Demographic inference based on Site frequency spectrum (SFS) – Part I

Vitor Sousa

CE3C – center for ecology, evolution and environmental changes

2018 WSPG Cesky Krumlov
26 Jan 2018

vmsousa@fc.ul.pt

UNIVERSITÄT BERN

centre for ecology, evolution and environmental changes

SIB Swiss Institute of Bioinformatics

Outline

Part I

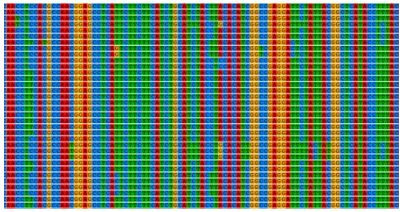
- Modeling demographic history: Population trees vs gene trees
- The SFS and coalescent trees
- Fastsimcoal2 principles

Part II

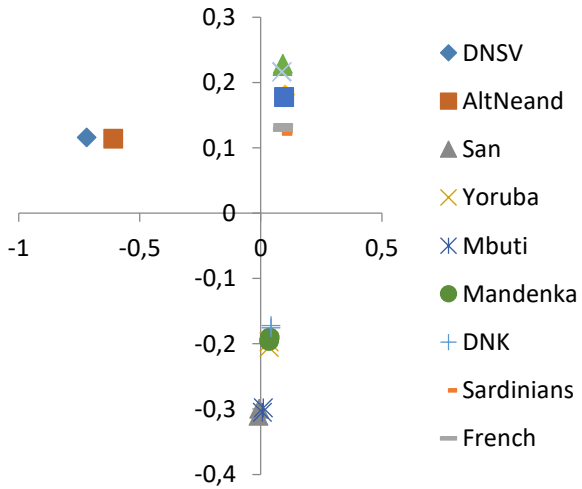
- Example of applications to different problems and types of data

Connecting observed patterns to the evolutionary processes

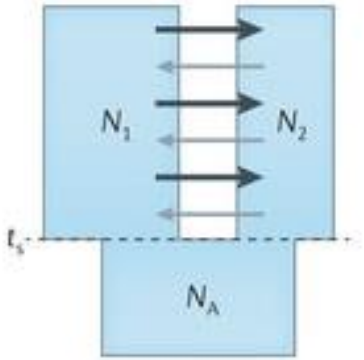
Genomic data



«Model-free» methods e.g. PCA



Model-based methods



Observed patterns

Evolutionary Processes:

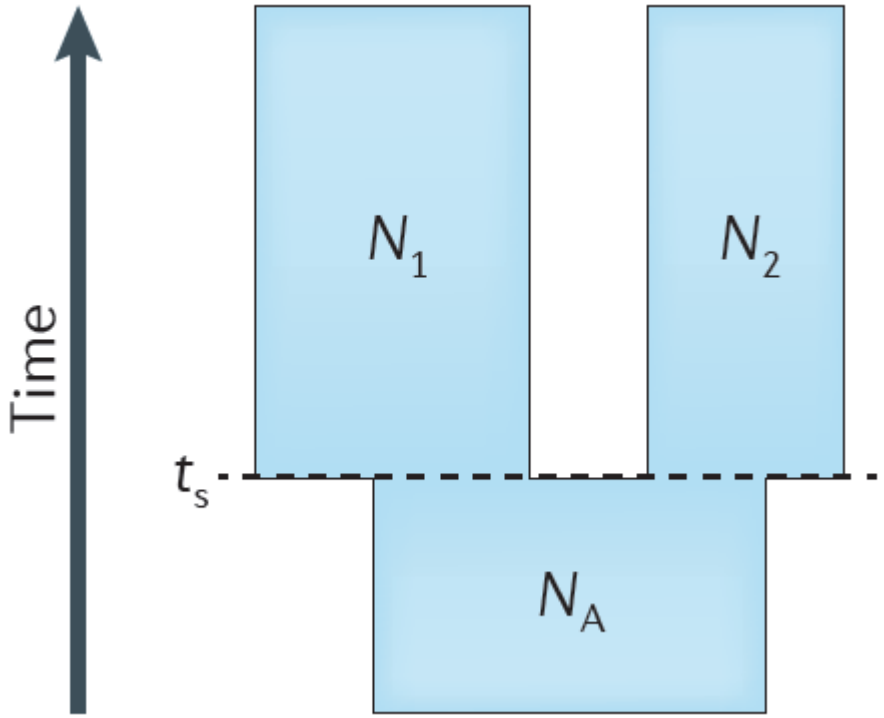
- Demography
- Selection
- Mutation
- Recombination

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

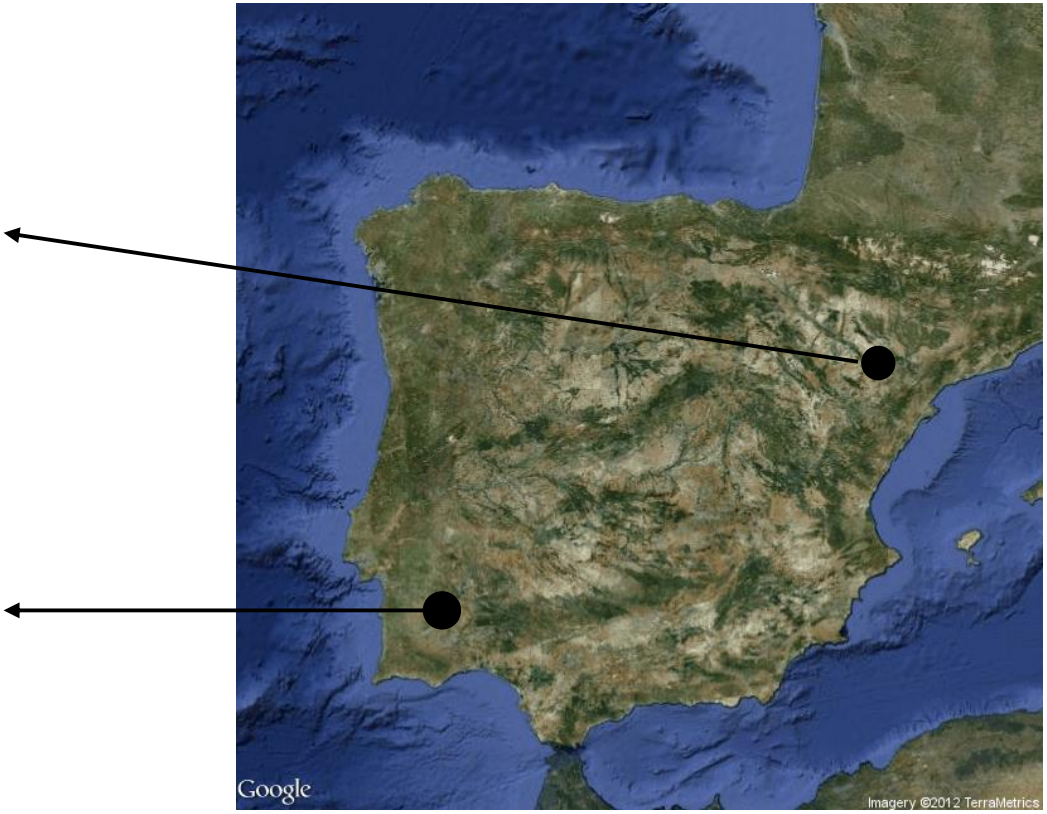
Population tree



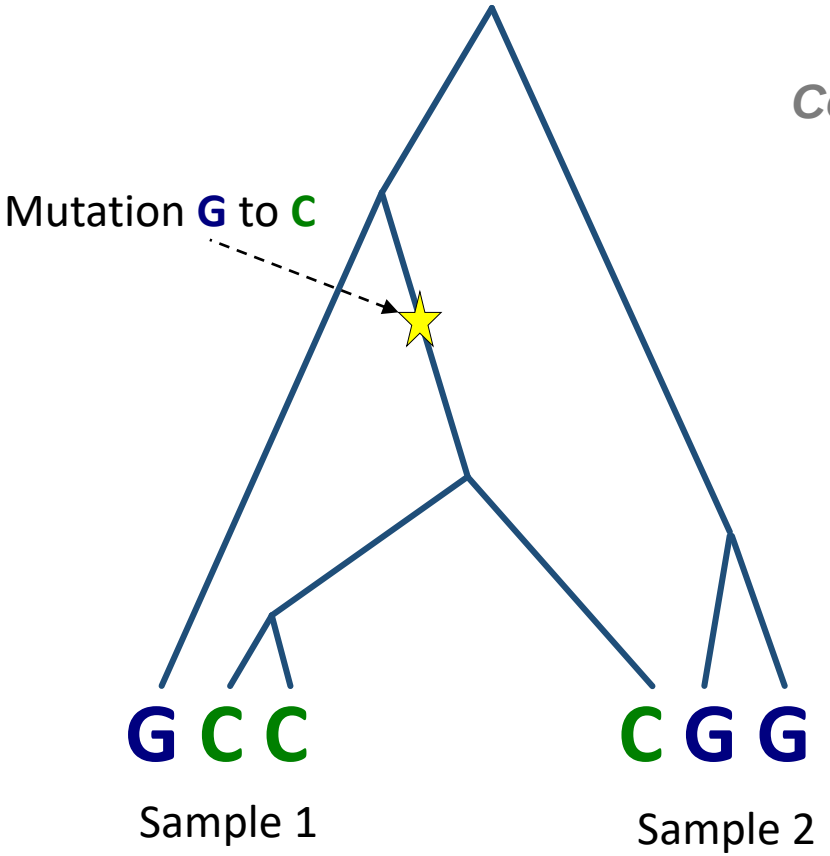
How to connect alleles with demographic history?

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

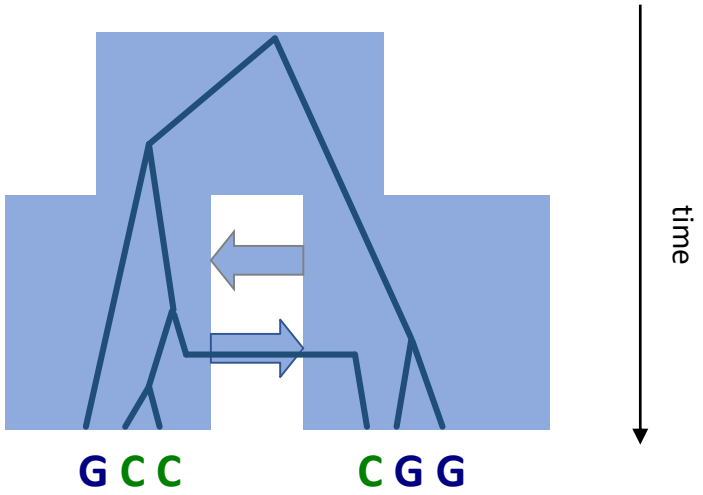


Gene trees reflect the species history

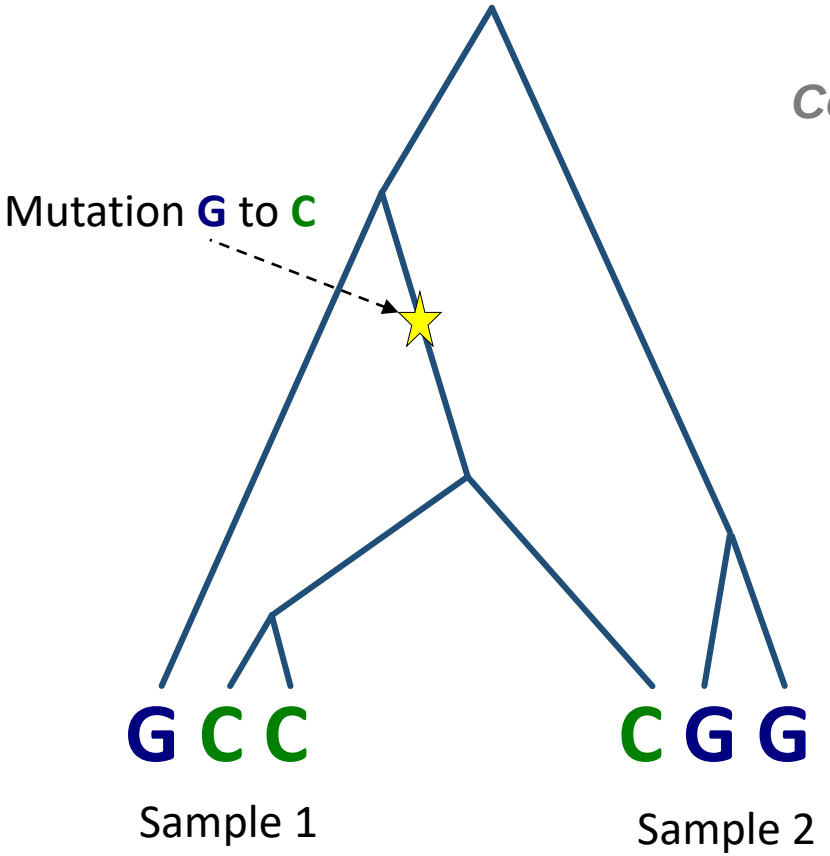


Population tree with migration

Compatible

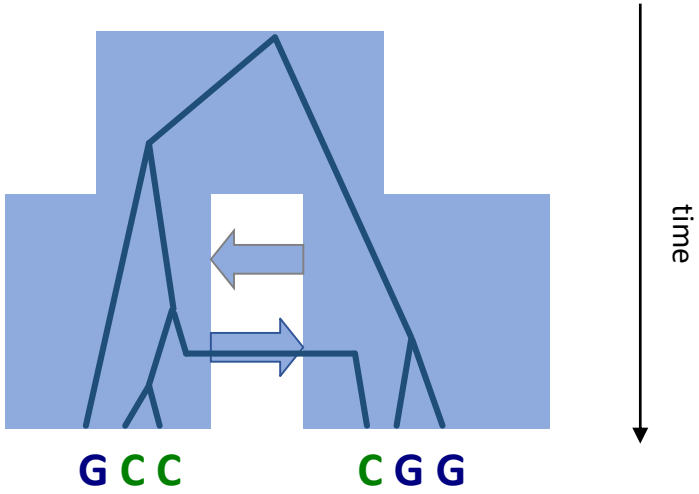


Gene trees reflect the species history

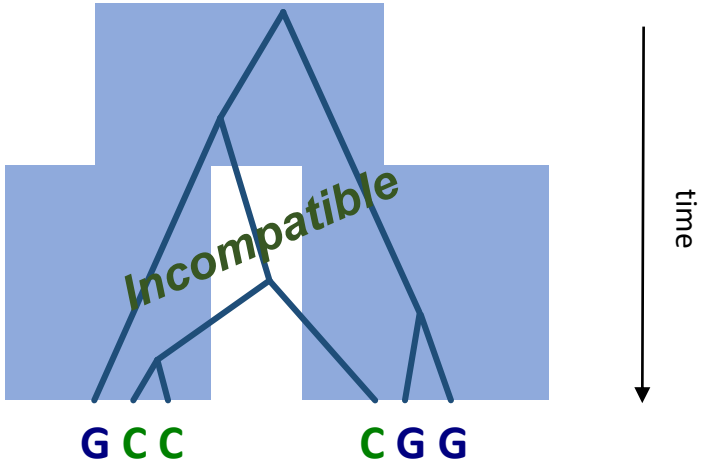


Population tree with migration

Compatible



Population tree with no migration



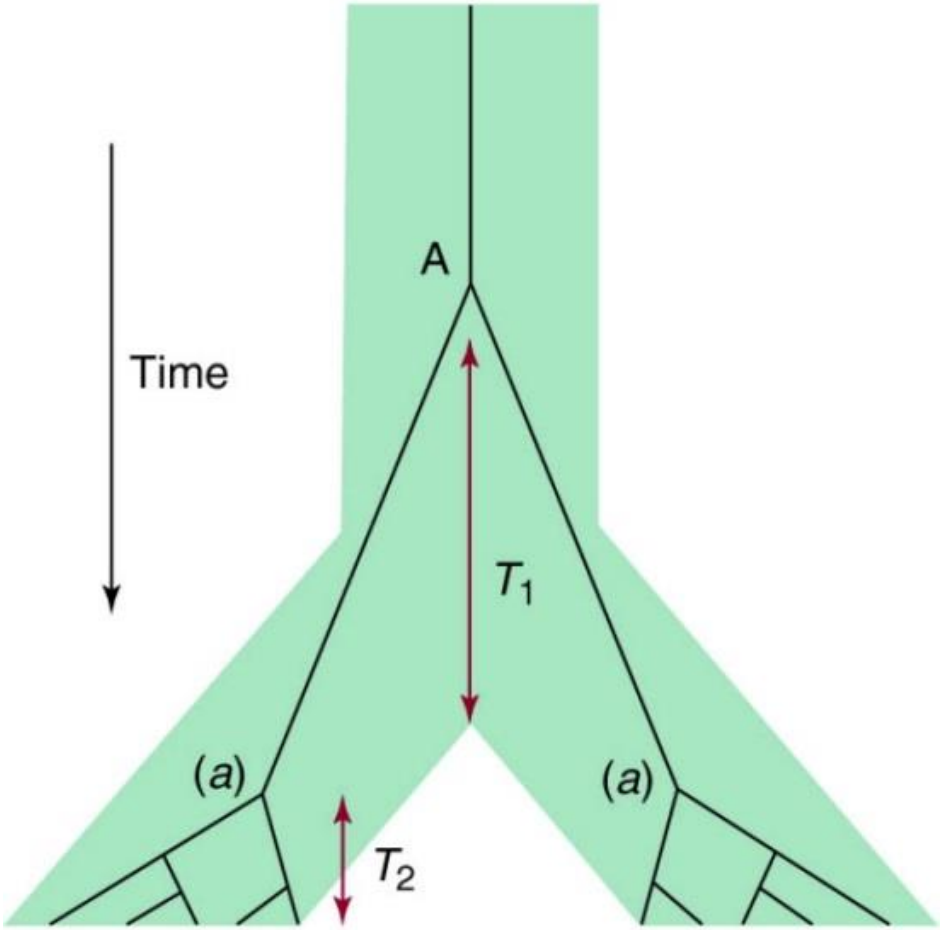
Gene trees vs. Population trees

Gene trees reflect the ancestry of genes within populations.

The relationship between populations is given by the **population tree**.

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

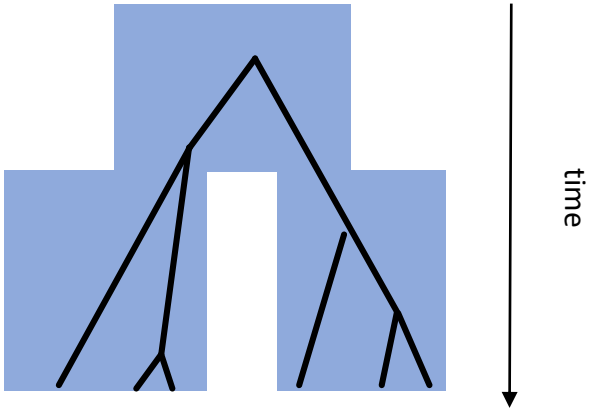
However, in the time scale of population genetics, gene trees at a particular region of the genome (locus) can be very different from the population tree.



Nichols (2001) TREE

Reconstructing the demographic history from genomic data

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

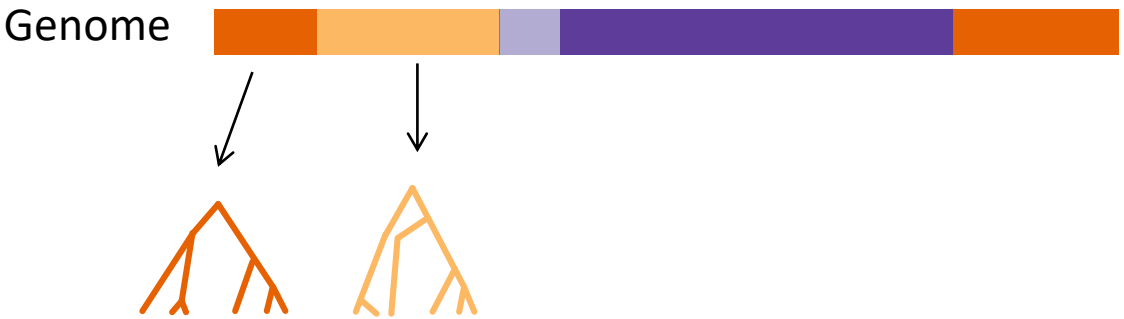


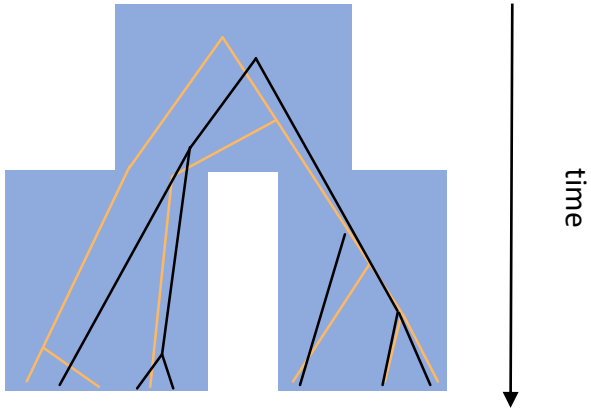
All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

- 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



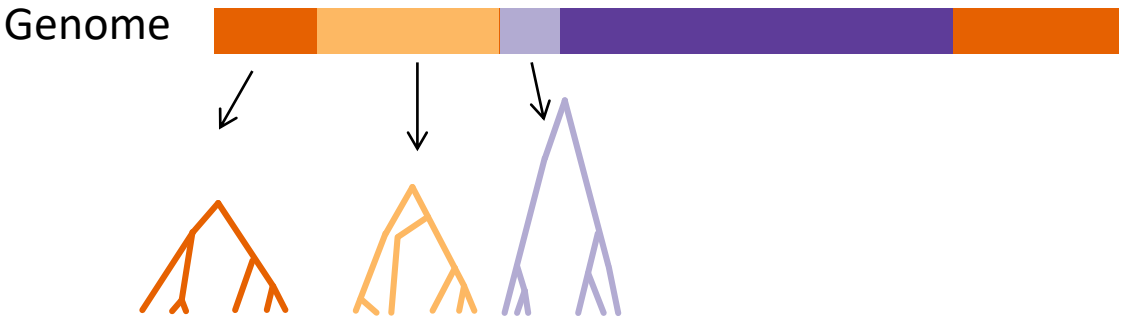


All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

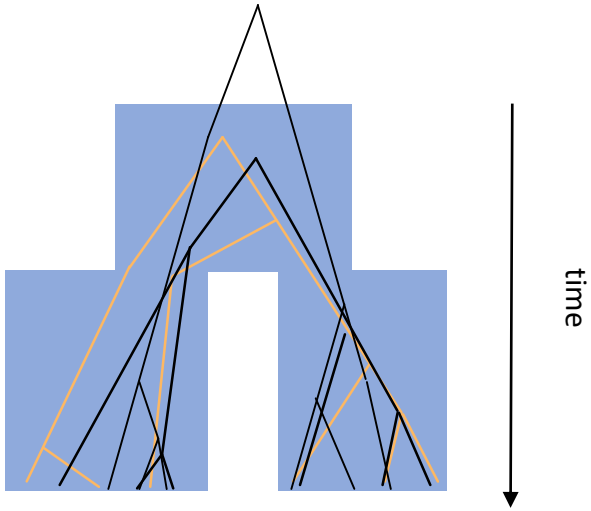
- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

Reconstructing the demographic history from genomic data

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



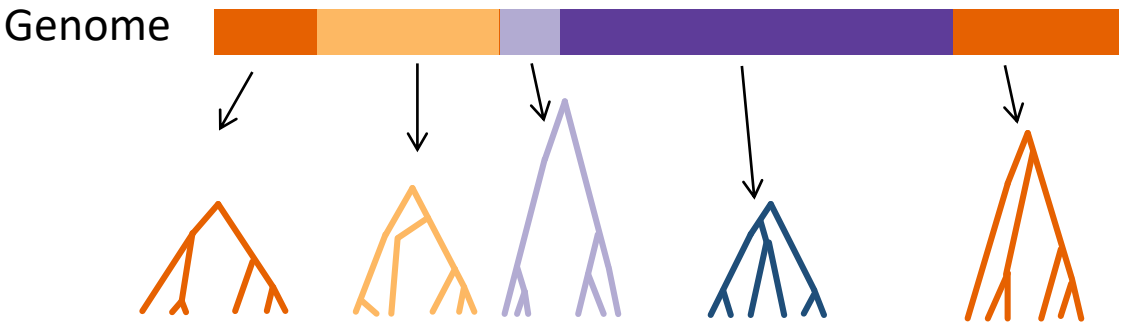
- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions



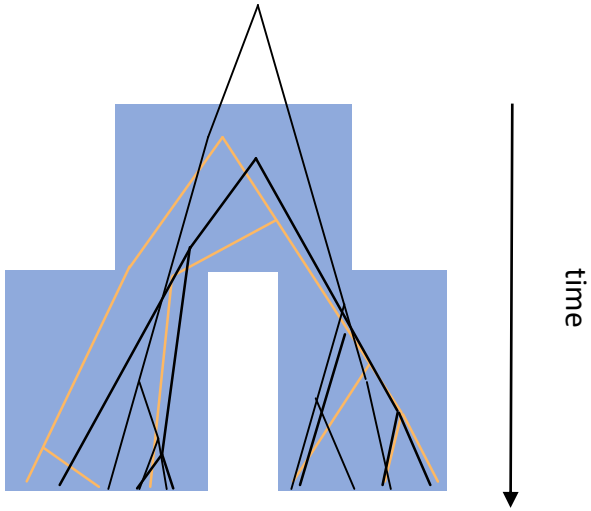
All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

Reconstructing the demographic history from genomic data

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



- 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.

Coalescent Theory

Classical population genetics theory tries to predict what will happen in the future of a given population. It is a **prospective approach**.

Coalescent theory is a retrospective approach to population genetics that describes the ancestry of sampled of genes, as a function of an underlying model (demographic history).

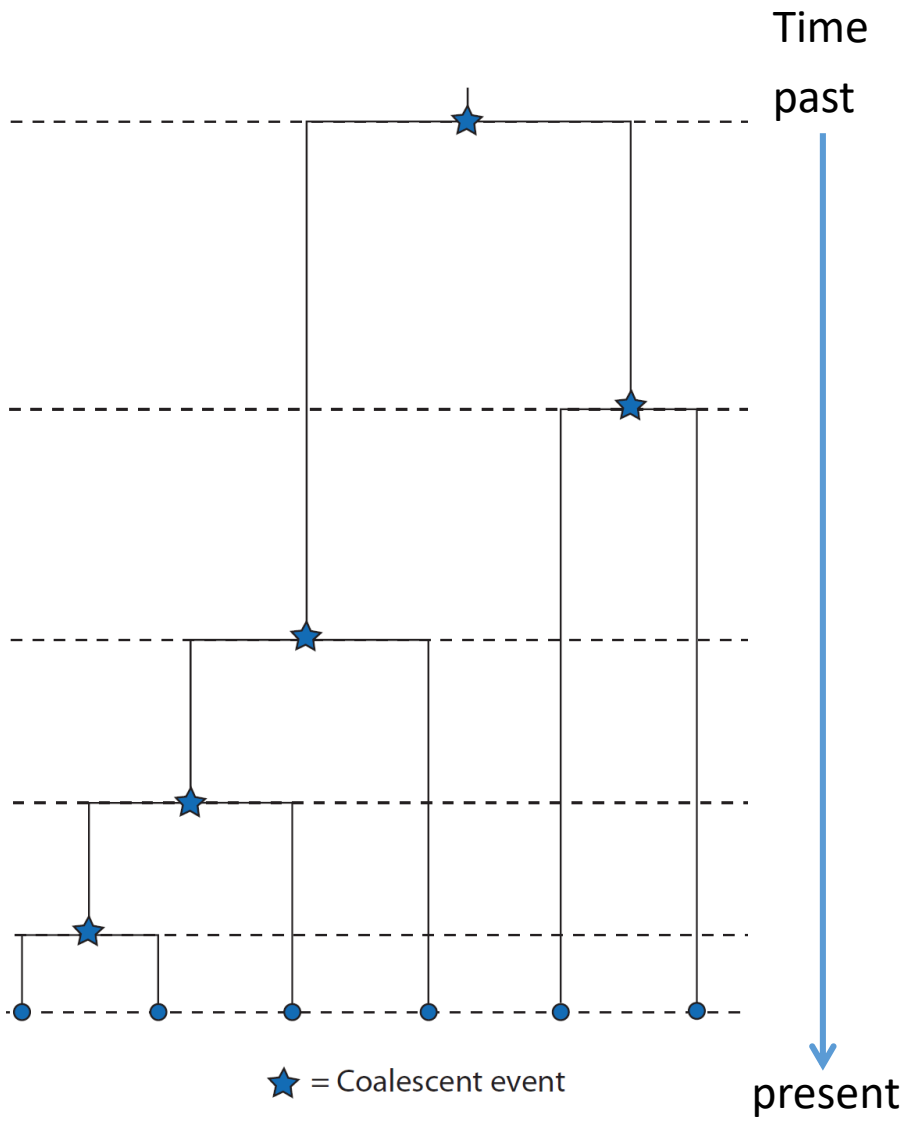
Coalescent theory provides a model for the **evolutionary forces** (genetic drift, migration, mutation) and **the sampling process** that affect the observed data.

This process has been formalized by J.C. Kingman in a series of seminal papers published in 1982, and independently formalized by Hudson (1983) and Tajima (1983).

The principle of the coalescent theory

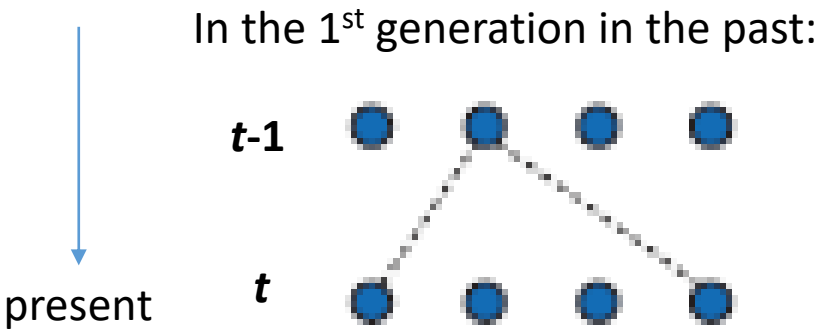
- For a given sample of individuals (gene copies or lineages) there is always a **gene tree** describing the ancestry of the sample.

- 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)

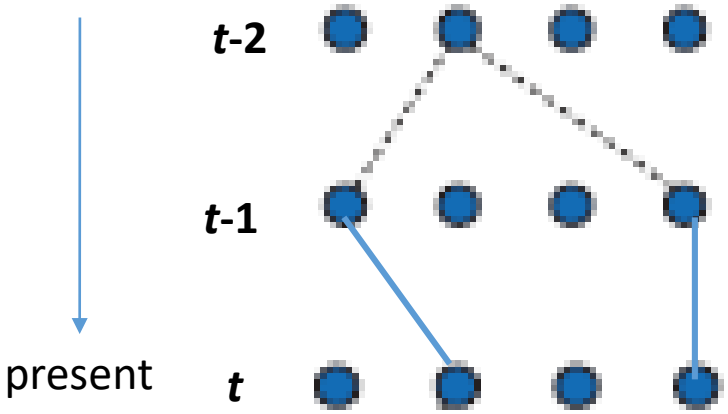


Pr(coal) = 1/2N
Pr(no coal) = 1-(1/2N)

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)

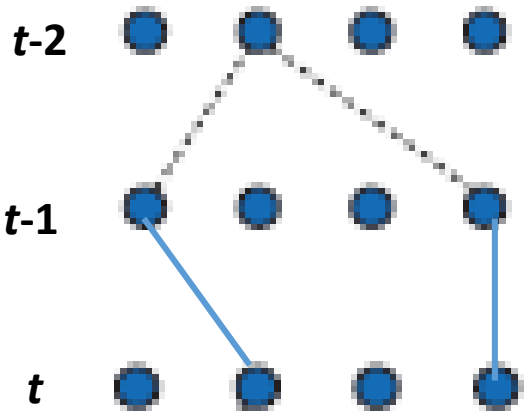
No coalescent
Pr(no coal) = 1-(1/2N)

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):

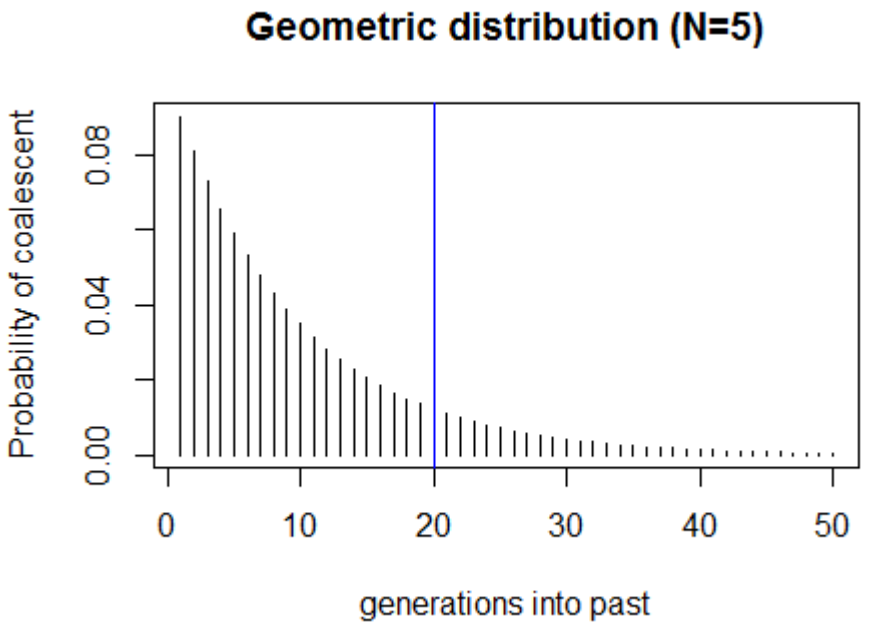
$$\Pr(\text{ coal at generation } t\ ) = \left(1-\frac{1}{2N}\right)^{t-1}\frac{1}{2N}$$

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:

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

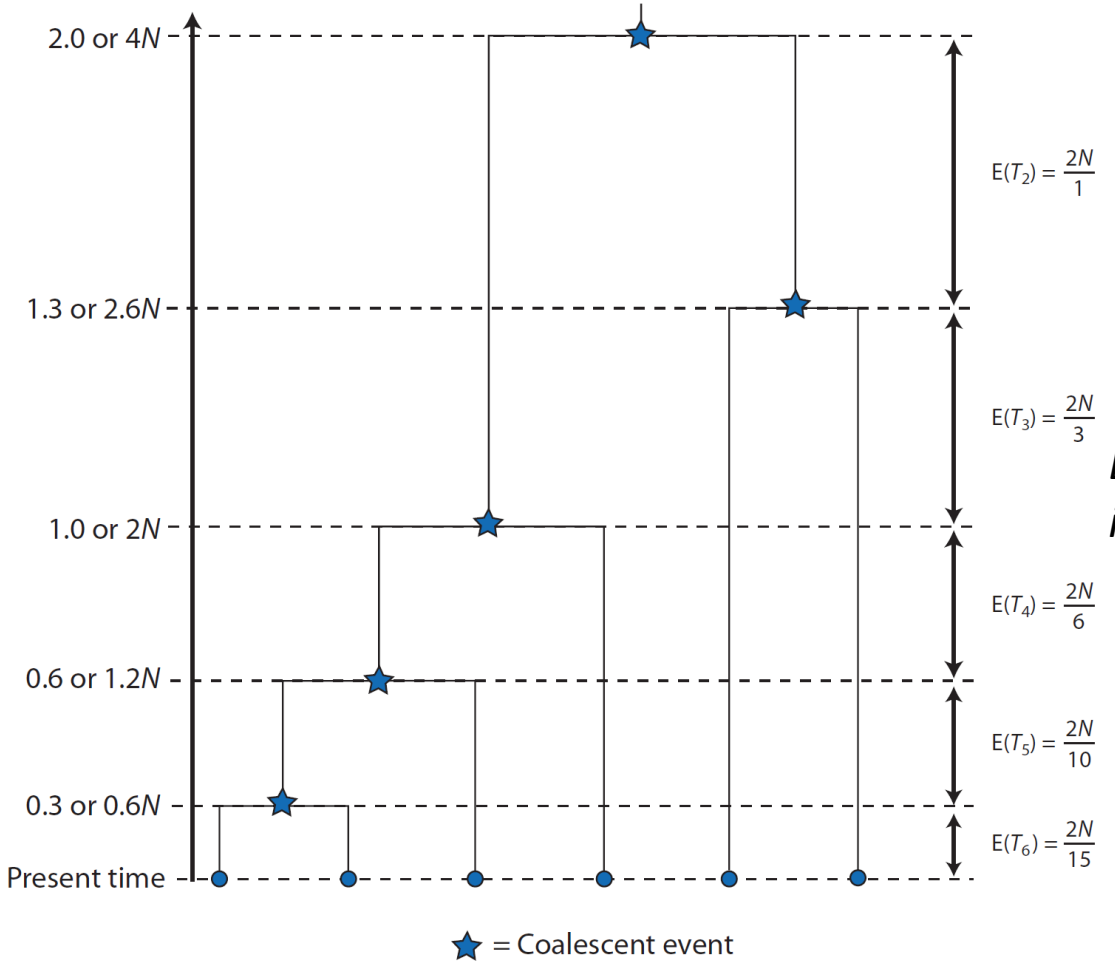
(HUGE) Variance:

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

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

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?

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

Five independent genomic regions from the **same constant size population**.

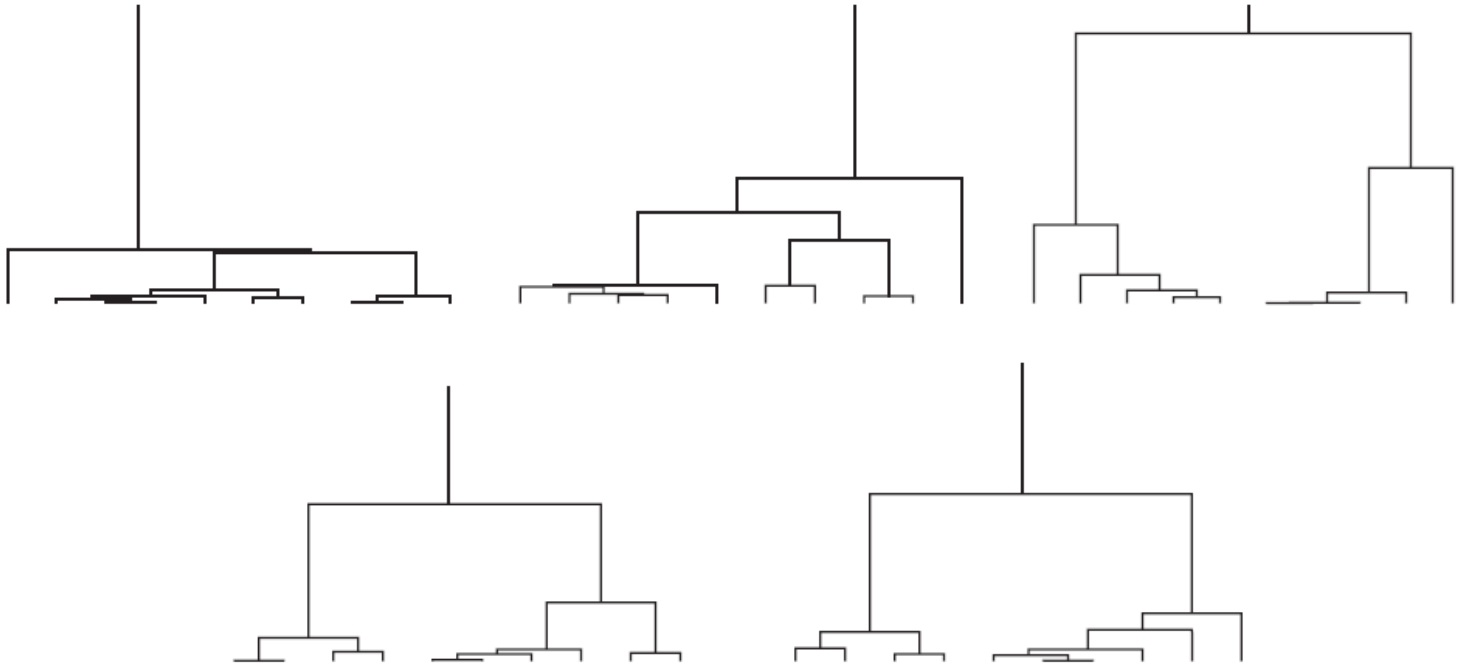
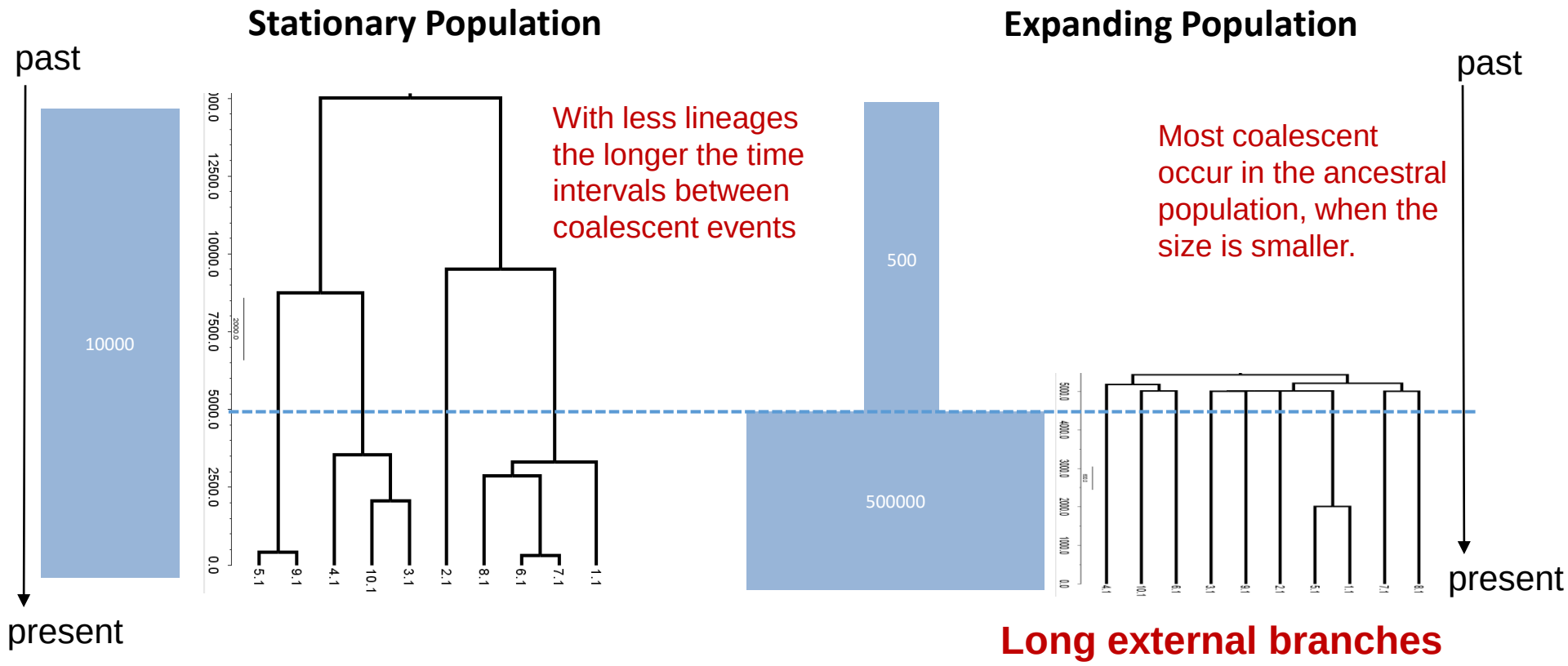


Figure 4.2 Five replicates of the coalescent process with constant population size for a sample of ten genes. Note the large variance in the time of the MRCA among replicates.

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

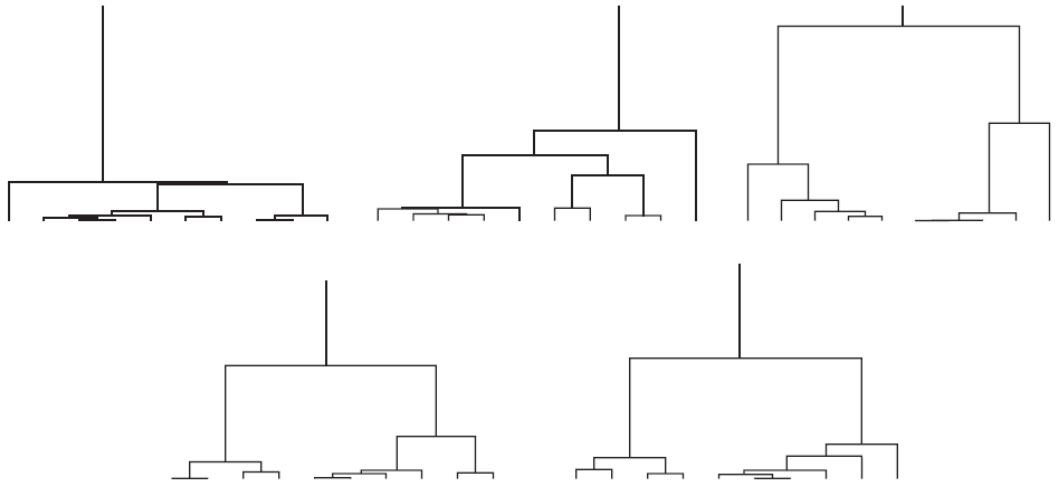


Figure 4.2 Five replicates of the coalescent process with constant population size for a sample of ten genes. Note the large variance in the time of the MRCA among replicates.

Expanding population
gene trees at five genome regions

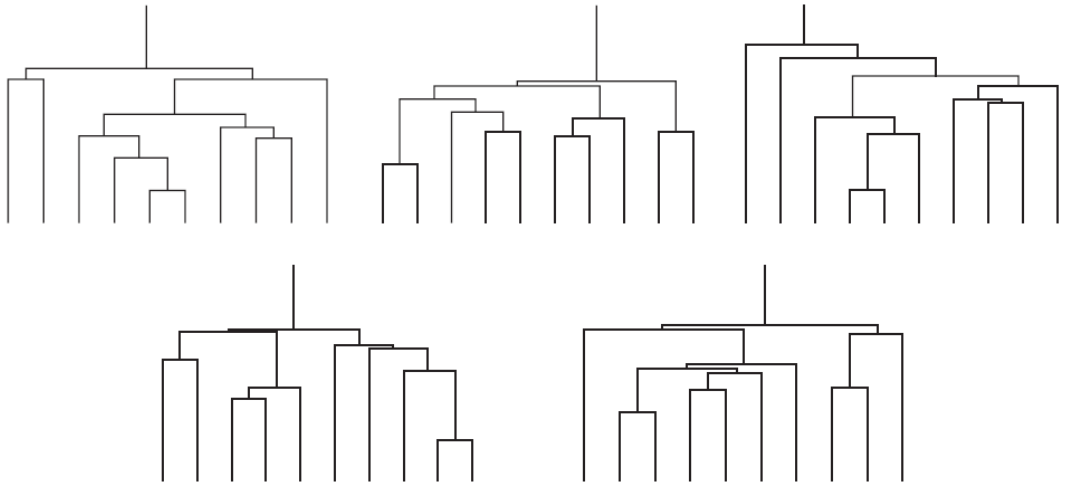
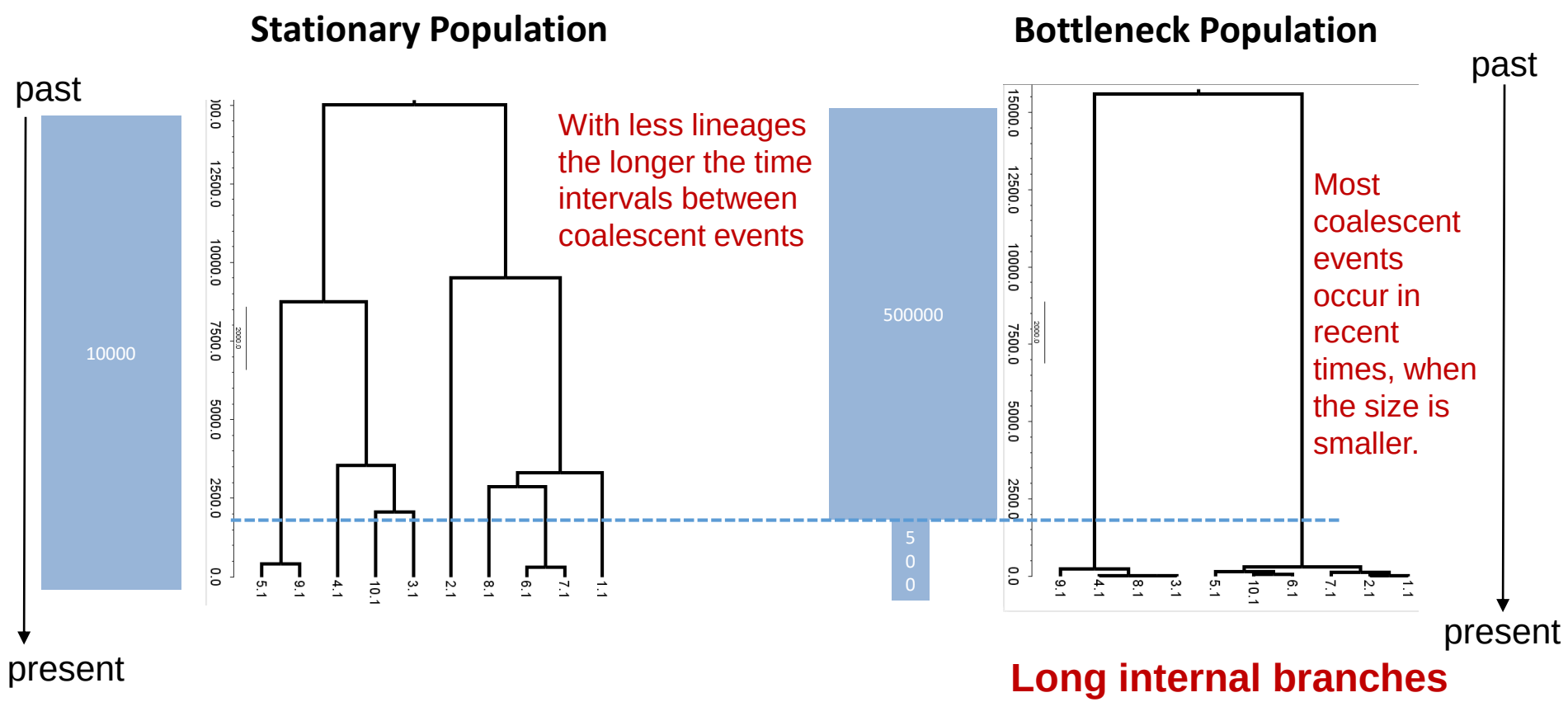


Figure 4.3 Five replicates of the coalescent with exponential growth, $\beta = 1000$, for a sample of $n = 10$ genes. Note the smaller variance in the time until the MRCA compared to the same quantity in Figure 4.2.

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

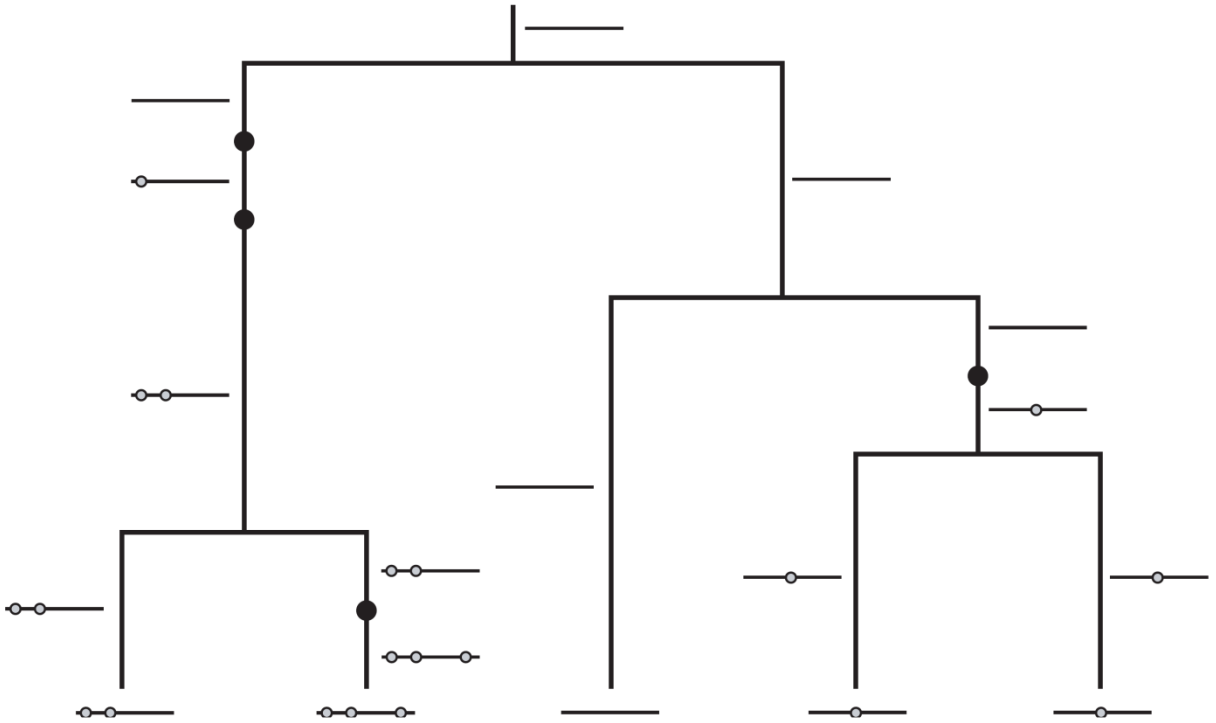
Gene trees for decreasing populations



- 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.

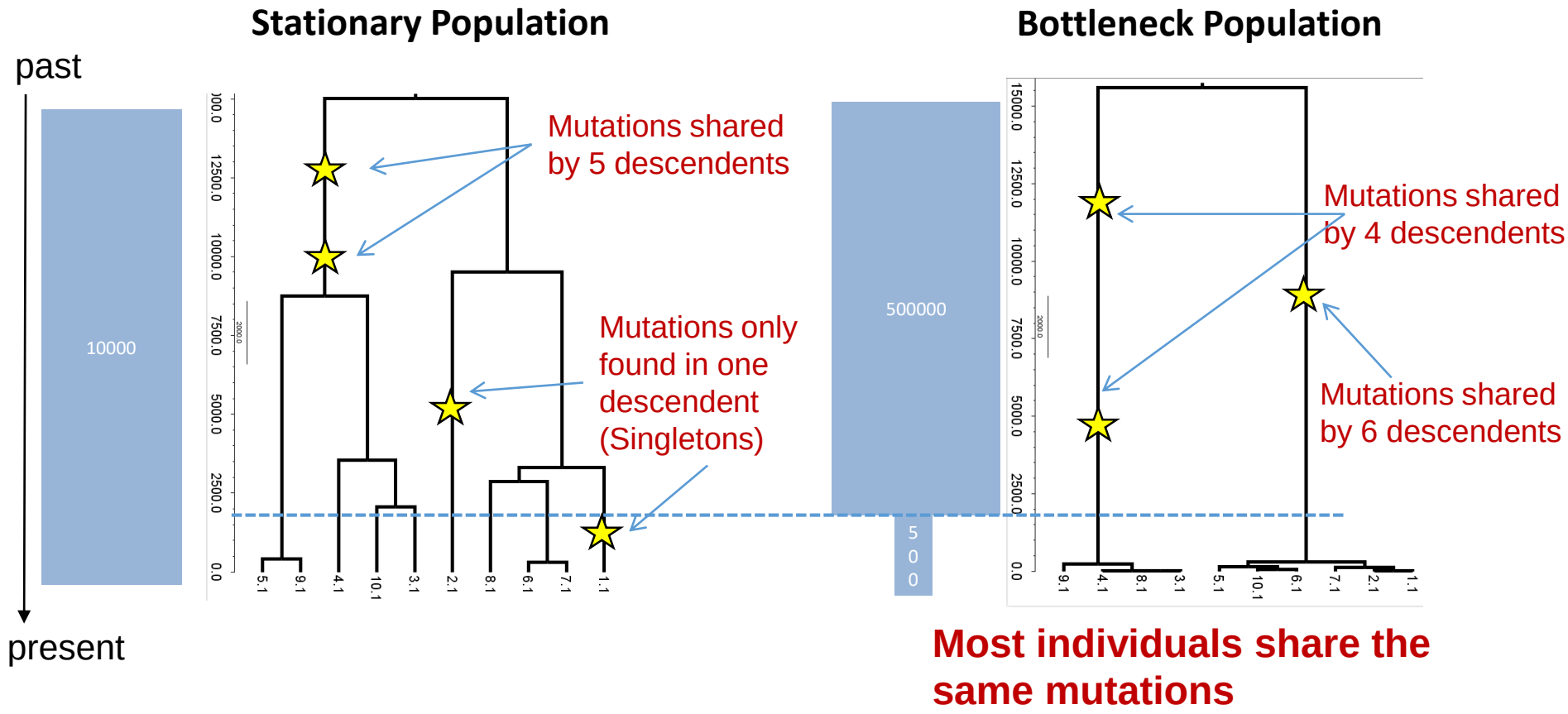


Mutations just accumulate along the branches of the tree according to a **Poisson process** with rate $\lambda_i = \mu t_i$ for the *i*-th branch of length t_i. The Poisson process is stochastic but it should be immediately **obvious that long branches will carry more mutations than short branches**

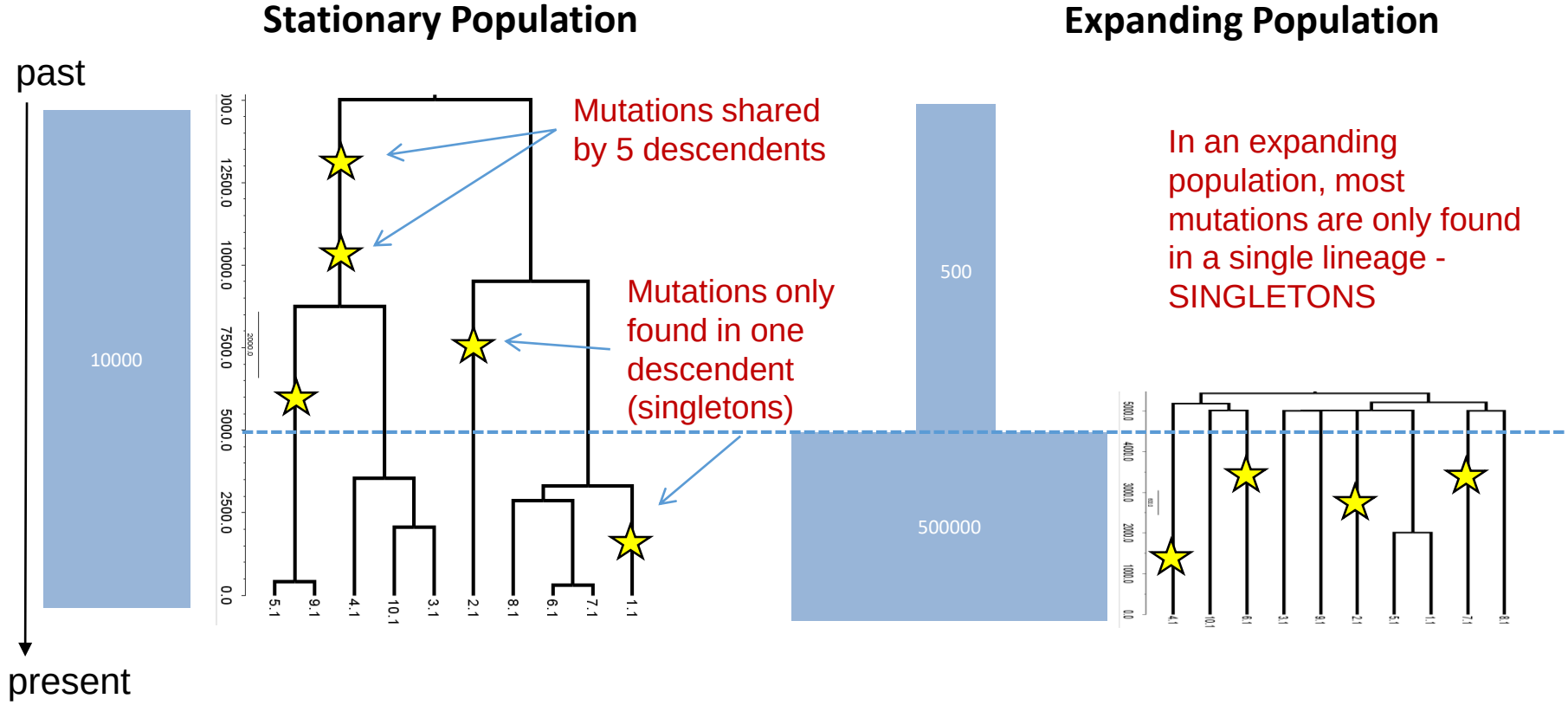
Hein et al (2004) Gene Genealogies, Variation and evolution

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.

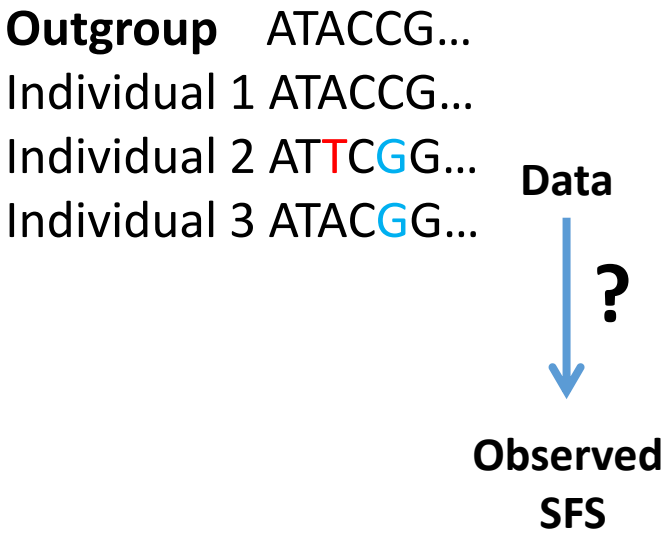


We expect less diversity in a bottlenecked population



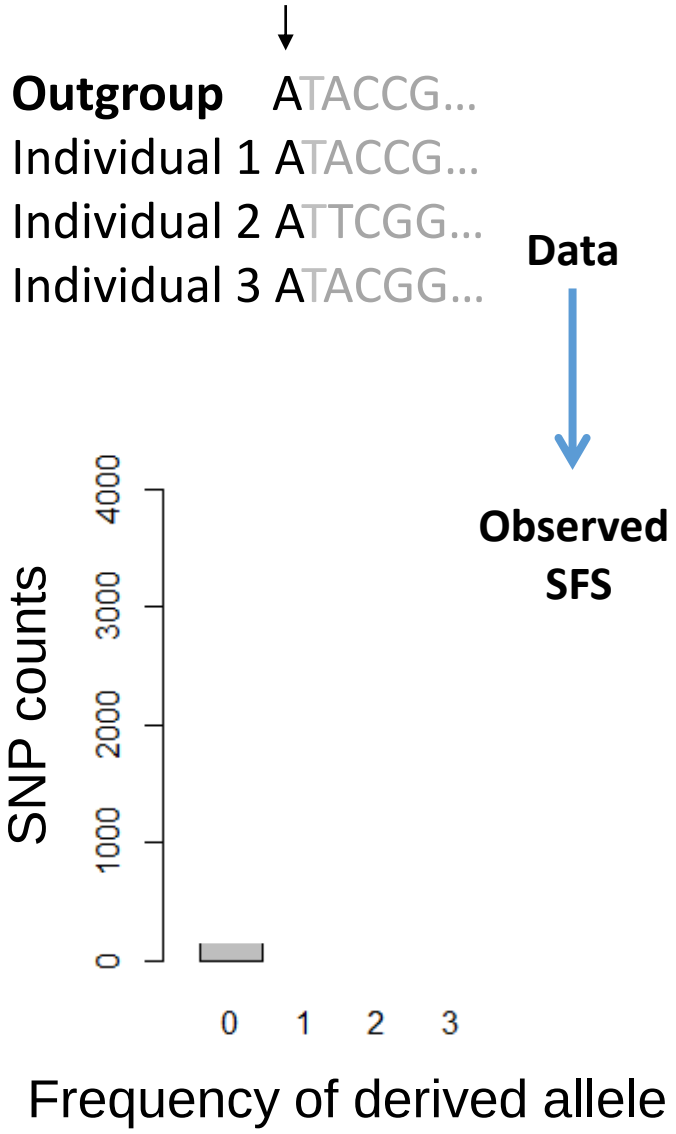
Site frequency spectrum (SFS)

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



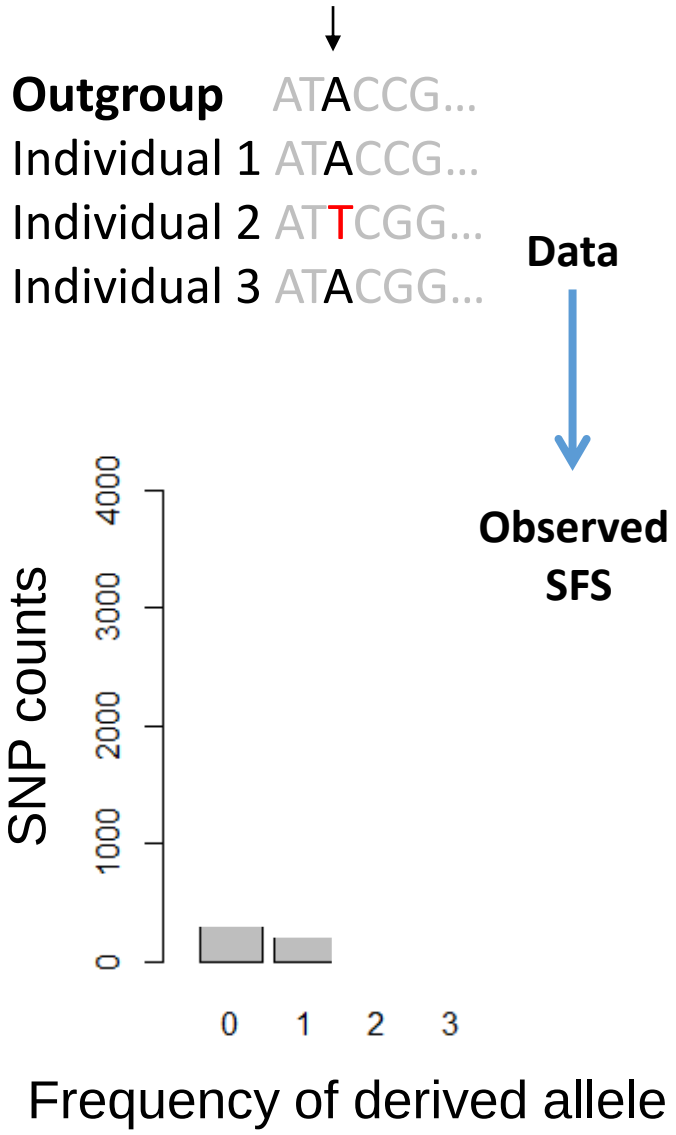
Site frequency spectrum (SFS)

- 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.

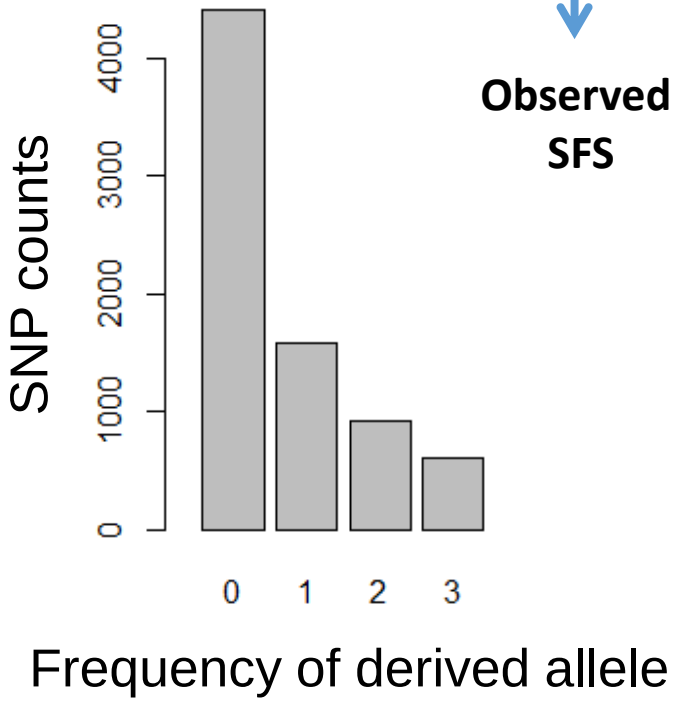
In a stationary population, the expected SFS relative frequencies are given by:

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

Fu and Li, 1993

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

Data



Observed SFS

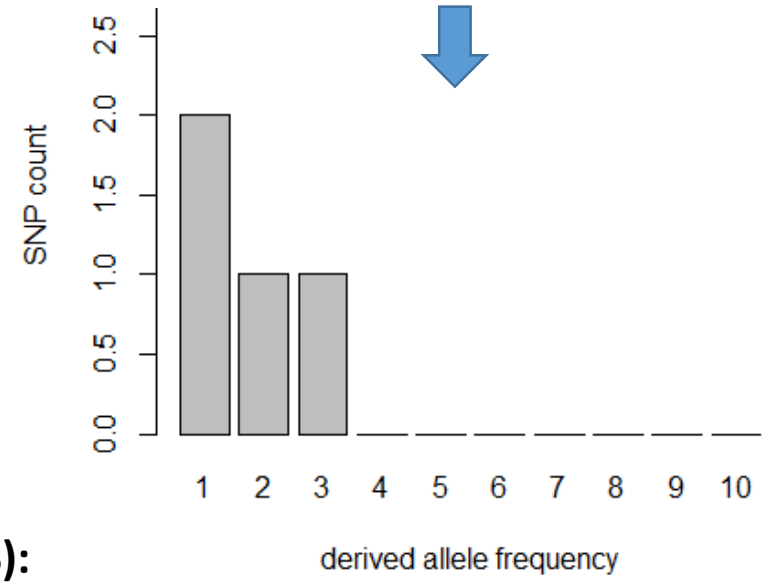
We can obtain the SFS from genotype call data

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



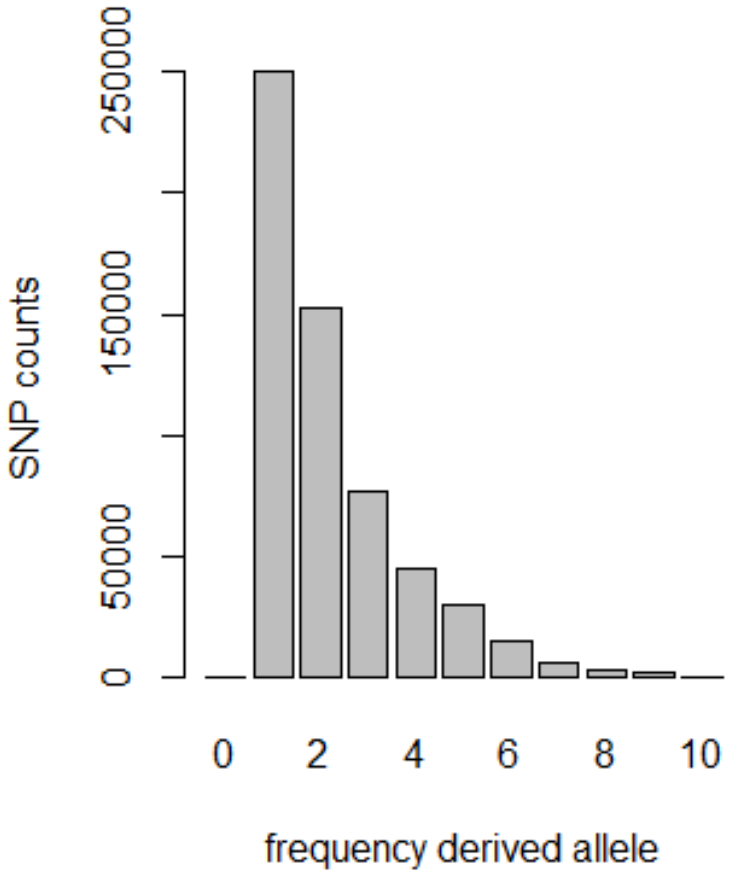
Observed SFS is a vector (1 dimensional SFS):

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.

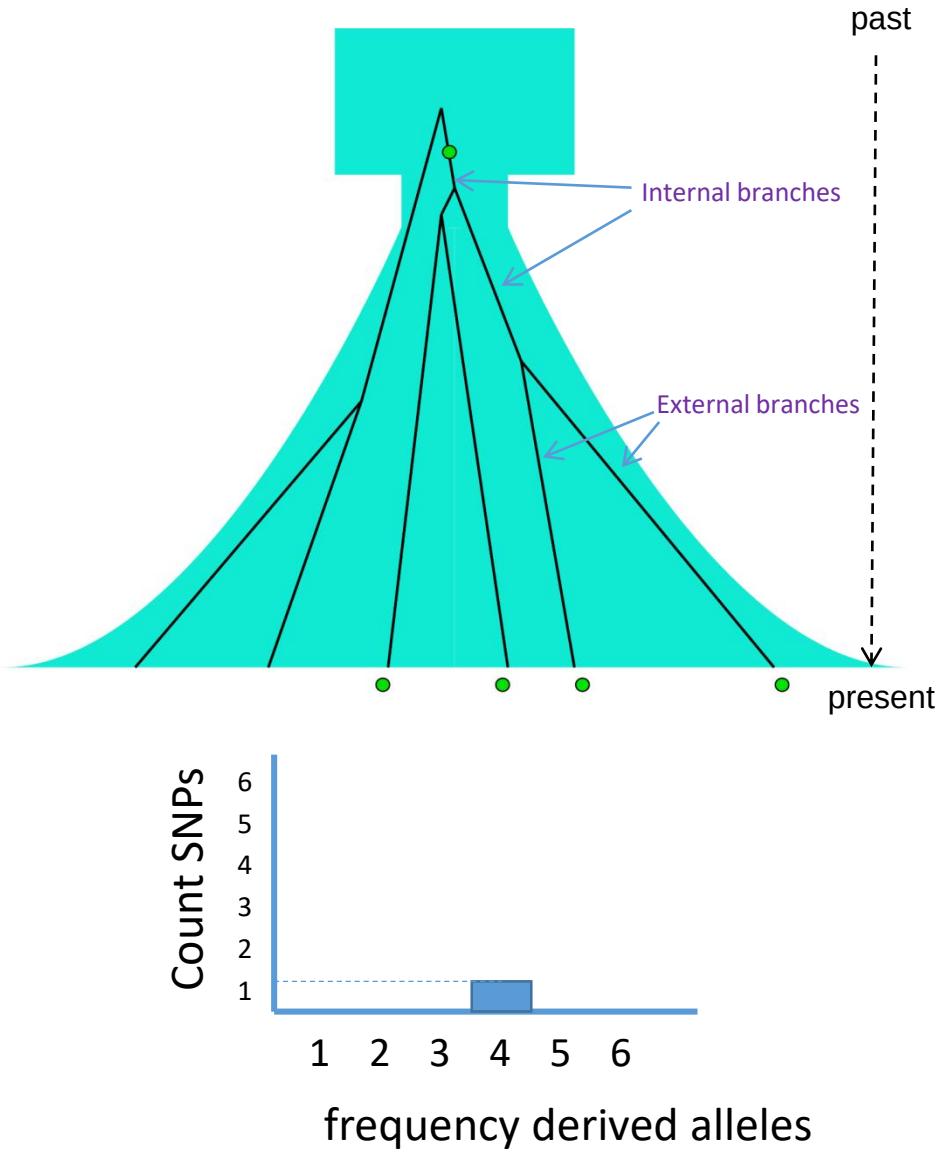


Observed SFS is a vector (1 dimensional SFS):

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

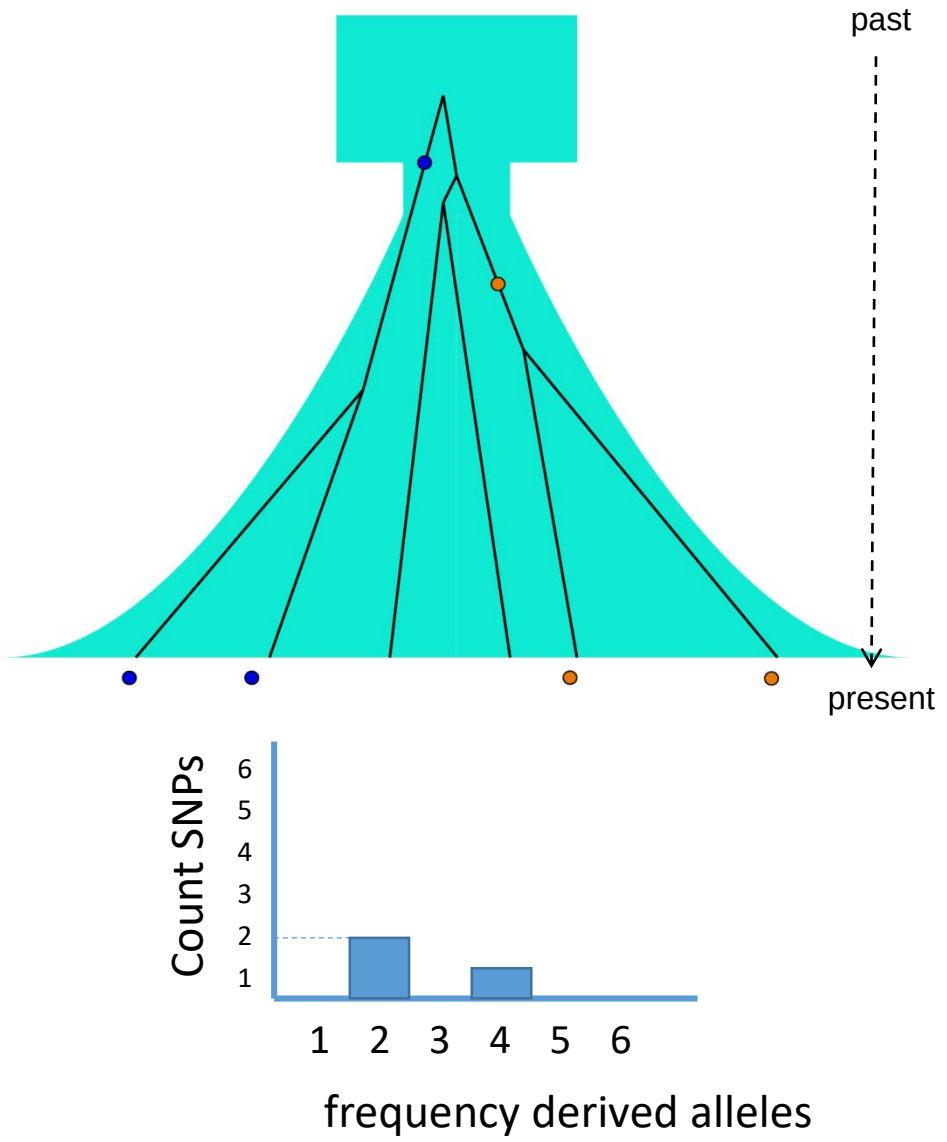
Coalescent and the SFS

- 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



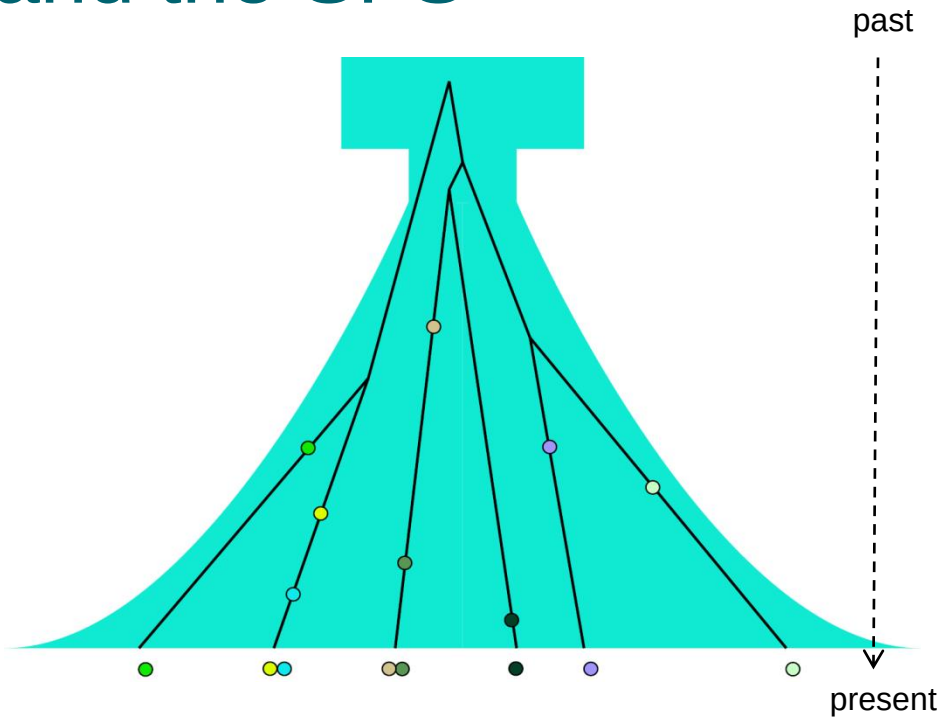
Coalescent and the SFS

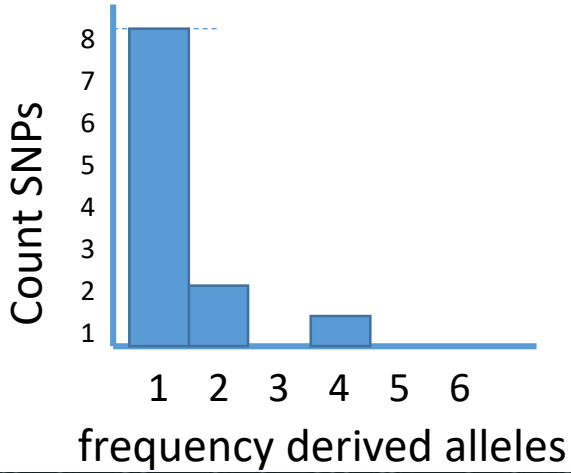
- A recent population growth following a bottleneck leads to gene trees with long external branches
- Very few mutations in the internal branches
- Most mutations in long external branches are only found in one lineage, resulting in an excess of singletons



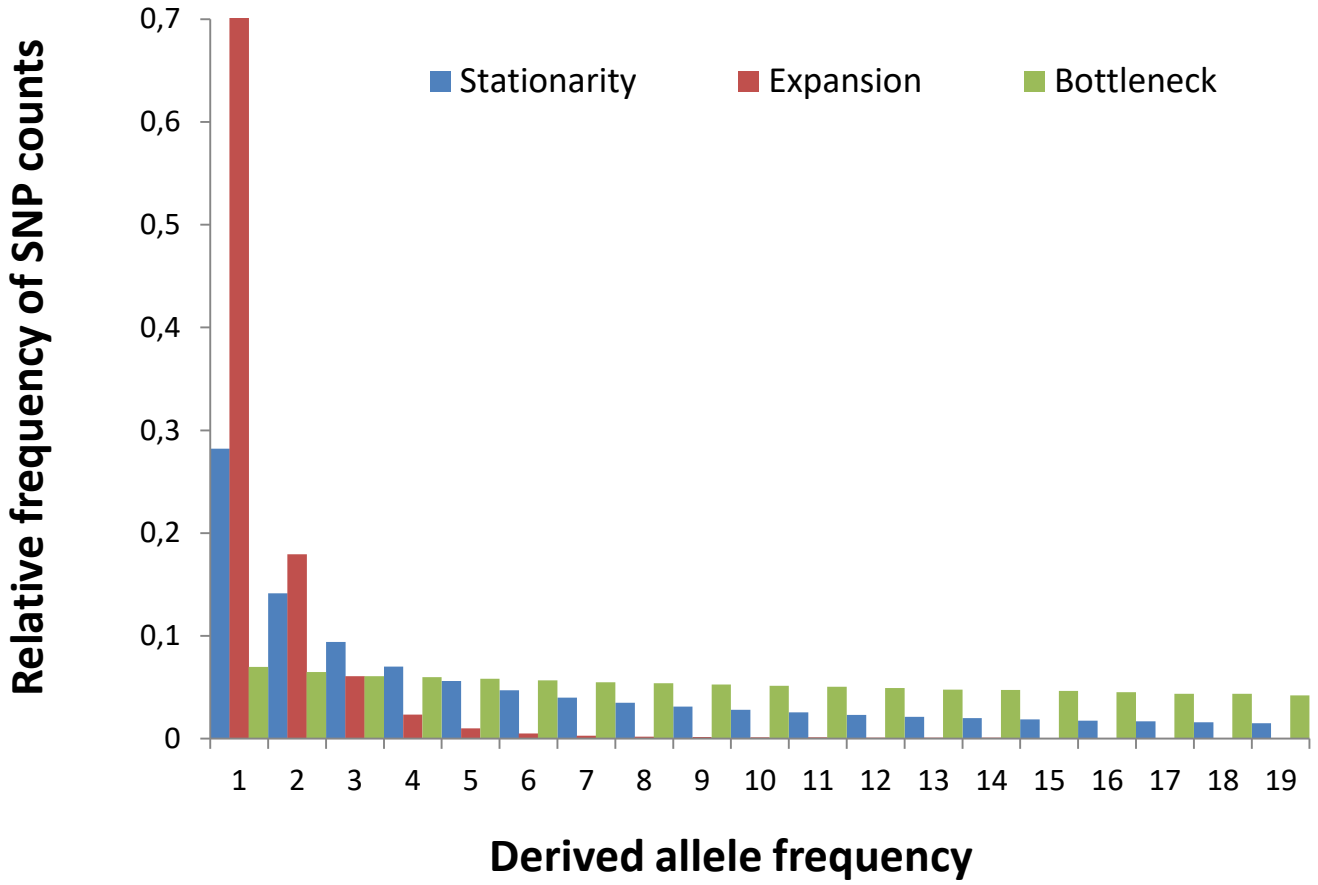
Coalescent and the SFS

- 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

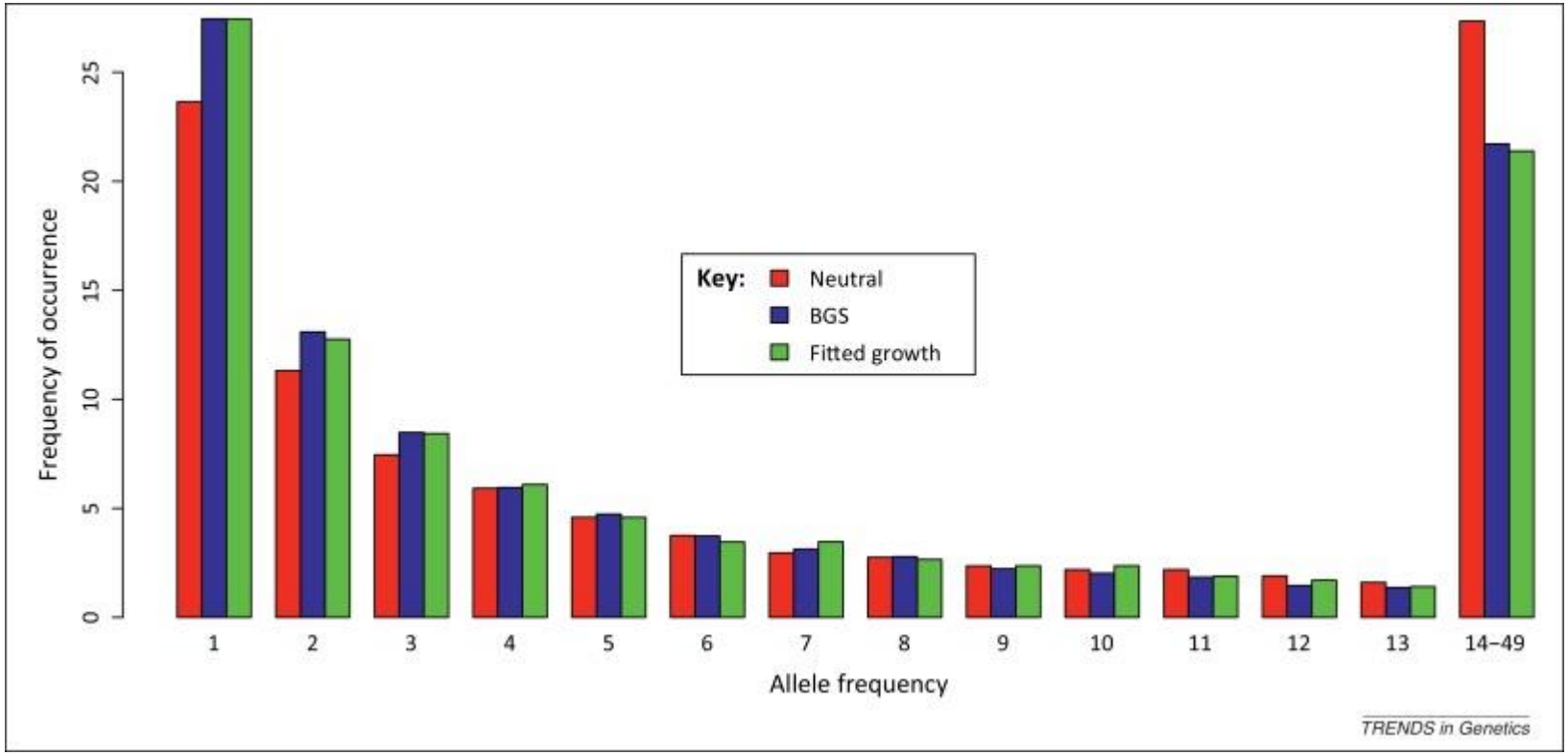




SFS depends on past demography



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

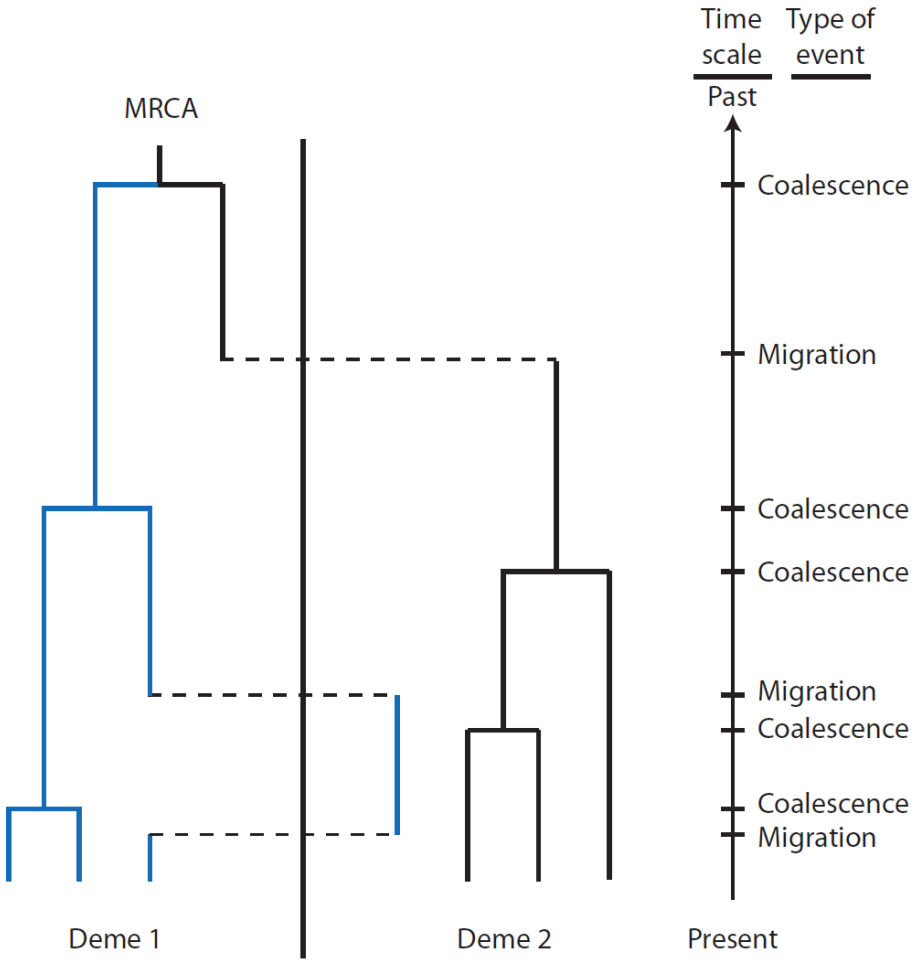
Migration events can be incorporated into gene trees.

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:

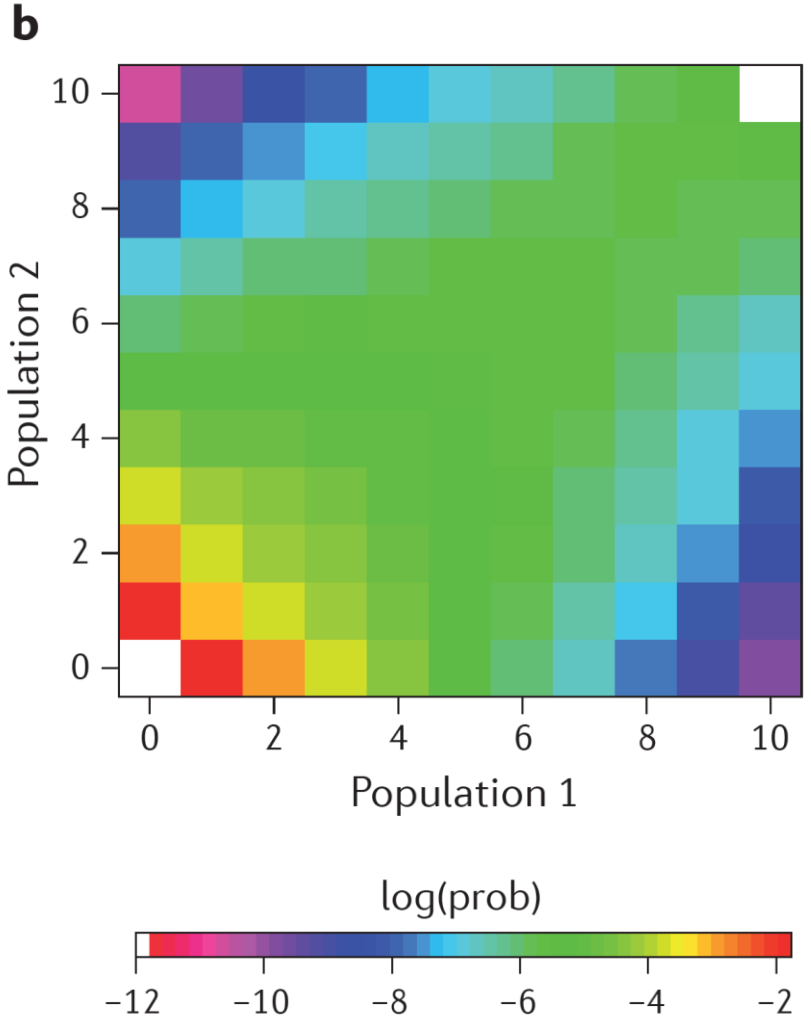
Pr(migrate) = $n_1 * m$

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.



F_{ST} in terms of coalescent times

$$F_{ST} = (f - f_w)/f$$

Where:
f is the average coalescent time among populations

f_w is the average coalescent time within populations

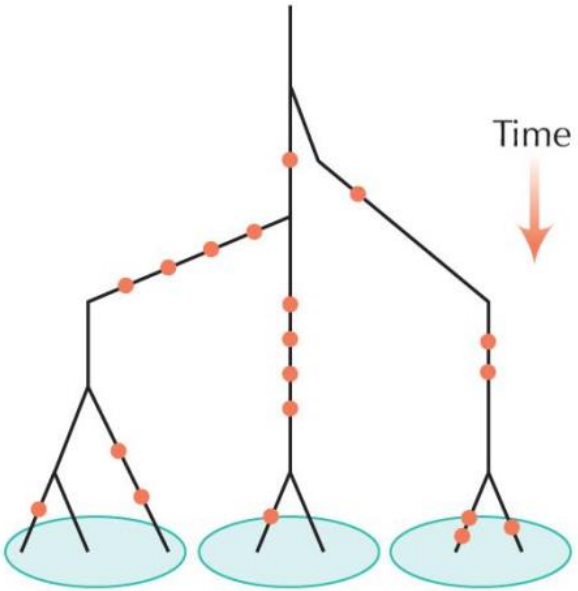


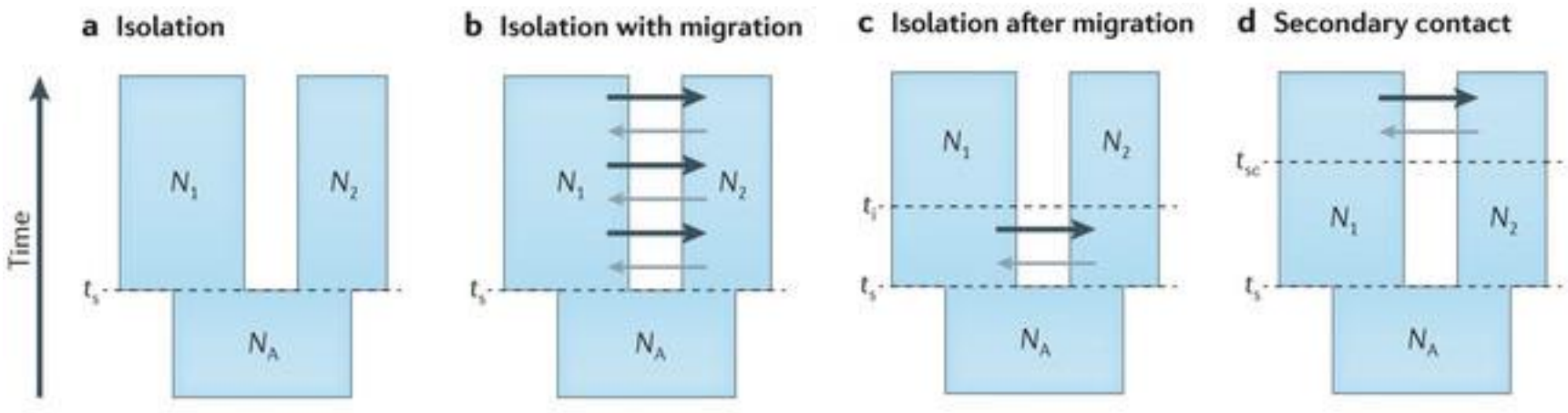
FIGURE 16.13. Wright's F_{ST} is related to the mean coalescence time between pairs of genes within demes, compared with the mean coalescence time between randomly chosen pairs: $F_{ST} = (\overline{T} - T_w)/\overline{T}$. These coalescence times, and hence F_{ST}, can be estimated from the number of mutations that separate each pair of genes (assuming the infinite sites model; see p. 424). In this example, seven genes are sampled from three demes; mutations are indicated by *red circles*. On average, there are 8.1 differences between pairs of genes sampled at random compared with 2.0 differences between genes within the same deme. Hence, F_{ST} is estimated to be $(8.1 - 2.0)/8.1 = 0.753$.

Derived vs Minor allele frequency spectrum

- So far, we have assumed that the allele frequency is the number of sequences with the derived allele frequency (unfolded SFS). We need information (outgroup) to determine the ancestral/derived state.
- If we do not have that information, we can work with the minor allele frequency (folded SFS). In this case, the allele with a lower frequency is treated as the reference.

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.

A model is represented by a population tree that reflects the past evolutionary history

Parameters:

Demography

- Population split times
- Migration rates
- Effective population sizes
- Temporal changes in migration rates and effective sizes

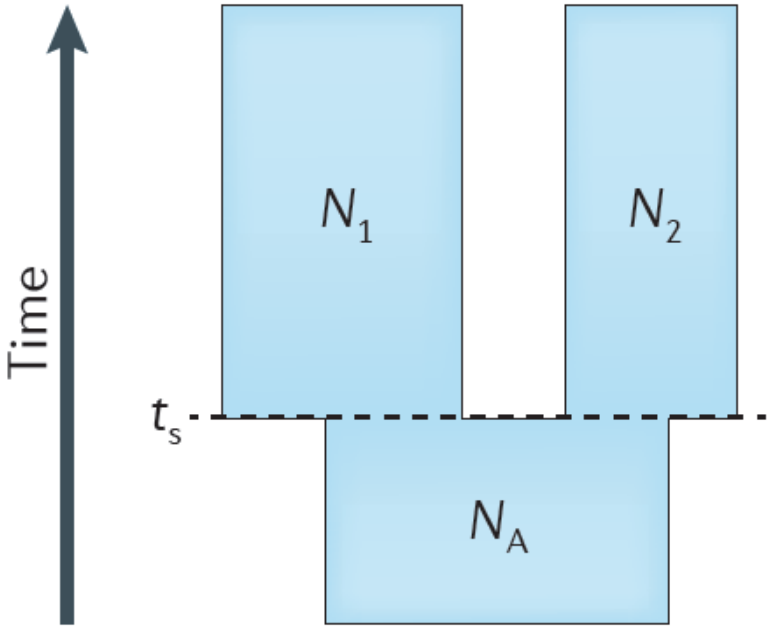
Selection

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

Genomic processes

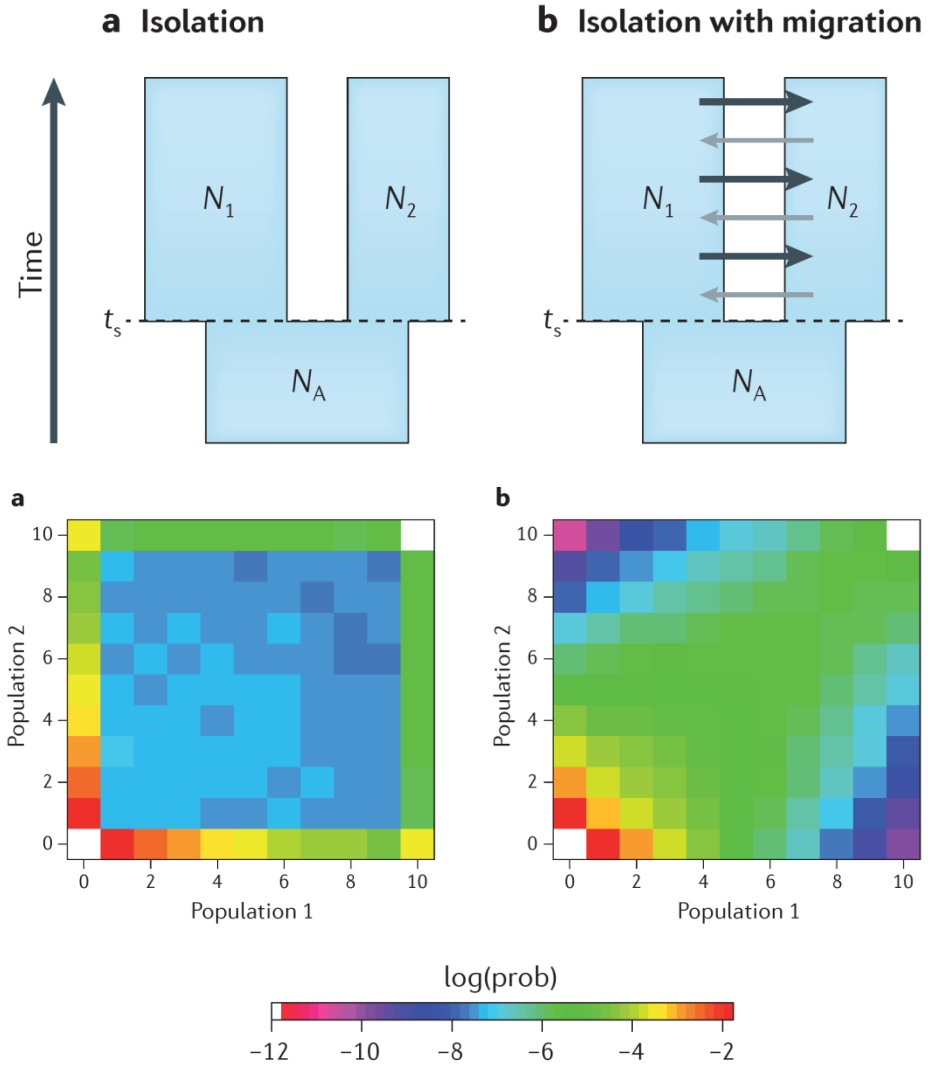
- Mutation rate
- Recombination rate

Population tree



Site frequency spectrum (SFS)

The SFS contains information about the demographic history of populations

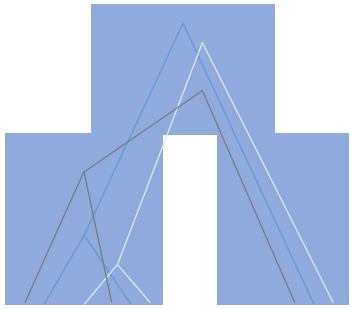


Sousa and Hey (2013) Nat. Rev. Genet.

Inferring the demographic history from the SFS

Genomic Data

Model



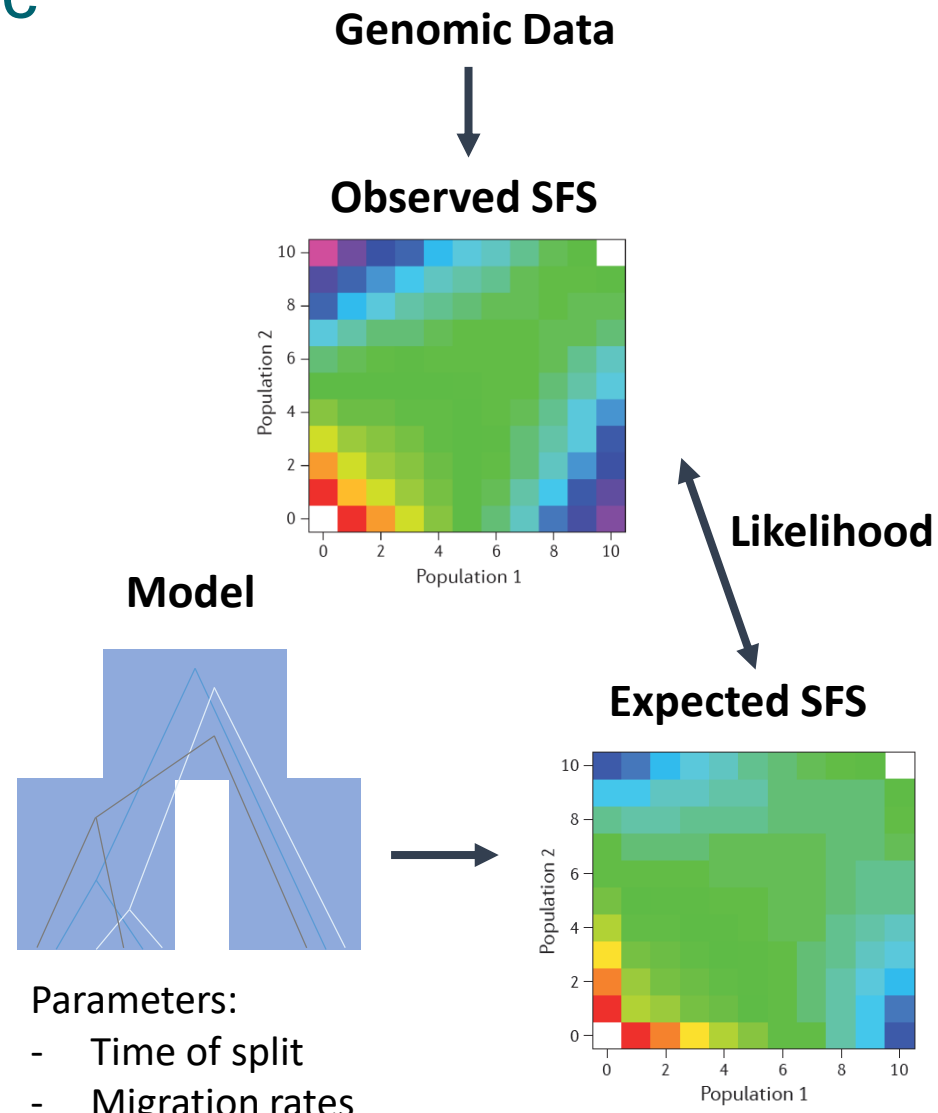
Parameters:

- 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

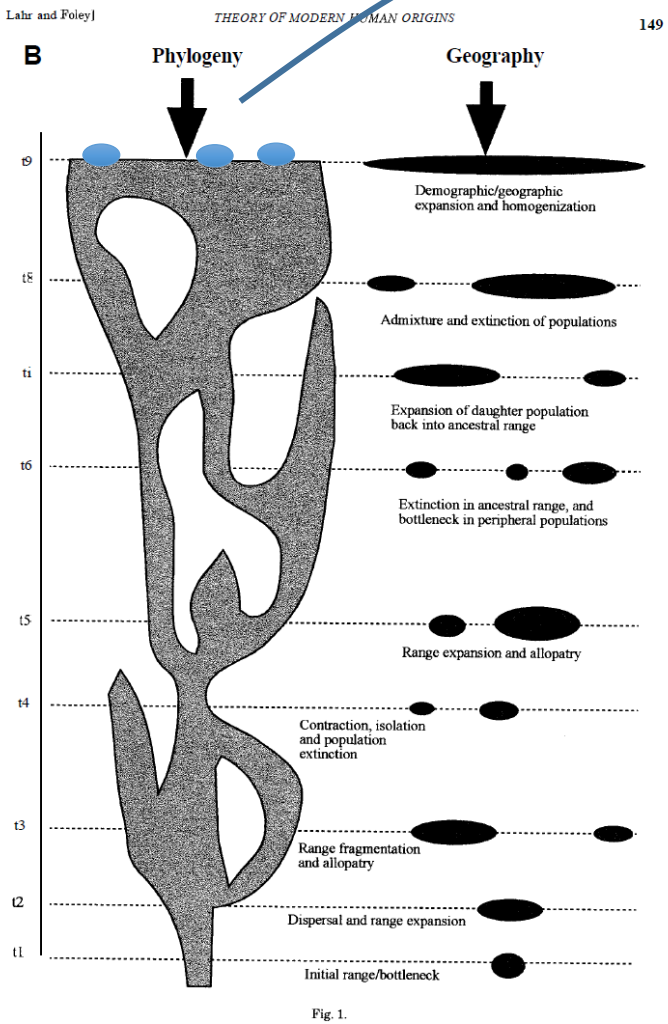
- There are different ways to obtain the expected SFS
  - Diffusion (forward in time)
  - Coalescent (backward in time)



Excoffier et al. (2013) *PloS Genetics*

Framework for demographic inference

"Truth"



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

Lahr and Foley, 1994

Estimating the SFS from the coalescent

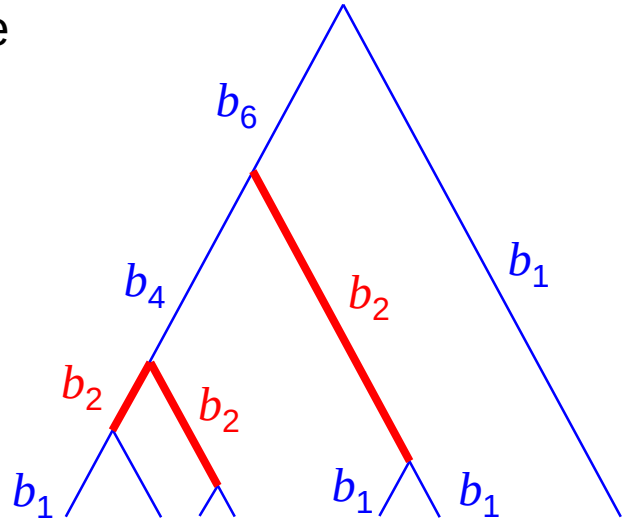
The probability of a SFS entry *i* can be estimated under a specific model θ from its expected coalescent tree as (Nielsen 2000)

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

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:

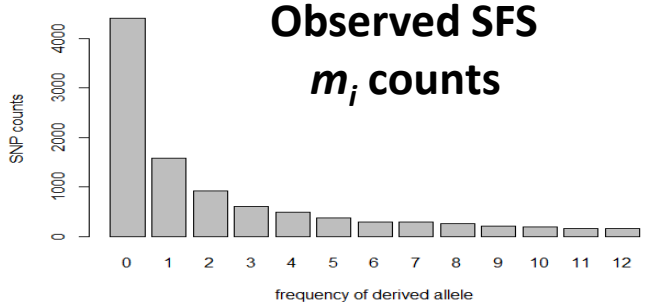
$$CL = \Pr(X \mid \theta) \propto P_0^{L-S} (1-P_0)^S \prod_{i=1}^{n-1} \hat{p}_i^{\,m_i}$$

probability of no mutation on the tree

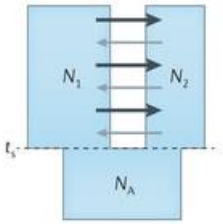
probability of at least one mutation in the tree

3 ingredientes for likelihood

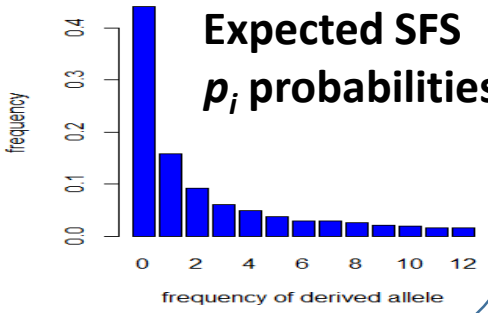
Observed SFS
m_i counts



Model



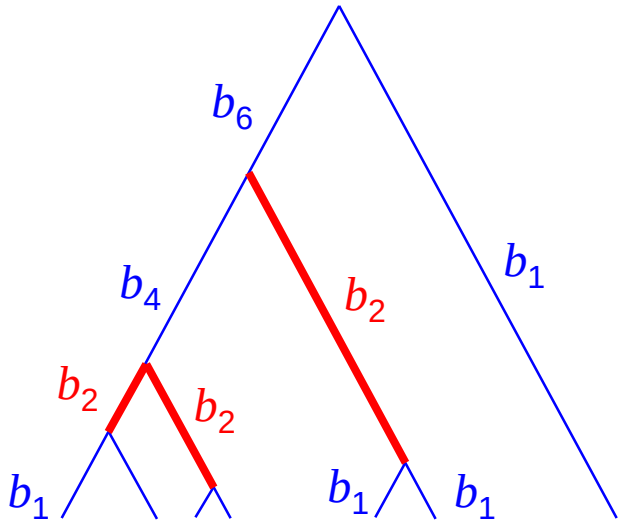
Expected SFS
p_i probabilities



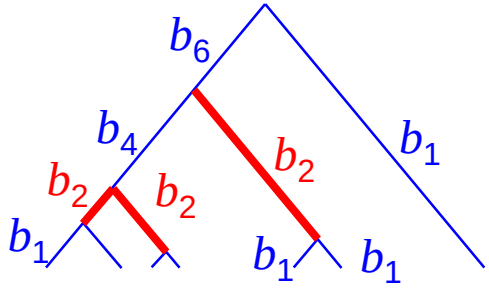
Composite likehood

These probabilities depend:
- Number of monomorphic sites
- A fixed and mutation rate

Everything is relative



T_L = total branch length



Frequency	0	1	2	3	4	5	6	7
SNP probability p_i	0	Sum(b_1)/ T_L	Sum(b_2)/ T_L	Sum(b_3)/ T_L	Sum(b_4)/ T_L	Sum(b_5)/ T_L	Sum(b_6)/ T_L	0

- The same expected SFS can be obtained in a large or small tree
- We need a mutation rate and the number of monomorphic sites to distinguish among the two!

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.

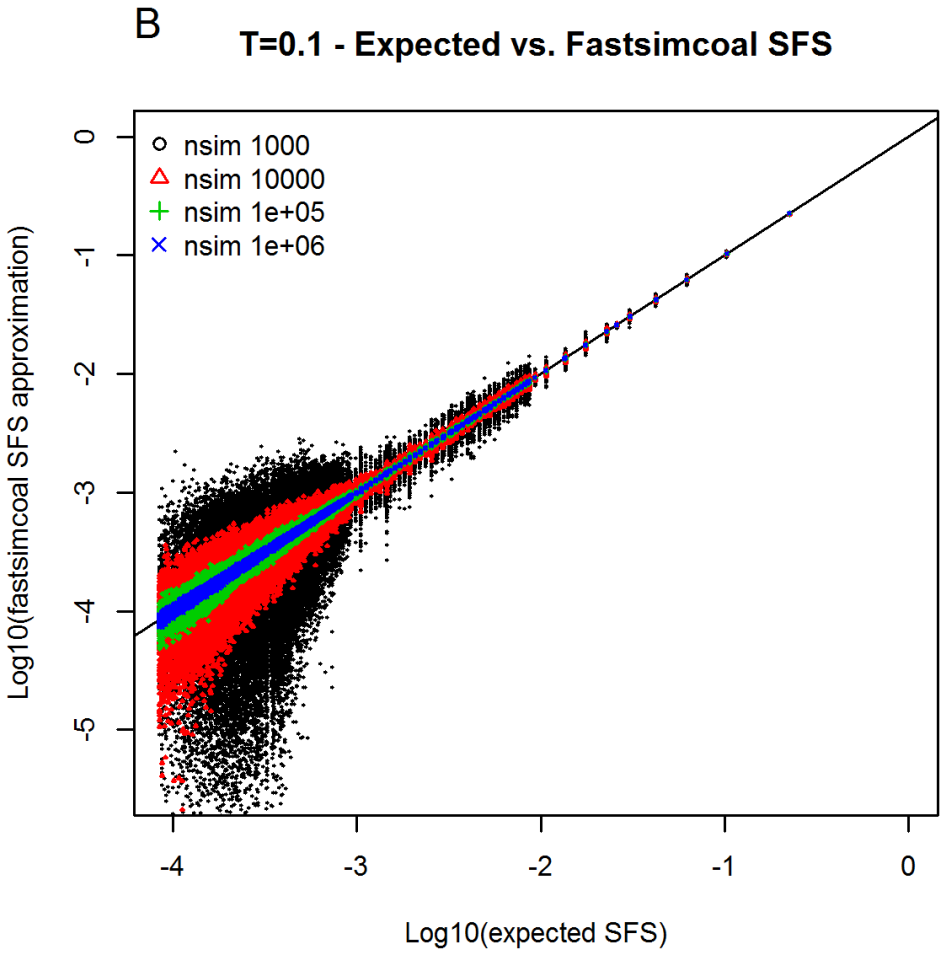
These probabilities can then be used to compute the composite likelihood of a given model as (Adams and Hudson, 2004)

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

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

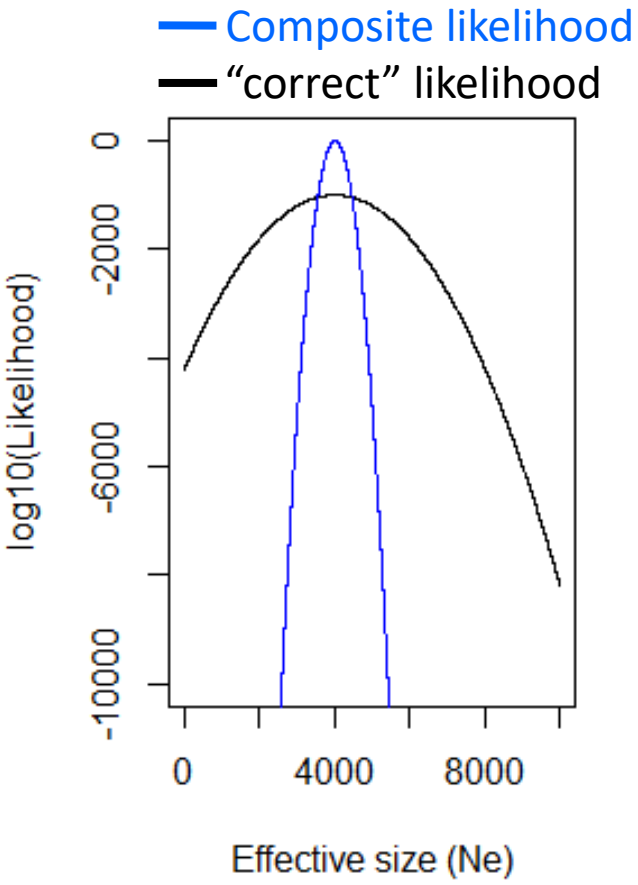


Excoffier et al. (2013) PloS Genetics; Chen (2012) TPB

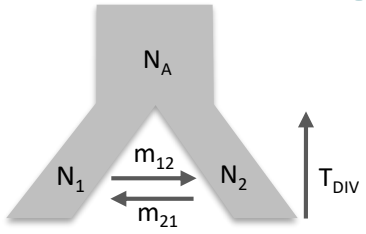
Properties of composite likelihoods

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

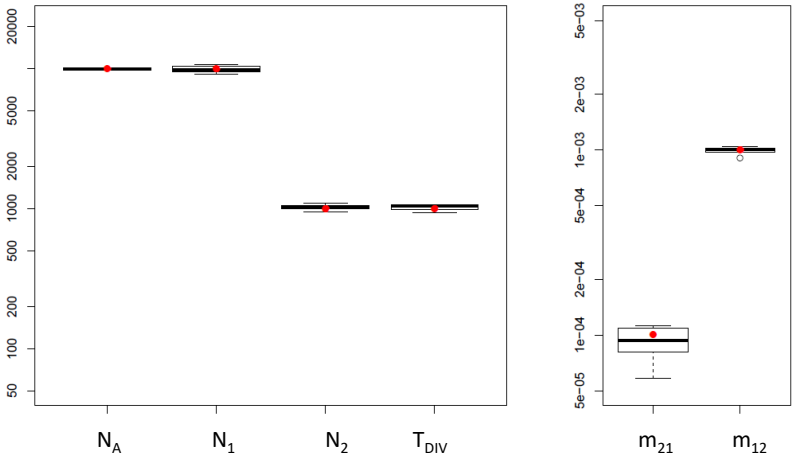


Comparisons of approaches

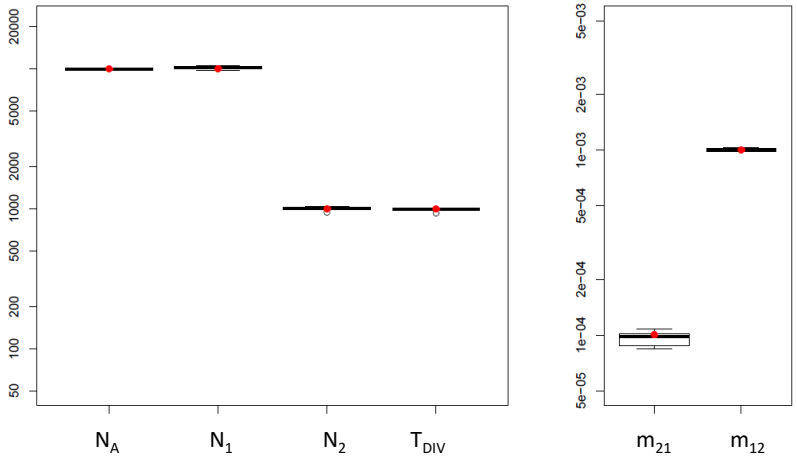


Simulation of 20 Mb data

fastsimcoal2



$\partial a \partial i$



Protocol for parameter estimation

1. **Get the observed SFS**:
   - derived SFS (DAF or unfolded SFS), when the ancestral state is known;
   - minor allele frequency SFS (MAF or folded SFS) when the ancestral state is unknown
2. Define the **demographic model**
3. **Estimate the parameters** – repeat 50-100 runs, and selecting the run with maximum likelihood
4. **Bootstrap** to obtain confidence intervals for each parameter – bootstrap 10-100 datasets, by repeating a few runs for each dataset
   - For datasets with linked sites use block-bootstrap, diving the genome into blocks

Potential problems

- Maximization of the CL is not trivial (precision of the approximation and convergence problems)
- Need to repeat estimations to find maximum CL
- Needs genomic data (several Mb), difficult to have gene-specific estimates
- Next-generation sequencing data must have high coverage (>10x) to correctly estimate SFS

Problems with estimation of demographic parameters from SFS

Can one learn history from the allelic spectrum?

Simon Myers[a], Charles Fefferman[b], Nick Patterson[a,*]

[a] *Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge MA 02142, United States*
[b] *Deptartment of Mathematics, Fine Hall, Washington Road, Princeton, NJ 08544, United States*

Theoretical Population Biology
www.elsevier.com/locate/tpb

Received 17 March 2007
Available online 30 January 2008