

Comparison of Tug-of-War Models Assuming Moran versus Branching Process Population Dynamics

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

[About eLife's process](#)

Reviewed preprint version 1

May 22, 2024 (this version)

Posted to preprint server

February 21, 2024


Sent for peer review

January 15, 2024

Khanh N. Dinh , Monika K. Kurpas, Marek Kimmel

Irving Institute for Cancer Dynamics and Department of Statistics, Columbia University, New York, NY, USA • Department of Systems Biology and Engineering, Silesian University of Technology, Gliwice, Poland • Departments of Statistics and Bioengineering, Rice University, Houston, TX, USA

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

Abstract

Mutations arising during cancer evolution are typically categorized as either ‘drivers’ or ‘passengers’, depending on whether they increase the cell fitness. Recently, McFarland et al. introduced the Tug-of-War model for the joint effect of rare advantageous drivers and frequent but deleterious passengers. We examine this model under two common but distinct frameworks, the Moran model and the branching process. We show that frequently used statistics are similar between a version of the Moran model and the branching process conditioned on the final cell count, under different selection scenarios. We infer the selection coefficients for three breast cancer samples, resulting in good fits of the shape of their Site Frequency Spectra. All fitted values for the selective disadvantage of passenger mutations are nonzero, supporting the view that they exert deleterious selection during tumorigenesis that driver mutations must compensate.

eLife assessment

This study uses numerical simulations to characterize and compare variants of two widely used mathematical models and then applies those models to inferring evolutionary parameters from breast cancer data. The copious numerical results will be of some interest to mathematical biologists working with similar models. The finding that many breast cancer mutations are mildly deleterious is **valuable** but the evidence supporting this claim is **incomplete** because the mathematical modelling and statistical methods are insufficiently justified and **inadequately** validated.

<https://doi.org/10.7554/eLife.94597.1.sa2>

1 Introduction

As demonstrated in the seminal paper [12], mutations in different cancers vary substantially in counts and patterns. These differences reflect distinct defects in DNA repair mechanisms, cancer exposures, and cell types. The authors also reported evidence for ‘driver’ mutations in about 120

genes, which contribute to tumorigenesis. However, the majority of somatic mutations are likely ‘passengers’, which do not have an effect on tumor progression. In the language of population genetics, driver mutations are selectively advantageous to cancers, while the passengers are at best neutral.

The concept of a model involving the joint effect of rare advantageous and frequent neutral or slightly deleterious mutations can be applied to describe evolution of cancer genes. To the best of our knowledge, such model was first introduced by McFarland and co-authors, in a series of publications [18, 19, 20], and named the Tug-of-War model to reflect the competition between driver and passenger mutations.

The original Tug-of-War model [18] assumed that the cell death rate increases with the number of cells in the population increasing, which creates a mechanism for limiting the eventual tumor size. In other papers [16, 17], a Moran model was used for the population process, which provides a strict bound on the number of proliferating cells (see relevant discussion in [16]). Another assumption of [18] was that driver mutations become instantly fixed in the population, which may be acceptable under very strong selection (for mathematical details, see Bobrowski et al. [2]), but in general it is not satisfied.

The literature includes many examples of comparisons of how mutation, drift and selection interact in different population dynamics frameworks such as branching process versus Moran model [3, 4] or Wright-Fisher model with population of varying size [5, 6]. Two possible modes of selection in cell populations are “crowding out” in which a faster-growing clone makes the slower-growing one rare to the point of being negligible, or “competitive replacement”, in which individual cells inhibit each other’s replacement by a descendant. Supercritical branching process models lead to the former, while the Moran and Wright-Fisher models to the latter. A version interpolating between these two mechanisms is the well-cited Gerrish and Lenski model [11]. We will return to these models in the Discussion.

In the present paper, we compare the Tug-of-War in the multitype Moran model with constant population size and a critical multitype branching process. The latter is conditional on non-extinction or other restrictions. We explore similarities and differences between the two types of selection in cell populations. This contributes to the ongoing discussion of which models are most appropriate for proliferating cell populations under drift, mutation, and selection.

We begin with mathematical definitions of the two versions of Tug-of-War process. Then we present simulation results, which demonstrate the differences between the long-term behavior of the two versions under different selection scenarios. Finally, we infer the selection coefficients for some breast cancer samples using the Moran framework, and cross-examine the fitted parameters with the branching process.

2 Models and Data

2.1 Moran process Tug-of-War models

Model descriptions in this section, are essentially summaries of the descriptions in [16] and to avoid redundancy, we summarize only the most essential features. For more detailed descriptions, see [16].

2.1.1 Model A

In this model (**Figure 1A** [↗](#)), we contextualize the Tug-of-War within the Moran model framework with multiple allelic types. We examine a population with a constant size of N cells. Each cell i is described by a pair of integers $\gamma_i = (\alpha_i, \beta_i)$, where α_i and β_i represent the quantities of drivers and passengers in its genotype, respectively. The cell's fitness is then defined as

$$f_i = f_i(\alpha_i, \beta_i) = (1 + s)^{\alpha_i} (1 - d)^{\beta_i}, \quad i = 1, \dots, N,$$

where $s > 0$ is the selective advantage associated with a driver mutation, while $d \in (0, 1)$ represents the selective disadvantage of a passenger mutation (selection coefficients, of driver and passenger mutations; see the Natural Selection chapter of the book by Durrett [\[9\]](#) [↗](#)). Under the time-continuous Markov Chain assumption, the time until the next death - replacement event is exponentially distributed with parameter $\Sigma_f = \sum_{i \in \{1, \dots, N\}} f_i$. The dying cell i is selected from a distribution biased by fitness, i.e., with probability mass function (pmf) $\{f_i / \Sigma_f, i = 1, \dots, N\}$. The cell j that replaces the dying cell is selected from the same distribution.

The time until the next mutation event is exponentially distributed with parameter $N\mu$, where μ is the mutation rate per cell. The cell selected to mutate is chosen uniformly among the N cells. Its state then changes from (α, β) to $(\alpha + 1, \beta)$ with probability $p \in (0, 1)$, or to $(\alpha, \beta + 1)$ with probability $1 - p$. In combination, the time to the next event is random and exponentially distributed with parameter

$$\Sigma_f + N\mu, \tag{1}$$

called the total rate of death - replacement and mutation events.

2.1.2 Model B

Model B (**Figure 1B** [↗](#)) is defined similarly to Model A, with the time until the next death - replacement event being exponentially distributed with parameter Σ_f . However, instead of being biased by fitness, the dying cell is chosen among all N cells uniformly. The time until the next event is likewise exponentially distributed with the same parameter in Equ. (1).

2.1.3 Model A versus Model B

As we noted in [\[16\]](#) [↗](#), the crucial difference between Model A and Model B lies in the expected fitness increment in the population after the death - replacement event. This is equal to the difference of $f_j - f_i$, where f_i and f_j are the fitnesses of the dead cell and the replacing cell, in the absence of mutations. The mutation-selection balance condition was derived in [\[2\]](#) [↗](#) in the form of

$$ps = (1 - p)d \tag{2}$$

When $ps > (1 - p)d$, drivers “dominate” over passengers, and the reverse occurs if $ps < (1 - p)d$. The expected fitness is intact after the death - replacement process in Model A, hence the expected fitness trend follows the mutation-selection condition. In Model B, the outcome is more complex (see Results and Discussion).

2.2 Branching process model

For the branching process model (**Figure 1C** [↗](#)), we consider a population consisting of $N(t)$ cells at time t . Similar to the Moran models, the fitness of a cell i of type (α_i, β_i) with α_i drivers and β_i passengers is defined by $f_i = (1 + s)^{\alpha_i} (1 - d)^{\beta_i}$.

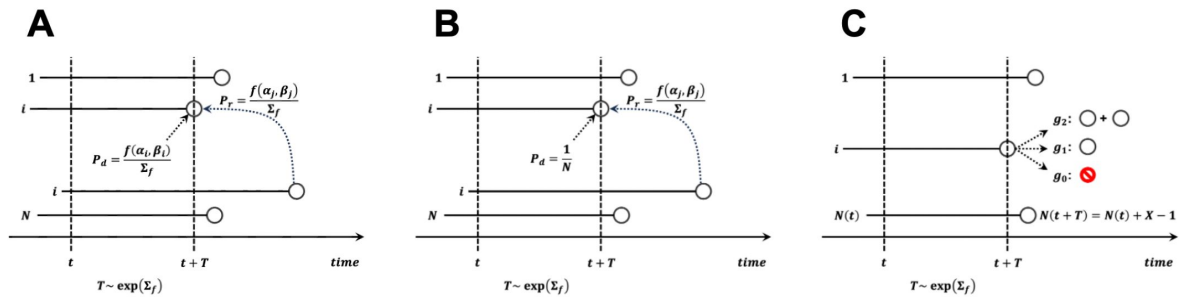


Figure 1

Graphical depiction of cell death and division events in **(A)** Moran A model, **(B)** Moran B model, and **(C)** branching process.
Notation for all models: t , current time; T , time to next event; $f(\alpha, \beta)$, fitness of cell with α drivers and β passengers; $\Sigma_f = \Sigma_k f(\alpha_k, \beta_k)$. **Notation for Moran models:** N , constant cell count in the process; i , cell dying and to be replaced; j , cell replacing cell i ; **Notation for branching process:** $N(t)$, cell count at time t ; i , cell chosen for division; $\{g_0, g_1, g_2\}$, progeny count distribution; X , progeny count of cell i ; $N(t+T) = N(t) + X - 1$.

Two possible event types can occur in the branching process: cell division and mutation. The time to the next cell division is exponentially distributed with rate $\Sigma_f = \Sigma_{i \in \{1, \dots, N(t)\}} f_i$, fitness sum of all $N(t)$ cells. The dividing cell is chosen from the $N(t)$ cells with probability weighted by fitness $\{f_i / \Sigma_f\}$, and its progeny count follows a given pmf $\{g_0, g_1, \dots\}$. If the progeny count is 0, the population loses the chosen cell. On the other hand, it survives if the progeny count is 1, and multiplies if the progeny count is more than 1. Additionally, the time to the next mutation event is exponentially distributed with rate $\mu N(t)$. A uniformly chosen cell acquires a driver and changes its type to $(\alpha + 1, \beta)$ with probability p , or it acquires another passenger to become type $(\alpha, \beta + 1)$ with probability $1 - p$.

2.2.1 Criticality and conditioning

One important difference between the Moran and branching process settings is that the total cell count at any time remains constant at N in the Moran models. The branching process model starts from the same cell count, i.e. $N(0) = N$, but it can change at random throughout time. For a direct comparison to Moran models, we may assume that $\mathbb{E}(N(t)) = N$. This is satisfied if we require that the mean progeny count of any cell is 1, i.e. that the branching process is critical [15, 15].

Even with criticality, at any time $N(t)$ can fixate at 0 (in which case the process enters extinction) or increase to a larger count. The probability of extinction increases as either t increases ([15], Section 3.3) or the cell fitness increases as a result of time scale change. We analyze the effects of two types of conditioning on the results of the branching process. The first type conditions the branching process on non-extinction, meaning $N(t_f) > 0$ at the final time t_f . The second type imposes a more stringent condition on the branching process, requiring that $N(t_f) \in [N - c, N + c]$ for some small constant c .

2.2.2 Distribution of progeny cell counts

We also study the impact of the distribution of progeny cell count $\{g_k, k = 0, 1, \dots\}$ on the outcomes of the branching process. In this study, we impose $g_k = 0$ for $k > 2$. Therefore, a chosen cell dies if $k = 0$, remains unchanged if $k = 1$, or divides into two progeny cells if $k = 2$. The criticality requirement, discussed in the previous section, is satisfied if $g_0 = g_2$. Note that the branching process setting discussed in this paper is equivalent to a birth-death process where each cell i with fitness f_i dies with rate $g_0 f_i$ or divides with rate $g_2 f_i$.

A common progeny count distribution is such that $g_0 = g_2 = 0.25$ and $g_1 = 0.5$, which is equivalent to a binomial distribution with rates $n = 2, p = 0.5$. For a direct comparison in simulations to the Moran models, the fitness in the branching process is scaled up by 4. Hence, the wait time until the next division event of a cell with fitness f_i is exponentially distributed with rate $4f_i$. This event has equal probability to be a cell death or a cell division, both at 0.25. Therefore, the model is equivalent to a birth-death process, where the birth and death rates are $0.25 \times 4f_i = f_i$. In comparison, in the Moran models, the death - replacement events also occur at rate f_i , resulting in two cells being chosen to divide and die, respectively. Because the event rates are now similar, it is easier to directly compare the model behaviors under different selection scenarios.

We will also investigate the effect of changing the progeny cell count distribution $\{g_0, g_1, g_2\}$. First, we retain the criticality by assuming $g_0 = g_2$, and analyze the branching process with different values for g_1 . We set $g_0 = g_2 = 0.5$ and $g_1 = 0$, which doubles the probabilities for cell division and death events. This is similar to increasing the birth and death rates in a birth-death process, hence we name this model “fast BP”. We also consider $g_0 = g_2 = 0.05$ and $g_1 = 0.9$ (“slow BP”), which decreases the cell division and death probabilities. Second, we investigate the supercritical branching process by setting $g_2 > g_0$. For each of these parameter sets, we will analyze the differences in sample statistics, compared to the binomial branching process and the Moran models.

2.3 Site Frequency Spectrum

As in other preceding paragraphs, in this section, we include a summary of the descriptions in [16]. For more details, see [16].

One of the common summary statistics of the sequence data is the so-called Site Frequency Spectrum. In a sequencing experiment with n cells, we can estimate for each novel somatic mutation call the number of cells carrying that mutation. The number $S_n(k)$ of mutations present in k cells is put into a vector $(S_n(1), S_n(2), \dots, S_n(n-1))$ called the Site Frequency Spectrum, abbreviated to SFS. Frequently number of cells that were sampled is not known, as for example in the bulk sequencing data. However, we can estimate the relative proportion of the mutant at each site, and so arrive at a frequency spectrum based on proportions with notation $S(x) = S(k/n)$, where x is treated as a continuous variable, such that $x \in (0, 1)$ (or $x \in (0, 1/2)$ if we consider diploid genome). The SFSs $S(k)$ and $S(x)$ are idealized versions of the empirical variant allele frequency (VAF) graph. It is convenient for reasons explained in [16] (Section 2.4) to employ the cumulative tail of the SFS $S(x)$

$$T(x) = \int_x^1 S(\xi) d\xi, \quad x \in [0, 1] \quad (3)$$

2.4 DNA Sequencing of Cell Samples from Breast Cancer Specimens

2.4.1 DNA sample collection and processing

Most of the details of DNA sample collection and processing in this section, overlap with those in [16] and we refer to the description there. We only mention here several basic details. Tissue samples from primary breast tumor were collected at the Department of Applied Radiology of the Maria Skłodowska-Curie National Research Institute of Oncology, Krakow Branch in Poland. The set of tumor and normal control samples called specimen G2 is HER2+ breast cancer, while sets described as G32 and G41 are triple-negative breast cancer type and luminal A type, respectively.

2.4.2 Removal of FFPE artifacts

Fixation of tissues in formalin leads to deamination of cytosine to uracil, which can be recognized by sequencing as C>T or G>A type modifications [8].

A significant portion of the variants detected in our WES data are a possible artifact of sample fixation in formalin. This is indicated primarily by the statistics of the number of variants of a specific type, where C:G>T:A definitely dominates (Table 1).

The reason for such a large number of this type of changes may be the duplication of variants related to deamination in PCR amplification (necessary in the case of WES, especially in the case of samples with a low amount of DNA).

Omitting all C:G>T:A variants would result in the loss of approximately 1/6 of the true variants. However, information about the frequency of reads with a specific orientation can be used to identify variants associated with the method of sample fixation. For this purpose, the SOBDetector [7] program was used. The software is based on the fact that formalin fixation most likely affects only one of the DNA strands (the C:G pair becomes the T:G pair) and therefore the paired-end next-generation sequencing approach can help this additional filtering step. By counting not only the number of reads supporting alternative alleles, but also the relative orientation of the reads (Forward-Reverse:FR or Reverse-Forward:RF), these FFPE artifacts will likely have a strand orientation bias toward one of the directions, while true mutations should have approximately the same number of FR and RF reads.

Patient ID	C:G>A:T	C:G>G:C	C:G>T:A	T:A>A:T	T:A>C:G	T:A>G:C	sum
G2	190	78	8379	87	430	84	9248
G32	86	47	6394	56	195	36	6814
G41	53	42	2088	44	176	38	2441

Table 1

Statistics of the number of variants of a specific type.

This work uses data in which the expected FFPE artifacts have been filtered out by SOBDDetector.

3 Results

3.1 Behavior of Moran and branching process models in extreme cases

We investigate the similarities and differences between Moran and branching process models. One thousand simulations are performed for each model, and each simulation starts with $N = 100$ cells with no mutations at $t_0 = 0$ under different values for the selection coefficients s and d , mutation rate μ and probability p of driver mutations (or equivalently probability $1 - p$ of passenger mutations). We then examine the statistics at final time $t_f = 100$ as well as during the entire time line $[t_0, t_f]$.

Five versions of branching processes are studied. This includes the branching process with $g_0 = g_2 = 0.25$ and $g_1 = 0.5$ conditioned on non-extinction (yellow), and the same branching process conditioned on the final population $N(t_f)$ restricted in $[90, 110]$ (orange). These models are termed “binomial BP” in the comparisons, since the progeny cell count distribution is binomial in this case. We also include the branching process with $g_0 = g_2 = 0.5$, $g_1 = 0$ (purple, termed “fast BP”) and $g_0 = g_2 = 0.05$, $g_1 = 0.9$ (cyan, termed “slow BP”), both conditioned on $N(t_f) \in [90, 110]$.

The fifth branching process model is conditioned on non-extinction with g_0 , g_1 and g_2 computed such that the population size is expected to double between $[t_0, t_f]$ under neutral evolution. It can be shown that g_0 and g_2 are required to satisfy

$$\begin{cases} g_2 - g_0 = \frac{\ln(2)}{t_f - t_0} \\ g_2 + g_0 = 1 - g_1 \end{cases}$$

We choose $g_1 = 0.5$, resulting in $g_0 = 0.2465$ and $g_2 = 0.2535$. This model is referred to as “supercritical BP” (gray) in the numerical comparisons. Finally, Moran A and Moran B are represented in dark blue and green, respectively.

Since we scale up the fitness by 4 in the branching process models to make them similar to the Moran models, in the results we scale the fitnesses down by 4, for more convenient comparisons. The division count, defined to be the total number of cell division events observed in a simulation, increases linearly with the total cell count and cell fitness, i.e. the rate at which cells divide. Since the expected cell count in a critical BP is identical to the Moran models, the division count in BP is 4 times higher than in the Moran models due to the fitness being scaled up. Therefore, we also downscale the BP division count by 4 to directly compare between different models.

3.1.1 Neutral evolution case

Figure 3 [↗](#) presents the simulated results for $s = d = 0$, which implies that all mutations are neutral, $\mu = 0.1$ and $p = 1/11$.

Since all cells have the same fitness, the formulations for Moran A and Moran B are identical. This is reflected in identical distributions among all statistics among the two variations (see Appendix C, **Table 2** [↗](#), dark blue for Moran A and green for Moran B). Moreover, the binomial BP conditioned on $N(t_f) \in [90, 110]$ also has the same distributions of allele counts and singleton counts, albeit with slightly higher variances (at final time in **Figure 3B-C** [↗](#) and throughout history in **Figure 3K-L** [↗](#)). The higher variances originate from wider distributions of event counts in the BP compared to Moran, these latter stemming from the fact that total cell count in BP varies

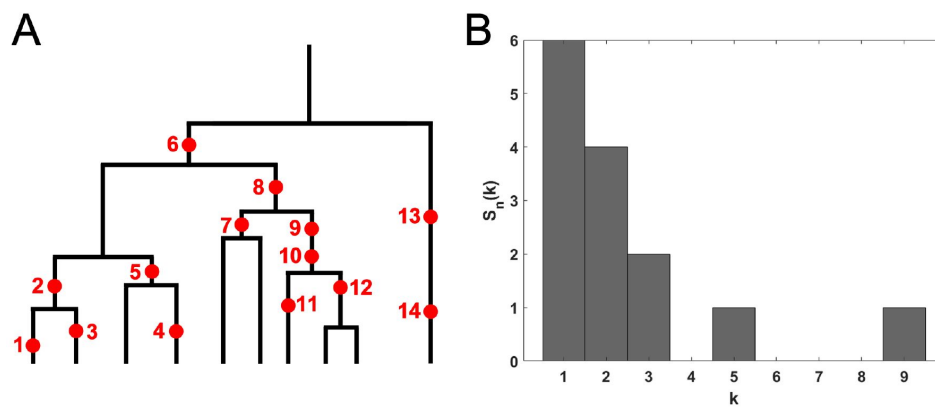


Figure 2

The Site Frequency Spectrum (SFS). **(A)** Genealogy of a sample of $n = 10$ cells includes 14 mutational events, denoted by red dots. Time is running down the page. Mutations 1, 3, 4, 11, 13, and 14 (total of 6 mutations) are present in a single cell, mutations 2, 5, 7 and 12 (total of 4 mutations) are present in two cells, mutations 9 and 10 (2 mutations) are present in three cells, mutation 8 (1 mutation) is present in five cells and mutation 6 (1 mutation) is present in 9 cells. **(B)** The resulting site frequency spectrum, $S_{10}(1) = 6$, $S_{10}(2) = 4$, $S_{10}(3) = 2$, $S_{10}(5) = 1$, and $S_{10}(9) = 1$, other $S_n(k)$ equal to 0.

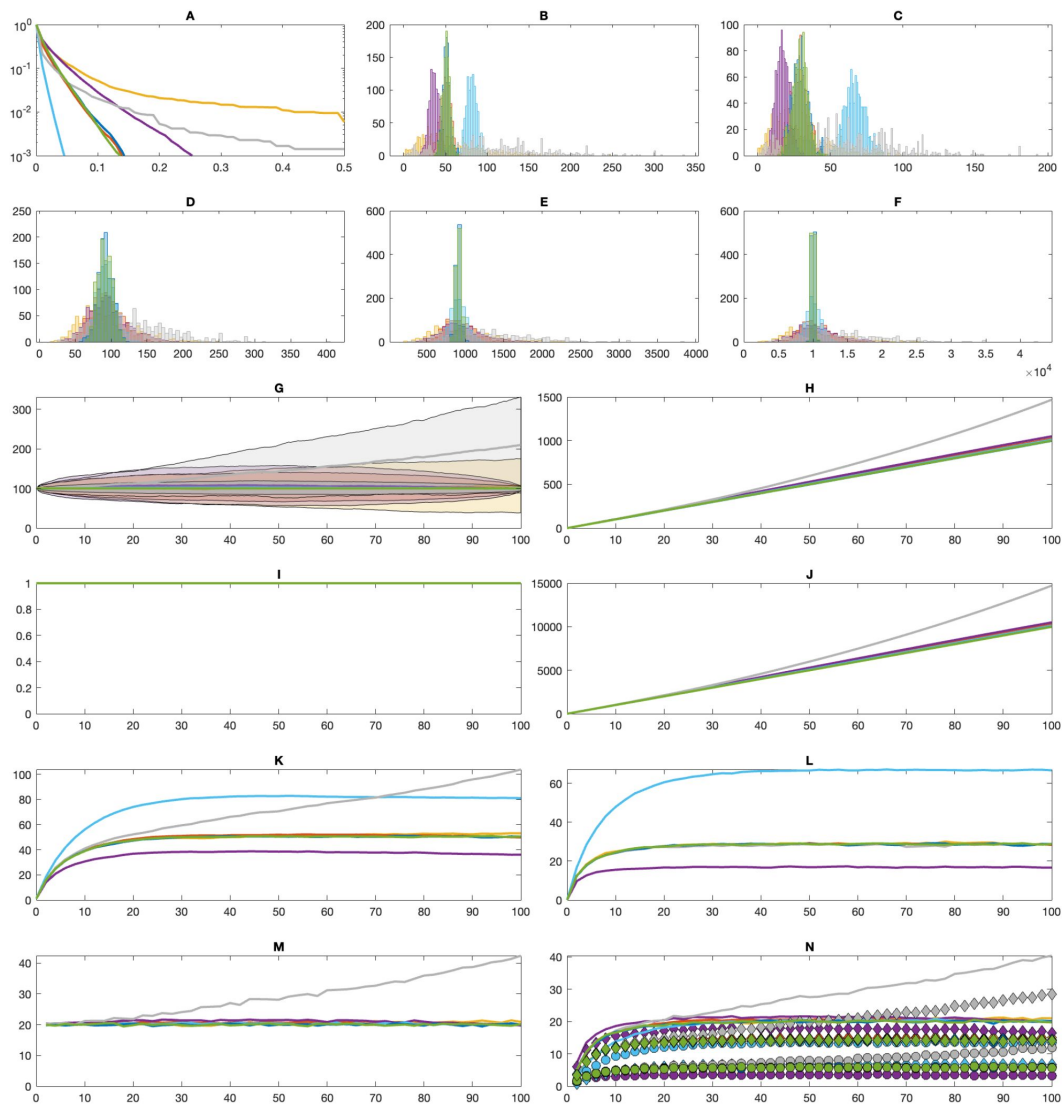


Figure 3

Comparisons between Moran and branching process (BP) models in the “neutral” setting. **(A)** Average cumulative tail of the mutational Site Frequency Spectra. **(B)** Distributions of allele counts at t_f . **(C)** Distributions of singleton counts at t_f . **(D-F)** Distributions of counts of driver mutations **(D)**, passenger mutations **(E)** and divisions **(F)** within $[t_0, t_f]$. **(G-N)** Trajectories of the averages over time of population sizes (+/-std) **(G)**, cumulative mutation counts **(H)**, fitness **(I)**, cumulative division/replacement counts **(J)**, allele counts **(K)**, percentage of singletons among all alleles **(L)**, allele birth counts **(M)** and allele death counts **(N)**. Allele death counts (lines) are categorized into mutation events (circles) and division/replacement events (diamonds). Dark blue = Moran A, green = Moran B, yellow = “binomial BP” with $g_0 = g_2 = 0.25$ (non-extinction), orange = “binomial BP” with $g_0 = g_2 = 0.25$ ($N(t_f) \in [90, 110]$), purple = “fast BP” with $g_0 = g_2 = 0.5$ ($N(t_f) \in [90, 110]$), cyan = “slow BP” with $g_0 = g_2 = 0.05$ ($N(t_f) \in [90, 110]$), gray = “supercritical BP” with $g_0 = 0.2465$, $g_1 = 0.5$ and $g_2 = 0.2535$ (non-extinction).

throughout time, as opposed to Moran where the total cell count remains constant (**Figure 3D-G**). Note that even though the statistics have higher variances, their averages are similar to the Moran models both throughout time and at the final time (Appendix C, **Table 2**, orange).

The impact of relaxing the conditioning of BP can be observed by comparing binomial BP conditioned on $N(t_p) \in [90, 110]$, binomial BP conditioned on non-extinction and supercritical binomial BP with $g_0 = 0.2465$, $g_1 = 0.5$ and $g_2 = 0.2535$, conditioned on non-extinction. As the population size can vary widely throughout time if conditioned only on non-extinction (**Figure 3G**), every statistics in the comparison has much higher variances (Appendix C, **Table 2**, yellow for critical BP and gray for supercritical BP). However, for critical BP, the averages remain faithful to both Moran models and the more stringently conditioned BP. Only in the case of supercritical BP do all statistics differ, even cumulative mutation count (**Figure 3H**) and cumulative division/replacement count (**Figure 3J**), which is strictly associated with rapidly growing population size.

We then evaluate the impact of changing the progeny cell count distributions, while retaining criticality. The fast BP has increased g_0 and g_2 and therefore is equal to a birth-death process with higher rates. This leads to higher variances in the population size throughout time compared to the binomial BP, even if similarly conditioned (**Figure 3G**, Appendix C: **Table 2**, purple). Importantly, the fast BP also results in both less alleles and lower percentage of singletons within all alleles (**Figure 3B-C, K-L**). Conversely, the slow BP is equivalent to a birth-death process with lower rates, whose population size therefore varies less than the binomial BP (**Figure 3G**). Both its allele count and percentage of singleton count is much higher than in the binomial BP (**Figure 3B-C, K-L**, Appendix C: **Table 2**, cyan).

Figure 3N details the rates at which alleles are lost from the population, divided into two categories: (a) cell deaths (division with no progeny cells in BP, or replacement in Moran), and (b) cell mutation (where the only remaining cell that carries the allele mutates into a different allele). The allele death counts, either combined or categorized to (a) or (b), are similar in all cases except supercritical BP and fast or slow BP. The rate of allele death is slightly lower for slow BP and slightly higher in the case of fast BP. They also have different categorized allele death counts. Compared to binomial BP, alleles are removed more frequently due to mutations in slow BP. Conversely, they are more likely to be removed due to failed divisions in fast BP. The supercritical BP is the only case in which allele death count does not reach plateau and is still increasing during the simulated time period, as the population grows exponentially.

3.1.2 Balanced evolution

Figure 4 showcases the simulated results for $s = 0.1$, $d = 0.01$, $\mu = 0.1$ and $p = 1/11$. As $ps = (1 - p)d$, condition (2) is satisfied, therefore the average fitness remains constant for Moran A model (**Figure 4I**). Moreover, the fitness in binomial BP also remains unchanged on average over time. As a result, the distributions for all statistics are similar between the balanced evolution setting and the neutral evolution setting (Appendix B, **Table 3**), discussed in the last section, for Moran A and binomial BP. The outcomes following modulation of the progeny cell count distribution or relaxing the conditioning for BP, likewise remain unchanged compared to the neutral evolution setting.

However, the fitness in Moran B increases over time instead of remaining constant (**Figure 4I**), which leads to slightly higher division count (**Figure 4F**) even though the mutation count, depending only on the population size which remains constant, is unchanged (**Figure 4D-E**). This results in slightly lower allele count and singleton count (**Figure 4B-C**), indicative of selective pressure.

Model	B	C	D	E	F
Dark blue	50.3±4.6	28.7±5.36	91.1±9.9	907.6±31.3	10005.5±207.0
Green	50.4±4.5	28.7±5.0	90.8±9.8	909.9±31.4	9992.3±206.1
Yellow	53.0±33.6	29.8±19.5	93.3±36.0	930±355.9	10235.7±3888.0
Orange	49.7±6.1	28.2±5.8	93.5±21.5	929.8±194.1	10261.6±2121.3
Purple	36.0±5.9	16.6±4.5	95.3±27.9	954.8±261.3	10498.6±2862.7
Cyan	81.0±6.2	66.5±7.5	92.1±13.3	919.9±95.0	10117.1±1003.4
Gray	103.8±58.7	58.9±33.8	133.1±53.4	1335.8±524.3	14723.8±5718.5

Table 2

Statistics (mean ± standard deviation) for **Figure 3** (neutral evolution). **Model color code:** dark blue = Moran A, green = Moran B, yellow = “binomial BP” with $g_0 = g_2 = 0.25$ (nonextinction), orange = “binomial BP” with $g_0 = g_2 = 0.25$ ($N(t_f) \in [90, 110]$), purple = “fast BP” with $g_0 = g_2 = 0.5$ ($N(t_f) \in [90, 110]$), cyan = “slow BP” with $g_0 = g_2 = 0.05$ ($N(t_f) \in [90, 110]$), gray = “supercritical BP” with $g_0 = 0.2465$, $g_1 = 0.5$ and $g_2 = 0.2535$ (non-extinction). **Statistics:** B = allele counts at t_f , C = singleton counts at t_f , D = driver mutation count within $[t_0, t_f]$, E = passenger mutation count within $[t_0, t_f]$, F = division count within $[t_0, t_f]$.

Model	B	C	D	E	F
Dark blue	50.7±4.5	29.0±5.1	91.3±9.5	910.8±27.1	9988.6±290.6
Green	47.7±5.3	26.6±5.3	89.7±9.8	910.6±29.8	10525.5±510.6
Yellow	50.3±34.8	28.5±20.0	92.6±38.6	922.8±375.5	10166.8±4151.7
Orange	50.3±6.2	28.7±5.8	91.9±21.2	925.2±192.8	10164.0±2118.3
Purple	36.5±5.8	17.1±4.6	95.3±28.4	956.7±265.8	10538.8±2947.8
Cyan	81.2±6.0	67.1±7.5	91.1±13.3	914.3±94.9	10055.1±981.3
Gray	97.7±54.4	55.3±31.1	130.4±50.6	1310.9±493.9	14413.9±5434.7

Table 3

Statistics for Figure 4 (balanced evolution)

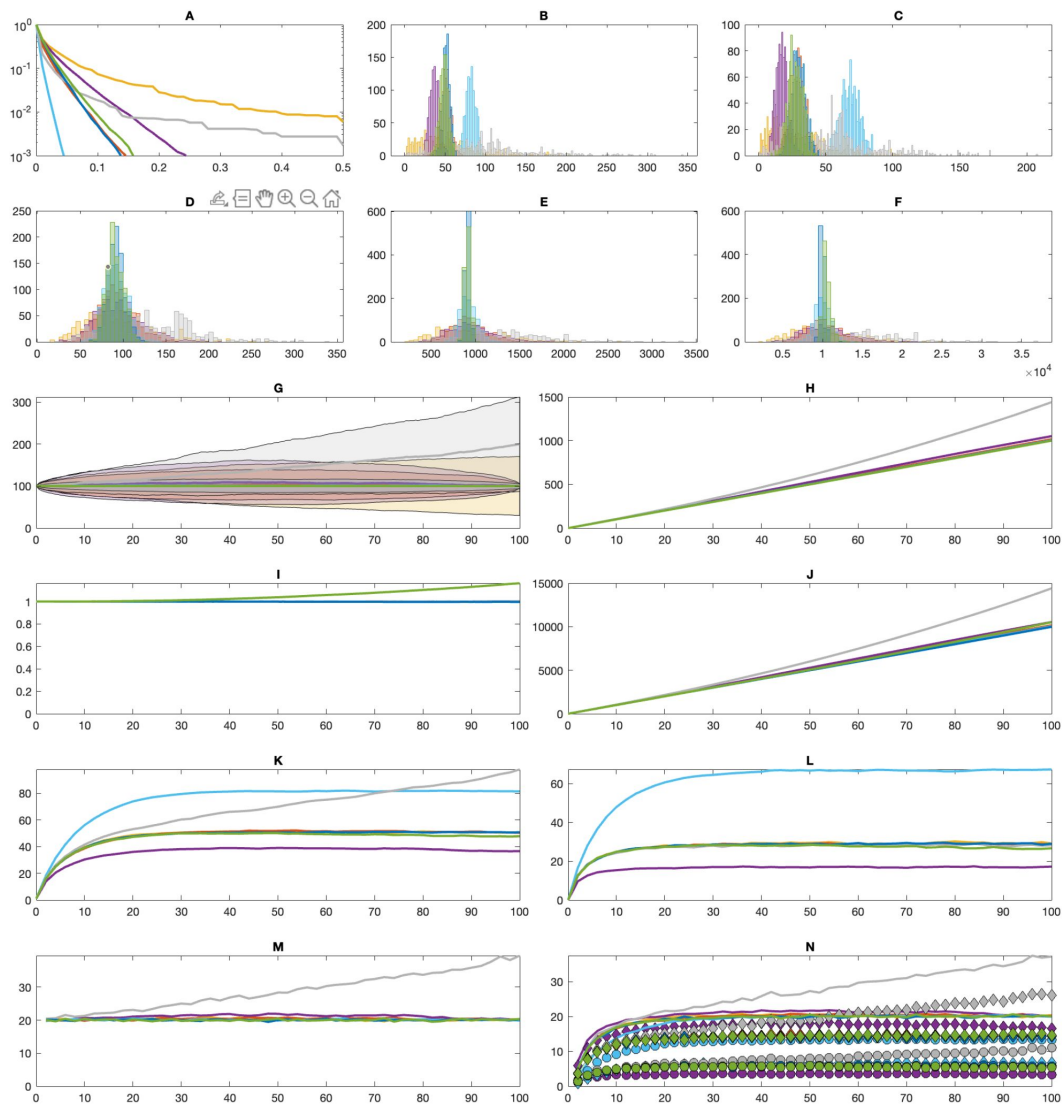


Figure 4

Comparisons between Moran and branching process (BP) models in the “balanced” setting. **(A)** Average cumulative tail of the mutational Site Frequency Spectra. **(B)** Distributions of allele counts at t_f . **(C)** Distributions of singleton counts at t_f . **(D-F)** Distributions of counts of driver mutations **(D)**, passenger mutations **(E)** and divisions **(F)** within $[t_0, t_f]$. **(G-N)** Trajectories of the averages over time of population sizes (+/-std) **(G)**, cumulative mutation counts **(H)**, fitness **(I)**, cumulative division/replacement counts **(J)**, allele counts **(K)**, percentage of singletons among all alleles **(L)**, allele birth counts **(M)** and allele death counts **(N)**. Allele death counts (lines) are categorized into mutation events (circles) and division/replacement events (diamonds). Dark blue = Moran A, green = Moran B, yellow = “binomial BP” with $g_0 = g_2 = 0.25$ (non-extinction), orange = “binomial BP” with $g_0 = g_2 = 0.25$ ($N(t_f) \in [90, 110]$), purple = “fast BP” with $g_0 = g_2 = 0.5$ ($N(t_f) \in [90, 110]$), cyan = “slow BP” with $g_0 = g_2 = 0.05$ ($N(t_f) \in [90, 110]$), gray = “supercritical BP” with $g_0 = 0.2465$, $g_1 = 0.5$ and $g_2 = 0.2535$ (non-extinction).

3.1.3 Driver domination case

To understand the consequences when the driver mutations are strongly selectively advantageous, we set $s = 0.25$, $d = 0$, $\mu = 0.1$ and $p = 1/10$ and compare the statistics in [Figure 5](#) and in Appendix B, [Table 4](#).

Because condition (2) is no longer satisfied, the fitness coefficients and mutation rates are not at equilibrium and fitness in Moran A increases over time ([Figure 5I](#)), leading to an increase in its replacement count ([Figure 5J](#)) as compared to the neutral or balanced evolution setting. As a result, the counts of both alleles and singletons are slightly lower in the selective evolution as compared to previous settings ([Figure 5B-C](#)). The same is also true for all remaining models. Remarkably, binomial BP still behaves identically to Moran A, differing only in population size over time ([Figure 5G](#)). As before, relaxing the conditioning on binomial BP leads only to higher variances of the statistics without changing their averages. Like in previous examples, supercritical BP has much higher averages and variances in all statistics than other models, both at the end of simulation as well as throughout time. Similarly as in the previous case, changing the progeny cell count distribution in BP while retaining criticality results in allele count and singleton count converging to different values ([Figure 5B-C, K-L](#)).

As in the balanced evolution setting, the only model differing in fitness from the remaining critical processes is Moran B ([Figure 5I](#)), this time resulting in twice as many replacements compared to other models ([Figure 5F, J](#)). In later moments of the simulation, the number of replacements is even higher than in the critical BP. Consequently, in Moran B, the counts of alleles and singletons decrease at a fast rate after reaching maximum values ([Figure 5B-C, K-L](#)).

3.1.4 Passenger domination case

Finally, we investigate the setting where passenger mutations are strongly deleterious, with parameters $s = 0$, $d = 0.5$, $\mu_d = 0.1$ and $p = 1/10$ corresponding to [Figure 6](#).

In Moran A, as cells accumulate increasingly more mutations, their fitness decrease to 0 because of the passenger mutations' deleterious coefficient ([Figure 6I](#)), therefore they stop dividing ([Figure 6J](#)). However, the mutation process depends only on the population size ([Figure 6G](#)) and therefore occurs at a constant rate throughout time ([Figure 6H](#)). The consequence is that every cell sooner or later would acquire a unique mutation, therefore the cell population almost consists only of singletons ([Figure 6B-C](#)).

As is the case for other settings, binomial BP matches the average statistics from Moran A throughout history and at the final time (Appendix B, [Table 5](#)). The same is true for relaxing the conditioning on binomial BP, which only increases the variances. However, unlike the selective evolution setting, altering the progeny cell count distribution does not change the steady state values for the allele and singleton counts, as both converge to the same distributions as binomial BP and Moran A ([Figure 6B-C, K-L](#)). Nonetheless, compared to these models, the fast BP converges faster and the slow BP takes more time to converge.

The deleterious evolution setting is the only scenario in which supercritical BP behaves similarly to most of the other models. While fitness tends to zero, cells stop dividing and the impact of the supercriticality is no longer significant.

Finally, as usual, Moran B has a much higher steady state fitness compared to other models ([Figure 6I](#)). Therefore, although the replacement count is lower than in the balanced or selective evolution settings, cells do not stop dividing as is the case with Moran A, because the fitness does not converge to 0 ([Figure 6E, I](#)). This results in much lower allele count and

Model	B	C	D	E	F
Dark blue	45.4±5.2	24.2±5.2	100.1±9.8	899.7±30.6	11455.7±772.4
Green	23.6±7.8	9.4±4.8	100.4±9.7	899.2±28.3	20767.0±7693.0
Yellow	44.6±29.5	24.0±16.2	98.2±37.9	888.3±331.5	11153.3±4416.3
Orange	45.3±6.3	24.3±5.5	103.2±25.2	928.1±207.0	11683.1±2716.7
Purple	32.3±5.9	14.1±4.5	105.0±33.3	942.4±287.8	11818.8±3751.9
Cyan	77.8±6.0	61.7±7.2	100.5±14.2	903.9±94.0	11441.5±1242.2
Gray	97.3±56.8	51.6±29.8	150.2±58.5	1341.1±506.0	17454.9±7228.5

Table 4

Statistics for Figure 5 [↗](#) (driver domination)

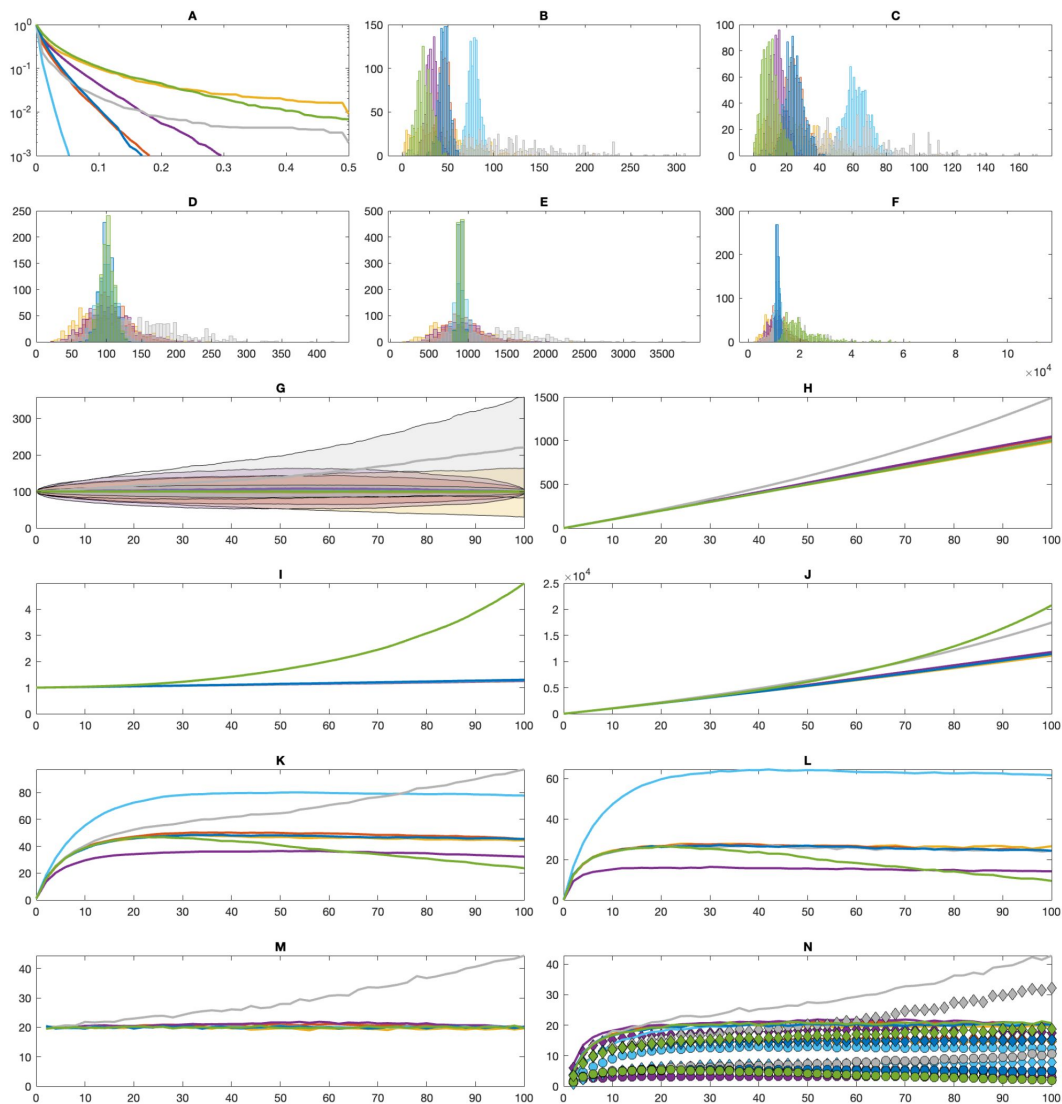


Figure 5

Comparisons between Moran and branching process (BP) models in the “selective” setting. **(A)** Average cumulative tail of the mutational Site Frequency Spectra. **(B)** Distributions of allele counts at t_f . **(C)** Distributions of singleton counts at t_f . **(D-F)** Distributions of counts of driver mutations **(D)**, passenger mutations **(E)** and divisions **(F)** within $[t_0, t_f]$. **(G-N)** Trajectories of the averages over time of population sizes (+/-std) **(G)**, cumulative mutation counts **(H)**, fitness **(I)**, cumulative division/replacement counts **(J)**, allele counts **(K)**, percentage of singletons among all alleles **(L)**, allele birth counts **(M)** and allele death counts **(N)**. Allele death counts (lines) are categorized into mutation events (circles) and division/replacement events (diamonds). Dark blue = Moran A, green = Moran B, yellow = “binomial BP” with $g_0 = g_2 = 0.25$ (non-extinction), orange = “binomial BP” with $g_0 = g_2 = 0.25$ ($N(t_f) \in [90, 110]$), purple = “fast BP” with $g_0 = g_2 = 0.5$ ($N(t_f) \in [90, 110]$), cyan = “slow BP” with $g_0 = g_2 = 0.05$ ($N(t_f) \in [90, 110]$), gray = “supercritical BP” with $g_0 = 0.2465$, $g_1 = 0.5$ and $g_2 = 0.2535$ (non-extinction).

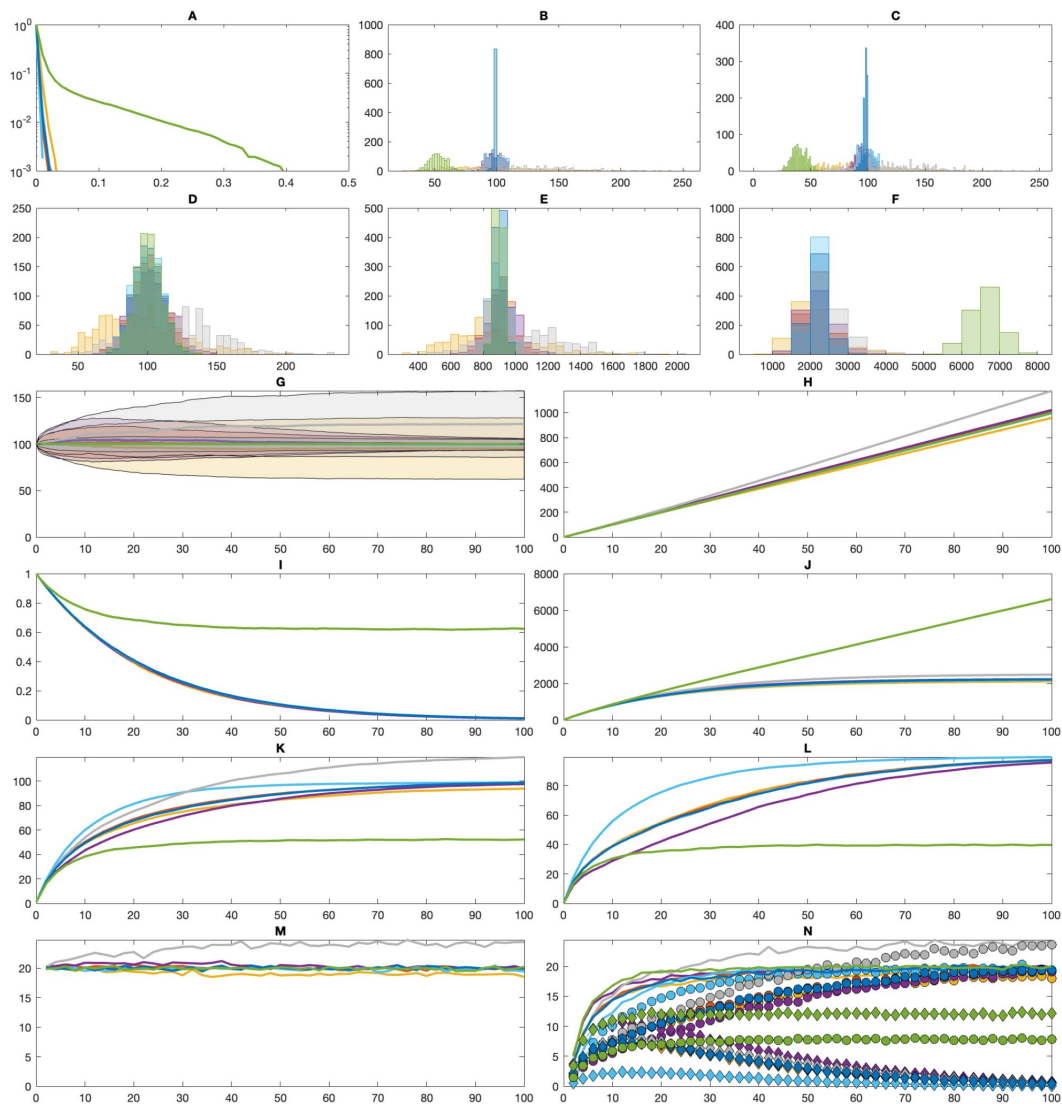


Figure 6

Comparisons between Moran and branching process (BP) models in the “deleterious” setting. **(A)** Average cumulative tail of the mutational Site Frequency Spectra. **(B)** Distributions of allele counts at t_f . **(C)** Distributions of singleton counts at t_f . **(D-F)** Distributions of counts of driver mutations **(D)**, passenger mutations **(E)** and divisions **(F)** within $[t_0, t_f]$. **(G-N)** Trajectories of the averages over time of population sizes (+/-std) **(G)**, cumulative mutation counts **(H)**, fitness **(I)**, cumulative division/replacement counts **(J)**, allele counts **(K)**, percentage of singletons among all alleles **(L)**, allele birth counts **(M)** and allele death counts **(N)**. Allele death counts (lines) are categorized into mutation events (circles) and division/replacement events (diamonds). Dark blue = Moran A, green = Moran B, yellow = “binomial BP” with $g_0 = g_2 = 0.25$ (non-extinction), orange = “binomial BP” with $g_0 = g_2 = 0.25$ ($N(t_f) \in [90, 110]$), purple = “fast BP” with $g_0 = g_2 = 0.5$ ($N(t_f) \in [90, 110]$), cyan = “slow BP” with $g_0 = g_2 = 0.05$ ($N(t_f) \in [90, 110]$), gray = “supercritical BP” with $g_0 = 0.2465$, $g_1 = 0.5$ and $g_2 = 0.2535$ (non-extinction).

Model	B	C	D	E	F
Dark blue	98.5±1.2	97.3±2.3	99.7±10.0	901.1±28.7	2216.4±236.6
Green	52.2±6.6	39.5±6.4	100.5±9.8	898.1±28.8	6610.3±395.6
Yellow	93.7±32.0	92.4±31.4	96.3±29.2	862.2±241.9	2106.2±579.5
Orange	98.4±5.9	97.1±6.1	100.0±12.1	904.2±75.6	2187.1±301.4
Purple	97.5±6.1	95.5±6.5	101.7±14.2	920.7±92.6	2226.7±406.0
Cyan	98.8±5.9	98.5±6.0	99.1±11.1	895.5±58.5	2183.8±186.4
Gray	119.6±34.7	117.9±34.1	117.3±29.4	1059.3±255.4	2477.0±573.8

Table 5

Statistics for Figure 6 [↗](#) (passenger domination)

singleton count (**Figure 6A-B, J-K**). Despite accumulating passenger mutation, after the initial period of dropping the average fitness stays at nonzero value, due to the drift process favoring fixation of clones with higher fitness.

In the deleterious evolution scenario, since cells stop dividing, the main reason of cells' death is mutation (**Figure 6N**), except model B in which cells keep dividing (**Figure 6J**).

3.2 Fitting breast cancer SFS

We use the Moran A model to fit the mutational SFS from 3 samples of breast cancer. We fix population size $N = 100$ cells, average time between mutation events $L = N\mu = 6$, probability of driver mutations $p = 0.01$, final time $t_f = 100$. We then vary the values for s and d , simulate the SFS from 1,000 simulations and compute the average SFS. For a given sample, we compute the cumulative tail of the SFS $S(f_i)$ i.e., the proportion of mutations occurring at frequencies $> f_i$, and similarly the average $\{S(f_i | s, d)\}$ for every combination of (s, d) . The reverse cumulative SFS is evaluated for mutations with frequency $f > 0.05$. The error for (s, d) is defined as

$$\sum_{i=1}^I |\log_{10} \bar{S}(f_i) - \log_{10} S(f_i | s, d)|$$

where I is the largest index such that $\bar{S}(f_i)$ and $S(f_i | s, d)$ are both positive.

Figures 7-9 present the fitting results for the SFS from the breast SFS data. The (s, d) combinations with low error exhibit a trade-off between driver and passenger mutations: the observed SFS can be simulated by Moran A with either low values for s and d , or high values for both (panel A in each figure). As a result, the 100 best (s, d) parameters (marked as squares) can be approximated by linear regression.

The range of best d parameter values does not vary among cases, with the G2 sample (**Figure 7**, panel A) having greater tolerance for changes in this parameter value. The impact of the s parameter on the SFS tail shape is significant: the range of the best fits varies between cases. The value of s is small for sample G2 (**Figure 7**, panel A) and slightly bigger for G32 (**Figure 8**, panel A). In the case of sample G41, the best fit was obtained using a relatively high s parameter value (**Figure 9**, panel A), indicating strong selection.

We compare the SFS from the 100 best (s, d) parameters against the breast cancer data-based SFS (**Figures 7-9**, panel B). The SFS from each sample is well fitted with Moran A, not only with the optimal parameter set but also with other (s, d) combinations. The fit is particularly good for the region of SFS with low frequency ($f < 0.2$). There are exponentially fewer mutations occurring at larger f , resulting in relatively higher discrepancy in the long tail of the SFS from Moran A. However, the overall shape of the observed SFS can be fitted well by the Moran A model.

We also compare the SFS from the binomial BP ($g_0 = g_2 = 0.25$, $N(t_f) \in [90, 110]$), using the 100 best (s, d) parameters from Moran A inference (**Figures 7-9**, panel C). Similarly to Moran A, the SFS from BP can fit the shape of the observed SFS tail well. This is consistent with our finding from the previous section that the binomial BP with tight conditioning on the final population size behaves similarly to the Moran A model, resulting in similar statistics under a range of selection scenarios.

4 Discussion and Conclusion

As it has been expected, Moran model A behaves comparably to the binomial BP conditioned on final population size being close to the initial count. This manifests in similar statistics under different extremes of selection scenarios (Section 3.1), including values that are observable in

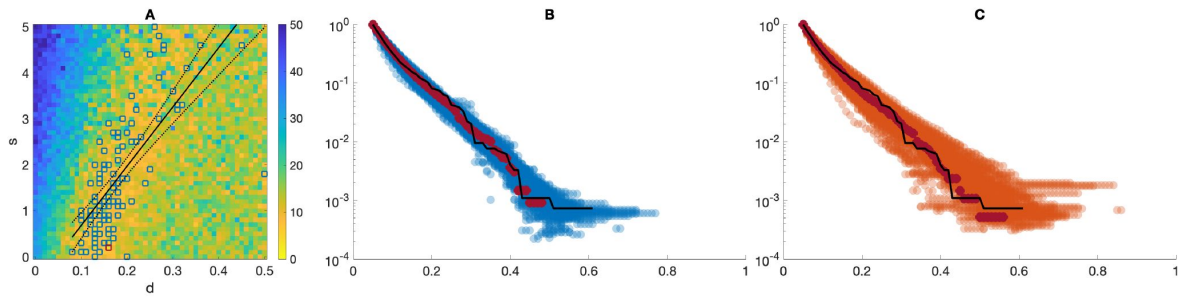


Figure 7

Results from fitting the SFS from sample G2. **(A)** Heatmap of error between data SFS tail and average SFS tail for distinct (s, d) combinations in Moran A. Dark blue squares: 100 best (s, d) combinations, with linear regression. Dark red square: best (s, d) combination. **(B)** Comparison between sample SFS (black line), average SFS under Moran A from 100 best (s, d) combinations (dark blue dots) and the best (s, d) combination (dark red dots). **(C)** Comparison between sample SFS (black line), average SFS under “binomial BP” ($g_0 = g_2 = 0.25, N(t_f) \in [90, 110]$) from 100 best (s, d) combinations (red dots) and the best (s, d) combination (dark red dots). Each SFS from Moran A or binomial BP is averaged from 1,000 simulations.

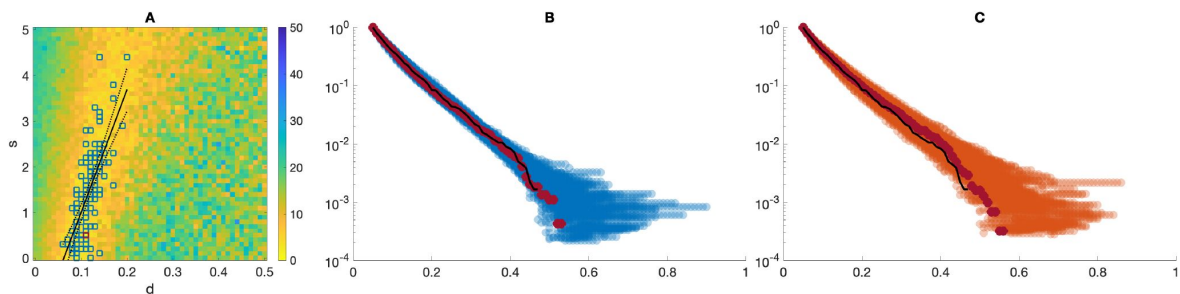


Figure 8

Results from fitting the SFS from sample G32. **(A)** Heatmap of error between data SFS and average SFS for distinct (s, d) combinations in Moran A. Dark blue squares: 100 best (s, d) combinations, with linear regression. Dark red square: best (s, d) combination. **(B)** Comparison between sample SFS (black line), average SFS under Moran A from 100 best (s, d) combinations (dark blue dots) and the best (s, d) combination (dark red dots). **(C)** Comparison between sample SFS (black line), average SFS under “binomial BP” ($g_0 = g_2 = 0.25, N(t_f) \in [90, 110]$) from 100 best (s, d) combinations (red dots) and the best (s, d) combination (dark red dots). Each SFS from Moran A or binomial BP is averaged from 1,000 simulations.

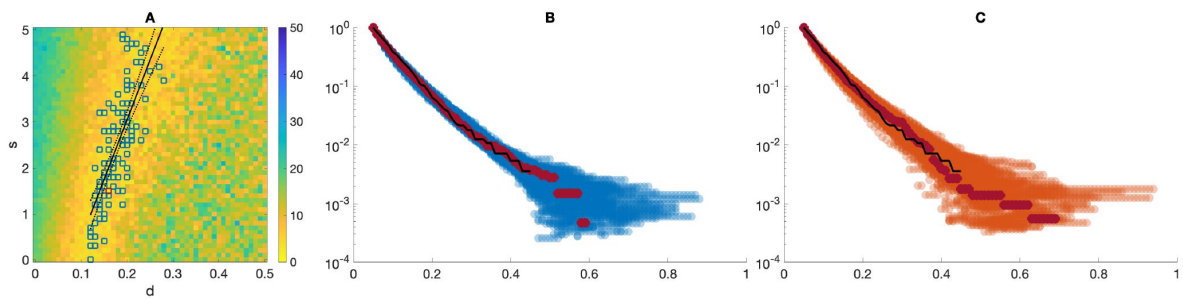


Figure 9

Results from fitting the SFS from sample G41. **(A)** Heatmap of error between data SFS and average SFS for distinct (s, d) combinations in Moran A. Dark blue squares: 100 best (s, d) combinations, with linear regression. Dark red square: best (s, d) combination. **(B)** Comparison between sample SFS (black line), average SFS under Moran A from 100 best (s, d) combinations (dark blue dots) and the best (s, d) combination (dark red dots). **(C)** Comparison between sample SFS (black line), average SFS under "binomial BP" ($g_0 = g_2 = 0.25, N(t_f) \in [90, 110]$) from 100 best (s, d) combinations (red dots) and the best (s, d) combination (dark red dots). Each SFS from Moran A or binomial BP is averaged from 1,000 simulations.

sequencing samples. Crucially, SFS fitting results for experimental breast cancer data (Section 3.2) are similar between the two models. This finding might hold mathematical importance, and requires further investigation. Moreover, interestingly, the similarity between Moran A and binomial BP becomes more pronounced as the latter is conditioned more tightly to resemble the constant population size expected in Moran models. When conditioned only on non-extinction, the population size of each BP realization may deviate significantly (**Figures 3-6G**). This leads to higher variances in allele and singleton counts (**Figures 3-6B-C, K-L**) as well as mutation and division/replacement counts (**Figures 3-5D-F, H-N**), although the means of these statistics remain similar (Appendix B, Tables 2-5). However, the high population size variance also results in different SFS in non-extinction BP compared to tightly conditioned BP and Moran model A (**Figures 3-6A**).

Our Moran model B, similar though not identical to the model introduced in [2], exhibits a phenomenon known as the drift barrier, which prevents the deleterious passenger mutations from dominating temporal trends in fitness, even under mutation-selection balance condition tipped in their favor ($sp < dq$). Indeed, under this condition, due to drift, the fitness may decrease or increase depending on how much smaller sp is than dq . In addition, fitness generally increases at mutation-selection equilibrium ($sp = dq$). On the contrary, trends in Moran model A and binomial BP model follow the mutation-selection condition.

The effect of increasing fitness in model B was already predicted in [2] and described in [16]. While in Moran A fitness stays constant (as expected), in the case of Moran B, clones with higher fitness are favored, even for the same initial conditions and in absence of new mutations (see section 3.1.1 in [16]). This behavior results from the difference between Moran A and Moran B in the expected change in population fitness after a death-replacement event. As shown in Eqs. (4) and (5) in [16], the expected fitness change is 0 in Moran A and is ≥ 0 in Moran B. In general, fitness in Moran A depends only on the balance between drivers and passengers, while the trends in Moran B are more complex, as explained mathematically and confirmed by simulations in [2]. The drift and selection pattern in Moran B biases it toward increasing fitness.

Among the BP variations, the supercritical model behaves differently in all cases compared to other models, due to the difference in population size growth rate (**Figures 3-5G**). The only exception is the deleterious evolution setting, in which the impact of supercriticality is less prominent, since the fitness being close to zero means cells stop dividing. In this scenario all statistics are much more similar to those obtained from other models (Appendix C, **Table 5**).

Our experiments with different progeny cell count distributions in BP show that the fast BP always has higher variances in the population size throughout time compared to the binomial BP, even if similarly conditioned (**Figures 3-6G**). The fast BP also results in both less alleles and lower percentage of singletons within all alleles (**Figures 3-5B-C, K-L**). This is not observed in case of deleterious evolution (**Figure 6B-C, K-L**), where there is only a difference in the rate of reaching the steady state, which is the same as for other models. On the other hand, the averages of mutation and division/replacement event counts (**Figures 3-6D-F** and Appendix C, **Table 2-5**) do not differ from averages for Moran model A and binomial BP. Reversely, the population size in slow BP varies less, and the allele count and percentage of singleton count is much higher, compared to the binomial BP.

There are features shared between all Moran models and BP variations across different selection scenarios. The more division/replacement events occur during a simulation, the less alleles and singletons we observe both at the sampling time point as well as throughout tumor history. This is especially pronounced in case of deleterious evolution, where in Moran B continuously dividing population under selective pressure prevents the accumulation of singletons (**Figure 6B-C, K-L**). Conversely, if the events occurring during a simulation are dominantly mutations, then the

population consists of more alleles and singletons. In conclusion, across models, higher selection is associated with less alleles and singletons, higher pace of allele death, and cumulative SFS with fat tail.

It seems relevant to note that the frequently cited reference by Gerrish and Lenski [11] introduces a model of competition in populations of constant size in an asexual population. From the reading of this corner-stone paper, it seems that it uses results from supercritical branching processes and then just scales them intuitively into the constant population size framework. This method seems not mathematically rigorous. Our comparison of Moran and branching process models identifies subtle but important differences between the two approaches. Overall, we have shown that the critical binomial BP and the Moran A model behave similarly in the Tug-of-War setting under distinct selection scenarios. This finding is relevant for improving simulating efficiency and optimizing model inference. Branching process and Moran model remain the two main stochastic modeling approaches in population genetics, where they provide the theoretical framework to uncover a tumor's history from sequencing snapshots. However, BP simulation is considerably more time-consuming, as the cell population size can change arbitrarily due to random fluctuations. This problem is exacerbated in critical or near-critical BP, which is applicable for modeling many cancers. In this setting, the BP often has high probability of extinction, hence the high fraction of simulations that have to be discarded makes model inference computationally costly. In such cases, it would be more efficient to employ Moran A, which we have shown to provide comparable sample statistics and which is easier to implement. However, more work is needed to establish the theoretical equivalence between the Moran A model and the critical BP, and if this compatibility breaks down under certain conditions.

Both the Moran A model and the binomial BP can fit the SFS tail in our breast cancer samples well. However, the inference is complicated by a wide range of selection coefficients that result in equally comparable SFS to the data. These coefficients exhibit a trade-off between driver and passenger mutations, as the same SFS can result from driver mutations being more advantageous if the passenger mutations are also more deleterious, and vice versa. Therefore, the mutational SFS alone is not adequate to differentiate between these different selection settings. Separately, we found that the inference for all of our samples requires $d > 0$, confirming the observations from McFarland et al. [19] that passenger mutations exhibit a deleterious effect during tumor progression.

Data availability

The breast cancer sequencing data can be found under <https://ega-archive.org/> with accession number: EGAD00001009081. Any queries should be directed to the corresponding author.

Funding

This research was funded by a subsidy for the maintenance and development of research potential BKM-581/RAU1/2023 (02/040/BKM23/1048) granted by the Polish Ministry of Science and Higher Education (M.K.K) and by Polish National Science Center grant 2021/41/B/NZ2/04134 (M.K.).

K.D. acknowledged the support from the Herbert and Florence Irving Institute for Cancer Dynamics and Department of Statistics at Columbia University.

Appendix

A Elements of mathematical population genetics

A.1 Wright-Fisher model and Moran model comparison

A1.1 Wright-Fisher model

There are two copies of each gene in a diploid cell. The copies can be of the same allele (e.g. AA or aa) or two different alleles (Aa). If the population consists of N diploid individuals, there exist $2N$ copies in total. In the accordance with Wright-Fisher model, in each generation random alleles are drawn with replacement from gene pool of size $2N$ – the generations do not overlap. Such reproduction scheme can be described mathematically as a discrete-time Markov chain – the future allele frequencies are dependent only on the present frequencies, not on those from past generations. Due to the randomness in the process, allele frequencies change at a rate which is inversely proportional to the population size. These fluctuations correspond to a process of genetic drift.

The state of the population in each generation can be described as the number of A alleles in the population, which can range from 0 (loss of allele A) to $2N$ (fixation of allele A and loss of allele a). The states 0 and $2N$ are called “absorbing states” because the population is not able to leave any of these (considering no mutation or migration events). In other cases, the transition probability can be calculated based on binomial distribution. In the population with i copies of allele A , the frequency of allele A is equal to $p = i/2N$ and the frequency of allele a is $q = 1 - p$. The probability of changing state from i copies of A to j copies of A (for $i, j = 0, 1, 2, \dots, 2N$) in one generation is [13]:

$$T_{ij} = \binom{2N}{j} p^j q^{2N-j} \quad (4)$$

Moran model

In the Moran model [21], in each step of algorithm, a single random allele x from haploid population of $2N$ individuals dies and is being “replaced” by another randomly chosen allele from the population (including x itself), thus ensuring that the population size remains constant. Unlike the Wright-Fisher model, where an allele can have up to $2N$ offspring, in the Moran model the allele can have 0 or 2 descendants. The time between birth-death events (the lifespan of an individual allele) is exponentially distributed with mean equal to 1 and generations do overlap.

As in the Wright-Fisher model, this reproduction scheme can be described mathematically by Markov chain. In this case, however, it is the continuous-time Markov chain with values in the set $\{0, 1, \dots, 2N\}$. Therefore, the mathematical solutions for the Moran model are easier to ascertain than the Wright-Fisher model. However, there are fewer time steps to be computed in the Wright-Fisher model. As a result, the simulations are more computationally convenient in the Wright-Fisher framework, compared to the Moran model. The Moran and Wright-Fisher models give qualitatively similar results, but genetic drift runs twice as fast in the Moran model [9].

The details regarding Moran model are described in section 2.1.

A.2 Infinitely many alleles version of Wright-Fisher model

For neutrality testing purposes we consider the “infinitely many alleles” (IAM) version of the Wright-Fisher model. This type of model is particularly useful in molecular population genetics and was inspired by molecular nature of the gene. Average gene is sequence of 3000 nucleotides

(A, G, T and C), so there are 4^{3000} possible sequences (alleles), the number which for practical purposes can be taken as infinity, what leads to infinitely many alleles model.

Most nucleotide mutations will lead to sequences not currently existing in the population, so in this case all mutants are assumed to be of a new allelic type - there is no reverse mutation. In such a model each allele will sooner or later be lost from the population.

A.2.1 Expected allele number

Under assumptions of the IAM, let us define $\mathbf{A} = (A_1, A_2, \dots, A_n)$ as the vector of the allelic types each of which is represented by exactly j genes in the sample. The following Ewens Sampling Formula (ESF) was derived in [10] and [14]

$$\mathbb{P}(\mathbf{A} = \mathbf{a}) = \frac{n! \theta \sum a_j}{1^{a_1} 2^{a_2} \dots n^{a_n} a_1! a_2! \dots a_n! S_n(\theta)}, \quad (5)$$

with

$$S_n(\theta) = \theta(\theta + 1)(\theta + 2) \dots (\theta + n - 1)$$

Furthermore, the probability distribution of number K of different alleles in the sample has the form

$$\mathbb{P}(K = k) = |S_n^k| \theta^k / S_n(\theta). \quad (6)$$

Under neutrality $\theta = n\mu$, where μ denotes mutation coefficient and n is the sample size. $|S_n^k|$ is the absolute value of a Stirling number of the first kind. The expression for the expected value of K can be derived from Equ. (6)

$$\mathbb{E}(K) = \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \dots + \frac{\theta}{\theta+n-1}, \quad (7)$$

with the corresponding expression for variance of K

$$\mathbb{V}(K) = \theta \sum_{j=1}^{n-1} \frac{j}{(\theta+j)^2}, \quad (8)$$

Equations (5) and (6) show jointly that the conditional distribution of the vector $\mathbf{A} = (A_1, A_2, \dots, A_n)$, given the value of K , is

$$\mathbb{P}\{\mathbf{A} = \mathbf{a} | K = k\} = \frac{n!}{|S_n^k| 1^{a_1} 2^{a_2} \dots n^{a_n} a_1! a_2! \dots a_n!}, \quad (9)$$

where $\mathbf{a} = (a_1, a_2, \dots, a_n)$. From (9) the procedure can be derived, which allows testing the null hypothesis that the alleles in the sample are selectively equivalent (neutral).

B Statistics in Section 3.1

Tables 2, 3, 4, 5 contain some statistics for Figures. 3, 4, 5, 6, respectively. The models are named according to their colors in the figures, which are listed in legend for Table 2 along with the statistics.

References

- [1] Athreya Krishna B, Ney Peter E, Ney PE (1972) **Branching processes**
- [2] Bobrowski Adam, Kimmel Marek, Kurpas Monika K, Ratajczyk Elżbieta (2023) **Moran process version of the tug-of-war model: Behavior revealed by mathematical analysis and simulation studies** *Discrete and Continuous Dynamical Systems-B* **28**:4532–4563
- [3] Burden Conrad J, Simon Helmut (2016) **Genetic drift in populations governed by a Galton–Watson branching process** *Theoretical population biology* **109**:63–74
- [4] Cox Alexander MG, Horton Emma, Villemonais Denis (2022) **Binary branching processes with Moran type interactions** *arXiv preprint*
- [5] Cyran Krzysztof A, Kimmel Marek (2010) **Alternatives to the Wright–Fisher model: The robustness of mitochondrial Eve dating** *Theoretical Population Biology* **78**:165–172
- [6] Dingli David, Traulsen Arne, Michor Franziska (2007) **(A)symmetric stem cell replication and cancer** *PLoS computational biology* **3**
- [7] Diossy Miklos, Sztupinszki Zsafia, Krzystanek Marcin, Borcsok Judit, Eklund Aron C, Csabai István, Pedersen Anders Gorm, Szallasi Zoltan (2021) **Strand Orientation Bias Detector to determine the probability of FFPE sequencing artifacts** *Briefings in Bioinformatics* **22**
- [8] Do Hongdo, Dobrovic Alexander (2015) **Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization** *Clinical chemistry* **61**:64–71
- [9] Durrett Richard (2008) **Probability models for DNA sequence evolution**
- [10] Ewens Warren J (1972) **The sampling theory of selectively neutral alleles** *Theoretical population biology* **3**:87–112
- [11] Gerrish Philip J, Lenski Richard E (1998) **The fate of competing beneficial mutations in an asexual population** *Genetica* **102**
- [12] Greenman Christopher *et al.* (2007) **Pat-terns of somatic mutation in human cancer genomes** *Nature* **446**:153–158
- [13] Hartl Daniel L, Clark Andrew G, Clark Andrew G (1997) **Principles of population genetics**
- [14] Karlin. Samuel (1972) **Addendum to a paper of W. Ewens** *Theor. Popul. Biol* **3**:113–116
- [15] Kimmel M, Axelrod DE (2015) **Branching processes in biology**
- [16] Kurpas Monika K, Kimmel Marek (2022) **Modes of selection in tumors as reflected by two mathematical models and site frequency spectra** *Frontiers in Ecology and Evolution* **10**
- [17] Kurpas Monika Klara, Jaksik Roman, Kuś Pawel, Kimmel Marek (2022) **Genomic analysis of SARS-CoV-2 Alpha, Beta and Delta Variants of Concern uncovers signatures of neutral and non-neutral evolution** *Viruses* **14**

- [18] McFarland Christopher D, Mirny Leonid A, Korolev Kirill S (2014) **Tug-of-war between driver and passenger mutations in cancer and other adaptive processes** *Proceedings of the National Academy of Sciences* **111**:15138–15143
- [19] McFarland Christopher D, Yaglom Julia A, Wojtkowiak Jonathan W, Scott Jacob G, Morse David L, Sherman Michael Y, Mirny Leonid A (2017) **The damaging effect of passenger mutations on cancer progression** *Cancer Research* **77**:4763–4772
- [20] McFarland Christopher Dennis (2014) **The role of deleterious passengers in cancer** *PhD thesis*
- [21] Moran Patrick Alfred Pierce, et al. (1962) **The statistical processes of evolutionary theory** *The statistical processes of evolutionary theory*

Article and author information

Khanh N. Dinh

Irving Institute for Cancer Dynamics and Department of Statistics, Columbia University, New York, NY, USA

For correspondence: knd2127@columbia.edu

ORCID iD: [0000-0002-0010-4251](https://orcid.org/0000-0002-0010-4251)

Monika K. Kurpas

Department of Systems Biology and Engineering, Silesian University of Technology, Gliwice, Poland

Marek Kimmel

Department of Systems Biology and Engineering, Silesian University of Technology, Gliwice, Poland, Departments of Statistics and Bioengineering, Rice University, Houston, TX, USA

Copyright

© 2024, Dinh et al.

This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

Editors

Reviewing Editor

Jennifer Flegg

The University of Melbourne, Melbourne, Australia

Senior Editor

Alan Moses

University of Toronto, Toronto, Canada

Reviewer #1 (Public Review):

This paper can be seen as an extension of a recent study by two of the same authors [1]. In the previous paper, the authors considered two variants of the Moran process, labelled Model

A and Model B, and examined differences between the evolutionary dynamics of these two models. They further described the site frequency spectra, expected allele counts, and expected singleton counts of these models, building on analytical results from prior studies, and used numerical simulations to investigate the models' evolutionary dynamics. Finally, they compared the site frequency spectra of the two models (using numerical simulations) to spectra derived from a small breast cancer data set (two sets of three samples).

In the new paper, the authors consider the same two Moran process variants (Model A and Model B) and some related branching processes. As before, they compare the site frequency spectra and various summary statistics of these models, but here they present only numerical simulations (except that some prior analytical results are summarized in Appendix A, which are never referred to in the main text and seem unconnected to the study). They then compare the site frequency spectra of these models (again using numerical simulations) to those derived from the same breast cancer samples as before and thus infer some evolutionary parameters.

The first main conclusion is that the critical branching process and the Moran process models behave similarly and generate similar site frequency spectra. This finding is unsurprising (indeed, the authors acknowledge that the result "has been expected"). For a reasonably large population size, the population size in the critical branching process has been shown to vary relatively little over time and the model is thus essentially a continuous time Moran process (see, for example, Equation 8.55 in ref 2). Nor is it surprising that the authors see stronger similarities when they select only the subset of branching process replicates in which the final population size is particularly close to the initial population size (this is because, in these replicates, the population size likely varies even less than usual).

The second main conclusion is that, although "the mutational SFS alone is not adequate" to quantify the strength of selection, "All fitted values for the selective disadvantage of passenger mutations are nonzero, supporting the view that they exert deleterious selection during tumorigenesis". Although the question of whether mildly deleterious mutations play an important role in cancer evolution is of considerable interest, it's debatable whether the results presented here help resolve the issue.

Many prominent researchers have called into question whether cancer evolutionary parameters can be reliably inferred from site frequency spectra (e.g., [3-7]), even using sophisticated statistical methods. The statistical approach used here (though not named as such in the paper) is a crude kind of approximate Bayesian computation. To improve the accuracy of the results, it would have been better to have set reasonably vague priors for the uncertain mutation rates, rather than fixing them arbitrarily. It would also have been better to have chosen a likelihood function explicitly based on an analysis of the sampling and error distributions, rather than just summing the absolute logged deviations. It is well known that "Checking the model is crucial to statistical analysis" and "A good Bayesian analysis, therefore, should include at least some check of the adequacy of the fit of the model to the data and the plausibility of the model for the purposes for which the model will be used" [8]. The authors' failure to describe any attempt to validate or check their model, using simulated data or otherwise, casts doubt on the reliability of their inferences.

Putting aside the potential biasing effects of sampling error, measurement error, and the limitations of the authors' statistical method, it is well established that both population growth and spatial structure profoundly alter the shape of site frequency spectra in ways that can mimic the effects of selection (e.g. [9-11]). Indeed, Figures 3, 4 and 5 show that the critical and super-critical branching processes generate markedly different site frequency spectra. It follows that if the population dynamics and spatial structure of the mathematical model used for inference don't match those of the biological process that produced the data then any inferred evolutionary parameter values will be unreliable. Breast cancer has two indisputable ecological features that shape its evolutionary dynamics: the cell population

expands by many orders of magnitude from a single cell, and the population is spatially structured. In the authors' mathematical model, the population size is initially 100 cells and either remains constant or varies little, and there is no spatial structure. These profound mismatches between model and data cast further doubt on what is supposed to be the paper's most important biological finding.

In this paper the authors offer no justification for their decision to model breast cancer as a non-growing, non-spatial cell population. Nor do they engage with the extensive recent literature on the challenges of inferring evolutionary parameters from cancer site frequency spectra (they cite none of the many relevant papers listed at <https://www.sottorivalab.org/neutral-evolution.html>). Their 2022 paper [1] claims that, "it sometimes makes sense to consider cancer growth in the framework of constant-population models. Our models correspond to the situation in which a constant population of N "healthy" stem cells is gradually replaced by a growing clone of transformed cells with increasing fitness." No evidence was presented to support this hypothesis regarding breast cancer progression. On the other hand, a wealth of evidence supports the consensus view that, in breast cancer and other human solid tumours, the number of cells with unlimited proliferative potential is several orders of magnitude greater than 100 and grows over time (e.g. [12]).

Analytic expressions for the site frequency spectra with neutral mutations are already known. It is well known that the site frequency spectrum of an exponentially growing population has a tail following a power law $S_k \sim k^{-2}$ [13, 14]. Similarly, it is known that for the critical branching process or the Moran process, the site frequency spectrum at equilibrium is $S_k \sim k^{-1}$ [13, 15]. Especially noteworthy yet uncited studies that use those results about site frequency spectra to make inferences based on sequencing data include ref 16, in which selection is inferred, and ref 17, in which evolutionary parameters of constant populations (healthy cell populations) are inferred.

Although the paper is well written, the figures are ineffective in communicating the results. As others have put it, "A figure is meant to express an idea or introduce some facts or a result that would be too long (or nearly impossible) to explain only with words" and "If your figure is able to convey a striking message at first glance, chances are increased that your article will draw more attention from the community" [18]. On the contrary, Figures 3, 4, 5 and 6 are bewilderingly complicated, crowded, and repetitive. These figures comprise no fewer than fifty-six plots, each containing numerous curves or histograms, spread across four pages. To compare the results of different scenarios, the reader is presumably expected to put these figures side by side and try to spot the differences, hampered by inconsistent axis ranges, absence of axis labels, absence of titles, absence of legends, and unreliable captions ("cyan" seems to refer to pale blue, and "orange" to something closer to red). For example, the only notable difference between Figures 3 and 4 is in the shape of a single green curve in panel I. In the main text of a published paper, one would expect fewer, more carefully curated figures drawing attention to salient features, so that the reader can infer the main results with minimal effort. The rest can be put in supplementary figures.

In summary, this paper adds somewhat to our understanding of some standard mathematical models; whether it tells us anything new about cancer is open to debate.

References

- (1) Kurpas, Monika K., and Marek Kimmel. "Modes of selection in tumors as reflected by two mathematical models and site frequency spectra." *Frontiers in Ecology and Evolution* 10 (2022): 889438.
- (2) Bailey, Norman TJ. *The elements of stochastic processes with applications to the natural sciences*. John Wiley & Sons, 1964.
- (3) Tarabichi, Maxime, et al. "Neutral tumor evolution?." *Nature Genetics* 50.12 (2018): 1630-1633.
- (4) McDonald, Thomas O., Shaon Chakrabarti, and Franziska Michor. "Currently available

bulk sequencing data do not necessarily support a model of neutral tumor evolution." *Nature Genetics* 50.12 (2018): 1620-1623.

(5) Balaparya, Abdul, and Subhajyoti De. "Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data." *Nature Genetics* 50.12 (2018): 1626-1628.

(6) Noorbakhsh, Javad, and Jeffrey H. Chuang. "Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures." *Nature Genetics* 49.9 (2017): 1288-1289.

(7) Bozic, Ivana, Chay Paterson, and Bartłomiej Waclaw. "On measuring selection in cancer from subclonal mutation frequencies." *PLoS Computational Biology* 15.9 (2019): e1007368.

(8) Neher, Richard A., and Oskar Hallatschek. "Genealogies of rapidly adapting populations." *Proceedings of the National Academy of Sciences* 110.2 (2013): 437-442.

(9) Gelman, Andrew, et al. *Bayesian data analysis* (Third Edition). Chapman and Hall/CRC, 2014.

(10) Fusco, Diana, et al. "Excess of mutational jackpot events in expanding populations revealed by spatial Luria-Delbrück experiments." *Nature Communications* 7.1 (2016): 12760.

(11) Noble, Robert, et al. "Spatial structure governs the mode of tumour evolution." *Nature Ecology & Evolution* 6.2 (2022): 207-217.

(12) Lawson, Devon A., et al. "Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells." *Nature* 526.7571 (2015): 131-135.

(13) Gunnarsson, Einar B., Leder, Kevin, and Foo Jasmine. "Exact site frequency spectra of neutrally evolving tumors: A transition between power laws reveals a signature of cell viability" *Theoretical Population Biology* 142 (2021) 67-90

(14) Durrett, Richard "Branching Process Models of Cancer" Springer (2015)

(15) Durrett, Richard "Probability Models for DNA Sequence Evolution" Springer Science & Business media (2008)

(16) Williams, Mark J. et al. "Quantification of subclonal selection in cancer from bulk sequencing data." *Nature Genetics* 50 (6). 895-903 (2018)

(17) Moeller, Marius E. et al. "Measures of genetic diversification in somatic tissues at bulk and single-cell resolution" *eLife* (2024) 12:RP89780

(18) Rougier, Nicolas P., Michael Droettboom, and Philip E. Bourne. "Ten simple rules for better figures." *PLoS Computational Biology* 10.9 (2014): e1003833.

<https://doi.org/10.7554/eLife.94597.1.sa1>

Reviewer #2 (Public Review):

Summary:

In this manuscript, the authors present a comparison of two models of cancer evolution with advantageous drivers and deleterious passengers: a fixed-population "Moran" model, and a "Branching Process" (BP) model with dynamic population size. The Moran model is more mathematically-tractable, but since cancer is a disease of uncontrolled growth, it is unclear to me how clinically-relevant it is to consider a model with constant population size. Intriguingly, both models can explain observed Site Frequency Spectrums (SFSs) in three breast cancers, which suggests that the Moran model may have some value. This distinction between the two models is addressed well.

Strengths:

The comparisons of the various BP models (extinction/non-extinction, and balanced/supercritical) are very interesting. The survivability of rare, fitness-disadvantaged clones has huge implications for treatment resistance in general - drug resistant clones are very often disadvantaged in the absence of drug. Clinical sequencing is, most decidedly, investigating population dynamics conditioned on non-extinction, however most published models do not condition on non-extinction - an unfortunate community oversight that this publication rectifies.

Site Frequency Spectrums in three breast cancers are measured with unprecedented resolution to my knowledge (allele abundances below one in a thousand).

Detailed description of the behavior of the various models.

Weaknesses:

I do not believe Moran B is a useful theoretical distinction between Moran A. Incorporating fitness effects into the birth process, instead of the death process, is generally mathematically equivalent when time is measured in generations (or cell divisions). Visible differences in the two models in Figures 2-6 by all accounts seem to be due to the fact that Moran B experiences more evolution in the balanced/driver-dominated case, and less evolution in the passenger dominated case. We generally do not use arbitrary time steps for this reason - we quantify time in 'generations'.

<https://doi.org/10.7554/eLife.94597.1.sa0>

Author Response:

eLife assessment

We thank the Editors for identifying qualified reviewers. We agree that the “evidence supporting this claim (that ‘many breast cancer mutations are mildly deleterious’) is incomplete”. Much more detail is needed to state this decisively and we do not claim completeness here. As far as validation, we carried out synthetic testing of the models as suggested by Reviewer #1 and the results seem good.

Reviewer #1:

We thank the Reviewer for a very thorough examination of not only the current paper but also our previous paper. We agree that the illustration material can be overwhelming and we plan to use the Reviewer’s advice in that matter. In addition, we originally put some textbook material in the Appendix, and arguably some of it may be considered superfluous.

Most of the references the Reviewer provides are known to us, although it is likely we should cite and discuss more. All of the above will be included in the revision we are planning.

The Reviewer is certainly correct that population growth and spatial effects play a major role in cancer. However, the effects of constraining environment are quite strong and the reality lies somewhere between the Moran and branching process models; exactly what we attempt to clarify. As for spatial effects, most tumors extracted in clinic are dissected in bulk and sub-sampling is rare, so the spatial information is rarely accessible.

The subsequent point of importance concerns the weak specificity of the site frequency spectra (SFS) with respect to the underlying genetic and demographic forces. This cannot be denied. However, we just meant to state that our SFS are consistent with a model involving slightly deleterious passengers.

Regarding the validation of the estimation procedures which is a point well-taken, we carried out synthetic testing of the models as suggested by Reviewer #1 and the results seem good. This will be discussed in full in the revision.

In our view, the most important remark is the one concerning scaling of the models. The Reviewer is certainly correct that 100 stem cells are insufficient to drive a realistic tumor. However, what we had in mind but not explained sufficiently, is that a sample of 100 cells corresponds to average-depth coverage in bulk sequencing. Therefore, the strict

interpretation is that the model mirrors what is observed in the sample. A more accurate approach would be to up-scale the model and then sample 100 cells from it. The Moran-type model can be up-scaled using diffusion approximation, and we hope to include these computations in the revision. The associated criticism concerning tumor growth seems less relevant, since we experimented with less or more stringent constraints in our models.

Reviewer #2:

We thank Reviewer #2 for studying our paper and some very positive comments. Among others, the Reviewer underscores the fact that the Moran-type model generates SFS concordant with the data (with all necessary reservations). The Reviewer concurs with us that conditioning on non-extinction is not very common in the literature, while it should be.

Similarly as the Reviewer, we are somewhat puzzled by the differences in behavior between models A and B. Model B seems more parsimonious, but Model A looks more similar to the critical or slightly supercritical branching process. We will work to clarify these observations.

<https://doi.org/10.7554/eLife.94597.1.sa3>