# The Accumulation of Beneficial Mutations and Convergence to a Poisson Process

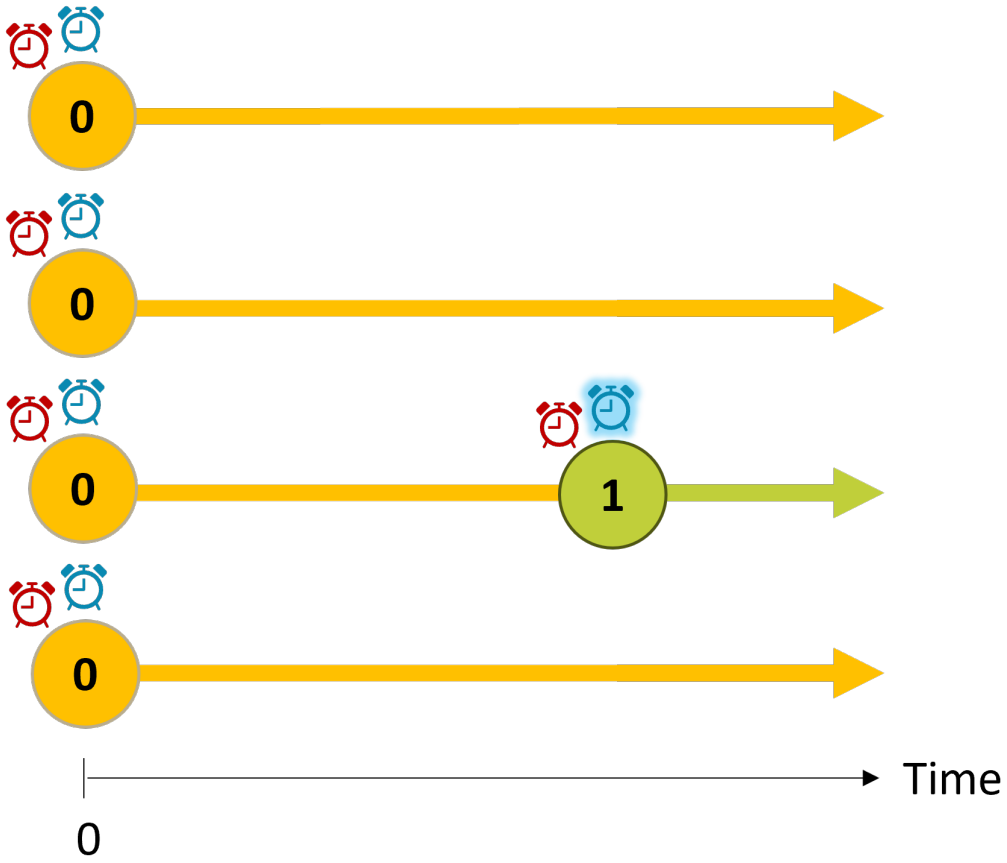
Nantawat Udomchatpitak

Mahidol University, Thailand
Joint work with Jason Schweinsberg from University of California San Diego, USA.

May 20, 2024

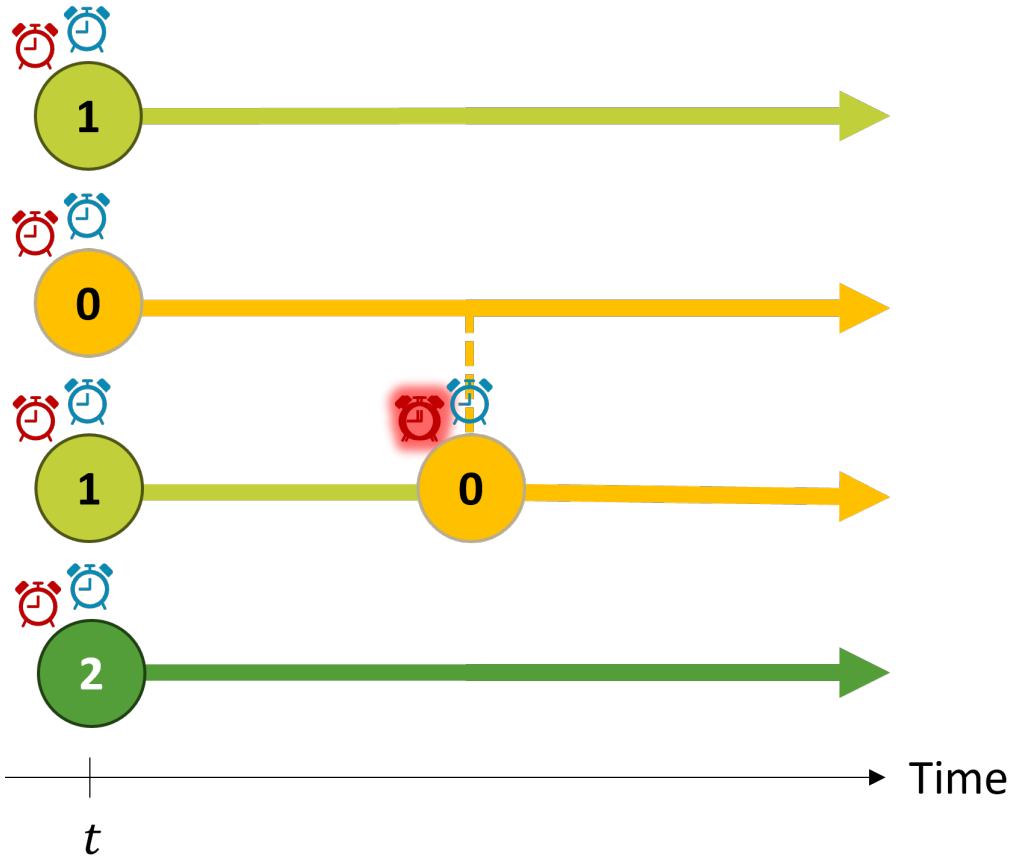
# Outlines

# The Model




- Population size: $N$.
- Mutation
  - Rate: $\mu_N$ per individual.
  - All mutations are beneficial.
  - The population starts without any mutations.
- Selection
  - Fitness of an individual: $(1 + s_N)^k$ where $k$ is the number of mutations that the individual has.
  - Death rate: 1 per individual.
  - Replacement of the death is randomly chosen proportional to the fitness.

# The Model




- Population size: $N$.
- Mutation
  - Rate: $\mu_N$ per individual.
  - All mutations are beneficial.
  - The population starts without any mutations.
- Selection
  - Fitness of an individual: $(1 + s_N)^k$ where $k$ is the number of mutations that the individual has.
  - Death rate: 1 per individual.
  - Replacement of the death is randomly chosen proportional to the fitness.

# The Case of a Single Mutation

Consider the case that one individual acquires a mutation, and no further mutations occur. Let $X_{0,N}(t)$ and $X_{1,N}(t)$ be the number of individuals at time $t$ with no mutations and with one mutation, respectively.

- The process $X_1$ jumps up by 1 at rate $X_{0,N}(t) \cdot \frac{(1+s_N)X_{1,N}}{X_{0,N}(t)+(1+s_N)X_{1,N}(t)}$.
- The process $X_1$ jumps down by 1 at rate $X_{1,N}(t) \cdot \frac{X_{0,N}}{X_{0,N}(t)+(1+s_N)X_{1,N}(t)}$.
- The ratio of the jump-up rate to the jump-down rate is $1 + s_N$.
- Standard results on asymmetric random walks yield that $X_1$ hits $N$ before 0 with probability

$$\frac{s_N}{(1+s_N)(1-(1+s_N)^{-N})},$$

  which is approximately $\frac{s_N}{1+s_N}$ if $(1+s_N)^N \to 0$ as $N \to \infty$.
- Given that the selective sweep occurs, the duration of the selective sweep is approximately $\frac{2}{s_N} \log N$.

# Assumptions on the Parameters

1. $\mu_N \ll \frac{1}{N \log N}$.
2. $s_N \sim N^{-\eta}$ where $\eta \in (0, 1)$ is a constant.

Reasons for the assumptions:

- Total mutation rate is $N\mu_N$.
- The probability that a mutation triggers a selective sweep is approximately $\frac{s_N}{1+s_N} \approx s_N$, provided that $s_N \ll 1$.
- Mutation that triggers a selective sweep occurs at rate $N\mu_N s_N$.
- The duration of a selective sweep is approximately $\frac{2}{s_N} \log N$.
- The assumption $\mu_N \ll \frac{1}{N \log N}$ implies that the waiting time for a mutation that triggers a selective sweep is much longer than the the duration of a selective sweep.

# Notations

- Let $X_{k,N}(t)$ be the number of individuals with exactly $k$ mutations at time $t$ for all nonnegative integers $k$ and all $t \geq 0$.
- Let $T_{k,N} = \inf\{t \geq 0 : X_{k,N}(t) > \frac{\log N}{s_N}\}$ for all positive integers $k$, and let $T_{0,N} = 0$.

An individual with exactly $k$ mutations will be called <u>type $k$</u> for all nonnegative integers $k$.

# The Main Result

## Theorem (Part 1)

Let $\eta \in (0,1)$. Assume that $\mu_N \ll \frac{1}{N \log N}$ and $s_N \sim N^{-\eta}$. Let $(\xi_k)_{k=1}^{\infty}$ be a sequence of independent random variables having the exponential distribution with mean one. Then for each fixed positive integer $K$, as $N \to \infty$ we have the convergence in distribution

$$\left(N\mu_N s_N (T_{k,N} - T_{k-1,N})\right)_{k=1}^{K} \Rightarrow (\xi_k)_{k=1}^{K}. \tag{1}$$

# The Main Result

## Theorem (Part 2)

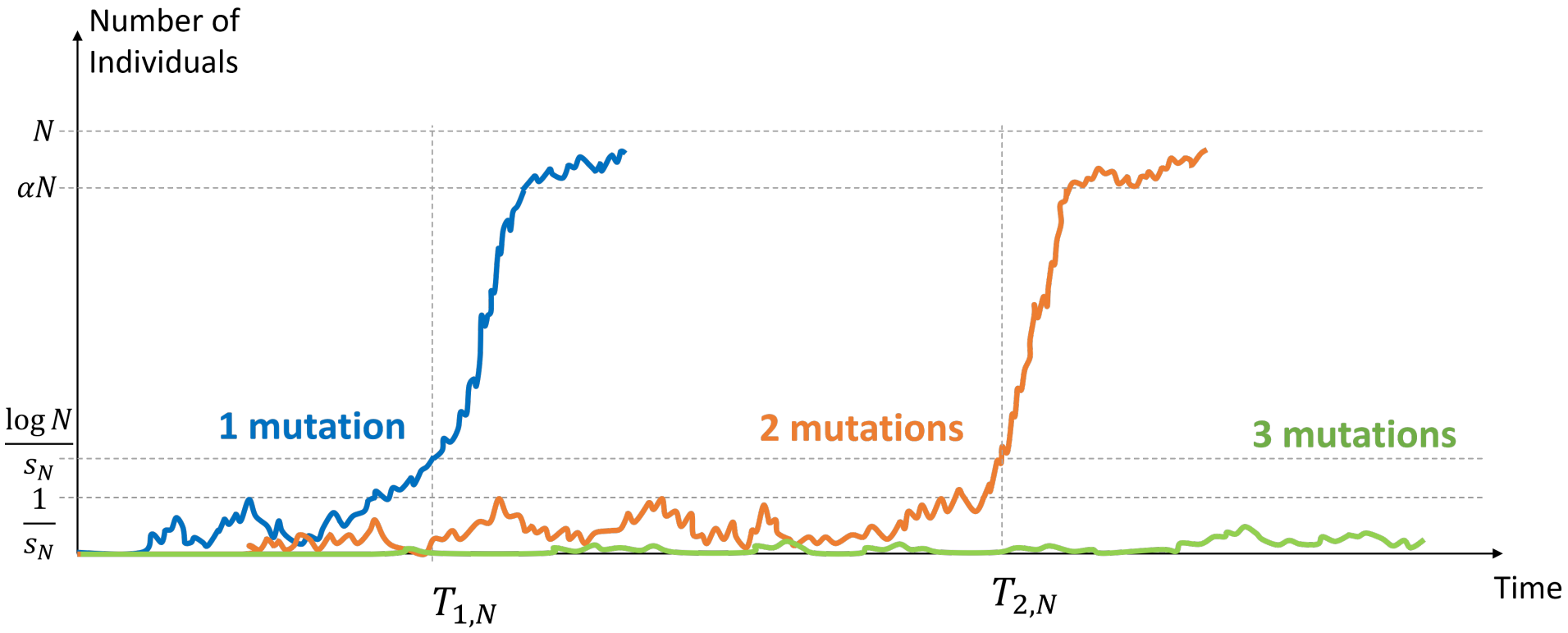
Furthermore, there exist positive constants $C_1$ and $C_2$ and a positive integer $\Delta$, all depending on $\eta$, such that for all nonnegative integers $k$, we have

$$\lim_{N\to\infty} P\left(X_{k,N}(t) \geq N - \frac{C_2 \log N}{s_N} \text{ for all } t \in \left[T_{k,N} + \frac{C_1 \log N}{s_N}, T_{k+1,N}\right)\right) = 1 \tag{2}$$

and

$$\lim_{N\to\infty} P\left(\sum_{j=k}^{k+\Delta} X_{j,N}(t) = N \text{ for all } t \in \left[T_{k,N} + \frac{C_1 \log N}{s_N}, T_{k+1,N}\right)\right) = 1. \tag{3}$$

# The Main Result



# The Main Result

## Corollary

For all $t \geq 0$, let

$$\overline{X}_N(t) = \frac{1}{N} \sum_{k=0}^{\infty} k X_k(t).$$

Then, the finite-dimensional distributions of the processes $(\overline{X}_N(t/(N\mu_N s_N)), t \geq 0)$ converge as $N \to \infty$ to the finite-dimensional distributions of a homogeneous rate one Poisson process.

# Related Works

- Lenski's long-term evolution experiment (LTEE)
  - Every day, a sample of the populations of *Escherichia coli* is chosen to populate the next generation.
  - Observe that the mean fitness over time is a concave function.
- Casanova, Kurt, Wakolbinger, and Yuan (2016) presented a model that explains the Lenski's experiment. They assume that $\mu_N \sim N^{-(1+a)}$ and $s_N \sim N^{-b}$ where $0 < b < 1$ and $a > 3b$.
- Our work suggests that the same results may still hold under the weaker assumption that $a > 0$.

# Related Works (Different Values of $\mu_N$ and $s_N$)

1. Case: $\mu_N \sim \frac{C}{N \log N}$ where $C$ is a positive constant.
   - See overlaps between selective sweeps.
   - Studied by Gerrish and Lenski (1998).
   - Recent rigorous study by Casanova, Hermann, dos Santos, Tobias, and Wakolbinger. (The manuscript is in preparation.)
2. Case: $\mu_N \sim N^{-a}$ for some $a \in (0, 1)$ and $s_N = s$ is a constant.
   - Durrett and Mayberry (2011) studied the same model as ours. They showed that if $T_k$ is the first time that an individual has $k$ mutations, then there is a constant $t_k$ such that as $N \to \infty$,

$$\frac{sT_k}{\log(1/\mu_N)} \to_P t_k.$$

# Related Works (Different Values of $\mu_N$ and $s_N$)

3. Case: $N^{-a} \ll \mu_N \ll s_N^b$ for all $a > 0$ and $b > 0$.
   - Schweinsberg (2017) studied a similar model and made the results of Desai and Fisher (2007), and Desai, Walczak, and Fisher (2013) rigorous.
4. Case: $\mu_N = \mu$ and $s_N = s$ are both positive constants.
   - Yu, Etheridge, and Cuthbertson (2010) studied a different model and showed that the mean fitness increases at rate, on average, bounded below by $O(\log^{1-\delta} N)$.
   - Kelly (2013) showed that the rate that the mean fitness increases is bounded above by $O(\log N/(\log\log N)^2)$.
5. Case: $\mu_N$ and $s_N$ are of order $\frac{1}{N}$.
   - Study by using diffusion approximation. (See section 8.1 in Durrett's *Probability Models for DNA Sequence Evolution*).

# The Main Ideas of the Proof

**We shall omit writing the subscript $N$.

1. Up to time $T_1$, couple the process $X_1$ with two processes that bound $X_1$ from above and from below such that after a time scaling, both processes become branching processes with immigration.
2. For $k = 2, 3, \ldots$, define $M_k(t)$ to be the number of type $k$ individuals who mutate from being type $k-1$ until time $t$. Then, show that all $M_k(T_1)$ are small that they cannot prevent type 1 from almost fixation.
3. Show that there is a constant $\Delta$ such that $M_{\Delta+1}(T_1) \ll 1$.
4. After $T_1$, type 1 quickly reaches the point of almost fixation.

# Rates

Define $S(t) = \sum_{k=0}^{\infty}(1+s)^k X_k(t)$, which is the total fitness at time $t$.

For $k \geq 1$, the process $X_k$ is a birth-death process with immigration.

- Birth: a non-type $k$ individual dies and is replaced by a type $k$. The birth rate is

$$(N - X_k(t))\frac{(1+s)^k X_k(t)}{S(t)} =: b_k(t)X_k(t).$$

- Death:
  - a non-type $k$ individual dies and is replaced by a type $k$, or
  - a type $k$ individual becomes type $k+1$.

  The death rate is

$$X_k(t)\left(1 - \frac{(1+s)^k X_k(t)}{S(t)}\right) + \mu X_k(t) =: d_k(t)X_k(t).$$

- Immigration: a type $k-1$ individual becomes type $k$. The immigration rate is

$$m_k(t) := \mu X_{k-1}(t).$$

# Bound $X_1$ from Above

Before time $T_1$, the majority of the population should be type 0. Then, if $0 \leq t < T_1$,

- $b_1(t) = \frac{(1+s)(N-X_1(t))}{S(t)} \approx 1 + s$,
- $d_1(t) = 1 - \frac{(1+s)X_1(t)}{S(t)} + \mu \approx 1$,
- $m_1(t) = \mu X_0(t) \approx \mu N$.

Note that $b_1(t) < (1+s)d_1(t)$. Also, for every constant $a$, and for sufficiently large $N$ depending on $a$,

$$m_1(t) \leq \mu N(1+a)d_1(t).$$

# Bound $X_1$ from Above

We can construct a new birth-death process with immigration $Y_1$ such that

- birth rate per individual is $(1+s)d_1(t)$,
- death rate per individual is $d_1(t)$,
- immigration rate is $\mu N(1+a)d_1(t)$,
- $X_1(t) \leq Y_1(t)$ for all $0 \leq t \leq T_1$.

After a time scaling, the process $Y_1$ becomes a process $\tilde{Y}_1$ in which

- birth rate per individual is $1+s$,
- death rate per individual is 1,
- immigration rate is $\mu N(1+a)$.

In $\tilde{Y}_1$, the extinction probability of the family of each immigrant is $\frac{1}{1+s}$. Hence, the immigrant whose family survives appears at rate $\frac{(1+a)N\mu s}{1+s}$.

# Bound $X_1$ from Below

Let $\gamma, \zeta \in (0,1)$. By pruning some births and deaths in the process $X_1$, we can construct a process $Z_1$ such that

- birth rate per individual is $(1+\gamma s)d_1(t)$,
- death rate per individual is $d_1(t)$,
- immigration rate is $\mu N(1-\zeta)d_1(t)$,
- $X_1(t) \geq Z_1(t)$ for all $0 \leq t \leq T_1$.

After a time scaling, the immigrant whose family survives appears at rate $\frac{(1-\zeta)\gamma N \mu s}{1+\gamma s}$.

**In the construction, we need a good lower bound of the total fitness $\sum_{k=0}^{\infty}(1+s)^k X_k(t)$. Hence, we need to show that

1. $T_1 < T_k$ for all $k = 2, 3, 4, \ldots$.
2. There is a positive integer $\Delta$ such that no type $\Delta$ appears before $T_1$.

# An Upper Bound on $M_k$

Let $C > 0$ be a constant. Given that $T_1 < \frac{C}{N\mu s}$,

$$\int_0^{T_1} X_1(t)dt \leq \frac{\log N}{s} \cdot \frac{C}{N\mu s} = \frac{C \log N}{N\mu s^2}.$$

Hence,

$$E\left[M_2(T_1) | \, T_1 < \frac{C}{N\mu s}\right] \leq \frac{C \log N}{N\mu s^2} \cdot \mu = \frac{C \log N}{N s^2}.$$

# An Upper Bound on $M_k$

If we consider the branching process that start with 1 individual, and each individual gives birth and dies at rate $(1+s)^k$ and 1, respectively,

1. the extinction probability is $1/(1+s)^k$, and
2. given that the process goes extinct, the process becomes a branching process in which each individual gives birth and dies at rate 1 and $(1+s)^k$, respectively. Hence, the expected number of individuals that live before the extinction is

$$\int_0^\infty e^{(1-(1+s)^k)t} dt = \frac{1}{(1+s)^k - 1} \leq \frac{1}{ks}.$$

# An Upper Bound on $M_k$

The probability that the families of all type 2 immigrants that appear before $T_1$ go extinct is approximately

$$(1+s)^{-k\cdot\frac{\log N}{Ns^2}} \to 1$$

as $N \to \infty$.

From $M_2(T_1) \leq O(\frac{\log N}{Ns^2})$, the expected number of type 2 individuals that live before the $T_1$ is bounded above by

$$O\left(\frac{\log N}{Ns^2}\right)\cdot\frac{1}{2s} = O\left(\frac{\log N}{Ns^3}\right).$$

Each type 2 individuals mutates to type 3 at rate $\mu$. Then,

$$M_3(T_1) \leq O\left(\frac{\log N}{Ns^3}\right)\cdot\mu = O\left(\frac{\mu\log N}{Ns^3}\right).$$

# An Upper Bound on $M_k$

Inductively, for $k \geq 2$, we have

$$M_k(T_1) \leq O\left(\frac{\mu^{k-2}\log N}{s^k N}\right).$$

It follows from the assumptions on $\mu$ and $s$ that

$$\frac{\mu^{k-2}\log N}{s^k N} \ll \frac{1}{N^{(1-\eta)k-1}(\log N)^{k-3}}.$$

Hence, there is a constant $\Delta$ that $M_{\Delta+1}(T_1) \ll 1$.

# References I


- E. Baake, A. González Casanova, S. Probst, and A. Wakolbinger (2019). Modelling and simulating Lenski's long-term evolution experiment. *Theor. Pop. Biol.* **127**, 58-74.
- É. Brunet, I. M. Rouzine, and C. O. Wilke (2008). The stochastic edge in adaptive evolution. *Genetics* **179**, 603-620.
- M. M. Desai and D. S. Fisher (2007). Beneficial mutation-selection balance and the effect of linkage on positive selection. *Genetics* **176**, 1759-1798.
- M. M. Desai, A. M. Walczak, and D. S. Fisher (2013). Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics* **193**, 565-585.
- R. Durrett (2008). *Probability Models for DNA Sequence Evolution*. 2nd ed. Springer, New York,

# References II


- R. Durrett and J. Mayberry (2011). Traveling waves of selective sweeps. *Ann. Appl. Probab.* **21**, 699-744.
- D. S. Fisher (2013). Asexual evolution waves: fluctuations and universality. *Journal of Statistical Mechanics: Theory and Experiment*, P01011.
- V. G. Gadag and M. B. Rajarshi (1992). On processes associated with a super-critical Markov branching process. *Serdica.* **18**, 173-178.
- P. J. Gerrish and R. E. Lenski (1998). The fate of competing beneficial mutations in an asexual population. *Genetica* **102/103**, 127-144.
- A. González Casanova, F. Hermann, R. Soares dos Santos, A. Tobiás, and A. Wakolbinger. In preparation.

# References III


- A. González Casanova, N. Kurt, A. Wakolbinger, and L. Yuan (2016). An individual-based model for the Lenski experiment, and the deceleration of the relative fitness. *Stochastic Process. Appl.* **126**, 2211-2252.
- B. H. Good, I. M. Rouzine, D. J. Balick, O. Hallatschek, and M. M. Desai (2012). Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc. Natl. Acad. Sci. USA* **109**, 4950-4955.
- B. H. Good, A. M. Walczak, R. A. Neher, and M. M. Desai (2014). Genetic diversity in the interference selection limit. *PLOS Genetics* **10**, e1004222.
- M. Kelly (2013). Upper bound on the rate of adaptation in an asexual population. *Ann. Appl. Probab.* **23**, 1377-1408.

# References IV


- M. Kimura and T. Ohta (1969). The average number of generations until the fixation of a mutant gene in a finite population. *Genetics* **61**, 763-771.
- J. Liu and J. Schweinsberg (2021). Particle configurations for branching Brownian motion with an inhomogeneous branching rate. *ALEA Lat. Am. J. Probab. Math. Stat.* **20**, 731-803.
- M. J. Melissa, B. H. Good, D. S. Fisher, and M. M. Desai (2022). Population genetics of polymorphism and divergence in rapidly evolving populations. *Genetics* **221**, iyac053.
- R. A. Neher and O. Hallatschek (2013). Genealogies in rapidly adapting populations. *Proc. Natl. Acad. Sci.* **110**, 437-442.
- M. Roberts and J. Schweinsberg (2020). A Gaussian particle distribution for branching Brownian motion with an inhomogeneous branching rate. *Electron. J. Probab.* **26**, 1-76.

# References V


- I. M. Rouzine, É. Brunet, and C. O. Wilke (2008). The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. *Theor. Pop. Biol* **73**, 24-46.
- J. Schweinsberg (2017). Rigorous results for a population model with selection I: evolution of the fitness distribution. *Electron. J. Probab.* **22**, no. 37, 1-94.
- J. Schweinsberg (2017). Rigorous results for a population model with selection II: genealogy of the population. *Electron. J. Probab.* **22**, no. 38, 1-54.
- F. Yu, A. Etheridge, and C. Cuthbertson (2010). Asymptotic behavior of the rate of adaptation. *Ann. Appl. Probab.* **20**, 978-1004.