Evolutionary Genomics

University of Bern, 2020

Vitor Sousa
vmsousa@fc.ul.pt

Evolutionary forces affecting the history of populations

Demography

- Past effective population sizes
- Past migration rates

Selection

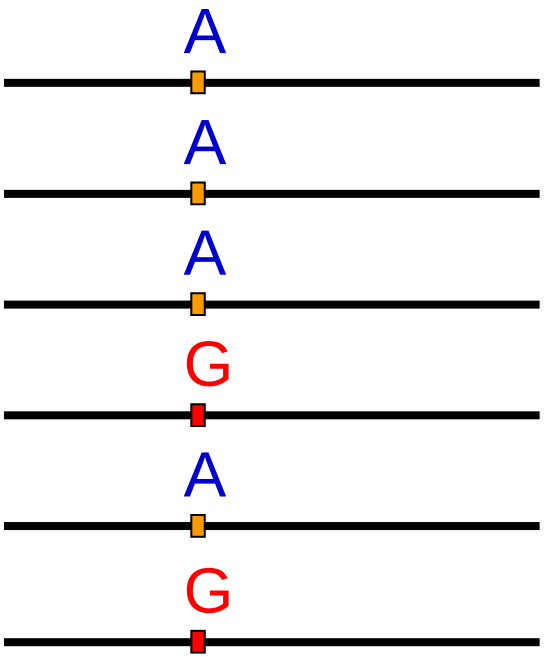
- Selective coefficient and type of selection (positive or negative)

Genomic processes

- Mutation rate
- Recombination rate

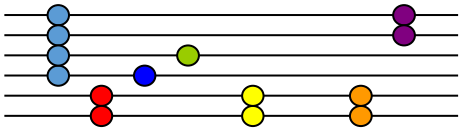
Mutations at the molecular level

Single Nucleotide Polymorphisms (SNPs)

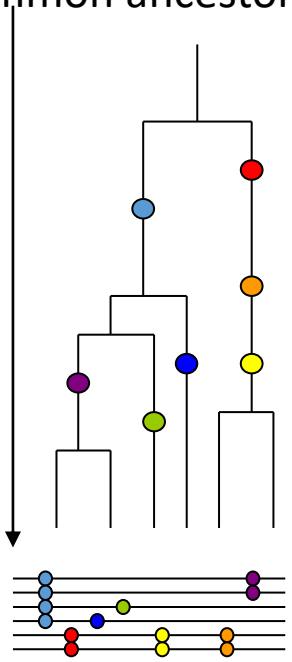


A SNP is simply a polymorphic nucleotide on a chromosome

When we compare DNA sequences, we find several polymorphic sites (SNPs) with variable frequencies



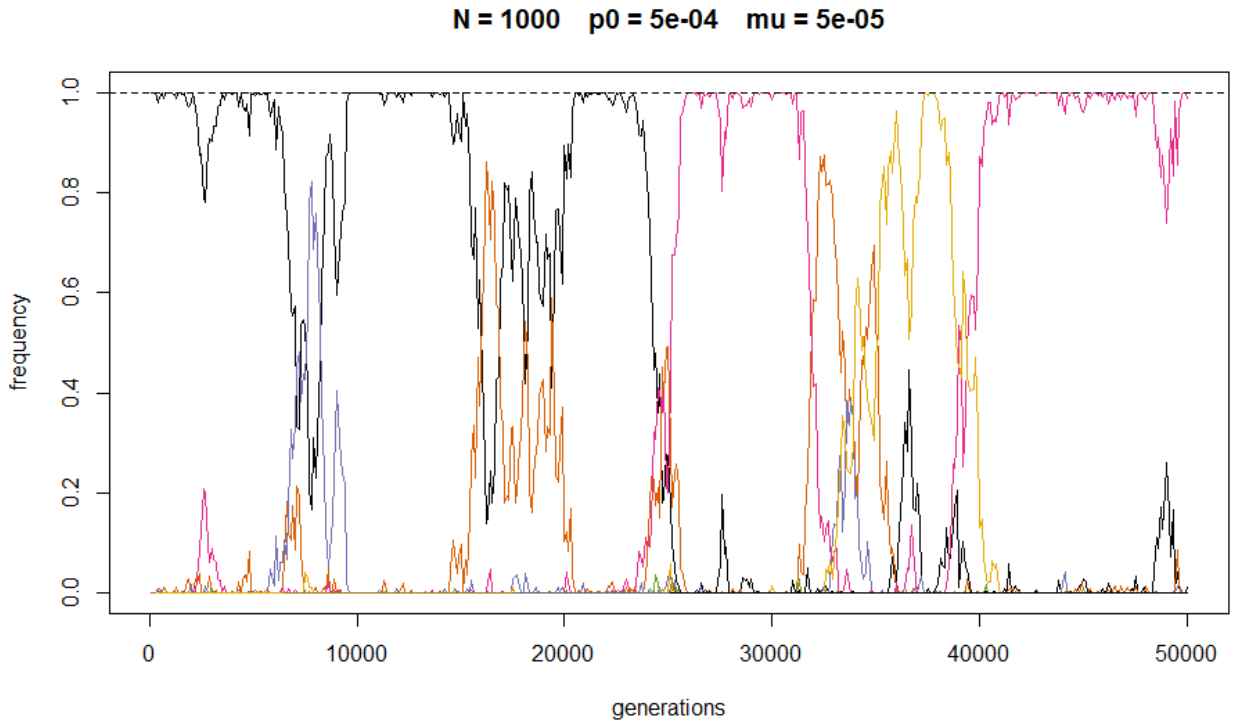
These SNPs have been produced by a series of mutations along the evolution of the DNA sequences since their most recent common ancestor



Mutation-drift equilibrium under the IAM

We have seen that a under the effect of genetic drift, the population will fix one allele. But this result assumes that no mutation occurs at all during this fixation process.

More realistically, like in the case of migration, the loss of alleles by genetic drift can be compensated by the creation of new alleles by mutation.

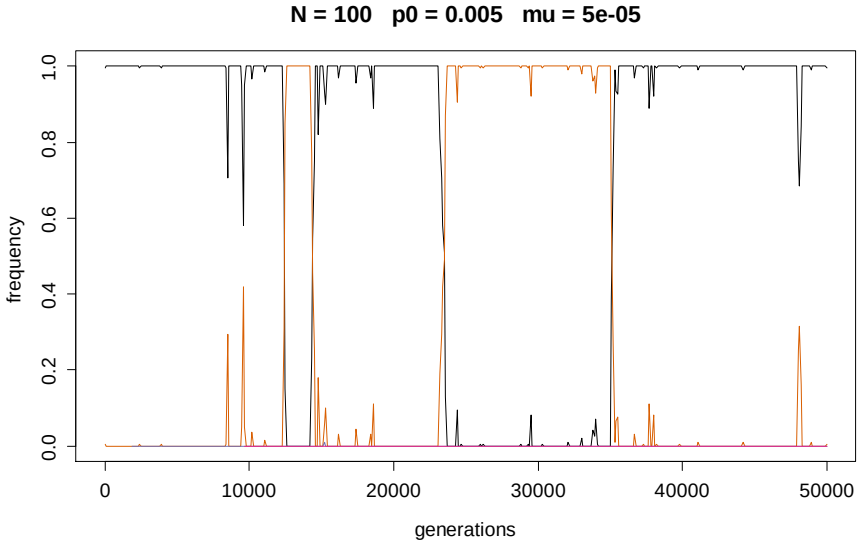


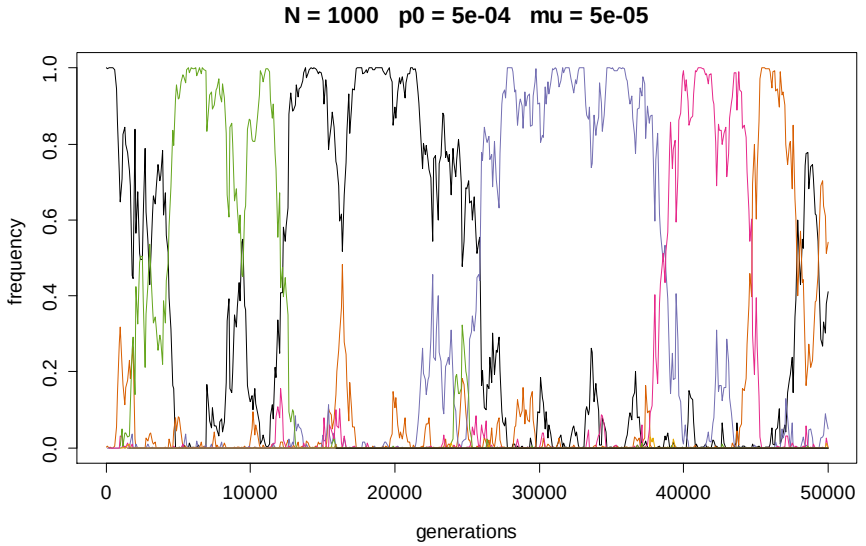
Evolution of allele frequencies over time under drift and mutations

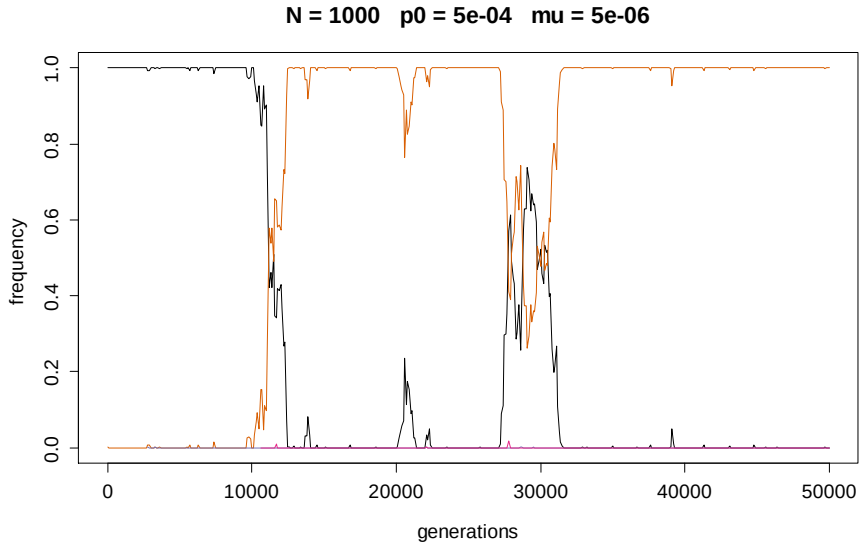
We can reach a steady state, with a constant influx of new mutations and the maintenance of a polymorphism

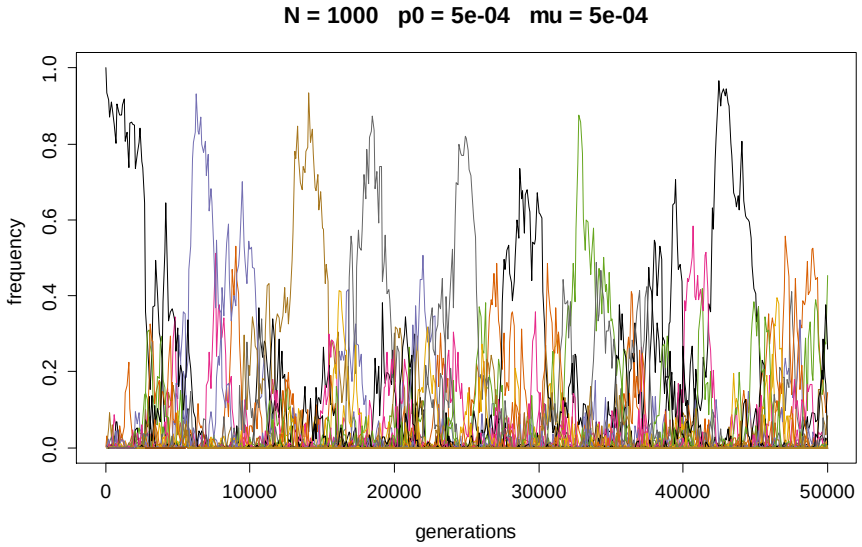
Mutation-drift equilibrium under the IAM

The amount of polymorphism will increase with the size of the population and the mutation rate (mu=μ)







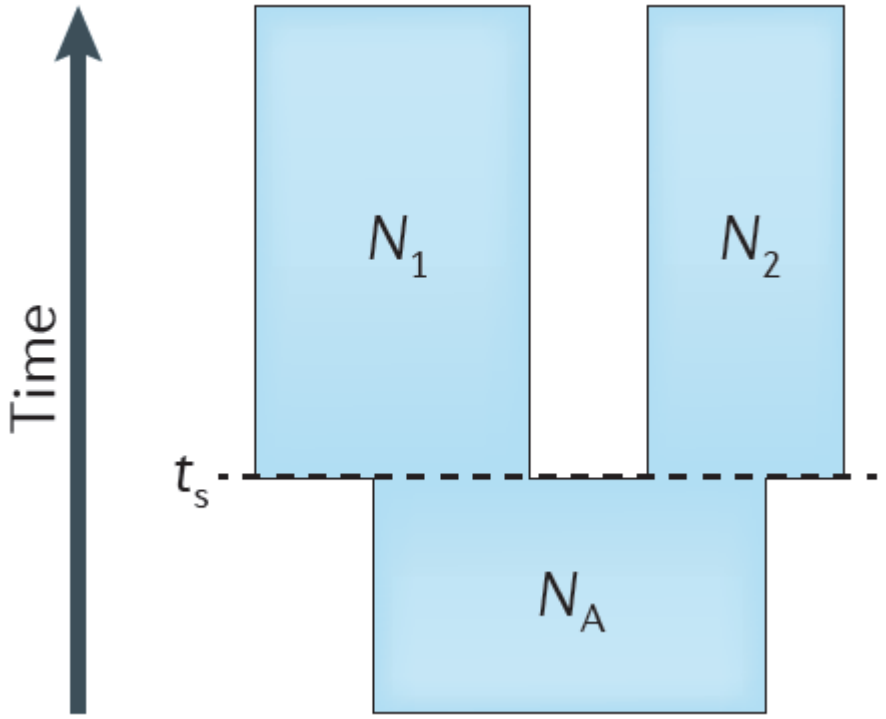


Demographic history of populations

Past demographic events:

- Population split
- Migration events
- Changes in effective population sizes (expansions or bottlenecks)
- Temporal changes in migration rates and effective sizes

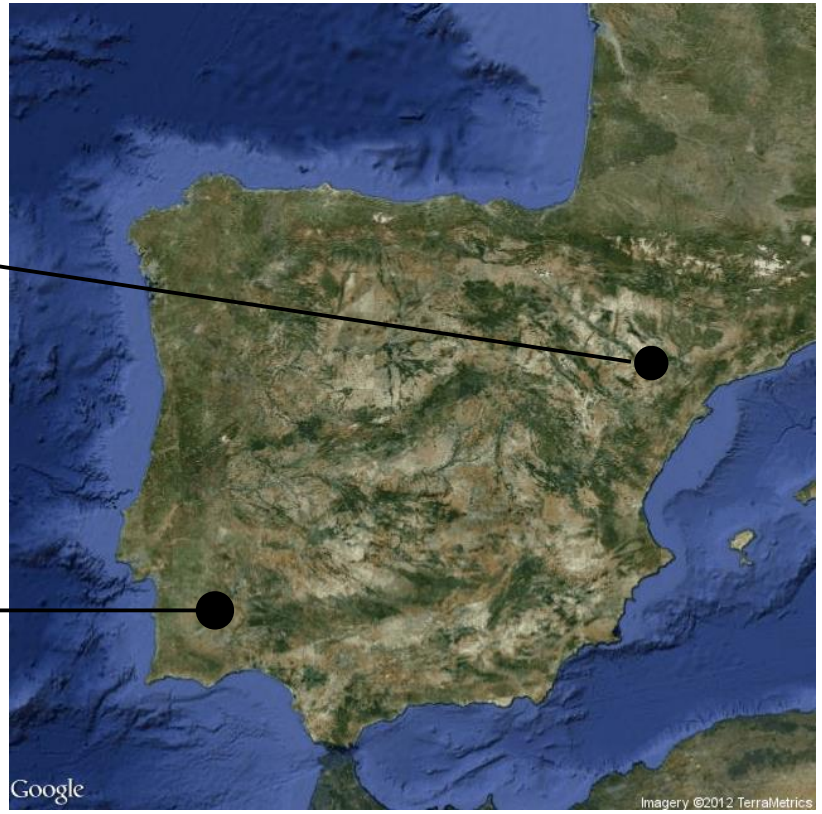
Population tree



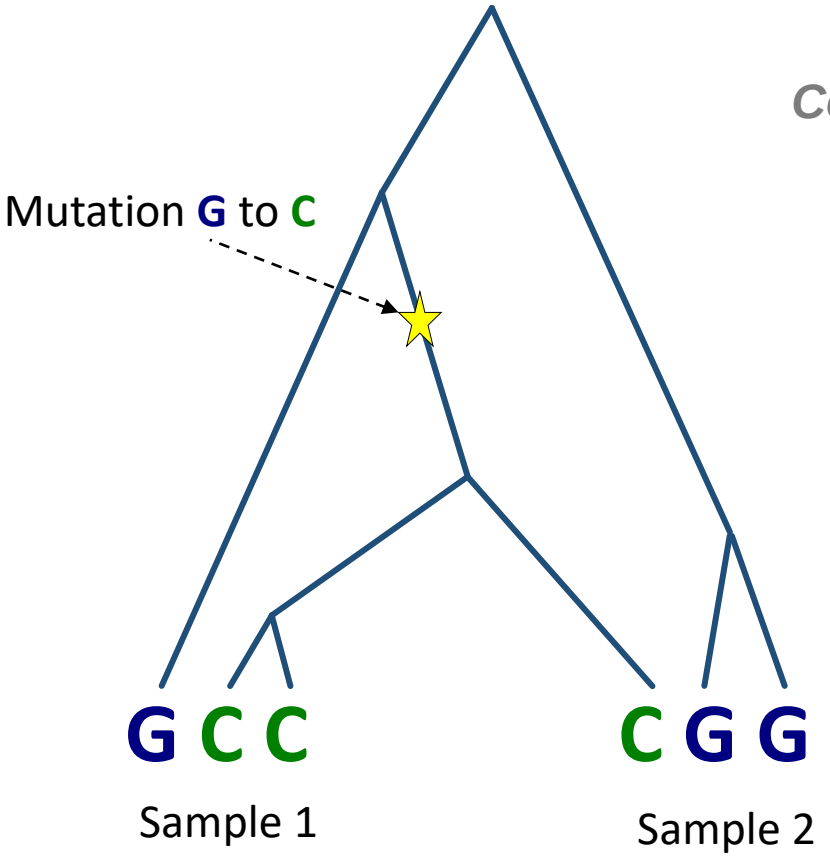
How to connect alleles with demographic history?

Sample site 1
ind1 AT**G**C – allele 1
ind2 AT**C**C – allele 2
ind3 AT**C**C – allele 2

Sample site 2
ind1 AT**C**C – allele 2
ind2 AT**G**C – allele 1
ind3 AT**G**C – allele 1

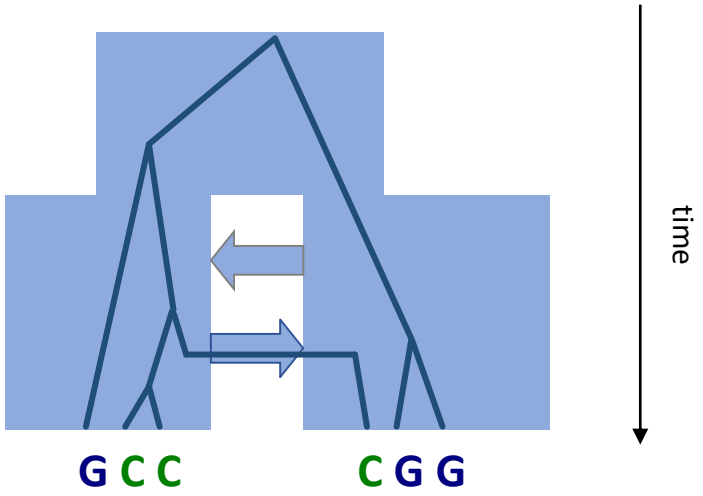


Gene trees reflect the species history

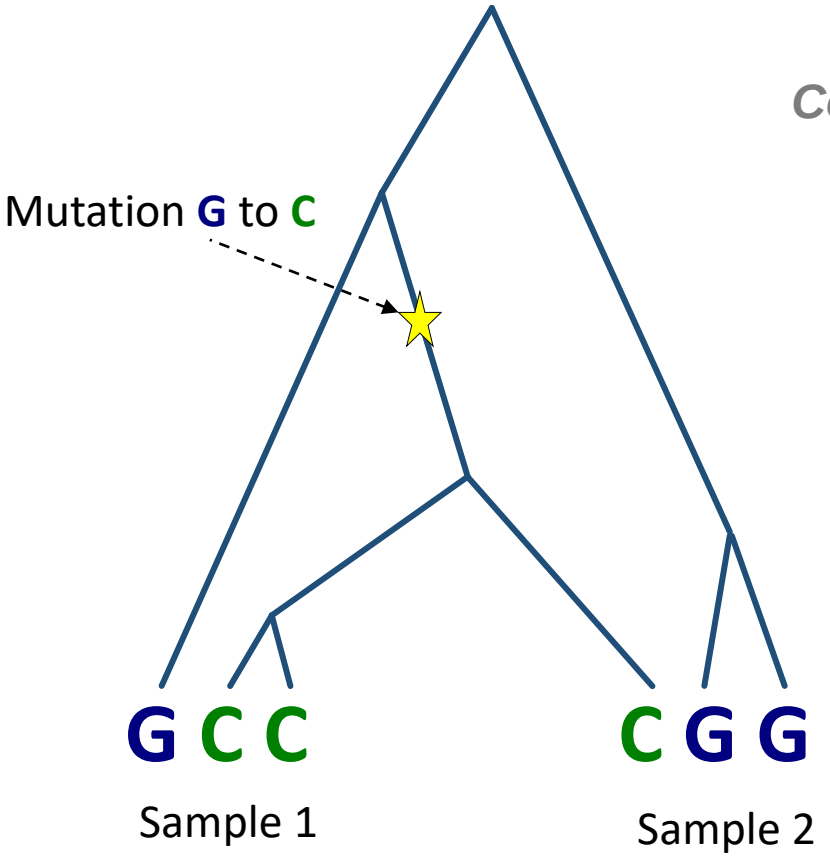


Compatible

Population tree with migration

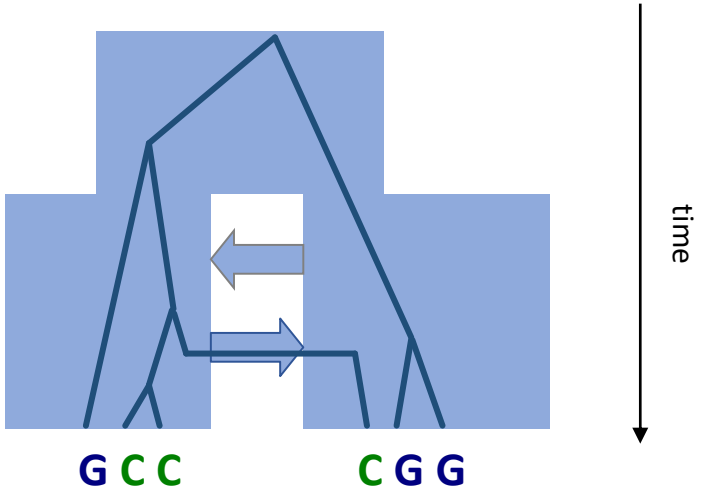


Gene trees reflect the species history

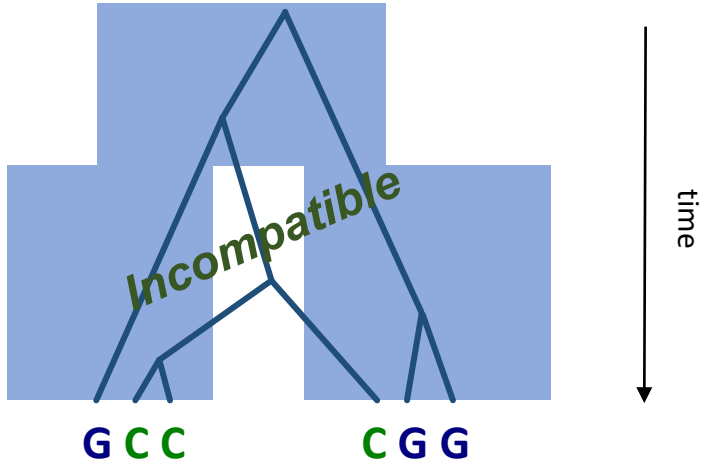


Population tree with migration

Compatible

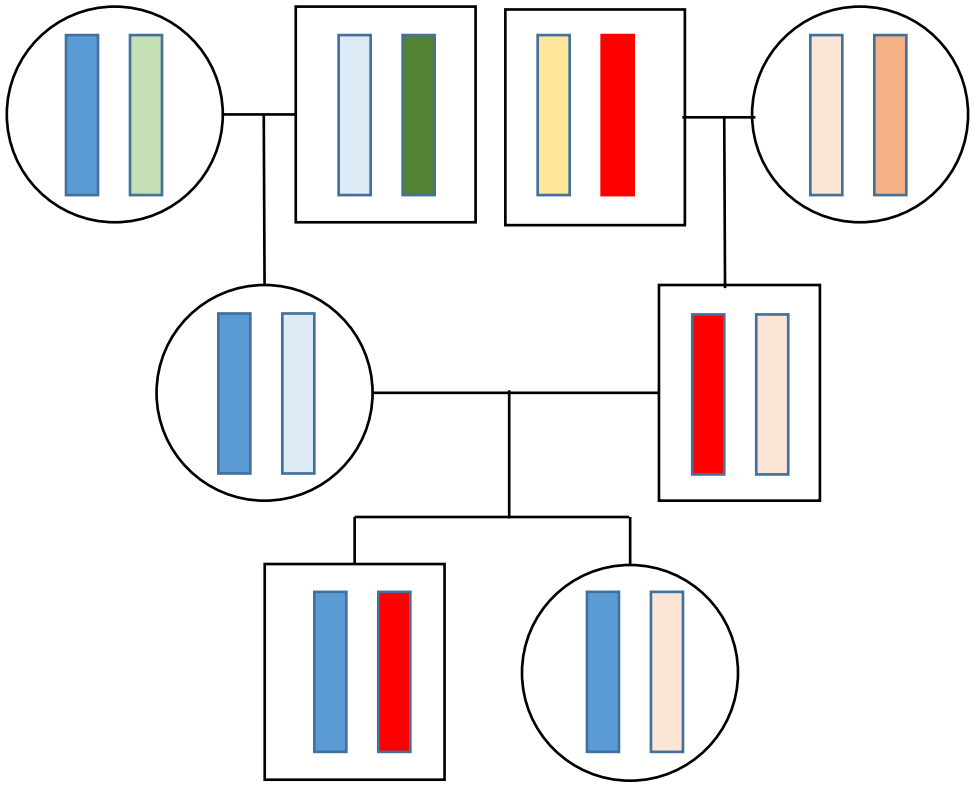


Population tree with no migration

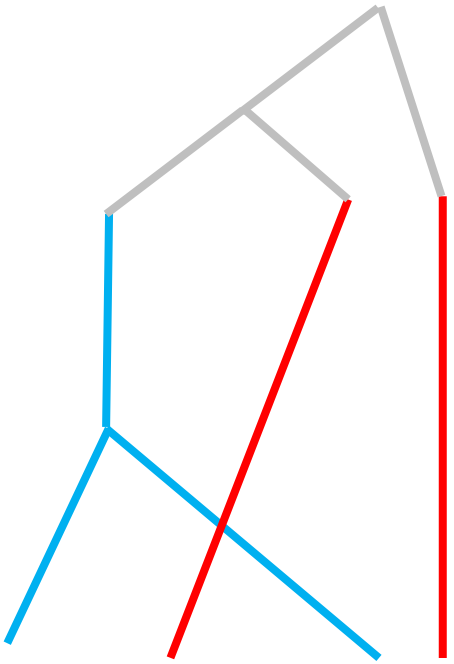


Gene trees within pedigrees

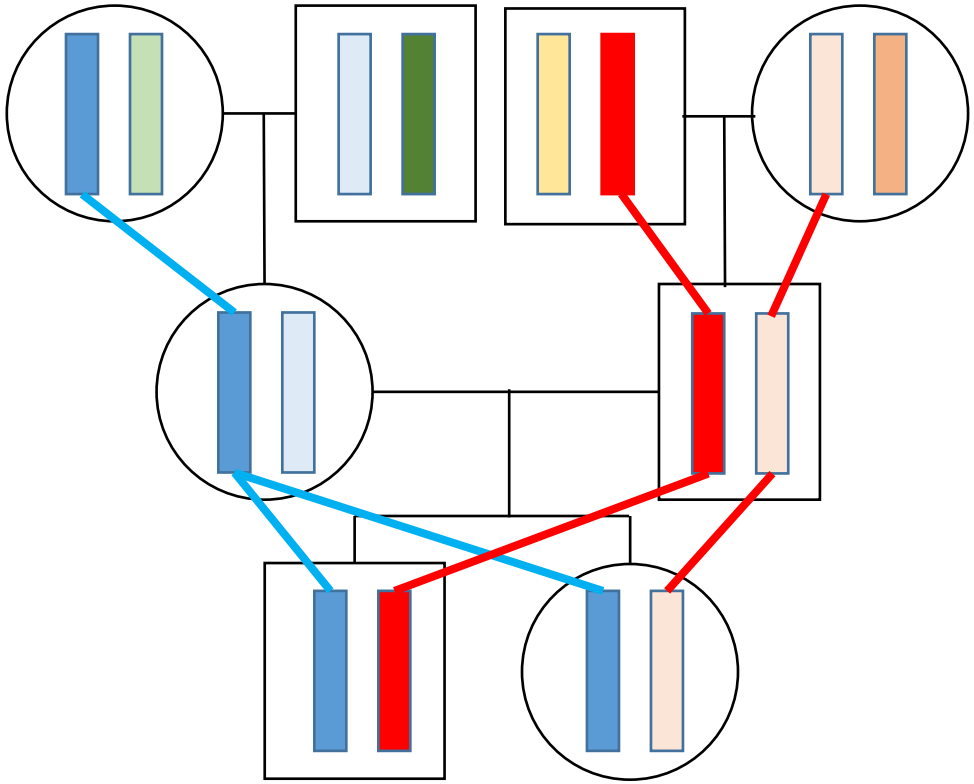
- **Gene trees** reflect the ancestral relationship of sampled gene copies
- For now, lets assume there is no mutations (**branch lengths do not reflect mutations in coalescent gene trees**!).
- Because of transmission of genes at each generation, at each position of the genome there is always a gene tree describing the relationship of gene copies in our sample.
- All individuals share the same pedigree, but gene trees can vary due to independent segregation and recombination



The same pedigree can lead to different gene trees across the genome

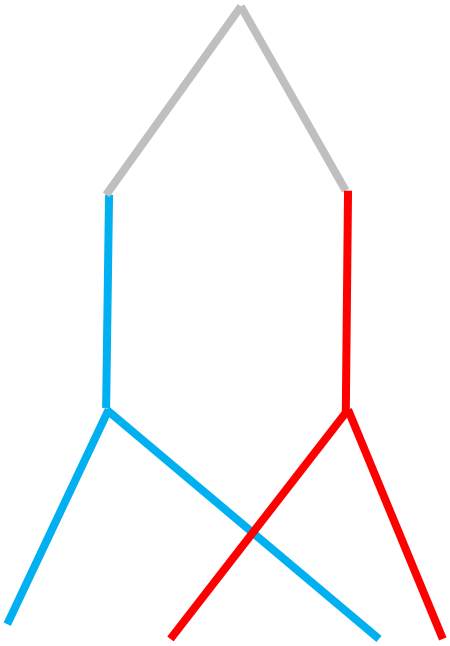


Gene tree of a sample of size n=4
(2 diploid individuals)

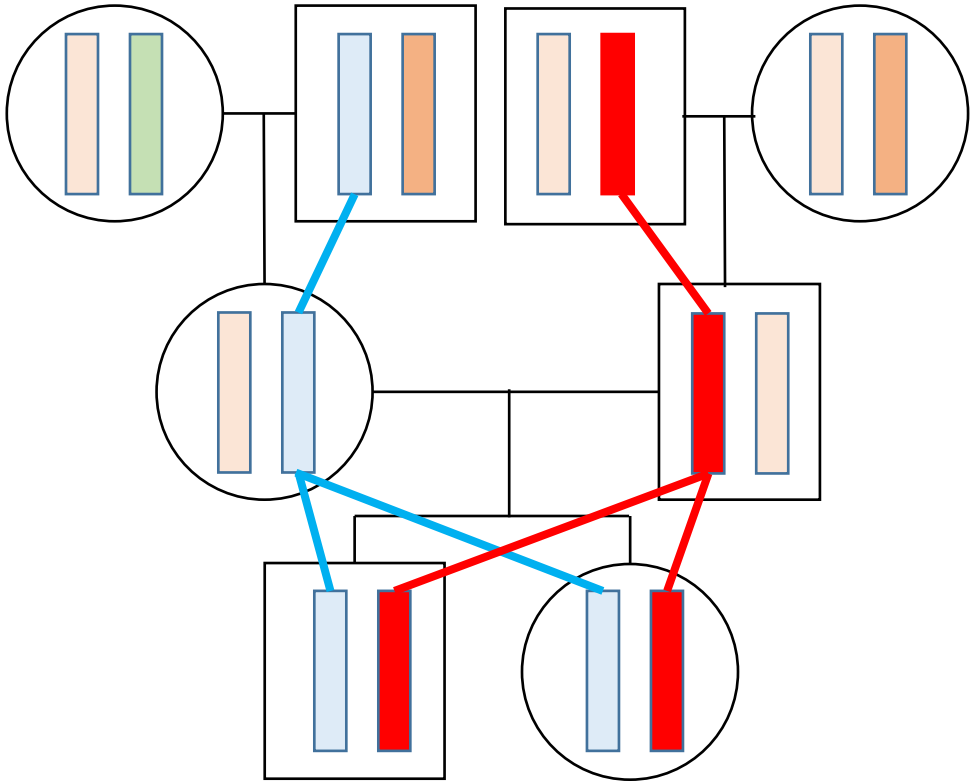


Chromosome 1

The same pedigree can lead to different gene trees across the genome



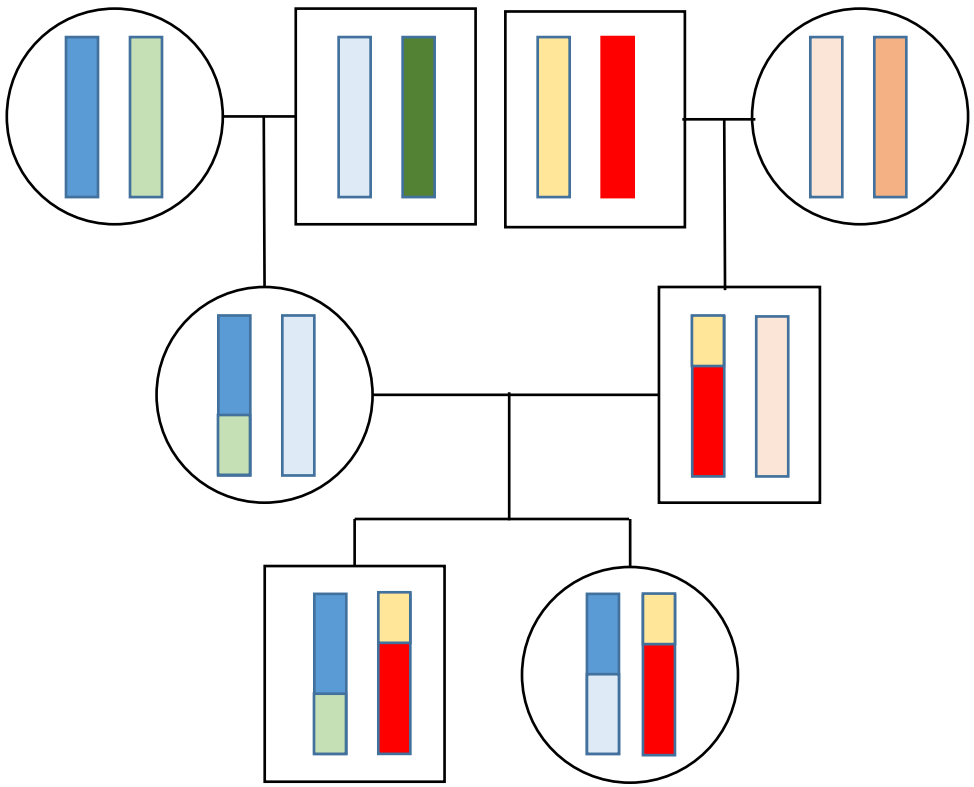
Gene tree of a sample of size n=4
(2 diploid individuals)



Chromosome 2

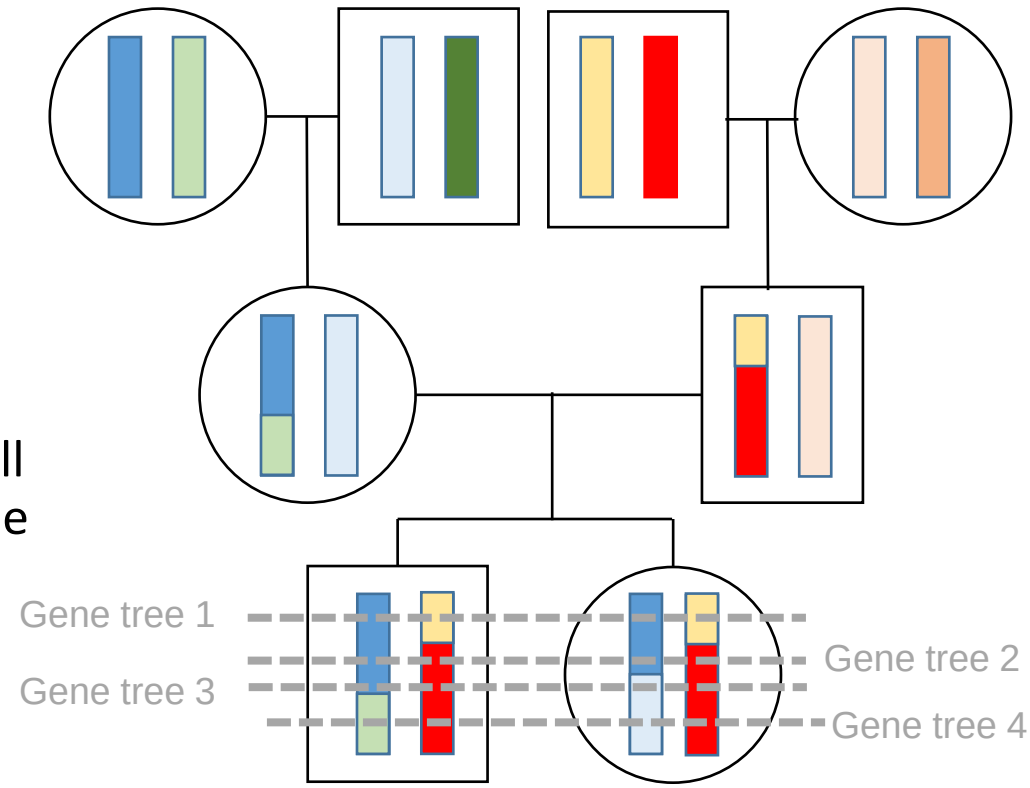
Gene trees and pedigrees

- Although we have the same pedigree, the gene trees at different loci will be different
- Ancestral chromosomes that did not contribute to our sample can be ignored
- With recombination, different regions of the chromosome will have different (correlated) gene trees



Gene trees and pedigrees

- Although we have the same pedigree, the gene trees at different loci will be different
- Ancestral chromosomes that did not contribute to our sample can be ignored
- With recombination, different regions of the chromosome will have different (correlated) gene trees



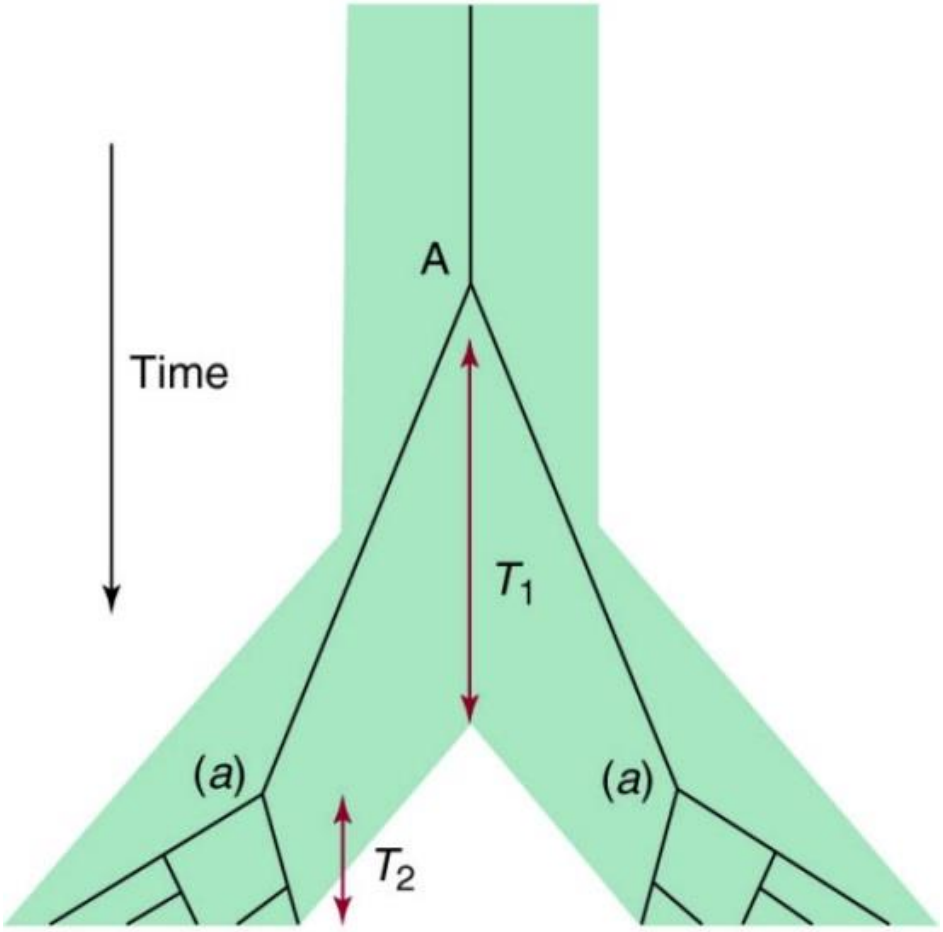
Gene trees vs. Population trees

Gene trees reflect the ancestral relationship of sampled gene copies.

The relationship between populations is given by the **population tree**. As with pedigrees, the population tree reflects the relationship between populations that is shared by all individuals.

In phylogenetics it is usually assumed that the gene tree reflects the population/species tree.

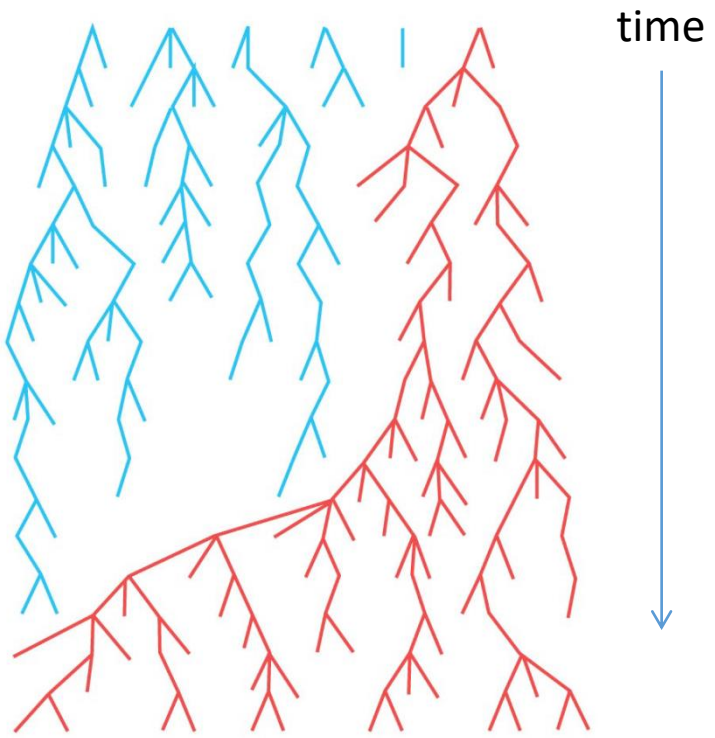
However, in the time scale of population genetics, gene trees at a particular region of the genome (locus) can be very different from the population tree.



Nichols (2001) TREE

Forward in time vs backward in time

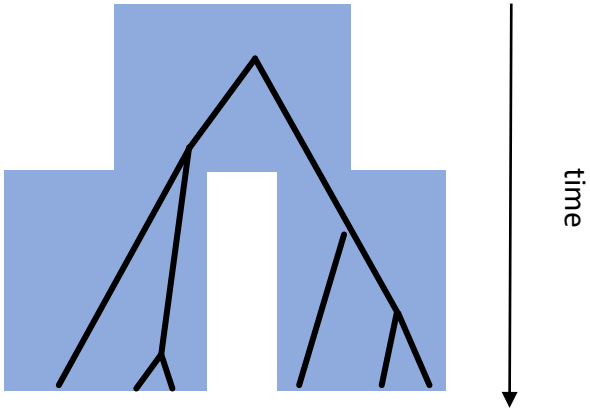
- Many results in population genetics come from studying how allele frequencies change forward in time
- But, when we look at samples from natural populations we want to know what happened in the past to explain what we see today
- **All sites in the genome have a gene tree describing the ancestral relationship of the lineages in a sample**



15.0, Genetic drift

Reconstructing the demographic history from genomic data

Because of recombination, different regions of the genome can have different gene trees

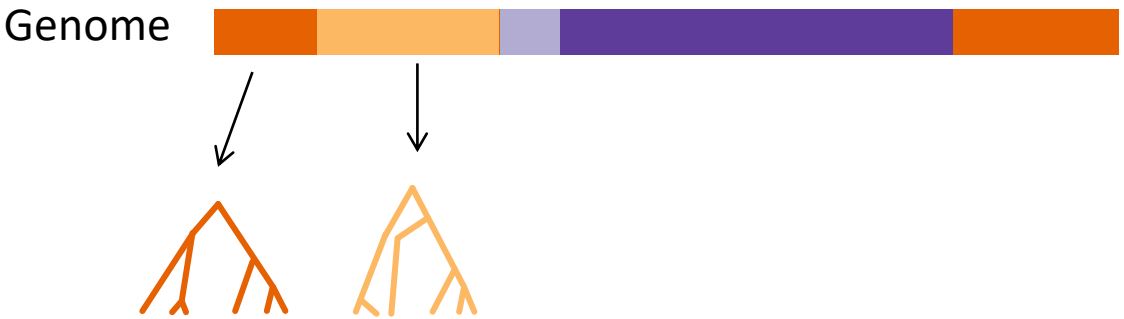


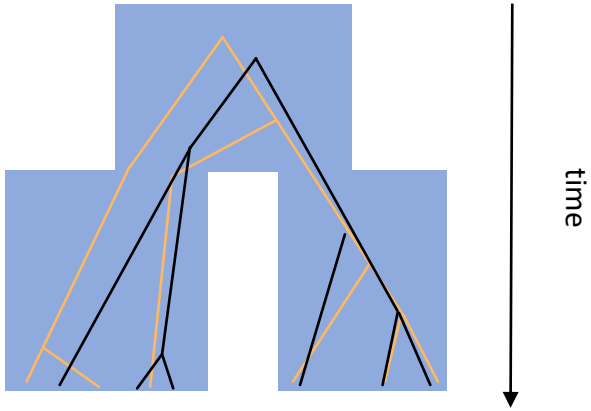
All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

Reconstructing the demographic history from genomic data

Because of recombination, different regions of the genome can have different gene trees



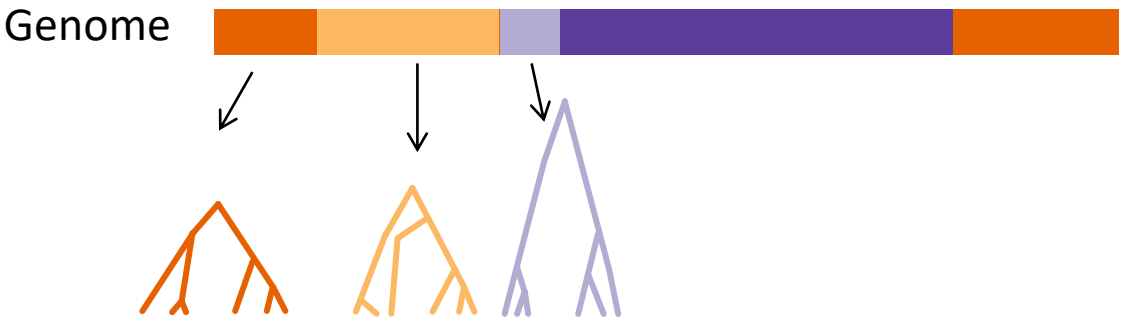


All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

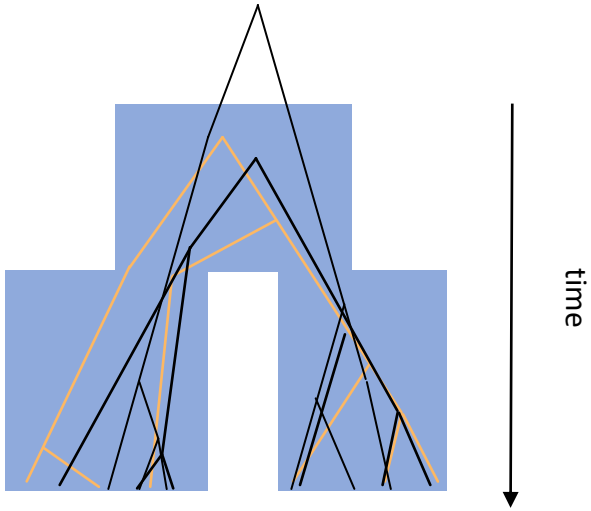
- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

Reconstructing the demographic history from genomic data

Because of recombination, different regions of the genome can have different gene trees



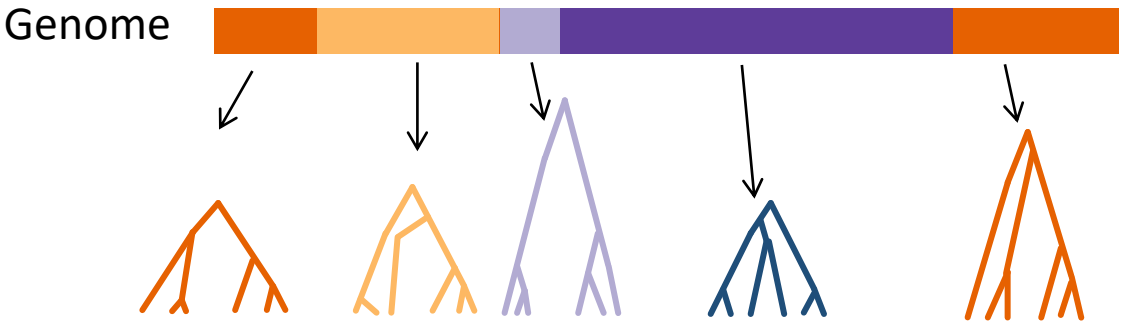
- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

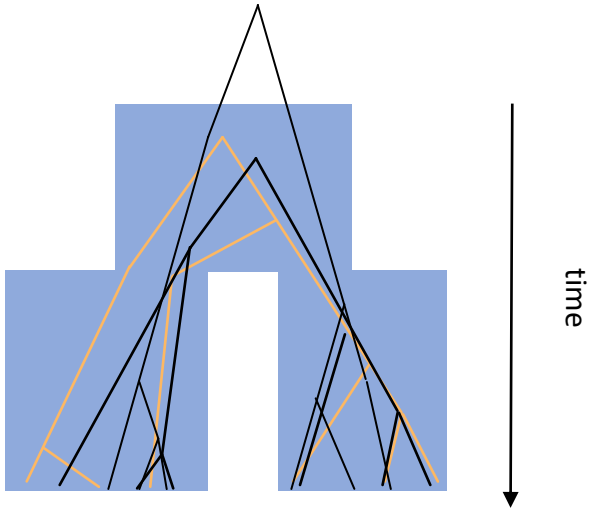


All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

Reconstructing the demographic history from genomic data

Because of recombination, different regions of the genome can have different gene trees





All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

Why do we care about demographic history?

Demography affects the efficiency of natural selection

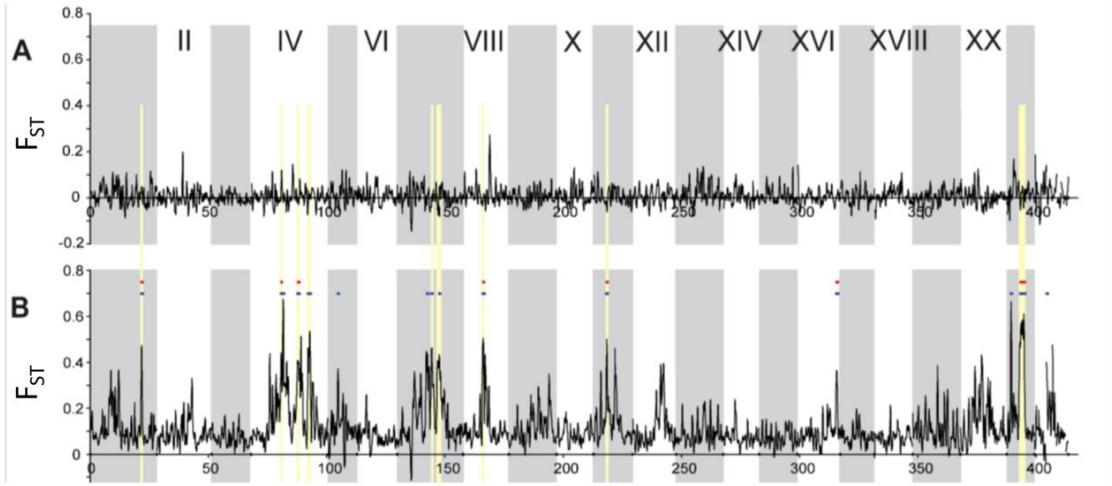
- Response to selection is different in small *vs* large populations, with *vs* without gene flow, etc.

Demographic history affects the genome-wide patterns

- It can be seen as a "null" model. Regions under selection are detected as outliers.

Between a pair of marine populations

Between marine and freshwater populations



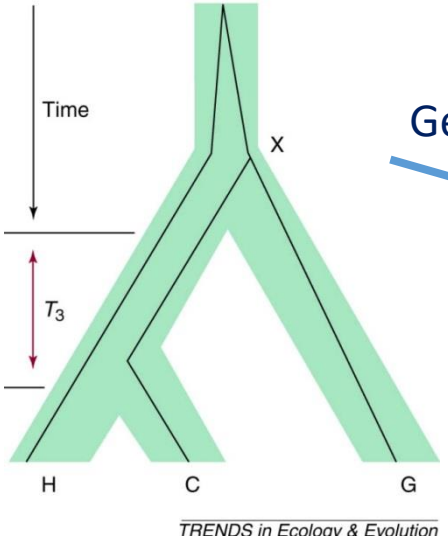
Chromosome position

Alternating white and grey indicate linkage groups (chromosomes)

Hohenlohe et al (2010) Plos Genetics

Intro to coalescent theory

- Coalescent theory provides a mathematical framework which describes the distribution of gene trees in populations
- However, while the tradition from phylogenetics is to estimate a tree and use the estimated tree to deduce evolutionary relationships, the population genetic tradition sees the tree as a random outcome of a population genetic process.



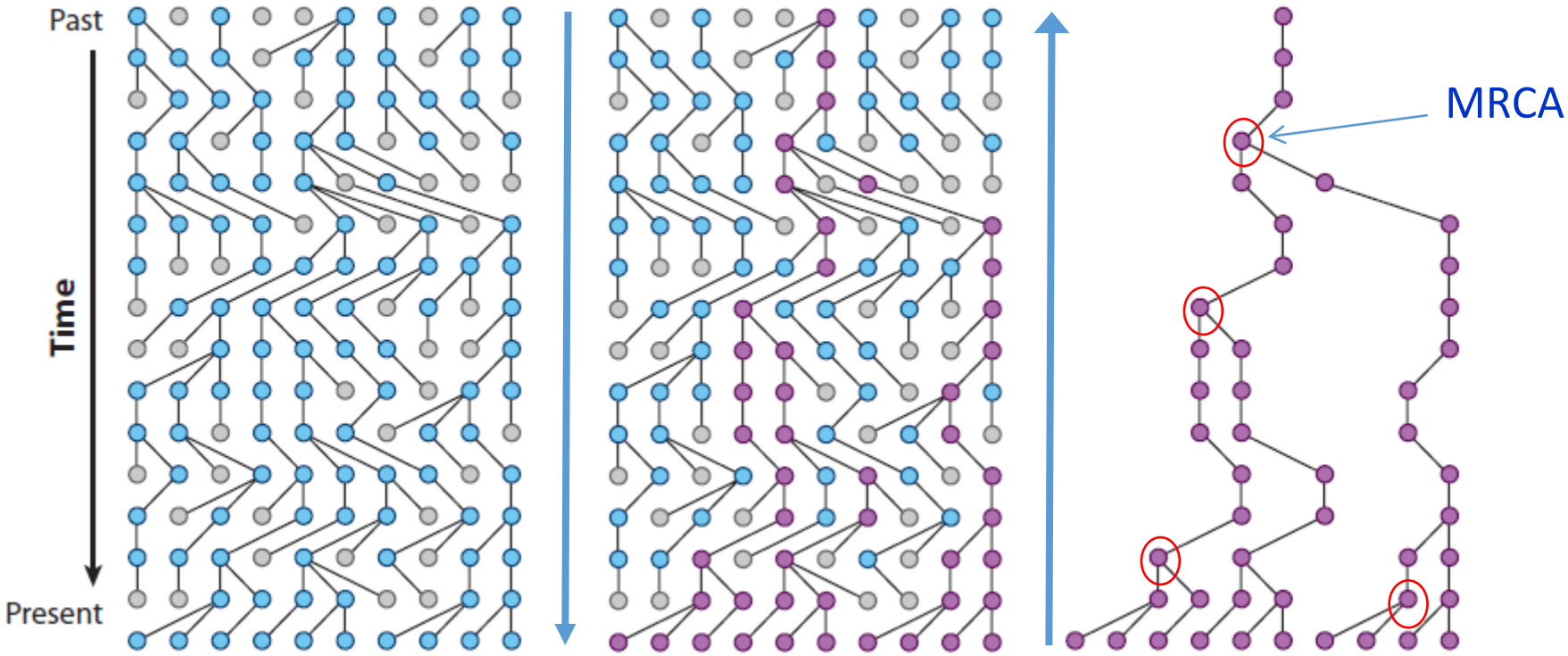
Nichols (2001) TREE

Genealogies within and among populations

Rosenberg and Nordborg (2002) Nat. Rev. Genetics

Coalescent backward process

Forward process Fixation process Coalescent process



MRCA: Most Recent Common Ancestor

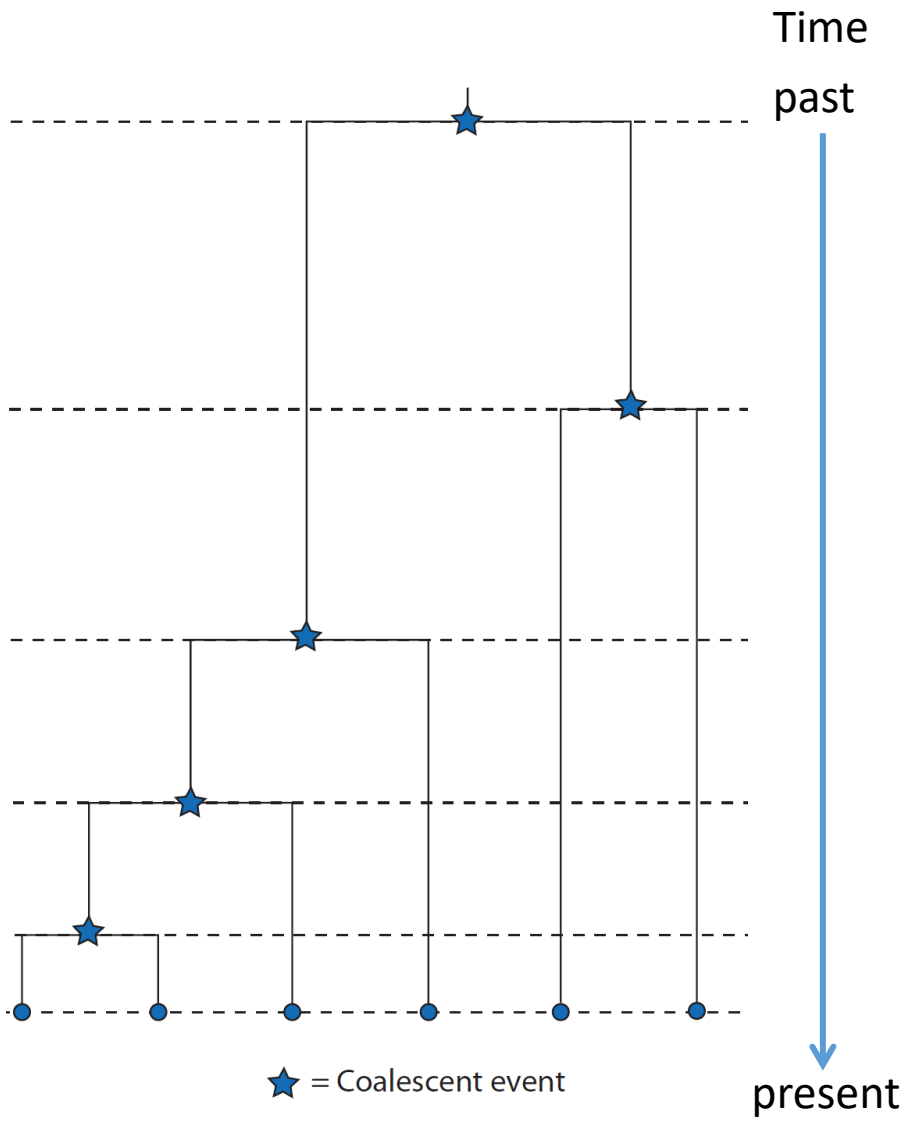
Legend:
○ Coalescent events

Grünwald and Goss. Annu. Rev. Phytopathol. 2011. 49:249–67

The principle of the coalescent theory

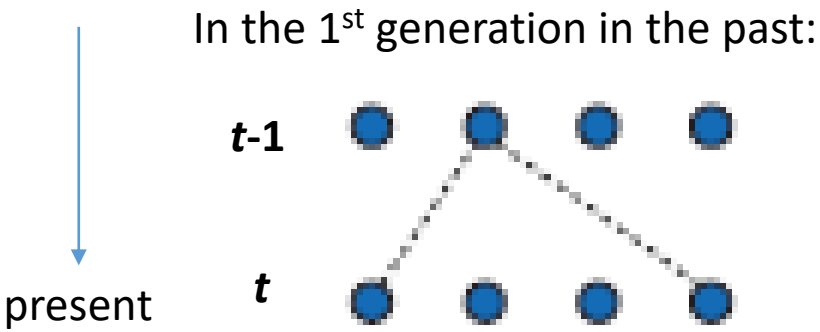
- For a given sample of individuals (gene copies or lineages) there is always a **gene tree** describing the ancestry of the sample.

- How many generations do we need to wait until they find an ancestor (i.e. a pair of lineages coalesce)?



Let's start simple: 2 lineages

In a population of N diploid individuals (2N gene copies)

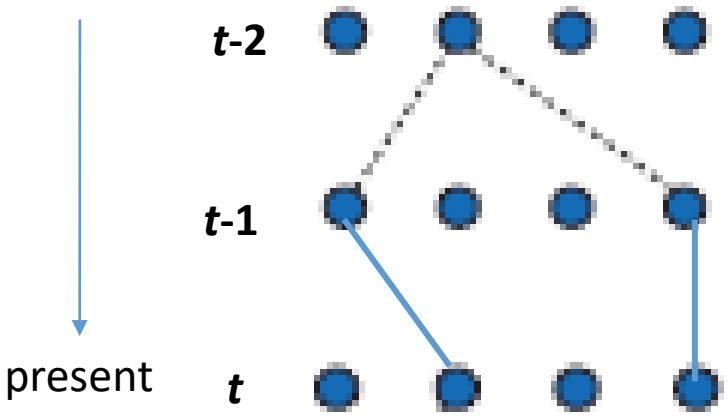


Pr(coal) = 1/2N
Pr(no coal) = 1-(1/2N)

Let's start simple: 2 lineages

In a population of N diploid individuals (2N gene copies)

In the 2nd generation in the past:



Pr(coal) = 1/2N
Pr(no coal) = 1-(1/2N)

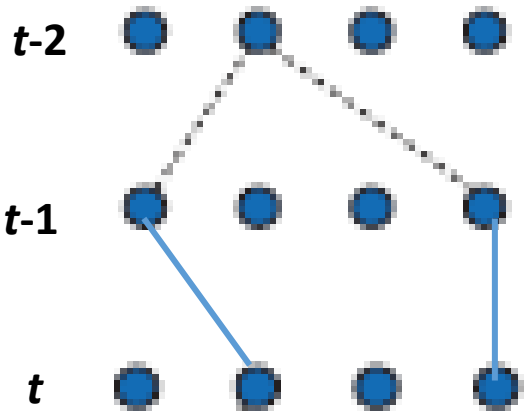
No coalescent
Pr(no coal) = 1-(1/2N)

In the t^{th} generation in the past the probabilities remain the same:

Pr(coal) = 1/2N
Pr(no coal) = 1-(1/2N)

Let's start simple: 2 lineages

In a population of N diploid individuals (2N gene copies)



Pr(coal) = 1/2N
Pr(no coal) = 1-(1/2N)

No coalescent
Pr(no coal) = 1-(1/2N)

Hence, the probability of coalescent at generation *t* follows a geometric distribution, with probability of success (1/2N):

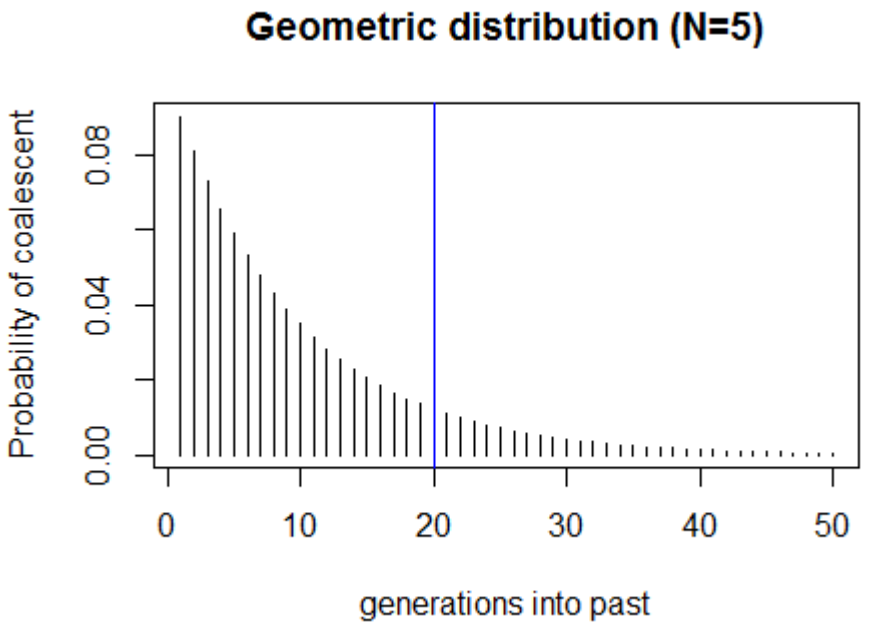
$$\Pr(\text{ coal at generation } t\ ) = \left(1-\frac{1}{2N}\right)^{t-1}\frac{1}{2N}$$

Expected time of coalescent for two lineages

Lets say T_2 is the random variable describing the time (in generations) until two lineages coalesce

$$\Pr(T_2 = t) = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}$$

T_2 follows a geometric distribution with probability of success p=(1/2N)



Expected time:

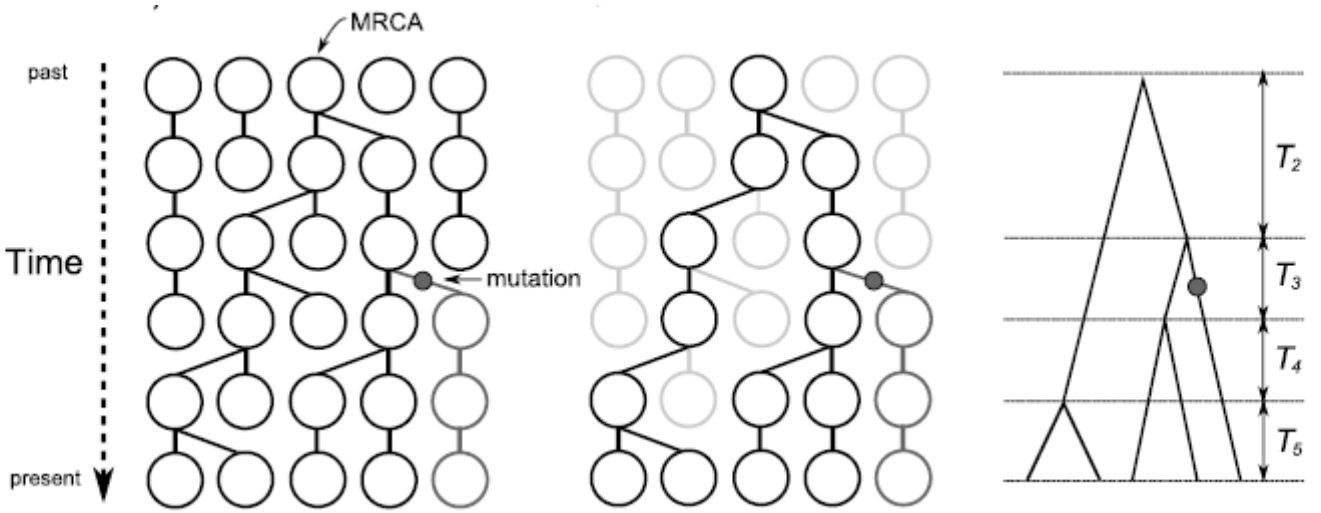
$$E[T_2] = (1/p) = 2N$$

(HUGE) Variance:

$$Var(T_2) = (1-p)/p^2 = 2N^2 - 2N$$

The expected time for the TMRCA of 2 lineages is 2N!

The coalescent process with **more than two lineages**



The coalescent process can thus be seen as a process going from *n* genes at the present time to a single gene sometimes in the past through a series of coalescent events. This last gene is said to be the **Most Recent Common Ancestor** (MRCA) of all genes of the sample.

$$n \to (n-1) \to (n-2) \to \cdots \to 2 \to 1\,(MRCA)$$

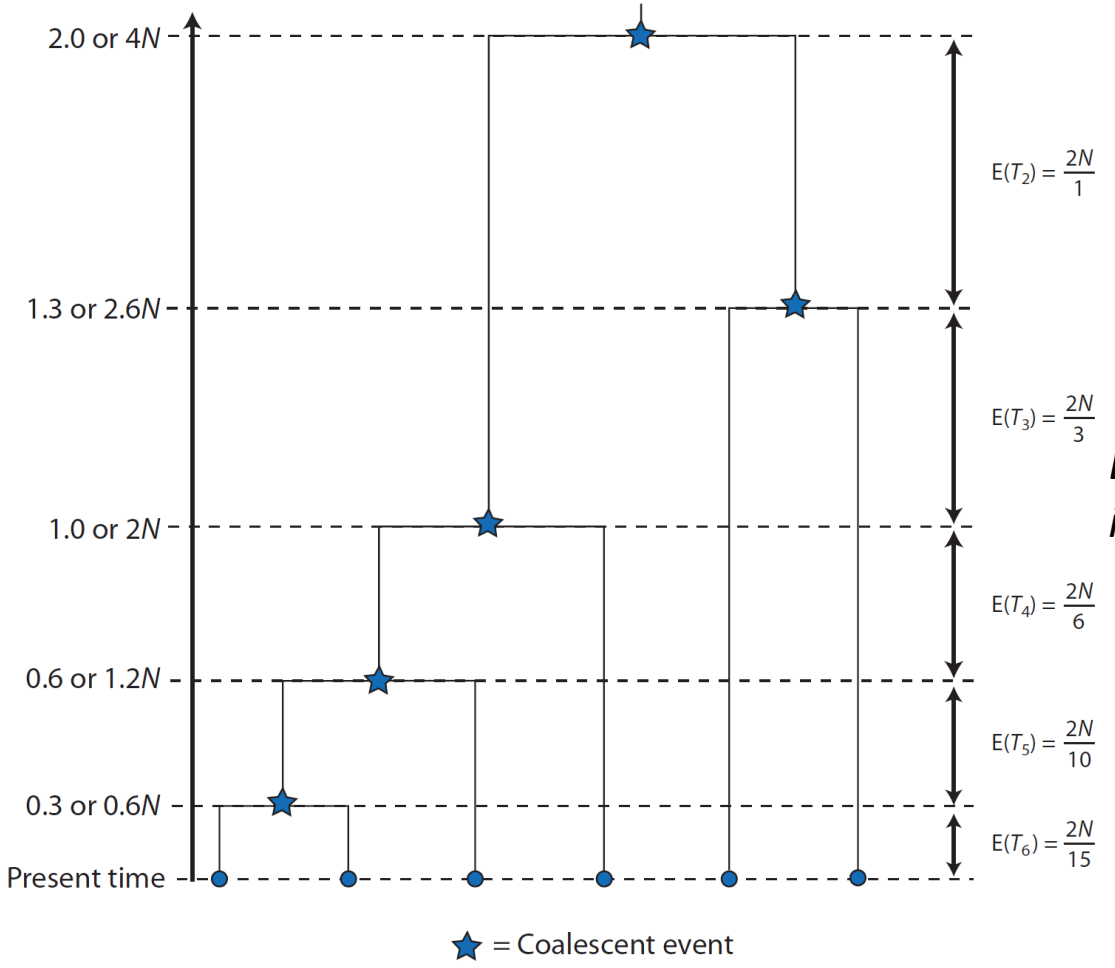
At each step the probability to pass from *n* gene lineages to *n*-1 lineages is $p_n = \binom{n}{2}\frac{1}{2N}$

Each generation, we have the same probability p_n of having a coalescent event. The time to the next coalescent event T_n can thus be considered as the **waiting time** until we have a coalescent event, which is therefore **geometrically distributed** with mean

$$E(T_j) = \frac{1}{p_n} = \frac{4N}{n(n\text{-}1)}$$

Expected coalescent times in a constant size population

For a sample of *n* lineages the expected TMRCA is 4N



Continuous time (in 2N units)

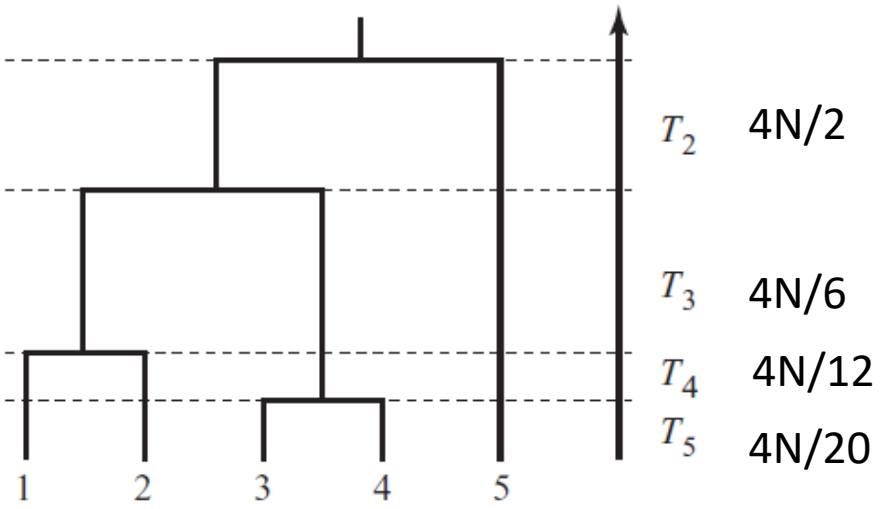
Discrete time in generations

- What are the longest branches we expect in a stationary population?
- Do we expect the relative branch length to differ in large and small populations?

Expected times in a coalescent tree

We can draw a coalescent tree with branch lengths proportional to their expected values as

At each coalescent event, pick two lineages at random to coalesce



Note that:

1) The coalescent process only gives us times between successive coalescent events, but not the topology. Assuming that all individuals are equally fit, **a pair of lineage is chosen at random for each coalescent event.**
2) Coalescent times increase with decreasing number of lineages: we expect long internal branch length and small external branch lengths
3) There is an enormous variance in the coalescent process

The expected time is 4N, but there is a large variance

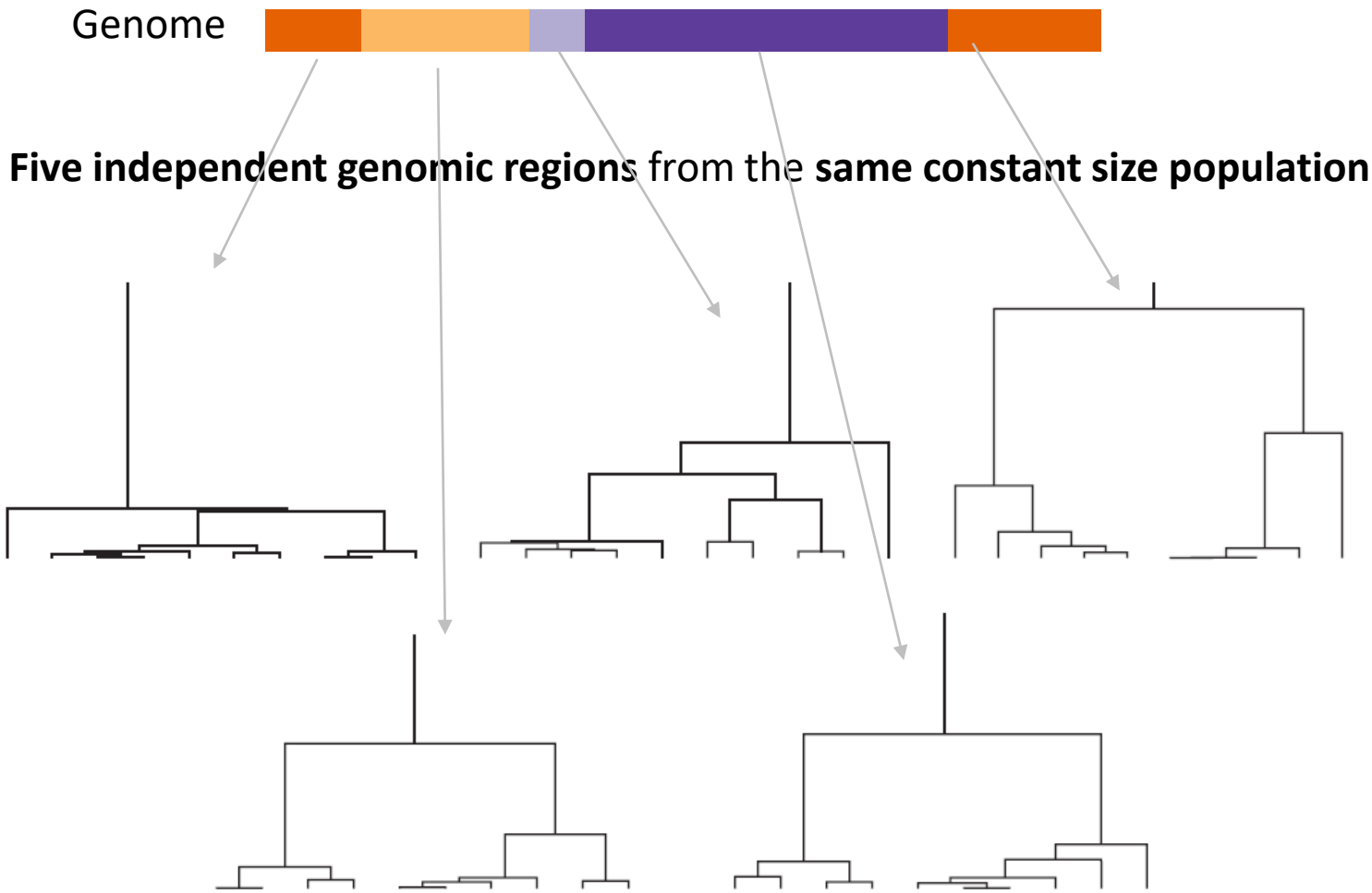
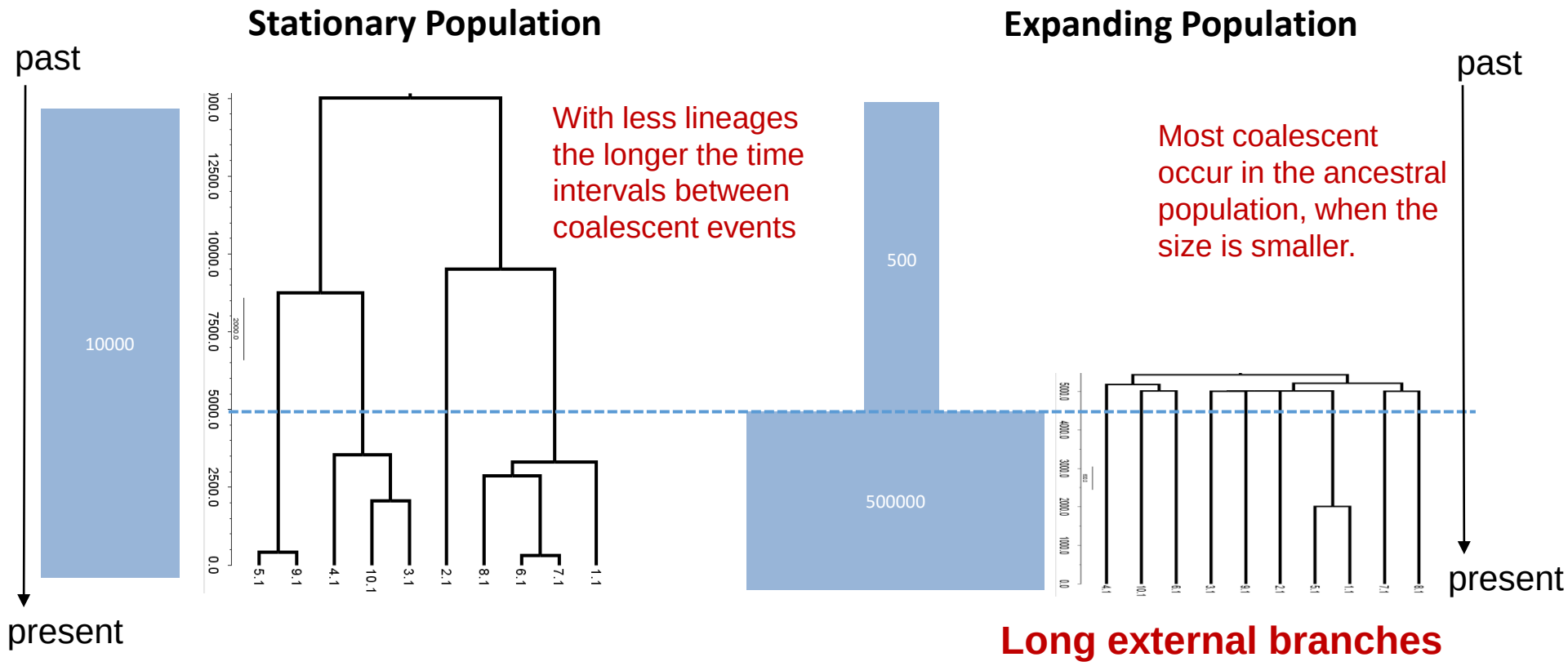


Figure 4.2 Five replicates of the coalescent process with constant population size for a sample of ten genes. Note the large variance in the time of the MRCA among replicates.

Hein et al (2004) Gene Genealogies, Variation and evolution

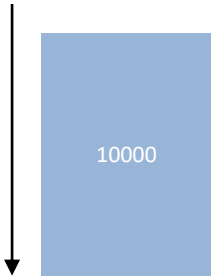
Gene trees in growing populations



- Coalescent rate is larger in smaller populations, and so we expect smaller intervals between coalescent events in smaller populations
- Coalescent rate is lower with a lower number of lineages, and so we expected larger intervals between coalescent events as the number of lineages decrease

Stationary population

gene trees at five genome regions
(all share same population history!)



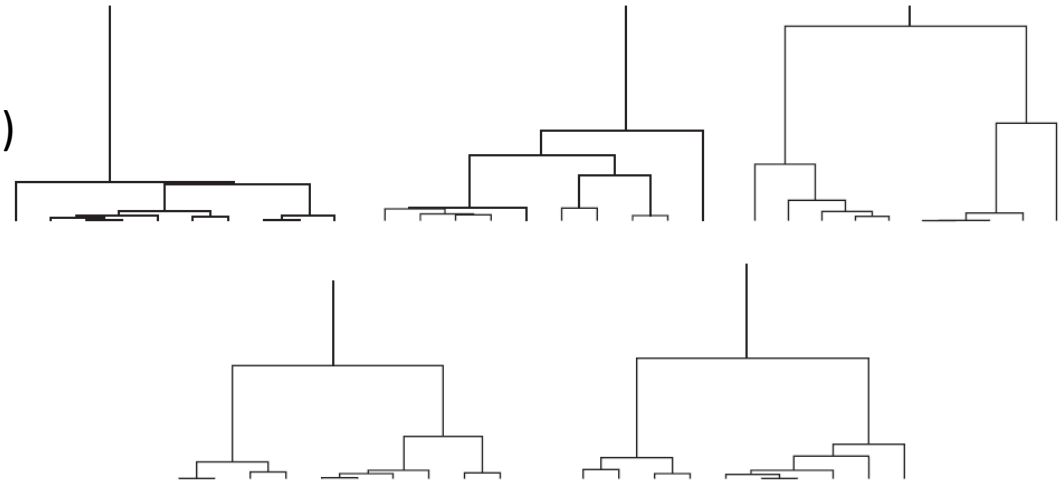
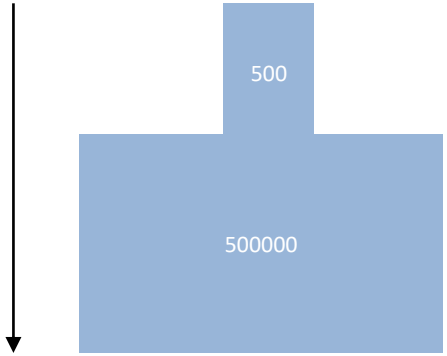


Figure 4.2 Five replicates of the coalescent process with constant population size for a sample of ten genes. Note the large variance in the time of the MRCA among replicates.

Expanding population

gene trees at five genome regions
(all share same population history!)



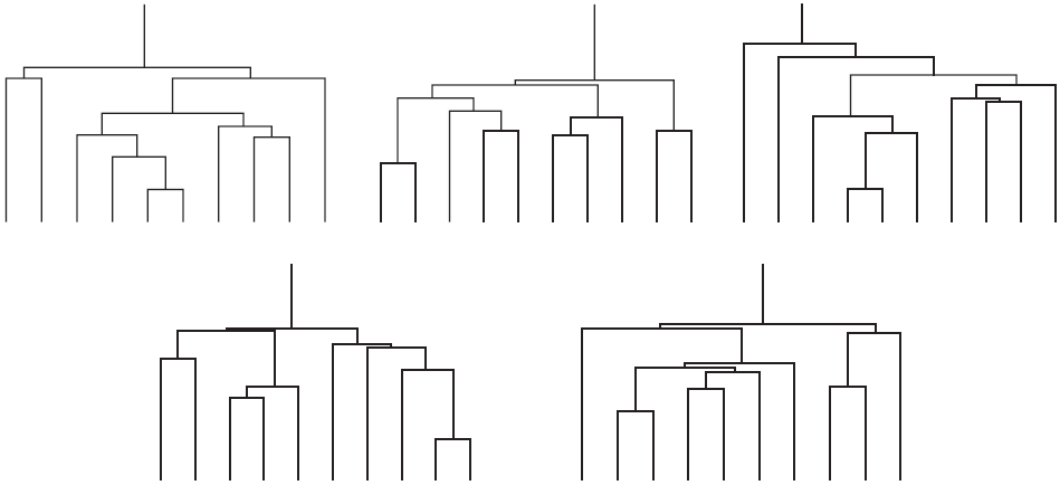
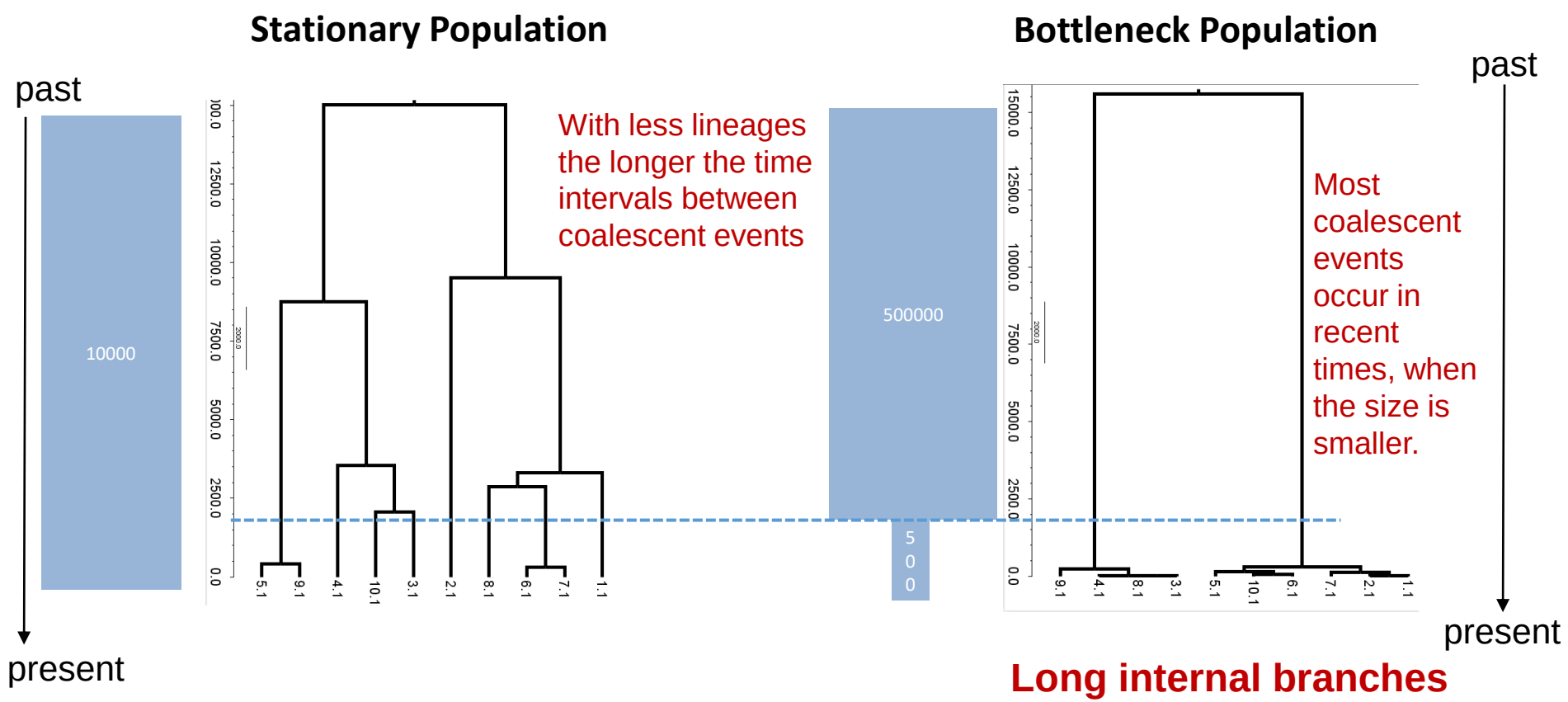


Figure 4.3 Five replicates of the coalescent with exponential growth, $\beta = 1000$, for a sample of $n = 10$ genes. Note the smaller variance in the time until the MRCA compared to the same quantity in Figure 4.2.

Hein et al (2004) Gene Genealogies, Variation and evolution

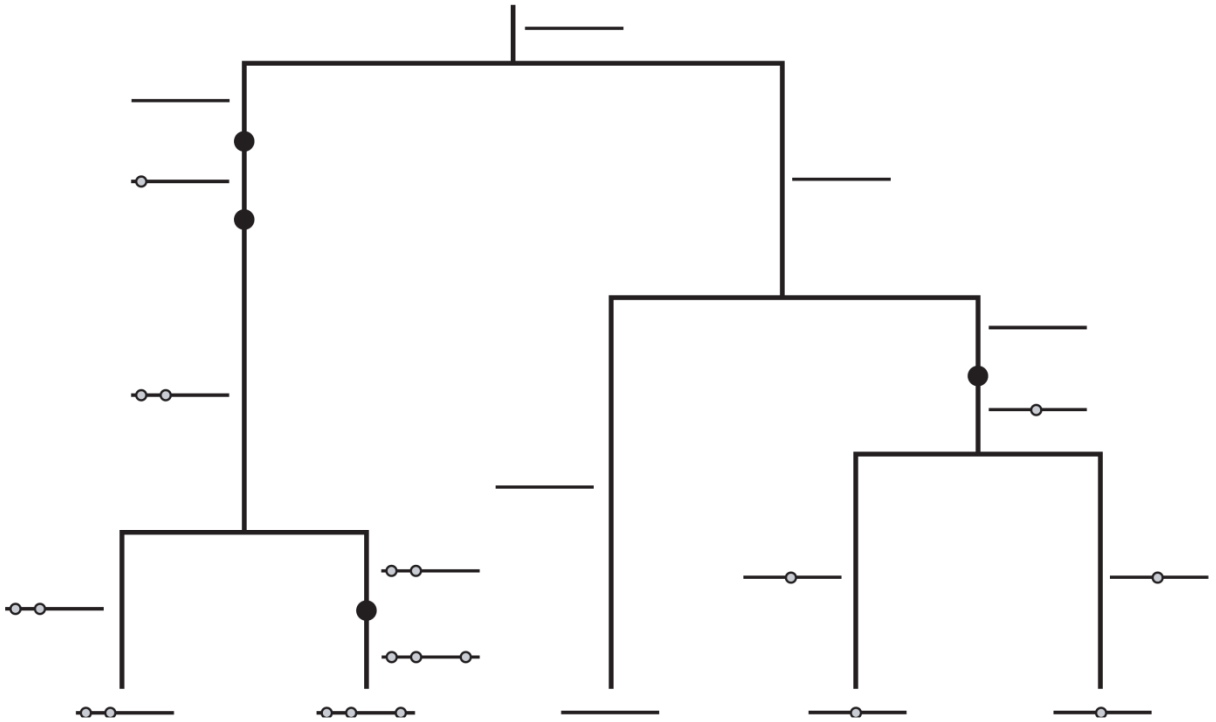
Gene trees for decreasing populations



- If we could observed directly the gene trees, we could easily reconstruct the population tree and the demographic history.
- But we do not observe gene trees…
- We can still learn about gene trees from the observed mutations and the allele frequencies in samples

Adding neutral mutations

The shape of neutral coalescent trees only depend on the population demography, and not on the mutational process. Assuming that all alleles have the same fitness, the mutational process can be modeled as an independent process superimposed on a realized coalescent tree.

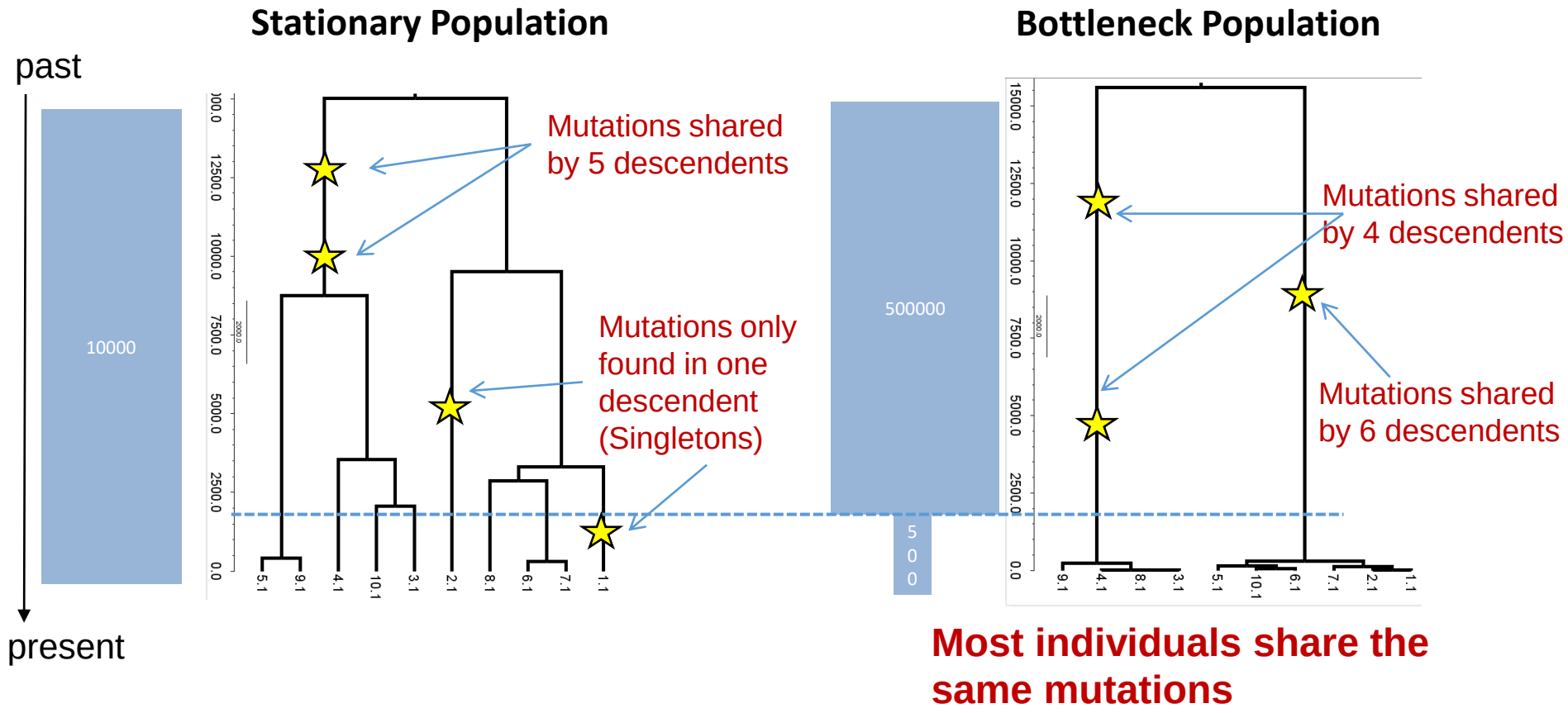


Mutations just accumulate along the branches of the tree according to a **Poisson process** with rate $\lambda_i = \mu t_i$ for the *i*-th branch of length t_i. The Poisson process is stochastic but it should be immediately **obvious that long branches will carry more mutations than short branches**

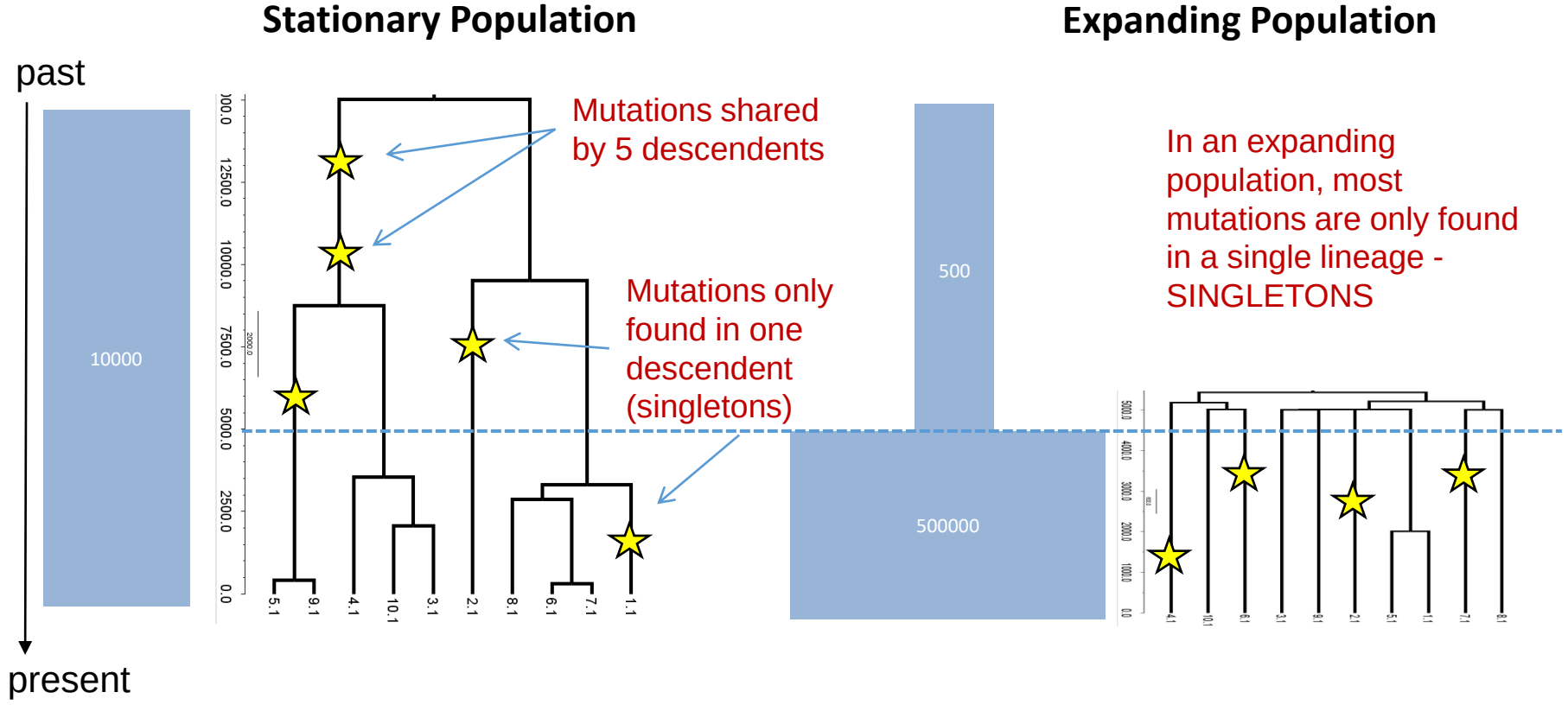
Hein et al (2004) Gene Genealogies, Variation and evolution

We expect less diversity in a bottlenecked population

- Mutations accumulate along the branches.
- The longer a given branch the more likely it becomes that a mutation have happened on it.

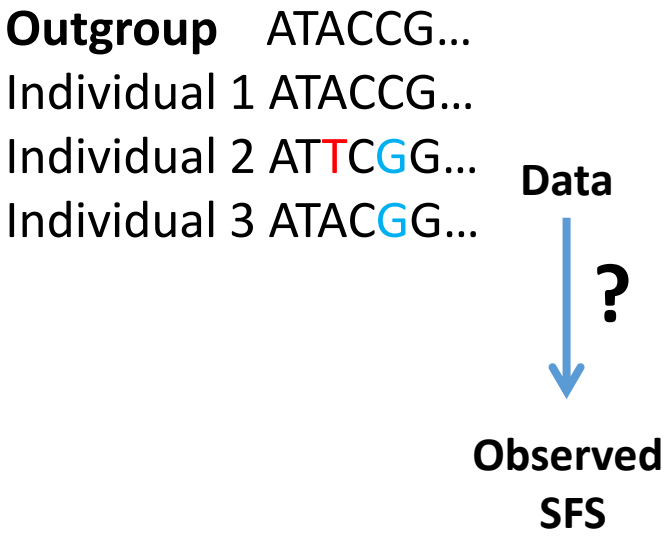


We expect less diversity in a bottlenecked population



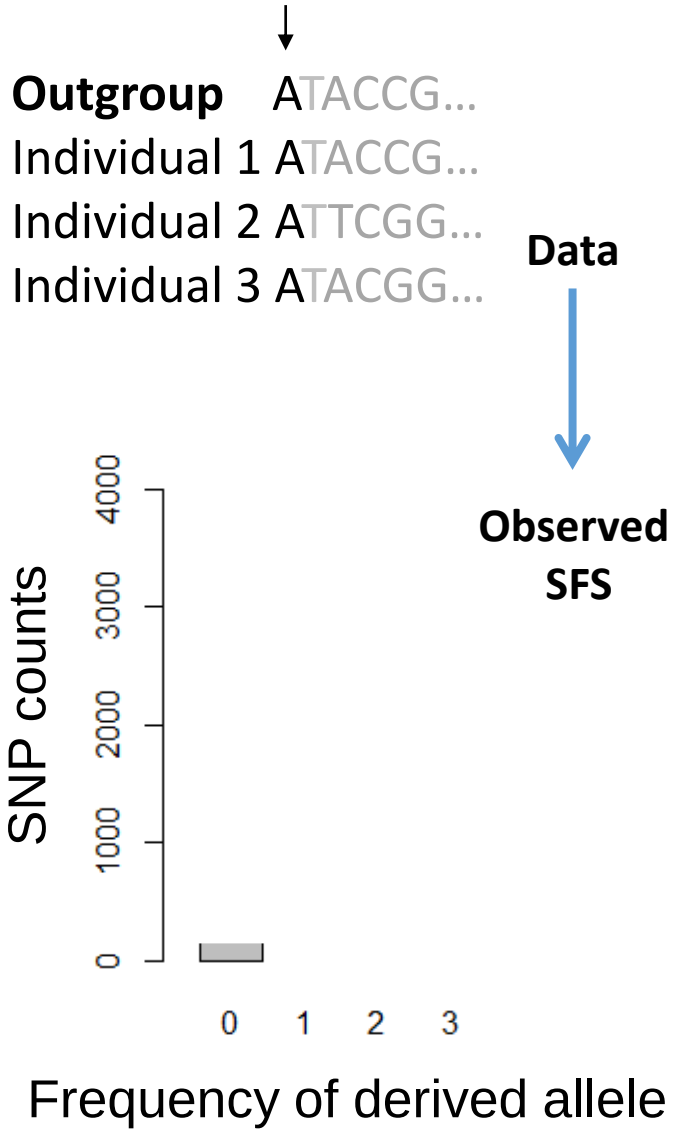
Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- Assuming a single population – 1Dimensional SFS



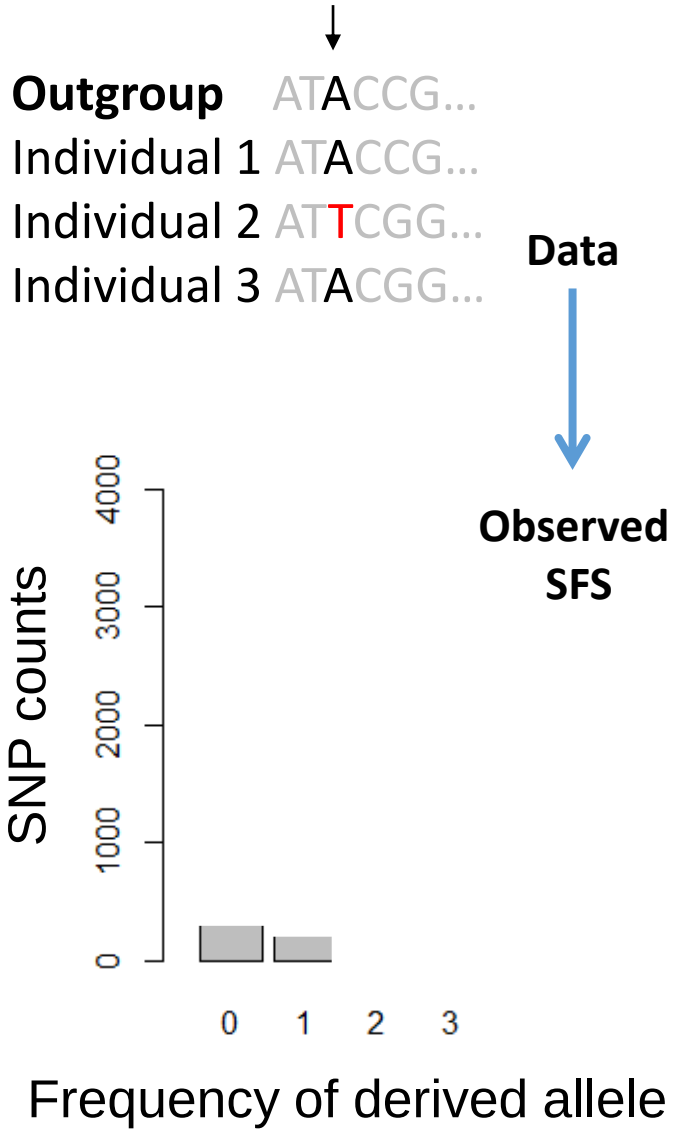
Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- Assuming a single population – 1Dimensional SFS



Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- Assuming a single population – 1Dimensional SFS



Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data

- Assuming a single population – 1Dimensional SFS

The SFS ignores information about linkage. It is best suited for the study of many unlinked (or recombining) DNA sequences.

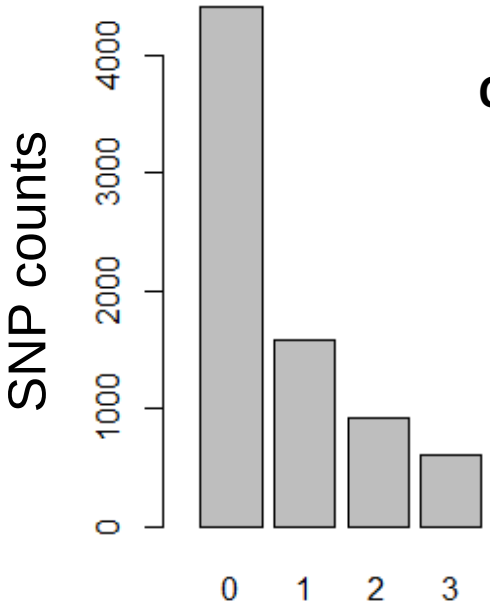
In a stationary population, the expected SFS relative frequencies are given by:

$$E(\xi_i) = \frac{\theta}{i}$$

Fu and Li, 1993

Outgroup ATACCG…
Individual 1 ATACCG…
Individual 2 ATTCGG…
Individual 3 ATACGG…

Data



Observed SFS

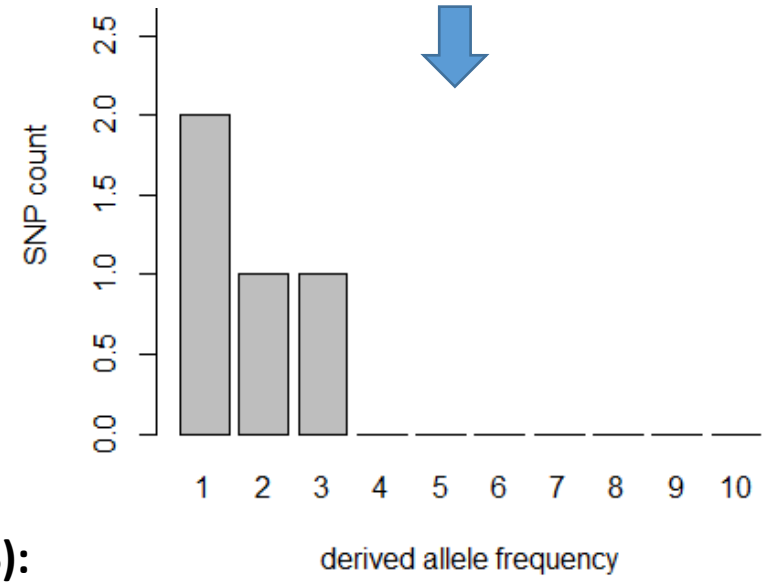
We can obtain the SFS from genotype call data

Genotypes:

- 0 homozygote for reference allele
- 1 heterozygote
- 2 homozygote for alternative allele

This can be done if we have enough depth of coverage (>10x)

	SNP1	SNP2	SNP3	SNP4
Individual 1	0	2	0	1
Individual 2	0	0	1	0
Individual 3	1	0	0	0
Individual 4	0	1	0	0
Individual 5	0	0	1	0



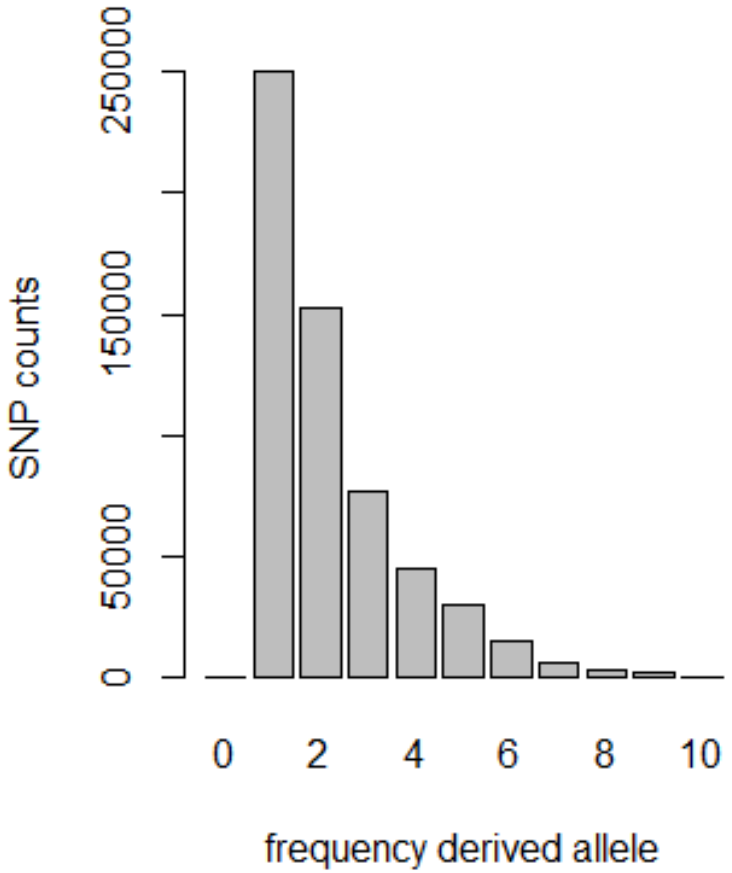
Observed SFS is a vector (1 dimensional SFS):

Frequency	0	1	2	3	4	5	6	7	8	9	10
SNP count	0	2	1	1	0	0	0	0	0	0	0

SFS from genotype call data

Even if we have millions of SNPs we can summarize the genomic data to 10 numbers with the SFS!

The size of the SFS depends on the number of sampled individuals.

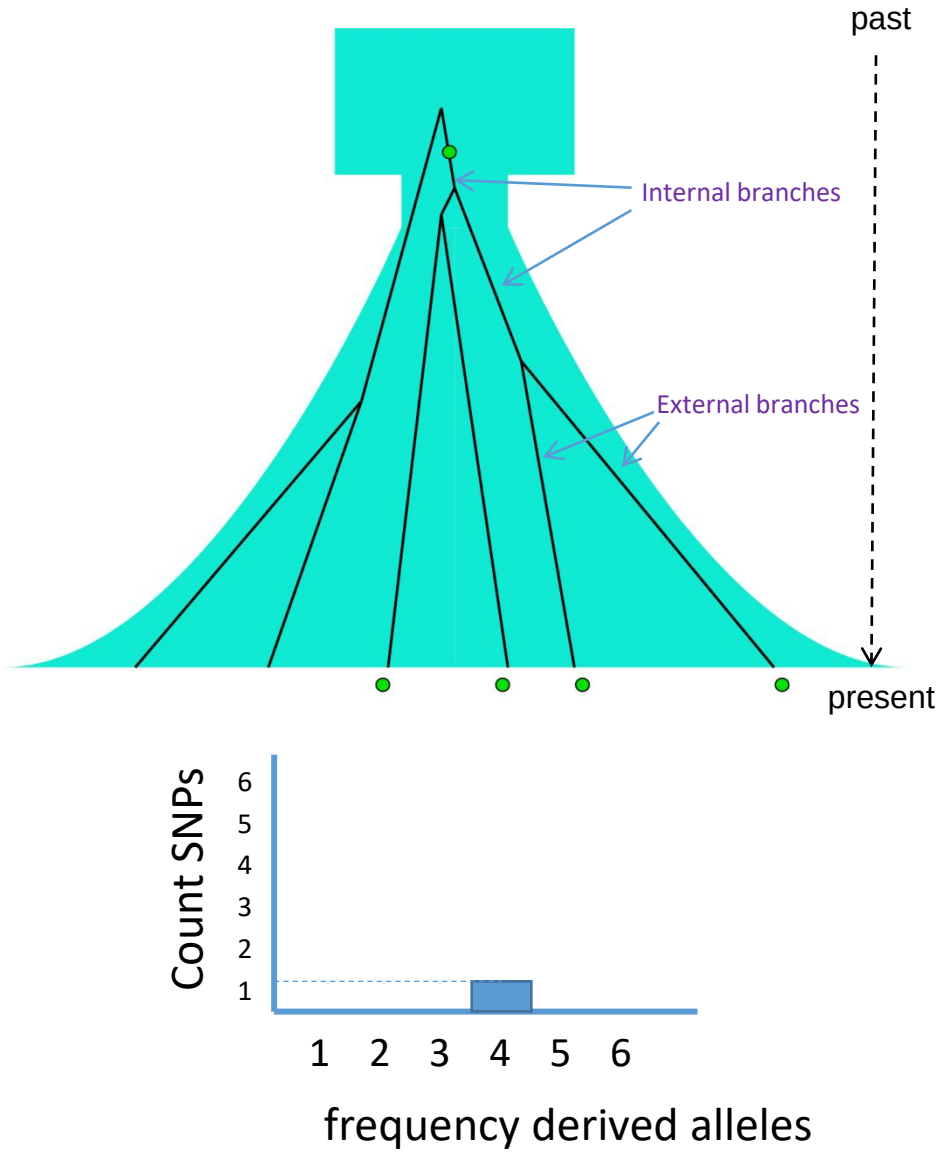


Observed SFS is a vector (1 dimensional SFS):

Frequency	0	1	2	3	4	5	6	7	8	9	10
SNP count	0	250,032	152,300	76,504	45,362	30,210	15,329	5,642	3,524	2,123	0

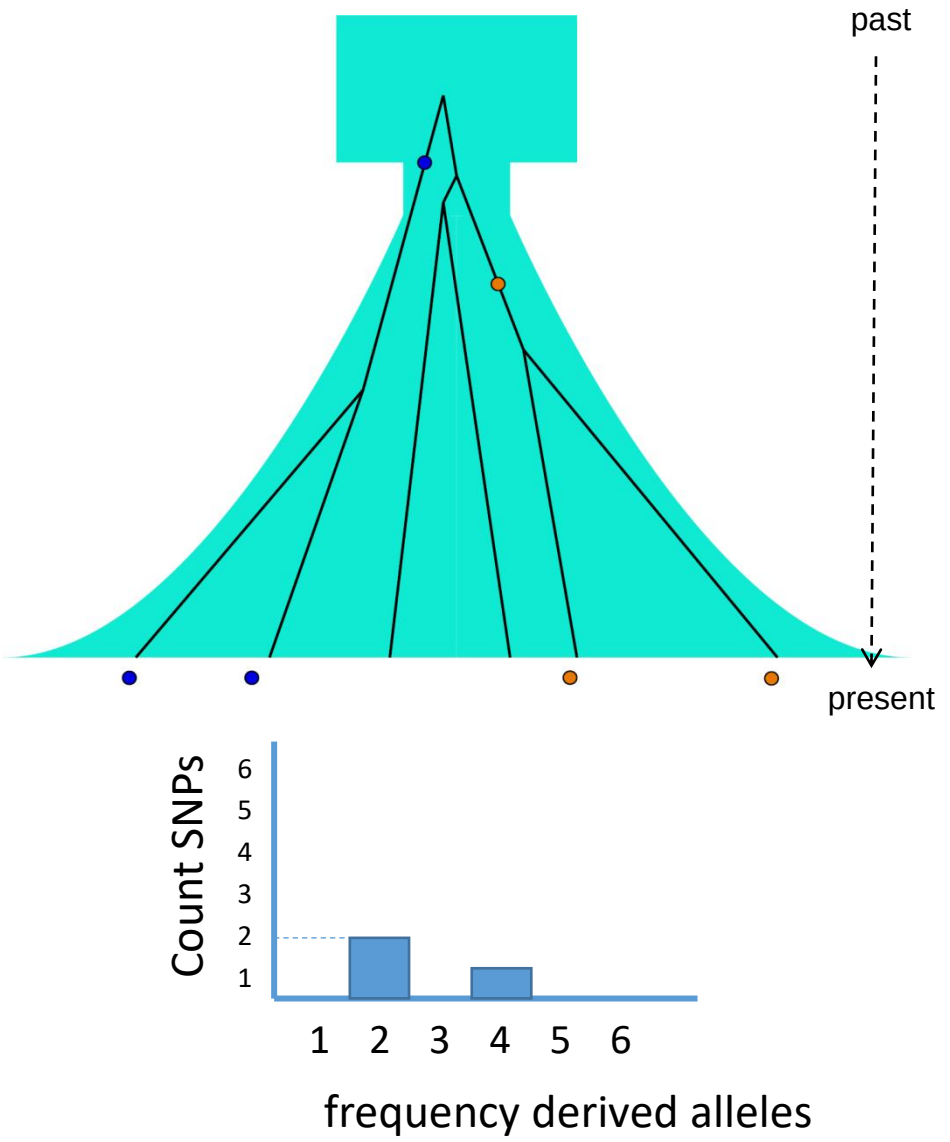
Coalescent and the SFS

- A recent population growth following a bottleneck leads to gene trees with long external branches
- Very few mutations in the internal branches
- Most mutations in long external branches are only found in one lineage, resulting in an excess of singletons



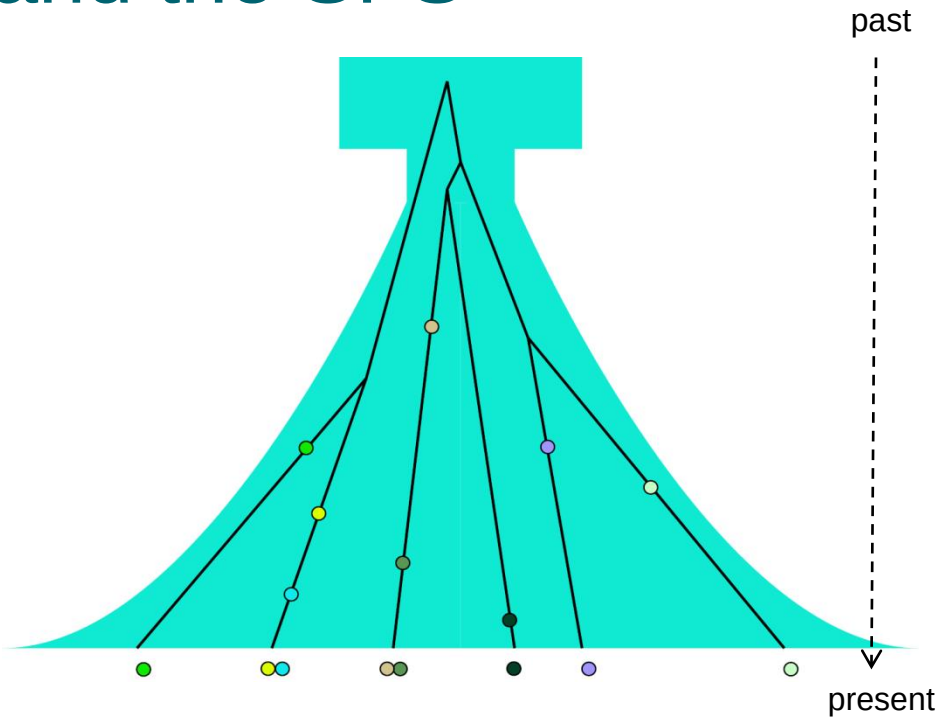
Coalescent and the SFS

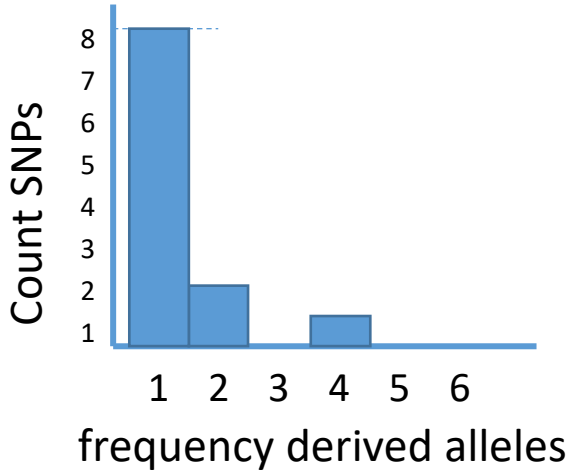
- A recent population growth following a bottleneck leads to gene trees with long external branches
- Very few mutations in the internal branches
- Most mutations in long external branches are only found in one lineage, resulting in an excess of singletons



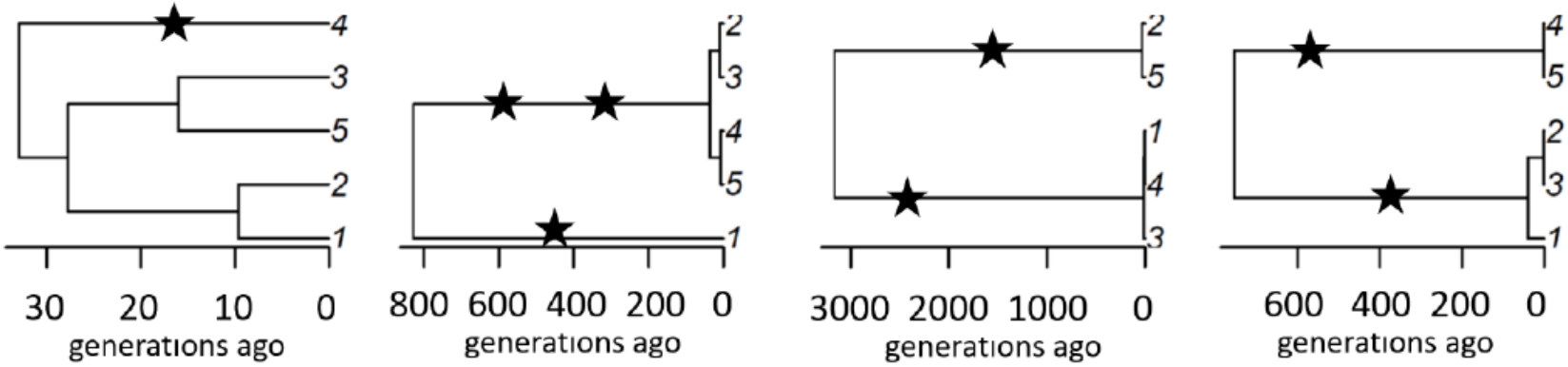
Coalescent and the SFS

- A recent population growth following a bottleneck leads to gene trees with long external branches
- Very few mutations in the internal branches
- Most mutations in long external branches are only found in one lineage, resulting in an excess of singletons

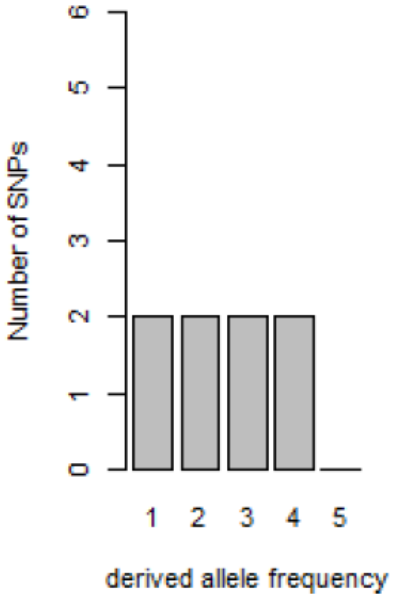




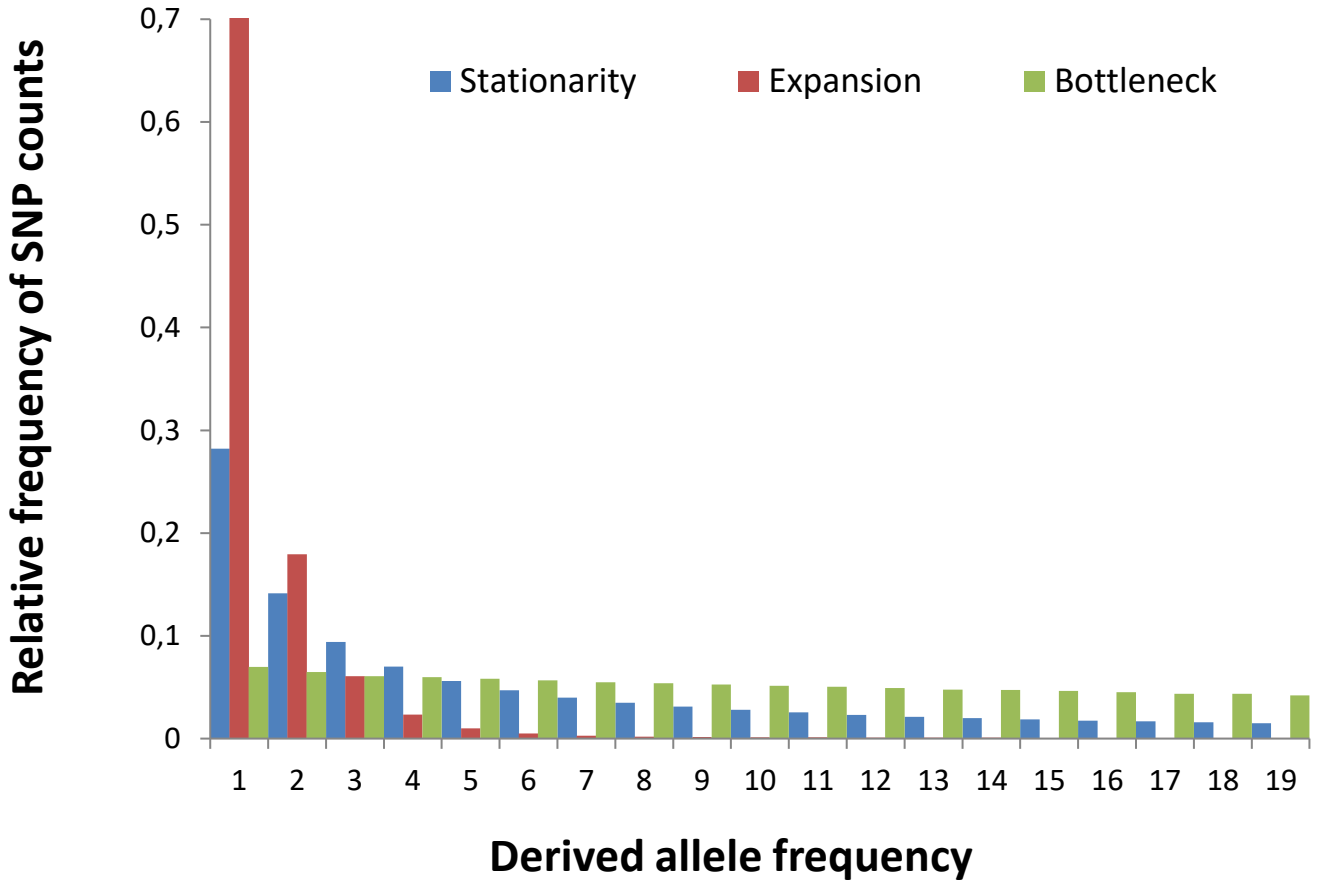
Coalescent gene trees at multiple independent sites and the SFS



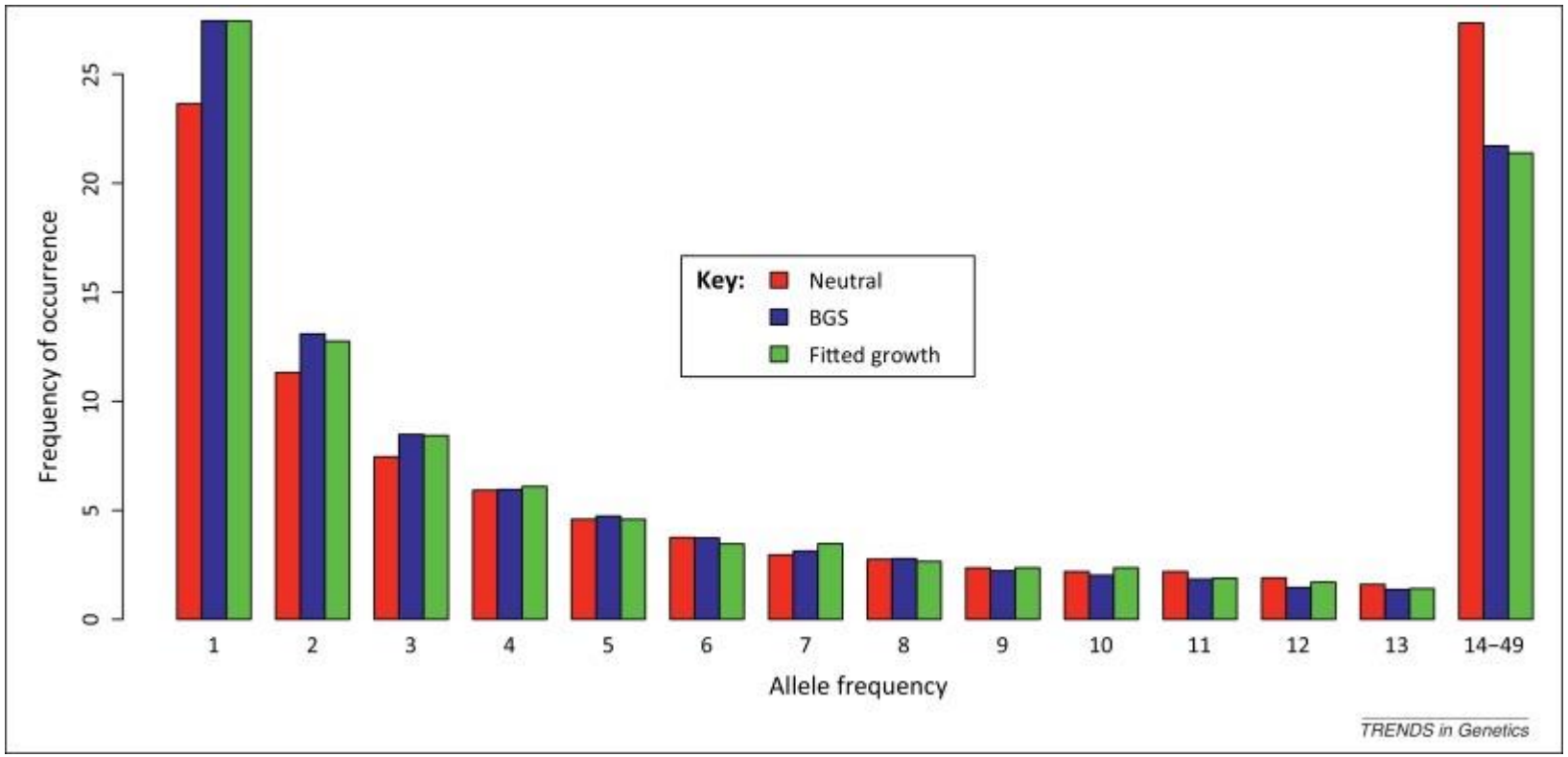
- The SFS is summed across loci
- Independent loci can have different gene trees, and different mutations and allele frequencies
- But, assuming neutrality, all sites in the genome reflect the population tree and the demographic history
- What can we say about the demographic events that lead to this SFS?
  - Bottleneck, expansion, constant size population?
  - Time of event?



SFS depends on past demography



Natural selection also affects the SFS



Background selection (BGS) leads to patterns similar to population expansion.

Bank et al (2014) Trends in Genetics

Population structure

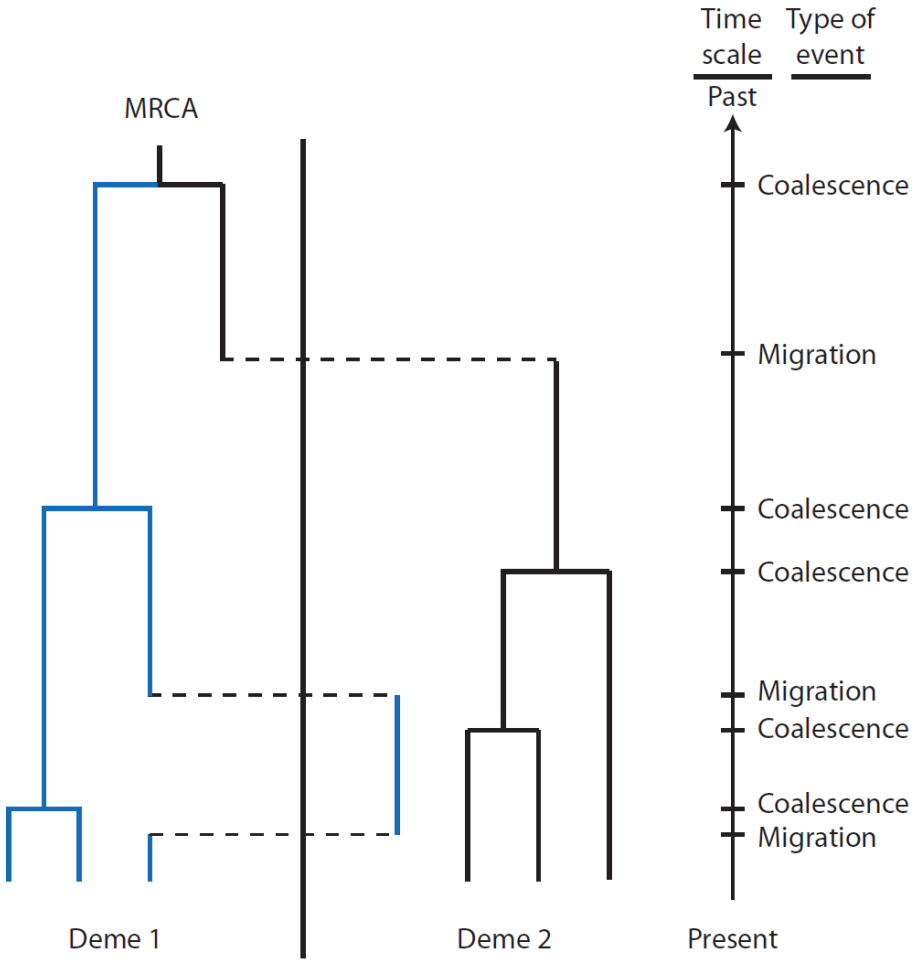
Migration events can be incorporated into gene trees.

Migration from Pop 2 to Pop 1, leads to lineages moving from Pop 1 to pop 2 backward in time.

At each generation, the probability of immigration into population 1 from population 2 is given by:

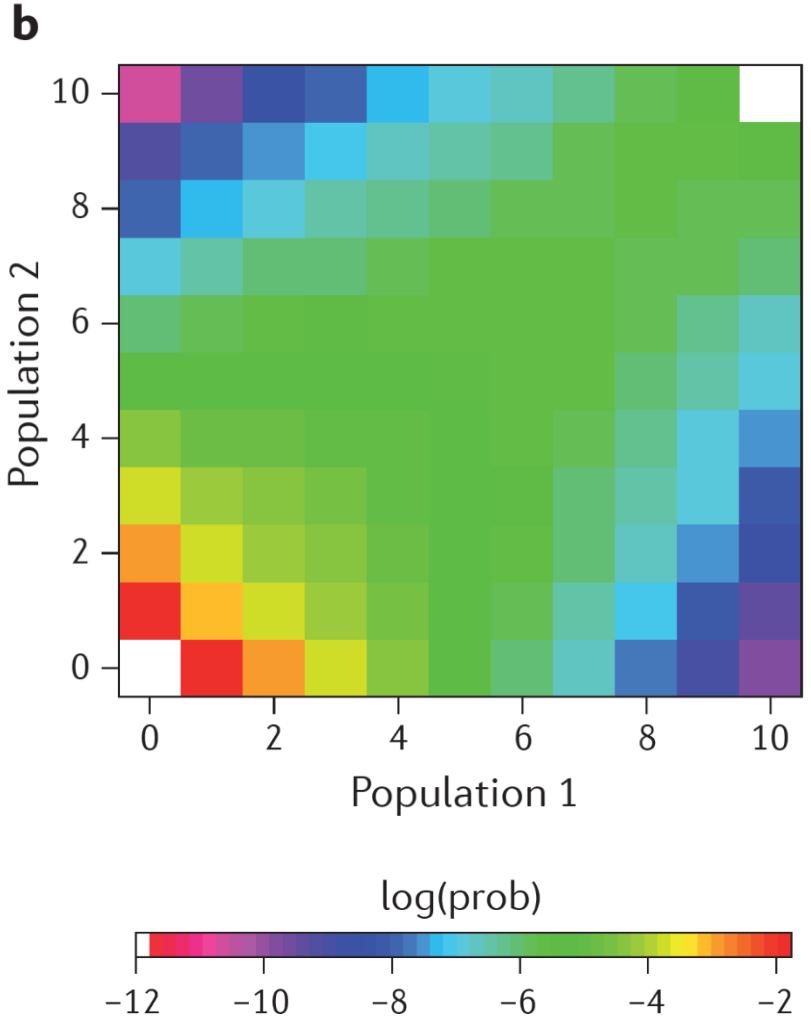
Pr(migrate) = $n_1 * m$

Where n_1 is the number of lineages in population 1, and m is the immigration rate.



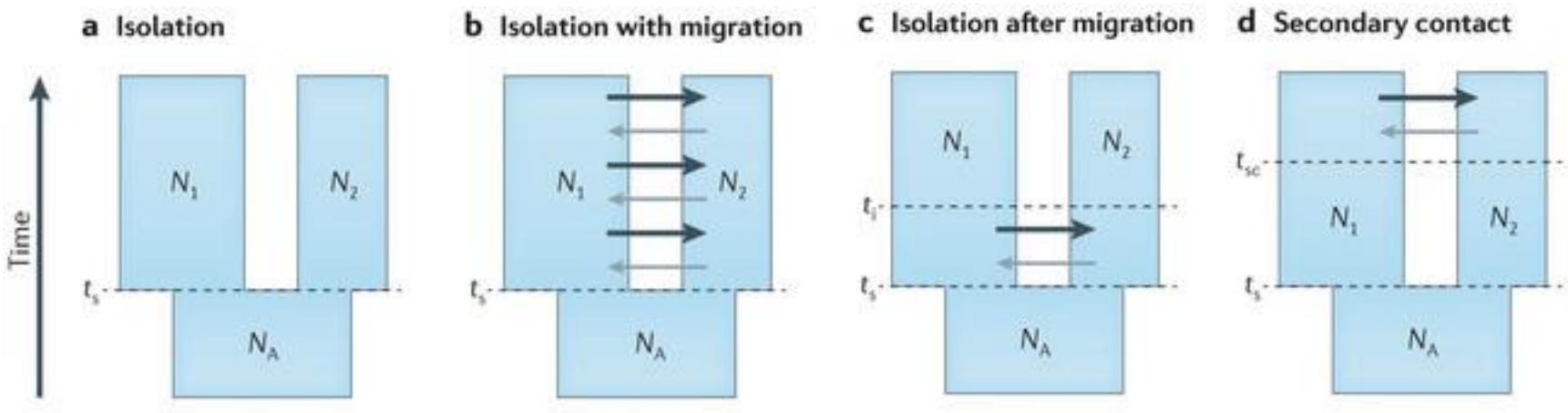
Site frequency spectrum from multiple populations (joint SFS)

- For a pair of populations – 2D SFS
  - Count the SNPs have a frequency of the derived allele of *i* in population 1, and of *j* in population 2
- We can extend this to 3D SFS, 4D SFS, etc.



Model based inference

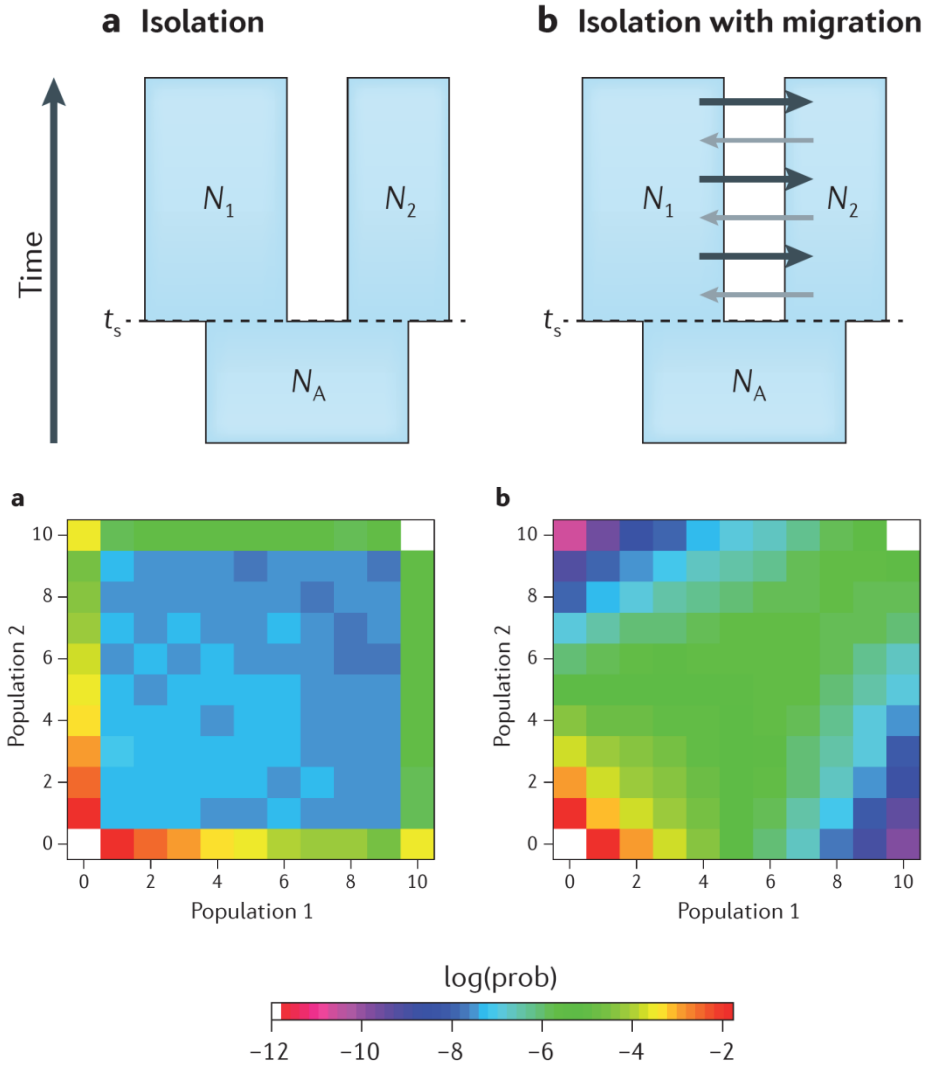
- What is the model that best fits the data?
- What are the most likely parameters of each model?



Sousa and Hey (2013) Nat. Rev. Gen.

Site frequency spectrum (SFS)

The SFS contains information about the demographic history of populations

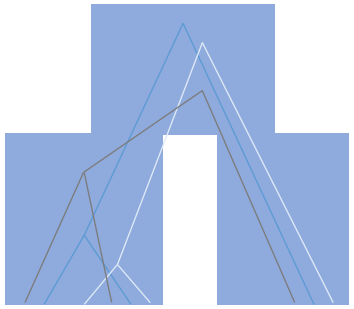


Sousa and Hey (2013) Nat. Rev. Genet.

Inferring the demographic history from the SFS

Genomic Data

Model



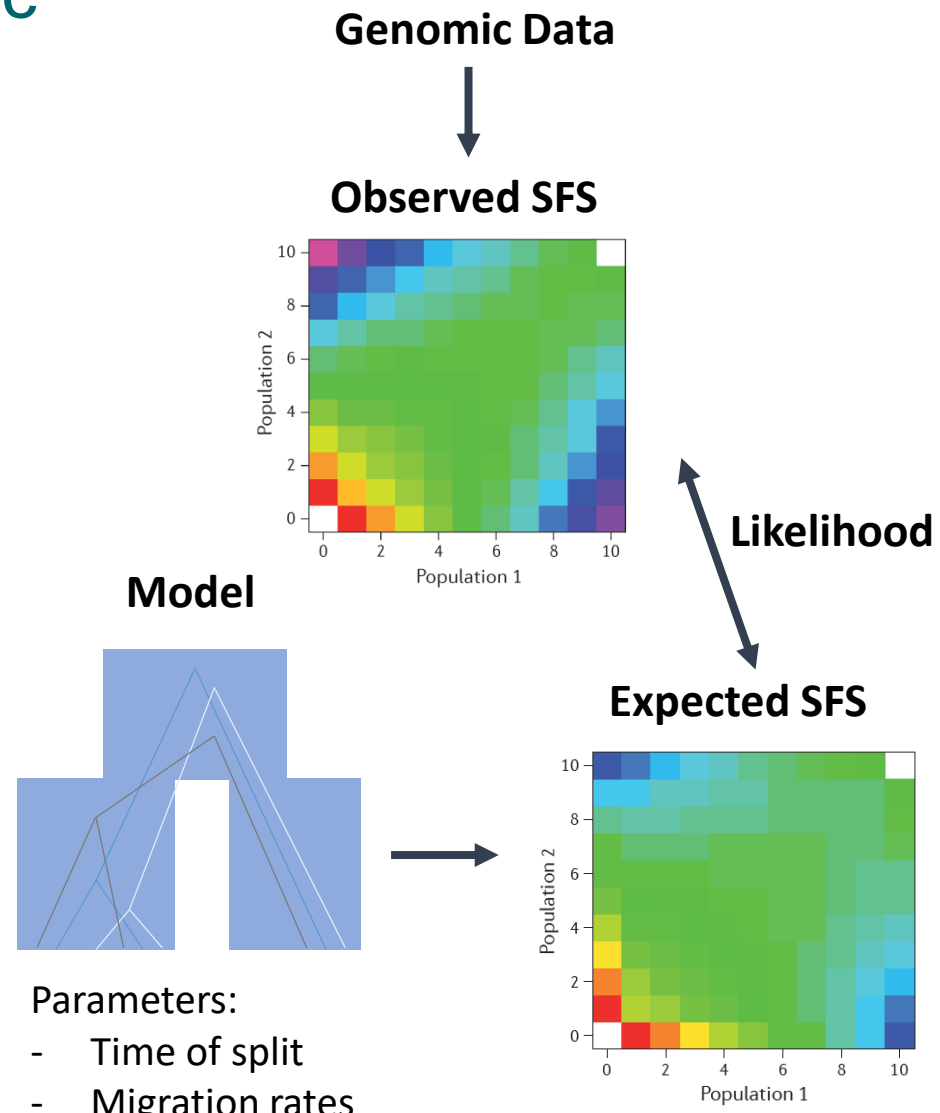
Parameters:

- Time of split
- Migration rates
- Effective sizes

Inferring the demographic history from the SFS

- The likelihood is easily computed based on the expected SFS under a given model

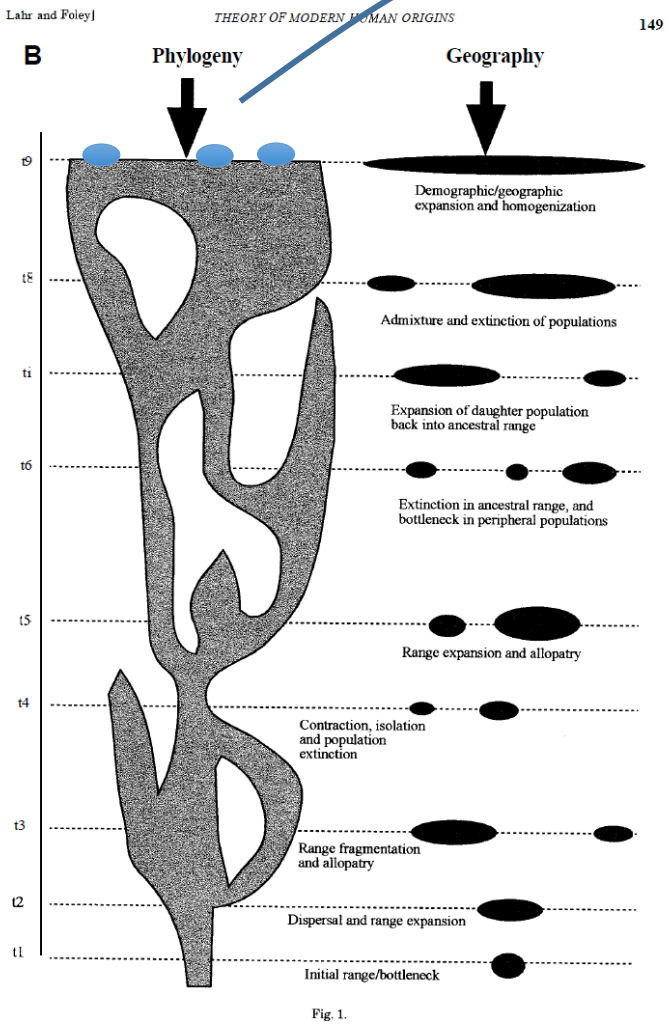
- There are different ways to obtain the expected SFS
  - Diffusion (forward in time)
  - Coalescent (backward in time)



Excoffier et al. (2013) *PloS Genetics*

Framework for demographic inference

"Truth"



Lahr and Foley, 1994

Sample genomic data

What generated the data?
Test specific hypotheses.

Define demographic scenarios

Models

Estimation of demographic parameters

"All models are wrong but some are useful"
George Box

Estimating the SFS from the coalescent

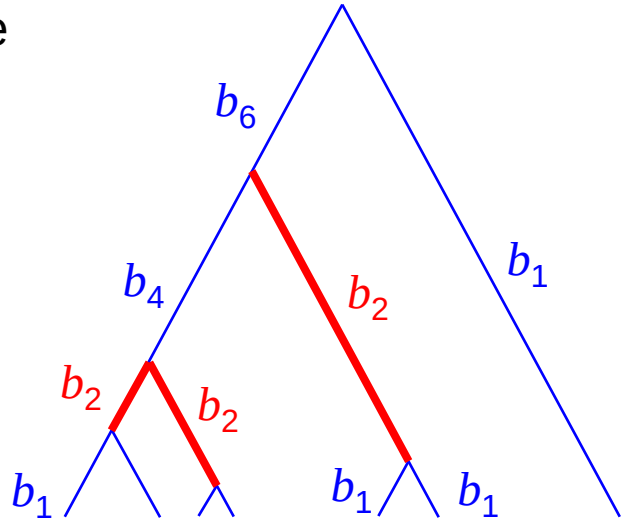
The probability of a SFS entry *i* can be estimated under a specific model θ from its expected coalescent tree as (Nielsen 2000)

$$p_i = \frac{E(t_i \mid \theta)}{E(T \mid \theta)}$$

Where t_i is the total length of all branches directly leading to i terminal nodes, and T is the total tree length.

It gives the relative probability that if a mutation occurs on one of these b_i branches, it will be observed *i* times in the sample

This is true under the infinite sites model. No more than 1 mutation per site, back mutations not allowed!



Composite likelihood

Even though we can have linked sites, we assume that all sites are independent. Given *S* polymorphic sites (SNPs) out of *L* sites (Adams and Hudson, 2004) the composite likelihood is:

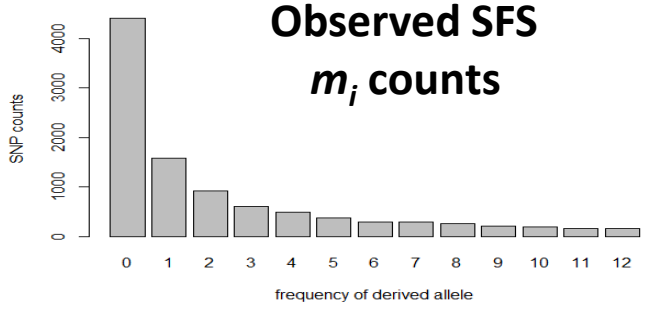
$$CL = \Pr(X \mid \theta) \propto P_0^{L-S} (1-P_0)^S \prod_{i=1}^{n-1} \hat{p}_i^{\,m_i}$$

probability of no mutation on the tree

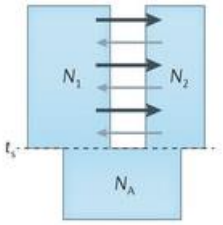
probability of at least one mutation in the tree

3 ingredientes for likelihood

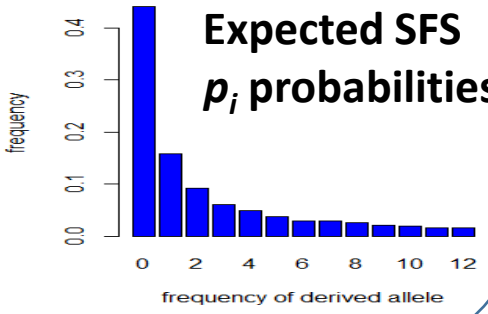
Observed SFS
m_i counts



Model



Expected SFS
p_i probabilities



Composite likehood

These probabilities depend:

- Number of monomorphic sites
- A fixed and mutation rate

Methods based on the SFS

Different ways to obtain the expected SFS p_i under different demographic models

- Coalescent-based
  - Multiple populations

    Fastsimcoal2 (Excoffier et al 2013 PLoS Genetics)

    Momi (Kamm et al 2015) and Momi 2

    Rarecoal (Schiffels et al 2016 Nat Genetics)
  - Single population

    Stairway plot (Liu and Fu, 2015 Nat Genetics)
- Diffusion-based

  Dadi (Gutenkunst et al 2009 PLoS Genetics)

  Multipop (Lukic and Hey 2012 Genetics)

  Jouganous et al (2017) Genetics

fastsimcoal2 program

- Fastsimcoal2 can estimate parameters from the SFS using coalescent simulations
- Maximum (composite) likelihood method
- Uses a conditional expectation (CEM) maximization algorithm to find parameter combinations that maximize the likelihood
- **It approximate the expected SFS** by performing coalescent simulations (>50,000)

Estimating the SFS and likelihoods with coalescent simulations

This probability $\boldsymbol{p_i}$ can then be estimated on the basis of Z simulations as

$$\hat{p}_i = \frac{\sum_{j}^{Z} \sum_{k \in \Phi_i} b_{kj}}{\sum_{j}^{Z} T_j}$$ where b_{kj} is the length of the k-th compatible branch in simulation j.

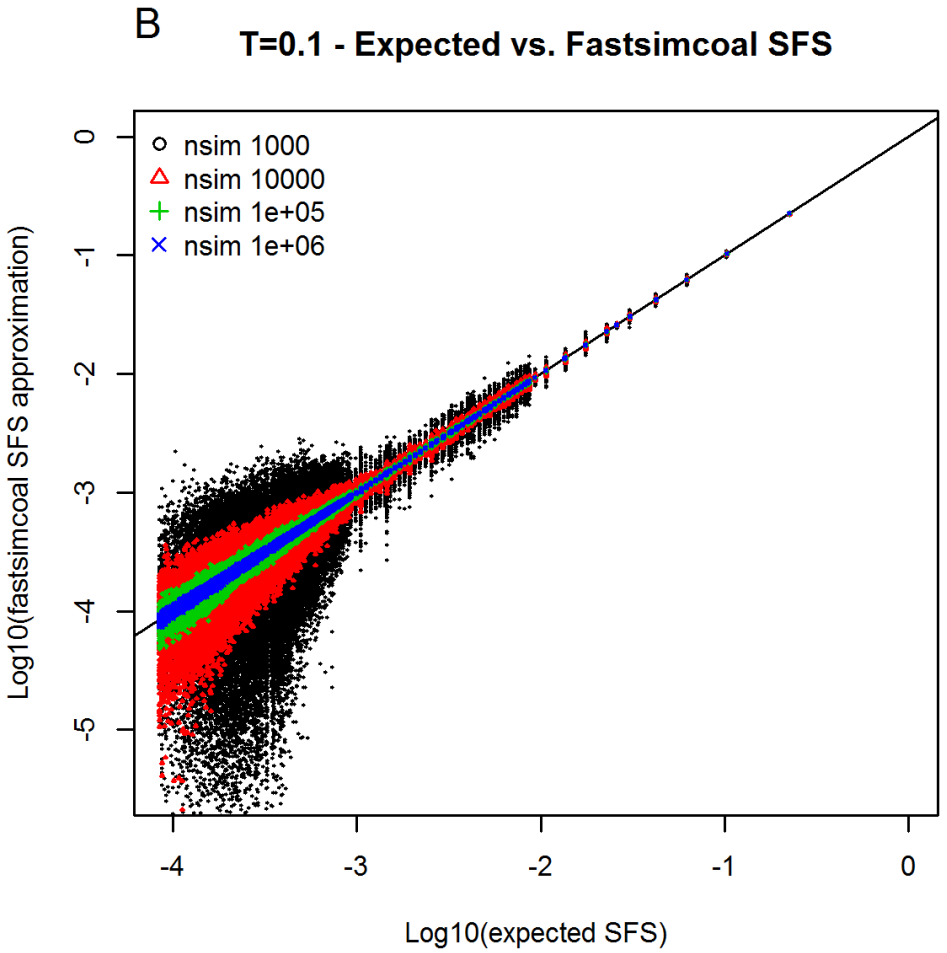
These probabilities can then be used to compute the composite likelihood of a given model as (Adams and Hudson, 2004)

$$CL = \Pr(X \mid \theta) \propto P_0^{L-S} (1 - P_0)^S \prod_{i=1}^{n-1} \hat{p}_i^{m_i}$$

where X is the SFS in a population sample of size n, S is the number of polymorphic sites, L is the length of the studied sequence, and P_0 is the probability of no mutation on the tree

Approximating the expected SFS with coalescent simulations

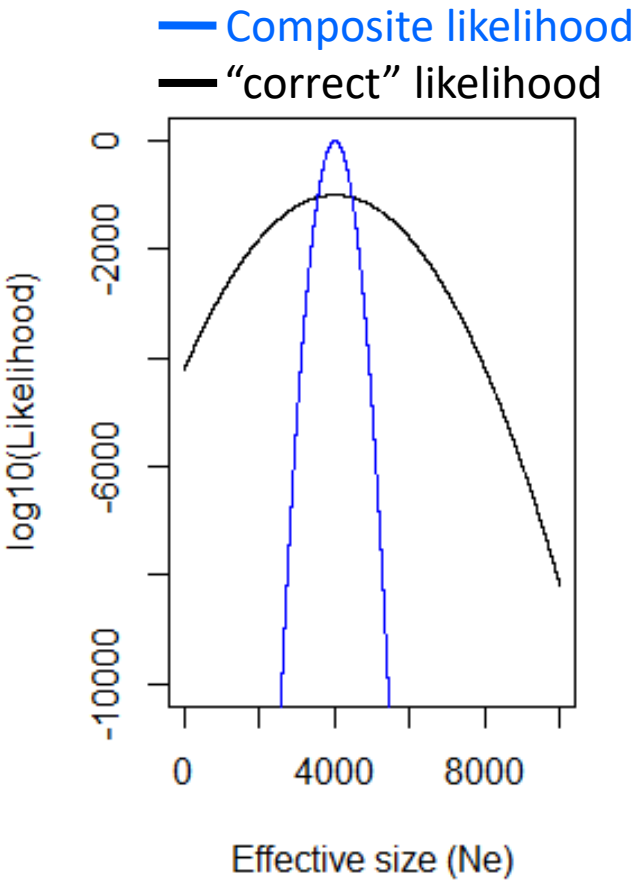
Increasing the number of simulations improves the approximation of the expected SFS



Properties of composite likelihoods

This composite likelihood (CL) is not a proper likelihood due to the non-independence of allele frequencies at linked sites.

- CL is maximized for the same parameters as full likelihood
- Can be used for parameter estimation
- Confidence intervals cannot be estimated from likelihood profile, need to bootstrap
- CL surface might be more complex than likelihood surface, and thus more difficult to explore and get the global maximum
- CL ignores information on linkage disequilibrium (recombination) between sites



Evolutionary forces affecting the history of populations

Demography

- Past effective population sizes
- Past migration rates

Selection

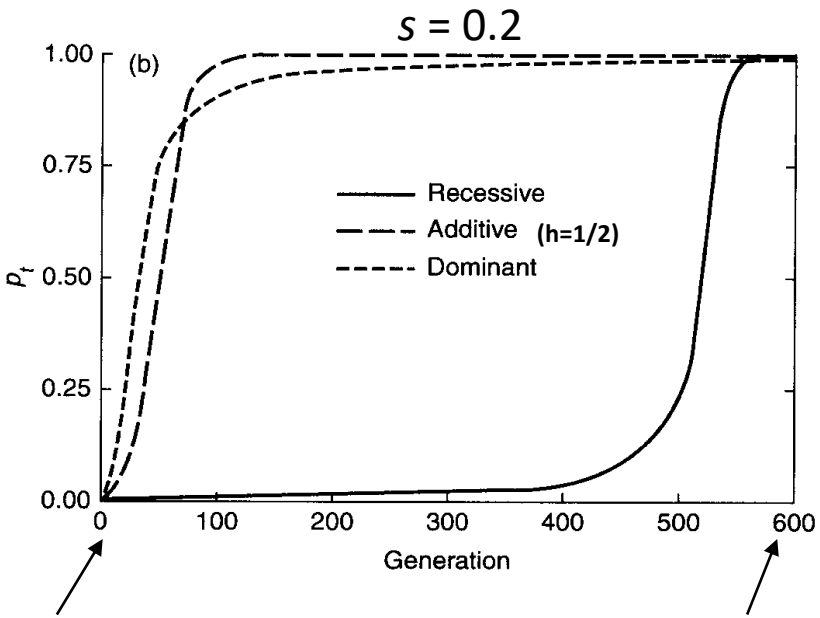
- Selective coefficient and type of selection (positive or negative)

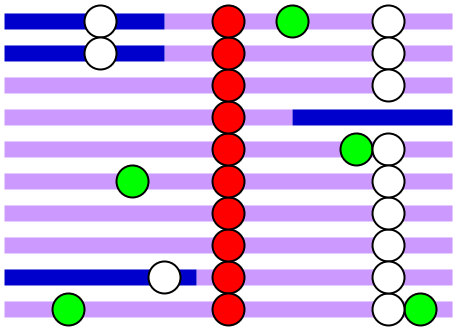
Genomic processes

- Mutation rate
- Recombination rate

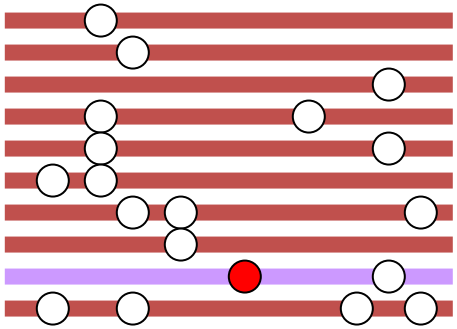
Effect of directional selection on molecular diversity

- Advantageous mutation
- Neutral mutations

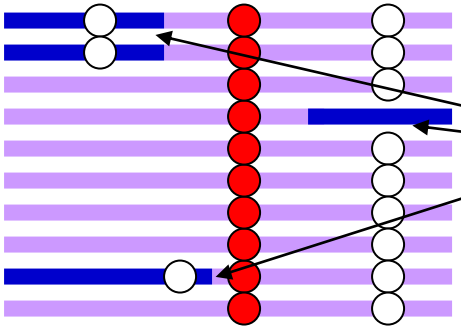




After some time, the original level of diversity is restored by mutation and recombination



Before selection, a new favorably selected mutation appears



Recombinant segments

After selection, the mutation has fixed and there is much less diversity around the selected site

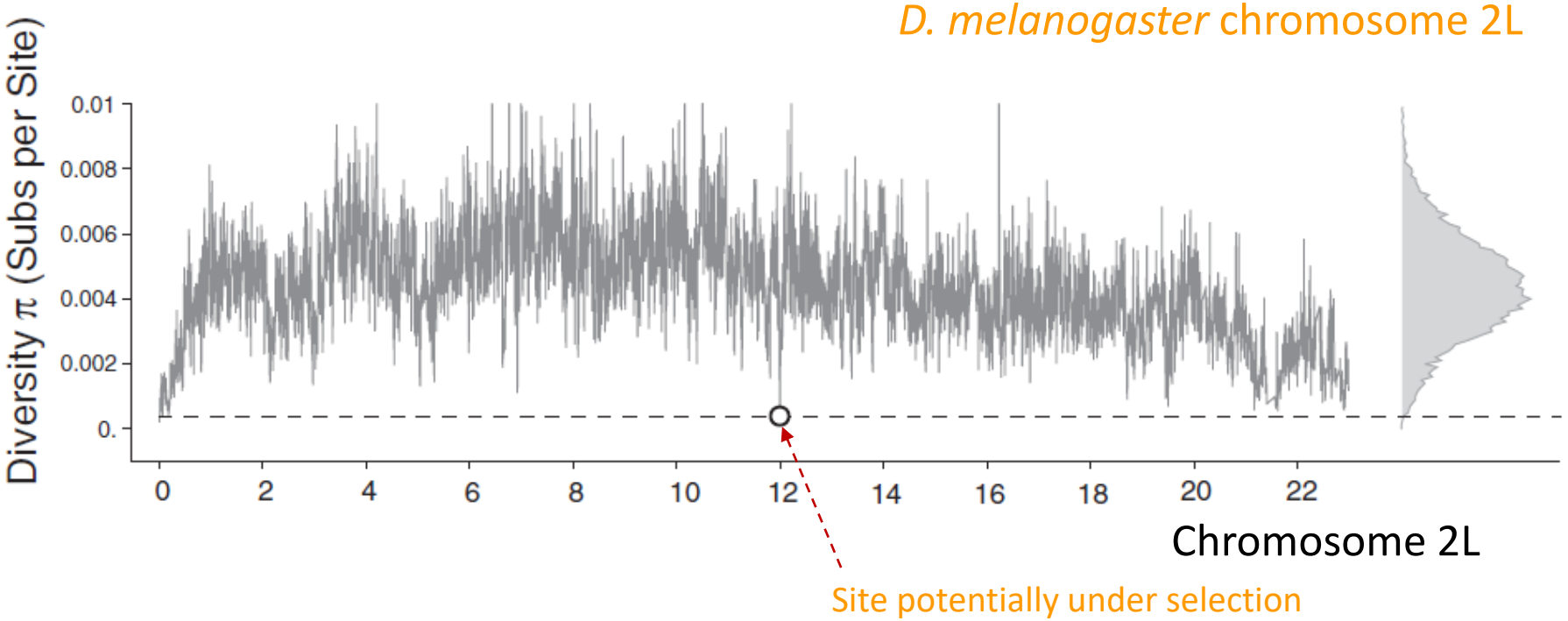
There has been a selective sweep

Genome scans to detect selection

One can measure pattern of diversity, e.g. θ_π along chromosomes to discover sites potentially under selection

θ_π measured in windows of 10 kb

D. melanogaster chromosome 2L



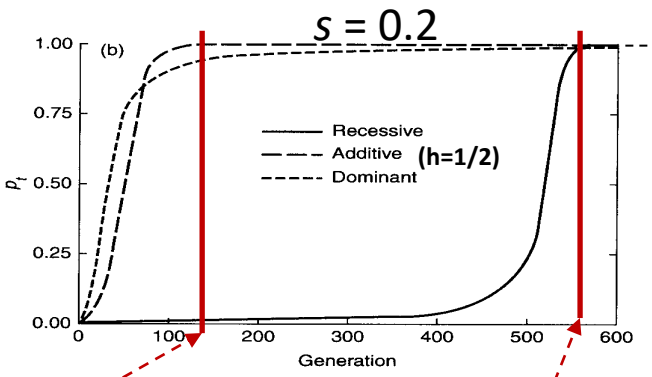
Quetzalcoatl gene formed by the fusion of two genes and which swept to fixation recently

Rogers et al. 2010 PNAS

Genomic diversity profiles

With genomic data, one can follow levels of diversity along the chromosome after a selective sweep

Diversity is measured at the end of the sweep



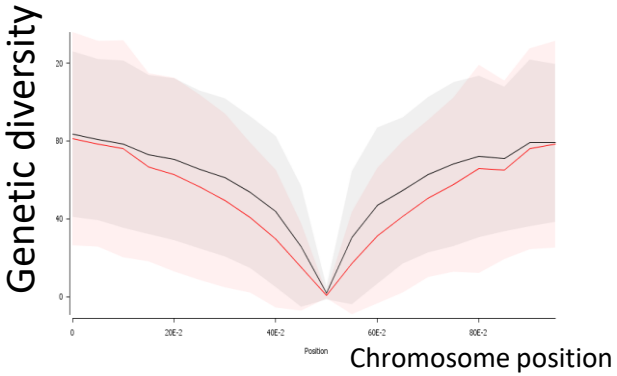
Positive selection

N=10000

$\theta = 100$

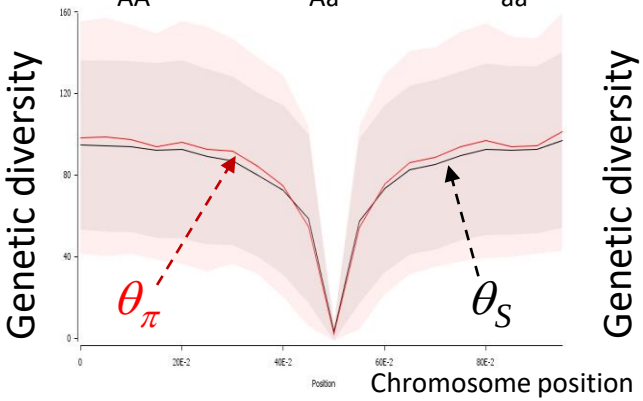
Additive

w_{AA}=1.2 w_{Aa}=1.1 w_{aa}=1



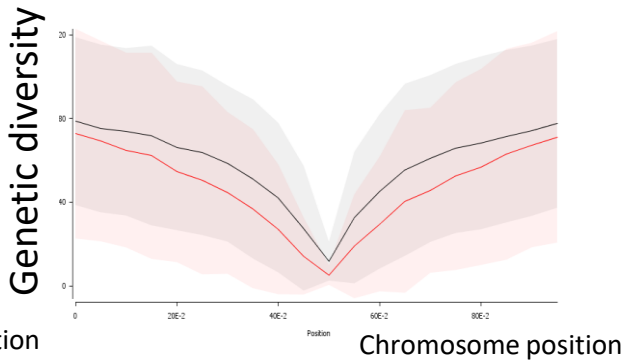
Recessive

w_{AA}=1.2 w_{Aa}=1 w_{aa}=1



Dominant

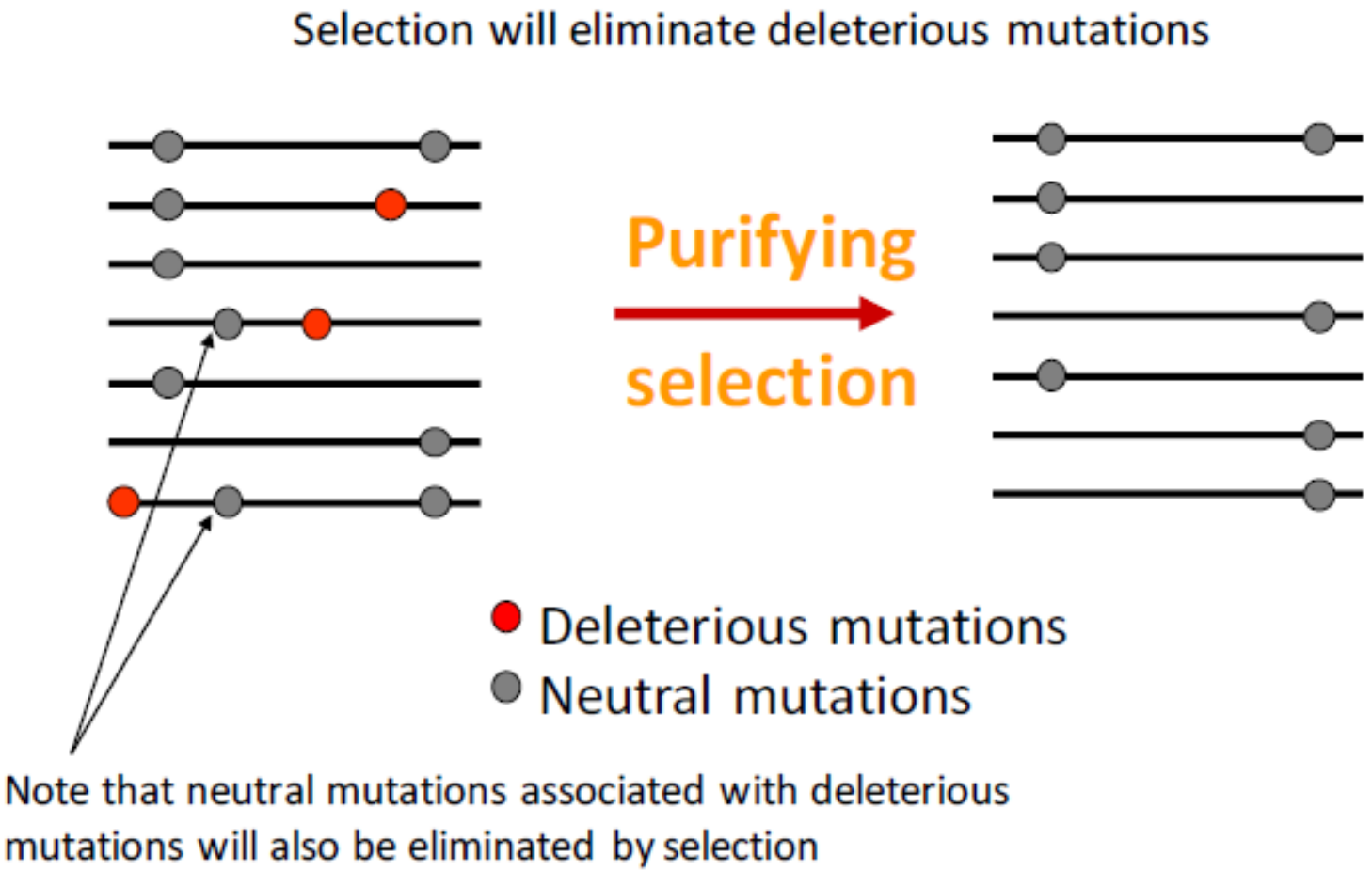
w_{AA}=1.2 w_{Aa}=1.2 w_{aa}=1



Signal of selection is less clear for dominant or additive mutations

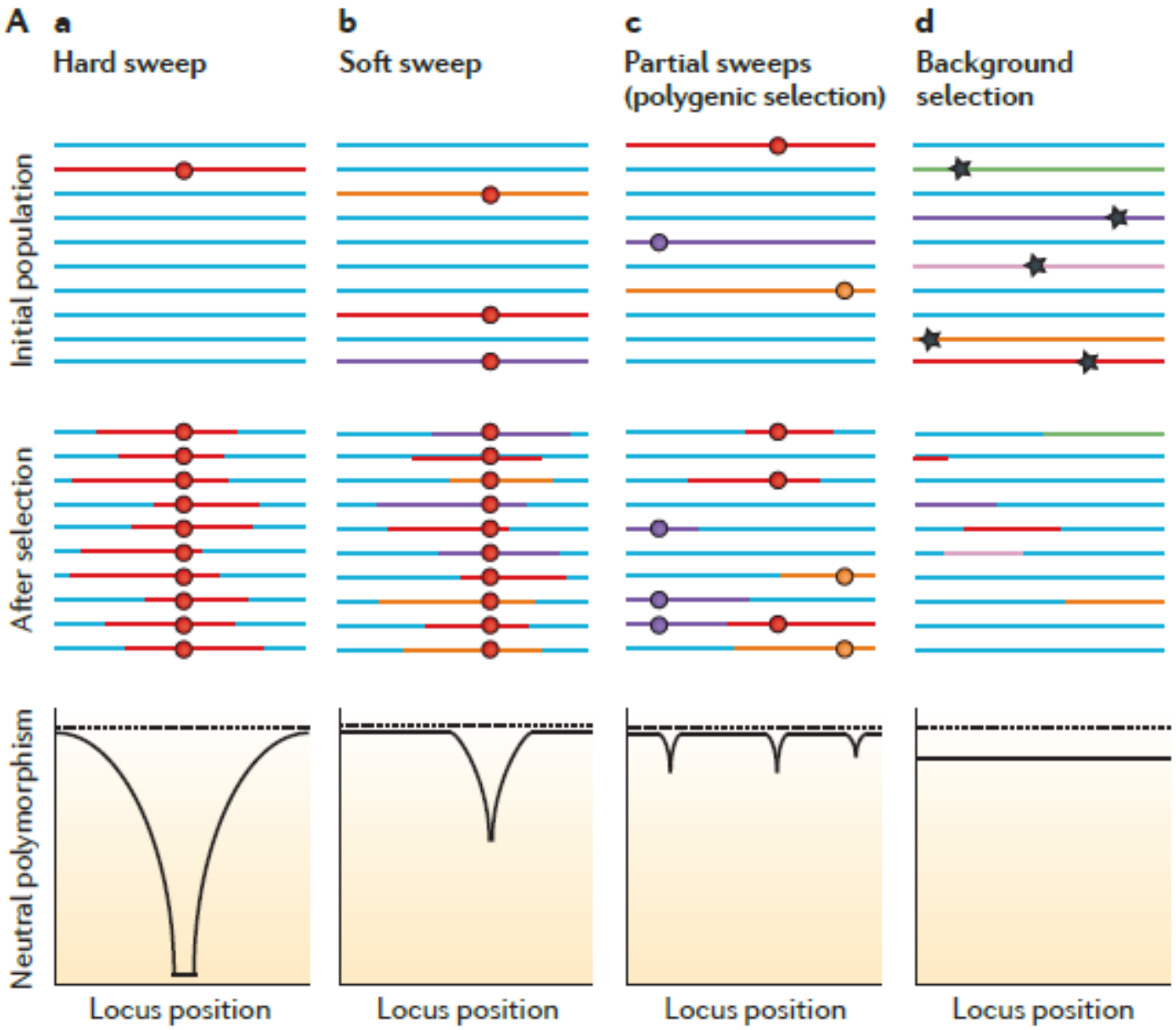
msmsPlay (http://www.mabs.at/ewing/msms/msmsplay.shtml)

Selection against deleterious mutations



- The effect of purifying selection on linked neutral variation is called **background selection**
- Background selection leads to a decreased diversity, as linked neutral variation is eliminated together with the deleterious mutations

Different forms of hitchhiking effects



As time goes by and recombination happens between haplotypes with and without selected alleles, we expect to see differences:

- selective sweeps lead to a localized effect, reducing genetic diversity around the selected locus.
- background selection will affect the entire region in a similar way.

Cutter and Payseur (2013) Nature Reviews Genetics

Evidence of linked selection (i.e. effects of selection on linked neutral sites) at the genomic level:

- Positive correlation of recombination rate with genetic diversity
- Neutral mutations in regions of low recombination are affected by positive selection (sweeps) and negative selection (background selection)
- Thus, regions of low recombination show less genetic diversity
- If al mutations were neutral, diversity should be the same in regions of low and high recombination

Cutter and Payseur (2013) Nature Reviews Genetics

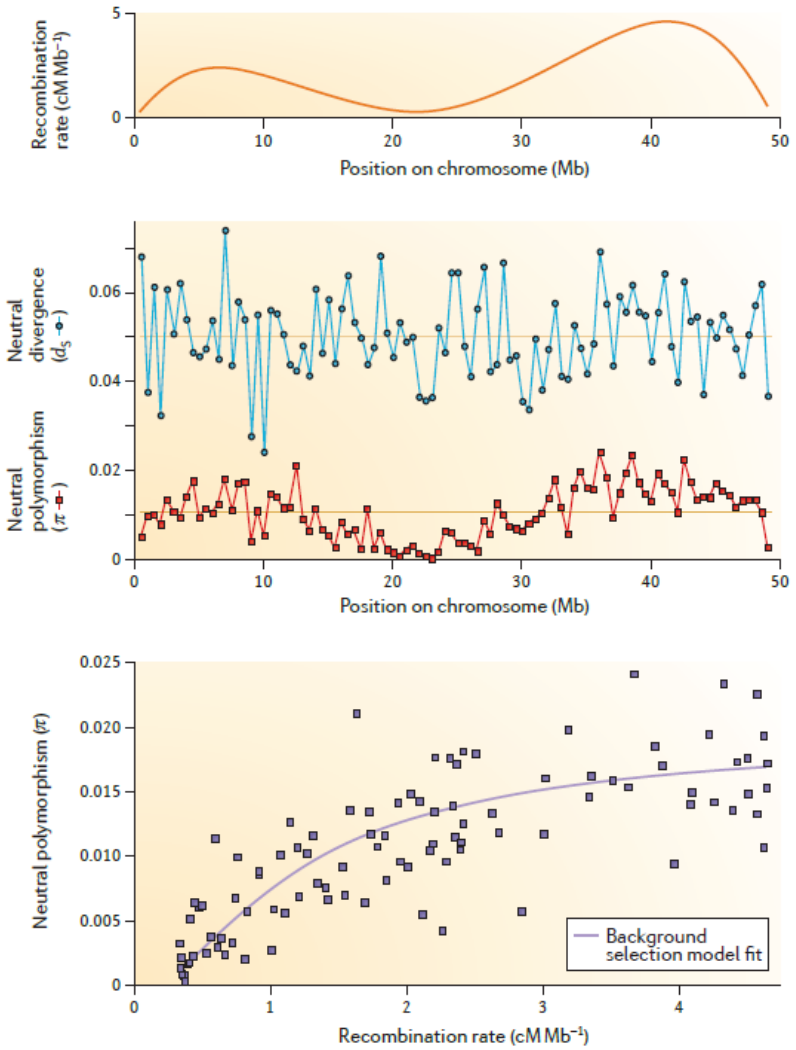
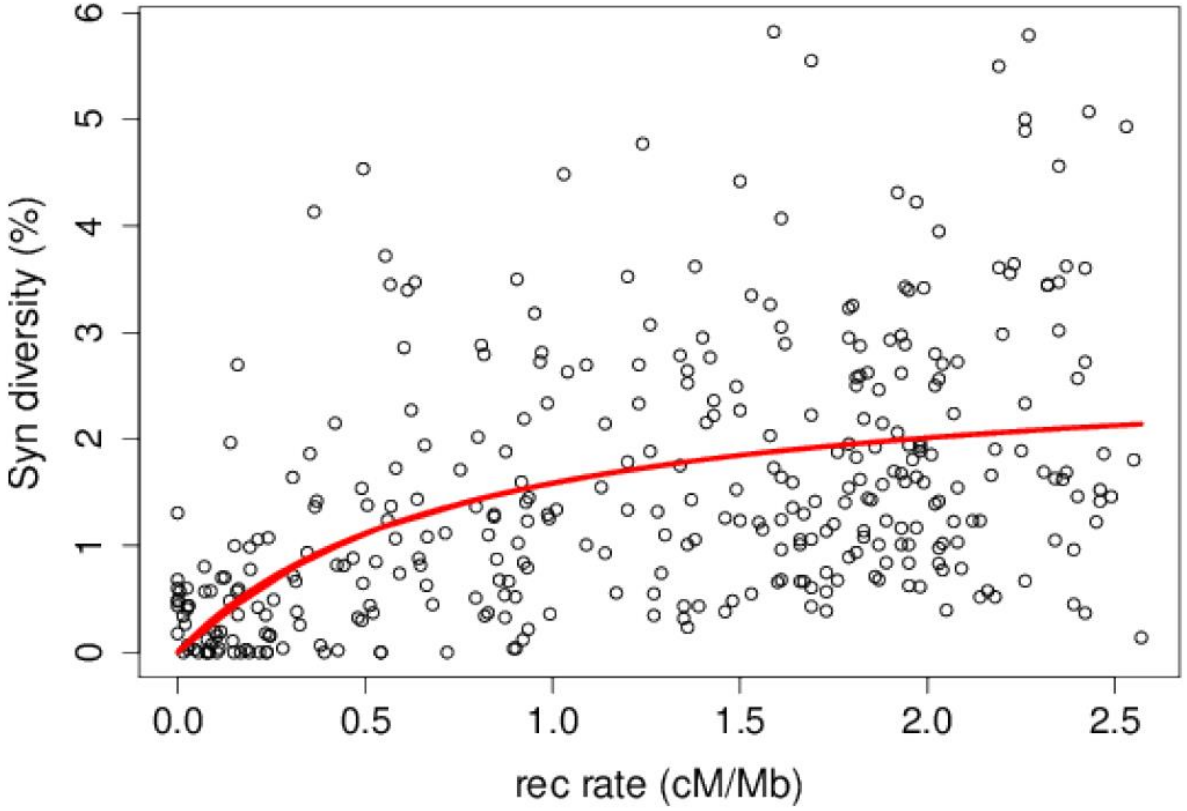


Figure 1 | **A hypothetical chromosome exhibiting a 'selection at linked sites' pattern.** Given the hypothetical recombination rate profile across the chromosome (top panel) and random variation in mutation rate, which is reflected in divergence at unconstrained sites (blue points, middle panel), measures of neutral polymorphism at unconstrained sites are predicted by recurrent genetic hitch-hiking and background selection to be lower in chromosomal regions with lower average recombination rates (red points, middle panel). Recurrent positive or negative selection would yield a positive association between polymorphism and recombination rate (bottom panel). These hypothetical data were generated assuming background selection, but recurrent hard sweeps would yield a qualitatively similar pattern. cM, centimorgans; Mb, megabases.

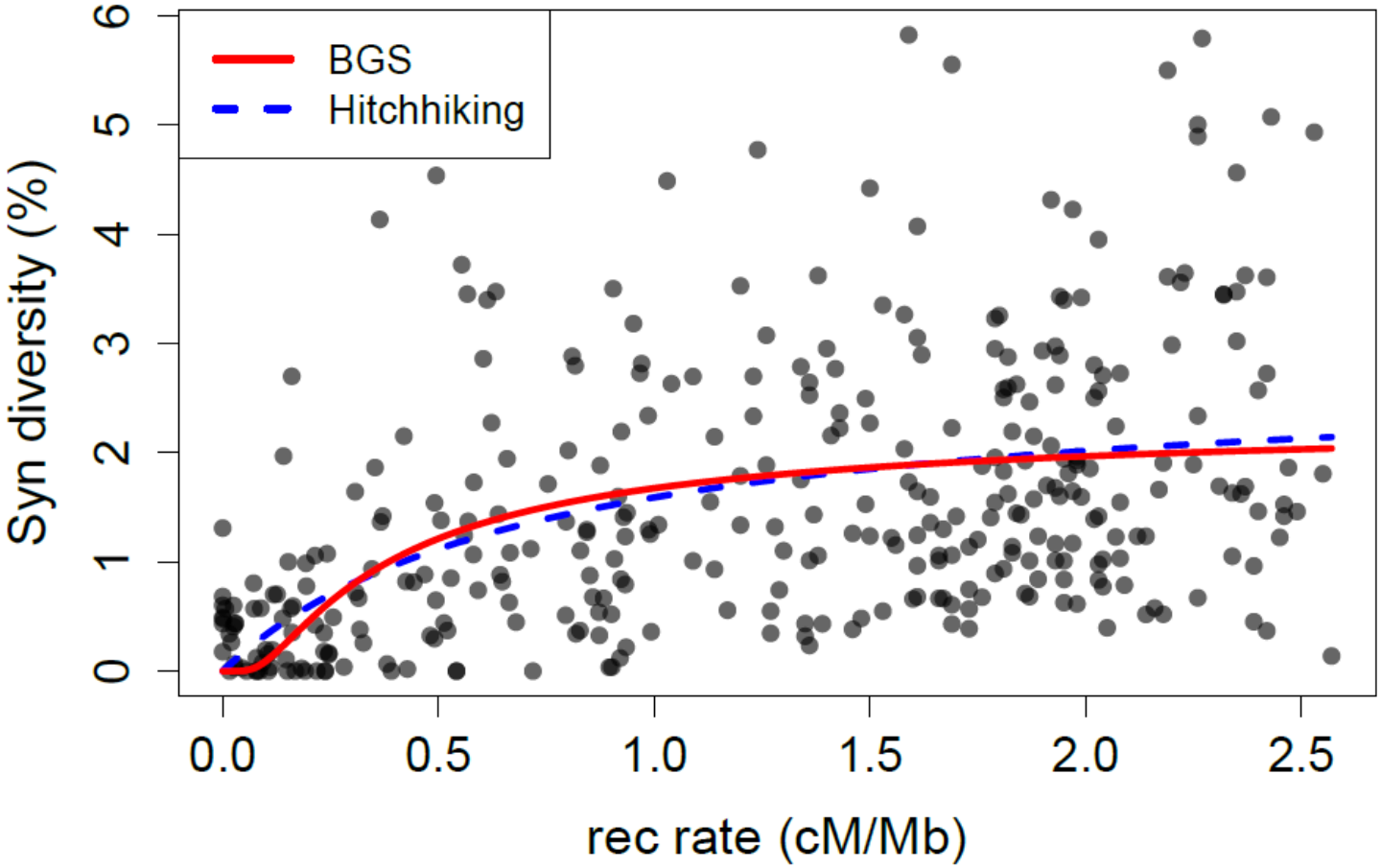
Evidence from *Drosophila melanogaster*

The relationship between (sex-averaged) recombination rate and synonymous site pairwise diversity in *Drosophila melanogaster*. The curve is the predicted relationship between and recombination rate, obtained by fitting a recurrent hitchhiking model



Coop (2019) Population Genetics Notes
https://github.com/cooplab/popgen-notes/

Challenge to distinguish effect of recurrent selective sweeps from background selection



Coop (2019) Population Genetics Notes
https://github.com/cooplab/popgen-notes/

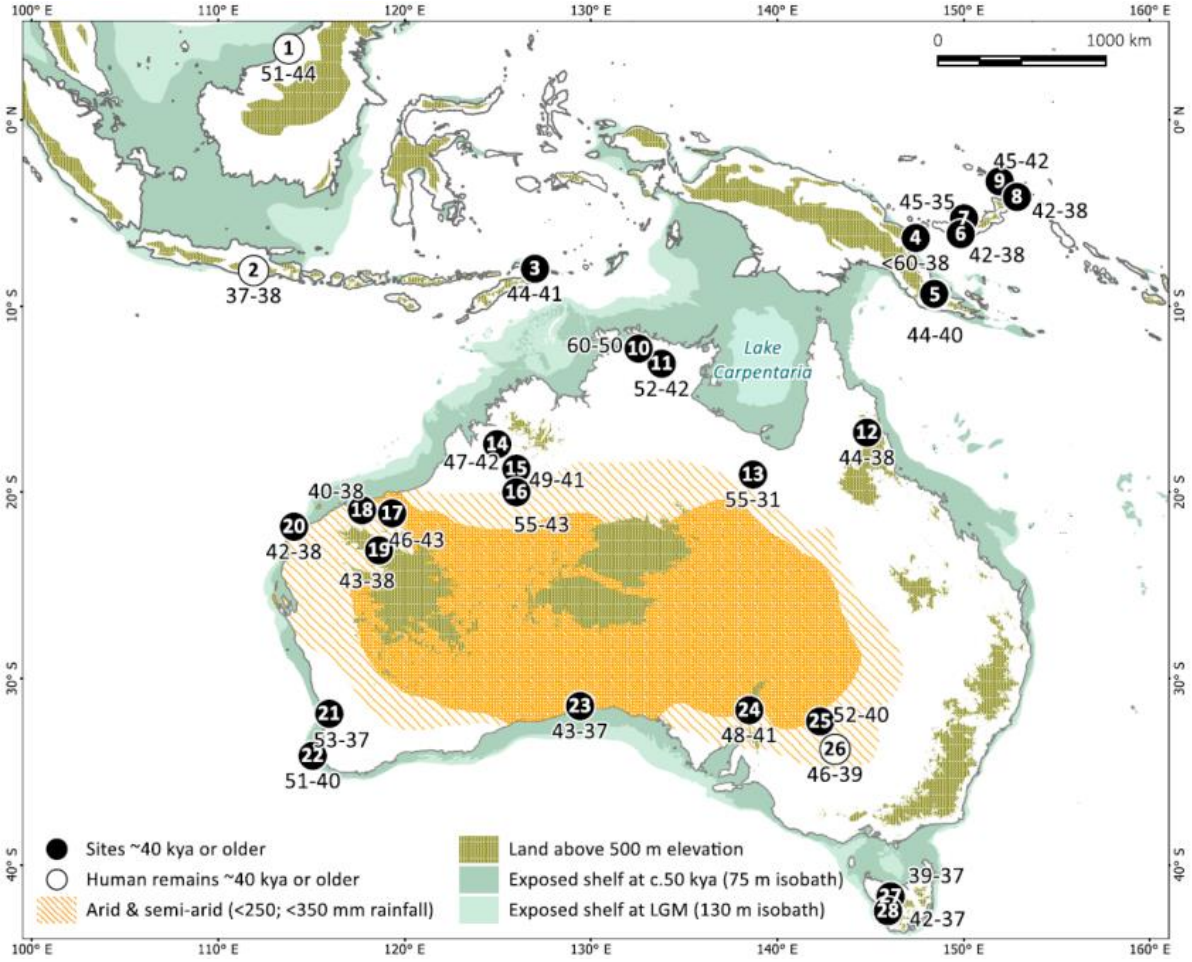
A genomic history of Aboriginal Australia

Anna-Sapfo Malaspinas[1,2,3]*, Michael C. Westaway[4]*, Craig Muller[1]*, Vitor C. Sousa[2,3]*, Oscar Lao[5,6]*, Isabel Alves[2,3,7]*, Anders Bergström[8]*, Georgios Athanasiadis[9], Jade Y. Cheng[9,10], Jacob E. Crawford[10,11], Tim H. Heupink[4], Enrico Macholdt[12], Stephan Peischl[3,13], Simon Rasmussen[14], Stephan Schiffels[15], Sankar Subramanian[4], Joanne L. Wright[4], Anders Albrechtsen[16], Chiara Barbieri[12,17], Isabelle Dupanloup[2,3], Anders Eriksson[18,19], Ashot Margaryan[1], Ida Moltke[16], Irina Pugach[12], Thorfinn S. Korneliussen[1], Ivan P. Levkivskyi[20], J. Víctor Moreno-Mayar[1], Shengyu Ni[12], Fernando Racimo[10], Martin Sikora[1], Yali Xue[8], Farhang A. Aghakhanian[21], Nicolas Brucato[22], Søren Brunak[23], Paula F. Campos[1,24], Warren Clark[25], Sturla Ellingvåg[26], Gudjugudju Fourmile[27], Pascale Gerbault[28,29], Darren Injie[30], George Koki[31], Matthew Leavesley[32], Betty Logan[33], Aubrey Lynch[34], Elizabeth A. Matisoo-Smith[35], Peter J. McAllister[36], Alexander J. Mentzer[37], Mait Metspalu[38], Andrea B. Migliano[29], Les Murgha[39], Maude E. Phipps[21], William Pomat[31], Doc Reynolds[40], Francois-Xavier Ricaut[22], Peter Siba[31], Mark G. Thomas[28], Thomas Wales[41], Colleen Ma'run Wall[42], Stephen J. Oppenheimer[43], Chris Tyler-Smith[8], Richard Durbin[8], Joe Dortch[44], Andrea Manica[18], Mikkel H. Schierup[9], Robert A. Foley[1,45], Marta Mirazón Lahr[1,45], Claire Bowern[46], Jeffrey D. Wall[47], Thomas Mailund[9], Mark Stoneking[12], Rasmus Nielsen[1,48], Manjinder S. Sandhu[8], Laurent Excoffier[2,3], David M. Lambert[4] & Eske Willerslev[1,8,18]

Nature(2016)

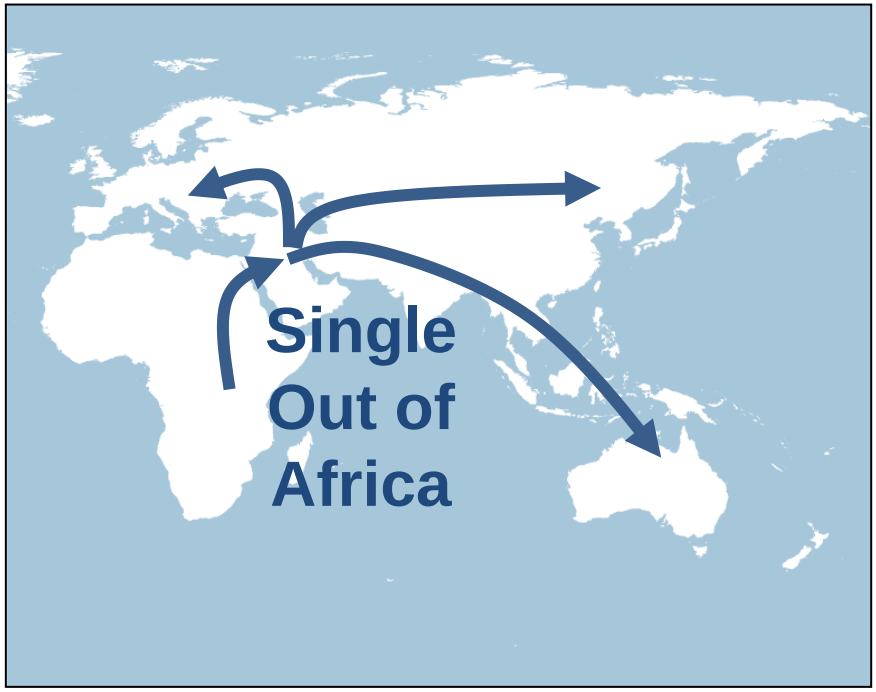
Ewaninga Rock Carvings Conservation Reserve, NT, Australia

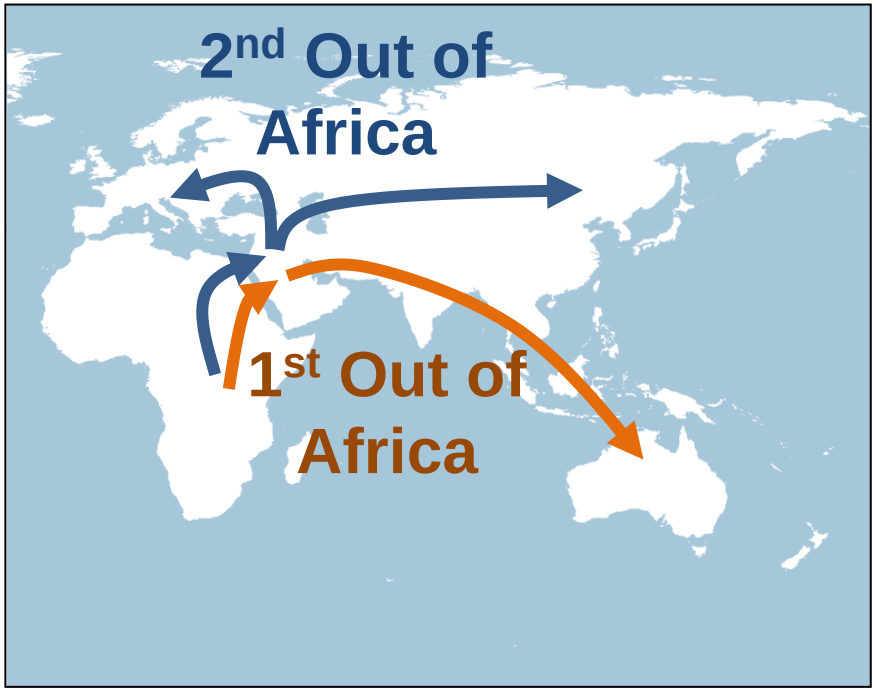
Australia harbors some of the oldest modern human remains outside Africa



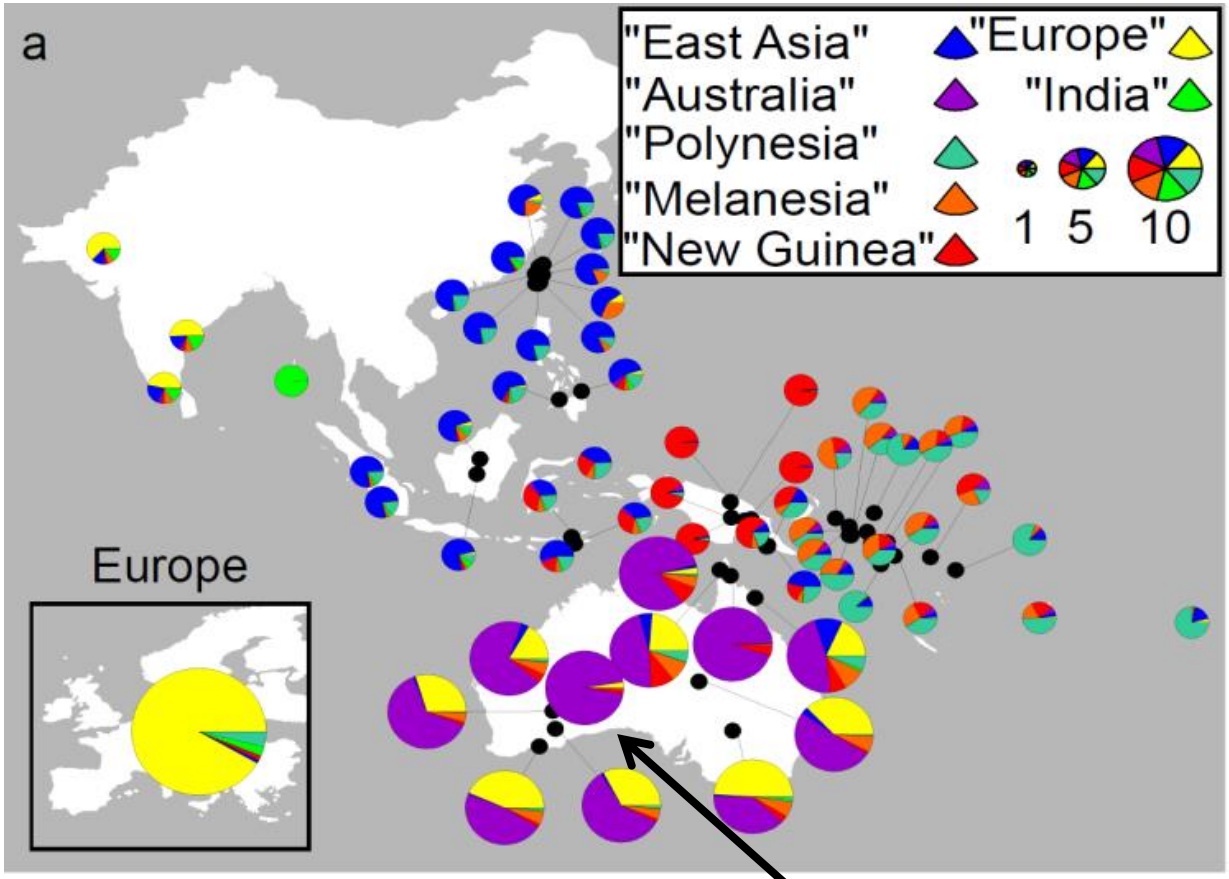
Many sites and remains dated to be older than 40 kya, suggesting a human settlement 47.5-55 kya

One wave out of Africa *vs* Two waves out of Africa





83 high-coverage Aboriginal Australians genomes

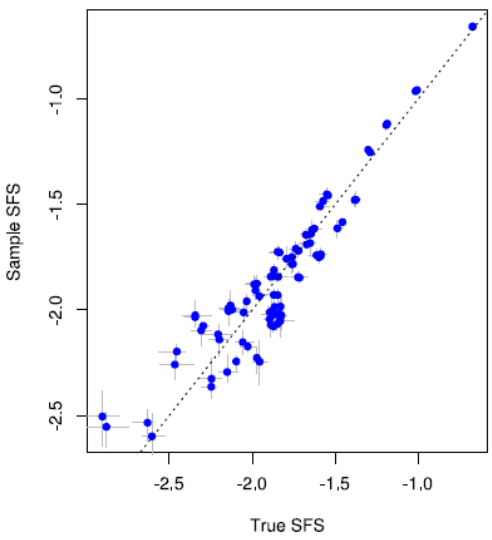


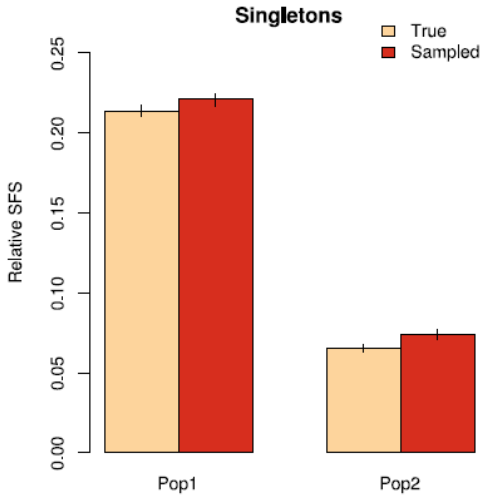
Average depth of coverage: 65x

A note on recovering the SFS from genomic data

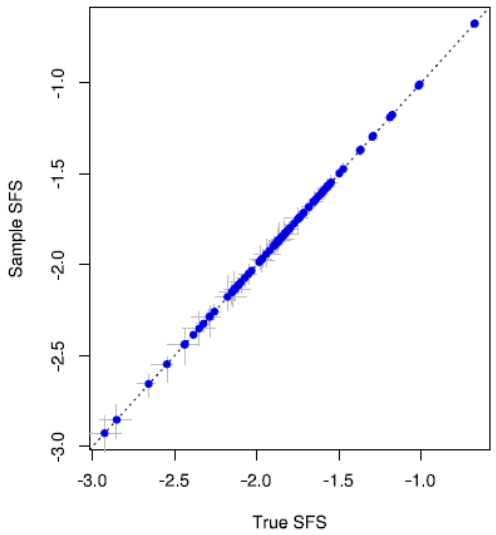
- Simulation study
- Low depth of coverage and missing data lead to biased SFS towards rare variants

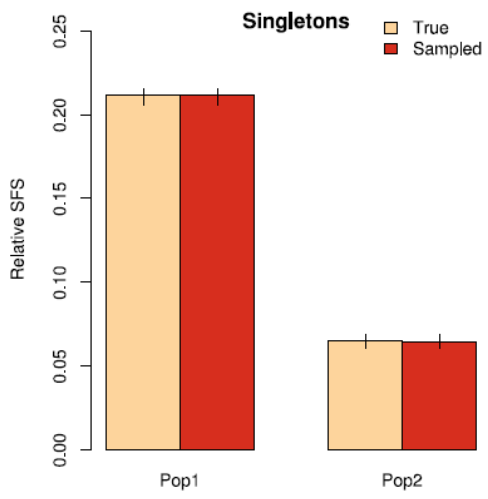
a) Low depth of coverage, no GQ filter, allowing missing data





b) Depth of coverage similar to observed data, GQ>30 filter, no missing dat





★ Archaic human genomes:
- 1 Neanderthal (~66 kya)
- 1 Denisovan (~52 kya)

Mutation rate assumed
1.25×10^{-8} /site/gen
Scally and Durbin (2012) *Nat. Rev. Genet.*

Generation time
29 years/gen
Fenner (2005) *Am. J. Phys. Anthropol.*

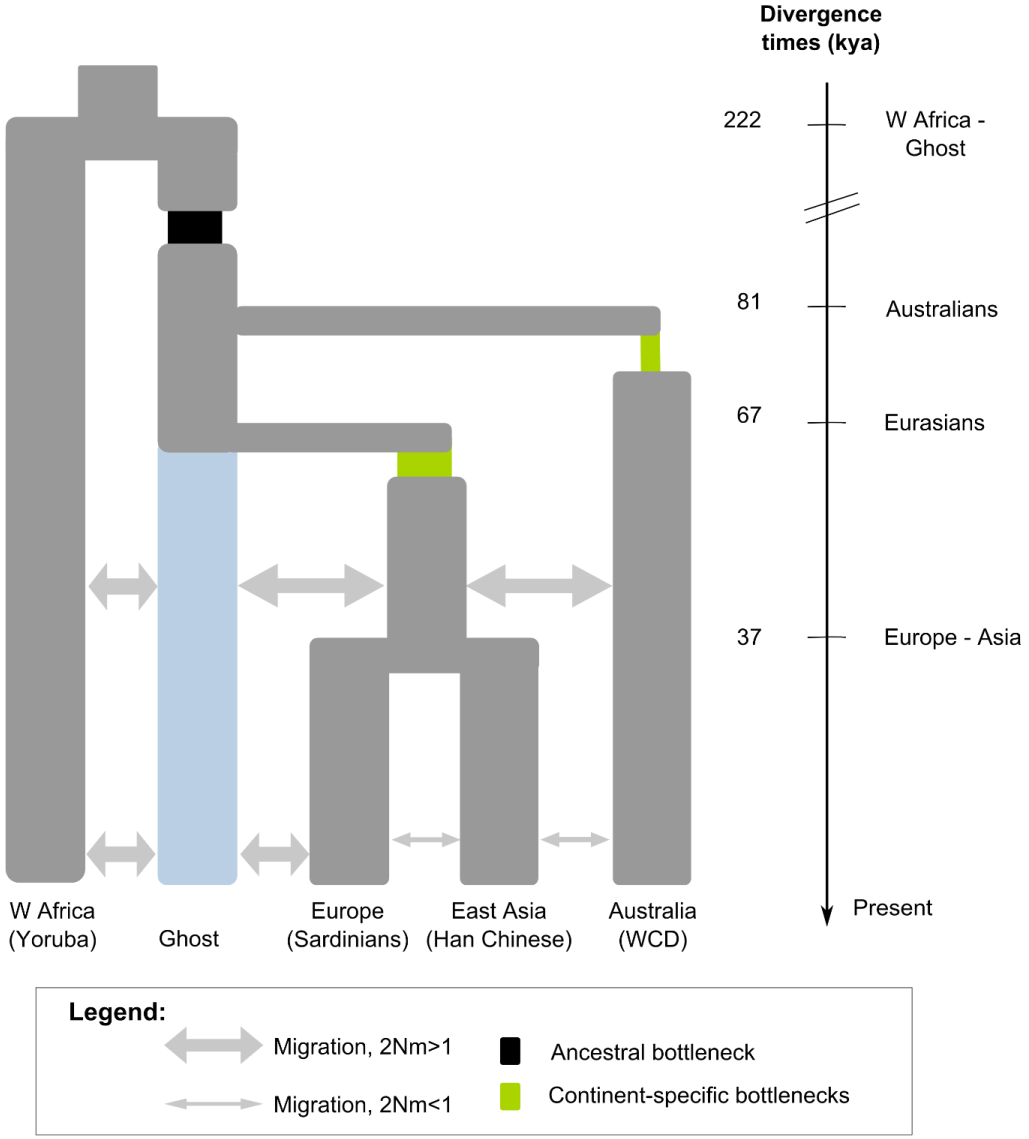
Since we want to infer demography we tried to minimize the number of sites affected by selection:

- 985 1Mb blocks outside genic regions and CpG islands (~4.3 Million SNPs)
- 5 dimensional SFS (16,875 entries)
- Confidence intervals obtained using block-bootstrap

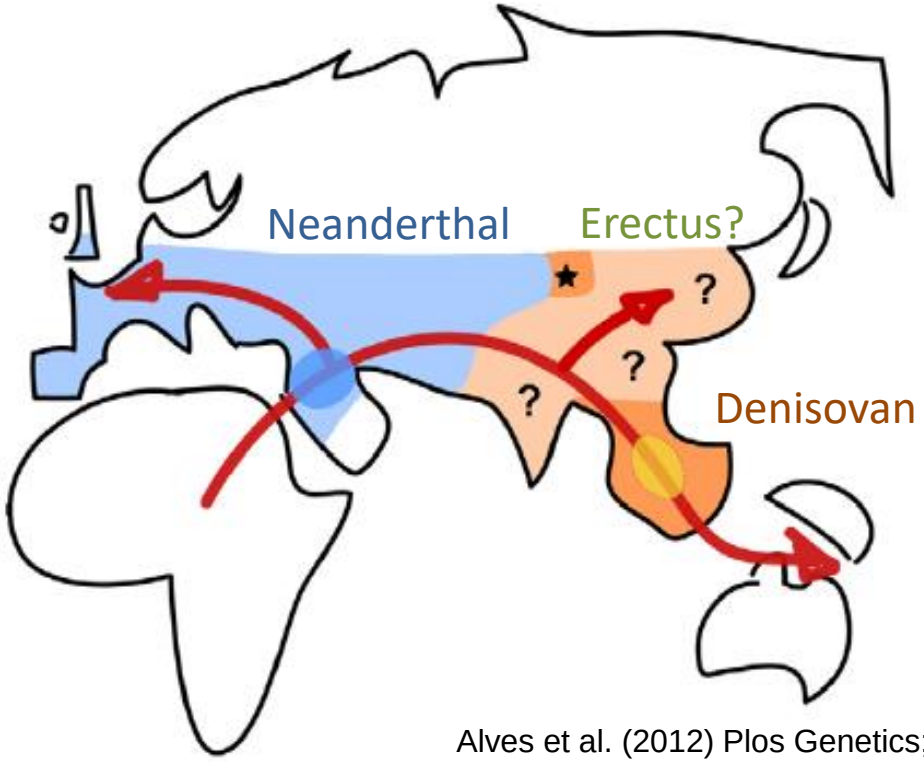
We always get results…

Evidence of two waves Out of Africa:

- Old split leading to colonization of Australia (81kya)
- More recent split leading to colonization of Eurasia (67 kya)



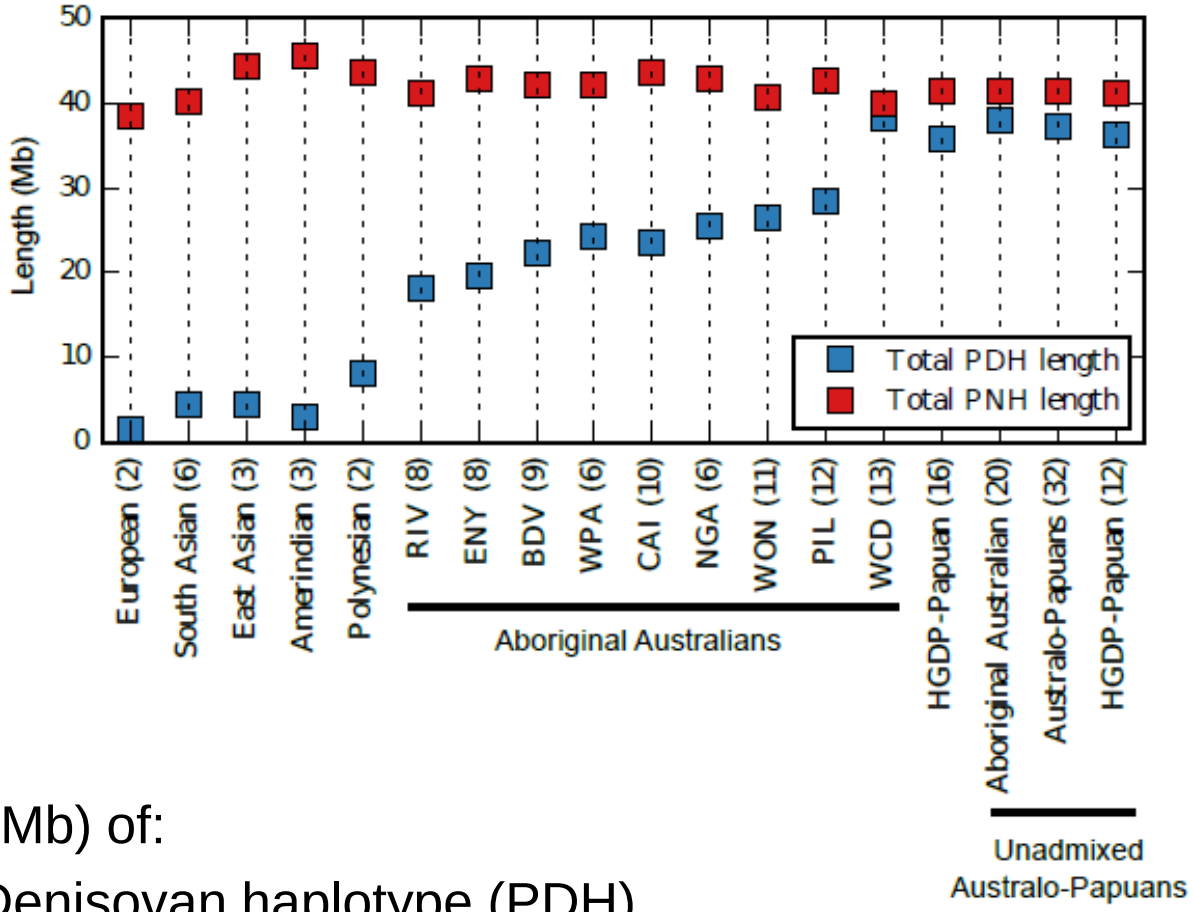
Towards a model incorporating Neanderthal and Denisovan admixture



Alves et al. (2012) Plos Genetics;

- Non-African populations: 1-4% estimated Neanderthal admixture
- Aboriginal Australians and New Guineans: 3-6% estimated Denisovan admixture
- Archaic admixture can affect times of split estimates

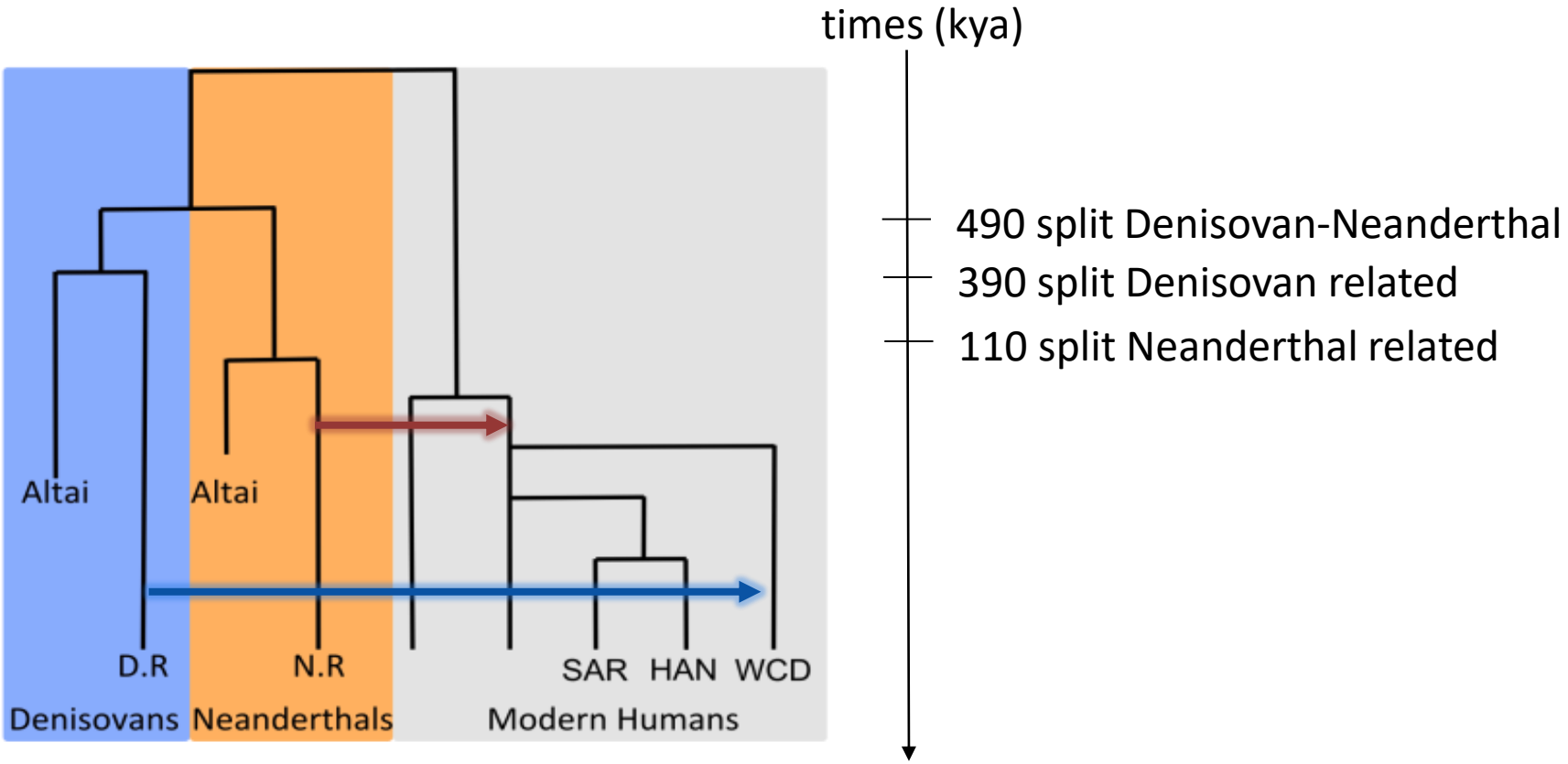
Evidence of archaic introgression



Total length (Mb) of:

- Putative Denisovan haplotype (PDH)
- Putative Neanderthal haplotypes (PNH)

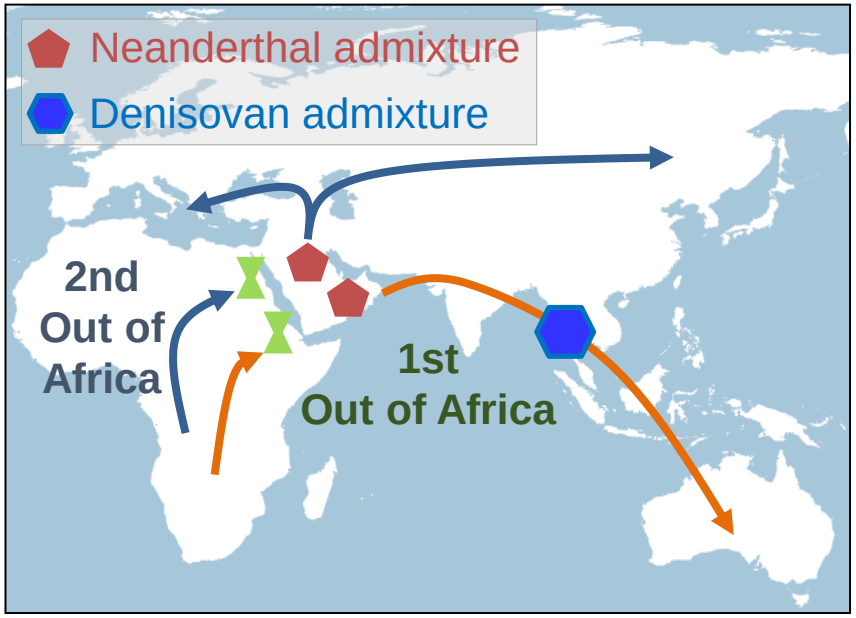
Accounting for shared ancestry of Neanderthal and Denisovan

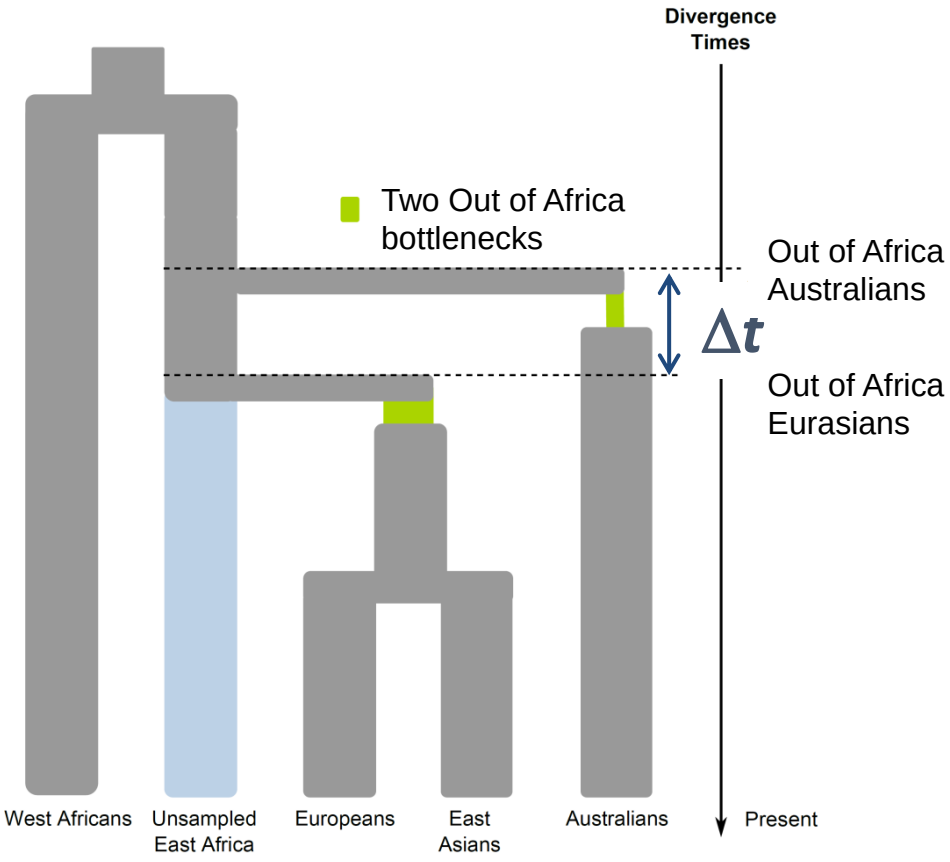


Admixture occurs between modern humans and:

- Denisovan-related (D.R.) population
- Neanderthal-related (N.R.) population

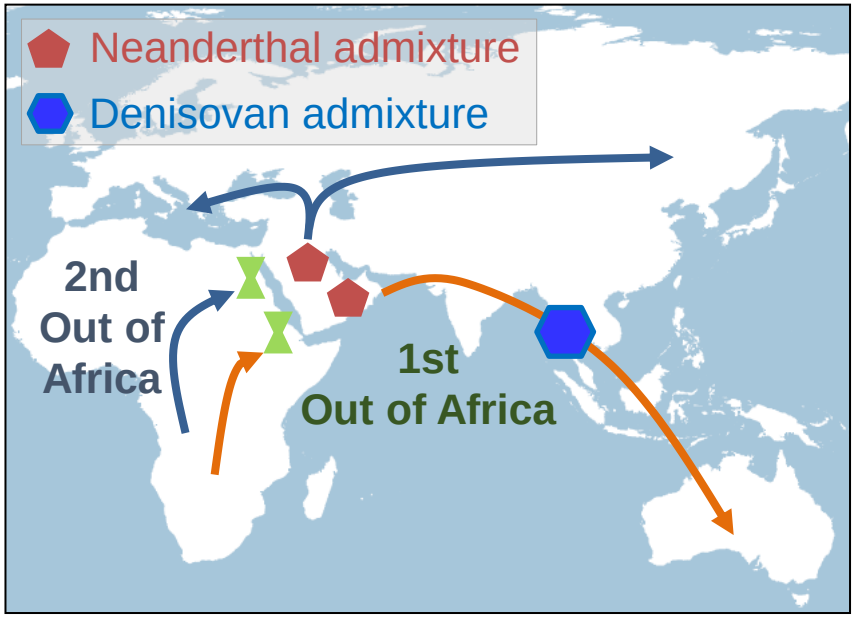
Two-waves out of Africa

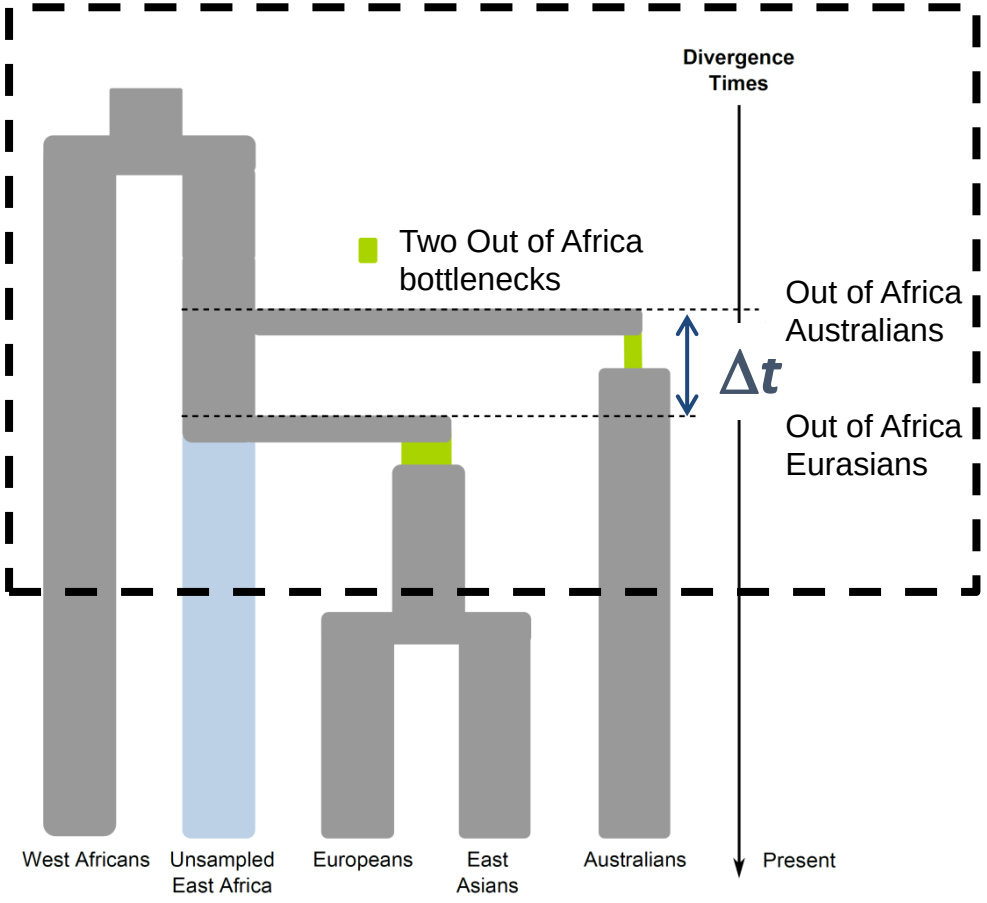




- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events

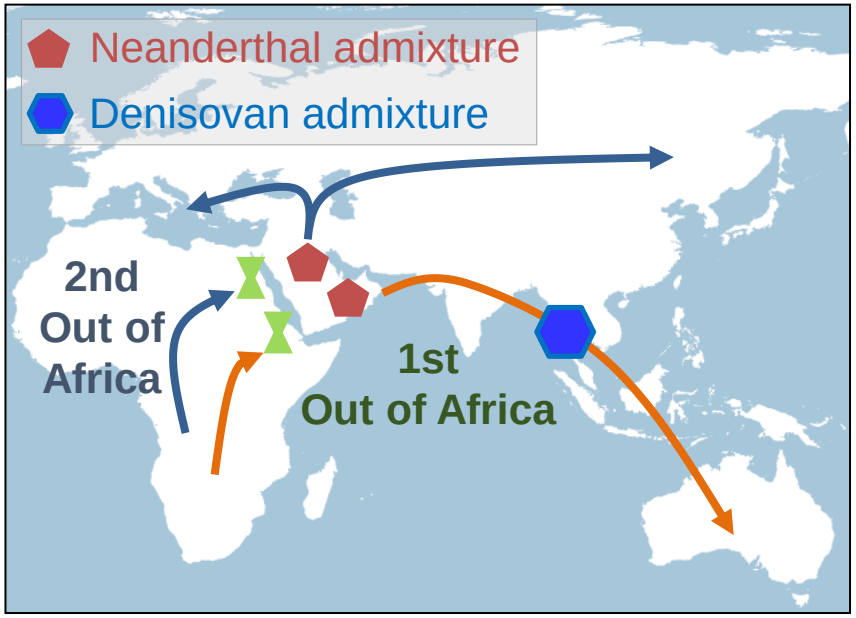
Two-waves out of Africa

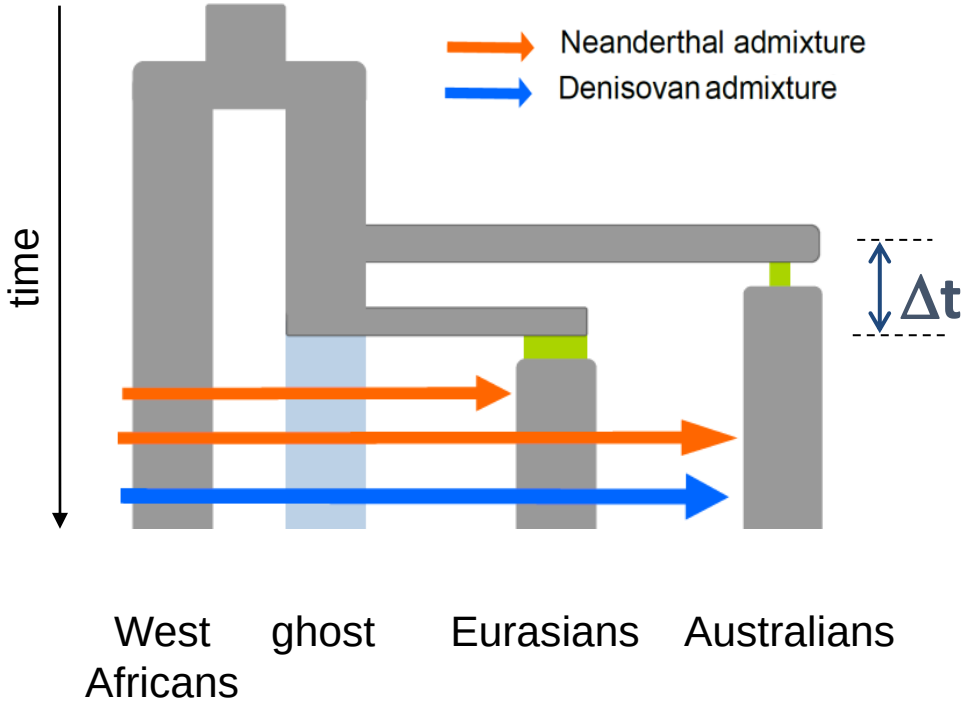




- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events

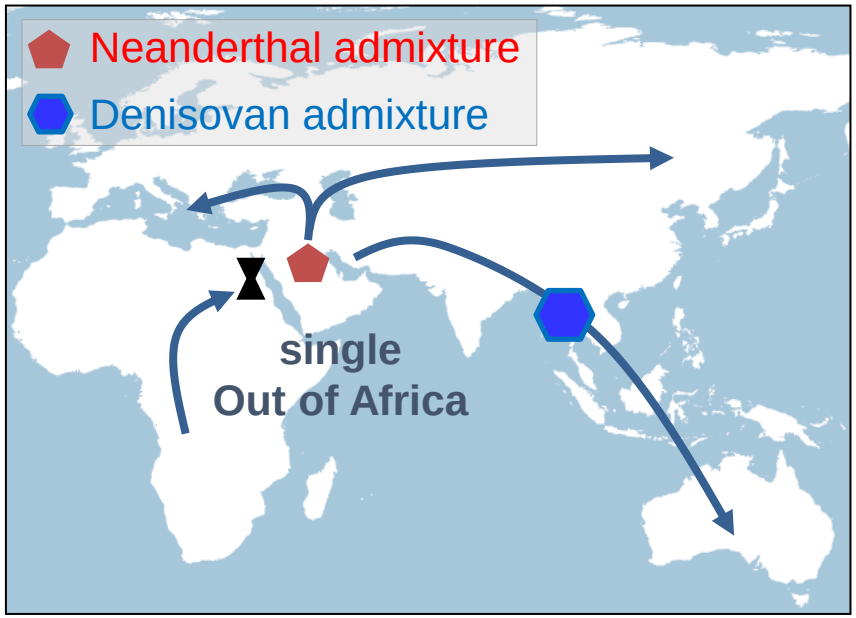
Two-waves out of Africa

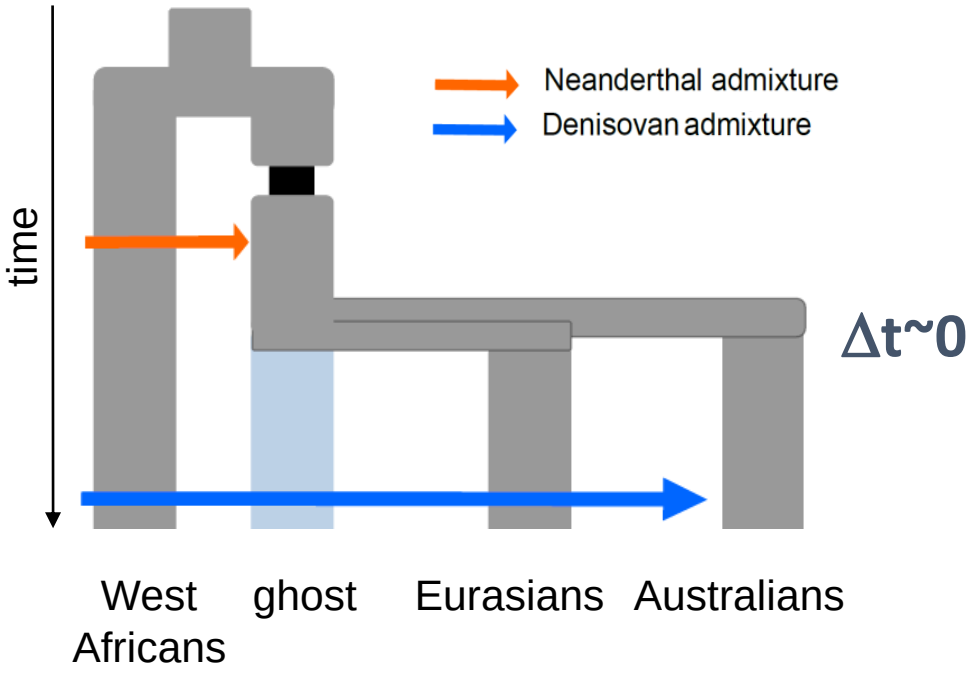




- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events

One wave out of Africa

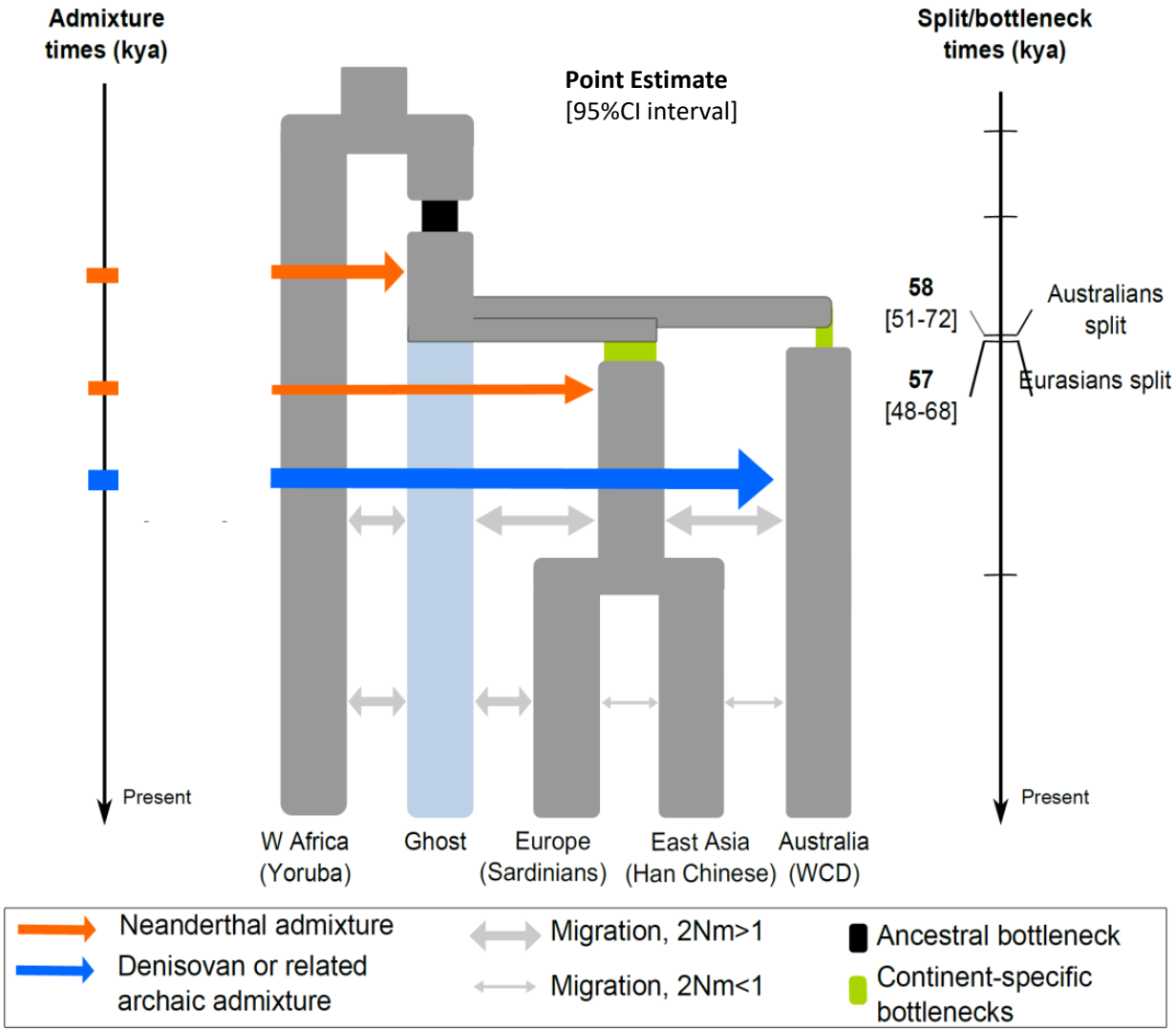




- Similar divergence times (Δt close to zero)
- One single bottlenecks associated with the Out of Africa events
- A major admixture pulse with Neanderthal

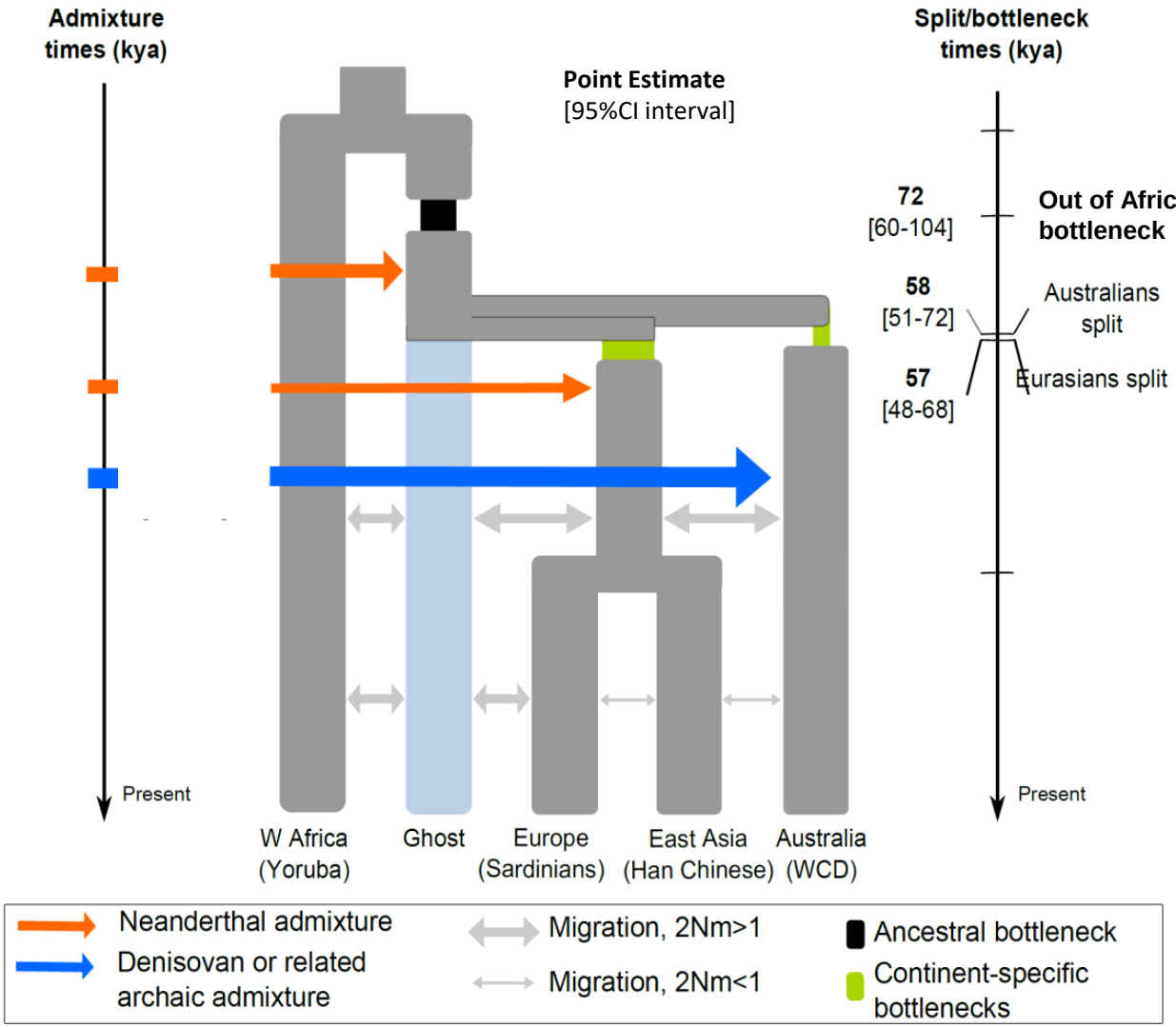
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)



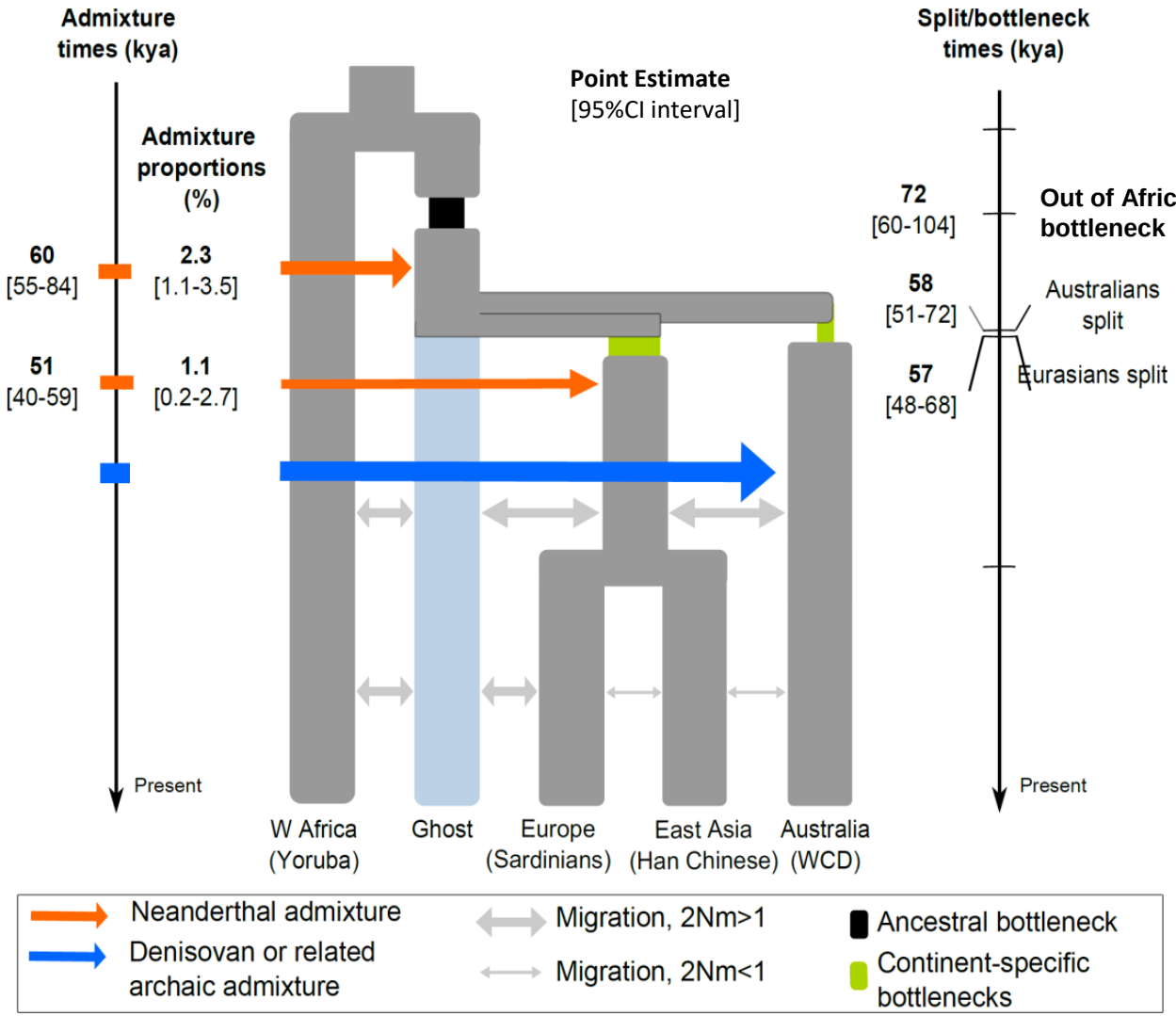
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)
- Bottleneck associated with the Out of Africa event



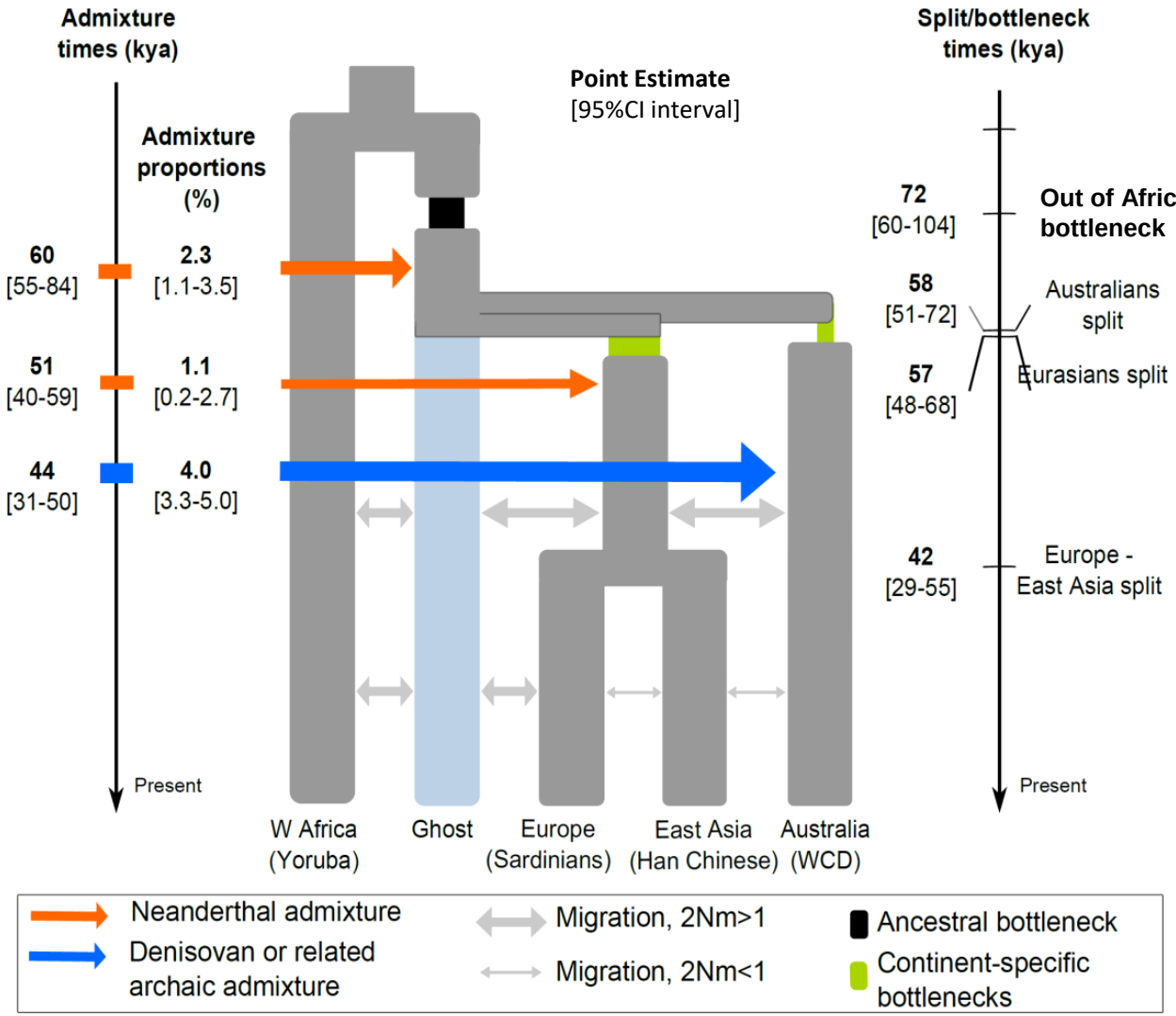
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)
- Bottleneck associated with the Out of Africa event
- A major admixture pulse with Neanderthal in ancestors of all non-Africans

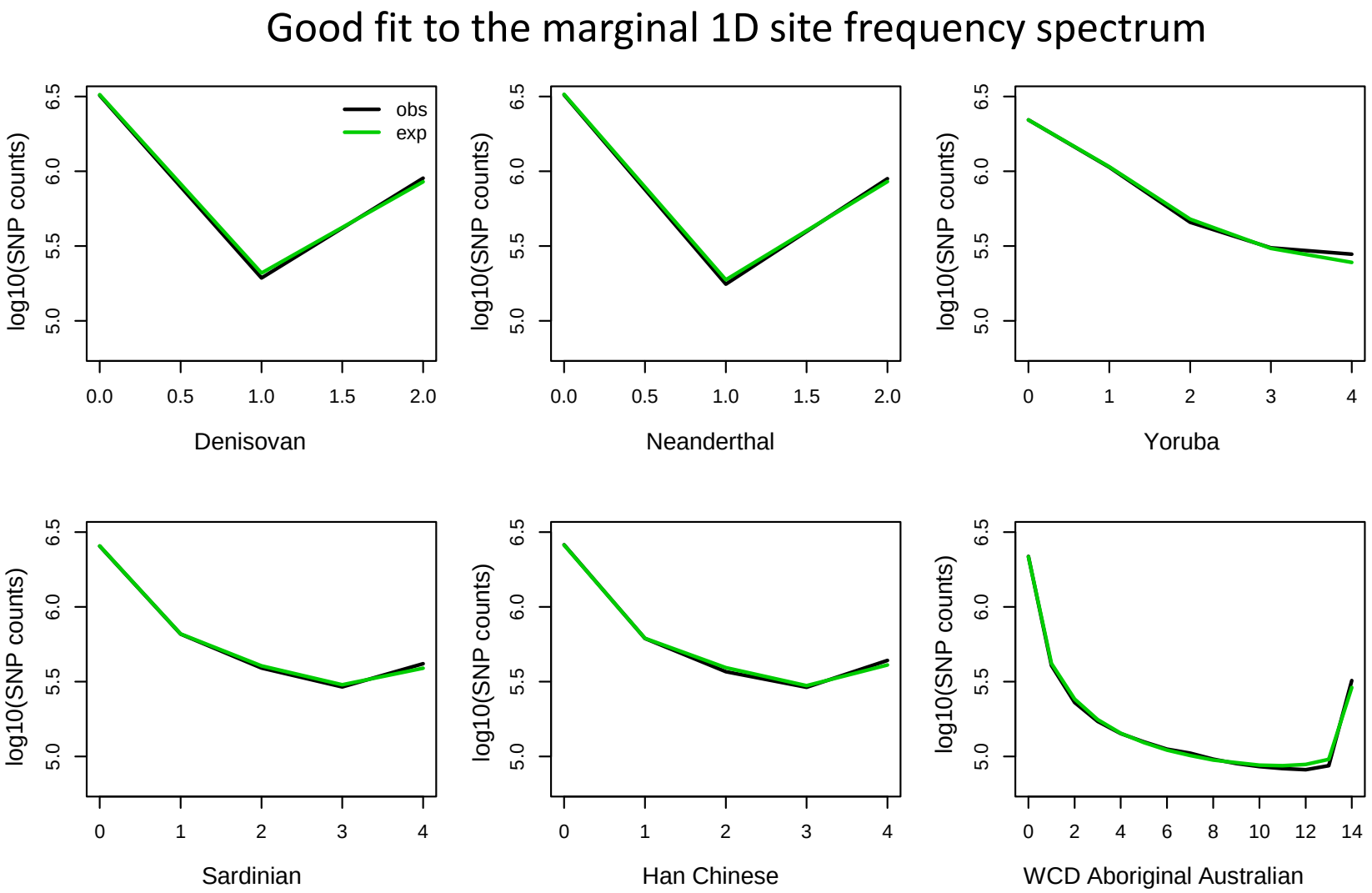


A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)
- Bottleneck associated with the Out of Africa event
- A major admixture pulse with Neanderthal in ancestors of all non-Africans



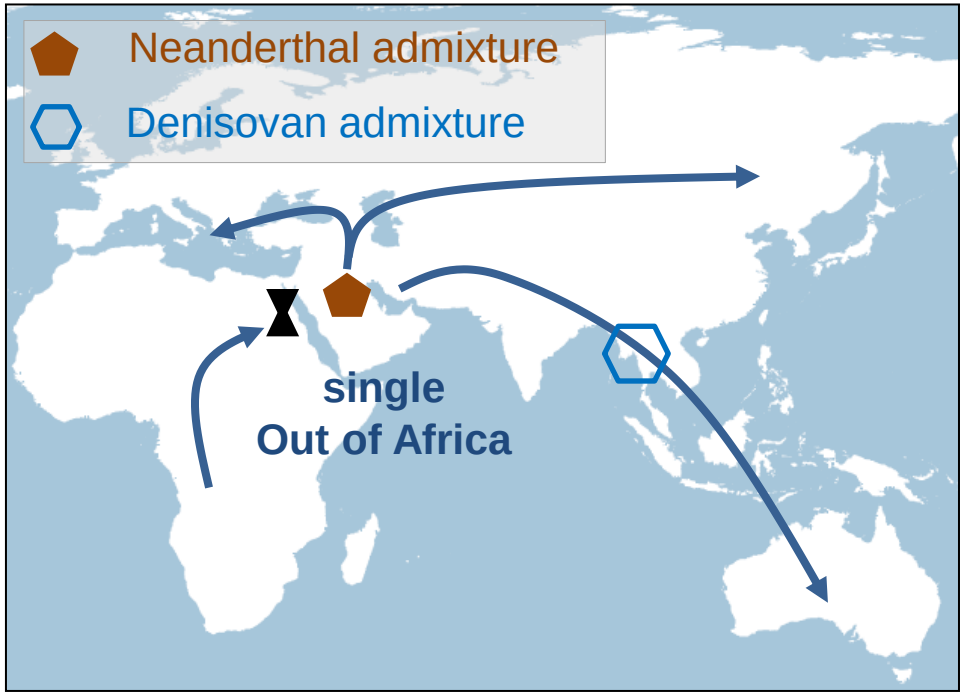
Model captures aspects about the observed data



Summary

Aboriginal Australians genomes support a single major wave out of Africa

- Accounting for archaic admixture with Neanderthal and Denisovan was crucial to understand population divergence
- Genomic data consistent with a single major dispersal event out of Africa (60-104 kya)
- Two major dispersal waves into Asia: Aboriginal Australians diverged 51-72 kya from Eurasians



Exercise 1
Simulate and draw a coalescent tree

With a piece of paper and using R studio to simulate random numbers draw a gene tree for a sample of 4 individuals taken from a population with size N=1000. Drawing trees in R can be complex. So, you will draw the tree in a piece of paper, but simulate the values for the time intervals using R.

1. Recall that the time intervals between coalescent events follow a geometric distribution. You can simulate random numbers from a geometric distribution with

```
rgeom(1, prob = coalescentProb)
```

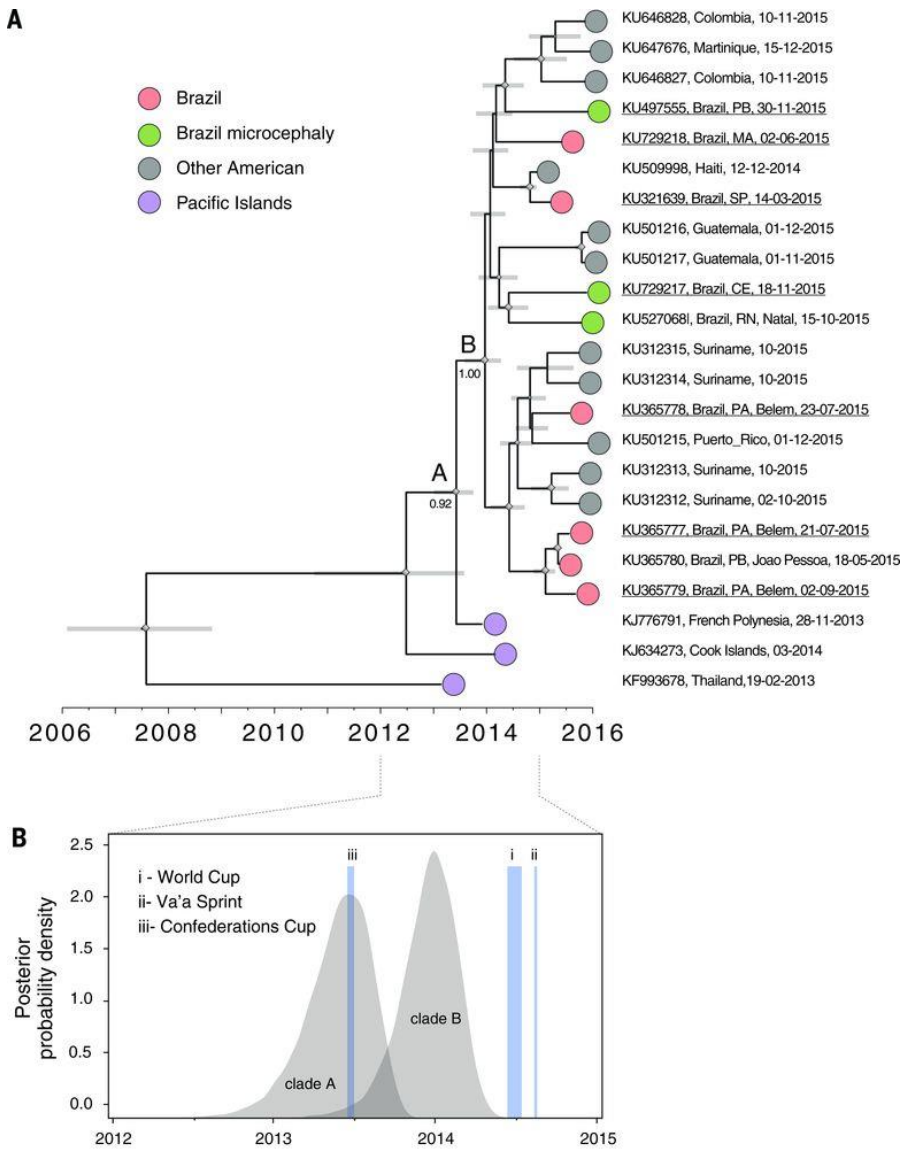
2. Save the time intervals T_4, T_3, T_2

3. You can decide what lineages to coalesce using the function sample()

```
sample(c(1,2,3,4), size=2, replace=F)
```

QUESTIONS:
- What is the TMRCA of your tree?
- What is the total branch length of your tree?
- Add two mutations to your tree. What is the more likely branch to have a mutation?

Examples of inferred gene trees



Zica virus estimated gene tree

Is this tree expected in a panmitic constant size population?

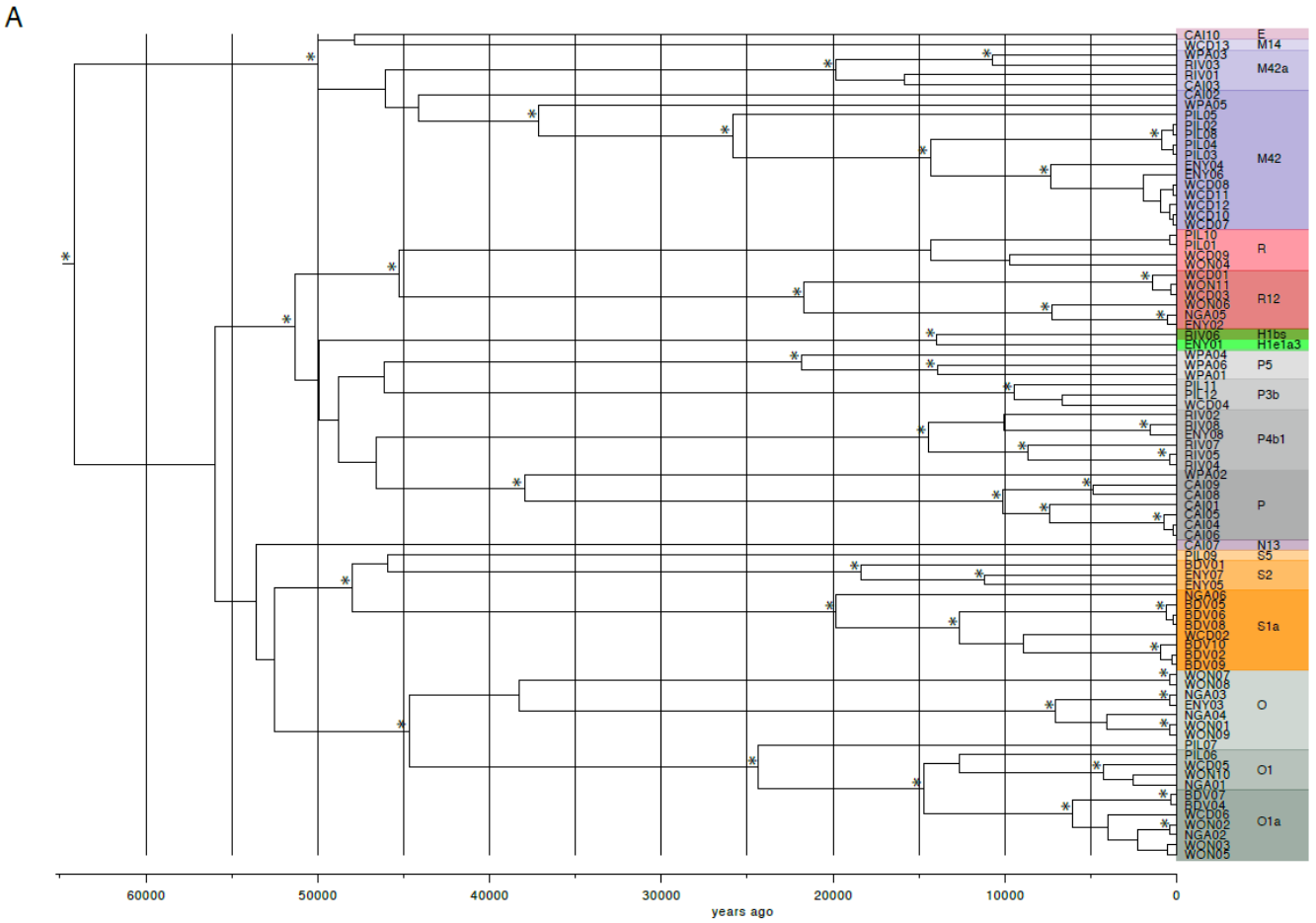
What does the TMRCA of each clade tell us?

Faria et al (2016) Science

Examples of inferred gene trees

Is this tree expected in a panmitic constant size population?

What does the TMRCA tell us?



mtDNA gene tree inferred for Aboriginal Australians

Malaspinas et al (2016) Nature