UNIVERSITÉ LIBRE DE BRUXELLES
Faculty of Sciences
Interuniversity Institute of Bioinformatics in Brussels

Modeling Tumor Heterogeneity

Milan MALFAIT

Master thesis submitted as partial fulfillment to obtain the degree of Master in Bioinformatics and Modeling

Promotor: T. Lenaerts
Supervisors: N. Mon Père, S. Vande Velde

Academic year 2017 - 2018

Acknowledgements

This thesis is the culmination of a year of work with the contribution of many people to whom I owe my sincere gratitude. First and foremost, I would like to thank my promotor, Prof. Tom Lenaerts, for giving me the opportunity to perform my master thesis at the $(IB)^2$ (Interuniversity Institute for Bioinformatics in Brussels) and for his guidance and advice throughout the year.

My deepest gratitude also goes out to my supervisors Nathaniel and Sylvie, for their guidance and support and always being their to help me whenever I needed it. I also would like to thank everyone else at $(IB)^2$ for keeping up the positive vibe and making it an enjoyable work environment.

Thank you to all my friends and fellow students for the good times and moral support and making the past few years at university such an amazing time!

To my parents, thank you for all the opportunities and support that you have given me, for inspiring me and pushing me forward. Finally, my thesis would not have been completed without the support of my girlfriend, thank you for believing in me and always being there to keep me motivated.

Contents

Chapter 1

Introduction

Despite significant advances in healthcare and medicine, cancer remains a major cause of death in modern society. The highly variable nature of this disease type makes developing efficient therapies particularly difficult. Approximately 95% of clinical trials fail between phase I and phase III [1]. It is now being recognized that tumor heterogeneity has a severe impact on the efficacy of drugs in humans and that its inadequate recognition by existing clinical models is a cause of the high failure rates [2]. Therefore, a deeper understanding of the evolutionary dynamics that underlie tumor progression are essential for the future development of new anti-cancer strategies.

According to the dominant view, the genetic diversity of a tumor is mainly driven by Darwinian evolution, in which positive selection acts upon those mutations that confer a fitness advantage for the cell [3, 4, 5]. The genotype of the resulting tumor will be dominated by those mutations that confer the highest fitness and the heterogeneity within the tumor will be relatively low. However, this hypothesis has recently been called into question, arguing that the observed intra-tumor heterogeneity (ITH) of some tumors is too high to be a consequence of only a Darwinian evolutionary process [6]. Instead, it seems that non-Darwinian evolution can also contribute significantly to the genetic diversity within a tumor. In this process – known as neutral evolution – mutations accumulate randomly and the evolutionary dynamics are purely stochastic, leading to a higher degree of ITH than would be expected from evolution under natural selection [7]. This view was supported by extensive sampling and bulk sequencing of various clinical tumors, which showed degrees of heterogeneity consistent with what is predicted by mathematical models describing neutral evolution [8, 9]. This has lead to an ongoing debate in cancer research about which evolutionary force dominates tumor development most [6].

Existing models from classic population genetics provide a suitable base for cancer evolution models to be built upon. One such model that has been successfully applied to describe the

dynamics of cancer cell populations is the Wright-Fisher model. Originally used to simulate genetic drift of a population of constant size, it can be adapted to predict the fate of individual cells within a tumor population undergoing random mutations [10]. In recent years, this strategy was used to model the accumulation of driver mutations and the waiting time to cancer [11], simulation of the mutator phenotype [12], modeling the evolutionary dynamics of tumor heterogeneity based on fitness advantages and positive selection (Darwinian evolution) [13] and finally, demonstrating the role of neutral (non-Darwinian) evolution [8].

To further elucidate the origins of genetic heterogeneity in tumor cell populations, a new model is proposed here. For this purpose, the model is developed so that it is able to simulate a growing population of cells that evolves from a homogeneous origin to a heterogeneous state in a way that is similar to biological reality. It is known that genetic heterogeneity arises from both the inherent stochasticity of growth (i.e. neutral evolution) and selective pressure. Therefore, the model takes both effects into account. The Wright-Fisher model is an appropriate choice for this objective, as it already incorporates stochasticity in population evolution and can be adapted to include natural selection [10]. In addition, it has the major advantage of not having to simulate each cell division separately, which makes it a powerful method to simulate large populations. However, a drawback of the Wright-Fisher model is that it assumes non-overlapping generations, whereby the previous generation dies out when a new one arises. This is not compatible with cancer, because cancer cells acquire a state approaching immortality, preventing them from dying normally [14]. The proposed model overcomes this issue by simply discarding the assumption of non-overlapping generations, resulting in a growing population subjected to stochasticity and selection. However, this poses a new problem, because now no cell death can occur at all. Because this phenomenon is known to often play an important role in evolutionary dynamics, cell death is introduced in the model as a separate mechanism acting in parallel to cell division, thereby retaining its influence on the evolution of the population. The novelty of the model lies in the fact that it uses the Wright-Fisher model to simulate the growth of genetically identical subpopulations within the tumor (*clones*), instead of considering each individual cell separately. This way, a general 'mean-field' model is obtained that allows efficient computation, while still enabling the study of different aspects of tumor evolution.

In brief, the constructed model is used to simulate tumor populations under different conditions, followed by an assessment of the general behavior of the model and the influence of the relevant parameters. The resulting genetic diversity is then compared between different levels of selective pressure and to existing models from the literature in an attempt to elucidate the relative contributions of stochasticity and selection to the evolutionary dynamics of cancer progression.

Chapter 2

State of the Art

2.1 The biology of cancer

The genomic complexity that serves as the blueprint of metazoan organisms has allowed them to evolve into a plethora of multicellular structures with divergent organizations. Individual cells that serve as the building blocks of these life forms each carry an identical copy of the complete genetic information. Furthermore, they can differentiate into specialized types that carry out many different tasks. To achieve this versatility and to keep cells in their appointed roles, sophisticated systems that regulate gene expression are required. In particular, cell division and the associated cell cycle are under strict control to ensure that the right cells divide at the right time, while other cells are prevented from proliferating when they are not needed to do so. These differences in cell cycles are regulated through a molecular network of genes and proteins and are essential to processes such as embryonic development, immune response and wound healing, which require active proliferation of cells. On the other hand, cells in tissues such as the nervous system are kept in a non-dividing state to safeguard their correct functioning [15]. However, defects can arise in these regulatory networks, and because all cells essentially have access to the complete genome, including the instructions for growth, this can have dire consequences. When a cell escapes the usual cell cycle control mechanisms, it can enter a state of unlimited proliferation, eventually leading to a large cell mass that is recognized as a tumor [14]. Some tumors can become malignant and acquire the ability to spread out and invade other part of the body, resulting in what is known as a metastatic cancer.

2.1.1 Tumor development is driven by genetic alterations

2.1.1.1 The ten hallmarks that underlie all forms of cancer

Cancer does not consist of one disease but rather forms a collection of very divergent pathological disorders that nevertheless share some common characteristics. To prevent cells from entering a malignant state, the immune system has evolved a range of defense mechanisms that actively monitor and contain any cells that show abnormal behavior [16]. Thus, in order for a cell to become cancerous, it has to acquire the abilities that allow it to escape these immune barriers. In their highly cited review article from 2000, D. Hanahan and R. A. Weinberg argue that all forms of malignant growth are governed by six 'hallmarks': (i) the ability to stimulate their own growth, (ii) insensitivity to anti-growth signals, (iii) evasion of programmed cell death, (iv) the ability to multiply indefinitely, (v) stimulation of blood-vessel formation (angiogenesis) to provide the tumor with nutrients, and (vi) the invasion of tissue and ability to spread throughout the body (metastasis) [17]. In their 2011 update, they defined four additional hallmarks: (vii) the reprogramming of metabolic pathways, (viii) avoidance of destruction by the immune system, (ix) increase of genomic instability, and (x) tumor-promoting inflammation; the latter two being described as "enabling characteristics" that promote the acquirement of the other hallmarks [18].

Tumor development can be understood as a multistep process in which cells have to accumulate most, if not all, of these hallmark properties over time. The age incidence of cancer is directly related to this. The risk of cancer increases significantly with older age, suggesting that most tumors need many years to develop [14]. The progression towards cancer occurs through somatic mutations in key genes, providing it with the necessary hallmarks to become malignant. Essentially, the single progenitor cell that lies at the origin of a tumor will have accumulated multiple of these oncogenic mutations over the lifetime of a patient before it becomes cancerous [19].

2.1.1.2 Cancer genes have the ability to transform healthy cells

It has long been recognized that tumors consist of cells that show unlimited proliferative potential. The regulatory mechanisms of their cell cycle are disrupted by genetic alterations, leading to abnormal growth [20]. The target genes of these modifications are usually components of the cell cycle regulation and are collectively known as 'cancer genes'. These can be categorized into two groups according to their normal function. Oncogenes are growth-stimulating genes that have become over-activated. The normal gene (known as a proto-oncogene) can be converted in three main ways. The first is translocation due to chromosomal rearrangement, which can

cause a gene to become located near an active promoter, leading to higher expression. The second, amplification, is the creation of multiple copies of a gene by a series of duplications, each of which can act as a source of the gene product. Finally, point mutations in the promoter region or the coding sequence of a growth-stimulating gene can either increase its expression or lead to hyperactivity of the encoded protein, respectively [15]. Tumor-suppressor genes on the other hand, normally inhibit cell division and prevent uncontrolled growth. Deactivation of these subsequently leads to the loss of a barrier between the healthy state of a cell and the cancer phenotype.

Two archetypal examples of a proto-oncogene and a tumor-suppressor gene are *ras* and *p53* respectively. The *ras* gene encodes a protein involved in a signaling cascade that stimulates cell division. Normally, a growth factor is required to initiate the signal, but certain point mutations in the coding sequence can convert the Ras protein to a hyperactive form that is able to maintain the signal, even in absence of the growth factor [21, 22]. The product of *p53* is a transcription factor that induces the synthesis of a series of proteins involved in cell apoptosis. Its expression is activated as a response to DNA damage that could otherwise potentially lead to tumor formation. Mutations that cause a diminished or complete loss of *p53* activity enable a cell to escape programmed cell death, specifically apoptosis [23].

Cancer gene mutations are usually classified as *driver* mutations, while mutations that do not contribute directly to cancer development are known as *passengers* [4]. Drivers like *ras* and *p53* are well known because they appear across many types of cancer. However, many cancer genes are restricted to specific cancers and their total number remains elusive. Current estimates are that more than 1% of all human genes can potentially contribute to cancer [24]. In mouse models, more than 2000 cancer genes have been identified [25]. The number of driver mutations required for tumor initiation is highly variable between different forms of cancer and ranges from as few as 3 up to 20, according to mathematical models of rate-limiting events based on age-incidence and somatic mutation data [26, 11]. These predictions are supported by biological experiments that determined the driver mutations which lead to tumor formation for numerous cancers [27, 28, 14]. The rate at which mutations occur in cancer cells varies greatly between different forms, although common estimates are in the range of 10^{-3} – 10^{-2} for the functional part of the genome per cell division [29, 30].

2.1.2 The clonal architecture of cancer

When a cell has acquired a sufficient number of driver mutations, its unlimited proliferation will lead to a vast amount of offspring, all of which will inherit the cancer-inducing mutations of the progenitor [3, 31]. This group of genetically identical cells is then referred to as a

clonal population. The driver mutations of the ancestral cell are termed 'clonal' as they are shared by all offspring. During further tumor development, more driver mutations can give rise to new clonal subpopulations. These will generally consist of a small fraction of the tumor population and are referred to as *subclones*, subpopulations that are in themselves clonal. A tumor usually consists of multiple clonal populations and each of them can in turn give rise to still other subclones when new mutations occur (Figure 2.1). Each subclone forms a subpopulation of identical cells that will compete for space and resources, further driving the evolution of the tumor [32, 33]. It is this architecture of clones and subclones that leads to the high genetic variability observed both within the same tumor (intra-tumor heterogeneity) and between different tumors (inter-tumor heterogeneity).

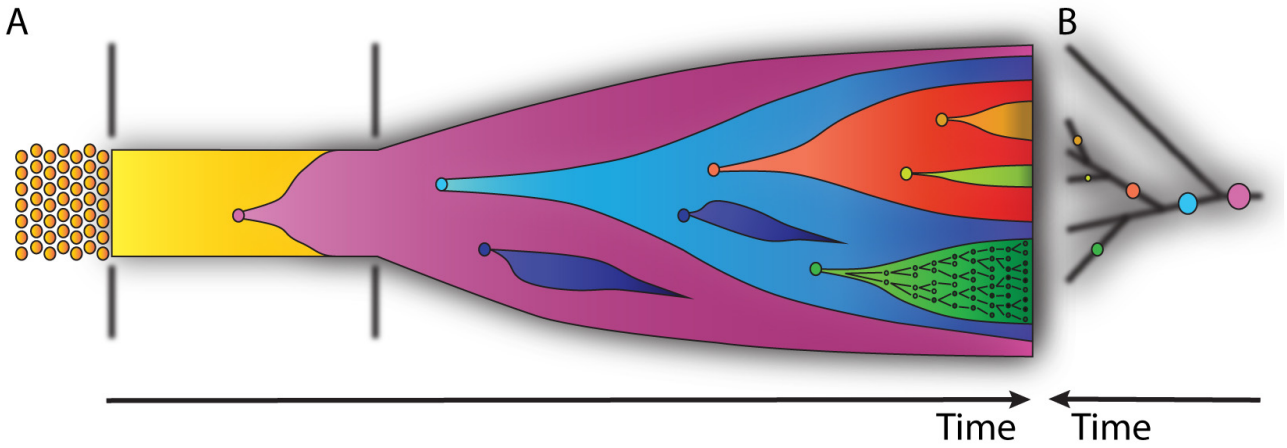

Figure 2.1. The clonal structure of cancer. (A) A cell in healthy tissue in homeostasis (in yellow) can undergo a driver mutation, leading to excessive proliferation of the selectively advantageous clonal progeny (in purple). During growth of the ancestral clone, new mutations can give rise to subclonal populations, which could potentially grow to dominate the population (light-blue), die out (dark blue) or remain subclonal (other colors). A more detailed view of the branching evolution of cells is given in the green subclone. (B) The phylogenetic tree corresponding with the population shown in A, depicting the clonal architecture of the tumor. Figure reprinted from Beerenwinkel *et al.* [34].

Passenger mutations do not contribute directly to the cancerous properties of its carrier cell. They occur randomly and are carried along with driver mutations during expansion of the clone, a concept that is known in evolution biology as 'hitchhiker' mutations. They do, however, contribute to tumor heterogeneity by increasing the genetic diversity between distinct subclones [4]. More importantly, it has been shown that some passenger mutations, present in only a small subset of the tumor, can be converted to driver mutations when the environment is altered [35, 36]. Therapeutics can fundamentally change the cancer ecosystem. Many tumor cells may be exterminated but pre-existing variants carrying passenger mutations that confer resistance

to the therapy now have the potential to thrive [37]. Furthermore, these cells can subsequently dominate the tumor, passing down their resistant genotype and lead to relapse of the cancer with a tumor that is now more adapted, with the possibility of having a higher malignancy, due to increased resistance [31]. This is an example of how the heterogeneous nature of a tumor can have severe clinical consequences.

In addition to the driver and passenger mutations, a third class of mutations can be defined as those that are able to increase the mutation rate of a cancer cell, known as 'mutator' lesions [38, 39]. These mutations increase the genetic instability of a clonal population and thereby accelerate the evolutionary dynamics. Cells that have accumulated such mutations will then acquire what is known as a *mutator phenotype*.

2.1.3 Is cancer evolution governed by a Darwinian model?

Darwin's theory of evolution states that populations evolve by natural selection of traits that confer a fitness advantage. In essence, fitness can be defined as the relative reproductive ability of a given genotype in its current environment. The source for selection to act on is provided by the heritable variability between individuals. Cancer can be regarded as such a system. Variation between tumor cells is created by the genetic alterations, whereby driver mutations provide cells with increased fitness and chance of survival. Successful clones outgrow others and come to dominate the population in a series of 'selective sweeps' (Figure 2.2A) [31]. However, due to the high mutation rates and large population sizes associated with tumors, driver mutations and related fitness advantages are usually distributed among different clones, leading to inter-clonal competition [40, 33]. This could explain the clonal diversity still present in mature tumors instead of a single dominant clone, although the extremely high heterogeneity observed in some tumors remains puzzling [41, 42, 43].

Modern views of evolution recognize another force that acts in parallel to Darwin's natural selection. Genetic drift is a non-Darwinian process that alters a population's gene pool by random sampling [44]. This neutral evolution is purely stochastic and as such, the genetic diversity that follows is generally larger (Figure 2.2B) [45]. Both mechanisms act conjointly in driving evolution. However, the relative contribution of the neutral stochastic effects have generally been underestimated. Accordingly, the prevailing view that tumor progression is governed exclusively by Darwinian evolution has also been called into question recently. One study using extensive tumor sampling argues that the level of observed intra-tumor heterogeneity (ITH) can only be explained by a neutral, non-Darwinian evolutionary process [8]. Another study shows that the frequency of subclonal mutations is more consistent with neutral dynamics [9]. Sottoriva *et al.* propose a more generalized model in which a tumor grows from a single clonal

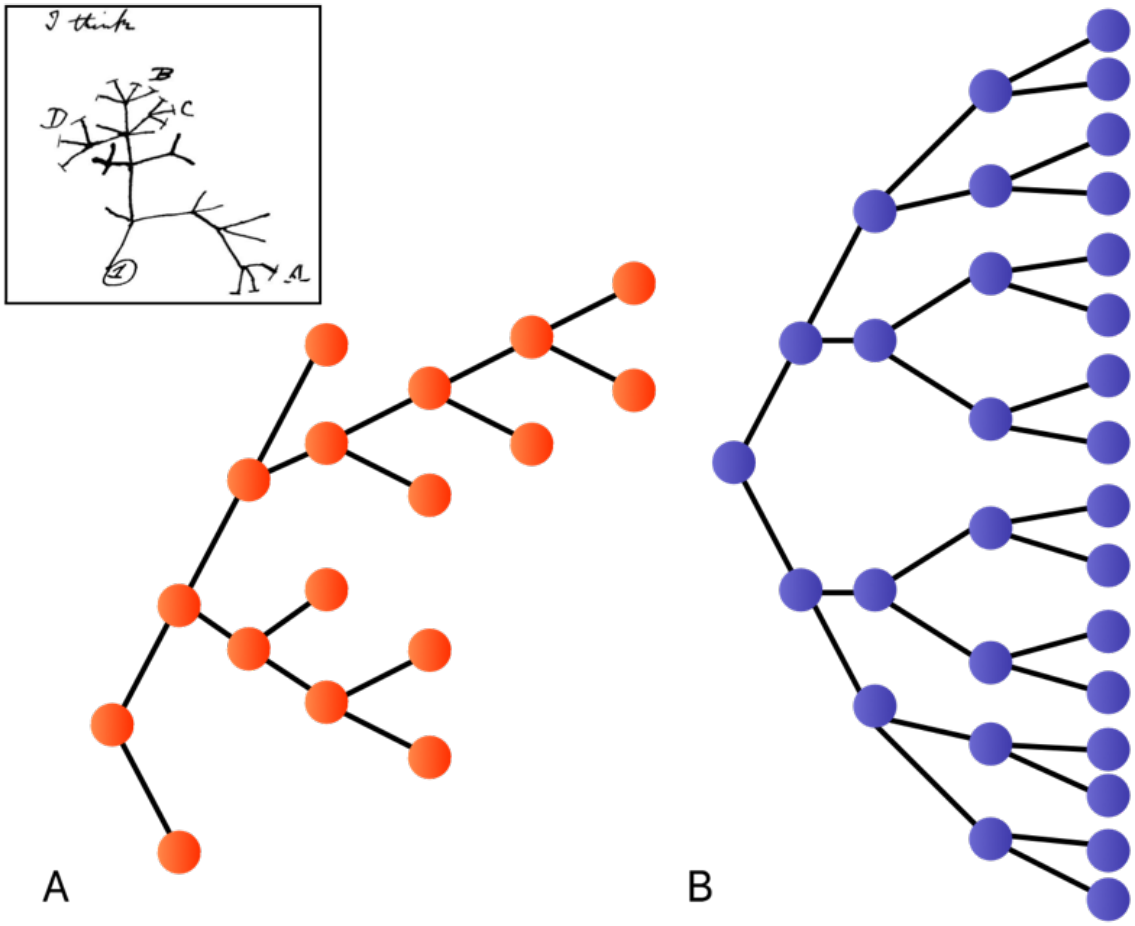

Figure 2.2. Different modes of evolution. (A) In Darwin's theory of evolution by natural selection, individuals that have higher fitness will have a higher probability of producing offspring (long branch), while others will go extinct. In the inset is a drawing from Darwin's notebook where he sketched his idea of evolution. (B) In a neutrally evolving population, no selective pressure is present and individuals thus reproduce at equal rate.

expansion that was selected early in development, followed by neutral growth that gives rise to a vast array of subclones that remain pervasive in the tumor [7]. Similar to this "Big Bang" model, the theory of cancer stem cells (CSCs) suggests that neutrally evolving subclones are sustained by a stem cell-like hierarchy [46, 47]. Indeed, it was found that the majority of cancer cells actually have limited proliferative potential, while a small subset consists of CSCs that are capable of extensive self-renewal and differentiation [48]. The debate about the contribution between selective and neutral evolution remains however controversial and requires more extensive research.

Much of the progression of a tumor's evolutionary dynamics depends on what happens early after transformation of the initial cancer cell. However, tumor initiation proves to be very impractical to observe directly, and the events that lie at the origin of the growth history remain largely unknown. At the same time, these first cell divisions and mutations dictate how

the tumor will progress. The eventual genomic profile and ITH highly depend on which clonal and subclonal mutations developed early on [49, 50]. Therefore, these events could provide important clues as to what the best strategies are to diagnose, prognose and ultimately treat cancer. Where experimental techniques become inadequate to measure early tumor initiation, predictive theoretical frameworks and the power of mathematical modeling could bring solace.

2.2 The mathematics of cancer

Mathematical modeling already has a long history of importance in cancer biology. One of the first successful applications was the study of cancer incidence by Nordling in the 1950's, which resulted in the postulation of the multi-stage theory of cancer and the subsequent discovery that cancer is a result of accumulating somatic mutations [51, 34]. Since then, the applications of mathematical theories on cancer evolution have expanded tremendously, in part due to the advent of next-generation sequencing and related techniques. The amount of data and accuracy generated by these new technologies have provided a wealth of information on which computational and statistical methods can be used to elucidate new aspects about cancer biology and quantitatively test existing theories [52]. Conversely, the development of new mathematical predictions is also essential to improve prevailing experiments and suggest new ones [53]. As with other evolutionary studies, the theoretical framework provided by mathematical modeling is a key step towards the full comprehension of cancer evolution.

2.2.1 Applications of mathematical modeling in cancer biology: an overview

An array of mathematical models can be applied in order to grasp the complex dynamics underlying cancer development. By definition, a model is a simplification of the real system to keep it mathematically tractable, while simultaneously being able to accurately predict experimental and clinical observations. Deterministic models can be applied to study the hierarchical organization of cancer tissue and its influence on therapeutic intervention, for example [54, 55]. However, in order to capture the variability and fluctuations in the growth and composition of tumors, stochastic models are usually more suitable to simulate the growth of a population and individual cell fates [56].

Branching processes are powerful methods that incorporate stochasticity and are frequently used to simulate the growth of populations in biology, including cancer [57]. In particular, models from population genetics provide appropriate frameworks to study aspects of tumor evolution such as initiation, progression and development of resistance against treatments [34].

A more detailed description of these methods is given in the following section.

A last set of models are those that study the microenvironment of a tumor, which plays an important role in its expansion [14]. Typically, these strategies discern between a discrete stochastic description of cell growth and continuous deterministic dynamics of extracellular factors [53]. Sophisticated hybrid models attempt to combine these two concepts and have been applied, among others, to study the formation of blood vessels during tumor angiogenesis [58] and the metastasis of bone cancers [59].

It is clear that the use of mathematical models is essential for a deeper understanding of the quantitative aspects of cancer. Particularly, because modern experimental techniques do not yet allow direct observation of the processes that forego the development of a tumor to a detectable malignancy, stochastic models based on branching processes can provide significant insights.

2.2.2 Modeling cancer evolution as a stochastic process

2.2.2.1 Branching processes and population genetics models

The power of branching processes lies within the assumption that the characteristic rates of each cell – birth, mutation and death – are independent of its surroundings, meaning a cell will behave identically when it is alone or part of a larger population [60]. This allows a considerable amount of simplification of the system, making it mathematically tractable. A branching process is Markovian in that each individual produces a random number of offspring (or dies) at a given timepoint, independent of previous events. Birth, mutation and death events occur stochastically with cell-specific rates and each mutation can give rise to a new cell type with possibly different rates [53]. This leads to a branched evolution of the population whereby mutations accumulate and new cell types arise as the tumor progresses (Figure 2.3A). Selective pressure can be implemented in this kind of model, for example, by decreasing the death rate with each driver mutation (or conversely, increasing the growth rate) as was done by Bozic *et al.* to predict the number of passenger and driver mutations that have accumulated in a tumor [61]. In addition, this study demonstrated that the selective advantage of driver mutations was surprisingly low, with only a 0.4% increase in growth rate.

Population genetics models such as the Moran and Wright-Fisher processes are based on branching principles and have been applied extensively to model populations in evolutionary biology [10]. Originally, these methods were constructed to describe populations of fixed size undergoing genetic drift [62]. The Moran model considers each cell individually and during each timestep one cell is randomly chosen to divide and possibly mutate, while another is chosen to die, keeping the total population size constant (Figure 2.3B). The probability of a cell to be

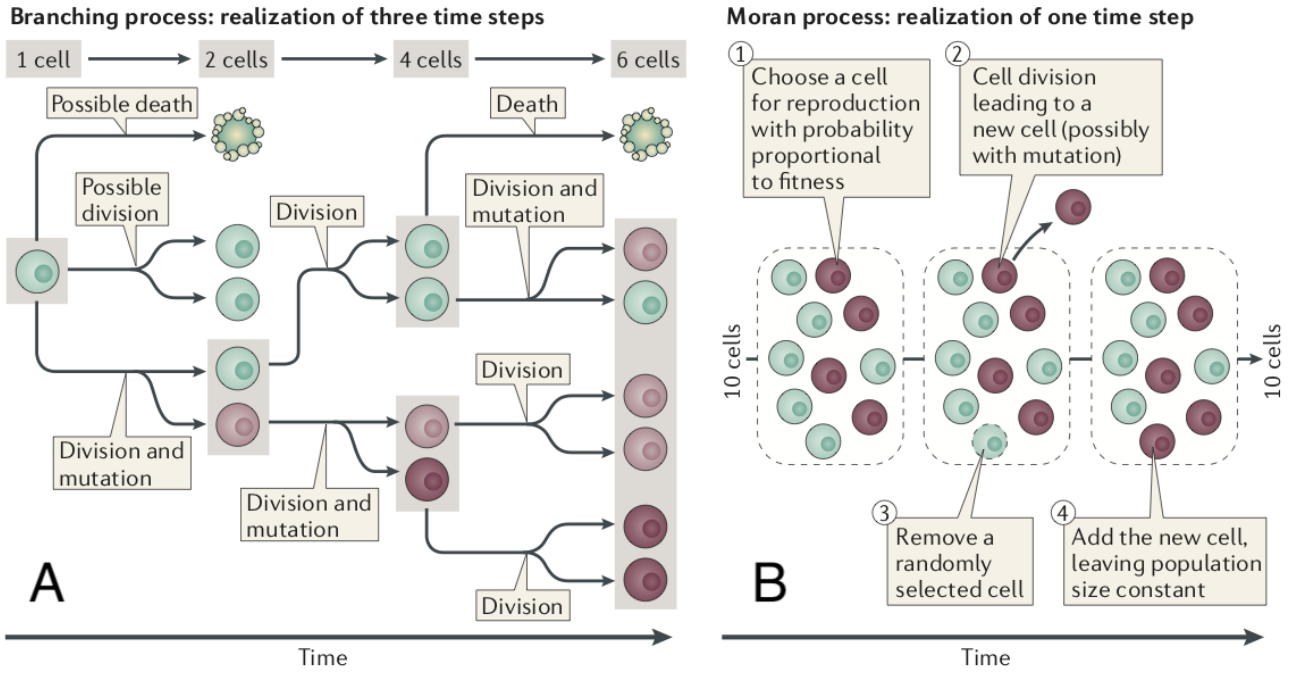

Figure 2.3. Stochastic models for cancer evolution. (A) In branching processes, each cell at a given timepoint can either divide, divide and mutate or die, each with a specific rate. Mutations give rise to new cell types that can have altered growth, death and mutation rates. (B) The Moran process describes a population of fixed size whereby at each timepoint a cell is selected to divide with probability proportional to its fitness, while another cell is chosen to die. The process is then repeated for each individual cell. Random mutations and increase in fitness advantage can also be incorporated in the model. Figure reprinted from Altrock *et al.* [53].

chosen to reproduce is proportionate to its fitness. In simplistic models, the increase in fitness is constant for each mutation. However, another approach considered a probability distribution of fitness levels from which a value is sampled for each mutation [63]. This method more realistically represents the fact that not all mutations confer the same fitness increase and might even be deleterious. Coupling this random distribution of fitness to a Moran process allowed the authors to simulate mutation accumulation and estimate the rate at which a cell initiates clonal expansion.

Because all individuals are treated separately, the Moran process can be seen as being *continuous* and its generations are *overlapping*. In contrast, the Wright-Fisher model assumes *non-overlapping* generations and is therefore discrete [10]. In essence, the model generates a population of size N at each generation by sampling N individuals with replacement from the previous generation, thereby keeping the population size constant (Figure 2.4). Consequently, gene frequencies are redistributed randomly at each generation and this leads to a simulation of neutral evolution by genetic drift [64]. Although originally described for fixed-size populations, both Moran and Wright-Fisher models can be adapted to simulate growth. In case of cancer, the continuous Moran process would be more realistic to simulate a growing cancer, because

it describes a dynamic population in which each cell is considered separately [65]. However, it is also computationally less efficient. Furthermore, the Moran and Wright-Fisher processes are highly similar in their properties [10]. Therefore, the latter is preferred for computer simulations and its potential in modeling cancer is being further explored.

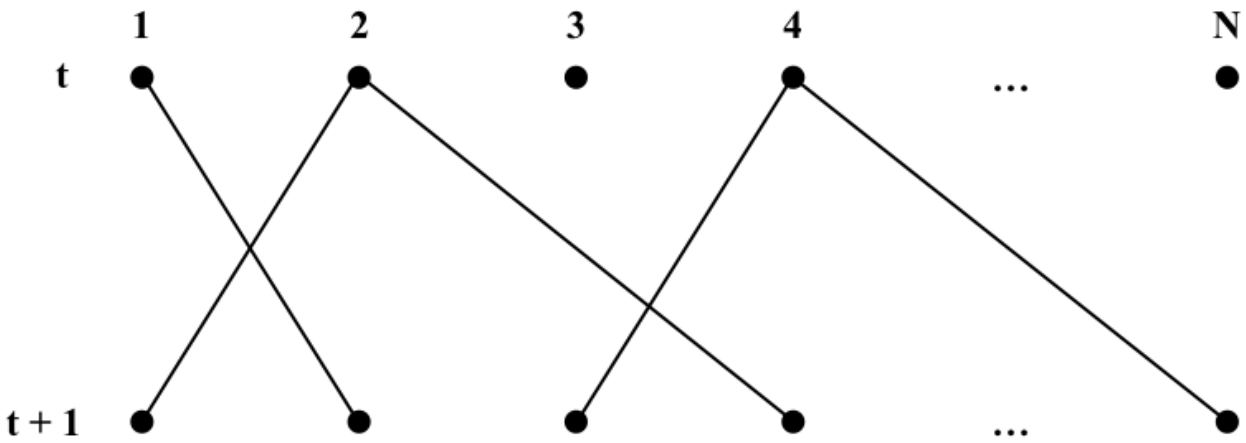

Figure 2.4. Schematic representation of the Wright-Fisher process. At generation *t*, the population consists of *N* types of individuals. For the population at generation *t + 1*, *N* individuals are sampled with replacement from the previous generation. Consequently, some types will now have an increased presence (2 and 4), while others have disappeared (3). This process was originally used to describe random genetic drift.

2.2.2.2 The Wright-Fisher model can be applied to simulate tumor growth

The Wright-Fisher process forms the basis of the model constructed in this thesis. It can easily be adapted to simulate a growing population of cells by assuming that the population size increases at each generation, usually exponentially in the case of cancer. The stochastic sampling of cell types then occurs on the expanded population [66]. The probability of a given genotype to be selected is proportional to its relative frequency and the population distribution is binomial or, in case of more than two types, multinomial [10]. Mutations can be incorporated into the model by allowing cells to change their genotype with a probability that is in correspondence with realistic mutation rates. In addition, a selective advantage can be assigned to a mutation in order to mimic Darwinian evolution. The fitness is then reflected by a weight that increases the probability of being sampled by the multinomial distribution [65].

This approach has been successfully used in a number of publications. Beerenwinkel *et al.* estimated the waiting time to cancer and suggested that up to 20 driver mutations could be

required for tumor initiation [11]. Another study succeeded in modeling the genetic instability of cancer cells caused by the mutator phenotype and its influence on tumor progression [12]. Finally, the Wright-Fisher model was applied to simulate a tumor growing under neutral evolution and lead to the postulation that the degree of heterogeneity observed in some cancer genomes is more related to a non-Darwinian mode of evolution [8].

Chapter 3

Aims

There is an ongoing debate in cancer research about the evolutionary forces driving cancer. Historically, the dominant view considers that tumor growth is governed by a Darwinian mode of evolution involving the selection of cells with increased fitness. However, this view is increasingly being called into question. The opposing view argues that stochasticity in the form of neutral drift dominates the evolution of at least a few forms of cancer. Both theories are supported by substantial evidence. It is likely that both forces act together, but the question remains how important the contribution of each one is. Adequate experimental methods exist that can retrieve accurate data from cancers at the molecular level. However, there is still a lack of suitable mathematical models that provide a theoretical framework in which this data can be analyzed quantitatively.

Therefore, a new model is constructed in this thesis, based on the Wright-Fisher process. This model is adapted to allow the simulation of growing tumor populations undergoing stochastic genetic drift and natural selection. For this, the biological process of cell division, cell death and mutation are incorporated as stochastic events, as well as the implementation of fitness weights to imitate selection. The goal is to obtain a model that is (i) mathematically tractable, to allow analytical analysis; (ii) general, so that it contains the key elements of evolution but can be adapted to study the influence of different parameters; and (iii) 'mean-field' (i.e. not required to simulate individual cells), since this makes it computationally more efficient. The potential of the model is investigated by applying common analyses on the simulation results and comparing with established results from the literature, with the goal of demonstrating the viability of the Wright-Fisher process to simulate growing tumor populations. The model is then used to assess the influence of selective pressure on tumor heterogeneity. Populations are simulated under different levels of selection and compared among each other and to a population grown under purely stochastic evolution. In doing so, this thesis aims to contribute to the ongoing research and debate concerning the evolutionary dynamics of cancer progression.

Chapter 4

Methods

4.1 Wright-Fisher model for a growing tumor population

A Wright-Fisher-based model is used to simulate the growing tumor population in a discrete stochastic manner [10]. As was briefly introduced in Section 2.2.2, the original Wright-Fisher process can be adapted to allow simulation of a growing population. Individual cells are still randomly sampled with replacement according to a multinomial distribution to produce offspring. However, in this case the progeny does not replace the previous generation, but rather is added to it to increase the population size, i.e. the selected cells are those that will *divide*. The sampling with replacement reflects the fact that stochastic fluctuations in growth rate can allow some cells to divide multiple times in the same generation. This approach preserves the original Wright-Fisher process but without letting the previous generation go extinct. However, the effect of cell death is now lost, although it usually plays an important factor as an evolutionary mechanism. Therefore, cell death is treated separately, whereby cells are randomly selected to die according to a hypergeometric distribution. Essentially, cells are sampled to die without replacement, as a cell cannot die more than once. Finally, the number of mutating cells is randomly sampled from the dividers according to a binomial distribution (Figure 4.1. Following the *infinite alleles* model, it is assumed that each mutation gives rise to a new unique subclone [67]. This new subclone will then inherit the characteristics of its parent clone in addition to the new mutation. These processes are described in detail in the sections below, both for a neutrally evolving population and in case of natural selection. It is assumed that the tumor population is initiated by a single cell that will form the ancestral clone. For each new subclone that arises, the clonal ancestry is recorded. The cells are grouped according to their clonal heritage and the selection of the number of dividers, deaths and mutations is based on the clone sizes. This prevents the need of storing each cell in memory, which can become infeasible for large populations. Instead, as all cells belonging to the same clone or subclone are identical

by definition, it suffices to track the clonal subpopulations and the number of cells contained within them. Generations, denoted by t, are here defined as the time-step corresponding to one cycle of selecting dividers, deaths and mutations for the entire population. The population parameters are summarized at the end of this section (Table 4.1).

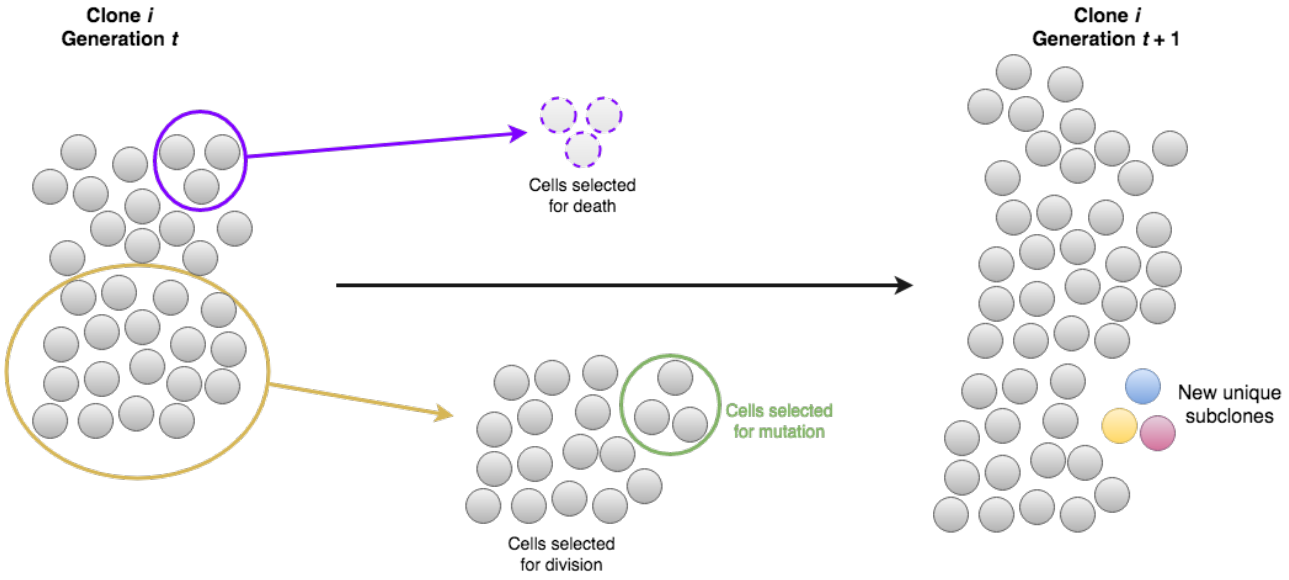

Figure 4.1. Schematic description of the model. A clone i at generation t undergoes a generational transition, whereby some cells are randomly selected for death and some for division. From the dividing cells in turn, a number of mutating cells are randomly sampled. Each mutation gives rise to a new unique subclone, following the infinite allele assumption. The clone at generation $t+1$ consists of the original cells from the previous generation, minus the number of deaths and plus the number of divisions. The mutated cells will form new subclones that are descendants from clone i.

4.1.1 Stochastic divisions

4.1.1.1 Neutral evolution

Following the Wright-Fisher model, the number of offspring in a generation is fixed to the total number of cells in the population. Specifically, in a population of K clones, the clone size $n_i(t)$ at generation t for each clone $i \in [1, 2, \ldots, K]$, is known. To simulate the size distribution of the population at the next generation $t+1$, the number of dividers $p_i(t+1)$ for each clone is stochastically selected from the total population size N. According to the multiple-allele Wright-Fisher model, the size for the next generation follows a multinomial distribution [10]. The equation has been adapted here to simulate the number of *dividing* cells:

$$
\begin{aligned}
&P\left\{p_1(t+1)=p_1,\ p_2(t+1)=p_2,\ \ldots,\ p_K(t+1)=p_K\right\} \\
&\quad=\frac{N!}{p_1!\,p_2!\ldots p_K!}\left(\frac{n_1(t)}{N}\right)^{p_1}\left(\frac{n_2(t)}{N}\right)^{p_2}\cdots\ \left(\frac{n_K(t)}{N}\right)^{p_K}
\end{aligned}
\tag{4.1}
$$

where $p_i(t+1)$ is the amount of cells that will be added to the current clone size $n_i(t)$. At generation $t+1$, the size of clone i will thus be $n_i(t+1)=n_i(t)+p_i(t+1)$, depicting stochastic growth.

However, this distribution does not allow efficient determination of $p_i(t+1)$, because each clone has to be considered simultaneously. It would be mathematically more tractable if each clone could be handled separately. Therefore, a new equation is derived. When considering the random selection of dividers $p_1(t+1)$ for any given clone, the probability distribution can be described by:

$$
P\left\{p_1(t+1)=p_1\right\}=\binom{N}{p_1}\left(\frac{n_1(t)}{N}\right)^{p_1}\left(1-\frac{n_1(t)}{N}\right)^{N-p_1}
\tag{4.2}
$$

This is a selection based on a biniomial distribution, in which there are N divisions available for this first clone with a probability equal to its relative fraction in the total population. If the number of dividers (p_1) for one clone is known, there will be $N-p_1$ divisions still available for all other clones. Consequently, the conditional probability that there will be $p_2(t+1)$ dividers for any other clone, given that p_1 divisions have already been selected, is:

$$
\begin{aligned}
&P\left\{p_2(t+1)=p_2 \mid p_1(t+1)=p_1\right\} \\
&\quad=\binom{N-p_1}{p_2}\left(\frac{n_2(t)}{N-n_1(t)}\right)^{p_2}\left(1-\frac{n_2(t)}{N-n_1(t)}\right)^{N-p_1-p_2}
\end{aligned}
\tag{4.3}
$$

This process can now be repeated for all subsequent clones. Thus, a general expression for the probability that a clone i will have $p_i(t+1)$ dividers is obtained:

$$
\begin{aligned}
&P\left\{p_i(t+1)=p_i \mid p_j(t+1)=p_j\ \forall\ j\in[1,2,\ldots,i-1]\right\} \\
&\quad=\binom{N-\sum_{j=1}^{i-1}p_j}{p_i}\left(\frac{n_i(t)}{N-\sum_{j=1}^{i-1}n_j(t)}\right)^{p_i}\left(1-\frac{n_i(t)}{N-\sum_{j=1}^{i-1}n_j(t)}\right)^{N-\sum_{j=1}^{i-1}p_j-p_i}
\end{aligned}
\tag{4.4}
$$

When $p_i(t+1)$ has been determined for all i up to $K-1$, then $p_K(t+1)$ for the last clone will be fixed as $p_K(t+1) = N - \sum_{j=1}^{K-1} p_j$, because all cells are distributed among the clones in the population ($N = \sum_{i=1}^{K} p_i(t+1)$). Therefore, the total distribution of dividers for K clones can be determined by calculating the distribution for $K-1$ clones, which can be expressed as the product of Equation (4.2) and the conditional probabilities given by Equation (4.4) for $i \in [2, \ldots, K-1]$:

$$\begin{aligned} &P\left\{p_1(t+1) = p_1,\ p_2(t+1) = p_2,\ \ldots,\ p_{K-1}(t+1) = p_{K-1}\right\} \\ &= P\left\{p_1(t+1) = p_1\right\} \cdot \prod_{i=2}^{K-1} P\left\{p_i(t+1) = p_i \mid p_j(t+1) = p_j\ \forall\ j \in [1, 2, \ldots, i-1]\right\} \end{aligned} \tag{4.5}$$

It can be shown that Equation (4.5) is equivalent to Equation (4.1) (for a detailed description, see appendix A). Therefore, Equation (4.4) can be correctly used to describe the probability distribution of the number of dividing cells for each clone at any generation. This allows simulation of the dividers $p_i(t+1)$ in an iterative fashion for all clones $i \in [1, 2, \ldots, K]$ by drawing from a binomial distribution:

$$p_i(t+1) \sim \mathrm{B}\left\{N(t) - \sum_{j=1}^{i-1} p_j(t+1),\ \frac{n_i(t)}{N(t) - \sum_{j=1}^{i-1} n_j(t)}\right\} \tag{4.6}$$

It is important to note that, as equations (4.5) and (4.1) are equivalent, the order in which the clones are processed by (4.6) does not matter.

4.1.1.2 With selection

Selection can be implemented in the model by assigning fitness advantages s_i to each clone. These values are then used as a weight for the probability of the clone being selected to divide, i.e. a higher fitness will result in an increased growth rate for the clone. Equation (4.6) for each clone now becomes:

$$p_i(t+1) \sim \mathrm{B}\left\{N(t) - \sum_{j=1}^{i-1} p_j(t+1),\ \frac{s_i n_i(t)}{\sum_{j=1}^{K} s_j n_j(t) - \sum_{j=1}^{i-1} s_j n_j(t)}\right\} \tag{4.7}$$

Similar to Foo *et al.*, a random distribution from which the fitness advantages are drawn is considered [63]. Here, the ancestral clone that initiates the population is set to have a fitness

advantage of 1. Each new subclone will then be assigned a new fitness weight by drawing from a *gamma distribution* around the weight of its parent clone (Figure 4.2). As such, the fitness of subclones will fluctuate around that of their parent clone. The choice for a gamma distribution is motivated by the fact that it is bounded at 0 so that fitness weights can never become negative and because it can be shaped to have a long positive tail, reflecting the possibility that a mutation can confer a significantly increased fitness. A gamma distribution is defined by its *shape* k and its *scale* θ, with the mean given by $k \cdot \theta$ and the variance $\sigma^2 = k \cdot \theta^2$:

$$P\left\{\text{Fitness of clone } i = s_i\right\} = \frac{1}{(k-1)!\ \theta^k}\ s_i^{k-1}\ e^{\frac{s_i}{\theta}} \tag{4.8}$$

In this case, the mean of the distribution is set to the fitness weight of the parent clone. To allow simulation of populations with different levels of selective pressure, a *selection factor* is defined, which will take the value of the scale θ for the gamma distribution. A higher selection factor will thus result in a higher selective pressure as there will be a larger variance in fitness weights. With θ defined as a parameter of the population and the mean equal to the weight of the parent clone (which will be denoted as s_i^* here), the gamma distribution from which the fitness weight s_i of a subclone i will be drawn, is given by:

$$s_i \sim \Gamma\left\{k = \frac{s_i^*}{\theta},\ \theta\right\} \tag{4.9}$$

4.1.2 Stochastic mutations

During cell division, the probability exists that mutations occur in the genome. Within each clone, some of the cells will mutate at each generations. These mutated cells will then carry an additional mutation to the cell's genotype and give rise to new subclones. This process can be modeled by randomly selecting the cells destined to mutate from the pool of dividers within each clone. In particular, a Bernouilli trial is performed for each cell to determine whether it will mutate or not, with a probability that is equal to the mutation rate u of the population. This results in a selection according to a binomial distribution of the number of mutating cells $m_i(t)$, at generation t in a clone i, from the cells that have divided at that generation ($p_i(t)$):

$$m_i(t) \sim \mathrm{B}\left\{p_i(t),\ u\right\} \tag{4.10}$$

whereby the mutation rate u is that of the entire functional part of the genome per cell division and is considered to be constant for the whole population.

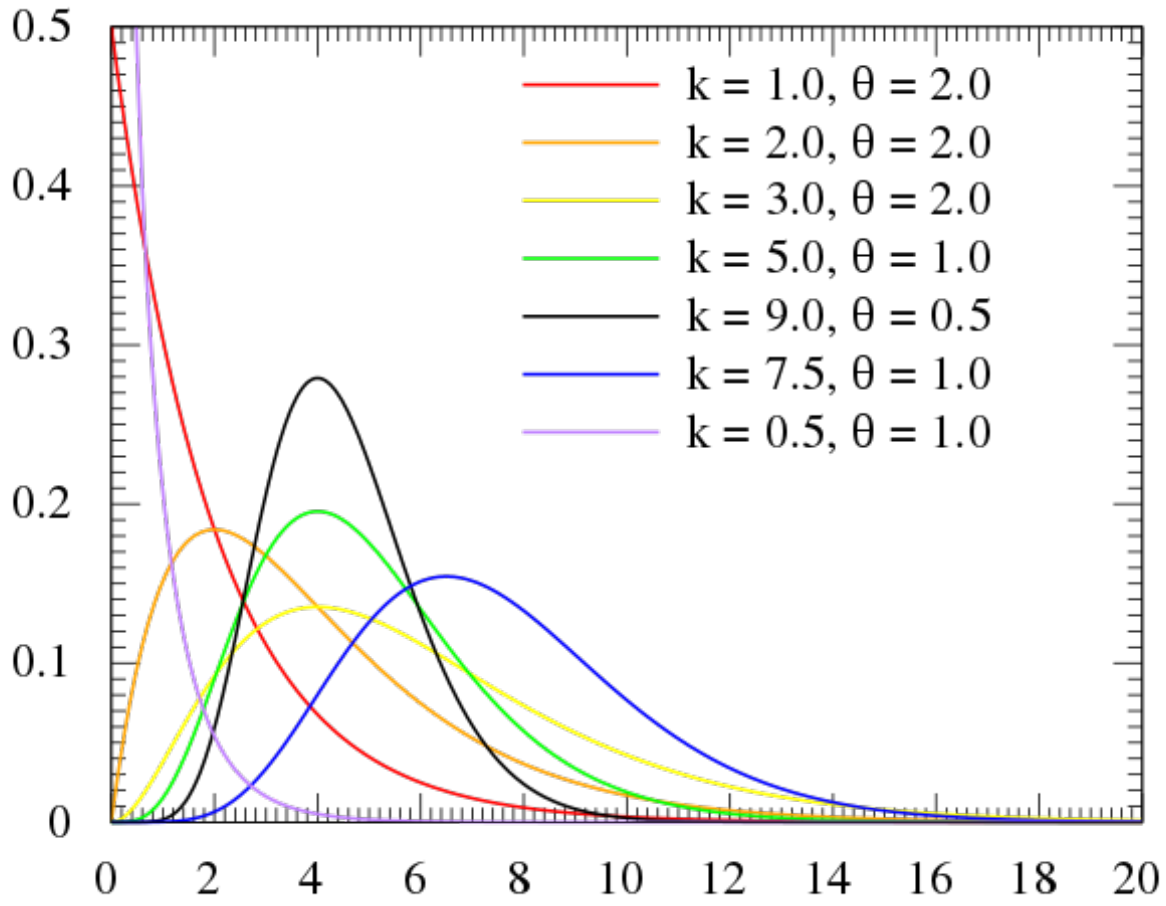

Figure 4.2. The gamma distribution. Illustration of the probability density function for gamma distributions with various values the parameters k and θ. Taken from `https://en.wikipedia.org/wiki/Gamma_distribution` (Accessed on 26/05/2018).

4.1.3 Stochastic deaths

There is always a chance that some of the cells will die due to a lack of resources or to targeting by external factors. To incorporate this in the model, it is considered that at each generation a fraction $0 < f < 1$ of the total population size N will not survive to the next generation (to prevent extinction, the fraction is kept smaller than 1). Similar to the number of divisions, this fixes the number of deaths to the total population size at each generation. A cell can only die once during a generation, as opposed to the possibility of multiple divisions. Therefore, the sampling now needs to occur without replacement and thus a hypergeometric distribution is used. The joint hypergeometric distribution to sample $d_i(t+1)$ deaths for each clone $i \in [1, 2, \ldots, K]$ in a population of size N, considering that fN deaths occur in total, is described as:

$$P\left\{d_1(t+1)=d_1,\ d_2(t+1)=d_2,\ \ldots,\ d_K(t+1)=d_K\right\} = \frac{\binom{n_1(t)}{d_1}\binom{n_2(t)}{d_2}\ \cdots\ \binom{n_K(t)}{d_K}}{\binom{N}{fN}} \tag{4.11}$$

whereby $n_i(t)$ is the size clone i at generation t. This equation poses the same computational difficulty as was encountered for the stochastic divisions with Equation (4.1). Therefore, a similar approach is taken to derive a new equation that allows sampling of each clone iteratively. The probability distribution for sampling $d_1(t+1)$ deaths for a first clone is:

$$P\left\{d_1(t+1)=d_1\right\} = \frac{\binom{n_1(t)}{d_1}\binom{N-n_1(t)}{fN-d_1}}{\binom{N}{fN}} \tag{4.12}$$

Following the same reasoning behind Equation (4.4), the pool of 'available' deaths for each clone is decreased by the number of deaths that were sampled for the previous clones, while the total population size from which to sample is decreased by the previous clones' sizes. Thus, when $d_1(t+1)$ is known for the first clone, the conditional distributions of the number of deaths for the second clone, and more general for each clone i are:

$$P\left\{d_2(t+1)=d_2 \mid d_1(t+1)=d_1\right\} = \frac{\binom{n_2(t)}{d_2}\binom{N-n_1(t)-n_2(t)}{fN-d_1-d_2}}{\binom{N-n_1(t)}{fN-d_1}} \tag{4.13}$$

$$P\left\{d_i(t+1)=d_i \mid d_j(t+1)=d_j\ \forall\ j\in[1,2,\ldots,i-1]\right\} = \frac{\binom{n_i(t)}{d_i}\binom{N-\sum_{j=1}^{i-1}n_j(t)-n_i(t)}{fN-\sum_{j=1}^{i-1}d_j-d_i}}{\binom{N-\sum_{j=1}^{i-1}n_j}{fN-\sum_{j=1}^{i-1}d_j}} \tag{4.14}$$

Similar to the case of stochastic divisions, when $d_i(t+1)$ has been determined for each clone except the last, then the number of deaths for the last clone K is fixed as $fN - \sum_{j=1}^{K-1} d_j$. It can again be shown that the serial product of equations (4.12) and (4.14) for all clones $i \in [1, 2, \ldots, K-1]$ is equivalent to Equation (4.11) for all K clones:

$$\begin{aligned} &P\left\{d_1(t+1) = d_1,\ d_2(t+1) = d_2,\ \ldots,\ d_K(t+1) = d_K\right\} \\ &= P\left\{d_1(t+1) = d_1\right\} \cdot \prod_{i=2}^{K-1} P\left\{d_i(t+1) = d_i \mid d_j(t+1) = d_j\ \forall\ j \in [1, 2, \ldots, i-1]\right\} \end{aligned} \tag{4.15}$$

The proof for this equivalence is described in appendix A. Equation (4.14) is therefore a correct way to describe the probability distribution of the number of deaths $d_i(t+1)$ for each clone i. During simulation of the tumor population, $d_i(t+1)$ is sampled iteratively from the hypergeometric distribution as follows:

$$d_i(t+1) \sim \mathrm{H}\left\{n_i(t),\ N - \sum_{j=1}^{i-1} n_j(t),\ fN - \sum_{j=1}^{i-1} d_j(t+1)\right\} \tag{4.16}$$

whereby H represents the hypergeometric distribution for a population of size N, with $n_i(t)$ the target subpopulation, $N - \sum_{j=1}^{i-1} n_j(t)$ the remainder of the population and $fN - \sum_{j=1}^{i-1} d_j(t+1)$ the number of samples to be drawn.

4.1.4 General growth equation

Finally, the evolution of each clone within the population can now be simulated by using equations (4.6), (4.10) and (4.16):

$$n_i(t+1) = n_i(t) + p_i(t+1) - m_i(t+1) - d_i(t+1) \tag{4.17}$$

In essence, the clone size is increased with the number of dividers minus the number of mutations (as these will form new subclones) and decreased by the number of deaths. Each mutation gives rise to a new subclone with a starting size of 1. The total population size at generation $t+1$ is thus given by:

$$N(t+1) = \sum_{i=1}^{K} n_i(t+1) + \sum_{j=1}^{M} n_{K+j}(t+1) \tag{4.18}$$

with $n_{K+j}(t+1) = 1$ and $M = \sum_{i=1}^{K} m_i(t+1)$ the total number of mutations.

The population that is thus obtained follows exponential growth. Its size will increase each time-step with a factor $(2 - f)$ as the total number of divisions equals N and the total number of deaths is equal to fN. Therefore, the following growth equation for the entire population is obtained:

$$N(t) = N_0 \cdot (2 - f)^t = N_0 \cdot e^{\gamma t} \tag{4.19}$$

where $\gamma = \ln(2 - f)$ is the growth rate. N_0 will generally be set to 1.

Table 4.1. Summary of the population parameters used in the model. The corresponding symbols and brief descriptions are provided.

Symbol	Parameter	Description
u	Mutation rate	Probability with which mutating cells (m_i) are sampled from the dividers (p_i), according to a *binomial distribution*.
f	Fraction of deaths	Fraction of the total population (N) that will die. Deaths for each subclone (d_i) are sampled from fN according to a *hypergeometric distribution*. Determines the overall population growth rate as $\gamma = \ln(2 - f)$.
θ	Selection factor	Scale of the *gamma distribution* (with mean = s_i^*, the weight of the parent clone) from which the selection weights (s_i) for each subclone are sampled.

4.2 Simulation data analyses

4.2.1 Artificial sampling method

Clinical data usually contain only a small fraction of the genotypes from the total tumor, due to limited sampling. Mutations with higher frequency will have a higher probability to be sampled. Therefore, an artificial sampling method is constructed to mimic the limited detection of clones for the simulated populations. The true clone sizes are known, so sampling can be performed by randomly selecting cells from the population with a probability proportionate to their relative size. This results in the following multinomial distribution for K clones:

$$P\{x_1,\ x_2,\ \ldots,\ x_K\} = \frac{N_S!}{x_1!\ x_2! \ldots x_K!} \left(\frac{n_1}{N_T}\right)^{x_1} \left(\frac{n_2}{N_T}\right)^{x_2} \ldots \left(\frac{n_K}{N_T}\right)^{x_K} \tag{4.20}$$

where x_i is the sampled number of cells of clone i, n_i the real clone size, N_S the total number of cells sampled and N_T the total number of cells in the original tumor. Equation (4.20) is the same distribution that was used for the number of dividers in the Wright-Fisher model (see

Section 4.1.1). A similar derivation can be conducted to achieve the following distribution, which can be used to sample each clone separately:

$$\begin{aligned} &P\left\{x_i \mid x_j \;\forall\; j \in [1, 2, \ldots, i-1]\right\} \\ &= \binom{N_S - \sum_{j=1}^{i-1} x_j}{x_i} \left(\frac{n_i}{N_T - \sum_{j=1}^{i-1} n_j}\right)^{x_i} \left(1 - \frac{n_i}{N_T - \sum_{j=1}^{i-1} n_j}\right)^{N_S - \sum_{j=1}^{i-1} x_j - x_i} \end{aligned} \tag{4.21}$$

Consequently, each clone can be sampled iteratively according to a binomial distribution based on the previously sampled clones:

$$x_i \sim \mathrm{B}\left\{N_S - \sum_{j=1}^{i-1} x_j,\; \frac{n_i}{N_T - \sum_{j=1}^{i-1} n_j}\right\} \tag{4.22}$$

For the simulations performed in this study, populations will generally be grown to a size of 10^8 cells and the sampling size will be set to 10^6 so that 1% of the tumor is covered, which is in accordance with many clinical studies.

4.2.2 Number of mutations per clone

The number of mutations that a clone carries is viewed in respect to the ancestral clone at the origin of the tumor. Thus, the number of mutations for the ancestral clone is regarded as being 0 and for each subclone as the number of accumulated mutations relative to the ancestral clone. The subclones that are directly descended (first clonal generation) from the ancestral clone will have one mutation, their own subclones (second clonal generation) will have two, and so forth. To get an estimate of the number of mutations per subclone, a weighted average is calculated as follows:

$$\overline{M} = \frac{\sum_{i=1}^{K} (n_i \cdot M_i)}{\sum_{i=1}^{K} n_i} \tag{4.23}$$

for K subclones, with n_i the size of each subclone and M_i the number of accumulated mutations for each subclone. Note that the size of the ancestral clone is not taken into account, as only subclones are considered.

4.2.3 Allele frequency spectrum

Allele frequencies are a measure for the relative abundance of specific genotypes or mutations within the population. For the simulated populations presented here they are defined as the fraction of cells carrying a certain mutation relative to the total population size. Each mutation event will have initiated a unique subclone and will be inherited by any further derived subclones in the branching evolution of the original subclone. The fraction of cells carrying a mutation i can thus be retrieved from the *family size* N_i of the original subclone that was initiated by mutation i:

$$\phi_i = \frac{N_i}{N} \tag{4.24}$$

The family size is hereby defined as the sum of the size of the original subclone and all cells that descended from it. In phylogenetic terms, the number of cells contained within the subtree rooted at the original subclone where the mutation first appeared. By definition, the allele frequency of the ancestral clone is 1, because all cells in the tumor population are descended from this clone.

4.2.4 Heterogeneity of the population: Simpson's diversity index

Simpson's diversity index gives the probability that two randomly selected entities belong to two different groups within a set of groups and can be used in a wide range of fields [68]. It is used here as a measure of the heterogeneity within the tumor population:

$$H = 1 - \sum_{i=1}^{K} \left(\frac{n_i}{\sum_{i=1}^{K} n_i} \right)^2 \tag{4.25}$$

with n_i the subclone sizes. It thus gives the probability that two randomly selected mutated cells are genetically different, i.e. carry different mutations.

4.2.5 Reconstruction of mutational timeline

A method to reconstruct the mutational timeline of a tumor population is proposed here and tested out on the simulated populations. It can be assumed that in a neutrally evolving population, the allele frequency of a mutation will remain virtually constant as all clones are considered to progress at equal rate. Therefore, the total population size N at the time that a mutation i occurred can be estimated as $1/\phi_i$, with ϕ_i the allele frequency of the mutation in the final

population, under the neutral evolution model [9]. For example, if the population size is 100 when a mutation occurs, its allele frequency will be 1/100, at the next generation both the population and the subclone carrying the mutation will have doubled in size, so the allele frequency is now 2/200 and so on. If the size of the population when a mutation occurred is known (from the final allele frequency), the timepoint can be retrieved from the growth Equation (4.19):

$$t_i = \frac{\ln(N(t_i))}{\gamma} = \frac{\ln(1/\phi)}{\gamma} \tag{4.26}$$

where γ is the growth rate of the population and ϕ_i is the allele frequency of the given mutation. This method has the disadvantage that it requires knowledge of the growth rate. Something which is not always readily measurable. However, the growth equation can be transformed to units of population size doublings T, by dividing t by the doubling time $\frac{\ln(2)}{\gamma}$, which cancels out the growth rate:

$$\begin{aligned} T &= t \cdot \frac{\gamma}{\ln(2)} \\ N(T) &= e^{\ln(2)T} \\ T_i &= \frac{\ln(N(T_i))}{\ln(2)} = \frac{\ln(1/\phi_i)}{\ln(2)} \end{aligned} \tag{4.27}$$

This allows to trace back the number of population size doublings T_i that took place before mutation i occurred. The errors for this reconstruction are then simply calculated by taking the difference between the reconstructed timepoints and the real ones, which are stored during simulation.

4.2.6 Multiple simulations: population scores

For each set of population parameters, 1000 simulations were run in order to perform statistical analysis. A few population 'scores' are defined to compare the different simulations. The *maximum allele frequency* is the highest occurring allele frequency value for a subclonal mutation within the population. Next, the *heterogeneity* of each simulated population is calculated with Simpson's diversity index as described before (Equation (4.25)). For these two scores, distributions are plotted in the form of histograms. The *number of mutations per clone* is calculated as described previously (see Section 4.2.2), averages and standard deviations are calculated for each number of mutations. A total average is also calculated for the entire set of simulations, according to Equation (4.23). Finally, the mutational timelines for each population is reconstructed (as described in Section 4.2.5). The median of the errors in the estimation of mutational events is calculated to get the *median absolute deviation (MAD)* of the reconstructed timepoints. The distribution of these MAD values is then used for comparisons.

Chapter 5

Results and Discussion

In the following sections, the constructed model is studied in detail. First, the mechanisms and general behavior of the model are assessed. The approach of modeling the selective advantages is discussed (Section 5.1.1) and two example simulations for a neutrally evolving population and one with selection are provided to display the main differences (Section 5.1.2). Further example runs of simulations with varying population parameters were performed for illustration. The influence of each parameter is briefly discussed to provide an overview of the different evolutionary outcomes that can be obtained from the model (Section 5.1.3). The impact of selective pressure on a tumor population is then more rigorously analyzed by performing multiple simulations with the same population parameters and applying statistical analyses (Section 5.2). Finally, the influence of the sampling method is briefly discussed (Section 5.3).

5.1 General model behavior

5.1.1 Modeling the selective advantages with a gamma distribution

The model to simulate selection employed here assumes that fitness values are randomly distributed among subclones around that of their parent clone. In essence, the fitness advantage is passed on to progeny but can be decreased by deleterious mutations or further increased by new driver mutations. A gamma distribution is used in this study and the selective advantages are implemented as changes in the division rate of a clone. Retrieving real selective advantage values from experimental data is not straightforward. Consequently, only a limited number of studies has attempted this and usually with simplified approximate models, as measuring the advantage of a mutation directly would require extensive sampling over long time periods from the tumor [31]. Therefore, the choice for a gamma distribution is somewhat arbitrary and is

mainly motivated by the fact that it can be adjusted to reflect a dispersal of values with a peak around a given mean and with a positive tail. This is what one could expect if fitness is inherited and new mutations have a small probability of conferring a higher advantage.

The fitness of the ancestral clone is set to 1 and so the selective weight of each subclone can be regarded as a change in division rate relative to the original clone. Within this setting, the level of selection can be viewed as the variance in the distribution of selection weights. Neutral evolution corresponds to a variance of 0, meaning all clones will have identical division rates. As described in the Methods (Section 4.1.1.2), the scale θ of the gamma distribution is used to parametrize the level of selection. The variance is given by $\sigma^2 = k \cdot \theta^2$ and as the mean of the distribution is set equal to the selective weight of the parent clone $s_i^* = k \cdot \theta$, the variance can be expressed as $\sigma^2 = s_i^* \cdot \theta$. Thus, the variance increases linearly with θ and for higher values, the chance that a subclone emerges with extremely high fitness increases. Therefore, θ is used as the *selection factor* parameter of a population. Different values of selection are explored, ranging from 0 (neutral) up to 3 (Figure 5.1). It is clear that for higher selection factors, the population indeed contains subclones that have an increasingly higher fitness than the ancestral clone. This way, the presented model allows the study of populations under varying levels of selective evolution. In Section 5.2, this aspect will be explored further by running multiple simulations with the same set of parameters for different selection factors.

5.1.2 Simulation of individual tumors

Simulated tumor populations were grown to a minimal size of $N = 10^8$ cells. This corresponds to a size of approximately 1 cm^3, the typical detection limit of a tumor [69]. A sample of 10^6 cells is taken according to the multinomial sampling method, with a detection limit of minimally 100 cells per clone (see Section 4.2.1). Generally, a constant mutation rate u and *selection factor* are assumed. A standard mutation rate of 10^{-3} was used, corresponding to the rate of mutations observed in the functional part of the genome in some common cancers [29, 30]. The *selection factor* determines the variance in the gamma distribution of selective advantages (see Section 4.1.1.2) and reflects the level of selective pressure exercised on the population. In case of a neutrally evolving tumor, this parameter is set to 0. The growth dynamics are controlled with the parameter *f*, which determines the total number of deaths at each generation as *fN* and is also kept constant during simulation. The fraction of deaths *f* was standard set to 0.3, resulting in a growth rate of $\ln(1.7) = 0.53$ (see Section 4.1.4), which is in the order of magnitude of realistic growth rates [70] and provides clear simulation results. The evolution of the distinct subclones is described by the Wright-Fisher process. This approach allows one to follow the progression of the tumor composition over time (Figure 5.2A-C). Here, the subclone *family sizes* are followed, which are defined as the sum of the sizes of a subclone

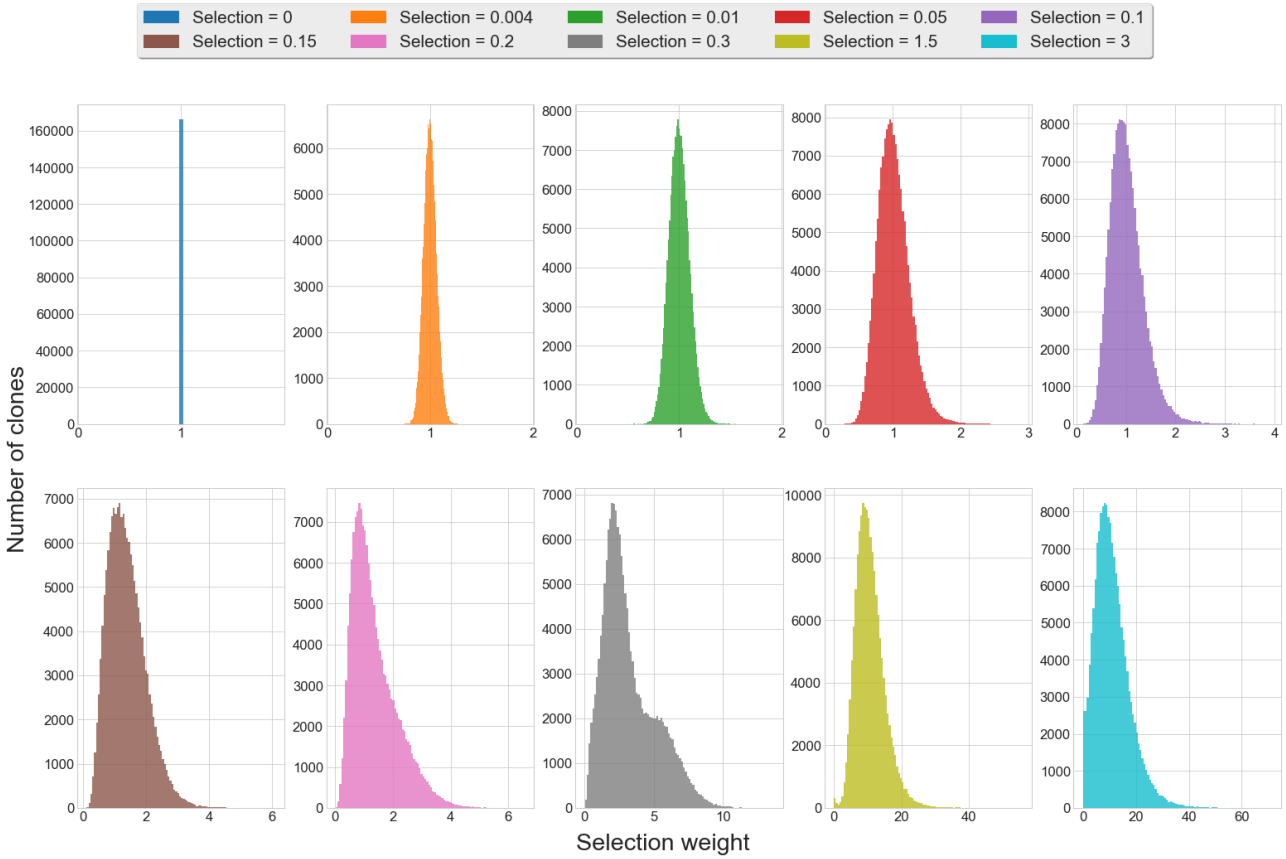

Figure 5.1. Distributions of fitness weights for different selection factors. For each level of selection, a population was grown to a minimal size of 10^8 with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.3$. The fitness weights of the subclones are plotted as histograms for each population.

and all its descendants. A phylogenetic tree can be reconstructed from the final population because the clonal ancestry of each clone is tracked (Figure 5.2B-D).

In these first example simulations (Figure 5.2), the neutrally evolving population shows a high number of first-generation subclones (those that are directly descended from the ancestral clone), while in case of selection a few subclones are selected early on. These successful subclones then evolve further and give rise to new subclones of their own, resulting in a 'deeper' phylogenetic tree, i.e. a population with more clonal generations. This also results in an increased number of accumulated mutations carried by each subclone (Figure 5.3C). In some cases, a favored subclone can take over the entire population and so an originally subclonal mutation can become clonal (i.e. shared by the entire population, e.g. the brown subclone in Figure 5.2C). This is also visible in the allele frequency spectrum, which shows the number of mutations that occur in a certain fraction of the population (Figure 5.3A-B). Mutations that reach an allele frequency close to 1 (i.e. carried by every cell) are common when selection is applied, while they were not observed for neutral evolution.

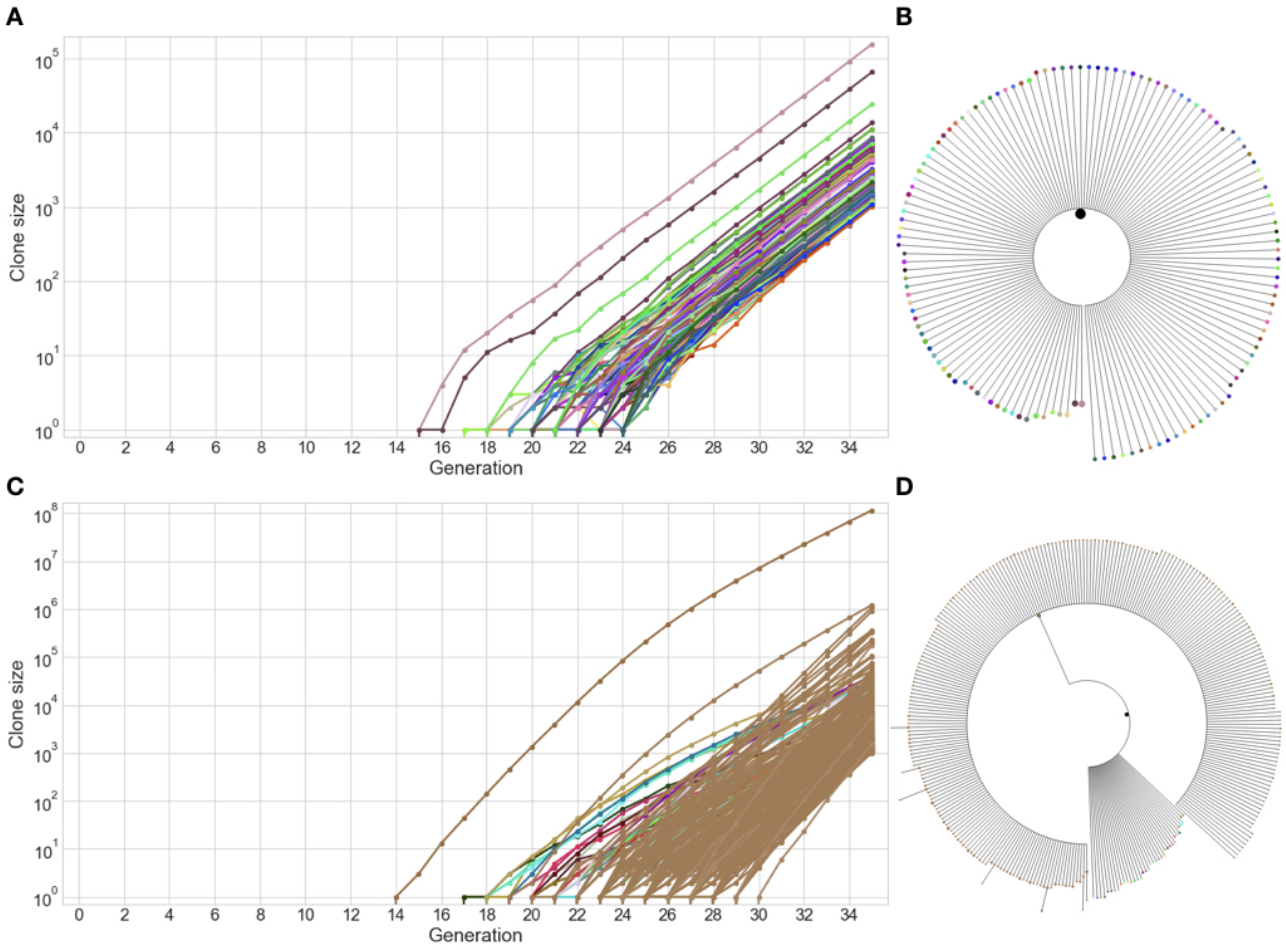

Figure 5.2. Simulation results for populations under neutral evolution and selection. Populations were grown to a minimal size of 10^8, with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.3$. In case of selection the selection factor was 0.3. A color code is used to discern the different subclones. Each clone has its own color, whereby later-generation subclones inherit the color of their parent clone in a lighter tint. (A) Evolution of subclone family sizes for the neutrally evolved population. (B) Phylogenetic tree of the neutrally evolved population, subclones are depicted by the same colors as in A, branch lengths correspond to the time of birth for each clone. (C) Evolution of subclone family sizes for the population under selective pressure. (D) Phylogenetic tree corresponding to the population from C. From C and D it is clearly visible that the brown-colored subclone is taking over the population (the first-generation subclone reaches a size of approximately 10^8, equal to the final population size, i.e. virtually all cells are descended from this clone).

According to Williams *et al.* [9], the cumulative number of mutations is inversely related to the allelic frequency ($M(\phi) \sim \frac{1}{\phi}$) when the tumor grows under neutral evolution conditions. This hypothesis is tested for the model presented here. The cumulative number of mutations per inverse allelic frequency is calculated based on the allele frequency spectrum. Each frequency represents a distinct mutation; these are inverted and then sorted in increasing order. The unique values of $1/\phi$ are determined and the number of times they appear is counted. When

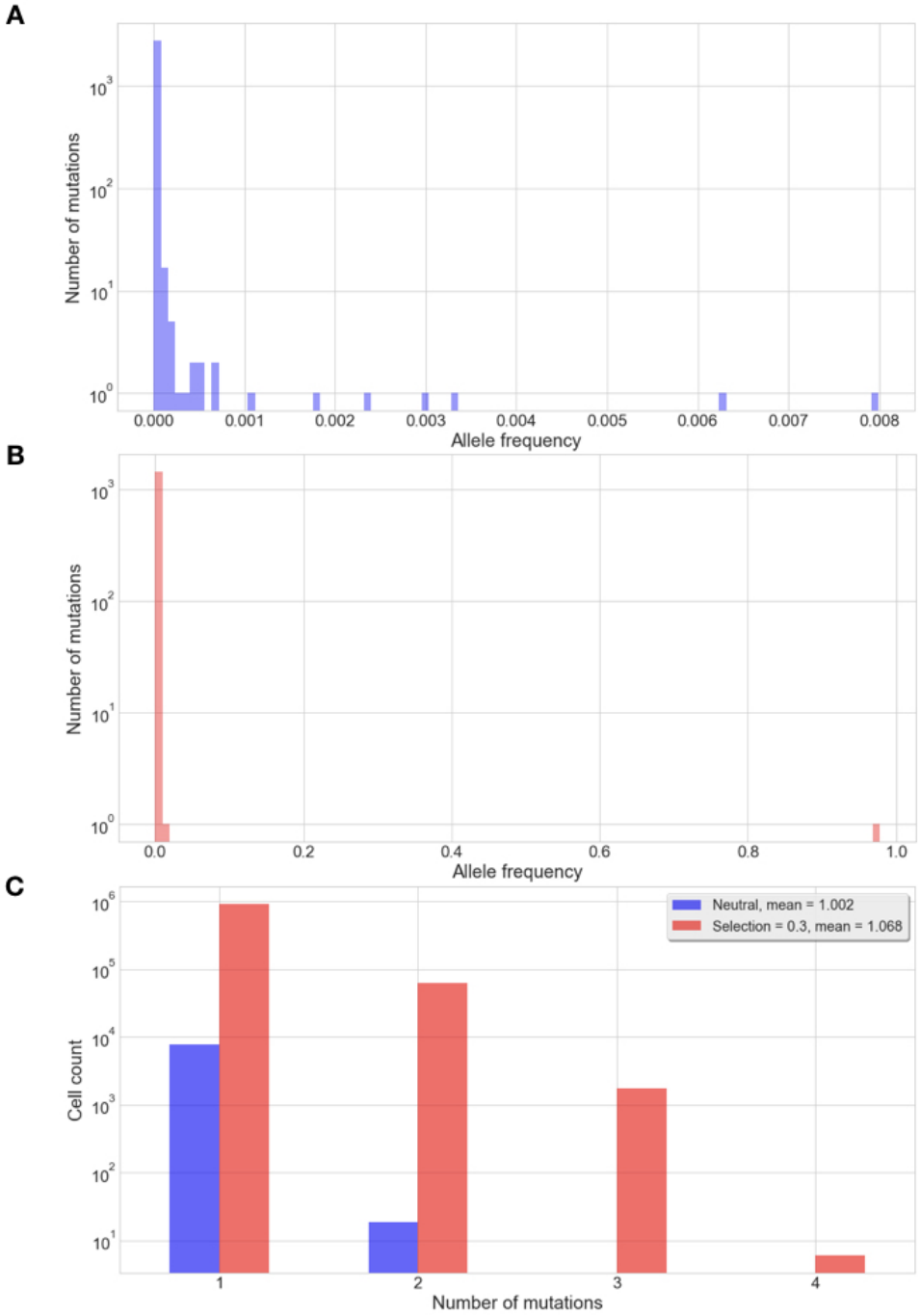

Figure 5.3. Allele frequencies and number of mutations for simulated populations. Populations were grown to a minimal size of 10^8, with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.3$. In case of selection the selection factor was 0.3. Prior to analysis, the populations were sampled using the multinomial sampling method at 1% and a detection limit of 100 cells. (A) Allele frequency spectrum for the neutrally evolved population. (B) Allele frequency spectrum for the population grown under selective pressure. (C) Distribution of the number of accumulated mutations per subclone for both populations. The weighted means are given in the legend.

taking the cumulative sum of these counts, one obtains the cumulative number of mutations $M(\phi)$ per allelic frequency. A fit by linear regression is then calculated between $M(\phi)$ and $1/\phi$. When applying the method from Williams *et al.*, a relatively good fit with high R^2 is obtained for the neutral population, while the selective population has a lower R^2 (although still high) and shows the same kind of bending in the curve that was reported in the original study (Figure 5.4). The simulations thus seem to agree with the model proposed by Williams *et al.* [9]. However, the results presented here suggest that the difference between neutral and selective evolution is not significant for this analysis.

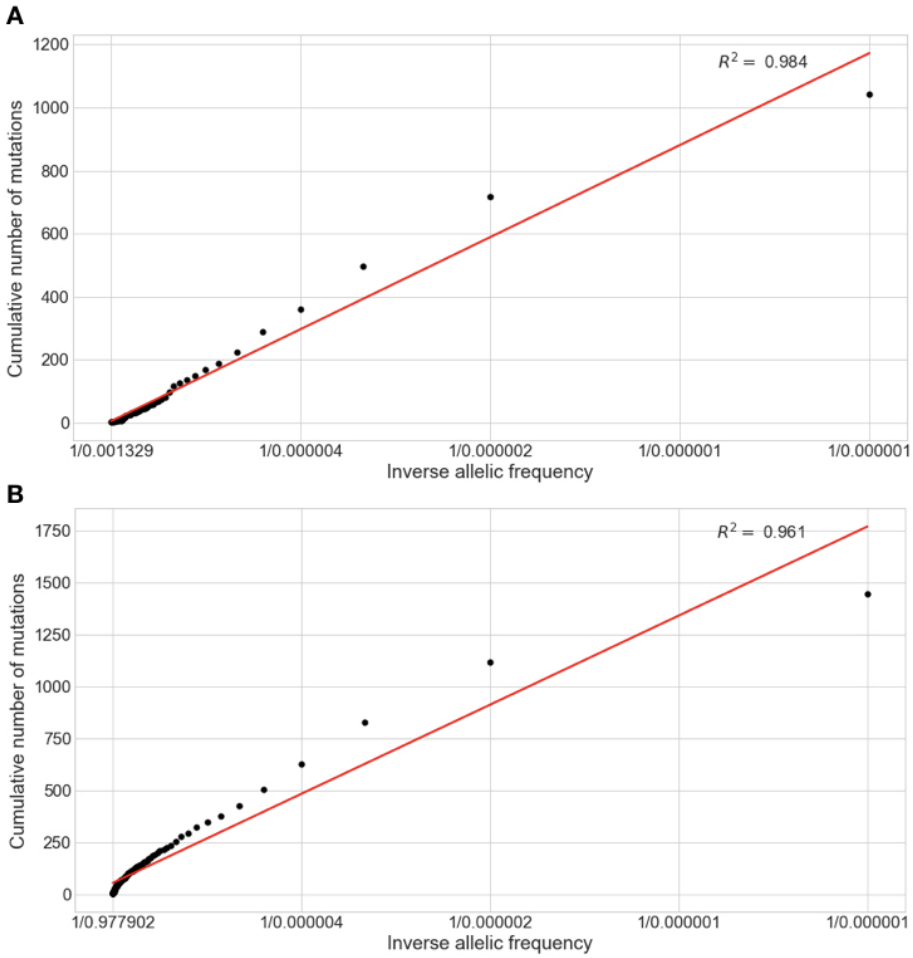

Figure 5.4. Fitting the accumulated distribution of mutations to the inverse allele frequencies. The method of Williams *et al.* [9] was used to assess the relation between the cumulative number of mutations and the inverse allele frequency by linear regression. Populations were grown to a minimal size of 10^8, with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.3$. In case of selection the selection factor was 0.3. Prior to analysis, the populations were sampled using the multinomial sampling method at 1% and a detection limit of 100 cells. (A) Fit for the neutrally evolved population. (B) Fit for the population grown under selective pressure. The R^2 values are given in the chart.

The heterogeneity of a population can be measured with Simpson's diversity index [68, 71, 72]. It is used here to calculate the probability that two randomly selected cells belong to different subclones and thus have divergent genetic backgrounds. A value of 0.96 is found for the neutral population, depicting a large variety in subclonal mutations. In case of selection, the index equals only 0.13. Thus, these first simulations indicate that the model predicts a higher degree of heterogeneity in absence of selection. This is in line with the results from previous studies [6, 8]. The proposed model therefore seems to follow the assumptions from established literature. Note however that the results discussed here are from single simulations for each population and only serve to provide an overview of the model's potential. In Section 5.2 a more elaborate investigation of the model is provided based on multiple simulations for varying degrees of selective pressure. In the following section, the different population parameters and their influence on heterogeneity are further explored.

5.1.3 Influence of population parameters

Parameters like the death rate, mutation rate and the selection factor can vary greatly from tumor to tumor. Therefore, several experimental simulations are carried out with tweaked parameters to investigate the different outcomes. By increasing the fraction of cells that dies at each generation ($f = 0.7$), the time for the population to reach its final size is prolonged (Figure 5.5A-B). This has two immediate consequences. Subclones have a lower chance of surviving to a size at which they can be detected and during simulation it was visible that many mutations that occur during growth go extinct by the time the tumor reaches its final size (results not shown). This might indicate that a lot of mutations that were present at a certain point during tumor development are not observable anymore at the time of diagnosis. On the other hand, a wider array of mutations can occur and subclones can evolve further because of the prolonged evolutionary dynamics. Especially when selection is applied, subclones with high fitness thrive even more and become dominant in the population. Indeed, it can be seen that in case of increased death rate, more subclones with high allele frequency are observed (Figure 5.5C-D). This implies that some subclones are predicted to prosper under an increased death rate by the model. A common problem in cancer therapy is the presence of subclonal mutations that confer resistance in a small fraction of the population, enabling these cells to evade destruction [37]. This increases the relative fitness of these subclones significantly, enabling them to outcompete other subclones. In addition, the overall death rate of the tumor will increase due to the change in environment when a treatment is applied. The simulations shown here suggest that the higher death rate could further increase the dominance of the resistant subclones, because those with lower fitness will be killed off at an increased rate and the prolonged dynamics gives the advantageous subclones more time to expand and take over the population.

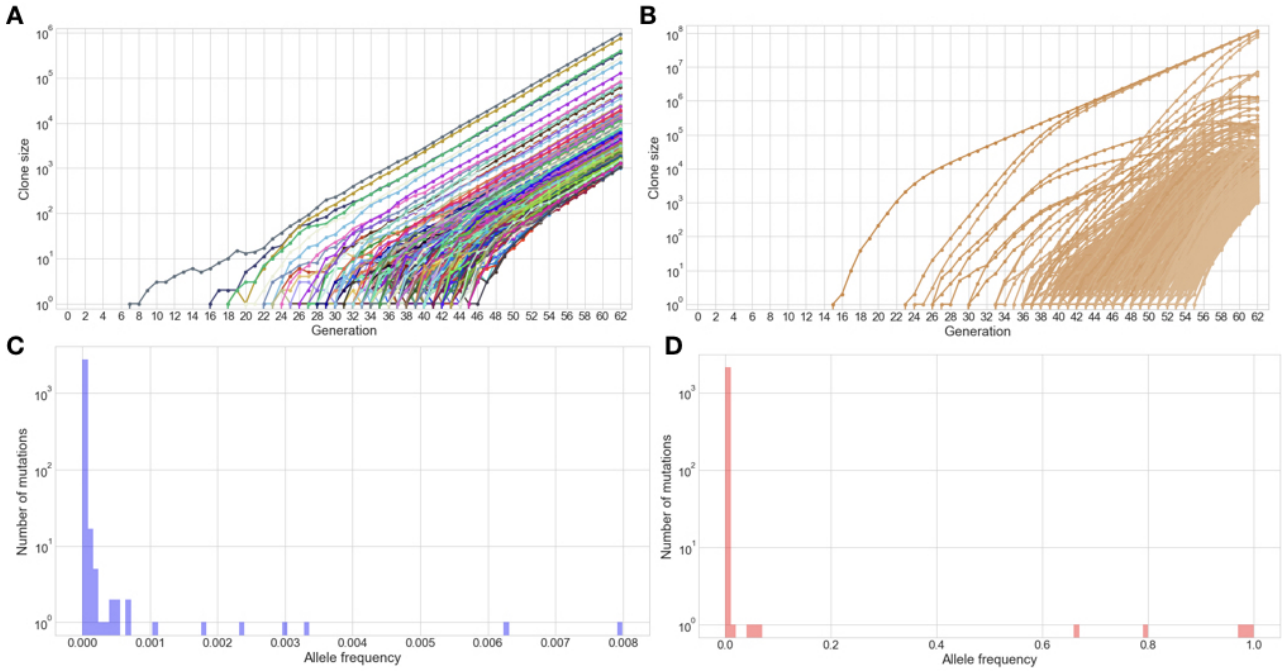

Figure 5.5. Populations grown under an increased death rate. Populations were grown to a minimal size of 10^8, with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.7$. In case of selection the selection factor was 0.3. Prior to analysis, the populations were sampled using the multinomial sampling method at 1% and a detection limit of 100 cells. (A-B) Evolution of subclone family sizes for neutral evolution and selective pressure, respectively. In B, it can be seen that the population is completely dominated by the light-brown subclone and its descendants. (C-D) Allele frequency spectra for neutral and selective populations, respectively. (E) Distribution of the number of accumulated mutations per subclone for both populations. The weighted means are given in the legend.

The standard mutation rate used for the simulations here is set to 10^{-3} per cell division for the functional genome, as this is a recurring value for common cancers such as colorectal and breast [29, 30]. However, simulations were also run with rates between 10^{-6} and 10^{-1} to explore the impact of the mutation rate on the final populations. As expected, higher mutation rates lead to an increased number of subclones and an increased number of mutations per subclone, both for neutral and selective evolution (Figure 5.6C). Strikingly, the effect of selection seems to be diminished under a high mutation rate. The observed allele frequencies are considerably lower (Figure 5.6B). In addition, the value for Simpson's diversity index was found to be virtually 1 for both the neutral and selective populations, indicating that the levels of heterogeneity are practically the same. This could imply that the differences between selection and neutral evolution disappear in case of high mutation rates. Consistently, another study that employed a Wright-Fisher-type model reported that increased selective advantages mainly reduced tumor development time in case of lower mutation rates, whereas for higher rates selection had a lesser effect [12].

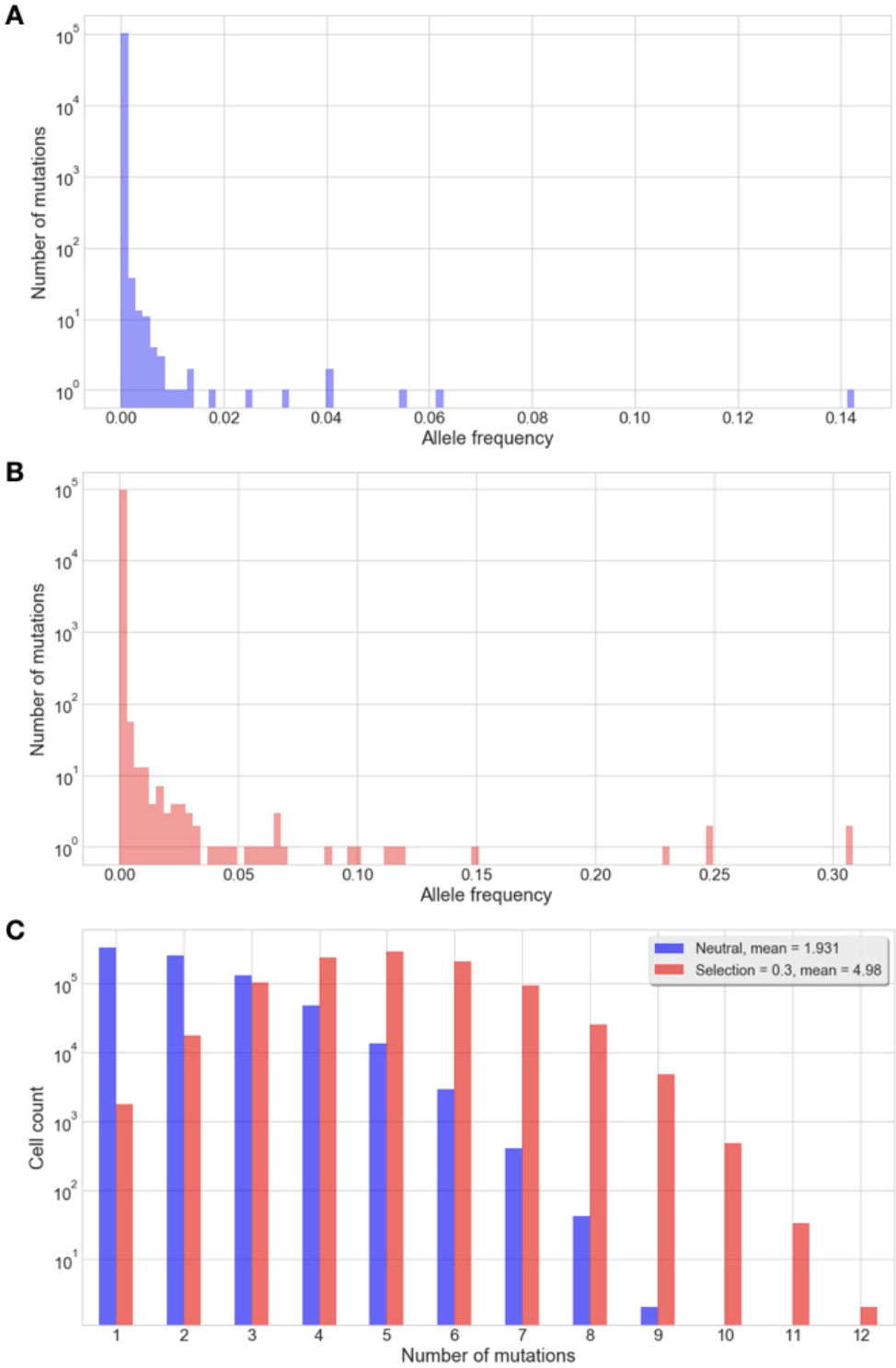

Figure 5.6. Populations grown under an increased mutation rate. Populations were grown to a minimal size of 10^8, with mutation rate $u = 10^{-1}$ and fraction of deaths $f = 0.3$. In case of selection the selection factor was 0.3. Prior to analysis, the populations were sampled using the multinomial sampling method at 1% and a detection limit of 100 cells. (A-B) Allele frequency spectra for neutral evolution and selective pressure, respectively. (C) Distribution of the number of accumulated mutations per subclone for both populations. The weighted means are given in the legend.

The influence of a mutator phenotype was also investigated, by adapting the model to allow the occurrence of mutator mutations. Such alterations increase the mutation rate of cancer cells by destabilizing genomic integrity [39]. They are implemented in this model by increasing the mutation rate of each new subclone relative to their parent clone with a constant factor. In this simplified model it is assumed now that every mutation increases the mutation rate. A value of 3 is used for the mutation rate-increasing factor. Thus, each new mutation will give rise to a subclone with a three-fold increased mutation rate relative to its parent. In the final population, subclones with mutation rates in the order of 10^{-1} can be found in case of selection, which is not uncommon in real tumors. For neutral evolution, the highest observed mutation rate is only 0.009, because the number of mutations per subclone is generally lower, as discussed before. Although this model of mutator mutations is clearly an oversimplification, the results provide a useful qualitative insight into their influence on the evolutionary dynamics. The mutator phenotype seems to have mostly an effect on the evolutionary dynamics in case of selection (Figure 5.7). The average number of mutations per subclone is higher than would be for a non-mutator phenotype (see also Figure 5.3C) and the heterogeneity according to Simpson's index is 0.48, an almost four-fold increase. The increased effect of mutator mutations under selective evolution could be explained by the fact that all mutations are considered to increase the mutation rate in this model. Thus, also those mutations that confer the highest fitness advantage will increase the mutation rate, leading to a more rapid acquisition of new driver mutations and an even higher fitness as well as increasing the mutation rate further. The notion that a mutator phenotype mainly plays a role under selective evolution was also found in previous work [12].

5.1.4 A method to reconstruct the mutational timeline

A population that grows in absence of any selective pressure can, in its simplest form, be considered to have constant growth rates and clonal fractions. This deterministic approach to model neutral evolution was employed by Williams *et al.* [9]. The authors exploited this assumption to reconstruct the mutational timeline of tumors that were shown to follow neutral evolution dynamics. The reconstruction method introduced by Williams *et al.* is further expanded here by deriving the points in time when each mutation occurred in units of doubling times, i.e. the number of times the population had doubled in size when the mutation occurred (see Section 4.2.5). It is important to note that this approach relies on the expected clone sizes under non-stochastic growth in the absence of selection. The method is tested on the simulated populations from the current model to assess the influence of selection and stochasticity in tumor evolution on the viability of the method. Using simulated populations has the advantage that the real timeline is exactly known and thus can be used to compare the reconstruction with and

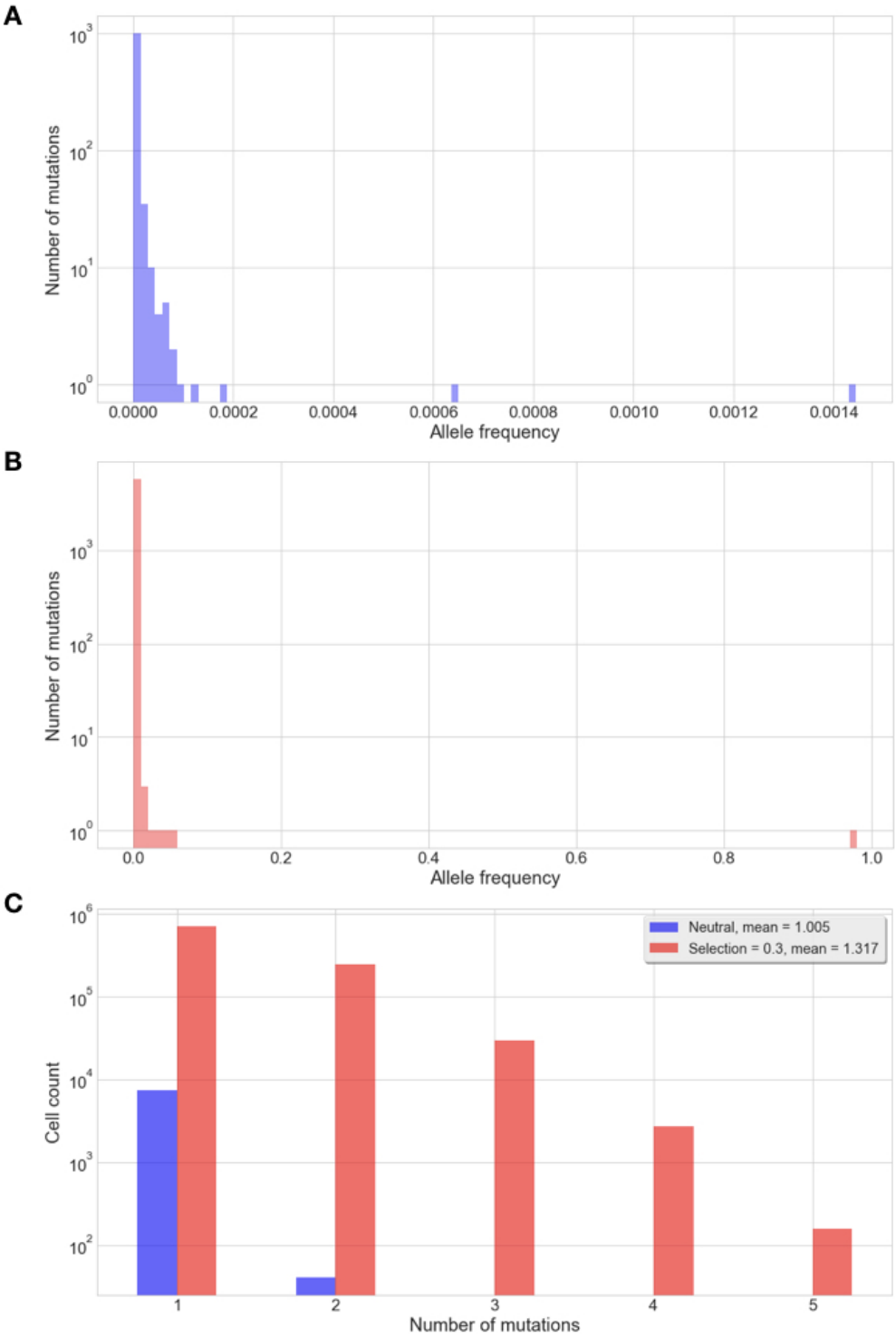

Figure 5.7. Influence of mutator phenotype. Populations were grown to a minimal size of 10^8, with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.3$. In case of selection the selection factor was 0.3. In addition, a mutation rate increasing factor of 3 was applied for each mutation. Prior to analysis, the populations were sampled using the multinomial sampling method at 1% and a detection limit of 100 cells. (A-B) Allele frequency spectra for neutral evolution and selective pressure, respectively. (C) Distribution of the number of accumulated mutations per subclone for both populations. The weighted means are given in the legend.

calculate errors. For a neutrally evolving population, the timeline can indeed be approximately reconstructed, with a median error of 1.03 (Figure 5.8A). In contrast, for a population that evolved under selective pressure, the median error is three times higher, at 3.09 (Figure 5.8B). In particular, some mutations are predicted to have occurred at the very start of tumor initiation, while in reality the first mutations occur after 10 population size doublings. This is caused by the fact that selection enables some mutations to become dominant in the population, which can result in an allele frequency of 1. As the method assumes that allele frequencies stay constant and estimates the population size as $1/\phi_i$ at the time the mutation occurred, these dominant mutations are predicted to have occurred when the population had a size of 1 cell, i.e. at initiation. This assumption is clearly not valid in case of selection. Therefore, this method could only be used on empirical data if it is known that the tumor evolved under mainly neutral evolution. Furthermore, it is interesting to note that in both cases, the mutations always seem to be predicted earlier than they actually appeared. The influence of selection on the errors in the timeline reconstruction is further investigated in Section 5.2.2.

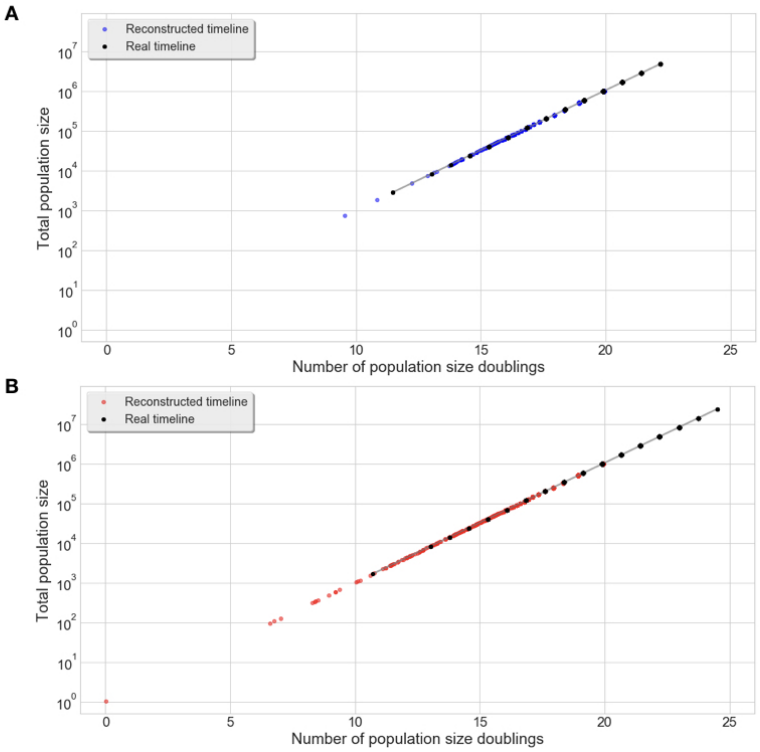

Figure 5.8. Reconstruction of mutational timelines. The mutational timelines were reconstructed for a neutrally evolved population (A) and a population grown under selective pressure (B). Populations were grown to a minimal size of 10^8, with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.3$. In case of selection the selection factor was 0.3. Each mutation event is depicted as a point in the graph, with the real timeline in black and the reconstruction in color.

5.2 The impact of selective pressure on heterogeneity

5.2.1 Multiple simulations under varying levels of selection

In order to assess the impact of selection on tumor evolution, 1000 simulations were run for the same set of parameters to monitor average behavior over multiple simulations. This was done for a range of selection factors from 0 up to 3. The higher values result in selective weights that are likely unrealistic in a biological context (Figure 5.1) but are given here as a theoretical demonstration. Each population is grown to a minimal size of 10^8, of which 1% is sampled (10^6 cells) with a detection limit of 100 cells. A constant mutation rate of 10^{-3} is used and the fraction of deaths at each generation (f) is set to 0.3. Different population characteristics are studied for comparison. To investigate whether any selective sweeps have taken place, whereby a specific subclone has outcompeted all others and taken over the population, the maximum observed allele frequency can be used. If such a dominant subclone is present, allele frequencies higher than 0.5 and up to 1 are expected. When considering the cumulative distribution function (CDF) of the maximum allele frequencies for all 1000 populations from a given set of simulations, the impact of selection becomes visible (Figure 5.9A). For neutrally evolving and low selection populations, allele frequencies above 0.1 are rare, while for selection factors above 0.3 the probability of observing dominant subclones increases drastically. For comparison, this analysis was also performed for 1000 simulations of populations that were grown to a final size of 10^4. Interestingly, the differences between the different levels of selection seem to diminish for smaller populations (Figure 5.9B).

The heterogeneity of the simulated populations is further assessed by calculating Simpson's diversity index for each of them. The variability in the index between simulations proves to be large, especially for higher selection factor values (Figure 5.10A). Neutral populations show a tendency towards high heterogeneity, but in case of selection the index values seem to be more evenly distributed between simulations. Thus, even with selection, populations can show considerable heterogeneity but in general it will be lower than for neutral evolution. These distributions could potentially be used to estimate the level of selection that has taken place during tumor development by comparing the observed heterogeneity from clinical data. A similar method has been used recently to detect neutral evolution in an extensively sampled tumor [8]. Their observation of a Simpson's diversity index value of 0.941 for a hepatocellular carcinoma is indeed more likely to be caused by neutral evolution according to the model presented here. Smaller populations (with final size 10^4) were also analyzed for comparison (Figure 5.10B). Again, the distributions for different selection factors are more similar.

Another key aspect of tumor evolution is the number of accumulated mutations carried by clones. After the first appearance of a cancer cell and initiation of the tumor, cells keep

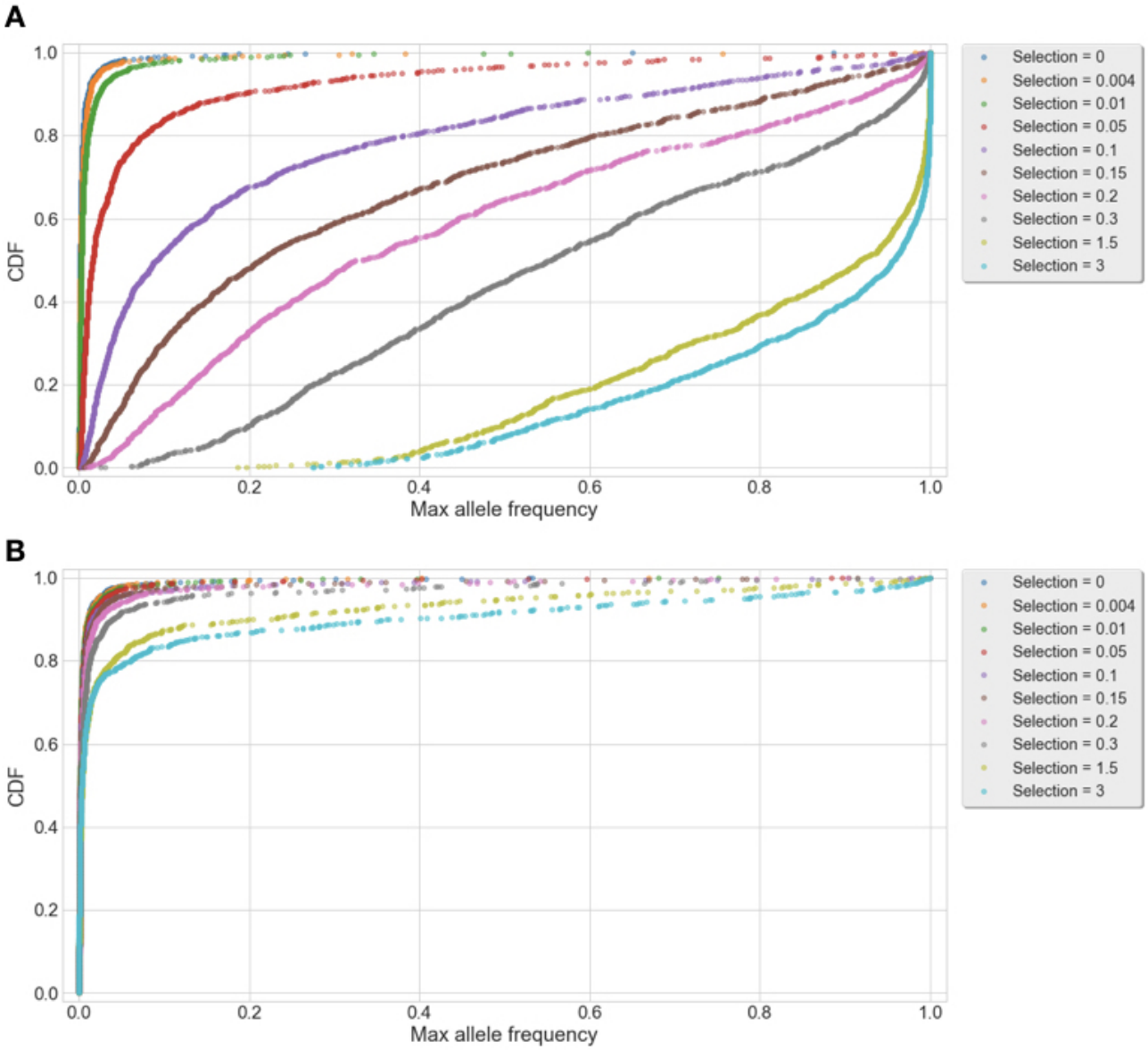

Figure 5.9. Cumulative distributions of maximum allele frequencies for different selection factors. For each level of selection, 1000 simulations were run for populations grown to a minimal size of 10^8 (A) or 10^4 (B), with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.3$. The maximum allele frequency from each population was obtained and the cumulative distribution function (CDF) for each level of selection was plotted.

accumulating mutations and thereby potentially develop new survival strategies. Each subclone will carry an additional unique mutation in respect to its parent clone under the infinite alleles model [67]. When considering that the ancestral clone has 0 mutations, it is possible to obtain the number of additional mutations each clone carries from a simulated population. By counting the number of cells in each clone and their associated number of mutations, a distribution is obtained that can be compared between different levels of selection (Figure 5.11A). The average number of mutations per subclone increases with higher selection, which is consistent with the

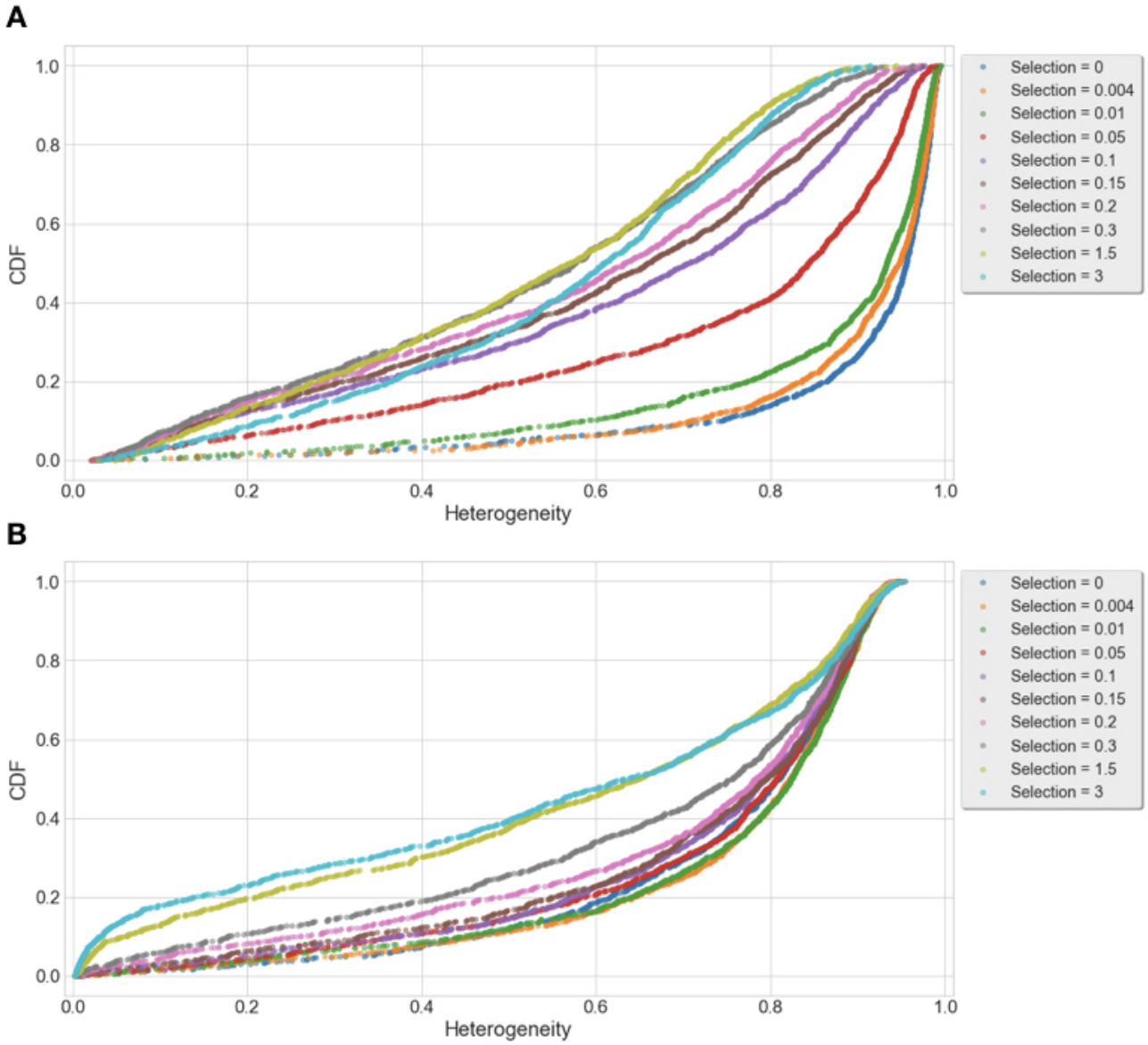

Figure 5.10. Cumulative distributions of heterogeneity for different selection factors. For each level of selection, 1000 simulations were run for populations grown to a minimal size of 10^8 (A) or 10^4 (B), with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.3$. The heterogeneity from each population was measured using Simpson's diversity index and the cumulative distribution function (CDF) for each level of selection was plotted.

deeper phylogenetic trees that were observed in Section 5.1.2. In contrast however, the total number of mutations actually decreases for higher selection (Figure 5.11B).

This is in line with the lower heterogeneity – and the fact that more mutations are shared between clones – associated with selective evolution. The increased number of mutations per clone can be explained by the fact that tumor development under selection is governed by selective sweeps brought about by clonal expansions [31]. Thus, every time a subclone emerges

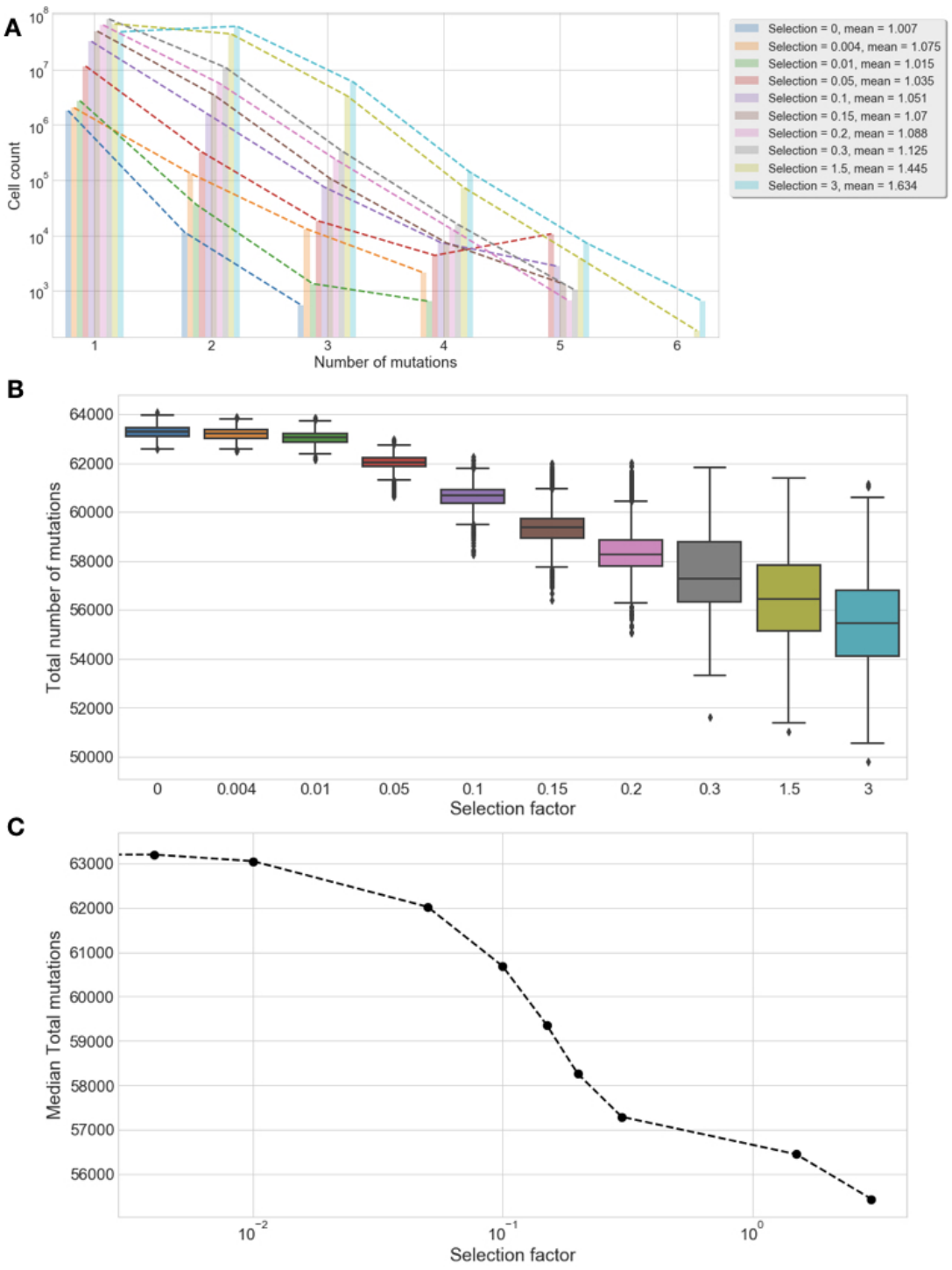

Figure 5.11. The distribution of the number of mutations for different selection factors. For each selection factor, 1000 simulations were run for populations grown to a minimal size of 10^8, with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.3$. (A) The number of accumulated mutations per subclone was obtained for each population and the averages of the 1000 populations per selection factor are displayed. The bar charts are connected by dashed lines to visualize the trends. The total weighted averages of mutations per clone are given in the legend. (B) The total number of mutations for each population was obtained and shown as a boxplot for each selection factor. (C) The medians from the total number of mutations are also plotted as a scatter plot in function of the selection factors, with the latter scaled logarithmically on the x-axis.

with significantly increased fitness, it will outcompete other clones and start to dominate the population, in turn giving rise to a new generation of subclones with additional mutations, and so on. Conversely, under neutral evolution the ancestral clone stays dominant, producing new subclones that each carry one additional mutation but rarely grow large enough to generate further clonal generations. What is interesting is that the number of total mutations remains relatively constant for neutral evolution and low selection factors and then starts dropping around selection factor 0.05 (Figure 5.11B). This could imply that there exists a threshold for the level of selection that needs to take place in order to produce a fundamentally different tumor population.

5.2.2 Timeline reconstruction errors for different levels of selection

As a final comparison between the different levels of selection, the errors in reconstructing the mutational timelines are studied (see also Section 5.1.4). For each simulated population, the errors are calculated as the absolute difference between the real and reconstructed time points of each mutation. The median of these errors is then taken to get the *median absolute deviation (MAD)* of the reconstruction errors for the whole population. The simulation for each selection factor is repeated 1000 times to produce statistically relevant results. The distribution of the MAD reconstruction errors can then be displayed using boxplots (Figure 5.12). As already suggested in Section 5.1.4, the errors in mutational timeline reconstruction indeed increase for higher selection factors. Yet, for the lower levels of selection (0.004 and 0.01), the errors seem to be comparable to those of neutral evolution. If real selection values are indeed that low, as some studies have suggested [11, 61, 73], it can be considered that even populations undergoing selective pressure can be reasonably approximated by models assuming neutral evolution. This implies that the timeline reconstruction method is still usable, even for populations undergoing selection. It would be interesting to see if the method can indeed be applied to empirical data and obtain an accurate prediction of the history of tumor development. However, this would require a dedicated experimental setup that allows validation of the predicted timeline and therefore lies beyond the scope of this thesis.

5.2.3 Realistic levels of selection resemble neutral evolution

It was proposed by Bozic *et al.* that driver mutations confer an increase of only 0.4% in growth rate. According to the model presented here, a population that grows under such a low level of selection shows almost identical behavior as for neutral evolution. In general, the results obtained in this study indicate that for populations with a size of up to 10^8 cells, there is no significant difference between neutrally evolving populations and those that grow

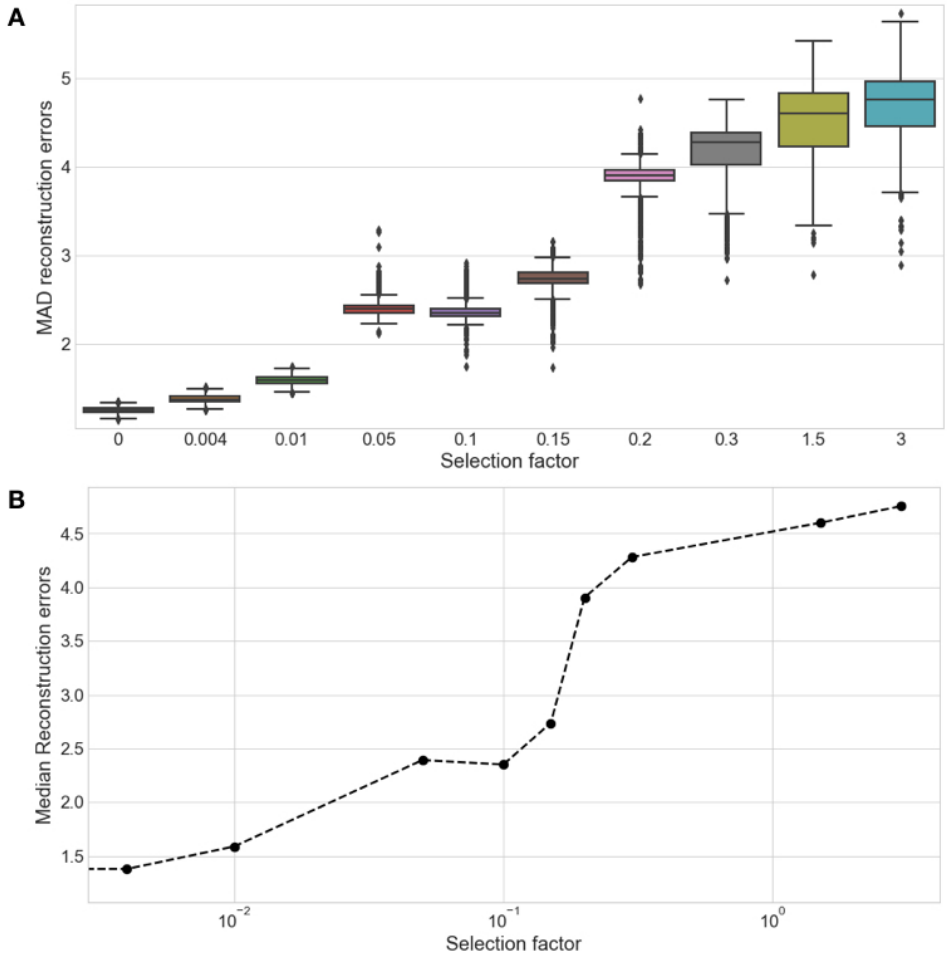

Figure 5.12. Errors in reconstruction of mutational timelines for different selection factors. For each level of selection, 1000 simulations were run for populations grown to a minimal size of 10^8, with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.3$. The mutational timelines were reconstructed for each population and the median absolute deviations (MAD) were calculated. (A) The MAD values for each population are plotted as boxplots for each selection factor. (B) The median MAD values are also plotted as a scatter plot in function of the selection factors, with the latter scaled logarithmically on the x-axis.

under realistic levels of selection. This also confirms the suggestion that neutral evolution and selection are similar, based on the small difference in R^2 values that was observed when fitting the cumulative number of mutations to the inverse allelic frequency (see Section 5.1.2 and Figure 5.4). In addition, it was shown for smaller populations that the various levels of selection display even smaller differences in evolutionary outcome (Figures 5.9B and 5.10B). Combined with the observation that higher mutation rates also diminish the effect of selection, as discussed in Section 5.1.3, this might provide a new context in which previous work can be evaluated. It should be noted, however, that the conclusions concerning the higher mutation rate are based on single simulations. No additional simulations were run for this parameter as this became computationally too demanding. The number of subclones that are generated

under a high mutation rate (10^{-1}) increases significantly, leading to a combinatorial explosion that poses memory problems, even for a single population of 10^8 cells.

Studies like those of Ling *et al.* and Williams *et al.* assumed the dominance of neutral evolution based on stochastic simulations [8, 9]. However, the population sizes from these studies are orders of magnitude lower than those used in this thesis (10^8) and the mutation rates higher. In addition, in Williams *et al.* no distinction is made between functional mutations and those that have no effect (synonymous), as was done here. Thus, according to the results obtained in the current study, it seems only logical that neutral evolution was detected in these previous studies, as selection has little impact at this scale. Furthermore, in the study of Williams *et al.* the theoretical predictions are compared with clinical data. However, in the analysis, the allele frequencies above 0.24 are discarded, because it is assumed that these are clonal mutations that were already present in the ancestral clone [9]. This filtering of the data is indeed necessary to ensure only subclonal mutations are kept from the sequencing data. However, the simulation results of the current study show that in case of selection, subclonal mutations can expand to sizes corresponding to allele frequencies far above the 0.24 threshold. This complicates analysis as one cannot be sure whether a high allele frequency mutation is ancestral or a subclonal mutation that has become dominant because of selection. By discarding this data, the allele frequency spectrum automatically gets a profile that resembles neutral evolution more closely. This shows the difficulty in determining the evolutionary dynamics from clinical data and the need for more advanced empirical methods and more discerning mathematical models.

5.3 Results are independent from the sampling method

In this last section, the influence of the sampling method on the analyses will be investigated. The multinomial sampling method is carried out with different sample sizes (10^5, 10^6 and 10^7) on two populations, one neutral and one selective (selection factor = 0.3), both grown to a final size of 10^8. Various analyses are carried out on the populations before and after sampling (Figure 5.13). No significant differences were observed between the sampled and original data, except for the expected reduction in number of cells and mutations. The profile of the allele frequency spectrum is retained, as well as the level of heterogeneity measured with Simpson's index. This shows that the random sampling of the data has no influence on the analysis of the population. In addition, this implies that when real tumors are properly sampled for sequencing, i.e. with enough randomness, the data can be considered to correctly represent the entire tumor population.

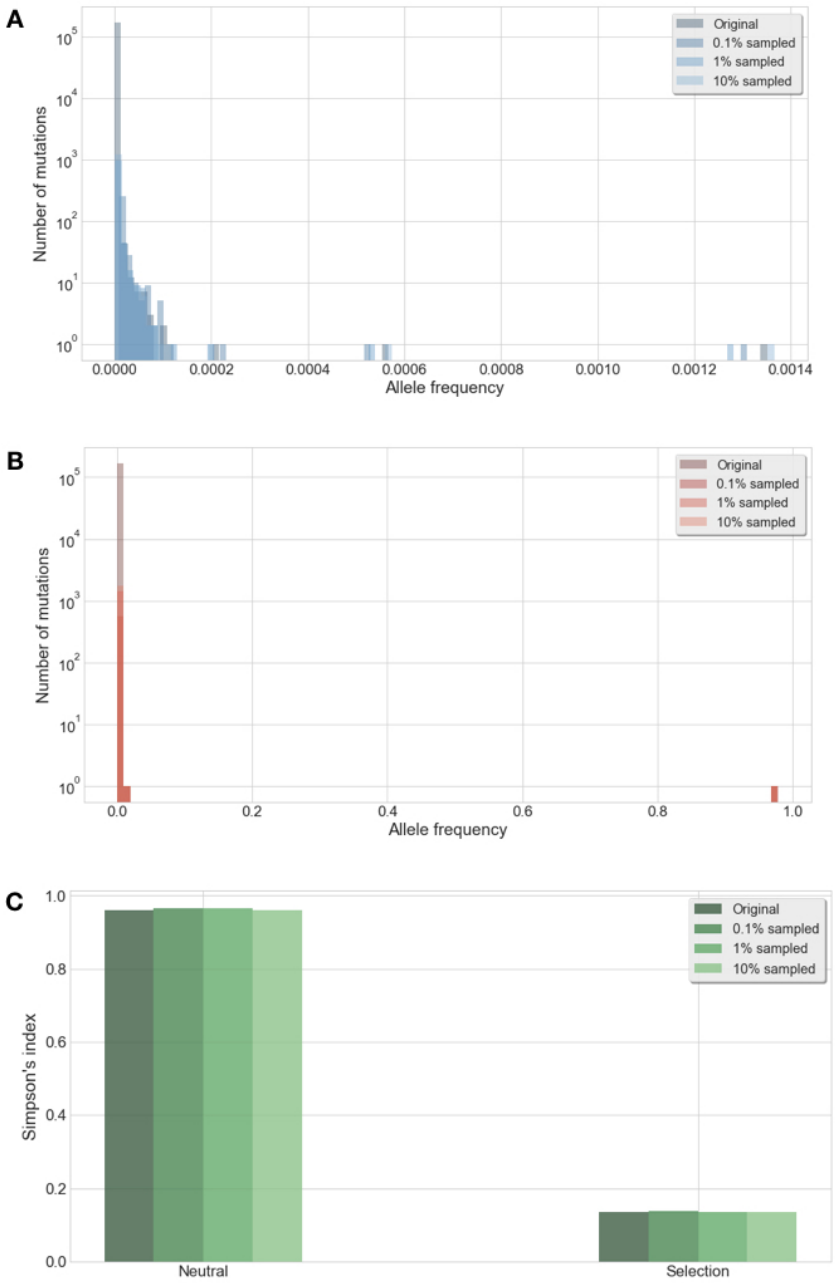

Figure 5.13. Comparative analysis between different levels of multinomial sampling. Two populations were grown to a minimal size of 10^8, with mutation rate $u = 10^{-3}$ and fraction of deaths $f = 0.3$. One was grown under neutral evolution, the other with a selection factor of 0.3. Each population was subjected to three sampling levels: 0.1%, 1% and 10%, each with a detection limit of 100 cells. (A-B) Allele frequency spectra for the neutrally evolved and selective populations, respectively. (C) Heterogeneity, measured with Simpson's index, for both populations.

Chapter 6

Conclusions and Future Perspectives

In this study, the genetic heterogeneity inherent to tumor populations was explored using mathematical modeling. A new model based on the Wright-Fisher process was created for this purpose and used to simulate both stochastic events and natural selection. This approach has the advantage that not all individual cell divisions need to be simulated separately. It can be assumed that this makes the model computationally more efficient, although no comparative analysis was performed with other models. This way, a general mathematically tractable model for tumor evolution was constructed.

Individual simulations were run to provide an overview of the general model behavior and show the main differences between populations governed solely by neutral evolution and those that are subject to a high level of selection. It was observed that neutral evolution leads to a more heterogeneous population with lower allele frequencies and mutation depth (the average number of mutations per subclone). In addition, the influence of different population parameters was investigated. A higher death rate was found to exacerbate the effect of selection and it was observed that the population was even more dominated by a single subclone and its descendants. On the other hand, a higher mutation rate diminished the effect of selection and produced populations similar to those under neutral evolution, in terms of subclonal fractions. The mutator phenotype (mutations that increase the mutation rate) was also simulated and the results suggest that this phenomenon mainly plays a role in populations under influence of selection. However, this process could be modeled more realistically by discerning between mutations that confer an altered mutation rate and those that do not. In general, these results show that the model produces similar predictions as those found in the literature.

The impact of selective pressure on the evolutionary outcome was investigated in more detail by running ensembles of multiple simulations with the same parameters to obtain statistically relevant results. It was found that populations that are subject to the low levels of selection

reported in previous studies actually resemble neutral evolution. This indicates that selection mainly plays a role for populations that have grown long enough to allow advantageous subclones to expand and become dominant. In contrast, there is no significant difference between selection and neutral evolution for populations with a size under at least 10^8 cells, according to the results presented here, which could explain why neutral evolution was observed in the studies of Ling *et al.* and Williams *et al.* [8, 9]. This does not necessarily contradict their conclusions or those of the Big Bang model of Sottoriva *et al.* – where it is argued that selection only plays a role early in tumor development, followed by neutral evolution [7] – but rather suggests that in order for the effects of selection to become observable, the tumor needs to have grown a sufficient amount of time to a considerable size. One important consequence of this is that the method to reconstruct the mutational timeline of a tumor (see Section 5.1.4) is still usable for populations that undergo selection, even though the method assumes a simplistic growth model that does not incorporate selection. Indeed, it was shown with a multitude of simulations that the errors in timeline reconstruction were comparably low for neutrally evolving populations and those under realistic levels of selection.

The results of this thesis provide a quantitative analysis of the proposed model, although they are mainly theoretical. Due to a lack of suitable datasets and time, no clinical data analysis was successfully performed to compare the model with. Ultimately, this would be essential to validate the predictions made by the model and is therefore encouraged for future work. In addition, a more analytical analysis of the model could be carried out, which should be possible because of the underlying simplicity. An improvement to the model would be to incorporate selective weights in the probability distribution of the stochastic deaths as well. This would result in differing death rates for each clone in addition to the altered division rates and would act as an additional selective force that is also expected to be present in nature. However, this is not straightforward to implement in the hypergeometric distribution that is used here for the stochastic deaths. Therefore, other methods could be explored to model this process. Finally, spatial and environmental effects could be added to the model to achieve a closer approximation of reality, because cancers are not independent entities but part of a complex ecosystem involving many types of cells. However, it should be considered that this might make the model overly complicated, losing mathematical tractability.

Many aspects of the evolutionary dynamics underlying tumor development remain unclear. Mathematical models provide a theoretical framework to quantitatively assess experimental data and are essential in cancer research. Ultimately, a combination of more sophisticated computational models and accurate experimental methods could provide the key to fully understanding cancer and, subsequently, the development of improved therapeutic strategies and personalized medicine.

Appendix A

Mathematical proofs

A.1 Probability distribution for the number of dividers

Proof for the equivalence of equations (4.1) and (4.5) from section 4.1.1. If a population of three clones with size $N = n_1 + n_2 + n_3$ is considered, the probability distribution of dividers (p_1, p_2, p_3) for the whole population according to equation (4.1) is given by:

$$P\{p_1,\ p_2,\ p_3\} = \frac{N!}{p_1!\ p_2!\ p_3!}\left(\frac{n_1(t)}{N}\right)^{p_1}\left(\frac{n_2(t)}{N}\right)^{p_2}\left(\frac{n_3(t)}{N}\right)^{p_3} \tag{A.1}$$

The selection of dividers for the first two clones separately can be expressed as (equations (4.2) and (4.3)):

$$P\{p_1\} = \binom{N}{p_1}\left(\frac{n_1(t)}{N}\right)^{p_1}\left(1 - \frac{n_1(t)}{N}\right)^{N-p_1} \tag{A.2}$$

$$P\{p_2 \mid p_1\} = \binom{N-p_1}{p_2}\left(\frac{n_2(t)}{N-p_1}\right)^{p_2}\left(1 - \frac{n_2(t)}{N-p_1}\right)^{N-p_1-p_2} \tag{A.3}$$

Because the distributions of dividers for the first two clones have been determined, p_3 for the third clone will be fixed, as $N = p_1+p_2+p_3$. Therefore, the distribution for the whole population is known when the distribution for p_1 and p_2 is known, which in turn can be expressed according to the law of *conditional probability* as:

$$P\{p_1,\ p_2\} = P\{p_1\} \cdot P\{p_2 \mid p_1\} \tag{A.4}$$

When writing out the full expression and considering $N = p_1 + p_2 + p_3$, equation (A.4) can be transformed to:

$$
\begin{aligned}
&P\{p_1\} \cdot P\{p_2 \mid p_1\} \\
&= \frac{N!}{p_1!(N-p_1)!} \left(\frac{n_1(t)}{N}\right)^{p_1} \left(1 - \frac{n_1(t)}{N}\right)^{N-p_1} \\
&\quad \cdot \frac{(N-p_1)!}{p_2!(N-p_1-p_2)!} \left(\frac{n_2(t)}{N-n_1(t)}\right)^{p_2} \left(1 - \frac{n_2(t)}{N-n_1(t)}\right)^{N-p_1-p_2} \\
&= \frac{N!}{p_1!\, p_2!\, p_3!} \left(\frac{n_1}{N}\right)^{p_1} \left(\frac{N-n_1}{N}\right)^{N-p_1} \left(\frac{n_2}{N-n_1}\right)^{N-p_1-p_3} \left(\frac{N-n_1-n_2}{N-n_1}\right)^{p_3} \\
&= \frac{N!}{p_1!\, p_2!\, p_3!} \left(\frac{n_1}{N}\right)^{p_1} \left(\frac{N-n_1}{N} \frac{n_2}{N-n_1}\right)^{N-p_1} \left(\frac{n_2}{N-n_1}\right)^{-p_3} \left(\frac{n_3}{N-n_1}\right)^{p_3} \\
&= \frac{N!}{p_1!\, p_2!\, p_3!} \left(\frac{n_1}{N}\right)^{p_1} \left(\frac{n_2}{N}\right)^{N-p_1} \left(\frac{n_3}{n_2}\right)^{p_3} \\
&= \frac{N!}{p_1!\, p_2!\, p_3!} \left(\frac{n_1}{N}\right)^{p_1} \left(\frac{n_2}{N}\right)^{p_2+p_3} \left(\frac{n_3}{n_2}\right)^{p_3} \\
&= \frac{N!}{p_1!\, p_2!\, p_3!} \left(\frac{n_1}{N}\right)^{p_1} \left(\frac{n_2}{N}\right)^{p_2} \left(\frac{n_3}{N}\right)^{p_3}
\end{aligned}
$$

Finally, equation (A.1) is found. The equivalence of equations (4.1) and (4.5) thus holds true for the case of three clones. By induction and because of the property $N = \sum_{i=1}^{K} p_i$, this proof can be extended to any number of clones K.

A.2 Probability distribution for the number of deaths

Proof for the equivalence of equations (4.11) and (4.15) from section 4.1.3. Again a population of three clones is considered with size $N = n_1 + n_2 + n_3$ and death rate f. The joint probability distribution of deaths (d_1, d_2, d_3) for the whole population according to equation (4.11) is given by:

$$
P\{d_1,\ d_2,\ d_3\} = \frac{\binom{n_1(t)}{d_1}\binom{n_2(t)}{d_2}\binom{n_3(t)}{d_3}}{\binom{N}{fN}} \tag{A.5}
$$

The probability distributions of deaths for the first two clones separately can be expressed by (equations (4.12) and (4.13)):

$$P\{d_1\} = \frac{\binom{n_1(t)}{d_1}\binom{N-n_1(t)}{fN-d_1}}{\binom{N}{fN}} \tag{A.6}$$

$$P\{d_2 \mid d_1\} = \frac{\binom{n_2(t)}{d_2}\binom{N-n_1(t)-n_2(t)}{fN-d_1-d_2}}{\binom{N-n_1(t)}{fN-d_1}} \tag{A.7}$$

The total number of deaths in the population is defined as $fN = d_1 + d_2 + d_3$ and therefore d_3 is fixed when d_1 and d_2 have been chosen. The joint probability distribution of deaths for a population of three clones can thus be described by considering the separate distributions of two clones. The latter can be expressed by the law of *conditional probability* as:

$$P\{d_1,\ d_2\} = P\{d_1\} \cdot P\{d_2 \mid d_1\} \tag{A.8}$$

The right-hand side of this equation can be transformed as follows, using the properties $fN = d_1 + d_2 + d_3$ and $N = n_1(t) + n_2(t) + n_3(t)$:

$$\begin{aligned} P\{d_1\} \cdot P\{d_2 \mid d_1\} &= \frac{\binom{n_1(t)}{d_1}\binom{N-n_1(t)}{fN-d_1}}{\binom{N}{fN}} \cdot \frac{\binom{n_2(t)}{d_2}\binom{N-n_1(t)-n_2(t)}{fN-d_1-d_2}}{\binom{N-n_1(t)}{fN-d_1}} \\ &= \frac{\binom{n_1(t)}{d_1}\binom{n_2(t)}{d_2}\binom{n_3(t)}{d_3}}{\binom{N}{fN}} \end{aligned}$$

This again leads to equation (A.5), proving the equivalence for the case of a population with three clones. By induction this can be extended to any number of clones K, as long as the condition $fN = \sum_{i=1}^{K} d_i$ is kept true.

Bibliography

[1] Arrowsmith, J. and Miller, P. Trial watch: phase II and phase III attrition rates 2011-2012. *Nature Reviews Drug Discovery*, 12(8):569–569, 2013.

[2] Unger, C. et al. Modeling human carcinomas: physiologically relevant 3d models to improve anti-cancer drug development. *Advanced Drug Delivery Reviews*, 79:50–67, 2014.

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

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

[5] Burrell, R.A. et al. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338, 2013.

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

[7] Sottoriva, A. et al. A big bang model of human colorectal tumor growth. *Nature Genetics*, 47(3):209, 2015.

[8] Ling, S. et al. Extremely high genetic diversity in a single tumor points to prevalence of non-darwinian cell evolution. *Proceedings of the National Academy of Sciences*, 112(47): E6496–E6505, 2015.

[9] Williams, M.J. et al. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48(3):238, 2016.

[10] Ewens, W.J. *Mathematical population genetics*, volume 27 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, NY, USA, 2nd edition, 2004.

[11] Beerenwinkel, N. et al. Genetic progression and the waiting time to cancer. *PLoS Computational Biology*, 3(11):e225, 2007.

[12] Datta, R.S. et al. Modelling the evolution of genetic instability during tumour progression. *Evolutionary Applications*, 6(1):20–33, 2013.

[13] Durrett, R. et al. Intratumor heterogeneity in evolutionary models of tumor progression. *Genetics*, 188(2):461–477, 2011.

[14] Weinberg, R. *The biology of cancer*. Garland Science, New York, NY, USA, second edition, 2013.

[15] Reece, J.B. et al. *Campbell biology*. Pearson, San Francisco, CA, USA, 9th edition, 2011.

[16] Blair, G. and Cook, G. Cancer and the immune system: an overview. *Oncogene*, 27(45): 5868, 2008.

[17] Hanahan, D. and Weinberg, R.A. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

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

[19] Greenman, C. et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446 (7132):153, 2007.

[20] Foulds, L. The experimental study of tumor progression: a review. *Cancer Research*, 14 (5):327–339, 1954.

[21] Tabin, C.J. et al. Mechanism of activation of a human oncogene. *Nature*, 300(5888):143, 1982.

[22] Bos, J.L. *ras* oncogenes in human cancer: a review. *Cancer Research*, 49(17):4682–4689, 1989.

[23] Harris, C.C. *p53* tumor suppressor gene: from the basic research laboratory to the clinic-an abridged historical perspective. *Carcinogenesis*, 17(6):1187–1198, 1996.

[24] Forbes, S.A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, 2017.

[25] Touw, I.P. and Erkeland, S.J. Retroviral insertion mutagenesis in mice as a comparative oncogenomics tool to identify disease genes in human leukemia. *Molecular Therapy*, 15(1): 13–19, 2007.

[26] Balmain, A. et al. How many mutations are required for tumorigenesis? implications from human cancer data. *Molecular Carcinogenesis*, 7(3):139–146, 1993.

[27] Schinzel, A.C. and Hahn, W.C. Oncogenic transformation and experimental models of human cancer. *Frontiers in Bioscience: A Journal and Virtual Library*, 13:71–84, 2008.

[28] Stephens, P.J. et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400, 2012.

[29] Jones, S. et al. Comparative lesion sequencing provides insights into tumor evolution. *Proceedings of the National Academy of Sciences*, 105(11):4283–4288, 2008.

[30] Wang, Y. et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155, 2014.

[31] Greaves, M. and Maley, C.C. Clonal evolution in cancer. *Nature*, 481(7381):306, 2012.

[32] Anderson, K. et al. Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature*, 469(7330):356, 2011.

[33] Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341): 90, 2011.

[34] Beerenwinkel, N., Greenman, C.D. and Lagergren, J. Computational cancer biology: an evolutionary perspective. *PLoS Computational Biology*, 12(2):e1004717, 2016.

[35] Roche-Lestienne, C. et al. Several types of mutations of the abl gene can be found in chronic myeloid leukemia patients resistant to sti571, and they can pre-exist to the onset of treatment. *Blood*, 100(3):1014–1018, 2002.

[36] Mullighan, C.G. et al. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*, 322(5906):1377–1380, 2008.

[37] Gilbert, L.A. and Hemann, M.T. DNA damage-mediated induction of a chemoresistant niche. *Cell*, 143(3):355–366, 2010.

[38] Loeb, L.A. A mutator phenotype in cancer. *Cancer Research*, 61(8):3230–3239, 2001.

[39] Loeb, L.A. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature Reviews Cancer*, 11(6):450, 2011.

[40] Leedham, S.J. et al. Individual crypt genetic heterogeneity and the origin of metaplastic glandular epithelium in human barrett's oesophagus. *Gut*, 57(8):1041–1048, 2008.

[41] Navin, N. et al. Inferring tumor progression from genomic heterogeneity. *Genome Research*, 20(1):68–80, 2010.

[42] Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, 366(10):883–892, 2012.

[43] Sottoriva, A. et al. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, 110(10):4009–4014, 2013.

[44] Lande, R. Natural selection and random genetic drift in phenotypic evolution. *Evolution*, 30(2):314–334, 1976.

[45] Durrett, R. Population genetics of neutral mutations in exponentially growing cancer cell populations. *The Annals of Applied Probability: An Official Journal of the Institute of Mathematical Statistics*, 23(1):230, 2013.

[46] Reya, T. et al. Stem cells, cancer, and cancer stem cells. *Nature*, 414(6859):105, 2001.

[47] Beck, B. and Blanpain, C. Unravelling cancer stem cell potential. *Nature Reviews Cancer*, 13(10):727, 2013.

[48] Driessens, G. et al. Defining the mode of tumour growth by clonal analysis. *Nature*, 488 (7412):527, 2012.

[49] Basanta, D. and Anderson, A.R. Exploiting ecological principles to better understand cancer progression and treatment. *Interface Focus*, 3(4):20130020, 2013.

[50] Vogelstein, B. et al. Cancer genome landscapes. *Science*, 339(6127):1546–1558, 2013.

[51] Nordling, C. A new theory on the cancer-inducing mechanism. *British Journal of Cancer*, 7(1):68, 1953.

[52] Scholz, M.B., Lo, C.C. and Chain, P.S. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current Opinion in Biotechnology*, 23(1):9–15, 2012.

[53] Altrock, P.M., Liu, L.L. and Michor, F. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, 15(12):730, 2015.

[54] Roeder, I. et al. Dynamic modeling of imatinib-treated chronic myeloid leukemia: functional insights and clinical implications. *Nature Medicine*, 12(10):1181, 2006.

[55] Werner, B., Dingli, D. and Traulsen, A. A deterministic model for the occurrence and dynamics of multiple mutations in hierarchically organized tissues. *Journal of The Royal Society Interface*, 10(85):20130349, 2013.

[56] Neyman, J. and Scott, E.L. Stochastic models of population dynamics. *Science*, 130(3371): 303–308, 1959.

[57] Durrett, R. *Branching process models of cancer*. Springer, Cham, Switzerland, 2015.

[58] McDougall, S.R., Anderson, A.R. and Chaplain, M.A. Mathematical modelling of dynamic adaptive tumour-induced angiogenesis: clinical implications and therapeutic targeting strategies. *Journal of Theoretical Biology*, 241(3):564–589, 2006.

[59] Araujo, A. et al. An integrated computational model of the bone microenvironment in bone-metastatic prostate cancer. *Cancer Research*, 74(9):2391–2401, 2014.

[60] Parzen, E. *Stochastic processes*, volume 24 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, USA, 1999.

[61] Bozic, I. et al. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–18550, 2010.

[62] Moran, P.A.P. Random processes in genetics. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 54, pages 60–71. Cambridge University Press, 1958.

[63] Foo, J., Leder, K. and Michor, F. Stochastic dynamics of cancer initiation. *Physical Biology*, 8(1):015002, 2011.

[64] Kimura, M. Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences*, 41(3):144–150, 1955.

[65] Beerenwinkel, N. et al. Cancer evolution: mathematical models and computational inference. *Systematic Biology*, 64(1):e1–e25, 2014.

[66] Niida, A., Iwasaki, W.M. and Innan, H. Neutral theory in cancer cell population genetics. *Molecular Biology and Evolution*, 2018.

[67] Durrett, R. *Probability models for DNA sequence evolution.* Springer Science & Business Media, New York, NY, USA, 2nd edition, 2008.

[68] Simpson, E.H. Measurement of diversity. *Nature*, 163(4148):688, 1949.

[69] Del Monte, U. Does the cell number 10^9 still really fit one gram of tumor tissue? *Cell Cycle*, 8(3):505–506, 2009.

[70] Mehrara, E. et al. Specific growth rate versus doubling time for quantitative characterization of tumor growth rate. *Cancer research*, 67(8):3970–3975, 2007.

[71] Almendro, V. et al. Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell Reports*, 6 (3):514–527, 2014.

[72] Park, S.Y. et al. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *The Journal of Clinical Investigation*, 120(2):636–644, 2010.

[73] Siegmund, K.D. et al. High DNA methylation pattern intratumoral diversity implies weak selection in many human colorectal cancers. *PloS ONE*, 6(6):e21657, 2011.