

Introduction to the coalescent

Magnus Nordborg

Gregor Mendel Institute

The Gregor Mendel Institute

- Basic research institute focused on plant biology
- Owned by the Austrian Academy of Sciences
- Located at the **Vienna BioCenter**, a campus with multiple institutions and companies (roughly 1700 researchers from at least 40 countries) located in the one of world's most liveable cities
- Excellent opportunities at all levels — apply!
- (Yes, it English-speaking. . .)

The Coming of Data

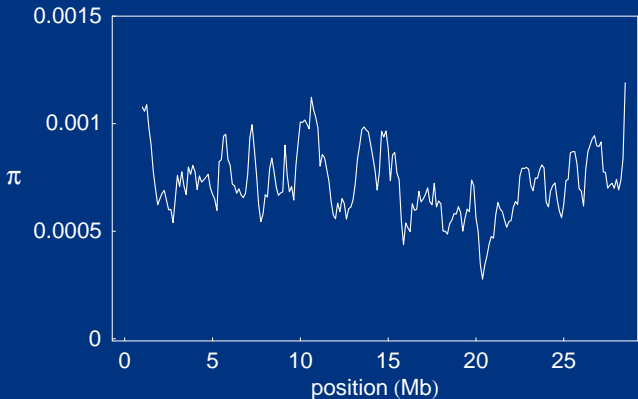
- Molecular polymorphism data started to become available in the 1960's
- For example, Esterase-2 in various *Drosophila* species:

Species	n	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇
<i>willistoni</i>	582	559	11	7	2	1	1	1
<i>tropicalis</i>	298	234	52	4	4	2	1	1
<i>equinoxialis</i>	376	361	5	4	3	3		
<i>simulans</i>	308	91	76	70	57	12	1	1

“We cranked the handle and nothing came out!”

- Classical population genetics had little to say about these data
- What should we expect to see? How many alleles? Which level of heterozygosity?
- To answer these questions, a model was needed. . .

Example: Polymorphism in the human genome (chromosome 21)



Making sense of sequence data

- What do we expect the level of polymorphism to be?
- What is the variance?
- What is the distribution along the genome? Coding, non-coding, *etc*?
- Does it matter if we sample from the same individual or from different individuals?
- Does it matter if we sample from the same geographic location or from different locations?
- What about differences between species?

The neutral model

- Kimura's Neutral Theory provided a useful null model...
- ...but the theory got hard because stochastic processes were needed
- *Genetic polymorphism data represent the outcome of a single, highly complex, non-repeatable evolutionary history*
- The stochastic process known as “the coalescent” presents a coherent statistical framework for analyzing genetic polymorphism data

History

The coalescent has played a central role population genetics for well over 30 years:

- Kingman (1982) — definitive mathematical treatment
- Hudson (1983) — recombination!
- Tajima (1983)
- Arguments anticipating the coalescent had appeared much earlier...

The first survey of DNA sequence polymorphism was published by Kreitman in 1983...

Importance

Coalescent models follow the genealogy (ancestry) of genes backward in time, starting from the present. This turns out to be a very powerful way of thinking about genetic polymorphism:

- elegant mathematics
- powerful simulation algorithms
- explicit likelihood calculations

An intuitive understanding of coalescent models is essential for anyone analyzing polymorphism data...

Three insights

The coalescent is based on three insights. One is mathematical and will be described in a moment; two are conceptual insights about selective neutrality:

- “state” can be separated from “descent”
- the properties of a sample depends only on *their* genealogy, which can be modeled backward in time

Some English

coalesce (*verb*) merge, combine, fuse — used here in the context of ancestral lineages

coalescence (*noun*) coincidence, equality, identity, sameness, fusion, junction — used here to denote the event of ancestral lineages coalescing

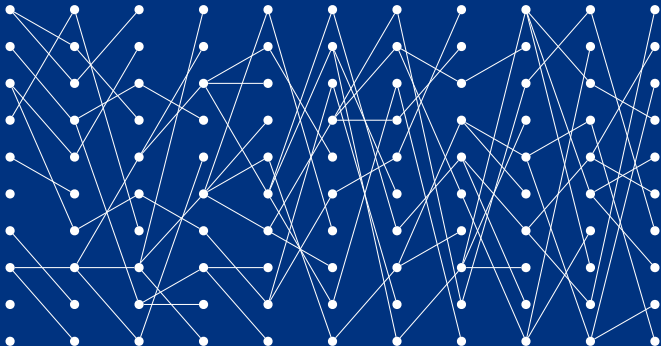
The (neutral) Wright-Fisher model

- A constant-size population of N clonal organisms
- Discrete generations
- Each new generation is formed by randomly sampling N parents with replacement from the current generation

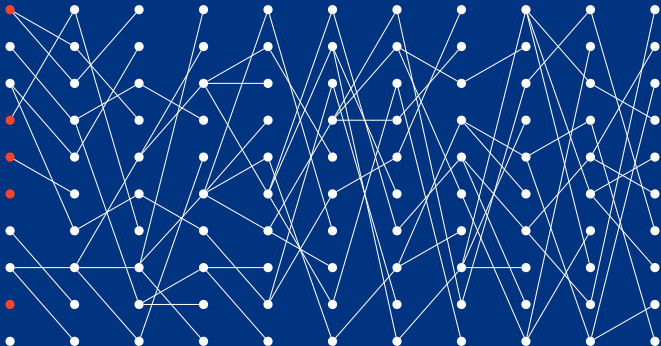
Thus:

- The number of offspring contributed by a particular individual is $\text{Bin}(N, 1/N)$
- The joint distribution of offspring numbers is symmetrically multinomial

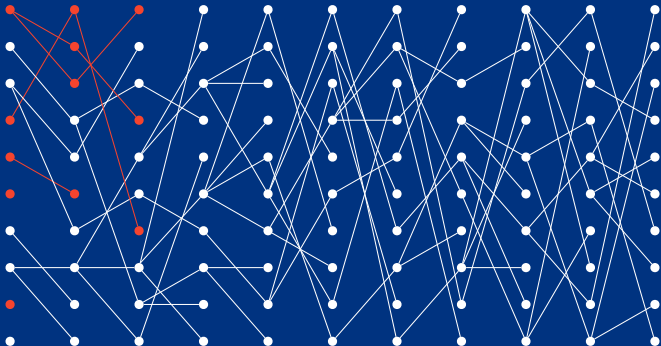
The Wright-Fisher model



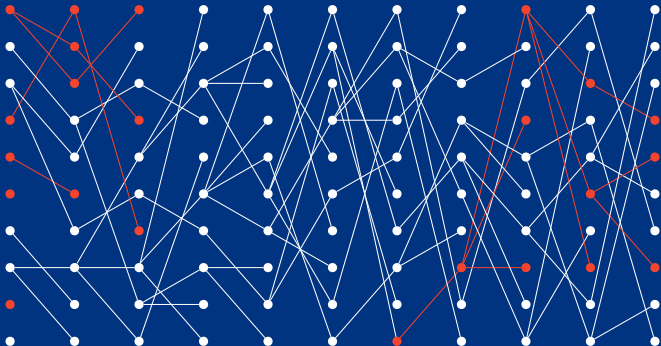
The Wright-Fisher model



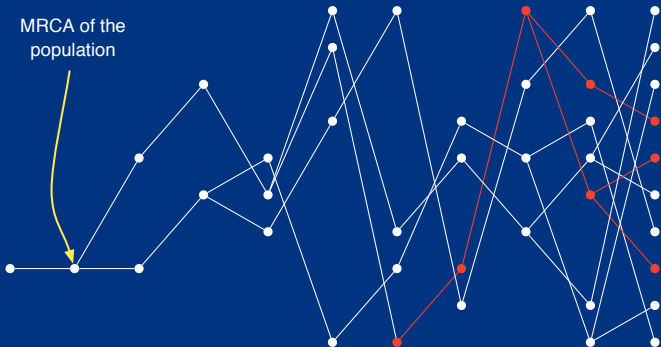
The Wright-Fisher model



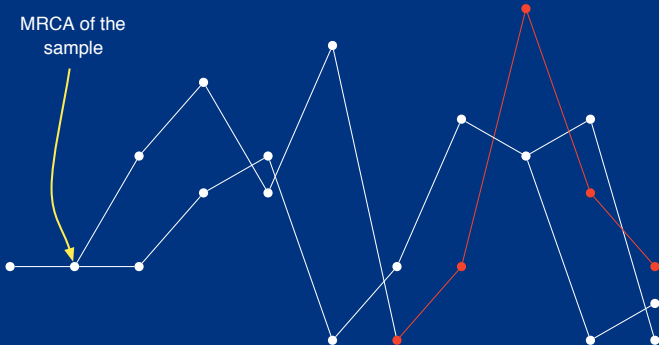
The Wright-Fisher model



The Wright-Fisher model



The Wright-Fisher model



Summary

Under neutrality, the joint effects of random reproduction (“genetic drift”) and mutation on the distribution of a sample may be modeled by:

1. generating the genealogy backward in time
2. superimposing mutations forward in time

The coalescent and classical population genetics

The main difference is one of perspective:

- Classical models were prospective: given some starting conditions, what will happen? This is useful for thinking about how evolution might work; less useful for interpreting data
- Coalescent models are retrospective: given the present, what could have happened? This is more natural for thinking about data

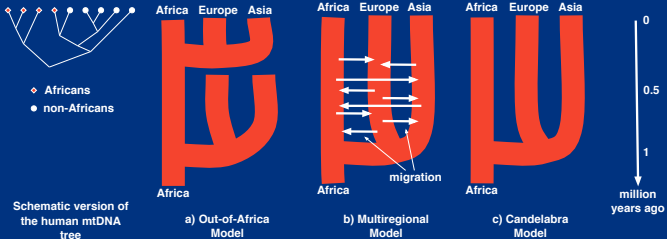
Implications for how we (should) view polymorphism data

Typically, we want to infer things about the evolutionary process that gave rise to the data. Almost always, this process affects only the genealogy. Then:

- The observed polymorphisms are of interest only because they contain information about the unobserved underlying genealogy
- The genealogy is of interest only because it contains information about the evolutionary process

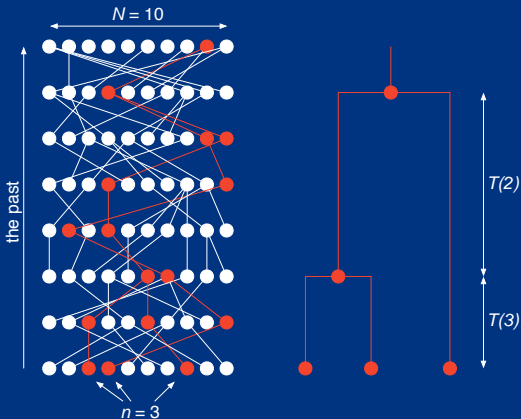
Note that no matter how many individuals we sample, there is still only a single genealogy. . .

Example: mitochondrial Eve

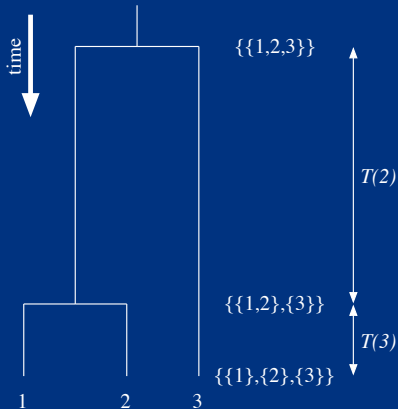


- We must consider the likelihood of the data under alternative models

The coalescent



Topology & branch lengths



Distribution of the topology

- Because of neutrality, individuals are equally likely to reproduce;
- therefore, all lineages must be equally likely to coalesce

For example,

$$\{\{1\}, \{2\}, \{3\}\} \text{ goes to } \begin{cases} \{\{1, 2\}, \{3\}\} \\ \{\{1, 3\}, \{2\}\} \\ \{\{2, 3\}, \{1\}\} \end{cases}$$

with equal probability of $1/3$

Distribution of the branch lengths

A pair of lineages coalesce one generation back in time with probability $1/N$ and stay distinct with probability $1 - 1/N$. The probability that they remain distinct for more than τ generations is

$$(1 - 1/N)^\tau.$$

The coalescence time is geometrically distributed with mean N .

This suggests a scaling approximation. . .

Scale time so that one unit of scaled time corresponds to N generations. Then the probability that two lineages stay distinct for more than t units of time is

$$\left(1 - \frac{1}{N}\right)^{[Nt]} \rightarrow e^{-t},$$

as $N \rightarrow \infty$. The coalescence time is exponentially distributed with mean 1 in the limit.

Now consider k lineages. The probability that none of them coalesce in the previous generation is

$$\prod_{i=0}^{k-1} \frac{N-i}{N} = \prod_{i=1}^{k-1} \left(1 - \frac{i}{N}\right) = 1 - \frac{\binom{k}{2}}{N} + O\left(\frac{1}{N^2}\right),$$

and the probability that more than two do so is $O(1/N^2)$.

Let $T(k)$ be the (scaled) time till the first coalescence event when there are k lineages. In the limit:

- $T(k)$ is exponentially distributed with mean

$$\frac{2}{k(k-1)}$$

- the probability that more than two lineages coalesce in a single generation can be neglected, so $T(k)$ is the time from k to $k-1$

Summary

The coalescent is continuous-time Markov process, which models the genealogy of a sample of n individuals (genes) as a random bifurcating tree, where the $n - 1$ coalescence times $T(n), T(n - 1), \dots, T(2)$ are mutually independent, exponentially distributed random variables. Each pair of lineages coalesces independently at rate 1, so the total rate when there are k lineages is “ k choose 2.”

What do coalescence trees look like?

Four simulated genealogies for $n = 6$:



Note:

- extreme variability, in all respects
- often dominated by deep branches

Deep branches often dominate

The expected time to the MRCA (the height of the tree) is

$$\mathbb{E}\left[\sum_{k=2}^n T(k)\right] = \sum_{k=2}^n \mathbb{E}[T(k)] = \sum_{k=2}^n \frac{2}{k(k-1)} = 2\left(1 - \frac{1}{n}\right),$$

while

$$\mathbb{E}[T(2)] = 1.$$

The expected time during which there are only two branches is greater than half the total expected tree height!

What do larger coalescence trees look like?

Four simulated genealogies for $n = 32$:



Increasing the sample size only adds twigs to the tree. An important consequence of this is that increasing the sample size is often surprisingly ineffective...

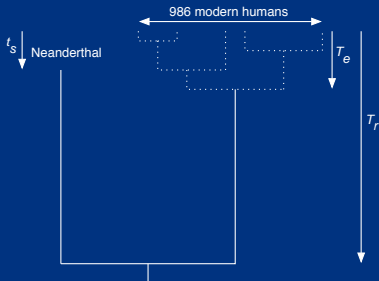
Example: How big a sample is needed to include the MRCA of everyone?

What is the probability that the MRCA of a sample of size n is the same as the MRCA of the entire population?

Answer:

$$\frac{n-1}{n+1}$$

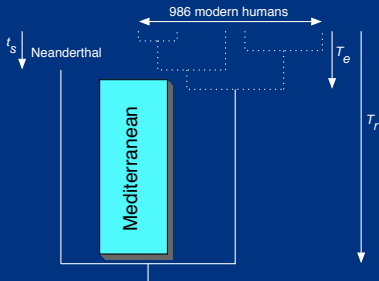
Example: ancient Neanderthal mtDNA



- Modern humans monophyletic
- $T_r > 4T_e$

Does this prove that Neanderthals and modern humans did not interbreed?

Example: ancient Neanderthal mtDNA



Assuming that they did interbreed, what is the probability of getting a tree like the one observed just by chance?

The probability is high: to rule out admixture, we need a large sample; to rule out ancient admixture, we need a large ancient sample. . .

The mutation rate

Let the per-generation probability that an allele mutates to another allele be u . The probability of no mutations τ generations back in time is then $(1 - u)^\tau$. Typically, u is small, and it makes sense to write

$$u = \frac{\theta}{2N},$$

where θ is the *mutation rate*. Using the same scaling as before, the probability of no mutations in t units of scaled time is

$$(1 - u)^{[Nt]} = \left(1 - \frac{\theta}{2N}\right)^{[Nt]} \rightarrow e^{-\theta t/2},$$

as $N \rightarrow \infty$. The time till the first mutation is exponentially distributed with mean $2/\theta$.

The probability of “identity by descent”

Two alleles are identical by descent if and only if they have descended from the same ancestral allele without mutation. In other words, if and only if they coalesce before either incurs a mutation. Coalescence and mutation occurs according to independent Poisson processes, so:

$$\mathbb{P}(\text{ibd}) = \frac{1}{1 + \theta/2 + \theta/2} = \frac{1}{1 + \theta}$$

Superimposing mutations

Going forward in time, mutations occur at rate $\theta/2$ along each branch of the tree. The number of mutations on a branch of length t is Poisson-distributed with mean $\theta t/2$.

The total length of the tree is

$$L_n = \sum_{j=2}^n jT(j).$$

Let S_n denote the total number of mutations on the tree: it has a mixed Poisson distribution, $S_n \sim \text{Po}(\theta L_n/2)$.

The expected number of mutations is:

$$\begin{aligned}\mathbb{E}(S_n) &= \mathbb{E}(\mathbb{E}(S_n|L_n)) \\ &= \mathbb{E}(\theta L_n/2) \\ &= \frac{\theta}{2} \mathbb{E} \sum_{j=2}^n jT(j) \\ &= \frac{\theta}{2} \sum_{j=2}^n j \mathbb{E}T(j) \\ &= \frac{\theta}{2} \sum_{j=2}^n j \frac{2}{j(j-1)} \\ &= \theta \sum_{j=1}^{n-1} 1/j\end{aligned}$$

Note that

$$\mathbb{E}(S_n) = \theta \sum_{j=1}^{n-1} 1/j \rightarrow \theta(\gamma + \log n),$$

as $n \rightarrow \infty$. Increasing the sample size adds few mutations.

The variance of the number of mutations can be shown to be:

$$\text{Var}(S_n) = \theta \sum_{j=1}^{n-1} 1/j + \theta^2 \sum_{j=1}^{n-1} 1/j^2.$$

Mutation models

Any form of mutation fits into our scheme. For example:

- There is a finite number of possible alleles. Whenever a mutation occurs, the type is determined by the matrix $U = (u_{ij})$, where u_{ij} is the probability that an allele of type i mutates to type j . Note that this allows the mutation rate to depend on the current allelic state.
- Infinite-alleles — each mutant allele is unique.
- Infinite-sites — each mutation hits a new site in a DNA sequence.

Estimating θ under the infinite-sites model

Tajima's estimator is based on the average number of pairwise differences:

$$\hat{\theta}_T = \frac{2}{n(n-1)} \sum_{i < j} S_{ij},$$

where S_{ij} is the number of differences between sequence i and j .

Watterson's estimator is based on the number of segregating (polymorphic) sites, S_n :

$$\hat{\theta}_W = S_n / \sum_{j=1}^{n-1} 1/j.$$

Both are unbiased, however,

$$\text{Var}(\hat{\theta}_T) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2$$

and

$$\text{Var}(\hat{\theta}_W) = \frac{\theta \sum_{j=1}^{n-1} 1/j + \theta^2 \sum_{j=1}^{n-1} 1/j^2}{\left(\sum_{j=1}^{n-1} 1/j\right)^2}.$$

Note that

$$\text{Var}(\hat{\theta}_T) \rightarrow \frac{1}{3}\theta + \frac{2}{9}\theta^2,$$

as $n \rightarrow \infty$. Not consistent! Watterson's estimator is, but

$$\text{Var}(\hat{\theta}_W) \propto \frac{1}{\log n}$$

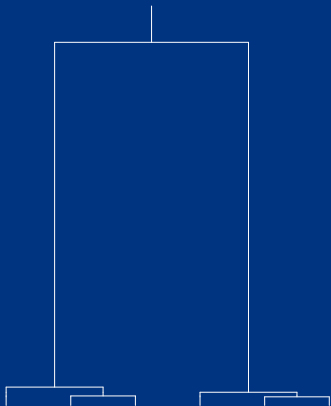
for large $n \dots$

How do we detect deviations from the standard model (e.g., growth)?

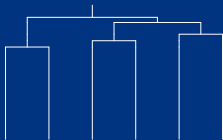
A common method relies on a test statistic known as Tajima's D_T :

$$D_T = \frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{\text{Var}(\hat{\theta}_T - \hat{\theta}_W)}}$$

Should be $N(0, 1)$.

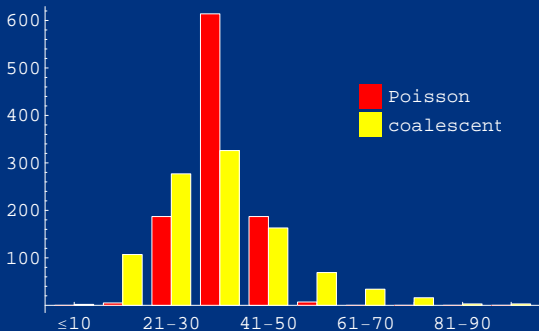


excess of common alleles: $D_T > 0$



excess of rare alleles: $D_T < 0$

The coalescent variance



The number of SNPs in 20 copies of 10 kb, 1000 runs

Review

- Population genetics data reflect a evolutionary process with a complicated dependence structure: this leads to an “evolutionary” variance component that cannot be taken into account using textbook statistics
- Under selective neutrality, we can model such data very efficient using the coalescent, which essentially:
 1. generates the genealogy backward in time
 2. superimposes mutations forward in time

Robustness of the coalescent

I introduced the coalescent using the haploid Wright-Fisher model. However, the coalescent arises as a limiting process for a wide range of neutral models, *provided time is scaled appropriately*.

We will investigate models where the coalescent is obtained in the limit, and models where it isn't.

Example

- In the W-F model, the variance of the number of offspring produced by an individual is $1 - 1/N \rightarrow 1$
- Consider an generalized version where the limiting variance is σ^2 , $0 < \sigma^2 < \infty$. This model converges to the coalescent provided time is measured in units of N/σ^2 generations.

Increased variance in reproductive success increases the rate of coalescence. A simple linear change in the time scale of the coalescent takes this into account.

A remarkable range of phenomena can be shown to have the same effect — which is good news or bad news, depending on your point of view...

Population genetics in and of itself only allows limited inference

In humans, we have $\hat{\theta} \approx 10^{-3}$. (Two randomly chosen human sequences differ less than once every 1,000 bp.) How shall we interpret this? Even under the very simplest model, we have

$$\theta = 4N/\sigma^2 u$$

(the additional “2” will be explained shortly). People want to know things about the actual population size: this cannot be estimated without external information.

Effective population size

Because the model in the example converges as a W-F model with size N/σ^2 , it is often said to have an “effective population size”,

$$N_e = N/\sigma^2.$$

This terminology has caused much confusion:

- All N_e 's are not created equal
- $N_e \neq N$!

Many phenomena can be modeled as a linear change to the time scale of the coalescent — but most cannot!

Variable population size

The rate of coalescence depends on the population size. If the population size varies, so will the rate of coalescence. Alternatively, we may let the time scale change. . .

Let $N(\tau)$ be the population size τ generations ago.

In a constant population:

- τ generations $\Rightarrow \tau/N$ units of coalescence time
- t units of coalescence time $\Rightarrow [Nt]$ generations

In a variable population:

- τ generations ago \Rightarrow

$$g(\tau) = \sum_{i=1}^{\tau} \frac{1}{N(i)}$$

units of coalescence time

- t units of coalescence time $\rightarrow [g^{-1}(t)]$ generations

Exponential growth

Consider a population that has grown rapidly so that, backwards in time, it shrinks according to

$$N(\tau) = N(0)e^{-\beta\tau}.$$

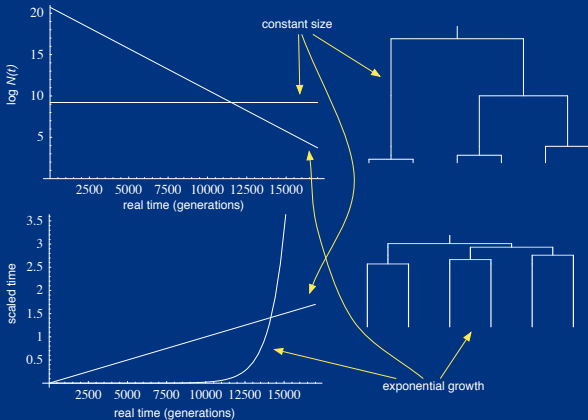
Then

$$g(\tau) \approx \int_0^\tau \frac{1}{N(s)} ds = \frac{e^{\beta\tau} - 1}{N(0)\beta},$$

and

$$g^{-1}(t) \approx \frac{\log(1 + N(0)\beta t)}{\beta}.$$

Exponential growth



Rule of thumb

In order to leave a trace in the pattern of polymorphism, phenomena have to have some duration on the coalescent time scale.

- Recent growth in a large population may be too recent
- Bottlenecks do not matter unless they last for a number of generations equal to the size of the reduced population

Population structure

- Most populations have geographical structure
- Many biological phenomena can be thought of as analogous to population structure
- Different kinds of structure may be important on different time scales

The structured Wright-Fisher model

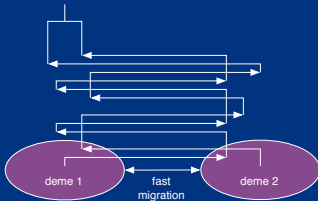
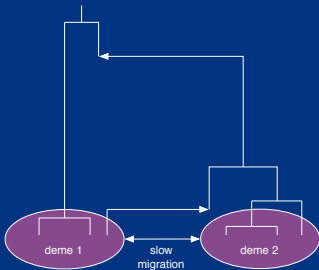
- M patches (or demes) of fixed sizes N_i , $i \in \{1, \dots, M\}$, so that $\sum_i N_i = N$
- Infinitely many propagules are produced each generation
- Propagules migrate independently of each other so that with probability m_{ij} a propagule from i ends up in j
- After migration, the finite number of adults in each patch are randomly chosen from the cloud of propagules

This model can equally well be studied backward in time. Each lineage “picks its parent” independently from the previous generation. The probability that a lineage current in i picks a parent in j is

$$b_{ij} = \frac{N_j m_{ji}}{\sum_k N_k m_{ki}}.$$

The general idea is to let $N \rightarrow \infty$, as before. The limiting behavior of the model depends on how the remaining parameters scale. . .

The two-deme model



Slow migration

Assume that M , $c_i = N_i/N$, and $B_{ij} = 2Nb_{ij}$, $i \neq j$, all remain constant as $N \rightarrow \infty$. Then, with time measured in units of N generations, the process converges to the so-called “structured coalescent”:

- each pair of lineages in patch i coalesce independently at rate $1/c_i$
- each lineage in i “migrates” independently to j at rate $B_{ij}/2$
- no other events are seen

Events occur according to independent Poisson processes. Let k_i be the number of lineages currently in patch i . The

waiting time till the first event is the sum of all rates, *i.e.*,

$$h(k_1, \dots, k_M) = \sum_i \left(\frac{\binom{k_i}{2}}{c_i} + \sum_{j \neq i} k_i \frac{B_{ij}}{2} \right).$$

When an event occurs, it is coalescence in patch i with probability

$$\frac{\binom{k_i}{2} / c_i}{h(k_1, \dots, k_M)}$$

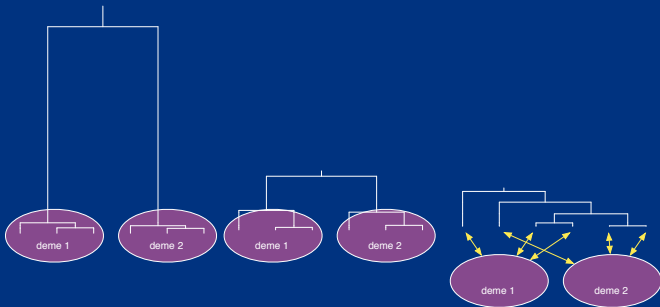
and a migration from i to j with probability

$$\frac{k_i B_{ij} / 2}{h(k_1, \dots, k_M)}$$

In the former case, a random pair in i coalesces and k_i decreases by one. In the latter case, a random lineage in i moves to j , k_i is decreased by one, and k_j increased by one.

Consequences of slow migration

- Coalescence within patches (isolation by distance)
- Increased mean and variance of coalescence times



Fast migration

The more migration, the less effect of subdivision. This is obvious, but leads to an interesting result in the limit. Make the same assumptions as before, except that the backward migration rates, b_{ij} , are no longer $O(1/N)$.

- lineages migrate back and forth infinitely fast on the coalescent time scale
- the rate at which pairs of lineages coalesce in a patch is determined by how often the patch is visited

Let π_i be the stationary probability that a lineage is in patch i . The standard coalescent is retrieved if time is measured in unit is N/α , where

$$\alpha = \sum_i \pi_i^2 / c_i.$$

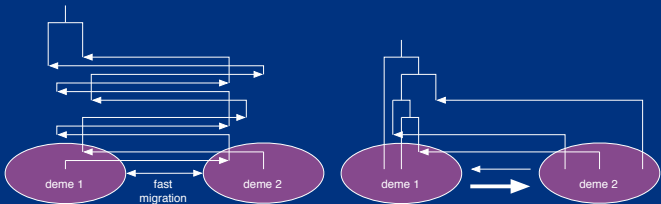
We have $\alpha \geq 1$, with equality if and only if

$$\sum_{j \neq i} N_i b_{ij} = \sum_{j \neq i} N_j b_{ji} \quad \forall i.$$

Coalescence occur faster unless emigration equals immigration everywhere!

What's the intuition behind this?

Source-sink environments



Going backward in time, lineages tend to spend more time in rich environments, and less in poor. This decreases the “effective population size” — parts of the population do not contribute to future generations.

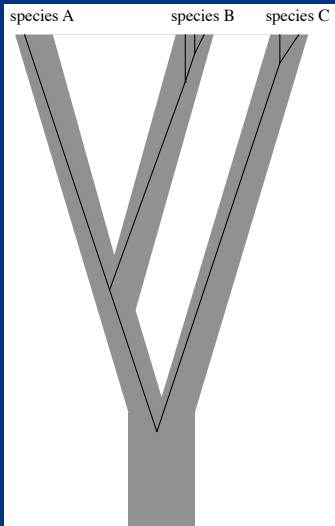
Is the island model a good model?

Two major problems:

- Most organisms show some form of isolation by distance — such models are hard, because they require density regulation
- Assumes constant demography on the coalescent time scale

Nonetheless, the model has many uses. . .

Gene trees and species trees



Sex

segregation — diploidy can be taken into account by changing the scaling from N to $2N$

recombination — matters...

Hermaphrodites and self-fertilization

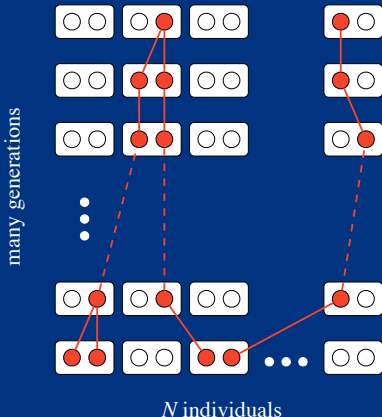
Looks like a haploid
population of size $2N$,
divided into N patches
of size 2:

$$M = N$$

$$N_i = 2$$

$$c_i = 2/N$$

We scale time in units
of $2N \dots$



On the coalescent time scale...

Pairs of lineages in different individuals “coalesce” into the same individual at rate 2 — whenever this happens, we have:

- a real coalescence with probability $1/2$
- two distinct lineages occupying the same individual with probability $1/2$

On a much faster time scale. . .

The fate of a pair of lineages in the same individual depends on how that individual was produced:

outcrossing — the lineages end up back in distinct individuals again

selfing — the lineages end up in the same individual again, and we have:

- a real coalescence with probability $1/2$
- two distinct lineages occupying the same individual with probability $1/2$

On the coalescent time scale, the fate of the two lineages will *instantaneously* either coalesce or end up back in

different individual. If the probability of selfing is S , then the probability of the former is:

$$F = \frac{S/2}{S/2 + 1 - S} = \frac{S}{2 - S}$$

Each time two lineages “coalesce” into the same individual, the probability that this results in actual coalescence is:

$$1/2 \times 1 + 1/2 \times F = (1 + F)/2$$

Since lineages “coalesce” into the same individual at rate 2, the total rate of coalescence is $1 + F$.

Summary

The Wright-Fisher model with diploid hermaphrodites converges to the standard coalescent as long as time is scaled in units of

$$2N_e = \frac{2N}{1 + F}$$

In obligate outcrossers, $F = 0$, and the correct scaling is $2N$.

Note that:

- Special considerations apply to samples
- These results correspond to classical notions of heterozygosity

Males and females

Consider a diploid population of N_m breeding males and N_f breeding females so that $N_m + N_f = N$.

- Can be thought of as a haploid population of size $2N$, divided into two patches of size $2N_m$ and $2N_f$ respectively, each of which is further subdivided into patches of size 2.
- Each lineage came from current or opposite sex in previous generation with equal probability $1/2$.
- With sex, all individuals are equally likely to be chosen

Looks (almost) like a structured Wright-Fisher with $M = 2$, $c_m = N_m/N$, $c_f = N_f/N$, and $b_{mf} = b_{fm} = 1/2$

Pairs of lineages in different individuals (regardless of sex) coalesce in the previous generation if and only if both came from:

- the same sex
- the same diploid individual within that sex
- the same haploid genome within that individual

This occurs with probability

$$\frac{1}{4} \times \frac{1}{N_m} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{N_f} \times \frac{1}{2} = \frac{N_m + N_f}{8N_m N_f},$$

or, in the limit $N \rightarrow \infty$, at rate $\alpha = (4c_m c_f)^{-1}$.

Summary

The Wright-Fisher model with diploid males and females converges to the standard coalescent as long as time is scaled in units of

$$2N_e = \frac{2N}{\alpha} = \frac{8N_m N_f}{N_m + N_f}$$

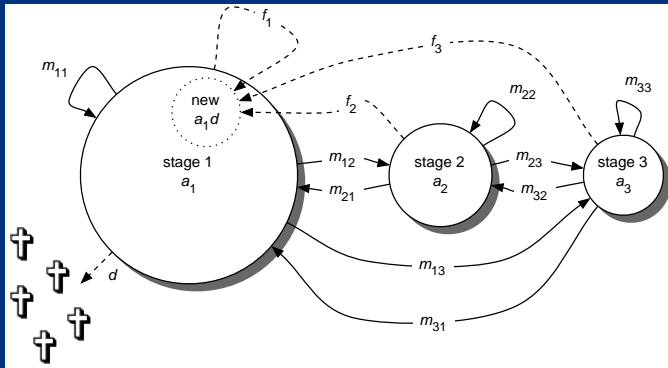
If $N_m = N_f = N/2$, the correct scaling is again $2N$.

Note that:

- Easy to extend to different variances in reproductive success for the sexes
- Easy to do sex-linked loci

Stage structure

Three life stages in *Fritillaria camtschatcensis*:



Stochastic demography

- If changes are slow, coalescent sees one state
- If changes are fast, coalescent averages over states
- If changes occur on the coalescent time scale, things get hairy...

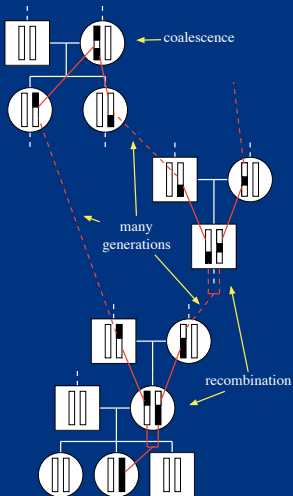
Why recombination?

Recombination makes analytical results impossible, but the need for models that include recombination is transparently obvious in the era of genomic data:

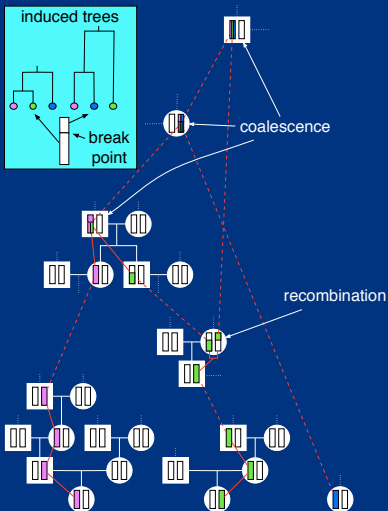
- linkage disequilibrium and GWAS
- scans for selection
- accurate estimation of demography

Recombination — backwards in time!

- Splits up lineages
- Never recombination with other ancestors



**Recombination
makes it possible
for linked sites to
have different
genealogies**



The ancestral recombination graph

Due to Griffiths & Marjoram (1996, 1997) [Hudson (1983) had the first coalescent with recombination]

Assumptions:

- let the per-generation probability of recombination in a segment be r
- define the *recombination rate*

$$\rho = \lim_{N \rightarrow \infty} 4Nr$$

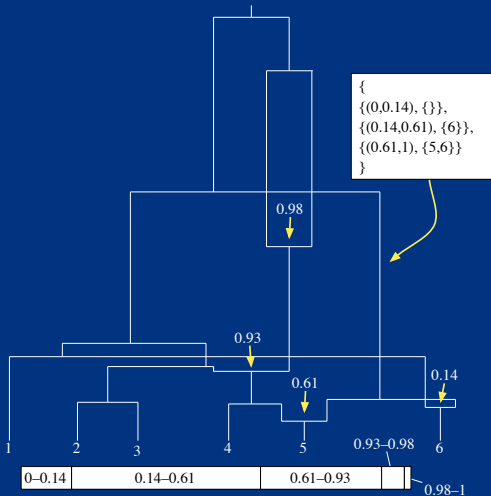
- measure time in units of $2N$ generations

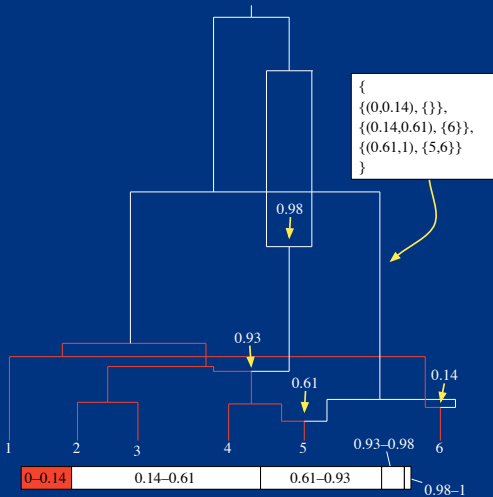
In the limit as $N \rightarrow \infty$:

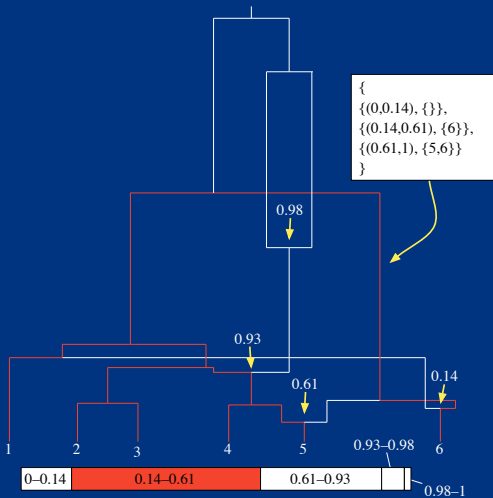
- each lineage independently undergoes recombination at rate $\rho/2$
- each pair of lineages independently coalesces at rate 1
- each recombination increases the number of lineages by one, each coalescence decreases the number of lineages by one
- when there are k lineages, the total rate of recombination is $k\rho$, and the total rate of coalescence is $\binom{k}{2}$
- the number of lineages is guaranteed to stay finite, and even hit one, occasionally

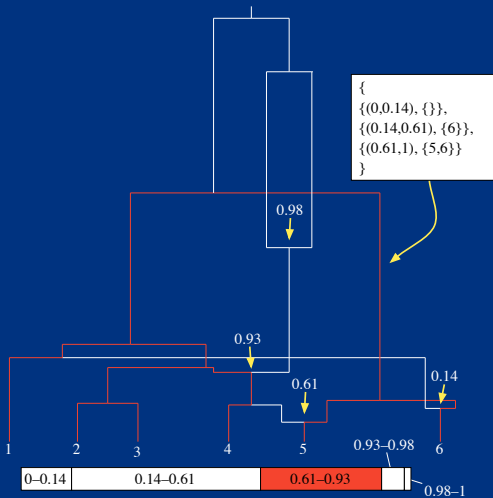
- the first time the number of lineages hits one, we have the Ultimate MRCA — which is very different from the MRCA for any point
- any model of recombination (or gene conversion) may be used: a particularly simple one (“infinite-sites”) has recombination occurring uniformly in a $(0,1)$ interval

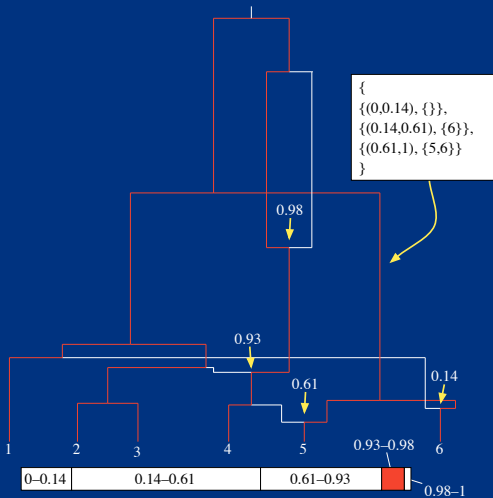
Here is an example for $n = 6$. .

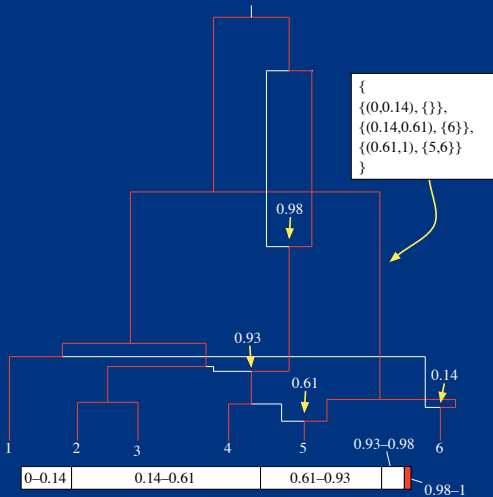












A walk through tree space



As we walk along the chromosome, the trees change — but only gradually. Linked trees are correlated, and the degree of correlation decreases with genetic distance. One manifestation of this is linkage disequilibrium...

Simulating data

1. Simulate the graph backward in time until the UMRCA is found (or, to save time, till all point-wise MRCAs have been found)
2. Add mutations forward in time

How common is recombination?

- If one 1 cM \sim 1 Mb, then $r \approx 10^{-8}$ per site
- The per-generation mutation probability, u , is estimated to be the same or lower
- Thus we should have $\rho \geq \theta \dots$

It follows that there will typically be at least as many recombination events in the history of a sample as there are segregating sites!

Detecting recombination

A number of tests for “detecting recombination” exist. These are designed to detect rare recombination events ($\theta \gg \rho$). Paradoxically, they become less powerful the more common recombination is.

A very simple and useful test is the “four gamete test”: if, for two di-allelic loci, all four gametes exist, then there must have been recombination between the loci (or a repeated mutation at one of the loci)

It can be shown that a very large fraction of recombination events cannot be detected even with infinite mutation rate. When θ and ρ are of the same magnitude, *there is no chance of estimating the genealogical graph.*

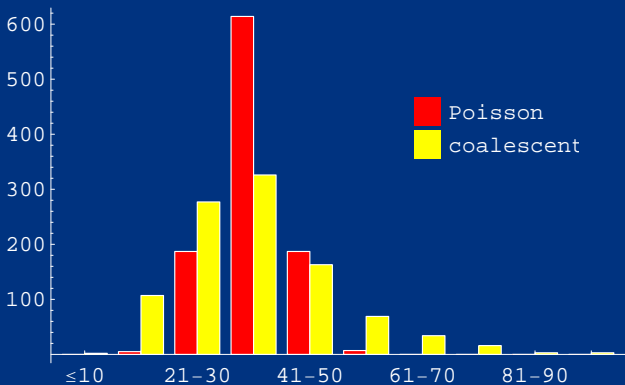
Note that phylogenetic methods will still produce a tree. . .

Overcoming the evolutionary variance

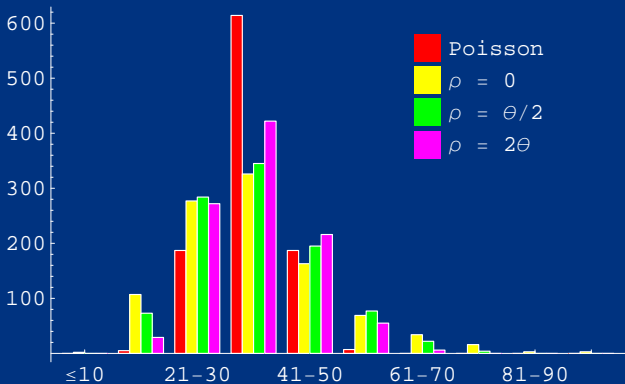
Recombination makes different parts of the genome increasingly independent: this reduces the evolutionary variance that is due the coalescent. Finally a sample size greater than 1!

For example, the variance of the estimates of θ are decreasing functions of ρ . . .

The number of SNPs in 20 copies of 10 kb, 1000 runs:



Same thing, with recombination:



Linkage disequilibrium (LD)

LD refers to the non-random association of alleles in haplotypes. Often thought of in terms of pairwise measures.

Consider two loci, \mathcal{A} and \mathcal{B} , with alleles A/a and B/b , respectively. Let P_A stand for the frequency of allele A , and so forth. Similarly, let P_{AB} stand for the frequency of the AB haplotype. Examples of LD measures include:

- $D = P_{AB} - P_A P_B$, the classical LD coefficient
- $|D'|$, the absolute value of $D = P_{AB} - P_A P_B$, normalized to take values between 0 and 1 regardless of the allele frequencies

- r^2 , the correlation in allelic state between the two loci as they occur in haplotypes
- $d^2 = (P_{Ba}/P_a - P_{BA}/P_A)^2$, which measures the association between the alleles at (marker) locus \mathcal{B} and the alleles at the (disease) locus \mathcal{A}

All of these are related to the standard χ^2 statistic for a 2×2 contingency table.

Why “linkage disequilibrium”?

The phrase “linkage disequilibrium” is one of the most misleading in population genetics. First of all:

- unlinked genes can be in LD
- linked genes are not necessarily in LD

Thus linkage disequilibrium is only indirectly associated with linkage. Most textbooks will note this. But what does it have to do with “equilibrium”?

Where does D come from?

Consider the dynamics of the four gamete frequencies P_{AB} , P_{Ab} , P_{aB} , and P_{ab} in a very large population. Let the recombination fraction between the two loci be r . Note first that:

$$\begin{aligned} D &= P_{AB} - P_A P_B \\ &= P_{AB} - (P_{AB} + P_{Ab})(P_{AB} + P_{aB}) \\ &= P_{AB} - P_{AB}^2 - P_{AB}P_{Ab} - P_{AB}P_{aB} - P_{Ab}P_{aB} \\ &= P_{AB}(1 - P_{AB} - P_{Ab} - P_{aB}) - P_{Ab}P_{aB} \\ &= P_{AB}P_{ab} - P_{Ab}P_{aB}. \end{aligned}$$

Assuming random mating, we have:

$$\begin{aligned}P'_{AB} &= (1 - r)P_{AB} + rP_AP_B \\&= (1 - r)P_{AB} + r(P_{AB} + P_{Ab})(P_{AB} + P_{aB}) \\&= (1 - r)P_{AB} \\&\quad + r(P_{AB}^2 + P_{AB}P_{Ab} + P_{AB}P_{aB} + P_{Ab}P_{aB}) \\&= P_{AB} \\&\quad - r(P_{AB} - P_{AB}^2 - P_{AB}P_{Ab} - P_{AB}P_{aB} - P_{Ab}P_{aB}) \\&= P_{AB} - r(P_{AB}[1 - P_{AB} - P_{Ab} - P_{aB}] - P_{Ab}P_{aB}) \\&= P_{AB} - r(P_{AB}P_{ab} - P_{Ab}P_{aB}) \\&= P_{AB} - rD\end{aligned}$$

Repeating this exercise for the other frequencies gives:

$$P'_{AB} = P_{AB} - rD$$

$$P'_{Ab} = P_{Ab} + rD$$

$$P'_{aB} = P_{aB} + rD$$

$$P'_{ab} = P_{ab} - rD$$

Thus the haplotype frequencies do not change if $D = 0$.

The single-locus allele frequencies never change:

$$\begin{aligned} P'_A &= P'_{AB} + P'_{Ab} \\ &= P_{AB} - rD + P_{Ab} + rD \\ &= P_{AB} + P_{Ab} \\ &= P_A \end{aligned}$$

How does D evolve? Well,

$$\begin{aligned}D' &= P'_{AB} - P'_A P'_B \\&= P_{AB} - rD - P_A P_B \\&= (1 - r)D\end{aligned}$$

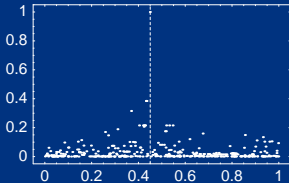
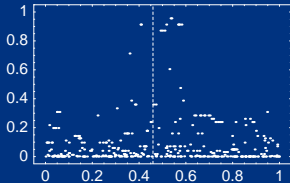
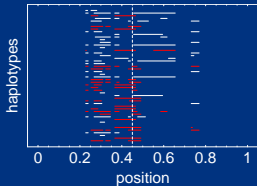
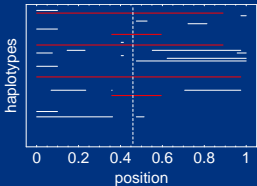
so D approaches zero at rate $1 - r$. After t generations,

$$D_t = (1 - r)^t D_0$$

so that, at “equilibrium”, $D = 0$ (whence “linkage disequilibrium”)

However, backward in time...

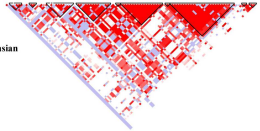
Decay of haplotype sharing



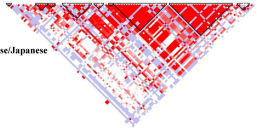
HapMap Project

- 100 kb of 8p23.1
- Less LD in Africa

Caucasian



Chinese/Japanese



Yoruba



Selection and the coalescent

Selection is difficult to model backward in time because it, by definition, couples descent and state. Nonetheless some progress has been made using two very different approaches:

- the “conditional structured coalescent” (Kaplan *et al*, 1988; 1989)
- the ancestral selection graph (Krone & Neuhauser, 1997)

The ancestral selection graph

- decouples descent from state by introducing branches (backward in time) that represent “virtual” ancestors
- mutations are added forward in time, and the graph pruned to a tree according to state

The Moran model

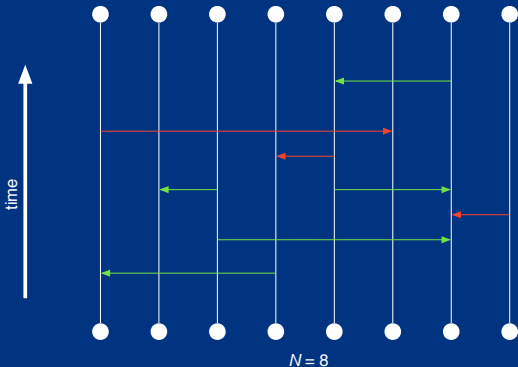
- N haploid individuals, continuous time
- individuals give birth, one at a time, at rates determined by their types
- the newly born individual replaces a randomly chosen one from the population
- a newly born individual mutates with probability u

Assume two types, A_1 , which reproduce at rate λ_1 , and A_2 , which reproduce at rate

$$\lambda_2 = \lambda_1(1 + s)$$

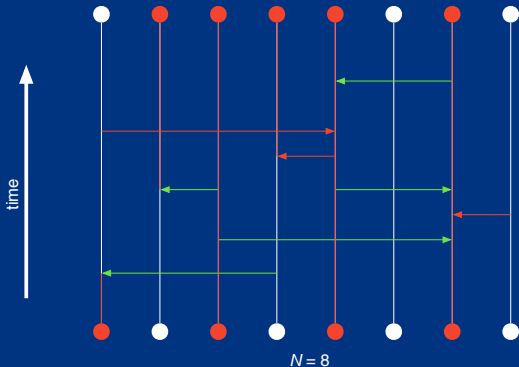
The biased voter model

Arrows are put down at rate λ_2 : $\mathbb{P}(\text{green}) = \lambda_1/\lambda_2$.



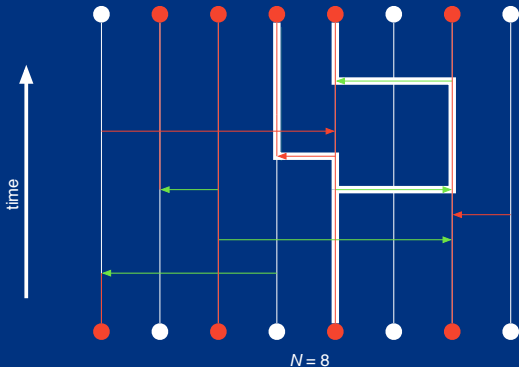
The biased voter model

Only reds (A_2) can give birth through red arrows.



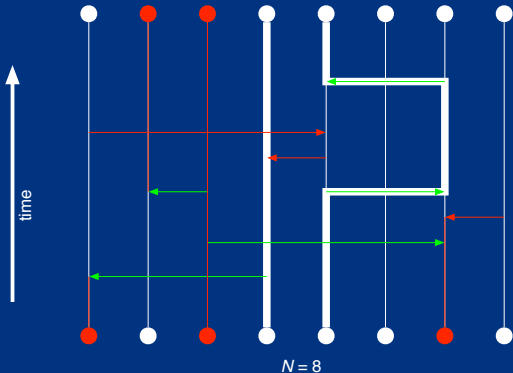
The dual process

Descent depends on state...



The dual process

Descent depends on state...



The ancestral selection graph

- Each lineage independently branches at a rate that depends on the maximum selective difference
- Each pair of lineages coalesces at rate 1.
- As was the case for the recombination graph, we are certain to hit an UMRCA eventually
- Assign a type to the the UMRCA and drop mutations in the usual manner, *pruning* branches to get a tree

The conditional structured coalescent

If we knew the past frequencies of the various alleles at the locus under selection, we can think of the population as divided into allelic classes, and consider genealogies within each class. Analogous to population subdivision: mutation plays the role of migration.

- may be reasonable if selection is strong
- allows us to study the effects of selection on the pattern of variation in the chromosomal region surrounding the selected locus
- useful in the context of “detecting selection” — what should we expect to see?

Balancing selection

Two alleles, A_1 and A_2 are maintained at constant frequencies p_1 and $p_2 = 1 - p_1$ by strong selection. Mutation back and forth occurs at scaled rate $\nu = 4Nv$.

Consider a segment of scaled length $\rho = 4Nr$ that contains the selected locus. Depending on the allelic state at this locus, the segment belongs to either the A_1 or the A_2 allelic class. Say that it belongs to A_1 . Trace it back a single generation. . .

- the segment “mutates” to the A_2 allelic class with probability

$$\frac{vp_2}{vp_2 + (1-v)p_1} = \frac{\nu}{4N} + O\left(\frac{1}{N^2}\right)$$

- the segment undergoes recombination with probability

$$r = \rho/(4N)$$

- otherwise, it remains in its class

If there was a recombination event, we have to know the genotype of the individual in which the recombination event took place. . .

Since the lineage is A_1 , the individual must have been either A_1A_1 or A_1A_2 . The probability that it was the latter is

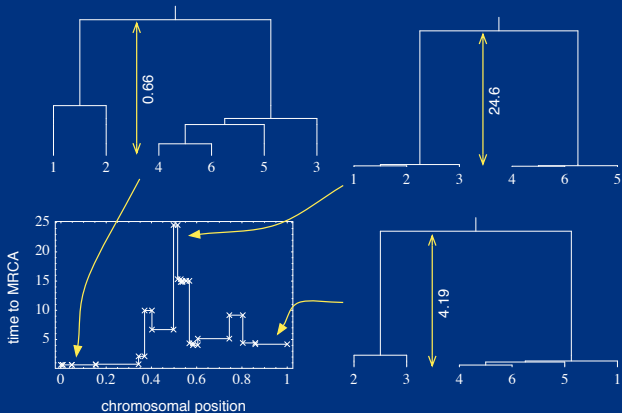
$$\frac{w_{12}x_{12}/2}{w_{12}x_{12}/2 + w_{11}x_{11}},$$

where x_{ij} and w_{ij} are the genotype frequencies and relative fitness values. Under certain simplifying assumptions, the above probability simplifies to $p_2 \dots$

If recombination occurs in a homozygote, both branches remain in the A_1 allelic class. If it occurs in a heterozygote, one branch (the one not containing the selected locus) jumps to A_2 .

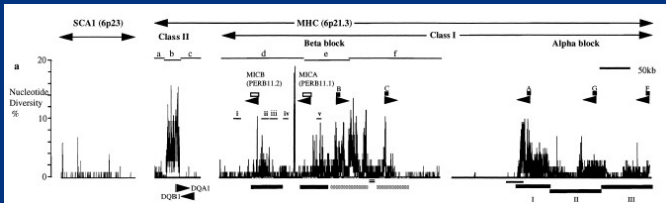
Coalescences occurs as usual for structured models, within allelic classes.

A peak of polymorphism is expected...



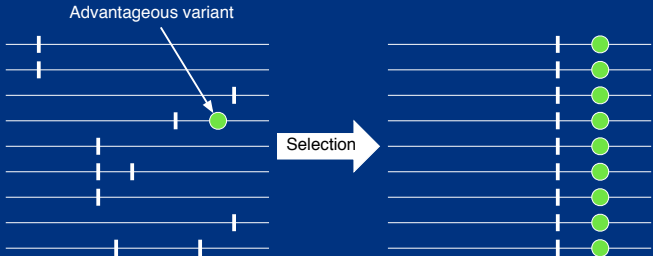
... and occasionally seen

Nucleotide diversity in the MHC region:



Positive directional selection

“Selective sweeps” wipe out polymorphism locally:



Strong directional selection can be modelled precisely as balancing selection, except that the sizes of the allelic class now change over time. Intuition:

- balancing selection looks like local population subdivision
- directional selection looks like local population growth

Selective sweeps in progress will also increase LD. The width of the affected region depends on the relationship between selection and recombination.

Example: maize and teosinte



teosinte



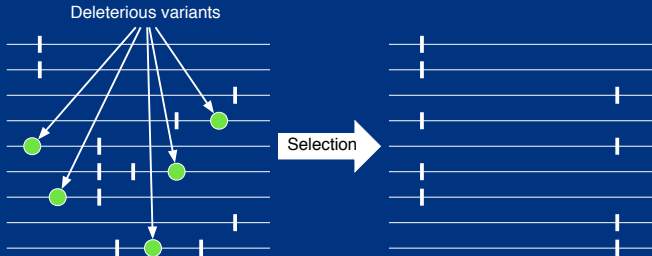
maize



maize with *tb1* mutation

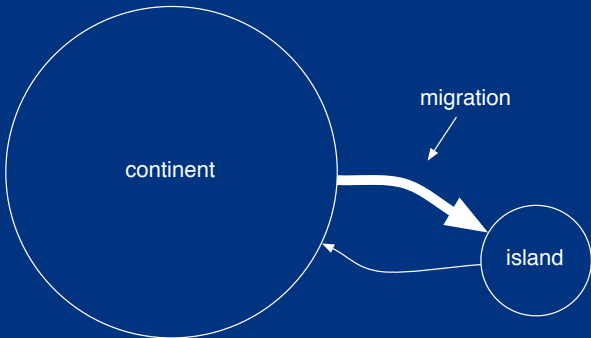
Purifying selection

Purifying selection decrease the level of variation:

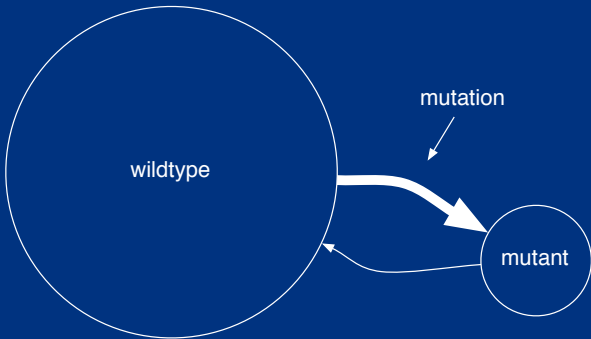


This kind of selection *may* behave like a decreased $N_e \dots$

Why purifying selection may reduce N_e



Why purifying selection may reduce N_e



Summary

- The conditional structured coalescent is very useful for providing intuition about what we might expect to see in the vicinity of a small number of strongly selected sites. There are lots of questions about the mathematical details, but they are of little interest to biologists. . .
- The ASG is not very useful at all, because no one cares about weak selection at a single locus. We do care about weak selection at individual loci on a genome-wide scale, but selection on a genome-wide scale is not weak. . .