation calls. This chapter is a version of the work first presented in the following publication:

Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nature Genetics. 2016 Mar;48(3):238-44.

Various critiques and debates arose from this work, a list of the original critique and our replies is provided below:

- Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data. Balaparya A, De S. Nature Genetics. Springer US; 2018 Sep 11;48:1?4.

- Reply to “Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data.” Williams MJ, Werner B, Heide T, Barnes CP, Graham TA, Sottoriva A. Nature Genetics. 2018 Dec 1;50(12):1628?30.

- Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution. McDonald TO, Chakrabarti S, Michor F. Nature Genetics. 2018 Oct 17;48:1?6

- Reply to "Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution." Werner B, Williams MJ, Barnes CP, Graham TA, Sottoriva A. Nature Genetics; 2018 Oct 19;48:1?8

- Neutral tumor evolution? Tarabichi M, Martincorena I, Gerstung M, Leroi AM, Markowetz F, PCAWG Evolution and Heterogeneity Working Group, Spellman PT, Morris QD, Lingjrde OC, Wedge DC, Van Loo P. Nature Genetics. 2018 Oct 29;481:306.

- Reply to "Neutral tumor evolution?" Heide T, Zapata L, Williams MJ, Werner B, Caravagna G, Barnes CP, Graham TA, Sottoriva A. Nature Genetics.2018 Oct 23;48:1?9

- Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures Noorbakhsh J., Chuang J.H. Nature Genetics, 2017, 49:1288-1289.

- Reply: Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures Williams M.J., Werner B., Barnes C.P., Graham T.A., Sottoriva A. Nature Genetics, 2017, 49:1289-1291

- Is the evolution in tumors Darwinian or non-Darwinian? Wang H., Chen Y., Tong D., Ling S., Hu Z., Tao Y., Lu X., Wu C. National Science Review, 2018, 0:1-3, doi: 10.1093/nsr/nwx076.

- Reply: Is the evolution in tumors Darwinian or non-Darwinian? Williams M.J., Werner B., Barnes C.P., Graham T.A., Sottoriva A. National Science Review, 2018, 0:1-3, doi: 10.1093/nsr/nwx131

Chapter 4

Quantifying sub clonal selection in human cancer

4.1 Introduction

In the previous chapter the focus was on understanding the expected allele frequency distribution under a neutral evolutionary model. In this chapter I will explore what this distribution would look like when sub-populations within a tumour are selected for. I will show that using a model of tumour evolution where populations within the tumour can have different fitnesses (expressed via differential growth rates between subclones), that we can measure the relative fitness differences between different subclones and the time when subclones appear. This is achieved using a combination of theoretical models inspired by population genetics, stochastic simulations which generate synthetic data and Bayesian inference.

Clonal selection during tumour evolution naturally leads to the outgrowth of a (sub)population of cells, ultimately leading to a subclone increasing in frequency relative to its frequency when it appeared. Due to the constant accumulation of mutations during division, any clonal outgrowths should be visible from the mutations present in the subclone (assuming a high enough mutation rate). What is sometimes observed in genome sequencing of bulk cancer tissues samples is then clusters of mutations at different variant allele frequencies, which are thought to arise from such clonal expansions. One of the first and perhaps most compelling

observations (due to the ultra high depth sequencing the study deployed) of this phenomenon is in Nik-Zainal *et al.,* 2012b. In this study a single breast cancer sample was deeply sequenced to 188X depth. This data is shown in Figure 4.1, mutations clearly cluster into different groups based on their frequencies. The highest frequency peak at frequency just below 0.4 represents the clonal mutations, that is mutations present in every cell in the tumour sample, while the clusters at lower frequencies are assumed to be a consequence of clonal expansions as described above. While the original study identified 3 clusters, I will show that in fact the lowest frequency cluster that can be seen by eye is not due to a clonal expansion but rather is a consequence of all within clone neutral mutations. In other words this is the $1/f$ tail interrogated in the previous chapter that is a natural consequence of mutation accrual during population growth.

Recognising that subclonal populations present as clusters in VAF space, many bioinformatic approaches have been developed to identify such clusters (Qiao *et al.,* 2014; Roth *et al.,* 2014; Miller *et al.,* 2014; Fischer *et al.,* 2014). However, none of these methods account for the accumulation of mutations within subclones together with identification of mutational clusters. Furthermore, no work has been done in attempting to explain how and when such subclones arise, from an evolutionary perspective what fitness advantages do subclones have relative the host tumour population and when did they emerge?

In this chapter I will show how integrating both neutral and non-neutral processes within the same framework allows us to measure these evolutionary parameters directly from the VAF distribution. Before proceeding to elucidate how this information is encoded I will summarise briefly the approach. The time a clone emerges can be inferred from both the number of mutations in the first cell that gave rise to the clone and the mutation rate in the tumour, dividing the former by the later gives us the number of divisions this founder cell of the clones experiences. Then extrapolating from the time the clone emerges to the time it takes to reach a certain frequency f it is possible to estimate the selective advantage of the clone. Fortunately these parameters are all available from the VAF distribution, I

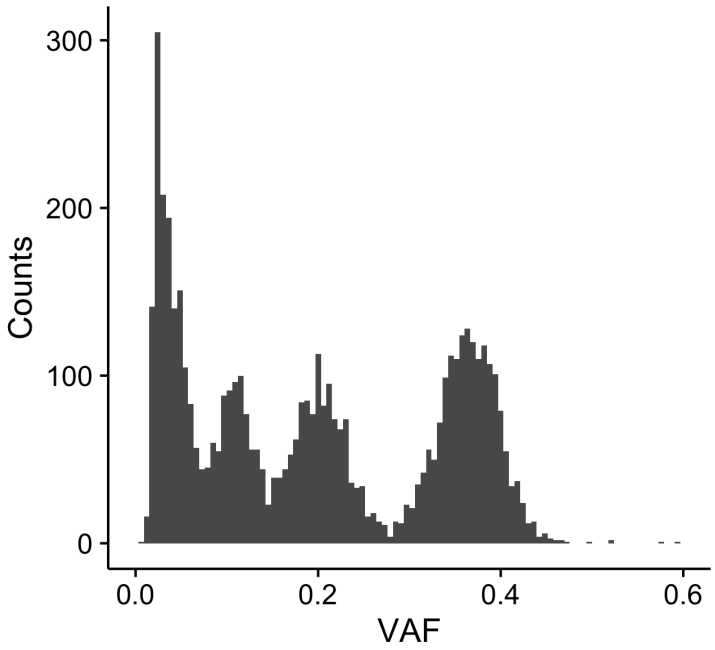

Figure 4.1: *Distribution of variant allele frequencies of a deeply sequenced breast cancer from Nik-Zainal et al*

showed in the previous chapter how the the mutation rate is encoded and here I will demonstrate that even when the VAF distribution displays subclonal clusters, it is still possible to accurately measure the mutation rate. The number of mutations in the founder cell of the clone is the number of mutations in the cluster, and the frequency of the clone is the mean frequency of the cluster. Taking all this together, it is then possible to estimate the age of a subclone and the relative fitness of the subclone compared to the host tumour population.

4.2 Simulating selected sub populations

In the previous chapter I used a discrete time stochastic simulation to demonstrate what the VAF distribution looks like under a neutral evolutionary model and then showed that when a fitter population is introduced we observe clusters of mutations in the VAF distribution, as in Figure 4.1. This simulation framework is somewhat limited for multiple reasons; at most 2 populations can be considered at one time, only relatively small fitness advantages can be considered ($1 + s < 2$) and finally, because the simulation is discrete in time mutants can only be introduced at discrete

Algorithm 3: Grow tumour via continuous time birth-death process

input : mutation rate - μ
maximum population size - N_{max}
number of subclones
time subclones introduced $t_{subclone}$
fitness advantage of subclones - s_i
birth rates for clone i - b_i
death rates for clone i - b_i

output: Frequency of mutations in the population

start with one cell, $N = 1$

while $N < N_{max}$ **do**
  $cell_i \leftarrow$ sample random cell
  r $\leftarrow Uniform(0, b_{max} + d_{max})$
  if $r < b_i$ **then**
    cell birth event
    new cell inherits parents genotype
    # of mutations in new cells $\leftarrow$ sample from $P_o(\mu)$
    $N = N + 1$
    $t = t + \Delta t$
    if $t == t_{subclone}$ **then**
      one daughter cell forms a new subclone with different growth rates
  if $b_i < r < d_i$ **then**
    cell death event
    $N = N - 1$
    $t = t + \Delta t$
    break
  if $r < b_i + d_i < r$ **then**
    no event
    $N = N$
    $t = t + \Delta t$
    break

Algorithm 4: Generate synthetic sequencing data from simulations

input : cells and mutations assigned to each cell, output from *Algorithm 1*
read depth, D
ploidy, π
Population size N_{max}
Detection limit, d (often use $d = 5/D$)
Over dispersion parameter, ρ
cellularity, c

output: Depth (D_i), read counts (R_i) and VAFs (VAF_i) for each mutation

calculate frequency of all mutations, f
for *i in mutations* **do**
  Sample depth: $D_i \sim Bo(n = N_{max}, p = D/N_{max})$
  Adjust frequency: $f_i = f_i \times c/\pi$
  if $f_i < d$ **then**
    remove mutation i
  Sample read counts: R_i $BetaBin(n = D_i, p = f_i, \rho)$
  Calculate VAF: $VAF_i = R_i/D_i$

generations rather than at any time or any population size. To investigate these non-neutral dynamics further the simulation framework from the previous chapter was modified such that it is continuous in time and has the ability simulate any number of subclones. A kinetic Monte Carlo algorithm as described in chapter 2 was used. The simulation algorithm is described in *Algorithm 3*, the output of which is mutations assigned to individual cells. This output then undergoes a process of empirically motivated sampling such that the sequencing of cancer samples is mimicked, (see *Algorithm 4*).

This simulation strategy produces synthetic data that mimics real sequencing data that can then be used to investigate what subclones that arise due to differential fitness effects should look like in typical sequencing data. As in the example from Nik Zainal et al. shown in Figure 4.1, we observe clusters of mutations corresponding to distinct subclonal populations, see Figure 4.2. These can be seen clearly with a comparison to the case where there is an absence of subclonal populations (ie neutral evolution).

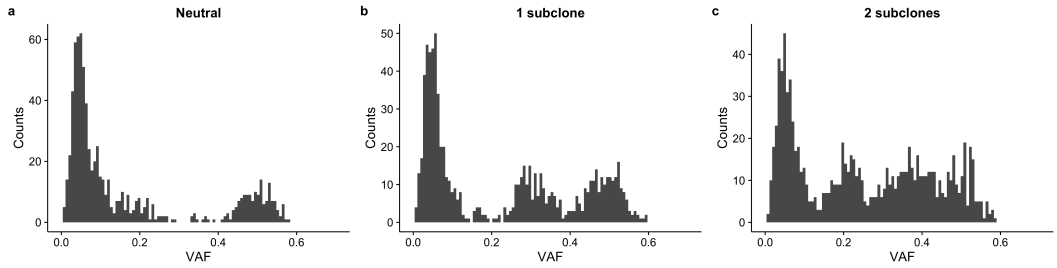

Figure 4.2: *Example of simulated sequencing data for the case of 0 subclones(neutral evolution), 1 subclone and 2 subclones*

4.3 Detecting subclonal clusters

Before proceeding to develop the framework which I will use to infer the evolutionary dynamics of non-neutrally evolving tumours, it is instructive to get a sense of what parameters of the evolutionary process lead to observing mutational clusters and thus deviations from that predicted from a neutral model. To do this I will show how some simple extensions to the model and statistical methodology of the previous chapter can lead to more robust metrics and enable identification of parameters that produce subclonal clusters, but first I'll demonstrate how current clustering methods are inadequate for this process.

4.3.1 Dirichlet process clustering

Many methods have been developed to identify subclonal clusters, all these methods however neglect the accumulation of within sub-clone mutations. This violates some important assumptions of these approaches, firstly that mutations with the same frequency are in the same sub-population and secondly that mutations can always be assigned to a particular sub-population. This invariably results in over-clustering of VAF distributions. To demonstrate this, I implemented one of the most popular clustering approaches, Dirichlet Process clustering (Dunson, 2009; Nik-Zainal *et al.,* 2012a; Roth *et al.,* 2016) and applied it to simulated data from the above model. Dirichlet process clustering is a Bayesian non-parametric clustering approach where the number of clusters is inferred directly from the data rather than specified *a priori*, further details of the statistical model can be found in Appendix A. Even accounting for the possibility that the low frequency $1/f$ may be identified

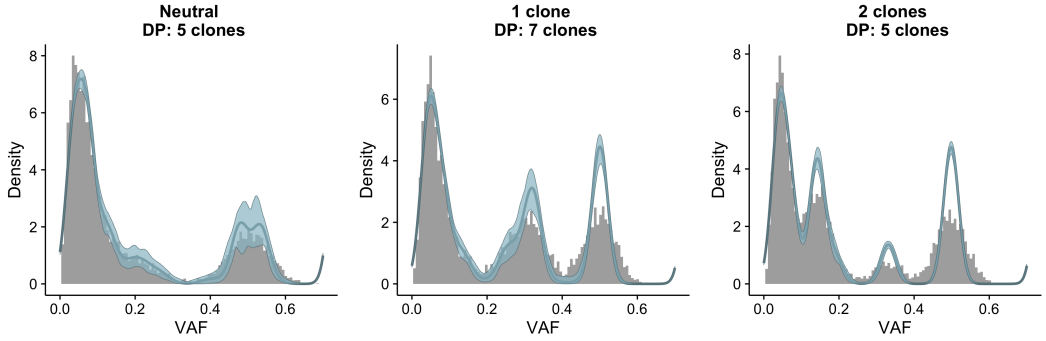

Figure 4.3: *Example fits of the dirichlet clustering algorithm for a neutral tumour, a tumour with a single subclone and a tumour with 2 subclones*

as a cluster, many more clusters are identified than are actually present, with 5 being identified in the neutral case, 7 in the 1 clone case and 5 in the 2 clone case, Figure 4.3. This is because these mutations are not drawn from a distribution with a single true underlying frequency as is assumed in this clustering approach but rather the true underlying frequencies of these mutations are genuinely different and cannot be assigned with a single underlying cluster. Applying Dirichlet clustering to 20 neutrally simulated tumours and 20 tumours with a single subclone finds that the average number of inferred subclones is 5.05 and 5.6 respectively, see Figure 4.4. For these methods to be improved upon, knowledge of the neutral growth processes needs to be integrated within these statistical frameworks, however a simple method to comparing the null distribution to data as described in the next section performs well when the aim is simply identifying deviations from the neutral distribution.

Due to limitations in sequencing technologies, not all subpopulations of cells that have fitness advantages will produce subclonal clusters in the VAF distribution. In 100X sequencing, mutations below a VAF of 5% are challenging to detect (Cibulskis *et al.,* 2013), thus often it is impractical to observe populations that are smaller than 10% of the tumour. Furthermore observing a cluster of mutations relies on a clone having accumulated a sufficient number of passenger mutations to produce clusters in the VAF distribution. Observing such mutational clusters will then necessarily depend on the evolutionary parameters in the tumour, principally the mutation rate, the relative growth rate of the subclone and the time the subclone emerges. With this in mind, I simulated many tumours with different parameters

and could then identify those parameters that lead to subclonal clusters and deviations from the neutral model. The approach taken here is similar to the previous chapter, where large deviations from the null neutral model are taken as evidence of subclonal clusters. In an attempt to increase the sensitivity of detecting such deviations I developed and evaluated the performance of 3 additional test statistics.

4.3.2 Metrics for detecting deviations from neutrality

First, recapping the null neutral model from the previous chapter, the model predicts that the cumulative number of mutations, $M(f)$ with a frequency, f is given by,

$$M(f) = \frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{max}}\right) \tag{4.1}$$

This equation can be transformed into an equation that is invariant to the mutation rate and death rate by dividing equation (4.1) by the maximum of $M(f)$ which occurs when $f = f_{min}$, that is the largest value of $M(f)$ occurs when f is small and given that we constrain the range of frequencies with which we integrate over, we know the smallest value of f. This gives us a normalized version of $M(f)$ which I

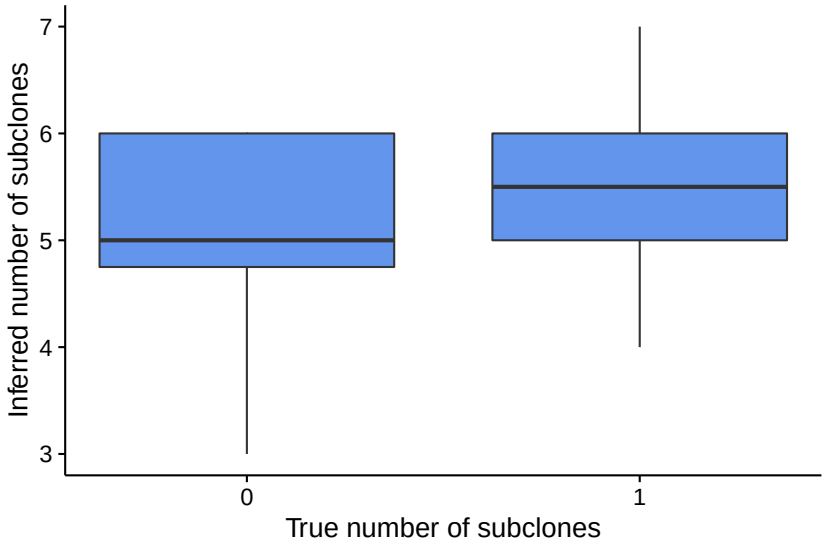

Figure 4.4: *The number of clusters identified by dirichlet clustering is always greater than the true number of clusters.*

will refer to as $\overline{M}(f)$,

$$\overline{M}(f) = \frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{max}}\right) \Big/ \frac{\mu}{\beta}\left(\frac{1}{f_{min}} - \frac{1}{f_{max}}\right)$$
$$\overline{M}(f) = \left(\frac{1}{f} - \frac{1}{f_{max}}\right) \Big/ \left(\frac{1}{f_{min}} - \frac{1}{f_{max}}\right) \tag{4.2}$$

By normalising any dataset so that the maximum value of $M(f)$ is equal to 1, any dataset can be compared with equation (4.2). I tested 3 test statistics based on this formulation of the model.

1. The Kolomogorov distance between the data and $\overline{M}(f)$, D_k
2. The area between the data and $\overline{M}(f)$, A
3. The euclidean distance between the data and $\overline{M}(f)$, d

I first evaluated the performance of these test statistics along with the R^2 test statistic from a linear model fit with equation (4.1) as described in the previous chapter. To do this, I ran a large number of neutral simulations and non-neutral simulations and then evaluated how different the distribution of the metrics were between neutral and non-neutral cases. The distributions were significantly different for all these metrics, see Figure 4.5, showing that detecting deviations from the neutral model enables identifying non-neutral tumours. Additionally, by framing the problem as a classification problem between classifying a simulated cancer as neutral or non-neutral it is possible to evaluate the performance using receiver operator characteristic curves. This showed unsurprisingly, that the ability to identify cancers with selected subpopulation depends on the size of the selected subclones, with subclones in the centre of the distribution causing larger deviations and hence were easier to detect, see Figure 4.6. This analysis indicated that the area test statistic had the highest AUC but the difference was minimal.

4.3.3 Evolutionary parameters of non-neutral tumours

Having shown that the area test statistic is the most performant, it was then used to characterise the range of evolutionary parameters that would lead to detectable

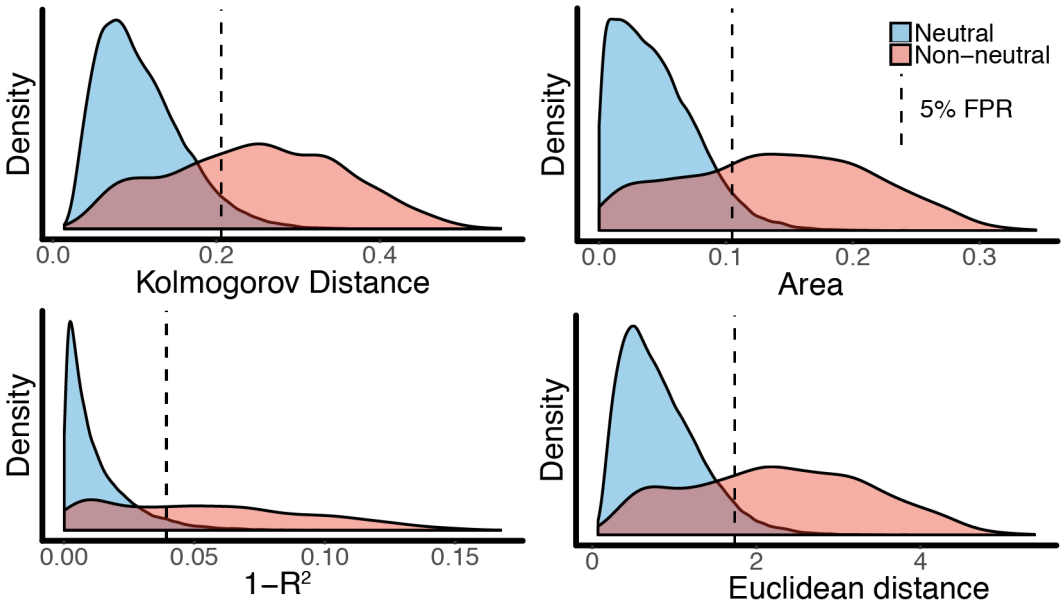

Figure 4.5: *Distribution of the 4 metrics for neutral and non-neutral tumours.*

subclonal mutational clusters. A large number of simulations with different mutation rates, different growth advantages and different times subclones emerge were generated, these simulations were than sampled using algorithm 4 to produce synthetic datasets equivalent to 100X sequencing. This showed that subclones must emerge early and have large fitness advantages to be observable in typical sequencing data, see Figure 4.7. Futhermore, if the subclone becomes dominant (greater than 90% frequency) then the tumour will appear neutral once again due to the VAF distribution then reporting on all neutral mutations within the subclone.

We see from this analysis that it is neutral passengers that are most informative for measuring evolution in the cancer genome as it is the neutral passenger mutations present in the founder cell of a subclone that will expand to a higher frequency and present as mutational clusters, while passenger mutations that accrue in cells during growth cause the characteristic $1/f$ tail that can be used to measure the mutation rate. This hitchhiking phenomenon in asexual populations is crucial in these approaches to quantify evolution in cancer (Fay & Wu, 2000; Gillespie, 2000)

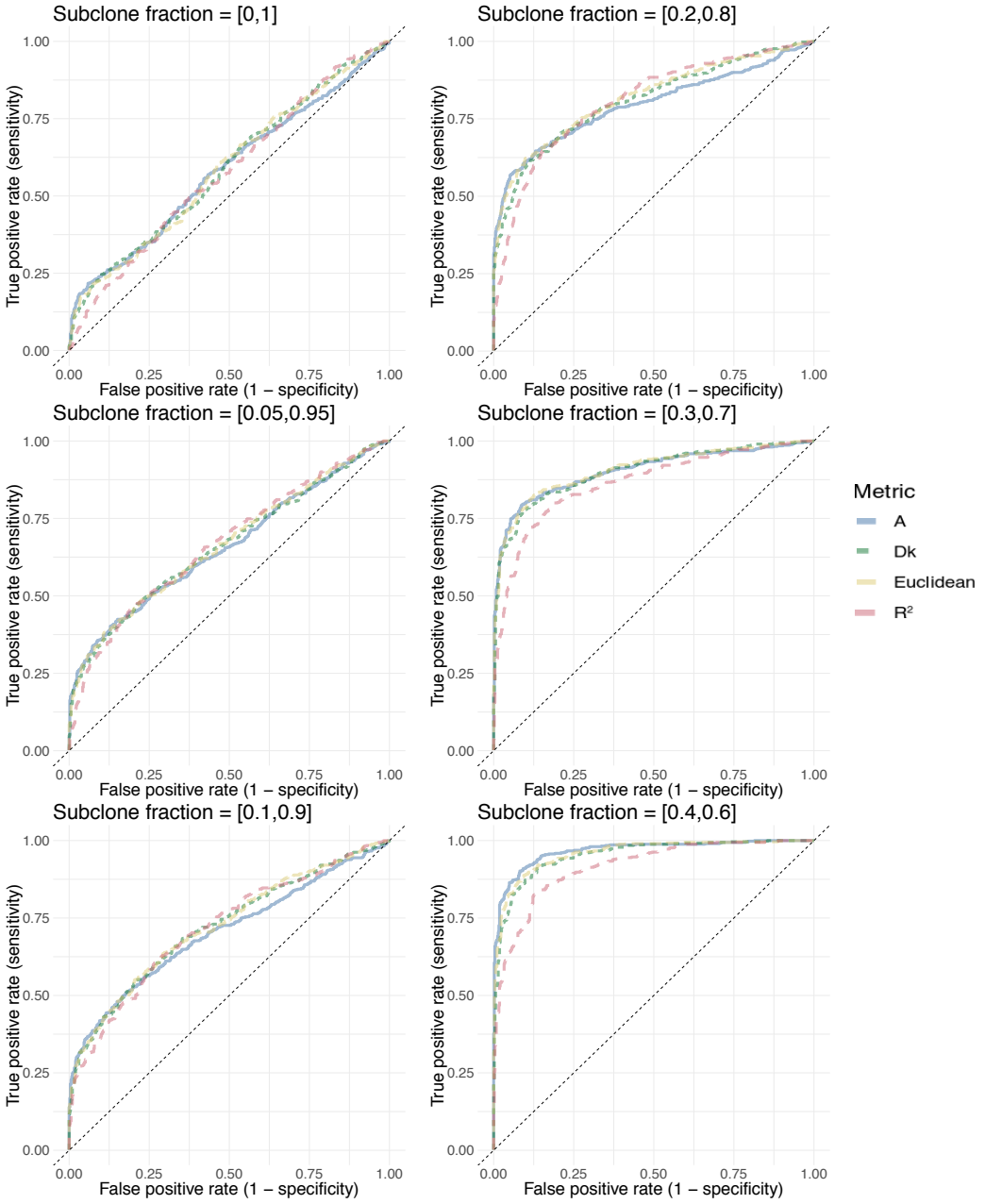

Figure 4.6: *ROC curves for all 4 metrics with different subclone sizes. When the subclones were centred toward the middle of the VAF distribution the metrics performed best.*

4.4 Mutational clusters encode evolutionary dynamics

Having demonstrated that selection of subclones within growing tumours produces subclonal clusters in the frequency distribution, and explored what evolutionary pa-

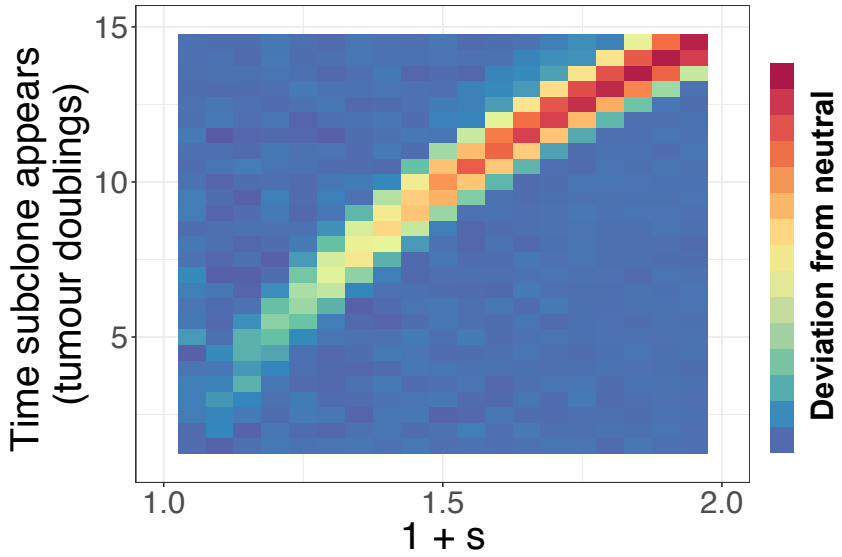

Figure 4.7: *Regions where we have detectable deviations from neutrality*

rameters lead to this observation in typical sequencing data, next I wanted to try and measure these values directly from the data. In this section I'll show how we can use passenger mutations in the subclonal cluster and passenger mutations from the $1/f$ tail to measure the fitness advantage of subclones and the time they appear. To do this I'll develop a mathematical model that describes the accumulation of mutations in a growing tumour followed by the expansion of a single cell in the tumour, and demonstrate how this can be used to infer the selective advantage and emergence time of subclones from properties measurable in the VAF distribution. In summary the mutation rate can be inferred from the low frequency $1/f$ tail, this together with the number of mutations in the cluster tells us the age of a subclone, as the mutation rate allows us to calibrate how long it should have taken for a single cell to accumulate the number of mutations we see in the cluster. Finally, the frequency of the subclone cluster tells us how much the subclone has expanded in the population allowing us to estimate the relative growth rate.

We begin at time $t_0 = 0$ with a single cell carrying a set of M_c, clonal mutations. This single cell is the most recent common ancestor of the sampled tumour. I will assume that the tumour grows exponentially growth and will measure time in units of tumour doublings. Therefore at each tumour doubling, a cell will acquire μ new mutations, which is the mutation rate per tumour doubling. This is equivalent to the

effective mutation rate discussed in the previous chapter. If a random cell is chosen from the tumour at some time t_1 it will have M_{sc} additional subclonal mutations which will be given by the product of the mutation rate, μ and the number of cell divisions, Γ_1,

$$M_{sc} = \mu\Gamma_1. \tag{4.3}$$

As we're interested in time rather than the number of divisions we first need to convert Γ_1 to the time, t_1. If we were considering discrete generations these would be equivalent, in the more realistic case of overlapping generations this is however not true. Let's consider $N_i(t)$ as the total number of cells that have completed i divisions at time t, we can then write a set of differential equations where $N_i(t)$ increases or decreases based on the birth and death rates b and d.

$$\begin{aligned}\frac{dN_0(t)}{dt} &= -(b+d)N_0(t)\\ \frac{dN_i(t)}{dt} &= -(b+d)N_i(t) + 2bN_{i-1}(t)\end{aligned} \tag{4.4}$$

Intuitively we can think of each N_i being a compartment, and b and d being the rates with which cells move from one compartment to another or are lost from the system. Cells can be lost from a compartment via death or via a birth event, where for each birth event in compartment $i-1$, 2 cells are gained in compartment i. Equations (4.4) have the following solution, given the initial condition $N_0(0) = 1$.

$$\begin{aligned}N_0(t) &= e^{-(b+d)t}\\ N_i(t) &= \frac{(2bt)^i}{i!}N_0(t)\end{aligned} \tag{4.5}$$

The pdf of this system can easily be found by recognising that the total population grows as $e^{(b-d)t}$, and dividing equation (4.5) by this expression.

$$P_i(t) = \frac{(2bt)^i}{i!}e^{-2bt} \tag{4.6}$$

This is a poisson distribution with mean $2bt$. Figure 4.8 shows the distribution of

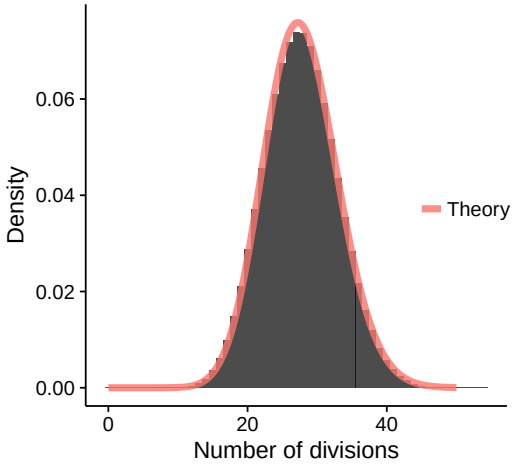

Figure 4.8: *Distribution of number of divisions for a tumour growing to size* 2^{20}. *Red line is the theoretical distribution predicted by* (4.6)

the number of divisions from a stochastic simulation where a tumour is grown to size 2^{20} along with the theoretical predictions from equation (4.6). If t is in units of tumour doublings then $b = log(2)$, we can write the expected mean number of divisions, Γ_1 experienced by a cell after time t_1 as:

$$\Gamma_1 = 2\log(2)t_1 \tag{4.7}$$

Therefore in our framework where we measure the number of divisions experienced by the founder cell of a subclone, we can measure the most probable time (in tumour doublings) that the subclone emerged. Returning to equation (4.3), we can now write down an expression that relates the number of mutations present in a cell picked at random from an exponentially growing population at time t_1.

$$M_{sc} = \mu \times 2\log(2)t_1 \tag{4.8}$$

Thus equation (4.8), solves the first part of our problem, how to measure the time a subclone emerges, next I'll move onto the second problem, how to measure the relative fitness of the subclone.

With the knowledge of when a subclone emerges, we can then attempt to quantify its fitness advantage. The VAF distribution provides one further piece of useful

information; the mean VAF of the cluster tells us the frequency of the subclone in the tumour population. Assuming the subclone starts out as a single cell in the tumour mass, we then know the initial frequency f_i of the subclone and the frequency of the subclone at sampling time t_{end}, which we will call f_{sc}. How fast the subclone expands from f_i to f_{sc} then informs us about the relative growth rates of the host tumour population and the subclonal population, or in other words the relative fitness.

Defining the fitness advantage of the subclone as the ratio of net growth rates we have:

$$1+s=\frac{\lambda_{sc}}{\lambda_{host}} \tag{4.9}$$

The subclone will then grow with a rate $(1+s)\lambda_{host}$. With this we can write down how the subclone frequency (f_{sc}) changes over time, which is the ratio of subclone population size vs total tumour population size.

$$f_{sc}(t_{end})=\frac{e^{\lambda(1+s)(t_{end}-t_1)}}{e^{\lambda(1+s)(t_{end}-t_1)}+e^{\lambda t_{end}}} \tag{4.10}$$

We can solve f_{sc} for s (by recognising that equation (4.10) is of the form $y=x/(x+c)$), which gives:

$$s=\frac{\log(\frac{f_{sc}}{1-f_{sc}})+\lambda t_1}{\lambda(t_{end}-t_1)} \tag{4.11}$$

Here I have demonstrated that we can infer t_1 using equation (4.8) and also that we can measure f_{sc}, unfortunately however the VAF distribution contains no information on t_{end}. However we can estimate t_{end} from estimating the population size of the tumour, which for a 5cm tumour will be $\approx 10^{10}$ cells. t_{end} can then be calculated via $2^{t_{end}}=(1-f)10^{10}$. In fact provided we assume a sufficiently (and realistically) large population size of $>10^9$ then the choice has minimal impact.

Taken together the VAF distribution allows us to measure the mutation rate of the tumour (μ), the frequency of a subclone (f_{sc}) and the number of mutations present in the first cell of a subclone (M_{sc}) and, using equations (4.3) and (4.11), infer the time the subclone emerges and the relative fitness advantage of the subclone

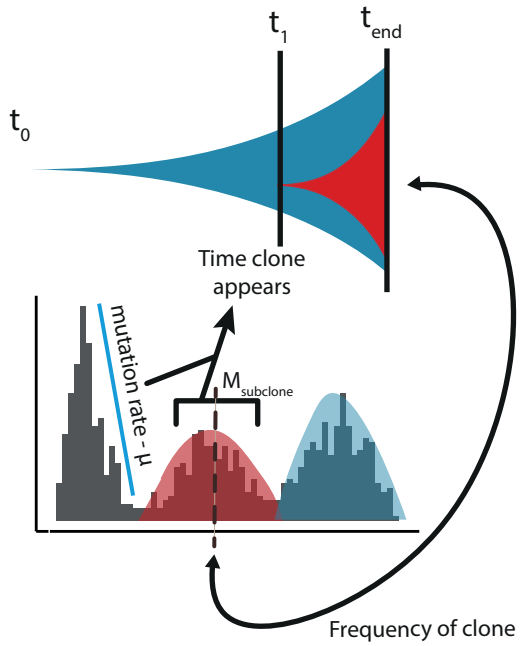

Figure 4.9: *Summary of how the VAF distribution encodes information on the evolutionary dynamics.*

compared to the host tumour population. A summary of how these measurements are acquired from the VAF distribution is provided in Figure 4.9.

4.4.1 Multiple subclones

The above model only applies to the case of a single subclone, some cancers may however have multiple subclonal populations. In the case where we have multiple subclones, these can be nested, ie one subclone emerges from within another or independent, see Figure 4.10 for an illustration.

First of all I'll consider the simpler case where clones are not nested, ie two clones arise independently within the host population. We will then have 2 equations describing how the frequency of the 2 clones increases in time.

$$f_1(t_{end}) = \frac{e^{\lambda s_1(t_{end}-t_1)}e^{-\lambda t_1}}{e^{\lambda s_1(t_{end}-t_1)}e^{-\lambda t_1} + e^{\lambda s_2(t_{end}-t_2)}e^{-\lambda t_2} + 1} \tag{4.12}$$

$$f_2(t_{end}) = \frac{e^{\lambda s_2(t_{end}-t_2)}e^{-\lambda t_2}}{e^{\lambda s_1(t_{end}-t_1)}e^{-\lambda t_1} + e^{\lambda s_2(t_{end}-t_2)}e^{-\lambda t_2} + 1} \tag{4.13}$$

Figure 4.10: *A case of nested subclones is shown on the left, independent subclones are shown on the right.*

We then arrive at the following for s_1 and s_2:

$$s_1 = \frac{\log(\frac{f_1}{1-f_1-f_2}) + \lambda t_1}{\lambda(t_{end} - t_1)} \tag{4.14}$$

$$s_2 = \frac{\log(\frac{f_2}{1-f_1-f_2}) + \lambda t_2}{\lambda(t_{end} - t_2)} \tag{4.15}$$

In the case of nested subclones, one subclone will grow inside the other thereby increasing the frequency of the major subclone. We define subclone 1 as the major subclone ($t_1 < t_2$) with subclone 2 growing inside, this also necessitates ($s_1 < s_2$) . The frequency of subclone 1 will therefore be given by:

$$f_1(t_{end}) = \frac{N_1(t_{end} - t_1) + N_2(t_{end} - t_2)}{N_1(t_{end} - t_1) + N_2(t_{end} - t_2) + N_H(t_{end})} \tag{4.16}$$

Proceeding as before we get:

$$s_1 = \frac{\log(\frac{f_1-f_2}{1-f_1}) + \lambda t_1}{\lambda(t_{end} - t_1)} \tag{4.17}$$

$$s_2 = \frac{\log(\frac{f_2}{1-f_1}) + \lambda t_2}{\lambda(t_{end} - t_2)} \tag{4.18}$$

We also have a modified equation for the time the subclones emerged.

$$M_1 = \mu\Gamma_1 \tag{4.19}$$

$$M_2 = \mu\Gamma_2 \tag{4.20}$$

Here Γ_2 is the number of divisions between t_1 (time subclone 1 appears) and t_2 (time subclone 2 appears). Meanwhile as subclone 1 is growing faster by a factor $1+s_1$, converting the number of divisions requires including this factor for the second subclone.

$$\Gamma_2 = t_1 + (1+s_1) \times 2\log(2)t_2 \tag{4.21}$$

While for t_1, we have as before:

$$\Gamma_1 = 2\log(2)t_1 \tag{4.22}$$

4.5 Statistical inference to measure s and t_1

To be able to measure s and t_1 in sequencing data from human cancers we need some way to extract all relevant parameters from the VAF distribution. One approach would be to use some kind of clustering approach as is commonly employed to this type of data, however these approaches are often prone to over clustering as was shown at the beginning of this chapter. An alternative approach that I take here is to use the simulation framework together with Approximate Bayesian Computation to fit the data and extract the relevant parameters. This has the benefit of directly modelling all within clone passenger mutations that can skew traditional clustering approaches and it is also straightforward to directly model the sequencing noise. Furthermore, using a stochastic simulation to fit the data means we can capture any stochastic effects in tumour growth, this is potentially important as tumour formation is a single realisation of a stochastic process, fitting some distribution that reflects average dynamics may miss out on some of these effects. Specifically, the algorithm used was Approximate Bayesian Computation Sequential Monte Carlo

(ABC SMC) with model selection, which is described in detail in Chapter 2. Using the model selection algorithm allows us to remain agnostic about the clonal structure of the tumour and to infer this directly from the data while simultaneously estimating the parameters governing the evolution of each subclone (s and t_1) and the parameters of the tumour as a whole (μ, M_c). I give equal prior probabilities to the number of subclones, however the more complex models with subclones are naturally penalised in this approach due to models with selected subclones having larger numbers of parameters.

ABC approaches often require a large amount of simulations to be performed to get good fits to the data. A number of choices and simplifications were made with the aim of making the simulations as computationally efficient as possible.

4.5.1 ABC SMC implementation

As our distance measure I used the euclidean distance between the cumulative distributions (unnormalized) of the target data and the simulated data. The cumulative distribution holds information on the shape of the distribution and also applies some degree of smoothing to mitigate sampling noise, making it an appropriate choice. Implementing the ABC SMC with model selection algorithm requires choosing the model perturbation kernel, the particle perturbation kernel and the prior distributions for parameters. The kernel ensures that the algorithm explores the space of both models and parameter values fully by perturbing the model and model parameters at each step. For the model perturbation kernel I used the following, where m^* is the sampled model and m is the perturbed model.

$$KM_t(m|m^*) = \begin{cases} 0.6, & \text{if } m = m^* \\ 0.4, & \text{if } m \neq m^* \end{cases} \tag{4.23}$$

I used uniform parameter perturbation kernels with limits determined from the range of parameter values from the previous population (Filippi *et al.,* 2013), for parame-

ter k, $KP_t(k|k^*) = U(k_i - \sigma, k_i + \sigma)$, where σ is given by:

$$\sigma = \frac{1}{2}(max(k)_{t-1} - min(k)_{t-1}) \tag{4.24}$$

Prior distributions and constants for all the parameters are shown in table 4.1. Finally, the ABC algorithm stops when one of the following criteria has been met.

1. $\frac{\varepsilon_t - \varepsilon_{t-1}}{\varepsilon_t} < 0.05$
2. Completed 5×10^6 simulations
3. 200 hours of computation time

I found that the type of data (WXS vs WGS) and the maximum number of clones I attempted to fit had a large effect on the computational cost and the time needed to get good fits. The above criteria were found to be sufficient for fitting WGS data with up to 2 subclones. Fitting WXS data with up to 1 subclone may for example only need 10^5 simulations to converge to a reasonable fit.

4.5.2 Computational efficiency

One of the challenges of ABC simulation based inference is being able to simulate the model efficiently enough so that potentially many millions of simulations can be performed in order to explore the parameter space fully. In an attempt to make the simulation as computationally efficient as possible I made a number of simplifications. Firstly, I neglected cell death in the model, as high cell death means a larger number of time steps are needed to reach a certain tumour size when compared to setting cell death to 0. As units of the parameters affected by cell death (the mutation rate and time) are all normalized by tumour doubling time this does not affect the inference on these parameters. Neglecting cell death does however reduce the stochasticity in the model, but despite this I found that the fitting methodology

1+s	μ	clonal mutations	t_1	N_{max}	birth rate	death rate
[1, 26.0]	[0.1, 500]	[1, 5000]	[3, 20]	10^4	log(2)	0.0

Table 4.1: *Limits on prior distributions and constant values for all parameters*

could appropriately capture the noise. Another simplification that was made was to simulate only small tumours when fitting. When fitting to the data, the aim is to infer all parameters that are encoded in the VAF distribution, tumour size is one parameter that cannot be inferred directly from the distribution and must be assumed. The only requirement in terms of population size is that the tumour grows large enough for enough mutations to accumulate prior to the expansion of any selected for subclone, the initial exploration of the parameter space indicated that subclones generally emerge early during tumour growth. Simulating large tumours with relatively small relative fitness advantages is equivalent to simulating small tumours with large fitness advantages. Opting for the latter vastly reduces the computational cost per simulation and selection coefficient values can be scaled appropriately using equation (4.11). In summary, the fitting methodology is then to use the ABC inference to measure the mutation rate, the number of clonal mutations and the number of mutations in any subclone and then to use equations (4.8) and (4.11) to construct posterior distribution for $1+s$ and t_1 for realistically large tumours.

The final simplification that was made was to include at most 2 subclones in the inference scheme as the addition of each subclone results in an ever larger parameter space to search. By eye all the data that was used appeared to have at most 2 subclonal populations. Furthermore I found that identifying 2 subclones with high confidence required ultra high-depth sequencing of $>$100X, identifying more than 2 subclones is likely therefore to require even higher depths, so restricting model to 2 subclones was deemed appropriate given the quality of data available.

4.5.3 Accurate recovery of parameters from simulated data

To confirm that the inference methodology correctly identifies the clonal composition of a tumour and the parameters of interest in what can be noisy sequencing data I first assessed the ability to accurately recover input parameters and the clonal composition from datasets generated from the model. First, I generated a virtual cohort of tumours with a single subclone, tumours were all grown to a size of 10^6 cells, with different mutation rates, different death rates, different selective advantages and subclone emergence times. The above fitting methodology was applied

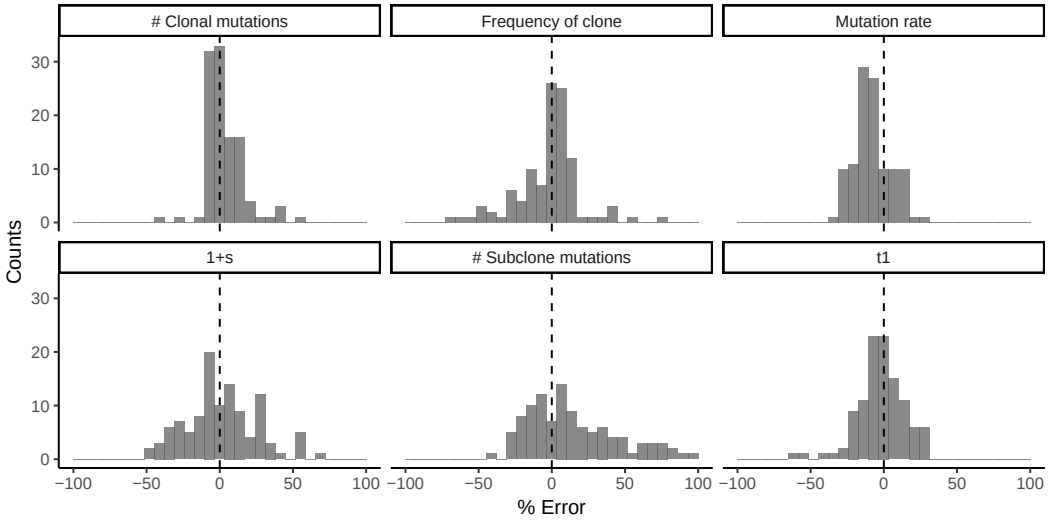

Figure 4.11: *% Error for all inferred parameters from a cohort of synthetic datasets where the ground truth is known.*

and the percentage error on inferred vs known parameter values was calculated. This showed that the inference scheme on average accurately recovers the input parameters despite some of the simplifications made, Figure 4.11.

Next, I confirmed that the the inference scheme could correctly infer the clonal structure of tumours, suspecting that the depth of coverage could in particular have an effect on the ability to resolve distinct subclones. Again I generated a virtual cohort of tumours, this time for 3 clonal architectures (neutral, 1 and 2 clones) at different depths (25X, 50X, 100X, 200X, 300X) and sequencing strategies (whole genome sequencing (WGS) vs whole exome sequencing (WXS)). 10 simulations for each of the different clonal architectures were generated and the same simulations then subjected to differing sampling procedures for a total of 300 virtual tumours at the prescribed sequencing depths and strategies. The parameters for sampling for the sequencing strategies were tuned so that the observed mutation burdens were of the order 100/exome and 10,000/whole genome at 100X depth respectively which is consistent with my observations of such data.

Figure 4.12 shows an example simulated dataset for each of the clonal architectures (neutral, 1 and 2 clones) with fitted distributions in red. This demonstrates that the inference scheme accurately fits the data and can infer the correct subclonal composition (inset shows posterior probabilities for the number of subclones). Ap-

plying the methodology to all 300 simulated datasets it was possible to accurately recover the clonal structure provided the depth of sequencing was sufficiently high. This is perhaps not surprising given that the subclonal peaks that are a readout of the clonal structure become progressively obscured as the depth decreases, shown in Figure 4.13.

4.6 Application to multiple cancers from different types

Having validated the ability to resolve evolutionary dynamics from the VAF distribution of sequenced tumours and explored the limitations of the approach, next I applied the method to a blood cancer, a lung cancer, a breast cancer, a cohort of metastases and the cohort of colon and gastric cancers from the previous chapter. The blood cancer sample was an acute myeloid leukaemia sample subjected to 300X whole genome sequencing with extensive validation of mutation calls. The mutation calls for the original study were used for the analysis (Griffith *et al.,* 2015). The lung cancer samples came from a multiregion sequencing study of lung adenocarcinoma, where multiple samples from the same tumours were subjected to high depth exome sequencing (Zhang *et al.,* 2014). The breast cancer sample is the sample discussed at the beginning of this chapter and shown in Figure 4.1. The colon cancer cohort and gastric cancer cohort are the same data used for the previous chapter.

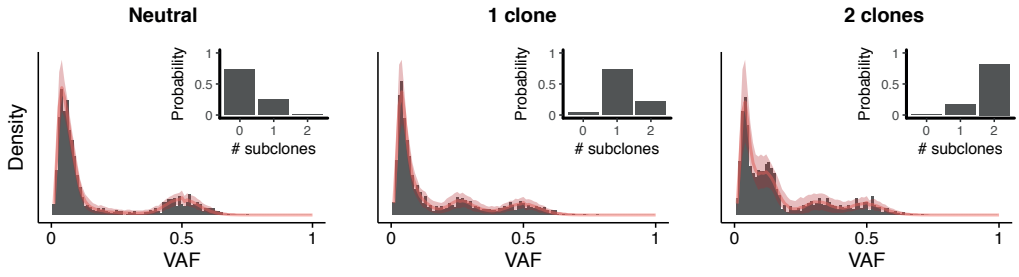

Figure 4.12: *Model fits to a neutral tumour, a tumour with 1 subclone and a tumour with 2 subclones. Red lines show the mean value and shaded red area the 95% credible interval, showing the model accurately recapitulates the data. Insets show the posterior probabilities for each of the subclones showing the inference framework accurately recovers the correct clonal structure.*

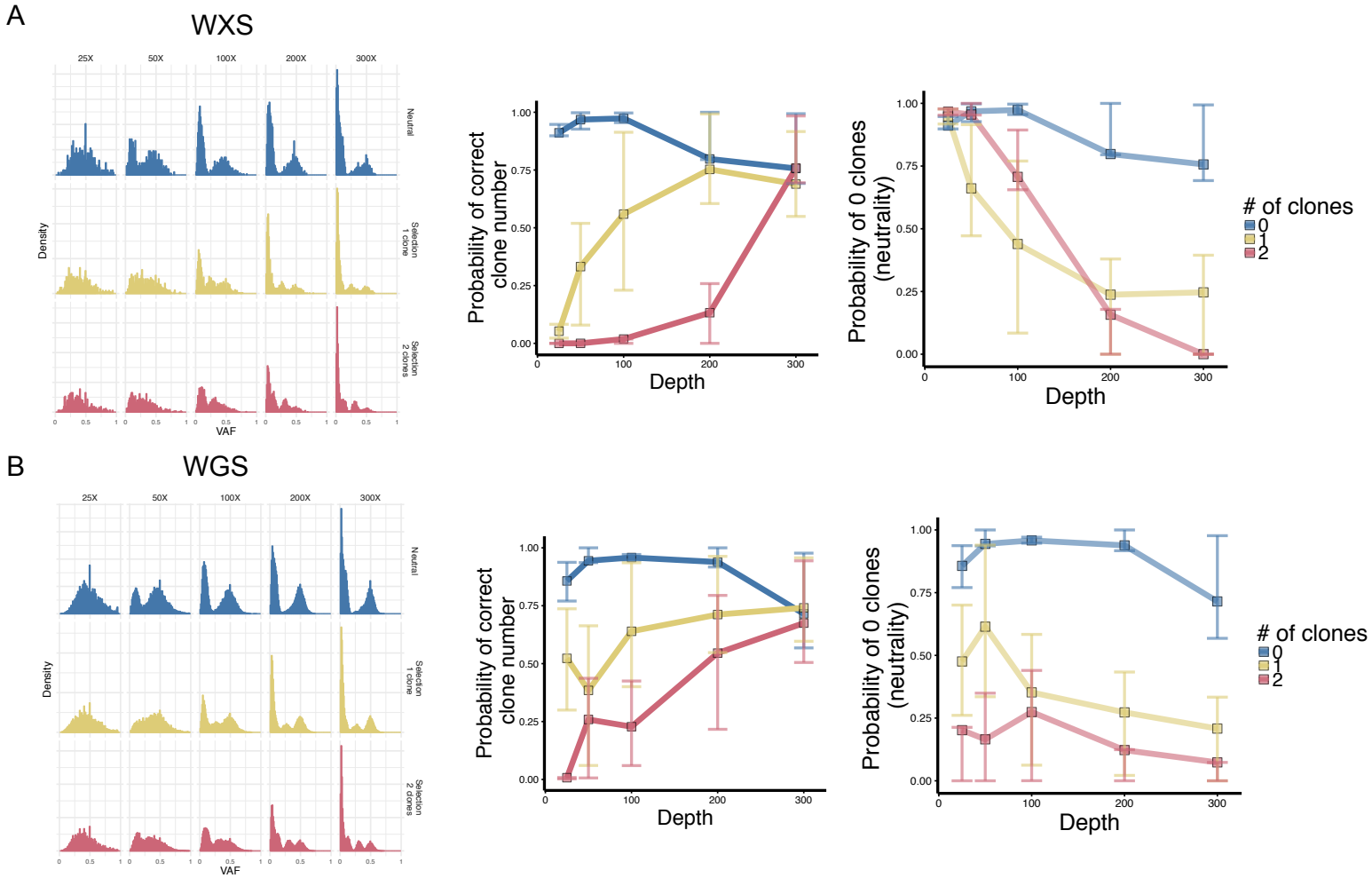

Figure 4.13: *A cohort of 30 synthetic tumours with 0, 1 and 2 subclones were generated and exposed to different sequencing strategies and depths. Left panels show how the clonal strcuture become progressively obscured as the depth decreases. Right panel shows the results of applying the statistical inference framework to these data demonstrating that high depth sequencing is required to resolve clonal structures in many cases.*

Data from the metastases study were kindly provided by the authors (Robinson *et al.,* 2017).

4.6.1 Data analysis

Mutect2 was used to call mutations in the breast cancer sample and lung cancer samples. The sequenza algorithm was used to infer allele specific copy number profiles for the breast cancer, lung cancers and gastric cancers. This allowed us to remove mutations falling in non-diploid regions of the genome, therefore mitigating the effects of copy number aberrations on the measured mutation frequencies. Sequenza also provides estimates of the cellularity of the sample, this was used to correct the VAF of mutations, such that a clonal mutation would be expected to be observed at $VAF = 0.5$. No copy number alterations were observed in the AML sample and cellularity estimates were provided in the original analysis. To remove copy number aberrations from the colon cancer cohort, paired SNP array data was used to identify non-diploid regions. For the lung cancers, colon cancers and gastric cancers samples with cellularity $< 50\%$ or where the number of mutations was less than 20 following filtering for diploid regions were removed. All other data was obtained from the original publications. For the metastases cohort the cellularity of the sample was fitted using the ABC approach. To further refine the cellularity estimates and measure the degree of dispersion in the data, both Binomial and Beta-Binomial models were fitted to the clonal cluster in the VAF distribution. Given a mutation i has a read count f_i, depth D_i then the Beta-Binomial model is given by,

$$f_i = BetaBin(n = D_i, p = 0.5, \rho). \tag{4.25}$$

$p = 0.5$ as we are fitting the clonal cluster, from fitting this model I could estimate the degree of dispersion ρ and tweak the cellularity estimates such that $p = 0.5$. Fitting was performed using Markov Chain Monte Carlo. I fitted these models to all our data and used the dispersion parameter ρ in the sampling algorithm.

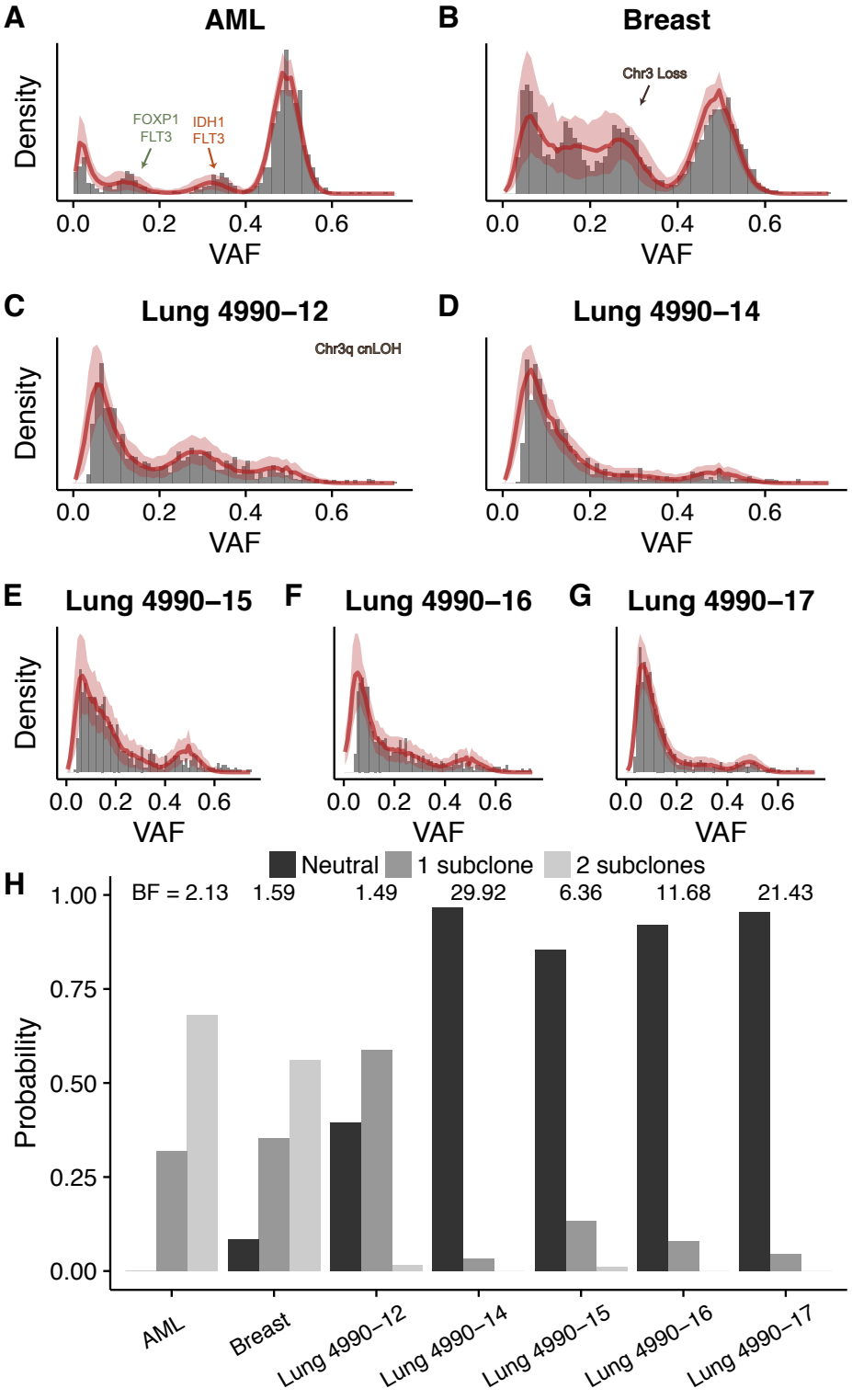

Figure 4.14: *Data (grey histograms) together with model fits, red line is mean value from ABC simulations and shaded ares is 95% credible intervals. Panel E shows the posterior probability for the different clonal structures.*

4.6.2 Results

I first applied the method to the AML, lung and breast cancer cases. Figures 4.14A-H shows the data and model fits for the most probable model. Bayesian model selection showed that the AML and breast cancer samples had evidence of 2 subclones, while 1 lung cancer sample had evidence of a subclone with the other 4 showing higher probability for the neutral model (0 subclones), Figure 4.14H. From the posterior distribution of the parameters, see Figure 4.16, subclones in these data all emerged relatively early (during the first 15 tumour doublings) and had relative fitness advantages of greater than 20%, with some as high as 100% ie a two fold increase in growth rate. Inferred mutation rates were variable, in particular the AML samples showed a mutation rate 100 fold less than the breast cancer. These observations are consistent with AML having a low mutation burden compared to breast and lung cancers (Lawrence *et al.,* 2014). For the AML case, putative subclonal drivers in *FLT3*, *IDH1* and *FOXP1*, all known drivers of carcinogenesis were found in the subclonal clusters providing a genetic mechanism for the increased fitness of these subclones. In the breast cancer case, the original study showed that one of the clusters had a subclonal deletion of chromosome 13 (Nik-Zainal *et al.,* 2012b). Meanwhile in the lung cancer case, the sample from patient 4990 that appeared to harbour a fitter subclone showed evidence of copy neutral LOH on chromosome 13, see Figure 4.17 for copy number profiles for all 5 samples.

Next I applied the model to the whole exome sequenced colon cancer cohort,

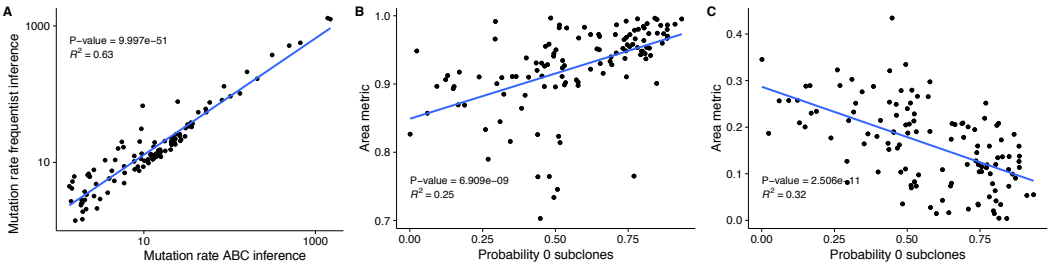

Figure 4.15: *Value from Bayesian inference compared with frequentist inference.* ***A*** *Mutation rates from the two approaches were also found to be highly correlated.* ***B*** *and* ***C*** *shows the correlation between the neutrality metrics and the posterior probabilities of different clonal structures.*

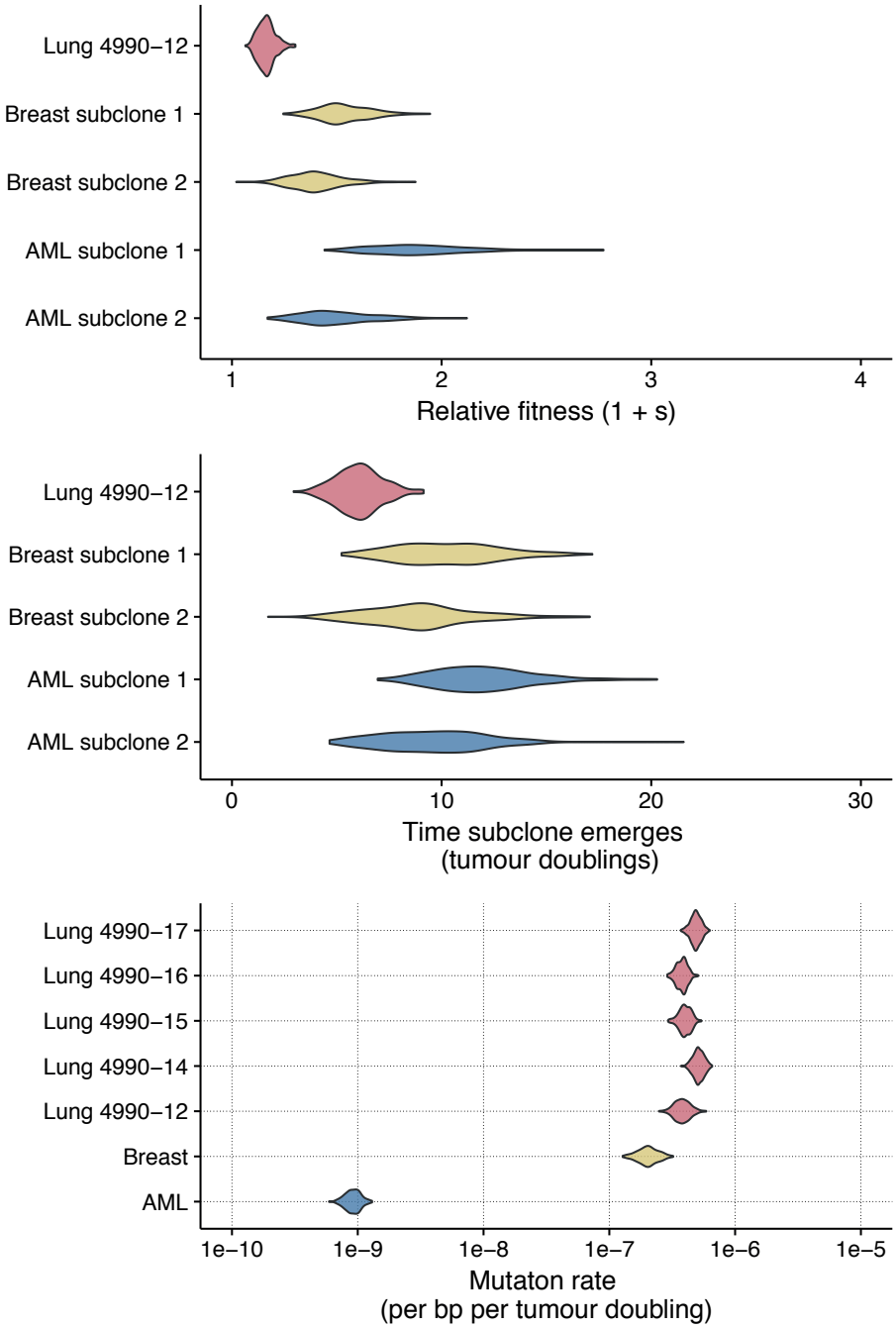

Figure 4.16: *Posterior distribution for inferred parameters for data presented in Figure 4.14*

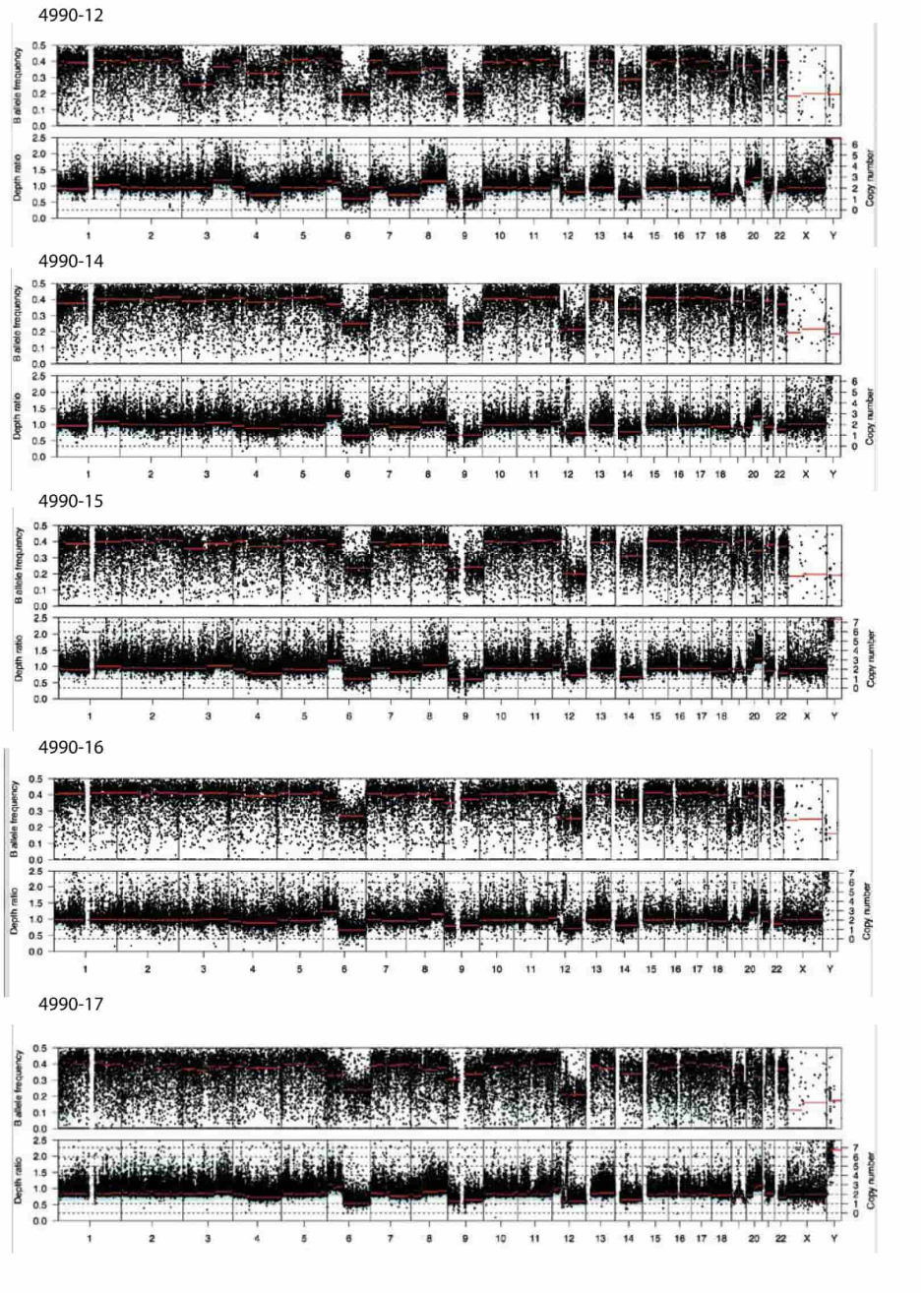

Figure 4.17: *Copy number profile of the lung adenocarcinoma samples showing sample 4990-12 appears to have a subclonal loss on Chr3.*

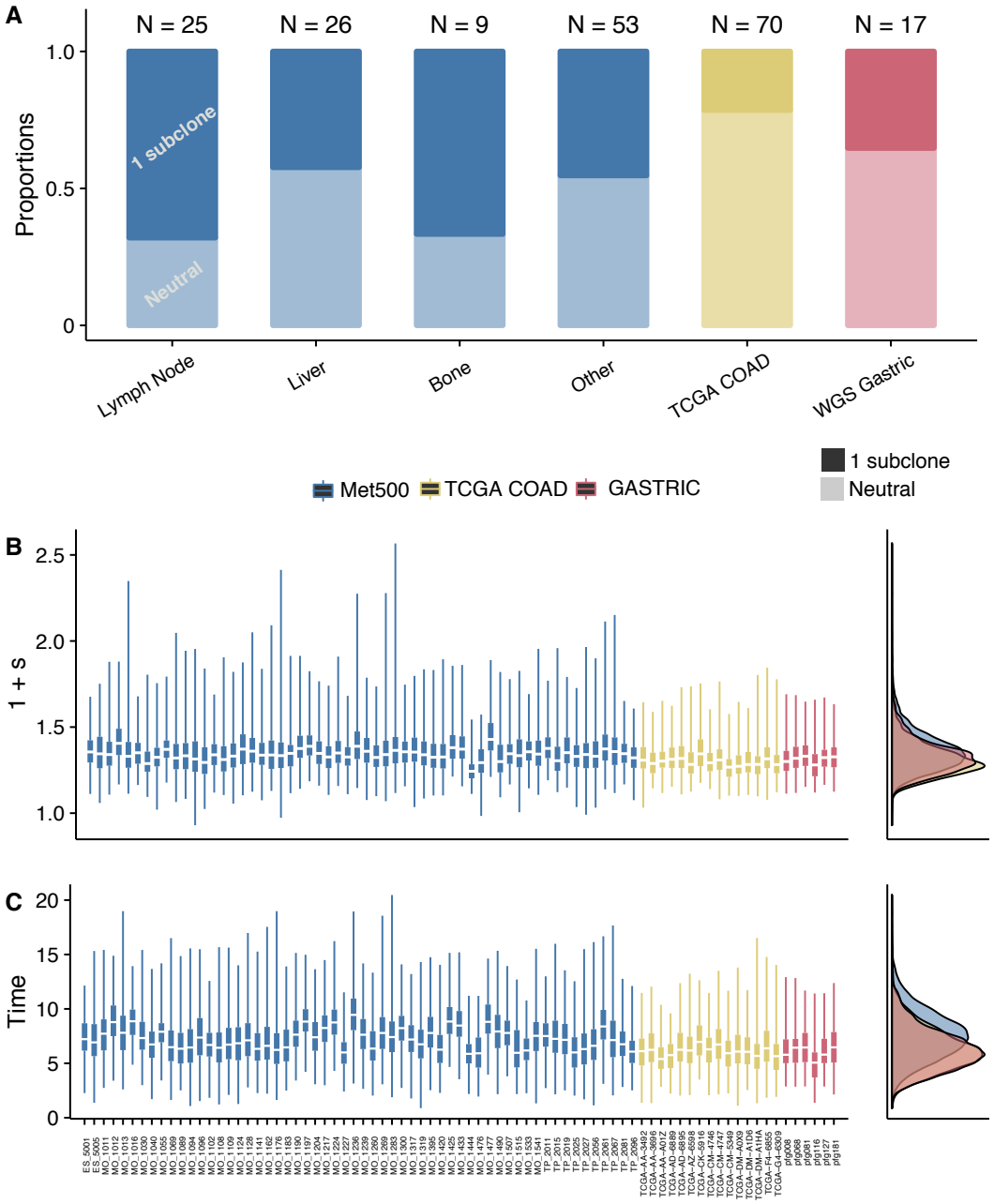

Figure 4.18: *Proportions of each type identified as neutral or non-neutral, **A**. For non-neutral cases the inferred time subclones emerged and their fitness is shown in **B** and **C***

the whole genome sequenced gastric cancer cohort and the whole exome sequenced metastases cohort. Data and model fits are shown in Figures 4.19, 4.20 and 4.21. 6/17 of the gastric cancers, 16/70 colon cancers and 58/113 of the metastasis cohort showed evidence of subclonal selection, again with subclones emerging early and

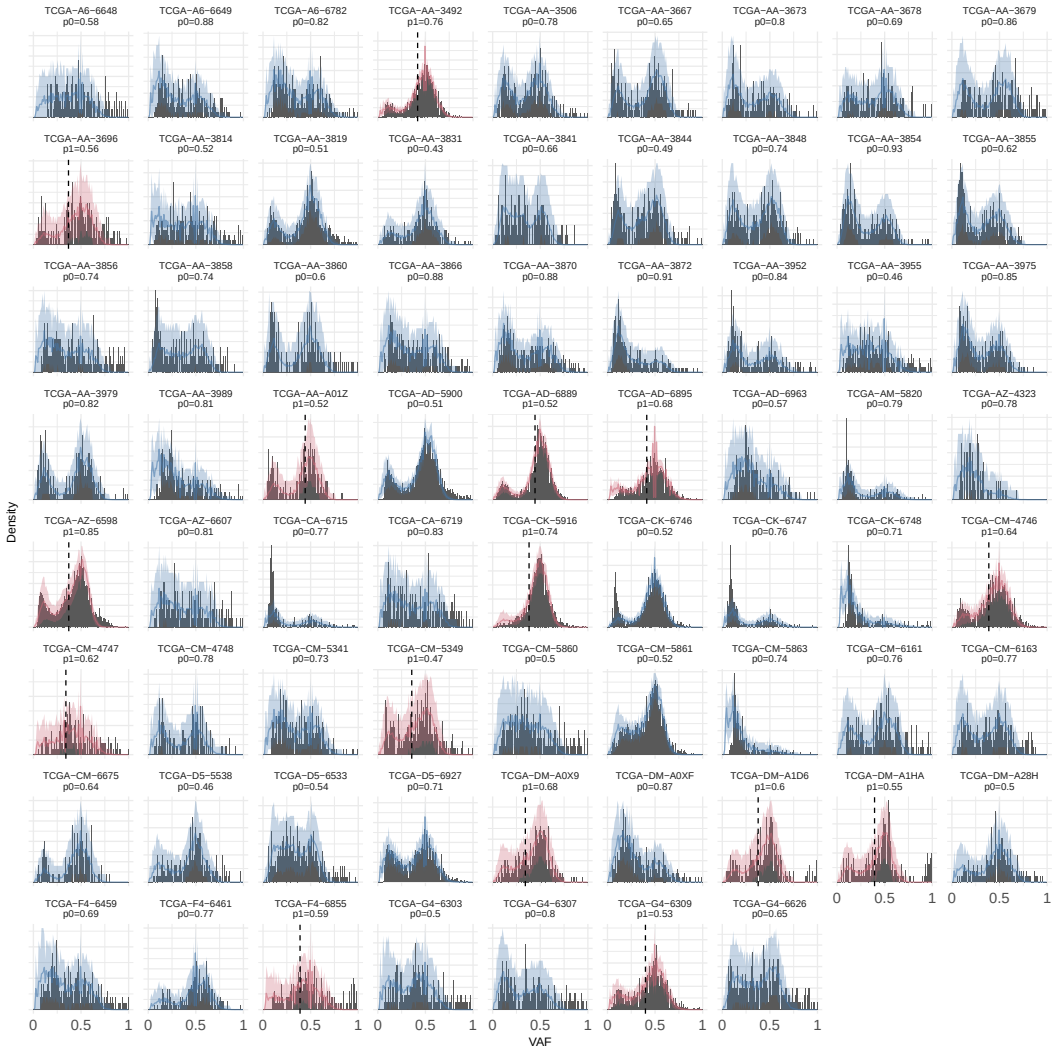

Figure 4.19: *Colon cancer model fits. Grey histograms are empirical VAF distributions, line is mean value from 500 simulations that fitted the data and shaded area is 95% interval. Blue tumours are those identified as neutral and red as those with subclonal selection. For those with a subclone, dashed line shows the mean of the subclonal cluster. Title of each panel shows sample name and probability of the assigned subclonal structure.*

having high fitness advantages 4.18. A higher proportion of the colon and gastric cancer samples were called as neutral with this method in comparison to the previous chapter. The principal reason for this is likely that using a stochastic simulation as was done here can capture deviations from the deterministic model presented in the previous chapter that are due to stochastic effects, ie neutral drift rather than selection, this phenomenon is more likely when the death rate in the tumour is high. These results suggest the possibility that tumours that are classified as non-neutral with the deterministic model but as neutral with the stochastic model have higher death rates, and also demonstrates that the emergence of subclonal clusters may not always be due to selection. Making the connection between the two approaches I found a significant correlation between the posterior model probabilities derived from the ABC approach and the R^2 and area between curves metric, while inference on the mutation rates were also highly correlated, Figure 4.15.

4.6.3 Predicting tumour evolution

Given that the effects of selection are deterministic, being able to measure the fitness advantage means we should be able to predict how the size of subclones might change over time. If one population in the tumour is growing at a faster rate than the rest of the tumour we can say with some confidence that given sufficient time and in the absence of further perturbations, that population will become the dominant clone. Mutations and drift on the other hand are stochastic process and are thus more difficult or even impossible to predict. For example imagine we have sampled a tumour at some time t_1 and found a subclonal population at a frequency f_1 within the tumour and have measured its fitness, we can then ask how long before that tumour becomes dominant and reaches a frequency f_2. This time ΔT can be expressed mathematically by the following (found from solving equation (4.10) for time),

$$\Delta T = \frac{\log(\frac{f_2}{1-f_2}) - \log(\frac{f_1}{1-f_1})}{\lambda s} \tag{4.26}$$

I implemented this idea using a simulated tumour. Here I sampled the tumour and measured the fitness and time the subclone emerged and then predicted how the

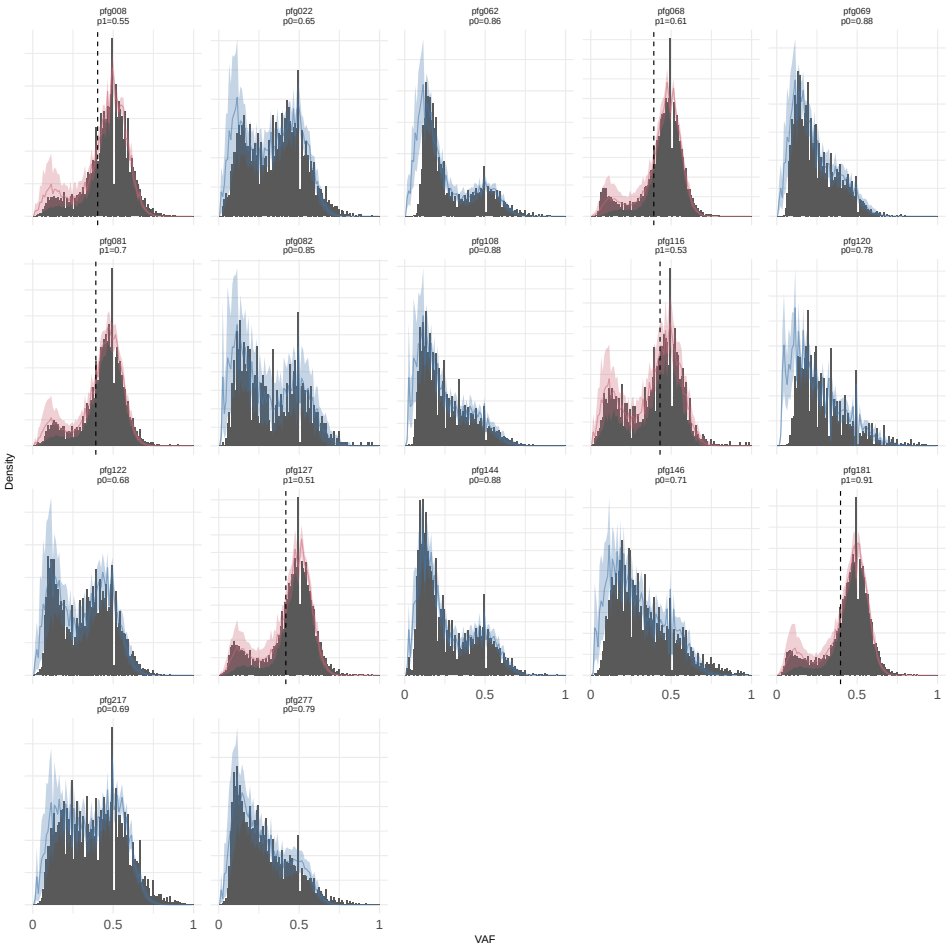

Figure 4.20: *Gastric cancer model fits. Grey histograms are empirical VAF distributions, line is mean value from 500 simulations that fitted the data and shaded area is 95% interval. Blue tumours are those identified as neutral and red as those with subclonal selection. For those with a subclone, dashed line shows the mean of the subclonal cluster. Title of each panel shows sample name and probability of the assigned subclonal structure.*

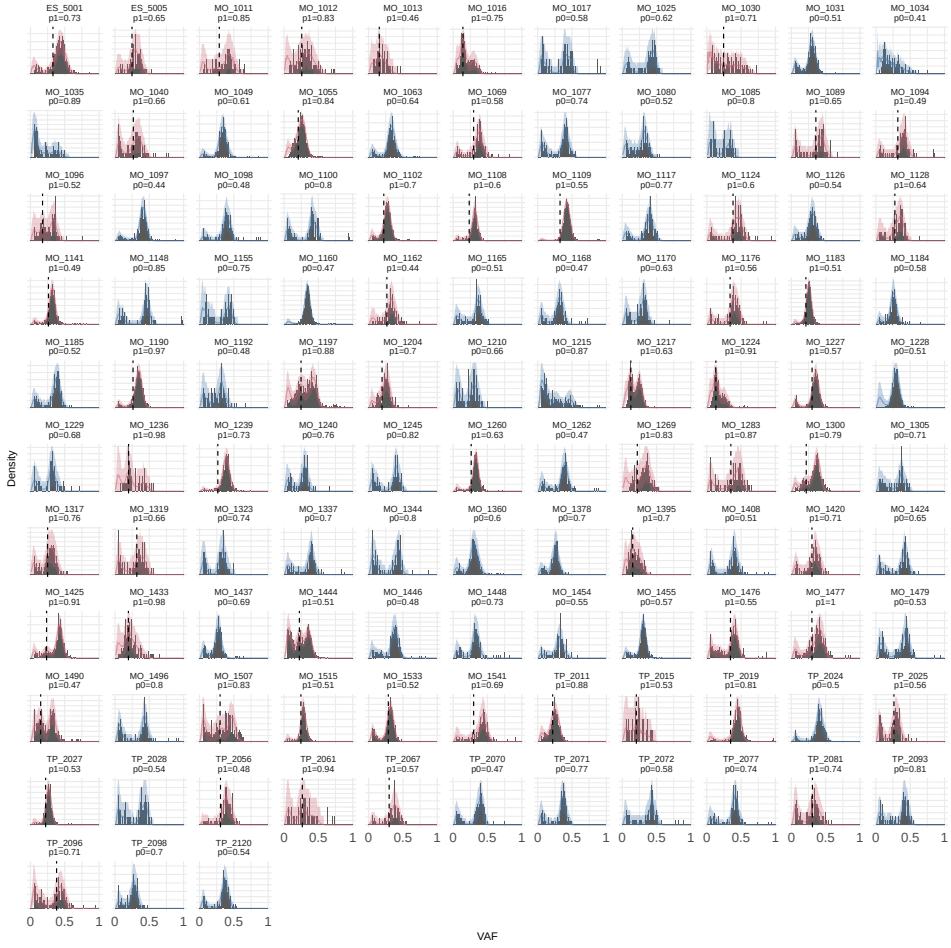

Figure 4.21: *Metastases model fits. Grey histograms are empirical VAF distributions, line is mean value from 500 simulations that fitted the data and shaded area is 95% interval. Blue tumours are those identified as neutral and red as those with subclonal selection. For those with a subclone, dashed line shows the mean of the subclonal cluster. Title of each panel shows sample name and probability of the assigned subclonal structure.*

subclone frequency would change over time. Figure 4.22B shows that the prediction and ground truth match well, Figure 4.22A shows the model fits to the tumour when it was sampled. The question then is how might these measurements be useful? One application might be in measuring the time scales of treatment resistance. Suppose a genetic screen has found a suspected treatment resistant population, measuring the fitness of the treatment resistant subpopulation could allow us to infer the length of time before this population sweeps through the population and becomes dominant. Another utility would be in implementing ecological or evolutionary inspired treatment strategies such as adaptive therapy. Here the goal is to maintain the coexistence of (potentially treatment resistant) subclones rather than to eradicate the tumour completely, and avoid selecting for the most aggressive cells (Gatenby, 2009; Enriquez-Navas *et al.,* 2016). With the kind of measurements described here, these type of treatments could be exactly optimised to based on the relative fitnesses of competing subclones.

4.7 Discussion

In this chapter I've shown how the frequency distribution of mutations from bulk sequencing data can be used to measure the evolutionary dynamics of subclones. Using a mathematical model of cancer growth I confirmed that subclonal clusters of mutations in the frequency distribution do arise due to selection and furthermore that properties of these clusters can be used to measure the time they emerge and their relative fitness advantage. This was validated using synthetic data generated from the model and then applied to sequencing data from human cancers. Importantly, the model demonstrates that a low frequency peak, or $1/f$ tail is a pervasive feature of bulk cancer genome sequencing due to the constant accumulation of mutations within all clones as the tumour grows.

The relative fitness advantages I measured at first appear strikingly high, particularly in relation to classical population genetics. The high values reported here are however, perhaps not unprecedented. Mutant *KRAS* and *APC* stem cells in the mouse intestine have been shown to have a 2-4 fold increase in the probability of

fixing in the stem cell niche compared to wild types (Vermeulen *et al.,* 2013), while *TP53* mutants in mouse epidermis show 10% bias toward self renewal (Klein *et al.,* 2010). Recently, TP53 mutants in cultured embryonic stem cells were measured to have double the growth rate of their wild type counterparts (Merkle *et al.,* 2017). Furthermore the values reported here at the extreme end of the values measured in experimental evolution systems where most values are measured to be small (Lenski & Travisano, 1994; Kassen & Bataillon, 2006). Thus, it may be the case that current sequencing standards only allow the detection of highly selected clones, and that many so called *mini-drivers* are present in a tumour but have minimal effect on

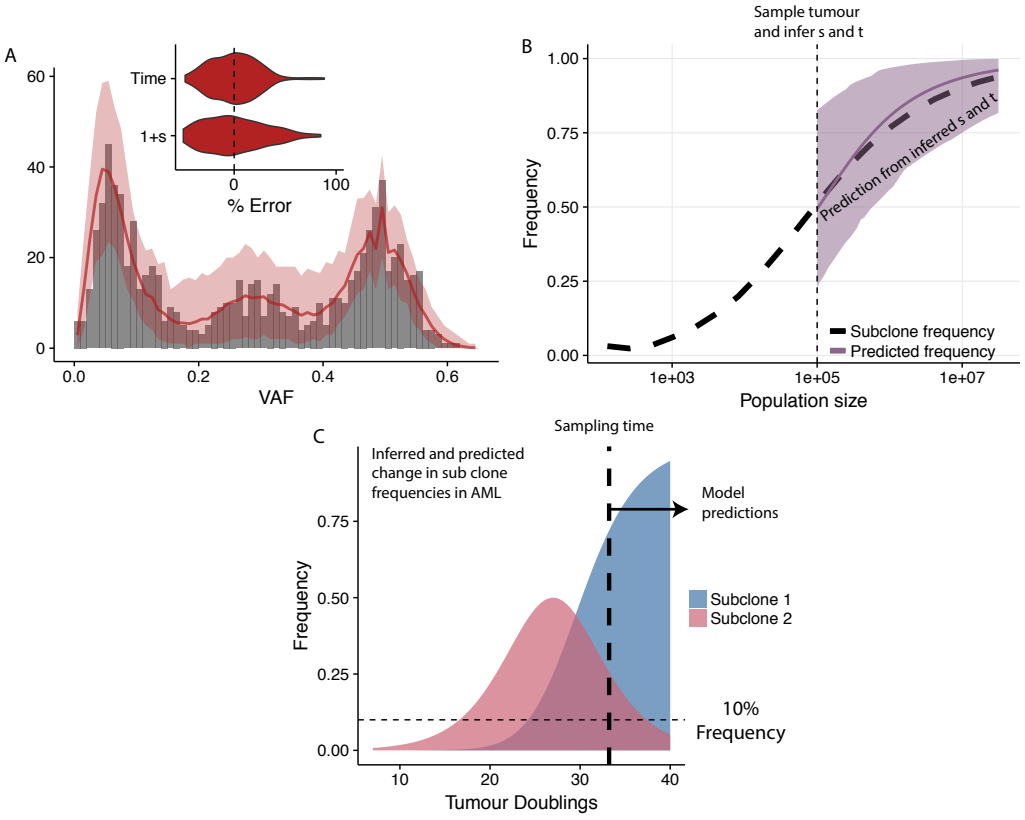

Figure 4.22: ***A*** *VAF of in silico tumour sampled at* 10^5 *cells was used to measure the fitness and time of emergence of a subclone.* ***B*** *These values were then used to predict the spread of the subclone as the tumour grew to* 10^7 *cells, showing the predictions matched the ground truth. Predictions were made by extrapolating the posterior distribution of* $1+s$ *using equations in the main text.* ***C*** *Using the same approach in the AML sample, where I measured* $1+s$*,* t_1 *and* t_2*, the model would predict that clone 2 would become dominant within 3-4 further tumour doublings while clone 1 may be too small to be detected.*

the clonal composition, and large effect subclones are required to dramatically alter the subclone architecture (Castro-Giner *et al.,* 2015). Indeed, modelling results show that if selection is too weak it is unlikely to be observable in bulk sequencing data, corroborating other studies (Sottoriva *et al.,* 2015; Sun *et al.,* 2017). A limitation of the approach is therefore that only those subclones that appear in a narrow window can be accurately measured as exemplified in Figure 4.7.

The model presented here is of course an abstraction of the dynamics during tumour growth. In particular, the model does not include spatial effects of the evolutionary process which may effect the accuracy of our measurements (Fusco *et al.,* 2016). In particular our framework cannot be used to quantify the degree of mixing within the tumour cell population, a phenomenon which has been shown to be a signature of effectively neutral dynamics (Sottoriva *et al.,* 2015). Our results however do demonstrate heterogeneity in the evolutionary process, multiple samples from the same tumours showed different evolutionary dynamics with one sample from the lung cancer dataset appearing to harbour a subclone while 4 others were consistent with a neutral model. Integrating spatial information into this type of approach could potentially lead to better estimates and more power to detect subclones, however given little is known on the dynamics of growth at a spatial scale this is currently challenging to implement. I also neglected the effect of cell death and assumed a small population size when fitting to data using the simulation. This was deemed an appropriate compromise between computational efficiency and model complexity. In future the ABC algorithm could be parallelised to enable more simulations per tumour or alternatively a statistical model which includes all relevant aspects such as the $1/f$ tail could be implemented that extracts all relevant parameters. This would overcome certain limitations of the ABC approach such as the large computational cost required to run the analysis. A statistical model which takes into account neutral $1/f$ distributed mutations together with an approach such as Dirichlet clustering would likely be orders of magnitude faster and would allow the approach to be applied in a wide range of circumstances.

This work also demonstrates the potential importance of stochastic effects dur-

ing tumour growth. Deviations from the deterministic model of neutral tumour growth can arise due to drift, and can in some cases lead to subclonal clusters, in particular if the death rate is high. This leads to a high proportion of colon and gastric cancers being classified as neutral compared to when only the deterministic model was applied to the data.

In summary this chapter has shown how integrating dynamical models of cancer evolution with genomic data enables quantifying the evolutionary process during tumour growth and facilitates mechanistic prediction with many potential applications.

4.8 Acknowledgements

This project was conducted in collaboration with Trevor Graham and Chris Barnes (my PhD supervisors), Andrea Sottoriva and Benjamin Werner as well as Timon Heide and Christina Curtis. I developed the mathematical framework, analysed the data and fitted the models with support from the above people. This chapter is a version of the work first presented in the following publication:

Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, Graham TA. Quantification of subclonal selection in cancer from bulk sequencing data. Nature Genetics. 2018 Jun;50(6):895?903.

Chapter 5

Population dynamics of dN/dS

5.1 Introduction

In this chapter I will take a different approach to the previous chapters and explore dN/dS as a method to identify selection in cancer genomes, although, as will become apparent, results from the previous two chapters can help inform value of dN/dS and how they should be interpreted. dN/dS has its roots in comparative species evolution as a method to uncover genetic loci under selection. In recent years the approach has been adapted and applied to data from somatic human tissues (both cancer and non-cancer) in an attempt to uncover signatures of selection (Martincorena *et al.,* 2015; Wu *et al.,* 2016; Martincorena *et al.,* 2017; Greenman *et al.,* 2006; Yang *et al.,* 2003; Weghorn & Sunyaev, 2017). The intuitive idea is that synonymous mutations can be used to estimate the rate of neutral substitutions (dS) and then by quantifying the abundance of non synonymous substitutions (dN) relative to this baseline, positive or negative selection at the level of the genome can be inferred. An over representation of non-synonymous mutations ($dN/dS > 1$) is a signature of positive selection as these are mutations that are expected to have functional consequences at the protein level. Meanwhile an under representation of non synonymous mutations ($dN/dS < 1$) is indicative of negative or purifying selection. $dN/dS = 1$ is the neutral expectation where the rate of non-synonymous and synonymous mutations are equivalent. dN/dS is an alternative approach to identifying selection in cancer that in theory does not rely on using the size of lineages and

their expectation under various evolutionary models. However I will show in this chapter how the population dynamics can influence measurements of dN/dS and confound its interpretation in cancer genomes. Another difference to the approach I have taken thus far is that dN/dS in cancer genomics is generally applied to cohorts of tumours and genome wide dN/dS or gene-level dN/dS across the cohort is reported. This differs from inferring levels of selection or the lack of it thereof on a sample by sample basis as was done in the previous two chapters.

dN/dS was originally developed to investigate evolutionary pressures on the genomes of divergent species (Goldman & Yang, 1994). Importantly, in this case dN/dS is a measure of the rate of substitutions across divergent lineages, or in other words the rate at which mutations arise and subsequently fix. Consideration of a Wright-Fisher process at long time limit then provides a straightforward mapping of values of dN/dS to the selection coefficient s (Nielsen & Yang, 2003):

$$\frac{dN}{dS} = \frac{2Ns}{1 - e^{-2Ns}} \tag{5.1}$$

Where N is the population size and s is the selection coefficient. Equation (5.1) is derived by considering the fixation probabilities of new neutral and non-neutral mutations and thus is appropriate for species evolution where genetic differences represent fixation events.

Later work used the same methodology to look at sequence evolution within the same species, such as polymorphisms within humans. However it has been shown that the interpretation and inference of selection with dN/dS when applied to segregating polymorphisms within the same species is not straightforward (Mugal *et al.,* 2013; Kryazhimskiy & Plotkin, 2008; Peterson & Masel, 2009). Over shorter time scales the dynamics of the process of fixation/loss ie the population genetics of the process becomes important. Kryazhimskiy & Plotkin, 2008 found that this means inferring the intensity of selection from dN/dS is problematic. That is the straightforward mapping between dN/dS and s given by equation (5.1) no longer hold. This is particularly striking in the case of large fitness effects and high mutation rates where dN/dS can be < 1 even for strong positive selection. The intu-

ition behind this contradictory result is that if an allele is undergoing a rapid sweep through a population then two individuals sampled randomly are likely to carry the allele (same dN in both samples) while neutral synonymous mutations may have accumulated independently in the two individuals (different dS).

This cautionary tale should serve as a warning in applying dN/dS in cancer cell populations given that time scales are short and that many mutations in cancer are only found in a subset of the cancer cells. This makes the process more qualitatively similar to segregating polymorphisms within species than fixed substitutions in distantly related species. Furthermore, peculiarities of evolution unique to cancer may introduce further difficulties (as well as advantages) with regard to the application of dN/dS to infer selection in cancer genomes. In particular, cancers grow, which as has been discussed in the previous chapters dilutes the effect of selection and much of the theoretical work devoted to dN/dS assume a fixed size Wright-Fisher model (Yang & Bielawski, 2000). Despite some of these concerns, using dN/dS in cancer also has some properties that make it appealing. Firstly there is no recombination in cancer so this does not need to be taken into account. Secondly, the ancestral genome is always known, as mutations in cancer are found by comparing to a normal reference sample. In species evolution this must be inferred from phylogenetic methods.

This chapter will focus on developing a model of dN/dS that takes into account the population dynamics of cancer. I will use stochastic simulations and theoretical models based on branching processes to to derive the expected distribution of dN/dS. In other words the goal is to derive an equivalent expression to equation (5.1) applicable to cancer. I will then apply this model to data from TCGA.

5.2 Methods

To investigate the role of population dynamics in the interpretation of dN/dS in cancer I modified and extended the simulation presented in the previous chapter so that driver mutations ($s \neq 0$) occur stochastically at rate μ_d, while the passenger mutations ($s = 0$) occur with rate μ_p. Driver mutations provide an increased fitness

advantage for cells and fitness effects combine multiplicatively, such that a cell with n drivers has fitness advantage $(1+s)^n$. Similarly to the model in the previous chapter, at time 0 a single cell (with no mutations) begins a birth death process which terminates when the population has reached some size N. A measure of dN/dS can then be obtained by counting the numbers of each type (N_d, N_p) of mutation in the cancer and scaling by their respective mutation rates.

$$\frac{dN}{dS} = \frac{N_d/\mu_d}{N_d/\mu_d} \tag{5.2}$$

This represents an idealised scenario where all non-synonymous mutations are drivers (ie they provide a fitness advantage), while this is almost certainly not true in cancer it represents a straightforward scenario with which to explore these dynamics. It is also possible with this model to explore negative selection by setting the fitness effect of mutations to be < 0. This simulation based model will be used to verify the prediction of a theoretical model of dN/dS for cancer evolution and to test its assumptions which I'll now derive in the next section.

5.3 Results

5.3.1 dN/dS for exponentially growing populations

Here I'll develop a theoretical model for dN/dS which is applicable to cancer evolution. The aim is to derive an equivalent equation to equation (5.1). Such a model needs to take into account the population growth dynamics and the existence of subclonal mutations which may rise and fall in frequency over time due to selection or drift. Fortunately, the theoretical framework of these dynamics (the Luria-Delbrück distribution and its extensions) has been well studied and can be adapted for cancer evolution. We have already encountered this distribution in Chapter 3 where using a deterministic model we showed that for neutral mutations, the cumulative number of mutations with a frequency $> f$ follows a $1/f$ power law. It is also possible to derive the expected distribution for non-neutral mutations under certain assumptions. Here I'll follow closely the model outlined in Kessler & Levine, 2014.

Kessler & Levine use branching processes to derive expressions for what they refer to as the clone size distribution for both neutral and non-neutral mutations (under certain limiting assumptions), however these expressions may more accurately be called the site frequency spectrum, as these expressions represent the expected number of mutants at a given clone size and depend on the mutation rate, and are therefore equivalent stochastic expressions to the model in Chapter 3. The derivation of some of these expressions is rather lengthy so I will discuss without proof the most relevant results. Briefly, solutions can be obtained for the relevant master equations using generating functions. I will call $C_N^{neut}(m)$ the expected number of lineages (or clones in the nomenclature of Kessler & Levine) of size m, given a total population size N for neutral mutations. Here m is an integer valued number, rather than a frequency as in Chapter 3. This exhibits a fat-tailed $1/m^2$ dependence which is well known in literature concerned with the Luria-Delbrück distribution (Zheng, 1999):

$$C_N^{neut}(m) = \frac{\mu_p N}{\beta}\frac{1}{m^2} \tag{5.3}$$

Here, μ_p is the neutral passenger mutation rate per division and as before β is the probability of a lineage surviving. In terms of a birth rate b and death rate d, $\beta = (b-d)/b$. Note that we recover the $1/m$ (or equivalently $1/f$) dependence of Chapter 3 by integrating the above lineage size distribution over m, $\int C_N^{neut}(m)dm \sim 1/m$. Kessler & Levine also derive expressions for the case when mutants have different growth rates compared to wild types. As in the previous chapter I'll define s as the ratio of net growth rates (λ) between wild type and mutants (or drivers) $1+s = \lambda_m/\lambda_w$. The site frequency spectrum for this two-type process with different growth rates is also found to be power law tailed but with exponent dependent on the relative fitness of the mutant population s.

$$C_N^{sel}(m) = \frac{N\mu_d}{\beta_d^{\frac{1}{1+s}}}\frac{b_p}{b_d}\frac{\Gamma(\frac{2+s}{1+s})}{m^{\frac{2+s}{1+s}}} \tag{5.4}$$

Where $\Gamma(x)$ is the gamma function, and b_p is the birth rate of the wildtype (ie the neutral passengers) lineages and b_d is the birth rate of lineages with a driver

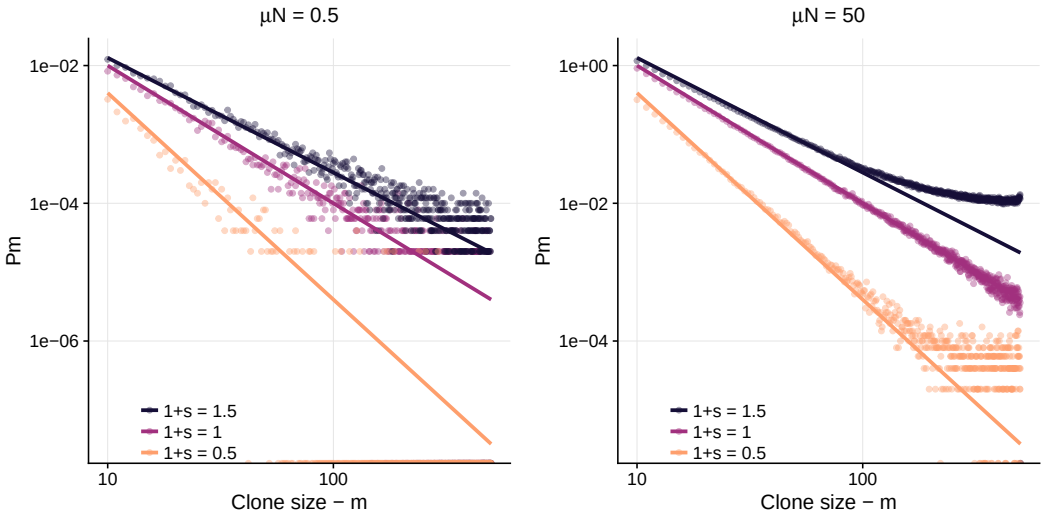

Figure 5.1: *Theoretical site frequency spectrum from equations* (5.4) *and* (5.3) *compared with simulations. Solid lines are theoretical results, points are simulations. 1,000 simulations were performed and the mean number of clones of each size m calculated. Theoretical results match simulations well for* $\mu N \sim 1$ *(left). Simulation parameters:* $N = 500$ *and* $\mu = 0.004$ *on the left and* $\mu = 0.1$ *on the right.*

mutation. As a sanity check we note the following:

$$\lim_{s \to 0} C_N^{sel}(m) \approx C_N^{neut}(m) \tag{5.5}$$

For neutral mutations equation (5.3) is valid for all mutation rates, however this is not the case for equation (5.4). When $s \neq 0$ and the mutation rate is high there is a significant probability of double mutants which are not accounted for in the above model. For $\mu N \sim 1$ and smaller equation (5.4) is a good approximation of the clone size distribution, but not for $\mu N \gg 1$. I checked the validity of these limits via simulations, see Figure 5.1. Assuming small mutation rates is not an unreasonable assumption for the emergence of driver mutations in cancer as the population size is large and the small number of driver mutations per tumour (<10) suggests the driver mutation rate is small.

Note that there is a dependence on the mutation rate in equations (5.3) and (5.4). The goal of dN/dS approaches is to quantify the excess or deficiency of mutations due to selection taking into account mutation rate variability. In dN/dS

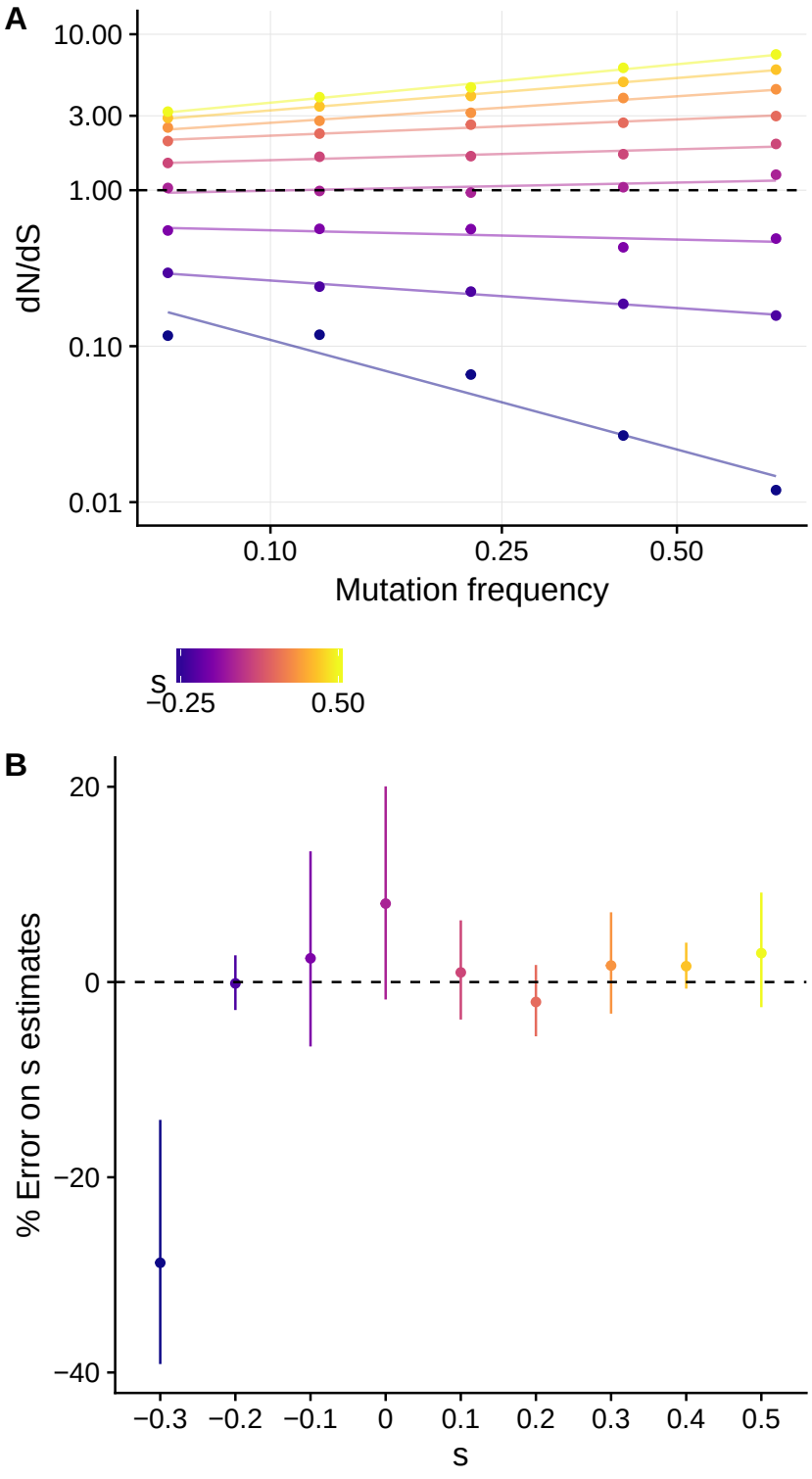

Figure 5.2: ***A*** *Simulations recapitulate predictions from Equations* (5.6) *and* (5.7) *that* dN/dS *should increase as a function of VAF. Each point is* dN/dS *calculated from a cohort of 5,000 tumours. Lines are regressions through the points. Simulation parameters: passenger mutation rate: 1/division driver mutation rate: 0.01/division, maximum population size:* 10^4*, birth rate of host* $= \log 2$*. Driver mutations increase birth rate of clone.* ***B*** *Extracting the regression coefficient enables accurately estimating the selection coefficient, apart from when s is strongly negative. Error bars are 95% confidence intervals from the regressions.*

methods, care is taken to accurately account for variability of mutation rates across the genome and the number of sites in order to isolate the role of selection from mutation. To address this in this theoretical approach we can take the ratio of equations (5.3) and (5.4) and normalize for the mutation rate to derive our dN/dS formula:

$$\frac{dN}{dS} = \frac{C_N^{sel}(m,s)/\mu_d}{C_N^{sel}(m,s=0)/\mu_p} = \frac{b_w}{b_m}\frac{\beta_w}{\beta_m^{\frac{1}{1+s}}}\Gamma\left(\frac{2+s}{1+s}\right)m^{\frac{s}{1+s}} \tag{5.6}$$

The expected excess increases as a function of clone size for positive selection and decreases for negative selection. For neutral mutations ($s=0$), $\frac{dN}{dS}=1$ as would be expected. Thus in cancer we would expect dN/dS to increase as a function of the lineage size for positively selected mutations and decrease for negatively selected mutations.

A straightforward way to modify equation (5.6) so that it can be directly applied to data is to linearize it by taking the log of both sides:

$$\log\left(\frac{dN}{dS}\right) = \frac{s}{1+s}\log(m) + C \tag{5.7}$$

This makes the gradient of the slope between when dN/dS is plotted against mutation lineage size a readout of the relative fitness advantage s. The intercept C meanwhile is a combination of many other parameters.

To confirm the predictions of equation (5.7) I used the simulation based model to generate a number of synthetic cohorts of 5000 tumours with varying fitness coefficients and then measured the dN/dS ratio at different lineage sizes across the cohort. Then I could test if the predictions of equation (5.7) were valid and that it is possible to accurately recover the selection coefficient. Figure 5.2 shows that this was indeed the case and a positive correlation between dN/dS and cancer cell fraction exists for positive selection and a negative correlation for negative selection.

This model neglects some potentially important features of cancers, it assumes all mutants have the same fitness, that double (or higher order) mutants do not occur and also neglects passengers hitchhiking to higher frequency on the back of driver mutations. So to confirm that the excess of drivers should increase as a function of lineage size when these simplifications are not held true and to investigate the accu-

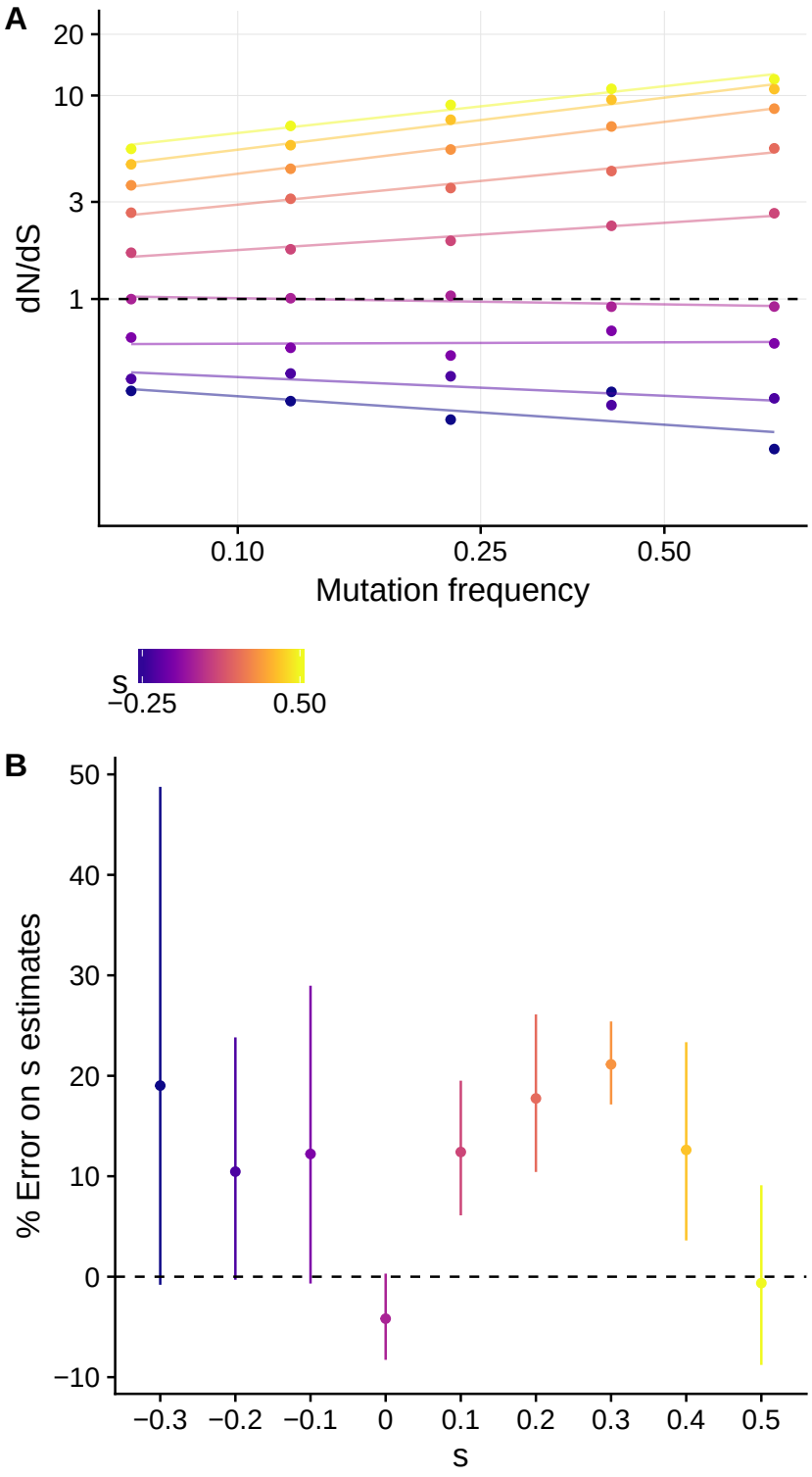

Figure 5.3: *Equivalent to Figure 5.2 but for a model with relaxed assumptions. Simulation parameters: passenger mutation rate: 1/division driver mutation rate: 0.01/division, maximum population size:* 10^4*, birth rate of host* $= \log 2$*. Driver mutations increase birth rate of clone and are drawn from an exponential distribution with mean s. Fitness is multiplicative so that a lineage with 2 driver mutations has fitness* $(1+s_1)(1+s_2)$

racy of inferring the selection coefficient when fitness is drawn from a distribution rather than fixed I relaxed some of the assumption. I modified the simulation to generate another cohort of tumours while relaxing the assumption. The fitness effect of a new driver mutation is now drawn from an exponential distribution with mean given by s and multiple driver mutations can accumulate within a lineage where the effect is multiplicative.

Figure 5.3 shows the results of using this more realistic model. In this case we see that there is a tendency to overestimate the selection coefficient but that this overestimation is within 20-30% of the true value suggesting that the model represents a reasonable approximation even with some limiting assumptions.

Having developed the theory and tested it on a cohort of simulated data next I wanted to test the theory on data. First of all however it will be instructive to discuss one particular aspect of cancer evolution not included in the theory. The theory only takes into account mutations that appear as the cancer grows and not any mutations that are present in the first cancer cell that gives rise to the tumour that is ultimately sampled. In reality cancers will have many clonal mutations (mutations present in every cell), and given that this first cancer cell by definition has a higher fitness and clonally expands we would expect these mutations to have $dN/dS>1$. This therefore will exacerbate the trend that our theory predicts of high dN/dS at high mutation frequencies.

Taking these two things together we can summarise how we would expect the dN/dS ratio to change over the VAF spectrum reported by deep sequencing of cancers. In summary we would expect high frequency mutations to have the highest dN/dS as this is where we would expect to see clonal mutations and mutations that have risen to a high frequency via selection. Low frequency mutations on the other hand would be expected to have the lowest dN/dS as these mutations will be dominated by neutral mutations where selection has not had enough time to act (as discussed in the previous chapter). In some tumours where the fitness advantage of a mutant is high enough and it emerges early enough we would observe mutational clusters which would inflate dN/dS at intermediate and higher frequencies when

over a whole cohort. We would also therefore expect that mutational clusters have $dN/dS>1$.

5.3.2 TCGA data

Data from the Cancer Genome Atlas was used to test the theoretical results. It is worth noting that this approach - looking for a correlation between lineage size and dN/dS - is attractive for reasons related to robustness of dN/dS measurements in addition. dN/dS measurements can be influenced by many confounding factors which influence the baseline mutation rate, in particular mutational biases such as those described in Alexandrov *et al.,* 2013 will alter the baseline expectation and need to be corrected for. This is a difficult challenge for sparse mutational data where assigning the correct mutational bias is difficult, however as we are looking for a trend this should be robust any such biases given that any errors associated with them should be consistent across the VAF spectrum.

In the previous 2 chapters, to ensure there was a straightforward mapping between the size of mutational lineages and the variant allele frequency I removed mutations falling in non-diploid regions. For this analysis however, utilising all the mutations is important because driver mutations are known to be rare and therefore it is likely to be important to use as much of the data as possible. This raises the problem that in regions of the genome where there are copy number gains or losses the straightforward mapping between VAF and lineage size is lost. Furthermore different purity of samples across a cohort will influence any analysis done across the whole cohort. Therefore for this analysis I converted VAF to cancer cell fraction (CCF) using copy number data and cellularity estimates. For mutation i with with variant allele frequency f_i, copy number CN_i at the locus and cellularity of the tumour c_T the CCF was calculated as follows:

$$CCF_i = \frac{CN_i \times f_i}{c_T} \tag{5.8}$$

TCGA data was downloaded using the TCGAbiolinks R package (Colaprico *et al.,* 2016). Mutation files from the Mutect2 mutation calling algorithm and copy

number segmentation data for 9950 cancers were downloaded. Cellularity estimates were obtained from Aran *et al.,* 2015. I then filtered for > 2 reads reporting the variant and > 9 reads coverage at each locus in both the tumour and normal sample and then calculated the CCF of mutations using equation (5.8). I removed samples where cellularity estimates were unavailable or where the cellularity was estimated to be $< 50\%$ so that there was good power to detect low frequency variants. I also removed cancer types that had < 100 samples. This led to 6694 samples from 17 cancer types suitable for analysis.

For all the analysis that follows mutations were binned into CCF bins so that dN/dS as a function of mutation frequency could be investigated. For inferring dN/dS values I used the *dndscv* R package (Martincorena *et al.,* 2017). The *dndscv* package applies corrections based on many potential confounding factors such as chromatin state, gene expression level and mutational signatures. To test the theory, dN/dS was calculated per bin and values plotted against average CCF values of mutations within the bins on a log-log scale where we would expect to see a linear relationship as in equation (5.7).

5.3.2.1 dN/dS across the whole genome and in cancer genes

Figure 5.4 shows this analysis for all mutations across the genome, for the majority of cancer types there is no significant correlation (using linear regression) between dN/dS and CCF, the only cancer types with a significant correlation are low grade glioma (LGG) and prostate adenocarcinoma (PRAD). Next I redid the analysis restricting dN/dS measurements to a set of of putative cancer genes (550 genes from the COSMIC cancer gene census), reasoning that this should result in increased power to detect signatures of positive selection as in (Martincorena *et al.,* 2017). For this restricted set of genes we do observe the predicted increase of dN/dS as a function of CCF for all cancer types except for kidney renal papillary cell carcinoma (KIRP), kidney renal clear cell carcinoma (KIRC) and melanoma (SKCM). From the regressions we can extract an estimate of the selection coefficient. A summary of this estimate across cancer types is shown in Figure 5.6. Following multiple comparison correction (Benjamini-Hochberg) of the p-values associated with the linear

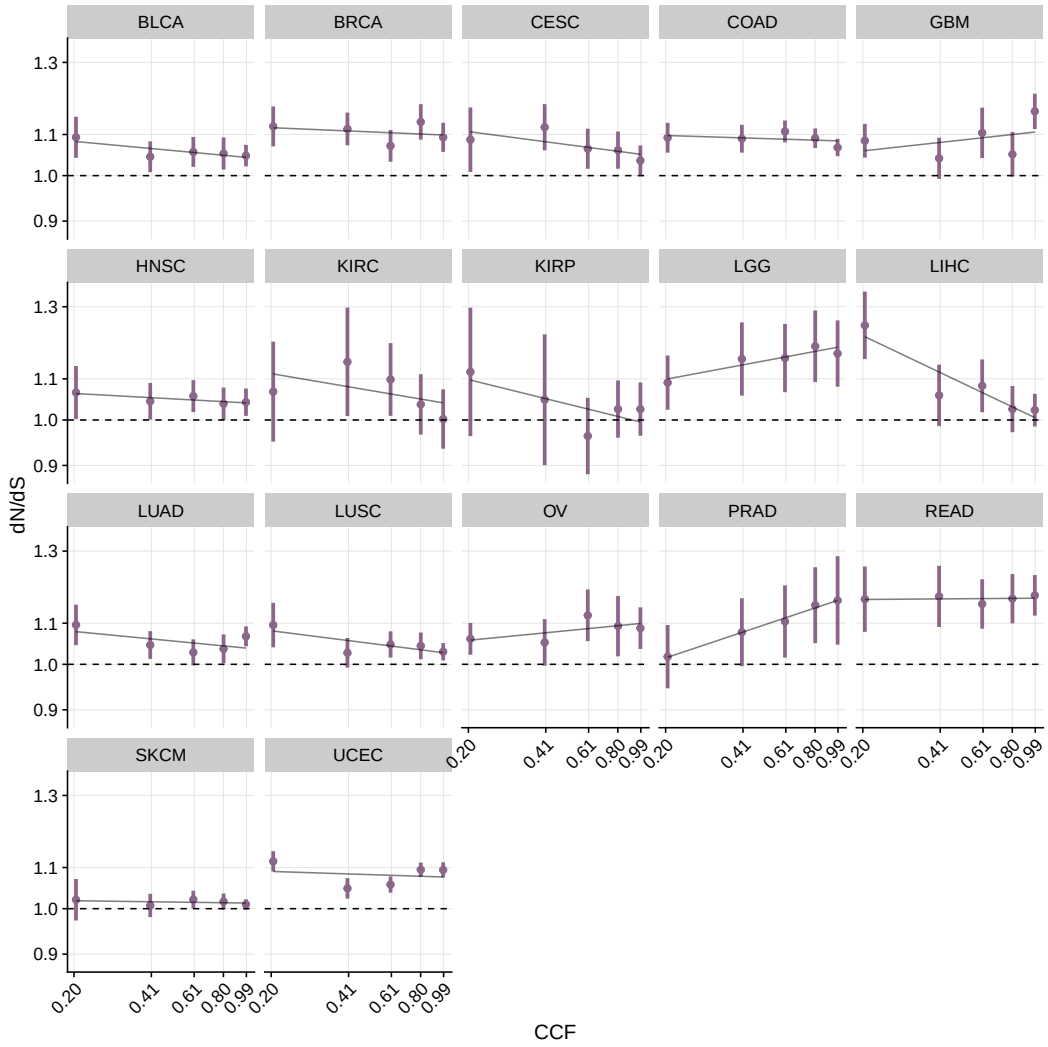

Figure 5.4: *Exome wide dN/dS calculated as a function of CCF across cancer types. Values plotted on a log-log plot. Black line shows the linear regression line between dN/dS and CCF.*

regression only 2 cancer types retained a significant ($q < 0.1$) correlation. However given the regressions were performed over only 5 data points it is perhaps not surprising. The fact that there appears to be a positive correlation in 14/17 cancer types suggests that this is a genuine phenomenon.

5.3.2.2 dN/dS for cell essential genes

Recent results from both Martincorena *et al.,* 2017 and Weghorn & Sunyaev, 2017 suggest that cancers can tolerate almost all mutations and negative selection is rare in cancer genomes. Both these papers look for an absence of mutations as a signa-

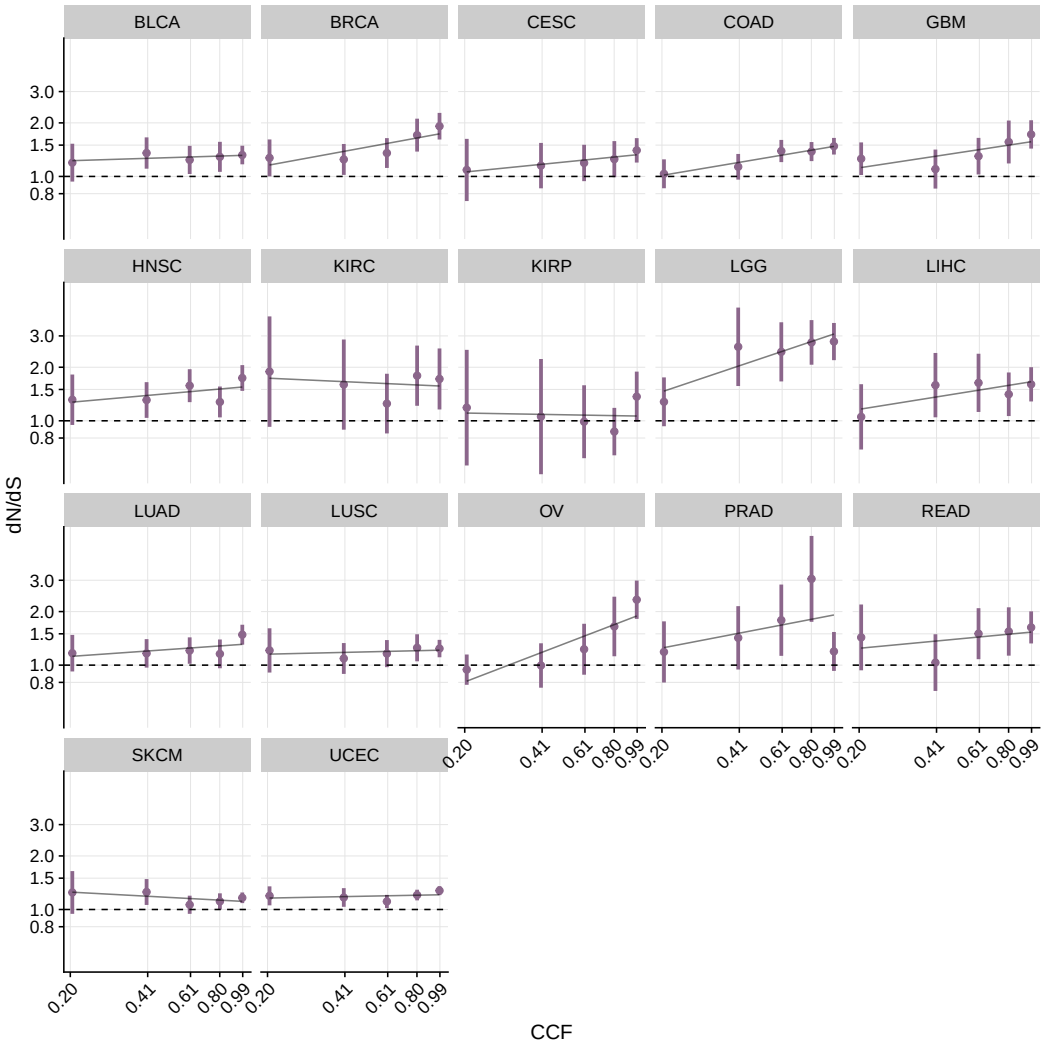

Figure 5.5: *dN/dS calculated as a function of CCF across cancer types for 550 putative driver mutations. Values plotted on a log-log plot. Black line shows the linear regression line between dN/dS and CCF.*

ture of negative selection. Deleterious mutations however do not necessarily cause cell death, but rather may retard the growth compared to other cancer cells. Equation (5.6) would predict that deleterious mutations that retard growth would be on average at lower frequencies compared to neutral mutations, thus negatively selected mutations should decrease in frequency as a function of dN/dS as observed in the simulation based model. To look at this, as in Martincorena *et al.,* 2017 I looked at dN/dS for genes classified as cell essential in Wang *et al.,* 2015 as a function of their lineage size. For most cancer types, this analysis revealed no significant correlations, except in liver hepatocellular carcinoma (LIHC) and lung squamous cell

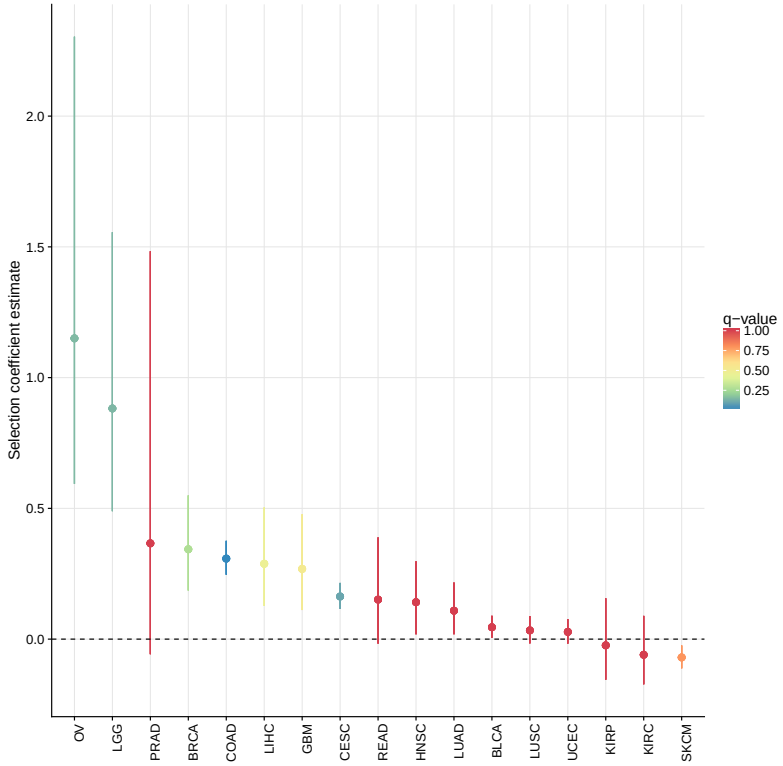

Figure 5.6: *Values from linear regression coefficients were used to estimate the selection coefficient based on equation* (5.7) *for all cancer types in Figure 5.5, colours donate the q-value from multiple testing of the regression p-values. Confidence intervals show +/- standard deviation from linear regression model.*

carcinoma, see Figures 5.9 and 5.10. This suggests that negative selection plays a less important role than positive selection in line with recent result from Martincorena *et al.,* 2017 and Weghorn & Sunyaev, 2017. Summarising the estimated selection coefficients for all cancer types across the whole exome, the cancer gene census panel and the cell essential gene panel show that only the cancer gene panel shows a clear departure from neutrality, Figure 5.7.

Restricting the analysis to haploid regions (logR < 1.5) does however reveal patterns of negative selection, see Figure 5.8 which compares dN/dS for mutations falling in diploid vs haploid regions. With this restricted set of mutations we do observe the expected decrease in dN/dS as a function of CCF for truncating mutations in cell essential genes.

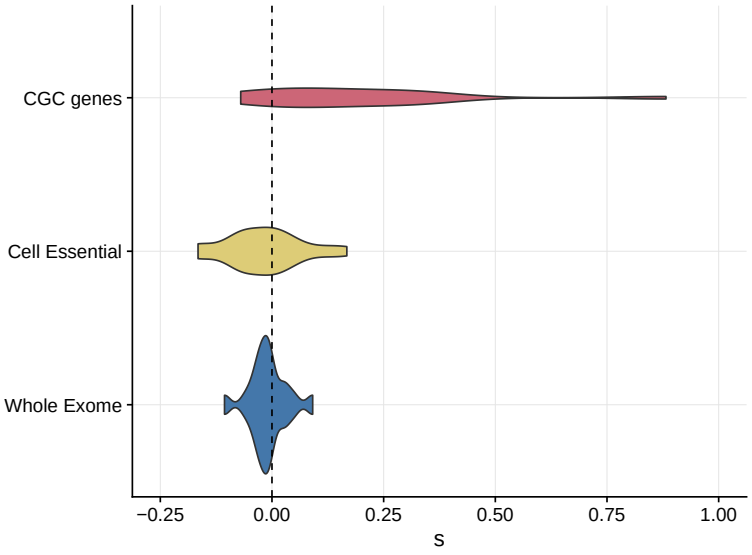

Figure 5.7: *Summary of the selection effects measured across cancer types in different sets of genes. Only putative driver genes from cosmic show a significant departure from neutrality.*

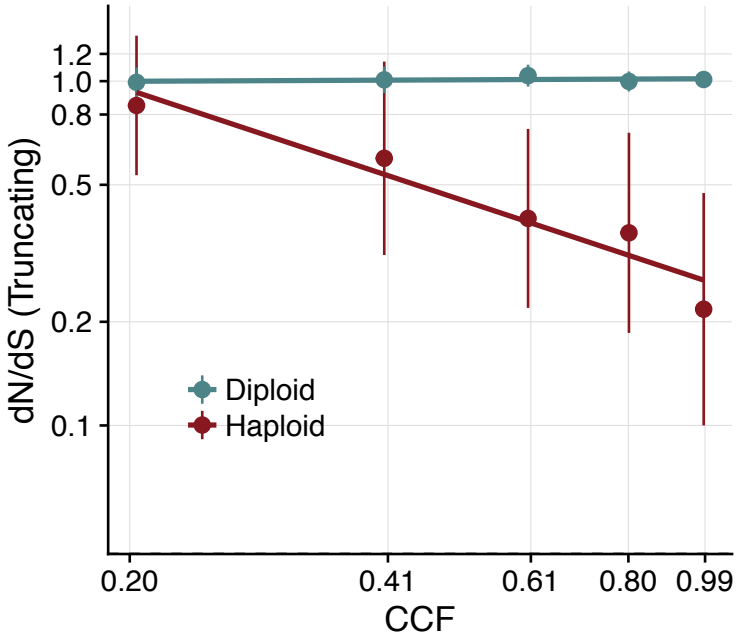

Figure 5.8: *Restricting the analysis to mutation in cell essential genes in haploid regions of the genome shows the effects of negative selection across the TCGA cohort.*

5.4 Discussion

Recent applications of dN/dS to cancer have shown that dN/dS is approximately 1 across the genome with a small number of genes with large dN/dS values (Wu

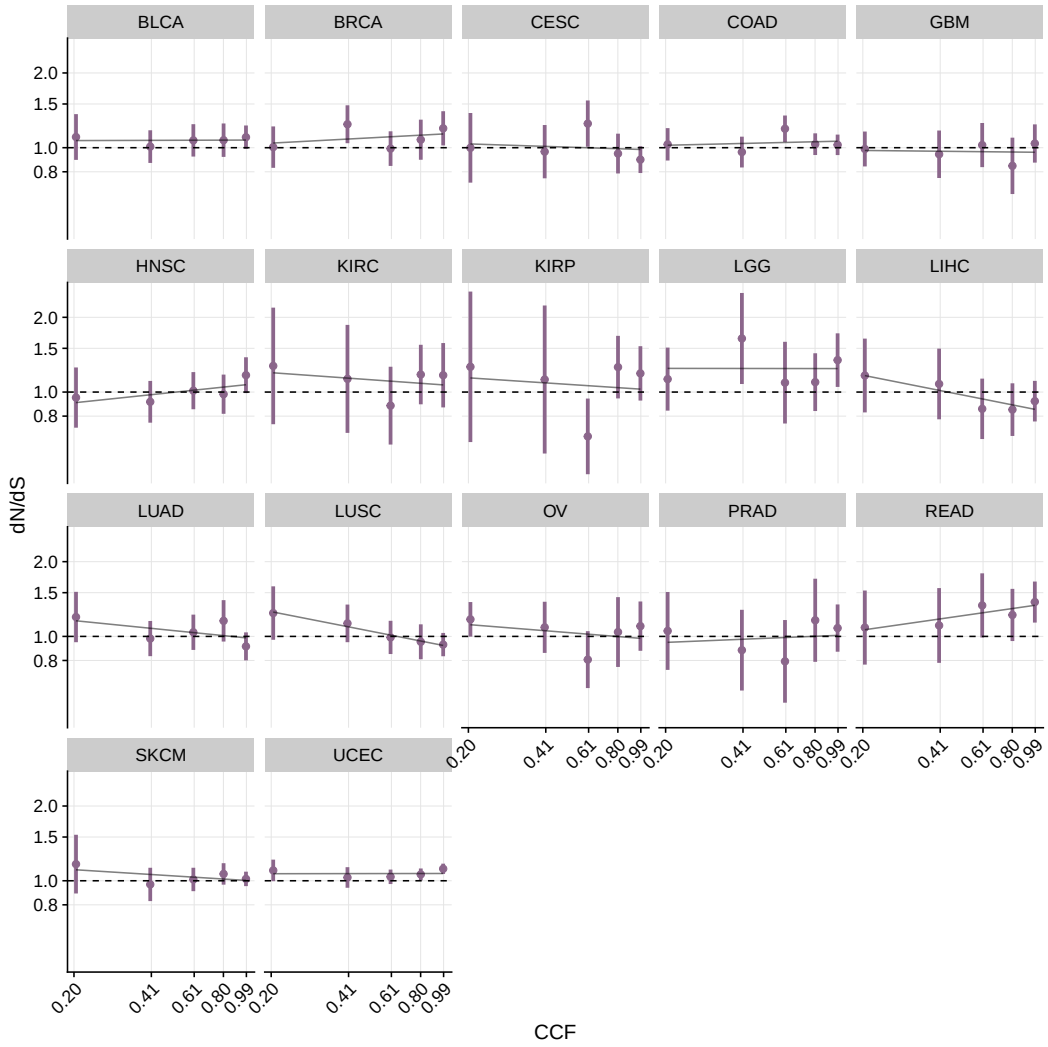

Figure 5.9: *dN/dS calculated as a function of VAF across cancer types for 1100 cell essential genes. Red line shows the linear regression line between dN/dS and VAF, each plot is annotated with R2 and p-values of the linear regression coefficient.*

et al., 2016; Martincorena *et al.,* 2017; Weghorn & Sunyaev, 2017). These papers also showed that positive selection was far more evident than negative selection in cancer genomes. These studies however failed to consider the population genetic consequence of positively or negatively selected mutations, which results in lineages rising to higher frequency (or decreasing in frequency) than would be expected under neutral evolution and provides another layer of information that can be incorporated into these analyses to uncover patterns of selection.

In this chapter I showed that combining comparative genomics with population genetics can provide further insight into patterns of cancer evolution across large

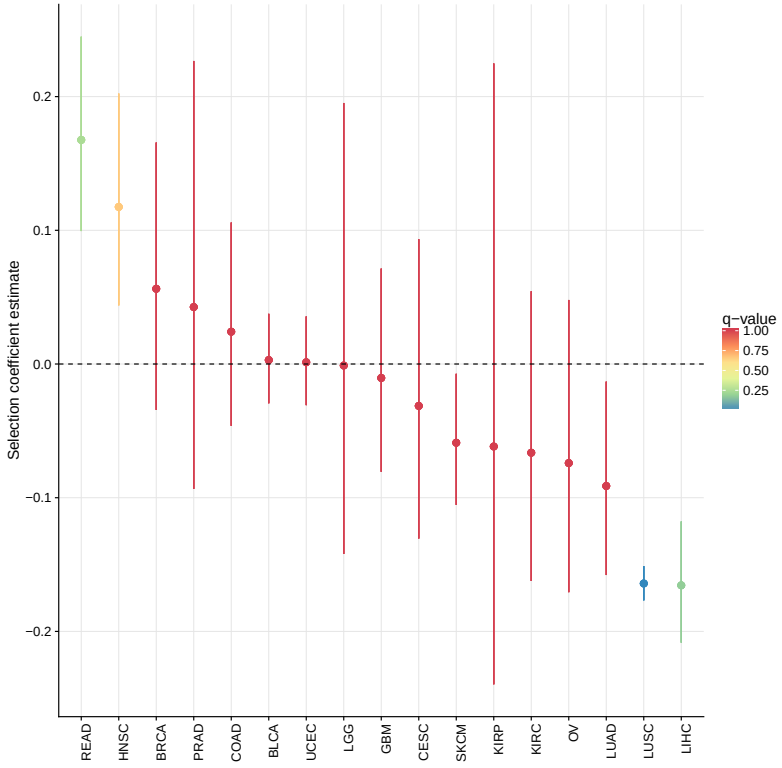

Figure 5.10: *Linear regression coefficients for all statistical test from Figure 5.9, colours donate whether regressions was significant or not. Confidence intervals show +/- standard deviation from linear regression model.*

cohorts of sequenced tumours. This approach allows for estimating the selection coefficients of non-neutral mutations which has so far been neglected in dN/dS approaches in cancer. These results also show that the clonality of mutations can influence measurements of dN/dS due to the population genetics of selection during tumour growth.

Nonetheless using dN/dS in the study of cancer evolution is a powerful method because (rather than despite) of some of the peculiarities of cancer as an evolutionary system. Unlike in species evolution where two or more diverged lineages are compared to calculate dN/dS, in cancer evolution the ancestral genome is always known because mutations in the cancer are identified by comparing to a normal tissue or blood sample. Another peculiarity is that there is no recombination and therefore mutations hitchhike. Together these mitigate some peculiar dynamics that can materialise when classical dN/dS approaches are applied to closely related lin-

eages. For example the observation by Kryazhimskiy & Plotkin, 2008 that it is possible to observe $dN/dS < 1$ for strong positive selection in some scenarios is mitigated in cancer because the comparison is between the present day genome and its ancestor not between the genomes of two present day lineages. Both the analytical results and simulated cohorts show that dN/dS maintains a monotonically increasing relationship as a function of s, but considerations of the clonality of mutations is important for its interpretation.

Despite these appealing properties some difficulties remain however. dN/dS cannot elucidate evolutionary pressures on a sample by sample basis and only reveals properties in cohorts of cancers. Other approaches such as using the site frequency spectrum are likely more powerful for these type of questions as demonstrated in the previous chapter. Also as others have shown, correctly accounting for mutational biases across the genome is important for having unbiased estimates which can be challenging. Also some care is needed in interpreting selection coefficients like we measure here, firstly not all non-synonymous mutations are pathogenic so the inferred selection coefficients are averages across both pathogenic and non-pathogenic mutations. Furthermore synonymous mutations may in some cases be pathogenic and non-neutral (Bailey *et al.,* 2014). Nevertheless the results here demonstrate that a combination of population genetics and comparative genomics results can help identify signatures of selection across cancer genomes.

Chapter 6

Stem cell dynamics in the human colon

6.1 Introduction

Thus far in this thesis I have used models of population cell dynamics to measure evolutionary processes in cancer. In particular the focus has been on the dynamics post transformation, ie how do evolutionary processes shape the cancers we ultimately observe in the clinic or lab given that they descend from a single ancestor some time in the past. This however says little about the population dynamics pre-transformation and in particular how mutations in physiologically normal tissue (may) alter the population dynamics as the tissue progresses to a malignant state. This chapter will explore these processes, with a focus on the architecture of stem cell populations in the colon. Understanding how stem cells in the colon regulate cell turnover in homeostasis is crucial for understanding how this process is dysregulated as tissues progress to cancer.

The colon is an ideal tissue system to investigate the population dynamics in normal tissue as it has a well characterised stem cell population. During normal physiological conditions the colonic epithelium is made up of small invaginations in the lining of the epithelium. These finger like protrusions contain cells responsible for the uptake of nutrients and the transport and extrusion of waste, the two principal functions of the intestine. Due to the harsh environment encountered in

the intestine, the tissue undergoes continuous renewal. The entire epithelia is replaced over the course of a week (Vermeulen & Snippert, 2014). This renewal is facilitated by a population of intestinal stem cells that reside at the base of the crypt. Colonic stem cells were first identified *in vivo* using lineage tracing experiments in mice, where a genetic marker such as a fluorescent protein is induced in a candidate stem cell and then followed over time (Blanpain & Simons, 2013). These lineage tracing experiments, showed that cells residing at the bottom of the crypt expressing the LGR5 marker, were capable of self renewal and were multipotent, that is their progeny could give rise to all cell types within the crypt (Barker *et al.,* 2007), thus satisfying the necessary condition of stemness. Later lineage tracing experiments demonstrated that multiple stem cells reside at the bottom of the crypt and that these cells are all equipotent, and stochastically replace each other in a process analogous to neutral drift in population genetics (Lopez-Garcia *et al.,* 2010; Snippert *et al.,* 2010).

Colonic stem cells have gained considerable interest in the context of cancer in recent years as stem cells are thought to be the cell of origin for colorectal cancer (Barker *et al.,* 2009), and driver mutations such as mutations in the tumour suppressor *APC* in colorectal cancer have been shown to disrupt the WNT signalling pathway which is of known importance in maintaining the stem cell niche. Colorectal cancers and their precursor lesions adenomas, are made up of glandular structures, reminiscent of the crypts in physiologically normal tissue, suggesting a similar cellular hierarchy in these lesions and the existence of a cancer stem cell population (Medema & Vermeulen, 2011). LGR5 positive cells can also be found in adenomas and cancers although these cells have been found to be more diffuse in the glandular structures and not as localised (Baker *et al.,* 2015). If the existence of cancer stem cells is *bone fide* then it has profound implications for the progression and treatment of the disease, as due to their self-renewal capacity eradicating the CSC pool is the crucial task.

Controversy remains however over whether a cancer stem cell is an intrinsic property of a cell and that self renewal is a property exclusive to a subset of cells

(Wright, 2012) or alternatively that CSC is a plastic phenotype that requires a supportive niche and that non-CSC can become CSC if in the right place at the right time.

6.2 Mutations as a clonal lineage marker

Experiments that have been used to quantify stem cell dynamics have generally been some variant of a pulse chase experiment, where a *pulse* is given to induce a label such as a fluorescent protein, and the label is then followed over time. Daughter cells retain the label of their parent cell and the change in clone size can then be tracked over time. Analysis of the results of these experiments showed that these dynamics were consistent with a small number, 5-10 of equipotent stem cells residing at the bottom of the crypt undergoing stochastic loss and replacement (Lopez-Garcia *et al.,* 2010; Snippert *et al.,* 2010). Similar experiments showed that if one of the stem cells carried an oncogenic mutation in the *KRAS* gene, then the dynamics conformed to a biased drift, where the *KRAS* mutants were more likely to take over the whole stem cell population, the possibility of stochastic loss of the mutant remains however (Vermeulen *et al.,* 2013). Using naturally occurring mitochondrial mutations as a mechanism of labelling lineages, it was later shown that stem cells in human colonic crypts also exhibited the characteristic neutral drift process (Baker *et al.,* 2014). Interestingly, neutral drift dynamics of stem cell populations appears to be a conserved phenomenon across multiple tissue types, suggesting a universal mechanism that tissues use to renew their cellular population and suppress the accumulation of mutations (Klein & Simons, 2011).

While shedding light on the stem cell architecture, these lineage tracing experiments have some limitations. Generally, they cannot be applied to human tissues and tissues cannot be followed for extended periods of time. Another approach as used in Baker *et al.,* 2014 is to use naturally occurring labels in human tissues, such as mutations in mitochondrial DNA that encode cytochrome c-oxidase, these mutations can be visualized in tissue with histo-chemical staining. As discussed at length in the previous chapters, mutations in genomic DNA can also serve as clonal marks

and serve as an alternative naturally occurring label with which to track population dynamics. Such an approach has been used in physiologically normal human tissue to measure stem cell dynamics in the skin (Simons, 2016).

Using mutations in genomic DNA together with deep sequencing as a read out of the lineage size distribution has a number of appealing features for measuring stem cell dynamics. It can be applied to (any) human tissue, a record is kept of the dynamics across the whole lifespan of the tissue and multiple lineages can be followed simultaneously. This is the approach I take in this chapter.

6.2.1 Sequencing data

With the aim of assessing the stem cell dynamics in the colon, single crypts from physiology normal tissue were isolated and DNA extracted. 2 crypts each from 2 healthy individuals were obtained. One individuals crypts were subjected to whole exome sequencing and the other individual to whole genome sequencing. This data was processed as described in Chapter 2. To extract mutation calls and call copy number states across the genome Mutect2 and sequenza were used respectively. A summary of the data, showing the mean depth of coverage and the number of mutations called is shown in Table 6.1. To reduce false positive mutations but retain a large number of true positives, different filtering of mutations was done in the two sequencing strategies due to the large differences in depth, particularly in the normal sample. For the whole exome sequencing, mutations were filtered based on the following criteria:

1. $\geq 20X$ coverage in both test and control samples
2. ≥ 5 reads reporting the variant in the test sample
3. ≤ 2 reads reporting the variant in the control sample
4. ratio of VAF between crypt and normal > 10
5. Mutations removed if the change was $C > A$

Mutations in the $C > A$ channel were removed given there was an abundance of low frequency $C > A$ mutations (see Figure 6.1), these are likely to be errors arising

Sample ID	Mean Coverage (Tissue)	Coverage (Blood)	# mutations	# exonic mutations
450 A2	120	280	96	57
450 A6	75	280	127	80
452 R1	32	30	2063	28
452 R2	38	30	3284	31

Table 6.1: *Summary of sequencing data from normal crypts*

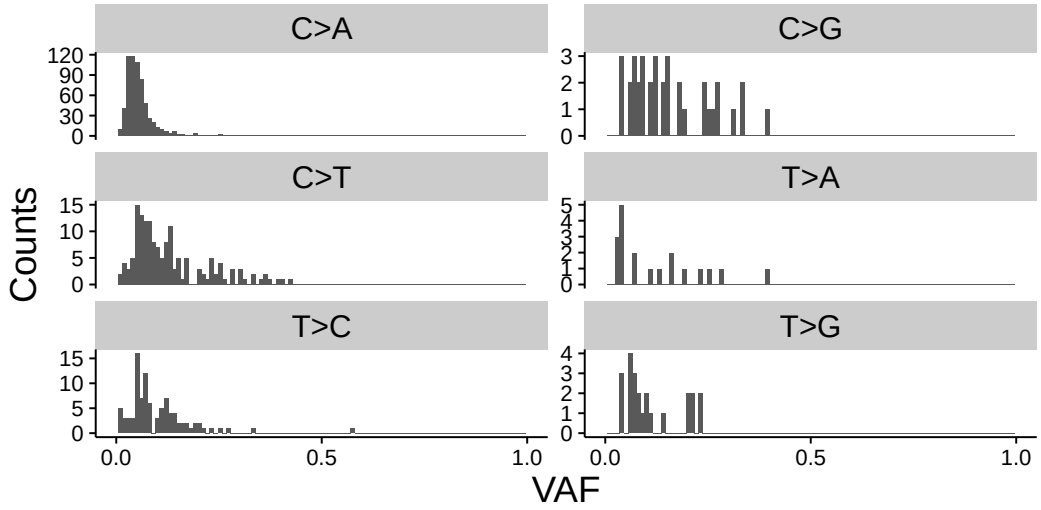

Figure 6.1: *Frequency distribution of mutations in 2 whole exome sequenced crypts stratified by mutation channel. There is an abundance of low frequency $C > A$ mutations which are likely to be errors arising from the library preparation and were removed from further analysis.*

due to DNA damage during library preparation (Chen *et al.,* 2017a; Costello *et al.,* 2013) and hence were removed. For the whole genome sequencing the following filtering criteria was used:

1. $\geq 20X$ coverage in both test and control samples
2. ≥ 5 reads reporting the variant in the test sample
3. 0 reads reporting the variant in the control sample

6.3 Neutral drift of equipotent stem cells

Given there is a substantial amount of experimental and theoretical work looking at the stem cell dynamics in normal tissue during homeostasis, particularly in mice

I first looked at the whether the sequencing data from normal crypts conforms to models of stem cell dynamics in the literature. In particular whether a model of neutral drift of an equipotent pool of stem cells fits this data. I'll first discuss and develop some small extensions to this theory before returning to the data to see whether this model can explain the sequencing data.

The neutral drift model of stem cells at the crypt base was originally demonstrated in Lopez-Garcia *et al.,* 2010. Much of the theoretical work was also developed here and has been applied subsequently with some modifications to other studies in mice (Ritsma *et al.,* 2014; Kozar *et al.,* 2013; Snippert *et al.,* 2013) and in humans (Baker *et al.,* 2014). I'll begin by introducing this theoretical model, and then will introduce some extensions to the model given that new lineages can be stochastically labelled over time by acquiring mutations.

In the original work presented in Lopez-Garcia *et al.,* 2010, stem cells are assumed to be equipotent and are orientated in such a why that they can only replace their nearest neighbours. This 1D arrangement can be thought of as a ring of stem cells where the loss of a stem cell results in one of its nearest neighbours replacing it with equal probability, these dynamics result in the number of stem cells staying constant throughout. Following the labelling of a single stem cell at time $t = 0$, we will define the probability of the clone having acquired a size n stem cells after a time t due to neutral drift of the stem cell population as $P_n(t)$, with loss-replacement rate λ. The master equation describing the time-evolution of this process can then be written as follows

$$\frac{dP_n(t)}{dt} = \lambda(P_{n+1} + P_{n-1} - 2P_n) - \lambda(\delta_{n,1} + \delta_{n,-1} - 2\delta_{n,0})P_0 + \lambda(\delta_{n,Ns+1} + \delta_{n,Ns-1} - 2\delta_{n,Ns})P_{N_s} + \delta_{n,1}\delta(t) \quad (6.1)$$

Where N_s is the number of stem cells occupying the niche. Here the first term describes the random walk of the clone size, while the second and third term express the potential for the clones to be lost and flushed out of the crypt or fix in the stem cell pool and hence the whole crypt population. The final term expresses the initial

condition of a single labelled stem cell at time $t = 0$. The solution to the above equation is given by (Lopez-Garcia *et al.,* 2010):

$$
\begin{aligned}
P_0(t) &= \frac{2}{N_s}\sum_{k=1}^{N_s=1}\cos^2\left(\frac{\pi k}{2N_s}\right)\left(1-e^{-4\lambda t\sin^2\left(\frac{\pi k}{2N_s}\right)}\right) \\
P_n(t) &= \frac{2}{N_s}\sum_{k=1}^{N_s=1}\sin\left(\frac{\pi kn}{N_s}\right)\sin\left(\frac{\pi k}{N_s}\right)e^{-4\lambda t\sin^2\left(\frac{\pi k}{2N_s}\right)} \\
P_{N_s}(t) &= \frac{2}{N_s}\sum_{k=1}^{N_s=1}(-1)^{k+1}\cos^2\left(\frac{\pi k}{2N_s}\right)\left(1-e^{-4\lambda t\sin^2\left(\frac{\pi k}{2N_s}\right)}\right)
\end{aligned}
\tag{6.2}
$$

At times $1 \ll \lambda t \ll N_s^2$ the clone size distribution exhibits a particular property termed scaling, in this scaling regime the shape of the clone size distribution scaled by the average clone size is conserved over time, and the above equations collapse to a simple exponential form. This scaling property has been shown to provide a quantitive test of stochastic loss and replacement of stem cells, as for example a strict hierarchical architecture with a *master* stem cell does not lead to scaling behaviour (Lopez-Garcia *et al.,* 2010). We would not expect the clone size distribution we measure from the variant allele frequency of mutations to conform to this simple scaling behaviour as the time scales over a human lifespan will be $\lambda t \gg N_s^2$, where λ has been measured to be of the order 0.2 per day in humans (Baker *et al.,* 2014). Furthermore, the VAF distribution should provide a readout of the number of fixed mutations within the crypt, meaning we can leverage information on both the number of fixed mutations and the number of partial mutations to infer the stem cell dynamics. This means the solution for the full clone size distribution is necessary. Furthermore, the above model only considers a single label in the crypt, we are interested in the case were labels (or rather mutations) are continuously generated. Kozar *et al.,* 2013 made some progress in this direction, however the question I wish to ask is slightly different, we wish to know how many mutations would we expect to see at a particular frequency, where as *Kozar et al.* where interested in the expected number of partial crypts vs fixed crypts. In the language of population genetics language, the number of mutations would we expect to see at a particular frequency is known site frequency spectrum which I will refer to as $C(t)$. To derive

$C(t)$ requires integrating the above clone size distributions over time, weighted by the mutation rate per division μ, and the loss/replacement rate, λ.

$$C_n(t) = \int_0^t \lambda \mu N_s P_n(t-\tau) d\tau. \tag{6.3}$$

Performing this integral for the above we arrive at the following.

$$C_n(t) = \mu \sum_{k=1}^{N_s-1} \frac{\sin\left(\frac{\pi k n}{N_s}\right)}{\tan\left(\frac{\pi k}{2N_s}\right)} \left(1 - e^{-4\lambda t \sin^2\left(\frac{\pi k}{2N_s}\right)}\right) \tag{6.4}$$

$$C_{N_s}(t) = \mu\lambda t + \frac{\mu}{2} \sum_{k=1}^{N_s-1} \frac{(-1)^k}{\tan^2\left(\frac{\pi k}{2N_s}\right)} \left(1 - e^{-4\lambda t \sin^2\left(\frac{\pi k}{2N_s}\right)}\right)$$

To confirm that the above solution is accurate, a Monte Carlo simulation of the above dynamics was implemented and compared it to the predictions of the above equations. Figure 6.2 shows there is excellent agreement between the theory and simulation.

These equations can be simplified at long times, $\lambda t \gg N_s^2$:

$$C_n(t) = \mu(N_s - n) \tag{6.5}$$

$$C_{N_s}(t) = \mu(\lambda t - \kappa)$$

Where $\kappa = \frac{1}{6}\left(N_s^2 - 1\right)$. Finally, we can turn this into the probability distribution function by dividing by $\sum_{n=1}^{N_s} C_n(t) = \frac{\mu}{2} N_s(Ns-1)$. Simulations compared with this theoretical result are shown in Figure 6.2.

$$P_n(t) = \frac{2(N_s - n)}{N_s(N_s - 1)} \tag{6.6}$$

Arriving at a result first presented in Simons, 2016. Additionally, we can calculate the mean clone size which is given by:

$$\bar{n} = \sum_{n=1}^{N_s-1} n P_n(t) = \frac{1 + N_s}{3} \tag{6.7}$$

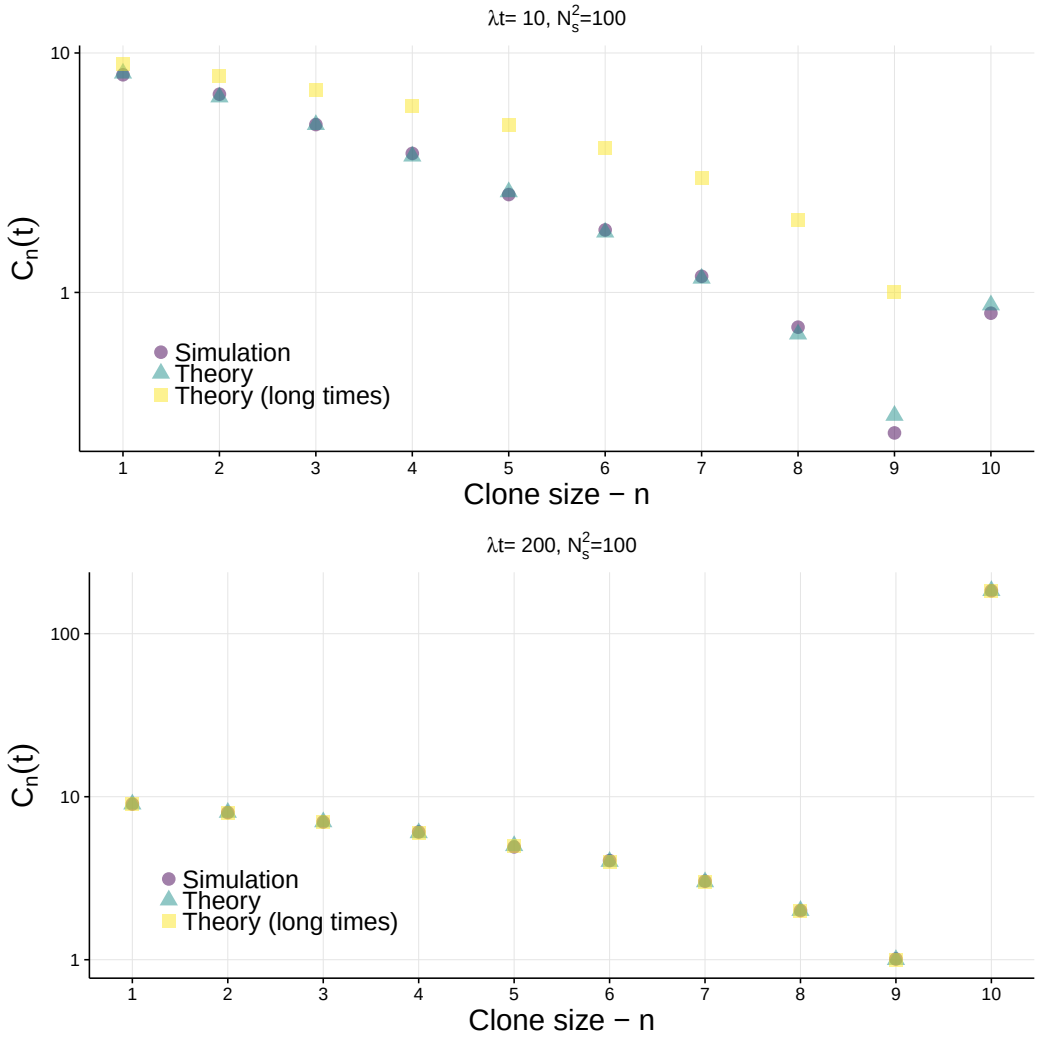

Figure 6.2: *Simulation and theory agree. Simulation parameters:* $\lambda = 1$, $\mu = 1$, $N_s = 10$, $t = 10$ *for the top panel and* $t = 200$ *for the bottom panel.*

This neutral drift model makes some simple predictions on what we would expect to observe in the sequencing data of crypts. i) A linear increase in the number of fixed mutations over time ii) A linear decrease in the number of partially fixed mutations as a function of frequency. Interestingly, the clone size distribution at long times only depends on the the number of stem cells N_s, and is invariant to the mutation rate, the loss replacement rate and time. Thus, given a sufficient amount of time, the clone size distribution for subclonal mutations reaches an equilibrium state where subclonal mutations are in effect *lost* through fixation but are compensated by the acquisition of new mutations.

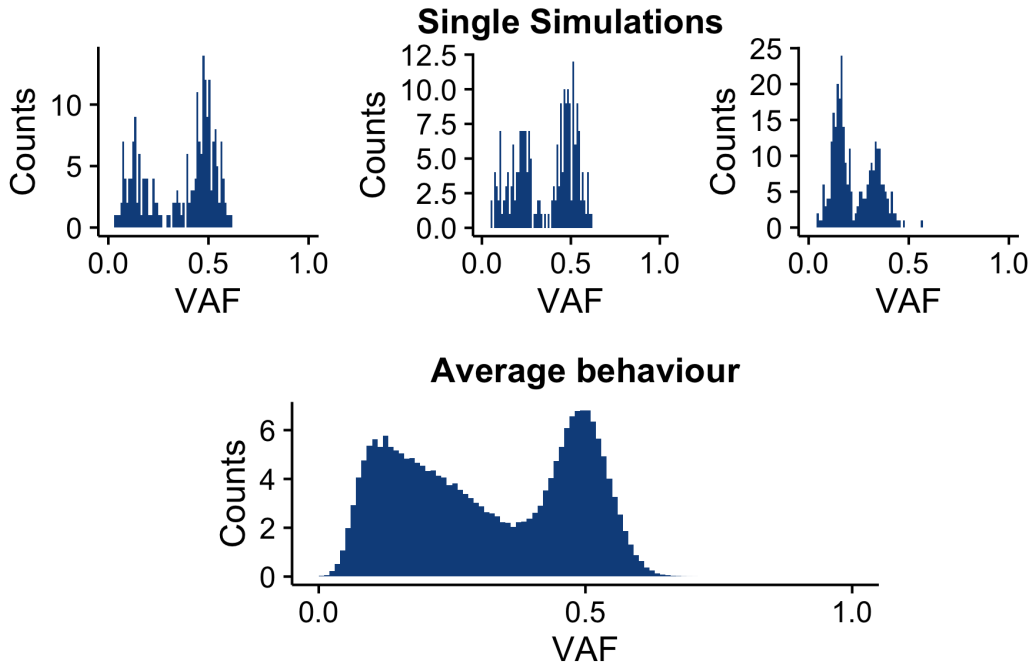

Figure 6.3: *Theory predicts that the distribution of partial mutation follows a linear dependence in clone size but that to identify the linear dependence requires averaging over many simulations.*

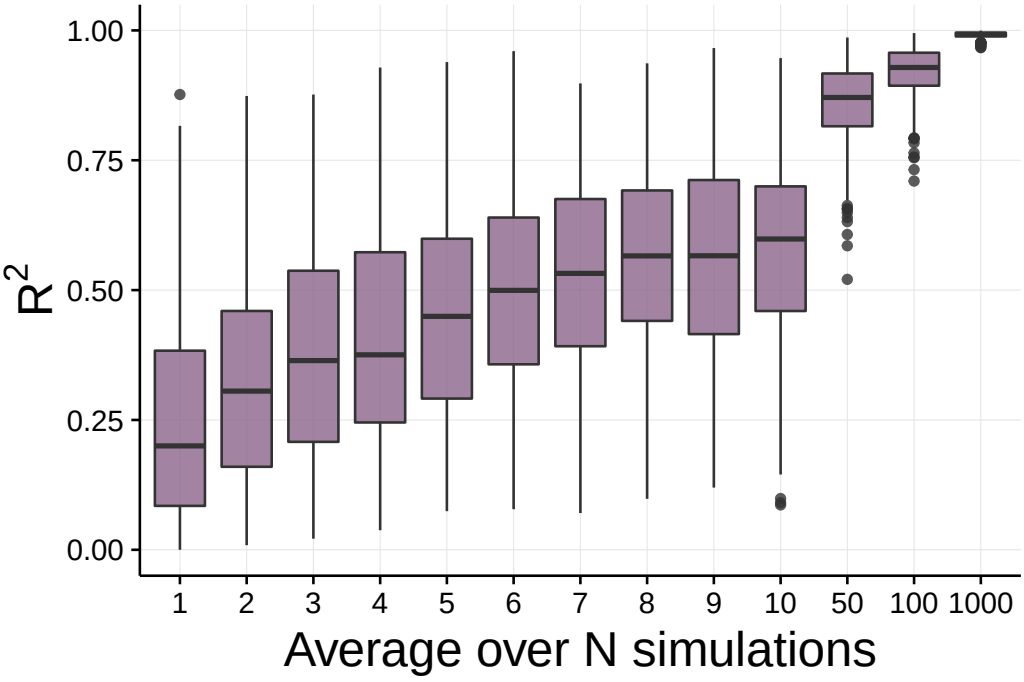

Figure 6.4: *Number of simulations needed to identify neutral drift dynamics. R^2 is the coefficient of determination for a linear model, so higher R^2 is a better fit for the neutral drift model. Averaging over more than 50 simulations robustly captures the neutral drift dynamics given the high R^2 values.*

Unlike previous lineage tracing experiments using fluorescent reporters that could build up the clone size distribution by averaging over many hundreds of repeats, sequencing data collected from a single crypt is a single realization of the underlying (stochastic) process, and captures the clone size distribution of one of these realisations at a single time point. Simulations of the process show that the data from a single crypt is unlikely to exhibit the features predicted by the analytical model of neutral drift. Simulations of the process vs the average behaviour over many simulations demonstrate that single realisations can look very different, while the average behaviour follows the predictions of the analytical results, ie a linear decrease in the number of mutations as a function of frequency, see Figure 6.5. To resolve the clone size distribution will therefore require averaging over many single crypts. Simulating the process N times then averaging the simulations shows that the linear relationship can only be identified robustly by averaging over 10's of simulations, Figure 6.4.

6.3.1 Expected number of mutations

Despite the limitations of requiring data from multiple crypts to resolve the clone size distribution, we can nevertheless get a sense of what the data from a single crypt would be expected to look like given some reasonable parameters. Table 6.2 shows inferred neutral drift parameters from 5 studies. Based on these values I used $\lambda = 0.1/day$ and $N_s = 8$ to simulate the neutral drift process with a mutation rate of $\mu = 10^{-9}$ per bp per division in equivalent whole exome sequencing and whole genome sequencing. Single realisations and the average behaviour are shown in Figure 6.5, demonstrating that we would expect to see many more clonal (fixed) mutations compared to subclonal mutations. Simulations were implemented using a Monte Carlo simulation of the neutral drift model where at each division mutations can accumulate followed by synthetic sequencing of the simulated mutation data as described in the previous chapter.

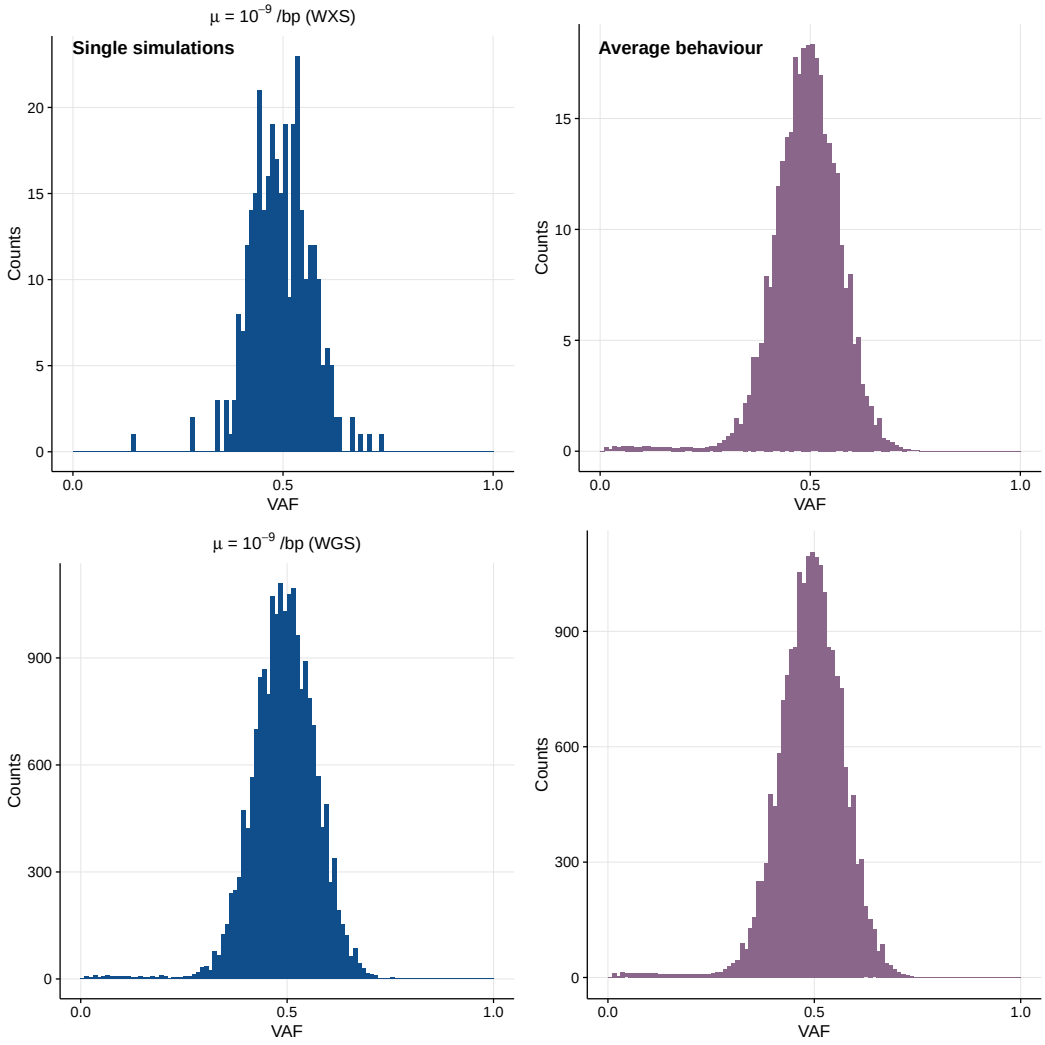

Figure 6.5: *Expected VAF distribution based on the classic neutral drift model with the following parameters: $N_s = 10$, $\mu = 10^{-9}$ per bp per division and $\lambda = 0.5$ per day (parameters based on previous estimates, see Table 6.2. Top figures are for whole exome sequencing data and bottom figure is for whole genome sequencing data. Average is over 100 equivalent simulations.*

6.3.2 Single crypt data

With some intuition on what we would expect the data to look I'll now return to the crypt sequencing data. In summary from the analytical derivation of the neutral drift model and simulations of the process we would expect to observe the following features in the sequencing data.

1. Large number of clonal mutations

2. Small number (relative to clonal mutations) of sub clonal mutations

N_s	λ (/day)	Ref.	Info
16	1.0	Lopez-Garcia *et al.,* 2010	Mouse model
5-7	0.1-0.3	Kozar *et al.,* 2013	Mouse model
8	0.24	Ritsma *et al.,* 2014	Mouse model
6	0.3 - 2.0	Baker *et al.,* 2014	Human tissue
-	0.1	Blokzijl *et al.,* 2016	Calculation

Table 6.2: *Inferred number of stem cells (N_s) and loss/replacement rate (λ) of neutral drift process from various studies. λ from Blokzilj study was inferred based on the reported value of 30 mutations per genome per year in the colon, a conservative estimate of the mutation rate of 3 per genome per division and equation [ref equation]*

3. Linear decrease in the number of subclonal mutations as a function of VAF

4. Linear increase of clonal mutations over time

Figure 6.6 shows the sequencing data from 2 crypts from 2 people with a healthy colon. Exome sequencing was used for sample 450 and whole genome sequencing was used for sample 452. The ages of the individuals were 45 and 50 respectively. Intriguingly none of the above predictions from the neutral drift model seem to be confirmed in this data. In particular there appear to be almost no clonal mutations (expected to be at VAF=0.5) and there are a large amount of subclonal mutations, the exact opposite of what would be expected from the neutral drift model.

There may be some technical issue due to the low quantity of DNA extracted from such a small tissue sample that could account for this. However this seems unlikely given that in these two samples, two different sequencing strategies were observed. We also observe similar mutation burdens across all of the crypts in exonic regions, see Figure 6.7.

Another explanation is that there are copy number changes which distort the relationship between lineage size and VAF (ie lineage size $\neq 2\times$VAF). The sequencing data however shows that crypts from both the whole exome sequencing experiment and the whole genome sequencing experiment are in a normal diploid state, see Figure 6.9. Furthermore the B-Allele Frequency (BAF) from germline SNPs cluster around 0.5 which also discounts the possibility of preferential capture

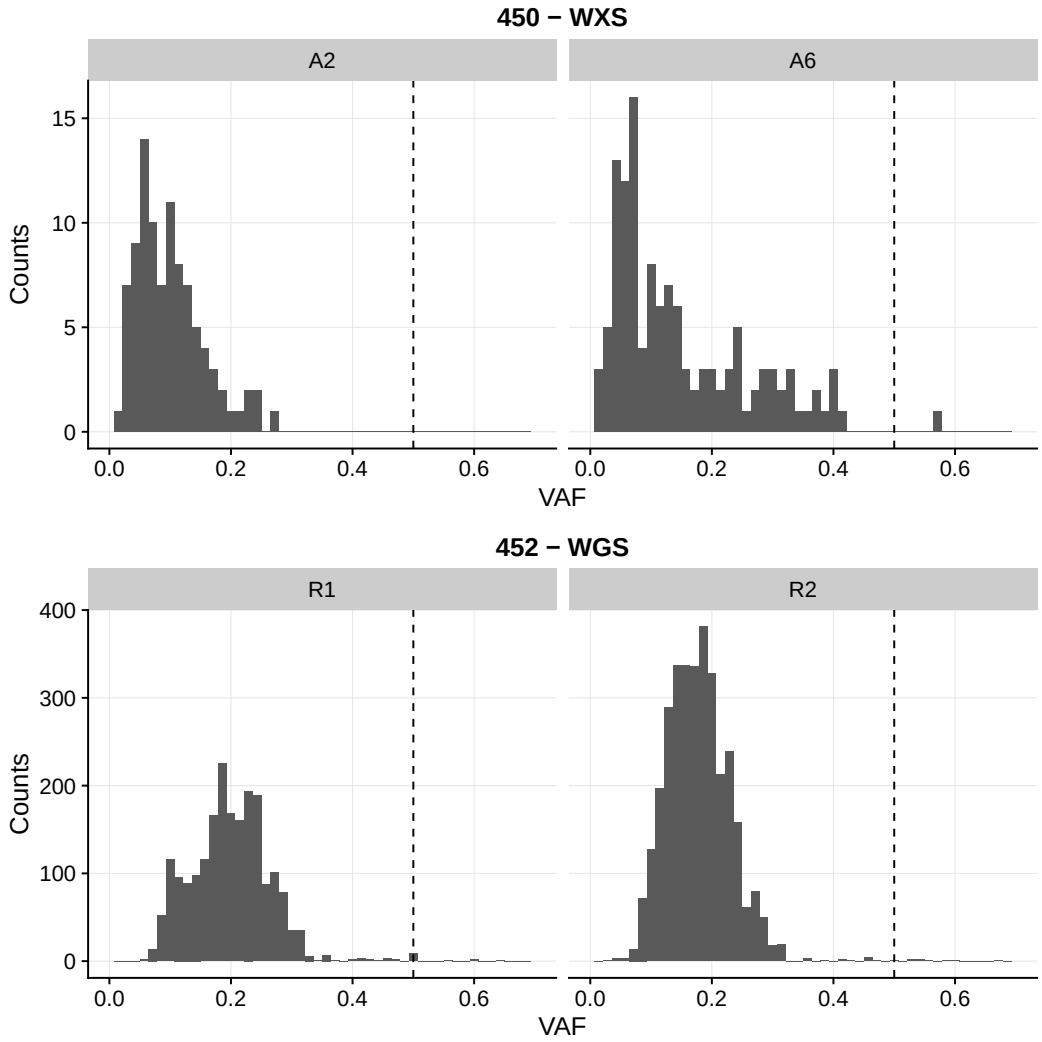

Figure 6.6: *VAF distribution for normal crypts from 2 patients*

of some alleles which may also result in skewed VAF distributions.

Low purity could account for the apparent low number of clonal mutations, where what appear to be subclonal mutations are in fact fixed mutations. Low purity could arise from large amounts of stromal tissue or immune cells being sequenced at the same time resulting in DNA from epithelial cells being reduced. The degree of contamination necessary such that most of the mutations that appear subclonal are in fact clonal in the crypt is $>50\%$ which also seems unlikely.

Assuming there are no technical issues with this data the other possibility which I will now explore in more detail is whether in fact the neutral drift model is inadequate. In all previous studies that have demonstrated neutral drift dynamics in

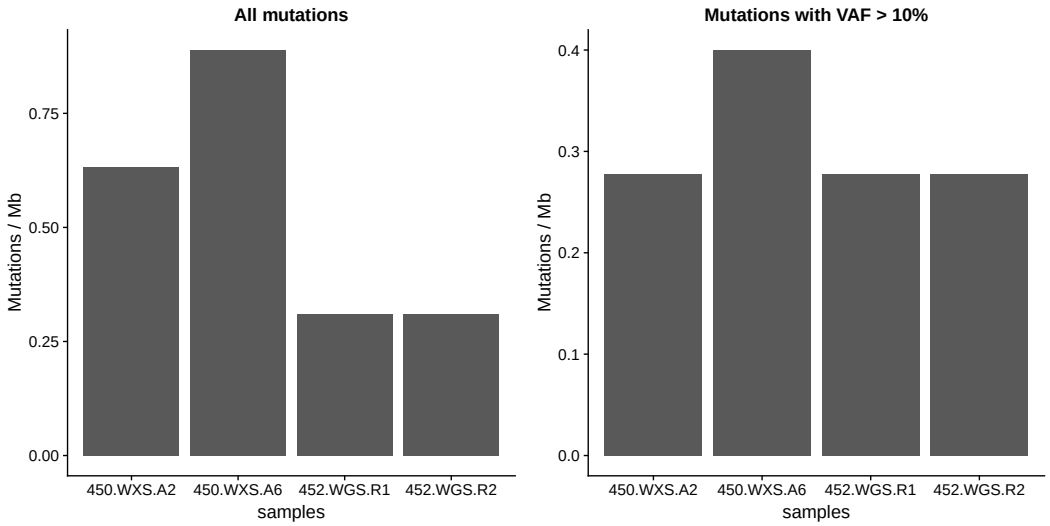

Figure 6.7: *Mutation burden in the 4 crypts. We would expect the higher depth exome sequencing to pick up more low frequency mutations, if we consider only mutation above frequency 10% which should be detected equally in both assays then then mutation burdens are highly similar across assays and patients.*

the colon, only a small number of labels (< 4) in a single crypt have been utilized, in a sense this is only informative on the expected clone size distribution at time scales over which these labels are lost or fixed. If for example there is some underlying dynamics that occur over longer time scales, then these methods may not have the temporal resolution to challenge the neutral drift model. One way that we could imagine altering the neutral drift dynamics so that the dynamics is altered at long times compared to short times is by introducing a population of slow cycling stem cells that are at the tip of the stem cell hierarchy. The contribution from these cells would be minimal over short times, but at long times such as the lifetime of a crypt may be significant. A slow cycling stem cell population could sporadically purge the crypt of mutations as due to the limited number of divisions a slow cycling stem cell would experience it would accumulate a limited number of mutations. This would therefore provide a powerful mechanism to further reduce the accumulation of mutations within the crypt which appears to be what we observe in the data; in homeostasis crypts seem to be able to minimise the acquisition of *fixed* mutations within the crypt.

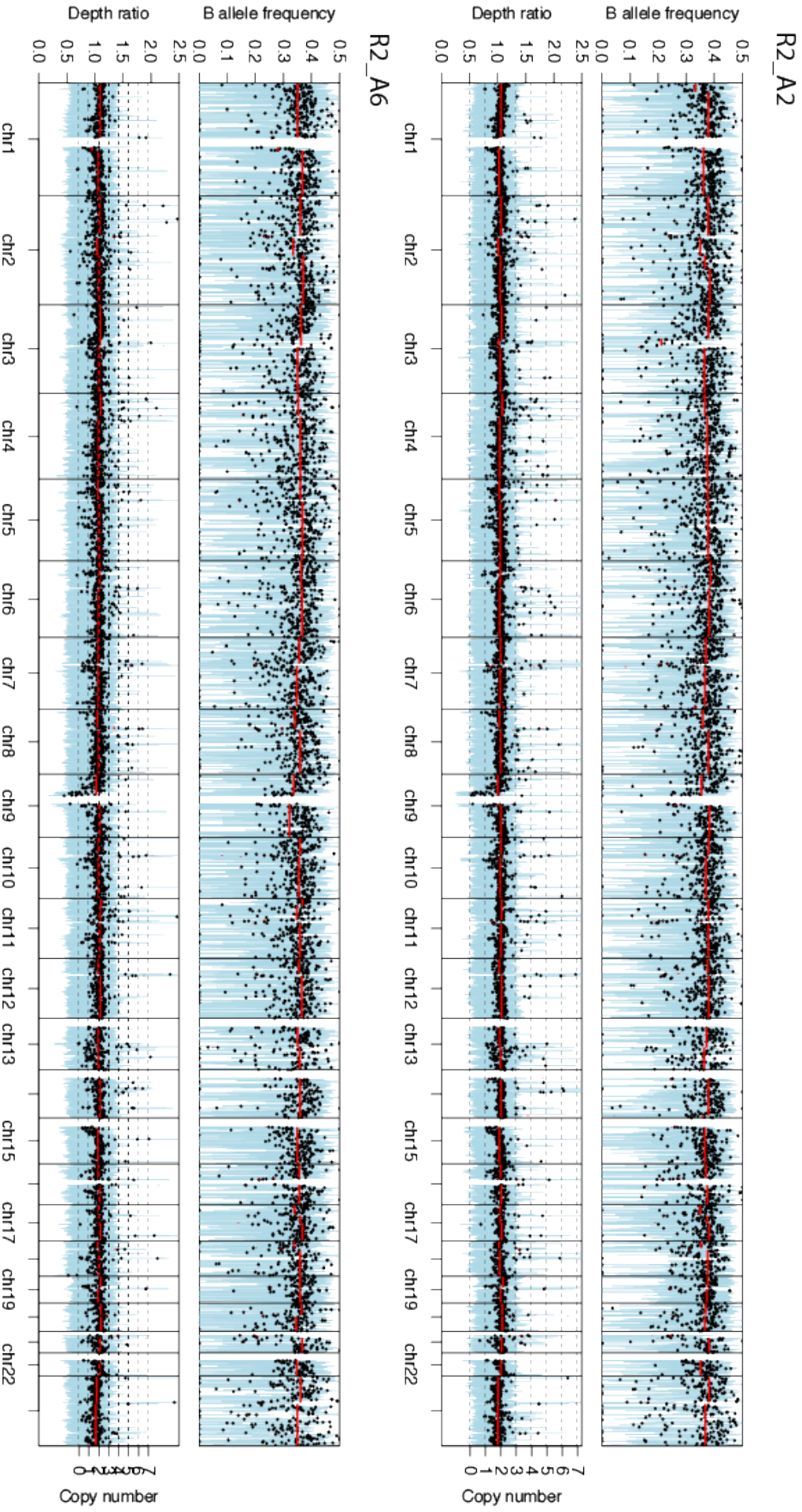

Figure 6.8: *Sample 450 (WXS) normal crypts copy number. Sequenza output showing BAF and depth ratios between crypt sequencing data, both are consistent with a normal diploid genome*

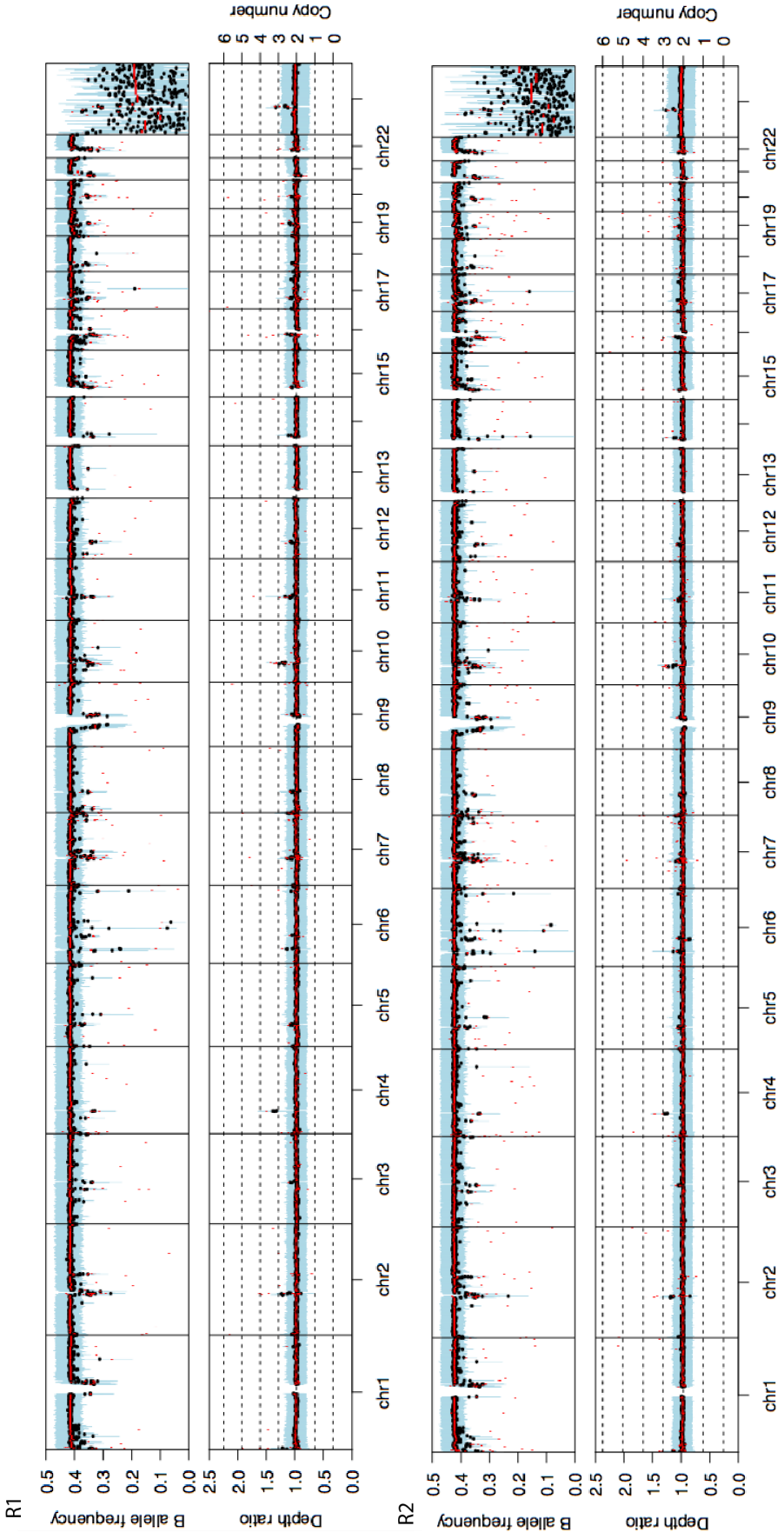

Figure 6.9: *Sample 452 (WGS) normal crypts copy number. Sequenza output showing BAF and depth ratios between crypt sequencing data, both are consistent with a normal diploid genome*

6.4 Slow cycling stem cells

I'll consider two models of slow cycling stem cells, the first will consider a single master stem cell which divides infrequently. In the second model I'll consider a model with multiple slow cycling stem cells which themselves undergo neutral drift, see Figure 6.10 for a schematic.

6.4.1 Single slow cycling *Master* stem cell

In the case of a model with a master stem cell the number of master stem cell division after a time t will follow a Poisson distribution with mean λ_M, the turnover rate of the master stem cell.

$$P_m(t) = \frac{(\lambda_m t)^n e^{-\lambda_m t}}{n!} \tag{6.8}$$

The expected number number of mutations accumulated in the master stem cell lineage C_m is then simply the product of the mutation rate per division μ and the mean of the above distribution ($= \lambda_m t$).

$$C_m(t) = \mu \lambda_m t \tag{6.9}$$

Thus we would expect a linear increase in mutations over time. As the master stem cell will never be lost this equation will describe the dynamics over long time scales. However over short time scales we would expect mutations to accumulate

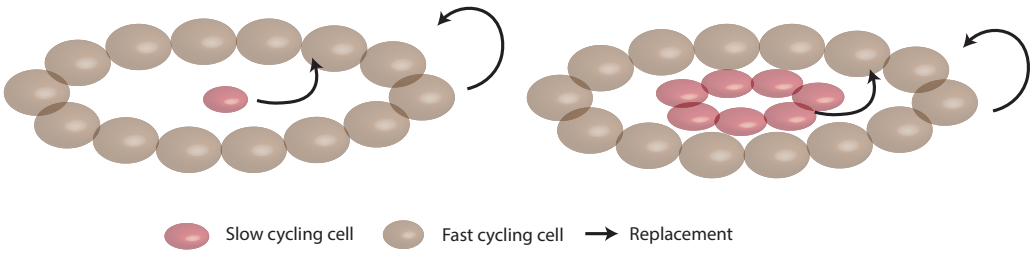

Figure 6.10: *Schematic of two alternate stem cell models. Brown cells are equipotent fast cycling stem cells, red cells are slow cycling cells which periodically replace the fast cycling cells. The model on the left has a single slow cycling cell, the model on the right has a number of slow cycling cells which can replace each other.*

Algorithm 5: Master stem model

input : Turnover rate of master stem cell (λ_M) loss/replacement rate of drifting population of stem cells (λ_s, λ_f), mutation rate per division μ and numbers of cells in slow and fast cycling population (N_s, N_f). Time when simulation ends t_{end}

output: List of mutations present in each cell

Initiate simulation with a single master stem cell and N_{drift} drifting stem cells. set time to $t = 0$

while $t < t_{end}$ **do**

1. randomly sample a cell
2. cell divides and replaces one of its neighbours with probability λ
3. if cell divides, daughter cells acquires $P_o(\mu)$ new mutations
4. if master stem cell divides, one of the daughter cell displaces a randomly chosen cell from the drifting compartment
5. update time $t = t + \Delta t$, where Δt is an exponentially distributed random variable

in the faster cycling neutral drift compartment. Any mutations exclusive to the neutral drift compartment will then be sporadically purged from the population so that mutation accumulation is determined by equation (6.9) over long times, the rate of these purges will depend on the ratio, R between the loss replacement rate in the neutral drift compartment and the turnover rate of the master stem cell: $R = \frac{\lambda}{\lambda_m}$. We can see this in Figure 6.11, which shows the mutation accumulation over time in a simulation of this process (simulation is described in Algorithm 5) , the thick black line show the predictions of equation (6.9) and the lighter coloured lines are single realisations of the simulation.

It is clear that this type of model can indeed provide an additional mechanism to suppress the accumulation of mutations but the questions remains whether it can explain the crypt sequencing data. Figure 6.12 which shows what we would expect data from this model to look like suggests not. While the total number of mutations is decreased, the number of clonally fixed mutations again far outweighs the number of subclonal partially fixed mutations, thus the master stem cell model also seems

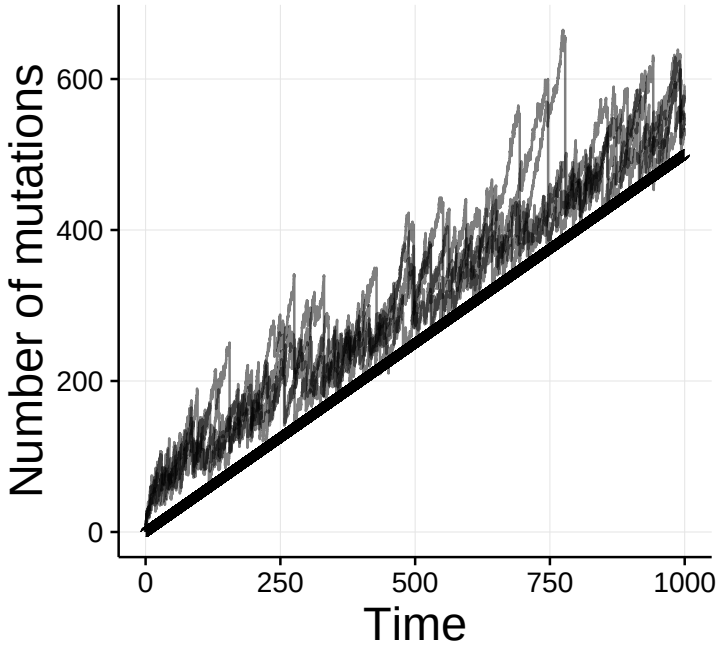

Figure 6.11: *Dynamics of mutation accumulation over time in the master stem cell model. Mutations accumulate in the fast cycling population of cells but are sporadically purged from the population when the slow cycling cell divides and replaces a fast cycling cell. If this cell drifts to fixation mutations accumulated in the fast cycling population are purged from the crypt.*

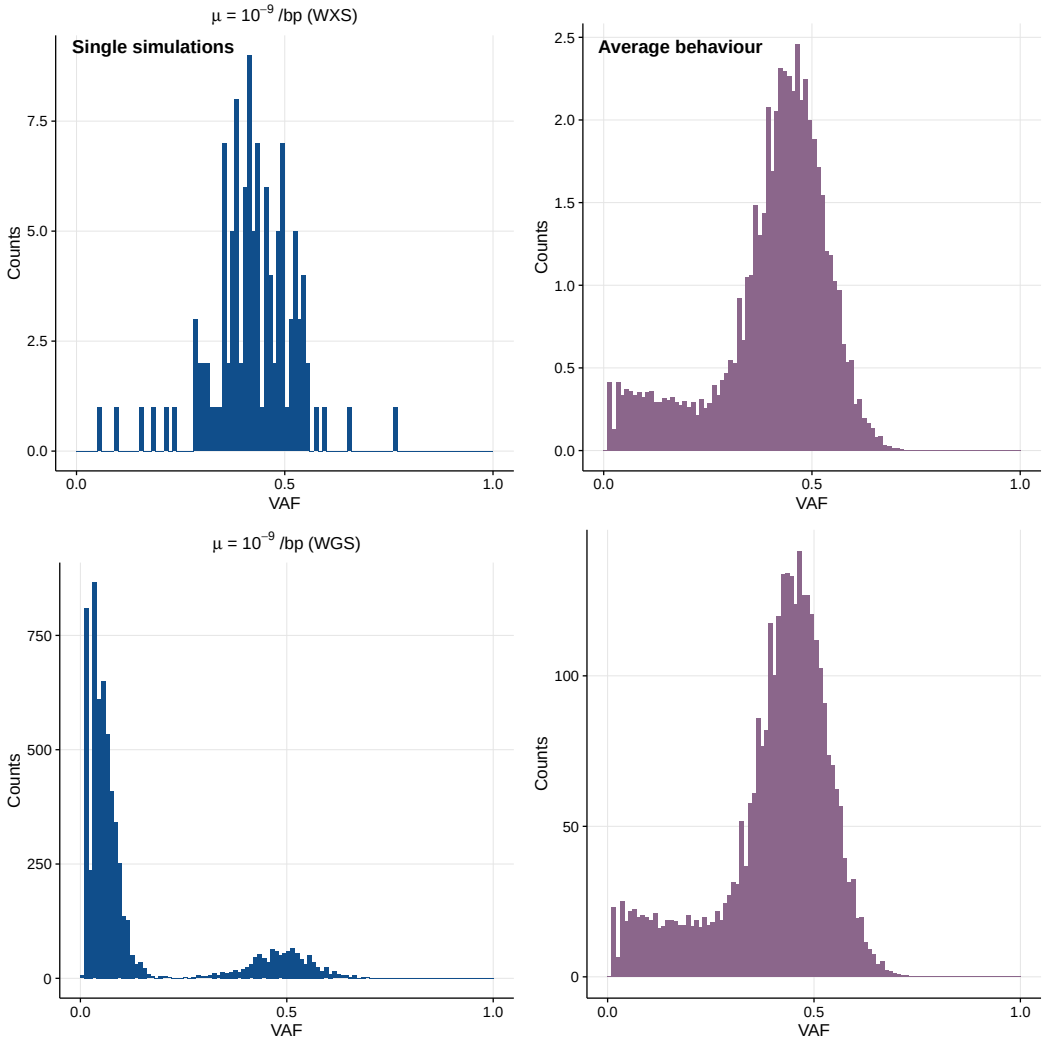

Figure 6.12: *Simulation parameters* $\mu = 10^{-9}$ *per bp (multiplied by* 45×10^6 *for WGS and* 3×10^9 *for WGS.*$N_{drift} = 8$. *Time of simulation* $= 45$*years.* $\lambda_{master} = 0.005$ *per day,* $\lambda_{drift} = 0.1$ *per day. Average simulations are from 1000 repeat runs.*

Algorithm 6: Hierarchical drift model

input : Loss replacement rate of slow and fast cycling population (λ_s, λ_f), mutation rate per division μ and numbers of cells in slow and fast cycling population (N_s, N_f). Time when simulation ends t_{end}

output: List of mutations present in each cell

Initiate simulation with N_s slow cycling stem cells and N_f fast cycling stem cells. set time to $t = 0$

while $t < t_{end}$ **do**

1. randomly sample a cell
2. cell divides and replaces one of its neighbours with probability λ
3. if cell divides, daughter cells acquires $P_o(\mu)$ new mutations i
4. f slow stem cell is displaced from slow compartment, the cell displaces a randomly chosen cell from the fast compartment
5. update time $t = t + \Delta t$, where Δt is an exponentially distributed random variable

to be unable to explain the patterns observed in the real data. Although it is possible to imagine a scenario that produces patterns more similar to what we observe by tweaking the mutation rate and replacement rates of the respective populations.

6.4.2 Hierarchical drift model

Rather than having a single slow cycling stem cell we'll now consider a two compartment model with a slow cycling population of cells and a fast cycling population of cells where both cell types undergo neutral drift to stochastically replace each other. It is straightforward to recognise that this could provide a further mechanism to restrict the accumulation of mutations as mutations would first need to fix in the slow cycling stem cell population before fixing in the fast cycling population. If the loss replacement rate in the slow cycling population is very slow this process can take a very long time. Furthermore we could still see similar dynamics as with the master stem cell model with mutations being sporadically purged when a division in the slow cycling population introduces a cell with limited mutations into the fast cycling population. Details of the algorithm used to simulate this process is described

in Algorithm 6.

Figure 6.13 shows what we would expect data from this model to look like, this model does show a reduction in clonal mutations while maintaining a reasonable number of subclonal mutations, this suggests that different stem cell architectures to the current standard view may be able to explain the data we observe. More data will be needed to resolve these models. Here, I have only demonstrated expectations from the model using a single set of parameters, the model will be sensitive to the choice of parameters. In particular the loss/replacement rates in the respective compartments can have a large effect on the distribution of mutation frequencies.

6.5 Discussion

In this chapter I've investigated stem cell dynamics in the human colonic crypt using sequencing data from two individuals of similar ages with healthy bowels. Intriguingly the clone size distribution reported by deep sequencing of crypts from these individuals did not conform to a model where a handful of equipotent stem cells reside at the bottom of the crypt and stochastically replace each other. This represents a paradox with previous studies which have demonstrated these dynamics in mice (Lopez-Garcia *et al.,* 2010) and humans (Baker *et al.,* 2015). However one mechanism that can reconcile these observations is the introduction of a population of slow cycling stem cells such that over short times the dynamics in the fast cycling population is important while at long times, ie over the life time of the crypt the dominant feature is the contribution of the slow cycling cells. Two recent studies support the idea that the long term dynamics of stem renewal in the colon are controlled by a slow cycling population of stem cells. Both studies estimated a loss/replacement rate of less than once per year (Nicholson *et al.,* 2018; Stamp *et al.,* 2018), which is a similar value to what was used in my alternative models for the slow cycling stem cell population. It is perhaps not unsurprising that stem cell dynamics in the human colon differs radically from the its murine counterpart given the large difference in life spans.

Further data will likely enable us to further refine these models, in particular

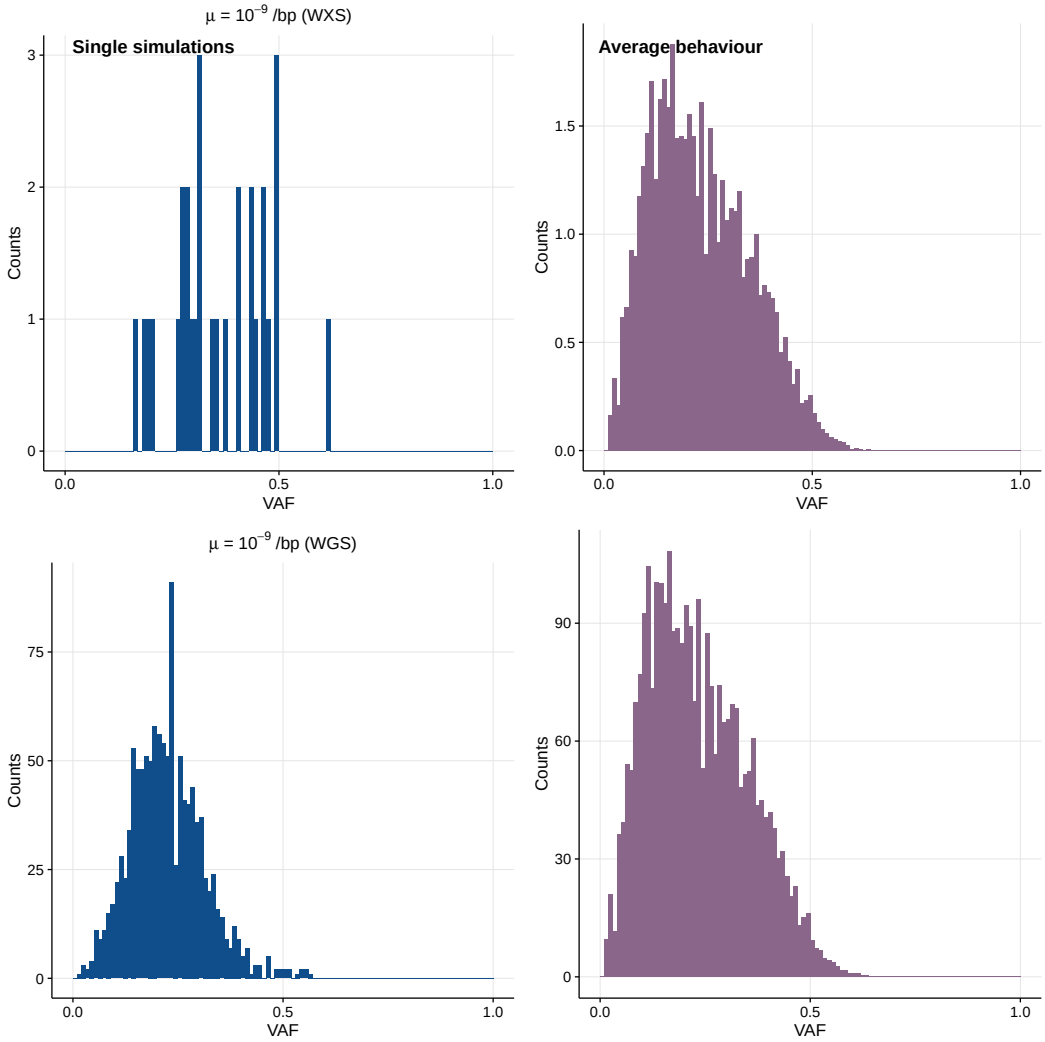

Figure 6.13: *Simulation parameters* $\mu = 10^{-9}$ *per bp (multiplied by* 45×10^6 *for WGS and* 3×10^9 *for WGS.*$N_s = 8$,$N_f = 30$. *Time of simulation* $= 45$ *years.* $\lambda_s = 0.002$ *per day,* $\lambda_f = 0.1$ *per day. Average simulations are from 1000 repeat runs.*

collecting samples from individuals of different ages will provide another means to test the predictions of the models. With this it will then be possible to probe how these dynamics change across disease states as the tissue progresses toward cancer which will be the focus of future work. With more data it should be possible to fit the models to the data and estimate plausible value for the parameters of the model such as the number of stem cells in the different compartments and their replacement rates. Furthermore sequencing of crypts across diseases states may elucidate how particular mutations change these dynamics and prime tissues for cancer initiation.

6.6 Acknowledgements

The single crypt sequencing data was generated by various members of the Evolution and Cancer lab including Ibrahim al-Bakir, Laura Gay, Annie Baker and Chris Kimberley.

Chapter 7

Summary and outlook

7.1 Summary

The aim of this thesis was to investigate evolutionary dynamics in the somatic evolution of human tissues with a particular focus on elucidating the role of selection in carcinogenesis. While there is little doubt that selection for certain phenotypes, or hallmarks of cancer (Hanahan & Weinberg, 2011) is what causes physiologically normal tissue to transform and become malignant, selection is often assumed rather than formally tested in cancer evolution. Furthermore, little is known about how and when natural selection operates in cancer evolution. Given that such alterations are the crucial determinants of how cancers evolve and adapt to their environment a better understanding of these dynamics critically important.

The field of population genetics provides a natural way to test for and quantify signatures of selection and has been applied successfully in species evolution. Population genetics, in simple terms is a mathematical description of how allele frequencies are expected to change due to the fundamental evolutionary forces selection, drift, mutation and recombination (not applicable in asexual evolution such as cancer). The field of population genetics is one of the few areas of biology whose foundations are quantitive theories. This is partly due to necessity, evolution proceeds over long times and cannot in most cases be observed directly. Similar issues exists in cancer where due to obvious ethical issues it is impractical to observe cancer evolution proceeding over time. In both scenarios the use of population genetics

allows for inferences on the past evolutionary dynamics from present day genomes. The aim of this thesis was to translate concepts and approaches that have proved successful in traditional population genetics to cancer evolution, in particular using the VAF distribution, the analog to the site frequency spectrum to uncover signatures of selection and therefore signatures of neutrality.

Despite this long history and the wealth of theoretical models with which to draw upon, for the most part traditional population genetics cannot be directly applied to the cancer setting due to cancer evolution not satisfying a number of assumptions commonly employed in classical population genetics. Much of the theoretical models assume fixed size populations while tumours evolve through a process of clonal evolution where the population grows over time. Furthermore because cancer evolution is asexual there is no recombination and mutations can hitchike (Gillespie, 2000; Fay & Wu, 2000). Thus a population genetic description of cancer evolution requires the theories to be adapted to account for these difference. Fortunately much of these models can be developed by drawing upon theories of branching processes and the Luria-Delbrück distribution.

In molecular species evolution, neutrality is the null hypothesis. In this paradigm it is assumed that neutrality (evolution in the absence of selection) adequately explains genetic diversity and changes in allele frequencies unless this null can be rejected. In cancer evolution, selection is often assumed rather than formally tested in this way, but neutrality can also provide a useful null in cancer evolution (Wu *et al.,* 2016). Having a null model to explain the genetic diversity which is observed across all cancer types is useful because it is a rigorous way to identify genetic alterations under selection, these being the important alterations that drive the disease and are responsible for resistance to treatment. With this in mind we developed a null model of cancer evolution which could be applied to the frequency spectrum of single nucleotide variants reported by deep sequencing experiments in cancer. Our null neutral model (described in detail in the first results chapter) assumes all mutations are neutral, exponential growth of the tumour and Poissonly distributed mutations. The expression we arrive at, that under neutrality

the cumulative number of mutations at a frequency f follows a $1/f$ distribution has been derived by others previously (Griffiths & Tavaré, 1998; Keller & Antal, 2015; Durrett, 2013b), these results are closely related to the Luria-Delbrück distribution (Kessler & Levine, 2014). Somewhat surprisingly this neutral model appeared to be a good description of the data in approximately 30% of the cancers we investigated. Other studies have also been unable to reject neutrality as a description of the available data in multi region sequencing studies (Ling *et al.,* 2015; Sun *et al.,* 2017), and a previous study showed that selection is suppressed in growing populations where the dynamics become effectively neutral (Sottoriva *et al.,* 2015). Hopefully the recognition that neutral evolutionary dynamics may often be a plausible explanation of genomic data in cancers will result in the research community testing for signatures of neutral evolution which is the natural null model rather than assuming selection. Selection for malignant phenotypes is what drives cancer so caution is warranted in ascribing selection to patterns of genomic diversity. I hope that a rigorous framework rooted in population genetics would enable more robust identification of patterns of selection.

What appeared initially as a somewhat surprising result begs the question, why do we find such a large proportion of cancers to be consistent with the neutral model? Firstly we must recognise some limitations of the data, bulk sequencing data will only report on mutations that are at high frequency and present in a number of cells on the order of a million. Thus there may well be interesting non-neutral evolutionary dynamics below this scale. In other words neutral dynamics may be a very good description at a macro scale of millions of cells but at a lower micro-scale of thousands or fewer cells selection may be in important force. Single cell sequencing technologies will be useful in unravelling these micro-scale dynamics in the future. Another reason for the preponderance of neutral tumours related to this and touched upon in Chapter 4 is that in growing populations, selection is attenuated. For a lineage to reach an appreciable frequency new mutational lineages are always playing catch up with pre-existing lineages. Put differently, the time scale of evolution is important, where mutations that appear late during tumour evolution

will start at a very small frequency in the population and may not have enough time to reach a detectable frequency even if highly selected. Another hypothesis is that the tissue structure has a role in suppressing selection. Colon cancers for example are composed of glands which contain on the order of 10,000 cells and colon cancers are thought to grow via gland fission. For mutations that possess a fitness advantage to spread in the population they must first fix in the gland before expanding via gland fission. This presents multiple barriers for mutations to spread and may suppress the effects of selection. Stomach and breast tissues amongst others have similar glandular architectures.

Although a sizeable number of tumours were found to be neutral in Chapter 3, a large number did not fit this model. This led to the natural question, what is happening in the other 70% of cancers. Could we learn something more about the process of natural selection in these cancers? To tackle this question I developed two approaches, one which could uncover selection on a sample by sample basis and a second which leverages data from cohorts of sequenced cancers to look at population level selection.

For the first approach I again used population genetics results from asexual organisms (Hartl & Clark, 2007) and developed a model of how the frequency of a selected subclone would be expected to change over time depending on its relative fitness advantage and the time the subclone emerged. By linking these dynamics to mutations accumulating in different lineages within the tumour I observed there would be characteristic patterns of subclonal selection observable in deep sequencing experiments. These patterns result in the frequency distribution deviating from the neutral prediction. Using a branching process simulation of cancer evolution and fitting this to the frequency spectrum using Bayesian inference I was then able to extract the necessary information to infer the relative fitness and time of emergence of subclones for non-neutral tumours. This approach measured strikingly high fitness advantages ($>$20%) for subclones under selection.

For the second approach I turned my attention to another method developed in molecular species evolution used to identify signatures of selection that has re-

cently been adapted for cancer (Weghorn & Sunyaev, 2017; Martincorena *et al.,* 2017), dN/dS the ratio of non-synonymous to synonymous mutations. The theoretical foundation of dN/dS assumes long evolutionary time and fixed size populations and does not take into account subclonal mutations of which there are many in cancer. By adapting the Luria-Delbrück distribution and taking the ratio of the expected distribution for selected mutations and neutral mutations I showed how dN/dS values would be expected to change across the frequency spectrum. This also allows for estimating the fitness effect across cohorts of cancers and uncovers how dN/dS is related to the relative fitness in growing populations. Differently from the first method this approach allows for identifying recurrent patterns of selection across cohorts of cancers but is limited in terms of quantifying evolutionary dynamics on a sample by sample basis. Future work may be able to combine these somewhat distinct approaches, by linking repeated evolutionary trajectories at the cohort level to selection for certain mutational lineages in individual patients.

The first 3 results chapters all point to selection for subclones in the cancers we ultimately observe being rare, much of the genetic alterations most likely occur in physiologically normal and pre-malignant tissue. With a view to this the last results chapter investigated population dynamics and mutation accumulation in physiologically normal tissue, in particular the colonic crypt. Given that cancer can be thought of as a perturbation of these dynamics, it is important to get a good understanding of how homeostasis is orchestrated and how genetic alterations may perturb this. Surprisingly, the data appeared at odds with the traditional view of stem cell dynamics within intestinal crypts. Evidence from lineage tracing in mouse models and in humans suggests that 6-8 stem cells reside at the base of the crypt and undergo a process of neutral drift. The sequencing data appeared at odds with this model given the abundance of partially fixed mutations and lack of fixed mutations, the opposite of what would be predicted from the neutral drift model. Other models were proposed that provide mechanisms for fixation of mutations to be strongly suppressed, however more data is required as these observation came from a limited number of samples and technical issues with the assay cannot completely be ruled out. The

approach outlined in this chapter provides a quantitative methodology to infer stem cell dynamics from sequencing data. In future this approach can be extended to study the effects of tumorigenic mutations and across disease states and understand how stem cell dynamics is modified.

In summary this thesis can be viewed as an investigation of the evolutionary dynamics of cancer at multiple scales. Firstly I looked at the scale of a bulk tumour, finding that many cancers exhibit patterns of neutral evolution in Chapter 3 and measuring the fitness advantage of detectable subclones in Chapter 4. Chapter 5 looked at a population scale, using dN/dS across cohorts of cancer samples as a readout of selection. Finally the last chapter, Chapter 6 investigated the dynamics of stem cells in the colon, ie looking at the dynamics at cellular resolution. Understanding evolution at all these scales, the tumour level, population level and cellular level will be required for a deeper understanding of tumour evolution. In the next and final section of the thesis I'll discuss a few avenues that I think may be fruitful in utilising quantitative measurements of these evolutionary dynamics in a clinically impactful way.

7.2 Outlook

This thesis has shown how quantitive theories together with genomics can be used to study evolution and population dynamics in human cancers and tissues. For evolutionary theory to have an impact clinically, quantitative approaches will likely be paramount. A small number of studies have already shown that evolutionary theory can help guide clinical decision making. Zhang *et al.,* 2017 showed that adaptive therapy, a treatment strategy inspired by ecology and evolution where the goal is to maintain the coexistence of resistant and susceptible cancer population so that they mutually repress each other significantly increases survival in prostate cancer therapy. A number of clinical trials along similar lines are currently in progress for different cancer types. Łuksza *et al.,* 2017 showed that using a model to measure the fitness of subclones based on their neoantigen burden was able to predict patient response to immunotherapy, and predictions were better for the model based approach

compared to solely using the mutation burden. Such studies suggest that application of evolutionary principles to treatment may be effective. Both these studies used quantitive models of cancer evolution. For such approaches to realise their potential, being able to quantitively measure the evolutionary dynamics in tumours will be paramount as these studies demonstrated. For the remainder of this chapter and thesis I'll discuss some areas where the application of theoretical models to cancer genomics might be fruitful.

A longstanding goal in evolutionary biology is to measure the distribution of fitness effects, this has been relatively unexplored in cancer where the importance of mutations tend to be ranked by their frequency in large cohorts. Recent applications of dN/dS to cancer go some way toward this goal and showed that mutations in some genes are highly selected (Martincorena *et al.,* 2017; Weghorn & Sunyaev, 2017). These studies however did not uncover much signal of negative selection, which appears at odds with the presentation of neo-antigens resulting in increased immune predation (McGranahan *et al.,* 2016) and that the burden of neo-antigens correlates with the efficacy of immunotherapies (Yarchoan *et al.,* 2017). Furthermore consideration of the population dynamics is important for interpreting dN/dS, particularly for subclonal mutations as discussed in Chapter 5. Further work is needed to unravel these complexities, experimental model systems may be useful here. Rogers *et al.,* 2018 for example used CRISPR/Cas9 constructs to engineer tumours with particular mutations in mice and measured their fitness with deep barcode sequencing. Extensions to this approach to quantify fitness of additional mutations and epistatic interactions between different driver mutations would be intriguing avenues to pursue. Another important aspect to consider is that fitness effects of mutations undoubtedly change under different environments, particularly in the context of treatment. Measuring how fitness effects change in different environments, particularly under treatment may provide novel insight into how treatment changes the fitness landscape in tumours. It would then be possible to elucidate if mutations that induce resistance tend to be pre-existing in the cancer and possibly the number and mechanism of resistance of these mutations. More complex models

are likely needed to unravel the complexities of these dynamics fitness landscapes or seascapes.

One way of viewing personalised medicine is how to predict and thus modify the treatment of an individuals cancer. From an evolutionary point of view this can be thought of as how to predict the evolutionary trajectory of a cancer and then crucially how this can best be perturbed for the benefit of the patient. The degree of predictability of evolutionary systems remains an open question in much of evolutionary biology (including cancer) (Lässig *et al.,* 2017). Being able to predict and then modify the treatment of an individuals cancer will require progress on all of the fronts discussed so far. From quantifying the distribution of fitness effects of mutations and how this changes across environments, to elucidating how resistance emerges and its mechanism. Such questions are perhaps beginning to be achievable.

As an illustration of how the work in this thesis may fit into this paradigm, in Chapter 4 I demonstrated how measuring the fitness of subclones in theory allows for predicting which subclone will come to dominate at some point in the future. This was only demonstrated in an *in silico* tumour however, so further work is needed to validate this approach in an experimental system. Furthermore, this type of analysis is somewhat limited as it assumes a constant environment and does not take into account the effects of treatment or the emergence of newer subclones. This type of approach could in theory be adapted to take into account changes in selective pressures due to treatment. Another interesting avenue may be to combine the population level dN/dS selection measures discussed in Chapter 5 with the tumour level selection measures from Chapter 4. The first approach provides an indication of how repeatable mutations in certain genes at a population scale, while the second approach enables estimating the selection coefficient in individual tumours. Combining these approaches may be more powerful than the approaches in isolation. Furthermore the approach outlined in Chapter 5 could be further developed to quantify selection pressures in normal and pre-malignant tissue and ultimately compared to results from malignant tissues as in Chapters 4 and 5.

In summary I hope that mathematical descriptions of tumour evolution together

with genomic data will facilitate a deeper understanding of genetic heterogeneity of cancers and enable strategies to treat and diagnose tumours inspired by evolutionary thinking to come to fruition. I hope that this thesis demonstrates in some small way how this approach can be utilised.

Appendix A

Dirichlet Process Clustering

In Chapter 4 I used a commonly used clustering approach to see how well the clonal composition of simulated tumours could be inferred from synthetic sequencing data. This method was first used in Nik-Zainal *et al.,* 2012b and is a form of Dirichlet Process clustering (Dunson, 2009). Subsequently many tools have been developed which use similar methodologies (Roth *et al.,* 2014; Miller *et al.,* 2014). I will outline the statistical model here following closely how it was originally described in Nik-Zainal *et al.,* 2012b. The assumptions are that mutations observed in a deep sequencing experiment are derived from an unknown number of subclones with unknown frequencies where the number of mutations associated with each subclone is also unknown. The goal is then to jointly estimate all these variables.

It is assumed that the number of reads, y_i reporting the ith mutation is Binomially distributed.

$$y_i \sim B_o(N_i, \pi_i) \tag{A.1}$$

Where N_i is the total number of reads covering the locus and pi_i is the fraction of tumour cells carrying the mutation. π can be any number between 0 and 1 and is modelled as coming from a Dirichlet process. Using the stick breaking representation of the Dirichilet process:

$$P = \sum_{h=1}^{\infty} \omega_h \delta_h, \text{with } \pi_h \sim P_0 \tag{A.2}$$

Here, ω_h is the weight of the hth mutation cluster, or equivalently the proportion of

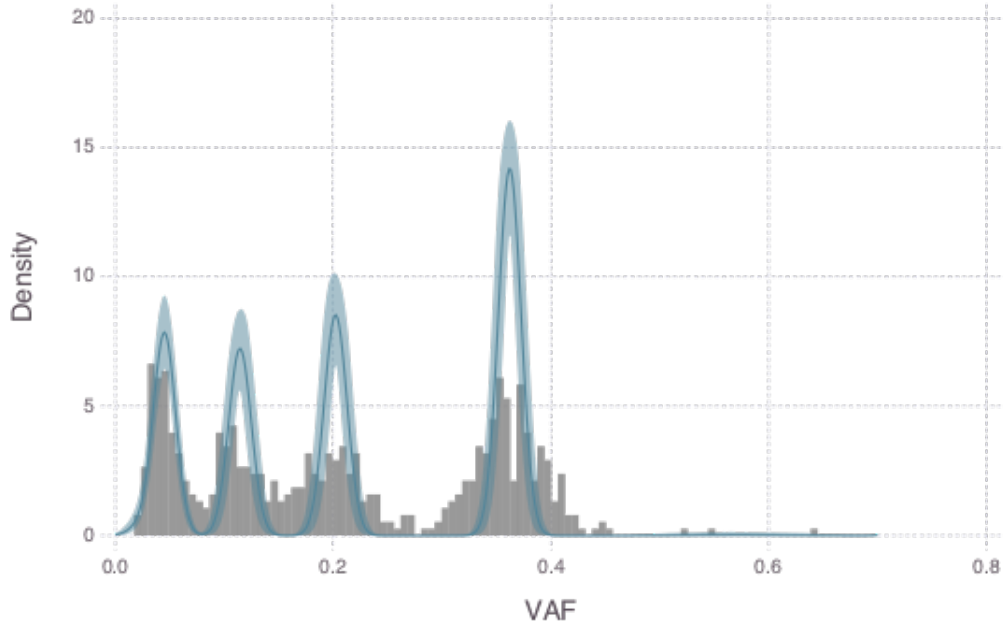

Figure A.1: *Dirichlet process clustering applied to Nik-Zainal data set. As in the original analysis my implementation of the statistical model found 4 clusters.*

mutations associated with that cluster, δ_h is a point mass at π. The stick-breaking formulation of the Dirichlet process can be captured using a Beta distribution as follows.

$$\omega_h = V_h \prod_{i<h} 1 - V_i \text{ with } V_h \sim Beta(1, \alpha) \tag{A.3}$$

Gibbs sampling can then be used to estimate the posterior distributions given the model above. This was implemented in Julia and the code is available here: `https://github.com/marcjwilliams1/DPclustering.jl`. As in Nik-Zainal et al, for the prior distributions I used $P_0 \sim U(0,1)$ and $\alpha \sim \Gamma(0.01, 0.01)$ and set an upper limit for h of 30. To confirm that my implementation works as expected, I ran the inference on one of the original datasets from Nik-Zainal *et al* and as in the original analysis found 4 clusters at the same frequency as was found in Nik-Zainal.

Appendix B

Publications

I have mentioned publications that I have contributed to during my PhD throughout where appropriate, this Appendix includes a full list.

B.1 First author publications

I obtained two first author publications during my PhD. These were presented in modified form in Chapter 4 and Chapter 5:

1. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nature Genetics. 2016 Mar;48(3):238?44.

2. Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, Graham TA. Quantification of subclonal selection in cancer from bulk sequencing data. Nature Genetics. 2018 Jun;50(6):895?903.

In addition to these primary research papers, I contributed to a number of discussions arising from Williams *et al.* 2016 which are listed below.

1. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A., Reply: Is the evolution of tumors Darwinian or non-Darwinian? NSR 2018 Jan 17;5(1):17?9.

2. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Reply: Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. Nature Genetics; 2017 Sep 1;49(9):1289?91.

3. Heide T, Zapata L, Williams MJ, Werner B, Caravagna G, Barnes CP, Graham TA, Sottoriva A. Reply to ?Neutral tumor evolution?.? Nature Genetics.2018 Oct 23;48:1?9.

4. Williams MJ, Werner B, Heide T, Barnes CP, Graham TA, Sottoriva A. Reply to ?Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data.? Nature Genetics. 2018 Dec 1;50(12):1628?30.

5. Werner B, Williams MJ, Barnes CP, Graham TA, Sottoriva A. Reply to ?Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution.? Nature Genetics; 2018 Oct 19;48:1?8.

B.2 Other papers

Below are some other papers to which I have contributed to by providing bioinfromatics support.

1. Baker A-M, Cross W, Curtius K, Bakir Al I, Choi C-HR, Davis HL, Temko D, Biswas S, Martinez P, Williams MJ, Lindsay JO, Feakins R, Vega R, Hayes SJ, Tomlinson IPM, McDonald SAC, Moorghen M, Silver A, East JE, Wright NA, Wang LM, Rodriguez-Justo M, Jansen M, Hart AL, Leedham SJ, Graham TA. Evolutionary history of human colitis-associated colorectal cancer. Gut. 2018 Jul 10;:gutjnl?2018?316191?11.

2. Temko D, van Gool IC, Rayner E, Glaire M, Makino S, Brown M, Chegwidden L, Palles C, Depreeuw J, Beggs A, Stathopoulou C, Mason J, Baker A-M, Williams M, Cerundolo V, Rei M, Taylor JC, Schuh A, Ahmed A, Amant F, Lambrechts D, Smit VT, Bosse T, Graham TA, Church DN, Tomlinson I. Somatic POLE exonuclease domain mutations are early events in sporadic endometrial and colorectal carcinogenesis, determining driver mutational landscape, clonal neoantigen burden and immune response. The Journal of Pathology. 2018 Jul;245(3):283? 96.

Bibliography

1. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. English. *Nature Genetics* **47,** 1402–1407 (Dec. 2015).
2. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. English. *Science* **354,** 618–622 (Nov. 2016).
3. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. English. *Nature* **500,** 415–421 (Aug. 2013).
4. Altrock, P. M., Liu, L. L. & Michor, F. The mathematics of cancer: integrating quantitative models. English. *Nature Reviews Cancer* **15,** 730–745 (Nov. 2015).
5. Alves, J. M., Prieto, T. & Posada, D. Multiregional Tumor Trees Are Not Phylogenies. English. *Trends in Cancer* **3,** 546–550 (Aug. 2017).
6. Anderson, A. R. A., Weaver, A. M., Cummings, P. T. & Quaranta, V. Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment. English. *Cell* **127,** 905–915 (Dec. 2006).
7. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. English. *Nature Medicine* **22,** 105–113 (Jan. 2016).
8. Antal, T. & Krapivsky, P. L. Exact solution of a two-type branching process: models of tumor progression. English. *Journal of Statistical Mechanics: Theory and Experiment* **2011,** P08018 (Aug. 2011).
9. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. English. *Nature communications* **6,** 8971 (Dec. 2015).

10. Araten, D. J. *et al.* A quantitative measurement of the human somatic mutation rate. English. *Cancer Research* **65,** 8111–8117 (Sept. 2005).

11. Armitage, P & Doll, R. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. English. *British journal of cancer* **11,** 161–169 (June 1957).

12. Baca, S. C. *et al.* Punctuated Evolution of Prostate Cancer Genomes. English. *Cell* **153,** 666–677 (Apr. 2013).

13. Bailey, N. *The elements of stochastic processes with applications to the natural sciences* 1964. <http://books.google.com/books?hl=en&lr=&id=yHPnwl4QOfIC&oi=fnd&pg=PA1&dq=The+Elements+of+Stochastic+Processes+with+Applications+to+the+Natural+Sciences&ots=DzjbVSVn1F&sig=UIcrNuxp1USRTHRQevh_UqtlW-w>.

14. Bailey, S. F., Hinz, A. & Kassen, R. Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. English. *Nature communications* **5,** 4076 (June 2014).

15. Baker, A.-M., Graham, T. A., Elia, G., Wright, N. A. & Rodriguez-Justo, M. Characterization of LGR5 stem cells in colorectal adenomas and carcinomas. English. *Scientific Reports* **5,** 25–8 (Mar. 2015).

16. Baker, A.-M. *et al.* Quantification of crypt and stem cell evolution in the normal and neoplastic human colon. English. *Cell Reports* **8,** 940–947 (Aug. 2014).

17. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. English. *Nature* **457,** 608–611 (Jan. 2009).

18. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449,** 1003–1007 (Oct. 2007).

19. Barnes, C. P., Filippi, S., Stumpf, M. P. H. & Thorne, T. Considerate approaches to constructing summary statistics for ABC model selection. English. *Statistics and Computing* **22,** 1181–1197 (2012).

20. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. English. *Genetics* **162,** 2025–2035 (Dec. 2002).

21. Beerenwinkel, N., Schwarz, R. F., Gerstung, M. & Markowetz, F. Cancer evolution: mathematical models and computational inference. English. *Systematic Biology* **64,** e1–e25 (Jan. 2015).

22. Beerenwinkel, N. *et al.* Genetic progression and the waiting time to cancer. English. *PLOS Computational Biology* **3,** e225 (Nov. 2007).

23. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. English. *Nature* **463,** 899–905 (Feb. 2010).

24. Betancourt, M. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv* (2017).

25. Bezanson, J, Edelman, A, Karpinski, S & Shah, V. Julia: A fresh approach to numerical computing. *SIAM*. doi:10.1137/141000671. <https://epubs.siam.org/doi/abs/10.1137/141000671>.

26. Bhang, H.-e. C. *et al.* Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. English. *Nature Medicine* **21,** 440–448 (May 2015).

27. Bielski, C. M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nature Genetics,* 1–11 (June 2018).

28. Blanpain, C. & Simons, B. D. Unravelling stem cell dynamics by lineage tracing. English. *Nature Reviews Molecular Cell Biology* **14,** 489–502 (Aug. 2013).

29. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature,* 1–17 (Oct. 2016).

30. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. English. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 18545–18550 (Oct. 2010).

31. Bozic, I., Gerold, J. M. & Nowak, M. A. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. English. *PLOS Computational Biology* **12,** e1004731 (Feb. 2016).

32. Campbell, B. B. *et al.* Comprehensive Analysis of Hypermutation in Human Cancer. *Cell,* 1–26 (Oct. 2017).

33. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. English. *Journal of Statistical Software* **76,** 1–32 (2017).

34. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. English. *Nature biotechnology* **30,** 413–421 (May 2012).

35. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nature Reviews Cancer,* 1–6 (Oct. 2015).

36. Champagnat, N., Lambert, A. & Richard, M. Birth and Death Processes with Neutral Mutations. English. *International Journal of Stochastic Analysis* **2012,** e569081–20 (Dec. 2012).

37. Chen, L., Liu, P., Evans, T. C. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. English. *Science* **355,** 752–756 (Feb. 2017).

38. Chen, Y., Tong, D. & Wu, C.-I. A New Formulation of Random Genetic Drift and Its Application to the Evolution of Cell Populations. English. *Molecular biology and evolution* **34,** 2057–2064 (Aug. 2017).

39. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. English. *Nature biotechnology* **31,** 213–219 (Mar. 2013).

40. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. English. *Nucleic Acids Research* **44,** e71–e71 (May 2016).

41. Costello, M *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. English. *Nucleic Acids Research* **41,** e67–e67 (Mar. 2013).

42. Davis, A. & Navin, N. E. Computing tumor trees from single cells. *Genome Biology,* 1–4 (May 2016).

43. Davis, A., Gao, R. & Navin, N. Tumor Evolution: Linear, Branching, Neutral or Punctuated? *BBA - Reviews on Cancer,* 1–58 (Jan. 2017).

44. De Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. English. *Science* **346,** 251–256 (Oct. 2014).

45. Del Moral, P., Doucet, A. & Jasra, A. Sequential Monte Carlo samplers. English. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68,** 411–436 (June 2006).

46. Didelot, X., Everitt, R. G., Johansen, A. M. & Lawson, D. J. Likelihood-free estimation of model evidence. English. *Bayesian Analysis* **6,** 49–76 (Mar. 2011).

47. Dobzhansky, T. *Nothing in biology makes sense except in the light of evolution* 1983. <https://philpapers.org/rec/DOBNIB>.

48. Driessens, G., Beck, B., Caauwe, A., Simons, B. D. & Blanpain, C. Defining the mode of tumour growth by clonal analysis. English. *Nature* **488,** 527–530 (Aug. 2012).

49. Dunson, D. B. in *Bayesian Nonparametrics* (eds Hjort, N. L., Holmes, C., Muller, P. & Walker, S. G.) 223–273 (Cambridge University Press, Cambridge, 2009). ISBN: 9780511802478. doi:10.1017/CBO9780511802478.008. <http://ebooks.cambridge.org/ref/id/CBO9780511802478A058>.

50. Durrett, R. & Schweinsberg, J. Approximating selective sweeps. English. *Theoretical population biology* **66,** 129–138 (Sept. 2004).

51. Durrett, R. Population genetics of neutral mutations in exponentially growing cancer cell populations. English. *The Annals of Applied Probability* **23,** 230–250 (2013).

52. Durrett, R. POPULATION GENETICS OF NEUTRAL MUTATIONS IN EXPONENTIALLY GROWING CANCER CELL POPULATIONS. English. *The Annals of Applied Probability* **23,** 230–250 (2013).

53. Durrett, R., Wai-Tong & Fan. Genealogies in Expanding Populations. arXiv: 1507.00918. <http://arxiv.org/abs/1507.00918> (July 2015).

54. Durrett, R., Foo, J., Leder, K., Mayberry, J. & Michor, F. Intratumor heterogeneity in evolutionary models of tumor progression. English. *Genetics* **188,** 461–477 (June 2011).

55. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518,** 422–426 (Feb. 2015).

56. El-Kebir, M., Satas, G. & Raphael, B. J. Inferring parsimonious migration histories for metastatic cancers. *Nature Genetics,* 1–13 (Apr. 2018).

57. Enriquez-Navas, P. M. *et al.* Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer. English. *Science Translational Medicine* **8,** 327ra24–327ra24 (Feb. 2016).

58. Ewens, W. J. *Mathematical Population Genetics 1: Theoretical Introduction* ISBN: 9780387218229. <https://books.google.co.uk/books?id=ZobfBwAAQBAJ> (Springer New York, 2012).

59. Favero, F *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. English. *Annals of oncology : official journal of the European Society for Medical Oncology* **26,** 64–70 (Jan. 2015).

60. Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. English. *Genetics* **155,** 1405–1413 (July 2000).

61. Fialkow, P. J. Clonal origin of human tumors. English. *Annual review of medicine* **30,** 135–143 (1979).

62. Fialkow, P. J. The origin and development of human tumors studied with cell markers. English. *The New England journal of medicine* **291,** 26–35 (July 1974).

63. Filippi, S., Barnes, C. P., Cornebise, J. & Stumpf, M. P. H. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. English. *Statistical applications in genetics and molecular biology* **12,** 87–107 (Mar. 2013).

64. Fischer, A., Vázquez-García, I., Illingworth, C. J. R. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. English. *Cell Reports* **7,** 1740–1752 (June 2014).

65. Fusco, D., Gralka, M., Kayser, J., Anderson, A. & Hallatschek, O. Excess of mutational jackpot events in expanding populations revealed by spatial Luria-Delbrück experiments. English. *Nature communications* **7,** 12760 (Oct. 2016).

66. Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature Genetics,* 1–15 (Aug. 2016).

67. Gatenby, R. A. A change of strategy in the war on cancer. English. *Nature* **459,** 508–509 (May 2009).

68. Gay, L., Baker, A.-M. & Graham, T. A. Tumour Cell Heterogeneity. English. *F1000Research* **5,** 238–14 (Feb. 2016).

69. Gelman, A, Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian data analysis* 2014. <http://amstat.tandfonline.com/doi/full/10.1080/01621459.2014.963405>.

70. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. English. *Nature Genetics* **46,** 225–233 (Mar. 2014).

71. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. English. *The New England journal of medicine* **366,** 883–892 (Mar. 2012).

72. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. English. *The Journal of Physical Chemistry* **81,** 2340–2361 (Dec. 1977).

73. Gillespie, J. H. Genetic Drift in an Infinite Population: The Pseudohitchhiking Model. English. *Genetics* **155,** 909–919 (June 2000).

74. Goldman, N & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. English. *Molecular biology and evolution* **11,** 725–736 (Sept. 1994).

75. Goode, D. L. *et al.* A simple consensus approach improves somatic mutation prediction accuracy. English. *Genome Medicine* **5,** 90 (2013).

76. Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. English. *Nature Reviews Genetics* **13,** 189–203 (Jan. 2012).

77. Greaves, M. & Maley, C. C. Clonal evolution in cancer. English. *Nature* **481,** 306–313 (Jan. 2012).

78. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical analysis of pathogenicity of somatic mutations in cancer. English. *Genetics* **173,** 2187–2198 (Aug. 2006).

79. Grelaud, A., Robert, C. P., Marin, J.-M., Rodolphe, F. & Taly, J.-F. ABC likelihood-free methods for model choice in Gibbs random fields. English. *Bayesian Analysis* **4,** 317–335 (June 2009).

80. Griffith, M. *et al.* Optimizing Cancer Genome Sequencing and Analysis. English. *Cell Systems* **1,** 210–223 (Sept. 2015).

81. Griffiths, R. C. & Pakes, A. G. An infinite-alleles version of the simple branching process. *Advances in applied probability.* doi:10.2307/1427033. <http://www.jstor.org/stable/1427033> (1988).

82. Griffiths, R. C. & Tavaré, S. The age of a mutation in a general coalescent tree. *Stochastic Models.* doi:10.1080/15326349808807471. <http://www.tandfonline.com/doi/abs/10.1080/15326349808807471> (1998).

83. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. English. *Cell* **144,** 646–674 (Mar. 2011).

84. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* English. ISBN: 9780878933082. <http://books.google.co.uk/books?id=SB1vQgAACAAJ&dq=intitle:Principles+of+Population+Genetics&hl=&cd=1&source=gbs_api> (Sinauer Associates Incorporated, 2007).

85. Hastie, T., Hast & Efron, B. Computer Age Statistical Inference, 1–495 (Oct. 2016).

86. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika.* <http://biomet.oxfordjournals.org/content/57/1/97.short> (1970).

87. Hu, Z., Sun, R. & Curtis, C. A population genetics perspective on the determinants of intra-tumor heterogeneity. *BBA - Reviews on Cancer,* 1–115 (Mar. 2017).

88. Humphries, A. *et al.* Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution. English. *Proceedings of the National Academy of Sciences of the United States of America* **110,** E2490–9 (July 2013).

89. Iwasa, Y., Nowak, M. A. & Michor, F. Evolution of resistance during clonal expansion. English. *Genetics* **172,** 2557–2566 (Apr. 2006).

90. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. English. *The New England journal of medicine,* NEJMoa1616288–13 (Apr. 2017).

91. Johnson, D. C. *et al.* Neutral tumor evolution in myeloma is associated with poor prognosis. English. *Blood* **130,** 1639–1643 (Oct. 2017).

92. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. English. *Nature* **502,** 333–339 (Oct. 2013).

93. Kassen, R. & Bataillon, T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. English. *Nature Genetics* **38,** 484–488 (Apr. 2006).

94. Keinan, A. & Clark, A. G. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. English. *Science* **336,** 740–743 (May 2012).

95. Keller, P. & Antal, T. Mutant number distribution in an exponentially growing population. English. *Journal of Statistical Mechanics: Theory and Experiment* **2015,** P01011 (Jan. 2015).

96. Kessler, D. A. & Levine, H. Large population solution of the stochastic Luria-Delbruck evolution model. English. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 11682–11687 (July 2013).

97. Kessler, D. A. & Levine, H. Scaling Solution in the Large Population Limit of the General Asymmetric Stochastic Luria–Delbrück Evolution Process. English. *Journal of Statistical Physics* **158,** 783–805 (Nov. 2014).

98. Kim, S., Jacob, L. & Speed, T. P. Combining calls from multiple somatic mutation-callers. English. *BMC bioinformatics* **15,** 154 (2014).

99. Klein, A. M. & Simons, B. D. Universal patterns of stem cell fate in cycling adult tissues. English. *Development* **138,** 3103–3111 (Aug. 2011).

100. Klein, A. M., Brash, D. E., Jones, P. H. & Simons, B. D. Stochastic fate of p53-mutant epidermal progenitor cells is tilted toward proliferation by UV B during preneoplasia. English. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 270–275 (Jan. 2010).

101. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. English. *Proceedings of the National Academy of Sciences* **68,** 820–823 (Apr. 1971).

102. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. English. *Genome Research* **22,** 568–576 (Mar. 2012).

103. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. English. *Cell* **152,** 1226–1236 (Mar. 2013).

104. Korolev, K. S. *et al.* Selective sweeps in growing microbial colonies. English. *Physical biology* **9,** 026008 (2012).

105. Kostadinov, R., Maley, C. C. & Kuhner, M. K. Bulk Genotyping of Biopsies Can Create Spurious Evidence for Hetereogeneity in Mutation Content. English. *PLOS Computational Biology* **12,** e1004413 (Apr. 2016).

106. Kozar, S. *et al.* Continuous Clonal Labeling Reveals Small Numbers of Functional Stem Cells in Intestinal Crypts and Adenomas. English. *Cell Stem Cell* **13,** 626–633 (Nov. 2013).

107. Kryazhimskiy, S. & Plotkin, J. B. The Population Genetics of dN/dS. English. *PLOS Genet* **4,** e1000304–10 (Dec. 2008).

108. Lan, X. *et al.* Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. English. *Nature* **549,** 227–232 (Sept. 2017).

109. Lässig, M., Mustonen, V. & Walczak, A. M. Predicting evolution. English. *Nature Ecology & Evolution* **1,** 77 (Feb. 2017).

110. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. English. *Nature* **505,** 495–501 (Jan. 2014).

111. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. English. *Nature* **499,** 214–218 (July 2013).

112. Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. English. *Proceedings of the National Academy of Sciences* **91,** 6808–6814 (July 1994).

113. Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. English. *Nature* **519,** 181–186 (Mar. 2015).

114. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013).

115. Ling, S. *et al.* Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. English. *Proceedings of the National Academy of Sciences of the United States of America* **112,** E6496–505 (Nov. 2015).

116. Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S. & Corander, J. Fundamentals and Recent Developments in Approximate Bayesian Computation. English. *Systematic Biology,* syw077–17 (Oct. 2016).

117. Lopez-Garcia, C., Klein, A. M., Simons, B. D. & Winton, D. J. Intestinal stem cell replacement follows a pattern of neutral drift. English. *Science* **330,** 822–825 (Nov. 2010).

118. Łuksza, M. *et al.* A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **27,** 1–20 (Nov. 2017).

119. Luria, S. E. & Delbrück, M. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. English. *Genetics* **28,** 491–511 (Nov. 1943).

120. Lynch, M. The frailty of adaptive hypotheses for the origins of organismal complexity. English. *Proceedings of the National Academy of Sciences* **104 Suppl 1,** 8597–8604 (May 2007).

121. Mackay, D. Information Theory, Inference, and Learning Algorithms, 1–640 (Sept. 2016).

122. Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. Markov chain Monte Carlo without likelihoods. English. *Proceedings of the National Academy of Sciences* **100,** 15324–15328 (Dec. 2003).

123. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. English. *Science* **348,** 880–886 (May 2015).

124. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell,* 1–35 (Oct. 2017).

125. Martinez, P. *et al.* Dynamic clonal equilibrium and predetermined cancer risk in Barrett’s oesophagus. *Nature communications* **7,** 1–10 (Aug. 2016).

126. McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R. & Mirny, L. A. Impact of deleterious passenger mutations on cancer progression. English. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 2910–2915 (Feb. 2013).

127. McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. English. *Science* **351,** 1463–1469 (Mar. 2016).

128. Mcpherson, A. *et al.* Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics,* 1–11 (May 2016).

129. McPherson, A. W., Chan, F. C. & Shah, S. P. Observing Clonal Dynamics Across Spatiotemporal Axes: A Prelude to Quantitative Fitness Models for Cancer. English. *Cold Spring Harbor perspectives in medicine,* a029603–14 (June 2017).

130. Medema, J. P. & Vermeulen, L. Microenvironmental regulation of stem cells in intestinal homeostasis and cancer. *Nature* **474,** 318–326 (June 2011).

131. Merkle, F. T. *et al.* Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations. *Nature,* 1–11 (Apr. 2017).

132. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. English. *The Journal of Chemical Physics* **21,** 1087–1092 (June 1953).

133. Miller, C. A. *et al.* SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. English. *PLOS Computational Biology* **10,** e1003665 (Aug. 2014).

134. Mugal, C. F., Wolf, J. B. W. & Kaj, I. Why Time Matters: Codon Evolution and the Temporal Dynamics of dN/dS. English. *Molecular biology and evolution* **31,** 212–231 (Oct. 2013).

135. Nicholson, A. M. *et al.* Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium. English. *Cell Stem Cell* **22,** 909–918.e8 (June 2018).

136. Nicholson, M. D. & Antal, T. Universal asymptotic clone size distribution for general population growth. arXiv: 1604.04936. <http://arxiv.org/abs/1604.04936> (Apr. 2016).

137. Nielsen, R. & Yang, Z. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. English. *Molecular biology and evolution* **20,** 1231–1239 (Aug. 2003).

138. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. English. *Cell* **149,** 979–993 (May 2012).

139. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. English. *Cell* **149,** 994–1007 (May 2012).

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

141. Ohtsuki, H. & Innan, H. Allele Frequency Spectrum in a Cancer Cell Population, 1–14 (Jan. 2017).

142. Okosun, J. *et al.* Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. English. *Nature Genetics* **46,** 176–181 (Feb. 2014).

143. Peterson, G. I. & Masel, J. Quantitative Prediction of Molecular Clock and Ka/Ks at Short Timescales. English. *Molecular biology and evolution* **26,** 2595–2603 (Oct. 2009).

144. Pino, M. S. & Chung, D. C. *The chromosomal instability pathway in colon cancer.* English. June 2010. doi:10.1053/j.gastro.2009.12.065. <http://linkinghub.elsevier.com/retrieve/pii/S0016508510001708>.

145. Poleszczuk, J. & Enderling, H. A High-Performance Cellular Automaton Model of Tumor Growth with Dynamically Growing Domains. English. *Applied mathematics* **5,** 144–152 (Jan. 2014).

146. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. English. *Molecular biology and evolution* **16,** 1791–1798 (Dec. 1999).

147. Qiao, Y. *et al.* SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. English. *Genome Biology* **15,** 443 (2014).

148. Queller, D. C. Fundamental Theorems of Evolution. English. *The American Naturalist* **189,** 345–353 (Apr. 2017).

149. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2016). <https://www.R-project.org>.

150. Ritsma, L. *et al.* Intestinal crypt homeostasis revealed at single-stem-cell level by in vivo live imaging. English. *Nature* **507,** 362–365 (Mar. 2014).

151. Robert, C. P., Cornuet, J.-M., Marin, J.-M. & Pillai, N. S. Lack of confidence in approximate Bayesian computation model choice. English. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 15112–15117 (Sept. 2011).

152. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. English. *Nature* **548,** 297–303 (Aug. 2017).

153. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29,** 24–26 (Jan. 2011).

154. Rogers, Z. N. *et al.* A quantitative and multiplexed approach to uncover the fitness landscape of tumor suppression in vivo. English. *Nature methods* **14,** 737–742 (July 2017).

155. Rogers, Z. N. *et al.* Mapping the in vivo fitness landscape of lung adenocarcinoma tumor suppression in mice. English. *Nature Genetics* **50,** 483–486 (Apr. 2018).

156. Ronen, R., Udpa, N., Halperin, E. & Bafna, V. Learning natural selection from the site frequency spectrum. English. *Genetics* **195,** 181–193 (Sept. 2013).

157. Roth, A. *et al.* Clonal genotype and population structure inference from single-cell tumor sequencing. English. *Nature methods* **13,** 573–576 (July 2016).

158. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. English. *Nature methods* **11,** 396–398 (Apr. 2014).

159. Rouhani, F. J. *et al.* Mutational History of a Human Cell Lineage from Somatic to Induced Pluripotent Stem Cells. English. *PLOS Genet* **12,** e1005932–15 (Apr. 2016).

160. S Datta, R., Gutteridge, A., Swanton, C., Maley, C. C. & Graham, T. A. Modelling the evolution of genetic instability during tumour progression. English. *Evolutionary Applications* **6,** 20–33 (Jan. 2013).

161. Salehi, S. *et al.* ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. English. *Genome Biology* **18,** 44 (Mar. 2017).

162. Sánchez-Danés, A. *et al.* Defining the clonal dynamics leading to mouse skin tumour initiation. *Nature,* 1–22 (July 2016).

163. Sandberg, A. A. & Hossfeld, D. K. Chromosomal abnormalities in human neoplasia. English. *Annual review of medicine* **21,** 379–408 (1970).

164. Schulze, T. P. Efficient kinetic monte carlo simulation. *Journal of Computational Physics.* <http://www.sciencedirect.com/science/article/pii/S0021999107004755> (2008).

165. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. English. *Nature* **488,** 504–507 (Aug. 2012).

166. Schwartz, R. & Schäffer, A. A. The evolution of tumour phylogenetics: principles and practice. English. *Nature Reviews Genetics* **18,** 213–229 (Apr. 2017).

167. Scott, J. & Marusyk, A. Somatic clonal evolution: A selection-centric perspective. *BBA - Reviews on Cancer,* 1–12 (Feb. 2017).

168. Seshadri, R, Kutlaca, R. J., Trainor, K, Matthews, C & Morley, A. A. Mutation rate of normal and malignant human lymphocytes. English. *Cancer Research* **47,** 407–409 (Jan. 1987).

169. Siegmund, K. D., Marjoram, P., Woo, Y.-J., Tavaré, S. & Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. English. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 4828–4833 (Mar. 2009).

170. Sievers, C. K. *et al.* Subclonal diversity arises early even in small colorectal tumours and contributes to differential growth fates. English. *Gut,* gutjnl–2016–312232–10 (Sept. 2016).

171. Simons, B. D. Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. English. *Proceedings of the National Academy of Sciences* **113,** 128–133 (Jan. 2016).

172. Snippert, H. J., Schepers, A. G., van Es, J. H., Simons, B. D. & Clevers, H. Biased competition between Lgr5 intestinal stem cells driven by oncogenic mutation induces clonal expansion. English. *EMBO reports* **15,** e201337799–69 (Dec. 2013).

173. Snippert, H. J. *et al.* Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells. English. *Cell* **143,** 134–144 (Oct. 2010).

174. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. English. *Nature Genetics* **47,** 209–216 (Mar. 2015).

175. Sottoriva, A., Barnes, C. P. & Graham, T. A. Catch my drift? Making sense of genomic intra-tumour heterogeneity. *BBA - Reviews on Cancer,* 1–17 (Jan. 2017).

176. Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. English. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 4009–4014 (Mar. 2013).

177. Sottoriva, A., Spiteri, I., Shibata, D., Curtis, C. & Tavaré, S. Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization. English. *Cancer Research* **73,** 41–49 (Jan. 2013).

178. Stamp, C. *et al.* Predominant Asymmetrical Stem Cell Fate Outcome Limits the Rate of Niche Succession in Human Colonic Crypts. *EBioMedicine* **31,** 166–173 (May 2018).

179. Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. English. *Science* **331,** 1553–1558 (Mar. 2011).

180. Sun, R. *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. English. *Nature Genetics* **49,** 1015–1024 (July 2017).

181. Sunnåker, M. *et al.* Approximate Bayesian Computation. English. *PLOS Computational Biology* **9,** e1002803–10 (Jan. 2013).

182. Tavaré, S, Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. English. *Genetics* **145,** 505–518 (Feb. 1997).

183. Temko, D., Tomlinson, I. P. M., Severini, S., Schuster-Böckler, B. & Graham, T. A. The effects of mutational processes and selection on driver mutations across cancer types. English. *Nature communications* **9,** 1857 (May 2018).

184. Thirlwell, C. *et al.* Clonality assessment and clonal ordering of individual neoplastic crypts shows polyclonality of colorectal adenomas. English. *Gastroenterology* **138,** 1441–54–1454.e1–7 (Apr. 2010).

185. Toni, T, Welch, D, Strelkowa, N, Ipsen, A & Stumpf, M. P. H. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. English. *Journal of The Royal Society Interface* **6,** 187–202 (Feb. 2009).

186. Toni, T. & Stumpf, M. P. H. Simulation-based model selection for dynamical systems in systems and population biology. English. *Bioinformatics (Oxford, England)* **26,** 104–110 (Jan. 2010).

187. Tsao, J. L. *et al.* Colorectal adenoma and cancer divergence. Evidence of multilineage progression. English. *The American Journal of Pathology* **154,** 1815–1824 (June 1999).

188. Tsao, J. L. *et al.* Genetic reconstruction of individual colorectal tumor histories. English. *Proceedings of the National Academy of Sciences* **97,** 1236–1241 (Feb. 2000).

189. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. English. *Current protocols in bioinformatics* **43,** 11.10.1–33 (2013).

190. Vermeulen, L. & Snippert, H. J. Stem cell dynamics in homeostasis and cancer of the intestine. English. *Nature Reviews Cancer* **14,** 468–480 (July 2014).

191. Vermeulen, L. *et al.* Defining stem cell dynamics in models of intestinal tumor initiation. English. *Science* **342,** 995–998 (Nov. 2013).

192. Vogelstein, B. *et al.* Cancer genome landscapes. English. *Science* **339,** 1546–1558 (Mar. 2013).

193. Waclaw, B. *et al.* A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. English. *Nature* **525,** 261–264 (Sept. 2015).

194. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. English. *Nucleic Acids Research* **38,** e164–e164 (Sept. 2010).

195. Wang, K. *et al.* Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. English. *Nature Publishing Group* **46,** 573–582 (June 2014).

196. Wang, Q. *et al.* Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. English. *Genome Medicine* **5,** 91 (2013).

197. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. English. *Science* **350,** 1096–1101 (Nov. 2015).

198. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. English. *Nature* **512,** 155–160 (Aug. 2014).

199. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. English. *Nature Genetics* **49,** 1–8 (Nov. 2017).

200. Werner, B., Traulsen, A., Sottoriva, A. & Dingli, D. Detecting truly clonal alterations from multi-region profiling of tumours. *Scientific Reports,* 1–9 (Mar. 2017).

201. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* ISBN: 978-0-387-98140-6. <http://ggplot2.org> (Springer-Verlag New York, 2009).

202. Wright, N. A. Stem cell identification—in vivo lineage analysis versus in vitro isolation and clonal expansion. English. *The Journal of Pathology* **227,** 255–266 (July 2012).

203. Wu, C.-I., Wang, H.-Y., Ling, S. & Lu, X. The Ecology and Evolution of Cancer: The Ultra-Microevolutionary Process. English. *Annual review of genetics* **50,** 347–369 (Nov. 2016).

204. Yang, Z & Bielawski, J. Statistical methods for detecting molecular adaptation. English. *Trends in Ecology & Evolution* **15,** 496–503 (Dec. 2000).

205. Yang, Z., Ro, S. & Rannala, B. Likelihood models of somatic mutation and codon substitution in cancer genes. English. *Genetics* **165,** 695–705 (Oct. 2003).

206. Yarchoan, M., Johnson, B. A., Lutz, E. R., Laheru, D. A. & Jaffee, E. M. Targeting neoantigens to augment antitumour immunity. English. *Nature Reviews Cancer* **17,** 209–222 (Feb. 2017).

207. Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. English. *Nature Medicine* **21,** 751–759 (July 2015).

208. Zahn, H. *et al.* Scalable whole-genome single-cell library preparation without preamplification. English. *Nature methods* **14,** 167–173 (Feb. 2017).

209. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. English. *Science* **346,** 256–259 (Oct. 2014).

210. Zhang, J., Cunningham, J. J., Brown, J. S. & Gatenby, R. A. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nature communications,* 1–9 (Nov. 2017).

211. Zhao, J. *et al.* Early mutation bursts in colorectal tumors. English. *PLOS ONE* **12,** e0172516–19 (Mar. 2017).

212. Zheng, Q. Progress of a half century in the study of the Luria–Delbrück distribution. *Mathematical biosciences.* doi:10.1016/S0025-5564(99)00045-0. <http://www.sciencedirect.com/science/article/pii/S0025556499000450> (1999).