Tug-of-War model of competition between advantageous and deleterious mutations

Marek Kimmel, Rice University
with
Adam Bobrowski, Lublin University of Technology
Khanh Dinh, Columbia University
Roman Jaksik, Silesian University of Technology
Monika Kurpas, Silesian University of Technology
Ren-Yi Wang, Rice University

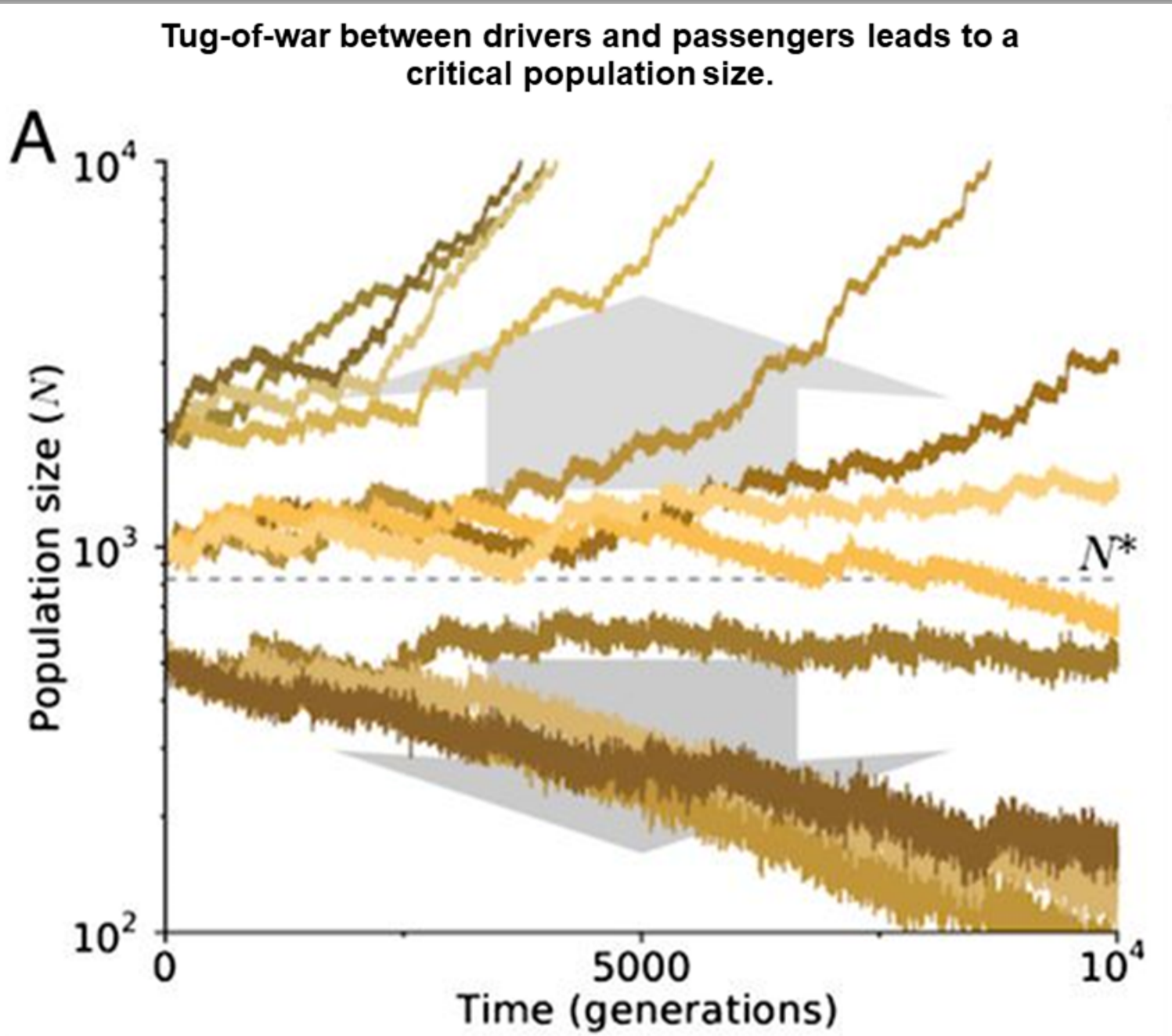
Dec 12, 2024

Tug-of-War

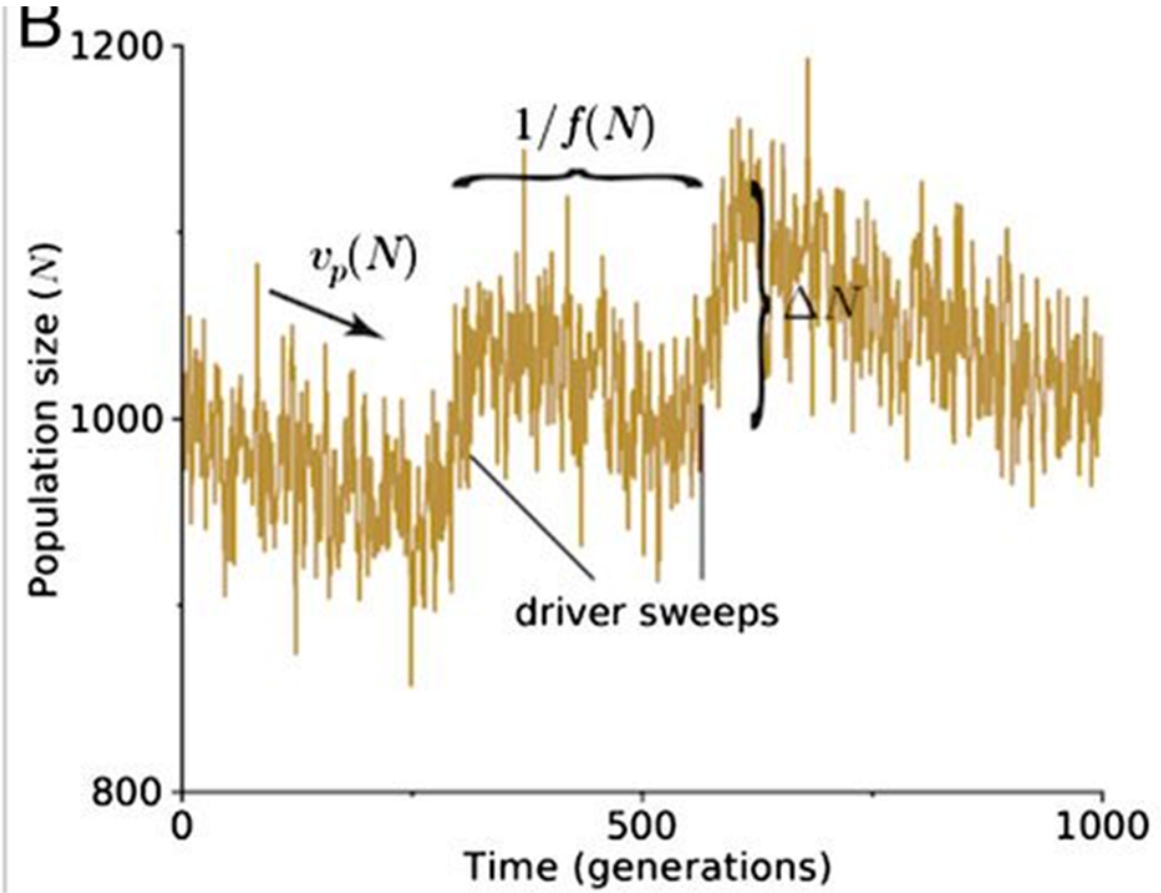
Outline

- Introduced by McFarland and co-authors to study evolution of cancer cell populations.
- We put the tug-of-war in the context of multitype Moran model and multitype branching process, which serve as mathematical framework for two different types of selection in cell populations: "competitive replacement", vs. "crowding out".
- We present mathematical definitions of the Moran model and branching process versions of tug-of-war and simulation result.
- We discuss applications to breast cancer and SARS-Cov-2 virus evolution.

Original Definition: *McFarland et al. (2014) PNAS*



Generic principle



Advantageous but **rare driver** mutation sweeps
are interspersed with periods of slowdown or decline caused by frequent but **mildly deleterious passenger** mutations

Experimental work

- A study by McFarland and co-authors in 2017 showed that passenger mutations are mildly deleterious and might inhibit or reverse tumour growth.
- In their experiment, authors discovered that passengers are capable to reduce proliferative fitness, slow tumour growth, and reduce metastasis progression.
- They also showed some other studies suggesting passenger mutations can trigger anticancer immune responses and the mechanism can be exploited to treat cancer.

Experiment in McFarland et al. (2017)

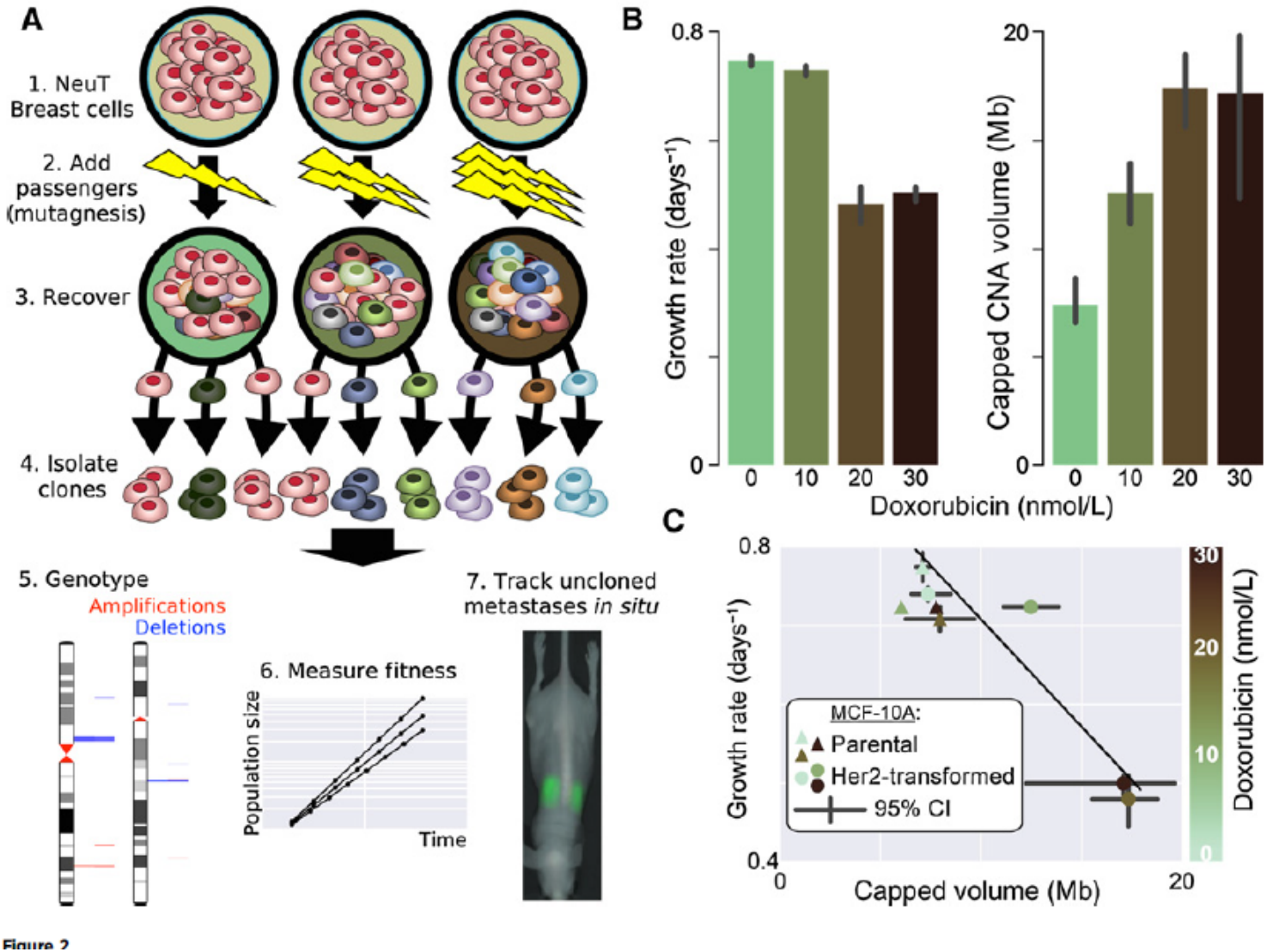


Figure 2

Moran Model: *Bobrowski et al. (2023) DCDS-B*

Population of a fixed number N of individuals

i-th characterized by a pair of integers $\gamma_i = (\alpha_i, \beta_i)$, the numbers of drivers and passengers in its genotype

fitness x_i of the i-th

$$x_i = x_i(\alpha_i, \beta_i) = (1+s)^{\alpha_i}(1-d)^{\beta_i},$$

where $s > 0$ and $d \in (0,1)$ selective values of driver over passenger mutations.

Population state is vector $\Gamma = (\gamma_1, \ldots, \gamma_N)$ of N pairs of non-negative integers.
State space is $S = \mathbb{Z}_+^{2N}$

$x = (x_1, \ldots, x_N)$ of fitnesses determines the future evolution of the population, under drift and selection pressure,

Moran Model: Drift/Selection

Time to the soonest selection/drift event is exponential with parameter $(N-1)\Sigma x$

where $\Sigma x = \sum_{k=1}^{N} x_k$.

After this time, i-th individual dies and is replaced j-th individual ($j \neq i$) wp. $\frac{x_j}{(N-1)\Sigma x}$.

This process then continues with Γ modified by replacing its i-th coordinate by its j-th coordinate.

Moran Model: Mutations

Moreover, each individual may, after an independent exponential time with parameter λ, undergo a mutation event, changing its state to either $(\alpha+1,\beta)$ or $(\alpha,\beta+1)$ with (conditional) probabilities $p \in (0,1)$ and $q = 1-p$, respectively.
Driver and passenger mutations form Poisson processes with parameters

$$\nu = \lambda p, \quad \mu = \lambda q,$$

As heuristically conjectured by Bobrowski, an equilibrium condition for the expectations of the process might have the form

$$s\nu = d\mu \tag{1}$$

which stems from condition $s\nu - d\mu = 0$, expressing no change in expected fitness per replacement/mutation event.
However, this condition is **inaccurate** ... drift will correct it ...

Theorem

Let f' be the state of the process right after drift and selection event of a population f. Then

$$E\, f' \geq f,$$

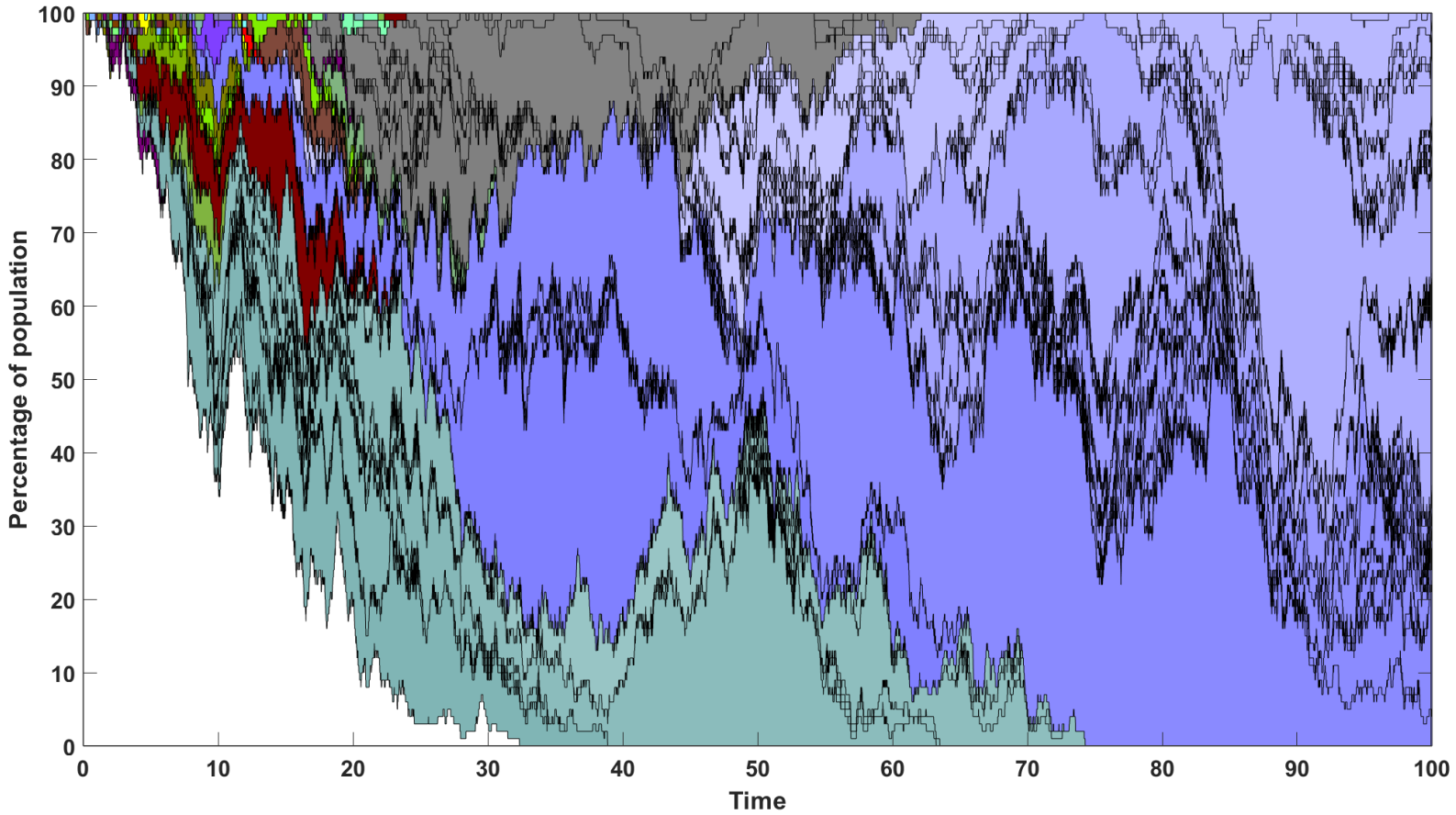
where E denotes expected value.

Proof Each event of replacing the ith coordinate of f by its jth coordinate is paired by the event in which the jth coordinate is replaced by the ith coordinate. The first of these events takes place with probability $f_j/|q_,|$, where $q_,$ is the diagonal element of the generator matrix in Equ. (**??**). Accordingly, $f' - f = f_j - f_i$, and the second event's characteristics are symmetrical. Therefore, $E\, f' - f$ equals

$$\sum_{i<j} \left[\frac{(f_j - f_i) f_j}{|q|} + \frac{(f_i - f_j) f_i}{|q|} \right] = \frac{1}{|q|} \sum_{i<j} (f_i - f_j)^2 \geq 0,$$

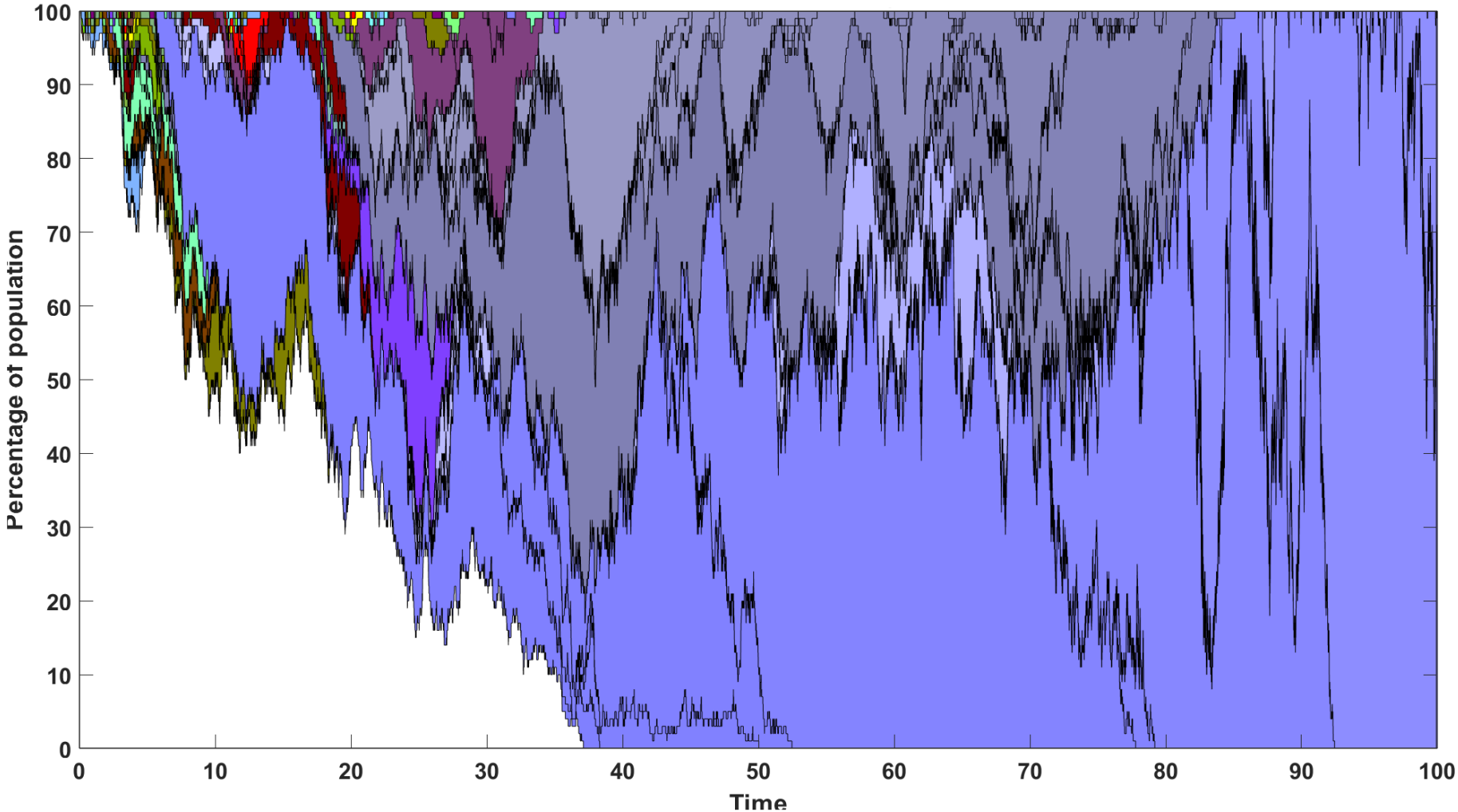
which completes the proof.

Simulated evolution of clones of drivers



Clones started by new drivers in the population of 100 cells.
Different shades of the same color distinguish subsequent clones sharing the same parent.
Neutrality ($s = d = 0$). New clones emerge and then become replaced.
Overall fitness fluctuates

Simulated evolution of clones of drivers



Advantageous drivers versus disadvantageous passengers ($s > d > 0$).

Overall population fitness increases, with the advantageous clone dominating with time.

The generator of the chain can be decomposed into the sum

$$\mathfrak{A} = \mathfrak{S} + \mathfrak{M}$$

of generators corresponding to drift/selection and mutation

The operator semigroup generated by $\mathfrak{S}$ tends to a limit (typical for genetic drift)

Instead of thinking of the chain in S governed by $\mathfrak{A} = \mathfrak{S} + \mathfrak{M}$ we may think of the chain governed by

$$\mathfrak{A}_\kappa =: \kappa\mathfrak{S} + \mathfrak{M}. \tag{2}$$

Asymptotics of Moran Model

$$\mathfrak{A}_\kappa =: \kappa\mathfrak{S} + \mathfrak{M}. \tag{3}$$

For each $\kappa > 0$, $\mathfrak{A}_\kappa$ is the generator of a semigroup, say $\{P_\kappa(t), t \geq 0\}$, of Markov operators. Thus, our task is that of characterizing the limit

$$\lim_{\kappa\to\infty} P_\kappa(t)x$$

where $x \in \ell^1$ is a distribution. We assume that

(i) $\lim_{t\to\infty} P_S(t)x =: \Pi x$ exists for all $x \in \ell^1$.

(ii) Π with domain equal to ℓ^1_0 is a generator in ℓ^1_0, where $\ell^1_0 =: \mathit{Range}\,\Pi$.

Kurtz's perturbation theorem states

$$\lim_{\kappa\to\infty} P_\kappa(t)x = e^{t\Pi\mathfrak{M}}\Pi x, \qquad t > 0. \tag{4}$$

In the limiting chain, any mutation becomes instantly fixed ...
Maybe not the right limit since we want only drivers to be fixed ...

Poor Man's proof

Claim: Given matrix exponent

$$\Phi(t) = e^{(M+\kappa D)t}$$

such that

$$e^{Dt} \to \Pi, \quad t \to \infty$$

we have

$$e^{(M+\kappa D)t} \to e^{\Pi M t}\, \Pi, \quad \kappa \to \infty$$

Proof Laplace transform of $\Phi(t)$

$$\hat{\Phi}(s) = (sI - (M + \kappa D))^{-1}$$

First, we find that

$$\Pi = \lim_{t\to\infty} e^{Dt} = \lim_{s\downarrow 0} s(sI - D)^{-1} = \lim_{\kappa\to\infty} (I - \kappa D)^{-1}$$

Poor Man's proof

$$\hat{\Phi}(s) = (sI - (M + \kappa D))^{-1}$$
$$= [sI + (s-1)(I - \kappa D)^{-1}\kappa D - (I - \kappa D)^{-1}M]^{-1}(I - \kappa D)^{-1}$$

and since

$$(I - \kappa D)^{-1}\kappa D = (I - \kappa D)^{-1} - I$$

this converges as $\kappa \to \infty$ to

$$(s\Pi - \Pi + I - \Pi M)^{-1}\Pi$$

which is also equal to

$$(sI - \Pi M)^{-1}\Pi$$

since $\Pi = \Pi^2$

And which is also the Laplace transform of $e^{\Pi M t}\ \Pi$

Branching Process formulation

Type space of the individual is $\mathbb{Z}_+^2$
Individual of type $\gamma = (\alpha, \beta)$ has α **drivers** and β **passengers**.

With usual independence assumptions
At rate ν, it gains a driver and becomes $(\alpha + 1, \beta)$-type.
At rate μ it gains a passenger and becomes type $(\alpha, \beta + 1)$
At rate $x(\alpha, \beta)$, it dies and splits into two (α, β)-type progeny

$$x(\alpha, \beta) = (1 + s)^{\alpha}(1 - d)^{\beta},$$

Each of the progeny survives with probability p.
Neutrality condition identical as in the Moran formulation (does not involve p)

Expectations

We can obtain expected-count expressions using backward equations for pgf's and Laplace transformation

$$\widehat{M}_{\gamma_1,\gamma_2}(\pi) = \mathcal{L}\left(M_{\gamma_1,\gamma_2}(t)\right)$$

denoting $\widehat{\rho}_{\gamma_1} = \frac{1}{\pi-(a(\gamma_1)(2p-1)-\nu-\mu)}$,

$$\widehat{M}_{(\alpha_1,\beta_1),(\alpha_2,\beta_2)} = \nu^{\alpha_2-\alpha_1} \cdot \mu^{\beta_2-\beta_1} \cdot \sum_{\omega\in A} \widehat{\mathrm{P}}_\omega, \text{ for } \alpha_1 \le \alpha_2, \beta_1 \le \beta_2$$

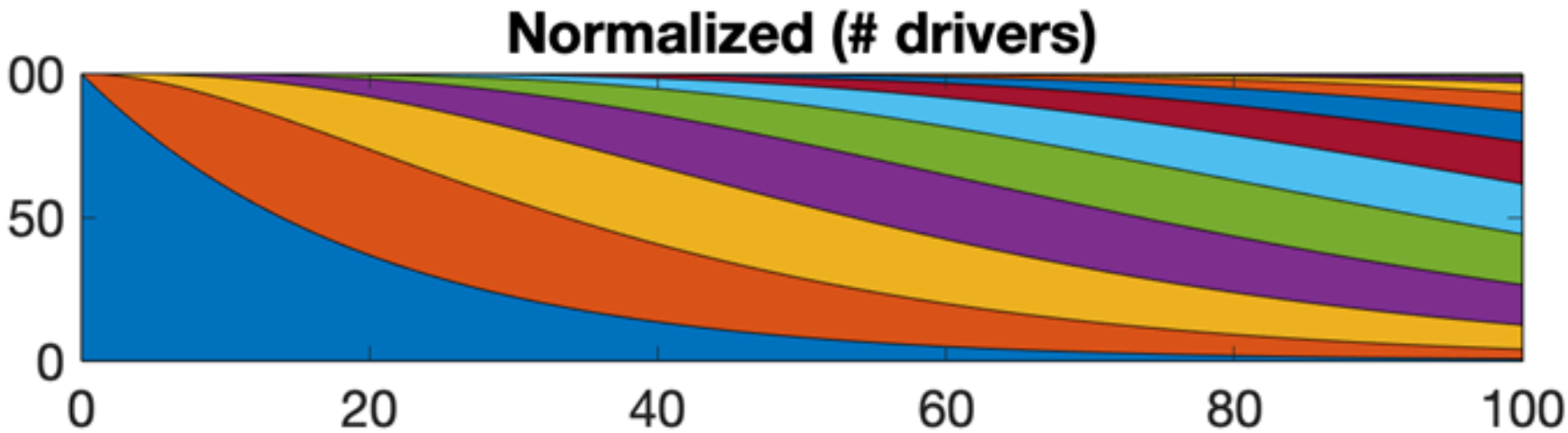
where $\widehat{\mathrm{P}}_\omega = \widehat{\rho}_{\omega_1} \cdot \widehat{\rho}_{\omega_2} \cdots \widehat{\rho}_{\omega_N}$
the set $\{\omega_1, \cdots, \omega_N\}$ is a path from γ_1 to γ_2, where each step is either one unit up or one unit to the right
$\begin{pmatrix} \alpha_2+\beta_2-\alpha_1-\beta_1 \\ \alpha_2-\alpha_1 \end{pmatrix}$ paths in set A
etc.

Expected counts do not reflect the clonal structure

Band structure

Each color is expected population percentage of cells with a given number of drivers

(blue = 0 drivers, orange = 1 driver, yellow = 2 drivers, . . .)



Although simulations show clonal structure similar as in the Moran formulation ...

Background

- A study by McFarland and co-authors in 2017 showed that passenger mutations are mildly deleterious and might inhibit or reverse tumour growth.
- In their experiment, authors discovered that passengers are capable to reduce proliferative fitness, slow tumour growth, and reduce metastasis progression.
- They also showed some other studies suggesting passenger mutations can trigger anticancer immune responses and the mechanism can be exploited to treat cancer.

Experiment in McFarland et al. (2017)

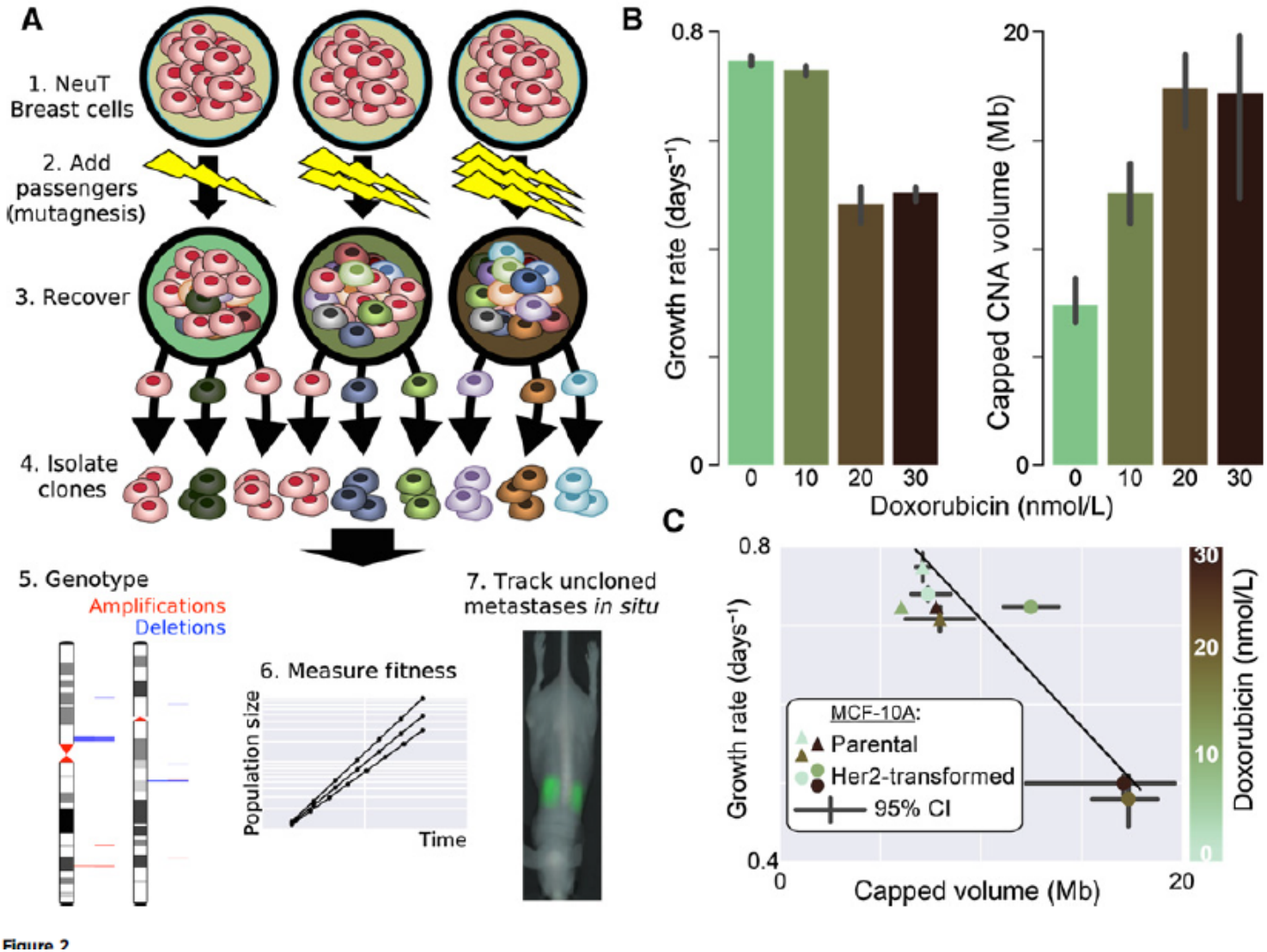
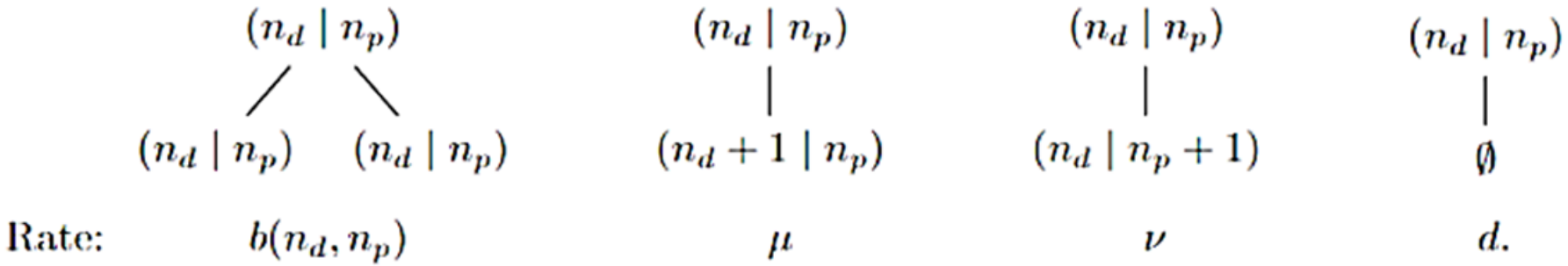


Figure 2

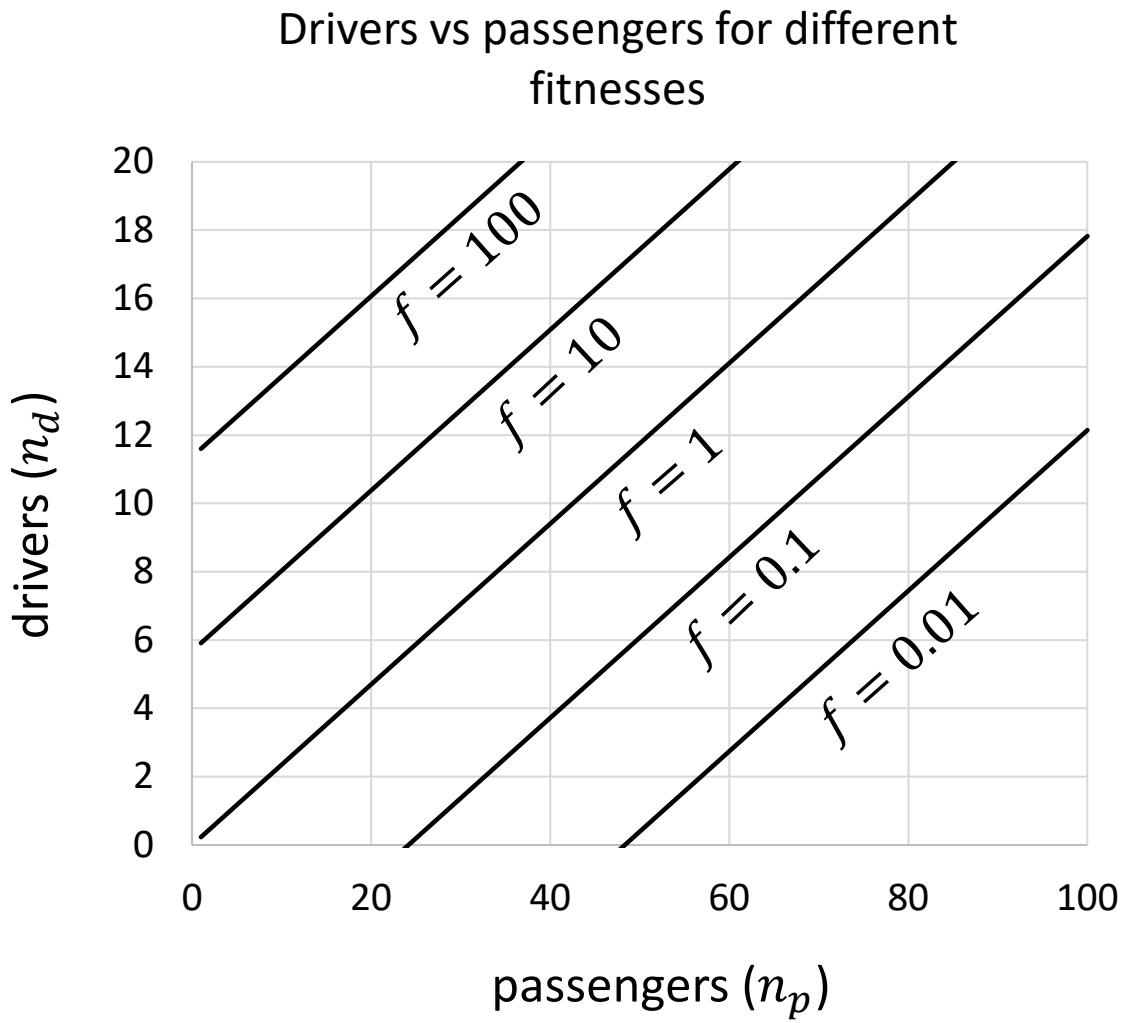
Branching Process model *Wang & Kimmel (2023) BMB*

- The cancer cell population is initiated by a single cell with birth rate (fitness) $b_0 > 1$ and the death rate is $d = 1$ for all cells.
- Cell types are categorized by number of drivers and passengers.
- A type (n_d, n_p) cell has fitness $b(n_d, n_p) = b_0 \frac{(1+s_d)^{n_d}}{(1+s_p)^{n_p}}$.
- $s_d > s_p$ are driver and passenger strengths; $\mu < \nu$ are mutation rates for driver and passengers.
- We also assume $\nu < d$ since mutations are rare events.
- This is a continuous-time Markov branching process with type space being $\mathbb{Z}_+^2$. Here, $\mathbb{Z}_+ = \{0, 1, 2, \cdots\}$.



Some graphics

Competition between rare advantageous driver mutations and frequent slightly deleterious passenger mutations



$$f = \frac{(1+s_d)^{n_d}}{(1+s_p)^{n_p}}$$

$s_d = 0.5, s_p = 0.1$

cell type = (s_p, s_d)

Reduced Process

- To reduce the dimension of the type space, We assume the effect of 1 driver mutation to the fitness of a cell is equivalent to that of L passenger mutations. That is, $\log_{(1+s_p)}(1+s_d) = L \in \mathbb{N}, L \geq 2$.
- This implies $b(n_d, n_p) = b_0(1+s_p)^{n_d L - n_p} = b_0(1+s_p)^i$ and type space is transformed into $\mathbb{Z}$.
- Therefore, the state space of the continuous process is $\mathbb{Z}^{\mathbb{Z}_+}$. By discarding states with zero probabilities, we may reduce the state space to a countable set that is still large enough to support the process.
- The process is non-explosive if we assume $d + \nu > \mu L$.

Embedded Process

- Before tackling the continuous process, we investigate the **embedded discrete branching (jump) process** with time unit being one generation.
- For a type $i = n_d L - n_p$ cell,

$$\mathbb{P}(\text{Divide}) = \frac{b(i)}{b(i)+\mu+\nu+d}, \mathbb{P}(\text{Dri. Mutation}) = \frac{\mu}{b(i)+\mu+\nu+d}$$
$$\mathbb{P}(\text{Pass. Mutation}) = \frac{\nu}{b(i)+\mu+\nu+d}, \mathbb{P}(\text{Die}) = \frac{d}{b(i)+\mu+\nu+d}.$$

- Therefore, the mean matrix for this process has the form

$$M_{i,i-1} = \frac{\nu}{\delta(i)}, M_{i,i} = 2\frac{b(i)}{\delta(i)}, M_{i,i+L} = \frac{\mu}{\delta(i)}, i \in \mathbb{Z}$$
$$\delta(i) = b(i)+\mu+\nu+d.$$

Extinction Probabilities

- For a countable-type branching process, there are two modes of extinction, global extinction and partial extinction.
- Partial extinction refers to extinction of each type and global extinction refers to extinction of the entire population.
- Note that the mean matrix M is irreducible and we show the convergence parameter R (inverse of spectral radius r) is $\frac{1}{2}$ by Theorem 6.1 in Seneta (1981).
- By Hautphenne et al. (2013), global extinction probabilities coincide with partial extinction probabilities and they are all strictly less than 1.
- To interpret this result, consider acquiring a passenger as death of a cell, then death rate becomes $\nu + d$; however, there is still a chance for a cell to acquire enough drivers so that the fitness $b(i)$ exceeds $\nu + d$ and this implies extinction is not an almost sure event.

Extinction Probabilities

- Let q_i denote the probability of (global) extinction if the population is initiated with a single type i cell.
- Probabilities of extinction q_i satisfy a difference equation. For all $i \in \mathbb{Z}$,

$$q_i = \frac{d}{\delta(i)} + \frac{\nu}{\delta(i)} q_{i-1} + \frac{b(i)}{\delta(i)} q_i^2 + \frac{\mu}{\delta(i)} q_{i+L}$$

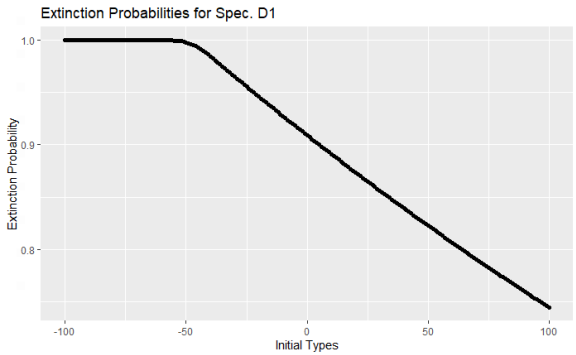
$$\Rightarrow 1 - q_i = \frac{\nu}{\delta(i)}(1 - q_{i-1}) + \frac{b(i)}{\delta(i)}(1 - q_i^2) + \frac{\mu}{\delta(i)}(1 - q_{i+L}).$$

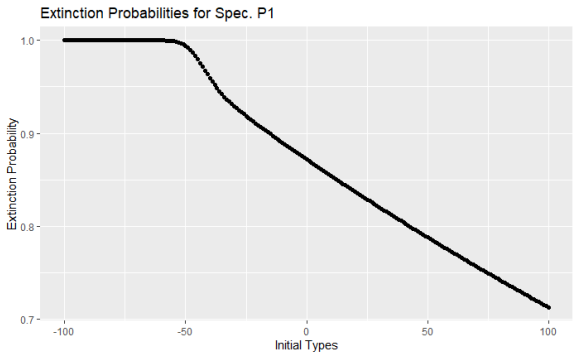
- An analytic solution to $(q_i)_{i=-\infty}^{\infty}$ is not feasible due to the quadratic term and inhomogeneity. By a coupling argument, we have $i > j$ implies $q_i \leq q_j$. Therefore, $\lim_{i\to\infty} q_i = 0$ and $\lim_{i\to-\infty} q_i = 1$.
- Hautphenne et al. (2013) Provide algorithms to compute parital/global extinction probabilities.

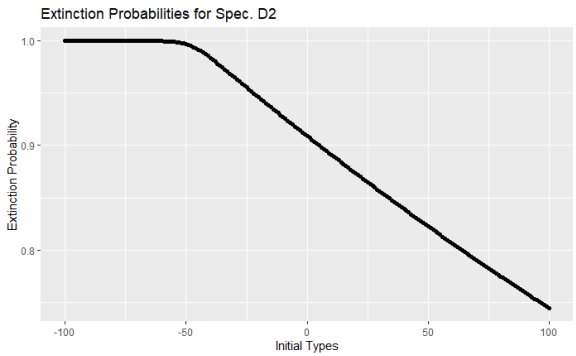
Parameter Specification

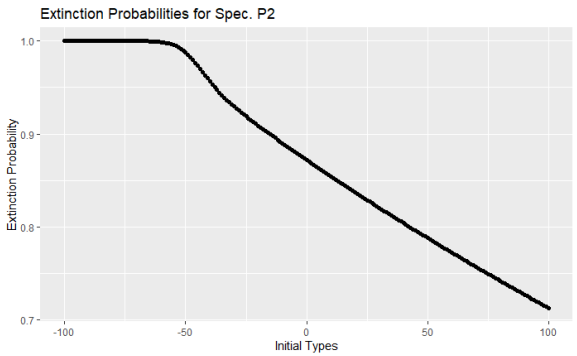
Parameter Specifications						
Parameters	b_0	μ	ν	s_p	L	$\frac{\nu}{\mu L}$
Specification $D1$	1.1	0.0251	0.05	0.002	2	≤ 1
Specification $D2$	1.1	0.011	0.05	0.002	5	≤ 1
Specification $D3$	1.1	0.0051	0.05	0.002	10	≤ 1
Specification $P1$	1.15	0.055	0.3	0.002	2	> 1
Specification $P2$	1.15	0.02	0.3	0.002	5	> 1
Specification $P3$	1.15	0.003	0.3	0.002	10	> 1

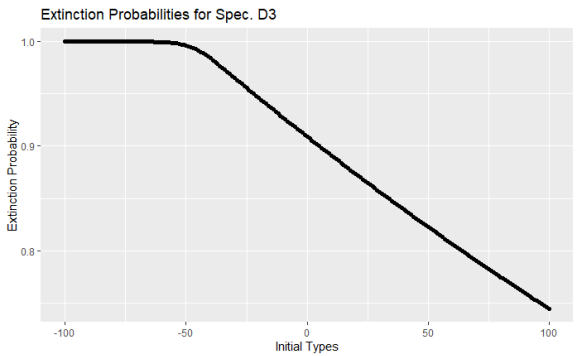
Computed Extinction Probabilities

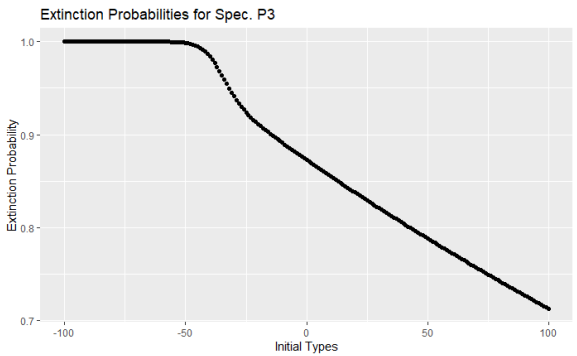












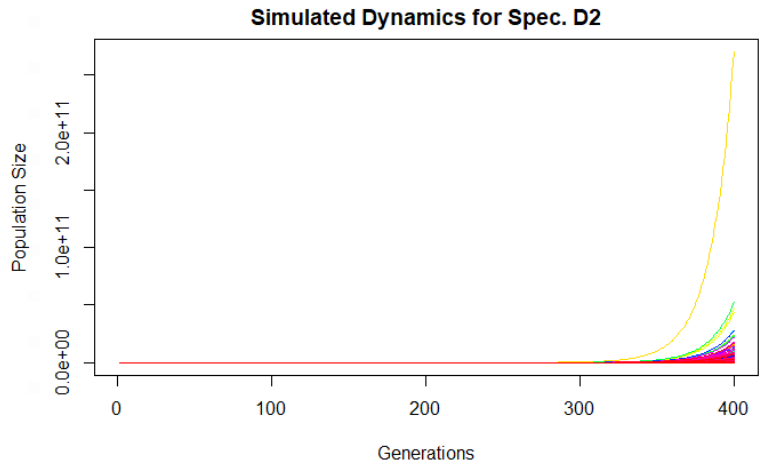
Connection between continuous-time and jump process

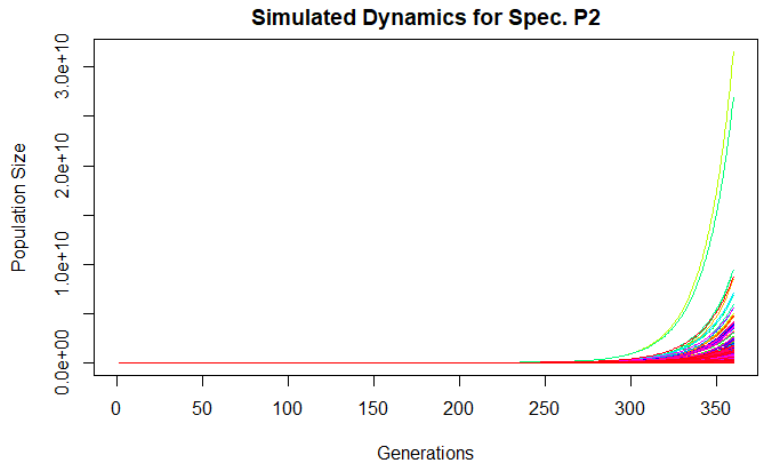
- Under the assumption of non-explosion ($\mu L < \nu + d$), the continuous process becomes extinct if and only if the embedded process becomes extinct.
- To see this, let A_n denote the event that the embedded process becomes extinct at or before the nth generation. That is, $\omega \in A_n$ implies $||\mathbf{E}_n(\omega)||_{\ell^1} = 0$. Since the number of generations n is finite, the number of jumps in the continuous process is also finite. Therefore, $\lim_{t\to\infty} ||\mathbf{Z}_t(\omega)||_{\ell^1} = 0$ for almost every $\omega \in A_n$.
- On the other hand, let B_t be the event that the continuous-time process becomes extinct at or before time $t \in \mathbb{Q}_+$. That is, $\omega \in B_n$ implies $||\mathbf{Z}_t(\omega)||_{\ell^1} = 0$. Since $[0, t]$ is a finite interval and the process is non-explosive, there are finitely many jumps in $[0, t]$ almost surely. Hence, the number of generations in $[0, t]$ is also finite a.s. and $\lim_{n\to\infty} ||\mathbf{E}_n(\omega)||_{\ell^1} = 0$ for almost every $\omega \in B_t$.

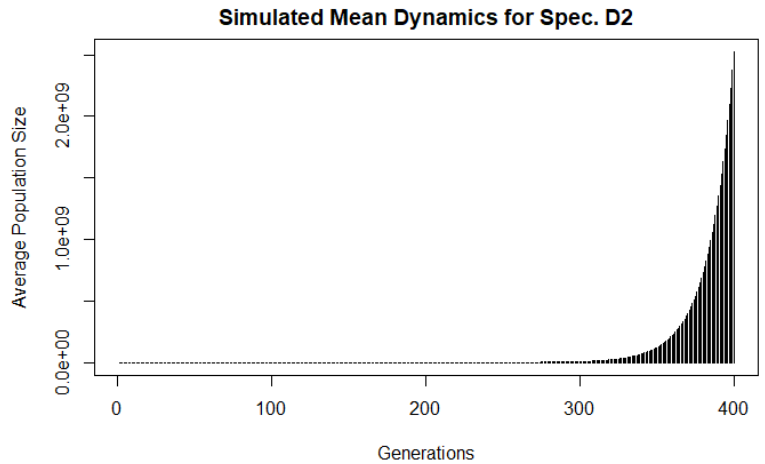
Mean Dynamics

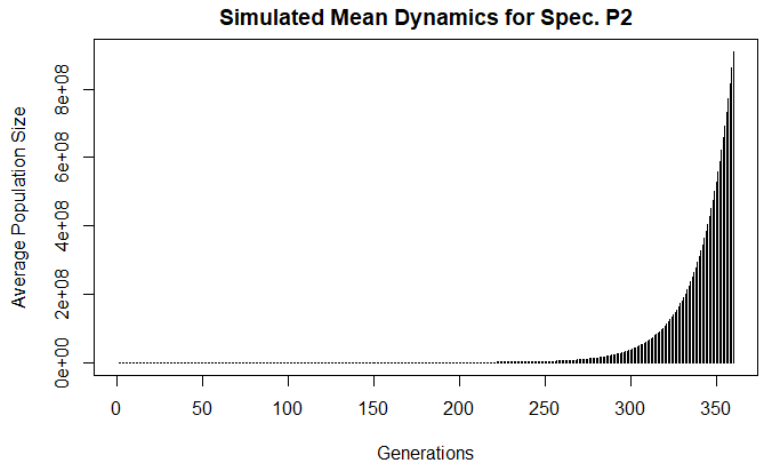
- The mean dynamics are unstable by exponential scaling and it can be shown that

$$(\frac{M}{c})^n \to \begin{cases} 0 \text{ elementwise if } c \geq 2 \\ \infty \text{ elementwise if } c < 2. \end{cases}$$









Type Transition

- So far, we do not observe any tug-of-war interaction between driver and passenger mutations on a population level. This motivates us to investigate transitions between types conditional on non-extinction.
- Focus on transitions between types along a surviving lineage (backbone of the process) and denote the process as (X_n), the transition probabilities are

$$T_{i,i-1} = \frac{\nu}{\delta(i)} \frac{1-q_{i-1}}{1-q_i}, T_{i,i} = \frac{b(i)}{\delta(i)} \frac{1-q_i^2}{1-q_i}, T_{i,i+L} = \frac{\mu}{\delta(i)} \frac{1-q_{i+L}}{1-q_i}.$$

- The type transition process can be made continuous in time and the jump chain $(\tilde{X}_n)$ has transition probability matrix $\tilde{T}$ such that

$$\tilde{T}_{i,i-1} = \frac{\nu}{\delta(i)(1-T_{i,i})} \frac{1-q_{i-1}}{1-q_i}$$
$$\tilde{T}_{i,i+L} = \frac{\mu}{\delta(i)(1-T_{i,i})} \frac{1-q_{i+L}}{1-q_i}.$$

- The holding time for state i follows an exponential distribution with rate $\delta(i)(1 - T_{i,i})$ and the continuous type-transition process is non-explosive since

$$\lim_{i\to\infty} \delta(i)(1 - T_{i,i}) = \mu + \nu, \lim_{i\to-\infty} \delta(i)(1 - T_{i,i}) = \mu + \nu + d$$
$$\Rightarrow \sup_{i\in\mathbb{Z}}\{\delta(i)(1 - T_{i,i})\} < \infty.$$

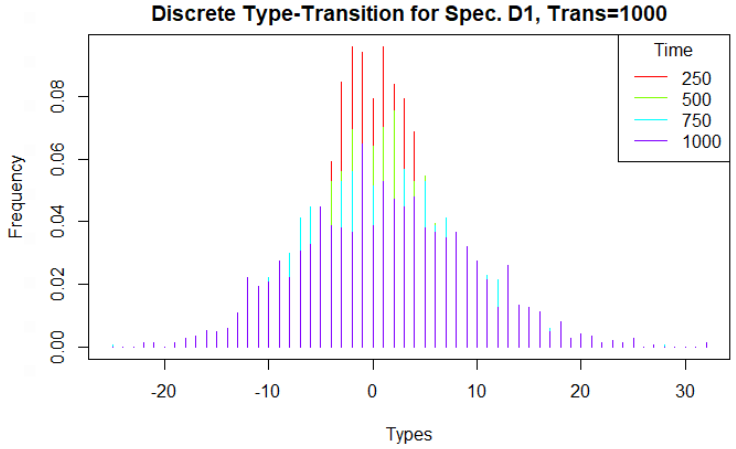
- When driver dominates ($\frac{\nu}{\mu L} \leq 1$), both (X_n) and $(\tilde{X}_n)$ are transient.
- When passenger dominates ($\frac{\nu}{\mu L} > 1$), the continuous type-transition admits a limiting distribution. Moreover, when $\frac{\nu}{\mu L} > (1 + s_p)^L = 1 + s_d$, (X_n) admits a limiting distribution as well.

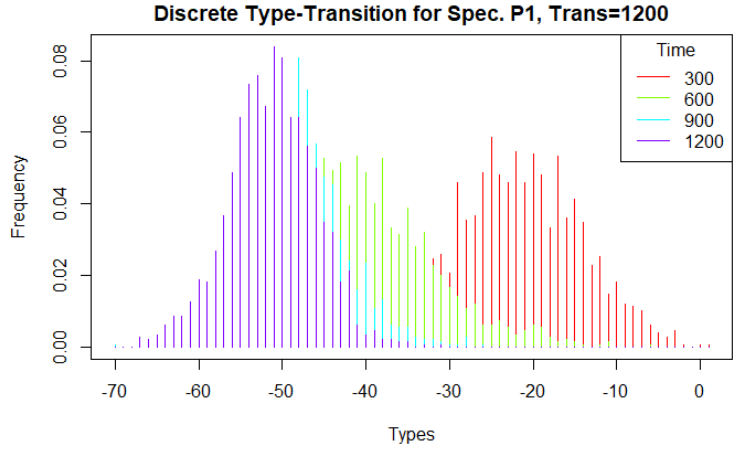
Driver Dominates

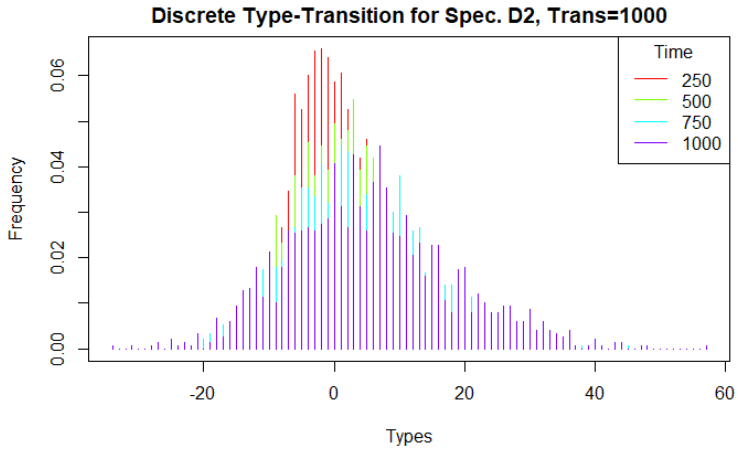
- $\mathbb{E}(X_{n+1} \mid X_n) = X_n + \frac{\mu L(1-q_{X_n+L})-\nu(1-q_{X_n-1})}{\delta(X_n)(1-q_{X_n})}$.

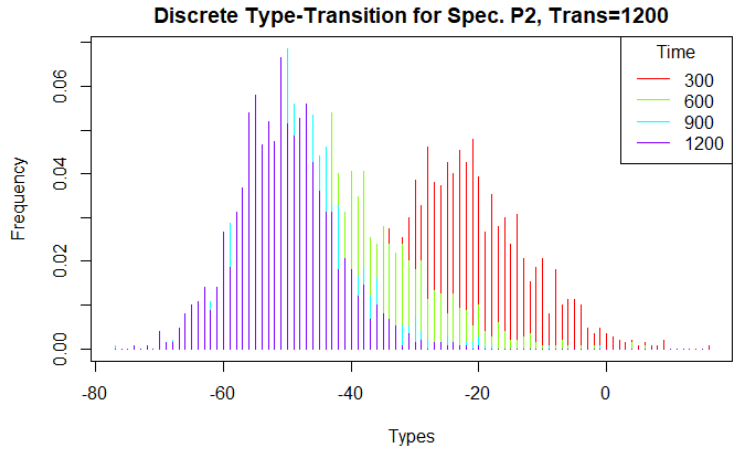
$$\begin{aligned} & \mu L(1-q_{i+L}) - \nu(1-q_{i-1}) > 0, \forall i \in \mathbb{Z} \\ \iff & \frac{\nu}{\mu L} < \frac{1-q_{i+L}}{1-q_{i-1}}, \forall i \in \mathbb{Z} \\ \iff & \frac{\nu}{\mu L} \leq 1. \end{aligned}$$

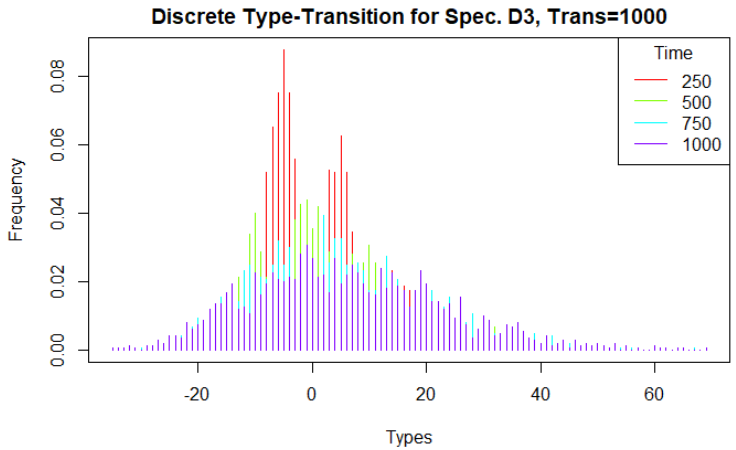
- When $\frac{\nu}{\mu L} \leq 1$, (X_n) is a submartingale and $\lim_{n\to\infty} \mathbb{E}(X_n) = \infty$.
- By Jensen's inequality, the average fitness $\mathbb{E}(b(X_n)) = \infty$ as well.
- This means tumour is growing more and more rapidly when driver dominates.
- Since both discrete and continuous type transition processes are transient, they do not admit limiting distributions.

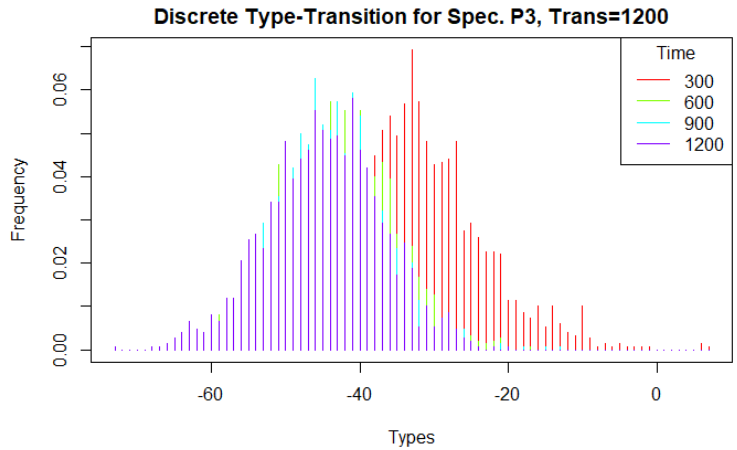

Discrete Type-Transition for Spec. D1, Trans=1000
Time
250
500
750
1000
Frequency
0.00
0.02
0.04
0.06
0.08
-20
-10
0
10
20
30
Types

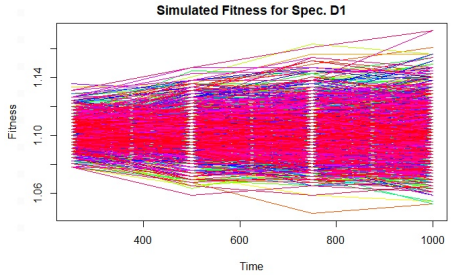

Discrete Type-Transition for Spec. P1, Trans=1200
Time
300
600
900
1200
Frequency
0.00
0.02
0.04
0.06
0.08
-70
-60
-50
-40
-30
-20
-10
0
Types

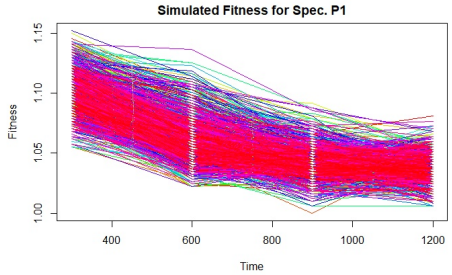

Discrete Type-Transition for Spec. D2, Trans=1000
Time
250
500
750
1000
Frequency
0.00
0.02
0.04
0.06
-20
0
20
40
60
Types

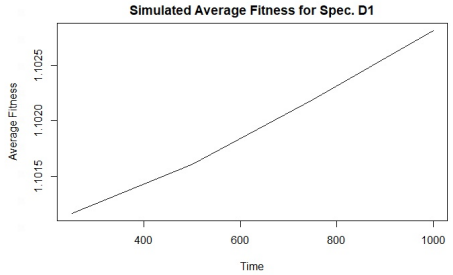

Discrete Type-Transition for Spec. P2, Trans=1200
Time
300
600
900
1200
Frequency
0.00
0.02
0.04
0.06
-80
-60
-40
-20
0
Types

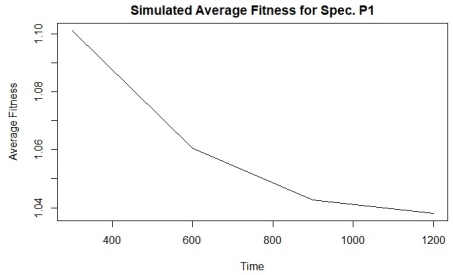

Discrete Type-Transition for Spec. D3, Trans=1000
Time
250
500
750
1000
Frequency
0.00
0.02
0.04
0.06
0.08
-20
0
20
40
60
Types

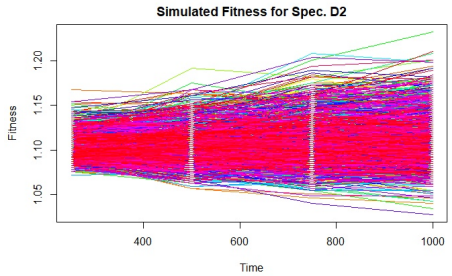

Discrete Type-Transition for Spec. P3, Trans=1200
Time
300
600
900
1200
Frequency
0.00
0.02
0.04
0.06
-60
-40
-20
0
Types

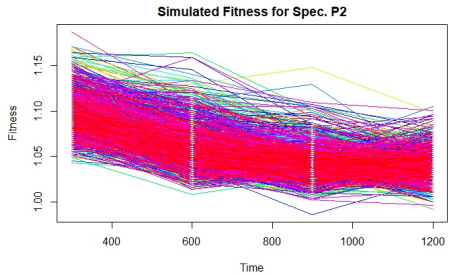

Simulated Fitness for Spec. D1
Fitness
Time

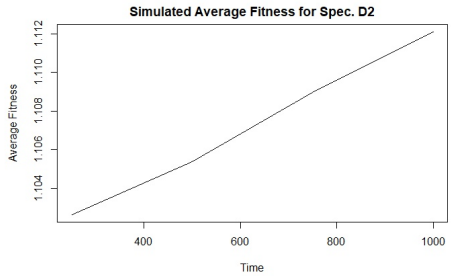

Simulated Fitness for Spec. P1
Fitness
Time

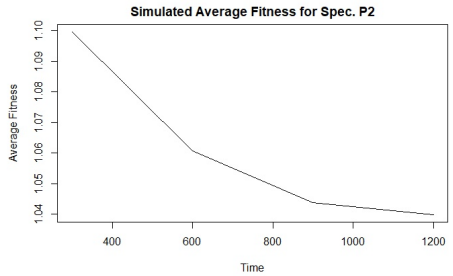

Simulated Average Fitness for Spec. D1
Average Fitness
Time

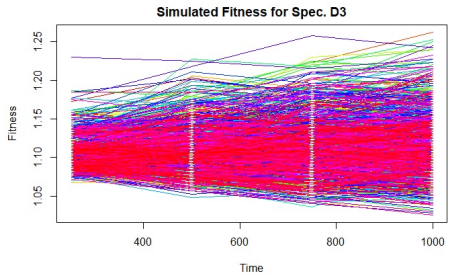

Simulated Average Fitness for Spec. P1
Average Fitness
Time

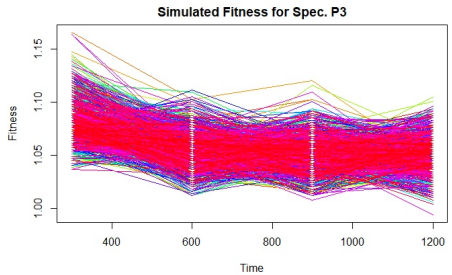

Simulated Fitness for Spec. D2
Fitness
Time

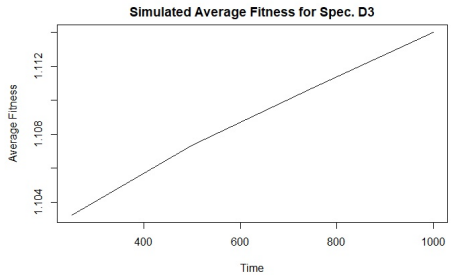

Simulated Fitness for Spec. P2
Fitness
Time

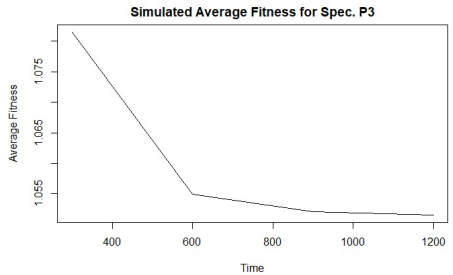

Simulated Average Fitness for Spec. D2
Average Fitness
Time


Simulated Average Fitness for Spec. P2
Average Fitness
Time


Simulated Fitness for Spec. D3
Fitness
Time


Simulated Fitness for Spec. P3
Fitness
Time


Simulated Average Fitness for Spec. D3
Average Fitness
Time


Simulated Average Fitness for Spec. P3
Average Fitness
Time

A Sufficient Condition for Transience

- We show the transience by the fact that a stochastic matrix is transient if and only if there exists a subinvariant measure that is not invariant.
- It turned out that a sufficient condition for the existence of the above measure is $q_i \leq \frac{d}{b(i)}$ for all $i \in \mathbb{Z}$ and it has a probabilistic interpretation.
- The extinction probability of a cell population initiated by a type i cell with mutations excluded from the model ($\mu = \nu = 0$) is $\min(1, \frac{d}{b(i)})$.
- Therefore, intuitively speaking, if driver dominates and fitnesses of cells are expected to increase, the extinction probability q_i should be upper bounded by $\min(1, \frac{d}{b(i)})$.

Passenger Dominates

- Recall that

$$1 - q_i = \frac{\nu}{\delta(i)}(1 - q_{i-1}) + \frac{b(i)}{\delta(i)}(1 - q_i^2) + \frac{\mu}{\delta(i)}(1 - q_{i+L})$$

$$\Rightarrow \frac{1 - q_i}{1 - q_{i-1}} = \frac{\nu}{\delta(i)} + \frac{b(i)}{\delta(i)}\frac{1 - q_i^2}{1 - q_{i-1}} + \frac{\mu}{\delta(i)}\frac{1 - q_{i+L}}{1 - q_{i-1}}$$

$$\Rightarrow \alpha = \frac{\nu}{\mu + \nu + d} + \frac{\mu}{\mu + \nu + d}\alpha^{L+1}, \alpha = \lim_{i \to -\infty} \frac{1 - q_i}{1 - q_{i-1}}.$$

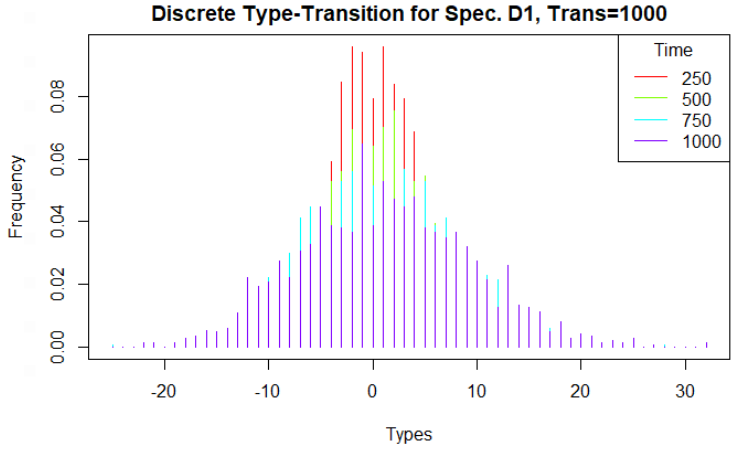
- The existence of the limit follows Levinson's fundamental theorem (discrete version) and one can show $\alpha > 1$.

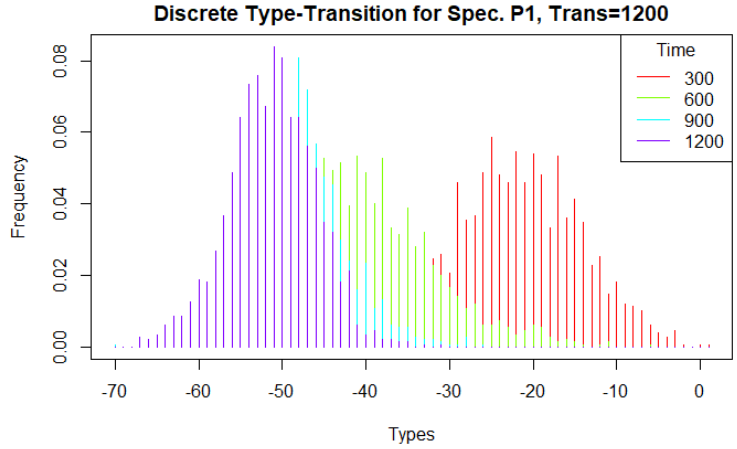
- If (X_n) is a supermartingale, $\frac{\nu}{\mu L} \geq \frac{1-q_{i+L}}{1-q_{i-1}}$ for all $i \in \mathbb{Z}$.
- Take $\lim_{i \to -\infty}$ on both sides, we obtain

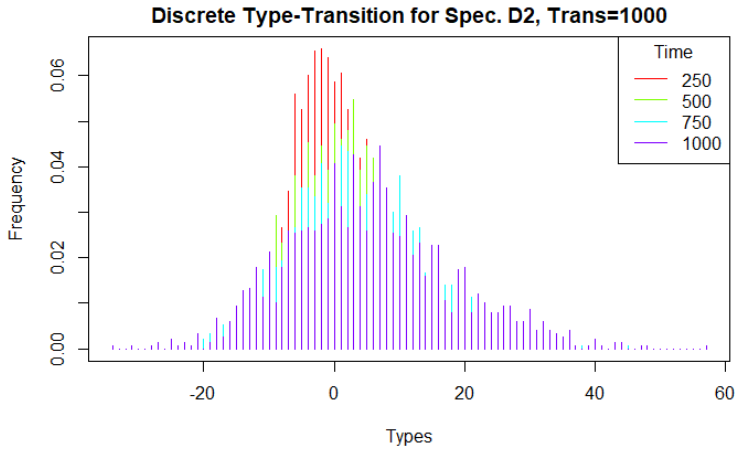
$$\frac{\nu}{\mu L} \geq \alpha^{L+1} = \frac{(\mu + \nu + d)\alpha - \nu}{\mu}$$

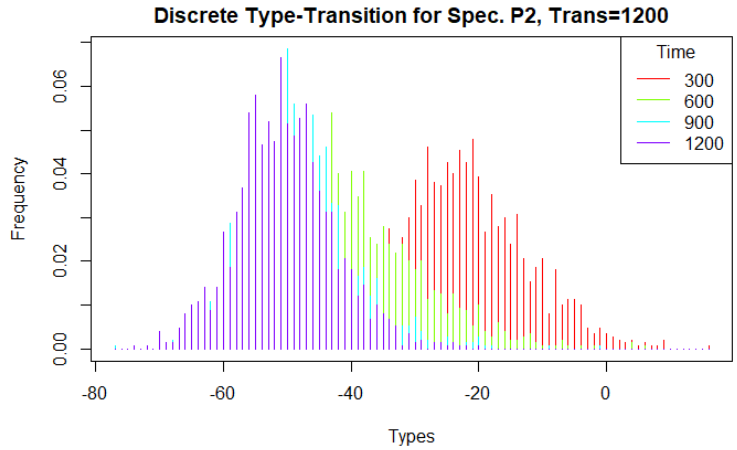
$$\Rightarrow \alpha \leq \frac{L+1}{L} \frac{\nu}{\mu + \nu + d} \leq \frac{3}{2} \frac{\nu}{\nu + d} < 1.$$

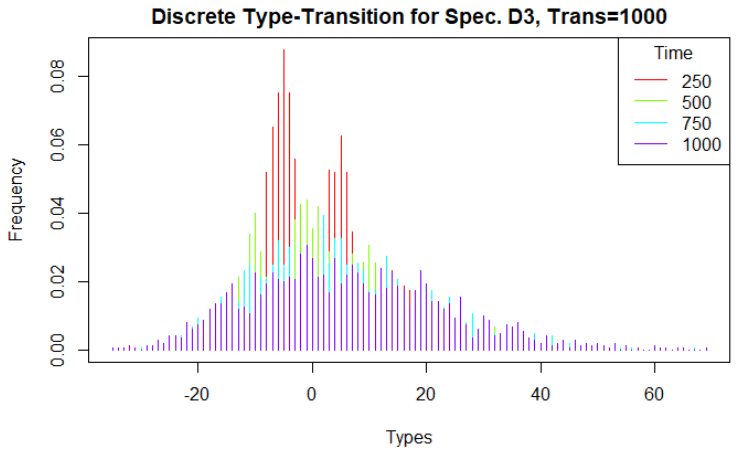
- Therefore, (X_n) can never be a supermartingale and there exists an index I such that $\mathbb{E}(X_{n+1} \mid X_n = i) > i$ for all $i < I$. This corresponds to the region where extinction probabilities change abruptly.
- Since the backbone is the process conditional on non-extinction, it can be viewed as the population filtered by natural selection.
- Natural selection tends to purge deleterious mutations.

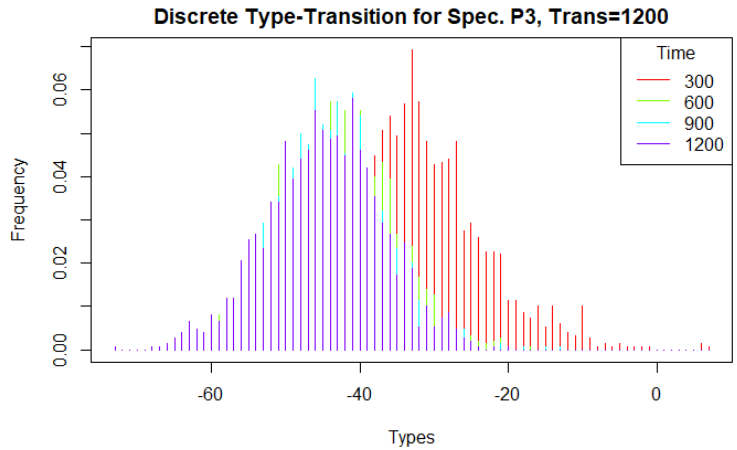

Discrete Type-Transition for Spec. D1, Trans=1000
Time
250
500
750
1000
Frequency
0.00
0.02
0.04
0.06
0.08
-20
-10
0
10
20
30
Types

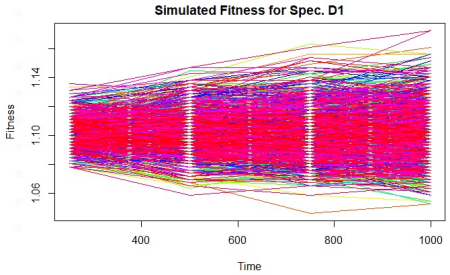

Discrete Type-Transition for Spec. P1, Trans=1200
Time
300
600
900
1200
Frequency
0.00
0.02
0.04
0.06
0.08
-70
-60
-50
-40
-30
-20
-10
0
Types

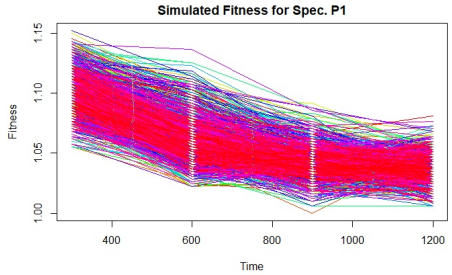

Discrete Type-Transition for Spec. D2, Trans=1000
Time
250
500
750
1000
Frequency
0.00
0.02
0.04
0.06
-20
0
20
40
60
Types

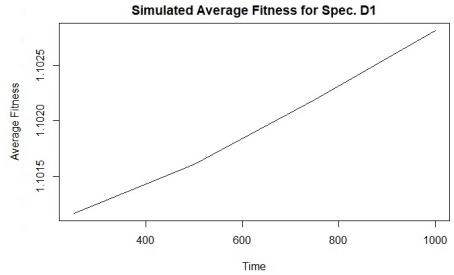

Discrete Type-Transition for Spec. P2, Trans=1200
Time
300
600
900
1200
Frequency
0.00
0.02
0.04
0.06
-80
-60
-40
-20
0
Types

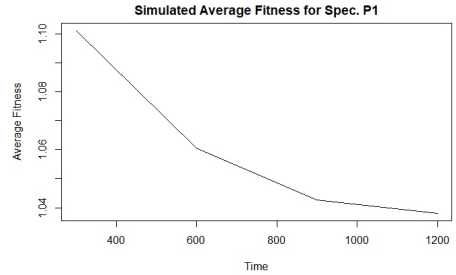

Discrete Type-Transition for Spec. D3, Trans=1000
Time
250
500
750
1000
Frequency
0.00
0.02
0.04
0.06
0.08
-20
0
20
40
60
Types

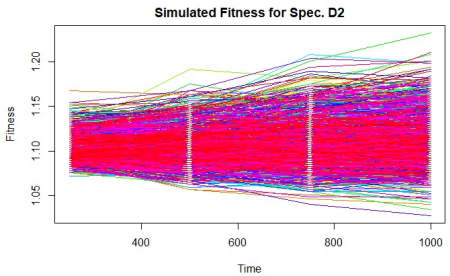

Discrete Type-Transition for Spec. P3, Trans=1200
Time
300
600
900
1200
Frequency
0.00
0.02
0.04
0.06
-60
-40
-20
0
Types

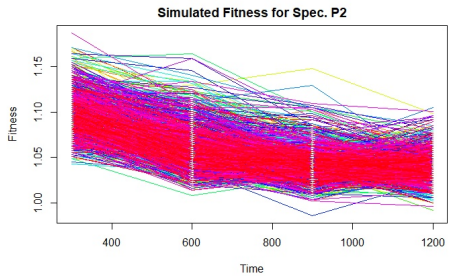

Simulated Fitness for Spec. D1
Fitness
Time

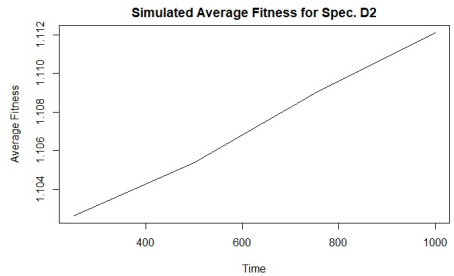

Simulated Fitness for Spec. P1
Fitness
Time

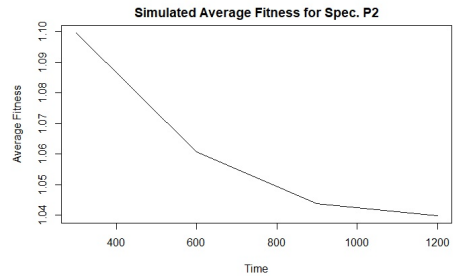

Simulated Average Fitness for Spec. D1
Average Fitness
Time

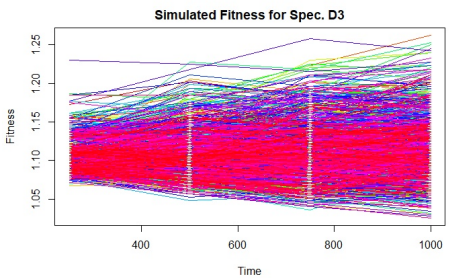

Simulated Average Fitness for Spec. P1
Average Fitness
Time

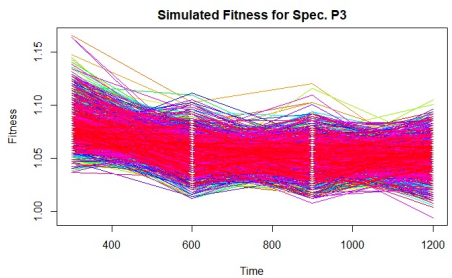

Simulated Fitness for Spec. D2
Fitness
Time

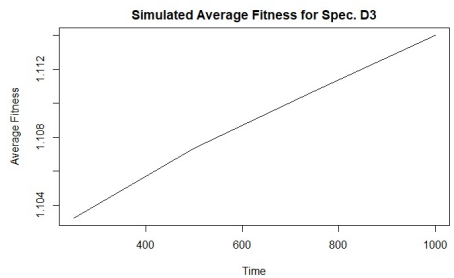

Simulated Fitness for Spec. P2
Fitness
Time

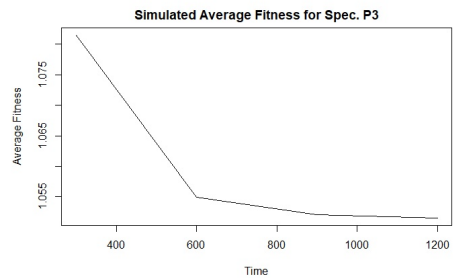

Simulated Average Fitness for Spec. D2
Average Fitness
Time

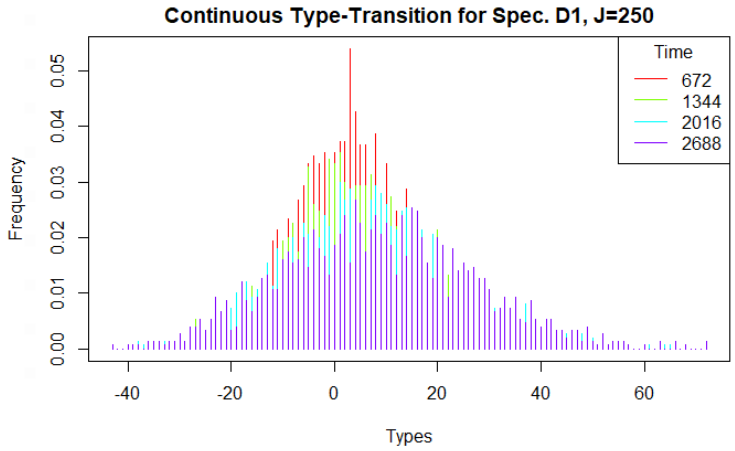

Simulated Average Fitness for Spec. P2
Average Fitness
Time

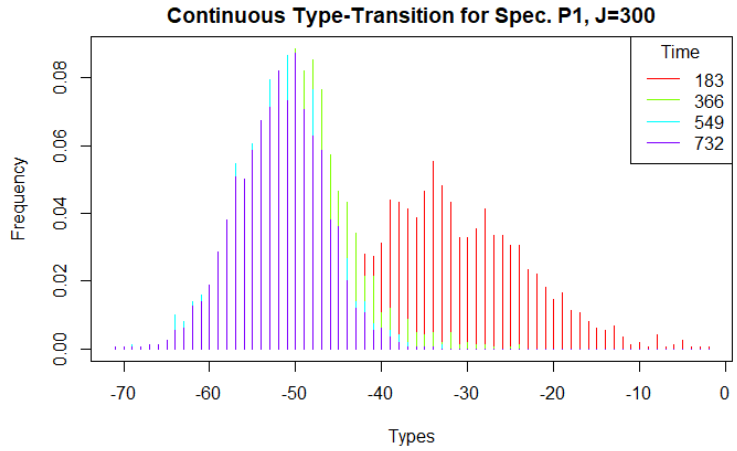

Simulated Fitness for Spec. D3
Fitness
Time

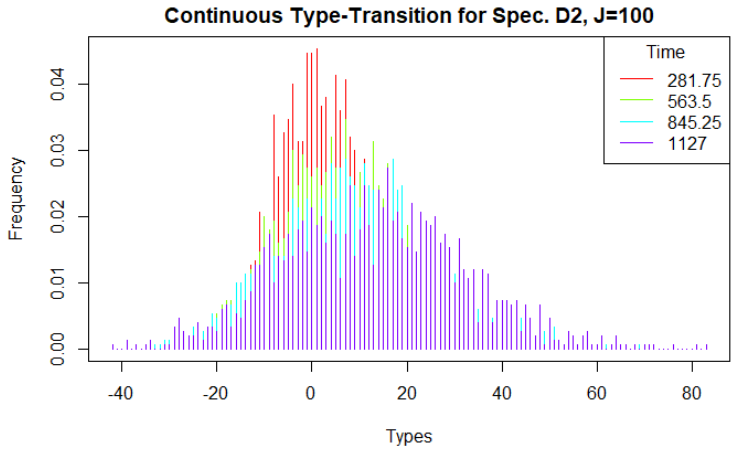

Simulated Fitness for Spec. P3
Fitness
Time

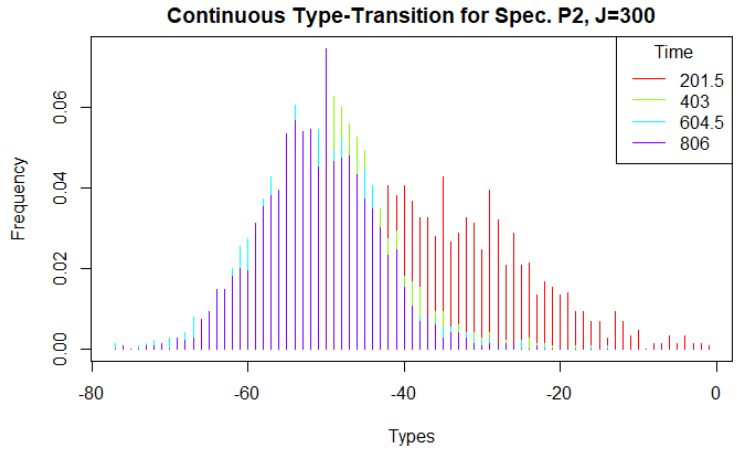

Simulated Average Fitness for Spec. D3
Average Fitness
Time

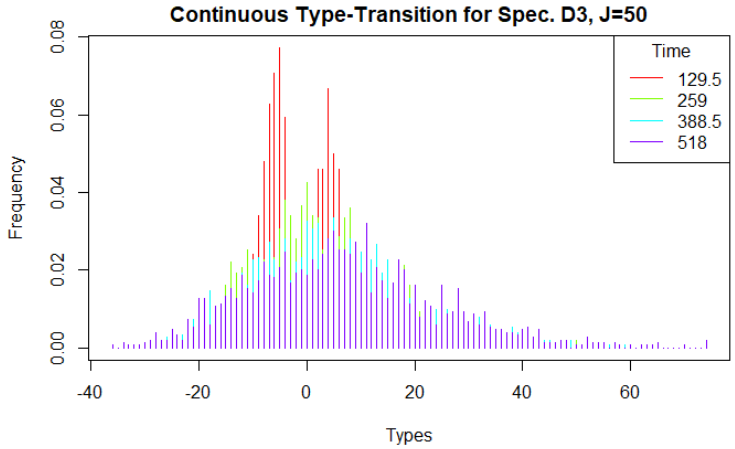

Simulated Average Fitness for Spec. P3
Average Fitness
Time

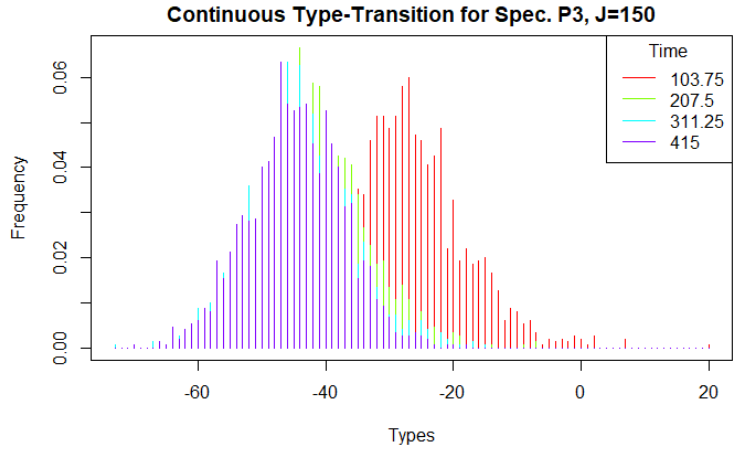

Continuous Type-Transition for Spec. D1, J=250
Time
672
1344
2016
2688
Frequency
0.00
0.01
0.02
0.03
0.04
0.05
-40
-20
0
20
40
60
Types


Continuous Type-Transition for Spec. P1, J=300
Time
183
366
549
732
Frequency
0.00
0.02
0.04
0.06
0.08
-70
-60
-50
-40
-30
-20
-10
0
Types

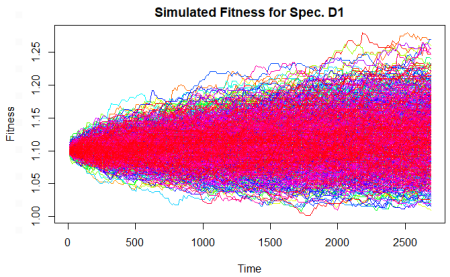

Continuous Type-Transition for Spec. D2, J=100
Time
281.75
563.5
845.25
1127
Frequency
0.00
0.01
0.02
0.03
0.04
-40
-20
0
20
40
60
80
Types

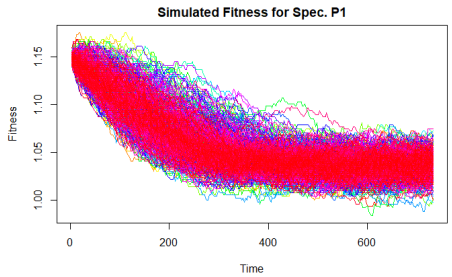

Continuous Type-Transition for Spec. P2, J=300
Time
201.5
403
604.5
806
Frequency
0.00
0.02
0.04
0.06
-80
-60
-40
-20
0
Types

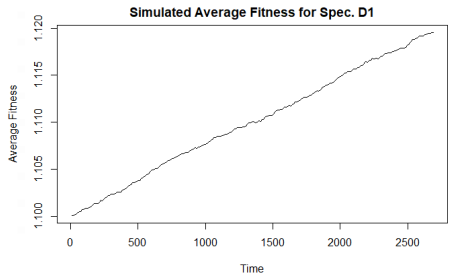

Continuous Type-Transition for Spec. D3, J=50
Time
129.5
259
388.5
518
Frequency
0.00
0.02
0.04
0.06
0.08
-40
-20
0
20
40
60
Types

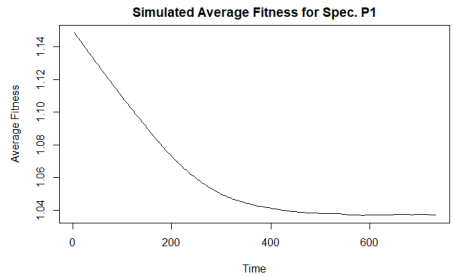

Continuous Type-Transition for Spec. P3, J=150
Time
103.75
207.5
311.25
415
Frequency
0.00
0.02
0.04
0.06
-60
-40
-20
0
20
Types

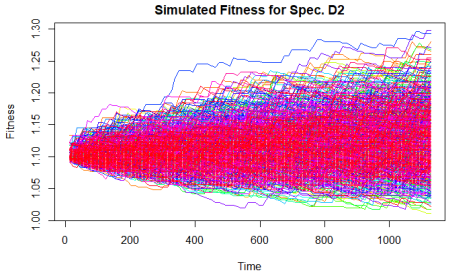
Continuous Type-Transition

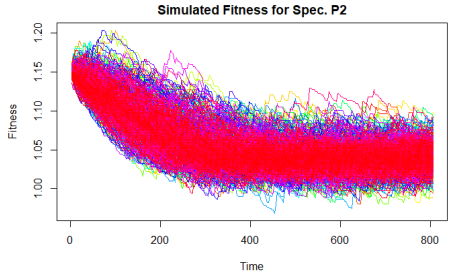


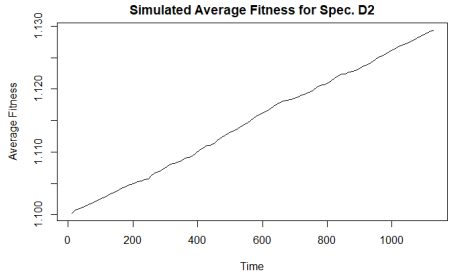


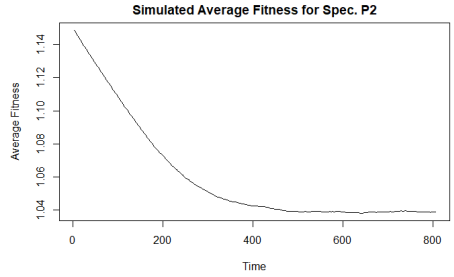


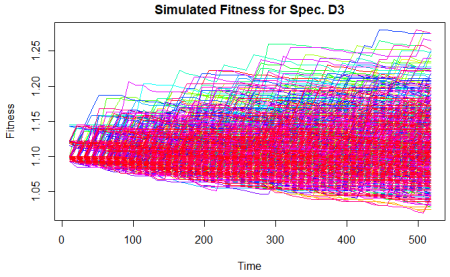


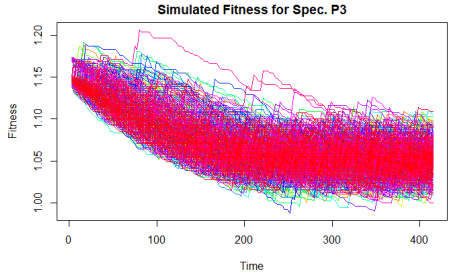


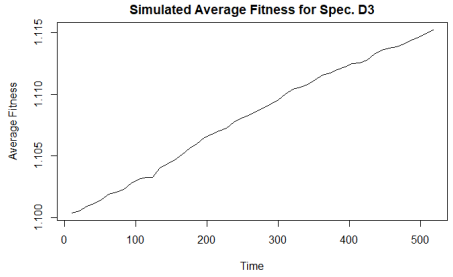


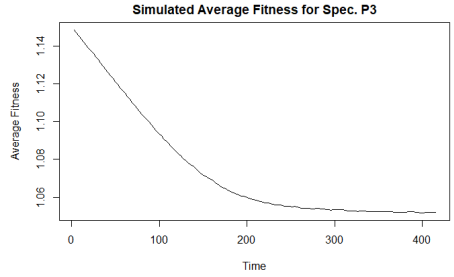












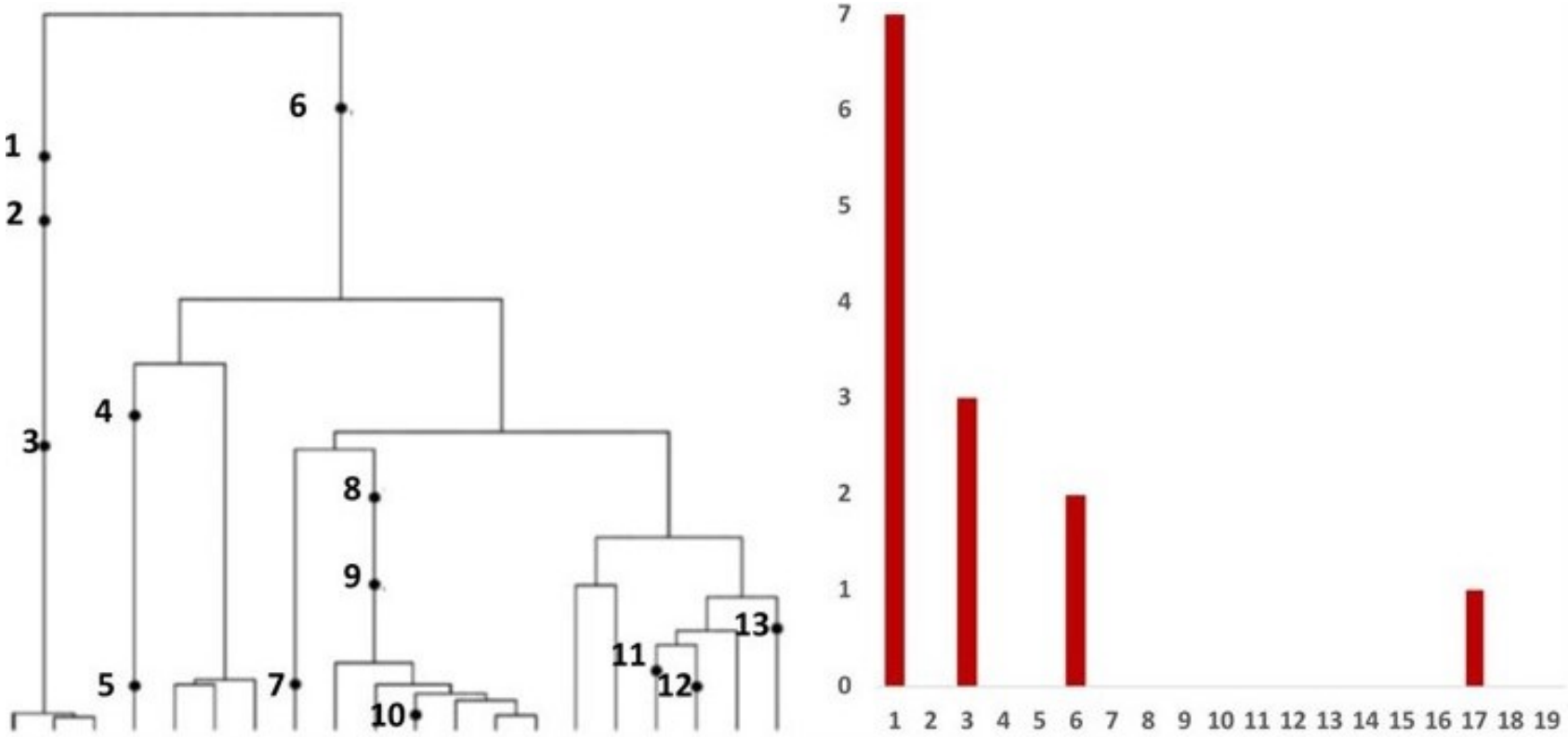
Some Remarks on the Proof

- The proof of positive recurrence of the type transition is long but it is easy to summarize.
- If is a proof by contradiction. If the jump chain for the continuous type transition is not positive recurrent, $\lim_{n\to\infty} \mathbb{P}(|\tilde{X}_n| > x) = 1$ for any fixed x.
- For x large enough, we only care about transition between types that are very large or very small.
- We approximate both cases by random walks and invoke a result in Spitzer (2001). The result states that a left-continuous random walk with a upper reflecting barrier is positive recurrent if the expected increment is positive.

Bibliography

- Bozic, I. et al. (2010). Accumulation of driver and passenger mutations during tumor progression. Proceedings of the National Academy of Sciences, 107(43), 18545-18550.
- Bozic, I., Gerold, J. M., & Nowak, M. A. (2016). Quantifying clonal and subclonal passenger mutations in cancer evolution. PLoS computational biology, 12(2), e1004731.
- Bodine, S., & Lutz, D. A. (2015). Asymptotic integration of differential and difference equations (Vol. 2129). New York: Springer.
- Durrett, R. (2015). Branching process models of cancer. In Branching process models of cancer (pp. 1-63). Springer, Cham.
- Hautphenne, S. et al. (2013). Extinction probabilities of branching processes with countably infinitely many types. Advances in Applied Probability, 45.4, 1068.
- McFarland, C. et al. (2017). The Damaging Effect of Passenger Mutations on Cancer Progression. Cancer research, 2017, 77.18: 4763-4772.
- Seneta, E. (1981). Nonnegative Matrices and Markov Chains, 2nd edn. Springer, NewYork.

Applications: Site frequency spectrum



Left: Genealogy of a sample of $n = 20$ individuals includes 13 mutational events, denoted by black dots. Mutations 4, 5, 7, 10, 11, 12, and 13 (total of 7 mutations) are present in a single individual, mutations 1, 2, and 3 (total of 3 mutations) are present in three individuals, mutations 8 and 9 (2 mutations) are present in six individuals, and mutation 6 (1 mutation) is present in 17 individuals.

Right: The observed site frequency spectrum, $S_{20}(1) = 7$, $S_{20}(3) = 3$, $S_{20}(6) = 2$, and $S_{20}(17) = 1$, other $S_n(k)$ equal to 0.

Kurpas and Kimmel (2022) Frontiers in ecology and evolution
Dinh et al (2024) E-life (Under revision)

Site Frequency Spectrum

Site Frequency Spectrum (SFS) is a summary statistics for mutations from sequence data:

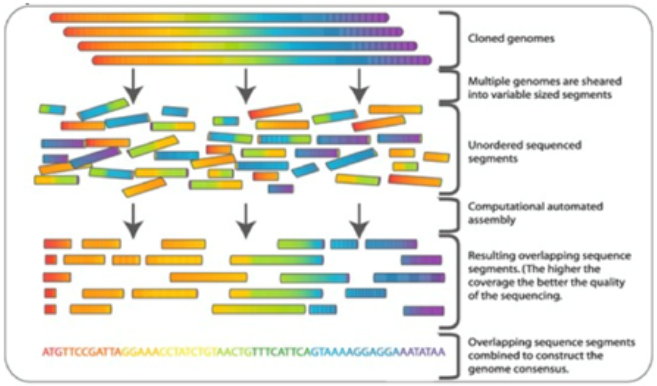
- k_i = count of cells mutated at site i
- $S_n(k)$ = count of mutations s.t. $k_i = k$
- $S_n(1)$ = singleton count

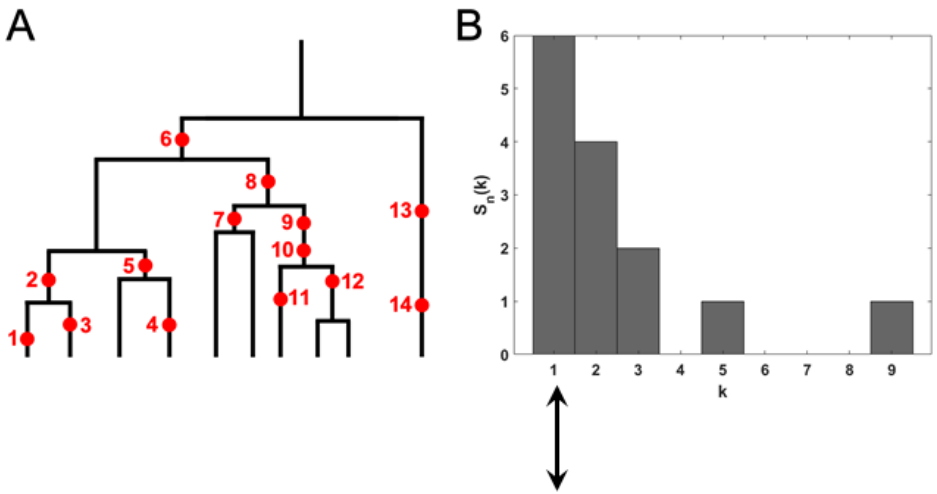
k_i is usually not known → approximated by Variant Allele Fraction (VAF):

$$k_i = \frac{a_i}{a_i + r_i} \in [0,1]$$

where a_i = alternate read count at site i
r_i = reference read count at site i

A: cell phylogeny with 14 mutations
B: corresponding SFS

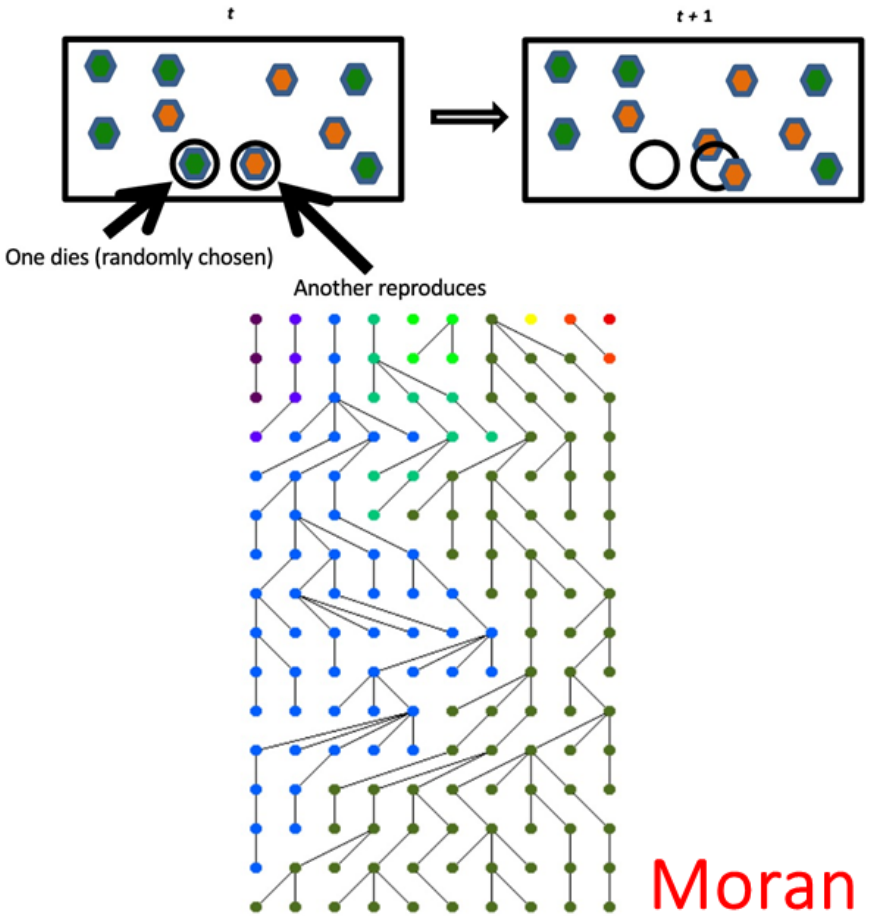


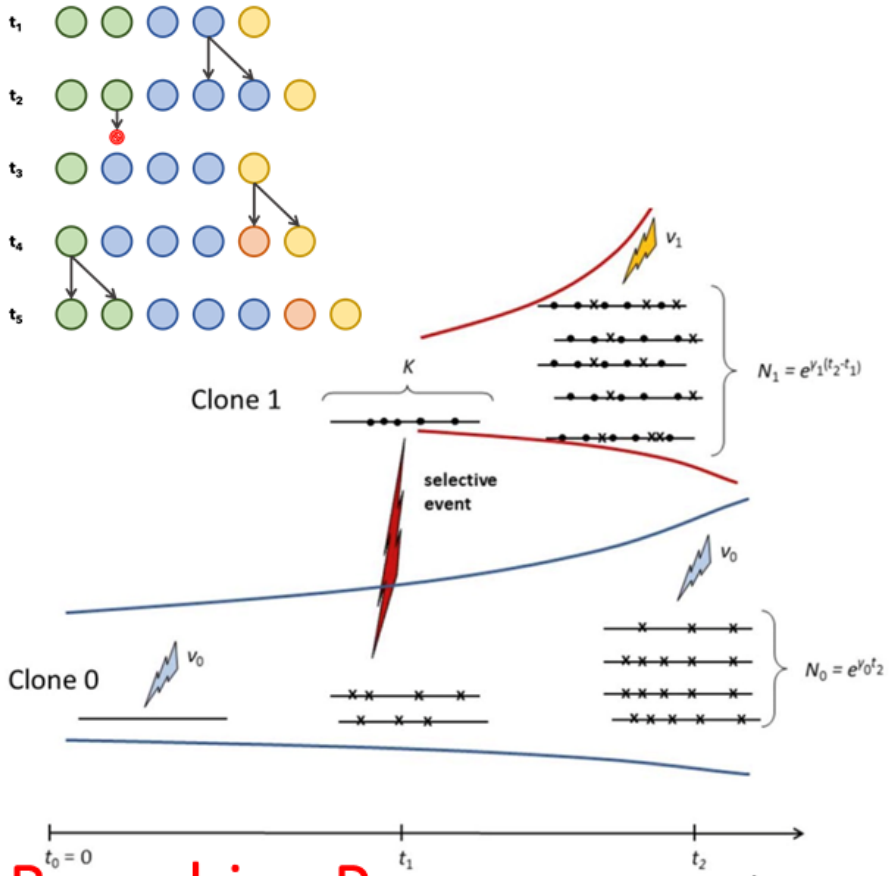


$S_n(1) = 6 \Leftrightarrow$ 6 mutations appear in 1 cell

Dinh KN, Jaksik R, Kimmel M, Lambert A, Tavaré S. Statist Sci 35(1)129-144

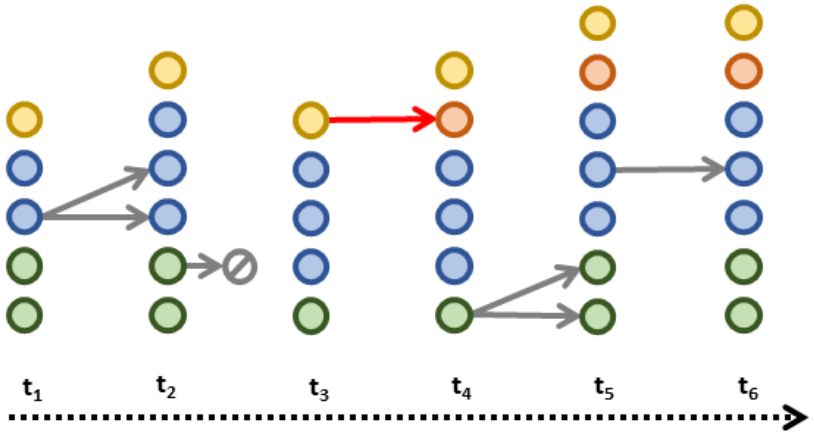
Mathematical models for cancer evolution





Moran vs Branching Process

Branching process



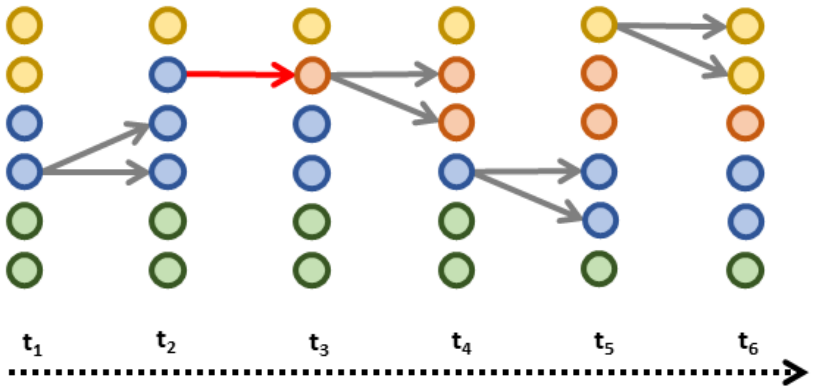
= clones (based on mutation profiles)
= division event
= mutation event

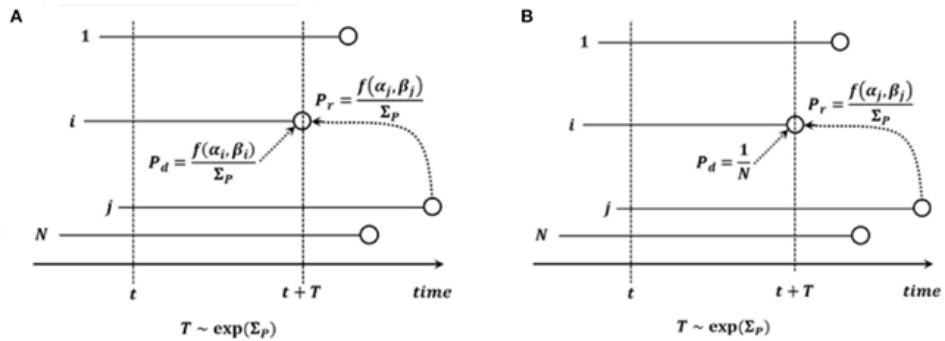
s = driver's selective advantage
d = passenger's deleterious disadvantage
μ = mutation rate
p = probability of driver mutation

- Cell with α drivers, β passengers:
  - time to next mutation $\sim$ $\mathbf{Exponential}(\mu)$
    - mutation = driver with probability p, passenger with probability $1-p$
  - time to next division $\sim$ $\mathbf{Exponential}(f(\alpha,\beta))$
    - fitness $f(\alpha,\beta) = (1+s)^{\alpha}(1-d)^{\beta}$
    - daughter count $\sim$ $\mathbf{Binomial}(n=2, p=0.5)$

6

Moran process





s = driver's selective advantage
d = passenger's deleterious disadvantage
μ = mutation rate
p = probability of driver mutation

- Mutations: same as Branching Process
- Time to next replacement $\sim \mathbf{Exponential}(\sum f) \rightarrow$ one cell replaces another cell
  - Cell to replace: $\sim f(\alpha, \beta)$
  - Cell to be replaced:
    - $\sim f(\alpha, \beta)$: Moran model A [fitness $\rightarrow$ fitness]
    - $\sim$ Uniform: Moran model B [uniform $\rightarrow$ fitness]

Balanced evolution, criticality and conditioning

Moran model A [fitness → fitness]

- Constant population size over time
- Equilibrium condition: $p \cdot s = (1-p) \cdot d \rightarrow$ expected fitness is constant ("balanced")

Moran model B [uniform → fitness]

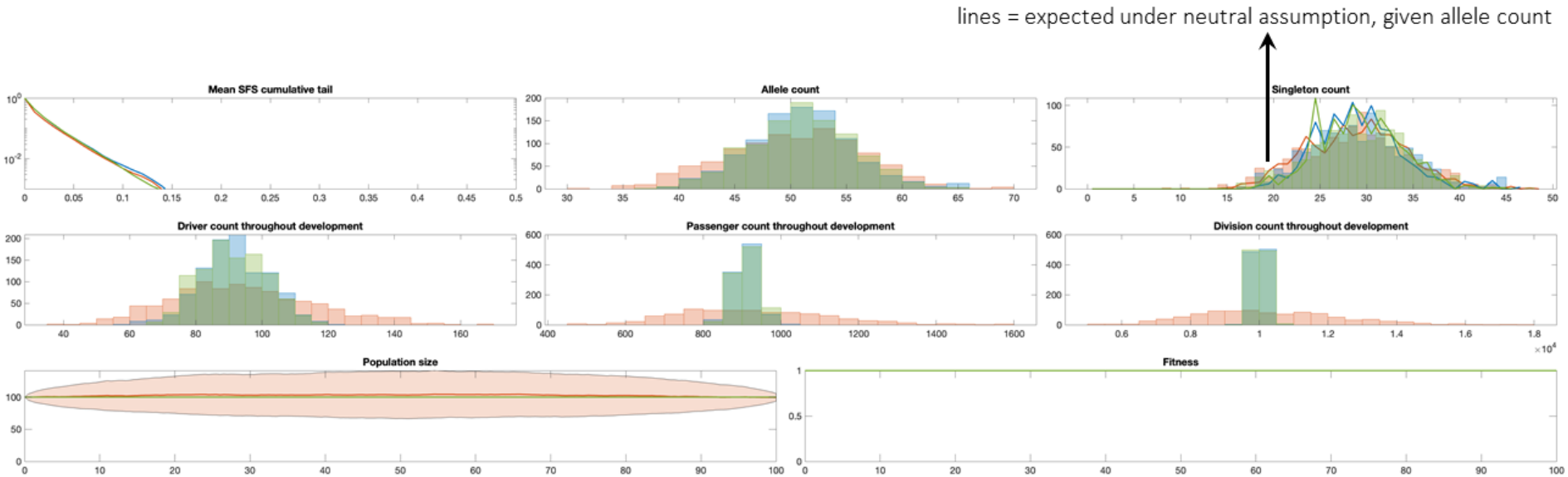
- Constant population size over time
- Expected fitness change ≥ 0 after every replacement event, independent of parameters
  $\Longrightarrow$ stronger selection
  $\Longrightarrow$ less clones/alleles, less singletons

Branching Process

- Population size is expected to remain unchanged, but size fluctuates in individual simulations
- For direct comparison to Moran: simulations accepted if $\left|1 - \frac{N(t_{end})}{N(t_0)}\right| \leq 0.1$

8

Examples - neutral evolution ($s = 0$, $d = 0$)

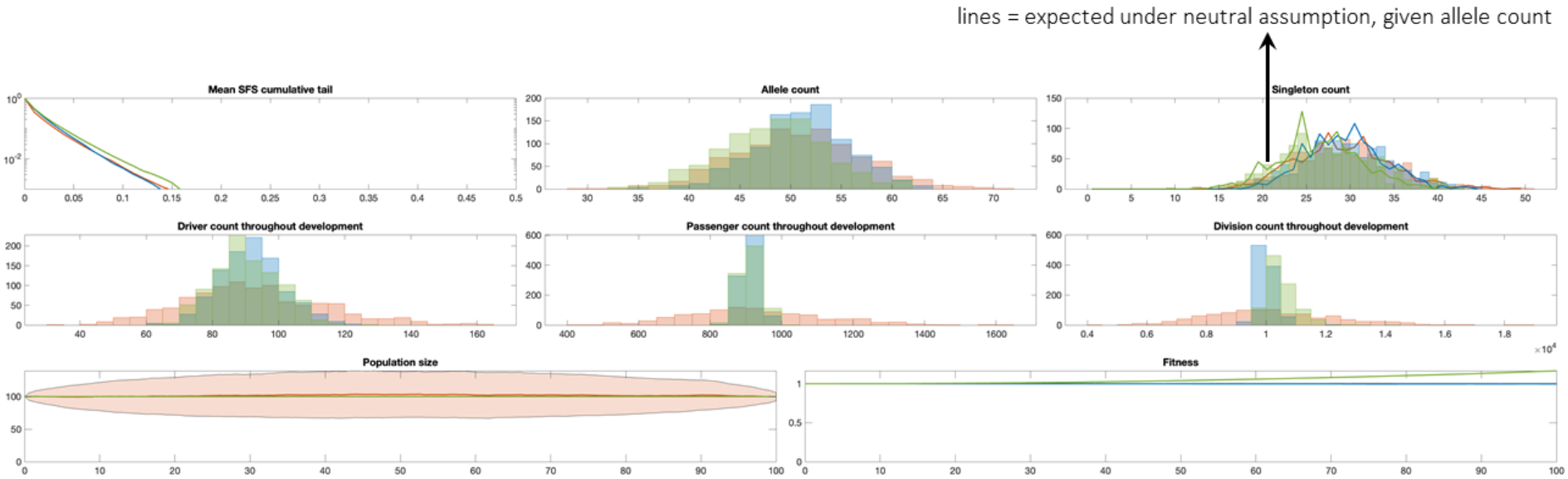


Orange= Branching Process

Blue = Moran model A [fitness → fitness]

Green = Moran model B [uniform → fitness]

Examples - balanced evolution ($s = 0.1$, $d = 0.01$)

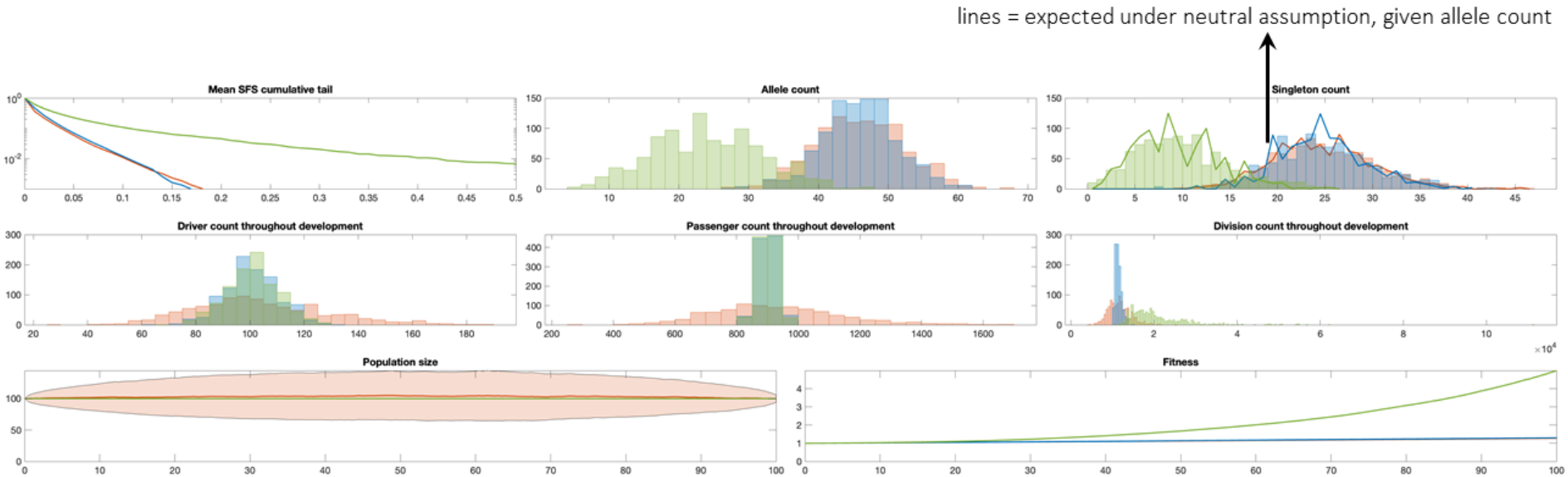


Orange= Branching Process

Blue = Moran model A [fitness → fitness]

Green = Moran model B [uniform → fitness]

Examples - selective evolution ($s = 0.25$, $d = 0$)

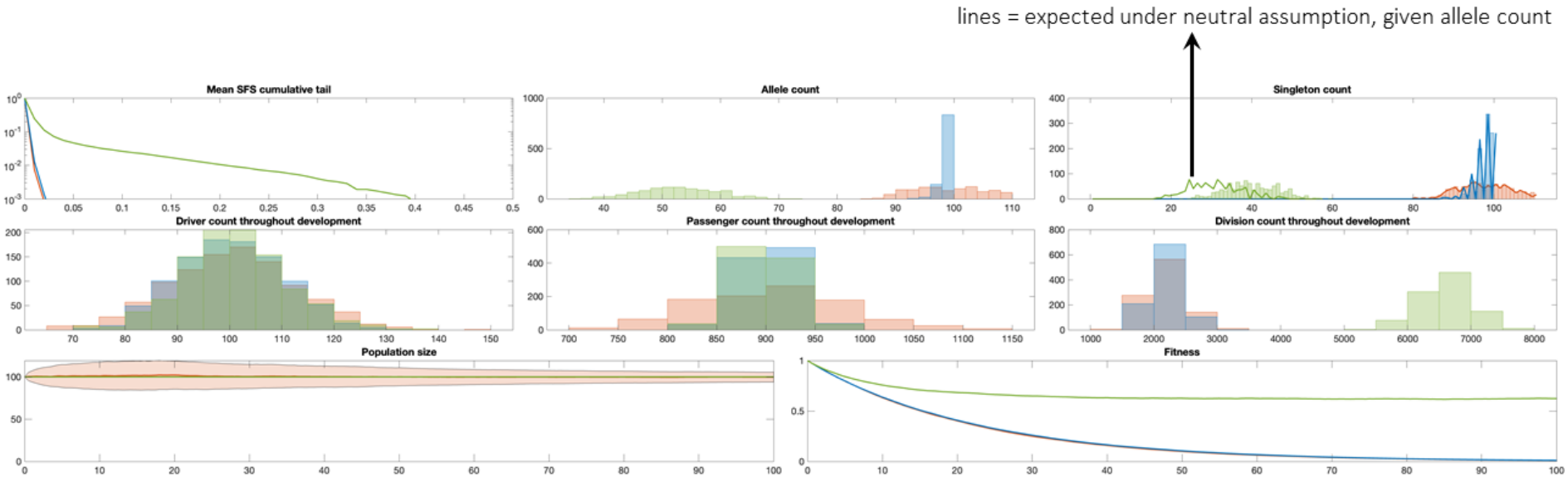


Orange= Branching Process

Blue = Moran model A [fitness → fitness]

Green = Moran model B [uniform → fitness]

Examples - deleterious evolution ($s = 0$, $d = 0.5$)



Orange= Branching Process

Blue = Moran model A [fitness → fitness]

Green = Moran model B [uniform → fitness]

12

Fitting for breast cancer WES data

- 6 patients (Maria Sklodowska-Curie National Research Institute of Oncology)
- 4 HER2+ breast cancer, 2 triple-negative breast cancer
- Paired samples from primary breast tumor + concurrent metastasis in lymph nodes
- Whole exome sequencing (WES)

Optimization objective:

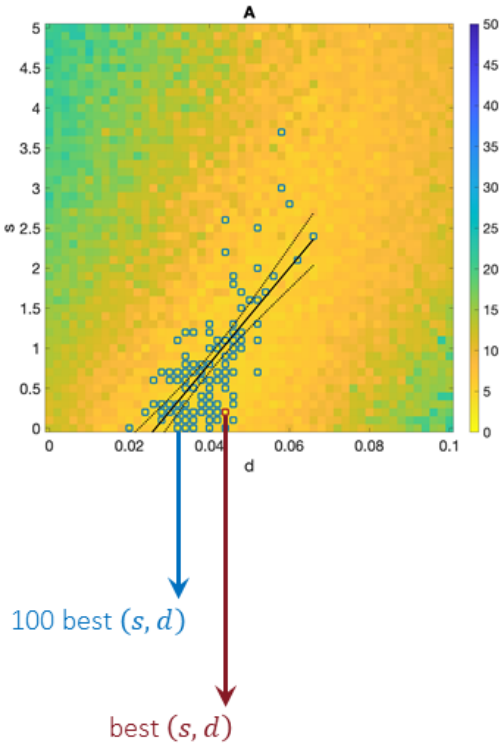
- $\{\hat{S}(f_i)\}$ = reverse cumulative SFS from data
- $\{S(f_i|s,d)\}$ = reverse cumulative SFS from model with parameters (s,d), averaged from 1000 simulations
- $\mathrm{Error(s,d)} = \sum\left|\log_{10}\hat{S}(f_i) - \log_{10}S(f_i|s,d)\right|$

Fitting strategy:

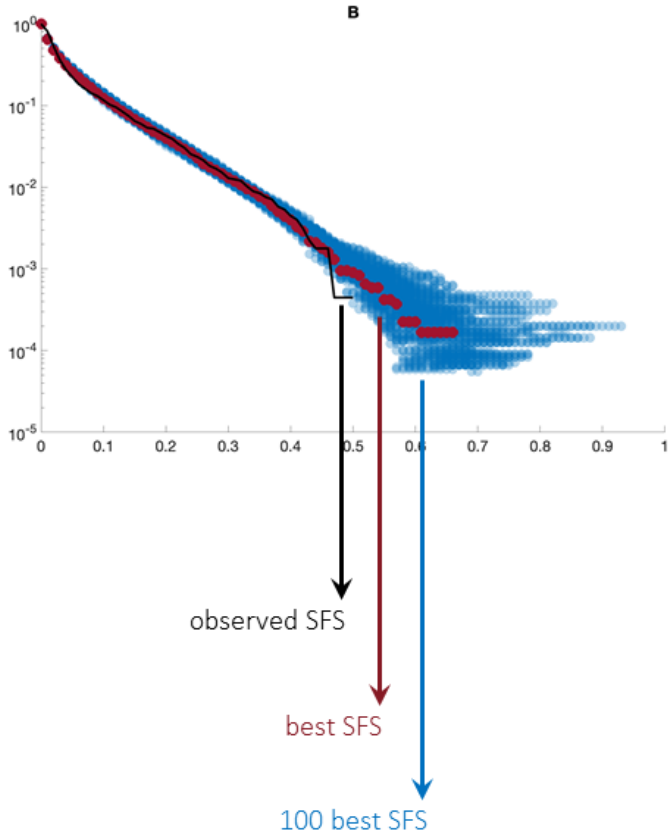
1. Examine error heatmap for Moran A [fitness → fitness] over wide ranges of (s,d)
2. Choose best-fitted $(\boldsymbol{s},\boldsymbol{d})$
3. Test same (s,d) with Branching Process

Fitting for breast cancer WES data

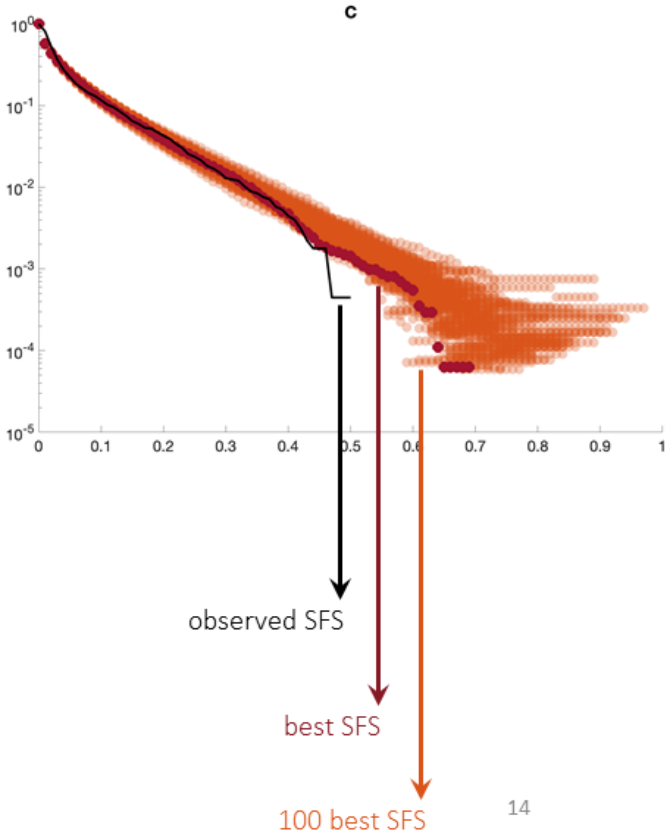
Error heatmap for Moran A



SFS comparison for Moran A



SFS comparison for Branching Process



14

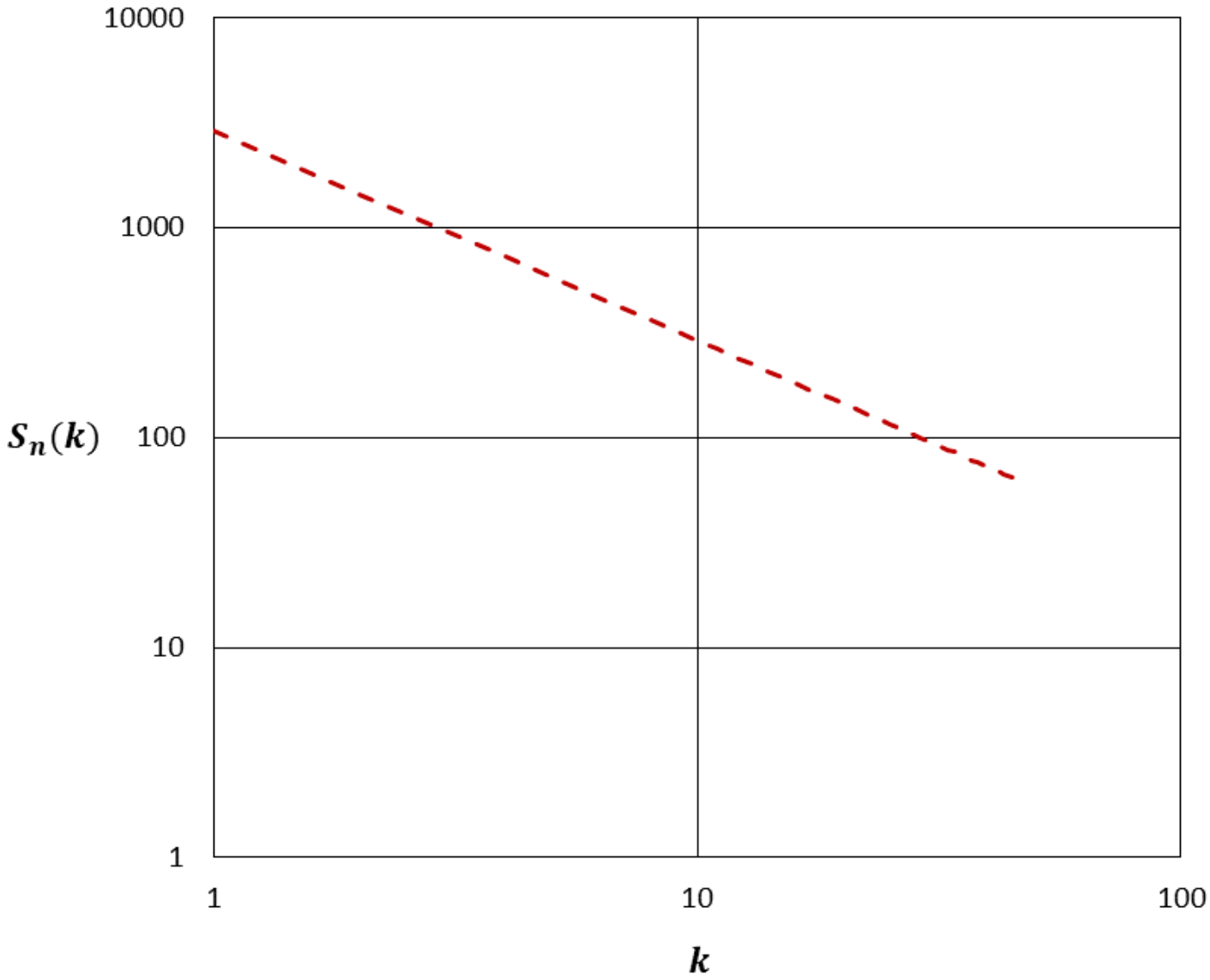
Phylogenetics of genomes of SARS-Cov-2-related viruses

Ca. 47 complete 32kb ssRNA genomes available
Analyses using BEAST 2 indicate:

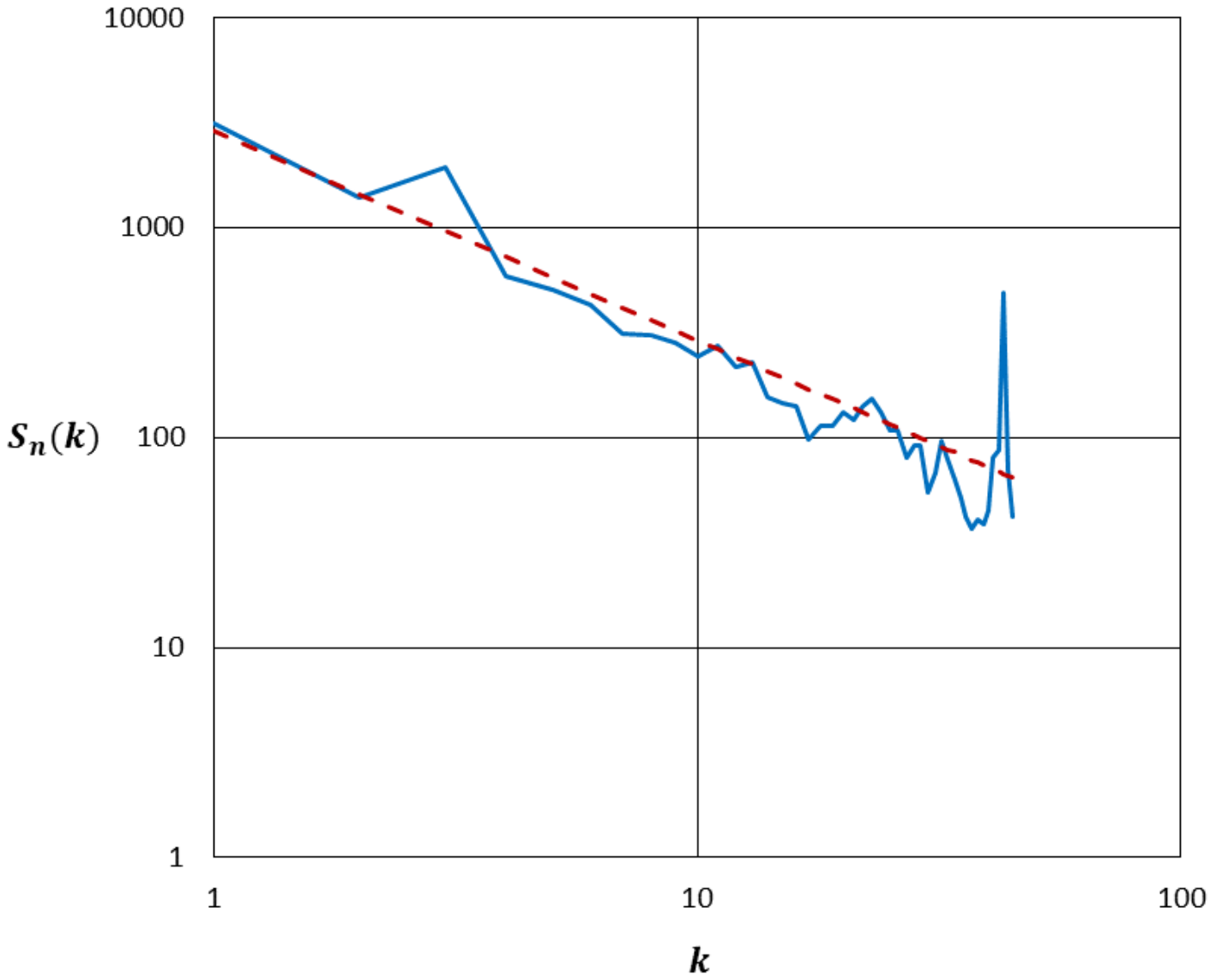
- Long-time coexistence with the hosts ($10^2 - 10^3$ yrs)
- Constant and rather small effective population size (100)
- Neutrality (?)

We may compare empirical SFS to expectations under different models
Under neutral Moran with constant pop size $\mathbb{E}(S_n(k) \sim 1/k$, hence slope -1 in log-log scale (Griffiths and Tavaré 1998)

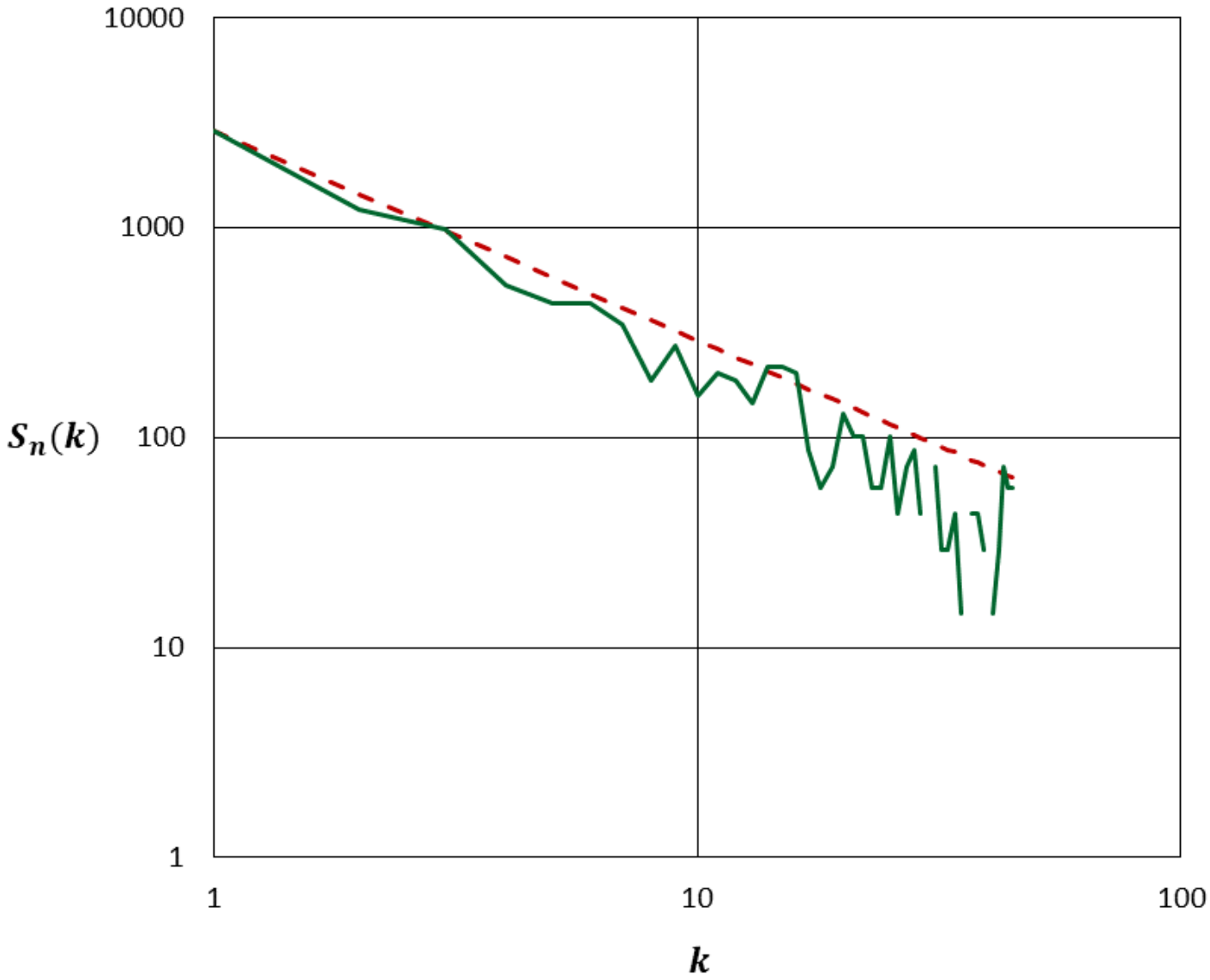
Expected



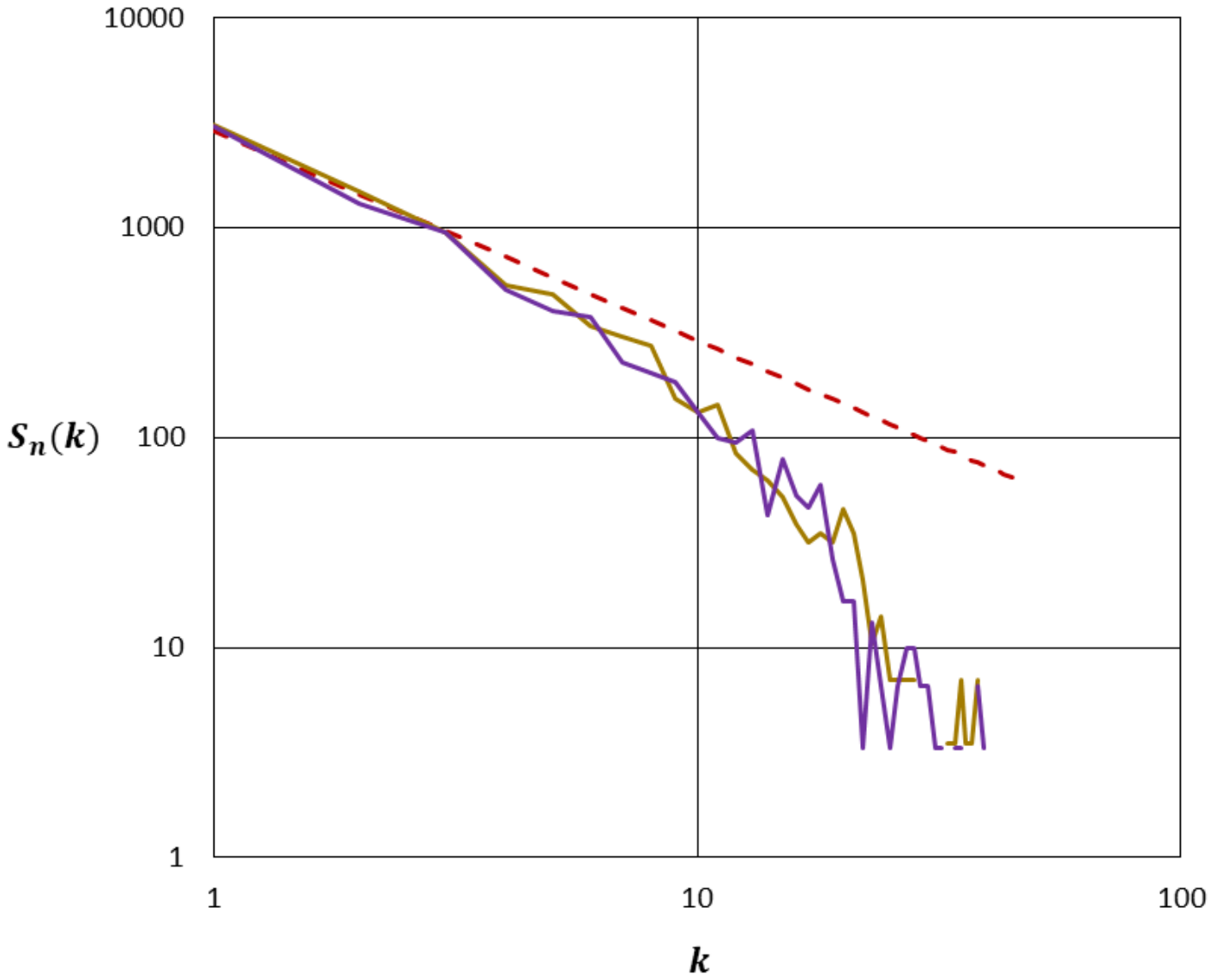
Empirical



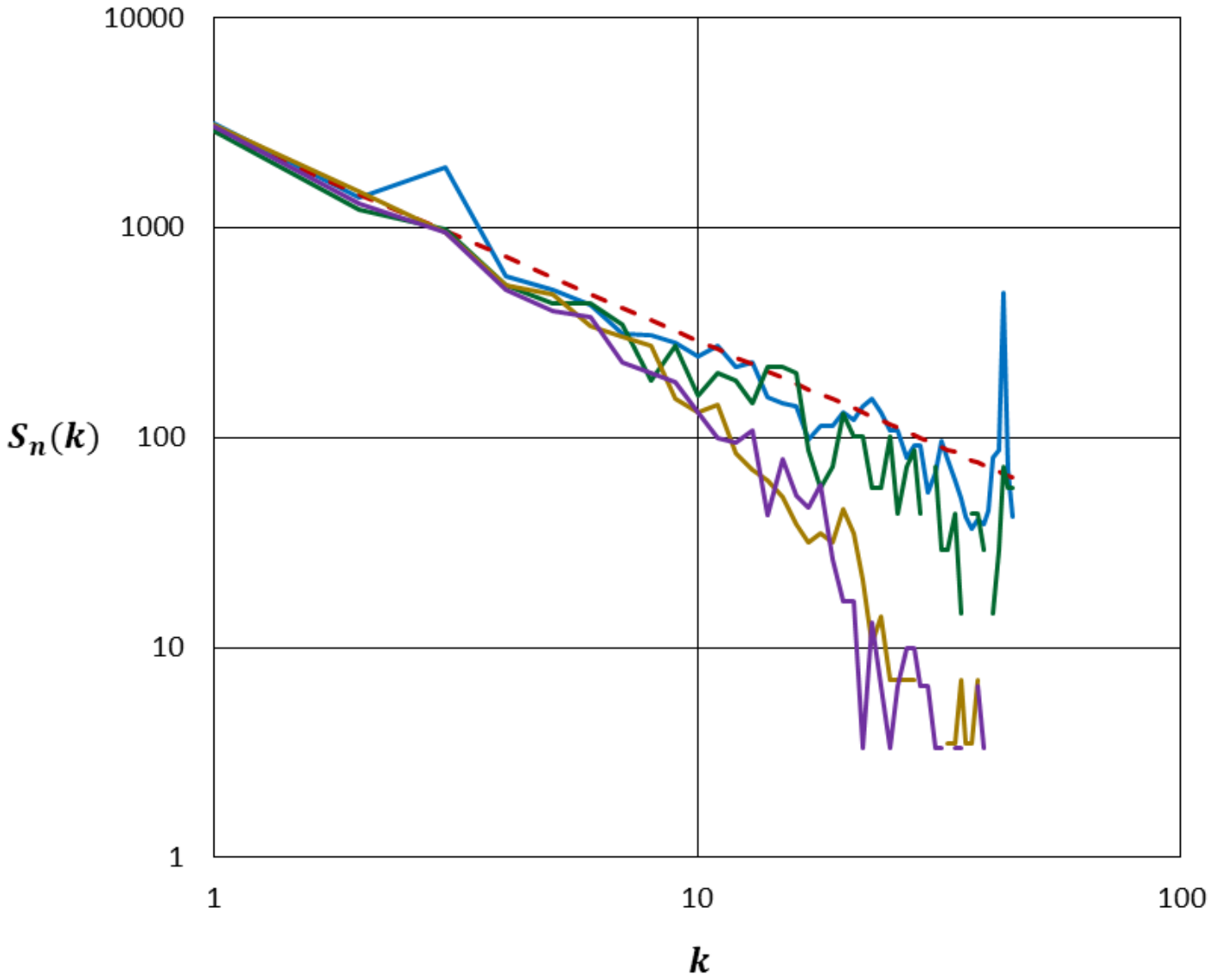
Tug-of-War $N = 100$, $s = 0.6$, $d = 0$, and $p = 0.5$



Tug-of-War $N = 100$, $s = 0$, $d = 0$, and $p = 0.5$



Interpretation?



97 weeks of SARS-Cov-2: *Kurpas et al. (2022) Viruses*

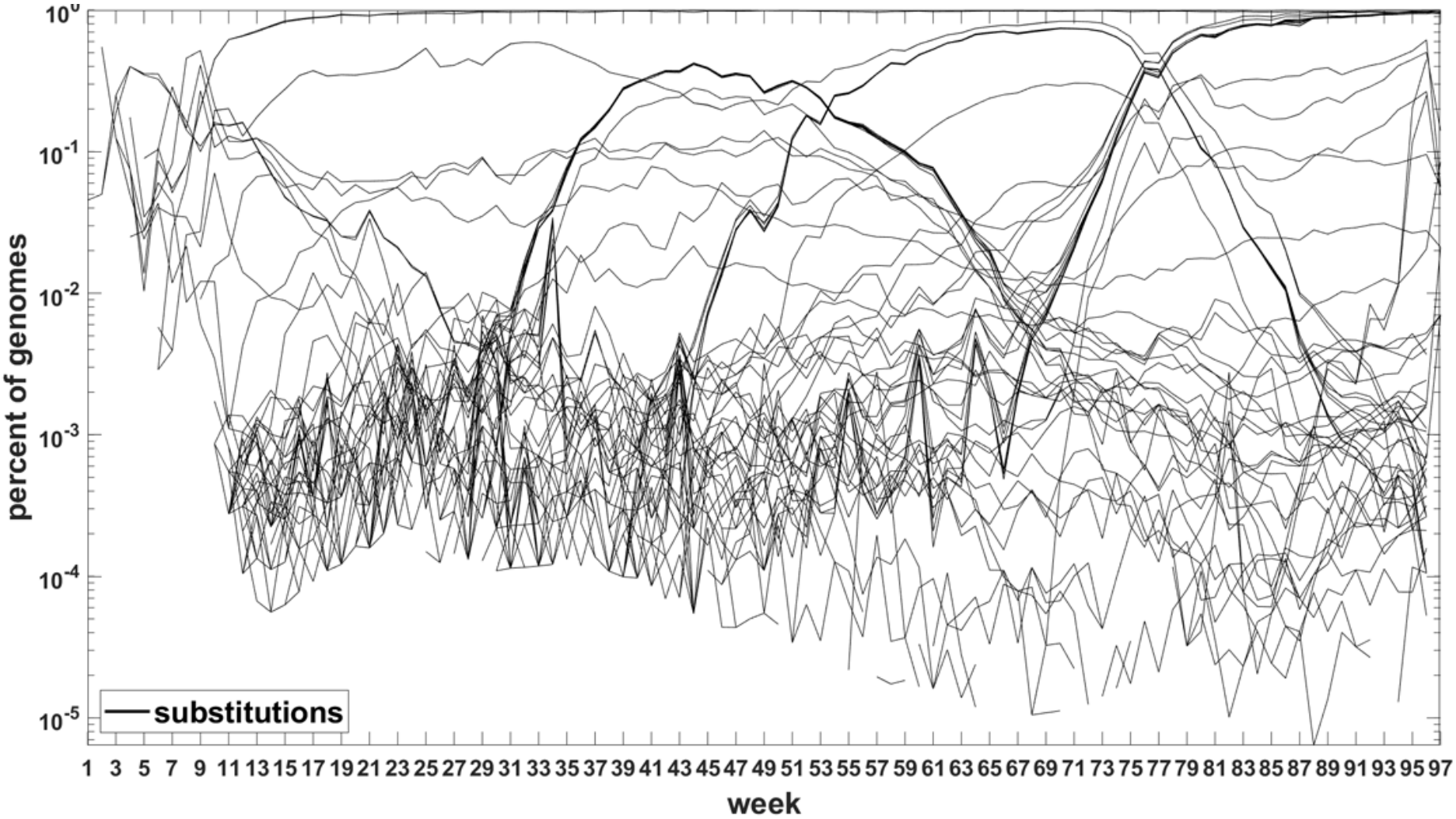


Figure: Time trajectories of frequencies of the nucleotide substitutions among top 1000 mutations observed in all genomes. Variant-defining sites excluded.

Tug-of-War Driver Clones: *Kurpas et al. (2022) Viruses*

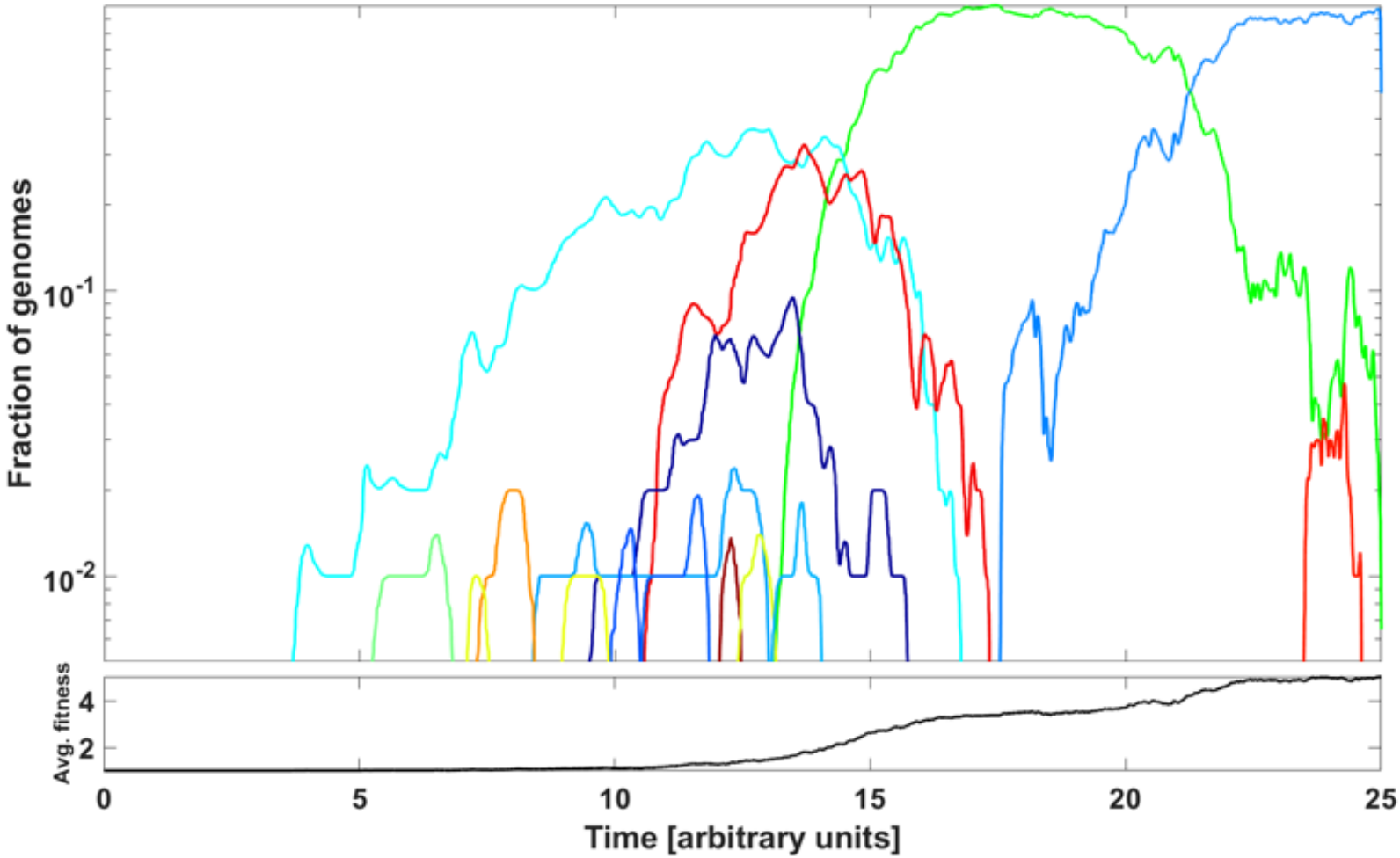


Figure: Frequencies of clones initiated by driver mutations in McFarland's Tug-of-War stochastic process (Model B) modeling evolution of a population of N genomes.