BIO-463
Genomics and bioinformatics

Lecture 5:
Population genetics: mutations, selection & drift

Prof. Anne-Florence Bitbol

EPFL

Schedule of this class

Lecture 1: Feb 18
Lecture 2: Feb 25
Lecture 3: March 4
Lecture 4: March 11
Lecture 5: March 18 – Assignment 1 available on March 20
Lecture 6: March 25 – Problem class devoted to **assignment 1**; deadline on **March 28**
Lecture 7: April 1
Lecture 8: April 8
Lecture 9: April 15 – Assignment 2 available on April 18
Lecture 10: April 29 – Problem class devoted to **assignment 2**; deadline on **May 2**
Lecture 11: May 6 – Mini-projects available on April 28; choose yours by May 6
Lecture 12: May 13
Lecture 13: May 20
Lecture 14: May 27 – Mini-project deadline on **May 30**

Information about the first assignment

- **Assignment released on March 20, problem class on March 25, deadline on March 28**

This assignment is a graded problem set, and will count for **25% of your final grade**

You can discuss with TAs and with fellow students about the problem set, but in the end, you should hand in a **personal solution**
Detected plagiarism will result in a reduction of your grade

The expected language is R, as in all the class BIO-463 – some R functions and libraries may be recommended in the problems

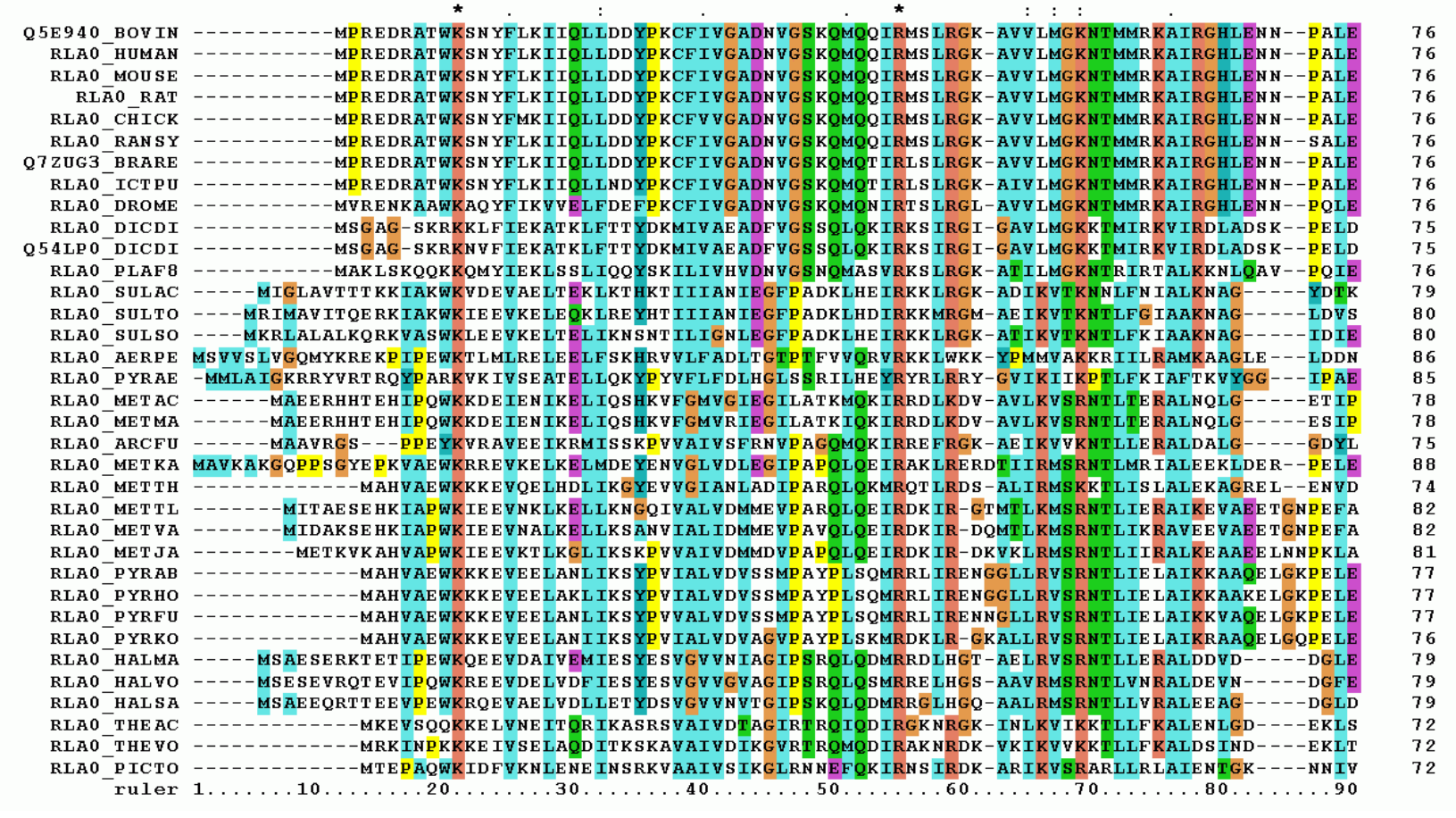
Please hand in your solution in two files:
• one should be your **source file**, with the following format: **.Rmd, .qmd**
• the other one should be the **html file** deriving from your source file

You will have to hand in your solution **via Moodle by Friday March 28**

Reminder: sequence data

- **Multiple sequence alignments**

Focus on amino-acid sequences of proteins (translated from the coding part of genomes)



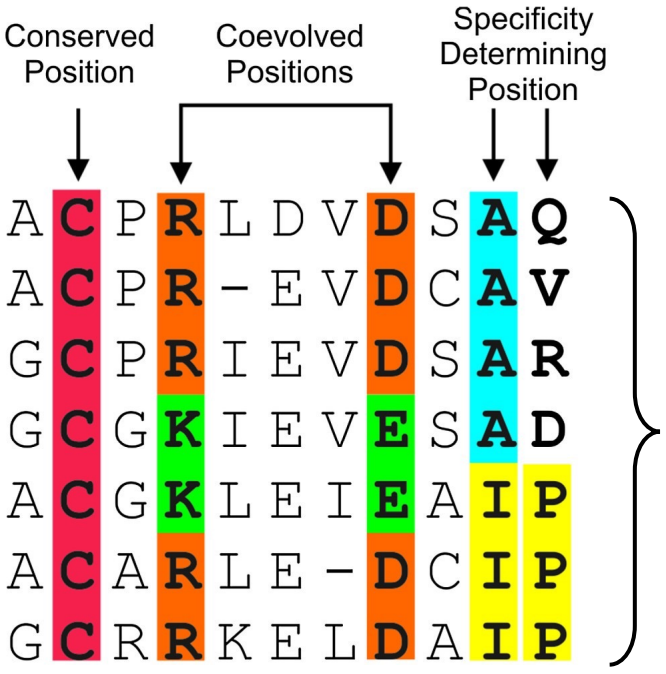
Acidic ribosomal protein P0 (first 90 positions) from several organisms

Row = sequence
Column = site (given position in 3D structure)

Colors = level of conservation

Reminder: sequence data

- **Multiple sequence alignments**



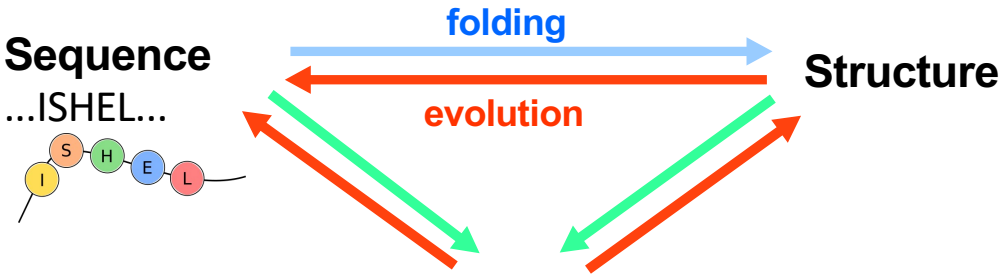
Multiple sequence alignment
of homologous protein sequences:
same ancestry,
similar function,
similar 3D structure

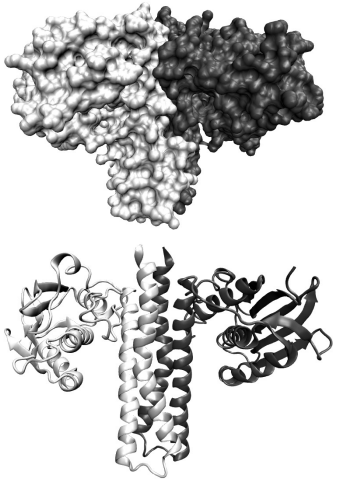
Teppa et al 2012

Special sites (e.g. highly conserved ones): signature of natural selection on these sites
We only observe sequences that have survived natural selection

Protein sequences and natural selection

- **Evolution of proteins**





Mutations act on sequences
BUT
selection acts on function

- Heteropolymers made of 20 types of amino-acids (monomers) → ~20^{100} possible proteins
- A given natural protein folds into a compact and (almost) unique 3D **structure**
- It has specific **interactions** with other molecules → **function**

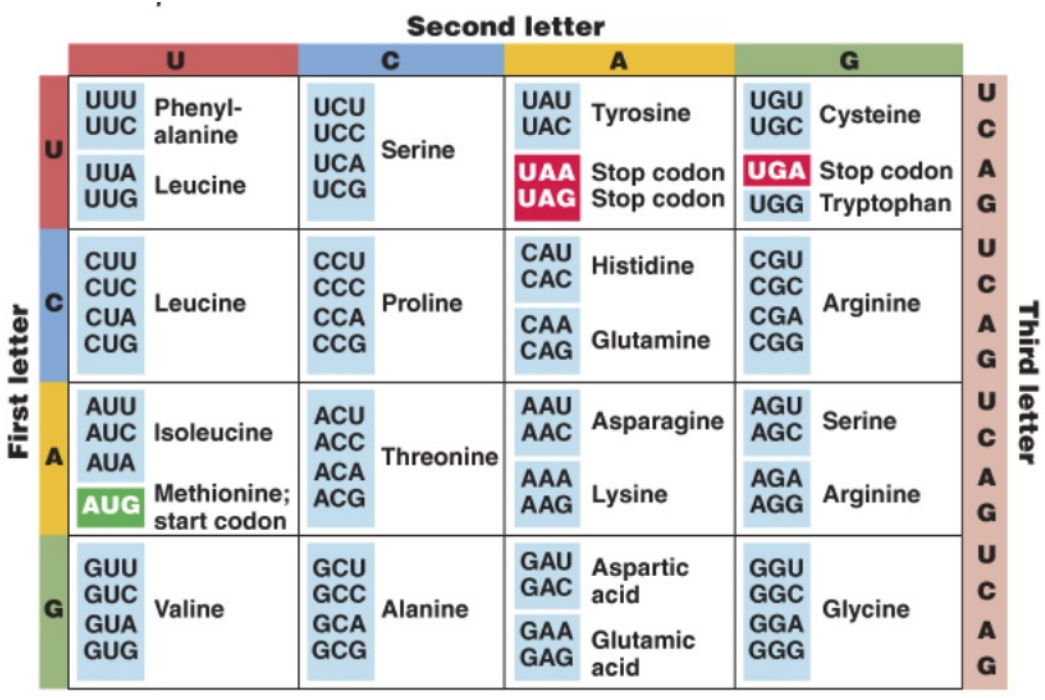
- Experiment: random proteins do not fold properly Socolich et al. (2005)

→ Natural proteins are special, due to natural selection for folding and function

Protein sequences and natural selection

- **A way to detect selection: dN/dS**

The genetic code has some redundancies:



First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC: Phenylalanine	UCU, UCC, UCA, UCG: Serine	UAU, UAC: Tyrosine	UGU, UGC: Cysteine	U C A G
	UUA, UUG: Leucine		UAA: Stop codon; UAG: Stop codon	UGA: Stop codon; UGG: Tryptophan	
C	CUU, CUC, CUA, CUG: Leucine	CCU, CCC, CCA, CCG: Proline	CAU, CAC: Histidine	CGU, CGC, CGA, CGG: Arginine	U C A G
			CAA, CAG: Glutamine		
A	AUU, AUC, AUA: Isoleucine	ACU, ACC, ACA, ACG: Threonine	AAU, AAC: Asparagine	AGU, AGC: Serine	U C A G
	AUG: Methionine; start codon		AAA, AAG: Lysine	AGA, AGG: Arginine	
G	GUU, GUC, GUA, GUG: Valine	GCU, GCC, GCA, GCG: Alanine	GAU, GAC: Aspartic acid	GGU, GGC, GGA, GGG: Glycine	U C A G
			GAA, GAG: Glutamic acid		

Some mutations are synonymous
→ they do not impact the protein sequence

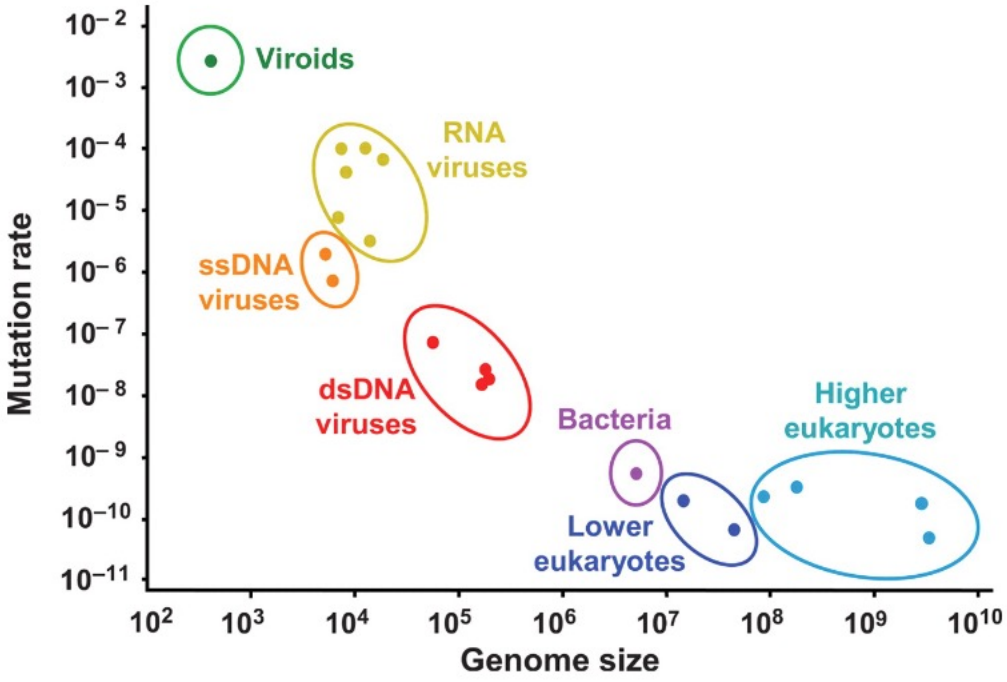
Selection can be assessed by comparing the rate of non-synonymous and synonymous mutations
Known as dN/dS (or Ka/Ks or ω)

dN/dS < 1 → selection to stay the same
dN/dS = 1 → no selection
dN/dS > 1 → selection to change

Mutations

- **How frequent are mutations?**

Mutation rates can be measured by the fluctuation test, inspired by the Luria-Delbrück experiment
They can also be measured by sequencing



Gago et al 2009

Mutation probabilities per base pair per replication (substitutions only)

Viruses have high mutation probabilities (~10^{-5})

Bacteria and eukaryotes have lower ones (~10^{-9})
Proofreading and error correction mechanisms allow to reach such low values

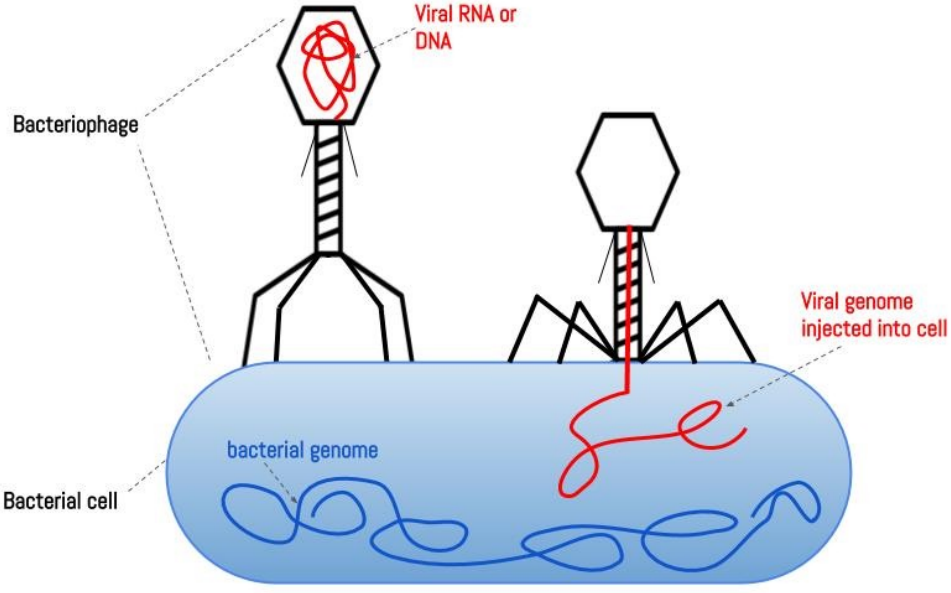
Mutations: Luria-Delbrück experiment

- **Luria-Delbrück experiment (1943)**

Phage and bacteria
(phage T1, obligately
lytic virus of *E. coli*)

By random mutations, bacteria can
develop resistance to phage infection

These mutants survive exposure
to phage

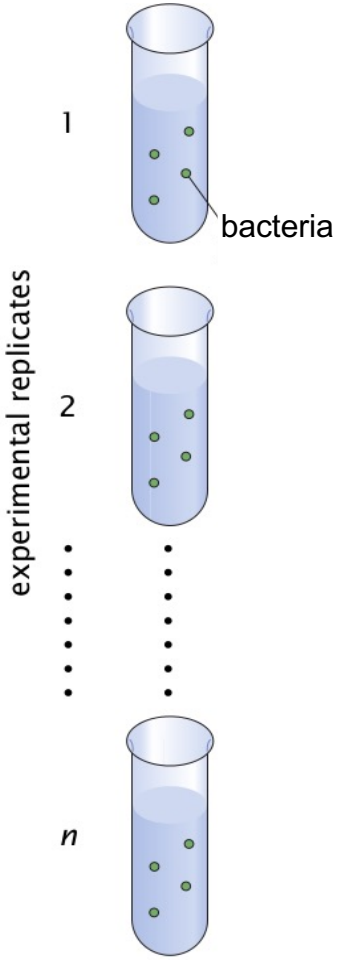


- Counting the bacteria that survive phage and using inference based on the **probability distribution** of the number of phage-resistant bacteria

→ Mutation rate estimate (for mutations giving resistance to phage)

Mutations: Luria-Delbrück experiment

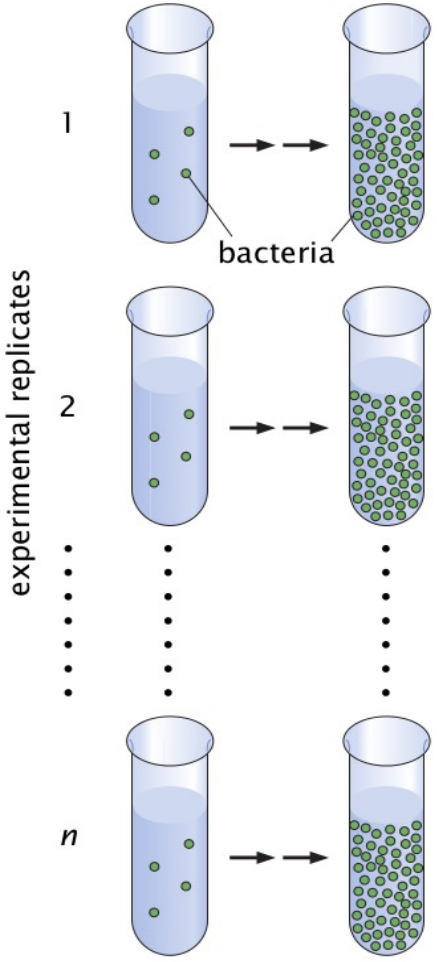
1. Prepare *n* separate identical cultures of the same bacteria



Mutations: Luria-Delbrück experiment

1. Prepare *n* separate identical cultures of the same bacteria

2. Let them grow

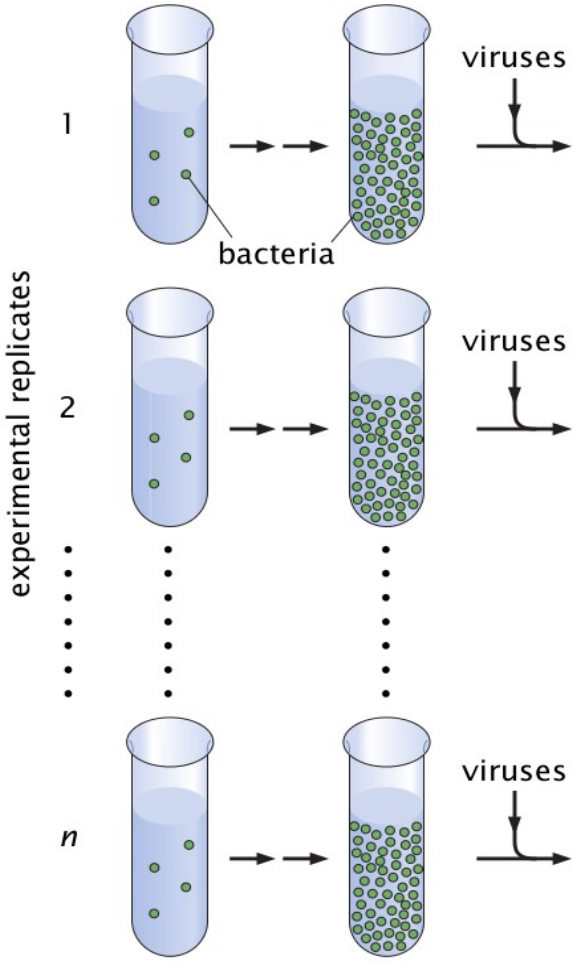


Mutations: Luria-Delbrück experiment

1. Prepare *n* separate identical cultures of the same bacteria

2. Let them grow

3. Add an excess of bacteriophage viruses (phage T1)
→ most bacteria die; only phage-resistant ones survive



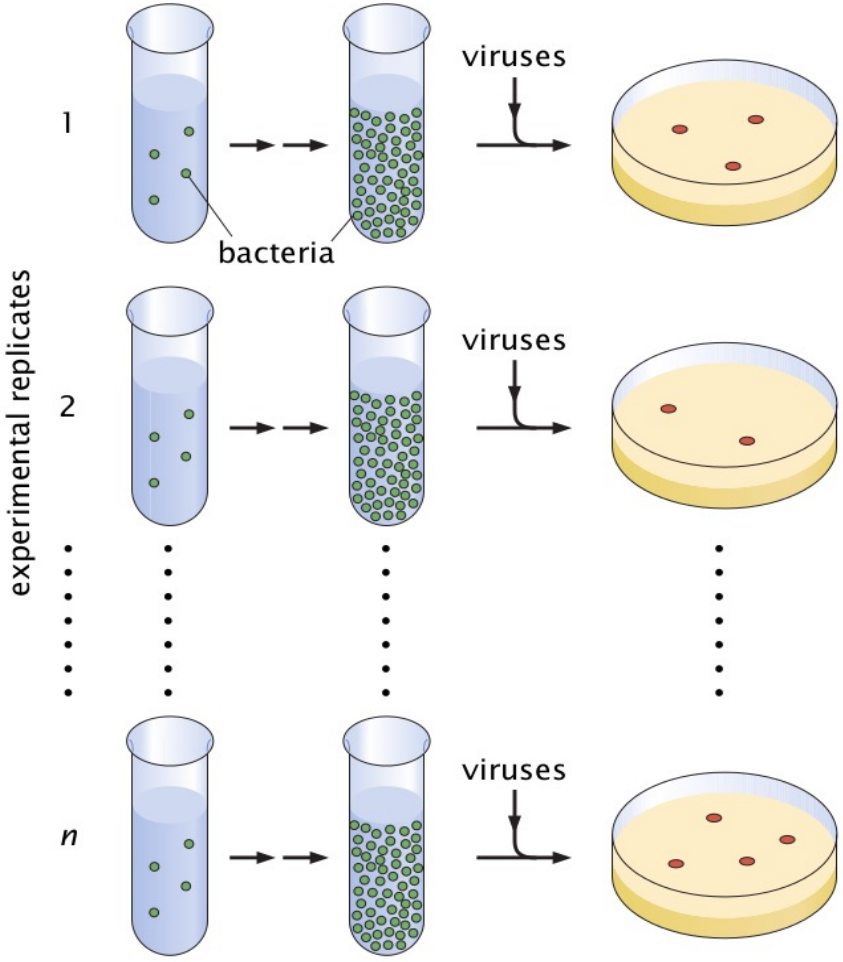
Mutations: Luria-Delbrück experiment

1. Prepare *n* separate identical cultures of the same bacteria

2. Let them grow

3. Add an excess of bacteriophage viruses (phage T1)
→ most bacteria die; only phage-resistant ones survive

4. To count the survivors, plate each culture separately → each survivor forms a colony



Mutations: Luria-Delbrück experiment

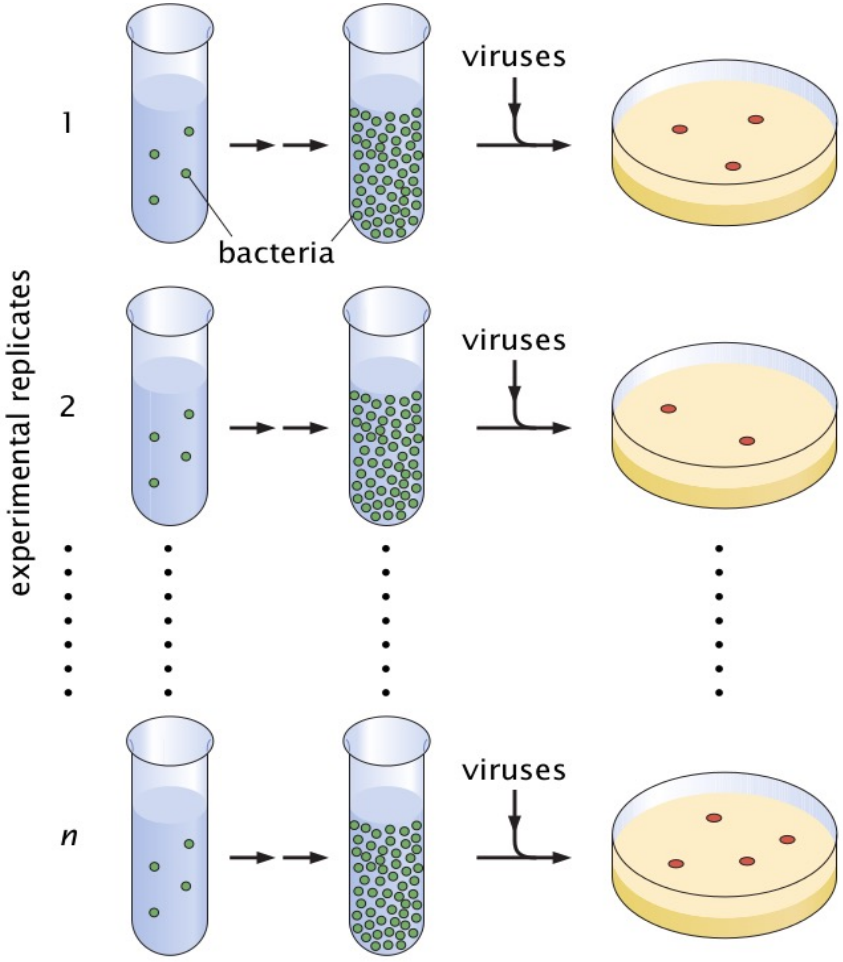
1. Prepare *n* separate identical cultures of the same bacteria

2. Let them grow

3. Add an excess of bacteriophage viruses (phage T1)
→ most bacteria die; only phage-resistant ones survive

4. To count the survivors, plate each culture separately → each survivor forms a colony

5. Count the number *m* of colonies growing in each plate → get *n* values of *m*



Fitness effects of mutations

- **Different measurements – most mutations are deleterious**

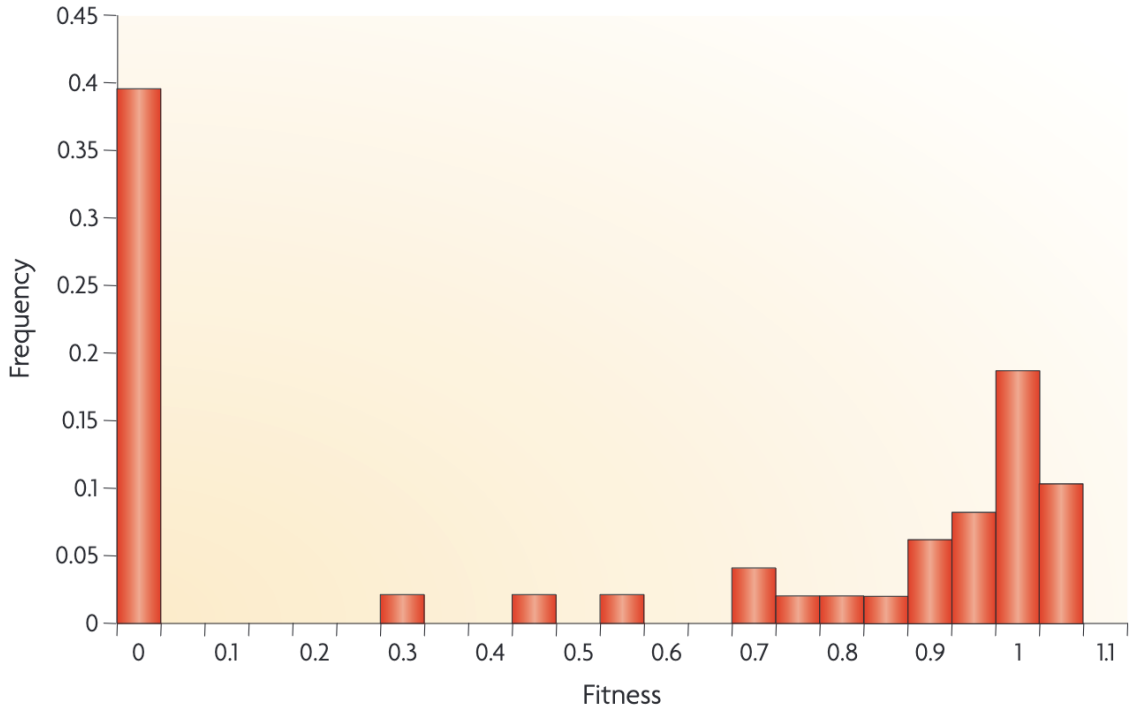


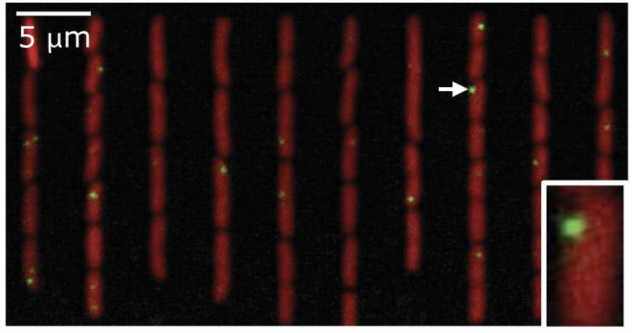
Figure 1 | **The distribution of fitness effects of random mutations in vesicular stomatitis virus.** In this experiment, random mutations were introduced into the virus, and the fitnesses of the mutants were compared against the unmutated wild type. A fitness of less than one indicates that the mutant was less fit than the wild type, so the mutation was deleterious. A fitness of zero indicates that no mutated progeny were recovered, and that the mutation was therefore lethal. Data from REF. 15.

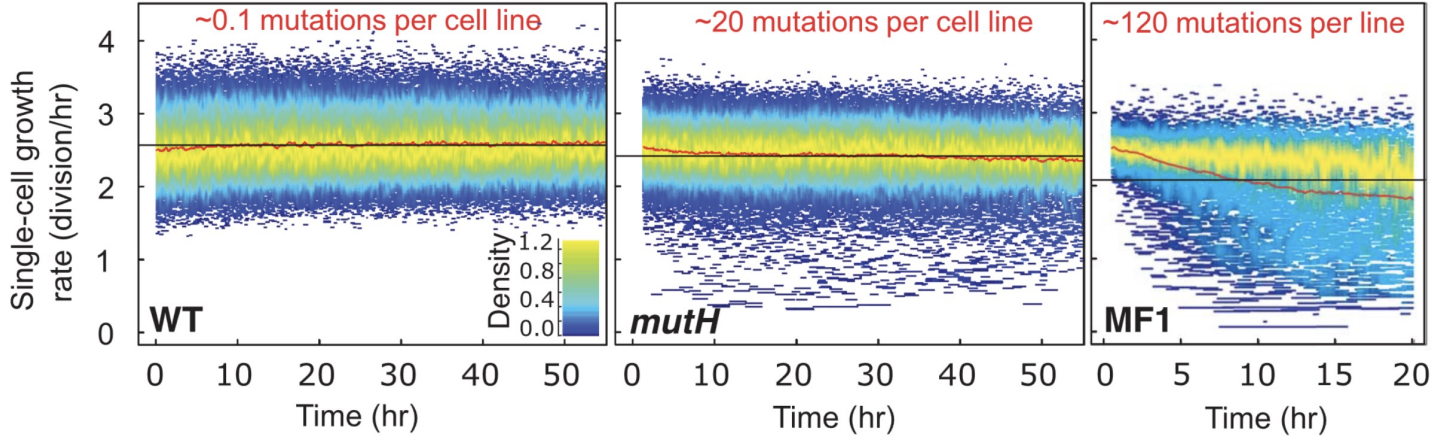
Sanjuan et al, 2004

Fitness effects of mutations

- **Different measurements – most mutations are deleterious**

Mutation accumulation in *E. coli* in a microfluidic mother machine – Robert et al, 2018

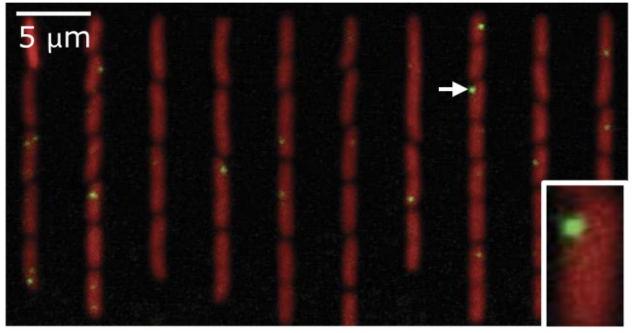


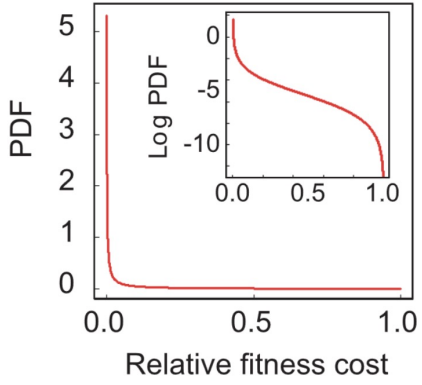


Fitness effects of mutations

- **Different measurements – most mutations are deleterious**

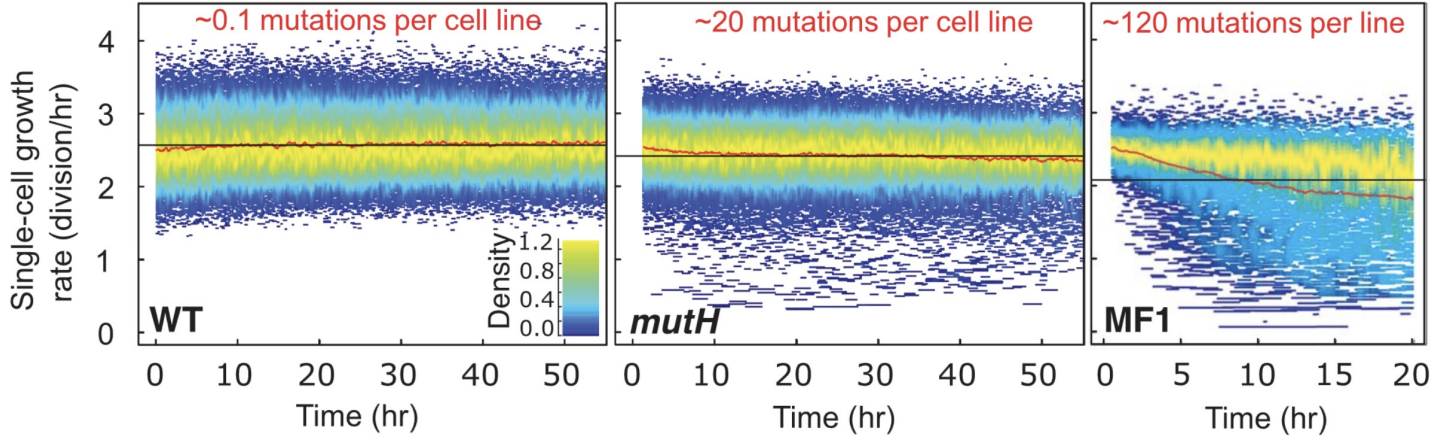
Mutation accumulation in *E. coli* in a microfluidic mother machine – Robert et al, 2018





Mean relative fitness cost: 3.1×10^{-3}

(+ 1% lethal mutations)



Large population: natural selection

Let us focus on the fate of one mutation: does it spread in the population?

- **Deterministic description for large populations**

Consider a population in exponential growth, with 2 types:

- B wild-types, division rate 1 (sets time unit)
- A mutants, division rate 1+s

$$\begin{cases} \frac{dA}{dt} = (1+s)A\,, \\ \frac{dB}{dt} = B\,, \end{cases}$$

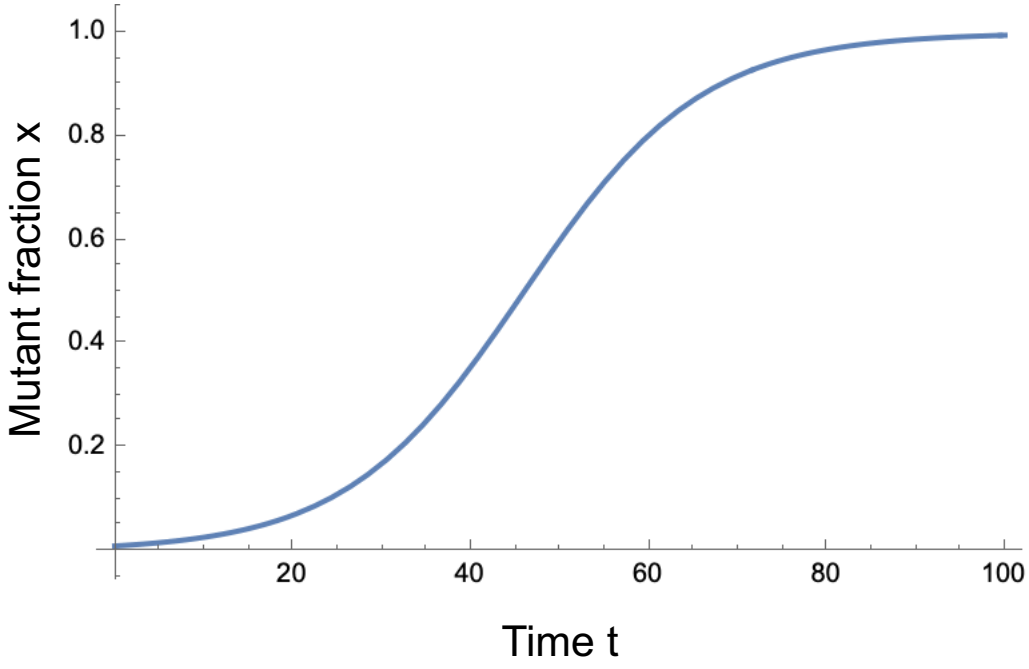
which gives $\frac{dx}{dt} = sx(1-x)$

Solution: $x(t) = \frac{x_0 e^{st}}{1 + x_0(e^{st}-1)}$

If s>0, mutant fraction grows toward a limit of 1 for large t (but does not reach it)

→ Natural selection: fitter type dominates (no coexistence at fixed x)

Example: x_0=0.01, s=0.1

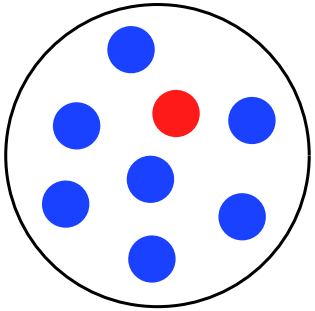


Finite population: genetic drift

What is the fate of a new mutation appearing in a finite-size population of haploid and asexual microorganisms (e.g. bacteria)?
First consider neutral mutants (no natural selection)

- **Population with finite and constant size N: Wright-Fisher model**

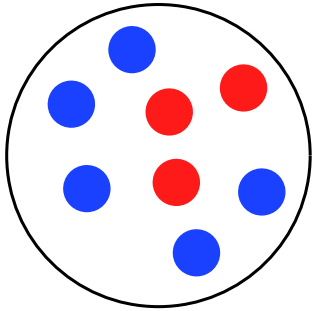
Generation n



k_n=1
$x_n=k_n/N=1/8$

Binomial sampling

Generation n+1



k_{n+1}=3
$x_{n+1}=k_{n+1}/N=3/8$

Non-overlapping generations
k_n: number of mutants in generation n

Next generation formed by binomial sampling

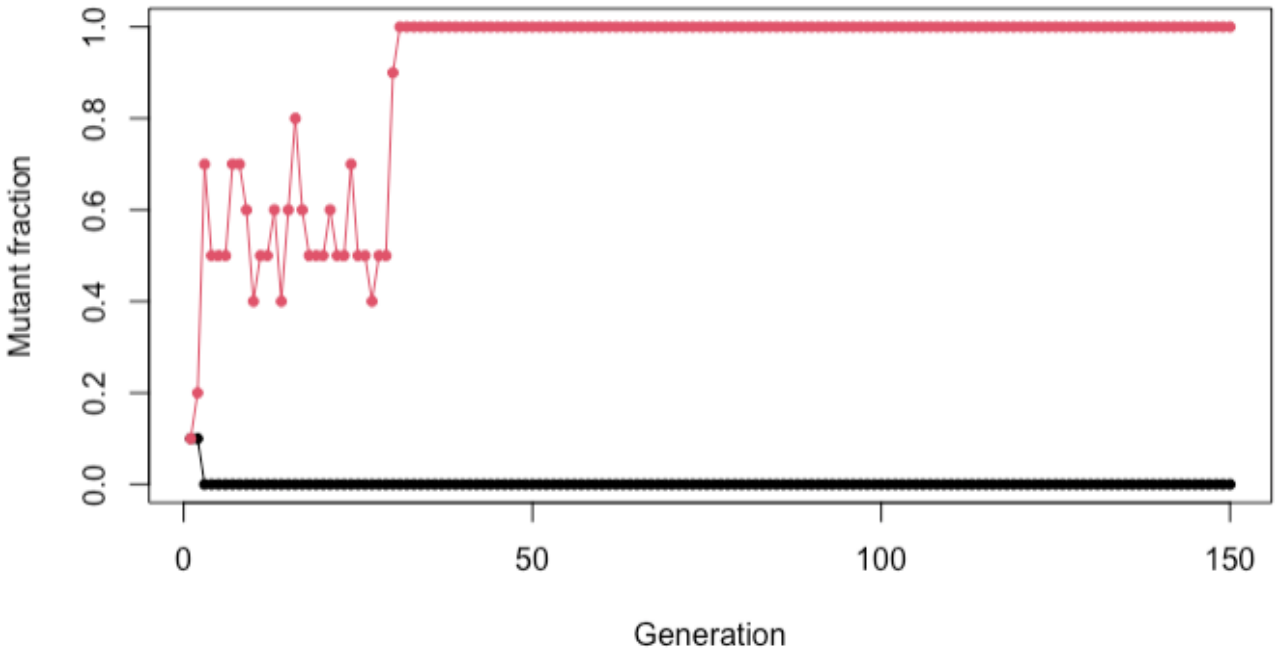
$$P(k_{n+1}) = \binom{N}{k_{n+1}} (x_n)^{k_{n+1}} (1-x_n)^{N-k_{n+1}}$$

Remark: other model: Moran model (one individual dies and one divides at each time step)

Finite population: genetic drift

- **Population with finite and constant size N: Wright-Fisher model**

Fraction of mutants over time (generation after generation): random walk



2 simulated trajectories
starting from k=1
with N=10

After a large number of generations, mutants either take over (fix) or disappear (go extinct)
This is due to finite-size fluctuations, called genetic drift
Fixation probability starting from 1 mutant: 1/N, by symmetry

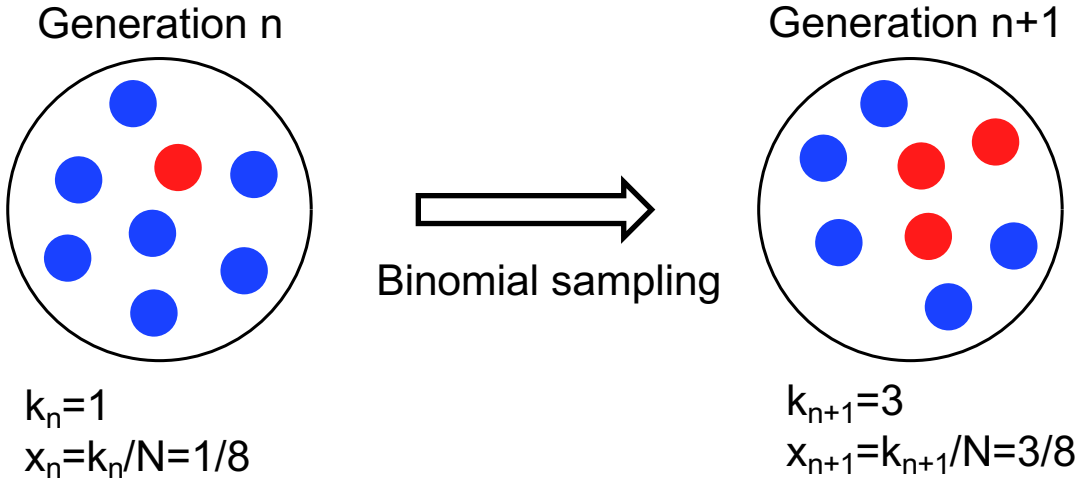
Finite population: genetic drift and selection

What happens if there is natural selection in a finite-size population?

Assume that mutants are more (or less) likely than wild-types to contribute to the next generation

Encode this in *fitnesses*: 1 for wild-types (reference), 1+s for mutants

- **Population with finite and constant size N: Wright-Fisher model**



Non-overlapping generations

k_n: number of mutants in generation n

Next generation formed by binomial sampling

$$P(k_{n+1}) = \binom{N}{k_{n+1}} (x'_n)^{k_{n+1}} (1 - x'_n)^{N - k_{n+1}}$$

$$x'_n = \frac{(1+s)x_n}{(1+s)x_n + 1 - x_n} = \frac{(1+s)x_n}{1 + sx_n}$$

What is the probability that a mutant fixes, if it has a selective advantage s?

Finite population: genetic drift and selection

- **Population with finite and constant size N: Wright-Fisher model**

Fixation probability starting from one mutant (1): branching process

Focus on the first sampling step, from initial state (called generation 1) to next generation (generation 2)

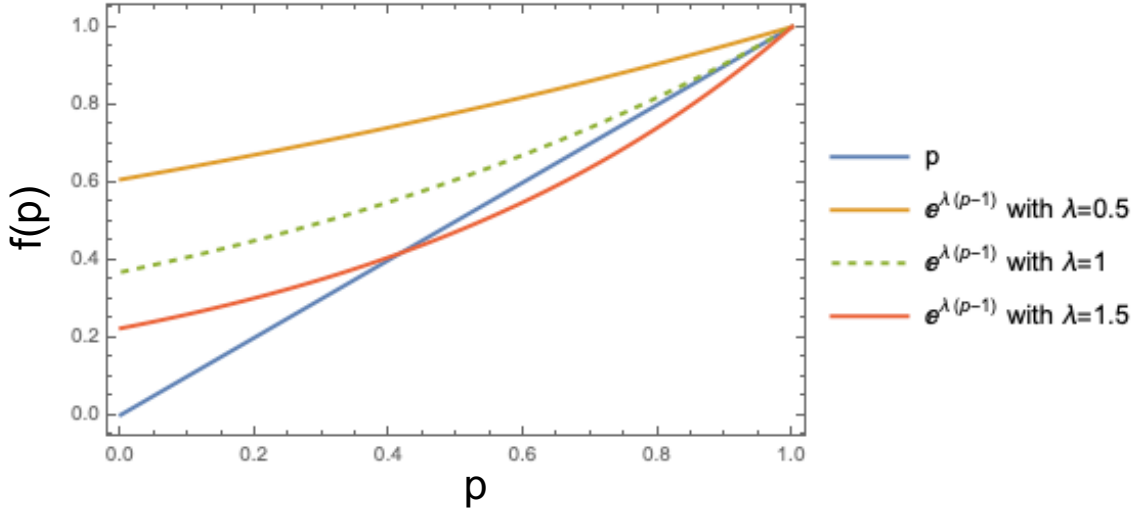
Poisson approximation: $P(k_2) = \binom{N}{k_2} (x_1')^{k_2} (1 - x_1')^{N - k_2} \approx \frac{\lambda^{k_2}}{k_2!} e^{-\lambda}$, with $\lambda = N x_1'$

$N \gg 1$, $x_1' \ll 1$ and $N x_1'$ is of order 1

after growth, starting from $x_1 = 1/N \ll 1$

Assuming that all mutant lineages are independent, the probability of extinction p of the mutant satisfies:

$$p = \exp\left[\lambda\,(p-1)\right]$$



Finite population: genetic drift and selection

- **Population with finite and constant size N: Wright-Fisher model**

Fixation probability starting from one mutant (1): branching process

Focus on the first sampling step, from initial state (called generation 1) to next generation (generation 2)

Poisson approximation: $P(k_2) = \binom{N}{k_2} (x_1')^{k_2} (1-x_1')^{N-k_2} \approx \frac{\lambda^{k_2}}{k_2!} e^{-\lambda}$, with $\lambda = N x_1'$

$N \gg 1$, $x_1' \ll 1$ and $N x_1'$ is of order 1

after growth, starting from $x_1 = 1/N \ll 1$

Assuming that all mutant lineages are independent, the probability of extinction p of the mutant satisfies:

$$p = \exp[\lambda (p-1)]$$

If $\lambda<1$ i.e. s<0, or if $\lambda=1$ i.e. s=0, the only solution is p=1: extinction is certain
If $\lambda>1$ i.e. s>0, another solution exists

Strategy to solve this equation and to obtain p if s>0: perform an expansion for small s

Starting from one mutant: $x_1 = 1/N \ll 1 \longrightarrow \lambda = N x_1' = N \frac{(1+s)x_1}{1+s x_1} = \frac{1+s}{1+s/N} = 1 + s - \frac{s}{N} + O\left(\frac{s^2}{N^2}\right)$

Assume $|s| \ll 1$ and $N|s| \gg 1 \longrightarrow \lambda = 1 + s + o(s^2)$

Then, to first order in s>0, we obtain p=1-2s, meaning that the probability of mutant fixation is 1-p=2s <<1

Finite population: genetic drift and selection

- **Population with finite and constant size N: Wright-Fisher model**

Fixation probability starting from one mutant (2): diffusion approximation

The branching process makes strong approximations, in particular N>>1: neglects finite population size
It gives a fixation probability 0 for s=0 – but we know that it is actually 1/N…
Nevertheless, it takes into account the fact that the mutant starts in small numbers

Diffusion approximation: another, more precise, approximation
Assumes |s|<<1 and N>>1 but includes term in 1/N, from binomial sampling variance $\Delta x_{n+1}^2 = \frac{x'_n(1-x'_n)}{N}$

It gives $\rho(1/N) = \frac{1-e^{-2s}}{1-e^{-2Ns}}$ for the mutant fixation probability ρ starting from one mutant (x_1=1/N)

In particular:

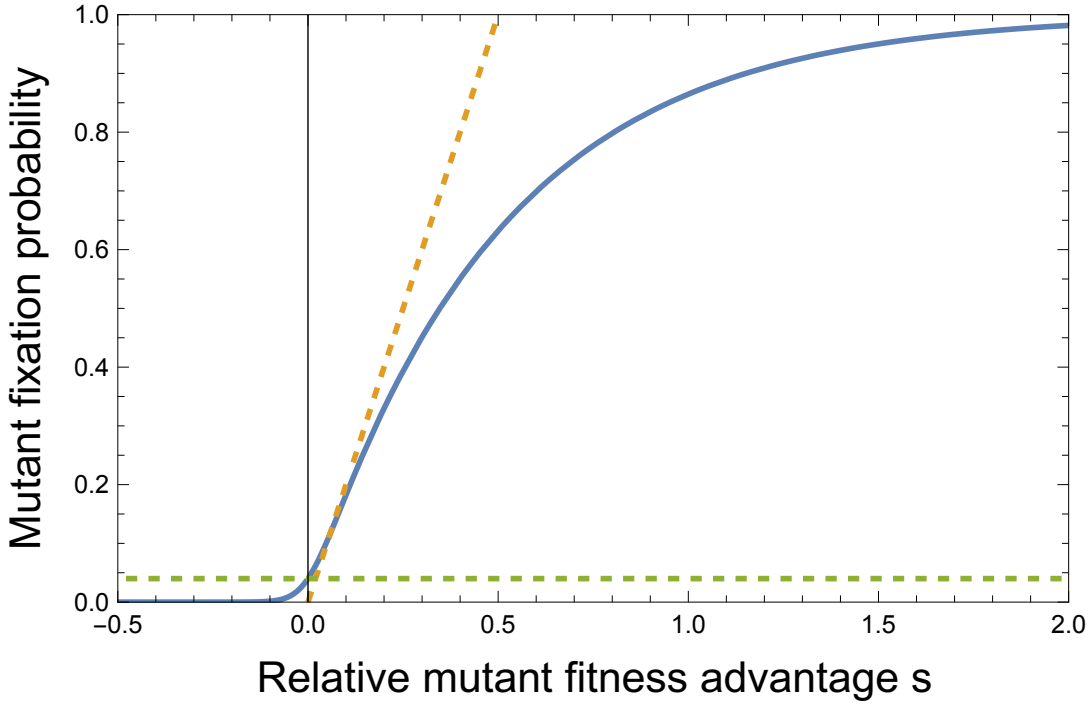
$|s| \ll 1/N \longrightarrow \rho(1/N) = \frac{1-(1-2s+O(s^2))}{1-(1-2Ns+O(N^2s^2))} \sim \frac{1}{N}$: "effectively neutral" regime

$N \gg 1$ while $s > 0$ but $s \ll 1$ and $Ns \gg 1 \longrightarrow \rho(1/N) = \frac{1-(1-2s+O(s^2))}{1-e^{-2Ns}} \sim 2s$: branching process regime

$N \gg 1$ while $s < 0$ but $|s| \ll 1$ and $N|s| \gg 1 \longrightarrow \rho(1/N) = \frac{1-(1-2s+O(s^2))}{1-e^{-2Ns}} \sim -2se^{2Ns} \to 0$: deleterious regime

Finite population: genetic drift and selection

- **Population with finite and constant size N: Wright-Fisher model**



N=25
Starting from one mutant

W–F diffusion $\rho(1/N) = \dfrac{1 - e^{-2s}}{1 - e^{-2Ns}}$

W–F BP $\rho(1/N) = 2s$

Neutral $\rho(1/N) = 1/N$

Important scale: N |s| - if N |s| << 1, effectively neutral regime
- if N |s| >> 1, selective regime

Finite population: genetic drift and selection

- **Population with finite and constant size N: Wright-Fisher model**

Fixation probability starting from a fraction x of mutants: diffusion approximation

$$\rho(x) = \frac{1 - e^{-2Nsx}}{1 - e^{-2Ns}}$$

Assume s>0 and Ns>>1. Then, $\rho(x) \approx 1 - e^{-2Nsx}$

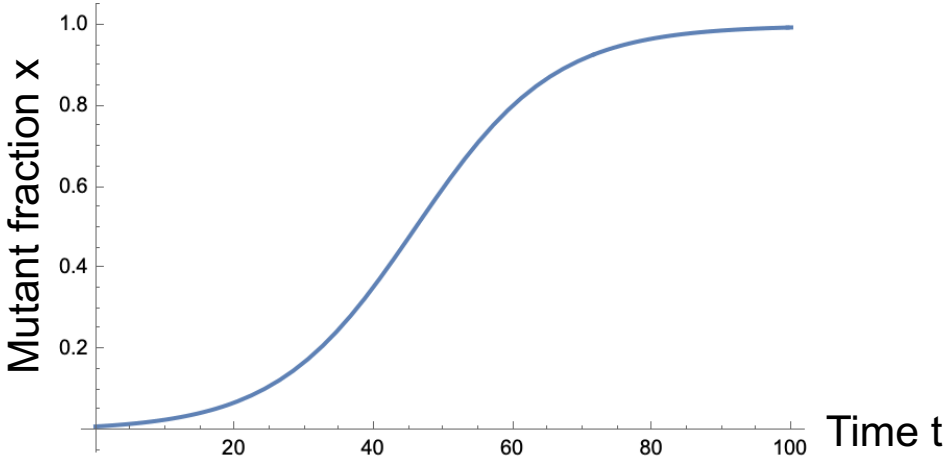
becomes small if Nsx>1, i.e. x>1/(Ns)

If a mutant with relative fitness advantage s>>1/N reaches a fraction x>1/(Ns), **it is very likely to fix**
For beneficial mutants, extinctions happen early, when they are in small numbers, due to fluctuations
Fixation timescale: ~1/s generations

Reminder: deterministic description:

This is OK if N>>1
and s>>1/N
and x>>1/(Ns)

Large population, sufficient selective advantage and sufficient mutant fraction



Finite population: genetic drift, selection and mutations

So far, we described the fate of one mutation appearing in a population
But other mutations may appear in the meantime: what is their effect?

- **Sequential fixation of mutations versus multiple mutations & clonal interference**

Mutation probability μ per individual and generation → total mutation probability $N\mu$ per generation

Beneficial mutant with $s>>1/N$ but $s<<1$:
Probability of fixation ~$2s$ → probability $2N\mu s$ per generation to have such a mutation that will then fix
Fixation timescale ~$1/s$ generations → if $2N\mu s<<s$ i.e. $N\mu<<1$, no new mutation appears during fixation

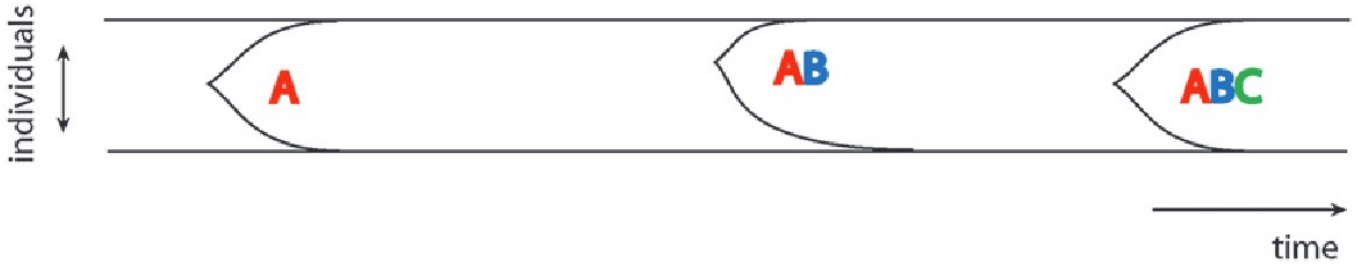
Effectively neutral mutant with $|s|<<1/N$:
Probability of fixation ~$1/N$ → probability μ per generation to have such a mutation that will then fix
Fixation timescale ~N generations → if $N\mu<<1$, no new mutation appears during fixation (same as above)

→ **If $N\mu<<1$, we can consider that mutations fix successively** – it is fine to focus on one at a time
If $N\mu>1$, new mutant lineages typically appear during the fixation process of a mutant
This is more complex (not described by our Wright-Fisher analysis)

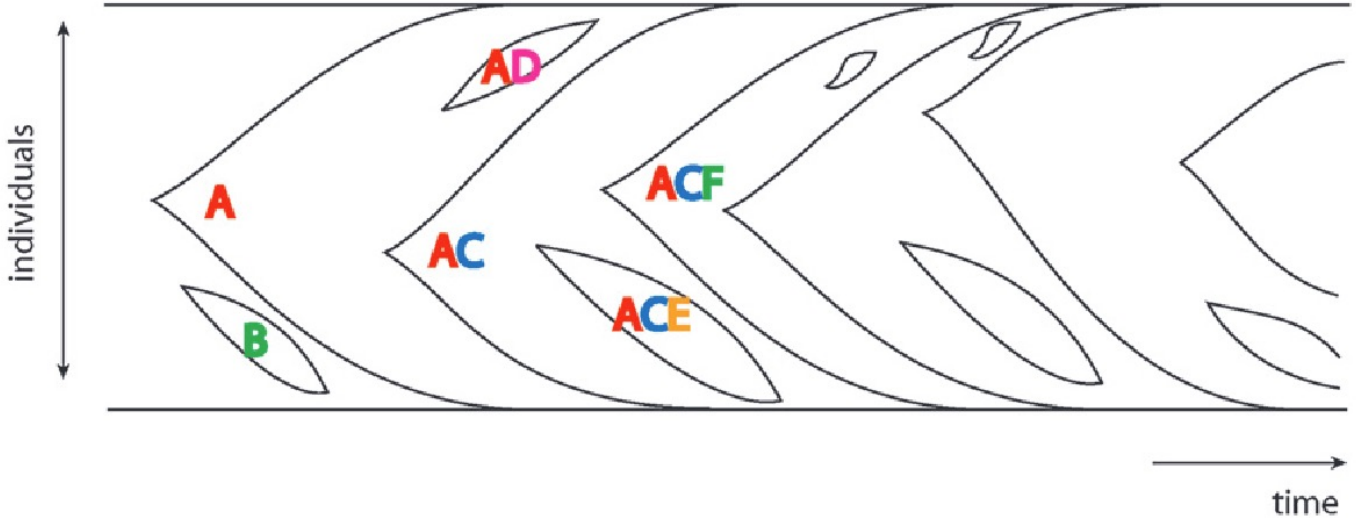
Finite population: genetic drift, selection and mutations

- **Sequential fixation of mutations versus multiple mutations & clonal interference**

Two different regimes – Desai and Fisher 2007



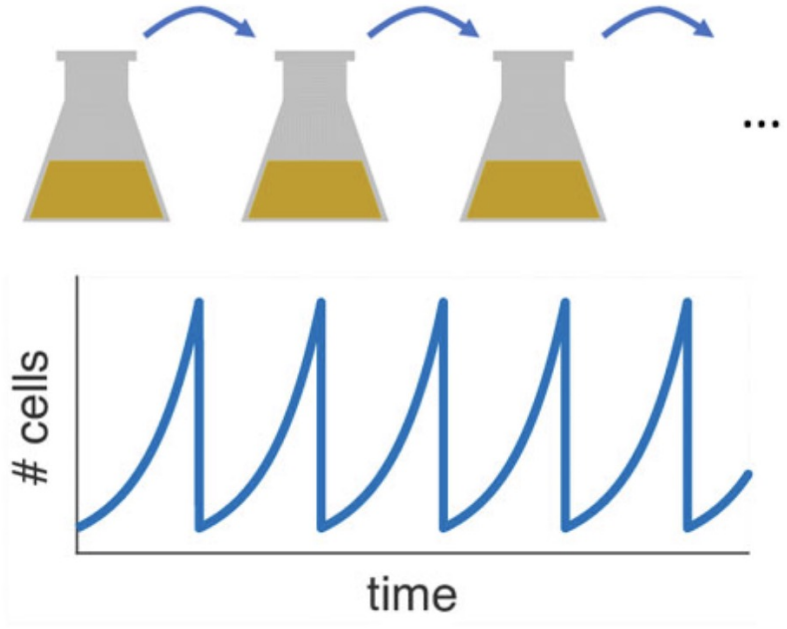
Nμ<<1: sequential fixation of mutations



Nμ>1: multiple mutations growing in frequency at the same time
→ "clonal interference":
if two beneficial mutations that would fix if alone appear at similar times, they compete, and only one fixes

Link to evolution experiments

- **Experimental protocol: serial transfer (or serial passage, or serial dilution)**



1- Cells are placed in culture medium and grow (exponentially – may or may not reach stationary phase)

2- Periodically, a small volume is sampled and placed in new medium – the rest is discarded → bottleneck

Phases 1&2 are repeated

Assume that the bottleneck has constant size K

Link to evolution experiments

- **Modeling serial transfer (or serial passage, or serial dilution)**

1- Growth phase → deterministic exponential growth with no death, starting from K cells

Starting from mutant fraction $x_n=k_n/K$ at bottleneck n, the fraction after growth reads $x'_n = \frac{x_n e^{st}}{1+x_n(e^{st}-1)}$

Introducing $\sigma = e^{st}-1$, we can write $x'_n = \frac{(1+\sigma)x_n}{1+\sigma x_n}$ → as in the Wright-Fisher model, with σ instead of s

2- Transfer / bottleneck → binomial sampling of K individuals from the grown population

Binomial sampling: $P(k_{n+1}) = \binom{K}{k_{n+1}} (x'_n)^{k_{n+1}} (1-x'_n)^{K-k_{n+1}}$ → as in the Wright-Fisher model

where k_{n+1} is the number of mutants at bottleneck n+1

Mutant fixation probability: as in the Wright-Fisher model, the diffusion approximation gives

$\rho(1/K) = \frac{1-e^{-2\sigma}}{1-e^{-2K\sigma}} = \frac{1-e^{-2st}}{1-e^{-2Kst}}$ starting from one mutant at a bottleneck (fraction 1/K)

K>>1, |σ|<<1 K>>1, |s|t<<1

Summary: effects at play

- **Different effects so far**
  - **Mutations:**
    - generate diversity
    - most have small fitness effects, most are deleterious
  - **Natural selection:**
    - acts upon random mutations
    - because of it, beneficial mutations tend to take over and fix
  - **Genetic drift:**
    - corresponds to stochastic fluctuations
    - arises from finite population size (total population or mutant population)
    - means that moderately beneficial mutations often do not fix – new mutations start in a single individual
- **Additional important effects**
  - Recombination (horizontal gene transfer; sexual reproduction)
  - Interactions between mutations – not just additive fitness effects
  - Specific interactions between individuals (beyond mere competition) – cooperativity, attacks…
  - Spatial population structure, migrations and genetic flow
  - Environmental variability