NTNU
Norwegian University of
Science and Technology

Stochastic modelling and analysis of neutral tumor evolution.

Can neutral tumor evolution be inferred from real data?

Pål Vegard Johnsen

Master of Science in Physics and Mathematics
Submission date: June 2018
Supervisor: Mette Langaas, IMF
Co-supervisor: Thea Bjørnland, IMF

Norwegian University of Science and Technology
Department of Mathematical Sciences

Preface

This master's thesis is completed as part of the course TMA4900 at The Department of Mathematical Sciences at The Norwegian University of Science and Technology (NTNU), in June 2018.

The aim of this thesis is to analyze a stochastic model for tumor evolution. The thesis is written in such a way that it is expected that the interested reader has some basic knowledge of cell biology as there are many definitions that will be presented. Anyhow, the definition list on page 64 will hopefully simplify the reading.

The topic was suggested by my supervisor Professor Mette Langaas and Co-Supervisor Thea Bjørnland as a result of recent research in this field, specifically in Williams et al. (2016). The everlasting importance of understanding the biological phenomenon of cancer, and the substantial ripple effects new knowledge can serve the entire world, made this topic extremely motivating for me.

I would like to express my deep gratitude to Professor Mette Langaas and PhD student Thea Bjørnland for your great guidance, support, motivation and curiosity during the whole process. It has been a pleasure to work with both of you.

Sammendrag

Vi utvikler en statistisk forgreiningsprosess med akkumulerende mutasjoner til å modellere nøytral tumorevolusjon. Ved å bruke denne modellen vil det vises i detalj hvilke approksimasjoner og forenklinger som er nødvendig for å utlede det samme uttrykket som i Williams et al. (2016), som karakteriserer en nøytral tumorevolusjon. Stokastiske simuleringer vil aktivt bli brukt for å validere approksimasjonene med hensyn på nøyaktighet.

Modellen vil så ta hensyn til DNA-sekvensiering, og en metode for å få forståelsen av unøyaktighetene av observerte allelfrekvenser vil bli utledet. Til slutt vil det bli argumentert for at det fremdeles er for mye usikkerhet til å utvikle en statistisk hypotesetest som evaluerer hvorvidt en tumor har utviklet seg nøytralt eller ikke.

Abstract

We develop a branching process with accumulating mutations to model neutral tumor evolution. Furthermore, by using this model, it will be shown in detail which approximations and simplifications that are necessary in order to deduce the same expression as in Williams et al. (2016) characterizing a neutral tumor evolution. Stochastic simulations will actively be used in order to validate approximations with respect to accuracy.

The model will then take DNA sequencing into account, and a method for incorporating the inaccuracies of observed variant allele frequencies will be developed.

Lastly, it will be argued that there are yet too much uncertainties in order to infer, using statistical hypothesis tests, whether a neutral tumor evolution actually is a good description of evolution for certain tumors.

Contents

1 Introduction

A tumor is an abnormal mass or tissue that arises as a result of uncontrolled growth of cells. It is thought that the uncontrolled growth of the tumor is initiated by a multistep process involving the occurrence of *somatic mutations*, which then can give a cell *selective advantages* in terms of larger probability of survival and reproduction than surrounding cells (Cooper 2000).

Inferring the evolution of a detected tumor is valuable with respect to prognosis (Williams et al. 2016). However, tumor evolution has shown to be difficult to characterize. One reason for this is that longitudinal measurements after detection of a tumor can be impractical or even unethical (Davis et al. 2017). Therefore, the evolutionary process of the tumor must often be inferred based on the characteristics of the tumor when it is first detected.
Another important reason why tumor evolution is difficult to characterize is *intra-tumor heterogeneity* (ITH) (Gerashchenko et al. 2013). This means, among other things, that within a tumor there is a large variation in genetic material in terms of somatic mutations among different tumor cells. In other words, there are many subsets of tumor cells in the tumor, often denoted as *subclones*, that share specific somatic mutations that other tumor cells do not possess.

In search of the causes of ITH, this has led to much discussion. A major theory is *clonal evolution* as first proposed in Nowell (1976), as a result of an ongoing acquisition of somatic mutations providing cells with new selective advantages during evolution. This can then lead to mutations that give cells selective advantages or, said in another word, increased *fitness*. These mutations are then called *driver mutations*, creating subclones where each subclone consist of cells that share a specific driver mutation. Clonal evolution can then be regarded to follow the theory of Darwinian natural selection, where the cells most adapted to the environment are most likely to survive and reproduce.

However, new research by Sottoriva et al. (2015) showed that the pattern of ITH may not necessarily need such complex causes as mentioned above. Sottoriva et al. (2015) showed that often, colorectal cancers predominantly grow as a single expansion consisting of a large number of intermixed subclones. From this observation, one can reason that the pattern of ITH is rather as a result of *passenger mutations* arising during the evolution. Passenger mutation does not alter the fitness of the cells. This was investigated further in Williams et al. (2016) leading to a *neutral tumor evolution* model, which we define as the following:

Definition 1.1 (**Neutral tumor evolution**)**.** In neutral tumor evolution all tumor cells have the same fitness. All mutations occurring after tumor initiation are therefore *passenger mutations.*

Using this definition, Williams et al. (2016) established a model and deduced an expression that characterizes a neutral tumor evolution as a function of the mutation- and growth rate, or as they call it: *The mutation rate per effective cell division.* The model is based on the observed value of the *variant allele frequency* (VAF) which in general can be defined as the following:

Definition 1.2 (**Variant allele frequency (VAF)**)**.** Given a sample of cells and a particular *allele* i positioned at a specific *locus.* Let S_i denote the number of times allele i appears in the sample, and let L denote the total number of copies of the locus in the sample. The *variant allele frequency*, VAF_i, for allele i is then given by

$$\text{VAF}_i = \frac{S_i}{L}. \tag{1}$$

Somatic mutations in the tumor can be detected by using *DNA sequencing* technology, and furthermore the corresponding VAFs can be estimated. In Williams et al.

(2016), the mutation rate per effective cell division is estimated from observed data based on the deduced expression characterizing neutral tumor evolution.

1.1 The aim of this thesis

The neutral tumor evolution model is a typical scientific example following Occam's razor (Kapp 1958), in being a model consisting of as few assumptions or parameters as possible in order to explain a phenomenon. The model of neutral tumor evolution in Williams et al. (2016) is based on a constant mutation rate, a constant growth rate, and an average number of chromosome sets equal to the ploidy π in each cancer cell. Furthermore, considering a continuous-time model, the average number of somatic mutations having VAF $\geq f$ in a neutrally evolving tumor, was shown to be proportional to $1/f$.

In this thesis, a discrete-time *branching process* with accumulating passenger mutations is developed. The primary aim of this thesis is to show in details how a similar expression as that given in Williams et al. (2016) can be deduced, denoted as $\hat{M}_N(f)$ in this thesis, using this branching process model. In this deduction, it will become clear which approximations are made, and each approximation will be validated using stochastic simulations.

As mentioned earlier, somatic mutations can be observed and their respective VAFs can be estimated using DNA sequencing. A statistical model for DNA sequencing similar to the one given in Sun et al. (2017) is developed. Using this model, observed data of VAFs are simulated for a neutrally evolving tumor. The simulated observed data are then compared with the theoretical expression deduced in Williams et al. (2016). In addition to this, the inaccuracies of the observed VAFs due to DNA-sequencing will be discussed.

In the end, the validity of using the theoretical results deduced in Williams et al. (2016) and in this thesis to infer if a tumor evolved neutrally, will be discussed.
In summary, the thesis will consist of the following tasks:

1. Develop a branching process with accumulating mutations to model neutral tumor evolution, and use this model to deduce the same expression as given in Williams et al. (2016). In addition the result will be validated using stochastic simulations.

2. Develop a statistical model for DNA sequencing, and compare simulated observed data from a neutrally evolving tumor with power law-distribution.

3. Discuss the validity of using our theoretical results to infer neutral tumor evolution from real data.

2 Statistical theory

The following statistical theory will be applied during this thesis. The notation and statistical theory is inspired by Ross (2014), Karr (1993) and Casella & Berger (2002).

2.1 Stochastic process

Definition 2.1. A *stochastic process*, $X(t), t \in T$, is a collection of stochastic variables. For each index $t \in T$, $X(t)$ is a stochastic variable.

Usually, t is interpreted as time, hence the notion of a *process*, but one may also think of it as a position in space. In general $X(t)$ denotes the *state* at time t, while $x(t)$ is the observation at time t. The process can be in a finite or infinite number of states given by the domain Ω. From the definition of a state, the domain Ω can also be called a *state space*. The stochastic process is said to be *discrete* if T is a countable set and *continuous* if T can take any real number in an interval. As t changes, the process may go from one state to another, a *transition*.

2.2 Discrete stochastic processes

If the stochastic process is discrete, instead of writing $X(t)$ for a discrete number t, a common notation is to write X_n to denote the stochastic variable giving the state of the process after n transitions. The reason for this is that the interest is in how the stochastic process changes and not when it changes. Let X_0 be the initial state.

In a stochastic process there is a probability of going from one specific state to another given by the *transition probabilities*. A fundamental idea of predicting the next transition, and therefore the next state, is to assume that this will be dependent on all the former states. For that reason, consider transition $n + 1$ from a discrete stochastic process and let $i, j \in \Omega$, where i is the state after n transitions and j is the state after transition $n+1$. The probability of going from state i to state j given all the former states $i, i_{n-1}, i_{n-2}, ..., i_0 \in \Omega$ is written as:

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, ..., X_0 = i_0). \tag{2}$$

This notation shows one typical part of the *model* of what governs the stochastic process. A *model* is a simplification of the real world and is made up of assumptions. The assumption here is that the probability of going from a random state $i \in \Omega$ to another random state $j \in \Omega$ is constant during the process (independent of n).

2.3 Markov chains

Definition 2.2. A *Markov chain* is a stochastic process where the *Markov property* holds, namely that the transition probability is only dependent on the present state.

A Markov chain is a special case of (2) with one additional assumption, namely that the transition probability of going from state $i \in \Omega$ to a state $j \in \Omega$ is only dependent on state i, and not the former states:

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = P(X_{n+1} = j | X_n = i). \tag{3}$$

This is called the Markov property. Another word for it is to say that the process is *memoryless.*

2.4 Branching process

A *branching process* is a special case of a discrete stochastic process with Markov property. Visualize a *population* of *individuals.* The interest is now how the population grows from one individual to many. The individuals behave independently of each others. The first individual in the population is defined to be *generation 0.* During one generation, an individual can give birth to $i \geq 0$ *offsprings* or *daughter cells* before dying. The probability for an individual to give birth to i new individuals is given by P_i. Here, $i \in \omega$ where the sample space ω is finite. See Figure 1 for a visualization of a branching process.

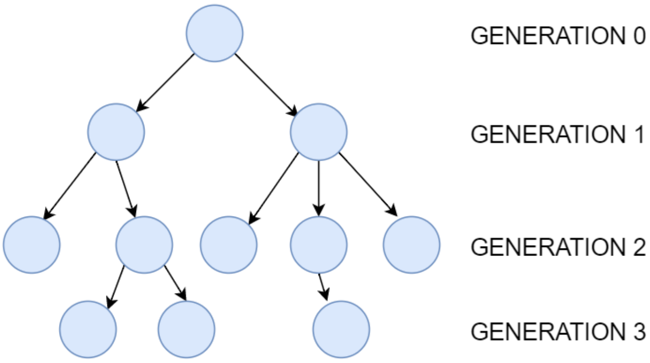

Figure 1: A cartoon of a branching process

The expected number of offsprings denoted by μ from an individual is given by:

$$\mu = \sum_{i \in \omega} i P_i. \tag{4}$$

The variance, σ^2, in the number of offsprings from an individual is given by:

$$\sigma^2 = \sum_{i \in \omega} (i - \mu)^2 P_i. \tag{5}$$

Let X_n be the stochastic variable containing the number of individuals after generation n where $n \in \{0, 1, 2, ...\}$. Generation 1 is the number of offsprings from generation 0 and, in general, generation n is the number of offsprings from generation $n - 1$. As X_n will only be dependent on the former generation, $n - 1$, by definition this is a Markov chain.

Let Z_k be the the stochastic variable giving the number of offsprings from individual k in generation $n - 1$, where $E[Z_k] = \mu$ and $\text{Var}\,(Z_k) = \sigma$ as given in (4) and (5). The total number of individuals in generation n is then given by:

$$X_n = \sum_{k=1}^{X_{n-1}} Z_k. \tag{6}$$

One interest could be to find the expected number of cells, $E[X_n]$ in generation n. By conditioning on X_{n-1} we find that:

$$\begin{aligned} E[X_n] = E[E[X_n|X_{n-1}]] &= E[\sum_{i=1}^{X_{n-1}} E[Z_i]] \\ &= E[\mu X_{n-1}] = \mu E[X_{n-1}], \end{aligned} \tag{7}$$

where μ is given in (4).

As $E[X_0] = 1$, then $E[X_1] = \mu$, $E[X_2] = \mu^2$ up to

$$E[X_n] = \mu^n. \tag{8}$$

The variance of X_n can be computed using the conditional variance formula (Ross (2014) Chapter 3 p. 112):

$$\operatorname{Var}(X_n) = E[\operatorname{Var}(X_n|X_{n-1})] + \operatorname{Var}(E[X_n|X_{n-1}]). \tag{9}$$

Looking at one specific generation n, using Equation (8) and by the definition of independence:

$$E\left[\operatorname{Var}(X_n|X_{n-1})\right] = E\left[\sum_{k=1}^{X_{n-1}} \operatorname{Var}(Z_k)\right] = E\left[X_{n-1}\sigma^2\right] = \sigma^2\mu^{n-1}.$$

Furthermore, by using that $\operatorname{Var}(aX) = a^2 \operatorname{Var} X$ for any constant a and any stochastic variable X, one can show that:

$$\operatorname{Var}(E[X_n|X_{n-1}]) = \operatorname{Var}(\sum_{k=1}^{X_{n-1}} E[Z_k]) = \operatorname{Var}(\mu X_{n-1}) = \mu^2 \operatorname{Var}(X_{n-1}),$$

showing a recurrence relation between X_{n-1} and X_n. Using this recurrence relation, the variance, $\operatorname{Var}(X_n)$ can in fact be derived:

$$
\begin{aligned}
\operatorname{Var}(X_n) &= E[X_{n-1}\sigma^2] + \operatorname{Var}(X_{n-1}\mu) \\
&= \sigma^2\mu^{n-1} + \mu^2 \operatorname{Var}(X_{n-1}) \\
&= \sigma^2\mu^{n-1} + \mu^2 \left(\sigma^2\mu^{n-2} + \mu^2 \operatorname{Var}(X_{n-2})\right) \\
&= \sigma^2(\mu^{n-1} + \mu^n) + \mu^4 \operatorname{Var}(X_{n-2}) \\
&= \sigma^2(\mu^{n-1} + \mu^n) + \mu^4 \left(\sigma^2\mu^{n-3} + \mu^2 \operatorname{Var}(X_{n-3})\right) \\
&\dots \\
&= \sigma^2(\mu^{n-1} + \mu^n + \ldots + \mu^{2n-2}) + \mu^{2n} \operatorname{Var}(X_0) \\
&= \sigma^2(\mu^{n-1} + \mu^n + \ldots + \mu^{2n-2}) \\
&= \sigma^2\mu^{n-1}(1 + \mu + \mu^2 + \ldots + \mu^{n-1}).
\end{aligned}
$$

The last expression can be seen as a *geometric* series which has an explicit form for $\mu \neq 1$ and is equal to n if $\mu = 1$. Therefore,

$$
\operatorname{Var}(X_n) = \begin{cases} \sigma^2\mu^{n-1}\left(\frac{1-\mu^n}{1-\mu}\right), & \text{if } \mu \neq 1 \\ n\sigma^2, & \text{if } \mu = 1. \end{cases} \tag{10}
$$

2.4.1 Probability of extinction

Let the probability for the population to die out be given by π_0. Starting with the first individual, the population may already die out if the first individual dies with probability P_0. If it does not die out, but creates offsprings, consider one of the children. This child gives offsprings with the same probability as its mother, and so the probability that this children or its offsprings will die out also has probability equal to π_0. Conditioning on whether the first cell dies with probability P_0 or gives i offsprings with probability P_i, the probability that the population dies out is given by:

$$
\pi_0 = \sum_{i=0}^{n} \pi_0^i P_i. \tag{11}
$$

Equation (11) results in solving $f(\pi) = 0$ for a polynomial function $f(\pi)$ which has possibly several solutions. However, it can be shown that the probability of extinction is the smallest possible solution of the polynomial equation that is greater than zero[1].

As seen from Equation (8), if μ given by Equation (4) is less than one, $\mu < 1$, then $\pi_0 = 1$ since $\lim_{n\to\infty} E[X_n] = \lim_{n\to\infty} \mu^n = 0$.

2.5 Discrete probability distributions

Definition 2.3. The Binomial distribution, $P(X = k; N, p)$, with $N, k \in \mathbb{N}_0$, $X \leq N$, and $p \in [0, 1]$ is given by:

$$P(X = k; N, p) = \binom{N}{k} p^k (1 - p)^{N-k}. \tag{12}$$

The binomial distribution is used to model the number of successes out of N trials when each trial is independent of the other trials and the probability of having a success is given by p. The mean value of X is Np.

Definition 2.4. The Poisson distribution, $P(X = k; \lambda)$, with $k \in \mathbb{N}_0$ and parameter $\lambda \in \mathbb{R}_{>0}$ is given by:

$$P(X = k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{13}$$

The Poisson distribution is used as a model for the number of occurrences in a period of time or in space. The mean value and the variance of X is given by λ.

Theorem 2.1. Consider both X and Y to be independent and Poisson distributed with rate λ_1 and λ_2 respectively, $X \sim \text{Poisson}(\lambda_1), Y \sim \text{Poisson}(\lambda_2)$. The sum $X + Y$ is then also Poisson distributed with rate equal to the sum of the rates, $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

[1] See Grinstead & Snell (2003) from page 378.

Theorem 2.2 (Poisson limit theorem). Consider the binomial distribution (12). Let $N \to \infty$ and $p \to 0$ such that the mean $\lambda = Np$ remains constant. Then the following approximation applies:

$$\binom{N}{k} p^k (1-p)^{N-k} \approx \frac{\lambda^k}{k!} e^{-\lambda} \tag{14}$$

The result of this Theorem is often called *law of rare events*. The reason for this is the fact that when the number of "trials" $N \to \infty$ while the probability, $p \to 0$ in such a way that Np is kept constant, the binomial distribution can be approximated as a Poisson distribution.

Definition 2.5. The Gamma-Poisson mixture distribution, $P(X = k|\psi, p)$ with $\psi \in \mathbb{R}^+$, and $p \in [0, 1]$ is given by:

$$P(X = k; N, p) = \frac{\Gamma(\psi + k)}{\Gamma(\psi)k!} p^k (1-p)^{\psi},$$

where ψ denotes the *size* and p is a probability of success. $\Gamma(x)$ denotes the Gamma-function[2]. The mean value of X is given by $\tau = \psi(1-p)/p$. The distribution can be reparameterized by setting $p = \psi/(\psi + \tau)$, meaning the distribution is described through the parameters τ and ψ. The distribution is then given by:

$$P(X = k; \psi, \tau) = \frac{\Gamma(\psi + k)}{\Gamma(\psi)k!} \left(\frac{\psi}{\psi + \tau}\right)^k \left(\frac{\tau}{\psi + \tau}\right)^{\psi}.$$

The variance, σ^2, as a function of τ and ψ is $\sigma^2 = \tau + \tau^2/\psi$.

In this thesis, $X \sim \text{GP}(\tau, \psi)$, will mean that the random variable X is Gamma-Poisson mixture distributed.

[2] See appendix B

2.6 Ordinary least squares method

Given explanatory variables $x_1, x_2, ..., x_n$ and response variables $y_1, y_2, ..., y_n$. Consider a simple linear regression model, $\hat{y}_i = \alpha x_i + \beta$. In simple linear regression, the coefficients α and β are determined using the ordinary least squares method by minimizing the following expression:

$$\min_{\alpha,\beta} \mathcal{Q}(\alpha, \beta) = \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha x_i - \beta)^2 .$$

This can be solved analytically (see for instance Walpole et al. (2014) on page 450), and coefficients given by $\hat{\alpha}$ and $\hat{\beta}$ are given by:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

and

$$\hat{\beta} = \bar{y} - \hat{\alpha}\bar{x}.$$

2.7 Hypothesis Testing

Definition 2.6. A hypothesis test is a rule that decides whether a given hypothesis called the *null hypothesis*, H_0, is <u>chosen</u> to be accepted or rejected given a sample of values, $\mathbf{X}$. The complement of H_0 is the *alternative hypothesis*, H_1, which is accepted if H_0 is rejected. More formally:

1. For a given set S of sample values, H_0 is accepted as true.

2. For sample values not part of the set S, but in the complement set, S^c, H_0 is rejected and so H_1 is accepted.

Remark the underline of the word chosen. It's up to the constructor of the hypothesis test what is acceptable enough in order to accept H_0. Accepting H_0 is not the same as proving H_0, but rather the test failed to conclude that H_0 is false.

2.8 Approximation methods

Definition 2.7. Let T_1,...,T_k be random variables with means $\theta_1, ..., \theta_k$ and define $\mathbf{T}$ = $(T_1, ..., T_k)$ and $\theta = (\theta_1, ..., \theta_k)$. Let g($\mathbf{T}$) be any differentiable function. Then the first-order Taylor series expansion given the observed values $\mathbf{t} = (t_1, ..., t_k)$ is given by:

$$g(\mathbf{t}) = g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(t_i - \theta_i) \ + \ \text{Remainder},$$

where $g_i'(\theta) = \frac{\partial}{\partial t_i} g(\mathbf{t})|_{t_1=\theta_1,...,t_k=\theta_k}$.

This expression for first-order Taylor series of a multivariate function gives rise to an approximation method for $E[g(\mathbf{T})]$. First, observing that $E(t_i - \theta_i) = 0$, the first-order approximation shows that:

$$E[g(\mathbf{T})] \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta) E[T_i - \theta_i] = g(\theta). \tag{15}$$

The method of approximations of expectancies and variances using Taylor series is often called the Delta method.

2.9 Conditional probability and conditional expectation for discrete distributions

Definition 2.8. Given the event A in addition to the mutually exclusive events $B_1, \dots, B_n$, whose probabilities sum to 1, $\sum_i^n P(B_i) = 1$. By the *law of total probability*, the probability of event A, $P(A)$, may be written:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \tag{16}$$

Definition 2.9. Given a discrete random variable X in addition to *any* discrete random variable Y. Then, by *the law of total expectation*:

$$E[X] = \sum_{y} E[X|Y = y]P(Y = y) = E_Y[E[X|Y]] \tag{17}$$

2.10 Moment generating function (MGF)

The moment generating function (MGF), $M_X(t)$, for a random variable X is by definition:

$$M_X(t) = E[e^{tX}], \tag{18}$$

for any $t \in \mathbb{R}$, where the expectation exists.

3 A branching process with accumulating mutations for modelling neutral tumor evolution

In this thesis, the growth of a tumor evolving neutrally as defined in 1.1 will be modelled as a branching process as described in Section 2. The idea behind this model is based on the assumption that all tumor cells are regarded to have equal *fitness*, meaning that all tumor cells during the whole evolution are equally likely to survive and reproduce. Strictly speaking, the tumor size should be regarded to change with one cell at a time, where each cell at some time point after birth takes a decision to either divide or die. However, as the tumor cells have equal *fitness*, the tumor cells can be expected to reproduce at same rate. Therefore, visualizing the growth of a tumor as a branching tree beginning with the first tumor cell in generation 0, the change in the number of tumor cells from generation to generation is regarded to reflect the change in tumor size in discrete time[3].

In addition to model the growth of the tumor, the accumulation of passenger mutations must also be modelled in order to model neutral tumor evolution as described in Definition 1.1. New mutations are allowed to happen in each generation. Having a model for both the growth of the tumor and the accumulation of passenger mutations, the aim is ultimately to show how to deduce the theoretical expression given in Williams et al. (2016), yielding a characteristic pattern for a neutrally evolving tumor.

3.1 Underlying assumptions of the model

Let X_i denote the number of tumor cells in generation i, and $X_0 = 1$ represent the first tumor cell. This first tumor cell has selective advantages, or in other words the

[3] See Appendix A for more mathematical motivation of the model.

probability for the cell to reproduce is higher than surrounding (nontumor-)cells. The tumor cell can then divide and the two new tumor cells, called *daughter cells*, are assumed to inherit the same selective advantages. This process then may continue creating a colony of tumor cells.

The fate of any tumor cell is assumed to be either to die or to divide into two new cells. Using the same notation as in Section 2, $P_0 \geq 0$, $P_2 \geq 0$ and $P_i = 0 \ \forall \ i \neq \{0, 2\}$. After a cell division, new mutations may occur. Here, only *point mutations* are considered. A new mutation occurring in one cell is assumed to be inherited by all the potential future daughter cells. The number of new mutations occurring in a daughter cell is modelled to be Poisson distributed. The reasoning behind this probability model may be explained in such a way that the average probability, p, for a point mutation to occur at at specific *locus* somewhere along the whole genome is very low, and the mutations are regarded to be independent of each other. Referring to Equation (12) in Section 2, as $N \to \infty$ and $p \to 0$ the binomial distribution can be approximated as a Poisson distribution as seen in Theorem 2.2. N will represent the large number of nucleotide bases in the genome. Denoting $\mathcal{P}$ as the number of mutations occurring in a daughter cell, one formally writes:

$$\mathcal{P} \sim \text{Poisson}(\lambda), \tag{19}$$

where λ is the average number of somatic mutations a new daughter cell possesses. The assumptions above are really as a result of the so-called Infinite Sites Model (ISM) (Kimura 1968). The model is based on three assumptions, where the first assumption is necessary in order for the two next assumptions to hold:

1. The genome is regarded to be a sequence of infinite number of nucleotides (the same as setting $N \to \infty$ above).

2. Every new mutation occurs at a novel site.

3. Once a mutation at a novel site has occurred, it is not allowed to mutate back to the origin.

By the definition of neutral tumor evolution in Definition 1.1, only *passenger mutations* are allowed to occur during evolution, as opposed to *driver mutations*. A passenger mutation does not change the *fitness* of a cell, while a driver mutation does. The fitness is a measure of reproductive success. Increasing the fitness of the cell will increase the cell's probability of dividing. The driver mutations are only allowed to arise prior to *tumor initiation* following The Big Bang model in Sottoriva et al. (2015).

One can now visualize the model in all its generality as in Figure 2.

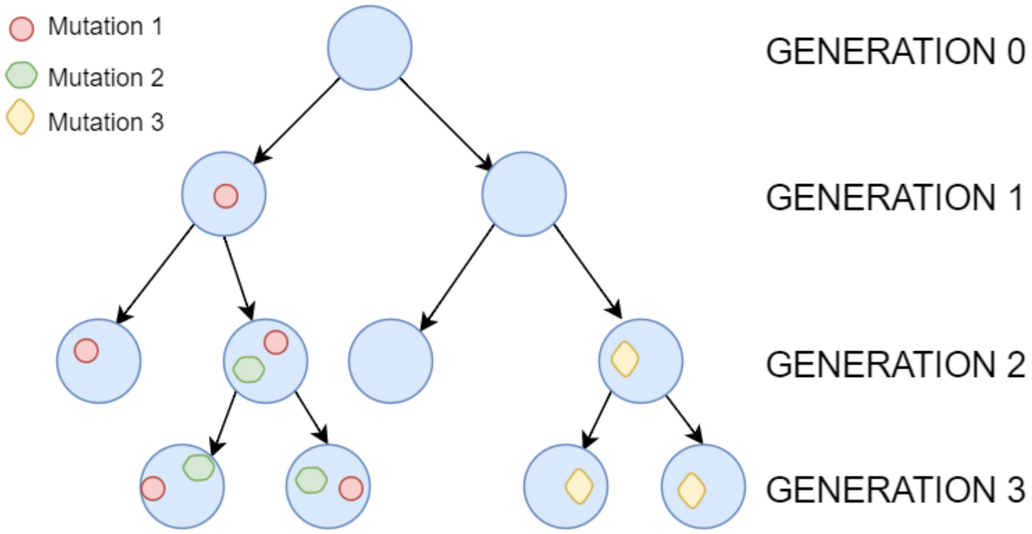

Figure 2: Drawing summarizing the branching process with accumulating mutations to model neutral tumor evolution.

3.2 Mathematical details of model

As described earlier, for neutral tumor evolution no driver mutations are assumed to occur during the process. This means that all cells have the same probability of dividing during the whole process. Visualizing this case as a branching process, let $P_2 = \beta$ and $P_0 = 1 - \beta$ be the probability for any cell to respectively divide or die, noting that $P_i = 0 \ \forall \ i \neq \{0, 2\}$. As described earlier, λ is the mutation rate given in Equation (19). Furthermore, the analysis for this model is based on the fact that $\beta > 0.5$ since a tumor will not grow large if $\beta < 0.5$ as shown in Section 2. From Equation (4), the expected number of offsprings from a tumor cell, μ, is given by

$$\mu = 0(1 - \beta) + 2\beta = 2\beta,$$

while the variance given in Equation (5), σ^2, is given by

$$\sigma^2 = (0 - \mu)^2(1 - \beta) + (2 - \mu)^2\beta = 4\beta(1 - \beta),$$

which is intuitively maximized for $\beta = 0.5$. From Equation (8), the expected number of tumor cells in generation i, $E[X_i]$, is given by

$$E[X_i] = (2\beta)^i, \tag{20}$$

while from Equation (9) the variance in generation i, $\mathrm{Var}(X_i)$ is given by

$$\mathrm{Var}(X_i) = \begin{cases} 4\beta(1-\beta)(2\beta)^{i-1}\left(\frac{1-(2\beta)^i}{1-2\beta}\right) & \text{if } \ \beta \neq 0.5 \\ 4i\beta(1-\beta) & \text{if } \ \beta = 0.5 \end{cases} \tag{21}$$

Let Z_j^i be the number of offsprings for a tumor cell j in generation i. Given the number of tumor cells X_i in generation i, the total number of tumor cells in the next generation, X_{i+1}, is given by $X_{i+1}|X_i = \sum_{j=1}^{X_i} Z_j^i$. Knowing the distribution of Z_j^i, one realizes that the number of cells that divide in generation i is binomial distributed.

Let D_j^i be the stochastic variable equal to 1 if cell j in generation i divides and 0 otherwise. Then $P(D_j^i = 1) = P(Z_j^i = 2) = \beta$ and $P(D_j^i = 0) = P(Z_j^i = 0) = 1 - \beta$, and the total number of cells that divide in generation i, D_i, is given by

$$D_i | X_i = \sum_{j=1}^{X_i} D_j^i \sim \text{Bin}(X_i, \beta).$$

Knowing how many cells that divide in generation i implies knowing how many cells there are in generation $i + 1$ given by:

$$X_{i+1} = 2D_i,$$

which is two times a binomial distributed variable. The parameter X_i in the distribution of D_i will also be 2 times a binomial distributed variable, $X_i = 2D_{i-1}$. This result shows that the number of cells in each generation, apart from the first, must be an even number.

We will now introduce how somatic passenger mutations accumulate during the growth of the tumor.

3.2.1 Variant allele frequences

Suppose a new passenger mutation arises in a specific locus in a cell in generation i with a total of X_i cells. In this model, it is regarded that each cell has π copies of this locus during the *whole* growth, and therefore in generation i, there are a total of πX_i copies of this locus. The immediate VAF, using Definition 1.2 with $L = \pi X_i$, for this specific somatic mutation is then $1/\pi X_i$. In other words, when a passenger mutation arises, the total number of cells in the population at that moment is approximated to be equal to the number of cells in the corresponding generation in the branching tree. As explained before, this approximation is based on the fact that all cells divide at the same rate. For future generations the VAF may change, but what is known for sure is that the VAF will be of the form:

$$VAF = \begin{cases} \frac{n}{m}, & \text{where } \frac{n}{m} < 1/\pi, \quad n, m \in \mathbb{N} \\ 0, & \text{if all cells having the mutation die out.} \end{cases} \tag{22}$$

The largest VAF a somatic mutation can have is when it is present in all tumor cells, which in this model with no copy number variations means a VAF of $1/\pi$.

3.2.2 Properites of surviving tumors

Since cells are allowed to die, the tumor can die as a whole before it has even been observed. Suppose a tumor cell has been initiated. From Equation (11), the probability that this cell or all future offsprings of this cell will die, ρ, is given by

$$\rho = \begin{cases} \frac{1}{\beta} - 1, & \text{for } \beta > 0.5 \\ 1, & \text{for } \beta \leq 0.5 \end{cases} \tag{23}$$

As only surviving tumors are observed, Equation (20) must be modified, conditioned on whether the tumor eventually dies out or not:

$$E[X_i] = E[X_i | \,\text{dies}]\rho + E[X_i | \,\text{survives}](1 - \rho). \tag{24}$$

For instance, the term $E[X_i | \,\text{dies}]$ is the expected value of the number of cells in generation i given that the tumor at some point will die out (the tumor may die before generation i or after). The term $E[X_i | \,\text{survives}]$ means the expected value of number of cells in generation i given that the tumor will increase to infinite size. This means for instance that $E[X_1 | \,\text{survives}] = 2$ for sure. Using Equation (20) and (24) this yields that $E[X_1 | \,\text{dies}] = 2(1 - \beta)$. This can also be computed:

$$\begin{aligned} E[X_1 | \,\text{dies}] &= 0P(X_1 = 0 | \,\text{dies}) + 2P(X_1 = 2 | \,\text{dies}) \\ &= 2\frac{P(X_1 = 2 \cap \text{dies})}{P(\text{dies})} = 2\frac{\beta(1/\beta - 1)^2}{(1/\beta - 1)} = 2(1 - \beta), \end{aligned}$$

which motivates Equation (24). Computations of $E[X_i | \,\text{dies}]$ for $i > 1$ become infeasible.

For tumors observed, generation i will naturally be large, and so without any formal proof the following limit must hold:

$$\lim_{i \to \infty} E[X_i |\, \text{dies}] = 0.$$

The *observed* expected number of cells in generation i, when i is large, can then be regarded to be given by:

$$E[X_i |\, \text{survives}] \approx \frac{E[X_i]}{1-\rho} = (2\beta)^i \frac{\beta}{2\beta - 1}. \tag{25}$$

As an example, consider the case where $\beta = 0.55$. Around 350 surviving populations were simulated reaching generation 150. The sample mean of the number of cells was computed for each generation and compared with Equation (25). In addition, in order to show how the variation in population changes for increasing generations, an interval equal to two sample standard deviation is added for each generation. See Figure 3.

3.3 Simulations

According to the mathematical model of neutral tumor evolution explained above, the only parameters needed in order to construct a tumor is the probability of division β and the mutation rate λ. However, even if the model only requires a few parameters, the corresponding distribution of VAFs of a growing tumor is complex except the unrealistic case were $\beta = 1$. This motivates for simulations of a growing tumor, starting with one tumor cell that may divide and produce two new tumor cells possibly with additional passenger mutations. This process may continue producing a surviving tumor consisting of a spectre of somatic mutations with different VAFs. An advantage of simulating neutral tumor evolution is that there is no need to simulate a large tumor consisting of billions of cells. As the behaviour of neutral tumor evolution is the same during the growth, the same pattern (VAF-distribution) will be shown no

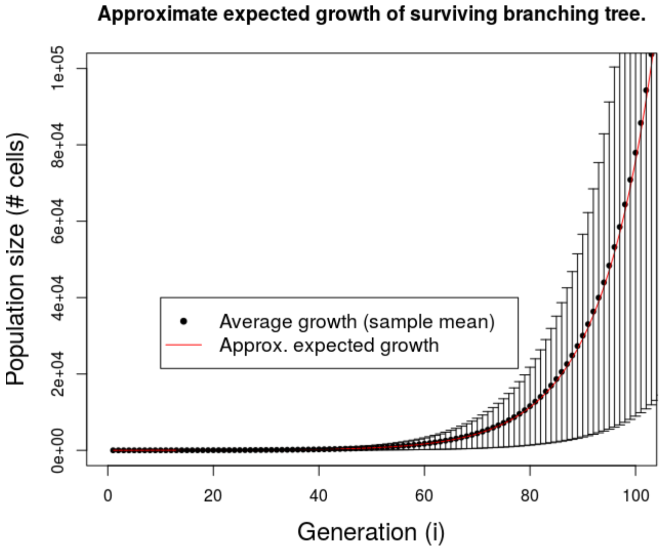

Figure 3: Plot shows that Equation (25) is a good approximation of expected population growth. Variation increases as the generation increases as can be seen by looking at the increase in length of the intervals for each respective sample mean, where each interval equals to two sample standard deviations.

matter the size of the simulated tumor. The simulation is built in the programming language R (R Core Team 2014). The code is available at `https://github.com/palVJ/tumorEvolution.git` and is also given at the end of this thesis.

The simulations will be used to either confirm analytic results or to explore further when analytic results seem infeasible. The choice of parameters is motivated by earlier work as given in Jones et al. (2007), where the average point mutation per base pair in colorectal cancers is estimated to be $\approx 5 \cdot 10^{-10}$. As there are $\approx 3 \cdot 10^{9}$ base pairs in the genome, the average *whole-genome* mutation rate per cell division is then estimated to be the product of these key numbers equal to around 1.5. Inspecting only *exome* regions which is about 1 % of the whole genome, the mutation rate per cell division is estimated to be around 0.015. A mutation rate of $\mu = 1.5$ is used in the simulations.

The analytical results will anyway be independent of what the value of the mutation rate is. The probability of cell division, β, must both intuitively and according to the model be larger than 0.5 in order for a tumor to grow large. As was done in Sun et al. (2017), $\beta = 0.55$ in all simulations.

3.4 A pure birth model

As a toy example, consider the case with $\beta = 1$ meaning all cells survive. The result of this special case gives fruitful information that can be exploited for the realistic case where $\beta \neq 1$. With this pure birth model and ISM, the VAF of a somatic mutation given in Definition 1.2 will be constant for all future generations once a new, and therefore unique, mutation has occurred. In order to see this, suppose a new mutation occurred in generation i which consist of 2^i cells as can be seen from Equation (8). As the mutation is unique only occurring in one cell out of total 2^i cells, the VAF in the pure birth case, $f_{pb}(i)$, with ploidy, π, is equal to:

$$f_{pb}(i) = \frac{1}{\pi 2^i} \tag{26}$$

As the cell possessing this mutation divides, two new cells will possess the unique mutation out of total 2^{i+1} cells in generation $i+1$. The VAF in generation $i+1$ is then:

$$f_{pb}(i) = \frac{2}{\pi 2^{i+1}} = \frac{1}{\pi 2^i}.$$

In other words, by the assumptions of ISM the VAF remains the same in all future generations. Suppose a tumor grows according to the pure birth model. Then, each recorded VAF belongs to a specific generation i in the branching tree, consisting of 2^i cells (by Equation (8)). Let $T_i^n(f_{pb}(i))$ denote the total number of unique somatic mutations appearing in generation i, and therefore as discussed above will have VAF equal to $f_{pb}(i)$ with $1 \leq i \leq n$. The model then says, using Equation (2.1), that:

$$T_i^n(f_{pb}(i)) \sim \text{Poisson}(2^i\lambda) \sim \text{Poisson}\left(\frac{\lambda}{\pi}\frac{1}{f_{pb}(i)}\right), \tag{27}$$

since $T_i^n(f_{pb}(i))$ is a result of mutations occurring in generation i, where each of the 2^i possibly mutated cells acquires a given number of mutations according to Equation (19) with rate λ. From Definition 2.4, the expected value of $T_i^n(f_{pb}(i))$ is:

$$E[T_i^n(f_{pb}(i))] = \frac{\lambda}{\pi}\frac{1}{f_{pb}(i)} \propto \frac{1}{f_{pb}(i)}. \tag{28}$$

Furthermore, as was done in Williams et al. (2016), the expected number of somatic mutations with VAF larger than some specified value $f_{pb}(k)$, given by $M(f_{pb}(k))$, can be computed:

$$\begin{aligned}
M(f_{pb}(k)) &= \sum_{i=1}^{k} E[T_i^n(f_{pb}(i))] = \frac{\lambda}{\pi}\sum_{i=1}^{k}\frac{1}{f_{pb}(i)} = \lambda\sum_{i=1}^{k}2^i = \lambda(\frac{1-2^{k+1}}{1-2}-1)\\
&= \lambda(2^{k+1}-2) = 2\lambda(2^k-1) = 2\lambda\left(\frac{1}{\pi f_{pb}(k)}-1\right) = 2\lambda\left(\frac{1}{\pi f_{pb}(k)}-\frac{\pi}{\pi}\right)\\
&= \frac{2\lambda}{\pi}\left(\frac{1}{f_{pb}(k)}-\pi\right) = \frac{2\lambda}{\pi}\left(\frac{1}{f_{pb}(k)}-\frac{1}{f_{pb}(0)}\right) = \frac{2\lambda}{\pi}\left(\frac{1}{f_{pb}(k)}-\frac{1}{f_0}\right),
\end{aligned} \tag{29}$$

where $f_0 = f_{pb}(0) = 1/\pi$ is the maximum VAF a somatic mutation can have corresponding to truncal somatic mutations already present from the first tumor cell. Hence, $M(f_{pb}(k))$ is proportional to $1/f_{pb}(k)$, $M(f_{pb}(k)) \propto 1/f_{pb}(k)$, with slope $2\lambda/\pi = \lambda$ since $\pi = 2$.

3.5 A birth and death model

When generalizing the case allowing tumor cells to die, meaning that $\beta < 1$, the number of cells in generation i is now stochastic, not deterministic as in the pure birth case. Furthermore, the tumor may eventually die out with probability ρ as seen in Section 2. However, in reality only surviving tumors are observed. From Equation (24):

$$E[X_i|\,\text{survives}] = \frac{E[X_i]}{1-\rho} - \frac{\rho}{1-\rho}E[X_i|\,\text{dies}], \tag{30}$$

Denote $Y_i = X_i|\,\text{survives}$. Furthermore, consider somatic mutations arising in generation i. Only some of the mutations will be present in the observed tumor since only some of the cells in generation i will produce surviving descendants with probability $(1-\rho)$. Let W_i denote the number of cells in generation i that will produce surviving descendants. Given $Y_i = y_i$, and since all cells are regarded to be independent of each other:

$$W_i|(Y_i = y_i) \sim \text{Bin}(y_i, (1-\rho)).$$

Using the conditional expectation rule:

$$E[W_i] = E[E[W_i|Y_i]] = E[(1-\rho)Y_i] = (1-\rho)E[Y_i].$$

Let T_i denote the number of somatic mutations created in generation i that will survive. This will be equal to the number of somatic mutations created in cells giving surviving descendents given by W_i. If $W_i = w_i$, then:

$$T_i|(W_i = w_i) \sim \sum_{k=1}^{w_i} \text{Poisson}(\lambda) = \text{Poisson}(w_i\lambda).$$

Once again using the conditional expectation rule, the expected number of surviving somatic mutations created in generation i is given by:

$$\begin{aligned} E[T_i] &= E[E[T_i|W_i]] = E[\lambda W_i] = \lambda E[W_i] \\ &= \lambda(1-\rho)E[Y_i] \\ &= \lambda(1-\rho)\left(\frac{E[X_i]}{1-\rho} - \frac{\rho}{1-\rho}E[X_i|\,\text{dies}]\right) \\ &= \lambda E[X_i] - \lambda\rho E[X_i|\,\text{dies}] \end{aligned} \tag{31}$$

Remember that X_i denotes the unconditional number of cells in generation i, where $E[X_i]$ is given in Equation (20), while $E[X_i|\,\text{dies}]$ is the expected number of tumor cells in generation i conditioned that the tumor eventually dies.

Using the same philosophy as Williams et al. (2016), one can estimate the expected number of total surviving somatic mutations that arose in all generations before or in generation k, given by $m_N(k)$:

$$\begin{aligned} m_N(k) &= \sum_{i=1}^{k} E[T_i] = \lambda \sum_{i=1}^{k} (2\beta)^i - \lambda\rho \sum_{i=1}^{k} E[X_i|\,\text{dies}] \\ &= \lambda \left(\frac{1-(2\beta)^{k+1}}{1-2\beta} - 1 \right) - \lambda\rho \sum_{i=1}^{k} E[X_i|\,\text{dies}] \\ &= 2\lambda \frac{\beta}{2\beta-1}(2\beta)^k - \frac{\lambda}{2\beta-1} - \lambda - \lambda\rho \sum_{i=1}^{k} E[X_i|\,\text{dies}] \\ &= \frac{2\beta\lambda}{2\beta-1}(2\beta)^k - \frac{2\beta\lambda}{2\beta-1} - \lambda\rho \sum_{i=1}^{k} E[X_i|\,\text{dies}]. \end{aligned} \tag{32}$$

Do notice that Equation (32) is exact. Without going too much into detail of the distribution of $X_i|\,\text{dies}$, it is reasonable to think that $E[X_i|\,\text{dies}]$ converges to zero as the generation i increases since it is given that the tumor will eventually die. As an example, consider the case where the cell division probability is $\beta = 0.55$. By Equation (23), the probability for a cell and its descendents to eventually die out is around 0.8. For a population reaching $x = 10$ cells at one point, the probability that this population will eventually die out is $\rho^{10} = 0.8^{10} \approx 0.1$ which is quite low. Therefore, most likely extinct populations does not reach a very large maximum population since the probability for a tumor to die out given the present number of tumor cells, $\rho^x = e^{log(\rho)x}$, decreases exponentially as the number of tumor cells x increases. With this argument in mind, this means that the following infinite sum:

$$\sum_{i=1}^{\infty} E[X_i|\,\text{dies}] = C(\beta), \tag{33}$$

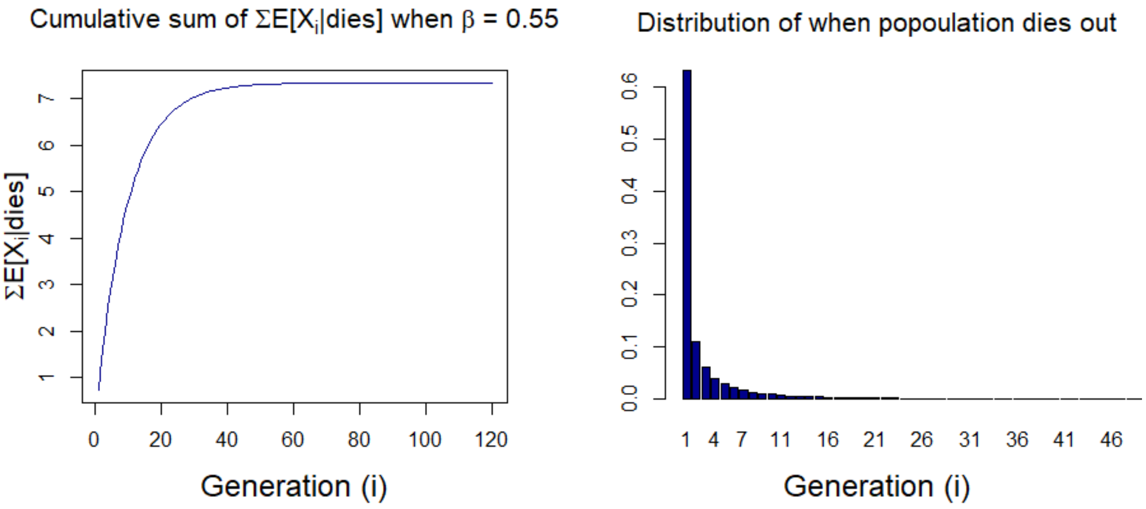

Figure 4: The cumulative sum $\sum_{i=1}^{k} E[X_i|dies]$ seems to converge, and is about 7 with the choice of $\beta = 0.55$.

for some constant $C(\beta)$ dependent on β, the cell division probability. The question is then how large $C(\beta)$ is, and how fast the sum converges. As an example, consider the case where $\beta = 0.55$. Simulating only extinct populations, the corresponding cumulative sum $\sum_{i=1}^{k} E[X_i|\,\text{dies}]$ as a function of generation i is plotted in Figure 4 showing that indeed the sum seems to converge[4]. A histogram showing the distribution of when the population dies is also given in Figure 4.

If the cumulative sum converges, then also:

$$\lim_{k\to\infty} m_N(k) = \frac{2\beta\lambda}{2\beta-1}(2\beta)^k - \frac{2\beta\lambda}{2\beta-1} - \lambda\rho C(\beta),$$

converges. In fact, given $\beta = 0.55(\rho \approx 0.8)$, Figure 4 shows that $C(\beta) \approx 7$. This means, given for instance that $\lambda = 1.5$, that the last term in Equation (32) converges

[4] The simulations should not be regarded to give exact results.

to about $\lambda\rho C(\beta) \approx 8.4$. The last term in Equation (32) is therefore expected to be negligible compared to the two first terms. In other words, when k is sufficiently large, $m_N(k)$ will in general increase exponentially. This property can be confirmed by simulations. Let $M_N(k)$ be the total number of surviving somatic mutations that arose in all generations less than or equal to some generation k[5], where $\beta = 0.55$ and $\lambda = 1.5$. 1000 tumors are simulated reaching last generation $j = 150$, and then $M_N(k)$ for each tumor for generation $k = 1 : 40$ is computed. For a given generation k, the mean $m_N(k) = E[M_N(k)]$ is estimated, given by $\hat{m}_N(k)$, by taking the sample mean from all tumors from the corresponding generation. The sample mean is an unbiased estimator no matter what the real distribution of $M_N(k)$ is. As the cumulative sum $\sum_{i=1}^{k} E[X_i|\,\text{dies}]$ is regarded to be negligible, instead define $\tilde{m}_N(k)$ to be:

$$\tilde{m}_N(k) = \frac{2\beta\lambda}{2\beta - 1}(2\beta)^k - \frac{2\beta\lambda}{2\beta - 1}, \tag{34}$$

Then, $\tilde{m}_N(k)$ is compared to $\hat{m}_N(k)$ from the 1000 tumors simulated. In addition the sample standard deviation of $M_N(k)$ for each generation is given having the purpose of showing the variation in the data. The result is given in Figure 5.

As one can see from Figure 5, $\tilde{m}_N(k)$ increases in the same way as $\hat{m}_N(k)$. Remember that $m_N(k)$ in Equation 32 is in fact exact, while $\hat{m}_N(k)$ is the sample mean with $E[\hat{m}_N(k)] = m_N(k)$ meaning that $\lim_{k\to\infty} E[\tilde{m}_N(k) - \hat{m}_N(k)] = \lambda\rho C(\beta)$ in general. It can also be seen from the figure that $\tilde{m}_N(k) > \hat{m}_N(k)$. Notice the large variation in $M_N(k)$ visualized as as interval equal to two times the sample standard deviation for each generation k.

[5] Do notice the difference between $m_N(k)$ and $M_N(k)$: $M_N(k)$ is stochastic, $m_N(k) = E[M_N(k)]$.

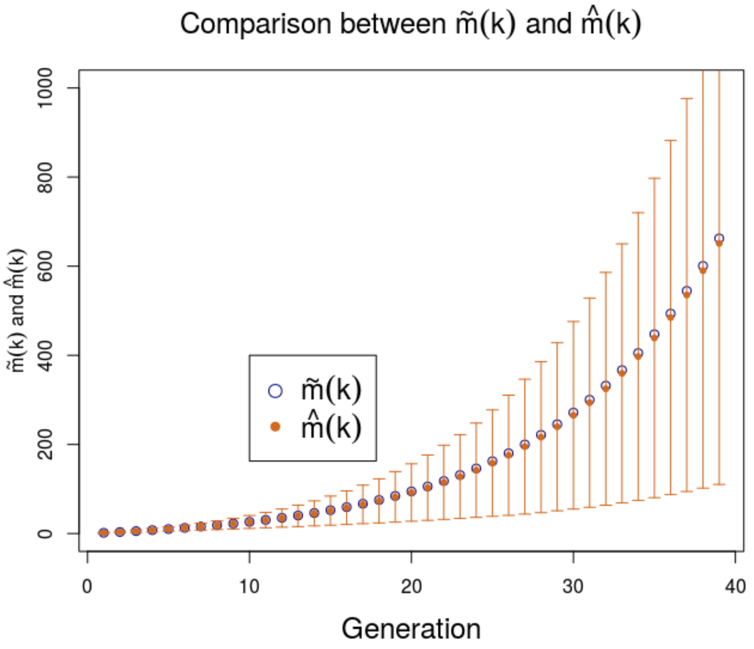

Figure 5: $\tilde{m}_N(k)$ (blue points) is a good approximation of $\hat{m}_N(k)$(orange points). As can be seen, even though $\tilde{m}_N(k) > \hat{m}_N(k)$, the difference is negligible for increasing generations.

3.5.1 How to adapt the model to observed data

In observed data of somatic mutations, there is of course no information of when each mutation occurred. Therefore, in order for the branching process model to be practical it must be independent of which generation each somatic mutation appeared, but rather be a function of the corresponding VAF in real data. Furthermore, the results so far has been based on expectancies. However, in reality this is not the kind of data one observes, but rather a single sample from an unknown statistical distribution, which in this case will be the VAF-distribution.

As a new somatic mutation arises in generation $i > 0$, let S_i^j be the number of cells having this mutation in generation j, and let Y_j denote the total number of cells in generation j, where it is known that both the tumor and the mutation survives. The VAF of this specific somatic mutation, given by f_i^j, in any future generation $j > i$ is given by:

$$f_i^j = \frac{S_i^j}{\pi Y_j},$$

Consider the expected VAF of this new mutation denoted by $E[f_i^j]$ after a generation j, for $j >> i$ meaning that the tumor has grown large. Using the Delta method as described in Section 2, where $E[S_i^j] \approx (2\beta)^{j-i}\frac{\beta}{2\beta-1}$ and $E[Y_j] \approx (2\beta)^j\frac{\beta}{2\beta-1}$ for large j from Equation (25), the expected VAF of this somatic mutation in first order approximation is given by:

$$E[f_i] = \lim_{j\to\infty} E[f_i^j] \approx \frac{E[S_i^j]}{\pi E[Y_j]} \approx \frac{(2\beta)^{j-i}}{\pi(2\beta)^j} = \frac{1}{\pi(2\beta)^i}, \tag{35}$$

which is independent of j just as in the pure birth case. Notice the resemblance with Equation (26). The difference is that Equation (26) is exact in the special case $\beta = 1$, while in Equation (35), f_i^j is a random variable, and the expected value is only approximated to first order. The idea is still the same, namely that somatic mutations arising at the same generation will be expected to have equal VAFs also in future because all cells grow at the same rate. This is just one way to explain neutral tumor evolution. An interesting question is how accurate Approximation (35) is when j is finite. For sure one must be certain that generation j is large enough compared to generation i, but regardless of this the first order approximation of the expected value will be biased. Notice that Approximation (35) is independent of the value of mutation rate λ, but as the number of arising somatic mutations per generation will tend to increase for increasing generations, the number of samples per generation will also tend to increase for increasing generations. In order to check the accuracy of Approximation (35), 1000 tumors with parameters $\beta = 0.55$ and $\lambda = 1.5$ are simulated where the last generation is $j = 150$. The VAF of each somatic mutation reaching last generation, and therefore assumed to survive[6], is stored along with the corresponding generation the mutation arose. Afterwards, given a generation i the sample mean for all surviving somatic mutations arising in generation i from all the 1000 tumors is computed for $i = 1$ to 40. The approximated mean to first order, $E[f_i]$, given in Approximation (35) is then compared to these sample means. A plot including the sample mean for each generation i, with a corresponding interval equal to two sample standard deviations is given in Figure 6.

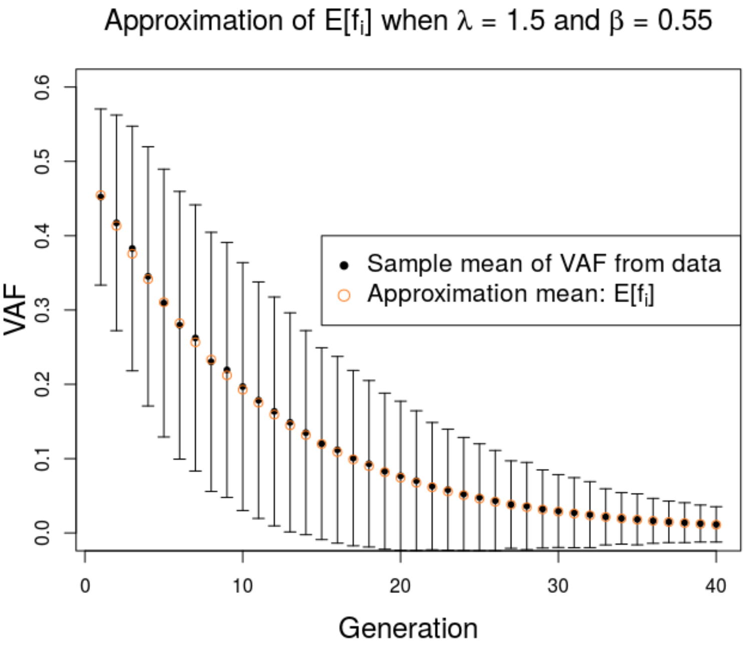

Figure 6: 1000 tumors are simulated with last generation $j = 150$. $E[f_i] = \frac{1}{\pi(2\beta)^i}$ is a quite good approximation of $E[f_i^j]$ when j is large. An interval equal to two standard deviations is plotted for the samples in each generation.

Figure 6 clearly shows that $E[f_i]$ is a good approximation of $E[f_i^j]$, especially for

[6] Note that this is an approximation

large generations, but also acceptable for lower generations. The less accurate sample means for low generations may be due to a smaller amount of samples per generation, larger variation in VAF and the fact that $E[f_i]$ is a first order approximation. A second order approximation of f_i^j is given by:

$$E\left[\frac{S_i}{\pi X_j}\right] \approx \frac{1}{\pi}\left(\frac{E[S_i^j]}{E[X_j]} - \frac{\mathrm{Cov}(S_i^j, X_j)}{E[X_j]^2} + \frac{E[S_i^j]}{E[X_j]^3}\mathrm{Var}(X_j)\right), \tag{36}$$

It is reasonable to think that the covariance term will be positive, since a tumor that grows large will also tend to have somatic mutations being present in a large amount of the cells. It is also reasonable that the covariance increases for decreasing generations since the growth of somatic mutations arising early is more dependent on the growth of the tumor as a whole. This means that the second term in Expression (36) will be smaller for increasing generations. The third and last term varies as $E[S_i^j]$ which will be smaller for increasing generation. Therefore, the second order approximation will converge to the first order approximation when increasing the generation.

Approximation (35) given by $E[f_i]$ can now be used to reformulate Equation (32). By setting $(2\beta)^k = 1/(\pi E[f_k])$ and assuming the cumulative sum given in Equation (33) is negligible, let $m_N(k)$ be approximated as:

$$\begin{aligned} m_N(k) &= 2\lambda\frac{\beta}{2\beta-1}(2\beta)^k - \frac{2\beta\lambda}{2\beta-1} - \lambda\rho\sum_{i=1}^{k}E[X_i|\,\mathrm{dies}] \\ &\approx \frac{2\beta\lambda}{\pi(2\beta-1)}\frac{1}{E[f_k]} - \frac{2\beta\lambda}{2\beta-1} - \lambda\rho\sum_{i=1}^{k}E[X_i|\,\mathrm{dies}] \\ &= \frac{\lambda}{1-\rho}\left(\frac{1}{E[f_k]} - 2\right) - \lambda\rho\sum_{i=1}^{k}E[X_i|\,\mathrm{dies}] \\ &= \frac{\lambda}{1-\rho}\left(\frac{1}{E[f_k]} - \frac{1}{f_0}\right) - \lambda\rho\sum_{i=1}^{k}E[X_i|\,\mathrm{dies}] \\ &\approx \frac{\lambda}{1-\rho}\left(\frac{1}{E[f_k]} - \frac{1}{f_0}\right), \end{aligned} \tag{37}$$

where it has been used that $\pi = 2$, ρ is given by Equation (23) and where $f_0 = 1/\pi = 1/2$ is defined to be the VAF corresponding to truncal mutations being present in the first tumor cell and therefore also in all future tumor cells. Notice the resemblance with the pure birth case in Equation (29). The only difference is that the variable $f_{pb}(k)$ in the pure birth model in Equation (29) is replaced by the expected value of the corresponding random variable in the birth-death model, denoted as $E[f_k]$. In addition, the mutation rate λ is now divided by $1 - \rho$ in Approximation (37). However, for pure birth, $\rho = 0$, so Approximation (37) coincides with Equation (29) in this case. From now on, let Approximation (37) be denoted as $\tilde{m}_N(E[f_k])$, namely:

$$\tilde{m}_N(E[f_k]) = \frac{\lambda}{1-\rho}\left(\frac{1}{E[f_k]} - \frac{1}{f_0}\right). \tag{38}$$

As a reminder, $E[f_k]$ is the expected VAF for a surviving somatic mutation arising in generation k in a large growing tumor. However, in reality one looks at observed data from a specific tumor with estimated VAFs of somatic mutations without knowing when each mutation occurred. How can $m_N(E[f_k])$ be applied to real data when it is both based on expectancy and dependent on when each mutation occurred? One idea is to look back at the pure birth case where $\beta = 1$. In that case, once a somatic mutation arises in generation k, the VAF of this mutation will stay constant during the whole evolution given by Equation (26). Somatic mutations arising after generation k will always have a lower VAF. For the birth and death case, this is no longer the case since a somatic mutation occurring in generation $k + 1$ may end up with a larger VAF than a somatic mutation occurring in generation k, or earlier, with probability larger than zero. This can also be observed by looking at the sample standard deviations given in Figure 6 showing that there is a large spread of possible VAF-values for each generation. However, by looking at $E[f_i]$ given in Approximation (35), a somatic mutation occurring in generation $k + 1$ is expected to have a smaller

VAF than a somatic mutation occurring in generation k. Having this in mind, one can expect that somatic mutations having VAF larger than $E[f_k]$ occurred in any generation earlier than generation k. Motivated by this observation, let $\hat{M}_N(E[f_k])$ approximate the cumulative number of somatic mutations having VAF equal to or larger than $E[f_k]$:

$$\hat{M}_N(E[f_k]) = \frac{\lambda}{1-\rho}\left(\frac{1}{E[f_k]} - \frac{1}{f_0}\right).$$

$\hat{M}_N(E[f_k])$ may be regarded as the "best guess" based on what is expected to happen. What is nice about $\hat{M}_N(E[f_k])$ is that it is as a function of a frequency and not a specific generation. There is however a problem about $\hat{M}_N(E[f_k])$, namely that $E[f_k]$ given in Approximation (35) is dependent on β, which is unknown in observed data. Despite this inconvenience, as $\hat{M}_N(E[f_k])$ changes inversely proportional to $E[f_k]$, $\hat{M}_N(E[f_k]) \propto 1/E[f_k]$, it is tempting to ask if this is also the case no matter what the VAF is, namely that $\hat{M}_N(f) \propto 1/f$. This finally leads to the following relation:

$$\hat{M}_N(f) = \frac{\lambda}{1-\rho}\left(\frac{1}{f} - \frac{1}{f_0}\right), \tag{39}$$

This relation is the corresponding result to what was found in Williams et al. (2016) for a continuous model. $\hat{M}_N(f)$ is now an estimator for $M_N(f)$, which denotes the cumulative number of somatic mutations having VAF $\geq f$ for a neutrally evolving tumor. Let $M(f)$ in general denote the observed cumulative number of somatic mutations having VAF $\geq f$. $M(f)$ *can* in fact be found from real sequencing data. Furthermore, the accuracy of $\hat{M}_N(f)$ given in Equation (39) can be inspected by looking at how $M_N(f)$ varies as a function of f of simulated tumors. According to Approximation (39), $\hat{M}_N(f) \propto 1/f$. Therefore, $M_N(f)$ should be approximately linear as a function of $1/f$ with slope equal to $\lambda/(1-\rho)$.

3.5.2 Validation of model using simulated data

In order to justify Expression (39), 20 tumors are simulated to reach last generation $j = 150$ for each set of given parameters: $(\beta = 0.6, \lambda = 1.5)$, $(\beta = 0.55, \lambda = 1.5)$, $(\beta = 0.55, \lambda = 1)$ and $(\beta = 0.55, \lambda = 2)$. For each set of parameters, $M_N(f)$ for each tumor is then plotted, see Figure 7. Included in the plot is the Expression $\hat{M}_N(f)$ given in Equation (39) with corresponding parameters. For each tumor simulated, $M_N(f)$ is fitted to a simple linear regression as a function of $1/f$, and the corresponding slope and intercept is estimated by ordinary least squares method[7].

The plots show indeed that $M_N(f)$ is approximately linear as a function of $1/f$. The sample mean of the slope for the region $1/f \in [0, 100]$ for each parameterized neutral tumor is investigated and compared with Expression (39) as shown in Table 1. It shows that the estimated slope given by $\lambda/(1-\rho)$ in $\hat{M}_N(f)$ does not differ very much from the sample mean of the slope. The sample standard deviation also shows that within tumors having the same parameters, the slope does not vary particularly much. According to Expression (39), the slope increases linearly as a function of λ when β is constant. A doubling of λ would therefore double the slope. This is also nearly the case when comparing the two tumors $(\beta = 0.55, \lambda = 1)$ and $(\beta = 0.55, \lambda = 2)$. Likewise, when λ is constant, the slope seems to change as a function of $1/(1-\rho) = \beta/(2\beta - 1)$.

From this analysis, it seems like the slope of $M_N(f)$ for a tumor that evolved neutrally is a good estimator for the expression $\lambda/(1-\rho)$, which Williams et al. (2016) calls *the mutation rate per effective cell division*. However, it is important to remember the approximations that are used in this thesis:

1. The population is approximated to grow as a discrete branching process.

[7] The R function lm() was applied.

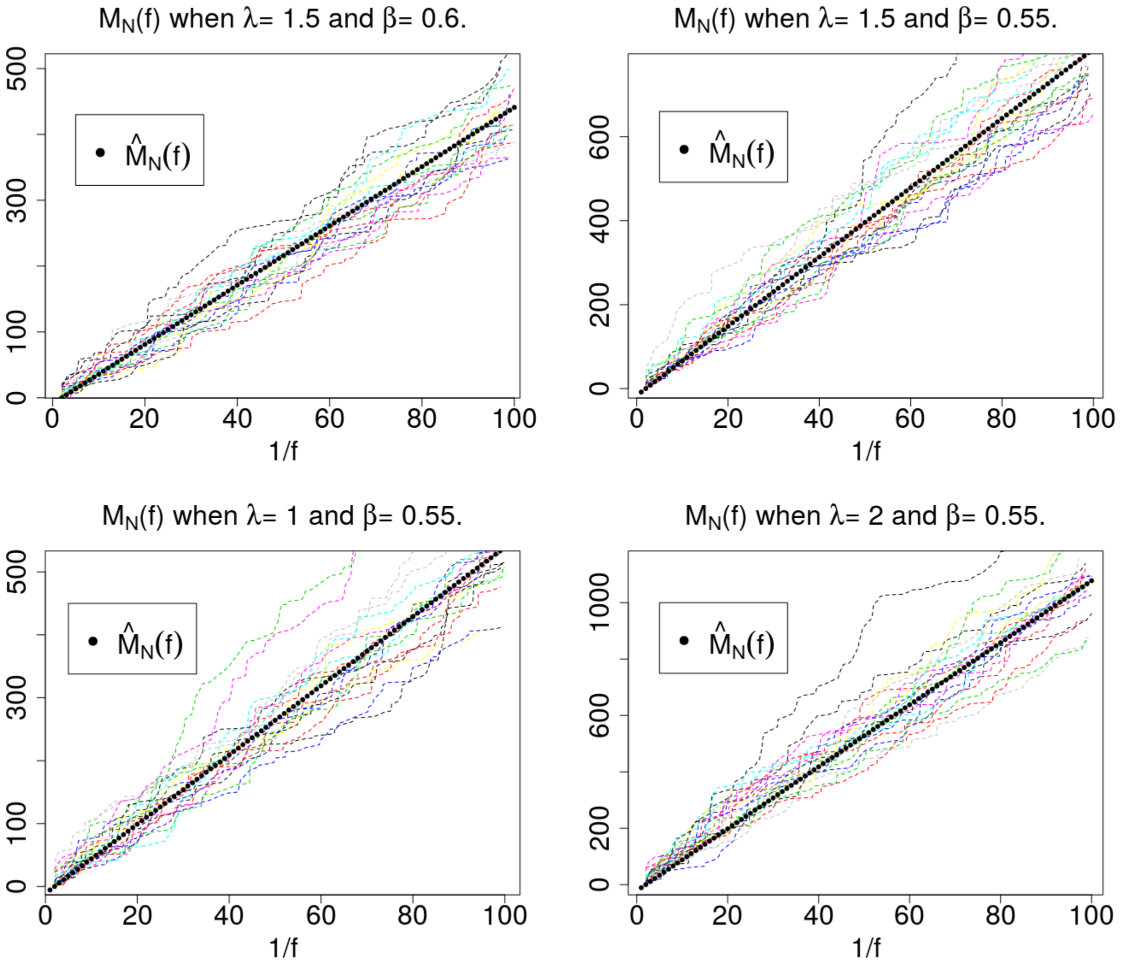

Figure 7: From the simulated tumors, neutral tumor evolution indicates that the function $M_N(f)$ is approximately linear as a function of 1/f. In addition, $\hat{M}_N(f)$ in black dots is a good estimation of $M_N(f)$ in these particular simulations.

β	λ	Sample mean of slope	Sample standard deviation	$\lambda/(1-\rho)$
0.55	1.5	8.03	1.1	8.25
0.6	1.5	4.33	0.5	4.5
0.55	1	5.47	0.9	5.5
0.55	2	10.61	1.4	11

Table 1: Table showing sample mean and estimated slope $\lambda/(1-\rho)$ for 20 tumors simulated reaching generation 150 for different parameters λ and β.

2. $E[f_i^j]$ is approximated to first order, when j is large, given by $E[f_i]$ in Equation (35).

3. The approximation of $m_N(k)$ given in (37) by using Approximation (35) and assuming the cumulative sum given in (33) is negligible.

4. The transition from $m_N(k)$ in (37) to $\hat{M}_N(E[f_k])$.

5. The transition from $\hat{M}_N(E[f_k])$ to $\hat{M}_N(f)$ given in Equation (39).

The first three approximations are based on the assumption of neutral tumor evolution, and according to the analysis summarized in Figure 4 and 6, the approximations seem to be appropriate. The inaccuracies due to the fourth and fifth approximation are more difficult to interpret. Yet, the simulations and corresponding plots in Figure 7 indicates that the slope of $M_N(f)$ still is a good estimate of the mutation rate per effective cell division in a neutrally evolving tumor.

Notice that only somatic mutations appearing after tumor initiation is accounted for when deducing $\hat{M}_N(f)$ as an approximation of $M_N(f)$. In reality, a tumor is complex consisting of both healthy cells and tumor cells, and somatic mutations that appeared both before and after tumor initiation.

4 The observation model

Until now, tumor evolution has only been investigated from a theoretical point of view. In order to test the goodness of our model, it must be applied to data from tissue samples surgically removed from tumors in patients. The tissue is analyzed using DNA sequencing which will be explained briefly below. Afterwards, similar to what is done in Williams et al. (2016), a statistical observation model where the process of DNA sequencing is included will be developed.

4.1 DNA sequencing

DNA Sequencing technology is used to determine the order of nucleotides in the DNA of the cells. The efficiency of the technology has improved a lot over the past decades partly as a result of a transition from sequential to *massively parallel* processing. Modern sequencing technologies are complex, and there are several approaches. The collection of these new technologies goes often by the name "Next-generation sequencing". Usually, the idea is to fragment the genome of the cells into small pieces. The fragments are then processed, and by using sequencing instruments, the order of nucleotide bases in the original fragments is then determined producing *reads*. Due to massively parallel processing, a large amount of reads can be produced simultaneously. Each read is then *mapped* to a *reference genome* which consists of a representative genome in healthy cells. In this way the position of the corresponding fragment in the genome can be inferred. The overlapping of different reads can furthermore be used to infer if a mutation has occurred, and in addition estimate the proportion of cells that has this mutation yielding an *observed* VAF. The number of nucleotide bases that is mapped to a specific locus in the genome is called the *read depth* (Lindner et al. 2013). An illustration summarizing the process is given in Figure 8. For whole-exome sequencing, only the exome of the genome is investigated.

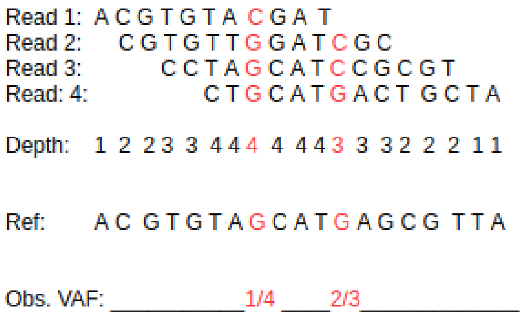

Figure 8: Illustration showing the idea behind sequencing. It also shows two fictive computations of VAF where mutations have occurred.

As was explained, the observed VAFs are only estimates. In order to analyze the goodness of any model, neutral evolution or not, the process from actual VAF-values to observed VAF-values must also be accounted for. This will be explained later on.

4.2 TCGA-data

In this thesis, the observation model developed will be illustrated using Next Generation sequencing data from The Cancer Genome Atlas (TCGA). TCGA is a collaboration between National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). The result is a large database consisting of sequencing data from 33 different types of cancer. All tumors from TCGA have been *resected* in order to provide as much as tissue from the tumor as possible. Some data are public, for instance the *Mutation Annotation Format*(MAF) in the data category *Simple Nucleotide Variation* consisting of estimated VAF and read depth for each discovered mutation at a single nucleotide base. This is the format that will be used in this thesis, therefore other aspects of the tumor such as copy number variations will not be covered. A MAF-file is an already preprocessed file based on raw sequencing data.

Preprosessing and estimating the VAFs from raw sequencing data will not be covered in this thesis. The MAF-files analysed are preprocessed using the *MuTect* method (Cibulskis et al. 2013). In this thesis, only the cancer type Colon Adenocarcinoma (COAD) will be investigated as this is one of the cancer types that are analyzed and compared with a neutral tumor evolution in Williams et al. (2016). Data from this specific cancer type will be denoted as TCGA-COAD data from now on.

4.3 The observation model - Sequencing taken into account

So far, only somatic mutations appearing *after* tumor initiation, so-called subclonal mutations, have been modelled. In real data, somatic mutations appearing *before* tumor initiation, so-called truncal mutations, will also be present in the tumor. The question is then how to distinguish between truncal and subclonal mutations. In addition to this, after resection not only tumor cells will be sequenced, but healthy cells too. How should this be accounted for? A resection may be only a part of the whole tumor. The heterogeneity of a tumor indicates that the geometric position of where the resection is done is also important. How to cope with this? Lastly, when the tumor cells are sequenced, what is the accuracy of estimated VAFs, and what does the accuracy depend on? In this section, these questions will be addressed resulting in a statistical observation model.

4.3.1 Introducing tumor purity and read depth into the model

When sequencing a tissue sample, it is important to recognize that this sample will consist of both healthy cells and tumor cells. However, the neutral tumor evolution model derived in Section 3 does not account for healthy cells, and the VAF for each mutation is only computed among tumor cells. The *tumor purity* is the proportion of tumor cells in a sample. For a tissue sample from a resection consisting of a total of

X cells, a given number of these cells are not tumor cells. In general, let κ denote the tumor purity. Then there are κX tumor cells in the sample. For a somatic mutation with VAF among only tumor cells denoted as $\mathcal{V}$, the real VAF of this somatic mutation in the tissue sample consisting of both tumor cells and healthy cells, VAF_{real}, is given by:

$$\text{VAF}_{\text{real}} = \kappa \mathcal{V}. \tag{40}$$

For a given number of tumor cells, the more healthy cells there are in the sample, the lower the real VAF will be for a given somatic mutation since the VAF is inversely proportional to the total number of cells as can be seen in Definition 1.2.

The real VAF of any somatic mutation can, as explained above, be estimated by DNA sequencing. The observed VAF after sequencing is in addition to the tumor purity also dependent on other parameters. One important parameter is the read depth for each somatic mutation recorded. A read depth of 100 for a given locus, often written 100x, will for example mean that a specific locus within the reference genome is mapped 100 times (100 nucleotide bases is recorded). A somatic mutation observed 10 times will then give an observed VAF of 10/100. The larger the VAF of a somatic mutation is, the more likely it to be observed among the reads. In fact, given a somatic mutation positioned at a specific locus with $\text{VAF}_{\text{real}} = \kappa \mathcal{V}$, the probability for any read covering this locus to include the somatic mutation can be regarded to be equal to VAF_{real}. For a read depth equal to $\mathcal{R}$, assuming the reads are independent, the number of times a somatic mutation will be observed, denoted as $\mathcal{O}$, is then binomial distributed:

$$\mathcal{O} \sim \text{Bin}(\kappa \mathcal{V}, \mathcal{R}).$$

Therefore, the distribution of observed VAFs after sequencing can then be given as:

$$\mathrm{VAF}_{\text{observed}} \sim \frac{\mathcal{O}}{\mathcal{R}}. \tag{41}$$

To link this to what has been done so far, $\mathcal{V}$ is the distribution of subclonal somatic mutations among only tumor cells. One model for this distribution may be the neutral tumor evolution as discussed theoretically in Section 3. What has not been discussed yet is how the read depth varies among the somatic mutations.

4.3.2 Read depth

When investigating sequencing data from TCGA, one observes a large variety in read depth. For instance, for a specific tumor from the cancer type "Colon Adenocarcinoma", Figure 9 shows the distribution of read depth among all the observed somatic mutations.

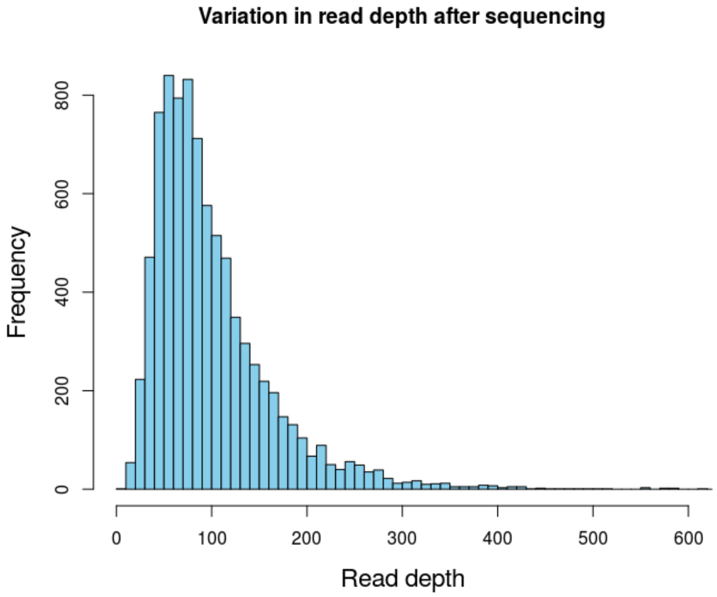

Figure 9: Histogram of observed read depths after sequencing from a TCGA-sample for the cancer type "Colon Adenocarcinoma" (COAD)

This large variation should be accounted for in the model in such a way that the read depth also is a random variable. In fact, based on some strong assumptions a known distribution can be applied. Let $\mathcal{G}$ denote the total number of base pairs of interest in the reference genome[8]. Furthermore, let $\mathcal{L}$ denote the length of each read (segment of nucleotide bases) and assume for now that the length is constant. Lastly, let $\mathcal{N}$ denote the total number of reads from the sequencing. Assume the position where each read is mapped to the genome follows a discrete uniform distribution. The probability for a given read to cover a specific nucleotide base in the reference genome is then $\mathcal{L}/\mathcal{G}$. Let C_b denote the number of times a given nucleotide base is covered, which is what is called the read depth. Assuming that each covering is in fact detected, then:

$$P(C_b = c) = \binom{\mathcal{N}}{c} \left(\frac{\mathcal{L}}{\mathcal{G}}\right)^c \left(1 - \frac{\mathcal{L}}{\mathcal{G}}\right)^{\mathcal{N}-c},$$

which is a binomial distribution. Using Theorem 2.2, assuming $\mathcal{N}$ is large and $\mathcal{L}/\mathcal{G}$ is small which is realistic in this case, the binomial distribution can be approximated by a Poisson distribution, namely[9]:

$$C_b \dot{\sim} \text{Poisson}\left(\frac{\mathcal{N}\mathcal{L}}{\mathcal{G}}\right).$$

The expression $\mathcal{N}\mathcal{L}/\mathcal{G}$ can be seen as the average read depth for each position. An average read depth of around 100 is normal for the tumors sequenced in the "TCGA-COAD"-project. In Figure 9, the average read depth is 133. Looking more closely at Figure 9, in this case the distribution has way more variance than a Poisson distribution, but still has a nice unimodal form. To improve the model further the randomness in both $\mathcal{N}$ and $\mathcal{L}$ should also be accounted for. In fact, by letting $\mathcal{N}\mathcal{L}/\mathcal{G}$ be gamma distributed, one can show that the read depth will follow a Gamma-Poission

[8] For whole-exome sequencing, $\mathcal{G} \sim 10^7$

[9] Read the sign $\dot{\sim}$ as "approximately distributed as".

mixture distribution[10]. This is a common way to approximate the distribution of read depth as for instance proposed in Miller et al. (2011) or Sun et al. (2017). The Gamma-Poisson mixture distribution has two parameters as described in Definition 2.5, the mean and the size parameter. What makes this distribution attractive in this particular case in comparison to the Poisson distribution, is that the variance can be chosen independently from the mean. In order to estimate the parameters in the Gamma-Poisson mixture distribution, one way is to use recorded read depth data to find the maximum likelihood estimates. However, this must be done numerically since there are no closed-form expressions for the corresponding estimators[11].

Equation (41) can then be generalized:

$$VAF_{\text{observed}} \sim \frac{\mathcal{O}}{\mathcal{R}}, \tag{42}$$

where $\mathcal{O} \sim \text{Bin}(\kappa\mathcal{V}, \mathcal{R})$ and $\mathcal{R} \sim \text{GP}(\tau, \psi)$, with mean τ and size ψ.

We now focus on how variation in read depth and tumor purity affect the actual VAF-distribution. What is important to recognize since the read depth is limited in size, is that the lower the VAF is for a mutation, the more likely it is for the mutation to not be observed or underestimated after sequencing. For the case where a tumor evolves neutrally, DNA sequencing with average read depth around 100 would only provide trustworthy values from early stages of the tumor evolution. By recognizing this fact, one can understand how tumor purity affects the actual VAF-distribution. From Equation (40), the lower the tumor purity is, the lower the real VAFs will be, making it even more likely for mutations not to be observed or underestimated. Therefore, a low tumor purity will also give larger inaccuracies in the observed VAFs.

[10] See appendix B for derivation.

[11] In this analysis, the R function *fitdistr* from the R package *MASS* is used to find the MLE-estimates.

One can now visualize the impact of read depth by comparing $M_N(f)$ from a simulated tumor evolving neutrally with $M_N(f)$ from observed VAFs given in Expression (42). This is done by first simulating a tumor with $\beta = 0.55$ and $\lambda = 1.5$[12]. Then, for a given tumor purity equal to $\kappa = 0.7$, $M_N(f)$ for the real VAF-values is plotted by using Equation (40). On the same plot, $M_N(f)$ for the *observed* VAFs given by Expression (42) is also added with mean read depth equal to 80, 90, 100, 110, 120, 130 and 190 respectively. The size parameter is kept constant equal to 2.5. See Figure 10.

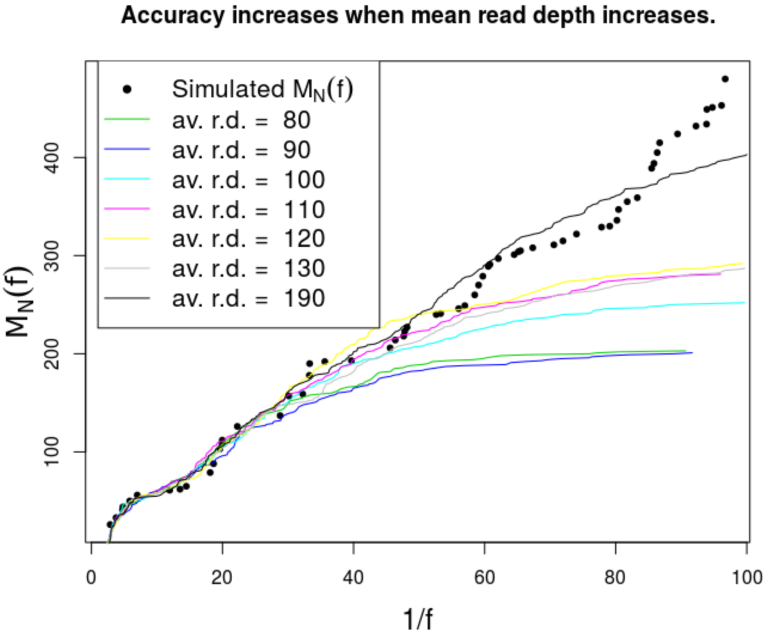

Figure 10: Simulated $M_N(f)$ from tumor evolving neutrally as black points versus observed $M_N(f)$ in colours after observation model given in Expression (42). "av. r.d" means average read depth.

[12] The tumor reaches generation $j = 150$. Remember that the size of the tumor does not matter in the case of neutral tumor evolution.

Figure 10 clearly shows that the larger the read depth, the more accurate is the observed $M_N(f)$. However, even for a large read depth equal to 130, VAFs less than or equal to 0.016 corresponding to 1/VAF equal to 60 or more, the observed $M_N(f)$ clearly underestimates the true $M_N(f)$. For read depth equal to 80, the observed $M_N(f)$ is only close to the correct $M_N(f)$ for VAF larger than around 0.05. Nothing has so far been said about the last parameter of the Gamma-Poisson mixture distribution, namely the size parameter. By Definition 2.5, the variance is inversely proportional to the size, and therefore the variance decreases when the size increases. The lower the variance is, the more positions across the reference genome will have read depth around the mean, and for that reason the observed $M_N(f)$ will be more accurate the larger the mean read depth and size-parameter is. However, from the TCGA-data the estimated size parameter, ψ, from different tumors do not differ very much as it usually varies around from $\psi \approx 2$ up to around $\psi \approx 3$. On the other hand, the mean value differs very much from $\tau \approx 70$ to $\tau \approx 200$. This shows a particularly important characteristic of the read depth distribution, as is the case for the Poisson distribution, that the variance increases when the mean increases. However, as there is a linear relation for the Poisson distribution, the variance increases faster for the read depth distribution[13]. The most important parameter in this model is therefore the mean read depth from the sequencing.

The read depth is important when considering the observed VAF-values. If the mean read depth is too small, then many low-frequent mutations will not be detected. As a measure showing the accuracy of the sequencing given that the distribution of read depth is Gamma-Poisson mixture distributed, consider the probability $P(\mathcal{O} = 0)$, where $\mathcal{O} \sim \text{Bin}(\kappa\mathcal{V}, \mathcal{R})$. $P(\mathcal{O} = 0)$ is the probability for a somatic mutation to not be

[13] The variance increases quadratically in terms of the Gamma-Poisson mixture distribution for instance as can be seen in Definition 2.5 on page 11.

observed. Using the law of total probability and conditioning on the real VAF value, $VAF_{real} = \kappa \mathcal{V} = v$:

$$P(\mathcal{O} = 0 | \, \mathrm{VAF}_{real} = v) = \sum_{r=0}^{\infty} P(\mathcal{O} = 0 | \mathcal{R} = r, \mathrm{VAF}_{real} = v) P(\mathcal{R} = r).$$

From Definition 12, $P(\mathcal{O} = 0 | \mathcal{R} = r, \mathrm{VAF}_{real} = v) = (1 - v)^r$ and $\mathcal{R}$ is Gamma Poisson distributed with mean τ and size ψ as given in Definition 2.5. Using that $(1 - v)^r = e^{r \log(1-v)}$, this shows that:

$$P(\mathcal{O} = 0 | \, \mathrm{VAF}_{real} = v) = \sum_{r=0}^{\infty} e^{r \log(1-v)} P(\mathcal{R} = r) = E[e^{\mathcal{R} t}] = M_{\mathcal{R}}(t),$$

where $M_{\mathcal{R}}(t)$ denotes the moment generating function (MGF) for the distribution of $\mathcal{R} \sim \mathrm{GP}(\tau, \psi)$ with $t = \log(1 - v)$. The MGF for the Gamma-Poisson mixture distribution can be derived[14] and is given by,

$$M_{\mathcal{R}}(t) = \left(\frac{\frac{\psi}{\tau + \psi}}{1 - (1 - \frac{\psi}{\tau + \psi}) e^t} \right)^{\psi},$$

where $t < -\log(1 - \frac{\psi}{\tau + \psi})$ in order for the MGF to exist. For $t = \log(1 - v)$ this is always satisfied, and therefore the probability for a mutation to not be observed is then given by:

$$P(\mathcal{O} = 0) = M_{\mathcal{R}}(\log(1 - v)) = \left(\frac{\frac{\psi}{\tau + \psi}}{\frac{\psi}{\tau + \psi} + \frac{\tau}{\tau + \psi} v} \right)^{\psi}. \tag{43}$$

Setting $v = 0$, corresponding to a non-existing mutation, $P(\mathcal{O} = 0) = 1$, as expected. Setting $v = 1$, corresponding to a truncal mutation gives $P(\mathcal{O} = 0) = P(\mathcal{R} = 0)$, which is the probability that the mutation exists, but no reads cover the position of the mutation. A graph of $P(\mathcal{O} = 0)$ for $\psi = 2.5$ and $\tau = 100$ is given in Figure 11.

[14] See Appendix C.

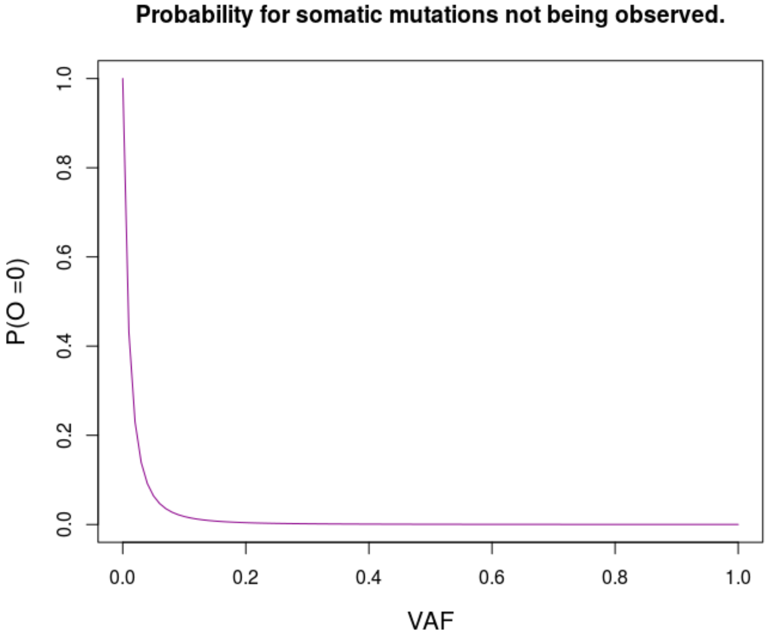

Figure 11: $P(\mathcal{O} = 0|\mathcal{R})$ when $\mathcal{R}$ is Gamma-Poisson mixture distributed with mean, $\tau = 100$, and size, $\psi = 2.5$.

This expression can now be used to see how low VAF one can get before exceeding a limit ϵ for which $P(\mathcal{O} = 0) > \epsilon$. Define $p_r = \frac{\psi}{\tau+\psi}$, then:

$$P(\mathcal{O} = 0) = \left(\frac{p_r}{p_r + (1-p_r)v}\right)^{\psi} > \epsilon \iff v < \frac{p_r(1-\epsilon^{1/\psi})}{(1-p_r)\epsilon^{1/\psi}}. \tag{44}$$

For instance, for $\psi = 2.5$, $\tau = 130$ and $\epsilon = 0.01$, the probability for not observing a mutation to be greater than $\epsilon = 0.01$, is then $v < 0.1$. Equation (43) can be used as a tool for deciding a lower bound for the observed VAFs. For instance, from Figure 11, $P(\mathcal{O} = 0)$ increases fast for $\text{VAF}_{\text{observed}} \leq 0.1$. Based on this observation, it would for instance be reasonable to only look at $\text{VAF}_{\text{observed}} \geq 0.1$.

4.3.3 Truncal somatic mutations

So far, only somatic mutations appearing *after* tumor initiation have been studied. What has not been evaluated yet is somatic mutations that appear *before* tumor initiation. These will be present in all tumor cells and are called *truncal mutations*[15]. As the interest is in tumor evolution, truncal mutations arising before tumor initiation are of no interest to us. However, the VAF-distribution after sequencing will be as a result of both subclonal and truncal mutations. Therefore, it is important to understand how to distinguish between these two type of mutations in order to find information that is only relevant for tumor evolution.

Assuming a constant copy number for all genes equal to the ploidy, truncal mutations already present from the first tumor cell will have VAF among only tumor cells equal to $\mathcal{V} = 0.5$. Assuming both the read depth, $\mathcal{R}$, and the tumor purity, κ, is known, the *observed* VAF of a truncal mutation is then given by:

$$\text{VAF}_{observed} \,|(\mathcal{V} = 0.5, \mathcal{R}, \kappa) \sim \text{Bin}(0.5\kappa, \mathcal{R})/\mathcal{R}, \tag{45}$$

where again $\mathcal{R} \sim \text{GP}(\tau, \psi)$. In other words, the VAF-distribution will then consist of a mixture of truncal mutations where $\mathcal{V} = 0.5$, and subclonal mutations where the distribution of $\mathcal{V}$ depends on how the tumor evolves.

4.4 Comparing observation model with observed data

As both DNA sequencing, tumor purity and truncal mutations has been described in the observation model, the theoretical shape of a VAF-distribution after sequencing a sample from a tumor that has evolved neutrally may now be visualized. In this

[15] Somatic mutations occurring after tumor initiation may also become present in all cell. See for instance Bozic et al. (2016)

example, a tumor is simulated to reach generation 150 in the branching tree. Because neutral tumor evolution behaves the same during the whole process, there is no need to simulate a particularly large tumor as the pattern remains the same, as described before. After having the original VAFs, VAF_{real}, before sequencing from the subclonal somatic mutations in the simulated tumor, a specified proportion of truncal mutations having $\mathcal{V} = 0.5$ is added representing somatic mutations being present before tumor initiation. The sample is then sequenced as described in Equation (42) yielding observed VAFs. The result with tumor purity $\kappa = 0.63$, mean read depth $\tau = 190$, and size parameter $\psi = 2.5$ is given in Figure 12 where it has been chosen that 50 % of the somatic mutations are truncal.

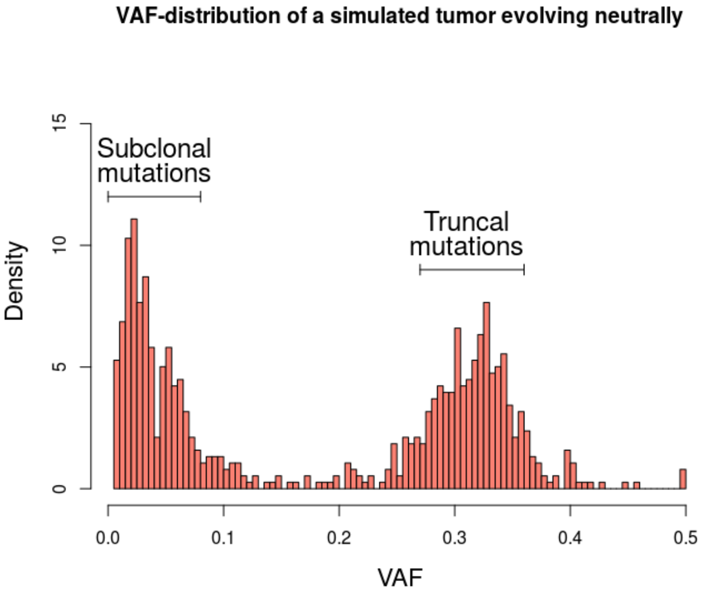

Figure 12: A fictive VAF-distribution from the observation model with $\kappa = 0.63$, $\tau = 188$, $\psi = 2.4$ and where 30 % of the somatic mutations are truncal. One clearly sees two peaks, one representing subclonal mutations to the left, and one representing truncal mutations to the right.

Figure 12 shows two peaks, where the peak to the left in the plot is a result of subclonal somatic mutations after neutral evolution, while the distribution to the right is a result of truncal somatic mutations arising before tumor initiation. When the observed VAF is approaching zero, one can see from the figure that the number of observed somatic mutations decreases, even though it should increase according to a neutral evolution model. However, as explained earlier, since the read depth is limited in size, somatic mutations with low VAF may be unobserved or at least underestimated. This observation is exactly the same as what can be seen in Figure 10. The peak representing the truncal somatic mutations seems to be quite symmetric. In addition, the two peaks are clearly separated from each other.

The theoretical shape of VAF-distribution from Figure 12 may now be compared with the observed VAF-distribution from a tissue sample coming from the TCGA-COAD project. An example is given in Figure 13.

In this case, it also shows a distribution with two peaks. According to the observation model, the peak at the right should represent truncal mutations, while the peak at the left should represent subclonal mutations. It also shows that the observed number of somatic mutations decreases when the VAF is approaching zero, just as predicted. Therefore, in this specific case the observation model described in Section 4 reflects core aspects with the observed VAF-distribution with real data in terms of equal shape in distribution. However, the peaks are not clearly separated from each other, but are rather intertwined. The mean read depth for this data set is 178.

There are also samples in TCGA-COAD that does not show a VAF-distribution consisting of two peaks, but only one peak. Even though this does not fit with a neutral tumor evolution, the reason may rather be either a low tumor purity, a low mean read depth or even a combination. In order to examine the evolution of the tumor, both the tumor purity and the mean read depth must be sufficiently large such that

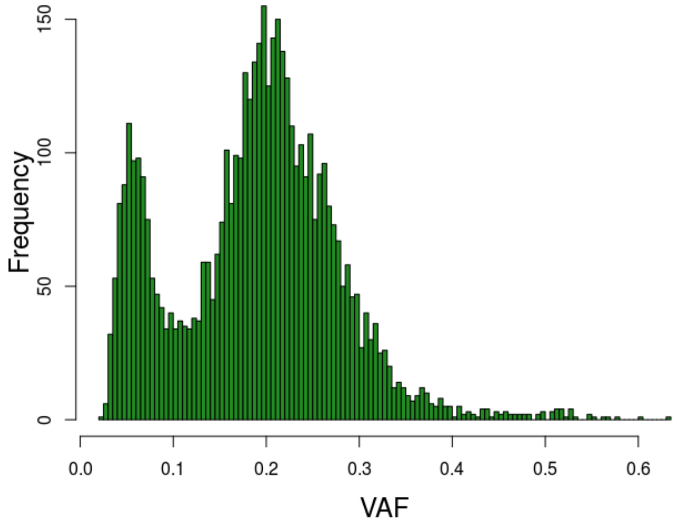

Figure 13: Histogram of observed VAFs from a TCGA-COAD sample showing two peaks, and therefore a candidate for neutral tumor evolution. According to the observation model, the one to the right represents truncal mutations, while the one to the left represents subclonal mutations.

low VAF-values can be examined. In addition to this, it is also important to have enough data in terms of observed somatic mutations. A large amount of the samples from TCGA-COAD consists of too few observed somatic mutations in order to infer anything about tumor evolution.

5 Discussion

In order to derive $\hat{M}_N(f)$ given in Equation (39) for neutral tumor evolution, there are several assumptions, simplifications and approximations made. These will be discussed thoroughly in this section.

Furthermore, even though the observation model for neutral tumor evolution described in this thesis can explain a pattern that resembles certain observed VAF-distributions from real data, this does not prove that a tumor actually evolved neutrally. This will also be discussed further.

5.1 Assumptions and simplifications

5.1.1 Constant exponential growth

According to the branching process model developed in Section 3, a tumor can in theory grow exponentially to infinite size. However, as the tumor grows it will eventually be exposed to external pressure from surrounding tissue or organs, and the need for oxygen and nutrients increases (Nishida et al. 2006). These factors should eventually reduce the tumor's growth. It may however be realistic that the tumor can grow approximately exponentially for a time period after tumor initiation, and therefore $\hat{M}_N(f)$ given in Equation (39) may be a realistic estimation of $M_N(f)$ at this moment. Assuming this is the case, a tumor may be detected during exponential growth or after exponential growth. If the tumor is detected after exponential growth, the question is then if $\hat{M}_N(f)$ will still be a good estimator of $M_N(f)$. The answer depends on how the growth varies during evolution, but with the following assumption the answer is yes: During evolution, after an exponential growth, the probability of cell division β per tumor cell can change as time goes by, but at any time point the probability of cell division is equal for all cells.

As an example where this assumption is applied, let the expected number of offsprings per cell in any generation i be dependent on the total number of cells, X_i, of the tumor at that moment given by $\mu(X_i)$. As the tumor grows approximately exponentially at the beginning, this will mean that $\mu(X_i) \approx \tilde{\mu}$ for some constant $\tilde{\mu}$ as long as X_i is small enough. During this evolution, $\hat{M}_N(f)$ will be a good estimator of $M_N(f)$ as can be seen in Figure 7. Furthermore, after exponential growth, consider a somatic mutation appearing in any generation i out of total X_i cells with corresponding VAF equal to $S/(\pi X_i)$. The expected VAF to first order approximation in the next generation will then be $S\mu(X_i)\Big/(\pi X_i \mu(X_i)) = S/(\pi X_i)$, resulting in the same VAF as before. Therefore, if the assumption holds, somatic mutations arising during exponential growth will relatively to each other produce a pattern similar to $\hat{M}(f)$ in Equation (39).

Letting the expected number of offsprings vary as a function of number of cells, tumor evolution can be modelled more realistically. For instance, in a logistic-growth model the tumor size eventually stabilizes, but this will again require more parameters making the model more complicated. In addition, it is reasonable that the growth of the tumor is dependent on more than the size of the tumor such as local surrounding pressure or access to oxygen and nutrients. In fact, there exist experimental data that support larger growth at the surface of the tumor than in the tumor's core (Waclaw et al. 2015). This has led to models based on peripheral growth, as can be seen in Sun et al. (2017). Despite the fact that neutral tumor evolution does not take any of these factors into account, nevertheless it may be a realistic model during early growth, and therefore applicable for early detected tumors.

5.1.2 A tissue sample from a tumor a good representation of the whole tumor

The TCGA-data are based on a sample resected from a tumor. However, the interest is to find out how the whole tumor evolved. Can a tissue sample consisting only of a part of the tumor represent the whole tumor? This question is again dependent on how the tumor evolved as analyzed in Sun et al. (2017), where virtual tumors both evolving neutrally and with selection were simulated in 3D, and virtual resections were taken on different geometrical regions. The analysis showed that simulated VAF-distributions from different resections of the tumor were more similar to each other for neutrally evolving tumors, and tumors with weak selection, than from tumors with strong selection. There is a reasonable explanation for this. In the case of neutral tumor evolution, all cells have the same fitness providing a somewhat predictive hierarchy of subclones, namely that mutations that arise early tend to have larger VAF than mutations that arise at a later stage, no matter the geometric position of the tumor. On the other hand, for a tumor with strong selection during its evolution, later arising mutations are more likely to get larger VAF than mutations that arose earlier due to the potential impact of driver mutations, increasing the fitness of some cells. It is therefore reasonable that there is a larger variation in VAF-distribution between samples from different geometric positions for tumors having strong selection, than in tumors that have weak selection or even neutral growth. What is interesting about the analysis of Sun et al. (2017) is that this may be one way of investigating if a tumor evolved neutrally or not by taking multiple resections from different regions of the tumor.

5.1.3 Ignoring copy number changes

In this thesis, copy number changes has not been taken into account, as can be seen in Section 3 where it is regarded that there are always π copies for each locus in each tumor cell. However, in reality genes or even whole chromosomes may be duplicated or deleted in some tumor cells. This would then affect the observed VAF. For instance, if there are more than π copies of some specific mutation in a given proportion of cells, the probability of detecting this mutation during sequencing would be larger than if there were only π copies in each cell. Therefore, copy number changes would violate with the deduction of $\hat{M}_N(f)$ in Equation (39). One way to handle this is to assume that π is the average number of copies for each locus as was done in Williams et al. (2016). This would then give the same equation of $\hat{M}_N(f)$ as in Equation (39). However, randomness in the copy number for each mutation would make even more variation in $M_N(f)$ for a neutrally evolving tumor.

There exist however computational methods to infer copy number variation for the observed mutations in the tissue sample (Zhao et al. 2013). In Williams et al. (2016) copy number changes are taken care of by only looking at diploid regions.

5.2 Can neutral tumor evolution be inferred from observed data?

As shown in this thesis, given the branching process model of neutral tumor evolution developed in Section 3, $\hat{M}_N(f)$ in Equation (39) was deduced to estimate $M_N(f)$ using several assumptions and approximations, where $M_N(f)$ is the cumulative number of somatic mutations in the tumor having VAF larger than f for a neutrally evolving tumor. It was shown that $\hat{M}_N(f) \propto 1/f$. The interest is now to look at the following situation: given $M(f) \propto 1/f$ from observed data, is it possible to infer if a tumor grew neutrally? Observe that $M(f)$ is the number of somatic mutations having VAF $\geq f$

for *any* evolving tumor, while $M_N(f)$ is specifically for a neutrally evolving tumor.

The alternative theory to neutral tumor evolution is clonal evolution, namely that driver mutations may arise during the evolution changing the fitness of certain tumor cells. These cells then may grow faster than surrounded tumor cells, creating subclones that may grow larger than subclones that arised earlier. However, the importance of a driver mutation should naturally depend on when it arises and how much it changes the fitness. For instance, it is reasonable to think that the earlier the driver mutation occurs, and the more it changes the fitness of a cell, the more impact the driver mutation will have on the evolution of the tumor. In weak selection, driver mutations have small impact on tumor evolution, and should therefore resemble neutral tumor evolution. In this case, it will be difficult to distinguish between neutral tumor evolution from tumor evolution with weak selection. An example is given in Tarabichi et al. (2017), where simulated tumors consisting of driver mutations were classified as a tumor evolving neutrally. Therefore, one should rather try to establish a hypothesis test inferring if a tumor evolved neutrally or with weak selection on one side, or if a tumor evolved with strong selection on the other side.

In order to establish a hypothesis test, it is important to recognize that $\hat{M}_N(f)$ deduced in this thesis, is a result of taking the expected value of all random variables. This means that $\hat{M}_N(f)$ in this thesis is based on the fact that everything is evolving as expected. However, from a statistical point of view, there can be a large difference between the value of a single observation and its expected value depending on the variance of the associated distribution. In our case, the single observation is the observed VAF for each detected somatic mutation, and by looking at Figure 6, there is a large variation of which VAF each somatic mutation acquire. Furthermore, by looking at the tumors evolving neutrally in Figure 7, there are some tumors where $M_N(f)$ does not appear particularly linear, but most of them in fact do in *average*.

$\hat{M}_N(f)$ is a result of averaging all tumors as seen in Figure 7. As can be seen from Table 1, there is also small variations in computed slopes from linear regression between different tumors with equal parameters, as can be seen by looking at the respective sample standard deviations for each set of parameters. From this point of view, the idea is to exploit the property of $\hat{M}_N(f)$ such that $M(f)$ from observed data of a tumor could be classified as a neutrally evolving tumor if $M(f)$ is proportional to $1/f$ with a given accuracy.

However, $\hat{M}_N(f)$ is deduced before the tumor is observed from DNA sequencing, and the presence of truncal mutations is not yet regarded. As can be seen in Figure 10, DNA sequencing affects the observation of $M_N(f)$ in such a way that linearity is violated for small VAF-values due to the increase of uncertainties when the VAF is getting too low. In addition, as can be seen in Figure 13, truncal mutations may have observed VAFs in the same region as subclonal mutations, creating an intermix of both subclonal and truncal mutations.

In Williams et al. (2016), this is dealt with by choosing a specific frequency interval of $f \in [0.12, 0.24]$ in the observed distribution of $M(f)$ for all tumors, with the purpose of making sure that too small VAF-values are discarded due to uncertainty in DNA sequencing, but also to make sure that truncal mutations are not included within the interval. The idea in Williams et al. (2016) is then to evaluate the linearity of $M(f)$ as a function of $1/f$ by using a simple linear regression where $f \in [0.12, 0.24]$, and then the coefficient of determination, often denoted as R^2, is used as the hypothesis test to infer whether a tumor evolved neutrally or not.

First and foremost, the problem about R^2 is that it is rather a measure of how better a prediction of a linear model is than by predicting a sample to be equal to the average of all samples[16]. In order to validate linearity of a function, one should

[16] See appendix about coefficient of determination

therefore not use R^2 for this purpose. Secondly, as $M(f)$ is a non-decreasing function anyway, R^2 will naturally tend to be closer to 1 than zero for a growing tumor. In Williams et al. (2016), this is accounted for by declaring tumors as neutral if $R^2 \geq 0.98$. However, as pointed out in Tarabichi et al. (2017) and also agreed upon by the authors of Williams et al. (2016) in their response to this paper (Williams et al. 2017), not rejecting a null hypothesis is not the same as declaring the null hypothesis to be true.

Therefore, by the property of $\hat{M}_N(f)$, namely that $\hat{M}_N(f) \propto 1/f$, one should rather develop the following hypothesis test:

$$H_0 : M(f) \propto 1/f$$

$$H_1 : M(f) \not\propto 1/f$$

Be that as it may, this is not a regular hypothesis test, as it is inferring whether a *function*, $M(f)$, behaves linearly as a function of $1/f$. In addition, due to the impact of sequencing and truncal mutations aforementioned, one is not free to choose in which interval to investigate linearity. In Williams et al. (2016), this is dealt with by choosing a specific interval $f \in [0.12, 0.24]$. This is a rather short interval of frequencies, and it is important to recognize that a test for linearity will improve the smaller the interval is, since any function will appear more linear in an interval, the smaller the interval is.

By this reasoning, one would like a larger interval of investigation than $[0.12, 0.24]$ in order to infer the evolution of the tumor, but at the same time assure that the uncertainties in observed VAFs are as low as possible within this interval. Increasing the interval from above would be difficult, as this increases the probability of including truncal mutations. Therefore, the simplest way to increase the interval would be to decrease the lower value of the interval. It is here the measure given in Equation

(43) comes in handy. If the read depth can be approximated to be a Gamma-Poisson mixture distribution, Equation (43) can then be used as a measure in order to grasp when the inaccuracies in observed VAFs begin to be considerable. From Figure 11, for a mean read depth of $\tau = 100$ and size parameter equal to $\psi = 2.5$, one can see that the inaccuracies explode for VAF ≤ 0.1. In this case, the lower limit of Williams et al. (2016) is in fact reasonable. As there are many observed data in TCGA that actually have a mean read depth of around 100, one is therefore only able to see a short period of the evolution of the tumor from these observed data. In order to investigate tumor evolution with more precision, and to conclude whether a tumor evolved neutrally or not, there is a need for a larger mean read depth than what is the case by now.

6 Conclusion

In this thesis, using a branching process with accumulating mutations to model neutral tumor evolution, it was shown in details how to deduce the same expression as deduced in Williams et al. (2016), namely an expression for the average number of somatic mutations having VAF $\geq f$, denoted here as $\hat{M}_N(f)$. Using stochastic simulation, the expression was validated.

Afterwards, a statistical model taking DNA sequencing into account was developed, and was used to investigate the pattern of observed VAFs from a tumor that evolved neutrally. An expression in order to grasp the inaccuracies of low-frequent VAFs was developed.

Lastly, inferring neutral tumor evolution from observed data was discussed, and it was argued that the observed data need a high accuracy in terms of larger mean read depth in order to properly investigate the evolution of tumor.

A Motivating the choice of a discrete birth and death branching process to model neutral tumor growth

As the cells are regarded to have equal fitness, each cell is expected to produce an equal amount of offsprings given by μ. By Equation (8), the expected number of cells, X_i, in generation i, given μ is then given by:

$$E[X_i] = (2\mu)^i.$$

By the fact of equal fitness, the cells are also expected to reproduce at the same rate. Consider the special case where each cell after birth takes a decision to either divide or die after some time $\mathcal{W}$, called waiting time. In this case, the expected number of cells, $X(i\mathcal{W})$ after time $i\mathcal{W}$, is then:

$$E[X(i\mathcal{W})] = (2\mu)^i. \tag{46}$$

Therefore, in this special case, the number of cells in a given generation is the expected number of cells after a specific discrete time point.
However, in reality, the cells will not use the same amount of time to either divide or die, but with equal fitness the difference in waiting time from cell to cell is regarded to be small. Therefore Equation (46) will then be an approximation for the total number of cells at a given time point.

B Derivation of Gamma-Poisson mixture distribution

Let $f(k|\lambda)$, $k \in [0, 1, 2, ...]$, be Poisson distributed with rate λ, namely that:

$$f(k|\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}.$$

Furthermore, let the parameter λ, $\lambda \in \mathbb{R}^+$, be Gamma distributed with shape, k, equal to r and scale, θ, equal to $p/(1-p)$, namely:

$$g(\lambda|k = r, \theta = p/(1-p)) = \frac{1}{\Gamma(r)(\frac{p}{1-p})^r}\lambda^{r-1}e^{-\lambda(1-p)/p}.$$

This is an example of a *compound probability distribution* which in general is a random variable following a parameterized distribution where some of the parameters are themselves random variable. The *unconditional* distribution of $f(k)$ is then given by:

$$\begin{aligned} f(k) &= \int_0^\infty f(k|\lambda)g(\lambda|k = r, \theta, \theta = p/(1-p))d\lambda \\ &= \int_0^\infty \frac{\lambda^k}{k!}e^{-\lambda}\left(\frac{1}{\Gamma(r)(\frac{p}{1-p})^r}\right)\lambda^{r-1}e^{-\lambda(1-p)/p}d\lambda \\ &= \left(\frac{1}{k!\Gamma(r)(\frac{p}{1-p})^r}\right)\int_0^\infty \lambda^{k+r-1}e^{-\lambda/p}d\lambda \\ &= \left(\frac{p^{k+r-1}}{k!\Gamma(r)(\frac{p}{1-p})^r}\right)p\int_0^\infty x^{k+r-1}e^{-x}dx \\ &= \left(\frac{p^{k+r}}{k!\Gamma(r)(\frac{p}{1-p})^r}\right)\Gamma(k+r) \\ &= \frac{\Gamma(r+k)}{\Gamma(r)k!}p^k(1-p)^r, \end{aligned} \tag{47}$$

where the substitution $x = \frac{\lambda}{p}$ is applied in the integral yielding an integral that by definition is the Gamma distribution, $\Gamma(z)$, given by:

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx.$$

As can be seen from Definition 2.5 on page 10, $f(k)$ is a Gamma-Poisson mixture distribution.

C Derivation of MGF for Gamma-Poisson mixture distribution

Given a Gamma-Poisson mixture distribution, $f(x)$, where $x \in \{0, 1, ...\}$, $p \in [0, 1]$ is the probability for success and $\psi \in \mathbb{R}^+$ is the size, the density function is:

$$f(x) = \frac{\Gamma(x+\psi)}{\Gamma(\psi)x!} p^{\psi}(1-p)^x.$$

It can be seen that $\sum_{x=0}^{\infty} f(x)$ sums to one using the following Taylor series for any ψ and $|k| < 1$:

$$\frac{1}{(1-k)^{\psi}} = \sum_{x=0}^{\infty} \frac{\Gamma(x+\psi)}{\Gamma(\psi)x!} k^x. \tag{48}$$

The MGF of the Gamma-Poisson mixture distribution, $E[e^{tx}]$, can also be derived using this infinite sum:

$$\begin{aligned} E[e^{tx}] &= \sum_{x=0}^{\infty} e^{tx} \frac{\Gamma(x+\psi)}{\Gamma(\psi)x!} p^{\psi}(1-p)^x = p^{\psi} \sum_{x=0}^{\infty} \frac{\Gamma(x+\psi)}{\Gamma(\psi)x!} \left(e^t(1-p)\right)^x \\ &= \frac{p^{\psi}}{(1-e^t(1-p))^{\psi}}, \end{aligned} \tag{49}$$

conditioning that $e^t(1-p) < 1$, which means that $t < -\log(1-p)$.

D Coefficient of determination

The following notation is in correspondence with Walpole et al. (2014).

Given explanatory variables $x_1, x_2, ..., x_n$ and response variables $y_1, y_2, ..., y_n$ in addition to a simple linear regression model, $\hat{y}_i = ax_i + b$, the *coefficient of determination*, $R^2 \in [0, 1]$, is given by:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}, \tag{50}$$

where SSE is the *sum of square errors*:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

while SST is the *total corrected sum of squares*:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. While SSE is a natural measure for the error of the simple linear regression model, SST should be seen as the corresponding error of the model where $y_i = \bar{y}$, namely a model where the response value, y_i, is constant and independent of the explanatory variable x_i. R^2 will therefore be close to 1 if a non-constant linear model fits better than a constant linear model and vice verca.

E Biological definitions in Cell biology

Allele A variant form of a gene.

Copy number variation A phenomenon where sections of the genome are repeated several times. This is natural, but it may also be as a result of somatic mutations.

Chromosome A double-strand of DNA situated in the cell nucleus encoded with genes. In humans, the somatic cells consist of 22 pairs of chromosomes plus two sex chromosomes.

DNA-replication A process where a double-stranded DNA is copied, producing two identical DNA-molecules. This process occurs prior to cell division.

DNA sequencing The process where the precise order of nucleotides within a DNA molecule is measured.

Driver mutation A mutation that alters the fitness of the cell.

Exome The complete set of regions in the genome that code information for protein synthesis. Each region is situated within a particular gene called and exon.

Fitness A measure of reproductive success.

Genome The complete set of genetic material in a cell.

ISM *Infinite Sites Model.* See for instance Kimura (1968).

ITH *Intratumor heterogeneity.* Large genetic variations within a tumor due to somatic mutations.

Locus The position of a gene or mutation on a chromosome.

Mapping The process of comparing a read with a reference genome in order to find the position in the reference genome that is most similar to the read.

Massively parallel Technical term used in computing to denote the use of a large number of processors in order to do coordinated computations in parallel.

Mutation The process of which the structure of a gene is changed due to rearrangements of one or more base units in the DNA.

Neutral tumor evolution theory See Definition 1.1 on page 2.

Next generation sequencing A modern sequencing technology where massive amounts of DNA-fragments are sequenced in parallel.

Nucleotide A compound consisting of a nucleoside and a phosphate group. The structural unit in DNA.

Passenger mutation A mutation not altering the fitness of the cell.

Point mutation A mutation only altering one base unit.

Ploidy The number of sets of chromosomes in a cell, often denoted by π. In normal human cells, there are two sets of each autosomal (non-sex) chromosome and so $\pi = 2$.

Proliferation Rapid reproduction of a cell.

Read depth The number of unique reads covering a particular locus after DNA sequencing. See drawing on page 39.

Reference genome A digital nucleic acid sequence database consisting of a representative set of genes for a given species.

Resection Surgical removal of part or all of a damaged organ or structure. The term is often used for removal of a tumor.

Selection A consequence due to the fact that individuals that adapt better to the environment, tend to have better chances of surviving in addition to produce surviving offsprings. Individuals may also be cells in this context.

Somatic cell Any cell in an organism other than sex cells (germ cells).

Somatic mutation A mutation that can occur in any cells of the body except for the germ cells.

Subclonal mutation A mutation in the tumor that is found in only some of the tumor cells.

Variant allele frequency (VAF) See Definition 1.2 on page 2.

Tumor initiation The process in which normal cells are transformed to cells capable of creating tumors.

Tumor purity The proportion of cells in a tumor sample that consists of tumor cells.

Truncal mutation A mutation that is found in all tumor cells.

References

Bozic, I., Gerold, J. M. & Nowak, M. A. (2016), 'Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution', *PLOS Computational Biology* **12**(2), e1004731.
URL: *https://doi.org/10.1371/journal.pcbi.1004731*

Casella, G. & Berger, R. L. (2002), *Statistical inference*, 2 edn, Brooks/Cole.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S. & Getz, G. (2013), 'Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples', *Nature Biotechnology* **31**, 213.
URL: *https://www.nature.com/articles/nbt.2514*

Cooper, G. M. (2000), *The Cell: A Molecular Approach*, 2 edn, Sinauer Associates.

Davis, A., Gao, R. & Navin, N. (2017), 'Tumor evolution: Linear, branching, neutral or punctuated?', *Biochimica et Biophysica Acta* **1867**(2), 151–161.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5558210/*

Gerashchenko, T. S., Denisov, E. V., Litviakov, N. V., Zavyalova , M. V., Vtorushin, S. V., Tsyganov, M. M., Perelmuter, V. M. & Cherdyntseva, N. V. (2013), 'Intratumor heterogeneity: Nature and biological significance', *Biochemistry (Moscow)* **78**(11), 1201–1215.
URL: *https://doi.org/10.1134/S0006297913110011*

Grinstead, C. M. & Snell, J. L. (2003), *Introduction to probability*, 2 edn, AMS.

Jones, S., Chen, W.-D., Parmigiani, G., Diehl, F., Beerenwinkel, F., Antal, T., Traulsen, A., Nowak, M. A., Siegel, C., Velculescu, V. E., Kinzler, K. W., Vo-

gelstein, B., Willis, J. & Markowitz, S. D. (2007), 'Comparative lesion sequencing provides insights into tumor evolution', *Proceedings of the National Academy of Sciences of the United States of America* **105**(11), 4283–4288.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2393770/*

Kapp, R. O. (1958), 'Ockam's razor and the unification of physical science', *The British Journal for the Philosophy of Science* **8**(32), 265–280.
URL: *http://www.jstor.org/stable/685852*

Karr, A. F. (1993), *Probability*, Springer-Verlag New York.

Kimura, M. (1968), 'The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations', *Genetics* **61**(4), 893–903.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1212250/*

Lindner, M. S., Kollock, M., Zickmann, F. & Renard, B. Y. (2013), 'Analyzing genome coverage profiles with applications to quality control in metagenomics', *Bioinformatics* **29**, 1260–1267.

Miller, C. A., Hampton, O., Coarfa, C. & Milosavljevic, A. (2011), 'Readdepth: A parallel R package for detecting copy number alterations from short sequencing reads', *PLOS ONE* **6**(1).

Nishida, N., Yano, H., Nishida, T., Kamura, T. & Kojiro, M. (2006), 'Angiogenesis in cancer', *Vascular Health and Risk Management* **2**(3), 213–219.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1993983/*

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

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Ross, S. M. (2014), *Introduction to Probability Models*, 11 edn, ACADEMIC PRESS.

Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D. & Curtis, C. (2015), 'A big bang model of human colorectal tumor growth', *Nature Genetics* **47**, 209.
URL: *https://www.nature.com/articles/ng.3214*

Sun, R., Hu, Z., Sottoriva, A., Graham, T. A., Harpak, A., Ma, Z., Fischer, J. M., Shibata, D. & Curtis, C. (2017), 'Between-region genetic divergence reflects the mode and tempo of tumor evolution', *Nature genetics* **49**(7), 1015–1024.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5643198/*

Tarabichi, M., Martincorena, I., Gerstung, M., Markowetz, F., Spellman, P. T., D. Morris, Q., Lingjærde, O. C., Wedge, D. C. & Van Loo, P. (2017), 'Neutral tumor evolution?', *bioRxiv* .
URL: *https://www.biorxiv.org/content/early/2017/06/30/158006*

Waclaw, B., Bozic, I., Pittman, M. E., Hruban, R. H., Vogelstein, B. & Nowak, M. A. (2015), 'A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity', *Nature* **525**(7568), 261–264.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782800/*

Walpole, R. E., Myers, R. H. & Myers, S. L. (2014), *Probability and Statistics for Engineers and Scientists*, 9 edn, PEARSON.

Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. (2017), 'Response to Tarabichi and colleagues'.
URL: *https://pubpeer.com/publications/70B52651A230B39F3A9EFDC50BAB8A*

Williams, M. J., Werner, B., P. Barnes, C., Graham, T. A. & Sottoriva, A. (2016), 'Identification of neutral tumor evolution across cancer types', *Nature Genetics* **48**(3), 238–244.
URL: *https://www.nature.com/articles/ng.3489*

Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. (2013), 'Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives', *BMC Bioinformatics* **14**(Suppl 11), S1–S1.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3846878/*

7 R-code

```
#Explanation of R-code:
#First, all functions used will be listed below. Afterwards, scripts are
     presented in order to use these functions and provide the same
    figures as presented in the thesis.

TumorGrowth = function(maxgen,beta,lambda){
#Simulate tumor until reaching generation maxgen.
#Discrete branching tree model of tumor evolution where each cell has
    initially the probability of
# "beta" to divide and "1-beta" to die, so only two options. For each
    cell division
# both daughter cells have the possibility to acquire new mutations. The
     number of passenger mutations is given by a Poisson distribution
    with rate "lambda".
#Mutations appearing at a very late stage are not expected to be
    detected. After "maxMuts" cells, the virtual tumor grows without
#adding new mutations, since the mutations are assumed not to be
    traceable anyway.
#The function returns the mutations for all cells, the generation where
    each mutation arose and the total number of cells.


  #Begins with one tumor cell
  NumberOfCells = 1



  gen = 0 #The present generation in the branching tree. The initial
    generation is by definition 0 (consisting of the first tumor cell)
```

```
#MutationsInEachCell is what is returned. For neutral evolution: A vector of lists, where each list represent a cell, and the
#list contains an atomic vector consisting of every mutation that has occured in this specific cell given by an integer.

#Each list in a vector represents a cell in this group. The list consists of all mutations this cell in this group has.
#Mutation 0 represents mutations already present at the first tumor cell.
MutationsInEachCell = 0
m = 1 #Every unique passenger mutation is labeled as an integer.

#For convenience, let the first cell be guaranteed to divide without new driver mutations:
NumberOfCells = 2
NumberOfCellsPerGroup = 2
gen = 1

#Mutations in daughter cells:
muts1 = rpois(1,lambda)
muts2 = rpois(1,lambda)
#Preallocate:
MutationsInEachCell = vector(mode = "list", length = NumberOfCells)
InWhichGenerationMutationsOccurred = vector(mode = "list", length = NumberOfCells)
if(muts1 >0){
  #add mutations:
  MutationsInEachCell[[1]] = c(0,m:(m+muts1-1))
  InWhichGenerationMutationsOccurred[[1]] = c(0,rep(gen,muts1))
  #update number of mutations that have occured:
  m = m + muts1
}
```

```
   #daughter cell is just the same as mother cell:
     else{
        MutationsInEachCell[[1]] = 0
        InWhichGenerationMutationsOccurred[[1]] = 0
     }
   #repeat:
    if(muts2 > 0){
      MutationsInEachCell[[2]] = c(0,m:(m+muts2-1))
      InWhichGenerationMutationsOccurred[[2]] = c(0,rep(gen,muts2))
      m = m + muts2
    }
     else{
        MutationsInEachCell[[2]] = 0
        InWhichGenerationMutationsOccurred[[2]] = 0
     }

#From now on, let the tumor grow stochastically according to a branching
       process until reaching maxsize cells (or dies out).

    while(gen <= maxgen){




      #Decide how many cells that divide in present generation (the other
      cells will eventually die):
      CellsThatDivided = rbinom(1, NumberOfCells, beta)

      #choose which cells divided:
      WhichCellsDivided = sample(NumberOfCells,CellsThatDivided,replace =
     FALSE)

      UpdateNumberOfCells = 2*CellsThatDivided
```

```
# If tumor dies out: return(0)
if(UpdateNumberOfCells == 0){
  return(0)
}

#Since tumor has not died out, the generation has increased by one
gen = gen + 1

#preallocate vector for updating mutations in every cell (assuming
this makes the code faster).
UpdateMutationsInEachCell = vector(mode = "list", length =
UpdateNumberOfCells)
UpdateInWhichGenerationMutationsOccurred = vector(mode = "list",
length = UpdateNumberOfCells)

# Update mutations in each cell:
# To avoid too much space of new mutations, stop to create new
mutations after a given generation given by maxMuts
temp = 1 #temporary variable in order to update each cell correctly.
if(gen <= maxMuts) {
  #Simulate the number of passenger mutations occuring in each of
the cells
  NrOfMutationsInEachCell = rpois(UpdateNumberOfCells,lambda)

  #Update mutations in each cell:
  for(j in WhichCellsDivided){
    #Create the two daughter cells coming from cell j:
    if(NrOfMutationsInEachCell[temp] > 0){
      UpdateMutationsInEachCell[[temp]] = c(MutationsInEachCell[[j
]],m:(m+NrOfMutationsInEachCell[temp]-1))
```

```
         UpdateInWhichGenerationMutationsOccurred[[temp]] = c(
InWhichGenerationMutationsOccurred[[j]],rep(gen,
NrOfMutationsInEachCell[temp]))
        m = m + NrOfMutationsInEachCell[temp]
      }
      else{
         UpdateMutationsInEachCell[[temp]] = MutationsInEachCell[[j]]
         UpdateInWhichGenerationMutationsOccurred[[temp]] =
InWhichGenerationMutationsOccurred[[j]]
      }
      if(NrOfMutationsInEachCell[temp+1] > 0){
         UpdateMutationsInEachCell[[temp+1]] = c(MutationsInEachCell[[j
]],m:(m+NrOfMutationsInEachCell[temp+1]-1))
         UpdateInWhichGenerationMutationsOccurred[[temp+1]] = c(
InWhichGenerationMutationsOccurred[[j]],rep(gen,
NrOfMutationsInEachCell[temp+1]))
        m = m + NrOfMutationsInEachCell[temp+1]
      }
      else{
         UpdateMutationsInEachCell[[temp+1]] = MutationsInEachCell[[j]]
         UpdateInWhichGenerationMutationsOccurred[[temp+1]] =
InWhichGenerationMutationsOccurred[[j]]
      }
      temp = temp + 2
    }
  }
 #else, there is no need to add mutations as they will not be
detectable.
 else{
   #All daughter cells inherit exactly the same mutations as their
parents:
    UpdateMutationsInEachCell = MutationsInEachCell[rep(
  WhichCellsDivided,2)]
    UpdateInWhichGenerationMutationsOccurred =
  InWhichGenerationMutationsOccurred[rep(WhichCellsDivided,2)]
    # for(j in WhichCellsDivided){
    # UpdateMutationsInEachCell[[temp]] = MutationsInEachCell[[j]]
    # UpdateMutationsInEachCell[[temp+1]] =  MutationsInEachCell[[j]]
    # temp = temp + 2
    # }
   }
  #Update parameters
   NumberOfCells = UpdateNumberOfCells
   MutationsInEachCell = UpdateMutationsInEachCell
   InWhichGenerationMutationsOccurred =
  UpdateInWhichGenerationMutationsOccurred
   }

   return(list(MutationsInEachCell,InWhichGenerationMutationsOccurred,
  NumberOfCells,gen))

}




RealDistributionOfVAF = function(tumorSample,gens,ploidy = 2){
#From the simulated tumor, compute the VAF for each somatic mutation.

   TotalNumberOfCells = tumorSample[[3]]
   TheSample = tumorSample[[1]]
   CorGen = tumorSample[[2]]
  #Get all mutation IDs in one single atomic vector:
```

```
  AllMutations = unlist(TheSample)
  #Get VAF of all somatic mutations using R-function table:
  RealVAF = table(AllMutations)/(ploidy*TotalNumberOfCells)
  #Get the corresponding generation where each mutation occurred:
  GenerationForEachVaf = unlist(CorGen)
  Vafs = vector(mode = "list", length = gens)
  for(i in 1:gens){
    #Find which mutations belong to the same generation (some
  mutations are equal as they belong to the same parent possesing this
   mutation)
    where = which(GenerationForEachVaf == i)
    #Find the unique mutations appearing in generation i
    muts = as.character(unique(AllMutations[where]))
    #Store the corresponding VAFs of the somatic mutations appearing
  in generation i.
    Vafs[[i]] = as.vector(RealVAF[muts])
  }
  return(Vafs)
}

#COMPUTE OBSERVED M(k) from tumor:
#M(k) is the cumulative number of somatic mutations arising before
  generation k that have survived.
RealMk = function(tumorSample,gens){
  #tumor contains cells and mutations in each cell in addition to
  which generation each surviving mutation appeared.
  #gens is the number of generations to look at.
  TotalNumberOfCells = tumorSample[[3]]
  TheSample = tumorSample[[1]]
  CorGen = tumorSample[[2]]
  #Get all mutation IDs in one single atomic vector:
  AllMutations = unlist(TheSample)
```

```
    #Get the corresponding generation where each mutation occurred:
    GenerationForEachVaf = unlist(CorGen)
    NrOfMutsAppearedInGenk = vector(mode = "numeric", length = gens)
    for(k in 1:gens){
      #Find which mutations belong to the same generation (some mutations are equal as they belong to the same parent possesing this mutation)
      where = which(GenerationForEachVaf == k)
      #Find the unique number of mutations appearing in generation i
      NrOfMutsAppearedInGenk[k] = length(unique(AllMutations[where]))
    }
    Mk = cumsum(NrOfMutsAppearedInGenk)
    return(Mk)
  }


DistributionOfPublicMutations = function(ID,readsFile, purity, sims,l){

  #Get read depth data.
  name = paste(readsFile,ID, sep = "/")
  name = paste(name, ".RData",sep = "")
  readsData = load(file = name)
  reads = get(readsData)
  #Compute MLE-estimators:
  library(MASS)
  par_estimators = fitdistr(x = reads, densfun = "negative binomial")
  size_par = as.numeric(par_estimators$estimate[1])
  mu_par = as.numeric(par_estimators$estimate[2])

  #Generate sims reads:
  r = rnbinom(n = sims, size = size_par, mu = mu_par)
  #For each read, generate an observed VAF-value:
```

```
  VAF_observed = rbinom(n = sims, size = r, prob = 0.5*purity)/r
  p = length(which(VAF_observed < 1))/sims
  hist(VAF_observed, breaks = 100, xlim = c(0,0.6))
  return(p)
}

DistributionOfExtinction = function(beta, maxgen){
#Count number of cells in each generation until extinction:
#population begins with one cell:
pop = 1
gen = 1
NrOfCellsInEachGen = vector(mode = "numeric", length = maxgen)
while(pop>0 && gen <= maxgen){
#Decide how many cells that divide in present generation (the other
    cells will eventually die):
CellsThatDivided = rbinom(1,pop, beta)

pop = 2*CellsThatDivided
# If tumor dies out: return(0)
if(pop == 0){
  return(NrOfCellsInEachGen)
}
else{
  NrOfCellsInEachGen[gen] = pop
}
gen = gen + 1
}
  #extinction not guarantied, return 0:
  return(0)
}

DistributionOfSurivingPopulations = function(beta,maxgen){
  #Count number of cells in each generation until maxgen
  #population begins with one cell:
  pop = 1
  gen = 1
  NrOfCellsInEachGen = vector(mode = "numeric", length = maxgen)
  while(gen <= maxgen){
    #Decide how many cells that divide in present generation (the other
    cells will eventually die):
    CellsThatDivided = rbinom(1,pop, beta)

    pop = 2*CellsThatDivided
    # If tumor dies out: return(0)
    if(pop == 0){
      return(0)
    }
    else{
      NrOfCellsInEachGen[gen] = pop
    }
    gen = gen + 1
  }
  return(NrOfCellsInEachGen)
}


TakeResection = function(tumorSample,par_mean,par_size,purity,
    truncalamount, ploidy = 2){
  #Function has input "tumorSample" from a virtual tumor created from
    tumorGrowth().
  #par_mean and size is the parameters of the negative binomial
    distribution (mean and size).
```

```
#Get all mutation IDs in one single atomic vector:
AllMutations = unlist(tumorSample[[1]])
S = tumorSample[[3]]
#Get VAF of all somatic mutations using R-function table:
RealVAF = (table(AllMutations)/(ploidy*S))*purity
#Frequencies that are too low have very little chance of beeing seen,
 especially in the region of interest [0.01,0.25].
#Therefore, for simplicity discard too low frequencies as they will
 not influence the great picture anyway:
RealVAF = RealVAF[RealVAF >= 0.001]

#Now, the OBSERVED frequencies of each somatic mutation is estimated
 assuming a Gamma-Poisson distributed read depth:
ObservedVAF = vector(mode = "numeric",length = length(RealVAF))
#Simulate read depth in each position:
ReadDepths = rnbinom(n = length(RealVAF), size = par_size, mu = par_
 mean )
for(i in 1:length(RealVAF)){

  if(ReadDepths[i] >= 10){
    NrOfMutationsRecorded = rbinom(1,ReadDepths[i],as.numeric(RealVAF[
 i]))

    if(NrOfMutationsRecorded >= 3){
      ObservedVAF[i] = NrOfMutationsRecorded/ReadDepths[i]
    }

  }

}
ObservedVAF = ObservedVAF[ObservedVAF!= 0]
```

```
  #Add truncal mutations
  NrOftruncalMutations = as.integer(((truncalamount)/(1-truncalamount))*
    length(ObservedVAF))
  truncalVAFs = vector(mode = "numeric",length = NrOftruncalMutations)
  #Read depths for each truncal mutation:
  truncalReadDepths = rnbinom(n = NrOftruncalMutations, size = par_size,
     mu = par_mean )
  for(i in 1:NrOftruncalMutations){

    if(truncalReadDepths[i] >= 10){
      NrOfMutationsRecorded = rbinom(n=1,size = truncalReadDepths[i],p
    =0.5*purity)

      if(NrOfMutationsRecorded >= 3){
        truncalVAFs[i] = NrOfMutationsRecorded/truncalReadDepths[i]
      }

    }

  }
  ObservedVAFs = c(ObservedVAF,truncalVAFs)
  ObservedVAFs = ObservedVAFs[ObservedVAFs != 0]

  return(ObservedVAFs)
}
#END OF FUNCTIONS APPLIED
###############################################

#HERE, THE FUNCTIONS ABOVE CAN BE APPLIED
#EACH SCRIPT IS SEPERATED BY HASHTAGS ###. Outcomment a script in order
    to directly use it in any R-script.
```

```
#Decide parameters here:
beta = 0.55
lambda = 1.5
maxgen = 150
maxMuts = 40
ploidy = 2
truncalamount = 0.5
# par_mean = c(80,90,100,110,120,130,190)
# par_size = c(2.5,2.5,2.5,2.5,2.5,2.5,2.5)
# purity = c(0.7,0.7,0.7,0.7,0.7,0.7,0.7)
par_mean = 178
par_size = 2.27
purity = 0.3

##################################################
#CREATE A TUMOR AND COMPUTE M(E[F_k])
  # k = 1:40
  # Efk = 1/(2*(2*beta)^k)
  # x = 1/Efk
  # mEfk = (lambda/(2-1/beta))*(1/Efk-2)
  # t = tumorGrowth(maxgen, beta, lambda, lambda_d, maxMuts, s)
  #    while(object.size(t)<70){
  #       t= tumorGrowth(maxgen, beta, lambda, lambda_d, maxMuts, s )
  #    }
  # TotalNumberOfCells = t[[3]]
  # TheSample = t[[1]]
  # #Get all mutation IDs in one single atomic vector:
  # AllMutations = unlist(TheSample)
  # #Get VAF of all somatic mutations using R-function table:
  # RealVAF = table(AllMutations)/(ploidy*TotalNumberOfCells)
  # RealVAF = RealVAF[RealVAF >= 0.01]
  # tab = rev(table(RealVAF))
  # Mf = as.vector(cumsum(tab))
  # recordedvafs = as.numeric(names(tab))
  # MEfk = vector(mode = "numeric", length = length(Efk))
  # for(i in 1:length(MEfk)){
  #   position = OrderNumericInRightPlace(recordedvafs,Efk[i])
  #   MEfk[i] = Mf[position]
  # }
  # Mflist = list(x,mEfk, MEfk)
  # save(Mflist, file = )
##########################################################

#COMPARE M(f) in real data vs. observed data with given read depth
    distribution:
#
#  library(foreach)
#  library(doParallel)
#  #Decide number of clusters here:
#  cl = makeCluster(3)
#  registerDoParallel(cl)
#
#  foreach(i = 1:7) %dopar% {
# o = TakeResection(t,S,par_mean[i],par_size[i],purity[i])
# save(o, file = )
#  }
#  stopCluster(cl)
#########################################################################


#ESTIMATE distribution of VAF here when neutral evolution (only looking
    at subclonal mutations):
NrOftumors = 200
 library(foreach)
 library(doParallel)
 #Decide number of clusters here:
 cl = makeCluster(5)
 registerDoParallel(cl)
foreach(i = 1:NrOftumors) %dopar% {
  t = tumorGrowth(maxgen,beta,lambda,lambda_d,maxMuts, s)
  while(object.size(t)<70){
    t= tumorGrowth(maxgen,beta,lambda,lambda_d,maxMuts, s )
  }
  r = RealDistributionOfVAF(tumorSample = t,gens = maxMuts)
  save(r, file = )
}
######################################################################

# #ESTIMATE EXPECTED GROWTH FOR SURVIVING TUMORS
# NrOftumors = 2000
# library(foreach)
# library(doParallel)
# #Decide number of clusters here:
# cl = makeCluster(5)
# registerDoParallel(cl)
# foreach(i = 1:NrOftumors) %dopar% {
#   sur = DistributionOfSurivingPopulations(beta,maxgen)
#   if(object.size(sur) > 100){
#     save(sur, file = )
#   }
# }

###################################################################
#FIND M(k) based on 200 tumors
# NrOftumors = 50
# gens = 40
#  library(foreach)
#  library(doParallel)
#  #Decide number of clusters here:
#  cl = makeCluster(5)
#  registerDoParallel(cl)
# foreach(i = 1:NrOftumors) %dopar% {
#   t = tumorGrowth(maxgen,beta,lambda,lambda_d,maxMuts, s)
#   while(object.size(t)<70){
#     t= tumorGrowth(maxgen,beta,lambda,lambda_d,maxMuts, s )
#   }
#   Mk = RealMk(t,gens)
#   save(Mk, file =  )
# }
################################################################

#ESTIMATE expected number of cells in each generation conditioned on
    extinction:
# maxgen = 120
# extinctMatrix = matrix(nrow = 5000000, ncol = maxgen)
# WhenDied = vector(mode = "numeric", length = 120)
# for(i in 1:5000000){
#   d = DistributionOfExtinction(beta,maxgen)
#   if(length(d)>1){
#     d = DistributionOfExtinction(beta,maxgen)
#     when = min(which(d == 0))
#     WhenDied[when] =  WhenDied[when] + 1
#     extinctMatrix[i,] = d
#   }
#
# }
# ex = colMeans(extinctMatrix, na.rm = TRUE)
# save(ex,file = )
# save(WhenDied, file = )

##############################################################

#CHECH DISTRIBUTION OF PUBLIC MUTATIONS:
# pm = DistributionOfPublicMutations(ID = "TCGA-NH-A5IV-01A-42D-A36X
    -10", readsFile = ,
#                                        purity = 0.8,sims = 1000000, l =
    0.1)



#v = VirtualSequenceAnalysis(200,2,0.01,0.5,200,2.4,0.72)
# tumor = tumorGrowth(maxsize,beta,lambda,lambda_d,maxMuts,s)
#
#    while(object.size(tumor)<70){
#      tumor = tumorGrowth(maxsize,beta,lambda,lambda_d,maxMuts, s )
#    }
# biopsy = TakeBiopsy(tumor,10000000,100,2,1)
# save(biopsy, file = )
# # #Simulate bunches of tumors in parallell:
# library(foreach)
# library(doParallel)
# #Decide number of clusters here:
# cl = makeCluster(3)
# registerDoParallel(cl)
#
# foreach(i = 1:11) %dopar% {
#
# #Call tumorGrowth() until a tumor has grown to desired size
#    tumor = tumorGrowth(maxsize,beta,lambda,lambda_d,maxMuts,s)
#    #If object size
#    while(object.size(tumor)<70){
#      tumor = tumorGrowth(maxsize,beta,lambda,lambda_d,maxMuts, s )
#     }
#    RealVAF = RealDistributionOfVAF(tumor)
#    #save somewhere:
#    save(RealVAF, file = )
# }
# stopCluster(cl)

####################################################################

#Create a fictive VAF-distribution
# tumor = tumorGrowth(maxgen,beta,lambda,lambda_d,maxMuts,s)
#
#    while(object.size(tumor)<70){
#      tumor = tumorGrowth(maxgen,beta,lambda,lambda_d,maxMuts, s )
#    }
# resection = TakeResection(tumor,par_mean, par_size,purity,
    truncalamount)
```