Measuring tumour evolution at the genetic and epigenetic level in individual patients from cancer genomic data

Submitted for the degree of

Doctor of Philosophy

Ketevan Chkhaidze

Institute of Cancer Research

University of London

The work contained in this thesis is my own work unless otherwise stated.

Abstract

High-throughput genomic data from cancers uncovered high intra and inter-tumour heterogeneity and subclonal architecture of cancer cell populations. If we consider cells as asexually reproducing individuals, we can apply evolutionary theory to cancer, as all three building blocks of evolutionary dynamics - replication, selection, and mutation - are also the defining characters of cancer development. Studying cancer evolution in humans is imperative to predict the course of the disease and develop better therapeutic strategies.

Currently, there is a lack of mathematical and computational models that describe the effects of clonal selection in cancer data for growing populations. One of the reasons being that selection depends on many factors, such as fitness, context, and spatial constraints. In this thesis, we developed a stochastic simulation model of spatial tumour growth from which we can generate the genomic data we expect under different conditions and sampling methods. The model enabled us to monitor the effects of sampling bias on cancer genomic data as well as the effects of spatial constraints on tumour growth dynamics.

In the second part of the thesis, we also tried to study the links between genetics and epigenetics that influence cancer formation and progression. We developed methods to test our hypothesis that different mutational processes (giving rise to distinct mutational signatures) are active in epigenetically different regions of the genome, and a model that infers times of different chromatin aberration events.

Overall, this thesis shows the importance of coupling mathematical and computational modelling with experiments to gain a better understanding of cancer initiation and progression and consequently achieve better clinical performance.

Acknowledgements

I would like to express my immense gratitude to my supervisor Andrea Sottoriva for his patience and support throughout these last four years. Thank you for attributing a lot of trust to my skills and always encouraging to be more persistent and confident. I am grateful for all the advice I received from you and how you were always motivated in my personal and scientific development. You helped me become a better thinker, scientist and person in general.

I would like to thank all the current and former members of the Sottoriva lab, to all my co-authors and collaborators, without whom I would not be able to be at this position. Especially, to Ben Werner, for his patience and extraordinary ability to explain abstract concepts and keeping me motivated to never stop exploring. Timon Heide, for all the great skills and coding tips I learned from him. Marc Williams and Haider Tari, for re-explaining concepts, helping me debug my code and just for all their support and motivation in my success.

And finally, I would like to thank my family and friends, for their encouragement, help and never-ending support. For believing in me and keeping me motivated when I needed it the most.

The study reported in chapter 3 has been made possible partly on the basis of the data that the Hartwig Medical Foundation has made available to the study (DR-078). Data from this cohort are available for academic research upon request (https://www.hartwigmedicalfoundation.nl/en).

Contents

List of Figures

List of Abbreviations

1D One-dimensional.

2D Two-dimensional.

3D Three-dimensional.

ABC Approximate Bayesian Computation.

ABC-SMC Approximate Bayesian Computation Sequential Monte Carlo.

AIC Akaike Information Criteria.

ATAC-seq Assay for Transposase-Accessible Chromatin using sequencing.

AUC Area Under the Curve.

BIC Bayesian Information Criteria.

CAT Computational Approach Test.

ChIP-seq Sequencing method that combines Chromatin ImmunoPrecipitation with massively parallel DNA sequencing.

CNV Copy-Number Variation.

CPT Change-Point Analysis.

ITH Intra-Tumour Heterogeneity.

MBIC Modified Bayesian Information Criteria.

MCMC Markov Chain Monte Carlo.

MRCA Most Recent Common Ancestor.

NGS Next-Generation Sequencing.

ODE Ordinal Differential Equations.

PDE Partial Differential Equations.

PELT Pruned Exact Linear Time.

RNA-seq Whole transcriptome shotgun sequencing.

SNV Single-Nucleotide Variant.

TMRCA Time to the Most Recent Common Ancestor.

VAF Variant Allele Frequency.

WES Whole-Exome Sequencing.

WGS Whole-Genome Sequencing.

WT Wild type.

Chapter 1

Introduction

1.1 Biology basics

1.1.1 Cancer biology

Cancer is a term for diseases mainly characterised by out-of-control cell growth. Cells obey a set of rules in order for organisms to function normally. Sometimes, due to damages to the cells' genetic material that produce mutations, or due to various environmental factors that affect cells' epigenetic features, cells start to behave abnormally. Abnormal cells can grow out of control forming tumours. There are two types of tumours – benign and malignant. A benign tumour cannot invade neighbouring tissues and thus does not spread to other parts of the body (or metastasize). Whereas a malignant tumour spreads to other parts of the body (sometimes to very distant parts through the lymphatic system or bloodstream) and becomes cancerous.

The reasons why cells escape their control mechanisms and start rapid proliferation are various. One example is viruses that can alter cells' proliferation properties. Discovery of the papillomavirus and its role in causing cervical cancer [1] lead to

the development of vaccination (and later won a Nobel prize in medicine). Another example is a study that found a statistical correlation between Burkitt's lymphoma and the Epstein-Barr virus, where the viruses caused to trigger mutations in specific oncogenes [2, 3]. This discovery formed the current view that cancer is a genetic disease mainly caused by mutational hits in different cancer-initiating genes. Usually, several mutational hits are necessary within a cell to start abnormal behaviour and lead to cancer progression. These discoveries established the multi-stage theory of carcinogenesis, which assumes that mutations accumulate within cells and eventually provide a fitness advantage to those cells [4].

Through multi-stage theory, several important discoveries have been made. One such is an observation where one mutation in a single gene was sufficient to cause retinoblastoma, but since the mutation is recessive (i.e. genes can be expressed in offspring only when inherited from both parents), thus both copies of the gene (also called alleles) needed to have the mutations for cancer initiation. Such recessive genes are referred to as tumour suppressor genes, with one such prominent gene being p53 that is mutated in almost half of all human cancers [5, 6]. Genes, where a single mutational hit is sufficient to change cells' proliferation properties, are called oncogenes. One such oncogene is the BCR-ABL translocation of the 9th and 22nd chromosome in hematopoietic stem cells. Although a single mutated gene is, in most cases, not sufficient to initiate cancer, rather multiple mutated oncogenes are necessary. The exact number is determined by tissue and cancer type.

1.1.2 Cancer genomics and epigenomics

Cancer has long been assumed to be a disease primarily routed in genetics. As described in the previous section, a series of disruptions to the cellular mechanisms need to take place for cancer to develop and progress. Such aberrations affect cellular proliferation, immortality, angiogenesis, cell death, invasion, and metastasis [7].

Besides, this sequence of disruptions needs to be thoroughly encoded so that oncogenic events accumulate and establish clonal lineages. We know that the nucleotide sequence is the basic level of genetic information. And changes in genetic information, also called genetic mutations such as single nucleotide variations, copy number alterations, insertions, deletions and recombinations, create sources for phenotypic variations.

Advances in molecular profiling approaches gave birth to a whole new field called cancer epigenomics that studies the role of epigenetics in cancer development [8, 9]. Studies began to find another crucial and complementary role in cancer initiation and progression to be the various processes that are involved in gene regulations (i.e. epigenetics). Epigenetic patterns of gene regulation and functionality can be heritable from one cell to the other but do not affect the sequence of the genome. Through epigenetics, scientists started to understand the mechanisms of how identical genotypes give rise to different phenotypes even when subject to the same environmental stimulus.

Epigenetic modifications can consist of methylation or acetylation changes or chromatin modifications. These epigenetic variations persist through multiple cell divisions and thus mark a powerful effect on cells phenotype. It has only recently been discovered and recognised the important role of mutations in epigenetic regulatory mechanisms and the variety of alterations to the epigenome in cancer cells [10].

The advances in the next-generation sequencing (NGS) technologies brought significant insights in understanding cancer development both on the genomic and epigenomic level. It is becoming easier and cheaper to generate massive amounts of NGS data that allow scientists to study the diverse variations present in individual patients. For cancer research, being able to sequence samples from tumour tissues as well as normal (germline DNA) and compare the patterns, has been crucial for un-

derstanding the somatic variants that initiate and regulate cancer development. In cancer we often use different genomic analysis [11], such as whole-exome sequencing (WES) that identifies mutations in coding regions of DNA, whole-genome sequencing (WGS) that studies patterns in non-coding DNA regions, transcriptome sequencing (RNA-seq) that tries to reconstruct gene expression patterns. All these technologies lead to the current understanding of how genetic and epigenetic mechanisms act co-operatively and have shared influence when it comes to acquiring different hallmarks of cancer initiation and development [7].

1.2 Evolutionary biology

"Nothing in cancer makes sense
except in the light of evolution"
- Mel Greaves

1.2.1 Molecular evolution

The field of molecular evolution studies how gene sequences change and evolve over time [12]. It uses the principles of evolutionary biology and population genetics to explain patterns of these changes. In this chapter, I will briefly introduce some of the main topics and concepts in molecular evolution, such as mutation, allele frequency, different models of evolution and how evolutionary forces influence genomic and phenotypic changes.

As in most scientific disciplines, the main goal of molecular evolution is to infer process from patterns. Such processes can be either evolution of individual organisms deduced from the changes of DNA (Deoxyribonucleic acid), or the processes of molecular evolution inferred directly from DNA variations. **DNA** is a molecule composed of two chains of sequences formed by four nucleotides (cytosine - C, guanine - G, adenine - C or thymine - T) that create a spiral known as the double helix [13]. The genetic instructions for growth, reproduction, development and functioning for all organisms are written through these chains of nucleotides.

For the genetic instructions to be activated, first DNA needs to be transcribed into **RNA**, which is also a chain of nucleic acids. **Transcription** is then followed by RNA **translation** into proteins (Figure 1.1). During translation, specific chains

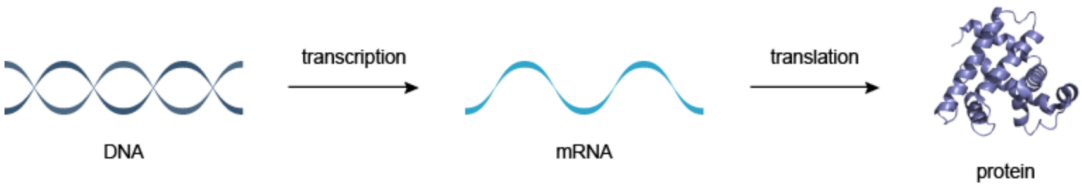


Figure 1.1: DNA transcription into RNA that is then translated to form proteins. (Figure source *wikipedia*)

of amino acids are formed, that fold into distinct proteins and perform different functions in the cell. DNA triplets, also known as **codons**, code for **Amino acids**. There are $4^3 = 64$ possible combinations of the 4 nucleotides, minus 3 that are stop codons that do not code for any amino acid, so 61 different codons in total that code for 20 different amino acids. Hence, there are some codons that code for the same amino acid. Different codons coding for the same amino acid are called **synonymous codons**.

DNA regions that code for proteins – **genes**, are usually split into **exons**, that are expressed during transcription, and **introns**, that are spliced out during messenger RNA formation. In addition, genes have **regulatory regions** like enhancers and promoters, that instruct time and position of DNA transcription into RNA for protein synthesis. The information system that governs the process of translation to form amino acids from the sequences of RNA, is called the **universal genetic code**. Although, this code is not completely universal as the mitochondrial genome uses codons in different ways; for instance, some stop codons code for amino acids in the mitochondrial DNA.

A somatic change in DNA sequence due to either a mistake during DNA copying or response to different environmental factors (cigarette smoking or exposure to ultraviolet light) is called a **mutation**. There are four different kinds of genetic variations: nucleotide insertion, deletion, inversion and substitution to another nu-

cleotide. Some mutations are harmful as they change the code so that proteins are no longer generated or have a modified harmful property. Others might have a positive effect and help the organism adapt to a changing environment. Most mutations are not harmful nor positive, but neutral – they don't cause any functional change of the genetic code. When a single nucleotide variation occurs in a protein-coding gene and creates a synonymous codon, this is referred as a **synonymous mutation** or silent substitution, whereas if it creates a nonsynonymous codon then it's called **nonsynonymous mutation**. The ones that result in termination codons are referred to as **nonsense mutations**.

With molecular evolution, we can understand the sequence of processes that are initiated by mutants in the population using principles of evolutionary biology and population genetics. **Population genetics** studies frequency dynamics of gene polymorphic sites over time. A gene locus can have different variants due to mutations passed on to the offspring. These variants at a given locus are termed **alleles**. In a diploid population, there are only two alleles. When both alleles are the same, the alleles are **homozygous**, and **heterozygous** otherwise. Allele frequencies of given mutations change over time. At the extremes, they either get lost in the population or reach fixation.

The rate of population divergence is governed by an underlying mutation rate, generation time steps and evolutionary forces such as competition between alleles, positive and negative selective pressures, the fitness of species with a given allele, genetic drift and population size. If a fitness of an individual carrying a certain allele is high, or in other words, an allele is fit in the population, it will be subject to positive selection. On the other hand, when an allele is less fit, it will be subject to negative selection. Sometimes, heterozygosity can be more advantageous feature over being homozygous. When heterozygosity is advantageous and maintained, this creates polymorphism and is termed **balancing selection**.

Fitness of an allele is determined by the organism's phenotype. Selective pressures often act on nonsynonymous substitutions of the coding regions. Synonymous mutations are mostly considered to be neutral. However, synonymous substitutions might not always be neutral, as some of the synonymous alterations of amino acids can cause changes in RNA secondary structure.

Another important concept from population genetics is the **effective population size**, that is defined as idealised population size (usually smaller than the real population size) with the assumption of perfect random mating and same gene frequency dynamics as of the real population under the study. There are deterministic and stochastic forces that coupled with the effective population size, influence the rate of fixation of a mutation.

If the effective population size is infinitely large and there are only deterministic evolutionary forces, then the changes in allele frequency will be determined only by the reproductive fitness of the variant in a given environment together with the constraints of the environment. In this context, only **natural selection** (survival and the reproductive difference between species due to their phenotypic differences) determines the changes in gene frequencies. In such cases, the exact gene frequencies can be predicted if fitness and environmental conditions are known. On the other hand, when the effective population size is small, random events, such as **genetic drift** (a given gene variant frequency change due to a random sampling of the organisms for the next generation) play the primary role in gene frequency dynamics.

In reality, evolution is never either deterministic or stochastic; it is rather the combination of both and depending on a given population size and selective forces, the interplay between natural selection and genetic drift influences the evolution of gene frequency. Although genetic mutations are random, they can influence fitness advantage of an allele, get selected by positive selective forces and eventually get fixed in the population sooner than it would under neutral evolution (assuming the

effective population size is large enough). Similarly, if a random genetic mutation results in an adaptive fitness disadvantage of an allele, it will experience negative selective pressures but can still get fixed in the population due to random genetic drift. The fixation of such a mutation will need more generations than under neutral conditions. In general, nonsynonymous mutations are subject to selective forces as they cause phenotypic changes of an organism, whereas synonymous mutations being neutral, can get fixed in the population due to random genetic drift only.

In "On the origin of species", Darwin introduced the main factors that dictate evolution. These are: environmental constraints, inheritable variations of traits that shaped the fitness of individuals, competitions between organisms and natural selection. After rediscovering Mendelian laws, it became clear that random genetic mutations are the main sources of variation and that natural selection was acting upon it and this way driving the evolution. As discussed above, mutations that result in an advantageous or disadvantageous trait for a given environment will get fixed or eliminated from the population, respectively. Also, changes in the environment can cause changes in the phenotypic traits of neutral mutations. Such model is referred to as **adaptive evolution** model; individuals with higher fitness advantage increase in frequency and become more adapted to the environment.

In 1968, Motoo Kimura introduced the **neutral theory of molecular evolution**, that states that despite positive selection being central to adaptation, the majority of mutations were not positively selected but rather neutral or negatively selected[14]. His main argument was that effective population size is very small compared to the magnitude of selective forces; hence, positive selection, although having the primary influence in shaping the genome of species, occurs rarely. Also, organisms are already so well adapted to the environment, that most of the nonsynonymous substitutions are deleterious and constantly removed from the population. Hence, fixation of a variation is mostly dictated by random genetic drift and thus

stochasticity plays the dominant role in evolution.

Kimura introduced a very simple description of the process of molecular evolution; if u is a mutation rate per gene per generation and N is the population size, then the number of new mutations per generation in a diploid population is $2Nu$. The probability of a new mutation in a population with $2N$ genes is $1/2N$ (which can be also called the probability of fixation). The **rate of substitution** K is then $2Nu * (1/2N) = u$. That is, the neutral rate of molecular evolution is equal to the rate of neutral mutation. This equation implies that genes, with different mutation rates or under selection, will have different rates of molecular evolution. This observation led to the prediction that the synonymous sites will evolve faster than nonsynonymous sites due to different functional constraints accompanying the later.

Different rates of substitutions have indeed been observed between different lineages of species[15]. The causes of these differences have not yet been fully understood, but this is studied under the **molecular clock hypothesis**. The molecular clock hypothesis is a consequence of the neutral model in a sense that if most mutations are neutral, then the majority of variation should have a clock-like behaviour. The hypothesis states that there is a positive linear relationship between the time of two species divergence and the amount of genetic divergence between them. The fact that evolution rates differ as discussed above, implies that there might be several molecular clocks that tick in parallel and at different rates.

The relationship between adaptive evolution and neutral evolution can be illustrated by fitness landscape plots (Figure 1.2), where x and y coordinates represent genetic variation and height indicates the fitness of an individual. Species will be driven towards fitness peaks by adaptive evolution by selecting the individuals with the advantageous mutations for the given environmental conditions. The paths towards the peaks will be determined by the slope of the fitness peaks in the given

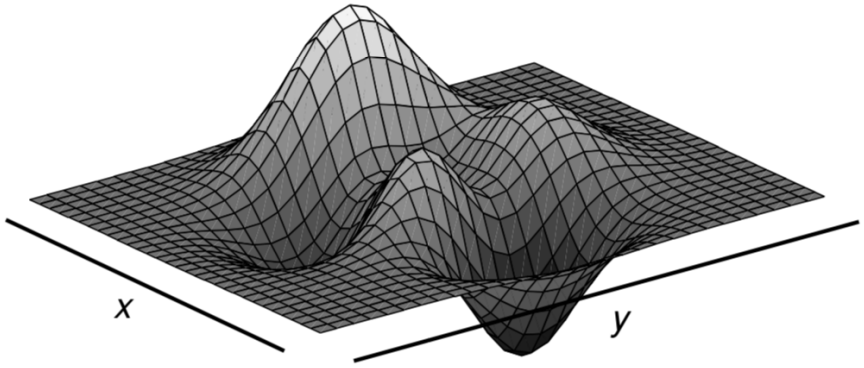


Figure 1.2: Fitness landscape where x and y coordinates represent genetic variation and height indicates to the fitness of an individual. Populations can take different paths when climbing up the adaptive peaks depending on the slope of the fitness. (Figure source *The Phylogenetic Handbook* [16])

environment, the effective population size and random genetic drift. Once the fitness peak is reached, neutral evolution will take over and mutations will get fixed only due to random genetic drift. The sharpness of fitness peaks determines the variation in the population; less sharp, more variation. However, environmental changes can alter the fitness landscape and activate back the adaptive evolutionary forces [16].

1.2.2 Cancer as an evolutionary process

Evolution is the gradual change of genetic characteristics of biological populations, which states that all species on earth share a common ancestor from whom they all descend and evolved over a long period of time. In other words, evolution places all life forms on an evolutionary tree, where the tree root is the common ancestor of all species and branches indicate speciation [6]. The recent technological advances in extracting data from multilayer biological processes (DNA, RNA, Proteins) uncovered high intra and inter-tumour genetic heterogeneity and subclonal architecture of cancer cell populations. Hierarchical subclonal cell arrangements resemble the evolutionary trees first introduced by Charles Darwin in 1837 [17] (Figure 1.3). If we

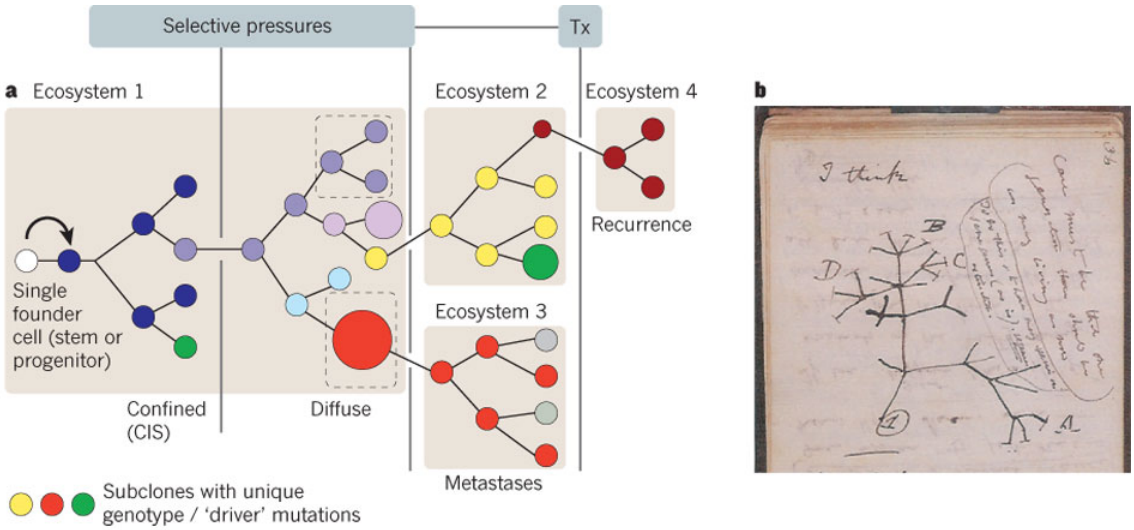


Figure 1.3: Subclonal branching architecture of cancer and Darwin's branching evolutionary tree of speciation. (Figure source *M. Greaves and C. Maley 2012* [17])

consider cancer cells as asexually reproducing individuals we can apply evolutionary theory to cancer, as all three building blocks of evolutionary dynamics - replication, selection, and mutation - are also the defining characters of cancer development [17, 18].

At the core of studying cancer development as an evolutionary process is the paradigm of molecular evolution, which as we introduced in the previous section, studies evolutionary dynamics at the DNA sequence level via linking Mendelian Genetics to Darwin's theory of natural selection and adaptation [19]. One of the key technologies from molecular evolution that makes it possible to observe spatial and temporal patterns of cancer development is genomic sequencing. Next generation sequencing technologies allow us to study cancer as an evolutionary process by tracing the stepwise accumulation of somatic mutations followed by sequential Darwinian selection [17].

Somatic mutations in cancer evolution are classified as drivers and passengers. Driver mutations cause cells to acquire advantageous phenotypes i.e. positive (Dar-

winian) selection (for instance proliferate more than other cells, or avoid cell death and increase survival) while passenger mutations are largely neutral - they don't affect the cell phenotype, but they can hitchhike with the driver mutation and thus increase in relative frequency in the population. A subset of cells that are driven to a higher relative frequency than expected under normal conditions by a driver mutation will also drag along all the previously acquired neutral i.e. passenger mutations. Usually, driver events are rare as it takes a significant amount of cell divisions and random mutations before a tumour suppressor gene is mutated (because the size of the coding DNA region is orders of magnitude smaller than the rest of the genome). In [20], passenger and driver mutations are compared to a train driver and passengers; the metaphor being that as GPS of the passengers in a train enables us to follow the location of the train, similarly, passenger somatic mutations let us trace the driver mutation. Passenger mutations are more abundant and hence more easily detectable, whereas, it needs considerable bioinformatic and statistical tools to detect driver mutations (as it is not very easy to detect phenotypic modifications at the cellular level). As such, passenger mutations are very important to study tumour cell population dynamics. This observation dates back in population genetics when Maynard Smith [21] termed the neutral passenger mutations hitchhiking mutations and highlighted their importance for studying gradual population changes.

Different subpopulations of cells in cancer are termed clones which is a parallel concept to sub-species from evolutionary biology. Providing a rigorous definition of a clone in cancer evolution is more challenging than in evolutionary biology. The main characteristic that discriminates between different species is their phenotype, which is relatively easier to measure than in cancer. In cancer evolution fully linking cancer genotype to its phenotype remains a challenging task as changes in genotype and phenotype at the cellular level do not often occur in a synchronous fashion, and their relationship is obscure and counterintuitive [12].

Throughout the thesis when using the term "clone" or "subclone" of a cancer cell population, I will be referring to the definition presented in [20]; where a clone is defined as "a group of cells with the same phenotype, which has expressed that phenotype consistently since their most recent common ancestor". Although it is hard to precisely measure phenotypic invariance over time, this definition is still providing a more rigorous explanation of the concept, as it specifies the common ancestor unifying the cells subgroup.

Studying the changes in relative frequencies of cancer cell subpopulations over time allows understanding the dynamics of cancer clone evolution. As discussed above, there are three main evolutionary processes that drive these dynamics: mutations, genetic drift and selection. Mutations and genetic drift are stochastic events, while selection is not random. Different combinations of these processes create different modes of cancer clone evolution dynamics such as branching, linear, punctuated, neutral evolution and natural selection.

When a specific lineage acquires fitness (proliferative and/or surviving) advantage as a response to some microenvironmental changes in order to adapt to the changing new environment, Darwinian positive (or adaptive) selection is said to be at play. Positive selection is one of the main forces of driving tumour progression [22]. Evolutionary forces can also act so that the lineages with decreased relative frequency get removed from the population. This is called negative or purifying selection and also contributes to progression, by, for example, removing potent neo-antigens [23].

While positive selection is rare, random mutations and genetic drift occur continuously over the lifetime of an organism. As they do not affect cells' phenotype, are regarded as neutral. When there are no selective forces acting on the population and most variations are neutral or negatively selected, such a dynamic is called neutral evolution [24]. Sometimes, when a positively selected cancer clone sweeps through the population, there is only one clone dominating the population, it will give rise

to another flow of neutral evolution until the next selective forces come into play.

Studying cancer development as an evolutionary process allows to better design treatment to potentially slow down and control its progression [17]. There has already been observed links to understanding cancer's evolutionary dynamics and its potential to patient clinical outcomes; clonal diversity of tumour cell population was correlated with tumour progression across many cancer types [25, 26, 27, 28, 29]. Tumours that show patterns of punctuated evolution were characterised to have less heterogenous driver mutations and more homogenous cell populations, and these tumours were prone to proliferate faster and seed metastases, and had worse clinical outcomes compared to the cases with more clonal diversity and subclonal aneuploidy as the later grew slowly and showed low gradual rate of driver mutation accumulation [30, 31, 32, 33]. There have also been studies that showed how the temporal order of mutation accumulation is associated with different clinical outcomes [33, 34, 35]. Cancers characterised by punctuated evolution, are the trickiest to treat as it is hard to detect such early events with strong metastatic potential, cases often referred to as "born to be bad". As shown by [36], preclinical models were able to detect metastatic dissemination before the malignancy was histologically identified. Studies have also shown how understanding the dynamics of treatment resistance, as a response to the selective forces caused by therapy interventions, can help preventing or delaying it [37, 38, 39]. As such, understanding cancer evolution dynamics promises to lead to better strategies for treatment intervention.

1.3 Mathematical/Computational modelling

"All models are wrong, but some are useful"
- George E. P. Box

Mathematical and computational models have been developed to describe complex systems using mathematics and simulations. When a model for a system is created, one can study the various elements of it, such as the interaction between its elements and eventually predict a probable outcome or the development of system dynamics. Modelling has been extensively used in cancer research as well, as it allows to quantify chemical and physical interactions during tumour initiation and development [40, 41]. When experimental procedures are accompanied by theoretical modelling, a better understanding of cancer clonal dynamics and microenvironmental factors can be achieved. Mathematical/computational models for cancer are usually classified as being deterministic versus stochastic and discrete versus continuous. In deterministic models, when initial conditions do not change, the general end state of the process also stays unchanged, whereas in stochastic models, there is randomness incorporated into the model and hence even with a fixed set of initial parameters, the end state will differ still. Discrete models treat cells as discrete entities and study their behaviour and interaction and tumour microsystem at the cell resolution, while continuous models consider concentrations of different cell types and study overall tumour morphology and distribution of nutrients, ignoring the roles of individual cell influences to the environment. In this section, I will briefly introduce agent-based modelling (an example of a discrete model) and the concept of stochastic simulations, followed by a very powerful technique in statistical modelling - Bayesian inference.

1.3.1 Differential equations

Physical quantities that change dynamically can be modelled by differential equations. A differential equation is a mathematical equation where the unknown variable is a function to be solved for. The equation describes a relation between the unknown function (for instance a physical quantity) to its derivatives (that usually represent the rates of change in the quantity). As such relations between a physical quantity and its rate of change is prevalent in many disciplines, differential equations are broadly used in various fields, such as physics, biology, engineering or economics.

While inventing the field of calculus, Newton and Leibniz were the first who introduced the use of differential equations. The following three primary types of differential equations were listed in [42] by Newton:

$$\frac{\partial y}{\partial x} = f(x), \quad \frac{\partial y}{\partial x} = f(x,y), \quad x_1 \frac{\partial y}{\partial x_1} + x_2 \frac{\partial y}{\partial x_2} = y \tag{1.1}$$

where he discusses the ways for solving these equations by infinite series and also the non-uniqueness of the solutions. Currently, there are several types of differential equations such as ordinal vs partial, linear vs non-linear, homogeneous vs inhomogeneous. Each type has its definition and properties of the equation that helps in choosing an approach for a solution to a problem that one intends to model as a differential equation system.

Differential equations are handy at describing how some dynamic systems or quantities change, although their usefulness depends on whether they can be solved or not. Unlike, for instance, algebraic equations, solving differential equations is a more difficult task as the solutions are not always obvious. It is sometimes unclear whether a given solution is unique or the equations can be at all solvable for a given system.

Differential equations can be used to model tumour progression, invasion and re-

sponse to therapeutic interventions. One basic and prominent model is the Malthusian law [43], that models tumour growths not at the individual cell level, but as a tumour cell population dynamics over time. When there is no treatment, the population tends to increase continuously, and when it gets large, individual cell contributions become negligible compared to the entire population. Hence, the population increase, if approximated, can be treated as continuous and a differentiable function of time. If $x(t)$ is a population size at time t, the tumour growth differential equation can be written as follows:

$$\frac{\partial x}{\partial t} = Kx, \quad K > 0. \tag{1.2}$$

Given the conditions $K > 0$ and $x(t_0) = x_0$, the equation has the following solution:

$$x = x_0 e^{K(t-t_0)} \tag{1.3}$$

which is the equation for exponential growth.

The Bernoulli equation can give a more realistic representation of tumour growth:

$$\frac{\partial x}{\partial t} = \alpha x - \beta x^2, \quad \alpha > 0, \beta > 0 \tag{1.4}$$

The βx^2 additional term is to minimise tumour expansion as time passes or to model therapy interventions such as radiation therapy or chemotherapy that have the exponential (quadratic) effect on individual cells. When β is very small compared to α, αx predominates and the tumour grows very rapidly – exponentially – for that time period. However, as the tumour cell population grows like diffusion, the βx^2 term becomes more dominant at a later time, and tumour growth rate starts to decrease. Tumour population growth dynamics modelled by equation (1.4) is called the logistic law of tumour growth.

Both of the above-described equations are examples of ordinal differential equa-

tions (ODE), i.e. they have one unknown function of one variable only. Cancer growth can also be modelled using partial differential equations (PDE). These kinds of models incorporate a different type of cell interactions or spatial constraints during tumour expansion. Experiments have shown that when a tumour grows by diffusion only, after some time, it reaches a dormant state [44]. At the early stages of tumour progression when the tumour cell population size is small, there are more nutrients available to tumour cells through diffusion, and hence almost all cells proliferate. While after some time when a tumour reaches a certain size for a given environment, it will create a scarcity for nutrients (especially in the centre of the tumour mass where it is harder for nutrients to reach the cells). Near the centre, cells will no longer proliferate and die, and thus only the cells on the tumour surface will carry on division and eventually tumour will grow only on its boundary.

Another example of the use of differential equations in tumour growth models are the deterministic models of well-mixed tumour cell populations. In well-mixed populations, stochastic effects are neglected and mean behaviour of the tumour's evolutionary process are modelled deterministically. One example of this approach is the following replicator equation [45] for describing frequencies of different genotypes:

$$\frac{\partial x_i}{\partial t} = x_i(f_i(x) - \varphi(x)), \quad i = 1, ..., n \tag{1.5}$$

where x_i is the frequency of the genotype i and $f_i = f_i(x_1, ..., x_n)$ denotes fitness which is a function of all other genotypes, and $\varphi(x) = \sum_j f_j(x)(x_j)$ is the average fitness of the population [45]. When fitness is constant over time $f_i(x) = f_i$, the replicator equation (1.5) is called the selection equation. In this scenario, the genotype with the highest fitness reaches fixation while all other genotypes go extinct (also termed survival of the fittest by [6]). The selection equation can be further extended to

account for mutation:

$$\frac{\partial x_i}{\partial t} = \sum_j x_j(f_j p_{ji} - \varphi(x)x_i), \quad i = 1, ..., n \tag{1.6}$$

where p_{ji} is the probability of a mutation from type j to type i. Using the equation one can predict the threshold mutation rate beyond which the genetic information is modified to the extent that the population can no longer be maintained [46, 47].

Differential equations cannot be applied to describe complex system dynamics that incorporate different levels of stochasticity. Besides, it is impossible to integrate the spatial dynamics and constraints of a system studied in space. Hence, we decided to develop a stochastic simulation model of cancer growth using the lattice-based modelling approach described below in the following sections.

1.3.2 Cellular automata

A cellular automaton is a lattice-based discrete model that consists of a grid of cells where each cell has a finite number of states. The grid can have any finite dimension. The model starts by defining a grid and placing cells on it with their predefined states. The set of surrounding cells for a given cell is called a cell neighbourhood. In a new generation, a new state is assigned to each cell simultaneously according to some fixed rule (usually described by a mathematical function). The cell state update depends on the cell's current state as well as its neighbourhood cell states. There are also models [48, 49] where cell states do not update simultaneously, but rather one by one either stochastically (stochastic cellular automaton) or again with some fixed but different rules per cell that are applied to each cell individually and asynchronously (asynchronous cellular automaton).

Cellular automaton models have been broadly used in cancer research [50]. One main advantage of cellular automata for cancer growth models is their ability to

model the fate of each cell individually and explicitly. Although for cancer, usually, a large ensemble of cells need to be simulated and can become computationally very expensive.

Using cellular automata one can model patterns and effects of cancer cell migration as well. For example, in [51] they simulated Moran's process on 1D and 2D grids and found that migration can stimulate a mutant cell's ability to invade an existing clone and take over the population. Their model also showed that the migratory phenotype tends to be selected for a large-scale cell death that can explain how chemotherapy provides a selection mechanism for highly invasive and migratory cancer cells.

More complex 3D models of tumour progression have also been introduced [52], that account for different vascular processes, blood flow, angiogenesis, nutrient and growth factor distribution, as well as cell movement and interaction between normal and tumour cells. The models like this can provide a means to study the spatiotemporal evolution of a tumour and its response to therapy.

1.3.3 Agent-based models

Agent-based models discretise a system into autonomous decision-making units called agents. There are a predefined set of rules that each agent follows and based on these rules and the surrounding environmental conditions, including competitions with other agents, makes a next step decision. The defining characteristic of agent-based modelling is the repetitive competitive interplay between the agents that are modelled using computational methods that study the dynamics of the system which is often impossible to model using mathematical formulas [53]. A very basic agent-based model can provide powerful insights into a complex system dynamics that it was designed to mimic.

One of the most effective features of agent-based models is the ability to cap-

ture the emergent phenomenon of a system. Reducing a system to its constituent parts is not always straightforward, as there are interactions between entities that lead to complex emergent patterns. Describing such individual-based behaviour and relationship is difficult with differential equations as they tend to miss important fluctuations. Also, agent behaviour can get so complex, that the differential equations describing it, will, in turn, get exponentially complex and intractable. Another main feature is that it is very simple to add stochasticity to the model. A desired source of randomness can be applied to the agent's behaviour at a desired time and location, in opposite to when random noise is applied to an aggregate differential equation. And one final benefit of agent-based models is that they are easy to programme and tune the model agent rules and attributes.

There are lattice-based or lattice-free instances of agent-based models. In lattice-based models, agents move on a spatially discretised grids (cellular automata is one such model that will be described into more details in chapter 2), while in lattice-free models, agent velocity and location are modelled by continuous variables that respond to the environmental forces. Agent-based modelling is widely used in cancer research as it captures well spatial effects of tumour growth dynamics and heterogeneity [54]. It has also been applied to model dynamics of different cancer types, such as melanoma [55], breast [56], colorectal [57], lung [58] [59], liver [60] and metastases [61].

1.3.4 Bayesian statistics

In 1763, Thomas Bayes, by introducing his "Bayes Theorem", established the second most widely used inference technique and reasoning in the field of statistics [62]. He essentially changed the approach and philosophy regarding statistical inference. In frequentist statistics, probabilities are the frequencies of random events that one can calculate after running repeated trials of an event for a long period of time [63].

Whereas for Bayesian thinkers, probabilities are beliefs that are not fixed and can be updated after new data is gathered. Also, while a frequentist statistician would try to remove as much uncertainty as possible by producing as accurate estimate as possible, a Bayesian statistician, on the other hand, would, in contrast, keep the uncertainty and refine it by updating his beliefs after seeing a new data generated as a new evidence.

Bayes Theorem derives a probability of an event by updating a prior knowledge of the event with a conditional probability of the event given the data:

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y) \tag{1.7}$$

The conditional probability $p(y|\theta)$ is called the likelihood function and it is the probability of seeing the newly generated data with a given parameter set θ. $p(\theta)$ is a prior knowledge/belief about the parameter and $p(y)$ - the probability of generating the data. $p(\theta|y)$ is the posterior (updated prior) probability about the event. There are different methods to derive and calculate the likelihood function using either analytical methods or resampling techniques that are computationally intensive.

While the frequentist approach suffers from sample size as the calculation of p-values and confidence intervals heavily depend on the number of performed or repeated experiments, the main criticism towards the Bayesian approach is regarding its use of an arbitrary prior when there is no prior knowledge about the system. Also, Bayesian inference techniques require a lot of computation but with the advent of computer power and development of new algorithms like Markov Chain Monte Carlo, Bayesian methods have started to get more popularity and are more broadly used than in the previous century when the field of statistical inference was dominated by the frequentist approach [64, 65].

We applied the Approximate Bayesian Computation (ABC) inference method to infer our stochastic simulation model parameters.

1.4 Thesis objectives

After the introduction and background presented here in Chapter 1, the remaining thesis consists of four other chapters. In Chapter 2, I first discuss the subclonal architecture of cancer and methods to trace selection in cancer evolution. Then introduce my initial approach of using phylogenetic tree analysis on tumour samples to measure selection. And finally, I present our study on how spatial constraints of a growing tumour impact our ability to infer cancer evolutionary dynamics.

Another important topic in cancer research is understanding the interlaced processes of genetics and epigenetics that are governing cancer cell behaviour. It has been shown that there are patterns of single base-pair substitutions within genomic binding sites that have been seen across all cancer types [66]. Motivated by this study and also by the work of Alexandrov et al. of mutational signatures in cancer [67], we decided to examine the distribution of different mutation types across epigenomic regions of breast cancer and measure the strength of associations between the two. I present this study and results in Chapter 3.

Chapter 4 is also on studying the link between the genetics and epigenetics of cancer but through the chromatin this time. Chromatin aberrations, such as changes in DNA methylation, histone modifications or distorted nucleosome remodelling, have been observed to be one of the sources of tumorigenesis. The time point estimates of different chromatin region modifications can be associated with external non-stochastic environmental factors and appropriate measures taken against further development of the disease. Here, I introduce the method I developed to estimate times to chromatin aberration events and the current limitations of its applicability due to the scarcity of the required relevant data.

I conclude my thesis with Chapter 5 which gives the overall discussion of the work presented in this thesis, summarises the main results and suggests future directions.

Chapter 2

Measuring clonal selection

2.1 Introduction

Cancer is an evolutionary process fuelled by genomic instability and intra-tumour heterogeneity (ITH) [17]. ITH leads to therapy resistance, arguably the biggest problem in cancer treatment today [68]. Recently, seminal studies have attempted to quantify ITH by either looking at subclonal mutations in deep sequencing data from single bulk samples [69, 70], or by taking multiple samples of the same tumour, the so-called multi-region sequencing approach (reviewed in [71]). Phylogenetic approaches are then used to reconstruct the ancestral history of cancer cell lineages [72]. However, one important difference between phylogenetic analyses in cancer and classical phylogenetic analyses of species is that each cancer sample is not a single individual, but a mixture of different cancer cell subpopulations and non-cancer cells [73].

The problem is usually tackled by performing subclonal deconvolution of the samples to separate the different subpopulations [69, 74]. However, these approaches do not account for the spatiotemporal dynamics that generated the data. To study the evolutionary dynamics of individual tumours, mathematical and computational

models of evolutionary processes are widely employed [40, 41, 75, 76]. Many of these models are rooted in theoretical population genetics, a field that quantifies the evolution of alleles in populations and that is central to the modern evolutionary synthesis [77]. More recently, spatial models have also been used [78, 79, 80, 81, 82, 83, 84, 85, 86, 87]. However, seldom have mathematical and computational models of cancer evolution been directly connected to next-generation sequencing data from human tumours. Recent works have shown that combining theoretical modelling and cancer genomic data allows for measurement of fundamental properties of the tumour evolutionary process in vivo, such as mutation rates and strength and onset of subclonal selection events [86, 88, 89].

In this chapter, I will first discuss the subclonal architecture of cancer and methods to trace selection in cancer evolution. I will then introduce our initial approach of using phylogenetic tree analysis on tumour samples to measure selection. Finally, I will present our study on how spatial constraints of a growing tumour impact our ability to infer cancer evolutionary dynamics (which has been published in PLoS Comput Biol.). We combine explicit spatial evolutionary modelling with a synthetic generation of multi-region bulk and single-cell data, thus providing a generative framework in which we know the evolutionary trajectories of all cells in a tumour and can examine the genomic patterns that emerge from the sampling experiment. We show that spatial constraints, stochastic spatial growth and sampling biases can have unexpected effects that confound both the interpretation and inference of the perceived evolutionary dynamics from cancer sequencing data. We also present a statistical inference framework that begins to account for some of these confounding factors and recover aspects of the cancer evolutionary dynamics from various types of multi-region sequencing data as well as single-cell data.

2.2 Subclonal architecture of cancer

2.2.1 Clone frequency distribution to trace selection

In chapter 1 we introduced three different forces that drive cancer evolution: mutation, genetic drift and selection. Even though mutation and genetic drift are purely random processes and selection is not, the latter generates the most complex patterns. There is still a lack of quantitative analytical models for selection in contrast to mutation and drift that have been modelled using Poisson and Markov processes, respectively. Currently, there are three main approaches for detecting selection that are based on clone frequency, patterns of mutational processes or phylogenetic tree analysis.

Clone frequency-based methods study the Variant Allele Frequency (VAF) distribution to detect an overrepresentation of a lineage compared to the expectations under neutral evolution. This approach has been motivated by the discoveries that the shape of the VAF distribution under neutral conditions and in a well-mixed population can be estimated analytically. Specifically, the distribution of $m(f)$ - number of mutations per allele frequency – follows $1/f^2$ [88, 90, 91, 92, 93]. If clonal evolution is under selective restraints, the frequencies of the driver and hitchhiking mutations will increase and hence the VAF distribution will deviate from neutral dynamics. As such neutral evolution can be treated as a null model against which selection could be tested and identified [94].

Detecting deviations from the neutral null model, however, suffers from a lack of power [89]. It is very hard to detect weak selection as it causes only slight changes in the VAF distribution and hence under the limited sequencing depth, selected subclones will likely get undetected, especially if they arise late during the tumour progression [89, 95]. Spatial constraints and sampling limitations also play a major

role in identifying selection.

Another approach to detecting clones whose frequencies vary disproportionally compared to the background population is analysing longitudinal samples per individual/tumour instance [96]. Longitudinal sampling would also help tackle the scenarios when a selected subclone sweeps through the population and reaches fixation after which the tumour evolution dynamics will revert back to neutral. However, collecting multiple samples over time has significant technical and ethical issues; taking a sample from a patient might be very important from the scientific point of view but unnecessary medically. Although, given the fast-paced development of the next generation sequencing technologies and associated price drop, analysing circulating cell-free DNA might be able to address the problem [97].

2.2.2 Intratumoral genetic heterogeneity and neutral evolution

Mutations cause genotypic and phenotypic variation, and consequently increase heterogeneity. Selection and genetic drift, on the other hand, make clones increase or decrease in relative frequency (some clones might go extinct while others get fixed in the population) and thus they reduce heterogeneity.

Unlike selection, which is a non-random process and has a larger effect in modifying lineage frequency, drift is an inherently random process. A lineage might produce more offspring than others just by a random chance when selecting individuals for the next proliferation step and thus increase in relative frequency. This effect will be stronger in a small population, while in a large population it will cause insignificant changes that might not be noticeable [98].

When some lineages show proliferative advantage or increased survival rate, then they are said to be driven by selective forces. When there is no such functional variation, then the population is said to be functionally homogeneous. In a func-

tionally homogeneous population, one might think that there is only one clone as the population is phenotypically stable. But the population can never be genotypically stable as mutations and genetic drift does occur constantly and thus they increase genetic variation and create heterogeneity. Such a scenario of population dynamics is called neutral evolution - functionally population is homogeneous while genetically heterogeneous [20]. Under such a scenario, the maximal genetic heterogeneity is created as there is no selection to remove variation. However, there is one peril to this case; sometimes environmental conditions can change so that a neutral genetic variation becomes functional and causes selection of some lineages. This observation, first made by Luria and Delbruck in 1943 in their famous experiment where they observed pre-existing resistance in bacteria [91], initiated the thinking that all variations are pre-existing at the origin.

Patients usually have different patterns of treatment response because each individual patient is unique with regard to their genome, microbiome, lifestyle, environment, disease history and exposure to drugs. When it comes to cancer, response to treatment gets even more complex to predict, as heterogeneity has been observed to be present not only between tumours but also between different subclones of the same tumour cell population [99]. Intra-tumour genetic heterogeneity and changes in the tumour microenvironment have been discovered to have a high impact on treatment response and disease progression [68, 100]. Measuring degrees of heterogeneity from experiments and clinical trials is difficult, more so the dynamics of spatial heterogeneity and as such computational modelling approaches are more broadly used to address the challenge. In the following sections, we will present such a computational approach that models spatial heterogeneity.

2.3 Phylogenetic analysis in cancer

2.3.1 Phylogenetic trees in species vs tumour evolution

Cancer cells divide and accumulate mutations. Phylogenetic analysis in cancer treats the clonal subpopulations as independent taxa and tries to apply different methods of phylogenetic tree reconstruction to infer the tumour cell phylogeny. The evolutionary history of cancer subclonal arrangements can then be used to test different hypotheses about cancer evolution [101]. However, even though cancer is an evolutionary process [17], some of the evolutionary characteristics are different in cancer versus species, especially when it comes to applying the same phylogenetic reasoning to cancer as to species.

There are four main areas aspects of which differ between cancer and species. These are types of aberrations, mutation rates, selection intensity and the level of heterogeneity [72]. Cancer evolution is often characterised with hypermutability (chromosomal instability, microsatellite instability, elevated point mutations, copy number variations, kataegis, chromothripsis, chromoplexy), plus mechanisms of hypermutability vary over time and between tumours which is not characteristic to species evolution.

The idea of constructing tumour phylogeny was first suggested by Tsao et al. [102] and then originally implemented by Desper et al. [103]. Following a decade of collaboration between evolutionary and computational biologists and many more implementations of similar analysis, lead to the establishment of the field called tumour phylogenetics. There are various methods used in inferring tumour phylogenies that differ by the type of data used (SNV vs CNV vs DNA methylation etc), type of study design (cross-patient, multi-region bulk sampling within one patient or single-cell sampling) and types of mathematical models applied. Most mathematical

and computational models for tumour tree inference have been adapted from species phylogenetics, such as maximum parsimony [104], distance based [105], maximum likelihood [106] or Bayesian probabilistic modelling approaches [12, 107, 108]. New methods have also been developed that account for the different characteristics of tumour versus species evolution [109, 110, 111, 112].

2.3.2 Bulk vs single-cell sequencing phylogenetic trees

Some studies collect multiple bulk samples per patient and perform phylogenetic tree inference to reconstruct the evolutionary history of a single tumour. Similarly, single-cell based studies infer evolutionary dynamics also within one tumour but using cell-to-cell variations.

When multiple bulk samples are sequenced from a single patient, they consist of a mixture of several cell lineages [104, 113, 114]. This is usually tackled by first reconstructing the subclonal architecture from the bulk samples, known as subclonal deconvolution, and then performing phylogenetic inference on the inferred subclones [115]. Tools developed for clonal deconvolution are SciClone [116], PyClone [117] and Clomial [118]. If subclonal deconvolution is not performed prior to tree inference, the estimated trees would resemble clustering of tumour bulk samples rather than its evolutionary history.

The development of single-cell tumour phylogenetics precedes single cell sequencing and was based on more limited profiling of single cells using microsatellite [119] of FISH [120] markers. For the first time single-cell sequencing to tumour phylogenetics was introduced by Navin et al. [105] and since then it became one of the prominent fields in cancer research. Since single-cell sequencing provides the means to infer genotypes of individual cells, it can provide a significant advancement in inferring tumour evolution (more so as subclonal deconvolution of tumour cell populations from bulk biopsies remains to be computationally challenging still).

However, techniques for single-cell sequencing are being improved; there are challenges to be overcome related to cell isolation, genome amplification to be scaled to whole-genome or whole-exome assays and high levels of allelic dropout [105, 121].

2.3.3 Tree balance indexes

As a preliminary analysis, we decided to apply tumour phylogenetic tree analysis for detecting selection using tree topology and branch length distributions.

One of the most used and well-studied measures of tree topology is balance index, which is a degree of similarity among the numbers of descendants that internal nodes produce per lineage [122]. A well-balanced tree represents a neutral, while a less balanced one - a selective tumour growth phylogeny. In our analysis we tested three different balance indexes [123]: (1) Sackin (sum of the depths of tree leaves), (2) Colles (sum of the net number of descendants of children nodes for all tree nodes) and (3) TCI (total cophenetic index - sum of the depths of the least common ancestor nodes for each tree node pairs):

$$S(T) = \sum_{i=1}^{n} \delta_T(i) \tag{2.1}$$

$$C(T) = \sum_{v \in V_{int}(T)} bal_T(v) \tag{2.2}$$

$$\Phi(T) = \sum_{a \leq i < j \leq n} \delta_T(LCA_T(i,j)) \tag{2.3}$$

where $\delta_T(v)$ is the depth of a node v in a phylogenetic tree T, which is the length (in number of arcs) of the unique path from the root r to v; $bal_T(v)$ is the balance value of v and is equal to $|K_T(v_1) - K_T(v_2)|$, where K is the number of descendants for a node and $v_{1,2}$ are the children nodes of the node v; $LCA_T(v_1, v_2)$ is the lowest common ancestor of a pair of nodes v_1, v_2 meaning that it is the unique common

ancestor of them that is a descendant of every other common ancestors of the nodes.

We computed the normalized versions for each balance index under two stochastic models of evolutionary tree growth - the Yule and the uniform models. In addition to the balance indices, we also calculated the mean branching time - the distances from each node to the tips averaged over all nodes, and the pairwise distances between the branch lengths for the pairs of nodes and tips separately (indicating for patterns of clustering if present). We compare these balance indexes for simulated neutral versus selective tumour single-cell sample trees in section 2.6 (Figure 2.21) using our simulation model, which is also introduced in the following sections.

2.3.4 Statistical methods on phylogenetic trees

Considering that the occurrence of a mutation per cell division can be modelled as a random event, we can apply several other statistical methods to detect selection on a phylogenetic tree. Namely, at a given time point t, if a clone divides at a rate λ, the accumulated mutations should follow the Poisson distribution with rate $\mu\lambda t$. Another clone with a different μ or λ, will have accumulated mutations with a different mean. To identify these differences, we employed the following two statistical methods of Poisson mean comparison: CAT [124] and Changepoint [125]. CAT is a permutation-based test that detects the differences between the unbalanced group means of the Poisson distributed random variables, whereas Changepoint identifies the differences in the rates of a non-homogeneous Poisson process.

Computational Approach Test - CAT

Let $X_{i1},...,X_{in_i} \sim Poisson(\gamma_i)$ where X_{ij} are the counts belonging to the i-th group and coming from the j-th sample, $i = 1,...,I,\ j = 1,...,N$ and γ_i is the Poisson rate parameter of the i-th group. Since the distribution of the sums of independent

Poisson random variables is still Poisson we have that $Y_i = \sum_{j=1}^{n_i} X_{ij} \sim Poisson(n_i\gamma_i)$ where n_i is the number of observations in the i-th group. Then the null hypothesis can be specified as follows: $H_0 : n_i\gamma_i = n_j\gamma_j$ for all $i \neq j$ $i, j = \overline{1,I}$ while the alternative: $H_A : n_i\gamma_i \neq n_j\gamma_j$ for at least one pair of $i \neq j$, $i, j = \overline{1,I}$.

The CAT procedure specifies the null hypothesis in the following way: $H_0 : \eta = 0$ versus $H_A : \eta > 0$, where η can be chosen from any suitable scalar measure; for our analysis we chose $\eta = \sum_{i=2}^{k} (\sqrt{\gamma_i} - \sqrt{\gamma_1 A_i})^2$ and $\eta = \sum_{i=2}^{k} |\sqrt{\gamma_i} - \sqrt{\gamma_1 A_i}|$ (k = number of groups, A_i i-th alternative hypothesis), as was justified by the authors to be the most appropriate measures for Poisson distributed data in the paper [124].

The following are the steps of the CAT algorithm:

STEP 1: $\hat{\eta}_{ML} = \sum_{i=2}^{k} (\sqrt{\hat{\gamma}_{i(ML)}} - \sqrt{\hat{\gamma}_{1(ML)} A_i})^2$ MLE estimate of η is calculated, where $\hat{\gamma}_i = Y_i/n_i$ is an MLE estimate of γ_i

STEP 2: Under the null hypothesis that $\gamma_i = \gamma_1 A_i$ $\forall i, i = \overline{2,I}$ the restricted MLE of γ_1 is calculated through the corresponding log-likelihood function for γ_1: $L = -\gamma_1 (\sum_{i=1}^{I} n_i A_i) + \sum_{i=1}^{I} Y_i ln(\gamma_1 n_i A_i) + C$

From which the restricted MLE of γ_i is obtained: $\gamma_{1(RML)} = \frac{\sum_{i=1}^{I} Y_i}{\sum_{i=1}^{I} n_i A_i}$

STEP 3: $Y_i \sim Poisson(n_i \gamma_{1(\hat{RML})})$ data is generated for $i = 1, ..., I$ a desirably large number of times, say - M and at each data generation step
$\eta_m = \sum_{i=2}^{I} (\sqrt{\hat{\gamma}_{i(ML)}^{(m)}} - \sqrt{\hat{\gamma}_{1(ML)}^{(m)} A_i})^2$ is calculated, where $\gamma_{i(ML)}^{(m)}$ is obtained from the simulated data in the m-th replication, $1 \leqslant m \leqslant M$

STEP 4: p-value for testing $H_0 : \eta = 0$ vs $H_A : \eta > 0$ is defined by:

$p_{CAT} = \sum_{m=1}^{M} I[\hat{\eta_{(m)}} > \eta_{(\hat{ML})}]/M$, where $I[.]$ is the indicator function.

Changepoint - CPT

Let $y = (y_1, ..., y_n)$ be the sequence of ordered data; Changepoint analysis aims to detect a time τ such that the sequences $(y_1, ..., y_\tau)$ and $(y_{\tau+1}, ..., y_n)$ have different statistical properties. There could be only one such a changepoint τ or multiple changepoints $\tau_1, ..., \tau_m$, and therefore there exist different single and multiple changepoint detection algorithms.

Single changepoint detection is performed as hypothesis testing, where $H_0 : m = 0$ and $H_1 : m = 1$ via likelihood ratio test statistic. Under the null hypothesis the maximum log-likelihood is $Log(p(y_{1:n}|\hat{\theta}))$ where $\hat{\theta}$ is MLE of parameters and $p()$ is the probability density function of the data, while under the alternative hypothesis the maximum log-likelihood for a given changepoint τ_1 is the $ML(\tau_1) = log(p(y_{1:\tau_1}|\hat{\theta}_1)) + log(p(y_{\tau_1+1} : n|\hat{\theta}_2))$. The test statistic then is as follows:

$$\lambda = 2[max_{\tau_1} ML(\tau_1) - Log(p(y_{1:n}|\hat{\theta}))] \tag{2.4}$$

Choosing the appropriate value for a threshold C, so that the test rejects the null hypothesis if the $\lambda > C$, is still an open research question. Currently, there are several penalty methods such as Asymptotic penalty which is equivalent to p-values, AIC, BIC, MBIC that can be specified to reach the decision on rejecting the null hypothesis or not [126].

For the multiple changepoints, there are several, optimization based methods, that try to find the maximum of $ML(\tau_{1:m})$ over all possible combinations of $\tau_{1:m}$. Three main algorithms that solve the optimization problem are used widely: binary segmentation, segment neighbourhood and pruned exact linear time (PELT) [125]. For our analyses, we applied PELT that incorporates dynamic programming techniques to obtain the optimal segmentation for m changepoints by reusing the information that calculated for $m-1$ changepoints and considering BIC (Bayesian

Information Criteria) penalty values obtained at each step.

2.3.5 Coalescent theory and common ancestor

Phylogenetic methods however do not fully account for the inherent randomness of evolution because the true genealogy of samples is not known. The coalescent theory models the past evolutionary forces on current genetic variation stochastically, assuming genealogy is random [127]. It traces back the Most Recent Common Ancestor (MRCA) of all tree nodes and uses only the individuals that are ancestral to the sample rather than keeping track of the entire population, and thus, is more computationally efficient than the phylogenetic methods [127]. One of the most important targets of modelling MRCA features is time i.e. given a set of nodes on a tree how much further backward does one need to go to encounter all present nodes' most recent common ancestor - the Time to the MRCA or TMRCA. We considered that estimating the TMRCA of subclones on a phylogenetic tree would be interesting as cells within a clone should have a more recent common ancestor than cells from two different clones, and hence selective subclones could be identified. In the following, I describe the technical procedure of deriving the posterior distribution for coalescent times for a growing population size.

Let W_j be the time when the sample has j distinct ancestors, $j = 2, ..., n$, then W_j has exponential distribution with parameter $j(j-1)/2$. Two important quantities associated with each phylogenetic tree are the height and the length of the tree denoted by T_n and L_n respectively $T_n = \sum_{j=2}^{n} W_j$ and $L_n = \sum_{j=2}^{n} jW_j$

In the coalescent theory the times at which mutations occur are modelled by a Poisson process of a constant rate $\theta/2$. Thus, for a given tree length L_n the number of mutations S_n on the tree follows the Poisson distribution: $P(S_n = k|L_n = l) = Pois(k, \theta l/2)$.

The posterior distribution of T_n is then modelled as the prior distribution of T_n

multiplied by the probability of observing data D conditional on a given time $T_n = t$:

$$f_{T_n}(t|D) \propto f_{T_n}(t)P(D|T_n = t) \tag{2.5}$$

And replacing D by the observed values of S_n gives us the exact form for the posterior:

$$f_{T_n}(t|S_n = k) \propto \int_0^\infty f_{T_nL_n}(t,l)Pois(k,l\theta/2)dl \tag{2.6}$$

where $f_{T_nL_n}(t,l)$ is the joint probability density function of T_n and L_n under the coalescent theory. The evaluation of $f_{T_n}(t|S_n = k)$ from (2.6) can be performed by stochastic simulation using rejection sampling algorithm [128].

We modified the standard coalescent simulation algorithm, as it assumes fixed population size, which is not the case for cancer cell populations. We scaled the generation time by the population size $\tau = \frac{1}{2N}$ also referred as coalescent time. Now let $N(t)$ be the population size as a function of t, then the amount of coalescent time traversed in going from generation i to $i+1$ is $\frac{1}{2N(i)}$, and from 1 to t - $g(\tau) = \sum_{i=1}^{t} \frac{1}{2N(i)}$. Given that $g(\tau)$ is a strictly increasing function, we can easily convert it and calculate the number of generations $\tau = g^{-1}(t)$ corresponding to τ units of coalescent times. For our simulations I used exponential growth - $N(t) = N_0 e^{\beta t}$ and derived the mapping from the coalescent times to the generation numbers:

$$g^{-1}(t) = \log\left[\frac{\log(a-t)}{\frac{j!}{2(j-2)!}\frac{e^{\beta s}}{\beta}} + 1\right]\frac{1}{\beta} \tag{2.7}$$

Thus, by the aforementioned simulation procedure, we can estimate the time of onset of subclonal cell populations for an exponentially growing population structure.

2.3.6 Detecting selection from phylogenetic trees

In the attempt of measuring subclone evolution on phylogenetic trees, we applied the CAT, CPT and modified TMRCA methods introduced above. The tree branch lengths are proportional to the number of mutations detected during the time passed since the two nodes diverged. The particular aim of our study is to investigate whether there is a fixed mutation rate for the whole tree or there are small sub-trees or groups of branches that have different rate parameters.

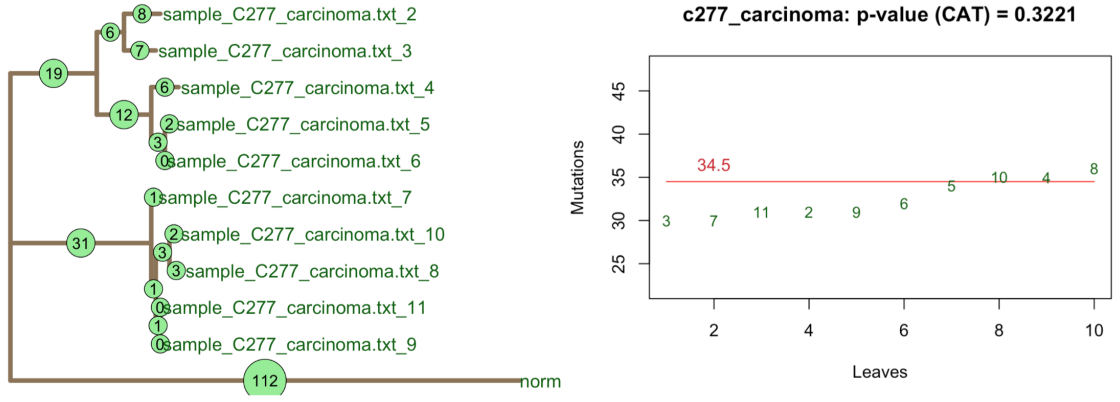
Branches that connect the root node to each leaf (outer) node were modelled as independent groups to identify subgroups within these groups. As the numbers of mutations from these branches are Poisson random variables, and the sum of the independent Poisson variables are still Poisson distributed, we model every leaf on a tree as a Poisson random variable with a rate parameter given by the sum of the component branch rates. Thus, the statistical reformulation of our study is to assess multiple comparison methods of Poisson means with limited sample size and unbalanced data, for which we used CAT.

We also modelled the number of mutations on a phylogenetic tree as a non-homogeneous Poisson process and estimated the times (referred as change points) when the rate parameter of the stochastic process changes. By detecting change points, we identify the subclones of a tree that have different mutation rates.

If there is a subgroup detected, either by CAT or CPT analyses, we split the trees into the detected subclones and estimate TMRCA for each subtree. Figure 2.1 shows an example of a real data tree (data source: [129] where they performed multi-region genome and exome sequencing of benign and malignant colorectal tumours) that visually looks balanced while both methods, CAT and CPT confirm that there is no significant sub-grouping present on the tree. In Figure 2.2 we can see an instance of an unbalanced tree, where CPT identified three major subgroups. As expected, the estimated MRCAs for each three subgroups differ considerably (Figure

2.3).

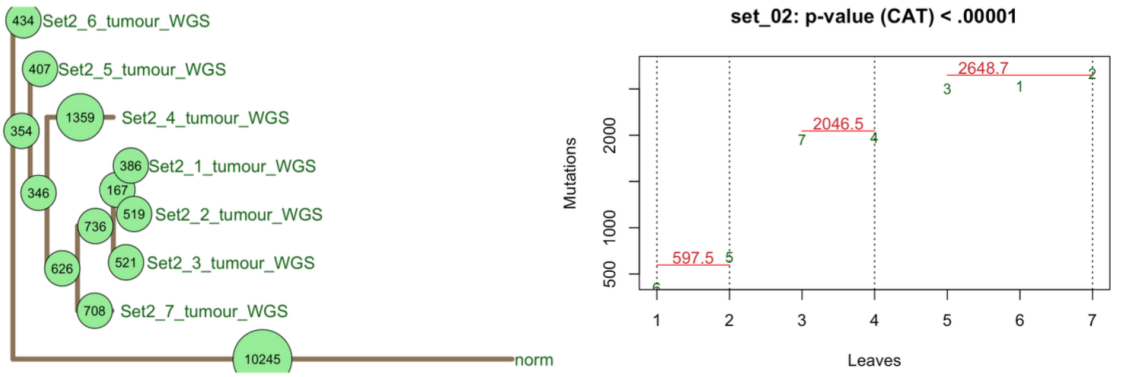
We found these methods are limited when applied to real data due to the limited amount of sampling. For this reason, we decided to take a different approach for measuring selection and developed a stochastic simulation model which is introduced and discussed in the subsequent sections.



(a) Phylogenetic tree

(b) CAT and CPT results

Figure 2.1: Balanced tree - an example of a real data phylogenetic tree that is inferred to be balanced both from CAT and CPT methods. **(a)** - maximum parsimony phylogenetic tree of C277 colon cancer patient; on the tips of the tree are the sample names, while the numbers in the green circles represent the accumulated mutations for the relevant nodes. **(b)** - C277 tree leaves ordered by the number of mutations; the horizontal red line represents the estimated overall group mean = 34.5 for all the leaves, indicating there was no changepoint/subgrouping detected within the leaves by CPT analysis. The p-value from the CAT analysis is 0.3258, which verifies the CPT result.



(a) Phylogenetic tree

(b) CAT and CPT results

Figure 2.2: Unbalanced tree - an example of a real data phylogenetic tree that is inferred to be unbalanced both from CAT and CPT methods. **(a)** - maximum parsimony phylogenetic tree of set2 colon cancer patient; on the tips of the tree are the sample names, while the numbers in the green circles represent the accumulated mutations per node. **(b)** - set2 tree leaves ordered by the number of mutations; the three horizontal red lines represent the estimated means 597.5, 2046.5 and 2648.7 of the three subgroups detected by CPT analysis. CAT returned the p-value << .000001 validating the CPT results.

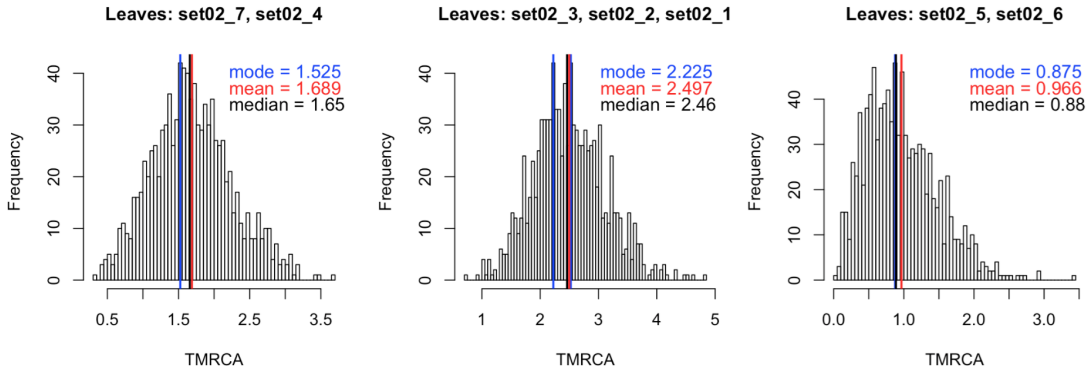


Figure 2.3: MRCA estimations - MCMC simulations of the MRCA distributions for three detected subtrees of the phylogenetic tree from Figure 2.2. Subtree (1) with leaves: set02-7 and set02-4 and estimated mean MRCA = 1.689, subtree (2) with leaves set02-3, set02-2 and set02-1 and estimated mean MRCA = 2.497, and subtree (3) with leaves set02-5 and set02-6 and estimated mean MRCA = 0.966. The estimated summary statistics differ significantly among all three subgroups.

2.4 Stochastic models of tumour expansion

(published in PLoS Comput Biol. 2019)

2.4.1 Stochastic birth-death processes

A stochastic process is a sequence of random variables that are defined on a common probability space and describe the evolution of some complex system. There are different types of stochastic processes with the most famous and widely used in many areas being Markov processes. The fundamental property of the Markov process is that conditional on the history of the system, the probabilistic state of the future does not depend on the past, but only on the present state of the system.

A branching process is a Markov process that models population dynamics where individuals in generation n produce a random number of offspring for generation $n+1$. In other words, it studies how population individuals reproduce and die independently but according to a specific probability distribution. There are many types of branching processes – discrete vs continuous time, single type vs multitype, population size dependent or independent, and have various applications from population biology and phylogenetics to cancer evolution studies [130, 131, 132, 133].

In 1960, Kendal, who in the famous Luria-Delbruck experiment allowed cells to grow as a birth-death stochastic process [134], established the stochastic Luria-Delbruck model, that since then has become the foundational approach for mathematically understanding cancer evolution. Kendal's stochastic birth-death model and its many extensions have been applied to study drug resistance [135, 136, 137], driver mutations[138, 139] and metastases[140, 141, 142].

The differential equation of the stochastic birth-death process assuming that in the time interval $(t, t+\delta t)$ an individual gives birth with probability $b\delta t$ and dies with $d\delta t$, is as follows:

$$\frac{\partial N_0(t)}{\partial t} = \partial N_1(t) \tag{2.8}$$

$$\frac{\partial N_n(t)}{\partial t} = b(n-1)N_{n-1}(t) - (b+d)nN_n(t) + d(n+1)N_{n+1}(t), \quad n \geqslant 1$$

where N_n is the population size at generation n. The solution of the equation is derived in [143] and is given by:

$$N_0 = \alpha, \quad N_n = (1-\alpha)(1-\beta)\beta^{n-1}, \quad n \geqslant 1 \tag{2.9}$$

where

$$\alpha = \frac{d(e^{(b-d)t}-1)}{be^{(b-d)t}-d}, \quad \beta = \frac{b(e^{(b-d)t}-1)}{be^{(b-d)t}-d} \tag{2.10}$$

Given this probability distribution, we can calculate the mean and variance of the population size at time t:

$$\mu_N = e^{(b-d)t}, \quad \sigma_N^2 = \frac{b+d}{b-d}e^{(b-d)t}(e^{(b-d)t}-1) \tag{2.11}$$

We can see that the standard deviation is of the orders of $e^{(b-d)t}$ indicating large expected variation in population size that makes it harder to infer the parameters of the system that depends on the population size. We will see an implementation of this process in our stochastic simulation model and the effects of small simulated population size on the inference we later performed.

2.4.2 The Gillespie algorithm

To model complex cell dynamics often methods from stochastic chemical reactions are applied. One such approach is the Gillespie algorithm which is the base algorithm

of our simulation model too and hence we will briefly describe the algorithm here first and its one variant of implementation that we applied for our model.

To describe the cell population dynamics in a hierarchically organised tissue, one needs to address the following question:

- If we assume that proliferation event is discretised, so that there is only one proliferating cell at a given time, and also there are different cell types in the population, which cell should proliferate next and what time does the system need to wait for this next proliferation event?

Let's assume we have n cell types and there are $X_i(t)$ cells of type i at time t, $i \in (1,...,n)$. Given some initial conditions $X(t_0) = x_0$, the goal then will be to estimate stochastic trajectories of the state space vector $X(t) = (X_1(t),...,X_n(t))$. Let's also assume that we have M possible reaction types $R_1,...,R_M$ (different mutation or proliferation pathways) in the system. Then the key question for modelling the system will be translated into estimating the probability of the next reaction type being j, $j \in (1,...,M)$ and taking place within the time interval of $(t+\tau, t+\tau+d\tau)$. Let's denote this probability by $P(\tau, j)d\tau$ and by r_j the number of distinct reaction combinations for reaction R_j at time t. Then if one follows the argumentation in [144, 145] one can easily derive the exact expression for the reaction probability density function:

$$P(\tau, j) = r_j c_j \exp[\sum_{l=1}^{M} r_l s_l \tau] \tag{2.12}$$

where c_j is the rate parameter of the reaction r_j.

Having this probability function, one can construct an exact stochastic simulation of the system. There are multiple ways for doing so, one of them being the method of first moments [145], that we applied for our model here. The simulation workflow of the method of the first moments is as simple as first drawing $\xi_1,...,\xi_M$ uniform

random numbers from the interval $[0,1]$ to construct M time intervals for the M reaction types:

$$\tau_j = \frac{1}{r_j c_j} \ln \frac{1}{\xi_j} \tag{2.13}$$

and then determining the cell and the reaction type to occur next by choosing the smallest τ_k from $k \in (1,...,n)$. The simulation step ends by incrementing the system flow time with this smallest chosen time interval.

2.4.3 Simulating tumour evolution

Here we develop and analyse a stochastic spatial cellular automaton model of tumour growth that incorporates cell division, cell death, random mutations and clonal selection. Each tumour simulation starts with a single "transformed" cell in the centre of either a 2D or a 3D lattice, and we model the resulting expansion of this first cancer cell. All events, such as cell proliferation, death, mutation and selection are modelled according to a Gillespie algorithm [146] the detailed derivation of one implementation of which is described in the preceding section above. In our model we account for different spatial constraints that are parameterised within our simulation. In order for a cell to divide, a new empty space for its progeny is required within the 8 neighbouring cells if we consider a 2D grid with Von Neumann neighbourhood. If no empty space is present, a cell can generate a new space by pushing neighbouring cells outwards (choosing a random direction of the push). In this scenario, the growth is "homogeneous" and all cells in the neoplasm can divide (Figure 2.4 A,B). Because all cells in the tumour can divide, this scenario leads to an overall exponential expansion (Figure 2.5 A,B). At some point during the simulation (Figure 2.4 A-D), within the original tumour population (blue cells), we introduce a new mutant (a new subclone – red cells) which may or may not have a selective

advantage. In the case of a neutral subclone (no selective advantage), the mutant cells divide exactly as all the other cells (Figure 2.4 A). We note that in this case, colouring a new subclone in red at a certain point during neutral growth is arbitrary, and equivalent to the marking of a lineage by a random neutral (passenger) mutation. In the case where the subclone has a fitness advantage, the mutant will, on average, grow more rapidly compared to the parental background clone, thus increasing in relative proportion over time (Figure 2.4 B and 2.5 B).

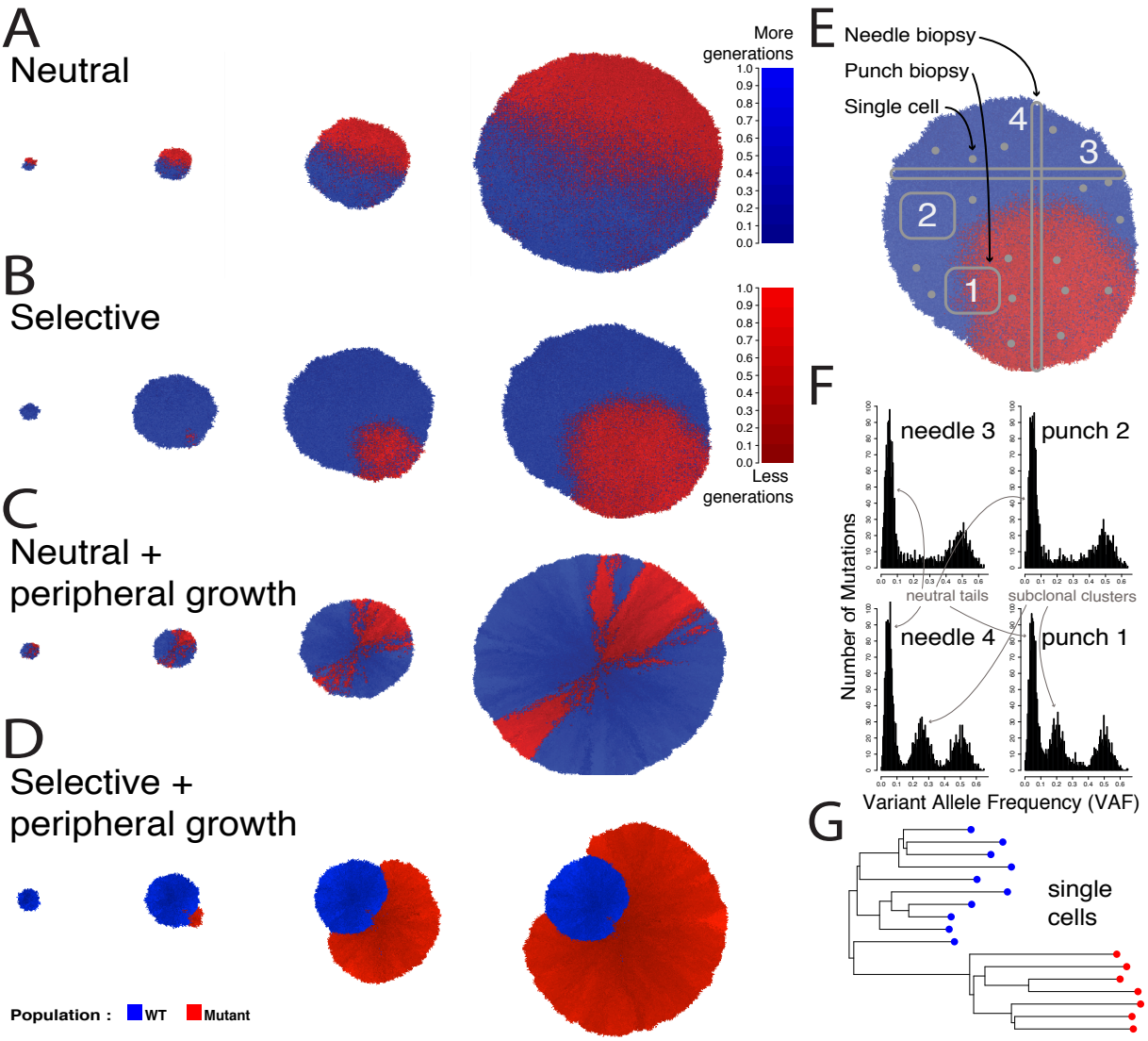


Figure 2.4: A spatial tumour growth model that simulates sequencing data. In our model we introduce a mutant at a given time t (blue = background clone; red = mutant subclone; shade is proportional to the number of generations the cell has gone through). (A) The new mutant subclone can have no fitness advantage (mutation is a passenger), giving rise to a neutrally growing neoplasm, or (B) have a fitness advantage $s > 0$ with respect to the background population (mutation is a driver), giving rise to differential selection in the tumour population. In addition, cells accumulate unique passenger mutations during each cell division. (C) In some tumours, it is likely that only cells close to the tumour border are able to proliferate due to the abundance of resources and space. We simulate this in our model as boundary driven growth, which gives rise to complex radial patterns. (D) When boundary driven growth is combined with selection, spatial effect can either amplify the growth of the new subclone, as in this exemplary case, or even decrease the effects of selection if the subclone, by chance, gets imprisoned behind the growing front. (E) In our simulation we also model the raising and spread of point mutations in the genome of cancer cells (all passengers and, when subclone is selective, one additional driver). We can simulate the sampling of punch biopsies (squares), needle biopsies (thin stripes) and single cells. (F) By simulating the noise and measurement errors of next-generation sequencing, we can generate synthetic realistic variant allele frequency distributions from the spatial simulations. (G) Single-cell data can also be simulated, in this case clearly showing the presence of a selected subclone demonstrated by the clade of "red" cells with a recent common ancestor.

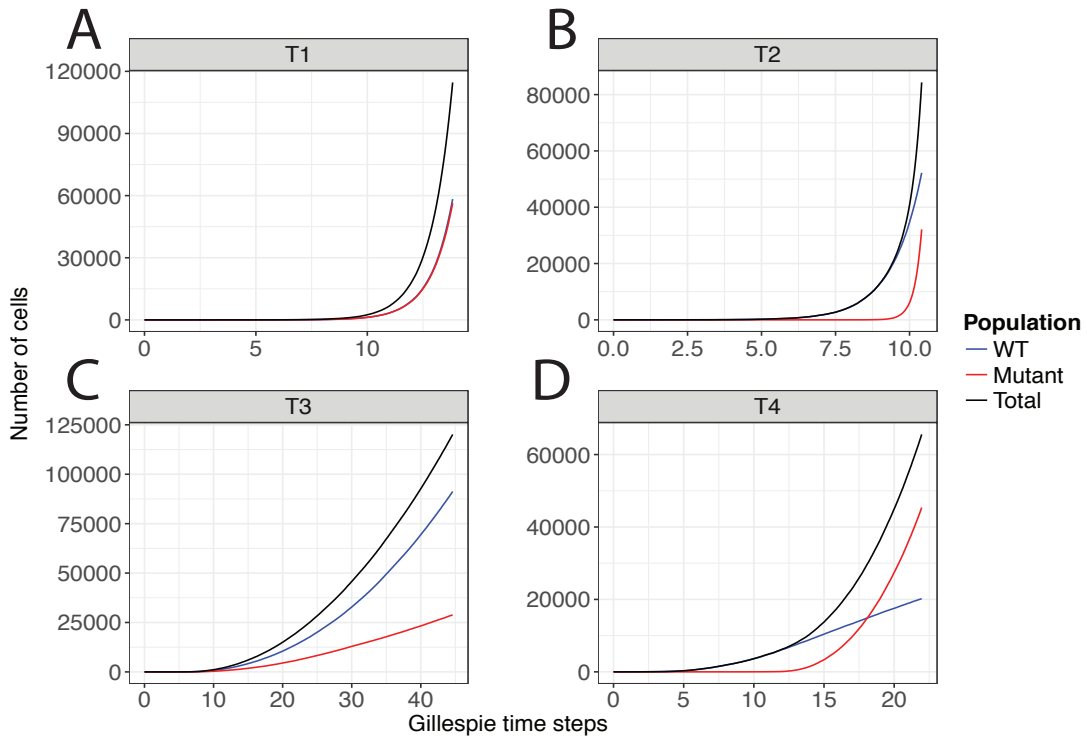


Figure 2.5: Growth curves. Tumour cell population growth curves for each of the representative cases: (A) neutral homogeneous, (B) selective homogeneous, (C) neutral boundary driven, (D) selective boundary driven. Wild type (WT) and mutant growth curves are plotted separately in addition to the whole population growth curves. Without the spatial constraints of our model, the growth curves are exponential as expected. (A, B) With the boundary driven growth the growth becomes polynomial. We can also see for the tumours with selection (B, D) how the mutant subpopulation outcompetes wild type cell population.

We also model "boundary driven" growth, where only cells that are sufficiently close to the border of the tumour can proliferate. Other cells may remain "imprisoned" in the centre of the tumour unable to proliferate because of the lack of empty space around them. Boundary-driven growth has been observed experimentally [147, 148, 149] as well as in model systems [150]. The magnitude of this effect is controlled in our simulation with the parameter a, which considers cell location

and defines the probability that a cell will push neighbouring cells to create empty spots depending on how far is the cell from the boundary (we present more details below). Boundary driven growth leads to a polynomial expansion (Figure 2.5 C). Importantly, in both the case of neutral mutants (Figure 2.4 C) and selected mutants (Figure 2.4 D), the spatial distribution of mutant cells in this scenario is strongly affected by the spatial constraints.

At each division, a cell has a certain probability to acquire additional somatic mutations, modelled with a Poisson distribution, with mean u, in line with many other previous models [40, 88, 89, 151, 152]. Notably, u is the average number of new somatic mutations per division for the whole genome of a single cell. We assume that both daughter cells can acquire mutations, that mutations are unique (infinite site model) and we neglect back mutations (infinite allele model). Finally, the large majority of mutations are assumed to be passengers (neutral), with a few driver alterations allowing for subclonal fitness advantages (e.g. subclonal populations in Figure 2.4 B,D). This is consistent with large-scale genomic sequencing studies indicating that in any given tumour, the number of driver events is generally small, while the number of passengers is often orders of magnitude larger [22, 151].

Importantly, our spatial model of tumour growth allows for the simulation of tissue sampling and genomic data generation. For instance, we can simulate the collection of punch biopsies where spatially localised chunks of tumour are collected (Figure 2.4 E). We can also simulate needle biopsies, where a long and thin piece of tissue is sampled (Figure 2.4 E). We can then simulate the genomic data generation process starting from the cells in the sample and the identification of somatic mutations. For example, we can simulate the sequencing at a given coverage using Binomial sampling of the alleles, the limits of low frequency mutation detection (e.g. minimum number of reads with a variant, minimum coverage), as well as non-uniformity of coverage leading to over-dispersion of the variant allele frequency

(VAF) of detected mutations. This allows generating realistic data from simulated tumours, e.g. in the case of the simulation of a diploid tumour with one selected subclone in Figure 2.4E, all needles and punch biopsies contained clonal mutations, shown as a cluster of variants around VAF=0.5 (Figure 2.4 F), and in the case of punch biopsy 1 and needle biopsy 4, also a subclonal cluster representing the growing subclone.

Details of the simulation model

We consider tumour cells as asexually reproducing individuals that die and divide with certain pre-defined probabilities. If b is the birth rate for each cell and d the death rate, then the growth of the population over time t is:

$$N(t) = e^{(b-d)t} \tag{2.14}$$

where $N(t)$ is the population size at time t, and $b-d$ is the net growth rate. At first, we assume that birth and death rates are constant over time, whereas the overall growth rate can vary over time due to the randomness of each birth or death event, as well as due to spatial constraints that can limit or promote cell division over time. We model spatial constraints with the boundary proliferation parameter a, which models the distance from the border of the tumour within which cells are allowed to proliferate even in the absence of space (by pushing neighbouring cells outwards). When $a \sim 1$ all cells can proliferate (homogeneous growth), and their growth is equivalent to an exponential expansion. When $a \sim 0$, cells can only proliferate if they have an empty space in their neighbourhood, resulting in only a small layer of cells at the tumour border being able to divide. In this case the growth curve can significantly deviate from equation (2.14).

In addition to cell division, we also model mutation and selection, where the latter can change birth and/or death rates. We model somatic mutations acquired

by each cell after division as a Poisson random variable – $Poisson(u)$, where u is the mean mutation rate. Thus, after each cell division, a random set of new unique mutations occur in each cell of the two cells resulting from the division. The majority of these mutations are passenger mutations and hence do not affect a cell's phenotype. However, they enable us to trace cell lineages uniquely in the final tumour. In addition, we also allow for driver mutation "events" that can lead to positive selection of a subpopulation of cancer cells: a driver event conveys a fitness advantage to that particular cell and its offspring, thus allowing the lineage to increase in frequency. Since we ask what is the distribution of mutations across space, rather than the expected waiting time of driver events as previously analysed [153], we introduce a driver mutation at a fixed time in our simulations, also to make simulations comparable and computationally efficient.

To simulate tumour growth in space with these four stochastic events – birth, death, mutation and selection – we have used a modification of the Gillespie algorithm [146].

Specifically, the simulation framework works as follows:

- **Initialization:** start with a 2D/3D grid with Von Neumann neighbourhood. Place the first tumour cell in the centre of the grid. Set time t=0.

Until a cell reaches a predefined grid boundary, repeat the following steps

1. Compute the reaction propensities according to the Gillespie algorithm. Each reaction event of birth (or death) has a functional form $f(x) = kx$; here x is the number of cells of type "x" (wild-type or mutant), and k is either the birth or death rate. The time of each event is obtained by sampling an exponential random variable with mean given by its propensity. The next event chosen is the one completing first (i.e., with smallest clock value, as in the so-called next reaction method [146]). Given the event, we increment time by its clock.

Note that these time steps do not correspond to population doubling times i.e. generations; doubling times can be retrieved scaling time by a factor $log(2)$.

2. If the next event is a cell division, we use a heuristic method to place the 2 daughter cells on the grid. We first replace the parent cell with the first daughter, and search for a suitable position to place the second daughter cell. We use a Von Neumann neighbourhood and check if any of the 8 (in 2D grid) neighbouring spots of the parent cell is empty; if one or more are, we locate the second cell in one of those spots at random. Otherwise, with a probability determined by a parameter a, we push all cells along a randomly chosen direction until we hit the grid boundary, and place the second daughter at the nearest emptied spot. With the parameter a we can model boundary driven growth, as it represents the fraction of the radius of the growing tumour where cells are allowed to proliferate; that is, $a = 0.2$ creates a tumour periphery of width equal to 20% of the whole tumour width in which cells are allowed to proliferate even without empty space by pushing neighbouring cells outwards (when $a = 1$, periphery width is 100%, every cell can always push and divide, and the tumour grows exponentially). When a cell divides, we generate passenger mutations by drawing a number from $Poisson(u)$. These mutations will be assigned to both daughter cells.

3. If the next event is cell death, we simply free the position allocated to the cell.

4. At the end of this step, we check if the clock is greater than the time of the next scheduled driver event t_{driver}; if it is, we convert a single wild type (WT) cell into a new mutant and increase its birth rate, or decrease its death rate. This will result in mutant cells having a proliferative advantage. To quantify the effect, we define the fitness s as: $1 + s = (b_{mutant} - d_{mutant})/(b_{wt} - d_{wt})$.

2.4.4 Simulating cancer genomic data generation

At the end of the simulation, we can collect bulk or single-cells and simulate sequencing data generation. Bulk samples are spatially separated tumour chunks 'cut out' from the tumour. We model two different shapes:

1. Squares, which are referred to as "punch biopsies"
2. Long thin rectangles that resemble a "needle biopsy"

A bulk sample is a set of adjacent cells from the final tumour population. Each cell has its unique ID, a position on a grid and its list of somatic mutations. From the sampled cells (in a bulk) joined list of mutations we can construct the Variant Allele Frequency (VAF) distribution as in a real sequencing experiment.

To construct a VAF distribution from a simulated bulk tumour sample, we mimic realistic next generation sequencing steps, specifically sequencing coverage and limits of detectability of low frequency mutations. We proceed as follows:

1. We generate (dispersed) coverage values for the input mutations by sampling a coverage from a Poisson distribution $D \sim Poisson(\lambda = Z)$ with mean λ equal to a desired sequencing depth.
2. Once we have sampled a depth value k for a mutation, we sample its frequency (number of reads with the variant allele) with a Binomial trial. We use $f \sim Binomial(n,k)$ where n is the proportion of cells carrying this mutation in the sample.

This procedure guarantees that the generated read counts reflect the proportions of mutations in the simulated tumour. To model limits of detection of a mutation, after resampling a mutation, we discard it if the corresponding number of reads

containing the variant allele are less than 5 (using the fixed coverage 100, which accounts for $a \sim 0.05$ minimum VAF).

We also performed single cell sequencing taking either random single cells across the whole tumour population, or from spatially structured biopsies (mimicking bulk tissue collection followed by single-cell isolation). We used the obtained single cells to construct maximum parsimony phylogenetic trees. In addition to single cell sequencing, we also model genotyping cells with a given list of mutations, corresponding to targeted sequencing of mutations found using e.g. exome or whole-genome sequencing. To implement this, we take one of the bulk samples as reference genotype and check for presence of each individual mutation in a random set of 200 cells. Similarly, we use the obtained genotyped single cells to infer phylogenetic trees and check how much the genotyped trees differ from the single cell trees.

2.5 Effects of spatial constraints and sampling bias

2.5.1 Spatial effects on bulk sequencing data

For each representative simulation of spatial constraints in Figure 2.4, we modelled the sampling of 6 punch biopsies (small square regions), 2 needle biopsies (long and thin regions), as well as hypothetically sampling the whole tumour. From each sample, we simulated the generation of 100x depth whole-genome data. Figure 2.6 A shows the variant allele frequency (VAF) distributions of samples from the neutral homogeneous growth case in Figure 2.4 A, with clonal mutations (truncal) in grey, subclonal mutations exclusive to the parental background clone in light blue and subclonal mutations within the mutant in pink. All samples show the characteristic $1/f^2$ distribution corresponding to neutral evolutionary dynamics [88], as one would

expect theoretically [93]. The Area Under the Curve (AUC) test for neutrality[89] ($p<0.05$ means neutrality is rejected) is reported on top of each VAF plot and shows that even in the presence of a spatial structure, homogeneous (exponential) neutral growth follows a $1/f^2$ distribution (Figure 2.6 A-i to A-iv). As we have shown previously, it is possible to recover the mutation rate per cell doubling from the $1/f^2$ neutral tail, which in this case without cell death was 10 mutations per division. This was correctly recovered in all samples from Figure 2.6 A (recovered mutation rate reported in each plot as u).

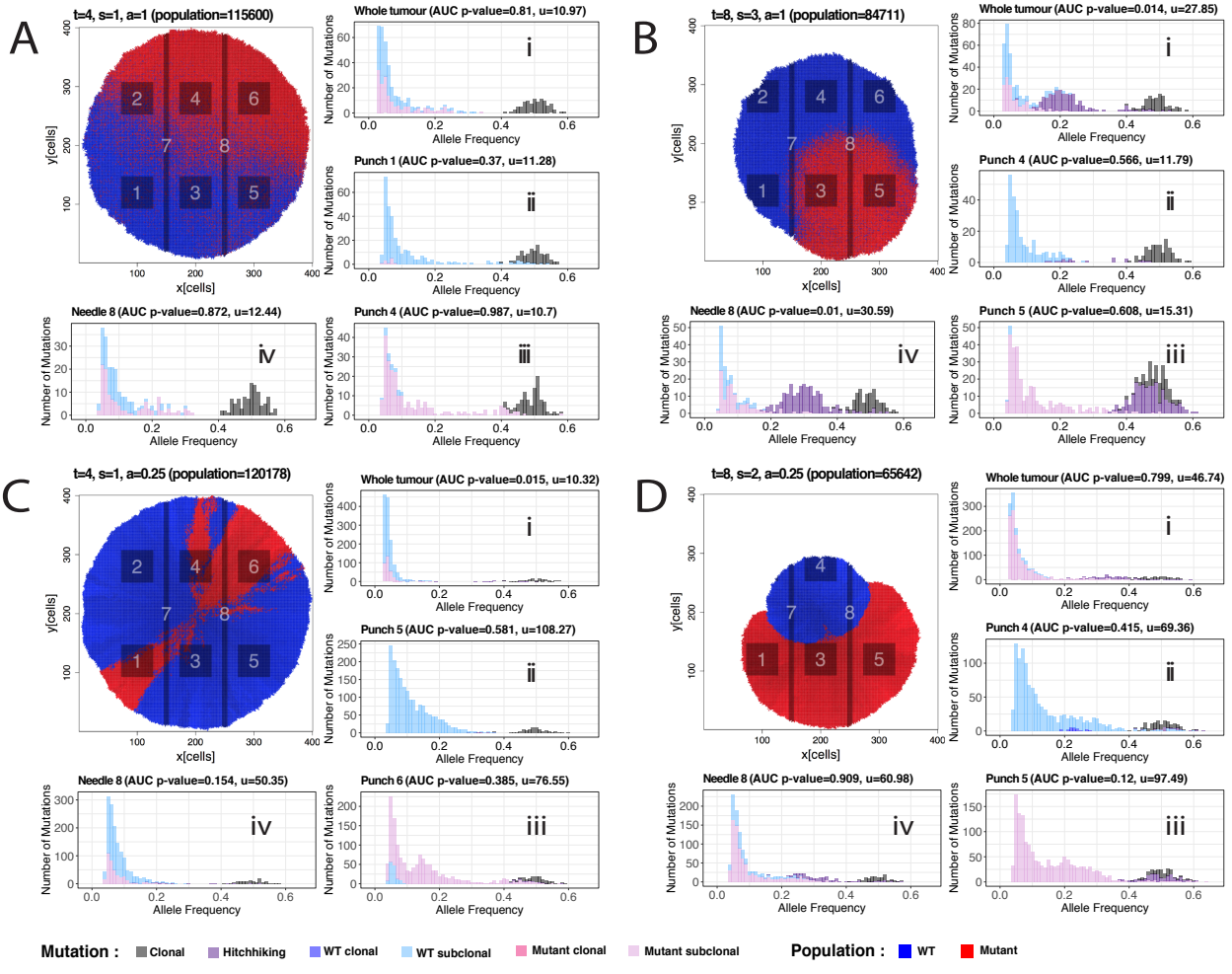


Figure 2.6: Variant allele frequency distributions of punch and needle biopsies from representative scenarios - (A) In the illustrative example of neutral homogeneous growth, a neutral mutant was introduced at generation time $t = 4$ with a selection coefficient of $s = 0$ (neutral) and homogeneous growth ($a = 1$). The mutation rate was $u = 10$. Tumour was simulated until $\sim 100K$ cells. From the final tumour, we sampled 6 punch biopsies (1-6), 2 needle biopsies (7-8) and a "whole-tumour" sample, and simulated $100\times$ whole-genome sequencing data. VAF distributions of each sample are shown (i-iv). (B) In this case, a differentially selected subclone with $s = 3$ was introduced at time $t = 8$ in a homogeneous growth scenario ($a = 1$) and $u = 10$. Final population size was $\sim 80K$ cells. In those samples where both the background and the mutant subclone were present (i and iv), the VAF distribution showed evidence of subclonal selection, with a subclonal cluster (purple) generated by mutations in the selected subclone that hitchhiked to high frequency due to selection. (C) In the case of neutral boundary driven growth, a new (neutral) mutant was introduced at $t = 4$ with $s = 1$ and boundary driven growth parameter $a = 0.025$. Even though the tumour grew neutrally, the spatial effects of boundary driven growth led to deviations from the neutral expected null under homogeneous growth. Moreover, clusters in the VAF spectrum are detectable in iii, where sampling bias produced an overrepresentation of a lineage that was not due to selection. (D) Boundary driven growth with selection (mutant introduced at $t = 8$ with $s = 2$ and $a = 0.025$) produced even more complex patterns of drift and sampling bias. The data represents tumour simulations in 2D space. Birth rate $b = 1$ in all simulations.

In the case of homogeneous growth with subclonal selection (Figure 2.6 B), neutrality could be rejected based in all those samples containing a mix of the background clone and the new subclone (Figure 2.6 B-i and B-iv, see subclonal cluster in purple). Specifically, needle 4 and punch 1 showed the expected signature of selection, with a subclonal cluster a consequence of the over-representation of passenger mutations in the expanded clone [69, 89]. The $1/f^2$-like tail resulting from the within-clone accumulation of passenger mutations remains in the frequency spectrum [89]. Specifically, in the plots in Figure 2.6 B we report the mutations that were present in the first subclone cell in purple. Those are mutations that increased in frequency by hitchhiking on the selected mutant. Importantly, we note that these mutations are not exclusive to the subclone but are also found in other lineages (e.g. in the "cousins" of the selected subclone). The same dynamics are observed if it is the death rate to decrease, rather than the birth rate to increase (Figure 2.7 A,B). Importantly, the cell death d not only increases the rate of genetic drift, as expected, but also the level of clonal intermixing due to the additional stochasticity introduced by high cell replacement (Figure 2.7 C-F, examples of neutral cases). Selection could not be detected in other spatially-distinct samples from the same tumour when they did not contain differentially selected populations, and either captured only the background clone (blue) or only the selected mutant (red) (Figure 2.6 B-ii and B-ii). This is correct as in those samples ITH is neutral.

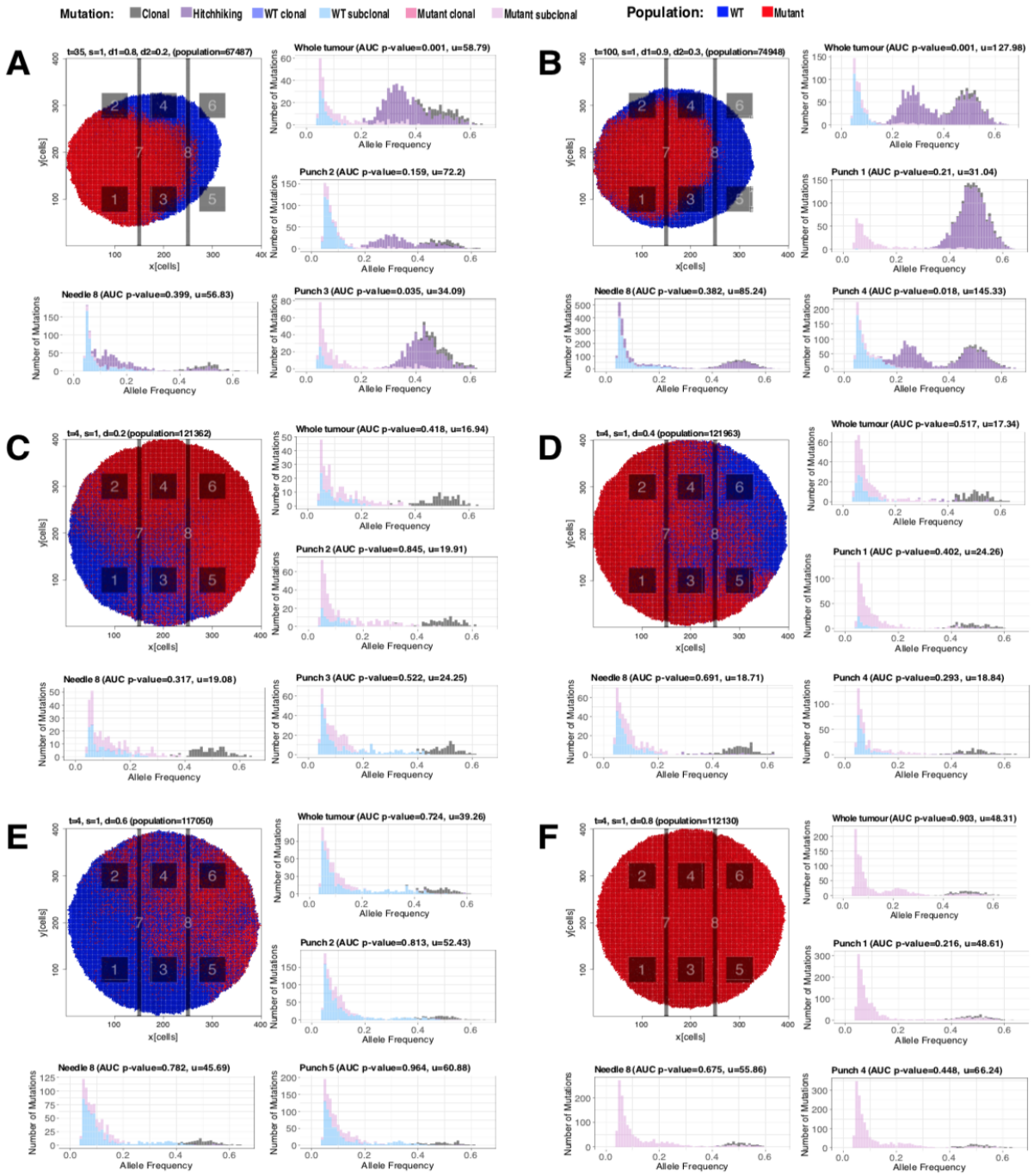

Mutation: Clonal Hitchhiking WT clonal WT subclonal Mutant clonal Mutant subclonal
Population: WT Mutant
A
t=35, s=1, d1=0.8, d2=0.2, (population=67487)
Whole tumour (AUC p-value=0.001, u=58.79)
Punch 2 (AUC p-value=0.159, u=72.2)
Needle 8 (AUC p-value=0.399, u=56.83)
Punch 3 (AUC p-value=0.035, u=34.09)
B
t=100, s=1, d1=0.9, d2=0.3, (population=74948)
Whole tumour (AUC p-value=0.001, u=127.98)
Punch 1 (AUC p-value=0.21, u=31.04)
Needle 8 (AUC p-value=0.382, u=85.24)
Punch 4 (AUC p-value=0.018, u=145.33)
C
t=4, s=1, d=0.2 (population=121362)
Whole tumour (AUC p-value=0.418, u=16.94)
Punch 2 (AUC p-value=0.845, u=19.91)
Needle 8 (AUC p-value=0.317, u=19.08)
Punch 3 (AUC p-value=0.522, u=24.25)
D
t=4, s=1, d=0.4 (population=121963)
Whole tumour (AUC p-value=0.517, u=17.34)
Punch 1 (AUC p-value=0.402, u=24.26)
Needle 8 (AUC p-value=0.691, u=18.71)
Punch 4 (AUC p-value=0.293, u=18.84)
E
t=4, s=1, d=0.6 (population=117050)
Whole tumour (AUC p-value=0.724, u=39.26)
Punch 2 (AUC p-value=0.813, u=52.43)
Needle 8 (AUC p-value=0.782, u=45.69)
Punch 5 (AUC p-value=0.964, u=60.88)
F
t=4, s=1, d=0.8 (population=112130)
Whole tumour (AUC p-value=0.903, u=48.31)
Punch 1 (AUC p-value=0.216, u=48.61)
Needle 8 (AUC p-value=0.675, u=55.86)
Punch 4 (AUC p-value=0.448, u=66.24)
x[cells]
y[cells]
Number of Mutations
Allele Frequency

Figure 2.7: Examples where selection is modelled by varying death rates instead of birth rates, and neutral growth under high cell death. Two examples where fitness advantage is modelled by decreasing cell death the mutant subpopulations and increasing for the wild type. (A) Death rate of mutant subpopulation is 0.2 while for the WT is 0.8. (B) Death rate of mutant subpopulation is 0.3 while for the WT is 0.9. (C-F) Examples of neutral growth with high cell death, which increases the level of genetic drift (especially noticeable in (F)) as well as the level of spatial intermixing due to stochasticity of cell replacement. Birth rate b was 1 in all simulations.

This initial spatial analysis produced similar results to the previous studies of well-mixed non-spatial models [88, 89]. We next investigated the effect of boundary driven growth. Here, only cells close to the borders grow, leaving other cells "imprisoned" inside the tumour mass, a pattern of gene surfing, or sometimes called genetic draft emerges, causing radial patterns of cells growing only at the front of the growing wave (Figure 2.6 C). This has been previously documented both theoretically and experimentally in bacteria [154], in mathematical models of tumour growth [80, 81, 155], as well as in cancer model systems, where the neutral expansion of the cancer cell population under boundary driven growth led to lineages growing just because they were "lucky" to be in the right place at the right time [149]. This has implications for the impact of the immune system during the evolution of a tumour, which exert a negative selection pressure on the cancer cell population through neoantigen recognition and removal [23], especially because neoantigen recognition is clone size dependent [156]. Importantly, boundary driven growth leads to non-exponential population dynamics [147, 148] that also impact the distribution of mutations between the centre and the periphery of a solid neoplasm, as shown in a case of liver cancer sampled at high resolution [157]. The accumulation of subclonal mutations in a neutrally expanding tumour under boundary driven growth is expected to follow a $1/f^2$ scaling form within most of the detectable frequency range ($f > 5\%$), although at low frequency deviations are expected [154]. This is largely driven by the increasing difference in mutational burden between the centre

and the border of the tumour, which could lead to rejection of the standard neutral expectation under exponential growth, as seen when the whole tumour is sampled with respect to when only a localised bulk/needle biopsy is collected (Figure 2.6 C).

Because the population is no longer homogeneously distributed however, this can lead to significant spatial bias, causing over- or under-representation of mutations in the VAF distributions solely due to spatial effects and not because of selection. This causes deviations from the neutral expectation of the mutant allele distributions that risk being wrongly interpreted as the consequence of on-going subclonal selection, as in Figure 2.6 C. In this scenario, we know that subclonal clusters (e.g. punch 6 in Figure 2.6 C-iii) are not differentially selected subclones, but the over-representation of alleles is solely induced by the spatial structure. Furthermore, even when we observe distributions that appear to follow the neutral expectation (AUC $p>0.05$), boundary driven growth results in much higher mutational loads than would be expected in the well mixed case. Here our inferred mutation rates are up to 10 times higher than the ground truth. This can be observed more explicitly in Figure 2.8, where we sample each representative tumour from the centre towards the periphery by taking samples along concentric circles (Figure 2.8 A) and compare the mutational loads of the samples (Figure 2.8 B). This was indeed observed in a case of neutrally growing liver cancer [157] and a similar phenomenon is also observed in species evolution [158].

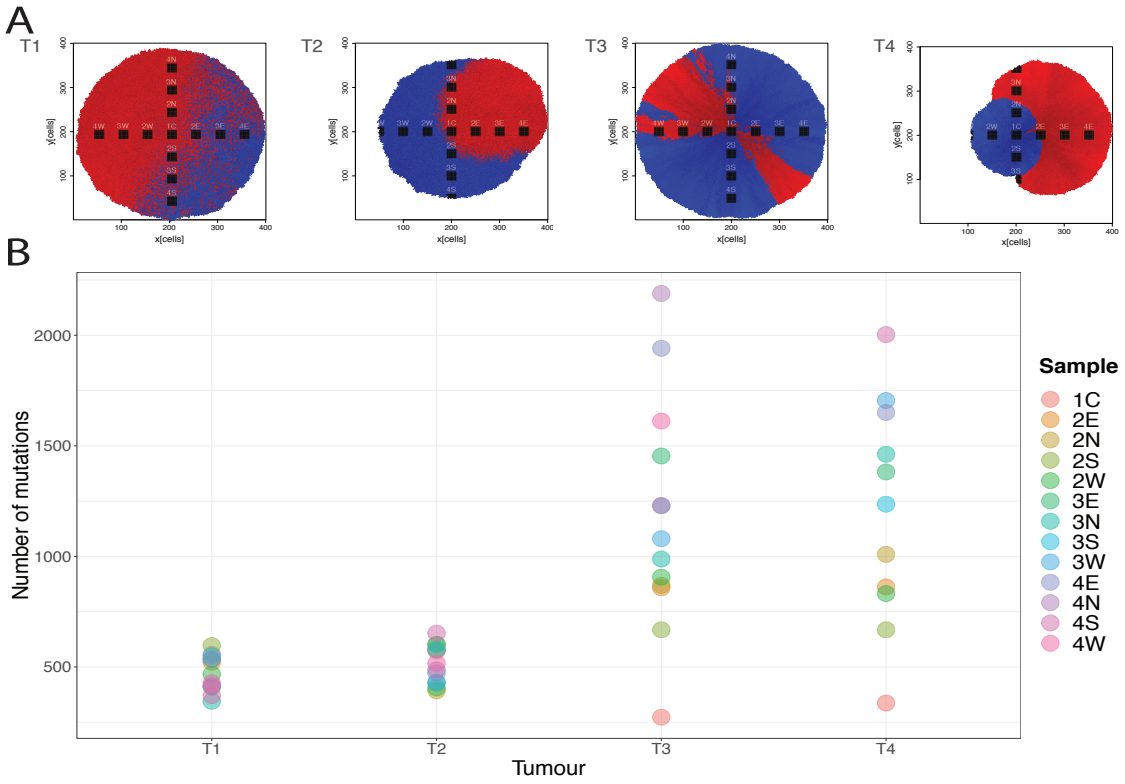


Figure 2.8: Mutational load comparison for different growth cases. (A) We sample each representative example tumours (T1 – neutral homogenous, T2 – selective homogenous, T3 – neutral boundary driven, T4 – selective boundary driven) from the tumour centre (bulk sample C1) towards the periphery following the concentric circles in four directions: W – west, E – east, N – north, S – south. The bulk indexes (2W, 3W, 4W) are proportional to the distance from the centre to the periphery. (B) We observe how number of mutations per bulk sample increases proportionally to the distance from the tumour centre in the case of boundary driven growth. Also, the total number of mutations are much higher for the constrained boundary driven growth than for the homogenous tumour due to increased cell turnover in the former case.

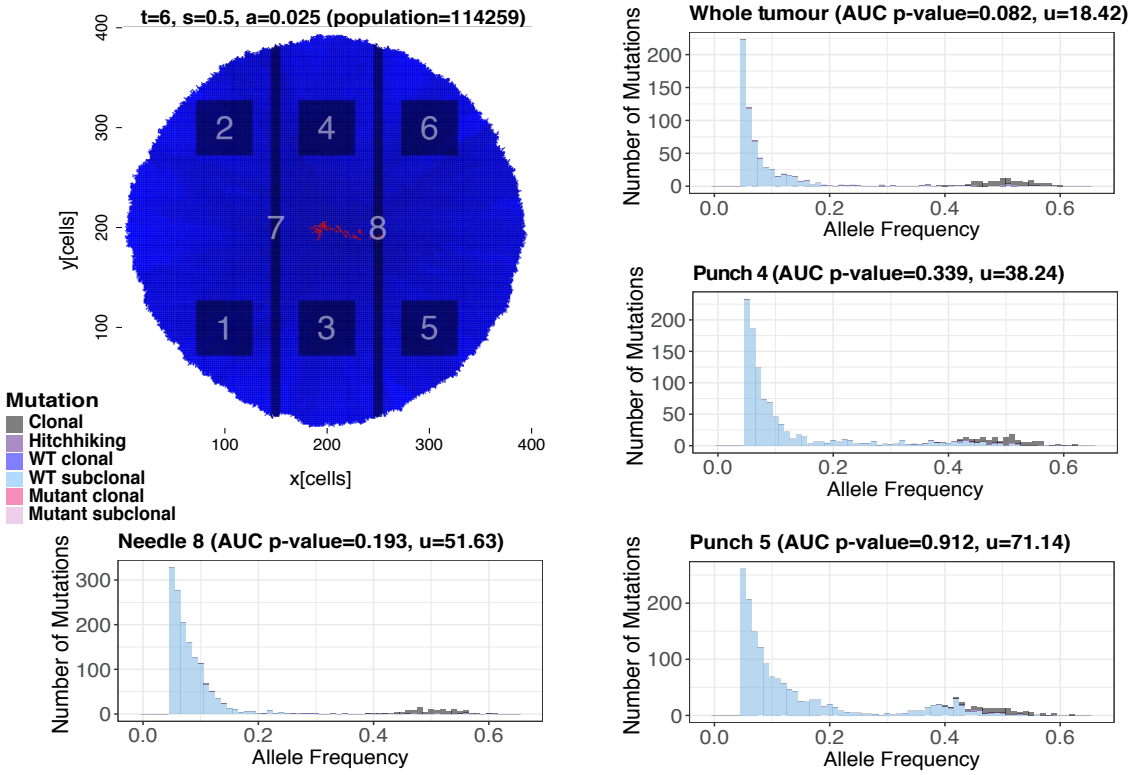


Figure 2.9: Example of imprisonment. Example of selective boundary driven growth when the driver mutant subpopulation gets trapped within the wild type population despite being fitter than the WT clone.

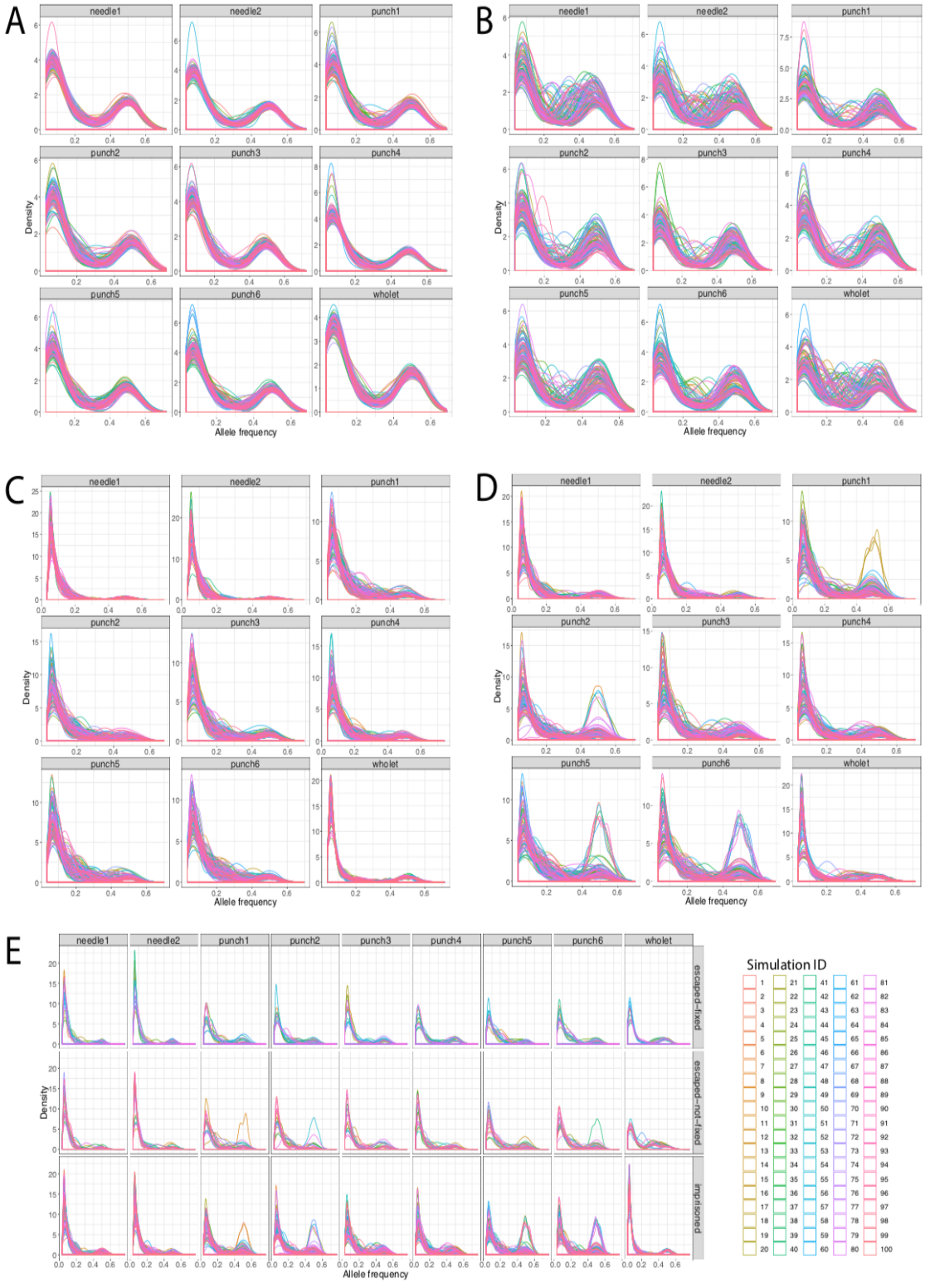

A
B
C
D
E
needle1
needle2
punch1
punch2
punch3
punch4
punch5
punch6
wholet
Density
Allele frequency
escaped−fixed
escaped−not−fixed
imprisoned
Simulation ID

Figure 2.10: The effect of stochasticity and sampling bias on the shapes of VAF distributions for the four representative scenarios. For each of the representative cases: (A) neutral homogeneous, (B) selective homogeneous, (C) neutral boundary driven, (D) selective boundary driven, we simulated 100 different runs of each case keeping the underlying parameters constant and varying only the random seed of the simulation. For each simulated tumour, we constructed needle and punch biopsy sample VAF distributions along with the whole tumour VAFs. Overall there is less variation among the distributions for neutral (A,C) versus selective (B,D) cases. In addition, punch biopsy VAFs scatter more than needle biopsy samples in comparison to the whole tumour VAF distributions. (E) We separated the VAF distributions for the selective boundary driven between cases where the new clone escaped and grew to fixation, versus escaped by not yet fixed (signature of ongoing subclonal selection), versus imprisoned (leading to neutral dynamics)

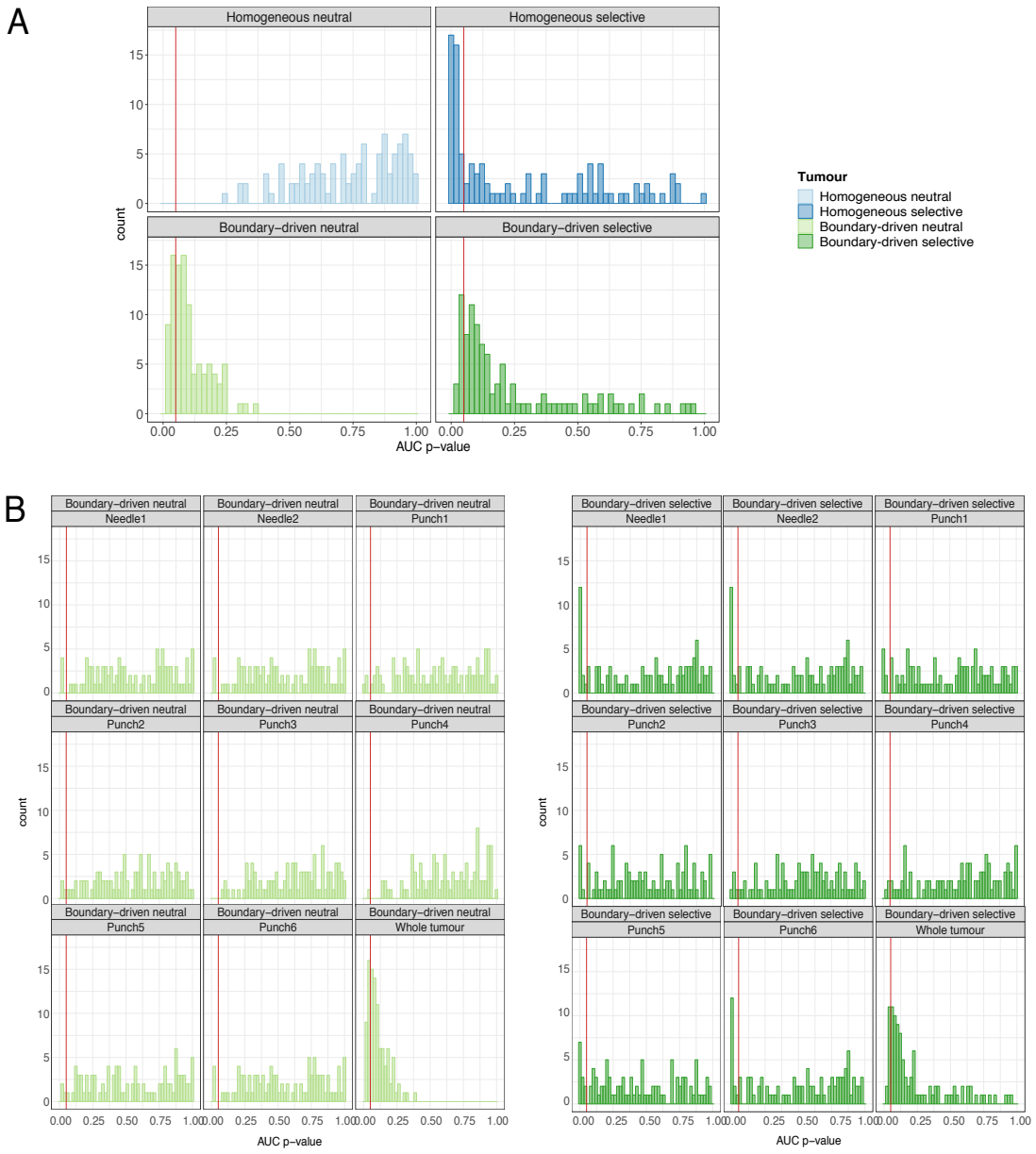


Figure 2.11: Distribution of AUC based neutrality test p-values. (A) We simulate 100 different tumours for each 4 representative growth models and fit $1/f$ test to their corresponding whole tumour sample VAFs. Reported are the distributions of p-values obtained from each test using the AUC statistics. (B) For the cases of boundary-driven growth modes we compared tests of neutrality using the whole-tumour sample versus punch/needle biopsies.

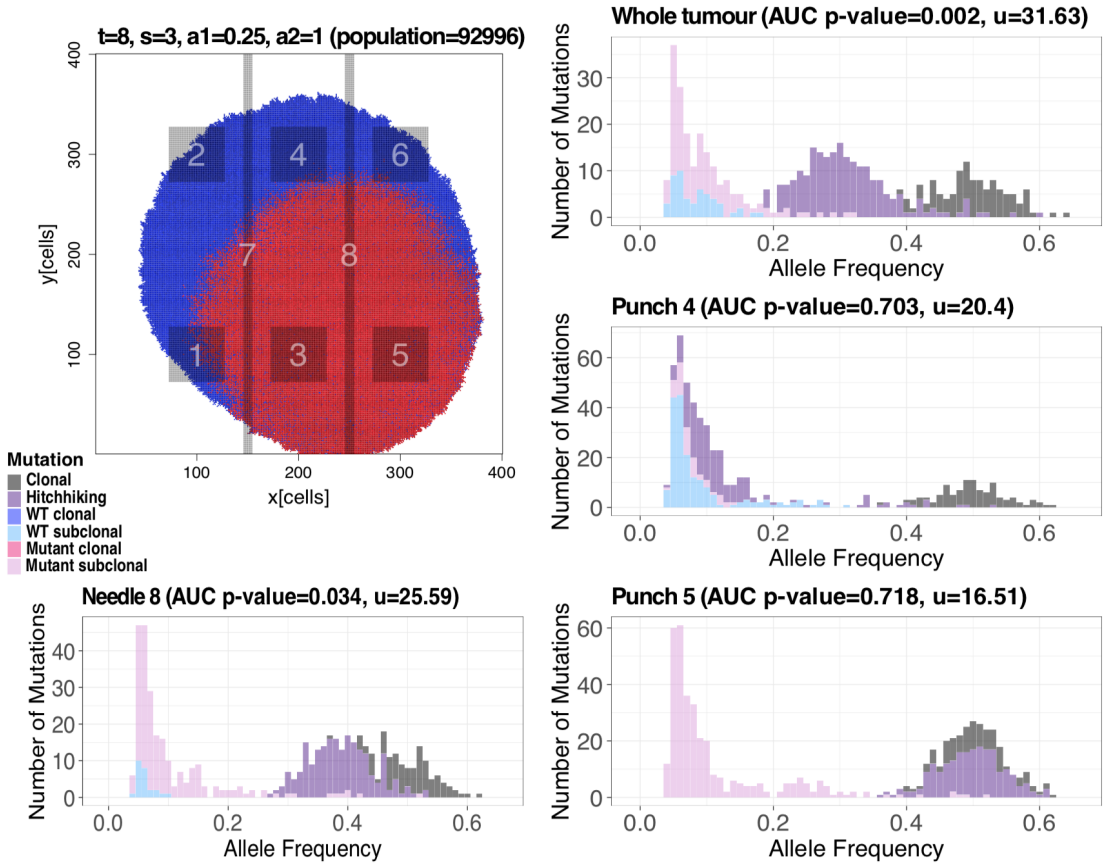


Figure 2.12: Example of selection when mutant subpopulation has higher push power instead than higher birth rate. Example of a selective exponential growth when the mutant subpopulation has higher "push power" than the wild type population.

If we combine boundary driven growth and subclonal selection the situation is further complicated: selective effects are now modulated by spatial constraints. In some cases, the selected mutant emerges and remains directly at the front of tumour growth. In this scenario the outgrowth caused by its selective advantage is amplified further just because it occurred at the growing front (Figure 2.6 D). In other cases, the selected mutant may, by chance, remain "imprisoned" within the tumour (assuming the mechanism of selective advantage is unable to overcome this spatial entrapment) and stops proliferating despite its selective advantage (e.g. Figure 2.9

4). In both these cases, further sampling biases occur. In the case of punch 5 for example (Figure 2.6 D-iii), where the new subclone is fixed (clone fraction=100%), there is an overrepresentation of a cluster of mutations that is only due to spatial drift and not selection. These dynamics are recapitulated in larger cohorts of simulated tumours with the same parameters (Figure 2.10). The distributions of p-values for the AUC measurements for all simulations for different modes of growth are illustrated in Figure 2.11 A. This figure shows that neutrality is accepted in the majority of homogeneous cases without selection, and it is rejected in the majority of homogeneous cases with selection. In the case of boundary driven growth things are more complicated. In Figure 2.11 B we show the AUC tests for neutrality applied to whole-tumour samples versus punch/needle biopsies. In the case of neutral boundary driven growth, neutrality is accepted in the majority of cases when we use localised punch/needle biopsies, but rejected when the whole-tumour sample is examined. This is due to the deviation from strict neutrality caused by boundary driven growth, that can be detected only when a large region of the tumour is sampled (and hence differences between centre and periphery of the tumour are captured). In the case of selective boundary driven growth, we observe similar dynamics but with the ability of rejecting neutrality if differential selection of the growing subclone is captured within the punch/needle sample. We note that under selective boundary driven growth, the subclone often remains imprisoned, leading to neutral-like dynamics. Similar dynamics to Figure 2.6 B are observed when positive selection is modelled as the probability of growing in the absence of space (pushing probability parameter a increased) rather than the increased birth rate. This leads to dynamics dominated by the homogeneous growth of the subclone rather than boundary growth of the background clone (Figure 2.12). Moreover, removal of the majority of cells (99%) by treatment leads to enhancement of outgrowth of selected clones due to competitive release (Figure 2.13 and Figure 2.14).

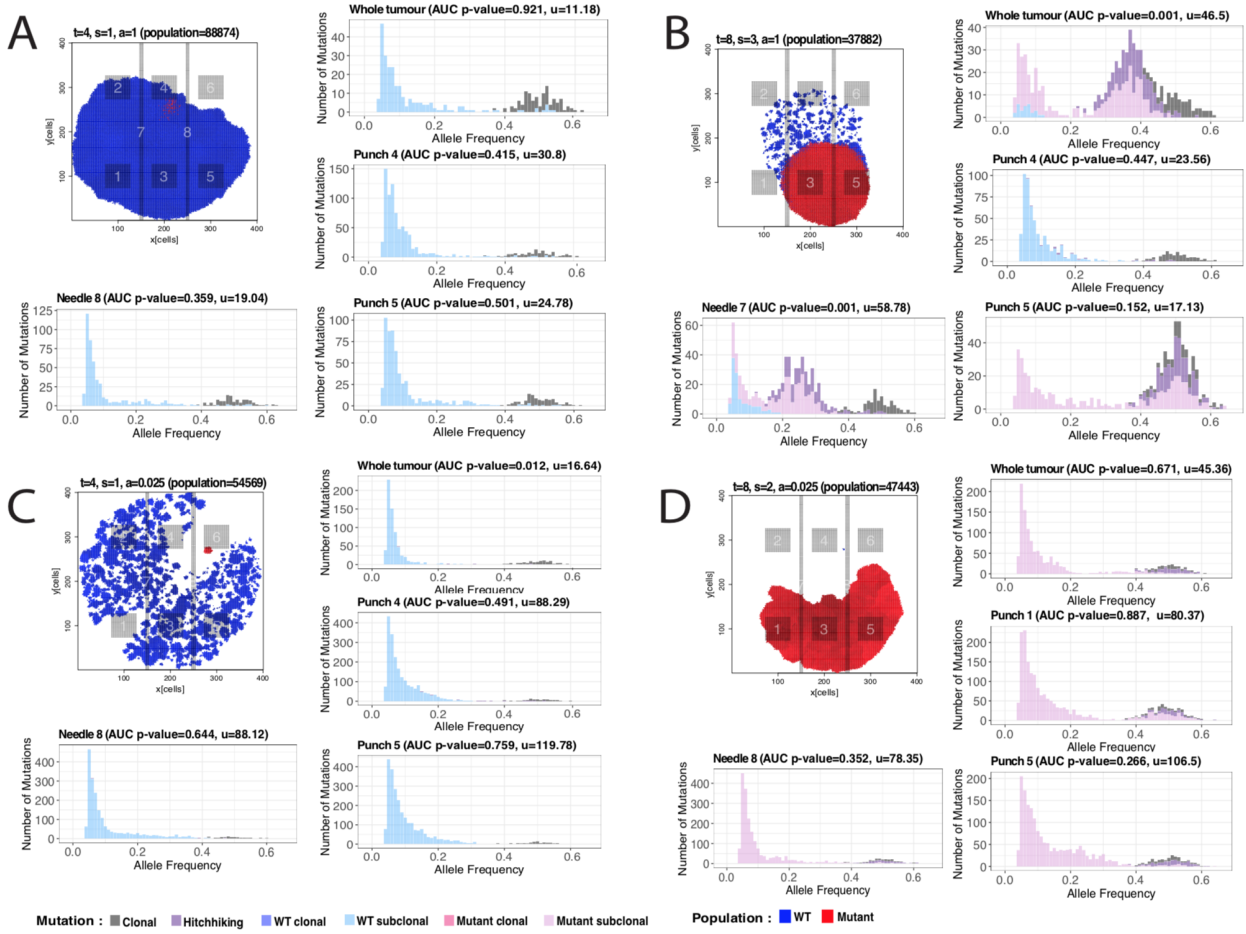


Figure 2.13: Killing 99% of cell population and re-growing tumours. For each of the representative cases: (A) neutral homogeneous, (B) selective homogeneous, (C) neutral boundary driven, (D) selective boundary driven, we simulated procedures of removing large cell population (here 99%) by the end of tumour growth and wait it to regrow to its original size.

We then looked at the pairwise VAF distributions between samples. The amount of subclonal mutations scattered through the frequency spectrum (Figure 2.15) and the number of subclonal clusters due to sampling bias and spatial drift was significant (e.g. Figure 2.15 D). As per ground truth, only the dark purple mutations should show a subclonal clustering pattern (e.g. Figure 2.15 B, punch 1). We found that scattered variants were mostly due to the effect of neutral lineages spreading in

space, and then subsampled in different ways in each tumour region. In the case of boundary driven growth, sampling bias produces evident clusters that do not correspond to differently selected subclones in the tumour. This makes the reconstruction of the true clonal phylogeny and its evolutionary interpretation problematic.

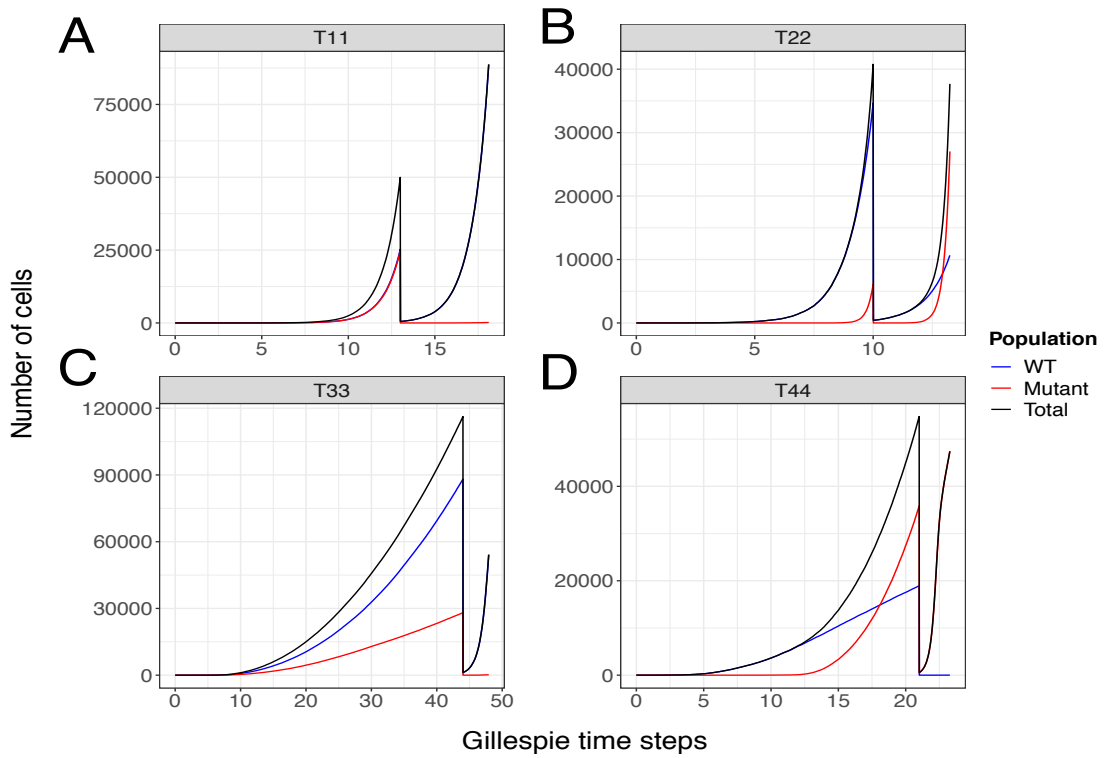


Figure 2.14: Growth curves through cell killing. Tumour cell population growth curves for each of the representative cases: (A) neutral homogeneous, (B) selective homogeneous, (C) neutral boundary driven, (D) selective boundary driven, where by the end of tumour growth we remove 99% of cell population and wait for the tumour to regrow to its original size.

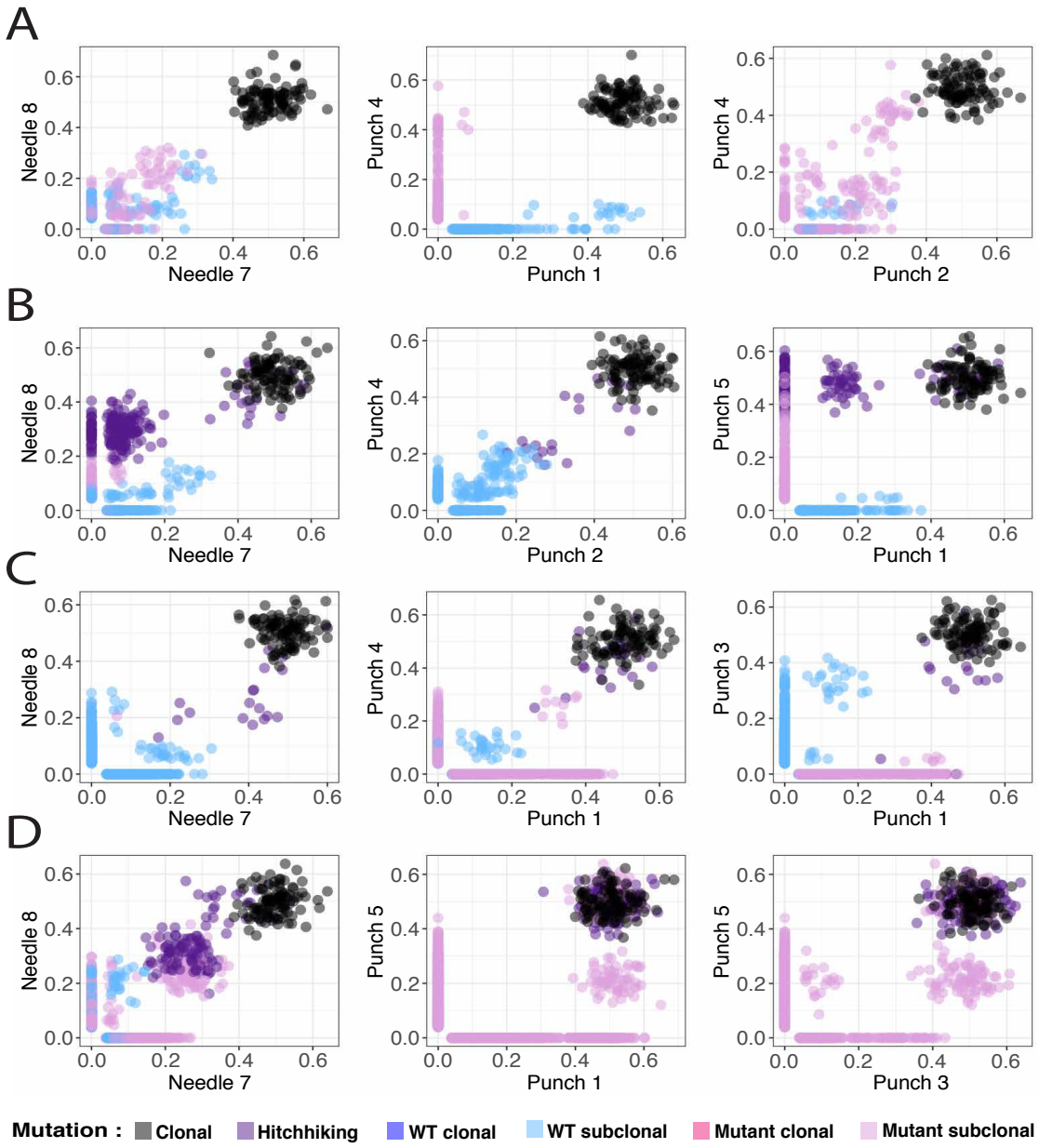


Figure 2.15: Sample vs sample scatterplots of mutations. For each of the representative cases: (A) neutral homogeneous, (B) selective homogeneous, (C) neutral boundary driven, (D) selective boundary driven, we report the scatterplots of somatic mutations in selected samples. Clearly, the presence of passenger subclonal mutations in the neutral tail of growing clones that spread in space as the tumour grows produces scattered variants (e.g. A). Even more striking is the formation of subclonal clusters of mutations particularly in the presence of boundary driven growth (e.g. C, D) where some lineages are over-represented not because of differential selection, but because of sampling bias and spatial drift.

2.5.2 Spatial effects on single-cell sequencing data

Most of the confounding factors we have described so far result from the limitations of bulk sequencing, where the genomes of many cells are convolved within samples. Single-cell sequencing does not suffer from this particular limitation and promises high-resolution cancer evolutionary analysis devoid of the drawbacks of bulk sequencing [121].

To examine the effect of single cell sequencing, we simulated whole-genome sequencing of 10 single cells taken at random from the tumour and reconstructed their phylogenetic relationship (Figure 2.16 A-i). For the neutral cases (Figure 2.16 A and C), the patterns are consistent with a typical "balanced" neutral tree, wherein all lineages contribute roughly equally to the final cell populations. In a balanced tree, the average distance between the trunk and each leaf of the tree is similar in each lineage. In cases with selection (Figure 2.16 B-i and D-i), the selected subclonal lineages are over-represented on the tree (as reflected in VAF distributions), as the red lineage is introduced at time $t = 8$ and would have been much smaller if it was not selected for. Here the average distance between trunk and any leaf is different in the background vs the new clone. The pattern is even clearer if we sample 400 single cells and performed WGS (Figure 2.16 B-ii and D-ii). We note that if we use randomly sampled single cell sequencing and plot the site frequency spectrum (frequency distribution of mutations within the population of sampled cells) we recapitulate the VAF distribution, including subclonal clusters and $1/f^2$ tails (Figure 2.17). This is because the site-frequency spectra derived from single-cell sequencing data corresponds to a VAF distribution.

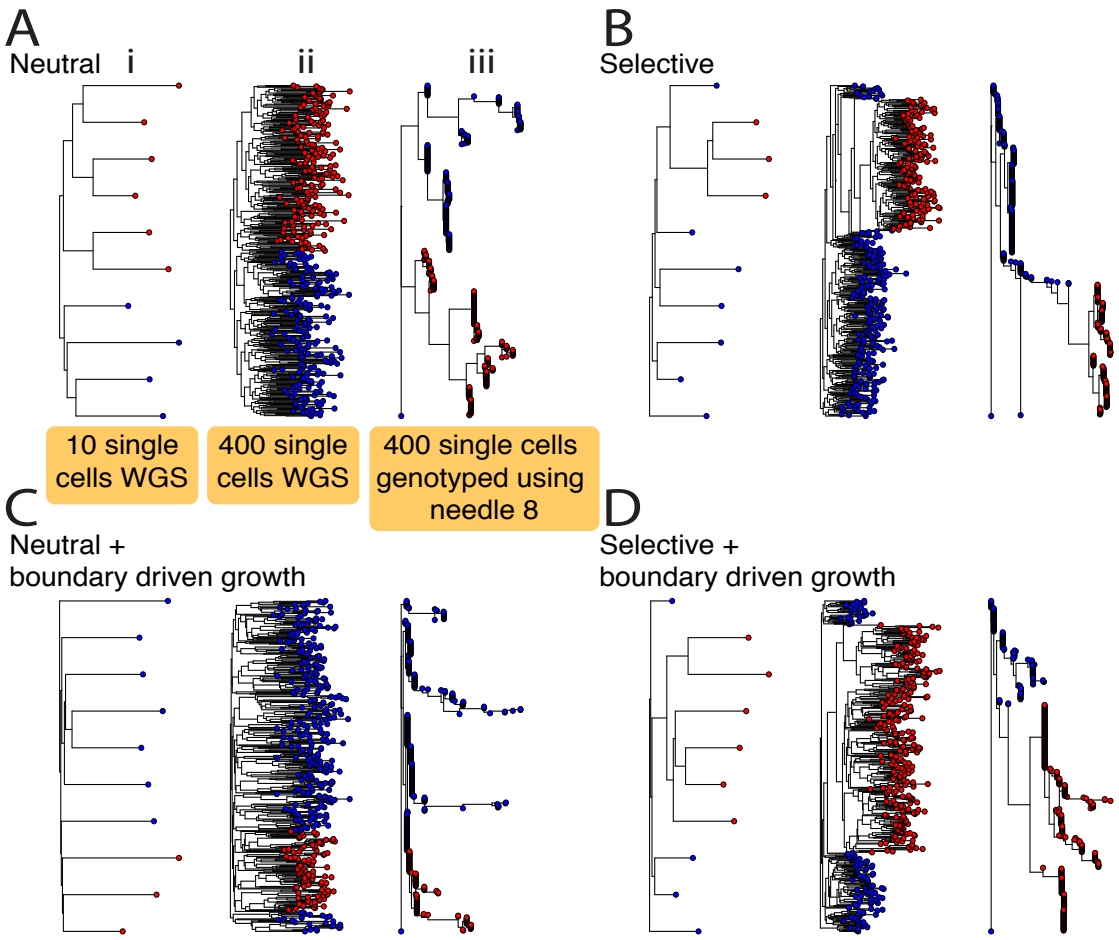


Figure 2.16: Single-cell sequencing data from spatial tumour simulations. (A) From each representative scenario we sampled 10 single-cells at random (i) as well as 400 single-cells at random (ii) and performed synthetic whole-genome sequencing. In both homogeneous (A) and boundary driven growth (C), single-cell sequencing significantly reduces the sampling bias that we found in bulk samples and the only overrepresented lineages were due to selection (B, D). However, due to the currently high error rate of single-cell sequencing, several studies rely on single-cell genotyping using mutations found in bulks. We simulated this by genotyping on 400 single-cells the mutations found at VAF>5% in needle biopsy 8 of each tumour (iii). The resulting trees are hard to interpret in terms of the clonal phylogeny due to the bias in the selection of variants to be genotyped.

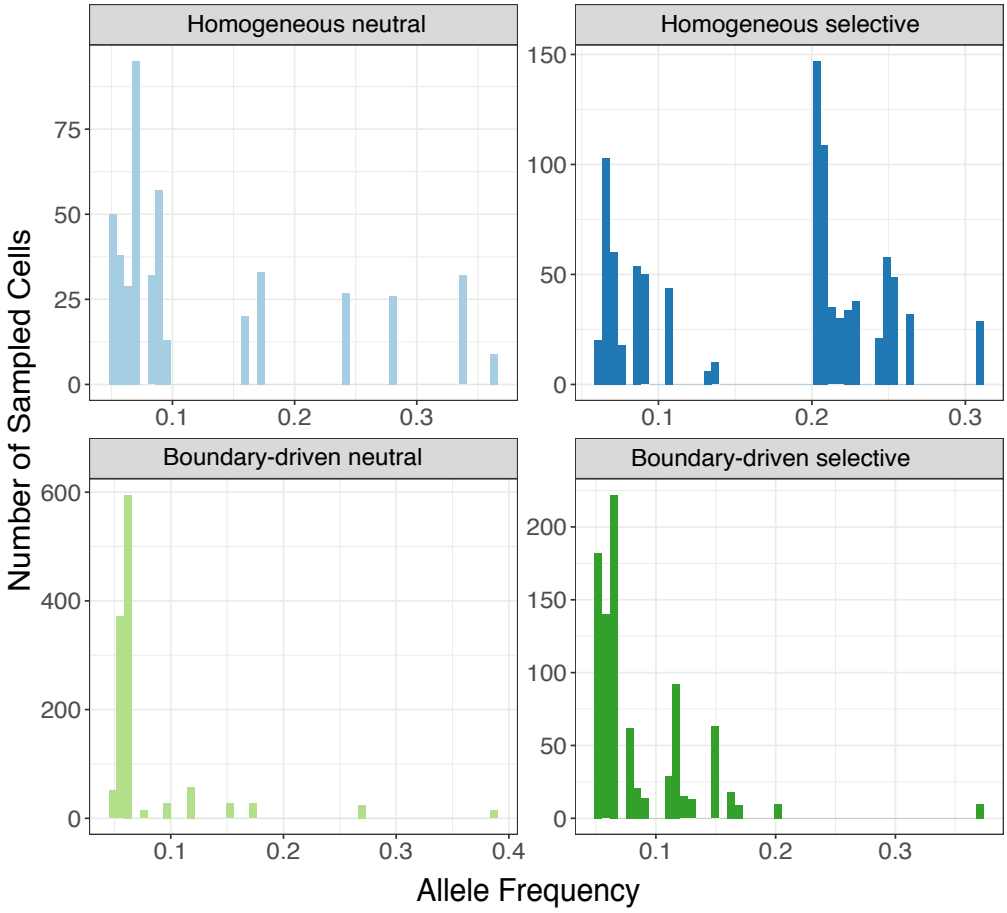


Figure 2.17: Allele frequency distributions derived from single cell sequencing. We construct the allele frequency distributions from sequencing the randomly sampled 400 single cells (same as in Figure 4) from the four representative tumour examples: T1 – neutral homogenous, T2 – selective homogenous, T3 – neutral boundary driven, T4 – selective boundary driven.

However, as whole-genome mutational profiling of single cells is still difficult due to allele dropout [111], often single-cell genotyping has to be performed instead [159]. In this approach, a bulk sample is sequenced and all mutations in that bulk sample are then tested in single cells for presence/absence of the mutation. Integrating bulk sequencing with single-cell information is extremely powerful [160], but requires careful interpretation of the results. In Figure 2.16 we show that this

approach, although informative, can lead to very distorted phylogenetic trees where branch lengths are heavily biased by the initial choice of mutations to be assayed, and consequently the signature of selection vs neutrality is not readily identifiable from these data alone.

Moreover, significant sampling bias is still apparent for single-cell sequencing when individual cells are not sampled uniformly at random from the whole tumour, but instead isolated in 'clumps' from different bulk samples. In Figure 2.18 we have simulated the collection of 4 single cells from each of the 6 punch biopsies in Figure 2.6 (these are the same simulations used to generate Figure 2.16). The trees are quite different from those sampled in Figure 2.16 and moreover, it is interesting to see how the underlying patterns of growth are reflected in the mixing of cells from different bulks. For instance, homogeneous growth leads to very high intermixing of cells in different bulks, whereas boundary driven growth tends to spatially segregate bulks. We have quantified the level of intermixing for different modes of growth in all our simulation cohort, highlighting this pattern (Figure 2.19). We have observed these patterns real data from carcinomas vs adenomas, where carcinomas were characterised by clonal intermixing, but adenomas were not [95]. Similar patterns of intermixing have also been found more recently using single-cell seeded organoid sequencing [161].

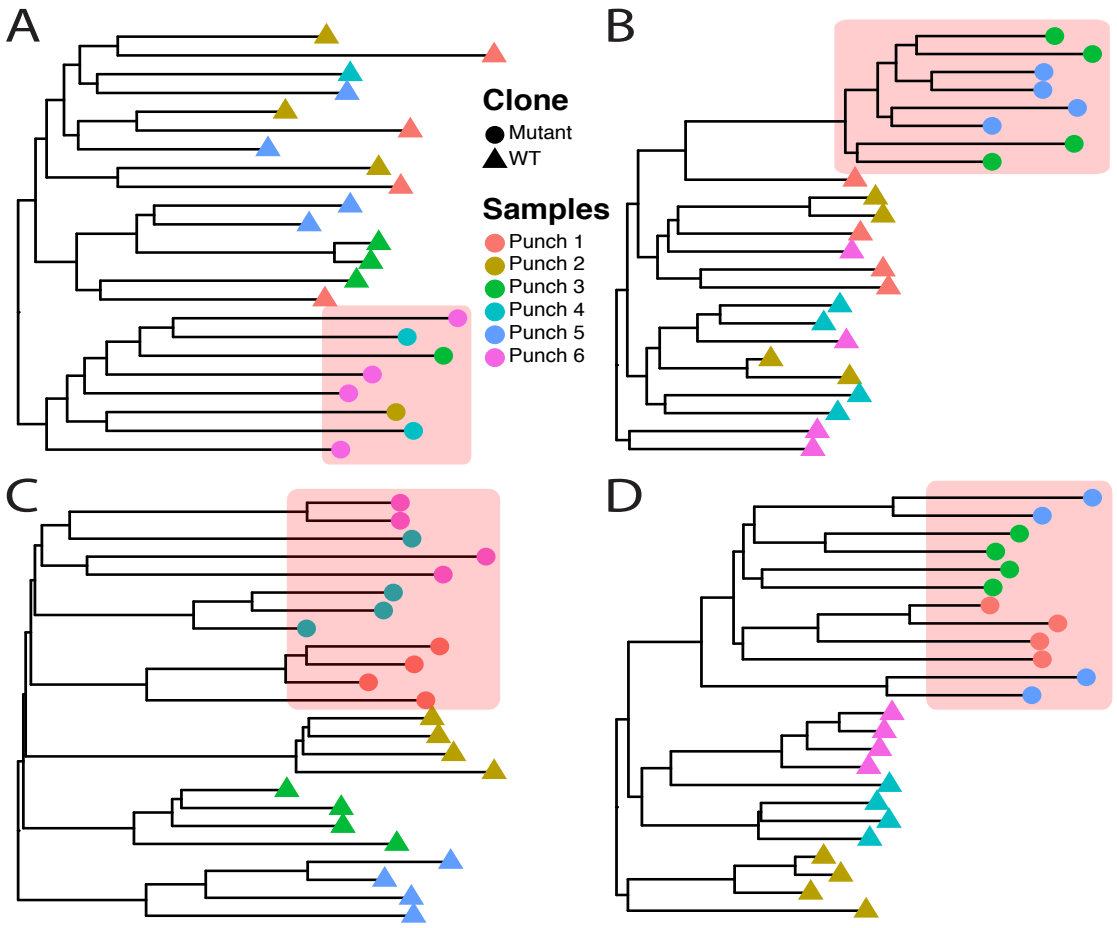


Figure 2.18: Biases of single-cell sequencing when cells are taken from spatially separated bulk samples. Whereas taking random cells from a tumour highly reduces sampling bias, this is often not how single cells from neoplasms are sampled. Often first small chunks of the tumour are dissected and then single-cells are isolated from those. (A) neutral homogeneous, (B) selective homogeneous, (C) neutral boundary driven, (D) selective boundary driven. For each of our representative examples, we simulated this type of sampling and show how this impacts severely on the phylogenetic tree and patterns of clonal intermixing. In particular, single-cell sampling from bulks alters the detected phylogenetic relationship of the cells because, since groups of cells come from spatially segregated regions, those appear more closely related than expected by chance. This is an important source of sampling bias that needs to be considered when analysing single-cell phylogenies. Cells coming from the 'red' mutant subclone are highlighted in the red shaded box.

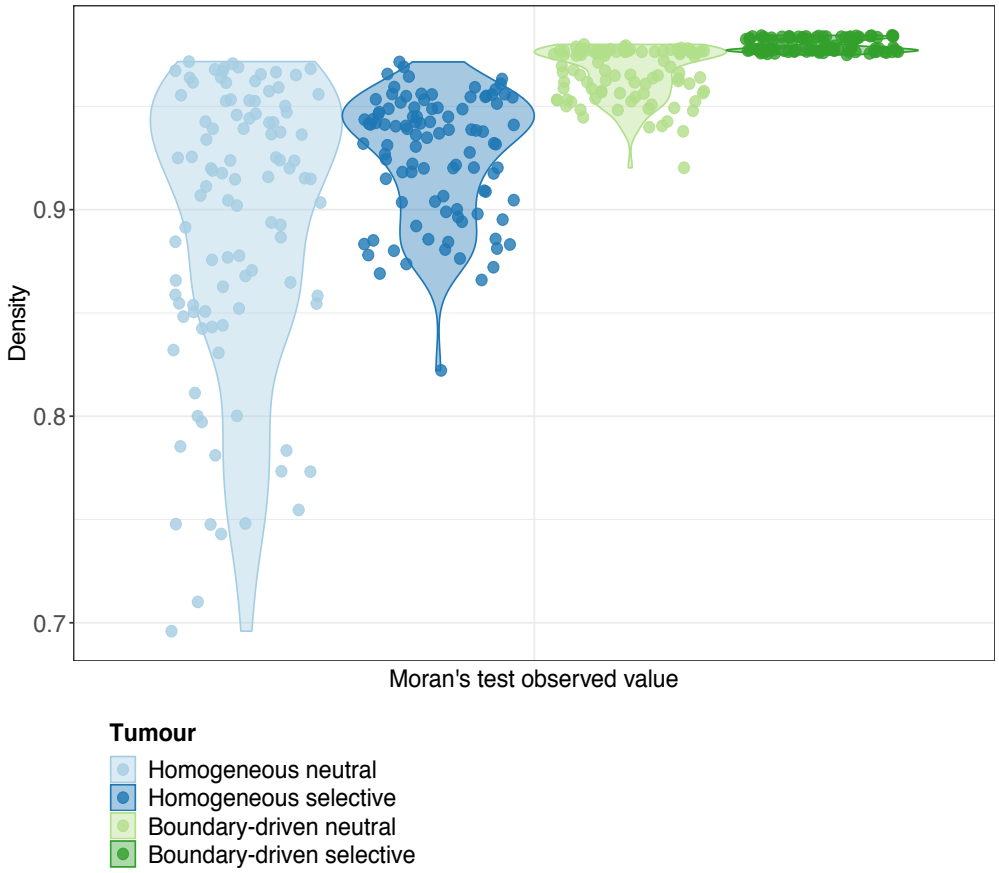


Figure 2.19: Distribution of Moran's test effect size. We simulate 100 different tumours for each 4 representative growth models and test intermixing of subpopulations within each simulation lattice using Moran's entropy-based test. Each individual test output significant p-values indicating to high spatial correlation between tumour cell types (mutant vs WT) and their location on tumour lattice. Although the test effect size (the observed values of the Moran's test statistic) differ as we can see from their distributions per model scenario.

2.6 Resolving spatial effects with Bayesian inference

2.6.1 Approximate Bayesian Computation

Due to the complexity captured by our spatial model of tumour growth, we do not have explicit formulas for the stationary probabilities of the stochastic process, and hence cannot derive a likelihood function. Thus, we have to use likelihood-free methods to perform statistical inference on the parameters and compute posterior distribution of the parameters $\boldsymbol{\theta}$.

Here we use Approximate Bayesian Computation (ABC) [162, 163] to infer the parameters of our model. ABC is based on the idea of scanning a large grid of plausible values for $\boldsymbol{\theta}$, and simulating the model many times with such parameters. Outputs of the model are stored and compared using a predefined set of summary statistics that are initially evaluated on real data. We can then rank sets of parameters that lead to the generation of synthetic data that are close to the observed data. We can estimate a posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{D})$ for the model parameters $\boldsymbol{\theta}$, using the available data $\boldsymbol{D}$ and the prior for $\boldsymbol{\theta}$. This method is computationally intensive, and requires running several hundred (ideally thousands or millions) simulations. In our case we have generated $\sim$74 million simulations that we use to perform the inference step.

There are different approaches to implement ABC, the simplest is rejection-sampling. More advanced implementations such as ABC with Markov Chain Monte Carlo (MCMC) can result in significant increases in efficiency. We implemented a simple rejection-sampling algorithm first, and then added Monte Carlo simulation techniques to speed up convergence. The simple ABC rejection-sampling algorithm consists of the following steps:

1. Sample parameter vector θ from a prior distribution $p(\theta)$.
2. Run the model with the given parameter set and generate the synthetic dataset
3. Evaluate the distance between the simulated dataset and the target data
4. If the distance is less than a desired threshold, accept the parameters.
5. Return to step 1 and repeat until N parameter values are accepted.

In this study we use uniform priors for all parameters: $u \sim Uniform(0,100)$, s,d,$a \sim Uniform(0,1)$, $t \sim Uniform(0,15)$. One of the most important factors that affect the ABC outcome is the number of simulations that one can afford to run, and the summary statistics chosen to evaluate the distance between a target and a simulated dataset. Summary statistics can be any quantitative measurement that captures the information from the multidimensional data without losing too much information. As for our distance metric, we use Euclidean and Wasserstein distances between summary statistics for different parameters as discussed below.

The Wasserstein metric estimates distance between probability distributions by treating each distribution as a unit amount of dirt piled up on a given metric space and calculates the minimum cost required to convert one pile into another. If x and y are two vectors we want to evaluate the distance of, first we calculate their empirical distribution functions $F(t) = \sum_{i=1}^{m} w_i^x \mathbb{I}\{x_i \leqslant t\}$ and $G(t) = \sum_{i=1}^{m} w_i^y \mathbb{I}\{y_i \leqslant t\}$ (for weights w_i^x and w_i^y we took $1/m$ and $1/n$ respectively), the Wasserstein distance is defined by evaluating the following:

$$W_p(F,G) = \left(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^p\right)^{1/p} \tag{2.15}$$

where we took $p = 1$ for our analysis. We used the R package transport [164] to implement the distance calculation.

We used different summary statistics for each sampling scheme. For punch biopsy, needle biopsy and whole tumour sampling – we used the VAF distribution to compute our summary statistics. For the whole tumour VAFs, our ABC procedure was similar to the one in [89]. For the bulk samples, since our model implements multi-region sampling, we first evaluate the multivariate VAF distribution (which is a joint probability distribution of all sampled bulk VAFs) and then calculated the Euclidean distance between the obtained empirical probability distribution vectors:

$$D_{Euclidean}(F_{sim.data}(VAF_{Bulk_1},...,VAF_{Bulk_N}),F_{target.data}(VAF_{Bulk_1},...,VAF_{Bulk_N})) \quad (2.16)$$

With single cell samples, we constructed phylogenetic trees per tumour and used different tree-based summary statistics to evaluate the distance. Since the inferred phylogenetic tree branch length is proportional to the number of unique mutations belonging to a node, we decided to compare the vectors of all branch lengths (between a simulated and target tumour trees) by computing the Wasserstein distance. For the subclone introduction time tdriver, death rate d and the boundary driven growth parameter a, we chose to compare the vectors of branching times for each node of the phylogenetic trees.

Due to computational costs, we are limited to run the ABC framework with a small tumour size (∼100k cells) or simulate smaller datasets per inference, both of which can significantly affect the outcome. To therefore speed up our ABC framework we implemented a Sequential Monte Carlo (SMC) algorithm [165] to increase the acceptance rate of the simple ABC rejection algorithm. Our ABC-SMC algorithm uses sequential importance sampling by running several rounds of resampling around the accepted parameters (correlating the rounds), and gradually decreasing the acceptance threshold while converging to the posterior distribution. This approach significantly increases the acceptance rate of the simulated datasets

[166].

Our implementation of the ABC-SMC algorithm is as follows:

1. Initialise the indicator to rounds r and the acceptance threshold ε
2. **If** r=1
   - 2.1. Run the simple ABC rejection algorithm (described above).
   - 2.2. Order the simulated parameters set according to their corresponding distance values.
   - 2.3. Keep the top Q per cent of the parameters.
3. **Else**
   - 3.1. Sample next particle $\theta = (u,t,s,d,a)$ from the accepted set of parameters from round $r-1$ with weights W_{r-1}.
   - 3.2. Perturb each sampled parameter p_i using uniform perturbation kernel $K = Unif(p_i - \sigma, p_i + \sigma)$, where $\sigma = \frac{1}{2}(max(p_i^{r-1}) - min(p_i^{r+1}))$.
   - 3.3. **If** $\pi(\theta) > 0$, keep θ; **Else** go to step 3.2.
   - 3.4. Simulate data from the model using the sampled particle θ.
   - 3.5. Calculate distance D between the target and the simulated data.
   - 3.6. **If** $D < \varepsilon$, keep θ; **Else** go to step 3.1.
4. Calculate the weights for all accepted particles $1 \leqslant j \leqslant N$:
   - 4.1. **If** $r = 1$, set $W_{(j,r)} = 1$
   - 4.2. **Else** $W_{(j,r)} = \frac{\pi(\theta_{(j,r)})}{\sum_{l=1}^{N} W_{(l,r-1)} K(\theta_{(l,r)} | \theta_{(l,r-1)})}$
5. Update the threshold ε to the top Q-th percentile of the accepted particles.

6. Repeat until ε is less than a desired convergence threshold.

Our ABC-SMC framework tries to recover all the parameters (referred as a particle in the algorithm above) at the same time. We notice that once one of the parameters converges, the acceptance rate decreases significantly. We then decided to fix the converged parameter at the inferred value (mode of its posterior) and re-run the inference varying the rest of the parameters until other parameters converge, and repeat the procedure. We found that this significantly improved the convergence speed. For the 2D inference in Figure 2.20 we started with $N = 100$ simulated particles, performed $r = 10$ rounds with quantile $Q = 0.5$, leading to ∼200k simulations for each parameter and ∼1M simulations in total. For the 3D inference in Figure 2.23, we started with $N = 1000$ simulated particles, performed $r = 10$ rounds with quantile $Q = 0.5$, leading to ∼2M simulations for each parameter and ∼10M simulations in total.

We developed an R package called CHESS (Cancer Heterogeneity with Spatial Simulations) that implements the following three sampling strategies for the inference:

1. Bulk samples (punch or needle biopsies) - ABCSMCwithBulkSamples()
2. Single cell sample phylogenetic trees - ABCSMCwithTreeSampleBL() and ABCSMCwithTreeSampleBT() (using Branch Lengths or Branching Times as summary statistics)
3. Whole tumour bulk sample - ABCSMCwithWholeTumour()

Depending on the strategy, a user would need to provide real or synthetic target data in the form of tumour bulk sample VAFs (list of R data.frames where each row should correspond to a unique mutation with the following columns: clone

(Clone type label set to 0), alt (Number of reads with the variant), depth (Sequencing depth), id (Unique mutation ID)), an array of whole tumour sample VAFs or single cell sampling phylogenetic trees. Alternatively, a user can provide a set of parameters (please refer to the package documentation for the details of each input parameter format) to simulate a synthetic target tumour to then recover these input parameters.

The functions output a sequence of files containing sets of inferred parameters corresponding to each SMC round (that can then be used to construct the posterior distributions for each parameter).

2.6.2 Inferring tumour growth model parameters

The spatial effects of drift and sampling bias one can observe are remarkable and represent a major challenge for the correct subclonal reconstruction of tumours growing in three-dimensional space. Due to the inherent complexity, analytical solutions to this problem that take space into the account remain challenging, although some attempts to tackle this difficult question are being undertaken [167]. Understanding the complex impact of spatially growing cell populations on the actual genomic data requires an approach based on computational simulations.

Here we devise a statistical inference framework, similar in spirit to what we previously proposed for well mixed populations [89], that aims at recovering the evolutionary parameters of each individual tumour from the type of data we have discussed so far. We constructed a test-set of 34 synthetic tumours simulated with different parameter values and assessed the error in recovering the parameters used to generate these tumours after statistical inference with an Approximate Bayesian Computation – Sampling Monte Carlo (ABC-SMC) approach [89, 162, 168, 169]. We used approximately one million simulation instances to perform parameter inference using uniform priors. We were particularly interested in comparing the

accuracy provided by the different spatial sampling methods in recovery evolutionary dynamics. We studied three different sets of tumours. In the first set, we investigated parameter recovery in tumours with homogeneous (exponential) growth, with and without selection but with no cell death. In the second set, we added stochastic cell death as an additional factor. In the third set, we studied cases of boundary driven growth where we also examined our ability to recover the extent of the boundary driven parameter a. In all three sets, we studied the differences in the ability to recover parameter if we used a single bulk sample of the whole tumour multi-region punch biopsies, multi-region needle biopsies or single-cell sequencing. Following the inference of the parameters, we calculate the percentage error for each parameter as a difference between the true parameter value and inferred parameter value (mode of a parameter posterior distribution) scaled by the true parameter value. Then we plot the distributions of the percentage errors for each parameter per growth model and sampling strategy in Figure 2.20.

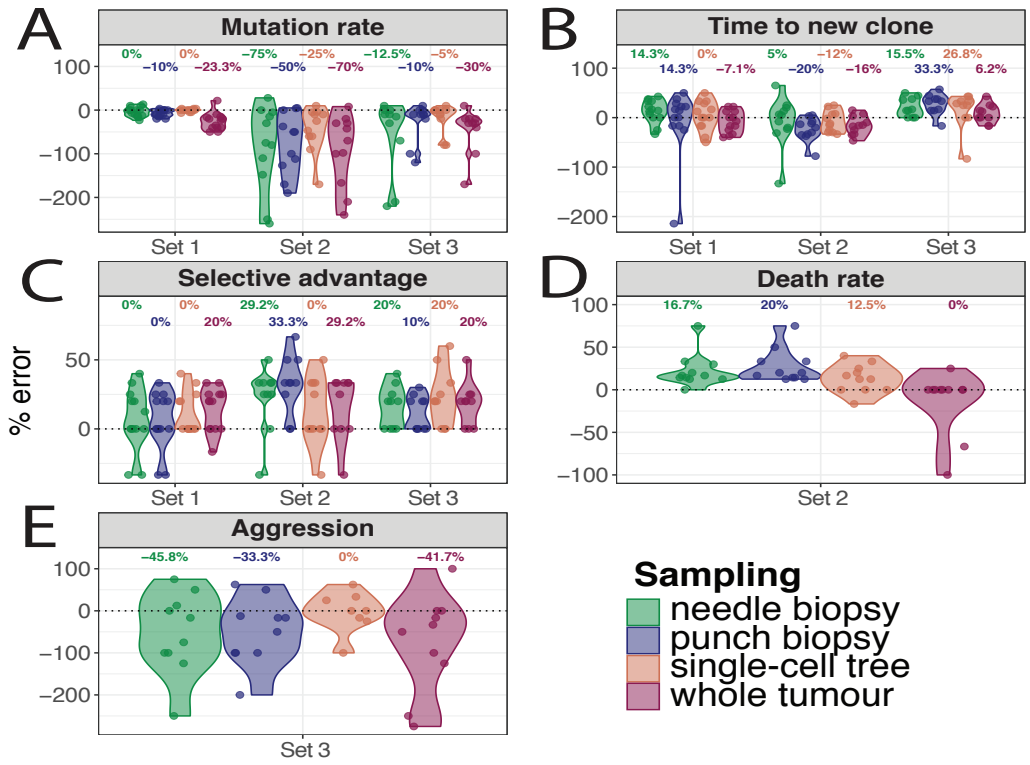


Figure 2.20: Statistical inference framework to recover evolutionary parameters. We combined our model with a statistical inference framework (Approximate Bayesian Computation – Sequential Monte Carlo) in order to infer the evolutionary parameters of selection and growth from the data. We tested this framework on 34 synthetic (target) tumours for which we generated genomic data. Our of these 34 target cases, 13 were characterised by homogeneous growth with no cell death (A, Set 1), 11 were homogeneous but with cell death (B, Set 2), and 10 where characterised by peripheral growth (C, Set 3), see all parameters in Table S1. We tested the ability to recovery parameters of 4 different sampling schemes: punch samples, needle biopsies, single cell phylogenetic trees and whole-tumour sampling (see Materials and Methods for details). We report the percentage error of the inference (true parameter value – inferred value based on the mode of the posterior probability) for each parameter and scenario. See prior parameter ranges in Table S2. (D) For the homogeneous stochastic cell death scenario (Set 2), we also report the error in recovering the death rate parameter d. (E) For the boundary driven growth scenario we report the error in recovering boundary driven growth parameter a (Set 3).

Not surprisingly, the scenario with exponential homogeneous growth without cell death was the one where the evolutionary parameters were the easiest to recover because spatial constraints were limited and the number of unknown parameters

lowest (Figure 2.20 A-C, "Set 1"). In particular, the percentage-error in recovering the mutation rate u was particularly low, especially using single-cell sequencing (Figure 2.20 A, "Set 1"). The mean percent error of the parameters t (Gillespie time when a new mutant is introduced) and s (selective coefficient of the new mutant), in the case of homogeneous growth were also within 20% and overall agrees with our previous observations in well-mixed populations [89]. The presence of stochastic cell death, even within a homogeneously growing tumour, introduced significant spatial and sampling biases (spatial drift) that led to a higher error in the recovery of the parameters (Figure 2.20 A-C, "Set 2"). Furthermore, some of the evolutionary parameters became unidentifiable (mutation and death rate). In this scenario, the best sampling strategies to recovery the death parameter d were single-cell sequencing or whole-tumour sequencing, reflecting the need to collect large population of cells for the correct estimation of this parameter (Figure 2.20 D). Boundary driven growth also introduced significant biases that led to higher percent-error values in the recovered parameters (Figure 2.20 A-C, "Set 3"). Here, single-cell sequencing was best in recovering the boundary driven growth parameter a (Figure 2.20 E). See Figure 2.21 for summary statistics from the simulations in Figure 2.20. Parameter dependency in the inference of t and s combinations is reported in Figure 2.22. We performed the same inference approach but with 3-dimensionally growing tumours using a test set of a single simulated "target" tumour and inferred the parameters using approximately 10 million simulated cancers and found similar results (Figure 2.23).

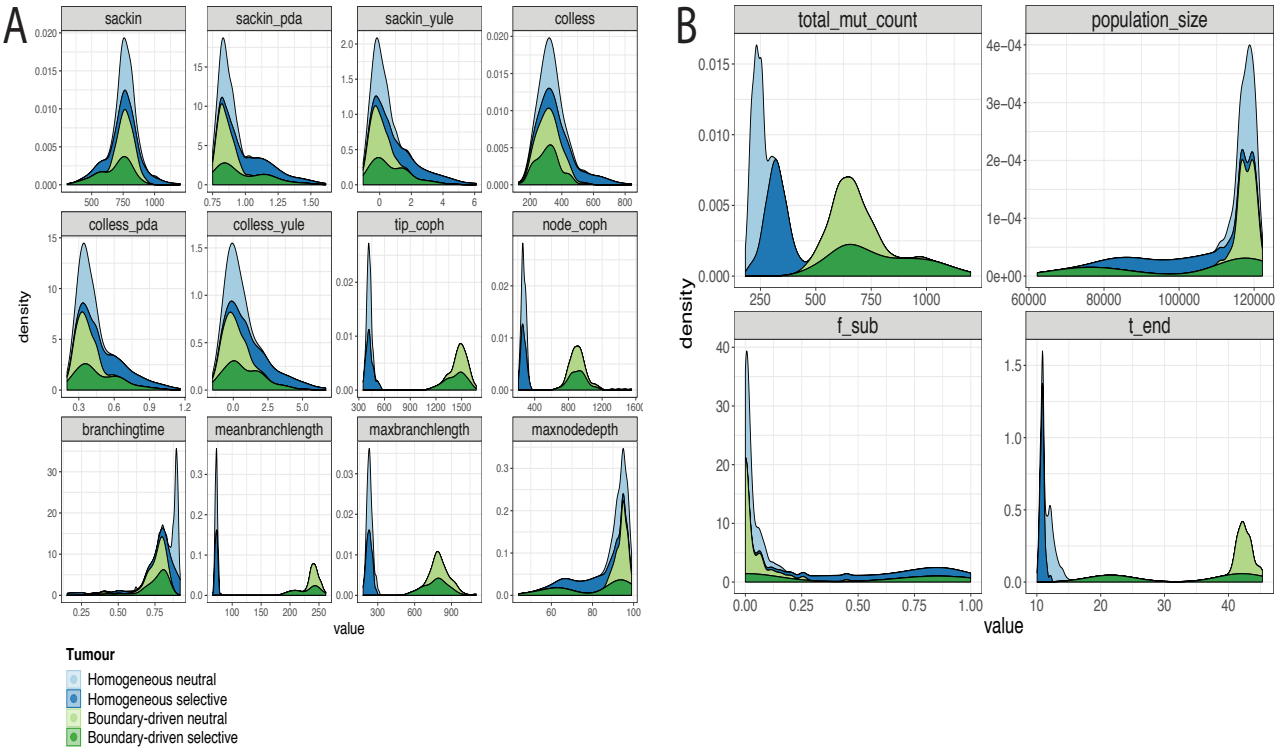


Figure 2.21: Comparing site frequency spectrum and phylogenetic tree balance index statistics for each representative scenario and sampling strategy. (A) Distributions of different summary statistics from single cell sampling (100x) phylogenetic trees for the four representative cases. The balance index-based statistics (Sackin, Colless with their different normalisation approaches – Yule, PDA) seem to have similar shapes among all four tumour cases, while tip and node Cophenetic distance-based statistics show different trends for neutral versus selective examples with not observable variation between homogenous and boundary driven tumours. Branch length-based statistics give similar results as cophenetic distances. Only one statistic, maximum node depth, tend to have longer flat tails for boundary driven tumours compared to homogenous tumour simulations. (B) For each of four tumour examples, we compare total number of passenger mutations and final population sizes along with the time the simulations finish and the final frequency of the new sub-population (introduced after a driver event).

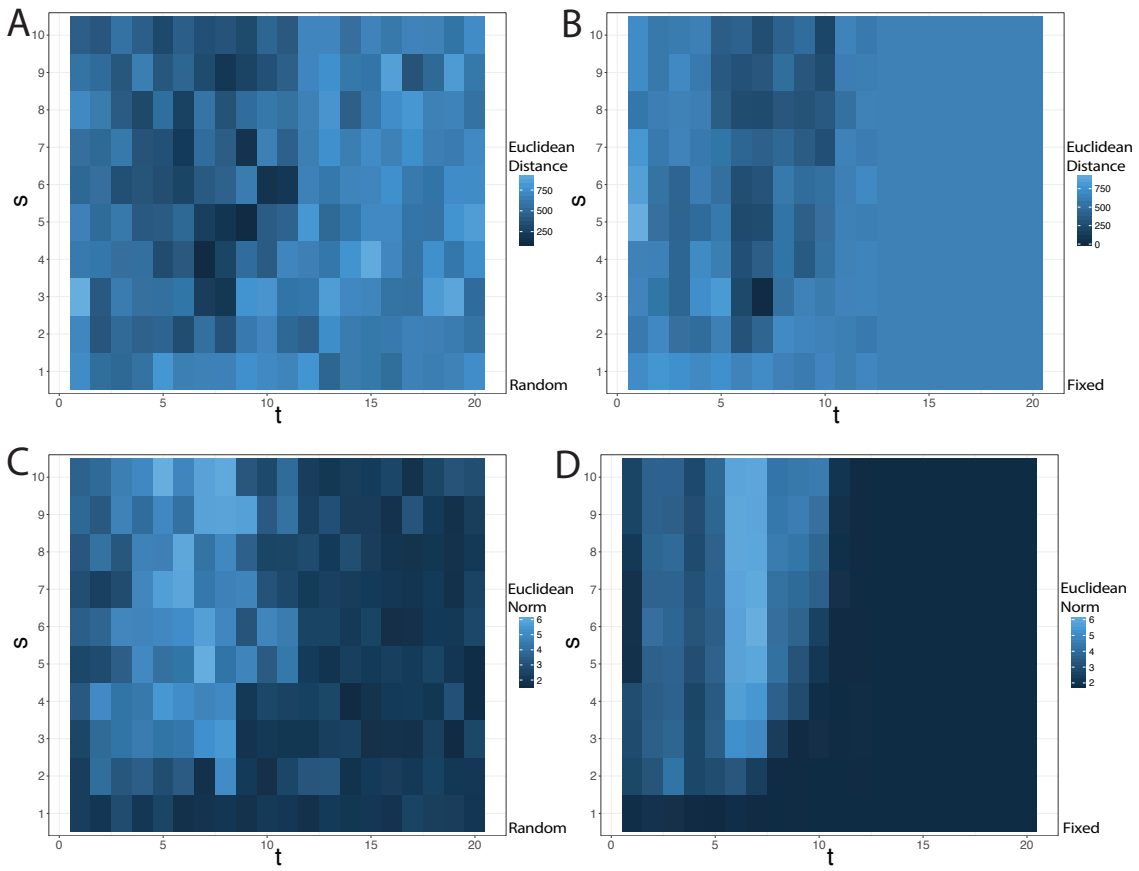


Figure 2.22: The effect of stochasticity on the dependence of t and s parameter combinations on the VAF distribution. To explore the interdependence of the parameter pair t and s, for their different values we simulate tumour growth while fixing all the other parameters (2D grid size= 400, $u = 10$, $d = 0$, $a = 1$). We summarised the obtained tumours by calculating either the Euclidean norm of the obtained whole tumour VAFs (C, D) or the calculating Euclidean distance between the cumulative VAF distributions of the simulated and a chosen target tumour (in this case target tumour parameters are $t = 7$, $s = 3$) (A, B). To reduce the effect of stochasticity we fix the random seed in (B) and (D) and they indeed showed less scattered patterns of (A) and (C) plots respectively.

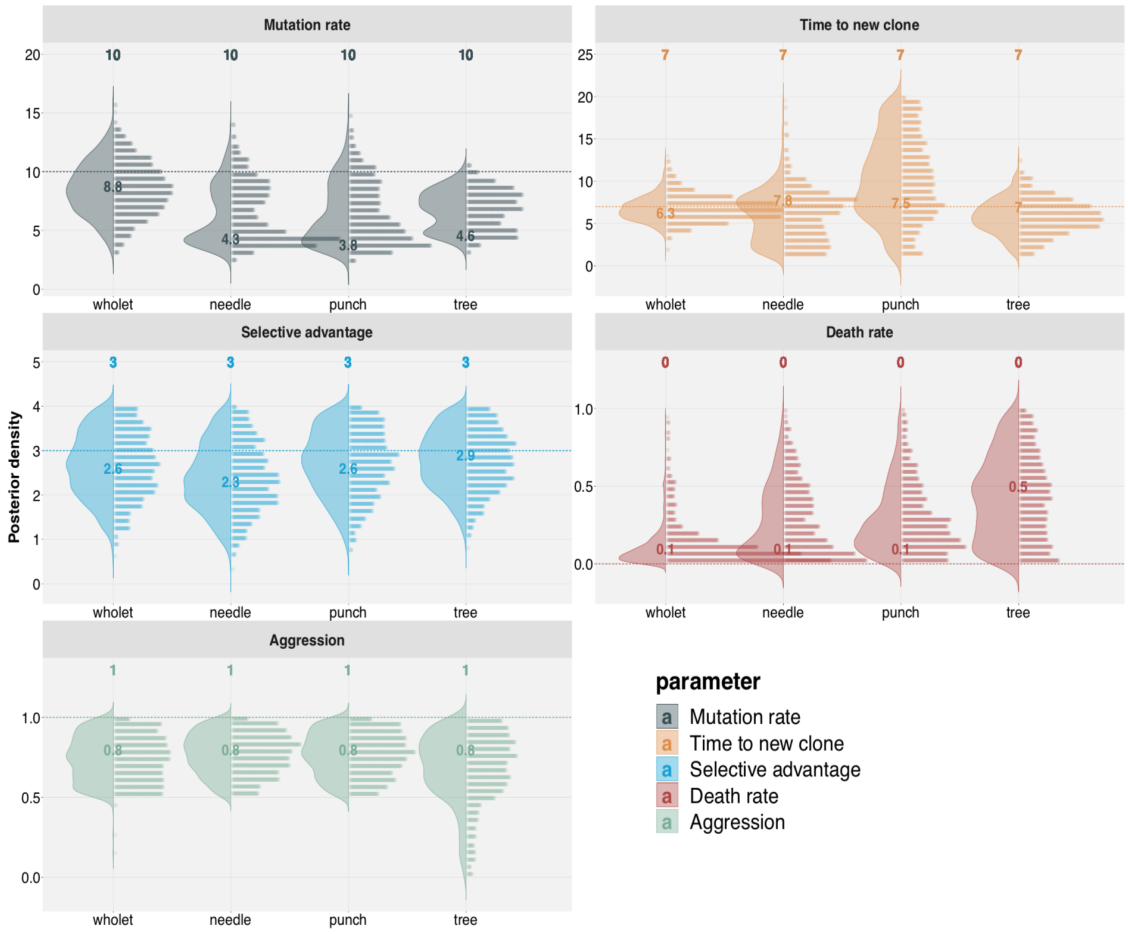


Figure 2.23: Posterior distributions for a 3D model. ABC-SMC inference for a selective homogenous growth simulation in 3D space. Real "target" values are reported as dashed lines. We run this ABC framework similarly to 2D simulations, where we recover each parameter at a time; first varying all parameters, once one is converged, fixing it at its inferred value and rerunning the simulation varying the parameters left to infer. Here we first recovered mutation rate, then time and selective advantage (together), and finally death rate and aggression (together as well). Similar to 2D models, our ABC framework with whole tumour sampling performs the best compared to other sampling strategies.

2.7 Discussion

It is now widely accepted that tumour growth is governed by evolutionary principles. Thus, recovering the evolutionary histories of tumours is essential to understanding of patient-specific tumour growth and treatment response. However, these analyses are inevitably based on limited information due to sampling biases, noise of known and unknown nature, lack of time resolved data amongst many others. Despite these limitations, many approaches based on single sampling, multi-region bulk profiling, or single cell sequencing have been developed. Information from such data is often derived using purely statistical bioinformatics methods such as clustering analyses, without consideration of the confounding underlying influence of the cellular mechanics of tumour growth. Here we explicitly investigated spatial effects on the evolutionary interpretation of typical multi-region sequencing data of tumours. We found that the effects of sampling bias and spatial distributions of spatially intermixed cell populations critically depend on the mode of tumour growth as well as the details of the underlying sampling and data generation procedure. Most surprisingly, we could observe clusters of over-represented alleles in the VAF distribution of some tumour samples that were indistinguishable from positively selected subclonal populations, despite emerging solely due to the spatial distribution of cells. Such clusters vary depending on how one samples a tumour, and would therefore cause a major challenge for the evolutionary interpretation of cancer genomic data based on subclonal reconstruction.

We furthermore presented a Bayesian inference framework to recover evolutionary parameters from our spatial distributions. Evolutionary parameters such as strength of selection or mutation rates may be important surrogate measurements of evolvability, and hence linked to progression and treatment resistance, as it has been demonstrated for the rates of chromosomal instability [25, 170]. Again, we observe

that our ability to precisely recover certain evolutionary parameters depend on the scenarios of tumour growth and spatial sampling strategies. However, we do believe that although complex, the situation is far from hopeless. More involved statistical frameworks based on first principles of tumour growth can help resolving some of the evolutionary parameters on an individualised patient basis. Importantly, careful spatial sampling and single-cell sequencing can mitigate some of the confounding issues. We do acknowledge that our model has some important limitations, such as the infinite allele assumption (which could be violated by copy number loss [111]). Also, for computational feasibility we mostly focus on 2D spatial analyses and of a relatively limited number of cells with respect to the billions of cells present in a human tumour. We also acknowledge that we do not offer a closed mathematical formulation for the distribution of alleles under spatial effects, which would be very useful but remains a very difficult problem that can only be tackled partially (e.g. [154]). Additionally, more realistic models of tumour growth dynamics that account for force fields between cells [171] have been developed that could improve on the study of spatial patterns of growth [87, 172]. For computational feasibility, especially in regards to the necessity of performing statistical inference on the data and generate thousands of simulations, we restricted our analysis to the stochastic cellular automaton model we propose here. Nevertheless, our approach highlights the importance of spatial modelling of real data and the impact of confounding factor in our estimate and understanding of tumour evolution.

Chapter 3

Linking mutational signatures to the epigenome

3.1 Introduction

Not just genetic but also epigenetic alterations are involved in tumorigenesis and cancer evolution. It is known that genetic alterations can cause epigenetic changes (e.g. mutations in chromatin modifier genes) [173, 174, 175]. And it is also known that the epigenetic configuration of the genome influences the accumulation of mutations due to different efficiency of mismatch repair genes that act in the presence or absence of chromatin (e.g. the associations found between chromatin structure and the corresponding mutational load [176]). The patterns of mutation accumulation in the genome can be studied using mutational signatures introduced by Alexandrov et al. in [67]. Here we hypothesize that different mutational processes, giving rise to distinct mutational signatures, are active in epigenetically different regions of the genome. To test our hypothesis, we need an epigenetic map of the regions such as promoters, enhancers, coding and non-coding DNA sections. Such epigenomic

maps are cancer type specific. Our collaborator, Luca Magnani, and his lab have derived such a mapping for breast cancer and hence we developed a method to test the enrichment of mutational signatures in different epigenetic regions of the breast cancer genome.

Breast cancer is the most common cancer type and one of the leading causes of death for women worldwide. Oestrogen-receptor (ER) positive cells are present in over 70% of tumours [177], making oestrogen receptor positive breast cancer the most common subtype. Although significant progress has been achieved through the use of hormonal therapies, less is known about its particular aetiology and evolution. Using the data from [178], where they profiled the active regulatory landscape of over 50 ER+ breast cancer patients using epigenetics-based assays, we partitioned the genome into functionally distinct epigenetic regions, such as regulatory, coding, repetitive, transcribed and not-transcribed regions. We analysed each of these partitions in over 560 whole genomes from ER+ and ER- cancers and identified the activity of different mutational signatures across the distinct genomic regions, between the two cancer subtypes [179].

3.2 Epigenomic annotations

Among all breast cancer cases, around 70% contain different amounts of estrogen-receptor (referred to as ERα-positive) cells, that play a significant role in the disease progression. In cell lines it has been show that parallel to genetic evolution, epigenetic changes also play a role in breast cancer progression and resistance to endocrine therapies (ET) [180]. Several studies showed that epigenetic information can modify gene transcription states during cell division [181, 182, 183, 184]. Epigenetic modifications can also interact with ERα associated pioneer factors and thus modify its binding to enhancers [185, 186]. Epigenetic regulatory regions have been also

successfully mapped and non-coding parts of DNA been annotated using epigenetic modifications of histone proteins [187, 188].

Acetylation of lysine 27 on histone 3 (H3K27ac) has been shown to be associated with promoters and enhancers of transcriptionally active genes [189, 190, 191]. In [178] they profiled 55 ERα-positive breast cancer samples using H3K27ac ChIP-seq and built the list of clinically relevant DNA regulatory regions. Regions enriched by H3K27ac were classified into 23,976 proximal promoters and 326,719 enhancers. To identify promoters, profiling four patients was enough, whereas for enhancers around 40 patients had to be profiled that reflects the 1:10 ratio of the identified enhancers and promoters. The analysis was also in agreement with several studies showing that enhancers are the main sources for cell-type-specific transcriptional differences. In the study, they also developed a sharing index – SI to indicate the number of patients sharing the H3K27ac signal at each specific location and annotated enhancers and promoters as a function of this index. This index corresponds to the level of recurrence in the cohort of a given epigenetic state. The majority of obtained enhancers were patient-specific i.e. their SI = 1, while active promoters had much higher values of SI. Hence, they showed that enhancers play a dominant role in forming the epigenetic heterogeneity of ERα-positive breast cancer.

We used these annotations of breast cancer and partitioned/annotated whole genome sequencing of breast cancer cases, the analysis of which is discussed in the upcoming sections.

3.3 Mutational signatures

Somatic mutations that affect cell growth and division are one of the main sources of cancer initiation and progression [192]. Defects in DNA repair mechanisms or environmental mutagens increase the rate of somatic mutations and hence elevate

the chance of cancer development. Different mutagenic processes cause different molecular lesions that by themselves are activating different repair mechanisms, and hence the whole process creates specific mutational spectra – referred to as 'mutational signatures' [67]. The signatures can help identify which mutagenic processes are active in the tumour, find specific characteristics of different tumour subtypes or even direct therapeutic interventions by being markers for therapeutic response [193].

Mutational signature is a specific distribution of the 96 possible substitution types (6 types of substitution x 4 types of 5' base x 4 types of 3' base) that occur at a base pair in the middle of a trinucleotide. Signatures are therefore inferred by identifying common patterns of mutations with their sequence context across many tumours. The first signature discovery method was based on Non-Negative Matrix Factorization (NMF) [67] (which tries to decompose the matrix of mutations (where rows correspond to samples and columns to mutational motifs i.e. SNV triplets) into two matrices of mutational spectra and their corresponding expression weights) and still remains to be the most broadly used tool among other signature discovery methods [194, 195, 196, 197]. Using the original NMF-based method and sequence analysis of over 7000 human cancers (mostly exome sequencing) Alexandrov et al. initially discovered 22 different mutational signatures. Then they later expanded the analysis to over 12000 human cancers and inferred 30 distinct mutational signatures (https://cancer.sanger.ac.uk/cosmic/signatures). This set of 30 signatures, also called COSMIC signatures, has since been the most widely used set of signatures. Recently, the analysis was expanded to include 23000 human tumours with 71 different cancer subtypes. This time they also included non-point mutations such as indels and dinucleotide (tandem) mutations and inferred 49 distinct signatures [198].

Examining signatures with their associated aetiology across different cancer types

has revealed different mutagenic characteristics specific to different carcinogenic processes. From the 30 distinct COSMIC signatures, around third was found to be associated to endogenous mutagenic processes, such as the activity of the APOBEC family of deaminases and deamination of 5-methylcytosine, defective DNA polymerases or defective DNA repair processes [199]. Some signatures have been associated with exposure to mutagenic agents, such as tobacco carcinogens, UV radiation, alkylating chemotherapy drugs, aristolochic acid and aflatoxin B1 carcinogens. The aetiology of almost half of the signatures still remains unknown. There have been several approaches to produce mutagen-induced signatures in vitro as identifying unknown signature aetiologies should provide better insights for understanding and better characterising different cancer progression pathways and eventually link it to patient clinical outcome.

3.4 Data analysis

3.4.1 Data annotation

As discussed in section 3.2 of epigenomic annotations, we partitioned the genome into functionally distinct categories, such as regulatory, coding, repetitive, transcribed and not-transcribed regions using the data from [178]. Figure 3.1 shows the region size distribution for the 9 obtained epigenetic regions. As we can see, the longest regions are non-regulatory non-coding DNA regions. Based on the SI (sharing index from the study [178]), that was used to differentiate between recurrent epigenetic regions across the genome and non-recurrent ones, we split the enhancer regions into three groups and labelled as unique, shared and recurrent if their corresponding SI number was equal to 1, was between 1 and 21, or above 21, respectively. Enhancer and promoter region sizes are much smaller compared to the other DNA regions. The distinctly smallest regions are the recurrent enhancers in non-coding regions.

For our further analysis, we took into account the length of each epigenomic region and scale the results accordingly to make sure the patterns we observed were not mainly influenced by the different region size distribution.

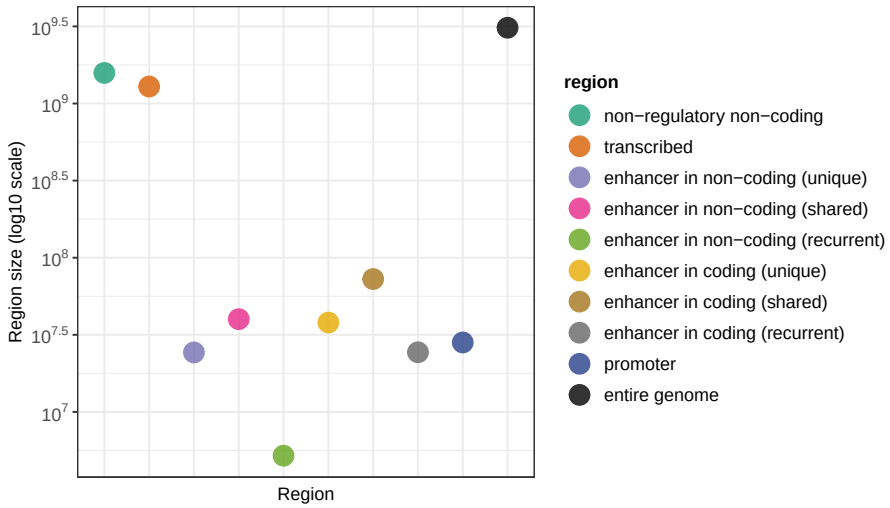


Figure 3.1: Epigenomic region size distribution. The region size distribution for the 9 epigenomic regions plus the entire genome. The longest regions are non-regulatory non-coding and coding DNA regions. Based on the SI (sharing index from the study [178] that indicates to the patient count where a given enhancer was observed to be active.) numbers we split the enhancer regions into three groups and labelled as unique, shared and recurrent if their corresponding SI number was equal to 1, was between 1 and 21, or above 21, respectively. Enhancer and promoter region sizes are much smaller compared to the other DNA regions. The distinctly smallest regions are the recurrent enhancers in non-coding regions.

Initially, we analysed three different sets of breast cancer data: two whole genomes (primary cancers - Nik-Zainal [179] and metastases - Hartwig [200]) and one whole exome (TCGA [201]) sequencing data. Figure 3.2 shows the distributions of mutational burden per base pair for the three datasets. There was no BRCA status available for Hartwig dataset, hence we labelled it as 'BRCAunknown'. We can see that the mutational burdens across the 10 genomic regions are more densely distributed in Hartwig than the other two datasets. Both in Nik-Zainal and TCGA,

there are higher mutational burdens across all regions in BRCA wild type (WT) patients compared to BRCA mutant. In our further analysis, we will only focus on BRCA WT patients (here we included them just for the comparison). Coding DNA regions, all categories of enhancers in coding regions and promoters have higher relative mutational burden compared to the non-coding both regulatory and non-regulatory DNA regions in Nik-Zainal dataset. From the TCGA dataset, we can see that enhancers have higher mutational burden than non-regulatory coding regions, while the entire exome has overall much higher mutational burden than each region separately. This could largely be due to higher rate of false positives in exome sequencing. To explore the mutational burden distribution in Hartwig dataset we plot it separately from the other two datasets (Figure 3.3). Here we do not see as much difference in the distribution of mutational burden between the regions, and overall the ER-positive patients have higher mutational burden compared to ER-negative.

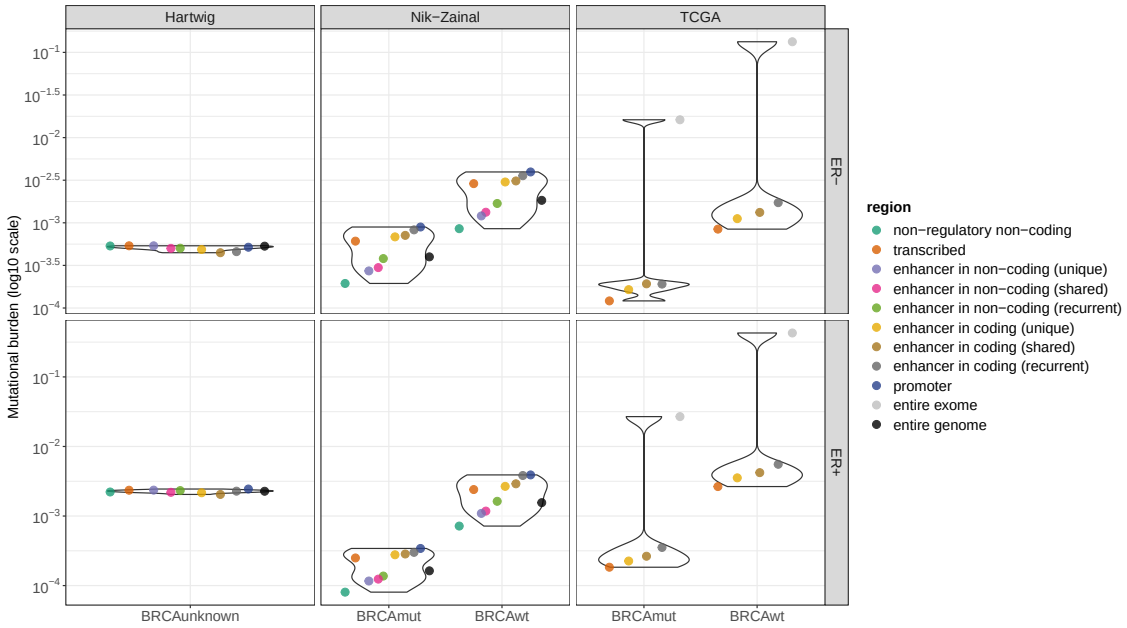


Figure 3.2: Distributions of mutational burden across the epigenomic regions. The distributions of mutational burden per base pair for the three datasets - Hartwig, Nik-Zainal, TCGA. There was no BRCA status available for Hartwig dataset, hence we labelled it as 'BRCAunknown'. BRCAwt and BRCAmut stand for BRCA wild type and mutant, respectivelly. The mutational burden across the 10 genomic regions are more densely distributed in Hartwig data than the other two datasets. Both in Nik-Zainal and TCGA, there is higher mutational burden across all regions in BRCA wild type patients compared to BRCA mutant. Regions with the higher relative mutational burden are promoters and enhancers in coding regions

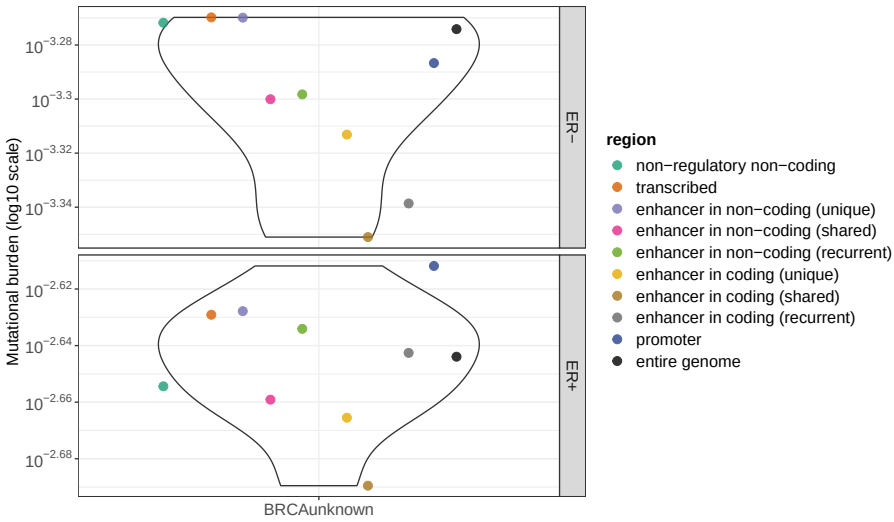


Figure 3.3: Distributions of mutational burden across the epigenomic regions in Hartwig dataset. To explore the mutational burden distribution in Hartwig dataset we plot it separately from the other two datasets (from Figure 3.2). Here we do not see as much difference in the distribution of mutational burden between the regions, and overall the ER-positive patients have higher mutational burden compared to ER-negative.

3.4.2 Signature activity

We run the signature activity analysis on the three breast cancer datasets using the COSMIC set of mutational signatures. Figure 3.4 shows normalised inferred signature weights for Hartwig, Nik-Zainal and TCGA datasets. In Hartwig dataset, we can see that signatures 2 and 13 are over-represented in ER-positive patients. ER-negative patients have dominantly signature 3 expressed, and also signature 2 but to

a lesser extent compared to the ER-positive patients. The aetiology of signature 2 is defined by its association to the activity of the AID/APOBEC family of cytidine deaminases. A similar characteristic is attributed to signature 13 but is associated with the process of converting cytosine to uracil (one of the main four bases found in DNA and RNA). Signature 3 is associated with failure of DNA double-strand break repair by homologous recombination (for the full list of all known signatures aetiologies please refer to the following link: https://cancer.sanger.ac.uk/cosmic/). Similarly, in the Nik-Zainal data, signature 2 is overrepresented in ER-positive patients and underrepresented in ER-negative. Although the difference is more evident if we look by BRCA wild type and mutant subgroups, where signature 2 is more dominant in BRCA wild type patients. Signature 3 is the dominant signature expressed in BRCA mutant patients compared to BRCA wild type, and similar to the Hartwig dataset, it is more prevalent in ER-negative patients than in ER-positive. TCGA data analysis showed a slightly different trend, which makes sense as it is a whole exome sequencing data. Here the prevalence of signature 2 has disappeared from almost all cohorts except ER-positive BRCA mutant patients. The most dominant signatures are number 3, 5 and 12, as well as signature 1 in ER-negative BRCA mutant subset (the aetiologies of signatures 5 and 12 are currently unknown).

Besides the overall signature activity patterns over the three breast cancer datasets, we can also notice patterns of signature activity specific to each epigenomic region. For instance, from Nik-Zainal data analysis, we can see that recurrent enhancer regions (both transcribed and non-coding) have higher activity level of signature 2 compared to other regions. Another interesting trend is of signature 1 (the aetiology of which is the correlation between the inferred mutational spectrum and the age of cancer diagnosis) which is equally present in all regions except the recurrent enhancers but in non-coding DNA only. To test the significance of these observed

patterns, we performed jackknife resampling based statistical analysis. In the remaining sections, I will present the details of the analysis and discuss the obtained results.

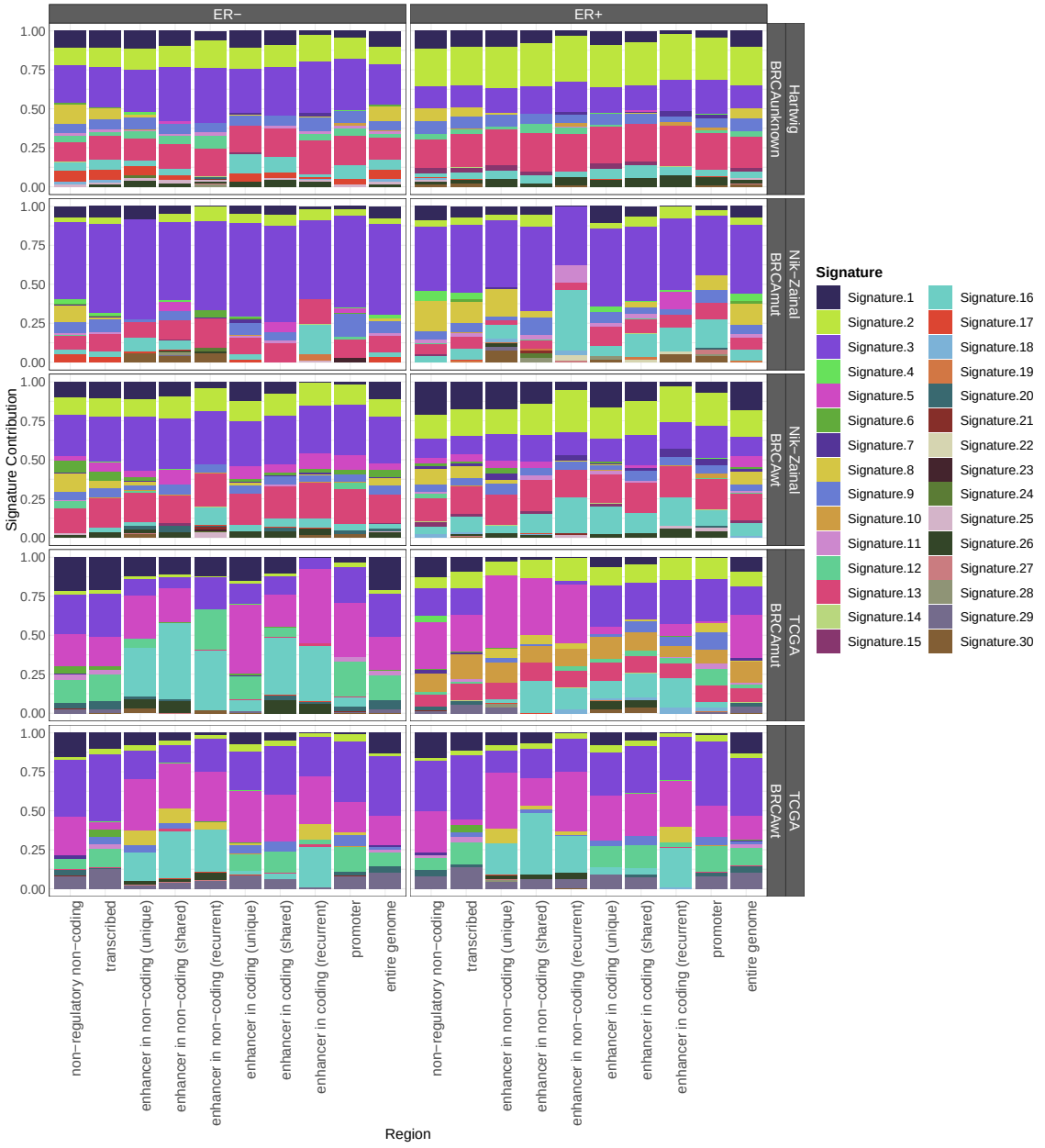


Figure 3.4: Signature activity weights per epigenomic region. Signature activity weights (normalised) per 10 epigenomic region. In Hartwig, signatures 2 and 13 are overrepresented in ER+ patients. ER- patients have dominantly signature 3 expressed, and also signature 2 but to a lesser extent compared to the ER+ patients. In Nik-Zainal, signature 2 is overrepresented in ER+ patients and underrepresented in ER-. Although the difference is more evident if we look by BRCA wild type and mutant subgroups, where signature 2 is more dominant in BRCA wild type patients. Signature 3 is more dominant in BRCA mutant patients compared to BRCA wild type, and similar to the Hartwig dataset, it is more prevalent in ER- patients than in ER+. In TCGA, signature 2 has disappeared from almost all cohorts except ER+ BRCA mutant patients.

3.5 Statistical modelling

3.5.1 Population level analysis

Initially, we performed the analysis on the patient level i.e. inferred signature activity patterns per patient individually and looked at the distribution of signature proportions over the entire cohort. Patient-level analysis suffers from not enough mutations per patient per region to infer mutational signatures appropriately. Hence, we decided to perform the analysis on the aggregated number of mutations over the entire patient cohort per dataset. Figure 3.4 presents the results of one such call for mutational signatures over the aggregated number of mutations. To test the significance of the observed patterns we performed bootstrapping analysis, specifically, jackknife resampling by regions and patients. Jackknife resampling by regions means leaving out 10% of each epigenomic region and rerunning the signature calling analysis 100 times. The jackknife patients method leaves out 10% of patients each time and repeats the analysis. If the obtained signature weight distributions are not very wide (i.e. each run consistently returns activity weights close to each other per signature) and the observed patterns are maintained, then we can conclude that the pattern should be statistically significant.

We can also randomise the annotations of the epigenomic regions themselves. That

is, we shuffle the regions; keep their size but change the annotation and again rerun the analysis 100 times. This allowed us to test the robustness of the signature activity patterns we observed per region (that they are not some consistent random noise). If after randomising the annotations and rerunning the analysis, patterns do change and signature weight distributions become wider, then we can conclude that the obtained signal should not be noise.

3.5.2 Jackknife resampling

As discussed in the previous section, to test the significance of the obtained results of different signature activity patterns in different epigenomic regions, we run jackknife resampling analysis. For this analysis, we only focus on the whole genome sequencing datasets – primary tumours (Nik-Zainal) and metastases (Hartwig). We run jackknife resampling in two different ways: resampling regions and resampling patients. Resampling regions with leaving out 10% of all regions each time and (re)inferring the signature activity patterns per run, allows us to monitor the significance and consistency of the signal we observe after a single run. Similarly, jackknife resampling by patients with leaving out 10% of the total cohort and rerunning the analysis, lets us generalise the obtained results and clear the doubt if any observed pattern is significantly affected by only a subset of patients (characterised by a high mutational load, for instance).

As in the main signature activity analysis, we split the dataset into 4 subgroups: BRCA WT vs BRCA mutant and ER-positive vs ER-negative (although for the Hartwig dataset, we did not have BRCA status, and split the dataset into 2 groups by ER status only). Here we only present ER-positive BRCA WT patient cohort analysis as we think they showed more interesting results (although we include each subgroup analysis - excluding BRCA mutant patients - in the appendix).

Figure 3.5 shows Nik-Zainal dataset ER-positive BRCA WT patient cohort analysis

jackknife by regions. We excluded signatures with the activity level less than 5% in all regions. From the figure, we can see that, now we have distributions of different signature activity weights per one signature per region rather than a single activity level per signature. As in the single run, the most interesting signature activity patterns were observed for signatures 1 (age) and 2 (APOBEC); the recurrent enhancer regions in coding and non-coding DNA are protected from signature 1 and enriched by signature 2. This pattern was maintained after running the jackknifing resampling of the regions and reinferring the signatures.

Similarly, Figure 3.6 shows Nik-Zainal dataset ER-positive BRCA WT patient cohort analysis jackknife by patients. Again, we excluded signatures with the activity level less than 5% in all regions and focus only on signatures 1 and 2. Jackknife by patient analysis gave very similar results to when running jackknife by region analysis; the observed pattern is maintained when rerunning the inference and leaving out 10% of patients per inference.

Very similar results were obtained for the Hartwig dataset as well. Figure 3.7 shows jackknife by region and Figure 3.7 jackknife by patient resampling results. The only difference between the Hartwig (metastases) and Nik-Zainal (primary breast cancer) datasets, is the higher activity level of signature 2 and lower activity of signature 1 in Hartwig compared to Nik-Zainal.

Overall, the most interesting observation is the enrichment of signature 2 (aetiology of which is the activity of the AID/APOBEC family cytidine deaminases) at enhancers of ER-positive breast cancer. This might indicate that there is a specific interplay of the oestregon receptor and the APOBEC enzyme through which oestrogen drives the accumulation of APOBEC-induced mutations in these cancer types.

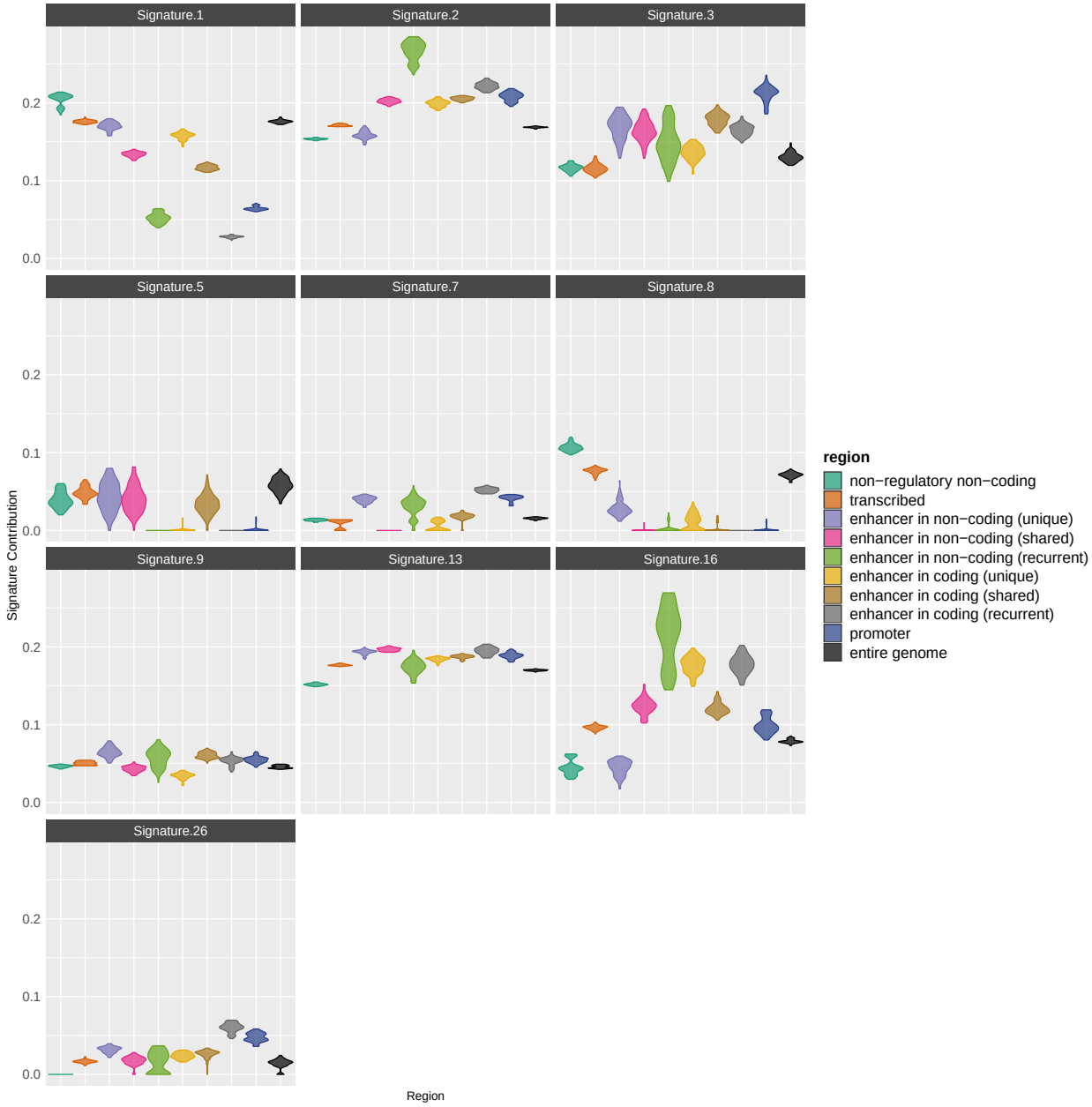


Figure 3.5: Jackknife resampling by regions (BRCA-WT, ER-positive) - Nik-Zainal. The distributions of different signature activity weights (per signature per epigenomic region) obtained after the jackknife resampling of regions and rerunning the signature activity analysis 100 times. As in the single run (Figure 3.4), the most interesting signature activity patterns were observed for signatures 1 (age) and 2 (APOBEC); the recurrent enhancer regions in coding and non-coding DNA are protected from signature 1 and enriched by signature 2. This pattern was maintained after running the jackknifing resampling of the regions and reinterring the signatures.

Figure 3.6: Jackknife resampling by patients (BRCA-WT, ER-positive) - Nik-Zainal. The distributions of different signature activity weights (per signature per epigenomic region) obtained after the jackknife resampling of regions and rerunning the signature activity analysis 100 times. As in the single run (Figure 3.4), the most interesting signature activity patterns were observed for signatures 1 (age) and 2 (APOBEC); the recurrent enhancer regions in coding and non-coding DNA are protected from signature 1 and enriched by signature 2. This pattern was maintained after running the jackknifing resampling of the regions and reinterring the signatures.

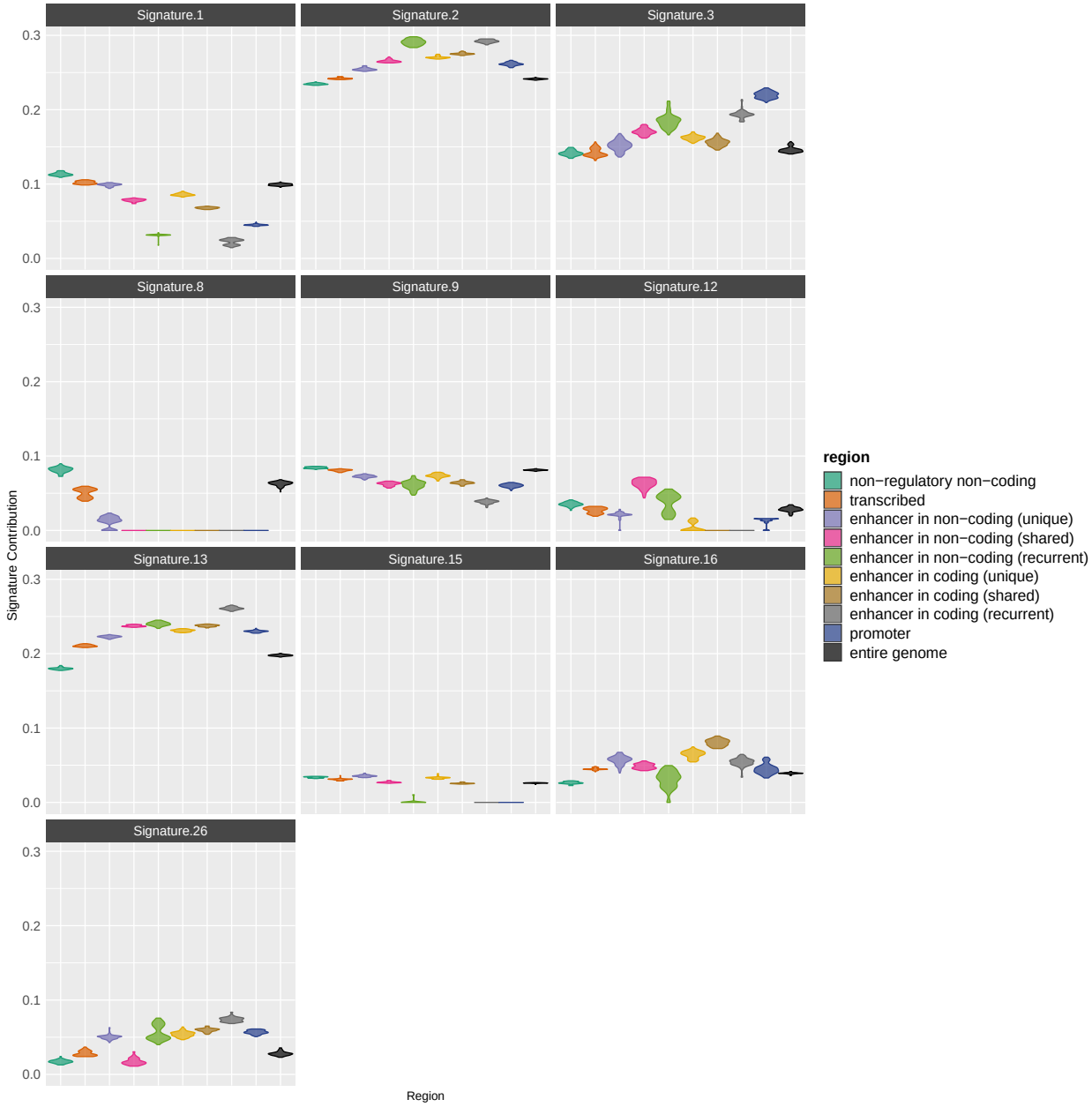


Figure 3.7: Jackknife resampling by regions (ER-positive) - Hartwig. The distributions of different signature activity weights (per signature per epigenomic region) obtained after the jackknife resampling of regions and rerunning the signature activity analysis 100 times. As in the single run (Figure 3.4), the most interesting signature activity patterns were observed for signatures 1 (age) and 2 (APOBEC); the recurrent enhancer regions in coding and non-coding DNA are protected from signature 1 and enriched by signature 2. This pattern was maintained after running the jackknifing resampling of the regions and reinterring the signatures.

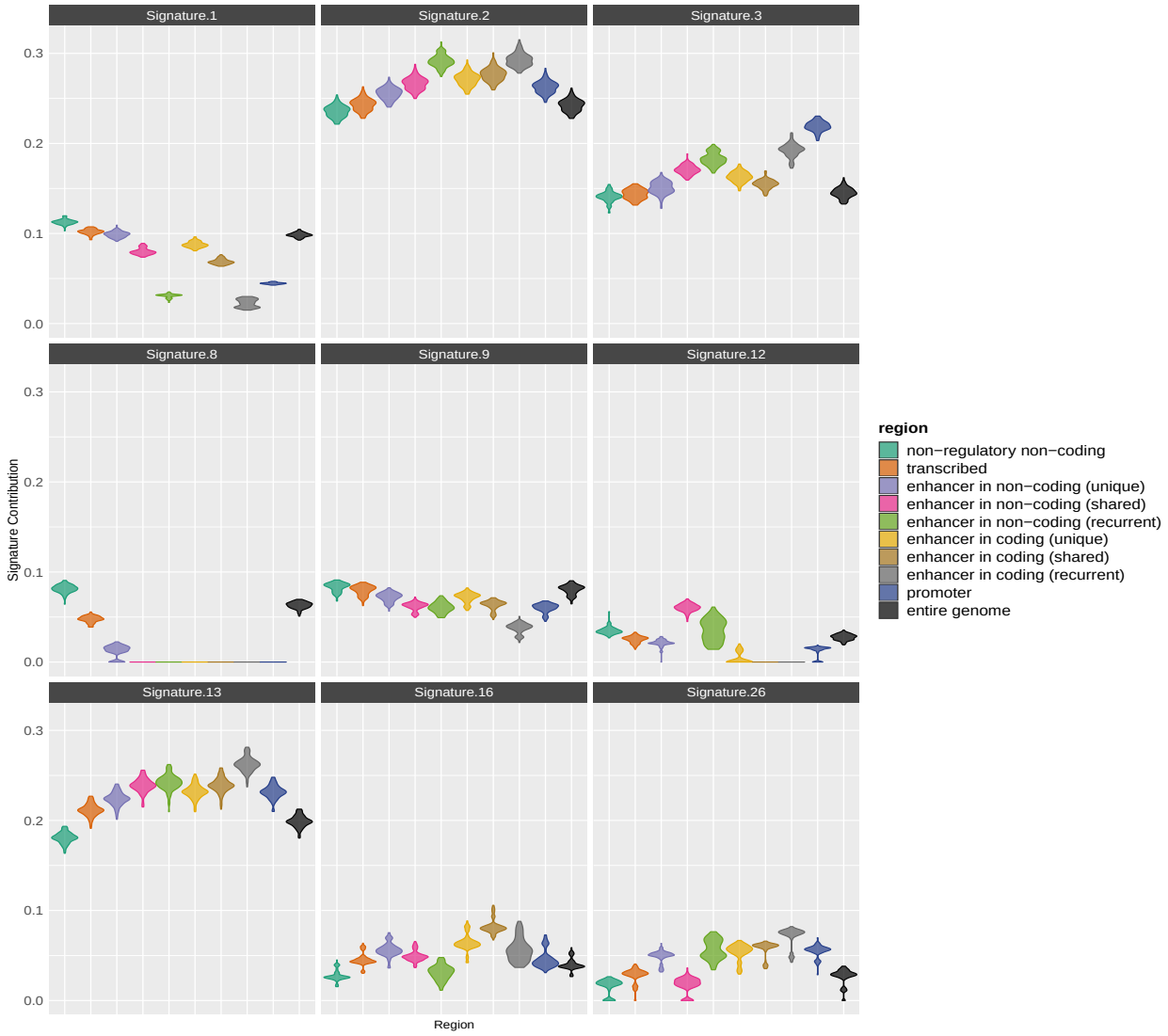


Figure 3.8: Jackknife resampling by patients (ER-positive) - Hartwig. The distributions of different signature activity weights (per signature per epigenomic region) obtained after the jackknife resampling of regions and rerunning the signature activity analysis 100 times. As in the single run (Figure 3.4), the most interesting signature activity patterns were observed for signatures 1 (age) and 2 (APOBEC); the recurrent enhancer regions in coding and non-coding DNA are protected from signature 1 and enriched by signature 2. This pattern was maintained after running the jackknifing resampling of the regions and reinterring the signatures.

3.5.3 Randomise annotations

To further test the significance of the obtained results and whether the observed signal is not caused by the specific distribution of the epigenomic regions, we decided to randomise the annotations. That is, we shuffled the regions keeping each region size constant and only swapping their annotations. If the results are significant and not a consistent noise (especially influenced by region size), after randomising the annotations and rerunning the inference for signature activity, the obtained patterns should disappear. This would confirm that the patterns we observed before are likely to be indeed due to different mutational signature activities in distinct epigenetic regions.

After rerunning the inference (100 times here as well), we saw that the observed patterns did disappear. Figures 3.9 and 3.10 show the results of the randomised annotation analysis of the Nik-Zainal BRCA-wt ER-positive and Hartwig ER-positive patient cohort datasets, respectively. From both figures, we can see that the previously observed patterns of signature activities per epigenomic region have disappeared. That is, we can no longer see the differences in the level of signature activity by the epigenomic regions. For each signature, we can see a very similar level of activity per epigenomic region. In addition, the overall distributions of different activity levels per signature (obtained from each resampling round) are wider than when running jackknife resampling analysis confirming the strength of the signal obtained when not randomising the regions.

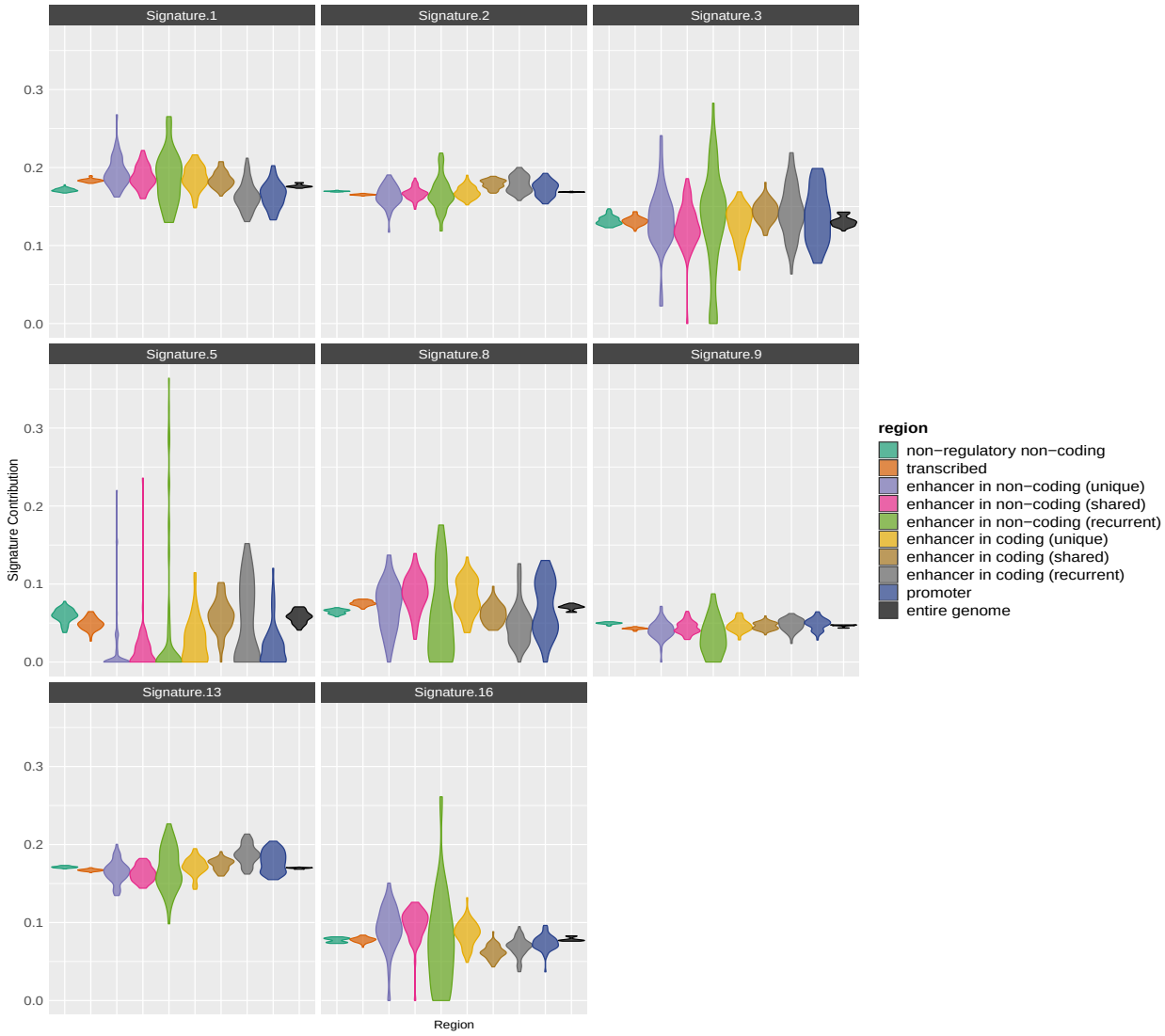


Figure 3.9: Randomised annotation analysis (BRCA-WT, ER-positive) - Nik-Zainal. The distributions of different signature activity weights (per signature per epigenomic region) obtained after randomly shuffling the regions (keeping each chunk size constant) and rerunning the signature activity analysis 100 times. The previously observed patterns of signature activity levels per epigenomic region is disappeared and the distributions look wider (confirming the strength of the signal obtained when not randomising the regions).

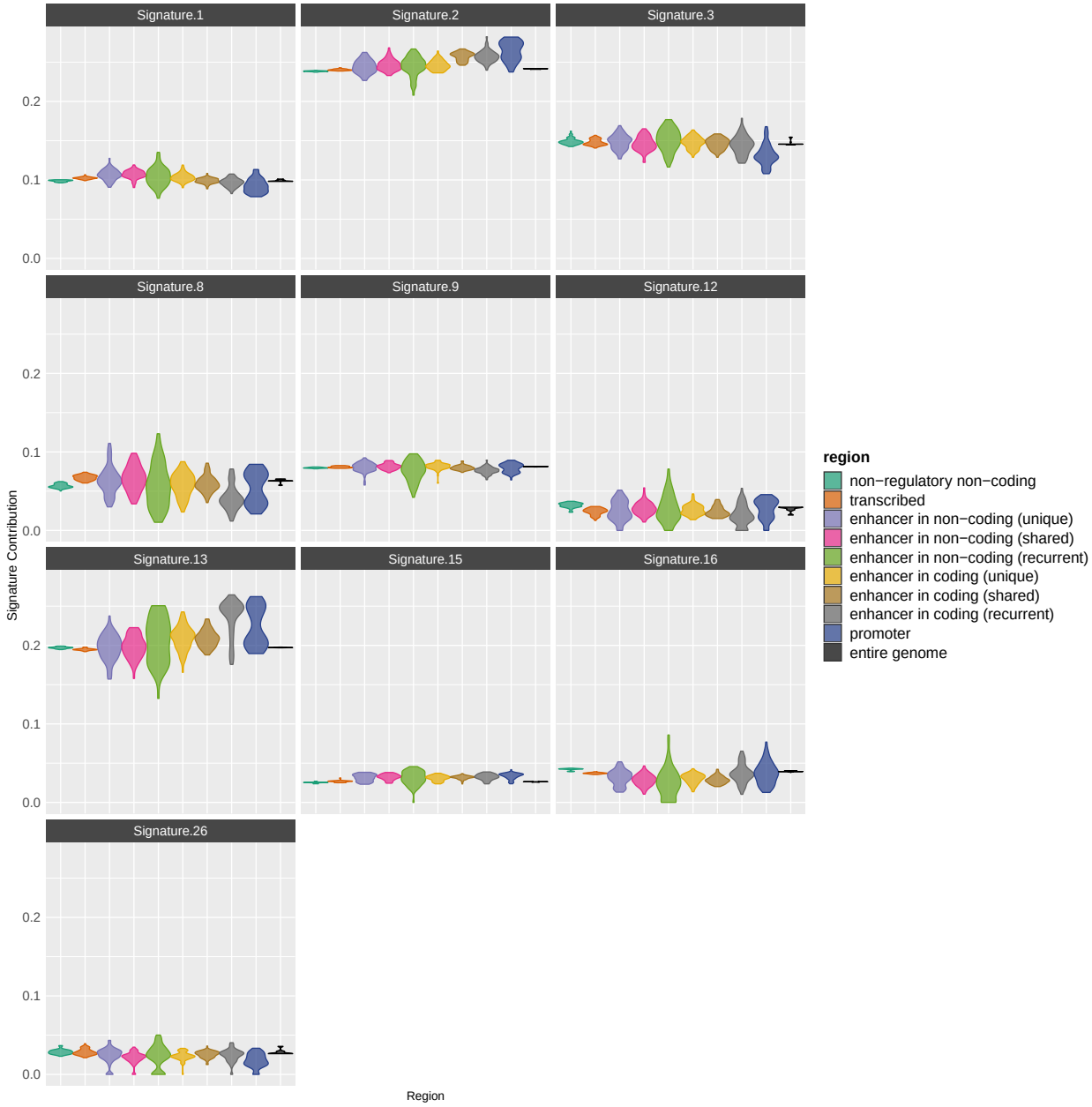


Figure 3.10: Randomised annotation analysis (ER-positive) - Hartwig. The distributions of different signature activity weights (per signature per epigenomic region) obtained after randomly shuffling the regions (keeping each chunk size constant) and rerunning the signature activity analysis 100 times. The previously observed patterns of signature activity levels per epigenomic region is disappeared and the distributions look wider (confirming the strength of the signal obtained when not randomising the regions).

3.6 Discussion

Genotype-phenotype interaction is becoming the ever-active research area [202, 203] thanks to the technological advances that produce data from multiple levels of biological processes (DNA, RNA, epigenetic markers, proteomics, metabolomics). These technological advances enabled important progress in cancer research as well. Mainly, in understanding the significant role of epigenetic alterations that are involved in tumorigenesis and cancer development. For instance, in [176] they found how chromatin structure influences the corresponding mutational load for a given genomic region. Also, several studies showed how mutations in chromatin modifier genes can lead to epigenetic alterations. Thus, understanding the link between cancer genetics and epigenetics using the novel technological advances and scientific discoveries is another important step in fully understanding the process of tumour initiation and progression.

Here we studied the associations between mutational signatures and epigenome. Namely, we found that there are different mutational processes (that give rise to different mutational signatures [67]) active in epigenetically distinct genomic regions. Our collaborator, Luca Magnani, and his lab have derived an epigenetic map of the genomic regions for breast cancer that we used to develop a method that tests the enrichment of mutational signatures in different epigenetic regions of the breast cancer genome. Using their annotations [178], we partitioned the genome into functionally distinct categories, such as regulatory, coding, repetitive, transcribed and not-transcribed DNA regions. To find the associations, we analysed three different sets of breast cancer - two whole genomes (primary cancers and metastases) and one whole-exome - sequencing data.

Our main finding was the enrichment of the COSMIC mutational signature 2 at enhancers of ER-positive breast cancer patients. The proposed aetiology of the sig-

nature 2 is the activity of the AID/APOBEC family of cytidine deaminases. This led us to further hypothesise that there is a specific interplay of the oestrogen receptor (ER) and the APOBEC enzyme through which oestrogen drives the accumulation of APOBEC-induced mutations in these cancers [204]. To validate the origin (APOBEC) of this mutational signature and its dependence on oestrogen exposure, Luca Magnani's lab has developed an in vitro system. They are maintaining a series of ER+ and ER- clones - with varying exposure to oestrogen - in a neutral evolutive regime and have been mapping the emerging mutations at a single base-pair resolution using whole-genome sequencing (WGS) at different time points. Analysis of contributing mutational signatures for each genomic section will reveal whether the APOBEC enzyme plays a role in the evolution of oestrogen-dependent cancers by specifically mutating regulatory regions.

Chapter 4

Timing epigenetic changes

4.1 Introduction

We know that every human cell contains approximately two metres long DNA within its five-micron nucleus [205]. This marvellous topological challenge is solved by hierarchical folding of DNA fragments around histone proteins that form nucleosomes. The nucleosomes are then tied together, like beads on a string (Figure 4.1), to form chromatin. Through chromatin, very long DNA molecules are packaged into a compact, dense shape and fit into a cell nucleus. The dense shape of the chromatin has other important roles, such as keeping DNA strands from becoming tangled, preventing DNA damage, regulating gene expression and DNA replication during cell division [205].

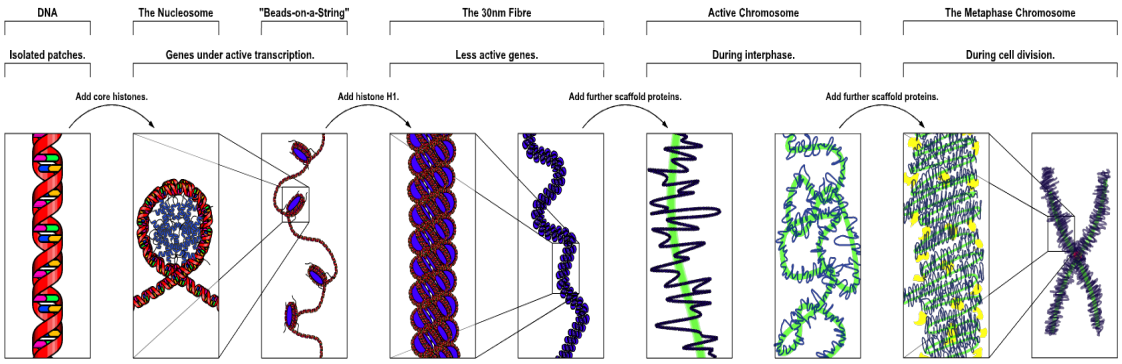


Figure 4.1: Chromatin structure. (Figure by Richard Wheeler at en.wikipedia.org [CC BY-SA 3.0)])

Although this process of DNA packaging into the cell nucleus via chromatin formation is a remarkable way of solving the topological challenge by each human cell, there are cases when the process gets defected. For instance, changes in DNA, histone modifications or distorted nucleosome remodelling (more details are discussed in the following section), often referred to as chromatin aberrations, can be involved in tumorigenesis [206].

On the other hand, it has been shown that mutations in the somatic cells are not distributed uniformly across the human genome but rather their distributions vary (up to fivefold) depending on specific genomic regions [207]. Several studies proposed that epigenomic organization of genomic regions are the main sources of how cancer somatic mutational landscape is formed [208, 209, 210, 211]. In [176] they compared the distribution of mutations from multiple samples of different cancer types to cell-type-specific epigenomic characteristics and found that chromatin accessibility and modification, as well as replication timing, explain over 80% of the variation in mutations rates depending on different cancer genomic regions. This study showed how the epigenetic configuration of the genome influences the accumulation of mutations due to different efficiency of mismatch repair genes that act in the presence or absence of chromatin. They labelled genome regions as open or

close depending on when the chromatin was unfolded or not, and found that there were distinct patterns of mutation accumulation in open vs closed chromatin regions. Figure 4.2 is borrowed from the paper [176] and shows the density of C to T mutations in melanoma alongside a 100-kb window profile of melanocyte chromatin accessibility. We can see how closed chromatin regions have a higher mutational load compared to open chromatin regions. This might be due to the activity of different epigenetic repair mechanisms that undo or prevent genome from mutations in open chromatin regions.

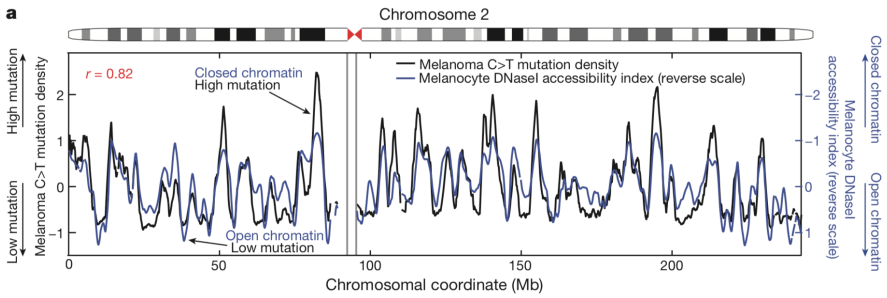


Figure 4.2: The example of the associations found between chromatin structure and the corresponding mutational load. The density of C to T mutations in melanoma alongside a 100-kb window profile of melanocyte chromatin accessibility ('DNase I accessibility index'; shown in normalized, reverse scale; high values correspond to less accessible chromatin and vice versa). Closed chromatin regions have a higher mutational load compared to open chromatin regions. This might be due to the activity of different epigenetic repair mechanisms that undo or prevent genome from mutations in open chromatin regions (Figure source [176])

The observation above suggests different mutation rates in open vs closed chromatin, which enables us to time chromatin somatic changes. Hence, we decided to develop a mathematical model that would infer times of chromatin aberration events. Specifically, when different chromatin aberration processes lead to disordered closing and opening of the chromatin regions. We tried to estimate the times when a given region was supposed to be open but remained closed or vice versa due to abnormal chromatin behaviour.

4.2 Chromatin organization and remodeling

As introduced above, chromatin is a complex of nucleic acids and proteins, which condenses to form chromosomes during eukaryotic cell division. Some of its primary functions are: first, packaging DNA into a smaller volume so that it fits in the cell and is prevented from damage; and second, regulating gene expression and DNA replication. Nucleosome (a complex of histone proteins) remodelling - the process of chromatin structure modification that enables gene expression - plays a crucial role in gene regulation. These remodelling processes (also called chromatin opening) are performed by biochemical modifications of histones (methylation, acetylation, phosphorylation) that alter the chromatin structure so that it is readily available for DNA transcription factors. The nucleosome bound chromatin regions before the remodelling takes place, are referred to as the closed chromatin regions. Different epigenetic aberrations cause disordered opening/closing of chromatin areas and hence affect gene expression, which results in unbalanced DNA accessibility, and later on in cancer formation and progression [212].

4.3 ATAC-seq data

ATAC-seq (Assay for Transposase-Accessible Chromatin with high-throughput sequencing) is a method for determining chromatin accessibility across the genome. Its main elements are hyperactive mutant transposases that are used to probe open chromatins by inserting their sequencing adapters into open regions of the genome. Transposases are enzymes catalysing the movement of transposons to other parts in the genome [213]. Figure 4.3, borrowed from wikipedia, shows a nice illustration of the method workflow; at the top, we can see the tightly packed, transcriptionally inactive (i.e. closed) chromatin and loosely packed, transcriptionally active (i.e. open)

chromatin regions. The mutant Tn5 transposase cuts out sufficiently long DNA via tagmentation which is a process of simultaneous fragmentation and tagging of the accessible DNA in the open chromatin regions. The tagged DNA fragments are then purified and amplified by PCR. Finally, they are sent for sequencing and ATAC-seq peaks corresponding to open chromatin regions are identified.

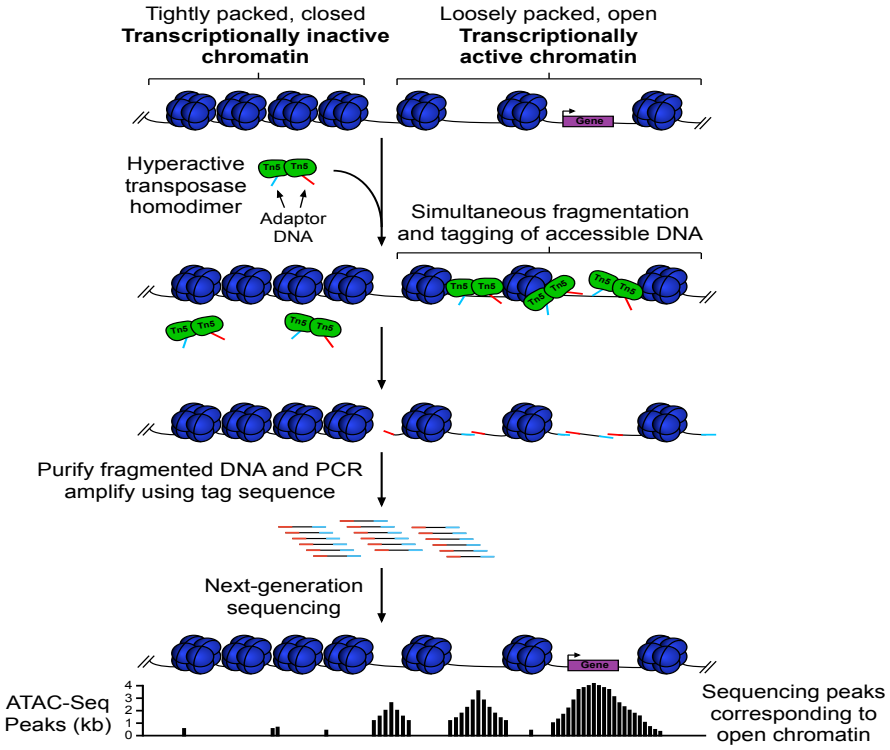


Figure 4.3: ATAC-seq workflow. (Figure source: cross-wiki upload from en.wikipedia.org [CC BY-SA 4.0])

4.4 Statistical modelling

Without loss of generality, we can set the time interval to when a given chromatin region is open and after which it becomes closed, to be (0, T). Assuming chromatin aberration time points can be normalised to be within the (0,1) interval, we developed a maximum likelihood estimation (MLE) framework that infers time points to each chromatin region aberration event based on the mutational burden measured per region. In the following sections, I describe the details of the model derivation, followed by data analysis and model application steps to real and synthetic datasets.

4.4.1 The MLE-based model derivation

Let us say we have R chromatin regions: $l_1, ..., l_R$. Out of these R regions, say N were observed to be undefined, and denote them as: $u_1, ..., u_N$. That leaves $R-N$ remaining regions to be either open or close. Say we count number of mutations per region; so we have a sequence of R mutation counts: $m_1, ..., m_R$. Each m_i is assumed to be generated by a corresponding Poisson distribution with some λ_i rate. We define three different sets of λ_i-s corresponding to the three types of chromatin regions:

$$\begin{aligned}
\lambda_o &= l_i(T-t_0)\mu_o, \ \ i \in Open \\
\lambda_c &= l_i(T-t_0)\mu_c, \ \ i \in Closed \\
\lambda_{u_i} &= l_i((\theta_i - t_0)\mu_o + (T-\theta_i)\mu_c), \ \ i = \overline{1,N}
\end{aligned} \tag{4.1}$$

where θ_i denotes the proportion of the time before an abberation occured in the i-th interval. So, each λ is a product of the length of the corresponding region - l_i, the mutation rate - either μ_o or μ_c assuming there are only two different mutation rates

corresponding to open vs close regions, and the time interval of the observation - $T - t_0$.

Now let us derive the likelihood function for seeing the data given the set of parameters we defined. Our data are the number of mutations per chromatin region. The probability of each mutation count can be calculated using the Poisson distribution function: $P(m|\lambda) = \frac{\lambda^m \exp(-\lambda)}{m!}$. The likelihood function for observing each m_i data point, then will simply be the product of each individual point's probability, assuming they are being generated independently. Let's denote the time points corresponding to the undefined chromatin regions (that we aim to infer) as θ_i-s, where $i = 1, ..., N$. Then the likelihood function will be the following:

$$\begin{aligned}\mathscr{L}(m_1, ..., m_R|\theta_i, ..., \theta_N) = &\prod_{i \in Open} \exp(-\lambda_o)\frac{\lambda_o^{m_i}}{m_i!} \times \\ &\prod_{i \in Closed} \exp(-\lambda_c)\frac{\lambda_c^{m_i}}{m_i!} \times \\ &\prod_{i \in Undefined} \exp(-\lambda_{u_i})\frac{\lambda_{u_i}^{m_i}}{m_i!} = \\ &\prod_{i \in Open} \exp(-l_i(T - t_0)\mu_o)\frac{(l_i\mu_o)^{m_i}}{m_i!} \times \\ &\prod_{i \in Closed} \exp(-l_i(T - t_0)\mu_c)\frac{(l_i\mu_c)^{m_i}}{m_i!} \times \\ &\prod_{i=1}^{N} \exp(-l_i((\theta_i - t_0)\mu_o + (T - \theta_i)\mu_c))\frac{(l_i((\theta_i - t_0)\mu_o + (T - \theta_i)\mu_c))^{m_i}}{m_i!}\end{aligned} \tag{4.2}$$

Without loss of generality, we can set $t_0 = 0$ and $T = 1$, and the likelihood function will take the following simpler form:

$$\mathscr{L}(m_1,...,m_R|\theta_i,...,\theta_N) = \prod_{i\in Open} \exp(-l_i\mu_o)\frac{(l_i\mu_o)^{m_i}}{m_i!}\times \prod_{i\in Closed} \exp(-l_i\mu_c)\frac{(l_i\mu_c)^{m_i}}{m_i!}\times \prod_{i=1}^{N}\exp(-l_i(\theta_i\mu_o+(1-\theta_i)\mu_c))\frac{(l_i(\theta_i\mu_o+(1-\theta_i)\mu_c))^{m_i}}{m_i!} \tag{4.3}$$

Now we need to find the parameter values that will maximise this likelihood function. The is an optimisation problem. To solve it, we need to first differentiate the likelihood function with respect to the parameters, set the obtained expression to zero, and then solve for the parameters. The function is not easy to differentiate, so we do the common trick by taking the natural logarithm of the expression and differentiating it instead of the actual likelihood function. This trick is accepted and works because the natural logarithm is a monotonically increasing function. This ensures that the maximum value of the log of the function will occur at the same value as for the original likelihood function. Hence, our likelihood function can be further simplified by converting it to the log-likelihood function:

$$L(\overrightarrow{m}|\overrightarrow{\theta}) = \sum_{i\in Open}(-l_i\mu_o+m_i\log(l_i\mu_o)-\log(m_i!))+ \sum_{i\in Open}(-l_i\mu_c+m_i\log(l_i\mu_c)-\log(m_i!))+ \sum_{i=1}^{N} l_i(\theta_i\mu_o+(1-\theta_i)\mu_c)+m_i\log(l_i(\theta_i\mu_o+(1-\theta_i)\mu_c)-\log(m_i!))+ \sum_{i=1}^{N}\gamma_i(1-\theta_i)+\sum_{i=1}^{N}\gamma_{N+i}\theta_i \tag{4.4}$$

The expression above is also called the Lagrangian given we have the following constraint for our optimisation problem: $0 \leq \theta_i \leq 1,\ i = 1,...,N$ and we added the last term with some constants γ which is called the Lagrangian multiplier. The first derivative of the log-likelihood function with respect to each parameter θ_i will then be the following:

$$\frac{\partial L}{\partial \theta_i} = -l_i(\mu_o - \mu_c) + \frac{m_i l_i(\mu_o - \mu_c)}{l_i(\theta_i \mu_o + (1-\theta_i)\mu_c)} - \gamma_i + \gamma_{N+i} \tag{4.5}$$

Then we apply the following Karush-Kuhn-Tucker (KKT) optimisation conditions (KKT conditions extend the method of Lagrange Multipliers when optimisation problems have constraints given by inequalities rather than equalities [214]): $\frac{\delta L}{\delta \theta_i} = 0$, $\gamma_i(1-\theta_i) = 0$, $\gamma_{N+i}\theta_i = 0$, $0 \leq \theta_i \leq 1$ and solve for θ_i.
Let's break down the conditions; consider first $\gamma_i(1-\theta_i) = 0$, if:

- $(1-\theta_i) = 0 \rightarrow$ we store $\hat{\theta}_i = 1$
- $\gamma_i = 0 \rightarrow$ we solve the equation (4) for θ, that gives $\hat{\theta}_i = \frac{m_i}{l_i(\mu_o - \mu_c) - \gamma_{N+i}} - \frac{\mu_c}{\mu_o - \mu_c}$;
  Now within the current condition, let's consider the next condition: $\gamma_{N+i}\theta_i = 0$, here if:
  - $\theta_i = 0 \rightarrow$ we store $\hat{\theta}_i = 0$
  - $\gamma_{N+i} = 0 \rightarrow$ we store $\hat{\theta}_i = \frac{m_i - l_i\mu_c}{l_i(\mu_o - \mu_c)}$

Hence, the solution workflow above leads us to the following three possible values for $\hat{\theta}$: $(1, 0, \frac{m_i - l_i\mu_c}{l_i(\mu_o - \mu_c)})$. The Maximum Likelihood Estimate of $\hat{\theta}$ will be the one of these derived values that maximizes the likelihood function given by (4.3).

4.5 Data analysis

4.5.1 Synthetic data analysis

Before applying our model to real data, we decided to first test it on a synthetically generated dataset. For our model, we need to have labelled chromatin regions. The labels are open, closed and the undefined ones. For region labelling usually ATAC-seq data is used, but for our simulations, we will randomly split and label the regions of a chromosome. Then, by assumption, we know that there should be a higher mutational burden in closed chromatin regions compared to open regions. Hence, we generate a random number of mutations per synthetic region from two different Poisson distributions; with a higher rate for closed regions and lower for open. Then there will be also regions that due to different chromatin aberration events became closed when they were supposed to be open and vice versa. We call such regions undefined and their mutational load will be between the high and low peaks of closed and open chromatin regions, respectively. So, within a defined time interval: (t_0, T) the pattern of mutation accumulation per chromatin region would look like the one presented in Figure 4.4.

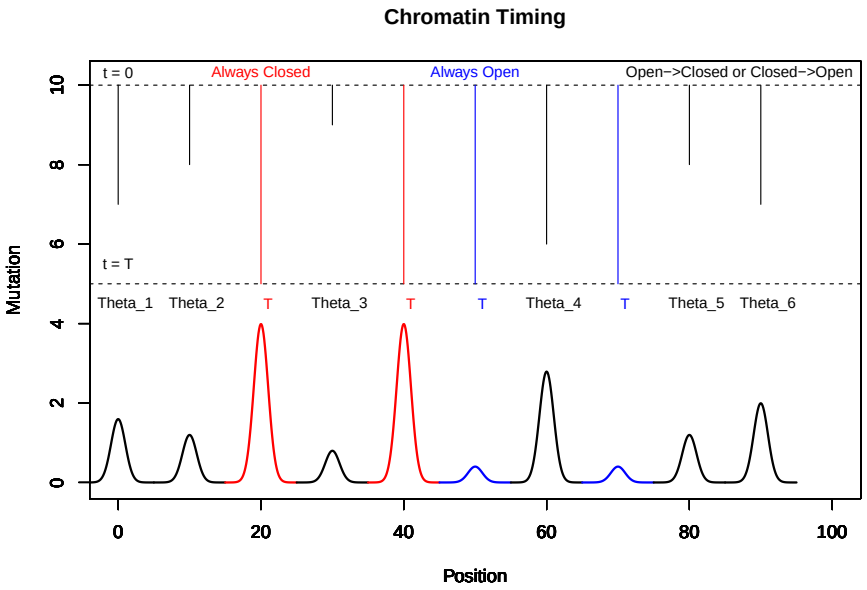


Figure 4.4: Chromatin timing model illustration. An illustration of chromatin region aberrations and their impact on mutation accumulation pattern per region. During a given time interval $(0, T)$ some regions were closed (low blue peaks) and hence accumulated a high number of mutations (high blue bars), and regions that were open (tall red peaks) and accumulated low number of mutations (low red bars). Due to different chromatin aberration events, some regions that were initially open became closed and vice versa, and we label them as undefined chromatin regions (presented here in black).

The synthetic data, to apply the model on, would thus consist of a number of accumulated mutations per region. As we already mentioned, we used two different rates for the Poisson distribution to generate mutations accumulating in open and closed chromatin regions. We aim to infer the times of those undefined regions when they initially were open and then became closed and vice versa. For this, we used a set of random rates for the Poisson distribution to generate the number of mutations corresponding to each undefined region individually. We then stored these randomly generated rates and try to infer them back using our model introduced in the previous section. We note these times by θ-s as in the model introduction. Figure 4.5 shows the results of inferred vs original θ values. We can see that the

model predictions are very close to the actual values with $R^2 = 0.98$.

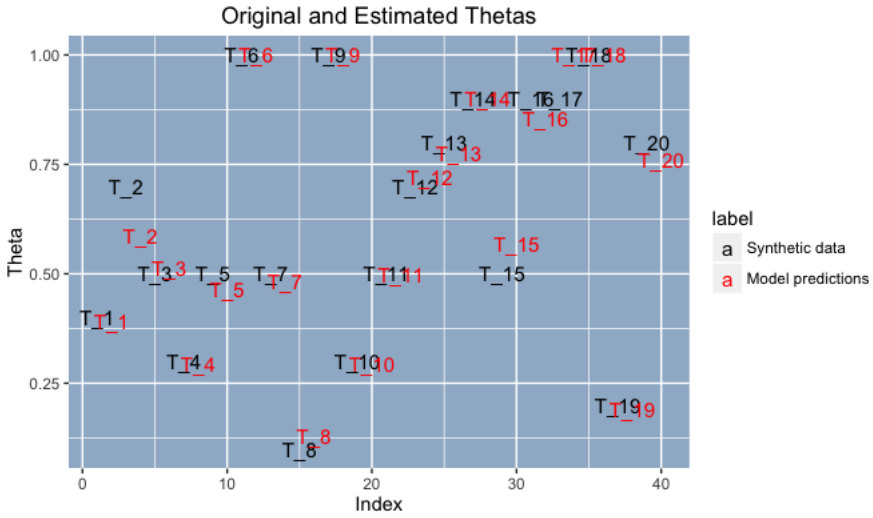


Figure 4.5: Chromatin timing - synthetic data vs model predictions. The figure shows the results of inferred vs original θ values. The model predictions are very close to the actual values.

4.5.2 WGS and ATAC-seq data analysis

In the previous sections, we described our model and its performance on a synthetic dataset. Here, we will show the analysis of an in-house generated real dataset and its limitations for model applicability.

The data consist of paired WGS and ATAC-seq multiple samples per patient. First, we tried to reproduce Figure 4.2 from [176] with our datasets. As in the figure, we slide 5Mb window on chromosome 2 and count the number of truncal (shared across all samples per patient) mutations (SNVs only) and sum-up ATAC-seq pileups per window. Then we plot these measurements on the same graph to compare. We analysed 3 colon cancer cases (one MSI and two MSS) with 5 samples per patient. Figure 4.6 shows the results. Unfortunately, our data do not look like the one from Polak et al. [176]. That is, we do not see the similar trend of a strong reverse association between the mutational load and chromatin accessibility index

per region.

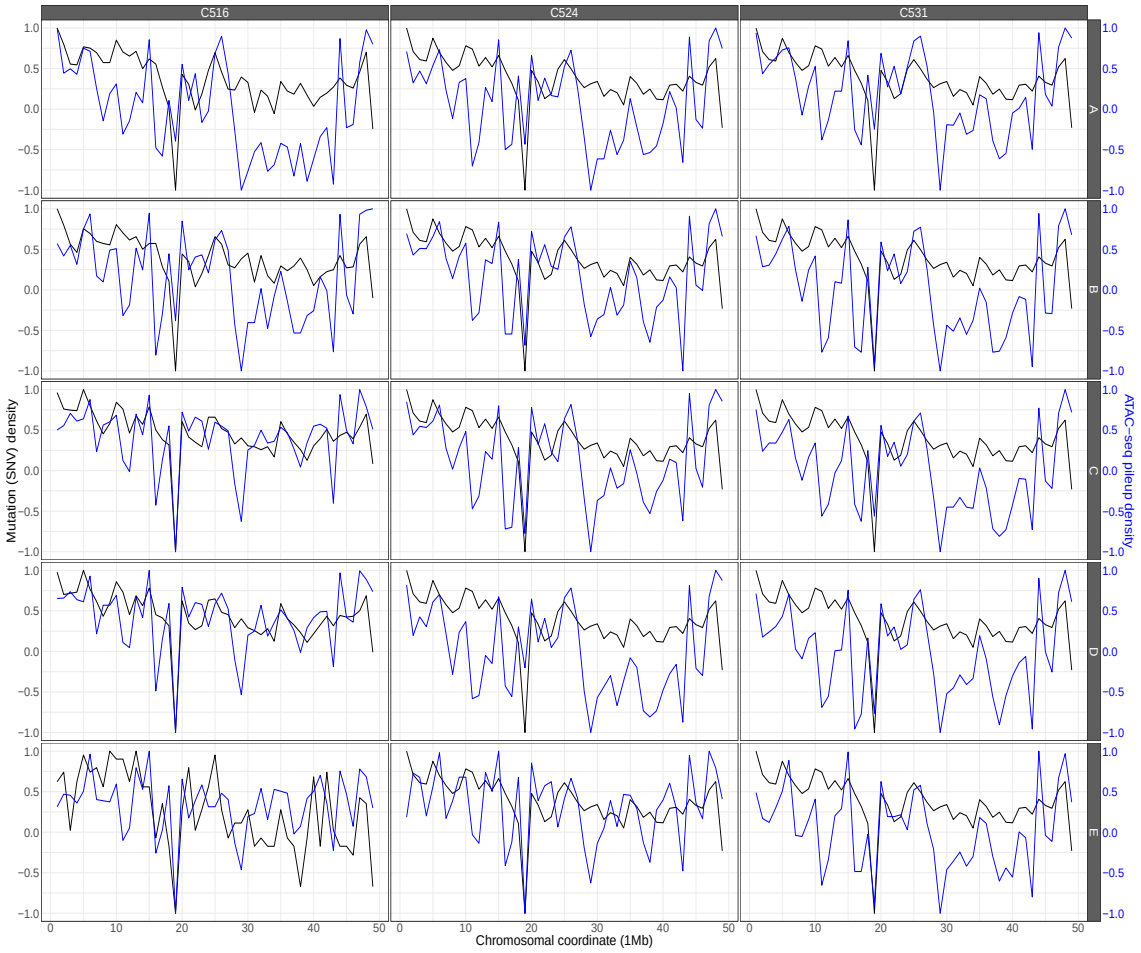


Figure 4.6: Mutational load vs ATAC-seq pileup density The figure is generated by sliding 5Mb window on chromosome 2 and counting the number of truncal (shared across all samples per patient) mutations (SNVs only) and summing up ATAC-seq pileups per window. We do not see the similar trend of a reverse association between the mutational load and chromatin accessibility index as shown in Figure 4.2 by Polak et al. in [176].

They report a significant negative correlation between the mutations per megabase and the density of chromatin accessibility index. We tested the correlations between the binned measurements and did not find any strong associations (Figure 4.7 shows the scatterplots of the measurements per sample per patient). We also checked if different window sizes would change the results; we repeated the analysis for 2Mb

and 1Mb windows, after which the obtained correlations increased but only very slightly (we present these results in the Appendix).

As such, since we could not reproduce the associations found in Polak et al. with the data we have, we could not use our model on data at current resolution. The negative association between chromatin regions (open vs closed) and the corresponding mutational load is the main assumption that we build our model on.

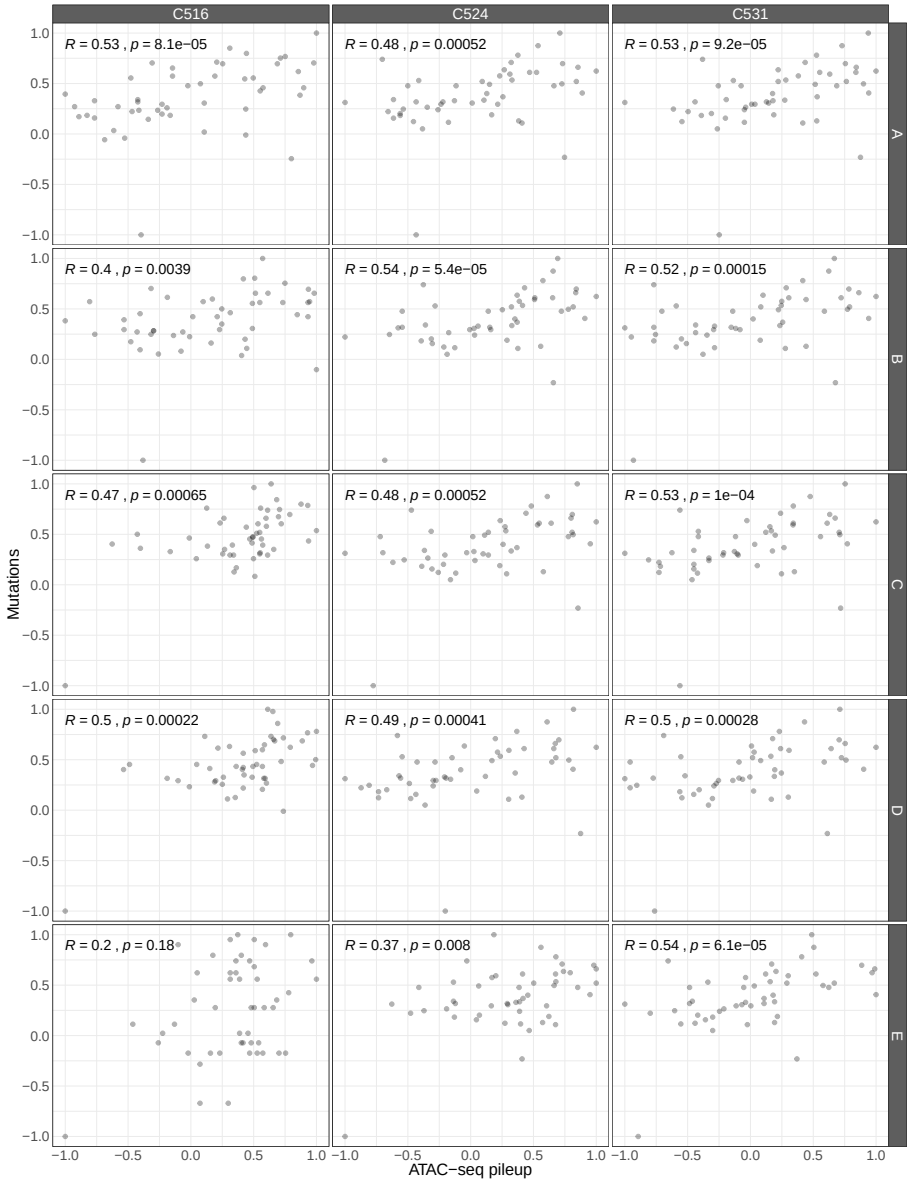


Figure 4.7: Mutational load vs ATAC-seq pileup scatterplots The figure is generated by sliding 5Mb window on chromosome 2 and counting the number of truncal (shared across all samples per patient) mutations (SNVs only) and summing up ATAC-seq pileups per window. There are significant but not strong associations found between the binned measurements that were tested using the Pearson corrolation test (the effect size and p-values of each test are reported on each scatterplot).

4.6 Discussion

We know that not only genetic, but epigenetic aberrations play an important role in tumorigenesis. The work presented in this chapter has been motivated by the studies that showed that mutations in the somatic cells are not distributed uniformly across the human genome but rather their distributions vary depending on specific genomic regions. Specifically, it has been observed that the epigenetic configuration of the genome influences the accumulation of mutations due to different efficiency of mismatch repair genes that act in the presence or absence of chromatin. In the study [176], they labelled genome regions as open or close depending on when the chromatin was unfolded or not, and found that there were distinct patterns of mutation accumulation in open vs closed chromatin regions.

We decided to develop a model that would infer the timing of chromatin aberration events based on the observation above. That is, if we had chromatin regions with the number of accumulated mutations per region, we would expect to see the following pattern: high number of mutations in closed chromatin region and low in open. In addition, due to different aberrations, there would be regions that initially were open and became closed or vice versa, and thus the number of mutations would fall between expected high (for close) and expected low (for open) numbers. We would label these regions as undefined. Based on having such a pattern of the mutation accumulation, we developed a Maximum Likelihood Estimation model for the times when such chromatin aberration events would occur.

Initially, we applied our model to a synthetically generated dataset. That is, we generated an expected pattern of mutation accumulation per chromatin region – open, closed or undefined, using Poisson distribution. We stored the artificial times of chromatin aberrations and inferred them back using our model. From Figure 4.5 we can see the results of the inference – high R^2 values and inferred time points close to the original synthetic data points.

Then we also tried to apply our model to a real dataset that was generated in house and consists of paired ATAC-seq and WGS samples of colorectal cancer patients. Unfortunately, we could not reproduce the results found in [176] with the data we have and thus were unable to test our model on it. The main building block assumption of our model is that we would have the reversed pattern of the mutation accumulation per chromatin region as reported in the study.

We think the model will need to be further developed and adjusted after it is applied to a real dataset (that reproduces the results from [176]). Timing chromatin events is important to understand the evolutionary history of a tumour and estimate the order of somatic changes at the epigenetic level that give rise to a tumour. This has been done in the context of copy number alterations (e.g. [215]).

Chapter 5

Summary and outlook

In this thesis, we developed models and methods to study different patterns and characteristics of cancer evolution by combining approaches from cancer genetics and epigenetics. In Chapter 1, we give a brief introduction to the field, the necessary background information for the reader to be able to follow the concepts discussed throughout the thesis, and the basic tools and methods from statistical and computational modelling that we used to develop our models.

In Chapter 2, we focus on studying spatial effects when interpreting multi-region sequencing data to infer the tumour evolutionary dynamics. We found that the effects of sampling bias and spatial distributions of spatially inter-mixed cell populations critically depend on the mode of tumour growth as well as the details of the underlying sampling and data generation procedure. We could observe clusters of over-represented alleles in the VAF distribution of some tumour samples that were indistinguishable from positively selected subclonal populations, despite emerging solely due to the spatial distribution of cells. Such clusters vary depending on how one samples a tumour, and would, therefore, cause a major challenge for the evolutionary interpretation of cancer genomic data based on subclonal reconstruction.

Also, in Chapter 2, we present a Bayesian inference framework to recover evo-

lutionary parameters from our stochastic simulation model. We observed that our ability to precisely recover certain evolutionary parameters depends on the scenarios of tumour growth and spatial sampling strategies. We think, more involved statistical frameworks based on first principles of tumour growth can help to resolve some of the evolutionary parameters on an individualised patient basis. Importantly, careful spatial sampling and single-cell sequencing can mitigate some of the confounding issues. We acknowledge that our model has some important limitations, for example, the infinite allele assumption. Also, for computational feasibility, we mostly focused on 2D spatial analyses and of a relatively limited number of cells with respect to the billions of cells present in a human tumour. Furthermore, we do not offer a closed mathematical formulation for the distribution of alleles under spatial effects, which would be very useful but remains a very difficult problem that can only be tackled partially (e.g. [154]). Furthermore, for computational feasibility, especially in regards to the necessity of performing statistical inference on the data and generate thousands of simulations, we restricted our analysis to the stochastic cellular automaton model. We think our approach highlights the importance of spatial modelling of real data and the impact of confounding factor in our estimate and understanding of tumour evolution.

Future versions of the model could help to guide optimal sample collection that would minimise the spatial biases in the data. Due to the current technical limitations of these types of approaches, we are still far from direct application in the clinic. Additional effort should also be directed towards the use of measurements from other clinical data, such as imaging, where estimations of necrosis, for example, can help parameterise computational models. However, we argue it remains extremely important to understand the confounding factors and spatial biases we expect to find in samples from which often we need to base clinical decisions on.

In Chapter 3, we studied the associations between mutational signatures and the

epigenome. Here we hypothesize that different mutational processes (giving rise to distinct mutational signatures) are active in epigenetically different regions of the genome. To test our hypothesis, we used an epigenetic map of the regions such as promoters, enhancers, coding and non-coding DNA sections of breast cancer genome (derived by our collaborator's lab) and developed a method that tests the enrichment of mutational signatures in these different epigenetic regions. We found that there are different mutational processes (that give rise to different mutational signatures active in epigenetically distinct genomic regions.

Our main finding in Chapter 3 was the enrichment of the COSMIC mutational signature 2 at enhancers of ER-positive breast cancer patients. The proposed aetiology of the signature 2 is the activity of the AID/APOBEC family of cytidine deaminases. This led us to further hypothesise that there is a specific interplay of the oestrogen receptor (ER) and the APOBEC enzyme through which oestrogen drives the accumulation of APOBEC-induced mutations in these cancers. To validate the origin (APOBEC) of this mutational signature and its dependence on oestrogen exposure, Luca Magnani's lab has developed an in vitro system. They are maintaining a series of ER+ and ER- clones - with varying exposure to oestrogen - in a neutral evolutive regime and have been mapping the emerging mutations at a single base-pair resolution using whole-genome sequencing (WGS) at different time points. Analysis of contributing mutational signatures for each genomic section will reveal whether the APOBEC enzyme plays a role in the evolution of oestrogen-dependent cancers by specifically mutating regulatory regions.

In Chapter 4, we tried to develop a model that would infer times of chromatin aberration events. Here, our work was motivated by the studies that showed that mutations in somatic cells are not distributed uniformly across the human genome but rather their distributions vary depending on specific genomic regions. Specifically, it has been observed that the epigenetic configuration of the genome influences

the accumulation of mutations due to different efficiency of mismatch repair genes that act in the presence or absence of chromatin. Currently, we could not apply our model to a relevant real dataset, as the one we have does not satisfy the assumptions for the model. The performance of the model was satisfactory when applied to a synthetically generated dataset. The future direction of this part of the thesis would be first generating the data that would resemble the assumptions we based our model on, then further tune the model to make better predictions. Chromatin aberrations, such as changes in histone methylation, histone modifications or distorted nucleosome remodelling, have been observed to be one of the sources of tumorigenesis. Timing these events is important to understand the evolutionary history of a tumour and estimate the order of somatic changes at the epigenetic level as well.

Mathematical modelling of cancer evolution is a growing field with a fast-expanding repertoire of models and approaches. The attention to the clinical and biological relevance of modelling approaches is necessary to ensure these efforts do not result in dead ends. I believe this thesis shows at least partly the importance of coupling mathematical and computational modelling with experiments to gain a better understanding of cancer initiation and progression, and consequently achieve better clinical performance.

Appendix A

Linking mutational signatures to the epigenome

A.1 Jackknife resampling by regions

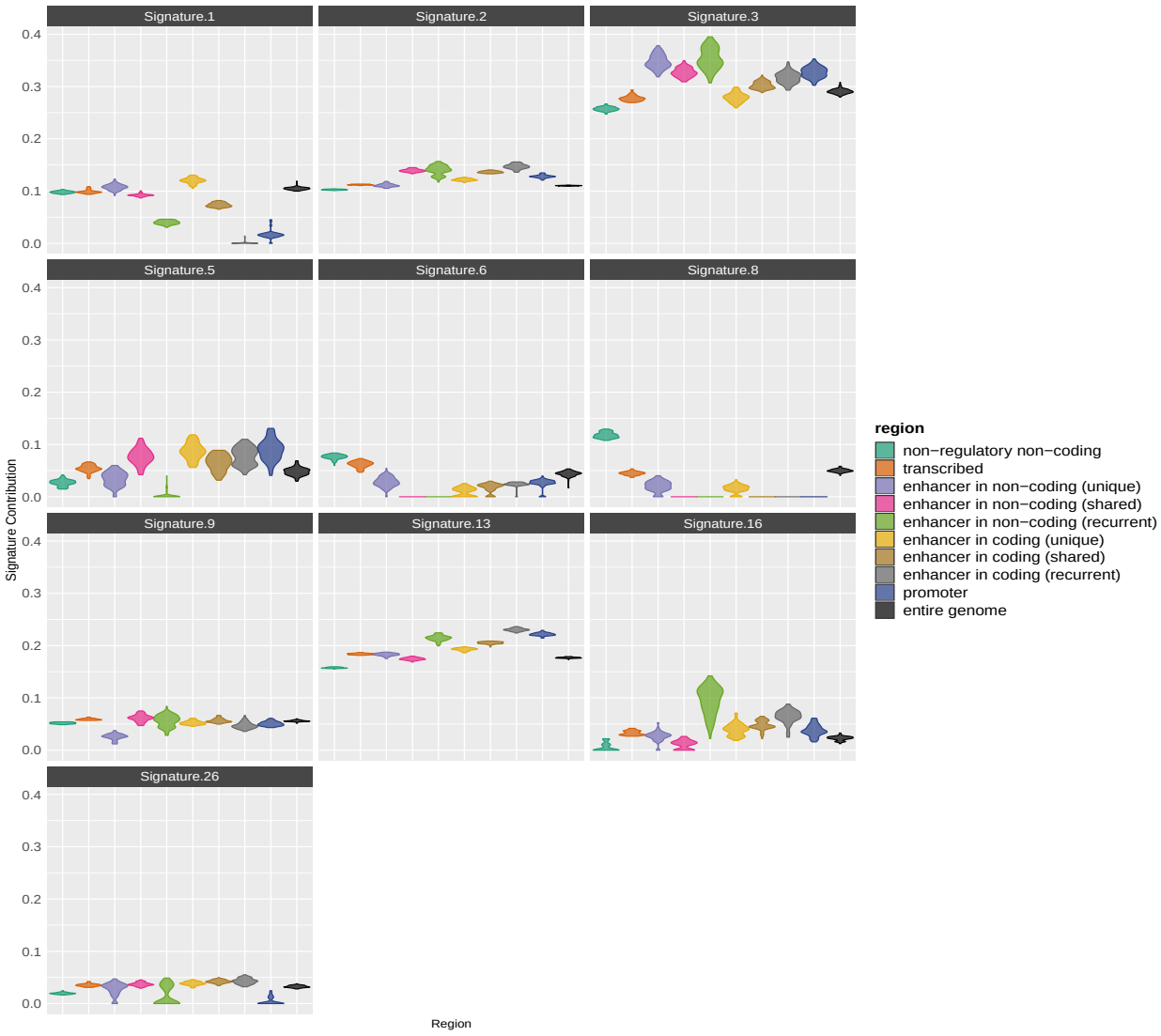


Figure A.1: Jackknife resampling by regions (BRCA-WT, ER-negative) - Nik-Zainal.

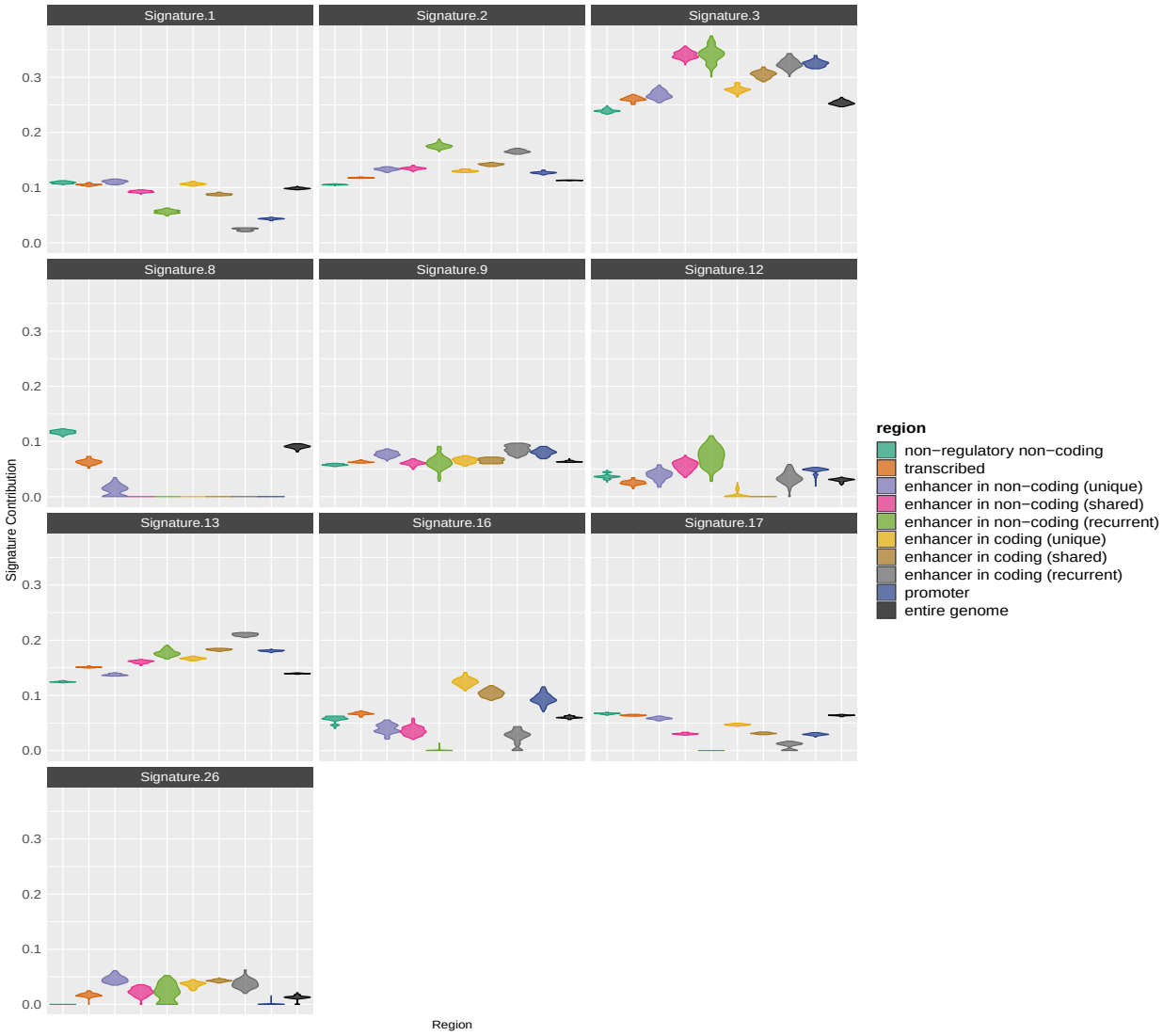


Figure A.2: Jackknife resampling by regions (ER-negative) - Hartwig.

A.2 Jackknife resampling by patients

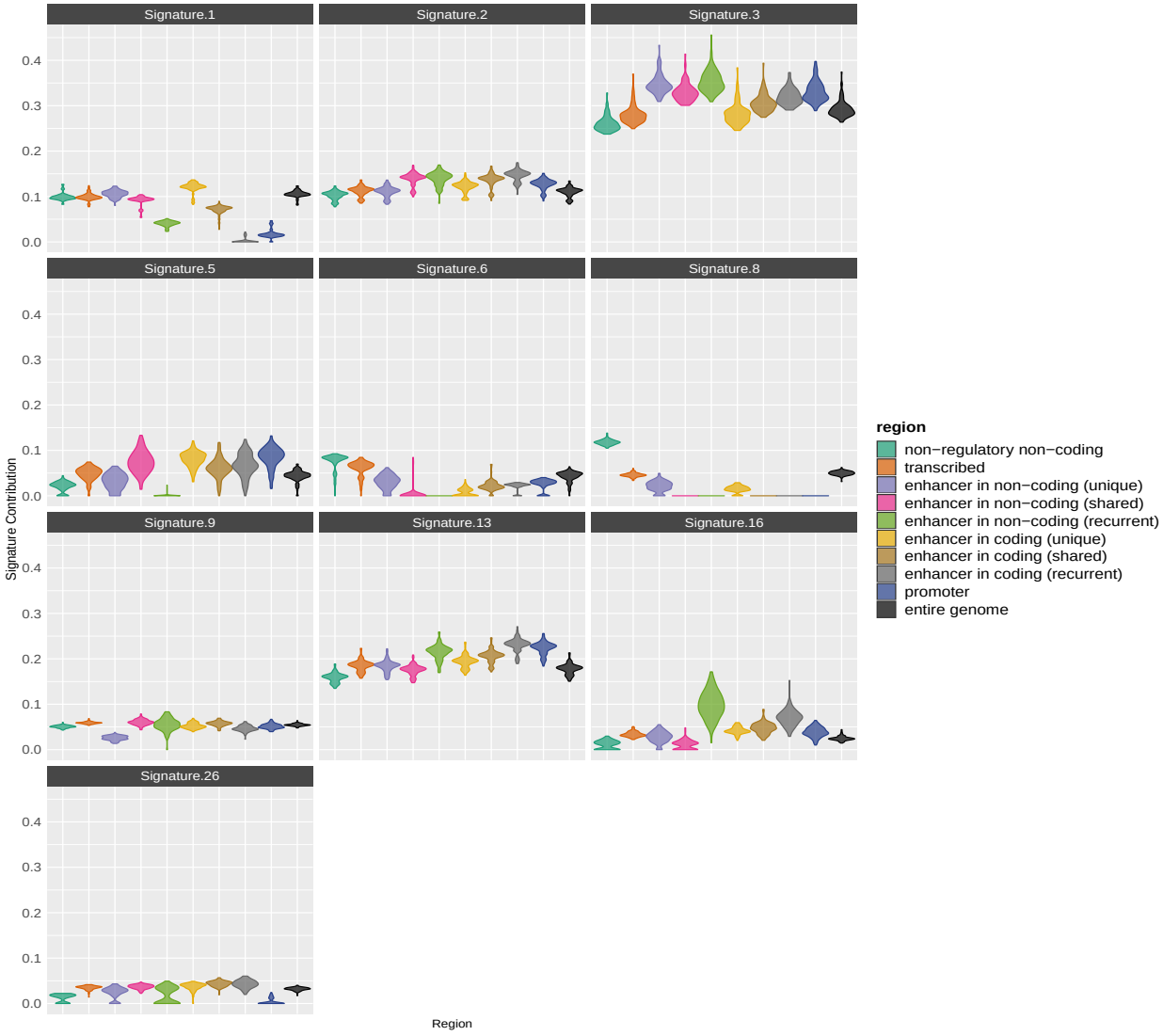


Figure A.3: Jackknife resampling by patients (BRCA-WT, ER-negative) - Nik-Zainal.

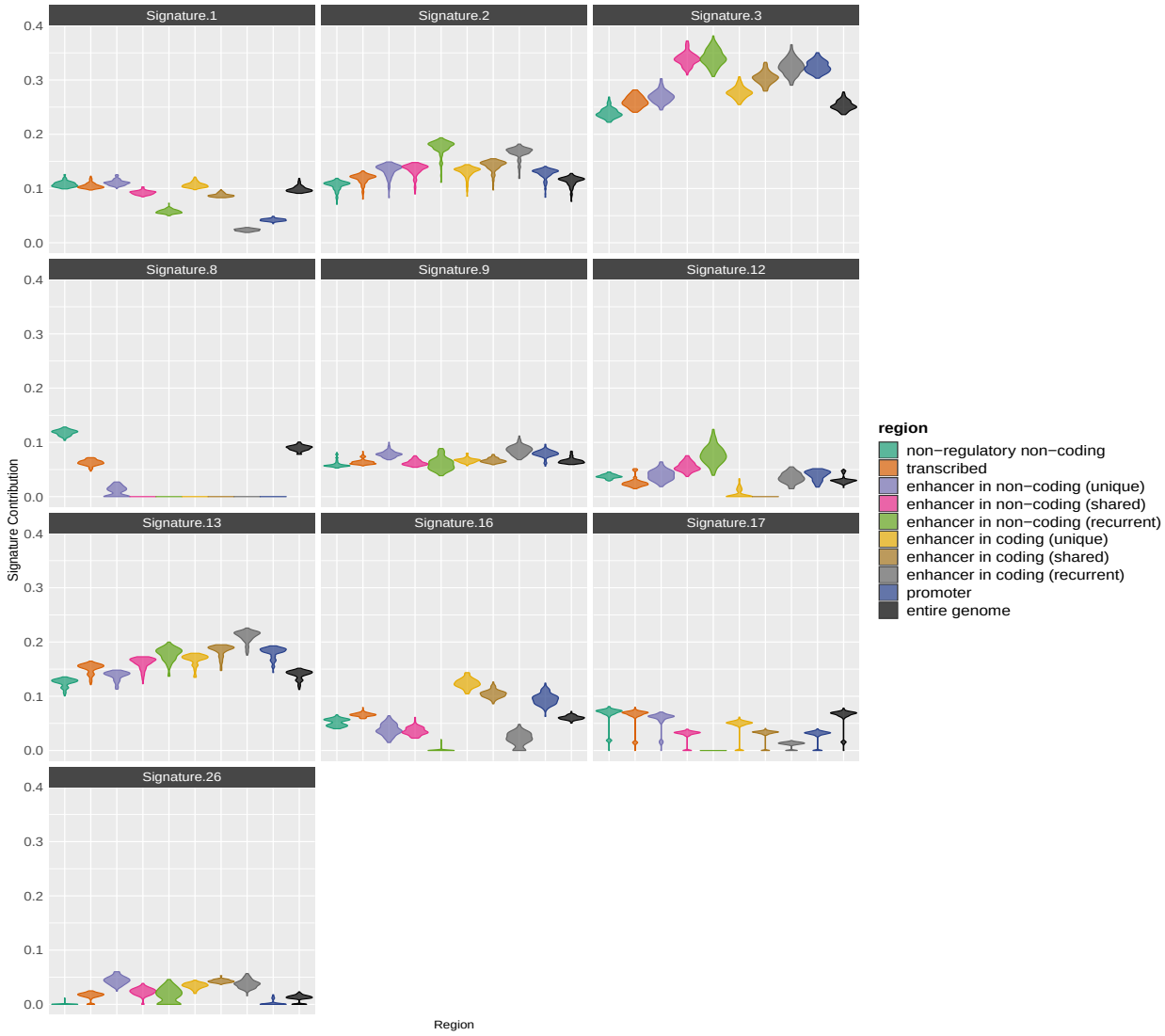


Figure A.4: Jackknife resampling by patients (ER-negative) - Hartwig.

A.3 Randomised annotations

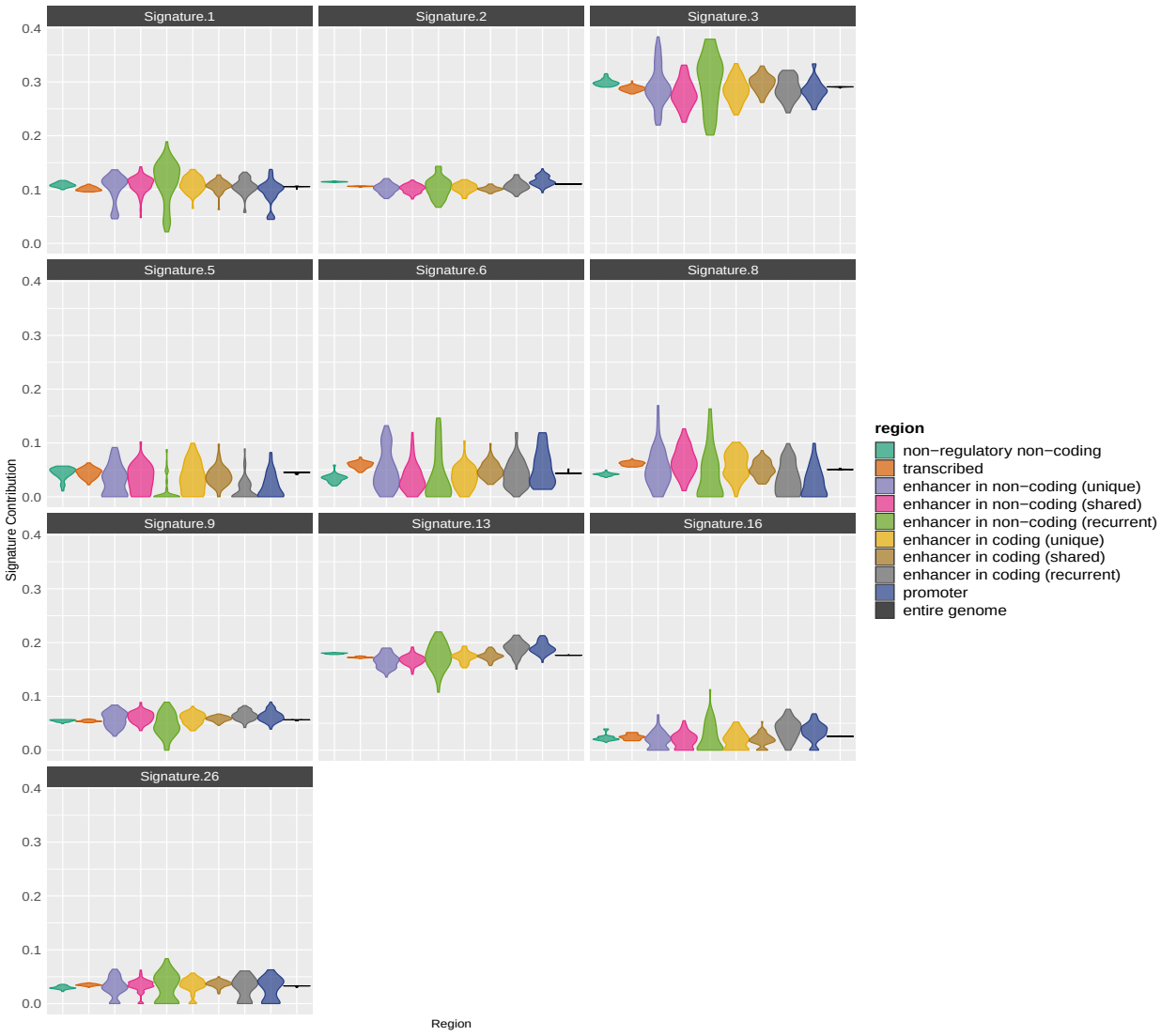


Figure A.5: Randomised annotations (BRCA-WT, ER-negative) - Nik-Zainal.

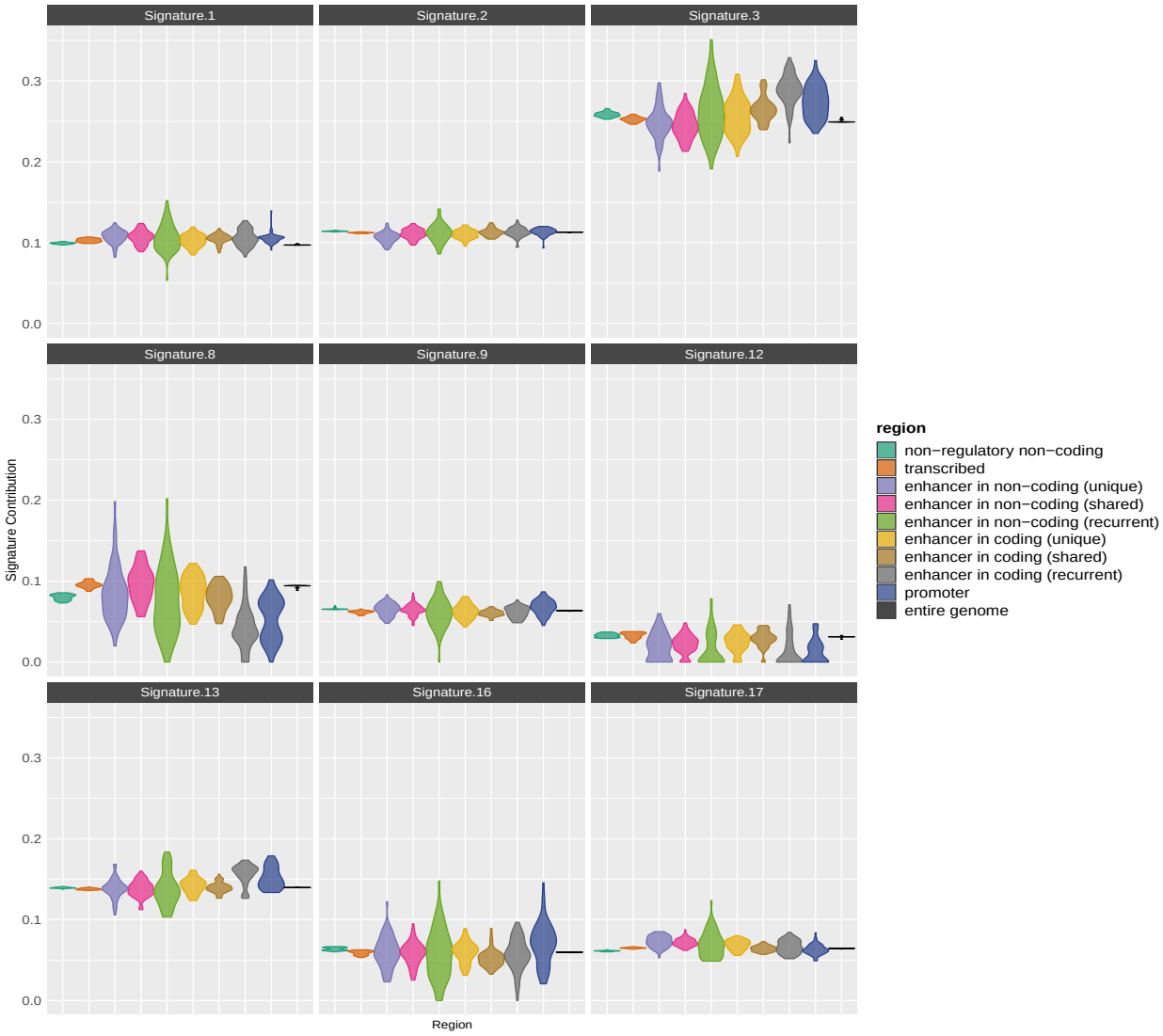


Figure A.6: Randomised annotations (ER-negative) - Hartwig.

Appendix B

Timing epigenetic changes

B.1 Mutational load vs ATAC-seq pileup

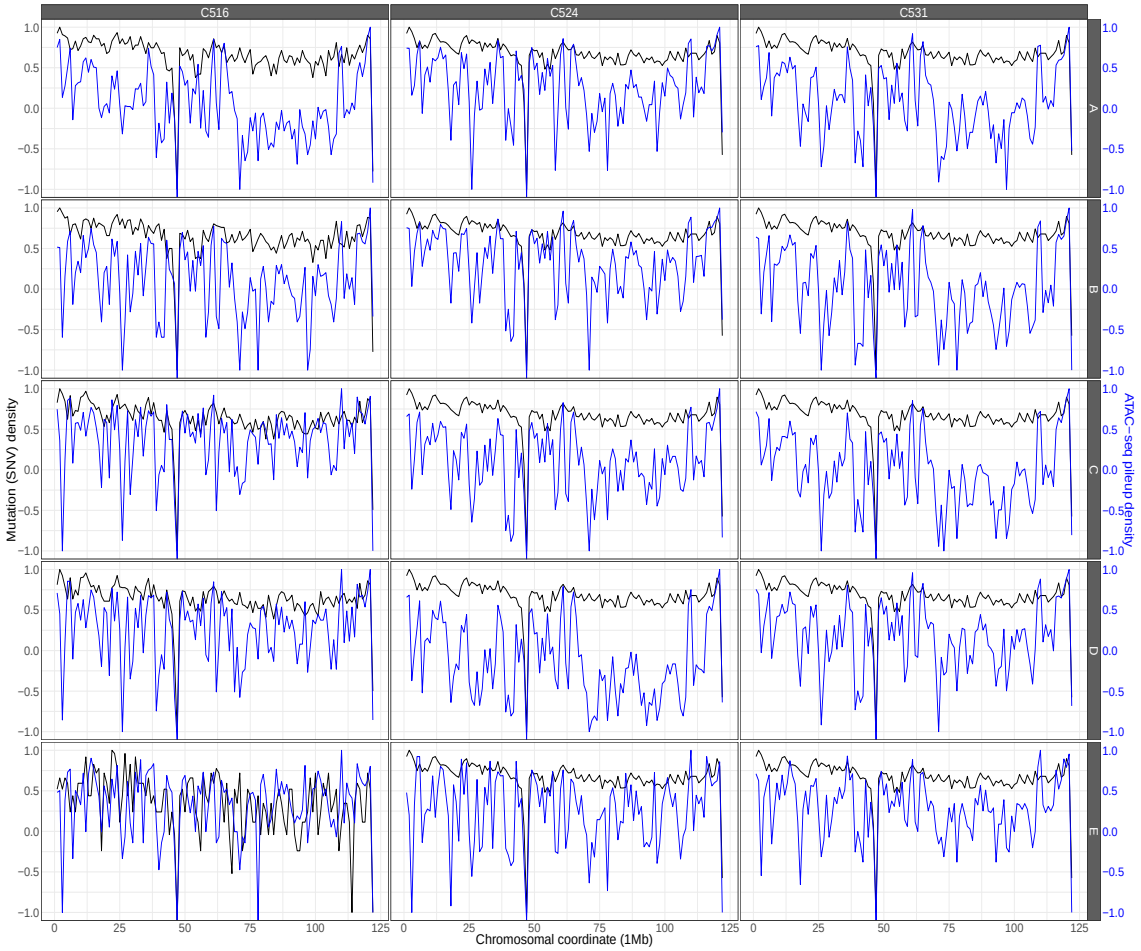


Figure B.1: Mutational load vs ATAC-seq pileup densities - 2Mb window

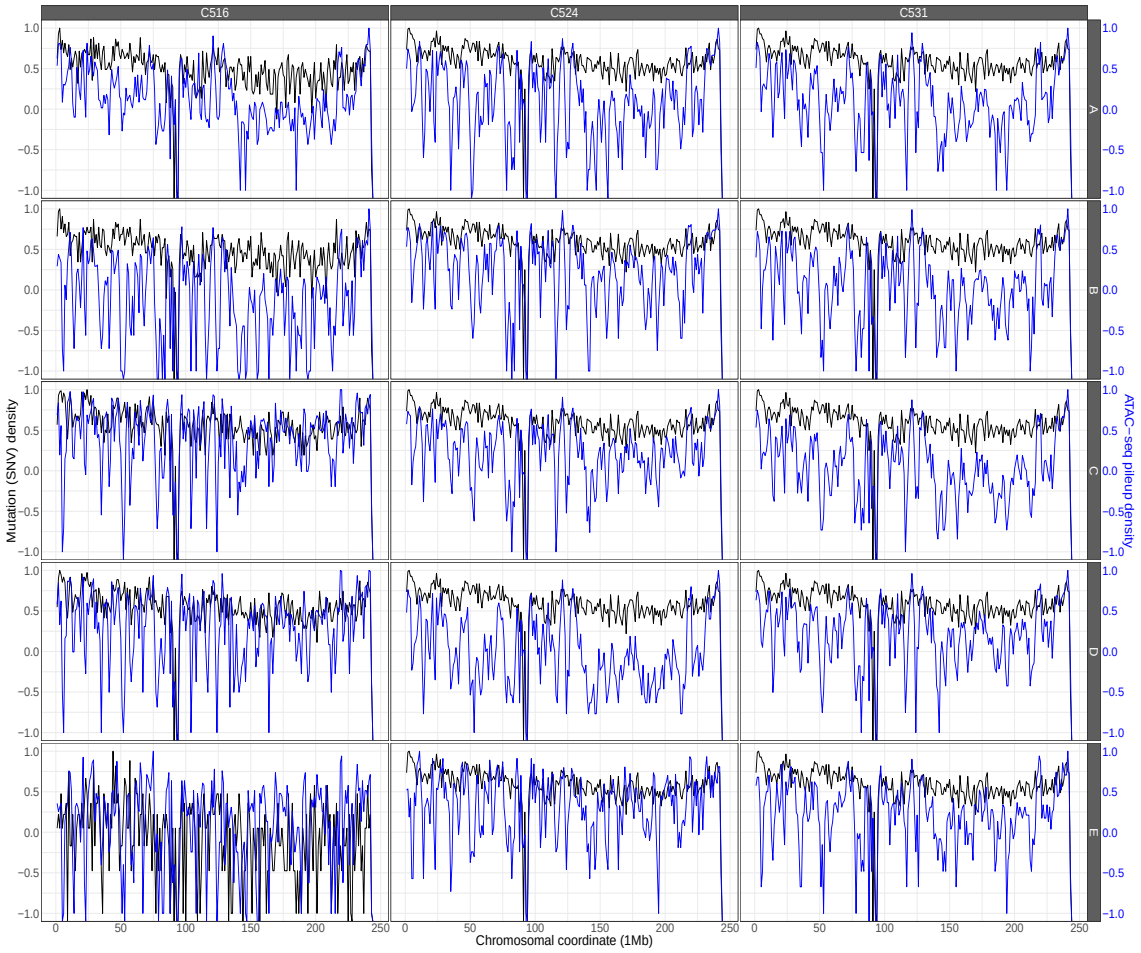


Figure B.2: Mutational load vs ATAC-seq pileup densities - 1Mb window

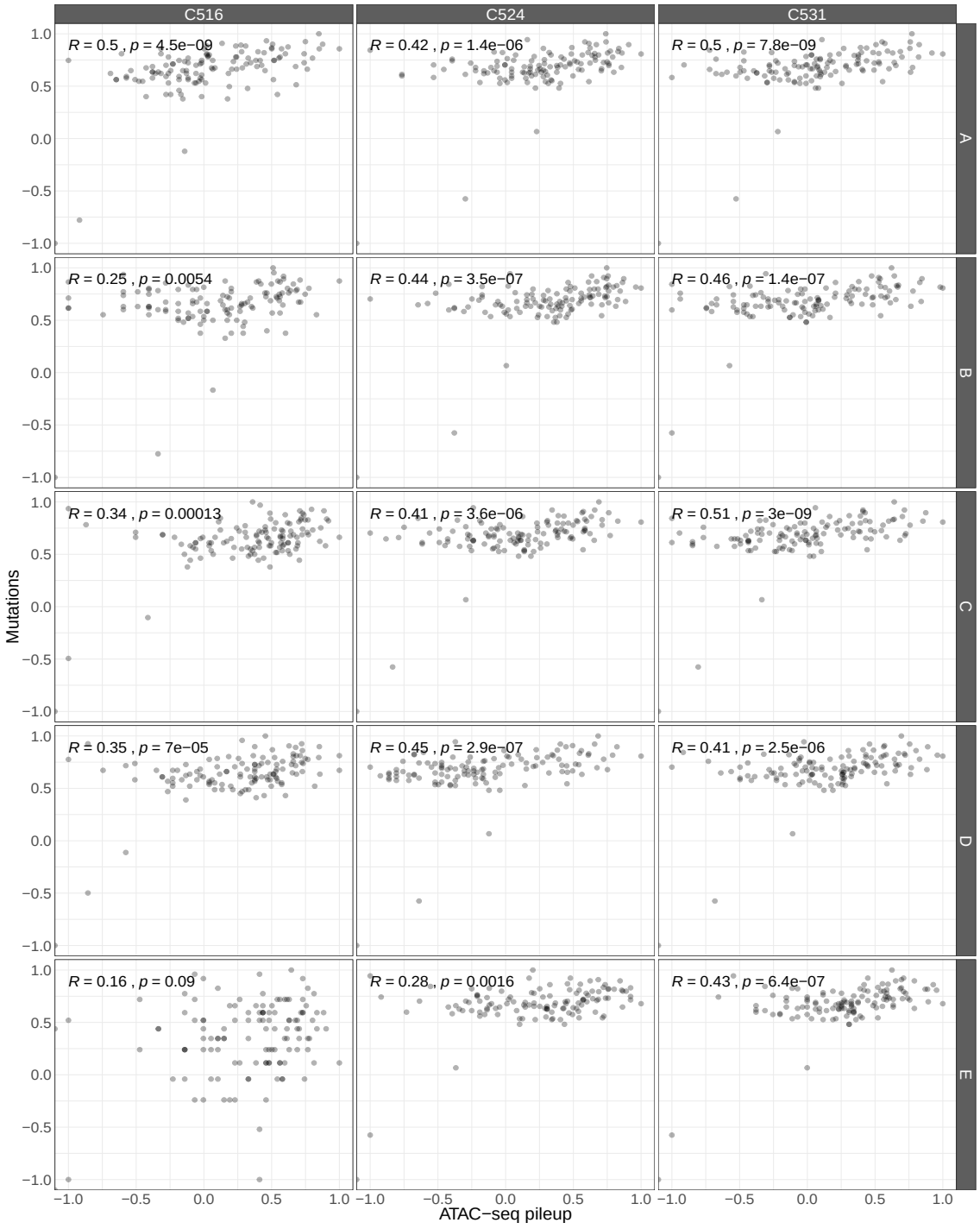


Figure B.3: Mutational load vs ATAC-seq pileup scatterplots - 2Mb window

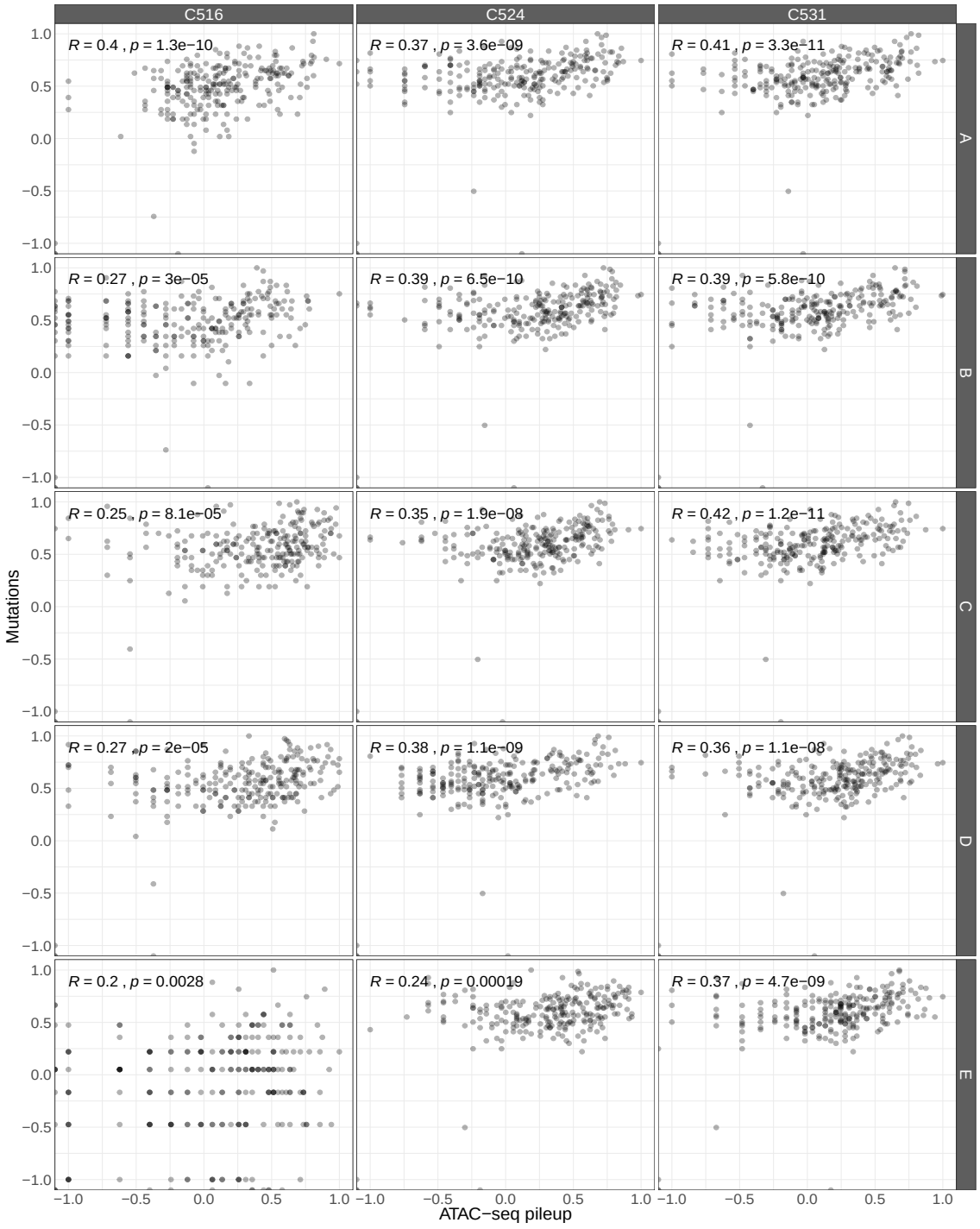


Figure B.4: Mutational load vs ATAC-seq pileup scatterplots - 1Mb window

Appendix C

Publications

Paper

The following paper has been peer-reviewed and was published as

Chkhaidze K., Heide T., Werner B., Williams M.J., Huang W., Caravagna G., Graham T.A., Sottoriva A. "Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data." *PLoS Comput Biol.* 15(7):e1007243 (2019).

Conference abstract

This conference abstract has been peer-reviewed and presented at AACR 2019.

Abstract 4232: Spatially constrained tumor growth affects the patterns of clonal selection and neutral drift in cancer genomic data.

DOI: 10.1158/1538-7445.AM2019-4232 Published July 2019

Proceedings: AACR Annual Meeting 2019; March 29-April 3, 2019; Atlanta, GA

Abstract

Quantification of the effect of spatial tumor sampling on the patterns of mutations detected in next-generation sequencing data is largely lacking. Here we use a spatial stochastic cellular automaton model of tumor growth that accounts for somatic mutations, selection, drift and spatial constrains, to simulate multi-region sequencing data derived from spatial sampling of a neoplasm. We show that the spatial structure of a solid cancer has a major impact on the detection of clonal selection and genetic drift from bulk sequencing data and single-cell sequencing data. Our results indicate that spatial constrains can introduce significant sampling biases when performing multi- region bulk sampling and that such bias becomes a major confounding factor for the measurement of the evolutionary dynamics of human tumors. We present a statistical inference framework that takes into account the spatial effects of a growing tumor and allows inferring the evolutionary dynamics from patient genomic data. Our analysis shows that measuring cancer evolution using next-generation sequencing while accounting for the numerous confounding factors requires a mechanistic model-based approach that captures the sources of noise in the data.

Citation Format: Kate Chkhaidze, Timon Heide, Benjamin Werner, Marc J. Williams, Weini Huang, Giulio Caravagna, Ann-Marie Baker, Trevor A. Graham, Andrea Sottoriva. Spatially constrained tumor growth affects the patterns of clonal selection and neutral drift in cancer genomic data [abstract]. In: Proceedings of the American Association for Cancer Research Annual Meeting 2019; 2019 Mar 29-Apr 3; Atlanta, GA. Philadelphia (PA): AACR; Cancer Res 2019;79(13 Suppl):Abstract nr 4232.

Bibliography

[1] H. zur Hausen. Papillomaviruses and cancer: from basic studies to clinical application. *Nat Rev Cancer.*, 2(5):342–350, 2002.

[2] D. A. Thorley-Lawson and A. Gross. Persistence of the epstein-barr virus and the origins of associated lymphomas. *N Engl J Med.*, 350(13):1328–1337, 2004.

[3] M.J. Allday. How does epstein-barr virus (ebv) complement the activation of myc in the pathogenesis of burkitt's lymphoma? *Semin Cancer Biol.*, 19(6):366–376, 2009.

[4] P. Armitage and R. Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer.*, 8(1):1–12, 1954.

[5] R.A. Weinberg. Tumor suppressor genes. *Science.*, 254(5035):1138–1146, 1991.

[6] M. A. Nowak. *Evolutionary dynamics: exploring the equations of life.* Harvard University Press, 2006.

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

[8] M. A. Dawson and T. Kouzarides. Cancer epigenetics: from mechanism to therapy. *Cell.*, 150(1):12–27, 2012.

[9] P. A. Jones and S.B. Baylin. The epigenomics of cancer. *Cell.*, 128(4):683–692, 2007.

[10] J. S. You and P. A. Jones. cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell.*, 22(1):9–20, 2012.

[11] M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature.*, 458(7239): 719–724, 2009.

[12] Z. Yang. *Computational molecular evolution.* Oxford University Press, 2006.

[13] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature.*, 171(4356):737–8., 1953.

[14] M. Kimura. Evolutionary rate at the molecular level. *Nature.*, 217(5129):624–6, 1968.

[15] S. Yi. Neutrality and molecular clocks. *Nature Education Knowledge*, 4(2):3, 2013.

[16] P. Lemey, M. Salemi, and A. M. Vandamme. *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing.* Cambridge University Press, 2009.

[17] M. Greaves and C. C. Maley. Clonal evolution in cancer. *Nature.*, 481:306–313, 2012.

[18] M. Greaves. Evolutionary determinants of cancer. *Cancer Discov.*, 5(8):806–820, 2015.

[19] S. Turajlic, A. Sottoriva, T. A. Graham, and C. Swanton. Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics*, 20:404–416, 2019.

[20] A. Sottoriva, C. P. Barnes, and T. A. Graham. Catch my drift? making sense of genomic intra-tumour heterogeneity. *Biochim Biophys Acta Rev Cancer.*, 1867(2):95–100, 2017.

[21] J. M. Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genet Res.*, 89(5-6): 391–403, 2009.

[22] I. Martincorena, K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, and P. J. Campbell. Universal patterns of selection in cancer and somatic tissues. *Cell.*, 171(5):1029–1041, 2017.

[23] L. Zapata, O. Pich, L. Serrano, F. A. Kondrashov, S. Ossowski, and M .H. Schaefer. Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biol.*, 19(1):1–67, 2018.

[24] T. A. Graham and A. Sottoriva. Measuring cancer evolution from the genome. *J Pathol.*, 241(2):183–191, 2016.

[25] C. C. Maley, P. C. Galipeau, J. C. Finley, V. J. Wongsurawat, X. Li, C. A. Sanchez, T. G. Paulson, P. L. Blount, R. A. Risques, P. S. Rabinovitch, and B. J. Reid. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet.*, 38(4):468–473, 2006.

[26] D. A. Landau, S. L. Carter, P. Stojanov, A. McKenna, K. Stevenson, M. S. Lawrence, C. Sougnez, C. Stewart, A. Sivachenko, L. Wang, Y. Wan, W. Zhang, S. A. Shukla, A. Vartanov, S. M. Fernandes, G. Saksena, K. Cibulskis, B. Tesar, S. Gabriel, N. Hacohen, M. Meyerson, E. S. Lander, D. Neuberg, J. R. Brown, G. Getz, and C. J. Wu. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell.*, 152(4):714–726, 2013.

[27] F. Nadeu, G. Clot, J. Delgado, D. Martín-García, T. Baumann, I. Salaverria, S. Beà, M. Pinyol, P. Jares, A. Navarro, H. Suárez-Cisneros, M. Aymerich, M. Rozman, N. Villamor, D. Colomer, M. González, M. Alcoceba, M. J. Terol, B. Navarro, E. Colado, Á. R. Payer, X. S. Puente, C. López-Otín, A. López-Guillermo, A. Enjuanes, and E. Campo. Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia.*, 32(3):645–653, 2018.

[28] E. A. Mroz, A. D. Tward, C. R. Pickering, J. N. Myers, R. L. Ferris, and J. W. Rocco. High intra-tumor genetic heterogeneity is related to worse outcome in head and neck squamous cell carcinoma. *Cancer.*, 119(16):3034–3042, 2013.

[29] R. F. Schwarz, C. K. Y. Ng, S. L. Cooke, S. Newman, J. Temple, A. M. Piskorz, D. Gale, K. Sayal, M. Murtaza, P. J. Baldwin, N. Rosenfeld, H. M. Earl, E. Sala, M. Jimenez-Linan, C. A. Parkinson, F. Markowetz, and J.D. Brenton. Spatial and temporal heterogeneity in high-grade serous ovarian cancer: A phylogenetic analysis. *PLoS Med*, 12(2):e1001789, 2015.

[30] M. J. Field, M. A. Durante, H. Anbunathan, L. Z. Cai, C. L. Decatur, A. M. Bowcock, S. Kurtenbach, and J. W. Harbour. Punctuated evolution of canonical genomic aberrations in uveal melanoma. *Nat Commun.9: 116*, 9(1):116, 2018.

[31] R. Gao, A. Davis, T. O. McDonald, E. Sei, X. Shi, Y. Wang, P. C. Tsai, A. Casasent, J. Waters, H. Zhang, F. Meric-Bernstam, F. Michor, and N. E. Navin. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet.*, 48(10): 1119–1130, 2016.

[32] S. C. Baca, D. Prandi, M. S. Lawrence, J. M. Mosquera, A. Romanel, Y. Drier, K. Park, N. Kitabayashi, T. Y. MacDonald, M. Ghandi, E. Van Allen, G. V. Kryukov, A. Sboner, J. P. Theurillat, T. D. Soong, E. Nickerson, D. Auclair, A. Tewari, H. Beltran, R. C. Onofrio, G. Boysen, C. Guiducci, C. E. Barbieri, K. Cibulskis, A. Sivachenko, S. L. Carter, G. Saksena, D. Voet, A. H. Ramos, W. Winckler, M. Cipicchio, K. Ardlie, P. W. Kantoff, M. F. Berger, S. B. Gabriel, T. R. Golub, M. Meyerson, E. S. Lander, O. Elemento, G. Getz, F. Demichelis, M. A. Rubin, and L. A. Garraway. Punctuated evolution of prostate cancer genomes. *Cell.*, 153(3):666–677, 2013.

[33] S. Turajlic, H. Xu, K. Litchfield, A. Rowan, S. Horswell, T. Chambers, T. O'Brien, J. I. Lopez, T. B. K. Watkins, D. Nicol, M. Stares, B. Challacombe, S. Hazell, A. Chandra, T. J. Mitchell, L. Au, C. Eichler-Jonsson, F. Jabbar, A. Soultati, S. Chowdhury, S. Rudman, J. Lynch, A. Fernando, G. Stamp, E. Nye, A. Stewart, W. Xing, J. C. Smith, M. Escudero, A. Huffman, N. Matthews, G. Elgar, B. Phillimore, M. Costa, S. Begum, S. Ward, M. Salm, S. Boeing, R. Fisher, L. Spain, C. Navas, E. Grönroos, S. Hobor, S. Sharma, I. Aurangzeb, S. Lall, A. Polson, M. Varia, C. Horsfield, N. Fotiadis, L. Pickering, R. F. Schwarz, B. Silva, J. Herrero, N. M. Luscombe, M. Jamal-Hanjani, R. Rosenthal, N. J. Birkbak, G. A. Wilson, O. Pipek, D. Ribli, M. Krzystanek, I. Csabai, Z. Szallasi, M. Gore, N. McGranahan, P. Van Loo, P. Campbell, J. Larkin, C. Swanton, and TRACERx Renal Consortium. Deterministic evolutionary trajectories influence primary tumor growth: Tracerx renal. *Cell.*, 173(3):595–610, 2018.

[34] C. A. Ortmann, D. G. Kent, J. Nangalia, Y. Silber, D. C. Wedge, J. Grinfeld, E. J. Baxter, C. E. Massie, E. Papaemmanuil, S. Menon, A. L. Godfrey, D. Dimitropoulou, P. Guglielmelli, B. Bellosillo, C. Besses, K. Döhner, C. N. Harrison, G. S. Vassiliou, A. Vannucchi, P. J. Campbell, and A. R. Green. Effect of mutation order on myeloproliferative neoplasms. *N Engl J Med.*, 372(7):601–612, 2015.

[35] G. Caravagna, Y. Giarratano, D. Ramazzotti, I. Tomlinson, T. A. Graham, G. Sanguinetti, and A. Sottoriva. Detecting repeated cancer evolution in human tumours from multi-region sequencing data. *Nat Methods.*, 15(9):707–714, 2018.

[36] A. M. Baker, W. Cross, K. Curtius, I. A. Bakir, C. H. R. Choi, H. L. Davis, D. Temko, S. Biswas, P. Martinez, M. J. Williams, J. O. Lindsay, R. Feakins, R. Vega, S. J. Hayes, I. P. M. Tomlinson, S. A. C. McDonald, M. Moorghen, A. Silver, J. E. East, N. A. Wright,

L. M. Wang, M. Rodriguez-Justo, M. Jansen, A. L. Hart, S. J. Leedham, and T. A. Graham. Evolutionary history of human colitis-associated colorectal cancer. *Gut.*, 68(6):985–995, 2019.

[37] A. Hochhaus, R. A. Larson, F. Guilhot, J. P. Radich, S. Branford, T. P. Hughes, M. Baccarani, M. W. Deininger, F. Cervantes, S. Fujihara, C. Ortmann, and H. D. Menssen. Long-term outcomes of imatinib treatment for chronic myeloid leukemia. *N Engl J Med.*, 376(10): 917–927, 2017.

[38] A. N. Hata, M. J. Niederst, H. L. Archibald, M. Gomez-Caraballo, F. M. Siddiqui, H. E. Mulvey, Y. E. Maruvka, F. Ji, H. E. Bhang, V. Krishnamurthy-Radhakrishna, G. Siravegna, H. Hu, S. Raoof, E. Lockerman, A. Kalsy, D. Lee, C. L. Keating, D. A. Ruddy, L. J. Damon, A. S. Crystal, C. Costa, Z. Piotrowska, A. Bardelli, A. J. Iafrate, R. I. Sadreyev, F. Stegmeier, G. Getz, L. V. Sequist, A. C. Faber, and J. A. Engelman. Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat Med.*, 22(3):262–269, 2016.

[39] S. Misale, R. Yaeger, S. Hobor, E. Scala, M. Janakiraman, D. Liska, E. Valtorta, R. Schiavo, M. Buscarino, G. Siravegna, K. Bencardino, A. Cercek, C. T. Chen, S. Veronese, C. Zanon, A. Sartore-Bianchi, M. Gambacorta, M. Gallicchio, E. Vakiani, V. Boscaro, E. Medico, M. Weiser, S. Siena, F. Di Nicolantonio, D. Solit, and A. Bardelli. Emergence of kras mutations and acquired resistance to anti-egfr therapy in colorectal cancer. *Nature.*, 486(7404): 532–536, 2012.

[40] P. M. Altrock, L. L. Liu, and F. Michor. The mathematics of cancer: integrating quantitative models. *Nat Rev Cancer.*, 15(12):730–740, 2015.

[41] H.M. Byrne. Dissecting cancer through mathematics: from the cell to the animal model. *Nat Rev Cancer.*, 10(3):221–230, 2010.

[42] I. Newton. Methodus fluxionum et serierum infinitarum (the method of fluxions and infinite series). *Opuscula, 1744, Vol. I. p. 66*, 1671.

[43] P. Rushton. Malthus's law and tumour growth. *Br Med J.*, 1(1321):778, 1886.

[44] S. Pradhan, J. L. Sperduto, C. J. Farino, and J. H. Slatercorresponding. Engineered in vitro models of tumor dormancy and reactivation. *J Biol Eng.*, 12:37, 2018.

[45] P. Schuster and K. Sigmund. Replicator dynamics. *J. Theor. Biol.*, 100:533–538, 1983.

[46] R.V. Solé and T.S. Deisboeck. An error catastrophe in cancer? *J Theor Biol.*, 228(1):47–54, 2004.

[47] Y. Brumer, F. Michor, and E.I. Shakhnovich. Genetic instability and the quasispecies mode. *J Theor Biol.*, 241(2):216–222, 2006.

[48] B. Schönfisch and A. de Roos. Synchronous and asynchronous updating in cellular automata. *Biosystems.*, 51(3):123–143, 1999.

[49] P. Arrighi, N. Schabanel, and G. Theyssier. Stochastic cellular automata: correlations, decidability and simulations. *arXiv:1304.7185*, 2013.

[50] A. Deutsch and J. Moreira. Cellular automaton models of tumor development: A critical review. *Ad Complex Syst.*, 5:247–267, 2002.

[51] C. J. Thalhauser, J. S. Lowengrub, D. Stupack, and N. L. Komarova. Selection in spatial stochastic models of cancer: migration as a key modulator of fitness. *Biol Direct.*, 5:21, 2010.

[52] Ho. Perfahl, H.M. Byrne, T. Chen, V. Estrella, T. Alarcon, A. Lapin, R. A. Gatenby, R. J. Gillies, M. C. Lloyd, P. K. Maini, M. Reuss, and M. R. Owen. Multiscale modelling of vascular tumour growth in 3d: the roles of domain size and boundary conditions. *PLoS One.*, 6(4):e14790, 2011.

[53] R. Axelrod. *The complexity of cooperation: agent-based models of competition and collaboration.* Princeton University Press, 1997.

[54] T. Alarcón, H. M. Byrne, and P. K. Maini. A mathematical model of the effects of hypoxia on the cell-cycle of normal and cancer cells. *J Theor Biol.*, 229(3):395–411, 2004.

[55] J. Wang, L. Zhang, C. Jing, G. Ye, H. Wu, H. Miao, Y. Wu, and X. Zhou. Multi-scale agent-based modeling on melanoma and its related angiogenesis analysis. *Theor Biol Med Model.*, 10:41, 2013.

[56] C. Bianca and M. Pennisi. The triplex vaccine effects in mammary carcinoma: A nonlinear model in tune with simtriplex. *Nonlinear Analysis: Real World Applications, Volume 13, Issue 4, Pages 1913-1940*, 13(4):1913–1940, 2012.

[57] J. N. Kather, J. Poleszczuk, M. Suarez-Carmona, J. Krisam, P. Charoentong, N. A. Valous, C. A. Weis, L. Tavernar, F. Leiss, E. Herpel, F. Klupp, A. Ulrich, M. Schneider, A. Marx, D. Jäger, and N. Halama. In silico modeling of immunotherapy and stroma-targeting therapies in human colorectal cancer. *Cancer Res.*, 77(22):6442–6452, 2017.

[58] M. Pennisi, F. Pappalardo, and S. Motta. Agent based modeling of lung metastasis-immune system competition. *ICARIS*, pages 1–3, 2009.

[59] N. Jagiella, B. Müller, M. Müller, I. E. Vignon-Clementel, and D. Drasdo. Inferring growth control mechanisms in growing multi-cellular spheroids of nsclc cells from spatial-temporal image data. *PLoS Comput. Biol.*, 12(2):e1004412, 2016.

[60] S. Hoehme, F. Bertaux, W. Weens, B. Grasl-KrauppJan, J. G. Hengstler, and D. Drasdo. Model prediction and validation of an order mechanism controlling the spatiotemporal phenotype of early hepatocellular carcinoma. *Bull Math Biol.*, 80(5):1134–1171, 2018.

[61] M. Pennisi, F. Pappalardo, A. Palladini, G. Nicoletti, P. Nanni, P. L. Lollini, and S. Motta. Modeling the competition between lung metastases and the immune system using agents. *BMC Bioinformatics. 2010, 11, S13*, 11(Suppl 7):S13, 2010.

[62] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Royal Society London*, 1763.

[63] D. V. Lindley. Statistical inference. *Journal of the Royal Statistical Society*, 16:30–76, 1953.

[64] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC; 3 edition, 2013.

[65] S. E. Fienberg. When did bayesian inference become "bayesian". *Bayesian Anal.*, 1(1):1–40, 2006.

[66] V. B. Kaiser, M. S. Taylor, and C. A. Semple. Mutational biases drive elevated rates of substitution at regulatory sites across cancer types. *PLoS Genet.*, 12(8):e1006207, 2016.

[67] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jäger, D. T. Jones, D. Jones,

S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, Australian Pancreatic Cancer Genome Initiative., ICGC Breast Cancer Consortium., ICGC MMML-Seq Consortium., ICGC PedBrain., J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton. Signatures of mutational processes in human cancer. *Nature.*, 500(7463): 415–421, 2013.

[68] N. McGranahan and C. Swanton. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell.*, 27(1):15–26, 2015.

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

[70] M. Griffith, C. A. Miller, O. L. Griffith, K. Krysiak, Z. L. Skidmore, A. Ramu, J. R. Walker, H. X. Dang, L. Trani, D. E. Larson, R. T. Demeter, M. C. Wendl, J. F. McMichael, R. E. Austin, V. Magrini, S. D. McGrath, A. Ly, S. Kulkarni, M. G. Cordes, C. C. Fronick, R. S. Fulton, C. A. Maher, L. Ding, J. M. Klco, E. R. Mardis, T. J. Ley, and R. K. Wilson. Optimizing cancer genome sequencing and analysis. *Cell Syst.*, 1(3), 2015.

[71] N. McGranahan and C. Swanton. Clonal heterogeneity and tumor evolution. *Cell.*, 168(4): 613–628, 2017.

[72] R. Schwartz and A. A. Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nat Rev Genet.*, 18(4):213–229, 2017.

[73] J. M. Alves, T. Prieto, and D. Posada. Multiregional tumor trees are not phylogenies. *Trends Cancer.*, 3(8):546–550, 2017.

[74] A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris. Phylowgs: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.*, 16:35, 2015.

[75] A. R. Anderson, C. J. Tomlin, J. Couch, and D. Gallahan. Mathematics of the integrative cancer biology program. *Interface Focus.*, 3(4):20130023, 2013.

[76] N. Beerenwinkel, R. F. Schwarz, M. Gerstung, and F. Markowetz. Cancer evolution: mathematical models and computational inference. *Syst Biol.*, 64(1):e1–25, 2014.

[77] D. L. Hartl and A. G. Clark. *Principles of population genetics.* OUP USA; 4 edition, 2006.

[78] A. R. Anderson, A. M. Weaver, P. T. Cummings, and V. Quaranta. Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment. *Cell.*, 127(5): 905–915, 2006.

[79] H. Enderling, A. R. Anderson, M. A. J. Chaplain, A. Beheshti, L. Hlatky, and P. Hahnfeldt. Paradoxical dependencies of tumor dormancy and progression on basic cell kinetics. *Cancer Res.*, 69(22):8814–8821, 2009.

[80] A. Sottoriva, J. J. Verhoeff, T. Borovski, S. K. McWeeney, L. Naumov, J. P. Medema, P. M. Sloot, and L. Vermeulen. Cancer stem cell tumor model reveals invasive morphology and increased phenotypical heterogeneity. *Cancer Res.*, 70(1):46–56, 2010.

[81] A. Sottoriva, L. Vermeulen, and S. Tavaré. Modeling evolutionary dynamics of epigenetic mutations in hierarchically organized tumors. *PLoS Comput Biol.*, 7(5):e1001132, 2011.

[82] J. G. Scott, D. Basanta, A. R. Anderson, and P. Gerlee. A mathematical model of tumour self-seeding reveals secondary metastatic deposits as drivers of primary tumour growth. *J R Soc Interface.*, 10(82):20130011, 2013.

[83] S. C. Massey, R. C. Rockne, A. Hawkins-Daarud, J. Gallaher, A. R. Anderson, P. Canoll, and K. R. Swanson. Simulating pdgf-driven glioma growth and invasion in an anatomically accurate brain domain. *Bull Math Biol.*, 80(5):1292–1309, 2018.

[84] M. Robertson-Tessi, R. J. Gillies, R. A. Gatenby, and A. R. Anderson. Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes. *Cancer Res.*, 75(8): 1567–1579, 2015.

[85] B. Waclaw, I. Bozic, M. E. Pittman, R. H. Hruban, B. Vogelstein, and M. A. Nowak. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature.*, 525(7568):261–264, 2015.

[86] R. Sun, Z. Hu, A. Sottoriva, T. A. Graham, A. Harpak, Z. Ma, J. M. Fischer, D. Shibata, and C. Curtis. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat Genet.*, 49(7):1015–1024, 2017.

[87] A. Ghaffarizadeh, R. Heiland, S. H. Friedman, S. M. Mumenthaler, and P. Macklin. Physicell: An open source physics-based cell simulator for 3-d multicellular systems. *PLoS Comput Biol.*, 14(2):e1005991, 2018.

[88] M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, and A. Sottoriva. Identification of neutral tumor evolution across cancer types. *Nat Genet.*, 48(3):238–244, 2016.

[89] M. J. Williams, B. Werner, T. Heide, C. Curtis, C. P. Barnes, A. Sottoriva, and A. G. Trevor. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat Genet.*, 50(6): 895–903, 2018.

[90] R. Durrett. Population genetics of neutral mutations in exponentially growing cancer cell populations. *Ann. Appl. Probab.*, 23(1):230–250, 2013.

[91] S. E. Luria and M. Delbruck. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics.*, 28(6):491–511, 1943.

[92] Y. E. Maruvka, D. A. Kessler, and N. M. Shnerb. The birth-death-mutation process: a new paradigm for fat tailed distributions. *PLoS One.*, 6(11):e26480, 2011.

[93] D. A. Kessler and H. Levine. Large population solution of the stochastic luria-delbruck evolution model. *Proc. Natl Acad. Sci. USA*, 110(29):11682–11687, 2013.

[94] M. Nei, Y. Suzuki, and M. Nozawa. The neutral theory ofmolecular evolution in the genomic era. *Annu Rev Genomics Hum Genet.*, 11:265–289, 2010.

[95] A. Sottoriva, H. Kang, Z. Ma, T. A. Graham, M. P. Salomon, J. Zhao, P. Marjoram, K. Siegmund, M. F. Press, D. Shibata, and C. Curtis. A big bang model of human colorectal tumor growth. *Nat. Genet.*, 47(3):209–216, 2015.

[96] S. F. Levy, J. R. Blundell, S. Venkataram, D. P. Petrov, D. S. Fisher, and G. Sherlock. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature.*, 519: 181–186, 2015.

[97] M. Murtaza, S. J. Dawson, K. Pogrebniak, O. M. Rueda, E. Provenzano, J. Grant, S. F. Chin, D. W. Y. Tsui, F. Marass, D. Gale, H. R. Ali, P. Shah, T. Contente-Cuomo, H. Farahani, K. Shumansky, Z. Kingsbury, S. Humphray, D. Bentley, S. P. Shah, M. Wallis, N. Rosenfeld, and C. Caldas. Multifocal clonal evolution characterized using circulating tumour dna in a case of metastatic breast cancer. *Nat. Commun.*, 6:8760, 2015.

[98] M. Lynch, M. S. Ackerman, J. F. Gout, H. Long, W. Sung, W. K. Thomas, and P. L. Foster. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet.*, 17(11): 704–714, 2016.

[99] H. Lehrach. Virtual clinical trials, an essential step in increasing the effectiveness of the drug development process. *Public Health Genomics.*, 18(6):377–371, 2015.

[100] A. Marusyk, V. Almendro, and K. Polyak. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer.*, 12(5):323–334, 2012.

[101] N. E. Navin and J. Hicks. Tracing the tumor lineage. *Mol Oncol.*, 4(3):267–283, 2010.

[102] J. L. Tsao, J. Zhang, R. Salovaara, Z. H. Li, H. J. Järvinen, J. P. Mecklin, L. A. Aaltonen, and D. Shibata. Tracing cell fates in human colorectal tumors from somatic microsatellite mutations: evidence of adenomas with stem cell architecture. *Am J Pathol.*, 153(4):1189–1200, 1998.

[103] R. Desper, F. Jiang, O. P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer. Inferring tree models of oncogenesis from comparative genomic hybridization data. *J Comput Biol.*, 6(1):37–51, 1999.

[104] E. C. de Bruin, N. McGranahan, R. Mitter, M. Salm, D. C. Wedge, L. Yates, M. Jamal-Hanjani, S. Shafi, N. Murugaesu, A. J. Rowan, E. Grönroos, M. A. Muhammad, S. Horswell,

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

[105] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W.R. McCombie, J. Hicks, and M. Wigler. Tumour evolution inferred by single-cell sequencing. *Nature.*, 472(7341):90–94, 2011.

[106] M. El-Kebir, G. Satas, L. Oesper, and B. J. Raphael. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Syst.*, 3(1):43–53, 2016.

[107] J. P. Huelsenbeck, F. Ronquist, R. Nielsen, and J. P. Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science.*, 294(5550):2310–2314, 2001.

[108] J. Felsenstein. *Inferring phylogenies.* OUP USA, 2003.

[109] S.A. Chowdhury, S.E. Shackney, K. Heselmeyer-Haddad, T. Ried, A.A. Schäffer, and R. Schwartz. Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations. *Bioinformatics.*, 29(13):189–198, 2013.

[110] R. F. Schwarz, A. Trinh, B. Sipos, J. D. Brenton, N. Goldman, and F. Markowetz. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput Biol.*, 10(4):e1003535, 2014.

[111] K. Jahn, J. Kuipers, and N. Beerenwinkel. Tree inference for single-cell data. *Genome Biol. 17, 96*, 2016.

[112] M. Nicoloau, A. J. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl Acad. Sci. USA*, 108:7265–7270, 2011.

[113] M. Gerlinger, A. J. Rowan, S. Horswell, M. Math, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum,

NQ. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal, and C. Swanton. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med.*, 366(10):883–892, 2012.

[114] J. Zhang, J. Fujimoto, J. Zhang, D. C. Wedge, X. Song, J. Zhang, S. Seth, C. W. Chow, Y. Cao, C. Gumbs, K. A. Gold, N. Kalhor, L. Little, H. Mahadeshwar, C. Moran, A. Protopopov, H. Sun, J. Tang, X. Wu, Y. Ye, W. N. William, J. J. Lee, J. V. Heymach, W. K. Hong, S. Swisher, I. I. Wistuba, and P. A. Futreal. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science.*, 346(6206):256–9, 2014.

[115] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger. Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res.*, 41(17):165, 2013.

[116] C. A. Miller, B. S. White, N. D. Dees, M. Griffith, J. S. Welch, O. L. Griffith, R. Vij, M. H. Tomasson, T. A. Graubert, M. J. Walter, M. J. Ellis, W. Schierding, J. F. DiPersio, T. J. Ley, E. R. Mardis, R. K. Wilson, and L. Ding. Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol.*, 10(8): e1003665, 2014.

[117] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nat Methods.*, 11(4):396–8, 2014.

[118] H. Zare, J. Wang, A. Hu, K. Weber, J. Smith, D. Nickerson, C. Song, D. Witten, C. A. Blau, and W.S. Noble. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput Biol.*, 10(7):e1003703, 2014.

[119] D. Frumkin, A. Wasserstrom, S. Itzkovitz, T. Stern, A. Harmelin, R. Eilam, G. Rechavi, and E. Shapiro. Cell lineage analysis of a mouse tumor. *Cancer Res.*, 68(14):5924–5931, 2008.

[120] G. Pennington, C. A. Smith, S. Shackney, and R. Schwartz. Reconstructing tumor phylogenies from heterogeneous single-cell data. *J. Bioinform. Comput. Biol.*, 5:407–427, 2007.

[121] N. E. Navin. The first five years of single-cell cancer genomics and beyond. *Genome Res.*, 25(10):1499–1507, 2015.

[122] M. G. Blum and O. Francois. On statistical tests of phylogenetic tree imbalance: The sackin and other indices revisited. *Math Biosci.*, 195(2):141–153, 2005.

[123] A. Mir, F. Rosselló, and L.A. Rotger. A new balance index for phylogentic trees. *Math Biosci. 241(1):125-36.*, 2013.

[124] C. H. Chang, N. Pal, and J. J. Lin. A note on comparing several poisson means. *Communications in Statistics - Simulations and Computation*, 39(8):1605–1627, 2010.

[125] R. Killick and I. A. Eckley. changepoint: an r package for changepoint analysis. *Journal of Statistical Software*, 58(3), 2014.

[126] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer; 2nd edition, 2016.

[127] N. A. Rosenberg and M. Nordborg. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet.*, 3(5):380–390, 2002.

[128] S. Tavare, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from dna sequence data. *Genetics.*, 145(2):505–518, 1997.

[129] W. Cross, M. Kovac, V. Mustonen, D. Temko, H. Davis, A. M. Baker, S. Biswas, R. Arnold, L. Chegwidden, C. Gatenbee, A. R. Anderson, V. H. Koelzer, P. Martinez, X. Jiang, E. Domingo, D. J. Woodcock, Y. Feng, M. Kovacova, T. Maughan, M. S:CORT Consortium, Jansen, M. Rodriguez-Justo, S. Ashraf, R. Guy, C. Cunningham, J. E. East, D. C. Wedge, L. M. Wang, C. Palles, K. Heinimann, A. Sottoriva, S. J. Leedham, T. A. Graham, and I. P. M. Tomlinson. The evolutionary landscape of colorectal tumorigenesis. *Nat Ecol Evol.*, 2(10):1661–1672, 2018.

[130] A. G. Pakes. Biological applications of branching processes. *Handbook of Statistics*, 21: 693–773, 2003.

[131] D. J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. *Statistical Science.*, 16(1):23–34, 2001.

[132] N. O'Connell. The genealogy of branching processes and the age of our most recent common ancestor. *Advances in Applied Probability.*, 27(2):418–442, 1995.

[133] R. Durrett. Branching process models of cancer. *Mathematical Biosciences Institute Lecture Series.*, 1.1:1–63, 2015.

[134] D. G. Kendall. Birth-and-death processes, and the theory of carcinogenesis. *Biometrika.*, 47(1-2):13–21, 1960.

[135] Y. Iwasa, M.A. Nowak, and F. Michor. Evolution of resistance during clonal expansion. *Genetics.*, 172(4):2557–2566, 2006.

[136] N Komarova. Stochastic modeling of drug resistance in cancer. *J Theor Biol.*, 239(3): 351–366, 2006.

[137] J. Foo and K. Leder. Dynamics of cancer recurrence. *The Annals of Applied Probability.*, 23(4):1437–1468, 2013.

[138] R. Durrett and S. Moseley. Evolution of resistance and progression to disease during clonal expansion of cancer. *Theor Popul Biol.*, 77(1):42–48, 2010.

[139] R. Durrett, J. Foo, K. Leder, J. Mayberry, and F. Michor. Evolutionary dynamics of tumor progression with random fitness values. *Theor Popul Biol.*, 78(1):54–66, 2010.

[140] F. Michor, M. A. Nowak, and Y. Iwasa. Stochastic dynamics of metastasis formation. *J Theor Biol.*, 240(4):521–530, 2006.

[141] H. Haeno and F. Michor. The evolution of tumor metastases during clonal expansion. *J Theor Biol.*, 263(1):30–44, 2010.

[142] M. D. Nicholson and T. Antal. Universal asymptotic clone size distribution for general population growth. *Bull Math Biol.*, 78(11):2243–2276, 2016.

[143] T. J. Bailey Norman. *The elements of stochastic processes with applications to the natural sciences.* John Wiley and Sons, 1990.

[144] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics.*, 22(4):403–434, 1976.

[145] D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem.*, 58:35–55, 2007.

[146] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, 1977.

[147] I. A. Rodriguez-Brenes, N. L. Komarova, and D. Wodarz. Tumor growth dynamics: insights into evolutionary processes. *Trends Ecol Evol (Amst).*, 28:597–604, 2013.

[148] C. Schröder and S. Rahmann. A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification. *Algorithms Mol Biol.*, 12:21, 2017.

[149] K. J. Lenos, D. M. Miedema, S. C. Lodestijn, L. E. Nijman, T. van den Bosch, X. Romero Ros, F. C. Lourenço, M. C. Lecca, M. van der Heijden, S. M. van Neerven, A. van Oort, N. Leveille, R. S. Adam, F. de Sousa E Melo, J. Otten, P. Veerman, G. Hypolite, L. Koens, S. K. Lyons, G. Stassi, D. J. Winton, J. P. Medema, E. Morrissey, M. F. Bijlsma, and L. Vermeulen. Stem cell functionality is microenvironmentally defined during tumour expansion and therapy response in colon cancer. *Nat Cell Biol.*, 20(10):1193–1202, 2018.

[150] M. van der Heijden, D. M. Miedema, B. Waclaw, V. L. Veenstra, M. C. Lecca, L. E. Nijman, E. van Dijk, S. M. van Neerven, S. C. Lodestijn, K. J. Lenos, N. E. de Groot, P. R. Prasetyanti, A. Arricibita Varea, D. J. Winton, J. P. Medema, E. Morrissey, B. Ylstra, M. A. Nowak, M. F. Bijlsma, and L. Vermeulen. Spatiotemporal regulation of clonogenicity in colorectal cancer xenografts. *Proc Natl Acad Sci USA.*, 116(13):6140–6145, 2019.

[151] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein, and M. A. Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci USA.*, 107(43):18545–50, 2010.

[152] R. Durrett and J. Schweinsberg. Approximating selective sweeps. *Theor Popul Biol.*, 66(2): 129–138, 2004.

[153] N. Beerenwinkel, T. Antal, D. Dingli, A. Traulsen, K. W. Kinzler, V. E. Velculescu, B. Vogelstein, and M. A. Nowak. Genetic progression and the waiting time to cancer. *PLoS Comput Biol.*, 3(11):e225, 2007.

[154] D. Fusco, M. Gralka, J. Kayser, A. Anderson, and O. Hallatschek. Excess of mutational jackpot events in expanding populations revealed by spatial luria–delbrück experiments. *Nat Comms.*, 7:12760, 2016.

[155] R. Kostadinov, C. C. Maley, and M. K. Kuhner. Bulk genotyping of biopsies can create spurious evidence for hetereogeneity in mutation content. *PLoS Comput Biol.*, 12(4):e1004413, 2016.

[156] R. S. Gejman, A. Y. Chang, H. F. Jones, K. DiKun, A. A. Hakimi, A. Schietinger, and D. A. Scheinberg. Rejection of immunogenic tumor clones is limited by clonal fraction. *eLife.*, 7: e41090, 2018.

[157] S. Ling, Z. Hu, Z. Yang, F. Yang, Y. Li, P. Lin, K. Chen, L. Dong, L. Cao, Y. Tao, L. Hao, Q. Chen, Q. Gong, D. Wu, W. Li, W. Zhao, X. Tian, C. Hao, E. A. Hungate, D. V. Catenacci, R. R. Hudson, W. H. Li, X. Lu, and C. I. Wu. Extremely high genetic diversity in a single tumor points to prevalence of non-darwinian cell evolution. *Proc Natl Acad Sci USA.*, 112(47):E6496–505, 2015.

[158] S. Peischl, I. Dupanloup, M. Kirkpatrick, and L. Excoffier. On the accumulation of deleterious mutations during range expansions. *Mol Ecol.*, 22(24):5972–5982, 2013.

[159] P. Eirew, A. Steif, J. Khattra, G. Ha, D. Yap, H. Farahani, K. Gelmon, S. Chia, C. Mar, A. Wan, E. Laks, J. Biele, K. Shumansky, J. Rosner, A. McPherson, C. Nielsen, A. J. Roth, C. Lefebvre, A. Bashashati, C. de Souza, C. Siu, R. Aniba, J. Brimhall, A. Oloumi, T. Osako, A. Bruna, J. L. Sandoval, T. Algara, W. Greenwood, K. Leung, H. Cheng, H. Xue, Y. Wang, D. Lin, A. J. Mungall, R. Moore, Y. Zhao, J. Lorette, L. Nguyen, D. Huntsman, C. J. Eaves, C. Hansen, M. A. Marra, C. Caldas, S. P. Shah, and S. Aparicio. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature.*, 518(7539):422–426, 2015.

[160] S. Salehi, A. Steif, A. Roth, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. ddclone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biol.*, 18:44, 2017.

[161] S. F. Roerink, N. Sasaki, H. Lee-Six, M. D. Young, L. B. Alexandrov, S. Behjati, T. J. Mitchell, S. Grossmann, H. Lightfoot, D. A. Egan, A. Pronk, N. Smakman, J. van Gorp, E. Anderson, S. J. Gamble, C. Alder, M. van de Wetering, P. J. Campbell, M. R. Stratton, and H. Clevers. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature.*, 556(7702):457–462, 2018.

[162] M. A. Beaumont. Approximate bayesian computation in evolution and ecology. *Annual Rev.*, 41:379–406, 2010.

[163] V. Plagnol and S. Tavaré. Approximate bayesian computation and mcmc. *Monte Carlo and Quasi-Monte Carlo Methods. Springer, Berlin, Heidelberg*, page 99–113, 2004.

[164] D. Schuhmacher, B. Bähre, C. Gottschlich, V. Hartmann, F. Heinemann, and B. Schmitzer. transport: Computation of optimal transport plans and wasserstein distances. *R package version 0.11-1*, 2019. URL `https://cran.r-project.org/package=transport`.

[165] T. Toni and Stumpf M. P. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics.*, 26(1):104–110, 2010.

[166] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface.*, 6(31):187–202, 2009.

[167] M. Nanda and R. Durrett. Genotype patterns in growing solid tumors. *bioRxiv: 390385.*, 2018.

[168] M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate bayesian computation in population genetics. *Genetics.*, 162:2025–2035, 2002.

[169] E. O. Buzbas and N.A. Rosenberg. Aabc: Approximate approximate bayesian computation for inference in population-genetic models. *Theor Popul Biol.*, 99:31–42, 2015.

[170] M. Jamal-Hanjani, G. A. Wilson, N. McGranahan, N. J. Birkbak, T. B. K. Watkins, S. Veeriah, S. Shafi, D. H. Johnson, R. Mitter, R. Rosenthal, M. Salm, S. Horswell, M. Escudero, N. Matthews, A. Rowan, T. Chambers, D. A. Moore, S. Turajlic, H. Xu, S. M. Lee, M. D. Forster, T. Ahmad, C. T. Hiley, C. Abbosh, M. Falzon, E. Borg, T. Marafioti, D. Lawrence, M. Hayward, S. Kolvekar, N. Panagiotopoulos, S. M. Janes, R. Thakrar, A. Ahmed, F. Blackhall, Y. Summers, R. Shah, L. Joseph, A. M. Quinn, P. A. Crosbie, B. Naidu, G. Middleton, G. Langman, S. Trotter, M. Nicolson, H. Remmen, K. Kerr, M. Chetty, L. Gomersall, D. A. Fennell, A. Nakas, S. Rathinam, G. Anand, S. Khan, P. Russell, V. Ezhil, B. Ismail, M. Irvin-Sellers, V. Prakash, J. F. Lester, M. Kornaszewska, R. Attanoos, H. Adams, H. Davies, S. Dentro, P. Taniere, B. O'Sullivan, H. L. Lowe, J. A. Hartley, N. Iles, H. Bell, Y. Ngai, J. A. Shaw, J. Herrero, Z. Szallasi, R. F. Schwarz, A. Stewart, S. A. Quezada, J. Le Quesne,

P. Van Loo, C. Dive, A. Hackshaw, C. Swanton, and TRACERx Consortium. Tracking the vvolution of non–small-cell lung cancer. *N Engl J Med.*, 376(22):2109–2121, 2017.

[171] J. W. Ponder and D. A. Case. Force fields for protein simulations. *Adv Protein Chem.*, 66: 27–85, 2003.

[172] G. R. Mirams, C. J. Arthurs, M. O. Bernabeu, R. Bordas, J. Cooper, A. Corrias, Y. Davit, S. J. Dunn, A. G. Fletcher, D. G. Harvey, M. E. Marsh, J. M. Osborne, P. Pathmanathan, J. Pitt-Francis, J. Southern, N. Zemzemi, and D. J. Gavaghan. Chaste: an open source c++ library for computational physiology and biology. *PLoS Comput Biol.*, 9:e1002970, 2013.

[173] A. Klosin, K. Reis, C. Hidalgo-Carcedo, E. Casas, T. Vavouri, and B. Lehner. Impaired dna replication derepresses chromatin and generates a transgenerationally inherited epigenetic memory. *Science Advances.*, 3(8):e1701143, 2017.

[174] Z. Herceg and P. Hainaut. Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. *Mol Oncol.*, 1(1):26–41, 2007.

[175] B. Sadikovic, K. Al-Romaih, J. A. Squire, and M. Zielenska. Cause and consequences of genetic and epigenetic alterations in human cancer. *Curr Genomics.*, 9(6):394–408, 2008.

[176] P. Polak, R. Karlić, A. Koren, R. Thurman, R. Sandstrom, M. Lawrence, A. Reynolds, E. Rynes, K. Vlahovičcek, J. A. Stamatoyannopoulos, and S. R. Sunyaev. Cell-of-origin chromatin journal shapes the mutational landscape of cancer. *Nature.*, 518(7539):360–364, 2015.

[177] X. Dai, L. Xiang, T. Li, and Z. Bai. Cancer hallmarks, biomarkers and breast cancer molecular subtypes. *J Cancer.*, 7(10):1281–1294, 2016.

[178] D. K. Patten, G. Corleone, B. Gyorffy, Y. Perone, N. Slave, I. Barozzi, E. Erdos, A. Saiakhova, K. Goddard, A. Vingiani, S. Shousha, L. S. Pongor, D. J. Hadjiminas, G. Schiavon, P. Barry, C. Palmieri, R. C. Coombes, P. Scacheri, G. Pruneri, and L. Magnani. Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer. *Nat Med.*, 24(9):1469–1480, 2018.

[179] S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, P. Van Loo, Y. S. Ju, M. Smid, A. B. Brinkman,

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

[180] V. T. Nguyen, I. Barozzi, M. Faronato, Y. Lombardo, J. H. Steel, N. Patel, P. Darbre, L. Castellano, B. Győrffy, L. Woodley, A. Meira, D. K. Patten, V. Vircillo, M. Periyasamy, S. Ali, G. Frige, S. Minucci, R. C. Coombes, and L. Magnani. Differential epigenetic reprogramming in response to specific endocrine therapies promotes cholesterol biosynthesis and cellular invasion. *Nat Commun.*, 6:10044, 2015.

[181] F. Falahi, C. Huisman, H. G. Kazemier, P. van der Vlies, K. Kok, G. A. Hospers, and M. G. Rots. Towards sustained silencing of her2/neu in cancer by epigenetic editing. *Mol Cancer Res.*, 11(9):1029–1039, 2013.

[182] F. Laprell, K. Finkl, and J. Müller. Propagation of polycomb-repressed chromatin requires sequence-specific recruitment to dna. *Science.*, 356(6333):85–88, 2017.

[183] X. Wang and D. Moazed. Dna sequence-dependent epigenetic inheritance of gene silencing and histone h3k9 methylation. *Science.*, 356(6333):88–91, 2017.

[184] R.T. Coleman and G. Struhl. Causal role for inheritance of h3k27me3 in maintaining the off state of a drosophila hox gene. *Science.*, 356(6333):pii: eaai8236, 2017.

[185] L. Magnani, J. Eeckhoute, and M. Lupien. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends Genet.*, 27(11):465–474, 2011.

[186] K. M. Jozwik and J. S. Carroll. Pioneer factors in hormone-dependent cancers. *Nat Rev Cancer.*, 12(6):381–385, 2012.

[187] Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 reference human epigenomes. *Nature.*, 518(7539):317–330, 2015.

[188] ENCODE Project Consortium. et al. An integrated encyclopedia of dna elements in the human genome. *Nature.*, 489(7414):57–74, 2012.

[189] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature.*, 473(7345): 43–49, 2011.

[190] W. A. Whyte, D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell.*, 153(2):307–319, 2013.

[191] N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet.*, 39(3):311–318, 2007.

[192] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, LA. Jr. Diaz, and K. W. Kinzler. Cancer genome landscapes. *Science.*, 339(6127):1546–1558, 2013.

[193] S. Wang, M. Jia, Z. He, and X. S. Liu. Apobec3b and apobec mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. *Oncogene.*, 37(29):3924–3936, 2018.

[194] J. S. Gehring, B. Fischer, M. Lawrence, and W. Huber. Somaticsignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics.*, 31(22):3673–5, 2015.

[195] Y. Shiraishi, G. Tremmel, S. Miyano, and M. Stephens. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet.*, 11(12):e1005657, 2015.

[196] A. Fischer, C. J. Illingworth, P. J. Campbell, and V. Mustonen. Emu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.*, 14(4): R39, 2013.

[197] D. Ramazzotti, A. Lal, K. Liu, R. Tibshirani, and A. Sidow. De novo mutational signature discovery in tumor genomes using sparsesignatures. *bioRxiv: 384834*, 2018.

[198] L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. NG. Ng, A. Boot, K. R. Covington, D. A. Gordenin, E. Bergstrom, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, G. Getz, S. G. Rozen, and M. R. Stratton. The repertoire of mutational signatures in human cancer. *bioRxiv: 322859*, 2018.

[199] L. B. Alexandrov and M. R. Stratton. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev.*, 24:52–60, 2014.

[200] P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. Bruijn, K. Duyvesteyn, S. Haidari, A. Hoeck, W. Onstenk, P. Roepman, C. Shale, M. Voda, H. J. Bloemendal, V. C. G. Tjan-Heijnen, C. M. L. Herpen, M. Labots, P. O. Witteveen, E. F. Smit, S. Sleijfer, E. E. Voest, and E. Cuppen. Pan-cancer whole genome analyses of metastatic solid tumors. *Nature.*, 575(7781):210–216, 2019.

[201] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, and Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Nat Genet.*, 45(10):1113–1120, 2013.

[202] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet.*, 16(2):85–97, 2015.

[203] T. M. Baye, T. Abebe, and R. A. Wilke. Genotype–environment interactions and their translational implications. *Per Med.*, 8(1):59–70, 2011.

[204] M. Periyasamy, H. Patel, C. F. Lai, V. T. M. Nguyen, E. Nevedomskaya, A. Harrod, R. Russell, J. Remenyi, A. M. Ochocka, R. S. Thomas, F. Fuller-Pace, B. Győrffy, C. Caldas, N. Navaratnam, J. S. Carroll, W. Zwart, R. C. Coombes, L. Magnani, L. Buluwela, and S. Ali. Apobec3b-mediated cytidine deamination is required for estrogen receptor action in breast cancer. *Cell Rep.*, 13(1):108–121, 2015.

[205] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell*. Garland Science; 4th edition, 2002.

[206] S. B. Baylin and J. E. Ohm. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction. *Nat Rev Cancer.*, 6(2):107–116, 2006.

[207] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.*, 499(7457):214–218, 2013.

[208] A. Hodgkinson, Y. Chen, and A. Eyre-Walker. The large-scale distribution of somatic mutations in cancer genomes. *Hum Mutat.*, 33(1):136–143, 2012.

[209] L. Liu, S. De, and F. Michor. Dna replication timing and higher-order nuclear journal determine single-nucleotide substitution patterns in cancer genomes. *Nat Commun.*, 4:1502, 2013.

[210] B. Schuster-Böckler and B. Lehner. Chromatin journal is a major influence on regional mutation rates in human cancer cells. *Nature.*, 488(7412):504–507, 2012.

[211] Y. H. Woo and W. H. Li. Dna replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat Commun.*, 3:1004, 2012.

[212] M. Ducasse and M. A. Brown. Epigenetic aberrations and cancer. *Mol Cancer.*, 5:60, 2006.

[213] W. S. Reznikoff. Transposon tn5. *Annu Rev Genet.*, 42:269–286, 2008.

[214] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific; 2nd edition, 1999.

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