DNA sequence evolution in the Yule process

David Cheek* and Tibor Antal†

School of Mathematics, University of Edinburgh

Abstract

We study a fundamental model of DNA evolution in a growing population of cells: cell divisions follow the Yule process, and each cell contains a sequence of nucleotides which can mutate at cell division. Following typical parameter values in bacteria and cancer cell populations, we take the mutation rate to zero and the final number of cells to infinity. We prove that almost every site (entry of the sequence) is mutated in only a finite number of cells, and these numbers are independent across sites. However independence breaks down for the rare sites which are mutated in a positive fraction of the population. The model is free from the popular but disputed infinite sites assumption. Violations of the infinite sites assumption are widespread while their impact on mutation frequencies is negligible at the scale of population fractions. Some results are generalised to allow for cell death, selection, and site-specific mutation rates. We also estimate mutation rates in a lung adenocarcinoma.

*dmcheek7@gmail.com
†tibor.antal@ed.ac.uk
Contents

1 Introduction 3

2 Model 6

3 Preliminaries 6
  3.1 Luria-Delbrück distribution 6
  3.2 Yule tree 7
  3.3 Mutation frequency notation 8
  3.4 Parameter regime 9

4 Mutation frequencies 9

5 Generalisations 12
  5.1 Model and notation 12
  5.2 Generalised Luria-Delbrück distribution 14
  5.3 Mutation frequencies 15
  5.4 Open problems 16

6 Proofs: mutations at finite numbers 17
  6.1 Counting genomes 17
  6.2 Construction 19
  6.3 Neglecting double mutations 26
  6.4 Convergence of genome counts 29
1 Introduction

A population of dividing cells with a mutating DNA sequence is ubiquitous in biology. We study a fundamental model of this process. Starting with one cell, cells divide according to the Yule process. That is, cells divide independently at a constant rate. As for DNA, we loosely follow classic models from phylogenetics [11, 19]: each cell contains a sequence of the nucleotides A, C, G, and T, and each site (entry of the sequence) can mutate independently at cell division. We also study some generalisations to site-specific and nucleotide-specific mutation rates, cell death, and selection (a cell’s division and death rates are a function of its DNA sequence).

Let’s discuss a specific motivation. In recent years, cancer genetic data has
been made available in vast quantities. One especially abundant type of data consists of mutation frequencies in individual tumours. These data take the form of a vector \((x_i)_{i \in S}\), where \(i \in S\) denotes genetic sites and \(x_i\) is the frequency of cells which are mutated at site \(i\). A natural question is: What does a simple mathematical model of tumour evolution predict for mutation frequencies?

Some remarkable works on the topic are [21, 4, 20, 6], which consider branching process and deterministic models of tumour evolution. They compare theory with data, estimating evolutionary parameters such as mutation rates. A central feature of their theory, and of countless other works, is the so-called infinite sites assumption (ISA). The ISA states that no genetic site can mutate more than once in a tumour’s lifetime; its simplicity drives its popularity. However recent statistical analysis of single cell sequencing data [15] shows “widespread violations of the ISA in human cancers”.

For a ‘non-ISA’ model of a growing population of cells, there is in fact a famous example: Luria and Delbrück [17] modelled recurrent mutations in an exponentially growing bacterial population. Subsequent works [16, 13, 9, 12, 14] (and others) adapted Luria and Delbrück’s model to branching processes and calculated mutation frequencies. These works describe only two genetic states, mutated or not mutated, effectively restricting attention to a single genetic site. In [5] we offered an account of one such model, proving limit theorems for mutation times, clone sizes, and mutation frequencies. We then briefly studied an extension to multiple genetic sites. Now we offer a self-contained sequel to [5], slightly adapting the model, and aiming for a deeper understanding of multiple sites.

The present work is narrower than [5] in that we study only one parameter regime. The studied parameter regime is popular in Luria-Delbrück-type models of bacteria and it makes sense for tumour modelling too. A detected tumour comprises in the region of \(10^9\) cells. On the other hand, the mutation rate per site per cell division is in the region of \(10^{-9}\) [10]. Therefore we take the final number of cells to infinity and the mutation rate to zero with their product finite.

The number of cells mutated at a given site (mutations are defined relative to the initial cell) converges to the Luria-Delbrück distribution. This recovers a well-known result of single-site models [16, 9, 14, 12, 5]. So a site is mutated
in only a finite number of cells, standing in contrast to the infinite total number of cells. Going beyond [16, 9, 14, 12, 5], we also study the rare event that a site is mutated in a positive fraction of cells. We show that, when appropriately scaled, this fraction of cells follows a power-law distribution.

Across sites, mutation frequencies are asymptotically independent. The independence leads to a many-sites law of large numbers. Specifically, the site frequency spectrum (empirical measure of mutation frequencies) converges to a deterministic measure concentrated at finite cell numbers. At positive fractions of cells, away from the mass concentration, independence breaks down and the site frequency spectrum converges to a Cox process. These results go beyond [7, 21, 4, 6, 5] who only give the expected site frequency spectrum, so our work contributes an appreciation of randomness.

We also assess the infinite sites assumption’s validity. A positive fraction of sites violate the ISA with probability one. For typical parameter values in tumours, the number of sites to violate the ISA is at least millions, or even billions. Thus our work agrees with [15]’s statistical analysis of single cell sequencing data, that violations of the ISA are widespread. Then there is the question of ISA violations’ impact on mutation frequencies. At the scale of cell numbers, mutation frequencies are significantly impacted. At the scale of population fractions, mutation frequencies are not impacted. The majority of cancer genetic data is not derived from single cell sequencing but rather from bulk sequencing [6], which is sensitive on the scale of population fractions. Therefore our work endorses the ISA-reliant analyses of bulk sequencing data [21, 4, 20, 6].

We believe that the studied model of DNA sequence evolution offers more value than just an assessment of the ISA. For one, there is the aesthetic appeal of an evolving sequence. More objectively, the sequence gives identities to genetic sites, inviting discussions on: (1) correlations between sites; (2) site-specific and nucleotide-specific mutation rates; (3) site-specific effects on cell division and death rates. Although our work is primarily intended to be a mathematical exploration, for an example of (2) in action we estimate mutation rates in a lung adenocarcinoma. As for (3), which asks for greater generality than the Yule process, our analysis is less developed and some questions are posed for further research.
2 Model

Here the model is stated in its simplest form. It comprises two parts.

1. Population dynamics: Starting with one cell, cells divide according to the Yule process. That is, cells divide independently at constant rate.

2. Genetic information: The set of nucleotides is $N = \{A, C, G, T\}$. The set of genetic sites is some finite set $S$. The set of genomes (or DNA sequences) is $\mathcal{G} = N^S$. Each cell has a genome, i.e. is assigned an element of $\mathcal{G}$. Suppose that a cell with genome $(v_i)_{i \in S} \in \mathcal{G}$ divides to give daughter cells with genomes $(V_i^{(1)})_{i \in S}$ and $(V_i^{(2)})_{i \in S}$. Conditional on $(v_i)$, the $V_i^{(r)}$ are independent over $i \in S$ and $r \in \{1, 2\}$, and

$$
\mathbb{P}
\left[
V_i^{(r)} = \psi | (v_i)
\right] = \begin{cases}
\mu/3, & \psi \neq v_i; \\
1 - \mu, & \psi = v_i.
\end{cases}
$$

It is also assumed that mutations occur independently for different cell divisions.

The model is generalised to cell death, selection, and nucleotide/site-specific mutation rates in Section 5.

3 Preliminaries

3.1 Luria-Delbrück distribution

Let $(Y_k)_{k \in \mathbb{N}}$ be an i.i.d. sequence of random variables with

$$
\mathbb{P}[Y_1 = j] = \frac{1}{j(j+1)}
$$

for $j \in \mathbb{N}$. Let $K$ be an independent Poisson random variable with mean $c$. The Luria-Delbrück distribution with parameter $c$ is defined as the distribution of

$$
B = \sum_{k=1}^{K} Y_k.
$$
It is commonly seen in its generating function form (e.g. [16, 24])
\[ E_z^B = (1 - z)^{c(z^{-1} - 1)}. \]

Although the distribution is named after Luria and Delbrück (due to their groundbreaking work [17]), it was derived by Lea and Coulson [16]. See [24] for a historical review.

The Luria-Delbrück distribution’s power-law tail was derived in [18].

Lemma 3.1. \( \lim_{m \to \infty} m \mathbb{P}[B \geq m] = c. \)

### 3.2 Yule tree

The set of all cells to ever exist, following standard notation, is
\[ T = \bigcup_{l=0}^{\infty} \{0, 1\}^l. \]

A partial ordering, \( \prec \), is defined on \( T \). For \( x, y \in T, x \prec y \) means that cell \( y \) is a descendant of cell \( x \). That is, \( x \prec y \) if

1. there are \( l_1, l_2 \in \mathbb{N}_0 \) with \( l_1 < l_2 \) and \( x \in \{0, 1\}^{l_1}, y \in \{0, 1\}^{l_2} \); and
2. the first \( l_1 \) entries of \( y \) agree with the entries of \( x \).

Note that \( \emptyset \in T \) and that \( \emptyset \prec x \) for any \( x \in T \setminus \{\emptyset\} \). So \( \emptyset \) is the initial cell from which all other cells descend. For further notation, write \( x \preceq y \) if \( x \prec y \) or \( x = y \). Also, write \( x0 \) and \( x1 \) for the daughters of \( x \in T \); precisely, if \( x \in T \) and \( j \in \{0, 1\} \), then \( xj \) is the element of \( \{0, 1\}^{l+1} \) whose first \( l \) entries are the entries of \( x \) and whose last entry is \( j \).

Let \( (A_x)_{x \in T} \) be a family of i.i.d. exponentially distributed random variables with mean 1. \( A_x \) is the lifetime of cell \( x \). The cells alive at time \( t \) are

\[ T_t := \left\{ x \in T : \sum_{y \prec x} A_y \leq t < \sum_{y \preceq x} A_y \right\}. \]
The proportion of cells alive at time $t$ which are descendants of cell $x$ (including $x$) is

$$P_{x,t} := \frac{|\{y \in \mathcal{T}_t : x \preceq y\}|}{|\mathcal{T}_t|}.$$ 

**Lemma 3.2.** For each $x \in \mathcal{T}$,

$$\lim_{t \to \infty} P_{x,t} = P_x := \prod_{\emptyset \prec y \preceq x} U_y$$

almost surely, where

1. the $U_y$ are uniformly distributed on $[0, 1]$;
2. for any $y \in \mathcal{T}$, $U_{y_0} + U_{y_1} = 1$;
3. $(U_{y_0})_{y \in \mathcal{T}}$ is an independent family.

### 3.3 Mutation frequency notation

When DNA is taken from a tumour, the tumour’s age is unknown, but one may have a rough idea of its size. Therefore we are interested in the cells’ genetic state at the random time

$$\sigma_n = \min\{t \geq 0 : |\mathcal{T}_t| = n\},$$

when the total number of cells reaches some given $n \in \mathbb{N}$.

Write $V^{\mu}(x) = (V^{\mu}_i(x))_{i \in \mathcal{S}} \in \mathcal{S}$ for the genome of cell $x \in \mathcal{T}$ (where $\mu$ is the mutation rate). So $(V^{\mu}(x))_{x \in \mathcal{T}}$ is a Markov-process indexed by $\mathcal{T}$ with transition rates given in Section 2. Write $V^{\mu}(\emptyset) = (u_i)_{i \in \mathcal{S}}$ for the initial cell’s genome. A genetic site is said to be mutated if its nucleotide differs from that of the initial cell. Write

$$B_i^{n,\mu} = |\{x \in \mathcal{T}_{\sigma_n} : V^{\mu}_i(x) \neq u_i\}|$$

for the number of cells which are mutated at site $i \in \mathcal{S}$ when there are $n$ cells in total. The quantity (1), and its joint distribution over $\mathcal{S}$, is the key object of our study.
3.4 Parameter regime

The number of cells in a detected tumour may be in the region of $n = 10^9$, whereas the mutation rate is in the region of $\mu = 10^{-9}$ [10]. The human genome’s length is around $|S| = 3 \times 10^9$. Very roughly,

$$n \approx \mu^{-1} \approx |S|.$$ 

Therefore we study the limits:

- $\mu \to 0$, $n\mu \to \theta < \infty$;
- $\mu \to 0$, $n\mu \to \theta < \infty$, $|S| \to \infty$ (sometimes with $|S|\mu \to \eta < \infty$).

Remark 3.3. Taking the number of sites to infinity is not to be confused with the infinite sites assumption.

4 Mutation frequencies

The first result shows that sites are typically mutated in only a finite number of cells, and that these numbers are independent across sites.

Theorem 4.1. As $\mu \to 0$ and $n\mu \to \theta < \infty$,

$$(B_{i:n:\mu}^{n})_{i \in S} \to (B_{i})_{i \in S}$$

in distribution, where the $B_i$ are i.i.d. and have Luria-Delbrück distribution with parameter $2\theta$.

Remark 4.2. Taking $|S| = 1$, Theorem 4.1 recovers results of single site models [16, 9, 12, 14, 5].

The site frequency spectrum is a popular summary statistic of genetic data. It is defined as the empirical measure of mutation frequencies:

$$\sum_{i \in S} \delta_{B_{i:n:\mu}}.$$ 

The site frequency spectrum sees a law of large numbers.
**Theorem 4.3.** As $\mu \to 0$, $n\mu \to \theta < \infty$, and $|S| \to \infty$,

\[
\frac{1}{|S|} \sum_{i \in S} \delta_{B_i^{n,\mu}} \to \Lambda
\]

in probability, where $\Lambda$ is the Luria-Delbrück distribution with parameter $2\theta$. Convergence is on the space of probability measures on the non-negative integers equipped with the topology of weak convergence.

Theorems 4.1 and 4.3 teach us that almost every site is mutated in only a finite number of cells. What about the rare sites which are mutated in a positive fraction of cells? Heuristically, the Luria-Delbrück distribution’s tail gives the probability that site $i$ is mutated in at least fraction $a$ of cells:

\[
P[n^{-1}B_i^{n,\mu} > a] \approx P[B_i \in (na, n)] \approx 2\mu(a^{-1} - 1). \tag{2}
\]

Approximation (2) is a hand-waving consequence of Theorem 4.1. Approximation (3) is due to Lemma 3.1. The next result offers rigour.

**Theorem 4.4.** Let $i \in S$ and $a \in (0, 1)$. As $\mu \to 0$ and $n\mu \to \theta < \infty$,

\[
\mu^{-1}P[n^{-1}B_i^{n,\mu} > a] \to 2(a^{-1} - 1).
\]

Theorem 4.4 and linearity of expectation yield the mean site frequency spectrum at positive fractions of the population.

**Corollary 4.5.** Let $a \in (0, 1)$. As $\mu \to \infty$, $n\mu \to \infty$, and $|S|\mu \to \eta < \infty$,

\[
E \sum_{i \in S} \delta_{n^{-1}B_i^{n,\mu}}(a, 1) \to 2\eta(a^{-1} - 1).
\]

The next result gives the distribution of the site frequency spectrum at positive fractions of the population.

**Theorem 4.6.** As $\mu \to 0$, $n\mu \to \theta < \infty$, and $|S|\mu \to \eta < \infty$,

\[
\sum_{i \in S} \delta_{n^{-1}B_i^{n,\mu}} \to \sum_{x \in \mathbb{P} \setminus \{\emptyset\}} M_x \delta_{P_x}
\]

in distribution. Convergence is on the space of measures on $[\epsilon, 1]$ equipped with the vague topology, where $\epsilon \in (0, 1)$ is arbitrary. The random variables which appear in the limit are:
Figure 1: The number of mutant cells with respect to a single site is simulated $10^5$ times. The parameters are $\mu = 10^{-3}$ and $n = 10^3$. The plot compares $\mathbb{P}[n^{-1}B_1^{n,\mu} \in (a, 1)]$ (simulation), $\mathbb{P}[n^{-1}B_1 \in (a, 1)]$ (Theorem 4.1), and $2\mu(a^{-1} - 1)$ (Theorem 4.4), for $a \in (0, 1)$. Simulation and Theorem 4.1 appear indistinguishable.
• $(M_x)$ is a family of i.i.d. Poisson($\eta$) random variables;

• $(P_x)$ is from Lemma 3.2 and is independent of $(M_x)$.

Remark 4.7. The mean site frequency spectrum, according to Theorem 4.6’s limit, is

$$\mathbb{E} \left[ \sum_{x \in \mathcal{T} \setminus \{\emptyset\}} M_x \delta_{P_x}(a, 1) \right] = 2\eta(a^{-1} - 1),$$

which recovers the limit of Corollary 4.5.

Remark 4.8. The variance of the site frequency spectrum, according to Theorem 4.6’s limit, is bounded below by

$$\text{Var} \left[ \sum_{x \in \mathcal{T} \setminus \{\emptyset\}} M_x \delta_{P_x}(a, 1) \right] \geq 2\eta(a^{-1} - 1).$$

In particular, the coefficient of variation tends to infinity as $a \uparrow 1$.

The details of Remarks 4.7 and 4.8 are given in Section 7.

5 Generalisations

Motivated by biological reality, we introduce some generalisations: cell death, selection, and heterogeneous mutation rates.

5.1 Model and notation

Starting with one cell, the cell population grows according to a continuous-time multitype Markov branching process. The types are the genomes, elements of $\mathcal{S} = \mathbb{N}^\mathcal{G}$. It will be helpful to classify different types of genetic site. Partition the sites into neutral and selective sites:

$$\mathcal{S} = \mathcal{S}_{\text{neut}} \cup \mathcal{S}_{\text{sel}}.$$
Figure 2: The site frequency spectrum is simulated a single time. The parameters are $\mu = 10^{-3}$, $n = 10^3$, and $|S| = 10^3$. The plot compares $\sum_{i \in S} \delta_{n^{-1}B_i^n}(a,1)$ (simulation), $|S|\mathbb{P}[n^{-1}B_1 \in (a,1)]$ (Theorem 4.3), and $2|S|\mu(a^{-1} - 1)$ (Corollary 4.5), for $a \in (0,1)$.
with $S_{\text{sel}} \neq \emptyset$. For a genome $v = (v_i)_{i \in S}$, write $v' = (v_i)_{i \in S_{\text{sel}}}$ for its restriction to the selective sites. Let $\alpha$ and $\beta$ be functions with domain $\mathbb{N}^{S_{\text{sel}}}$ and range $[0, \infty)$. A cell with genome $v$ divides at rate $\alpha(v')$ (to be replaced by two daughter cells) and dies at rate $\beta(v')$.

The initial cell is said to have genome $u$, which is assumed to give a positive growth rate: $\alpha(u') > \beta(u')$.

Consider a cell with genome $(v_i)_{i \in S}$ dividing to give daughter cells with genomes $(V_i^{(1)})_{i \in S}$ and $(V_i^{(2)})_{i \in S}$. Conditional on $(v_i)$, the $V_i^{(r)}$ are independent over $i \in S$ and $r \in \{1, 2\}$, and

$$
P \left[ V_i^{(r)} = \psi | (v_i) \right] = \mu_i^{v_i, \psi}.
$$

Slightly adapting previous notation, write

$$
\mu = \left( \mu_i^{\chi, \psi} \right)_{i \in S, \chi, \psi \in \mathbb{N}}
$$

for the collection of mutation rates. Now let’s state the notation for mutation frequencies (for brevity, unlike in Section 3.3, we shall do so in words). Write $B_i^{n, \mu}$ for the number of cells which are mutated at site $i$ when $n$ cells are first reached conditioned on the event that $n$ cells are reached.

### 5.2 Generalised Luria-Delbrück distribution

Let $a, b \geq 0$. Let $(Y_k(\cdot))_{k \in \mathbb{N}}$ be an i.i.d. sequence, where $Y_1(\cdot)$ is a birth-death branching process with birth and death rates $a$ and $b$ respectively and initial condition $Y_1(0) = 1$. Let $(\xi_k)_{k \in \mathbb{N}}$ be an i.i.d. sequence of exponentially distributed random variables with mean 1, independent of the $Y_k(\cdot)$. Let $K$ be an independent Poisson random variable with mean $c$. The generalised Luria-Delbrück distribution with parameters $(a, b, c)$ is defined as the distribution of

$$
B = \sum_{k=1}^{K} Y_k(\xi_k).
$$

It is commonly seen in its generating function form (e.g. [12, 14, 5])

$$
E z^B = \exp \left( c \frac{b}{a} \left[ 1, \frac{1}{1+\frac{1}{a-b}} \right] \left[ 1, \frac{b/a - z}{1 - z} \right] \right),
$$

14
where $F$ is Gauss’s hypergeometric function.

Taking parameters $(1, 0, c)$ recovers the Luria-Delbrück distribution with parameter $c$.

### 5.3 Mutation frequencies

To begin, Theorem 4.1 is generalised. The genomes whose only difference from the initial cell’s genome is at site $i \in S$,

$$G_i = \{ v \in S : \forall j \in S, (u_j \neq v_j \iff i = j) \},$$

will play a crucial role.

**Theorem 5.1.** Take $\mu_{i}^{\chi,\psi} \to 0$ and $n\mu_{i}^{\chi,\psi} \to \theta_{i}^{\chi,\psi} < \infty$ for all $i \in S$ and $\chi, \psi \in N$ with $\chi \neq \psi$. Then

$$(B_{i}^{n,\mu})_{i \in S} \to \left(\sum_{v \in G_i} X_v\right)_{i \in S}$$

in distribution, where the $X_v$ are independent and have generalised Luria-Delbrück distributions with parameters

$$\left(\frac{\alpha(u')}{\alpha(u') - \beta(u')}, \frac{\beta(u')}{\alpha(u') - \beta(u')}, \frac{2\alpha(u')\theta_{i}^{u_i,v_i}}{\alpha(u') - \beta(u')}\right).$$

In the next result, which generalises Theorem 4.3, we keep the number of selective sites finite while taking the number of neutral sites to infinity. For this limit, mutation rates require consideration. Partition the set of neutral sites:

$$S_{\text{neut}} = \bigcup_{j \in J} S(j),$$

such that mutation rates and the initial genome’s nucleotides are homogeneous on $S(j)$ ($J$ is just some indexing set). Write $\mu_{i}^{\chi,\psi}(j) = \mu_{i}^{\chi,\psi}$ for the mutation rates of the sites $i \in S(j)$. Write $u(j) = u_i$ for the initial genome’s nucleotide at the sites $i \in S(j)$.
Theorem 5.2. Take $\mu^{\chi,\psi}(j) \to 0$, $n\mu^{\chi,\psi}(j) \to \theta^{\chi,\psi}(j) < \infty$, $|S_{\text{neut}}| \to \infty$, and $|S(j)|/|S_{\text{neut}}| \to q(j)$, for all $j \in J$ and $\chi, \psi \in N$ with $\chi \neq \psi$. Then

$$\frac{1}{|S|} \sum_{i \in S} \delta_{B_{i,\mu}} \Rightarrow \sum_{j \in J} q(j) \Lambda(j)$$

where the $\Lambda(j)$ are generalised Luria-Delbrück distributions with parameters

$$\begin{pmatrix}
\alpha(u') / (\alpha(u') - \beta(u')) & 2\alpha(u') \sum_{\psi \in N \setminus \{u(j)\}} \theta^{u(j),\psi}(j) / (\alpha(u') - \beta(u')) \\
\beta(u') / (\alpha(u') - \beta(u')) & \alpha(u') / (\alpha(u') - \beta(u'))
\end{pmatrix}.$$ 

Convergence is on the space of probability measures on the non-negative integers equipped with the topology of weak convergence.

5.4 Open problems

To generalise Theorem 4.6 we conjecture the following.

Conjecture 5.3. Take $\mu^{\chi,\psi}(j) \to 0$, $n\mu^{\chi,\psi}(j) \to \theta^{\chi,\psi}(j) < \infty$, and $\mu^{\chi,\psi}(j)|S(j)| \to \eta^{\chi,\psi}(j) < \infty$, for all $j \in J$ and $\chi, \psi \in N$ with $\chi \neq \psi$. Then

$$\sum_{i \in S} \delta_{B_{i,\mu}} \Rightarrow \sum_{x \in \mathcal{X}} \sum_{r=1}^{R_x} M_{x,r} \delta_{P_{x}}$$

in distribution. Convergence is on the space of measures on $[\epsilon, 1]$ equipped with the vague topology, where $\epsilon \in (0, 1)$ is arbitrary. The random variables which appear in the limit are:

- $(P_x)_{x \in \mathcal{T} \setminus \{\emptyset\}}$ is distributed as in Lemma 3.2 (and Theorem 4.6), and $P_0 = 1$;
- $(R_x)_{x \in \mathcal{T} \setminus \{\emptyset\}}$ is an i.i.d. family of geometric random variables with parameter $((\alpha(u') - \beta(u'))/\alpha(u' + \beta(u')))$, and $R_0$ is independent of $(R_x)$ but with $R_0 \overset{d}{=} R_0 - 1$;
- $(P_x)$ is independent of $(R_x)$ if and only if $\beta(u') = 0$;
• \((M_{x,r})_{x \in \mathbb{Z}, r \in \mathbb{N}}\) is an i.i.d. family of Poisson random variables with mean 
\[ \sum_j \sum_{\psi \neq u(j)} \eta^{u(j), \psi}(j), \] independent of \((P_x, R_x)\).

See the Appendix for a heuristic derivation of Conjecture 5.3, which is based on a Yule spinal decomposition of the branching process.

Selection in cancer is a major research topic, and there have been attempts to infer selection from cancer genetic data [3, 20, 6]. Pertinently, Theorem 5.2 and Conjecture 5.3 suggest that selection may not be visible in mutation frequency data, which according to [21] is the case for around 1/3 of tumours. However we have assumed that the number of selective sites is kept finite. According to [3], there are \(3.4 \times 10^4\) selective sites at which mutations can positively affect growth rate. Thus insight could be gleaned, for example, by taking \(|S_{\text{sel}}| \to \infty\) with \(\mu^\gamma |S_{\text{sel}}| \to \eta\) for \(\gamma \in (0, 1]\).

6 Proofs: mutations at finite numbers

In this section we prove results on mutations present in only a finite number of cells. In Subsections 6.1 to 6.5 we prove Theorems 4.1 and 5.1. In Subsection 6.6 we prove Theorems 4.3 and 5.2.

6.1 Counting genomes

Assuming mutation rates \(\mu = (\mu_{i,x}^\chi, \psi)_{i \in \mathcal{S}; x, \psi \in \mathbb{N}}\), write
\[ X^\mu_v(t) \]
for the number of cells with genome \(v \in \mathcal{G}\) at time \(t \geq 0\). (Recall that the initial condition is \(X^\mu_v(0) = \delta_{u,v}.\)) Write
\[ \sigma_n^\mu = \min \left\{ t \geq 0 : \sum_{v \in \mathcal{G}} X^\mu_v(t) = n \right\} \]
for the time at which \(n \in \mathbb{N}\) cells are reached, and use the convention \(\min \emptyset = \infty\).
Recall from (4) that $G_i$ is the subset of genomes with exactly one mutation which is at site $i$. Write

$$G_{\geq 2} = \{ v \in G : |\{ i \in S : v_i \neq u_i \}| \geq 2 \}$$

for the subset of genomes with at least two mutations.

**Theorem 6.1.** Take $\mu_i^{\chi,\psi} \to 0$ and $n\mu_i^{\chi,\psi} \to \theta_i^{\chi,\psi} \in (0, \infty)$ for $i \in S$ and $\chi, \psi \in \mathbb{N}$ with $\chi \neq \psi$. Then

$$[(X_v^{\mu_n}(\sigma_n))_{v \in G} | \sigma_n^\mu < \infty] \to (X_v)_{v \in G}$$

in distribution, where the $X_v$ are independent and distributed according to:

- $X_u = \infty$;
- for $v \in G_i$, $X_v$ has generalised Luria-Delbrück distribution with parameters
  $$\left( \frac{\alpha(v')}{\alpha(u') - \beta(u')} , \frac{\beta(v')}{\alpha(u') - \beta(u')} , \frac{2\alpha(u')\theta_i^{u_i,v_i}}{\alpha(u') - \beta(u')} \right)$$;
- for $v \in G_{\geq 2}$, $X_v = 0$.

At first glance, Theorem 6.1 says that cells with the initial genome dominate the population while cells with at least two mutated sites are non-existent. However simulations and biology tell the opposite story, that cells typically have many mutated sites. This apparent contradiction comes because, while the population size and mutation rate reciprocal converge to infinity, the number of sites is kept finite. So the result only makes sense if one is considering a small subset of the billions of sites.

Mutation frequencies are given by

$$B_i^{n,\mu} = \left[ \left( \sum_{v \in G, v_i \neq u_i} X_v^{\mu_n}(\sigma_n) \right) \bigg| \sigma_n^\mu < \infty \right] .$$

Therefore Theorem 6.1, via the continuous mapping theorem, implies Theorems 4.1 and 5.1.
Theorem 6.1’s proof is rather lengthy. So, before jumping in with the technical details, let’s give an overview.

In Subsection 6.2 we present a construction of \((X^\mu_n)_{v \in \mathcal{G}}\). The construction will ultimately illuminate the importance of various subpopulations and the mutations between them. Of particular importance is the primary subpopulation, which is defined as those unmutated cells which have an unbroken lineage of unmutated cells going back to the initial cell. The primary subpopulation, in the limit, grows deterministically and exponentially.

In Subsections 6.3 and 6.4 we show that several events are negligible: primary cells divide to give two mutant daughters; primary cells mutate at multiple sites at once; mutated cells receive further mutations, including backwards mutations. With these events neglected, the situation is pleasingly simplified. The primary subpopulation seeds, as a Poisson process with exponential intensity, single-site mutant subpopulations. The mutant subpopulations grow without further mutations, independently. This gives independent Luria-Delbrück distributions. Finally, in Subsection 6.5 we condition on the event that the population reaches \(n\) cells.

Although the proof’s overview may sound simple, the details are less so. The reader will find that the random time \(\sigma^\mu_n\) shoulders a large responsibility for complexity.

### 6.2 Construction

Additional notation to be used in the proof: for \(v \in \mathcal{G}\),

\[
\epsilon_v = (\delta_{v,w})_{w \in \mathcal{G}}
\]

is the element of \((\mathbb{N}_0)^\mathcal{G}\) denoting that there is one genome \(v\) and zero other genomes. Also, for an interval \(I \subset [0, \infty]\) and a space \(R\), write \(\mathbb{D}(I, R)\) for the space of cadlag functions from \(I\) to \(R\), which is equipped with the usual Skorokhod topology.

Let

\[
[\mu_n]_{n \in \mathbb{N}} = \left(\mu^\chi_\psi_{n,i}\right)_{i \in \mathcal{S}, \chi, \psi \in \mathbb{N}}
\]
be a sequence of mutation rates. Assume that
\[ \lim_{n \to \infty} n\mu_{\chi,\psi}^n = \theta_i^{\chi,\psi} < \infty \]
for \( \chi \neq \psi \).

Fix \( n \in \mathbb{N} \). For \( v, w \in \mathcal{G} \), write
\[ p_n(v, w) = \prod_{i \in S} \mu_{u_i,v_i}^{n,i} \mu_{u_i,w_i}^{n,i} \]
for the probability that a cell with genome \( u \) which divides, gives daughters with genomes \( v, w \) (which implies that we have assumed an ordering of the daughters - the first has genome \( v \) and the second has genome \( w \)). Now the construction of \( (X^\mu_n, (\sigma^\mu_n))_{v \in \mathcal{G}} \) begins. For the foundational step, introduce the following random variables on a fresh probability space.

1. \((Z^n(t))_{t \geq 0}\) is a birth-death branching process with birth and death rates
\[ \alpha_n := \alpha(u')p_n(u, u) \]
and
\[ \beta_n := \beta(u') + \alpha(u') \sum_{v, w \in \mathcal{G}\setminus\{u\}} p_n(v, w). \]
The initial condition \( Z^n(0) = 1 \) is assumed.

2. For \( j \in \mathbb{N} \),
\[ E^n_j \]
are \( \{\emptyset\} \cup (\mathcal{G}\setminus\{u\})^2 \)-valued random variables, with
\[ \mathbb{P}[E^n_j = \emptyset] = \frac{\beta(u')}{\beta_n}, \]
and for \( v, w \in \mathcal{G}\setminus\{u\} \)
\[ \mathbb{P}[E^n_j = (v, w)] = \frac{\alpha(u')p_n(v, w)}{\beta_n}. \]
3. For $v \in \mathcal{S}\setminus\{u\}$ and $j \in \mathbb{N}$,

$$Y^n_{v,j}(\cdot)$$

is a $(\mathbb{N}_0)_{\mathcal{S}}$-valued Markov process, with the same transition rates as $(X^{\mu_n}_x(\cdot))_{x \in \mathcal{S}}$ (defined in (5)) and with the initial condition $Y^n_{v,j}(0) = e_v$.

4. For $v, w \in \mathcal{S}\setminus\{u\}$ and $j \in \mathbb{N}$,

$$Y^n_{v,w,j}(\cdot)$$

is a $(\mathbb{N}_0)_{\mathcal{S}}$-valued Markov process, with the same transition rates as $(X^{\mu_n}_x(\cdot))_{x \in \mathcal{S}}$ and with the initial condition $Y^n_{v,w,j}(0) = e_v + e_w$.

5. For $v \in \mathcal{S}$,

$$(N_v(t))_{t \geq 0}$$

are Poisson counting processes with rate 1.

The random variables

$$[Z^n(\cdot), E^n_j, Y^n_{v,j}(\cdot), Y^n_{v,w,j}(\cdot), N_v(\cdot)]$$

are assumed to be independent ranging over $v, w, j$.

Let’s explain the meaning of the random variables introduced so far. $Z^n(\cdot)$ represents the ‘primary’ subpopulation - which we define as the type $u$ cells whose ancestors are all of type $u$. That is to say, there is an unbroken lineage of type $u$ cells between any primary cell and the initial cell. The rate, $\alpha_n$, that a primary cell gives birth to another primary cell, is simply the type $u$ division rate multiplied by the probability that no mutation occurs in either daughter cell. The rate, $\beta_n$, that a primary cell is removed, is the rate that a type $u$ cell dies plus the rate that a type $u$ cell divides to produce two mutant daughter cells.

The $E^n_j$ describe what happens at the $j$th downstep in the primary subpopulation trajectory. If $E^n_j = \emptyset$, then the downstep is a primary cell death. If $E^n_j = (v, w)$, then the downstep is a primary cell dividing to produce two mutant daughter cells of types $v$ and $w$.

Sometimes a primary cell divides to produce one primary cell and one mutant cell of type $v$. For the $j$th time that this occurs, $Y^n_{v,j}(t)$ is the vector which
counts the cells with each genome amongst the descendants of that type \( v \) cell, time \( t \) after its birth.

Sometimes a primary cell divides to produce two mutant cells of types \( v \) and \( w \). For the \( j \)th time that this occurs, \( Y_{v,w,j}^n(t) \) is the vector which counts the cells with each genome amongst the descendants of the two mutants time \( t \) after their birth.

The \( N_v(\cdot) \) will soon be rescaled in time to represent the times at which primary cells divide to produce one primary cell and one cell with genome \( v \).

The random variables introduced so far, seen together in (7), provide all the necessary ingredients for the construction of \((X_{v,\mu_n}^n, \sigma_n^\mu_n)_{v\in\mathcal{G}}\). Now we build upon these founding objects, defining further random variables.

6. For \( v \in \mathcal{G}\setminus\{u\} \) and \( t \geq 0 \),

\[
K_v^n(t) = N_v \left( 2p(u, v)\alpha(u') \int_0^t Z_n^m(s) ds \right). \quad (8)
\]

7. For \( j \in \mathbb{N} \) and \( v \in \mathcal{G}\setminus\{u\} \),

\[
T^n_{v,j} = \min \{ t \geq 0 : K_v^n(t) = j \}.
\]

8.

\[
S^n_1 = \min \{ t \geq 0 : Z^n(t) - Z^n(t^-) = -1 \},
\]

and then for \( j > 1 \), recursively,

\[
S^n_j = \min \{ t > S^n_{j-1} : Z^n(t) - Z^n(t^-) = -1 \}.
\]

(Here \( Z^n(t^-) := \lim_{s\uparrow t} Z^n(s) \).)

9. For \( v, w \in \mathcal{G}\setminus\{u\} \),

\[
T^n_{v,w,1} = \min \{ S^n_j : j \in \mathbb{N}, E^n_j = (v, w) \},
\]

and then for \( j > 1 \), recursively,

\[
T^n_{v,w,j} = \min \{ S^n_j : j \in \mathbb{N}, S^n_j > T^n_{v,w,j-1}, E^n_j = (v, w) \}.
\]
10. For \( v, w \in \mathcal{G}\setminus\{u\} \), and \( t \geq 0 \),

\[
K^n_{v,w}(t) = |\{j \in \mathbb{N} : T^n_{v,w,j} \leq t\}|
\]

Let’s explain the meaning of the new random variables. The \( K^n_v(t) \) specify the number of times before time \( t \) that primary cells have divided to produce one primary cell and one type \( v \) cell. Let’s check that this interpretation makes sense. Conditioned on the trajectory of \( Z^n(\cdot) \), \( K^n_v(\cdot) \) is certainly a Markov process, and increases by 1 at rate \( 2p_n(u,v)\alpha(u')Z^n(t) \) - i.e. the rate at which primary cells divide multiplied by the probability that exactly one daughter cell is primary and one is type \( v \).

\( S^n_j \) is the time of the \( j \)th downstep of the primary subpopulation size. Then \( T^n_{v,w,j} \) is the time of the \( j \)th primary cell division which produces cells of types \( v \) and \( w \). \( K^n_{v,w}(t) \) is the number of primary cell divisions before time \( t \) which produce cells of types \( v \) and \( w \).

At last the construction reaches its denouement.

11. For \( t \geq 0 \),

\[
\mathcal{X}^n(t) = Z^n(t)e_u + \sum_{v \in \mathcal{G}\setminus\{u\}} K^n_v(t) \sum_{j=1}^{K^n_v(t)} Y^n_{v,j}(t - T^n_{v,j}) + \sum_{v,w \in \mathcal{G}\setminus\{u\}} K^n_{v,w}(t) \sum_{j=1}^{K^n_{v,w}} Y^n_{v,w,j}(t - T^n_{v,w,j}).
\]

12. \( \sigma_n = \min\{t \geq 0 : |\mathcal{X}^n(t)| = n\} \),

where \( |\cdot| \) is the \( l_1 \)-norm on \( \mathbb{R}^\mathcal{G} \).

Upon reflection it should be clear that \( \mathcal{X}^n(\cdot) \) has the same distribution as \( (\mathcal{X}^{\mu_n}(\cdot))_{v \in \mathcal{G}} \). Both objects are Markov processes on \((\mathbb{N}_0)^\mathcal{G}\), whose initial conditions and transition rates coincide. Next we show that certain elements of the construction converge in distribution.
Lemma 6.2. As $n \to \infty$,
\[
(e^{-\lambda_n t} Z^n(t))_{t \in [0, \infty]} \to (e^{-\lambda t} Z^*(t))_{t \in [0, \infty]}
\]
in distribution, on the space $\mathbb{D}([0, \infty], \mathbb{R})$. Here
\[
\lambda_n := \alpha_n - \beta_n
\]
is the growth rate of the primary cell population,
\[
\lambda := \alpha(u') - \beta(u')
\]
is the large $n$ limit of $\lambda_n$, and $Z^*(\cdot)$ is a birth-death branching process with birth and death rates $\alpha(u')$ and $\beta(u')$.

Remark 6.3. The processes of Lemma 6.2 are defined on the timeline extended to include infinity. For $n$ large enough that $\lambda_n > 0$,
\[
e^{-\lambda_n \infty} Z^n(\infty) := \lim_{t \to \infty} e^{-\lambda_n t} Z^n(t) = W^n,
\]
and
\[
e^{-\lambda \infty} Z^*(\infty) := \lim_{t \to \infty} e^{-\lambda t} Z^*(t) = W^*.
\]
The limits $W^n$ and $W^*$ exist and are finite almost surely, which is a classic branching process result [2].

Proof of Lemma 6.2. Convergence in finite dimensional distributions is immediate. To show tightness we shall use Aldous’s criterion [1]. His result is for the time interval $[0, 1]$; so let’s identify $[0, 1]$ with $[0, \infty]$, by $t(s) = -\lambda_n^{-1} \log(1 - s)$ for $s \in [0, 1]$. We check tightness of the sequence of martingales $(M^n(s))_{s \in [0, 1]}$ defined by
\[
M^n(s) := e^{-\lambda_n t(s)} Z^n(t(s)) = (1 - s)Z^n(-\lambda_n^{-1} \log(1 - s)).
\]
Suppose that $(\rho_n)$ is a sequence of stopping times with respect to $(M^n(\cdot))$, and $(\delta_n)$ is a positive sequence converging to zero. Then, writing $\mathcal{F}_{\rho_n}$ for the sigma-algebra generated by $M^n(\cdot)$ up to time $\rho_n$,
\[
E[(M^n(\rho_n + \delta_n) - M^n(\rho_n))^2 | \mathcal{F}_{\rho_n}] = E[M^n(\rho_n + \delta_n)^2 | \mathcal{F}_{\rho_n}] - M^n(\rho_n)^2
\]
\[
= \delta_n \frac{\alpha_n + \beta_n}{\lambda_n} M^n(\rho_n),
\]
where the last equality comes thanks to the fact that
\[ M^n(s)^2 - \frac{\alpha_n + \beta_n}{\lambda_n} (1 - s) M^n(s) \]
is a martingale. Now,
\[
\mathbb{E}[(M^n(\rho_n + \delta_n) - M^n(\rho_n))^2] = \delta_n \frac{\alpha_n + \beta_n}{\lambda_n} \mathbb{E}M^n(\rho_n)
\]
\[ = \delta_n \frac{\alpha_n + \beta_n}{\lambda_n}. \]

Take \( n \to \infty \) to see that \( M^n(\rho_n + \delta_n) - M^n(\rho_n) \) converges to zero in \( L_2 \) and hence in probability, thus satisfying Aldous's criterion.

\[ \square \]

**Lemma 6.4.** As \( n \to \infty \),
\[ Y^n_{v,j}(\cdot) \to Y_{v,j}(\cdot) e^v \]
in distribution, where \( Y_{v,j}(\cdot) \) is a birth-death branching process with birth and death rates \( \alpha(v') \) and \( \beta(v') \) and initial condition \( Y_{v,j}(0) = 1 \). Convergence is on the space \( \mathbb{D}([0, \infty), \mathbb{R}^\mathbb{N}) \).

**Proof.** It is enough to note that the transition rates converge (see for example page 262 of [8]). \( \square \)

**Lemma 6.5.** As \( n \to \infty \),
\[ \left( \sum_{j \leq kn^{3/2}} 1_{\{E^n_j \neq \emptyset\}} \right) \to (0) \]
in distribution, on the space \( \mathbb{R}^\mathbb{N} \).

**Proof.** Note that
\[ \lim_{n \to \infty} n^{3/2} \mathbb{P}[E^n_j \neq \emptyset] = 0. \]
Then
\[
\mathbb{P} \left[ \sum_{j \leq kn^{3/2}} 1_{\{E^n_j \neq \emptyset\}} = 0; k = 1, \ldots, r \right] = \mathbb{P} \left[ E^n_j = \emptyset; j \leq rn^{3/2} \right] \]
\[ = (1 - \mathbb{P} \left[ E^n_j \neq \emptyset \right])^{rn^{3/2}} \]
\[ \to 1. \]
Remark 6.6. The number $3/2$ which appears in Lemma 6.5 is not special. It only matters that $3/2 \in (1, 2)$. The relevance of the result will be seen in Section 6.3.

We are yet to say how the random variables in (7) are jointly distributed over $n \in \mathbb{N}$. In fact, the choice of this joint distribution over $n \in \mathbb{N}$ has no relevance to the statement of Theorem 6.1. Hence the choice can be freely made, in a way that streamlines the proof. We assume that:

$$
\lim_{n \to \infty} (e^{-\lambda_n t} Z_n(t))_{t \in [0, \infty]} = (e^{-\lambda t} Z^*_t(t))_{t \in [0, \infty]} \tag{9}
$$

almost surely, on the space $\mathbb{D}([0, \infty], \mathbb{R})$;

$$
\lim_{n \to \infty} (Y_{v,j}^n(t))_{t \in [0, \infty)} = (Y_{v,j}(t)e_v)_{t \in [0, \infty)} \tag{10}
$$

almost surely, on the space $\mathbb{D}([0, \infty), \mathbb{R}^G)$, for $v \in \mathcal{G}\backslash\{u\}$ and $j \in \mathbb{N}$; and

$$
\left( \sum_{j \leq kn^{3/2}} 1_{\{E_j^n \neq \emptyset\}} \right)_{k \in \mathbb{N}} \to (0) \tag{11}
$$

almost surely, on the space $\mathbb{R}^\mathbb{N}$.

To justify that it is possible to have constructed the random variables in such a way that (9), (10), and (11) hold, one can bring in Skorokhod’s Representation Theorem, to use with the Lemmas 6.2, 6.4, and 6.5.

6.3 Neglecting double mutations

Call the event that a primary cell divides to produce two mutant cells a ‘double mutation’. Recall that double mutations are represented by the events $\{E_j^n = (v, w)\}$, which occur at the times $S_j^n$ when the primary cell population steps down in size. In order to comment on double mutations, we will first prove a rather crude upper bound for the number of downsteps in the primary cell population trajectory. Write

$$
\tau_n := \min\{t \geq 0 : Z^n(t) \in \{0, n\}\}, \quad \tag{12}
$$

26
for the time at which the primary cell population hits 0 or \( n \). Write
\[
D_n := \left| \{ j \in \mathbb{N} : S^n_j \leq \tau_n \} \right|
\]
for the number of downsteps in the primary cell population before time \( \tau_n \).

**Lemma 6.7.**
\[
\sup_{n \in \mathbb{N}} n^{-3/2} D_n < \infty
\]
almost surely.

**Proof.** For each \( n \in \mathbb{N} \), let \((R^n_j)_{j \in \mathbb{N}}\) be a sequence of i.i.d. random variables with
\[
\mathbb{P}[R^n_j = x] = \begin{cases} 
\frac{\alpha_n}{\alpha_n + \beta_n}, & x = 1; \\
\frac{\beta_n}{\alpha_n + \beta_n}, & x = -1;
\end{cases}
\]
so
\[
\left( 1 + \sum_{j=1}^{k} R^n_j \right)_{k \in \mathbb{N}}
\]
is a random walk, whose distribution matches the steps of \( Z^n(\cdot) \). Write
\[
\rho_n = \min \left\{ k \in \mathbb{N} : 1 + \sum_{j=1}^{k} R^n_j \in \{0, n\} \right\}
\]
for the number of steps until the walk hits \( n \) or 0. Then the number of downsteps before hitting \( n \) or 0 is
\[
D_n \overset{d}{=} \sum_{j=1}^{\rho_n} 1_{\{R^n_j = -1\}} \leq \rho_n.
\]
Therefore we can bound the tail of \( D_n \)'s distribution:
\[
\mathbb{P}[D_n > n^{3/2}] \leq \mathbb{P}[\rho_n > n^{3/2}].
\]
But \( \{ \rho_n > n^{3/2} \} \subset \{ 1 + \sum_{j=1}^{\lfloor n^{3/2} \rfloor} R^m_j < n \} \), so

\[
\mathbb{P}[D_n > n^{3/2}] \leq \mathbb{P} \left[ 1 + \sum_{j=1}^{\lfloor n^{3/2} \rfloor} R^m_j < n \right] \\
\leq \mathbb{P} \left[ \left( \sum_{j=1}^{\lfloor n^{3/2} \rfloor} R^m_j - \lfloor n^{3/2} \rfloor \lambda_n \right) > (\lfloor n^{3/2} \rfloor \lambda_n + 1 - n)^2 \right] \tag{13}
\leq (\lfloor n^{3/2} \rfloor \lambda_n + 1 - n)^{-2} \text{Var} \left[ \sum_{j=1}^{\lfloor n^{3/2} \rfloor} R^m_j \right] \tag{14}
\leq cn^{-3/2}, \tag{15}
\]

for some constant \( c > 0 \). Inequality (13) holds for large enough \( n \) and Inequality (14) is Chebyshev’s inequality. Finally, (15) gives that

\[
\sum_{n \in \mathbb{N}} \mathbb{P}[D_n > n^{3/2}] < \infty,
\]

and the result is proven by Borel-Cantelli. \( \square \)

Now it is to be seen that double mutations occurring before time \( \tau_n \) can be neglected.

**Lemma 6.8.** Let \( v, w \in \mathcal{G} \setminus \{u\} \). As \( n \to \infty \),

\[
K^n_{v,w}(\tau_n) \to 0
\]

almost surely.

**Proof.** From Lemma 6.7, \( C := \sup_{n \in \mathbb{N}} n^{-3/2} D_n < \infty \). Then

\[
K^n_{v,w}(\tau_n) = \sum_{j=1}^{D_n} 1\{E^n_j = (v, w)\} \\
\leq \sum_{j=1}^{\lfloor Cn^{3/2} \rfloor} 1\{E^n_j = (v, w)\} \\
\leq \sum_{j=1}^{\lfloor Cn^{3/2} \rfloor} 1\{E^n_j \neq \emptyset\}.
\]
By (11) this converges to zero as $n \to \infty$. \hfill \qed

### 6.4 Convergence of genome counts

The purpose of this section is to show that $X^n(\sigma_n)$ converges when conditioned on the event $\{W^* > 0\}$ ($W^*$ is defined in Remark 6.3). The times $\tau_n$ (defined in (12)) will play the role of a helpful stepping stone in the proof.

**Lemma 6.9.** Condition on $\{W^* > 0\}$. Then, almost surely,

1. there exists $n_0$ such that for all $n \geq n_0$, $Z^n(\tau_n) = n$; and
2. $\lim_{n \to \infty} \tau_n = \infty$.

**Proof.** To see the first statement, observe that there exists $n_0$ such that for all $n \geq n_0$, $W^n > W^*/2 > 0$. For such $n$, $\lim_{t \to \infty} Z^n(t) = \infty$, and hence $Z^n(\cdot) > 0$. To see the second statement, suppose for a contradiction that there exists a bounded subsequence $(\tau_{n_k}) \subset [0, C]$. Then, for large enough $k$,

$$n_k = Z^{n_k}(\tau_{n_k}) \leq \sup_{n \in \mathbb{N}} \sup_{t \in [0, C]} Z^n(t).$$

The left hand side of the inequality is unbounded over $k$. On the other hand, the right hand side, which does not depend on $k$, is finite thanks to (9). \hfill \qed

**Lemma 6.10.** Condition on $\{W^* > 0\}$. Suppose that $(a_n)$ is a real sequence which converges to infinity, with

$$a_n \leq \tau_n$$

and

$$\lim_{n \to \infty} (a_n - \tau_n) = l \in [-\infty, 0]$$

almost surely. Then for each $t \in \mathbb{R}$, almost surely,

$$K^n_v(a_n + t) \to \begin{cases} K^*_v(l + t), & v \in \mathcal{G}_i; \\ 0, & v \in \mathcal{G}_{\geq 2}; \end{cases}$$

where

$$K^*_v(s) = N_v(2\lambda^{-1} \alpha(u') \theta_i^{u_i, v_i} e^{\lambda s}).$$
Proof. Thanks to (9):

1. for any sequence \((t_n)\) which converges to infinity, \(\lim_{n \to \infty} e^{-\lambda_n t_n} Z^n(t_n) = W^*\); and

2. \(\sup_{n \in \mathbb{N}} \sup_{t \in [0, \infty]} e^{-\lambda_n t} Z^n(t) < \infty\).

Therefore, for \(t \in \mathbb{R}\),

\[
\frac{1}{n} \int_0^{a_n + t} Z^n(s) ds = \int_{-a_n}^t \frac{Z^n(a_n + s)}{e^{\lambda_n (a_n + s)}} e^{\lambda_n (a_n - \tau_n + s)} ds
\]

\[
\to \int_{-\infty}^t e^{\lambda (l + s)} ds = \lambda^{-1} e^{\lambda (l + t)}
\]

as \(n \to \infty\), by dominated convergence. Note also that

\[
\lim_{n \to \infty} n p_n(u, v) = \begin{cases} \theta_i u, & v \in S_i; \\ 0, & v \in S_{\geq 2}. \end{cases}
\]

Then the result follows straight from the definition of \(K^n_v(\cdot)\) in (8). \(\square\)

**Lemma 6.11.** Condition on \(\{W^* > 0\}\). Suppose that \((a_n)\) satisfies the conditions of Lemma 6.10. Then, almost surely,

\[
\lim_{n \to \infty} (X^n(a_n) - Z^n(a_n)) e_u = \sum_{i \in S} \sum_{v \in S_i} e_v \sum_{j=1}^{K^n_v(l)} Y^v_{v,j}(l - T^*_v,j),
\]

where the \(Y^v_{v,j}(\cdot)\) are from Lemma 6.4, the \(K^n_v(\cdot)\) are from Lemma 6.10, and \(T^*_v,j = \min\{t \in \mathbb{R} : K^n_v(t) = j\}\).

Proof. Recall that

\[
X^n(a_n) - Z^n(a_n) e_u = \sum_{v \in S \setminus \{u\}}^{K^n_{v,w}(a_n)} \sum_{j=1}^{K^n_{v,w}(a_n)} Y^n_{v,j}(a_n - T^n_{v,j})
\]

\[
+ \sum_{v,w \in S \setminus \{u\}}^{K^n_{v,w}(a_n)} \sum_{j=1}^{K^n_{v,w}(a_n)} Y^n_{v,w,j}(a_n - T^n_{v,w,j}).
\]
The ‘double mutation’ term in (16) converges to zero, because
\[ K_{n,v,w}^n(a_n) \leq K_{n,v,w}^n(\tau_n), \]
which converges to zero by Lemma 6.8. The ‘single mutation’ term in (16) converges to the required limit due to Lemma 6.10 and (10).

**Lemma 6.12.** Condition on \( \{W^* > 0\} \).

\[ \lim_{n \to \infty} (\sigma_n - \tau_n) = 0 \]
almost surely.

**Proof.** By Lemma 6.9, for large enough \( n \), \( Z^n(\tau_n) = n \). So \( |X^n(\tau_n)| \geq n \), and hence \( \sigma_n \leq \tau_n \). Therefore
\[ \liminf_{n \to \infty} (\sigma_n - \tau_n) \leq 0. \]

Suppose, looking for a contradiction, that
\[ \liminf_{n \to \infty} (\sigma_n - \tau_n) = l \in [-\infty, 0). \]

Take a subsequence with
\[ \lim_{k \to \infty} (\sigma_{n_k} - \tau_{n_k}) = l. \]

Then, by Lemma 6.11,
\[ |X^n(\sigma_{n_k}) - Z^n(\sigma_{n_k})e_a| \]
converges, and so must be a bounded sequence. However it is also true that, taking \( k \to \infty \),
\[ |X^n(\sigma_{n_k}) - Z^n(\sigma_{n_k})e_a| = n_k - Z^n(\sigma_{n_k})e_a \]
\[ = n_k \left( 1 - \frac{Z^n(\sigma_{n_k})e^{\lambda_{n_k} \tau_{n_k}}}{e^{\lambda_{n_k} \sigma_{n_k}}} \right) \]
\[ \sim n_k (1 - e^{\lambda_l}), \]
which is unbounded. \( \square \)
Lemma 6.13. Condition on \( \{ W^* > 0 \} \).

\[
\lim_{n \to \infty} (X^n(\sigma_n) - Z^n(\sigma_n)e_u) = \sum_{i \in S} \sum_{v \in S_i} e_v \sum_{j=1}^{K^*_v(0)} Y_{v,j}(-T^*_v),
\]

almost surely.

Proof. Combine Lemmas 6.11 and 6.12. \qed

The limit of Lemma 6.13 is a vector of independent generalised Luria-Delbrück distributions:

\[
\sum_{i \in S} \sum_{v \in S_i} e_v \sum_{j=1}^{K^*_v(0)} Y_{v,j}(-T^*_v) \overset{d}{=} (X_v)_{v \in S \setminus \{u\}},
\]

where the \( X_v \) are as stated in Theorem 6.1. To complete the proof of Theorem 6.1 we need to show that conditioning on \( \{ W^* > 0 \} \) can be translated to conditioning on \( \{ \sigma_n < \infty \} \), which is the subject of the next subsection.

### 6.5 Conditioning on reaching \( n \) cells

In order to connect \( \{ W^* > 0 \} \) and \( \{ \sigma_n < \infty \} \), the next result is the key. It states that these events are approximately the same for large \( n \).

Proposition 6.14.

1. \( \lim_{n \to \infty} \mathbb{P}[W^* > 0, \sigma_n = \infty] = 0 \), and
2. \( \lim_{n \to \infty} \mathbb{P}[W^* = 0, \sigma_n < \infty] = 0 \).

Let’s break the proof of Proposition 6.14 into several lemmas; the idea is that the random variable \( W^n \) be used as an intermediary.

Lemma 6.15.

\[
\lim_{n \to \infty} \mathbb{P}[W^* > 0, W^n = 0] = 0.
\]
Proof. If $W^* > 0$, then there exists $n_0$, such that for all $n \geq n_0$

$$W^n > \frac{W^*}{2}.$$ 

So

$$\lim_{n \to \infty} 1_{\{W^* > 0, W^n = 0\}} = 0.$$ 

Therefore, by dominated convergence,

$$\mathbb{P}[W^* > 0, W^n = 0] = \mathbb{E}1_{\{W^* > 0, W^n = 0\}} \to 0.$$ 

Lemma 6.16.

$$\mathbb{P}[W^n > 0, \sigma_n = \infty] = 0.$$ 

Proof. If $W^n > 0$, then $\lim_{t \to \infty} X^n(t) = \infty$, and so $\sigma_n < \infty$. 

Proof of Part 1 of Proposition 6.14.

$$\mathbb{P}[W^* > 0, \sigma_n = \infty] = \mathbb{P}[W^* > 0, \sigma_n = \infty, W^n = 0] + \mathbb{P}[W^* > 0, \sigma_n = \infty, W^n > 0] \leq \mathbb{P}[W^* > 0, W^n = 0] + \mathbb{P}[\sigma_n = \infty, W^n > 0] \to 0$$

as $n \to \infty$, by Lemmas 6.15 and 6.16. 

The structure for the proof of Part 2 of Proposition 6.14 is much the same as that of Part 1. However the details will require a little extra work.

Lemma 6.17.

$$\lim_{n \to \infty} \mathbb{P}[W^* = 0, W^n > 0].$$ 

Proof. Let $\epsilon > 0$. If $W^* = 0$, then there exists $n_0$ such that for all $n \geq n_0$

$$W^n < \epsilon.$$
So
\[ \lim_{n \to \infty} 1\{ W^* = 0, W^n \geq \epsilon \} = 0. \]

Note that
\[ 1\{ W^* = 0, W^n > 0 \} = 1\{ W^* = 0, W^n \in (0, \epsilon) \} + 1\{ W^* = 0, W^n \geq \epsilon \}, \]
and that
\[ \sup_{n \in \mathbb{N}} \mathbb{P}[W^n \in (0, \epsilon)] \leq C\epsilon, \]
for some \( C > 0 \). Now, using dominated convergence,
\[ \lim_{n \to \infty} \sup \mathbb{P}[W^* = 0, W^n > 0] \leq C\epsilon. \]

But \( \epsilon > 0 \) was arbitrary, giving the result. \( \square \)

**Lemma 6.18.**
\[ \lim_{n \to \infty} \mathbb{P}[W^n = 0, \sigma_n < \infty] = 0. \]

**Proof.** If the primary population size never reaches \( n \) and there are never any mutations, then the total population size never reaches \( n \). That is, if \( Z^n(\tau_n) = 0, K^n_v(\cdot) = 0 \) and \( K^n_{v,w}(\cdot) = 0 \) for all \( v, w \in \mathcal{G} \setminus \{u\} \), then
\[ \sup_{t \geq 0} |X^n(t)| < n, \]
which means that \( \sigma_n = \infty \). Equivalently,
\[
\{\sigma_n < \infty\} \subset \{Z^n(\tau_n) = n\} \cup \{\exists v, K^n_v(\cdot) \neq 0\} \cup \{\exists (v, w), K^n_{v,w}(\cdot) \neq 0\} \\
= \{Z^n(\tau_n) = n\} \cup \{\exists v, K^n_v(\cdot) \neq 0\} \\
\cup \{\exists (v, w), K^n_{v,w}(\cdot) \neq 0, Z^n(\tau_n) = 0\}. 
\]

It follows that
\[
\mathbb{P}[W^n = 0, \sigma_n < \infty] \leq \mathbb{P}[W^n = 0 | Z^n(\tau_n) = n] \\
+ \sum_{v \in \mathcal{G} \setminus \{u\}} \mathbb{P}[K^n_v(\cdot) \neq 0 | W^n = 0] \\
+ \sum_{v, w \in \mathcal{G} \setminus \{u\}} \mathbb{P}[K^n_{v,w}(\cdot) \neq 0 | Z^n(\tau_n) = 0]. \tag{17}
\]
We will show that each term of the right hand side of Inequality (17) converges to zero. Firstly,
\[
P[W^n = 0 | Z^n(\tau_n) = n] = \left( \frac{\beta_n}{\alpha_n} \right)^n,
\]
which is the probability that \( Z^n(\cdot) \), if starting at size \( n \), eventually goes extinct; this clearly converges to zero.

Secondly,
\[
P \left[ \sup_t K_v^n(t) | W^n = 0 \right] = E \left[ N_v^n \left( 2p_n(u,v) \alpha(u') \int_0^\infty Z^n(s) ds \right) | W^n = 0 \right]
\]
\[
= E \left[ N_v^n \left( 2p_n(u,v) \alpha(u') \int_0^\infty Z^n(s) ds \right) | Z^n(\cdot) \right] | W^n = 0 \right] = 2p_n(u,v) \alpha(u') \int_0^\infty Z^n(s) ds | W^n = 0 \right] \to 0,
\]
because \( p_n(u,v) \to 0 \). Hence
\[
P \left[ \sup_t K_v^n(t) \neq 0 | W^n = 0 \right] \to 0.
\]

Lastly,
\[
P[K_v^n(\cdot) \neq 0 | Z^n(\tau_n) = 0] = \frac{P[K_v^n(\tau_n) \neq 0 | Z^n(\tau_n) = 0]}{P[Z^n(\tau_n) = 0]}.
\]
But \( P[K_v^n(\tau_n) \neq 0] \) converges to zero by Lemma 6.8, while \( P[Z^n(\tau_n) = 0] \) converges to \( P[W^* = 0] > 0 \) by (9).
Proof of Part 2 of Proposition 6.14. Just as for Part 1,
\[
\mathbb{P}[W^* = 0, \sigma_n < \infty] = \mathbb{P}[W^* = 0, \sigma_n < \infty, W^n > 0] + \mathbb{P}[W^* = 0, \sigma_n < \infty, W^n = 0] \\ 
\leq \mathbb{P}[W^* = 0, W^n < 0] + \mathbb{P}[\sigma_n < \infty, W^n = 0] \to 0
\]
as \( n \to \infty \), by Lemmas 6.17 and 6.18.

Corollary 6.19 (to Proposition 6.14). For any sequence of events \((H_n)_{n \in \mathbb{N}}\),
\[
\lim_{n \to \infty} \mathbb{P}[H_n, \sigma_n < \infty] = \lim_{n \to \infty} \mathbb{P}[H_n, W^* > 0]
\]
if the limit exists.

Proof. Partition the event \(\{H_n \cap (W^* > 0 \cup \sigma_n < \infty)\}\) in two ways to obtain
\[
\mathbb{P}[H_n, \sigma_n < \infty] + \mathbb{P}[H_n, W^* > 0, \sigma_n = \infty] \\ 
= \mathbb{P}[H_n, W^* > 0] + \mathbb{P}[H_n, W^* = 0, \sigma_n < \infty],
\]
and take \( n \to \infty \).

Finally we are in a position to prove Theorem 6.1.

Proof of Theorem 6.1. For any \( R \subset (\mathbb{N}_0 \cup \{\infty\})^g \),
\[
\lim_{n \to \infty} \frac{\mathbb{P}[\chi^n(\sigma_n) \in R, \sigma_n < \infty]}{\mathbb{P}[\sigma_n < \infty]} = \lim_{n \to \infty} \frac{\mathbb{P}[\chi^n(\sigma_n) \in R, W^* > 0]}{\mathbb{P}[W^* > 0]}
\]
(18)
\[
= \mathbb{P}[(X_v)_{v \in g} \in R],
\]
(19)
where (18) is due to Corollary 6.19 and (19) is due to Lemma 6.13.
6.6 Law of large numbers

**Proof of Theorem 5.2 (and Theorem 4.3).** The expected site frequency spectrum is given by

\[ E \left[ |S|^{-1} \sum_{i \in S} \delta_{B_i^{n,\mu}} \{k\} \right] = |S|^{-1} \sum_{i \in S} P[B_i^{n,\mu} = k] \]

\[ = |S|^{-1} \sum_{i \in S} P[B_i^{n,\mu} = k] \]

\[ + |S|^{-1} \sum_{j \in J} \sum_{i \in S(j)} P[B_i^{n,\mu} = k], \]

for \( k \in \mathbb{N}_0 \). Theorem 5.1 says that \( P[B_i^{n,\mu} = k] \to P[B(j) = k] \), while \( |S_{sel}|/|S| \to 0 \) and \( |S(j)|/|S| \to q(j) \). Hence

\[ E \left[ |S|^{-1} \sum_{i \in S} \delta_{B_i^{n,\mu}} \{k\} \right] \to \sum_{j \in J} q(j)P[B(j) = k]. \]

The variance is

\[ \text{Var} \left[ |S|^{-1} \sum_{i \in S} \delta_{B_i^{n,\mu}} \{k\} \right] = |S|^{-2} \sum_{i \in S} \text{Var}[1_{\{B_i^{n,\mu} = k\}}] \]

\[ + |S|^{-2} \sum_{i,j \in S \atop i \neq j} \text{Cov}[1_{\{B_i^{n,\mu} = k\}}, 1_{\{B_j^{n,\mu} = k\}}] \]

\[ \leq |S|^{-1} + \max_{i,j \in S \atop i \neq j} \text{Cov}[1_{\{B_i^{n,\mu} = k\}}, 1_{\{B_j^{n,\mu} = k\}}]. \]

Because \( S_{sel} \) and \( J \) are fixed sets and the random variables are exchangeable over \( S(j) \), the maximum is taken over a fixed set. Theorem 5.1 says that the covariances converge to zero.

\[ \square \]

7 Proofs: mutations at positive fractions

In this section we return to the basic Yule process setting, proving results on mutations present in a positive fraction of cells. In Subsection 7.1 we prove
Theorem 4.4 and also prove an upper bound for mutation frequencies. In 7.2 we prove Lemma 3.2 and another result concerning cell descendant fractions. In 7.3 we determine mutation frequencies under the infinite sites assumption. In 7.4 we show that the infinite sites assumption is accurate, concluding the proof of Theorem 4.6. In 7.5 we give details of Remarks 4.7 and 4.8.

7.1 Single site mutation frequencies and an upper bound

Write
\[ B_i^\mu = \{ x \in T : V_i^\mu(x) \neq u_i \} \]

for the cells which are mutated at site \( i \in S \), and
\[ \tilde{B}_i^\mu = \{ y \in T : \exists x \in B_i^\mu, x \preceq y \} \]

for their descendants. Recall that \( T_{\sigma_n} \) are the cells alive when the total number of cells reaches \( n \). Note the inequality
\[ \tilde{B}_n^{\mu,a} := |\tilde{B}_n^\mu \cap T_{\sigma_n}| \geq |B_n^\mu \cap T_{\sigma_n}| = B_n^{\mu,a}. \] (20)

The goal of this subsection is to prove Theorem 4.4 and the following closely related result, which will later play a crucial role in the proof of Theorem 4.6.

**Proposition 7.1.** Let \( i \in S \) and \( a \in (0, 1) \). As \( \mu \to 0 \) and \( n\mu \to \theta < \infty \),
\[ \mu^{-1}\mathbb{P}[n^{-1}	ilde{B}_n^{\mu,a} > a] \to 2(a^{-1} - 1). \]

For this subsection we are always talking about a single site \( i \in S \); for convenience, let’s drop the subscript \( i \) from the notation. To begin the proofs of Theorem 4.4 and Proposition 7.1, observe that \((B_n^{\mu})_{n \in \mathbb{N}}\) and \((\tilde{B}_n^{\mu})_{n \in \mathbb{N}}\) are Markov processes on the nonnegative integers with transition probabilities
\[ \mathbb{P}[B_{n+1}^{\mu} = k | B_n^{\mu} = j] \]
\[ = \begin{cases} \frac{j}{n}(\mu/3)^2, & k = j - 1; \\ \frac{j}{n}2\mu(1 - \mu/3) + \frac{n-j}{n}(1 - \mu)^2, & k = j; \\ \frac{j}{n}(1 - \mu/3)^2 + \frac{n-j}{n}2\mu(1 - \mu), & k = j + 1; \\ \frac{n-j}{n}\mu^2, & k = j + 2. \end{cases} \] (21)
\[ P[\hat{B}_{n+1,\mu} = k | \hat{B}^{\mu} = j] = \begin{cases} \frac{2 \mu^j n^{j-1}}{n^2} (1 - \mu)^2, & k = j; \\ \frac{2 n^{j-1} \mu^2 (1 - \mu)}{n^2}, & k = j + 1; \\ \frac{n^{j-1} \mu^2}{n^2}, & k = j + 2. \end{cases} \] (22)

The key idea of the proof will be to condition on the number of cells when the first mutant (with respect to site \( i \)) arises. For this purpose, introduce

\[ \xi^\mu = \min\{n \in \mathbb{N} : B^{\mu} > 0\} \]

for the total number of cells when the first mutant cell arises. Let

\[ \Xi_k^\mu = \{B^{\xi^\mu} = k\} \]

be the event that the first cell division to see a mutation gives \( k \) mutant cells, for \( k \in \{1, 2\} \).

**Lemma 7.2.** As \( \mu \to 0 \),

\[ \mu^{-1} P[\xi^\mu = j, \Xi_k^\mu] = \begin{cases} 2, & k = 1; \\ 0, & k = 2 \end{cases} \]

**Proof.** The probability that the first \( j - 2 \) cell divisions give no site \( i \) mutations multiplied by the probability that the \((j - 1)\)th cell division gives exactly one mutant daughter is

\[ P[\xi^\mu = j, \Xi_1^\mu] = (1 - \mu)^{2j-3} 2 \mu. \]

Similarly

\[ P[\xi^\mu = j, \Xi_2^\mu] = (1 - \mu)^{2j-4} \mu^2. \]

Divide by \( \mu \) and take \( \mu \to 0 \). \( \square \)

The next result gives conditional mutation frequencies.

**Lemma 7.3.** Let \( a > 0 \). As \( \mu \to 0 \) and \( n\mu \to \theta < \infty \),

\[ P[n^{-1} B^{n,\mu} > a | \xi^\mu = j, \Xi_1^\mu] \to (1 - a)^j \]

and

\[ P[n^{-1} \hat{B}^{n,\mu} > a | \xi^\mu = j, \Xi_1^\mu] \to (1 - a)^j \]

39
Proof. All random variables in this proof are conditioned on the event \( \{ \xi^\mu = j, \Xi_1^\mu \} \) (which for brevity we won’t continue to write). In particular, the processes \((B^r,\mu)_{r \geq j}\) and \((\hat{B}^r,\mu)_{r \geq j}\) have initial conditions \(B^j,\mu = \hat{B}^j,\mu = 1\) and transition probabilities (21) and (22). Directly calculating from the transition probabilities, for \( r \geq j \),

\[
\mathbb{E}\left[ r^{-1}B^r,\mu \right] \leq \mathbb{E}\left[ (r + 1)^{-1}B^{r+1},\mu \right] \\
\leq \mathbb{E}\left[ r^{-1}B^r,\mu \right] + \frac{2\mu(1 + \mu)}{r + 1}.
\]

(23)

Combining (23) with

\[
\mathbb{E}\left[ j^{-1}B^j,\mu \right] = j^{-1},
\]

we have that, for \( n \geq j \),

\[
\frac{1}{j} \leq \mathbb{E}\left[ n^{-1}B^n,\mu \right] \leq \frac{1}{j} + 2\mu(1 - \mu) \sum_{r=j+1}^{n} r^{-1}.
\]

Therefore, as \( \mu \to 0 \) and \( n\mu \to \theta \),

\[
\mathbb{E}\left[ n^{-1}B^n,\mu \right] \to j^{-1}.
\]

In just the same manner,

\[
\mathbb{E}\left[ n^{-1}\hat{B}^n,\mu \right] \to j^{-1}.
\]

Consider the single mutant cell present when the total number of cells is \( j \). Write \( D^n \) for the number of cells which have descended from this mutant cell when the total number of cells is \( n \geq j \). The process \((D^n, n - D^n)_{n \geq j}\) is just Polya’s urn. Clearly

\[
\mathbb{E}[n^{-1}D^n] = j^{-1}.
\]

Moreover a well-known result (e.g. [7]) says that, as \( n \to \infty \), \( n^{-1}D^n \) converges to a Beta random variable. That is, for \( a \in (0, 1) \),

\[
\mathbb{P}[n^{-1}D^n > a] \to (1 - a)^{j-1},
\]

(24)

which is exactly the limit we wish to show for \( n^{-1}B^n,\mu \) and \( n^{-1}\hat{B}^n,\mu \). To show that \( n^{-1}D^n \), \( n^{-1}B^n,\mu \), and \( n^{-1}\hat{B}^n,\mu \) share the same limiting distribution,
we will show that their differences converge to zero in expectation. The inequality
\[ D^n \leq \hat{B}^{n,\mu} \]
gives that
\[ E|\hat{B}^{n,\mu} - D^n| = E n^{-1} \hat{B}^{n,\mu} - E n^{-1} D^n \to 0. \]
The inequality
\[ B^{n,\mu} \leq \hat{B}^{n,\mu} \]
gives that
\[ E|n^{-1} \hat{B}^{n,\mu} - n^{-1} B^{n,\mu}| = E n^{-1} \hat{B}^{n,\mu} - E n^{-1} B^{n,\mu} \to 0. \]

Lemma 7.4. Consider \((\mu_n)_{n\in\mathbb{N}}\) with \(n \mu_n \to \theta < \infty\). Then
\[ \sup_{n \in \mathbb{N}} P[n^{-1} B^{n,\mu_n} > a|\xi^{\mu_n} = j, \Xi^{\mu_n}_k] \leq c j^{-2}, \]
where \(c > 0\) does not depend on \(j,k\).

Proof. The first inequality is immediate. We prove the second. In this proof we will always condition on \(\{\xi^{\mu_n} = j, \Xi^{\mu_n}_k\}\), although won’t write it. Directly calculating from the transition probabilities (22), for \(r \in \{j,..,n\}\),
\[ E[(\hat{B}^{r+1,\mu_n}/(r + 1))^2] - E[(\hat{B}^{r,\mu_n}/r)^2] \leq r^{-2} b, \]
where \(b > 0\) is some constant independent of \(n,r,j,k\). This leads to, again for \(r \in \{2,..,n\}\),
\[ E[(\hat{B}^{r,\mu_n}/r)^2] \leq E[(\hat{B}^{j,\mu_n}/j)^2] + b \sum_{k=j}^{r-1} k^{-2} \]
\[ = (k/j)^2 + b \sum_{k=j}^{r-1} k^{-2} \]
\[ \leq b' j^{-1}, \]
where $b'$ is another constant. (In the following, $b''$ and $b'''$ will also be constants.) A similar calculation for third moments gives

$$
\mathbb{E}[(\hat{B}^{r,\mu_n}/r)^3] - \mathbb{E}[(\hat{B}^{r,\mu_n}/r)^2] \leq b'' r^{-4} \mathbb{E}[(\hat{B}^{r,\mu_n}/r)^2] \leq b''' r^{-2} j^{-1},
$$

from which it follows that

$$
\mathbb{E}[(\hat{B}^{n,\mu_n}/n)^3] = \mathbb{E}[(\hat{B}^{j,\mu_n}/j)^3] + b''' j^{-1} \sum_{r=j}^{n-1} r^{-2} \\
\leq j^{-3} + b''' j^{-2} \\
\leq c j^{-2}.
$$

Apply Markov’s inequality to conclude. \hfill \Box

**Proof of Theorem 4.4 and Proposition 7.1.** For $a \in (0,1)$,

$$
\mu^{-1} \mathbb{P}[n^{-1} B^{n,\mu} > a] = \sum_{j=2}^{\infty} \sum_{k=1}^{2} \mu^{-1} \mathbb{P}[\xi = j, \Xi_k] \mathbb{P}[n^{-1} B^{n,\mu} > a | \xi = j, \Xi_k]. \quad (25)
$$

Lemmas 7.2, 7.3, and 7.4, along with the Dominated Convergence Theorem, show that the limit of (25) is

$$
2 \sum_{j=1}^{\infty} (1-a)^j = 2(a^{-1} - 1).
$$

The same argument works for $\hat{B}^{n,\mu}$. \hfill \Box

### 7.2 Cell descendant fractions

Here we are concerned with the $P_{x,t}$ (the fraction of cells alive at time $t \geq 0$ which are descendants of cell $x \in J$).

Lemma 3.2 shows that $\lim_{t \to \infty} P_{x,t} = P_x$ almost surely and gives the distribution of the $P_x$. Those readers who are familiar with branching processes...
may regard the result for a single \( x \in \mathcal{T}\setminus\{\emptyset\} \) as trivial. The joint distribution over \( x \in \mathcal{T}\setminus\{\emptyset\} \), however, is less obvious. It is surely already known, but we do not know of a reference. Now we give a proof.

**Proof of Lemma 3.2.** Write

\[
D_x = \{ y \in \mathcal{T} : x \preceq y \}
\]

for the descendants of cell \( x \in \mathcal{T} \), and write

\[
D_{x,t} = D_x \cap \mathcal{T}_t
\]

for the descendants of cell \( x \in \mathcal{T} \) which are alive at time \( t \geq 0 \). Observe that

\[
D_{x,\sum_{y \prec x} A_y + t} = \left\{ y \in \mathcal{T} : x \preceq y, \sum_{x \preceq z \preceq y} A_z \leq t < \sum_{x \preceq z \preceq y} A_z \right\}.
\]

Hence

\[
\left( |D_{x,\sum_{y \prec x} A_y + t}| \right)_{t \geq 0}
\]

is measurable with respect to the sigma-algebra generated by \((A_y)_{y \in D_x}\), and has the same distribution as

\[
\left( |D_{\emptyset,t}| \right)_{t \geq 0} = \left( |\mathcal{T}_t| \right)_{t \geq 0}.
\]

It follows that

\[
limit_{t \to \infty} e^{-t} |D_{x,\sum_{y \prec x} A_y + t}| =: W_x
\]

almost surely, where \( W_x \sim \text{Exp}(1) \); moreover if \( x, y \in \mathcal{T} \) are such that \( D_x \cap D_y = \emptyset \), then \( W_x \) and \( W_y \) are independent. In particular, \( W_{x0} \) and \( W_{x1} \) are independent. Now,

\[
\lim_{t \to \infty} \frac{|D_{x0,t}|}{|D_{x,t}|} = \lim_{t \to \infty} \frac{|D_{x0,t}|}{1 + |D_{x0,t}| + |D_{x1,t}|} = \frac{W_{x0}}{W_{x0} + W_{x1}} =: U_{x0}
\]

almost surely, and \( U_{x0} + U_{x1} = 1 \). A standard calculation shows that \( U_{x0} \) is uniformly distributed on \((0,1)\): for \( u \in (0,1) \),

\[
P[U_{x0} < u] = \int_0^\infty \int_0^\infty e^{-y}e^{-z}dydz = u.
\]
It remains to show independence of the $U_{x0}$. Another standard calculation shows that

$$U_{x0} = \frac{W_{x0}}{W_{x0} + W_{x1}}$$

is independent of

$$W_{x0} + W_{x1} :$$

for $(u, v) \in (0, 1) \times (0, \infty)$,

$$\mathbb{P}[U_{x0} < u, W_{x0} + W_{x1} < v] = \int_0^u \int_{z(1-u)/u}^{v-z} e^{-y} e^{-z} dy dz = u(1 - (1 + v)e^{-v}) = \mathbb{P}[U_{x0} < u] \mathbb{P}[W_{x0} + W_{x1} < v].$$

Now fix $l \in \mathbb{N}$. Because $U_{x0}$ and $W_{x0} + W_{x1}$ are measurable with respect to the sigma-algebra generated by $(A_{y})_{y \in \mathcal{D}_x \{x\}}$, we have that

$$[(U_{x0})_{|x|=l}, (W_{x0} + W_{x1})_{|x|=l}, (A_{x})_{|x|\leq l}]$$

forms an independent family of random variables.

Finally we complete the proof by induction. Suppose that $(U_{x0})_{x \in \mathcal{T}; |x| < l}$ is an independent family. Observing that for any $x \in \mathcal{T}$,

$$W_x = e^{-A_x}(W_{x0} + W_{x1}),$$

we have that $(U_{x0})_{x \in \mathcal{T}; |x| < l}$ is measurable with respect to the sigma-algebra generated by

$$[(W_{x0} + W_{x1})_{|x|=l}, (A_{x})_{|x|\leq l}].$$

Then, thanks to the independence of (26), $(U_{x0})_{x \in \mathcal{T}; |x| \leq l}$ forms an independent family.

Next comes a technical result whose value will become apparent in the next subsection.

**Lemma 7.5.** Let $\epsilon \in (0, 1)$. The set

$$\{x \in \mathcal{T}\{\emptyset\} : \exists t \geq 0, P_{x,t} > \epsilon\}$$

is almost surely finite.
Proof. Let $x \in \mathcal{T} \setminus \{\emptyset\}$. Note that $(P_{x,t})_{t \geq 0}$ is a submartingale and that $\mathbb{E}[(P_x)^2] = 3^{-|x|}$, where $|x|$ is the generation of $x$ (that is, $x \in \{0,1\}^{|x|}$). Then, by Doob’s inequality,

$$\mathbb{P}[\exists t \geq 0, P_{x,t} > a] = \mathbb{P}[(\sup_{t \geq 0} P_{x,t} > \epsilon) \leq \epsilon^2 3^{-|x|}].$$

Now

$$\sum_{x \in \mathcal{T} \setminus \{\emptyset\}} \mathbb{P}[\exists t \geq 0, P_{x,t} > \epsilon] \leq \sum_{l \in \mathbb{N}} \epsilon^{-2} (2/3)^l \times \infty,$$

so the Borel-Cantelli lemma concludes the proof. \( \square \)

7.3 Mutation frequencies under the infinite sites assumption

Enumerate the elements of $\mathcal{T}$,

$$\mathcal{T} = (x_k)_{k \in \mathbb{N}},$$

in such a way that

$$x_j < x_k \implies j < k. \quad (27)$$

Let’s give an example of such an enumeration: map $(x(r))_{r=1}^l \in \{0,1\}^l \subset \mathcal{T}$ to $2^l + \sum_{r=1}^l 2^{l-r} x(r)$.

Assuming a mutation rate $\mu$, write

$$\phi_i^\mu = \min\{x \in \mathcal{T} : V_i^\mu(x) \neq u_i\}$$

for the first cell (with respect to the enumeration) which sees a mutation at site $i \in S$.

Remark 7.6. $\phi_i^\mu$ has geometric distribution:

$$\mathbb{P}[\phi_i^\mu = x_k] = \mu(1 - \mu)^{k-1}.$$

In this subsection we are concerned with $P_{\phi_i^\mu, \sigma_n}$, which is the fraction of cells alive at time $\sigma_n$ (when $n$ total cells are reached) which are descendants of
cell $\phi_i^\mu$. Phrased another way, $P_{\phi_i^\mu, \sigma_n}$ is the fraction of cells alive at time $\sigma_n$ which are mutated at site $i$ under the infinite sites assumption.

Next we give an infinite-sites analog of Theorem 4.6.

**Proposition 7.7.** As $n \to \infty$, $\mu \to 0$, and $|S|\mu \to \eta < \infty$,

$$\sum_{i \in S} \delta_{P_{\phi_i^\mu, \sigma_n}} \to \sum_{x \in \mathcal{T}\setminus\{\emptyset\}} M_x \delta_{P_x}$$

in distribution. Convergence is on the space of measures on $[\epsilon, 1]$ equipped with the vague topology, for arbitrary $\epsilon \in (0, 1)$.

The proof of Proposition 7.7 will require us to count the number of sites which see their first mutation at cell $x \in \mathcal{T}\setminus\{\emptyset\}$ (with respect to the enumeration); write

$$M_{x}^{\mu, S} = |\{i \in S : \phi_i^\mu = x\}|.$$

**Lemma 7.8.** As $\mu \to 0$ and $|S|\mu \to \eta < \infty$,

$$(M_{x}^{\mu, S})_{x \in \mathcal{T}\setminus\{\emptyset\}} \to (M_x)_{x \in \mathcal{T}\setminus\{\emptyset\}}$$

in distribution, where the $M_x$ are i.i.d. Poisson($\eta$) random variables.

**Proof.** The initial cell is $x_1 = \emptyset$. The number of sites which mutate in cell $x_2$, $M_{x_2}^{\mu, S}$, is binomially distributed with parameters $S$ and $\mu$. This converges to a Poisson($\eta$) random variable. Now, for induction, suppose that

$$\lim_{n \to \infty} \left( M_{x_j}^{\mu, S} \right)^{k}_{j=2} = (m_j)^{k}_{j=2}$$

in distribution, where the $M_x$ are i.i.d. Poisson($\eta$) random variables. Then

$$\mathbb{P} \left[ \left( M_{x_j}^{\mu, S} \right)^{k+1}_{j=2} = (m_j)^{k+1}_{j=2} \right] = \mathbb{P} \left[ M_{x_{k+1}}^{\mu, S} = m_{k+1} \left| M_{x_j}^{\mu, S} \right. \right.^{k}_{j=2} = (m_j)^{k}_{j=2} \right] \times \mathbb{P} \left[ \left( M_{x_j}^{\mu, S} \right)^{k}_{j=2} = (m_j)^{k}_{j=2} \right]. \quad (28)$$

Due to the property (27) of the enumeration, $M_{x_{k+1}}^{\mu, S}$ conditioned on the event $(M_{x_j}^{\mu, S})^{k}_{j=2} = (m_j)^{k}_{j=2}$ is just a binomial random variable with parameters $|S| - \sum_{j=2}^{k} m_j$ and $\mu$. Therefore (28) converges as required. \qed
Proof of Proposition 7.7. Observe that

\[
\sum_{i \in S} \delta_{P_{\phi_i,\sigma_n}} = \sum_{x \in T \setminus \{\emptyset\}} M_x^{\mu, S} \delta_{P_x,\sigma_n}.
\]

Note that the \(M_x^{\mu, S}\) are independent of the \(P_{x,t}\), and condition on the \(P_{x,t}\):

\[
\left[ \sum_{x \in T \setminus \{\emptyset\}} M_x^{\mu, S_n} \delta_{P_{x,\sigma_n}} \left| (P_{x,t})_{x \in T \setminus \{\emptyset\}, t \geq 0} \right. \right] \rightarrow \left[ \sum_{x \in T \setminus \{\emptyset\}} M_x \delta_{P_x} \left| (P_{x,t})_{x \in T \setminus \{\emptyset\}, t \geq 0} \right. \right]
\]

in distribution, thanks to Lemmas 3.2, 7.5, and 7.8. Conditional convergence implies convergence.

\[\square\]

Lemma 7.9. Let \(a \in (0, 1)\). As \(n \to \infty\), \(\mu \to 0\), and \(|S|\mu \to \eta < \infty\),

\[
\mathbb{E} \left[ \sum_{i \in S} \delta_{P_{\phi_i,\sigma_n}} (a, 1) \right] \to 2\eta(a^{-1} - 1).
\]

Proof. First,

\[
\liminf \mathbb{E} \left[ \sum_{i \in S} \delta_{P_{\phi_i,\sigma_n}} (a, 1) \right] \geq \mathbb{E} \left[ \lim \sum_{i \in S} \delta_{P_{\phi_i,\sigma_n}} (a, 1) \right] = \mathbb{E} \left[ \sum_{x \in T \setminus \{\emptyset\}} M_x \delta_{P_x} (a, 1) \right] = 2\eta(a^{-1} - 1),
\]

by Fatou’s lemma, Proposition 7.7, and Remark 4.7. Second,

\[
\mathbb{E} \left[ \sum_{i \in S} \delta_{P_{\phi_i,\sigma_n}} (a, 1) \right] \leq \mathbb{E} \left[ \sum_{i \in S} \delta_{\hat{B}_t^{n,\mu}} (a, 1) \right] \to 2\eta(a^{-1} - 1),
\]

by the inequality \(P_{\phi_i,\sigma_n} \leq n^{-1} \hat{B}_t^{n,\mu}\) and then Proposition 7.1. \[\square\]
7.4 The infinite sites assumption is accurate

Lemma 7.10. Let \( a \in (0, 1) \). As \( n \to \infty \), \( \mu \to 0 \), and \( |S| \mu \to \eta < \infty \),
\[
\mathbb{E} \left| \sum_{i \in S} \delta_{P_{\phi^n, \sigma_n}}(a, 1) - \sum_{i \in S} \delta_{n^{-1}B^n_{i, \mu}}(a, 1) \right| \to 0.
\]

Proof. The inequalities
\[
n^{-1}B^n_{i, \mu} \leq n^{-1}\hat{B}^n_{i, \mu} \geq P_{\phi^n, \sigma_n},
\]
imply that
\[
\sum_{i \in S} \delta_{n^{-1}B^n_{i, \mu}}(a, 1) \leq \sum_{i \in S} \delta_{n^{-1}\hat{B}^n_{i, \mu}}(a, 1) \geq \sum_{i \in S} \delta_{P_{\phi^n, \sigma_n}}(a, 1).
\]
Hence
\[
\mathbb{E} \left| \sum_{i \in S} \delta_{P_{\phi^n, \sigma_n}}(a, 1) - \sum_{i \in S} \delta_{n^{-1}B^n_{i, \mu}}(a, 1) \right|
\leq \mathbb{E} \left| \sum_{i \in S} \delta_{n^{-1}\hat{B}^n_{i, \mu}}(a, 1) - \sum_{i \in S} \delta_{P_{\phi^n, \sigma_n}}(a, 1) \right|
+ \mathbb{E} \left| \sum_{i \in S} \delta_{n^{-1}B^n_{i, \mu}}(a, 1) - \sum_{i \in S} \delta_{n^{-1}\hat{B}^n_{i, \mu}}(a, 1) \right|
= 2\mathbb{E} \sum_{i \in S} \delta_{n^{-1}\hat{B}^n_{i, \mu}}(a, 1) - \mathbb{E} \sum_{i \in S} \delta_{P_{\phi^n, \sigma_n}}(a, 1)
- \mathbb{E} \sum_{i \in S} \delta_{n^{-1}B^n_{i, \mu}}(a, 1),
\]
which converges to zero thanks to Corollary 4.5, Proposition 7.1, and Lemma 7.9.

Proof of Theorem 4.6. Let \( I_1, \ldots, I_k \subset [\epsilon, 1] \) be intervals. Writing \( \| \cdot \| \) for the \( l_1 \)-norm on \( \mathbb{R}^k \),
\[
\mathbb{E} \left| \left( \sum_{i \in S} \delta_{n^{-1}B^n_{i, \mu}}(I_j) \right)^k - \left( \sum_{i \in S} \delta_{P_{\phi^n, \sigma_n}}(I_j) \right)^k \right|
= \sum_{j=1}^k \mathbb{E} \left| \sum_{i \in S} \delta_{n^{-1}B^n_{i, \mu}}(I_j) - \sum_{i \in S} \delta_{P_{\phi^n, \sigma_n}}(I_j) \right|
\to 0,
\]
48
due to Lemma 7.10. Therefore

$$\left(\sum_{i \in S} \delta_{n^{-1}B_{i}^{0,\mu}}(I_j)\right)_{j=1}^{k}$$

and

$$\left(\sum_{i \in S} \delta_{\phi_{\mu}^{-1}P_{\phi_{\mu}}}(I_j)\right)_{j=1}^{k}$$

share the same limiting distribution, if it exists. This limiting distribution, by Proposition 7.7, is

$$\left(\sum_{x \in \mathcal{T}\setminus\{\emptyset\}} M_x \delta_{P_{x}}(I_j)\right)_{j=1}^{k}$$

as required. $\square$

### 7.5 Mean and variance of the site frequency spectrum

Finally, let’s check Remarks 4.7 and 4.8. Note that

$$-\log(P_x) = -\sum_{y \leq x} \log(U_y)$$

is a sum of i.i.d. mean-1 exponentially distributed random variables, which is just a gamma random variable with parameters $|x|$ and 1. Then

$$P[x > a] = P[-\log(P_x) < -\log(a)]$$

$$= \int_{-\log(a)}^{-\log(a)} \frac{s^{|x|-1}e^{-s}}{(|x| - 1)!} ds.$$
So

\[
\mathbb{E} \left[ \sum_x M_x \delta P_x (a, 1) \right] = \eta \sum_x \mathbb{P}[P_x > a] \\
= \eta \sum_x \int_{-\log(a)}^{-\log(1/a)} \frac{s|x|^{-1} e^{-s}}{(|x| - 1)!} ds \\
= \eta \int_{-\log(a)}^{0} \sum_{l \in \mathbb{N}} 2^l s^{l-1} e^{-s} ds \\
= 2\eta \int_{0}^{-\log(a)} e^s ds \\
= 2\eta (a^{-1} - 1).
\]

As for the variance of the site frequency spectrum,

\[
\text{Var} \left[ \sum_x M_x \delta P_x (a, 1) \right] = \text{Var} \left[ \mathbb{E} \left[ \sum_x M_x \delta P_x (a, 1)|P_x \right] \right] \\
+ \mathbb{E} \left[ \text{Var} \left[ \sum_x M_x \delta P_x (a, 1)|P_x \right] \right] \\
\geq \mathbb{E} \left[ \text{Var} \left[ \sum_x M_x \delta P_x (a, 1)|P_x \right] \right] \\
= \mathbb{E} \left[ \sum_x \delta P_x (a, 1) \text{Var} [M_x] \right] \\
= \eta \mathbb{E} \left[ \sum_x \delta P_x (a, 1) \right].
\]

8 Infinite sites assumption violations

The infinite sites assumption (ISA) is a popular modelling assumption, stating that each genetic site can mutate at most once during the population’s evolution. There are influential and insightful analyses of tumour evolution which rely on the ISA, for example [21, 4, 20, 6]. However, recent statistical
analysis of single cell sequencing data shows “widespread violations of the ISA in human cancers” [15]. Thus it is unclear to what extent [21, 4, 20, 6]’s analyses can be trusted. Our studied model of DNA sequence evolution does not use the ISA and invites a theoretical assessment of the ISA’s validity. We claim that ISA violations are extremely prevalent. On the other hand, ISA violations’ impact on mutation frequencies is negligible at the scale of population fractions, which is made clear in Section 7. Thus our work stands in support of [21, 4, 20, 6, 15].

Let’s check the prevalence of ISA violations. For simplicity, consider the most basic version of the model, which was introduced in Section 2. Building upon notation of Section 3.2, write

$$T(n) = \{ x \in T : \exists y \in T_{\sigma_n}, x \prec y \}$$

for the set of ancestors of those cells alive at time $\sigma_n$ (when the total number of cells reaches $n$). Write

$$X_i^{n,\mu} = \left| \{(x, x_j) : x \in T(n), j \in \{0, 1\}, V_i^\mu(x) \neq V_i^\mu(x_j)\} \right|$$

for the number of times that site $i$ mutates up to time $\sigma_n$. Observe that $X_i^{n,\mu}$ is binomially distributed with parameters $2n - 2$ and $\mu$. Site $i$ is said to violate the ISA if $X_i^{n,\mu} \geq 2$, which occurs with probability

$$p(n, \mu) = \mathbb{P}[X_i^{n,\mu} \geq 2] = 1 - (1 - \mu)^{2n-2} - (2n - 2)\mu(1 - \mu)^{2n-3}.$$ 

Then the number of sites to violate the ISA,

$$|\{i \in S : X_i^{n,\mu} \geq 2\}|,$$

is binomially distributed with parameters $|S|$ and $p(n, \mu)$. If parameter values are indeed in the region of $n = 10^9$ and $\mu = 10^{-9}$, then the expected proportion of sites to violate the ISA is in the region of 0.5. This means that the expected number of sites to violate the ISA may be in the billions. Even if very conservative parameter estimates were plugged in, the number of violations is still massive. In fact violations are even more common if one considers cell death. Suppose that cells divide at rate $\alpha$ and die at rate $\beta$. Then to go from a population of 1 cell to $n$ cells requires approximately
$n\alpha/(\alpha - \beta)$ cell divisions, where the factor $\alpha/(\alpha - \beta)$ may be as large as 100 [4].

Note that our model only considers point mutations; it does not, for example, consider deletions of genomic regions, which are thought to be a significant cause of ISA violations [15].

9 Estimating mutation rates

In this section we wish to give the reader a light flavour of mutation frequency data and its relationship to the model. We estimate mutation rates in a lung adenocarcinoma.

9.1 Diploid perspective

Before presenting data, an additional ingredient needs to be considered: ploidy. Normal human cells are diploid. That is, chromosomes come in pairs. Therefore a particular mutation may be present zero, one, or two times in a single cell. It should be said that the story is far more complex in tumours, with chromosomal instability and aneuploidy coming into play. Even so, many tumour samples display an average ploidy not so far from two (for example see Figure (1a) of [22]). We imagine an idealised diploid world.

To illustrate the diploid structure, label the genetic sites as

$$S = \{1, 2\} \times \{1, \ldots, L\},$$

for some $L \in \mathbb{N}$. The first coordinate of a site $(i, j) \in S$ states on which chromosome of a pair the site lies, and the second coordinate refers to the site’s position on the chromosome. Mutations at sites $(1, j)$ and $(2, j)$ are typically not distinguished in data. In the original model set up, mutations were defined as differences to the initial cell’s genome. Let’s slightly improve that definition. Now a genome $v \in \mathcal{G}$ is said to be mutated at site $(i, j) \in S$ if $v_{i,j} \neq r_j$, where $(r_j)_{j=1}^L$ is some reference. Then data is simplistically stated
in the model’s language as

\[ F_j = \frac{1}{2n} \sum_{i=1}^{2} B_{i,j}^{n,\mu} \]  

(29)

for \( j = 1, \ldots, L \). That is, the total number of mutations at position \( j \) divided by the total number of chromosomes which contain position \( j \).

### 9.2 A lung adenocarcinoma

The mutation frequency data of a lung adenocarcinoma was made available in [23] (499017, Table S2). The data is plotted in Figures 3 and 4. It must be noted that a multitude of different data shapes exist, some looking nothing like this one. Our aim is only to give a flavour, hence just a single tumour’s data is presented.

Our method to estimate mutation rates is, to a large extent, inspired by [21, 4]. Their attention is restricted to a subset of mutations. They ignore mutations at frequency less than 0.1, saying that their detection is too unreliable. They ignore mutations above frequency 0.25, in order to neglect mutations present in the initial cell (which are relatively few). We do the same.

Write

\[ M(a, b) = |\{ j \in \{1, \ldots, L\} : F_j \in (a, b)\}| \]

for the number of mutations with frequency in \((a, b) \subset (0.1, 0.25)\). Then, adapting Corollary 4.5 to (29), the expected number of mutations with frequency in \((a, b)\) is

\[ EM(a, b) \approx \mu|S|(a^{-1} - b^{-1}). \]  

(30)

Under different models, [21, 4] derive the same approximation (30). They estimate the mutation rate \( \mu \) by applying a linear regression to (30). We simplify matters even further. Our estimator for \( \mu \) is

\[ \hat{\mu} = \frac{M(0.1, 0.25)}{6|S|}, \]  

(31)
Figure 3: A histogram of mutation frequencies from a lung adenocarcinoma.

Figure 4: The number of mutations (of the lung adenocarcinoma) whose frequency is in the interval $(0.1, x)$, for $x \in (0.1, 0.25)$. 
which (30) says is asymptotically unbiased. Now let’s calculate $\hat{\mu}$ for the data example. The data shows mutations on the exome, which has rough size $|S| = 3 \times 10^8$ [4]. And the number of mutations in the specified frequency range is $M((0.1, 0.25)) = 112$. This gives

$$\hat{\mu} = 6.2 \times 10^{-8}.$$ 

Next let’s consider mutation rate heterogeneity. Write $\mu_\chi$ for the rate that nucleotide $\chi \in \mathbb{N}$ mutates. Partition the genetic sites:

$$S = S_A \cup S_C \cup S_G \cup S_T,$$

where

$$S_\chi = \{i \in S : u_i = \chi\}$$

is the set of sites which are represented by nucleotide $\chi$ in the initial cell. Just as before,

$$\hat{\mu}_\chi = \frac{M_\chi((0.1, 0.25))}{6|S_\chi|}$$

is an unbiased estimator for $\mu_\chi$. The data gives

$$(\hat{\mu}_A, \hat{\mu}_C, \hat{\mu}_G, \hat{\mu}_T) = (0.7, 12.8, 15.0, 1.5) \times 10^{-8}.$$ 

This method could easily be extended to offer more detail, for example to estimate the rate at which nucleotide $A$ mutates to $C$ or to estimate mutation rates on different chromosomes.

The just presented statistical analysis is of course simple and brief. We recommend [6] for a far more comprehensive statistical analysis of mutation frequency data (although their infinite-sites framework does not consider mutation rate heterogeneity).

**Appendix**

A heuristic derivation of Conjecture 5.3 is given.

First we argue that, in the conjecture’s limit, selection is unimportant. Write

$$S_{sel} = \{v \in S : \exists i \in S_{sel}, v_i \neq u_i\}$$


55
for the set of genomes which are mutated at a selective site. Write
\[ Q^\mu_{\text{sel}}(t) = \frac{\sum_{v \in G_{\text{sel}}} X^\mu_v(t)}{\sum_{v \in G} X^\mu_v(t)} \]
for the proportion of cells at time \( t \geq 0 \) whose genomes are mutated at a selective site. Then, according to Theorem 6.1,
\[ (Q^\mu_{\text{sel}}(\sigma^\mu_n)|\sigma^\mu_n < \infty) \to 0 \]
in probability. Therefore we neglect selection.

Cells divide and die at rates \( \alpha(u') \) and \( \beta(u') \), which we now abbreviate to \( \alpha \) and \( \beta \). Some cells have an ultimately surviving lineage of descendants. Other cells eventually have no surviving descendants. Name these cells immortal and mortal respectively. In a supercritical birth-death branching process, it is well-known (eg. [7]) that the immortal cells grow as a Yule process and the mortal cells grow as a subcritical branching process. An immortal cell divides to produce two immortal cells at rate \( \alpha - \beta \), or it divides to produce one immortal and one mortal cell at rate \( 2\beta \). A mortal cell divides at rate \( \beta \) to produce two mortal cells, or it dies at rate \( \alpha \). Because the process is conditioned to reach a large population size, let’s assume that the initial cell is immortal.

The notation of Section 3.2, \( \mathcal{T} = \bigcup_{i=0}^{\infty} \{0, 1\}^i \) and its partial ordering \( \prec \), will be used to represent the immortal cells. Let \( (A_x)_{x \in \mathcal{T}} \) be i.i.d. \( \text{Exp}(\alpha - \beta) \) random variables, which represent the times for immortal cells to divide to produce two immortal cells. The immortal cells at time \( t \geq 0 \) are
\[ \mathcal{T}_t = \left\{ x \in \mathcal{T} : \sum_{y \prec x} A_y \leq t < \sum_{y \preceq x} A_y \right\}. \]
The immortal descendants of \( x \in \mathcal{T} \) are
\[ \mathcal{D}^I_x = \{ y \in \mathcal{T} : x \preceq y \}. \]
The number of immortal descendants of cell \( x \) at time \( t \) is
\[ D^I_{x,t} = |\mathcal{D}^I_x \cap \mathcal{T}_t|. \]
Let \( ((R_x(t))_{t \geq 0})_{x \in T} \) be i.i.d. Poisson processes with rate \( 2\beta \). Write \( R_{x,r} = \min\{t \geq 0 : R_x(t) = r\} \) for \( i = 1, \ldots, R_x(A_x) \). Then the seeding times of mortal cells are

\[
S_{x,r} = \sum_{y \prec x} A_y + R_{x,r}.
\]

Each seeding event initiates a subpopulation of mortal cells; let \( (Y_{x,r}(t))_{t \geq 0} \) be i.i.d. birth-death branching processes with birth and death rates \( \beta \) and \( \alpha \). Then the number of mortal descendants of \( x \) at time \( t \) is

\[
D^M_{x,t} = \sum_{y \in D_x} \sum_{r=1}^{R_y(A_y)} 1_{\{t-S_{y,r} \geq 0\}} Y_{y,r}(t - S_{y,r}).
\]

The number of descendants of \( x \) at time \( t \) is

\[
D_{x,t} = D^I_{x,t} + D^M_{x,t}.
\]

The next result shows the long-term proportion of a cell’s descendants which are immortal. The result is a basic consequence of classic branching process theory [2], and was mentioned in its specific form by [7].

**Lemma A.1.** There is \( c \in (0, \infty) \) with

\[
\lim_{t \to \infty} \frac{D^I_{x,t}}{D_{x,t}} = c \quad \text{almost surely.}
\]

We use Lemma A.1 to see the number of descendants of a cell as a proportion of the total population.

**Lemma A.2.** For \( x \in T \setminus \{\emptyset\} \),

\[
\lim_{t \to \infty} \frac{D_{x,t}}{D_{\emptyset,t}} = P_x
\]

almost surely, where the \( P_x \) are as in Lemma 3.2.

**Proof.** By Lemma 3.2 and Lemma A.1,

\[
\frac{D_{x,t}}{D_{\emptyset,t}} = \frac{D_{x,t}}{D^I_{x,t}} \frac{D^I_{x,t}}{D_{\emptyset,t}}\frac{D^I_{x,t}}{D^I_{\emptyset,t}} D^I_{\emptyset,t}
\]

converges to the required limit. \( \square \)
Let’s look at mutations. In the proof of Theorem 4.6 it was shown that the number of new mutations to arise at a cell’s birth is approximately Poisson. Here, with heterogeneous mutation rates, the number of new mutations to arise at a cell’s birth is approximately Poisson with mean

\[ \eta := \sum_{j \in J} \sum_{\psi \in N \setminus \{u(j)\}} \eta^{u(j), \psi}(j). \]

Each \( x \in \mathcal{T} \setminus \{\emptyset\} \) witnesses 1 + \( R_x(A_x) \) cell divisions. So the number of new mutations to arise at \( x \) is

\[ \sum_{r=0}^{R_x(A_x)} M_{x,r}, \]

where the \( M_{x,r} \) are i.i.d. Poisson random variables with mean \( \eta \). In the proof of Theorem 4.6 it was also shown that a mutation which arises in cell \( x \) will have approximate frequency \( P_x \). Here, thanks to Lemma A.2, the situation appears identical. It only remains to discuss mutations arising in mortal cells. Any subpopulation of cells which descended from a mortal cell must eventually die out. Hence mutations arising in mortal cells are negligible when compared to the infinite total population size.

### Acknowledgements

We thank Michael Nicholson, Trevor Graham, Marc Williams, and Martin Nowak for discussions. David Cheek was supported by The Maxwell Institute Graduate School in Analysis and its Applications, a Centre for Doctoral Training funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016508/01), the Scottish Funding Council, Heriot-Watt University and the University of Edinburgh.

### References

[1] D. Aldous. Stopping times and tightness. *Annals of Probability*, 6:335–340, 1978.
[2] K. B. Arthreya and P. Ney. *Branching Processes*. Dover Publications, 2004.

[3] I. Bozic et al. Accumulation of driver and passenger mutations during tumor progression. *PNAS*, 107:18545–18550, 2010.

[4] I. Bozic, J. M. Gerold, and M. A. Nowak. Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLOS Computational Biology*, 12(2):e1004731, 2016.

[5] D. Cheek and T. Antal. Mutation frequencies in a birth-death branching process. *Annals of Applied Probability*, 28:3922–3947, 2018.

[6] K. N. Dinh, R. Jaksik, M. Kimmel, A. Lambert, and S. Tavaré. Statistical inference for the evolutionary history of cancer genomes. *bioRxiv*, 2019.

[7] R. Durrett. Population genetics of neutral mutations in exponentially growing cancer cell populations. *Annals of Applied Probability*, 23(1):230–250, 2013.

[8] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. New York: Wiley, 1986.

[9] A. Hamon and B. Ycart. Statistics for the Luria-Delbrück distribution. *Electronic Journal of Statistics*, 6:1251–1272, 2012.

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

[11] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. *New York: Academic Press*, pages 21–132, 1969.

[12] P. Keller and T. Antal. Mutant number distribution in an exponentially growing population. *Journal of Statistical Mechanics: Theory and Experiment*, P01011, 2015.

[13] D. G. Kendall. Birth-and-death processes, and the theory of carcinogenesis. *Biometrika*, 47(1-2):13–21, 1960.
[14] D. A. Kessler and H. Levine. Scaling solution in the large population limit of the general asymmetric stochastic Luria-Delbrück evolution process. *Journal of Statistical Physics*, 158(4):783–805, 2015.

[15] J. Kuipers, K. Jahn, B. J. Raphael, and N. Beerenwinkel. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome research*, 2017.

[16] D. E. Lea and C. A. Coulson. The distribution of the numbers of mutants in bacterial populations. *Journal of Genetics*, 49(3):264–285, 1949.

[17] S. E. Luria and M. Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 48(6):419–511, 1943.

[18] A. Pakes. Remarks on the Luria-Delbrück distribution. *Journal of Applied Probability*, 30:991–994x, 1993.

[19] S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[20] M. J. Williams et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, 50:895–903, 2018.

[21] M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, and A. Sotoriva. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48:238–244, 2016.

[22] T. I. Zack et al. Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45:1134–1140, 2013.

[23] J. Zhang et al. Intra-tumor heterogeneity in localized lung adenocarcinomas delineated by multi-region sequencing. *Science*, 346:256–259, 2014.

[24] Q. Zheng. Progress of a half century in the study of the Luria-Delbrück distribution. *Mathematical Biosciences*, 162(1-2):1–32, 1999.