Population Genetics and Evolution - III

The Mechanisms of Evolution: Drift

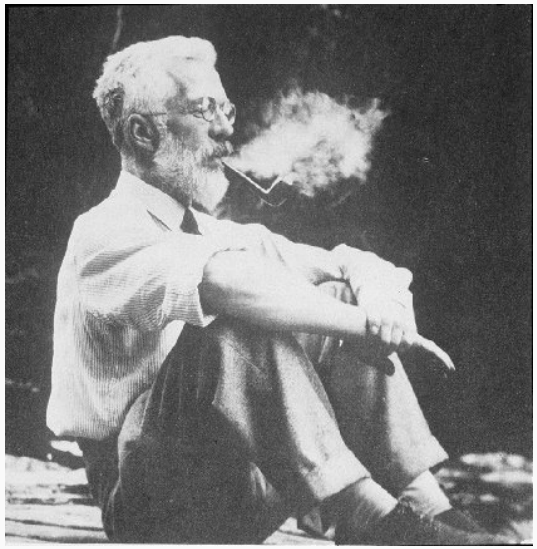
São Paulo / January 2019

SMRI (Italy)
luca@peliti.org

Drift

The Population Genetics Triad

Sewall Wright



Ronald A. Fisher

Motoo Kimura

Finite population

The Wright-Fisher model

- Population size N, number n_k of individuals of type k, $k = 1, \ldots, r$, with fitness w_k
- Nonoverlapping generations
- Given the composition vector $\boldsymbol{x} = (x_i)$, $x_i = n_i/N$, the numbers n'_k in the next generation are distributed according to

$$\mathrm{Prob}(n'_1, \ldots, n'_r) = \frac{N!}{n'_1! \cdots n'_r!} \xi_1^{n'_1} \cdots \xi_r^{n'_r}$$

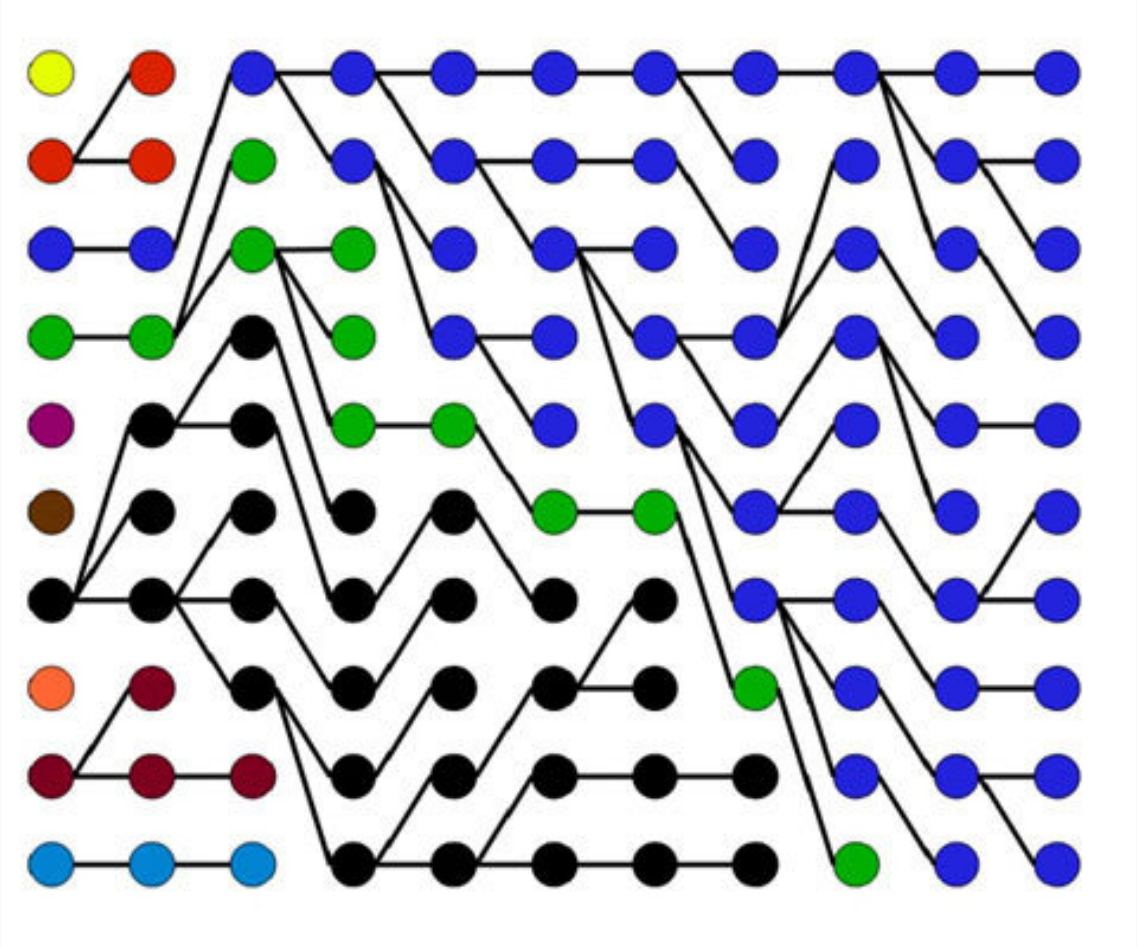
where

$$\xi_k = \frac{x_k w_k}{\sum_j x_j w_j}$$

- Thus n'_k is approximately distributed as a Gaussian with mean $N\xi_k$ and variance $N\xi_k(1 - \xi_k)$

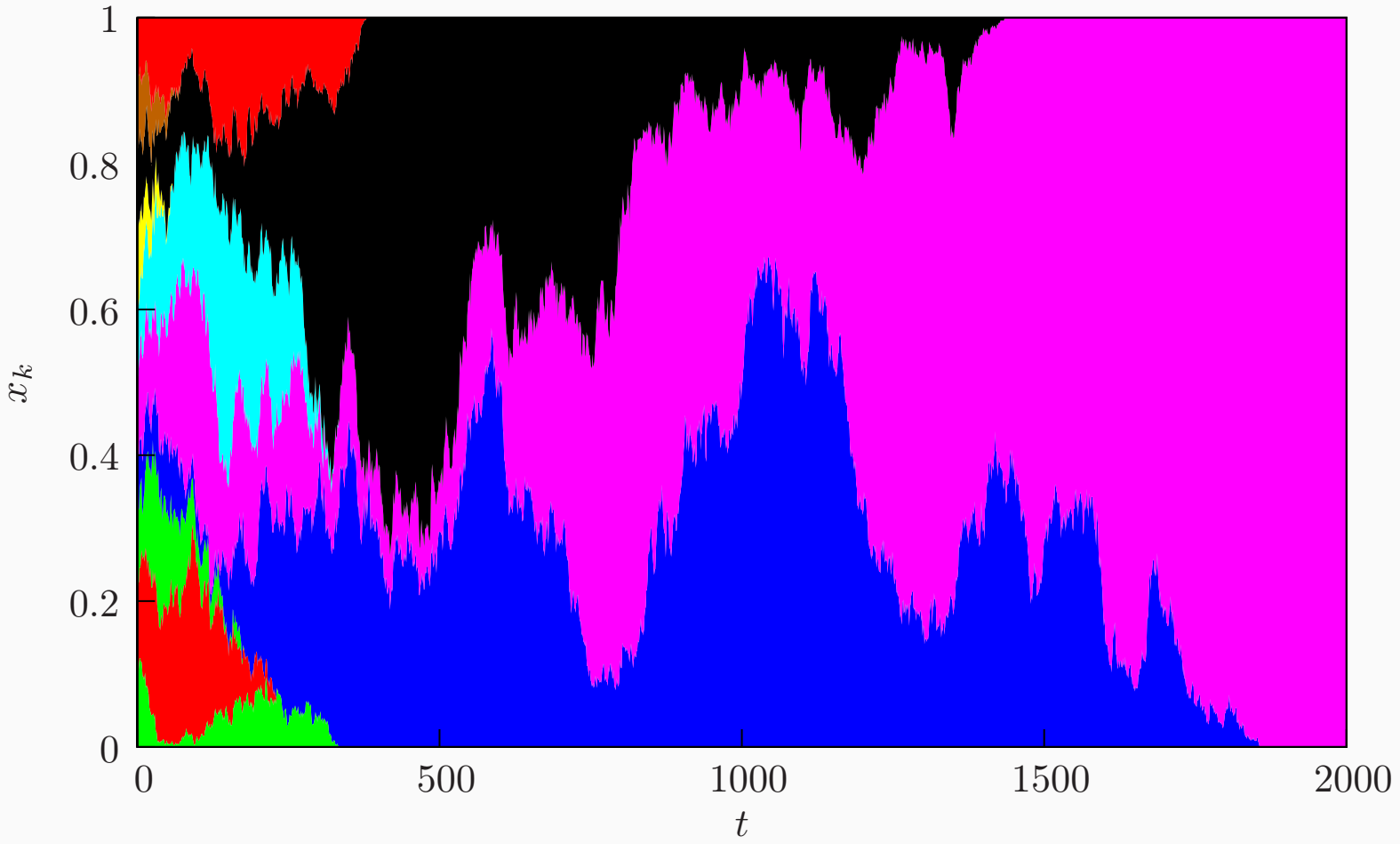
Finite population

The Wright-Fisher model



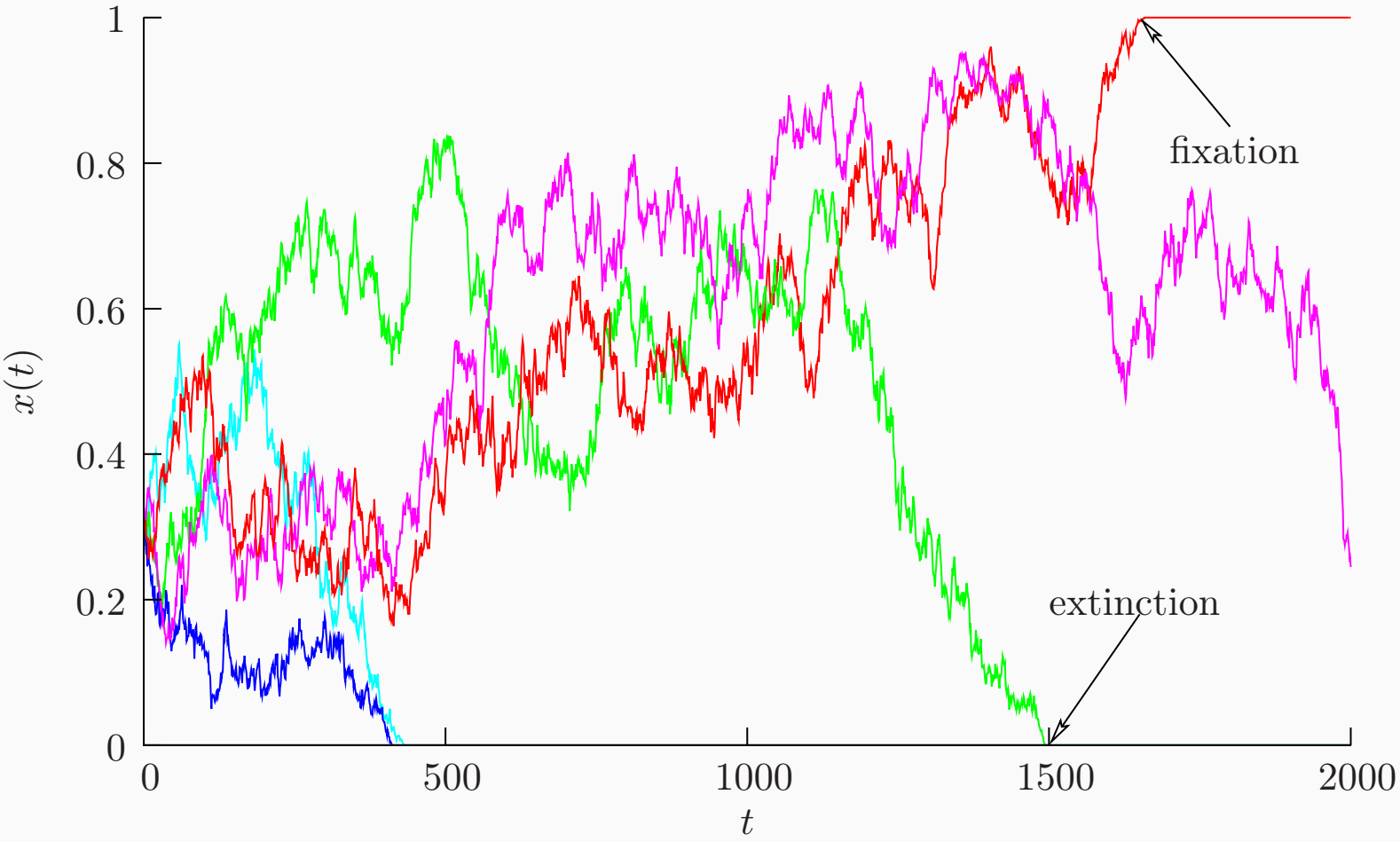
Finite population

The Wright-Fisher model: one realization (neutral)



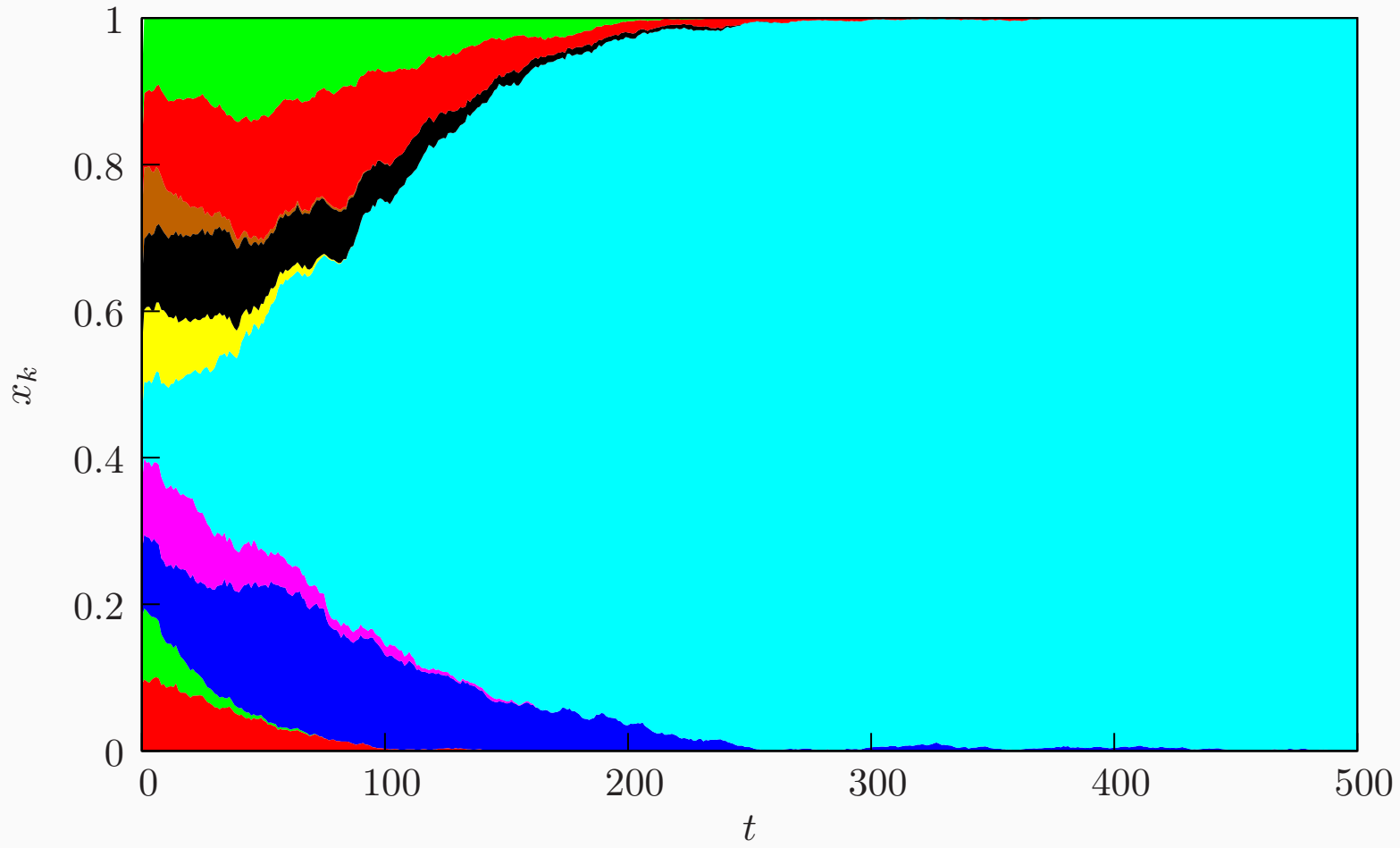
Finite population

The Wright-Fisher model: several realizations (neutral)



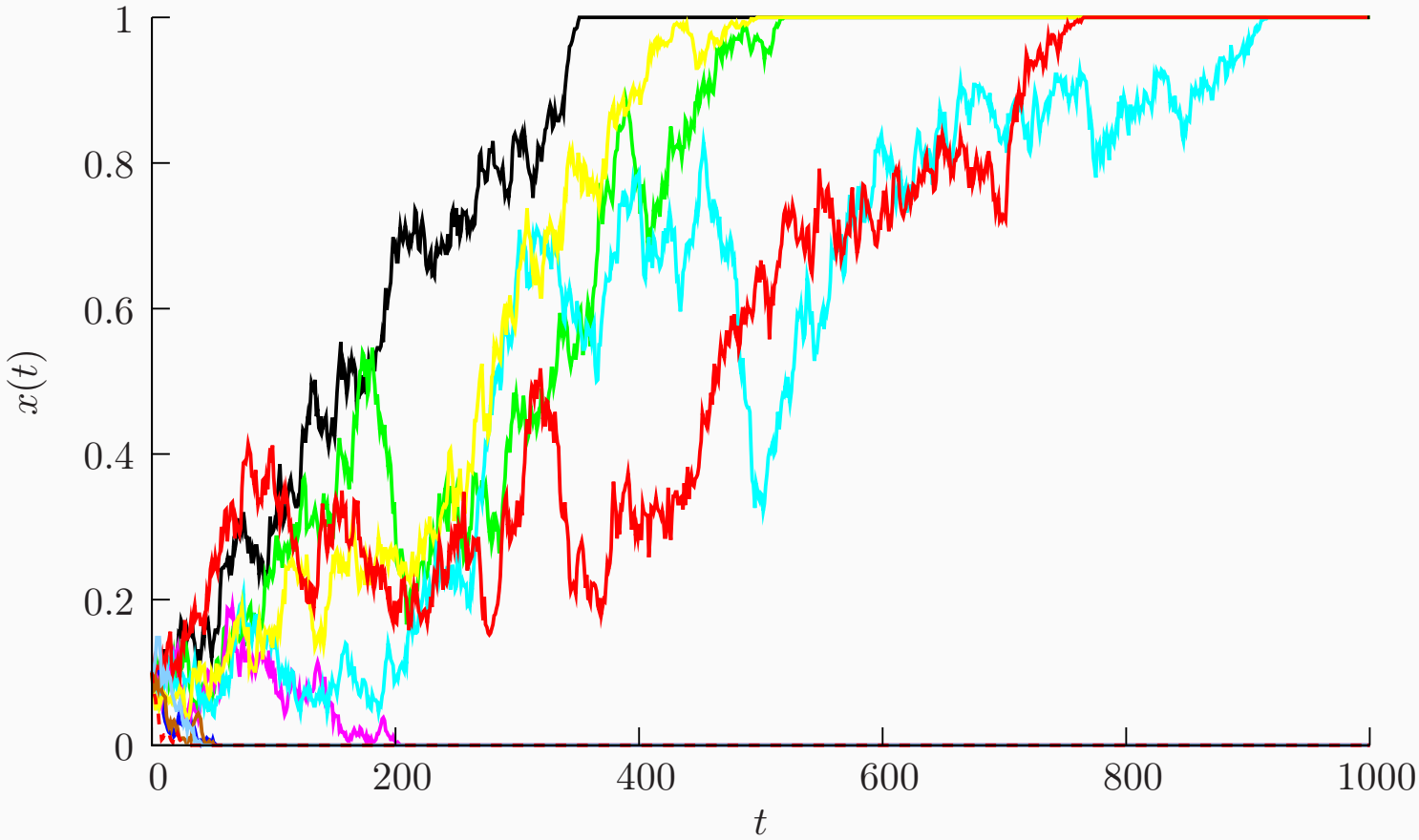
Finite population

The Wright-Fisher model: one realization (selective: $N = 10\,000$, $w_k \in \{1.0, 1.1\}$, $x_k(0) = 0.1$)



Finite population

The Wright-Fisher model: several realizations (selective: $N = 500$, $s = 0.01$, $x(0) = 0.1$)



Fixation in 5 cases out of 10

Drift

...it is often convenient to consider a natural population not so much as an aggregate of living individuals as an aggregate of gene ratios. Such a change of viewpoint is similar to that familiar in the theory of gases...

R. A. Fisher, 1953

Drift

We will start our discussion from the simplest situation where the gene frequency fluctuates from generation to generation because of the random sampling of gametes in a finite population. Since Wright's work, the term drift has become quite popular among biologists. However, in the mathematical theory of Brownian motion, the term drift originally connotes directional movement of the particle; therefore in our context the adjective random should be attached to it.

M. Kimura, 1964 (abridged)

Drift

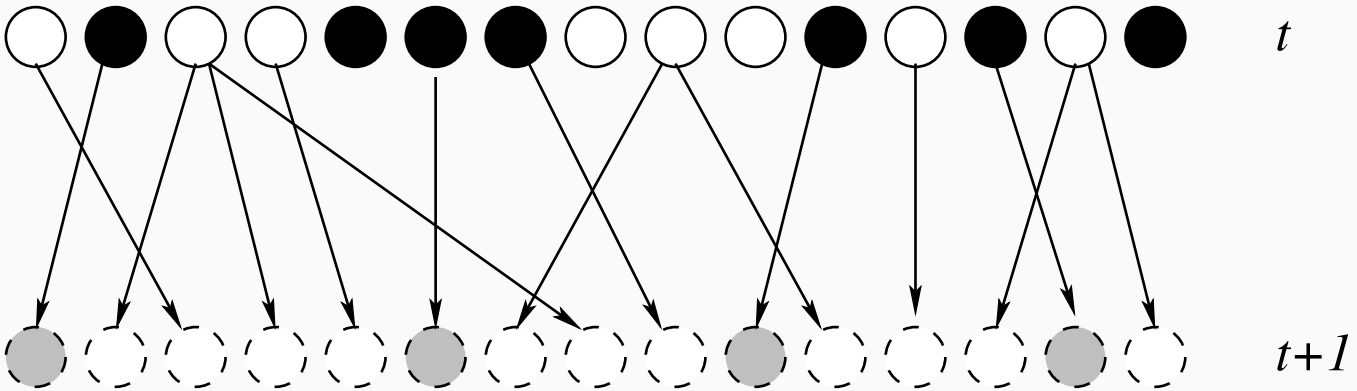
- Finite population implies different outcomes for different experiments in the same conditions (lack of self-averaging)
- Necessity to describe an ensemble of populations
- Use of the theory of Markov processes
- Simplification by means of diffusion equations

Random drift in the neutral case

- Population of N haploid individuals, 2 neutral alleles: A, a
- Frequency of the A allele: $x = n_{\mathrm{A}}/N$
- Wright-Fisher model: At each time step, each individual i of the new generation picks up a parent at random and copies it

Random drift in the neutral case

The Wright-Fisher model



Random drift in the neutral case

- Probability that $n_{\mathrm{A}}(t+1) = n$, given $n_{\mathrm{A}}(t) = Nx(t)$:

$$p_n(t+1) = \binom{N}{n} (x(t))^n (1-x(t))^{N-n}$$

- Assume $N \gg 1$, $\frac{1}{N} \ll x \ll 1 - \frac{1}{N}$, then

$$\mathrm{Prob}\left(x(t+1){=}x\right) \propto \exp\left(-\frac{(x-x(t))^2}{2Nx(t)(1-x(t))}\right)$$

- $\Delta x(t) = x(t+1) - x(t)$:

$$\langle \Delta x(t) \rangle = 0 \qquad \left\langle (\Delta x(t))^2 \right\rangle = \frac{x(t)(1-x(t))}{N}$$

The diffusion equation

Fokker-Planck equation:

$$\boxed{\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}\left(\langle \Delta x \rangle_x \, p(x,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\left\langle \Delta x^2 \right\rangle_x \, p(x,t)\right)}$$

In our case

$$\frac{\partial p}{\partial t} = \frac{1}{2N}\frac{\partial^2}{\partial x^2}\left(x(1-x)\, p(x,t)\right)$$

The solution in the neutral case

- Set $p(x,t \mid x_0,0) = \sum_n c_n(x_0)\chi_n(x)\,\mathrm{e}^{-\lambda_n t/(2N)}$
- Eigenvalue equation:

$$x(1-x)\chi_n''(x) + (1-2x)\chi_n'(x) + \lambda_n\chi_n(x) = 0$$

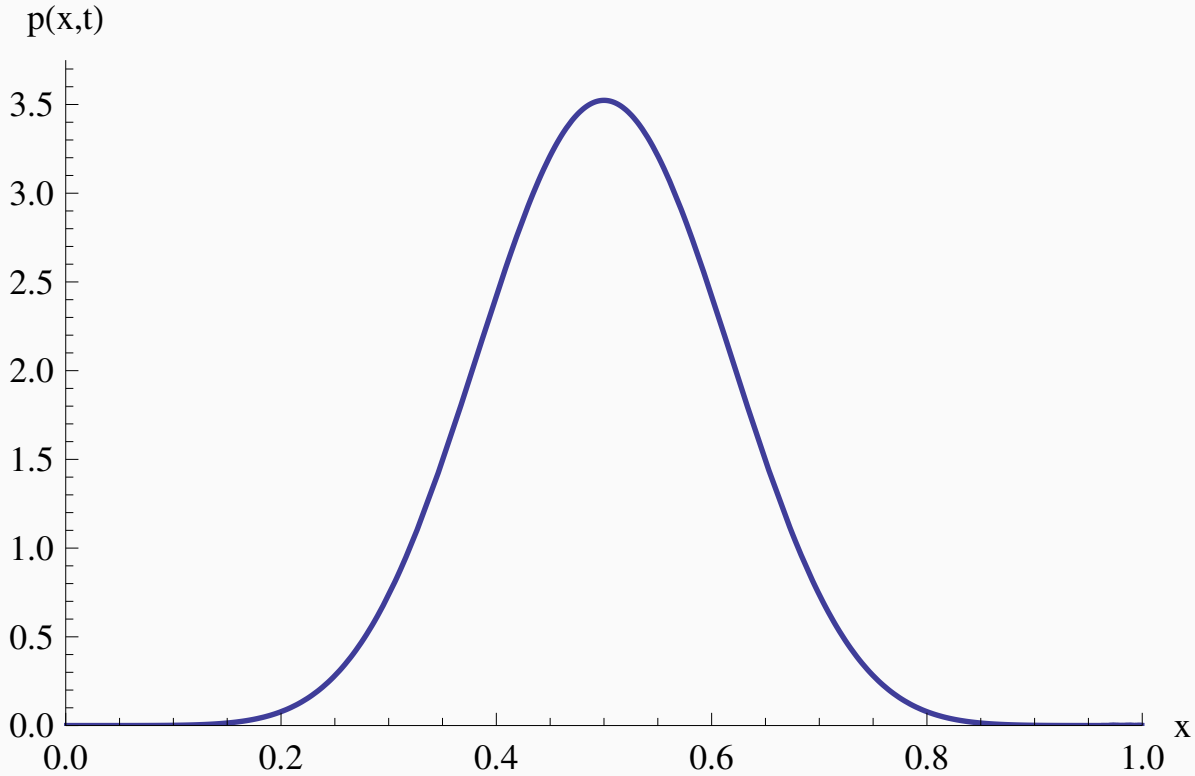
- Boundary conditions: $x = 0, 1$ are singular points; we require $\chi_n(0,1)$ finite $\forall n$
- Initial condition:

$$p(x,0 \mid x_0,0) = \sum_n c_n(x_0)\chi_n(x) = \delta(x-x_0)$$

- Solution in terms of hypergeometric functions:

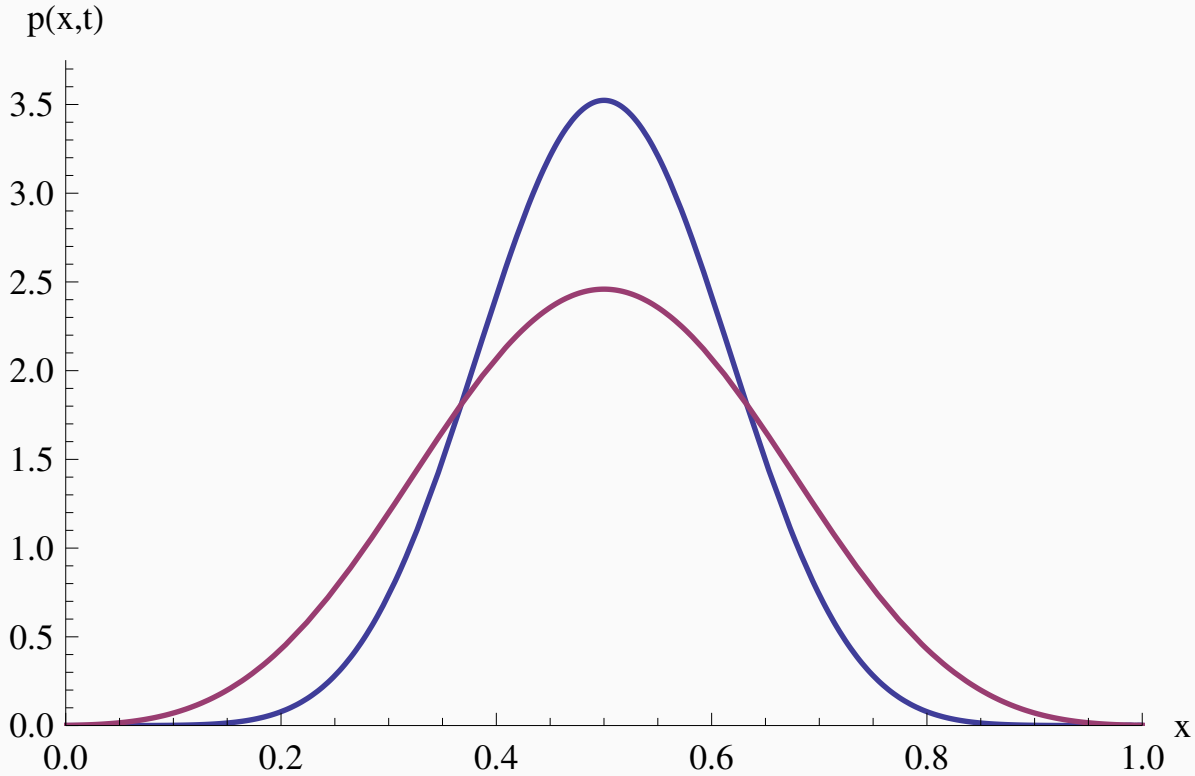
$$\chi_n(x) = F(1-n, n+2, 2, x) \qquad \lambda_n = n(n+1)$$

The solution in the neutral case



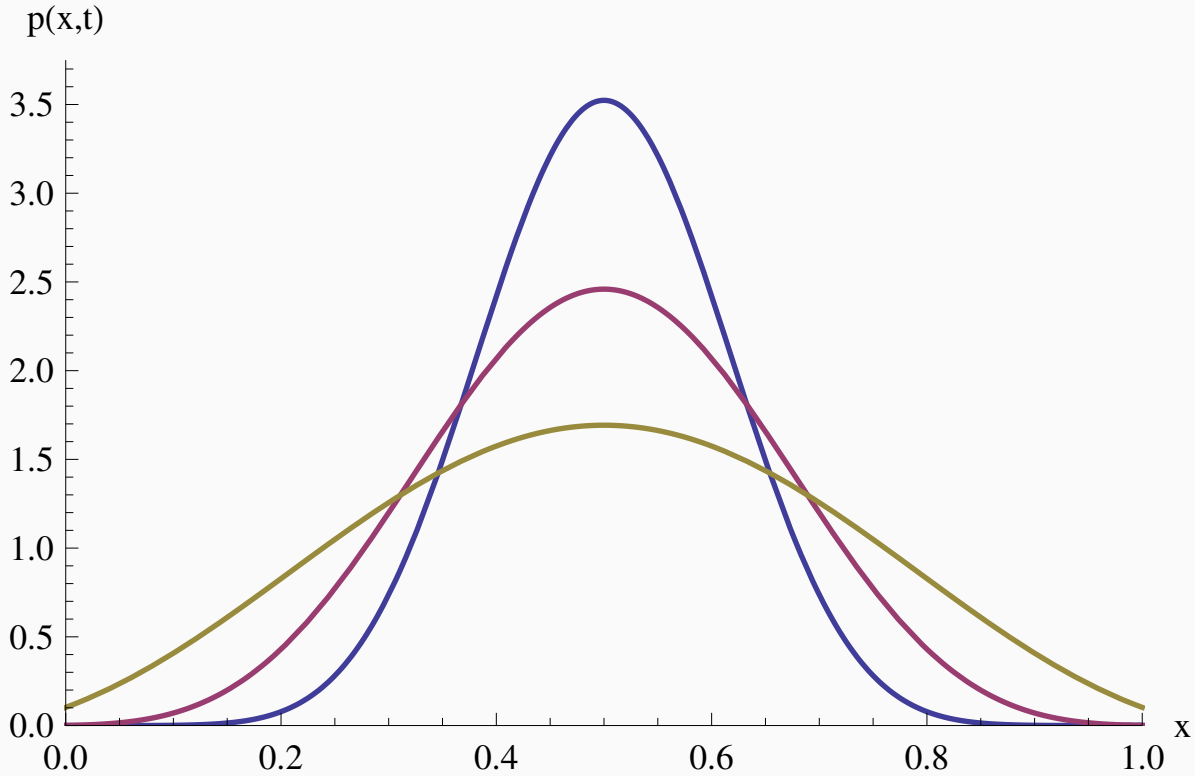
$t = 0.05N$

The solution in the neutral case



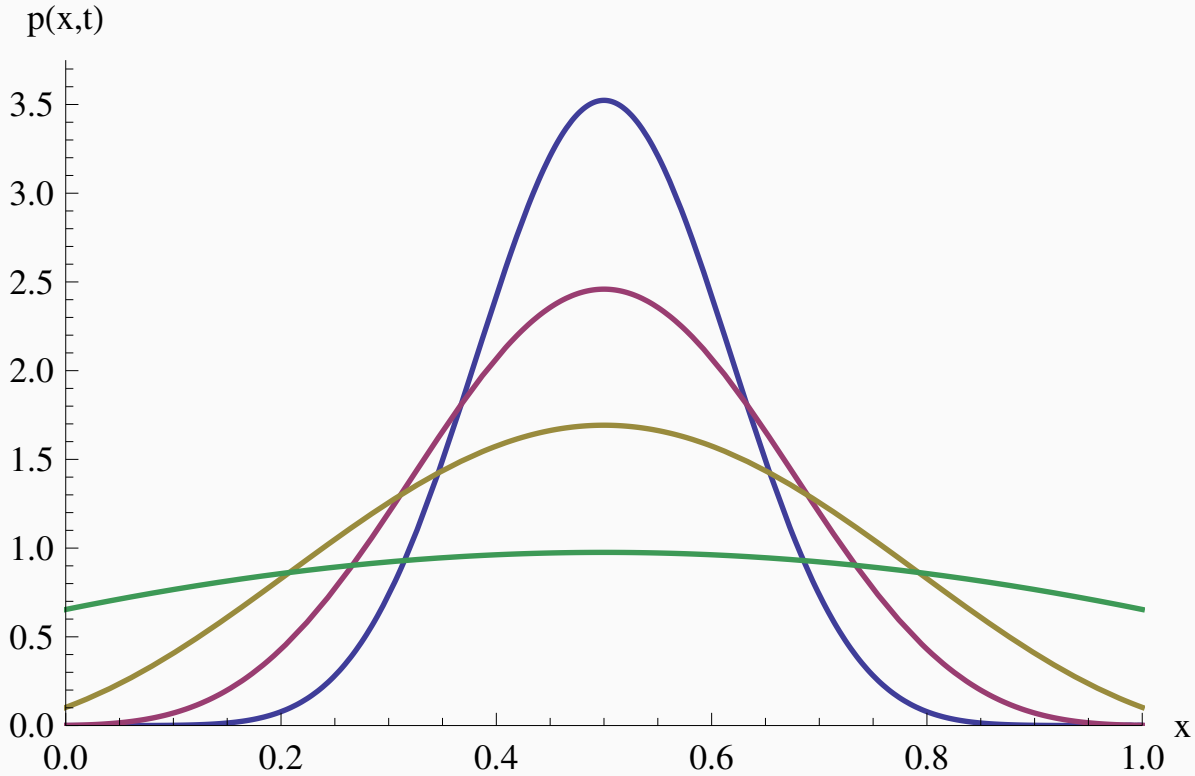
$t = 0.1N$

The solution in the neutral case



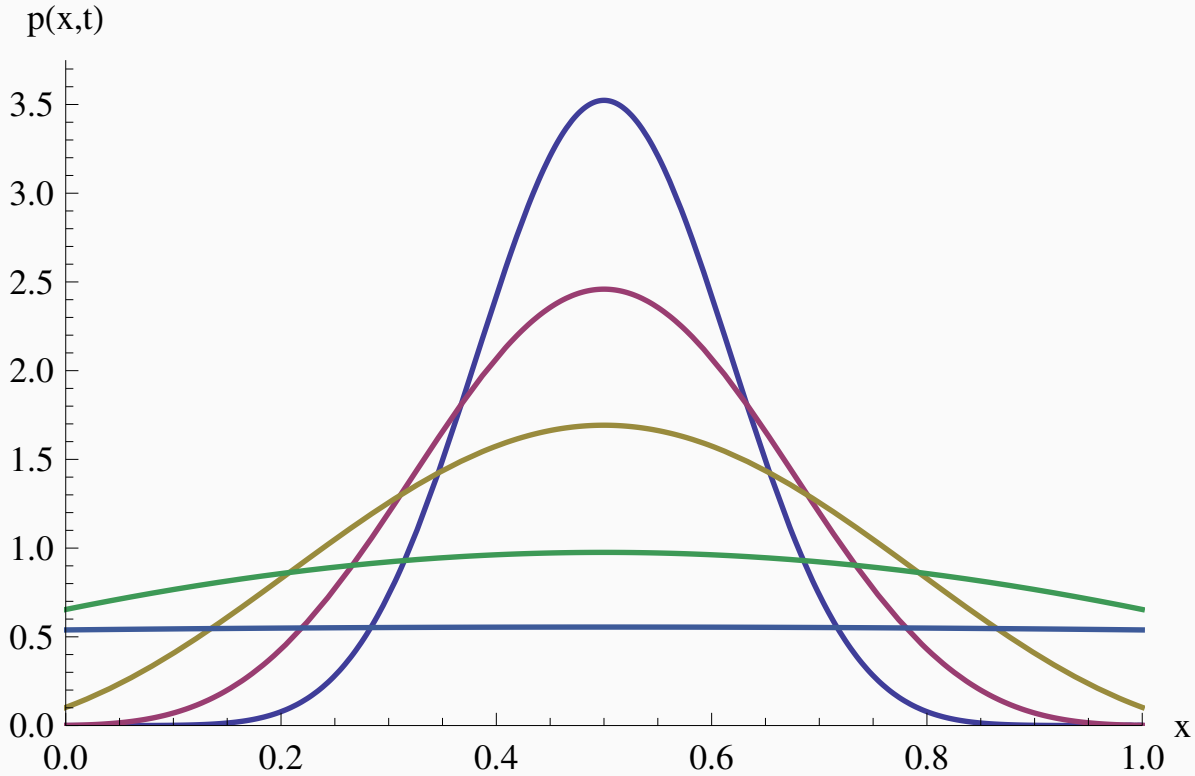
$t = 0.2N$

The solution in the neutral case



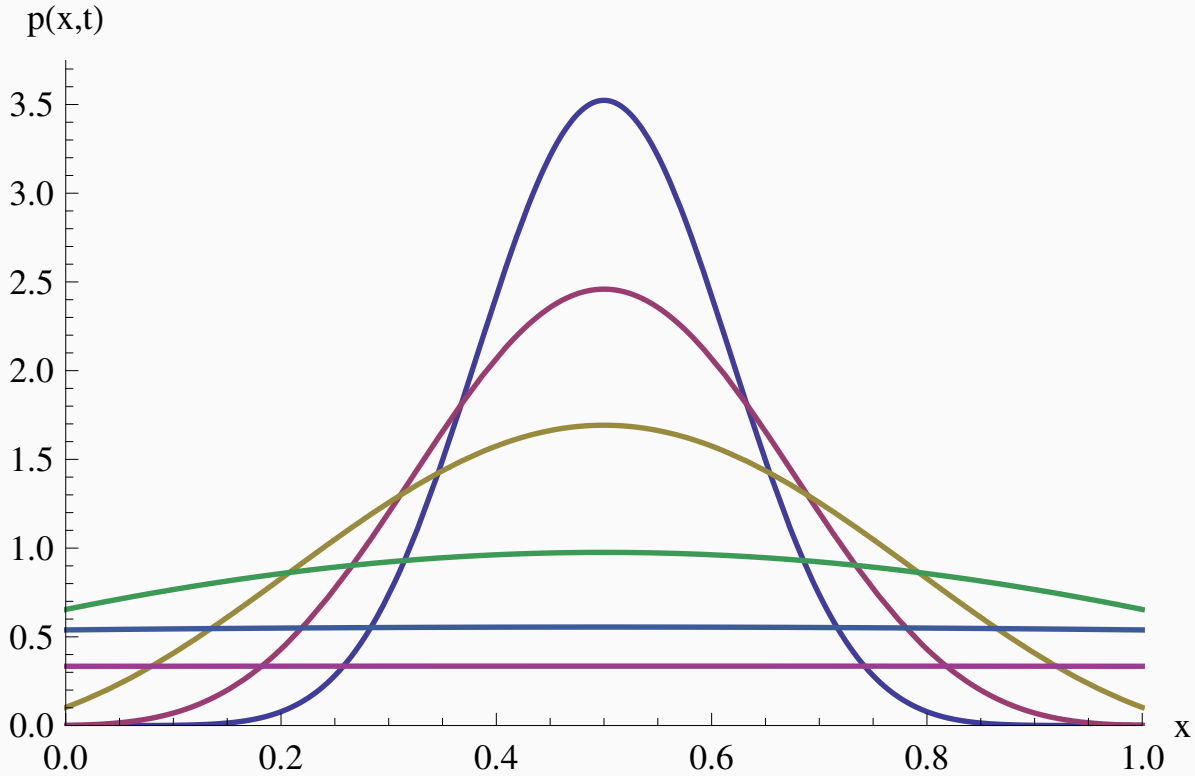
$t = 0.5N$

The solution in the neutral case



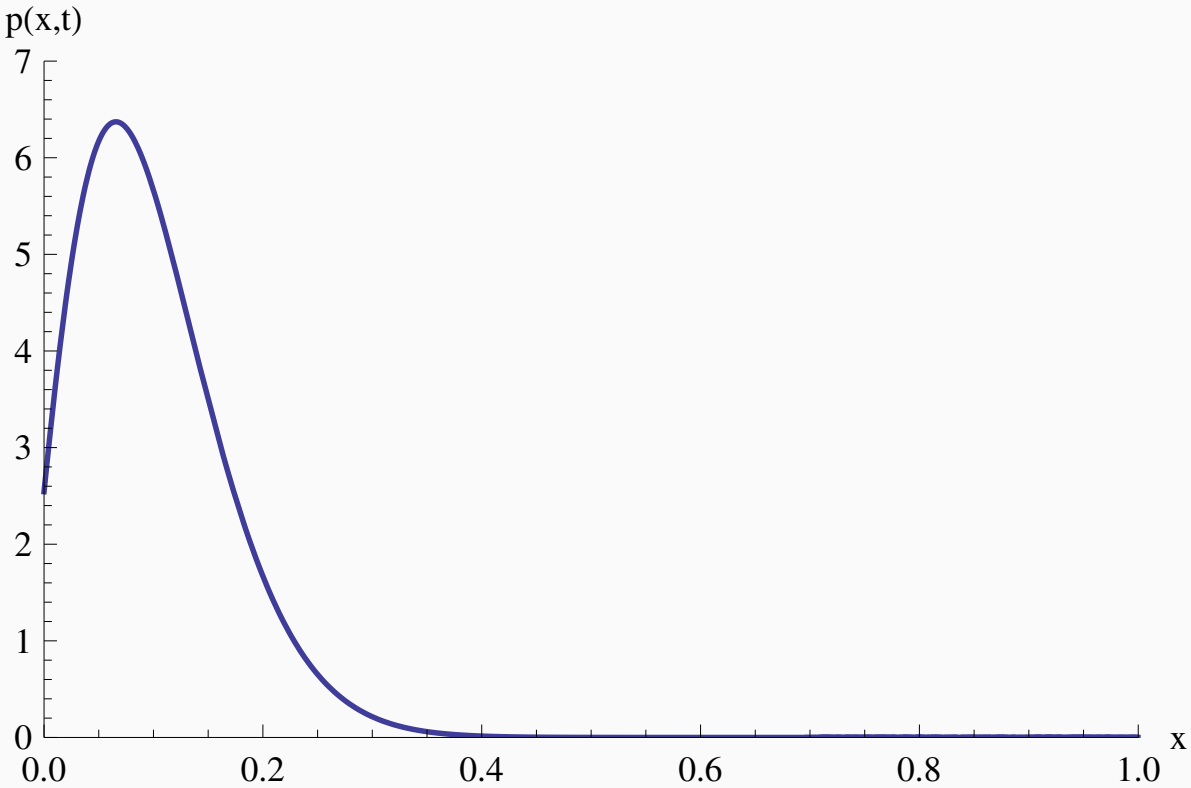
$t = N$

The solution in the neutral case



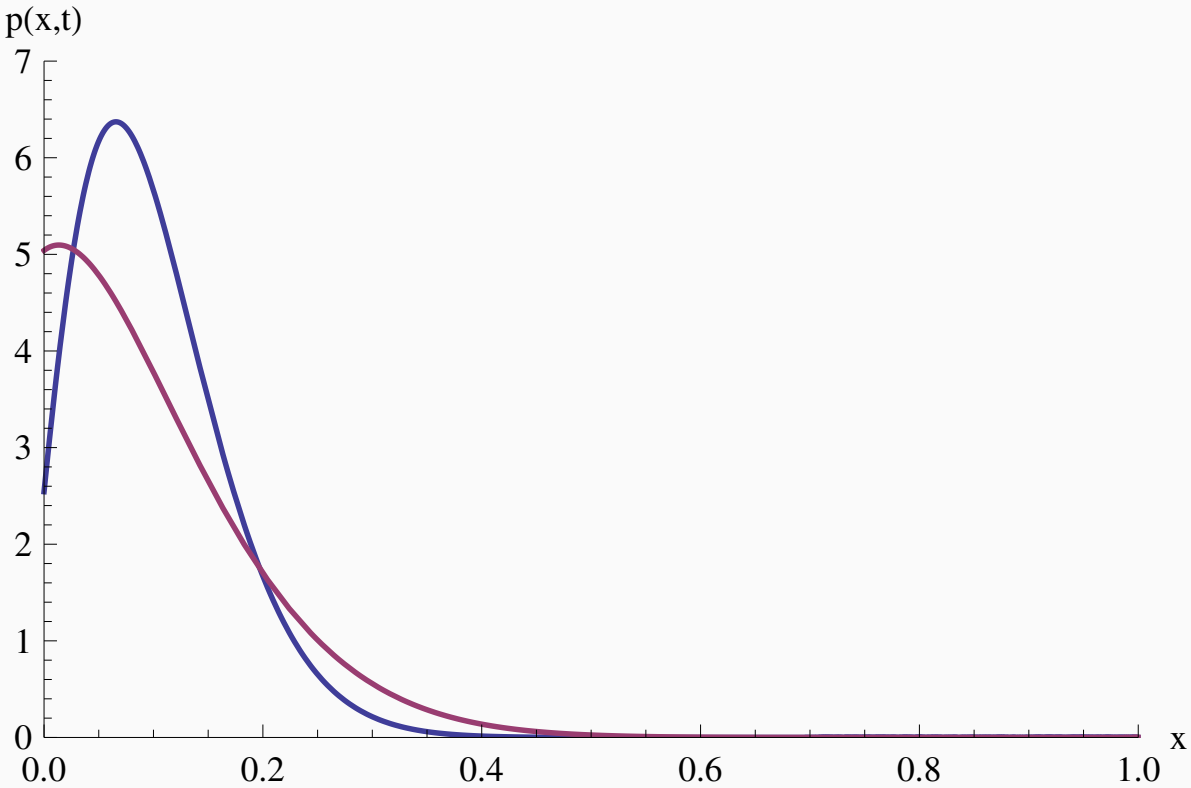
$t = 1.5N$

Initial condition $x(0) = 0.1$



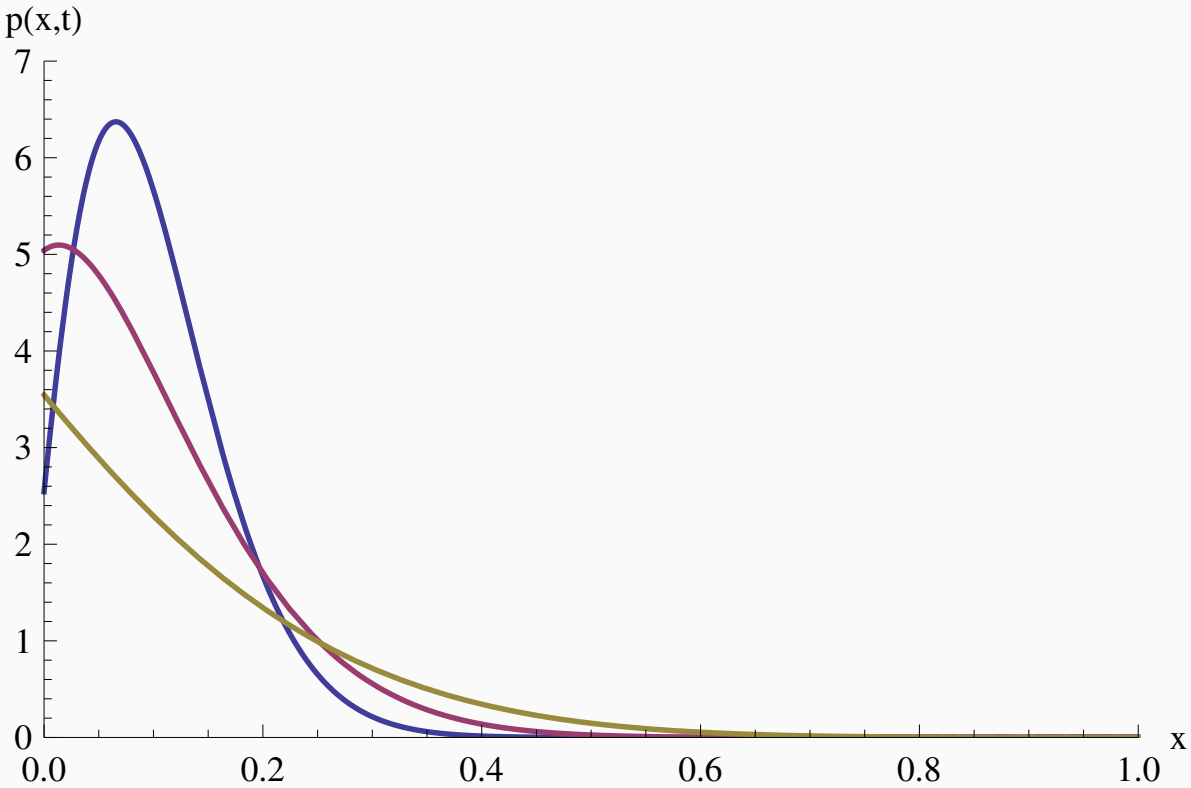
$t = 0.05N$

Initial condition $x(0) = 0.1$



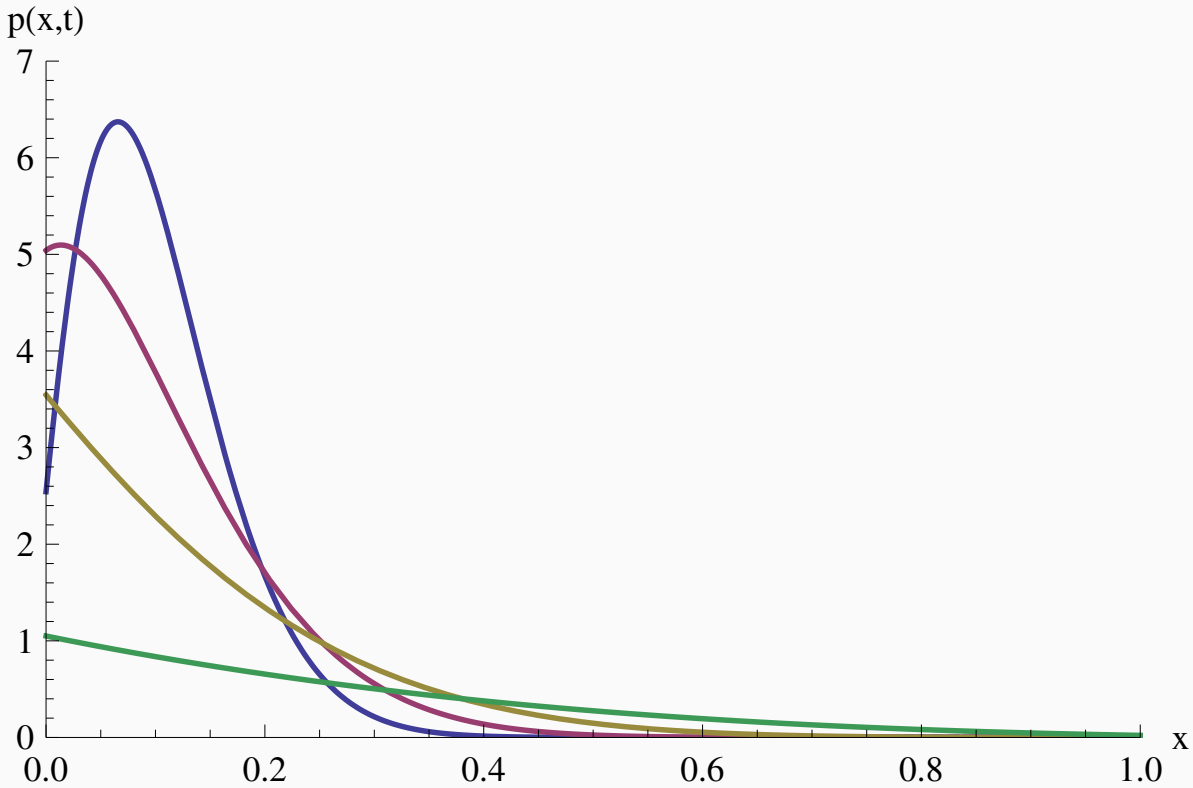
$t = 0.1N$

Initial condition $x(0) = 0.1$



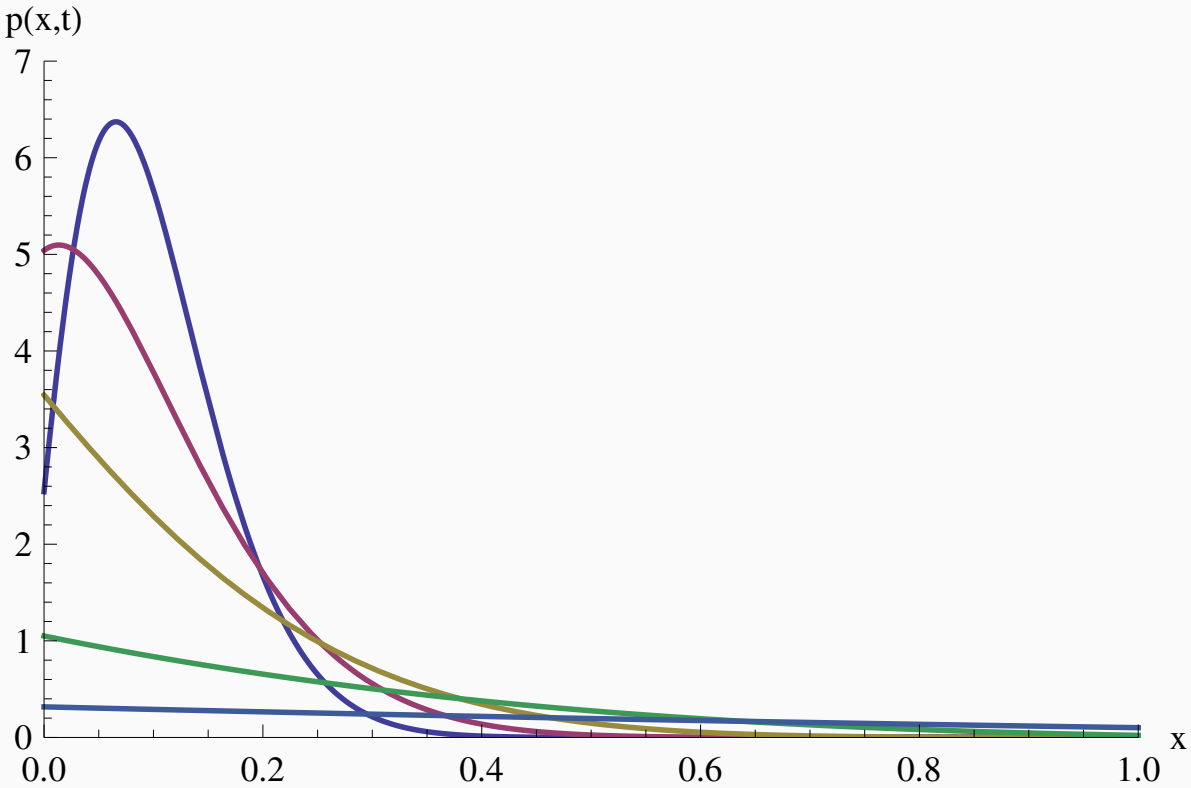
$t = 0.2N$

Initial condition $x(0) = 0.1$



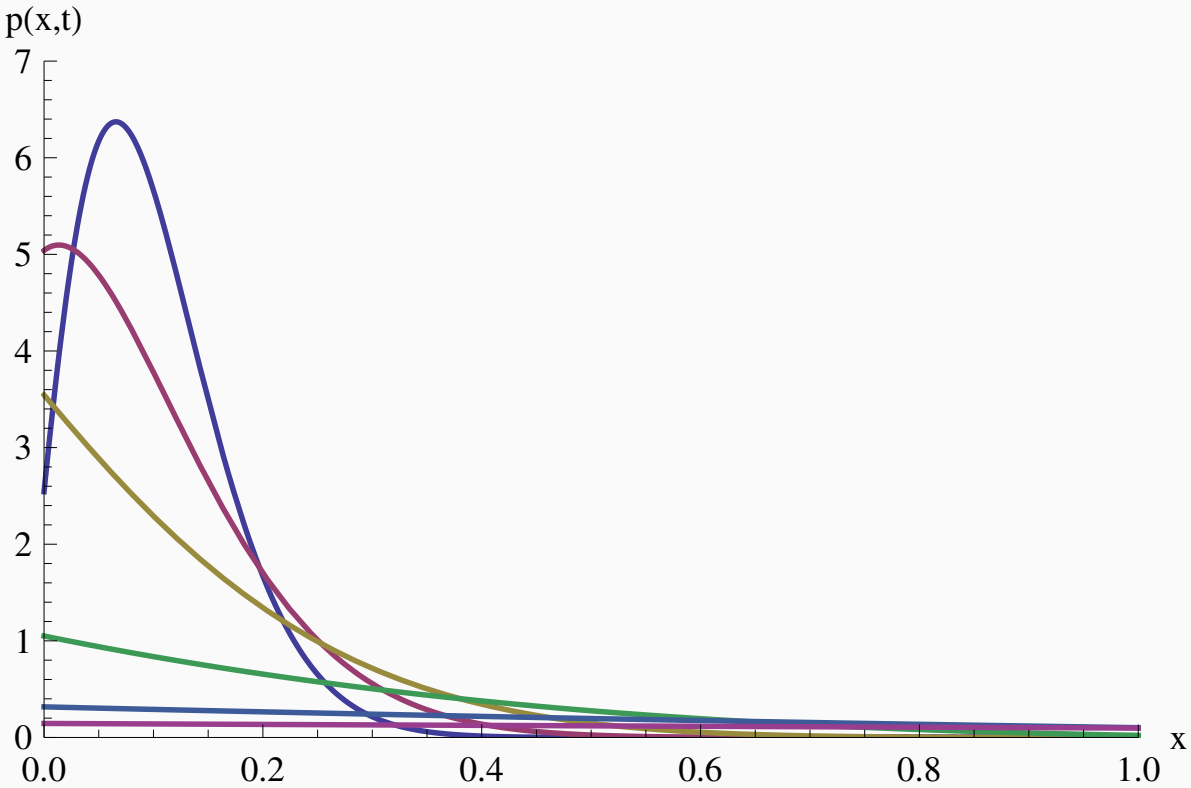
$t = 0.5N$

Initial condition $x(0) = 0.1$



$t = N$

Initial condition $x(0) = 0.1$



$t = 1.5N$

Results

- $p(x,t)$ decays exponentially: $p(x,t) \simeq 6x(0)(1-x(0))\mathrm{e}^{-t/N}$ for $t \gg N$
- Probability that A and a coexist at generation t: $\Omega(t) = \int_0^1 \mathrm{d}x\, p(x,t)$ decays with the same rate ($p(x,t)$ is flat)
- However, $p(x,t)$ becomes flat later when $x(0) \neq \frac{1}{2}$
- What is the probability of fixation of allele A as a function of $x(0)$?

The backward equation

- $p(x, t \mid x_0, t_0)$: Conditional probability that $x(t) = x$ given that $x(t_0) = x_0$
- Consider the effect of a single-generation sampling near t_0: $x(t_0 + 1) = x_0 + \Delta x_0$
- Equation for $p(x, t \mid x_0, t_0)$:

$$\boxed{-\frac{\partial p}{\partial t_0} = \langle \Delta x_0 \rangle_{x_0} \frac{\partial p}{\partial x_0} + \frac{1}{2} \langle \Delta x_0^2 \rangle_{x_0} \frac{\partial^2 p}{\partial x_0^2}}$$

- In our case

$$-\frac{\partial p}{\partial t_0} = \frac{x_0(1 - x_0)}{2N} \frac{\partial^2 p}{\partial x_0^2}$$

The fixation probability

- $P(t, x_0, t_0) = p(1, t \mid x_0, t_0)$: probability of being fixed by time t
- "Ultimate" fixation probability: $p^{\mathrm{fix}}(x_0) = \lim_{t\to\infty} P(t, x_0, t_0)$
- From the backward equation we obtain

$$\frac{\mathrm{d}^2 p^{\mathrm{fix}}}{\mathrm{d}x_0^2} = 0 \qquad x \in [0, 1]$$

- Boundary conditions: $p^{\mathrm{fix}}(x_0{=}0) = 0$ and $p^{\mathrm{fix}}(x_0{=}1)$
- Solution:

$$p^{\mathrm{fix}}(x_0) = x_0$$

Wright-Fisher model with selection

- Population of N haploid individuals, two alleles A and a
- Fitnesses: w_{A}, w_{a}
- Probability that an individual with allele A is chosen as a parent:

$$\xi_{\mathrm{A}} = \frac{n_{\mathrm{A}} w_{\mathrm{A}}}{\sum_{j=1}^{N} w_j} = \frac{n_{\mathrm{A}} w_{\mathrm{A}}}{n_{\mathrm{A}} w_{\mathrm{A}} + n_{\mathrm{a}} w_{\mathrm{a}}} = \frac{x\, w_{\mathrm{A}}}{x w_{\mathrm{A}} + (1-x) w_{\mathrm{a}}}$$

- Probability that $n_{\mathrm{A}}(t+1) = n$:

$$p_n(t+1) = \binom{N}{n} \xi_{\mathrm{A}}^{n} \left(1 - \xi_{\mathrm{A}}\right)^{N-n}$$

- Average and variance:

$$\begin{aligned} \langle x_{\mathrm{A}}(t+1) \rangle &= \xi_{\mathrm{A}} \\ \left\langle \left(x_{\mathrm{A}}(t+1) - \langle x_{\mathrm{A}}(t+1) \rangle \right)^2 \right\rangle &= \xi_{\mathrm{A}} \left(1 - \xi_{\mathrm{A}}\right) / N \end{aligned}$$

Selection and drift

> If the first human infant with a gene for levitation were struck by lightning in its pram, this would not prove the new genotype to have low fitness, but only that the particular child was unlucky.
>
> John Maynard Smith

Selection and drift

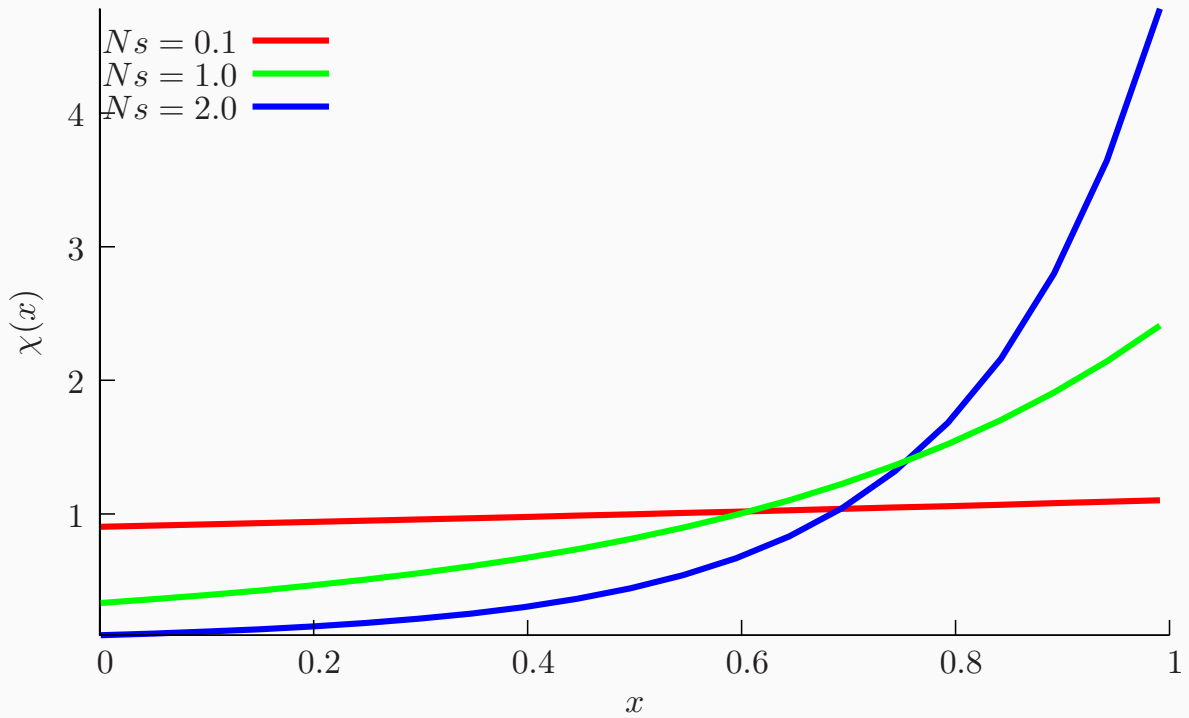
- Set $w_{\mathrm{A}} = 1 + s$, $w_{\mathrm{a}} = 1$, $s \ll 1$
- Then $\xi_{\mathrm{A}} = x w_{\mathrm{A}} / (x w_{\mathrm{A}} + w_{\mathrm{a}}(1-x)) = (1+s)x/(1+sx)$
- Then $\langle \Delta x \rangle_x = \langle x(t+1) \rangle - x = sx(1-x)/(1+sx) \simeq sx(1-x)$ and $\langle \Delta x^2 \rangle \simeq (x(1-x)/N)$
- Diffusion equation for $p(x,t)$:

$$\frac{\partial p}{\partial t} = -s \frac{\partial}{\partial x} \left(x(1-x)p \right) + \frac{1}{2N} \frac{\partial^2}{\partial x^2} \left(x(1-x)p \right)$$

- Solution in terms of spheroidal functions...
- Asymptotically $p(x,t) \propto \chi(x)\, \mathrm{e}^{-\lambda t/N}$

Solution with selection

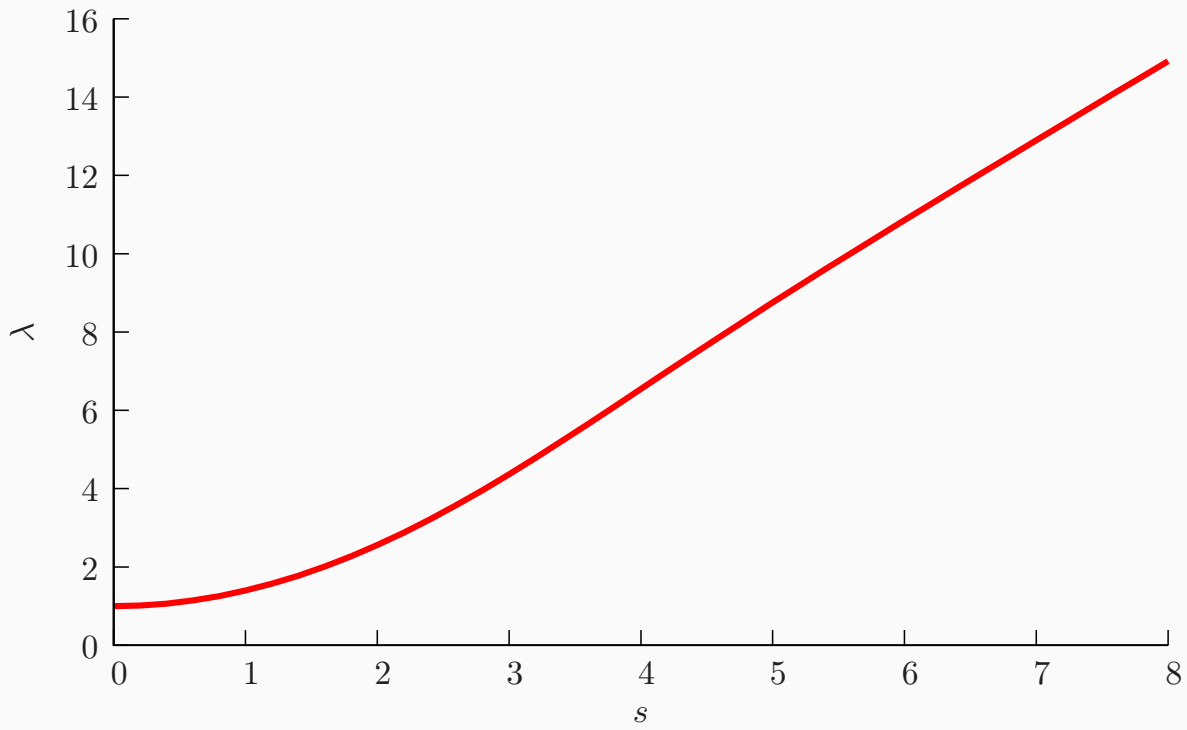
The long-living eigenfunction:



The leading eigenfunction $\chi(x)$ for several values of s

Solution with selection

The decay rate:



Leading eigenvalue λ as a function of Ns; decay rate: λ/N

The fixation probability with selection

- The backward equation:

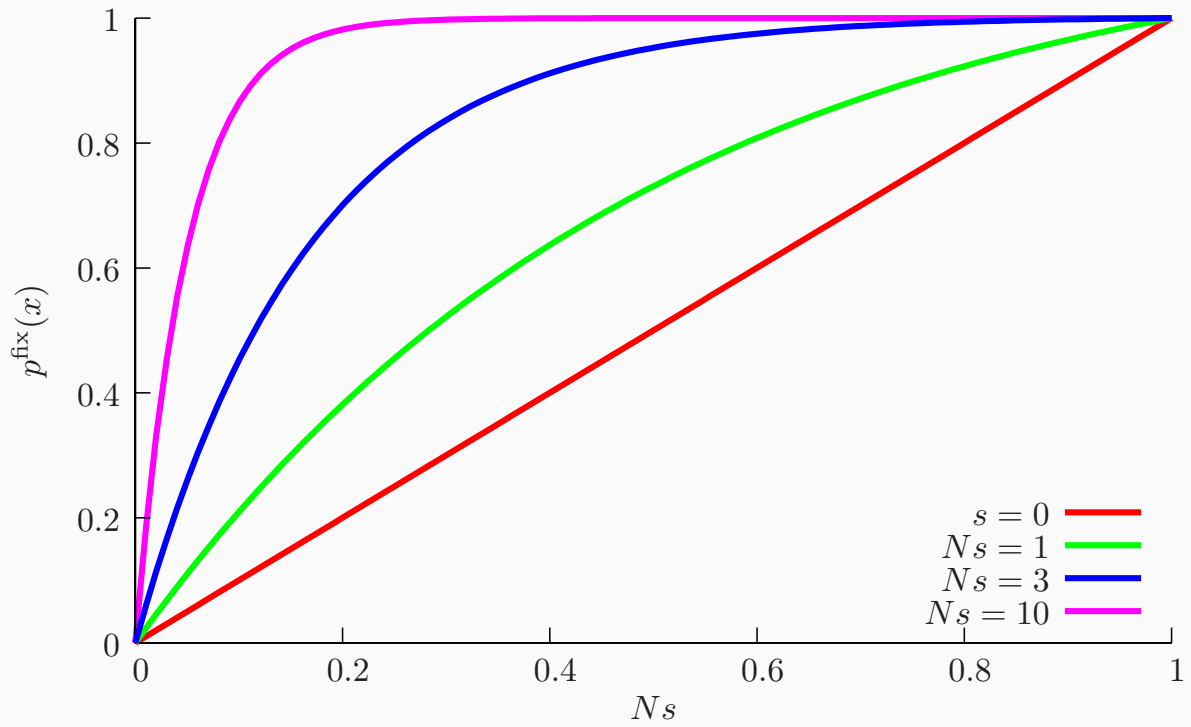
$$\frac{\partial p}{\partial t_0} = s x_0 (1 - x_0) \frac{\partial p}{\partial x_0} + \frac{x_0 (1 - x_0)}{2N} \frac{\partial^2 p}{\partial x_0^2}$$

- Stationary solution:

$$\begin{aligned} \frac{\partial p^{\text{fix}}}{\partial x_0} &= C_1 \mathrm{e}^{-2Nsx_0} \\ p^{\text{fix}}(x_0) &= C_0 - C_1 \mathrm{e}^{-2Nsx_0} \\ &= \frac{1 - \mathrm{e}^{-2Nsx_0}}{1 - \mathrm{e}^{-2Ns}} \end{aligned}$$

- In particular, for $s \to 0$, $p^{\text{fix}} \to x_0$

The fixation probability with selection



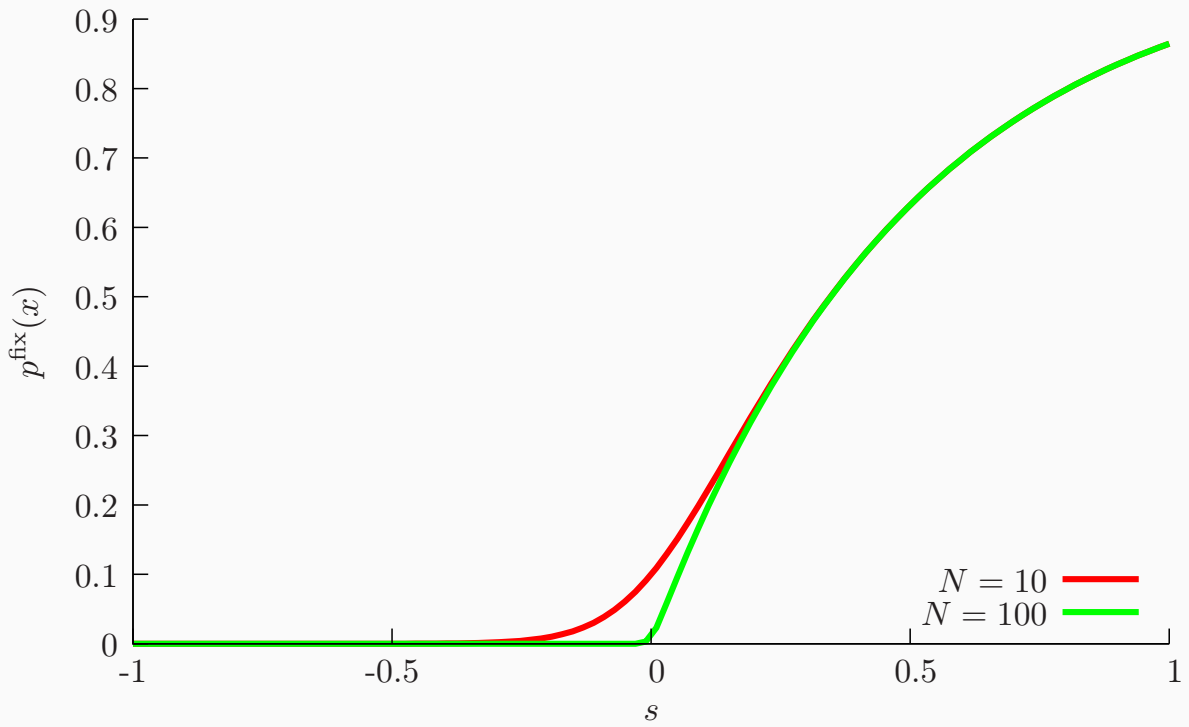
Fixation probability of a single mutant

- For a single mutant $x_0 = \frac{1}{N}$
- Thus

$$\boxed{p^{\mathrm{fix}} = \frac{1 - \mathrm{e}^{-2s}}{1 - \mathrm{e}^{-2Ns}}}$$

- Limits:
  - $s > 0$, $Ns \gg 1$: $p^{\mathrm{fix}} \simeq 1 - \mathrm{e}^{-2s}$ (for $s \ll 1$, $p^{\mathrm{fix}} \simeq 2s$)
  - $s < 0$, $|Ns| \gg 1$, $p^{\mathrm{fix}} \simeq 0$
  - $|Ns| \lesssim 1$, $p^{\mathrm{fix}} \simeq \frac{1}{N}$

Fixation probability of a single mutant



Frequency needed to obtain fixation

- How large must be x to be "almost sure" that a beneficial mutant fixes?
- Solve
$$p^{\mathrm{fix}}(x^*) = 1 - \gamma$$
- For $Ns \gg 1$ we have $p^{\mathrm{fix}}(x) \simeq 1 - \mathrm{e}^{-2Nsx}$, thus
$$x^* = -\frac{\log \gamma}{2Ns} \qquad \text{or} \qquad n^* = -\frac{\log \gamma}{2s}$$
- The fate of the mutant is determined in its initial phase, where it undergoes a branching process—the size of N is irrelevant!

Substitution rate

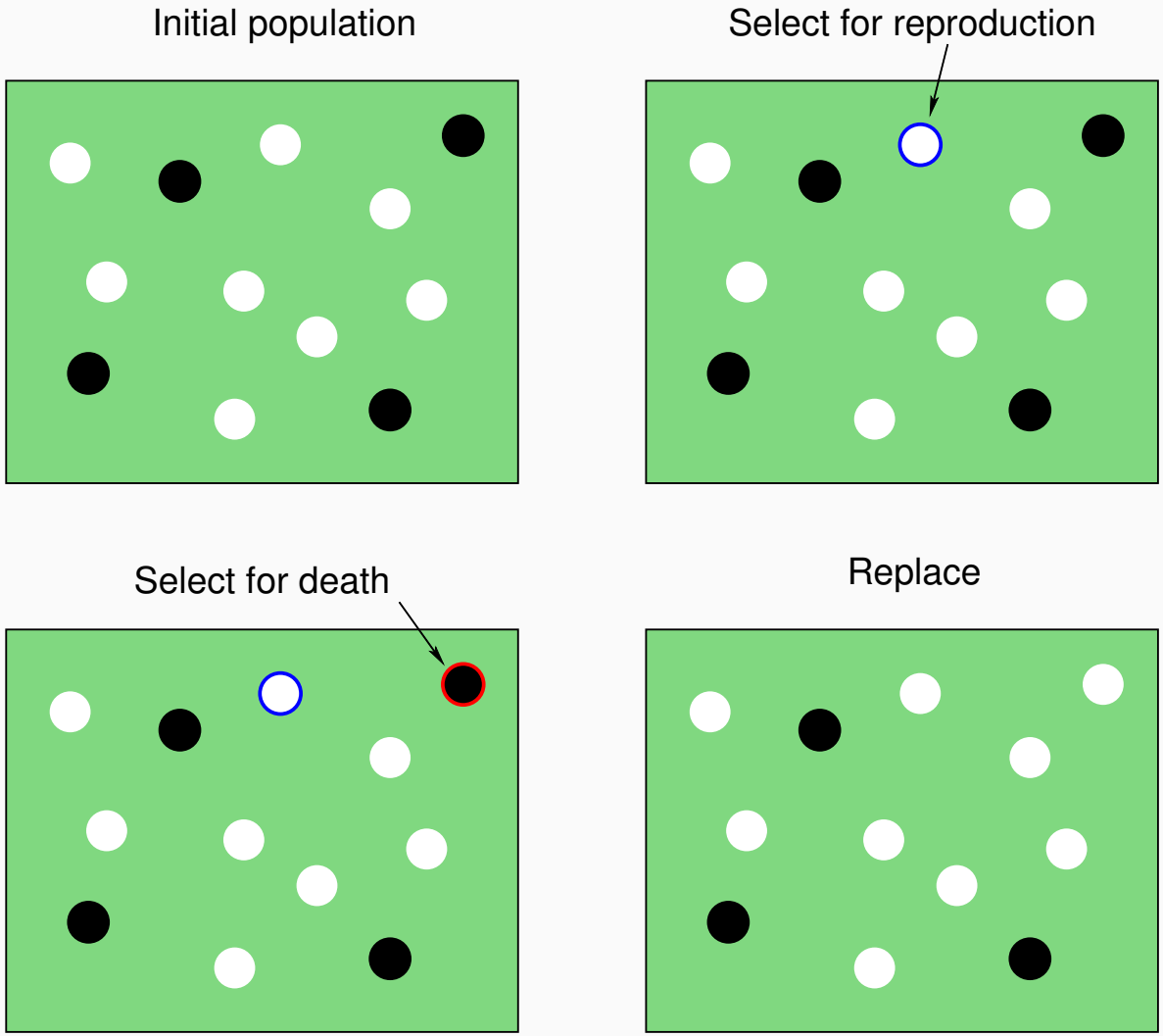
- For a new mutant, $x_0 = \frac{1}{N}$
- For a neutral mutant, $s = 0$, thus $p^{\text{fix}} = x_0 = \frac{1}{N}$
- If u is the mutation probability per genome and generation, the expected number of mutants per generations is uN
- Of these, only a fraction $\frac{1}{N}$ reaches fixation, i.e., produces a substitution
- Therefore the rate ν of neutral substitutions in a population with mutation rate u is equal to u:

$$\boxed{\text{substitution rate} = \text{mutation rate}}$$

independently of the population size N

The Moran model

Overlapping generations individual-based model:



The Moran model

- Selection: $p_{\mathrm{kill}}(\mathrm{A}) = 1 - s$, $p_{\mathrm{kill}}(\mathrm{a}) = 1$
- $\Delta t = \frac{1}{N}$; $\Delta n_{\mathrm{A}} \in \{-1, 0, +1\}$
- Probabilities:

$$
\begin{aligned}
P_{-1} &= \underbrace{\frac{n_{\mathrm{a}}}{N}}_{\mathrm{Prob}_{\mathrm{repr}}(a)} \underbrace{(1-s)\frac{n_{\mathrm{A}}}{N}}_{\mathrm{Prob}_{kill}(A)} \\
&= (1-s)x(1-x) \\
P_{+1} &= \frac{n_{\mathrm{A}}}{N}\frac{n_{\mathrm{a}}}{N} = x(1-x) \\
P_0 &= 1 - (P_{+1} + P_{-1})
\end{aligned}
$$

The Moran model

- Thus, for $\Delta t = \frac{1}{N}$, $s \ll 1$:

$$\begin{aligned} \langle \Delta n_{\mathrm{A}} \rangle &= P_{+1} - P_{-1} = sx(1-x) \\ \left\langle (\Delta n_{\mathrm{A}})^2 \right\rangle &= P_{+1} + P_{-1} = (2-s)x(1-x) \simeq 2\,x(1-x) \end{aligned}$$

- The diffusion equation for the Moran model:

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x}\left(sx(1-x)p\right) + \underbrace{\frac{1}{N}}_{=1/2N \text{ for WF}} \frac{\partial^2}{\partial x^2}\left(x(1-x)p\right)$$

- The devil (or God?) is in the details...

Adaptation and drift

Mustonen and Lässig, 2005–2010

Finite population of size N, r alleles, Moran model. Effects of mutation and selection:

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = \sum_k \Gamma_{jk} \frac{\partial \Phi}{\partial x_k}; \qquad \Phi = \langle f \rangle_x + \sum_\alpha \mu_\alpha \log x_\alpha$$

$$\Gamma_{jk}(\boldsymbol{x}) = \begin{cases} -x_j x_k, & \text{if } j \neq k \\ x_j(1-x_j), & \text{if } j = k \end{cases} \qquad \Gamma \text{ positive definite}$$

Adaptation and drift

Mustonen and Lässig, 2005–2010

- Random drift: $x \longrightarrow x + \xi$

$$\left\langle \xi^j \right\rangle_{\boldsymbol{x}} = 0; \qquad \left\langle \xi^j \xi^k \right\rangle = 2\frac{\Gamma_{jk}(\boldsymbol{x})}{N}$$

- Fokker-Planck equation for the pdf $P(\boldsymbol{x})$:

$$\begin{aligned} \frac{\partial P}{\partial t} &= \sum_{jk} \frac{\partial}{\partial x_j} \left[-\frac{\partial \Phi}{\partial x_k} \left(\Gamma_{jk} P \right) + \frac{1}{N} \frac{\partial}{\partial x_k} \left(\Gamma_{jk} P \right) \right] \\ &= \sum_{jk} \frac{\partial}{\partial x_j} \Gamma_{jk} \left(-\frac{\partial \tilde{\Phi}}{\partial x_k} P + \frac{1}{N} \frac{\partial P}{\partial x_k} \right) \end{aligned}$$

Adaptation and drift

Mustonen and Lässig, 2005–2010

- $\tilde{\Phi} = \Phi - \frac{1}{N} \log \det \Gamma$; $\qquad\qquad \det \Gamma = \prod_\alpha x_\alpha$
- Stationary solution:

$$\begin{aligned} P^{\mathrm{eq}}(\boldsymbol{x}) &\propto \mathrm{e}^{N\tilde{\Phi}} = (\det \Gamma)^{-1}\, \mathrm{e}^{N\Phi} = P_0\, \mathrm{e}^{N\langle f \rangle_{\boldsymbol{x}}} \\ P_0(\boldsymbol{x}) &\propto \prod_\alpha x^{-1+N\mu_\alpha} \end{aligned}$$

- Thus, for a static fitness function f,

$$[N \langle f \rangle_{\boldsymbol{x}}]^{\mathrm{eq}}_{\mathrm{av}} = \int \mathrm{d}\boldsymbol{x}\; P^{\mathrm{eq}}(\boldsymbol{x}) \log \frac{P^{\mathrm{eq}}(\boldsymbol{x})}{P_0(\boldsymbol{x})} = \underbrace{D_{\mathrm{KL}}\left(P^{\mathrm{eq}} \| \; P_0\right)}_{\text{Kullback-Leibler divergence}}$$

$$D_{\mathrm{KL}}(p\|q) = \sum_k p_k \log \frac{p_k}{q_k} \qquad (1)$$

cAMP-response protein binding loci in E. Coli

Mustonen and Lässig, 2005

- Factor binding sites are short DNA sequences which bind activating factors
- Small mutation rates: $\mu N \ll 1 \Rightarrow$ Population becomes monomorphic ($\boldsymbol{x} = (x_\alpha) \to \delta_{\alpha\beta}$)

$$p_\beta = \mathrm{Prob}\left(\boldsymbol{x} = \delta_{\alpha_\beta}\right) \propto \mathrm{e}^{N f_\beta}$$

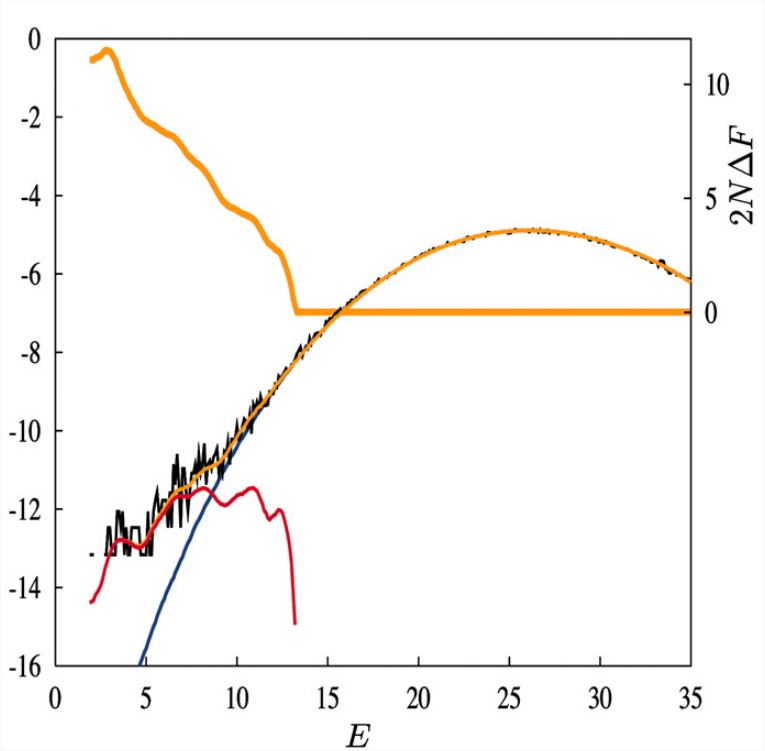
- It is reasonable to assume that their fitness depends on their binding energy E
- One can expect a linear model for $E(\sigma)$, $\sigma = (\sigma_1, \ldots, \sigma_\ell)$, $\sigma_i \in \{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$

$$E(\sigma) = \sum_{i=1}^{\ell} \epsilon_i(\sigma_i) \qquad \text{with } \epsilon_i(\sigma) = \epsilon_0 \log \frac{q_i(\sigma)}{p_0(\sigma)}$$

$p_0(\sigma)$: background nucleotide frequency

cAMP-response protein binding loci in E. Coli

Mustonen and Lässig, 2005



Log histogram $P(E)$ of binding energy E for 520 729 CRP-binding loci in E. Coli. Compared with $P(E) = (1-\lambda)P_0(E) + \lambda P_0(E)\mathrm{e}^{2NF(E)}$. The inferred form of $2NF(E)$ is also plotted. (W-F model)

Thank you!

References i

- G. H. Hardy, Mendelian proportions in a mixed population, Science **28** 49–50 (1908)
- W. Weinberg, Über den Nachweis der Vererbung beim Menschen, Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg **64** 368–382 (1908)
- B. de Finetti, Considerazioni matematiche sull'ereditarietà mendeliana, Metron **6** 29–37 (1926)
- John H. Gillespie, Population Genetics: A Concise Guide (2nd ed.) (Baltimore: Johns Hopkins Press, 2004)
- M. Eigen, Self-organization of matter and the evolution of biological macromolecules, Naturwissenschaften **58** 465–523 (1973)

References ii

- C. O. Wilke, Quasispecies theory in the context of population genetics, BMC Evolutionary Biology **5** 44 (2005)
- J. J. Bull and C. O. Wilke, Theory of Lethal Mutagenesis for Viruses, Journal of Virology **81** 2930–2939 (2006)
- J. J. Bull and C. O. Wilke, Lethal Mutagenesis of Bacteria, Genetics **180** 1061–1070 (2008)
- S. Crotty, C. E. Cameron and R. Andino, RNA virus error catastrophe: Direct molecular test by using ribavirin Proc. Natl. Acad. Sci. USA **98** 6895 (2001)
- M. Kimura, The neutral theory of molecular evolution (Cambridge: Cambridge U. P., 1983)
- R. A. Fisher, The Genetical Theory of Natural Selection (Oxford: Clarendon Press, 1930)

References iii

- R. A. Fisher, Population genetics, Proc. Roy. Soc. London, Ser. B **141** 510-523 (1953)
- S. Wright, Evolution in Mendelian populations, Genetics **16** 97-159 (1931)
- M. Kimura, Diffusion models in population genetics, J. Appl. Prob. **1** 177-232 (1964)
- J. Maynard Smith, Evolutionary Genetics (Oxford: Oxford U. P., 1989)
- P. A. P. Moran, Random processes in genetics, Mathematical Proceedings of the Cambridge Philosophical Society **54** 60-71 (1958)
- T. Ohta, Population size and rate of evolution, J. Mol. Evol. **1** 305-314 (1972)

References iv

- V. Mustonen and M. Lässig, From fitness landscapes to seascapes: Non-equilibrium dynamics of selection and adaptation, Trends Genet **25** 111–119 (2009)
- V. Mustonen and M. Lässig, Fitness flux and ubiquity of adaptive evolution, Proc. Natl. Acad. Sci. USA **107** 4248–4253 (2010)