THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Genetic composition of an exponentially growing cell population

OPEN ACCESS

Genetic composition of an exponentially growing cell population

David Cheek[*] and Tibor Antal[†]

Abstract

We study a simple model of DNA evolution in a growing population of cells. Each cell contains a nucleotide sequence which randomly mutates at cell division. Cells divide according to a branching process. 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 nucleotide 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. For illustration we estimate mutation rates in a lung adenocarcinoma.

---

[*]Program for Evolutionary Dynamics, Harvard University, dmcheek7@gmail.com
[†]School of Mathematics, University of Edinburgh, tibor.antal@ed.ac.uk

Contents

1   Introduction

A population of dividing cells with a mutating DNA sequence is ubiquitous in biology. We study a simple model of this process. Starting with one cell, cells divide and die according to a supercritical branching process. As for DNA, we loosely follow classic models from phylogenetics [16, 25]. 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 are interested in the sequence distribution when the population reaches many cells.

Let's discuss a specific motivation. In recent years, cancer genetic data has been made available in great quantities. One especially common type of data consists of mutation frequencies in individual tumours. These data take the

3

form of a vector $(x_i)_{i\in\mathcal{S}}$, where $i \in \mathcal{S}$ denotes genetic sites and x_i is the frequency of cells which are mutated at site i. To make sense of such data in terms of tumour evolution, simple mathematical models can be helpful.

Some important works on the topic are [27, 5, 26, 8]. They 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. The assumption's simplicity drives its popularity. However recent statistical analysis of single cell sequencing data [20] 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 [22] modelled recurrent mutations in an exponentially growing bacterial population. Subsequent works [21, 18, 14, 17, 19] (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 [7] 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 a sequence of genetic sites. Now we offer a self-contained sequel to [7], slightly adapting the model, and aiming for a deeper understanding of the sequence distribution.

In [7] we studied several parameter regimes. In the present work by contrast, we study only one parameter regime which is the most biologically relevant. We take the final number of cells to infinity and the mutation rate to zero with their product finite. This limit is relevant because a detected tumour has around 10^9 cells while the mutation rate per site per cell division is around 10^{-9} [15]. This limit is also standard in Luria-Delbrück-type models of bacteria.

Now we introduce our main results. 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 [21, 14, 19, 17, 7]. So a site is mutated in only a finite number of cells, standing in contrast to the infinite total number of cells. Going beyond [21, 14, 19, 17, 7], we also study the rare event that a site is mutated in

4

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 [10, 27, 5, 8, 7] who only give the expected site frequency spectrum, so our work contributes an appreciation of randomness.

Our results are not all at the same level of generality. For sites mutated in a positive fraction of cells, results are proven for a zero death rate and homogeneous division and mutation rates. For sites mutated in a finite number of cells, results are proven for sequence-dependent death, division, and mutation rates.

We also assess the infinite sites assumption's validity. Our results say that for typical parameter values, the number of sites to violate the ISA is at least millions, or even billions, in a single tumour. Thus our work agrees with [20]'s statistical analysis of single cell sequencing data which says that ISA violations are widespread. It should be emphasised however that ISA violations do not neccessarily invalidate the ISA. One of our results says that ISA violations do not impact mutation frequencies viewed at the scale of population fractions. Bulk sequencing data, which is the majority of cancer genetic data [8], views mutation frequencies at the scale of population fractions. Therefore our work endorses analyses of bulk sequencing data which are reliant on the ISA, such as [27, 5, 26, 8].

Before commencing the paper, let's note that there are a wealth of other works on mutations in branching processes. Especially common are infinite alleles models, for example [13, 6, 23, 9], where each individual in the population has an allele which can mutate to alleles never before seen in the population. In an infinite alleles model, a mutation always deletes an individual's ancestral genetic information. In an infinite sites model on the other hand, a mutation never deletes ancestral genetic information; mutations simply accumulate. The DNA sequence model which we study sits between those extremes.

5

The paper is structured as follows. In Section 2, we introduce the model in its simplest form. In Section 3, we give notation and preliminary ideas. In Section 4, we present the paper's main results. In Section 5, we give generalisations and open questions. In Section 6, we prove results on sites mutated in a finite number of cells. In Section 7, we prove results on sites mutated in a positive fraction of cells. In Section 8, we discuss the infinite sites assumption's validity. In Section 9, we consider data from a lung adenocarcinoma and estimate mutation rates.

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

$$B = \sum_{k=1}^{K} Y_k. \tag{1}$$

It is commonly seen in its generating function form (e.g. [21, 30])

$$\mathbb{E}z^B = (1-z)^{c(z^{-1}-1)}. \tag{2}$$

The connection between (1) and (2) is made explicit in [7] for example. Although the distribution is named after Luria and Delbrück (due to their groundbreaking work [22]), it was derived by Lea and Coulson [21]. See [30] for a historical review.

The Luria-Delbrück distribution's power-law tail was derived in [24].

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

$$\mathcal{T} = \cup_{l=0}^{\infty}\{0,1\}^l.$$

A partial ordering, $\prec$, is defined on $\mathcal{T}$. For $x, y \in \mathcal{T}$, $x \prec y$ 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 $\emptyset \in \mathfrak{T}$ and that $\emptyset \prec x$ for any $x \in \mathfrak{T}\backslash\{\emptyset\}$. 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 $x \in \mathfrak{T}$; precisely, if $x \in \mathfrak{T}$ 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 $(A_x)_{x \in \mathfrak{T}}$ 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

$$\mathfrak{T}_t := \left\{ x \in \mathfrak{T} : \sum_{y \prec x} A_y \leq t < \sum_{y \preceq x} A_y \right\}.$$

The proportion of cells alive at time t which are descendants of cell x (including x) is

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

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

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

almost surely, where

1. *the U_y are uniformly distributed on $[0,1]$;*

2. *for any $y \in \mathfrak{T}$, $U_{y0} + U_{y1} = 1$;*

3. *$(U_{y0})_{y \in \mathfrak{T}}$ is an independent family.*

Lemma 3.2 will be proved in Section 7.

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 $V^\mu(x) = (V_i^\mu(x))_{i \in \mathcal{S}} \in \mathcal{G}$ for the genome of cell $x \in \mathcal{T}$ (where μ 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. Note that, according to this definition, a site which mutates and then sees a reverse mutation to its initial state is not considered to be mutated. Write

$$B_i^{n,\mu} = |\{x \in \mathcal{T}_{\sigma_n} : V_i^\mu(x) \neq u_i\}| \tag{3}$$

for the number of cells which are mutated at site $i \in \mathcal{S}$ when there are n cells in total. The quantity (3), 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 $\mu = 10^{-9}$ [15]. The human genome's length is around $|\mathcal{S}| = 3 \times 10^9$. Very roughly,

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

Therefore we study the limits:

- $n \to \infty$, $\mu \to 0$, $n\mu \to \theta < \infty$;

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

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

4   Main results

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 $n \to \infty$ and $n\mu \to \theta \in [0, \infty)$,*

$$(B_i^{n,\mu})_{i \in \mathcal{S}} \to (B_i)_{i \in \mathcal{S}}$$

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

Remark 4.2. *Taking $|\mathcal{S}| = 1$, Theorem 4.1 recovers results of single site models [21, 14, 17, 19, 7].*

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

$$\sum_{i \in \mathcal{S}} \delta_{B_i^{n,\mu}}.$$

The site frequency spectrum sees a law of large numbers.

Theorem 4.3. *As $n \to \infty$, $n\mu \to \theta \in [0, \infty)$, and $|\mathcal{S}| \to \infty$,*

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

in probability, where Λ is the Luria-Delbrück distribution with parameter 2θ. 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:

$$\mathbb{P}[n^{-1} B_i^{n,\mu} > a] \approx \mathbb{P}[B_i \in (na, n)] \tag{4}$$
$$\approx 2\mu(a^{-1} - 1). \tag{5}$$

Approximation (4) is a hand-waving consequence of Theorem 4.1. Approximation (5) is due to Lemma 3.1. The next result offers rigour.

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

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

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 $n \to \infty$, $n\mu \to \theta \in [0, \infty)$, and $|\mathcal{S}|\mu \to \eta \in [0, \infty)$,*

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

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

Theorem 4.6. *As $n \to \infty$, $n\mu \to \theta \in [0, \infty)$, and $|\mathcal{S}|\mu \to \eta \in [0, \infty)$,*

$$\sum_{i \in \mathcal{S}} \delta_{n^{-1}B_i^{n,\mu}} \to \sum_{x \in \mathcal{T} \setminus \{\emptyset\}} M_x \delta_{P_x}$$

in distribution, with respect to the vague topology on the space of measures on $(0, 1]$. That is, the measure applied to a finite collection of closed intervals in $(0, 1]$ sees joint convergence. The random variables which appear in the limit are:

- *(M_x) is a family of i.i.d. Poisson(η) 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.
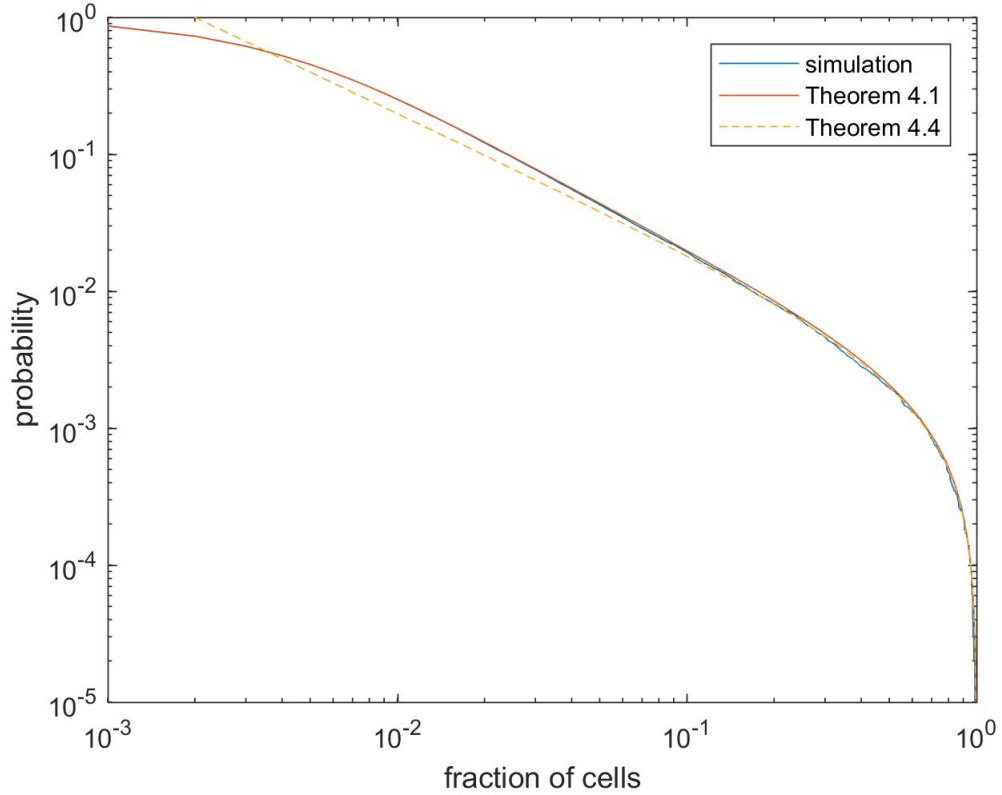
Figure 1: The number of mutant cells with respect to a single site is simulated 10^5 times. The parameters are $\mu = 10^{-3}$ and $n = 10^3$. The plot compares $\mathbb{P}[n^{-1}B_1^{n,\mu} \in (a,1)]$ (simulation), $\mathbb{P}[n^{-1}B_1 \in (a,1)]$ (Theorem 4.1), and $2\mu(a^{-1} - 1)$ (Theorem 4.4), for $a \in (0,1)$. Simulation and Theorem 4.1 appear indistinguishable.

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

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

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].

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 $\mathcal{G} = \mathbb{N}^{\mathcal{S}}$. It will be helpful to classify different types of genetic site. Partition the sites into *neutral* and *selective* sites:

$$\mathcal{S} = \mathcal{S}_{\text{neut}} \cup \mathcal{S}_{\text{sel}},$$

with $\mathcal{S}_{\text{sel}} \neq \emptyset$. For a genome $v = (v_i)_{i \in \mathcal{S}}$, write $v' = (v_i)_{i \in \mathcal{S}_{\text{sel}}}$ for its restriction to the selective sites. Let α and β be functions with domain $\mathbb{N}^{\mathcal{S}_{\text{sel}}}$ 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 $(v_i)_{i \in \mathcal{S}}$ dividing to give daughter cells with genomes $(V_i^{(1)})_{i \in \mathcal{S}}$ and $(V_i^{(2)})_{i \in \mathcal{S}}$. Conditional on (v_i), the $V_i^{(r)}$ are independent over $i \in \mathcal{S}$ and $r \in \{1, 2\}$, and

$$\mathbb{P}\left[V_i^{(r)} = \psi | (v_i)\right] = \mu_i^{v_i, \psi}.$$
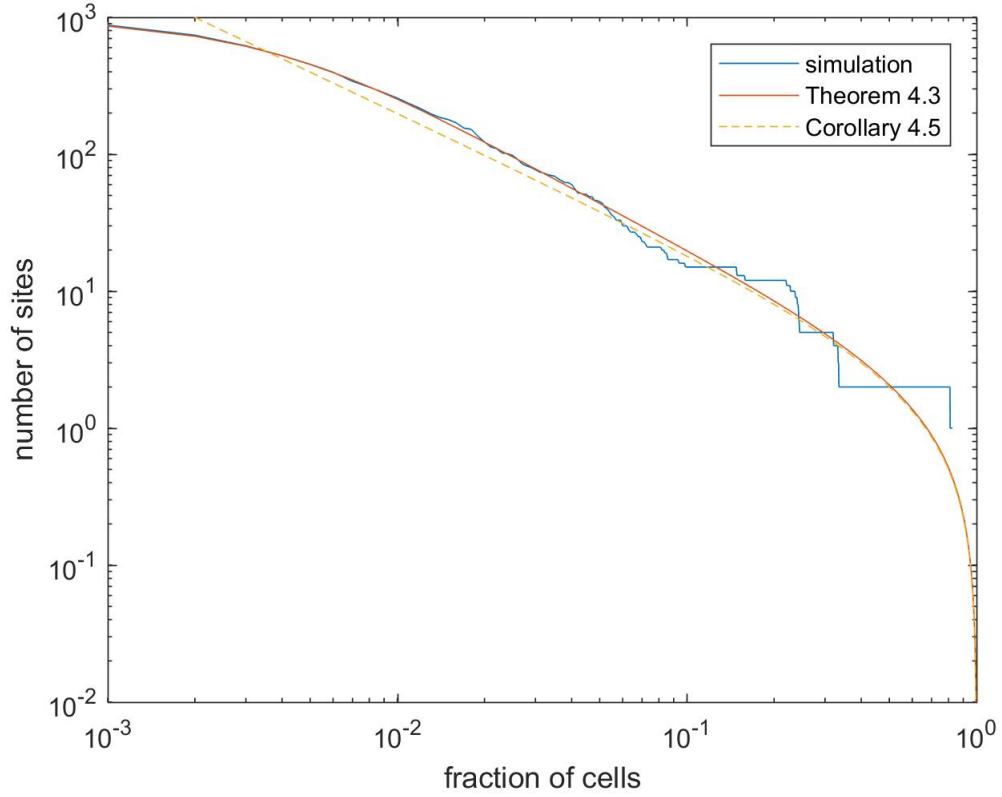
13

Figure 2: The site frequency spectrum is simulated a single time. The parameters are $\mu = 10^{-3}$, $n = 10^3$, and $|\mathcal{S}| = 10^3$. The plot compares $\sum_{i \in \mathcal{S}} \delta_{n^{-1}B_i^{n,\mu}}(a, 1)$ (simulation), $|\mathcal{S}|\mathbb{P}[n^{-1}B_1 \in (a, 1)]$ (Theorem 4.3), and $2|\mathcal{S}|\mu(a^{-1} - 1)$ (Corollary 4.5), for $a \in (0, 1)$.

Slightly adapting previous notation, write

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

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 $(\xi_k)_{k \in \mathbb{N}}$ be an i.i.d. sequence of exponentially distributed random variables with mean λ^{-1}. 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 K be a Poisson random variable with mean c. The (ξ_k), $(Y_k(\cdot))$, and K are independent. The *generalised Luria-Delbrück distribution* with parameters (λ, a, b, c) is defined as the distribution of

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

Its generating function

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

when $a > b$ is seen in [17, 19, 7]. Here F is Gauss's hypergeometric function.

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

The generalised Luria-Delbrück distribution with parameters $(\lambda, \lambda a, \lambda b, c)$, for $\lambda > 0$ and $a, b, c \geq 0$, does not depend on λ. So one could define the distribution with 3 rather than 4 parameters. We choose 4 for a cleaner interpretation of results.

5.3 Results

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

$$\mathcal{G}_i = \{v \in \mathcal{G} : \forall j \in \mathcal{S}, (u_j \neq v_j \iff i = j)\}, \tag{6}$$

will play a crucial role.

Theorem 5.1. *Take* $n \to \infty$ *and* $n\mu_i^{\chi,\psi} \to \theta_i^{\chi,\psi} \in [0,\infty)$ *for all* $i \in \mathcal{S}$ *and* $\chi, \psi \in \mathcal{N}$ *with* $\chi \neq \psi$. *Then*

$$(B_i^{n,\mu})_{i \in \mathcal{S}} \to \left(\sum_{v \in \mathcal{G}_i} X_v \right)_{i \in \mathcal{S}}$$

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

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

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:

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

such that mutation rates and the initial genome's nucleotides are homogeneous on $\mathcal{S}(j)$ (J is just some indexing set). Write $\mu^{\chi,\psi}(j) = \mu_i^{\chi,\psi}$ for the mutation rates of the sites $i \in \mathcal{S}(j)$. Write $u(j) = u_i$ for the initial genome's nucleotide at the sites $i \in \mathcal{S}(j)$.

Theorem 5.2. *Take* $n \to \infty$, $n\mu^{\chi,\psi}(j) \to \theta^{\chi,\psi}(j) \in [0,\infty)$, $|\mathcal{S}_{neut}| \to \infty$, *and* $|\mathcal{S}(j)|/|\mathcal{S}_{neut}| \to q(j)$, *for all* $j \in J$ *and* $\chi, \psi \in \mathcal{N}$ *with* $\chi \neq \psi$. *Then*

$$\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \delta_{B_i^{n,\mu}} \xrightarrow{p} \sum_{j \in J} q(j)\Lambda(j)$$

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

$$\left(\alpha(u') - \beta(u'), \alpha(u'), \beta(u'), \frac{2\alpha(u')}{\alpha(u') - \beta(u')} \sum_{\psi \in \mathcal{N} \setminus \{u(j)\}} \theta^{u(j),\psi}(j) \right).$$

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 to a non-zero death rate, selection, and heterogeneous mutation rates, we conjecture the following.

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

$$\sum_{i \in \mathcal{S}} \delta_{n^{-1} B_i^{n,\mu}} \to \sum_{x \in \mathcal{T}} \sum_{r=1}^{R_x} M_{x,r} \delta_{P_x}$$

in distribution, where convergence is in the same sense as Theorem 4.6. 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 $P_\emptyset = 1$;*

- *$(R_x)_{x \in \mathcal{T} \setminus \{\emptyset\}}$ is an i.i.d. family of geometric random variables with parameter $(\alpha(u') - \beta(u'))/(\alpha(u') + \beta(u'))$, and $R_\emptyset$ is independent of (R_x) but with $R_\emptyset \stackrel{d}{=} R_0 - 1$;*

- *(P_x) is independent of (R_x) if and only if $\beta(u') = 0$;*

- *$(M_{x,r})_{x \in \mathcal{T}, r \in \mathbb{N}}$ 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 [4, 26, 8]. Pertinently, Theorem 5.2 and Conjecture 5.3 suggest that selection may not be visible in mutation frequency data, which according to [27] is the case for around $1/3$ of tumours. However we have assumed that the number of selective sites is kept finite. According to [4], there are 3.4×10^4 selective sites at which mutations can positively affect growth rate. Thus insight could be gleaned, for example, by taking $|\mathcal{S}_{\mathrm{sel}}| \to \infty$ with $\mu^\gamma |\mathcal{S}_{\mathrm{sel}}| \to \eta$ for $\gamma \in (0, 1]$.

6   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 (where $\mathcal{S}$ is finite). In Subsection 6.6 we prove Theorems 4.3 and 5.2 (where $|\mathcal{S}|$ tends to infinity).

6.1   Counting genomes

Assuming mutation rates $\mu = (\mu_i^{\chi,\psi})_{i \in \mathcal{S}; \chi\psi \in \mathcal{N}}$, write

$$X_v^\mu(t) \tag{7}$$

for the number of cells with genome $v \in \mathcal{G}$ at time $t \geq 0$. (Recall that the initial condition is $X_v^\mu(0) = \delta_{u,v}$.) Write

$$\sigma_n^\mu = \min\left\{ t \geq 0 : \sum_{v \in \mathcal{G}} X_v^\mu(t) = n \right\}$$

for the time at which $n \in \mathbb{N}$ cells are reached, and use the convention $\min \emptyset = \infty$.

Recall from (6) that $\mathcal{G}_i$ is the subset of genomes with exactly one mutation which is at site i. Write

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

for the subset of genomes with at least two mutations.

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

$$\left[(X_v^\mu(\sigma_n^\mu))_{v \in \mathcal{G}\setminus\{u\}} | \sigma_n^\mu < \infty \right] \to (X_v)_{v \in \mathcal{G}\setminus\{u\}}$$

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

- *for $v \in \mathcal{G}_i$, X_v has generalised Luria-Delbrück distribution with parameters*
$$\left(\alpha(u') - \beta(u'), \alpha(v'), \beta(v'), \frac{2\alpha(u')\theta_i^{u_i,v_i}}{\alpha(u') - \beta(u')} \right);$$

- *for $v \in \mathcal{G}_{\geq 2}$, $X_v = 0$.*

Theorem 6.1 says that 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.

The mutation frequencies are

$$(B_i^{n,\mu})_{i \in \mathcal{S}} = \left(\sum_{\substack{v \in \mathcal{G} \\ v_i \neq u_i}} X_v^\mu(\sigma_n^\mu) \right)_{i \in \mathcal{S}}$$

conditional on the event $\{\sigma_n^\mu < \infty\}$. 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 $(X_v^\mu(\sigma_n^\mu))_{v \in \mathcal{G}}$. 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 σ_n^μ shoulders a large responsibility for complexity.

6.2 Construction

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

$$e_v = (\delta_{v,w})_{w \in \mathcal{G}}$$

is the element of $(\mathbb{N}_0)^{\mathcal{G}}$ denoting that there is one genome v and zero other genomes.

Let

$$[\mu_n]_{n \in \mathbb{N}} = \left[\left(\mu_{n,i}^{\chi,\psi} \right)_{i \in \mathcal{S}; \chi,\psi \in \mathcal{N}} \right]_{n \in \mathbb{N}}$$

be a sequence of mutation rates. Assume that

$$\lim_{n \to \infty} n \mu_{n,i}^{\chi,\psi} = \theta_i^{\chi,\psi} \in [0,\infty)$$

for $\chi \neq \psi$.

Fix $n \in \mathbb{N}$. For $v, w \in \mathcal{G}$, write

$$p_n(v,w) = \prod_{i \in \mathcal{S}} \mu_{n,i}^{u_i,v_i} \mu_{n,i}^{u_i,w_i} \tag{8}$$

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 $(X_v^{\mu_n}(\sigma_n^{\mu_n}))_{v \in \mathcal{G}}$ 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

   $$\beta_n := \beta(u') + \alpha(u') \sum_{v,w \in \mathcal{G}\backslash\{u\}} p_n(v, w).$$

   The initial condition $Z^n(0) = 1$ is assumed.

2. For $j \in \mathbb{N}$,
   $$E_j^n$$

   are $\{\emptyset\} \cup (\mathcal{G}\backslash\{u\})^2$-valued random variables, with

   $$\mathbb{P}[E_j^n = \emptyset] = \frac{\beta(u')}{\beta_n},$$

   and for $v, w \in \mathcal{G}\backslash\{u\}$

   $$\mathbb{P}[E_j^n = (v, w)] = \frac{\alpha(u') p_n(v, w)}{\beta_n}.$$

3. For $v \in \mathcal{G}\backslash\{u\}$ and $j \in \mathbb{N}$,
   $$\mathcal{Y}_{v,j}^n(\cdot)$$

   is a $(\mathbb{N}_0)^{\mathcal{G}}$-valued Markov process, with the same transition rates as $(X_x^{\mu_n}(\cdot))_{x \in \mathcal{G}}$ (defined in (7)) and with the initial condition $\mathcal{Y}_{v,j}^n(0) = e_v$.

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

   $$\mathcal{Y}_{v,w,j}^n(\cdot)$$

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

21

5. For $v \in \mathcal{G}$,
$$(N_v(t))_{t \geq 0}$$
are Poisson counting processes with rate 1.

The random variables
$$\left[Z^n(\cdot), E_j^n, \mathcal{Y}_{v,j}^n(\cdot), \mathcal{Y}_{v,w,j}^n(\cdot), N_v(\cdot) \right] \tag{9}$$
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, α_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, β_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_j^n describe what happens at the jth downstep in the primary subpopulation trajectory. If $E_j^n = \emptyset$, then the downstep is a primary cell death. If $E_j^n = (v, w)$, 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 jth time that this occurs, $\mathcal{Y}_{v,j}^n(t)$ 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 jth time that this occurs, $\mathcal{Y}_{v,w,j}^n(t)$ 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 (9), provide all the

22

necessary ingredients for the construction of $(X_v^{\mu_n}(\sigma_n^{\mu_n}))_{v\in\mathcal{G}}$. Now we build upon these founding objects, defining further random variables.

6. For $v \in \mathcal{G}\backslash\{u\}$ and $t \geq 0$,

$$K_v^n(t) = N_v\left(2p_n(u,v)\alpha(u')\int_0^t Z^n(s)ds\right). \qquad (10)$$

7. For $j \in \mathbb{N}$ and $v \in \mathcal{G}\backslash\{u\}$,

$$T_{v,j}^n = \min\{t \geq 0 : K_v^n(t) = j\}.$$

8.

$$S_1^n = \min\{t \geq 0 : Z^n(t) - Z^n(t^-) = -1\},$$

and then for $j > 1$, recursively,

$$S_j^n = \min\{t > S_{j-1}^n : Z^n(t) - Z^n(t^-) = -1\}.$$

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

9. For $v, w \in \mathcal{G}\backslash\{u\}$,

$$T_{v,w,1}^n = \min\{S_j^n : j \in \mathbb{N}, E_j^n = (v,w)\},$$

and then for $j > 1$, recursively,

$$T_{v,w,j}^n = \min\{S_j^n : j \in \mathbb{N}, S_j^n > T_{v,w,j-1}^n, E_j^n = (v,w)\}.$$

10. For $v, w \in \mathcal{G}\backslash\{u\}$, and $t \geq 0$,

$$K_{v,w}^n(t) = |\{j \in \mathbb{N} : T_{v,w,j}^n \leq t\}|.$$

Let's explain the meaning of the new random variables. The $K_v^n(t)$ 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 $Z^n(\cdot)$, $K_v^n(\cdot)$ is certainly a Markov process, and increases by 1 at rate $2p_n(u,v)\alpha(u')Z^n(t)$ - i.e. the rate

23

at which primary cells divide multiplied by the probability that exactly one daughter cell is primary and one is type v.

S_j^n is the time of the jth downstep of the primary subpopulation size. Then $T_{v,w,j}^n$ is the time of the jth primary cell division which produces two mutant cells of types v and w. Note that a primary cell division which produces two mutant cells neccessarily coincides with a downstep in the primary subpopulation size. $K_{v,w}^n(t)$ is the number of primary cell divisions before time t which produce cells of types v and w.

The reader might question why we have decided to construct the 'single mutation' times and the 'double mutation' times so differently. The reason for the difference is that single and double mutations will play different roles in the limit, and require different techniques for the proof.

At last the construction reaches its dénouement.

11. For $t \geq 0$,

$$
\begin{aligned}
\mathcal{X}^n(t) \;=\; & Z^n(t)e_u \\
& + \sum_{v \in \mathcal{G} \setminus \{u\}} \sum_{j=1}^{K_v^n(t)} \mathcal{Y}_{v,j}^n(t - T_{v,j}^n) \\
& + \sum_{v,w \in \mathcal{G} \setminus \{u\}} \sum_{j=1}^{K_{v,w}^n(t)} \mathcal{Y}_{v,w,j}^n(t - T_{v,w,j}^n).
\end{aligned}
$$

12.
$$
\sigma_n = \min\{t \geq 0 : |\mathcal{X}^n(t)| = n\},
$$
where $|\cdot|$ is the l_1-norm on $\mathbb{R}^{\mathcal{G}}$.

Note that $\mathcal{X}^n(\cdot)$ has the same distribution as $(X_v^{\mu_n}(\cdot))_{v \in \mathcal{G}}$; both objects are Markov processes on $(\mathbb{N}_0)^{\mathcal{G}}$, whose initial conditions and transition rates coincide.

Next we will show that certain elements of the construction converge in distribution. Convergence will sometimes be in the Skorokhod sense. For notation, write $\mathbb{D}(I, R)$ for the space of càdlàg functions from an interval

24

$I \subset [0, \infty)$ to a metric space R (which will always be complete and separable). The space $\mathbb{D}(I, R)$ is equipped with the standard Skorokhod topology. Less standard, we will also consider the space $\mathbb{D}([0, \infty], R)$, which is defined by identification with $\mathbb{D}([0, 1], R)$. Let's be specific. Define $\omega : [0, 1] \to [0, \infty]$ by

$$\omega : s \mapsto \begin{cases} -\log(1 - s), & s \in [0, 1); \\ \infty, & s = 1. \end{cases}$$

For $f \in \mathbb{D}([0, 1], R)$ define $\hat{w}(f) = f \circ \omega^{-1}$. Then the space $\mathbb{D}([0, \infty], R)$ is the image of $\hat{w}$ equipped with the induced topology. Note that according to this definition, for all $z \in \mathbb{D}([0, \infty], R)$, $\lim_{t \to \infty} z(t)$ exists. In fact we will only consider $z \in \mathbb{D}([0, \infty], R)$ with $\lim_{t \to \infty} z(t) = z(\infty)$.

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 λ_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 at $t = \infty$. 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 [3].

Proof of Lemma 6.2. The transition probabilities of the $\left(e^{-\lambda_n t} Z^n(t)\right)_{t \in [0,\infty]}$ and $\left(e^{-\lambda t} Z^*(t)\right)_{t \in [0,\infty]}$ are well-known [3, 11, 7]. These transition probabilities depend continuously on the birth and death rates, so finite-dimensional convergence is given. To show tightness we shall use Aldous's criterion [1]. Extend $\omega(s)$ to $s \in [0,2]$ by setting $\omega(s) = \omega(1) = \infty$ for $s \in [1,2]$. Write

$$M^n(s) = e^{-\lambda_n \omega(s)} Z^n(\omega(s))$$

for $s \in [0,2]$. Let (ρ_n) be a sequence of $[0,1]$-valued stopping times with respect to $(M^n(\cdot))$. Let (δ_n) be a positive deterministic sequence converging to zero. Then, writing $\mathcal{F}_{\rho_n}$ for the sigma-algebra generated by $M^n(\cdot)$ up to time ρ_n,

$$
\begin{aligned}
\mathbb{E}[(M^n(\rho_n + \delta_n) - M^n(\rho_n))^2 | \mathcal{F}_{\rho_n}] &= \mathbb{E}[M^n(\rho_n + \delta_n)^2 | \mathcal{F}_{\rho_n}] - M^n(\rho_n)^2 \\
&= \left(e^{-\lambda_n \omega(\rho_n)} - e^{-\lambda_n \omega(\rho_n + \delta_n)}\right) \\
&\quad \times \frac{\alpha_n + \beta_n}{\lambda_n} M^n(\rho_n),
\end{aligned}
$$

where the last equality comes thanks to the fact that

$$M^n(s)^2 + \frac{\alpha_n + \beta_n}{\lambda_n} e^{-\lambda_n \omega(s)} M^n(s)$$

is a martingale. But

$$
\begin{aligned}
e^{-\lambda_n \omega(\rho_n)} - e^{-\lambda_n \omega(\rho_n + \delta_n)} &= (1 - \rho_n)^{\lambda_n} - 1_{\{1 - \rho_n - \delta_n \geq 0\}}(1 - \rho_n - \delta_n)^{\lambda_n} \\
&\leq \max\{\lambda_n \delta_n, \delta_n^{\lambda_n}\}.
\end{aligned}
$$

Now,

$$
\begin{aligned}
\mathbb{E}[(M^n(\rho_n + \delta_n) - M^n(\rho_n))^2] &\leq \max\{\lambda_n \delta_n, \delta_n^{\lambda_n}\} \frac{\alpha_n + \beta_n}{\lambda_n} \mathbb{E}M^n(\rho_n) \\
&= \max\{\lambda_n \delta_n, \delta_n^{\lambda_n}\} \frac{\alpha_n + \beta_n}{\lambda_n}.
\end{aligned}
$$

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 $n \to \infty$,*

$$\mathcal{Y}_{v,j}^n(\cdot) \to Y_{v,j}(\cdot) e_v$$

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 $\mathbb{D}([0,\infty), \mathbb{R}^{\mathcal{G}})$.

26

Proof. It is enough to note that the transition rates converge (see for example page 262 of [12]). □

Lemma 6.5. *As $n \to \infty$,*

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

in distribution, on the space $\mathbb{R}^{\mathbb{N}}$.

Proof. Note that
$$\lim_{n \to \infty} n^{3/2} \mathbb{P}[E_j^n \neq \emptyset] = 0.$$
Then

$$
\begin{aligned}
\mathbb{P}\left[\sum_{j \leq kn^{3/2}} 1_{\{E_j^n \neq \emptyset\}} = 0; k = 1, .., r \right] &= \mathbb{P}\left[E_j^n = \emptyset; j \leq rn^{3/2} \right] \\
&= \left(1 - \mathbb{P}\left[E_j^n \neq \emptyset \right] \right)^{\lfloor rn^{3/2} \rfloor} \\
&\to 1.
\end{aligned}
$$

□

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.*

Lemma 6.7. *As $n \to \infty$,*

$$\left[(e^{-\lambda_n t} Z^n(t))_{t \in [0, \infty]}, \left((\mathcal{Y}_{v,j}^n(t))_{t \in [0, \infty)} \right)_{v \in \mathcal{G} \setminus \{u\}, j \in \mathbb{N}}, \left(\sum_{j \leq kn^{3/2}} 1_{\{E_j^n \neq \emptyset\}} \right)_{k \in \mathbb{N}} \right]$$

converges in distribution to

$$\left[(e^{-\lambda t} Z^*(t))_{t \in [0, \infty]}; \left((Y_{v,j}(t)e_v)_{t \in [0, \infty)} \right)_{v \in \mathcal{G} \setminus \{u\}, j \in \mathbb{N}}, (0)_{k \in \mathbb{N}} \right]$$

on

$$\mathbb{D}([0, \infty], \mathbb{R}) \times \mathbb{D} \left([0, \infty), \mathbb{R}^{\mathcal{G}} \right)^{\mathcal{G} \setminus \{u\} \times \mathbb{N}} \times \mathbb{R}^{\mathbb{N}},$$

where the $Z^*(\cdot)$ and $Y_{v,j}(\cdot)$ are independent.

27

Proof. The convergence seen in Lemmas 6.2, 6.4, and 6.5 is joint convergence over the product space due to independence. □

We are yet to say how the random variables in (9) 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:

$$\lim_{n \to \infty} (e^{-\lambda_n t} Z^n(t))_{t \in [0,\infty]} = (e^{-\lambda t} Z^*(t))_{t \in [0,\infty]} \tag{11}$$

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

$$\lim_{n \to \infty} (\mathcal{Y}^n_{v,j}(t))_{t \in [0,\infty)} = (Y_{v,j}(t)e_v)_{t \in [0,\infty)} \tag{12}$$

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

$$\left(\sum_{j \leq kn^{3/2}} 1_{\{E^n_j \neq \emptyset\}} \right)_{k \in \mathbb{N}} \to (0)_{k \in \mathbb{N}} \tag{13}$$

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 (11), (12), and (13) hold, one can bring in Skorokhod's Representation Theorem, to use with Lemma 6.7.

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 $\{E^n_j = (v, w)\}$, which occur at the times S^n_j 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

$$\tau_n := \min\{t \geq 0 : Z^n(t) \in \{0, n\}\}, \tag{14}$$

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 τ_n.

Lemma 6.8.
$$\sup_{n \in \mathbb{N}} n^{-3/2} D_n < \infty$$

almost surely.

Proof. For each $n \in \mathbb{N}$, let $(R_j^n)_{j \in \mathbb{N}}$ be a sequence of i.i.d. random variables with

$$\mathbb{P}[R_j^n = x] = \begin{cases} \alpha_n/(\alpha_n + \beta_n), & x = 1; \\ \beta_n/(\alpha_n + \beta_n), & x = -1; \end{cases}$$

so

$$\left(1 + \sum_{j=1}^{k} R_j^n \right)_{k \in \mathbb{N}}$$

is a random walk, whose distribution matches that of the discrete-time embedded chain of $Z^n(\cdot)$. Write

$$\rho_n = \min \left\{ k \in \mathbb{N} : 1 + \sum_{j=1}^{k} R_j^n \in \{0, n\} \right\}$$

for the number of steps until the walk hits n or 0. Then the number of downsteps before hitting n or 0 is

$$D_n \stackrel{d}{=} \sum_{j=1}^{\rho_n} 1_{\{R_j^n = -1\}} \leq \rho_n.$$

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 $\{\rho_n > n^{3/2}\} \subset \{1 + \sum_{j=1}^{\lfloor n^{3/2} \rfloor} R_j^n < n\}$, so

$$
\begin{aligned}
\mathbb{P}[D_n > n^{3/2}] &\leq \mathbb{P}\left[1 + \sum_{j=1}^{\lfloor n^{3/2} \rfloor} R_j^n < n\right] \\
&\leq \mathbb{P}\left[\left(\sum_{j=1}^{\lfloor n^{3/2} \rfloor} R_j^n - \frac{\lfloor n^{3/2} \rfloor \lambda_n}{\alpha_n + \beta_n}\right)^2 \right. \\
&\qquad \left. > \left(\frac{\lfloor n^{3/2} \rfloor \lambda_n}{\alpha_n + \beta_n} + 1 - n\right)^2\right] \\
&\leq \left(\frac{\lfloor n^{3/2} \rfloor \lambda_n}{\alpha_n + \beta_n} + 1 - n\right)^{-2} \mathrm{Var}\left[\sum_{j=1}^{\lceil n^{3/2} \rceil} R_j^n\right] \\
&\leq cn^{-3/2},
\end{aligned}
$$

(15)

(16)

(17)

for some constant $c > 0$. Inequality (16) holds for large enough n and Inequality (16) is Chebyshev's inequality. Finally, (17) gives that

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

and the result is proven by Borel-Cantelli. $\qquad\square$

Now it is to be seen that double mutations occurring before time τ_n can be neglected.

Lemma 6.9. *Let* $v, w \in \mathcal{G}\backslash\{u\}$. *As* $n \to \infty$,

$$
K_{v,w}^n(\tau_n) \to 0
$$

almost surely.

Proof. From Lemma 6.8, $C := \sup_{n\in\mathbb{N}} n^{-3/2} D_n < \infty$. Then

$$
\begin{aligned}
K^n_{v,w}(\tau_n) &= \sum_{j=1}^{D_n} 1_{\{E^n_j=(v,w)\}} \\
&\leq \sum_{j=1}^{\lceil Cn^{3/2}\rceil} 1_{\{E^n_j=(v,w)\}} \\
&\leq \sum_{j=1}^{\lceil Cn^{3/2}\rceil} 1_{\{E^n_j\neq\emptyset\}}.
\end{aligned}
$$

By (13) this converges to zero as $n \to \infty$. $\qquad\square$

6.4 Convergence of genome counts

The purpose of this section is to show that $\mathcal{X}^n(\sigma_n)$ converges when conditioned on the event $\{W^* > 0\}$ (W^* is defined in Remark 6.3). The times τ_n (defined in (14)) will play the role of a helpful stepping stone in the proof.

Lemma 6.10. *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 (11). $\quad\square$

Lemma 6.11. *Condition on $\{W^* > 0\}$. Suppose that $(a_n)_{n \in \mathbb{N}}$ is a sequence of real-valued random variables on the same probabiity space as everything else, with*

$$\lim_{n \to \infty} a_n = \infty,$$

$$a_n \le \tau_n$$

for each n, and

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

almost surely. Then, almost surely,

$$\lim_{n \to \infty} K_v^n(a_n) = \begin{cases} K_v^*(l), & v \in \mathcal{G}_i, i \in \mathcal{S}; \\ 0, & v \in \mathcal{G}_{\ge 2}; \end{cases}$$

where

$$K_v^*(s) = N_v(2\lambda^{-1}\alpha(u')\theta_i^{u_i,v_i}e^{\lambda s}).$$

Moreover, for $v \in \mathcal{G}_i$ and $j \in \mathbb{N}$,

$$\lim_{n \to \infty} \left(a_n - T_{v,j}^n \right) = l - T_{v,j}^*$$

almost surely, where

$$T_{v,j}^* = \min\{s \in \mathbb{R} : K_v^*(s) = j\}.$$

Proof. Let $t \in \mathbb{R}$. Since $Z^n(\tau_n) = n$,

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

Thanks to (11):

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

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

So, using dominated convergence,

$$\lim_{n\to\infty} n^{-1} \int_0^{a_n+t} Z^n(s)ds = \lambda^{-1} e^{\lambda(l+t)}$$

almost surely. Note also that

$$\lim_{n\to\infty} n p_n(u,v) = \begin{cases} \theta_i^{u_i,v_i}, & v \in \mathcal{G}_i; \\ 0, & v \in \mathcal{G}_{\geq 2}. \end{cases}$$

Then

$$\lim_{n\to\infty} 2 p_n(u,v)\alpha(u') \int_0^{a_n+t} Z^n(s)ds = 2\lambda^{-1}\alpha(u')\theta_i^{u_i,v_i} e^{\lambda(l+t)}.$$

Hence, recalling (10),

$$K_v^n(a_n+t) = N_v\left(2p_n(u,v)\alpha(u') \int_0^{a_n+t} Z^n(s)ds\right)$$

converges almost surely to

$$K_v^*(l+t) = N_v\left(2\lambda^{-1}\alpha(u')\theta_i^{u_i,v_i} e^{\lambda(l+t)}\right),$$

because $N_v(\cdot)$ is almost surely continuous at any fixed point.

Finally we check convergence of the $a_n - T_{v,j}^n$. Let $\epsilon > 0$. For sufficiently large n,

$$K_v^n(a_n - l + T_{v,j}^* + \epsilon) = K_v^*(T_{v,j}^* + \epsilon) \geq j;$$

so

$$a_n - l + T_{v,j}^* + \epsilon \geq T_{v,j}^n,$$

or equivalently

$$a_n - T_{v,j}^n \geq l - T_{v,j}^* - \epsilon.$$

The argument is now repeated for an upper bound. For sufficiently large n,

$$K_v^n(a_n - l + T_{v,j}^* - \epsilon) = K_v^*(T_{v,j}^* - \epsilon) < j;$$

so

$$a_n - l + T_{v,j}^* - \epsilon < T_{v,j}^n,$$

or equivalently

$$a_n - T_{v,j}^n < l - T_{v,j}^* + \epsilon.$$

$\square$

Lemma 6.12. *Condition on $\{W^* > 0\}$. Suppose that (a_n) satisfies the conditions of Lemma 6.11. Then, almost surely,*

$$\lim_{n\to\infty} \left(\mathcal{X}^n(a_n) - Z^n(a_n)e_u\right) = \sum_{i\in\mathcal{S}}\sum_{v\in\mathcal{G}_i} e_v \sum_{j=1}^{K_v^*(l)} Y_{v,j}(l - T_{v,j}^*),$$

where the $Y_{v,j}(\cdot)$ are from Lemma 6.4 and the $K_v^*(\cdot)$ and $T_{v,j}^*$ are from Lemma 6.11.

Remark 6.13. *By definition, $T_{v,j}^* \leq l$ for $j = 1, .., K_v^*(l)$. So the limit in Lemma 6.12 is well defined.*

Proof of Lemma 6.12. Recall that

$$\mathcal{X}^n(a_n) - Z^n(a_n)e_u = \sum_{v\in\mathcal{G}\setminus\{u\}} \sum_{j=1}^{K_v^n(a_n)} \mathcal{Y}_{v,j}^n(a_n - T_{v,j}^n) \qquad (18)$$

$$+ \sum_{v,w\in\mathcal{G}\setminus\{u\}} \sum_{j=1}^{K_{v,w}^n(a_n)} \mathcal{Y}_{v,w,j}^n(a_n - T_{v,w,j}^n).$$

The 'double mutation' term in (18) converges to zero, because

$$K_{v,w}^n(a_n) \leq K_{v,w}^n(\tau_n),$$

which converges to zero by Lemma 6.9. As for the 'single mutation' term in (18), Lemma 6.11 says that the $K_v^n(a_n)$ and $a_n - T_{v,j}^n$ converge to $K_v^*(l)$ and $l - T_{v,j}^*$, while (12) says that the $\mathcal{Y}_{v,j}^n(\cdot)$ converge to $e_v Y_{v,j}(\cdot)$. $\qquad\square$

Lemma 6.14. *Condition on $\{W^* > 0\}$.*

$$\lim_{n\to\infty} (\sigma_n - \tau_n) = 0$$

almost surely.

Proof. By Lemma 6.10, for large enough n, $Z^n(\tau_n) = n$. So $|\mathcal{X}^n(\tau_n)| \geq n$, and hence $\sigma_n \leq \tau_n$. Therefore

$$\liminf_{n\to\infty}(\sigma_n - \tau_n) \leq 0.$$

34

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 6.12,

$$|\mathcal{X}^n(\sigma_{n_k}) - Z^n(\sigma_{n_k})e_u|$$

converges, and so must be a bounded sequence. However it is also true that, taking $k \to \infty$,

$$
\begin{aligned}
|\mathcal{X}^n(\sigma_{n_k}) - Z^n(\sigma_{n_k})e_u| &= n_k - Z^{n_k}(\sigma_{n_k}) \\
&= n_k\left(1 - \frac{Z^{n_k}(\sigma_{n_k})}{e^{\lambda_{n_k}\sigma_{n_k}}}\frac{e^{\lambda_{n_k}\tau_{n_k}}}{Z^{n_k}(\tau_{n_k})}e^{\lambda_{n_k}(\sigma_{n_k} - \tau_{n_k})}\right) \\
&\sim n_k(1 - e^{\lambda l}),
\end{aligned}
$$

which is unbounded. $\qquad\square$

Lemma 6.15. *Condition on $\{W^* > 0\}$.*

$$\lim_{n \to \infty}(\mathcal{X}^n(\sigma_n) - Z^n(\sigma_n)e_u) = \sum_{i \in \mathcal{S}}\sum_{v \in \mathcal{G}_i}e_v\sum_{j=1}^{K_v^*(0)}Y_{v,j}(-T_{v,j}^*),$$

almost surely.

Proof. By Lemma 6.14, $(\sigma_n) = (a_n)$ satisfies the conditions of Lemma 6.11 with $l = 0$. Then Lemma 6.12 gives the result. $\qquad\square$

Let's look at the limit in Lemma 6.15. For $v \in \mathcal{G}_i$, $K_v^*(0)$ is Poisson distributed with mean $2\lambda^{-1}\alpha(u')\theta_i^{u_i,v_i}$. Conditional on $K_v^*(0)$, the times $(-T_{v,j}^*)_{j=1}^{K_v^*(0)}$, unordered, are i.i.d. exponentially distributed random variables with mean λ^{-1}. So

$$\sum_{j=1}^{K_v^*(0)}Y_{v,j}(-T_{v,j}^*)$$

35

has generalised Luria-Delbrück distribution with parameters

$$\left(\lambda, \alpha(v'), \beta(v'), 2\lambda^{-1}\alpha(u')\theta_i^{u_i, v_i} \right).$$

Therefore the limit of Lemma 6.15 is a vector of independent generalised Luria-Delbrück distributions:

$$\sum_{i \in \mathcal{S}} \sum_{v \in \mathcal{G}_i} e_v \sum_{j=1}^{K_v^*(0)} Y_{v,j}(-T_{v,j}^*) \overset{d}{=} \sum_{v \in \mathcal{G} \setminus \{u\}} e_v X_v,$$

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 6.16.

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 6.16 into several lemmas; the idea is that the random variable W^n be used as an intermediary.

Lemma 6.17.
$$\lim_{n \to \infty} \mathbb{P}[W^* > 0, W^n = 0] = 0.$$

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

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

So

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

36

Therefore, by dominated convergence,

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

$\square$

Lemma 6.18.
$$\mathbb{P}[W^n > 0, \sigma_n = \infty] = 0.$$

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

Proof of Part 1 of Proposition 6.16.

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

as $n \to \infty$, by Lemmas 6.17 and 6.18. $\square$

The structure for the proof of Part 2 of Proposition 6.16 is much the same as that of Part 1. However the details will require a little extra work.

Lemma 6.19.
$$\lim_{n \to \infty} \mathbb{P}[W^* = 0, W^n > 0] = 0.$$

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

$$W^n < \epsilon.$$

So
$$\lim_{n \to \infty} 1_{\{W^* = 0, W^n \geq \epsilon\}} = 0$$

almost surely. Then by dominated convergence,

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

37

Meanwhile for each n,

$$\mathbb{P}[W^n \in (0, \epsilon)] = \frac{\lambda_n}{\alpha_n}\left(1 - e^{-\frac{\lambda_n}{\alpha_n}\epsilon}\right) \le \epsilon$$

(the distribution of W^n is seen in [3, 7]). Therefore

$$\limsup_{n\to\infty} \mathbb{P}[W^* = 0, W^n > 0]$$
$$\le \limsup_{n\to\infty} \mathbb{P}[W^* = 0, W^n \ge \epsilon] + \limsup_{n\to\infty} \mathbb{P}[W^n \in (0, \epsilon)]$$
$$\le \epsilon.$$

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

Lemma 6.20.
$$\lim_{n\to\infty} \mathbb{P}[W^n = 0, \sigma_n < \infty] = 0.$$

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 $Z^n(\tau_n) = 0$, $K_v^n(\cdot) = 0$ and $K_{v,w}^n(\cdot) = 0$ for all $v, w \in \mathcal{G}\backslash\{u\}$, then

$$\sup_{t\ge 0} |\mathcal{X}^n(t)| < n,$$

which means that $\sigma_n = \infty$. Equivalently,

$$\{\sigma_n < \infty\} \subset \{Z^n(\tau_n) = n\} \cup \{\exists v, K_v^n(\cdot) \ne 0\} \cup \{\exists(v, w), K_{v,w}^n(\cdot) \ne 0\}$$
$$= \{Z^n(\tau_n) = n\} \cup \{\exists v, K_v^n(\cdot) \ne 0\}$$
$$\cup \{\exists(v, w), K_{v,w}^n(\cdot) \ne 0, Z^n(\tau_n) = 0\},$$

where the equality relies on the fact that $\{Z^n(\tau_n) = 0\} \cup \{Z^n(\tau_n) = n\}$ covers the whole probability space. It follows that

$$\mathbb{P}[W^n = 0, \sigma_n < \infty] \le \mathbb{P}[W^n = 0 | Z^n(\tau_n) = n]$$
$$+ \sum_{v\in\mathcal{G}\backslash\{u\}} \mathbb{P}[K_v^n(\cdot) \ne 0 | W^n = 0]$$
$$+ \sum_{v,w\in\mathcal{G}\backslash\{u\}} \mathbb{P}[K_{v,w}^n(\cdot) \ne 0 | Z^n(\tau_n) = 0]. \quad (19)$$

38

We will show that each term of the right hand side of Inequality (19) converges to zero. Firstly,

$$\mathbb{P}[W^n = 0 | Z^n(\tau_n) = n] \;=\; \left(\frac{\beta_n}{\alpha_n}\right)^n,$$

which is the probability that $Z^n(\cdot)$, if starting at size n, eventually goes extinct; this clearly converges to zero.

Secondly,

$$\mathbb{E}\left[\sup_t K_v^n(t) \Big| W^n = 0\right]$$
$$= \mathbb{E}\left[N_v^n\left(2p_n(u,v)\alpha(u')\int_0^\infty Z^n(s)ds\right)\Big| W^n = 0\right]$$
$$= \mathbb{E}\left[\mathbb{E}\left[N_v^n\left(2p_n(u,v)\alpha(u')\int_0^\infty Z^n(s)ds\right)\Big| Z^n(\cdot)\right]\Big| W^n = 0\right]$$
$$= \mathbb{E}\left[2p_n(u,v)\alpha(u')\int_0^\infty Z^n(s)ds \Big| W^n = 0\right]$$
$$= 2p_n(u,v)\alpha(u')\int_0^\infty \mathbb{E}[Z^n(s)|W^n = 0]ds$$
$$= 2p_n(u,v)\alpha(u')\int_0^\infty e^{-\lambda_n s}ds$$
$$\to 0,$$

because $p_n(u,v) \to 0$. Hence

$$\mathbb{P}\left[\sup_t K_v^n(t) \neq 0 \Big| W^n = 0\right] \to 0.$$

Lastly,

$$\mathbb{P}[K_{v,w}^n(\cdot) \neq 0 | Z^n(\tau_n) = 0] \;=\; \mathbb{P}[K_{v,w}^n(\tau_n) \neq 0 | Z^n(\tau_n) = 0]$$
$$\leq \; \frac{\mathbb{P}[K_{v,w}^n(\tau_n) \neq 0]}{\mathbb{P}[Z^n(\tau_n) = 0]}.$$

But $\mathbb{P}[K_{v,w}^n(\tau_n) \neq 0]$ converges to zero by Lemma 6.9, while $\mathbb{P}[Z^n(\tau_n) = 0]$ converges to $\mathbb{P}[W^* = 0] > 0$ by (11). □

39

Proof of Part 2 of Proposition 6.16. Just as for Part 1,

$$
\begin{aligned}
\mathbb{P}[W^* = 0, \sigma_n < \infty] \;=\; & \mathbb{P}[W^* = 0, \sigma_n < \infty, W^n > 0] \\
& + \mathbb{P}[W^* = 0, \sigma_n < \infty, W^n = 0] \\
\leq\; & \mathbb{P}[W^* = 0, W^n < 0] \\
& + \mathbb{P}[\sigma_n < \infty, W^n = 0] \to 0
\end{aligned}
$$

as $n \to \infty$, by Lemmas 6.19 and 6.20. $\qquad\square$

Corollary 6.21 (to Proposition 6.16). *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

$$
\begin{aligned}
& \mathbb{P}[H_n, \sigma_n < \infty] + \mathbb{P}[H_n, W^* > 0, \sigma_n = \infty] \\
& = \mathbb{P}[H_n, W^* > 0] + \mathbb{P}[H_n, W^* = 0, \sigma_n < \infty],
\end{aligned}
$$

and take $n \to \infty$. $\qquad\square$

Finally we are in a position to prove Theorem 6.1.

Proof of Theorem 6.1. For any $R \subset (\mathbb{N}_0)^{\mathcal{G}\setminus\{u\}}$,

$$
\begin{aligned}
\lim_{n\to\infty} & \frac{\mathbb{P}\left[(X_v^{\mu_n}(\sigma_n))_{v\in\mathcal{G}\setminus\{u\}} \in R, \sigma_n < \infty\right]}{\mathbb{P}[\sigma_n < \infty]} \\
& = \lim_{n\to\infty} \frac{\mathbb{P}\left[(X_v^{\mu_n}(\sigma_n))_{v\in\mathcal{G}\setminus\{u\}} \in R, W^* > 0\right]}{\mathbb{P}[W^* > 0]} \qquad (20) \\
& = \mathbb{P}[(X_v)_{v\in\mathcal{G}\setminus\{u\}} \in R], \qquad\qquad\qquad\qquad\qquad (21)
\end{aligned}
$$

where (20) is due to Corollary 6.21 and (21) is due to Lemma 6.15.

$\qquad\square$

6.6 Law of large numbers

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

$$
\mathbb{E}\left[|\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}}\delta_{B_i^{n,\mu}}\{k\}\right] = |\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}}\mathbb{P}[B_i^{n,\mu}=k]
$$
$$
= |\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}_{\mathrm{sel}}}\mathbb{P}[B_i^{n,\mu}=k]
$$
$$
+|\mathcal{S}|^{-1}\sum_{j\in J}\sum_{i\in\mathcal{S}(j)}\mathbb{P}[B_i^{n,\mu}=k], \qquad (22)
$$

for $k\in\mathbb{N}_0$. The penultimate term of (22) vanishes:

$$
|\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}_{\mathrm{sel}}}\mathbb{P}[B_i^{n,\mu}=k]\to 0
$$

because $|\mathcal{S}_{\mathrm{sel}}|/|\mathcal{S}|\to 0$. The last term of (22) can be written as

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

where $\mathbb{P}[B^{n,\mu}(j)=k]=\mathbb{P}[B_i^{n,\mu}=k]$ for $i\in\mathcal{S}(j)$. But

$$
\frac{|\mathcal{S}(j)|}{|\mathcal{S}|}\to q(j),
$$

while Theorem 5.1 implies that

$$
\mathbb{P}[B^{n,\mu}(j)=k]\to\Lambda(j)\{k\}.
$$

Therefore the expected site frequency spectrum converges:

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

The variance is

$$
\begin{aligned}
\mathrm{Var}\left[|\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}}\delta_{B_i^{n,\mu}}\{k\}\right] &= |\mathcal{S}|^{-2}\sum_{i\in\mathcal{S}}\mathrm{Var}[1_{\{B_i^{n,\mu}=k\}}] \\
&\quad +|\mathcal{S}|^{-2}\sum_{\substack{i,j\in\mathcal{S}\\i\neq j}}\mathrm{Cov}[1_{\{B_i^{n,\mu}=k\}},1_{\{B_j^{n,\mu}=k\}}] \\
&\leq |\mathcal{S}|^{-1}+\max_{\substack{i,j\in\mathcal{S}\\i\neq j}}\mathrm{Cov}[1_{\{B_i^{n,\mu}=k\}},1_{\{B_j^{n,\mu}=k\}}].
\end{aligned}
$$

Because $\mathcal{S}_{\mathrm{sel}}$ and J are finite sets and the random variables are exchangable over $\mathcal{S}(j)$, the maximum is taken over a finite set. Theorem 5.1 says that the covariances converge to zero. $\qquad\square$

7 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 can offer an approximation for mutation frequencies, 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

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

for the cells which are mutated at site $i\in\mathcal{S}$, and

$$
\hat{\mathcal{B}}_i^\mu = \{y\in\mathcal{T}:\exists x\in\mathcal{B}_i^\mu,x\preceq y\}
$$

for their descendants. Recall that $\mathcal{T}_{\sigma_n}$ are the cells alive when the total number of cells reaches n. Note the inequality

$$
\hat{B}_i^{n,\mu} := |\hat{\mathcal{B}}_i^\mu\cap\mathcal{T}_{\sigma_n}| \geq |\mathcal{B}_i^\mu\cap\mathcal{T}_{\sigma_n}| = B_i^{n,\mu}. \tag{23}
$$

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 $i \in \mathcal{S}$ and $a \in (0,1)$. As $n \to \infty$ and $n\mu \to \theta \in [0, \infty)$,*

$$\mu^{-1} \mathbb{P}[n^{-1} \hat{B}_i^{n,\mu} > a] \to 2(a^{-1} - 1).$$

For this subsection we are always talking about a single site $i \in \mathcal{S}$; for convenience, let's drop the subscript i from the notation. To begin the proofs of Theorem 4.4 and Proposition 7.1, fix μ, and observe that $(B^{r,\mu})_{r \in \mathbb{N}}$ is a Markov process on the nonnegative integers with transition probabilities

$$\mathbb{P}[B^{r+1,\mu} = k | B^{r,\mu} = j]$$
$$= \begin{cases} \frac{j}{r} (\mu/3)^2, & k = j - 1; \\ \frac{j}{r} 2(\mu/3)(1 - \mu/3) + \frac{r-j}{r}(1 - \mu)^2, & k = j; \\ \frac{j}{r}(1 - \mu/3)^2 + \frac{r-j}{r} 2\mu(1 - \mu), & k = j + 1; \\ \frac{r-j}{r} \mu^2, & k = j + 2. \end{cases} \quad (24)$$

Here, j/r is the probability that one of the j mutant cells divides, while $\mu/3$ is the probability that a mutant's daughter reverts to the unmutated state. The process $(\hat{B}^{r,\mu})_{r \in \mathbb{N}}$ has transition probabilities

$$\mathbb{P}[\hat{B}^{r+1,\mu} = k | \hat{B}^{r,\mu} = j]$$
$$= \begin{cases} \frac{r-j}{r}(1 - \mu)^2, & k = j; \\ \frac{j}{r} + \frac{r-j}{r} 2\mu(1 - \mu), & k = j + 1; \\ \frac{r-j}{r} \mu^2, & k = j + 2. \end{cases} \quad (25)$$

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

$$\xi^\mu = \min\{r \in \mathbb{N} : B^{r,\mu} > 0\}$$

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

$$\Xi_j^\mu = \{B^{\xi^\mu,\mu} = j\}$$

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

Lemma 7.2. *For $r \in \mathbb{N}$,*

$$\lim_{\mu \to 0} \mu^{-1} \mathbb{P}[\xi^\mu = r, \Xi_j^\mu] = \begin{cases} 2, & j = 1; \\ 0, & j = 2. \end{cases}$$

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

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

Similarly

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

Divide by μ and take $\mu \to 0$. $\qquad\square$

The next result gives conditional mutation frequencies.

Lemma 7.3. *Let $a > 0$. As $n \to \infty$ and $n\mu \to \theta \in [0, \infty)$,*

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

and

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

Proof. Calculating from the transition probabilities (24),

$$\mathbb{E}\left[B^{s+1,\mu} | B^{s,\mu} = k\right] = k + s^{-1} k (1 - 8\mu/3) + 2\mu.$$

So

$$\mathbb{E}\left[(s + 1)^{-1} B^{s+1,\mu} | B^{s,\mu} = k\right] = s^{-1} k + s^{-1}(s + 1)^{-1} \left(2s - \frac{8}{3} k\right) \mu,$$

and hence

$$\begin{aligned} s^{-1} k - 2s^{-1} \mu \quad &\leq \quad \mathbb{E}\left[(s + 1)^{-1} B^{s+1,\mu} | B^{s,\mu} = k\right] \\ &\leq \quad s^{-1} k + 2s^{-1} \mu. \end{aligned} \tag{26}$$

44

For the rest of the proof we will condition on the event $\{\xi^\mu = r, \Xi_1^\mu\}$. That is, we will consider the processes $(B^{s,\mu})_{s \geq r}$ and $(\hat{B}^{s,\mu})_{s \geq r}$ conditioned on $B^{r,\mu} = \hat{B}^{r,\mu} = 1$. Write

$$\mathbb{E}_{r,1}[\cdot] = \mathbb{E}[\cdot | \xi^\mu = r, \Xi_1^\mu]$$

for the conditional expectation. From (26), for $s \geq r$,

$$\begin{aligned}
\mathbb{E}_{r,1}\left[s^{-1}B^{s,\mu}\right] - 2s^{-1}\mu &\leq \mathbb{E}_{r,1}\left[(s+1)^{-1}B^{s+1,\mu}\right] \\
&\leq \mathbb{E}_{r,1}\left[s^{-1}B^{s,\mu}\right] + 2s^{-1}\mu.
\end{aligned} \tag{27}$$

Combining (27) with

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

we have that, for $n \geq r$,

$$r^{-1} - 2\mu \sum_{s=r}^{n-1} s^{-1} \leq \mathbb{E}_{r,1}\left[n^{-1}B^{n,\mu}\right] \leq r^{-1} + 2\mu \sum_{s=r}^{n-1} s^{-1}.$$

Therefore, as $\mu \to 0$ and $n\mu \to \theta$,

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

In just the same manner,

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

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

$$\mathbb{E}_{r,1}\left[n^{-1}D^n\right] = r^{-1}.$$

Moreover a well-known result (e.g. [10]) says that, as $n \to \infty$, $n^{-1}D^n$ converges to a Beta random variable. That is, for $a \in (0,1)$,

$$\mathbb{P}_{r,1}[n^{-1}D^n > a] \to (1-a)^{r-1}, \tag{28}$$

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. The inequality

$$D^n \leq \hat{B}^{n,\mu}$$

gives that

$$\mathbb{E}_{r,1}|n^{-1}\hat{B}^{n,\mu} - n^{-1}D^n| = \mathbb{E}_{r,1}n^{-1}\hat{B}^{n,\mu} - \mathbb{E}_{r,1}n^{-1}D^n \to 0.$$

The inequality

$$B^{n,\mu} \leq \hat{B}^{n,\mu}$$

gives that

$$\mathbb{E}_{r,1}|n^{-1}\hat{B}^{n,\mu} - n^{-1}B^{n,\mu}| = \mathbb{E}_{r,1}n^{-1}\hat{B}^{n,\mu} - \mathbb{E}_{r,1}n^{-1}B^{n,\mu} \to 0.$$

$\square$

Lemma 7.4. *Consider $(\mu_n)_{n\in\mathbb{N}}$ with $n\mu_n \to \theta \in [0,\infty)$. Then*

$$\sup_{n\in\mathbb{N}} \mathbb{P}[n^{-1}B^{n,\mu_n} > a | \xi^{\mu_n} = r, \Xi_j^{\mu_n}] \quad \leq \quad \sup_{n\in\mathbb{N}} \mathbb{P}[n^{-1}\hat{B}^{n,\mu_n} > a | \xi^{\mu_n} = r, \Xi_j^{\mu_n}]$$

$$\leq \quad cr^{-2},$$

where $c > 0$ does not depend on r, j.

Proof. The first inequality is immediate. We prove the second. From the transition probabiities (25),

$$
\begin{aligned}
\mathbb{E}\left[(\hat{B}^{s+1,\mu_n})^2 | \hat{B}^{s,\mu_n} = k\right] &= k^2\left(1 + 2s^{-1} - 4\mu_n s^{-1}\right) \\
&\quad + k\left(4\mu_n + s^{-1}(1 - 2\mu_n - 2\mu_n^2)\right) \\
&\quad + 2\mu_n + 2\mu_n^2 \\
&\leq k^2(s+1)^2 s^{-2} + 4n\mu_n + 2\mu_n + 2\mu_n^2 + 1.
\end{aligned}
$$

So, for $s \leq n$,

$$\mathbb{E}\left[(s+1)^{-2}(\hat{B}^{s+1,\mu_n})^2 | \hat{B}^{s,\mu_n} = k\right] \leq s^{-2}k^2 + s^{-2}c_1, \tag{29}$$

46

where $c_1 > 0$ is a constant which does not depend on k, n, s. Now let's condition on $\{\xi^{\mu_n} = r, \Xi_j^{\mu_n}\}$, again writing $\mathbb{E}_{r,j}$ for the conditional expectation. From (29), for $s \in \{r, .., n\}$,

$$\mathbb{E}_{r,j}[(s+1)^{-2}(\hat{B}^{s+1,\mu_n})^2] - \mathbb{E}_{r,j}[s^{-2}(\hat{B}^{s,\mu_n})^2] \le s^{-2}c_1.$$

This leads to, for $s \in \{r, .., n\}$,

$$
\begin{aligned}
\mathbb{E}_{r,j}[s^{-2}(\hat{B}^{s,\mu_n})^2] &\le \mathbb{E}_{r,j}[r^{-2}(\hat{B}^{r,\mu_n})^2] + c_1 \sum_{t=r}^{s-1} t^{-2} \\
&= r^{-2}j^2 + c_1 \sum_{t=r}^{s-1} s^{-2} \\
&\le c_2 r^{-1},
\end{aligned}
\tag{30}
$$

where c_2 is a constant which does not depend on j, n, r, s. (In the following, $c_3, .., c_6$ will also be constants.) Calculating third moments from the transition probabilities,

$$
\begin{aligned}
\mathbb{E}\left[(\hat{B}^{s+1,\mu_n})^3 | \hat{B}^{s,\mu_n} = k\right] &= k^3 \left(1 + 3s^{-1}(1 - 2\mu_n - 4\mu_n^2)\right) \\
&\quad + k^2 \left(6\mu_n + 3s^{-1}(1 - 2\mu_n - 6\mu_n^2)\right) \\
&\quad + k \left(6\mu_n + 6\mu_n^2 - s^{-1}(1 + 2\mu_n + 6\mu_n^2)\right) \\
&\quad + 2\mu_n + 6\mu_n^2 \\
&\le k^3(s+1)^3 s^{-3} + c_3 k^2 s^{-1} + c_4 k n^{-1}.
\end{aligned}
$$

Then

$$\mathbb{E}\left[(s+1)^{-3}(\hat{B}^{s+1,\mu_n})^3 | \hat{B}^{s,\mu_n} = k\right] \le s^{-3}k^3 + c_3 s^{-4}k^2 + c_4 s^{-2} n^{-1}.$$

Hence

$$
\begin{aligned}
&\mathbb{E}_{r,j}\left[(s+1)^{-3}(\hat{B}^{s+1,\mu_n})^3\right] - \mathbb{E}_{r,j}\left[s^{-3}(\hat{B}^{s,\mu_n})^3\right] \\
&\le c_3 s^{-2} \mathbb{E}_{r,j}[s^{-2}(\hat{B}^{s,\mu_n})^2] + c_4 s^{-2} r^{-1},
\end{aligned}
$$

which combined with (30) gives that

$$
\begin{aligned}
\mathbb{E}_{r,j}[n^{-3}(\hat{B}^{n,\mu_n})^3] &= \mathbb{E}_{r,j}[r^{-3}(\hat{B}^{r,\mu_n})^3] + c_5 r^{-1} \sum_{s=r}^{n-1} s^{-2} \\
&\le r^{-3} + c_6 r^{-2} \\
&\le c r^{-2}.
\end{aligned}
$$

Apply Markov's inequality to conclude. □

Proof of Theorem 4.4 and Proposition 7.1.

$$\mu^{-1}\mathbb{P}[n^{-1}B^{n,\mu} > a]$$

$$= \sum_{r=2}^{\infty}\sum_{j=1}^{2}\mu^{-1}\mathbb{P}[\xi^{\mu} = r, \Xi_j^{\mu}]\mathbb{P}[n^{-1}B^{n,\mu} > a|\xi^{\mu} = r, \Xi_j^{\mu}]. \qquad (31)$$

Lemmas 7.2, 7.3, and 7.4, along with the Dominated Convergence Theorem, show that the limit of (31) is

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

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

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 $x \in \mathcal{T}$).

Aldous [2], in a different language to ours, gives a similar result to Lemma 3.2. Rather than adapting his result, we now give a distinct proof of Lemma 3.2.

Proof of Lemma 3.2. Write

$$\mathcal{D}_x = \{y \in \mathcal{T} : x \preceq y\}$$

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

$$\mathcal{D}_{x,t} = \mathcal{D}_x \cap \mathcal{T}_t$$

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

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

48

Hence
$$\left(\left|\mathcal{D}_{x,\sum_{y\prec x}A_y+t}\right|\right)_{t\geq 0}$$
is measurable with respect to the sigma-algebra generated by $(A_y)_{y\in\mathcal{D}_x}$, and has the same distribution as
$$\left(\left|\mathcal{D}_{\emptyset,t}\right|\right)_{t\geq 0} = \left(\left|\mathcal{T}_t\right|\right)_{t\geq 0}.$$

It follows that
$$\lim_{t\to\infty} e^{-t}\left|\mathcal{D}_{x,\sum_{y\prec x}A_y+t}\right| =: W_x$$
almost surely, where $W_x \sim \mathrm{Exp}(1)$; moreover if $x, y \in \mathcal{T}$ are such that $\mathcal{D}_x \cap \mathcal{D}_y = \emptyset$, then W_x and W_y are independent. In particular, W_{x0} and W_{x1} are independent. Now,
$$\lim_{t\to\infty}\frac{|\mathcal{D}_{x0,t}|}{|\mathcal{D}_{x,t}|} = \lim_{t\to\infty}\frac{|\mathcal{D}_{x0,t}|}{1+|\mathcal{D}_{x0,t}|+|\mathcal{D}_{x1,t}|} = \frac{W_{x0}}{W_{x0}+W_{x1}} =: U_{x0}$$
almost surely, and $U_{x0} + U_{x1} = 1$. A standard calculation shows that U_{x0} is uniformly distributed on $(0,1)$: for $u \in (0,1)$,
$$\mathbb{P}[U_{x0} < u] = \int_0^\infty \int_{z(1-u)/u}^\infty e^{-y}e^{-z}dydz = u.$$

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 $(u,v)\in(0,1)\times(0,\infty)$,
$$
\begin{aligned}
\mathbb{P}[U_{x0} < u, W_{x0}+W_{x1} < v] &= \int_0^{uv}\int_{z(1-u)/u}^{v-z}e^{-y}e^{-z}dydz \\
&= u(1-(1+v)e^{-v}) \\
&= \mathbb{P}[U_{x0} < u]\mathbb{P}[W_{x0}+W_{x1} < v].
\end{aligned}
$$

Now fix $l \in \mathbb{N}$. Because U_{x0} and $W_{x0}+W_{x1}$ are measurable with respect to the sigma-algebra generated by $(A_y)_{y\in\mathcal{D}_x\setminus\{x\}}$, we have that
$$\left[(U_{x0})_{|x|=l}, (W_{x0}+W_{x1})_{|x|=l}, (A_x)_{|x|\leq l}\right] \tag{32}$$

49

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

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

Then, thanks to the independence of (32), $(U_{x0})_{x \in \mathcal{T}: |x| \le l}$ 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*

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

is almost surely finite.

Proof. For $t \ge 0$, let $\mathcal{F}_t$ be the sigma-algebra generated by $(\mathcal{T}_s)_{s \in [0,t]}$. For $t \ge s \ge 0$, conditional on $\mathcal{F}_s$, $(P_{y,t})_{y \in \mathcal{T}_s}$ is exchangable. So for $y \in \mathcal{T}$,

$$\mathbb{E}[1_{\{y \in \mathcal{T}_s\}} P_{y,t} | \mathcal{F}_s] = \frac{1_{\{y \in \mathcal{T}_s\}}}{|\mathcal{T}_s|}.$$

Now let $x \in \mathcal{T} \backslash \{\emptyset\}$. We have

$$P_{x,t} \ge 1_{\{|\mathcal{D}_{x,s}| > 0\}} P_{x,t} = \sum_{y \in \mathcal{D}_{x,s}} P_{y,t}$$

and hence

$$\mathbb{E}[P_{x,t} | \mathcal{F}_s] \ge P_{x,s}.$$

That is, $(P_{x,t})_{t \ge 0}$ is a submartingale with respect to $(\mathcal{F}_t)_{t \ge 0}$. Then by Doob's inequality,

$$\mathbb{P}[\exists t \ge 0, P_{x,t} > \epsilon] = \mathbb{P}[\sup_{t \ge 0} P_{x,t} > \epsilon] \le \epsilon^{-2} \mathbb{E}[(P_x)^2].$$

But P_x is simply a product of $|x|$ independent Uniform$(0, 1)$ random variables, where $|x| \in \mathbb{N}$ is the generation of x (that is, $x \in \{0, 1\}^{|x|}$). So $\mathbb{E}[(P_x)^2] = 3^{-|x|}$. Hence

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

Now

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

so the Borel-Cantelli lemma concludes the proof. $\qquad\qquad\qquad\qquad$ □

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

$$x_j \prec x_k \implies j < k. \tag{33}$$

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 μ, 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 $i \in \mathcal{S}$.

Remark 7.6. ϕ_i^μ 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 σ_n (when n total cells are reached) which are descendants of

51

cell ϕ_i^μ. Phrased another way, $P_{\phi_i^\mu, \sigma_n}$ is the fraction of cells alive at time σ_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 $|\mathcal{S}|\mu \to \eta \in [0, \infty)$,*

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

in distribution, with respect to the vague topology on the space of measures on $(0, 1]$.

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

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

Lemma 7.8. *As $\mu \to 0$ and $|\mathcal{S}|\mu \to \eta \in [0, \infty)$,*

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

in distribution, where the M_x are i.i.d. Poisson(η) random variables.

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

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

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

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

$$\times \mathbb{P}\left[\left(M_{x_j}^{\mu, \mathcal{S}} \right)_{j=2}^k = (m_j)_{j=2}^k \right]. \tag{34}$$

Due to the property (33) of the enumeration, $M_{x_{k+1}}^{\mu, \mathcal{S}}$ conditioned on the event $\left(M_{x_j}^{\mu, \mathcal{S}} \right)_{j=2}^k = (m_j)_{j=2}^k$ is just a binomial random variable with parameters $|\mathcal{S}| - \sum_{j=2}^k m_j$ and μ. Therefore (34) converges as required. $\qquad\square$

Proof of Proposition 7.7. Fix a sequence of sets of sites $(\mathcal{S}_n)_{n\in\mathbb{N}}$ and a sequence of mutation rates $(\mu_n)_{n\in\mathbb{N}}$ with $\mu_n|\mathcal{S}_n| \to \eta$. Apply Skorokhod's Representation Theorem to Lemma 7.8 to obtain random variables $(M_x^n)_{x\in\mathcal{T}\setminus\{\emptyset\},n\in\mathbb{N}}$ and $(M_x')_{x\in\mathcal{T}\setminus\{\emptyset\}}$ which satisfy

1. $(M_x^n)_{x\in\mathcal{T}\setminus\{\emptyset\}} \overset{d}{=} (M_x^{\mu_n,\mathcal{S}_n})_{x\in\mathcal{T}\setminus\{\emptyset\}}$ for each $n \in \mathbb{N}$;

2. $(M_x')_{x\in\mathcal{T}\setminus\{\emptyset\}} \overset{d}{=} (M_x)_{x\in\mathcal{T}\setminus\{\emptyset\}}$; and

3. $\lim_{n\to\infty}(M_x^n)_{x\in\mathcal{T}\setminus\{\emptyset\}} = (M_x')_{x\in\mathcal{T}\setminus\{\emptyset\}}$ almost surely.

Put the M_x^n, M_x' on the same probability space as the $P_{x,t}, P_x$ so that the M_x^n, M_x' are independent of the $P_{x,t}, P_x$. Then

$$\sum_{i\in\mathcal{S}} \delta_{P_{\phi_i^\mu,\sigma_n}} = \sum_{x\in\mathcal{T}\setminus\{\emptyset\}} M_x^{\mu_n,\mathcal{S}_n} \delta_{P_{x,\sigma_n}}$$
$$\overset{d}{=} \sum_{x\in\mathcal{T}\setminus\{\emptyset\}} M_x^n \delta_{P_{x,\sigma_n}}.$$

Let $I_1, .., I_k \subset (0,1]$ be closed intervals. Then

$$\left(\sum_{i\in\mathcal{S}} \delta_{P_{\phi_i^\mu,\sigma_n}}(I_j)\right)_{j=1}^k \overset{d}{=} \left(\sum_{x\in\mathcal{T}\setminus\{\emptyset\}} M_x^n \delta_{P_{x,\sigma_n}}(I_j)\right)_{j=1}^k. \tag{35}$$

Lemma 3.2 says that P_{x,σ_n} converges to P_x; and P_x does not lie on the boundaries of the I_j with probability one, so the summands of (35) converge pointwise. Meanwhile Lemma 7.5 says that the sum is over a finite subset of $\mathcal{T}\setminus\{\emptyset\}$. $\qquad\square$

Lemma 7.9. *Let* $a \in (0,1)$. *As* $n \to \infty$, $\mu \to 0$, *and* $|\mathcal{S}|\mu \to \eta \in [0,\infty)$,

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

Proof. First,

$$\liminf \mathbb{E}\left[\sum_{i\in\mathcal{S}}\delta_{P_{\phi_i^n,\sigma_n}}(a,1)\right] \geq \mathbb{E}\left[\lim\sum_{i\in\mathcal{S}}\delta_{P_{\phi_i^n,\sigma_n}}(a,1)\right]$$

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

$$= 2\eta(a^{-1}-1),$$

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

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

$$\to 2\eta(a^{-1}-1),$$

by the inequality $P_{\phi_i^n,\sigma_n} \leq n^{-1}\hat{B}_i^{n,\mu}$ and then Proposition 7.1. □

7.4 The infinite sites assumption approximation

Lemma 7.10. *Let $a \in (0,1)$. As $n \to \infty$, $n\mu \to \theta \in [0,\infty)$, and $|\mathcal{S}|\mu \to \eta \in [0,\infty)$,*

$$\mathbb{E}\left|\sum_{i\in\mathcal{S}}\delta_{P_{\phi_i^n,\sigma_n}}(a,1) - \sum_{i\in\mathcal{S}}\delta_{n^{-1}B_i^{n,\mu}}(a,1)\right| \to 0.$$

Proof. The inequalities

$$n^{-1}B_i^{n,\mu} \leq n^{-1}\hat{B}_i^{n,\mu} \geq P_{\phi_i^\mu,\sigma_n},$$

imply that

$$\sum_{i\in\mathcal{S}}\delta_{n^{-1}B_i^{n,\mu}}(a,1) \leq \sum_{i\in\mathcal{S}}\delta_{n^{-1}\hat{B}_i^{n,\mu}}(a,1) \geq \sum_{i\in\mathcal{S}}\delta_{P_{\phi_i^\mu,\sigma_n}}(a,1).$$

54

Hence

$$\mathbb{E}\left|\sum_{i\in\mathcal{S}}\delta_{P_{\phi_i^n,\sigma_n}}(a,1)-\sum_{i\in\mathcal{S}}\delta_{n^{-1}B_i^{n,\mu}}(a,1)\right|$$

$$\leq\mathbb{E}\left|\sum_{i\in\mathcal{S}}\delta_{n^{-1}\hat{B}_i^{n,\mu}}(a,1)-\sum_{i\in\mathcal{S}}\delta_{P_{\phi_i^n,\sigma_n}}(a,1)\right|$$

$$+\mathbb{E}\left|\sum_{i\in\mathcal{S}}\delta_{n^{-1}\hat{B}_i^{n,\mu}}(a,1)-\sum_{i\in\mathcal{S}}\delta_{n^{-1}B_i^{n,\mu}}(a,1)\right|$$

$$=2\mathbb{E}\sum_{i\in\mathcal{S}}\delta_{n^{-1}\hat{B}_i^{n,\mu}}(a,1)-\mathbb{E}\sum_{i\in\mathcal{S}}\delta_{P_{\phi_i^n,\sigma_n}}(a,1)-\mathbb{E}\sum_{i\in\mathcal{S}}\delta_{n^{-1}B_i^{n,\mu}}(a,1),$$

which converges to zero thanks to Corollary 4.5, Proposition 7.1, and Lemma 7.9. □

Proof of Theorem 4.6. Let $I_1,..,I_k\subset(0,1]$ be closed intervals. Writing $||\cdot||$ for the l_1-norm on $\mathbb{R}^k$,

$$\mathbb{E}\left\|\left(\sum_{i\in\mathcal{S}}\delta_{n^{-1}B_i^{n,\mu}}(I_j)\right)_{j=1}^k-\left(\sum_{i\in\mathcal{S}}\delta_{P_{\phi_i^\mu,\sigma_n}}(I_j)\right)_{j=1}^k\right\|$$

$$=\sum_{j=1}^k\mathbb{E}\left|\sum_{i\in\mathcal{S}}\delta_{n^{-1}B_i^{n,\mu}}(I_j)-\sum_{i\in\mathcal{S}}\delta_{P_{\phi_i^\mu,\sigma_n}}(I_j)\right|$$

$$\to0,$$

due to Lemma 7.10. Therefore

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

and

$$\left(\sum_{i\in\mathcal{S}}\delta_{P_{\phi_i^\mu,\sigma_n}}(I_j)\right)_{j=1}^k$$

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

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

55

as required. ☐

7.5   Mean and variance of the site frequency spectrum

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

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

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

$$
\begin{aligned}
\mathbb{P}[P_x > a] &= \mathbb{P}[-\log(P_x) < -\log(a)] \\
&= \int_0^{-\log(a)} \frac{s^{|x|-1}e^{-s}}{(|x|-1)!}ds.
\end{aligned}
$$

So

$$
\begin{aligned}
\mathbb{E}\left[\sum_x M_x \delta_{P_x}(a,1)\right] &= \eta \sum_x \mathbb{P}[P_x > a] \\
&= \eta \sum_x \int_0^{-\log(a)} \frac{s^{|x|-1}e^{-s}}{(|x|-1)!}ds \\
&= \eta \int_0^{-\log(a)} \sum_{l \in \mathbb{N}} 2^l \frac{s^{l-1}e^{-s}}{(l-1)!}ds \\
&= 2\eta \int_0^{-\log(a)} e^s ds \\
&= 2\eta(a^{-1} - 1).
\end{aligned}
$$

56

As for the variance of the site frequency spectrum,

$$
\begin{aligned}
\mathrm{Var}\left[\sum_x M_x \delta_{P_x}(a,1)\right] &= \mathrm{Var}\left[\mathbb{E}\left[\sum_x M_x \delta_{P_x}(a,1)\Big|(P_x)\right]\right] \\
&\quad + \mathbb{E}\left[\mathrm{Var}\left[\sum_x M_x \delta_{P_x}(a,1)\Big|(P_x)\right]\right] \\
&\geq \mathbb{E}\left[\mathrm{Var}\left[\sum_x M_x \delta_{P_x}(a,1)\Big|(P_x)\right]\right] \\
&= \mathbb{E}\left[\sum_x \delta_{P_x}(a,1)\mathrm{Var}\left[M_x\right]\right] \\
&= \eta\mathbb{E}\left[\sum_x \delta_{P_x}(a,1)\right].
\end{aligned}
$$

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 [27, 5, 26, 8]. However, recent statistical analysis of single cell sequencing data shows "widespread violations of the ISA in human cancers" [20]. Thus it is unclear to what extent [27, 5, 26, 8]'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.

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

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

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

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

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

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

Then the number of sites to violate the ISA,

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

is binomially distributed with parameters $|\mathcal{S}|$ 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 α and die at rate β. 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 [5].

Depite the apparent prevalence of ISA violations, our results suggest that their impact on mutation frequencies is negligible at the scale of population fractions. Importantly, bulk sequencing data is only sensitive on the scale of population fractions. Our theoretical work stands in support of the data-driven works [27, 5, 26, 8, 20].

Note however 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 [20].

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 [28]). We imagine an idealised diploid world.

To illustrate the diploid structure, label the genetic sites as

$$\mathcal{S} = \{1, 2\} \times \{1, .., L\},$$

for some $L \in \mathbb{N}$. The first coordinate of a site $(i, j) \in \mathcal{S}$ 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 $(i, j) \in \mathcal{S}$ 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

$$F_j = \frac{1}{2n} \sum_{i=1}^{2} B_{i,j}^{n,\mu} \tag{36}$$

for $j = 1, .., L$. 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 [29] (499017, Table S2). The data is plotted in Figures 3 and 4. This is just one tumor to illustrate our results. A broader picture of data is seen in [27, 5]. They analysed hundreds of tumors. Around 1/3 of the tumors were said to have a power-law distribution for mutation frequencies, resembling the one we consider.
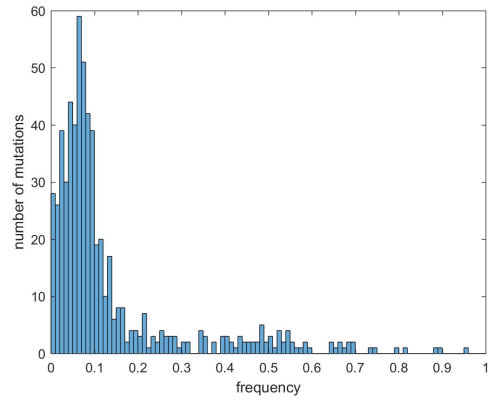
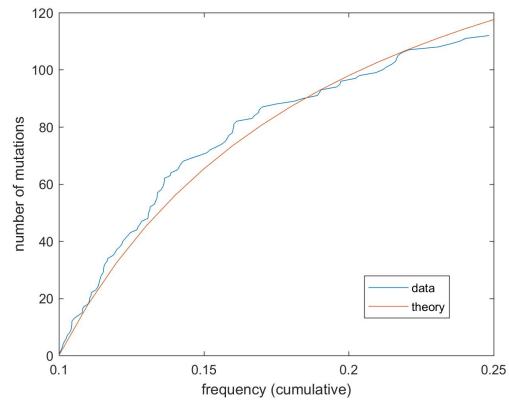Figure 3: A histogram of mutation frequencies from a lung adenocarcinoma.

Figure 4: The number of mutations (of the lung adenocarcinoma) whose frequency is in the interval $(0.1, x)$, for $x \in (0.1, 0.25)$.

Our method to estimate mutation rates is, to a large extent, inspired by [27, 5]. 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

$$\mathcal{M}(a, b) = |\{j \in \{1, .., L\} : F_j \in (a, b)\}|$$

for the number of mutations with frequency in $(a, b) \subset (0.1, 0.25)$. Then, adapting Corollary 4.5 to (36), the expected number of mutations with frequency in (a, b) is

$$\mathbb{E}\mathcal{M}(a, b) \approx \mu |\mathcal{S}| (a^{-1} - b^{-1}). \tag{37}$$

Under different models, [27, 5] derive the same approximation (37). They estimate the mutation rate μ by applying a linear regression to (37). We simplify matters even further. Our estimator for μ is

$$\hat{\mu} = \frac{\mathcal{M}(0.1, 0.25)}{6|\mathcal{S}|}, \tag{38}$$

which (37) 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 $|\mathcal{S}| = 3 \times 10^8$ [5]. And the number of mutations in the specified frequency range is $\mathcal{M}(0.1, 0.25) = 112$. This gives

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

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

$$\mathcal{S} = \mathcal{S}_A \cup \mathcal{S}_C \cup \mathcal{S}_G \cup \mathcal{S}_T,$$

where

$$\mathcal{S}_\chi = \{i \in \mathcal{S} : u_i = \chi\}$$

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

$$\hat{\mu}_\chi = \frac{\mathcal{M}_\chi(0.1, 0.25)}{6|\mathcal{S}_\chi|}$$

61

is an unbiased estimator for μ_χ. 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 [8] for a far more comprehensive statistical analysis of mutation frequency data. However their infinite-sites framework does not consider mutation rate heterogeneity.

Appendix

A heuristic 'proof' of Conjecture 5.3 is given.

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

$$\mathcal{G}_{\mathrm{sel}} = \{v \in \mathcal{G} : \exists i \in \mathcal{S}_{\mathrm{sel}}, v_i \neq u_i\}$$

for the set of genomes which are mutated at a selective site. Write

$$Q_{\mathrm{sel}}^\mu(t) = \frac{\sum_{v \in \mathcal{G}_{\mathrm{sel}}} X_v^\mu(t)}{\sum_{v \in \mathcal{G}} X_v^\mu(t)}$$

for the proportion of cells at time $t \geq 0$ whose genomes are mutated at a selective site. Then, according to Theorem 6.1,

$$(Q_{\mathrm{sel}}^\mu(\sigma_n^\mu)|\sigma_n^\mu < \infty) \to 0$$

in probability. Therefore we neglect selection.

Cells divide and die at rates $\alpha(u')$ and $\beta(u')$, which we now abbreviate to α and β. 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. [10]) 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 2β. A mortal cell divides at rate β to produce two mortal cells, or it dies at rate α. 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, $\mathfrak{T} = \cup_{l=0}^{\infty}\{0,1\}^l$ and its partial ordering $\prec$, will be used to represent the immortal cells. Let $(A_x)_{x \in \mathfrak{T}}$ be i.i.d. $\mathrm{Exp}(\alpha - \beta)$ 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

$$\mathfrak{T}_t = \left\{ x \in \mathfrak{T} : \sum_{y \prec x} A_y \leq t < \sum_{y \preceq x} A_y \right\}.$$

The immortal descendants of $x \in \mathfrak{T}$ are

$$\mathcal{D}_x^I = \{y \in \mathfrak{T} : x \preceq y\}.$$

The number of immortal descendants of cell x at time t is

$$D_{x,t}^I = |\mathcal{D}_x^I \cap \mathfrak{T}_t|.$$

Let $((R_x(t))_{t \geq 0})_{x \in \mathfrak{T}}$ be i.i.d. Poisson processes with rate 2β. Write $R_{x,r} = \min\{t \geq 0 : R_x(t) = r\}$ for $r = 1, .., R_x(A_x)$. 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 β and α. Then the number of mortal descendants of x at time t is

$$D_{x,t}^M = \sum_{y \in \mathcal{D}_x} \sum_{r=1}^{R_y(A_y)} 1_{\{t - S_{y,r} \geq 0\}} Y_{y,r}(t - S_{y,r}).$$

The number of descendants of x at time t is

$$D_{x,t} = D_{x,t}^I + D_{x,t}^M.$$

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 [3], and was mentioned in its specific form by [10].

63

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

$$\lim_{t \to \infty} \frac{D^I_{x,t}}{D_{x,t}} = c$$

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 $x \in \mathfrak{T} \backslash \{\emptyset\}$,*

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

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

Proof. By Lemma 3.2 and Lemma A.1,

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

converges to the required limit. $\qquad\square$

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

$$\eta := \sum_{j \in J} \sum_{\psi \in \mathcal{N} \backslash \{u(j)\}} \eta^{u(j),\psi}(j).$$

Each $x \in \mathfrak{T} \backslash \{\emptyset\}$ witnesses $1 + R_x(A_x)$ cell divisions, while $\emptyset$ witnesses $R_\emptyset(A_\emptyset)$ cell divisions (one less because there is not a cell division associated to the initiation of $\emptyset$). So the number of new mutations to arise at x is

$$\begin{cases} \sum_{r=0}^{R_x(A_x)} M_{x,r}, & x \neq \emptyset; \\ \sum_{r=1}^{R_x(A_x)} M_{x,r}, & x = \emptyset; \end{cases}$$

64

where the $M_{x,r}$ are i.i.d. Poisson random variables with mean η. 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.

Declarations of interest

None.

Acknowledgements

References

[1] D. Aldous. Stopping times and tightness. *Annals of Probability*, 6:335–340, 1978.

[2] D. Aldous. Probability distributions on cladograms. *Random Discrete Structures. The IMA Volumes in Mathematics and its Applications*, 76, 1996.

[3] K. B. Athreya and P. Ney. *Branching Processes*. Dover Publications, 2004.

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

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

[6] N. Champagnat and A. Lambert. Splitting trees with neutral Poissonian mutations I: small families. *Stochastic Processes and their Applications*, 122, 2010.

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

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

[9] J. Duchamps and A. Lambert. Mutations on a random binary tree with measured boundary. *Annals of Applied Probability*, 28:2141–2187, 2018.

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

[11] R. Durrett. *Branching Process Models of Cancer*. Springer, 2014.

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

[13] R. Griffiths and A. Pakes. An infinite-alleles version of the simple branching process. *Advances in Applied Probability*, 20:489–524, 1988.

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

[15] 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.

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

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

[18] D. G. Kendall. Birth-and-death processes, and the theory of carcinogenesis. *Biometrika*, 47(1-2):13–21, 1960.

[19] 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.

[20] 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.

[21] 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.

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

[23] T. McDonald and M. Kimmel. A multitype infinite-allele branching process with applications to cancer evolution. *Journal of Applied Probability*, 52:864–876, 2015.

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

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

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

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

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

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

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