Modeling evolutionary dynamics

Deterministic growth

Exponential Growth

- Assume a population of replicators (for example bacteria) sitting in a well-stirred flask with plenty of food.
- Assume that for bacteria of type *i* there is a probability r_i per unit time that they replicate. This parameter is the *growth rate*.
- The number of bacteria of each type *i* will then exponentially grow as:

$$\frac{dn_i(t)}{dt} = r_i n_i(t) \Leftrightarrow n_i(t) = n_i(0)e^{r_i t}$$

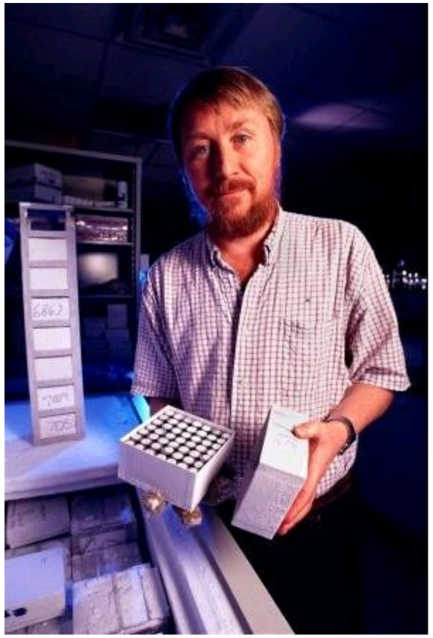
- We can also look at the dynamics of the *relative* proportions of individuals of types *i* and *j.*

$$\frac{n_i(t)}{n_j(t)} = \frac{e^{r_i t} n_i(0)}{e^{r_j t} n_j(0)} = e^{(r_i - r_j)t} \frac{n_i(0)}{n_j(0)}$$

- Whenever $r_i > r_j$ the ratio will grow exponentially and when $r_i < r_j$ it will decrease exponentially such that soon, the population will be *dominated* by the type with the highest growth-rate.

- We find for the *proportions:* $\rho_i(t) = \dfrac{n_i(t)}{\sum\limits_j n_j(t)} = \dfrac{e^{r_i t} n_i(0)}{\sum\limits_j e^{r_j t} n_j(0)} = \dfrac{e^{r_i t} \rho_i(0)}{\sum\limits_j e^{r_j t} \rho_j(0)} = \dfrac{e^{r_i t}}{\left\langle e^{rt} \right\rangle} \rho_i(0)$

Evolution in a controlled environment



Richard Lenski's long-term evolution experiment

1. Prepare a small flask with a fluid containing a *fixed* amount of nutrients.
2. Seed with some *E. coli* cells.
3. Grow overnight.
4. Take 1% of the fluid and cells out of the flask.
5. Put into a new flask with identical amount of fluid and nutrients.
6. Go to step 3.

The experiment has been running since February 1988, celebrated 75'000 generations in 2022.

- We can approximate what happens overnight by assuming that all bacterial types in the flask grow exponentially, *until the food runs out.* After that they just sit there.

- If we further assume that independent of their type, bacteria consume the same amount of food per reproduction, then a fixed amount of food supports the production of a fixed number of bacterial cells, say *N*.

- Thus, each night the population grows from 0.01N to N.

Evolution in a controlled environment

- Initially all bacteria are the same and grow at the same rate *r*
- Assume that at some time point, a mutant appears with growth rate *r+dr*
- Over one night we have: Wildtype: $n \rightarrow ne^{rt_1}$

  Mutant: $m = 1 \rightarrow e^{(r+dr)t_1}$

  where t_1 is the amount of time until the food ran out during the first night

- During the second night:

  Wildtype: $0.01ne^{rt_1} \rightarrow 0.01ne^{rt_1}\ e^{rt_2} = 0.01ne^{r(t_1+t_2)}$

  Mutant: $0.01e^{(r+dr)t_1} \rightarrow 0.01e^{(r+dr)t_1}\ e^{(r+dr)t_2} = 0.01e^{(r+dr)(t_1+t_2)}$

Evolution in a controlled environment

- Fist night: Wildtype: $n \rightarrow ne^{rt_1}$

  Mutant: $m = 1 \rightarrow e^{(r+dr)t_1}$

- Second night: Wildtype: $0.01ne^{rt_1} \rightarrow 0.01ne^{rt_1}\ e^{rt_2} = 0.01ne^{r(t_1+t_2)}$

  Mutant: $0.01e^{(r+dr)t_1} \rightarrow 0.01e^{(r+dr)t_1}\ e^{(r+dr)t_2} = 0.01e^{(r+dr)(t_1+t_2)}$

- After *k* nights, when the total growth time is $t = t_1 + t_2 + \ldots + t_k$:

  Wildtype: $0.01^k ne^{rt}$

  Mutant: $0.01^k e^{(r+dr)t}$

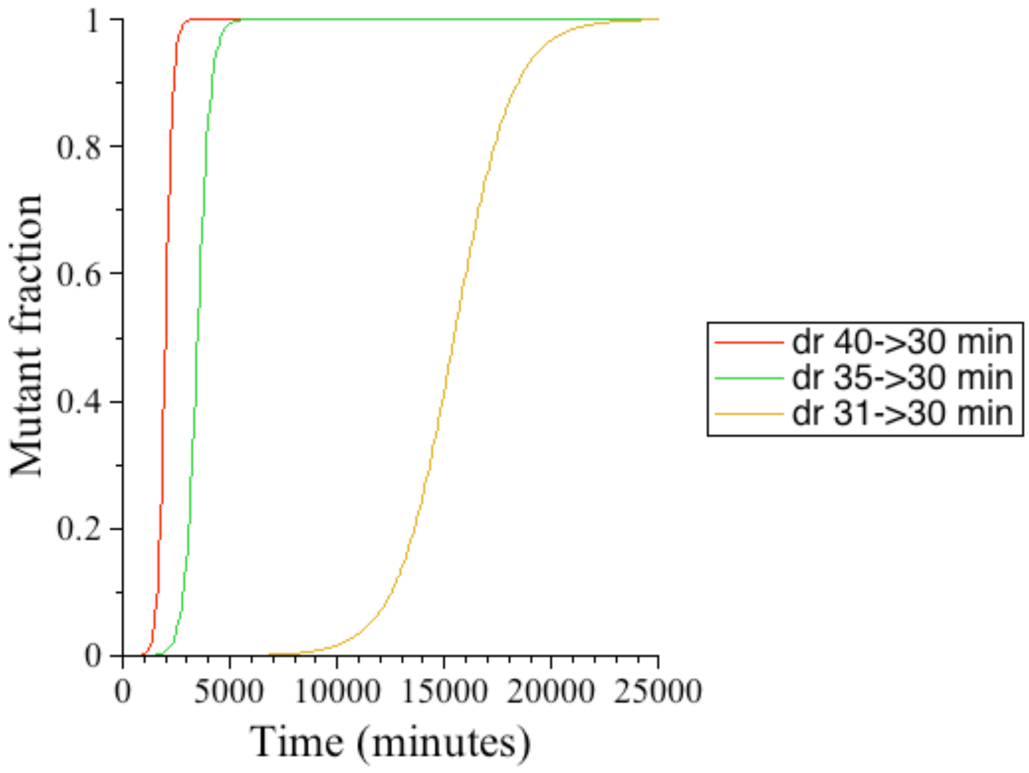
  The *fraction* of mutants after growth-time *t* is:

$$\rho(t) = \frac{m(t)}{m(t)+n(t)} = \frac{e^{(r+dr)t}}{e^{(r+dr)t}+ne^{rt}} = \frac{e^{(dr)t}}{e^{(dr)t}+n}$$

Evolution in a controlled environment

$$\rho(t) = \frac{e^{(dr)t}}{e^{(dr)t} + n}$$

Exponential take-over of a beneficial mutant (*dr* > 0)



n = 100,000 (in Lenski's experiment)

from 40 to 30 minutes doubling:
dr = 0.0058/min.
from 35 to 30 minutes doubling:
dr = 0.0033/min.
from 31 to 30 minutes doubling:
dr = 0.0007/min.

Time to go from 1/(n+1) to n/(n+1) is:

$$\frac{e^{(dr)t}}{e^{(dr)t}+n} = \frac{n}{n+1} \Longrightarrow t = \frac{2\log(n)}{dr}$$

For our example, assuming 10 hrs growth per night:
3986 min -> 6.6 days
6976 min -> 11.6 days
30893 min -> 51.5 days

Evolution in a controlled environment

Summary

Proportion of mutants as a function of time: $\rho(t) = \dfrac{e^{(dr)t}}{e^{(dr)t} + n}$

If the mutation is beneficial $(dr > 0)$, the mutant will ”take over” the population.

If the mutation is deleterious $(dr < 0)$, the proportion of the mutant in the population will keep decreasing.

If the mutant is ‘neutral’, without any selection advantage $(dr = 0)$, its proportion in the population would not change in time, will remain $\rho(t) = \frac{1}{1+n}$

Note: we can predict precisely how the proportion of the mutant changes in time (hence deterministic growth). In particular, for a neutral mutation, we would predict no change over time.

Do we now understand evolution?

In real life, populations are composed of relatively small numbers of discrete individuals. The dynamics of the frequency of a neutral mutant that appears in such a population is not described well by the model we just discussed.

Furthermore, if only 1% of the 4.6 * 10^6 positions in the *E.coli* were neutral, the number of distinct genomes with the same phenotype would be $4^{46000 =}$ 5.74 * 10^{27694}, much larger than the size of the cultures we encounter, which neutral genotype is present in the culture at a given time and in what frequency can only be determined by studying the stochastic dynamics of finite populations.

Stochastic dynamics in finite populations

Stochastic dynamics in finite populations

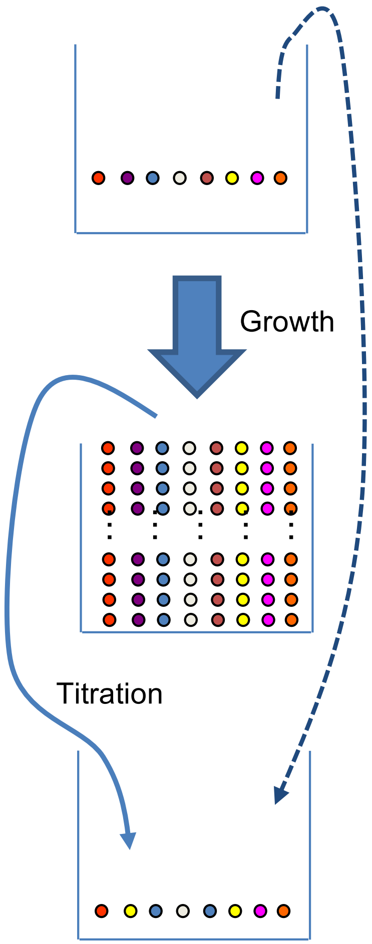
- Let's imagine a population evolving from one generation to the next, the population size remaining constant, as in Lenski's experiment.

- Assume for simplicity that all members of the population have the same growth rate *r.*

- Assume also that at the start each individual has a unique genotype.

- Thus, over night, all members of the population grow by *exactly* the same amount (this underestimates fluctuations but simplifies the mathematical treatment). The population goes from N to 100 *N,* and each individual goes from 1 to 100 copies (ignore mutations for now as well).

- In the morning we take a *random sample* of size *N* of from the 100 *N* individuals.



This dynamics gives a population composition that is roughly equivalent to what we would get if we chose each member of the new population to be a copy of a randomly chosen member of the previous population.

Stochastic dynamics in finite populations

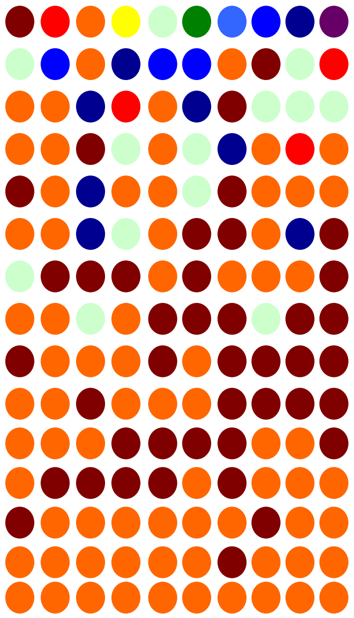
- Let's imagine a population evolving from one generation to the next, the population size remaining constant, as in Lenski's experiment.

- Assume for simplicity that all members of the population have the same growth rate *r.*

- Assume also that at the start each individual has a unique genotype.

- Thus, over night, all members of the population grow by *exactly* the same amount (this underestimates fluctuations but simplifies the mathematical treatment). The population goes from N to 100 *N,* and each individual goes from 1 to 100 copies (ignore mutations for now as well).

- In the morning we take a *random sample* of size *N* of from the 100 *N* individuals.



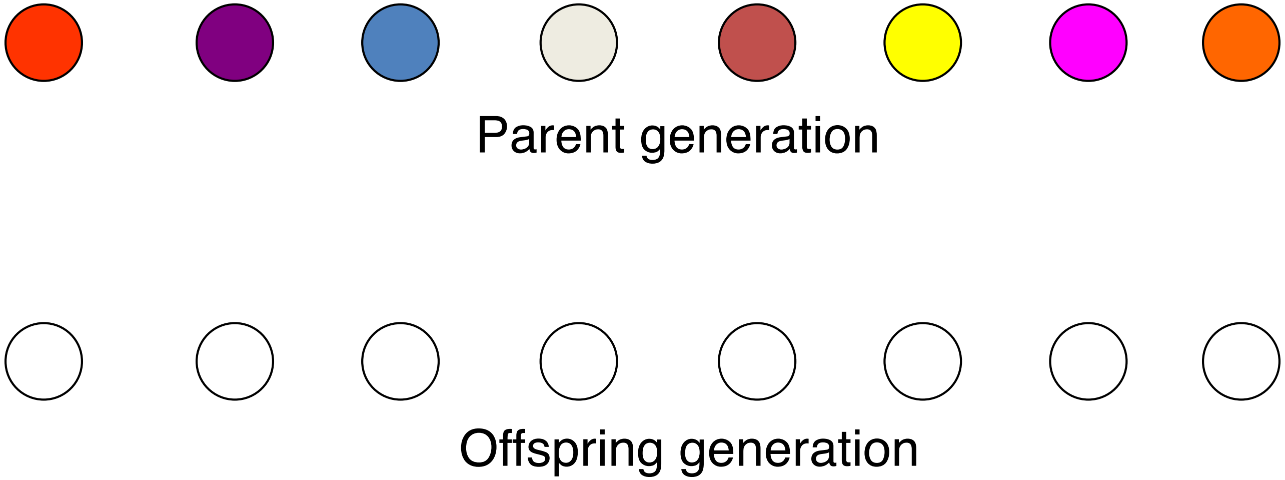
What do you notice about the dynamics of genotypes?

Genetic drift

The frequency of any gene variant in the population will vary in time due to random sampling, a phenomenon which is called *genetic drift*.

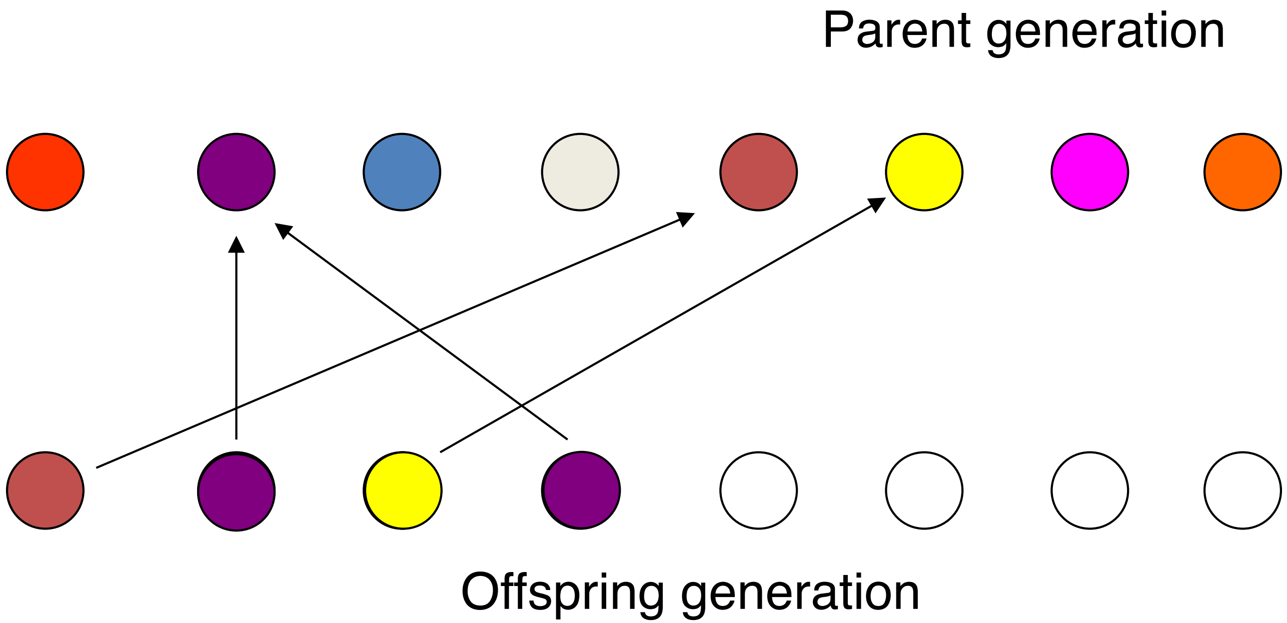
How does this work?

We start with a population of fixed size, each individual having its own genotype



Each individual has the same reproductive success on average =>
Each offspring individual in the new generation has a parent chosen randomly from the parent generation.

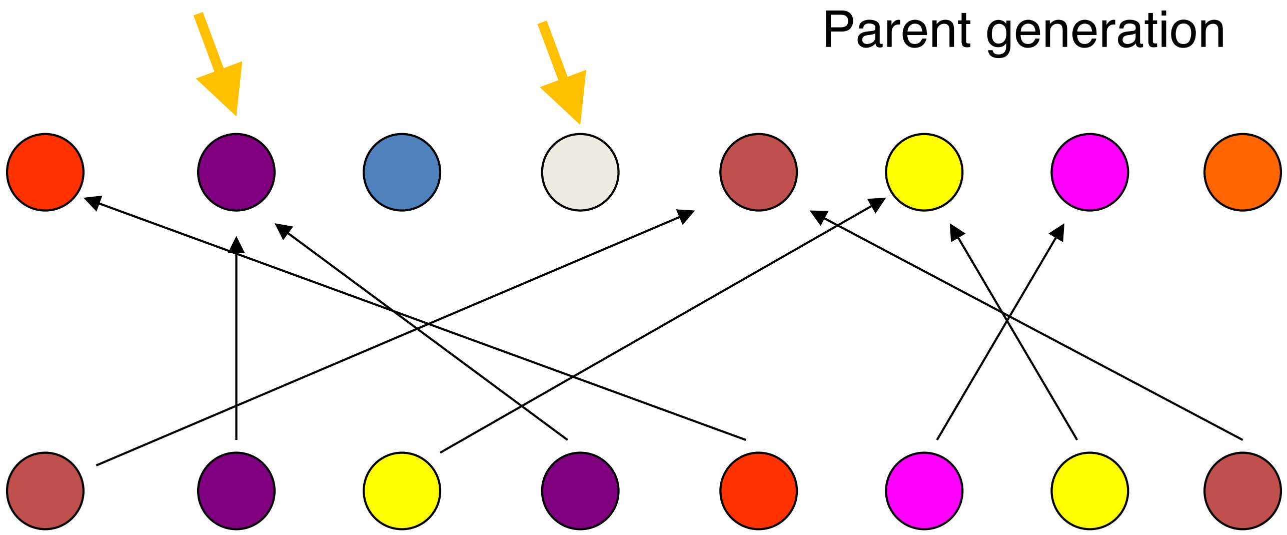
Genetic drift



Each individual has the same reproductive success on average =>
Each offspring individual in the new generation has a parent chosen randomly from the parent generation.

Genetic drift

Some parents have no offspring, some have multiple

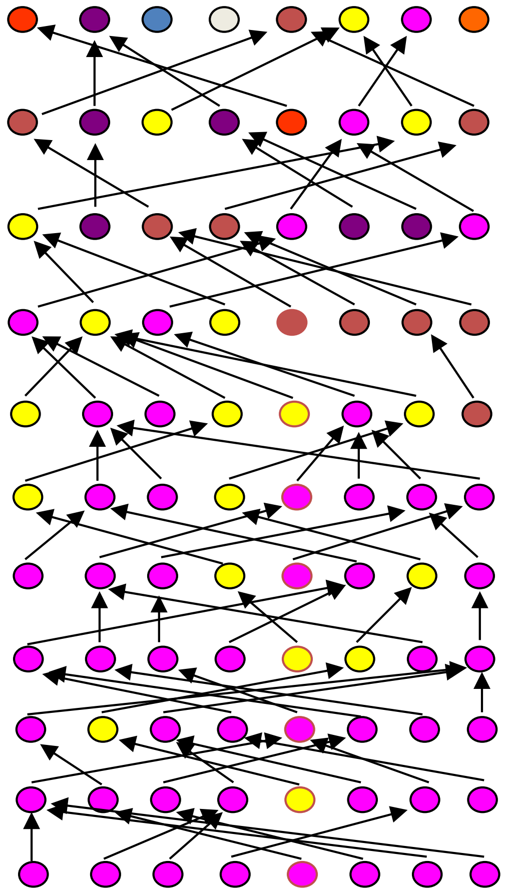


Offspring generation

Each individual has the same reproductive success on average =>
Each offspring individual in the new generation has a parent chosen randomly from the parent generation.

Genetic drift

Eventually all individuals in the population stem from a single individual



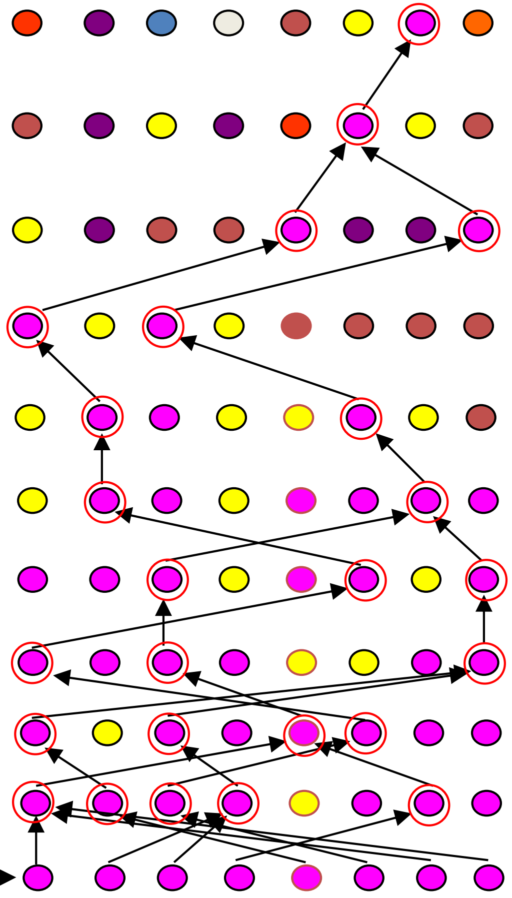
All genetic variation automatically disappears from the population even without fitness differences

How long does it take until the population converges on a single genotype?

Modeling genetic drift in finite populations

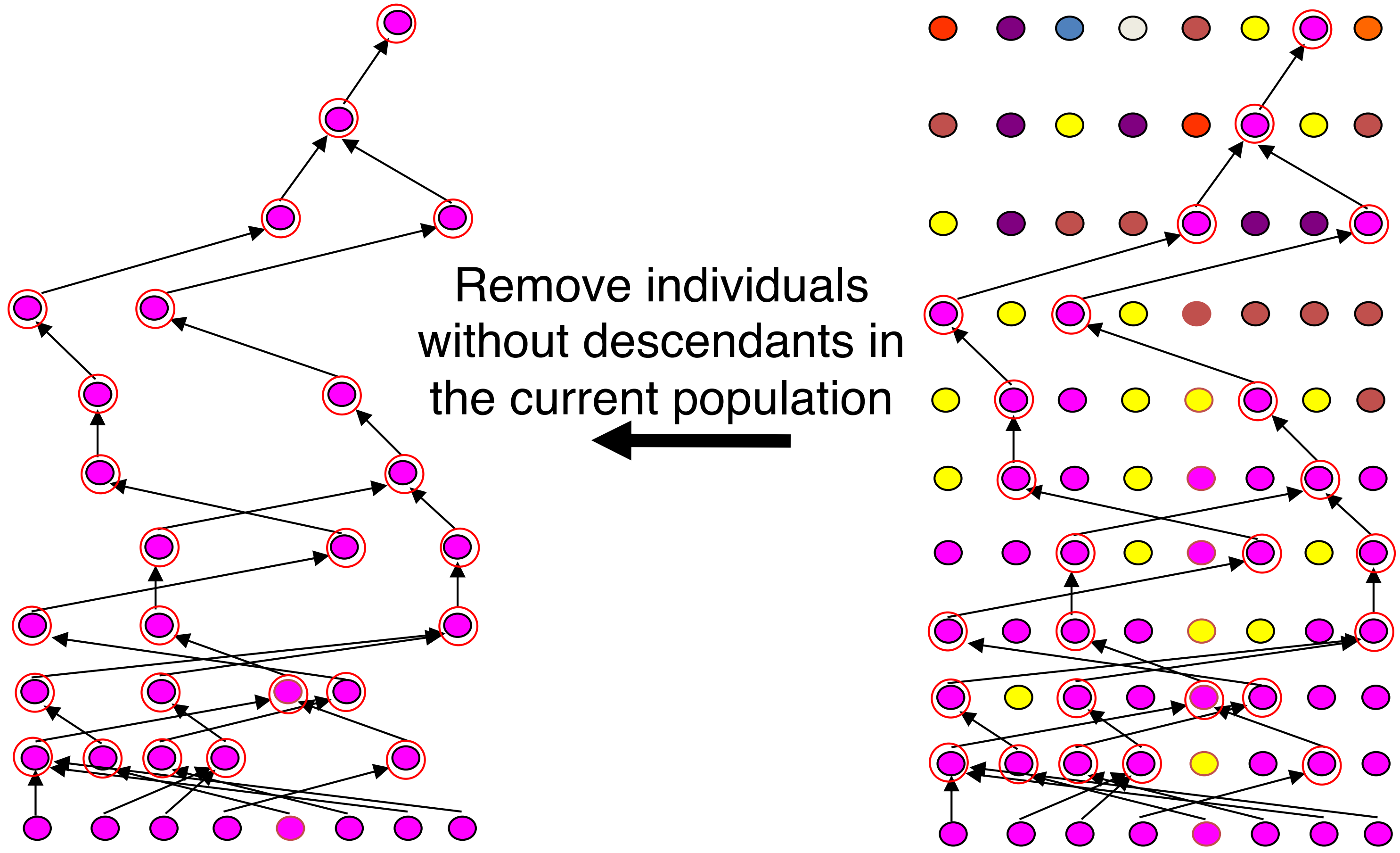
Highlighted are all individuals that have at least one descendant in the current population.

The arrows show how the current population traces back in time.
All other arrows are removed.



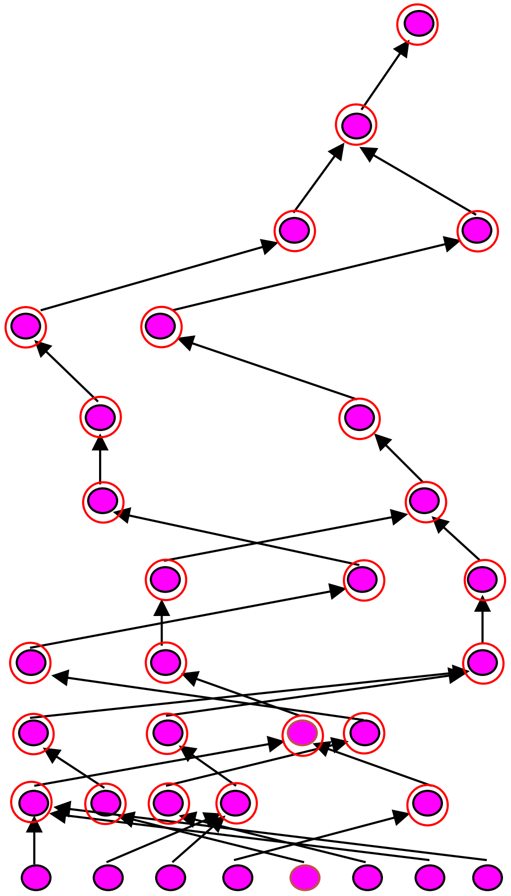
Current Population

Modeling genetic drift in finite populations

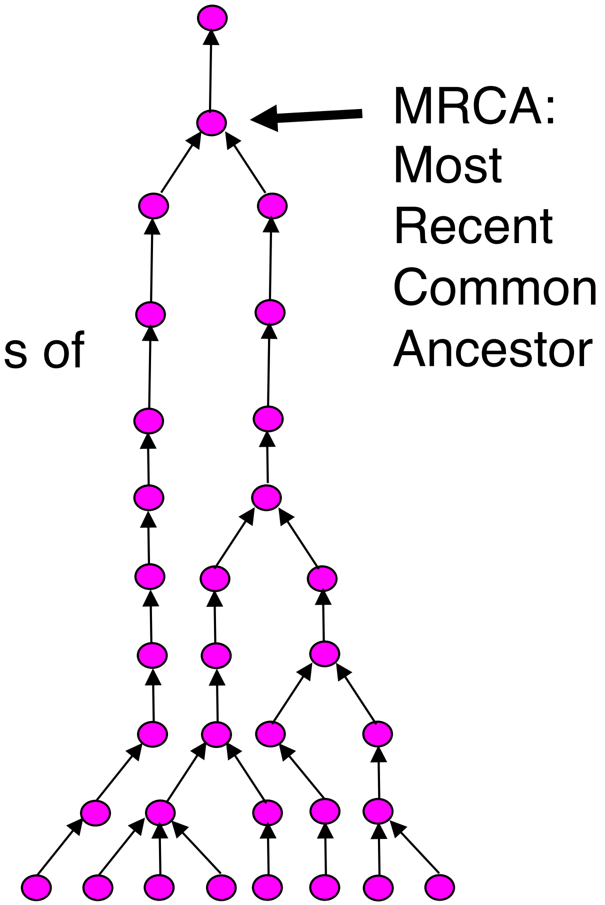


Remove individuals without descendants in the current population

Modeling genetic drift in finite populations



Rearrange the horizontal locations of the individuals

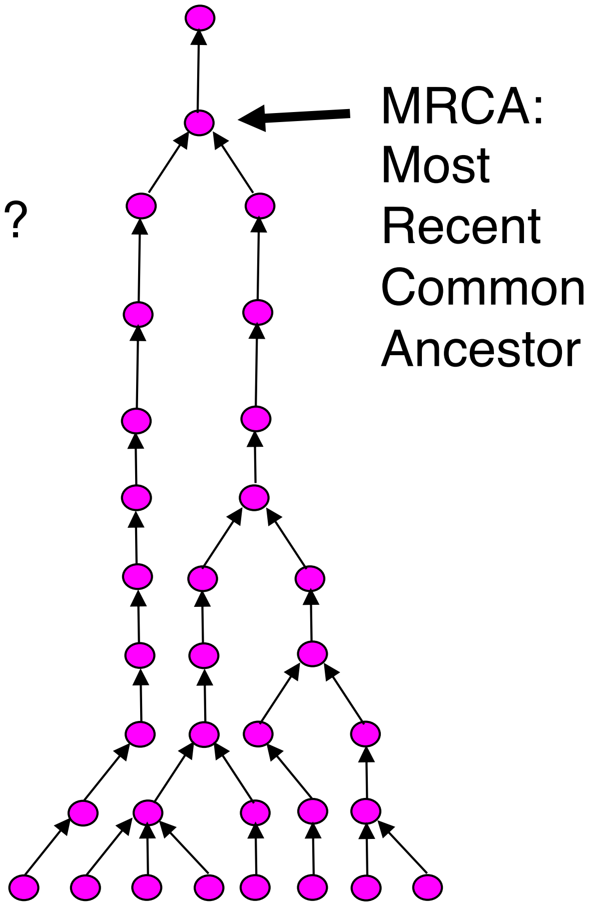


MRCA:
Most
Recent
Common
Ancestor

Modeling genetic drift in finite populations

How far back in time do we have to go to find the common ancestor of all individuals of today?

Answer provided by coalescent theory

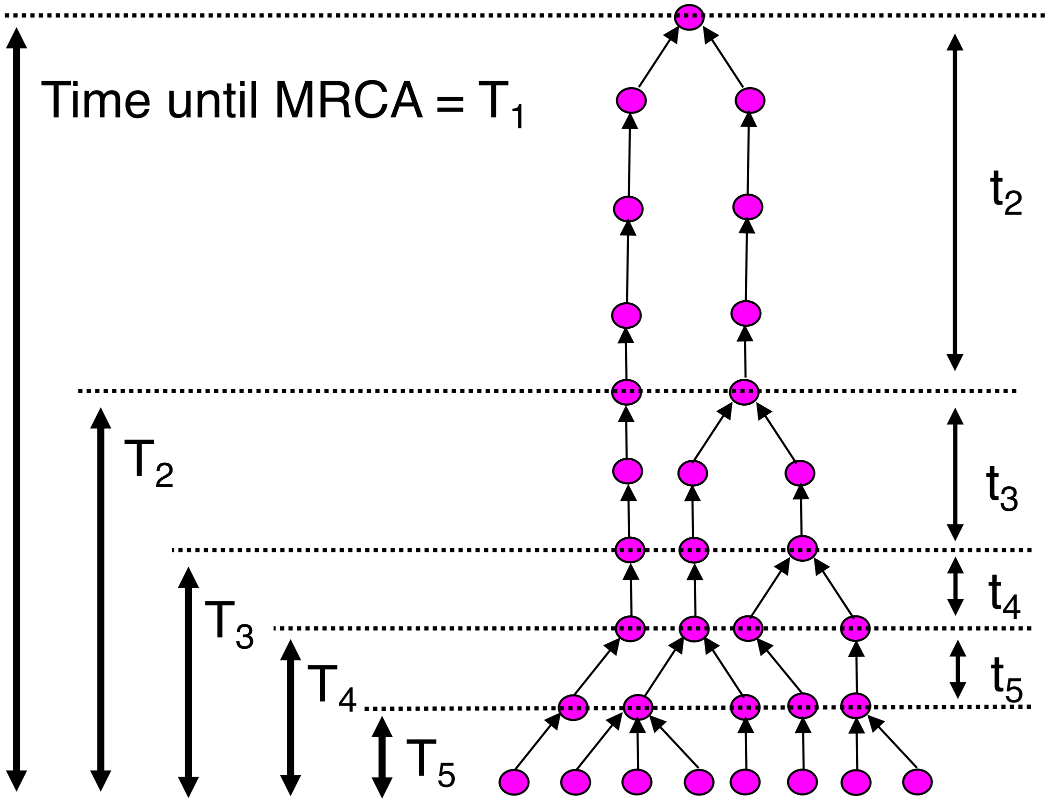


MRCA:
Most
Recent
Common
Ancestor

Coalescent theory

Approach:

- After going back a little bit in time one generally finds that only pairs of lineages *coalesce.*
- We can describe the tree by the times t_k that the population had *k* parallel lineages, i.e. times between *k* ancestors and *k-1* ancestors.



Coalescent theory

We will determine the distributions $p_k(t_k)$ of the time between the coalescence to *k* lineages and coalescence to *k-1* lineages. That is, we will determine the probability to spend t_k generations with *k* ancestors.

Start at the first generation in the past at which there were *k* ancestors.

For there to also be k ancestors in the previous generation, all *k* individuals need to have a different parent in the previous generation.
The probability that this happens in the population of *N* individuals is:

$$P(all\ k\ different) = 1\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{k-1}{N}\right)$$

Multiplying through and keeping only the terms of first order in *N* we get:

$$1-\frac{1}{N}(1+2+\cdots(k-1)) = 1-\frac{k(k-1)}{2N}$$

That is,

$$P(all\ k\ different) = 1\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{k-1}{N}\right) = 1-\frac{k(k-1)}{2N}+0\left(\frac{1}{N^2}\right)$$

Coalescent theory

$$P(all\ k\ different) = 1\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{k-1}{N}\right) = 1-\frac{k(k-1)}{2N}+0\left(\frac{1}{N^2}\right)$$

The probability that all *k* remain separate for at least *t* time steps is

$$P_k(t) \approx \left(1-\frac{k(k-1)}{2N}\right)^t$$

Recalling that the Taylor expansion of the exponential is

$e^x = 1 + x + \frac{x^2}{2!} + \cdots$ and that when $x \ll 1$ we can approximately write $e^x = 1 + x$

If we write $x = -\frac{k(k-1)}{2N}$, then $1-\frac{k(k-1)}{2N} \approx e^{-\frac{k(k-1)}{2N}}$

Thus, $P_k(t) \approx \left(1-\frac{k(k-1)}{2N}\right)^t = \left(e^{-\frac{k(k-1)}{2N}}\right)^t = e^{-\frac{k(k-1)t}{2N}}$

Coalescent theory

$$P(all\ k\ different) = 1\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{k-1}{N}\right) = 1-\frac{k(k-1)}{2N}+0\left(\frac{1}{N^2}\right)$$

The probability that all *k* remain separate for at least *t* time steps is

$$P_k(t) \approx e^{-\frac{k(k-1)t}{2N}}$$

The probability that all *k* remain separate for exactly *t* time steps is

$$p_k(t) = e^{-\frac{k(k-1)t}{2N}} - e^{-\frac{k(k-1)(t+1)}{2N}} = e^{-\frac{k(k-1)t}{2N}}\left(1-e^{-\frac{k(k-1)}{2N}}\right)$$

The average time that all *k* remain separate is

$$\int_{t=0}^{\infty} te^{-\frac{k(k-1)t}{2N}}\left(1-e^{-\frac{k(k-1)}{2N}}\right) = \left(1-e^{-\frac{k(k-1)}{2N}}\right)\int_{t=0}^{\infty} te^{-\frac{k(k-1)t}{2N}}$$

Coalescent theory

The average time that all *k* remain separate is

$$\int_{t=0}^{\infty} te^{-\frac{k(k-1)t}{2N}}\left(1-e^{-\frac{k(k-1)}{2N}}\right) = \left(1-e^{-\frac{k(k-1)}{2N}}\right)\int_{t=0}^{\infty} te^{-\frac{k(k-1)t}{2N}}$$

$1-e^{-\frac{k(k-1)}{2N}}$ = $\frac{k(k-1)}{2N}$ $\quad$ $\int_{t=0}^{\infty} te^{-\frac{k(k-1)t}{2N}}$ we integrate by parts

Starting from $\int g(t)f'(t) = g(t)f(t) - \int g'^{(t)}f(t)$

and denoting $g(t) = t$ and $f'(t) = e^{-\frac{k(k-1)t}{2N}}$

$$\int_{t=0}^{\infty} te^{-\frac{k(k-1)t}{2N}} = \left.\frac{te^{-\frac{k(k-1)t}{2N}}}{-\frac{k(k-1)}{2N}}\right|_{0}^{\infty} - \left.\frac{e^{\frac{k(k-1)t}{2N}}}{-\frac{k(k-1)}{2N}}\right|_{0}^{\infty}$$

$$\int_{t=0}^{\infty} te^{-\frac{k(k-1)t}{2N}}\left(1-e^{-\frac{k(k-1)}{2N}}\right) = \frac{k(k-1)}{2N}\frac{1}{\left(\frac{k(k-1)}{2N}\right)^{2}} = \frac{2N}{k(k-1)}$$

Coalescent theory

The average time that all *k* remain separate is

$$\langle t_k \rangle = \frac{2N}{k(k-1)}$$

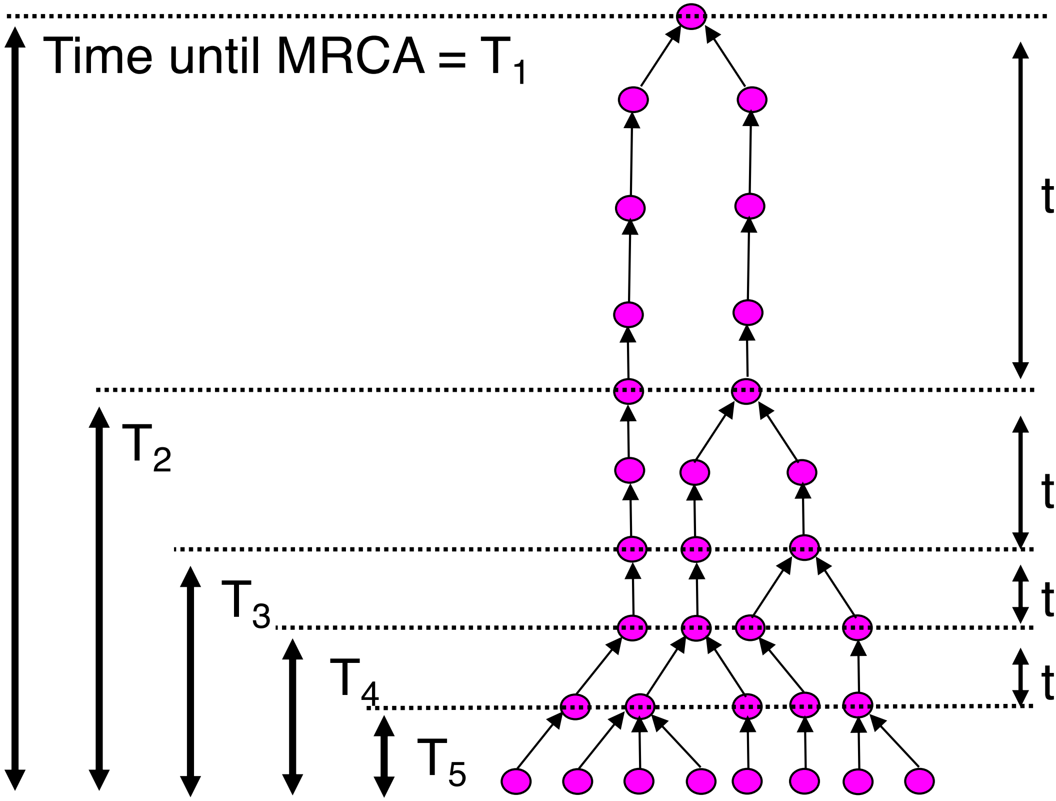
The times t_k are roughly exponentially distributed with the above mean length

This corresponds to the population size *N* divided by the number of pairs.

Time to the most recent common ancestor

Time T_1 to the MRCA is given by the sum over all time t_k:

$$T_1 = \sum_{k=2}^{N} \langle t_k \rangle = \sum_{k=2}^{N} \frac{2N}{k(k-1)} = \sum_{k=2}^{N} 2N\left(\frac{1}{k-1} - \frac{1}{k}\right) = 2N\left(1 - \frac{1}{N}\right) \approx 2N$$



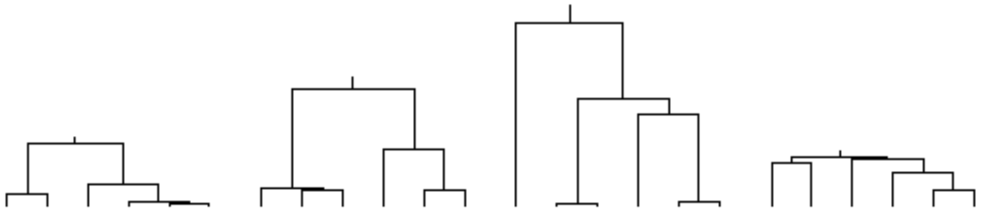
Time to the most recent common ancestor

Time T_1 to the MRCA is given by the sum over all time t_k:

$$T_1 \approx 2N$$

Thus, all members of the current population will have a common ancestor on average 2*N* generations in the past.

Of course there will be large variation from case to case:

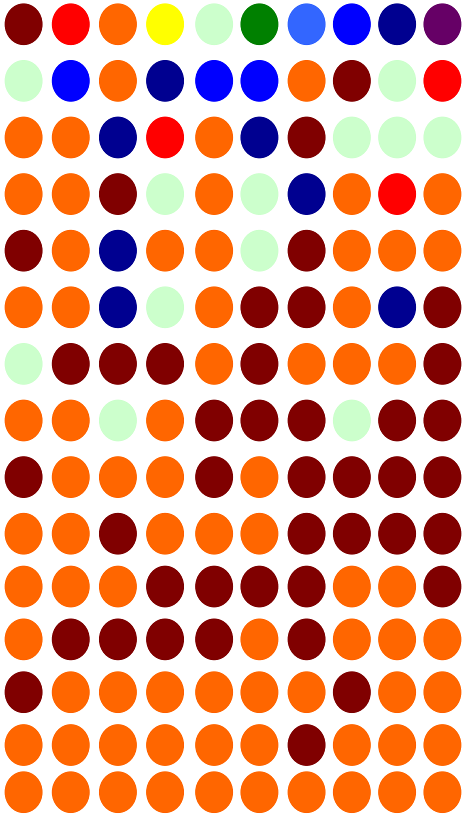


Coalescent theory studies the shape of the trees in more general models, including population size changes, periods of selection, etc.

What we discussed so far is:

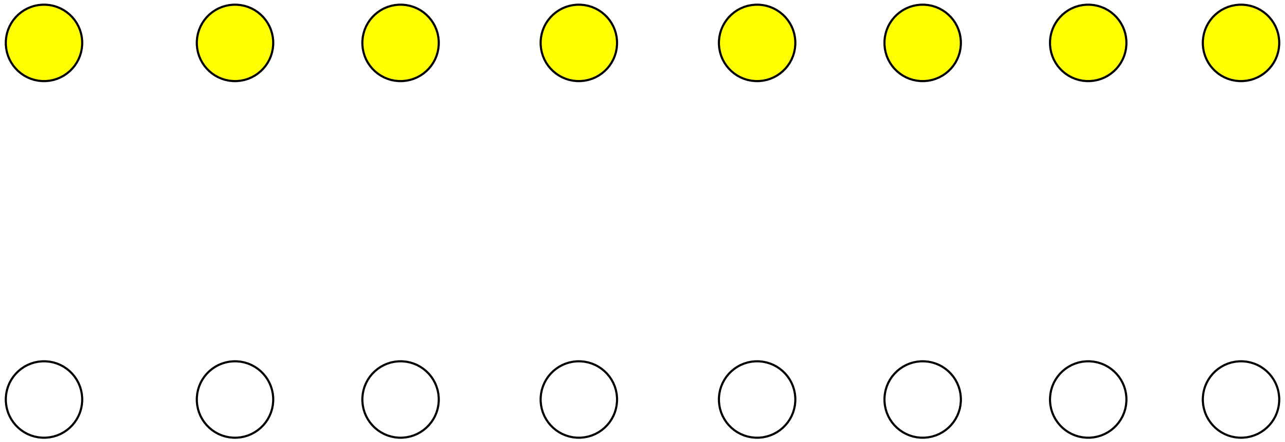
Genetic drift – loss of genetic diversity in finite populations as a result of sampling

- time since the most recent common ancestor – impact of population size



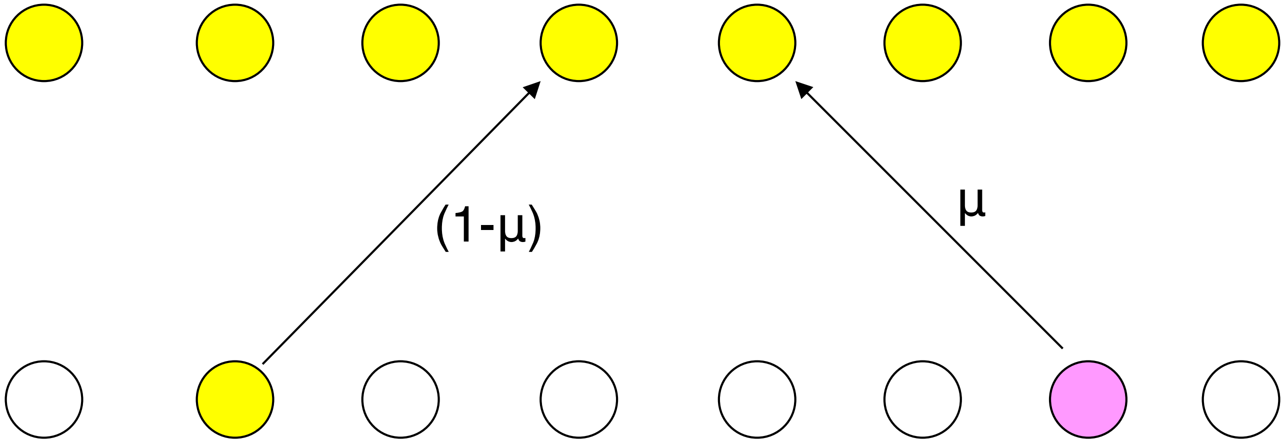
How does genetic diversity look like if mutations also occur?

Genetic drift with mutations



We again assume that all genotypes have the same fitness

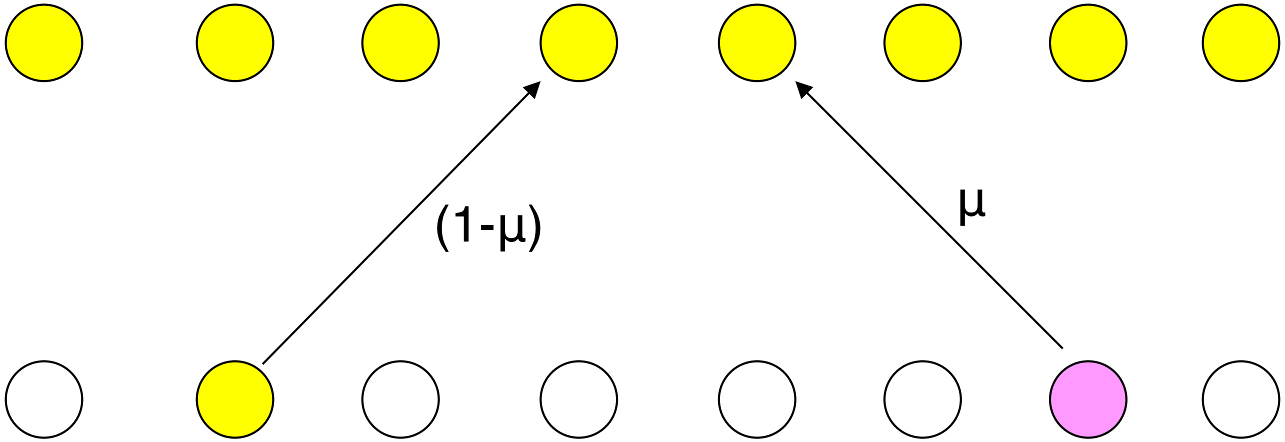
Genetic drift with mutations



We again assume that all genotypes have the same fitness

Each time an individual reproduces there is a probability β of mutation per base.
Thus, the probability that one of the L letters mutates is $L\beta$
Denote the mutation rate per individual by μ

Genetic drift with mutations

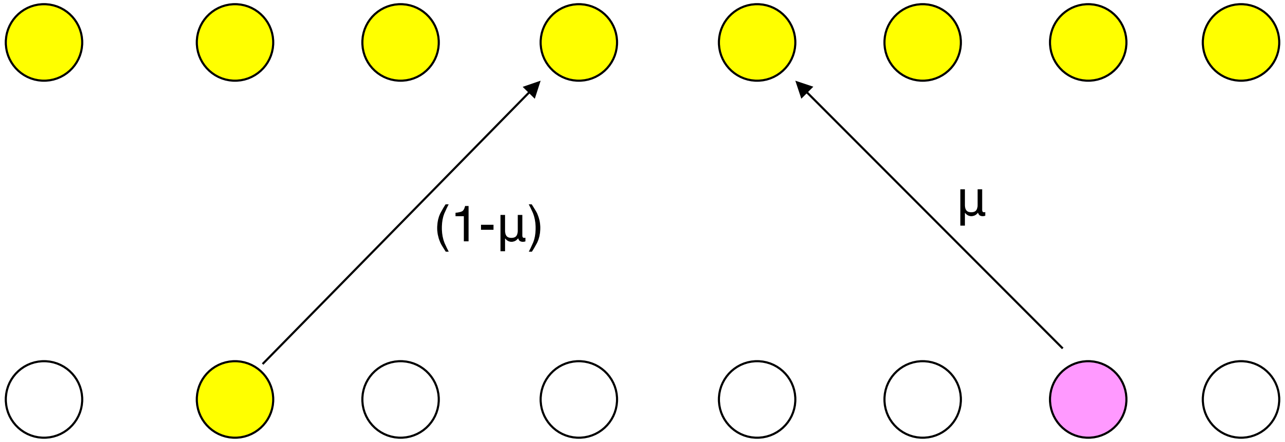


We again assume that all genotypes have the same fitness

Each time an individual reproduces there is a probability β of mutation per base.
Thus, the probability that one of the L letters mutates is $L\beta$
Denote the mutation rate per individual by μ

We further assume that every time a new letter mutates a *new genotype* is generated. We indicate this by giving the mutant a new color

Genetic drift with mutations



We again assume that all genotypes have the same fitness

Each time an individual reproduces there is a probability *β* of mutation per base. Thus, the probability that one of the *L* letters mutates is *Lβ*
Denote the mutation rate per individual by *μ*

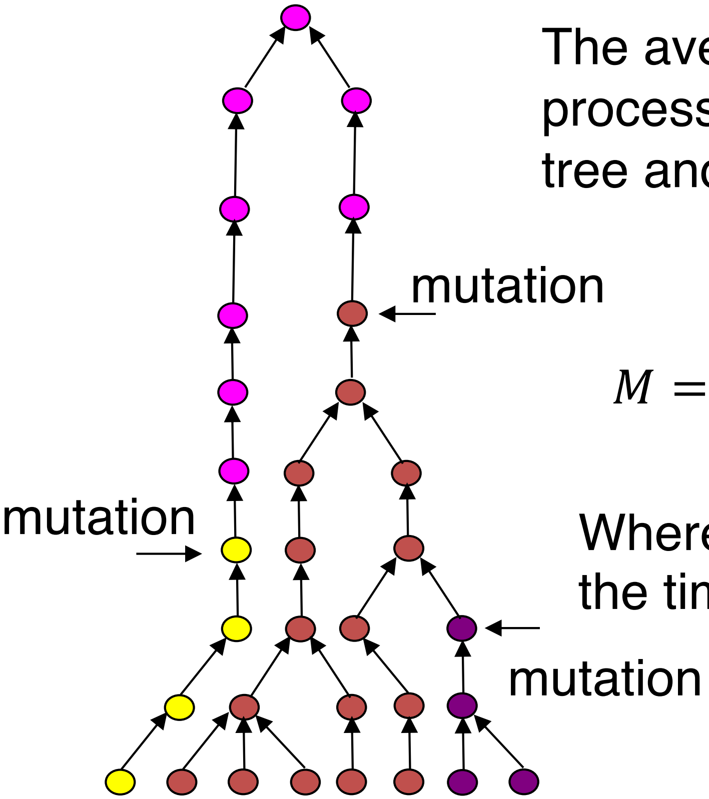
We further assume that every time a new letter mutates a *new genotype* is generated. We indicate this by giving the mutant a new color

To see what happens at longer times we ‘overlay’ this process on the coalescent

Genetic drift with mutations

The total amount of variation is proportional to the total number of mutations in the tree

The average number of mutations that occurred during the process is proportional to the number of individuals in the tree and the mutation rate per replication of each individual:

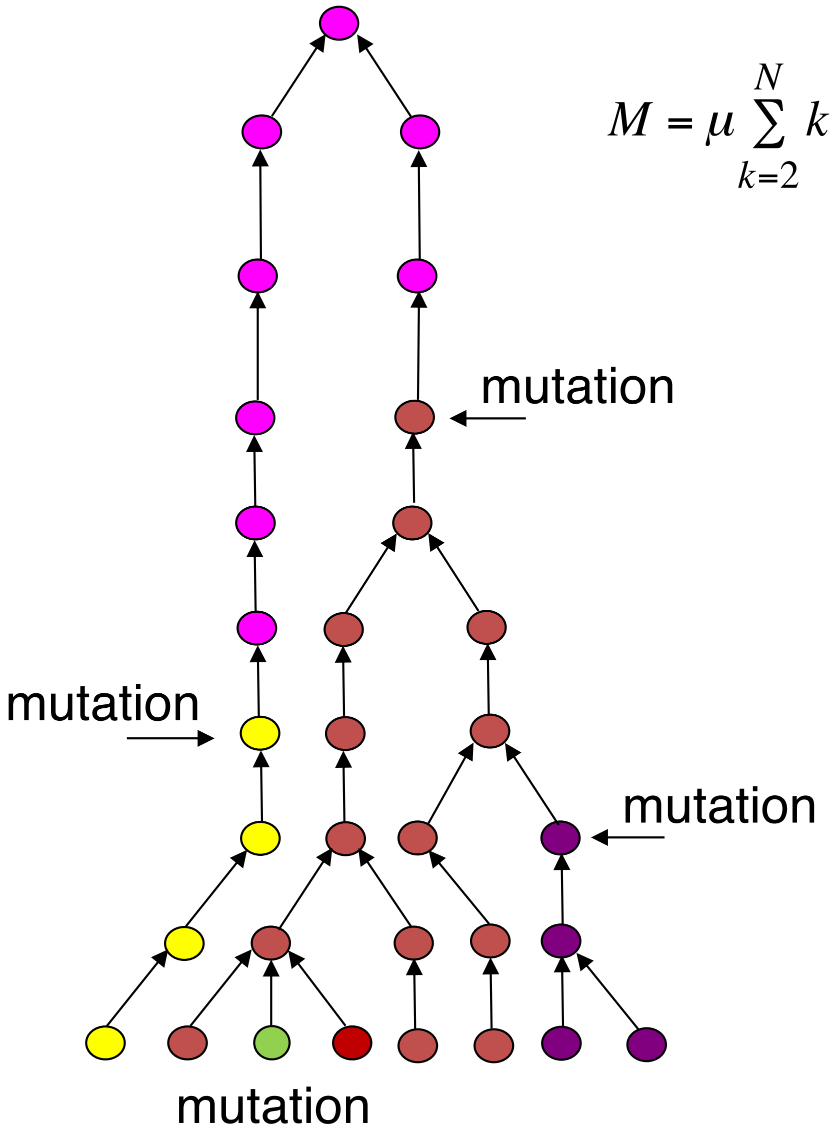


$$M = \mu \sum_{k=2}^{N} k\langle t_k\rangle = \mu \sum_{k=2}^{N} \frac{2N}{k-1} = 2\mu N \sum_{k=2}^{N} \frac{1}{k-1} \approx 2\mu N log(N)$$

Where $k\langle t_k\rangle$ denotes that k individuals were present during the time that the number of lineages went from k to k-1

Thus, the essential quantity controlling the amount of genetic variation that one observes is thus given by the *product μN* of the mutation rate *μ* and the population size *N*.

Average number of mutations

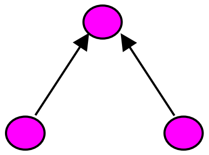


$$M = \mu \sum_{k=2}^{N} k \langle t_k \rangle = \mu \sum_{k=2}^{N} \frac{2N}{k-1} = 2\mu N \sum_{k=2}^{N} \frac{1}{k-1} \approx 2\mu N \log(N)$$

This is the number of mutations that occurred in this tree, not the number of differences we observe if we compare two individuals in the current population.

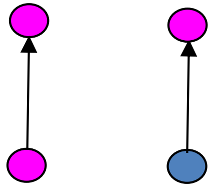
Can we relate the observed differences to the actually mutations that occurred in this phylogeny?

Following the genealogies of two individuals



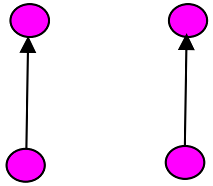
$\frac{1}{N}$

The pair coalesces



2μ

A mutation occurs



$1-\frac{1}{N}-2\mu$

Nothing happens

n mutations occur before the coalescence: *n* mutations, *k* ‘nothing happens’ (with k from 0 to ∞) and finally, coallescence:

$$\sum_{k=0}^{\infty}(2\mu)^n\left(1-2\mu-\frac{1}{N}\right)^k\frac{1}{N}\frac{(k+n)!}{k!\,n!}=\left(\frac{2\mu N}{2\mu N+1}\right)^n\frac{1}{2\mu N+1}$$

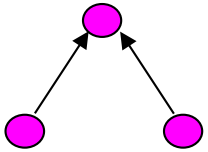
Which comes out the same as:

$$P(\text{mutation}\,|\,\text{something happens})=\frac{2\mu}{2\mu+\frac{1}{N}}=\frac{2\mu N}{2\mu N+1}$$

$$\left(\frac{2\mu N}{2\mu N+1}\right)^n\frac{1}{2\mu N+1}$$

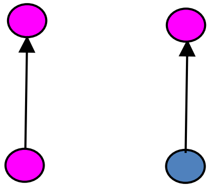
$$P(\text{coalescence}\,|\,\text{something happens})=1-\frac{2\mu}{2\mu+\frac{1}{N}}=\frac{1}{2\mu N+1}$$

Average pairwise distance



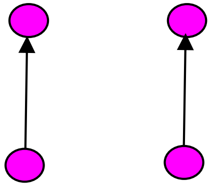
$\frac{1}{N}$

The pair coalesces



2μ

A mutation occurs



$1-\frac{1}{N}-2\mu$

Nothing happens

The probability that n mutations occur before the coalescence: $\left(\frac{2\mu N}{2\mu N+1}\right)^n \frac{1}{2\mu N+1}$

For the number of expected differences between two randomly chosen individuals

Let $x = 2\mu N$ and $y = \frac{x}{x+1}$. Then we can rewrite $\left(\frac{2\mu \mathrm{N}}{2\mu \mathrm{N}+1}\right)^n \frac{1}{2\mu \mathrm{N}+1} = y^n(1-y)$

and the expected number of mutations is

$$\langle n\rangle = \sum_i iy^i\left(1-y\right) = y(1-y)\sum_i iy^{i-1} = y(1-y)\frac{d}{dy}\sum_i y^i = y(1-y)\frac{d}{dy}\left[\frac{1}{1-y}\right] = \frac{y(1-y)}{(1-y)^2} = \frac{y}{1-y}$$

Making the back - substitution we have $\langle n\rangle = 2\mu N$

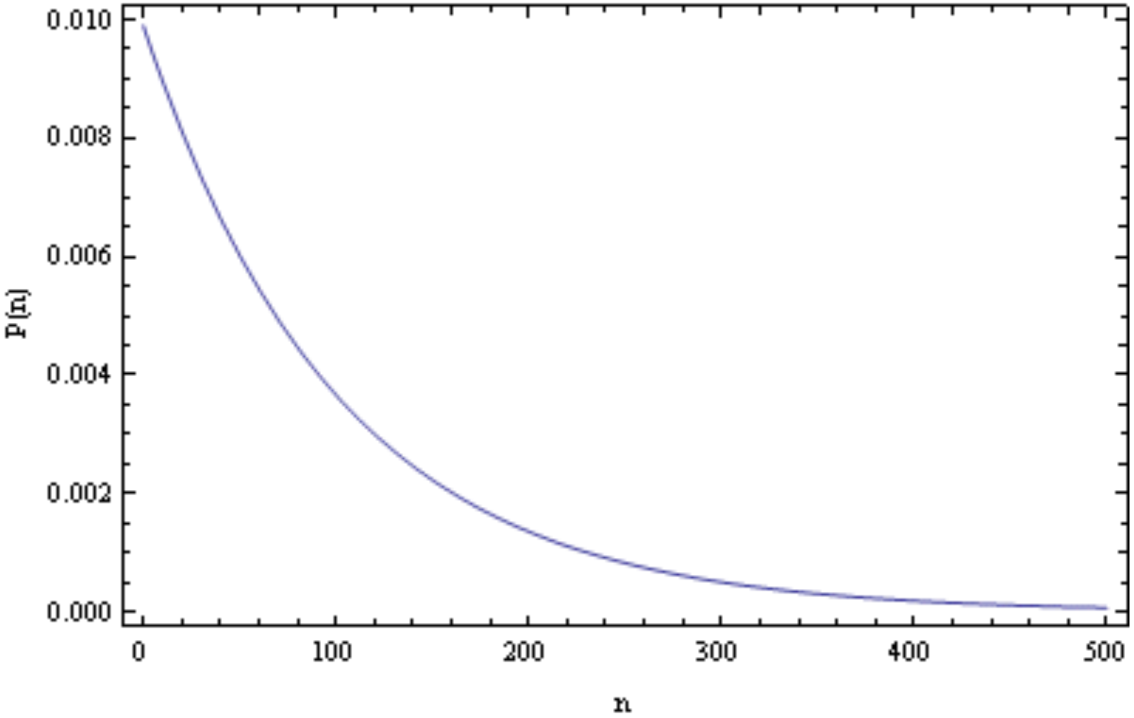
Pairwise distances

For neutral mutations in Lenski's experiment:

β $-$ mutation rate per base – 10^{-9}

L_e $-$ effective length of the genome – 1% of $4.6 * 10^6$ positions that are neutral

N $-$ population size $- 100'000$



$$\langle n \rangle = 2\mu N = 2\beta L_e N = 2 \cdot (10^{-9} \cdot 0.01 \cdot 4.6 \cdot 10^6) \cdot 10^5 = 9.2$$

Neutral evolution of a *single* site

- When we look at a single position the mutation rate is simply *β*.

- When picking a pair of random individuals from a population of size *N* the probability that the site has mutated *n* times since their common ancestor is:

$$P(n) = \left(\frac{2\beta N}{2\beta N + 1} \right)^n \frac{1}{2\beta N + 1}$$

And the average number of mutations at a given position for Lenski's experiment:

$$\langle n \rangle = 2\beta N = 2 \cdot 10^{-9} \cdot 10^5 \approx 2 \cdot 10^{-4}$$

Thus, at a single site almost all individuals will have *the same letter*, i.e. one letter will dominate the population.

We will now study the process by which the population can switch from one dominant letter to another.

To do this we use a slightly different stochastic model of evolution called the *Moran* model.

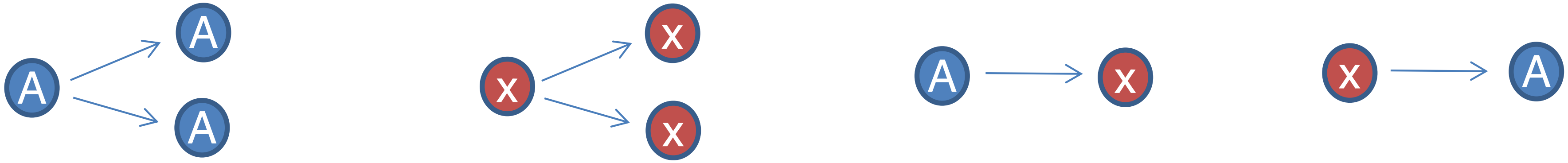
Moran model

- Evolutionary dynamics in a population of constant size: every time one individual reproduces another one is removed, to keep a fixed population size, N.

- Focus on a single position in the genome, and assume that individuals with letter A replicate at rate σ and all others have replication rate 1.
- Assume that initially we have $(N-n)$ individuals with letter A and n individuals with another letter ("mutants").
- Assume that, per unit time there is a probability μ that an individual undergoes a mutation at the chosen position.

- During a time interval dt the following events happen with indicated probabilities:

An A individual duplicates | Another type duplicates | An A-type mutates | Another type mutates



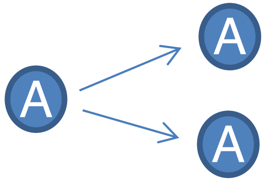
$\sigma(N-n)dt$ | ndt | $\mu(N-n)dt$ | $\frac{\mu n}{3}dt$

At each duplication a randomly chosen individual is removed:

$\left(1-\frac{n}{N}\right)$ | $\frac{n}{N}$

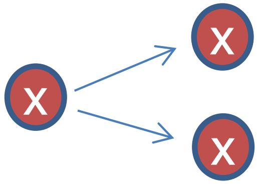
Moran model: selection and drift

An A individual duplicates



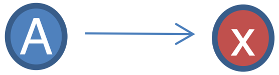
$$\sigma(N-n)dt$$

Another type duplicates



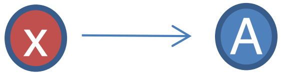
$$ndt$$

An A-type mutates



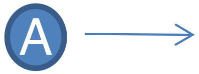
$$\mu(N-n)dt$$

Another type mutates

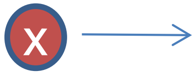


$$\frac{\mu n}{3}dt$$

At each duplication a randomly chosen individual is removed:



$$\left(1-\frac{n}{N}\right)$$

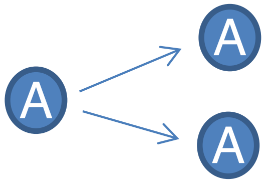


$$\frac{n}{N}$$

Deriving instantaneous rates of change in mutant frequency

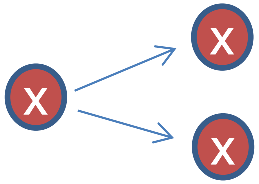
Moran model: selection and drift only, $\mu = 0$

An ‘A’ individual duplicates



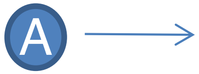
$\sigma(N-n)dt$

Another type duplicates

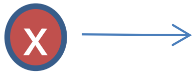


ndt

At each duplication a randomly chosen individual is removed:



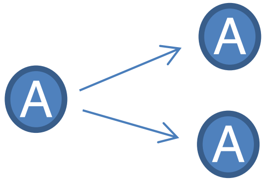
$\left(1-\frac{n}{N}\right)$



$\frac{n}{N}$

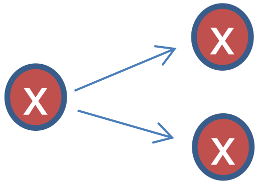
Moran model: selection and drift

An A individual duplicates



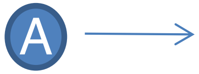
$\sigma(N-n)dt$

Another type duplicates

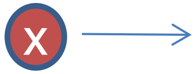


ndt

At each duplication a randomly chosen individual is removed:

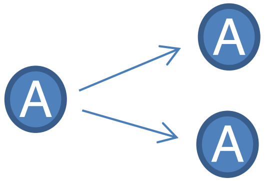


$\left(1-\frac{n}{N}\right)$

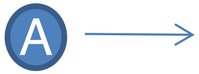


$\frac{n}{N}$

We then have:

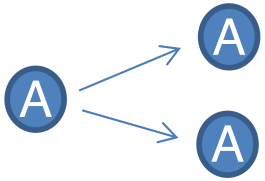


$\sigma(N-n)dt$

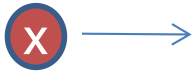


$\left(1-\frac{n}{N}\right)$

Scenario that decreases mutant frequency by 1/*N*

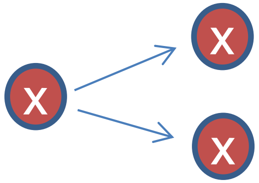


$\sigma(N-n)dt$

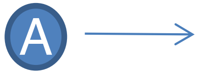


$\frac{n}{N}$

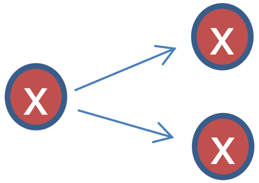
Scenario that increases mutant frequency by 1/*N*



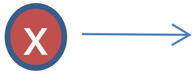
ndt



$\left(1-\frac{n}{N}\right)$



ndt



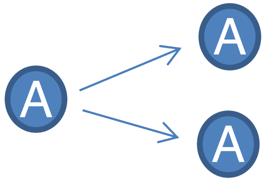
$\frac{n}{N}$

$$T\left(f,\delta f=-\frac{1}{N},dt\right)=\sigma(N-n)\frac{n}{N}dt=\sigma n(1-f)dt$$

$$T\left(f,\delta f=+\frac{1}{N},dt\right)=n\left(1-\frac{n}{N}\right)dt=nf(1-f)dt$$

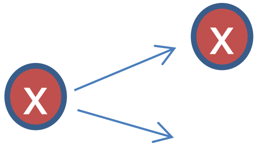
Moran model: selection, mutation, drift

An A individual duplicates



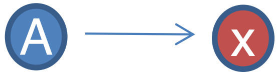
$\sigma(N-n)dt$

Another type duplicates



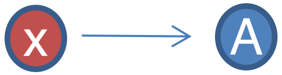
ndt

An A-type mutates



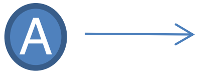
$\mu(N-n)dt$

Another type mutates

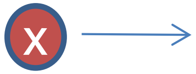


$\frac{\mu n}{3}dt$

At each duplication a randomly chosen individual is removed:



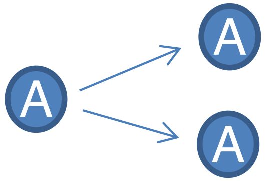
$\left(1-\frac{n}{N}\right)$



$\frac{n}{N}$

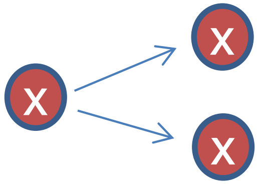
Or, in terms of $f=\frac{n}{N}$

An A individual duplicates



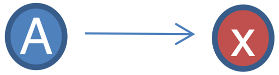
$N\sigma(1-f)dt$

Another type duplicates



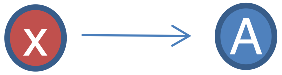
$Nfdt$

An A-type mutates



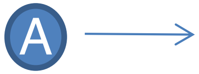
$N\mu(1-f)dt$

Another type mutates

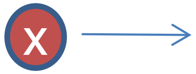


$\frac{\mu}{3}Nfdt$

At each duplication a randomly chosen individual is removed:



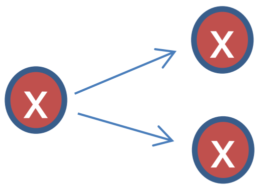
$(1-f)$



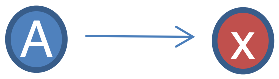
f

Selection, mutation and drift in the Moran model

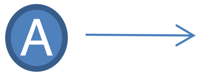
Scenarios that increase the mutant frequency by 1/*N*



$Nf\,dt$

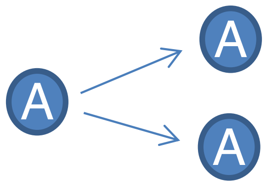


$N\mu(1-f)dt$

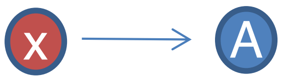


$(1-f)$

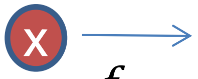
Scenarios that decrease the mutant frequency by 1/*N*



$N\sigma(1-f)dt$



$\frac{\mu}{3}Nf\,dt$



f

Probability of changing by 1 individual during time dt:

$$T\left(f, \delta f = -\frac{1}{N}, dt\right) = N\left[\sigma f(1-f) + \frac{\mu}{3} f\right] dt$$

$$T\left(f, \delta f = +\frac{1}{N}, dt\right) = N[f(1-f) + \mu(1-f)]dt$$

Probability of fixation

How like it is that a mutant that appears in the population will eventually take over the entire population?

Diffusion models: probability of fixation

The analysis that we will now perform is an example of the *diffusion approximation* in population genetics, introduced by the population geneticist Motoo Kimura.

Definitions:

$f =$ Fraction of the population with the 'mutant' genotype.

$\pi(f)$= Probability that the mutant will eventially take over the population given that it starts from a fraction f.

$T(f,\delta f,dt)$ = Probability that the fraction changes from f to $f+\delta f$ in a small interval dt.

The fixation probability distribution obeys the *Master equation*:

$$\pi(f) = \int T\left(f,\delta f,dt\right)\pi\left(f+\delta f\right)d\delta f$$

That is, the probability of fixation starting from fraction f is the probability to transfer to $f+\delta f$ times the probability to fix from $f+\delta f$, integrated over all possible changes δf.

Probability of fixation

The fixation probability distribution obeys the *Master equation*:

$$\pi(f) = \int T(f, \delta f, dt)\pi(f + \delta f)d\delta f$$

Since the changes δf are *small* in a small time interval dt we can expand in δf

Recall the Taylor expansion of a function $f(x)$ that is infinitely differentiable about a value a:

$$f(x) = f(a) + (x-a)\frac{f'(a)}{1!} + (x-a)^2\frac{f''(a)}{2!} + \cdots = \sum_{i=0}^{\infty}(x-a)^i\frac{f^{(i)}(a)}{i!}$$

In our case,

$$\pi(f + \delta f) = \pi(f) + \delta f\ \pi'(f) + \frac{(\delta f)^2}{2}\pi''(f) + \cdots$$

Probability of fixation

The fixation probability distribution obeys the *Master equation*:

$$\pi(f) = \int T(f, \delta f, dt)\pi(f + \delta f)d\delta f$$

Since the changes δf are *small* in a small time interval dt we can expand in δf:

$$\pi(f) = \int T(f, \delta f, dt)\left[\pi(f) + \delta f\, \pi'(f) + \frac{(\delta f)^2}{2}\pi''(f)\right]d\delta f$$

And rewriting

$$\pi(f) = \int T(f, \delta f, dt)\, \pi(f)d\delta f + \int T(f, \delta f, dt)\delta f\, \pi'(f)\, d\delta f + \int T(f, \delta f, dt)\frac{(\delta f)^2}{2}\pi''(f)d\delta f$$

$$\pi(f) = \pi(f)\int T(f, \delta f, dt)\, d\delta f + \pi'(f)\int \delta f\, T(f, \delta f, dt)\, d\delta f + \pi''(f)\int \frac{(\delta f)^2}{2}T(f, \delta f, dt)\, d\delta f$$

Using $\int T(f, \delta f, dt)\, d\delta f = 1$ and defining the average change and average squared-change

$$\int \delta f\, T(f, \delta f, dt)\, d\delta f = \langle \delta f \rangle_f \qquad \int (\delta f)^2 T(f, \delta f, dt)\, d\delta f = \langle (\delta f)^2 \rangle_f$$

We have

$$\pi(f) = \pi(f) + \pi'(f)\,\langle \delta f \rangle_f + \pi''(f)\frac{\langle (\delta f)^2 \rangle_f}{2} \quad \text{or} \quad \pi'(f)\,\langle \delta f \rangle_f + \pi''(f)\,\frac{\langle (\delta f)^2 \rangle_f}{2} = 0$$

Probability of fixation

$$\pi'(f)\,\langle\delta f\rangle_f + \pi''(f)\,\frac{\langle(\delta f)^2\rangle_f}{2} = 0$$

Define $X(f) = \pi'(f)$

We then have $X(f)\,\langle\delta f\rangle_f + X'(f)\,\frac{\langle(\delta f)^2\rangle_f}{2} = 0$ or $\frac{X\prime(f)}{X(f)} = -2\frac{\langle\delta f\rangle_f}{\langle(\delta f)^2\rangle_f}$

$$d[log(X(f))] = -2\frac{\langle\delta f\rangle_f}{\langle(\delta f)^2\rangle_f}\,df \Leftrightarrow log\big(X(f)\big) = -2\int\frac{\langle\delta f\rangle_f}{\langle(\delta f)^2\rangle_f}\,df + c$$

Finally, $\pi'(f) = X(f) = C\;exp\left(-2\int\frac{\langle\delta f\rangle_f}{\langle(\delta f)^2\rangle_f}\,df\right)$

And $\pi(f) = C' + C\int exp\left(-2\int\frac{\langle\delta f\rangle_f}{\langle(\delta f)^2\rangle_f}\,df\right)df$

Probability of fixation

$$\pi(f) = C' + C\int exp\left(-2\int \frac{\langle \delta f\rangle_f}{\langle(\delta f)^2\rangle_f}df\right)df$$ Let's call $-2\frac{\langle \delta f\rangle_f}{\langle(\delta f)^2\rangle_f} = g(f)$

Then $\int exp\left(-2\int \frac{\langle \delta f\rangle_f}{\langle(\delta f)^2\rangle_f}df\right)df$ can be written as $\int e^{\int g(f)\,df}df = \int e^{G(f)}df$

We also have some boundary conditions:

Starting from $f = 0$ the mutant will never take over the population, thus $\pi(0) = 0$

Starting from $f = 1$ the mutant has taken over the population, thus $\pi(1) = 1$

In other words

$C' + C\int e^{G(f)}df\Big|_0 = 0$ and $C' + C\int e^{G(f)}df\Big|_1 = 1$

This gives us the constants

$$C = \frac{1}{\int e^{G(f)}df\Big|_1 - \int e^{G(f)}df\Big|_0} \qquad C' = \frac{-\int e^{G(f)}df\Big|_0}{\int e^{G(f)}df\Big|_1 - \int e^{G(f)}df\Big|_0}$$

Probability of fixation

$$C = \frac{1}{\int e^{G(f)} df\big|_1 - \int e^{G(f)} df\big|_0} \qquad C' = \frac{-\int e^{G(f)} df\big|_0}{\int e^{G(f)} df\big|_1 - \int e^{G(f)} df\big|_0}$$

$$\pi(f) = C' + C\int exp\left(-2\int \frac{\langle \delta f\rangle_f}{\langle (\delta f)^2\rangle_f} df\right) df = C' + C\int e^{G(f)} df$$

$$\pi(f) = \frac{-\int e^{G(f)} df\big|_0}{\int e^{G(f)} df\big|_1 - \int e^{G(f)} df\big|_0} + \frac{\int e^{G(f)} df\big|_f}{\int e^{G(f)} df\big|_1 - \int e^{G(f)} df\big|_0} = \frac{\int e^{G(f)} df\big|_f - \int e^{G(f)} df\big|_0}{\int e^{G(f)} df\big|_1 - \int e^{G(f)} df\big|_0}$$

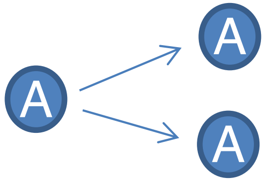
This leads us to $\pi(f) = \dfrac{\int_0^f e^{-2\int_0^x \frac{\langle \delta f\rangle_y}{\langle (\delta f)^2\rangle_y} dy} dx}{\int_0^1 e^{-2\int_0^x \frac{\langle \delta f\rangle_y}{\langle (\delta f)^2\rangle_y} dy} dx}$

We now work this out for the single-site Moran model that we introduced.

For this, we need to determine $\langle \delta f\rangle_y$ and $\langle (\delta f)^2\rangle_y$ for the Moran model.

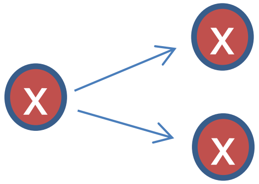
Moran model: selection and drift

An ‘A’ individual duplicates



$\sigma(N-n)dt$

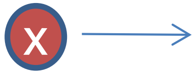
Another type duplicates



ndt

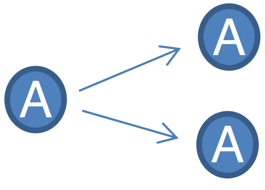
At each duplication a randomly chosen individual is removed:

$\left(1-\frac{n}{N}\right)$

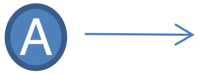


$\frac{n}{N}$

We then have:

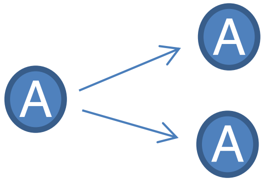


$\sigma(N-n)dt$

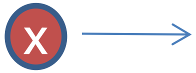


$\left(1-\frac{n}{N}\right)$

Scenario that decreases mutant frequency by 1/*N*

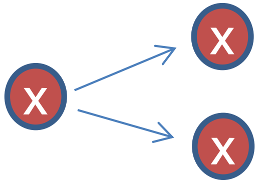


$\sigma(N-n)dt$

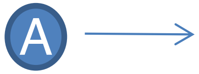


$\frac{n}{N}$

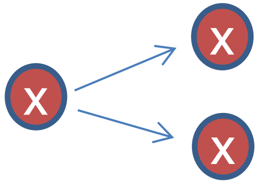
Scenario that increases mutant frequency by 1/*N*



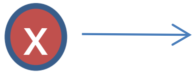
ndt



$\left(1-\frac{n}{N}\right)$



ndt



$\frac{n}{N}$

$$T\left(f, \delta f=-\frac{1}{N}, dt\right)=\sigma(N-n)\frac{n}{N}dt=\sigma n(1-f)dt$$

$$T\left(f, \delta f=+\frac{1}{N}, dt\right)=n\left(1-\frac{n}{N}\right)dt=nf(1-f)dt$$

Moran model: selection and drift

$$\delta f = -\frac{1}{N} \qquad T\left(f, \delta f = -\frac{1}{N}, dt\right) = \sigma(N-n)\frac{n}{N}dt = \sigma n(1-f)dt$$

$$\delta f = +\frac{1}{N} \qquad T\left(f, \delta f = +\frac{1}{N}, dt\right) = n\left(1-\frac{n}{N}\right)dt = nf(1-f)dt$$

Thus, the averages we need are:

$$\langle \delta f \rangle_f = \frac{1}{N} nf(1-f) - \frac{1}{N}\sigma n(1-f) = f(1-f)(1-\sigma)dt$$

$$\langle (\delta f)^2 \rangle_f = \frac{1}{N^2} nf(1-f) + \frac{1}{N^2}\sigma n(1-f) = \frac{f(1-f)(1+\sigma)}{N}dt$$

And finally the ratio: $\frac{\langle \delta f \rangle_f}{\langle (\delta f)^2 \rangle_f} = N\frac{1-\sigma}{1+\sigma}$

Moran model: selection and drift

Substituting $\frac{\langle\delta f\rangle_f}{\langle(\delta f)^2\rangle_f} = N\frac{1-\sigma}{1+\sigma}$ in $\pi(f) = \frac{\int_0^f e^{-2\int_0^x \frac{\langle\delta f\rangle_y}{\langle(\delta f)^2\rangle_y}dy}dx}{\int_0^1 e^{-2\int_0^x \frac{\langle\delta f\rangle_y}{\langle(\delta f)^2\rangle_y}dy}dx}$ gives:

$$\int_0^x \frac{\langle\delta f\rangle_y}{\langle(\delta f)^2\rangle_y}dy = \int_0^x N\frac{1-\sigma}{1+\sigma}dy = Nx\frac{1-\sigma}{1+\sigma}$$

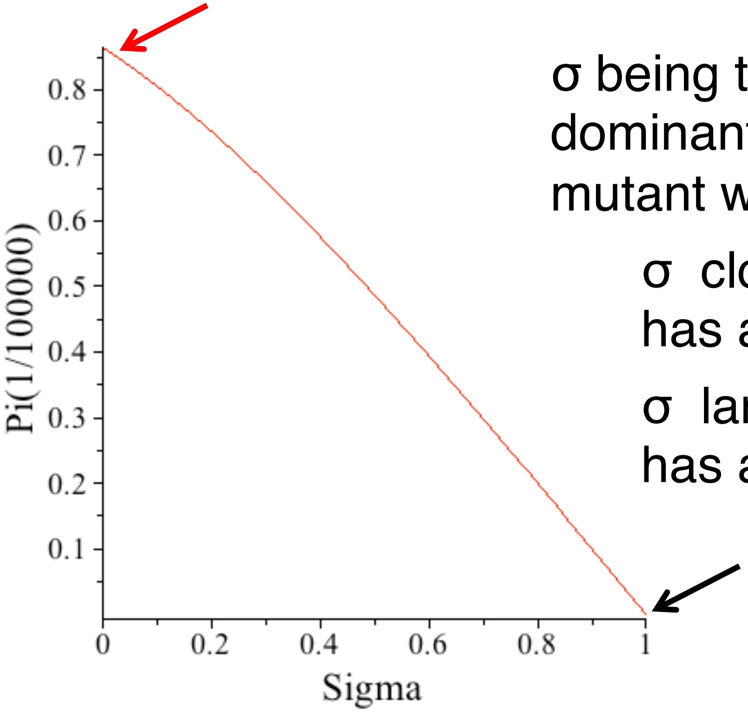
$$\int_0^f e^{-2Nx\frac{1-\sigma}{1+\sigma}}dx = \frac{1}{-2N\frac{1-\sigma}{1+\sigma}}\left(e^{-2Nf\frac{1-\sigma}{1+\sigma}} - 1\right) \quad \int_0^1 e^{-2Nx\frac{1-\sigma}{1+\sigma}}dx = \frac{1}{-2N\frac{1-\sigma}{1+\sigma}}\left(e^{-2N\frac{1-\sigma}{1+\sigma}} - 1\right)$$

Thus, $\pi(f) = \frac{1-e^{2Nf\frac{\sigma-1}{\sigma+1}}}{1-e^{2N\frac{\sigma-1}{\sigma+1}}}$ and starting from a single mutant, $f = \frac{1}{N}$,

$$\pi\left(\frac{1}{N}\right) = \frac{1-e^{2\frac{\sigma-1}{\sigma+1}}}{1-e^{2N\frac{\sigma-1}{\sigma+1}}}$$

Selection and drift in the Moran model

$$\pi\left(\frac{1}{N}\right) = \frac{1 - e^{2\frac{\sigma-1}{\sigma+1}}}{1 - e^{2N\frac{\sigma-1}{\sigma+1}}}$$



σ being the relative replication rate of the dominant genotype with respect to the mutant which has replication rate 1,

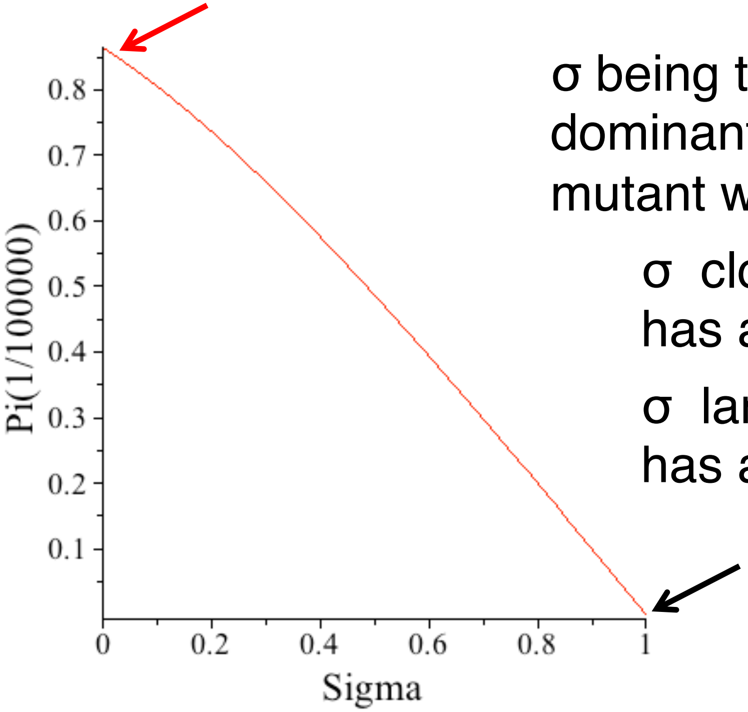
σ close to 0 means that the mutant has a large fitness *advantage*

σ larger than 1 means that the mutant has a large fitness *disadvantage*

- We find that even a mutant with a large fitness advantage (advantageous mutation) has a less than 1 probability of fixation -> it may need to appear more than once in the population before it takes over.
- If the mutant has a fitness disadvantage (deleterious mutation) it will have a small probability of fixation.

Selection and drift in the Moran model

$$\pi\left(\frac{1}{N}\right) = \frac{1 - e^{2\frac{\sigma-1}{\sigma+1}}}{1 - e^{2N\frac{\sigma-1}{\sigma+1}}}$$



σ being the relative replication rate of the dominant genotype with respect to the mutant which has replication rate 1,

σ close to 0 means that the mutant has a large fitness *advantage*

σ larger than 1 means that the mutant has a large fitness *disadvantage*

For mutations with small deleterious effect we write $\sigma - 1 = s$ and then we have

$$e^{\frac{2(\sigma-1)}{\sigma+1}} = e^{\frac{2s}{2+s}} \qquad e^{\frac{2N(\sigma-1)}{\sigma+1}} = e^{\frac{2Ns}{2+s}}$$

Since for $s \approx 0$, $\frac{2s}{2+s} \approx s \Rightarrow \pi\left(\frac{1}{N}\right) \approx \frac{1-e^{s}}{1-e^{Ns}}$

and because $1 - e^{s} \approx -s \Rightarrow \pi\left(\frac{1}{N}\right) \approx \frac{s}{e^{Ns}-1} = \frac{1}{N}\frac{Ns}{e^{Ns}-1}$

Selection and drift in the Moran model

Because $\lim_{s \to 0} \left(\frac{1}{N} \ \frac{Ns}{e^{Ns}-1} \right) = \lim_{s \to 0} \left(\frac{1}{N} \ \frac{N}{Ne^{Ns}} \right) = \frac{1}{N}$

neutral mutations ($s = 0$) spread with probability $\frac{1}{N}$

The effect of selection (s small, but not 0), depends on the effective parameter Ns

$$\pi\left(\frac{1}{N}\right) = \frac{1}{N} \frac{Ns}{e^{Ns}-1}$$