RICE UNIVERSITY

Modeling Clonal Evolution with Branching Processes

by

Thomas McDonald

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Dr. Marek Kimmel, Chair
Professor of Statistics, Rice University

Dr. Dennis Cox
Professor of Statistics, Rice University

Dr. Herbert Levine
Hasselmann Professor of Bioengineering,
Rice University

HOUSTON, TEXAS
JUNE 2015
Abstract

Modeling Clonal Evolution with Branching Processes

by

Thomas McDonald

Clonal evolution in tumorigenesis assumes a tumor arises from a single ancestor cell followed by a sequence of mutations that increase fitness in clones. Waves of clonal expansions take place with each new mutation contributing to the selection of the cell. Not all mutations affect a clone’s growth dynamics, and some have a neutral effect on cell growth. We create a branching process model of clonal evolution that distinguishes between selective (driver) and neutral (passenger) mutations. In this model, cells grow according to distributions unique to their set of mutations, and further mutations can arise as stochastic events during replication. The models are introduced as infinite-allele multitype branching processes, and we prove asymptotic results about the number and distribution of neutral alleles. The model can be used to determine the growth rate of a tumor and the clonal subpopulations within, and we discuss how this model may be used for estimation of growth parameters in clones. Using variant frequency data from a tumor at a single time point, estimation methods can then be used to rebuild the evolutionary history of a tumor. We discuss how the rate of growth of the number of neutral alleles can be used as a molecular clock to help us in this task.
This work is dedicated to my girlfriend, Kate, and to my parents.
ACKNOWLEDGEMENTS

I would like to thank Dr. Marek Kimmel, my advisor, for his expertise, advice, and patience. He took me in to work in branching processes, and provided me with the resources to learn and contribute to the field. I appreciate his assistance in writing and editing my work, and his memory of almost any paper or study he ever read which made literature reviews and searching for results much easier.

I would also like to thank the other members of my committee, Dr. Dennis Cox and Dr. Herbert Levine, for their input and direction along the way, and to Dr. Sharon Plon for her support and knowledge of the biology. She was always willing to explain the concepts I wasn’t familiar with during our meetings, and patiently made sure I could keep up, for which I am very thankful.

My thanks goes to Dr. Peter Olofsson for introducing me to this topic and continuing to check in on me. Without his first undergraduate class in Stochastic Processes, I cannot say where I would currently be. He also provided me the opportunity to work with him on my first research project over summer before graduate school while I also worked as a poolside waiter at a local hotel.

The friends and peers in my program also deserve a huge amount of credit for their help in our classes and the insightful discussions we had throughout the program. I thank Casper and Nicolas Flores for being
friends and a major source of advice as we navigated the program together. Emilian Vankov, Christine Peterson, Stephanie Hicks, and Sharon Chiang were all sources of support and friendship as well, and I am grateful for each of them.

I would like to acknowledge the help provided by my family: my parents for supporting my goals and encouraging me to continue my education, and my brothers, Sam and Joey, for years of competition which encouraged me to always work harder. Finally, I would like to thank my girlfriend, Kate Castilleja, for her help, support, and love the last six years.
Abstract

Acknowledgements

1 Introduction to Branching Processes

1.1 A background of branching processes

1.2 Outline of Dissertation

2 Cancer Evolution in Tumorigenesis

2.1 Motivation: Transition from severe congenital neutropenia to secondary AML

2.2 Population Dynamics of Tumorigenesis

2.3 Cancer Progression

2.4 Clonal Evolution

3 Multitype infinite-allele branching processes

3.1 Introduction to infinite-allele branching processes

3.2 The multitype infinite-allele Galton Watson process

3.3 Continuous Time Markov Branching Process

3.4 General Branching Process and Results

3.5 Proof of Concept Examples

3.6 Applications to Cancer Evolution

4 Application: A Model of Clonal Evolution

4.1 Introduction
4.2 Modeling the Number of Driver and Passenger Mutations . . . . . . . . 133
4.3 Estimating Cancer Progression . . . . . . . . . . . . . . . . . . . . . . . 146
4.4 Creating a Branching Process Model of Cancer Evolution . . . . . . . 157
4.5 Problems and Future Directions . . . . . . . . . . . . . . . . . . . . . . . 172
4.6 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 180

References 182
List of Figures

2.1 Clonal Evolution of Tumors ........................................ 28
2.2 Roeder’s Agent-Based Model ...................................... 39
2.3 Simulation of the Agent-Based Model ............................. 40
2.4 Logistic Birth-Death-Migration Process State Space .......... 41
2.5 Possible Actions in Logistic BDM Process ........................ 44
2.6 Quasistationary Joint Distribution from Simulations of Logistic BDM Process .................................................. 45
2.7 Quasistationary Marginal Distributions from Simulations of Logistic BDM Process .................................................. 46
2.8 Comparison of Cell Counts in Agent-Based and Logistic BDM Models ................................................................. 48
2.9 Possible evolutions of Tumor ......................................... 49
2.10 Different Routes to Achieve 3 Mutations ......................... 54
2.11 If we have an idea of the path taken, we can assume values of the mean matrix are zero, reducing the number of parameters in the model. This particular example leads to a reduction from 8 possible types to 4 types. 55
2.12 Process to Infer Clonal Phylogeny ............................... 57
2.13 Schematic of the Process to Link Data to Branching Processes .......................... 61

3.1 Infinite-Allele Multitype Galton-Watson Realization ............... 76
3.2 Number of Labels for 4 Type Irreducible Process .................. 87
3.3 Number of Labels for 4 Type Reducible Process ................... 89
3.4 Frequency Spectrum for 2-type Irreducible Process ................ 90
3.5 Histogram of Frequency Spectrum for 2-type Irreducible Process .... 91
3.6 Infinite-Allele Multitype Continuous Time Realization ............. 102
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.7</td>
<td>Number of Labels for 4-Type Irreducible Continuous Time Process</td>
<td>128</td>
</tr>
<tr>
<td>3.8</td>
<td>Number of Labels for 4-Type Reducible Continuous Time Process</td>
<td>130</td>
</tr>
<tr>
<td>4.1</td>
<td>A Single-Type Realization</td>
<td>137</td>
</tr>
<tr>
<td>4.2</td>
<td>Theoretical passengers in Glioblastoma</td>
<td>141</td>
</tr>
<tr>
<td>4.3</td>
<td>Drivers and passengers in Glioblastoma</td>
<td>142</td>
</tr>
<tr>
<td>4.4</td>
<td>Drivers and passengers in Lung Cancer</td>
<td>144</td>
</tr>
<tr>
<td>4.5</td>
<td>Curated Drivers and passengers in Lung Cancer</td>
<td>145</td>
</tr>
<tr>
<td>4.6</td>
<td>Growth in number of passenger alleles</td>
<td>146</td>
</tr>
<tr>
<td>4.7</td>
<td>Schematic for estimation of parameters</td>
<td>148</td>
</tr>
<tr>
<td>4.8</td>
<td>Correspondence Analysis</td>
<td>150</td>
</tr>
<tr>
<td>4.9</td>
<td>Correspondence Analysis</td>
<td>152</td>
</tr>
<tr>
<td>4.10</td>
<td>Example Trees</td>
<td>156</td>
</tr>
<tr>
<td>4.11</td>
<td>Example Trees</td>
<td>158</td>
</tr>
</tbody>
</table>
List of Tables

4.1 Number of passenger mutations from Bozic model . . . . . . . . . . . . . . . . . . 139
4.2 Number of passenger mutations from infinite allele model . . . . . . . . . . . . . . 139
4.3 Allele Frequencies from Estimated Trees . . . . . . . . . . . . . . . . . . . . . . 157
4.4 T staging for cancers . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 168
4.5 Distribution of T staging and tree topologies . . . . . . . . . . . . . . . . . . . . . 169
4.6 Approximate size of each tumor . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 169
Chapter 1

Introduction to Branching Processes

The branching process is a stochastic process originating with Francis Galton and Reverend Henry William Watson’s investigation of family lines of British nobility. Since then it became its own subject of study, taking on numerous applications within genetics, population biology, and atomic theory. A basic treatment of the process describes the growth of a population based on the properties of the identical individuals the population comprises. The number of offspring, type, and time of lifetime of individuals are treated as random variables. Individuals are independent of each other and time can be discrete or continuous. These features make branching processes convenient to use in the study of cell proliferation and tumorigenesis, and we can attempt to understand the growth and composition of a population of cells.

The ultimate goal is to determine how different subpopulations of cells with a specific set of mutations grow in a tumor in order to find the most aggressive mutations. To achieve this, we introduce two multitype branching processes that allow us to distinguish between driver and passenger mutations. In doing so, asymptotic results are established as is usually done in branching processes, so that the dynamics of the population could be understood over long periods of time. This will become useful for describing the composition of these populations, and showing that cells in
these tumors could continue and mutating into populations with higher fitness if left unchecked.

We give an introduction to branching processes in this chapter and state necessary definitions required throughout this thesis. These results are used later to help establish the processes that were formed to model clonal evolution and cancer progression, and are fundamental in the proofs for that associated theorems.

1.1 A background of branching processes

We introduce branching processes with a quick treatment of the original and most basic branching process: the Bienaymé-Galton-Watson (BGW) Process. The BGW process begins with $Z(0)$ individuals acting as ancestors. Each ancestor individual live for a single generation before splitting into a random number of offspring according to a given probability mass function, $\{p_k, k = 0, 1, \ldots\}$. Generations are non-overlapping, and offspring live for a single generation as well before splitting into their own offspring according to the same distribution. This process of individuals splitting continues ad infinitum [1]. At the most basic level, individuals split independently of each other or their parents. The particles are counted by the Markov chain $\{Z(n), n = 0, 1, 2, \ldots\}$ where time, $n$, is a nonnegative integer representing generations. We assume $Z(0) = 1$ unless stated differently. The BGW process can be explicitly defined by this Markov chain and its transition function.

\textbf{Definition 1} (BGW process). A Bienaymé-Galton-Watson (BGW) process is a Markov chain $\{Z(n), n = 0, 1, 2, \ldots\}$ on $0 \cup \mathbb{Z}^+$ with the transition function

$$P(i, j) = P(Z(n + 1) = j | Z(n) = i) = \begin{cases} p_j^i, & \text{if } i \geq 1 \\ \delta_{0j}, & \text{if } i = 0 \end{cases}$$
where \( p^*_j \) is the \( i \)-fold convolution of the probability mass function \( \{p_j; j = 0, 1, \ldots \} \).

Each individual in generation \( n \) can be given an arbitrary order. We then define \( \xi_j^{(n)} \) as the nonnegative random variable describing the number of offspring of the \( j^{th} \) individual of generation \( n \). Each \( \xi_j^{(n)} \) is independent and identically distributed.

The identifying feature of branching processes is they satisfy the aptly named branching property. That is the total number of individuals in a generation \( n + 1 \) is equal to the combined number of offspring of each individual in the previous generation, \( n \) [2]. In mathematical terms,

\[
Z(n + 1) = \sum_{j=1}^{Z(n)} \xi_j^{(n)}.
\]

Each individual initiates its own process, so we can also consider the property in terms of the subprocesses initiated by first-generation individuals (offspring of the ancestors). If we let \( Z^{(j)}(n) \) be the process initiated by the \( j^{th} \) individual of generation 1, then

\[
Z(n + 1) = \sum_{j=1}^{Z(1)} Z^{(j)}(n).
\]

A common tool in analyzing branching processes is the probability generating function (p.g.f.)

\[
f(s) \equiv E[s^{Z(1)}] = \sum_{j \geq 0} p_j s^j, \quad |s| \leq 1
\]

where \( p_j \equiv P[Z(1) = j | Z(0) = 1] \) is the offspring reproduction mass function. This is the same p.g.f. for all individuals in the process. The distribution of the branching process evolves according to the iterates of the p.f.g. \( f(s) \). Thus, \( f^{(n+1)}(s) = f(f^{(n)}(s)) \) is the p.g.f. for the \((n + 1)^{st}\) generation, and we define \( f^{(0)}(s) = s \) as the p.g.f. for the ancestor in the case of a single ancestor. Note that if \( Z(0) = i \), then \( f^{(1)}(s) = [f(s)]^i \). We can use the p.g.f. to determine the moments of \( Z(1) \) which can
be extended to determine the expectation of the process. Letting $m = E[Z(1)]$ and $\sigma^2 = Var[Z(1)]$,

$$m = f'(1-) \quad \text{and} \quad \sigma^2 = f''(1-) + m - m^2.$$ 

The mean and variance of the process, $Z(n)$ is

$$E[Z(n)] = m^n \quad \text{and} \quad Var[Z(n)] = \begin{cases} \frac{\sigma^2 m^{n-1}(m^n-1)}{m-1}, & m \neq 1 \\ n\sigma^2, & m = 1 \end{cases}$$

which are useful in describing the growth of the process.

Another property to consider is the probability of extinction of a process. Letting $q(i) = P[Z(i) = 0, \text{for some } 1 \leq i \leq n]$, we can show that $q = \lim_{n \to \infty} q(n)$ is the smallest root to the equation $s = f(s)$ for $s \in [0,1]$. Thus, we can employ the p.g.f. to determine properties about asymptotics of the process like the probability that a process eventually goes extinct. In fact results about extinction can be determined based on the value of $m$. We summarize these results in the following theorem which is provided in Kimmel and Axelrod [1].

**Theorem 1.** If $m \leq 1$ then $q = 1$, and if $m > 1$ then $q < 1$.

Growth properties about a BGW process can classified based on the value of $m$. A process is called *subcritical*, *critical*, or *supercritical* if its mean is less than, equal to, or greater than 1 respectively. These terms describe whether we expect a process to grow and escape extinction, or almost surely become extinct. The above theorem states that critical and subcritical processes will become extinct. This theorem and its multitype counterpart (described later) have important implications when modeling cell growth and tumorigenesis as a branching process. They essentially allow us to determine how we expect a process to grow based on only the offspring p.g.f.
We should expect hematopoeisis and stem cell growth can be modeled as a critical branching process since such a process has no expected growth. Tumorigenesis is due to unregulated growth of cancer cells, and this becomes a problem when the fitness of cancer cells becomes greater than that of normal cells, so a process modeling growth of cancer cells should be represented by a supercritical process.

**Markov continuous-time branching processes**

The Galton-Watson process makes a strong assumption that generations only last for a single unit of time. In reality, generations tend to overlap, so the Markov continuous-time branching process allows for overlapping generations by assuming individuals life for a random lifetime before splitting into offspring according to their p.g.f. $f(s)$. We begin by defining the process in one-dimension before extending to multiple types. We state important definitions and results for these types of processes that will be used later.

**Definition 2.** A Markov continuous-time branching process is a stochastic process \( \{Z(t), t \geq 0\} \) with state space \( 0 \cup \mathbb{Z}^+ \) having the transition function \( P_{ij}(t) \) such that

\[
\sum_{j=0}^{\infty} P_{ij}(t)s^j = \left( \sum_{j=0}^{\infty} P_{ij}(t)s^j \right)^i \quad \forall i \geq 0, |s| \leq 1.
\]

The transition probability stated in the definition characterizes the branching property that defines a branching process, that the probability function for the offspring of \( i \) individuals is equal to the product probability functions for the offspring of a single individual. Another way to write the branching property in terms of the process itself if

\[
Z(t + \Delta t) = \sum_{i=1}^{Z(\Delta t)} Z^{(i)}(t),
\]

that states that the number of individuals alive at a time \( t + \Delta t \) is equal to the sum
of the individual trees initiated by each living individual at time $\Delta t$ after proceeding for $t$ time.

In such a process, a single individual begins as the ancestor at time 0 and lives for a length of time that is a random variable with exponential distribution with parameter $a$ and mean $a^{-1}$. Upon death, it splits into $k$ offspring with probability $p_k$. The number of offspring is described by the same p.g.f. as in the Galton-Watson case:

$$f(s) = \sum_{j \geq 0} p_j s^j, \quad |s| \leq 1.$$  

We define the p.g.f. of the continuous-time process as

$$F(s; t) = E[s^{Z(t)}] = \sum_{k \geq 0} P_{1k}(t)s^k,$$

and the generating function $u(s) = a(f(s) - s)$. We still define the mean number of offspring from a single individual as $m = f'(1-)$, and we let $\lambda = a(m - 1)$. From the branching property, Kolmogorov equations are derived showing that $F(s; t)$ satisfies

$$\frac{\delta}{\delta t} F(s; t) = u(s) \frac{\delta}{\delta s} F(s; t)$$

$$\frac{\delta}{\delta t} F(s; t) = u(F(s; t)).$$

These two equations are the forward and backward (respectively) Kolmogorov equations with the boundary condition $F(s; 0) = s$ stating a single ancestor begins the process. The unique solution to either of these equations gives the p.g.f. for the process. Many times this solution is not solvable, or can lead to nonlinear differential equations that require numerical methods to solve. The main results from one-dimensional continuous time processes that we use are the mean and Kesten-Stigum limit theorems that we show are similar to the Galton-Watson case. First, let
the mean of the process be

\[ m(t) = \frac{\delta}{\delta s} F(s; t)|_{s=1}. \]

The backward Kolmogorov equation can be differentiated with respect to \( s \) to get the equation

\[ \frac{d}{dt} m(t) = \lambda m(t) \]

which has the solution \( m(t) = \exp[\lambda t] \). Thus the one-dimensional Markov process has exponential growth. In fact, we define criticality of the process by the value \( \lambda \). If \( \lambda < 0 \), the process is subcritical. If \( \lambda = 0 \), the process is critical, and if \( \lambda > 0 \) we call the process supercritical. The behavior of such processes are the same as in the discrete case, with almost sure extinction for \( \lambda \leq 0 \). The conditions based on \( \lambda \) are actually the same conditions as the Galton-Watson case if we use \( m \) instead. Finally, we introduce the theorem about the long term behavior of \( Z(t) \).

**Theorem 2.** There exists a random variable, \( W \) such that

\[ \lim_{t \to \infty} Z(t) e^{-\lambda t} = W \text{ almost surely.} \]

This theorem states that the process \( Z(t) \) grows at an exponential rate. From here, we have a sufficient introduction to lead us to multitype continuous-time processes that are used in this paper.

**Multitype Galton Watson Processes**

The simple BGW process can be extended to include multiple types of individuals, where an individual may have progeny that are other types with other growth capabilities. This allows us to incorporate mutations or different cell types in our model, where cells may have different reproduction laws. Individuals still reproduce indepen-
dently, but the identical distribution assumption is relegated to within each type. For a $k$-type BGW process, $k \in \mathbb{Z}^+$, we define $\{Z(n), n = 0, 1, 2, \ldots \}$ as a $k$-dimensional Markov chain with $Z(n) = (Z_1(n), Z_2(n), \ldots, Z_k(n))$. $Z(n)$ is a sequence of random vectors of length $k$ that counts the number individuals alive of each type at generation $n$.

In generation $n$, the $j^{th}$ $i$-type individual gives birth to $\xi_{j,i}^{(n)} = (\xi_{j,i,1}^{(n)}, \xi_{j,i,2}^{(n)}, \ldots, \xi_{j,i,k}^{(n)})$ progeny according to the reproduction law $\{p_{j,i} \in (0 \cup \mathbb{Z}^+)^k\}$. The branching property still holds in the multitype setting,

$$Z(n+1) = \sum_{i=1}^{k} \sum_{j=1}^{Z_i(n)} \xi_{i,j}^{(n)},$$

and probability generating functions can be used to determine properties about the process based on individual reproduction laws[3]. The branching process states that the number of individuals in generation $n+1$ of each type is equal to the vector sum of the offspring of all individuals in generation $n$.

We denote the p.g.f. of the reproduction law for an $i$-type individual

$$f_i(s) = E[s_1^{Z_1(1)} s_2^{Z_2(2)} \ldots s_k^{Z_k(k)} | Z(0) = \epsilon_i] = \sum_{j_1,j_2,\ldots,j_k \geq 0} p_{j_1,j_2,\ldots,j_k} s^{j}$$

where $p_j = p_{j_1,j_2,\ldots,j_k} = P[Z_1(1) = j_1, \ldots, Z_k(1) = j_k | Z(0) = \epsilon_i]$. Also, we denote the vector $f(s) = (f_1(s), f_2(s), \ldots, f_k(s))$ as the p.g.f. for the entire process. With this, we can give the p.g.f. for the process with an $i$-type ancestor at any generation by
the iterates of the p.g.f.:

\[ f^{(0)}(s) = (f_1^{(0)}(s), f_2^{(0)}(s), \ldots, f_k^{(0)}(s)) = s, \]

\[ f^{(1)}(s) = f(s), \quad \text{and} \]

\[ f^{(n+1)}(s) = f(f^{(n)}(s)) = f^{(n)}(f(s)) \]

The formal definition for a multitype Galton-Watson process is as follows:

**Definition 3** (k-type BGW process). A k-type BGW process is a Markov chain \( \{Z(n), n = 0, 1, 2, \ldots \} \) on \( \mathbb{Z}^{n+} \) with transition function

\[ P(i,j) = P(Z_{n+1} = j|Z_n = i) \]

which is equal to the coefficients of \( s^j \) in \( f(s)^i \).

The moments of the multitype BGW process are determined by the moments of the reproduction process. Let \( m_{ij} = E[Z_j(1)|Z(0) = e_i] \) be the expected number of \( j \)-type offspring from an \( i \)-type parent. Then we define the mean matrix \( M \) as

\[
M = [m_{ij}]_{1 \leq i, j \leq k} = \begin{bmatrix}
  m_{11} & m_{12} & \cdots & m_{1k} \\
  m_{21} & m_{22} & \cdots & m_{2k} \\
  \vdots & \vdots & \ddots & \vdots \\
  m_{k1} & m_{k2} & \cdots & m_{kk}
\end{bmatrix}.
\]

Similar to the single-type process, the moments are determined by the partial derivative of the p.g.f.,

\[ m_{ij} = \frac{\delta}{\delta s_j} f_i(1). \]

Note then that \( E_i[Z_j(n)] = \frac{\delta}{\delta s_j} f_i^{(n)}(1) \) which is used to show the mean of the process is \( E[\mathbf{Z}(n)|\mathbf{Z}(0) = \mathbf{e}_i] = \mathbf{Z}(0)M^n \).
Results about the growth and probability of extinction of the process can be described, but we introduce a few definitions first. Let $m_{ij}^{(n)}$ be the $(i, j)^{th}$ entry of the matrix $M^n$. We call a process irreducible or positive regular if given any type of ancestor, there is positive probability that any other type of individual can arise in the process.

**Definition 4** (Irreducible process). The $k$-type BGW process $\{Z(n), n = 0, 1, \ldots \}$ is irreducible if for any $i, j = 1, 2, \ldots, k$, there exists some $n \geq 0$ such that $m_{ij}^{(n)} > 0$.

A process that isn’t positive regular is called reducible or singular.

**Definition 5** (Reducible branching process). The $k$-type BGW process $\{Z(n), n = 0, 1, \ldots \}$ is reducible if there exists some $i, j = 1, 2, \ldots, k$, such that $m_{ij}^{(n)} = 0$ for all $n \geq 0$.

A reducible process means there exists a situation in which a specific type of particle may never be created depending on the current state. Most of the theory of multitype branching processes relies on the assumption of an irreducible process because it contains nice properties and the asymptotics of the entire chain can be considered at once. Nair and Mode studied and created theory for singular cases which have similar rules but contain certain conditions that break the Markov chain up into subsets\footnote{[4]}. The subsets then have their own asymptotic results. In tumorgenesis, it makes sense to use reducible processes since mutations can yield cancer cells represented by different types in the branching process. The probability of these cells eventually mutating back to normal cells is close to negligible, so we generally think that mutations only go in one direction. While not always true, this assumption allows us to put constraints on our p.g.f. when estimating the mass functions. However, theory about reducible processes relies on results from irreducible processes, so we continue under that assumption.
We define $\rho$ as the largest positive eigenvalue of the matrix $M$. This is also referred to as the *spectral radius* of $M$. Criticality of the process depends on the value of $\rho$. When $\rho > 1$, the process is called *supercritical*. A *subcritical* process has $\rho < 1$, and a *critical* process has $\rho = 1$. Since $\rho$ is the largest eigenvalue of the mean matrix, this describes the expected growth of the entire process. As one type gets larger, we expect all other types to follow the same pattern. The corresponding left and right eigenvectors for $\rho$ are $v$ and $u$ which remains consistent throughout the thesis. In this case, $u$ is a row vector of $k$ elements. These vectors are typically normalized such that $u \cdot v = 1$ and $u \cdot 1 = 1$.

Our interest in the growth of a multitype process is intertwined with the probability that a process goes extinct (all types have no living individuals). We denote the probability of eventual extinction $P(Z(n) = 0$ for some $n \geq 0) = q = (q_1, q_2, \ldots, q_k)$ where $q_j$ is the extinction probability of the process started by an ancestor of type $j$. Note that $q$ is the only solution to $s = f(s)$ [2]. The following lemma is the multitype analogue to Theorem 1.

**Lemma 1.** Assume $\{Z(n)\}$ is an irreducible BGW process with spectral radius $\rho$. If $\rho \leq 1$ then $q = 1$, and if $\rho > 1$ then $q < 1$.

The lemma states conditions for when a process goes extinct almost surely. In the case of irreducible processes, one type cannot go extinct on its own unless all types are doomed to extinction since the probability to revive it is still present. This is not necessarily the case with reducible processes.

The last basic properties to introduce with Galton-Watson processes are the asymptotics about how a process grows. We know that the mean grows according to $M^n$, but we can also show almost sure convergence of the process itself when normalized properly. We are mainly interested in supercritical processes, so only consider this case as we present the Kesten-Stigum Theorem for irreducible multi-
type Galton-Watson processes. The reducible version is presented later in the text. For this theorem, and throughout the rest of the text, we denote the expectation of a process with a single $i$-type ancestor as $E_i[Z(n)]$, or as a vector subscript when considering multitype types of ancestors, $E_i[Z(n)]$.

**Theorem 3** (Kesten-Stigum). Assume $E_i[Z_j(1) \log Z_j(1)] < \infty$ for all $1 \leq i, j \leq k$. Also assume $\{Z(n)\}$ is an irreducible branching process with $\rho > 1$. Then

$$\lim_{n \to \infty} \frac{Z(n)}{\rho^n} = vW \text{ almost surely.}$$

$W$ is a nonnegative random variable such that $P_i(W > 0) = 1 - q_i > 0$ and

$$E_i[W] = u_i \text{ for } 1 \leq i \leq k.$$  

Another interesting result in multitype processes that comes from these limit laws is the proportion of types approach the corresponding components of $v$. That is,

$$\lim_{n \to \infty} \frac{Z(n)}{u \cdot Z(n)} \to v \text{ almost surely.}$$

This is used when we consider the long-term proportion of individual cell-types and estimation of parameters based on these proportions in data. These laws become useful in determining the long-run nature of branching processes. They will also be used to set up possible estimators for the proportion of types, which we use along with the growth rate and current size to determine the time spent since initiation of the process.
Multitype Continuous Time Markov Branching Processes

Last, we include some information about the multitype analog of the Markov continuous-time process which the other processes lead up to. This will be one of our main tools in fitting clonal evolution models because it allows for overlapping generations while still giving us the convenience of exponential lifetimes, and as a result, the Markov property for these processes. Like the previous sections, this work comes from Athreya and Ney which sets up a good foundation for these processes as it does with the previous ones[2].

We begin by assuming a population can have \( k \) possible types for its individuals. A type \( i \) ancestor at time 0 lives for a random amount of time with an exponential distribution and rate parameter \( a_i \). Thus the vector of rates for the lifetimes of all individuals is \( \mathbf{a} = (a_1, \ldots, a_k) \). The stochastic process \( \{Z(t), t \geq 0\} \) counts the number of individuals of each type at time \( t \). \( \{Z(t), t \geq 0\} \) has the state space \( \mathbb{Z}^{+k} \) and transition probabilities \( P(i,j,t) = P(Z(s+t) = j | Z(s) = i) \) that satisfies the branching property,

\[
\sum_{j \in \mathbb{Z}^{+k}} P(i,j,t)s^j = \prod_{p=1}^{k} \left[ \sum_{j \in \mathbb{Z}^{+k}} P(e_p,j,t)s^j \right]^{i_p}.
\]

This condition essentially means the number of individuals in a process at time \( t \) started by \( i \) ancestors is equal to the number of individuals in \( |i| \) individual processes begun by a single ancestor with the types distributed according to \( i \). In other words, we can write the branching process at a time \( t + u \) as the sum of the individual branching processes left to grow for time \( u \) if they existed at \( t \), or

\[
Z(t + u) = \sum_{i=1}^{k} \sum_{l=1}^{Z_i(t)} Z_i^{(l)}(u).
\]
We have similar definitions for the probability generating functions of a multitype continuous-time Markov branching process as we do to the multitype BGW and one-dimensional markov processes. First, an individual with type $i$ lives for an exponentially distributed amount of time as described above before splitting into $j$ offspring with probability $p_{ij}$. Define the offspring probability generating function,

$$ f(s) = (f_1(s), \ldots, f_k(s)) $$

where $f_i(s) = \sum_j p_{ij} s^j$. Also, let $u_i(s) = a_i [f_i(s) - s_i]$. The probability generating function for the offspring of a type $i$ individual at $t$ is

$$ F_i(s, t) = E[s^{Z(t)} | Z(0) = e_i] = \sum_j P(e_i, j, t) s^j. $$

The Kolmogorov forward and backward equations for a system of differential equations, and are

$$ \frac{\delta}{\delta t} F_i(s, t) = \sum_{j=1}^p u_j(s) \frac{\delta}{\delta s_j} F_i(s, t) $$

and

$$ \frac{d}{dt} F_i(s, t) = u_i(F(s, t)). $$

These equations set up a linear system of differential equations which may sometimes be solved. There are many cases when the system can not be solved analytically though, even for simple branching processes. One such situation that we ran into in cell reproduction is a birth-death-mutation process. We assume cells of a type $i$ splits at death into 2 daughter cells of its own type or 1 of its own type and one of a different type. Alternatively, a cell may die and have no offspring. Such a situation leads to the formation of a system of quadratic ordinary differential equations known as Riccati Matrix Equations. This model does not fit into the class of Riccati equations with
a known solution. Solving these analytically requires knowing one solution in order to find others. This implies dealing with even Markov continuous-time processes can get increasingly difficult quickly.

However, we can proceed to discuss moments and asymptotic results which still allow us to determine some results. We assume the process is irreducible, which has the same definition as before. This assumption has been relaxed and limit theorems can be determined, but they have analogous results as the multitype BGW process[4].

First, define the mean number of type $j$ offspring from a type $i$ ancestor

$$m_{ij}(t) = E[Z_j(t)|Z(0) = e_i].$$

We can then set up the mean matrix of the process as

$$M(t) = [m_{ij}(t)]_{1 \leq i,j \leq k}.$$

From the theory on multitype Galton-Watson processes, we have the mean of the offspring p.g.f., $m_{ij} = \frac{d}{ds} f_i(s)$ which was used to construct the matrix $M$. Now, define the matrix $A = D_\mathbf{a}(M - I)$ where $D_\mathbf{a}$ is the diagonal matrix consisting of $\mathbf{a}$ in the diagonals, and $I$ is the $k$-dimensional identity matrix. Then it is known that

$$M(t) = \exp(At).$$

The matrix exponential is defined as $\exp(B) = \sum_{i=0}^{\infty} \frac{B^i}{i!}$.

Since $M(t)$ serves as the mean matrix at $t$, spectral theory about the growth of the process relies on eigenvalues of $M(t)$ instead of $M$. These are still closely related to $M$ through $A$ as described. Define the largest real eigenvalue of $A$ as $\lambda$ which has left and right eigenvectors $\mathbf{v}$ and $\mathbf{u}$. The eigenvalue $\lambda$ for $A$ implies that the largest
eigenvalue of $M(t) = \exp(At)$ is $e^M$ with the same left and right eigenvectors. We require $\mathbf{u}$ and $\mathbf{v}$ to be normalized such that $\mathbf{v}\mathbf{u} = 1$ and $\mathbf{1}^T\mathbf{u} = 1$. This becomes important when we use asymptotic results to show the frequency of types converges to $\mathbf{v}$, so it acts as a probability distribution.

An interesting note about the spectral radius, or largest real eigenvalue, or the mean matrix is its interpretation in the context of branching processes. Assuming an irreducible process again, the spectral radius dictates the mean rate of growth of the process. With this, we can attach the theory of branching processes to the theory of Malthusian growth, that populations will experience exponential growth. Thus, if a process escapes extinction early on, where stochasticity is most important, it should eventually follow more deterministic growth. More importantly, all types within the model will grow together due to movement between and within types. This is illustrated in the following results about the mean of a process.

The first convergence result we consider is the growth rate of the mean matrix. The result,

$$\lim_{t \to \infty} \frac{M(t)}{e^{-\lambda t}} = \mathbf{uv}$$

shows that the matrix $\mathbf{uv}$ is the result of normalizing the mean by its largest eigenvalue. This result can be considered more generally as a statement about limits of matrix powers and their spectral radii, or

$$\lim_{n \to \infty} \frac{B^n}{\rho^n} = \zeta \eta$$

where $\rho$ is the spectral radius of $B$ with associated left and right eigenvectors $\eta$ and $\zeta$. This result is standard for irreducible matrices and even holds in the reducible case[5].

As with all processes, we are interested in the criticality of multitype continuous
time processes. The results for criticality can be restated in different ways, and are actually the same as in the discrete case since the only difference is the scale of time. We define a process as supercritical, critical, or subcritical if \( \lambda > 0 \), \( \lambda = 0 \), or \( \lambda < 0 \). This will actually associate with the value of the spectral radius for the mean matrix, \( M \) and how it compares to 1. Critical and subcritical processes almost surely go extinct, and supercritical processes go extinct with probability \( q_i \) given an \( i \)-type ancestor. The value of \( q = (q_1, \ldots, q_k) \) is the solution to the equation \( u(s) = 0 \). A supercritical process goes to infinity if it does not go to zero, and for irreducible processes, this means all types should go to infinity together.

### 1.2 Outline of Dissertation

The rest of the dissertation is organized into three categories: models of clonal progression and evolution, the mathematical foundation of our model, and an application where we attempt to link progression to evolution using a branching process model. First, Chapter 1 provides background on the current methods used in systems biology to estimate tumor growth and selective advantage. We introduce relevant studies that attempt to model cancer progression and clonal evolution from a population dynamics point of view, and some that model clonal heterogeneity, and discuss their strengths and a few shortcomings we wish to address with our model. We include

Chapter 2 introduces the infinite-alleles branching process model from Griffiths and Pakes, Pakes, Taïb, and the applications by Wu and Kimmel[6][7][8][9]. Then we dive into the bulk of the thesis in a mathematical description of the multitype infinite-alleles branching process we created to model the accumulation of passenger and driver mutations in clonal evolution.

Finally, Chapter 3 consists of an application of our model where we use variant allele frequency data along with another model that infers the evolutionary hierarchy
of mutations in order to make inferences about growth rates and mutation rates using our model. We show how our model helps add on to previous knowledge and models of clonal evolution. If we can understand the phylogenetic structure of the model as well as knowledge about driver and passenger mutations, we should be able to reconstruct the evolutionary history of a tumor beyond its phylogeny. We can also estimate tumor growth and passenger mutation rate, or the rate of background mutations.

We conclude with some limitations of our model, including a discussion of the data used and ideal data for such a model. We believe we are currently limited by the type of data available, and discuss how this model could be more relevant and improved upon as we get a finer scope for data (e.g. single-cell sequencing, longitudinal studies, samples from many locations). We also include some topics for future research.
Chapter 2

Cancer Evolution in Tumorigenesis

Clonal Evolution of the Hematopoietic System as a Motivation for Tumorigenesis

The hematopoietic system produces enough blood over a lifetime to sustain an individual, balancing proliferation and self-renewal with differentiation. Because the hematopoietic system comprises multiple cell types, a hierarchy is established starting with the most primitive cells, hematopoietic stem cells. These have the ability to reconstitute the entire hematopoietic system. The next cells in the hierarchy are progenitor cells which still have self-renewal capabilities but are more differentiated than HSCs. Clonal expansion can occur in these cells, allowing for exponential growth so that the system can maintain itself in the event of injury or loss of cells[10]. A trade-off occurs between self-renewal capability and differentiation. However, when disregulation occurs via mutations and genetic alterations, acute leukemias can arise. Stem cells or progenitors can convert to a leukemic stem cell, initializing the spread of leukemia[11]. Multiple genetic changes must occur for leukemia to proliferate in the body. Loss of differentiation should be accompanied by increased proliferative capacity. It has been suggested these two functions occur as separate events, leading
to at least two hits required to initialize leukemia[10]. Loss of differentiation ability leads to a high number of immature blast cells that cannot function as mature cells, and end up building up in the form of a tumor (note that leukemic tumors are not solid localized tumors, but a large number of leukemic cells). Acute leukemia is heterogeneous due to the number of different genetic aberrations that may occur, which is why we have different markers and therapeutic targets. A common acute myeloid leukemia oncogene is the FLT3-ITD mutation which leads to increased proliferation and reduced apoptosis.

Clonal expansion is not limited to healthy cells or the hierarchy of differentiated cells. Mutations can also spread via clonal expansion, creating leukemia that consists of a heterogeneous population of cells, beginning with its ancestor, the cell of origin, which designates the initiating event of leukemia and cancer in general[12]. It is unclear how to determine the cell of origin in acute leukemias, so scientists need to understand the role of oncogenes in the progression of cancer.

Clonal evolution is typically depicted as a linear succession of events that cause and expansion leading to another event, and so on. From the point of view of the youngest clone, it makes sense to think of tracing the ancestry of its clones this way. However, in reality the acquisition of mutations is stochastic and can occur in any proliferating cell. Thus clones that are older continue to accumulate mutations even as newer clones arise, creating a nonlinear evolution of tumors. Many mutations and aberrations do not confer a growth advantage and are discarded. Those that are selected against should eventually die unless they acquire another mutation that gives clones an advantage before death. Others have no advantage and thrive or are discarded due to drift. Only a small minority have an advantage, and even these may not thrive in small numbers, when stochasticity plays a larger role. Clonal heterogeneity plays a large role in cancer evolution. Since it is a nonlinear process, it
provides a very heterogeneous population of clones for selection to work on, making
the cancer more likely to survive therapy. There is also the possibility for interactions
to occur between different cancer cells and networks between clones may arise[13].
The large diversity may also keep certain clones with functional properties at bay,
such as drug-resistant clones. A good example is chronic myeloid leukemia, where
treatment with imatinib reduces the cancer population, but relapse can occur. When
it does, cells may be resistant to imatinib. The drug resistant cells did not necessarily
arise as a response to treatment, but were present before treatment, but had the
same or lower fitness than other CML cells. Removing sensitive CML cells removed
competition and drug-resistant clones proliferated.

2.1 Motivation: Transition from severe congenital
neutropenia to secondary AML

Severe congenital neutropenia(SCN) is a syndrome characterized by a low number
of neutrophils. The number of myeloid precursor cells in the bone marrow are large
enough, but these numbers dwindle as the cells become further differentiated to ma-
ture neutrophils[14]. A low number of circulating neutrophil counts is also observed.
A common feature among patients suffering from SCN is an increased rate of apopto-
sis in neutrophils, and maturation ceases between the promyelocyte and myelocyte
stages of hematopoiesis[15][16]. SCN is a life-threatening condition due to the risk of
bacterial infection from the decreased number of neutrophils. In fact, Rosenberg et
al. reported half of the patients died within their first year from sepsis prior to the
discovery of a treatment[16].

Granulocyte colony stimulating factor (G-CSF) began to be administered to pa-
tients to stimulate granulocyte production which increases the proliferation rate of
granulocytes and mobilization out of the marrow. Thus neutrophil production is increased as a result, allowing these number to offset the number that die by apoptosis or become arrested during differentiation. Trials with G-CSF treatment showed an increase in mature neutrophil numbers and a decrease in infections, implying the treatment is successful with respect to SCN [17]. However, the increased production of granulocytes may not be without consequence, and an association has been found with SCN and progression to myelodysplastic syndrome and later acute myeloid leukemia[16][15]. A study by Beekman et al. found progression to AML has an overall cumulative incidence of 20% after 15 years of G-CSF treatment[14]. The reason for progression to AML is unclear, but it has been proposed that patients with SCN are more susceptible to developing AML due to genetic factors. An alternative idea, and the one that will be explored further, is that G-CSF treatment causes higher proliferation rates, and as a result, can contribute to the acquisition of additional mutations. Eventually, the acquired mutations can increase cell fitness, initiating leukemogenesis [18]. After progression from SCN to AML, mutations were found in 80% of patients, and cell line models show that a common set of defects in cells from SCN patients may cause a halt in differentiation and increased proliferation rates. The question of what causes the progression to AML is also difficult to determine because of the high mortality rates of SCN-affected patients at early ages. Death would occur prior to the formation of leukemia even in susceptible patients.

The hypothesis of how SCN progresses to secondary AML can be generalized to clonal evolution that occurs during tumorigenesis. As a population develops and maintains itself, mutations occur. Many of these mutations do not affect the cell fitness and are labeled passengers. Eventually, a mutation can occur that increases cell fitness initiating an event that leads to cancer. The cell that contains this mutation becomes the ancestor of the clonal population, and then goes on to split and further
mutate. It is important to note that that initiating mutant cell may also contain the previous mutations of its parents, but these are not considered drivers since they do not affect cell fitness. It has been shown, however, that there is a correlation between the number of passenger mutations that exist and age of the patient[19]. This may suggest some correlation between the number of passenger mutations that accumulate and the number of driver mutations, or progression of the cancer.

Further, a hierarchy of mutations exists as clones give rise to daughters with the same set of mutations. The clonal architecture of cancer populations has become the interest of numerous studies recently, and there are attempts to determine the timing or order of mutations in clonal populations. The order can lend information relevant to treatment options and genetic targets for therapy. Beyond the qualitative information given from sequencing, understanding how mutations affect the population dynamics of these populations could help us understand heterogeneity and rates of growth of cancer populations. A population dynamic point of view that includes information about population genetics can make a more complete model and warrants investigation, which is the aim of the study. We give background about an approach used to determine the hierarchy with respect to the mutations in the development of AML.

**Beekman et al. 2012 Study**

A current approach is to try look at the accumulation of mutations and attempt to determine which are driver and passenger mutations in the progression to leukemia during G-CSF treatment. Beekman et al collected hematopoietic stem cell samples from a single patient throughout his progression from SCN to AML. The patient developed AML after 17 years of G-CSF therapy [14]. Using next generation sequencing, the study found 12 nonsynonymous somatic mutations in leukemic blasts
at the end of the study. Three mutations occurred during early treatment, 15 years prior to AML diagnosis, and a clonal expansion of cells with these three mutations occurred which was discovered 9 years prior to diagnosis of AML. Additionally, AML cells were found within this population with nine additional mutations. A second hit to a previously mutated gene allowed for G-CSF independence.

The study shows a sequential gain of mutations in the development of AML from SCN. Other subpopulations of cells in the samples showed distinct mutations during the early phase, but these mutations were not present in the later phase, suggesting the subpopulations died out. In this study, it isn’t determined whether the cause of the mutations are due to clonal expansion and mutations encouraged by G-CSF or whether the patients were already susceptible to these mutations prior to treatment. Also, the authors do not conclude which mutations were drivers, increasing the fitness of the clonal cells. In order to determine this, the authors call for more studies of a similar nature to determine a similar set of mutations in other patients affected by SCN.

Implications of the study suggest leukemogenesis from SCN is a multistep evolution [18]. A dominant mutant is formed with higher fitness, outgrowing other mutant cells to become the main population of leukemic cells. Of course, the cells continue to evolve, so passenger mutations and other drivers may be added on affecting this fitness of the dominant population as time progresses. This suggests the heterogeneity and distribution of allele frequencies may hint at the sequence of mutations that occurred up to the point of sequencing. Our interest in this lies in studying the path of leukemogenesis and how the accumulation of mutations may lead to more frequent mutations. We study this from a population dynamics perspective, using branching processes to create our model.
2.2 Population Dynamics of Tumorigenesis

In most systems in the body, the number of cells is about constant and regulated by signaling, apoptosis, and other mechanisms. Typically cell fitness is equal among all cells in the same population, so only neutral selection, or genetic drift, occurs. That is, no particular cell should outcompete others, and cell genomes should ideally be homozygous. However as cell division occurs, somatic mutations can arise which can change the fitness of the mutant cells and the growth and differentiation capabilities [20]. Mutations that do not affect the fitness are known as passenger mutations, or neutral mutations, and may go unnoticed from a population dynamics perspective (the number of cells remain regulated), but create a more heterozygous population. These cells have no selective advantage over others that differ only by a similar set of passenger mutations [21]. Genetic drift still occurs within clonal populations that have a differing set if alleles due only to passenger mutations. Genetic drift is the random changes in cell populations with neutral mutations[22]. We say that given a set of mutations with the same selective advantage, the random fluctuations in subpopulations that differ on a set of passenger mutations are due to genetic drift. Thus, passenger mutations may arise and disappear throughout the lifetime of an individual stochastically, or some alleles with mutations may become dominant clones. Since there shouldn’t be any fitness advantage between these alleles, the outcome is random. Those mutations that change the fitness are known as driver mutations, and these can lead to the expansion of a subpopulation of cells different from the original population. Driver mutations can affect fitness by changing the rate of proliferation, resistance to apoptotic signals, or resistance to differentiation[23][24]. A mutation to a cell type with higher fitness can become the initiating factor of tumorigenesis, or leukemogenesis in the hematopoietic system, and can lead to the accumulation to further mutations and changes to the clonal population’s proliferative capabilities.
This initiating mutation becomes the ancestor and precursor to clonal evolution, or the sweeping waves of expansion that lead to a heterogeneous selectively more fit subpopulation we know as a tumor.

The current view of cancer progression is that a multistep accumulation of mutations leads to transformation of healthy cells into malignant cancer cells [25]. The transition happens over multiple replications, so cells evolve as proliferation occurs. Multiple sites on chromosomes may contain a mutation during intermediate steps en route to malignancy. Sequencing studies have shown that genomes undergo a large number of changes, but many of these mutations may be passengers, and do not affect cell fitness, and these studies have been backed by various stochastic models [26]. Many mutations may even occur before the initiating event, defined as the first driver mutation that leads to leukemia, and the number of mutations may be correlated with age [19]. These mutations prior to the initiating event are most likely unrelated to the pathogenesis, making them passenger mutations.

In order to distinguish between passenger and driver mutations more concretely, we give a definition to fitness, or selective advantage. We will define fitness of populations as the expected number of offspring per parent. Without distinguishing by type, fitness can be clearly linked to criticality in single-type branching processes. In multitype processes fitness can refer to the entire population based on the criticality (value of the spectral radius of the mean offspring matrix), or we can consider different subtypes and their fitness relative to other types. Regardless of the process, fitness has an easy connection to the parameters of a branching process. In order to maintain a constant number, the fitness in normal cells should be one, and they should change when regulatory signals encourage more proliferation or apoptosis (an example is injury repair). The process of leukemogenesis occurs during normal hematopoiesis, when a hematopoietic stem cell acquires a driver mutation. This driver mutation
gives this particular cell a fitness advantage over other cells, and from it expands a subpopulation of cells that contain this advantage as well as all passenger mutations that occurred up to the birth of this cell[27]. As time progresses, this clone develops new mutations-passengers and drivers-which can give rise to new subpopulations and we see evolution take over until we have a heterogeneous population of cells.

However, determining which mutations are passengers and which are drivers can be a difficult question. Further, finding a common set of driver mutations can be even more difficult due to the heterogeneity among patients. Another interesting study considered patients with myelodysplastic disorders and the transition to secondary AML. Next generation sequencing was performed on seven patients to sequence the genetic changes that occur during the progression from MDS to secondary AML. The authors performed whole-genome sequencing to identify somatic mutations during the MDS phase and s-AML phase of each patient. Interestingly, most mutations present in AML samples were also present in MDS samples along with an extra subset of mutations specific to AML, further supporting the hypothesis that leukemogenesis is due to the accumulation of mutations in clonal populations. Also, in all seven samples the founding MDS clone persisted, even though it was sometimes outcompeted by daughter subclones. The dominant secondary AML clone was even derived from the founding clone. New clones carried forward the existing driver and passenger mutations, creating a hierarchy of clonal evolution. Figure 2.1 gives a good illustration of the process of clonal evolution. The authors do not account for the possibility of mutants that do not confer a fitness advantage, leading to small subpopulations that grow and die temporarily. The mutations are broken up in clusters instead of individual mutations, but it appears each cluster contains a set of drivers that allow for larger fitness. The transition to MDS/sAML cells occurs in the second cluster of mutations, but the clinical diagnosis relies on myeloblast percentage, so detection
and diagnosis occurs later than the appearance of an ancestral MDS clone.

Figure 2.1: The clonal evolution of a patient from MDS to sAML suggests AML arises out of the accumulation of mutations that lead to subtypes which confer a fitness advantage of their parents to persist. The mutations are arranged in clusters according to their sequence or appearance and type[28].

However, the authors state that mutations in new clones must confer some growth advantage to persist in the population and compete with ancestral clones, suggesting that there is a possibility for subclones to exist which do not confer an advantage, but these clones will not survive. This study suggests that analyzing the clonal architecture can yield extra information about the pathogenesis of secondary AML from MDS, and more generally, leukemogenesis[28].

The rest of this chapter will focus on aspects of cancer progression and clonal evolution that need to be considered for us to use branching process models to understand the heterogeneous subpopulations that form a tumor and their individual parameters that lead to progression. We consider models of progression that have been studied for with varying number of mutations and discuss the state of cancer progression models. We then discuss some models of clonal evolution that attempt to determine the phylogeny of mutations in a tumor population. These models are useful because they can help develop the structure of a branching process offspring.
matrix in order to aid our estimation process. We will hold off on their application to our model until a later chapter.

2.3 Cancer Progression

In the original “Hallmarks of Cancer” paper, the authors established multiple properties of tumors as a result of the cells. Each of these properties are functional in how a tumor develops as a population, but specific properties are based on individual cell properties that result in tissue-wide changes [25]. We are interested in the change in fitness of overall cells, and how those changes effect the entire population. The changes come from some of the hallmarks set forth in the paper. Notably, cell fitness can changes due to evading apoptosis, resisting differentiation, and increased proliferative capacity. The functional impact of these types of mutations are more persistent growth of the subpopulation possessing these characteristics. We refer to these types of mutations as driver mutations. Studying the accumulation of driver mutations and their corresponding growth gives us an idea of which mutations might be important. In fact, Merlo et al. [21] suggested determining the actual growth rates of somatic mutations would lead to fundamental biomarkers for cancer progression. That is, we target the most aggressive cells due to specific mutations that make them grow the fastest or avoid apoptotic signals. Models of genetic cancer progression have attempted to understand the impact in growth of tumors using stochastic processes. While still mathematically advanced in their scope, many of these models use a number of simplifying assumptions to get a tractable solution. The most common assumption seen that we will further describe is the functional impact of additional mutations is the same and not unique to the actual mutation. In other words, studies treat all mutations as the same, and consider selection and fitness as a function of the number of mutations rather than the type of driver mutation.
Number of Driver Mutations to Cancer

Only a few drivers are necessary to lead to cancer [26], but we see many more mutations present due to passengers. Most early studies only consider the impact of drivers, but recent studies have started to include passenger mutations to determine if a relationship between the two exist (as in a mutator phenotype) [19]. Passenger mutations should not have any effect on the growth parameters. Recent studies have shown this is not entirely true, and passenger mutation accumulation may have a deleterious effect on cancer progression [29], but we’ll continue under the assumption that there is not effect.

Early models of cancer suggested multiple stages occur due to several successive changes that lead to higher mortality rates in patients with cancer [30]. By only looking at mortality rates among multiple types of cancers, the authors of this study show that cancer tends to progress where the logarithm of the rate of incidence is proportional to the log of the time. The authors infer that the time spent in each step towards cancer, $X_i$ is exponential with a rate of $u_i$ each year where $i$ refers to the step. Therefore, the probability that the $k^{th}$ step occurs in $(t, t + dt)$ where the steps occur in no particular order is $\frac{e^{-u_i t}}{(k-1)!} dt \prod_{i=1}^{k} u_i$ for small $t$, concluding that the slope of the log-log curve of age versus incidence was one minus the number of steps in the model, $k$. The model only applied to patients in a specific age group (25-74 years). This landmark model helped set up the idea that multiple lesions were responsible for the abnormal growth of tumors. Knudson became famous for his “two-hit hypothesis”, that two lesions were required in a gene in order to lead to cancer [31]. Later, Fearon and Vogelstein suggested that multiple unique events were required for the initiation and development of cancer [32]. They determined that mutations in around 4-5 genes were required in colorectal cancer, and Luebeck and Moolgavkar determined these steps to be the knockout of APC, activation of KRAS, and inactivation of TP53.
Out of these classic papers, numerous mathematical models for the population dynamics of a tumor undergoing multiple events (mutations) have been proposed. The stochastic versions of these models usually make the assumption of either constant populations (Wright-Fisher and Moran models) or exponentially growing populations (branching processes). Our focus is on branching process models, where we do not limit the tumor size.

Classic carcinogenesis models describe multiple clonal expansions that are preferentially selected for. These clones occur successively, being subtypes of their ancestors. In these models, heterogeneity is an important theme to the evolution of the cancer genome. Current research attempts to understand how cancer evolves and changes in space and time, providing two different types of studies: geographical and longitudinal. Geographical sampling occurs at a single time point and looks across multiple sites within a tumor, while longitudinal sampling is sampling at multiple times throughout the history of the cancer. We are mainly concerned with how a cancer evolves globally, so our interest is more concerned with longitudinal sampling. However, most characterizations of human cancers occur at a single point in time, representing multiple years of somatic mutations without a time-scale to determine the order and path of evolution that occurred. Attempts have been made to construct the order of the mutations based on the heterogeneity and proportion of clones at single time points in the population, and models estimating phylogeny are becoming more prevalent.

Another model of clonal evolution considers the population genetics of tumor suppressor genes and the evolution of a population of cells toward inactivating these genes. The authors assume a population of cells of size $N$ which is held constant in the model. There are three types of cells in the model, type 0, 1, and 2 which represent normal cells, cells with one of the TSG alleles inactivated, and cells
with both inactivated respectively. Mutation only occurs in one direction and we only consider a single gene, so type 0 can mutate to type 1 and type 1 can mutate to type 2. Events (births or mutations) occur according to a Moran process, and the birth rate for each type is $1, r$, and $a$. The mutation rate for a type 0 cell to become type 1 is $u_1$ and the mutation rate for type 1 to type 2 is $u_2$. Mutations occur at the events. An event occurs after an exponentially distributed random time. The type of cell that reproduces is randomly selected proportionate to the rates, and the type that dies is randomly selected with equal probability. The new daughter cell is of the same type with probability $1 - u_i$ or mutates to the next type with probability $u_i$ immediately at reproduction. The goal of both studies is to find expressions for the time until initiation of cancer (the time until a type 2 individual first arrives) conditioning on the type of the original population. The first study introduces a phenomenon known as stochastic tunneling, where a population consisting of only type 0 cells becomes a population of only type 2 cells with type 1 cells never dominating the population in between. They describe the inactivation of tumor suppressor genes as a normal population having fitness equal to 1 before a single allele is inactivated, dropping the fitness of the cell ($r \leq 1$), and ultimately inactivation of the second allele increases the fitness of the cells, so $a > 1$. At large $N$, some properties of the model are similar to branching process properties, and these models can be applied to mutations where genetic instability leads to increasing fitness, so each mutation makes the time to next mutation shorter.

The same group use a similar model to describe the evolution of resistance during clonal expansion where sensitive tumor cells grow and can acquire resistance [40]. A particular situation where resistant arises is in chronic myeloid leukemia, where growth can occur over a number of years and allow for imatinib-sensitive cells to acquire resistance prior to detection. If this occurs, chemotherapy will only remove
sensitive cells and resistant ones will continue to thrive, resulting in a relapse. Instead fixing the population at $N$ like the Moran model, the model is a branching process that grows until detection of a tumor which occurs when the population reaches size $N$. The authors create a 2-type birth-death process where type 1 comprises sensitive cells that can split or die at exponential rates ($r$ or $d$) and mutate upon splitting with probability $u$, and type 2 cells which are resistant tumor cells that can also split or die at exponentially distributed times (with rates $a$ and $b$ respectively). Type 2 cells cannot mutate back to type 1 cells. The goal is to determine the probability that a particular size tumor contains resistant mutants and the distribution of these mutants. Unlike the Luria-Delbrück model, the authors account for cell death as an event, but the authors show their model converges to this one in cases where they do not allow cell death. The relative fitness of resistant to sensitive cells in the model is $\alpha = \frac{a-b}{r-d}$, which determines whether or not resistant cells are selected for with respect to sensitive cells. Advantageous fitness occurs when $\alpha > 1$ and deleterious fitness occurs when $\alpha < 1$. There is no advantage in fitness when $\alpha = 1$. The process ends when extinction (no cells) or detection ($N$ cells) occurs. The authors also extend this model to include a second level of resistance, allowing type 1 cells to become type 2 in order to determine the probability of the presence of resistant mutations when a tumor reaches a specific size [41]. The additional level is useful in accounting for situations where multiple therapeutic options exist. Chronic myeloid leukemia can be treated with imatinib and dasatinib, which have different targets on the BCR-ABL domain, so two hits at different locations would be required to alter the genes enough to confer complete resistance in CML. A potentially important result that holds in these studies that is important is large death rates increase the chance of accumulating mutations in the model since cells that don’t mutate will have a high probability of dying, so the detection size is not reached. Large cell turnover will lead
to more cell divisions before reaching detection.

These particular models outline a general strategy of modeling and simulating birth-death processes by determining the Kolmogorov Equations for the model and attempting to solve and/or simulate. Simulations can take advantage of the Gillespie algorithm which is used in biochemical systems and systems biology\[42\]. The general idea is to reduce the birth-death branching process to a continuous-time Markov process where the next event occurs in a time that is exponentially distributed with a rate proportional to the current state of the system (i.e. the number of individuals currently alive). After the event time is determined, the type of event (birth/death/mutation) occurs based on a probability proportional to the number and rates of each type. Thus the simulation of the process is simplified by removing the need to simulate each individual’s lifetime and offspring throughout the process and only considering the next event globally.

A final model mentioned using the same scheme as above allows \( k \) mutations to occur where each type can split, die, or mutate into the next type\[43\]. The approach used is more rigorous than the previous studies so that other values can be found, such as the waiting time until the first type-k cell and the growth of the initial type of cells as a martingale. Letting \( Z_0(t) \) be the number of type-0 cells alive at \( t \) and \( \lambda_0 \) be the Malthusian parameter for type-0 cells, \( E[Z_0(t)] = e^{\lambda_0 t} \), so \( W(t) = e^{-\lambda_0 t}Z_0(t) \) is a martingale. By conditioning on non-extinction, the author establishes that \( (W(t)|Z_0(t) > 0) \rightarrow V_0 \sim \text{exponential}(\lambda_0/a_0) \) where \( a_0 \) is the birth rate for type-0 cells. The modification \( Z_0^*(t) = e^{\lambda_0 t}V_0 \) is used to simplify calculations. The author approximates the distribution for \( \tau_k \), the time until the appearance of the first type-k cell as \( P(\tau_k > t) \approx (1 + Ke^{\lambda_0 t})^{-1} \), where \( K \) is determined by the birth, death, and mutation rates of the process.

More recently, models incorporate passenger and driver mutations to incorporate
the effect of driver mutations on fitness with the heterogeneity added as a result of passenger mutations. These models have different types of cells that arise from mutations, typically having higher fitness as the mutations increase. One study uses a branching process to model the accumulation of driver and passenger mutations [26].

The model is a similar but simpler version of the one proposed. The discrete time branching process begins with a single individual mutant, representing the initiation of a tumor. At each time step (or generation), a type-$j$ cell splits with probability $b_j$ or dies with probability $d_j = 1 - b_j = 1/2(1 - s)^j$. The variable $s$ can be seen as the fitness advantage of the $j + 1^{st}$ mutation over the $j^{th}$. Further, the one of the daughters may mutate to become a type-$j + 1$ cell with probability $u$. Thus the offspring generating function for a type-$j$ individual is

$$f_j(s) = d_j + ub_j s_j s_{j+1} + (1 - u)s_j^2.$$  

Each offspring may also have a passenger mutation with probability $u$. The average number of type-$j$ cells conditioning on nonextinction grows with $\frac{1}{1 - q_j} (b_j (2 - u))^{\tau/T}$ where $q_j$ is the probability of extinction, $\tau$ is the time since the first appearance of a type $j$ cell and $T$ is the generation time. The authors use this to determine the average time until the first type-$j$ cell appears conditioned on survival. They also note that passenger mutations, which have no effect on fitness, only accumulate in individuals. At generation $t$, the distribution of the number of passenger mutations in an individual is binomial($t, v$) since at each generation individuals have an independent probability of having a passenger mutation. Using these results, the authors find the number of passenger mutations that accumulate with $k$ driver mutations is $n = \frac{v}{2\tau} \log \frac{4ks^2}{u^2} \log k$.

A continuous time formula is also provided, which is analogous to the discrete time one.

A similar model is used to describe the transition from severe congenital neutrope-
nia to secondary acute myeloid leukemia to determine if stochasticity in the process can account for the variation of times at transition seen in the data\textsuperscript{[44]}. The model is a Galton Watson process that allows for splitting or death after every generation and upon splitting, can mutate to the next type. A difference in the calculations with this model is the authors do not use expected time to next mutation in order to find the number of cells in the $k^{th}$ clone, allowing for variability during the process. However, the authors point out that the model and the previous one simplifies the process of leukemogenesis since it assumes each mutations increases fitness, which is not necessarily true\textsuperscript{[14]}. The model also does not account for passenger mutations and the number of passenger mutations associated with a specific number of drivers.

Lastly, Tomasetti et al. propose that half or more passenger mutations arise prior to tumor initiation, making it necessary to account for the growth of normal non-mutant cells in the population\textsuperscript{[19]}. This model is based on the idea that cells grow through clonal expansions from a single cell until reaching the proper number where proliferation rates decrease and cells divide only to maintain numbers. During this time, passenger mutations exist and subpopulations have neutral selection, experiencing only genetic drift. At some point a cell is born with higher fitness and expands, becoming the initiating event for cancer. By this time, many point mutations should have already occurred, and a correlation should be found between the number of mutations and age. The expected number of mutations at a time $t$, $E[X_S(t)] = Sut$, where $S$ is the number of DNA base pairs sequenced, $u$ is the probability per base pair of a mutation, and $t$ is the number of divisions before tumor initiation. This means the expected number of mutations in the first cancer initiating cell is $E[X_S(t)]$ instead of 0, which is commonly assumed in previous models. The study brings up the necessity for including growth and development of the population prior to initiation.
Population-dependent model of Hematopoiesis

The models for evolution of resistance lead us to try and apply multitype branching processes to hematopoiesis with the assumption of a protective niche that places hematopoietic stem cells (HSCs) in an essentially quiescent phase. It is hypothesized that the niche protects cells from chemotherapy and allowing for a relapse in leukemia after leaving the niche and reproducing after chemotherapy. Evidence of regulation of HSCs in the bone marrow lead to the idea of a microenvironment being present, although the exact role is less clear. The osteoblastic niche and vascular niche are the two proposed for HSCs, and studies have found signaling and adhesion occurs between HSCs and these cells. Along with protection, the HSC niche is thought to create homeostasis in blood by inducing dormancy in HSCs. When numbers drop, HSC cells can leave the niche and begin proliferating again until sufficient cells are created, when cells reenter the niche and become dormant again[45]. A functional definition of tissue stem cells was proposed to emphasize the functionality of cells and allow for more flexibility. The definition states tissue stem cells are functionally undifferentiated, home to an appropriate environment, proliferate, are capable of producing a large number of differentiated cells, can self-renew and self-maintain, can regenerate functional tissue, and reserve some flexibility in the use of those options[46]. This definition allows for the idea of stem cell plasticity, where a stem cell can become more or less capable of proliferation and differentiation while still being functionally similar[47]. This allows some flexibility and reversibility in the proliferative capacity of stem cells. This definition forms the basis of a stochastic model of stem cell organization.

Roeder and Loeffler proposed an agent-based model of stem cell organization that fulfills the functional definition of hematopoietic stem cells, provides a protective niche, and accounts for differentiation. The model, pictured in Figure 2.2 defines
cells by their age (time since last split), environment, and attachment affinity. The attachment affinity is a numeric term that describes a cell’s proliferative potential, or in terms of the model, is used to determine the probability of moving to the niche and becoming dormant. Cells proliferate by splitting every 48 hours in the growth environment, GE-Ω, or remain the same in the dormant environment, GE-A. At each time step, the affinity decreases while in the growth environment or increases in the dormant environment by a factor of $d$ or $r$ respectively. A cell can switch from GE-Ω to GE-A at and time point with probability $\alpha$ and can witch back with probability $\omega$, which are both functions of the affinity and the current number of cells in the other environment. Specifically, for a cell,

$$\alpha = \frac{a}{a_{max}} \left( A + B \cdot \exp\left( C \frac{N_A}{N_{max}} \right) \right)^{-1} + D \quad \text{and} \quad \omega = \frac{a_{min}}{a} \left( A + B \cdot \exp\left( C \frac{N_N}{N_{max}} \right) \right)^{-1} + D$$

where $a \in (a_{min}, a_{max})$, $N_{max}$ is a scaling factor for the number of cells, and $N_{A/Ω}$ is the number of cells existing in GE-A/Ω. The result creates a model where cells that are in GE-Ω tend to remain in proliferation and become lose their attachment affinity before becoming differentiated. However, this probability is balanced by the number of cells in GE-A, so as GE-Ω loses cells to differentiation, they are pulled from GE-A and proliferate until proper numbers are restored, at which point GE-A numbers are restored. The model is built on a heuristic set of assumptions, and become hard to analyze. For instance, the probability of transitioning is based on a sigmoidal function without much rationale, and the affinity concept is very abstract. The resulting model become difficult to analyze mathematically. However, the model allowed the authors to break up leukemogenesis into three parts to model relapse of CML: genesis of chronic myeloid leukemia, chemotherapy, and the relapse by altering parameter values in the model.

Figure 2.3 shows a result of simulation using the Roeder model with a set of
parameters provided for testing. The growth environment is in green while the niche is in blue. When both environments reach a maximum, new events should differentiate (red), which we consider as leaving the system in our model. These will eventually die after a few expansions. Roeder describes the model in three stages. A growth stage which leads to a plateau until time reaches 4000. At this point, Treatment occurs which kills a percentage of the growth environment every time step for 2000 time points. Finally, treatment is ended, and we see the same population increase as before. This shows that the niche serves as a pool to feed the growth compartment when it is under duress. That is, the niche does not necessary directly feel the effects of the treatment, but there is a flow of cells from the niche to the growth environment. When treatment is finished, the cells that are still alive have the opportunity to grow again. Even if all cells in the growth environment were targeted, cells in the quiescent environment can still lead to renewal of the population.

Since this model is built of heuristic concepts and is mainly simulation driven, we attempted to simplify the model by finding a suitable stochastic process that could
Figure 2.3: A test simulation of the Roeder model shows that cells in the growth environment, quiescent (niche) environment, and differentiated cells under treatment with a relapse as treatment is quit.

mimic these results. Since migration between environments can occur at any time in the model by Roeder, we decided to use a birth-death process with a migration component to allow this in our model as well. The model is density dependent, so the probability of migration or birth in the next $\Delta t$ is based on the current number of cells in the system. At time $t$, suppose there are $n$ cells in the growth environment and $m$ cells in the niche environment. Cells in the growth environment split at a rate $\lambda(n, m) = \lambda_0 n(1 - n/K_1)$, migrate to the niche at a rate $\mu_1(n, m) = \mu_1^* n(1 - m/K_2)$, or differentiate and leave the system at a rate of $\nu(n, m) = \nu m$. Cells in the niche migrate to the growth environment with rate $\mu_2(n, m) = \mu_2^* m(1 - n/K_1)$. $K_1$, and $K_2$ are the carrying capacities of the growth and niche environments. Figure 2.5 illustrates the potential actions of cells in the model. We reduced the agent-based model’s probability of migration to a function of the current density only, removing the affinity term. If we $X(t) = (X_1(t), X_2(t))$ be the number of cells in the growth and quiescent environments respectively, then we can write out the probability for changes in states in the model:
Figure 2.4: The state space for a 2-type logistic birth-death-migration process as described with a growth and quiescent environment.

- \( P[X(t + \Delta t) = (n + 1, m) | X(t) = (n, m)] = \lambda(n, m)\Delta t + o(\Delta t) \)

- \( P[X(t + \Delta t) = (n - 1, m + 1) | X(t) = (n, m)] = \mu_1(n, m)\Delta t + o(\Delta t) \)

- \( P[X(t + \Delta t) = (n + 1, m - 1) | X(t) = (n, m)] = \mu_2(n, m)\Delta t + o(\Delta t) \)

- \( P[X(t + \Delta t) = (n - 1, m) | X(t) = (n, m)] = \nu_1(n, m)\Delta t + o(\Delta t) \).

This leads to the state space seen in figure 2.4. \( X(t) \) is a two-dimensional Markov chain with state space

\[ \{(n,m), n = 0, \ldots, K_1, m = 0, \ldots, K_2\}. \]
From here, we can write the Kolmogorov equations for the model,

\[
\frac{dp_{n,m}}{dt} = \lambda(n - 1) \left(1 - \frac{n - 1}{K_1}\right) p_{n-1,m}(t) + \mu_1(n + 1) \left(1 - \frac{m - 1}{K_2}\right) p_{n+1,m-1}(t) + \\
+ \mu_2(m + 1) \left(1 - \frac{n - 1}{K_1}\right) p_{n-1,m+1}(t) + \nu(n + 1) p_{n+1,m}(t) - \\
- \left(\lambda n(1 - \frac{n}{K_1}) + \mu_1 n(1 - \frac{m}{K_2}) + \mu_2 m(1 - \frac{n}{K_1}) + \nu n\right) p_{n,m}(t),
\]

for \( n = 0, 1, 2, \ldots, K_1 \) and \( m = 0, 1, 2, \ldots, K_2 \) with the initial condition \( p_{i,0}(0) = 1 \).

The Kolmogorov equations can be used to derive the p.g.f., MGF, and determine moments. Differential equations for the first moments are

\[
\frac{dE[X_1(t)]}{dt} = \lambda E[X_1(t)] \left(1 - \frac{E[X_1(t)]}{K_1}\right) - \nu E[X_1(t)] - \\
- \mu_1 E[X_1(t)] \left(1 - \frac{E[X_2(t)]}{K_2}\right) + \mu_2 E[X_2(t)] \left(1 - \frac{E[X_1(t)]}{K_1}\right) - \\
- \frac{\lambda}{K_1} Var[X_1(t)] + \left(\frac{\mu_1}{K_2} - \frac{\mu_2}{K_1}\right) Cov[X_1(t), X_2(t)]
\]

\[
\frac{dE[X_2(t)]}{dt} = \mu_1 E[X_1(t)] \left(1 - \frac{E[X_2(t)]}{K_2}\right) - \mu_2 E[X_2(t)] \left(1 - \frac{E[X_1(t)]}{K_1}\right) - \\
- \left(\frac{\mu_1}{K_2} - \frac{\mu_2}{K_1}\right) Cov[X_1(t), X_2(t)]
\]

where \( X_1(t) \) is the number of cells in the growth environment and \( X_2(t) \) is the number in the niche.

The problem that comes up with these processes is the issue of nonlinearity. We see second moment terms in the moment differential equations which require their own set of equations to solve. Moment closure methods exist can be used to approximate solutions to the differential equation by letting higher moments equal zero and solving the system with respect to the lower moments. If we remove the second moment terms
from the equation, we’ll have the deterministic system:

\[
\frac{dX_1(t)}{dt} = \lambda X_1(t) \left(1 - \frac{X_1(t)}{K_1}\right) - \nu X_1(t) \\
- \mu_1 X_1(t) \left(1 - \frac{X_2(t)}{K_2}\right) + \mu_2 X_2(t) \left(1 - \frac{X_1(t)}{K_1}\right)
\]

\[
\frac{dX_2(t)}{dt} = \mu_1 X_1(t) \left(1 - \frac{X_2(t)}{K_2}\right) - \mu_2 X_2(t) \left(1 - \frac{X_1(t)}{K_1}\right)
\]

with initial conditions \(X_1(0) = i\) and \(X_2(0) = 0\). In order to solve for our parameters, we find the equilibrium solution by setting \(\frac{dX_1(t)}{dt} = \frac{dX_2(t)}{dt} = 0\). At equilibrium,

\[
X_1 = \frac{K_1 \left(\lambda_0 - \nu_0\right)}{\lambda_0}
\]

and

\[
X_2 = \frac{K_1 K_2 \mu_1^* \left(\lambda_0 - \nu_0\right)}{K_1 \mu_1^* \left(\lambda_0 - \nu_0\right) + K_2 \mu_2^* \nu_0}.
\]

If we let \(R_0 = \lambda_0/\nu_0\) and \(R_1 = \mu_1^*/\mu_2^*\), then we know that a steady state solution exists when \(R_0 > 1\) and \((0, 0)\) is a solution for \(R_0 \leq 1\). Also, \(X_1\) approaches its carrying capacity \(K_1\) as \(R_0 \to \infty\) and \(X_2\) approaches \(K_2\) as \(R_1 \to \infty\). After determining the equilibrium solutions, we could solve for \(nu_0\) as a function of \(\lambda_0\) or \(\mu_2^*\) as a function of \(\mu_1^*\). Estimating \(\lambda_0\) and \(\mu_1^*\) became an issue and a proper solution was not fully explored. For comparison to the agent-based model, we fit the model by setting the steady state values to those of the agent-based model and graphically checking the mean curves.

Quasistationary results were also explored using results from Darroch and Seneta [48]. Since \((0, 0)\) is an absorbing state, we want to understand the stationary distribution of the Markov chain conditioning on nonextinction. If a continuous time Markov chain exists with infinitesimal transition probability matrix \(R\) having a submatrix \(C\)
Figure 2.5: A simple description of actions that can occur in the model. Birth ($\lambda$), migration ($\mu_1$ or $\mu_2$), and death ($\nu$) are the three potential actions and rate which is based on the current state of the system.

for the transient states such that

$$R = \begin{bmatrix} 0 & 0' \\ q_0 & C \end{bmatrix}$$

then the transition matrix $P(t) = [p_{i,j}(t)] = exp\{Rt\}$. Further, if $Q(t) = exp\{Ct\}$, then

$$P(t) = \begin{bmatrix} 1 & 0' \\ p_0(t) & Q(t) \end{bmatrix}$$

so the quasistationary distribution can be solved in a method similar to that for solving a stationary distribution. However, $Q(t)$ is substochastic, so normalizing it to make it stochastic is necessary. $C$ has an eigenvalue, $\rho < 0$, with the largest real part. Suppose $\rho$ has left eigenvector $\nu$. Then $\nu$ can be normalized by $\tilde{\nu} = (\nu'1)^{-1}\nu$. If the quasistationary distribution exists, it is $\tilde{\nu}$. Also, the mean time
Figure 2.6: Multiple simulations using the birth-death-migration process show an approximately normal quasistationary distribution in the joint distribution.

until absorption can be given by \( p(0)'[-C^{-1}]1 \) where \( p(0) \) is this initial distribution. Applying this to our model requires reducing the model down to a single dimension. Since the number of states is finite, this is possible and not very complicated, but it is important to note there are \((K_1 + 1) \cdot (K_2 + 1)\) possible states for the Markov process, so approximations become desirable such as diffusion approximations. Figures 2.6 and 2.7 show a simulated quasistationary distribution after multiple simulations. We see an approximately multivariate normal distribution between the growth \((X_1)\) and quiescent \((X_2)\) cells.

Comparing the results of the model to those of the agent based model in Figure 2.8, we can see similar structure in the two curves. The blue trajectories represent the number of cells in the growth compartment while the green trajectories represent
Figure 2.7: Multiple simulations using the birth-death-migration process result an approximately normal quasistationary distributions in the marginals. Normal curves are plotted in blue with the same mean and variance as the data to show similarity.
the number of cells in the niche. The parameters could be adjusted some so the ODE solutions which approximate the mean of the birth-death-migration process increases at a faster rate in the growth compartment. One noticeable difference is the variance in the different models. We see much larger variance in the niche environment than the growth environment of the agent-based model, yet the growth environment has the larger variance in the stochastic process model. It should be noted though, that the formulation of the agent-based model did not emphasize the variance of the process when fitted to data, so our model may not be entirely inaccurate. Unfortunately, we did not have data to test our results.

Some complications with the model will arise when trying to determine the quasistationary distribution. The large number of dimensions will create a large transition matrix, so that even if the carrying capacities are relatively small, matrix computations will take long. In reality though, our carrying capacities will be extremely large, making computation of eigenvectors and an inverse matrix unfeasible. The agent-based model created was difficult to use and extract information such as growth rates from as well as cell counts at specific times. Thus extracting data from the agent based model for comparison was not possible, and visual comparisons could only be made. Preliminary mathematical results and the development of a diffusion process based on the birth-death migration process were presented as posters. This model extends previous results about evolution of resistance by trying to create a more parsimonious version of the agent-based model described above. The model assumes we are only interested in counting cancer cells with specific mutations. For CML, this would be the BCR-ABL tyrosine kinase inhibitor that can undergo imatinib treatment. The model can be extended to models of cancer evolution by adding an extra dimension to each compartment to reflect additional mutations. However, this model can become complicated very quickly when trying to determine transitions between
Figure 2.8: Multiple simulations using the agent-based model (top) and the birth-death-migration process (bottom) show similar structures, but slight differences in the increasing portions occur.
Figure 2.9: Tumors evolve over time as new mutations give rise to subclones with varying levels of fitness. Mutations that do not affect fitness (passengers) are shown in gray. Those that increase fitness (drivers) are given as different colors. Two evolving landscapes from tumors can lead to similar subclonal frequencies over time. Figure taken from [21]

mutation state spaces along with compartment state spaces.

2.4 Clonal Evolution

Tumors contain a heterogeneous population of cells that result from a sequence of somatic mutations [49]. Intratumor heterogeneity is a byproduct of clonal evolution, where somatic mutations initiate new subpopulations in the tumor which have the capability for differing levels of fitness based on the new mutations (driver or passenger mutations). We show this process of evolution in Figure 2.9 [21]. In the figure, we see subclones with driver mutations as colored populations and those with passenger mutations as gray populations. The previous sections discuss the effect of each of these types of mutations which is illustrated in this figure. However, our interest lies in the age and frequency of each population. We see here than the yellow clone is the oldest, and these mutations should persist in all subclones as a result. Any subpopulations born will contain their new mutations as well as the mutations of their descendants.

From a clinical perspective, the search for therapeutic targets would be most
effective if those targets are either the most aggressive or most ubiquitous variants in the population[50]. According to the theory of clonal evolution, the mutations seen in most of the population should also be the oldest mutations from a phylogenetic sense, and these mutations should be present in all subclonal populations as well. Thus targeting these would impact all subclones. Because of this, methods from evolutionary biology are employed to determine the clonal architecture in tumors[21]. However, we are faced with some challenges in determining the order and phylogeny of mutations that make this task more difficult. First, we do not always have access to complete information. Figure 2.9 shows two possible evolutions of a tumor containing neutral and driver mutations. In this model, we are considering the frequency of somatic mutations over time. Unfortunately, we may only have information at a few points of time, at different locations, or even on at the time of diagnosis and removal of a tumor. The figure shows two different evolutions that may lead to the final result. To counter this, we make some assumptions going forth.

The infinite sites assumption states that there are such a large number of locations for a somatic mutation to occur that the probability of seeing the same mutation in two different clones is infinitesimally small [51]. Thus, we can assume the left hand figure is more likely since the right hand side has the same mutation (orange) occurring in different clones from the ancestor clone (yellow). This assumption constrains any models we create.

We also assume the tumor derives from a single clone. That is, a single cell serves as an ancestor to the entire tumor population [52]. A driver mutation occurs in this particular cell which gives it higher fitness than normal cells. As a result, further mutations may arise with varying levels of fitness. However, the assumption relies on that first tumor-initiating cell not going extinct. There can be multiple opportunities for a cell to become the precursor cell, and any driver mutation in normal cells may
lead to a subpopulation with higher fitness, but the subpopulation has to persist and give rise to further mutations that are characteristic of cancer cells [49].

Given multiple sequences at different time points, we can add other constraints to the phylogeny. We show a few models that use sequencing studies at multiple time points in the following subsection. Having multiple time points (or locations) will allow us to determine the age and/or growth rates of clones relative to each other. For example, if we consider the time halfway through the left hand side of Figure 2.9 and at the end, we see new mutations in the end time relative to the other time point. This tells us the mutations are younger and gives an idea of their age. It will also give an idea of the rate of growth relative to the earlier clones. Samples across space can achieve the same result. We can consider multiple sites or at metastases and use the heterogeneity at these sites to help infer the phylogeny of the locations as well as the mutations within [53]. As a tumor grows, we expect cells to be surrounded with neighbors from the same clone. A new clone may develop and grow, but they should not mix in a solid tumor, so neighboring cells will be similar. Thus, while the entire tumor is heterogeneous with respect to the clones, specific points should contain cells from the same or only a couple clones. This finer resolution in time or space simplifies our ability to find a phylogeny.

Resolution can also vary at specific times or locations due to the sampling methods. Different sequencing methods give us different distributions for variants in a population. The highest resolution available is from single-cell sequencing, where a population of cells is sequenced at the cellular level. That is, one sample from the population refers to the DNA from a single cell. Though a relatively new technology, this allows us to consider joint distributions with respect to the frequencies of all variants. We can then infer correlations and determine orderings based on the presence and absence of variants in each cell. Hierarchical clustering can be used to determine
the clusters and their order. Cells that are very different in terms of their variants should be further apart, and those that have similar sets should be close. Ideally the oldest variants should be near the top. We could create trees of different resolutions as well, down to individual cells, but we would be more interested in creating clusters with a similar set of driver mutations.

On the other end of the spectrum lies the most available types of sequencing data: whole-genome and whole-exome sequencing studies use a sample from a tumor and perform sequencing on this. This method gives frequencies of somatic mutations in the entire sample without distinguishing between the individual cells. Thus, we are given marginal frequencies for individual variants without correlations. These methods are used for data collection in The Cancer Genome Atlas, where large amounts of data exists. However, we need to use inferential methods to determine the phylogenies and recognize the variability that may exist.

We could create a branching process model of clonal evolution and progression that accounts for clonal evolution, but would phase some problems without determining the phylogeny first. Such a process would be multitype, where each type would represent a different clone. However, if we have \( K \) possible mutations, there are \( 2^K \) possible clones that could arise, so without any other information we would have a large process. If we assume irreversibility of mutations, we would need to create a branching process with a mean matrix that is an upper diagonal block matrix with the form:

\[
M = \begin{bmatrix}
M_0 & M_{01} & 0 & \ldots & 0 \\
0 & M_1 & M_{12} & \ldots & 0 \\
0 & 0 & M_2 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & M_K
\end{bmatrix}
\]
where the rows and columns can be broken up based on the set of mutations (indexed by the natural numbers). \( M_0 \) would simply refer to the mean number of normal offspring from normal parents. \( M_n, n \neq 0 \) is a \( K - n \)-element diagonal matrix representing the mean number of offspring of the same clone, and \( M_{n,n+1}, n \neq 0 \) is a matrix representing the mean number of offspring with an additional mutation given by the indices in \( M_{n,n+1} \). For example, if we consider a population consisting of three types of mutations and normal cells (\( \emptyset, A, B, C \)), we can have 8 possible clones: \( \emptyset, A, B, C, AB, AC, BC, ABC \). Thus, we can see any evolutions along the path given in Figure 2.10. The resulting mean offspring matrix from this process is

\[
M = \begin{pmatrix}
\emptyset & A & B & C & AB & AC & BC & ABC \\
\emptyset & m_{0,0} & m_{0,1} & m_{0,2} & m_{0,3} & 0 & 0 & 0 & 0 \\
A & 0 & m_{1,1} & 0 & 0 & m_{1,4} & m_{1,5} & 0 & 0 \\
B & 0 & 0 & m_{2,2} & 0 & m_{2,4} & 0 & m_{2,6} & 0 \\
C & 0 & 0 & 0 & m_{3,3} & 0 & m_{3,5} & m_{3,6} & 0 \\
AB & 0 & 0 & 0 & 0 & m_{4,4} & 0 & 0 & m_{4,7} \\
AC & 0 & 0 & 0 & 0 & 0 & m_{5,5} & 0 & m_{5,7} \\
BC & 0 & 0 & 0 & 0 & 0 & 0 & m_{6,6} & m_{6,7} \\
ABC & 0 & 0 & 0 & 0 & 0 & 0 & 0 & m_{7,7}
\end{pmatrix}
\]

Creating a model this way would require us to estimate \( \sum_{i=1}^{n} \binom{n}{i} (i+1) = 2^{n-1}(n+2) \) total variables in the mean matrix if no constraints are placed. Considering \( m_{ij} = \frac{d}{ds_j} f_i(s)|_{s=0} \), the number of variables to estimate with limited data becomes staggering. Thus results would be dependent on the proper resolution of the data, or using other methods to determine the phylogeny and creating an informed mean matrix based on these methods. Thus, we consider the phylogeny and evolution of the clone as two problems where we separate the ordering of clones from the dynamics of clonal
Figure 2.10: A model with three possible variants has the above possible evolutions given by any of the paths beginning with a cell with no mutations represented by $\emptyset$. This means, at the time of sequencing, it is possible to have any combination of genomes from the path taken.
Figure 2.11: If we have an idea of the path taken, we can assume values of the mean matrix are zero, reducing the number of parameters in the model. This particular example leads to a reduction from 8 possible types to 4 types.

growth. For example, consider the above example, and we assume a particular path was followed as given in Figure 2.11. This particular evolution would tell us that the only types of cells we expect to see would be normal, $B$, $AB$, and $ABC$, reducing our typespace by half. In terms of the mean matrix, our informed mean matrix would be

$$
M = \begin{pmatrix}
0 & B & AB & ABC \\
\emptyset & m_{0,0} & m_{0,1} & 0 & 0 \\
B & 0 & m_{1,1} & m_{1,2} & 0 \\
AB & 0 & 0 & m_{2,2} & m_{2,3} \\
ABC & 0 & 0 & 0 & m_{3,3}
\end{pmatrix}
$$

and our number of of parameters to estimate is reduced to 7. Thus, information about the phylogeny simplifies our process.

Evolutions do not have to be linear. Branching can occur, but we do not expect to see any merging of paths after branching events occur due to the infinite sites assumption. Models of clonal evolution that we discuss take this into account when trying to estimate a phylogeny.
Models of Clonal Evolution

Models of evolution can be simple or complicated based on the resolution of data. If we have multiple time points or sites and very few mutations to consider, we can infer evolution by ranking the frequencies or presence/absence of mutations at certain sites and times. For example, if we consider two mutations at two time points and mutation A is present at both times, and mutation B is present in only one time, we assume A and B occurred at different time points which is an obvious inference. If we look at it in space, the same conclusion can be inferred, that the mutation with higher frequency is older.

Reid et al [55] considers the order of mutations and dependence by looking across patients for consistent orderings with the goal of creating a model of the events that lead to cancer in Barrett’s esophagus. Single cell sequencing was also used to determine the evolutionary history of a tumor in the breast by sampling multiple sites of the tumor and determining the ploidy of cells to determine major subpopulations [54]. The phylogenetic tree was reconstructed be using breakpoints determined by changes in copy number in the associated dendrogram. Most methods with a large number of mutations may require clustering of variants into common clones prior to creating a phylogenetic tree [56] to reduce the number of clones to order. Doing so assumes mutations occur simultaneously, or within enough time such that only one dominant subclone arises from each set of similarly clustered mutations.

These methods described work well based on having multiple samples or consistent orderings across patients. Instead of having data with this resolution, we consider situations when we only have the worst resolution: a single sample from the tumor with a list of variants found and their corresponding sequences in the entire sample. We are not able to make inferences about specific pairs of mutations without some error. Methods are available that use only variant allele frequencies in order to de-
Figure 2.12: The basic method for inferring clonal evolution from single nucleotide variants involves obtaining frequency data for each variant, determining primary clusters for clones, and estimating evolutions based on variant frequencies. This can be used to estimate clonal frequencies. Figure from [57].

termine clonal phylogeny. Figure 2.12 shows the overarching schematic for inferring phylogenies common to most methods [57]. The three main steps involve obtaining frequencies for variants, clustering variants into common subclones, and recreating the evolutionary history based on the frequencies. Certain assumptions are made to determine the orderings, but the main idea that remains is clones with larger frequencies should be older, and a split into two different daughter clones can only occur if the sum of the daughter frequencies is less than that of the parent[52].

We cover in depth one particular algorithm that we will use later with real data. Phylosub attempts to recreate the evolutionary history in a single person using a single or multiple tumor samples[52]. It makes the same assumptions mentioned before. Every tumor arises from a single cell, each variant only appears once (infinite-sites assumption), and tumor cells that derive from ancestors gain an advantage in fitness,
reflected by the rate of growth or increased probability of splitting vs. cell death. The model recreates the phylogeny as a binary tree. For a set of three single nucleotide variants, two possible graphs can be created assuming the largest frequency is the root. There is either a single leaf, and the tree shows linear evolution, or two leaves extend from the root, and branched evolution occurs. The latter case can only occur if the sum of the frequencies of the leaf SNVs is not greater than the frequency of the root SNV. However, these are not the only rules that can determine the topology of a tree. For instance, because branching could occur does not mean that it has to, so other methods need to be used to determine the tree structure of the phylogeny.

Phylosub puts a nonparametric prior over all directed trees representing the major subclone lineages (after clustering). Each node in the tree is an associated subset of SNVs present in the lineage, but not present in the ancestors of that clone. The node is also associated with the population frequency of cells with that specific genotype. The genotype is determined from the SNV group of that clone as well as the genotype of its ancestors. The result of Phylosub is a MCMC model that gives the likelihood of a given tree using read counts for each SNV. The data, read counts, is given as the parameters $a_i$ which is the number of reads without the variant $i$ and $d_i$ for the number of reads mapping to the $i^{th}$ loci. The variant allele frequency is then $1 - a_i/d_i$.

For $N$ total variants, the variants are clustered using a Bayesian finite mixture of $K$ components in order to put similar variants in the same cluster. The clustering is done hierarchically using a two stick-breaking procedure in order to create a rooted tree structure. This tree structure serves as the prior to model allele count data, accounting for the possibility of sequencing errors. Phylosub is an MCMC method where a single Gibbs sampler involves sampling cluster assignments, stick lengths for the tree, hyper parameters for the model, and then SNV frequencies. Sampling SNV frequencies involves the constraints mentioned above as well. Instead of sam-
pling the frequencies for each variant though, they sample the weights of each clone, which can be found by subtracting the frequencies of the descendental variants. Then, the variant frequencies can be rebuilt by adding the weights of a clone and all its descendants. A Metropolis-Hastings algorithm is used in this step to determine the posterior distribution for the clonal frequencies. Along with this, the authors create a partial order plot for all variants. This plot gives an idea of the hierarchy of the variants and order of their evolution individually. This is based off all runs of the Gibbs sampler and associated likelihoods, and does not include clusters of variants. They also given the top trees and clusters based on having the best likelihood, which we can use to construct a mean matrix for our branching process as discussed in the following section.

**Connecting Clonal Evolution and Progression: Using Clonal Evolution Results to Inform a Branching Process Model**

We could use branching processes to create a model of clonal evolution, or use previously created models to determine the structure of the mean matrix and process. Branching process models can then be used with the informed model to estimate the rates left. There are a couple connections needed in order to get the data from variant allele frequencies in a way that works with branching processes. The frequencies are marginal probabilities. A branching process at any time gives the number of cells with respect to a type, where types are cells with a specific set genome.

Fortunately, reconstructing a phylogeny allows us to determine joint frequencies from the marginals which is done by Phylosub [52]. If we have two variants at different loci, A and B, then let A belong in the ancestor clone and the descendental clone has the genome with AB. The frequency of the type AB cells is the same as the VAF for B. The frequency of the type A cells is the VAF for type A minus the VAF for B.
That is, leaves have frequencies equal to the estimated VAF, and ancestors require subtracting all descendent frequencies. The end results are frequencies of all types, serving as our data.

We would also need to determine how to get branching process results in the form of frequencies. Assuming a $k$-type branching process, the frequency for the $i^{th}$ type is

$$\delta_i(t) = \frac{Z_i(t)}{|Z(t)|}.$$ 

If we can determine the distribution function for $\delta(t)$ given the time of the process, we could use this to make inferences about the parameters and create confidence intervals.

Our data in this case suffers from a few problems to address. First, our data is whole genome or exome data at surgery. We are given allele frequencies and can use previous methods to infer a phylogeny. We then need to determine the amount of time a tumor has grown. Using the asymptotic property in multi type branching processes,

$$\frac{Z(t)}{e^{at}} \to Wv \text{ almost surely}$$

tells us that given the spectral radius, time, and left eigenvector of the matrix, we can approximate the number of cells of each type. The left eigenvector actually represents the long-term proportion of cells of each type, so we can use the frequencies from the data for this [2]. Another way of writing it is in the form of the mean matrix that we’ve worked with earlier. Given a type $i$ ancestor,

$$e_i^T M(t) \sim e^{at} u_i v.$$ 

The process is outlined in Figure 2.13. We would then need to determine how to find $\alpha$ or $t$ in order to use the approximation. We approach this problem more
Figure 2.13: The schematic shows the general approach to estimating the necessary parts to fit the allele frequencies into a branching process model. Estimation of frequencies and tumor sizes come from previous works, and branching processes can be used to determine the subclonal growth rates based on inferred sizes. Figure includes parts from [57].

in depth in the final chapter. Before doing so, we include information that can aid in our approximation of $t$ by using the number of passenger mutations in a tumor as a molecular clock. For this, we create a new branching process model that takes advantage of the infinite-allele theory to allow passengers without creating a process with a large type space.
Multitype infinite-allele branching processes

3.1 Introduction to infinite-allele branching processes

The aim of this chapter is to introduce extensions of infinite-allele branching process models previously studied by Griffiths and Pakes [6][7] and Taib [8]. These works only consider single-type branching processes, so neutral evolution can only be considered. We wish to determine how the processes work when we allow selection by introducing a multitype branching process, resulting in mutations that lead to drift or selection depending on their type. We do so for three different types of branching processes. First we consider the most basic discrete-time Bienaymè-Galton-Watson process. Then we allow for continuous time Markov process which imposes a restriction on the distribution individual lifetimes to exponentially distributed lifetimes. Finally, we relax all assumptions and create a general infinite-allele branching process to show asymptotic results still hold. As we relax our assumptions on individual lifetimes, we also do so on the probability of a progeny being a new allele. Our first case we assume this probability is the same regardless of type, but then we allow for
type-dependent probabilities. The multitype infinite-allele Galton-Watson process material comes from a paper published in the Journal of Applied Probability [58]. The general branching process model and its applications come from a manuscript currently under review [59]. These two papers and the model and applications make up the central focus of this dissertation.

**Single type infinite-alleles branching processes**

We begin with a review of the models previously created in the single-type case. These papers go much more in depth with results than we are concerned, so we only introduce the processes and state the results we extended.

Suppose we have a simple Bienaymè-Galton-Watson process with \( Z_0 = i \) ancestors, \( \{Z_n; n = 0, 1, 2, \ldots \} \). We assume that upon splitting, each offspring has probability \( \mu \) of having a previously unseen mutation. The assumption extends from the infinite-type Wright-Fisher model mentioned in Ewens [60]. In these cases we say a mutation occurs when the offspring is different from its parent, without worrying about the details of how it is different. For now, we assume the same rules of a normal BGW process. That is, the offspring distribution is \( f(s) = E[s^{Z_1}] \) with mean \( m = f'(1-) \) and the added rule for the mutation. Any new offspring that are not mutated have the ancestor allele, so we can define the offspring p.g.f. for the ancestor allele as \( h(s) = f(\mu + (1 - \mu)s) \). Note that the ancestor process is a branching process on its own. Thus it has its own criticality based on the mean number of same-allele offspring, \( M = (1 - \mu)m \). The same allele offspring p.g.f. can be used to count the number of different alleles living in the population by counting those individuals which are not of the same type of their parent. This approach is used in the multitype setting we create, so we introduce it here to aid in our explanation later. First, define the indicator \( I_{j,r,n-r} = I(\text{the } j^{th} \text{ individual in generation } r \text{ has at least one offspring} \)
alive with the same allele in generation \( n \). If we have \( i \) ancestors, then \( I_{j,0,n} \) is one if any ancestor has a same-allele descendent alive in generation \( n \). Doing this allows us to count the total number of labels currently existing based on whether a new allele in generation \( r \) persists for \( n - r \) more generations. Then \( \{ K_n, n = 0,1,\ldots \} \) is the process counting the total number of labels existing at generation \( n \) and can be written

\[
K_n = I_{1,0,n} + \sum_{r=1}^{n} \sum_{j=1}^{Z_r} I_{j,r,n-r}.
\]

The expectation of this uses the ancestor p.g.f. and is written

\[
E[K_n] = 1 - h_n(0) + u \sum_{r=0}^{n-1} m^{n-r}(1 - h_r(0)).
\]

As is typical in branching processes, limit theorems are established for the growth of normalized versions of these processes, which are usually martingales. These are stated to compare to our versions we found earlier.

**Lemma 2.** Suppose a process has a single ancestor. Then,

\[
\frac{E[K_n]}{m^n} \rightarrow A = u \sum_{n=0}^{\infty} m^{-n}(1 - h_n(0)) \text{ as } n \rightarrow \infty.
\]

Lemma 2 is used with convergence theory in branching processes (given in the introduction) to establish the a.s. convergence of \( \frac{K_n}{Z_n} \).

**Theorem 4.** Assuming \( Z_n \) is a supercritical process with a single ancestor and \( E[Z_n \log Z_n] < \infty \), then

\[
\frac{K_n}{Z_n} \rightarrow A \text{ almost surely as } n \rightarrow \infty
\]

on the set of nonextinction, \( \{ Z_n \rightarrow \infty \} \).
These results establish the growth rates of the allele process. Essentially, we see similar exponential growth as the number of individuals, but proportional to them according to the rate of mutations as well as the probability of nonextinction of the ancestor processes as given by $A$. With this knowledge, we could make large number approximations of $\mu$, or knowing $\mu$ we could determine other properties of the population given samples for $Z_n$ and $K_n$. These two results are extended in the multitype discrete time version along with a few other results [58].

Now we consider the continuous time version. Suppose a continuous time Markov branching process, $\{Z_t, t \geq 0\}$ with offspring p.g.f. $f(s)$, mean offspring $m = f'(1-)$ and the same probability of a mutation to a new neutral allele, $\mu$. Also, suppose the lifetime of an individual is exponentially distributed with rate $a$ and let $\lambda = a(m-1)$ be the Malthusian parameter associated with the process. Pakes studies the growth of the expected number of alleles in this case, showing similar growth properties as the continuous time process. The continuous-time version presents a little more difficult challenge in counting generations. In the BGW process, generations are synchronized, so counting the number of alleles alive at any time requires going through all generations and counting all individuals born with a new label. In the continuous time version, we can do the same, but we need to know the splitting times and number of offspring at those times, essentially considering the embedded discrete time process by standardizing the individual lifetimes to unit length. Let $K_t$ count the number of alleles represented at time $t$. We define the split times $T_1, T_2, \ldots$ for the process $Z_t$ and $N_t$ is the number of splits up to time $t$. Also, $U_n$ is the number of offspring produced at $T_n$. These random variables can be used to split the population into discrete events where we can count all individuals only once. We define the indicators $I_{n,j}(t)$ to be one if the $j^{th}$ individual born at $T_n$ is a new allele and has same-type offspring living at time $t$. For consistency, define $T_0 = 0$. Then we can formulate the
expression for the number of living alleles

\[ K_t = I_{0,1}(t) + \sum_{n=1}^{N_t} \sum_{k=1}^{U_n} I_{n,j}(t). \]

Essentially, the process counts offspring by going through all splits rather than generations. Finding the distribution for splitting times can be a problem, but we know the distribution of the next splitting time given the current number of individuals, \( (T_n - T_{n-1} | Z_T) \), is exponentially distributed with rate \( aZ_{T_{n-1}} \). The expectation of \( K(t) \) is

\[ E[K_t] = 1 - F(0; t) + amue^{\lambda t} \int_0^t e^{-\lambda u} (1 - F(0; u)) du, \]

and as \( t \to \infty \), the expected number of alleles converges as given in Lemma 3.

**Lemma 3.**

\[ \frac{K_t}{e^{-\lambda t}} \to amue \int_0^{\infty} e^{-\lambda u} (1 - F(0; u)) du \text{ as } t \to \infty \]

The frequency spectrum of the process is defined as the expectation of \( \alpha_j(t) \), the number of labels represented by \( j \) individuals alive at \( t \). Similar theorems hold for the frequency spectrum in the discrete and continuous cases, but the results are unsurprising, and simple extensions of the identity,

\[ K_t = \sum_{j \geq 1} \alpha_j(t). \]

One problematic area in Pakes’ work is the lack of almost sure versions of these results in the continuous case. The issue comes from the distributions of the splitting times and the difficulty associated with these connections. However, general branching processes can be used to establish those results with a few modifications. Since the extension for general processes counted by random characteristics to our actual model is simple, we leave the theory of general processes for later and introduce it in
terms of our specific model.

**Birth-death infinite allele processes**

The previous papers work for general branching processes and give asymptotic results as time gets large. These results are not explicit due to different possible offspring distributions and the inability to find distributions at specific times. However, some specific offspring distribution in explicit forms for the parameters in Lemma 3. The birth-death process is investigated in [61] and the limiting frequency spectrum is found for \( M \geq 1 \), or when the ancestor process is supercritical. The birth-death process has birth and death rates \( \alpha \) and \( \beta \) with the same mutation probability \( \mu \). We show a similar result for \( m = 1 \) which implies \( M < 1 \). In this situation, we have a specific branching process that can exist, when the probability of a split equals that of a death. Consider a binary splitting Markov process with \( f(s) = ps^2 + (1-p) \) where \( p = \frac{1}{2} \) to guarantee the process is critical. The continuous-time Markov process with this generating function is the same as a birth-death-mutation process. The offspring pgf for like-type individuals is

\[
h(s) = f(\mu + (1-\mu)s) = p(\mu + (1-\mu)s)^2 + (1-p).
\]

Note that the offspring mean of this process is \( M = 1 - \mu \). We define the p.g.f. for the ancestor process \( H(s; t) = E[s^{Z_t}] = \sum_{j \geq 0} q_j(t)s^j \). Also, define \( \phi_t(j) = E[\alpha_j(t)] \) as the frequency spectrum, or expected number of labels represented by \( j \) individuals at \( t \). The limiting mean frequency spectrum is

\[
\Phi_j = \lim_{t \to \infty} \Phi_j(t) = \frac{am \mu G_j(\lambda)}{am \mu \int_0^\infty e^{-\lambda x}(1-q_0(x))dx} = \frac{G_j(\lambda)}{\int_0^\infty e^{-\lambda x}(1-q_0(x))dx}
\]
where $G_j(\lambda) = \int_0^\infty e^{-\lambda t} q_j(t) dt$ and $\Phi_j(t) = \frac{\phi_i(j)}{E[K_t]}$. We can think of the limiting mean frequency spectrum as a long run distribution for the number of labels having a certain number.

Let $A^2 = (1-p) + p\mu^2 = \frac{1}{2} + \frac{1}{2}\mu^2$ and $B^2 = p(1-\mu)^2 = \frac{1}{2}(1-\mu)^2$. Then we can write the backward Kolmogorov equation for $H(s,t)$

$$\frac{\delta F}{\delta t} = a(A^2 + B^2) \left[ \frac{1}{A^2 + B^2} (A^2 + B^2 F^2) - F \right].$$

The solution to the process is given in Athreya and Ney [2] as

$$H(s, t) = \frac{A^2(1-s) - (A^2 - B^2 s) e^{-ct}}{B^2(1-s) - (A^2 - B^2 s) e^{-ct}},$$

where $c = a(B^2 - A^2) = -a\mu < 0$. The pgf can be rewritten as the form of the mixture of two geometric pgfs

$$H(s, t) = w_1 \frac{\tilde{p}}{1-s(1-\tilde{p})} + w_2 \frac{s\tilde{p}}{1-s(1-\tilde{p})}.$$

The resulting values for $q_j(t)$ are

$$q_0(t) = \frac{A^2(1-e^{-ct})}{B^2 - A^2 e^{-ct}}, \text{ and}$$

$$q_j(t) = \frac{(B^2 - A^2)^2 [B^2(1-e^{-ct})]^{j-1} e^{-ct}}{(B^2 - A^2 e^{-ct})^{j+1}}, \quad j \geq 1.$$

**Theorem 5.** For a critical single-type infinite allele birth-death process,

$$\Phi_j = \frac{\mu}{B^2 j(A^2 \frac{\mu}{B^2})^j(1 - A^2 \frac{\mu}{B^2}) \log \left( \frac{\frac{\mu^2}{B^2} - 1}{\frac{A^2}{B^2}} \right)}.$$

**Proof.** We first derive $G_j = \int_0^\infty e^{-\lambda t} q_j(t) dt$, or the Laplace transform of $q_j(t)$. At
\[ \lambda = 0, \text{ we note that } (B^2 - A^2) = -\mu < 0 \text{ since } \mu \in [0, 1]. \text{ Also } A^2/B^2 \geq 1. \text{ Further, } A^2/B^2 = 1 \text{ at } \mu = 0 \text{ and } A^2/B^2 \to \infty \text{ as } \mu \to 1 (\text{important later}). \text{ Instead of solving for } G_0(\lambda), \text{ we solve for the denominator of } \Phi_j, \int_0^\infty e^{-\lambda t}(1-q_0(t))dt \text{ at } \lambda = 0 \text{ which is simply } \int_0^\infty 1-q_0(t)dt: \]

\[
\int_0^\infty 1-q_0(t)dt = \int_0^\infty 1 - \frac{A^2(1-e^{-ct})}{B^2-A^2e^{-ct}}dt \\
= \int_0^\infty \frac{B^2-A^2}{B^2-A^2e^{-ct}}dt \\
= \frac{B^2-A^2}{B^2} \int_0^\infty \frac{1}{1 - \frac{A^2}{B^2}e^{a\mu t}}dt
\]

(\text{Let } v = e^{a\mu t} \text{ so } \frac{dt}{dv} = (a\mu v)^{-1})

\[
= \frac{B^2-A^2}{a\mu B^2} \int_1^\infty [(1 - \frac{A^2}{B^2}v)v]^{-1}dv \\
= \frac{B^2-A^2}{a\mu B^2} \int_1^\infty \frac{1}{v} + \frac{\frac{A^2}{B^2}}{1 - \frac{A^2}{B^2}v}dv \\
= \frac{B^2-A^2}{a\mu B^2} \log \left( \frac{v}{1 - \frac{A^2}{B^2}v} \right) \bigg|_1^\infty \\
= \frac{B^2-A^2}{a\mu B^2} \cdot \lim_{t \to \infty} \log \left( \frac{v(1 - \frac{A^2}{B^2})}{1 - \frac{A^2}{B^2}v} \right) \\
= \frac{B^2-A^2}{a\mu B^2} \log \left( \frac{\frac{A^2}{B^2} - 1}{\frac{A^2}{B^2}} \right)
\]

since \( \lim_{t \to \infty} \log \left( \frac{v(1-k)}{1-kv} \right) = \log \left( \frac{k-1}{k} \right) \). The function is defined for \( \mu \in (0, 1) \) and negative. Note in the final solution, \( B^2 - A^2 < 0 \), so the solution is positive, as expected.
\[ \frac{G_j(0)}{10} = \int_{0}^{\infty} q_j(t)dt = \int_{0}^{\infty} \frac{(B^2 - A^2)^2[B^2(1 - e^{-c})]^j}{(B^2 - A^2 c e^{-c})^{j+1}} e^{-c} dt \]

\[ = (B^2 - A^2)^2 \int_{0}^{\infty} \frac{(B^2)^{j-1}(1 - e^{-c})^{j-1}}{(B^2)^{j+1}(1 - e^{-c})^{j+1}} e^{-c} dt \]

\[ = (B^2 - A^2)^2 \int_{0}^{\infty} \frac{(1 - e^{-c})^{j-1}}{(1 - e^{-c})^{j+1}} e^{-c} dt \]

\[ = (B^2 - A^2)^2 \int_{0}^{\infty} \frac{(1 - e^{\alpha \mu})^{j-1}}{(1 - e^{\alpha \mu})^{j+1}} e^{\alpha \mu} dt \]

(\text{Let } v = e^{\alpha \mu}, \text{ so } \frac{dt}{dv} = (a \mu v)^{-1})

\[ = \frac{(B^2 - A^2)^2}{a \mu B^4} \int_{1}^{\infty} -v^{-1} \frac{(1 - v)^{j-1}}{(1 - e^{\alpha \mu v})^{j+1}} vdvdv \]

\[ = \frac{(B^2 - A^2)^2}{a \mu B^4} \int_{1}^{\infty} (1 - v)^{j-1}(1 - e^{\alpha \mu v})^{-j-1}. \]

Let \( k = \frac{A^2}{B^2} \) and proceed by integrating by parts with \( x = (1 - v)^{j-1} \) and \( dy = (1 - kv)^{-j-1} dv \). Then we can solve for the integral.

\[ \int_{0}^{1} (1 - v)^{j-1}(1 - kv)^{-j-1} = \]

\[ = \frac{(1 - v)^{j-1}(1 - kv)^{-j}}{k} \bigg|_{0}^{\infty} + \frac{j - 1}{k} \int_{0}^{1} (1 - v)^{j-2}(1 - kv)^{-j} dv \]

\[ = 0 + \frac{j - 1}{k} \int_{0}^{1} (1 - v)^{j-2}(1 - kv)^{-j} dv \]

Continue integrating by parts \( j \)-2 more times the same way

\[ = \cdots = 0 + \frac{1}{jk^{j}(1 - k)}. \]

Note each time we integrate by parts in the form \( \int_{1}^{\infty} x(t)y(t)dt = x(t)y(t)|_{1}^{\infty} - \int_{1}^{\infty} y(t)x'(t)dt \), the term \( x(t)y(t) \) has the functional form \( (1 - v)^{j-i}(1 - kv)^{-j-(i-1)} \) for
the $i^{th}$ iteration. This has a value of 0 at $v = 1$ and a limit of 0 as $v \to \infty$. Thus

$$G_j(0) = \frac{(B^2 - A^2)^2}{cB^4} \left( \frac{1}{j\left(\frac{A^2}{B^2}\right)^j(1 - \frac{A^2}{B^2})} \right) = \frac{B^2 - A^2}{aB^4\left[j\left(\frac{A^2}{B^2}\right)^j(1 - \frac{A^2}{B^2})\right]}.$$ 

We can plug back in for $\Phi_j$, to get

$$\Phi_j = \frac{B^2 - A^2}{aB^4\left[j\left(\frac{A^2}{B^2}\right)^j(1 - \frac{A^2}{B^2})\right]} \frac{B^2 - A^2}{a\mu B^4 \log \left(\frac{A^2}{B^2} - 1\right)}$$

$$= \frac{\mu}{B^2 j\left(\frac{A^2}{B^2}\right)^j(1 - \frac{A^2}{B^2}) \log \left(\frac{A^2}{B^2} - 1\right)}.$$

\[\Box\]

The limiting variance frequency spectrum is defined as

$$\xi_j = \lim_{t \to \infty} \eta_{i,t}(j)$$

where $\eta_{i,t}(j) = \text{Var}_i[a_j(t)]$ is the variance frequency spectrum [61]. The variance frequency spectrum for the above case is

$$\eta_j(t) = q_j(t)[1 - q_j(t)] +$$

$$+ m^2 \mu^2 \left[ C(t) + (\lambda + a)e^\lambda \int_0^t e^{-\lambda x} C(x)dx \right] +$$

$$+ am \mu e^\lambda \int_0^t e^{-\lambda x} q_j(x)dx + ia(\sigma^2 - m)\mu^2 e^\lambda \int_0^t e^{-\lambda x} q_j^2(x)dx$$
with

\[ C(t) = a \int_0^t [q_j(x) + (\sigma^2 + m^2)\beta_1^2(x)]e^{-a(t-x)}dx \]

\[ - \left[ a \int_0^t q_j(x)e^{-a(t-x)}dx \right]^2 - \left[ am \int_0^t \beta_1(x)e^{-a(t-x)}dx \right]^2 \]

and \( \beta_1(x) = ae^{\lambda x} \int_0^x e^{-\lambda u}q_j(u)du \). We can reduce some of these pieces to simplify the equation. Solving for \( \int_0^\infty q_j(x)dx \) gives us a closed-form for \( \beta_1(x) \).

\[
\int_0^t q_j(x)dx = \frac{(B^2 - A^2)^2}{B^4} \int_0^t \frac{(1 - e^{\alpha \mu x})^{j-1}e^{\alpha \mu x}}{(1 - \frac{A^2}{B^2}e^{\alpha \mu x})^j + 1}dx
\]

\[
= \frac{(B^2 - A^2)^2}{B^4a\mu} \int_0^\infty e^{\alpha \mu t} \frac{(1 - v)^{j-1}(1 - \frac{A^2}{B^2}v)^{-(j+1)}}{j(k-1)(1 - \frac{A^2}{B^2}e^{\alpha \mu t})^j} dv
\]

using the same substitution and parts method as above. Then,

\[ \beta_1(x) = \left( \frac{(B^2 - A^2)^2}{B^4a\mu} \right) \frac{(1 - e^{\alpha \mu x})^j}{j(k-1)(1 - \frac{A^2}{B^2}e^{\alpha \mu x})^j}. \]

This can be plugged in to \( C(t) \) above to get a rather large expression for the variance frequency spectrum.

### 3.2 The multitype infinite-allele Galton Watson process

#### Introduction

We begin with an introduction of the multitype version of the infinite-allele Bienaymé-Galton-Watson process introduced by Griffiths and Pakes [6]. We extend the
process from a single type of individuals to a $k$-type process to allow differing growth parameters between types. We define $\{Z(n), n = 0, 1, 2, \ldots \}$ as the $k$-dimensional stochastic process where $Z(n) = (Z_1(n), \ldots, Z_k(n))$ is the number of particles of each type in the $n^{th}$ generation, $G_n$. The process $\{Z(n)\}$ is a typical multitype BGW process where each type represents a different set of the so-called driver mutations which may confer a growth advantage to the mutant cell clone [62]. Probabilities of driver mutations occurring are then represented in the offspring probability generating function (p.g.f.) of $\{Z(1)\}$, $f(s) = (f_1(s), \ldots, f_k(s))$. That is, given a type $i$ individual, the offspring p.g.f. is

$$f_i(s) = \sum_{j_1, \ldots, j_k \geq 0} p_i(j_1, \ldots, j_k) s_1^{j_1} \ldots s_k^{j_k}, \quad s_i \in [0, 1]$$

where $p_i(j_1, \ldots, j_k) = P[Z(1) = (j_1, \ldots, j_k) | Z(0) = e_i]$. The transition probabilities of splitting into different types in the p.g.f. describe driver mutation probabilities. We denote the mean offspring matrix by $M$. Let $\rho$ be the spectral radius of $M$ with left and right eigenvectors $v$ and $u$. Results concerning differences in driver mutations, criticality, and asymptotics of $Z(n)$ follow the standard theory of multitype processes and can be found in Athreya and Ney [2] and Mode [63].

Within this multitype process, we incorporate the possibility for newly born offspring to have a passenger mutation with probability $\mu \in (0, 1)$ regardless of type. This extends the infinite-allele idea introduced by Griffiths and Pakes, where each individual branching process initiates an infinite-allele branching process. Since we can have a large number of passenger (selectively neutral) mutations that only affect heterogeneity, we do not distinguish between alleles that have a different set of passenger mutations in $Z(n)$ or its p.g.f. $f^{(n)}(s)$. Instead we only track their count. To avoid confusion, we use the term type to distinguish individuals that differ with respect to driver mutations and have a particular type with respect to the branching
process \( \{Z(n)\} \). That is \( Z_i(n) \) and \( Z_j(n) \) count different types within the population. When an individual has an offspring that undergoes a passenger mutation, we use the same terminology as Taib [8] and say the offspring has a different *label*. Every passenger mutation event leads to a new and unique label. Thus individuals can be distinguished by their *type* and *label*, but only the type influences growth rates.

We are interested in a particular quantity, \( \mathbf{K}(n) = (K_1(n), \ldots, K_k(n)) \), the \( k \)-dimensional vector with \( K_i(n) \) equal to the number of labels carried by \( i \)-type individuals in \( G_n \). \( \{\mathbf{K}(n)\} \) is a stochastic process with \( \mathbf{K}(0) = e_i \) representing the only label present of the \( i \)-type ancestor counted by \( Z(0) \). We will also make use of the branching process \( \{\tilde{Z}(n)\} \) which we call the ancestor process of \( \{Z(n)\} \). \( \tilde{Z}(n) \) counts the number of individuals in generation \( n \) that have the same label as the ancestor, or never undergo a passenger mutation. We define the p.g.f. for the ancestor offspring process of an \( i \)-type individual

\[
H_i(s) = E[s^{\tilde{Z}(1)}|\tilde{Z}(0) = e_i] = f_i(\mu + (1 - \mu)s_1, \ldots, \mu + (1 - \mu)s_k).
\]

The \( k \)-dimensional vector \( \mathbf{H}(s) \) is the offspring p.g.f. for the ancestor process. In both the normal individual process and the ancestor process the p.g.f. in the \( n \)th generation is the \( n \)th iterate of the p.g.f., denoted \( f^{(n)}(s) \) and \( \mathbf{H}^{(n)}(s) \) respectively. We also denote the mean matrix of the ancestor process \( \tilde{M} = (1 - \mu)M \).

We count the number of labels for a particular type by counting individuals with specific characteristics. Define the indicator \( I_{m,i,n} = 1 \) if the \( m \)-th \( i \)-type individual in \( G_n \) has a new label different from its parent. Also, define the indicator \( J_{m,i,r,n-r}(j) = 1 \) if some \( j \)-type individual in \( G_n \) has a label initiated by the \( m \)-th \( i \)-type individual in \( G_r \). It follows then that

\[
J_{m,i,0,n}(j) = I(\tilde{Z}_j(n) > 0|\tilde{Z}(0) = e_i)
\]
and further

\[ E[J_{m,i,0,n}(j)] = P(\tilde{Z}_j(n) > 0|\tilde{Z}(0) = e_i) = 1 - H_i^{(n)}(1 - e_j). \]

If \( \mathcal{F}_n \) is the natural filtration with respect to \( \{Z(n)\} \), then

\[ E[I_{m,i,r}J_{m,i,r,n-r}(j)|\mathcal{F}_r] = \mu(1 - H_i^{(n-r)}(1 - e_j)). \]

We can express the number of labels of a certain type in terms of the number of individuals in the population that are ancestors to new labels and not yet extinct. Given a type \( \alpha \) ancestor,

\[ K_j(n) = J_{1,\alpha,0,n}(j) + \sum_{r=1}^{n} \sum_{k=1}^{k} \sum_{m=1}^{m} I_{m,i,n}J_{m,i,r,n-r}(j). \]

The expectation given a type \( \alpha \) ancestor is then

\[
E_\alpha[K_j(n)] = E_\alpha[J_{1,\alpha,0,n}(j)] + \sum_{r=1}^{n} \sum_{i=1}^{i} E_\alpha[\sum_{m=1}^{m} I_{m,i,n}J_{m,i,r,n-r}(j)]
\]

\[
= 1 - H_\alpha^{(n)}(1 - e_j) + \mu \sum_{r=0}^{n-1} e^{r} M^{n-r}(1 - \mathbf{H}^{(r)}(1 - e_j)) \]

(3.1)

which we simplify using conditional expectation and rewriting the indices in the sum.

To aid in our development of the process, we give a simple example in Figure 3.1. We begin with a single type 1 ancestor. As the process unfolds, we see changes to new types represented by different colors. When a new label occurs, we start a new process as given by a new shaded area where that individual becomes the ancestor to the process. Thus, to count the number of labels of a specific type, we count the number of gray areas that contain at least 1 cell of the type in question at the generation in question. For example, in generation 6, we get \( K(n) = (1,2,3,3) \).
Figure 3.1: An example of the multitype infinite-allele Galton Watson Process shows how we can count types (colors) and neutral alleles (gray trees).
Irreducible $M$

In this section we suppose the mean matrix, $M$, is irreducible. Also suppose $\rho$ is the spectral radius of $M$ with left and right eigenvectors $v$ and $u$ normalized so that $vu = 1$ and $1^\top u = 1$. The eigenvector $v$ is a row vector, and $u$ is a column vector. Let us define the constant

$$A_j = \mu \sum_{r=0}^{\infty} v \rho^{-r} \left[ 1 - H^{(r)}(1 - e_j) \right].$$

Note that this constant is finite regardless of criticality of $\rho$. This yields the following lemma about the limit of the expectation as given in Equation (3.1).

Lemma 4. Given an irreducible process starts with an ancestor of type $\alpha$,

$$\lim_{n \to \infty} \frac{E_\alpha[K_j(n)]}{E_\alpha[Z(n)u]} = A_j$$

Proof. First note that $E_\alpha[Z(n)u] = e_\alpha^\top M^n u = e_\alpha^\top \rho^n u$.

If $\rho > 1$, then $(1 - H_\alpha^{(\alpha)}(1 - e_j))\rho^{-n} \to 0$. If $\rho \leq 1$, then $(1 - H_\alpha^{(\alpha)}(1 - e_j)) \to 0$ since extinction occurs almost surely.

The second term of the sum from (3.1) can be rewritten as

$$\frac{\mu}{e_\alpha^\top \rho^n u} \sum_{r=0}^{n-1} e_\alpha^\top M^{n-r}(1 - H^{(r)}(1 - e_j)) = \frac{\mu e_\alpha^\top M^n}{e_\alpha^\top \rho^n u} \sum_{r=0}^{n-1} M^{-r}(1 - H^{(r)}(1 - e_j))$$

For irreducible $M$, $\lim_{n \to \infty} (M \rho^{-1})^n = uv$. Also,

$$1 - H_\alpha^{(\alpha)}(1 - e_j) \leq E_\alpha[\tilde{Z}_j(n)] = e_\alpha^\top \tilde{M}^n e_j$$
by Markov’s inequality. Since \( \tilde{M} = (1 - \mu)M \), then \( M^{-r}\tilde{M}^r = (1 - \mu)^r I \), so

\[
\sum_{r=0}^{n-1} M^{-r}(1 - H^{(r)}(1 - e_j)) \leq \sum_{r=0}^{n-1} M^{-r}\tilde{M}^r e_j \leq \sum_{r=0}^{n-1} (1 - \mu)^r e_j \rightarrow \frac{1}{\mu} e_j \text{ as } n \rightarrow \infty.
\]

(3.2)

This series converges absolutely, leading to

\[
\lim_{n \rightarrow \infty} \frac{E_{\alpha} [K_j(n)]}{e^T_{\alpha} \rho^u u} = \frac{\mu e^T_{\alpha} \rho^u u}{e^T_{\alpha} \rho^u u} \sum_{r=0}^{\infty} M^{-r}(1 - H^{(r)}(1 - e_j))

= \mu \sum_{r=0}^{\infty} \rho M^{-r}(1 - H^{(r)}(1 - e_j))

= \mu \sum_{r=0}^{\infty} \rho \rho^{-r}(1 - H^{(r)}(1 - e_j))
\]

giving the desired result for all \( \rho > 0 \).

\[\square\]

If we define \( \Omega_a = \{ Z(n) > 0, n = 0, 1, 2, \ldots \} \) as the set of nonextinction, then we can show the limiting behavior of \( K_j(n) \) for supercritical processes (\( \rho > 1 \)) converges almost surely to \( A_j \) conditionally on nonextinction.

**Theorem 6.** If \( E_i[Z_j(1) \log Z_j(1)] < \infty \) for all \( 1 \leq i \leq k \) and \( 1 \leq j \leq k \) and \( 1 < \rho < \infty \), and if the process is started by a \( \alpha \)-type ancestor, then

\[
\lim_{n \rightarrow \infty} \frac{K_j(n)}{Z(n)u} = A_j \quad \text{a.s. on } \Omega_a.
\]

**Proof.** Define the variable

\[
K_j(r, n-r) = \sum_{i=1}^{k} \sum_{m=1}^{I} I_{m,i,r} \alpha_{i,r}(j)
\]
which is the number of new $j$-type labels in $G_r$ that are still represented in $G_n$. Note that $K_j(n) = \sum_{r=0}^{n} K_j(r, n - r)$. Then for any fixed $n'$,

$$\left(Z(n) u\right)^{-1} \sum_{r=1}^{n'-1} K_j(r, n - r) \leq \left(Z(n) u\right)^{-1} \sum_{r=1}^{n'-1} Z_j(r) \rightarrow 0 \text{ a.s. on } \Omega_a.$$  

We can represent $K_j(n)$ as a sum and show each of the summands converge. First we show that

$$\sum_{r=0}^{n} \frac{K_j(r, n - r)}{Z(n) u} = \sum_{r=0}^{n-n'} \frac{K_j(n-r, r)}{Z(n) u} = \sum_{r=0}^{n-n'} \frac{K_j(n-r, r)}{Z(n-r) u} \left(\frac{Z(n-r) u}{Z(n) u}\right).$$  

(3.4)

Now, conditioning on $Z(n-r)$, $E[K_j(n-r, r)|Z(n-r)] = \mu \left(Z(n-r)[1 - H^{(n-r)}(1 - e_j)]\right)$ is a sum of i.i.d. random variables. Noting that $Z(n)/(Z(n) u) \rightarrow v$ a.s. on $\Omega_a$ as given in Athreya and Ney [2, Theorems 1 and 4 on p. 193] and proved by Kurtz et al. [64], we can use a strong law for random sums to get almost sure convergence $K_i(n-r, r)/(Z(n) u) \rightarrow \mu v[1 - H^{(n-r)}(1 - e_i)]$ on $\Omega_a$ as $n \rightarrow \infty$ and $r = O(1)$.

Hoppe shows in Theorem 2.1 that for $1 < \rho < \infty$, there exists a sequence of positive vectors, $\{c_n\}$ and scalars $\{\gamma_n\} = \{v c_n\}$ such that for each $\alpha$, if $Z(0) = e_\alpha$ then

$$\lim_{n \rightarrow \infty} Z(n) c_n = W_\alpha \text{ a.s.},$$  

(3.5)

$$\lim_{n \rightarrow \infty} \gamma_n/\gamma_{n+1} = \rho,$$  

(3.6)

$$\lim_{n \rightarrow \infty} c_n/\gamma_n = u,$$  

(3.7)

and

$$\lim_{n \rightarrow \infty} \gamma_n Z(n) = W_\alpha v \text{ a.s.}$$  

(3.8)

where $W = (W_1, \ldots, W_K)$ is a nonnegative random variable if $E[Z_i(1) \log Z_i(1)] < \infty$ for $i = 1, \ldots, k$ [65]. Now, since $\gamma_n Z(n) u \rightarrow W_\alpha$ a.s. by (3.8), we make use of (3.5)-
(3.7) to get
\[
\frac{Z(n-1)u}{Z(n)u} \sim \frac{W_n \gamma_n}{W_n \gamma_n^{-1}} \to \rho^{-1},
\]
implying
\[
\frac{Z(n-r)u}{Z(n)u} \to \rho^{-r} \text{ a.s.}
\]
as \(n \to \infty\). Since \(K_{r(n-r)} \leq \frac{Z(n-r)}{Z(n-r)u} \), each summand of the right hand side of (3.4) is dominated by \(\rho^{-r}\). Thus the series is dominated by the geometric series that converges to \((1 - \rho)^{-1}\), so we can use the dominated convergence theorem to show the series on the right hand side of (3.4) converges to \(A_j\) almost surely on \(\Omega_a\) as \(n \to \infty\). The assertion follows.

\[\square\]

**Reducible M**

We now remove the assumption of irreducibility from the process to model more realistic scenarios. Mutations usually have low probability of being reversed, so if we assume the probability is 0 (as often is the case), we want to understand how the number of labels grows within each type. We are able to group the types of the branching process into equivalence classes that form irreducible subprocesses. This allows a reordering of matrix \(M\) as a block lower triangular matrix with blocks along the diagonal being irreducible. Results about the process can then be ascertained based on results for the blocks.

Define the equivalence classes \(\{C_a\}_{a=1,\ldots,t}\) as \(C_a = \{i, j \in 1, 2, \ldots, k; m_{i,j}^{(n_1)} > 0\text{ and } m_{j,i}^{(n_2)} > 0 \text{ for some } n_1 \text{ and } n_2\}\). That is, the equivalence classes are created by separating types into groups where each type in the group communicates with every other type in the same group \([66]\). We are able to order the indices of types and permute the mean matrix \(M\) according to the equivalence classes by imposing an
order on the classes such that if \( b > a \), then \( m_{i,j}^{(n)} > 0 \) for all \( i \in C_b \) and \( j \in C_a \). The resulting mean matrix after permutation is

\[
M = \begin{bmatrix}
M_{1,1} & 0 & 0 & \ldots & 0 \\
M_{2,1} & M_{2,2} & 0 & \ldots & 0 \\
M_{3,1} & M_{3,2} & M_{3,3} & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
M_{l,1} & M_{l,2} & M_{l,3} & \ldots & M_{l,l}
\end{bmatrix}
\]

with \( M_{a,b} = (m_{i,j})_i \in C_a, j \in C_b \). If we limit a process to types within a subclass, then \( Z_{C_a}(n) = \{Z_i(n), i \in C_a\} \) is an irreducible subprocess of \( Z(n) \) having mean offspring matrix \( M_{a,a} \). Given an ancestor with type \( \alpha \in C_a \), results about \( Z_{C_a}(n) \) and \( K_{C_a}(n) \) follow those of the previous section. We show here the results for a 2-class example and note that examples with more than 2 classes can be developed analogously.

Suppose we have a reducible BGW process \( Z(n) \) with mean matrix

\[
M = \begin{bmatrix}
M_{1,1} & 0 \\
M_{2,1} & M_{2,2}
\end{bmatrix}
\]

with irreducible \( M_{i,1} \) and \( M_{2,2} \), and \( M_{2,1} \neq 0 \). Let \( M_{i,i} \) have spectral radius \( \rho_i \) with associated left and right eigenvectors \( \mathbf{u}_i \) and \( \mathbf{v}_i \) for \( i = 1, 2 \). Let \( \rho \) be the spectral radius of \( M \), which is the maximum of \( \rho_1 \) and \( \rho_2 \). If \( \rho = \rho_1 > \rho_2 \), then the eigenvectors associated with \( \rho \) are \( \mathbf{v} = (\mathbf{v}_1, 0) \) and \( \mathbf{u} = (\mathbf{u}_1, (\rho I - M_{2,2})^{-1}M_{2,1}\mathbf{u}_1) \). If \( \rho = \rho_2 > \rho_1 \), then \( \mathbf{v} = (\mathbf{v}_2M_{2,1}(\rho I - M_{1,1})^{-1}, \mathbf{v}_2) \) and \( \mathbf{u} = (0, \mathbf{u}_2) \). These are used as eigenvectors when we show the limits hold for the reducible cases. We present an analogous lemma to Lemma 1 for the expectation of the number of labels, \( K_j(n) \) in a 2-class process with the mean matrix as above. We also limit ourselves to an ancestor of a type in
the class $C_2$ to avoid trivial results that can arise. Also, we expect ancestors of cell populations to have no somatic mutations, but have the ability to gain mutations. We do not expect mutant cells to give rise to daughter cells without those mutations since the reversing of mutations is very rare. Because of this, we require the reducibility assumption for $M$. Another case can occur when $\rho_1 = \rho_2$. Different convergence results exist in this situation which can be determined analogously based on Theorem 2.3 of Kesten and Stigum [66]. This situation is less likely to occur in cancer evolution, so it is not discussed here.

**Lemma 5.** Suppose $\rho$ is the spectral radius of $M$ and it is simple. Given a reducible BGW process with an ancestor of type $\alpha \in C_2$,

$$\lim_{n \to \infty} \frac{E_\alpha[K_j(n)]}{E_\alpha[Z(n)u]} = A_j$$

with $u$ and $v$ defined as above based on whether $\rho = \rho_1$ or $\rho = \rho_2$, and

$$A_j = \mu \sum_{r=0}^{\infty} v^r \left[ 1 - H^{(r)}(1 - e_j) \right].$$

**Proof.** Under the conditions above, $\lim(\frac{M}{\rho})^n = uv$ as $n \to \infty$ where $vu = 1$ [5]. The remaining calculations of Lemma 4 then hold with $u$ and $v$ determined by whether $\rho = \rho_1$ or $\rho = \rho_2$. \qed

The almost sure convergence of $K_j(n)$ is also extended to the reducible case. Suppose again we consider a $\alpha$-type ancestor with $\alpha \in C_2$. Let $\Omega_a$ still be defined as the set of nonextinction.

**Theorem 7.** Assume $E[Z_i(1) \log Z_i(1)|Z(0) = e_a] \leq \infty$ for $i = 1, 2, \ldots, k$. Also, let
$\rho$ be the spectral radius of $M$ and assume it is simple. Then

$$\lim_{n \to \infty} \frac{K_j(n)}{Z(n)u} = A_j \ a.s. \ on \ \Omega_a.$$  

with $u$ and $v$ defined as above.

We first introduce a corollary to Kesten and Stigum’s almost sure convergence of $Z(n)$ that we use in our proof.

**Corollary 1.** Suppose $\rho$ is the spectral radius of $M$, the mean matrix for $Z(n)$ with eigenvalues $u$ and $v$ as given above depending on whether $\rho = \rho_1$ or $\rho = \rho_2$. Then

$$\lim_{n \to \infty} \frac{Z(n)}{Z(n)u} = v \ a.s. \ on \ \Omega_a.$$  

**Proof.** Theorem (2.1) of Kesten and Stigum states that

$$\lim_{n \to \infty} \frac{Z(n)}{\rho^n} = w \cdot v$$

almost surely where $w$ is a scalar random variable [66]. This implies that $\lim_{n \to \infty} \frac{\rho^n}{Z(n)u} = w^{-1}$ and further $\lim_{n \to \infty} \frac{Z(n)}{Z(n)u} = \lim_{n \to \infty} \frac{\rho^n}{\rho^n Z(n)u} = v$ almost surely on $\Omega_a$. \[Q.E.D.\]

**Proof of Theorem 7.** Since $\alpha \in C_2$, the proof is similar to that of Theorem 1 with $u$ and $v$ defined as above based on whether $\rho = \rho_1 > \rho_2$ or $\rho = \rho_2 > \rho_1$. We use the reducible versions of the Kesten and Stigum theorems [66, Theorem 1.1] in place of the irreducible case to show almost sure convergence of the number of individuals. The calculations from the proof of Theorem 6 above then hold with the modified $u$ and $v$. \[Q.E.D.\]
Frequency Spectrum

Let $\alpha_i(j, n)$ be the number of $i$-type labels in generation $n$ represented by $j$ individuals currently living in generation $n$. We will denote the expectation of this term as the frequency spectrum for type $i$. The term $\alpha_i(j, n)$ can be expressed as a sum of indicators via an approach similar to determining the total number of labels. Define the indicators,

$$I_{j,i}(n) = I(\text{the ancestor has } j \text{ } i\text{-type descendants with the same label in generation } n)$$

and

$$I_{j,i,l,k}(r, n - r) = I(\text{the } l\text{th } k\text{-type individual in generation } r \text{ is a new label and has } j \text{ } i\text{-type descendants with the same label in generation } n).$$

Given an ancestor of type $\alpha$, $E[I_{j,i}(n)] = P[\tilde{Z}_i(n) = j|\tilde{Z}_\alpha(0) = 1]$, which we denote $q_{\alpha,i}^{(n)}(j)$. The values of $q_{\alpha,i}^{(n)}(j)$ over $i$ and $j$ make up the coefficients to $H_\alpha^{(n)}(s)$. We can write $\alpha_i(j, n)$ in terms of the previous indicators,

$$\alpha_{j,i}(n) = I_{j,i}(n) + \sum_{r=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{Z_k(r)} I_{j,i,l,k}(r, n - r)$$

allowing us to derive an expression for the frequency spectrum, $\phi_i(j, n) \equiv E[\alpha_i(j, n)]$.

The frequency spectrum can be simplified to

$$\phi_i(j, n) = E[I_{j,i}(n)] + \sum_{r=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{Z_k(r)} I_{j,i,l,k}(r, n - r)$$

$$= q_{\alpha,i}^{(n)}(j) + \sum_{r=1}^{n} \sum_{k=1}^{K} \mu q_{k,i}^{(n-r)}(j)e'_\alpha M'e_k$$
using conditional expectation. Now denote the vector $q_i^{(r)}(j) = [q_{1,i}^{(r)}(j), \ldots, q_{K,i}^{(r)}(j)]$. Then,

$$
\phi_i(j, n) \equiv E[\alpha_i(j, n)] = q_{\alpha,i}^{(n)}(j) + \sum_{r=1}^{n} \mu e'_{\alpha} M^{r} q_i^{(n-r)}(j) = q_{\alpha,i}^{(n)}(j) + \mu \sum_{r=0}^{n-1} e'_{\alpha} M^{n-r} q_i^{(r)}(j). \quad (3.9)
$$

**Theorem 8.** Let $u$ and $v$ be the right and left eigenvectors associated with $\rho$. Then

$$
\lim_{n \to \infty} \frac{\phi_i(j, n)}{Z(n)u} = \sum_{r=0}^{\infty} \mu \rho^{-r} v q_i^{(r)}(j) \quad (3.10)
$$

and

$$
\Psi(i, j) = \lim_{n \to \infty} \frac{\phi_i(j, n)}{E[K_i(n)]} = \sum_{r=0}^{\infty} \mu \rho^{-r} v q_i^{(r)}(j) / A_i.
$$

**Proof.** The proof is essentially the same as that of Lemma 4. We use the fact that $q_{\alpha,i}^{(n)}(j) \to 0$ as $n \to \infty$ in (3.9), and the sum is bounded. □

In this case, $\Psi(i, j)$ is the long-run frequency of $i$-type labels having $j$ individuals. This provides an idea of the distribution of labels having different individuals.

**Proof of Concept Simulations**

We consider two different 4-type branching processes with similar p.g.f.s to illustrate the almost sure convergence results. Each process contains cells undergoing reproduction via binary fission or death, and the probability of a new allele at each generation
is $\mu = 5 \times 10^{-4}$. The first process is irreducible with p.g.f.

$$f_1(s) = 0.45 + 0.03s_1s_2 + 0.02s_1s_3 + 0.5s_1^2,$$
$$f_2(s) = 0.51 + 0.06s_1s_2 + 0.04s_2s_3 + 0.39s_2^2,$$
$$f_3(s) = 0.56 + 0.04s_2s_3 + 0.05s_3s_4 + 0.35s_3^2$$
and
$$f_4(s) = 0.5 + 0.03s_2s_4 + 0.06s_3s_4 + 0.4s_4^2.$$

The mean matrix is

$$M = \begin{bmatrix}
1.05 & 0.03 & 0.02 & 0 \\
0.06 & 0.88 & 0.04 & 0 \\
0 & 0.04 & 0.79 & 0.05 \\
0 & 0.03 & 0.07 & 0.9
\end{bmatrix}.$$  

The spectral radius of this process is $\rho = 1.0617$ with left and right eigenvectors
$$v = [1.3974, 0.2728, 0.1554, 0.0480]$$
and
$$u = [0.6637, 0.2291, 0.0451, 0.0620]^\top.$$  

Thus, the process is supercritical, and growth is expected. Numerically evaluating $A$ gives

$$A = [0.0035, 0.0016, 0.0012, 0.0005]^\top.$$  

We ran 100 simulations of the process beginning with a single type 1 ancestor, showing the sample paths for each type in Figure 3.2. We condition on nonextinction by removing simulations that become extinct and only using the remaining simulations. A horizontal red line is superimposed at the respective value of $A_i$ according to the type, $i$. Over 140 generations we see convergence of nearly all paths to the limit, $A_i$.

We created a similar 4-type process, but adjusted the probabilities so that types 3 and 4 form a class that can feed into types 1 and 2, but not in the other direction. The p.g.f. for the process is
100 simulations showing almost sure convergence of each type to $A_i$

Figure 3.2: The paths of the process $K(n)$ for 100 simulations shows convergence to $A_i$, $i = 1, 2, 3, 4$ which is given by the red line. We scale the paths with a log$_{10}$ transformation to see convergence over the 140 generations.

\[
\begin{align*}
  f_1(s) &= 0.47 + 0.03s_1s_2 + 0.5s_1^2, \\
  f_2(s) &= 0.51 + 0.1s_1s_2 + 0.39s_2^2, \\
  f_3(s) &= 0.54 + 0.02s_2s_3 + 0.09s_3s_4 + 0.35s_3^2 \text{ and} \\
  f_4(s) &= 0.4 + 0.08s_2s_4 + 0.07s_3s_4 + 0.45s_4^2.
\end{align*}
\]
The mean matrix is

$$M = \begin{bmatrix}
1.03 & 0.03 & 0 & 0 \\
0.1 & 0.88 & 0 & 0 \\
0 & 0.02 & 0.81 & 0.09 \\
0 & 0.08 & 0.07 & 1.05
\end{bmatrix}.$$ 

After breaking the matrices up into submatrices and determining the spectral radii of each, we find the process is supercritical with spectral radius $\rho = 1.0739$ since $\rho_2 > \rho_1$. The results of Kesten and Stigum lead to vectors $v = [1.9053, 0.8359, 0.3262, 1.2298]$ and $u = [0, 0, 0.2543, 0.7457]^T$. We numerically evaluate $A$ to get

$$A = [0.0055, 0.0039, 0.0017, 0.0032]^T.$$ 

The results of 100 simulations are given illustrating our theorem in Figure 3.3 with a red horizontal line at the value of $A_i$. Note that these simulations are run under the condition of nonextinction and are initiated with a single type 4 ancestor in generation 0. The figure shows similar results in the reducible case that holds by adjusting the eigenvectors according to Kesten and Stigum’s results.

Finally, we show the results of the convergence of the frequency spectrum, Theorem 8, in a 2-type simulation starting with a single type 1 ancestor. Again we set the probability of a new label to $\mu = 5 \times 10^{-4}$. Our p.g.f. for this process is:

$$f_1(s) = 0.5s_1 + 0.2s_1s_2 + 0.3s_1^2 \text{ and }$$

$$f_2(s) = 0.5s_2 + 0.4s_1s_2 + 0.1s_2^2.$$
100 simulations showing almost sure convergence of each type to $A_i$

Figure 3.3: The paths of the process $K(n)$ for 100 simulations in a reducible BGW shows convergence to $A_i$, $i = 1, 2, 3, 4$ which is given by the red line. We scale the paths with a $\log_{10}$ transformation to see convergence over the 140 generations.

The process is supercritical with mean matrix

$$M = \begin{bmatrix} 1.3 & 0.2 \\ 0.4 & 1.1 \end{bmatrix}$$

having spectral radius of $\rho = 1.5$ and eigenvectors $v = [4/3, 2/3]$ and $u = [1/2, 1/2]^T$.

The results for the convergence of the normalized frequency spectrum can be seen in Figure 3.4. We scaled the process using a $\log_{10}$ transformation to better illustrate the
convergence and difference in each curve. Each color represents the average number of labels represented by $j$ individuals after 100 simulations. In each case, the value converges to the numerical solution given on the right hand side of the theorem. The two plots refer to both types in the process. We also show the results in the 40th generation, after convergence has occurred, in Figure 3.5. The black curves show the results of simulations of the process represented by the left hand of Equation (3.10) in Theorem 8 and the red curve shows the results of calculating the right hand side to show convergence. In both cases, there is very little error after the 40 generations.

![Frequency Spectrum Converging for Different Values of $j$](image)

Figure 3.4: The simulation of the frequency spectrum for type-1 and type-2 individuals shows convergence to the analytical formula given by the horizontal lines for each number of individuals.

### Applications to Cancer Evolution

The current view of cancer progression is that the multistep accumulation of somatic mutations leads to the transformation of healthy cells into malignant cancer cells with
higher fitness \[49\]. In terms of multitype branching processes, this can be represented by varying the parameters and offspring probabilities for each type of cell, where higher fitness cells have a higher probability of splitting, ensuring supercriticality of the process and particular types. Because of the way cells proliferate, we are mostly concerned with binary fission, where each offspring may survive, die, or mutate even though our theory holds for general multitype processes. The transition from normal cells to supercritical cancer cells occurs over multiple replications, and waves of expansion are observed. While the probability of a mutation occurring in a single cell is small, years of replication in a large number of cells makes the likelihood of cancer initiation greater. Sequencing studies show that genomes undergo a large number of changes, but most mutations are neutral (the so-called ”passenger mutations”) and do not affect cell fitness \[26\]. In fact, one modeling study determined that half or more somatic mutations occur prior to the cancer initiating event \[19\]. This means the
prior mutations are either passenger mutations, or even if they are driver mutations they do not lead to cells with high enough fitness to overcome normal cells.

The multitype infinite-allele branching process allows modeling of both passenger and driver mutations. We referred to the subpopulations that have different fitness as different types in the model, and to the subpopulations with the same fitness but different genomes as different labels. Cells of different types have different sets of driver mutations in their ancestry, while those with different labels have different sets of passenger mutations. Our results from the model show that the number of mutations grows exponentially and at a rate proportional to the number of individuals alive. Previous studies [26] attempted to determine the correlation between the number of passenger mutations and driver mutations, which we can determine by the number of labels for each type in our model. Alternatively, we can create a more specific process where the type of individual refers to the number of driver mutations present and $K_i(n)$ would represent the total number of passenger alleles associated with $i$ driver mutations. Such a model would allow us to directly compare results to the previous studies. However, the model constrains us to assuming driver mutations are not unique in their effect on growth rates. Our model allows us to get around this constraint and adds more flexibility without requiring the high number of dimensions associated with considering each mutation (driver or passenger) as a specific type.

Markov Continuous Time The idea of crossing the infinite-alleles model of population genetics with branching processes originated with Griffiths and Pakes, who used the framework of a Bienaymé-Galton-Watson process [6]. In this model, an individual lives for a single generation before splitting according to the offspring probability generating function (p.g.f.) into a random number of offspring. Each offspring has the probability of initiating a new allele, or label, which is previously not seen in the population. Griffiths and Pakes obtained asymptotic results for the
total number of labels and the frequency spectrum, or total number of labels being represented by $j$ individuals. From here, two directions were taken to generalize the model for other scenarios. First, Pakes created a continuous-time Markov version of the infinite-alleles process, allowing individuals to live for a random lifetime having an exponential distribution[7]. This work was further generalized by Taïb, who employed general branching processes to investigate asymptotics of an infinite-allele process given random lifetimes and reproduction processes of individuals[8]. Doing so allows for the consideration of lifetimes of individuals that are not exponential, and strong convergence theorems are found that were left incomplete in the Markov process. We recently created a multitype version of the infinite-allele BGW process [58]. This model allows for multiple types with different offspring p.g.f. while still allowing for the presence of labels that do not affect the growth rate. We argued this model has applications in cancer biology and clonal evolution, where driver mutations (i.e. types) affect growth rates of populations, but passenger mutations (i.e. labels) do not. However, these passenger mutations affect heterogeneity in the population, and can be used as a molecular clock. We show asymptotic results for the number of labels and spectral radius in this process. Beyond that, the next step is to consider a continuous time version of the multitype BGW process, further generalizing the works of Pakes and Taïb.

We introduce two continuous-time multitype branching processes with infinite collection of labels. The extensions allow us to consider different rates of growth in branching processes as an effect of having different types. Along with this, the process allows any progeny individual born to be labeled with some probability $\mu$. Anytime an individual is given a label, the label is unique and new in the population to this individual and its descendants. Individuals of the same type with different labels do not have different offspring distributions or lifetime distributions, so they
are considered identically distributed. Thus we have a countably infinite number of possible labels in the population, but we do not need to track them along with the types, simplifying our process from an infinite-type process that would achieve the same goals.

The paper begins with the simpler of the models: the continuous-time Markov case. An $i$-type individual existing in this $k$-type process lives for an exponential lifetime. Upon death, it splits into individuals of other types according to its offspring distribution, $f_i(s)$. With probability $\mu$ each offspring is given a new label independently of each other. Otherwise it has the same label as its parent. We show the growth of labels in this process and the frequency spectrum have similar asymptotic properties as the single-type and discrete-time cases. For this process, we make the simplifying assumption that the mean matrix $M$ of the offspring distribution is irreducible. We provide a simple example of a birth-death-mutation model where an individual can either die or split into one of its own and one of any other types with a given probability.

We then introduce a general branching process version of the infinite-alleles multitype model. This model makes no assumption about the lifetime distribution or offspring distribution of individuals except that they follow a given distribution and point process. In this case an individual can live for a random amount of time. Instead of splitting at the end of its life, this individual can give birth to offspring of different types according to a point process. Upon birth, an offspring can have a new unique label with a given probability unique to the parent type. Asymptotic results from the theory of general branching processes are presented specific to the label process and frequency spectrum. Finally, we use the general process to develop results for the Bellman-Harris process, a framework sufficient for consideration of cell growth.
The result of these models create a general theory of infinite-allele branching processes that can be applied to clonal evolution and growth in tumor populations. The conclusion discusses the possible applications of these models for studying tumor evolution.

3.3 Continuous Time Markov Branching Process

Suppose we have a $k$-type branching process with offspring probability generating function $f(s)$ and mean matrix $M$ made up of elements $m_{ij} = \frac{\delta}{\delta s_i} f_i(s)|_{s=1}$. A type $i$ individual lives for an exponentially distributed amount of time with rate parameter $a_i$ before splitting into $j = (j_1, \ldots, j_k)$ offspring with probabilities given by the coefficients of $f_i(s)$. Upon splitting, each offspring can initiate a unique label with probability $\mu$ or remain the same label as their parent with probability $1 - \mu$.

If $\{Z(t), t \geq 0\}$ is the $k$-type branching process that counts the number of individuals of each type alive at time $t$, define $\tilde{Z}(t)$ as the ancestor branching process, or the process of individuals with the same label as the ancestor. The transition probabilities for $\tilde{Z}(t)$ are the coefficients of the p.g.f.

$$A_r(s; t) = E[s^{\tilde{Z}(t)} | Z(0) = e_r] = \sum_{j_1, \ldots, j_k} \tilde{q}_{rj}(t)s^j.$$ 

Following the notation in Pakes [7] define the ancestor p.g.f. for a type $i$ individual

$$H_i(s) = f_i(\mu + (1 - \mu)s)$$

and

$$h_i(s) = a_i(H_i(s) - s_i)$$

so that the ancestor generating functions for the continuous time process, $A(s; t)$,
satisfy the Backward Kolmogorov Equations,

\[ \frac{dA_i(s; t)}{dt} = h_i[A(s; t)]. \]

We’ll define the marginal probabilities of \( \tilde{q}_{r,j}(t) \) as \( q_{r,i,j}(t) \equiv P(\tilde{Z}_i(t) = j|Z(0) = e_r) \) which are the coefficients to the generating process \( A_r(1 - e_i(1 - s_i); t). \)

Frequency Spectrum and Number of Labels, \( K_i(t) \) We begin our study of the process by defining the frequency spectrum and the number of labels in the process. Let \( \alpha_{i,j}(t) \) be the number of \( i \)-type labels which have \( j \) individuals alive at time \( t \). As \( t \) grows, this value over its summation across the index \( j \) gives us a distribution of the long-run proportion of labels for the type \( i \).

Define the random sequences \( T_1, T_2, \ldots \) as the successive splitting times of \( Z(t) \), \( N(t) \) as the number of splits in \((0, t]\), and \( U_{n,i} \) as the total number of \( i \) type offspring produced at \( T_n \). Define the following two indicator functions:

\[ I_{0,r,i,j}(t) = I(\text{the ancestor is type } r \text{ and has } j \text{ type } i \text{ individuals alive with the same label at time } t) \]

\[ I_{n,m,l,i,j}(t) = I(\text{the } m^{th} \text{ } l \text{ type individual born at time } T_n \text{ is a new label and has } j \text{ } i \text{-type individuals of the same label at time } t). \]

Then \( \alpha_{i,j}(t) \) can be expressed as a sum of the indicators over all individuals in all
types and generations in the process,

\[
\alpha_{i,j}(t) = I_{0,r,i,j}(t) + \sum_{n=1}^{N(t)} \sum_{t=1}^{k} \sum_{m=1}^{U_{n,l}} I_{n,m,l,i,j}(t - T_n).
\]

That is, we go through each individual that was born in the population up to time \(t\) and only count the ones that initiated new labels.

The expectation of the frequency spectrum, \(\phi_{i,j}(t) = E[\alpha_{i,j}(t)]\) can then be found by finding the expectation of the individual parts of the expression, which is useful for our goal of finding asymptotic results. First,

\[
E[I_{0,r,i,j}(t)] = P[\tilde{Z}_i(t) = j|Z(0) = e_r] = q_{r,i,j}(t)
\]

where \(q_{r,i,j}(t)\) is a marginal probability for the transition probabilities of type \(i\) individuals with the ancestor label at \(t\) as we defined above. We also have

\[
E[I_{n,m,l,i,j}(t - T_n]|T_n] = \mu P[\tilde{Z}_i(t) = j|Z(0) = e_l] = \mu q_{l,i,j}(t - T_n),
\]

and if \(U_n = (U_{n,1}, \ldots, U_{n,k})\), then

\[
E[U_n|the \ n^{th} \ split \ is \ of \ a \ type \ w \ individual] = e^T_n M.
\]

Using this, the expectation of the frequency spectrum at time \(t\) for a process
initiated by a single type $r$ ancestor can be rewritten as

$$E_{e_r}[\alpha_{i,j}(t)] = q_{r,i,j}(t) + E_{e_r} \left[ \sum_{n=1}^{N(t)} \sum_{l=1}^{k} \sum_{m=1}^{I_{n,m,i,j}} I_{n,m,i,j}(t - T_n) \right]$$

$$= q_{r,i,j}(t) + \sum_{l=1}^{k} E_{e_r} \left[ \sum_{n=1}^{N(t)} E[U_{n,l}] \mu_{q_{l,i,j}}(t - T_n) \right]$$

$$= q_{r,i,j}(t) + \mu E_{e_r} \left[ \sum_{n=1}^{N(t)} \sum_{l=1}^{k} E[U_{n}] q_{l,i,j}(t - T_n) \right]$$

$$= q_{r,i,j}(t) + \mu E_{e_r} \left[ \sum_{n=1}^{N(t)} E[U_n] q_{i,j}(t - T_n) \right]$$

where $q_{i,j}(t) = (q_{1,i,j}(t), \ldots, q_{k,i,j}(t))^T$ is the $k \times 1$ vector of marginal probabilities indexed by the ancestor type. Let $\delta_{n,w}$ be the indicator that the $n^{th}$ split was a type $w$ individual and it’s accompanying probability is $\rho_{n,w}$. The expectation in the expression can be simplified further to

$$E_{e_r} \left[ \sum_{n=1}^{N(t)} E[U_n] q_{i,j}(t - T_n) \right] = E_{e_r} \left[ \sum_{n=1}^{N(t)} E[U_n] \delta_{n,w} = 1] \rho_{n,w} q_{i,j}(t - T_n) \right]$$

$$= E_{e_r} \left[ \sum_{n=1}^{N(t)} \sum_{w=1}^{k} e_w^T M \rho_{n,w} q_{i,j}(t - T_n) \right]$$

$$= E_{e_r} \left[ \sum_{n=1}^{N(t)} \rho_n^T M q_{i,j}(t - T_n) \right]$$

where $\rho_n = (\rho_n,1, \ldots, \rho_n,k)^T$. Define this expectation term

$$\beta_r(t) = E_{e_r} \left[ \sum_{n=1}^{N(t)} \rho_n^T M q_{i,j}(t - T_n) \right].$$
Note that if we had $z_0$ ancestors, then $\beta_{z_0}(t) = z_0 \beta_r(t)$, so we only write $\beta$ as a function of a single ancestor. If we consider the vector $\beta(t) = (\beta_1(t), \ldots, \beta_k(t))^T$ and we had $z_0$ ancestors, then $\beta_{z_0}(t) = z_0 \beta(t)$. Conditioning on the time and type of the first split gives

$$E_{e_1} \left[ \sum_{n=1}^{N(t)} \rho_n^T M q_{i,j}(t - T_n)|T_1, U_1 \right] =$$

$$= \rho_1^T M q_{i,j}(t - T_1) + U_1 \left( E_{e_1} \left[ \sum_{n=2}^{N(t)} \rho_n^T M q_{i,j}(t - T_n)|T_1 \right] \right) =$$

$$= e_1^T M q_{i,j}(t - T_1) + U_1 \left( E_{e_1} \left[ \sum_{n=2}^{N(t)} \rho_n^T M q_{i,j}(t - T_n)|T_1 \right] \right)$$

where $T_n' = T_n - T_1$ and $N'(t)$ is the number of splits in $(T_1, t]$. This then leads us to an expression for $\beta_r(t)$ by taking the expectation of the above expression:

$$\beta_r(t) = \int_0^t \left( e_1^T M q_{i,j}(t - u) + e_1^T M \beta(t - u) \right) f_{T_1}(u) du$$

$$= \int_0^t \left( e_1^T M q_{i,j}(t - u) + e_1^T M \beta(t - u) \right) a_r e^{-a_r u} du$$

$$= \int_0^t \left( e_1^T M q_{i,j}(u) + e_1^T M \beta(u) \right) a_r e^{-a_r(t - u)} du \iff$$

$$e^{a_r t} \beta_r(t) = \int_0^t \left( e_1^T M q_{i,j}(u) + e_1^T M \beta(u) \right) a_r e^{a_r(u)} du \iff$$

$$a_r e^{a_r t} \beta_r(t) + e^{a_r t} \beta'_r(t) = (a_r e_1^T M q_{i,j}(t) + a_r e_1^T M \beta(t)) e^{a_r t} \iff$$

$$\beta'_r(t) = a_r e_1^T M q_{i,j}(t) + a_r e_1^T M \beta(t) - a_r \beta_r(t)$$
which is equivalent to

$$
\beta_r'(t) = a_r e_r^T M q_{i,j}(t) + a_r e_r^T M \beta(t) - a_r \beta_r(t)
$$

so our solution satisfies the linear system of differential equations:

$$
\beta'(t) = D_a M q_{i,j}(t) + D_a (M - I) \beta(t)
$$

with the initial condition $\beta(0) = 0$. The matrix $D_a$ is the diagonal matrix with entries from $a$, the splitting rates. The solution to this is

$$
\beta(t) = \int_0^t e^{(t-u)D_a(M-I)} D_a M q_{i,j}(u) du
$$

$$
= e^{tD_a(M-I)} \int_0^t e^{-uD_a(M-I)} D_a M q_{i,j}(u) du
$$

$$
\beta_r(t) = e_r^T e^{tD_a(M-I)} \int_0^t e^{-uD_a(M-I)} D_a M q_{i,j}(u) du
$$

Note that $D_a(M - I)$ is the infinitesimal generator of the semigroup $\{M(t); t \geq 0\}$ where $M(t)$ is the mean matrix of the process $Z(t)$. Thus we get the final result for the frequency spectrum with a single $r$ type ancestor:

$$
\phi_{i,j}(t) = E[\alpha_{i,j}(t)] = q_{r,i,j}(t) + e_r \mu e^{tD_a(M-I)} \int_0^t e^{-uD_a(M-I)} D_a M q_{i,j}(u) du.
$$
If we check this with the 1-type case, we get

\[
E[\alpha_j(t)] = q_j(t) + \mu \beta(t)
\]

\[
= q_j(t) + \mu a^t e^{a(t-1)} \int_0^t e^{-u a(m-1)} q_j(u) du
\]

\[
= q_j(t) + \mu a^t \int_0^t e^{-u \lambda} q_j(u) du
\]

which is similar to the result of the single type case [7].

Along with the frequency spectrum, we define \( K_i(t) \) as the number of labels present at time \( t \) for the \( i \)th type. In this case we can write \( K_i(t) = \sum_{j \geq 1} \alpha_{i,j}(t) \) and elucidate the various results. First, we define the marginal probability vector \( \mathbf{q}_i,(t) = A(1 - \mathbf{e}_i; t) = (q_{1,i},(t), \ldots, q_{k_i,i},(t))^T \), which is the probability of having 0 \( i \)-type individuals with the ancestor label at time \( t \). The \( r \)th element of the vector refers to the probability for an ancestor of type \( r \). We can write the expression for \( K_i(t) \) in terms of the sums of indicators as well

\[
K_i(t) = \sum_{j \geq 1} I_{0,r,i,j}(t) + \sum_{n=1}^{N(t)} \sum_{l=1}^{k} \sum_{m=1}^{U_{n,l}} I_{n,m,l,i,j}(t - T_n).
\]

Since \( K_i(t) \) counts all labels represented by more than one individual, it is the same as the sum of the frequency spectrum over the index \( j \). Thus, the expectation of \( K_i(t) \) follows easily by using the complement of the probability of no individuals \((1 - q_{r,i},)\) in place of \( q_{r,i,j},\)

\[
E_r[K_i(t)] = 1 - q_{r,i},(t) + e_r \mu e^{tD_a(M-1)} \int_0^t e^{-uD_a(M-1)} D_a M(1 - \mathbf{q}_i,(u)) du.
\]

Figure 3.6 shows a realization of the time continuous process. This process con-
Figure 3.6: An example of the multitype infinite-allele continuous-time Markov Process shows how we can count types (colors) and neutral alleles (gray trees). A lightning bolt represents a new label is initiated in the individual, so it becomes an ancestor to a new labeled process.

tains three types distinguished by their color. At any split, a new label can be initiated with probability $\mu$ represented by the lightning bolt. These labels initiate their own new process which grows with a distribution identical to the original process (given the same ancestor type).

* 

Asymptotics of $\phi_{i,j}(t)$ and $K_i(t)$ Define $A \equiv D_a(M - I)$, and let $\lambda$ be the spectral radius of $A$ with associated left and right eigenvectors $v$ and $u$. The branching process is considered supercritical if $\lambda > 0$, critical when $\lambda = 0$, and subcritical when $\lambda < 0$. Note the eigenvalue of $e^{At}$ is $e^{\lambda t}$ with left and right eigenvectors $v$ and $u$. Then using
the fact that
\[ \lim_{t \to \infty} e^{At}e^{-\lambda t} = uv \]
we get the following results about the asymptotics of the expectation:

**Lemma 6.** Suppose we have a branching process initiated by a single type \( r \) individual. Then
\[ \lim_{t \to \infty} \frac{\phi_{ij}(t)}{e^{t\lambda}} = e_r^T\mu uvD_aM \int_0^\infty e^{-u\lambda}q_{ij}(u)du \]
and
\[ \lim_{t \to \infty} \frac{E[K_i(t)]}{e^{t\lambda}} = e_r^T\mu uvD_aM \int_0^\infty e^{-u\lambda}(1 - A_i(u))du \]

*Proof.* The result can be found by plugging into the expressions as follows
\[ \lim_{t \to \infty} e^{-t\lambda}\phi_{ij}(t) = e_\alpha^T\mu uv \int_0^\infty e^{-uA}D_aMq_{ij}(u)du \]
\[ = e_\alpha^T\mu uvD_aM \int_0^\infty e^{-u\lambda}q_{ij}(u)du, \]
and \( \lim_{t \to \infty} e^{-t\lambda}E[K_i(t)] = e_\alpha^T\mu uv \int_0^\infty e^{-uA}D_aM(1 - A_i(u))du \]
\[ = e_\alpha^T\mu uvD_aM \int_0^\infty e^{-u\lambda}(1 - A_i(u))du. \]

Note that \( \frac{1-\eta_{r,i}(t)}{e^{-t\lambda}} \to 0 \) regardless of the value of \( \lambda \) by Markov’s inequality. We make use of the eigenvector relation \( ve^{-tA} = ve^{-t\lambda} \) to reduce the second term of the expression. \( \square \)

We can also decide to normalize the process by \( E[Z(t)u] \) as in the discrete time version[58]. From general theory on multitype branching processes, given a type \( r \) ancestor, \( \lim_{t \to \infty} Z(t)ue^{-\lambda t} = W_r \) with \( E[W_r] = u_r = e_r^T u \) [2]. When the ancestor is
type $r$,

$$\lim_{t \to \infty} \frac{E[K_i(t)]}{E[Z(t)u]} = \mu v \int_0^\infty e^{-uA} D_a M (1 - A_i(u)) du$$

$$= \mu v \int_0^\infty e^{-u\lambda} D_a M (1 - A_i(u)) du$$

$$= \mu v D_a M \int_0^\infty e^{-u\lambda} (1 - A_i(u)) du$$

since $ve^{-xA} = ve^{-\lambda}$. The frequency spectrum has a similar result with

$$\lim_{t \to \infty} \frac{\phi_{ij}(t)}{E[Z(t)u]} = \mu v D_a M \int_0^\infty e^{-u\lambda} q_{ij}(u) du.$$  

Stronger results about the convergence of $K_i(t)$ exist, but are not proved here. Instead, we leave them as applications of the asymptotic results in the general branching process case, shown later.

**Example: Birth-Death-Mutation Process**

We attempt to illustrate the above with an example using a multitype birth-death-mutation process where individuals live for a random amount of time before either dying or splitting into 2 individuals where one must be of the same type. We assume a type $i$ individual has an exponentially distributed lifetime with mean $a_i^{-1}$. At the end of it’s lifetime it splits into another type $i$ individual and a type $j$ individual with probability $p_{ij}, j = 1, \ldots, k$. Note that a type $i$ individual can split into two of its own type with probability $p_{ii}$. Alternatively, the individual can die with probability $p_{i0} = 1 - \sum_{j=1}^k p_{ij}$. The p.g.f. for the associated Galton-Watson process is then

$$f_i(s) = p_{i0} + \sum_{j=1}^k p_{ij} s_i s_j.$$
We need to determine the p.g.f. for the continuous time ancestor process, \( A_i(s; t) \). We use the backward Kolmogorov equation to begin solving this, where the solution will end up in the form of a system of Ricatti equations.

\[
\frac{d}{dt} A_i(s; t) = h_i(A(s; t))
\]

\[
= a_i [H_i(A(s; t)) - A_i(s; t)]
\]

\[
= a_i [f_i(\mu + (1 - \mu)A(s; t)) - A_i(s; t)]
\]

\[
= a_i \left[ p_{i0} + \sum_{j=1}^{k} p_{ij} [\mu + (1 - \mu)A_i(s; t)][\mu + (1 - \mu)A_j(s; t)] - A_i(s; t) \right]
\]

\[
= a_i \left[ p_{i0} + \sum_{j=1}^{k} p_{ij} \mu^2 + \sum_{j=1}^{k} p_{ij} (1 - \mu) \mu A_i(s; t) + \sum_{j=1}^{k} p_{ij} (1 - \mu) \mu A_j(s; t) - A_i(s; t) \right]
\]

\[
= a_i \left[ p_{i0} + (1 - p_{i0}) \mu^2 + [(1 - p_{i0})(1 - \mu) \mu - 1 + (1 - \mu) \mu p_{ii}] A_i(s; t) + \right.
\]

\[
+ \sum_{j \neq i} p_{ij} (1 - \mu) \mu A_j(s; t) + \sum_{j=1}^{k} p_{ij} (1 - \mu)^2 A_i(s; t) A_j(s; t) - A_i(s; t) \right]
\]

\[
= a_i \left[ \kappa_i + C_{ii} A_i(s; t) + \sum_{j \neq i} C_{ij} A_j(s; t) + \sum_{j=1}^{k} \Gamma_{ij} A_i(s; t) A_j(s; t) \right]
\]

with

\[
\kappa_i = p_{i0} + \mu^2 (1 - p_{i0}),
\]

\[
\Gamma_{ij} = p_{ij} (1 - \mu)^2,
\]

and

\[
C_{ij} = \begin{cases} 
(1 + p_{ii} - p_{i0}) \mu (1 - \mu) - 1 & : i = j \\
\frac{p_{ij} \mu (1 - \mu)}{p_{i0}} & : i \neq j 
\end{cases}
\]
leading to the system of differential equations

\[
\frac{d}{dt} A(s; t) = D_a [\kappa + CA(s; t) + \text{diag}(A(s; t)) \Gamma A(s; t)]
\]

\[
\frac{d}{dt} A(s; t)_{(k \times 1)} = \text{diag}(a)_{(k \times k)} \left[ \kappa_{(k \times 1)} + C_{(k \times k)} A(s; t)_{(k \times 1)} + \text{diag}(A(s; t))_{(k \times k)} \Gamma_{(k \times k)} A(s; t)_{(k \times 1)} \right]
\]

with initial conditions \( A_i(s; 0) = s_i^{I(i\text{-type ancestor})} \).

While we do not get a formal solution for this, we can solve numerically to find a solution. For this process, the mean matrix is

\[
M = \begin{bmatrix}
2p_{11} & p_{12} & \ldots & p_{1k} \\
p_{21} & 2p_{22} & \ldots & p_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
p_{k1} & p_{k2} & \ldots & 2p_{kk}
\end{bmatrix}
\]

and infinitesimal generator

\[
A = \begin{bmatrix}
2a_1p_{11} - 1 & p_{12} & \ldots & p_{1k} \\
p_{21} & 2a_2p_{22} - 1 & \ldots & p_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
p_{k1} & p_{k2} & \ldots & 2a_kp_{kk} - 1
\end{bmatrix}.
\]

Even in a simplified splitting process with mutations, closed form results about the multitype continuous-time process are still elusive, and must be determined numerically.
3.4 General Branching Process and Results

Introduction to general branching processes

A more expansive theory on the single-type infinite allele model was introduced by Taib following the theory of a Jagers-Crump-Mode process\[8][67]. This generalization allows us to consider individuals that can give birth during their life without the individual necessarily having an exponential lifetime or splitting. The probability of a mutation to a new label is also unique to an individual’s type which is a more accurate model of mutations in cancer, as opposed to a baseline rate of new labels we use before and in the BGW scenario.

We begin by introducing the theory of general branching process. We define an individual of a population as \( x = (x_1, x_2, \ldots, x_n) \in \mathbb{N}^n, n \in \mathbb{N} \) as the \( x_n \)th child of the \( x_{n-1} \)th child of the \( \ldots \) of the \( x_1 \)th child of 0, where 0 is the ancestor of the population. The dimension of \( x \) is \( n(x) \) which can also be called the generation of \( x \).

We can also label an individual \( xy \) for the individual where the first \( n \) components are \( x \) and the last \( m \) components are \( y \). The first \( n - k \) ancestors of \( x \) are denoted \( x[k] = (x_1, \ldots, x_{n(x) - k}) \). Let \( I \) be the set of all descendants of the population,

\[
I = \bigcup_{n \geq 0} \mathbb{N}^n
\]

where \( \mathbb{N}^0 = 0 \). This set represents all possible conceivable individuals in the population, so any individual that can exist is an element of this set. We say that \( 0x = x0 = x \).

At birth, an individual inherits a life history, \( \omega \), from the space \((\Omega, \mathcal{A})\) containing all possible life histories. There exists a sequence \((\tau(k, \omega))_{k \in \mathbb{N}}\) of random variables representing the age of the individual with life history \( \omega \) at the time of birth of her \( k \)th child. At birth a daughter inherits a type, \( r \), from the space \((S, \mathcal{S})\) which can
affect the choice of life history. The probability measure $P_r \in \{P_s; s \in S\}$ on the set $(\Omega, \mathcal{A})$ determines the life history of an individual. We define the reproduction point process $\xi$ on $S \times \mathbb{R}^+$ by

$$
\xi(A \times B, \omega) := \#\{i \in \mathbb{N}; \rho(i, \omega) \in A, \tau(i, \omega) \in B\}
$$

where $\rho(i, \omega)$ is a measurable function giving the type of the $i^{th}$ child of the individual with life history $\omega$. Additionally, $\lambda : S \times \Omega \to \mathbb{R}^+$ is a random variable on defined as the life length of an individual, and the mutation index, $\gamma : \mathbb{N} \times \Omega \to \{0, 1\}$ describes the label of an offspring by

$$
\gamma(k, \omega) = I(\text{the } k^{th} \text{ child of an individual with life history } \omega \text{ has the same label as the individual}).
$$

The sample space for the branching process, or population space for a population started by ancestor 0 is constructed as $(S \times \Omega^I, S \times \mathcal{A}^I)$ where an outcome consists of a starting type for the ancestor and the life history of each individual in the population, $I$. Any individual, $\mathbf{x}$ in the set $I$ has a life history $\omega \in (S \times \Omega^I)$ that describes its life, including the births and types of offspring. The probability measure is $P_r$ where the ancestor is of type $r$. The evolution of the population is defined in terms of the birth times of individuals where $\sigma_\mathbf{x}$ is the birth time of the individual $\mathbf{x}$. The function is defined recursively with $\sigma_0 = 0$ for the birth time of the ancestor, and $\sigma_{\mathbf{xk}} = \sigma_\mathbf{x} + \tau(k, \omega_\mathbf{x})$. A shift operator, $S_\mathbf{x} : S \times \Omega^I \to S \times \Omega^I$, $\mathbf{x} \in I$ is introduced, mapping $(s_0, \{\omega_\mathbf{y}; \mathbf{y} \in I\})$ to $(s_\mathbf{x}, \{\omega_{\mathbf{xy}}; \mathbf{y} \in I\})$ which effectively treats $\mathbf{x}$ as an ancestor of its own process. Thus $S_0$ is the actual population process itself since it refers to the ancestor, 0. We also let $\rho_\mathbf{x} = \rho(\mathbf{x}_{n(\mathbf{x})}, \omega_{\mathbf{x}[1]})$ denote the type of the individual $\mathbf{x}$ and $\gamma_\mathbf{x} = \gamma(\mathbf{x}_{n(\mathbf{x})}, \omega_{\mathbf{x}[1]})$ denote the label of the individual $\mathbf{x}$. A special case is noted in
Jagers and Nerman for splitting processes, i.e. Galton-Watson processes, Bellman-Harris processes, and any we’ve been interested up to this point. In these cases, an individual does not reproduce until death, so \( \sigma_1 = \sigma_2 = \cdots = \sigma_k \) which is also equal to \( \sigma + \lambda \) where \( \lambda \) is the age of \( x \) at its death, \( \lambda : (S \times \Omega \rightarrow \mathbb{R}^+) \).

With information about individuals and their life histories set up, we can now begin to define branching processes in the population as processes that count contributions from individuals. That is, we say a branching process at time \( t \) is found by considering all individuals born up to time \( t \). For a single individual, we use a function to determine its contribution to the population known as a \textit{random characteristic}.

\textbf{Definition 6 (Random Characteristic).} A \textit{random characteristic} is a real-valued measurable function, \( \chi : S \times \Omega \times \mathbb{R} \rightarrow \mathbb{R}^+ \) associated with the individual \( x \) by \( \chi_x(a) = \chi \circ S_x(a) \) at age \( a \in \mathbb{R} \).

Characteristics are usually nonnegative because we are interested in considering contributions of individuals (i.e. whether the individual was born, number of children, etc.). We use characteristics to count by summing over the characteristics for all individuals in the population space, \( I \). A branching process counted by the characteristic \( \chi \) can be written as the sum

\[
Z^\chi(t) = \sum_{x \in I} \chi_x(S_x, t - \sigma_x)
\]

where the random characteristics provide the contribution of each individual to the branching process at their age at time \( t \). In this case, we note that \( S_x \) contains the type and the lives of \( x \) and her progeny. We give a few simple examples that are commonly used. First, a simple characteristic is

\[
\chi(a) = I_{\mathbb{R}^+}(a)
\]
which is the indicator that the age of an individual is positive, or the individual is born ($a \in \mathbb{R}^+$). That is, if for an individual, $x$, $a \in [t - \sigma_x, \infty)$, then the individual was born before $t$. Using this characteristic, and counting over the entire population leads to the branching process

$$Y(t) = Z^\chi(t)$$

that counts all births in the population up to $t$. A second example counts the number of individuals alive at time $t$, or the most commonly used branching process. Let $\chi(a) = I_{(0,\lambda_\omega)}(a)$ which is the indicator that an individual was born, but hasn’t died as of time $t$. In this case,

$$Z(t) = Z^\chi(t).$$

These two examples are the simplest versions which don’t consider the type or even children of the individual. More complicated examples can be created based on the progeny or ancestors of an individual. We call a characteristic that only relies on $\omega_x$, or the life history of $x$, an individual characteristic. Asymptotic theorems are created by Jagers that require a process to be counted by individual characteristics [68]. We face situations where we need to rewrite a characteristic in order to make it an individual characteristic when we introduce the infinite-allele general branching process. From here, we can establish the well-known branching property, that the number of individuals at a time $t$ is equal to the contribution of the ancestor along with the processes of its daughter up to time $t$, or

$$Z^\chi(t) = \chi_0(t) + \sum_{j \in \mathbb{Z}^+} Z^\chi(t - \sigma_j)$$

where $Z^\chi(t - \sigma_j)$ is the process created by the $j^{th}$ daughter of 0.

We need to define other mappings as well about the expectations of the reproduction point process[69]. First, the reproduction kernel, or expected number of children
is
\[ \mu(r, A \times B) = \int_\Omega \xi(A \times B, \omega) P_r(d\omega) = E_r[\xi(A \times B)], r \in S, A \in \mathcal{S}, B \in \mathcal{B} \]
or \[ \mu(r, ds \times dt) = E_r[\xi(ds \times dt)]. \]

We also define \( \mu(\alpha, r, ds \times dt) = e^{-\alpha t} \mu(r, ds \times dt) \).

This will be useful for notational purposes when setting up Laplace transforms of our characteristics.

Define the composition operation, \( \ast \), as the transition on \( S \) and convolution on \( \mathbb{R}^+ \) such that
\[ \mu \ast \mu(r, A \times B) = \int_{S \times \mathbb{R}^+} \mu(s, A \times (B - t)) \mu(r, ds \times dt) \]
and the \( n \)-fold composition is defined as \( \mu^{*n}(r, A \times B) \). The series of \( n \)-fold compositions defined as the measure \( \nu \equiv \sum_{n=0}^{\infty} \mu^{*n} \). For \( k = 0 \), we have \( \mu^{*0}(r, A \times B) = 1_{A \times B}(s, 0) \). In the single type case, we do not need to include the transition on the state space, \( S \), and instead define \( \ast \) as the convolution on the positive real line. However, we’re interested in the multitype process. Additional information about the kernel, irreducibility, and the Malthusian parameter, \( \alpha \), can be found in Jagers, but is not relevant to our processes here[69].

Given the reproduction kernel, \( \mu \), we assume the population is Malthusian, so there exists an \( \alpha > 0 \) such that
\[ \hat{\mu}_\alpha(r, ds) = \int_{\mathbb{R}^+} \mu_\alpha(r, ds \times dt) \]
has Perron root one. Under conditions outlined in the previous sources, \( \hat{\mu}_\alpha(r, ds) \) has eigenmeasure \( \pi \) given by
\[ \pi(A) = \int_S \hat{\mu}_\alpha(r, A) \pi(dr) \]
and eigenfunction

\[ h(r) = \int_{S} h(s) \mu_{\alpha}(r, ds), \quad r \in S \]

where both are normed so that \( \int_{S} h(s) \pi(ds) = 1 \).

Last we define the mean age at child bearing as

\[ \beta = \int_{S \times S \times \mathbb{R}^{+}} t e^{-\alpha t} h(s) \mu(r, ds \times dt) \pi(dr). \]

and the intrinsic martingale,

\[ w_t = \sum_{\{x \in I, \sigma_{x} \leq t \leq \sigma_{x}'\}} e^{-\alpha \sigma_{x}} h(\rho_{x}) \]

which has the limit \( \lim_{t \to \infty} w_t = w \). Most of these definitions will be used for asymptotic results about the infinite-alleles model.

**Asymptotic Results from General Branching Processes**

Limit theorems were created for branching processes described by random characteristics that are used to show convergence of the infinite-allele case\[68\]. These convergence theorems actual show solutions for the previous processes we described from a different direction. That is, we could have set up the Galton-Watson process and Markov continuous-time process as applications of the general process as in Taïb \[8\].

In the Markov continuous-time case, one result, almost-sure convergence, could not be shown. This theory helps complete the missing result there. The theorems are set up in terms of single-type processes, but the extension to multitype only involves adjusting the characteristics and kernel to reflect multiple dimensions. First, we as-
sume a supercritical general branching process with a Malthusian parameter, \( \alpha \) that exists, and \( \mu \) is non lattice. Last, we assume \( E[\hat{\xi}(\alpha) \log \hat{\xi}(\alpha)] < \infty \). We also define \( \beta \) as the expected age at childbearing, or

\[
\beta = \int_0^\infty ue^{-\alpha u} \mu(du) < \infty.
\]

If these assumptions hold, we get the following results:

**Theorem 9** (Convergence of Expectation). Let \( \chi \) be a random characteristic such that

\[
\sum_{n \geq 0} \sup_{n \leq u \leq n+1} e^{-\alpha u} E[\chi(u)] < \infty
\]

and

\[
E[\chi] \text{ is a.e. continuous},
\]

then

\[
\lim_{t \to \infty} \frac{E[Z^\chi(t)]}{e^{at}} = \frac{E[\hat{\chi}(\alpha)]}{\alpha \beta}.
\]

**Theorem 10** (Almost Sure Convergence). Under the same assumptions of Theorem 9, then

\[
\lim_{t \to \infty} \frac{Z^\chi(t)}{Y(t)} = E[\hat{\chi}(\alpha)] \text{ on } \{Y(t) \to \infty\} \text{ in probability.}
\]

Theorems 9 and 10 show similar convergence results to what we’ve previously shown. First, the expectation converges as \( t \) gets large, and also the normalized process converges on the nonextinction set. The difference in this case is we normalize by the number of individuals ever born instead of by the expected rate of growth of the process, \( e^{at} \). Almost sure convergence also exists when there exists some values \( \alpha' < \alpha \) such that \( E[\sum_{t \geq 0} e^{-\alpha' t} \chi(t)] < \infty \), and \( \hat{\mu}(\alpha') < \infty \). This result might be more important for us, because it allows us to find almost sure limits of ratios by using the Continuous mapping theorem for almost sure convergence. That is,
Corollary 2 (Almost sure convergence of ratios). Let $\chi$ and $\chi'$ be two difference characteristics under the same assumptions as above. Then,

$$\lim_{t \to \infty} \frac{Z^\chi(t)}{Z^{\chi'}(t)} \to \frac{E[\hat{\chi}(\alpha)]}{E[\hat{\chi}'(\alpha)]}$$

almost surely as $t \to \infty$ on $\{Y(t) \to \infty\}$.

Corollary 2 becomes useful when dealing with sequencing data given in the application. We do not have cell counts available to us. Instead, we have allele frequencies, so having convergence of frequencies can be more useful for estimation. Finally, we present convergence of the process normalized by its mean growth.

Theorem 11. Assuming the conditions of Theorems 9 and 10 are met, then there exists a random variable, $w$

$$\lim_{t \to \infty} \frac{Z^\chi(t)}{e^{\alpha t}} \to \frac{E[\hat{\chi}(\alpha)]w}{\alpha \beta}$$

on $\{Y(t) \to \infty\}$ in $L^1$ and in probability.

Thus we have a similar, but weaker, version of convergence as a tool we can use to show our processes converge in the general process. These results are important because they show that as time gets sufficiently large, the process begins to grow deterministically. We can use this information and the assumption of long enough time to make estimates for certain growth rates, times, and parameters in our model. These estimates are discussed in the application. First, we need to create a general infinite-allele branching process and show we can come up with a random characteristic version of the process that satisfies the above conditions. The work for a single type process has been covered, but we extend to the multitype setting[8].
The infinite-allele general branching process

With the basics for a general branching process, we can begin to reimagine the infinite-allele branching process in terms of a general branching process. For this, we consider a multitype general process, but need to define a process for labeling offspring as ancestor or mutant with the same set up as the discrete and continuous Markov cases. Above, we used the function $\gamma_x$ to define the label of an individual which we can incorporate type into. This way, we can allow different types to have different mutation probabilities, further generalizing the previous processes even more. For an individual, we have the reproduction point process of type $s$ offspring, $\xi(s, t)$ which can be split into offspring with the same label as the parent or a new label not yet seen in the population. The process is split into $\tilde{\xi}(s, t)$ for mutant children and $\tilde{\xi}(s, t)$ for nonmutant children with type $s$. We can write $\tilde{\xi}(s, t)$ as

$$\tilde{\xi}(s, t) = \sum_{i \in \mathbb{N}} \gamma(i, \omega_0) 1_{[0, \infty)}(t - \tau(i, \omega_0)) 1_{\{s\}}(\rho(i, \omega_0))$$

and $\xi(s, t) = \tilde{\xi}(s, t) + \tilde{\xi}(s, t)$. The reproduction kernel can also be split into its label and unlabeled parts, $\mu(r, A \times B) = \tilde{\mu}(r, A \times B) + \tilde{\mu}(r, A \times B)$. The branching process initiated by an individual $0$ can then be split into the processes from all unlabeled offspring and the processes from all labeled offspring. The branching property still exists in these daughter population processes.

Extinction and the Ancestor Label

The ancestor label forms its own subprocess which can be determined by using the process $\tilde{\xi}$ in place of $\xi$. We can use results about this process and its fate to determine other processes, $Z^\chi(t)$, based on the characteristic that an individual has its parent label, $\chi$. We are interested in the probability of the ancestor label going extinct and
the probability of seeing no individuals of a specific type with the ancestor label (not the label is not necessarily extinct in this event). First, we consider the branching process counting the number of individuals with the ancestor label. This process has types in $S$, and we count the total number of individuals alive at time $t$ by the vector

$$\tilde{Z}(t) = \text{the number of living individuals carrying}$$

$$\text{the ancestor label at time } t$$

which contains the components

$$\tilde{Z}_s(t) = \text{the number of } s\text{-type living individuals carrying}$$

$$\text{the ancestor label at time } t$$

Then we have the probability of extinction

$$\tilde{q}_{r,0}(t) = P(\tilde{Z}(t) = 0|\tilde{Z}(0) = e_r)$$

and the probability of having no $s$-type individuals with the ancestor label alive at time $t$

$$q_{r,s,0}(t) = P(\tilde{Z}_s(t) = 0|\tilde{Z}(0) = e_r).$$

These are defined the same way as the Markov model presented above.

**Total Number of Labels**

Define $N_i(t)$ as the total number of $i$-type labels ever up to time $t$ excluding the ancestor label. We can define a random characteristic to write the number of individuals as a branching process. This essentially comes down to counting every type $s$ individual that experiences a mutation, but must be defined in a way that is a valid
characteristic. If we define the characteristic $\chi$ as

$$\chi_x(i, u) = \begin{cases} 
1 & \text{if } x \text{ is a type } i \text{ mutant and } u \geq 0 \\
0 & \text{otherwise}
\end{cases}$$

then the life history in $S_x$ depends on the mother of the individual $x$, so $\chi_x' \neq \chi' \circ S_x$.

We instead rewrite the characteristic to count the number of offspring of an individual that are mutants and take the sum of these characteristics (which can take a value in $\mathbb{N}$). Note the characteristic

$$\chi_x(i, u) = \sum_{j \in \mathbb{N}} (1 - \gamma(j, \omega_x))1_{\{i\}}(\rho(j, \omega_x))1_{\mathbb{R}^+}(u - \tau(j, \omega_x))$$

allows us to write $\chi_x = \chi \circ S_x$ so that

$$N_i(t) = \sum_{x \in I \setminus \{0\}} \chi_x(i, t - \sigma_x) = \sum_{x \in I} \chi_x(i, t - \sigma_x) = Z^x(t).$$

Rewriting the characteristic this way allows us to write the mutant offspring process as $\tilde{\xi}_0(i, t) = \chi(i, t)$. Then we can use the same technique in Taïb [8] and Jagers and Nerman [68] to show

$$E_r[N_i(t)] = E[\chi_x] * \nu(r, S \times [0, t])$$

with $E_r[\chi(t)] = \tilde{\mu}(r, S \times [0, t])$.

Since $\chi$ is an individual characteristic that is a.e. continuous, the assumptions to prove convergence of the process $N_i(t)$ is satisfied. Let

$$\tilde{\xi} = \int_{S \times \mathbb{R}^+} e^{-\alpha t}h(s)\xi(ds \times dt)$$
and

\[ E_\pi[\cdot] = \int_S E_s[\cdot] \pi(ds). \]

**Theorem 12.** Suppose the ancestor is of type \( \rho_0 = r \in S \). If the process is supercritical (\( \alpha > 0 \)) and \( E_\pi[\xi \log \xi] < \infty \), then

1. \( \frac{E_r[N_i(t)]}{e^{\alpha t}} \to h(r) \frac{E_\pi[\hat{\chi}(\alpha)]}{\alpha \beta} \) as \( t \to \infty \) where \( E_r[\hat{\chi}(\alpha)] = \hat{\mu}_\alpha(r, S) \).

2. \( \frac{N_i(t)}{e^{\alpha t}} \to \frac{E_r[\hat{\chi}(\alpha)]}{\alpha \beta} w \) in \( L^1[\mathbb{P}_s] \) for \( \pi \)-almost all \( s \in S \) as \( t \to \infty \) where \( E_r[w] = h(r) \).

3. \( \frac{N_i(t)}{Y(t)} \to E_\pi[\hat{\chi}(\alpha)] \) a.s. on the nonextinction set \( \{Y(t) \to \infty\} \)

**Proof.** These results are restatements of those about branching processes counted by random characteristics in Jagers and Nerman [68] as applied to the branching process \( N_i(t) \). By showing the characteristic \( \chi \) is an individual characteristic and a.e. continuous, the results hold. 

\( \square \)

**Current labels with living individuals,** \( K_i(t) \)

Define \( K_i(t) \) as the number of labels represented by at least one type \( i \) individuals that is alive at time \( t \) without including the ancestral label. Note that this is different than the Markov case because we do not include the ancestral label here. \( K \) is counted with the characteristic

\[ \chi'_X(u) = 1(X \text{ has a new label with at least one living type } i \text{ descendent with the same label at age } u \geq 0) \]

which like before is not a true characteristic (\( \chi_X \neq \chi \circ S_X \)). Instead, we can count characteristics based on and individual's children rather than looking at an individual
directly in order to get the characteristic

\[ \chi_x(a) = \sum_{j \in \mathbb{N}} (1 - \gamma(j, \omega_x))1_{\mathbb{R}^+} (a - \tau(j, \omega_x))1_{[1,\infty)}(\tilde{Z}_i(a) \circ S_{xj}) \]

where \( \tilde{Z}_i(a) \) is defined as the number of \( i \) type individuals with the ancestral label alive at time \( a \). The term is rewritten as a process using a different characteristic by

\[ \tilde{Z}_i(t) = \sum_{x \in I} \chi^{\dagger}_x(t - \sigma_x) \]

with

\[ \chi^{\dagger}_{1,x}(a) = 1_{[0,\lambda_x]}(a)1_{\{i\}}(\rho(x_{n(x)}, \omega_{x[1]})). \]

Note that since this is still in terms of the daughters’ types and the branching processes started by the daughter, \( \chi_x \) is still a characteristic. However, it is not an individual characteristic since it depends on descendants of the individual. From here, we can show the basic convergence theorems hold.

**Theorem 13.** \( \frac{E[K_i(t)]}{e^{st}} \to h(s) \frac{E_x[\hat{\chi}(\alpha)]}{\alpha^3} \) as \( t \to \infty \) where \( E_x[\hat{\chi}(\alpha)] = \hat{\mu}_\alpha(r, S)(1 - \hat{q}(S, i, \alpha)). \)

**Proof.** Since \( \chi \) is a random characteristic that is continuous almost everywhere, the convergence theorems for general branching processes hold. Thus we only need to show the result of the expectation. The final step occurs since we have the Laplace transform of a convolution.
\begin{equation*}
E_r[\hat{\chi}(\alpha)] = E \left[ \int_0^\infty e^{-\alpha t} \chi(t) dt \right] \\
= E \left[ \int_0^\infty e^{-\alpha t} E \left[ \sum_{j \in \mathbb{N}} (1 - \gamma(j, \omega_0))1_{\mathbb{R}^+} (t - \tau(j, \omega_0))1_{[1,\infty)}(\tilde{Z}_i(t) \circ S_j) \right] dt \right] \\
= \int_0^\infty e^{-\alpha t} \int_0^t \int_S \tilde{\mu}(r, ds' \times du) P(\tilde{Z}_{i,t-u} > 0|\rho_0 = s') dt \\
= \int_0^\infty e^{-\alpha t} \int_0^t \int_S \tilde{\mu}(r, ds' \times du) (1 - q_{s',i,0}(t-u)) dt \\
= \hat{\mu}_\alpha(r, S)(1 - \hat{q}(S, i, \alpha))
\end{equation*}

where \( \hat{q}(s', i, \alpha) = \int_0^t e^{-\alpha u} q_{s',i,0}(u) du \) is the Laplace transform.

Also, since Jagers shows characteristics do not necessarily need to be individual for convergence in \( L^1 \) but must have left and right limits, the results from Theorem 12 hold for this characteristic as well [68].

**Lemma 7.**

\[
\frac{K_i(t)}{e^{\alpha t}} \rightarrow \frac{E_\pi[\hat{\chi}(\alpha)]}{\alpha \beta} w
\]

in \( L^1[P_s] \) for \( \pi \)-almost all \( s \in S \) as \( t \to \infty \) where \( E_\pi[w] = h(s) \) and \( E_\pi[\hat{\chi}(\alpha)] = \hat{\mu}_\alpha(s, S)(1 - \hat{q}_n(S, i, \alpha)) \).

**Proof.** Same as proof for Theorem 12. \( \square \)

More useful is the ratio result that states the branching process with respect to a characteristic normalized by all births converges almost surely on the nonextinction set.

**Lemma 8.** \( \frac{K_i(t)}{Y(t)} \rightarrow E_\pi[\hat{\chi}(\alpha)] \) a.s. on \( \{ Y(t) \to \infty \} \).

With this, we can use long run approximations for allele frequencies by taking advantage of the almost sure convergence.
Corollary 3.

$$\frac{Z^\chi(t)}{Z^{\chi'}(t)} \rightarrow \frac{E_\pi[\bar{\chi}(\alpha)]}{E_\pi[\bar{\chi}'(\alpha)]}$$

as \(t \to \infty\) a.s. on \(\{Y(t) \to \infty\}\).

Applications for this can be used when considering the variant allele frequencies from sequencing data, where we might have some proportion of alleles present at a time, \(t\). In such a case, we could be interested in determining the proportion of alleles of a specific mutation set with respect to the marginals for a specific mutation in that set. Another useful aspect about this result is we are already conditioning on nonextinction. Allele frequency data requires this assumption upon collection.

**Frequency spectrum**

A quick extension to the number of living labels of a type \(i\) is to consider the frequency spectrum. For the general branching process, we redefine the frequency spectrum slightly to consider a possible range of numbers for descendants. This allows us to consider tail probabilities of the process, or potentially group a number of types. If the type space were continuous, we would need to consider transitions on the space with respect to the probability measure instead of treating it like the product of transition matrices. By relaxing our assumptions on the frequency spectrum, we’re allowing for more general, and complicated models to be built from this. Define the process \(\alpha_{i,\Gamma}(t)\) as the number of \(i\)-type labels represented by \(j\) descendants, \(j \in \Gamma\), at time \(t\) without including the ancestor label. Note that \(\Gamma = \mathbb{Z}^+\) gives the total number of labels alive at \(t\), or \(K_i(t)\). Also, \(\Gamma = \{j\}\) gives the frequency spectrum as described in Griffiths and Pakes [6] and McDonald and Kimmel [58]. Our generalization might lead to an easier extension for tail probabilities of long-run frequencies. This process is written using the characteristic
\[ \chi'_{\Gamma, x}(u) = 1 \text{ (if } x \text{ has a new label with } j, j \in \Gamma \text{ living type-} i \text{ descendents with the same label at age } u \geq 0) \]

which suffers from the same issue that \( \chi_x \neq \chi \circ S_x \). Like before, a quick rewrite of the characteristic is in order, so

\[ \chi_x(\Gamma, a) = \sum_{j \in \mathbb{N}} (1 - \gamma(j, \omega_x))1_{\mathbb{R}^+}(a - \tau(j, \omega_x))1_{\Gamma}(\bar{Z}_i(a) \circ S_{xj}). \]

Our interest here lies in \( \phi_{i, \Gamma}(t) = E[\alpha_{i, \Gamma}(t)] \). Expectation convergences should hold and should be similar to the results from the general process with different probability values, \( \hat{q}(s, i, j, \alpha) = \int_0^\infty e^{-\alpha u}q_{s, i, j}(u)du \) in place of \( 1 - \hat{q}(s, i, \alpha) \). This might become interesting when we consider long-run frequencies of labels using the ratio convergence results. That is, we can consider the characteristics \( \chi_x(\Gamma, a) \) and \( \chi_x(\mathbb{Z}^+, a) \) look at long term proportions from the asymptotic results of

\[ \frac{Z^\pi(t)}{K_i(t)} \]

on the set of nonextinction \( \{Y(t) \rightarrow \infty\} \).

**Corollary 4.**

\[ \frac{Z^\pi(t)}{K_i(t)} \rightarrow \frac{E_x[\hat{\chi}(\alpha)]}{E_x[\hat{\chi}'(\alpha)]} \]

as \( t \rightarrow \infty \) a.s. on \( \{Y(t) \rightarrow \infty\} \) where \( E_x[\hat{\chi}(\alpha)] = \hat{\mu}_\alpha(r, S)\hat{q}(S, i, \Gamma, \alpha) \) and \( E_x[\hat{\chi}'(\alpha)] = \hat{\mu}_\alpha(r, S)(1 - \hat{q}(S, i, \alpha)) \).
Example: Bellman-Harris Process

We consider a $k$-type Bellman-Harris process where individuals split upon death according to a given probability function. Upon splitting, an individual may initiate a new label or retain the label of her $i$-type mother with probability $\mu_i$. For an individual $x \in I$ with life history $\omega$, we define the random variable $\lambda_\omega$ as the lifetime of the individual and $B(\omega) = (B_1(\omega), \ldots, B_k(\omega))$ as the number of offspring born of each type. Let us also suppose the lifetime and number of daughters are independent of each other and depend only on the type of the individual. Then the birth process is $\xi(t, \omega) = (\xi_1(t, \omega), \ldots, \xi_k(t, \omega))$. Note that in terms of our previous definition of the birth process, we have $\xi_i(t, \omega) = \xi([i] \times [0, t], \omega)$. Our birth process then is

$$\xi(t, \omega) = \begin{cases} 
0 & \text{if } t < \lambda_\omega \\
B(\omega) & \text{if } t \geq \lambda_\omega
\end{cases}$$

If we suppose $x$ is type $i$ (or using the notation above, $\rho_\omega = i$), then let $\lambda_{\omega x}$ have the distribution function $G_i(t)$ and define $m_{ij} = E[B_j(\omega_x)]$. Then the reproduction kernel is

$$m_{ij}(t) \equiv \mu(i, \{j\} \times [0, t]) = E_i[\xi_j(t, \omega)] = m_{ij}G_i(t).$$

As before, we define $\gamma_x$ as the label of the individual $x$ where

$$\gamma(i, \omega_x) = \gamma_{xi} = I(\text{the } i^{th} \text{ offspring of } x \text{ has the same label}).$$

Then, upon the death and splitting of $x$, all $B(\omega_x)$ offspring have the same label independently with probability $\mu_i$ (assuming $\rho_\omega = i$). Otherwise they initiate a new label. We distinguish the reproduction processes separately by

$$\xi(t, \omega) = \tilde{\xi}(t, \omega) + \tilde{\xi}(t, \omega)$$
where \( \tilde{\xi}(t, \omega) = \sum_{i=1}^{B(\omega)} \gamma(i, \omega)I(t \geq \lambda_\omega) \) is the ancestor label reproduction process. Likewise, the reproduction kernel can be written as the sum of ancestor and new labelled individuals,

\[
m_{ij}(t) = m_{ij(m)}(t) + m_{ij(n)}(t) = \mu_i m_{ij} G_i(t) + (1 - \mu_i)m_{ij} G_i(t).
\]

For a vector \( \epsilon = \epsilon_1, \ldots, \epsilon_k \), let \( D_\epsilon \equiv \text{diag} (\epsilon) \). Define the matrix \( H(\alpha) \)

\[
H(\alpha) = (\hat{\mu}_\alpha(i, \{j\})) = \int_{\mathbb{R}^+} e^{-\alpha t} m_{ij}(dt).
\]

and

\[
H_m(\alpha) = D_\mu H(\alpha)
\]

with \( i, j = 1, 2, \ldots, k \). Then the Malthusian parameter, \( \alpha \), is the value such that \( H(\alpha) \) has spectral radius equal to one. If \( \alpha \) exists, then there also exist left and right eigenvectors \( \eta \) and \( \zeta \) of \( H(\alpha) \) such that \( \eta H(\alpha) = \eta \) and \( H(\alpha)\zeta = \zeta \). We normalize the vectors such that \( \eta \zeta = 1 \) and \( \zeta 1 = 1 \). The eigenfunctions, \( \pi(s) \) and \( h(s) \) are the \( s^{th} \) components of the eigenvectors \( \eta \) and \( \zeta \) to link the theory above to this example, and \( v_j = \eta_j(1 - \hat{G}_j(\lambda)) \)[63]. We also have that

\[
\beta = \int t e^{-\alpha t} h(j) \mu(i, dj) \times dt) \pi(di)
\]

\[
= \int_{\mathbb{R}^+} t e^{-\alpha t} \eta M D G_i(dt) \zeta
\]

\[
= \eta \int_{\mathbb{R}^+} t e^{-\alpha t} M D G_i(dt) \zeta.
\]

If we consider the process that counts the number of labels currently alive with type \( i \), \( K_i(t) \), we use the characteristic from section 4. Recall the expectation of
the Laplace for the characteristic at $\alpha$ and reduce it for the Bellman-Harris case as follows:

\[ E_s[\hat{\chi}(\alpha)] = \int_S \hat{\mu}_m(s, ds') (1 - \hat{q}(s', i, \alpha)) \]
\[ = e_s H_m(\alpha)(1 - \hat{q}(\cdot, i, \alpha)) \]
\[ = e_s D_\mu H(\alpha)(1 - \hat{q}(\cdot, i, \alpha)) \]

and

\[ E_\pi[\hat{\chi}(\alpha)] = \int_S E_s[\hat{\chi}(\alpha)] \pi(ds) \]
\[ = \eta D_\mu H(\alpha)(1 - \hat{q}(\cdot, i, \alpha)). \]

We then see the asymptotic results unfold:

\[ \frac{K_i(t)}{e^{\alpha t}} \rightarrow \frac{\eta D_\mu H(\alpha)(1 - \hat{q}(\cdot, i, \alpha))}{\alpha \beta} w \]

in $L^1$ where $E[w] = \zeta_i$. This is very closely related to the results from the Galton-Watson [58] and continuous-time Markov case.

3.5 Proof of Concept Examples

We show the above results and possible extensions using two different birth-death-mutation processes. We show convergence in mean as well as illustrate the convergence of each path (almost sure) conditioning on nonextinction. First, we consider a 4-type irreducible branching process with the same lifetimes as an application of the Markov case. We were unable to find almost sure convergence results in the Markov case without the aid of general branching processes, but the simulations show it still
In our first example, we consider a 4-type process with lifetime rates of 1 for all types, and the offspring p.g.f.

\[
f_1(s) = 0.45 + 0.05s_1s_2 + 0.04s_1s_3 + 0.46s_1^2 \\
f_2(s) = 0.51 + 0.06s_1s_2 + 0.04s_2s_3 + 0.39s_2^2 \\
f_3(s) = 0.56 + 0.04s_2s_3 + 0.05s_3s_4 + 0.35s_3^2 \\
f_4(s) = 0.50 + 0.03s_2s_4 + 0.07s_3s_4 + 0.40s_4^2
\]

which has mean matrix

\[
M = \begin{bmatrix}
1.01 & 0.05 & 0.04 & 0 \\
0.06 & 0.88 & 0.04 & 0 \\
0 & 0.04 & 0.79 & 0.05 \\
0 & 0.03 & 0.07 & 0.9
\end{bmatrix}
\]

giving the generator

\[
A = M - I = \begin{bmatrix}
0.01 & 0.05 & 0.04 & 0 \\
0.06 & -0.12 & 0.04 & 0 \\
0 & 0.04 & -0.21 & 0.05 \\
0 & 0.03 & 0.07 & -0.1
\end{bmatrix}
\]

The mean matrix \(e^{At}\) has spectral radius equal to \(e^{\lambda t}\) where \(\lambda\) is the spectral radius of \(A\) which is \(\lambda = 0.0345\). Since \(\lambda > 0\), the process is supercritical. We have associated left and right normalized eigenvectors

\[
v = \begin{bmatrix} 1.368 & 0.5596 & 0.3529 & 0.1312 \end{bmatrix} \quad \text{and} \quad u = \begin{bmatrix} 0.6051 & 0.2501 & 0.0585 & 0.0862 \end{bmatrix}^T
\]
We show the results of 100 simulations of a process begun by a single individual in Figure 3.7. Convergence in the mean happens fairly quickly, but the 3\textsuperscript{rd} type shows a small deviation at the end which could be due to a few paths jumping near the end. More simulations should smooth this out. The black line represents the mean and the blue line are margins for 2 standard deviations on each side of the mean. The thinner standard deviations suggest convergence of paths to a point as given by the almost sure convergence result from general branching processes. Running the process for a longer period of time would show the paths converging, or beginning the process with a different number of ancestors in each type may show this as well. Unfortunately, a numerical solution becomes intractable since we are faced with the 4-type system of Riccati equations as we showed above.

While a numerical solution for the 4-type process is intractable, Antal and Krapivsky \cite{70} found a closed form solution for a 2-type reducible process in terms of hypergeometric functions. While our results are proved for the irreducible case, the reducible case models cancer growth and clonal evolution better since we want irreversible mutations. Thus we present simulations of a 2-type process following the scheme with a possible solution \cite{70}. We assume a two-type process where type 1 can split into two of its own, one of its own and one mutant, or die. The second type can divide into its own type, or die. We make a small change and allow different growth rates. Type 1 individuals live for an exponential time with rate $a_1 = 1.2$, while type 2 individuals have rate $a_2 = 1$. The p.g.f. for this process is

\begin{align*}
  f_1(s) &= 0.45 + 0.05s_1s_20.5s_1^2 \\
  f_2(s) &= 0.6 + 0.4s_2^2
\end{align*}
Figure 3.7: The paths of the process $K(n)$ for 100 simulations shows convergence of the mean.

and the mean offspring matrix for a division is

$$M = \begin{bmatrix} 1.05 & 0.05 \\ 0 & 0.8 \end{bmatrix}.$$
The generator for this matrix is

\[
A = \begin{bmatrix}
1.2 & 0 \\
0 & 1 \\
\end{bmatrix}
\begin{bmatrix}
1.05 & 0.05 \\
0 & 0.8 \\
\end{bmatrix}
- I_4
= \begin{bmatrix}
0.06 & 0.06 \\
0 & -0.2 \\
\end{bmatrix}.
\]

Thus the spectral radius is \( \rho = 0.06 \) for \( A \), or the process grows according to the spectral radius of \( e^{At} e^{0.06t} \). The left and right eigenvectors have a different form for the reducible case. Like in the discrete time case, our eigenvectors have the form

\[
v = \begin{bmatrix} v_1 & v_1 A_{1,2} (\rho I - A_2)^{-1} \end{bmatrix}
\quad \text{and} \quad u = \begin{bmatrix} u_1 & 0 \end{bmatrix}^T.
\]

Where we break the matrices up into their irreducible classes. The subscripts refer to the classes used. In this simple case, classes consist of only a single type, so the submatrices are really just elements and we get the eigenvectors

\[
v = \begin{bmatrix} 1 & 0.2308 \end{bmatrix}
\quad \text{and} \quad u = \begin{bmatrix} 1 & 0 \end{bmatrix}^T.
\]

In this situation, normalizing the number of labels at time \( t \), \( K(t) \) by \( Z(t)u \) is simply dividing by the number of type 1 individuals at time \( t \).

We show the results of the 2-type process in Figure 3.8 where we’ve conditioned on nonextinction. The same pattern occurs where the mean converges in the reducible case as in the irreducible case. In fact, this simulation is even more convincing, but makes sense because the first type follows a single type process. Convergence for the single type process is given in Pakes [7]. We have solutions for the p.g.f. for the number of individuals, \( Z(t) \) in this particular process. We can adjust this p.g.f. solution to determine \( A(s; t) \), the ancestor process that is used to determine the limit shown in the simulations. Even in a simple two-type case, we have a solution that consists of multiple hypergeometric functions that need to be integrated, requiring
quadrature methods or further reductions of the solution given.

![100 simulations showing convergence of $K_i/Z \cdot u$](image)

**Figure 3.8**: The paths of the process $K(n)$ for 100 simulations shows convergence of the mean.

### 3.6 Applications to Cancer Evolution

Our previous paper discussed the application to cancer evolution when considering a Galton-Watson branching process as the underlying model\[58\]. However, there are issues that occur under this scenario that would make modeling impossible outside the artificial framework of laboratory of cell lines. Tumor cells grow at different rates,
and generations are overlapping, so the Galton-Watson model provides a simplified version of the truth that can not be tested in vivo. We present two new models because the next step is to consider Markov lifetimes which can allow for easier analytical solutions. Further, we relax the assumptions to allow a general lifetime distribution. Our results show that the number of passenger alleles grows with the size of the tumor regardless of the lifetime distribution, at least in expectation.

These results allow to determine tumor age and clonal history of a tumor by considering only the allele frequencies based on variant allele frequency data and estimating growth rates and mutation probabilities based on tumor size. If we assume the probability of accumulating passenger mutations (labels) does not change over time, we should be able to treat the number of passenger mutations as a surrogate for time itself. This can even be done without the assumption of a single baseline probability and instead with a type dependent probability, $\mu_i$ according to our general process set up. We should then be able to infer the history of the tumor, its age, and its composition by using the allele frequencies for driver mutations and the number of passenger mutations at a single point in time, or multiple time points. We will see these estimation methods implemented in our application to inferring tumor growth in the following chapters.
4.1 Introduction

The previous chapters introduced multitype branching process models with infinitely many alleles. We use this model as a basis for determining population dynamics of tumor growth. We assume two types of mutations: driver and passenger mutations. A set of driver mutations will be associated with a particular type in terms of the branching process. That is, cells with a specific genome make up a type different from those of another genome that varies on the set of drivers. However, passenger mutations will be treated differently. We make the infinite sites assumption \cite{60} that states that any passenger mutation that occurs in the process has not been seen until the current time. That is, each passenger mutation is never repeated and unique only to the cell that experienced the event and all its descendants. Since we’ve discussed how passenger mutations are much more common than driver mutations \cite{19}, this is not an unreasonable assumption to make.

There is a distinction that needs to be made between mutations and alleles that could affect our model under certain subsets of the parameter space. We assume there are infinitely many locations for passenger mutations to occur according to the
infinite sites model. If we consider cell types based on a set of passenger mutations, then there are even more possible genomes. Since there are not necessarily an infinite number of mutations, this assumption holds as long as the mutation rate is small enough, or we have enough sites. With a large enough mutation rate, we could eventually see mutations at the same site at two different time points. Under further conditions we make a stronger link between the alleles and passenger mutations. If we assume mutations are not reversible, then we could map every new mutation to the genome of the cell that gains it. In this way, a model that considers the number of alleles should be very similar to one that counts the number of passenger mutations. Passenger mutations will make up labels in the branching process model. However, the model makes a few more assumptions that we’ll need to work with in order to obtain reasonable results. We will list the assumptions necessary, and spend time discussing how we approach satisfying these assumptions.

4.2 Modeling the Number of Driver and Passenger Mutations

We first consider the situation that has been addressed previously, where drivers are considered indistinguishable and have the same effect on growth. Bozic et al. (2010) create a k-type Galton-Watson reducible branching process where each type represents the number of driver mutations accumulated in the tumor [26]. This model assumes the evolution of the tumor is linear, or that each new mutation in the population comes from the previous clone and initiates a wave of clonal expansion. This pattern neglects any possible branching evolution to make a simpler mode. The assumption seems valid if we assume the final tumor consists mainly of mutations that only increase in fitness as they are added. To ensure this, the study defines the selective
advantage, $s$, so that the probability that a type $k$ cell (a cell with $k$ driver mutations) dies is $d_k = 1/2(1 - s)^k$, and the probability of splitting at the end of a generation is $b_k = 1 - d_k$. Further, the probability that a driver mutation occurs is $u$, so an individual asymmetrically splits with probability $ub_k$. The resulting p.g.f. for a type $i$ individual is

$$f_i(s) = d_i + ub_is_is_{i+1} + (1 - u)b_is_i^2$$

with mean matrix

$$M = \begin{bmatrix}
(2 - u)b_1 & ub_1 & 0 & \ldots & 0 \\
0 & (2 - u)b_1 & ub_1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 2b_k
\end{bmatrix}.$$

The authors of the paper approximate the rate of a driver mutation occurring is $3.4 \times 10^{-5}$ for each cell division by assuming the rate of a point mutation is $5 \times 10^{-10}$ per base pair per division and approximately 34,000 positions where driver mutations can occur. Also, the probability of a passenger mutation occurring is approximated to $3.15 \times 5 \times 10^{-3}$. Based on the data, the authors also estimate $s$ to be 0.4%.

The first type in this situation is the tumor initiating cell which contains a single mutation, so the process is supercritical since $d_1 = 1/2(1 - s)^k = 1/2(1 - 0.004)^1 = 1/2(0.996) = 0.498$. However, the authors show simulated runs of the process and show in each situation the number of cells increases in all 6 runs. The probability that a process with a single type goes extinct is the root of the equation $s = (1 - u)b_1s^2 + ub_1s + d_1$ which has a solution at 0.9964, so around 4 out of every 1000 runs should survive. Thus, we assume the results from the study are conditioned on nonextinction. The assumption makes sense considering we only see patients with tumors currently present (i.e. the tumor does not die on its own). In order to
determine the average number of simulations, rather than use a model similar to our infinite-allele model, the study states that any living cell in generation $n$ should have gone through $n$ divisions, so the number of passenger mutations for that individual cell should be binomial distributed with parameters $n$ and $v$, where $v$ is the probability of a passenger mutation occurring. This probability is for a single cell living at generation $n$, and does not consider the total number of passenger mutations in the population at time $n$. For instance, if we consider a process that goes for 2 generations leading to 4 total cells, a single cell in generation 2 can have up to 2 mutations. This is the same for the other 3 cells of the same generation, but we have to get rid of double-counted cells. Thus, we have up to 6 possible chances for a mutation to occur, so the probability of seeing $x$ passenger mutations in the population would be

$$\binom{6}{x} v^x (1 - v)^{n-x}$$

as opposed to the

$$\binom{2}{x} v^x (1 - v)^{n-x}$$

mutations that could exist in a single cell after two generations of splitting. To determine the proper number of passenger mutations associated with the next driver then, we would need to consider the union of all cells alive at $n$ of a particular type, not including the ancestors.

To further illustrate this point, we can count the number of labels or mutations present in the realization of a single-type process given in Figure 4.1. This realization includes red circles where a passenger mutation occurred during splitting. If we look at the cell labeled $t5$, we see its history includes 3 mutations in the 5 splits of its ancestors. Since it went through 5 splits, there are a maximum of 5 possible mutations present. The Bozic model states that the number of mutations present
in the 5\textsuperscript{th} generation should have a binomial distribution with \( n = 5 \), so a maximum of 5 mutations are allowed. This would be true if our sample contained only cells that had the same history as \( t5 \). It also oversimplifies the model to assume independent histories and multiple the binomial random variable by the number of individuals alive at time \( t \), since this would double count mutations shared between two individuals with a common ancestor. Instead, we must count among all individuals the way we propose in our model. The tree contains 12 splits producing 24 total chances for a mutation to occur in an offspring. If we consider the single-type version of my process, the number of different alleles at the final time would be 9, the same as the number of surviving mutations and the normal cell types. We could separate the final nodes into sets based on the distinct neutral alleles. The sets would be \( \{t13\}, \{t12\}, \{t8\}, \{t7, t4\}, \{t6\}, \{t5\}, \{t3\}, \{t2\}, \{t1, t9, t10\} \). That is, this model produces 8 subclones, which our method would accurately count.

From a branching processes perspective, our multitype infinite-allele model allows us to count the number of mutations that occurred in the whole population. It also accounts for lineages that die off, where the alleles are no longer present in the population. We consider the number of passengers given for a particular set of drivers in the study used. In order to correlate the number of passengers and drivers, we could pick a time point and look at the number of alleles associated with a particular type. The study above uses the mean time when the \( k\textsuperscript{th} \) driver appeared to determine the number of passengers we should find in a population with \( k \) drivers. Using

\[ \tau_k = \frac{1}{ks} \log \frac{2ks}{u} \]

from [26] to determine the mean time (in generations) to the birth of the first \( k\textsuperscript{th} \) driver type, and our equation for the mean number of alleles that are type \( k \) at


Figure 4.1: A single type branching process can be used as an example for counting passenger mutations in the different models.

generation $n$ in our model,

$$E_1[K_k(n)] = 1 - H_1^{(n)}(1 - e_j) + \mu \sum_{r=0}^{n-1} e_1^T M^{n-r}(1 - H^{(r)}(1 - e_j)).$$

Table 4.1 and 4.2 show the number of passengers predicted with respect to $\tau_k$, the time for the first cell with $k - 1$ drivers to produce a surviving cell with $k^{th}$ drivers. The probability of seeing a passenger mutation in any daughter cell after a split determined from the model is estimated at about 0.0158 per daughter per division.
In the first table, we show the expected number of passenger mutations according to the model of Bozic, et al., at the time of the first $k + 1$ type mutation that does not go extinct, $n(T_k)$. We distinguish this value from similar values in our model. First, we define $K_i(T_k)$ as the number of alleles that appeared in type $k$ cells at the time of the first type $k + 1$ allele. This sum of this should be comparable to the value $n(T_k)$ in the previous model. The difference in the two is we account for multiple lineages that may grow, so passenger mutations do not have to occur in a linear manner, and expansions can arise in a branched manner in passenger mutations as well as drivers.

The very large exponential growth in passenger mutations in Table 4.2 suggests the passenger probability might be too large relative to the data. Alternatively, even though the first type $k + 1$ allele may exist, does not mean there is no more growth in previous types, so those should be considered as well. The value $Y(T_k)$ counts the total number of passenger alleles with respect to all previous clones up until the time point $T_k$. For example, even if a type 4 clone appears and expands at a faster rate than a type 3 does not mean there are no more passenger mutations occurring with respect to the type 3 clones. This may not matter much in early types, but as the number of types increase, this has a drastic effect on the number of alleles, but should more accurately represent real life since previous clones do not stop growing, but instead get overtaken by fitter clones. Early clones can still contribute to the number of mutations, and even the total number of cells. If we consider the exponential growth of the process itself as given by the growth of the mean matrix, $M^n$, these large numbers should not seem too unreasonable as the ratio of alleles to total cells remains low, given by $Y(T_k)/E[|Z(T_k)|]$. For instance, the expected number of cells at 2134 generations is $E[|Z(n)|] \approx 5.1 \times 10^3$. Thus $E[Y_2(n)]/E[|Z(n)|] \approx 0.09$. This value should reach an asymptote similar to a previously described theorem. Thus the number of passenger alleles in the process is held at bay relative to the total number
of cells. The growth, which seems to fast, is dictated in the model by the selective advantage, $s$, as well as the probability of a new driver and new passenger during division.

$$T_k E_1[Z(T_k)] = \frac{n(T_k)}{E_1[Z(T_k)]}$$

Table 4.1: The expected number of passenger mutations in the model at the expected time of the introduction of the $k^{th}$ driver mutation from Bozic et al [26].

<table>
<thead>
<tr>
<th>$k$</th>
<th>$T_k$</th>
<th>$n(T_k)$</th>
<th>$E_1[Z(T_k)]$</th>
<th>$\frac{n(T_k)}{E_1[Z(T_k)]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1365</td>
<td>21.50</td>
<td>222</td>
<td>0.10</td>
</tr>
<tr>
<td>2</td>
<td>2134</td>
<td>33.61</td>
<td>$5.1 \times 10^3$</td>
<td>0.006</td>
</tr>
<tr>
<td>3</td>
<td>2681</td>
<td>42.23</td>
<td>$7.48 \times 10^4$</td>
<td>$5.6 \times 10^{-4}$</td>
</tr>
<tr>
<td>4</td>
<td>3109</td>
<td>48.97</td>
<td>$1.32 \times 10^6$</td>
<td>$3.7 \times 10^{-5}$</td>
</tr>
<tr>
<td>5</td>
<td>3463</td>
<td>54.54</td>
<td>$2.44 \times 10^7$</td>
<td>$2.2 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

Table 4.2: The expected number of passenger mutations in the infinite-allele model at the expected time of the introduction of the $k^{th}$ driver mutation.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$T_k$</th>
<th>$E_1[Z(T_k)]$</th>
<th>$K_k(T_k)$</th>
<th>$\sum_{i=1}^{k} K_i(T_i)$</th>
<th>$Y(T_k)$</th>
<th>$\frac{Y(T_k)}{E_1[Z(T_k)]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1365</td>
<td>222</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>0.099</td>
</tr>
<tr>
<td>2</td>
<td>2134</td>
<td>$5.1 \times 10^3$</td>
<td>42.3</td>
<td>64.4</td>
<td>505.6</td>
<td>0.099</td>
</tr>
<tr>
<td>3</td>
<td>2681</td>
<td>$7.48 \times 10^4$</td>
<td>60.5</td>
<td>125</td>
<td>7.3 $\times 10^3$</td>
<td>0.098</td>
</tr>
<tr>
<td>4</td>
<td>3109</td>
<td>$1.32 \times 10^6$</td>
<td>75.8</td>
<td>200.7</td>
<td>1.27 $\times 10^5$</td>
<td>0.096</td>
</tr>
<tr>
<td>5</td>
<td>3463</td>
<td>$2.44 \times 10^7$</td>
<td>90.1</td>
<td>290.7</td>
<td>2.3 $\times 10^6$</td>
<td>0.094</td>
</tr>
</tbody>
</table>

Table 4.1 show the results from the model from Bozic et al. [26]. The issue that we see is with the expected number of mutations relative to the number of cells that should be alive at each time. This value decreases rapidly, suggesting that the number of passenger mutations slows down relative to the number of cells. Results from using our infinite alleles model in Table 4.2 shows the expected number of passenger mutations grows exponentially, but the number relative to the number of cells converges. The problem with making the models consistent with each other may come from incorrect values for the probability of a mutation (driver or passenger), or the given selective advantage conferred by a new driver mutation.

Our two different approaches to this looks at the number of expected alleles in
each clone on its own and its cumulative number, $K_k(T_k)$ and the number if previous clones keep growing as a means to connect the previous model to ours, and ours to data. This number grows with respect to the diagonal values of $M$, so $K_2(T_2) = c$ are the number of alleles that exist in type 2 clones at the time when the first surviving type 3 cell is born. We do not include the growth of type 1 cells in this. We also count the cumulative number of mutations from previous clones. The other way we count, $Y_k(T_k)$ does not assume the previous types stop growing once a new type enters the population, so we see growth from new types along with the previous types. This leads to exponential growth that is much faster since the previous types, 1, 2, \ldots, $k - 1$ will be being to grow at their respective times, $T_1, T_2, \ldots, T_{k-1}$, but then continue growing for $T_k, T_{k+1}, \ldots$.

An easier way to compare our results is to look at the function for the expected number of alleles. Bozic et al. assume a constant rate of accumulation of passenger mutations, so that the number in generation $n$ should be $n(t) = vt$ where $t$ is the number of generations. Ours is given as

$$E_1[K_k(n)] = 1 - H_1^{(n)}(1 - e_j) + \mu \sum_{r=0}^{n-1} e_i^T M^{n-r} (1 - H^{(r)}(1 - e_j)).$$

The results from each of these curves is given in Figure 4.2. The infinite allele model begins with a single type 1 allele based on the ancestor before seeing a large drop as the number of individuals will increase relative to the number of expected alleles. In this situation, the equation is dominated by the first summand, described as the probability of nonextinction of the ancestor. However, eventually the second summand of the equation dominates, and we see growth that should begin to look exponential as given by the asymptotic theorems.
We look at data itself to determine the rate of passenger mutations with respect to the number of driver mutations. The algorithm, CHASM, is used by Bozic et al. to determine which mutations are considered drivers and passengers. CHASM is an algorithm that uses random forests to prioritize somatic missense mutations which could be considered drivers. The method is trained on a known set of mutations (drivers and passengers)\cite{71}\cite{72}. We show the number of driver and passenger mutations predicted by CHASM with data for glioblastoma from The Cancer Genome Atlas in Figure 4.3. In this case, we notice that the trend appears linear with respect to the number of drivers predicted by CHASM and number of passengers. We also use Polyphen 2 which predicts the functionality of a mutation where high scores are considered damaging \cite{73}.
Figure 4.3: The number of passenger mutations associated with drivers in Glioblastoma (given by a CHASM FDR of $\leq 0.2$ or Polyphen2 score of 1) shows an approximately linear increase. Using Polyphen 2 results in a similar pattern.
However Figure 4.4 shows the same plot for lung adenocarcinoma using the same
cutoffs for CHASM and Polyphen2, and we see an exponential increase in the number
similar to what we would predict from the infinite-alleles model. We used 542 patients
from the TCGA Lung Adenocarcinoma dataset consisting of 192018 unique variants.
Within this data, we took a curated set of known mutations from various studies[74]
which were experimentally verified. The exponential growth with respect to the
number of passengers shares a form similar to our model, but we get a different result
if we consider only the curated set of mutations. Figure 4.5 shows the number of
variants associated with drivers where we only consider drivers from the curated set
that were experimentally verified. In this case, there is no increase in the number of
passengers as the number of drivers increases. This goes against our model as well as
the model results from Bozic et al. We also note that our model and that from Bozic
et al. give the number of passenger mutations at a specific time when the $k + 1^{st}$
driver appears while this data could be from anytime between the $k^{th}$ and $k + 1^{st}$
drivers since it comes from patients at a time of surgery for a tumor.
Figure 4.4: The number of passenger mutations associated with drivers (given by a CHASM FDR of $\leq 0.2$ or Polyphen score of 1) shows an increase according to a convex function and linear increase until we get to the tumors with a large number of drivers.
Figure 4.5: The number of passenger mutations associated with drivers where the driver mutations are only variants that were experimentally verified.

The expected growth of the alleles during this time should be exponential. Curves were calculated for this and are given in Figure 4.6. One area of consideration is it seems the number of alleles begins to growth before the expected appearance of the first mutation for each type as calculated by the study. Their approximation seems to be conditioned on nonextinction of each particular type, which could account for the difference. However, this should also affect the number of passengers. My work from the previous sections shows the asymptotics in expectation and almost surely conditioned on nonextinction for a supercritical process should both approach the same value, so the differences here should not be as large as they are calculated.
Figure 4.6: The number of passenger mutations increases exponentially in time. The data is transformed on the right with a $\log_{10}$ transformation since the larger types blow up quickly.

### 4.3 Estimating Cancer Progression

Instead of assuming all mutations have the same effect, we assume driver mutations are unique, and the effect of including a new mutation is different for each mutation. For instance, tumor suppressors may cause the probability of dying instead of splitting to be lower. Other mutations may lead to shorter lifetimes instead. This departs from the previous model by assuming an evolution that involves the specific genomes
of the tumor at any time. We use the multitype infinite allele model to determine the growth rate of subclones to determine which set of driver mutations are responsible for different rates of growth. The data we have access to comes from TCGA, where we have a list of single nucleotide variants and their associated frequencies for each patient. Variant allele frequencies are similar to marginal distributions for each of the variants, so we cannot directly use these probabilities as data for our branching process. Instead, we need to determine the joint probabilities associated with all variants together. We do this by reconstructing the phylogeny of variants to determine an ordering and which groups of mutations evolved together or as descendants. This method can be used for all variants (passengers and drivers), but we need to distinguish between the two before inputting them into our model as well, so we should do this before estimating the phylogeny to make the estimation problem more efficient.

We go through the steps necessary to estimate the rates for each subclone beginning with the data given by TCGA. This involves determining which mutations are drivers and the phylogeny of the mutations using an algorithm such as Phylosub to get the frequencies from each subclonal population. Along with the frequencies, we use information about the size of the tumor in order to estimate the rate of growth of the tumor, corresponding to the spectral radius of the branching process. Since these cells population have to reach a large size for detection, we are able to use our asymptotic results as approximations. We also present likelihood based methods for estimating parameters from a branching process and Approximate Bayesian Computation methods that are used to find distributions for the parameters. This process is shown in Figure 4.7. Since data is in the form of frequencies, most of the work involves getting the data into a form we for use in the branching process. Either we need to present the data in the form of cell counts, or we need to use estimation methods for branching processes based on allele frequencies. We give a method for
each of these directions in this section, but first we explain how to set the data up.

Figure 4.7: The process to estimate the parameters for cancer progression from whole exome or whole genome sequencing involves determining the subtypes present and the overall growth of the population before implementing the branching process model.

**Determining Drivers from Variants**

Sequencing studies in the TCGA data give variant allele frequencies for all variants sequenced by reporting the number of reference and alternate alleles of each variant in each patient. We use data from lung adenocarcinoma which contains around 223,000 variants from 542 patients. This data has a mean number of 411 variants per patient with a standard deviation of about 414. Various methods are used to determine which of these variants are considered driver mutations as opposed to passengers.

First, compiled lists are given based on experimental studies that have determined possible driver variants[74]. We use these driver mutations as our only set of drivers and try to determine an ordering of mutations for a specific patient. Ideally, we want
to have a few patients with a similar set of mutations in order to validate our results. Figure 4.8 presents the absence and presence of proposed driver variants based on starting position in the genome (as opposed to the change in base pairs). We ran a correspondence analysis to reorder the rows and columns to get a matrix that gives blocks in the diagonal. The results should show which patients and variant locations correspond with each other best. Our first problem comes from the fact that no patients have more than 2 proposed drivers as given by this list. Further, while many patients share a single driver mutation, not many seem to share more than one driver. The right side of the figure reduces the patient set down to those who only have 2 or more mutations. We see that about 3 patients share the same set of two variants which are only a single base-pair apart. Otherwise, no two patients share a common set of drivers. We conclude then that either lung adenocarcinoma is initiated and develop from 1-2 driver mutations, or the compiled set of drivers is incomplete. This implies we should consider other ways to determine driver mutations from the set.
Figure 4.8: We consider the presence/absence of known drivers in all patients. A correspondence analysis uses the eigenvectors associated with the largest eigenvalues less than one of a matrix to determine the orderings. The lower panel is only from patients that have 2 or more drivers present.
CHASM uses random forests to determine functionality of variants trained on a set of variants that are known drivers to keep sensitivity high and known passengers to keep specificity high. We used CHASM to find functional drivers in the set of patients that had at least one mutation curated from the above analysis. There were 232 patients used with at least one of the drivers. We kept all mutations that had a false discovery rate q-value of 0.05 or lower as well as the other known drivers in order to create a short list of variants. There were 473 variant locations in this list among all patients and 260 of those were unique, so we still do not see a lot of conserved variants. We ran a correspondence analysis on this set to see if we could identify any clusters. The results are given in Figure 4.9. Again, we do not see many clusters of mutations, although a number of single mutations are shared among patients. This again means that there are a number of possible single nucleotide variants that could lead to cancer, and they are not necessarily always required. It appears based on the clustering that we could identify a subset of ”required” variants where we should expect to see at least one of them in any lung adenocarcinoma.
Figure 4.9: We consider the presence/absence of drivers in all patients with at least one of the mutations given from the curated set and a false discovery rate of 0.05 or lower. A correspondence analysis uses the eigenvectors associated with the largest eigenvalues less than one of a matrix to determine the orderings. The lower panel is patients that have 2 or more drivers present.
An alternative suggestion that we might check would be to look at regions of genes. Because genes can be very large and variants in different locations might have different effects, we should just use genes in place of mutations. However, we might consider regions of base pairs as a single mutation, implying that any mutation in nearby areas that are deemed functional by CHASM may have a similar effect on growth rates.

For Lung Adenocarcinoma, it turns out only a few mutations may be necessary to cause tumorigenesis, so we may only need to consider the effect of 1-3 mutations. Many times these mutations happen in EGFR or KRAS (personal correspondence with Dr. Ivan P. Gorlov at Geisel School of Medicine, Dartmouth College). We assume these mutations compiled from Lovly et al. will serve as our driver mutations and supplement this set with drivers found by CHASM with a false discovery rate of 0.2\cite{74}. That is, we only consider the patients that had at least one driver mutation from the compiled list. Patients can have other drivers as well selected by CHASM, but each mutation must occur more than once in the population. The results from this are given in the second plot of Figure 4.9. Even in this case, we only see once patient with 3 mutations. Many patients have a mutation in KRAS and EGFR, but the patient with 3 mutations has variants in ATM, PIK3CA, and BRAF.

**Estimating the Phylogeny of Clones**

Of the 542 patients, 214 had at least one driver based on the above criteria. 205 patients had one driver mutation, 8 had two mutations, and a single patient had three. Since we are estimating a phylogeny for driver mutations, we could order any set of mutations for most of the patients since they only have two. Instead, we ignore the clustering above for now, and just consider the set of given driver mutations and look at a few patients with more than three of the given drivers in different genes.
Based on this criteria, we have 59 patients who had at least 3 driver mutations with at least one driver that was curated and the others from the curated set or selected by CHASM. There was a mean of 4 driver mutations in this set. Phylosub works for determining the clonal evolution within individual patients, so we ran Phylosub for each patient to produce an expected evolutionary history of variants along with an estimate for the allele frequencies of each type. The 59 patients resulted in five different possible tree structures. Most patients had a linear evolutionary history, but a branching history was estimated in a few patients, suggesting a nonlinear succession of mutations. We give examples of each of the trees in Figure 4.10. All trees only consisted of three possible levels, so the associated branching processes are not too complex.

The results show the evolutionary history of the tumor. The root is the ancestor type in this model. Types that arise from the root contain all variants from the ancestors along with the new ones listed. These variants persist in the population through descendant clones as well. These trees are estimated using the allele frequencies along with some constraints as described in the previous chapters. These constraints ensure that clones with smaller frequencies should be daughters of those with larger frequencies, and if two daughter clones have frequencies with a sum greater than that of a parent, they must be arranged linearly instead of as a branch. One misleading aspect of the pictures is that it appears that the only variants or types we should see at the time of sequencing and detection may be the final leaves of the tree. In fact, we should see all cell types at sequencing, so we could see cells with ancestor variants without the variants in leaves. The numbers below the labels refer to the value $\phi$ according to the Phylosub algorithm. This value is essentially the estimated frequency with respect to a particular variant. In order to convert this to a frequency with respect to a particular clone, we subtract the sum of $\phi$ in all daughters.
As an example, we consider the tree with four nodes that undergoes branching (associated with TCGA-44-2666, or tree number 4 in Figure 4.10). In this case, we begin with the leaves. Leaf 1 has variants in BRAF with a value of $\phi = 0.469$. The other leaf has a variant in ADARB1 and DAPK1 with $\phi = 0.049$. Since these are leaves, the values of $\phi$ refer to the actual clonal frequencies. However, the parent node with unknown variants has $\phi = 0.717$ so the clonal frequency is $0.717 - (0.469 + 0.049) = 0.199$. The ancestor node has a frequency of $1 - 0.717 = 0.283$. Table 4.3 summarizes these results. We use these trees to construct generating functions for associated branching processes.
Figure 4.10: These trees are examples of the five different estimated evolutionary tree topologies resulting from the Phylosub algorithm on the 59 TCGA patients with Lung Adenocarcinoma.
<table>
<thead>
<tr>
<th>Node</th>
<th>$\phi$</th>
<th>Clonal Frequency</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0.283</td>
<td>&lt;ancestor&gt;</td>
</tr>
<tr>
<td>1</td>
<td>0.717</td>
<td>0.199</td>
<td>&lt;none&gt;</td>
</tr>
<tr>
<td>2</td>
<td>0.469</td>
<td>0.469</td>
<td>BRAF</td>
</tr>
<tr>
<td>3</td>
<td>0.049</td>
<td>0.049</td>
<td>ADARB1, DAPK1</td>
</tr>
</tbody>
</table>

Table 4.3: A summary of the frequencies for the specific variants and the clones containing those variants as well as the variants in ancestors.

4.4 Creating a Branching Process Model of Cancer Evolution

The estimated evolutionary histories provide information to construct branching processes. With the above examples, we have 5 possible branching processes where one is 2-type, one is 3-type, two are 4-type, and one is 5-type. We’ll focus on the two 4-type processes since these generalize previous types a bit, but also can be further generalized. For these cases, we have either linear or branched evolution.
Figure 4.11: These trees are two examples of the different estimated evolutionary trees resulting from the Phylosub algorithm given 4 distinct clones (nodes).

Model Assumptions

For these models, we’ll assume cells live for an exponentially distributed lifetime with parameter $a$ (mean $a^{-1}$) before splitting. A cell of type $i$ may either die or split, and splitting can be into two of the same type $i$, or one type $i$, and one of a different type, $j$, with $j > i$. For these two trees we’ve shown, we get the following possible offspring
generating functions

\[
\begin{align*}
    f_0(s) &= p_{00} + p_{01}s_0s_1 + p_{02}s_0^2 \\
    f_1(s) &= p_{10} + p_{11}s_1(1/2s_2 + 1/2s_3) + p_{12}s_1^2 \\
    f_2(s) &= p_{20} + p_{22}s_2^2 \\
    f_3(s) &= p_{30} + p_{32}s_3^2.
\end{align*}
\]

assuming the probability of a type 1 cell becoming a type 2 and 3 cell is the same, and

\[
\begin{align*}
    f_0(s) &= p_{00} + p_{01}s_0s_1 + p_{02}s_0^2 \\
    f_1(s) &= p_{10} + p_{11}s_1s_2 + p_{12}s_1^2 \\
    f_2(s) &= p_{20} + p_{21}s_2s_3 + p_{22}s_2^2 \\
    f_3(s) &= p_{30} + p_{32}s_3^2.
\end{align*}
\]

The associated mean offspring matrices would be

\[
M_1 = \begin{bmatrix}
    2p_{02} + p_{01} & p_{01} & 0 & 0 \\
    0 & 2p_{12} + p_{11} & \frac{1}{2}p_{11} & \frac{1}{2}p_{11} \\
    0 & 0 & 2p_{22} & 0 \\
    0 & 0 & 0 & 2p_{32}
\end{bmatrix}
\]
and

\[
M_2 = \begin{bmatrix}
2p_{02} + p_{01} & p_{01} & 0 & 0 \\
0 & 2p_{12} + p_{11} & p_{11} & 0 \\
0 & 0 & 2p_{22} + p_{21} & p_{21} \\
0 & 0 & 0 & 2p_{32}
\end{bmatrix}.
\]

Since our processes are Markov and continuous time, both processes have generators \( A = a(M - I_4) \) where \( I_4 \) is the four-dimensional identity matrix. Further, the mean process for all types is \( M(t) = \exp(At) \) where

\[
M(t) = m_{ij}(t) = E[Z_j(t)|Z_i(0) = e_1].
\]

These results were covered in the previous chapters and in Athreya and Ney [2]. If we substitute in \( m_{ij} \) for the \((i, j)\)th element of \( M_1 \) or \( M_2 \) for simplicity, we can express \( M_{0j}(t) \) for any \( j \) as a linear combination of exponential functions,

\[
M_{0j}(t) \propto \sum_{l=0}^{k} C_l e^{\alpha(l)(m_{ll} - 1)}.
\]

A four-type process like the two we are showing can be written out, but this can quickly become more complicated. For the first process we have
\[ M_{00}(t) = e^{ta(m_{00}^{-1})} \]
\[ M_{01}(t) = \frac{m_{01}}{m_{11} - m_{00}} [e^{ta(m_{11}^{-1})} - e^{ta(m_{00}^{-1})}] \]
\[ M_{02}(t) = \frac{m_{01}m_{12}}{(m_{22} - m_{11})(m_{11} - m_{00})(m_{22} - m_{00})} [(m_{11} - m_{00})e^{ta(m_{22}^{-1})} + \]
\[ + (m_{00} - m_{22})e^{ta(m_{11}^{-1})} + (m_{22} - m_{11})e^{ta(m_{00}^{-1})}] \]
\[ M_{03}(t) = \frac{m_{01}m_{13}}{(m_{33} - m_{11})(m_{11} - m_{00})(m_{33} - m_{00})} [(m_{11} - m_{00})e^{ta(m_{33}^{-1})} + \]
\[ + (m_{00} - m_{33})e^{ta(m_{11}^{-1})} + (m_{33} - m_{11})e^{ta(m_{00}^{-1})}] \]

for the first process associated with \( M_{11} \). The mean process associated with the second branching process is the same for \( M_{00}(t) \), \( M_{01}(t) \), and \( M_{02}(t) \). However, \( M_{03}(t) \) has a different form which is much longer since it depends on all previous types.

\[ M_{03}(t) = \frac{m_{01}m_{12}m_{23}}{(m_{33} - m_{22})(m_{22} - m_{11})(m_{33} - m_{11})(m_{33} - m_{00})(m_{22} - m_{00})(m_{11} - m_{00})} \times \]
\[ \times \left( e^{ta(m_{33}^{-1})} \left[ m_{22}^2(m_{11} - m_{00}) + m_{11}^2(m_{00} - m_{22}) + m_{00}^2(m_{22} - m_{11}) \right] + \right. \]
\[ + e^{ta(m_{22}^{-1})} \left[ m_{33}^2(m_{00} - m_{11}) + m_{11}^2(m_{33} - m_{00}) + m_{00}^2(m_{11} - m_{33}) \right] + \]
\[ + e^{ta(m_{11}^{-1})} \left[ m_{33}^2(m_{22} - m_{00}) + m_{22}^2(m_{00} - m_{33}) + m_{00}^2(m_{33} - m_{22}) \right] + \]
\[ + \left. e^{ta(m_{00}^{-1})} \left[ m_{33}^2(m_{11} - m_{22}) + m_{22}^2(m_{33} - m_{11}) + m_{11}^2(m_{22} - m_{33}) \right] \right) \].

The covariance matrix function of a \( k \)-type continuous-time Markov process started by a type 0 individual is the solution to the linear differential equation

\[ \frac{d}{dt} D(t) = A^T D(t) + D(t) A + \sum_{i=1}^{k} m_{0i}(t) B^{(i)} \]
where

\[ D(t) = [d_{ij}(t)] = [E[Z_i(t)Z_j(t)|Z_0(t) = e_1], \]
\[ B^{(i)} = [b_{jk}^{(i)}] \]
\[ b_{jk}^{(i)} = a \left( \frac{\delta^2 f_i(s)}{\delta s_j \delta s_k} \right)_{s=1} + \delta_{jk} - (m_{ij} - \delta_{ij})\delta_{ik} - (m_{ik} - \delta_{ik})\delta_{ij} - \delta_{ij}\delta_{ik} \]

which has the solution

\[ D(t) = M(t)^T M(t) + \int_0^t M(t - \tau)^T (\sum_{i=1}^k m_{0i}(\tau)B^{(i)}) M(t - \tau) d\tau \]

as given by Athreya and Ney [2]. This will become important because we require the first two moments of a process in order find estimates for the frequencies of each clone.

The matrices \( B_{jk}^{(i)} \) for the either process are simplified to the following function

\[ b_{jk} = \begin{cases} 
  a & \text{if } i \neq j = k \\
  2ap_{00} & \text{if } i = j = k \\
  0 & \text{o.w.} 
\end{cases} \]

If we consider the solutions for the first of our two-type processes, where branched evolution occurs, the number of terms for \( D(t) \) becomes very large. Fortunately, these are exponential functions, so a solution exists to the integral where each term has the form of \( \int_0^t e^{C(t-\tau)}d\tau = C^{-1}(e^{Ct} - 1) \), so the final solution has the form \( \sum_i C_i^{-1}(e^{C_i t} - 1) \).

However, we found the expression \( M(t - \tau)^T (\sum_{i=1}^k m_{0i}(\tau)B^{(i)}) M(t - \tau) \) in the 4-type branched evolution process leads to a matrix where entry \( D_{(1,1)}(t) \) has 8 terms and \( D_{(4,4)}(t) \) has over 100 without simplification. Thus, we are unlikely to provide estimates in even these simple cases. Even a 3-type branched evolution process can
result in up to 25 terms for the linear combination of exponential functions.

To further illustrate this, we considered a two-type process with the following p.g.f.

\[ f_1(s_1, s_2) = p_{10} + p_{11}s_1 s_2 + p_{12}s_1^2 \]
\[ f_2(s_1, s_2) = p_{20} + p_{12}s_1^2 \]

and exponentially distributed lifetimes with rate \( a \). Such a process results in a mean matrix of

\[
M(t) = \begin{bmatrix}
    e^{p_{01} + 2p_{02} - 1} & \frac{p_{01}}{p_{01} + 2p_{02} - 2p_{12}} (e^{p_{01} + 2p_{02} - 1} - e^{2p_{12} - 1}) \\
    0 & e^{2p_{12} - 1}
\end{bmatrix}
\]

The resulting expression for

\[ D(t) = M(t)^T M(t) + \int_0^t M(t - \tau)^T (\sum_{i=1}^k m_{0i}(\tau) B^{(i)}) M(t - \tau) d\tau \]

is found by breaking each term into respective parts. First, define the constants:

\[ C_1 \equiv p_{01} + 2p_{02} - 1 \]
\[ C_2 \equiv 2p_{12} - 1 \]
\[ C_3 \equiv p_{01} + 2p_{02} - 2p_{12} = C_1 - C_2. \]

The matrix inside the integral, called \( D^{(2)}(t - \tau) \), is symmetric with the following terms,
\[ [D^{(2)}(t - \tau)]_{(1,1)} = e^{(t-\tau)2C_1} \left( 2ap_{00}e^{\tau C_2} - \frac{p_{01}}{C_3}(e^{\tau C_1} - e^{\tau C_2}) \right) \]

\[ [D^{(2)}(t - \tau)]_{(1,2)} = \frac{p_{01}}{C_3} \left[ e^{(t-\tau)C_1} \left( 2ap_{00}e^{\tau C_1} + \frac{ap_{01}}{C_3}(e^{\tau C_1} - e^{\tau C_2}) \right) \right] \left( e^{(t-\tau)C_1} - e^{(t-\tau)C_2} \right) \]

\[ [D^{(2)}(t - \tau)]_{(2,2)} = \frac{p_{01}^2}{C_3^2} \left[ (e^{(t-\tau)C_1} - e^{(t-\tau)C_2})^2 \left( 2ap_{00}e^{\tau C_1} + \frac{ap_{01}}{C_3}(e^{\tau C_1} - e^{\tau C_2}) \right) \right] + (e^{(t-\tau)C_2})^2 \left( a e^{\tau C_1} + \frac{2ap_{01}p_{10}}{C_3}(e^{\tau C_1} - e^{\tau C_2}) \right) \]

Integrating the first term over \([0,t]\) results in

\[ \int_0^t e^{(t-\tau)2C_1} \left( 2ap_{00}e^{\tau C_2} - \frac{p_{01}}{C_3}(e^{\tau C_1} - e^{\tau C_2}) \right) d\tau = \]

\[ = \left( 2ap_{00} + \frac{ap_{01}}{C_3} \right) e^{t2C_1} \left( \frac{1 - e^{tC_2}}{C_1} \right) - \frac{ap_{01}}{C_3} e^{t2C_1} \left( \frac{e^{(tC_2-C_1)} - 1}{C_2 - 2C_1} \right) . \]

The term \([D^{(2)}(t - \tau)]_{(2,2)}\) integrates to a 24 term expression, which we omit because it becomes too complicated. Thus, we reached a point where trying to find the variances of multitype processes becomes exceedingly difficult. This occurs in only a 2-type process, so other methods should be considered (possibly using simulations). However, assuming we could determine the first two moment functions, methods are available that use these to estimate relative frequencies which we discuss in the following section.

**Likelihood Based Estimation of Clonal Frequencies**

We can attempt to estimate the parameters of the probability generating function using likelihood functions based on the relative frequencies of each type at a point in time, \(t\) [75]. This method is an extension of the work from Yakovlev and Yanev that establishes central limit theorems for relative proportions in branching processes.
The theory is developed for multitype branching processes that can be discrete or continuous, and assumes an ancestor population of size $N$, where $N$ grows.

If we consider the $k$-type branching process $\{Z(t), t \geq 0\}$, then define the total number of cells at time $t$ as

$$U(t) = |Z(t)|$$

and the proportion of type $i$ cells as

$$\Delta_i(t) = \frac{Z_i(t)}{U(t)}$$

don the nonextinction set, $\{U(t) \geq 0\}$. If we consider a situation where we have $N$ ancestors of the same type, then let $Z(t; N)$, $U(t; N)$, and $\delta(t; N)$ be the associated processes with the same meanings as above.

Above we discussed mean and variance of some processes in continuous-time. If we restrict ourselves to type 0 ancestors, then let $m_j(t) = m_{1j}(t)$ be the mean process, $m_\bullet(t) = \sum_j m_j(t)$, and $C_{ij}(t) = Cov(Z_i(t), Z_j(t)) = D_{ij}(t) - m_i(t)m_j(t)$. For simplicity, we call to the diagonal elements of $\sigma^2_j(t) \equiv C_{jj}(t)$. The expected relative proportion of each type is defined as

$$p_i(t) = \frac{m_i(t)}{m_\bullet(t)}.$$

With these definitions, the following Central Limit Theorems are given in Yakovlev and Yanev[75]. First, define $a_{ii}(t) = \sigma_i(t)(1 - p_i(t))$, $a_{ij}(t) = -\sigma_i(t)p_j(t)$, and $r_{ij}(t) = Cor(Z_i(t), Z_j(t)) = C_{ij}(t)/(\sigma_i(t)\sigma_j(t))$.

**Theorem 14.** Assume $\sigma^2_i(t) < \infty$ for all $i = 1, \ldots, k$. Then, if we define the random variable

$$W_i(t; N) = m_\bullet(t)\sqrt{N} \left(\Delta_i(t; N) - p_i(t)\right),$$
we have the following results as $N \to \infty$:

1. $W_i(t; N) \overset{d}{\to} Y_i(t)$, where
   \[ Y_i(t) \sim N(0, S_i^2(t)), \]
   \[ S_i^2(t) = \sum_{j,l=1}^{k} r_{jl}(t)a_{ji}(t)a_{ki}(t). \]

2. For any vector of size $\kappa < k$, let $u(i) \in \{1, \ldots, k\}$,
   \[ (W_{u_1}(t; N), \ldots, W_{u_\kappa}(t; N)) \overset{d}{\to} (Y_{u_1}(t), \ldots, Y_{u_\kappa}(t)), \]
   where
   \[ (Y_{u_1}(t), \ldots, Y_{u_\kappa}(t)) \sim MN(0, \text{Cov}(Y_i(t), Y_j(t))). \]

The covariance matrix is

\[ \text{Cov}(Y_i(t), Y_j(t)) = [a_{ij}(t)]^T[r^{(d)}(t)][a_{ij}(t)] \]

with the indices $i, j$ defined according to the subset from $u_1, \ldots, u_\kappa$.

This result states that for large enough populations, the relative proportion of each type is approximately normally distributed with mean equal to the expected relative frequency, $p_i(t)$ and a variance that can be calculated from the variance functions for the process. Thus, we can come up with estimates for the relative proportion of each type at a time $t_0$ along with confidence intervals for it based on our data. In order to do so, we need to determine more information about the tumor, including its age and size. This information is used to determine $t$ for the relative proportions.
Estimating Tumor Growth Rates and Time

Figure 4.7 shows the process in which we try to estimate parameters of the branching process. We’ve gone through and determined the phylogeny and made a connection between that and estimating relative frequency of the associated branching process. However, we have other variables that are still missing that should be filled in. First, we need to know the age of the tumor at sequencing. We are also interested in the size of the tumor. These two variables should be directly dependent on each other and the Malthusian parameter of the population. That is, we have the following asymptotic result:

$$\lim_{t \to \infty} \frac{Z(t)}{e^{\lambda t}} = W v \text{ a.s.}$$

where $e^{\lambda t}$ is the largest eigenvalue of $M(t) = \exp(At)$, and $v$ is the associated left eigenvector. Note that $E[W] = u_0$, the element of the right eigenvector with the same index as the ancestor.

Yanev presents a method to estimate the unknown age of a single type process given only the size in a process with $n$ ancestors as

$$\hat{t}(n) = \log \frac{Z_t(n)}{Z_0(n)} / \log(m).$$

We propose a similar estimator for the multitype continuous time case,

$$\hat{t} = \frac{\log U(t)}{\lambda}$$

where $U(t)$ is the total number of cells in the tumor[76].

Using the asymptotic result above, we note that $|Z(t)| \sim e^{\lambda t} W |v|$. Since $v$ is the long-run proportion of each type in the process, and $uv = 1$, $E[|Z(t)|] \sim e^{\lambda t} u_0 |v| < e^{\lambda t}$. Thus we can put a lower bound on $t$ of $\log E[|Z(t)|]/\lambda$, so given a tumor size,
<table>
<thead>
<tr>
<th>Stage</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>\leq 3\text{ cm}</td>
</tr>
<tr>
<td>T1a</td>
<td>\leq 2\text{ cm}</td>
</tr>
<tr>
<td>T1b</td>
<td>2 – 3\text{ cm}</td>
</tr>
<tr>
<td>T2</td>
<td>3 – 7\text{ cm}</td>
</tr>
<tr>
<td>T2a</td>
<td>3 – 5\text{ cm}</td>
</tr>
<tr>
<td>T2b</td>
<td>5 – 7\text{ cm}</td>
</tr>
<tr>
<td>T3</td>
<td>&gt; 7\text{ cm or invasion in certain regions}</td>
</tr>
</tbody>
</table>

Table 4.4: The stage of a lung adenocarcinoma primary tumor can be related to a range of sizes.

$U(t)$, our estimator above should be biased. More information about $u_0$ could lead to more certain results.

This then gives us a way to estimate $t$ provided we have estimates of $U(t)$, the total tumor size, and $\lambda$, the growth rate. The total tumor size should not be a large problem, except the information we are given from TCGA is the volume of the sample instead of the tumor. We are also given the stage of the tumor which can be converted to a size, and as a result, a volume and/or cell count.

According to the American Joint Committee on Cancer, the staging of a primary tumor is divided into categories based on the size and other attributes\[77\]. Table 4.4 gives the staging for the primary tumor (T) along with the range of sizes in the greatest dimension. Patient information about cancer staging is used to give an estimate of the tumor size.

In our data, a subset of 59 patients are distributed among the various stages. The distribution of stages and structures of trees is given in Table 4.5. The table shows that larger structures (more nodes) seem to have higher staging associated with them while smaller structures seem do not seem to be associated with stage. The structure does not necessarily say anything about the number of drivers. Instead, it gives an idea of the number of different clones, where one clone may have multiple drivers that arose simultaneously. Higher resolution of the data or more samples could help
us get more information about the evolutionary structure, but we are limited to heterogeneity within individuals patients.

\[
\begin{array}{cccccccc}
\text{nodes} & \text{2 nodes} & \text{3 nodes: linear} & \text{4 nodes: branched} & \text{4 nodes: linear} & \text{5 nodes: branched} \\
\hline
- & 0 & 4 & 0 & 0 & 0 \\
\text{T1} & 1 & 1 & 0 & 0 & 0 \\
\text{T1a} & 0 & 4 & 0 & 0 & 0 \\
\text{T1b} & 2 & 12 & 3 & 1 & 0 \\
\text{T2} & 4 & 5 & 1 & 0 & 1 \\
\text{T2a} & 2 & 3 & 0 & 0 & 0 \\
\text{T2b} & 3 & 1 & 0 & 0 & 0 \\
\text{T3} & 3 & 3 & 0 & 0 & 0 \\
\text{T4} & 4 & 4 & 0 & 0 & 0 \\
\end{array}
\]

Table 4.5: The structure of the evolutionary history based on the estimated trees for each tumor seems to be related to the stages for structures with more nodes. Each node is a clone or subclone estimated using Phylosub.

For the sake of finding a specific number in terms of the size, we can use the midpoint of each stage’s range. We also assume the geometry of each tumor is spherical to estimate the volume. We get \( V = \frac{4}{3}\pi r^3 \) cubic centimeters as the volume of the tumor where \( r \), the radius, is half the midpoint of each stage’s range. We also estimate about \( 10^9 \) cells per gram, or cubic centimeter as is standard. Thus, our estimated sizes and cell count with respect to each stage are:

\[
\begin{array}{cccccccc}
\text{Size (cm)} & \text{T1} & \text{T1a} & \text{T1b} & \text{T2} & \text{T2a} & \text{T2b} & \text{T3} \\
\text{Volume (cm}^3\text{)} & 2 & 1.5 & 2.5 & 5 & 4 & 6 & 7 \\
\text{\( E[\hat{t}] \) (years)} & 4.18 & 1.76 & 8.18 & 65.45 & 33.51 & 113.1 & 179.59 \\
\end{array}
\]

Table 4.6: The size of each tumor is determined by taking the midpoint for the range of sizes in each stage and using this as the diameter. The conversion to a volume and cell count assumes \( 10^9 \) cells per gram.

Lastly, we have to determine how we might estimate the growth rate of a tumor in order to estimate the time. We can look at historical studies for lung adenocarcinoma to determine the doubling time for a tumor. Using this and assuming exponential growth along with finding the tumor doubling size, we can determine the tumor’s growth rate. Given a tumor doubling time for a cancer, \( t_d \), the growth rate is found
Thus, we can estimate the value $\lambda$ from previous studies without using our own data, which requires time. Tumor doubling time varies between patients and can be represented as a random variable with an associated distribution due to differences in growth rates, tumor stage, and other factors. Usuda et al. give a mean doubling time in adenocarcinoma of 223.1 days with a standard deviation of 209.3 days [78]. Alternatively, Kanashiki et al. give a mean doubling time of 177 days [79]. Using this and our estimate for the age, $t$, we can get the expected age of each tumor as

$$E[t] = \frac{\log U(t)}{\lambda} = \frac{E[t_d] \log U(t)}{\log 2}.$$ 

The age of tumors for each stage are given in Table 4.6 as well using the mean doubling time of 177 days. Note that this is a crude estimate to get parameters for our data, and it does not take into account the variability.

Better parameters can be found by collecting data from multiple time points as discussed in a later section, or even making $t_d$ and $\lambda$ a random variable. Doing so would change the results of our branching process theory though. If the spectral radius was considered a random variable, then so should the parameters of the original probability generating function, $f(s)$. This approach may lead to more accurate results since we know that there is variation in each patient in the growth and death rates of cells that goes beyond the type of cells and clones present. However, doing so includes adding additional complexity to a model that already became complex very quickly.

Another direction to estimating is to use the number of passenger mutations as a molecular clock. We showed above that the number of passengers should increase ex-
ponentially in time if we follow the infinite-allele assumption. We have the expression for the growth of the number of alleles for clones of type $i$ as

$$E[K_i(t)] \sim e^{\lambda t} e^{T_{\mu u D_M}} \int_0^t e^{-u\lambda} (1 - A_i(u)) du.$$  

This equation, proven earlier shows the number of alleles grows exponentially proportional to the growth of the population size ($e^{\lambda t}$) and the probability of passenger mutations ($\mu$). This result relies on knowledge of the parameters since the p.g.f., $(1 - A_i(u))$, must be solved numerically. This value is the probability of having no $i$-type cells alive at time $u$ in a process. Because of this, it would be better used for validation, or to estimate $\mu$ given knowledge of the parameters from before.

Last, if we had access to single cell data, we could use the information in a single cell to predict the number of splits, and as a result, parameters from the original distribution. A cell that splits $n$ times should have $X$ passenger mutations where $X \sim \text{Binomial}(n, \mu)$. Since the process is Markov, we should be able to approximate the mean number of generations given the age of the tumor and estimated splitting time.

**Checking the Infinite-alleles assumption**

It is important to consider one of the main assumptions of our infinite-alleles branching process - that every passenger mutation that occurs has not been seen before in the population. We could check this by looking at the frequency of all passenger mutations across tumors. If there are infinitely many sites for a mutation to occur, an individual tumor should not have the same mutation occur twice. Thus, across multiple tumors (or patients), the same assumption should hold. Using the TCGA Lung Adenocarcinoma data set, we consider all mutations (including drivers). There
are 192,017 variants present in this data set that are characterized by Polyphen2. Of these, 7887 variants are seen more than once which is about 4% of the population of variants. If we only consider variants that have a Polyphen2 score having a low enough value so it is characterized as 'possibly damaging' or 'benign’, then there are 146,748 total variants with 5956 variants that are seen at least twice. This value is still close to 4% of the population. Of these, there are 5593 variants seen twice, and 363 seen more than twice. In fact, 10 variants are observed in 20 patients. A single variant is observed in 59 patients. The particular gene the variant occurred in was HSD17B7P2 which has a Polyphen2 score of 0.009 and was classified as 'benign'.

However, since 96% of the variants were unique among patients, we assume that we will not see the same mutation occur twice in the same tumor, satisfying the infinite allele assumption. Because this is satisfied, we should be able to use the growth in passenger mutations (or alleles) over time as a means to determine the rate of tumor growth given a baseline mutation probability per cell replication event.

4.5 Problems and Future Directions

We showed the beginnings of a practical example to estimate the p.g.f. for tumor growth in lung adenocarcinoma. Unfortunately, we show that such results become complicated quickly, and there also appear to be too many variables to estimate given the limited scope of data. Having single-cell sequencing data would give higher resolution and give more accurate estimates by removing some steps about inferring phylogenies. Even with single-cell sequencing data, we are left with moment functions that grow exponentially fast in the number of terms with respect to the number of clones. Even though these are linear combinations of exponential functions, each coefficient contains higher order polynomials of \( p \) which would make estimation with them unsolvable. Essentially,
One of the problems with the methods above are the intractable likelihoods and generating functions that come from continuous time branching processes. We need to determine the first and second moments in order to get estimates for the relative proportion of each type of clone, but these second moment functions become large very quickly. The results are always a linear combination of exponential functions, but even a four-type case resulted in an expression with many terms. Instead of taking a frequentist approach to finding parameters, we could take a Bayesian approach and take advantage of the simplicity of simulating a branching process. Approximate Bayesian Computation is one set of methods we propose for finding a set of parameters.

Bayesian Estimation and Approximate Bayesian Computation

We are interested in estimation of the offspring probability generating function of a multitype branching process conditional only on the ancestor and $N^{th}$ generations. We use approximate Bayesian computation and modify the Gibbs Sampler described in Gonzalez et al. (2013) in order to simulate the missing data of the middle generations. We show this approach in discrete time since we consider the process by generations.

Define the $k$-type Galton-Watson branching process $\{Z(n), n = 0, 1, \ldots \}$ where $Z(n) = (Z_1(n), \ldots, Z_k(n))$. Suppose individuals of type $i$ has offspring distribution $p_i = (p_{i,s})$ whose support is $\mathcal{I}_i$ (or $\mathcal{I}$ for simplicity where we assume the same support for all $i$). That is, for $s \in \mathcal{I}$, $p_{i,s} = P[Z(1) = s | Z(0) = e_i]$. At each generation, there is probability $\mu$ that an offspring undergoes a new neutral mutation regardless of type. Denote the process of labels of each type $\{K(n), n = 0, 1, \ldots \}$ and the cumulative number of labels $\{S(n), n = 0, 1, \ldots \}$. Note that $S_i(n) = \sum_{r=1}^{n} K_i(r)$. We first show a simple extension to Gonzalez et al. (2013) under the infinite allele
model to create a Gibbs’ sampler in order to estimate $p$ and $\mu$ using the number of cells and cumulative number of labels up to generation $N$ as our data. Further, we consider the scenario where data is missing, or only the final generation is known and extend the Gibbs sampler. Last, we create a hybrid Gibbs Sampler with Approximate Bayesian Computation methods which allow us to use $\{K(n)\}$ instead of $\{Z(n)\}$.

**Gibbs sampler using complete data**

Gonzalez et al. (2008) show a Gibbs sampler to approximate the posterior distribution for the probability generating function of a multitype Galton Watson process. Let $\eta_{ij}(n)$ be the $k$-dimensional offspring vector for the $j^{th}$ type $i$ individual in the $n^{th}$ generation. The process $\{Z(n)\}$ has the branching property

$$Z(n + 1) = \sum_{i=1}^{k} \sum_{j=1}^{Z_i(n)} \eta_{ij}(n).$$

Also, given a vector $s \in \mathcal{S}$, define

$$Z_i(n,k) = \sum_{j=1}^{Z_i(n)} I(\eta_{ij}(n) = s).$$

This term represents the number of type $i$ individuals in generation $n$ having the offspring vector $s$. Note that this happens for each individual with probability $p_{i,s}$ as stated before independent of other individuals. Thus the distribution for $Z_i(n,s)|(Z(n),p) \sim \text{Multinomial}(Z_i(n),p_i)$. If we include conditioning on the size of the next generation, then we normalize the probabilities such that

$$P[(Z_i(n,s), s \in \mathcal{S}, i = 1, \ldots, k)|(Z(n), Z(n + 1), p)] = 0 \text{ if } Z(n+1) \neq \sum_{i=1}^{k} \sum_{s \in \mathcal{S}} sZ_i(n,s).$$

That makes sampling at each generation a matter of sampling from a multinomial
distribution under the condition that the sum over all types is equal to the number of individuals in the next generation. Gonzalez et al. (2008) then show a Dirichlet prior can be used for $p_i$ with parameters $\alpha_i$ to give a posterior distribution that is Dirichlet with parameters for $p_i$ of $\alpha_i + \sum_{r=0}^{N-1} Z_i(r, s), s \in S_i$ where $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ik})$.

Since $S_i(n)$ is the cumulative number of new labels for a type $i$ individual in generation $n$, define $Y_i(n) = S_i(n) - S_i(n - 1)$ which counts the number of new labels in the current generation. Conditioned on $Z_i(n)$ and $\mu$, the random variable $Y_i(n) | (Z_i(n), \mu) \sim \text{Binomial}(Z_i(n), \mu)$. Using a Beta prior for $\mu$ with parameters $\beta_1$ and $\beta_2$ is the natural choice since mutations occur independently with the same probability. Then if $L(N, Y) = (Y_i(n), i = 1, \ldots, k, n = 0, \ldots, N)$ and $L(N, X) = (Z_i(n), i = 1, \ldots, k, n = 0, \ldots, N)$ represent the history of the processes,

$$P[\mu | (L(N, Y), L(N, Z))] \propto P[\mathcal{L}(N, Y) | (\mu, L(N, Z))] P(\mu | L(N, Z))$$

$$= P[\mathcal{L}(N, Y) | (\mu, L(N, Z))] \pi(\mu)$$

$$= \prod_{i=1}^k \prod_{n=1}^N P[Y_i(n) | (Z_i(n), \mu)] \pi(\mu),$$

so the posterior distribution is Beta with parameters $\beta_1 + \sum_{i=1}^k \sum_{n=1}^N Y_i(n)$ and $\beta_2 + \sum_{i=1}^k \sum_{n=1}^N (Z_i(n) - Y_i(n))$.

We can use the same Gibbs sampler to estimate $p$ for the process and we have a posterior distribution for $\mu$ as well.

**Gibbs sampler with missing data**

Assume $Z_i(n)$ and $Y_i(n)$ is known only for a few generations, including $n = 0$ and $n = N$. We show results for when only 2 generations are known and this can be extended to multiple generations by breaking it up. Note that as before, $Z_i(n, s) | (Z(n), p) \sim \text{Multinomial}(Z_i(n), p_i)$. Further, we can construct the next gen-
eration from the current generation using the identity

\[ Z(n + 1) = \sum_{i=1}^{k} \sum_{s \in S_i} sZ_i(n, s), \]

so we get the joint distribution for the number of offspring in each generation by using this and the fact that the chain is Markov by

\[
P((Z_i(n, s_i), i = 1, \ldots, k, n = 0, \ldots, N - 1)|(Z(N), Z(0))) = \\
= P((Z_i(N - 1, s))|(Z(N), Z_j(n, s_j), j = 1, \ldots, k, n = 0, \ldots, N - 2)) \times \\
\times P((Z_j(n, s_j), j = 1, \ldots, k, n = 0, \ldots, N - 2)|(Z(N), Z(0))) \\
= P((Z_i(N - 1, s))|(Z(N), Z_j(N - 2, s_j), j = 1, \ldots, k)) \times \\
\times P((Z_j(n, s_j), j = 1, \ldots, k, n = 0, \ldots, N - 2)|(Z(N), Z(0))) \\
= \prod_{n=1}^{N-1} \prod_{i=1}^{k} P((Z_i(n, s))|(Z(n), Z_j(n - 1, s_j), j = 1, \ldots, k)) \times P((Z_i(0, s))|(Z(0)))
\]

normalized so that

\[
P((Z_i(n, s_i), i = 1, \ldots, k, n = 0, \ldots, N)|(Z(N), Z(0))) = 0 \text{ if } Z(N) \neq \sum_{i=1}^{k} \sum_{s \in S_i} sZ_i(N - 1, s).
\]

That is, we only need to condition on the final generation, so when simulating, we only keep iterations where the simulated final generation equals the number of individuals in the final generation for the data.

Since \( Y_i(n)|(Z_i(n), \mu) \) has a binomial distribution, we can simulate these values in a Gibbs sampler to get a posterior distribution for \( \mu \) under the condition that \( \sum_{n=0}^{N} Y(n) = S(n) \). Pseudocode for the Gibbs sampler is given:

Fixed \( p_i^{(0)} \sim \text{Dirichlet}(\alpha_i), i = 1, \ldots, k \)

Fixed \( \mu^{(0)} \sim \text{Beta}(\beta_1, \beta_2) \)
while $l \leq M$ do

for $n = 0$ to $N - 1$ do

Generate $Z_i^{(l)}(n, s_i) \sim Z_i(n, s_i) | (Z_i^{(l)}(n), p_i^{(l)})$, $i = 1, \ldots, k, s_i \in \mathcal{S}_i$

$Z^{(l)}(n + 1) \leftarrow \sum_{i=1}^k \sum_{s_i \in \mathcal{S}_i} s Z_i^{(l)}(n, s)$

Generate $Y_i^{(l)}(n + 1) \sim Y_i(n + 1) | (Z_i^{(l)}(n + 1), \mu^{(l)})$

end for

Generate $p_i^{(l)} \sim p_i | (p_i^{(l-1)}, Z^{(l)}(n, s), n = 0, \ldots, N - 1, s \in S)$

Generate $\mu^{(l)} \sim \mu | (\mu^{(l-1)}, Y(n), n = 0, \ldots, N)$

$l \leftarrow l + 1$

end while

The need to condition on the final generation can cause this sampler to take a long time, especially if $N$ is large. When running the sampler, instead of analytically normalizing to determine the p.d.f. based on $(Z(N), K(N))$, we just rerun a simulation when

$$Z(N) \neq \sum_{i=1}^k \sum_{s \in \mathcal{S}_i} s Z_i(N - 1, s)$$

or

$$\sum_{n=0}^N Y(n) \neq S(n).$$

We discuss a method that incorporates this Gibbs sampler by approximating this in order to speed up the time to iteration next.

**Approximate Bayesian Computation Hybrid**

To get around this problem of requiring the simulations to have the same exact results as the data is to only require them to be “close”. That is, we only rerun simulations when

$$\max(\delta_1(Z(N), \sum_{i=1}^k \sum_{s \in \mathcal{S}_i} s Z_i(N - 1, s)), \delta_2(\sum_{n=0}^N Y(n), S(n)) > \epsilon$$
for some $\epsilon$ which we designate as our threshold. The functions $\delta_1$ and $\delta_2$ can be defined in a number of ways, including the 1-norm, 2-norm, and others we mention in the following section. For example, using the 2-norm, we accept an iteration of the sampler when

$$\max \left( \left\| \mathbf{Z}(N) - \sum_{i=1}^{k} \sum_{s \in \mathcal{S}} s \mathbf{Z}_i(N - 1, s) \right\|_2, \left\| \sum_{n=0}^{N} \mathbf{Y}(n) - \mathbf{S}(n) \right\|_2 \right) < \epsilon.$$ 

This ensures that the simulation gives results that are "close enough" to accept.

This takes into account the methods from Approximate Bayesian Computation by requiring a criterion to accept simulations, but still using aspects of the Gibbs sampler and altering the posterior distribution function based on each run. This method allows the simulations to run faster than a pure Gibbs sampler, but makes use of every iteration to help determine the posterior distribution for the parameters $p$ and $\mu$.

**Sequential Monte Carlo**

We can use an adaptive sampler to increase the efficiency of the ABC algorithms. The idea is to sample for a sequence of posterior distributions based on a decreasing value of $\epsilon$. The benefit of this is a speedup in the sampling since we limit the prior each time and slowly reduce the tolerance, so we are closing in on the target distribution. As such, less samples are rejected based on the prior distribution. We then use the posterior as a prior for the next sampler which has a smaller value of epsilon. The algorithm comes from Beaumont et al. (2012) and is as follows:

At iteration $t = 1$

for $i = 1$ to $M$ do

repeat
Simulate $\theta_i^{(1)} \sim \pi(\theta)$ and $(Z(N), K(N)) \sim f((Z(N), K(N)|\theta_i^{(1)})$ by running a sample process given the parameters $\theta_i^{(1)}$.

until $\rho((Z(N), K(N)), (Z \ast (N), K \ast (N))) \leq \epsilon_1$

set $\omega_i^{(1)} = 1/M$

end for

Take $\Sigma = 2\text{Cov}(\theta_i^{(1)})$.

for $t = 2$ to $T$ do

for $i = 1$ to $M$ do

repeat

Sample $\theta_i^{*}$ from $\theta_j^{(t-1)}$ with weights $\omega_j^{(t-1)}$.

Generate $\theta_i^{(t)} \sim N(\theta_i^{*}, \Sigma_{t-1})$ and $(Z(N), K(N)) \sim f((Z(N), K(N)|\theta_i^{(t)})$

until $\rho((Z(N), K(N)), (Z \ast (N), K \ast (N))) \leq \epsilon_t$

Set $\omega_i^{(t)} = \pi(\theta_i^{(t)})/\sum_{j=1}^{N} \omega_j^{(t-1)} \cdot \phi(\Sigma_{t-1}^{-1/2}(\theta_i^{(t)} - \theta_j^{(t-1)}))$

end for

Set $\Sigma_t = 2\text{Cov}(\theta_i^{(t)})$

end for

In order to run this properly, we have to transform our parameters such that they can come from a normal distribution. That is, we let $\theta_i^{(t)}$ be the vector $(q_{1i}^{(t)}, \ldots, q_{ki}^{(t)}, \mu_i^{(t)*})$.

Let $\mu_i^{(t)*} = \text{logit}(\mu_i^{(t)})$, so this allows the support of the distribution to be the real line instead of $[0, 1]$. We set $q$ as a multinomial logit transformation, or

\[
p_{1,1} = \frac{\exp(q_{1,1})}{1 + \sum_{j=1}^{s-1} \exp(q_{1,j})}
\]

and

\[
p_{1,s} = \frac{1}{1 + \sum_{j=1}^{s-1} \exp(q_{1,j})}
\]
That is, given a parent of type \( i \) with p.g.f. \( f_i(s) \) having coefficients \((p_1, \ldots, p_s)\), we rescale them to \( q \) using a logit transformation so that we are able to have a support equal to the real line and can use a multivariate normal for our prior. This way, \( \theta^{(t)}_i \) can be sampled from a multivariate normal distribution and transformed back to the respective values of \( p_i \) and \( \mu \) without worrying about boundaries. The rest of the algorithm then works out fine.

4.6 Conclusion

The methods presented in this section show how clonal evolution can be modeled with multitype branching processes. While multitype branching processes are intuitive methods for modeling dividing cellular populations, the area hasn’t been developed as much as their single-type counterparts. Also, models become complex very quickly, so there is a divergence from what is theoretically possible and what can be shown computationally. Simple adjustments to the models result in differential equations without general solutions, so each process must be considered on a case-by-case basis.

We are also faced with the problem of data. The most ubiquitous type of data (Whole-genome or whole-exome sequencing) provides the coarsest resolution if space and time are not taken into account during collection. We are only able to see marginal proportions and have to infer correlations and evolutions of the different clones. This inference adds more noise to our model before even trying to model clonal dynamics which could be avoided once single-cell sequencing data becomes more available. The problems discussed that are not based on the process itself can all be remedied, and the outlook for the next few years is optimistic that these models will become relevant and useful as single cell data is available. For now, the models are best used in cases where a few clones are assumed, as we showed in lung adenocarcinoma.

We presented a branching process in discrete and continuous times that are used as
models for clonal evolution in this thesis which account for unique types of mutations that affect growth as well as mutations that do not. We argue that these processes can be used to model cancer growth in a way that hasn’t been approach much yet, where we assume driver mutations have a specific effect on tumor growth that is unique to that mutation. Using this assumption generalizes previous work done with cancer progression and branching processes. Along with this effect, we also include processes that count the number of passenger mutations we should see, assuming a baseline passenger mutation rate. Doing so leads to the idea that passenger mutations serve as a molecular clock. Thus, tumors with more passenger mutations either grow faster or are older.

The method still has some simplifying assumptions, such as the idea of a single baseline mutation rate. Because of this, we generalized our processes to a Jagers-Crump-Mode branching process where we make no assumptions about lifetimes or the mutation rate for passengers with respect to each type. The generalization creates a framework for future studies with stronger assumptions, and more importantly, shows us the asymptotic results presented in our discrete and continuous Markov processes still hold regardless of the underlying distribution.
References


[58] T. O. McDonald and M. Kimmel, “A multitype infinite-allele branching process with applications to cancer evolution,” *Journal of Applied Probability*, vol. 52, 2015. 3.1, 3.1, 3.2, 3.3, 3.4, 3.4, 3.6

[59] ———, “Multitype infinite-allele branching processes in continuous time,” 2015. 3.1


