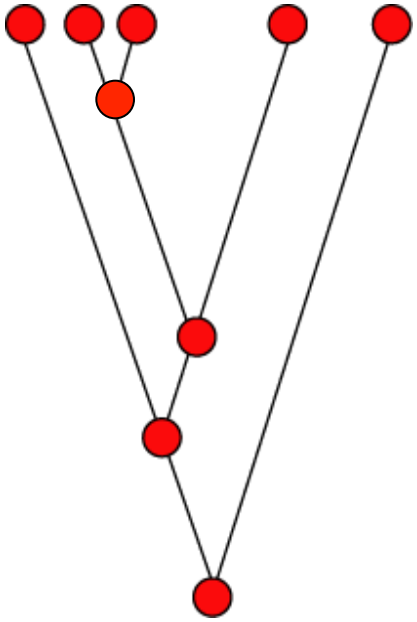
Introduction to Coalescent Theory

Mick Elliot & Arne Mooers

What is the coalescent?

The coalescent is a model of the distribution of gene divergence in a genealogy



It is widely used to estimate population genetic parameters such as population size, migration rates and recombination rates in natural populations

It was originally formulated as the “n-coalescent” by Kingman (1982). Others refer to it as the “Kingman coalescent” or just the “coalescent”

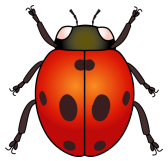
The coalescent model is derived from a simple population genetic model, and the easiest way to understand what it is and how it works is to follow the basic derivation

Kingman 1982

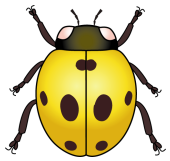
The Wright-Fisher Population Model

Consider a biallelic gene in a diploid organism

As a visual aid, the wing-cases of the ladybirds below are coloured to represent the alleles carried by each individual



Two “red” alleles



Two “yellow” alleles

A “red” and a “yellow” allele

The Wright-Fisher Population Model

Start with a population of size N

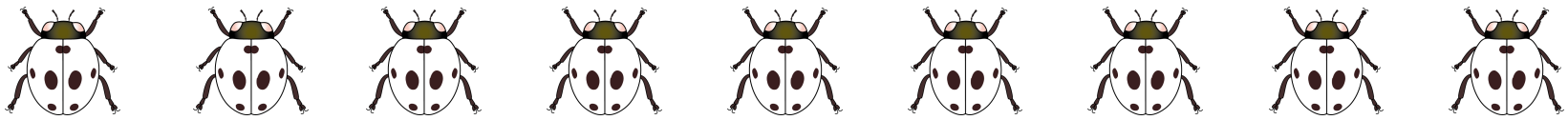
Generation 1

The Wright-Fisher Population Model

Start with a population of size N

As soon as an individual dies it is replaced by a new offspring, so the population size remains constant

Generation 2



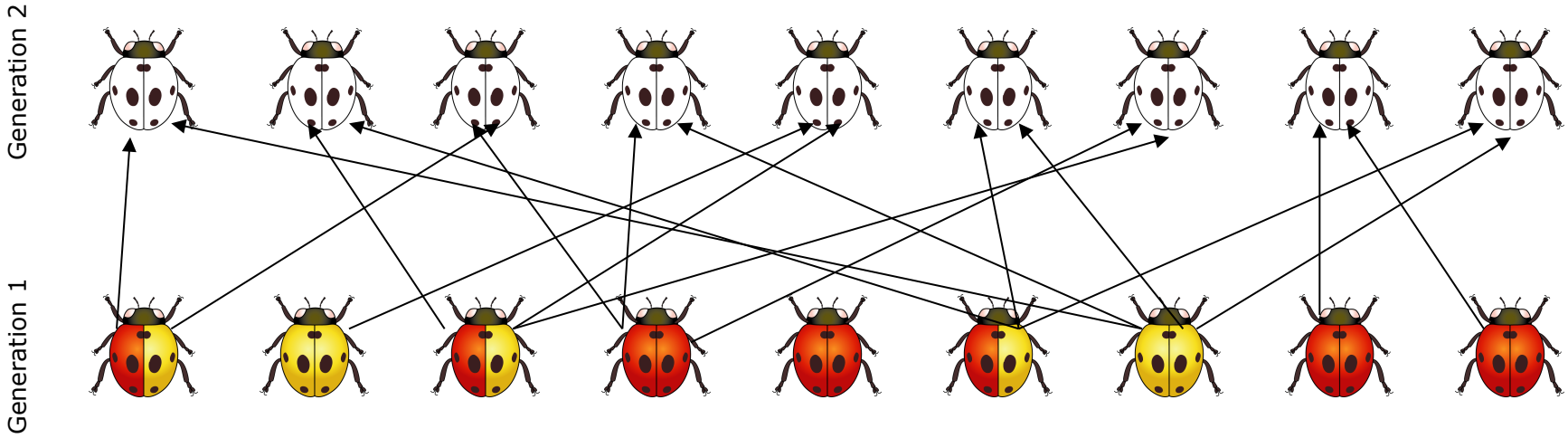
Generation 1

The Wright-Fisher Population Model

Start with a population of size N

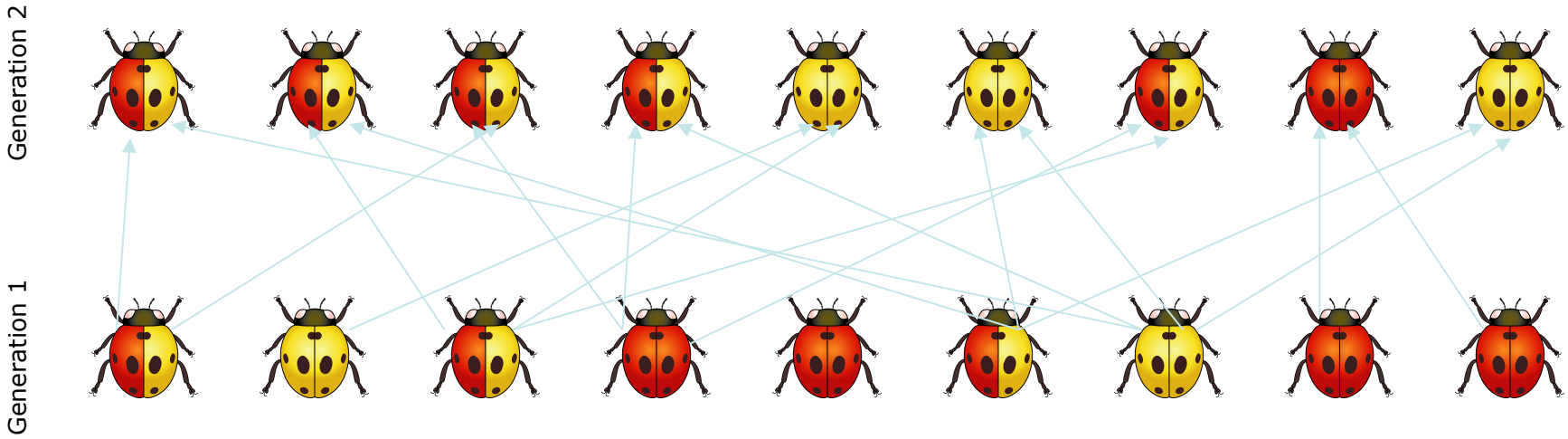
As soon as an individual dies it is replaced by a new offspring, so the population size remains constant

Each individual releases many gametes, and new individuals are drawn **randomly** from the gamete pool



The Wright-Fisher Population Model

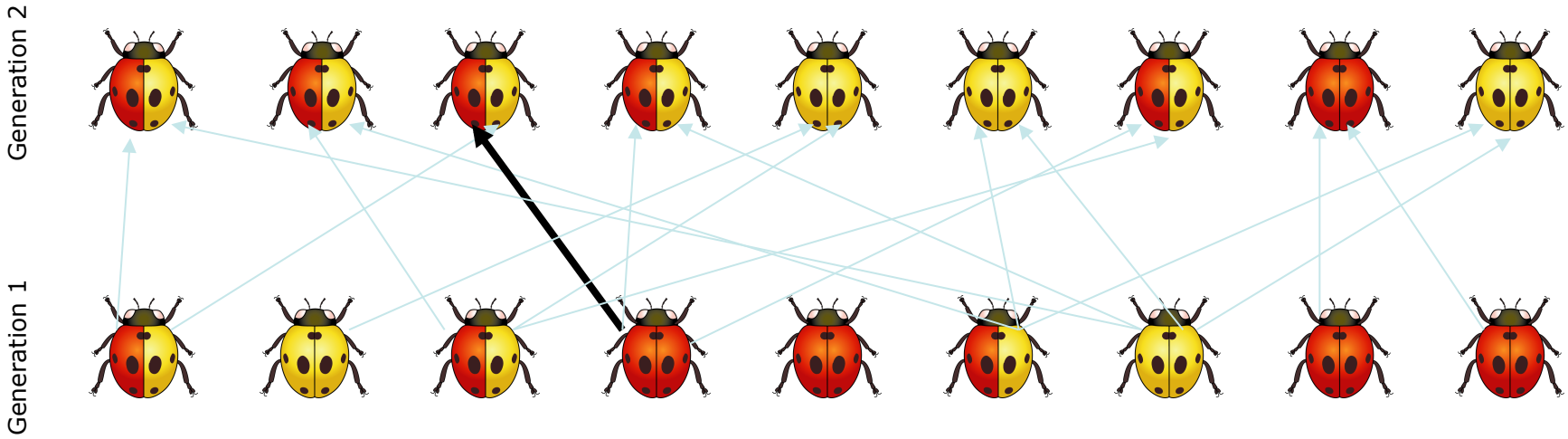
Sewall Wright made an important observation



The Wright-Fisher Population Model

Wright and fisher made an important observation

Probability that an allele in G2 has a parent in G1 = 1

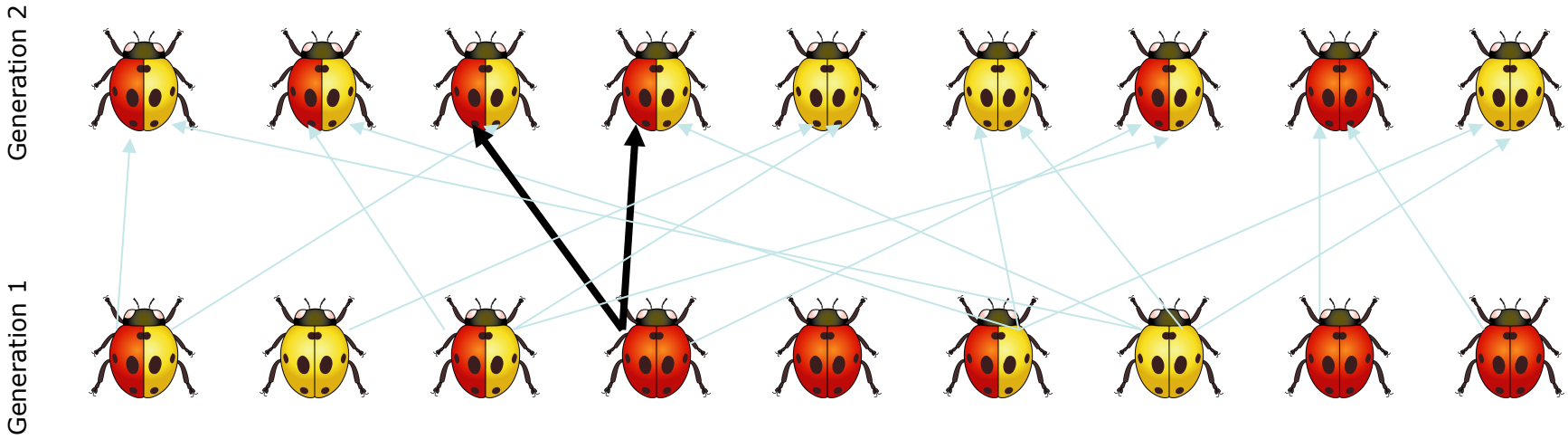


The Wright-Fisher Population Model

Wright and fisher made an important observation

Probability that an allele in G2 has a parent in G1 = 1

Probability that a random allele in G2 has *the same* parent in G1 = 1/2N



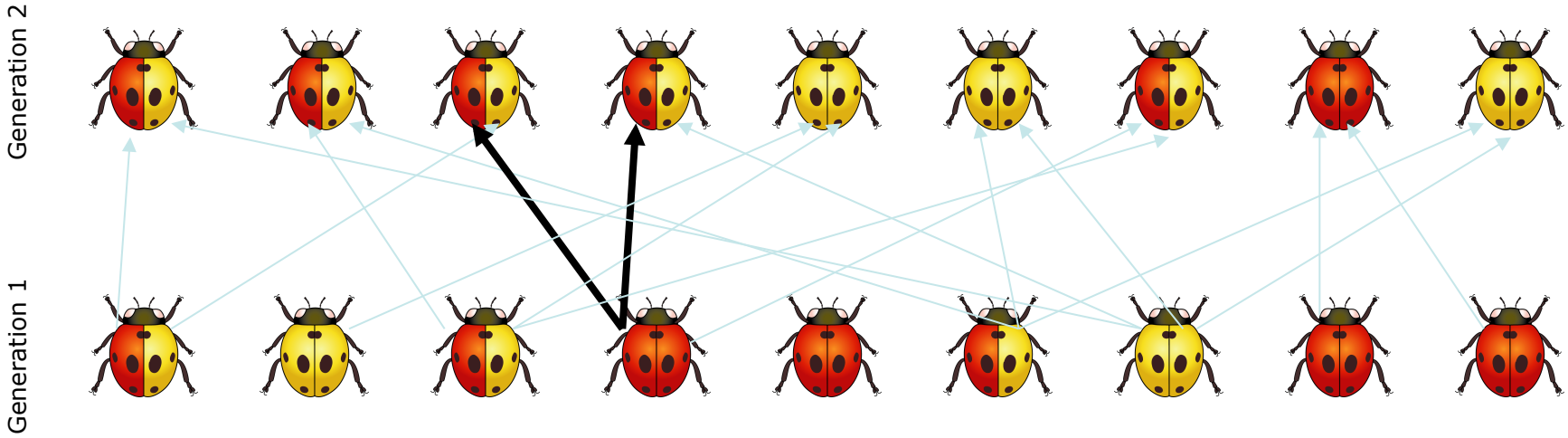
The Wright-Fisher Population Model

Wright and fisher made an important observation

Probability that an allele in G2 has a parent in G1 = 1

Probability that a random allele in G2 has *the same* parent in G1 = 1/2N

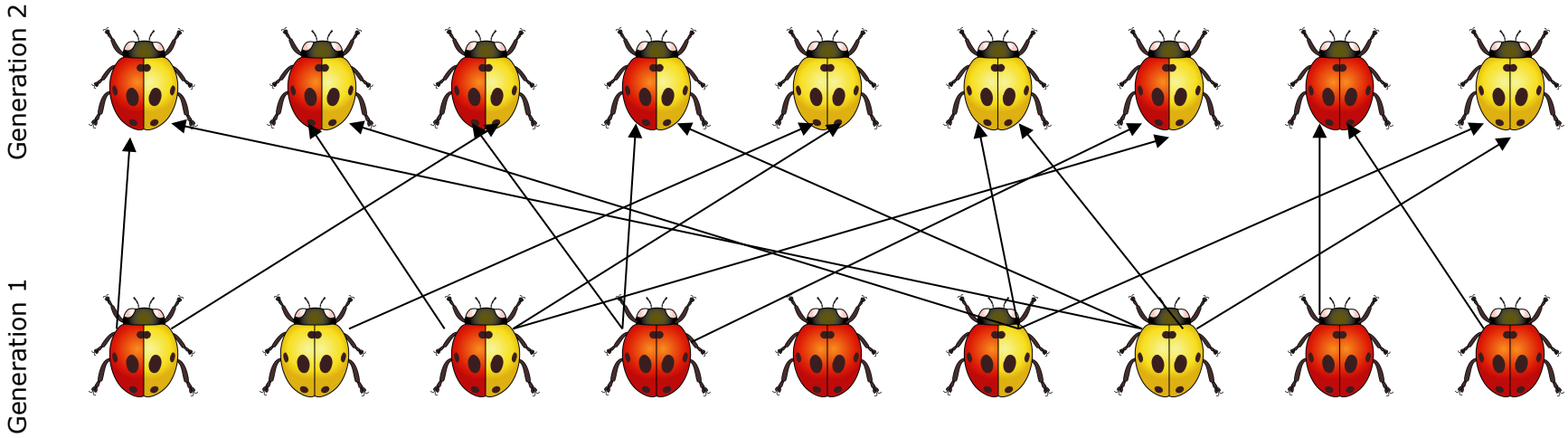
So the probability that **two copies of a gene came from the same copy** in the previous generation is **1/2N**



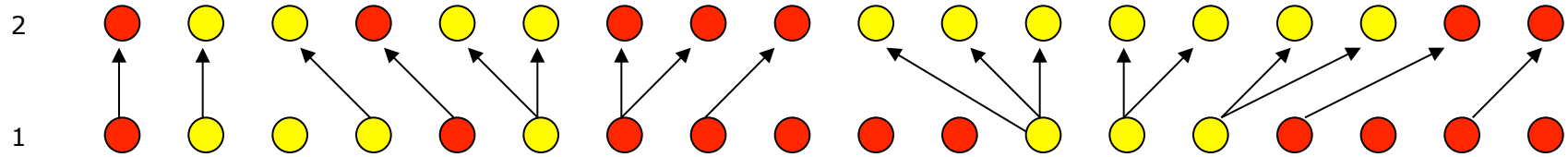
The Wright-Fisher Population Model

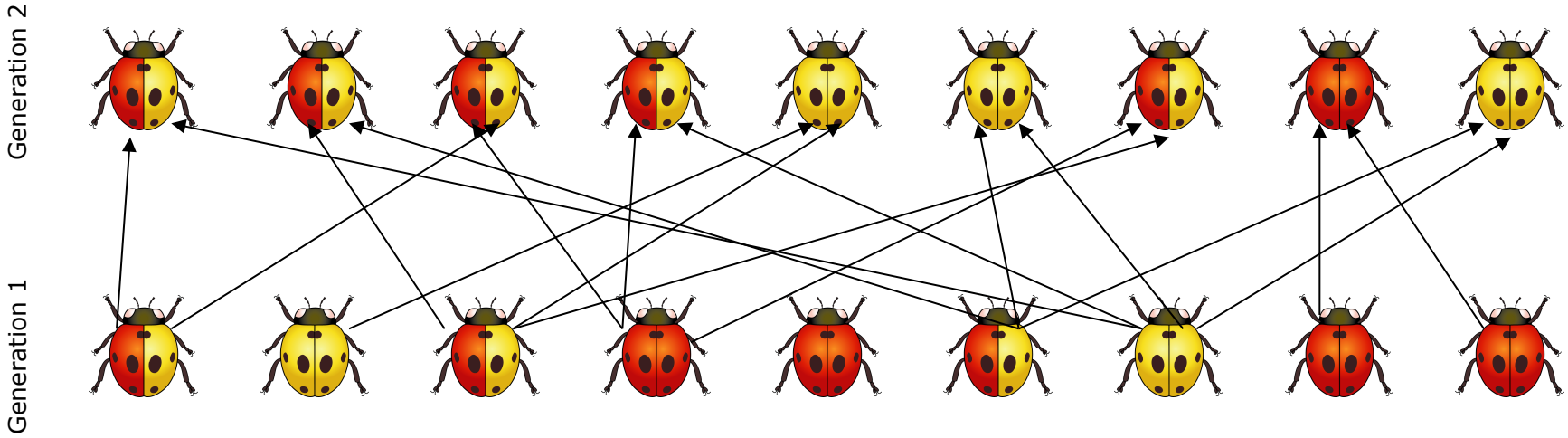
The arrows in this diagram contain a **genealogy of genes**

We can reveal this genealogy by redrawing the diagram in terms of gene copies rather than individuals (or alleles)



The Wright-Fisher Population Model



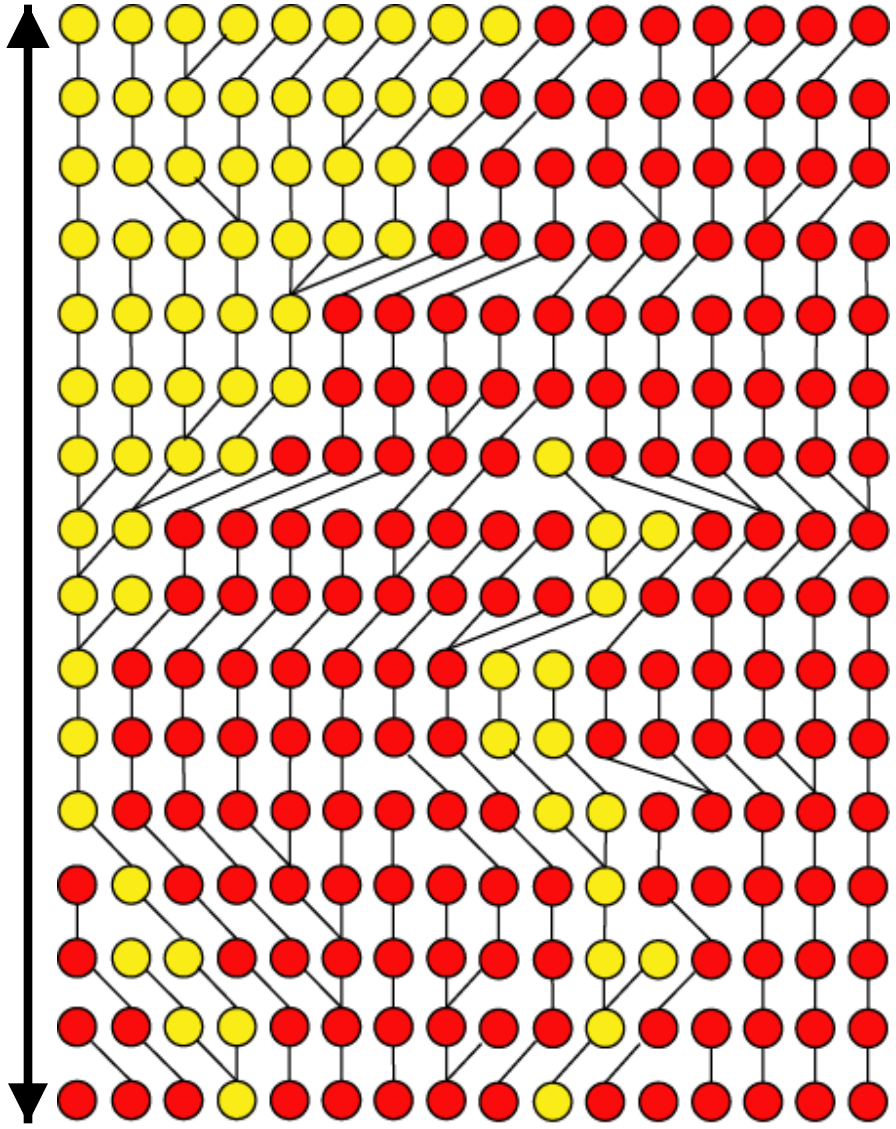


Evolutionary biologists
Analyse evolution backwards in time from the present

Base their research on a *sample* of extant individuals rather than knowledge of an entire population

Do not know initial population parameters (estimating these parameters may be the purpose of the research)

Are concerned with the **coalescence** of extant genes

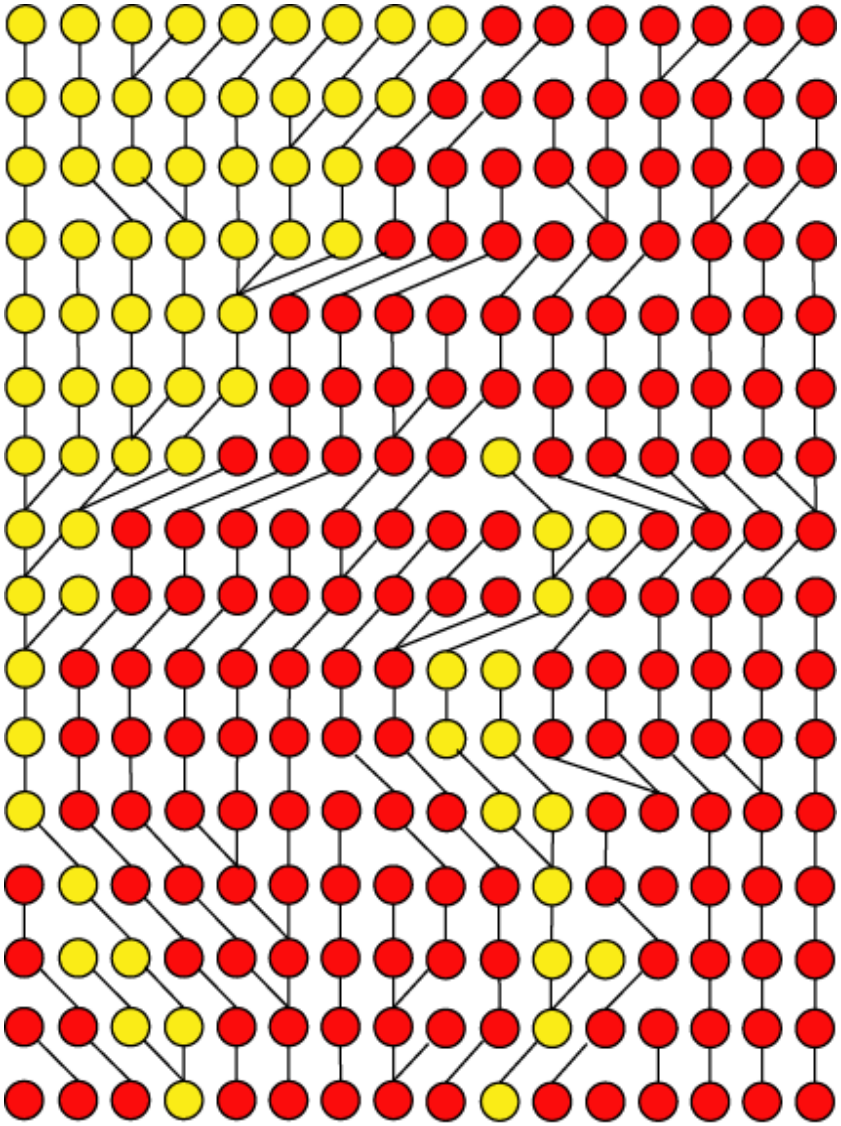


The coalescent

It is a model of the distribution of coalescent events on a gene genealogy

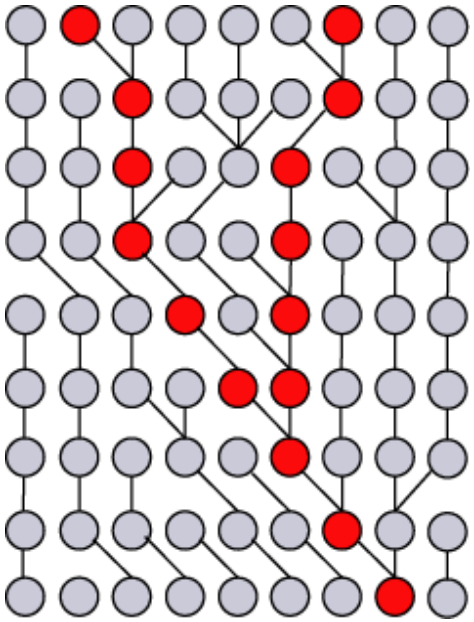
Based on a sample of extant gene copies and equipped with our favourite model of evolution, we use the coalescent to estimate population genetic parameters associated with coalescent events

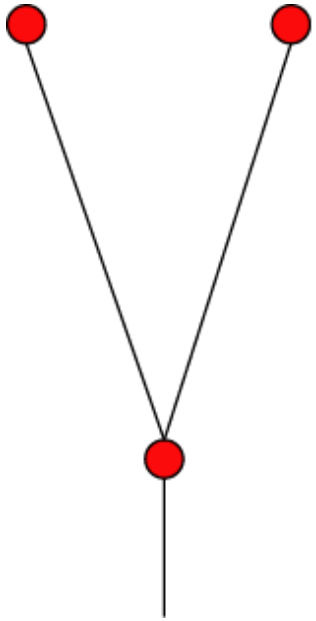
i.e. when was the most recent common ancestor of existing gene copies? What was the population size at the time of the coalescent event? How was the population changing before and after the coalescent event? How frequently do gene copies "go extinct"? What migration regime was operating in the historic population?



The Coalescent

We're going to stick with the Wright-Fisher model for a while



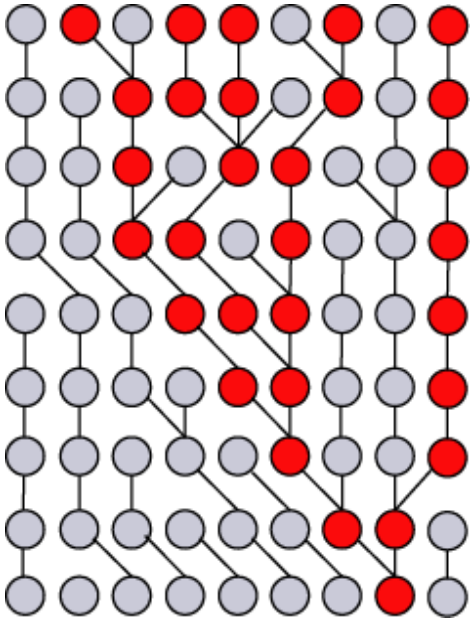


P(Coalesces 1 generation ago)	1/2N
P(Coalesces 2 generations ago)	(1-1/2N) * 1/2N
P(Coalesces 3 generations ago)	$(1-1/2N)^2$ * 1/2N
P(Coalesces 4 generations ago)	$(1-1/2N)^3$ * 1/2N
P(Coalesced *t* generations ago)	$(1-1/2N)^{t-1}$ * 1/2N

Coalescence of **two gene copies** follows a *geometric distribution* with mean **2N**

The Coalescent

So much for two gene copies. What about *k* gene copies?



There are k(k-1)/2 distinct pairs of genes that could coalesce

The probability that **one** of these coalesces in the previous generation is given by

$$P(\text{coalescence}) = \frac{k(k-1)}{2} * \frac{1}{2N}$$

Number of pairs of gene copies

Probability that a pair coalesces

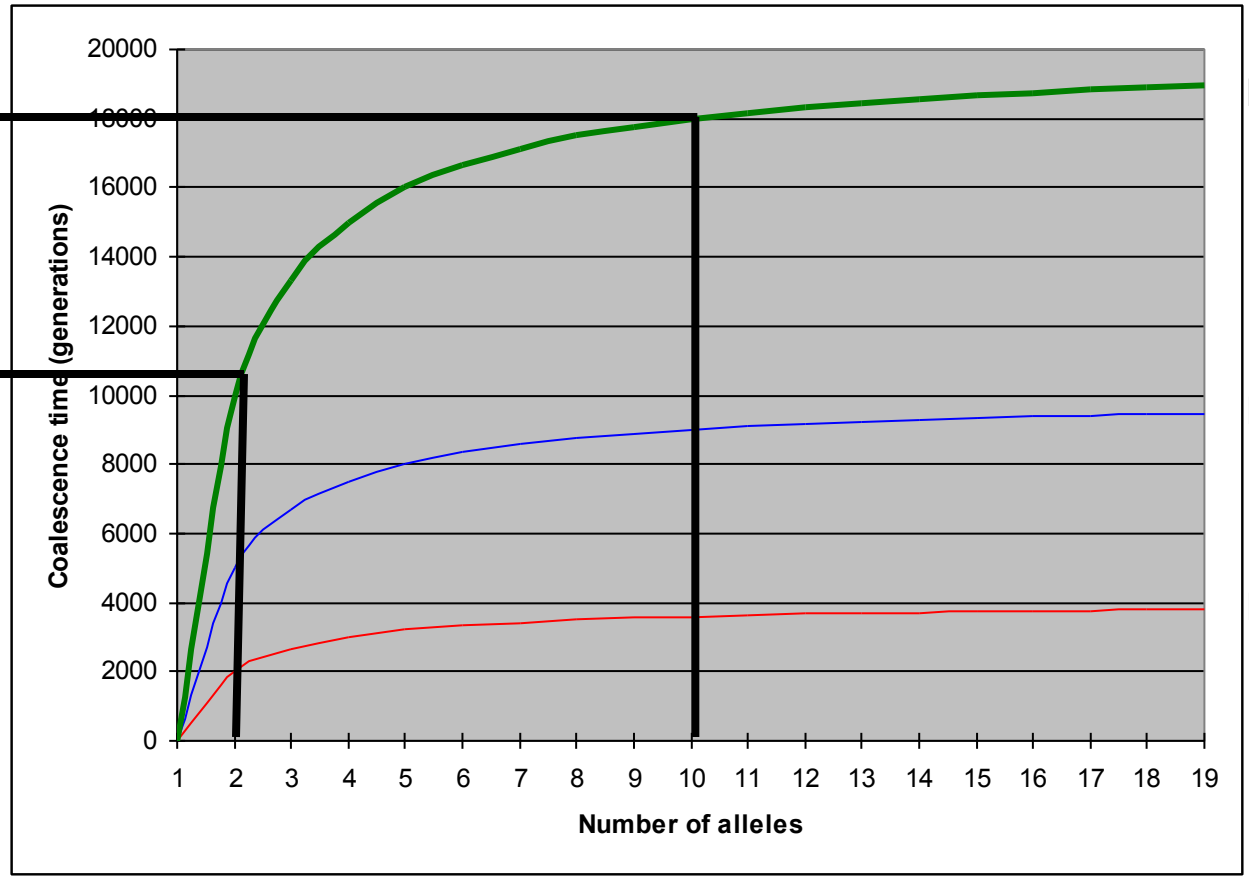
Can carry through the math – answer is 4N(1-1/k)
(or 2x what it is for a pair)

Properties of the Coalescent

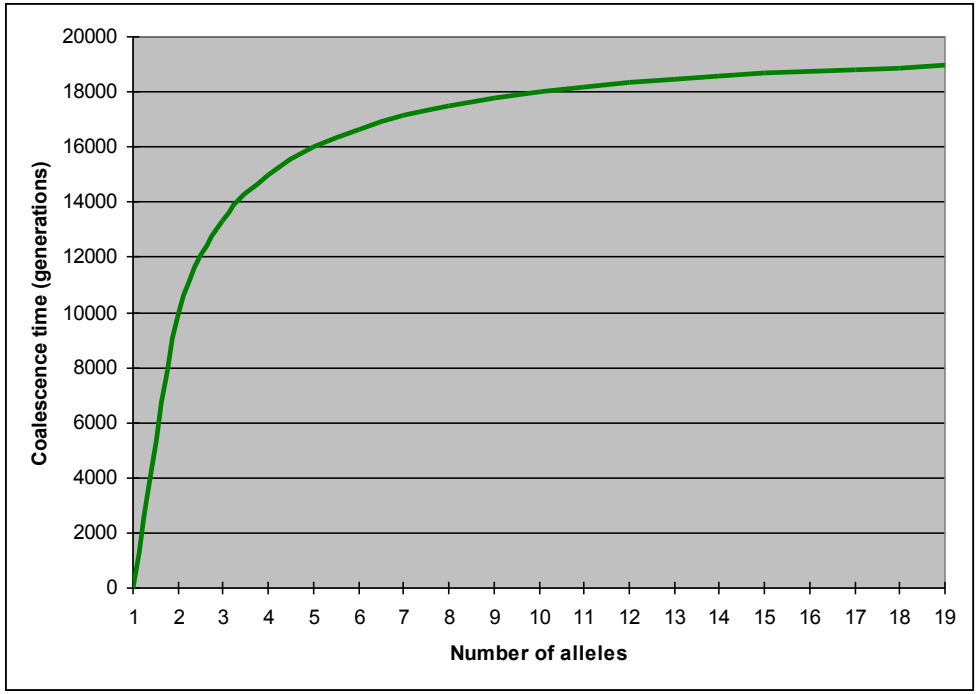
We start with 20 alleles and wait for them to coalesce until we reach the most recent common ancestor of all alleles

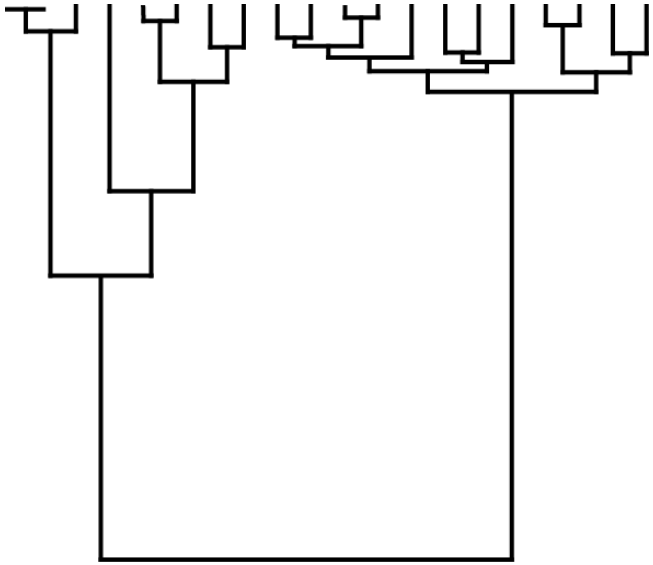
Half the alleles coalesce in the first 10% of time

50% of the total coalescence time is spent waiting for the last pair of alleles to coalesce!



Properties of the Coalescent





This means that coalescent trees are top-heavy!

Properties of the Coalescent

The fact that most branches coalesce at the top of the tree means that deep tree nodes can be inferred from a small number of gene copies

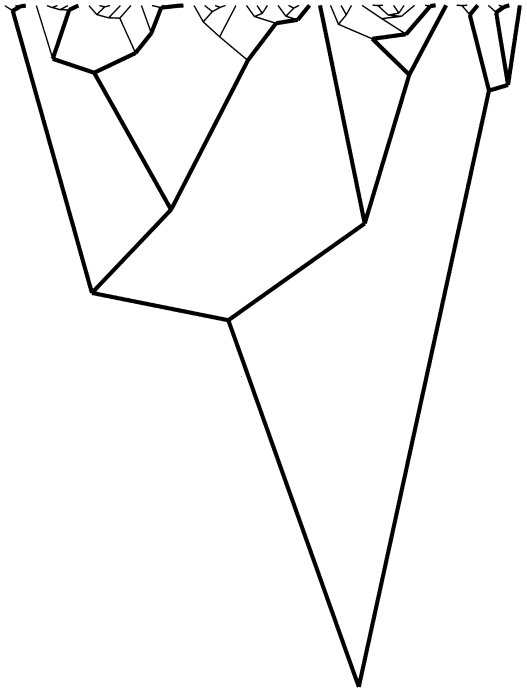
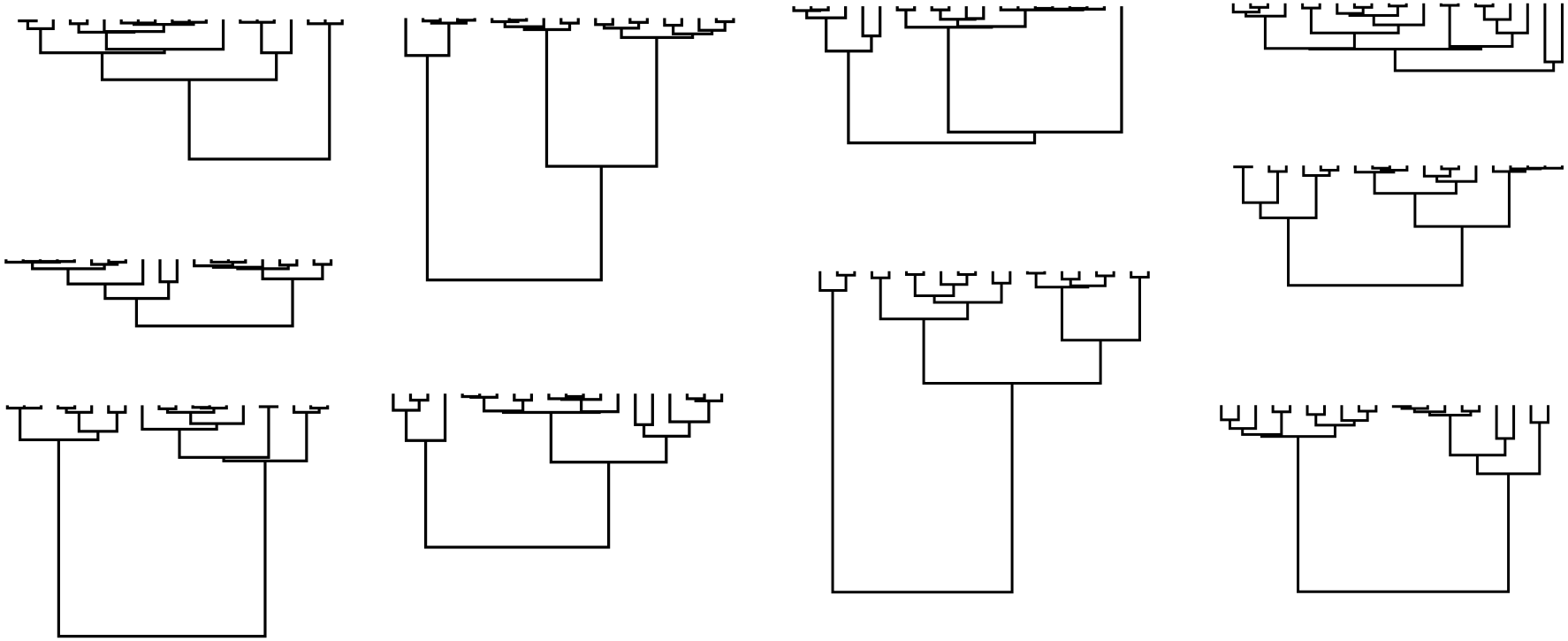


Figure 26.6: A sample genealogy of 50 gene copies, with the ancestry of a random 10 of them indicated by bold lines. Note that adding 40 more gene copies to the sample discloses no new lines in the bottom part of the diagram.

Properties of the Coalescent

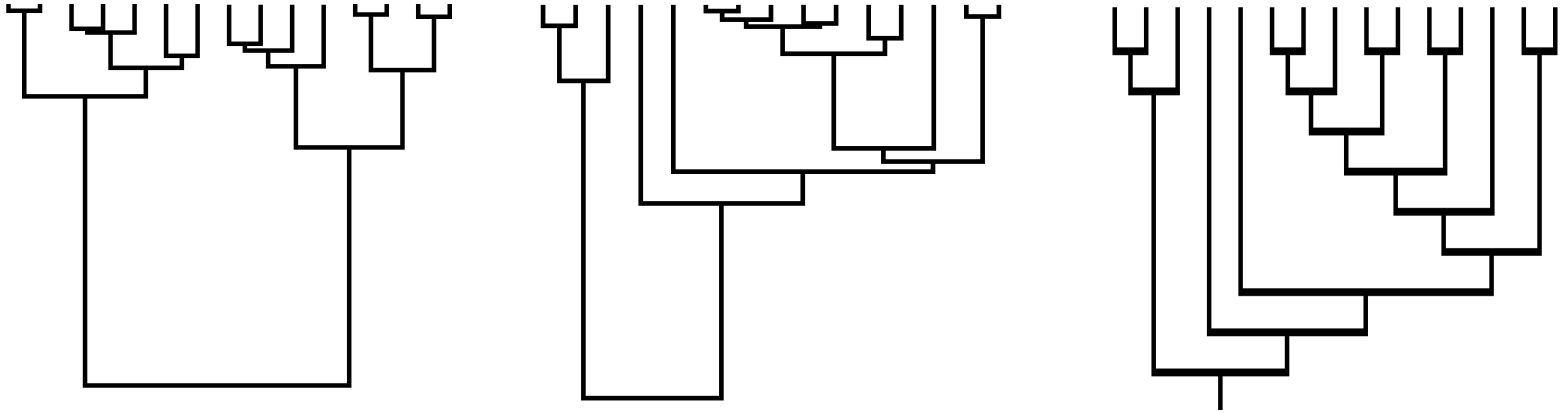
The exponential nature of the time between coalescent events makes the coalescent distribution very noisy. These are tree simulated under a stochastic version of the coalescent with an identical N and k.



Properties of the Coalescent

The coalescent can be used to simulate a large number of possible genealogies. **Some of these genealogies are more likely than others**.

The most likely tree is one in which each coalescence event occurs exactly at the expected time according to the coalescent distribution. The further the topology of the simulated tree is from the expected distribution of the coalescent, the less likely it is to be the REAL history of population coalescence.



High likelihood

medium likelihood

low likelihood

http://www.mesquiteproject.org

Properties of the Coalescent

What is the coalescence *rate* per unit time?

We saw earlier that there are $\frac{k(k-1)}{2}$ possible pairs of alleles that could coalesce

There are 2N alleles in a diploid population

So the average rate of coalescence is $\frac{k(k-1)}{2}$ / 2N

$= k(k-1)/4N$

Summary of the basic coalescent

Expected coalescence time for *k* alleles is exponentially distributed

with a mean ≈ **4N** and coalescence rate of **k(k-1)/4N**
for diploid populations

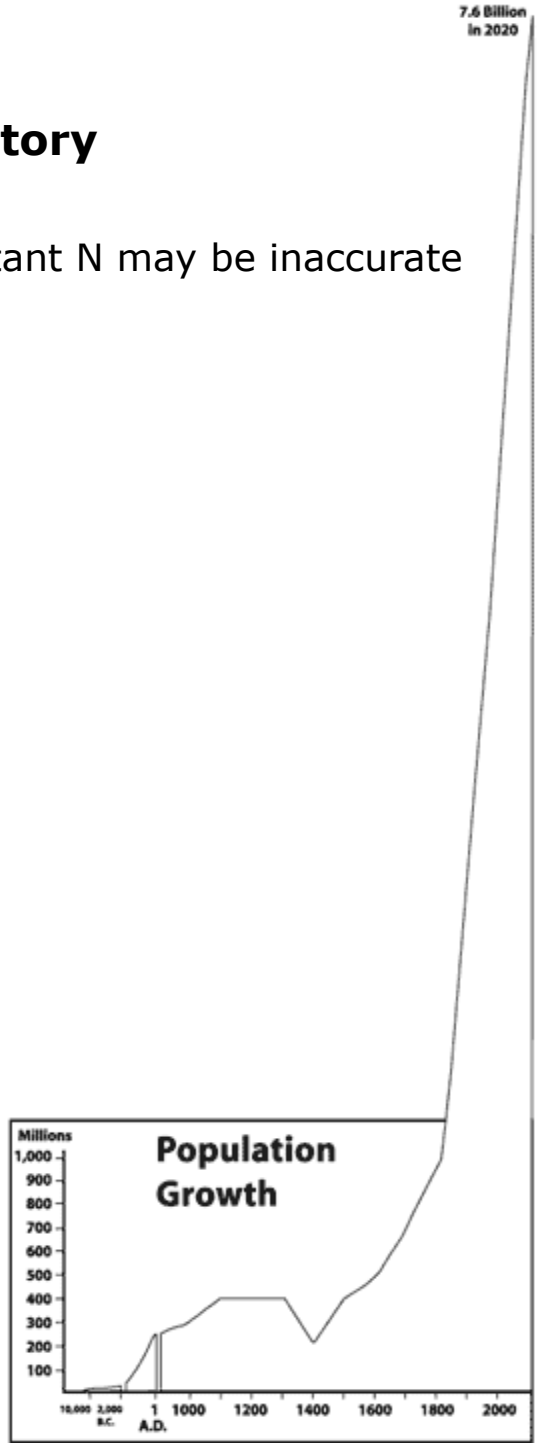
with a mean ≈ **2N** and coalescence rate of **k(k-1)/2N**
for haploid populations

with a mean ≈ $\mathbf{2N_f}$ and coalescence rate of $\mathbf{k(k-1)/2N_f}$
for populations of mitochondria

when *k* is large

Inference of ancestral population history

The Fisher-Wright model's assumption of constant N may be inaccurate

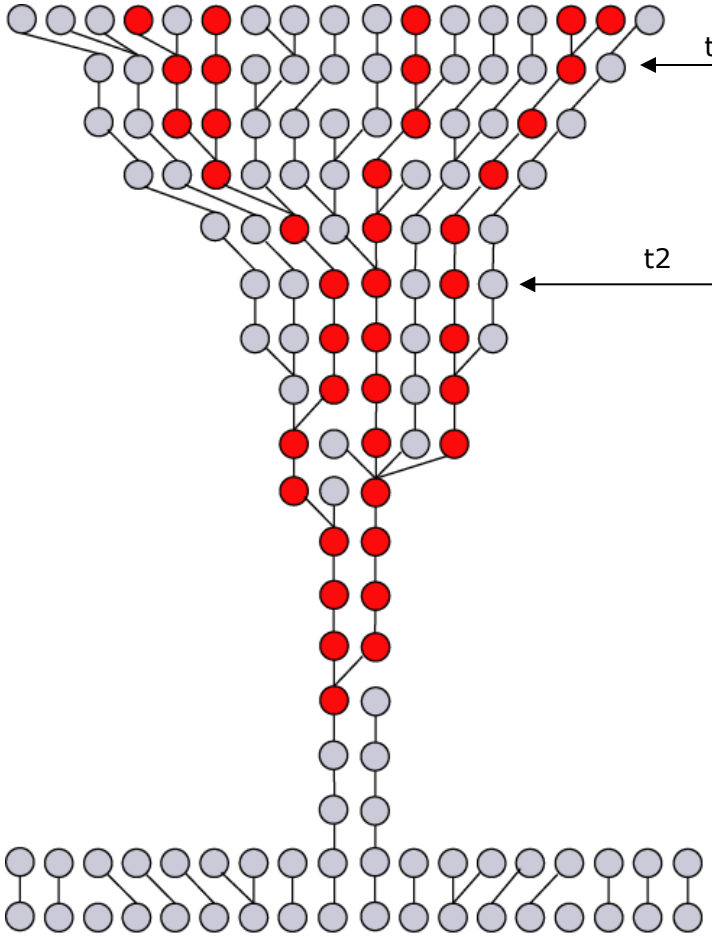


http://www.pbs.org

Inference of ancestral population history

The Fisher-Wright model's assumption of constant N may be inaccurate

Changes to ancestral population sizes are of interest to evolutionary biologists



2N=14, so coalescence rate is k(k-1)/4N
= k(k-1)/28

2N=7, so coalescence rate is k(k-1)/4N
= k(k-1)/14

Coalescence occurs twice as fast when when N is reduced by 50%

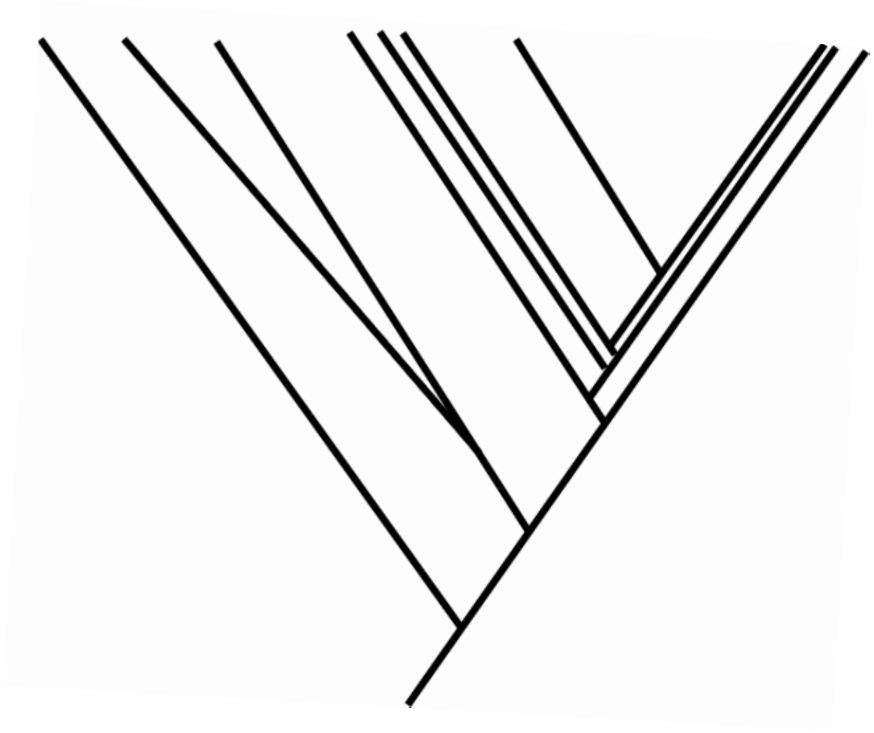
This gives the effect of time travelling twice as fast at time t2 as it does at time t1, **under constant N**

Mick’s basic conceptual understanding of coalescence times and population size...

You sample a gene from 10 members of a population

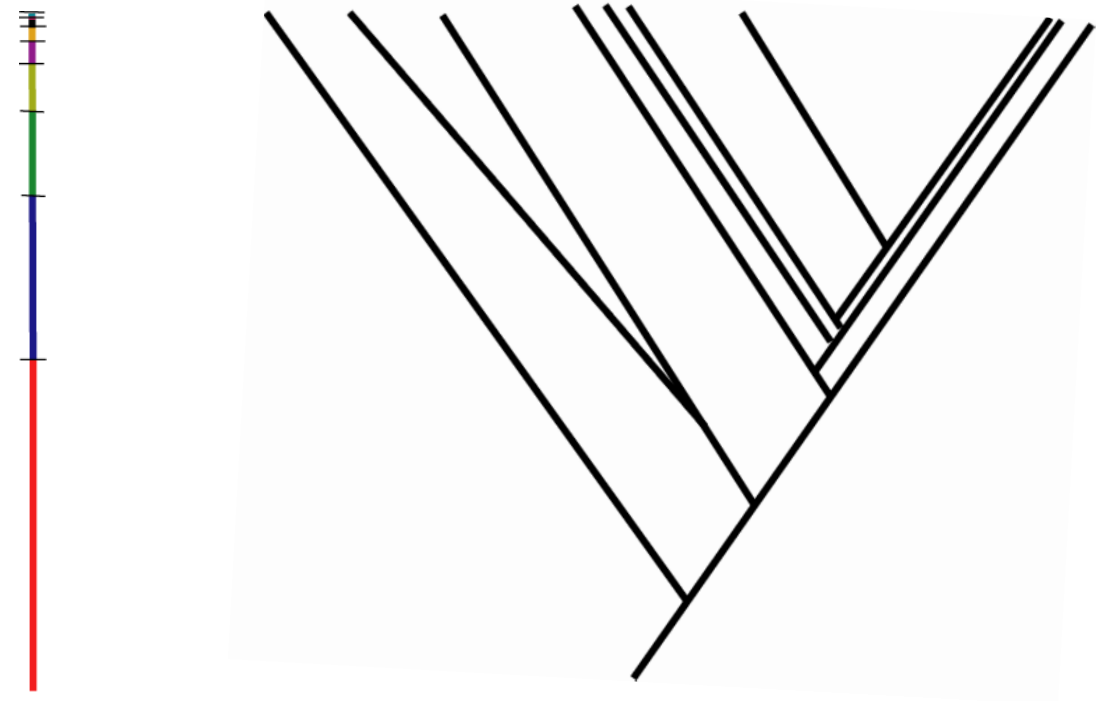
Mick's basic conceptual understanding of coalescence time and population size...

You estimate a phylogeny for these 10 members of the population…



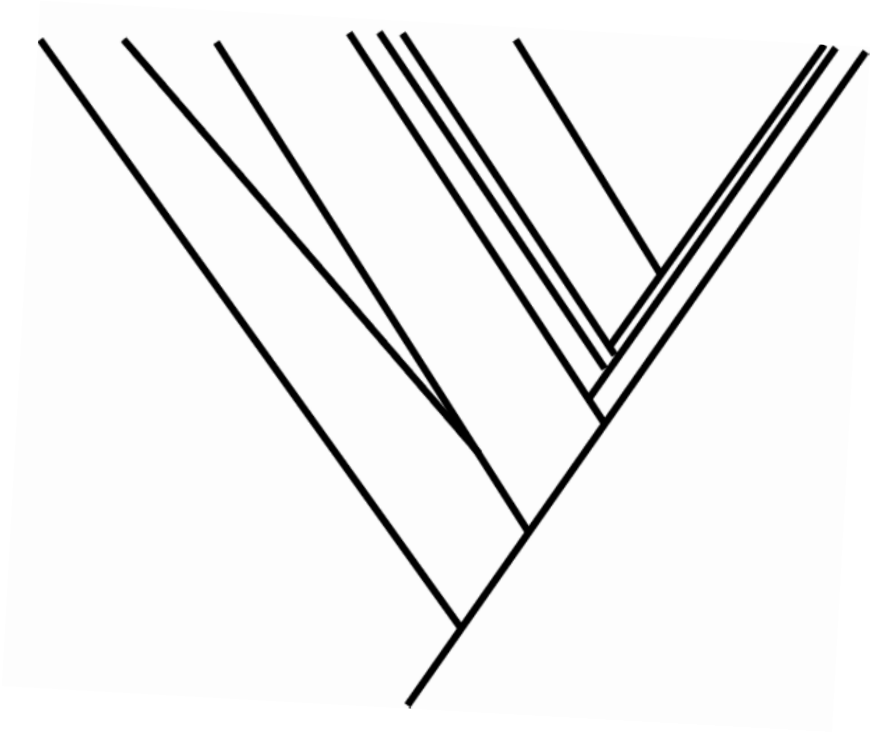
Mick's basic conceptual understanding of coalescence time and population size...

But the most likely coalescent tree for these genes looks very different!



Mick's basic conceptual understanding of coalescence time and population size...

But the most likely coalescent tree for these genes looks very different!



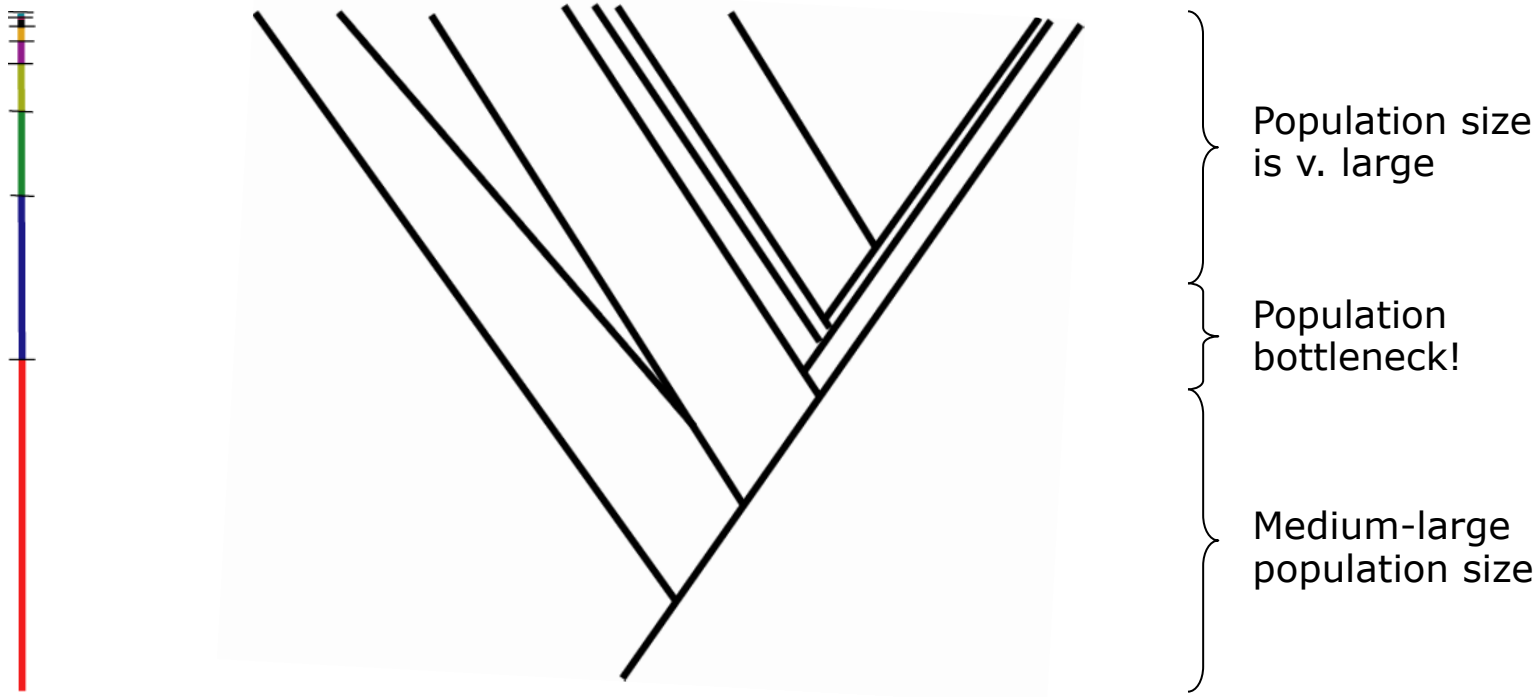
Coalescence is happening much slower than expected

Coalescence is happening much faster than expected

Coalescence is happening moderately faster than expected

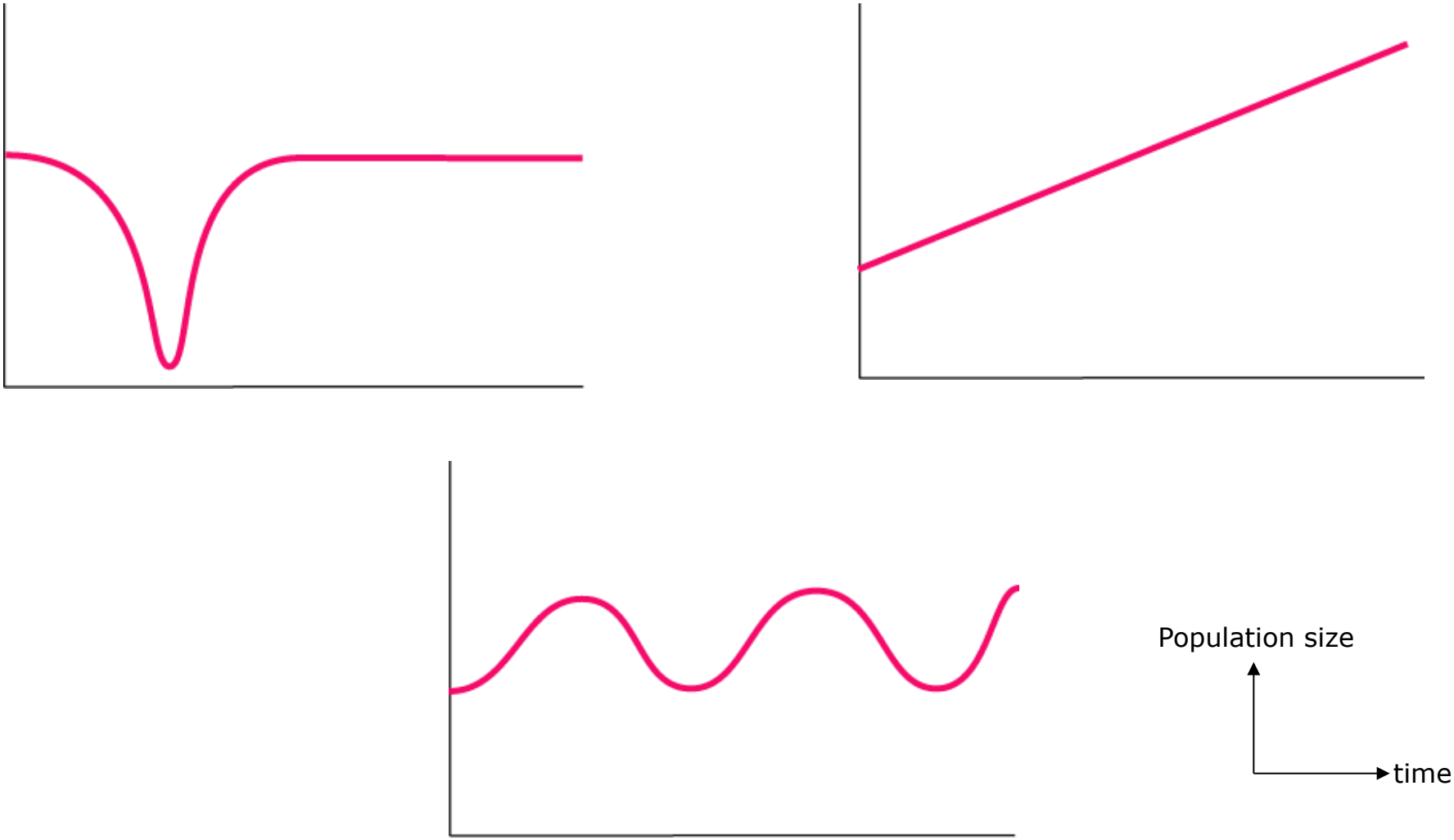
Mick's basic conceptual understanding of coalescence time and population size...

But the most likely coalescent tree for these genes looks very different!

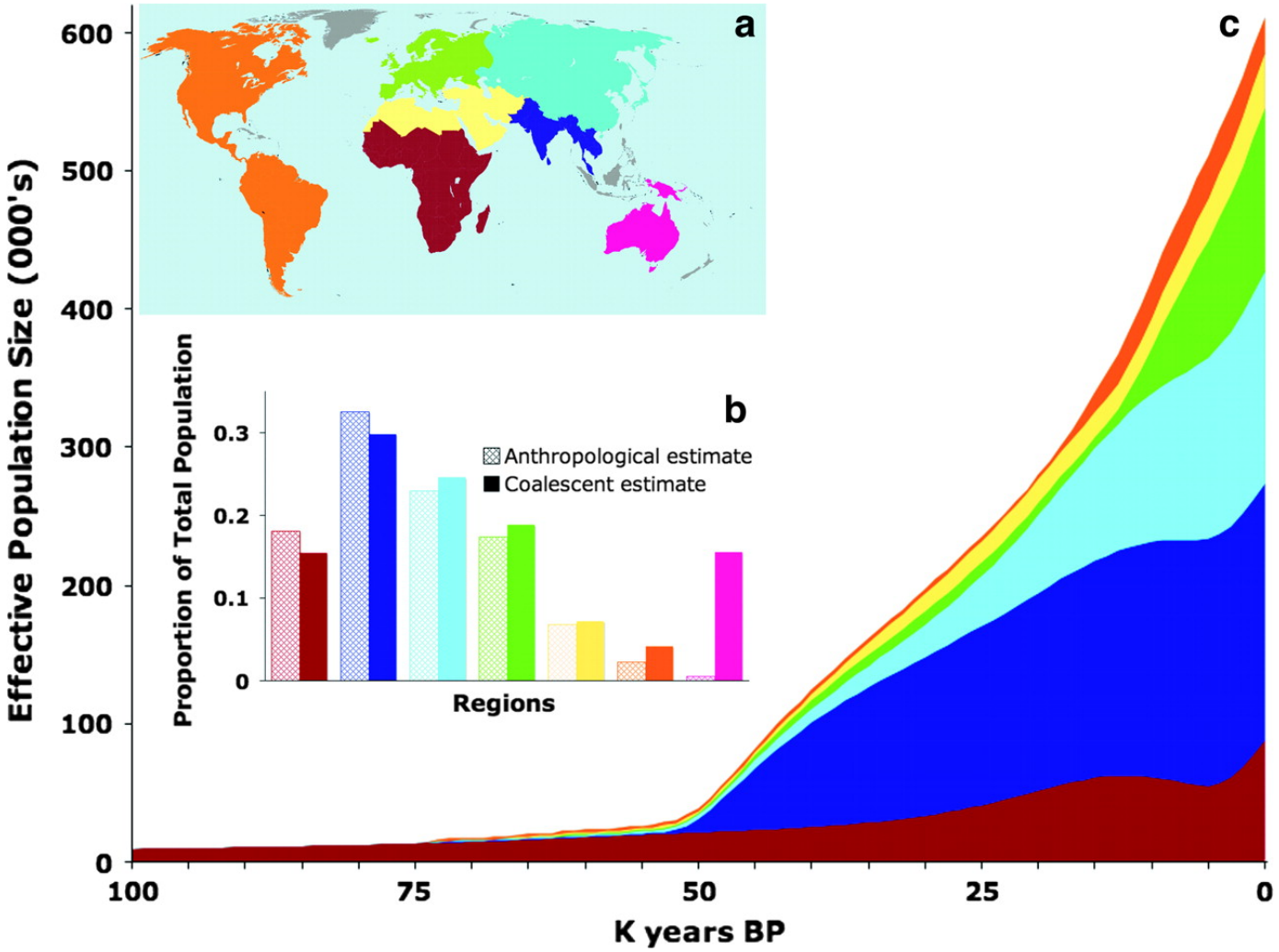


Inference of ancestral population history

We can use this method for any model of population size change that can be integrated with respect to t

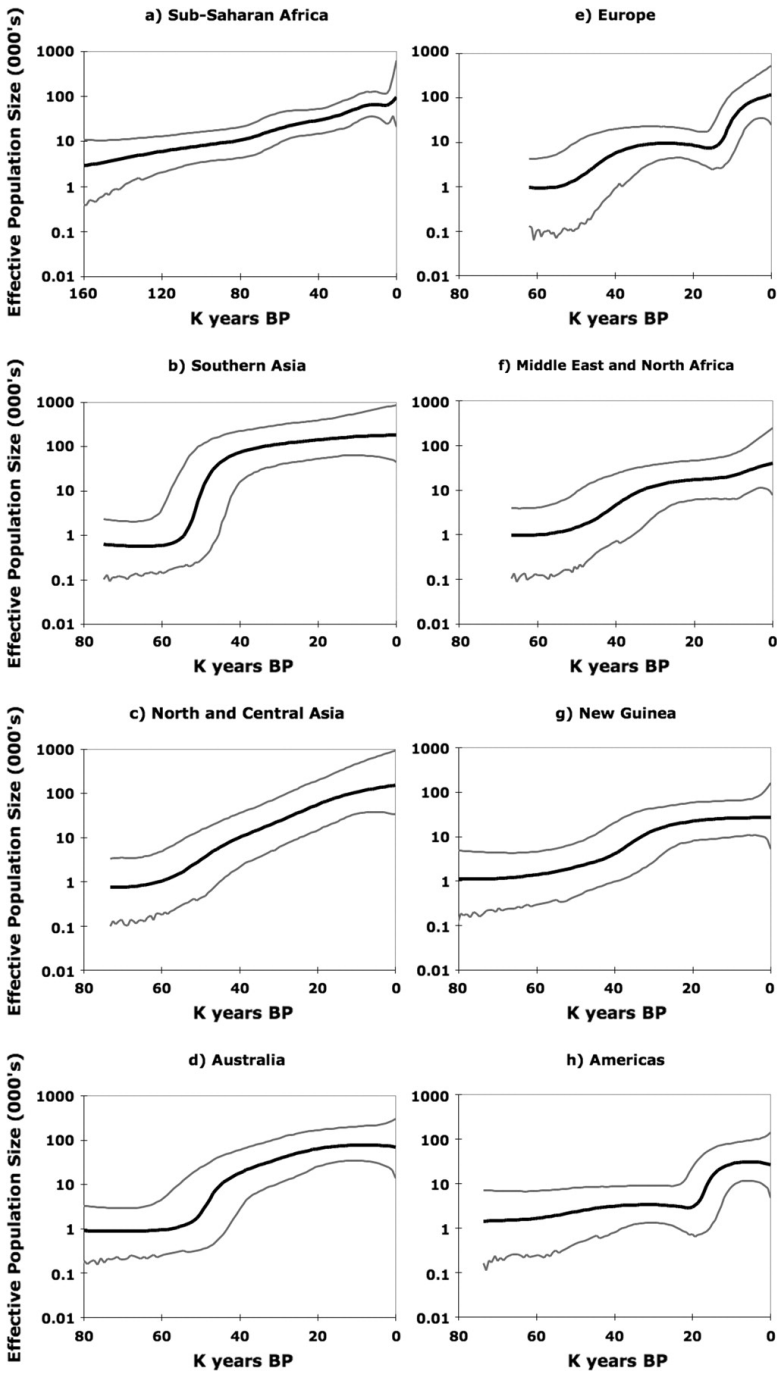


Comparative analysis of relative regional population sizes through time.



Atkinson Q D et al. Mol Biol Evol 2007;25:468-474

MBE MOLECULAR BIOLOGY AND EVOLUTION



mtDNA Variation Predicts Population Size in Humans and Reveals a Major Southern Asian Chapter in Human Prehistory

Atkinson Q D et al. Mol Biol Evol 2007;25:468-474

Bayesian Skyline Plots of effective population size through time.

MBE MOLECULAR BIOLOGY AND EVOLUTION