Pan-cancer Reconstruction of Clonal Evolution in 1,800 Patients Using the Discrete-Time Branching Process

Luis Eduardo Lara-Gonzalez
(ORCID: 0000-0003-3897-353X)

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

May 2020

Faculty of Medicine, Dentistry and Health Sciences
Sir Peter MacCallum Department of Oncology
The University of Melbourne

Abstract

Intra and inter-tumour heterogeneity poses a challenge for associating molecular and immunohistochemical markers with clinical outcomes. Sequencing technologies has enabled detailed assessment of tumour heterogeneity, facilitating the genomic characterisation of tumours. Whilst such technologies have revealed mutational landscapes and have identified key driver alterations for tumorigenesis, pan-cancer clonal evolution reconstructions are lacking. In order to bridge this gap, I used discrete-time branching models to derive biological insights into tumour progression and reconstructed the clonal evolution in 1,800 patients, successfully linking mutations with growth patterns of disease progression.

I first modified a discrete time-branching process to account for individual clonal subpopulations and derived analytical solutions for expectation and variance of both clonal and tumour expansions. Additionally, I derived the expected time for any given clone to successfully expand as $\hat{\tau}$, and with the use of these analytical solutions, I showed the likely driver and clonal compositions of the tumours and their phylogenies.

Secondly, I generated a database of results from four different versions of time-branching process models that covered multiple parameters. Here I identified how an increase in diversity arises by both increased mutation rate and reduced fitness. I further corroborated that total number of drug resistant cells is directly proportional to lineage extinction probability (δ) and tumour size as shown in previous studies. I also showed that this effect can be extrapolated to other types of functional passenger mutations involved in cancer-specific mortality. Moreover, I showed how commonly used sequencing cut-offs limit the accurate inference of tumour's average selective advantage and driver mutation rate.

Thirdly, I identified that a minimum distance metric can provide accurate fits of simulated cancer cell fractions to real patient tumour data. This metric showed at least 80% accuracy to identify the initial parameters of s and u and at least 40% accuracy to recover the correct evolutionary trajectory.

Fourthly, I applied this fitting procedure to reconstruct the evolutionary trajectories of 1,800 tumours from different cancer sequencing studies. The best fits derived suggests that the most likely parameters for the evolution of solid tumours are high driver mutation rates and weak driver effects of fitness.

Fifthly, using The Cancer Genome Atlas cohort, I identified an association between predicted degree of clonality and survival, and found branched topologies are common in malignancies with adverse prognosis. In the TRACERx non-small-cell lung cancer cohort, I identified that clonal reconstructions agreed with previously reported phylogenies. Additionally, using data from the Breast International Group 1-98, I identified the role of tumour fitness in determining clinical outcome, and the evolutionary dynamics of TP53 and PIK3CA mutations conducive to distant metastasis.

Finally, using data from a metastatic melanoma patient collected through the CASCADE melanoma study, I was able to propose a pattern of dissemination from the primary to metastatic sites in the liver and brain based on the phylogenies recovered from my data-fitting procedure. This study demonstrates the power of the discrete-time branching process in reconstructing tumour evolution, and its potential to uncover insights in the dynamics of tumour growth that are missed by current methods.

Declaration

This is to certify:

1. This thesis comprises only original work towards this PhD degree expect where indicated in the Preface.
2. Due acknowledgments had been made in the text and all other materials.
3. This thesis is less than 100,00 words excluding tables, references and appendices, as approved by the Research Higher Degrees Committee.

Luis Eduardo Lara-Gonzalez

May 2020

Preface

The work presented here was carried out at the Peter MacCallum Cancer Centre between April 2016 and May 2020. The following people have contributed to my work:

Vivien Yeung in developing analysis on the expectation and variance of growth in the *k*-driver discrete-time branching process.

Carl Gu in validating analytical solution of expected time to growth in the *k*-driver discrete-time branching process.

Shu Wen Xu in developing solutions for the inhomogeneous differential equation for the discrete-time branching process reported in Appendix A.2.2.

Farid Kaveh in post-analysis of the minimum Euclidian distance method with false discovery correction by neighbouring. To determine how the number of cancer cell fractions affect the quality of the fits. His contribution helped in the generation of Figure 4.7.

The Breast International Group 1-98 and Sherene Loi who provided the samples of this study as described in Chapter IV.

Ismael Vergara who shared the cases of the CASCADE melanoma patient described in Chapter IV.

The Nectar research cloud that facilitated the computation of the database of branching process models generated in Chapter III.

The National Computation Infrastructure (NCI) for allowing me to use their high-performance computation machines to generate simulations of tumour development that helped the generation of the database described in Chapter III.

Cancer Therapeutics CRC for awarding me a top-up scholarship during my PhD studies.

I gratefully acknowledge **El Consejo Nacional de Ciencia y Tecnologia CONACyT** of Mexico for awarding me a top-up scholarship to support my PhD studies overseas.

Acknowledgements

My deep and honest appreciation to David Goode for all his support, invaluable advice and guidance during my post-graduate studies. Thank you for being a crucial mentor in my development as a scientist.

I would like to thank Sherene Loi for the supervision and guidance to better understand the clinical role in my project. Thank you for having a significant impact in my PhD project.

I would also like to thank Davide Ferrari and Anthony Papenfuss for the supervisor and the statistical and computational input. Thank you, your feedback, support and resources played a key role in developing my thesis.

A very special thanks to all my colleagues at Goode and Loi labs for their immense support during my PhD studies.

A very special thank you to my dear friends that are always supportive and helped me in numerous ways. A heartfelt thanks to Sara, Valeria, David Stasiak, Kat, Chris, Daniel Thomas, Chad, Ursula, Matt, Rae, Fey, Silvia, Clelia, Rosie, Enrique, Alexandra, Paola, Daniela, Eduardo, Ismael, Michael, Richard, and PhD cohort of 2016 for all the encouraging support.

And finally, to my Mum, Dad, Fernando and Adriana for their loving support, thank you for always encouraging me to persevere and achieve my goals.

Table of Contents

List of Figures

List of Tables

Table 1 Mathematical Symbols

Variable	*Meaning*
s	Fitness or selective advantage.
k	Number of driver alterations, or index of the element of interest.
u	Driver mutation rate, number of loci that can cause a clonal expansion per base pair per cell division.
t	Generation, arbitrary unit of time if the average division rate is not known
b	Proliferation rate, this is a function of fitness s
d	Death rate/stagnation probability.
λ	Net growth of a subpopulation or clone
δ	Extinction probability (d/b)
P	Average division rate of the cell type to be modelled
S *or* S_k	Cells in subpopulation containing k driver mutations
C , $C_{k,i,j}$ *or* C_*	Cells clone that emerged from a given lineage *
N	Total tumour cells
B	Number of newborn cells of a given subpopulation or clone at generation t
D	Number of cell deaths of a given subpopulation of clone at generation t
M	Number of driver mutants of a given subpopulation of clone at generation t
τ, τ_k, $\tau_{k,i,j}$ *or* $\tau_{k,*}$	Expected time in which progeny $k+1$ of subpopulation of clone k successfully emerges.
$\hat{\tau}$, $\hat{\tau}_k$, $\hat{\tau}_{k,*}$ *or* $\hat{\tau}_{k,i,j}$	Approximation of τ for a given subpopulation of clonal offspring
w_{wt}	Fitness background of the wildtype cell assumed to be 1
β	Detectability threshold in the range of [0,1]
β_k	Carrying capacity depending on k
β_N	Carrying capacity depending on tumour size
t_N	Target tumour size
ε	Calibration parameter in the range [0,1]
c	c^{th} clone coming from a given parental clone
μ	Net growth rate of the Galton-Watson branching process
σ^2	Variance of a subpopulation or at time $t=1$ or $t-\tau_{k,*}=1$
μ_r	Drug resistance rate
t_ε	Point in which and emerging clone reaches the size of its parent

Chapter I

1 Cancer as an Evolutionary Disease

Cancer is a collection of related diseases that manifest as multiple forms of abnormal growth. Cancer progression is fuelled within cells by the accumulation of key (epi)-genetic changes that confer novel traits described in the so-called hallmarks of cancer [1, 2]. These novel traits render cancerous cells robust and adaptable to environmental change, displaying a remarkable capacity to acquire resistance to different therapeutic interventions [3]. Understanding how the molecular alterations operate in concert to provide such phenotypic complexity has created a paradigm shift towards studying cancer as an evolutionary disease, aiming to associate molecular alterations with clinical presentation.

The evolutionary perspective in cancer research gained recognition during the 1950s with the Armitage and Doll model. The model suggested that the cancer incidence observed in epidemiological records requires a sequence of distinct events associated with age for the onset of cancer [4, 5]. Further research consolidated cancer as a multistep process with potential genetic origins leading to studies investigating the role of molecular alterations in cancer.

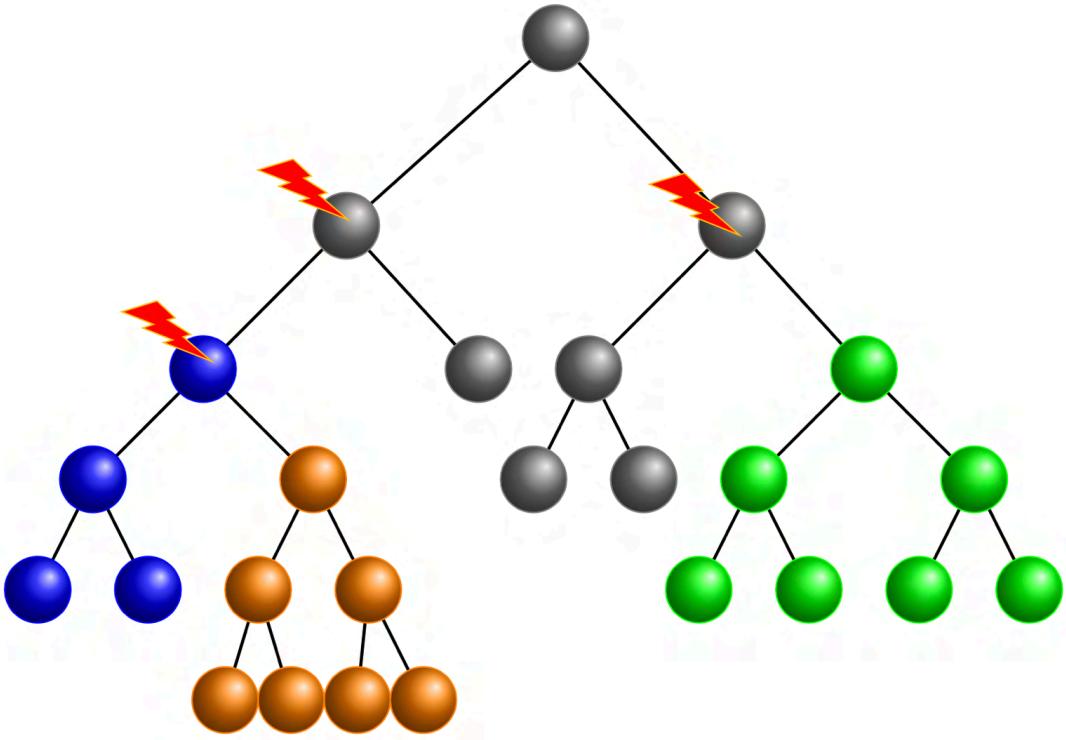


Figure 1.1 Clonal evolution model. Branches represent symmetric cellular division and colours represent the different clones. Driver alterations in red cause the emergence of a new clone with increased proliferation that can outcompete its predecessors.

Further studies showed systematic transformations in neoplastic growth, such as karyotypic aberrations [6] and cytoband losses which culminated in the clonal evolution theory introduced by Peter Nowell in 1976 [7]. The seminal work of Nowell suggests that through the accumulation of genetic and epigenetic alterations, cells are, subject to their tissue microenvironment, iteratively selected to expand and compete for resources, resulting in tumour formation and dissemination of cancerous cells.

As outlined in Figure 1.1, the accumulation of alterations promotes cellular proliferation creating competitive clonal subpopulations as observed in distinctive histopathological stages of tumour development [8].

The impact of the evolutionary framework of Peter Nowell led to experimental efforts to validate the theory [9, 10], not only corroborating that cancer is a genomic disease but also in identifying the oncogenic drivers responsible for cancer. Fearon et al. [11] aligned the evolutionary framework with a solid experimental model showing how mutational activation of oncogenes and loss of tumour suppressors resulted in progressive tumour development in colorectal cancer.

Since then, the evolutionary framework with genomic models was assumed to be representative of multiple malignancies in which genomic alterations promote neoplastic growth that compromises organ-specific functions [12-16].

2 Principles of Evolution Applied to Cancer

Subsequent refinements of the clonal theory postulated that the mechanism of evolution is by Darwinian natural selection [17] and highlighted that diversity of the cancerous phenotype is not only genomic [13, 18-21] but also non-genomic [22-26].

Two global processes can shape the cancerous phenotype: genomic alterations that confer novel traits (e.g. increase in proliferation) and non-genomic alterations that are caused by changes in gene regulation and cellular-environmental interactions [27].

Evidence for the role of genomic alteration has been provided by pan-cancer sequencing studies that have identified driver somatic mutations in multiple malignancies that affect clinical outcome [28-30]. Similarly, evidence of the non-genomic role has been inferred from studies where therapeutic resistance cannot be explained by genomic alterations or the acquisition of them in the therapeutic window [24, 26, 31-33].

Both genomic and non-genomic factors are influenced by random events [34, 35]. Genetic alterations accumulate randomly displaying an aggressive phenotype, and through selection said phenotypes are oriented toward an evolutionary trajectory that allows them to exploit their tissue microenvironment [36-38].

However, phenotypes can, change depending on the environment and without acquiring genomic alterations. Cancerous cells surrounded by rich vascularisation and growth factors can display the same phenotype as a cell that has accumulated many driver alterations. As a consequence, genomic and non-genomic factors complicate the association of genotypes with phenotypes [39-41], having a direct impact on connecting mutational changes with clinical outcome.

Population genetics models of cancer are aware of said effect [42, 43], efforts have been made to determine somatic mutation rates in cancer and connect them in an evolutionary framework to reconstruct tumour and clonal evolution.

Chowell et al. [44] designed a clonal evolution model using the known driver mutation rate in cancer ($3.14\text{x}10^{-5}$ reported by [43]), aiming to study tumour formation by simulating clonal and tumour growth with different initial conditions. These conditions are initial fitness (also

known as average selective advantage) and average driver mutation rate, both determine how fast clones and the tumour expand.

They showed that all parameter combinations displayed heavily skewed clonal distributions. Although the role of the environment is not explicitly modelled, their work highlighted that the process of clonal evolution is skewed, with a few clones representing most of the tumour composition and the majority of clones describing a minimum fraction of it.

They additionally modelled the acquisition of pre-existing drug resistant mutations, and even when conditioned by pre-existing drug resistance, the skewed clonal distributions remained present. This showed that selection is variable and cellular fitness is context-dependent. As exemplified in Figure 1.2, clones display different fitness distributions depending on the environment. Environment A selects for clones with faster expansion by driver accumulation, while environment B selects for pre-existing drug resistant clones, in both cases the distributions are skewed.

A framework to understand clonal evolution in cancer has been proposed by multiple authors [23, 24, 35, 44, 45]. Their core elements can be interpreted as illustrated in Figure 1.2. The association of genotype to phenotype is influenced by genomic and non-genomic factors.

The genomic factors occur within cells by alterations in the DNA (single nucleotide variation, structural variation and copy number changes) and induce a phenotype that is not reversible as depicted with the one-direction arrow in the landscape Fig. 1.2 [44, 45]. The non-genomic factors occur within cells by epigenetic alterations, and outside cells by cellular interaction with other cells and their environment. The non-genomic factors can reverse the phenotype as shown by the two-sided arrow in the landscape of Fig.1.2.

Genomic alterations that are neutral in a given environment can be functional in a different environment. For instance, Figure 1.2 shows environment A to be the equivalent of primary formation in which dominant clones are the ones that accumulate more driver alteration in a period of time. When treatment is introduced, exemplified in environment B, the dynamics change, and selection favours cells that have pre-existing drug resistance. Alterations that are functional in environment B occur randomly in environment A with neutral fitness effects, and their presence is to guarantee survival in sudden environmental changes (known as bet-hedging) [35].

A transitory drug resistance phenotype can be achieved non-genomically by gene expression noise occurring cyclically in around 1% or more of cells as a secondary survival mechanism [23, 24]. In this context, regardless of the environment**,** cells are switching between phenotypic states for the sake of survival [35, 36, 46, 47].

In the example in Figure 1.2 the green clone has a moderate fitness in environment A due to the accumulation of two driver alterations. When the environment changes the green clone switches its phenotype to a drug resistance mode (shown as the one-sided arrow because it was acquired by genetic alteration). This is known as phenotypic plasticity: the capability of a genotype to produce multiple phenotypes [39, 48-50]. This is different to phenotypic diversity which is the number of phenotypes in a given environment.

The clonal evolution model of Chowell et al. [44] allows the modelling of environment A, describing tumour formation by the accumulation of drivers. If the drug resistance mutation

rate is known, the model can provide estimates of how many cells are drug resistant at any time and provide information about the fitness distribution in environment B as exemplified in Figure 1.2 with the green clone.

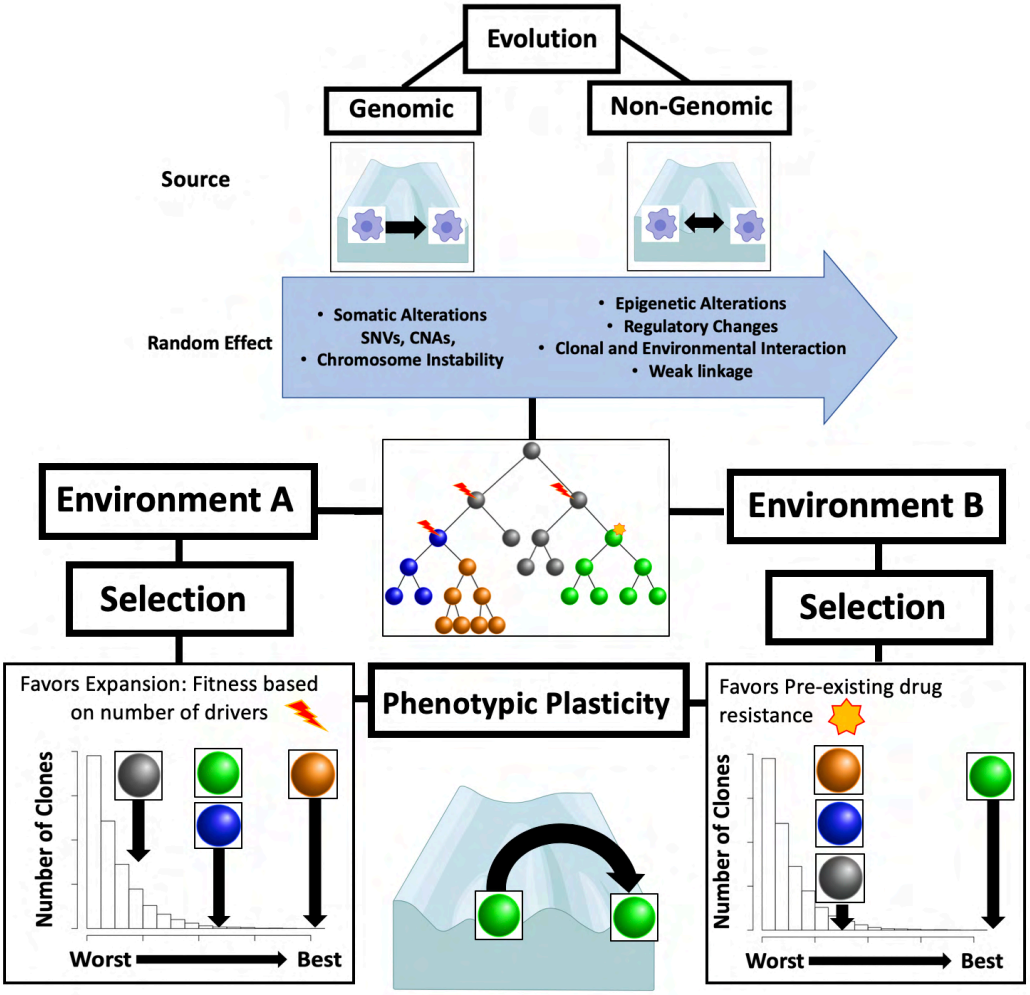


Figure 1.2 Cancer as an evolutionary process. Two sources influence cancerous cells: genomic and non-genomic alterations that occur randomly. Genomic events induce phenotypic states through selection to a given environment. Genetic alterations introduce a phenotype in a given environment which is not reversible, as illustrated by the one-sided arrow. Non-genomic alterations introduce a phenotype in a given environment that can be reversible as illustrated with the two-sided arrow. Environmental changes exemplified as A going to B can modify the evolutionary trajectory, making clones switch their phenotype. In this example the green clone is going to dominate in environment B by harbouring the drug resistance mutation.

3 Clinical Challenges of Tumour Heterogeneity and Treatment Resistance

Inter- and intratumor heterogeneity is a product of clonal evolution [51-53] affected by the random effects as shown in Figure 1.2. It varies between patients and between tumours, limiting the capacity to measure the mutational landscape and identify recurrent evolutionary trajectories[30, 53]. The common denominator is that heterogeneity is a marker of therapeutic failure [29, 30] rooted in genotypic and phenotypic variation.

Genetic heterogeneity is the consequence of random accumulation of mutations that are not negatively selected occurring at different rates [43, 54-59]. Somatic mutations can act as

drivers or be neutral depending on the environment [60, 61]. Drivers increase fitness that results in greater odds of survival while a subset of neutral mutations may be advantageous when environmental conditions change [23, 62] (e.g. drug resistance as in Figure 1.2).

Phenotypic diversity is the result of the alterations within the cell and how the cell interacts with the environment, resulting in a complex genotype/phenotype map that is many-to-many [35, 63-65] and not one-to-one. For instance, cells deprived of growth factors by their surrounding microenvironment may be forced to undergo regulatory changes to reprogram their metabolism for survival's sake, changing their phenotype regardless of their genotype [46, 66, 67].

Figure 1.3 shows the schematic of tumour inter- and intratumor heterogeneity. At the molecular level, genetic and epigenetic changes accumulate randomly. Then, cellular interactions with the environment will render a phenotype subject to positive, neutral or negative selection [68-70]. As a result, genotypic and phenotypic heterogeneity is stochastically determined and varies within tumours (intratumour) and between patients (intertumour).

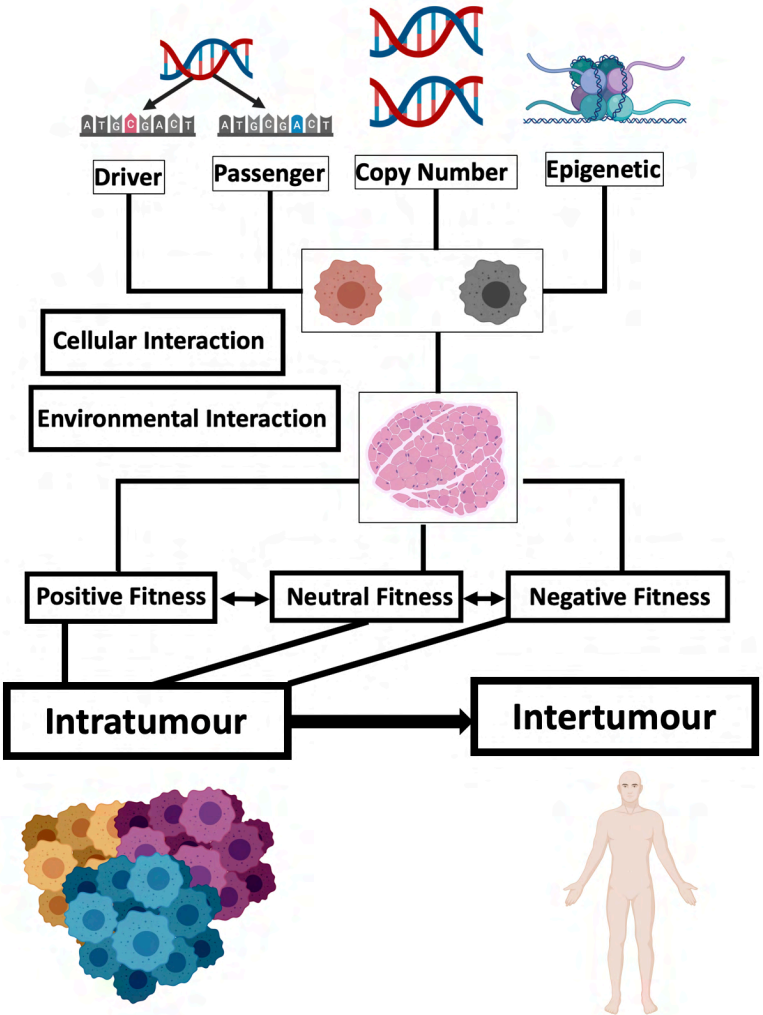


Figure 1.3 Tumour heterogeneity. Molecular alterations accumulated within cells, such as driver, passenger, copy number and epigenetic results in a phenotype. This phenotype can be further influenced by cellular and environmental interactions modifying its fitness as positive, neutral or negative. This dynamic results in a mixture of genotypes and phenotypes stochastically determined within tumours and between patients.

Tumour heterogeneity allows multiple evolutionary trajectories through selection and the co-existence of cells with diverse fitness, thus providing survival advantage in oscillatory environments or in extreme conditions such as therapeutic interventions [35, 71]. This poses a challenge, how to measure it and how to connect alterations with phenotypes. The former because heterogeneity is stochastic and tissue micro environments in the tumour can alter the genotype-phenotype map [35, 62]. The latter, because measuring the main alterations of primary tumours might not reveal the actual biological program(s) responsible for cancer-specific mortality [44, 45, 72].

A secondary challenge is how to consistently capture the comprehensive mutational landscape of a tumour [73]. How many patients, how many tumour regions and optimal sequencing coverage are still open questions [74]. On top of that, it is not clear yet when it is crucial to measure heterogeneity: at primary formation [29, 75, 76], after first line treatment [77], after adjuvant treatment[78], at metastatic disease [79] or during all stages of disease progression.

Although tumour heterogeneity has prognostic power, these important questions that impact upon experimental discovery and design remain unanswered. As a result, it is not yet clear how-to best measure tumour heterogeneity and extract the relevant signal to elucidate the relationship of genomic mutations, evolutionary patterns and clinical outcome.

4 Measuring Mutational Heterogeneity and Clonality

Phenotypic heterogeneity is difficult to measure, as it requires studying clonal subpopulations under specific perturbations to identify a signal of interest (e.g. regulatory change) [33, 80, 81]. On the other hand, genotypic heterogeneity is easier to measure, which can be done by immunohistochemistry, fluorescence activated cell sorting (FACS), mass cytometry (CyTOF), and sequencing. DNA Sequencing has the greatest resolution due to its capability to survey the entire genome at the bulk or single cell level.

The advent of sequencing technologies has enabled the measurement of genetic heterogeneity, providing inferences about tumour evolution and disease progression associating mutational profiles with clinical outcome [82].

Overall, next-generation sequencing technologies measure the mutational prevalence in a sample at a given coverage with the particular sensitivity of the assay used. The coverage of the assay refers to the fraction of the genome assayed. It can be whole genome, whole exome or just some genes of interest. The depth refers to how many reads of a given nucleotide got sequenced, which will impact variant measurability. Bulk sequencing measures a mixture of genomes in a sample that can come from healthy cells or cancerous cells, and the mixture is sequenced altogether.

After sequencing is done, variant and copy number calling is performed to evaluate the penetrance of alterations present in the sequencing sample [83, 84]. Numerous tools exist with different degrees of sensitivity and specificity and some are designed to work better than others in certain sequencing assays. Different studies provide their pipelines with parameters and modifications done in the tools to replicate their results [75, 76]; there is no gold-standard defined.

In sequencing studies, critical decisions have to be made about sequencing coverage, depth and the number of samples of the tumour to sequence. Usually increasing the coverage decreases

the depth which has different implications. For instance, in clinical diagnostics sequencing panels of hundreds of genes are used to measure if oncogenes and tumour suppressors are present [85], whereas in research, whole exome assays are preferred to discover mutational patterns.

The number of samples of a tumour is a critical measure because clonal subpopulations may be located in different areas of the tumour and their measure can represent clinical relevance [74]. For instance, in The Cancer Genome Atlas (TCGA) it is reported that ~30% of the samples are neutral, meaning that only a single clonal subpopulation was measured [86]. Thus, neutral samples can lead to an underrepresentation of the tumour composition.

Once sequencing and variant calling has been performed using a particular assay, driver annotation and clonality assessment can be conducted to evaluate the potential number of clonal expansions present in the sample(s) [74]. Then, an estimation of tumour's phylogeny aligned with clonal fitness driver mutation rate can be achieved as shown by [87, 88].

In order to perform clonality assessment, the first step is to adjust the variant allelic fractions (VAF) to cancer cell fractions (also known as adjusted cellular prevalence, CCF) [89]. This means correcting the VAFs for ploidy to approximate the cellular prevalence of variants in the sample or tumour [90]. Then, the cancer cell fractions are grouped by their frequencies to identify mutational clusters which are a proxy of the (hypothetical) number of clonal sweeps that have occurred in the tumour [91, 92].

The CCFs of single snapshot sequencing methods can be highly biased due to heterogeneity and their prevalence may only be reflective for the biopsy and not for the tumour. Sequencing methods such as rep-seq [93] and multi-region can bypass that proven and provide CCFs representative of the tumour.

Due to inheritance segregation (shown in eq. 1.1), founder lineage alterations are enriched, hence diluting the frequency of the most recent driver events rendering difficult to identify and delineate from sequencing artefacts. Bozic et al. [94] using a branching process showed the impact of inheritance segregation, revealing that clonal detectability is inherently biased towards lower or higher variant allele fractions.

The challenge in clonality assessment is the accurate detection of mutational clusters from allelic copy number adjusted frequencies. This is affected by the mixture of clonal populations harbouring different alterations present in the sample(s), introducing bias in recovering the number of clones and establishing their relationship.

The biological sources of bias are due to the intrinsic dynamics of tumour evolution that limit the accurate recovery of the mutational clusters [94]. These include:

1) Spatial heterogeneity: different regions of the tumours can harbour different subpopulations, requiring multiple samples or sequence-specific assays. Failing to capture a representative sample of the tumour will lead to an underrepresentation of the number of clones, affecting downstream analyses [74, 90, 95].
2) Differential genome-wide copy number change: tumour cells can acquire differential copy number changes such as loss of heterozygosity (LOH) or genome doublings complicating the accurate estimation of cancer cell fractions. Failure to perform

accurate correction can lead to spurious mutational clusters that misrepresent the clonality of the tumour [96, 97].

3) Purity: non-cancerous cells are contained in the sample limiting the detectability of clones [98].
4) The mutational process of evolution: certain tumours evolve predominantly by either point mutations or copy number alterations [28, 97, 99, 100]. If the sequencing assay and the clonality tool are not well suited to the main evolutionary process of the tumour this can lead to a restricted representation of the measurement of interest [101]. For instance, measuring clonality by point-mutations when the tumour evolves by copy number alterations may not reveal the signal of interest.

The biological sources of variation can be reduced by a careful study design and collecting multiple samples with increased coverage and depth. This enables a better ploidy correction and a better representation of the clonal landscape [75].

The technical biases are due to specific use of bioinformatic tools for mutational and clonality calling, each having their different sensitivities and specificities. These depend on the parameters used to run the tools, the type of sequencing assay and the sequencer used.

Popular clonality tools robust for multiple sequencing assays that recover clonality based on point mutations are PyClone [91], SciClone [102] and ExPANdS [92].

- PyClone is a Bayesian hierarchical model providing a probabilistic interpretation of clustering. It assigns a beta-binomial distribution to every mutation in a Dirichlet process and establishes mutational clusters by a Markov Chain Monte Carlo (MCMC) sampling. It provides flexible prior definitions based on copy number change. A new version has been published that improves on its performance and application [103].
- SciClone, only handles copy number neutral and LOH-free loci (diploid), but improves performance by using a Variational Bayesian Beta mixture model on the mutational frequencies as opposed to the Dirichlet process with MCMC sampling used in PyClone. Another key advantage of the Variation Bayesian Beta mixture is its probabilistic assignment for every mutation that is used to generate robust clusters.
- In contrast, ExPANdS models cellular frequencies as probability distributions that are subsequently clustered as a proxy for modelling clonal expansions, under the assumption that cellular frequencies within the same range can potentially originate from the same clonal expansion. This approach relies on enough hitchhiker/passenger mutations being acquired during the clonal expansions to support the signal of the driver alteration(s) behind the clonal sweep.

The previous tools provide different approaches to identify mutational clusters that are driven by somatic variants. When the driver events are driven by focal or cytoband copy number alterations, as described by [97] as class *C* tumours, ploidy correction is limited by multiple copy number states within a locus and a copy-number specific method should be preferred [101].

In such cases, clonality based on copy number calling can be done for whole genome or whole exome sequencing assays. Clonality assessment can be performed by tools like TITAN [104], THetA [105], CloneCNA [106] or Battenberg[107]. These tools use similar probabilistic frameworks based on the Dirichlet process or hidden Markov models to approximate fractions

of copy number change in the genome. Fractions are subsequently clustered as a proxy of clonal sweeps caused by driver copy number alterations.

Although multiple clonality tools are available, careful consideration has to be made to align the sequencing assay with the clonality caller and the mutational process driving the tumour progression. Failing to align these will result in incomplete representation of clonality, limiting power to associate them with patient outcomes [101].

5 Tumour Evolution Reconstruction

The previous sections showed tumours as evolving ecosystems that manifest genotypic and phenotypic heterogeneity that fluctuates according to environmental pressures. Genotypic heterogeneity is one the main measurements given the capability of next generation sequencing to survey the genome. The advantage of sequencing technologies is their capability for downstream analysis to recover clonality based on point mutations, genome-wide copy number alterations or both to reconstruct the clonal evolution of tumours.

Figure 1.4, shows the outline for reconstructing tumour evolution. A tumour sample is sequenced followed by mutation calling. Next, evaluation of neutrality has to be made to define the subsequent analysis [86, 88].

If the sample is neutral, it means that only one subpopulation was measured and therefore the information that can be recovered is the number of driver alterations, fitness and mutation rate [58, 86]. If the sample has more than one subpopulation then information about clonal ancestry and relatedness can be inferred based on cancer cell fractions by clonality tools such as PyClone or ExPANdS [108].

In the case of neutrality, Bozic et al. [58] provides a framework that can be used to fit the neutral tail of the variant allelic frequency distribution, as highlighted in red in Figure 1.4. To use their approach, variant allelic frequencies are assumed to be unaffected by LOH and there must be a sufficient number of mutations with frequencies in the range of 0.12 to 0.25.

In the case that the sample is not neutral, clonality evaluation is the initial step in reconstructing clonal evolution. Once the mutational clusters are generated a deconvolution algorithm needs to resolve the inheritance and relatedness of the samples [73, 100, 101, 109-112].

The cancer cell fraction measured by clonality tools is a function of the founder clone that introduced the mutation and its progeny that inherited the mutation. The complication comes because the founder and progeny are of unknown proportions and can take many arrangements, this effect increases in trying to determine the relationship in all cancer cell fractions.

However, clonal relatedness can be solved manually, by published algorithms, or recursively by testing all combinations of cancer cell fractions by inheritance shown in equation (1.1) [113].

Equation (1.1) states that the value of the observed cancer cell fraction Z, is the sum of the parent X_{CCF} that introduced the driver alteration and its subsequent progeny Y_{CCF}. Due to limitations in measuring clonal subpopulations, it can be seen that the sum of X_{CCF} and Y_{CCF} must be less or equal to Z, but not higher (unless a certain tolerance for noise is allowed). Thus, this condition has to be met for all clones captured, and this method can be used as a

quality control to check phylogeny reconstructions [101]. CloneEvol uses the inheritance property of eq. 1.1 to solve the relatedness of the measured ploidy corrected variant frequencies (CCFs) [113].

$$Z \approx X_{CCF} + \sum_{i=1}^{N} Y_{i_{CCF}} \tag{1.1}$$

Multiple approaches can be used to process the clonality signal. These fall into three broad categories, 1) use of pre-computed mutational clusters from tools like PyClone and ExPANdS to solve the deconvolution problem and reconstruct a phylogenetic tree [91, 92, 102, 113-116]. 2) performing simultaneous estimation of the cancer cell fractions and deconvolution of the phylogeny [117-121] or 3) applying a population genetics model to fit a set of initial conditions for tumour evolution to a given variant allelic frequency distribution [44, 87, 88].

In category 1) we have methods that use the mutation clusters predicted by tools such as PyClone or ExPANdS. Such approaches use combinatorial, likelihood or the inheritance methods as shown in equation (1.1) to establish a phylogenetic tree. Commonly used tools of this type are SCHISM [114], TrAp [115] and ClonEvol [113]. These tools provide the phylogenetic tree(s) that can be inferred from the sequenced sample, often producing numerous possible topologies.

In category 2) the methods are all-in-one, each with their own algorithms for identifying mutation clusters by correcting allelic fractions for ploidy, while at the same time applying a deconvolution model or similar to equation (1.1) to recover a phylogenetic tree. The limitation is that these are constrained to whole exome or whole genome sequencing assays. Popular tools in this category include PhyloSub [117], PhyloWGS [118], BitPhylogeny [119], Canopy [120] and SPRUCE [121].

All of these tools have specific requirements for their operation. Jiang et. al. [120], creators of the Canopy algorithm provide a good comparative summary of differences between the main tools for clonality and phylogeny assessment. Canopy, PhyloWGS and SPRUCE are good tools for estimating clonality and phylogenies when multiple samples are taken, the sequencing assay is whole exome and the tumours are driven by point mutations. Similar to category 1 methods, the phylogenies are based on the sequenced sample(s).

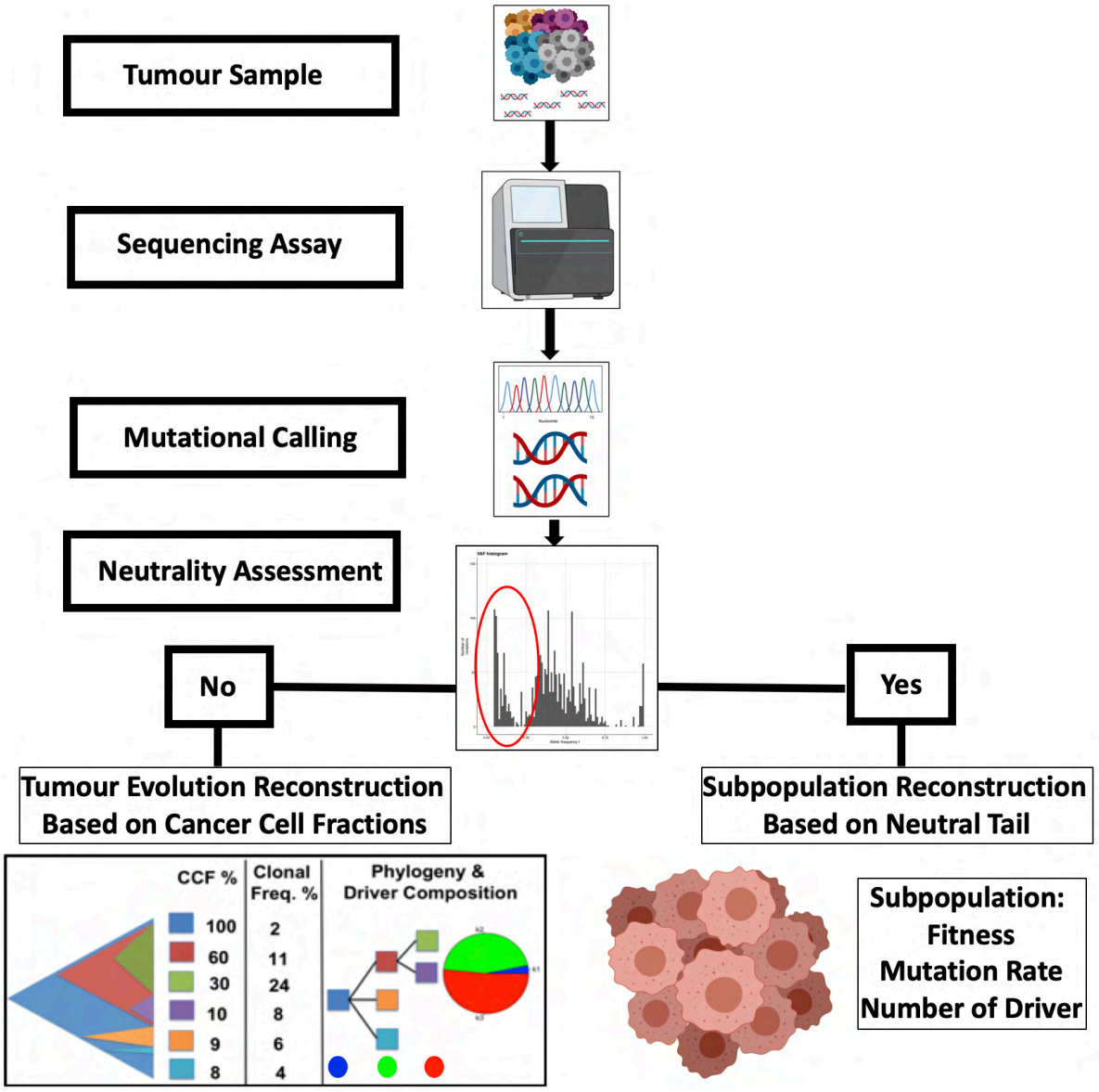


Figure 1.4 Outline of approaches for reconstructing tumour evolution. Tissue samples are collected and sequenced with a given assay, followed by mutational calling. After mutational calling is done, an evaluation of neutrality can be made. If the sample is neutral, information can only be obtained from the measured subpopulation. If the sample is not neutral, downstream analysis can be done to reconstruct its clonal evolution.

In category 3) we have the population genetics models. This is an emerging field that acknowledges that clonal evolution reconstruction can be severely affected by the sequencing assay performed. By the use of population genetics models their goal is to identify the parameters of the model (initial fitness and mutation rate) that recreate the observed variant allelic fraction distribution. With this information, inferences about the number of clones, their fitness and mutation rate can be made [44, 87, 88].

One criticism from the population genetics developers to data-driven approaches is the lack of evolutionary modelling integration in their algorithms. Specifically, how the so-called passenger tail can lead to false positive clones if not filtered [87]. This will result in more phylogenies and incorrect conclusions about clonal evolution.

Two approaches have been published so far. The first one uses the branching process with the infinite alleles assumption to parametrise tumour fitness and mutation rate to explain the distribution of variant allele frequencies. The model is geared by a Bayesian computation strategy [88]. The second one is a generalisation of this first approach. Two components of a mixture model are fitted to the observed variant allele frequency distribution: a neutral signal using a Pareto Type-I distribution, and one or more positive selection signal(s) using the Beta distribution(s). The two components define a mixture model that is inferred by an expectation maximisation likelihood framework, to identify the parameters of the mixture that best describe the input data [87].

The potential of the branching process to provide an evolutionary history of tumours when data is missing is the central topic of my thesis and will be discussed in detail in subsequent chapters. My approach for reconstructing tumour evolution is similar to that of Williams et al. [88] as it uses the mathematical branching process forward in time. However, I used the simulated driver cancer cell fractions to approximate the evolutionary history of real tumours by comparing to pre-computed cancer cell fractions obtained from a clonality caller (e.g. PyClone, SciClone or ExPANdS).

6 Birth and Death Process

Although there are significant mathematical models for tumour progression, my focus is on only the branching process and its potential to reconstruct tumour evolution using sequencing data [89, 122-130].

The backbone of the branching process is the birth and death process. It aims to define the events of interest at the single cell level occurring at the same time at generation t. In the context of tumour evolution, the interest is in how selection influences cellular growth. Thus, at a given time t, cell division occurs with probability b and cell death with probability $d = 1 - b$.

In neoplastic growth, the main interest is when $b > d$ and the conditions that maintain or increase it, which are the accumulation of driver alterations occurring with probability u. Accounting for driver alteration events the birth/death process has the following outcomes: successful division defined with probability $b(1 - u)$, death as $d = 1 - b$ and division with driver alteration as bu, as shown in Figure 1.5.

It is worth mentioning that u explicitly models the alterations that increase fitness; other authors have explored the opposite. How the accumulation of alterations (e.g. passenger mutations) reduces cellular fitness is not going to be considered here [122, 130].

The model is flexible enough to incorporate neutral events of interest as shown in Figure 1.2. For instance, Chowell et al. [44] included drug-resistance with neutral fitness in primary tumour formation. Mutations were assumed to occur with probability $\mu_r = 10^{-8}$ and were used to estimate the number of pre-existing drug-resistant clones at diagnosis.

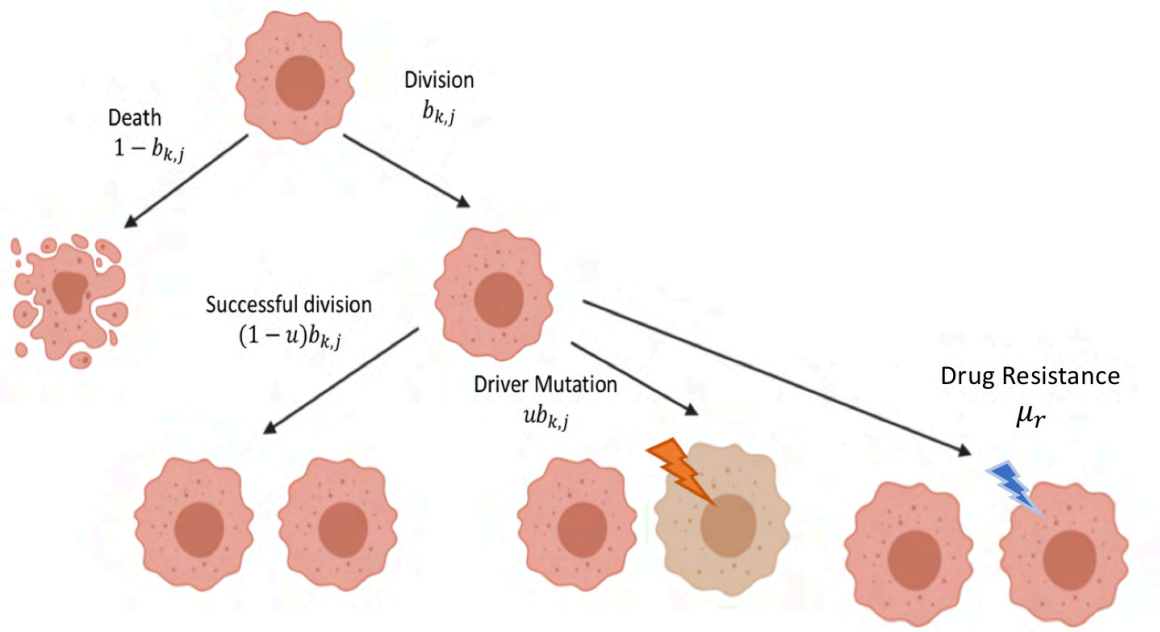


Figure 1.5 Birth and death process. At a given discrete time t, division and death are modelled with given probabilities. Divisions are then evaluated to describe the events of successful division or division with a driver mutation. Neutral fitness mutations can be incorporated at given rates such as drug resistance.

With the birth and death process authors have modelled clonal expansions as a function of the passenger/driver makeup by evaluating every single cell [131]. This results in a considerable computational cost because at every time t, all cells individually have to be updated. To bypass this problem, Bozic et al. [43] scaled the evaluation of every cell to subpopulations by sampling from a multinomial distribution, resulting in a faster computation allowing for easier implementation.

7 The Discrete-Time Branching Process

The branching process has great flexibility to model genotypic heterogeneity [132-136], phenotypic heterogeneity and environmental fluctuations. The scope is going to be restricted in using the discrete-time branching process reported by Bozic and Chowell [43, 44] to study tumour evolution. This means that the environment modelled remains constant throughout evolution, and the average selective advantage conferred by driver alterations is a parameter and not a variable.

Bozic et al. [43] implemented the discrete-time branching process, expanding the birth and death model to describe the number of distinct subpopulations that have accumulated *k* drivers. By analysing the passenger-driver ratio in the human tumours, they concluded that the average selective advantage conferred by every driver accumulation is low, $s = 0.004 \pm 0.0004$.

With the discrete-time branching process authors were able to predict the growth of familial adenomatous polyposis in two clinical cases [43]. Despite the success of the branching process a comprehensive evaluation to predict growth in multiple malignancies is lacking.

Chowell et al. [44] recently expanded Bozic's [43] to describe the clonal composition of tumours. Although both models have the same mathematical construct, they have different

outputs and complexities. One models *k*-driver subpopulations and the other clonal subpopulations highlighted in the middle panel of Figure 1.6.

As shown in Figure 1.6 top panel, Bozic [43] and Chowell [44] models are rooted in the birth and death process. Both models describe tumour progression in different ways, Bozic by tracking *k*-driver subpopulations and Chowell by tracking individual clones. These methods for updating cumulative fitness are interchangeable in both branching processes, as indicated by the dashed arrows in Figure 1.6 medium panel. In the Bozic model [43] changes in fitness are additive, this means that fitness increases by a constant value for every driver alteration as depicted by a straight blue line in the bottom panel left plot. In the Chowell model [44] fitness changes were implemented by the stickbreaking strategy, which samples from an exponential distribution parameterised by the parental fitness, depicted by a smoother curve in the bottom panel right plot.

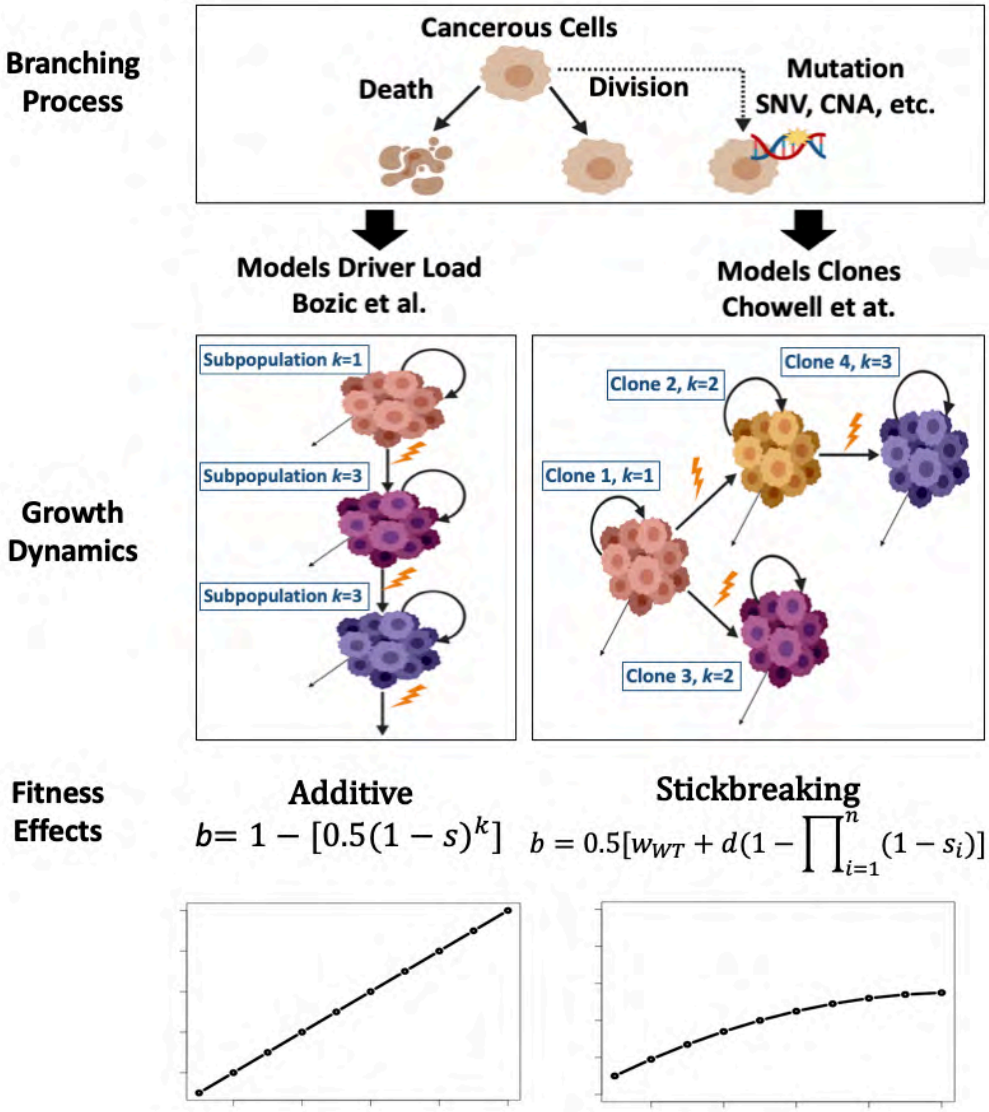


Figure 1.6 Discrete-time branching process to reconstruct tumour evolution. Top panel: the birth and death process leading to the Bozic and Chowell models. Medium panel: the differences between the Bozic and Chowell models, one implementing k-driver subpopulation abundance and the other clonal subpopulations. Bottom panel: the difference in fitness updates that can be interchangeable in the models. Plots indicate the characteristic fitness trajectories of additive fitness model vs the stickbreaking model.

As illustrated in Figure 1.6, growth and inheritance dynamics should be incorporated. The goal of the growth dynamics is to describe how the number of cancerous cells changes over time. It is necessary to generate a time-dependant function for tumour growth, limited by a carrying capacity if necessary. As mentioned, the function can be modified to model individual cells, sub-clonal populations or any mutational process interest (epigenetic alterations, copy number changes, etc.).

The inheritance dynamic should describe how a given driver alteration k changes the rate of growth by modifying s from parent to progeny. For its implementation it is necessary to define a fitness effect function such as the additive fitness or stickbreaking model, as discussed below in Section 9. The effect of the fitness gain is depicted in Figure 1.7 as $f_2(t) > f_1(t)$.

Time is a key component of modelling tumour evolution and essential in deriving analytical solutions. Of special interest is τ, the time interval during which new progeny arise and successfully expand, as growth of the progeny can be simplified from that point onwards. τ is also relevant as a proxy of clonal competition, provided there is enough time for progeny to outcompete their parents.

As noted by Bozic et al. [137] new data suggests that certain tumours have more pronounced fitness effects suggesting that the additive fitness model may be underpowered in certain scenarios. The stickbreaking fitness model is a relevant candidate to mitigate said effect as it allows for variation of the selective advantage of every new k driver alteration [44, 138].

The main difference in both models is that the additive fitness model scales the based on the founder fitness whereas the stickbreaking scales based on how close is the founder fitness to the hypothesised upper boundary. Chowell et al. [44] implemented the stickbreaking model by sampling from an exponential distribution of fitness effects of mean $\bar{s}$ (0.1, 0.01 & 0.005) and then adjusted to the hypothesized upper boundary (assumed to be *1*). Sampling from the exponential distribution allows a differential effect in the selective advantage which in turn can explain the emergence of hyper selected clones (big leaps in fitness acquisition are considered with low probability).

8 Growth Dynamics in The Discrete-Time Branching Process

It is necessary to define growth dynamics in terms of the selective advantage/net growth s and in the context of proliferation rate b and death rate d. In general, $s = b - d$ where $b + d = 1$ and fitness $w = 1 + s$. These parameters should produce a unique growth function for a given clone or subpopulation based on its value of s or w.

Figure 1.7 is an example of an implementation of growth dynamics. A founder subpopulation expands according to $f_1(t)$. At time τ subpopulation $f_1(t)$ acquires a driver mutation leading to faster growing subpopulation $f_2(t)$. The total tumour growth $N(t)$ is the sum of the two expanding subpopulations.

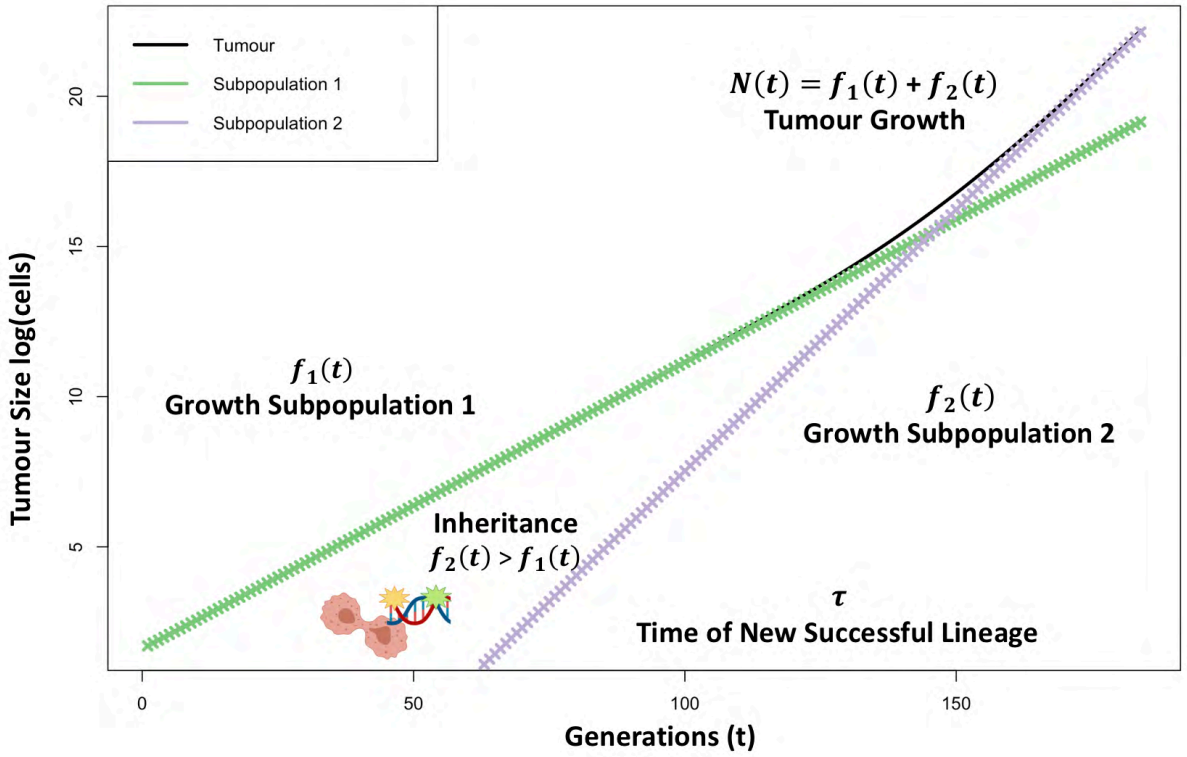


Figure 1.7 Growth dynamics. Subpopulation 1 has a net-growth given by *s* coloured in green. When a cell acquires a driver alteration at time τ, the net-growth in subpopulation 2 increases relative to subpopulation 1, eventually overtaking its parent. The total tumour size is the sum of both subpopulations.

Time is relative to number of generations or cell divisions occurring at discrete time t. However, if the average division rate P is known, then the generations can be scaled accordingly to reflect days, months, years or the time-unit of interest.

Figure 1.8 shows a schematic representation of the Bozic [43] and Chowell [44] models with their expansion equations. It can be seen that the Chowell [44] model increases the computational cost rapidly by tracking every clonal subpopulation as compared to the Bozic [43] model. However, the simplicity of the Bozic [43] model does not allow for direct comparison to sequencing data.

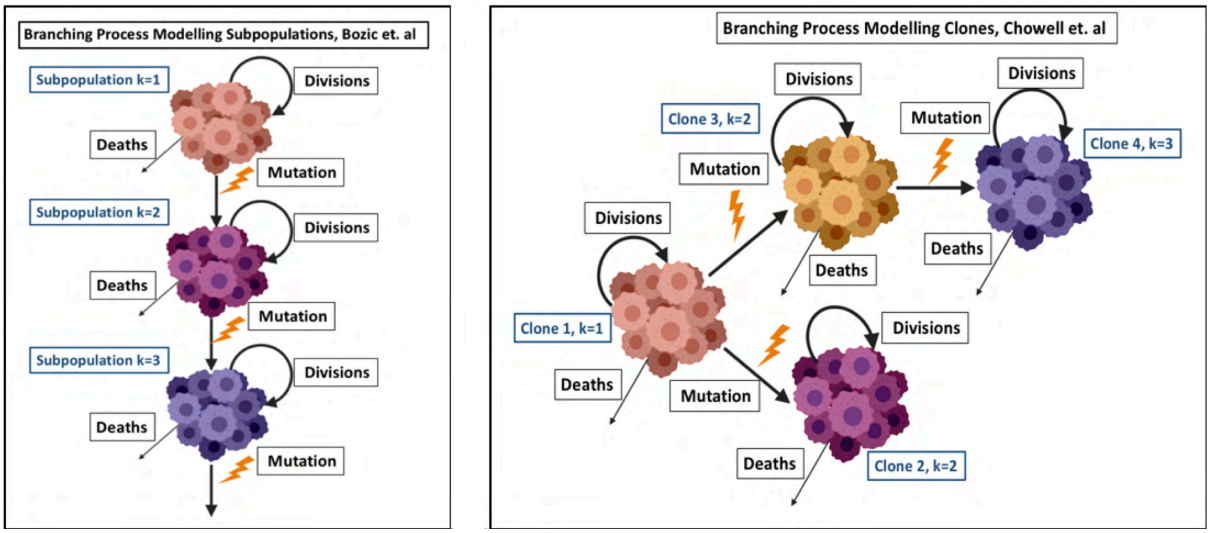


Figure 1.8 Growth dynamics in the discrete-time branching models. The Bozic model accounts for births, deaths, mutants coming from its predecessor and mutants leaving by the accumulation of an extra driver. This is accounted for every subpopulation *S*. The Chowell model accounts for births, deaths, and driver mutants becoming a new clone *C*. This subtle difference allows the Chowell model to track every clone directly, but increases the computational cost.

In each model, the sizes of individual subpopulation or clones are updated as follows:

Bozic et al. [43] $$S_k(t+1) = S_k(t) + B_k - D_k + M_{k-1} \quad (1.2)$$

Chowell et al. [44] $$C_{k,i,j}(t+1) = C_{k,i,j}(t) + B_{k,i,j} - D_{k,i,j} \quad (1.3)$$

Where $S_k(t+1)$ refers to the size of a subpopulation with k drivers being evaluated at a given time t, S_k is the result of the births B_k, deaths D_k and driver mutants coming from predecessor M_{k-1}. Similarly, $C_{k,i,j}(t+1)$ refers to a clone being evaluated at a given time t containing k driver mutations coming from parental lineage i and progeny j. It is worth mentioning that symmetric division is assumed and values M_k and $M_{k,i,j}$ are not required for equations 1.2 and 1.3.

As reported by Bozic [43], updates to random variables B, D, M are taken from instances of a multinomial distribution with parameters defined by the following equations.

[43] $$[B_k, D_k, M_k] \sim Multinom(S_k(t), [b_k(1-u), d_k, b_k u]) \quad (1.2a)$$

[44] $$[B_{k,i,j}, D_{k,i,j}, M_{k,i,j}] \sim Multinom(C_{k,i,j}(t), [b_{k,i,j}(1-u), d_{k,i,j}, b_{k,i,j}u]) \quad (1.3a)$$

9 Incorporating Driver and Passenger Mutations in the Discrete-Time Branching Process

In the branching process, mutations can be drivers or neutral, as outlined in Figure 1.9. Bozic [43] applied a driver mutation rate of $u = 3.4\text{x}10^{-5}$, which describes the rate of emergence of new clones, depicted with unique colours in Figure 1.9.
The driver signal is only affected by driver alterations and their fraction corresponds to the number of cells within the clone. For example, the blue clone in Figure 1.9 has only 4 cells which corresponds to a fraction of $4/25 = 0.16$. Thus, in this model of clonal evolution, the driver signal is only affected by frequencies of the clones that carry the drivers and the progeny

of the clones that inherit those mutations. In this example, the dissemination of driver mutations is indicated with letters with the most prevalent mutations being, in order, A (grey) first, followed by B (blue) then C(green) and finally D (orange).

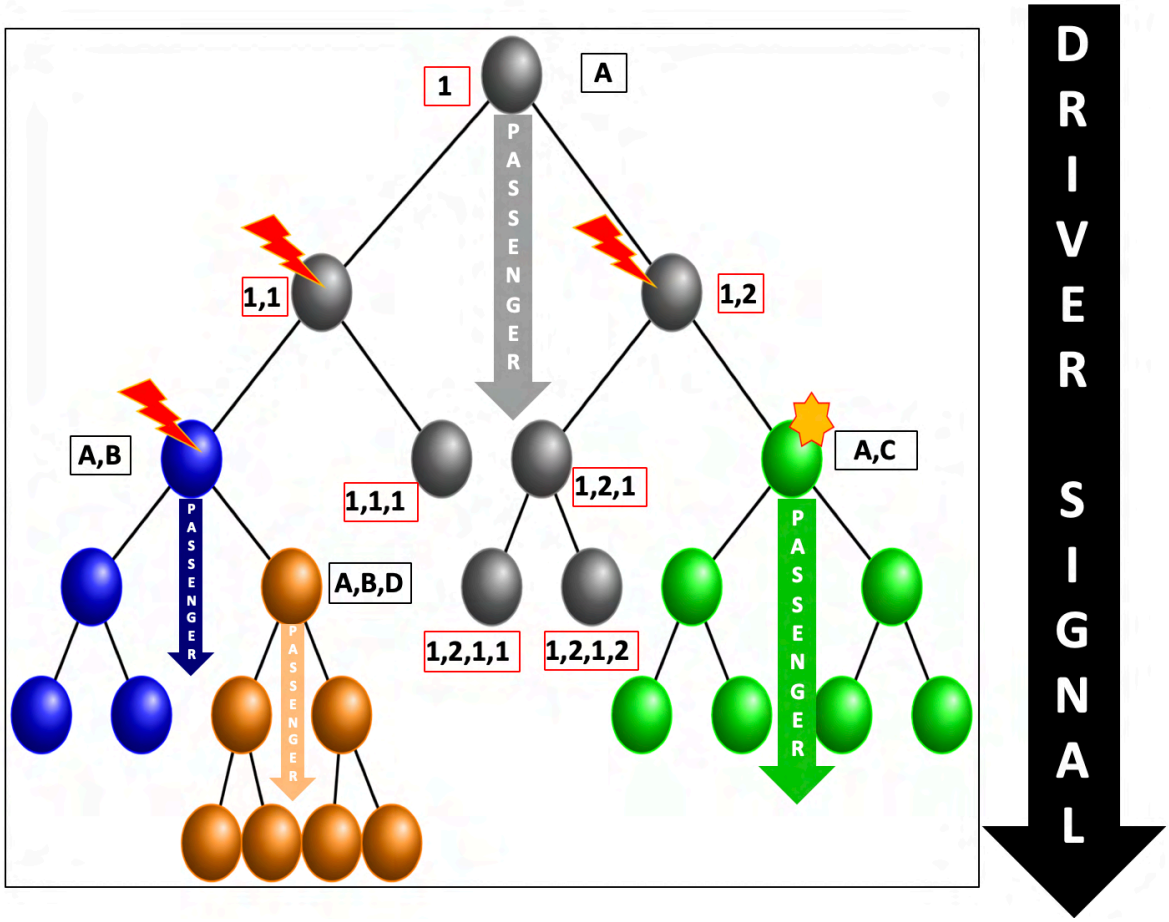


Figure 1.9 Driver and passenger signal accumulation. Driver mutations are indicated with letters and neutral mutations by numbers, with clones coloured according to the set of driver mutations they carry. The driver signal is generated by the accumulation of driver alterations, while the passenger signal occurs stochastically in every clone as exemplified with numbers in the grey clone. Neutral fitness alterations such as the yellow star one can be explicitly modelled to account for the prevalence of drug resistance mutations.

The passenger signal has neutral fitness and is reported at a rate of $\nu = 0.016$ occurring in all cells within a clone, denoted by numbers in Figure 1.9. This signal is responsible for the so-called neutral tail in the variant allelic frequency distributions. Due to the high passenger mutation rate the passenger signal can bias clonality tools by introducing noise or false positive clones [87] producing incorrect estimates of clonal frequencies.

Because the passenger accumulation occurs within a clone with given fitness, the signal of the passenger tail is often portrayed as an exponential decay function. If multiple clones are present the overlap in their passenger signals create a mixed signal. In this example the passenger prevalence of the grey clone is *1*; *1,2*; *1,2,1* etc. Similar signals will emerge in subsequent clones. The branching process can approximate the median number of passengers by recording the value of d/b of every driver clone with the framework provided by Bozic et al. [58].

As mentioned, passenger mutations with potential to confer drug resistance can be incorporated into clones at a given rate. In Figure 1.9 this is exemplified by the yellow star mutation in the green clone. Therefore, besides modelling clonal evolution by the accumulation of driver alterations, it provides an estimate of the number of drug resistant cells and how the fitness distribution will change if the environment changes by treatment.

The reported range of pre-existing drug resistance probability at cell division ranges from 10^{-8} - 10^{-9} which is measured by the number of mutated loci that can confer a drug resistance trait [45, 139-141] . However, tumours have different first lines of treatment involving different drug agents with diverse action mechanisms introducing variability in the tumour-specific drug resistance rate. In other words, it is expected that tumours solely treated with chemotherapy have a different drug resistance rate than those only treated by targeted therapies or immunotherapies. Bozic et al. [123] derived analytical solutions for the average number of resistance clones and the median number of drug resistance cells in a clone with k drivers that can refine the results reported by Chowell et al. [44]. In other words, it can model the abundance of drug resistance cells that emerge at a given rate μ_r without simulations. Other authors provide complementary analytical solutions on the number of drug resistance cells considering the effective loci [45] that can be useful in cases of malignancies treated with a particular type of anticancer agent.

10 Inheritance Dynamics in the Discrete-Time Branching Process

Bozic [43] and Chowell [44] models also differ in the way accumulated driver mutations alter fitness. Bozic modelled by additive fitness, this means that fitness changes are proportional to number of accumulated drivers k [43]. The additive fitness is then,

Bozic [43]

$$d_k = \frac{1}{2}(1-s)^k \qquad b_k = 1 - d_k \tag{1.4}$$

As a result, as the number of drivers k increases, d_k is reduced inducing stronger and more frequent selective sweeps. However, additive fitness effects have been criticised for their lack of consistency in recreating the patterns observed in experimental evolution [138].

For that reason, Chowell implemented in the clonal branching process the stickbreaking strategy [44], where fitness effects are scaled by the distance of the current fitness and a hypothesized upper fitness boundary (assumed to be 1) [138]. With this adjustment, the stickbreaking deviates from the additive and multiplicative strategies enabling for a better sampling of the fitness landscape, capable to replicate experimental evolutionary patterns [138].

The definition of the stickbreaking model in the context of clonal evolution provided by Chowell [44] is defined as following,

Chowell [44]

$$b_{k,i,j} = \frac{1}{2}\left[w_{wt} + a\left(1 - \prod\nolimits_{k,i=1}^{n}(1 - f_{k,i,j})\right)\right] \qquad d_{k,i,j} = 1 - b_{k,i,j} \tag{1.5}$$

Where w_{wt} is the fitness background, which is assumed to be 1, and a is a scaling factor for the upper boundary, which is also assumed to be 1. H. Allen et al. [142] showed that the distribution of fitness effects is exponential, meaning for every mutation the fitness effect has to account for this as $f_{k,i,j} \sim exponential(s_{k-1,i,j})$. Then it has to be scaled by the fitness background as shown in the previous equation. The sampling effect in the stickbreaking model makes fitness effects vary from simulation to simulation but converge to a similar trend, as shown in Figure 1.10.

Figure 1.10 presents a side by side comparison of the fitness effects in both models. The left-hand side shows how under additive fitness effects, fitness increases are proportional to the number of accumulated drivers k, resulting in a straight line. Although it makes tracing fitness changes in the tumour tractable, it provides a narrow fitness landscape by limiting the number of fitness combinations possible [142].

The right-hand side shows how the stickbreaking model of fitness works. First, fitness effects are sampled from an exponential distribution with mean s, adjusted by the fitness background and a hypothetical upper fitness value. This results in richer combinations of fitness effects that allow for different clonal dynamics.

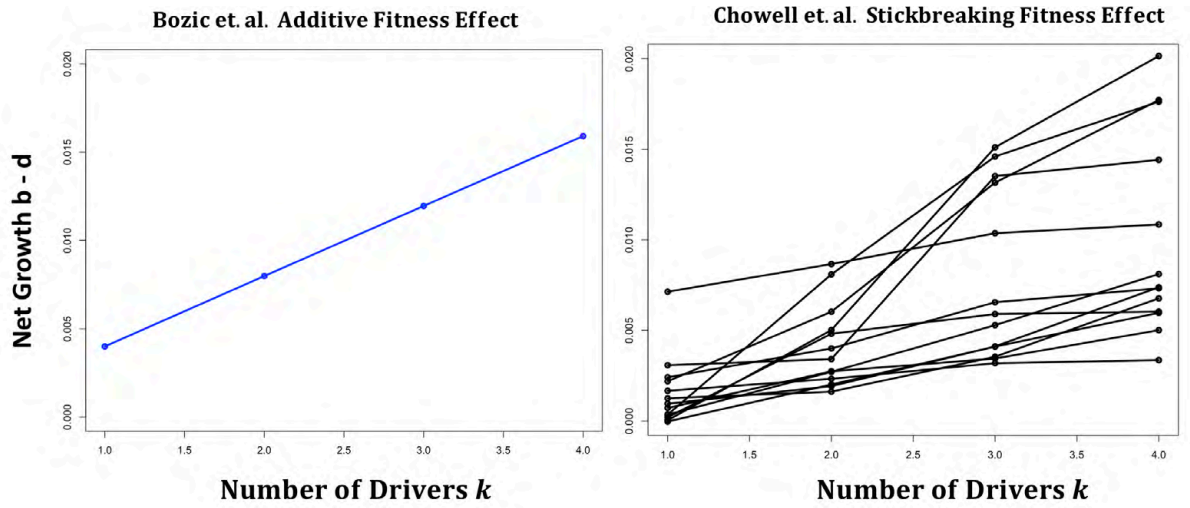


Figure 1.10 Fitness effects. The left-hand side shows the additive fitness effect used in the Bozic model showing that fitness increases are proportional to k. The right-hand side displays 15 samples of the stickbreaking model used by Chowell, showing the different trajectories in fitness accumulation that result in a richer fitness landscape. Both plots used the same starting value of $s = 0.004$.

11 Assumptions of the Discrete-Time Branching Process

As with all models, branching process models come with assumptions and limitations that have to be accounted for to ensure their correct implementation:

1. The parameters s and u are averages.
2. Infinite alleles assumption: every driver alteration leads unique subpopulation/clone that has not appeared before.
3. The exact type of driver alteration does not matter as long as u captures the overall average rate.
4. Only driver mutations can induce clonal sweeps or fitness changes.

5. Passenger mutations are neutral and not directly modelled.
6. Carrying capacities are not modelled nor are spatial or microenvironmental pressures.
7. The tumour is initiated by as single founder subpopulation/clone.
8. Can model epistasis by parametrising s and w_{wt} and a in the stickbreaking model.

12 Making Inferences About Tumour Evolution Using the Discrete-Time Branching Process

Relevant mathematical derivations have been developed using the continuous time branching process. Some of their formulations can be applied in the discrete case, such as expected subpopulation growth and waiting times for clonal lineage to emerge [42, 143-145]. However, the continuous time branching process is outside of the scope of my thesis.

Bozic et al. [43] validated the model qualitatively in two clinical studies in patients with familiar adenomatous polyposis. They were able to recover the age of the patients at clinical diagnosis, the number of polyps and the average diameter of said polyps.

Subsequent use of the branching process aimed to model the accumulation of drug resistance mutations [45, 136, 140]. In this study [146] authors aimed to estimate if resistance of EGFR-blockade is pre-existing or not. Using the branching process, they concluded that the pre-existing drug resistance cells are present prior to therapy.

Analytical solutions of pre-existing resistance has been made using the branching process by the evaluation of the total number of cell divisions and the extinction probability δ [45, 123, 136].

Likewise, Chowell et al. [44] used the clonal branching process to evaluate the emergence of drug resistant clones in primary tumour formation, and similarly identified that pre-existing drug resistant cells are present prior to treatment. They showed how the number of drug resistant cells is influenced by the starting value average selective advantage s. Lower values of s increased the number of drug resistant cells.

The branching process was also used to predict the number of clonal and subclonal passenger mutations [58, 88]. The authors derived analytical solutions, showing that the sequencing depth and coverage affects not only the measurability of the number of passengers, but also the lineage extinction probability δ. Their framework was able to fit the passenger tail of different cases in TCGA, providing the median number of clonal and subclonal passengers.

Analytical solutions of the expected clonal growth using the branching process were provided by Reiter et al.[147]. They used a simplified version of the branching process to evaluate clonal evolution considering two driver subpopulations. Their framework provides a baseline to solve the branching process analytically.

The branching process has been used to study metastasis[126]. In this study [148] the growth and dynamics of pancreatic cancers was evaluated based on how SMAD4 mutations conferred metastatic potential. They were able to identify the probability of clonal dissemination conducive of metastasis in 228 patients. This approach is similar to the one derived by Durrett et al. [42] by considering non-metastatic cells as a type-0 branching process with given parameters of expansion, and metastatic potential cells as type-1with fitness parameters greater than type-0. Once type-1cells successfully metastasise, they evolve as type-2. By accelerating

their expansion by a cumulative fitness increase, they provided estimates of the probability of type-2 cells at diagnosis and how to determine their sizes. In this study, the branching process was used to determine metastatic seeding patterns and the authors developed a tool to recover the phylogeny of metastasis and links to tissue specific lesions [149].

In the context of using the branching process to reconstruct tumour evolution, Williams et al. [88] is the only study that applied the branching process to predict the variant allelic frequency distribution in multiple cancer subtypes used as input in their approach. Their model accounted for the passenger tail and mutations under positive selection, and fit both parts simultaneously. This is the only approach that is able to use all the information of the variant allelic frequency distribution to infer the number of clones, tumour fitness and their mutation rate.

13 Knowledge Gaps and Contribution to the Field

During the last decade, the field of oncology has acknowledged the importance of reconstructing tumour evolution and its contribution in clinical outcome. To this end flexible and robust methods that allow computational modelling to be direct linked with sequencing and clinical data in meaningful ways are required.

Current approaches are data-driven and rely on multiple samples and whole exome/genome assays, making them difficult to deploy in longitudinal studies. The inference made by those tools can be biased by the interference of the passenger signal, affecting the adequate reconstruction of clonal evolution [87].

Many potential applications of clonal evolution models of the type reported by Chowell [44] remain unexplored. They can empower clonal evolution reconstruction using sequenced tumours and provide evolutionary histories at the patient level. This new path has potential to provide better understanding of tumour development, resistance and recurrence.

In addition, such an approach could also aid in the design of treatment regimens by simulating how treatment affects the evolutionary trajectory and how scheduling of dosing could be improved. This could be scaled up by simulating clinical trials and fitting predictions from simulated tumours to sequencing data collected in real-world trials.

In this thesis I will explore the potential of the clonal branching process for reconstruction of tumour evolution in a clinically meaningful context.

Justification, Hypothesis and Aims

Cancer is a main killer globally [150, 151]. There has been significant progress in therapeutic design and understanding cancer evolutionary dynamics. However, how best to measure tumour heterogeneity and the root causes of therapeutic failure and resistance remain open challenges.

With the advent of high-throughput sequencing technologies, large scale sequencing studies have been deployed to study most types of cancer, uncovering the mutational landscapes of multiple malignancies and identifying key drivers of tumorigenesis. Despite the widely recognised relevance of reconstructing clonal evolution to establish its association with prognosis and resistance, most sequencing projects are not well powered enough to recover the clonal architecture of tumours.

Therefore, novel approaches are required to link computational modelling, population genetics and genomics data with clinical data. Combining these elements will provide a testable framework to recover tumour evolution to provide a picture on tumour progression and the different evolutionary paths of sequenced tumours. Connecting clonal evolutionary paths with clinical manifestation allows the identification of conditions that associate with malignancy and help the development of predictive and prognostic models.

The robustness and flexibility of the clonal branching process makes it a strong candidate to provide tools that can meet these goals. Consequently, the ultimate goal of this project is to evaluate the potential of the clonal branching process as a testable framework to reconstruct tumour evolution.

Hypothesis: The discrete-time branching process can be used to infer the evolutionary history of real tumours. It presents a flexible and testable framework that connects mutational process with growth dynamics and provides biological insight and clinical value.

Aim I: Derive analytical solutions for the additive fitness branching process in the k-compartment and clonal configurations.

Objectives:

- Calculate expectation and variance of clonal subpopulations and total tumour size at any time t.
- Determine the detectable driver composition at a given tumour size N.
- Determine detectable phylogenies and number of clones for different values of s and u.

Aim II: Modify the clonal branching process to account for changes in the average driver mutation rate and driver selective advantage to study how these influence simulation outcomes.

Objectives:

- Implement and generate representative samples from three versions of the clonal branching process with: 1) additive fitness, 2) stickbreaking fitness and 3) additive fitness paired with accelerated driver mutation rates.
- Establish the effect of the initial values of s and u on simulation outcomes, such as expansion times and degree of diversity.
- Establish likely clonal and driver compositions of simulated tumours.

- Using simulated tumours with clones greater or equal than 10% cancer cell fraction, evaluate the degree of overlap and phylogenies between parameter combinations of s and u.
- Establish the detectable clonal and k-driver compositions tumours when only clones greater or equal than 10% cancer cell fraction can be measured.
- Study the distribution of drug resistance cells and driver makeup between parameter combinations of s and u.
- Implement a neutral evolution model to study how the passenger signal is accumulated for different values of s and k.

Aim III: Identify a robust statistical method for comparison of cancer cell fractions from simulated and sequenced tumours.

Objectives:
- Compare distribution-free with minimum distance methods to fit cancer cell fractions under different sources of bias.
- Use a sample of the models in Aim II to generate a truth set and a test set to benchmark the statistical methods.
- Compare the methods by adding noise and removing information to replicate the biases that occur in measuring clonal subpopulation with sequencing.
- Identify the best performing method to accurately identify parameters s and u and the correct simulation.

Aim IV: Apply the branching process models to reconstruct tumour evolution in clinical sequencing studies using the statistical method identified in Aim III.

Objectives:
- Reconstruct tumour evolution in TCGA, TRACERx NSCLC, BIG 1-98 and CASCADE melanoma.
- Evaluate how the reconstructions associate with clinical variables and clinical outcome (distant recurrence or death).
- Recover tumour phylogenies and identify differences between subtypes and clinical outcomes.
- Associate the mutational events in the phylogeny with tumour growth and predicted number of drug resistant cells.

Chapter II

1 Background

As reported by Bozic et. al. [43] and Chowell et. al. [44], the discrete-time branching process is a computational stochastic model capable of predicting the clonal composition of tumours, yet a pan-cancer validation is needed to determine its clinical impact. Although the value of mathematical modelling in reconstructing tumour evolution is widely accepted, a framework to align it with clinical and genomic data for evolutionary inference remains lacking. Evolutionary inference of clonal evolution provides estimates of tumour expansion rates and degree of clonal diversity and connects mutational processes with growth patterns. Determining the proportions of clonal subpopulations within tumours and their phylogenetic structure allows identification of molecular vulnerabilities such as truncal mutations that may be targeted with current therapeutics.

In this chapter I will provide analytical approximations of the computational branching process of subpopulation/clonal growth, timeframes for successful clonal lineages to emerge and provide estimates of driver frequencies and clonal composition. I will discuss the relevance of the analytical solutions in the context of clinical applications and the conditions where analytical solutions are preferred over computational simulations. I will show the expected clonal makeup and phylogenies that can be recovered as an example of the use of the analytical solutions to recover evolutionary properties of tumours. Applicability limitations by model assumptions are going to be discussed and how to identify future expansion for the models.

2 Outline

First, I will derive analytical solutions of the expectation and variance of the *k*-th driver model reported by Bozic et al. [43] and the clonal branching process, similar to Chowell et al. [44] but using an additive fitness update for simplicity.

Second, using the analytical solutions of the Bozic model [43], I will estimate the measurable driver tumour composition assuming a representative sample of the tumour (e.g. rep-seq [93]) in which subpopulation detection is proportional to its size.

Third, I will estimate the waiting times for any given clone to emerge using estimator $\hat{\tau}_{k,i,j}$ to derive analytical solutions for the clonal model with additive fitness.

Fourth, using the derived analytical solutions, I will estimate the measurable clonal tumour compositions predicted by the clonal model assuming a representative sample of the tumour, and how that information is presented in clonality assessment tools. I will report the expected number of detectable clonal subpopulations under different values of average selective advantage s and average driver mutation rate u at milestone tumour sizes 1 cm^3 and 4 cm^3.

Finally, I will describe the expected phylogenies from a different combination of parameters and how to align these findings with clonality tools for tumour evolution reconstruction in patients.

3 Introduction

Intratumour heterogeneity in cancer has many layers of complexity, requiring a comprehensive approach to connect the biological determinants of tumour progression with a testable framework that links experimental, clinical and sequencing studies in a useful way. As such, mathematical evolutionary modelling has the potential to bridge this gap by means of using the branching process model of tumour evolution to predict the growth patterns and clonal compositions of diverse human malignancies. However, an in-depth validation and evaluation to its clinical application remains lacking.

As shown in Chapter I, once growth and inheritance dynamics have been properly implemented for the branching processes, inferences about tumour growth can be made. Branching process models can be used to explore population-specific properties and their impact in total tumour growth as well as to estimate driver and passenger diversity.

Figure 2.1 shows the difference between simulations from two versions of the branching process run with the same initial conditions of $s = 0.004$ and $u = 3.4\text{x}10^{-5}$. The Bozic model [43], on the left-hand side, is driven by subpopulations with 2 and 3 drivers (coloured in green and dark blue, respectively). Moreover, it shows the emergence of subpopulations with higher number of drivers that expand during the later stages of tumour development.

The Chowell model [44] agrees with the Bozic model [43], resulting in a tumour mostly composed by three 2-driver clones ($C_{1,147}$; $C_{1,141}$; $C_{1,149}$,) and one 3-driver clone ($C_{1,147,35}$). However, the advantage is that it provides the clonal relationship that can recover the tumour phylogeny providing the exact prevalence of a mutation or genotype of interest.

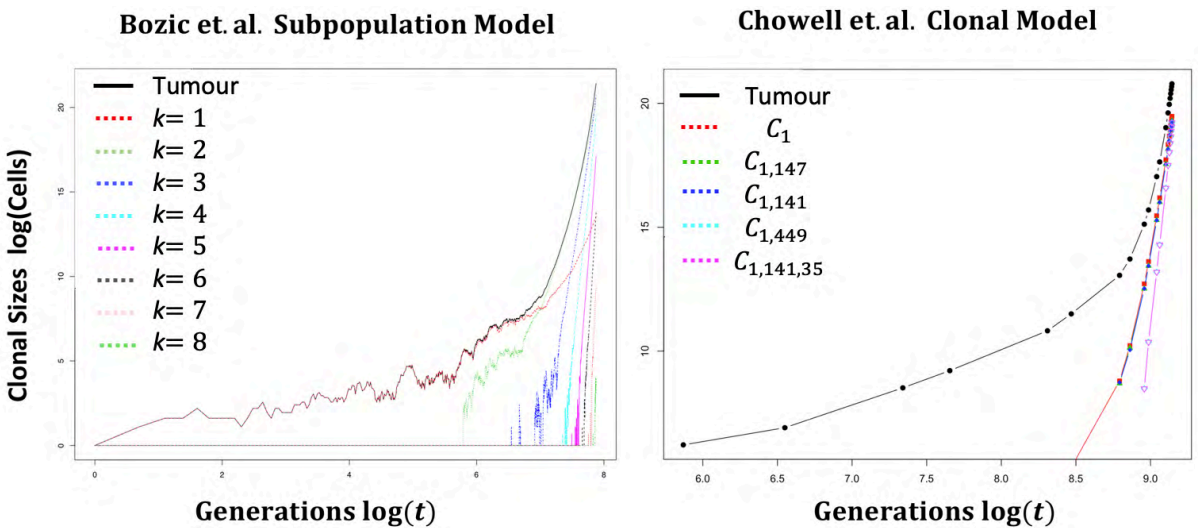


Figure 2.1 Growth dynamics of two branching process models. Both models used the same initial conditions, $s = 0.004$ and $u = 3.14\text{x}10^{-5}$. Left panel shows the Bozic model, with the black line indicating total tumour size, calculated as the sum all subpopulations. Individual subpopulations are displayed with uniquely coloured dashed lines. Right panel show the Chowell model, displaying the growth of just the selected subpopulations. The black solid line represents total tumour size as the sum of the clones. Individual clones are depicted by different colours.

As shown in Figure 2.1 both branching process models display similar properties given the same initial values of s and u. My interest is in approximating the main behaviour of the model without simulations. The goal is to determine the possible ranges of s and u in cancer that can be applied to different stages of disease progression. Although Bozic et al. [43] suggest the average selective advantage s of a typical cancer driver mutation is $\sim$0.004, other studies indicate that certain oncogenic drivers can amplify selection [88, 152-154]. Therefore, exploring the parameter combinations will provide a sense on the range of potential values to use for subsequent simulations and analyses.

A considerable amount of sequencing studies and initiatives such as the TCGA have collected single snapshot samples for tumours, but due to spatial heterogeneity single samples may be underpowered to represent the tumour architecture and their mutational prevalence [90]. Clinical markers can be superior or complementary to sequencing in establishing the current state of the disease requiring investigation on how to integrate them into the branching process [155].

A further aim is to evaluate the use of analytical solutions of the branching process to provide a global view of the tumour when limited genomic information is available. What information can be provided by clinical markers to the analytical model in order to gain evolutionary knowledge is an outstanding question in the field.

In this chapter, we detach the modelling from the sequencing perspective to learn the main behaviour of the tumour with different combinations of s and u. Subpopulation or clonal detectability are subject to minimal fraction of the tumour. This is only reflective when an ideal sample of the tumour is collected which is achievable in certain conditions (e.g. [93, 155]). The objective is to replicate conditions where no sequencing or genomic information is available and information about diversity has to be made by other means, such as clinical and histopathology markers.

Instead of solving the clonal model of Chowell et al. [44] with the stickbreaking algorithm, I will use the k-subpopulation model and its expansion into the clonal model as these two models provide different granularity that can be comparable (e.g. both can provide the k-driver load).

4 Hypothesis and Aims

Hypothesis: The branching process can be used to establish the expected clonal composition of tumours. Analytical solutions of the branching process alleviate the computational requirements of extensive simulations to characterise tumour dynamics.

Aim I: Establish the driver content of tumours for different combinations of s and u by:

- Deriving analytical solutions from the model of Bozic et al. [43].
- Apply the analytical solutions to identify the k-driver tumour content simulating the conditions of an ideal tissue collection scenario.

Aim II: Establish the detectable clonal composition of tumours and tumour phylogenies for different combinations of s and u by:

- Deriving analytical solutions of the Bozic et al. [43] and a clonal model with additive fitness.
- Recursively estimating the time of detectable clonal sweeps, constrained by a tumour fraction threshold to recover tumour phylogenies.

5 Methods

The expectation and variance of the Bozic et al. [43] were developed by using the reported solutions for the discrete-time branching process for founder subpopulation $S_{k=1}$ as a baseline [147, 156]. The expectation and variance for any subpopulation S_k represents the waiting time for a successful lineage to expand from its parent S_{k-1} at time τ_k as illustrated by Bozic et al. [43] and shown in Fig. 1.7. Solutions are provided in the corresponding sections and further details are on the Appendix A.2.1. A deterministic solution for the expectation is proposed in Appendix A.2.2.

All clones in the clonal model with additive fitness have the same mathematical form as $S_{k=1}$, though the waiting times for any clone to emerge $\tau_{k,i,j}$ have to be approximated. I used the definition of Cheek et al. [145] who showed that the mutational events occur as a Cox process to define estimator $\hat{\tau}_{k,i,j}$. This enables to approximate the time of emerging clones considering the accumulated probability of new driver clones as $b_k u[b_k(2-u)]$.

Analytical solutions were compared with computational simulations for validation, specific details of sample sizes and parameter used are provided in the corresponding sections.

Using the analytical solutions of the *k*-subpopulation model, the expected driver load was estimated for tumours at sizes 1 – 4 cm^3 which corresponds to ~ 1 – 4 billion cells, assuming a representative sample. In this scenario, the detectability of cellular driver load in the tumour has to be greater or equal than a β fraction of the tumour. This means that subpopulations S_k can be detectable if their fraction in the tumour is greater or equal than $E[N(t_N)] * \beta$ for $0 < \beta < 1$.

In the k-subpopulation model (Fig 1.8 left) the expected driver composition is aggregated into subpopulations with the same driver load limiting the phylogenetic reconstruction of the tumour. To resolve the clonal structure of the expected driver load at tumour sizes 1 – 4 cm^3, the analytical solutions of the clonal model were used applying the same restriction $E[N(t_N)] * \beta$.

The expected tumour driver load and clonal compositions used the analytical solutions forward in time evaluating if the target tumour size is achieved every generation $t = t + 1$. The number of subpopulations S_k or clones $C_{k,ij}$ evaluated at every iteration were subject to their emergence at time τ or $\hat{\tau}_{k,i,j}$. Details of the parameter combinations used are provided in the corresponding sections. Appendix 2.2.5 provides a step-by-step implementation.

6 Relationship Between Expectation and Variance of the Founder Subpopulation $S_{k=1}$ in the Bozic Model and $C_{k,i,j}$ in the Clonal Model

Deriving expectation and variance for key parameters in the branching process summarises how the initial conditions affect simulation outcomes, and approximates the main evolutionary trajectories without performing computational simulations. This is useful when the parameters of interest require considerable computational time to generate or to identify the likelihood of initial conditions of non-primary tumours (i.e. metastases, cell lines, patient derived xenografts, etc).

From there, predicted model outcomes can be made in a patient-specific context to understand clinical implications. For instance, identifying analytical solutions is valuable in situations when genomic information is not available and comparisons have to be done using clinical measurements such as Ki-67, tumour size, differential gene expression, etc.

The best place to start when deriving analytical solutions for the discrete-time branching process is by summarising the growth dynamics of the founder subpopulation $S_{k=1}$ (Fig.2.2 and eq.1.2), namely the size of $S_{k=1}$ at time $t = t + 1$, as $E[S_{k=1}(t+1)]$ and $Var(S_{k=1}(t+1))$ in the Bozic model [43].

The mathematical notations for this founder subpopulation ($S_{k=1}$) have the advantage of being applicable to every clone in a clonal model ($S_{k=1}$ is relatable in form to $C_{k,i,j}$). This means that identifying the expectation and variance for the founder subpopulation of the Bozic model [43] has the same mathematical form in the clonal model as shown in Figure 2.2. The indexing of the clonal model is in reference to the position of a given clone in the phylogeny. Where k represents the number of drivers, i the parental lineage and j the progeny of clone $C_{k,i,*}$ as shown in Fig 2.6 later in the chapter.

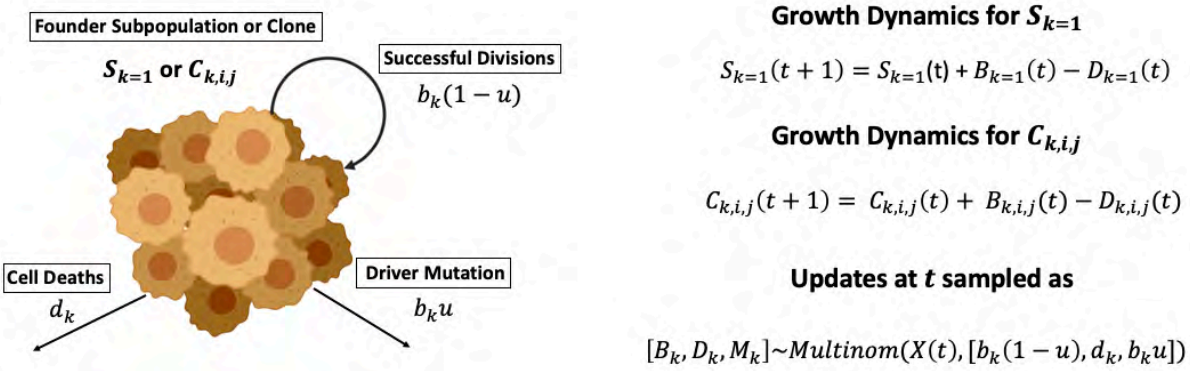


Figure 2.2 Growth dynamics of the branching process models. Left, the description of the events of interest and their respective probabilities in the branching process. Both the subpopulation and clonal (modification suggested by Chowell et al.) models have the same mathematical form. Right, the expansion equations for the founder subpopulation in the Bozic model and the expansion in the clonal model. Both branching process models use the multinomial distribution to update their parameters.

As mentioned, one of the assumptions of the model is that the growth dynamics for subpopulations or clones are exponential given $b > d$ for all cells. Decreases or fluctuations in fitness by negative selection are not modelled. Martincorena et al. [70] showed in a pan-cancer analysis including 7,664 samples, that positive selection is more prominent that negative selection during tumour development. The equivalent has been showed at the expression level [157], thus it is reasonable to assume that modelling only positive fitness gains has power to recreate tumour formation.

As such, changes in growth dynamics (in the models evaluated) are only due to stochastic drift associated with the value of the extinction probability, $\delta = \frac{d}{b}$, for every subpopulation or clone.

The branching process has known general solutions for the expectation and variance for models of cellular division where it is assumed that cells are of the same type and no migration occurs [156]. Derivations of the expectation and variance can be obtained by applying the moments

of the probability generating function [158, 159], or by conditional expectation in the growth dynamics equations shown in Figure 2.2 (further details and derivations Appendix A.2.1). The expectation and variance in the Galton Watson process with subpopulation X and net-growth μ are the following subject to $\mu > 1$,

$$E[X_n] = \mu^n$$

$$Var(X_n) = \frac{\mu^{n-1}(1-\mu^n)\sigma^2}{1-\mu}$$

With $E[X_1] = \mu$ and $Var(X_1) = \sigma^2$, in our models we are only interested in the case where $\mu > 1$ given $b > d$ with no carrying capacity. The $Var(X_n)$ is written in terms of the expectation evaluated at different times which simplifies its implementation.

The general solutions mentioned above can be used as a baseline to derive $E[S_{k=1}(t+1)]$ and $Var\big(S_{k=1}(t+1)\big)$. As described by Bozic et al. [43], the branching process models tumour progression and not tumour initiation, thus identifying $E[S_{k=1}(t+1)]$ and $Var\big(S_{k=1}(t+1)\big)$ does not require modelling how healthy cells become cancerous, but does implicitly assume a monoclonal origin. Therefore, the expectation of $E[S_{k=1}(t+1)]$ for the Bozic model [43] accounting for the multinomial sampling expressed in terms of $b_{k=1}$ is,

$$E[S_{k=1}(t+1)] = S_{k=1}(t) + S_{k=1}(t)[b_{k=1}(1-u) - 1 + b_{k=1}]$$

$$E[S_{k=1}(t+1)] = S_{k=1}(t)[b_{k=1}(2-u)] \tag{2.1}$$

It can be seen from eq. 2.1 how the recursive form of the expectation will lead to an increasing polynomial term after few iterations. It can be simplified by expressing in terms of b, leading to a simplified closed form solution: $E[S_{k=1}(t+1)] = S_0[b_{k=1}(2-u)]^t$, where S_0 is the initial number of cells, to account for successful lineages the $S_0 = 1/(1-\delta_{k=1})$. As a result, the expected value for the founder subpopulation is the following,

$$E[S_{k=1}(t+1)] = \frac{[b_{k=1}(2-u)]^t}{(1-\delta_{k=1})} \tag{2.2}$$

Reiter et. al.[147] approximated $E[S_{k=1}(t+1)]$ by expressing it in terms of the selective advantage s, as $E[S_{k=1}(t+1)] = \frac{1}{2s_0}(1+s_0)^t$. However, this expression is restricted to cases where $s_0 \ll 1, s_1 \ll 1$, and $u \ll s_0$, otherwise both expected solutions are equivalent.

It is worth reiterating that expectation and variance of $S_{k=1}$ can be applied to all clones by definition of the clonal model construction, thus new driver mutants $C_{k,ij}$ are evaluated independently. The clonal model has the following expectation,

$$E\big[C_{k,i,j}(t+1)\big] = C_{k,i,j}(t-\tau_{k-1}) + S_k(t-\tau_{k-1})(b_k(1-u) - 1 + b_k)$$

$$E\big[C_{k,i,j}(t+1)\big] \approx \left(\frac{\big[b_{k,i,j}(2-u)\big]^{t-\hat{\tau}_{k-1,i,j}}}{1-\delta_{k,i,j}}\right) \tag{2.3}$$

$$t - \hat{\tau}_{k-1,i,j} \geq 0$$

The distinction between $E[S_{k=1}(t+1)]$ and $E\left[C_{k,ij}(t+1)\right]$ is in the indexing of time $t - \hat{\tau}_{k-1,i,j}$ which refers to the time when a successful clone $C_{k,ij}$ expands using estimator $\hat{\tau}$. This is a distinction from reported value of τ from the k-subpopulation model that only considers the first k founder lineages, the clonal model requires a generalisation of τ for every clone in the position of the phylogenetic tree however.

Instead of deriving the analytical solution of $\tau_{k,i,j}$ for the clonal model, I derived its estimator $\hat{\tau}_{k,i,j}$ which uses a calibration term ε that allows for better alignment with computer simulations later discussed in the chapter.

With the known form of the variance of the branching process in terms of its expectation previously mentioned, the variance for $S_{k=1}$ and $C_{k,ij}$ takes the following form,

$$Var\big(S_{k=1}(t+1)\big) = \sigma^2(1-\delta_{k=1})^{-1}[b_{k=1}(2-u)]^{t-1}\left(\frac{1-(1-\delta_{k=1})^{-1}[b_{k=1}(2-u)]^{t}}{1-(1-\delta_{k=1})^{-1}[b_{k=1}(2-u)]}\right)$$

$$Var\big(S_{k=1}(t+1)\big) = \sigma^2 E[S_{k=1}(t)]\left(\frac{1-E[S_{k=1}(t+1)]}{1-E[S_{k=1}(1)]}\right) \tag{2.4}$$

$$Var\left(C_{k,i,j}(t+1)\right) \approx \sigma^2(-\delta_{k,i,j})^{-1}[b_{k,i,j}(2-u)]^{t-\hat{\tau}_{k-1,i,j}-1}\left(\frac{1-(1-\delta_{k,i,j})^{-1}[b_{k,i,j}(2-u)]^{t-\hat{\tau}_{k-1,i,j}}}{1-(1-\delta_{k,i,j})^{-1}([(2-u)]}\right)$$

$$Var\left(C_{k,i,j}(t+1)\right) \approx \sigma^2 E[C_{k,i,j}(t-\hat{\tau}_{k-1,i,j}-1)]\left(\frac{1-E\left[C_{k,i,j}\left(t-\hat{\tau}_{k-1,i,j}\right)\right]}{1-E[C_{k,i,j}(t-\hat{\tau}_{k-1,i,j}=1)]}\right) \tag{2.5}$$

$$t - \hat{\tau}_{k-1,i,j} \geq 0$$

Where σ^2 is the variance associated at the first iteration of subpopulation or clonal update, $Var\big(S_{k=1}(t=1)\big) = \sigma^2$ and $Var\left(C_{k,i,j}\left(t-\hat{\tau}_{k-1,i,j}=1\right)\right) \approx \sigma^2$.

With the values of the expectation and variance it is possible to approximate the growth of founder subpopulation in the Bozic model [43], and for any particular clone in the clonal model. This enables investigation of how different values of the average selective average s and average driver mutation rate u impact growth dynamics.

As mentioned, the growth dynamics of the models studied here have stochastic drift as a function of the extinction probability δ, impacting their evolutionary trajectories. Thus, it is relevant to determine how well the expectation and variance capture the model dynamics as δ changes.

To investigate the effect of the extinction probability δ and stochastic drift in the k-subpopulation model, the following figure shows how the expectation (in black) and the standard deviation (in red) of $S_{k=1}$ compared to a sample of simulations (in blue) with starting

values of s: $\{0.1, 0.01, 0.001\}$ and a fixed value of the driver mutation rate $u = 3.4\text{x}10^{-5}$ use in by Bozic et al. [43].

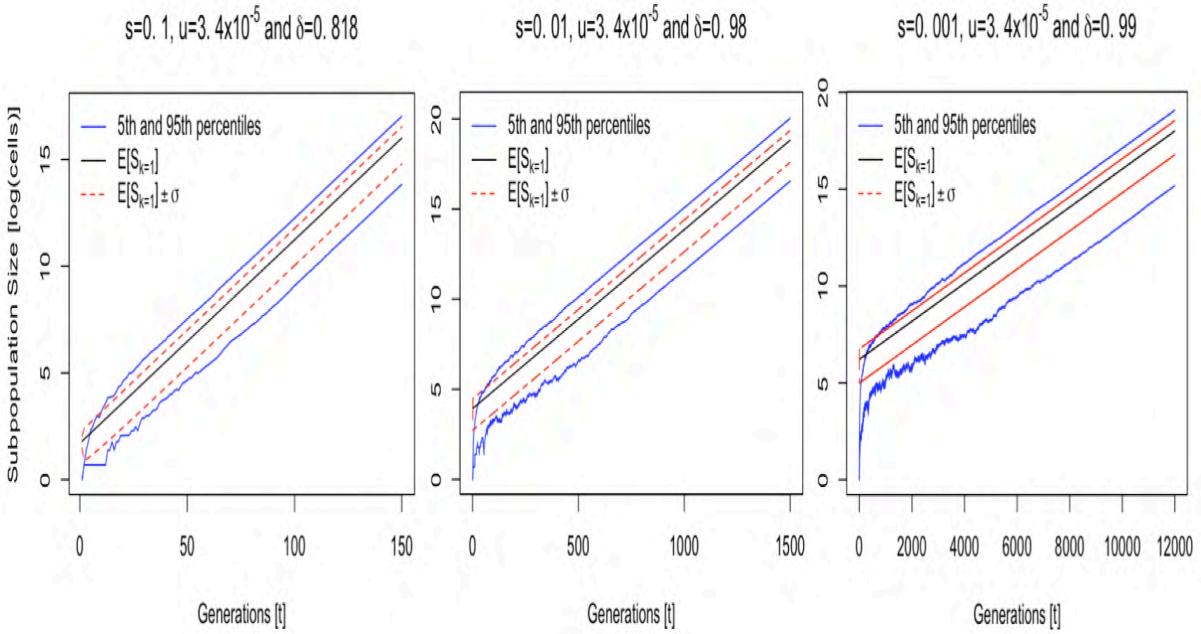


Figure 2.3 Growth of $S_{k=1}$ considering multiple values of δ. All plots have u= $3.4\text{x}10^{-5}$ and s: $\{0.1, 0.01, 0.001\}$. In blue the 5th and 95th percentiles of 100 (successful) simulations, black is the expectation using eq. 2.2 and red the standard deviation from the mean using eq. 2.4.

Increases in the extinction probability δ not only increased the number of generations that the subpopulation/clone requires to grow but also the number of evolutionary paths (higher time for tumour development increases the potential evolutionary paths the tumour can take). As a result, the net-growth expansion is proportional to s, whereas variability and drift are inversely proportional to s. As a consequence, the analytical solutions provide an estimate of the main behaviour of the bulk of the simulations and also reveal the degree of stochastic variation subject to initial conditions of s.

Although growth is an exponential process, I derived analytical solutions assuming a carrying capacity shown in Appendix A.2.3. Bozic et al. [43], showed that the branching process with additive fitness model is a good approximation of the carrying capacity model, given this simplification the exponential growth model is preferred in certain cases.

With solutions for the founder subpopulation $S_{k=1}$ in the Bozic model [43], the next goal is to obtain the general form of S_k. However, as subpopulations $k \geq 1$ emerge later in time, a generalisation of their expectation considering a time shift needs to be introduced.

7 Expectation and Variance of S_k in the Bozic Model

Bozic et. al. [43] and Durrett et al. [42] have provided estimates for the waiting time in the discrete and continuous branching processes, denoted by τ and σ respectively, and showed their values for multiple combinations of s, k and u. Therefore, using the analytical solution of $\tau_k \approx \frac{T}{ks} log\left(\frac{2ks}{u}\right)$, the expectation for a subpopulation expansion takes the following general form,

$$E[S_k(t+1)] \approx E[S_k(t-\tau_{k-1})] + E[B_k] - E[D_k] + E[M_{k-1}] \qquad (2.6)$$
$$t - \tau_{k-1} \geq 0$$

Here $t - \tau_{k-1}$ is the adjusted time reference relative to the expansion of subpopulation S_k. The indexing of τ is in terms of the progeny $k+1$, meaning that the expected time of emergence of the founder subpopulation with one driver alteration $S_{k=1}$ is τ_0 and the equivalent for a two-driver subpopulation $S_{k=2}$ is τ_1.

Now, it is possible to express the expected values in terms of lineage emergence for the Bozic model [43] as following,

$$E[S_k(t+1)] \approx S_k(t-\tau_{k-1}) + S_k(t-\tau_{k-1})(b_k(1-u) - 1 + b_k) + S_{k-1}(t-\tau_{k-2})u$$

$$E[S_k(t+1)] \approx \left(\frac{[b_k(2-u)]^{t-\tau_{k-1}}}{1-\delta_k}\right) + \left(\frac{[b_{k-1}u(2-u)]^{t-\tau_{k-2}}}{1-\delta_{k-1}}\right) \qquad (2.7)$$

$$t - \tau_{k-1} \geq 0\,;\; t - \tau_{k-2} \geq 0$$

For cases where $k > 1$, the expansion term takes into consideration two components, the expansion of the subpopulation with expectation as shown in eq. 2.2 and the entry of cells accumulating for the *k*th new driver mutation shown in Fig. 1.8, $S_{k-1}(t-\tau_{k-2})b_{k-1}u$. It is worth mentioning that the time indexing for the expectation of S_k cannot be negative $t - \tau_k \geq 0$, which limit the number of possible k subpopulations when a target tumour size N is required.

As mentioned earlier in the chapter, the variance can be expressed in terms of the expectation evaluated at different times, the $Var(S_k(t+1))$ takes the following form,

$$Var\big(S_k(t+1)\big) = \sigma^2 E[S_k(t)]\left(\frac{1 - E[S_k(t+1)]}{1 - E[S_k(1)]}\right) \qquad (2.8)$$

With the expectation and variance of S_k and $C_{k,i,j}$ inferences about the number of driver subpopulations in the tumour and clonality can be made considering different experimental conditions.

8 Waiting Times of Successful Clones and the Estimator $\hat{\tau}_{k,i,j}$

To connect the expectation of clones and tumour to growth in the clonal model, it is required to derive the expected time in which a clone emerges ($\tau_{k,i,j}$) and does not die out by drift, hence successfully starts expansion.

A generalisation of τ for the continuous-time branching process was shown by Durrett et al. [42] that alternatively can be used from the approximation proposed here. Bozic et al. [43] showed the general form of the waiting time as $\sum_{t=1}^{\tau_{k/T}} \left(\frac{1-\delta_{k+1}}{1-\delta_k}\right) ub_k[b_k(2-1)]^t = 1$, where T represents the average division rate used to scale time. This general form can be parametrised to account for the general case as illustrated in eq. 2.9.

Further development in the branching process models by Cheek et al. [145] showed that mutational events in this type of model occur as a Cox process with probability bu. For every clone the introduction of new mutants occurs at a rate of $bu[b(2-u)]$. Because the driver mutation rate u is low and the number of cells required to introduce a new driver mutant is large, the process can be generalised as a Poisson point process with probability $b_{k,i,j}u[b_{k,i,j}(2-u)]^t$ occurring in a time interval $[1, x]$. The waiting time $\tau_{k,i,j}$ for the c[th] clone descending from parent $C_{k,i,j}$ is a random stopped sum as shown in eq. 2.9 with $E[X] = \tau_{k,i,j}$.

Instead of finding analytically the value of $\tau_{k,i,j}$, I used the approximation to identify at which time t inside the interval $[1, x]$, the cumulative sum of $b_{k,i,j}u[b_{k,i,j}(2-u)]^t$ takes integer values reflecting the new emerging lineage ($c = 1$ first successful lineage, $c = 2$ second successful lineage, etc.) shown in equation 2.9a. As discussed by Durrett et al. [42], the τ estimator from the Bozic model [43] is slightly biased due to numeric approximation and may benefit from a correction term. As a result, including a calibration term ε, is one of the main advantages of the $\hat{\tau}_{k,i,j}$ estimator that can be used for fine-tuning.

$$\sum_{t=1}^{X} b_{k,i,j}u[b_{k,i,j}(2-u)]^t = c \tag{2.9}$$

$$\hat{\tau}_{k,i,j} = \underset{1 \le t \le x}{\operatorname{argmin}}\left(\left|cumsum(b_{k,i,j}u[b_{k,i,j}(2-u)]^t) - c - \varepsilon\right|\right) \tag{2.9a}$$

To validate the estimates of $\hat{\tau}_{k,i,j}$ using eq. 2.9a, predictions were compared with 300 computer simulations with parameters $u = 1\text{x}10^{-5}$, $s: \{0.1, 0.01, 0.001\}$, $k: \{1, \ldots, 5\}$ and with τ defined in its approximation form as $\tau \approx \frac{T}{ks} log\left(\frac{2ks}{u}\right)$ for the *k*-subpopulation model. It is worth mentioning that for comparing with Bozic's τ the value of c of eq 2.9a is set to $c = 1$ as we are looking for the first founder lineages.

The best calibration was obtained when $c - \varepsilon = 0.4$ with minimum deviation from the median of simulations shown in Figure 2.4, even when no calibration was done the $\hat{\tau}_k$ predictions were inside the 25[th] and 75[th] percentiles and closely related to τ (Figure S2.3).

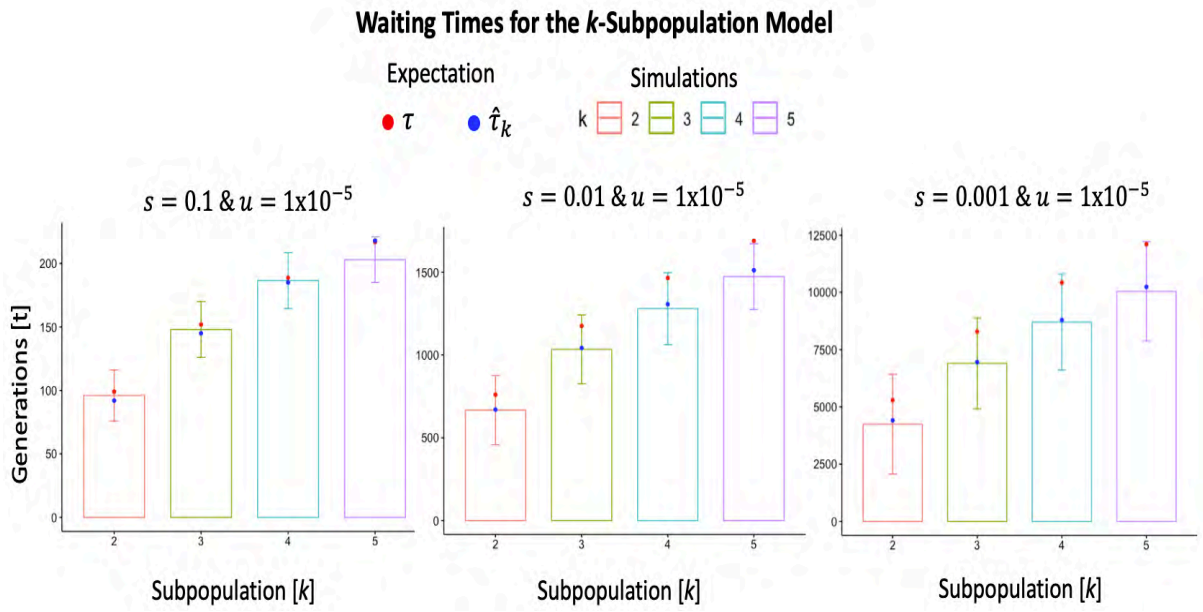


Figure 2.4 Comparison of $\hat{\tau}_k$, τ and computer simulations with the k-subpopulation model. 300 (successful) simulations were used to contrast expected solutions of τ and $\hat{\tau}_k$ with initial parameters $u = 1\text{x}10^{-5}$, s:$\{0.1, 0.01, 0.001\}$ & k:$\{1,\ldots,5\}$. Error bars represent the 25th and 75th percentiles of simulations, in red the reported value of τ and in blue $\hat{\tau}_k$ with $\varepsilon = 0.4$.

It can be seen in Figure 2.4 that $\hat{\tau}_k$ (blue) is better aligned with computer simulations than τ (red) as it is inside the 25th and 75th percentiles of simulations (error bars) for most of the k-subpopulations, and it is closer to the median of the simulations. The effect of drift is evident in the previous figure showing that as s decreases more variability is introduced and more simulations are required to represent the global effect of the input parameters s and u. The main advantage of the $\hat{\tau}_k$ estimator was in the calibration parameter ε that can help to deviate from τ in an acceptable range achieving a better global fit if required. However, caution has to be made to avoid overfitting by taking a considerable sample, which is dependant of s, u and k.

A similar experiment was conducted for $\hat{\tau}_{k,i,j}$ using the clonal model with additive fitness, 150 simulations of the clonal model were generated with parameters $u = 1\text{x}10^{-5}$ and s:$\{0.1, 0.01\}$ to evaluate the alignment with the simulations. The descendants from the founder clone $C_{1,*}$ the first 2-driver clone $C_{1,2,*}$ were evaluated. In this experiment no calibration was done ($\varepsilon = 0$) showing a good fit within 25th and 75th percentiles (error bars).

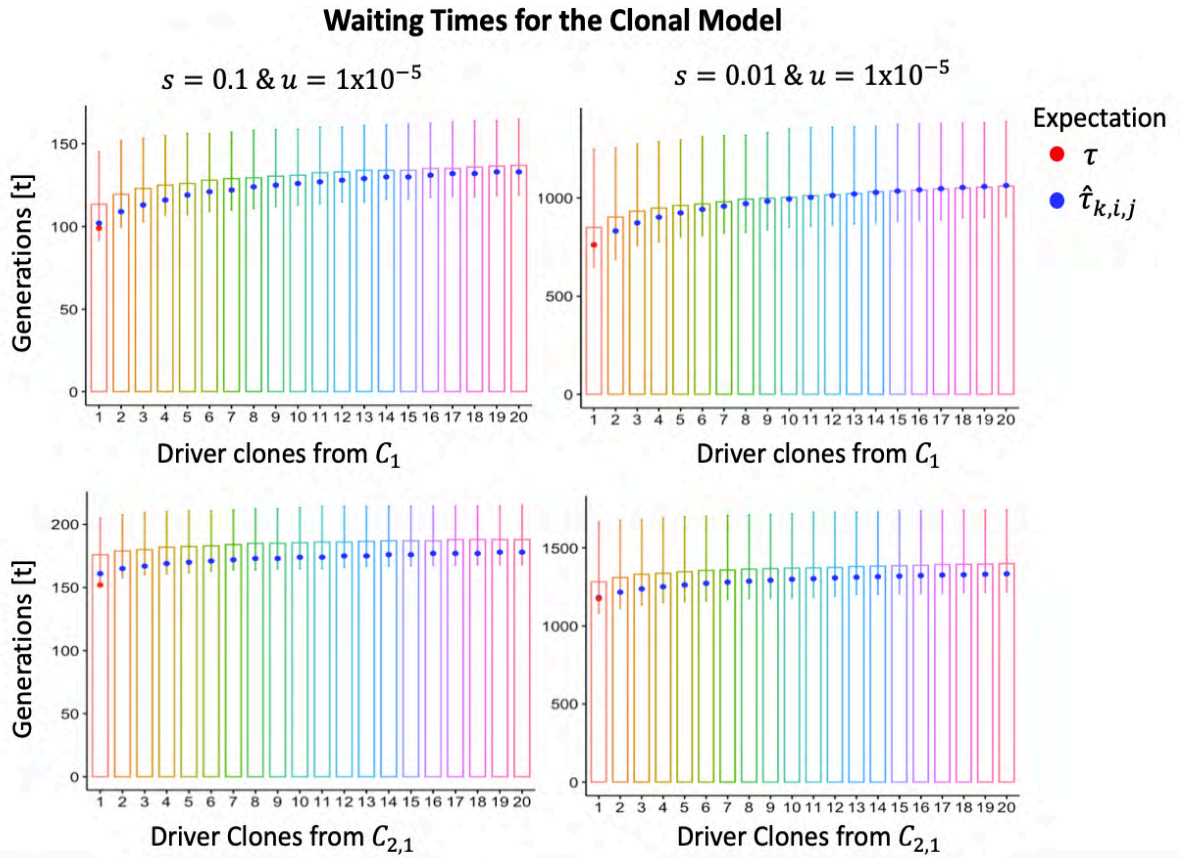


Figure 2.5 Comparison of $\hat{\tau}_{k,i,j}$ and computer simulations using the clonal model with additive fitness. 150 (successful) simulations were used to contrast expected solutions of τ and $\hat{\tau}_k$ with initial parameters $u = 1\text{x}10^{-5}$, s: $\{0.1, 0.01\}$ & k: $\{1, \ldots, 20\}$. Error bars represent the 25th and 75th percentiles of simulations, in red the reported value of τ and in blue $\hat{\tau}_k$ with $\varepsilon = 0.0$.

Aligning Equation 2.9a with simulations for the k-subpopulation and clonal models with additive fitness can be used to learn the rate at which clonal waves occur. Figure 2.6 shows an example of how the clonal sweeps occur using $\hat{\tau}_{k,i,j}$ and $E[C_{k,i,j}]$. In black, the founder clone introduces a new lineage at $\hat{\tau}_1$ coloured in blue and shortly after subsequent mutant lineages $\hat{\tau}_{1,*}$ are spawned. The introduction of mutants accelerates as it is proportional to the expansion of the parental clone. The red and green lines represent clones containing 3 drivers, each of which split independently from the first 2-driver lineages, eventually the 3-driver clones outcompete the 2-driver lineages around generation 2,000.

In addition, it shows how these clones can be allocated in a phylogeny. The founder clone C_1, coloured black, starts its expansion at $t = 1$. At τ_1 (or $\hat{\tau}_1$) the first successful lineage with two drivers emerges, $C_{2,1}$, represented by the top blue line. In the interval from $\hat{\tau}_{1,1}$ to $\hat{\tau}_{1,n}$ subsequent clonal waves from the founder clone C_1 emerge, also coloured blue. Similarly, new driver mutants from clone $C_{2,1}$ emerge in the interval of $\hat{\tau}_{2,1,1}$ to $\hat{\tau}_{2,1,n}$ and are coloured in red. Equivalently, new driver mutants of clone $C_{2,2}$ which is the second lineage clone from the founder, are also introduced in the interval $\hat{\tau}_{2,2,1}$ to $\hat{\tau}_{2,2,n}$, coloured in green. In this fashion, $\hat{\tau}_{k,i,j}$ provides an alternative approach for estimating waiting times in the clonal model.

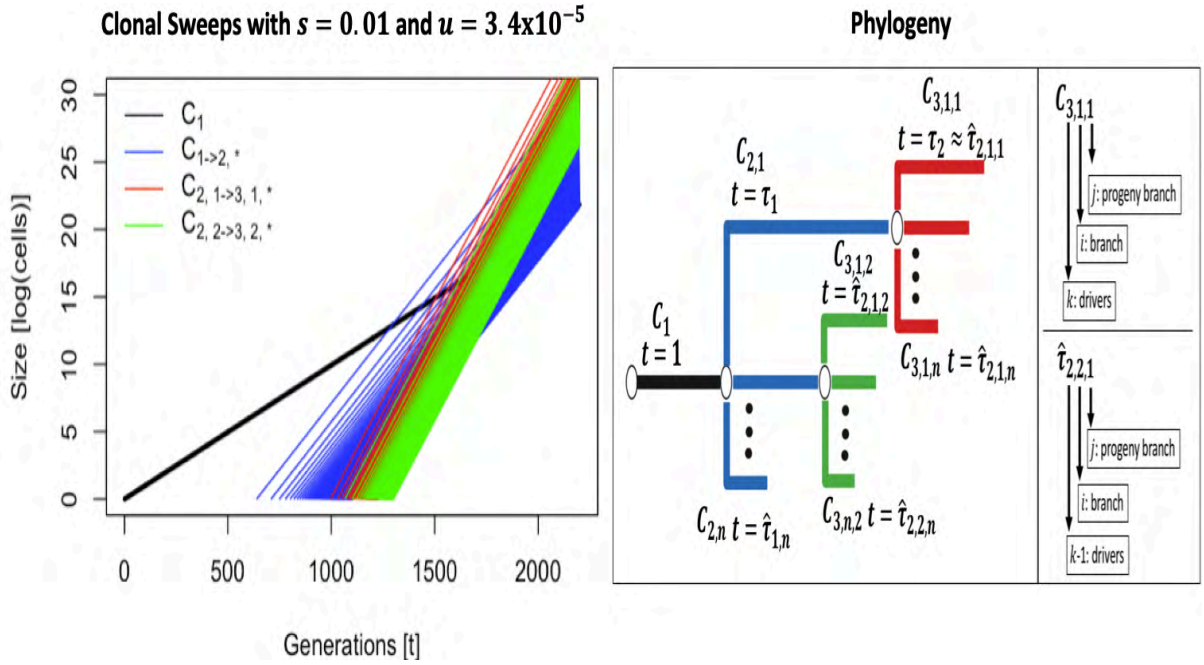


Figure 2.6 Timing of clonal sweeps as predicted by $\hat{\tau}_{k,i,j}$. The left-hand side plot shows the times at which clones are expected to arise and expand according to $\hat{\tau}_{k,i,j}$. The branch with the founder clone is labelled in black. The founder introduces 2-driver clones, coloured in blue, and 3-driver clone originate from the first 2-driver lineage, coloured in red. Later 3-driver clones are from the second 2-driver lineage, coloured in green. The right-hand side shows the phylogeny associated with these clonal sweeps, with arrows indicating the lineage of the clones in the phylogeny.

It is possible now to evaluate the expected growth of a given clone $C_{k,i,j}$ with the estimator $\hat{\tau}_{k,i,j}$, the next figure is an example of how the analytical solutions can predict the average behaviour of the stochastic simulations. To exemplify the use of the analytical solutions, 100 computer simulations were generated with parameters $s = 0.1$ and $u = 1\text{x}10^{-5}$ to evaluate the first 3 clones that emerged from the founder clone $C_{k=1}$.

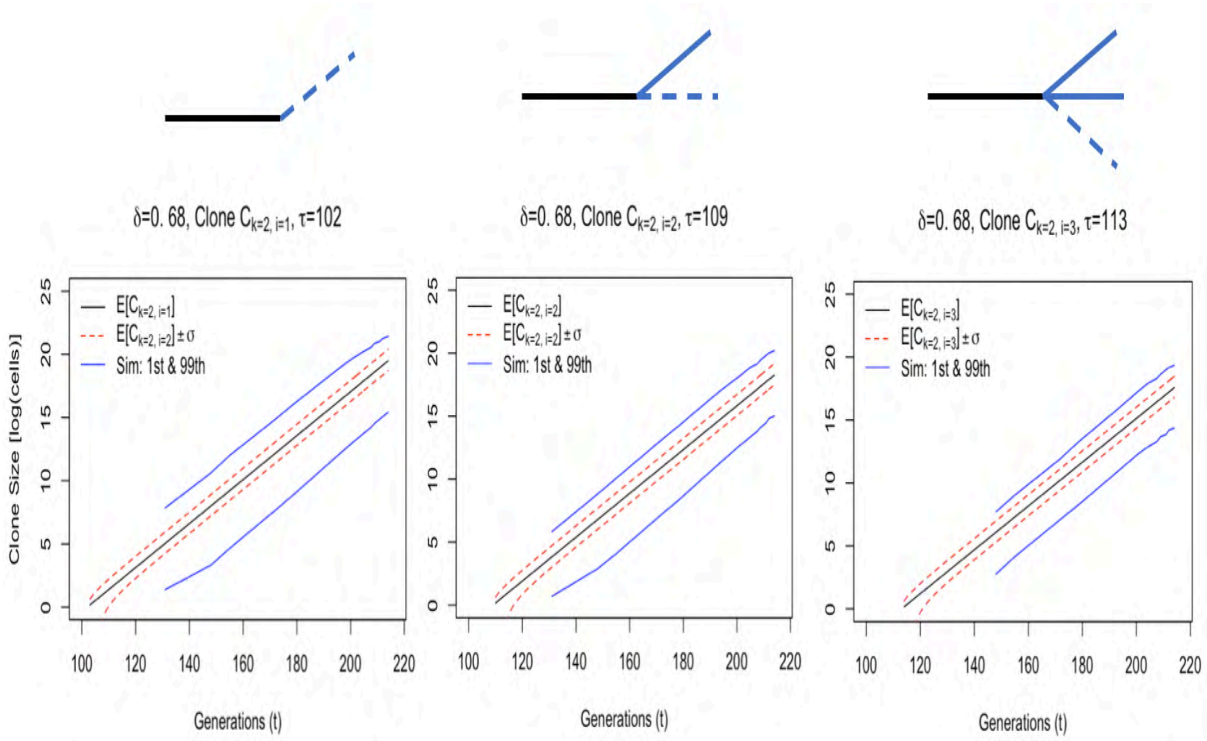


Figure 2.7 Growth of the first, second and third two-driver clones, $C_{k=1,j=\{1,2,3\}}$, considering multiple values of δ. All plots have $u = 1\text{x}10^{-5}$ and $s = 0.1$. In blue the 1th and 99th percentiles of 100 (successful) simulations, black is the expectation using eq. 2.3 and red the standard deviation from the mean using eq. 2.5. On top of every is the tree topology, the founder $C_{k=1}$ subpopulation is coloured black and its progeny $C_{k=1,*}$ coloured blue. The dashed branch represents the clone whose growth is being evaluated.

The analytical solutions were able to represent the main behaviour of the simulations providing detailed information on a specific branch of the tree and can be used to predict future behaviour.

With the expected value of S_k, $C_{k,i,j}$ and $\hat{\tau}_{k,i,j}$ the k-subpopulation and clonal models can be used to approximate tumour composition at any given time, with the governing assumptions imposed by the models. The next sections will focus the expectation of the tumour, the expected driver abundance and expected number of clonal subpopulations.

9 Tumour Subpopulation Composition at Expected Tumour Size $E[N(t+1)]$ in the Bozic Model

With the expectation and variance of the k-subpopulation and clonal models, it is now possible to derive the expected tumour size $E[N(t+1)]$ as a function of the number of subpopulations or clones expanding in the tumour at time t. The tumour composition based on initial conditions of s, u and τ for the k-subpopulation and clonal models is the following,

$$E[N(t+1)] = \sum_{k=1}^{K} E[S_k(t+1)] \approx \sum_{k=1}^{K} \left(\frac{[b_k(2-u)]^{t-\tau_{k-1}}}{1-\delta_k} \right) + \mathbb{I}_{k>1} \left(\frac{[b_{k-1}u(2-u)]^{t-\tau_{k-2}}}{1-\delta_{k-1}} \right) \quad (2.10)$$
$$t - \tau_{k-1} \geq 0 \; ; \; t - \tau_{k-2} \geq 0$$

$$E[N(t+1)] = \sum_{k=1}^{K} E\left[C_{k,i,j}(t+1)\right] \approx \left(\frac{\left[b_{k,i,j}(2-u)\right]^{t-\hat{\tau}_{k-1,i,j}}}{1-\delta_{k,i,j}} \right) \quad t - \tau_{k-1,i,j} \geq 0 \tag{2.10a}$$

Note that $Var(N(t+1))$ can be estimated using the solutions of the expectations as exemplified previously. With equations 2.9 and 2.9a it is possible to approximate tumour growth in a given interval $t: a \leq t \leq b$, with the restriction that only subpopulations (or clones) that have emerged within the interval $t: 1 \leq t \leq b$ should be evaluated. Figure 2.8 shows the expected tumour growth for the *k*-subpopulation model using eq. 2.10 over the interval of time [0, 4500], within the interval 6 subpopulations have emerged at times τ_k:{1, 1671, 2593, 3252, 3761, 4181} that were used for the approximation. The expectation and standard deviation are compared with the 5th and 95th percentiles of 300 simulations generated with $s = 0.004$ and $u = 1\text{x}10^{-5}$.

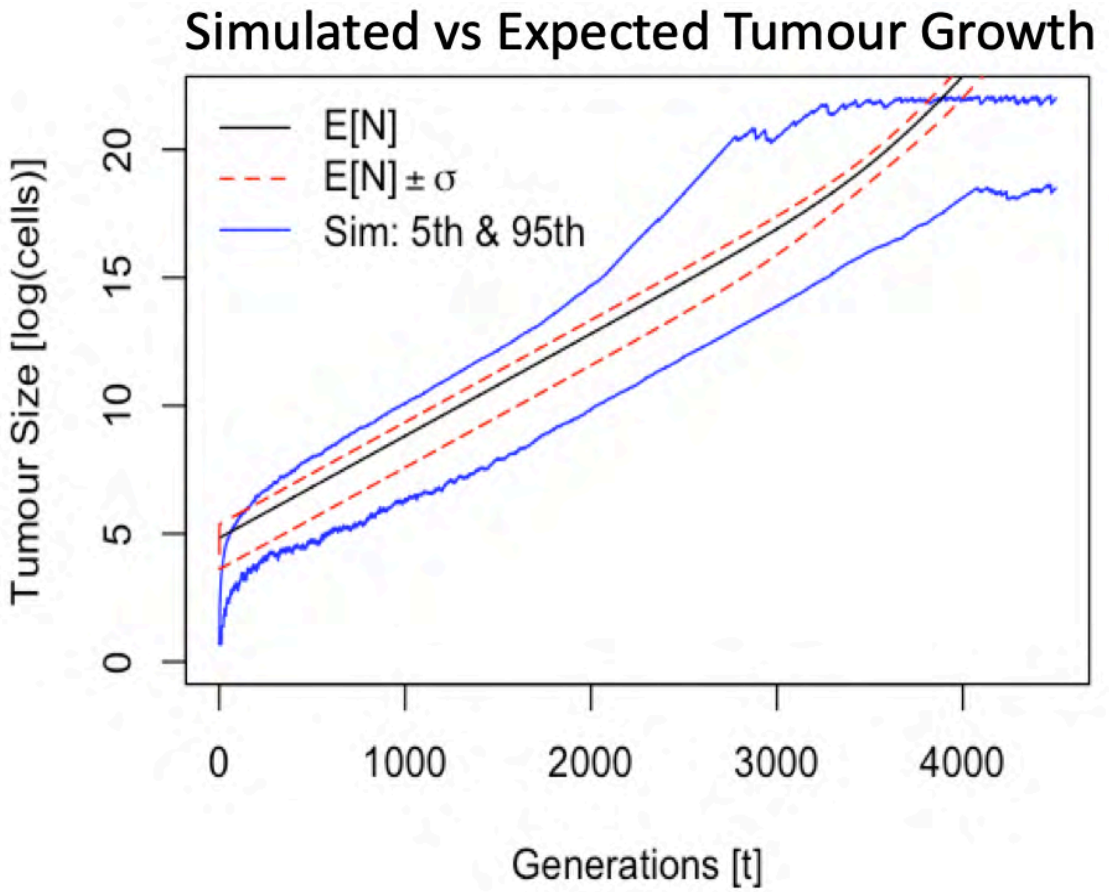


Figure 2.8 Expected tumour growth $E[N(t+1)]$ by subpopulation aggregation. In blue the 5th and 95th percentiles of 300 (successful) simulations, with initial parameters $s = 0.004$ and $u = 1\text{x}10^{-5}$. In black is the $E[N(t+1)]$ using eq. 2.10, in red and red the standard deviation from the mean using eq. 2.4 for all *k*.

It can be seen that the analytical solutions derived for the *k*-subpopulation model are aligned with the simulations. An equivalent experiment was done to validate the clonal model, the following figure is the validation of the analytical solutions of the clonal model using eq. 2.10a, 200 simulations were generated with parameters $s = 0.1$ and $u = 1\text{x}10^{-5}$. A maximum of 70,000 2-driver clones were allowed from the founder clone if their emergence time was in the

range from generation 1 to 200, similarly a maximum of 5,000 clones were allowed from 2-driver lineages if their emergence was the generation timeframe (1 to 200).

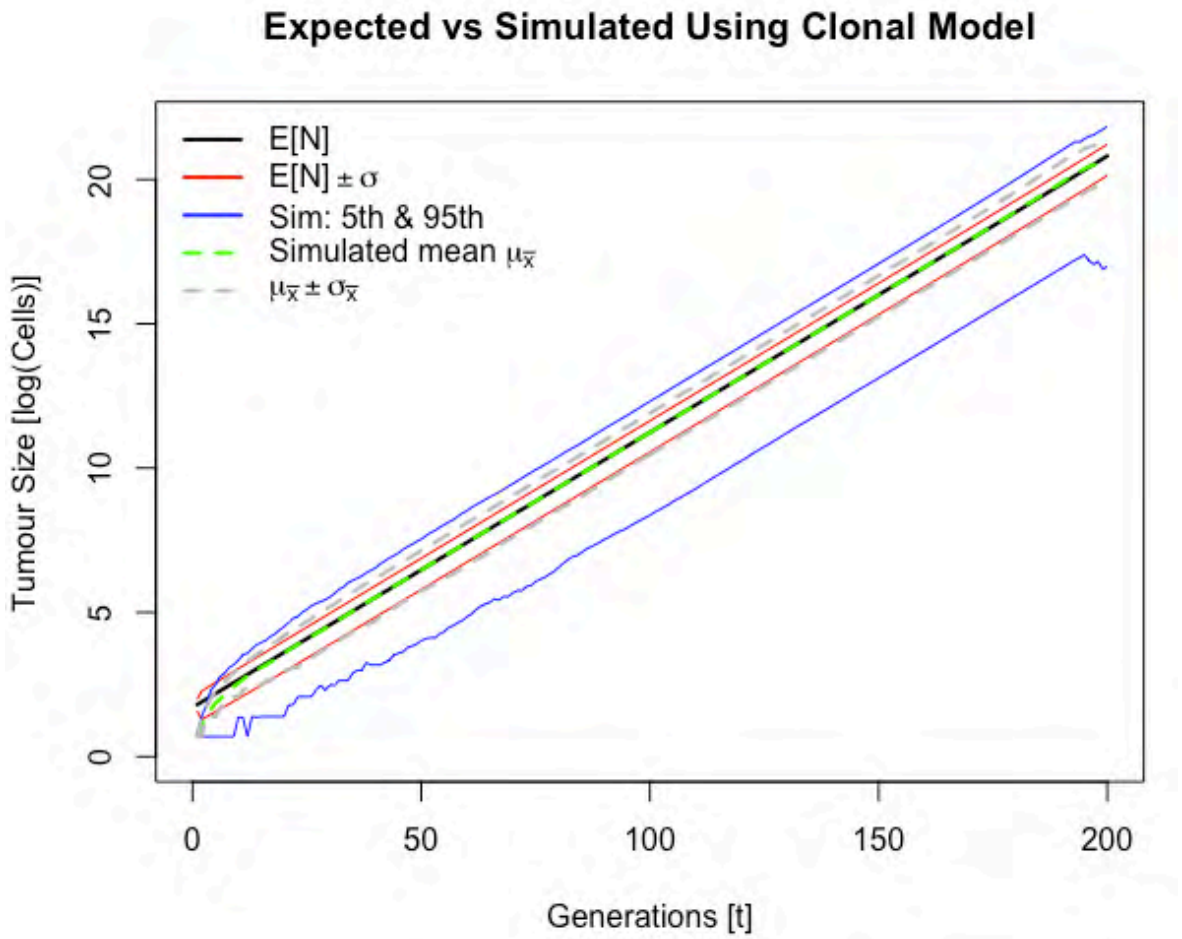


Figure 2.9 Expected tumour growth $E[N(t+1)]$ by clonal aggregation. In blue the 5th and 95th percentiles of 200 (successful) simulations, with initial parameters $s = 0.1$ and $u = 1\text{x}10^{-5}$. In black is the $E[N(t+1)]$ using eq. 2.10a, in red and red the standard deviation from the mean using eq. 2.5 for all k. In green is the sample mean from the simulations and in grey the standard error from simulations.

Figure 2.9 shows agreement between the simulations and analytical solutions using the analytical solution of the clonal model. It shows how all solutions can be incorporated together to represent tumour growth.

In this section I was able to incorporate the analytical solutions to model individual subpopulations, clones, their emergence times and the total tumour size. The next sections will use the analytical solutions to provide the expected driver and clonal composition of the tumour.

10 Expected Driver Composition in Tumours

The analytical solutions derived allow the estimation of the expected driver subpopulation composition of a tumour under certain initial conditions. Of clinical interest is the evolutionary features of tumours at different stages of disease progression, such as tumour burden, heterogeneity, proliferation, etc.

In relating tumour size with number of cells, it has been widely accepted that human malignancies are clinically detectable (palpable) have reached ~ 1 cm^3 ≈1 g wet weight [160]. At 1 cm^3 it was estimated that a lesion contains 10^9 cells and represents two-thirds of its growth [160, 161], and the relationship scales linearly (e.g. a 2 cm^3 lesion contains ~ 2 billion cells etc.). Although, this assumption is widely used in the branching process models [42, 45, 123], as suggested by [160], the relationship may be over estimated by an order of magnitude for epithelial tumours [160]. In this thesis, the traditional 1 cm^3 ≈10^9 cells for clinical detection is going to be assumed.

In the context of breast cancer, tumours are palpable between 1.5 – 2 cm diameter which corresponds to stages T0 – T1 and lesions above 2 cm diameter to stages T2 – T4 [162]. For the purpose of this thesis I will assume two milestone sizes with clinical relevance considering breast as benchmark:

- 1 cm^3 (1 billion cells) to reflect tumours with low risk of dissemination.
- 4 cm^3 (4 billion cells) for tumours at greater risk of dissemination [2, 163].

The main purpose of using analytical solutions, is to compare how clonal competition dynamics and the timeframes associated with changes in tumour size and malignancy (fitness) occur between early and late diagnosis.

The analytical solutions derived so far provide a framework to relate clinical tumour size and their proliferation (e.g. Ki-67) with simulated fitness to gain a global overview of the malignancy of the tumour. By applying the analytical framework, information about evolutionary features can be obtained when the tumour size, proliferation and average division rate is known but there is no sequencing information available.

To investigate the changes in k-subpopulation dominance at milestone sizes (1 & 4 cm^3), equation 2.10 was evaluated forward in time with input parameters N, s, u and k, evaluating at every iteration $t = t + 1$ if the input value of N is achieved by summing the sizes of all subpopulations S_k. The input parameters used to be in line with the results from Bozic et al. [43] were,

- N:1 billion and 4 billion cells.
- s: $\{0.1, 0.01 \text{ \& } 0.001\}$
- u: $3.14\text{x}10^{-5}$
- k: 1 to 10

As shown in Figure 2.10, the main difference in $E[N]$ over the range of values of s is in the time taken to attain detectable sizes of 1 cm^3 and 4 cm^3. For instance, tumours with $s = 0.001$, require 11,800 to 13,800 more generations to grow up to 4 cm^3 as compared to tumours where drivers confer a strong selective advantage $s = 0.1$. This indicates that tumours where driver mutations confer only a weak selective advantage ($s = 0.001$) require a fast division rate (1 or 2 days) in order to occur within a typical human lifespan. In contrast, tumours with a strong selective advantage ($s = 0.1$) can have considerably longer periods for cell division and still plausibly grow to detectable size within a typical human lifespan.

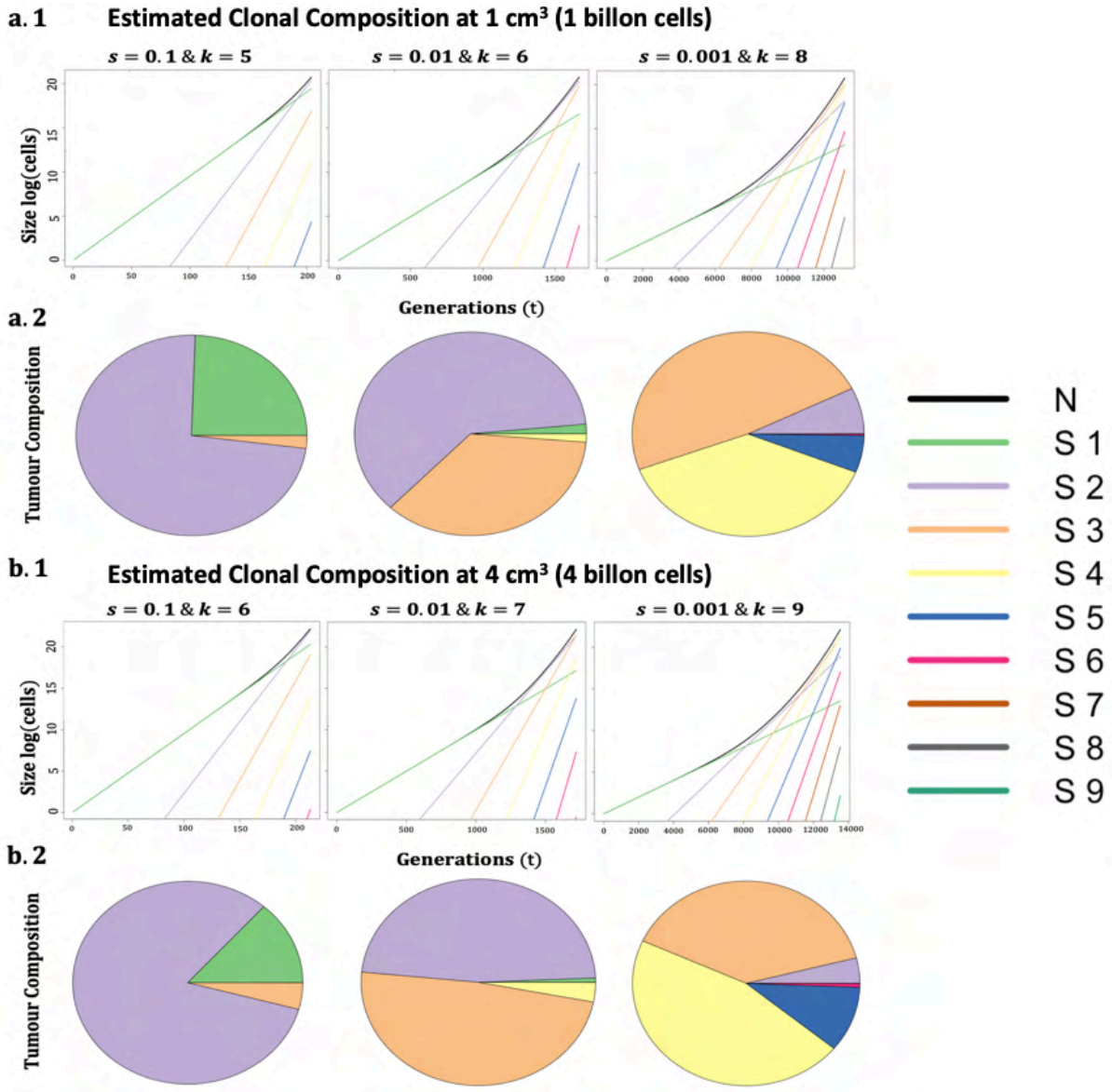


Figure 2.10 k-driver subpopulation in tumours. **a.1** and **b.1,** expected growth of tumour and *k*-driver subpopulations from Equation 2.9 at tumour sizes 1 and 4 cm^3, providing as an input N, k and s with $u = 3.14\text{x}10^{-5}$. **a.2** and **b.2**, pie charts of the fractions of the subpopulations at tumour sizes 1 and 4 cm^3.

The analytical solutions suggest that the main change in *k*-driver subpopulations from 1 cm^3 to 4 cm^3 size tumours, is the addition of an extra driver subpopulation and minor changes in proportions. The model indicates that subpopulation structure of a tumour is already mostly well-defined prior to reaching the assumed diagnosable size. As a result, though difference in tumour size leads to a higher driver content, the frequency of the new driver subpopulations is small, complicating their detection. Hence, it is expected that phylogeny reconstruction in tumour sizes of 1 cm^3 through 4 cm^3 will render similar phylogenetic topologies (e.g. limiting by $\geq 10\%\, CCF$) in unbiased samples (e.g. rep-seq and consensus multi-region) on exponential growth dynamics.

11 Expected Driver Composition Relative to Tumour Size

The previous section used the analytical solutions to measure the expected driver abundance at two milestone sizes, 1 and 4 cm^3. In this section concerns estimating abundance of driver mutations within the tumour when a restriction on the detectable subpopulation size is imposed. This restriction is relative to tumour size to reflect the number of drivers within cells that are measured if a representative sample to reflect *k*-subpopulation proportion is taken.

For instance, assuming exponential growth, Figure 2.11 a.2 and b.2 showed that the expected driver load in cells was in the rage of 1 – 6 drivers, though the tumour is mostly composed of 1 – 2 driver cell subpopulations.

In this section I assumed a scenario in which no sequencing is available and the only markers to measure tumour evolution are based on clinicopathological markers. Therefore, the interest is in the driver load of cells to represent tumour fitness. In the context of breast cancer, analytical solutions can be compared with a representative Ki-67 measurement of the whole tumour to establish tumour's fitness [155].

As shown by Chowell et al. [44], only few dominant clones are above the ~ 10% frequency ($\sim 10^8$ cells). This implies that at milestone sizes 1 and 4 cm^3 detectable *k*-driver subpopulations contain at least 100 and 400 million cells when a restriction of 10% is imposed on detection of clones (relative tumour size). Therefore, a 5 – 10% cut-offs was selected to compare with outcomes of the clonal model to approximate a CCFs, but do not necessarily correspond to realistic cut-offs. However, it is assumed that detectability of *k*-driver subpopulation is relative to cellular driver load dominance in clinicopathological markers such as Ki-67, and the effect of (spatial)-heterogeneity can be controlled with multiple samples [164].

The threshold for detectability is denoted as β, which is a tumour fraction in the range of $0 < \beta < 1$.The main interest is in knowing the range of *k* -subpopulations subject to detectability restriction $E[S_k(t_N)] \geq \beta E[N(t_N)]$, where t_N is the time when the tumour achieves its target size (e.g. 1 or 4 cm^3 sizes), and β is pre-specified .To investigate the changes in *k*-subpopulation dominance at these target sizes with a restriction imposed, the experiment from the previous section was expanded to consider different values of u: $3.14\text{x}10^{-2} - 3.14\text{x}10^{-7}$ allowing only subpopulations with frequencies greater than or equal to $\beta E[N(t_N)]$.

As suggested in Figure 2.10, driver subpopulation composition remains similar when the tumour is in the 1 cm^3 through 4 cm^3 size range when exponential growth is assumed. A similar effect is also shown in Figure 2.11, minimum differences are observed comparing tumour sizes restricted by β. Detailed values of detectable driver subpopulation composition are in Supplementary Table 2.1.

The minimal difference in the values shown in Figure 2.11 again indicate how subpopulation structure is already well defined before the 1 cm^3 size. This implies that minimum variability in the driver load composition is expected in certain malignancies when a representative sample of the tumour is taken. Additionally, highlights the relevance of using the branching process to study early mutational events and their dominance at diagnosis.

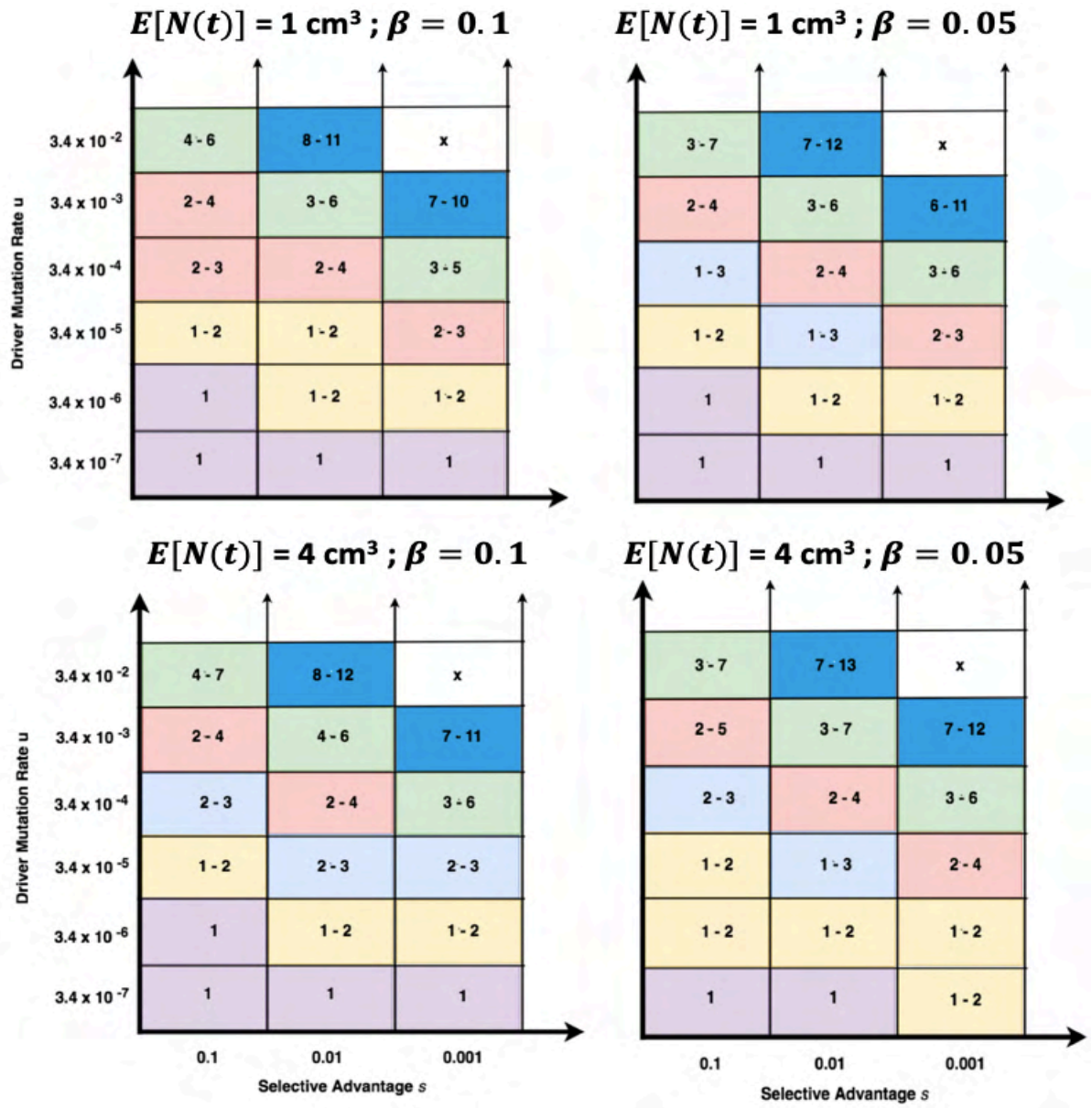


Figure 2.11 Expected range of number of detectable driver subpopulations under two levels of constraint on detection. Parameters are colour coded by similarity. Subpopulations required to be greater or equal than $E[N] * \beta$ to be detected and counted in these tables.

Figure 2.11 demonstrates how cells with higher driver load mostly come from simulations where the selective advantage is weak ($s = \{0.01, 0.001\}$) and the driver mutation rate u is high ($u = \{3.14\text{x}10^{-5}, 3.14\text{x}10^{-4}, 3.14\text{x}10^{-3}\ \&\ 3.14\text{x}10^{-2}\}$)–a similar trend was observed by [165].

Jointly, Figures 2.11 and 2.12 suggest that with the likely driver load in tumour cells ranges from 2 – 3 driver alterations, though, to align this finding with experimental evidence requires a representative sample of the tumour which may require careful sampling of biopsies for pathology [164, 166].

12 Expected Clonal Composition in Tumours Relative to Tumour Size

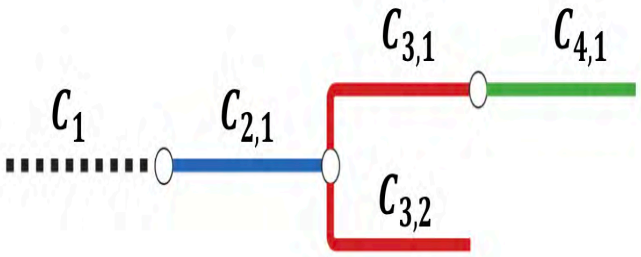
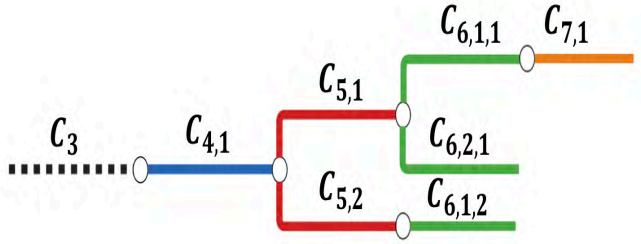
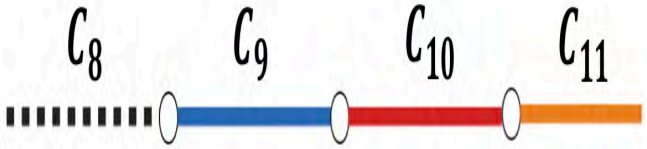
The previous section used the analytical solution of the *k*-subpopulation model to describe the expected driver abundance when a detectability restriction β is applied. The objective of this section is to expand previous results by assigning the expected clonal structure using the analytical solutions of the clonal model.

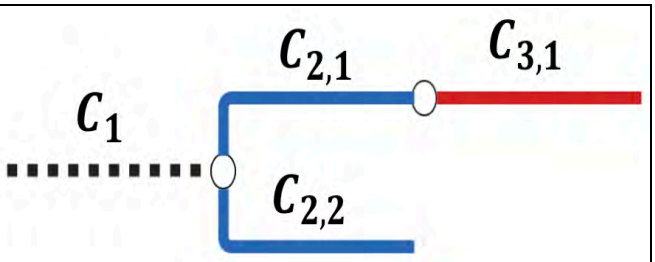
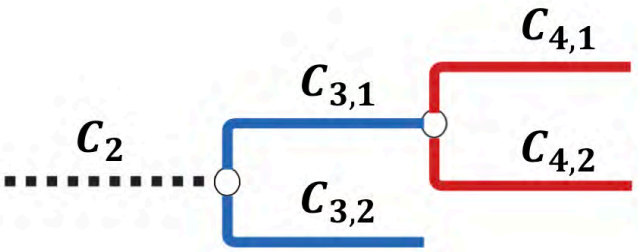
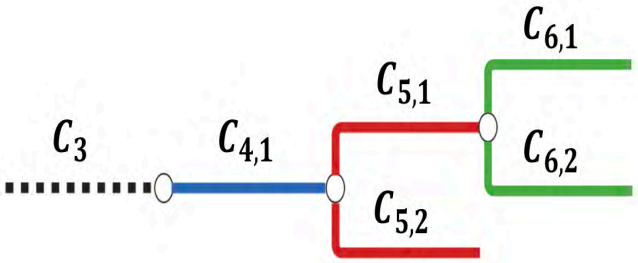
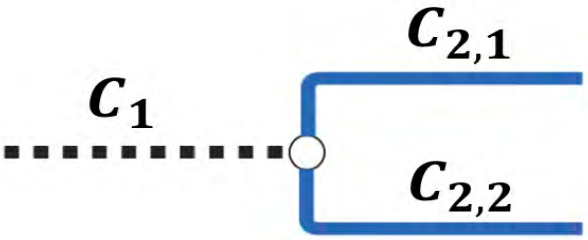
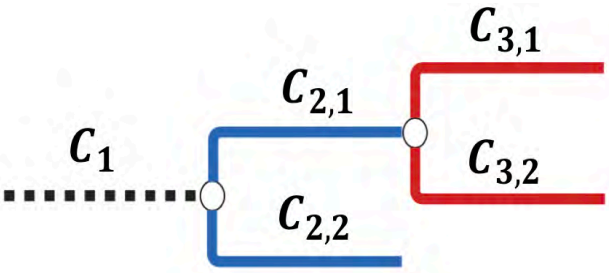
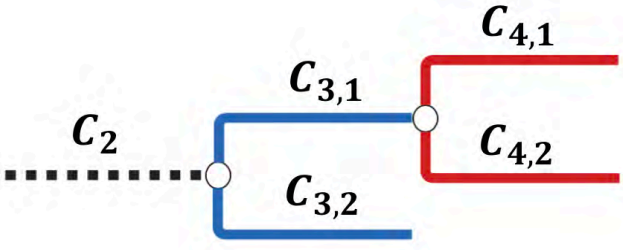
Therefore, an ideal scenario is assumed again in which a representative sample of the tumour is collected and a clonal size detection threshold is applied. As a result, the phylogenies derived in this section are approximations of phylogenies that can be measured by representative sequencing. Clinicopathological markers are related to a phenotypic evaluation of the stage of the tumour (e.g. Ki-67, grade, apoptotic index), whereas sequencing requires inference of the phenotype from genotype (e.g. PyClone clusters as proxy of clonal sweeps).

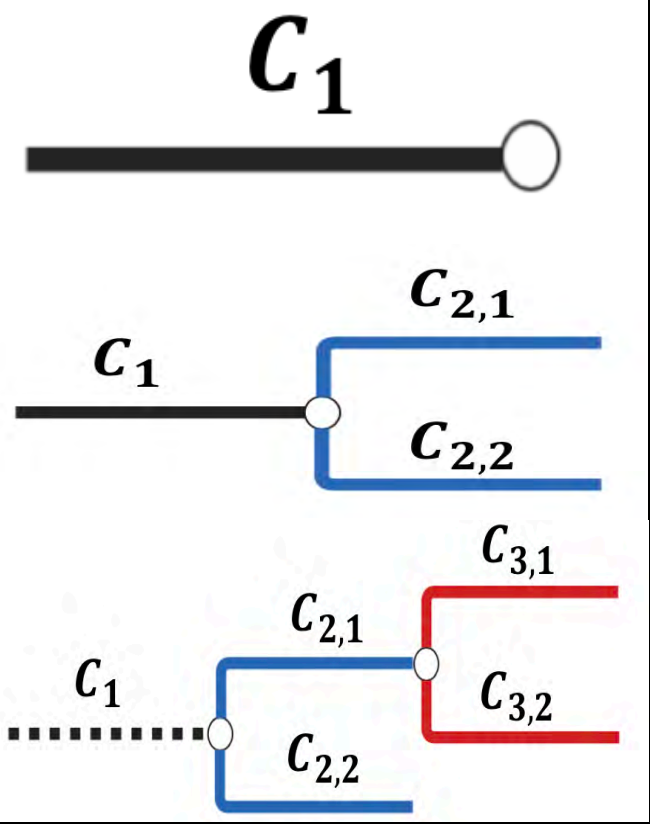
The experimental design was repeated using the analytical solutions of the clonal model. In order to estimate the number of detectable clonal lineages, it is necessary to recursively estimate equations 2.3 and 2.10.a for every clone using until the target tumour size is achieved. Then phylogeny is reconstructed by the clones that are greater or equal than $\beta E[N(t_N)]$.

Table 2.1 shows the expected phylogenies for a tumour of 4 cm^3 size with $\beta = 0.1$ under different combinations of values for *s* and *u*.

Table 2.1 Likely Detectable Phylogenies and Number of Clones at 4 cm^3 Tumour Size

u	*s*	*Clones*	*k-range*	*Clones at 4 cm^3 with $\beta = 10\%$ Phylogenies*
$\mathbf{3.4x10^{-3}}$	0.1	5	2-4	
$\mathbf{3.4x10^{-3}}$	0.01	10	4-7	
$\mathbf{3.4x10^{-3}}$	0.001	11	9-11	

$\mathbf{3.4x10^{-4}}$	0.1	4	2-3	
$\mathbf{3.4x10^{-4}}$	0.01	6	3-4	
$\mathbf{3.4x10^{-4}}$	0.001	8	4-6	
$\mathbf{3.4x10^{-5}}$	0.1	3	2	
$\mathbf{3.4x10^{-5}}$	0.01	5	2-3	
$\mathbf{3.4x10^{-5}}$	0.001	6	3-4	

$\mathbf{3.4x10^{-6}}$	0.1	1	1	
$\mathbf{3.4x10^{-6}}$	0.01	3	1-2	
$\mathbf{3.4x10^{-6}}$	0.001	5	2-3	

Dashed lines represent *k*-driver subpopulations that are not directly detectable but can be inferred by inheritance.
Colours refer to the number of drivers per clone.
Notation of clones $C_{k,i}$ refers to the index *k* as the number of drivers, and index *i* is the branch number.

Table 2.1 shows the expected phylogenies for different combinations of u and s for tumours of 4 cm^3 size with a restriction in clonal frequency $\beta E[N(t_N)]$. It can be seen that there are minor differences between the outcomes of the *k*-subpopulation and the clonal model for the range of *k,* Table 2.1 column *k*-range and Fig. 2.11 bottom left (4 cm^3 and $\beta = 0.1$).

The clonal model also provides an estimate of the number of clusters that a clonality tool such as PyClone and ExPANdS would generate from a sample of a tumour with a given set of mutational parameters (assuming a representative sampling). I compared the expected number of detectable clones with data from Andor et al. [29], who assessed clonality in TCGA data from multiple malignancies using the tools PyClone and ExPANdS.

Using ExPANdS as a clonality caller, it can be seen that tumours with high clonality such as stomach, lung, bladder and melanoma are more likely to match simulations with high mutation rates and moderate/weak average selective advantages, ~ 5- 9 clones in Fig 2.12 vs $u = 3.4\text{x}10^{-3}$ & s: $\{0.1, 0.01, 0.001\}$; $u = 3.4\text{x}10^{-4}$ & s: $\{0.01, 0.001\}$; $u = 3.4\text{x}10^{-5}$ & s: $\{0.01, 0.001\}$ and } and $u = 3.4\text{x}10^{-6}$ & $s = 0.001$. Similarly, tumours with low clonality such as thyroid, prostate, kidney and low-grade glioblastoma are predicted to have moderate to low driver mutation rates and strong to moderate selective advantages, ~2-3 clones in Fig. 2.12 vs $u = 3.4\text{x}10^{-5}$ & $s = 0.1$ and $u = 3.4\text{x}10^{-6}$ & $s = 0.01$. Finally, the rest of the malignancies can be assigned to multiple parameter combinations.

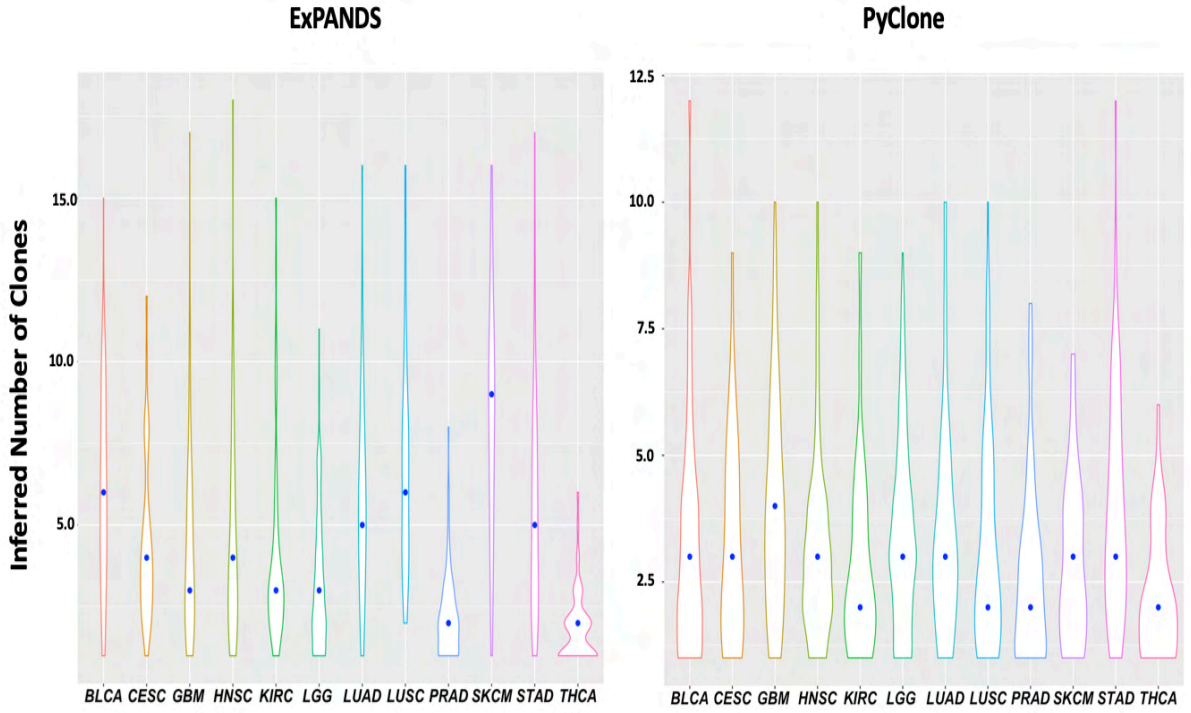


Figure 2.12 Pan-cancer clonality distribution. Inferred clonality reported by Andor et al., using ExPANdS and PyClone, dots represent median values

PyClone and ExPANdS show significant divergence in their distributions of predicted numbers of clones, making difficult to establish consistent patterns in clonal tumour composition. A second concern is that the study of Andor et al. [29] uses single snapshot sequencing which may lead to a biased representation of the clonal structure of tumours. This will impact the comparison with mathematical models like the branching process as clonal clusters (e.g. the dominant clone in the sample) can have an accumulated number of *k*-drivers precluding their fitness estimation.

For instance, THCA shows low clonality using PyClone and ExPANdS that may lead to a false interpretation of the total number of driver mutations in the sample and their assignment to fitness. The dominant clone in the sample (background) could be an evolved clone with a higher accumulated *k* but their low clonality may give the impression of the opposite. Driver annotation is a key metric that clonality tools such as PyClone and ExPANdS do not use to corroborate their cluster calling that can improve evolutionary inference as suggested by [70].

Clonality evaluation alone (e.g. PyClone and ExPANdS) may not provide a complete picture of the effective number of clonal expansions in the sample requiring additional information to infer *s* and *u*.

Overall, this highlights the need to validate the robustness and validity of common clonality tools as well as the challenge of inferring clonal evolution using sequencing technologies, which introduce multiple source of bias that can influence any inferences made about clonal tumour evolution [87, 90].

The branching process can provide additional information to clonality tools to reconstruct tumour's phylogeny and estimate *s* and *u*.

13 Discussion

The main goal of deriving the analytical solutions for the *k*-subpopulation and clonal models is to have a simple framework to translate clinicopathological and molecular markers to reflect the evolutionary landscape of the tumour. Moreover, to add different layers of complexity when needed such as carrying capacity, variation in drug resistant mutation rates, variation in driver mutation rates, etc.

Here I have provided analytical solutions for the *k*-subpopulation and clonal models assuming exponential growth and additive fitness. Moreover, I showed how using the estimator $\hat{\tau}_{k,i,j}$ enables one to approximate the time in which any given clone successfully expands, making it the core of the analytical solutions of the clonal branching process. Solutions to the multi-type branching Galton-Watson process are provided by McDonald et al. [167] considering driver and passenger mutations, though their model requires modification to represent tumour phylogeny and clonal-specific passenger load.

Analytical solutions enable one to approximate the main behaviour of the stochastic process without the need for simulation, which in many cases can be computationally intensive, e.g., when $u = 3.14\text{x}10^{-3}$ and $s = 0.001$. Additionally, scenarios in which the biology of the tumour is uncharacterised or in an advanced state, the analytical solutions provide an approximation of the plausible values of s, u, τ, carrying capacity, etc.

The average selective advantage s impacts the timeframe of tumour growth. Assuming division rates of 1 - 5 days, tumours that grow with a high average selective advantage $s = 0.1$ are expected to grow over a range of 0.5 – 2.7 years, moderate $s = 0.01$, in a range 5 - 27.3 years and weak $s = 0.001$ in a range of 34 – 164 years. Therefore, this indicates that the plausible range of tumour development is captured in the range of s [0.001, 0.1]. Diversity in the expansion growth patterns is proportional to the extinction probability δ as it defines the net-growth rate of the tumour, with higher values of δ more affected by drift. Thus, δ defines the range of evolutionary trajectories that can be generated by initial conditions of s and u. In other words, the net growth of the tumour is associated with the amount of diversity. Our findings are in agreement with Durret et al. [165] showing with a multi-type branching process that heterogeneity is a function of the age of the tumour and the maximum attainable fitness of emerging driver lineages. In their model, mutations occur at a fixed rate per unit of time and not at cell division, thus showing that the main result would not be affected by how driver mutations are implicitly modelled.

There rate of introduction of new driver mutants is modelled differently from Durret et al. [165] In their model fitness is not additive, allowing new clones to have a fitness gradient. However, there is intersection, in that clonal dominance transitions to a higher value of *k* are measurable in a window of time associated with the strength of *s*. This highlights the relevance of studying tumours at common diagnosable sizes and the timeframe associated with parameters *s* and *u* as it can be used to align tumour sizes with predicted driver load and clonal dominance. Furthermore, if representative samples are collected, it can reveal if the model assumptions are appropriate to represent the behaviour of interest.

Analytical solutions of the *k*-driver subpopulation model indicated overlap of the expected driver load within cells (~2 – 4) across different parameter combinations displaying different phylogenetic topologies. Therefore, if a representative sample is taken, an approximation of tumour's *s* and *u* can be done with measurements of the driver load clonality and their average

proliferation in primary tumours. Andor et al. [29] identified that the number of clones did not have a linear association with survival, instead the presence of more than two clones was associated with worse overall survival. Consequently, single snapshot sequencing may require special treatment to maximise its power in predictive models.

Markers of proliferation such as the apoptotic index or Ki-67 are indicators of fitness that can be aligned with analytical solutions of the branching process and help the inference of s and u. Clinical markers have unexplored potential detaching the need of heavily relying on sequencing. Therefore, a secondary motivation in deriving analytical solutions was to provide a simplified representation of the discrete-time branching process to allow an easier comparison with alternative markers of tumour progression.

The analytical solutions showed that expected subpopulation or clonal tumour compositions are defined before reaching the minimum diagnosable size of 1 cm^3, suggesting that clonal architecture of a tumour is established early on in its evolution. This result is representative of liquid tumours or malignancies where carrying capacity is not dependant of k. It is reasonable to assume that in early stages (T0 – T1/2) the effect of the global carrying capacity is reduced without a major impact on the key findings.

Results from Andor et al. [29] indicated the typical clonality distributions for different cancer subtypes that were assayed by whole genome sequencing are around 3 – 4 clones. , This is not far off the 5 - 6 detectable clones reported by Bozic et al.[43] when $s = \{0.001 \ \& \ 0.01\}$ and $u = 3.14\text{x}10^{-5}$. Although because of spatial heterogeneity within tumours one sample may not be representative of the clonal landscape [74] it seems likely the selective advantage of driver mutations in many human malignancies will fall in the range of $[0.01, 0.001]$. The challenge in sequencing studies is the accurate measurement of k as pointed out by [87].

The analytical solutions here provide a fully parametrised framework to study individual or a mixture of subpopulation or clones for different mutational processes. For instance, it can model the number of drug resistance cells that a single clone can accumulate in a given period of time with an accumulated or initial value of s as shown here [123]. The analytical solutions described here can be used to approximate phylogenies of other mutational processes such as copy number or epigenetic alterations. For instance, Laughney et al. [54] reported a missegregation rate of $5.2\text{x}10^{-3}$ that leads to expected phylogenies containing 5 – 11 clones, while 3 – 8 clones could result from a CpG methylation rate of 10^{-4} to $2\text{x}10^{-5}$ [55, 168, 169]. This highlights that mutational processes in cancer (point mutation, copy number and epigenetic) are mixed, occurring at different rates that combine to form the overall average driver mutation rate u.

The variance of the phylogenies can be computed with the framework described here. The advantage of computing this variation is in determining with branches are more likely to change. However, it is expected that the phylogenies with more variation are the ones that have lower average selective advantage s and higher average driver mutation rate u. A secondary advantage is that the phylogenies derived here, can be recovered when no bulk sequencing data is available but other measures of clonal dominance or proportion have been collected (e.g. fluorescence in situ hybridization (FISH), FACS, CyTOF, single-cell sequencing, RNA-seq, etc.). These techniques are commonly used to measure markers of interest and can provide enough information to compare with the results values presented here.

Given that cancer is a collection of dynamic diseases, a one-size-fits-all model is underpowered to capture all aspects of tumour evolution. There are implicit assumptions on the solutions presented here that introduce limiting factors on their applicability. The conditions assumed here better describe tumours with reduced effect of carrying capacity (i.e. exponential growth), standardised effects in selective advantage conferred by driver alterations and no spatial constraints. Therefore, liquid tumours and solid malignancies at early stage may be replicative of the dynamics of the solutions provided. However, one of the main goals of identifying the analytical solution of the additive fitness model branching process is to have a simple framework that allows for scalability to features that are not considered or over simplified. Controlled experimental conditions are crucial to reveal tumour-type-specific properties that would require specific modelling to incorporate into the current solutions.

Tumour heterogeneity introduces a challenge on how to sample the tumour. Non-representative samples may reflect a biased evolutionary state of the tumour with consequences for downstream analyses (e.g. single sample sequencing). Integration of molecular and clinicopathological markers are required in the field for the branching process to have a significant translational impact. The CTS5 [170-172] score is an example of a model integrating clinical markers with prognostic and predictive impact in breast cancer. Similarly, a recent clonal evolution model has shown power to predict progression-free survival demonstrating the role of evolutionary theory in informing clinical manifestation [173].

Future work needs to evaluate how to integrate clinical, histopathological, and molecular markers with the branching process to gain evolutionary knowledge and determine its prognostic and predictive power.

Moving forward with the analytical solutions, Chapter III is going to explore the outcomes of computer simulations with different parameters of *s* and *u* to determine the expected tumour compositions and phylogenies.

13 Appendix and Supplementary Figures and Tables

A.2.1 Model definitions, Expectation & Variance of the Discrete-Time Branching Process

The clonal evolution model is a Markov chain of the form of Galton-Watson process. Every time step a cell divides into two cells $P(X_j = 2) = b_j$, dies or stagnates with probability $P(X_j = 0) = d_j$ with $b_j + d_j = 1$. Driver mutations reduce the probability of death as an additive factor of s such that $d_j = \frac{(1-s)^j}{2}$ and driver events occur with probability $P(X_j = 1) = b_j(1-u)$. The following table shows the differences from the k-subpopulation model and the clonal model.

$\mathbf{P(X_j = n)}$	***K*-SUBPOPULATION**	**CLONAL**
$\mathbf{P(X_j = 2)}$	$b_j(1-u)$	$b_j(1-u)$
$\mathbf{P(X_j = 1)}$	$b_j u;\ X_{j-1}b_{j-1}u$	$b_j u$
$\mathbf{P(X_j = 0)}$	d_j	d_j

Recalling that branching process have generating functions of the form,

$$f^{(i)}(s_1, s_{2,}\ldots, s_n) = \sum_{i=0}^{\infty} p_i\, s^i \qquad\qquad |s| \leq 1$$

For the number of offspring produced by a cell is given by generating function for the *k*-subpopulation model,

$$f^{(i)}(s_1, s_{2,}\ldots, s_n) = d_j + b_j u s_j s_{j+1} + b_j(1-u)s^2$$

The equivalent probability generating function for the clonal model is,

$$f^{(i)}(s_1, s_{2,}\ldots, s_n) = d_j + b_j u s_j + b_j(1-u)s^2$$

With $0 \leq s_\alpha \leq 1$ and $\alpha = 1,2,\ldots$ The model is updated recursively for the k-subpopulation model as,

$$S_k(t+1) \;=\; S_k(t) \;+ B_k \;-\; D_k + M_{k-1};\ \text{With } S_1(0) = 1.$$

With the total number of cells in the tumour as $N(t+1) = \sum_{k=1} S_k(t+1)$. The clonal model is updated as,

$$C_{k,i,j}(t+1) \;=\; C_{k,ij}(t) \;+ B_k \;-\; D_k;\ \text{With } \; C_{1,*}(0) = 1.$$

With the total number of cells in the tumour as $N(t+1) = \sum_{\{k,i,j\}=1} C_{k,i,j}(t+1)$. Where $[B, D, M] \sim multinom(X(t),\ [b_k(1-u), d_k, b_k u])$

In the context of tumour growth, we are interested in the super-critical cases of the branching process where $1+b-d > 1$.

Expectation

Cells X_j have probability extinction probability of $\delta_j = {}^{d_j}\!/_{b_j}$, with δ_j is possible to evaluate the expected number of cells of a surviving lineage as $E[X_j^+] = {}^{1}\!/_{1-\delta_j}$.

We are interested in the expected value of a subpopulation after n generations $E[Z_n]\ of\ a$ branching process $\{Z_0, Z_1, \ldots Z_n\}$ with initial condition $Z_0 = 1$ and offspring distribution Y such that $Z_n = Y_1 + \ldots + Y_{Z_{n-1}}$. The expectation of the offspring is known as $E[Y] = \mu$. Then,

$$E[Z_n] = E\big[E[Z_n \mid Z_{n-1}]\big] = E\left[\sum_{i=1}^{Z_{n-1}} Y_i\right] = E[\mu Z_{n-1}]$$

$$E[Z_n] = \mu^n E[Z_0] = \mu^n$$

The clonal $C_{k,i,j}$ and the initial founder S_1 for the clonal and k-subpopulation models have the same form and therefore μ is derived of a single individual as,

$$E\left[C_{k,i,j}(t=1)\right] = C_{k,i,j}(t=0)\left[b_{k,i,j}(1-u) - 1 + b_{k,i,j}\right]$$

$$E\left[C_{k,i,j}(t=1)\right] = b_{k,i,j}(2-u) = \mu_{k,i,j}$$

$$E[S_1(t=1)] = b_k(2-u) = \mu_k$$

With the previous statement it is possible to define the expected number of cells in the clonal an k-subpopulation models $E[N(t+1)]$ as,

$$E[N(t+1)] = \sum_{\{k,i,j\}=1} C_{k,i,j}(t+1) = \sum_{\{k,i,j\}=1} \frac{\left[b_{k,i,j}(2-u)\right]^{t-\hat{\tau}_{k,i,j}}}{1-\delta_{k,i,j}}$$

$$t - \hat{\tau}_{k-1,i,j} \geq 0$$

$$E[N(t+1)] = \sum_{k=1} S_k(t+1) = \sum_{k=1} \left(\frac{[b_k(2-u)]^{t-\tau_k}}{1-\delta_k} \right) + \mathbb{I}_{k>1}([b_{k-1}u(2-u)]^{t-\tau_{k-1}})$$

$$t - \tau_{k-1} \geq 0 \; ; \; t - \tau_{k-2} \geq 0$$

Where τ is the time when a new driver lineage successfully expands and $\hat{\tau}_{k,i,j}$ the estimator for the clonal case.

Variance

With the variance of the offspring distribution of the branching process $Var(Y) = \sigma^2$, is possible to determine the variance of the branching process by the law of total variance as,

$$Var(Z_n) = E[Var(Z_n|Z_{n-1})] \; + \; Var(E[Z_n|Z_{n-1}]) = \sigma^2 E[Z_{n-1}] + \mu^2 Var(Z_{n-1})$$

$$Var(Z_n) \; = \; \sigma^2(\mu^{2n-2} + \mu^{2n-3} + \ldots + \mu^{n-1})$$

The $Var(Z_n)$ can be seen as a geometric series and in the supercritical case where $1 - b - d > 1$ the variance is,

$$Var(Z_n) \; = \sigma^2 \mu^{n-1} \left(\frac{1-\mu^n}{1-\mu} \right)$$

For the *k*-subpopulation and clonal models the variance is estimated as,

$$Var(C_{k,i,j}(t=1)) = b_{k,i,j}(1-u)[1 - (b_{k,i,j}(1-u))]$$

$$Var(S_1(t=1)) \; = b_k(1-u)[1 - (b_k(1-u)]$$

$$Var(C_{k,i,j}(t+1)) = \left[\left(\frac{1}{1-\delta_{k,i,j}} \right) b_{k,i,j}(1-u)[1 - (b_{k,i,j}(1-u))] \right] \mu^{n-1} \left(\frac{1-\mu^n}{1-\mu} \right)$$

$$Var(S_k(t+1)) = \left[\left(\frac{1}{1-\delta_k}\right)(b_k(1-u)[1-(b_k(1-u))] + \mathbb{I}_{k>1} u b_{k-1}(1-ub_{k-1}))\right]\mu^{n-1}\left(\frac{1-\mu^n}{1-\mu}\right)$$

Term $\left(\frac{1}{1-\delta_k}\right)$ is to account for successful lineages.

A.2.2 Expected Number of Cells in Bozic Model by Inhomogeneous First-Order Differential Equation

We begin by considering a deterministic analogue of the Bozic model [43]. Let $Z_k(n)$ denote the number of type-k cells at time-step n, the model evolves setting $d_k = \frac{(1-s)^{k+1}}{2}$ according to:

$$Z_0(n+1) = Z_0(n)(1+(1-u)(1-d_0)-d_0) = Z_0(n)(2-u)(1-d_0)$$
$$Z_k(n+1) = Z_k(n)(2-u)(1-d_0) + u(1-d_{k-1})Z_{k-1}(n)$$

Solving the first linear first-order difference equation, we have:

$$Z_0(n) = Z_0(0)\big((2-u)(1-d_0)\big)^n = \big((2-u)(1-d_0)\big)^n$$

Note that the particular solution associated to an inhomogenous first-order differential equation $Z(n+1) = kZ(n) + Ak'^n$, is: $Z(n) = \frac{Ak'^n}{k'-k}$, confirmed by direct substitution:

$$\frac{k'Ak'^n}{k'-k} = \frac{kAk'^n}{k'-k} + Ak'^n$$
$$\frac{k'}{k'-k} = \frac{k}{k'-k} + 1$$

Therefore, the solution for $Z_1(n)$ is:

$$Z_1(n) = A_{11}\big((2-u)(1-d_1)\big)^n + \frac{u(1-d_0)\big((2-u)(1-d_0)\big)^n}{(2-u)(d_1-d_0)}$$

Where the A_{11} coefficient is chosen so that $Z_1(0) = 0$ – that is, the coefficients add to zero. Form a matrix A_{kj}, containing the coefficient for $\big((2-u)(1-d_j)\big)^n$ in the analytic solution for $Z_j(n)$. The algorithm for generating the k-th row of this matrix, given the (k-1)-th row is:

- Multiply the $A_{(k-1),j}$ entry from the preceding row, by $\frac{u(1-d_{k-1})}{(2-u)(d_k-d_j)}$, and copy this into the A_{kj} position.

Set A_{kk} to be the negative of the sum of the other entries on the row

In this way, the analytical solution for the deterministic analogue of the Bozic model can be expressed:

$$Z_k(n) = \sum_{j=0}^{k} A_{kj}\left((2-u)\big(1-d_j\big)\right)^n$$

A.2.3 Adding Size Dependant Carrying Capacity

One of the effects not modelled in branching processes is the effects on clonal proliferation by a carrying capacity based on size, the aim of the carrying capacity is to limit the expansion as resources become scarce. A k-dependant β_k or N-dependant β_N carrying capacity can be added on the model as a proportional penalty in a recursive form,

$$\beta_k = (1 + ks) - ks\left(\frac{S_k(t) - 1}{N_k^* - 1}\right)$$

$$S_k(t+1) = S(t) * \beta_k; \quad \beta_k \geq 1$$

Where N_k^* is the number of cells at which cells are not allowed to expand anymore and their expansion plateaus.

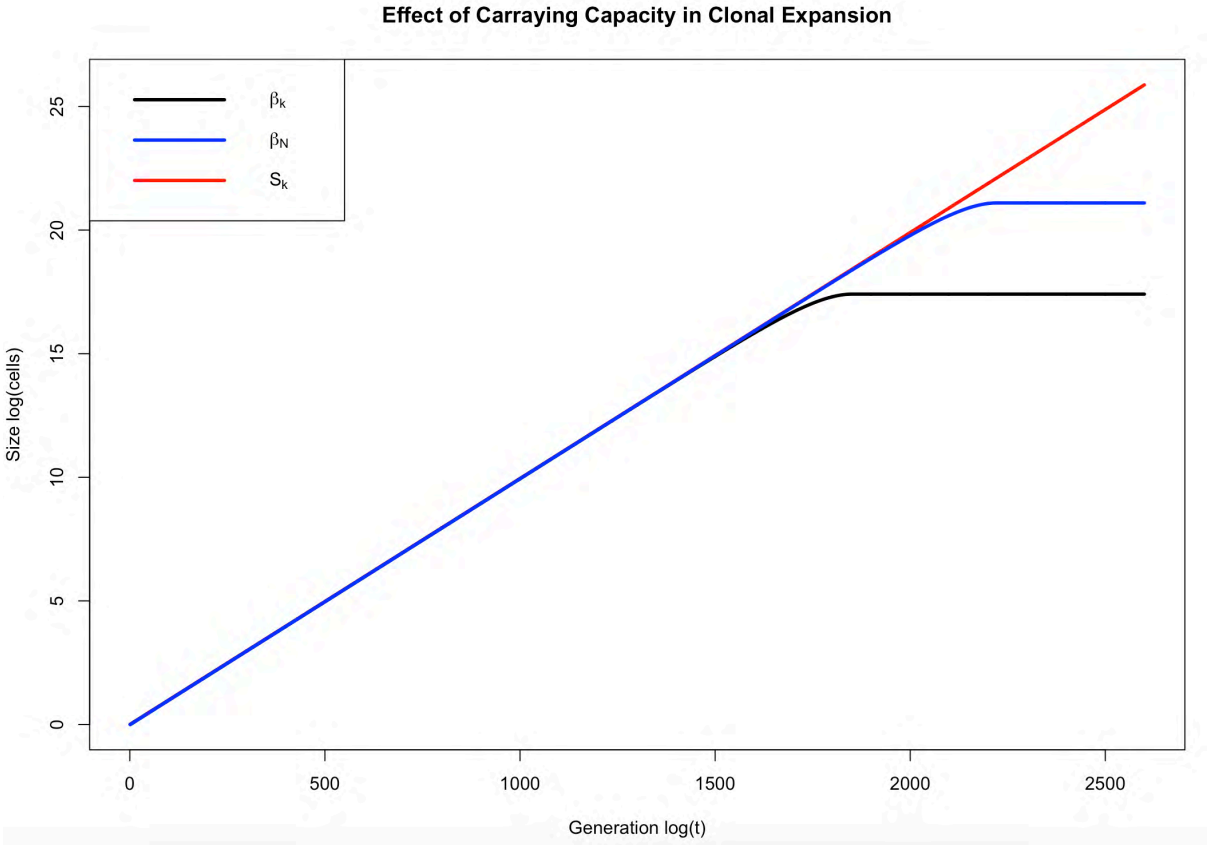


Figure S2.1 Carrying capacities. Clonal growth was generated using the parameters reported by Bozic et al., $s = 0.004$ and $u = 3.14\text{x}10^{-5}$ with $k = 1$.

A.2.4 Critical Crossover Point of Clonal Waves

We can identify the expected point in time t_ε when two different subpopulations are equal in size by solving for t in this setting $(b_k(2-u))^{t-\tau_{k-1}} = [b_{k+1}(2-u)]^{t-\tau_k}$,

$$t_\varepsilon = \frac{\tau_{k-1}\, log([b_k(2-u)]^{t-\tau_{k-1}}) - \tau_k[b_{k+1}(2-u)]^{t-\tau_k})}{log(b_k(2-u)) - log(b_{k+1}(2-u))}$$

It is expected that $S_{k+1} \geq S_k$ for $t \geq t_\varepsilon$, the following picture is an example of the estimated crossover point.

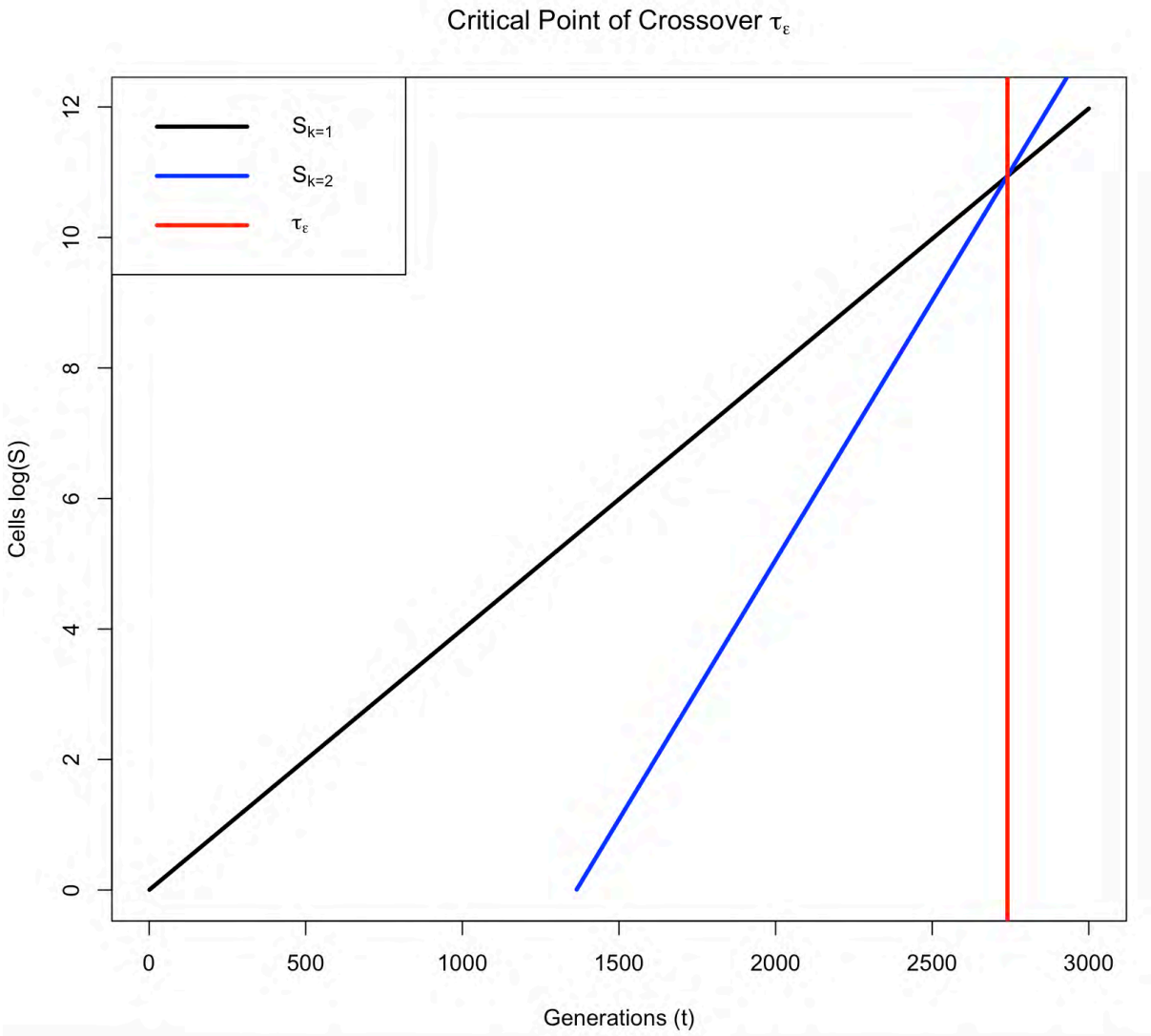


Figure S2.2 Crossover point. Clonal growth was generated using the parameters reported by Bozic et al., $s = 0.004$ and $u = 3.14\text{x}10^{-5}$. Black, the expected growth of subpopulation S_1. Blue, expected growth of subpopulation S_2. Red, the time τ_ε when S_2 outcompetes S_1.

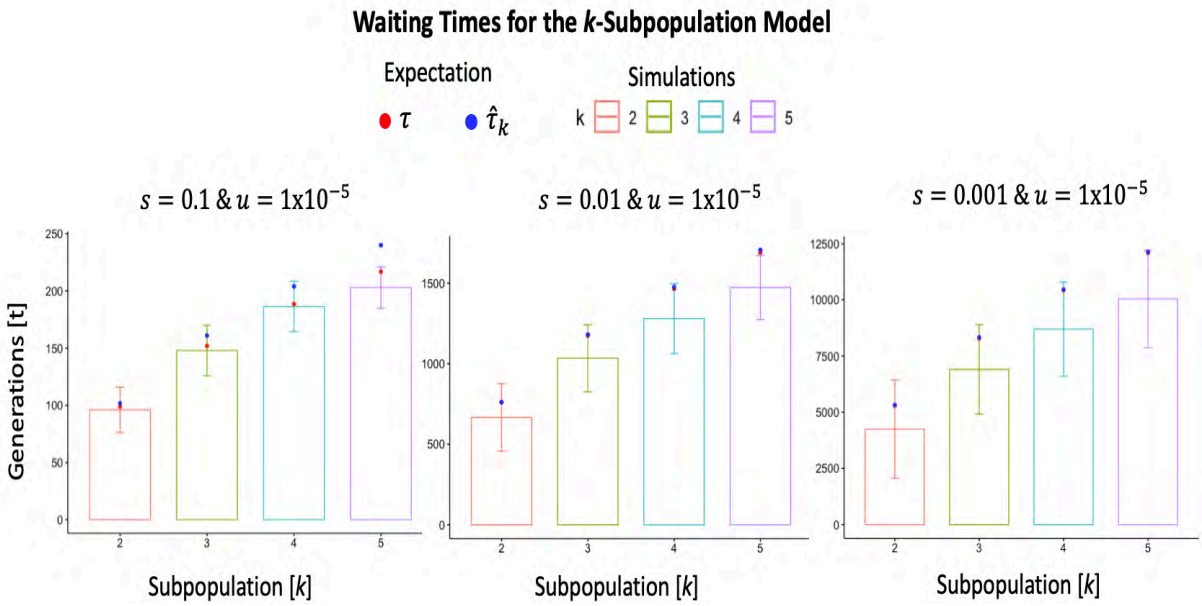


Figure S.2.3 Comparison of $\hat{\tau}_k$, τ and computer simulations with the *k*-subpopulation model. 300 (successful) simulations were used to contrast expected solutions of τ and $\hat{\tau}_k$ with initial parameters $u = 1\text{x}10^{-5}$, $s: \{0.1, 0.01, 0.001\}$ & $k: \{1, \ldots, 5\}$. Error bars represent the 25th and 75th percentiles of simulations, in red the reported value of τ and in blue $\hat{\tau}_k$ with $\varepsilon = 0.0$.

A.2.5 Procedure to Recover Expected Value of Clonal or Tumour Size

With the expectation equations derived, it is possible to formalise a procedure that can recover the proportions of unique subclonal populations given an arbitrary tumour size. The procedure requires an inheritance rule based on the selective advantage *s* [43, 44], the total number of drivers *K* to recover the tumour, driver mutation rate *u*, division time *P* and the target tumour size *N*.

Step 1) Initialise the model with provided input parameters *k* and *s* (for instance with Bozic),

$$d_k = 0.5(1 - s)^k$$

$$b_k = 1 - d_k$$

Step 2) Estimate values of τ_k, by [42, 43] or $\hat{\tau}_k$,

$$\tau_k \approx \underset{1 \le t \le x}{\text{argmin}}(|cumsum(b_k u[b_k(2-u)]^t) - c - \varepsilon|) = \hat{\tau}_k$$

Step 3) Identify the *K-th* proportion that approximates *N*. We have 2 cases, when *N* is achieved by a single S_k or when *N* is achieved by a mixture of S_k. In the case one only one *k* we have,

$$E[N(t+1)] = \left(\frac{1}{1-\delta_k}\right)[b_k(2-u)]^t$$

We can solve for *t* provided $b_k > d_k$ and obtain t_n, the number of generations required for a single S_k to grow at a given size *N*,

$$t_n = \frac{log\left(N(t)\right)}{log\left(b_k(2-u)\right)};$$

$$E[N(t+1)] \approx \left(\frac{1}{1-\delta_k}\right)[b_k(2-u)]^{t_n}$$

For the second case $k > 1$ evaluate from $t{:}\{1, t_m\}$ if the tumour target N is achieved restricting for subpopulations in their interval $S_k{:}\{\tau_{k-1}, t_m\}$ and clones $C_{k,i,j}{:}\{\hat{\tau}_{k-1,i,j}, t_m\}$.

$$[K,t] \approx \min_{1\le t\le t_m}\left(\left|\left(\sum_{k=1}^{K} E[S_k(t+1)]\right) - N\right|\right)$$

$$[K,t] \approx \min_{1\le t\le t_m}\left(\left|\left(\sum_{k=1}^{K} E\left[C_{k,i,j}(t+1)\right]\right) - N\right|\right)$$

Then choose all $E[S_k(t)] \ge N\beta$ and $E\left[C_{k,i,j}(t)\right] \ge N\beta$ to determine the number of subpopulations or clones detectable by fraction β.

Supplementary Table 2.1 Expected Driver Subpopulation Composition of Tumours

u	*s*	*K at 1 cm³*	*k at 1 cm³* β *=10%*	*k at 1 cm³* β *=5%*	*K at 4 cm³*	*k at 4 cm³* β *=10%*	*k at 4 cm³* β *=5%*
$3.4x10^{-3}$	0.1	10	2 - **4**	2 - **4**	11	2 - **4**	2 - **5**
$3.4x10^{-3}$	0.01	12	3 - **6**	3 - **6**	13	4 - **6**	3 - **7**
$3.4x10^{-3}$	0.001	20	7 - **10**	6 - **11**	22	7 - **11**	7 - **12**
$3.4x10^{-4}$	0.1	7	2 - **3**	1 - **3**	7	2 - **3**	2 - **3**
$3.4x10^{-4}$	0.01	7	2 - **4**	2 - **4**	8	2 - **4**	2 - **4**
$3.4x10^{-4}$	0.001	10	3 - **5**	3 - **6**	11	3 - **6**	3 - **6**
$3.4x10^{-5}$	0.1	5	1 - **2**	1 - **2**	5	1 - **2**	1 - **2**
$3.4x10^{-5}$	0.01	5	1 - **2**	1 - **3**	5	2 - **3**	1 - **3**
$3.4x10^{-5}$	0.001	6	2 - **3**	2 - **3**	7	2 - **3**	2 - **4**
$3.4x10^{-6}$	0.1	3	**1**	**1**	4	**1**	1 - **2**
$3.4x10^{-6}$	0.01	4	1 - **2**	1 - **2**	4	1 - **2**	1 - **2**
$3.4x10^{-6}$	0.001	4	1 - **2**	1 - **2**	5	1 - **2**	1 - **2**

K: total number of the driver subpopulations occurring in the tumour,
k: the measurable driver subpopulations according to their proportions in the tumour
The maximum number of accumulated drivers that can be measured at the tumour fraction detectability limit β for each set of parameters are shown in bold

Chapter III

1 Outline

This chapter describes the implementation of clonal evolution models based on the branching process, explores and compares their properties and assesses their relevance for clinical applications and clonal reconstruction.

The following clonal models were implemented:

1) **Additive fitness**: changes in s are proportional to k, and u remains unchanged. This is going to be used as a baseline.
2) **Stickbreaking**: values of s are sampled from a fitness distribution bounded by parental fitness, u remains unchanged. This will explore how more variation in s changes outcome.
3) **Increased mutation**: s is similar to the additive fitness model, u can change with 75% probability in second driver lineages. This will explore the effect of changing the mutation rate on evolutionary trajectories.
4) **Neutral**: to study the accumulation of stochastic passenger signals for different values of s and k. This will explore how the number of passengers accumulates over time.

With the models implemented the goal was to explore how different combinations of s and u affect tumour clonal compositions, driver heterogeneity and growth dynamics by simulating multiple instances of the branching process and analysing their patterns.

All models were simulated with multiple parameter combinations. Numerous snapshots were saved to record features of interest during the course of each simulation, generating a database of positive selection and neutral evolution. By taking snapshots of numerous replicates per parameter combination, I aim to parallel the diversity and complexity in tumour heterogeneity seen in the clinic.

As a result, I have generated the largest database of results from branching process models of tumour evolution, comprising of 14,470 tumours in total. The database is intended to be used to better understand clonal evolution, drug resistance, the consequences of heterogeneity and for comparison to data from cancer sequencing studies.

First, I will describe the implementation and technical aspects of the different models, such as the number of replicates generated per parameter combination and the properties of the tumour stored at every snapshot.

Second, I will describe the main features of all the simulations, focusing in two milestone tumour sizes, 1 cm^3 (early detection) at and 4 cm^3 (late detection). Said features include metrics of tumour expansion, diversity and clonal composition, as measured by our current technologies, phylogenies and similarities between simulated cancer cell fractions (CCF) in the context of patient classification.

Third, I will show the dynamics of clones that acquired drug resistance during tumour expansion and identify conditions that increase the odds of their emergence, such as reduced average selective advantage s.

Finally, I will calculate the number of detectable passengers at different sequencing cut-offs in cases when sequenced samples are neutral and how that information can be used to establish parameters s and u.

As a result, I will show how the mathematical models of tumour evolution can provide biological insight by elucidating how driver mutations are linked with tumour development and clonal heterogeneity.

2 Introduction

Accurate reconstruction of tumour evolution remains a significant challenge due to the limitations of measuring tumours at the molecular level. Multiple studies have focused on modelling the initial conditions that fuel tumorigenesis, namely the fitness and mutation rates of diverse malignancies and their association with prognosis. Yet a testable framework to connect biological hypothesis with experimental data remains lacking.

With the recent advances in clonal evolution modelling said framework is now possible and forms the focal point of my project. Its backbone is the discrete-time branching process model, which allows parametrisation of tumours based on their average selective advantage *s* and average driver mutation rate *u*. Therefore, how the values of *s* and *u* govern tumour progression in both primary and metastatic settings need to be established.

Current estimates of the average selective advantage *s* for somatic driver point mutations range from 0.004 to 0.02, as reported for different malignancies such as ovarian cancer, colorectal cancer and glioblastoma [42, 43]. This discrepancy highlights the current need to identify an approach to approximate tumour fitness at the subtype and patient levels.

Similarly, mutation rates in cancer vary significantly across processes that influence clonal composition in tumours [17]. As shown in Table 3.1, the reported average driver mutation rate $3.4\text{x}10^{-5}$ consider only point mutation changes, but copy number alterations and genome doublings can also affect fitness [96, 97]. Consequently, it is necessary to consider a range of average driver mutation rates to $3.4\text{x}10^{-6}$ - $3.4\text{x}10^{-4}$ [130], to accommodate all mutational processes that fuel tumour growth.

Table 3.1 Reported Mutation Rates in Cancer

Process	**Rate**	**Reference(s)**
Passenger	0.016	[58]
Missegregation	$2.7\text{x}10^{-3} - 6\text{x}10^{-3}$	[54]
CpG Methylation	10^{-4}	[168, 169]
Amplification/Deletion	$10^{-5} - 10^{-4}$	[174]
SNV Driver	$3.4\text{x}10^{-5}$	[43]
Resistance	10^{-9}	[44, 45]
Point Mutation	$10^{-10} - 10^{-9}$	[175, 176]

As a result, defining the ranges of *s* and *u* that capture realistic values for mutational process and driver fitness effects for diverse malignancies is an essential goal for modelling clonal evolution.

Ciriello et al. [97] established the role of point mutations and copy number alterations in a number of cancer subtypes in The Cancer Genome Atlas (TCGA). Interestingly they showed that malignancies evolve by one of the two mutational processes measured, point mutations or copy number but rarely by both. A subsequent study in TCGA of estimates of clonality distributions from two standard clonality assessment tools, PyClone and ExPANdS indicated there was prognostic power in the number of inferred clones identified by ExPANdS but not by PyClone. Similar inconsistencies were found in clonality assessments in pan-cancer studies [30, 177].

This exemplifies the necessity of further investigation into the set of conditions and mutational processes that can recapitulate the observed clonal compositions of tumours reported in pan-cancer sequencing studies, and how these are associated with disease progression and survival.

This challenge motivates the development of a clonal evolution database comprising a comprehensive sample of possible tumours in a combinatorial space covering realistic values of s and u. This database enables investigation of the role of tumour fitness and diversity under different conditions, for comparison with real tumours. It aims to provide a comprehensive landscape of clonal evolution to gain better understanding of how clonal fitness and driver alterations manifest in intratumoral heterogeneity and drug resistance phenotypes.

The overarching goal of this chapter is to provide the foundation for a testable framework that allows the comparison of experimental with simulated data for determining the evolutionary properties of tumours.

3 Hypothesis and Aims

Hypothesis: The clonal branching process can be used as a tool to investigate the main dynamics of tumour formation, heterogeneity, drug resistance and phylogenetic reconstruction.

Aim I: Generate a database of simulated tumours using the branching process under three different positive selection models and a neutral evolution model.

Aim II: Perform exploratory data analysis on the collected results from Aim I and establish association between starting model parameters and simulation results.

Aim III: Investigate the biological implications by comparing the models to identify patterns of tumour growth, intratumor heterogeneity, emergence of drug resistance and clonal compositions of tumours.

4 Methods

Data collection and sample generation is described in the next Section (5). The different benching process models were implemented in Python 2.7. Simulated tumours generated multiple files:

1. Growth dynamics: updated every iteration, containing: time, tumour size, clones, Shannon entropy, Shannon equity, Rich-Gini Simpson index of diversity, number of drug resistant clones.
2. Summary values: generated at milestone sizes, containing: number of iterations, tumour size, number of years, number of clones, total mutations, total resistance mutations,

fraction of the tumour explained by the top 100 clones, Rich-Gini Simpson index of diversity, Shannon entropy, Shannon equity, tumour fitness, average proliferation and average δ.

3. Subpopulation values: generated at milestone sizes, containing: k, number of clones and number of cells.
4. Properties of the top 100 clones: generated at milestone sizes, containing for every clone: clone id, number of drivers, frequency in the tumour, CCF relative to tumour, CCF relative to the sample, fitness, drug resistant status, number of cells and δ.
5. Properties of drug resistant clones: generated at milestone sizes, containing for every clone: clone ID, resistance id, number of drivers, born time, fitness, cells and δ.

Using the properties of the files mentioned above, the boxplots of the properties of the tumours at sizes 1 cm^3 and 4 cm^3 were generated.

Detectable driver ant tumour compositions are generated by filtering clones in the top 100 files below a certain CCF of interest, i.e. 10% and 5%. Once the filtering is performed, all the clones for the replicates of s and u are merged to generate their distributions or establish tumour phylogenies.

Recurrent phylogenies were recovered using the clone ID generated in the properties of *the top 100 clones* files. Every clone has an ID that allows it to be associated with its position in the phylogeny. The description of the ID was shown in Figure 2.8. Every clone is labelled by the number of drivers, parental clone ID and its position in the phylogeny branch as $C_{k,i,j}$. The proportions of the phylogenies were calculated as a fraction of the total for a given parameter combination of s and u.

5 Data Collection and Database Storage

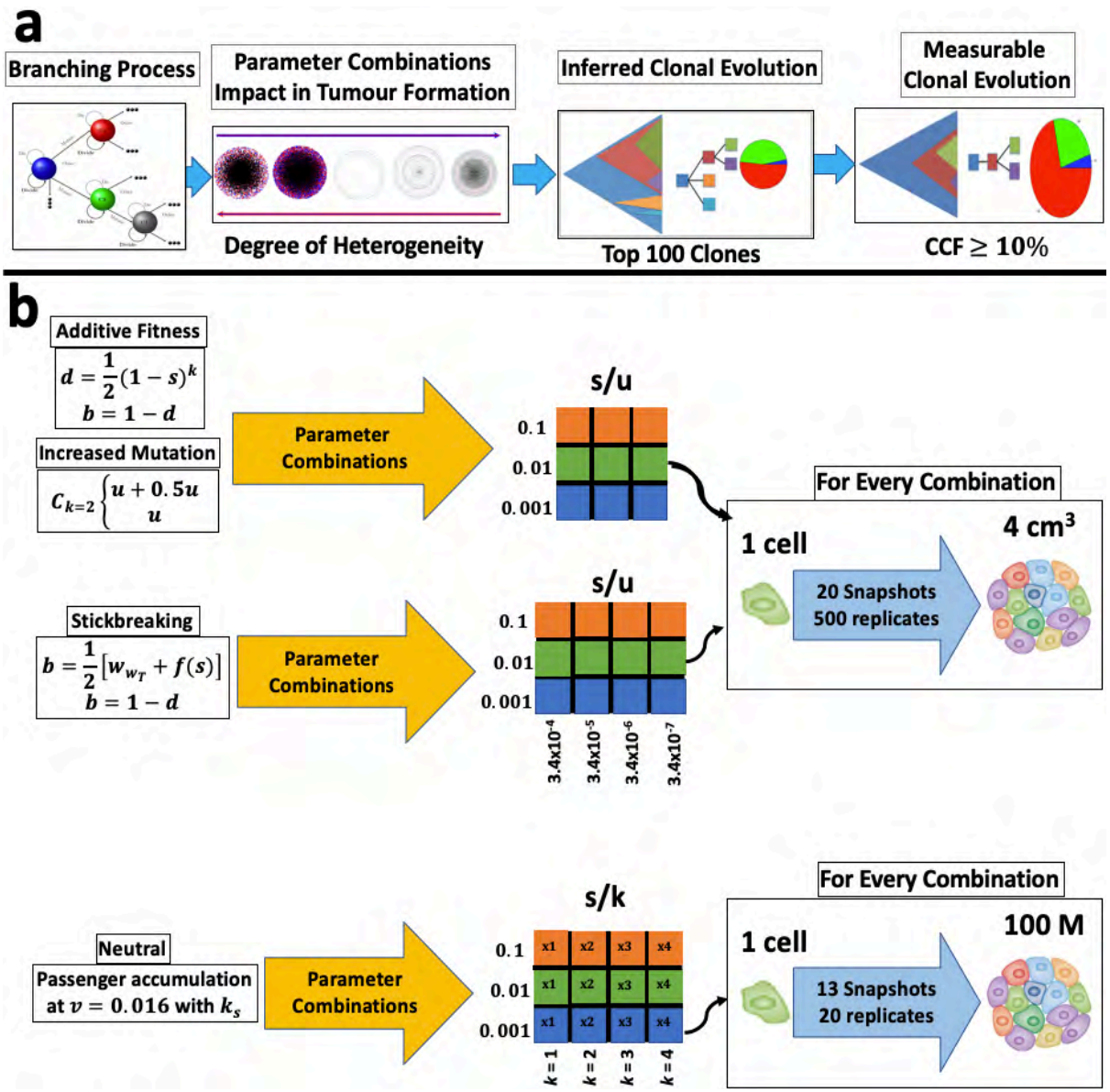


Figure 3.1 Tumour evolution models using the discrete branching process. **a,** the main goal of simulating the tumour evolution models with different parameters is to explore the effects on diversity and how the simulations may be used to infer measurable clonality. **b**, Description of the data generation process for the different models implemented here, showing the different parameter combinations used, numbers of replicates and numbers of snapshot samples saved for each.

Although analytical solutions provide the expected behaviour of the parameters of interest, they do not provide the complete picture of the diversity of clonal trajectories attainable through *in silico* simulation. The advantage of using computational models of tumour evolution is they complement methods that infer clonal composition from sequencing data and determine the evolutionary properties of real tumours, as displayed in Figure 3.1.a.

To investigate the role of multiple parameter combinations of average selective advantage s and average mutation rate u in clonal evolution using the branching process, I implemented the following models,

1. **Additive Fitness**: Expands the model of Bozic et al. [43], to record diversity at the clonal level. The average selective advantage s changes proportional to k and the average driver alteration u remains unchanged.
2. **Stickbreaking model**: Proposed by of Chowell et al. [44], it aims to model more stable fitness effects than the additive fitness model. The average selective advantage s changes by sampling from a fitness distribution bounded by the parental fitness and a hypothetical maximum (defined as 1). The average driver mutation rate u remains unchanged.
3. **Increased mutation rate model**: Changes in the average selective advantage s are the same as the additive fitness model. The average driver alteration rate u can change as $C_{k=2*} = u + 0.5u$ with probability of 75%. This model aims to describe the impact of changing the mutation rate in the dynamics of tumour growth.
4. **Neutral evolution**: aims to describe the passenger accumulation per subpopulation at a passenger mutation rate of $\nu = 0.016$ for every k. This model will be used to evaluate if the passenger signal can be used to decimate between parameters.

Collectively these models cover different possibilities of tumour development. The additive fitness model is going to be used a reference because fitness changes in clonal subpopulations occur in an additive fashion.

The stickbreaking model allows for more variation in fitness to accommodate more complex cases when constant accumulation cannot describe tumour development, this model is going to be used as the gold standard due to the ability of the stickbreaking algorithm to replicate the patterns of evolution [138].

While the stickbreaking model changes the average selective advantage s to allow more variation, the increased mutation rate model aims to model how elevated mutation rates u that can arise as a consequence of alterations in tumour suppressor genes.

To explore the translational application of the tumour evolution models, I modelled selective advantages ranging from one order of magnitude above the maximum selective advantage coefficient reported by Durett et al. [136, 152] down to the value reported by Bozic et al. [43] ($s = 0.001 - 0.1, 0.001$ instead of 0.004 to have steady changes). In the context of the average driver mutation rate u, to capture the main two mutational process, copy number alteration and point mutations, I chose a range of $u = \{3.4\text{x}10^{-7} - 3.4\text{x}10^{-4}\}$ as shown in in Table 3.1.

Figure 3.1, described how simulation data were generated. For models 1 and 3, I generated nine parameter combinations of u and s, changing u by orders of magnitude over the range u: $\{3.4\text{x}10^{-7}, 3.4\text{x}10^{-6}, 3.4\text{x}10^{-5}\}$, and changing s by an order of magnitude as well, over the range $s = \{0.001, 0.01 \ \& \ 0.1\}$. For model 2 (stickbreaking), I generated twelve parameter combinations, as this model can be treated as the gold standard, introducing a higher mutation rate parameter $u = 3.4\text{x}10^{-4}$. The neutral evolution model was run with same range of s but changing the number of drivers k at the start from $k = \{1, \ 2, 3 \text{ and } 4\}$ expanding the population from a single starting cell to 100 million cells.

I generated 500 replicates for every parameter combination, each simulation started with one cell and was stopped when the tumour reached a size of ≥4 billion cells.

A reduced sample was generated from 3 parameter combinations of the increased mutation rate model where reaching 4 billion cells size was too computationally and memory intensive to

simulate. The affected parameter combinations are the following, $s = 0.01$ with $u = 3.4\text{x}10^{-5}$ with 250 replicates; $s = 0.001$ with $u = 3.4\text{x}10^{-5}$ with 68 replicates and $s = 0.001$ with $u = 3.4\text{x}10^{-6}$ with 412 replicates.

During the evolution of each simulated tumour, twenty records are made of the clonal structure and key features of the tumour at selected milestone sizes ($5\text{x}10^2$, $10\text{x}10^2$, $5\text{x}10^3$, $10\text{x}10^3$, $5\text{x}10^4$, $10\text{x}10^4$, $5\text{x}10^5$, $10\text{x}10^5$, $5\text{x}10^6$, $10\text{x}10^6$, $5\text{x}10^7$, $10\text{x}10^7$, $5\text{x}10^8$, $1\text{x}10^9$, $1.5\text{x}10^9$, $2\text{x}10^9$,$2.5\text{x}10^9$,$3\text{x}10^9$,$3.5\text{x}10^9$,$4\text{x}10^9$).

Every snapshot stores data on the top 100 clonal linages by size and the following features,

1) Metrics of heterogeneity (as described in the appendix): Number of (resistant) clones and their sizes, the Rich-Gini Simpson index of diversity (RGS), Shannon entropy, Shannon equity, and what fraction of the tumour was comprised of the top-100 clones.
2) Metrics of tumour dynamics: Total number of driver mutations, total number drug-resistant mutations, driver tumour composition, tumour size, number of generations elapsed, average proliferation, tumour fitness and the average survival probability of the tumour.
3) Clonal metrics: Lineage relationship, clonal emergence time, number of drivers, fitness, expected number of passengers, drug-resistance status, and survival probability.
4) Metrics for sequencing comparison: Cancer cell fractions relative to the tumour (all clones measured) and cancer cell fractions relative to the top 100 clones.

Similar outcomes of the clonal branching process were reported by Chowell et al. [44]. For instance, they showed tumours contain an abundance of low-frequency clones that is not detectable by current sequencing assays. They also found that drug-resistance is more likely to occur in tumours with low fitness. In this chapter I will show results that replicate the Chowell et al. study and expand upon their work.

6 Milestone Tumour Sizes for Evaluation, 1 cm^3 & 4 cm^3

All simulations generated in the positive selection models start with one cell and increase to 4 billion cells, at which point they are stopped. Two key milestone tumour sizes, 1 cm^3 (1 billion cells) and 4 cm^3 (4 billion cells), were selected for analysis. A justification of the tumour sizes was provided in Chapter II section 10. To complement this diagnosable range, I also provide the complete growth pattern of the features of interest to assess their consistency over time.

All simulations begin with a cell having a preselected set fitness advantage proportional to *s*. Tumour initiation is not modelled. Rather each simulation begins at a point where the founder cell has acquired a mutation that provides a fitness advantage. This could represent a primary tumour or a metastatic growth.

For instance, a metastasis that has accumulated a certain number of driver mutations is expected to have an increased fitness relative to the primary and thus can be simulated by starting with a higher average selective advantage *s*. Therefore, the initial fitness and the value of *k* is relative to the genomic background and context of interest. A clone with starting fitness of $s = 0.001$ that has accumulated 10 drivers is equivalent to a clone with $s = 0.01$ with only one driver. Accounting for this provides great flexibility to study all stages of disease progression in a single self-contained model.

7 Exploring the Properties of The Additive Fitness Model

Bozic et al. [43] originally reported using the discrete-time branching process as a tool to recreate tumour growth to infer clonal evolution and describe driver mutation abundance. In this section, I will expand on their model by showing the properties the clonal additive fitness model, focusing on the variables that describe expansion, diversity and tumour composition and their potential use in the clinic.

7.1 Number of Generations and Fitness as Predictors of Average Selective Advantage *s*

Tumour size and markers of proliferation (such as Ki-67) are common features measured in the clinic to estimate rates of tumour growth. The main goal here is to evaluate which of the simulation variables that correlate with the average selective advantage *s* and average driver mutation *u* may also have clinical relevance.

As shown in Chapter II Figure 2.10, the input parameters *s* and *u* are associated with the number of generations required to reach a certain tumour size as they have an impact of net proliferation rate $\lambda_{k,i,j} = b_{k,i,j} - d_{k,i,j}$. Figure 3.2.a supports the analytical solutions showing clear distinctive timeframes of tumour development among input parameters *s* and u.

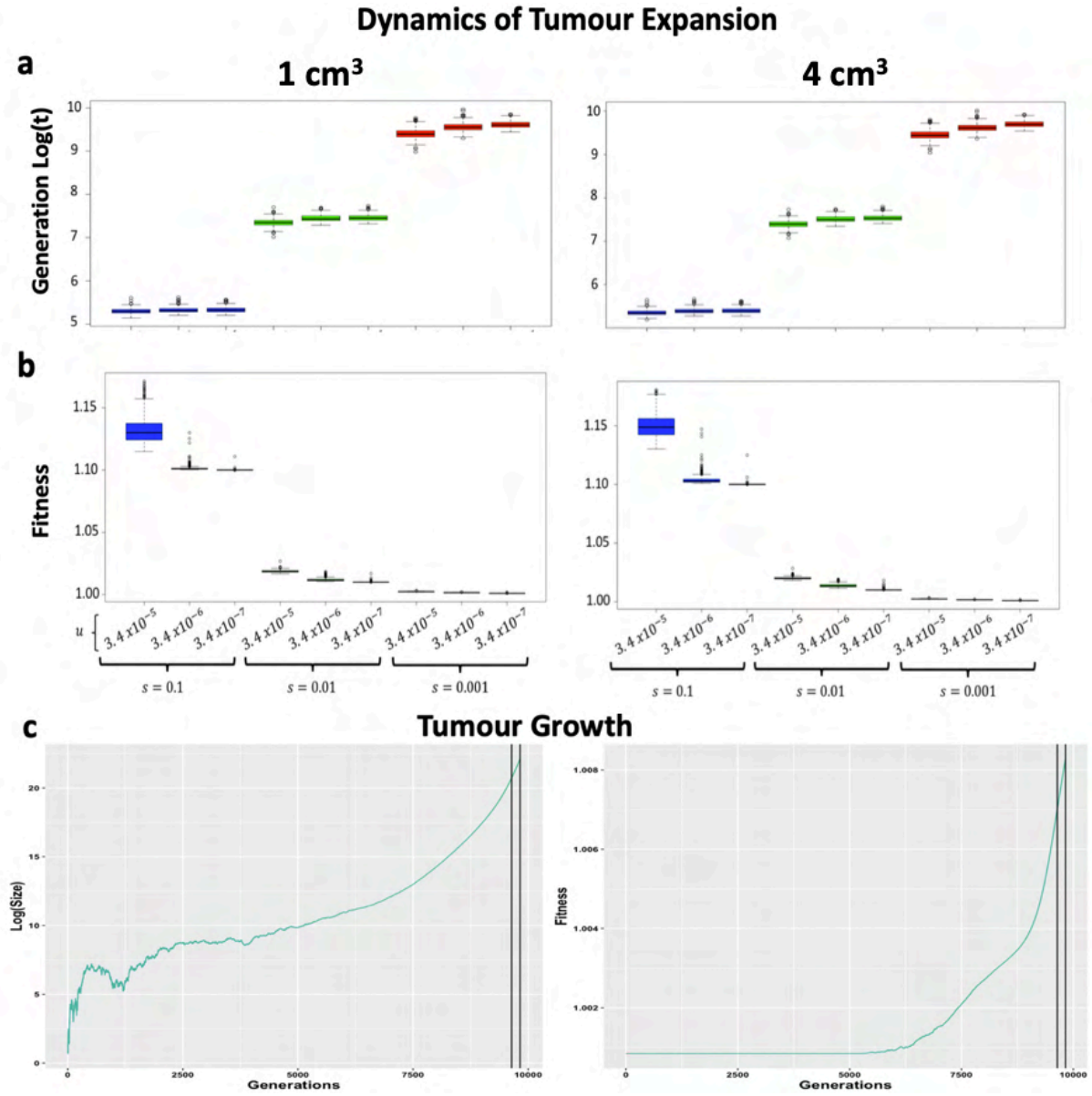


Figure 3.2 Dynamics of tumour expansion in the additive fitness model. Boxplots show measured features recorded at the indicated milestone sizes across all simulations. **a,** distribution of the number of generations required to reach the milestone tumour size. **b**, fitness distributions of all simulations, where fitness is calculated as the average fitness across all cells. **c,** example simulation with $s = 0.001$ & $u = 3.4\text{x}10^{-5}$. Vertical black lines indicate milestone sizes 1 cm^3 and 4 cm^3. Tumour growth, is plotted on the left, and the increase of fitness over time on the right.

Total simulation time is measured in generations, as reported in Figure 3.2.a. In order to apply this information to a particular cancer subtype, it is necessary to rescale in years or months. To transform the time scale, one must estimate the average division rate of the tumour of interest and adjust the generations accordingly. For instance, Bozic et al. [43] used division rates of 3 – 4 days which allowed them to compare against observed rates of neoplastic formation.

Figure 3.2 shows that the number of generations (or tumour age) and tumour fitness cluster according to the value of the average selective advantage s. In Figure 3.2.c, it can be seen that the values of both variables at 1 cm^3 and 4 cm^3 do not change significantly over time.

As a result, this finding suggests that both metrics, the number of generations and fitness, are indicative of the average selective advantage *s* and can have classification power to aid in the determination of *s* and *u* in the additive fitness model.

7.2 Simulated Metrics of Detectable Diversity Cannot Predict Average Driver Mutation Rate *u*

Tumour heterogeneity is a widely accepted indicator of therapeutic failure and disease progression. However, despite its relevance, efforts to determine its origins and its role in dissemination and resistance remain incomplete. This is in part because studying tumours in early stages in human models across patients is difficult, if not impossible, with our current detection techniques.

With the objective of providing a better understanding the dynamics of tumour heterogeneity, I evaluated metrics of intratumoral heterogeneity in my models with respect of clonal diversity.

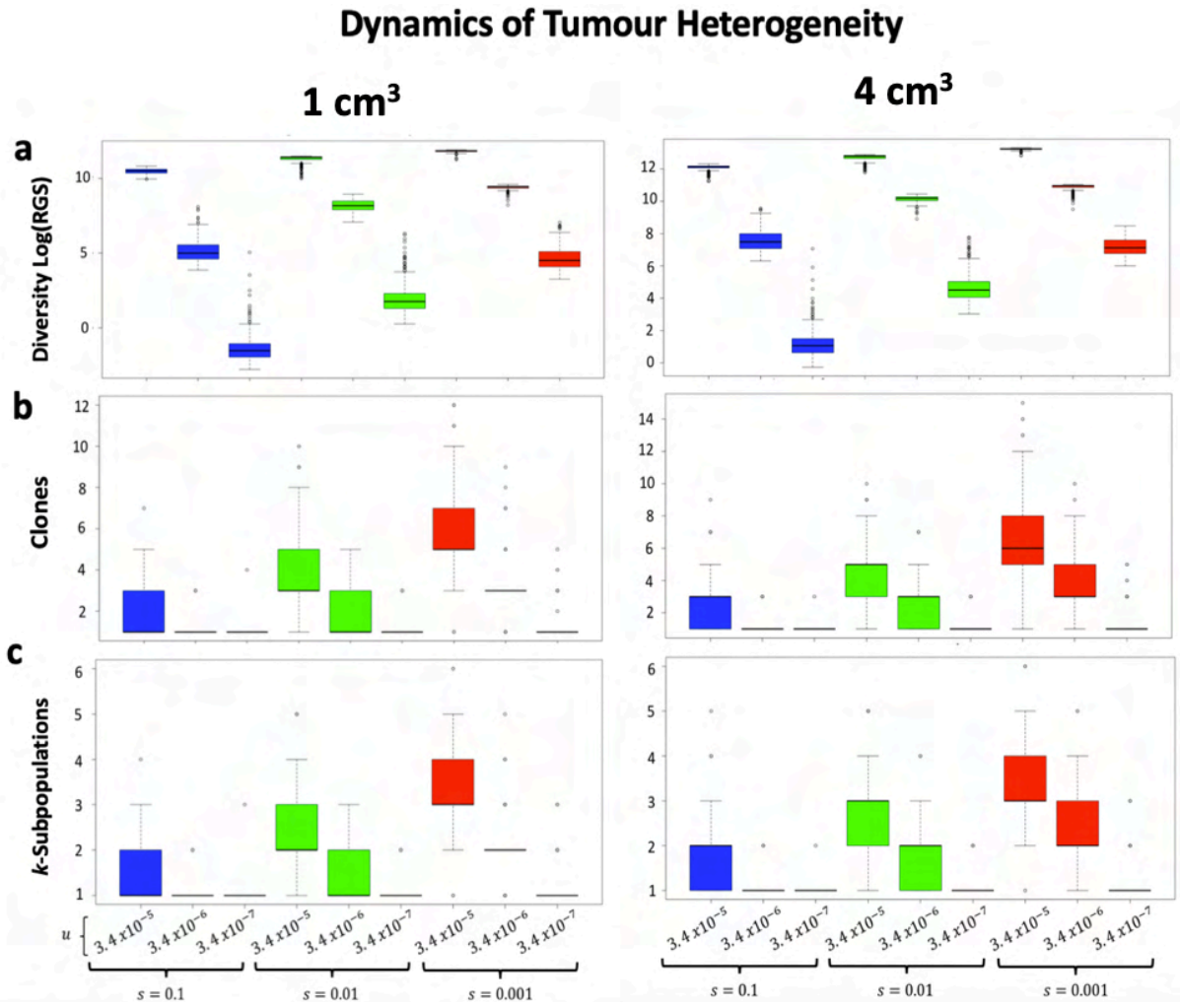


Figure 3.3 Boxplots of dynamics of tumour heterogeneity in the additive fitness model at 1 cm³ and 4 cm³. Boxplots are taken at milestone sizes in all the simulations generated. **a,** distribution of the Rich-Gini Simpson index of diversity when all clones are measured. **b**, distribution of the number of detectable clones using a 10% CCF cut-off. **c**, distribution of the number of detectable driver subpopulations using a 10% CCF cut-off.

The number of clones, RGS (Fig 3.3.a), Shannon entropy and Shannon equity all correlated with and clustered with the average driver mutation rate u when information for all clonal subpopulations is used. However, in real life it is impossible to measure all clonal frequencies within a tumour. Therefore, to replicate real life scenarios I restricted the analysis to only the heterogeneity detectable at CCF frequency cut off or 10% or greater, which is comparable to typical limits on sequencing resolution in many studies (Fig 3.3.b and Fig 3.4).

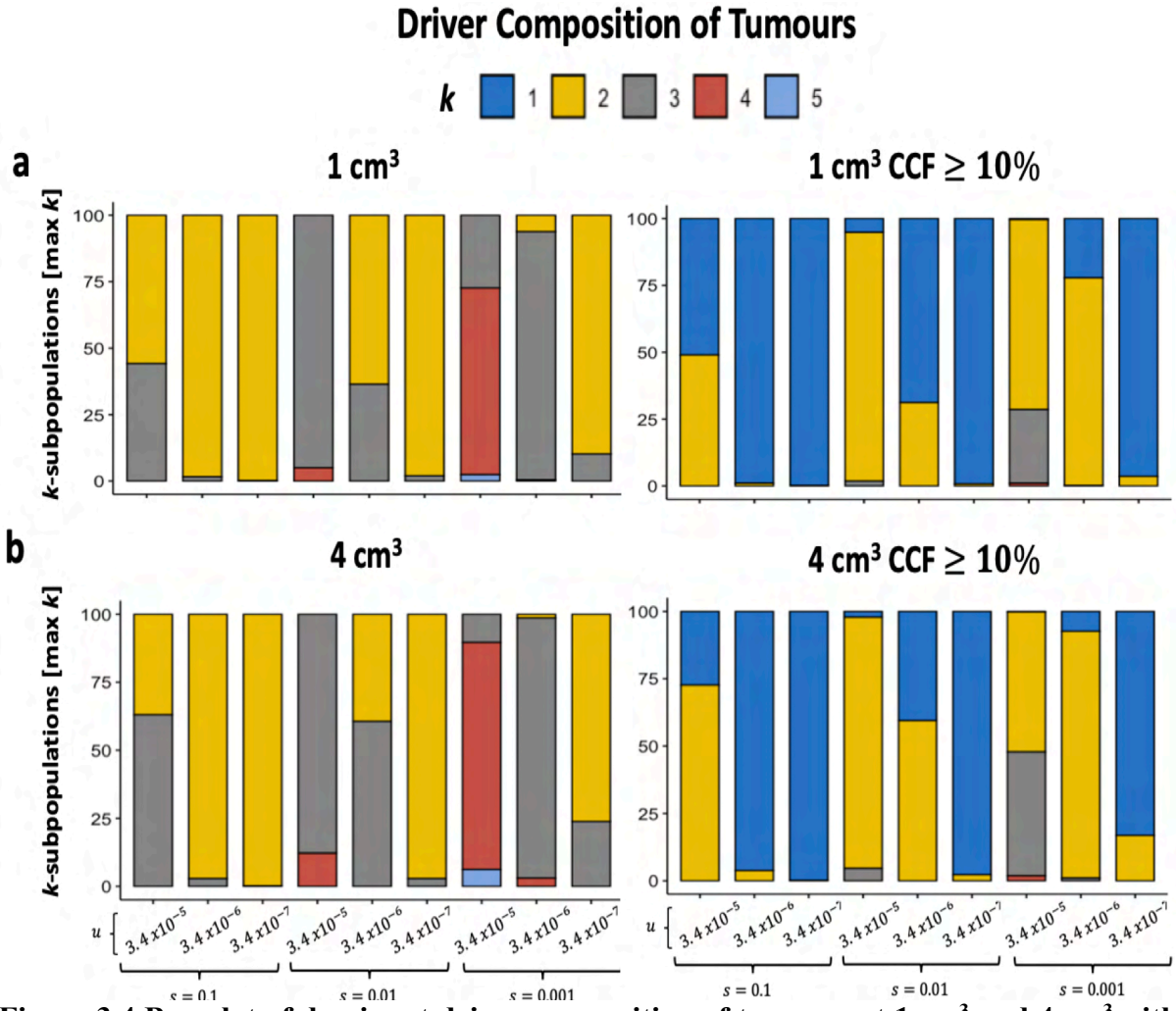


Figure 3.4 Bar plot of dominant driver composition of tumours at 1 cm^3 and 4 cm^3 with and without applying a 10% cancer cell fraction cut-off. Bar plots are taken at milestone sizes in all the simulations generated with and without the standard sequencing cut-off. **a,** comparison at 1 cm^3, **b**, comparison at 4 cm^3. All samples take the dominant k of the simulation.

As can be seen by comparing panels a and b in Figure 3.4, the difference in driver abundance detection is by one unit, which means that most of the driver makeup of the tumours would be detectable by standard sequencing approaches (in conditions subject to model assumptions).

When all clonality is measured, the diversity metrics such as RGS, Shannon entropy and Shannon equity cluster with the changes of the average driver mutation u as illustrated in Figure 3.3.a. This suggests that intratumoral heterogeneity provides a good approximation of the average driver mutation rate u, though this effect is only apparent when a significant amount of heterogeneity beyond the reach of current technologies is included. Even RGS diversity in the top-100 clones showed considerable overlap between parameter values.

As a result, clustering of parameters with average driver mutation rate u is lost when restricting to measurable clonality, as exemplified in Figure 3.3.b where there is significant overlap the number of clones detected at different values of u.

Changes in tumour size do not significantly change the inferred number of clones or the diver composition as seen in Figures 3.3.b, 3.3.c and the left-hand side of Figure 3.4. The effect of increasing the average driver mutation rate manifests in simulation results with higher clonality and more diversity in their driver composition.

As shown in Chapter II, the driver compositions predicted by the analytical solutions are in concordance with the stochastic simulations of the additive fitness model and provide validation for the analytical solutions derived in the previous chapter (supplementary tables 3.1 and 3.2).

In summary, the diversity metrics RGS, Shannon entropy and Shannon equity do not display classification power for determining average mutation rate u in the additive fitness model.

7.3 Limitations in Clonal Composition Determination Using Sequencing Technologies

Clonality tools require an unbiased estimate of the frequency of the mutation in the sample or tumour (CCF), i.e. ploidy correcting the VAFs. VAF measures the fraction of chromosomes that carry a given mutation, whereas CCF measures the faction of cells in the sample or tumour that carry said mutation. As shown in eq. 1.1 due to inheritance the CCF in bulk sequencing is the measurement the lineage frequency (parental and progeny). Certain phylogenetic tools such as PhyloWGS resolve the mixture of parental and progeny to build phylogenetic topologies. Other commonly used tools require subsequent tree construction (e.g. PyClone and ExPANdS). One of the advantages of the model is the capability to decompose the CCFs into clonal frequencies to indicate the prevalence of genotypes in the sample, aiding phylogenetic tree reconstruction.

With the simulation outcomes, an intuitive question arises, in which scenarios the CCFs are not required to resolve the clonal mixture (eq. 1.1) and the CCF is indicative of the clonal size. To investigate the role of tumour composition and establish its association with the average driver mutation rate u, I evaluated the relationship between cancer cell fractions and clonal frequency.

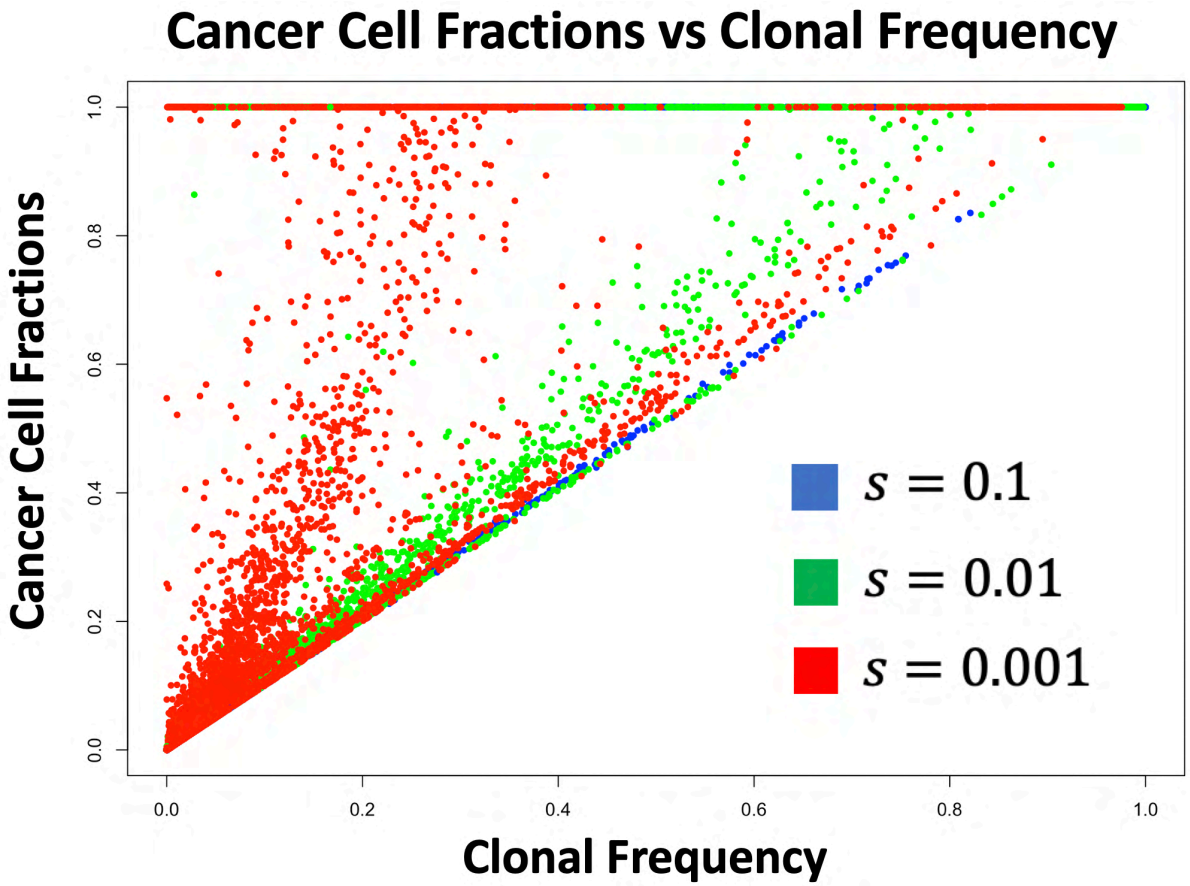


Figure 3.5 Cancer cell fraction vs clonal frequency in the top 100 clones. Scatterplot shows how clonal frequencies translate to cancer cell fraction in all tumours simulated with the additive fitness model at a 4 cm^3 tumour size. Every dot represents a clone of the top-100 per parameter combination *s* and *u*.

Figure 3.5 shows that tumours with strong selective advantage $s = 0.1$ do not require resolution of the inheritance proportion of eq. 1.1 (measurement is only parental). In this scenario, the ploidy adjusted frequency represents clonal size (or dominance, in this scenario CCFs are likely to represent parental clones and their CCF is equivalent to their genotype).

For tumours with moderate and weak selective advantage s:$\{0.01,0.1\}$, the mixture of proportions of eq. 1.1 has to be resolved to establish the genotypic or clonal prevalence within the sample. For instance, assuming $s = 0.001$, after bulk sequencing a mutation X was found to have a $CCF = 1$, in this scenario is likely that founder clone where mutation X arose is outcompeted by its progeny and may not even be present in the sample. This can be seen in the way the maximum cancer cell fraction (1.0) spans multiple clonal frequencies in Figure 3.5. Moreover, there is a high degree of overlap between simulations with different average selective advantages s which illustrates the non-linearity of cancer cell fractions as a function of clonal frequency.

Next, I evaluated the number of cells that accumulated *k*-driver alterations and what proportion of the total tumour size each take up respectively at the default detectable threshold 10% CCFs cut-off. The 10% (or 0.1) CCF threshold for whole exome sequencing (~100x) to avoid false positives from the sequencer or sample prep-noise (e.g. FFPE) [178-180]. Although, this cut-off is the suggested standard, it does not account for the degree of copy number change that

my affect clonal lineages at borderline thresholds limiting even more an accurate reconstruction of clonal evolution solely with clonality methods.

It is worth mentioning these are different from the clonal proportions reported in Chapter II, as those refer to the number of total cells and not to the measurable proportions. Instead of focusing on the total number of cells, I focus on the inference on clonal composition based on the default detectable sequencing threshold of 10% CCF to replicate the challenges faced in determining clonal architecture in a real-world scenario.

To better explain the problem, Figure 3.6 compares the actual driver composition for the top-100 clones versus the measurable tumour composition ≥10% CCF cut-off (assuming a representative sample). In this example, the 500 simulations with parameter combination $s = 0.001$ and $u = 3.4\text{x}10^{-5}$ at 4 cm^3 size was used. In this snapshot, the top 100 clones describe 82% of the total tumour composition and the remaining composition is distributed in 593,655 clones with reduced number of cells.

Figure 3.6.a describes how the number of clones with *k* drivers in the top-100 compares to frequencies of *k*-driver subpopulations at 10% CCF frequency detection. It can be seen within top-100 clones that two lineages with 2 driver alterations spawned, and 87 3-driver clones, but due to inheritance inflation (eq. 1.1) only two clones are detectable in rep-seq sequencing, the founder and a 2-driver clone.

In Figure 3.6.b when the cancer cell fraction cut-off was set to 5%, requiring considerable sequencing depth and careful clonality calling. Two previously undetectable 3-driver clones are now evident at the 5% threshold, which will provide a more complete picture of the tumour composition. This illustrates how the problem of inheritance complicates efforts to recover clonal ancestry given the limits of current sequencing technologies, variant calling and clonality calling methods.

Figure 3.6.c further illustrates this point, showing that the 3-driver clones comprise the majority of the tumour but each are so small that only the 2-driver clones can be detected when the CCF cut-off is applied.

In summary, Figure 3.6.c shows the cellular driver load of the tumour which is dominated by $k = 2$ and $k = 3$, but in the process of (rep-)sequencing only few clones can be detected by restricting detection to $\geq 10\%$ CFF (equivalent to ~50% VAF in diploid tumours) due to inheritance. Figure 3.6.b shows that all clones contain mutations labelled as *1* and *1,22* leading to a 100% CCF detection by inheritance even when founder clone $k = 1$ represents a minimal fraction of the tumour (blue Fig. 3.6.c). Bozic et al. [94] showed that this effect is prominent and leads to bimodal driver variant frequency distributions. This highlights the relevance of having clinical markers to improve alignment with the branching process and the methodological challenges in sequencing studies to recover tumour evolution.

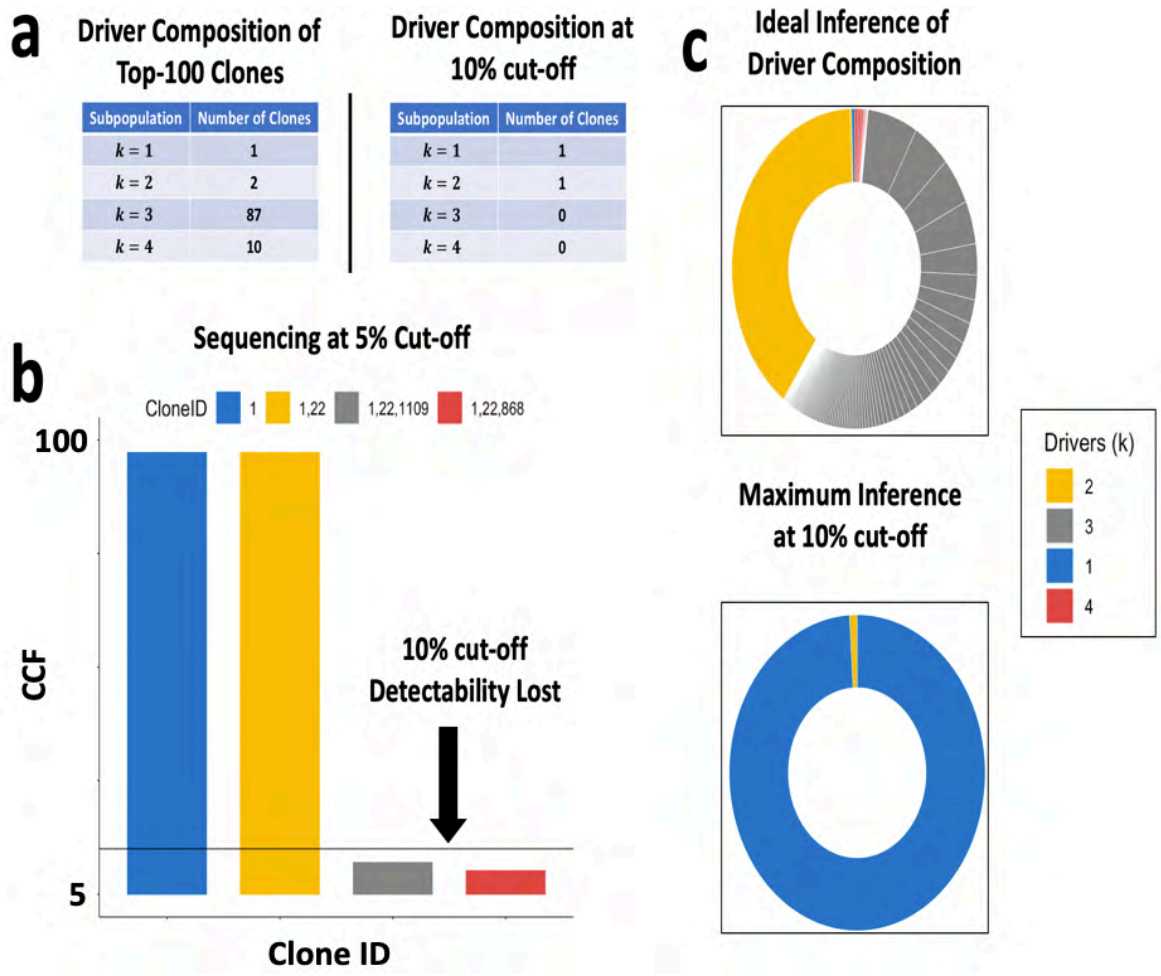


Figure 3.6 Limitations of sequencing technologies to recover clonal composition. a, table describing the number of accumulated drivers and the number of clones among the top-100 clones vs those detectable at with a 10% CCF cut-off. **b**, the cancer cell fractions of the sample at a 5% CCF cut-off colour coded by clone ID. Numbers refer to the parental lineage, with the last number referring to the clone ID. **c**, pie charts of the driver composition of the top-100 clones vs 10% cut-off colour coded by *k*. Separations in the chart indicate the cancer cell fraction proportion.

Figure 3.5 shows how cancer the cell fraction overlap between parameters introduces further challenges for reconstructing tumour evolution using evolutionary models. Figure 3.6 shows how reconstructing tumour evolution is affected by inheritance, sequencing depth, variant and clonality calling methods. For instance, the phylogeny reconstruction that can be made by Figure 3.6.b at 10% CCF frequency cut-off is single branched (1→22). In contrast, at a 5% CCF cut-off the reconstruction can be single branched as (1→22→109→868) or divergent (1→22→[109,868]).

By accounting for these two challenges, Chapter IV will discuss how to compare simulated vs real cancer cell fractions to reconstruct tumour evolution.

7.4 Distribution of Measurable Driver Composition in the Additive Fitness Model

Section 7.3 showed the limitations in sequencing using one sample as an example, in this section I use all simulations to evaluate driver composition and its variance, and to assess the degree of overlap between parameter combinations.

I used all the simulations in the additive fitness model, filtered based on the default cut-off of 10% CCF and evaluated the proportion of cells containing *k* driver alterations in the remaining clones. The next figure shows the distributions per parameter combination of the expected driver compositions of the tumour.

Figure 3.7 shows clonal lineages that have accumulated 3-driver alterations can only occur when the average selective advantage is weak, $s = 0.001$ and the average driver mutation rate is high, $3.4\text{x}10^{-5}$. In contrast, 2-driver clonal lineages, marked by an asterisk (*) in the figure are likely to happen with strong selection, $s = 0.1$, or at moderate to lower average driver mutation rates, $3.4\text{x}10^{-6}$ & $3.4\text{x}10^{-7}$. In the context of clonal competition, it can be seen that only when the selective advantage is moderate or weak ($s = \{0.001 \text{ \& } 0.01\}$) 2-driver clones can overtake their predecessors and become the dominant measurable subpopulation.

As observed in Figure 3.4, the driver composition remains practically the same at milestone sizes with and without a CCF cut-off. With Figure 3.7 showing within the variation in detectable driver compositions the 2-driver composition is the most likely to be observed. Again, the risk of applying the 10% CCF cut-off is having underrepresentation of clonal composition as shown in Figure 3.6 when a representative sample is obtained.

The following section will explore the degree of overlap of the cancer cell fractions at different CCF cut-offs between simulation parameters and their implications for estimating the initial average selective advantage s and average derive mutation rate u.

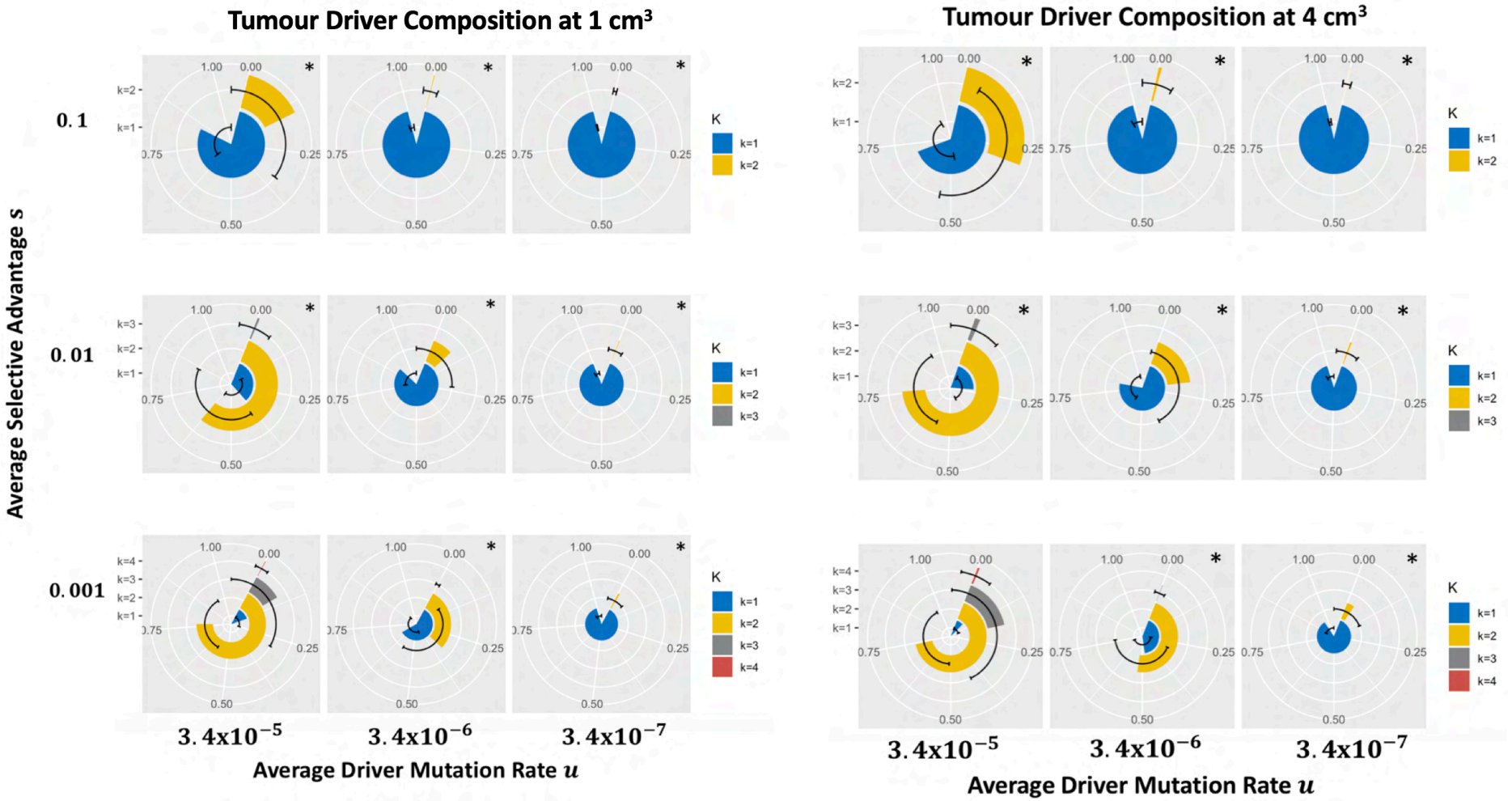


Figure 3.7 The measurable *k*-driver composition of tumours at milestone sizes in the additive fitness model. Pie charts with error bars indicating the *k*-driver tumour composition colour coded by driver subpopulation *k*. Numbers represent the proportion of the tumour. The pie charts marked with an asterisk (*) show that most 1-2 driver subpopulations can be detected.

7.5 Evaluating the Degree of Overlap in Measurable Cancer Cell Fractions

The observation that cancer cell fractions overlap across parameters can make it problematic to determine tumour fitness and mutation rate using the branching process. Assessing the degree of uniqueness between simulation outcomes is the motivation for this section, in other words, what is the degree of similarity on CCF distributions within and across initial parameters s and u.

To investigate the extent of the overlap in cancer cell fractions, I compared all cancer cell fractions against each other using a similarity score in the additive fitness model at 4 cm^3 (a total of the 4,500 simulations) with a degree tolerance ε of 5% to account for sequencing noise. Details of the similarity score and its implementation are described in Appendix 3.2. The following heatmap shows the degree of similarity scoring all vs all cancer cell fractions:

It can be seen that there is a significant overlap among parameters, as indicated by the red gradient intensity in Figure 3.8 at both the 1% and 10% CCF cut-offs. The lowest average driver mutation rate $u = 3.4\text{x}10^{-7}$ had the most overlap due to the low clonality of those simulations.

However, certain sections of the heatmap show the opposite pattern, as highlighted in Figure 3.8. These represent parameter combinations with reduced overlap in CCFs. Here, the additional number of detectable clones provides better separation between simulations with strong versus moderate/weak selective advantage.

The parameters in the likelihood reported by Bozic et al.[43] ($s \sim 0.001$ and $u = 3.4\text{x}10^{-5}$) as being relevant for most cancer subtypes, show CCF overlap with results from all s values when the average driver mutation rates are $3.4\text{x}10^{-6}$ and $3.4\text{x}10^{-5}$. This illustrates the importance of exploring a range of average driver mutation rates.

To validate this observation, I applied the same metric to CCF data generated from real tumours for the pan-cancer study of estimated clonality in TCGA conducted by Andor et al. [29], who used ExPANdS as their main clonality caller tool, Figure 3.9.

As expected, there is similarity in cancer cell fractions between different types of malignancies in TCGA. Two clusters highlighted with red intensity and flagged with arrows in Figure 3.8.a stand out, one containing bladder cancer (BLCA), lung adenocarcinoma (LUAD), lung squamous cell adenocarcinoma (LUSC), skin cutaneous cell melanoma (SKCM) and stomach adenocarcinoma (STAD) subtypes and the other thyroid (THCA) and prostate adenocarcinoma (PRAD) and low-grade glioma (LGG). Interestingly subtypes with each of these clusters also have similar clonality distributions as depicted in Figure 3.9.b.

The degree of overlap between parameters at the 10% CCF cut-off introduces a difficult challenge for the accurate estimation of tumour fitness and average driver selective advantage. However, the database generated here can be used as an approximation method to evaluate which parameter combinations are more likely to explain cancer cell fractions obtained through sequencing, accounting for constraints on coverage and depth.

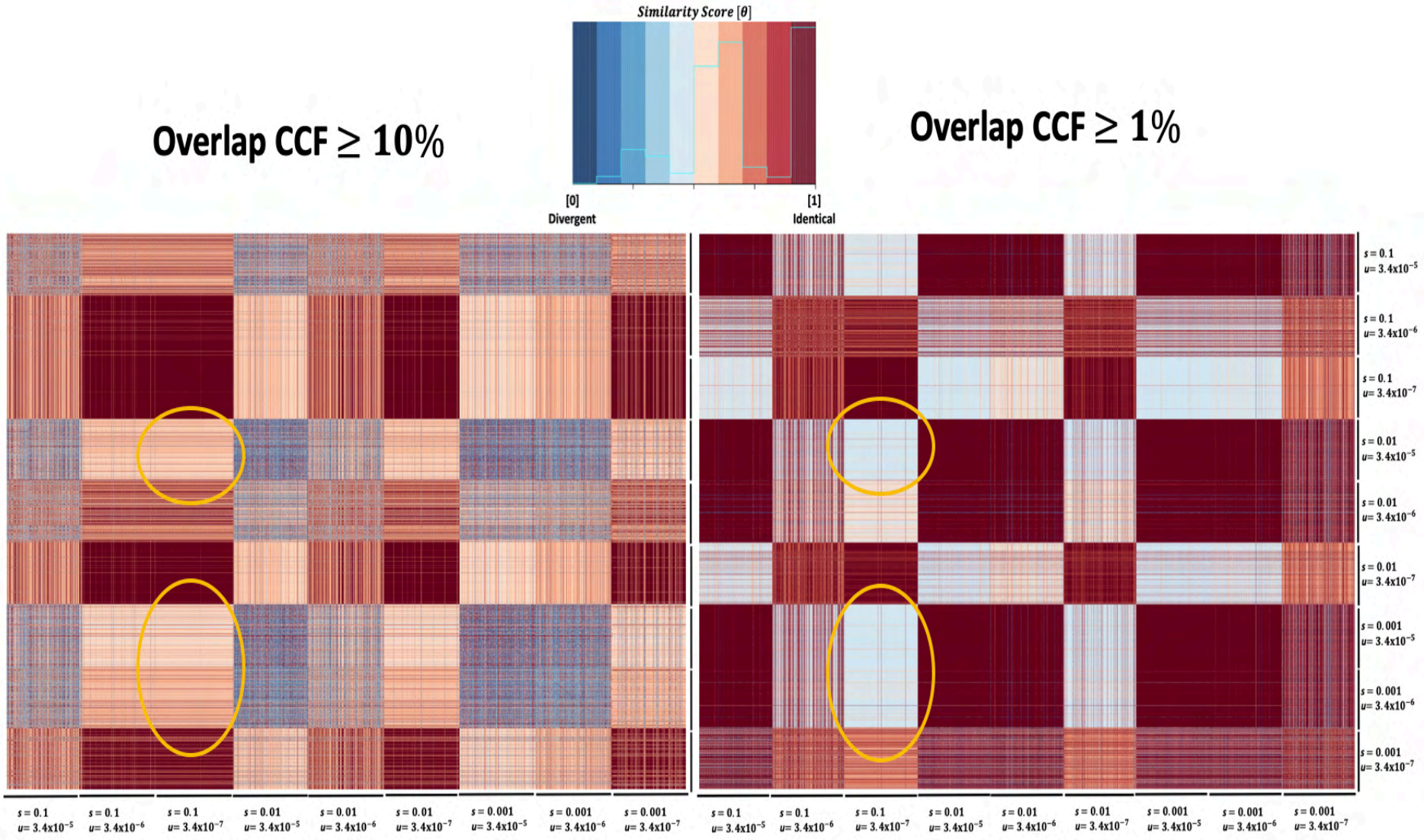


Figure 3.8 Heatmap of similarity in detectable cancer cell fractions at 10% and 1% CCF cut-off. Colour gradient represents the degree of similarity with blue being the lowest and red the highest. Yellow circles highlight areas of low overlap.

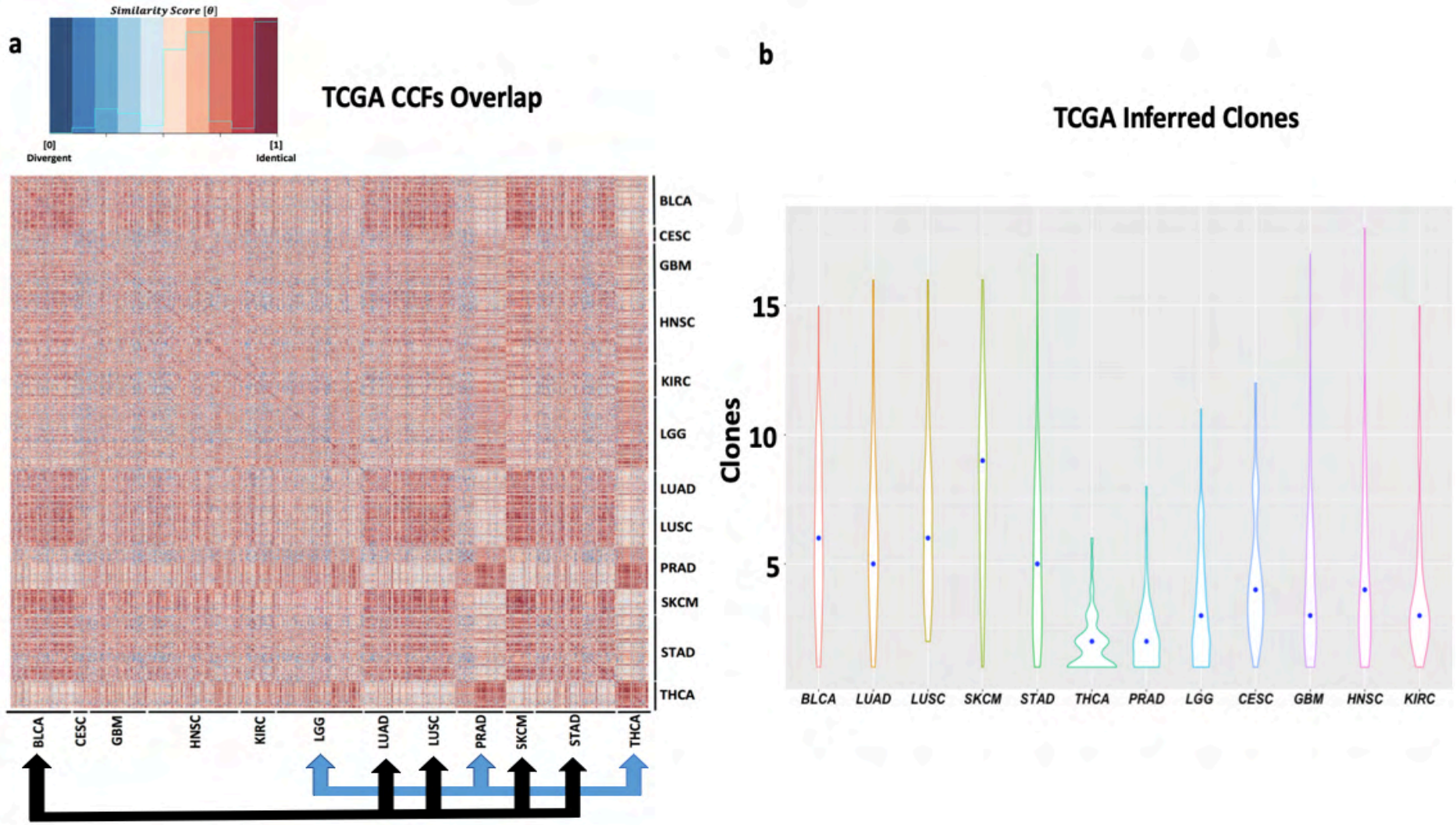


Figure 3.9 Heatmap of similarity in detectable cancer cell fractions and inferred clonality in TCGA. a, colour gradient represents the degree of similarity with blue being the lowest and red the highest. Arrows indicate subtypes with high similarity. **b,** number of inferred clones with a precision >0.7 identified by ExPANdS. Blue dots represent median values.

As a result, the additive fitness model has considerable degree of overlap between parameters that can cause difficulty in precisely determining a given tumour's fitness and driver mutation rate, particularly at the commonly used sequencing CCF cut-off. This effect seems to be also present in human malignancies highlighting the need to integrate clinical and molecular markers.

The following sections describe the differences between the stickbreaking and increased mutation rate models relative to the additive fitness model.

8 Comparing the Additive Fitness to the Stickbreaking Fitness Model

I next evaluated the metrics of tumour dynamics using the stickbreaking fitness model. In contrast to the additive fitness model, new driver mutations in the stickbreaking model are sampled from a predefined distribution of fitness effects. Said distribution is bounded by the parental fitness and a hypothetical ceiling assumed to be 1, resulting in a richer fitness landscape [138, 142].

The goal was to evaluate how applying a stickbreaking fitness model may alter the trends shown in the previous section from the additive fitness model. I repeated those same analyses aiming to determine which variables have classification power and associate with the average selective advantage s and average driver mutation rate u.

In contrast to the results from the additive fitness model (Figure 3.2), the parameters describing tumour expansion overlap as shown in Figure 3.10. The overlap of parameters is the key additional feature of the stickbreaking model as it allows more variation in development times and tumours fitness. The downside is reduced ability to get classification power from the number of generations and predicted fitness.

While the number of generations in both models are similar, the stickbreaking model has a greater range in fitness distributions and variance. Moreover, the stickbreaking model shows an increase in overall fitness at the 4 cm^3 size relative to the additive fitness model. Despite different fitness ranges in both modes, tumours grow in a similar timeframe in both models.

The dynamics of tumour heterogeneity of the stickbreaking model have the same pattern as the additive fitness model (Figure S3.1.a) when using the complete metrics of tumour diversity such as RGS. Similarly, metrics of heterogeneity at the default frequency cut-off 10% of CCF have reduced classification power due to the degree of overlap (Figure S3.1.a and Figure S3.1.b). The number of k-driver subpopulations are similar at milestone sizes (Figure S3.1.c) and the number of k-driver subpopulations does not change significantly with and without CCF cut-off, as observed in the additive fitness model shown in Figure 3.3.

As a result, the number of generations and fitness have reduced classification power to infer average selective advantage s and average driver mutation rate u as compared to the additive fitness model. However, the stickbreaking model can provide more diverse patterns of clonal evolution by creating greater variation in cumulative fitness as the number of accumulate driver mutations grows.

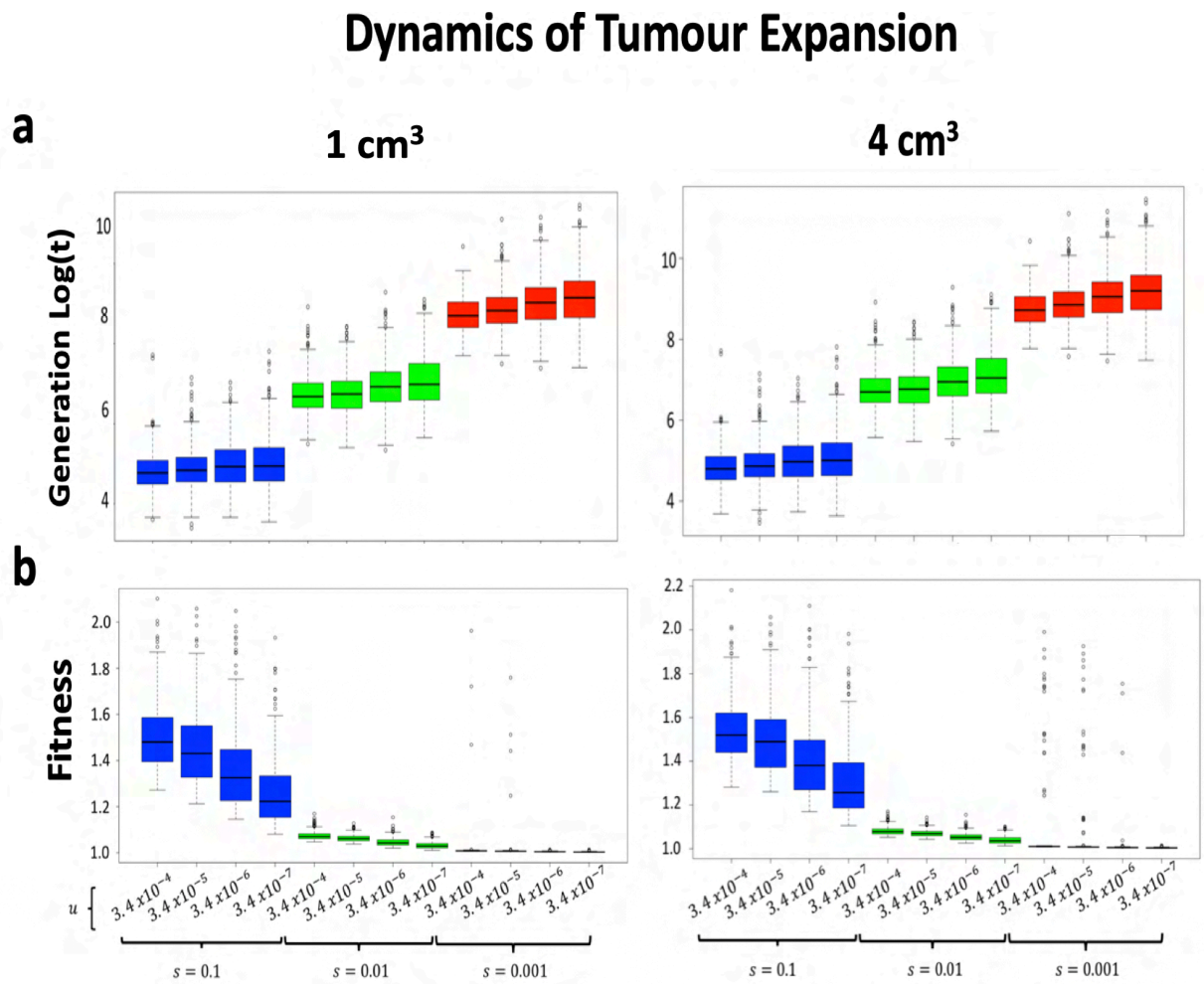


Figure 3.10 Dynamics of tumour expansion in the stickbreaking fitness model. Boxplots show measured features recorded at the indicated milestone sizes in all the simulations generated. **a,** distribution of the number of generations required to reach the milestone tumour size. **b**, fitness distributions of all simulations where fitness is calculated as the average of all cells.

8.1 Distribution of Measurable Driver Composition in the Stickbreaking Model

The differences in fitness distributions between the additive fitness model and the stickbreaking model (as described in Section 8 and shown in Fig. S3.1) motivate exploration of the differences in driver composition between models, as it will have an impact in defining approaches to compare simulation results to estimates from sequencing data to recover clonal ancestry.

Similar to the additive fitness model, significant overlap between parameter combinations is also observed in the stickbreaking model, with 2-driver subpopulations more likely to occur with strong selection $s = 0.1$ and moderate to low average driver mutation rate $u = \{3.4\text{x}10^{-7}\ \&\ 3.4\text{x}10^{-6}\}$. In addition, 3-driver subpopulations are likely to occur when the average fitness is moderate or low, $s = \{0.001\ \&\ 0.01\}$ and the average mutation rate is high $u = \{3.4\text{x}10^{-5}\ \&\ 3.4\text{x}10^{-4}\}$.

Overall, fitness increases in the stickbreaking model manifest in a greater *k*-driver mutational burden relative to the additive fitness model, caused by variation in the average selective

advantage *s*. Although the stickbreaking model does not have strong classification power, its benefit comes in describing clonal trajectories that otherwise would not be captured by the additive fitness model.

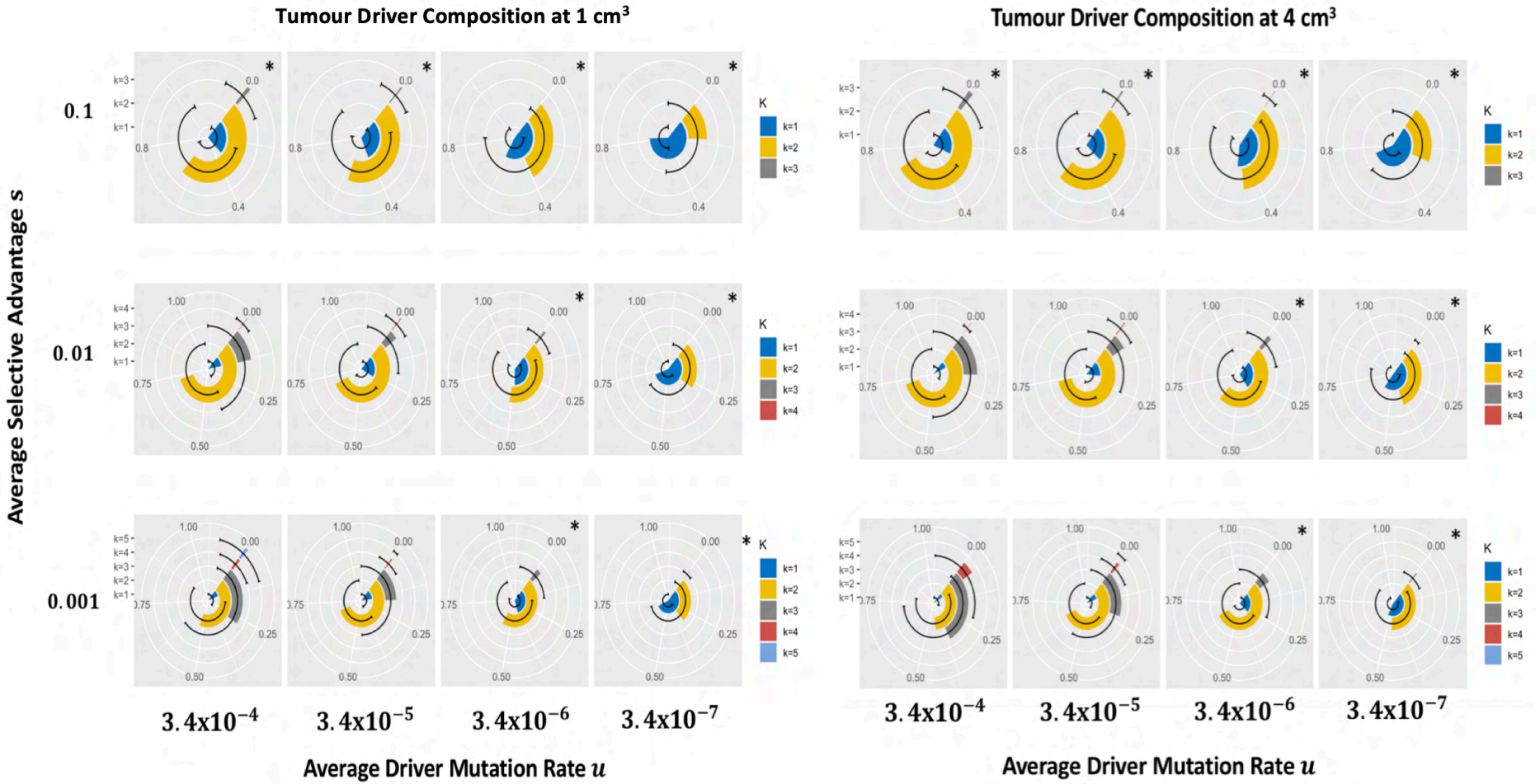


Figure 3.11 The measurable *k*-driver composition of tumours at milestone sizes in the stickbreaking model. Pie charts with error bars indicating the *k*-driver tumour composition colour coded by driver subpopulation *k*. Numbers represent the proportion of the tumour. The pie charts marked with an asterisk (*) show that most 1-2 driver subpopulations can be detected.

9 Evaluating the Increased Mutation Model

Sections 7 and 8 explored the properties of the additive fitness and stickbreaking models to examine how differences in fitness effects may change the course of tumour evolution. These models keep the average driver alteration rate *u* unchanged. To evaluate the impact of changing the average driver mutation rate *u*, the increased mutation model was designed.

The increased mutation rate model is similar to the additive fitness in that changes in the average selection advantage occur proportional to *k,* as shown in Figures S3.2 and S3.3, with similar distributions observed in the additive fitness model.

The increased mutation rate model increases the average driver mutation rate u in 2-driver lineages according to the formula $u = u + 0.5u$, to mimic the effects of hypermutators or the loss of tumour suppressors that maintain genome integrity such as TP53.

The same exploratory analyses of the metrics of expansion and diversity described in the previous two sections were repeated on the results from the increased mutation rate model.

The effects of the increased mutation rate model are manifested in the metrics of intratumor heterogeneity, with a greater RGS diversity than the additive fitness and stickbreaking models, Figure S3.a. The number of clones is similar to the additive fitness model when tumour size is 1 cm^3 but differs when size reaches 4 cm^3. With respect to the stickbreaking model, the number of clones is smaller due to the increase of diversity, Figure S3.b. Similar to the previous models the number of *k*-subpopulations does not change significantly by tumour size nor by detection threshold.

As shown in Figure S3.4, the increased mutation model shows overlaps in the distributions of tumour driver composition with both the additive fitness and stickbreaking models. Similarities in driver composition with the additive fitness model are expected because cumulative fitness increases are done in the same way in both models. However, said similarity is in the strong to moderate outcomes, $s = \{0.1, 0.01\}$, because the expansion rates are faster and this does not allow emerging mutants to fully sweep to fixation.

Similarities with the stickbreaking model were seen when the average selective advantage was weak $s = 0.001$, due to differential cumulative fitness which is more evident when $s = 0.001$. The main difference in both outcomes is in the diversity. Although at $s = 0.001$ both models show similarity the increased mutation rate model is less descriptive of tumour composition due to its greater heterogeneity.

The increased mutation rate model shows unique properties of driver composition that are in between those of the additive fitness model and the stickbreaking model.

10 Summary of the Comparisons of Positive Selection Models

So far, I have individually described properties of each model at milestone tumour sizes 1 cm^3 and 4 cm^3 using the additive fitness model as a baseline for comparison. Table 3.2 provides a summary of the ranges of the outcomes of interest in the models.

Table 3.2 Ranges of Main Outcomes in the Tumour Evolution Models

Variable	*Additive Fitness*	*Stickbreaking*	*Increased Mutation*
Generations Log(t)	5 - 10 *	**5 - 11**	5 - 10
Fitness (s)	1 - 1.15 *	**1 – 2**	1 - 1.20 *
Log(RGS)	0 – 12 *	**0 - 14**	**0 – 14 ***
Detectable Clones	1 - 14	**4 - 20**	4 - 15
Drivers at 10%	1 - 4	**1 - 5**	1 - 4

Ranges consider maximum and minimum values in the Figures 3.2, 3.3, 3.7, 3.10, 3.11, S3.1, S3.2, S3.3 and S3.4.
* indicates that the variable can be used to distinguish the input parameters.
Bolded are the highest values per category.

It can take extra generations for a tumour to develop under the stickbreaking model. Although these cases were outliers, slow growing tumours in this model are indicative of an unusually low fitness, requiring additional accumulation of drivers to accelerate expansion. Outlier values also occur for RGS diversity. The overall range is similar to that of the increased mutation model, however the median values are lower which translates into better detection of clonality and driver composition.

The increased mutation rate model has the highest RGS diversity which limits the fraction of clones that can be detected. Increases in the mutation rate manifest a slightly higher fitness increase, though the number of generations and driver composition is the same as the additive fitness model.

As a result, the stickbreaking and the increased mutation rate models can provide complementary clonal histories to the additive fitness model encompassing multiple process of tumour development.

The previous sections focused in studying simulation outcomes at tumour milestone sizes 1 cm^3 and 4 cm^3. Figures 3.12, 3.13 and 3.14 record the key metrics of the tumour at every time and show all the simulation time-series coloured by average selective advantage. They show that the patterns observed at the milestone sizes are consistent during the course of tumour progression for all models.

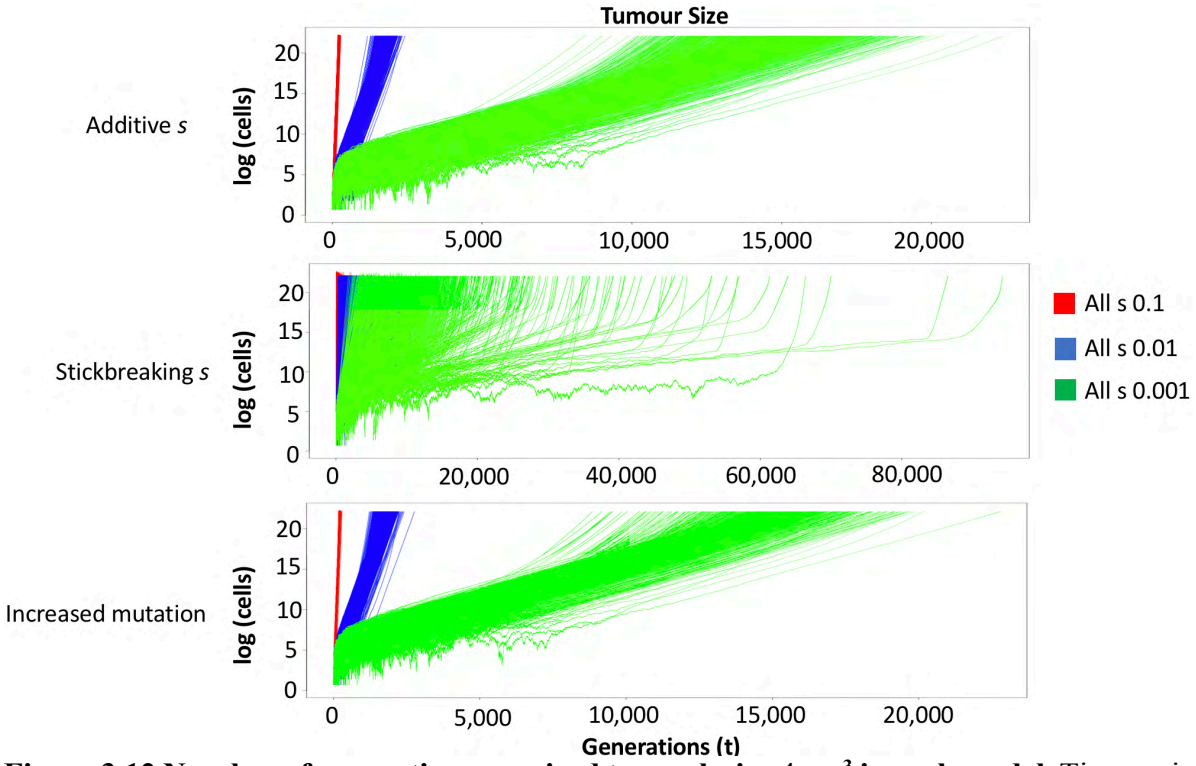


Figure 3.12 Number of generations required to reach size 4 cm^3 in each model. Time series is generated by recording the variable of interest at every time t for all positive selection models. Growth curves are coloured according to the average selection coefficient used in the simulation.

Figure 3.12 shows how in the additive fitness and increased mutation rate models, the value of s can be determined from tumour size and the number of generations based on the differences between growth curves. The divergence occurs early when the tumour is less than a million cells.

It also shows that the stickbreaking model has outlier simulations where tumour development can take a considerable amount of time, with simulations spanning 30,000 generations being less frequent. Interestingly, these outlier simulations indicate tumours can be developing but the lifespan is not enough to manifest symptoms. This is because the division rate has to be faster (12-24 hours) to accommodate a human lifespan, or malignancies need to be detected at early stages.

Figure 3.13 shows the relationship between fitness and tumour development, showing how a strong average selective advantage (coloured in red) causes tumour development in faster timeframes. In this setting tumours can have longer periods of average division (e.g. 4 - 7 days) and manifest in a couple of years or decades. Similarly, with a moderate selective advantage (coloured in blue), there can be flexible average division rates that still give a realistic timeframe for tumour development, years or decades. In contrast, the weak average selective advantage (coloured in green) requires faster division rates, on the order of 1 - 2 days. For instance, tumour taking 15,000 generations to grow up to 4 cm^3, would require ~41 – 82 years to reach that size with average division rates of 1-2 days.

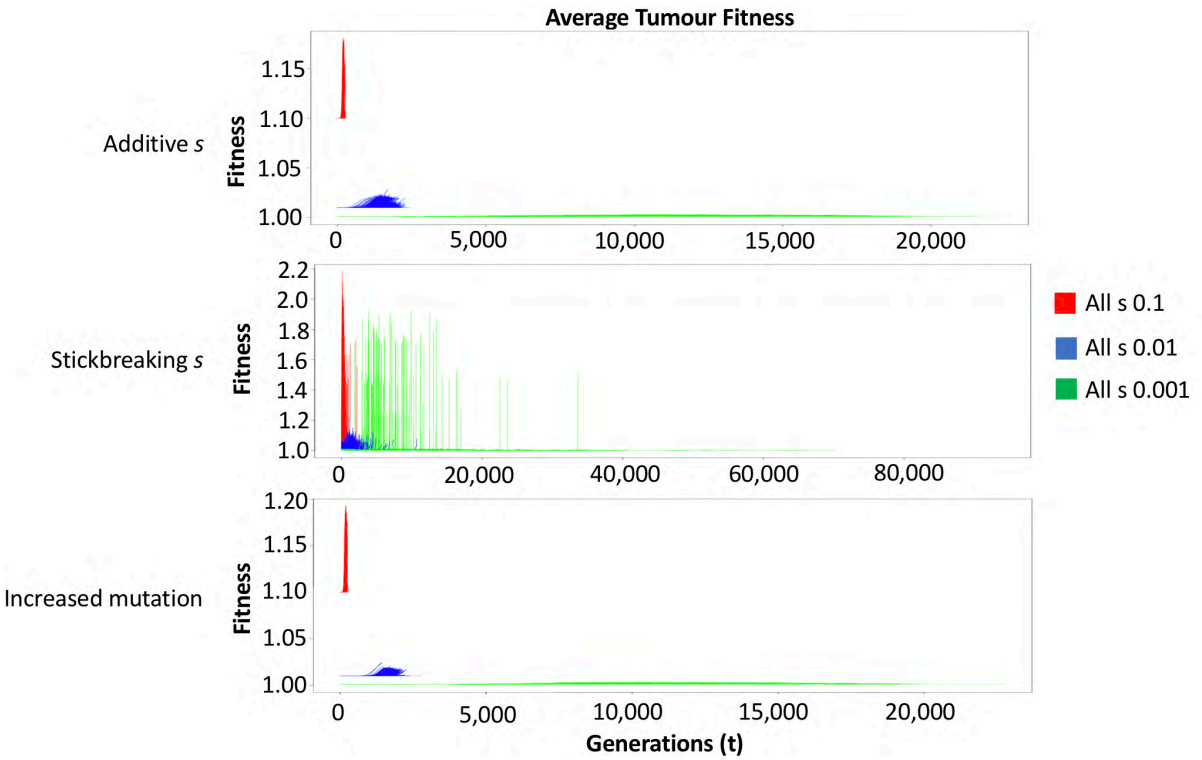


Figure 3.13 Fitness changes over time in each model. Time series is generated by recording the variable of interest at every time t for all positive selection models.

Similar to Figure 3.12, in Figure 3.14, measures of diversity can distinguish starting values of the average selective advantage s.

It shows how the models with $s = 0.001$ can recreate scenarios of higher and lower diversity with different clonality, explaining why cancer cell fractions overlap between parameters, as shown in Figure 3.8. Furthermore, it highlights how diversity emerges in specific timeframes for different values of s, indicating its relevance as a marker of the length of time it took a tumour to develop.

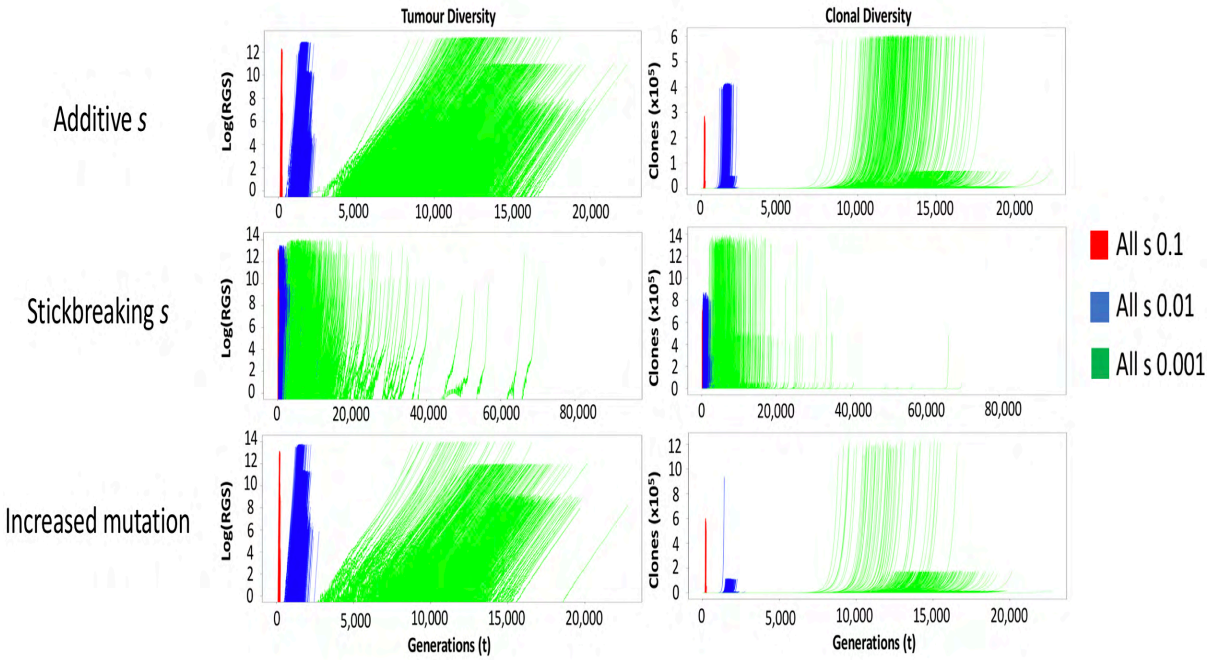


Figure 3.14 RGS diversity and number of clones over time. Time series is generated by recording the variable of interest at every time t for all positive selection models.

With the information provided in Table 3.2 and Figures 3.12, 3.13 and 3.14 it can be seen that the stickbreaking model has the largest variation in simulation outcomes providing numerous clonal trajectories that can be used for downstream analysis, and demonstrate the strong impact that changing s has on simulation outcomes.

The increased mutation rate model shows how mutation rate changes influence RGS diversity, fitness and clonal detection, and shows how increasing the average driver mutation rate u can have the effect of boosting fitness and clonal diversity.

The additive fitness model provides a simplified approximation of tumour development that although powerful has the limitation of not describing key features of tumour development such as hyperselection, hypermutants, the role of tumour suppressors, etc. As seen, adding variation in s and u can render multiple outcomes informative about the connection between mutational processes with tumour growth dynamics and clonal architectures.

10.1 Summary Comparison of Detectable Driver Composition

The positive selection models have revealed limitations on determining the initial conditions of s and u due to the overlap of cancer cell fractions due to the stochasticity of the models and limits on sequencing depth, Figures 3.5 and 3.8.

At the depth, variant- and clonality calling of 10% CCF, the main distinction that can be made regarding likely driver composition is to separate tumours 2 drivers versus those that have 3 or more drivers.

In Figure 3.15 , we see two clusters of similarity based on initial conditions, one with strong s and weak u ($s = 0.1$ & $u = \{3.4x10^{-7}, 3.4x10^{-6}\}$, coloured blue), and one with moderate/weak s with high u ($s = 0.001, 0.01$ & $u =$

$\{3.4x10^{-5}, 3.4x10^{-4}\}$, coloured brown). The results of the additive fitness model are in concordance with the analytical solutions given in Chapter II in Figure 2.10 and Table 2.1.

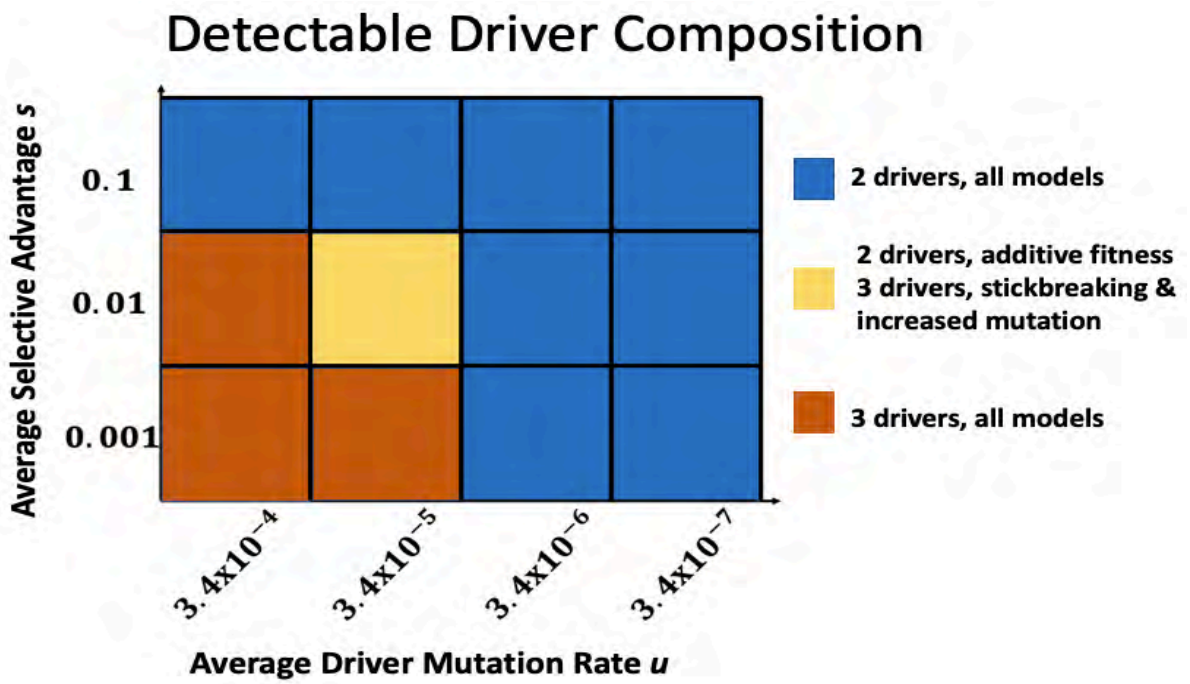


Figure 3.15 Summary of detectable driver composition across all positive selection models. Number of drivers accumulated by the dominant *k*-driver population with each parameter combination are coloured coded accordingly the dominant value of *k* and the differences between model implementations.

The information provided in Figure 3.15 relates the likely driver composition in the positive selection models with a 10% CCF cut-off. However, this is not directly inferred by sequencing and the association with the number of detectable clones is needed for the accurate recovery of clonal evolution. Based on data from Figures 3.15, 3.3b, S1.b and S3.b, in Table 3.3 it shows how the number drivers relates to their clonal composition.

Table 3.3 Ranges of Number of Clones in the Tumour Evolution Models

Detectable Drivers	Additive Fitness	Stickbreaking	Increased Mutation
2	1 – 5 [1]	1 – 5 [2]	1 – 5 [1]
2-3	3 – 8 [5]	4 - 10 [5]	6 – 11[6]

Ranges consider inter quantile range at a 4 cm^3 tumour size in Figures 3.3b, S3.1.b and S3.3.b.
Values in brackets represent the consensus median value.

The ranges reported in Table 3.3 are within the range with the cancer cell fractions reported by Andor et al. [29] (Figure 3.8), suggesting that increased clonality is linked to moderate to weak average selective advantage s and high average driver mutation rate u. A similar result was obtained in Chapter II with the analytical solutions.

Assuming a representative sample and exponential growth, the models suggests that at the commonly used CCF cut-off of 10%, cancer cell fractions can be used to determine the initial conditions of s and u. Strong s and weak u vs moderate/weak s with high u give distinct values for number of detectable clones, with increased clonality associated with the latter.

11 Clinical Implications for Pre-existing Drug-Resistant Cells

The origins of cellular drug resistance in cancer are still under debate. With multiple treatment approaches and interventions, drug resistance requires special consideration, being tumour-specific and drug-specific in many cases. Furthermore, some patients may not be eligible for a first-line therapy, adding complexity on how to characterise the rates of emergence of drug resistance in certain tumours. Additionally, the sequence and combination of drugs can impact the rates of emergence of resistant variants limiting a global generalisation. Bozic et al. provides a good summary of some of the considerations and implications when modelling drug resistance here [45], providing key analytical solutions applicable to the branching process.

In this thesis instead to aiming to recapitulate cancer-specific drug resistance rates, I used approximated value from [139] $\mu_r = 1\text{x}10^{-8}$ that Chowell et al. [44] incorporated in the clonal branching process as a general baseline. Chowell et al. [44] used this rate to evaluate the frequency of clones carrying at least one drug resistance mutation without fitness changes in the absence of treatment. It is worth noting the framework provided by Bozic et al. [123] that is incorporated here allows for estimates of the median number of drug resistance cells that emerge for a given clone $C_{k,i,j}$ at any time t based on any therapy-specific mutation rate, without need for running additional simulations.

I incorporated the drug resistance mutation rate in all the positive selection models to explore the dynamics of the emergence of drug resistant subpopulations during clonal evolution. Every drug resistance mutation in the model (primary formation) has a neutral fitness effect and once acquired resistance mutations are inherited by all cells in that lineage. With framework from Bozic et al. [123] deviations from neutrality can be made for the number of drug resistance cells without the need to re-run simulations. With this framework, outcomes of sensitive clones can be used to estimate the number of drug resistant cells within clones at different drug resistance rates μ_r.

Assuming neutrality in the fitness effects of drug resistance cells during primary formation, the number of drug resistant clones is proportional to tumour size as shown in Figure 3.16. As illustrated in Figure 3.17, drug resistant cells are already present when the tumour is 1 cm^3 in all models and further tumour development increases the prevalence of those cells. It can be seen in Figure 3.17 that increasing the average driver mutation rate u can increase the number of drug resistant cells in the stickbreaking model when, the number of resistant clones increases as well as the number of cells. This is further supported in the supplemental Figure S3.5, under $u = 3.4\text{x}10^{-4}$. In this scenario, drug resistant cells can acquire their own driver mutations, further accelerating their expansion.

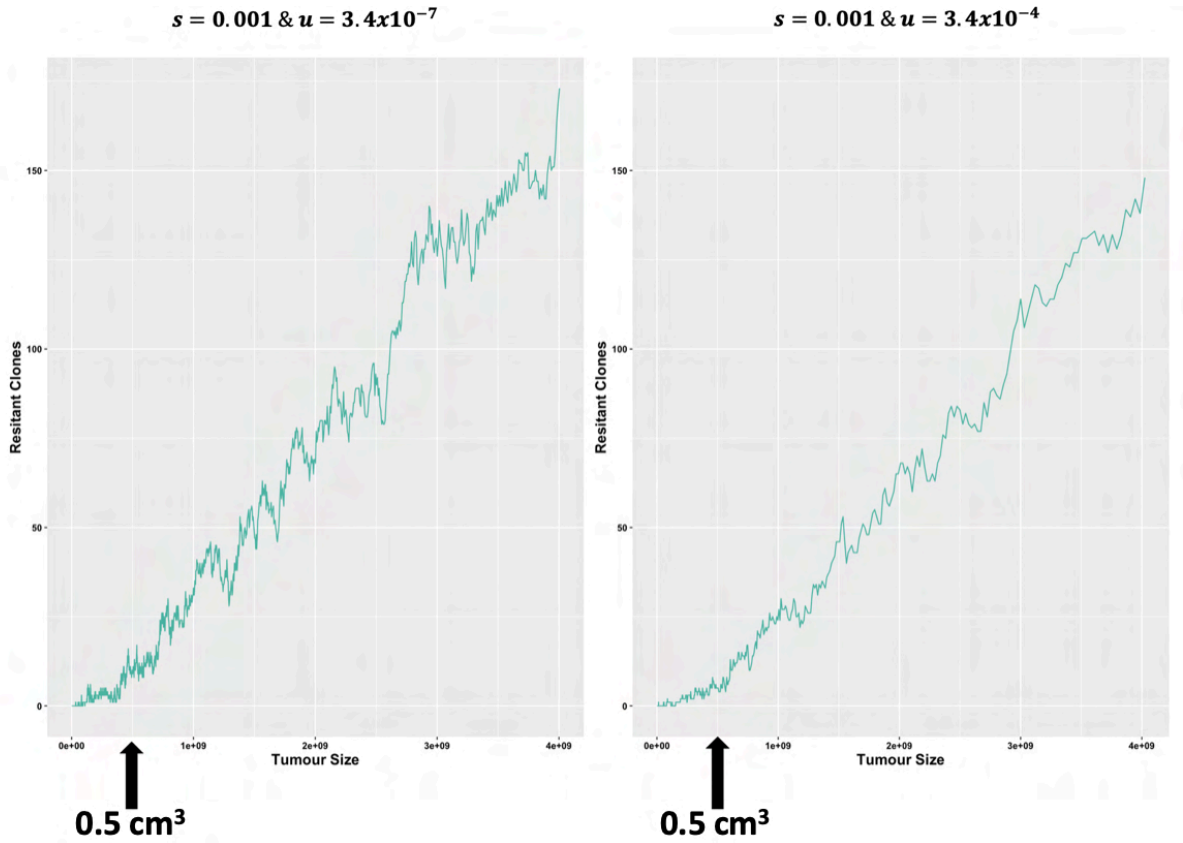


Figure 3.16 Relationship between number of resistant clones and tumour size. Time series of two different simulations of the stickbreaking process showing the proportional relationship between the number of resistant clones and tumour size. Arrows depict the tumour size at which drug resistance starts to emerge.

The main condition associated with the number of drug resistant cells is drift. High drift or reduced average selective advantage s increases the number of cellular divisions which increases the odds of drug resistance as compared to other parameters. This is shown in Figure 3.17 with parameters $s = 0.001$ and $u = 3.14\text{x}10^{-7}$ displaying higher numbers of drug resistant cells. The pattern is consistent across time as shown in Figure S3.6.

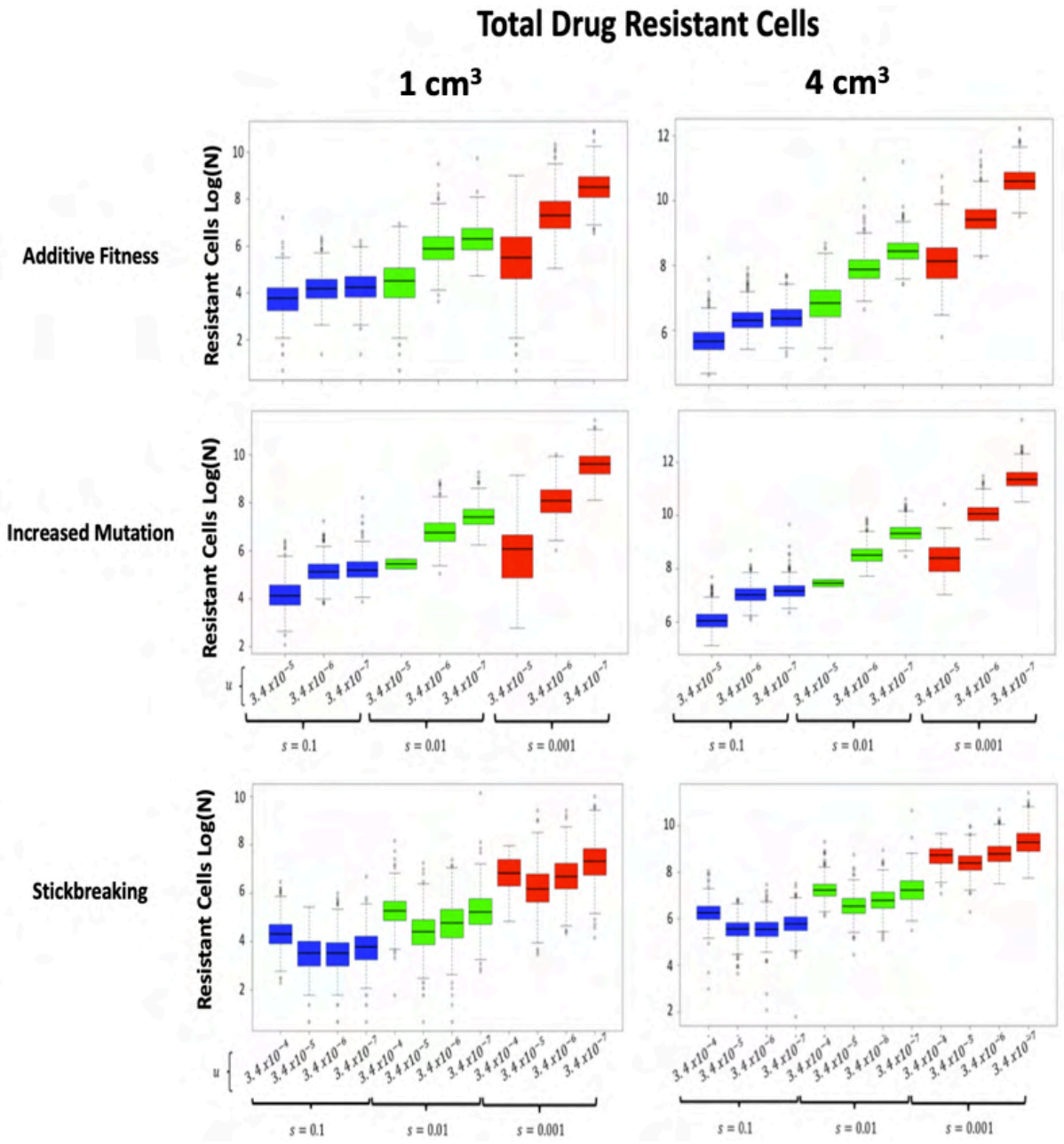


Figure 3.17 Total number of drug resistant cells. Boxplots are taken at milestone sizes considering all simulations in the positive selection models colour coded by average selective advantage *s*.

At milestone sizes, the number of drug resistance cells in the tumour comprise less than 0.001% of all cells composition which represents a significant challenge for the development of diagnostic methods to detect the presence of these cells. Drug resistant clones typically emerge when the tumour is around 5 mm^3 in size as exemplified with the black arrows in Figure 3.16, because drug resistance is proportional to tumour size the values shown in Figure 3.16 can be extrapolated to all parameters.

The *k*-driver composition of drug resistant clones is likely to be 1 - 2 for strong and moderate *s* and 1 – 3 for weak *s* as seen in Figure 3.18, the effect of increasing the average driver mutation rate *u* is in increase the number of drivers.

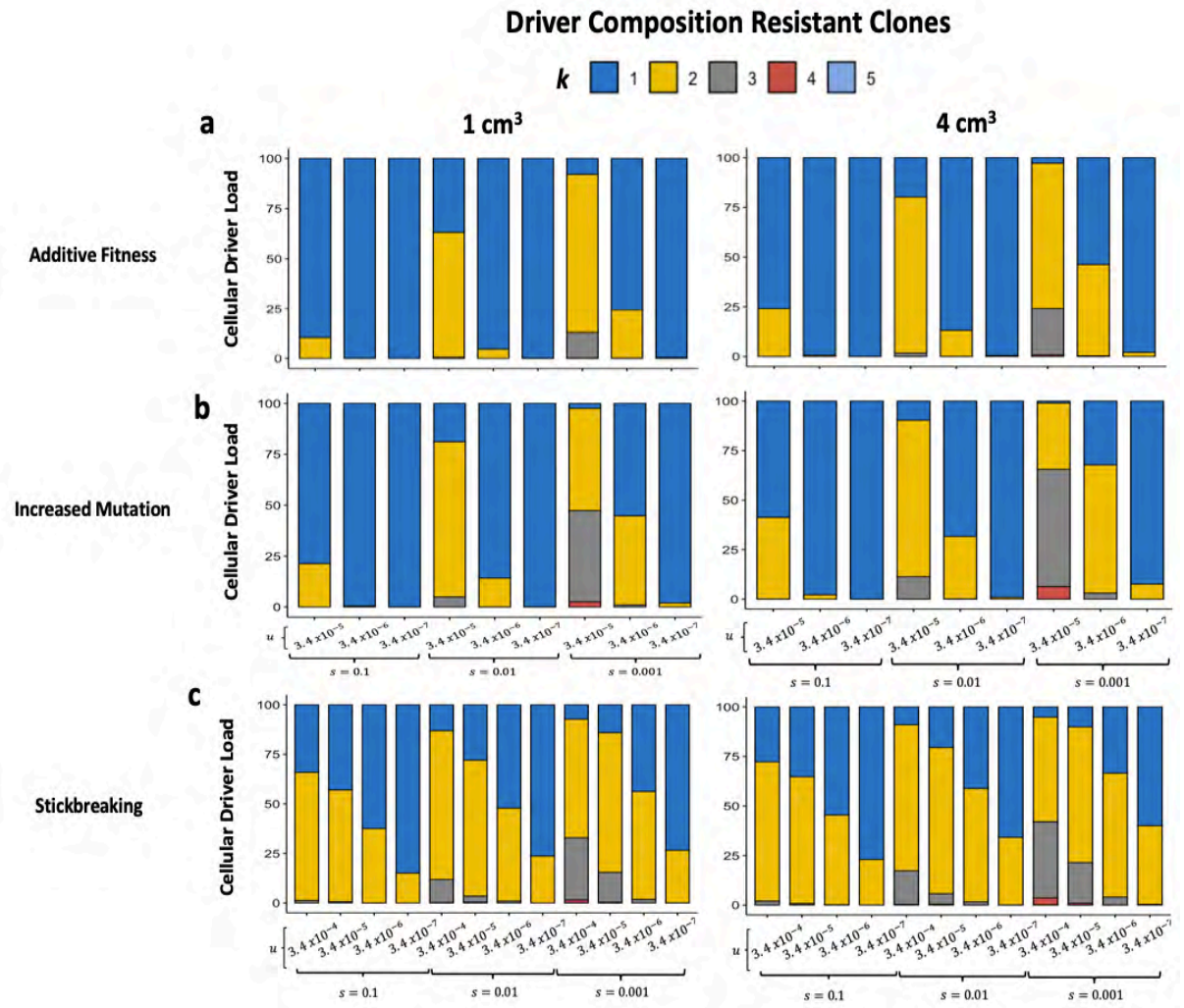


Figure 3.18 Bar plot of *k*-driver composition of drug resistant clones Boxplots are taken at milestone sizes considering all simulations in the positive selection models colour coded by average selective advantage *s*.

Considering neutral fitness effects, drug resistance is proportional to tumour size and represents less than 0.001% of all cells in the tumour. The number of drug resistant cells increases with drift, meaning lower the values of *s* result in greater numbers of drug resistant cells (excluding $u = 3.14\text{x}10^{-4}$).

12 Tumour Phylogenies Inferred from Cancer Cell Fractions ≥ 10%

A key advantage of the computational models used in this study is that they enable direct phylogeny reconstruction as a by-product of the snapshots collected during the simulations, allowing extraction of the precise measurable clonality at a given sequencing frequency cut-off (in a rep-seq condition). In this section, I will show the structures of phylogenies based on clones detectable at 10% CCF cut-off under each model and parameter set, and show how recurrent topologies can occur under different starting conditions.

The values reported in Figure 3.19 are in concordance with the median consensus values of detection clusters listed in Table 3.3 as determined from Figure 3.15. Furthermore, Figure 3.15 showed that at the 10% CCF cut-off most of the parameter combinations lead to tumours composed of clones with 2 driver mutations, suggesting they will primarily have simple

topologies. This is evident in Figure 3.19, with all pie charts having the two-clone nested topology depicted with phylogeny code B. This exemplifies why at this cut-off there is considerable overlap within cancer cell fractions because all parameter combinations give rise to similar topologies with simple branching patterns, as shown in Figure 3.8.a.

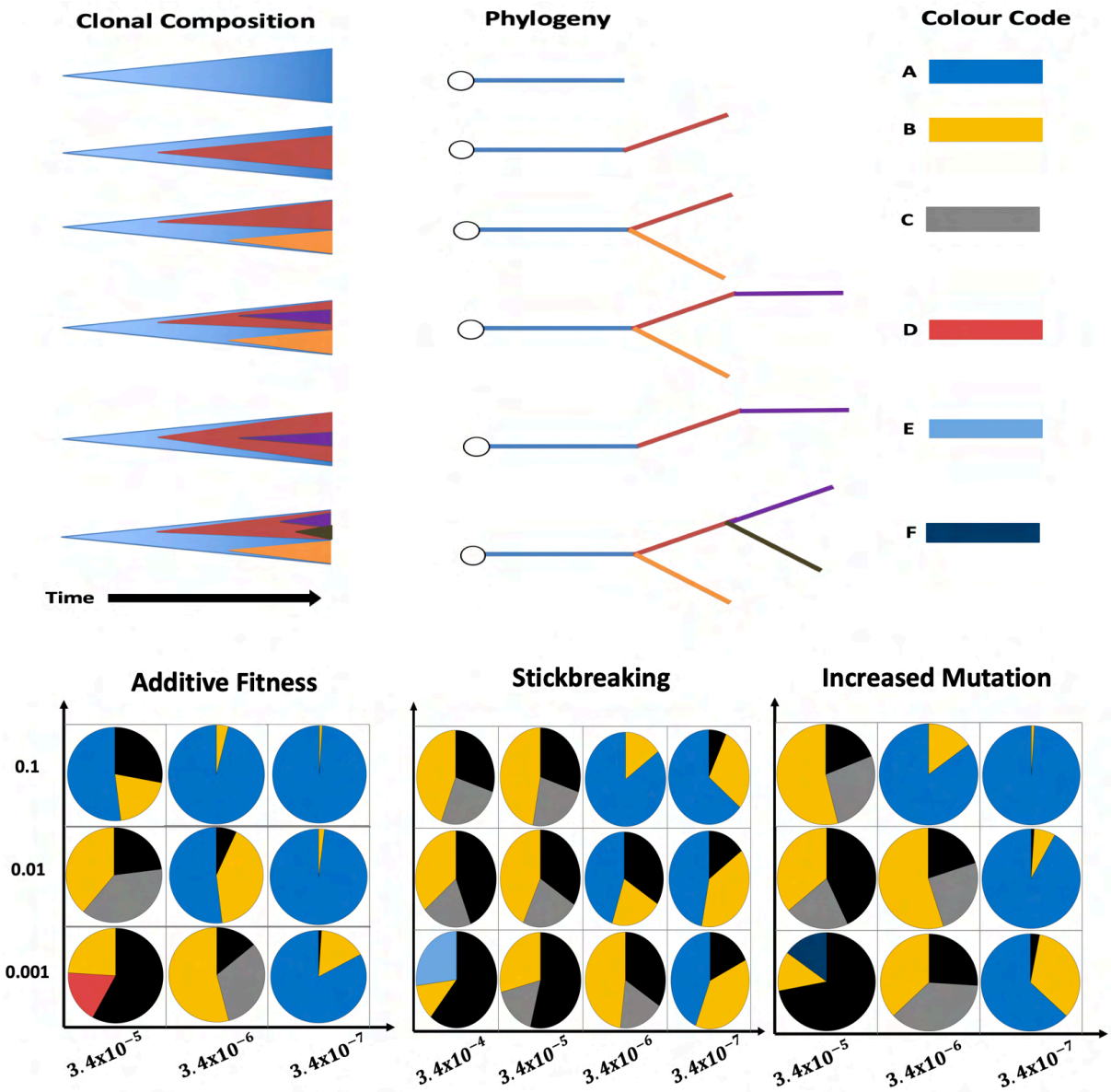


Figure 3.19 Summary of recurrent detectable driver composition in all positive selection models (assuming rep-seq). Phylogenies were determined using all simulations at the 4 cm^3 size. Top section shows Muller plots with their respective phylogenies, and topologies are assigned to a colour code and a letter for guidance. Bottom section, pie charts showing the frequency of the recurrent topologies colour coded according to the reference in the top section. Black sections in the pie charts represent the frequency of other topologies not depicted here.

As shown in Figure 3.19 a monoclonal phylogeny (A with blue colour code) occurs under 50% of the parameter combinations, most often when the driver alterations confer a strong selective advantage or there is moderate to lower average driver mutation rate, i.e., $s = 0.1$ & $u = 3.4\text{x}10^{-7}, 3.4\text{x}10^{-6}$.

Two nested driver clone phylogeny (B with colour code yellow) occurs in all cases at varying frequencies. Such topology can present a problem when determining the starting values of average selective advantage *s* and average driver mutation rate *u*. The expected values reported in Chapter II suggest that this phylogeny is more likely to occur under parameter combinations of $s = 0.01$ & $u = 3.4\text{x}10^{-6}$ and $s = 0.1$ & $3.4\text{x}10^{-5}$ in the additive fitness model.

Phylogeny C, colour coded grey, occurs in 40% of the parameter combinations. It does not occur when the average driver mutation rate *u* is low $3.4\text{x}10^{-7}$.

Phylogenies D and F contain clones with 3 driver alterations and occur in the additive fitness and increased mutation rate models with parameters $s = 0.001$ & $u = 3.4\text{x}10^{-5}$. However in the stickbreaking model, 3 driver clones only appear with phylogeny E, under parameters $s = 0.001$ & $u = 3.4\text{x}10^{-4}$. With these parameters, there is great diversity in the number of different observed topologies beyond the ones shown here, as indicated by the black segments in the pie charts. A benefit of the high diversity seen at these parameters is it provides a broad set of unique trees across simulations that facilitates the inference of the true values of *s* and *u* based on cancer cell fractions.

At a 10% CCF cut-off there are 6 recurrent topologies that can be solved manually. More diversity in tumour phylogenies could be recovered by increasing sequencing resolution. As shown in Figure 3.19 using as an example the additive fitness model, more clones can be detected and hence different topologies can be established with lower CCF cut-offs. However, the variability recovered at 5% CCF cut-off does not represent a significant increase over the number of recurrent topologies recovered, with only 3 parameter combinations (framed in red) showing any difference. At 1% CCF cut-off or less however, more clones can be measured providing more unique phylogenies and hence better discrimination between parameters as shown in Figure 3.20.b.

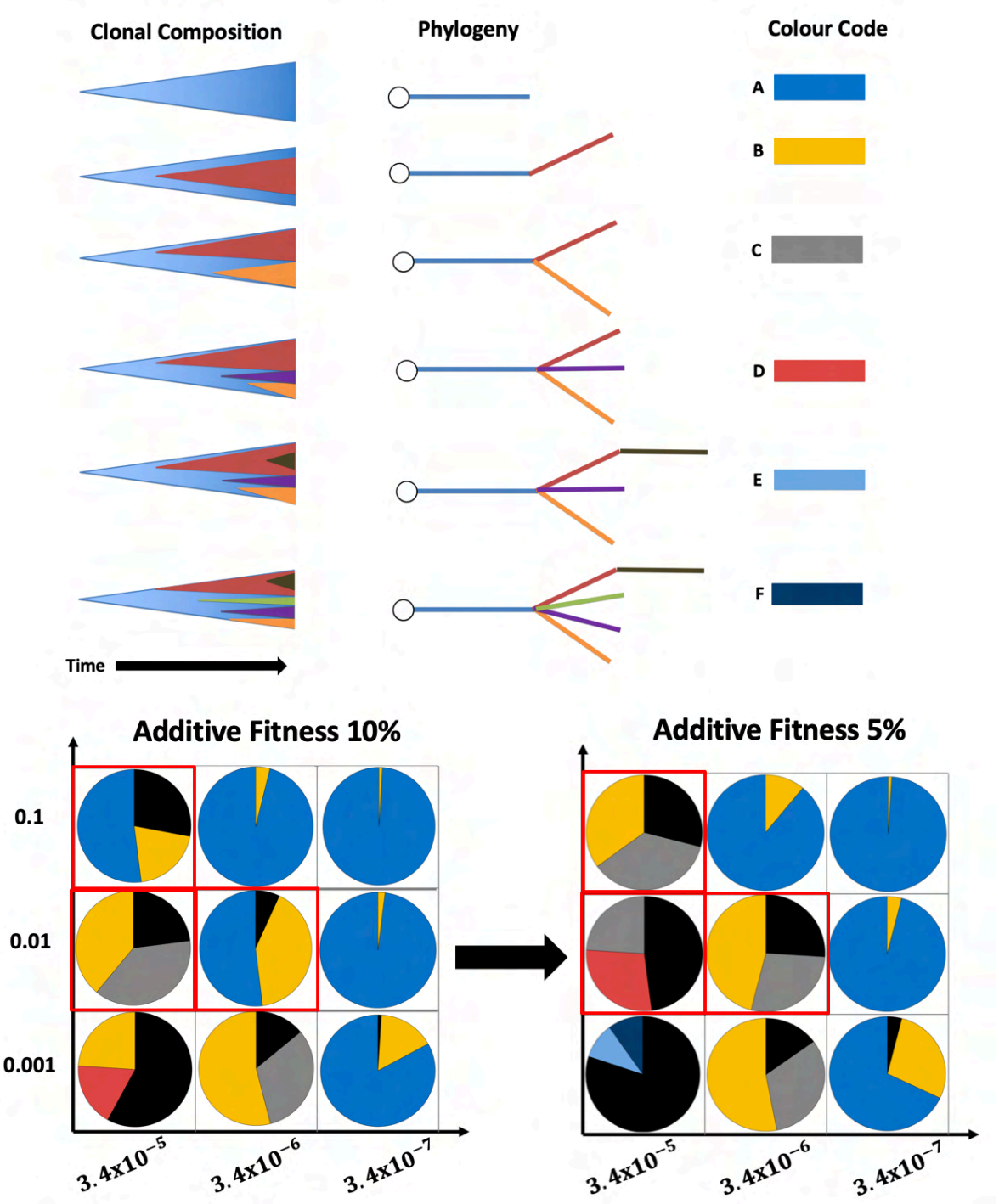


Figure 3.20 Summary of recurrent detectable driver composition in the additive fitness model at the 10% and 5% CCF cut-offs (assuming rep-seq). Phylogenies were determined using all simulations at the 4 cm^3 size. Lowering the CCF threshold to 5% only noticeably changes the distribution of phylogenies for three parameter combinations. Top section shows Muller plots with their respective phylogenies, and topologies are assigned to a colour code and a letter for guidance. Bottom section, pie charts showing the frequency of the recurrent topologies colour coded according to the reference in the top section. Black sections in the pie charts represent the frequency of other topologies not depicted here. Highlighted in red are parameter combinations that show different topologies at the lower frequency.

Jointly, Figures 3.19 and 3.20 show that variation in observed phylogenies is not only a function of variant- and clonality calling but is also affected by how the combination of parameters shape clonal evolution. For instance, strong s with low u ($s = 0.1$ & $u = 3.4\text{x}10^{-7}$), leads mostly to single or two branched topologies (phylogenies A and B). In other parameter combinations such as weak s and high u ($s = 0.001$ & $u = 3.4\text{x}10^{-5}, 3.4\text{x}10^{-4}$) would have to have taken on such topology at some point during clonal evolution before acquiring additional branches. Therefore, parameter combinations can dynamically intersect at different timepoints.

This shows why weak s and high u are the most variable combinations, because during clonal evolution they not only it spans similarity in the cancer cell fractions but also share topologies with other parameter combinations of s and u.

In summary, the number of unique topologies increases proportional to u and inversely proportional to s with 1 to 2 branch topologies being the most prevalent across parameters and models. Weak average selective advantage $s = 0.001$ results in the greatest degree of variability in topologies due to slow net-growth rate, this suggesting that increased clonality is likely to be originate under conditions with strong s and low u ($s = \{0.001, 0.01\}$ & $u = \{3.4\text{x}10^{-5}$ & $3.4\text{x}10^{-4}\}$). This was mentioned in the in Chapter II with the increased extinction probability δ rendering diversity in the growth patterns and reported by Durrett et al. [165].

13 Neutral Evolution

The previous sections have described positive selection models, which are more useful when at least two driver subpopulations can be detected from a sequencing sample. As shown from analysis of distortions in variant allelic frequency distributions [86, 88] in data from TCGA , the overall fraction of neutral samples is as high as 30%. In those scenarios the most informative component of the variant allelic frequency distribution comes from the neutral tail, which contains information about the passenger signal, as shown in Figure 3.21.b. This signal accumulates stochastically with neutral fitness during clonal expansion. The passenger signal can be used to estimate the initial values of average selective advantage s and average driver mutation rate u.

Although the neutral mutations do not provide a fitness gain, the excessive accumulation of passengers is believed to reduce fitness as suggested by MacFarland et al.[130] . A secondary effect of in passenger accumulation is the presence of mutations that may confer gain-of-function when environmental conditions change (e.g. treatment). A good example of this is drug resistance. It is an event with low probability that is apparent by environmental intervention such as treatment.

This process can be simulated with the branching process by setting an initial fitness s that does not change over time, and recording the passenger propagation with mutation rate $\nu = 0.016$ (the reported value from Bozic et al. [58]), as illustrated in Figure 3.21.a. To evaluate the strength of the passenger signal, I computed the dynamics of its accumulation using the branching process for different values of s and k for tumours of up to 100 million cells.

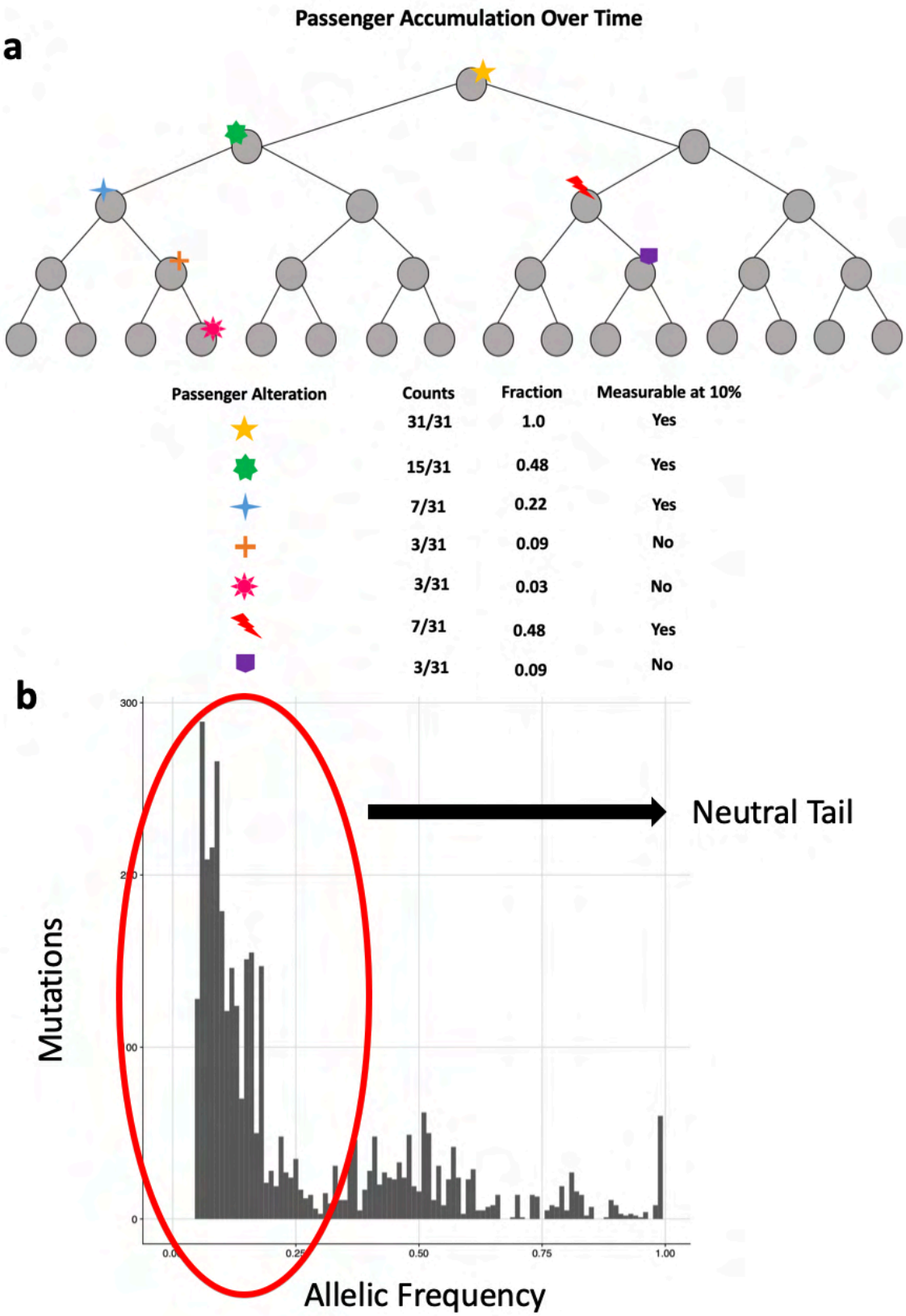


Figure 3.21 Dynamics of passenger accumulation under neutral tumour evolution. **a**, example of the passenger accumulation implemented in the neutral fitness model. **b,** an example of a variant allele frequency distribution from a neutral sample with the neutral tail highlighted in red.

In addition, the positive selection models can recover the passenger signal using the fixed number of passengers as $C_n = {\nu(1-\alpha)}/{(1-\delta)\alpha}$ provided by Bozic et al. [58]. Here, α is the sequencing detection threshold of CCFs, expected to be ~10% in whole exome assays, and δ is the lineage survival probability as $\delta = d_k/b_k$. Figure 3.22 shows the fixed number of passengers calculated for the positive selection models.

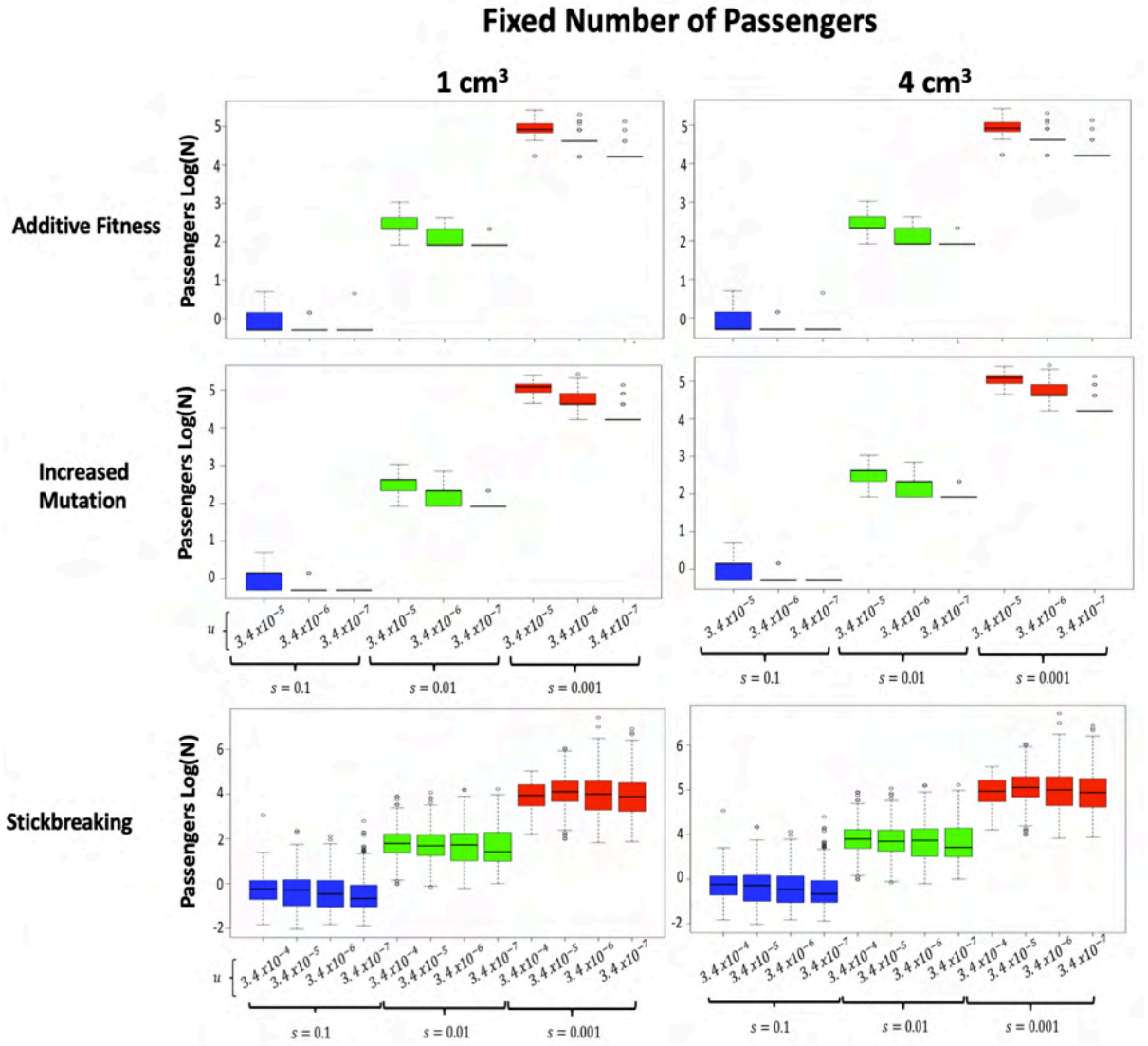


Figure 3.22 Summary of number of detectable passenger mutations in all models. Boxplots are taken at milestone sizes considering all simulations in the positive selection models colour coded by average selective advantage *s*.

The passenger signal increases with reductions in the average selective advantage s. This is translated in a high extinction probability δ causing the net-growth rate to be slow and requiring more cellular divisions to expand. As a result, the passenger accumulation increases over time. The number of passengers is also somewhat affected by the average driver mutation rate as it influences the number of detectable clones, and therefore the passenger signal can aid in determining parameters s and u from neutral samples.

To measure the strength of passenger signal I used the neutral fitness model, as outlined in Figure 3.21.a. As illustrated in Supplementary Tables 3.1, 3.2 and 3.3, depth of sequencing

determines the allele frequency resolution and imposes limits on the measurement of the passenger signal. A 1% CCF cut-off (representative of ultra-deep sequencing assays) is the most informative but this is difficult to achieve. Consequently, sequencing depths that cannot resolve confidently 5% - 10% frequency variants are not sufficient to reveal a solid passenger signal at clonal sizes of 100 million, with bigger subpopulation sizes required to enrich the signal.

Although these values may change if the model is allowed to expand to sizes up to 1 billion, they provide a picture of how critical sequencing depth is for obtaining accurate estimates as to the true values of the parameters of interest, as well as how representative a biopsy sample may be of the complete clonal landscape of a given tumour.

In summary, the stochastic passenger accumulation can be used an alternative approach for samples showing evidence of neutrality. It can also aid in determining parameters s and u, however the accurate estimation of the passenger signal is influenced by measurement biases arising from such factors as tumour purity, sequencing coverage and depth.

14 Applications for Reconstructing Tumour Evolution

One of the main motivations for generating such a large collection of tumour simulations, is to have a database of possible tumour evolutionary trajectories that can be used to infer clonal ancestry and relatedness through comparison to ploidy corrected sequenced data from tumour samples. As shown in Figure 3.1.a, the goal is to provide a complete picture of clonal evolution that is lost due to reduced sensitivity of current sequencing depths and variant calling sensitivities.

This is possible by comparing the clonality estimates from tools such as PyClone or ExPANdS with the data from my simulations. The comparison can assign tumours to one of the two clusters identified in Figure 3.15. The first cluster is the one that display strong fitness and/or low mutation rates ($s = 0.1$ & $u = \{3.4\text{x}10^{-7}, 3.4\text{x}10^{-6}\}$). The other cluster is the one that has moderate to weak fitness effects and high mutation rates ($s = \{0.001, 0.01\}$ & $u = \{3.4\text{x}10^{-5}, 3.4\text{x}10^{-4}\}$). Chapter IV will explore different statistical methods for comparing simulated cancer cell fractions with real ploidy corrected cancer cell fractions.

The positive selection models recorded the top-100 clonal subpopulations in the simulations at every snapshot, which together explain ~60 – 100% of the clonal composition of the simulated tumours providing enough information to have a broader picture of clonal evolution. The parameters reported by Bozic et al. [43] as being representative, $s = 0.001$ & $u = 3.4\text{x}10^{-5}$, have more variability in clonality and can encompass multiple tumour architectures and evolutionary trajectories.

Real samples that have an increased number of inferred clones such as SKCM in the Andor et al. [29] study in Figure 3.8.b, are better explained by moderate/weak average selective advantage s and high driver mutation rate u. Chapter IV is going to explore if the diversity in simulation outcomes for these parameters help to be identify such evolutionary trajectories when comparing simulated vs real cancer cell fractions.

Neutral samples can provide information on s and u by comparing the passenger tail with the fixed number of passengers, equation using the δ values of clones as shown in Figure 3.22.

Alternately, one can fit equation for the fixed number of passengers given in Section 13 to the neutral tail distribution, as suggested by Bozic et al [58].

Chapter IV will explore a framework to reconstruct tumour evolution based on simulation results considering the caveats and limitations observed in this chapter.

15 Discussion

The models of clonal evolution I present here agree with commonly accepted features of tumour development, with 3 - 6 alterations required to fuel cancerous growth [11, 181]. In addition, they suggest the existence of considerable heterogeneity, in concordance with genome wide pan-cancer studies and tumour manifestations in the clinic [29, 30].

The stickbreaking model has the main advantage of exploiting the fitness landscape due its stability of cumulative fitness sampling, and its reported robustness in recovering evolutionary patterns [138, 142]. The increased mutation rate model showed that changes in the mutation rate not only increase diversity but also approximate the simulation outcomes from the stickbreaking model.

Although the additive fitness model does well represent the dynamics of clonal evolution, fixed changes in cumulative fitness will struggle to capture dynamics that require fitness-shifts such as the *Big-Bang* evolutionary mode.

Metrics associated with proliferation and time have strong association with average selective advantage s and can aid in establishing the true fitness of a tumour. This is because s has a direct effect on proliferation and lineage survival. Therefore, clinical markers related to proliferative capacity, such as Ki-67, could be used to compare with fitness estimates from the positive selection models that can be used as a prognostic model.

In contrast, the variables associated with diversity at the detectable sequencing threshold do not provide enough power to infer to s and u. This is due to the high number of competing clonal subpopulations present at different times and with different fitness, which favours the emergence of a few dominant subpopulations. This impacts parameter estimation at the default sequencing detection level of 10% CCF cut-off.

It is possible to differentiate cases with strong fitness and/or a weak average driver mutation rate from those with weak selection and a high average driver mutation rate. Classification of a tumour based on more specific and unique patterns would be possible with data powered to go below the 5% - 1% CCF cut-off, but this is usually unfeasible in longitudinal studies with current sequencing technologies and their associated costs.

The main complication observed in the models tested, was the overlap in CCF distributions between different parameter combinations s and u. This implies that similar evolutionary modes (monoclonal, oligoclonal, polyclonal, etc.) can be achieved with different parameters under representative sampling (e.g. rep-seq) at the 10% CCF cut-off. In other words, there is not enough uniqueness in the CCF distributions that allow discrimination between evolutionary patterns when a representative sample is taken (at 10 and 5 % CCF cut-off). This can limit our inference in experimental studies for the accurate assignment of evolutionary parameters s and u using sequencing data (e.g. ploidy corrected and the clonality called). This highlights again

the relevance of using clinical markers to inform beyond the genotypic information obtained by 'omics technologies.

The applicability of the models used here in reconstructing tumour evolution will be hampered due to the overlap of cancer cell fractions between different sets of parameters. They can however be used to determine if tumour fitness is strong or moderate to weak if a tumour's age or proliferation rate is known. Improvements in classifying tumours can be made by incorporating the signal from passenger mutations when available, and the number of passengers can provide further evidence supporting estimates of the key evolutionary parameters s and u.

At the rate of $\mu_r = 1\text{x}10^{-8}$ the emergence of modelled drug resistant cells occurs when most tumours have reached 0.5 cm^3 in size, this is the critical mass at which drug resistant variants emerge. Therefore, mutational events occurring at the low rate of $\mu_r = 1\text{x}10^{-8}$ used here are helpful to illustrate that tumour size and drift δ can promote abundance of these cells, which in turn, can limit therapeutic interventions if disseminated. As a result, the exponential growth models of tumour evolution used in this chapter suggest that both increases in tumour size and the effects of drift can introduce novel traits associated with cellular robustness and adaptability.

Under the models tested, drug resistance is proportional to tumour size and increases with drift. This is because net growth is slow, requiring more cellular divisions. hence, the odds of resistant cells emerging and proliferating increases given the longer window of time compared with higher values of *s*. Drug resistant cells appear when the tumour is 0.5 cm^3 in size (~2/3 of tumour's age) and originate from lineages carrying 1 - 2 drivers that are abundant during that period, resulting in drug resistant cells with an initial lower fitness relative to the tumour expansion. As a result, drug resistant cells require roughly similar time-frame for clinical detection (e.g. palpable at 1 cm^3).

The number of generations it takes to go from 0.5 cm^3 to the ending 4 cm^3 size used in my simulations corresponds to the final ~10% of the total time of tumour development. This is in agreement with estimates from Goldie & Coleman [182] about the frequency of pre-existing drug resistance (in chemotherapy) cells before diagnosis [45, 125, 183].

The relationship between tumour size and drug resistance suggests that therapeutics should aim to reduce the maximum the number of cells as much as possible in tumours with weak average selective advantage s. This is because the net growth of sensitive cells is slower allowing for resistant cells to achieve higher abundance as compared to other values of *s,* in which sensitive cells propagate faster to detection and may not disseminated yet.

For instance, this can explain why long-term adjuvant treatment in estrogen receptor (ER)-positive early breast cancer may show clinical benefit and reduced mortality after 10 years [184, 185]. This comes from a study evaluating the mortality of tumour size in breast cancer in node positive (N1+) cases, which also showed increases in tumour size increase almost linearly with the risk of cancer-specific mortality in ER-positive and -negative cases–implying potential activation of a tumour dissemination program. A similar effect as found in the node negative cases (N0), suggesting the relevance of tumour size as a proxy of drug resistance/dissemination [186].

On the other hand, the stickbreaking model can be used to explain why bigger tumours do not necessarily represent increased risk (e.g. in the malignancies where an increased tumour size does not impact prognosis), with a potential explanation being that tumours with very low fitness are faced with severe drift. Then, after many attempts, hyper-selection occurs causing aggressive expansion within a short period of time. Despite the emergence of drug resistant clones, the accelerated progression does not provide enough time for dissemination. This suggest that fitness leaps can be beneficial (therapeutically) by accelerating the rate of clinical detection and consequently, avoiding dissemination of emerging resistance cells from the primary site.

Carrying capacities can also represent a clinical risk because sustained proliferation of tumours at a restricted size can cause accumulation of drug resistance mutations, as shown by Bozic et al. [187]. Chapter II provided analytical solutions of carrying capacities if their effect is relevant to study.

My findings on the association of drug resistance with reduced fitness in simulated data sets aligns with studies that suggest increased genome-wide alterations are associated with a poor prognosis [75, 188]. For instance, it is widely known that alterations to TP53 increase genomic instability and are associated with therapeutic failure. Based on observations from the models implemented here, tumour suppressor alterations such as TP53 may lead to subsequent mutations that reduce fitness by increasing overall mutational burden. The extra cellular divisions required to overcome this fitness reduction may increase the odds of drug resistance.

In contrast, certain oncogenes can accelerate expansion and provide therapeutic benefit indirectly. For instance, in breast cancer HER2 and PIK3CA enriched malignancies have better prognosis to those that have accumulated genome-wide copy number alterations or TP53 mutations [189]. Other studies that support this argument have shown the accumulation of such as passengers mutations reduces cellular fitness [122, 130] and consequently increases the odds of drug resistance emerging.

However, cancer is not a linear disease and the exact associations of phenotypes with genotypes may vary within tumours, between patients and amongst subtypes. Therefore, it is important to determine how molecular alterations under the evolutionary framework can explain biological determinants relevant to clinical manifestation. The simulations shown in this chapter can aid in evaluating evolutionary properties of tumour progression that can be subsequently tested with experimental data (e.g. clinicopathological and molecular markers).

The process of drug resistance can be generalised as the effects of passenger mutations that do not have an impact on tumour expansion but then become advantageous in a different environmental context. Such alterations, although at very low frequency, can fuel cancer specific mortality traits, such as resistance, dissemination or biological programs related to therapeutic failure.

In summary, the models suggest that tumours with high average selective advantage can elevate risk for patients by the aggressive expansion of clonal subpopulations, disrupting biological functions. However, lower average selective advantages can also put patients at risk by enabling accumulation of low-frequency passenger gain-of-function alterations that can act as a reservoir for potential drug resistance when environment changes.

To study drug resistance with the branching process, it is important to model tumour-specific and drug target specific dynamics. There are analytical solutions models that can be incorporated to the simulated tumours to expand or incorporate to current simulations, e.g. [45]. For instance, at a target tumour size *N*, a given clone $C_{k,i,j}$ has achieved certain size and potentially some cells within the clone had acquired drug resistance. With the analytical solution provided by Bozic et al. [123], it is possible to estimate the number of drug resistance cells at a given rate μ_R within that clone without re-simulating with parameters *s* and *u*.

The dynamics of the models suggest that increased mutational burden is associated with lower values of *s*, in which *u* may not have a strong effect. Lower values of *s* will increase the number of cell divisions required to reach a given tumour size enriching passenger accumulation, while increasing *u* will introduce new clones diluting the passenger signal providing less measurable mutational burden.

High mutational burden has clinical importance as a biomarker of prognosis for certain malignancies [190], and for some subtypes it is a biomarker of response to immune checkpoint blockade [191-193]. Applying an evolutionary view to understand the association of tumour mutational burden with prognosis may elucidate the conditions under which it can be used as a biomarker, as well as in the understanding of why is a good biomarker.

A recent model of solid tumours by Noble et al. [173] aimed to characterise the dynamics of tumour heterogeneity with survival. In their model, to reflect a solid tumour development, growth is constrained by a carrying capacity deviating from the exponential assumption implicit in the models studied here. The authors explored similar ranges of fitness effects *s* (multiplicative fitness effects) and driver mutation rates *u,* identifying that the mean cell division rate had positive correlation in predicting future growth rates. This is supported in the models, by showing how tumours with different starting values of *s* take well defined timeframes for tumour development, i.e. the net growth rate $\lambda_0 = b_0 - d_0$ has classification power.

To establish progression-free survival, authors evaluated the time required for drug resistant cells to develop once the tumours has reached a certain size. They then eliminated the sensitive cells and studied the time taken for the resistant subpopulation growth to achieve a given size. They showed that higher diversity and higher driver mutation rates are associated with faster relapses. Within the conditions that the authors modelled, we expect that progression free-survival will be inverted in our models for diversity (low s=0.001 and low s:{0.01, 0.001} high u:{ $3.14\text{x}10^{-4}$, $3.14\text{x}10^{-5}$}), meaning lower diversity will lead to faster relapses. Although higher values of *s* are associated with lower abundance of resistance cells, once those are present, the time to growth will be proportional to the magnitude of *s*. Similarly, once drug resistant cells are present higher values of *u* will boost expansion leading to faster relapses. This was evident in the stickbreaking model showing increased number of drug resistant cells when the driver mutation was high ($u = 3.14\text{x}10^{-4}$). If we assume that resistant cells disseminate relative to their abundance at the primary site, then it would be expected that our results align with the findings of Noble et al. Lower values of *s* will lead to more clonal diversity and consequently a higher accumulation of drug resistance cells with greater diversity of drug resistance mechanisms. However, drug resistance cells with reduced *s* will take longer to be detectable due to slow proliferation, manifesting in longer clinical responses relative to higher values of *s*.

This reiterates the relevance of studying tumour evolution with different experimental designs to make evident the dynamics that can potentially explain real observations. Further investigation on how to align experimental data with evolutionary models for predictive and prognostic purposes are required.

The database generated here with the positive selection models can be used to approximate tumour fitness and average driver mutation rate and provide biological insights about tumour composition and clonal evolution that cannot be made with sequencing and clonality tools alone.

16 Appendix and Supplementary Figures and Tables

A.1 Metrics of Diversity

Rich-Gini Simpson Diversity Index

There are multiple parameters of potential use in measuring tumour heterogeneity, as highlighted by [194]. When the number of species (clones) is large, classic measures of diversity like Shannon's lose critical power for additive partitioning of species diversity, becoming non-informative. To tackle this problem a modified version of the Gini-Simpson index was proposed called the Rich-Gini-Simpson quadratic index of biodiversity (RGS). This index was shown to overcome the problem the large species diversity while preserving all features of the Gini-Simpson index. The following is the equation of RGS,

$$RGS(t) = N(t)\sum_{i}\frac{X_i}{N(t)}\left(1-\frac{X_i}{N(t)}\right) = N(t)\left(1-\sum_{i}\left(\frac{X_i}{N(t)}\right)^2\right)$$

Where $N(t)$ is the tumour size at generation *t* and X_i is the number of cells of clone *i.*

Due to the robustness of the RGS metric in large population diversity I used this as the preferred metric of tumour heterogeneity.

Shannon Entropy

Entropy was measured using the clonal proportion of the tumour $N(t)$with X_i cells for all clones as following,

$$H = -\sum_{i=1}^{C}\frac{X_i}{N(t)}\log\left(\frac{X_i}{N(t)}\right)$$

Shannon Equity

Shannon equity was computed based on the entropy *H* as following,

$$H_e = \frac{H(t)}{\log\left(Z(t)\right)}$$

Where *Z* represents the number of clones in the tumour at generation *t*.

A.2 Similarity Score for Comparing Cancer Cell Fractions

The similarity score was implemented as following,

1. Let X and Y be cancer cell fractions with respective sizes, $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_M$.
2. Evaluate all pairwise intersections of X and Y as $1_{x_i \cap y_{i \pm \epsilon}}$ to generate a matrix of intersections $\boldsymbol{Z}$ of size NxM.
3. With the matrix of pairwise intersections $\boldsymbol{Z}$ generated, evaluate a global fit by evaluating θ as following,

$$\theta = 0.5\left(\frac{1}{N}\sum_{i}^{N}\sum_{k}^{K} Z_{i,k} + \frac{1}{M}\sum_{j}^{M}\sum_{k}^{K} Z_{j,k}\right)$$

The value of θ measures the degree of similarity between cancer cell fraction vectors X and Y in a range from 0 to 1 with a tolerance of measurement error of $\pm\varepsilon$. To model scenarios presented in practice by sequencing erroI set the tolerance ε to be 5%.

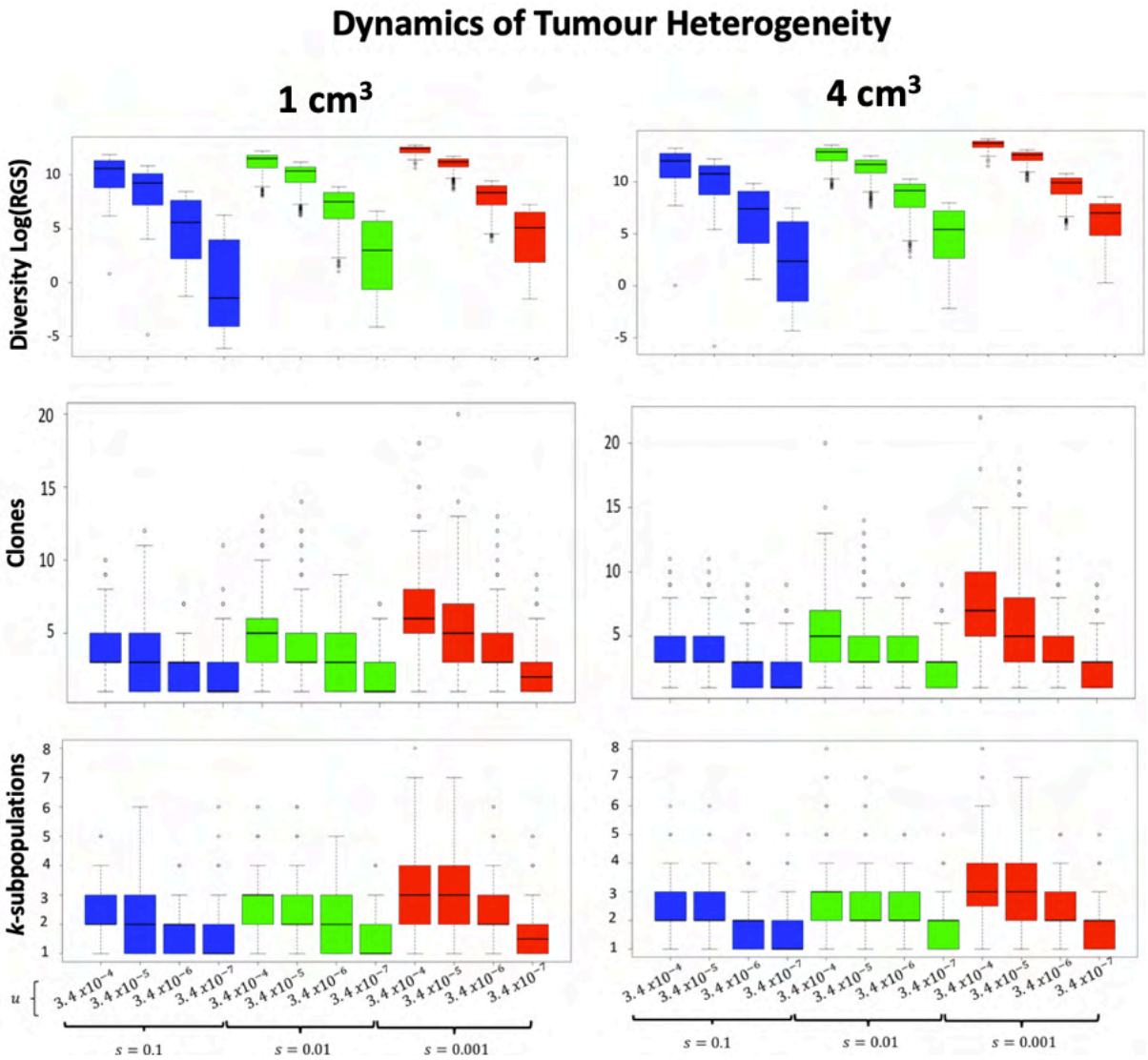


Figure S3.1 Boxplots of dynamics of tumour heterogeneity in the stickbreaking fitness model. Boxplots summarize data collected at the indicated milestone sizes from all simulations generated. **a,** distribution of RGS when all clones are measured. **b**, distribution of the detectable number of clones detectable above a 10% CCF sequencing cut-off. **c**, distribution of the detectable driver accumulation with the same 10% CCF sequencing cut-off.

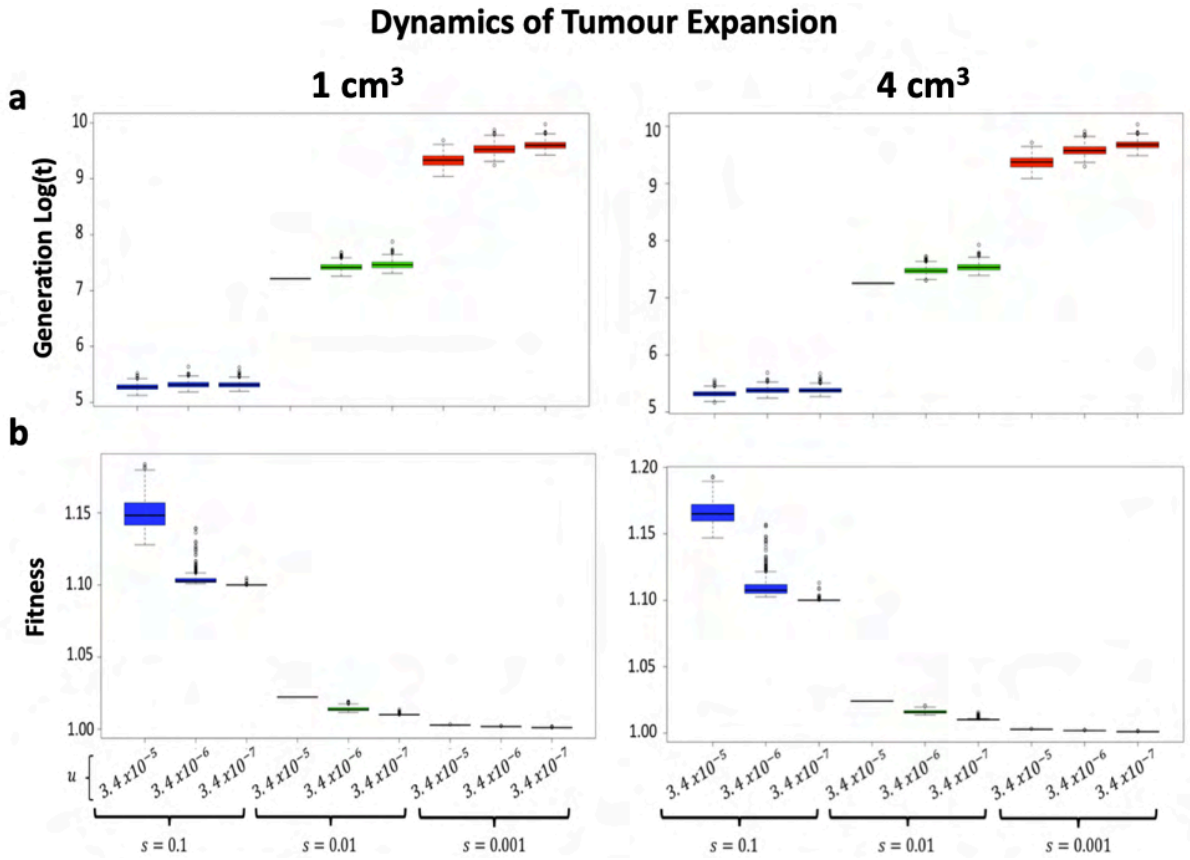


Figure S3.2 Boxplots of dynamics of tumour expansion in the increased mutation rate model. Boxplots summarize data collected at the indicated milestone sizes from all simulations generated. **a,** distribution from all simulations of the number of generations required achieved the milestone tumour size. **b**, fitness distributions of all simulations.

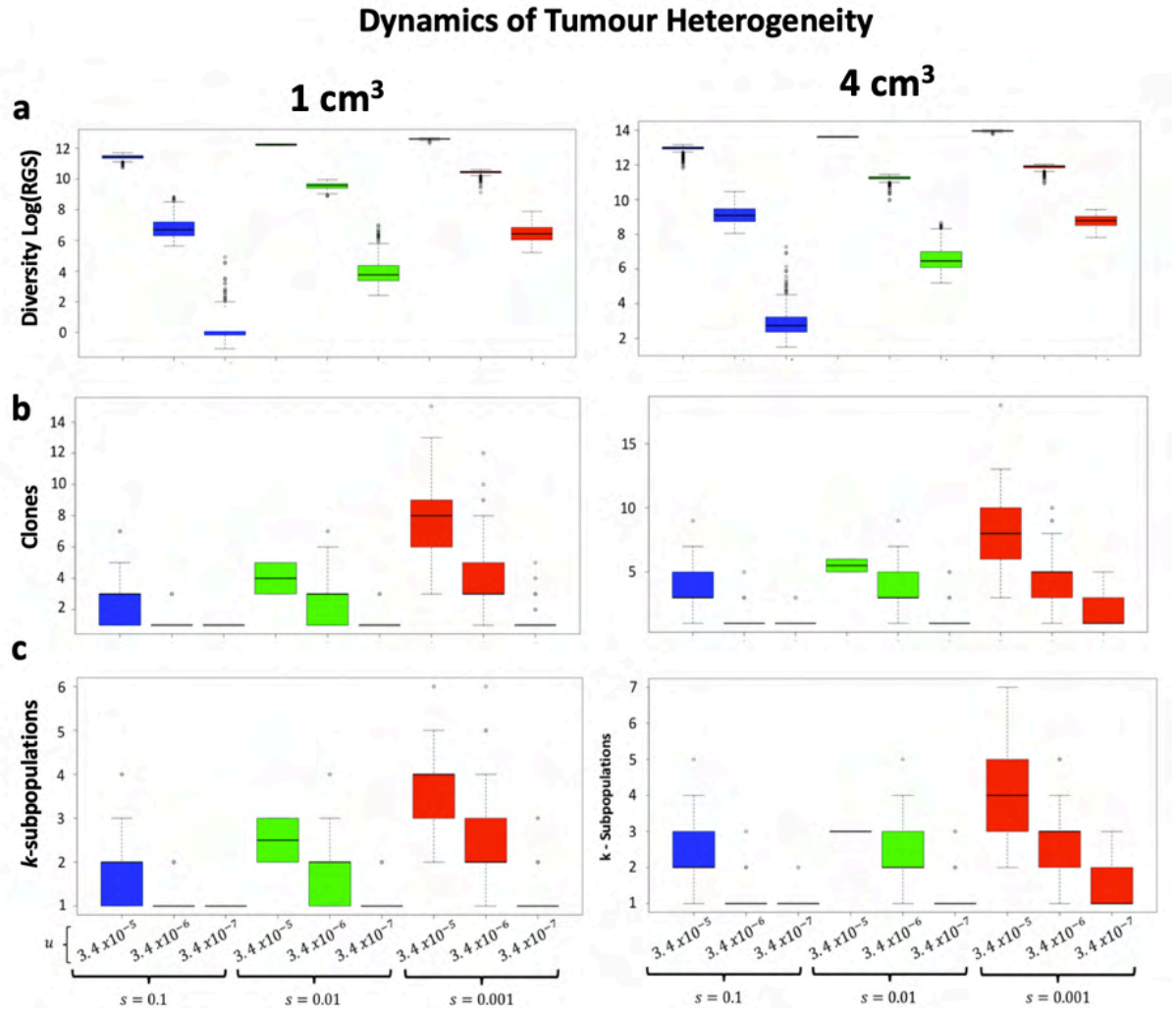


Figure S3.3 Boxplots of dynamics of tumour expansion in the increased mutation rate model. Boxplots summarize data collected at the indicated milestone sizes from all simulations generated. **a,** distribution of RGS when all clones are measured. **b**, distribution of the detectable number of clones using 10% CCF sequencing cut-off. **c**, distribution of the detectable driver accumulation using 10% CCF sequencing cut-off.

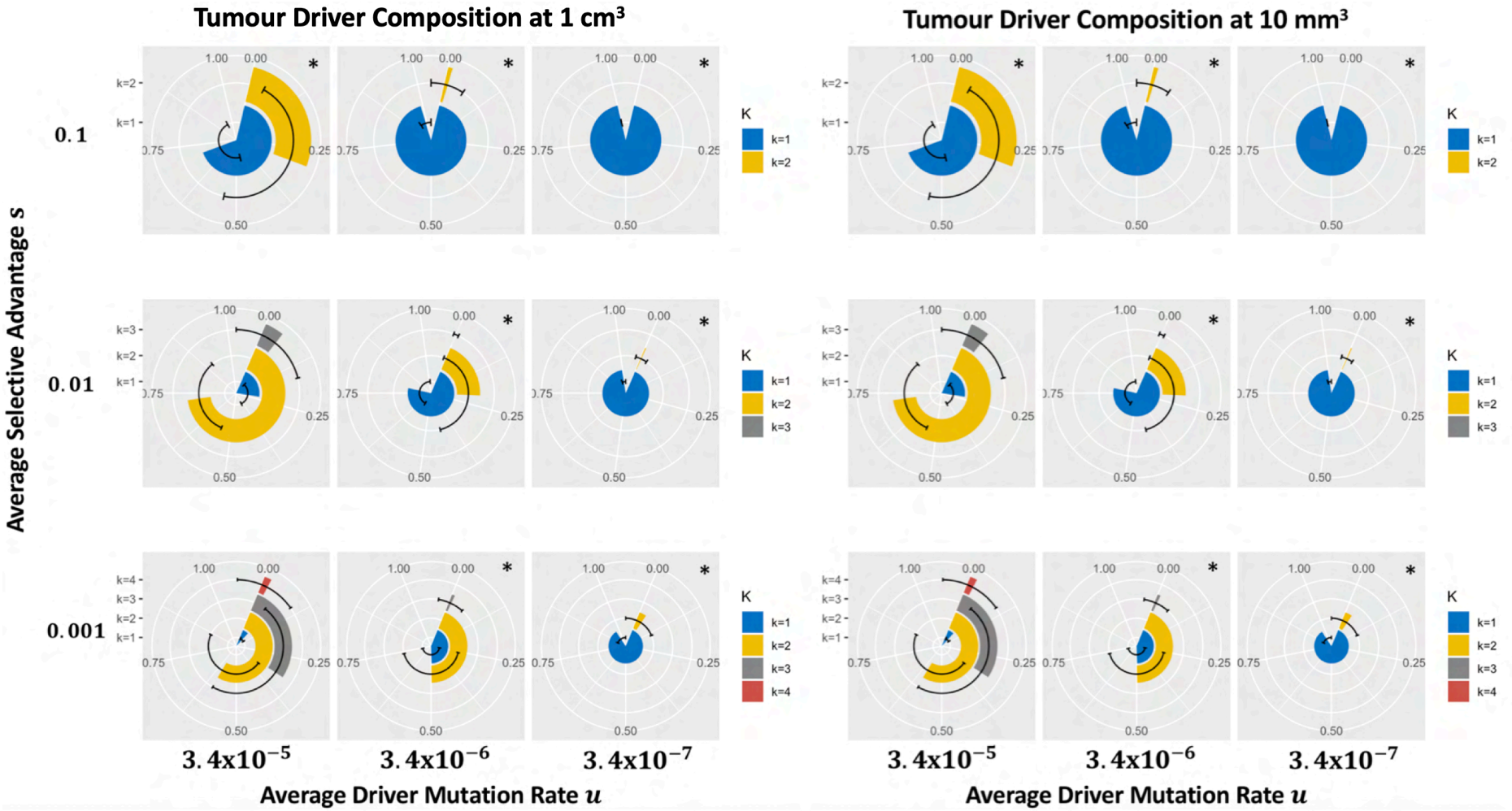


Figure S3.4 k-driver measurable composition of tumours at milestone sizes in the increased mutation model. Pie charts with error bars indicating the k-driver tumour composition colour coded by driver subpopulation *k*, numbers represent the proportion of the tumour in scale of 0.0 to 1.0. The pie charts tagged with (*) showed that most likely 1-2 driver subpopulations can be detected.

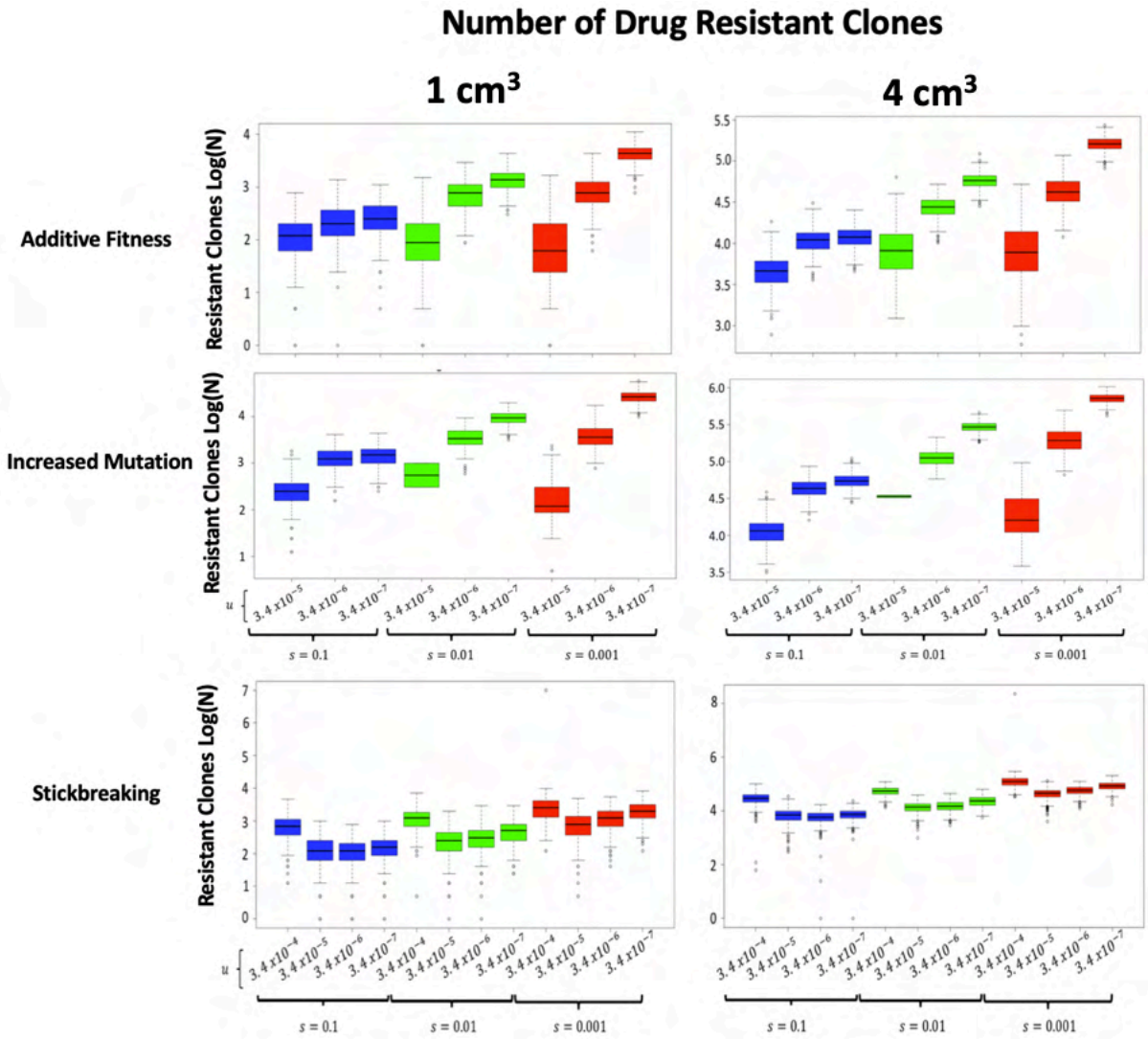


Figure S3.5 Total number of drug resistant clones. Boxplots summarize data collected at the indicated at milestone sizes considering all simulations in the positive selection models colour coded by average selective advantage *s*.

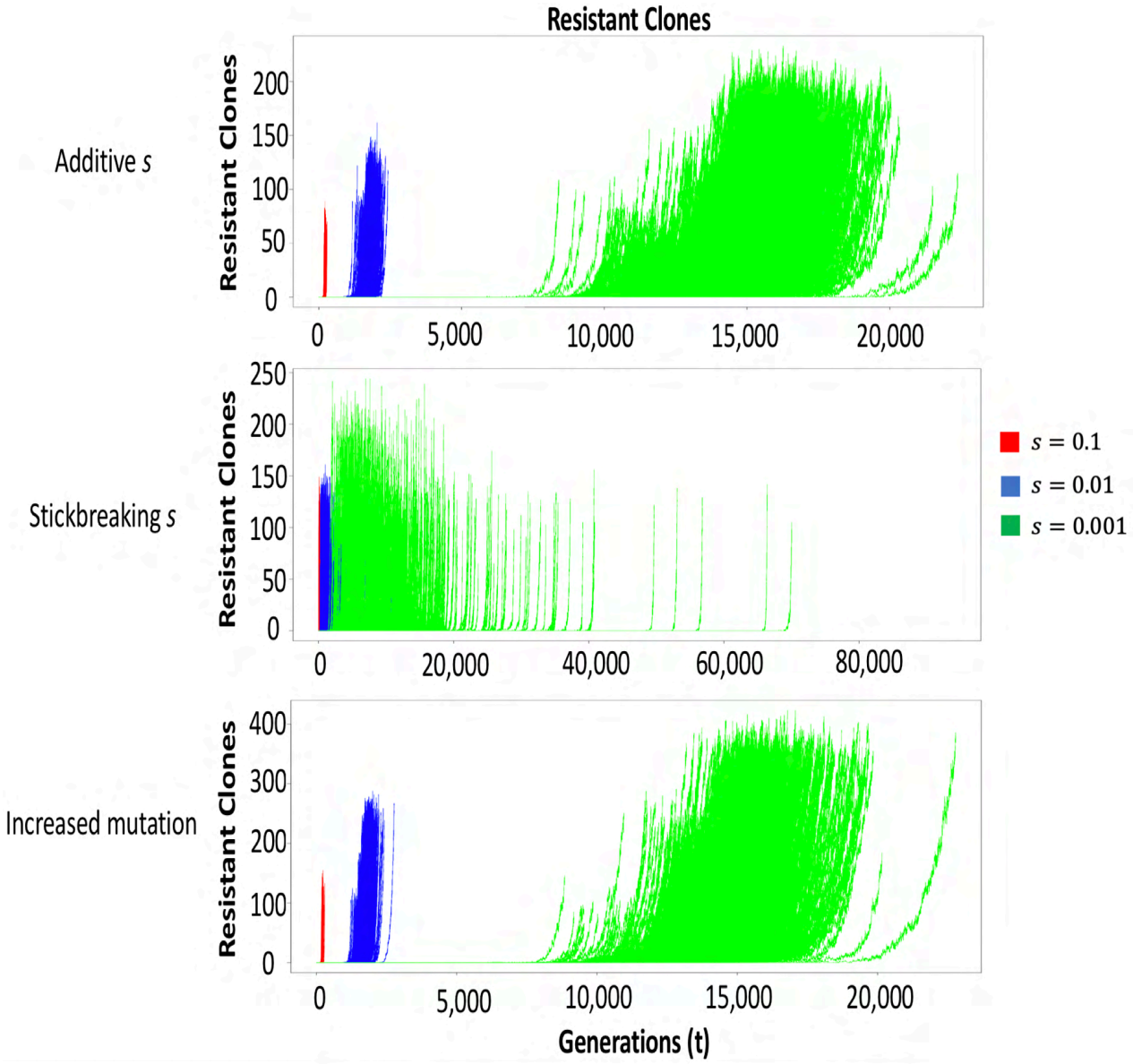


Figure S3.6 Drug resistance clones over time. Time series is generated by recording the variable of interest at every time t for all positive selection model.

Supplementary Table 3.1 Analytical vs Simulations

s	u	Simulations Range *k* at 1 cm^3	Analytical Range *k* at 1 cm^3	Simulations Range *k* at 4 cm^3	Analytical Range *k* at 4 cm^3
0.1	$3.4x10^{-5}$	2-3	1-3	2-3	1-3
0.01	$3.4x10^{-5}$	3-4	1-4	3-4	1-4
0.001	$3.4x10^{-5}$	3-5	2-6	3-5	3-6

Analytical results are obtained from Figure 2.10
Simulation results are obtained from Figure 3.4

Supplementary Table 3.2 Analytical vs Simulations with 10% CCF cut-off

s	u	Simulations Range *k* at 1 cm^3	Analytical Range *k* at 1 cm^3	Simulations Range *k* at 4 cm^3	Analytical Range *k* at 4 cm^3
0.1	$3.4x10^{-5}$	1-2	1-2	1-2	1-2 [2]
0.01	$3.4x10^{-5}$	1-3	1-2	1-3	2-3 [2-3]
0.001	$3.4x10^{-5}$	2-4	2-3	2-4	2-3 [3-4]
0.1	$3.4x10^{-6}$	1-2	1	1-2	1 [1]
0.01	$3.4x10^{-6}$	1-2	1-2	1-2	1-2 [1-2]
0.001	$3.4x10^{-6}$	1-2	1-2	1-2	1-2 [2-3]
0.1	$3.4x10^{-7}$	1	1	1	1
0.01	$3.4x10^{-7}$	1-2	1	1-2	1
0.001	$3.4x10^{-7}$	1-2	1	1-2	1

Analytical results are obtained from Figure 2.11 and values inside brackets from Table 2.1
Simulation results are obtained from Figure 3.4

Supplementary Table 3.3 Median Number of Passengers at Different Tumour Sizes with 10% CCF Frequency Resolution

Parameter/Cells	*500*	*1,000*	*5,000*	*10,000*	*50,000*	*100,000*	*500,000*	*1 M*	*5 M*	*10 M*	*50 M*	*100 M*
$s = 0.1$	c	c	c	c	c	c	c	c	c	c	c	c
$s = 0.2$	c	c	c	c	c	c	c	c	c	c	c	c
$s = 0.3$	c	c	c	c	c	c	c	c	c	c	c	c
$s = 0.4$	c	c	c	c	c	c	c	c	c	c	c	c
$s = 0.01$	2	2	2	2	2.5	2	2	2.5	2	2.5	2	2
$s = 0.02$	2	2	2	2	2	2	2	2	2	2	2	2
$s = 0.03$	c	c	c	c	c	c	c	c	c	c	c	c
$s = 0.04$	c	c	c	c	c	c	c	c	c	c	c	c
$s = 0.001$	2	3	4	4	4	3.5	3	3	2	2	2	2
$s = 0.002$	2	3	3	3	3.5	3.5	3	3	3	3	2.5	2.5
$s = 0.003$	2	3	3	3.5	4	4	3	4	3	3	3	3.5
$s = 0.004$	2	3	3	3	3	3	3	3	3.5	3.5	3	3

c only 1 passenger detected or clonal.

Supplementary Table 3.4 Median Number of Passengers at Different Tumour Sizes with 5% CCF Frequency Resolution

Parameter/Cells	*500*	*1,000*	*5,000*	*10,000*	*50,000*	*100,000*	*500,000*	*1 M*	*5 M*	*10 M*	*50 M*	*100 M*
$s = 0.1$	c	c	c	c	c	c	c	c	c	c	c	c
$s = 0.2$	c	c	c	c	c	c	c	c	c	c	c	c
$s = 0.3$	c	c	c	c	c	c	c	c	c	c	c	c
$s = 0.4$	c	c	c	c	c	c	c	c	c	c	c	c
$s = 0.01$	2.5	3	5	4	5.5	5	5	5	5	5	5	5
$s = 0.02$	2	2	3	3	3	3	3	3	3	3	3	3
$s = 0.03$	c	2	1.5	2	c	c	c	c	c	c	c	c
$s = 0.04$	2	2	2	2	2	2	2	2	2	2	2	2
$s = 0.001$	3	4	7.5	8	8	7.5	7	6	5	5	4	3
$s = 0.002$	3	4	6	6	6	6	6.5	7	6	6	6	6
$s = 0.003$	3	4	5.5	6	6	6.5	6	6.5	6	6	6.5	7
$s = 0.004$	3	4	5.5	6	6	6	6	6	7	7	7	7

c only one passenger detected or clonal.

Supplementary Table 3.5 Median Number of Passengers at Different Tumour Sizes with 1% CCF Frequency Resolution

Parameter/Cells	*500*	*1,000*	*5,000*	*10,000*	*50,000*	*100,000*	*500,000*	*1 M*	*5 M*	*10 M*	*50 M*	*100 M*
$s = 0.1$	3.5	4	5	6	6	6	6	6	6	6	6	6
$s = 0.2$	3	2	3	3	3	3	3	3	3	3	3	3
$s = 0.3$	2.5	2	2	2	2	2	2	2	2	2	2	2
$s = 0.4$	2	2	2	2	2	2	2	2	2	2	2	2
$s = 0.01$	6	8	16	17.5	21.5	23.5	24	24.5	24	23.5	24	24.5
$s = 0.02$	5	7	12	13.5	15	15	17	27	17	17	17.5	17.5
$s = 0.03$	4	6	9.5	11	11.5	11.5	12	11.5	12.5	12.5	11.5	12
$s = 0.04$	4	5.5	8	7.5	9.5	9.5	9	9	9.5	10	9.5	9.5
$s = 0.001$	6	9	23.5	29	37	39.5	42	40	41	40.5	38	39.5
$s = 0.002$	6.5	10	20.5	25	35.5	35	37	38	38.5	38	37	38
$s = 0.003$	6	10	20	24.5	33.5	34.5	36.5	37	36	36	36.5	37
$s = 0.004$	7	10	21.5	24	30.5	30.5	35	35	34.5	34.5	33	34

Chapter IV

1 Outline

The main goal of this chapter is to reconstruct tumour evolution in multiple bulk sequencing studies using the three discrete branching process models described and implemented in Chapter III.

- Additive fitness model: the average selective advantage s increases by a constant value with introduction of every new driver mutation.
- Stickbreaking: s changes by a predefined fitness distribution subject to the parental fitness but the average driver mutation rate u remains unchanged.
- Increased mutation rate model: the average selective advantage s is the same as the additive fitness model and the average driver mutation rate u can increase in 2-driver mutants with probability of 0.75.

The models explore different hypotheses of tumour development, considering scenarios in which the average selective advantage s and driver mutation rate u can deviate from the default model (additive fitness), allowing for additional variation to better explain observed patterns in tumour sequencing data.

To find an effective method for comparison of simulation and sequencing data I evaluated multiple statistical methods including distribution-free, goodness-of-fit and minimum distance approaches. To replicate the technical and biological biases of measuring a mixture of clonal genomes in bulk sequencing, I benchmarked the statistical methods considering different magnitudes of noise contamination, missing information and a combination of both.

From all the methods compared, minimum distance outperformed the rest in all categories. The metric I designated as minimum distance A corrected for false discovery rate was subsequentially used for reconstructing tumour evolution.

I used 4 studies with a total of 1,800 cases to test the minimum distance method. The studies encompass different cancer subtypes with different bulk sequencing assays. Excluding CASCADE, all of them have clinical annotations that served as quality control and survival data that allowed association of fitness and RGS diversity with clinical outcome. The following is a summary of the studies:

- A pan-cancer study of whole exome sequencing from 1,165 TCGA patients by Andor et al. [29] containing primary and metastatic samples.
- TRACERx Non-small-cell lung (NSCLC) [76], a prospective study of 96 patients with multi region ultra-deep whole exome sequencing of primary samples of non-small cell lung cancer.
- Breast International Group 1-98 (BIG) [195], a retrospective study with amplicon sequencing of 500 genes from primary samples of 538 post-menopausal woman with hormone receptor positive HER2-negative breast cancer.
- CASCADE melanoma, which tracked disease progression from primary formation until death in in an individual with melanoma, through multi region whole exome sequencing of primary samples and metastases.

Overall, the minimum distance method gave results that aligned well with genomic and clinical data, providing biological insights and demonstrating its power to reconstruct tumour evolution, even in challenging cases with limited amplicon sequencing. It provides an evolutionary framework to complement data-driven clonality tools such as PyClone and ExPANdS by recovering clonal histories and tumour growth dynamics.

The main outcomes in reconstructing tumour evolution with the minimum distance methods are as follows:

First, it gave biological insights on how clonal distributions and phylogenies are associated with overall survival in TCGA.

Second, it approximated the phylogenies in TRACERx NSCLC. Although clinical significance was not found with predicted fits, it showed a trend in which increased fitness and increased RGS diversity can potentially associate with recurrence or death, suggesting 18% of cases had a *Big-Bang* evolutionary mode.

Third, the method showed how mutational profile is significantly associated with a mechanism of distant recurrence as a function of fitness and genomic instability in BIG 1-98.

Finally, it was able to identify patterns of dissemination in the CASCADE melanoma patient by simulation convergence in primary and metastatic sites.

As a result, the discrete time branching process with the minimum distance method can be used as an alternative tool for clonal evolution reconstruction in different bulk sequencing assays that can provide unique biological insights into tumour growth dynamics.

2 Introduction

Reconstructing tumour evolution remains a significant challenge due to the complexity of measuring a mixture of clonal genomes at the molecular level. Efforts to tackle this problem can be grouped in three categories:

1. Data-driven models. These are based on ploidy correction and aim to estimate the number of clonal expansions from sequencing samples. Commonly used algorithms are PyClone, SciClone and ExPANdS. Limitations include sensitivity to biases introduced by signatures of neutral evolution in samples and LOH with genome doublings that can hamper ploidy correction leading to inaccurate predictions.
2. Data-driven models with evolutionary frameworks. Besides ploidy correction, this approach applies an evolutionary framework to establish clonal ancestry, such as in the commonly used tools PhyloWGS and Canopy. The limitation in this approach is the requirement of whole exome/genome multi-region sequencing assays that represent a financial burden for large-cohort and longitudinal studies.
3. Evolutionary framework driven models. This is an emerging area in clonal evolution reconstruction in which the use of population genetics models accounting for fitness and mutation rate aim to describe the distribution of variant allele frequencies or clonal proportions. Only two algorithms are published, SubClonalSelection and MOBSTER.

It is essential to reconstruct tumour evolution prospectively to measure how treatment affects clonal evolution trajectories, to monitor disease progression and to associate clonal

architectures with clinical outcome. However, there is no all-in-one tool to pair clonal reconstruction with tumour growth dynamics. Therefore, the aim of this Chapter is to bridge-this gap by evaluating the potential of the discrete time branching process to perform this task.

The approached that I designed falls into the third category of the models listed above, with the evolutionary framework being the discrete time branching process which recovers clonal ancestry with tumour growth dynamics. I implemented 3 models, as described in Chapter III, using as the default the additive fitness model because it does not change the average driver mutation rate *u,* and the average selective advantage *s* changes proportional to accumulation of *k.* This leads to restricted tumour fitness configurations by the controlled increase in fitness.

Aiming to relax this effect, the stickbreaking model provides more flexibility as changes in fitness are sampled from a distribution bounded by parental fitness leading to a better sampling of the fitness landscape. However, the stickbreaking model does not change the average driver mutation rate, and to explore its role on predicting tumour evolution, the increased mutation rate model was designed. The model increases the mutation rate by a factor of $0.5u$ starting from 2-driver lineages that will propagate the increased mutation rate to their progeny.

With the three models implemented covering multiple aspects of tumour evolution, and a database generated with numerous instances of the branching process models. A robust statistical method is required to compare cancer cell fractions from simulated data against those observed in patient sequence data.

The initial sections of this chapter will address this, with goodness-of-fit and minimum distance methods compared and benchmarked. The comparison will consider scenarios that replicate biological variability and technical biases to measure clonal subpopulations from bulk sequencing. The second half of the chapter elaborates on the quality and biological insights provided by applying the best performing minimum distance fitting procedure to data from cancer sequencing studies.

I will discuss the strengths and limitation of the designed approach and recommend guidelines for its use on other datasets. I will elaborate on how clinical and genomic variables can be used as quality control measures, and the contribution of the evolutionary variables to further understanding of the dynamics of tumour evolution in different malignancies.

3 Hypothesis and Aims

Hypothesis: Incorporating evolutionary modelling into data-driven models can refine tumour evolution reconstruction. It is possible to approximate the average selective advantage *s* and average driver mutation *u* by comparing simulated to real cancer cell fractions to approximate the evolutionary history of real tumours.

Aim I: Explore multiple fitting statistics that allows the comparison of observed against simulated cancer cell fractions and benchmark these under different conditions that replicate the biological and technical biases involved in measuring tumour diversity at the molecular level.

Aim II: Having identified the optimal statistical method, apply it to different studies covering different cancer subtypes, mutational profiles and sequencing assays.

Aim III: Evaluate the quality of the fits by comparing to clinical and genomic variables, and establish the power of the designed approach to provide biological insights into clinical outcomes and tumour growth dynamics.

4 Methods

4.1 Methods: Fitting Procedure and Benchmarking

The goal of the fitting procedure is to compare observed (patient) vs simulated (branching process models) cancer cell fractions (CCFs). Simulated CCFs were evaluated in all simulations using equation 1.1 on their clonal frequencies, and only the top-100 CCFs per simulation were used to compare with patient data (10, 5 & 1% cut-offs were used).

Distribution-free methods such as the the Cramér-von Mises criterion (CvM) and the Kullback-Leibler (KL) divergence were compared with the Euclidean distance. For the CvM and KL methods, the empirical density was evaluated due to the low number of clones at the 10% CCF cut-off. The density implementation was done by the *density* function in R and then probability density was evaluated using the *approxfun* (R) in the range of [0,1] with increases of 0.01 with default parameters. The Cramér-von Mises criterion was implemented as the ranked method instead of evaluating the integral of the cumulative density function, as shown in Appendix A.4.1. The Kullback-Leibler divergence was implemented using the KLD function in R; the statistic was the average of $(KLD[p(y)||p(x)] + KLD[p(x)||p(y)])/2$.

The minimum distance method is described in section 5.2 and compares all-vs-all (i.e. observed vs simulated) CCFs without requiring a density estimation. In the cases where observed and simulated CCFs have different number of clones, the direction of minimisation is relevant leading to 2 different metrics, MDA and MDB. Details of the Euclidean distance implementation are described in section 5.2

The benchmarking was performed using the stickbreaking model, selecting all simulations at the 4 cm^3 tumour size to enrich the CCF signal as much as possible. To match the sequencing resolution, simulated clones below the 10% CCF were removed. This was considered the truth set with a total of ~6,000 tumours. Every tumour in the truth set was compared to all the 120,000 snapshots available in the stickbreaking model comparing the 11 statistical methods (details in section 5.3). The 120,000 snapshots were ranked by their statistical score, and in some cases further corrected for false discovery. False discovery was implemented by minimising adjacent snapshots of the best fit, e.g. if the best fit is a 2 cm^3, the final score is by averaging snapshots 1.5, 2 and 2.5 cm^3. As a result, the top 10 scoring parameters were evaluated for accuracy to determine the best method to compare with patient data. This was repeated for 10 experiments that aimed to replicate the sources of variation in measuring clonal populations, which included different magnitudes of noise contamination, missing information and a combination of both.

Benchmarking plots in identifying the starting values of s and u considered if the predicted parameter is in the top 10 scoring statistics. Subject to the accurate guess of s and u, the accuracy of the correct simulation defined if the predicted parameters contain the correct simulation replicate ID (1- 500).

4.2 Methods: Fitting TCGA Patients

CCFs extracted from data published by Andor et al. [29], who applied ExPANdS and PyClone to estimate clonality from TCGA data from diverse cancer types excluding breaast (BLCA, CESC, GBM, HNSC, KIRC, LGG, LUAD, LUSC, PRAD, SKCM, STAD & THCA), were fit to CCFs from the simulated tumours described in Chapter III. Details are provided in Table 4.1.

Fits were performed using the MDA.N method (details in section 5.2 and 5.3). Comparison of simulated and patient data was performed by filtering the simulated clones below 10% of CCF. A heatmap of predicted fits was generated by the consensus of the best scoring statistic in all the subtypes with both clonality callers (PyClone and ExPANdS) for all branching process models. Distributions of the predicted number of clones, RGS and fitness used the best scoring simulation per cancer subtype.

Recurrent phylogenies in TCGA were generated by using the best scoring simulations of the stickbreaking process (preferred method) for every subtype and identifying recurrent topologies. Percentages displayed are relative to the subtypes. Top recurrent simulations plots were generated by using the best scoring simulations for every branching process independently, and identified the frequency of simulations with the same s, u and iteration ID. Percentages displayed are relative to the subtypes and total of the cohort

4.3 Methods: Fitting TRACERx NSCLC Patients

CCFs from simulated tumours were fit to clonality estimates produced by Turajlic et al.[76] who applied PyClone to multi-region whole exome data from 100 NSCLC tumours (deep-sequencing, medina depth ~426x). Details provided in Table 4.1.

Fits were performed using the MDA.N method. Comparison of simulated and patient data was performed by filtering the simulated clones below 1% of CCF. A global CCF was obtained by averaging variant CCF if it was present in multiple sites. A heatmap of predicted fits was generated by the consensus of the best scoring statistic for all branching process models using the PyClone calls reported in the study. Distributions of the predicted number of clones, RGS and fitness used the best scoring simulation.

Recurrent phylogenies in TRACERx NSCLC were generated by using the best scoring simulations considering all models independently. Top recurrent simulation plots were generated by using the best scoring simulations for every branching process independently, and identified the frequency of simulations with the same s, u and iteration ID. Percentages are relative to the total of the cohort.

The 4 clusters representative of TRACERx NSCLC were generated manually by visual selection. Once the categories were defined, the best scoring simulation was selected for all phylogenies of the category and a visual selection of the best approaching phylogeny from the 3 positive selection models was chosen.

Cox proportional hazards model were stratified by adjuvant therapy using the survival package of R 3.6.2.

4.4 Methods: Fitting BIG 1-98 Patients

The evaluation of the most frequent cytoband alterations in BIG 1-98 [172] and METABRIC [168] was performed by sub setting the cohort on their status responders vs distant recurrences (BIG 1-98) or deaths (METABRIC). SNV alterations from BIG 1-98 came from Foundation Medicine and in the reported calls from METABRIC. Cytoband alterations were estimated with the weighted Genome Integrity Index [196] per cytoband.

Mutational calling was performed by VarDict [197] with VEP annotation to corroborate Foundation Medicine [198] calls [199]. High quality calls were used and variants were filtered by ExAC frequency greater than 0.01% and Condel score. Validation of the driver effect of the mutations was performed by using the maximum likelihood dN/dS method [70] (performed with the tool *dndscv* available in R 3.6.2).

Copy number calls were performed using CopywriteR [200]. Calls were winsorized before binary segmentation [201, 202]. Only the autosomal calls were used.

wGII heatmaps were generated with all the cytoband calls. Hierarchical clusters were configured with Manhattan distance and Ward.D2 agglomeration method. Kaplan Meier used the clusters generated by the heatmaps.

Weighted univariate and multivariate Cox proportional hazards models were evaluated in BIG 1-98 stratifying by treatment arm and luminal status. In METABRIC, univariate and multivariate Cox proportional hazards models were stratified by luminal status and if the patients received chemotherapy or not. wGII was averaged for every patient.

The BIG 1-98 study [172, 199] sequenced 538 patients with 140 distant recurrence events in a panel of 287 driver-genes, the sequencing was performed by Foundation Medicine on formalin fixed paraffin embedded samples T5 targeted panel (variants with median above 150x were used) [198].

Clonality calling in ExPANdS was performed with the default values and a significant filter of 0.7. PyClone was run with the recommended values when no matched normal information is available. CCFs were adjusted to represent the fraction in the tumour as $CCF_i/max(CCF)$. Mutational clusters with low number of mutations < 5 were removed.

Single clone evaluation (neutral samples) was performed with the *neutralitytesr* tool [103] in R 3.6.2 and the maximum likelihood dN/dS method [70].

Fits were performed using the MDA.N method. Comparison of simulated and patient data was performed by filtering the simulated clones below 5% of CCF. A heatmap of predicted fits was generated by the consensus of the best scoring statistic in all the subtypes with both clonality callers (PyClone and ExPANdS) for all branching process models. Distributions of the predicted number of clones, RGS and fitness used the best scoring simulation.

Fitting the passenger tail was used with the Bozic et al. [58] median number of passenger alterations solution that were fitted to the passenger tail using least squares technique.

Recurrent phylogenies in BIG 1-98 were generated by using the best scoring simulations considering all models independently. Top recurrent simulations plots were generated by using

the best scoring simulations for every branching process independently, and identified the frequency of simulations with the same s, u and iteration ID. Percentages are relative to the total of the cohort.

Weighted Cox proportional hazards were stratified by mutational profile (No TP53 and PIKCA, TP53, PIK3CKA and PT53+PIK3CA) treatment arm and luminal status.

4.5 Methods: Fitting CASCADE Melanoma

One patient from the cohort was used to evaluate the method longitudinally using primary, circulating DNA and metastases samples [203]. 3 multi-region samples from the primary and 5 metastases were evaluated from whole exome sequencing with a mean depth of ~132x. Mutational calling was performed by external collaborators who ascertained high quality confidence in the variant and copy number calling with an accuracy of 90% [203].

Clonality calling in ExPANdS was performed with the default values and a significant filter of 0.7. CCFs were adjusted to represent the fraction in the tumour as $CCF_i/max(CCF)$.

Fits were performed using the MDA.N method. Comparison of simulated and patient data was performed by filtering the simulated clones below 10% of CCF. Heatmaps of predicted fits were generated by the consensus of the best scoring statistic for all branching process models using the ExPANdS calls for every sample individually. Distributions of the predicted number of clones, RGS and fitness used the best scoring simulation.

Recurrent phylogenies in the CASCADE melanoma were generated by using the best scoring simulations considering all models independently for primary and metastases. Top recurrent simulations plots were generated by using the best scoring simulations for every branching process independently, and identified the frequency of simulations with the same s, u and iteration ID in primary and metastases samples.

ExPANdS primary and metastases clustering was performed in all samples together with default values and quality filtering at 0.7. The dendrogram plot was configured to only show the dominant clone for better depiction.

5 Statistical Methods for Comparing Simulated vs Real Cancer Cell Fractions

Statistical fitting to cancer cell fraction data is a complex problem due to the different sources of bias such as sequencing depth, noise, breadth and depth of sequencing assays, number of samples, spatial heterogeneity, etc. In this section I tackle this problem by evaluating multiple statistical approaches to compare simulated cancer cell fractions originating from the discrete time branching process against those measured from sequenced tumour samples.

Because cancer cell fractions are influenced by inheritance, they are not related to clonal size (equation 1.1, unless $s = 0.1$), and therefore quantification of tumour evolutionary parameters has to be done through estimation the of initial conditions for the average selective advantage s and average driver mutation u.

The main difficulty in determining the initial conditions is that the simulated distribution from the branching process $F(\cdot;\theta)$ does not have an explicit form. Thus, estimated parameters $\hat{\theta}$

cannot be easily obtained by derivative-based optimization methods, such as Fisher scoring approaches.

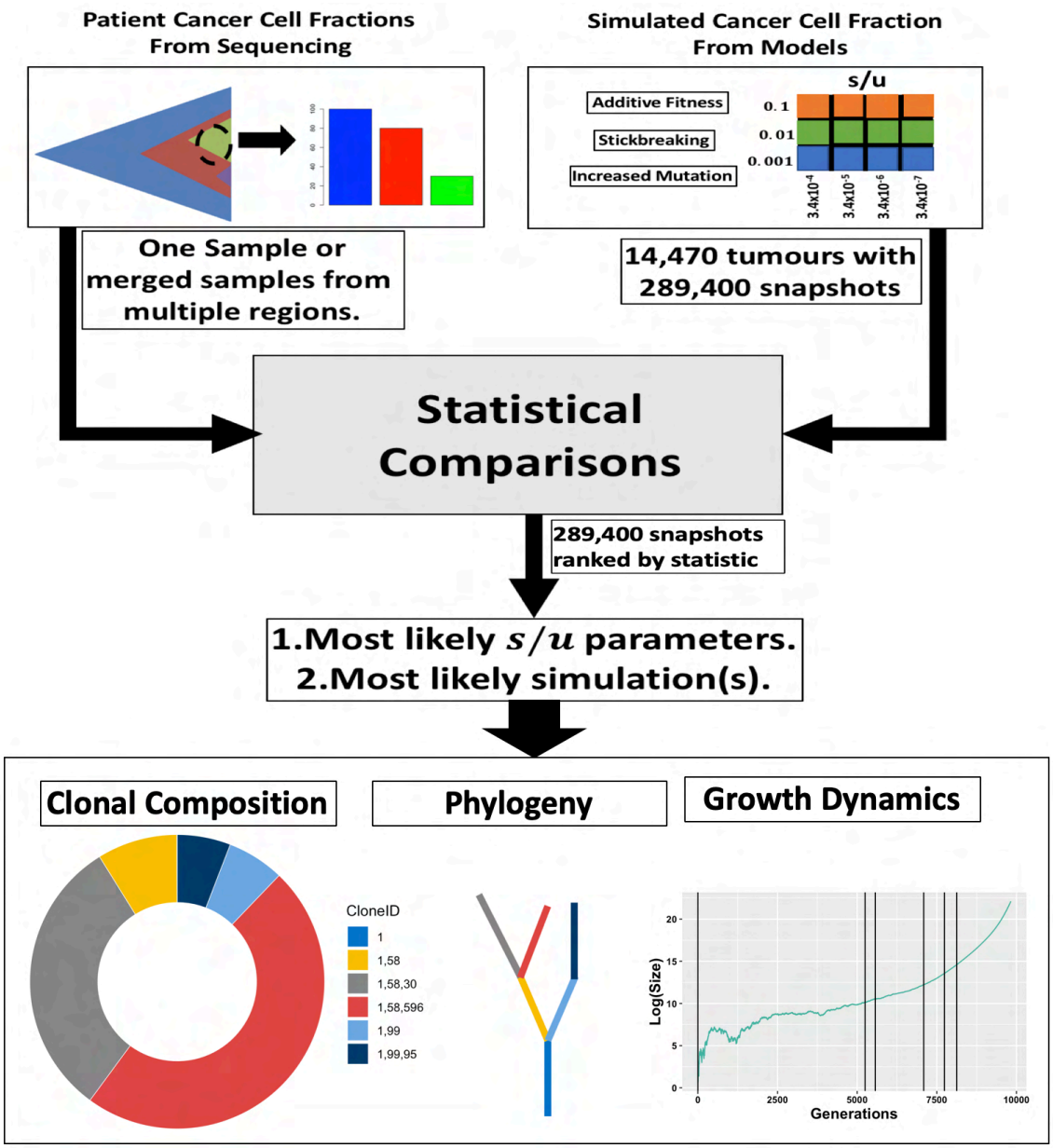


Figure 4.1 Framework for comparing of observed cancer cell fractions to simulated data. Cancer cell fractions from ploidy corrected sequenced tumours are compared to all simulated snapshots. Comparisons to the snapshots will provide a score that is corrected for false discovery then used to reconstruct the clonal evolution of the sequenced sample. The clonal evolution reconstruction will recover the clonal and driver compositions in the tumour and its phylogeny, growth dynamics and number of drug resistant clones.

The approach consists of selecting the best fitting simulation, and using the simulation properties to describe how clonal evolution proceeded in the sequenced sample.

The chosen statistical methods described here were used to compare every simulated snapshot of the positive selection models $F(\cdot;\theta)$, to the observed data $F_\theta(x)$ to approximate the parameters of interest θ, as shown in Figure 4.1. Based on all fits the best scoring statistic is going to be used to recover tumour phylogeny and reconstruct clonal growth dynamics. In a

nutshell, to the main aim here is to minimise a statistic based on a selection criterion and use the top scoring model to describe the properties of the tumour.

However, in Chapter III I showed that at the default sequencing cut off 10% (assuming rep-seq) comparisons with real cancer cell fractions (CCFs) can be assigned to one of two clusters. The first cluster displays strong fitness and/or low mutation rates ($s = 0.1$ & $u = 3.4\text{x}10^{-7}, 3.4\text{x}10^{-6}$) while the second cluster has moderate to weak fitness and high mutation rates ($s = 0.001, 0.01$ & $u = 3.4\text{x}10^{-5}, 3.4\text{x}10^{-4}$). This effect is going to be accounted in the fitting by limiting comparisons to only those CCFs above 5% or above 10%.

5.1 Distribution-Free Goodness-of-Fit Statistics

The reason for using distribution-free statistical methods is to account for the acute skewness of cellular cell fractions that is introduced by inheritance. Chowell et al. [44] showed heavy tailed distributions result from multiple parameter combinations using the stickbreaking model. Therefore, symmetric goodness-of-fit statistics such as the t-test were excluded from this analysis.

Multiple distribution free methods were considered but only the Cramér-von Mises criterion (CvM) and the Kullback-Leibler divergence (KL) were carried forward for testing. The 2-way CvM is a robust goodness-of-fit statistic that compares whether two observations are equal or not using their empirical distribution densities. It can be computed by comparing the ranks of two observations as suggested by [204]. The advantages of the CvM include good sensitivity in mixture distributions, robustness to noise and better power than the Kolmogorov-Smirnov test [205, 206]. The KL divergence, akin to the Bhattacharyya distance, evaluates the similarity of two probability densities. It can be viewed as a minimum distance in probability space.

Distribution-free methods require a minimum sample size greater than the number of CCFs typically detectable at the traditional sequencing cut-off 10%. For that reason, the empirical distribution function was used from the raw cancer cell fractions as $F(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$. Details of both models and their implementation are provided in Appendix A.4.1.

5.2 Minimum Distance Metrics

Key limitations of distribution-free goodness-of-fit statistics for CCFs include reduced sample sizes and missingness of the data, which can lead to the tests being underpowered when the empirical density cannot provide enough information about the CCF distribution. The extent of such effects in sequencing data are determined by multiple factors, including tumour purity, the number of regions sampled, overall sequencing coverage and depth, etc. Thus, alternative methods of comparing cancer cell fractions using minimum Euclidean distance were explored. The advantage of minimum distance is that it does not requires density estimation and provides a simple, direct comparison between cancer cell fractions.

With two samples, simulated and observed CCFs are consider vectors of lengths *N* and *M*, such that $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_M$. These are arranged in a matrix to evaluate their Euclidean distance as following,

$$\boldsymbol{D} = \begin{bmatrix} \sqrt{(x_1 - y_1)^2} & \cdots & \sqrt{(x_1 - y_M)^2} \\ \vdots & \ddots & \vdots \\ \sqrt{(x_N - y_1)^2} & \cdots & \sqrt{(x_N - y_M)^2} \end{bmatrix}$$

$\boldsymbol{D}$ is a matrix of consisting in the combinatorial Euclidean distance of all X and Y pairs which enables the comparison between simulated and observed cancer cell fractions of unequal sizes. In the cases where X and Y are the same size, the diagonal elements of $\boldsymbol{D}$ are equal to zero, and minimising by row or column does not matter. In cases where X and Y have different sizes a choice has to be made, whether to minimise by rows or by columns depending on fitting to X or Y, $\bar{d}_r$ or $\bar{d}_c$ respectively.

Therefore, we define as minimum distance A (MDA) when the aim is to evaluate the overlap of $\bar{d}_r = X \cap Y$ and vice versa, minimum distance B (MDB) when the aim is to perform the complimentary comparison to evaluate the overlap of $\bar{d}_c = Y \cap X$.

5.3 Benchmarking Performance of Model Fitting Methods

Before applying the statistical methods introduced in Sections 5.1 and 5.2, their accuracy must be evaluated under multiple conditions to replicate the sources of variation affecting the measurement of clonal genomes. However, due to the lack of publicly available validated test-sets to replicate sequencing distributions, I used a sample of the positive selection models described in Chapter III. An outline of the benchmarking experimental design is shown in Figure 4.2 [74].

I selected all simulations at the 4 cm^3 tumour size using the stickbreaking model as a truth set (12 parameter combinations of s and u with 500 replicates each, leading to a total of 6,000 tumours), filtered out clones below the 10% CCF cut-off and removed cases that had only one clone above this threshold. Since the stickbreaking model has the largest range of parameter combinations and is considered the gold standard here, I considered its benchmarking results to be applicable to the additive fitness and increased mutation rate models as well.

For each of the 6,0000 tumours in the truth set, 120,000 comparisons were evaluated with each of the ten statistical methods (listed below). This was then repeated in the eleven experiments also described below. This represented a significant computational burden, therefore only the stickbreaking model was benchmarked.

For every tumour compared, the 120,000 simulations were ranked by their statistical scores and, for those requiring it, were further corrected for false discovery. As a result, the top ten scoring simulations were chosen for reporting its accuracy.

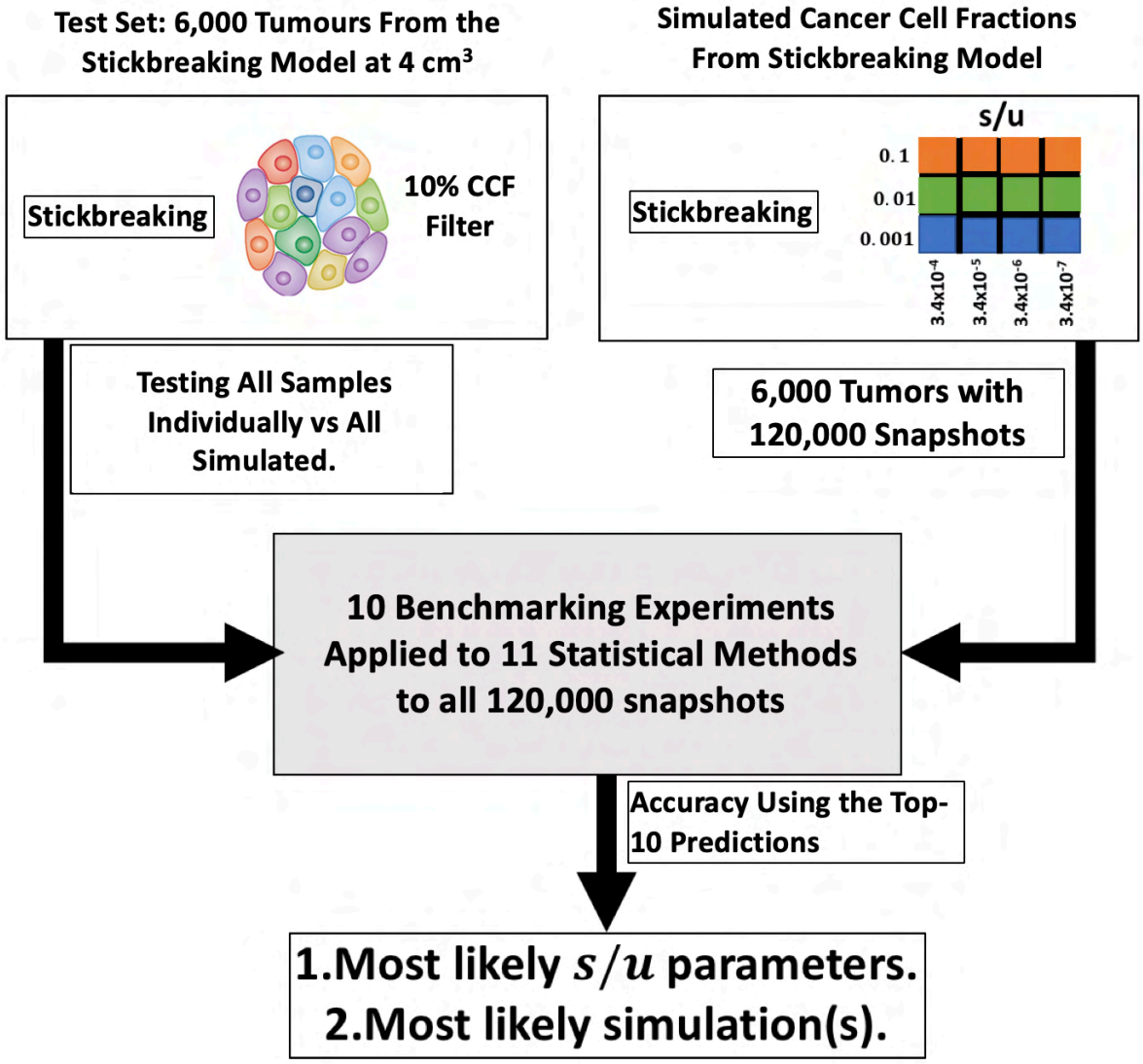


Figure 4.2 Benchmarking experimental design. Simulated snapshots at the 4 cm^3 size were filtered to consider only the clones that are above 10% CCF as a truth set. Truth set tumours were compared to all available simulated snapshots of the stickbreaking model (120,000) to benchmark the statistical methods.

To allow for flexibility in the model fitting, accuracy was estimated according to whether the true parameters of interest were found within the top 10 predictions. The same criterion was applied for the determining whether the correct simulation (correct parameter and replicate) was obtained. This is to provide multiple candidate fits, similar to widely used phylogenetic tools [113, 115, 118].

The main sources of variation considered for statistical comparison are the degree of noise, the missing cancer cell fractions due to sampling bias and a combination of both.

I designed the following experiments to evaluate how well the correct parameter set and simulations are recovered in the test set under the following conditions:

1. Unadjusted original dataset: models an ideal sequencing scenario with no noise, to evaluate how well the statistical method does compared to random guessing.

2. Added noise: to determine robustness of the methods in the presence of noise contamination.
    - Low: simulates a low degree of DNA degradation/artefacts.
    - Moderate: simulates a medium degree DNA degradation/artefact.
    - High: models a high degree of DNA degradation/artefacts.
3. Missing data (missingness): to determine the effect of incomplete cancer cell fraction distributions.
    - Removing one clone: simulates scenarios when almost all detectable heterogeneity was measured such as in multi region sequencing.
    - Removing two related clones by inheritance: simulates limited clonal information, such as in single region sequencing.
4. Missingness and noise: simulates more realistic scenarios in for measuring the detectable clonal landscape.
    - Removing one clone and low noise: simulates a multi-region with low DNA degradation/artefacts.
    - Removing one clone and high noise: simulates a multi-region with high DNA degradation/artefacts.
    - Removing two clones and low noise: simulates a single-region with low DNA degradation/artefacts.
    - Removing two clones and high noise: simulates a single-region with high DNA degradation/artefacts.

Noise contamination was introduced as a uniform process using the *jitter* function in R statistical software version 3.6.0. The *jitter* function adds noise as follows:

$$Z = uniform\left(N, \frac{-\alpha d}{5}, \frac{-\alpha d}{5}\right)$$

Where α is the degree of noise, set to $\alpha = \{0.01, 0.05 \ \& \ 0.1\}$ for low, moderate, and high noise contamination. The value of d represent the smallest difference between adjacent unique cancer cell fraction values which was set to the default option.

To explore the potential of combining methods I explored ensembles of equal weight using KL divergence combined with a minimum distance technique.

As many comparisons are performed the risk of false discovery is high, so to minimise this effect I averaged fits to three contiguous neighbour snapshots for consistency. As shown in Chapter III, detectable clonal composition does not change significantly over time. Thus, using the neighbour snapshots allows filtering of false positive cases and facilitates the minimisation task. I applied this strategy for all methods.

The following are the statistical methods evaluated,

1. Cramér-von Mises criterion (CvM).
2. Kullback-Leibler divergence (KL).
3. Minimum distance: experimental ∩ simulated $\bar{d}_r$ (MDA).
4. Minimum distance: simulated ∩ experimental $\bar{d}_c$ (MDB).
5. Kullback-Leibler divergence + minimum distance $\bar{d}_r$ (KL.MDA).
6. Kullback-Leibler divergence + minimum distance $\bar{d}_c$ (KL.MDB).
7. Cramér-von Mises criterion minimising neighbours (CvM.N).

8. Kullback-Leibler divergence minimising neighbours (KL.N).
9. Minimum distance: experimental ∩ simulated $\bar{d}_r$ minimising neighbours (MDA.N).
10. Minimum distance: simulated ∩ experimental $\bar{d}_c$ minimising neighbours (MDB.N).
11. Kullback-Leibler divergence + minimum distance simulated ∩ experimental $\bar{d}_c$ minimising neighbours (KL.MDB.N).

In summary, 10 experiments evaluating 11 statistical methods were applied to the 14,470 tumours at 4 cm^3 against all 289,400 simulations. The output is a ranked list of best-fit simulations allowing for the 'true' simulation to be in the top 10 predictions, with accuracy measured by how many cases were correct out of the total number evaluated.

5.4 Benchmarking Results

The first experiment, testing if the comparison techniques can recover a sample without missing information and noise, showed minimum error across all methods, as depicted in Figure S4.1. Although this scenario was used to check the correct implantation of the methods it also illustrates the performance under ideal conditions.

In the experiments with added noise, Figure 4.3, most of the methods achieved 80% accuracy in detecting the correct set of parameters. In the cases with low noise, most of the methods had an accuracy of ~70%. However, the accuracy dropped to ~50% when it came to finding the correct simulation in the moderate and high noise tests, middle and bottom panels of Figure 4.3.

Excluding MDB, the worst performing method, all statistical methods had similar accuracy. As a result, most of the statistical methods can handle noise contamination.

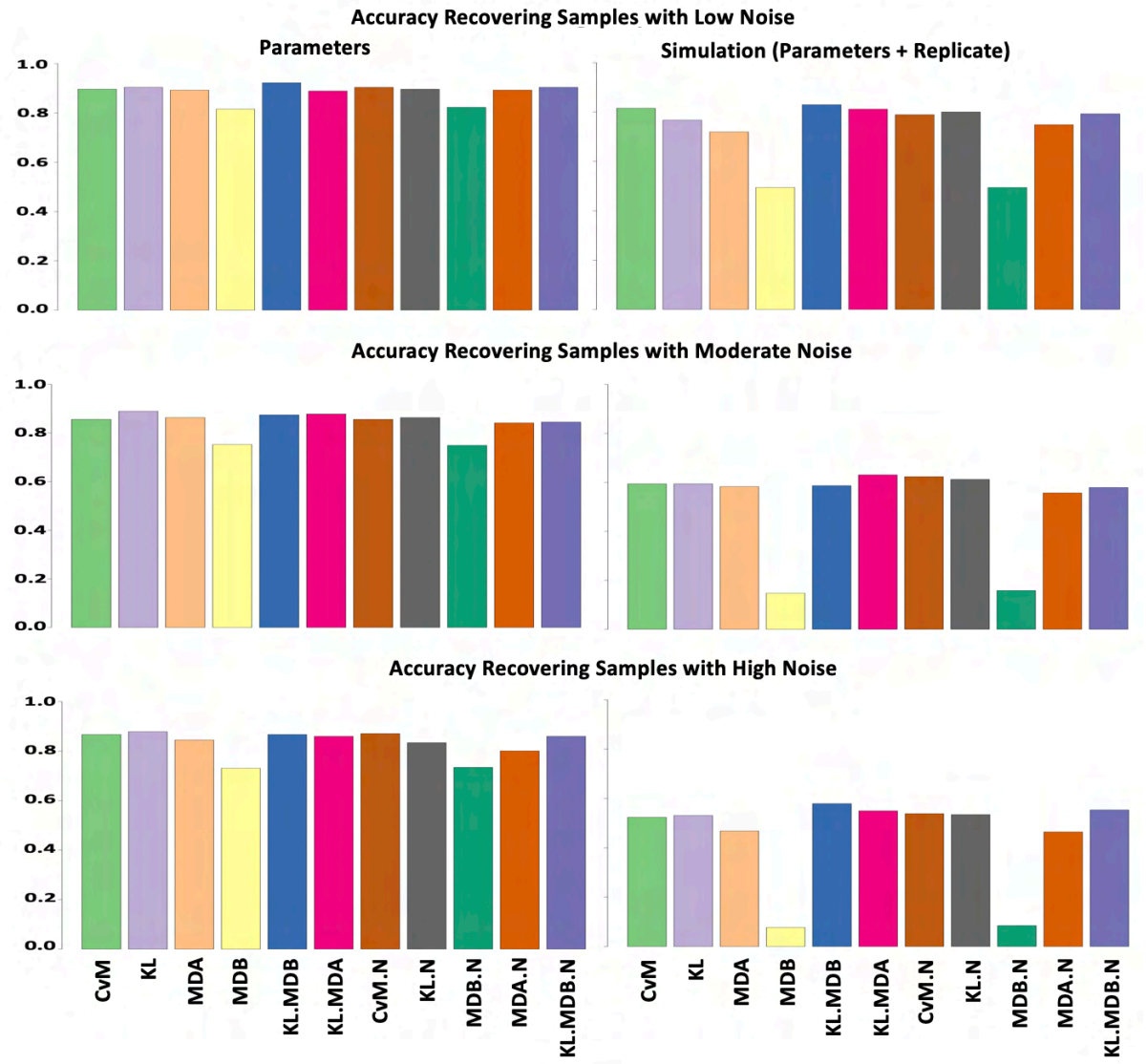


Figure 4.3 Benchmarking added noise experiments. Left bar plots show accuracy for recovering the simulation parameter within the top-10 fits, and the right bar plots show the equivalent for specifically recovering the correct individual simulation.

In the missing data experiments, MDA and MDA.N far outperformed the rest of the statistical methods, identifying both the correct set of parameters and the correct simulation with more than 90% accuracy, as shown in the right-hand side bar plots in Figure 4.4. It can be seen that distribution-free goodness-of-fit methods struggle to recover the correct simulation when values are missing. This is because the shape of the distribution is significantly affected by missing data and sample size. As a result, MDA and MDA.N are the preferred methods for missing data comparisons.

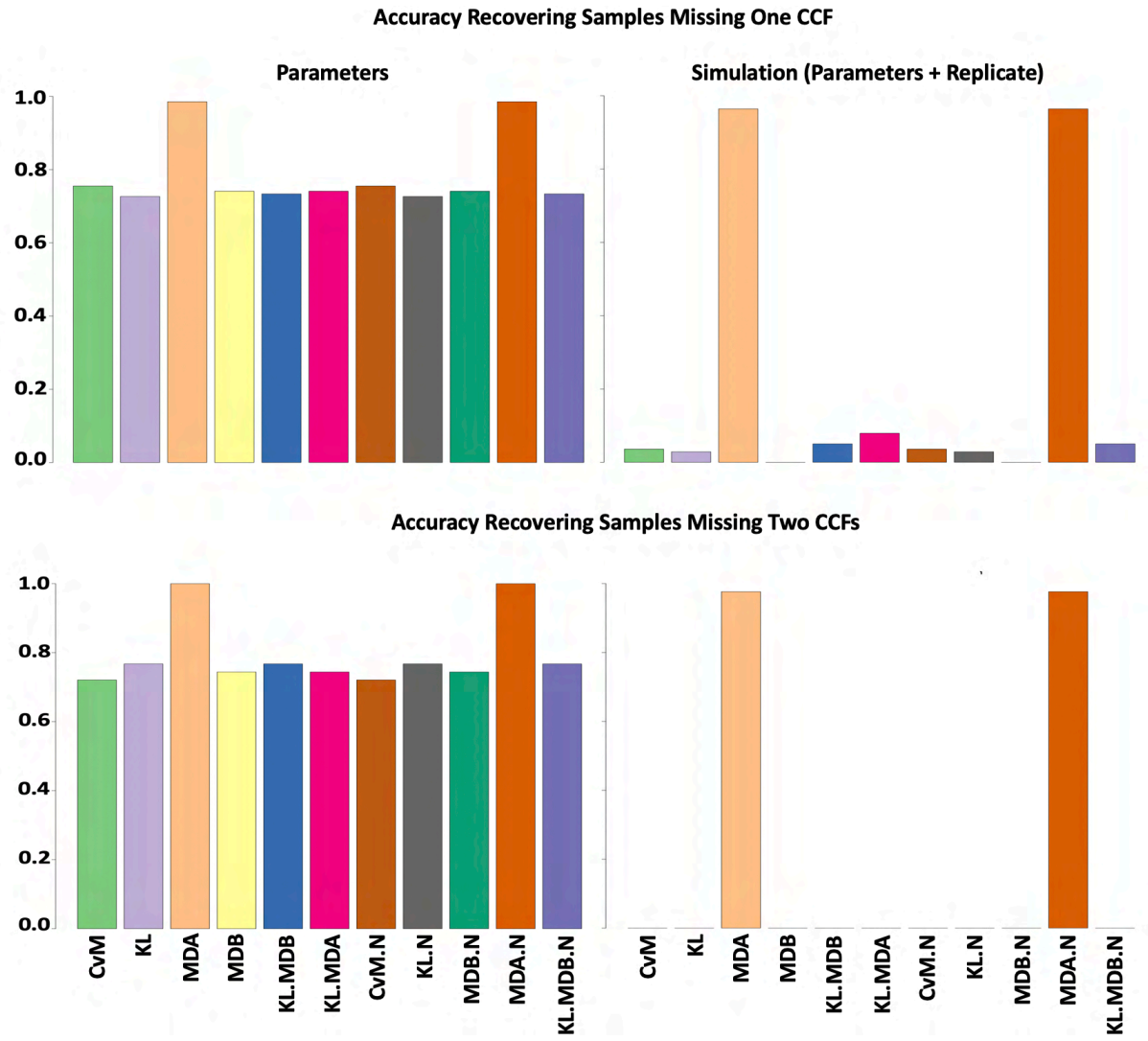


Figure 4.4 Benchmarking fits for missing-data experiments. Left bar plots show accuracy for recovering the simulation parameter within the top-10 fits, and the right bar plots show the equivalent for specifically recovering the correct individual simulation.

In the missing data plus noise experiments, similar results were observed with MDA and MDA.N outperforming the rest. However, in this scenario the accuracy of MDA and MDA.N is ~80% in finding the correct parameters in all experiments, as Figures 4.5 and 4.6 demonstrate. MDA.N is superior than MDA when the degree of noise is high, as the results for the missing one plus high noise experiment, Figure 4.6. As a result, even in the worst-case scenario MDA.N has ~40% accuracy in finding the correct simulation while all of the other methods struggle, displaying much lower accuracy when the degree of noise contamination is high.

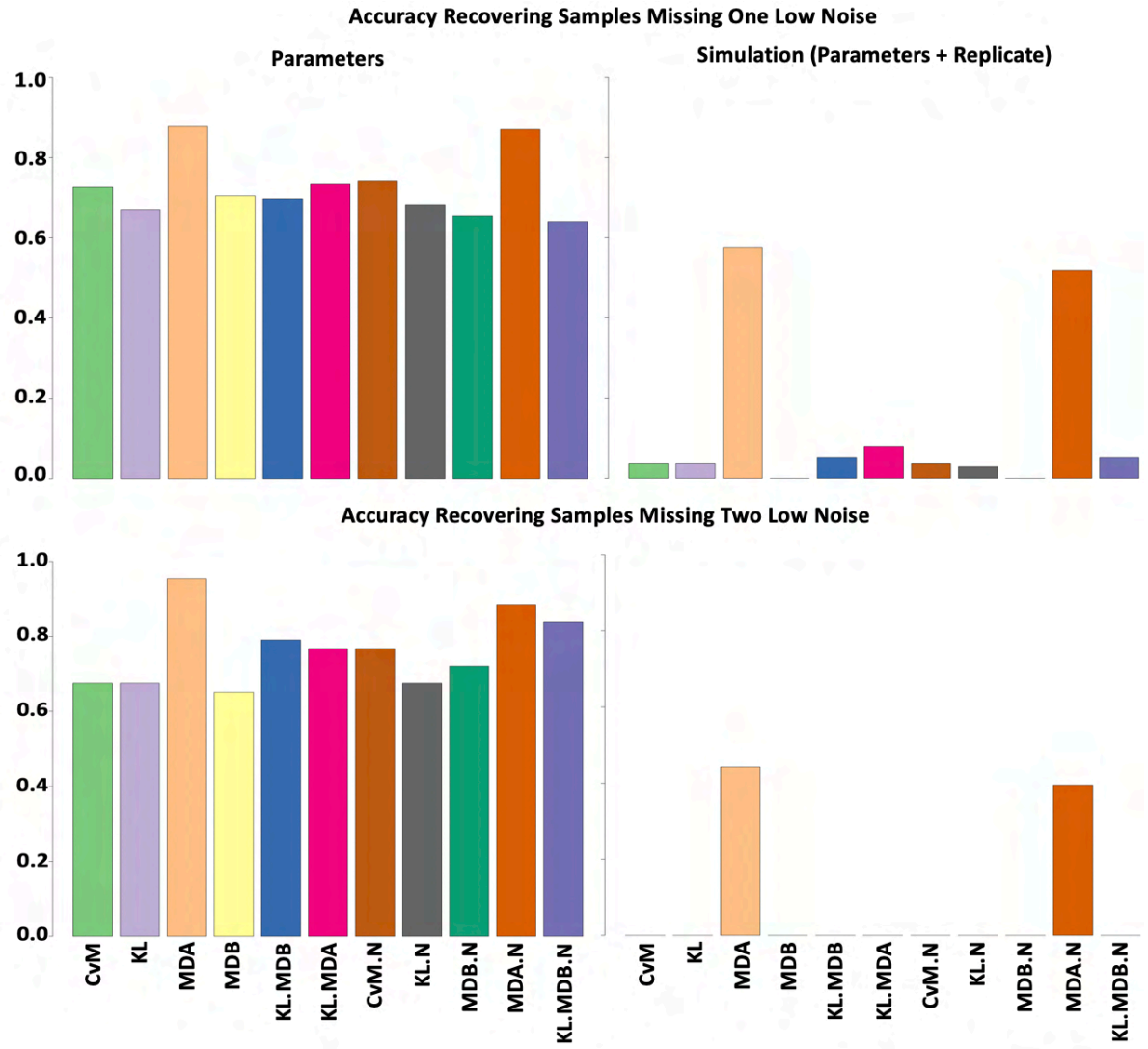


Figure 4.5. Benchmarking missing data experiments with high noise. Left bar plots show accuracy for recovering the simulation parameter within the top-10 fits, and the right bar plots show the equivalent for specifically recovering the correct individual simulation.

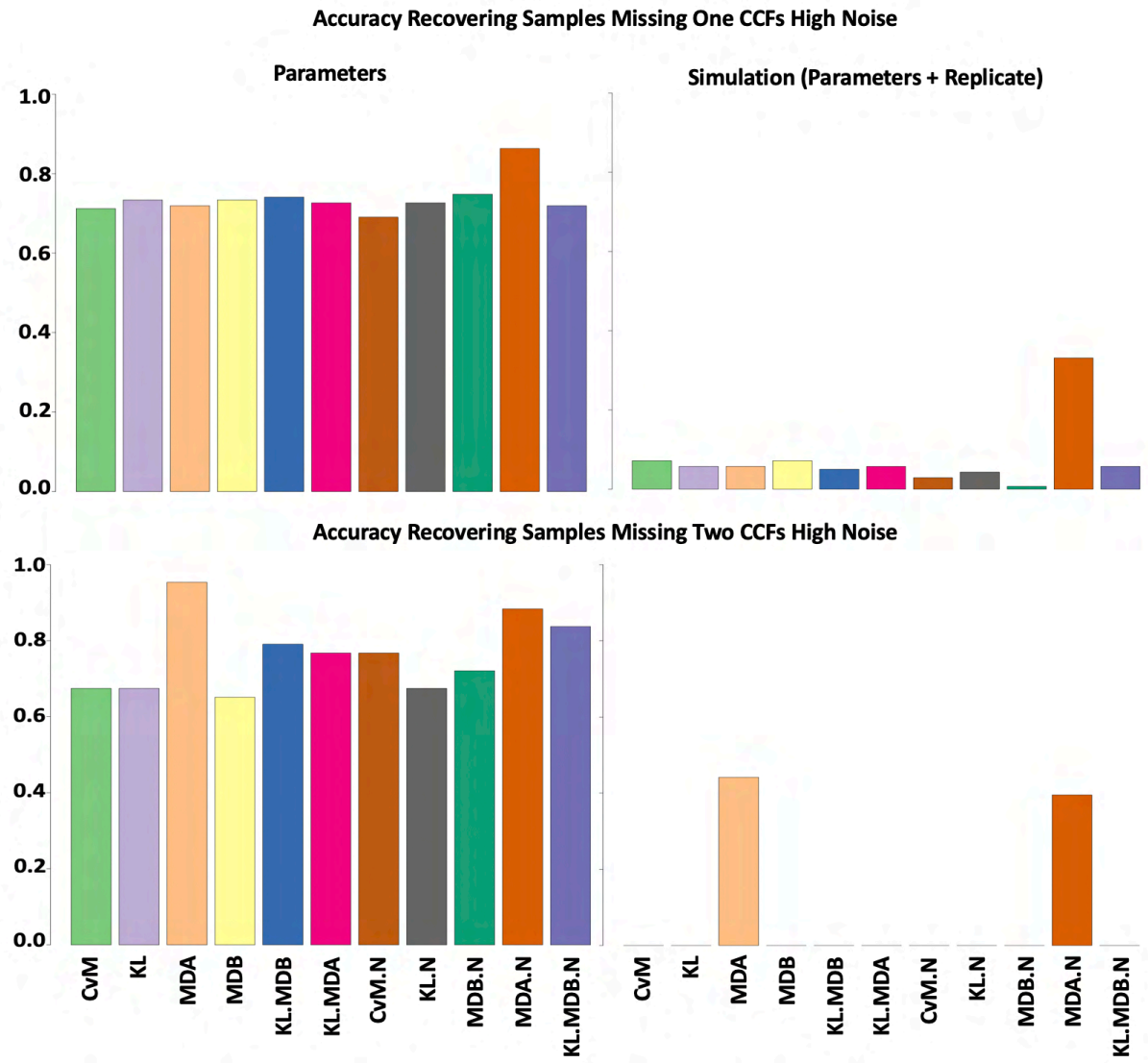


Figure 4.6 Benchmarking missing data experiments with high noise. Left bar plots show accuracy for recovering the simulation parameter within the top-10 fits, and the right bar plots show the equivalent for specifically recovering the correct individual simulation.

The results of the benchmarking showed MDA.N as the most reliable all-round statistic with at least ~80% accuracy in predicting the correct parameters, and ~40% accuracy in specifically finding the correct simulation. Further ensembles using MDA.N with KL and CvM were tested but only showed marginal gain.

As a result, MDA.N was selected as the statistical method to use for comparing simulated to real cancer cell fractions for all subsequent analyses.

5.5 Effect of Cancer Cell Fraction Length on Fits

The low accuracy of the statistical methods when data are missing is compounded by the degree of overlap of simulated cancer cell fractions, as shown in Chapter III. The lower the number of detectable CCFs in a given sample the higher the overlap between different simulations, resulting in reduced accuracy, Figure 4.7. To assess the impacts of this, the benchmarking process was repeated as described in Section 5.3 with only the MDA.N method, and the accuracy was evaluated based on number of CCFs used for fitting.

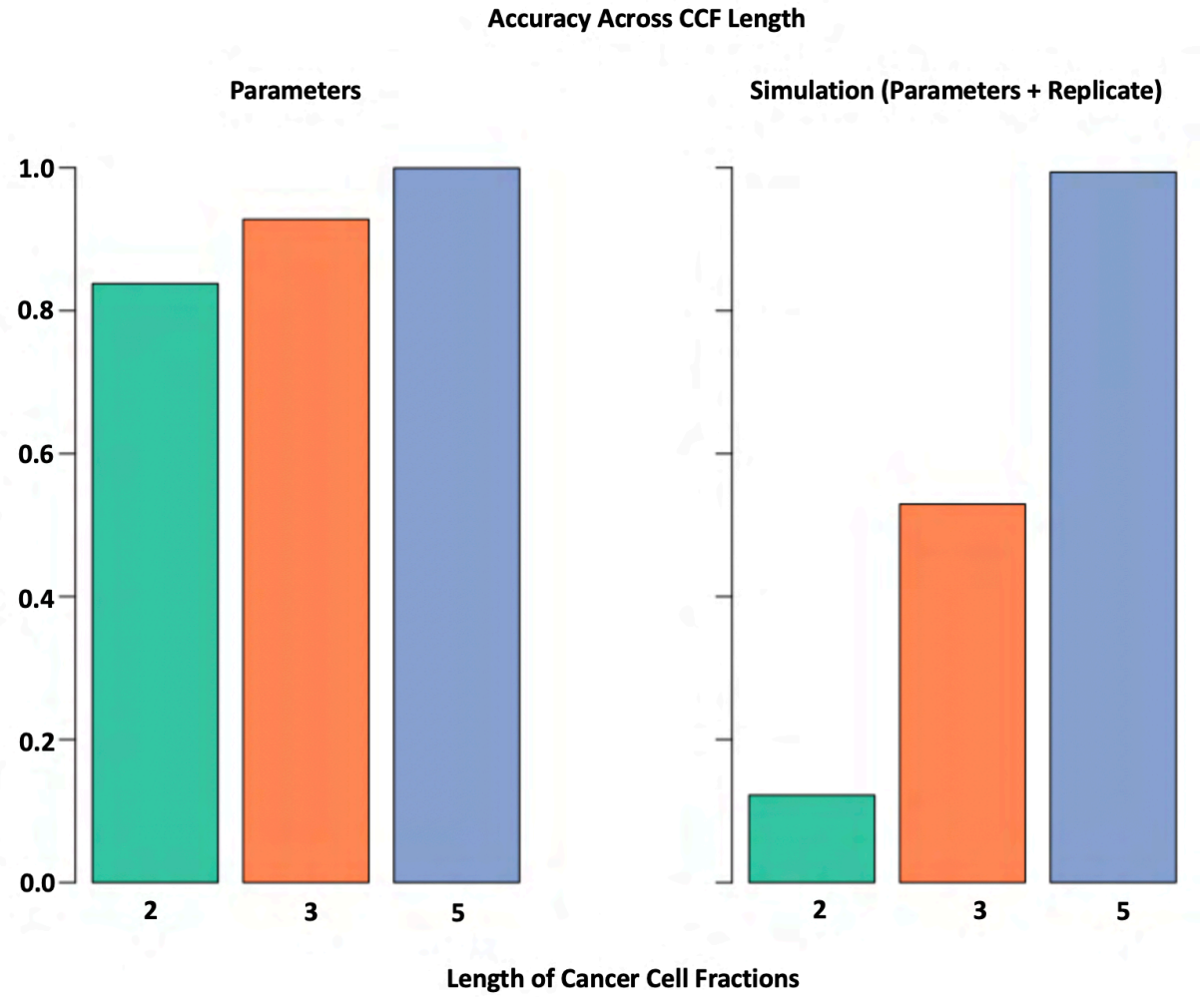


Figure 4.7 Relationship between length of cancer cell fractions and overall accuracy. Left bar plots display accuracy of parameter fitting as a function of the number of detected cancer cell fractions, and the right bar plots display the equivalent for accuracy in recovering the correct simulation.

Accuracy can be improved by increasing the sequencing depth or by sampling multiple regions which both provide better resolution to determine the CCF pattern unique to the sample. However, in cases where certain cancer cell fractions are missing MDA.N displayed the best performance (Fig 4.6).

5.6 Sequential Approximation of the Fitting Procedure

The fitting procedures described in the previous sections are based on comparisons to previously simulated realisations of fixed parameter combinations of *s* and *u.* As a result, the fits are approximations of the true values. However, the estimated parameters $\hat{\theta}$ can be improved by sequential approximation through iterative application of the preferred statistical method.

This process can be computationally intensive, requiring multiple rounds of branching process simulations for changing parameter combinations of s and u as new realisations are required to approximate θ. This reiterates the fact of why the CCF comparison method was chosen.

Although high performance computing approaches can improve this task, and 289,400 simulations to can be used identify the best starting point, this approach is only recommended when multiple samples were taken and deep sequencing was done, such as with the TRACERx NSCLC data set.

The following are the steps for the sequential approximation fitting,

1. Approximate $\hat{\theta}$ with the statistical method of interest using the positive selection sample of 289,400 tumour snapshots to determine the range of *s* and *u.*
2. At $t = 0$, draw a random sample of parameters using the range observe in 1 $\left\{\hat{\theta}_b^{(0)}, 1 \le b \le B\right\}$ from the *p-variate* normal distribution $N_p\left(\theta^{(0)}, \Sigma^{(0)}\right)$.
3. Set $t = t + 1$, draw observations $\left\{y_b^{(t)} \sim F\left(\cdot; \hat{\theta}^{(t)}, 1 \le b \le B\right)\right\}$ and compute the corresponding statistical method (MDA.N is preferred).
4. Compute the ρ-quantile $\tilde{c} = c^{(t)}_{(\lfloor B(1-\rho) \rfloor)}$, where $c^{(t)}_{(1)} < \cdots < c^{(t)}_{(B)}$ are ordered statistics from step 2. Then update the parameter values as,

$$\theta^{(t)} = \frac{\sum_{b=1}^{B} \hat{\theta}_b^{(t-1)} I(c_b^{(t)} > \tilde{c})}{\sum_{b=1}^{B} I(c_b^{(t)} > \tilde{c})}, \Sigma^{(t)} = \frac{\sum_{b=1}^{B} I(c_b^{(t)} > \tilde{c})(\hat{\theta}_b^{(t-1)} - \theta^{(t)})(\hat{\theta}_b^{(t-1)} - \theta^{(t)})^2}{\sum_{b=1}^{B} I(c_b^{(t)} > \tilde{c})}$$

5. Repeat steps 2-4 until $\frac{\|\theta^{(t)} - \theta^{(t-1)}\|}{\|\theta^{(t-1)}\|} < \varepsilon$ for some tolerance level $\varepsilon > 0$.

The following figure shows an example of the sequential approximation fitting using simulated data as a test set with 100 CCFs as input. The goal was to identify parameters $s = 0.004$ and $u = 3.2\text{x}10^{-5}$ extracting the CCFs of one realisation of a tumour of size 2 cm^3.

In this example, the CvM statistic was used for a sample of 80 parameter combinations and 200 replicates every update with an error tolerance of 5%. To accelerate the computation a parallel approach was implemented with 8 cores working collectively.

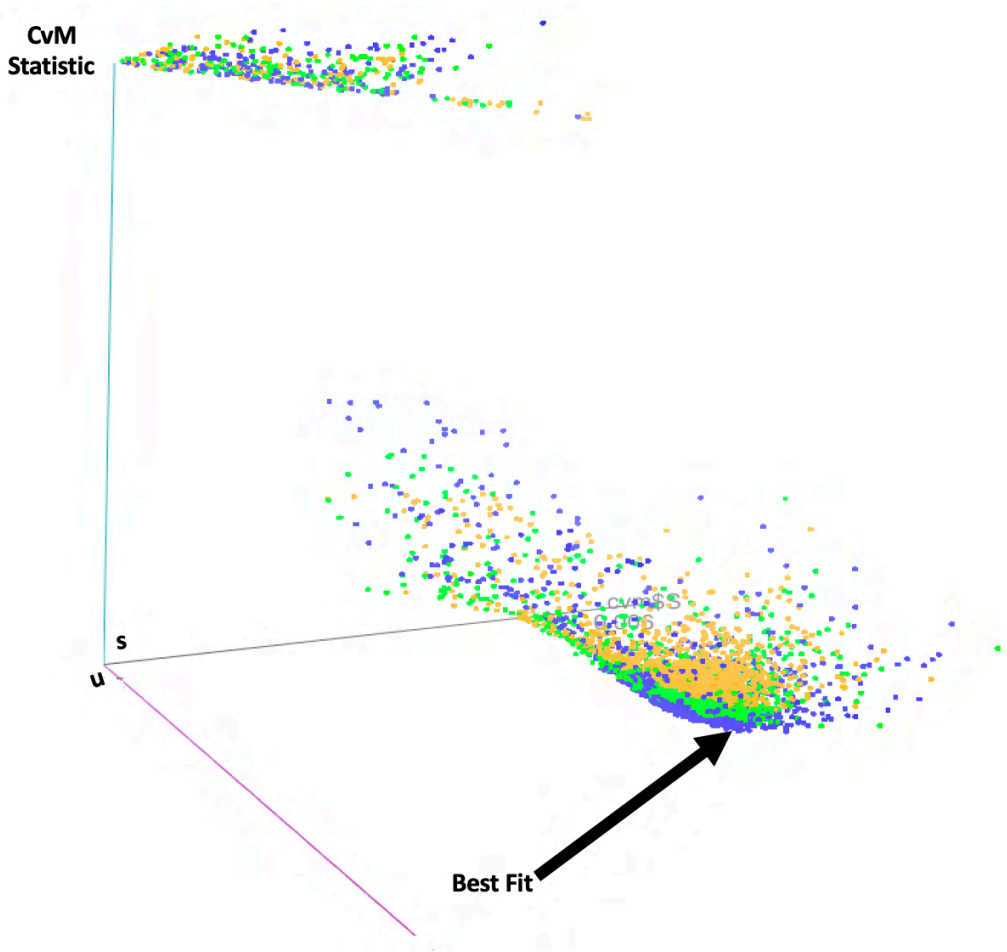


Figure 4.8 Sequential approximation of a test case. Axis represent the predicted parameters and the z axis the CvM statistic of 4,800 simulations. The colours orange, green and blue represent the final three iterations (18, 19 and 20).

The best fit was obtained after 20 batches and 32,000 simulations with estimated parameters $\hat{\theta}$ as $\hat{s} = 0.00389$ and $\hat{u} = 3.28\text{x}10^{-5}$, and as shown in Figure 4.3 with the colour gradient. The sequential approximation stopped when the rate of change of the ρ-quantile showed minimum change.

Although, the sequential approximation can identify the values of s and u, if the number of CCFs obtain from sequencing is below 30 the sequential approximation may not converge. Therefore, the single-comparison method accepting the top 10 MDA.N shown in the previous sections is preferred, and is going to be used in fits in the next sections.

6 Reconstructing Tumour Evolution in Real Datasets

In the previous section I identified that MDA.N, the minimum Euclidean distance with neighbour averaging for false discovery correction, displayed the best performance for identifying correct simulation parameters as well as the specific individual simulation used as the 'truth set' across all benchmarking experiments.

With the intention of reconstructing tumour evolution, the next aim is to apply the MDA.N method to real datasets representing different cancer subtypes and different sequencing assays

and correlate the simulations with clinical outcome. The fitting strategy to compare the simulations with real sequencing data was illustrated in Figure 4.1.

I analysed four studies covering a range of cancer subtypes, sequencing assays and coverage depths. Some of the studies have longitudinal data that can be used to establish association of inferred parameters with survival or distant recurrence. The following table is a summarises the datasets evaluated here.

Table 4.1 Studies Used to Reconstruct Tumour Evolution

Study	*Seq. Assay*	*N*	*Multi-region*	*Clonality Tool*	*Tumour Type*	*Subtype*	*Longi-tudinal*
TCGA	Whole exome	1,165	No	PyClone and ExPANdS	Primary and Metastasis	Multiple	No
TRACERx NSCLC	Whole exome	96	Yes	PyClone	Primary	Non-small cell lung	Yes
BIG 1-98	Amplicon, 287 genes	538	No	PyClone and ExPANdS	Primary	Breast	Yes
CASCADE melanoma	Whole exome	1	Yes	ExPANdS	Primary and Mets.	Melanoma	Yes

TCGA and TRACERx NSCLC are publicly available studies which made estimates of cancer cell fraction distributions available for download, avoiding the need to perform variant calling, copy number estimation and inference of clonality [29, 76].

In contrast, BIG 1-98 and CASCADE are private in-house projects that required extensive bioinformatics pre-processing before generation of clonality estimates. These datasets are not available for download.

The MDA.N method was applied each study with different aims in each case, depending the particular conditions and scenarios relevant to that study.

The TCGA cohort will establish the power of the MDA.N method to reconstruct tumour evolution in whole exome sequencing assays from multiple malignancies with diverse mutational signatures. It will investigate,

- The similarity of clonality distributions between subtypes and their relation with overall survival.
- How the particular program used to estimate clonality affects to results of the fitting procedure.
- Whether recurrent phylogenies can be found across subtypes.

TRACERx NSCLC will establish how well the MDA.N method performs in fitting high-depth (1%) whole exome sequencing assays with multi-region samples. This study will also inform

- How well the method performs in the presence of genome wide copy number alterations that are not explicitly modelled.

- Association with clinicopathological factors and survival.
- Degree of concordance with phylogenies previously reported by the TracerX consortium.

Data from the BIG 1-98 cohort will be used to explore if the MDA.N method can reconstruct tumour evolution from a panel covering a limited slice of genome (500 genes) but sequenced to high depth from formalin-fixed paraffin embedded (FFPE) samples. This study will inform

- How well the method performs in the presence of high noise due to DNA degradation in FFPE.
- How well the method preforms with limited sequencing coverage.
- Association with clinicopathological factors and distant recurrence.

CASCADE melanoma will explore the capability of the MDA.N method to identify similarity between primary and metastases of samples collected over time and aid in determining patterns of clonal dissemination.

As described in Figure 4.1, every patient in the aforementioned studies was compared to all the 289,400 snapshots of the 14,470 tumours, and the likelihood of the parameters established via neighbouring consensus for false discovery correction.

Moreover, due to the degree of overlap of cancer cell fractions in Chapter III, it is expected that most samples will fit one the two clusters, either the strong fitness and/or average driver mutation rate or the moderate/weak fitness and high average driver mutation rate ($s = 0.1$ & $u = \{3.4\text{x}10^{-7}, 3.4\text{x}10^{-6}\}$ vs $s = \{0.001, 0.01\}$ & $u = \{3.4\text{x}10^{-5}, 3.4\text{x}10^{-4}\}$).

6.1 Estimating Average Selective Advantage *s* and Average Driver Mutation Rate *u* in TCGA

Tumours differ not only by subtype but also according to their mutational signatures, mutational load and prognosis. Although, many clinical and molecular differences have been reported for different cancer subtypes, an evolutionary assessment incorporating fitness and driver mutation rates is lacking.

The main goals of the work covered in this section are twofold. First, to evaluate the capability of MDA.N to reconstruct tumour evolution from whole exome sequencing assays from multiple malignancies and determine its correlation with overall survival. Second, to establish evolutionary patterns across malignancies reported in TCGA.

To achieve these goals, I used data from Andor et al. [29], who measured the relationship of heterogeneity and survival in TCGA and reported the cancer cell fractions and clonality distributions of multiple malignancies and their relationships to heterogeneity and survival. Supplementary Table 4.1 provides specific survival details per cancer subtype.

They used two programs to estimate clonal frequencies, PyClone and ExPANdS, and found moderate concordance in the inferred clonal distributions (Spearman $\rho = 0.77$). Therefore, it is expected that fits using both tools should show similarity as well.

Of the two main simulation clusters (Figure 3.15), TCGA tumours consistently displayed solid fits to the high clonality subset ($s = 0.001$ & $u = \{3.4\text{x}10^{-4}, 3.4\text{x}10^{-5}\}$) rather than to the low-clonality cluster, as shown in Figure 4.9. Results are for the consensus calls of all the 3 positive selection models (additive fitness, stickbreaking and increased mutation rate) using both clonality callers (PyClone or ExPANdS) in all TCGA tumours. Thus, not only are the identified values of s and u within the ranges reported by Bozic et al. [43] but also the predictions were not affected by which clonality caller was used.

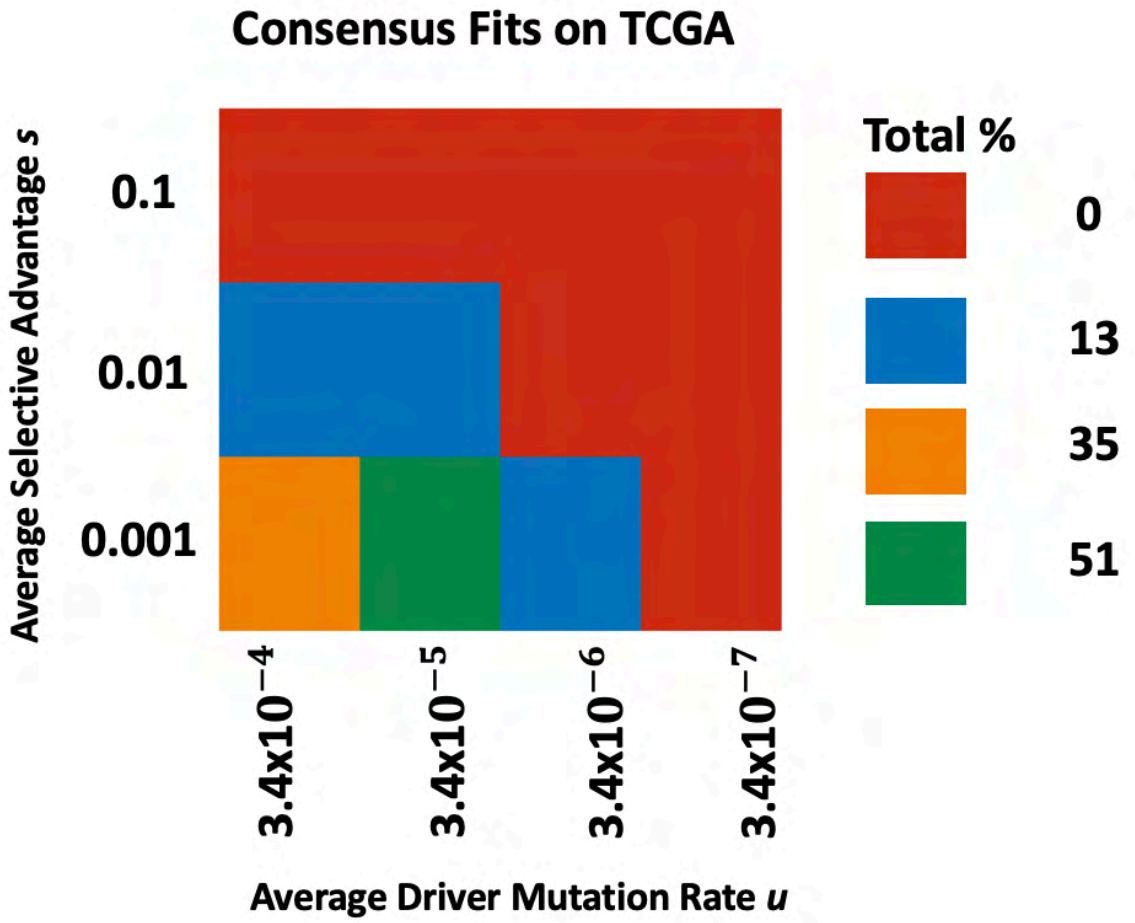


Figure 4.9 Fitting results for TCGA. Heatmap of consensus fits for subtypes using both clonality tools PyClone & ExPANdS. Intensity of the heatmap corresponds to the percentage of fits matched to each combination of s and u out of the total measured.

Next I compared the distribution of inferred number of clones from the Andor et al. vs the best scoring fits for every tumour to identify similarity.

The following table shows the range of median numbers for clones at tumour sizes 1 cm^3 and 4 cm^3 (including all simulations as described in Chapter III) for the most frequently matched parameter combination ($s = 0.001$ & $u = 3.4\text{x}10^{-5}$, green in Figure 4.9). The number of median clones in the table can be used to determine how well the simulations approximate the inferred number of clones displayed in Figure 4.10.

Table 4.2 Median Number of Clones at 1 cm^3 and 4 cm^3 with $s = 0.001$ and likelihood of 51%

Model	u	m_C *at 1 cm^3*	m_C *at 4 cm^3*
Additive fitness	$3.4\text{x}10^{-5}$	5	6
Stickbreaking	$3.4\text{x}10^{-5}$	5	5
Stickbreaking	$3.4\text{x}10^{-4}$	6	7
Increased mutation rate	$3.4\text{x}10^{-5}$	8	9

m_C: Median number of clones greater than a 10 CCFs cut-off

Ranges in Table 4.2 slightly vary from the distributions of ExPANdS, PyClone and the predicted fits, Figure 4.10. However, this can be explained by errors from the clonality callers, potentially due to sequencing artefacts, leading to 1-2 missing clones as was reported by Andor et al. [29]. Sampling of only a single region of the tumour and uncertainty in parameter approximation could also contribution to this discordance.

Comparing the predicted number of clones in Figure 4.10.b with the expected median value of their parameters in Table 4.2, suggests that the likely range of s is between 0.01 and 0.001, in agreement with the ranges reported by Bozic et al. [43].

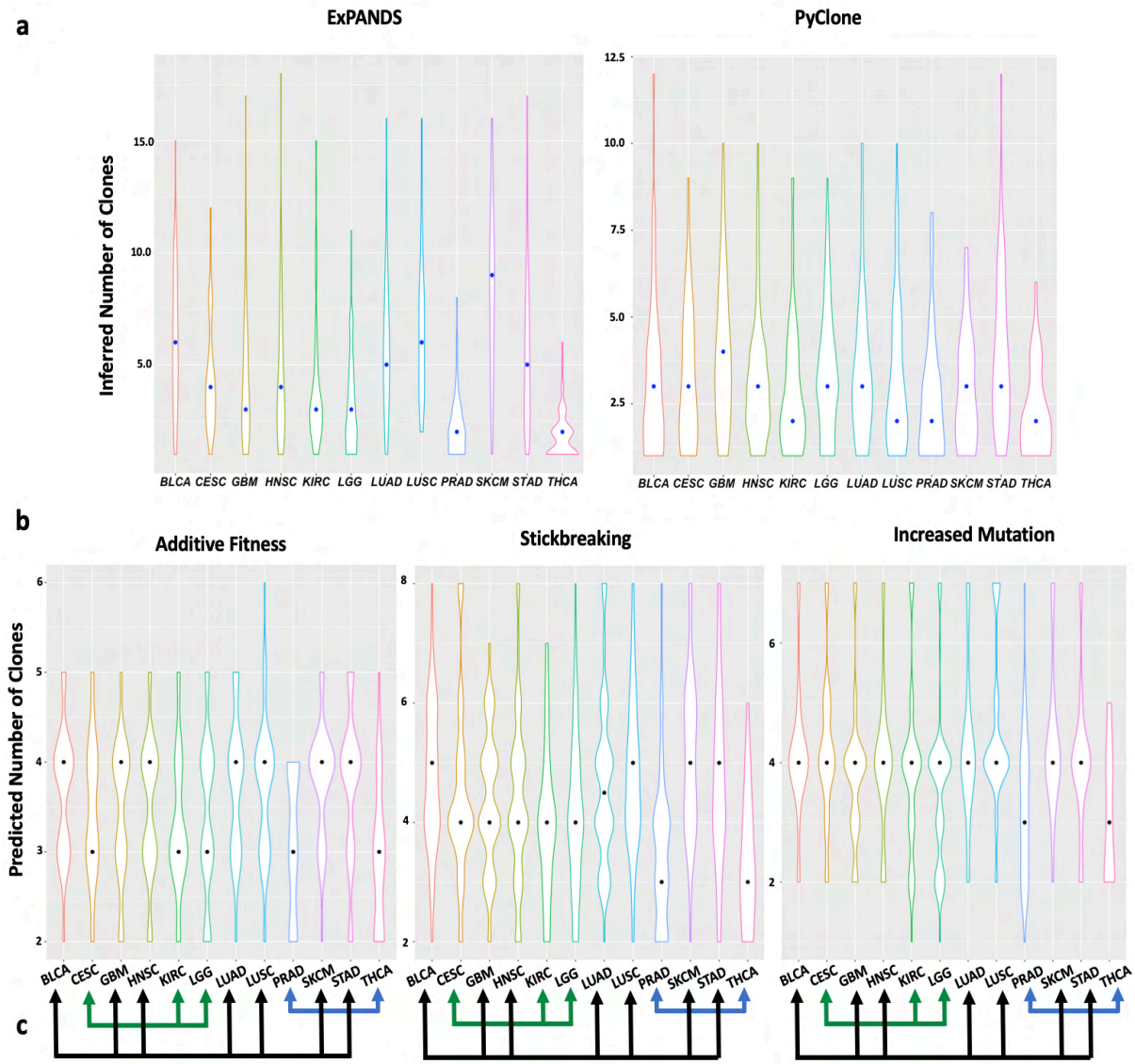


Figure 4.10 Inferred and predicted clonality distributions. a, shows the inferred clonality reported by Andor et al., using ExPANdS and PyClone, with dots representing median values. **b**, shows the predicted clonal distributions inferred from best fits to positive selection models, with simulated clones below 10% CCF filtered out with dots representing median values. **c**, the relationship to overall survival colour coded by prognosis, blue: good prognosis, green: moderate prognosis and black: poor prognosis.

Despite this slight variation, the fits are in concordance with the clonality callers. The predicted median values in most of the subtypes in all the models is 4, shown as black dots in Figure 4.10 panel b, which is similar in both clonality caller distributions. Furthermore, the relative differences in numbers of clones between subtypes closely matches those predicted by both the ExPANdS and PyClone distributions.

As a result, the observed distribution patterns suggest our model fits follow trends consistent with the genomic data, and therefore it is expected that other predicted evolutionary variables will as well.

Next, I studied the association of the estimated fits with clinical outcome. Clustering TCGA subtypes according the shapes of their detectable clonal distributions matched well with the

overall survival rates recorded for each subtype. Three clusters were detected, each with a similar range of survival. These clusters are delineated by blue, green and black in Figure 4.10 panel c as follows:

1. Good prognosis: THCA-PRAD (OS[%]; 96 and 98), coloured in blue.
2. Medium prognosis: LGG-KIRC-CESC (OS[%];75.5,66.8, 77), coloured in green
3. Bad prognosis: BLCA-LUAD-LUSC-STAD-HNSC-SKCM-GBM (OS[%];54, 58, 57, 61, 58,50,17), coloured in black.

The number of clones in Figure 4.10 is an indicator of tumour diversity as shown in Chapter III. The next question is if the overall survival is associated with tumour fitness. Figure 4.11 shows that fitness is proportional to the overall survival, the higher the fitness the better the prognosis which explains clonality distributions observed in Figure 4.10.b.

Median fitness was measured by weighting the detectable clones according to their fraction as shown in Equation (4.1).

$$m_{(1+s)} = median\left(\frac{C_{k,i}(1+s_{k,i})}{\sum_{k,i} C_{k,i}}\right) \tag{4.1}$$

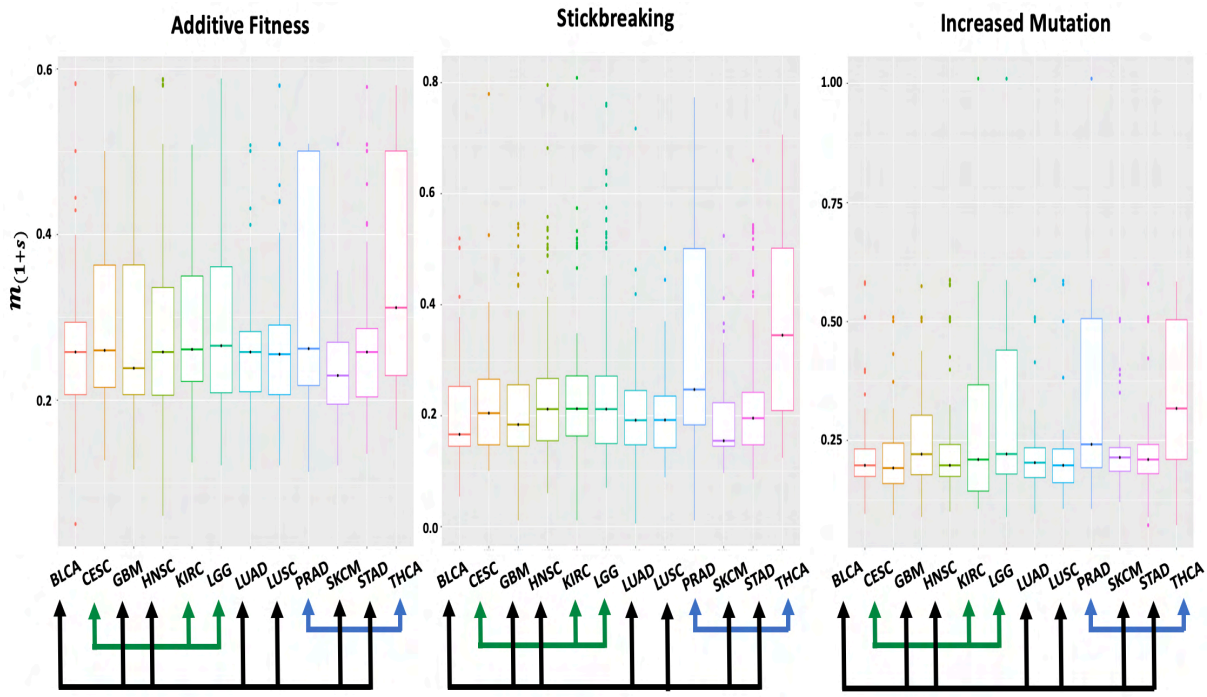


Figure 4.11 Predicted weighed fitness $m_{(1+s)}$. Black dots indicate median values. Arrows reflect the relationship to overall survival colour coded by prognosis, blue: good prognosis, green: medium prognosis and black: bad prognosis.

In Figure 4.11 the three models agree that THCA-PRAD have an increased median fitness which explains the reduced clonality in Figure 4.10. Relative to THCA-PRAD, the cluster containing LGG-KIRC-CESC subtypes have moderate median fitness which is more evident in the additive fitness and stickbreaking models. The rest of the subtypes (BLCA-LUAD-LUSC-STAD-HNSC-SKCM-GBM) have a reduced median fitness relative to THCA-PRAD.

Likewise, Figure 4.12 shows the RGS diversity which displays the same trend as in Figure 4.11, lower values of RGS are indicative of less heterogeneity, hence better prognosis.

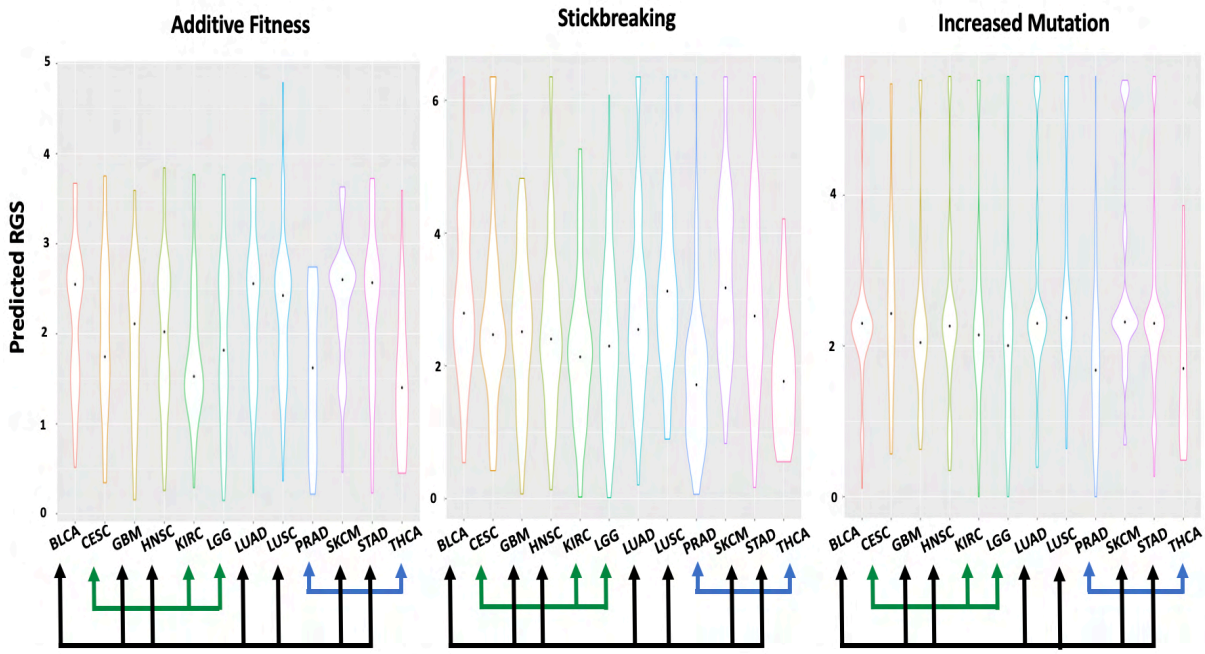


Figure 4.12 Predicted RGS diversity. Black dots indicate median values. Arrows reflect the relationship to overall survival colour coded by prognosis, blue: good prognosis, green: medium prognosis and black: bad prognosis.

Jointly, results in Figures 4.10, 4.11 and 4.12 suggest the relationship between number of detectable clones, fitness and RGS diversity with overall survival. Simulations of tumours in TCGA suggest that lower number of detectable clones, lower diversity and/or higher fitness is likely to explain a medium to good prognosis. Subsequent sections will explore how the evolutionary variables manifest in tumour phylogenies.

6.2 Recurrent Phylogenies and Clonal Evolution Reconstruction in TCGA

To further explore the relationship between heterogeneity and survival observed in the previous section, I identified the most recurrent topologies for each of the 3 positive selection models (additive fitness, stickbreaking and increased mutation rate).

The topologies were generated from the top-scoring simulation from the list of the top 10 candidates for every fit, and simulated clones below a 10% CCF were filtered out. Then, phylogenies are evaluated per cancer subtype and the top 3 recurrent topologies displayed in Figure 4.13.

Overall, the top 3 phylogenies describe between 41 – 60% of the top matching simulations per subtype, indicated in Figure 4.13. For example, the top 3 topologies of the SKCM represent 46% of their total whereas in LUAD 48%. The remaining topologies were present at low frequencies and the number of unique phylogenies is proportional to the variance of the detectable clonality distribution (Figure 4.10.b). For instance, LUSC, SKCM and STAD have a greater number of topologies as compared to the rest.

Different subtypes often showed similarity in their most frequent topologies. The degree of similarity was also consistent no matter which branching process model was used for fitting, shown as *Total* in Figure 4.13. Furthermore, it can be seen that the branched topologies are associated with worse survival as they are more frequent in the black cluster. In contrast, subtypes such as THCA and PRAD that have few detectable clones, were fit to less branched topologies potentially because they have less heterogeneity. The phylogenies observed in TCGA provide orthogonal evidence of the relationship of model fits with genomic data and clinical outcomes.

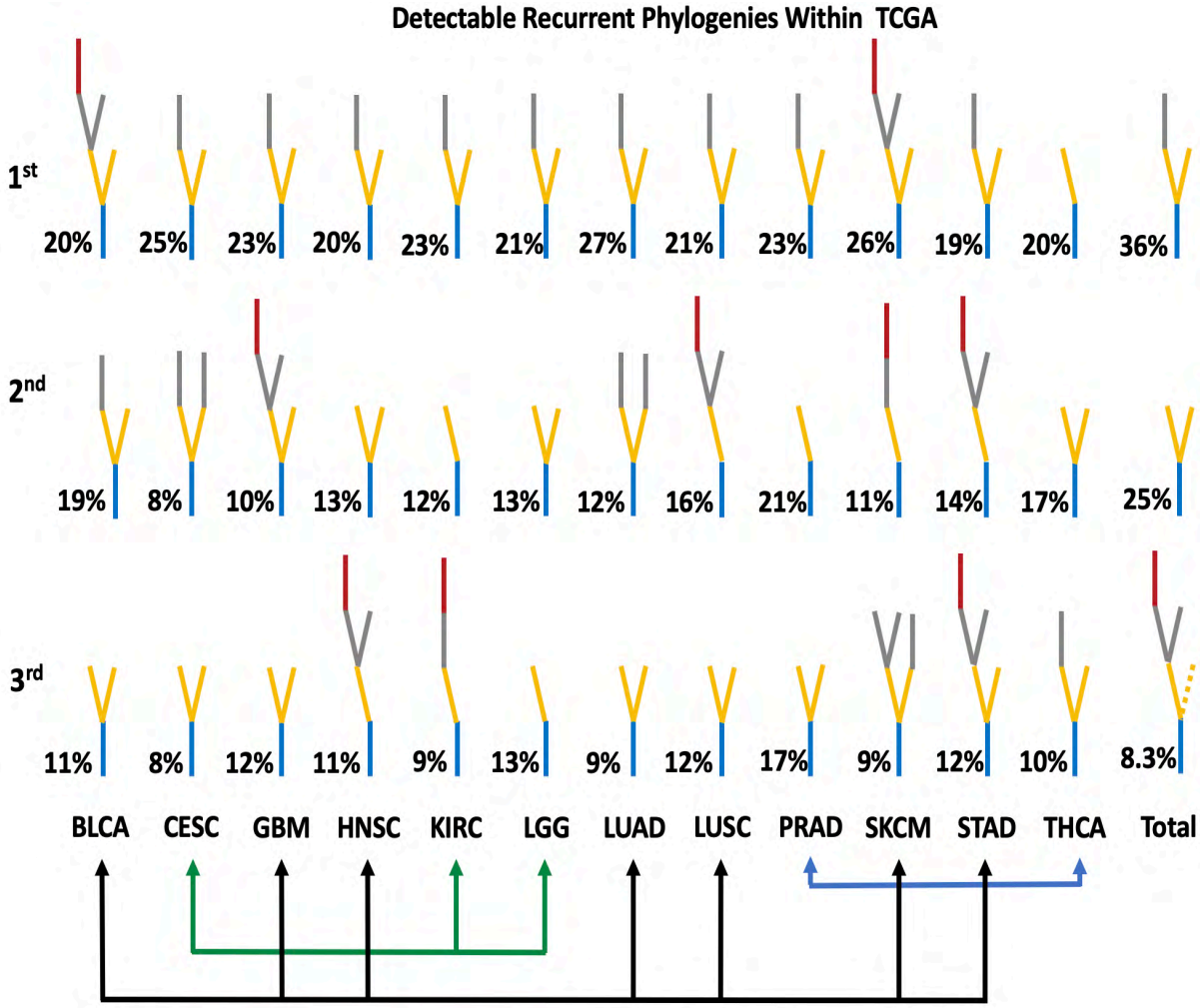


Figure 4.13 Top recurrent phylogenies in TCGA using the stickbreaking model. Phylogenies are ranked by frequency and percentages indicate the prevalence of the topology. Colours represent the number of drivers accumulated, blue, yellow, grey and red represent clones with 1, 2, 3 and 4 drivers. Arrows reflect the relationship to overall survival colour coded by prognosis, blue good prognosis, green medium prognosis and black bad prognosis. Total indicates percentages of the top-3 recurrent topologies across all subtypes combined. In third place there were two topologies of similar frequency, one marked by a solid line and the other with a solid plus dotted branch.

Not only were recurrent topologies observed across different models, but also certain specific simulations were recurrently identified as the best fits for multiple TCGA tumours as well. The most frequent recurrent simulation had the parameters $s = 0.001$ & $u = 3.4\text{x}10^{-5}$.

The following pages describe the evolutionary reconstructions of the top recurrent tumour simulation in each model. Clones below 1% CCF were removed for the additive fitness and stickbreaking models to better depict the structure of the phylogeny, whereas the increased mutation rate model this cut-off was set at 5% due to it having an increased number of clones.

Figure 4.14 shows the top recurrent simulation of the additive fitness model, that was identified in ~10% of all the fits to TCGA samples across all subtypes. In this simulation, tumour fitness increases at around generation 6,000, though the diversity surges at generation 9,000, which corresponds to the tumour reaching 2.5% of its final size (~100 million cells). At generation 9,000 only 25 drug resistant clones are present but from that time onwards, this number surges proportional to the RGS diversity.

When the tumour reaches nearly 500,000 cells (around generation 6,000), half of the tumour clonal architecture has been defined, as indicated by vertical black lines that mark point of lineage emergence in the *Generations vs Log(Size)* plot. This corresponds to just 0.001% of the complete course of tumour development in this simulation.

When the minimum detection limit is achieved, past the 10,000th generation, the clonal architecture is already established and subsequent progeny are below the detection threshold. Therefore, measurable clonal trajectories are defined when the tumour is less than 1% of its final size 4 cm^3. Interestingly, this effect was present in all the predicted simulations.

The tumour is mostly comprised of the 3-driver subpopulation which also manifests in the abundance of passenger/clonal and driver compositions, Figure 4.14 panel b. Lineage 1,77 is biggest in the tumour, containing 56% of all cells, and arose very early in tumour development when the tumour had around 50,000 cells in total.

Using the expected time for this lineage to merge, τ , as described in Chapter II, reveals the lineage emerged earlier than expected mean time of emergence for these parameters (3,447 vs 4,074) which provided ample time for further progeny to colonize and became detectable. This created a pattern of a branched evolutionary mode in which diverse clonal subpopulations emerge in sweeps. In this case clone 1,77,18 was the most dominant in the end, with 36% of the tumour composition.

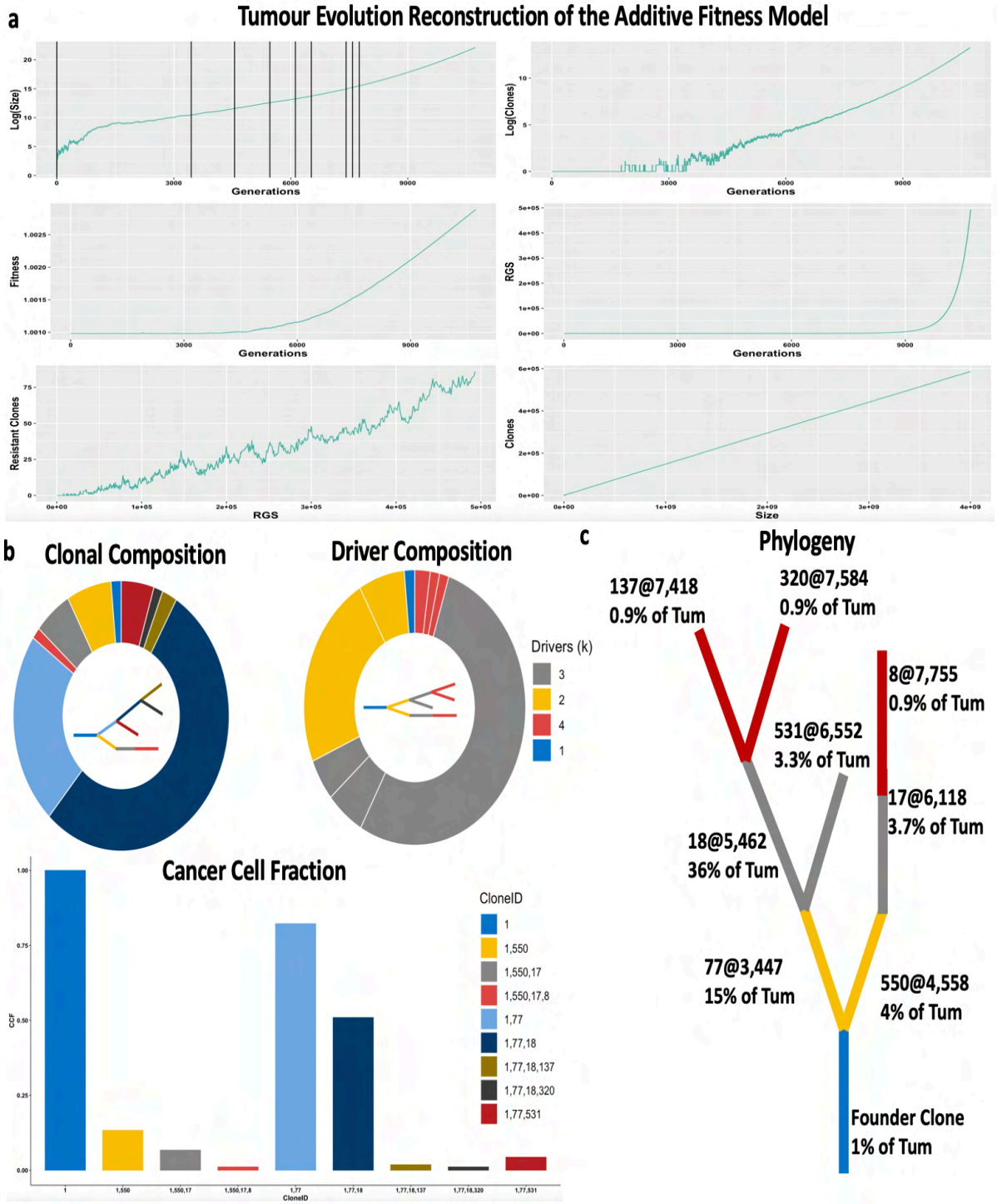


Figure 4.14 Top recurrent simulation in the additive fitness model. a, describes the tumour growth dynamics using the variables tumour size, generations, number of clones, fitness, RGS and resistant clones. Black vertical lines in the Log(Size) vs Generations indicate τ, the time when the detectable successful clones emerged and expanded. **b**, describes the clonal, and driver composition as a pie chart. Colours in the clonal composition indicate individual clones and in the driver composition the driver abundance. Cancer cell fraction bar plot is colour coded by the detectable clones. **c**, the phylogeny of driver clones for this simulation, colour coded by number of accumulated drivers k. Each clone is labelled with an ID number followed by @ with τ, the time of emergence of that clone. The proportion of the final tumour size represented by that clone is indicated underneath.

Figure 4.15 shows the properties of the commonly fit simulation from the stickbreaking model. This simulation was found in 17% of the fits and shows a different dynamic from the most recurrent simulation from the additive fitness model. Here lineage 140 acquired a strong driver to emerge from the background of the low fitness founder subpopulation. The fitness advantage was disseminated in further progeny within a short period leading to clonal dominance. This pattern is reflective of a *Big-Bang* evolutionary mode in which most driver mutations appear within a brief window of time, leading to subsequent expansion to a detectable size.

This simulation depicts the founder clone was strongly influenced by drift, as the timeframe of growth is three times higher than the top recurrent additive fitness simulation, with a greatly prolonged stochastic phase of tumour growth, Figure 4.15 panel a *Log(Size) vs Generation* plot. It was not until a strong driver emerged, lineage 140, that the tumour displayed a more aggressive expansion.

Similar to the additive fitness model, driver mutations emerged before the minimum detection limit, although in the stickbreaking model the *Big-Bbang* event occurred when the tumour was around 1.2% of its final size, which corresponds to ~50 million cells. This shows the relevance of the stickbreaking model in capturing different clonal dynamics that are it unlikely to occur in the additive fitness model.

The dominant clone in this simulation is a 4-driver clone corresponding to 40% of all cells. Jointly all the clones describe 80% of the tumour, providing more information compared to the case of additive fitness model. This is expected for a *Big-Bang* evolutionary mode, where multiple regions of the tumour show the same driver makeup. After a burst of driver alterations, fitness is gained accelerating tumour expansion causing further sub-clones to be undetectable.

Similar to the additive fitness model, the emergence of drug resistance clones occurs as a consequence of an increase in diversity at around generation 30,000, a point equivalent to the 1.2% of the tumour's complete development to 4 cm^3. This shows that the emergence of resistance clones is more strongly correlated with diversity than with time. Increases in diversity suggest increases in the number of divisions per time and tumour size, resulting in higher odds of resistance clones.

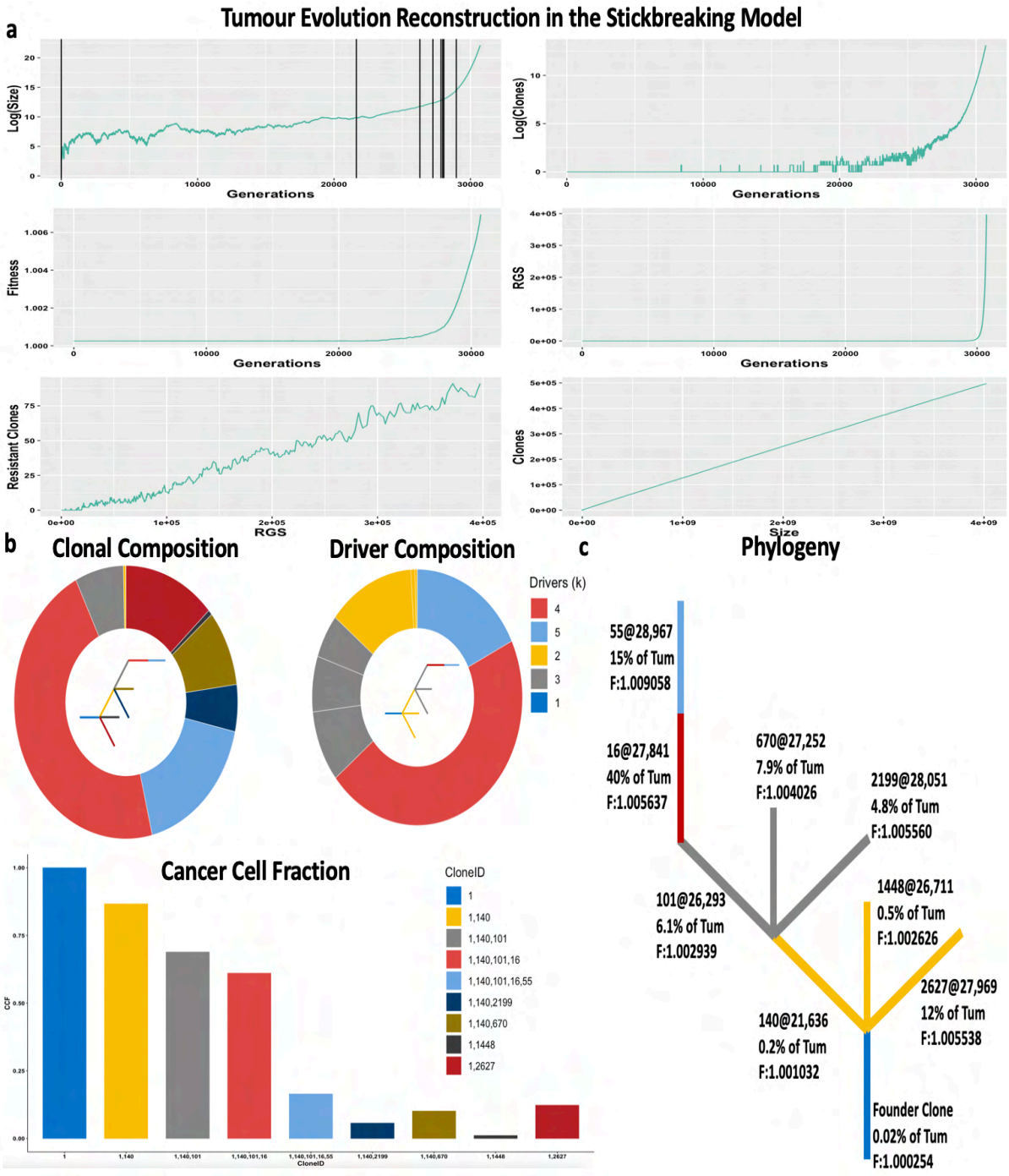


Figure 4.15 Top recurrent simulation in the stickbreaking model. a, describes the tumour growth dynamics using the variables tumour size, generations, number of clones, fitness, RGS and resistant clones. Black vertical lines in the Log(Size) vs Generations indicate τ, the time when the detectable successful clones emerged and expanded. **b**, describes the clonal and driver composition as a pie chart. Colours in the clonal composition indicate individual clones and in the driver composition the driver abundance. Cancer cell fraction bar plot is colour coded by the detectable clones. **c**, the phylogeny of driver clones for this simulation, colour coded by number of accumulated drivers k. Each clone is labelled with an ID number followed by @ with τ, the time of emergence of that clone. The proportion of the final tumour size represented by that clone is indicated underneath with their fitness.

The increased mutation rate model shows a similar dynamic and phylogeny to the additive fitness model, Figure 4.16. However, due to the computational cost of simulating this model at parameters $s = 0.001$ & $u = 3.4\text{x}10^{-5}$ there are few replicates as compared to the other models, resulting in a recurrent simulation appearing in 25% of the fits.

To analyse this recurrent simulation the cut off was set to 5% because of the large number of existing subpopulations at 1%. Most of the diversity comes from 2-3 driver clones, a consequence of the hyper mutation process, Figure 4.16 panel b (yellow branches in the phylogeny). Thus, the effect of increasing the mutation rate is translated into more heterogeneity and reduced clonal dominance, as observed in the RGS diversity over time, Figure 4.16 panel a. The most abundant clone represents 26% of the tumour and collectively all the clones describe ~43% of the tumour composition, lower than for the other models.

This demonstrates how the change in driver mutation rate impacts diversity, leading to more proliferative stress which increases the number of drug resistance clones as compared to the other models.

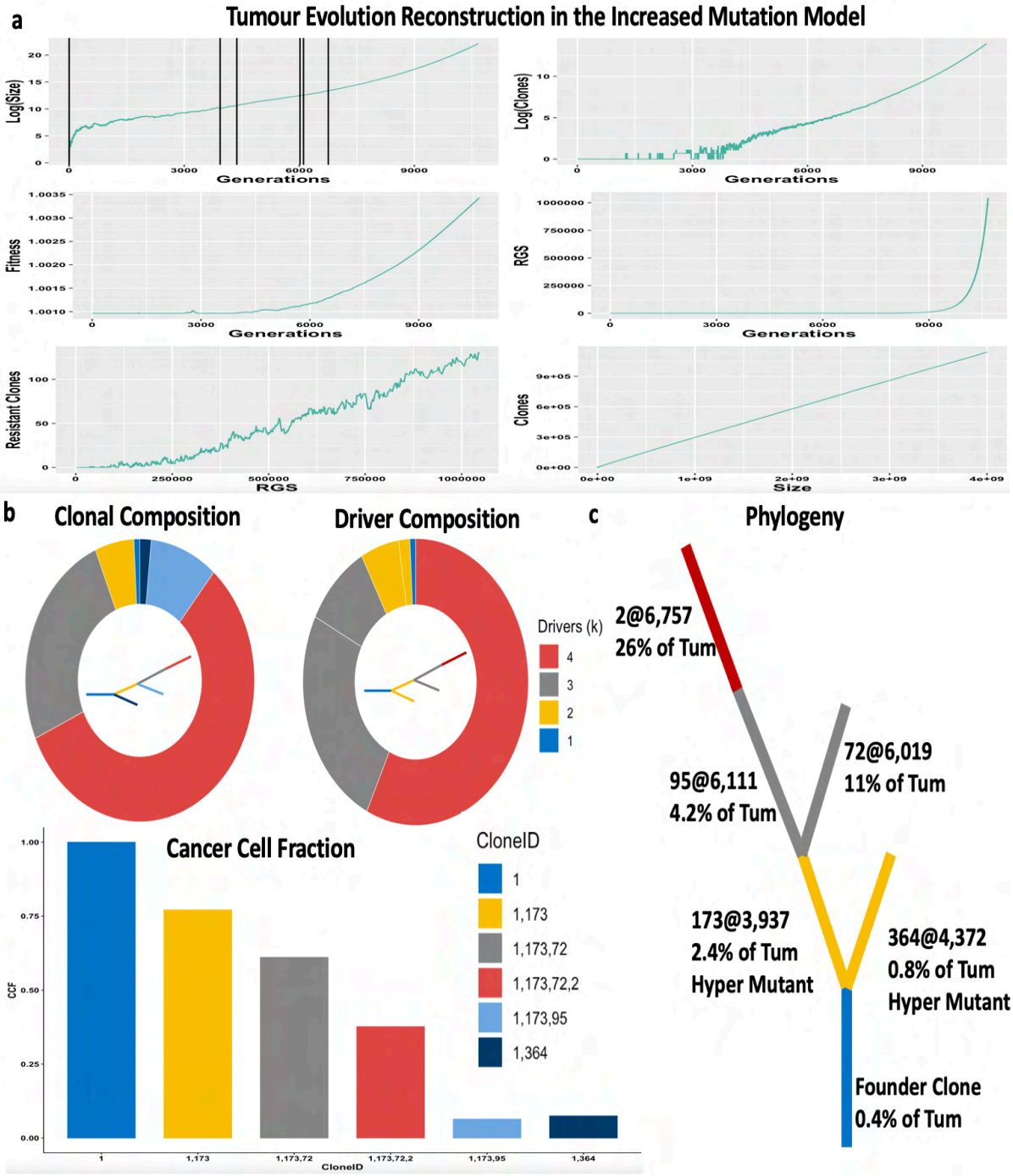


Figure 4.16 Top recurrent simulation in the increased mutation rate model. a, describes the tumour growth dynamics using the variables tumour size, generations, number of clones, fitness, RGS and resistant clones. Black vertical lines in the Log(Size) vs Generations indicate τ, the time when the detectable successful clones emerged and expanded. **b**, describes the clonal and driver composition as a pie chart. Colours in the clonal composition indicate individual clones and in the driver composition the driver abundance. Cancer cell fraction bar plot is colour coded by the detectable clones. **c**, the phylogeny of driver clones for this simulation, colour coded by number of accumulated drivers k. Each clone is labelled with an ID number followed by @ with τ, the time of emergence of that clone. The proportion of the final tumour size represented by that clone is indicated underneath with the status if the clone is hyper mutant or not.

Reconstructing tumour evolution in TCGA showed that the MDA.N method for fitting replicates predictions from ExPANdS and PyClone. The distributions of the detectable clonality and RGS correlate with clinical outcome. Cancer subtypes that were more likely to fit to more highly branched phylogenies were more likely to be those with lower overall survival.

Multi-branch topologies are an indicator of driver heterogeneity, which in turn is the consequence of accumulation of driver alterations in the tumour, resulting in increased replicative stress and emergence of resistance clones. Therefore, tumour size and RGS are better indicators of change of pre-existing drug resistance than is the length of time taken for a tumour to reach a certain size.

The additive fitness and increased mutation rate models replicate branched evolutionary modes because fitness variation is fixed in these models. Besides the branching evolution mode, the stickbreaking model can also recover the *Big-Bang* mode, which was present in at least 17% of the cases. Due to the earlier mutational driver events in the *Big-Bang* mode, single regions samples can provide a good approximation of the total composition of the tumour as compared to the branching modes.

6.3 Estimating Average Selective Advantage *s* and Average Driver Mutation Rate *u* in TRACERx NSCLC

TRACERx NSCLC is longitudinal study prospectively investigating the influence of tumour heterogeneity in non-small cell lung cancer on clinical outcome. The study performed ultra-deep whole exome sequencing (capable of detecting mutations found at 1% frequency) on multiple tumour regions in 100 early-stage cases. Deaths were recorded for 26 patients.

The clinical implications of several measures of intra-tumoural heterogeneity, including clonality, mutational co-occurrence, genome-wide copy number alterations were evaluated using Cox proportional hazards. In this section the main goal is to use the MDA.N method to reconstruct tumour evolution from multi-region, high-depth sequencing and assess the potential prognostic power of the information provided by the top fits.

TRACERx NSCLC identified that genomic instability and tumour heterogeneity is associated with an increased risk of death, thus it is expected that our model fits may show similar trend. The following figure shows the number of unique clusters computed with PyClone by the TracerX group, and the number of annotated drivers in the cohort grouped by clonal TP53 status.

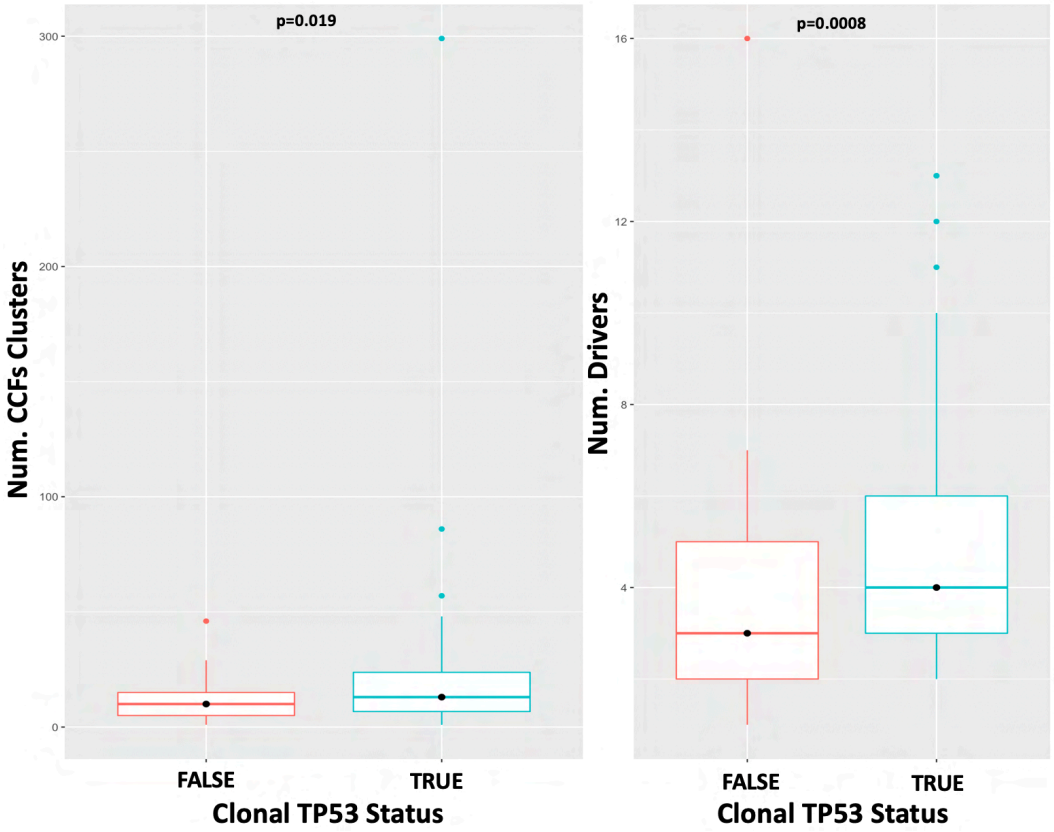


Figure 4.17 Boxplots of number of clusters and number of drives grouped by TP53 status in TRACERx.

It can be seen that clonal TP53 mutations increase driver heterogeneity, blue boxplots in Figure 4.17, and increases in the number of PyClone clusters or drivers are indicative of clonal sweeps. Chapter III showed that increases in driver heterogeneity are the result of increasing the driver mutation rate u or lowering the selective advantage ($s = 0.001$). There is considerable mutational burden in NSCLC, however, there is discrepancy between the median number of unique CCF clusters and the number of drivers, around 12 vs 4.

To compare with TRACERx NSCLC, the number of unique CCF clusters were fed into the fitting procedure using the MDA.N comparison technique. In this setting clones below 1% CCF were removed to match the sequencing depth of the study and fits were done to the average consensus of all regions

Similar to the best fits found for TCGA data, TRACERx NSCLC are better described by the high clonality cluster (Figure 3.15) with parameters, $s = 0.001$ & $u = 3.4\text{x}10^{-5}, 3.4\text{x}10^{-4}$, highlighted in orange and red in Figure 4.18.

The median number of detectable clones in the TracerX NSCLC cohort is 10, almost twice as many as for TCGA LUAD and LUSC subtypes observed in TCGA, Figure 4.10. In this cohort, RGS diversity is higher than the values in TCGA because its high-depth sequencing and multi-region sampling enable detection of greater number of clones.

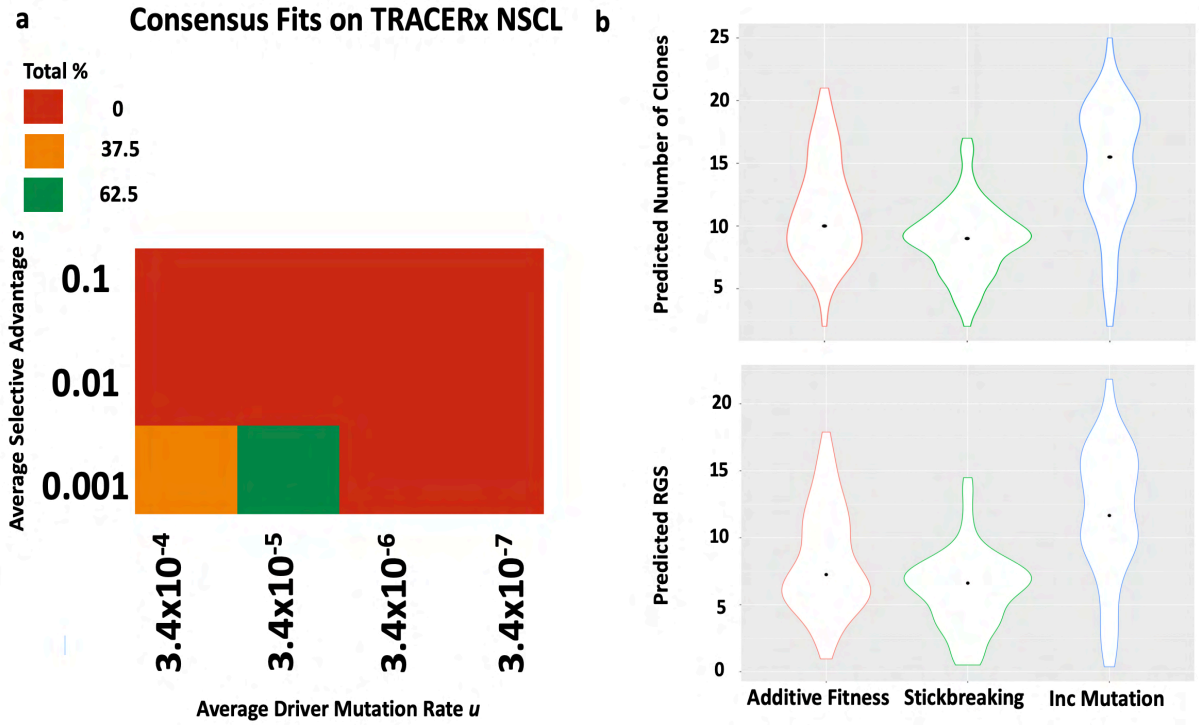


Figure 4.18 Fitting results in TRACERx. a, heatmap of consensus fits of all models from clonality calls done with PyClone. **b**, distribution of number of detected clones and diversity RGS for every model.

Panel a in Figure 4.18 corroborates what the modelling results in Chapter III suggested, that driver heterogeneity occurs with low average selective advantage s and high average driver mutation rate u.

Fits in TRACERx show concordance with subtypes LUSC and LUAD in TCGA, albeit with a slight difference. TCGA had cases that fit to a moderate average selective advantage $s = 0.01$ but that could be caused by the lack of regions sampled. PyClone results and presence of TP53 mutations in TRACERx are more consistent with low average selective advantage, $s = 0.001$. Further analysis to evaluate the quality of the fits will be presented in the next sections.

6.4 Recurrent Phylogenies and Clonal Evolution Reconstruction in TRACERx

In this section I am going to explore the similarities between phylogenies predicted from the best fit simulations and the ones reported by the TRACERx study, along with the overall patterns of the predicted and observed phylogenetic trees.

The best fits to TRACERx showed high RGS diversity, which translates to branched topologies and numerous outcomes. Given the 1% CCF resolution and the multi-region sampling, topologies in TRACERx are more branched than TCGA. Figure 4.19 shows the most recurrent topologies of the different models at 1% CCF. It can be seen they occur less frequently than observed in the results for the LUSC and LUAD subtypes in TCGA. This is due to improved resolution afforded by higher sequencing depth and multiple region sampling, which means more distinct topologies can be found for the TRACERx samples, reducing overlap.

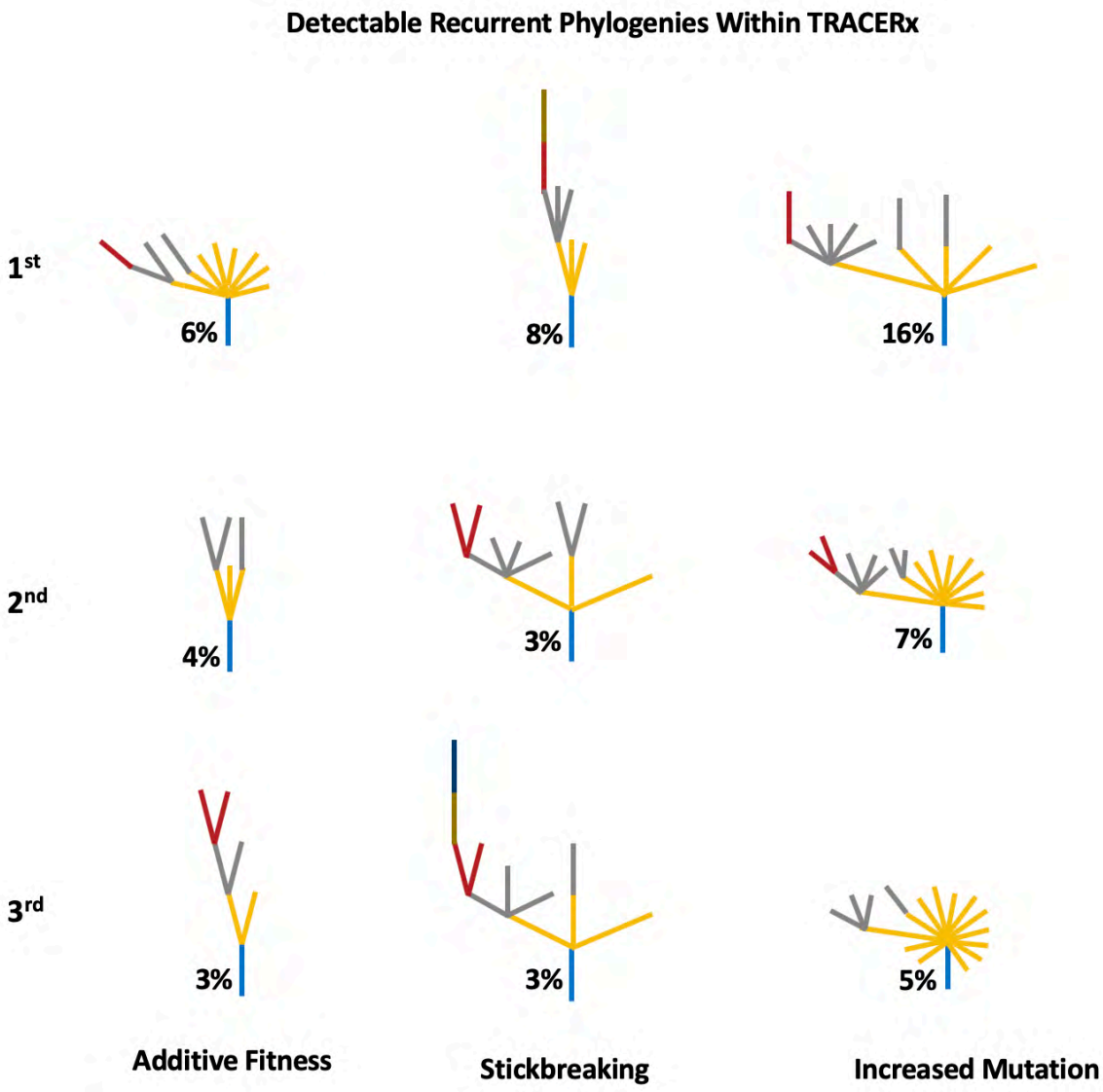


Figure 4.19 Recurrent phylogenies in TRACERx NCLC. Phylogenies are ranked by their frequencies, shown as percentages. Branches are coloured according to the number of drivers accumulated on each branch. Blue, yellow, grey and red represent clones with 1, 2, 3 and 4 drivers, respectively. The top fits to each of the branching process models are shown.

Similar to TCGA, the top recurrent simulations in TRACERx NCLC occurred with parameters $s = 0.001$ & $u = 3.4\text{x}10^{-5}$ in all models. The frequency of most recurrent simulation was 8%, 18% and 16% for the additive fitness, stickbreaking and increased mutation rate models, respectively. The most recurrent simulations in to additive fitness and increased mutation rate models were different between TRACERx NCLC and TCGA. However, the most recurrent stickbreaking simulation was the same one as seen in TCGA, Figure 4.15, which suggesting 18% of TRACERx NCLC cases also follow a *Big-Bang* evolutionary mode.

The top recurrent simulation in the additive fitness model, Figure 4.20, shows a branched evolution with 51% of the tumour expansion driven by 3 and 4 driver subpopulations from descended from lineage 67. Accounting for all clones, the detectable CCFs describe 73.5% of the total composition of the tumour, illustrating the advantage of deep sequencing and multi-region sampling.

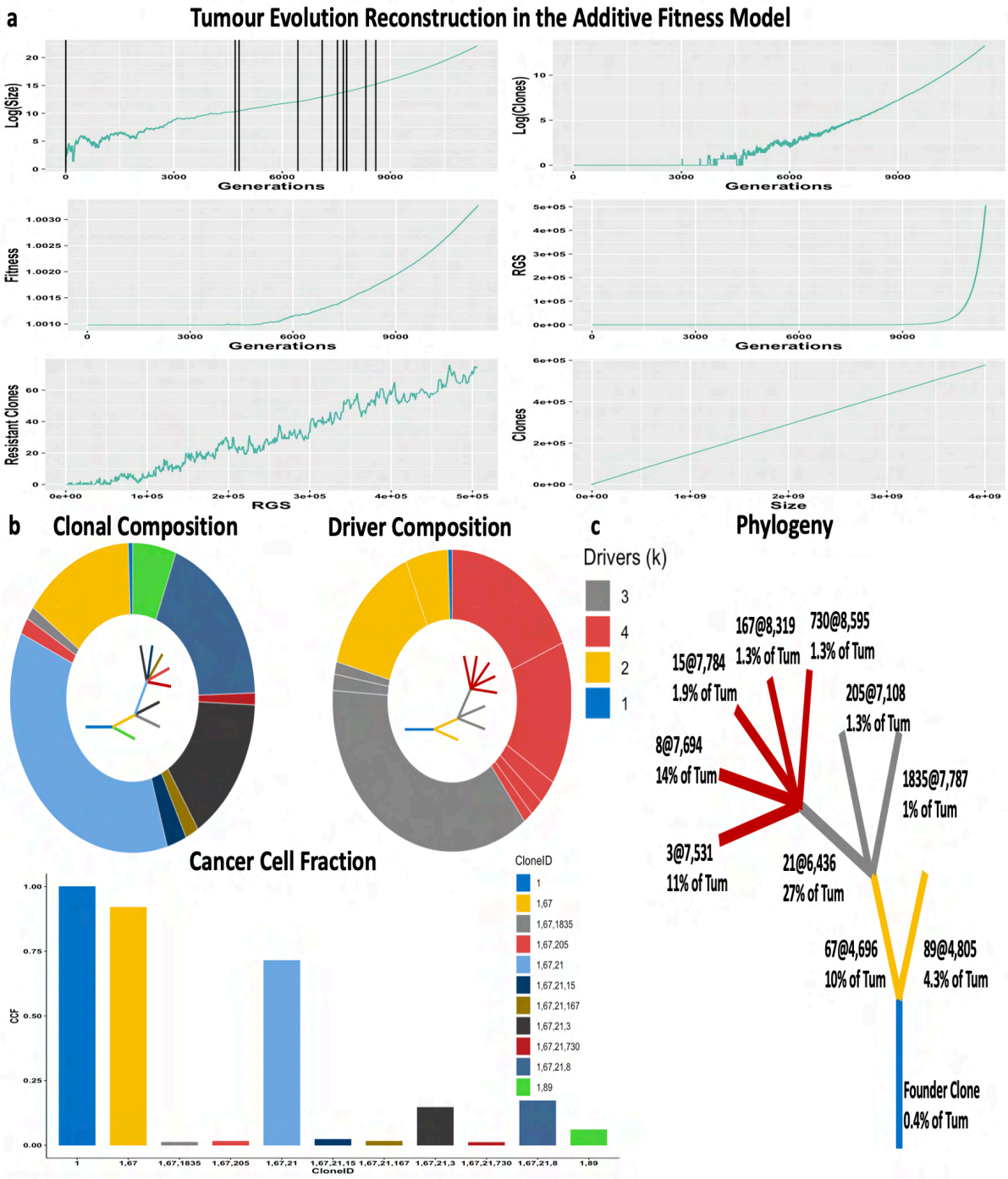


Figure 4.20 Top recurrent simulation in the additive fitness model. a, describes the tumour growth dynamics using the variables tumour size, generations, number of clones, fitness, RGS and resistant clones. Black vertical lines in the Log(Size) vs Generations indicate τ, the time when the detectable successful clones emerged and expanded. **b**, describes the clonal and driver composition as a pie chart. Colours in the clonal composition indicate individual clones and in the driver composition the driver abundance. The bar plot shows cancer cell fractions of all detectable clones colour coded as in the pie chart above. **c**, the phylogeny of driver clones for this simulation, colour coded by number of accumulated drivers k. Each clone is labelled with an ID number followed by @ with τ, the time of emergence of that clone. The proportion of the final tumour size represented by that clone is indicated underneath.

As observed in the TCGA samples, tumour clonal architecture is defined prior to achieving a clinically detectable size and the emergence of drug resistance clones occurs late in tumour development after generation 9,000. In this simulation, the tumour is driven by lineage 67 as a clonal branch that spawns multiple 4-driver clones coloured in red. The times of emergence of this lineage occurred around the expected times τ. This is a good example of the branched evolutionary mode representative of the additive fitness model.

The top increased mutation rate simulation, Figure 4.21, shows that detectable mutations appeared in a short window of 3,000 generations as a result of 2 driver hyper mutants. The effect of an increasing mutational burden in a short period is the appearance of a dominant four-driver clone, (clone 1, 23, 29, 7) in the phylogeny that comprises 23% of the tumour.

Both heterogeneity and number of clones are greater in the top additive fitness simulation, with clones detectable at the applied cut-off (5%) covering 52% of the tumour composition. As observed with the recurrent simulation in TCGA Figure 4.16, the number of resistant clones is higher than other models due to the replication stress induced by driver accumulation.

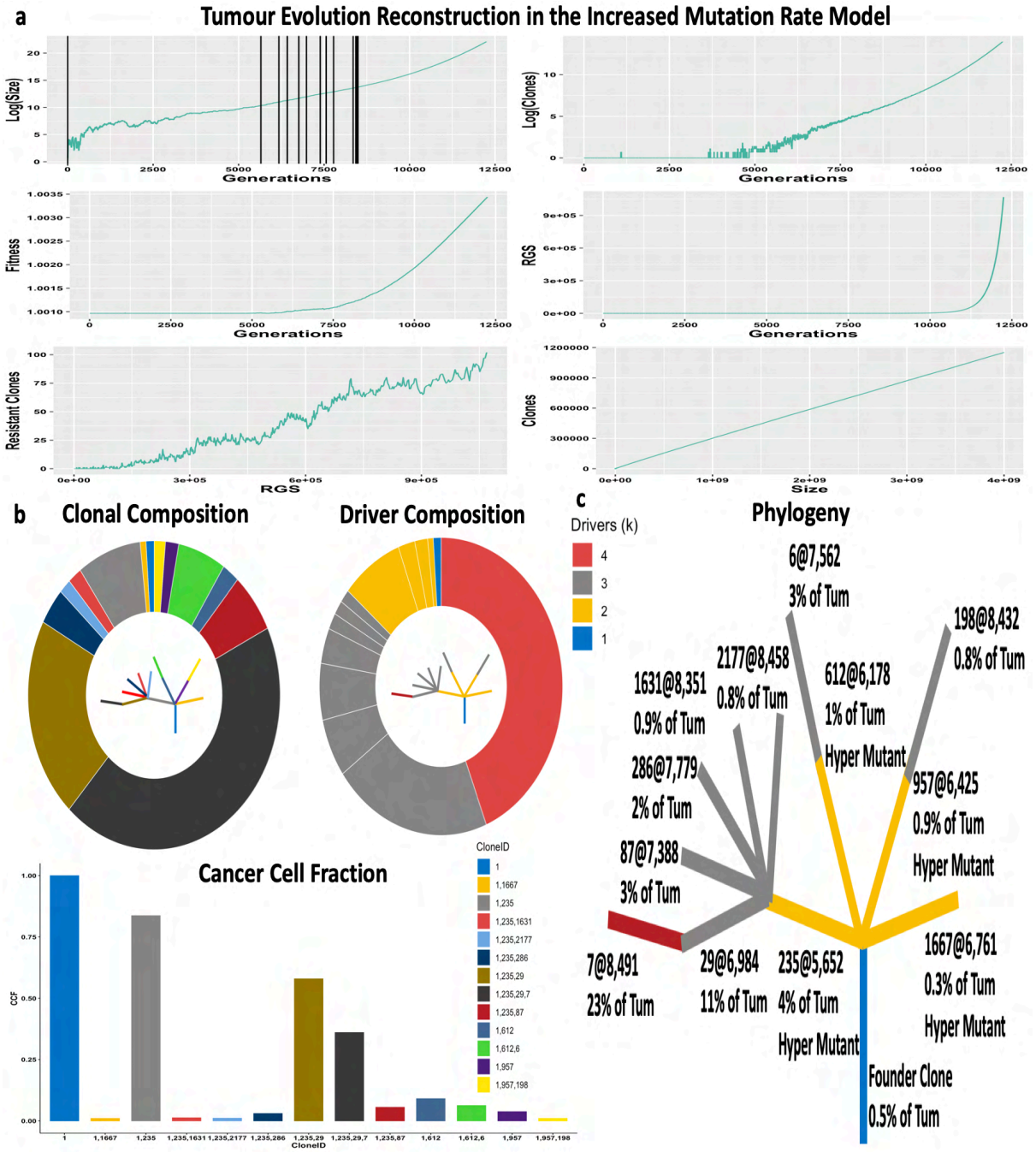


Figure 4.21 Top recurrent simulation in the increased mutation rate model. a, describes the tumour growth dynamics using the variables tumour size, generations, number of clones, fitness, RGS and resistant clones. Black vertical lines in the Log(Size) vs Generations indicate τ, the time when the detectable successful clones emerged and expanded. **b**, describes the clonal and driver composition as a pie chart. Colours in the clonal composition indicate individual clones and in the driver composition the driver abundance. The bar plot shows cancer cell fractions of all detectable clones colour coded as in the pie chart above. **c**, the phylogeny of driver clones for this simulation, colour coded by number of accumulated drivers k. Each clone is labelled with an ID number followed by @ with τ, the time of emergence of that clone. The proportion of the final tumour size represented by that clone is indicated underneath alongside whether the clone is hyper mutant or not.

An additional analysis done by TracerX group was the phylogenetic reconstruction for the 100 cases of the cohort. The phylogenies were constructed by averaging the regions and incorporating the unique number of PyClone clusters with copy number calls, resulting in hybrid phylogenies, TRACERx NSCLC supplemental material Figure S4.1.

I clustered TRACERx NSCLC topologies based on similarity, I identified 4 groups: single branched, symmetric, one-sided lineage and complex as shown in Figure 4.22. These groups can be defined by their mutational profiles. The single branched category is composed only of single nucleotide variations whereas branches in the symmetric and one-sided topologies categories are ~50% mixed with copy number alterations. In contrast, branches with the complex topology have a ~75% prevalence of copy number alterations.

I compared their reported topologies with the best fits from the 3 positive selection models to establish the degree of similarity for every patient. Phylogenies shown under 'Fitting Pattern' in Figure 4.22 were generated by the consensus of the best scoring fits of the 3 positive selection models for each category (single branch, symmetric, one-sided and complex). This helps to establish how copy number alterations influence the morphology of the phylogenetic tree of the best fit simulations.

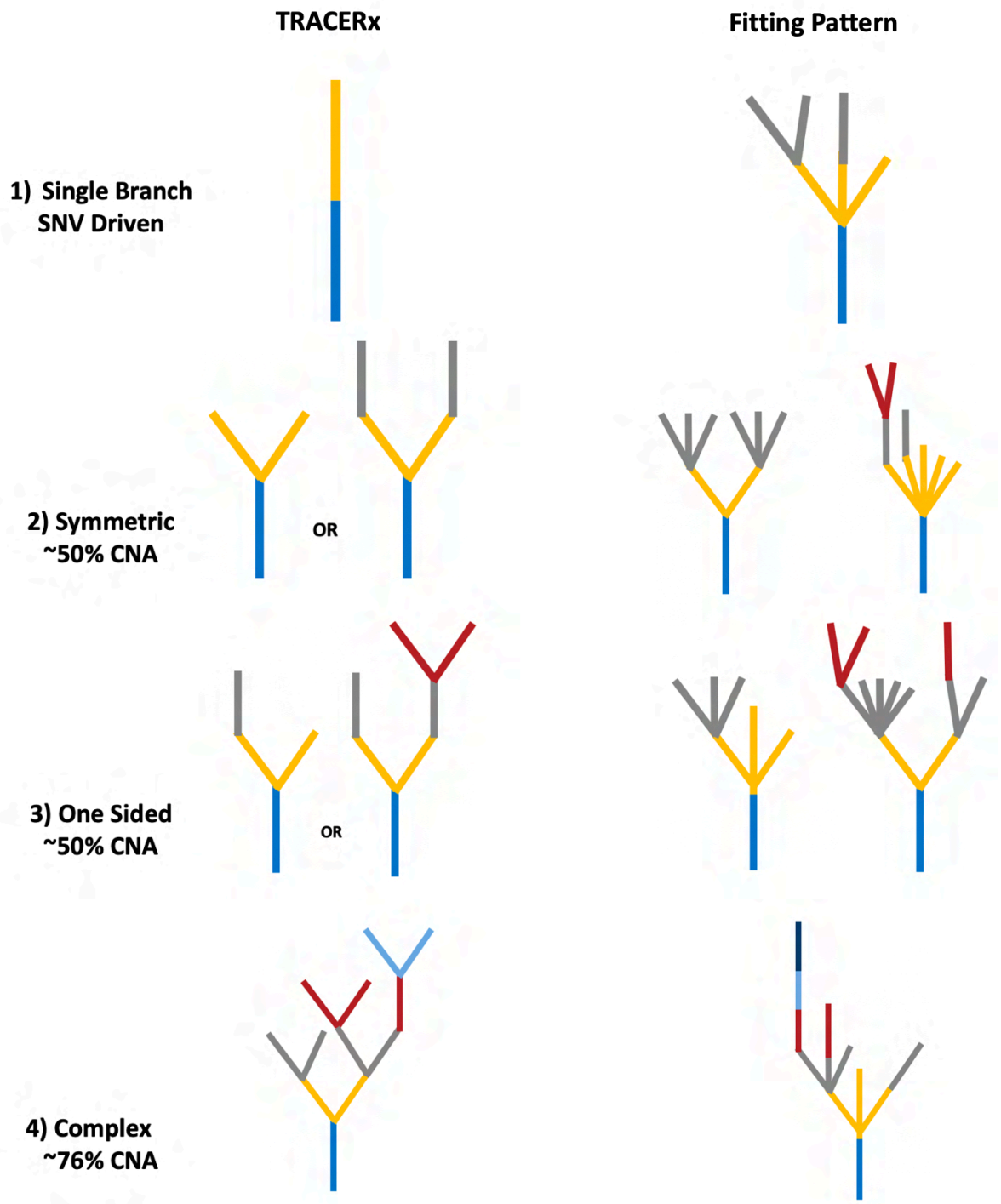


Figure 4.22 Approximation of TRACERx NSCL phylogenies based on topology. Left side shows the category and branch composition with the type of topology in TRACERx NSCL. Right hand side shows consensus topologies based on fits to the positive selection models. Phylogenies are colour coded by numbers of drivers accumulated along a given branch.

Overall, the trend of the fits is consistent with the global shape of the topologies in TRACERx NSCL. The number of branches seems to be proportional to the fraction of copy number, suggesting increased copy number is linked to high average driver mutation rate, hence I obtained fits to the high mutation rate simulations.

Single branch topologies were approximated as multi-branched by the models but with a strong prevalence of one lineage, these cases are ones with lower values of RGS diversity. Symmetric and one-sided topologies were more branched than expected, however the symmetric structure tended to be preserved. Complex topologies differ by 2-4 branches. Because a considerable number of clones were measured, false positive clones can be introduced by passenger tail or by the influence of copy number alterations with different selective advantage gains.

Potential reasons to explain the discrepancy between our model fits and the phylogenies inferred by TRACERx NSCL can be explained by a linear-to-branched evolutionary mode. In this mode, clonal sweeps occur in a truncal fashion until the fitness increase is substantial enough to branch out as seen in Figure 4.20. PyClone cannot recover these clonal sweeps as they get lumped into the same PyClone cluster resulting in a founder subpopulation with a higher value of fitness and k. This can explain the discrepancies in phylogenies for the single branching and symmetric cases as our approach of fitting to simulations can distinguish different clones emanating from the founder subpopulation that may be missed by PyClone.

Neutral samples (only one subpopulation measured) can bias the number of inferred clones, if multi-region samples contain only one clone. PyClone will be affected by the false positive clones in the passenger tail. When false positive clones are measured the number of potential phylogenies increases complicating the accurate recovery of the tumour phylogeny. This can affect all categories as it introduces noise into phylogenetic reconstruction.

A biological source of bias is hyper-selection, when fitness increases are higher than expected. TRACERx NSCLC is highly affected by both copy number alterations and whole genome doublings. Cytoband alterations and whole genome doublings can cause significant and rapid increases in fitness leading to a less predictable evolutionary dynamic. In this case the stickbreaking model can provide simulations that approximate this through their sampling of fitness effects, an example of which is displayed in Figure 4.15. Hyper-selection is expected to affect phylogenies in the symmetric, one-sided and complex categories most due to its increased composition in copy number alterations.

My model fits approximate the phylogenies in TRACERx despite the copy number alteration process not being explicitly modelled. The low frequency in recurrent topologies is caused by a more complete sampling of the clonal landscape revealing specific clonal trajectories. This is a function of the high-depth sequencing resolution, multi-region sampling and that predicted values of s and u lead to more heterogenous outcomes as shown in Chapter III. As mentioned previously, the *Big-Bang* evolutionary mode was inferred for at least 18% of the cases who were predicted to simulation illustrated in Figure 4.15. This evolutionary mode is suspected to affect cases with high copy number alterations leading to complex phylogenies described in Figure 4.22.

6.5 Association of Fits with Clinical Outcome in TRACERx NSCLC

This section is going to explore the relationship of the model fits with clinical outcomes, through integration and comparison of evolutionary variables with clinical and genomic data.

TRACERx NSCLC reported increased copy number and genome doublings are associated with an increased risk of death. However, the number of drivers or unique PyClone clusters reported showed no significant association with clinical outcome, Figures S4.2 and S4.3. Driver alterations and number of PyClone clusters in recurrences or deaths was were lower in the

responder group. Similarly, stratifying by clonal TP53 status did not show significant difference nor increased risk in clinical outcome as shown in Figure 4.23.

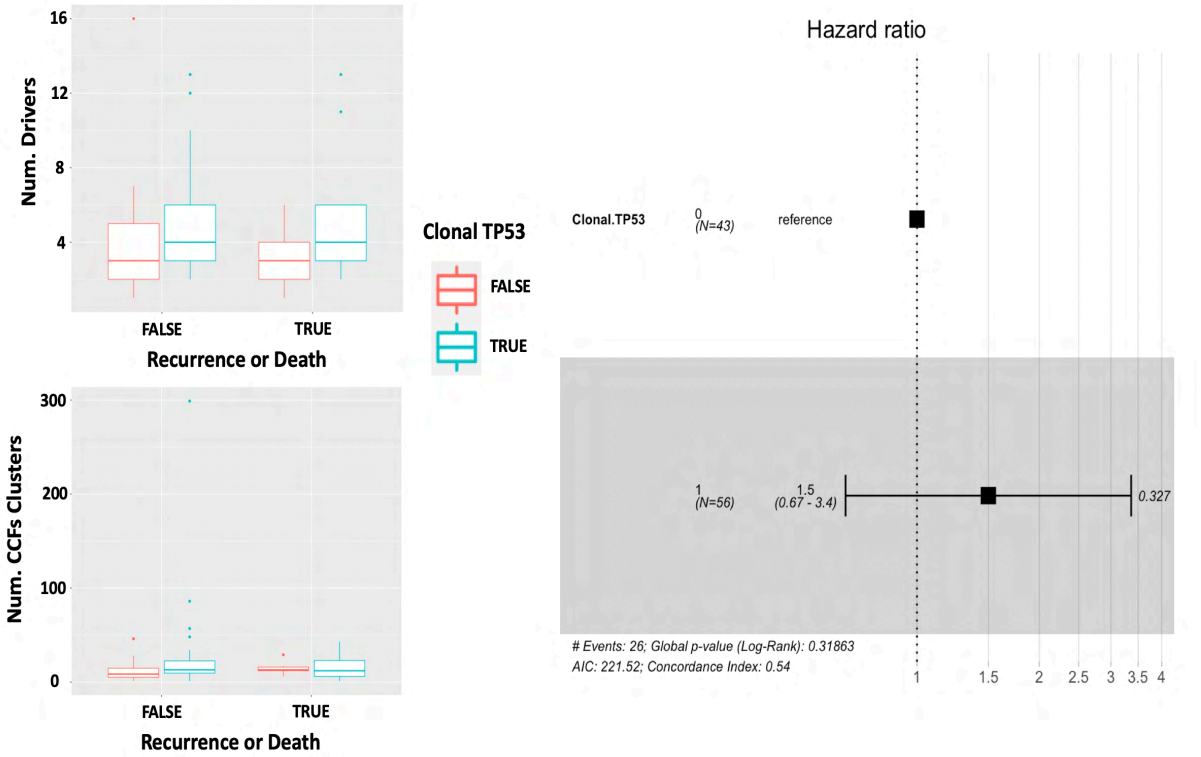


Figure 4.23 Association of clonal TP53 mutations with clinical outcome, the number of reported drivers and PyClone clusters in the TRACERx NSCLC cohort. Forest plot shows the Cox proportional hazards model of the presence or absence of TP53 mutations stratified by adjuvant treatment. p-value is next to upper confidence interval whisker.

It can be seen from Figure 4.23 that the number of drivers grouped by clonal TP53 status show similar distributions in responders and distant recurrences and deaths. The number of CCF clusters in recurrences or deaths likewise did not show a difference. The presence of TP53 mutations did not show significant association with recurrence or death as shown in the forest plot. The predicted tumour diversity as measured by RGS also did not differ between samples after stratification by clonal TP53 status and recurrence or death, and this metric was not associated with clinical outcome, Figure 4.24 panel b.

Similarly, no significant association were found for evolutionary variables such as the number of predicted clones, fitness and driver abundances or higher vs lower inferred mutation rates ($3.4\text{x}10^{-4}$ vs $3.4\text{x}10^{-5}$). This is expected given the clinical variables, number of drivers, number of PyClone clusters or clonal TP53 status were not significantly associated either.

Copy number alterations in TRACERx NSCLC were significantly associated with distant recurrence or death. As a result, the only marker that showed significant association with distant recurrence or death was whole genome doublings (reported in their study) that are not explicitly captured in the simulations. For this reason, it is expected that the fits to the simulations here do not have significant association with distant recurrence or death.

The stickbreaking and the increased mutation rate models showed the same trend, though the latter shows the strongest effect in the Cox proportional hazards model, Figure 4.24 panel b. Likewise, TP53 status does not have an impact on inferred RGS diversity, Figure 4.24 panel a, i.e., cases with a TP53 mutation do not show increased RGS diversity. This was not expected as Figure 4.23 showed that cases with TP53 increase the number of drivers.

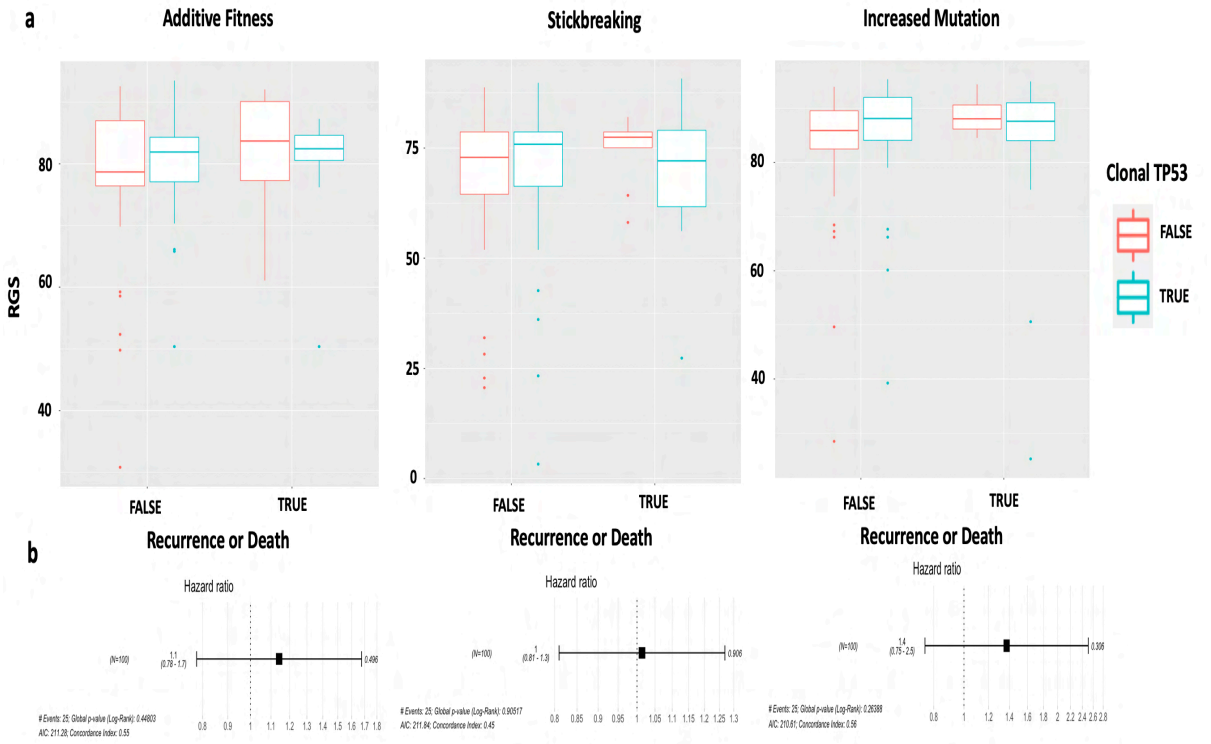


Figure 4.24 Association of RGS diversity with clinical outcome. **a**, boxplots of RGS diversity grouped by distant recurrence or death and clonal TP53 status for all models. **b**, Forest plot showing the univariate Cox proportional hazards for RGS diversity.

The model fits to the TRACERx NSCLC cohort are similar to those for the LUSC and LUAD TCGA cohorts, and fits of predicted topologies provide a global fit consistent with reported results However, no significant association between the inferred evolutionary properties could be made with the key clinical outcomes of recurrence or death.

6.6 Estimating Average Selective Advantage *s* and Average Driver Mutation Rate *u* in The Breast International Group 1-98

BIG 1-98 is a clinical trial of post-menopausal woman with hormone receptor positive HER2-negative breast cancer that aimed to compare the clinical benefit of adjuvant tamoxifen versus letrozole. The trial had a four-arm design: tamoxifen alone, letrozole alone or with sequences of 2 year of one drug followed by 3 years of the other. The study enrolled more than 8,000 women. As part of the study, amplicon sequencing of 287 driver-genes was performed by Foundation Medicine on formalin fixed paraffin embedded samples from 538 patients, including 140 distant recurrence events.

Single nucleotide variant and insertion/deletion calls provided by Foundation Medicine were used in the analyses as described here [198]. Variant calls were cross-validated using VarDict [197]. Genome wide copy number calling was performed using off-target reads with

winsorization [201, 202] for outlier removal followed by circular binary segmentation using CopywriteR [200]. The copy number calls were used for ploidy correction during clonality analysis in ExPANdS and PyClone.

BIG 1-98 represents a challenge for reconstructing tumour evolution and establishing probable phylogenies due to it being restricted to single biopsy samples, the limited number of genes sequenced, and noise introduced by FFPE-induced DNA degradation. However, amplicon sequencing should still detect and enrich for the main driver mutations. As driver frequencies are the main element used for fitting to simulation results, it is expected that it should be possible to obtain an approximation of the evolutionary history of many BIG1-98 patients through our approach.

Previous analysis of BIG 1-98 to identify the molecular drivers of the disease showed that amplifications in 11q14 and 8p11 associate with a significant increase of distant metastasis [199], suggesting clinical outcomes in this breast cancer subtype are influenced by CNAs. The following figure shows the mutational prevalence of recurrent alterations in both cohorts. Complete mutational frequencies can be found in supplemental Figures S4.4 and S4.5.

To corroborate the prevalence of copy number events I turned to data from the Molecular Taxonomy Group of Breast Cancer (METABRIC) study [207], which established the genomic stratification in breast cancer in a study comparison of 1,643 cases. I selected copy number profiles from the post-menopausal receptor positive HER2-negative cases, Figure 4.25.

Although death was the chief endpoint in METABRIC but distant recurrence was used in BIG 1-98. Both studies showed pervasive influence of genome wide copy number alterations with significant association in clinical outcome (distant recurrence or death). The effect of genome wide copy number alterations was superior to effect of the accumulation of SNVs. This explains why in BIG 1-98 no individual genes were significantly associated with distant recurrence in weighted Cox models [199]. In addition, TP53 is the gene enriched for mutations in both studies, which highlights the role of genomic instability in this subtype.

To further explore the role of genome-wide coy number alterations in this cohort, I used the weighted genome wide integrity index (wGII) [196], a weighted metric used to summarise the fraction of the genome altered based on segmented copy number data. wGII showed classification power to discriminate the cohort based on clinical outcome in both studies as indicated by the Kaplan-Meir estimator on clusters of accumulated genome-wide alteration, Figure 4.26.

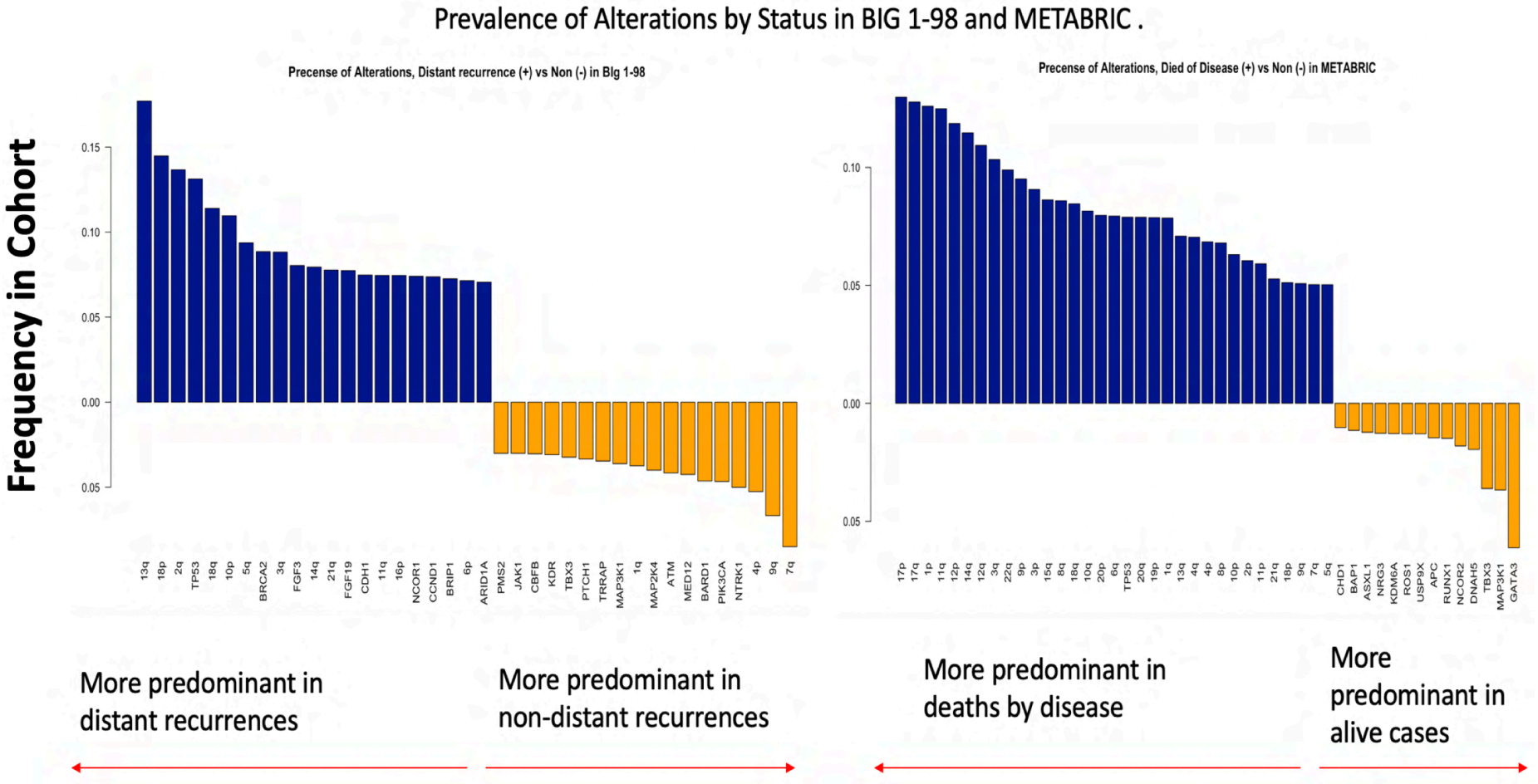


Figure 4.25 Top frequent alterations in BIG 1-98 and METABRIC. Copy-number alterations more frequent distant recurrences in BIG 1-98 or deaths in METABRIC are shown in blue and the CNAs more prevalent in responders shown in orange. Frequencies represent the fraction of the total samples in the indicated category (responders vs distant recurrences or death).

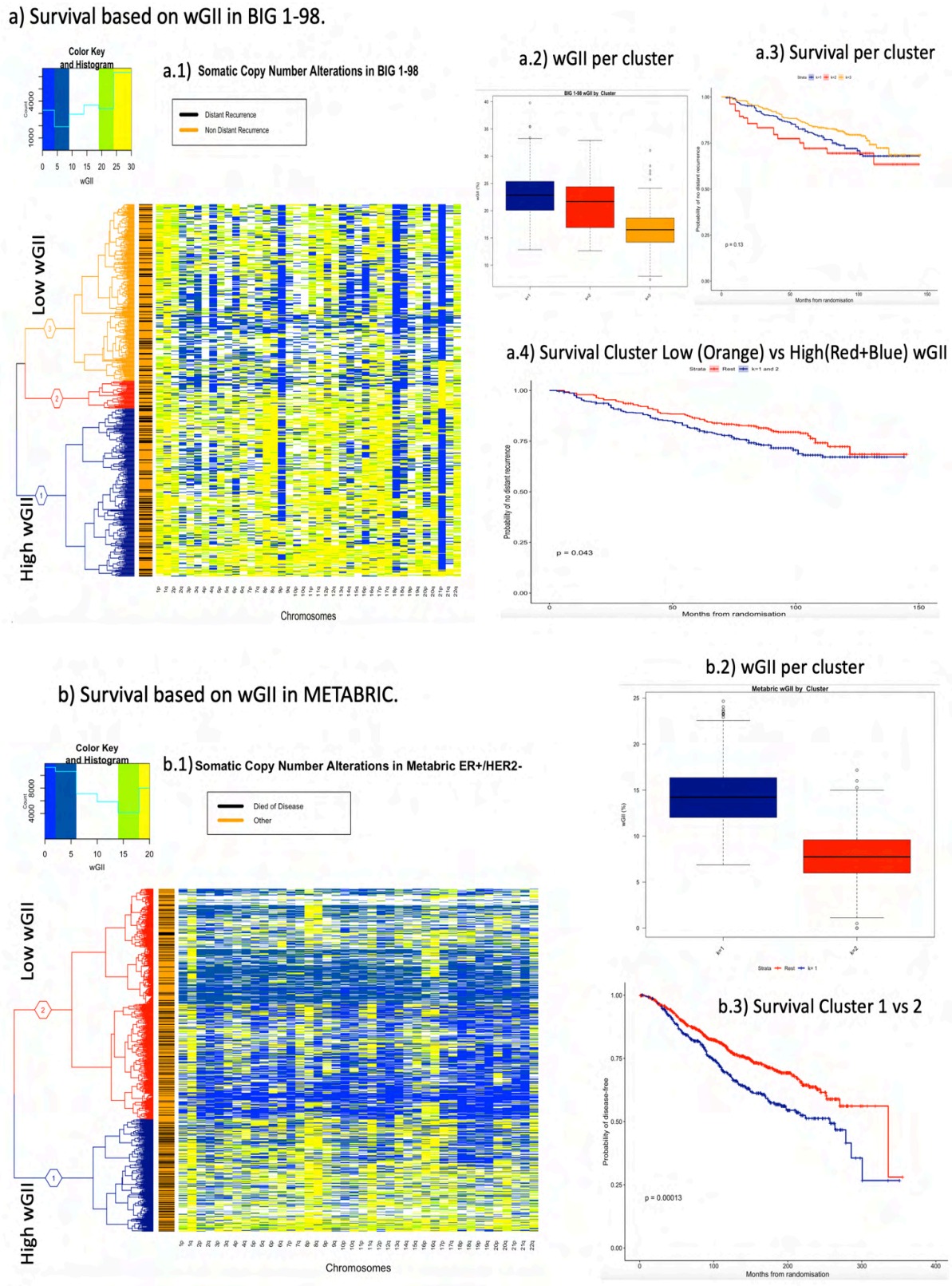


Figure 4.26 Association of wGII with clinical outcome in BIG 1-98 and METABRIC. Heatmaps display the wGII for each chromosomal arm as percentages, coloured by gradient, blue (low) to green/yellow (high). Boxplots show the wGII distributions of the clusters identified in the dendrograms. Kaplan Meier curves evaluate the survival according to cluster membership.

The wGII of a sample showed significant association with clinical outcome in both studies in univariate and multivariate weighted Cox models as shown in Table 4.3. All cytobands were evaluated for significance as shown in Supplementary Tables 4.2, 4.3, 4.4 and 4.5.

Table 4.3. Cox Proportional Hazards in BIG 1-98 and METABRIC

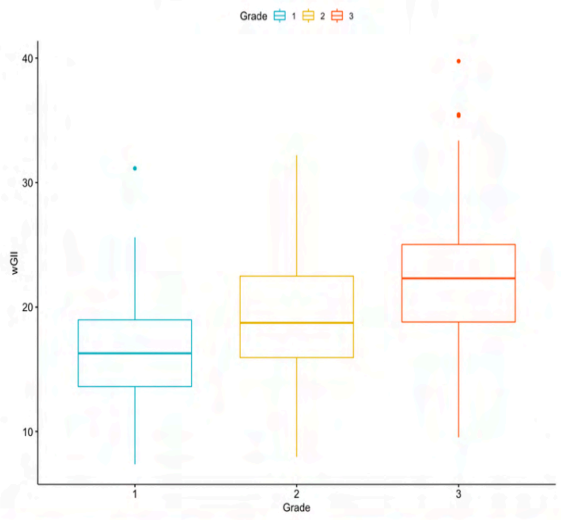
Study	*wGII HR (95%CI)*	*p-value*	*zph*
Univariate BIG -98	1.06 (1.02 – 1.09)	<0.001	0.245
Multivariate BIG 1-98	1.05 (1.01 – 1.09)	0.02	0.767
Univariate METABRIC	1.06 (1.04 – 1.09)	<0.001	0.461
Multivariate METABRIC	1.03 (1.00 – 1.09)	0.02	0.46

Note: zph is a test the proportional hazards assumption for each covariate and the column shows its p-value.

In addition, wGII was significantly associated with tumour grade in both studies as shown in Figure 4.27, with grade increasing proportional to wGII. In BIG 1-98 the wGII additionally associated with the number of nodes, TP53 status, number of copy number calls, shown in Figure 4.28. These variables were significant in Cox models, as along with Ki-67 and tumour size, which had weak correlation with wGII ($\rho = 0.34[0.26 - 0.42]$ & $\rho = 0.11\ [0.027 - 0.20]$). These clinical annotations were not reported in METABRIC.

Association of wGII with Clinicopathological Factors

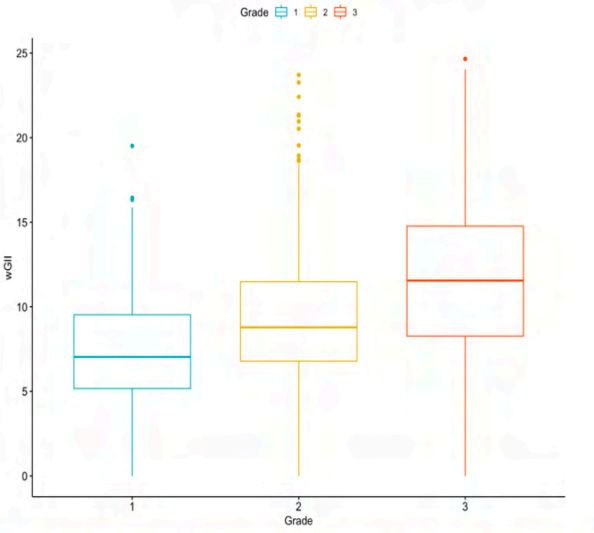
BIG 1-98: wGII vs Grade



Pairwise Comparison using Wilcoxon		
	Grade = 1	Grade = 2
Grade = 2	<0.001	-
Grade = 3	<0.001	<0.001

KW: $\chi^2_2 = 74.60$; P<0.001

Metabric: wGII vs Grade



Pairwise Comparison using Wilcoxon		
	Grade = 1	Grade = 2
Grade = 2	<0.001	-
Grade = 3	<0.001	<0.001

KW: $\chi^2_2 = 123.75$; P<0.001

Figure 4.27 Association of wGII with tumour grade in BIG 1-98 and METABRIC.

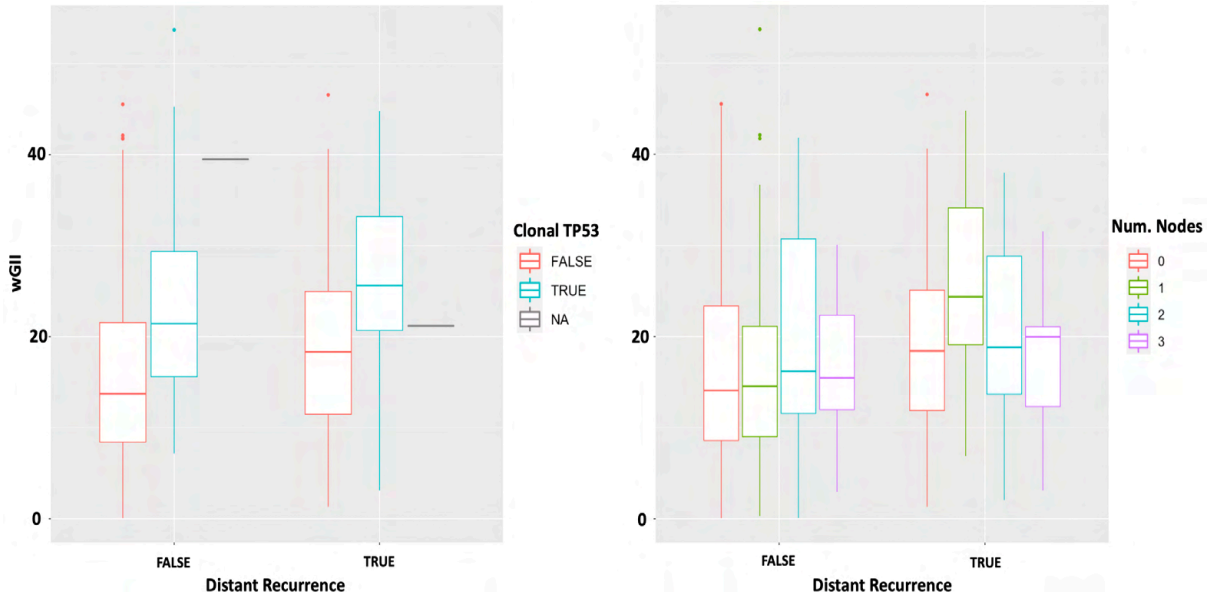


Figure 4.28 Association of wGII with Clonal TP53 and Number of Nodes in BIG 1-98.

The number of SNV drivers reported by Foundation Medicine in BIG 1-98 was not correlated with the wGII. Therefore, copy number alterations may play a larger role in driving tumour evolution in BIG 1-98, resulting in the following implications for reconstructing tumour evolution,

1. Genome wide copy number alterations driving the evolution of some of the tumours in this cohort may not have been detected by amplicon sequencing, thus skewing the fitting procedure toward incorrect predictions.
2. Although the reported mutation rate for CNAs is around $1\text{x}10^{-3} - 1\text{x}10^{-4}$, the rate at which driver CNAs appear is not known.

Based on available variant allele frequencies for every patient, I evaluated neutral evolution by consensus using the approach of Williams et al. with *neutralitytesr* tool [88]. The tool measures the prevalence of the passenger tail and provides four neutrality metrics: area under the curve, Kolmogorov distance, means distance and R^2. If one of the 4 statistics indicated neutrality, the sample was considered to be neutral.

As a result, 56% of the cases in BIG 1-98 showed evidence of having only one detectable driver subpopulation. The limitation in amplicon sequencing is that some of the 56% neutral samples may not be neutral, but the amplicon sequence data was not able to pick up any signals of positive selection. Nevertheless, the number of cases with evidence of positive selection carried forward for model fitting was 237, with 65 having distant recurrence events.

Next, I applied the MDA.N comparison technique to the ExPANdS and PyClone clonality distributions from these 237 samples. To match to the depth of the sequencing assay, simulated clones with CCFs below 5% were removed. The most likely fits covered the same combination of parameters observed in the other cohorts. That is, moderate/weak average selective advantage with high driver mutation rate, $s = \{0.001, 0.01\}$ & $u = \{3.4\text{x}10^{-5}, 3.4\text{x}10^{-4}\}$, with >90% of tumours matching to a simulation with $s = 0.001$. Again, predictions from both ExPANdS and PyClone gave similar results.

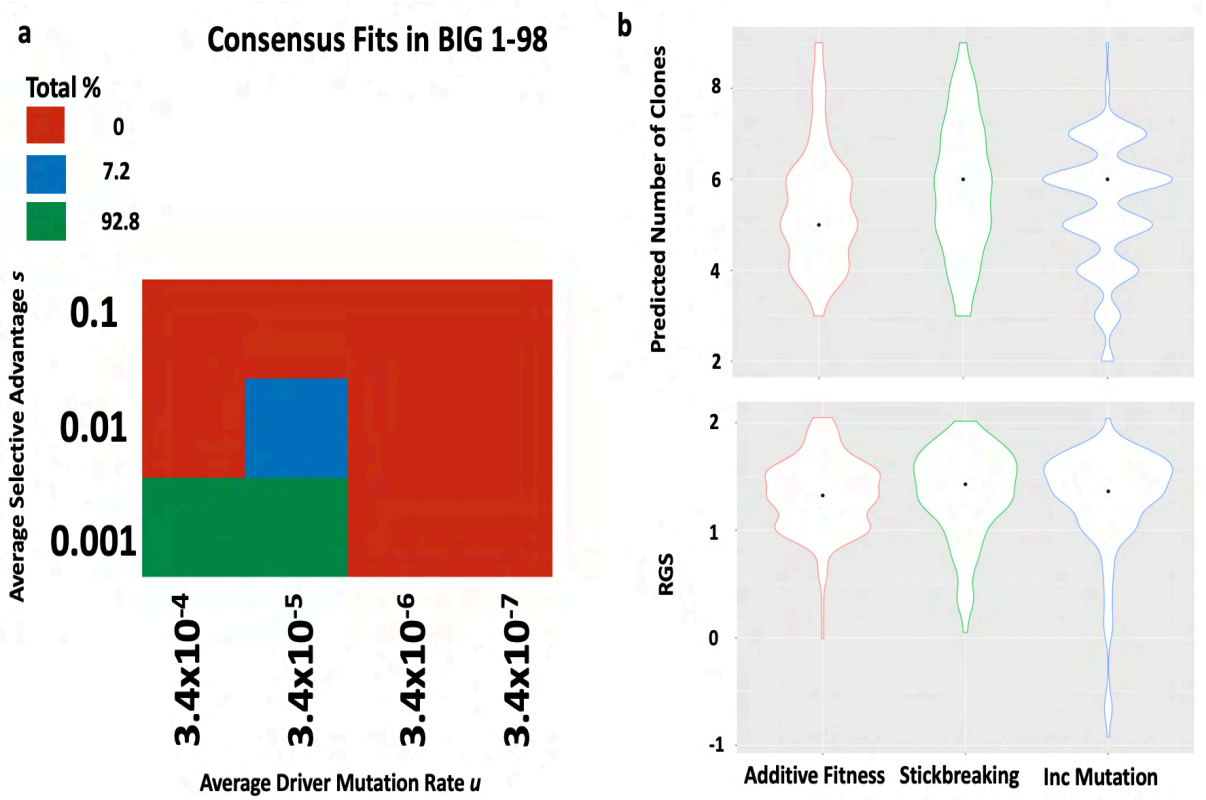


Figure 4.29 Fitting results in BIG 1-98. a, heatmap of consensus fits of all subtypes using both clonality tools. **b**, Distribution of number of detected clones and diversity RGS for every model.

The effects of amplicon sequencing translate into high clonality but low RGS diversity. This is because more clones can be detected at a ≥5 % CCF and false positive calls are unlikely to occur. The sequencing coverage of 500 genes in the panel is unlikely to have a prominent passenger tail. This was observed in PyClone or ExPANdS clusters with few clusters that were filtered out as per the tool's guideline. The fact that the fits are in the same likelihood as in TCGA and TRACERx, indicates that the MDA.N technique can fit to amplicon sequencing data.

I next evaluated whether the obtained fits correlate with the mutation calls made by Foundation Medicine, for cross-validation. The 500 genes in the panel were selected for their roles in cancer and thus are expected to influence disease progression to some extent. Indeed, most of the altered genes showed evidence of positive selection using the maximum likelihood dN/dS method [70](performed with the tool *dndscv* available in R).

For quality control, I correlated the predicted number of driver mutations in the best fit simulations for each sample with the observed number of driver point-mutation calls, copy number calls and the sum of both. It is expected that the higher the correlation the better the quality of the fit. However, the additive fitness was uncorrelated with the mutational patterns in BIG 1-98 and the stickbreaking model showed weak correlation with SNVs, whereas the increased mutation model showed apparent correlation with all calls, as shown in the following table.

Table 4.4 Correlation of Predicted vs Observed Drivers in BIG 1-98

Model	*SNV Calls*	*CNA Calls*	*All Calls*
Additive fitness	-0.024	0.047	0.0129
Stickbreaking	0.153*	-0.046	0.089
Increased Mutation	0.117*	0.145*	0.194*

* showed Pearson correlation coefficient $p < 0.05$.

The fits in BIG 1-98 showed consistency with the average selective advantage and average driver mutation rate observed in the previous studies. The wGII is significantly correlated with clinical outcome and also associated with the clonal TP53 status, suggesting is associated in driver mutation rate. This may explain why fits to the increased mutation rate model better align to the mutational calls from BIG 1-98. Thus, results from fits to the increased mutation rate model were used in downstream association analyses.

6.7 Validation of Predicted Fits Using Neutral Cases in BIG 1-98

The majority of the cases in BIG 1-98 showed evidence of neutrality. In these cases, the passenger signal is enriched because only one subpopulation was measured at the CCF cut-off of 5%. The passenger tail can be used to infer the likely values of s, u and k.

The passenger signal is proportional to the number of divisions that occurred in a clone, this can be calculated with the expectation as $\frac{\nu C_{k,i,j}}{(1-\delta_{k,i,j})}$ as suggested by Bozic et al. [58]. To support the parameters observed in Figure 4.29 it is possible to test the likelihood of observing a given passenger tail for combinations of s and k. This can be achieved by minimising the least squares error to the median number of passengers reported by Bozic et al. [58] with the passenger tail as shown in Figure 4.30.

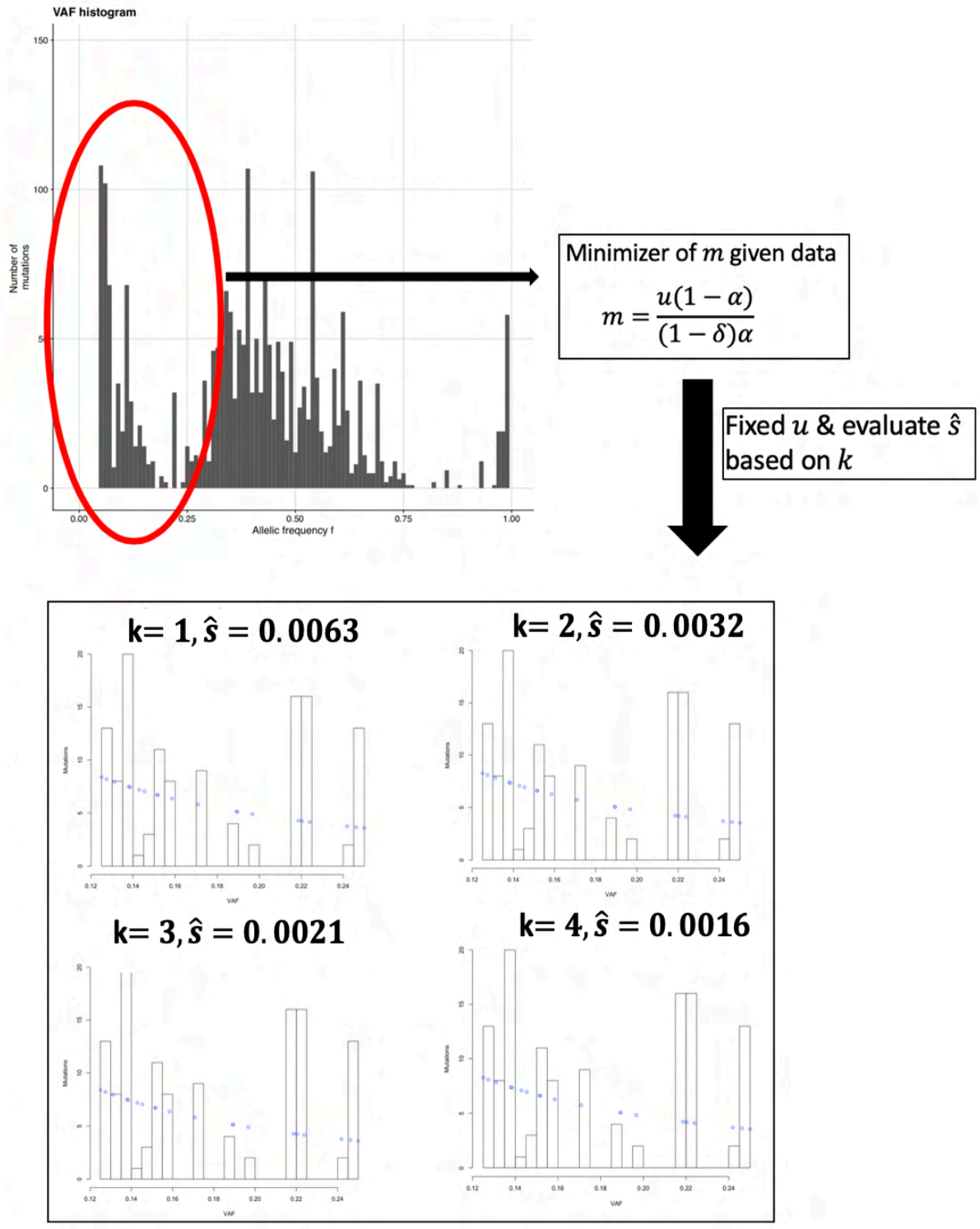


Figure 4.30 Fitting neutral cases in BIG 1-98. The neutral tail is extracted from the VAF distribution within a range from 0.1 to 0.3, subject to vales of u, k and s to identify the best fit that minimises the mean square error given the data.

Neutral cases support the fits identified by the positive selection models, as shown in Figure 4.25, with the range of the predict $s = \{0.0063 - 0.0016\}$ compatible with the finding $s = 0.001$ was the best fit for the positive selection models.

6.8 Recurrent Phylogenies and Clonal Evolution Reconstruction in BIG 1-98

Next, I evaluated the recurrent topologies in the positive selection cases of the cohort. To replicate the depth of targeted sequencing, clones below 5% CCF were removed. I identified recurrent topologies at low frequency in all models with similar shapes as shown in in Figure 4.31.

The top topologies of the distant recurrence cases are flagged as NR (non-responder). While these intersect with the topologies of responders, topologies of NR cases tend to be more branched which suggests fitness differences between responder and non-responder groups.

To compare the topologies obtained in Figure 4.31 with available tools, I used ClonEvol, TrAP, PhyloWGS on measurements of CCFs from PyClone and ExPANdS. Showing mostly single branch phylogenies (~93% of the patients), missing the branching nature of some of the samples in particular the NR. This is expected due to the lack of multi-region sampling and reduced coverage of the genome. Our approach bypasses this problem by allowing the branching process to infer the evolutionary history when data are limited or missing.

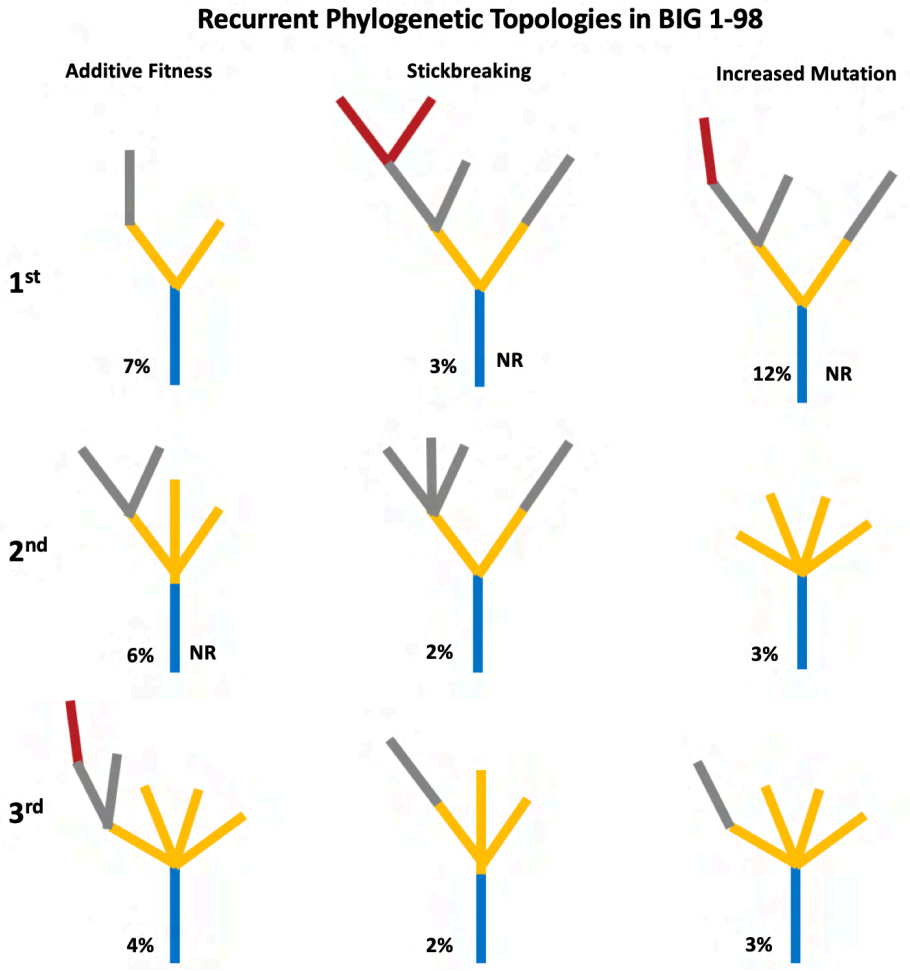


Figure 4.31 Recurrent phylogenies in BIG 1-98. Phylogenies are ranked by frequency and percentages indicate the prevalence of the topology; NR indicates the topology is present in samples from patients with distant recurrences. Branches are coloured according to the number of accumulated drivers. Blue, yellow, grey and red represent clones with 1, 2, 3 and 4 drivers, respectively.

The top recurrent simulation from the additive fitness model occurred in 7% of the BIG 1-98 samples that had undergone positive selection. This simulation was run with a moderate selective advantage, $s = 0.01$ and a high average driver mutation rate, $u = 3.4\text{x}10^{-5}$, Figure 4.32. This the first top recurrent simulation with a selective advantage greater than $s = 0.001$. The age of the tumour is around 5.4 to 27 years, based on division rates of 1 to 5 days per cell division. This simulation occurs exclusively in the responders group, hence the low degree of branching in its phylogenetic tree.

The increased average selective advantage resulted in faster tumour growth and reduced RGS diversity due to the reduced total number of cell divisions. The tumour is mostly composed of a 3-driver clone. In total, all detectable clones greater than 1% CCF made up 80% of the overall tumour composition. The founder clone was still present, representing 8% of the tumour. This simulation suggests that patients who responded to therapy had tumours with higher proliferation rates.

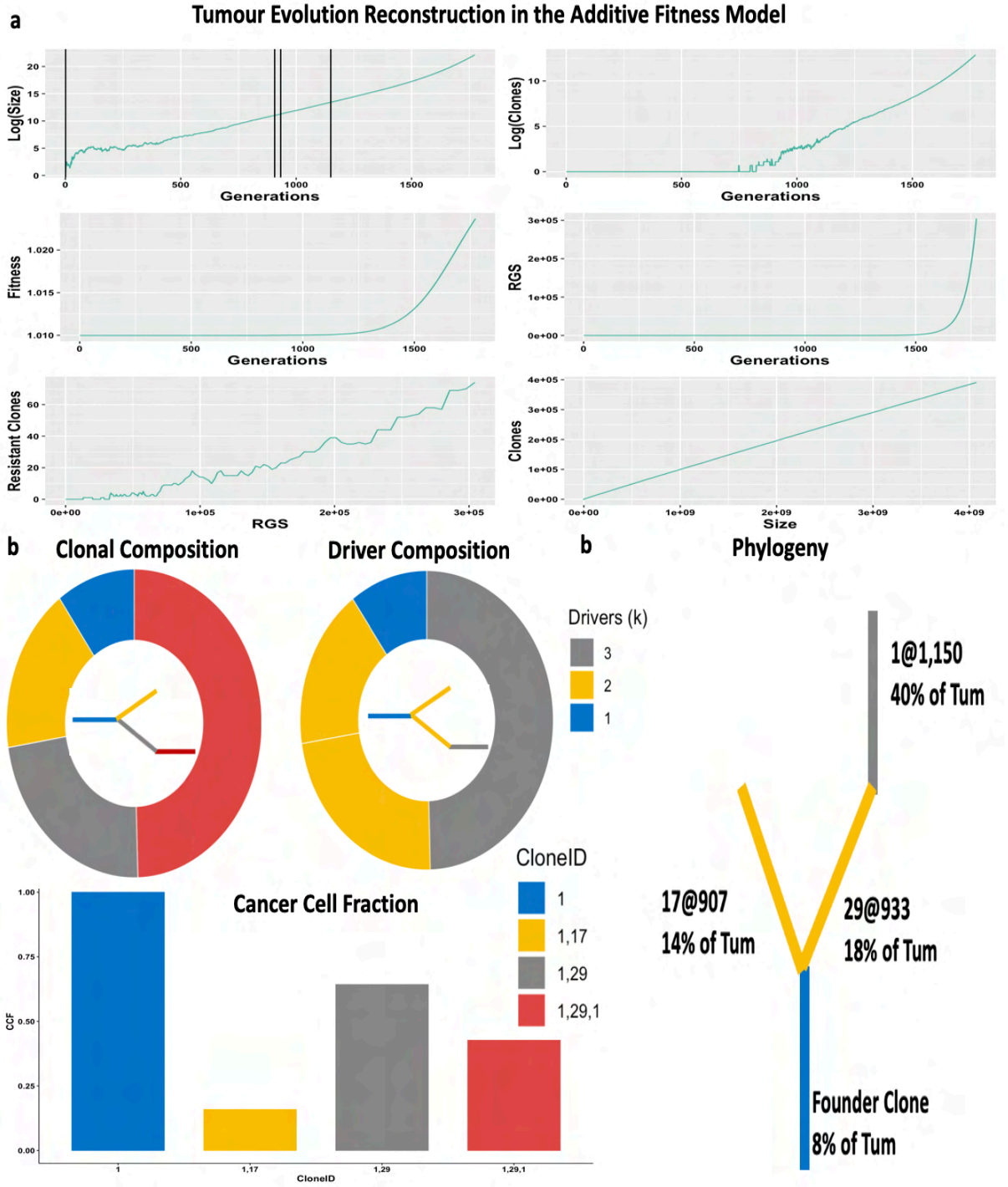


Figure 4.32 Top recurrent simulation in the additive fitness model. a, describes the tumour growth dynamics using the variables tumour size, generations, number of clones, fitness, RGS and resistant clones. Black vertical lines in the Log(Size) vs Generations indicate τ, the times at which the detectable successful clones emerged and expanded. **b**, describes the clonal and driver composition as a pie chart. Colours in the clonal composition indicate individual clones and in the driver composition the driver abundance. The bar plot shows cancer cell fractions of all detectable clones colour coded as in the pie chart above. **c**, Phylogeny of driver clones for this simulation, colour coded by number of accumulated drivers k. Each clone is labelled with an ID number followed by @ with τ, the time of emergence of that clone. The proportion of the final tumour size represented by that clone is indicated underneath.

The most frequent simulation in the stickbreaking model, Figure 4.33, was fit to 5% of all samples in total and to 9% of the patients with distant recurrences. It had starting parameters $s = 0.001$ and $u = 3.4\text{x}10^{-4}$. This simulation is the same as the top-recurrent topology in the additive fitness model as shown in Figure 4.31 but at a different time-point.

Here, the most abundant clone is a 4-driver clone comprising 40% of the tumour, which explains why clonality tools recover a single branched topology. In this simulation the most frequent clone (clone 2,882) has high probability of being sampled by sequencing. If an additional subpopulation is sampled then a plausible inference is a single branched topology but if not a neutral mode of evolution would be inferred. For instance, if clones 2,882 and 1,112 are sampled, 3 cellular cell fractions are going to be recovered, two for the sampled clones and one at 100% which is the shared ancestry. Thus, two topologies can be recovered, a *Y* shape or a single branch.

With a generation time of ~16,000 this tumour should have a fast division rate, 1-2 days maximum, which would take 43 – 87 years to grow. A division rate of 3 days or more for tumour would lead to an implausibly long timeframe.

In this simulation most of the surviving driver mutations occurred within a small window of ~2,000 generations, boosting the fitness and diversity of the tumour, accelerating its expansion and consequently increasing the number of potentially drug resistant clones.

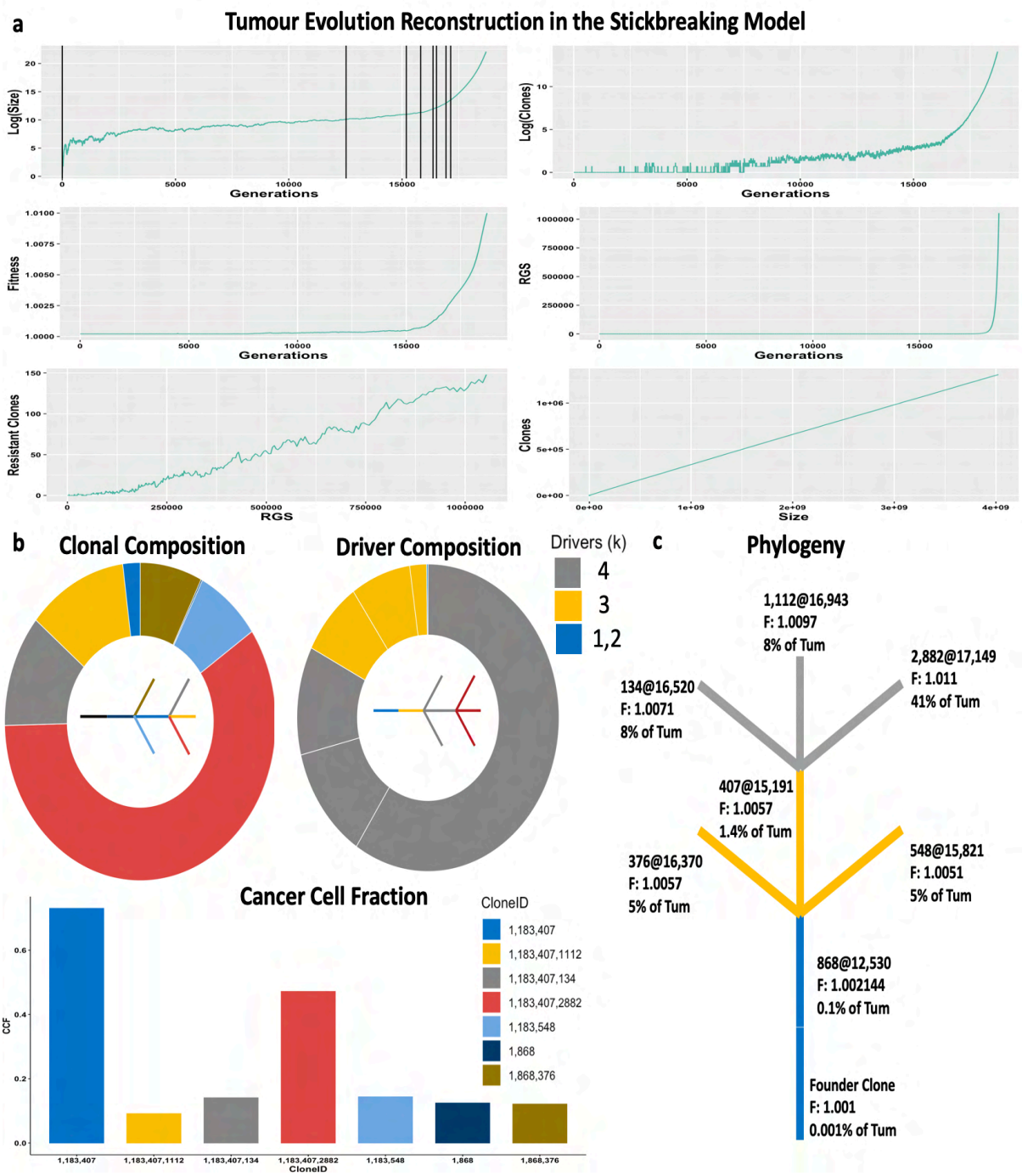


Figure 4.33 Top recurrent simulation in the stickbreaking model. a, describes the tumour growth dynamics using the variables tumour size, generations, number of clones, fitness, RGS and resistant clones. Black vertical lines in the Log(Size) vs Generations indicate τ, the time when the detectable successful clones emerged and expanded. **b**, describes the clonal and driver composition as a pie chart. Colours in the clonal composition indicate individual clones and in the driver composition the driver abundance. The bar plot shows cancer cell fractions of all detectable clones colour coded as in the pie chart above. **c**, Phylogeny of driver clones for this simulation, colour coded by number of accumulated drivers k. Each clone is labelled with an ID number followed by @ with τ, the time of emergence of that clone. The proportion of the final tumour size represented by that clone is indicated underneath with their fitness (F).

For the increased mutation rate model, the most recurrent fit matched to 13% of BIG 1-98 cohort, including 9% of the patients with distant recurrences, as shown in Figure 4.34. This simulation was run with parameters $s = 0.001$ and $u = 3.4\text{x}10^{-5}$.

This simulation resulted in a high degree of intratumoral heterogeneity, with detectable clones comprising just under half (49.2%) of the whole tumour. It showed a branched topology with the dominant clone having a frequency of 18% in the tumour.

The fitness of this tumour at size 4 cm^3 had increased to the 3 times its starting value, due to the abundance of 3- and 4- driver subpopulations. Fitness increased gradually after generation 6,000 resulting from the emergence of the first 3 clonal lineages that survived until the end of the simulation. However, the RGS diversity increased considerably after 9,000 generations, indicating that the process hypermutants seeding new driver progeny, increasing the probability of drug resistant clones emerging.

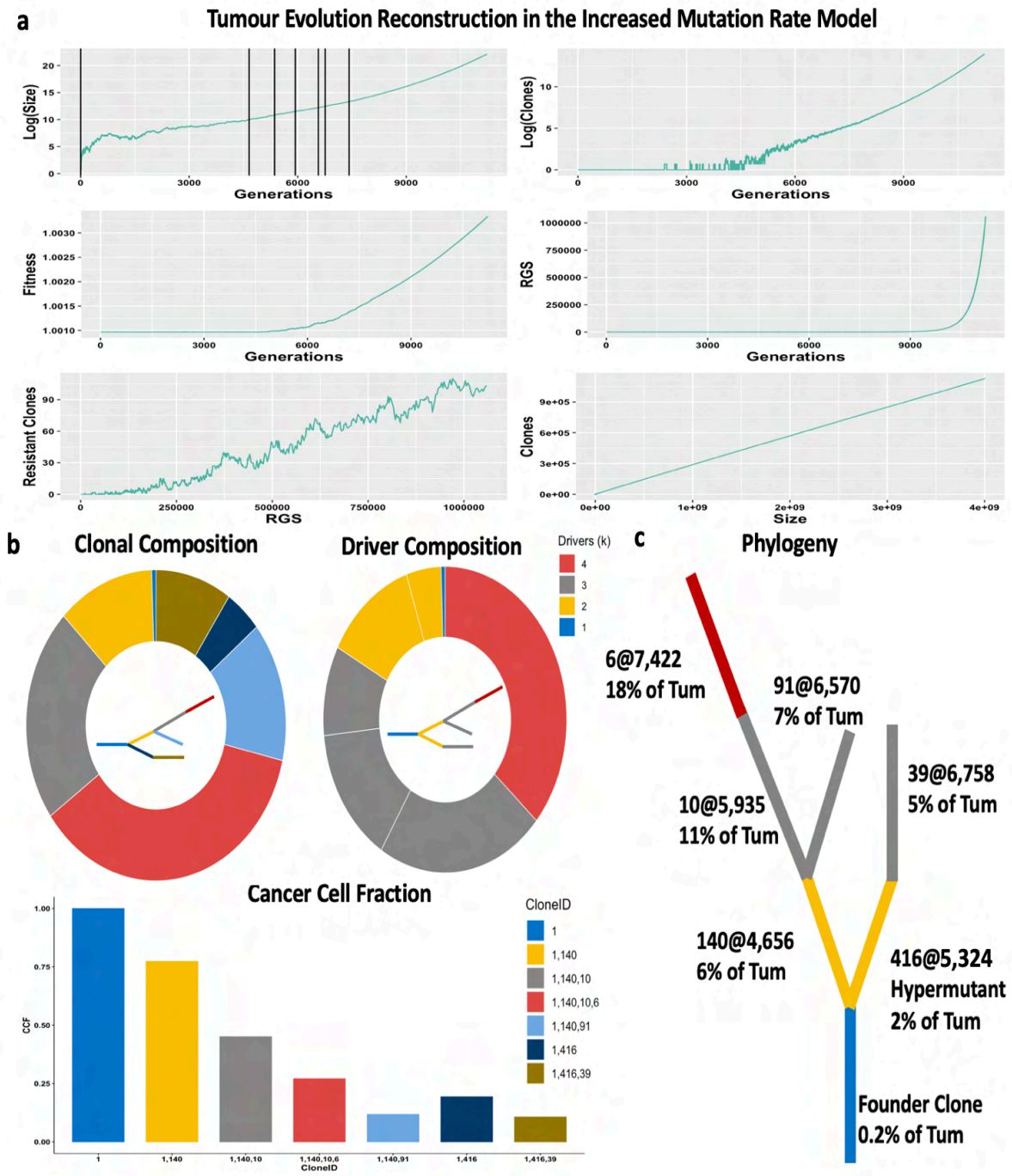


Figure 4.34 Top recurrent simulation in the increased mutation rate model. a, describes the tumour growth dynamics using the variables tumour size, generations, number of clones, fitness, RGS and resistant clones. Black vertical lines in the Log(Size) vs Generations indicate τ, the time when the detectable successful clones emerged and expanded. **b**, describes the clonal and driver composition as a pie chart. Colours in the clonal composition indicate individual clones and in the driver composition the driver abundance. The bar plot shows cancer cell fractions of all detectable clones colour coded as in the pie chart above. **c**, the phylogeny of driver clones for this simulation, colour coded by number of accumulated drivers k. Each clone is labelled with an ID number followed by @ with τ, the time of emergence of that clone. The proportion of the final tumour size represented by that clone is indicated underneath with the status if the clone is hyper mutant or not.

The topologies generated provide a better depiction of clonal evolution than the ones recovered by common clonality estimation tools. Branched topologies were 50% more prevalent in samples from cases that had distant recurrence, consistent with both a higher average driver mutation rate $u = 3.4\text{x}10^{-4}$ in the stickbreaking model fits and the correlation of number Foundation Medicine calls with predictions from the increased mutation rate model.

The stickbreaking and increased mutation rate models showed increased number of drug resistant clones compared to the additive fitness model, which indicates the role of increased genomic instability and emergence of drug resistance in distant recurrence cases whose primary tumours more closely follow the characteristics of these models.

6.9 Association of Fits with Clinical Outcome in BIG 1-98

With predicted topologies established for each patient, the next step is to associate the patient fits with distant recurrence using weighted Cox proportional hazards models. However, as mentioned tumours from patients with distant recurrences tend to have higher fraction of genome (wGII). Therefore, it is relevant to first explore if wGII is associated with mutational status in TP53 and PIK3CA to determine whether the presence of pathogenic mutations in these should be accounted for in the survival analysis.

Point mutations were selected to by high quality and call and precited deleterious (non-synonymous) by Condel score using the Foundation Medicine calls cross validated with VarDict. Then corroborated the deleterious effect using the maximum likelihood dN/dS method (performed with the tool *dndscv* available in R). Copy number alteration reported by Foundation Medicine had a significant prevalence in the cases with log_2 rations greater than > 1.

As Figure 4.35 shows, that the presence of a TP53 mutation increases the wGII whereas in cases with a PIK3CA mutation it is reduced. Similarly, the association between PIK3CA mutations and reduced wGII is sustained in cases with mutations in both TP53 and PIK3CA. Furthermore, wGII being even lower than in cases without any detected mutations in TP53 or PIK3CA. Therefore, the mutational status of these genes influences the degree of wGII in tumours from the BIG 1-98 cohort, with TP53 mutations causing increased genomic instability while PIK3CA mutations appearing to be linked to fewer CNAs.

The total number of alterations, both SNVs and CNAs, reported Foundation Medicine appears to also be related to distant recurrence, as shown in Figure 4.35. It can be seen that cases with presence of a TP53 mutation have elevated genomic instability, with increased fraction of genome altered (wGII) and, in distant recurrences, increased numbers of total mutations. A similar effect occurs in the presence of dual TP53 and PIK3CA mutations, but is wGII is lower than when TP53 alone is mutated. In contrast, the median number of alterations in groups with PIK3CA mutations or the absence of either TP53 or PIK3CA mutations was not affected by distant recurrence status.

The genomic data suggest the following dynamics,

1. In patients without mutations in either TP53 or PIK3CA, distant recurrence may be a consequence of accumulated wGII, SNVs, CNAs or some combination of these.
2. Cases with TP53 mutations show increased mutational burden with copy-number alterations being the more prevalent mutational process.

3. Cases with PIK3CA mutations tend have reduced wGII and fewer mutations overall, whether distant recurrence was observed or not.
4. In cases with mutations in both TP53 and PIK3CA genomic instability and mutational burden are both elevated for those with distant recurrence as compared to those without, but this effect is not as strong as it is in cases that only have TP53 mutations alone.

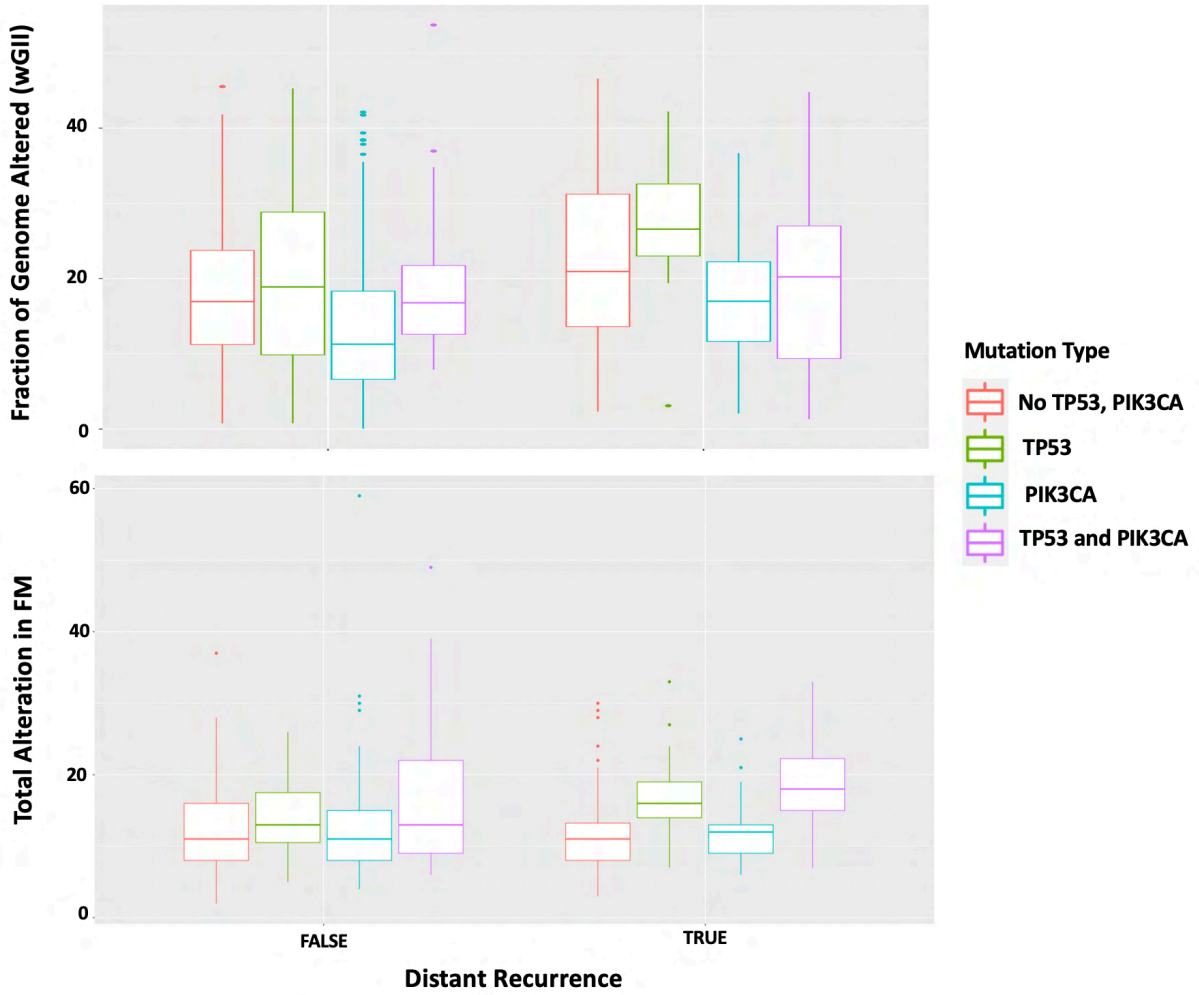


Figure 4.35 Mutational profile association in BIG 1-98.

To further explore the role of mutational burden on the predicted fits to the database of simulations, I applied Equation 4.1 to all simulated clones above 5% CCF for every patient to establish the fitness of each tumour. The increased mutation rate model was selected because it showed the best correlation with variant allele frequencies reported by Foundation Medicine, Table 4.4. As seen in Figure 4.36 for mutational burden, TP53 mutations can be linked to increased fitness in cases with distant recurrence. Furthermore, there was a significant association between increased fitness and distant metastasis in weighted Cox models.

The associations suggest the following dynamics:

1. In cases with no mutations in either TP53 or PICK3CA predicted fitness is not related to whether there was distant recurrence. This suggests distant recurrences in this group are subject to factors not directly related to positive selection, such as fraction of passenger mutations that can confer resistance.

2. Primary tumours from cases with TP53 mutations and distant recurrence show increased fitness relative to tumours cases that have not spread. This may be explained by:
    i. Mutations to TP53 increase driver mutation rates to $u = \{3.4\text{x}10^{-5}, 3.4\text{x}10^{-4}\}$ as happens in the increased mutation rate model.
    ii. Pronounced branched evolution where higher *k* subpopulations dominate.
    iii. A *Big-Bang* evolutionary mode where a boost in fitness is achieved within a small window of time.
3. Cases with PICK3CA mutations and distant recurrence show similar median fitness to those without but have reduced variance in fitness. This suggest sustained low net-growth resulting in replication stress over the long-term, increasing the odds of accumulating passenger mutations that can confer drug resistance. These tumours may be driven by accumulation of SNVs and focal CNAs rather than by large-scale genome-wide copy number alterations.
4. Cases with distant recurrence and mutations to both TP53 and PIK3CA show a minor median fitness increase relative to those with distant recurrence, but there is considerable variance. This suggests such pairs of mutations have a balanced effect on the fitness with alterations to TP53 increasing fitness but alterations to PIK3CA decreasing it.

As a result, the molecular profile influences the evolutionary mode by affecting the rates of expansion, increasing risk of passenger gain-of-function programs related to resistance or by exerting a combined effect on both. Tumour fitness is associated with clinical outcome for the cohort as a whole as shown in Figure 4.36.

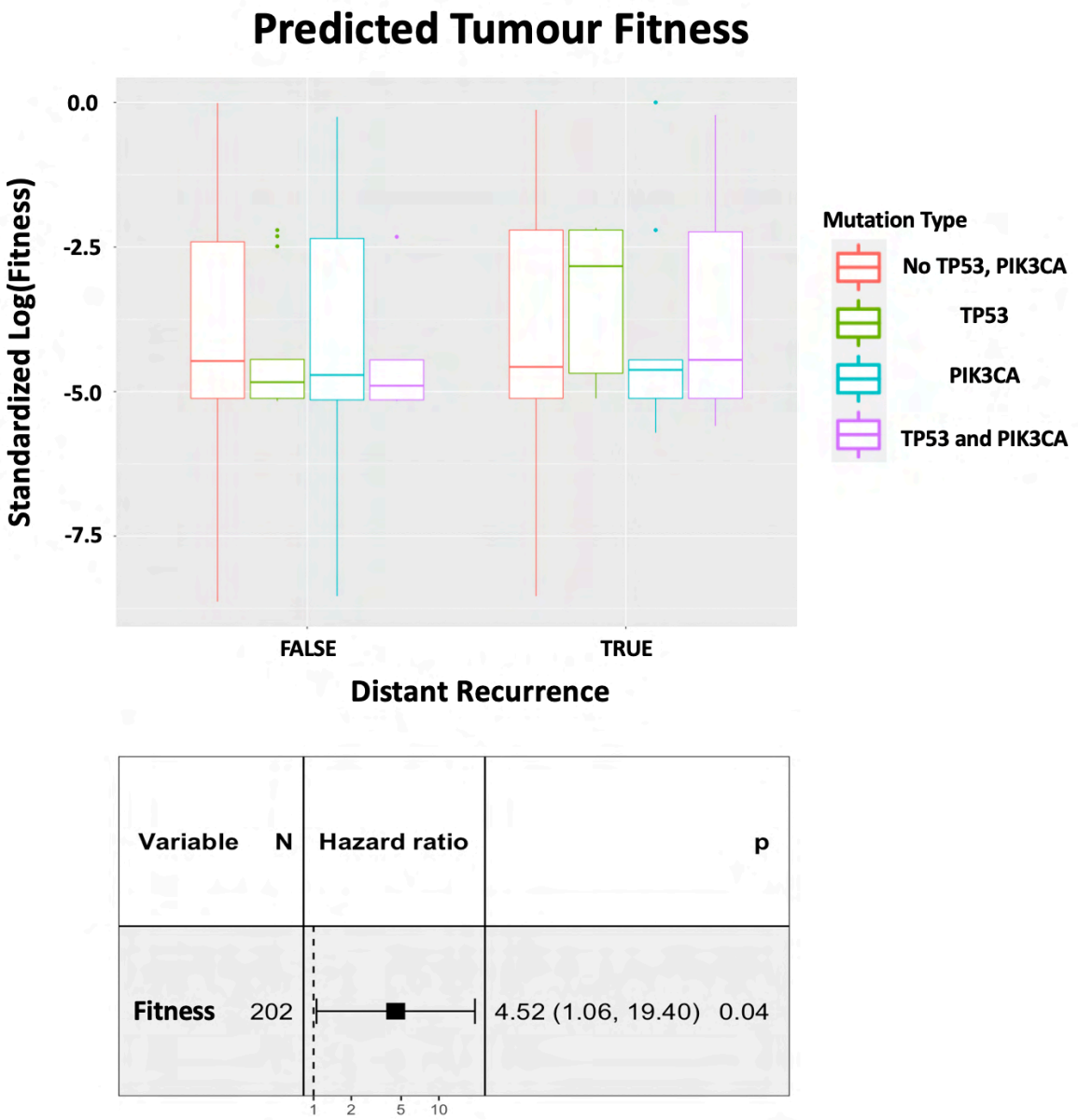


Figure 4.36 Association with clinical outcome with the predicted fitness in the increased mutation rate model. Predicted fitness of the best scoring simulations in the increased mutation rate model, fitness values were standard to make their effect more obvious. Cox models were adjusted by treatment arm and mutational profile.

The fitness increase in distant recurrences of TP53 mutant cases is consistent with the increase of Ki-67 reported for these tumours, as shown in Figure 4.37. Conversely, Ki-67 measurements in cases with PIK3CA mutations or the absence of TP53 and PIK3CA mutations appear unrelated to distant recurrence status. This provides an orthogonal validation that the fits of the increased mutation rate model are aligned with the genomic data, Figure 4.35.

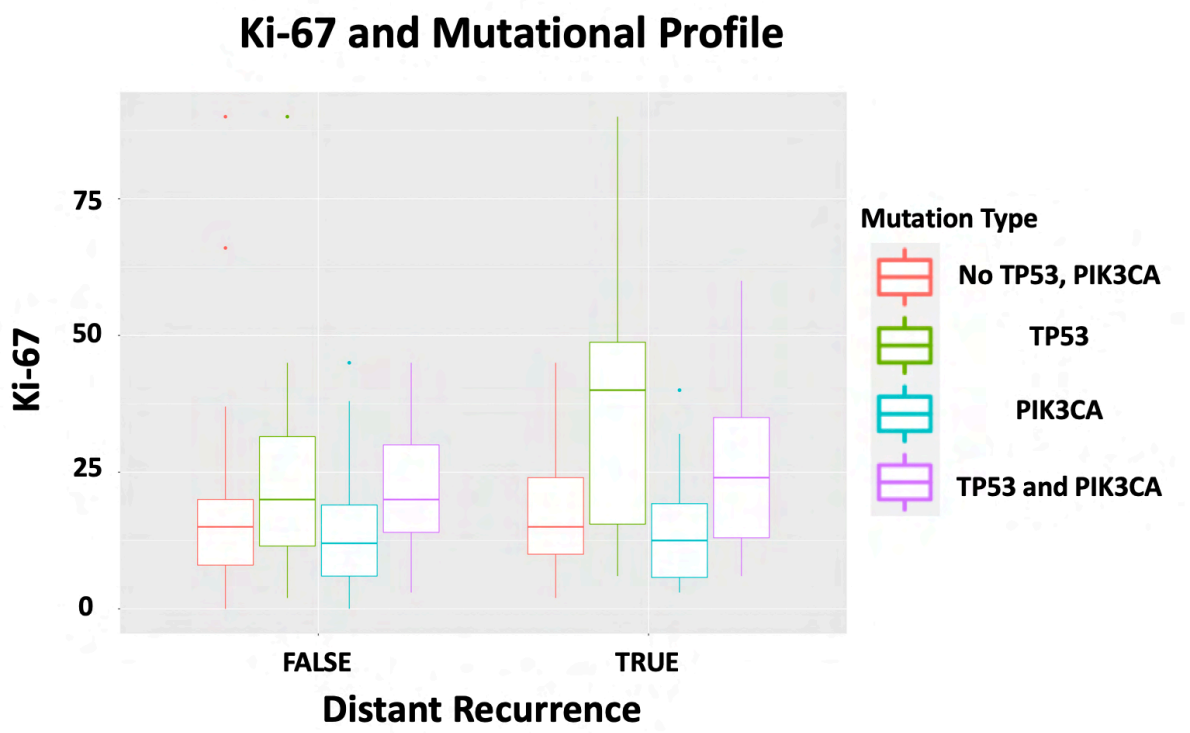


Figure 4.37 Ki-67 and mutational profile grouped by clinical outcome.

In summary, BIG 1-98 shows how the branching process with the MDA.N fitting technique is capable of reconstructing tumour evolution in studies with limited coverage and increased noise. It shows how model based reconstruction can improve phylogeny reconstruction compared to standard tools. Moreover, it shows how distant recurrences with TP53 mutations could be a consequence of fitness increases driven by such mutations, while recurrences with PIK3CA mutations and or recurrence in the absence of mutations to TP53 or PIK3CA are more likely due to natural disease progression by neutral passengers that confer functional advantages during treatment such as drug resistance mutations.

6.10 Estimating Average Selective Advantage *s* and Average Driver Mutation Rate *u* in CASCADE Melanoma

The main objective of this section is to evaluate the ability of the MDA.N method to identify similarity and divergence in patterns of dissemination and/or clonal makeup in matched samples of primary tumours and metastases, based on whole exome sequencing.

To this end, I evaluated data from a patient enrolled in the cancer collection after death program (CASCADE) who succumbed to metastatic melanoma. A total of 10 samples were collected from this patient: 4 from the primary tumour located on the skin of the right lower back, and 5 metastases, 3 from the liver and 2 from the brain. Two additional blood samples were taken after the initial stage of treatment to evaluate minimal residual disease.

One of the primary samples displayed evidence of neutrality and thus was removed from the analysis. The circulating tumour samples were also discarded from the analysis as the models do not account for that process. To match the whole exome sequencing assay, simulated clones with CCF below 10% were removed from the analysis and clonality assessment was done by ExPANdS.

As shown in Figure 4.38, the fits for the CASCADE melanoma samples are consistent with the values identified for the TCGA SKCM subtype. Excluding the Brain Left sample, which showed a moderate average selective fitness of $s = 0.01$, the average selective advantage observed was weak, $s = 0.001$. Similarly, excluding the liver right sample with a moderate average driver mutation rate $u = 3.14\text{x}10^{-6}$ the remaining samples have a high average driver mutation, $u = \{3.14\text{x}10^{-5}, 3.14\text{x}10^{-6}\}$. These values are within the range of the likelihoods reported in Chapter III, Figure 3.15.

Heatmaps in Figure 4.38 are the consensus fits for all 3 positive selection models combined using the best scoring simulations that minimised the MDA.N statistic. It can be seen there is generally a high level of concordance between the models.

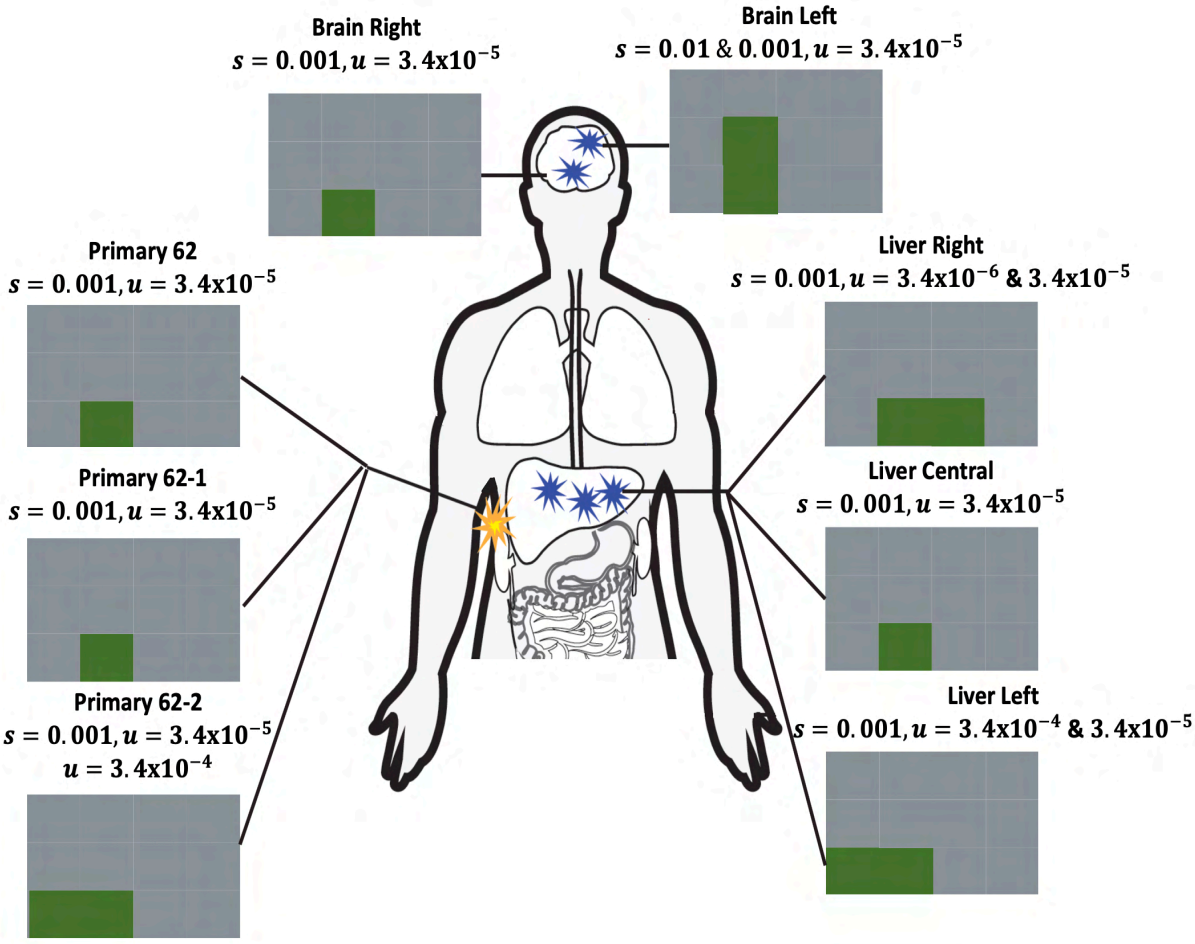


Figure 4.38 Fits on the CASCADE project. Heatmaps describing the consensus of best fits of the 3 positive selection models, with green the most likely area of the fits. Predicted parameters are given at the top of the heatmap for each tumour site.

Figure 4.39 shows the distribution of the best scoring MDA.N fits in the primary and metastases samples for the 3 positive selection models. The fits had similar distributions of predicted number of clones and RGS diversity. However, the predicted median number of clones is slightly lower than for the SKCM samples from TCGA (3 vs 4, 5) and the RGS diversity differs slightly as well (below 2 in CASCADE and greater than 2 in TCGA).

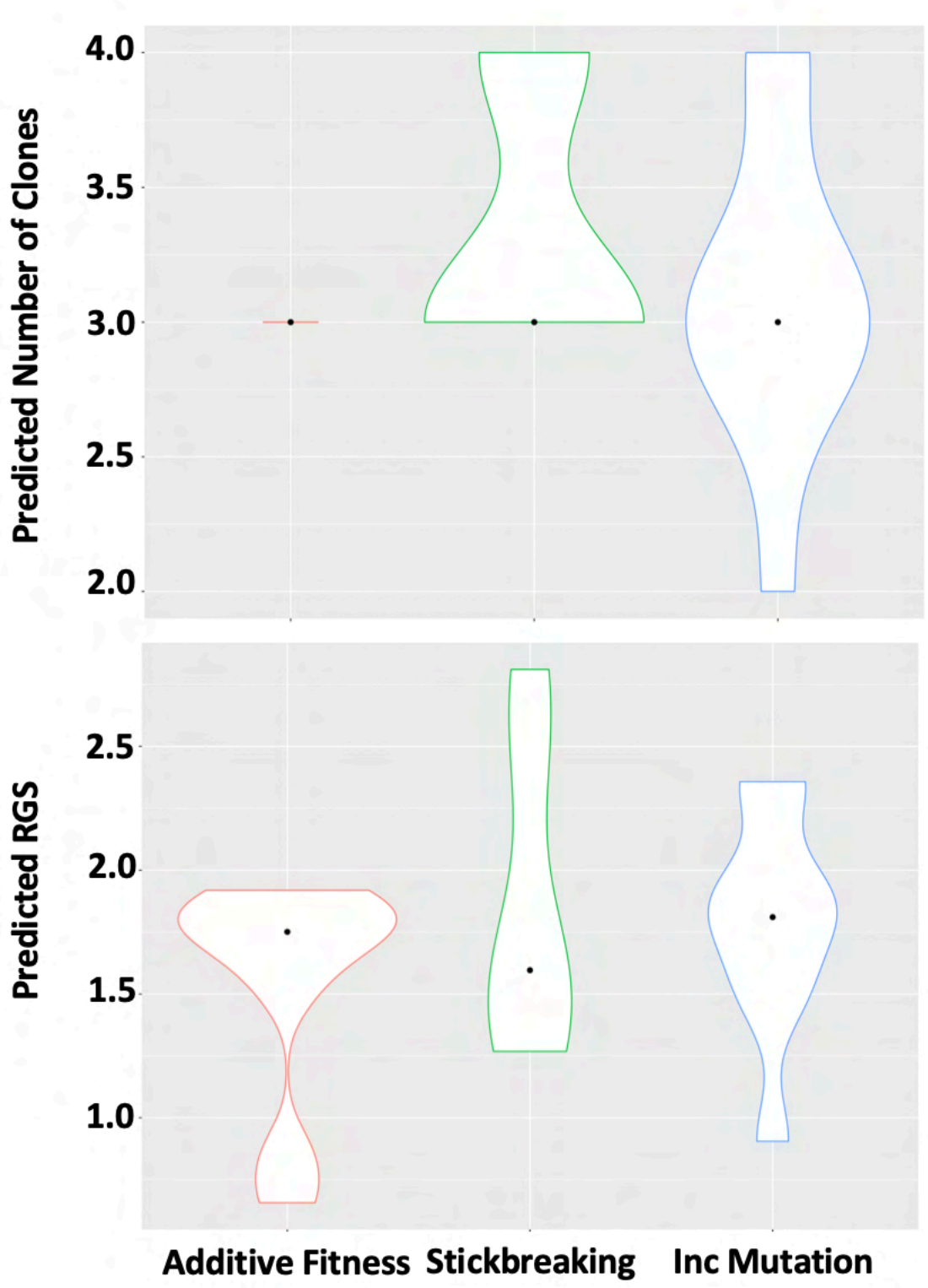


Figure 4.39 Distribution of detectable clones and RGS diversity. Distribution of number of detected clones and RGS diversity for all the positive selection models based on the top scoring simulations.

Melanoma is known to have a strong mutational burden, harbouring a considerable mutational load that is believed to build up before carcinogenesis creating a long trunk in the phylogenetic tree that eventually diversifies due to new driver clones [52, 208, 209].

Therefore, the increased RGS diversity shown as increased variance in the stickbreaking model–which is higher than the additive fitness and increased mutation rate models–leads fewer measurable subpopulations with the applied 10% CCF cut-off. This seems to suggest two possible explanations for the discordance in the fits to the stickbreaking and the other two models:

1. The stickbreaking model is representing the long trunk as a single branch explaining the reduced number of detectable clones.
2. The additive and increased mutation models are overfitting the data, leading to higher apparent clonal diversity.

The predictions of the likely values of *s* and *u* for primary and metastatic samples in CASCADE melanoma, are the same to those observed in the other studies in the previous sections.

6.11 Recurrent Phylogenies and Clonal Evolution Reconstruction in CASCADE Melanoma

Next I evaluated the topologies of every model independently to study the evolutionary relationship between primary and metastatic samples. There is a recurrent match between one specific stickbreaking model simulation and the Primary 61-1, Liver Central and Brain Right samples. However, each sample fit best to a different snapshot in time from this simulation, Figure 4.35. The Primary 61-1 sample fit to an advanced snapshot (3 cm^3) and metastatic Liver Central and Brain Right matched an earlier snapshot (2.5 cm^3). Additionally, this same simulation was one of the top 3 most recurrent in the SKCM subtype in TCGA in Figure 4.13

Primary 61-1, Liver Central and Brain Right fit to the same snapshot (3.5 cm^3) in the increased mutation rate model. This simulation turns out to most recurrent simulation in BIG 1-98 as well with the properties already mentioned, Figure 4.34. There were no recurrent simulations from the additive fitness model.

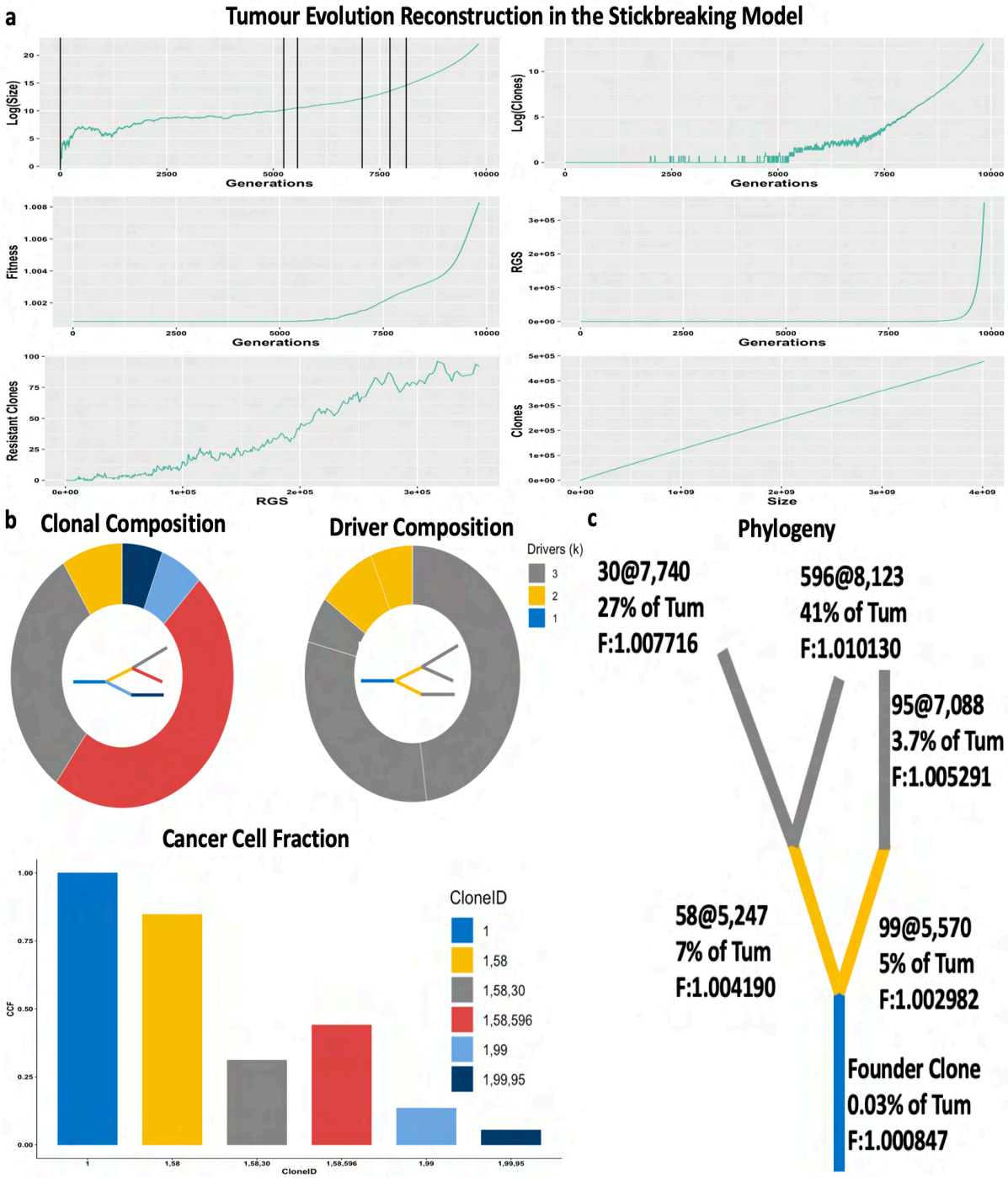


Figure 4.40 Top recurrent simulation in the stickbreaking model. a, describes the tumour growth dynamics using the variables tumour size, generations, number of clones, fitness, RGS and resistant clones. Black vertical lines in the Log(Size) vs Generations indicate τ, the time when the detectable successful clones emerged and expanded. **b**, describes the clonal and driver composition as a pie chart. Colours in the clonal composition indicate individual clones and in the driver composition the driver abundance. The bar plot shows cancer cell fractions of all detectable clones colour coded as in the pie chart above. **c**, the phylogeny of driver clones for this simulation, colour coded by number of accumulated drivers k. Each clone is labelled with an ID number followed by @ with τ, the time of emergence of that clone. The proportion of the final tumour size represented by that clone is indicated underneath with their fitness (F).

The recurrent simulation in the stickbreaking model shows one main lineage comprising 75% of the tumour, driving its expansion. This is a highly dominant lineage, as jointly all detectable clones together explain 83% of the composition of the whole tumour. This simulation has a timeframe that suggests division rates of 2-3 days. The degree of diversity in this simulation is lower than seen in other recurrent simulations from the stickbreaking model. This is likely because fitness increased 8 times more from the initial value at the final size 4 cm^3 meaning this tumour was proliferating faster and thus accumulating less diversity.

Figure 4.41 shows the inferred topologies and evolutionary parameters for each tumour from this patient. It suggests how patterns of dissemination to distant metastatic sites could potentially have occurred.

1. The Brain Right and Liver Central metastases disseminated from Primary 61-1, The similarity in fitness between these sites suggests dissemination occurred in parallel, indicated with a double asterisk (**). However, it is possible the tumour metastasized to one site and then the other, e.g., migration from primary to Liver Central and then to the Brain Right or vice versa.
2. The Liver Right and Brain Left metastases share the same topology, suggesting they share the same ancestry, shown with one asterisk (*). The Brain Left has an increased fitness and the Liver Right has a lower diver mutation rate which suggest the liver metastasis migrated to the brain. The topologies of the metastases suggest that they migrated from the primary when it was at an earlier stage of its development dominated by 2-driver clone.
3. The topology of the Liver Left is concordant with all topologies of the different regions within the primary, only possibly at an earlier evolutionary stage when the there was only a single detectable 3-driver population within the primary.

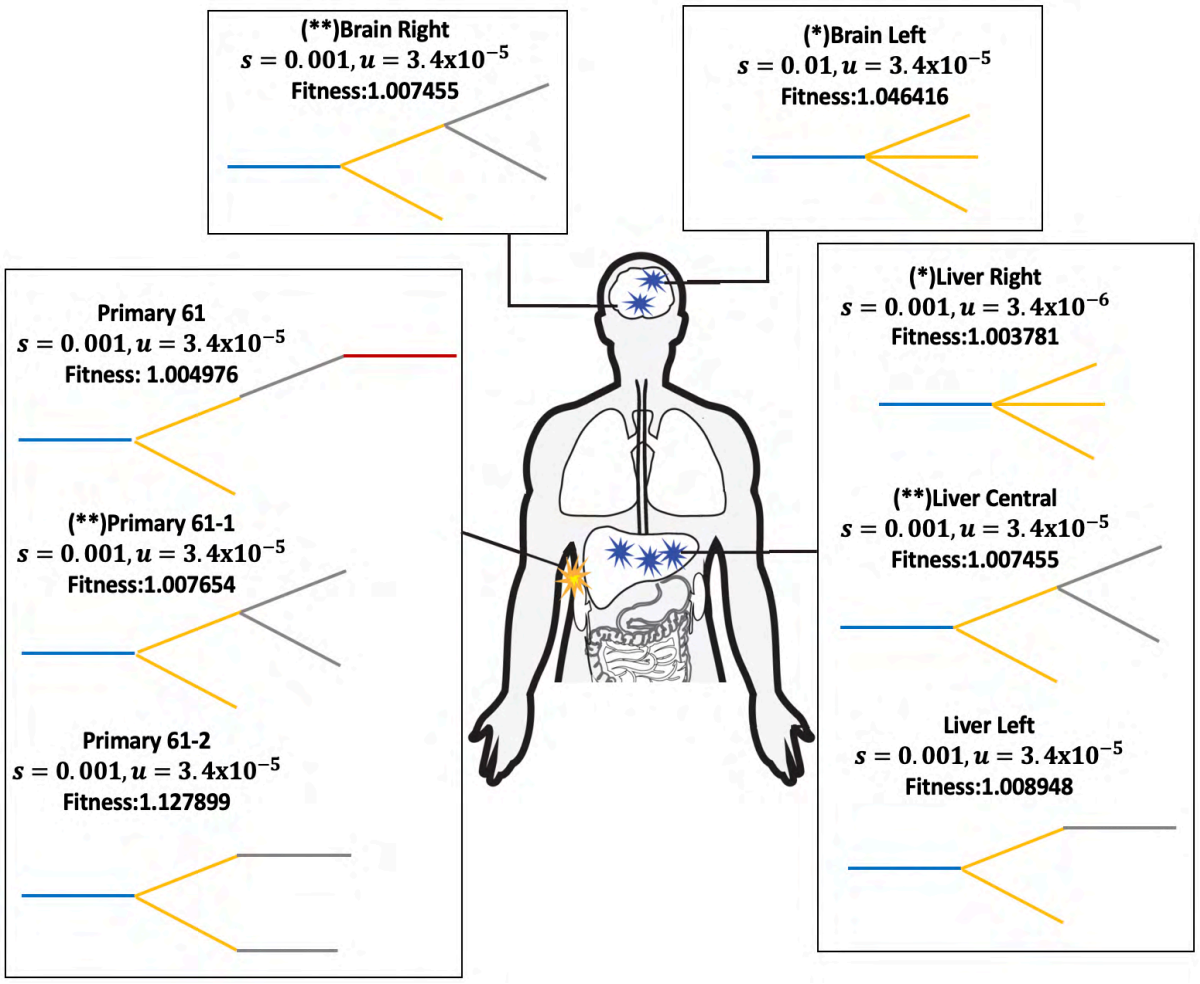


Figure 4.41 Recurrent phylogenies predicted by the stickbreaking model. Phylogenies are labelled with the predicted parameters and their corresponding fitness. Phylogenies are colour coded by number of accumulated drivers: blue, yellow, grey and red correspond to 1, 2, 3 and 4 drivers, respectively. With double asterisk (**) are the phylogenies that have the same initial parameters of s and u and replicate. With one asterisk (*), samples that share the same topology but come from different simulations.

When primary and metastatic samples are available, ExPANdS provides a function that while establishing the clonality, also clusters samples according to their similarity. Fig 4.42 shows the clustering done by ExPANdS using primary and metastatic samples. In support of the observations shown in Figure 4.41, the Brain Right, Liver Central and Primary 61-1 samples cluster in one group and the Liver Right with the Brain Left cluster in a different one.

Figure 4.42 Cluster of samples done by ExPANDdS. Dendrogram is based on cancer cell fraction similarity and copy number change. The dendrogram is sectioned by clusters of similarity with the phylogenetic topology. Phylogenies are colour coded by number of accumulated drivers: blue, yellow, grey and red correspond to 1, 2, 3 and 4 drivers, respectively. Dashed lines in the phylogeny indicate the consensus topology of the primary sample. Primary 61-3 contains only one clone and was discarded form the analysis. Labels on the tips of the phylogeny are the sample names followed by the dominant CCF named (SP), the CCF is displayed without purity correction.

The cluster outlined in black contains samples from the Primary 61, Primary 61-3 and Liver Left, indicating Liver Left disseminated from one of those primary sites. The cluster outlined in red contains samples that were fit to the same simulation, flagged with the ** in Figure 4.41. Samples collected from the patient's blood to evaluate circulating tumour DNA are also in that cluster, which reinforces the idea dissemination occurred from the Primary 61-1. The cluster outlined in purple contains the samples from the Brain Left and Liver Right that share similar topology, flagged with one asterisk (*) in Figure 4.41.

Although the topologies inferred for each site are different, the evolutionary relationships between sites predicted by results from fitting to the increased mutation model match those based on the stickbreaking model.

Figure 4.43 shows recurrent topologies from the increased mutation rate model. The concordance between samples Primary 61-1, Liver Central and Brain Right remains. The Liver Right and Brain Left samples also show similarity, as they did in the stickbreaking model and ExPANdS clustering. The topology for Liver Left sample remains the same as it was for the stickbreaking model. This provides more support for the correctness of the fits to the simulations from the stickbreaking model.

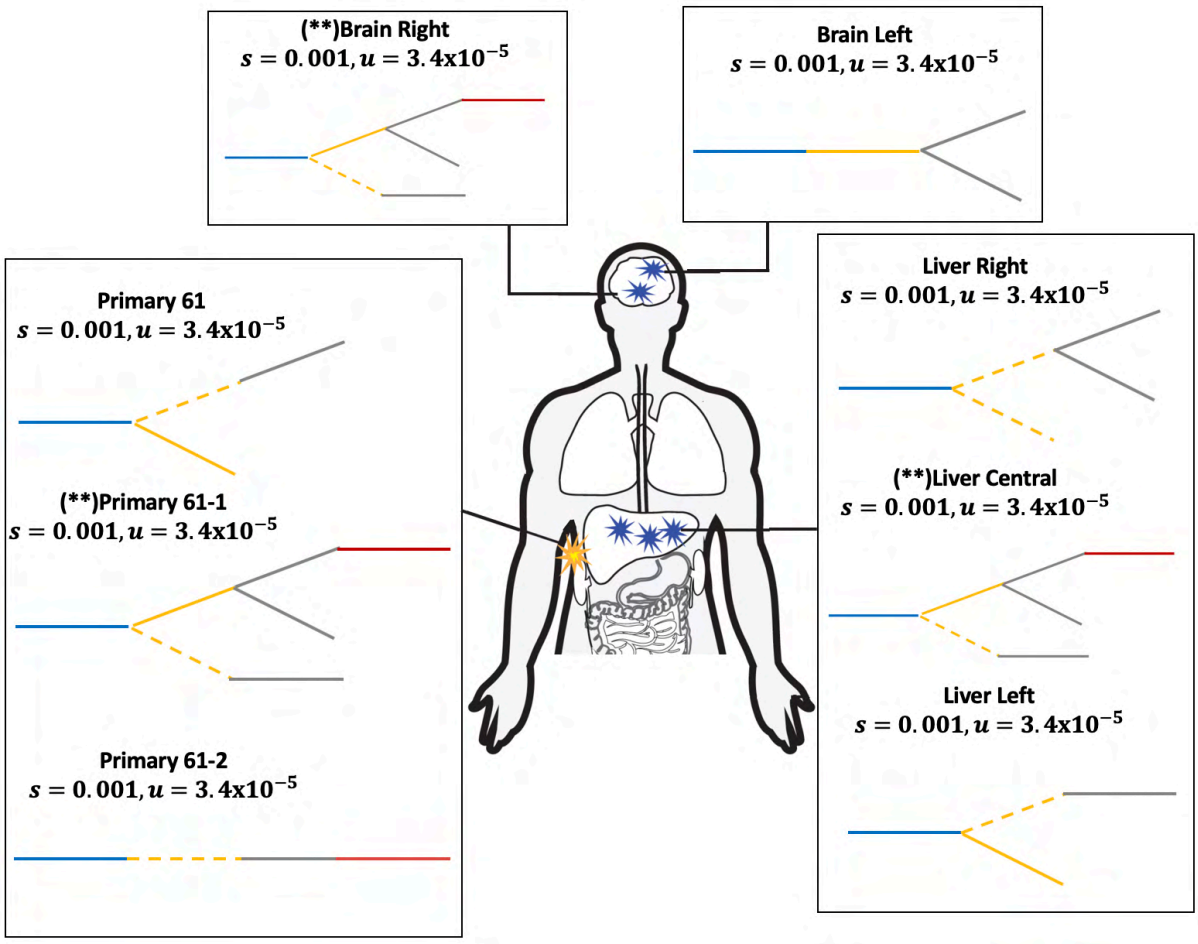


Figure 4.43 Recurrent phylogenies predicted by the increased mutation rate model. Phylogenies are labelled with the predicted parameters and their corresponding fitness. Phylogenies are colour coded by number of accumulated drivers: blue, yellow, grey and red correspond to 1, 2, 3 and 4 drivers, respectively. Dashed lines represent hyper mutant lineages. With double asterisk (**) are the phylogenies that have the same initial parameters of s and u and replicate.

Fitting to the branching process with the MDA.N method showed consistency in its fits between the TCGA SKCM subtype and the CASCADE melanoma study. It was able to identify similarities in the topologies that indicate plausible patterns of dissemination of metastatic disease. The consistent patterns observed between the stickbreaking and increased mutation

rate models indicate that the MDA.N method produces fits that align well to the data. This was confirmed with ExPANdS, which clustered samples according to their predicted topologies.

7 Discussion

In this chapter I developed a procedure to reconstruct tumour evolution in patients and use it to gain biological insights from patient tumour samples. I identified a way to compare simulation results from the branching process models developed in Chapter III with data from sequenced tumours based on ploidy corrected cancer cell fractions (CCFs). This was achieved by implementing a statistical fitting approach based on the minimum Euclidean distance that was able to handle noise and missing information. I applied this fitting strategy to 1,800 tumours from numerous cancer subtypes sequenced with varying genome wide coverages and depths. The predictions done in the 1,800 patients showed the power of the branching process to provide the evolutionary trajectories and growth dynamics in tumours with clinical value.

Tumour proregression starts with *weak* fitness. The fits in 1,800 tumours corroborated that the parameters of weak average selective advantage are the most common amongst solid tumours ($\sim s = \{0.001, 0.01\}$. I also found evidence that the higher driver rates tested here $u = \{3.14\text{x}10^{-5}, 3.14\text{x}10^{-4}\}$) are most biologically plausible. These findings align with those reported by Bozic et al. [43]. Interestingly, these parameters can provide a rich variety of evolutionary trajectories that can describe the extensive heterogeneity observed in human malignancies. Furthermore, by the time tumours achieve diagnosable sizes (1 cm^3 and 4 cm^3), they have accumulated a plethora of genetic alterations that can activate biological programs of cancer-specific mortality–such as pre-existing drug resistance.

Connecting tumour sizes with fitness dynamics is crucial to understanding how low-frequency passenger alterations influence dissemination and drug resistance. Tumours that acquired increased fitness during their evolution required fewer cell divisions, reducing the odds of low-frequent passenger alterations such as drug resistance to emerge. Therefore, tumours that do not increase in fitness may represent greater risk to patients as they can accumulate more drug resistant cells.

Evidence of high clonal diversity is associated with bad prognosis was seen in TCGA. The main reason for this is the lack of a single strong driver leading to the appearance of multiple competing clones. The distributions of the number of detectable clones, fitness and RGS diversity clustered with survival. Recurrent topologies across subtypes were identified in which branched phylogenies were more frequent in subtypes with poor prognosis, whereas *simple* topologies with less clonal subpopulations were observed in good prognosis subtypes. This is in line with other analyses of the relationship between inferred number of clones and heterogeneity [29, 30]

In TRACERx NSCLC, excluding stage and tumour size, clinical and genomic variables did not show significant association with clinical outcome, nor with predicted fitness or RGS diversity. Although the fits approximate the phylogenies reported by the study, the observed trend suggests that increases in fitness and RGS diversity can be associated with recurrence or death. This supports the role of tumour heterogeneity as a proxy for the prevalence of low-frequent passenger alterations such as drug resistance. The role of genome doubling and chromosomal instability are key processes that fuel tumour growth and affect clinical outcome [96]. Even though these processes were not explicitly modelled, the fits approximated these effects.

In BIG 1-98, the increased mutation rate model showed the best correlation with the patterns observed in the mutational profiles of the genomic data. The model indicated that distant recurrences with TP53 are the result of fitness increases by genome-wide copy number alterations or accumulation of single nucleotide variations. In contrast, distant recurrences with PIK3CA did not show fitness increases, indicating that in these cases recurrence may occur through natural disease progression driven by the stochastic emergence of biological programs such as drug resistance. Although the more positive prognosis indicated by PIK3CA mutations is known [189] I showed in BIG 1-98 that PIK3CA can represent a risk if the tumour stays in a weak fitness mode, as this will accumulate diversity because of the extra number of cell divisions required for the tumour to grow to a detectable size.

In the CASCADE melanoma patient, the stickbreaking and increased mutation rate models identified a possible pattern of dissemination from the primary to the liver and brain metastases, with fits converging to the same simulation. The prediction was validated with clustering performed using ExPANdS. The recurrent phylogeny in primary and metastatic sites suggests that treatment did not cause additional driver heterogeneity. Although this is just one patient, it illustrates the power of using the branching process in tumour evolution reconstruction.

Using the different branching process models provided valuable insights into many aspects of tumour evolution In BIG 1-98, it showed that an increased mutation rate model best approximated evolutionary trajectories with correlated with clinical outcome in this cohort. In TRACERx NSCLC and TCGA use of the stickbreaking model showed a *Big-Bang* evolutionary mode could explain a substantial portion of tumours. Such a finding is more likely to be achieved with the stickbreaking model, through its random sampling of s, which allows a large fitness leaps to occur stochastically.

Tumours can switch from a weak expansion to an aggressive expansion that can represent a higher risk course of disease. The stickbreaking and increased mutation rate models identified cases that start off with a low fitness founder clone dominated by the effects of drift that grow slowly. Increases in fitness can happen through low probability events (stickbreaking model) or hyper mutants (increased mutation rate model) switching the trajectory into an aggressive tumour expansion.

A good example was shown in the analysis of the BIG 1-98 cohort, with recurrent fits to a simulation in the stickbreaking model where the tumour stays in a weak phase, increasing the load of drug resistance cells, before expanding aggressively to increase its fitness by 10 times above its initial value. This illustrates the valuable information that can be gained by connecting mutational processes with growth development dynamics to provide a better understanding of disease progression.

Approximating of tumour evolution using the discrete time branching process is possible. It provides an alternative framework to connect mutational process with clonal growth dynamics. It expands on the range of evolutionary trajectories that may explain an observed mutational profile. It provides great flexibility and robustness by handling noise and missing information to estimate the key initial evolutionary parameters average selective advantage s and average driver mutation rate u. In addition, it can provide biological insights into key processes that drive clonal evolution.

The reconstruction of 1,800 patients with multiple cancer subtypes and sequencing assays shows that jointly, the use of the positive selection models can inform the process of tumour development. Future work involving the branching process should consider modifications to the average selective advantage s and average driver mutation u that were not explored in the models, to sample the full landscape of clonal configurations possible during tumour evolution.

8 Appendix and Supplementary Figures and Tables

A.4.1 Goodness-of-Fit-Statistics

Let $X_1, \ldots, X_n$ be i.i.d. observations on a d-dimensional random vector X (e.g. patient CCFs) with distribution function (cdf) $F(x)$ whilst $Y\ (\theta)$ denotes a d-dimensional random vector (e.g. simulated CCFs) with cdf $F_\theta(x)$ indexed by the parameter vector $\theta \in \Theta \subseteq R^p$. The former random sample represents experimental observation(s) derived from sequencing a given tumour while the latter sample represents simulated data generated from the evolutionary models at multiple points with input parameters θ.

Therefore, the goal of the goodness-of-fit statistics is to evaluate if two samples simulated and observed, $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_M$, come from the same distribution and to find the minimizer of the goodness-of-fit divergence, θ. As a result, every simulated instance of $F(\cdot;\theta)$ is compared to observation $F_\theta(x)$ to approximate θ.

The first statistic evaluated is the Cramér-von Mises criterion of two samples defined as

$$T_{N+M}(\theta) = \left[\frac{NM}{N+M}\right] \int \left\{\hat{F}_N(x) - F_M(x;\theta)\right\}^2 dH_{N+M}(x)$$

As shown by Anderson T W. [204] the integral from of the statistic can be evaluated by a comparing the ranks as following,

$$T_{N+M}(\theta) = \frac{U}{NM(N+M)} - \frac{4MN-1}{6(M+N)}$$

Where

$$U = N\sum_{i=1}^{N}(r_i - i)^2 + M\sum_{j=1}^{M}(s_j - j)^2$$

Where $r1 <\cdots< rN$ are the ranks of x in the combined sample and $s1 <\cdots< sM$ the ranks of y in the combined sample.

The second statistic used is the Kullback-Leibler divergence which measures the difference between to density functions. Because the divergence of $D_{KL}\left(\hat{F}_N(x) \parallel F_M(x;\theta)\right)$ is different from the divergence of $D_{KL}\left(F_M(x;\theta) \parallel \hat{F}_N(x)\right)$ I took the average of them as following,

$$D_{KL} = \frac{1}{2}\left(\int_{-\infty}^{\infty} \hat{F}_N(x) ln \frac{\hat{F}_N(x)}{F_M(x;\theta)} dx + \int_{-\infty}^{\infty} F_M(x;\theta) ln \frac{F_M(x;\theta)}{\hat{F}_N(x)} dx\right)$$

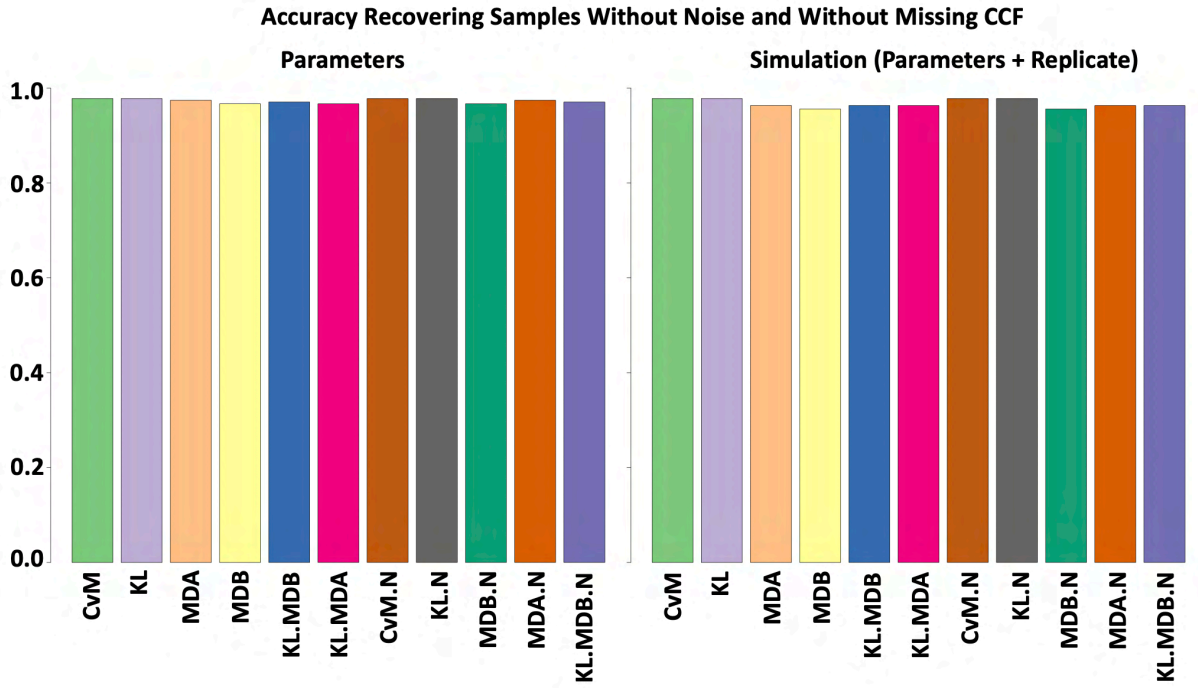


Figure S4.1. Benchmarking of original parameters. Left bar plots are the accuracy results of recovering the simulation in the top-10 fits, and the right bar plots are the equivalent for recovering the correct simulation iteration.

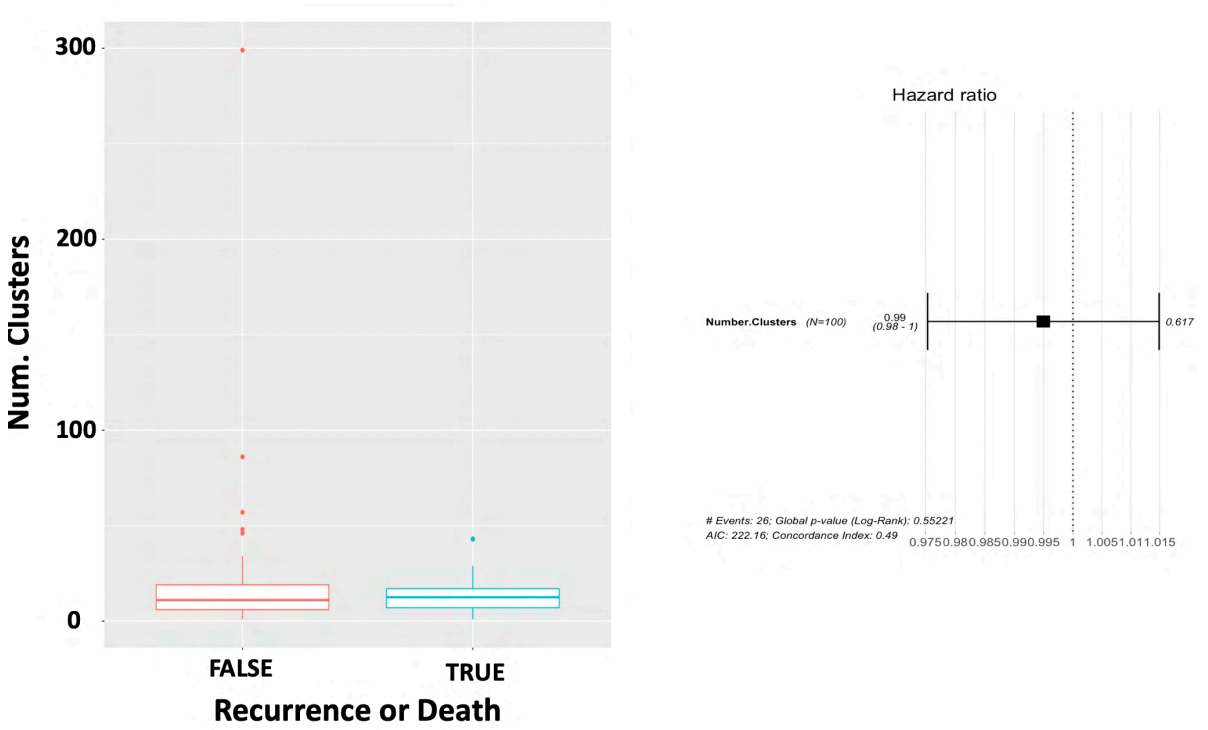


Figure S4.2 Association of the number of PyClone clusters with recurrence or death.

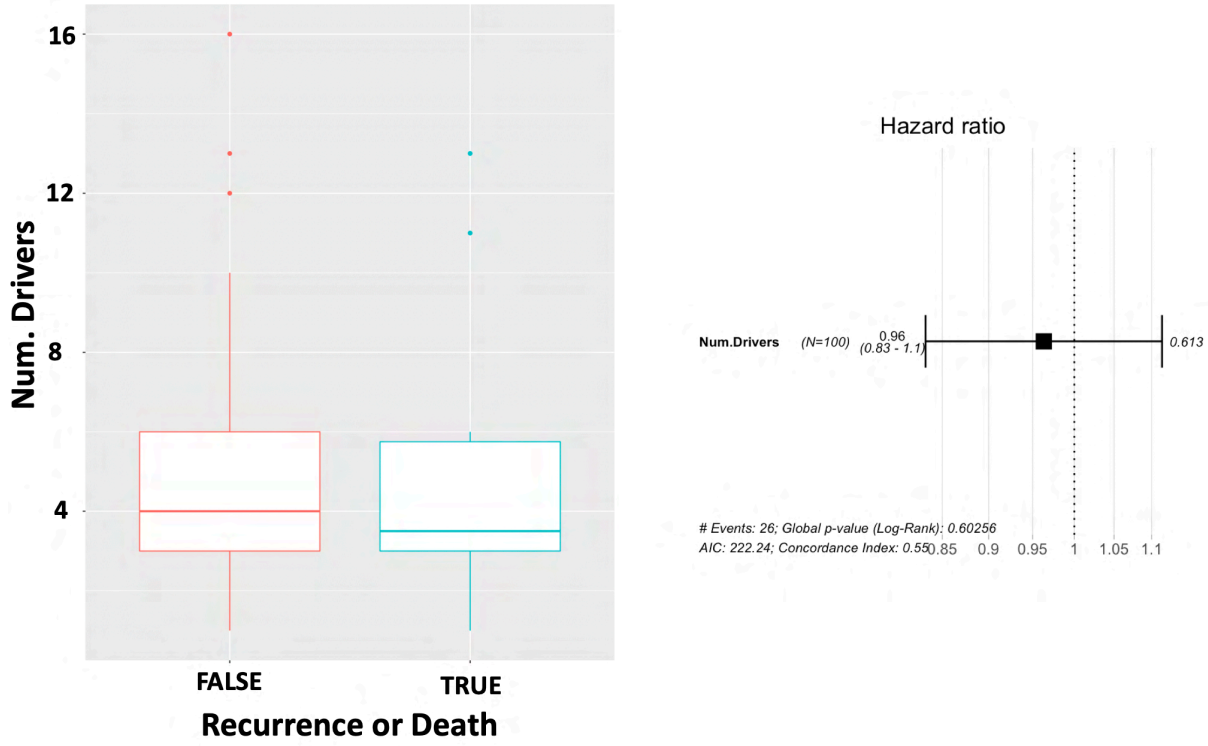


Figure S4.3 Association of the number of drivers and PyClone clusters with recurrence or death.

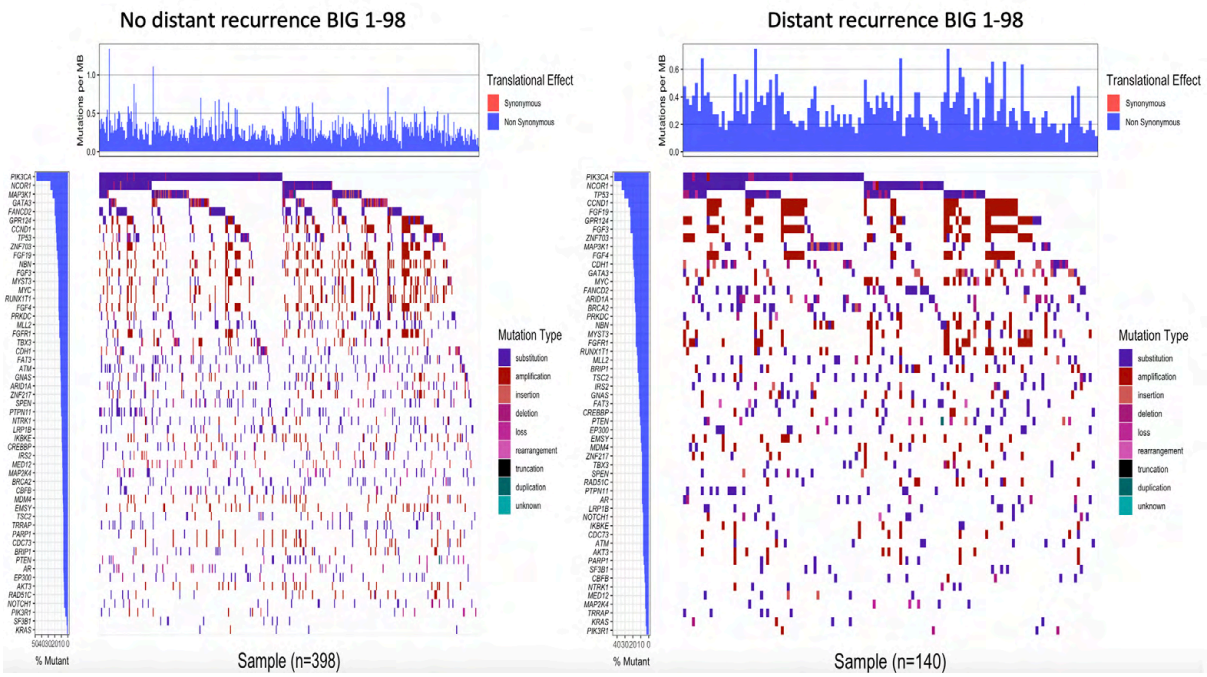


Figure S4.4 Mutational frequencies in BIG 1-98 grouped by clinical outcome.

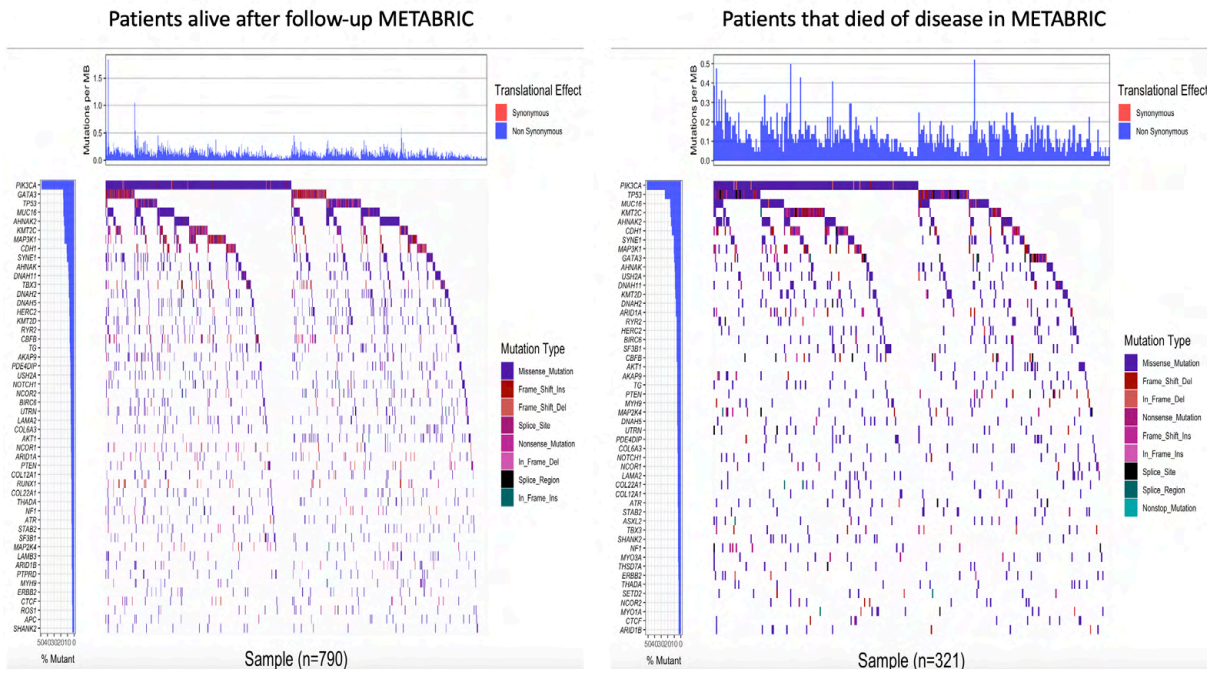


Figure S4.5 Mutational Frequencies in METABRIC grouped by clinical outcome.

Supplementary Table 4.1 TCGA Subtype Description and Overall Survival

Name	*Cancer Subtype*	*Overall Survival (%)*
BLCA	Bladder urothelial adenocarcinoma	54
CESC	Cervical squamous cell carcinoma	77
GBM	Glioblastoma Multiforme	17
HNSC	Head and neck squamous cell carcinoma	58
KIRC	Kidney renal clear cell carcinoma	66.8
LGG	Brain low grade glioma	75.5
LUAD	Lung adenocarcinoma	58
LUSC	Lung squamous cell carcinoma	57
PRAD	Prostate adenocarcinoma	98
SKCM	Skin cutaneous melanoma	50
STAD	Stomach adenocarcinoma	61
THCA	Thyroid carcinoma	96

Supplementary Table 4.2 Univariate Analysis BIG 1-98 wGII by Cytoband, N=538

Variable	**HR (95%CI)**	**p.val**	**zph**
wGII	1.06 (1.02 – 1.09)	<0.001	0.245
1p	1.01 (0.98 - 1.03)	0.6	0.03
1q	1.01 (1.0 - 1.02)	0.2	0.52
2p	1.02 (0.99-1.05)	0.2	<0.001
2q	1.03 (1.01-1.05)	0.005	0.94
3p	1.02 (1.0 - 1.04)	0.03	0.039
3q	1.01 (1.0 - 1.02)	0.2	0.05
4p	1.02 (0.99 - 1.05)	0.2	0.155
4q	1.01 (1.00 - 1.03)	0.2	0.39
5p	1.01 (1.00 - 1.02)	0.1	0.804
5q	1.02 (1.00 - 1.05)	0.04	0.76
6p	1.00 (0.99 - 1.01)	0.6	0.84
6q	1.01 (1.00 - 1.03)	0.04	0.85
7p	1.01 (0.99 - 1.03)	0.6	0.002
7q	0.99 (0.98 - 1.01)	0.6	0.02
8p	1.00 (1.00 - 1.01)	0.04	0.46
8q	1.00(1.00-1.01)	0.3	0.69
9p	1.00 (0.99 - 1.02)	0.6	0.142
9q	1.00 (0.98 - 1.03)	0.8	0.48
10p	1.02 (1.01 - 1.03)	<0.001	0.0508
10q	1.04 (1.01 - 1.06)	0.002	0.0733
11p	1.01 (1.00 - 1.02)	0.2	0.52
11q	1.01 (1.00 - 1.01)	0.1	0.97
12p	1.00 (0.99 - 1.01)	0.7	0.504
12q	1.01 (0.99 - 1.02)	0.7	0.3
13q	1.01 (1.00 - 1.02)	0.005	0.223
14q	1.01 (1.00 - 1.02)	0.05	0.684
15q	1.01(1.00 - 1.01)	0.05	0.457

16p	1.00 (0.99 - 1.01)	0.9	0.56
16 q	1.00 (0.98 - 1.02)	0.8	0.58
17p	1.00(0.98 - 1.01)	0.08	0.72
17q	1.00 (0.99 - 0.01)	0.5	0.581
18p	1.00 (1.00 - 1.01)	0.3	0.945
18q	1.01 (1.00 - 1.02)	0.06	0.502
19p	0.99 (0.98 - 1.01)	0.4	0.81
19q	1.00 (0.98 - 1.03)	0.7	0.432
20p	1.00 (0.99 - 1.01)	0.9	0.376
20q	1.00 (0.99 - 1.02)	0.4	0.51
21p	1.00 (1.00 - 1.00)	0.2	0.4
21q	1.01(0.99 - 1.02)	0.5	0.0125
22q	1.03 (1.01 - 1.05)	0.002	0.043

Supplementary Table 4.3 Univariate Analysis in METABRIC wGII by Cytoband N=1174

Variable	**HR (95%CI)**	**p.val**	**zph**
wGII	1.06 (1.04 – 1.09)	<0.0001	0.461
1p	1.03 (1.01 - 1.04)	<0.001	0.03
1q	1.01 (1.00 - 1.02)	0.13	0.04
2p	1.02 (1.01 - 1.03)	0.002	0.0554
2q	1.00 (0.99 - 1.02)	0.7	0.922
3p	1.01 (1.00 - 1.03)	0.02	0.558
3q	1.02 (1.01 - 1.03)	0.002	0.34
4p	1.01 (1.00 - 1.03)	0.05	0.53
4q	1.01 (1.00 - 1.03)	0.05	0.377
5p	1.01 (0.99 - 1.02)	0.3	0.91
5q	1.01 (0.99 - 1.02)	0.2	0.725
6p	1.00 (0.99 - 1.02)	0.5	0.5
6q	1.01 (1.00 - 1.02)	0.03	0.96
7p	1.01 (1.00 - 1.02)	0.09	0.789
7q	1.00 (0.99 - 1.01)	0.7	0.535
8p	1.02 (1.01 - 1.02)	<0.001	0.82
8q	1.01 (1.00 - 1.01)	0.002	0.706
9p	1.02 (1.01 - 1.03)	0.002	0.351
9q	1.00 (0.99 - 1.01)	0.9	0.89
10p	1.02 (1.00 - 1.03)	0.004	0.943
10q	1.02 (1.00 - 1.03)	0.02	0.88
11p	1.01 (1.00 - 1.02)	0.02	0.043
11q	1.01 (1.00 - 1.02)	<0.001	0.0501
12p	1.03 (1.01 - 1.04)	<0.001	0.22
12q	1.02 (1.01 - 1.03)	<0.001	0.726
13q	1.01 (1.00 - 1.02)	0.1	0.823
14q	1.01 (1.00 - 1.03)	0.03	0.16
15q	1.01 (1.00 - 1.01)	0.2	0.44
16p	1.0 (0.99 - 1.01)	1	0.89
16 q	1.01 (1.00 - 1.02)	0.1	0.66
17p	1.02 (1.01 - 1.03)	<0.001	0.66
17q	1.01 (1.00 - 1.02)	0.008	0.2
18p	1.02 (1.00 - 1.03)	0.02	0.32

18q	1.01 (1.00 - 1.02)	0.2	0.713
19p	1.02 (1.00 - 1.03)	0.006	0.968
19q	1.01 (1.00 - 1.02)	0.09	0.77
20p	1.01 (1.00 - 1.02)	0.03	0.95
20q	1.01 (1.00 - 1.01)	0.005	0.78
21p	1.00 (0.99 - 1.02)	0.6	0.87
21q	1.01 (0.99 - 1.02)	0.4	0.23
22q	1.00 (1.00 - 1.03)	0.02	0.81

Supplementary Table 4.4 Multivariate Analysis BIG 1-98 wGII by Cytoband, N=538

Variable	HR (95%CI)	p.val	zph
wGII	1.05 (1.01 – 1.09)	0.02	0.767
1p	1.03 (1.01 - 1.04)	<0.001	0.03
1q	1.01 (1.00 - 1.02)	0.13	0.04
2p	1.02 (1.01 - 1.03)	0.002	0.0554
2q	1.00 (0.99 - 1.02)	0.7	0.922
3p	1.01 (1.00 - 1.03)	0.02	0.558
3q	1.02 (1.01 - 1.03)	0.002	0.34
4p	1.01 (1.00 - 1.03)	0.05	0.53
4q	1.01 (1.00 - 1.03)	0.05	0.377
5p	1.01 (0.99 - 1.02)	0.3	0.91
5q	1.01 (0.99 - 1.02)	0.2	0.725
6p	1.00 (0.99 - 1.02)	0.5	0.5
6q	1.01 (1.00 - 1.02)	0.03	0.96
7p	1.01 (1.00 - 1.02)	0.09	0.789
7q	1.00 (0.99 - 1.01)	0.7	0.535
8p	1.02 (1.01 - 1.02)	<0.001	0.82
8q	1.01 (1.00 - 1.01)	0.002	0.706
9p	1.02 (1.01 - 1.03)	0.002	0.351
9q	1.00 (0.99 - 1.01)	0.9	0.89
10p	1.02 (1.00 - 1.03)	0.004	0.943
10q	1.02 (1.00 - 1.03)	0.02	0.88
11p	1.01 (1.00 - 1.02)	0.02	0.043
11q	1.01 (1.00 - 1.02)	<0.001	0.0501
12p	1.03 (1.01 - 1.04)	<0.001	0.22
12q	1.02 (1.01 - 1.03)	<0.001	0.726
13q	1.01 (1.00 - 1.02)	0.1	0.823
14q	1.01 (1.00 - 1.03)	0.03	0.16
15q	1.01 (1.00 - 1.01)	0.2	0.44
16p	1.0 (0.99 - 1.01)	1	0.89
16 q	1.01 (1.00 - 1.02)	0.1	0.66
17p	1.02 (1.01 - 1.03)	<0.001	0.66
17q	1.01 (1.00 - 1.02)	0.008	0.2
18p	1.02 (1.00 - 1.03)	0.02	0.32
18q	1.01 (1.00 - 1.02)	0.2	0.713
19p	1.02 (1.00 - 1.03)	0.006	0.968
19q	1.01 (1.00 - 1.02)	0.09	0.77
20p	1.01 (1.00 - 1.02)	0.03	0.95
20q	1.01 (1.00 - 1.01)	0.005	0.78

21p	1.00 (0.99 - 1.02)	0.6	0.87
21q	1.01 (0.99 - 1.02)	0.4	0.23
22q	1.00 (1.00 - 1.03)	0.02	0.81

Supplementary Table 4.5 Multivariate Analysis in METABRIC wGII by Cytoband N=1103

Variable	**HR (95%CI)**	**p.val**	**zph**
wGII	1.03 (1.00 – 1.09)	0.02	0.46
1p	1.02 (1.00 - 1.03)	0.02	0.028
1q	1.01 (1.00 - 1.02)	0.1	0.26
2p	1.02 (1.00 - 1.03)	0.02	0.204
2q	0.99 (0.98 - 1.01)	0.4	0.962
3p	1.00 (0.99 - 1.02)	0.5	0.56
3q	1.01 (0.99 - 1.02)	0.3	0.226
4p	1.01 (0.99 - 1.02)	0.3	0.172
4q	1.01 (0.99 - 1.02)	0.5	0.686
5p	1.00 (0.99 - 1.01)	1	0.414
5q	1.00 (0.99 - 1.01)	0.9	0.809
6p	0.99 (0.97 - 1.01)	0.2	0.477
6q	1.00 (0.99 - 1.02)	0.6	0.82
7p	1.00 (0.99 - 1.01)	0.9	0.311
7q	1.00 (0.99 - 1.01)	0.8	0.506
8p	1.01 (1.00 - 1.02)	0.1	0.296
8q	1.00 (0.99 - 1.01)	0.9	0.578
9p	1.01 (0.99 - 1.02)	0.3	0.387
9q	0.98 (0.97 - 1.0)	0.02	0.65
10p	1.01 (1.00 - 1.02)	0.2	0.535
10q	1.01 (1.00 - 1.03)	0.2	0.971
11p	1.00 (0.99 - 1.02)	0.6	0.055
11q	1.01 (1.00 - 1.02)	0.02	0.0128
12p	1.02 (1.00 - 1.03)	0.02	0.261
12q	1.02 (1.01 - 1.04)	<0.001	0.446
13q	1.00 (0.99 - 1.01)	0.9	0.29
14q	1.01 (1.00 - 1.02)	0.1	0.927
15q	1.00 (0.99 - 1.01)	0.6	0.57
16p	1.00 (0.99 - 1.01)	0.9	0.714
16 q	1.00 (0.99 - 1.02)	0.4	0.99
17p	1.01 (1.00 - 1.03)	0.02	0.42
17q	1.01 (1.00 - 1.01)	0.3	0.446
18p	1.01 (0.99 - 1.02)	0.4	0.502
18q	0.99 (0.98 - 1.01)	0.5	0.42
19p	1.01 (1.00 - 1.02)	0.2	0.85
19q	1.00 (0.99 - 1.02)	0.4	0.925
20p	1.00 (0.99 - 1.01)	0.4	0.81
20q	1.01 (1.00 - 1.01)	0.1	0.706
21p	1.00 (0.99 - 1.01)	0.9	0.95
21q	1.00 (0.98 - 1.02)	1	0.207
22q	1.01 (1.00 - 1.02)	0.1	0.76

zph test the proportional hazards assumption for each covariate

Chapter V

1 Summary of Main Findings

Tumour heterogeneity limits our understanding of the biology of cancer. It makes it difficult to align clinical and molecular markers to prognostic, predictive and therapeutic models. As a result, replicating the process of tumour evolution by means of the discrete-time branching process can reveal information that cannot otherwise be measured with current immunohistochemistry and sequencing technologies.

I was able to corroborate in 1,800 patients the ranges of average selective advantage s and driver mutation rate u that likely explain human malignancies, ($s = \{0.001,0.01\}$ & $u = \{3.14\text{x}10^{-5}, 3.14\text{x}10^{-4}\}$) as showed by Bozic el al. [43]. Additionally, I demonstrated the power of the discrete-time branching process models to connect mutational processes with tumour growth dynamics. Using these models, I reconstructed the clonal histories of patients that provided biological insight with clinical value.

1.1 The Discrete-Time Branching Process is a Robust and Flexible Model to Reconstruct Tumour Evolution in Real Patients

In all the tumours analysed, from multiple cancer subtypes sequenced by different assays using two clonality callers, the reported values of s and u were corroborated ($s = \{0.001,0.01\}$ & $u = \{3.14\text{x}10^{-5}, 3.14\text{x}10^{-4}\}$). These parameters were corroborated even in 56% of the neutral cases (only one clone captured) by fitting the neutral tail to the equation reported by Bozic et al. [58].

Although, the type of sequencing assay is important for clonal detectability, the discrete-time branching process models were able to handle amplicon sequencing covering a limited number of genes (as shown in BIG 1-98). This illustrates the power and flexibility of the discrete-time branching process to reconstruct evolutionary histories in any form of bulk sequencing assay.

In TCGA, the predicted clonality of the positive selection models clustered with the reported overall survival (Figure 5.1.a). This showed that overall survival was influenced by RGS diversity and tumour fitness, which was present in the predicted recurrent phylogenies per subtype. The degree of *branching* clustered with subtypes of similar survival are shown in Figure 5.1.a.

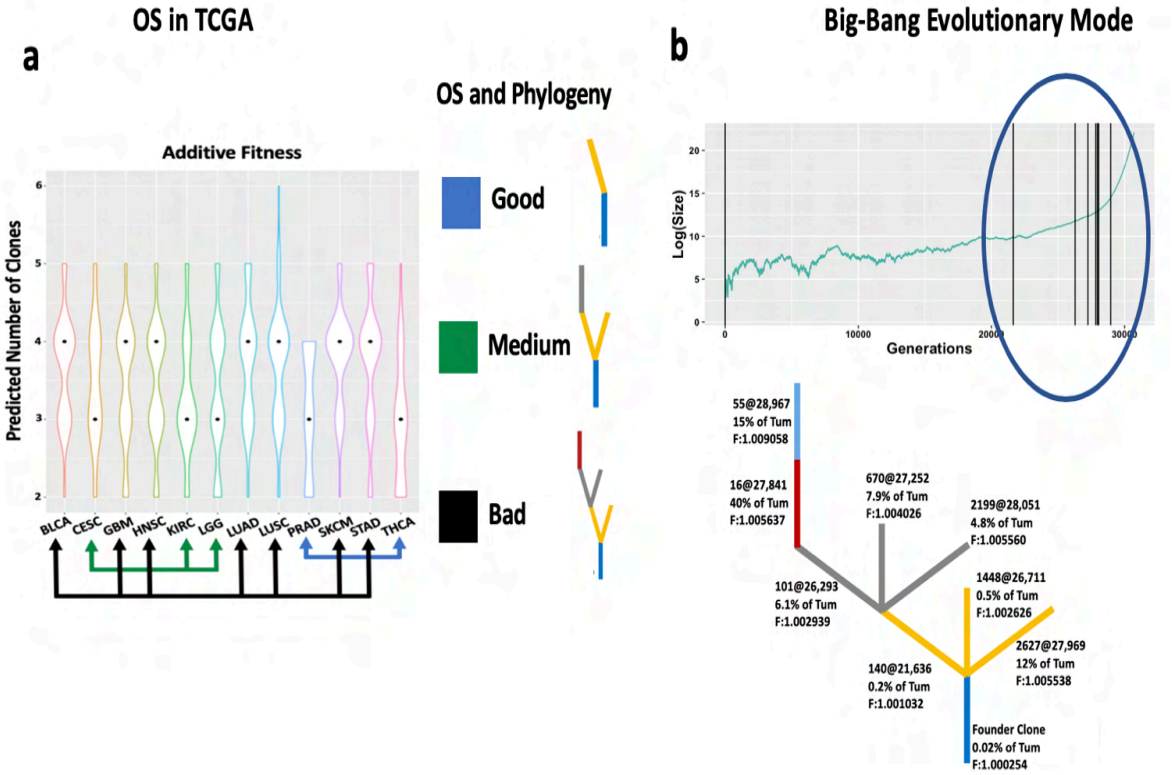


Figure 5.1 Fits on TCGA. a, the predicted clonality distribution of the additive fitness model using the minimum Euclidean distance method. Subtypes are clustered by overall survival and the top recurrent phylogeny of their category. **b**, *Big-Bang* evolutionary mode was identified in 17% of the cases. In the top figure, black lines refer to the emergence of new driver clonal expansions. The bottom figure shows phylogeny annotation for every clone, including fitness, fraction of tumour, clone ID and time of successful expansion. Number of drivers are colour coded.

A subset of the highly branched topologies, ~17% of cases, showed a *Big-Bang* evolutionary mode, as in the example shown in Figure 5.1.b. In this evolutionary trajectory, the founder clone starts out with low fitness and slow growth, but in a short period the tumour acquires a burst of driver alteration, increasing its fitness and making a few clones highly abundant in the tumour.

The consequence of this *Big-Bang* evolutionary mode is that the tumour stays in a low fitness state for a considerable amount of time, allowing for low-frequency drug resistance mutations to emerge. Then, the tumour switches to an aggressive proliferation by the fitness increase, making subsequent driver alterations undetectable. As a consequence, the tumour can take advantage of the weak evolutionary mode by the accumulation of drug resistant clones. Then it switches to an aggressive state by the rapid and sudden accumulation in fitness, leading to a tumour formation with increased proliferation higher in pre-existing drug resistant clones.

TRACERx NSCLC reported the phylogenies for all the cases combining single nucleotide variations and copy number alterations. The branching process models were able to handle this scenario in which the driver signal is mixed with driver somatic point mutations and driver copy number alterations. The predicted phylogenies showed a good representation overall, having a discrepancy of 2 - 4 branches that can be accounted for by the false positive calls of

the passenger tail. In this cohort, approximately 18% of the cases showed a *Big-Bang* evolutionary mode as described previously.

Some of the recurrent topologies in TRACERx NSCLC showed a high abundance of hyper mutants as shown in Figure 5.2.a and that the increased mutation rate was able to be recovered. This effect is likely to occur due to the increased mutational burden caused by genome-wide copy number alterations. In this simulation, detectable clones occur in a short window similar to the *Big-Bang*, but they remain at low frequencies. This simulation was the top recurrent in the increased mutation rate model which showed the early expansion of hypermutant lineage 235, giving it an advantage to seed more driver diversity than its competitors (Figure 5.2.a).

A similar effect was present in the top recurrent simulation additive fitness model with lineage 67 outcompeting the rest (Figure 5.2.b). As a result, it is most likely to infer a single branch phylogeny for these tumours. This type of long trunk which diversifies late is reported for lung adenocarcinoma McGranahan et al. [52].

TRACERx NSCLC highlighted how modifications to the discrete-time branching process, such as the increased mutation rate model, can provide additional information about the likely clonal trajectories of the tumours–which provided significant association with clinical outcome in BIG 1-98.

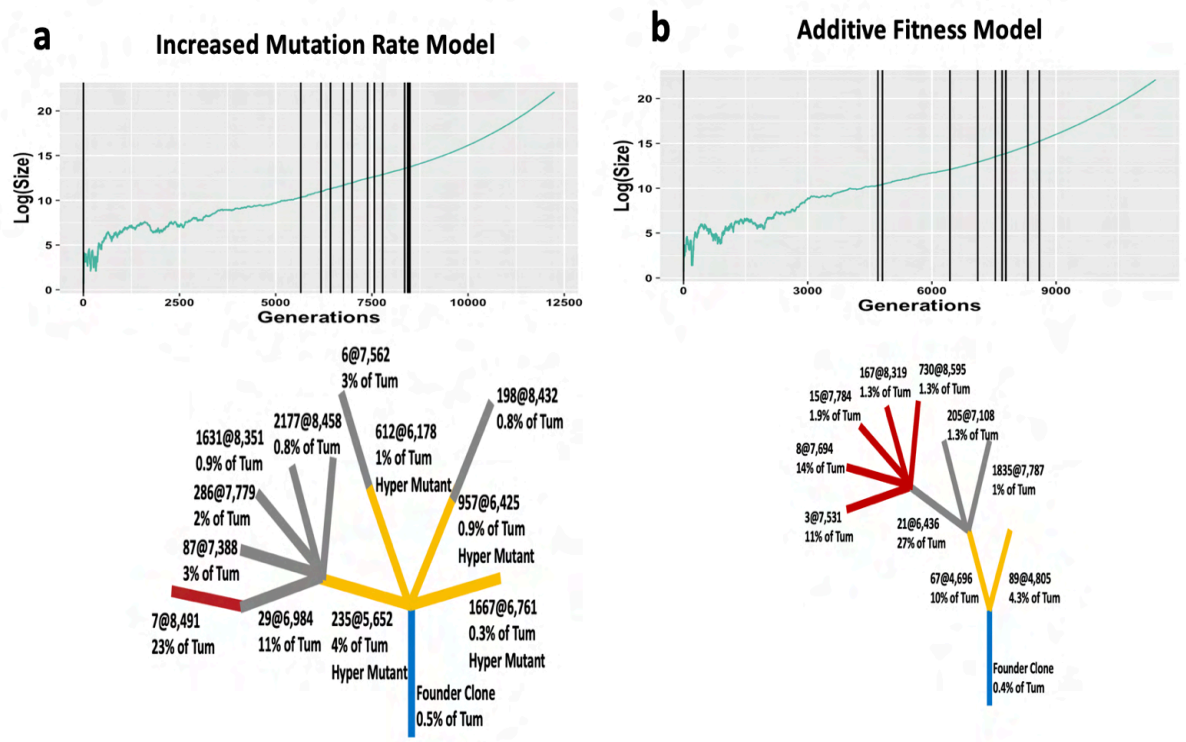


Figure 5.2 Special cases on TRACERx. Recurrent simulations of the (**a**) increased mutation rate and (**b**) additive fitness models. In the top part, tumour growth with black lines indicating times of new driver clones expanding. In the bottom part, the phylogeny with annotations for every clone, including hyper mutant status (for **a**), fitness, fraction of the tumour, clone ID and time of successful expansion colour coded by number of drivers: blue, yellow, grey and red correspond to $k = 1, 2, 3$ & 4.

Both simulations in Figure 5.2 reflect the pattern that is suspected in this cancer subtype with a long branch that eventually acquires diversity. The key point is that this phylogeny is measurable and thus better explained by weak average selective advantage *s*, high driver alteration *u* or both.

BIG 1-98 showed an interesting dynamic in the cases with TP53 mutations, PIK3CA mutations, a combination of both or neither. It showed that fitness is significantly associated with distant recurrence in weighted Cox models stratified by mutational profile. The cases with no TP53 and PIK3CA can have genome wide copy number alterations and accumulations of point mutations. As shown in Figure 5.3 in red, these cases have similar distributions in the distant recurrences and non-distant recurrences. Here distant recurrences are caused by the random activation of low frequency neutral fitness programs such as drug resistance. Ki-67 remains similar suggesting no increased proliferation.

TP53 cases coloured in green showed an increased median fitness in the distant recurrences which was supported by the values Ki-67. Alterations in cases with TP53 can be genome wide copy number alterations and point mutations. Cases with TP53 displayed a fitness switch as they started with a low fitness, acquiring a sizeable fraction of potentially drug resistant cells. This suggest that TP53 can lead to macro-evolutionary leaps as reported by [16, 28, 96, 99, 210]. The risk of PIK3CA mutations is that it is a weak driver if the tumour stays in that stage, and thus during the course of time, advantageous passenger alterations such as drug resistance may accumulate progressively. This explains why there is less fitness variance in distant recurrence cases with PIK3CA.

In contrast, PIK3CA mutation in blue is a weak driver [189] that did not co-occur with genome wide copy number alterations in BIG 1-98 and METABRIC. The risk of PIK3CA mutations is that it is weak driver if the tumour stays in that stage, and thus during the course of time, advantageous passenger alterations such as drug resistance may accumulate progressively. This explains why there is less fitness variance in distant recurrence cases with PIK3CA.

Cases with TP53+PIK3CA showed increased variance in fitness and Ki-67. Genome wide copy number alterations were less than in TP53 alone, indicating that PIK3CA did not allow cells to become fully genomically unstable.

The use of the discrete-time branching process was able to provide biological insight with clinical value by showing how somatic alterations manifest in fitness changes that significantly associated with clinical outcome.

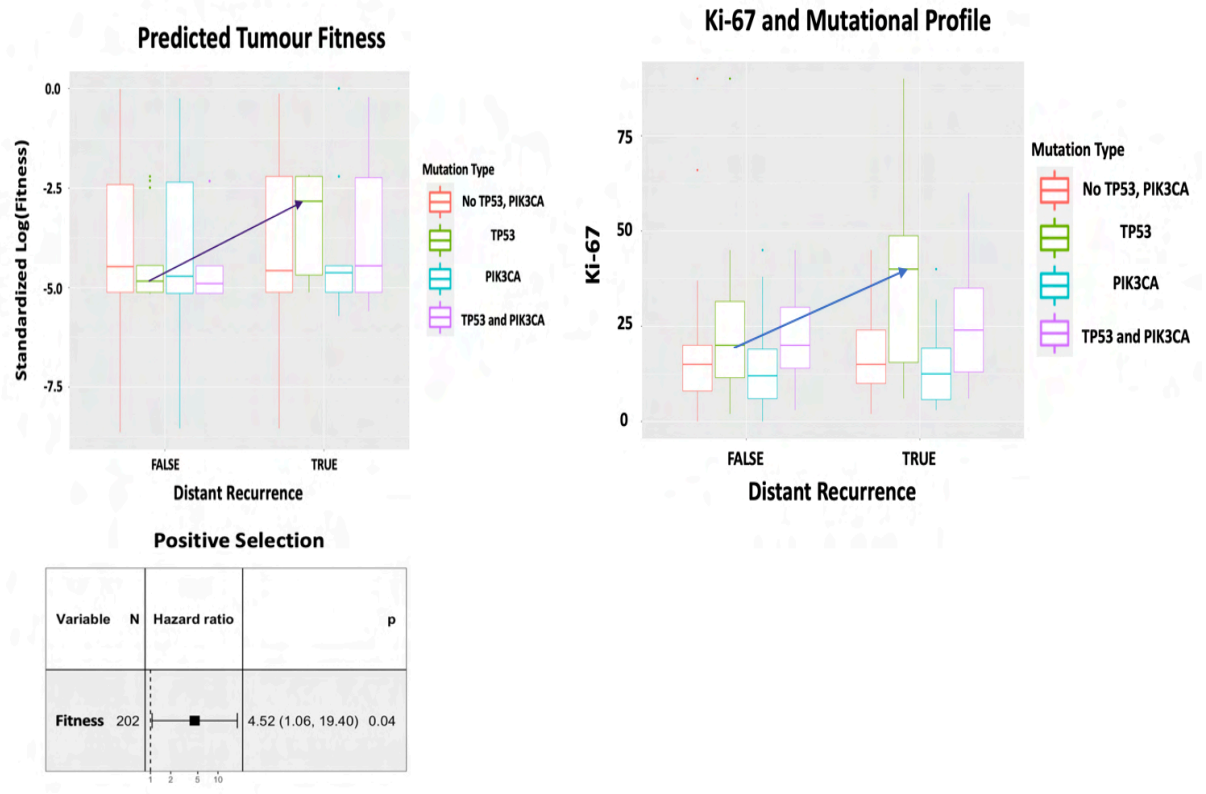


Figure 5.3 Mutational effects in BIG 1-98. Top section shows the distributions of predicted fitness based on mutational profile and the distribution of Ki-67. Bottom section shows the forest plot of fitness effects using a weighted Cox model stratified by mutational profile.

In CASCADE melanoma showed a pattern of clonal dissemination from primary to metastases. The stickbreaking model showed the same evolutionary trajectory in the primary '61-1', Liver Central and Brain Right samples (all converged in the same simulation). The same occurred in the increased mutation rate model. The suspected dissemination trajectory was from the primary to the liver and from the liver to the brain. Moreover, both models recover the same primary phylogeny when all the regions are merged together.

The discrete-time branching process was able to recover a pattern of dissemination in primary and metastases by reconstructing their phylogenies. This shows the power of the method to identify similarities between samples (primary and metastasis) that are affected by different sources of bias in sampling collection (purity, evolutionary time, etc.).

1.2 The Discrete-Time Branching Process Provides Biological Insight

Using the branching process models, I identified that a tumour's age has predictive power to identify the initial value of s (accounting for average division rates from 1 -5 days) as suggested by [165]. Distinctive patterns between different parameter combinations were observed in the additive fitness model and the increased mutation rate model, though the stickbreaking model had around 25% degree of overlap between parameter combinations. A similar effect was observed in the total number of clones, which has predictive power to determine the initial value of s. However, obtaining a representative sample of clonal diversity remains a significant challenge.

I showed that cancer cell fractions of simulation outcomes with different combinations of initial conditions of s and u overlapped at the 10% CCF cut-off used in standard sequencing assays. The impact of this overlap is the accurate determination of initial values of s and u in real tumours. Moreover, I showed that this overlap occurs in real cancer cell fractions as shown in the TCGA dataset identifying subtypes with similar cancer cell fraction values. Accounting for this effect, I showed that at the 10% CCF cut-off, tumours can be classified into two categories: strong fitness with low mutation rate ($s = 0.1$ & $u = \{3.14\text{x}10^{-7}, 3.14\text{x}10^{-6}\}$ or moderate/weak fitness with high mutation rate vs $s = \{0.001, 0.01\}$ & $u = \{3.14\text{x}10^{-5}, 3.14\text{x}10^{-4}\}$). The latter being the most likely parameter combination in human malignancies.

Overall, s controls the expansion rate of clones/tumour by defining their net-growth, depicted in colours blue and grey in the heatmap in Figure 5.4.a. The amount of diversity is influenced by the combination of s and u. The former is due to its control of the total number of expansions and the latter is due to increasing the frequency of new driver mutants.

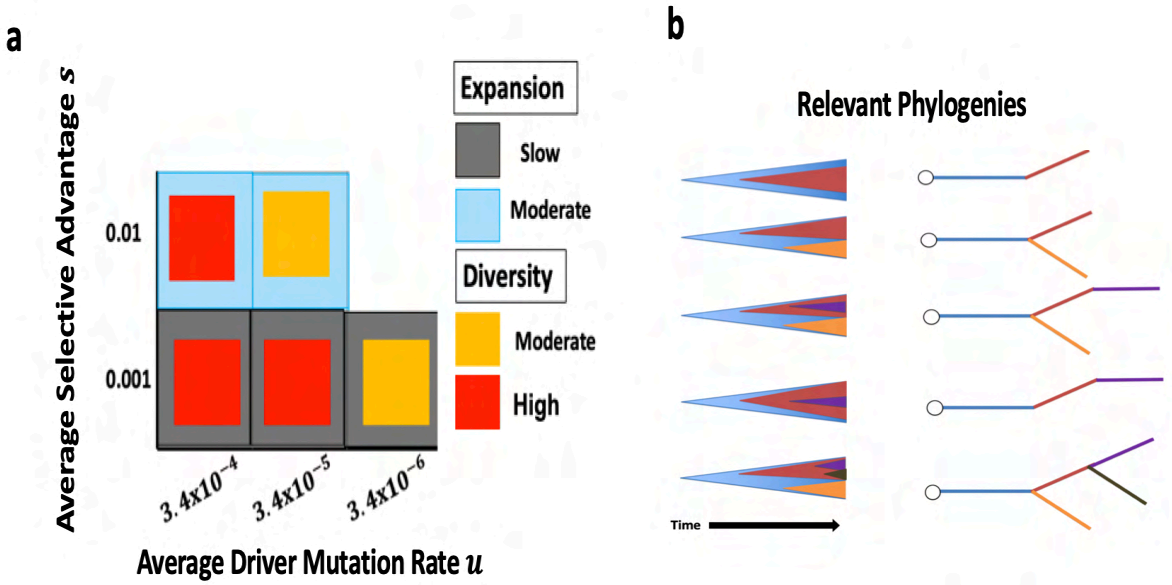


Figure 5.4 Growth dynamics of s and u at a 10% CCF cut-off representing human malignancies. Results of the likely values of s and u in human malignancies. **a,** the dynamics of expansion and diversity for each parameter combination. **b**, Muller plot and recurrent topologies colour coded by clone.

In addition, I identified the recurrent phylogenies in the three models implemented at the aforementioned cut-off. The lower value of s and the higher value of u leads to higher clonal diversity with unique number phylogenies, shown in Figure 5.4.b. The predicted recurrent topologies suggested by the branching process models had concordance with the fits performed in TCGA.

I showed that the number of drug resistant cells correlates with tumour size, with 0.5 cm^3 being the critical size in which the number of drug resistance cells start to emerge, shown in Figure 5.5.a.

The average selective advantage s is inversely proportional to the increase in number of drug resistant cells, because it determines the net-growth and number of divisions. Thus, s

modulates the increase of the number of drug resistant cells. The risk of a weak s is due to an increased accumulation of drug resistant cells that limits therapeutic efficacy.

The average driver mutation rate u does not influence the number of drug resistant cells, though it can impact their driver makeup. As illustrated in Figure 5.5.b, most cases show that the expected accumulated number of drivers is two, excluding weak s and high u combinations that carry three drivers.

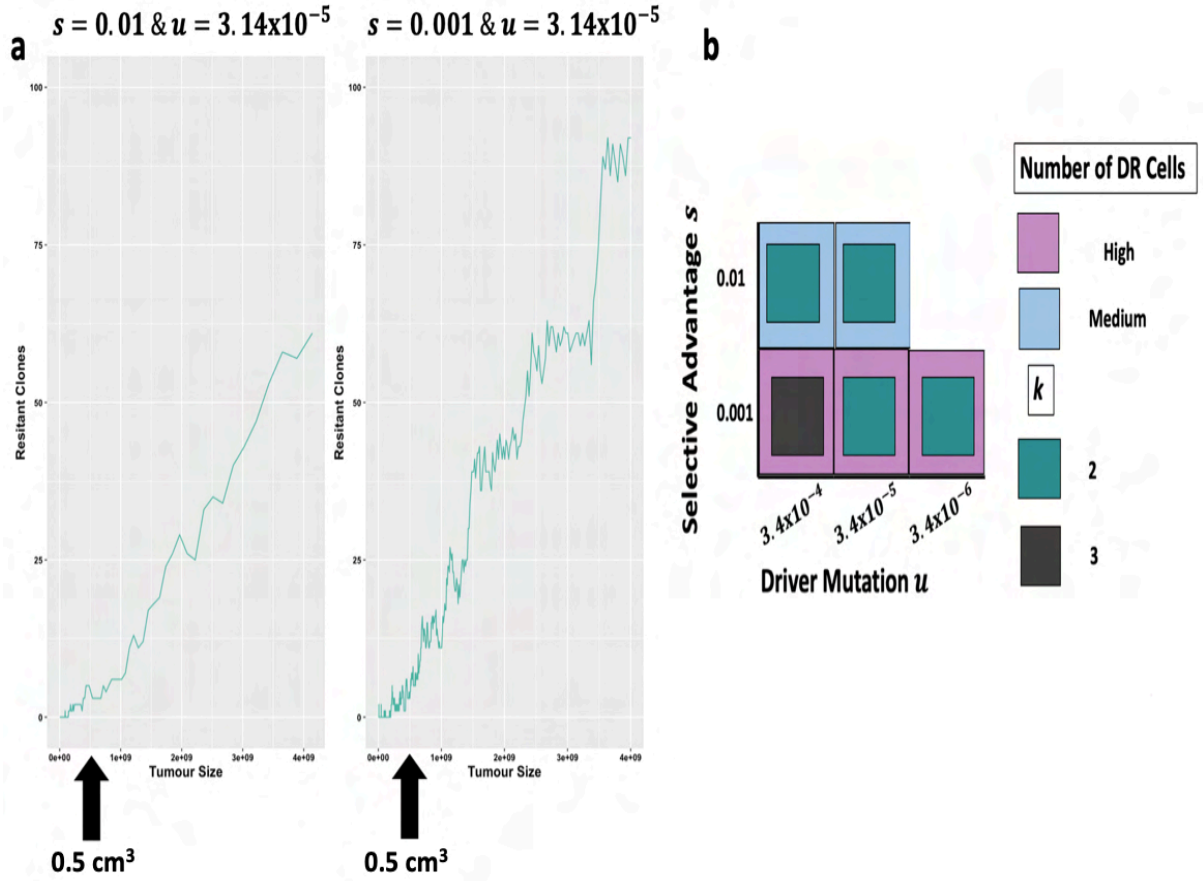


Figure 5.5 Drug resistance dynamics in the positive selection models. The values of s are based on likely values of human malignancies. **a**, shows the emergence of drug resistant clones over time for different values of s. **b**, colours purple and blue refer to the amount of drug resistant cells expected by the value of s. Colours green and grey refer to the number of accumulated drivers.

I showed the prevalence of the detectable passenger signal at the 10% CCF cut-off using different combinations of s and u. The number of passenger alterations is a function of the clone size and the expected number of passengers is $\nu C_{i,j,k}$ as reported by Bozic et al. [58]. However, because the passenger mutation rate is very high and its effect neutral, the passenger signal is more dynamic than the driver signal. Hence its detectability varies according to the number of clones measured and the sequencing assay.

The measurement of the number of passengers depends on the sequencing coverage, depth, number of clones measured and lineage survival probability δ. Therefore, even when the total number of passengers is proportional to clonal size, δ affects how the passenger signal is accumulated. As a result, the measurable passenger signal is proportional to δ because of the

number of divisions and lineage fixations. This was shown with the simulated number of passengers using the neutral model and with the median number of passengers reported by Bozic et al. [58].

The discrete time branching process models provide a way to better understand the role of the CCF in clonal detectability and allows for the simulation of the conditions in which a deeper CCF is beneficial. This can improve longitudinal study design to determine the most appropriate sequencing assay and regions to collect.

In addition, the branching process models provide a database that can be used to simulate tumour heterogeneity and help the development of clonality tools. A test dataset of benchmark multiple clonality tools and phylogenetic models is needed in the field.

1.3 Analytical Solutions Can Be Applied to Different Mutation Process in Cancer

I derived analytical solutions for the clonal additive fitness model that described expected tumour size estimated by accounting for all clones. My approach avoids computing the inhomogeneous first-order differential equation by estimating $\hat{\tau}_k$, the time in which any given clone is going to successfully expand. With $\hat{\tau}_k$, I was able to identify expected tumour phylogenies at any tumour fraction composition. I provided as an example the expected measurable topologies at a 10% tumour fraction composition shown in the following figure.

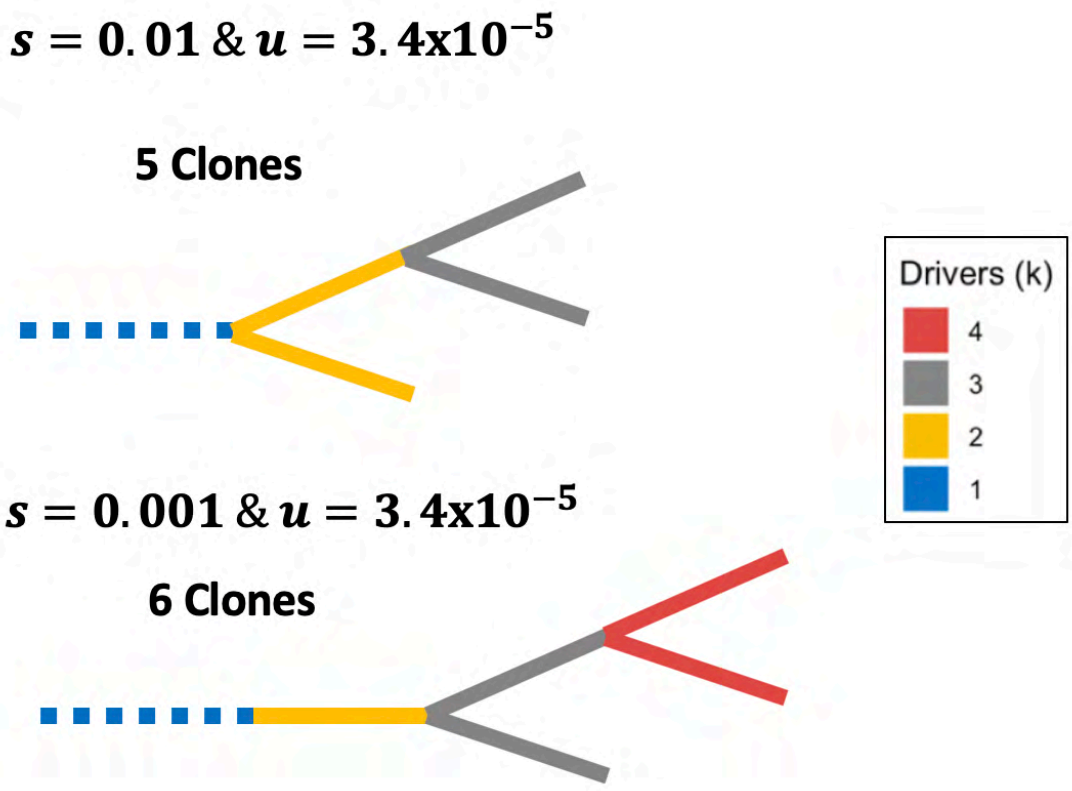


Figure 5.6 Most likely topologies at a 10% tumour fraction composition. Topologies were generated with the expected solutions of $E[C_{i,j,k}(t+1)], E[N(t+1)]$ and $\hat{\tau}_k$ as shown in Chapter II. Colours represent the number of accumulated drivers colour coded by the number of accumulated drivers. A dashed line means that the clone is undetectable, though due to inheritance, its driver alteration can be recovered.

Figure 5.6 shows the expected topologies for the moderate and weak average selective advantage s that can be recovered based on mutations with >10% CCF (the commonly used sequencing cut-off). This figure exemplifies the expected topologies that can be measured in ideal bulk sequencing. However, deviations of these topologies can occur due to the technical and biological biases that complicate measuring clonal subpopulations in tumours. The phylogenies recovered using the analytical solutions are in concordance with the number of clones and drivers of the simulations of the additive fitness model.

The analytical solutions can be modified to explore topologies at different tumour fraction compositions and can be applied to any parameter combinations of s and u at any tumour size N. The analytical solutions can also predict the phylogenies of different mutational processes, such as copy number or epigenetic alterations, if their driver mutation rate is known.

Although, Bozic et al. [43] showed that the exponential model is a good approximation of tumour growth, I provided the analytical solution considering a carrying capacity that can be used to better replicate solid tumour formation.

2 Discussion

Clonal evolution is a skewed process with few clones describing most of the tumour composition. Even though these clones dominate tumour expansion, their clonal makeup and relatedness is difficult to recover as a result of inheritance bias. Intratumor heterogeneity further complicates clonal evolution reconstruction when using sequencing technologies, by providing a large array of mutations that do not have a role in describing positive selection. Current clonality tools are prone to this effect and do not correct for it, affecting clonal ancestry reconstruction and phylogeny.

The discrete-time branching process in a clonal configuration tackles this problem by replicating clonal evolution. It does this by recreating the mutational frequencies affected by inheritance that are comparable to ploidy corrected frequencies obtained by sequencing. However, the way in which to compare simulated and sequenced cancer cell fractions is poorly understood.

I showed that by minimising the Euclidean distance between simulated and sequenced cancer cell fractions, the initial values of average selective advantage s and average driver mutation u can be determined, thus allowing for the approximation of real tumour evolution with the discrete-time branching process. Moreover, depending on how the sequenced sample is affected by noise and sampling bias, the correct evolutionary trajectory can be recovered. With the use of a fitting procedure using the minimum Euclidean distance, I was able to corroborate the range of the average selective advantage s suggested by Bozic et al. [42, 43]. Specifically, the fits performed on 1,800 patients suggest that the value of s is approximately 0.001 to 0.01 and the range of the average driver mutation rate u is approximately $3.14\text{x}10^{-6}$ to $3.14\text{x}10^{-4}$. With these ranges of parameters, a significant number of unique evolutionary trajectories descriptive of the numerous clonal configurations observed in the clinic can be recovered.

The discrete-time branching process for the association of the random process of accumulating mutations with clonal growth dynamics, enabling a better depiction of clonal evolution which is currently lacking. This was shown in the cases that displayed a *Big-Bang* evolutionary mode, 17% in TCGA and 18% in TRACERx NSCLC, and additionally is reported in different subtypes [86, 110]. The relevance of this evolutionary pattern is in the consistent similarity of

the genetic makeup caused by a burst of driver mutations occurring in a short window of time, leading to dominant clones representing a significant portion of the tumour.

This vulnerability is caused by truncal main alterations which allow an alignment to globally impact therapeutic targets. In BIG 1-98, the stickbreaking model showed a recurrent pattern occurring in 9% of the distant relapses. The founder clone started with very low fitness, potentially acquiring passenger mutations such as drug resistance. Next, hyper selection occurred by cytoband alterations, making an evolutionary leap in fitness and rendering the tumour aggressive. These examples demonstrate an advantage of using the branching process in reconstructing tumour evolution that is not available in current approaches.

In addition, the growth dynamics can be correlated with markers of disease progression, such as stage, grade, nodal status and Ki-67, that enable molecular and clinical markers to be linked. I showed the impact of reconstructing tumour evolution in three studies. 1) in TCGA, I showed how the fits correlated with overall survival in the recovered clonality distributions and in the phylogenetic topologies. 2) in BIG 1-98, I showed the role of fitness in distant recurrence cases harbouring TP53 that were also manifested in their Ki-67 status. Although the clinical benefit of tumours with PIK3CA is widely known [189], in the BIG 1-98 cohort, distant recurrences with PIK3CA are due to the lack of fitness allowing for passenger alterations such as drug resistance to emerge. 3) in CASCADE melanoma, I showed a pattern of dissemination from the primary site to the liver and brain that converged in the parameters and evolutionary trajectory.

The models of tumour evolution were designed to be compared at any instance of the disease progression. A significant amount of cancer research utilises mice models and cell lines in which the models reported here can approximate their evolution if the sequencing is available. The number of simulations and snapshots were generated with a comprehensive range to account for most of the tumour growth manifestations. Simulations higher than the average selective advantage of $s = 0.1$ do not represent a computational burden and are easy to generate. The features stored in the positive selection models, such as RGS diversity, number of drivers and clonal and cancer cell fractions, can be used to correlate with clinical markers in order to improve the quality of the fits. Tumour size, stage, grade and nodal status provide information about disease progression that can be used to select an evolutionary trajectory.

I showed that the discrete-time branching process can be used to approximate tumour evolution in patients using sequencing data. I also showed the relevance in evaluating neutrality in cases where a single subpopulation was measured. For example, the values of the average selective advantage s and average driver mutation rate u can be recovered by fitting data to the passenger tail as shown by Bozic et al. [43] and in BIG-1-98. The values of s and u from pre-computed simulations can provide a picture of clonal makeup. If clinical markers of diversity, are available, they can be used to improve evolutionary trajectory predictions.

When more than one subpopulation was measured, allelic frequencies had to be converted to cancer cell fractions by ploidy correction and, if possible, have the passenger tail removed. With the cancer cell fractions recovered, the comparison with the three positive selection models can be done using the minimum Euclidean method with false discovery correction.

The discrete-time branching process can also aid in recovering tumour evolution when no sequencing is available. This can be achieved by using the analytical solutions of expectation and variance of clonal and tumour sizes by means of $\hat{\tau}_k$. In this setting, the clonal frequencies

can be compared to a commonly used technique that measures clonal diversity such as single cell sequencing, Fluorescence in situ hybridization, FACS or CyTOF. These techniques can provide the number of unique molecular drivers that can be used as a proxy for driver alterations and aid in selecting the values of s and u that best correlate with the observations made.

There is only one study that uses the branching process to make inferences about clonal evolution and estimates about fitness and mutation rates [88]. They used the branching process to predict the variant allelic frequency distribution and their approach was applied to multiple cancer subtypes, including selected cases from TRACERx NSCLC. Their reported fitness values for this cohort agrees with the values recovered using my approach. However, the mutation rates differ, as they reported a lower mutation rate of $1x10^{-7}$to $1x10^{-6}$. This difference may be due to the number of false positive clones in the clonality calls. TRACERx NSCLC did not correct for the passenger tail, therefore leading to an increased clonality which may explain the increased mutation rate in our fits.

3 Future Work

3.1 Future Work: Developing Analytical Solutions

The framework developed here can be used as a backbone to further develop analytical solutions using the branching process. Analytical solutions can be implemented to build phylogenies based on copy number, epigenetic and passenger alterations if their mutation rates are known.

An area in which the branching process can further be explored is by modelling tumour microenvironments. In order to achieve this, it is necessary to have an estimate of microenvironmental fitness variation to calibrate the model to. If this value is unknown, the branching process can provide an estimate of this variation. For instance, this can be achieved by running the model with a different approach. In order to achieve this, you need to run the model with a different approach, where the input of the model becomes a target-fitness distribution. Importantly, this target fitness distribution can be acquired using the simulation outcomes I provided using the three positive selection models. Additionally, in order to recover this target distribution, s and u must vary randomly as opposed to being input parameters. Thus, the target fitness distribution is described by an array of clonal makeups, with varying values of s and u, as a proxy of microenvironmental selection.

3.2 Future Work: Improving the Comparison Between Simulated and Real Cancer Cell Fractions

Fits can be improved by using clinical data, such as tumour size, and by adding clinical variables that provide uniqueness to the fits, allowing for finer approximations. Although the branching process showed it can recover tumours qualitatively, a validation in estimating real tumour sizes will improve the accuracy of the model.

Additionally, fits can be improved by removing false positive clonality calls by pruning the passenger tail. The algorithms used by previous authors to determine the cancer cell fractions, PyClone and ExPANdS, are prone to this effect.

Ultimately, the ploidy correction can be used without running clonality callers before being fed into the minimum Euclidean distance method.

A new approach can be implemented by fitting the driver and passenger signals together. This requires fitting the diver signal first and is subject to the clones measured in order to identify which clones better recover the passenger tail. In this way, all the sequencing information can be used to establish clonal relationships.

3.3 Future Work: Adding More Models and Increase the Number of Samples

Simulations can be expanded to sizes greater than 4 cm^3. This increased mutation rate model requires a higher sample size, but generating instances of it is computationally intense. Similarly, the passenger model requires to be modelled in larger clonal sizes, up to two billion cells, in order to be comparable to the passenger tail distributions.

So far, we have generated three models of positive selection, however it is worth exploring a model that changes *s* and *u* stochastically for every driver accumulated, similar to [127]. This will explore more evolutionary trajectories and potentially be more realistic in approximating the process of clonal evolution. However, due to the potential diversity in this model, a bigger sample size is required to have a representation of the parameter that can be used to compare with sequencing.

3.4 Future Work: Compare to More Patient Samples and Include Pre-Clinical Models

Although I tested the model in 1,800 human tumours, a larger sample size will be beneficial to establishing consistent evolutionary signatures per cancer subtype. The breast cancer subtype from TCGA was missing from the analysis of Andor et al. and can substantially complement the pan-cancer analysis.

The model can be extrapolated to fit mice and cell line tumours, as the branching process describes tumour evolution regardless of the organism and tumour type. The database generated in Chapter III was designed to replicate malignancies with different starting values of fitness. Therefore, evolved cancer cells such as patient derived xenografts or organoids used to study tumour progression both *in-vitro* and *in-vivo*, can also be modelled with the current framework.

3.5 Future Work: Single Cell Technology and Circulating Tumour DNA (ctDNA)

The model implemented here can be extrapolated to single cell sequencing technologies. The challenge in this assay is to determine the number of unique clusters harbouring *k* driver alterations to compare with the branching process. However, it provides a better signal because it avoids the underrepresentation of mutational frequencies present in bulk sequencing.

The three branching process models described here can be applied to ctDNA, as the benchmarking showed that minimum distance comparisons can deal with missing information associated with ctDNA samples.

The branching process can help to determine if ctDNA samples can be linked to a given phylogeny/clone from the primary. This can be used to monitor disease progression and minimum residual disease and evaluate which clones are not responding to therapy.

3.6 Future Work: Dissemination and Metastasis

The branching process can provide estimates and likelihood of tumour dissemination. This can be achieved by random sampling of the primary tumour and by evaluating the likelihood of growth and driver makeup. This can be helpful in studies such as CASCADE, in which primary tumour and metastases are sequenced, enabling the calculation of time when clones disseminate and metastasise. This will allow for the prediction of a window of risk for a given cancer subtype.

In addition, this process can provide estimates of how early cellular dissemination occurs by comparing the time in which metastases were detected. This approach can be used as a predictive and prognostic model.

3.7 Future Work: Evolutionary Trajectories and Treatment

Fits performed in a primary tumour can be used to study multiple realisations of the tumour forward in time, predict tumour progression and establish how many different evolutionary trajectories can occur.

In addition, it will provide a baseline to model treatment regimens and narrow down the number of drugs and doses that can be used for testing in mice. If multiple snapshots of the tumour are taken, those can evaluate how the evolutionary dynamics of resistance changes.

4 Significance

This is the first study that explores the capability of the discrete-time branching process to reconstruct tumour evolution using patient data. It is also the first study to corroborate the values of the selective advantage s identified by Bozic et al. [43] in a representative sample of tumours. Moreover, it connects mutational process to growth dynamics of disease progression which current tools lack.

Also, it is the second study to provide a framework to recover expected tumour phylogenies analytically with the discrete-time branching process, accounting for any allelic frequency cut-off [149].

This study generated the largest database of simulated tumours storing multiple properties such as cancer cell fractions, diversity and clonal sizes. This can be used to explore multiple statistical methods to compare simulated versus real cancer cell fractions. Moreover, it can be used as a test set to evaluate clonality callers and phylogenetic tools.

It is the second tool to provide a framework to reconstruct tumour evolution in targeted sequencing with limited number of measured genes [211]. Similarly, it is the second study to recover tumour phylogenies using a sequencing approach with a limited number of genes. This type of sequencing is highly used in clinical diagnostics and can provide additional information about the properties of the tumour that can have clinical benefit.

Overall, the strength of the discrete-time branching process is in replicating clonal evolution and tumour growth. By identifying the starting values of s and u using sequencing data, we can approximate a patient evolutionary trajectory Ultimately, this allows the model to provide a picture of the hidden diversity that is not measurable using current sequencing technologies.

References

1. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* cell, 2000. **100**(1): p. 57-70.
2. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* cell, 2011. **144**(5): p. 646-674.
3. Housman, G., et al., *Drug resistance in cancer: an overview.* Cancers, 2014. **6**(3): p. 1769-1792.
4. Armitage, P. and R. Doll, *The age distribution of cancer and a multi-stage theory of carcinogenesis.* British journal of cancer, 1954. **8**(1): p. 1.
5. Armitage, P. and R. Doll, *A two-stage theory of carcinogenesis in relation to the age distribution of human cancer.* British journal of cancer, 1957. **11**(2): p. 161.
6. DiPaolo, J.A. and N.C. Popescu, *Relationship of chromosome changes to neoplastic cell transformation.* The American journal of pathology, 1976. **85**(3): p. 709.
7. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976. **194**(4260): p. 23-28.
8. Weinberg, R.A. and R.A. Weinberg, *The biology of cancer*. 2013: Garland science.
9. Nunney, L., *Lineage selection and the evolution of multistage carcinogenesis.* Proceedings of the Royal Society of London. Series B: Biological Sciences, 1999. **266**(1418): p. 493-498.
10. Barrett, M.T., et al., *Evolution of neoplastic cell lineages in Barrett oesophagus.* Nature genetics, 1999. **22**(1): p. 106-109.
11. Fearon, E.R. and B. Vogelstein, *A genetic model for colorectal tumorigenesis.* cell, 1990. **61**(5): p. 759-767.
12. Vogelstein, B. and K.W. Kinzler, *The multistep nature of cancer.* Trends in genetics, 1993. **9**(4): p. 138-141.
13. Greaves, M. and C.C. Maley, *Clonal evolution in cancer.* Nature, 2012. **481**(7381): p. 306.
14. Nowell, P.C., *Mechanisms of tumor progression.* Cancer research, 1986. **46**(5): p. 2203-2207.
15. Cahill, D.P., et al., *Genetic instability and darwinian selection in tumours.* Trends in cell biology, 1999. **9**(12): p. M57-M60.
16. Gerlinger, M., et al., *Cancer: evolution within a lifetime.* Annual review of genetics, 2014. **48**: p. 215-236.
17. Fortunato, A., et al., *Natural selection in cancer biology: from molecular snowflakes to trait hallmarks.* Cold Spring Harbor perspectives in medicine, 2017. **7**(2): p. a029652.
18. Negrini, S., V.G. Gorgoulis, and T.D. Halazonetis, *Genomic instability—an evolving hallmark of cancer.* Nature reviews Molecular cell biology, 2010. **11**(3): p. 220-228.
19. Loeb, L.A., *Mutator phenotype may be required for multistage carcinogenesis.* Cancer research, 1991. **51**(12): p. 3075-3079.
20. Loeb, L.A., *A mutator phenotype in cancer.* Cancer research, 2001. **61**(8): p. 3230-3239.
21. Tabassum, D.P. and K. Polyak, *Tumorigenesis: it takes a village.* Nature Reviews Cancer, 2015. **15**(8): p. 473-483.
22. Anderson, A.R., et al., *Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment.* Cell, 2006. **127**(5): p. 905-915.
23. Pisco, A.O. and S. Huang, *Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: 'What does not kill me strengthens me'.* British journal of cancer, 2015. **112**(11): p. 1725-1732.

24. Pisco, A.O., et al., *Non-Darwinian dynamics in therapy-induced cancer drug resistance.* Nature communications, 2013. **4**: p. 2467.
25. Brock, A., H. Chang, and S. Huang, *Non-genetic heterogeneity—a mutation-independent driving force for the somatic evolution of tumours.* Nature Reviews Genetics, 2009. **10**(5): p. 336-342.
26. Hugo, W., et al., *Non-genomic and immune evolution of melanoma acquiring MAPKi resistance.* Cell, 2015. **162**(6): p. 1271-1285.
27. Shoval, H., et al., *Tumor cells and their crosstalk with endothelial cells in 3D spheroids.* Scientific reports, 2017. **7**(1): p. 1-11.
28. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration.* Nature genetics, 2013. **45**(10): p. 1134-1140.
29. Andor, N., et al., *Pan-cancer analysis of the extent and consequences of intratumor heterogeneity.* Nature medicine, 2016. **22**(1): p. 105.
30. Morris, L.G., et al., *Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival.* Oncotarget, 2016. **7**(9): p. 10051.
31. Bell, C.C., et al., *Targeting enhancer switching overcomes non-genetic drug resistance in acute myeloid leukaemia.* Nature communications, 2019. **10**(1): p. 1-15.
32. Rambow, F., J.-C. Marine, and C.R. Goding, *Melanoma plasticity and phenotypic diversity: therapeutic barriers and opportunities.* Genes & Development, 2019. **33**(19-20): p. 1295-1318.
33. Su, Y., et al., *Phenotypic heterogeneity and evolution of melanoma cells associated with targeted therapy resistance.* PLoS computational biology, 2019. **15**(6): p. e1007034.
34. Lloyd, M.C., et al., *Darwinian dynamics of intratumoral heterogeneity: not solely random mutations but also variable environmental selection forces.* Cancer research, 2016. **76**(11): p. 3136-3144.
35. Nichol, D., et al., *Stochasticity in the genotype-phenotype map: implications for the robustness and persistence of bet-hedging.* Genetics, 2016. **204**(4): p. 1523-1539.
36. Huang, S., *Reprogramming cell fates: reconciling rarity with robustness.* Bioessays, 2009. **31**(5): p. 546-560.
37. Bashashati, A., et al., *Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling.* The Journal of pathology, 2013. **231**(1): p. 21-34.
38. Caravagna, G., et al., *Algorithmic methods to infer the evolutionary trajectories in cancer progression.* Proceedings of the National Academy of Sciences, 2016. **113**(28): p. E4025-E4034.
39. Gupta, P.B., et al., *Phenotypic plasticity: driver of cancer initiation, progression, and therapy resistance.* Cell stem cell, 2019. **24**(1): p. 65-78.
40. Dirkse, A., et al., *Stem cell-associated heterogeneity in Glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment.* Nature communications, 2019. **10**(1): p. 1-16.
41. Huang, S., *Genetic and non-genetic instability in tumor progression: link between the fitness landscape and the epigenetic landscape of cancer cells.* Cancer and Metastasis Reviews, 2013. **32**(3-4): p. 423-448.
42. Durrett, R., *Branching process models of cancer*, in *Branching process models of cancer*. 2015, Springer. p. 1-63.
43. Bozic, I., et al., *Accumulation of driver and passenger mutations during tumor progression.* Proceedings of the National Academy of Sciences, 2010. **107**(43): p. 18545-18550.

44. Chowell, D., et al., *Modeling the subclonal evolution of cancer cell populations.* Cancer research, 2018. **78**(3): p. 830-839.
45. Bozic, I. and M.A. Nowak, *Resisting resistance.* 2017.
46. Gallaher, J.A., et al., *Spatial heterogeneity and evolutionary dynamics modulate time to recurrence in continuous and adaptive cancer therapies.* Cancer research, 2018. **78**(8): p. 2127-2139.
47. Korolev, K.S., J.B. Xavier, and J. Gore, *Turning ecology and evolution against cancer.* Nature Reviews Cancer, 2014. **14**(5): p. 371-380.
48. Xue, B. and S. Leibler, *Benefits of phenotypic plasticity for population growth in varying environments.* Proceedings of the National Academy of Sciences, 2018. **115**(50): p. 12745-12750.
49. Dufty Jr, A.M., J. Clobert, and A.P. Møller, *Hormones, developmental plasticity and adaptation.* Trends in Ecology & Evolution, 2002. **17**(4): p. 190-196.
50. Price, T.D., A. Qvarnström, and D.E. Irwin, *The role of phenotypic plasticity in driving genetic evolution.* Proceedings of the Royal Society of London. Series B: Biological Sciences, 2003. **270**(1523): p. 1433-1440.
51. Burrell, R.A., et al., *The causes and consequences of genetic heterogeneity in cancer evolution.* Nature, 2013. **501**(7467): p. 338-345.
52. McGranahan, N. and C. Swanton, *Clonal heterogeneity and tumor evolution: past, present, and the future.* Cell, 2017. **168**(4): p. 613-628.
53. Raynaud, F., et al., *Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability.* PLoS genetics, 2018. **14**(9): p. e1007669.
54. Laughney, A.M., et al., *Dynamics of tumor heterogeneity derived from clonal karyotypic evolution.* Cell reports, 2015. **12**(5): p. 809-820.
55. Sottoriva, A., L. Vermeulen, and S. Tavaré, *Modeling evolutionary dynamics of epigenetic mutations in hierarchically organized tumors.* PLoS computational biology, 2011. **7**(5).
56. Jones, S., et al., *Comparative lesion sequencing provides insights into tumor evolution.* Proceedings of the National Academy of Sciences, 2008. **105**(11): p. 4283-4288.
57. Nowak, M.A., et al., *The role of chromosomal instability in tumor initiation.* Proceedings of the National Academy of Sciences, 2002. **99**(25): p. 16226-16231.
58. Bozic, I., J.M. Gerold, and M.A. Nowak, *Quantifying clonal and subclonal passenger mutations in cancer evolution.* PLoS computational biology, 2016. **12**(2).
59. Smith, J.C. and J.M. Sheltzer, *Systematic identification of mutations and copy number alterations associated with cancer patient prognosis.* Elife, 2018. **7**: p. e39217.
60. De, S. and S. Ganesan, *Looking beyond drivers and passengers in cancer genome sequencing data.* Annals of Oncology, 2017. **28**(5): p. 938-945.
61. Pon, J.R. and M.A. Marra, *Driver and passenger mutations in cancer.* Annual Review of Pathology: Mechanisms of Disease, 2015. **10**: p. 25-50.
62. Ardaševa, A., et al., *Evolutionary dynamics of competing phenotype-structured populations in periodically fluctuating environments.* Journal of Mathematical Biology, 2020. **80**(3): p. 775-807.
63. McEvoy, J., *Evolutionary game theory: lessons and limitations, a cancer perspective.* British journal of cancer, 2009. **101**(12): p. 2060-2061.
64. Dingli, D., et al., *Cancer phenotype as the outcome of an evolutionary game between normal and malignant cells.* British Journal of Cancer, 2009. **101**(7): p. 1130-1136.

65. Park, C.C., M.J. Bissell, and M.H. Barcellos-Hoff, *The influence of the microenvironment on the malignant phenotype.* Molecular medicine today, 2000. **6**(8): p. 324-329.
66. Kam, Y., et al., *Sweat but no gain: inhibiting proliferation of multidrug resistant cancer cells with "ersatzdroges".* International journal of cancer, 2015. **136**(4): p. E188-E196.
67. Huang, S., *Non-genetic heterogeneity of cells in development: more than just noise.* Development, 2009. **136**(23): p. 3853-3862.
68. Merlo, L.M., et al., *Cancer as an evolutionary and ecological process.* Nature reviews cancer, 2006. **6**(12): p. 924-935.
69. Gatenby, R.A. and T.L. Vincent, *An evolutionary model of carcinogenesis.* Cancer research, 2003. **63**(19): p. 6212-6220.
70. Martincorena, I., et al., *Universal patterns of selection in cancer and somatic tissues.* Cell, 2017. **171**(5): p. 1029-1041. e21.
71. Jolly, M.K., et al., *Phenotypic plasticity, bet-hedging, and androgen independence in prostate cancer: Role of non-genetic heterogeneity.* Frontiers in oncology, 2018. **8**: p. 50.
72. Foo, J. and F. Michor, *Evolution of acquired resistance to anti-cancer therapy.* Journal of theoretical biology, 2014. **355**: p. 10-20.
73. Turajlic, S., et al., *Resolving genetic heterogeneity in cancer.* Nature Reviews Genetics, 2019. **20**(7): p. 404-416.
74. Abécassis, J., et al., *Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data.* PloS one, 2019. **14**(11).
75. Jamal-Hanjani, M., et al., *Tracking the evolution of non–small-cell lung cancer.* New England Journal of Medicine, 2017. **376**(22): p. 2109-2121.
76. Turajlic, S., et al., *Tracking cancer evolution reveals constrained routes to metastases: TRACERx renal.* Cell, 2018. **173**(3): p. 581-594. e12.
77. Shu, Y., et al., *Circulating tumor DNA mutation profiling by targeted next generation sequencing provides guidance for personalized treatments in multiple cancer types.* Scientific reports, 2017. **7**(1): p. 1-11.
78. Abbosh, C., et al., *Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution.* Nature, 2017. **545**(7655): p. 446-451.
79. Savas, P., et al., *The subclonal architecture of metastatic breast cancer: results from a prospective community-based rapid autopsy program "CASCADE".* PLoS medicine, 2016. **13**(12).
80. Flavahan, W.A., E. Gaskell, and B.E. Bernstein, *Epigenetic plasticity and the hallmarks of cancer.* Science, 2017. **357**(6348): p. eaal2380.
81. Liau, B.B., et al., *Adaptive chromatin remodeling drives glioblastoma stem cell plasticity and drug tolerance.* Cell stem cell, 2017. **20**(2): p. 233-246. e7.
82. Cieslik, M. and A.M. Chinnaiyan, *Global genomics project unravels cancer's complexity at unprecedented scale*. 2020, Nature Publishing Group.
83. Xu, C., *A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data.* Computational and structural biotechnology journal, 2018. **16**: p. 15-24.
84. Zhang, L., et al., *Comprehensively benchmarking applications for detecting copy number variation.* PLoS computational biology, 2019. **15**(5): p. e1007069.
85. Doig, K.D., et al., *PathOS: a decision support system for reporting high throughput sequencing of cancers in clinical diagnostic laboratories.* Genome medicine, 2017. **9**(1): p. 38.

86. Williams, M.J., et al., *Identification of neutral tumor evolution across cancer types.* Nature genetics, 2016. **48**(3): p. 238.
87. Caravagna, G., et al., *Model-based tumor subclonal reconstruction.* BioRxiv, 2019: p. 586560.
88. Williams, M.J., et al., *Quantification of subclonal selection in cancer from bulk sequencing data.* Nature genetics, 2018. **50**(6): p. 895-903.
89. Beerenwinkel, N., et al., *Cancer evolution: mathematical models and computational inference.* Systematic biology, 2015. **64**(1): p. e1-e25.
90. Dentro, S.C., D.C. Wedge, and P. Van Loo, *Principles of reconstructing the subclonal architecture of cancers.* Cold Spring Harbor perspectives in medicine, 2017. **7**(8): p. a026625.
91. Roth, A., et al., *PyClone: statistical inference of clonal population structure in cancer.* Nature methods, 2014. **11**(4): p. 396.
92. Andor, N., et al., *EXPANDS: expanding ploidy and allele frequency on nested subpopulations.* Bioinformatics, 2014. **30**(1): p. 50-60.
93. Litchfield, K., et al., *Representative sequencing: unbiased sampling of solid tumor tissue.* Cell reports, 2020. **31**(5): p. 107550.
94. Bozic, I., C. Paterson, and B. Waclaw, *On measuring selection in cancer from subclonal mutation frequencies.* PLoS computational biology, 2019. **15**(9): p. e1007368.
95. Chkhaidze, K., et al., *Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data.* PLoS computational biology, 2019. **15**(7): p. e1007243.
96. Bielski, C.M., et al., *Genome doubling shapes the evolution and prognosis of advanced cancers.* Nature genetics, 2018. **50**(8): p. 1189-1195.
97. Ciriello, G., et al., *Emerging landscape of oncogenic signatures across human cancers.* Nature genetics, 2013. **45**(10): p. 1127-1133.
98. Aran, D., M. Sirota, and A.J. Butte, *Systematic pan-cancer analysis of tumour purity.* Nature communications, 2015. **6**: p. 8971.
99. Dewhurst, S.M., et al., *Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution.* Cancer discovery, 2014. **4**(2): p. 175-185.
100. Cross, W.C., T.A. Graham, and N.A. Wright, *New paradigms in clonal evolution: punctuated equilibrium in cancer.* The Journal of pathology, 2016. **240**(2): p. 126-136.
101. Schwartz, R. and A.A. Schäffer, *The evolution of tumour phylogenetics: principles and practice.* Nature Reviews Genetics, 2017. **18**(4): p. 213.
102. Miller, C.A., et al., *SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution.* PLoS computational biology, 2014. **10**(8).
103. Gillis, S. and A. Roth, *PyClone-VI: scalable inference of clonal population structures using whole genome data.* BMC bioinformatics, 2020. **21**(1): p. 1-16.
104. Ha, G., et al., *TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data.* Genome research, 2014. **24**(11): p. 1881-1893.
105. Oesper, L., A. Mahmoody, and B.J. Raphael, *THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data.* Genome biology, 2013. **14**(7): p. R80.
106. Yu, Z., A. Li, and M. Wang, *CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data.* BMC bioinformatics, 2016. **17**(1): p. 310.

107. Nik-Zainal, S., et al., *The life history of 21 breast cancers.* Cell, 2012. **149**(5): p. 994-1007.
108. Graham, T.A. and A. Sottoriva, *Measuring cancer evolution from the genome.* The Journal of pathology, 2017. **241**(2): p. 183-191.
109. Davis, A., R. Gao, and N. Navin, *Tumor evolution: Linear, branching, neutral or punctuated?* Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 2017. **1867**(2): p. 151-161.
110. Sun, R., Z. Hu, and C. Curtis, *Big Bang tumor growth and clonal evolution.* Cold Spring Harbor perspectives in medicine, 2018. **8**(5): p. a028381.
111. Ramazzotti, D., et al., *CAPRI: efficient inference of cancer progression models from cross-sectional data.* Bioinformatics, 2015. **31**(18): p. 3016-3026.
112. Loohuis, L.O., et al., *Inferring tree causal models of cancer progression with probability raising.* PloS one, 2014. **9**(10).
113. Dang, H., et al., *ClonEvol: clonal ordering and visualization in cancer sequencing.* Annals of oncology, 2017. **28**(12): p. 3076-3082.
114. Niknafs, N., et al., *SubClonal hierarchy inference from somatic mutations: automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing.* PLoS computational biology, 2015. **11**(10).
115. Strino, F., et al., *TrAp: a tree approach for fingerprinting subclonal tumor composition.* Nucleic acids research, 2013. **41**(17): p. e165-e165.
116. Flensburg, C., et al., *SuperFreq: Integrated mutation detection and clonal tracking in cancer.* PLoS computational biology, 2020. **16**(2): p. e1007603.
117. Jiao, W., et al., *Inferring clonal evolution of tumors from single nucleotide somatic mutations.* BMC bioinformatics, 2014. **15**(1): p. 35.
118. Deshwar, A.G., et al., *PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors.* Genome biology, 2015. **16**(1): p. 35.
119. Yuan, K., et al., *BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies.* Genome biology, 2015. **16**(1): p. 36.
120. Jiang, Y., et al., *Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing.* Proceedings of the National Academy of Sciences, 2016. **113**(37): p. E5528-E5537.
121. El-Kebir, M., et al., *Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures.* Cell systems, 2016. **3**(1): p. 43-53.
122. McFarland, C.D., L.A. Mirny, and K.S. Korolev, *Tug-of-war between driver and passenger mutations in cancer and other adaptive processes.* Proceedings of the National Academy of Sciences, 2014. **111**(42): p. 15138-15143.
123. Bozic, I. and M.A. Nowak, *Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers.* Proceedings of the National Academy of Sciences, 2014. **111**(45): p. 15964-15968.
124. Gatenby, R.A., et al., *Adaptive therapy.* Cancer research, 2009. **69**(11): p. 4894-4903.
125. Komarova, N.L. and D. Wodarz, *Drug resistance in cancer: principles of emergence and prevention.* Proceedings of the National Academy of Sciences, 2005. **102**(27): p. 9714-9719.
126. Altrock, P.M., L.L. Liu, and F. Michor, *The mathematics of cancer: integrating quantitative models.* Nature Reviews Cancer, 2015. **15**(12): p. 730-745.
127. McDonald, T.O. and F. Michor, *SIApopr: a computational method to simulate evolutionary branching trees for analysis of tumor clonal evolution.* Bioinformatics, 2017. **33**(14): p. 2221-2223.

128. Gatenbee, C.D., et al., *Niche engineering drives early passage through an immune bottleneck in progression to colorectal cancer.* bioRxiv, 2019: p. 623959.
129. Tan, W.-Y., W. Ke, and G. Webb, *A stochastic and state space model for tumour growth and applications.* Computational and Mathematical Methods in Medicine, 2009. **10**(2): p. 117-138.
130. McFarland, C.D., et al., *Impact of deleterious passenger mutations on cancer progression.* Proceedings of the National Academy of Sciences, 2013. **110**(8): p. 2910-2915.
131. S. Datta, R., et al., *Modelling the evolution of genetic instability during tumour progression.* Evolutionary applications, 2013. **6**(1): p. 20-33.
132. Athreya, K.B. and P. Jagers, *Classical and modern branching processes*. Vol. 84. 2012: Springer Science & Business Media.
133. Crump, K.S. and D.G. Hoel, *Mathematical models for estimating mutation rates in cell populations.* Biometrika, 1974. **61**(2): p. 237-252.
134. Haeno, H., Y. Iwasa, and F. Michor, *The evolution of two mutations during clonal expansion.* Genetics, 2007. **177**(4): p. 2209-2221.
135. Beerenwinkel, N., et al., *Genetic progression and the waiting time to cancer.* PLoS computational biology, 2007. **3**(11).
136. Durrett, R. and S. Moseley, *Evolution of resistance and progression to disease during clonal expansion of cancer.* Theoretical population biology, 2010. **77**(1): p. 42-48.
137. Bozic, I. and C.J. Wu, *Delineating the evolutionary dynamics of cancer from theory to reality.* Nature Cancer, 2020. **1**(6): p. 580-588.
138. Nagel, A.C., et al., *Stickbreaking: a novel fitness landscape model that harbors epistasis and is consistent with commonly observed patterns of adaptive evolution.* Genetics, 2012. **190**(2): p. 655-667.
139. Komarova, N.L., J.A. Burger, and D. Wodarz, *Evolution of ibrutinib resistance in chronic lymphocytic leukemia (CLL).* Proceedings of the National Academy of Sciences, 2014. **111**(38): p. 13906-13911.
140. Bozic, I., et al., *Evolutionary dynamics of cancer in response to targeted combination therapy.* elife, 2013. **2**: p. e00747.
141. Katouli, A.A. and N.L. Komarova, *Optimizing combination therapies with existing and future CML drugs.* PloS one, 2010. **5**(8): p. e12300.
142. Orr, H.A., *The distribution of fitness effects among beneficial mutations.* Genetics, 2003. **163**(4): p. 1519-1526.
143. Antal, T. and P. Krapivsky, *Exact solution of a two-type branching process: models of tumor progression.* Journal of Statistical Mechanics: Theory and Experiment, 2011. **2011**(08): p. P08018.
144. Antal, T. and P. Krapivsky, *Exact solution of a two-type branching process: clone size distribution in cell division kinetics.* Journal of Statistical Mechanics: Theory and Experiment, 2010. **2010**(07): p. P07028.
145. Cheek, D. and T. Antal, *Mutation frequencies in a birth–death branching process.* The Annals of Applied Probability, 2018. **28**(6): p. 3922-3947.
146. Diaz Jr, L.A., et al., *The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers.* Nature, 2012. **486**(7404): p. 537-540.
147. Reiter, J.G., et al., *The effect of one additional driver mutation on tumor progression.* Evolutionary Applications, 2013. **6**(1): p. 34-45.
148. Haeno, H., et al., *Computational modeling of pancreatic cancer reveals kinetics of metastasis suggesting optimum treatment strategies.* Cell, 2012. **148**(1-2): p. 362-375.
149. Reiter, J.G., et al., *Reconstructing metastatic seeding patterns of human cancers.* Nature communications, 2017. **8**(1): p. 1-10.

150. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2019.* CA: a cancer journal for clinicians, 2019. **69**(1): p. 7-34.
151. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2020.* CA: A Cancer Journal for Clinicians, 2020. **70**(1): p. 7-30.
152. Brown, P.O. and C. Palmer, *The preclinical natural history of serous ovarian cancer: defining the target for early detection.* PLoS medicine, 2009. **6**(7).
153. Gruber, M., et al., *Growth dynamics in naturally progressing chronic lymphocytic leukaemia.* Nature, 2019. **570**(7762): p. 474-479.
154. Baker, A.-M., et al., *Quantification of crypt and stem cell evolution in the normal and neoplastic human colon.* Cell reports, 2014. **8**(4): p. 940-947.
155. Karthik, G.-M., et al., *Intra-tumor heterogeneity in breast cancer has limited impact on transcriptomic-based molecular profiling.* BMC cancer, 2017. **17**(1): p. 1-11.
156. Lange, K., *Branching Processes*, in *Applied Probability*. 2010, Springer. p. 217-245.
157. Ostrow, S.L., et al., *Cancer evolution is associated with pervasive positive selection on globally expressed genes.* PLoS Genet, 2014. **10**(3): p. e1004239.
158. Haccou, P., et al., *Branching processes: variation, growth, and extinction of populations*. 2005: Cambridge university press.
159. Athreya, K.B., *Branching process.* Encyclopedia of Environmetrics, 2006. **1**.
160. Del Monte, U., *Does the cell number 109 still really fit one gram of tumor tissue?* Cell cycle, 2009. **8**(3): p. 505-506.
161. Devita Jr, V.T., R.C. Young, and G.P. Canellos, *Combination versus single agent chemotherapy: a review of the basis for selection of drug treatment of cancer.* Cancer, 1975. **35**(1): p. 98-110.
162. Edge, S.B. and C.C. Compton, *The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM.* Annals of surgical oncology, 2010. **17**(6): p. 1471-1474.
163. Erdi, Y.E., *Limits of tumor detectability in nuclear medicine and PET.* Molecular imaging and radionuclide therapy, 2012. **21**(1): p. 23.
164. Besusparis, J., et al., *Impact of tissue sampling on accuracy of Ki67 immunohistochemistry evaluation in breast cancer.* Diagnostic pathology, 2016. **11**(1): p. 1-10.
165. Durrett, R., et al., *Intratumor heterogeneity in evolutionary models of tumor progression.* Genetics, 2011. **188**(2): p. 461-477.
166. Thakur, S.S., et al., *The use of automated Ki67 analysis to predict Oncotype DX risk-of-recurrence categories in early-stage breast cancer.* PLoS One, 2018. **13**(1): p. e0188983.
167. McDonald, T.O. and M. Kimmel, *A multitype infinite-allele branching process with applications to cancer evolution.* Journal of Applied Probability, 2015. **52**(3): p. 864-876.
168. Chan, M.F., et al., *Reduced rates of gene loss, gene silencing, and gene mutation in Dnmt1-deficient embryonic stem cells.* Molecular and cellular biology, 2001. **21**(22): p. 7587-7600.
169. Sottoriva, A., et al., *Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization.* Cancer research, 2013. **73**(1): p. 41-49.
170. Sgroi, D.C., et al., *Prediction of late distant recurrence in patients with oestrogen-receptor-positive breast cancer: a prospective comparison of the breast-cancer index (BCI) assay, 21-gene recurrence score, and IHC4 in the TransATAC study population.* The lancet oncology, 2013. **14**(11): p. 1067-1076.

171. Cuzick, J., et al., *Effect of anastrozole and tamoxifen as adjuvant treatment for early-stage breast cancer: 10-year analysis of the ATAC trial.* The lancet oncology, 2010. **11**(12): p. 1135-1141.
172. Regan, M.M., et al., *Assessment of letrozole and tamoxifen alone and in sequence for postmenopausal women with steroid hormone receptor-positive breast cancer: the BIG 1-98 randomised clinical trial at 8· 1 years median follow-up.* The lancet oncology, 2011. **12**(12): p. 1101-1108.
173. Noble, R., et al., *When, why and how tumour clonal diversity predicts survival.* Evolutionary applications, 2020. **13**(7): p. 1558-1568.
174. Tlsty, T.D., B.H. Margolin, and K. Lum, *Differences in the rates of gene amplification in nontumorigenic and tumorigenic cell lines as measured by Luria-Delbrück fluctuation analysis.* Proceedings of the National Academy of Sciences, 1989. **86**(23): p. 9441-9445.
175. Araten, D.J., et al., *A quantitative measurement of the human somatic mutation rate.* Cancer research, 2005. **65**(18): p. 8111-8117.
176. Tomasetti, C., B. Vogelstein, and G. Parmigiani, *Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation.* Proceedings of the National Academy of Sciences, 2013. **110**(6): p. 1999-2004.
177. Noorbakhsh, J., et al., *Distribution-based measures of tumor heterogeneity are sensitive to mutation calling and lack strong clinical predictive power.* Scientific reports, 2018. **8**(1): p. 1-12.
178. Escalona, M., S. Rocha, and D. Posada, *A comparison of tools for the simulation of genomic next-generation sequencing data.* Nature Reviews Genetics, 2016. **17**(8): p. 459.
179. Chen, Z., et al., *Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency.* Scientific reports, 2020. **10**(1): p. 1-9.
180. Salcedo, A., et al., *A community effort to create standards for evaluating tumor subclonal reconstruction.* Nature biotechnology, 2020. **38**(1): p. 97-107.
181. Gerstung, M., et al., *The evolutionary history of 2,658 cancers.* Nature, 2020. **578**(7793): p. 122-128.
182. Goldie, J., *The somatic mutation theory of drug resistance: The "Goldie-Coldman" hypothesis revisited.* Principles and Practice of Oncology, PPO Updates, 1989. **3**: p. 1-12.
183. Komarova, N., *Stochastic modeling of drug resistance in cancer.* Journal of theoretical biology, 2006. **239**(3): p. 351-366.
184. Davies, C., et al., *Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial.* The Lancet, 2013. **381**(9869): p. 805-816.
185. Gray, R.G., et al., *aTTom: Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years in 6,953 women with early breast cancer*. 2013, American Society of Clinical Oncology.
186. Yu, K.-D., et al. *Effect of large tumor size on cancer-specific mortality in node-negative breast cancer*. in *Mayo Clinic Proceedings*. 2012. Elsevier.
187. Avanzini, S. and T. Antal, *Cancer recurrence times from a branching process model.* PLoS computational biology, 2019. **15**(11).
188. Kumaran, M., et al., *Germline copy number variations are associated with breast cancer risk and prognosis.* Scientific reports, 2017. **7**(1): p. 1-15.

189. Zardavas, D., et al., *Tumor PIK3CA genotype and prognosis in early-stage breast cancer: a pooled analysis of individual patient data.* Journal of clinical oncology, 2018. **36**(10): p. 981-+.
190. Wu, H.-X., et al., *Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers.* Annals of translational medicine, 2019. **7**(22).
191. Snyder, A., et al., *Genetic basis for clinical response to CTLA-4 blockade in melanoma.* New England Journal of Medicine, 2014. **371**(23): p. 2189-2199.
192. Rizvi, N.A., et al., *Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer.* Science, 2015. **348**(6230): p. 124-128.
193. Marabelle, A., et al., *Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study.* The Lancet Oncology, 2020. **21**(10): p. 1353-1365.
194. Guiasu, R.C. and S. Guiasu, *The Rich-Gini-Simpson quadratic index of biodiversity.* Natural Science, 2010. **2**(10): p. 1130.
195. Group, B.I.G.-C., *A comparison of letrozole and tamoxifen in postmenopausal women with early breast cancer.* New England Journal of Medicine, 2005. **353**(26): p. 2747-2757.
196. Endesfelder, D., et al., *Chromosomal instability selects gene copy-number variants encoding core regulators of proliferation in ER+ breast cancer.* Cancer research, 2014. **74**(17): p. 4853-4863.
197. Lai, Z., et al., *VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research.* Nucleic acids research, 2016. **44**(11): p. e108-e108.
198. Frampton, G.M., et al., *Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing.* Nature biotechnology, 2013. **31**(11): p. 1023.
199. Luen, S.J., et al., *Association of somatic driver alterations with prognosis in postmenopausal, hormone receptor–positive, HER2-negative early breast cancer: a secondary analysis of the BIG 1-98 randomized clinical trial.* JAMA oncology, 2018. **4**(10): p. 1335-1343.
200. Kuilman, T., et al., *CopywriteR: DNA copy number detection from off-target sequence data.* Genome biology, 2015. **16**(1): p. 49.
201. Nilsen, G., et al., *Copynumber: efficient algorithms for single-and multi-track copy number segmentation.* BMC genomics, 2012. **13**(1): p. 591.
202. Nilsen, G., K. Liestol, and O. Lingjaerde, *Copynumber: Segmentation of single-and multi-track copy number data by penalized least squares regression.* R package. version, 2013. **1200**.
203. Vergara, I.A., et al., *Evolution of late-stage metastatic melanoma is dominated by aneuploidy and whole genome doubling.* Nature communications, 2021. **12**(1): p. 1-15.
204. Anderson, T.W., *On the distribution of the two-sample Cramer-von Mises criterion.* The Annals of Mathematical Statistics, 1962: p. 1148-1159.
205. Xiao, Y., A. Gordon, and A. Yakovlev, *The-Version of the Cramér-von Mises Test for Two-Sample Comparisons in Microarray Data Analysis.* EURASIP Journal on Bioinformatics and Systems Biology, 2006. **2006**(1): p. 85769.
206. Razali, N.M. and Y.B. Wah, *Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests.* Journal of statistical modeling and analytics, 2011. **2**(1): p. 21-33.

207. Mukherjee, A., et al., *Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort.* NPJ breast cancer, 2018. **4**(1): p. 1-9.
208. Birkeland, E., et al., *Patterns of genomic evolution in advanced melanoma.* Nature communications, 2018. **9**(1): p. 1-12.
209. Goto, T., et al., *Understanding intratumor heterogeneity and evolution in nsclc and potential new therapeutic approach.* Cancers, 2018. **10**(7): p. 212.
210. Ni, X., et al., *Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients.* Proceedings of the National Academy of Sciences, 2013. **110**(52): p. 21083-21088.
211. Nieboer, M.M., et al., *TargetClone: A multi-sample approach for reconstructing subclonal evolution of tumors.* PloS one, 2018. **13**(11).