MUTATION FREQUENCIES IN A BIRTH–DEATH BRANCHING PROCESS

BY DAVID CHEEK[1] AND TIBOR ANTAL

University of Edinburgh

First, we revisit the stochastic Luria–Delbrück model: a classic two-type branching process which describes cell proliferation and mutation. We prove limit theorems and exact results for the mutation times, clone sizes and number of mutants. Second, we extend the framework to consider mutations at multiple sites along the genome. The number of mutants in the two-type model characterises the mean site frequency spectrum in the multiple-site model. Our predictions are consistent with previously published cancer genomic data.

1. Introduction. Luria and Delbrück's famous work of 1943 combined mathematical modelling with experiment [28]. They considered an exponentially growing population of bacterial cells which is sensitive to attack by a lethal virus. The bacteria may mutate to become resistant to the virus. Lea and Coulson [27] obtained a probability distribution for the number of mutants, commonly known as the Luria–Delbrück distribution. The distribution has seen empirical evidence and become a standard tool for the estimation of mutation rates in bacteria [33]. While early formulations of the model were semi-deterministic, stochastic cell growth was subsequently incorporated (see [35] for a review). Notably, Kendall allowed for cells to grow as birth–death branching processes [21].

Kendall's two-type branching process, often referred to as the stochastic Luria–Delbrück model, has been foundational in the mathematical understanding of cancer evolution. The model and various extensions have been used to study drug resistance [5, 13, 16, 24], driver mutations [10, 11] and metastasis [7, 14, 29, 30], for example. As introduced by Kendall, wild-type (type A) and mutant (type B) cells are assumed to divide, die and mutate independently of each other, according

to

$$
\begin{cases}
A \to AA, & \text{rate } \alpha_A; \\
A \to \varnothing, & \text{rate } \beta_A; \\
A \to AB, & \text{rate } \nu; \\
B \to BB, & \text{rate } \alpha_B; \\
B \to \varnothing, & \text{rate } \beta_B.
\end{cases}
$$

Whether the model represents the emergence of drug resistance in cancer or bacteria, the total number of mutants is of key interest. In recent years, [1, 2, 15, 16, 22, 23, 25] derived exact and approximate distributions for the number of mutants at fixed times and population sizes.

Our first objective is to offer a mathematically rigorous account of the two-type model, looking at the number of mutants, mutation times and clone sizes (a clone is a subpopulation of mutant cells initiated by a mutation). Both previously known and new results are presented. We explore small mutation limits and long-term almost sure convergence. Specialising to neglect cell death, we give some exact distributions.

Our second objective is to introduce a neutral model of cancer evolution, which keeps track of mutations at S sites on the genome. A site refers to a base pair. In our multiple-site model, each cell is labelled by a sequence $(z_1, \ldots, z_S) \in \{0, 1\}^S$, where $z_i = 1$ means that the cell is mutated at site i. The number of mutants with respect to a particular site follows the two-type model. Thus many of the two-type results are applicable in the multiple-site setting.

A standard summary statistic of genomic data is the site frequency spectrum. It is defined as the number of sites who see mutations in k cells, for $k \in \mathbb{Z}_{\geq 0}$. We prove that the mean site frequency spectrum can be approximated with a generalisation of the Luria–Delbrück distribution. This result is consistent with cancer genomic data presented in [4, 34].

Of course, many works have attempted to predict mutation frequencies in cancer. Prominent examples are [4, 8, 31, 34], who gave approximations for the mean site frequency spectrum in a population of cancer cells. Every one of these works and countless others have used the infinite sites assumption, which says that each mutation occurs at a unique site. However, recent statistical analysis of cancer genomic data has refuted the validity of this simplification [26]. We do not use the infinite sites assumption, and make a theoretical argument against it.

The rest of the paper is organised as follows. In Section 2, we introduce the two-type model. In Section 3, we present long-term almost sure convergence results. In Section 4, we define and study the large population small mutation limit. In Section 5, we look at the large time small mutation limit. In Section 6, we present results on the number of mutants at a finite population size. In Section 7, we introduce the multiple-site model and present results on the site frequency spectrum. In Section 8, we discuss the multiple-site model in relation to recent works and

data. In Section 9, we present proofs of our main results. See the Appendix for a generalisation of the results of Sections 4 and 5.

2. Two-type model. The wild-type cells grow as a linear birth–death process $(A(t))_{t \geq 0}$, with birth and death rates α_A and β_A, respectively. That is to say, $(A(t))_{t \geq 0}$ is a continuous time Markov process on $\mathbb{Z}_{\geq 0}$ with transition rates

$$i \mapsto \begin{cases} i + 1, & \text{rate } i\alpha_A; \\ i - 1, & \text{rate } i\beta_A. \end{cases}$$

The initial number of wild-type cells $A(0) \in \mathbb{N}$ is fixed. The mutation rate is $\nu > 0$. Mutation events occur as a Cox process $(K(t))_{t \geq 0}$ with intensity $(\nu A(t))_{t \geq 0}$. The mutation times are

$$T_i := \inf\{t \geq 0 : K(t) = i\}$$

for $i \in \mathbb{N}$. Each mutation event initiates a clone which grows as a linear birth–death process with birth and death rates α_B and β_B. Clones grow independently of the wild-type growth and mutation times, and are represented by the i.i.d. processes $(Y_i(t))_{t \geq 0}$ for $i \in \mathbb{N}$, with $Y_i(0) = 1$. The total mutant population size at time t is

$$B(t) = \sum_{i=1}^{K(t)} Y_i(t - T_i).$$

Write $\lambda_A = \alpha_A - \beta_A$ and $\lambda_B = \alpha_B - \beta_B$ for the fitnesses of the wild-type and mutant cells. We shall only be concerned with the case of supercritical wild-type growth, $\lambda_A > 0$.

Note that the process counting the number of cells, $(A(t), B(t))_{t \geq 0}$, is a Markov process on $\mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$, with transition rates

$$(i, j) \mapsto \begin{cases} (i + 1, j), & \text{rate } i\alpha_A; \\ (i - 1, j), & \text{rate } i\beta_A; \\ (i, j + 1), & \text{rate } i\nu + j\alpha_B; \\ (i, j - 1), & \text{rate } j\beta_B. \end{cases}$$

We are interested in the process at a fixed time t, and at the random times

$$\sigma_n := \inf\{t \geq 0 : A(t) + B(t) \geq n\}$$

and

$$\tau_n := \inf\{t \geq 0 : A(t) \geq n\},$$

for $n \in \mathbb{N}$. Trivially, $\sigma_n \leq \tau_n$.

A classic application of the model is the emergence of drug resistance in cancer. Here, type A and B cells represent drug sensitive and resistant cells, respectively.

While the age of a tumour is typically unknown, its size can be measured. Thus the times σ_n are relevant.

Another interpretation of the model is metastasis. Here, type A cells make up the primary tumour, and the clones represent secondary tumours. In this case, the times τ_n are relevant.

3. Large time and population limits. Keeping the mutation rate fixed, the long-term behaviour of the model is mostly already well understood. Durrett and Moseley [11] study the case $\lambda_A < \lambda_B$. Janson [17] studies a broad class of urn models, which encompasses Kendall's model in the case $\lambda_A > \lambda_B$. We do not present results as detailed as Janson's. Our aim for this section is not to offer a comprehensive study, but rather bring together basic results which give valuable insight.

First, we make note of a classic result:

(1) $$\lim_{t \to \infty} e^{-\lambda_A t} A(t) = W$$

almost surely (see [3] or [9]). Here,

$$W \overset{d}{=} \sum_{i=1}^{A(0)} \chi_i \psi_i,$$

where the $\chi_i \sim \text{Bernoulli}(\lambda_A/\alpha_A)$ and the $\psi_i \sim \text{Exponential}(\lambda_A/\alpha_A)$ are independent.

REMARK 3.1. The event that the wild-type population eventually becomes extinct agrees with the event $\{W = 0\}$ almost surely.

We see a trichotomy, depending on the relative fitness of wild-type and mutant cells. Part 1 of Theorem 3.2 is a special case of [17], Theorem 3.1, and part 3 is [11], Theorem 2.

THEOREM 3.2 (Large time limit). *The following limits hold almost surely*:

1. *For $\lambda_A > \lambda_B$*,

$$\lim_{t \to \infty} e^{-\lambda_A t} B(t) = \frac{\nu}{\lambda_A - \lambda_B} W.$$

2. *For $\lambda_A = \lambda_B$*,

$$\lim_{t \to \infty} t^{-1} e^{-\lambda_A t} B(t) = \nu W.$$

3. *For $\lambda_A < \lambda_B$*,

$$\lim_{t \to \infty} e^{-\lambda_B t} B(t) = V.$$

The limit random variable W comes from (1). The limit random variable V is $[0, \infty)$-valued with mean

$$\mathbb{E}[V] = \frac{A(0)\nu}{\lambda_B - \lambda_A}.$$

The full distribution of V is given in [2], Section 4.3, which we do not state here for the sake of brevity.

For $\lambda_A \geq \lambda_B$, conditioned on wild-type nonextinction, any individual clone ultimately makes up zero proportion of the mutant population. That is to say, conditioned on $W > 0$,

$$\lim_{t \to \infty} \frac{Y_i(t - T_i)}{B(t)} = 0$$

almost surely. We say that the mutant population is driven by the wild-type growth. This is seen in the limit random variables' dependence on W.

For $\lambda_A < \lambda_B$, early arriving clones make an important contribution to the mutant population. Conditioned on $W > 0$,

$$\lim_{t \to \infty} \frac{Y_i(t - T_i)}{B(t)} = \frac{X_i e^{-\lambda_B T_i}}{V}$$

almost surely. Note that if $W > 0$, then $V > 0$ [11]. The X_i are i.i.d. with distribution $\chi^B \psi^B$, where $\chi^B \sim \text{Bernoulli}(\lambda_B/\alpha_B)$ and $\psi^B \sim \text{Exponential}(\lambda_B/\alpha_B)$ are independent. We say that the mutant population is driven by the clone growth.

To see the asymptotic behaviour of the number of mutations, simply consider $\alpha_B = \beta_B = 0$ in Theorem 3.2:

$$\lim_{t \to \infty} e^{-\lambda_A t} K(t) = \frac{\nu}{\lambda_A} W$$

almost surely.

As corollaries to Theorem 3.2, we obtain large population limits. Note that conditioned on $W > 0$, $\lim_{n \to \infty} \tau_n = \lim_{n \to \infty} \sigma_n = \infty$ almost surely.

COROLLARY 3.3 (Large wild-type population limit). *Conditioned on $W > 0$, the following limits hold almost surely*:

1. *For $\lambda_A > \lambda_B$,*

$$\lim_{n \to \infty} n^{-1} B(\tau_n) = \frac{\nu}{\lambda_A - \lambda_B}.$$

2. *For $\lambda_A = \lambda_B$,*

$$\lim_{n \to \infty} (n \log(n))^{-1} B(\tau_n) = \frac{\nu}{\lambda_A}.$$

3. *For $\lambda_A < \lambda_B$,*

$$\lim_{n \to \infty} n^{-\lambda_B/\lambda_A} B(\tau_n) = V W^{-\lambda_B/\lambda_A}.$$

COROLLARY 3.4 (Large total population limit). *Conditioned on $W > 0$, the following limits hold almost surely.*

1. *For $\lambda_A > \lambda_B$,*

$$\lim_{n \to \infty} n^{-1} B(\sigma_n) = \frac{\nu}{\lambda_A - \lambda_B + \nu}.$$

2. *For $\lambda_A = \lambda_B$,*

$$\lim_{n \to \infty} n^{-1} \log(n) (n - B(\sigma_n)) = \frac{\lambda_A}{\nu}.$$

3. *For $\lambda_A < \lambda_B$,*

$$\lim_{n \to \infty} n^{-\lambda_A/\lambda_B} (n - B(\sigma_n)) = V^{-\lambda_A/\lambda_B} W.$$

Note that $n - B(\sigma_n) = A(\sigma_n)$. In case 1, the wild-type and mutant cells come to coexist in a constant ratio. In cases 2 and 3, the mutant cells eventually dominate the overall population, with

$$\lim_{n \to \infty} n^{-1} B(\sigma_n) = 1 \tag{2}$$

almost surely.

4. Large population small mutation limit. A tumour may comprise around 10^9 cells upon detection, with mutation rates per base pair per cell division estimated as 5×10^{-10} in colorectal cancer [19], for example. Hence, a biologically relevant limit can be found by taking the population size to infinity and the mutation rate to zero, while keeping their product fixed.

Suppose that $(\nu_n)_{n \in \mathbb{N}}$ is a sequence of mutation rates satisfying

$$\lim_{n \to \infty} n \nu_n = \theta, \tag{3}$$

for some $\theta \in (0, \infty)$. For each $n \in \mathbb{N}$, consider the two-type model with mutation rate ν_n. For the wild-type population, mutant population, clone sizes, number of mutations and mutation times, write $A^{(n)}(\cdot)$, $B^{(n)}(\cdot)$, $Y_i^{(n)}(\cdot)$, $K^{(n)}(\cdot)$, and $T_i^{(n)}$, respectively. Write

$$\sigma_n' := \inf\{t \geq 0 : A^{(n)}(t) + B^{(n)}(t) \geq n\}$$

and

$$\tau_n' := \inf\{t \geq 0 : A^{(n)}(t) \geq n\}.$$

First, we see a connection between the times τ_n' and σ_n' in the large n limit.

PROPOSITION 4.1. *Conditioning on $\tau_n' < \infty$,*

$$\tau_n' - \sigma_n' \to 0$$

in probability, as $n \to \infty$.

All of our large population small mutation limit results will hold both in terms of the wild-type population and total population size. That is to say, using τ_n' or σ_n' as the time variable will yield the same distributions in the large n limit. To save writing each result twice, we introduce the sequence (ρ_n), which may refer to (τ_n') or (σ_n').

Underlying all subsequent results of this section is that the times of mutation centered about (ρ_n) converge.

THEOREM 4.2 (Mutations times).   *Conditioning on $\rho_n < \infty$,*

$$K^{(n)}(\rho_n + t) \to K^*(t)$$

in finite dimensional distributions, as $n \to \infty$. $K^*(t)$ is a Poisson process on $\mathbb{R}$ with intensity $\theta e^{\lambda_A t}$.

A direct consequence of Theorem 4.2 is that for each $i \in \mathbb{N}$, conditioning on $\rho_n < \infty$,

$$T_i^{(n)} - \rho_n \to T_i^* := \inf\{t \in \mathbb{R} : K^*(t) = i\}$$

in distribution, as $n \to \infty$. In particular, T_1^* has Gumbel distribution:

$$\mathbb{P}[T_1^* \geq t] = \exp\left(-\frac{\theta}{\lambda_A} e^{\lambda_A t}\right).$$

Next, we look at the number of mutants.

PROPOSITION 4.3 (Number of mutants).   *Conditioning on $\rho_n < \infty$,*

$$B^{(n)}(\rho_n + t) \to B^*(t) := \sum_{i=1}^{K^*(t)} Y_i\left(t - T_i^*\right)$$

in finite dimensional distributions, as $n \to \infty$. The $Y_i(\cdot)$ and $K^*(\cdot)$ are independent.

In particular, $B^{(n)}(\rho_n)$ converges in distribution to

$$(4) \qquad\qquad B^* = B^*(0) \overset{d}{=} \sum_{i=1}^{K^*} Y_i(\xi_i),$$

where $K^* = K^*(0) \sim \text{Poisson}(\theta/\lambda_A)$, and ξ_i are i.i.d. Exponential(λ_A) random variables independent of the $Y_i(\cdot)$ and $K^*(\cdot)$.

Here, ξ_i corresponds to the age of a randomly selected clone, and $Y_i(\xi_i)$ the size of the clone. From [3], page 109,

$$(5) \qquad\qquad \mathbb{E}\left[z^{Y_i(t)}\right] = \frac{\beta_B(z-1) - e^{-\lambda_B t}(\alpha_B z - \beta_B)}{\alpha_B(z-1) - e^{-\lambda_B t}(\alpha_B z - \beta_B)},$$

and so

$$
\begin{aligned}
(6) \qquad r(z) : &= \mathbb{E}\big[z^{Y_i(\xi_i)}\big] \\
&= \int_0^\infty \mathbb{E}\big[z^{Y_i(t)}\big] \lambda_A e^{-\lambda_A t}\, dt \\
&= 1 - (1 - q_B) F\left(\begin{matrix} 1, \lambda_A/\lambda_B \\ 1 + \lambda_A/\lambda_B \end{matrix}; \frac{q_B - z}{1 - z} \right).
\end{aligned}
$$

The function F is Gauss's hypergeometric function, and $q_B = \beta_B / \alpha_B$, which is a clone's ultimate extinction probability if $q_B \le 1$. The third equality of (6) can be seen by making a change of variable $s = e^{-\lambda_B t}$, and then using a standard integral representation for F (e.g., [20], C.8).

Clearly, B^* is a compound Poisson random variable (4), and has generating function

$$
(7) \qquad \mathbb{E}\big[z^{B^*}\big] = \exp\left(\frac{\theta}{\lambda_A} \big(r(z) - 1\big) \right).
$$

This recovers recent results of Kessler and Levine [23] who provided a heuristic derivation of this expression, and Keller and Antal [20] who derived it for a deterministic exponentially growing wild-type population. Its large θ limit appeared in Durrett and Moseley [11] for $\lambda_A < \lambda_B$ (see [20] for a discussion). If $\alpha_B = \lambda_A$ and $\beta_B = 0$, (7) reduces to the Luria–Delbrück distribution [27]:

$$
\mathbb{E}\big[z^{B^*}\big] = (1 - z)^{\frac{\theta}{\lambda_A}(z^{-1} - 1)}.
$$

REMARK 4.4. For $\lambda_B > 0$, the generating functions (6) and (7) yield power law tails:

$$
\lim_{k \to \infty} k^{1 + \lambda_A/\lambda_B} \mathbb{P}\big[Y_i(\xi_i) = k\big] = \frac{\lambda_A}{\lambda_B} (1 - q_B)^{1 - \lambda_A/\lambda_B} \Gamma(1 + \lambda_A/\lambda_B)
$$

and

$$
\lim_{k \to \infty} k^{1 + \lambda_A/\lambda_B} \mathbb{P}\big[B^* = k\big] = \frac{\theta}{\lambda_B} (1 - q_B)^{1 - \lambda_A/\lambda_B} \Gamma(1 + \lambda_A/\lambda_B),
$$

which are given in [20, 23, 30].

Of potential interest is the number of clones of a given size, perhaps above some lower limit for reliable detection. Let I be a subset of $\mathbb{Z}_{\ge 0}$. Consider

$$
C_I^{(n)}(t) = \sum_{i=1}^{K^{(n)}(t)} 1_{\{Y_i^{(n)}(t - T_i^{(n)}) \in I\}}(t),
$$

giving the number of clones whose size is in I at time t.

PROPOSITION 4.5 (Number of clones of a given size).   *Conditioning on $\rho_n < \infty$,*

$$C_I^{(n)}(\rho_n) \to C_I^* \sim \text{Poisson}\left(\frac{\theta}{\lambda_A}\mathbb{P}[Y_i(\xi_i) \in I]\right)$$

in distribution, as $n \to \infty$.

Consider

$$M^{(n)}(t) = \max_{1 \le i \le K^{(n)}(t)} Y_i^{(n)}(t - T_i^n),$$

giving the size of the largest clone at time t.

PROPOSITION 4.6 (Size of largest clone).   *Conditioning on $\rho_n < \infty$,*

$$M^{(n)}(\rho_n) \to M^* = \max_{1 \le i \le K^*} Y_i(\xi_i)$$

in distribution, as $n \to \infty$. Here, $\mathbb{P}[M^* \le k] = \exp(-\frac{\theta}{\lambda_A}\mathbb{P}[Y_i(\xi_i) > k])$.

For an example, we take the simplest choice of mutant cell growth: $\beta_B = 0$ and $\alpha_B = \lambda_A$. The number of clones above size k is

$$C_{\{i \in \mathbb{N}: i \ge k\}}^* \sim \text{Poisson}\left(\frac{\theta}{\lambda_A k}\right).$$

The size of the largest clone is

$$\mathbb{P}[M^* \le k] = \exp\left(-\frac{\theta}{\lambda_A(k+1)}\right).$$

REMARK 4.7.   In this section, we have considered a limit in which the product of the population size and mutation rate, $\theta = n\nu$, remains finite. It should be noted that alternative limits are also possible here. For example, Kessler and Levine [22] investigate large θ. In a different twist, Hamon and Ycart [15], Theorem 1.1, take the initial population size to infinity, the time of measurement to infinity, and the mutation rate to zero.

5. Large time small mutation limit.   Here, we investigate results similar to Section 4, but with a view to approximating the process at a fixed time rather than population size. Let (t_n) be a sequence of nonrandom times converging to infinity, and (ν_n) a sequence of mutation rates satisfying

$$\lim_{n \to \infty} e^{\lambda_A t_n}\nu_n = \eta,$$

for some $\eta \in (0, \infty)$. For each $n \in \mathbb{N}$, consider the two-type model with mutation rate ν_n. We use the superscript (n) notation established in Section 4.

PROPOSITION 5.1 (Mutation times).   *As $n \to \infty$,*

$$K^{(n)}(t_n + t) \to K^\circ(t)$$

in finite dimensional distributions. $K^\circ(t)$ is a Cox process on $\mathbb{R}$ with intensity $W\eta e^{\lambda_A t}$, where W is distributed as (1).

A direct consequence of Proposition 5.1 is that for each $i \in \mathbb{N}$,

$$T_i^{(n)} - t_n \to T_i^\circ := \inf\{t \in \mathbb{R} : K^\circ(t) = i\}$$

in distribution, as $n \to \infty$.

PROPOSITION 5.2 (Number of mutants).   *As $n \to \infty$,*

$$B^{(n)}(t_n + t) \to B^\circ(t) = \sum_{i=1}^{K^\circ(t)} Y_i(t - T_i^\circ),$$

in finite dimensional distributions. The $Y_i(\cdot)$ are independent of $K^\circ(\cdot)$.

Observe that

$$(8) \qquad B^\circ = B^\circ(0) \stackrel{d}{=} \sum_{i=1}^{K^\circ} Y_i(\xi_i),$$

where $K^\circ = K^\circ(0)$ conditioned on W is Poisson distributed with mean $W\eta/\lambda_A$. The generating function of B° is

$$
\begin{aligned}
\mathbb{E}[z^{B^\circ}] &= \mathbb{E}\left[\exp\left(\frac{W\eta}{\lambda_A}(r(z) - 1)\right)\right] \\
&= \left(\frac{\lambda_A^2 - \beta_A \eta(r(z) - 1)}{\lambda_A^2 - \alpha_A \eta(r(z) - 1)}\right)^{A(0)},
\end{aligned}
$$
(9)

where $r(z)$ is the clone size generating function, given by (6).

REMARK 5.3.   For $\lambda_B > 0$, the generating function (9) yields the same power-law tail as (6) and (7) (see Remark 4.4):

$$\lim_{k \to \infty} k^{1 + \lambda_A/\lambda_B} \mathbb{P}[B^\circ = k] = \frac{A(0)\eta}{\lambda_B}(1 - q_B)^{1 - \lambda_A/\lambda_B} \Gamma(1 + \lambda_A/\lambda_B).$$

The number of clones of a given size and the size of the largest clone can be seen in the large time small mutation limit. Simply replace K^* with K° in Propositions 4.5 and 4.6.

Finally, we comment that the large time small mutation limit justifies a common approximation of the model, in which the wild-type population grows as $(We^{\lambda_A t})_{t \in \mathbb{R}}$. Here, $B^\circ(\cdot)$ corresponds to $Z_1^*(\cdot)$ defined in [11], for example.

6. Finite size results. For simplicity, we consider $(A(0), B(0)) = (1, 0)$ in this section. However, it should not be too difficult to extend to arbitrary initial cell numbers.

We are able to give the distribution of $B(\tau_n)$ in the special case of no wild-type cell death.

PROPOSITION 6.1. *For $\beta_A = 0$,*

$$
(10) \qquad B(\tau_n) \overset{d}{=} \sum_{i=1}^{n-1} \sum_{j=1}^{K_i(\xi_i)} Y_{i,j}(U_{i,j}\xi_i),
$$

where $(K_i(t))_{t \geq 0}$ are Poisson processes with intensity ν, $Y_{i,j}(\cdot) \overset{d}{=} Y_i(\cdot)$, $\xi_i \sim$ Exponential(α_A), and $U_{i,j} \sim$ Uniform[0, 1], which are all independent.

To interpret (10), let us consider a randomly selected type A cell, labelled i, of the $n - 1$ cells present just before time τ_n. The cell has been alive for time ξ_i, and initiated $K_i(\xi_i)$ mutant clones, with mutation times $(1 - U_{i,j})\xi_i$ for $j = 1, 2, \ldots, K_i(\xi_i)$. The clone sizes are $Y_{i,j}(U_{i,j}\xi_i)$.

The mean number of mutant cells at time τ_n is

$$
\mathbb{E}\big[B(\tau_n)\big] = \begin{cases} \dfrac{(n-1)\nu}{\alpha_A - \lambda_B}, & \lambda_B < \alpha_A; \\ \infty, & \lambda_B \geq \alpha_A. \end{cases}
$$

The generating function of $B(\tau_n)$ is

$$
\mathbb{E}\big[z^{B(\tau_n)}\big] = \left[\int_0^\infty \alpha_A e^{-\alpha_A t} \exp\left(\nu t \int_0^1 \mathbb{E}\big[z^{Y_{i,j}(ut)}\big] - 1 \, du \right) dt \right]^{n-1}
$$

$$
= \left[\frac{1}{1 + \frac{\lambda_B \nu}{\alpha_A \alpha_B}} F\left(\begin{matrix} 1, \nu/\alpha_B \\ 1 + \nu/\alpha_B + \alpha_A/\lambda_B \end{matrix}; \frac{q_B - z}{q_B - 1} \right) \right]^{n-1},
$$

where $\mathbb{E}[z^{Y_{i,j}(ut)}]$ is given by (5). The computation is lengthy but straightforward; one can apply the integral expression [20], C.8, for the hypergeometric function, and the identity [20], C.10. As in Remarks 4.4 and 5.3, for $\lambda_B > 0$,

$$
(11) \qquad \lim_{k \to \infty} k^{1 + \alpha_A/\lambda_B} \mathbb{P}\big[B(\tau_n) = k\big] \in (0, \infty)
$$

exists. The limit can be obtained using the method of [20], Section 6 (which is based on [12]), but is too cumbersome to include here. Power-law tails have often appeared in two-type branching processes, but were generally considered to be an artefact of approximation [11, 35].

REMARK 6.2. Contrary to (11), moments of $B(\tau_n)$ are finite in the standard semi-deterministic version of the model (e.g., [27] and [20]).

Next, specialising further to neglect wild-type and mutant death, we connect the distributions of the $B(\sigma_n)$ and $B(\tau_n)$.

LEMMA 6.3. *For $\beta_A = \beta_B = 0$, and integers $0 \le k < n$,*

$$\mathbb{P}[B(\sigma_n) \le k] = \mathbb{P}[B(\tau_{n-k}) \le k].$$

A similar result was given by Janson [18], Lemma 9.1, for a different class of urn models. Although Lemma 6.3 can be combined with Proposition 6.1 to determine the distribution of $B(\sigma_n)$, it does not seem likely that a tractable explicit expression can be obtained in general. However, for neutral mutations, Angerer was able to solve a recursion for the probabilities $\mathbb{P}[B(\sigma_n) = k]$ [1], Corollary 2.2.

PROPOSITION 6.4 (Angerer). *For $\alpha_A + \nu = \alpha_B$ and $\beta_A = \beta_B = 0$,*

$$\mathbb{P}[B(\sigma_n) = k] = \sum_{i=1}^{n-k} (-1)^{n-i} \binom{n-k-1}{i-1} \binom{i\frac{\alpha_A}{\alpha_B} - 1}{n-1}.$$

7. Multiple site model and site frequency spectrum.
In the case of neutral mutations, we extend the two-type model to consider mutations at multiple sites on the genome.

The overall population $(C(t))_{t \ge 0}$ grows as a birth–death branching process. Cells divide and die at rates a and b, where $a > b$. Consider S sites, labelled $i \in \{1, \dots, S\}$. Each cell is labelled by some $(z_1, \dots, z_S) \in \{0, 1\}^S$, where $z_i = 1$ corresponds to a mutation at site i. Initially, there are an arbitrary number of cells all with label $(0, \dots, 0)$.

The mutations are modelled to occur in such a way that the number of mutants with respect to a particular site follows the two-type model. At each division event, the parent cell dies, and two daughter cells are produced. The daughter cells inherit the parent's mutations and may receive further mutations. Suppose that site i is not already mutated in the parent cell. With probability $1 - \mu$ site i does not receive a mutation in either daughter cell. With probability μ exactly one of the daughter cells receives a mutation at site i.

To state this more precisely, let us consider a parent cell with label $(z_1, \dots, z_S) \in \{0, 1\}^S$ dividing. The two daughter cells have labels $(Z_1^{[1]}, \dots, Z_S^{[1]})$ and $(Z_1^{[2]}, \dots, Z_S^{[2]})$, where for each i

$$(Z_i^{[1]}, Z_i^{[2]}) = \begin{cases} (z_i, z_i), & \text{probability } 1 - \mu; \\ (\min\{z_i + 1, 1\}, z_i), & \text{probability } \mu/2; \\ (z_i, \min\{z_i + 1, 1\}), & \text{probability } \mu/2. \end{cases}$$

REMARK 7.1. For our purposes, we do not need to specify the joint distribution of $(Z_i^{[1]}, Z_i^{[2]})_{i=1}^S$.

REMARK 7.2. As in Kendall's model, we neglect back mutations, and neglect the event that a cell division sees both daughter cells receiving the same mutation (see [25] for a biological justification).

For each $i \in \{1, \ldots, S\}$, let $B_i(t)$ be the number of cells at time t with $z_i = 1$.

Now we establish the connection between the multiple-site model and two-type model. Put $\alpha_A + \nu = \alpha_B = a$, $\beta_A = \beta_B = b$, $\nu = \mu a$. Then for each i,

$$\left(C(t) - B_i(t), B_i(t)\right)_{t \geq 0} \overset{d}{=} \left(A(t), B(t)\right)_{t \geq 0}.$$

The site frequency spectrum is defined to be the number of sites who see mutations in a given number of cells, that is, the sequence

$$\left(\sum_{i=1}^{S} 1_{\{B_i(t)=k\}}\right)_{k \in \mathbb{Z}_{\geq 0}}.$$

By linearity of expectation, the mean site frequency spectrum is determined by

$$(12) \qquad \mathbb{E}\left[\sum_{i=1}^{S} 1_{\{B_i(t)=k\}}\right] = S\mathbb{P}[B(t) = k].$$

Antal and Krapivsky [2] found the distribution of $B(t)$, by solving the Kolmogorov equations. For brevity, we do not state their result. To see the mean site frequency spectrum at a fixed population size, define

$$\sigma_n := \inf\{t \geq 0 : C(t) \geq n\},$$

as in the two-type model. Then for $b = 0$ and $C(0) = 1$,

$$\mathbb{E}\left[\sum_{i=1}^{S} 1_{\{B_i(\sigma_n)=k\}}\right] = S\sum_{i=1}^{n-k}(-1)^{n-i}\binom{n-k-1}{i-1}\binom{i\frac{\alpha_A}{\alpha_B}-1}{n-1},$$

by Proposition 6.4.

Let us return to the general setting of $b \geq 0$ and $C(0) \in \mathbb{N}$. We briefly comment on the long term behaviour of the site frequency spectrum. The number of sites who are mutated in a given number of cells converges to zero: for any $k \geq 1$

$$\lim_{t \to \infty} \sum_{i=1}^{S} 1_{\{B_i(t)=k\}} = 0,$$

almost surely. Let $x \in [0, 1)$. The number of sites who are mutated in at least proportion x of the population converges to S:

$$\lim_{n \to \infty} \sum_{i=1}^{S} 1_{\{B_i(\sigma_n) \geq xn\}} = S,$$

almost surely, due to (2).

Next, we look at the large population/time small mutation limits. Take a sequence of mutation probabilities $(\mu_n)_{n\in\mathbb{N}}$. For each $n \in \mathbb{N}$, consider the multiple-site model with mutation probability μ_n, birth rate a and death rate $b_n = b(1 - \mu_n)$ (with $\lambda = a - b > 0$). Write $C^{(n)}(t)$ for the population size and $B_i^{(n)}(t)$ for number of site i mutants at time t. Write

$$\sigma_n' := \inf\{t \geq 0 : C^{(n)}(t) \geq n\}.$$

The purpose of choosing the sequence of death rates (b_n) in this way is to allow for a straightforward adaptation of the two-type results.

PROPOSITION 7.3 (Large population small mutation limit). *Suppose that* (μ_n) *satisfies*

$$\lim_{n\to\infty} n\mu_n a = \theta,$$

for some $\theta \in (0, \infty)$. *Then*

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{S} 1_{\{B_i^{(n)}(\sigma_n')=k\}}\Big| \sigma_n' < \infty\right] = S\mathbb{P}[B^* = k],$$

where B^* *is distributed according to* (4) *with* $\alpha_A = \alpha_B = a$ *and* $\beta_A = \beta_B = a$.

PROPOSITION 7.4 (Large time small mutation limit). *Take a sequence of times* (t_n) *converging to infinity, with*

$$\lim_{n\to\infty} e^{\lambda t_n} \mu_n a = \eta,$$

for some $\eta \in (0, \infty)$. *Then*

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{S} 1_{\{B_i^{(n)}(t_n)=k\}}\right] = S\mathbb{P}[B^\circ = k],$$

where B° *is distributed according to* (8) *with* $\alpha_A = \alpha_B = a$ *and* $\beta_A = \beta_B = a$.

REMARK 7.5. Our approximations for the mean site frequency spectrum have power-law tails:

$$\lim_{k\to\infty} k^2 S\mathbb{P}[B^* = k] = \frac{S\theta}{\lambda}$$

and

$$\lim_{k\to\infty} k^2 S\mathbb{P}[B^\circ = k] = \frac{S\eta C(0)}{\lambda},$$

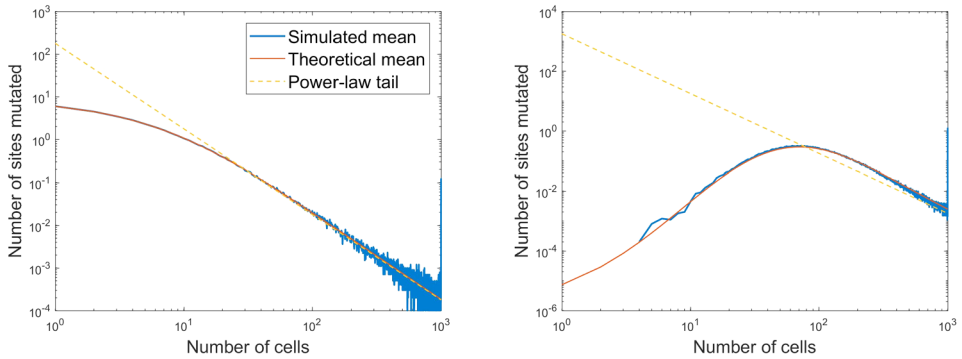which are special cases of Remarks 4.4 and 5.3.

FIG. 1. *Simulated and theoretical expected site frequency spectrum, with $a = 0.25$, $b = 0.18$, $S = 50$, $C(0) = 1$, $n = 10^3$. Two different mutation rates are plotted: $\mu = 10^{-3}$ (left) and $\mu = 10^{-2}$ (right). The average has been taken over 10^4 simulations in each case.*

Since the size, rather than age, of a tumour can be observed, we are most interested in the large population small mutation limit. To give the reader an idea of its appearance, in Figure 1 the mean site frequency spectrum as given by Proposition 7.3 is plotted. The theoretical result is compared to simulations, with birth, death and scaled mutation rates taken from biological literature. In particular, we consider $a = 0.25$ and $b = 0.18$ (per day), which were estimated in colorectal cancers by [6]. According to [19], θ may be of the order of a; we consider two different values for θ in this region. We take a relatively small population size of $n = 10^3$ and number of sites $S = 50$, so that computation time is reasonable. It is expected that taking larger n and fixed θ will give an even closer fit between theory and simulations.

8. Discussion. From a single cancer cell, a tumour may grow to comprise billions of cells. Mutations can occur at cell divisions, ultimately leading to great genetic diversity within a single tumour. With the advent of next-generation DNA sequencing, vast quantities of cancer genomes have been sequenced. Data has been made publicly available through the Cancer Genome Atlas and International Cancer Genome Consortium, for example. Considerable efforts have been made in recent years to explain observed mutation patterns with mathematical models, and from the observed mutation patterns to infer the evolutionary history of tumours.

Striking examples are Williams et al. [34] and Bozic et al. [4], who consider deterministic and branching process models, respectively. They both derive that the expected frequency of mutations occurring in x proportion of cells has density proportional to x^{-2} (away from 0). In [34], 323 out of 904 cancers considered are deemed to fit the x^{-2} power-law. In [4], 14 out of 42 cancers are deemed to fit the power-law.

The models of [4, 8, 31, 34] all used the infinite sites assumption, which states that each site can mutate at most once over the lifetime of a tumour. Statistical anal-

ysis of cancer genomic data refutes this assumption [26]. Furthermore, we make a theoretical argument against the infinite sites assumption in the branching process setting. According to Proposition 4.2, the number of times a particular site has mutated before the population size reaches n is approximately Poisson$(n\nu/\lambda_A)$. Therefore, the infinite sites simplification may be appropriate when $n\nu/\lambda_A$ is much smaller than 1. However, [34] estimated effective mutation rates, ν/λ_A, of single base pairs to be in the region of $10^{-7} - 10^{-6}$. If a detected tumour comprises $10^8 - 10^9$ cells (e.g., [4]), then $n\nu/\lambda_A$ is not sufficiently small.

In Proposition 7.3, we have shown that the mean site frequency spectrum can be approximated by a well-known generalisation of the Luria–Delbrück distribution. The distribution's x^{-2} tail agrees with theoretical predictions and data in [4, 34]. But our predictions disagree at the lower end of the frequency spectrum. Due to unreliable data, [4, 34] did not make a model-data comparison for mutations occurring in less than 10% of cells.

In upcoming work, we extend the multiple-site model to nonneutral mutations.

9. Proofs.

Proofs for Section 3.

PROOF OF THEOREM 3.2. For part 2, one needs to observe that

$$\left(e^{-\lambda_A t} B(t) - t e^{-\lambda_A t} \nu A(t)\right)_{t \geq 0}$$

is a martingale with respect to the obvious filtration, and is bounded in L_2.

For part 1, the reader may refer to [17] for a full proof in a more general and notation-heavy setting. For the reader's convenience, we offer the essence of Janson's proof here. Crucially,

$$\left(M(t)\right)_{t \geq 0} = \left(e^{-\lambda_B t} B(t) - \frac{\nu}{\lambda_A - \lambda_B} e^{-\lambda_B t} A(t)\right)_{t \geq 0}$$

is a martingale. Janson obtains bounds for the probabilities

$$\mathbb{P}\left[\sup_{t \in [n-1, n]} \left|e^{(\lambda_B - \lambda_A)t} M(t)\right| > \varepsilon\right],$$

via Doob's martingale inequality, and then applies the Borel–Cantelli lemma. $\square$

PROOF OF COROLLARY 3.4, PART 2. First, rewrite

$$\frac{\log(n) A(\sigma_n)}{n} = \frac{\log(A(\sigma_n) + B(\sigma_n)) A(\sigma_n)}{A(\sigma_n) + B(\sigma_n)}$$

$$= \frac{1}{\sigma_n}\left[\log\left(\frac{A(\sigma_n) + B(\sigma_n)}{\sigma_n e^{\lambda_A \sigma_n}}\right) + \log(\sigma_n) + \lambda_A \sigma_n\right]$$

$$\times \frac{e^{-\lambda_A \sigma_n} A(\sigma_n)}{\sigma_n^{-1} e^{-\lambda_A \sigma_n}(A(\sigma_n) + B(\sigma_n))}.$$

Then apply Theorem 3.2 and (1), to take $n \to \infty$. $\square$

The remaining parts of Corollaries 3.3 and 3.4 can be proven in a similar manner.

Proofs for Sections 4 and 5. For each $n \in \mathbb{N}$, the joint distribution of

$$(13) \qquad \left(A^{(n)}(\cdot), B^{(n)}(\cdot), \left(Y_i^{(n)}(\cdot)\right)_{i \in \mathbb{N}}, K^{(n)}(\cdot), \left(T_i^{(n)}\right)_{i \in \mathbb{N}}, \sigma_n', \tau_n'\right)$$

has been specified, with respect to the mutation rate ν_n. Note that the distributions of $A^{(n)}(\cdot)$ and $Y_i^{(n)}(\cdot)$ do not depend on n. We will construct the sequence (13) ranging over $n \in \mathbb{N}$ on a single probability space $(\Omega, \mathcal{F}, \mathbb{P})$ in a way that allows weak convergence to be shown via almost sure convergence.

On $(\Omega, \mathcal{F}, \mathbb{P})$ define the independent processes $(A(t))_{t \geq 0}$, $(Y_i(t))_{t \geq 0}$ for $i \in \mathbb{N}$, and $(N(t))_{t \geq 0}$. As one would expect we take $A(\cdot) \overset{d}{=} A^{(n)}(\cdot)$ and $Y_i(\cdot) \overset{d}{=} Y_i^{(n)}(\cdot)$. Take $N(\cdot)$ to be a Poisson counting process with intensity 1.

Define the mutation counting process by

$$K^{(n)}(t) = N\left(\int_0^t \nu_n A(s)\, ds\right).$$

The mutation times are given by

$$T_i^{(n)} = \inf\{t \geq 0 : K^{(n)}(t) = i\}.$$

The total mutant population is

$$B^{(n)}(t) = \sum_{i=1}^{K^{(n)}(t)} Y_i\left(t - T_i^{(n)}\right).$$

So the only dependence on n comes from the mutation times. As before, define

$$\sigma_n' = \inf\{t \geq 0 : A(t) + B^{(n)}(t) \geq n\}$$

and

$$\tau_n' = \inf\{t \geq 0 : A(t) \geq n\}.$$

The large population small mutation limit results all involve conditioning on $\sigma_n' < \infty$ or $\tau_n' < \infty$. Lemmas 9.1 and 9.2 will demonstrate that the results can be equivalently formulated by instead conditioning on nonextinction of the wild-type population.

LEMMA 9.1. *Suppose that* $(E_n)_{n \in \mathbb{N}}$ *and* $(F_n)_{n \in \mathbb{N}}$ *are sequences of events, such that*:

1. $\forall n \in \mathbb{N} \; (F_n \supset F_{n+1})$,
2. $\bigcap_{n \in \mathbb{N}} F_n = F$, *and*
3. $\mathbb{P}[F] > 0$.

Then

$$\lim_{n\to\infty} \mathbb{P}[E_n|F_n] = \lim_{n\to\infty} \mathbb{P}[E_n|F],$$

if it exists.

PROOF. Write

$$\mathbb{P}[E_n|F_n] = \frac{\mathbb{P}[F]}{\mathbb{P}[F_n]}\mathbb{P}[E_n|F] + \frac{\mathbb{P}[E_n \cap F_n \setminus F]}{\mathbb{P}[F_n]},$$

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

LEMMA 9.2.

$$\{W > 0\} = \bigcap_{n\in\mathbb{N}}\{\tau'_n < \infty\} = \bigcap_{n\in\mathbb{N}}\{\sigma'_n < \infty\},$$

where W is given by (1).

PROOF. That $\{W > 0\} \subset \bigcap_{n\in\mathbb{N}}\{\tau'_n < \infty\}$ and $\bigcap_{n\in\mathbb{N}}\{\tau'_n < \infty\} \subset \bigcap_{n\in\mathbb{N}}\{\sigma'_n < \infty\}$ should be clear. We show that

$$\bigcap_{n\in\mathbb{N}}\{\sigma'_n < \infty\} \subset \{W > 0\}.$$

Indeed, fix $\omega \in \{W = 0\} = \{\exists t \geq 0, A(t) = 0\}$.
Then

$$\int_0^\infty A(t)\,dt < \infty.$$

So one can choose sufficiently large $n \in \mathbb{N}$ such that both

$$v_n \int_0^\infty A(t)\,dt < \sup\{t \geq 0 : N(t) = 0\},$$

and

$$\tau'_n = \infty.$$

In this case, we must have

$$\omega \in \{\forall t \geq 0, K^{(n)}(t) = 0\} \cap \{\tau'_n = \infty\} \subset \{\sigma'_n = \infty\}. \qquad \square$$

We will now prove the results of Section 4 by conditioning on $W > 0$.

LEMMA 9.3. *Conditioning on $W > 0$, as $n \to \infty$,*

$$K^{(n)}(\tau'_n + t) = N\left(\int_{-\tau'_n}^t v_n A(\tau'_n + s)\,ds\right) \to N\left(\int_{-\infty}^t \theta e^{\lambda_A s}\,ds\right) =: K^*(t)$$

and

$$T_i^{(n)} - \tau'_n \to T_i^* = \inf\{t \in \mathbb{R} : K^*(t) = i\}$$

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

PROOF.   That $A(\cdot)$ is cadlag and satisfies (1), are enough to see that

$$\sup_{t \geq 0} \frac{A(t)}{e^{\lambda_A t}} < \infty,$$

and

$$\sup_{n \in \mathbb{N}} \frac{e^{\lambda_A \tau_n'}}{A(\tau_n')} < \infty$$

almost surely. Now write

$$v_n A(\tau_n' + t) = n v_n \frac{A(\tau_n' + t)}{e^{\lambda_A(\tau_n' + t)}} \frac{e^{\lambda_A \tau_n'}}{A(\tau_n')} e^{\lambda_A t}.$$

It becomes apparent that

$$\lim_{n \to \infty} v_n A(\tau_n' + t) = \theta e^{\lambda_A t},$$

and for all $t \in \mathbb{R}$

$$\sup_{n \in \mathbb{N}} v_n A(\tau_n' + t) \leq L e^{\lambda_A t}$$

almost surely, for some positive random variable L. Then, using dominated convergence and the fact that $N(\cdot)$ is almost surely continuous at $\int_{-\infty}^{t} \theta e^{\lambda_A s}\, ds$, we are done.   □

Lemma 9.3 corresponds to Theorem 4.2 in the case $(\rho_n) = (\tau_n')$. Lemmas 9.4 and 9.5 extend the result to $(\rho_n) = (\sigma_n')$.

LEMMA 9.4.   *Conditioning on $W > 0$,*

$$\sup_{n \in \mathbb{N}} B^{(n)}(\sigma_n') < \infty$$

almost surely.

PROOF.   For each $n \in \mathbb{N}$, consider the process

$$\hat{B}^{(n)}(t) = \sum_{i=1}^{K^{(n)}(t)} \hat{Y}_i(t - T_i^{(n)}),$$

where

$$\hat{Y}_i(t) = \sup_{s \in [0,t]} Y_i(s).$$

With probability 1,

$$\lim_{n \to \infty} \hat{B}^{(n)}(\tau_n') \in [0, \infty)$$

exists, by Lemma 9.3 and the almost sure continuity of the $\hat{Y}_i$ at $-T_i^*$. Finally,

$$B^{(n)}(\sigma_n') \leq \hat{B}^{(n)}(\sigma_n') \leq \hat{B}^{(n)}(\tau_n') \leq \sup_{n \in \mathbb{N}} \hat{B}^{(n)}(\tau_n') < \infty,$$

using the monotonicity of $\hat{B}^{(n)}(\cdot)$.   □

LEMMA 9.5.   *Conditioning on $W > 0$,*

$$\tau_n' - \sigma_n' \to 0$$

almost surely.

PROOF.   Consider a positive sequence (a_n), such that:

1. $\lim_{n \to \infty} a_n = \infty$, and
2. $\lim_{n \to \infty} (n - a_n)/n = 1$.

For example $a_n = n^{1/2}$ will do. Since

$$e^{\lambda_A(\tau_n' - \tau_{n-a_n}')} = \frac{W e^{\lambda_A \tau_n'}}{A(\tau_n')} \frac{A(\tau_n')}{n} \frac{n}{n - a_n} \frac{n - a_n}{A(\tau_{n-a_n}')} \frac{A(\tau_{n-a_n}')}{W e^{\lambda_A \tau_{n-a_n}'}},$$

we have that as $n \to \infty$

$$\tau_n' - \tau_{n-a_n}' \to 0.$$

By Lemma 9.4,

$$B^{(n)}(\sigma_n') \leq a_n$$

for sufficiently large n. For such n,

$$A(\sigma_n') \geq n - a_n,$$

so

$$\sigma_n' \geq \tau_{n-a_n}',$$

and hence

$$0 \leq \tau_n' - \sigma_n' \leq \tau_n' - \tau_{n-a_n}'.$$ □

PROOF OF THEOREM 4.2.   Combine Lemmas 9.3 and 9.5 to see that conditioning on $W > 0$,

$$\lim_{n \to \infty} K^{(n)}(\rho_n + t) = K^*(t)$$

almost surely, for each $t \in \mathbb{R}$. Convergence in finite dimensional distributions follows. Then apply Lemmas 9.1 and 9.2 so that we may instead condition on the $\rho_n < \infty$.   □

PROOFS OF PROPOSITIONS 4.3, 4.5 AND 4.6.   By Lemmas 9.3 and 9.5, conditioning on $W > 0$,

$$\lim_{n \to \infty} \rho_n + t - T_i^{(n)} = t - T_i^*$$

almost surely. Use that the Y_i are almost surely continuous at $t - T_i^*$. Then apply Lemmas 9.1 and 9.2, to condition on $\rho_n < \infty$.   □

PROOF OF PROPOSITION 5.1.   It needs to be shown that for each $t \in \mathbb{R}$

$$N\left(\int_{-t_n}^{t} v_n A(t_n + s)\, ds\right) \to N\left(\int_{-\infty}^{t} W \eta e^{\lambda_A s}\, ds\right)$$

almost surely, as $n \to \infty$. Indeed, writing

$$v_n A(t_n + s) = v_n e^{\lambda_A t_n} \frac{A(t_n + s)}{e^{\lambda_A (t_n + s)}} e^{\lambda_A s},$$

one sees that $v_n A(t_n + s)$ converges to the appropriate limit and is dominated by a multiple of $e^{\lambda_A s}$.   □

Proposition 5.2 follows by continuity.

Proofs for Section 6.   We first make note of a classic result, which can be found in [32].

LEMMA 9.6.   *Assume that $\beta_A = 0$. For each n, $(\tau_n - \tau_k)_{k=1}^{n-1}$ has the same distribution as a collection of $n - 1$ i.i.d. Exponential(α_A) random variables, which are ordered by size.*

PROOF OF PROPOSITION 6.1.   For each $i \in \mathbb{N}$, let $(T_{i,j})_{j \in \mathbb{N}}$ be the occurrence times of a homogeneous Poisson process on $[0, \infty)$ with intensity v. These are the mutation times corresponding to one particular wild-type cell present from time 0. Noting that

$$A(t) = \sum_{i=1}^{\infty} 1_{[\tau_i, \infty)}(t),$$

it is apparent that the mutation times of all wild-type cells are distributed according to

$$(\tau_i + T_{i,j})_{i,j \in \mathbb{N}}.$$

The number of mutants at time t is

$$B(t) \stackrel{d}{=} \sum_{i,j \in \mathbb{N}} 1_{\{t - \tau_i - T_{i,j} \geq 0\}} Y_{i,j}(t - \tau_i - T_{i,j}) = \sum_{i \in \mathbb{N}} 1_{\{t - \tau_i \geq 0\}} D_i(t - \tau_i),$$

where

$$D_i(t) = \sum_{j \in \mathbb{N}} 1_{\{t-T_{i,j} \geq 0\}} Y_{i,j}(t - T_{i,j}) \stackrel{d}{=} \sum_{j=1}^{K_i(t)} Y_{i,j}(U_{i,j}t).$$

The $D_i(\cdot)$ are i.i.d. Now, using Lemma 9.6,

$$B(\tau_n) \stackrel{d}{=} \sum_{i=1}^{n-1} D_i(\tau_n - \tau_i) \stackrel{d}{=} \sum_{i=1}^{n-1} D_i(\xi_i),$$

and by substituting $D_i(\cdot)$ the result is obtained. $\quad\square$

PROOF OF LEMMA 6.3. We will show that the events $\{B(\sigma_n) \leq k\}$ and $\{B(\tau_{n-k}) \leq k\}$ are equal, using the monotonicity of $A(\cdot)$ and $B(\cdot)$ and the fact that $A(\sigma_n) + B(\sigma_n) = n$. First, assume that $B(\sigma_n) \leq k$. Then $A(\sigma_n) \geq n - k$, so $\sigma_n \geq \tau_{n-k}$ and, therefore, $B(\tau_{n-k}) \leq k$. Now assume that $B(\sigma_n) > k$. Then $A(\sigma_n) < n - k$, so $\sigma_n < \tau_{n-k}$, and hence $B(\tau_{n-k}) > k$. $\quad\square$

Proofs for Section 7.

PROOFS OF PROPOSITIONS 7.3 AND 7.4. For $i \in \{1, \ldots, S\}$ and $n \in \mathbb{N}$, $[C^{(n)}(\cdot) - B_i^{(n)}(\cdot)]$ is a birth–death branching process with birth and death rates $a(1 - \mu_n)$ and $b(1 - \mu_n)$. But we wish to make use of the proofs for the two-type model. Hence we will rescale time by a factor of $(1 - \mu_n)$.

On a fresh probability space put a birth–death process, $(\tilde{A}(t))_{t \geq 0}$, with birth and death rates a and b. Put an independent Poisson process, $(\tilde{N}(t))_{t \geq 0}$, with intensity 1. And for $i, n \in \mathbb{N}$ put the birth–death processes $(\tilde{Y}_i^{(n)}(t))_{t \geq 0}$, which we ask to satisfy:

1. The $\tilde{Y}_i^{(n)}(\cdot)$ have birth and death rates $a/(1 - \mu_n)$ and b, and initial condition $\tilde{Y}_i^{(n)}(0) = 1$.

2. For each n, the $\tilde{Y}_i^{(n)}(\cdot)$ are independent ranging over i.

3. The $\tilde{Y}_i^{(n)}(\cdot)$ are independent of $\tilde{A}(\cdot)$ and $\tilde{N}(\cdot)$.

4. For each i, $\lim_{n \to \infty} \tilde{Y}_i^{(n)}(\cdot) = \tilde{Y}_i(\cdot)$ exists almost surely, in the standard Skorokhod sense on the space of cadlag functions $\mathbb{D}[0, \infty)$.

Define $\tilde{K}^{(n)}(t) = \tilde{N}(\frac{a\mu_n}{1-\mu_n} \int_0^t \tilde{A}(s)\,ds)$ and $\tilde{T}_i^{(n)} = \inf\{t \geq 0 : \tilde{K}^{(n)}(t) = i\}$. Define

$$\tilde{B}^{(n)}(t) = \sum_{i=1}^{\tilde{K}^{(n)}(t)} \tilde{Y}_i^{(n)}(t - \tilde{T}_i^{(n)}),$$

and then $\tilde{\sigma}_n = \inf\{t \geq 0 : \tilde{A}(t) + \tilde{B}^{(n)}(t) \geq n\}$.

We have just defined a slight adaptation of the framework used for the small mutation limits of the two-type model. The proofs for the two-type model are readily adapted to this new setting. Here, the mutation rates are $a\mu_n/(1 - \mu_n)$. Suppose that the μ_n satisfy the condition of Proposition 7.3, then $\lim_{n\to\infty} na\mu_n/(1 - \mu_n) = \theta$. Follow the proof of Proposition 4.3 to see that

$$\lim_{n\to\infty} \mathbb{P}[\tilde{B}^{(n)}(\tilde{\sigma}_n) = k | \tilde{\sigma}_n < \infty] = \mathbb{P}[B^* = k].$$

Then use that

$$\left(B_i^{(n)}(t)\right)_{t\geq 0} \stackrel{d}{=} \left(\tilde{B}^{(n)}((1 - \mu_n)t)\right)_{t\geq 0}$$

and

$$\sigma_n' \stackrel{d}{=} \tilde{\sigma}_n/(1 - \mu_n),$$

to obtain

$$\lim_{n\to\infty} \mathbb{P}[B_i^{(n)}(\sigma_n') = k | \sigma_n' < \infty] = \mathbb{P}[B^* = k].$$

This gives Proposition 7.3.

Similarly, if the μ_n and t_n satisfy the conditions of Proposition 7.4, then $\lim_{n\to\infty} e^{\lambda t_n/(1-\mu_n)} a\mu_n/(1 - \mu_n) = \eta$. Follow the proof of Proposition 5.2 to see that

$$\lim_{n\to\infty} \mathbb{P}[\tilde{B}^{(n)}(t_n/(1 - \mu_n)) = k] = \mathbb{P}[B^\circ = k].$$

Then

$$\lim_{n\to\infty} \mathbb{P}[B_i^{(n)}(t_n) = k] = \mathbb{P}[B^\circ = k],$$

as required. □

APPENDIX: GENERALISED TWO-TYPE MODEL

Here, we present a generalisation of Kendall's model in which the results of Sections 4 and 5 are valid. The broader framework encompasses more general branching processes as well as semi-deterministic versions of the model (e.g., [11, 20, 27]).

Consider the model defined in Section 2. Relax the requirement that $A(\cdot)$ and the $Y_i(\cdot)$ need to be birth–death branching processes. Instead let $A(\cdot)$ and the $Y_i(\cdot)$ be $[0, \infty)$-valued cadlag processes. Demand further that there exists $\lambda_A > 0$ and a nonnegative random variable W with:

1. $\lim_{t\to\infty} e^{-\lambda_A t} A(t) = W$, and
2. $\{W = 0\} = \{\exists T > 0, \forall t \geq T, A(t) = 0\}$,

almost surely. We claim that Theorem 4.2 and Propositions 4.1, 4.3, 4.5, 4.6, 5.1 and 5.2 remain valid.

To see this, only the proof of Lemma 9.3 requires additional work. Observe that conditioning on $W > 0$:

- $\tau_n < \infty$ for each $n \in \mathbb{N}$, and
- $\lim_{n \to \infty} \tau_n = \infty$.

Then we need Lemma A.1.

LEMMA A.1. *Conditioning on $W > 0$,*

$$\lim_{n \to \infty} n^{-1} A(\tau_n) = 1$$

almost surely.

PROOF. *Fix $\omega \in \{W > 0\}$ and $\varepsilon > 0$. Then there exists $T > 0$ such that for all $t > T$,*

$$\left| e^{-\lambda_A t} A(t) - W \right| \leq \varepsilon.$$

There is some $N \in \mathbb{N}$ such that for any integer $n \geq N$, $\tau_n > T$. Now, for all such n,

$$\left| e^{-\lambda_A \tau_n} A(\tau_n^-) - W \right| = \lim_{t \uparrow \tau_n} \left| e^{-\lambda_A t} A(t) - W \right| \leq \varepsilon,$$

where $A(\tau_n^-) = \lim_{t \uparrow \tau_n} A(t)$. That is to say,

$$\lim_{n \to \infty} e^{-\lambda \tau_n} A(\tau_n^-) = W.$$

Finally,

$$1 \leq \frac{A(\tau_n)}{n} \leq \frac{A(\tau_n)}{A(\tau_n^-)} = \frac{A(\tau_n)}{e^{\lambda_A \tau_n}} \frac{e^{\lambda_A \tau_n}}{A(\tau_n^-)} \to 1. \qquad \square$$

REFERENCES

[1] ANGERER, W. P. (2001). An explicit representation of the Luria–Delbrück distribution. *J. Math. Biol.* **42** 145–174. MR1816347

[2] ANTAL, T. and KRAPIVSKY, P. L. (2010). Exact solution of a two-type branching process: Clone size distribution in cell division kinetics. *J. Stat. Mech. Theory Exp.* **7** P07028. MR2720343

[3] ATHREYA, K. B. and NEY, P. E. (2004). *Branching Processes*. Dover Publications, Inc., Mineola, NY. Reprint of the 1972 original [Springer, New York; MR0373040]. MR2047480

[4] BOZIC, I., GEROLD, J. M. and NOWAK, M. A. (2016). Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLoS Comput. Biol.* **12** e1004731.

[5] BOZIC, I. et al. (2013). Evolutionary dynamics of cancer in response to targeted combination therapy. *eLife* **2** e00747.

[6] DIAZ JR., L. A. et al. (2012). The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486** 537.

[7] DINGLI, D. et al. (2007). The emergence of tumor metastases. *Cancer Biol. Ther.* **6** 383–390.

[8] DURRETT, R. (2013). Population genetics of neutral mutations in exponentially growing cancer cell populations. *Ann. Appl. Probab.* **23** 230–250. MR3059234

[9] DURRETT, R. (2015). *Branching Process Models of Cancer. Mathematical Biosciences Institute Lecture Series. Stochastics in Biological Systems* **1**. Springer, Cham. MR3363681

[10] DURRETT, R., FOO, J., LEDER, K., MAYBERRY, J. and MICHOR, F. (2010). Evolutionary dynamics of tumor progression with random fitness values. *Theor. Popul. Biol.* **78** 54–66.

[11] DURRETT, R. and MOSELEY, S. (2010). Evolution of resistance and progression to disease during clonal expansion of cancer. *Theor. Popul. Biol.* **77** 42–48.

[12] FLAJOLET, P. and SEDGEWICK, R. (2009). *Analytic Combinatorics*. Cambridge Univ. Press, Cambridge. MR2483235

[13] FOO, J. and LEDER, K. (2013). Dynamics of cancer recurrence. *Ann. Appl. Probab.* **23** 1437–1468. MR3098438

[14] HAENO, H. and MICHOR, F. (2010). The evolution of tumor metastases during clonal expansion. *J. Theoret. Biol.* **263** 30–44. MR2980717

[15] HAMON, A. and YCART, B. (2012). Statistics for the Luria–Delbrück distribution. *Electron. J. Stat.* **6** 1251–1272. MR2988447

[16] IWASA, Y., NOWAK, M. A. and MICHOR, F. (2006). Evolution of resistance during clonal expansion. *Genetics* **172** 2557–2566.

[17] JANSON, S. (2004). Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stochastic Process. Appl.* **110** 177–245. MR2040966

[18] JANSON, S. (2006). Limit theorems for triangular urn schemes. *Probab. Theory Related Fields* **134** 417–452. MR2226887

[19] JONES, S. et al. (2008). Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl. Acad. Sci. USA* **105** 4283–4288.

[20] KELLER, P. and ANTAL, T. (2015). Mutant number distribution in an exponentially growing population. *J. Stat. Mech. Theory Exp.* **1** P01011. MR3314726

[21] KENDALL, D. G. (1960). Birth-and-death processes, and the theory of carcinogenesis. *Biometrika* **47** 13–21. MR0112761

[22] KESSLER, D. A. and LEVINE, H. (2013). Large population solution of the stochastic Luria–Delbrück evolution model. *Proc. Natl. Acad. Sci. USA* **110** 11682–11687. MR3104978

[23] KESSLER, D. A. and LEVINE, H. (2015). Scaling solution in the large population limit of the general asymmetric stochastic Luria–Delbrück evolution process. *J. Stat. Phys.* **158** 783–805. MR3311481

[24] KOMAROVA, N. (2006). Stochastic modeling of drug resistance in cancer. *J. Theoret. Biol.* **239** 351–366. MR2223739

[25] KOMAROVA, N. L., WU, L. and BALDI, P. (2007). The fixed-size Luria–Delbruck model with a nonzero death rate. *Math. Biosci.* **210** 253–290. MR2364819

[26] KUIPERS, J., JAHN, K., RAPHAEL, B. J. and BEERENWINKEL, N. (2017). Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res.* **27** 1885–1894.

[27] LEA, D. E. and COULSON, C. A. (1949). The distribution of the numbers of mutants in bacterial populations. *J. Genet.* **49** 264–285.

[28] LURIA, S. E. and DELBRÜCK, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **48** 419–511.

[29] MICHOR, F., NOWAK, M. A. and IWASA, Y. (2006). Stochastic dynamics of metastasis formation. *J. Theoret. Biol.* **240** 521–530. MR2248436

[30] NICHOLSON, M. D. and ANTAL, T. (2016). Universal asymptotic clone size distribution for general population growth. *Bull. Math. Biol.* **78** 2243–2276. MR3567698

[31] OHTSUKI, H. and INNAN, H. (2017). Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. *Theor. Popul. Biol.* **117** 43–50.

[32] RÉNYI, A. (1953). On the theory of order statistics. *Acta Math. Acad. Sci. Hung.* **4** 191–231. MR0061792

[33] ROSCHE, W. A. and FOSTER, P. L. (2000). Determining mutation rates in bacterial populations. *Methods* **20** 4–17.

[34] WILLIAMS, M. J., WERNER, B., BARNES, C. P., GRAHAM, T. A. and SOTTORIVA, A. (2016). Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48** 238–244.

[35] ZHENG, Q. (1999). Progress of a half century in the study of the Luria–Delbrück distribution. *Math. Biosci.* **162** 1–32. MR1726870

SCHOOL OF MATHEMATICS
UNIVERSITY OF EDINBURGH
EDINBURGH EH9 3FD
UNITED KINGDOM
E-MAIL: d.m.cheek@sms.ed.ac.uk
          tibor.antal@ed.ac.uk