

# Demographic inference based on Site frequency spectrum (SFS) using fastsimcoal2

Vitor Sousa

CE3C – center for ecology, evolution and environmental changes

PGDH19

Population Genetics and Demographic History:  
model-based approaches

GTPB

14 May 2019

vmsousa@fc.ul.pt



*u<sup>b</sup>*

---

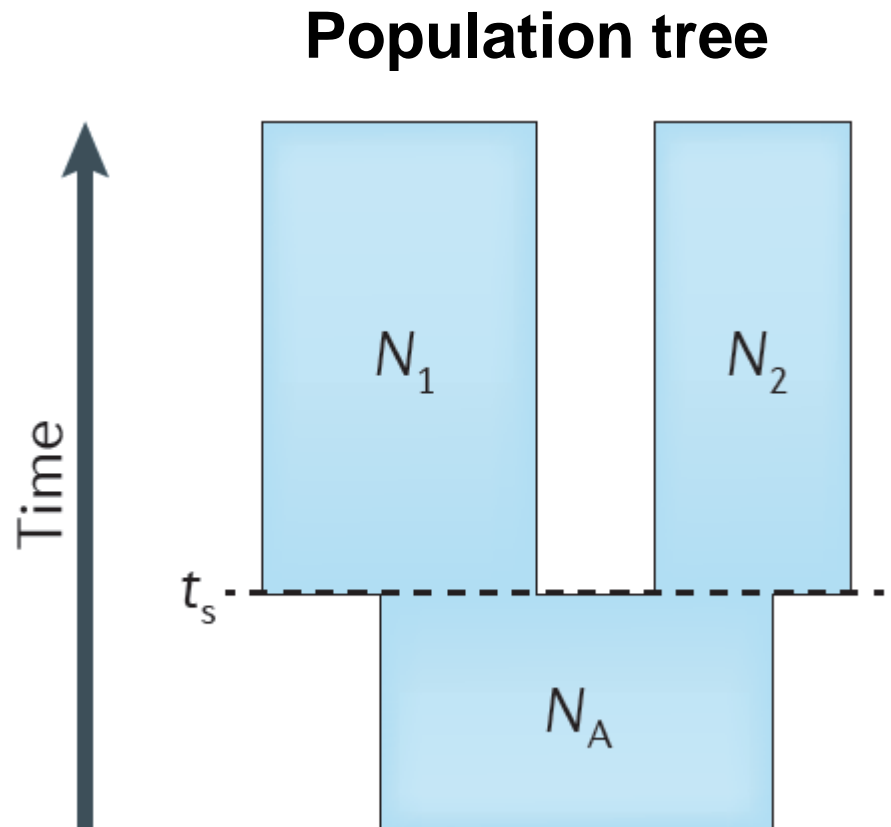
<sup>b</sup>  
UNIVERSITÄT  
BERN



# Demographic history of populations

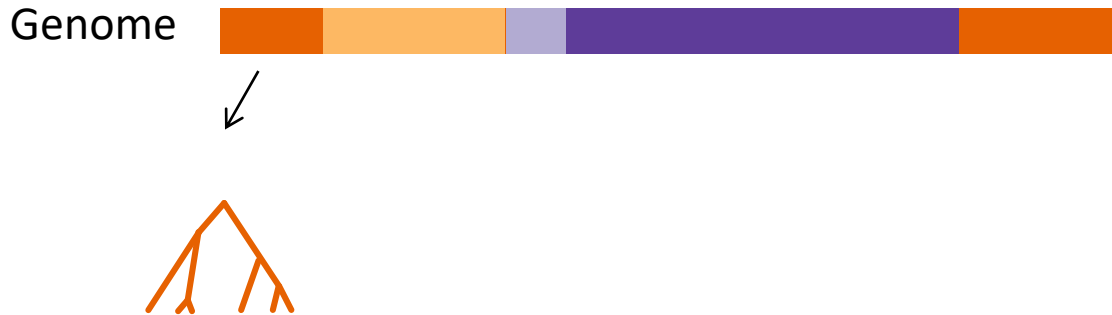
Past demographic events:

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

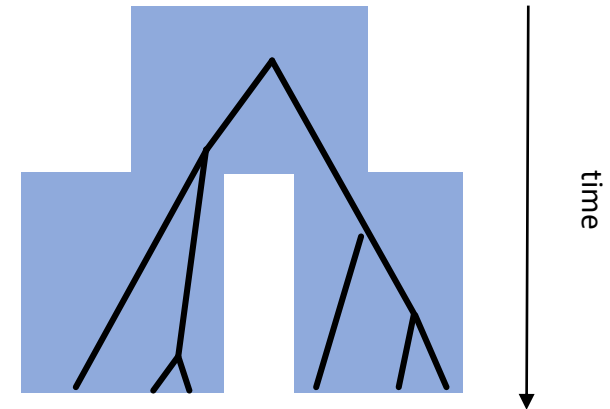


# Reconstructing the demographic history from genomic data

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



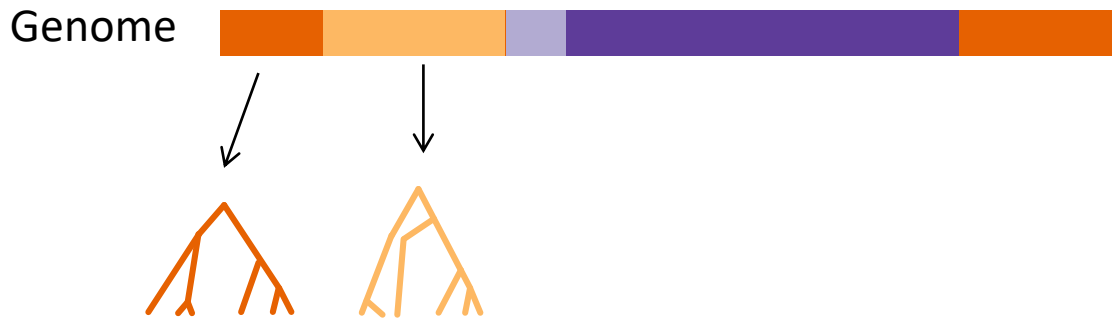
- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions



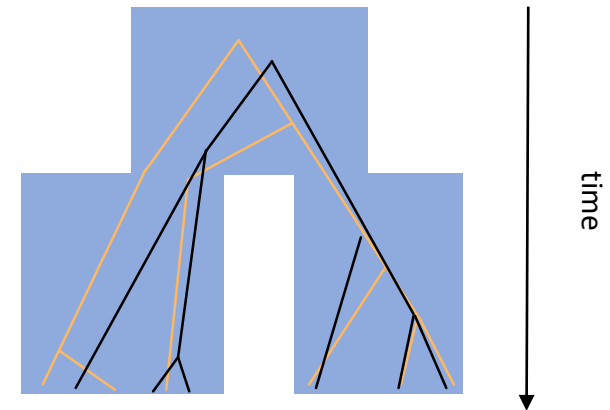
All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

# Reconstructing the demographic history from genomic data

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



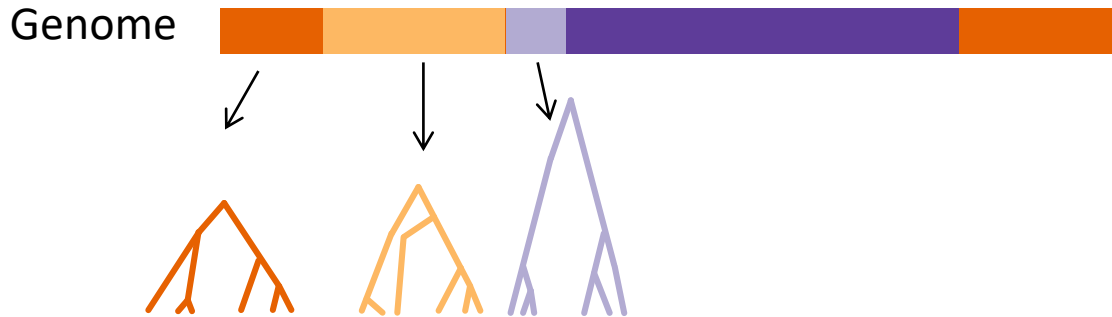
- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions



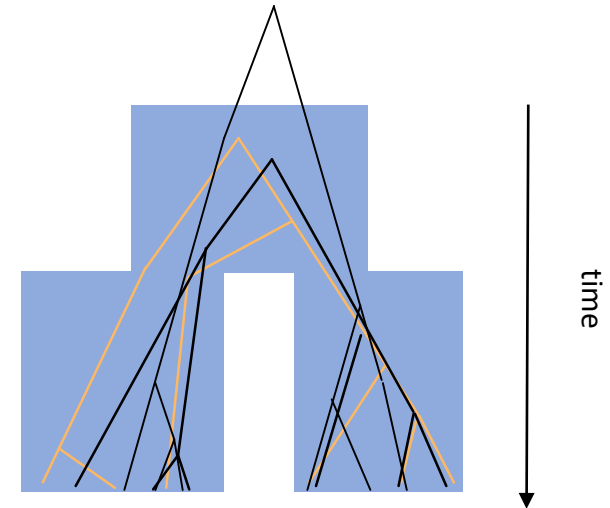
All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

# Reconstructing the demographic history from genomic data

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



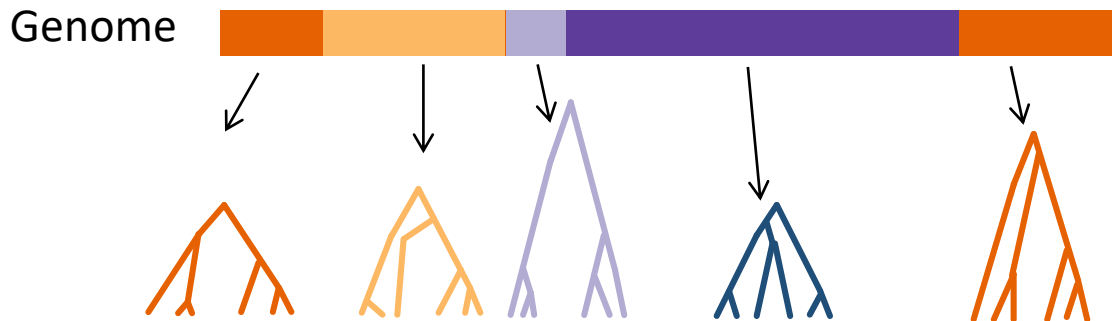
- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions



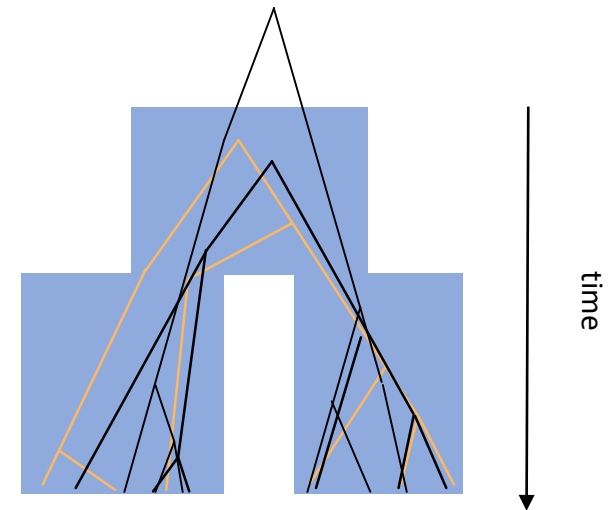
All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

# Reconstructing the demographic history from genomic data

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

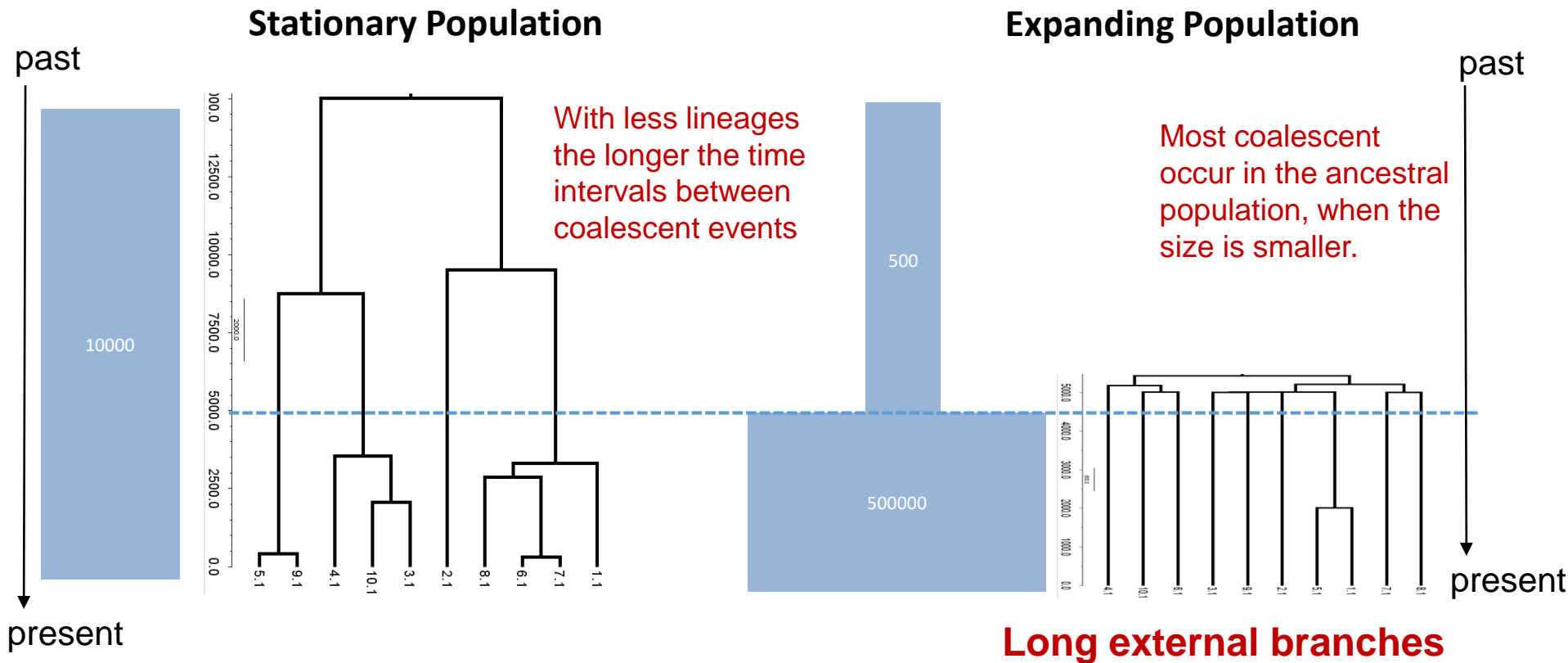


- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions



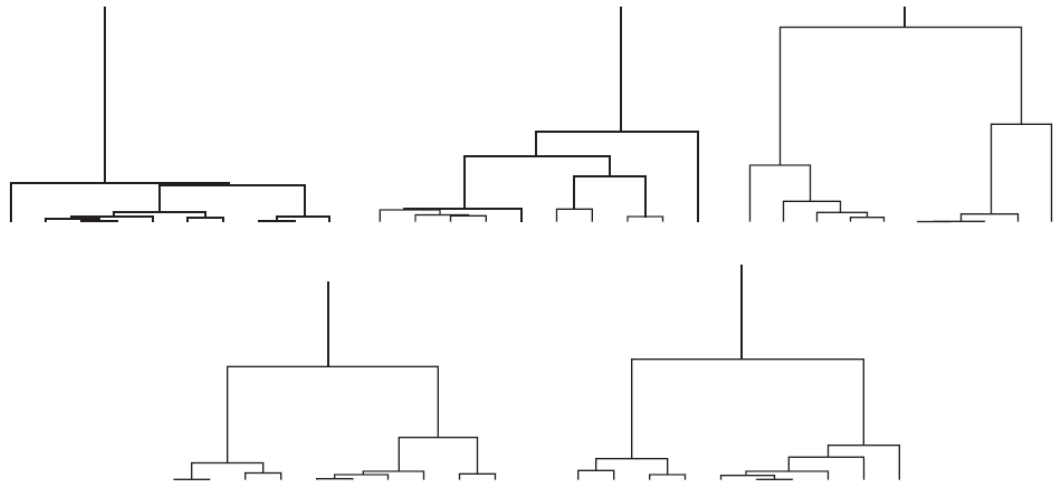
All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

# Gene trees in growing populations



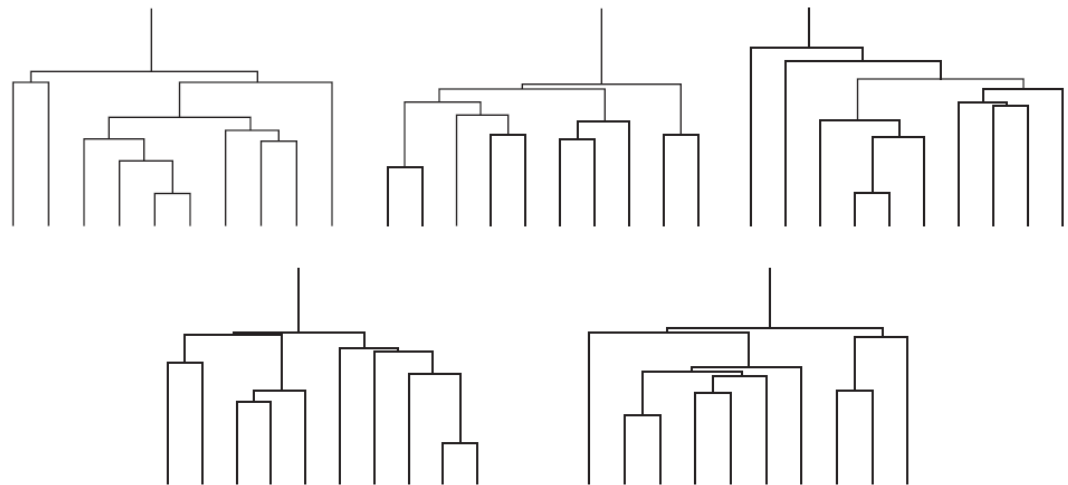
- Coalescent rate is larger in smaller populations, and so we expect smaller intervals between coalescent events in smaller populations
- Coalescent rate is lower with a lower number of lineages, and so we expected larger intervals between coalescent events as the number of lineages decrease

**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

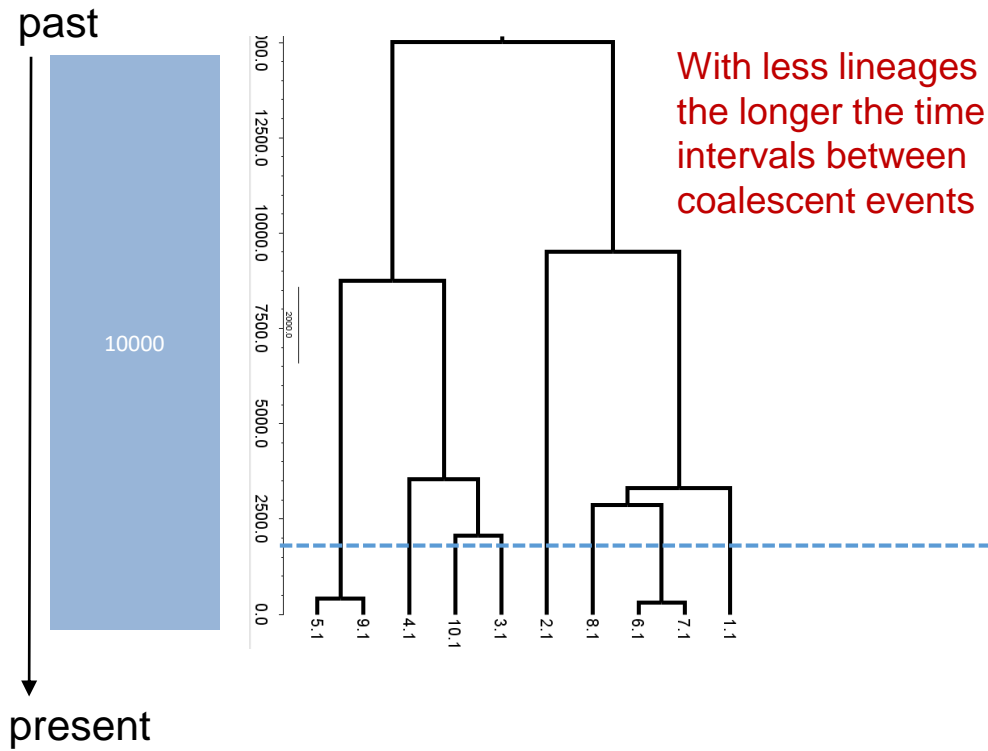


**Figure 4.3** Five replicates of the coalescent with exponential growth,  $\beta = 1000$ , for a sample of  $n = 10$  genes. Note the smaller variance in the time until the MRCA compared to the same quantity in Figure 4.2.

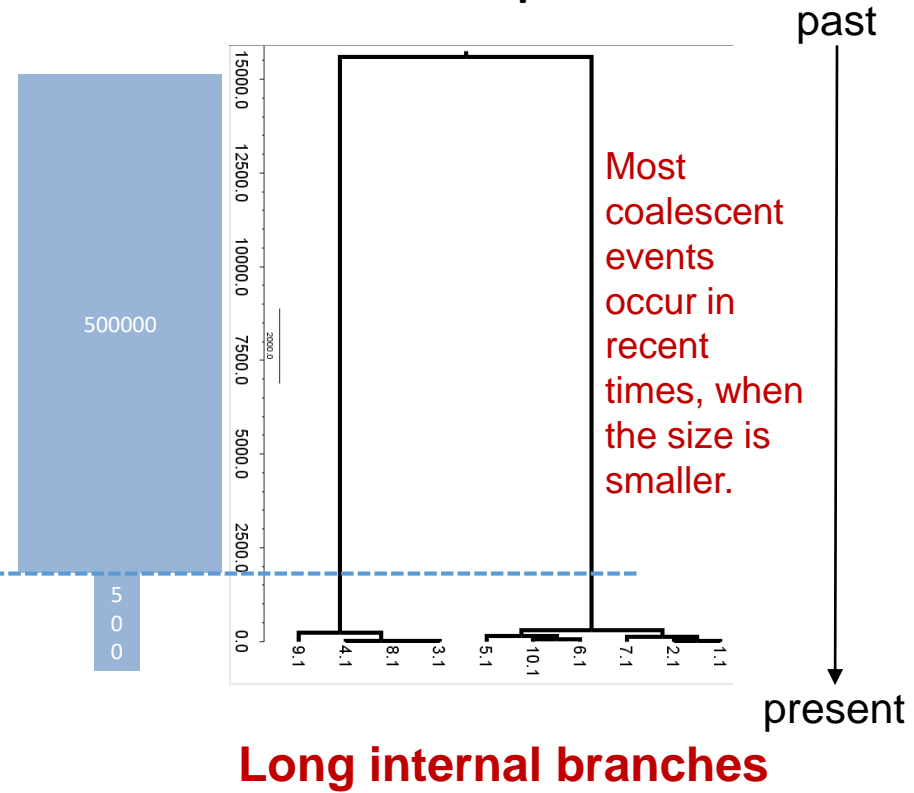


# Gene trees for decreasing populations

## Stationary Population



## Bottleneck Population

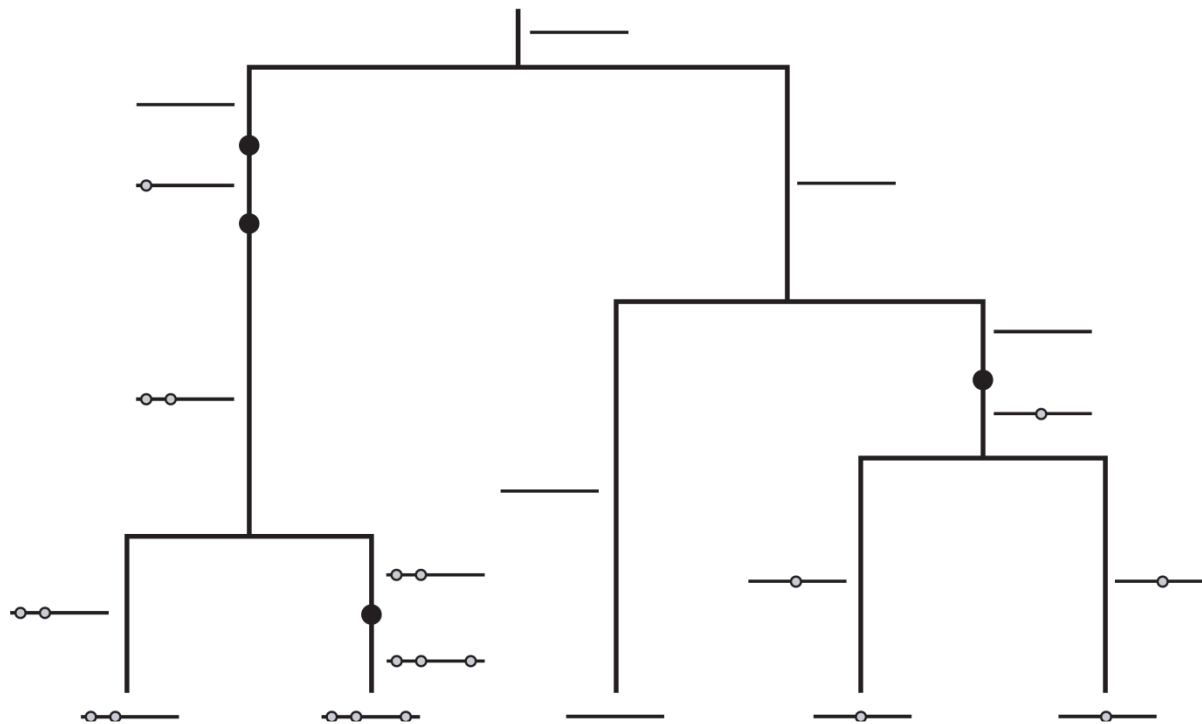


- If we could observe directly the gene trees, we could easily reconstruct the population tree and the demographic history.
- But we do not observe gene trees...
- We can still learn about gene trees from the observed mutations and the allele frequencies in samples



# Adding neutral mutations

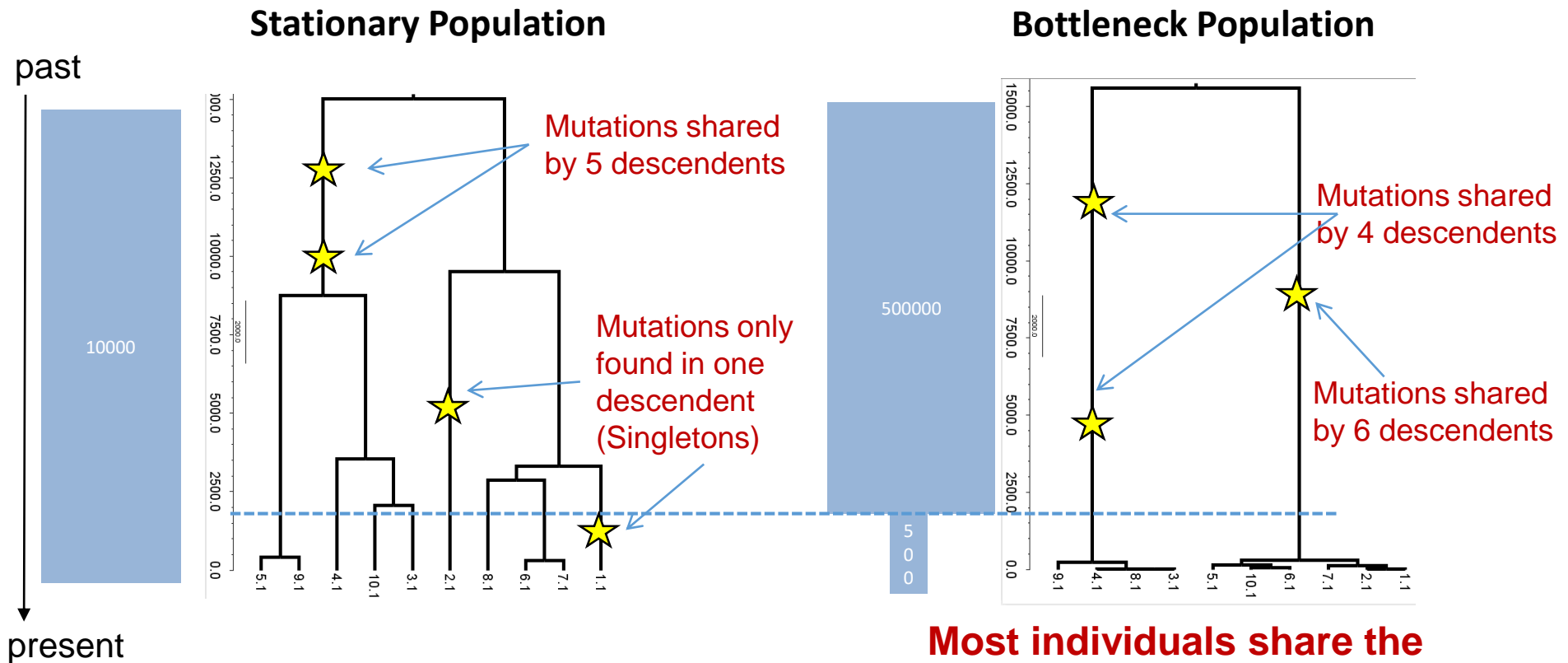
The shape of neutral coalescent trees only depend on the population demography, and not on the mutational process. Assuming that all alleles have the same fitness, the mutational process can be modeled as an independent process superimposed on a realized coalescent tree.



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**

# 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.

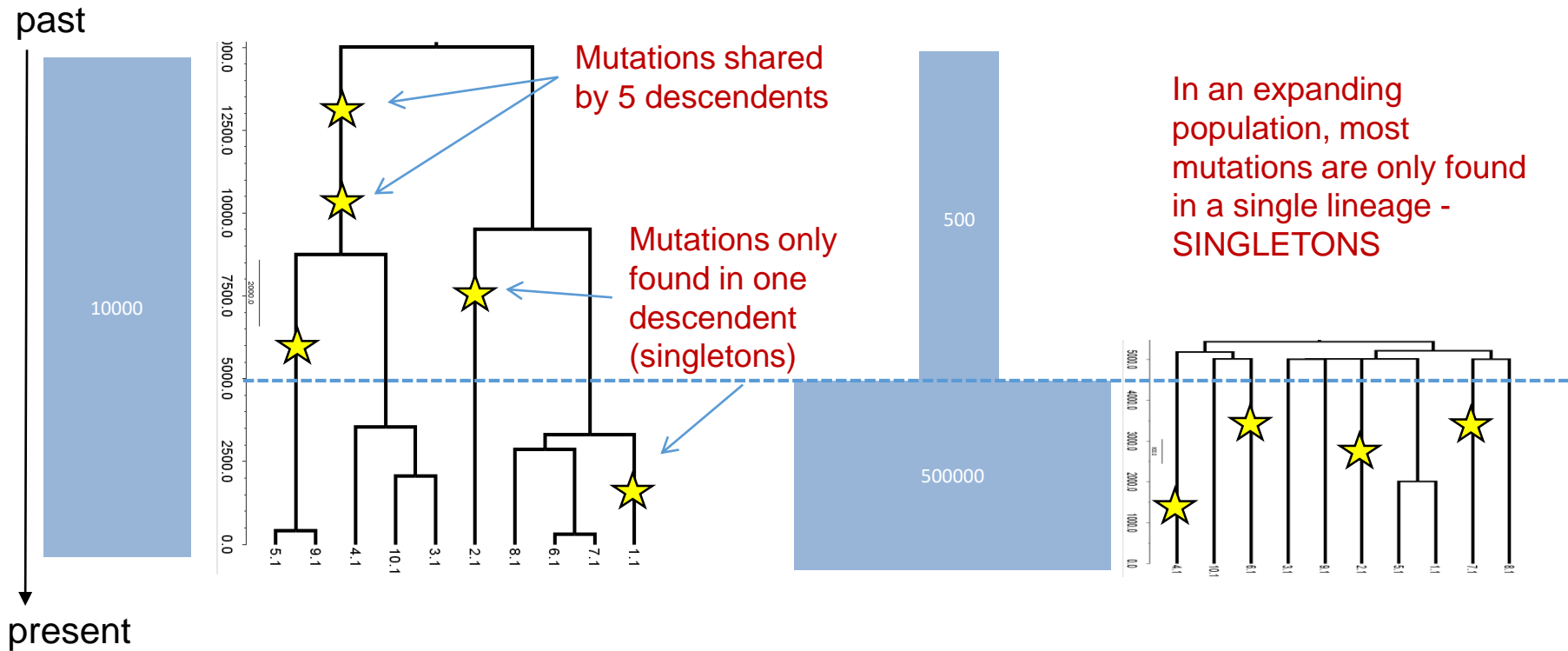


**Most individuals share the same mutations**

# We expect less diversity in a bottlenecked population

## Stationary Population

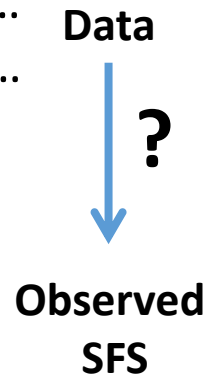
## Expanding Population



# Site frequency spectrum (SFS)

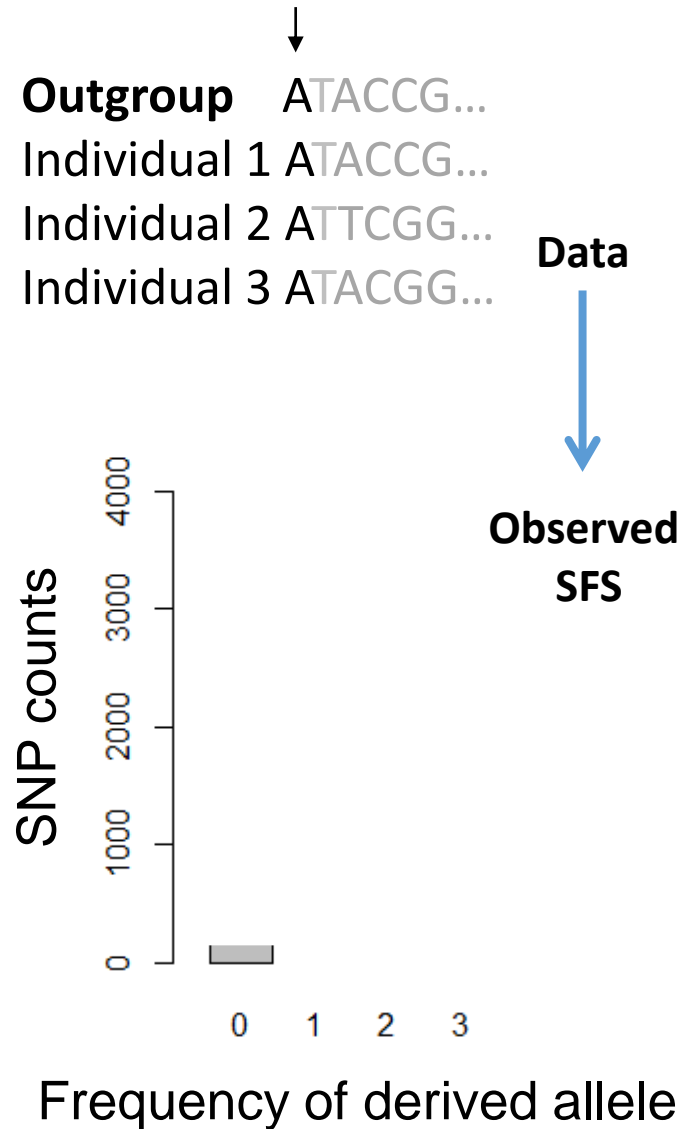
- The SFS summarizes efficiently genome-wide data
- Assuming a single population – 1Dimensional SFS

**Outgroup** ATACCG...  
Individual 1 ATACCG...  
Individual 2 ATTCGG...  
Individual 3 ATACGG...



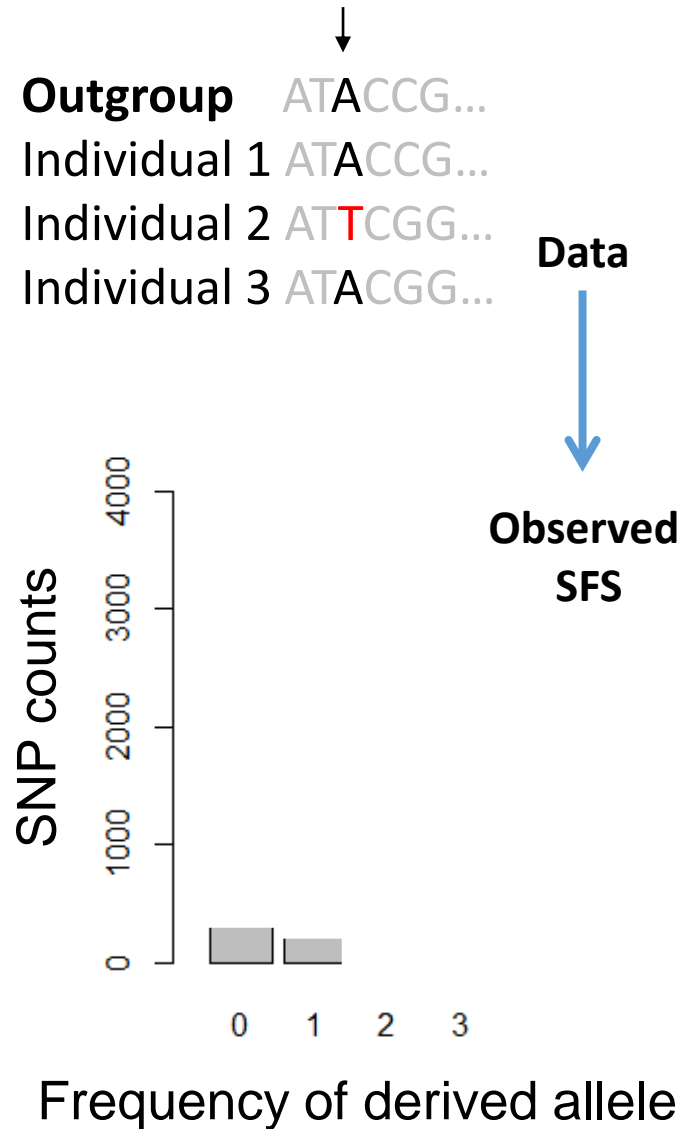
# Site frequency spectrum (SFS)

- 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:

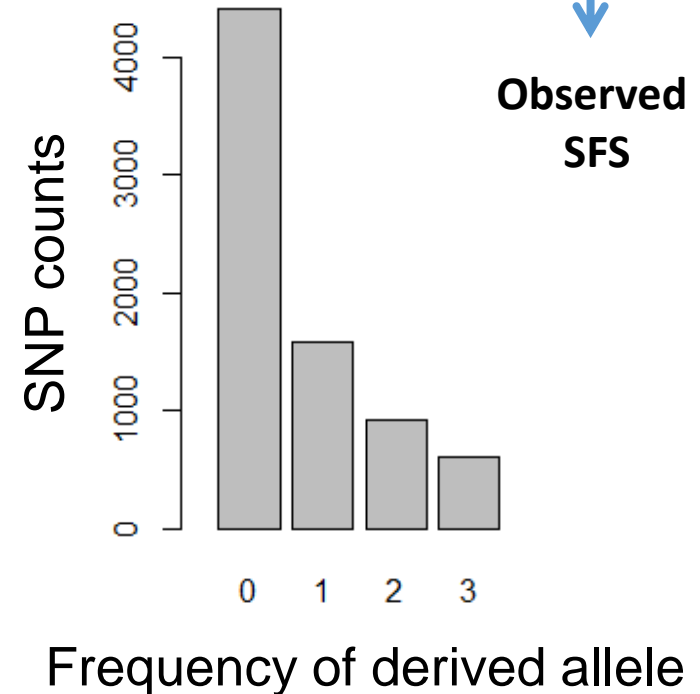
$$E(\xi_i) = \frac{\theta}{i} \quad \text{Fu and Li, 1993}$$

**Outgroup** ATACCG...  
Individual 1 ATACCG...  
Individual 2 ATTCGG...  
Individual 3 ATACGG...

Data



**Observed SFS**



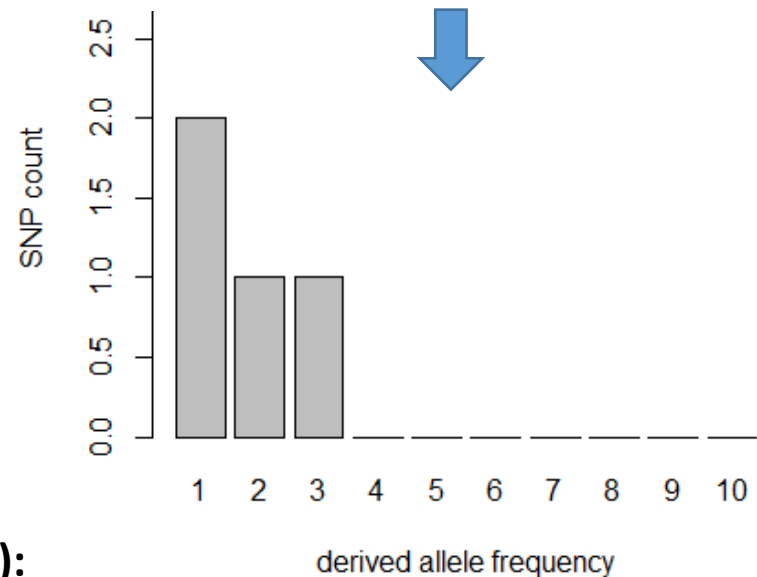
# We can obtain the SFS from genotype call data

## Genotypes:

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

	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

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



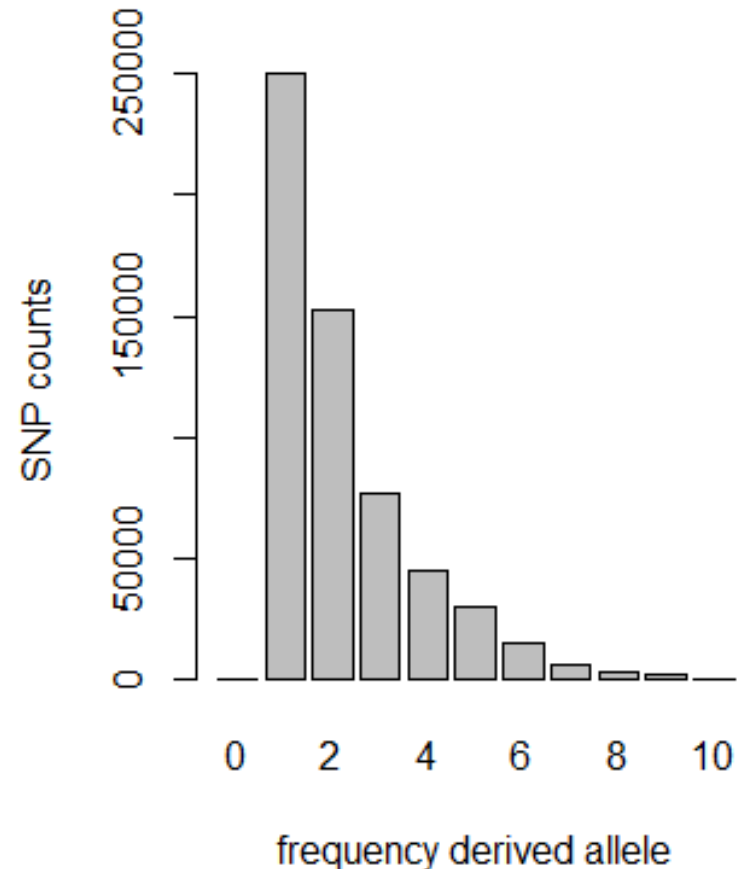
## Observed SFS is a vector (1 dimensional SFS):

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.

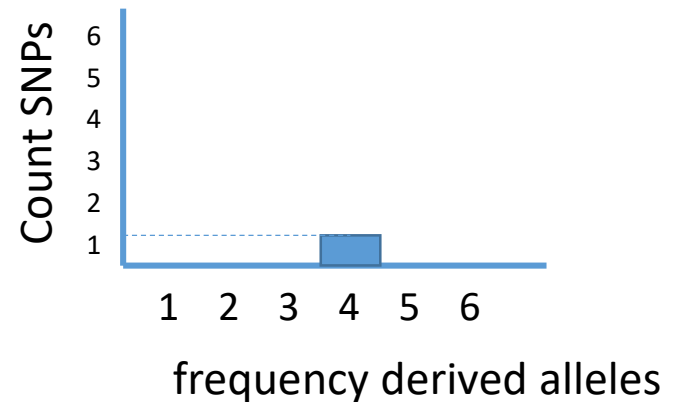
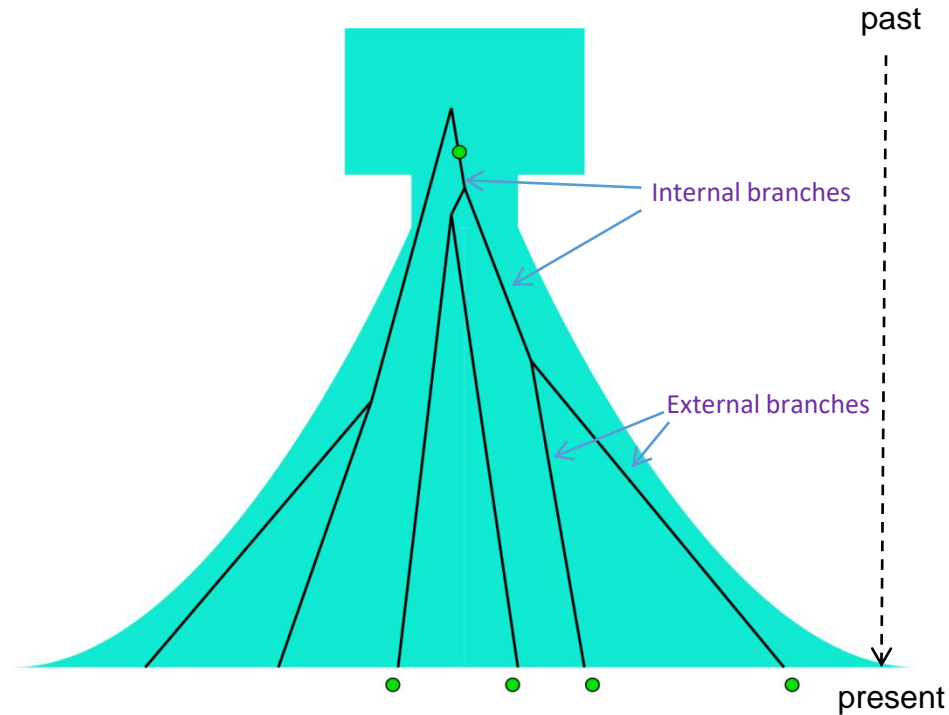


**Observed SFS is a vector (1 dimensional SFS):**

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

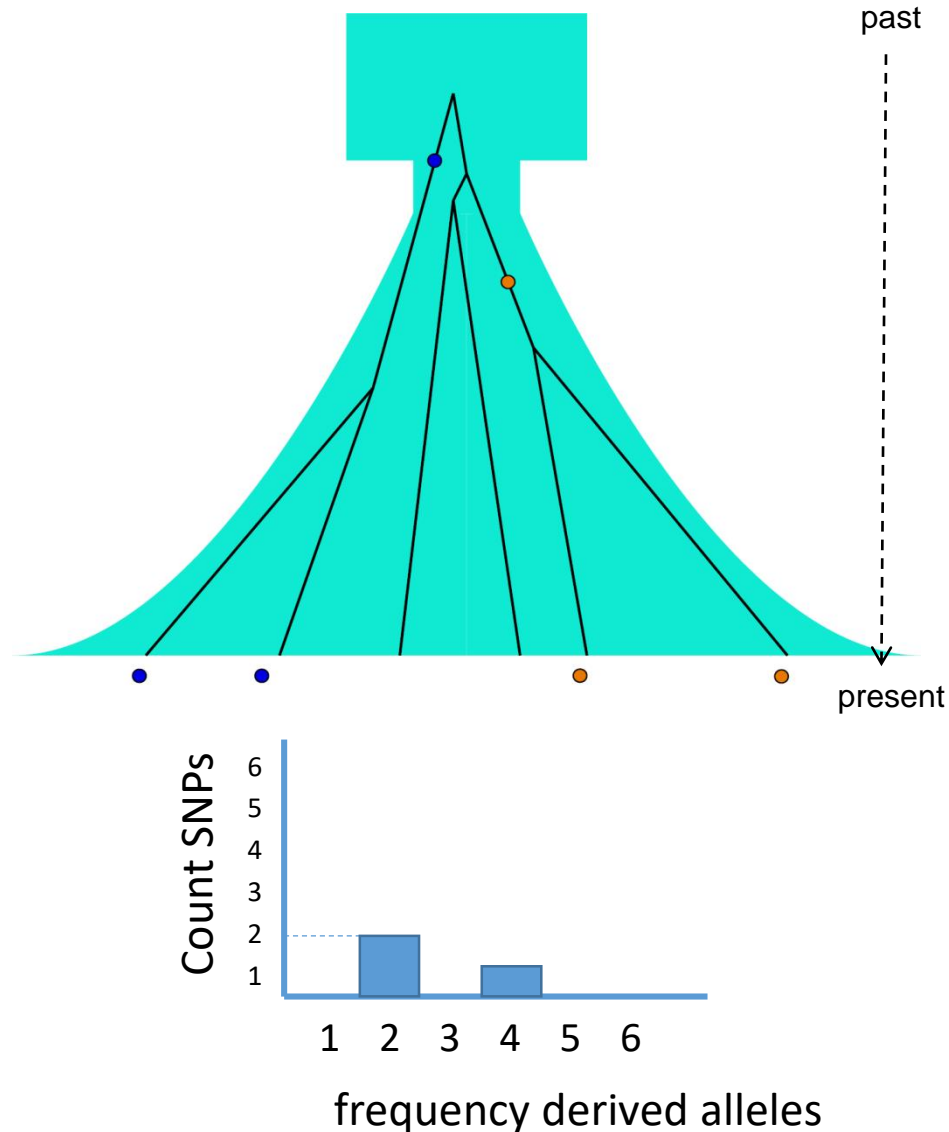
# Coalescent and the SFS

- 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



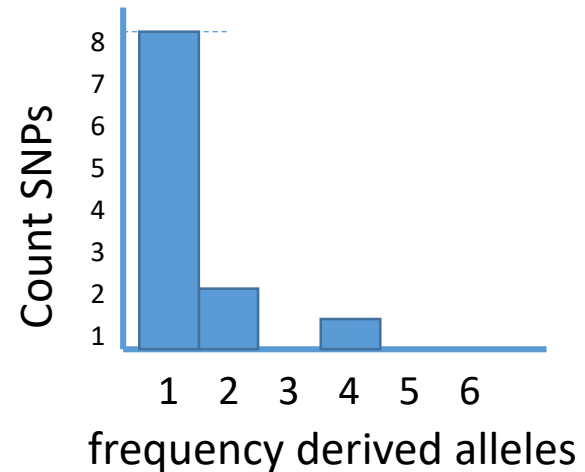
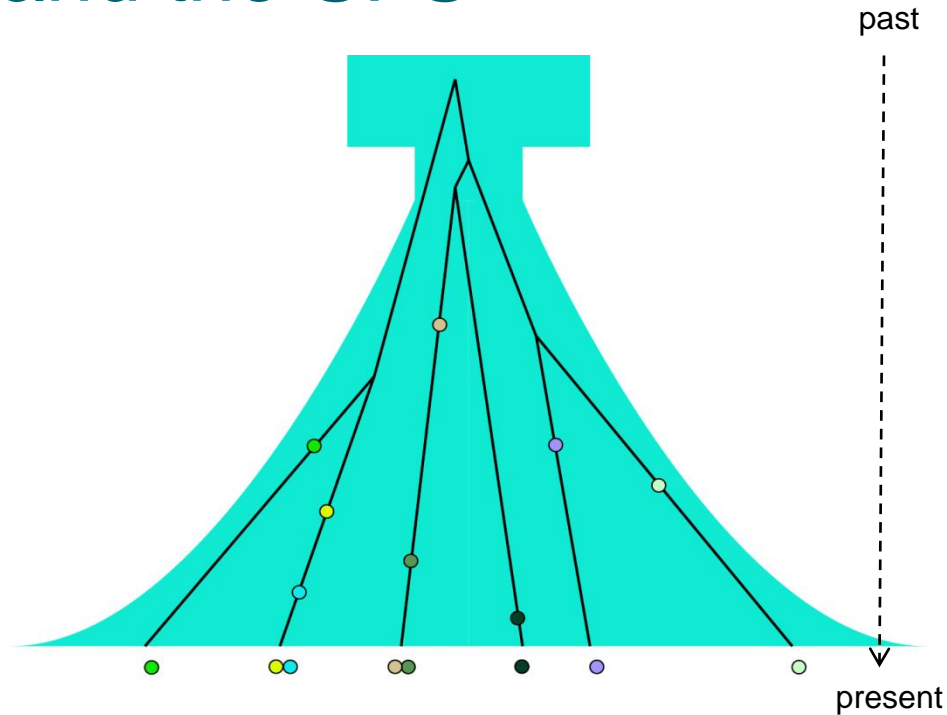
# Coalescent and the SFS

- 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



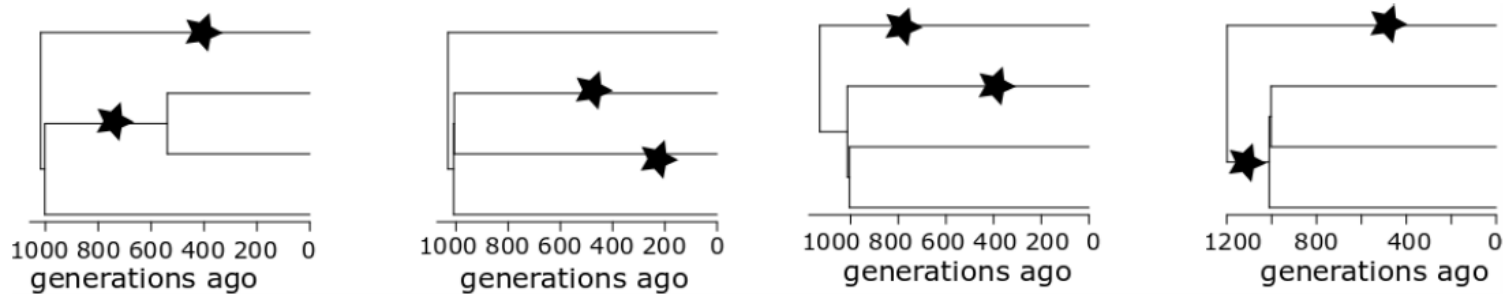
# Coalescent and the SFS

- 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

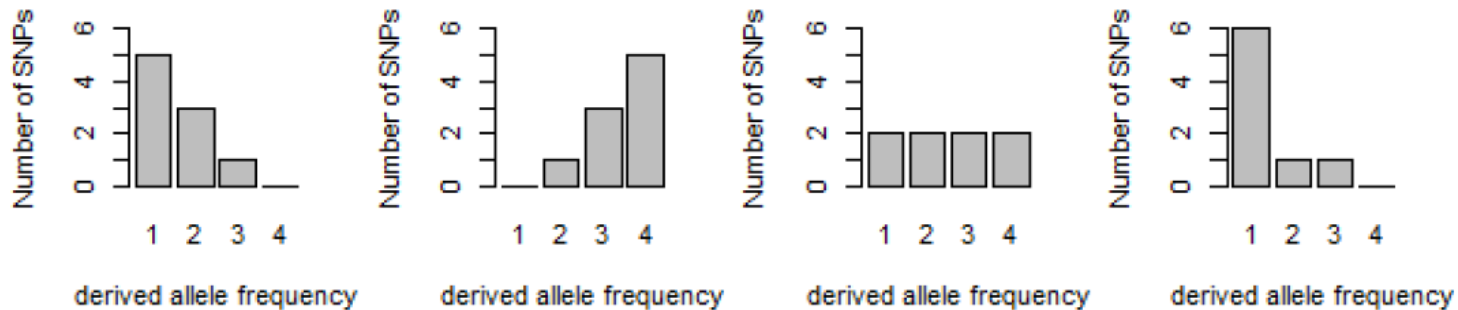


# Exercise - SFS from gene trees at multiple loci

Consider the following gene trees that were simulated for 4 genes with 10,000 bp for a population that went through an expansion. Mutations are indicated with a star.



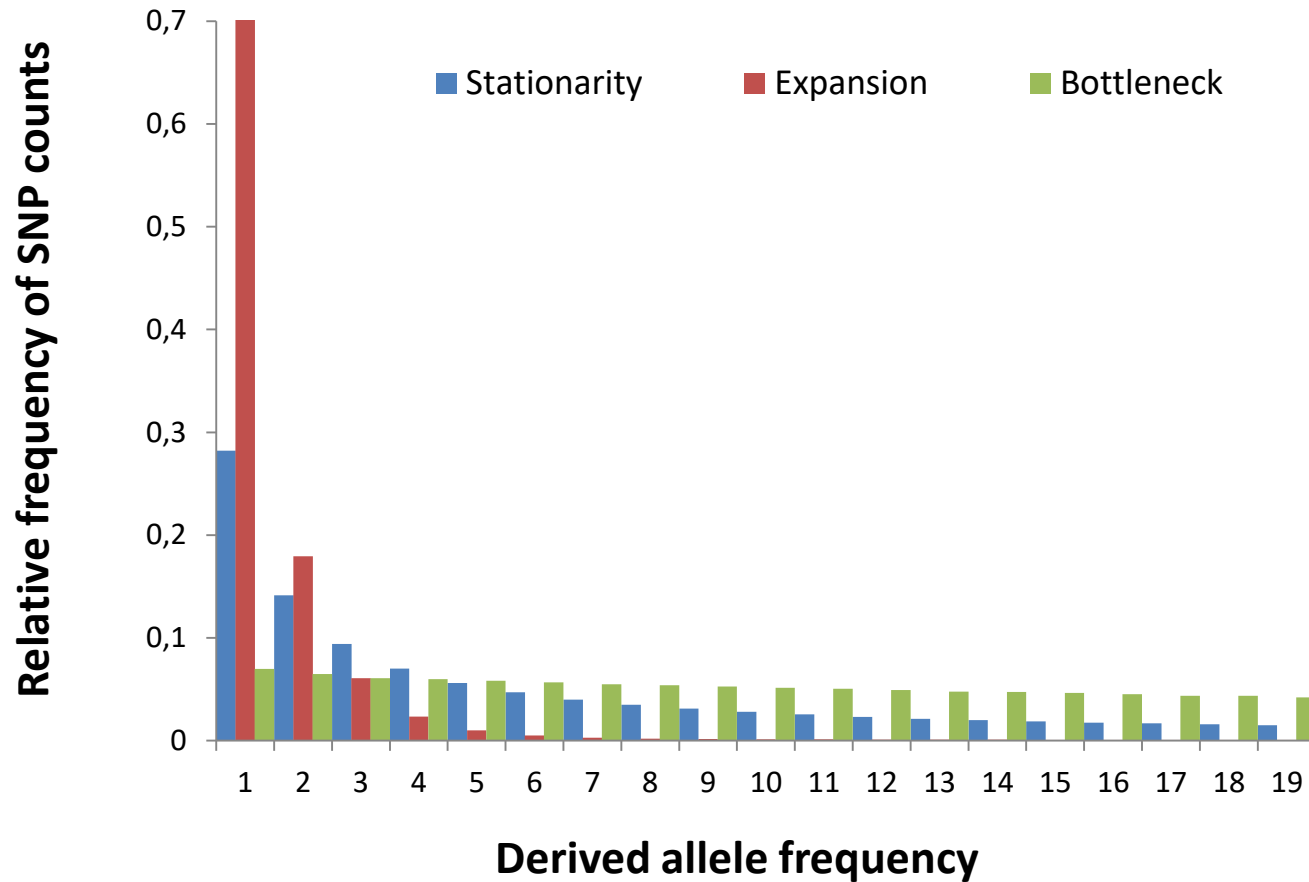
1. Summing across gene trees, what is the site frequency spectrum (SFS) compatible with the above gene trees?



A.	B.	C.	D.
----	----	----	----

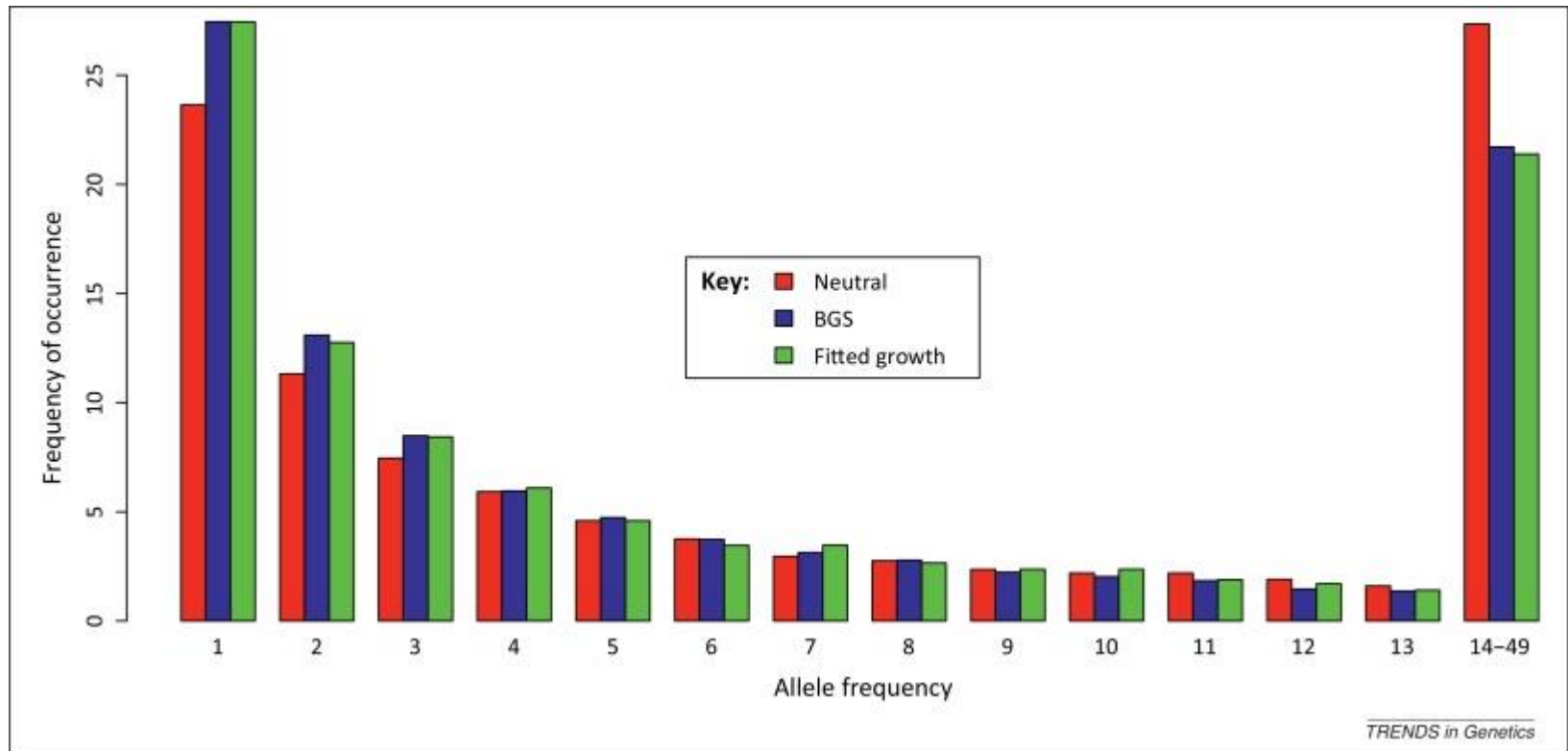
2. Given the gene trees for the 4 genes, what can we learn about the time of expansion?

# SFS depends on past demography





# Natural selection also affects the SFS



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

# Population structure

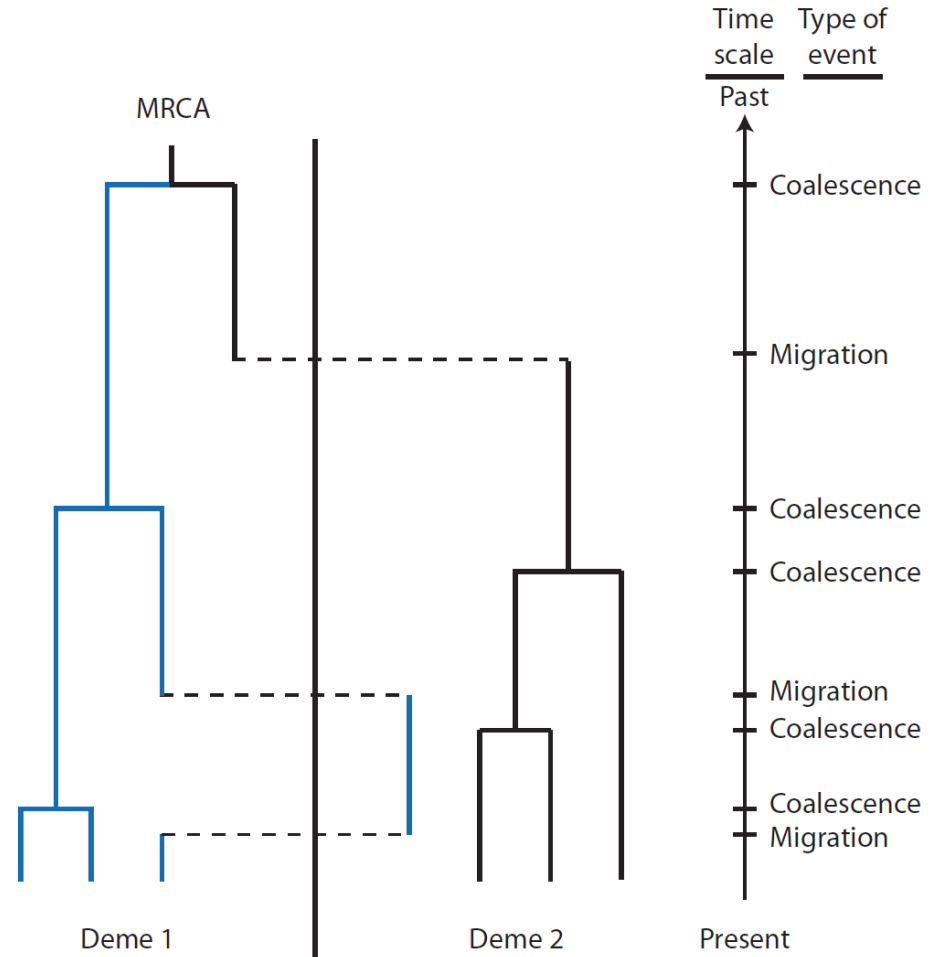
Migration events can be incorporated into gene trees.

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:

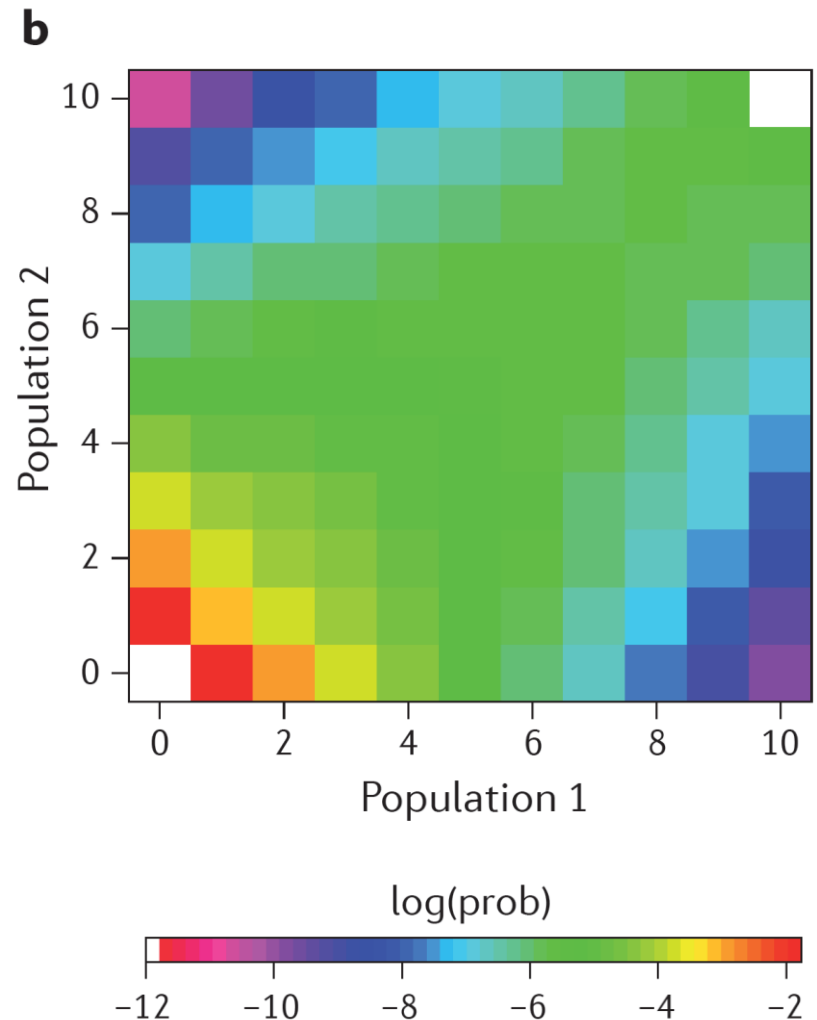
$$\Pr(\text{migrate}) = n_1 * m$$

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.

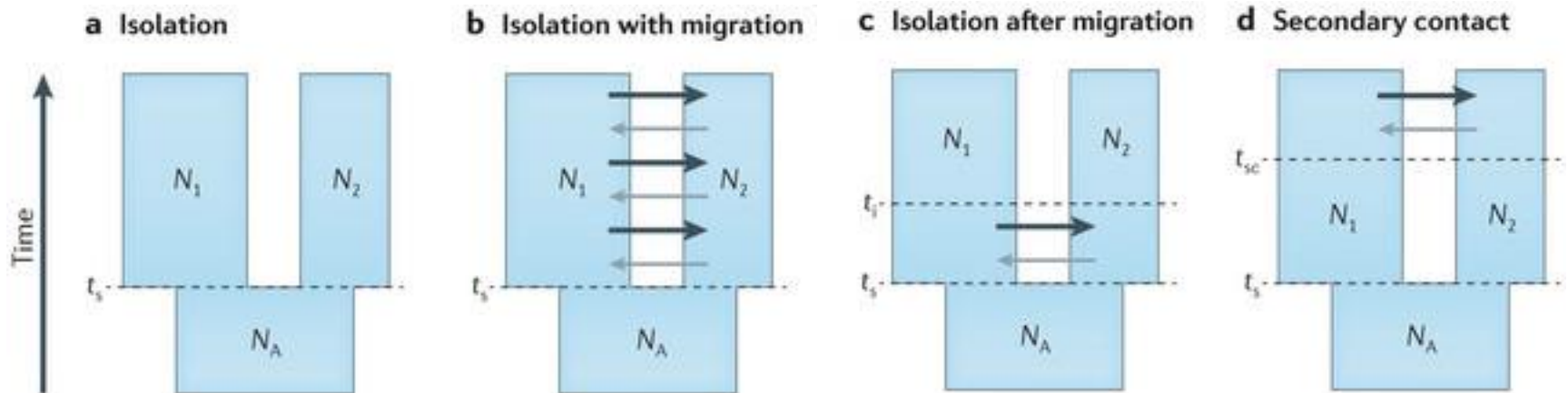


# Derived vs Minor allele frequency spectrum

- 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?

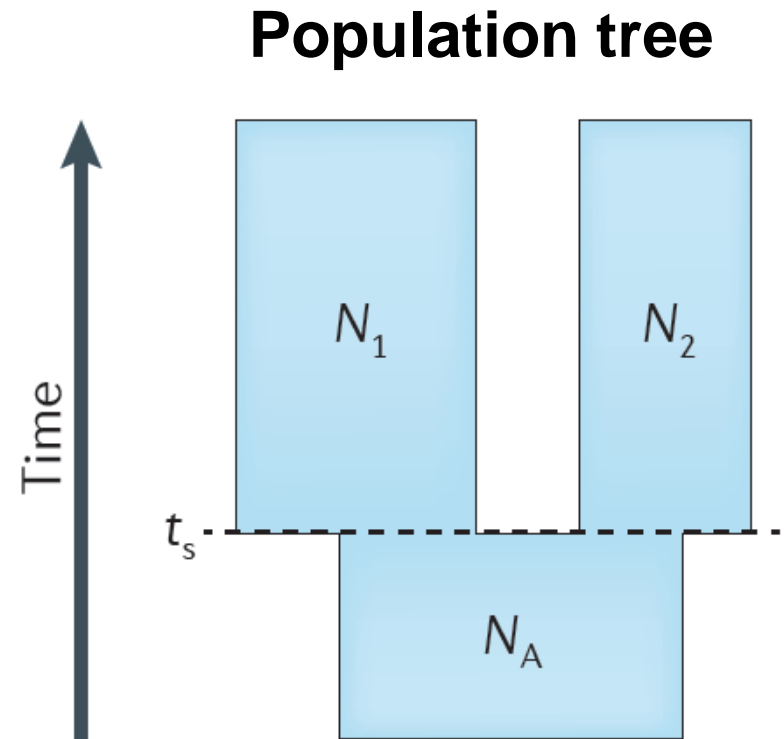


Sousa and Hey (2013) Nat. Rev. Gen.

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

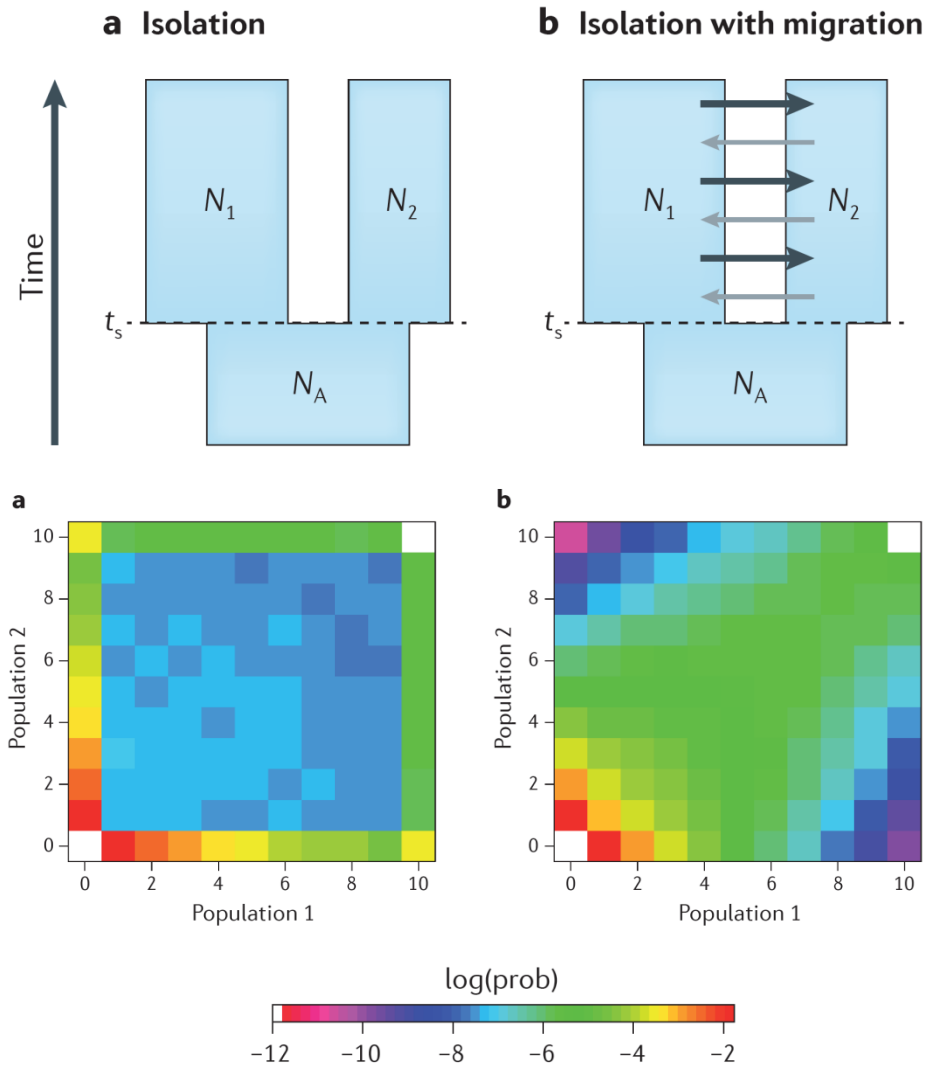
## Parameters:

- 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



# Site frequency spectrum (SFS)

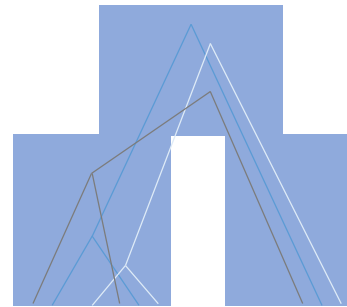
The SFS contains information about the demographic history of populations



# Inferring the demographic history from the SFS

Genomic Data

Model



Parameters:

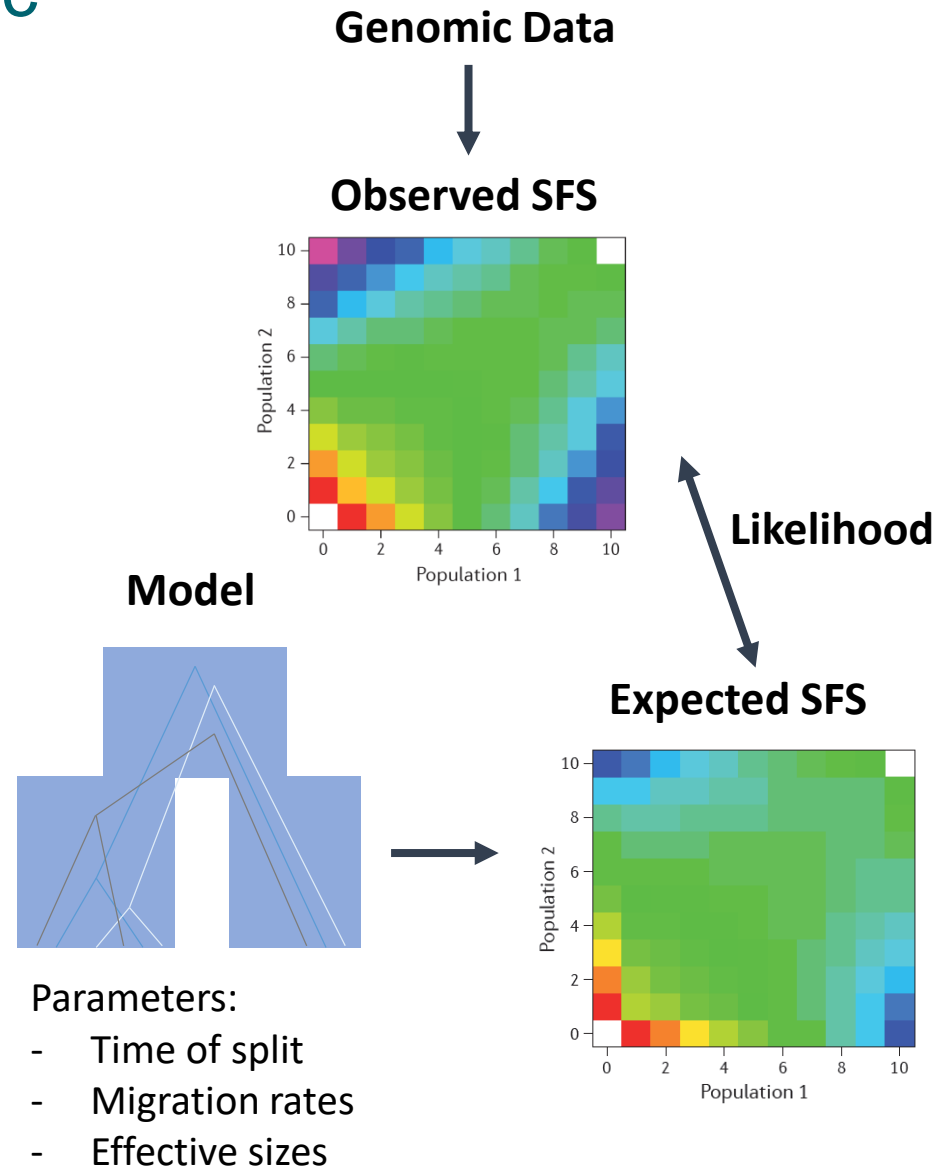
- Time of split
- Migration rates
- Effective sizes





# 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)



# Framework for demographic inference

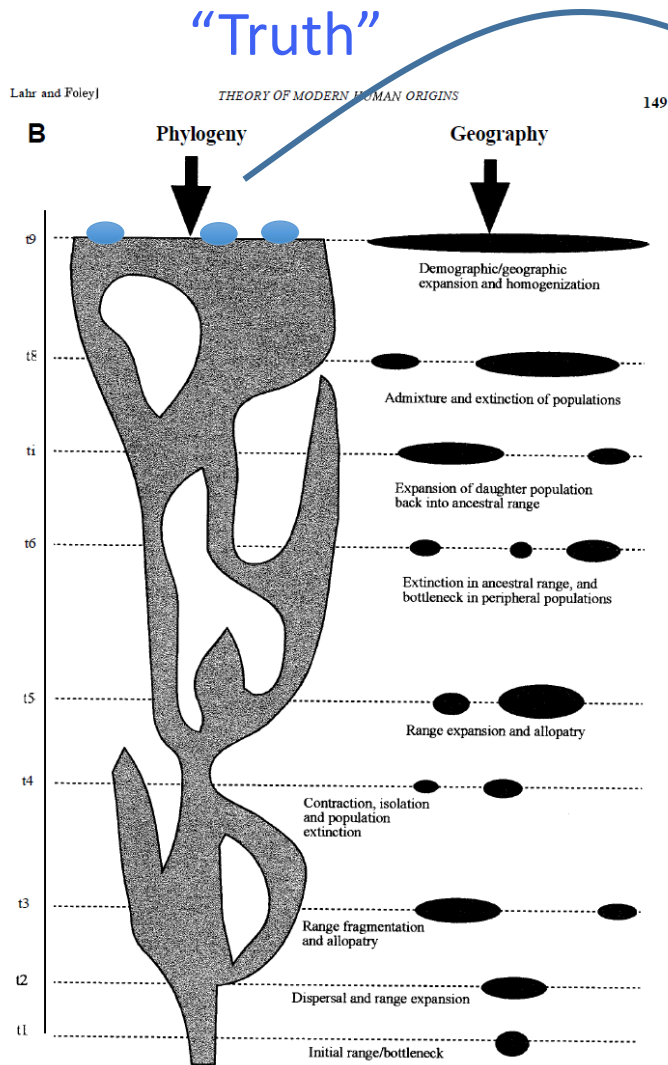


Fig. 1.

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

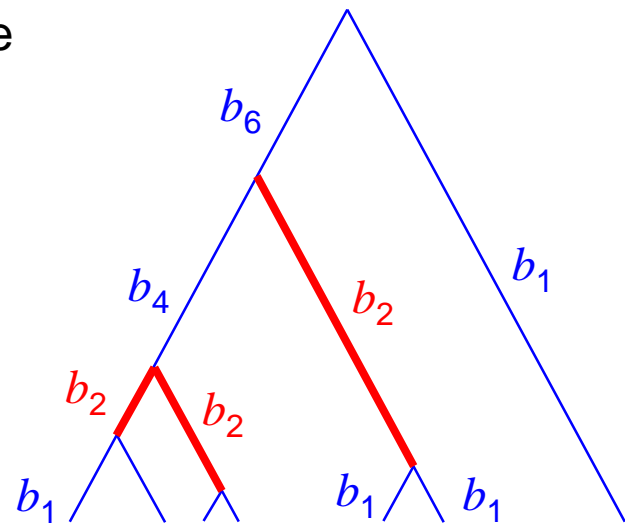
The probability of a SFS entry  $i$  can be estimated under a specific model  $\theta$  from its expected coalescent tree as (Nielsen 2000)

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

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:

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

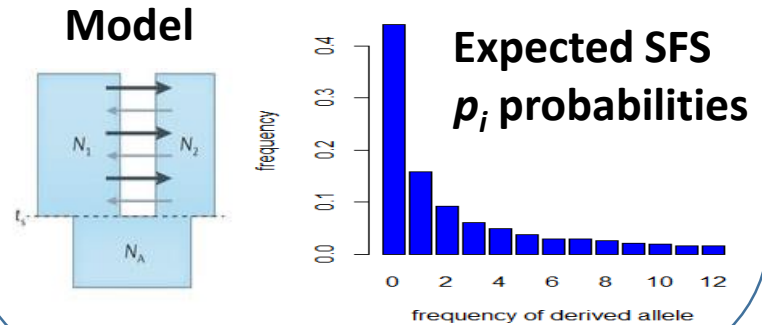
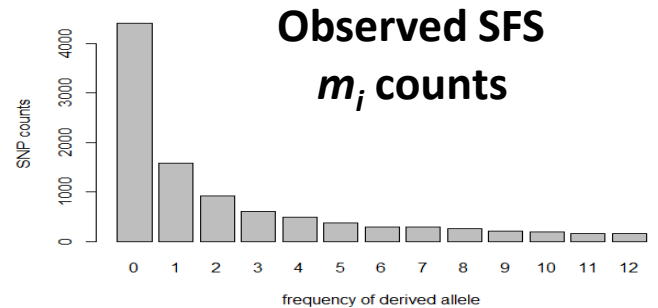
probability of no mutation on the tree

probability of at least one mutation in the tree

These probabilities depend:

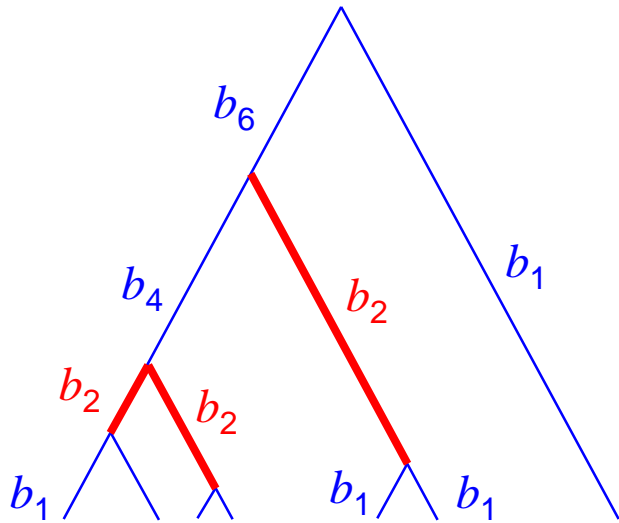
- Number of monomorphic sites
- A fixed and mutation rate

3 ingredients for likelihood

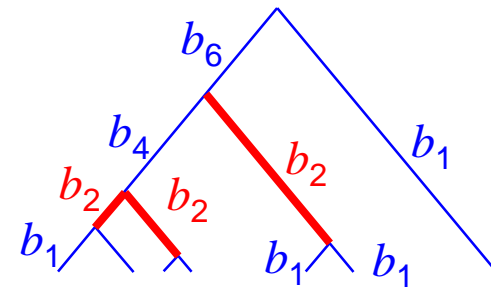


Composite likelihood

# Everything is relative



$T_L$  = total  
branch length



Frequency	0	1	2	3	4	5	6	7
SNP probability $p_i$	0	$\text{Sum}(b_1)/T_L$	$\text{Sum}(b_2)/T_L$	$\text{Sum}(b_3)/T_L$	$\text{Sum}(b_4)/T_L$	$\text{Sum}(b_5)/T_L$	$\text{Sum}(b_6)/T_L$	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  $p_i$  can then be estimated on the basis of  $Z$  simulations as

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

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

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

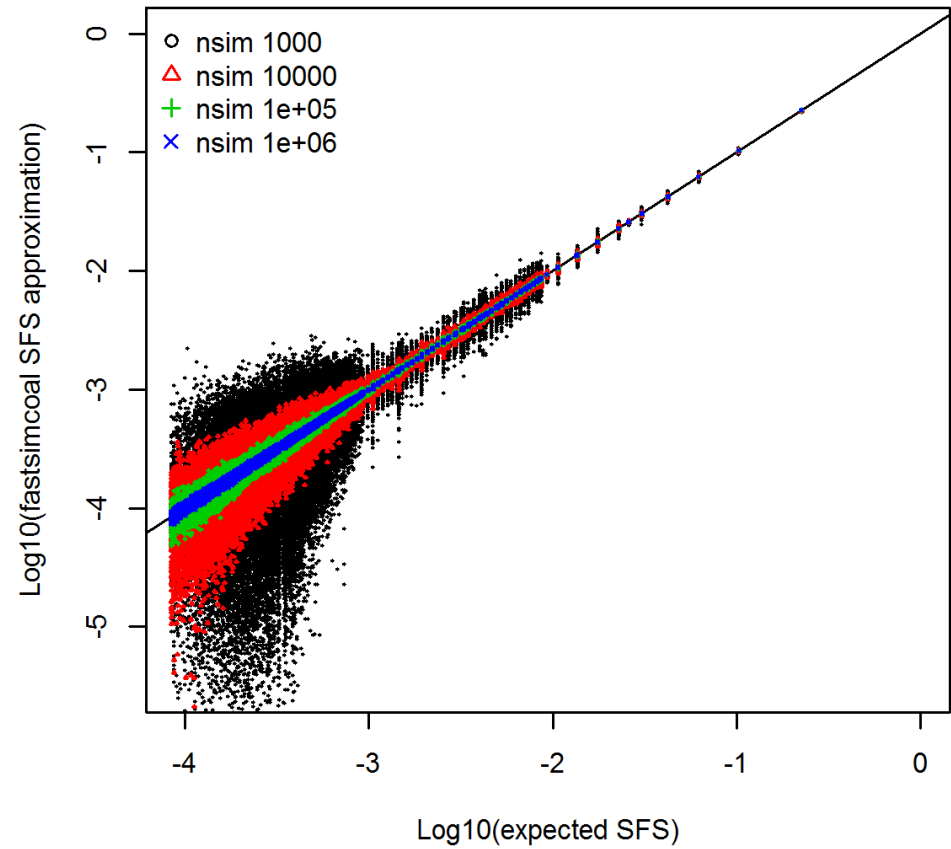
where  $X$  is the SFS in a population sample of size  $n$ ,  $S$  is the number of polymorphic sites,  $L$  is the length of the studied sequence, and  $P_0$  is the probability of no mutation on the tree



# Approximating the expected SFS with coalescent simulations

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

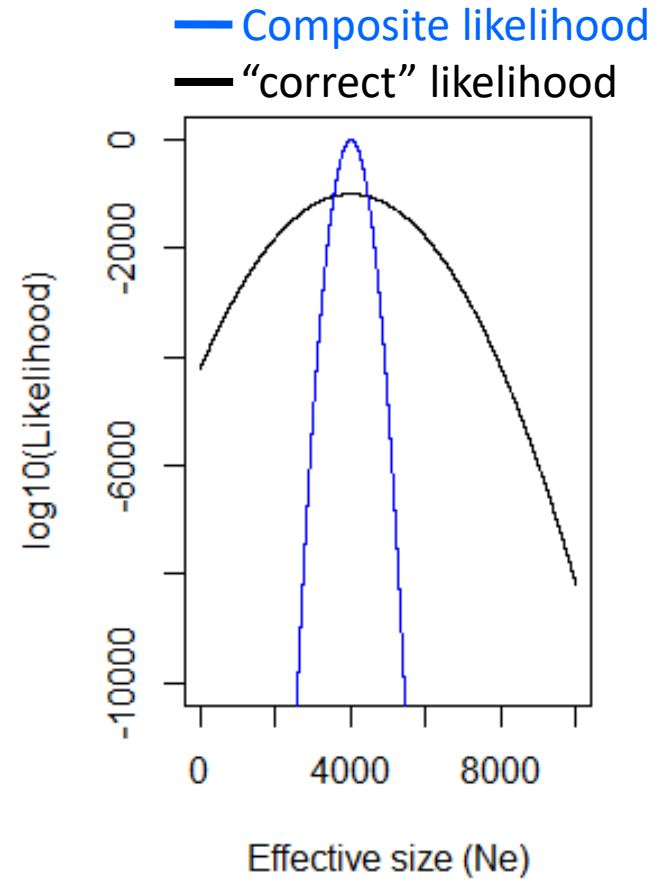
**B** T=0.1 - Expected vs. Fastsimcoal SFS



# Properties of composite likelihoods

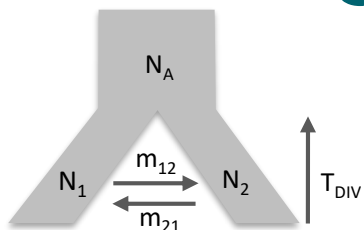
This composite likelihood (CL) is not a proper likelihood due to the non-independence of allele frequencies at linked sites.

- 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

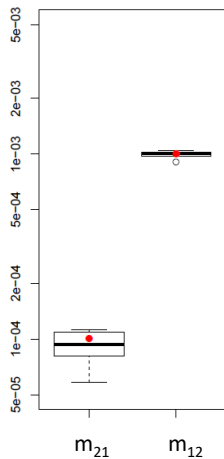
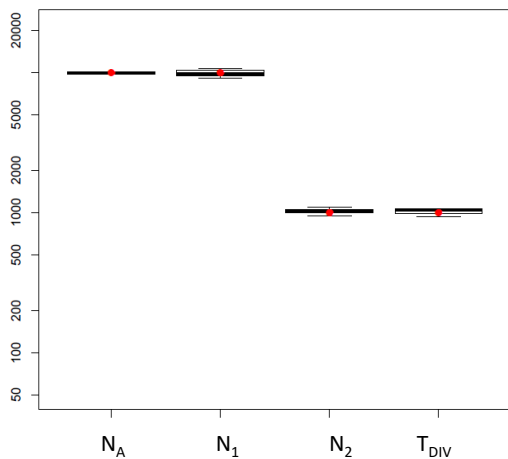


# Comparisons of approaches

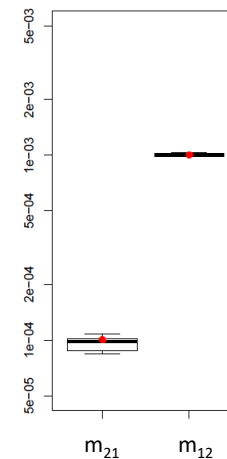
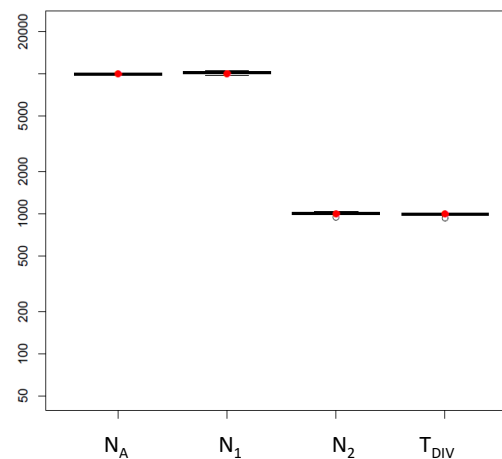
Simulation of 20 Mb data



*fastsimcoal2*



*ada*



# 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, dividing 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<sup>a</sup>, Charles Fefferman<sup>b</sup>, Nick Patterson<sup>a,\*</sup>

<sup>a</sup> Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge MA 02142, United States

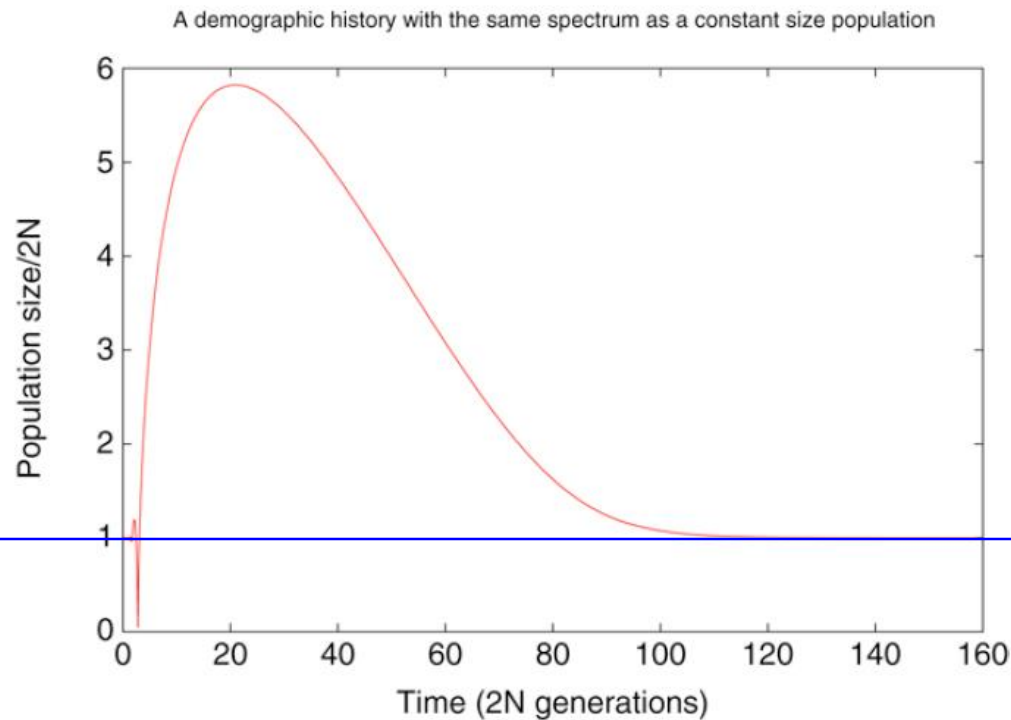
<sup>b</sup> Department of Mathematics, Fine Hall, Washington Road, Princeton, NJ 08544, United States

**Theoretical  
Population  
Biology**

[www.elsevier.com/locate/tpb](http://www.elsevier.com/locate/tpb)

Received 17 March 2007

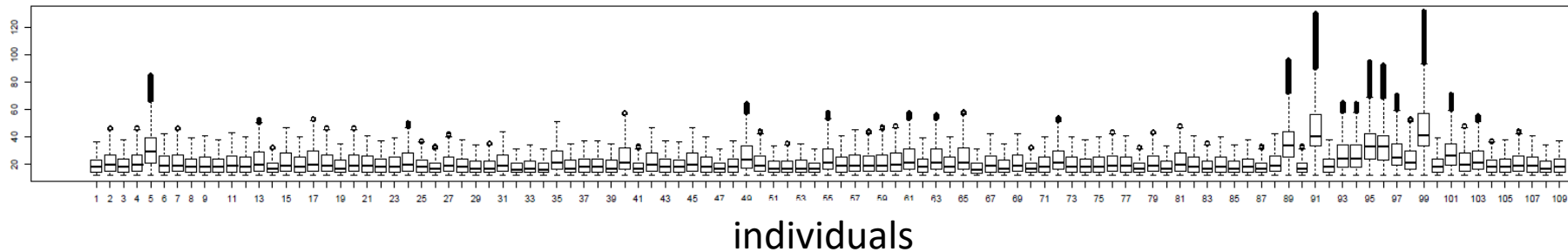
Available online 30 January 2008



# The depth of coverage varied considerably across individuals

DP (depth of coverage)

Example of the DP distribution for each individuals, for individuals with mean DP>12



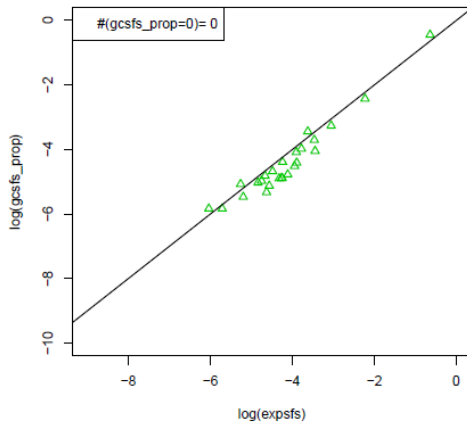
- Applying the same threshold for all individuals can lead to biases
- Apply a filter on DP for each individual

# Effect of DP filters on the SFS

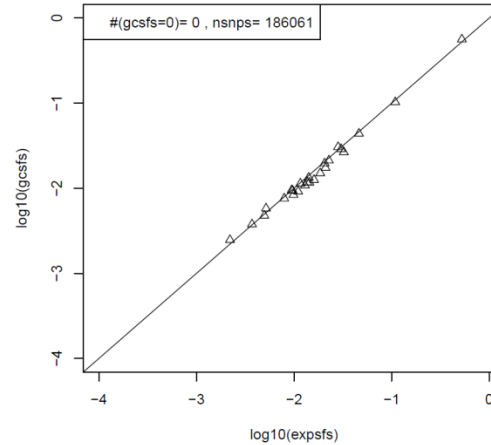
## Simulation study

SFS based on called genotypes

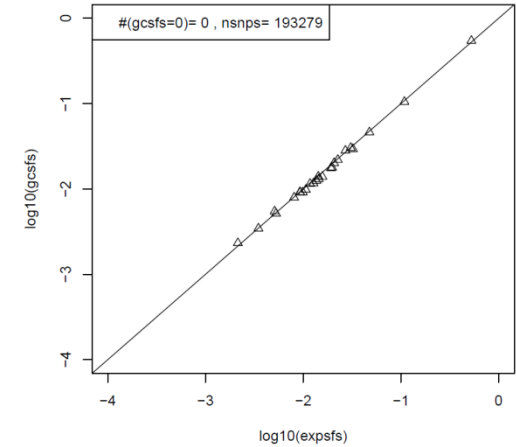
DP > 10



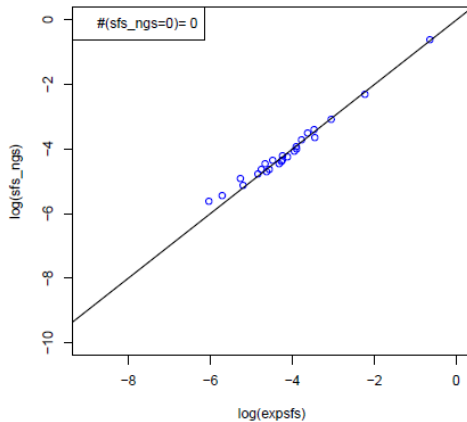
DP > 15



DP > 20



SFS accounting for genotype uncertainty (ANGSD)



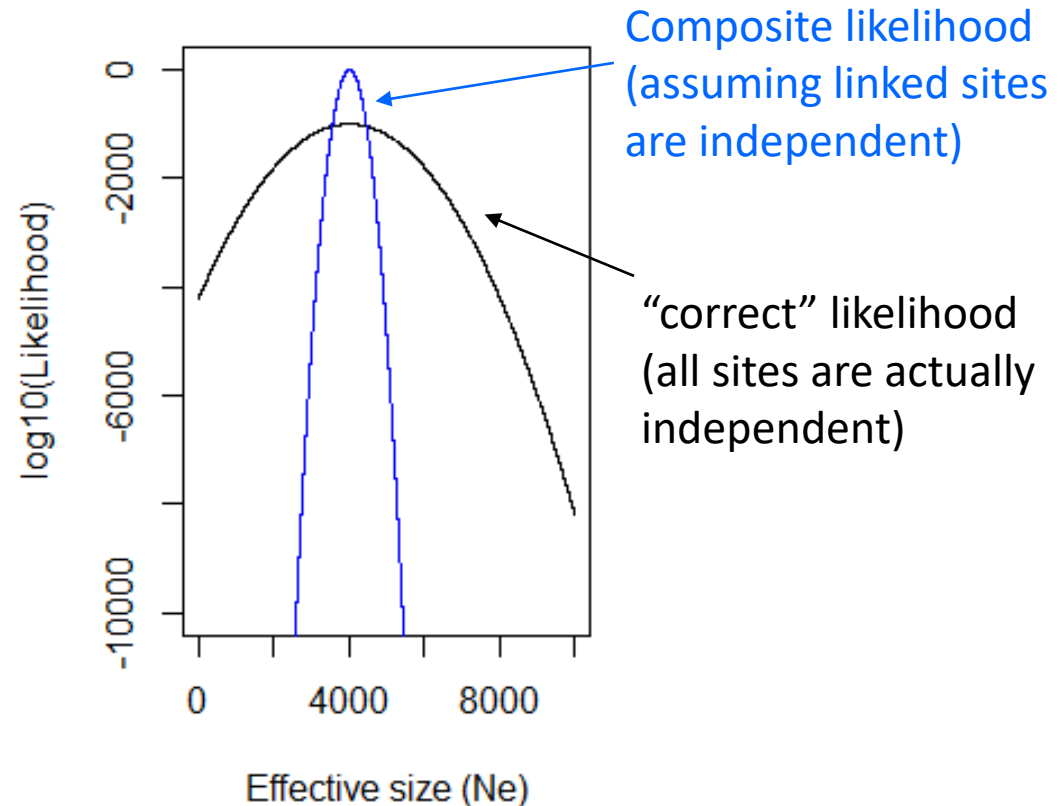
Simulated 2 pops SFS sampling 4 diploids from each pop, 200000 SNPs, mean coverage=**10x**, error rate=0.01. Simulated with correlated allele frequencies model ( $F_{ST}=(0.275, 0.01)$ )

With DP>15 we have a very good approximation to the correct SFS, even when using the called genotypes



# Comparing models with composite likelihoods

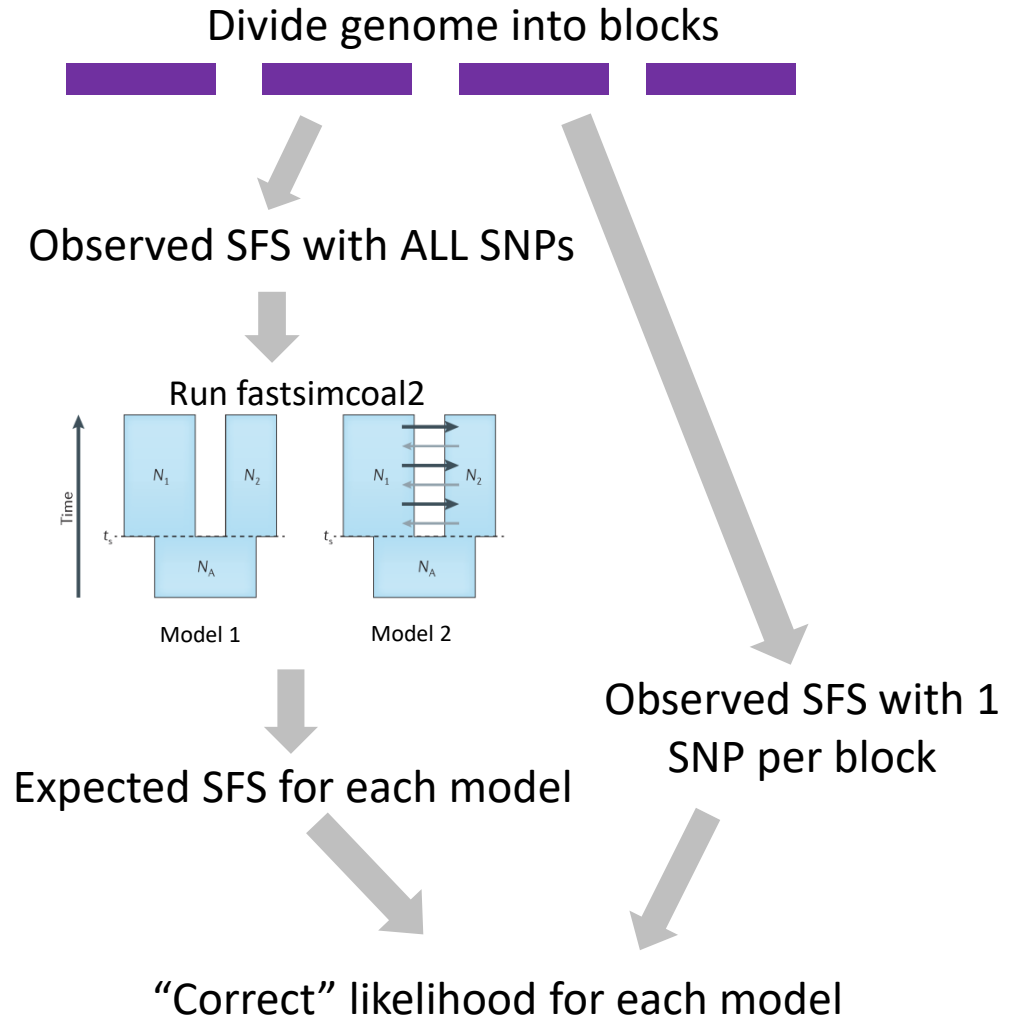
- Fastsimcoal2 likelihood is “correct” if all SNPs are independent
- We can then compare the model likelihoods using Akaike Information Criterion (AIC)



Composite likelihood provide unbiased maximum likelihood parameter estimates, but the likelihoods are inflated

# A strategy to compare models

1. Divide the dataset into LD blocks.
2. Create a dataset with all SNPs (including linked SNPs)
3. For each model, obtain the parameters that maximize the likelihood (this is ok even with linked sites!) and the corresponding expected SFS
4. Create a dataset with “independent” SNPs (1 SNP per RAD tag)
5. Given the expected SFS of each model, compute the “correct” likelihood for each model with the dataset with independent SNPs
6. Compare models with AIC



# Protocol for model comparison based on AIC when we have independent SNPs

- Get the observed SFS
- Define the alternative models
- Perform 50-100 runs under each model
- Select the runs with maximum likelihood under each model
- Compute the AIC (Akaike information criteria) for each model
- Select the model with minimum AIC

# FASTSIMCOAL2 INPUT FILES

Vitor Sousa

[vmsousa@fc.ul.pt](mailto:vmsousa@fc.ul.pt)

**PGDH18**

# Examples of observed SFS

## 1PopExpInst20Mb\_DAFpop0.obs

```
1 observations
d0_0    d0_1    d0_2    d0_3    d0_4    d0_5    d0_6    d0_7    d0_8    d0_9    d0_10
19973842 24630    810     173     145     111     88      84      61      56      0
```

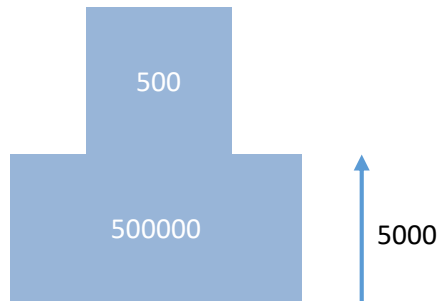
## 2PopDivMigr20Mb\_jointDAFpop1\_0.obs

```
1 observations
      d0_0    d0_1    d0_2    d0_3    d0_4    d0_5
d1_0 19985747 8350   1628   360    62     8
d1_1  9660    0      0      0      0
d1_2  4790    0      0      0      0
d1_3  3280    0      0      0      0
d1_4  2490    0      0      0      0
d1_5  1760    13     18     13     19     0
```

## 2PopDiv20Mb\_jointDAFpop1\_0.obs

```
1 observations
      d0_0    d0_1    d0_2    d0_3    d0_4    d0_5
d1_0 19985547 8211   1415   316    55    10
d1_1  1266    101    37     16     5     1
d1_2  61142    20     8      2      0
d1_3  48631    12     5      0      0
d1_4  47915    9      2      3      1
d1_5  1189     46     22     19     18     0
```

# Parameter estimation settings files



1PopExpInst20Mb

Additional files necessary to estimate parameters

## Estimation file

1PopExpInst20Mb/1PopExpInst20Mb.est

```
// Search ranges and rules file
// *****

[PARAMETERS]
//#isInt? #name #dist.#min #max
//all Ns are in number of haploid individuals
1 NPOP logunif 1000 1e7 output
1 NANC logunif 10 1e5 output
1 TEXP unif 10 1e5 output

[RULES]

[COMPLEX PARAMETERS]

0 RESIZE = NANC/NPOP hide
```

## Template file

1PopExpInst20Mb/1PopExpInst20Mb.tpl

```
//Parameters for the coalescence simulation program : fsimcoal2.exe
1 samples to simulate :
//Population effective sizes (number of genes)
NPOP
//Samples sizes and samples age
10
//Growth rates: negative growth implies population expansion
0
//Number of migration matrices : 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix index
1 historical event
TEXP 0 0 0 RESIZE 0 0
//Number of independent loci [chromosome]
1 0
//Per chromosome: Number of contiguous linkage Block: a block is a set of contiguous loci
1
//per Block:data type, number of loci, per generation recombination and mutation rates and optional parameters
FREQ 1 0 2.5e-8 OUTEXP
```

# INPUT files for fastsimcoal2: Defining an evolutionary model with TPL files

Number of samples  
to simulate

**2PopDivMigr10Loci.par**

```
//Parameters for the coalescence simulation program : fsimcoal2.exe
```

```
2 samples to simulate :
```

```
//Population effective sizes (number of genes)
```

```
NPOP1
```

```
NPOP2
```

```
//Samples sizes and samples age
```

```
5
```

```
5
```

```
//Growth rates: negative growth implies population expansion
```

```
0
```

```
0
```

Growth rates

```
//Number of migration matrices : 0 implies no migration between demes
```

```
2
```

Migration  
matrices

```
//Migration matrix 0
```

```
0 0
```

```
MIG10 0
```

```
//Migration matrix 1: No migration
```

```
0 0
```

```
0 0
```

Historical events

```
//historical event: time, source, sink, migrants, new RELATIVE deme size of sink, new growth rate,
```

```
NEW migration matrix index
```

```
2 historical event
```

```
TMIG 0 0 0 1 0 1
```

```
TDIV 1 0 1 RELSIZE 0 1
```

```
//Number of independent loci [chromosome]
```

```
1 0
```

FREQ indicates SFS  
data

```
//Per chromosome: Number of contiguous linkage Block: a block is a set of contiguous loci
```

```
1
```

```
//per Block:data type, number of loci, per generation recomb. and mut. rates and optional parameters
```

```
FREQ 1 0 2.5e-8 OUTEXP
```

Define mutation rate!!

OUTEXP indicates we want to output the expected SFS

# TPL files

TPL define the model using parameter tags.

These files are very important! Check carefully all the definitions. Errors in the TPL file are difficult to detect and imply the model specification is incorrect! This means that all inferences will be wrong, and also that all parameter estimates will be incorrect!

## Defining population sizes and sample sizes

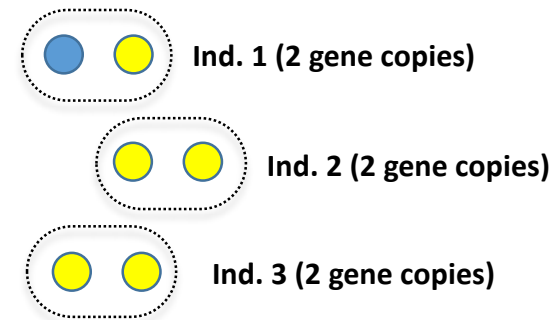
### 2PopDivMigr10Loci.par

```
//Parameters for the coalescence simulation program : fsimcoal2.exe
2 samples to simulate :
//Population effective sizes (number of genes)
NPOP1
NPOP2
//Samples sizes and samples age
6
6
//Growth rates: negative growth implies population expansion
0
0
```

Parameter tags

Population effective sizes are given in number of gene copies. For a diploid species with  $N=500$  individuals, this corresponds to a  $2N=1000$  gene copies, as each individual carries two gene copies at any given site.

The sample size is also given in gene copies. The value of 6 means that we sampled 3 diploid individuals.



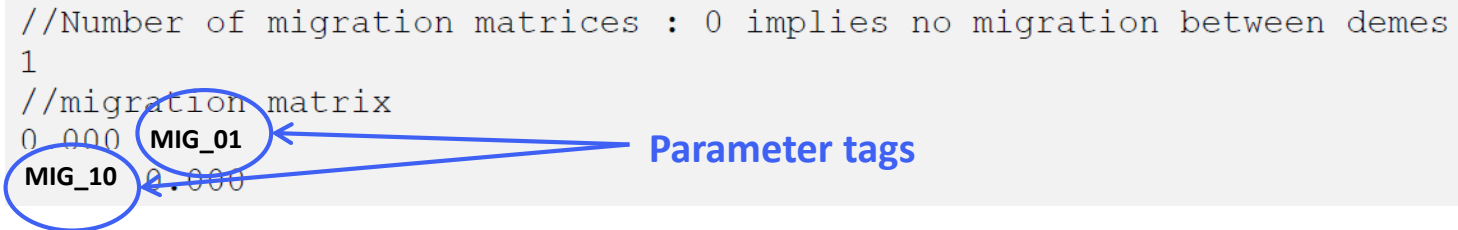


# TPL files

## MIGRATION

```
//Number of migration matrices : 0 implies no migration between demes
1
//migration matrix
0 0.000 MIG_01
MIG_10 0.000
```

Parameter tags



The migration matrix can be asymmetric, and in the case the entry  $m_{ij}$  list the **migration rates backward in time** from population  $i$  to population  $j$ . The above-mentioned matrix states that, for each generation backward in time, any gene from population 0 has probability MIG\_01 to be sent to population 1, and that a gene from population 1 has a probability MIG\_10 to move to population 0.

If no migration matrix is defined, no migration is assumed between populations.

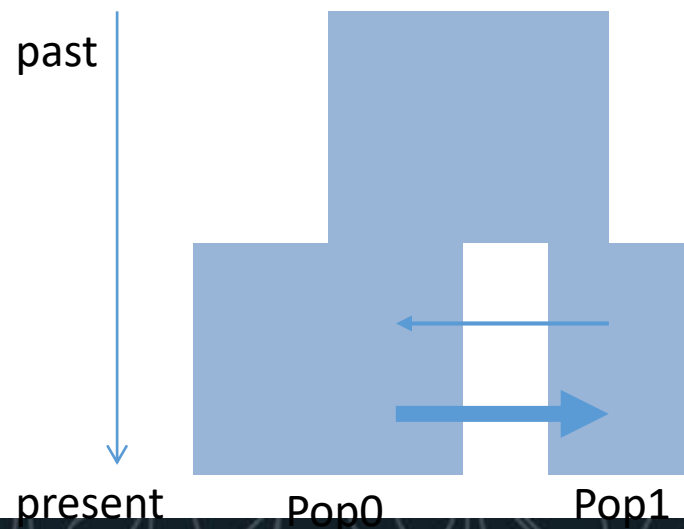
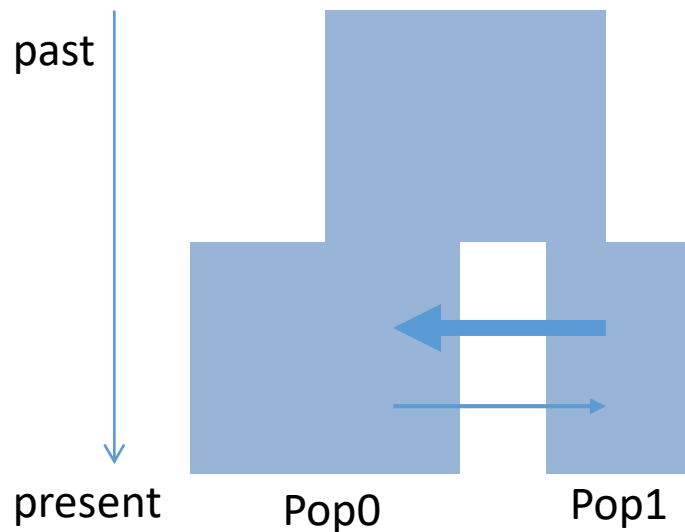
1PopStationary10Loci.par

```
//Number of migration matrices : 0 implies no migration between demes
0
```

# A note on looking backward in time

Assuming that we look forward in time and that the size of the arrows are proportion to the migration rate, to what model does the following migration matrix corresponds to?

```
//Number of migration matrices : 0 implies no migration between demes  
1  
//migration matrix  
0.000 0.005  
0.001 0.000
```

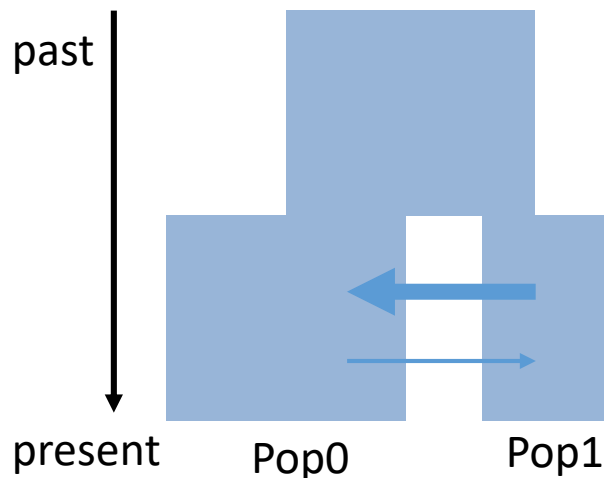


# A note on looking backward in time

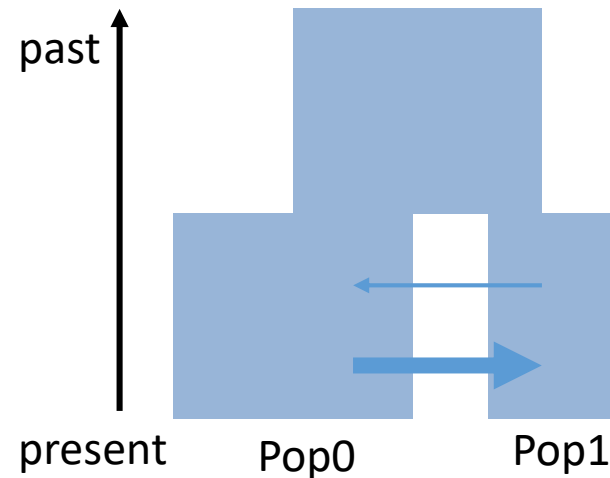
Assuming that we look forward in time and that the size of the arrows are proportion to the migration rate, to what model does the following migration matrix corresponds to?

```
//Number of migration matrices : 0 implies no migration between demes  
1  
//migration matrix  
0.000 0.005  
0.001 0.000
```

**Note that in the PAR and TPL files everything is backward in time!!**



**This is the correct model forward in time, meaning there are more migrants moving from pop1 to pop0 each generation.**



**Backward in time this is the model. Lineages are more likely to move from pop0 to pop1.**

# Historical events in fastsimcoal2

Historical events can be used to:

- Change the size of a given population
- Change the growth rate of a given population
- Change the migration matrix to be used between populations
- Move a fraction of the genes of a given population to another population. This amounts to implementing a (stochastic) admixture or introgression event.
- Move all genes from a population to another population. This amounts to fusing two populations into one looking backward in time.
- One or more of these events at the same time

**Defining the historical events is crucial to have a correct model!**



# Historical events (backward in time)

Each historical event is coded with a line with the following arguments

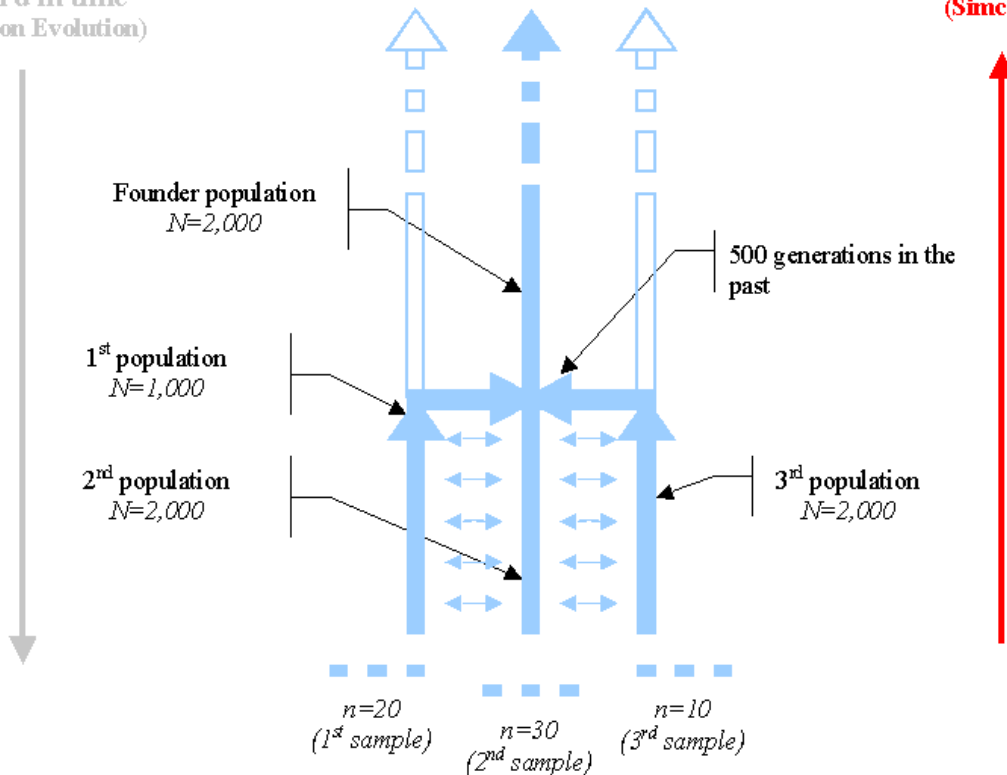
**time**, **source**, **sink**, **migrants**, **new deme size**, **new growth rate**, **migration matrix index**

500 0 1 1 1 0 1  
 500 2 1 1 1 0 1

500 generations ago, 100% (migrants=1.0) of lineages in pop0 (source =0) migrated to pop1 (sink=1). The size of the sink (pop1) remained the same (new deme size=1.0, i.e. N2=2000). The new growth rate is zero. The migration rate that is active after the event is given in the migration matrix 1.

Forward in time  
(Population Evolution)

Backward in time  
(Simcoal2)



# Historical events (backward in time)

Each historical event is coded with a line with the following arguments

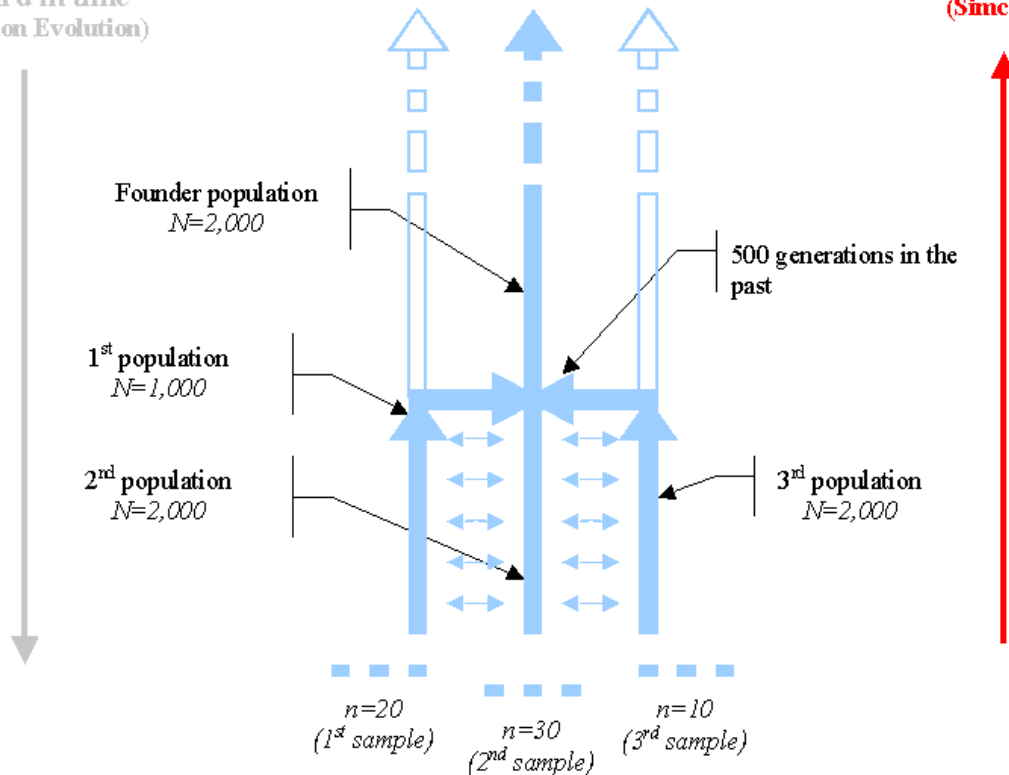
**time**, **source**, **sink**, **migrants**, **new deme size**, **new growth rate**, **migration matrix index**

500 0 1 1 1 0 1

500 2 1 1 1 0 1

Forward in time  
(Population Evolution)

Backward in time  
(Simcoal2)



500 generations ago, 100% of lineages (**migrants=1.0**) in **pop2** (**source =2**) migrated to **pop1** (**sink=1**). The size of the sink (pop1) remained the same (**new deme size=1.0**, i.e.  $N_2=2000$ ). The new growth rate is zero. The migration rate that is active after the event is given in the migration matrix 1.

# Historical events in fastsimcoal2

## Change the size of a given population

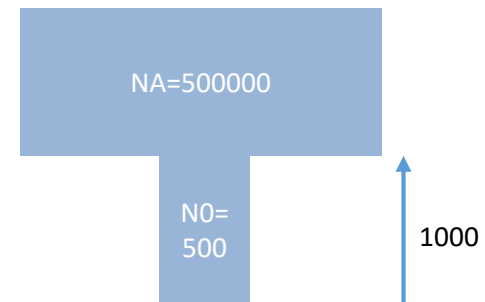
### 1PopContrInst10Loci.par

```
//Parameters for the coalescence simulation program : fsimcoal2.exe
1 samples to simulate :
//Population effective sizes (number of genes)
1000
//Samples sizes and samples age
10
//Growth rates: negative growth implies population expansion
0
//Number of migration matrices : 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix index
1 historical event
1000 0 0 0 1000 0 0
```



- 1000 generations ago, 0% (migrants=0) of lineages in pop0 (source) migrated to pop1 (sink). This means that 100% of lineages remained in pop0.
- The sink population (pop0) has a size 1000 larger after the event (new size=1000). Given that  $N_0=500$  diploids at time zero, it implies that  $N_A=500000$  diploids.
- The migration matrix valid after the event is the migration rate 0. Since it is not defined it implies no migration.

Recent instantaneous demographic contraction



1PopContrInst10loci.par

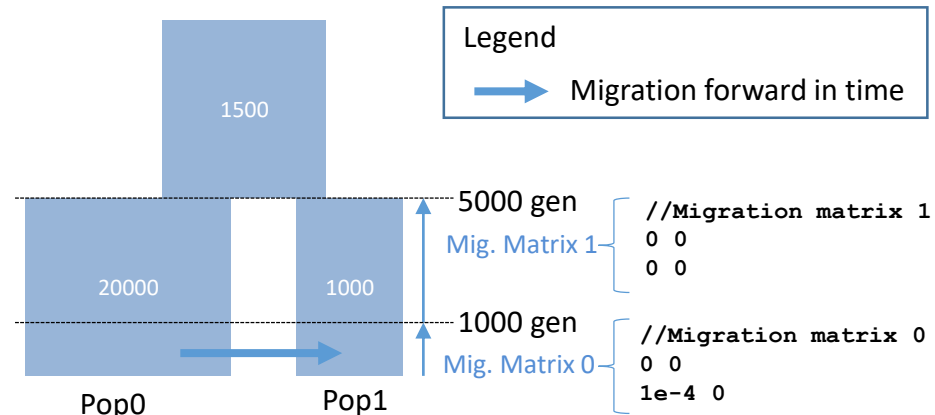
# Historical events in fastsimcoal2

Change the migration matrix to be used between populations

2PopDivMigr10Loci.par

```
//Number of migration matrices : 0 implies no migration between demes
2
//Migration matrix 0
0 0
1e-4 0
//Migration matrix 1: No migration
0 0
0 0
//historical event: time, source, sink, migrants, new RELATIVE deme size, new growth rate, migration
matrix index
2 historical event
1000 0 0 0 1 0 1
5000 1 0 1 1.5 0 1
```

- At generation 1000 in the past, 0% (migrants=0) of lineages migrated from pop0 (source=0) to pop1 (sink=0).
- After the historical event, the deme size of the sink population (pop1) remained the same (new deme size=1).
- After the historical event the growth rate was set to zero.
- After the historical event the migration rate matrix was set to matrix 1, i.e. no migration between populations.





# Historical events in fastsimcoal2

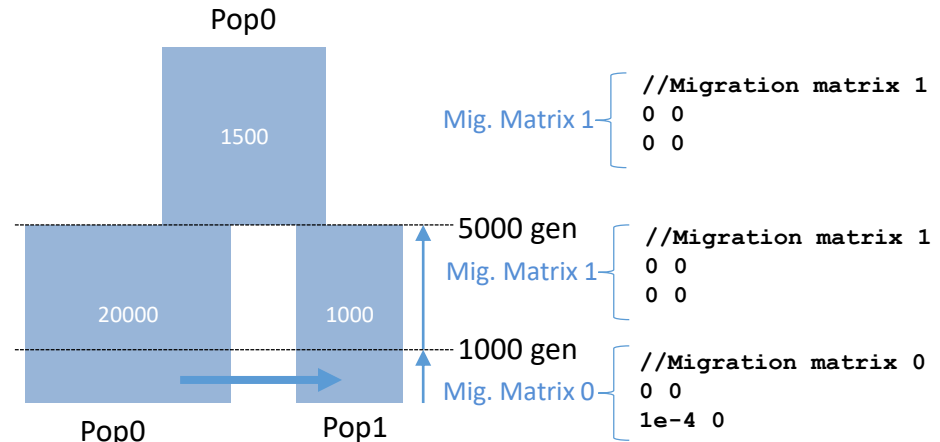
## Population split (merge populations going backwards in time)

### 2PopDivMigr10Loci.par

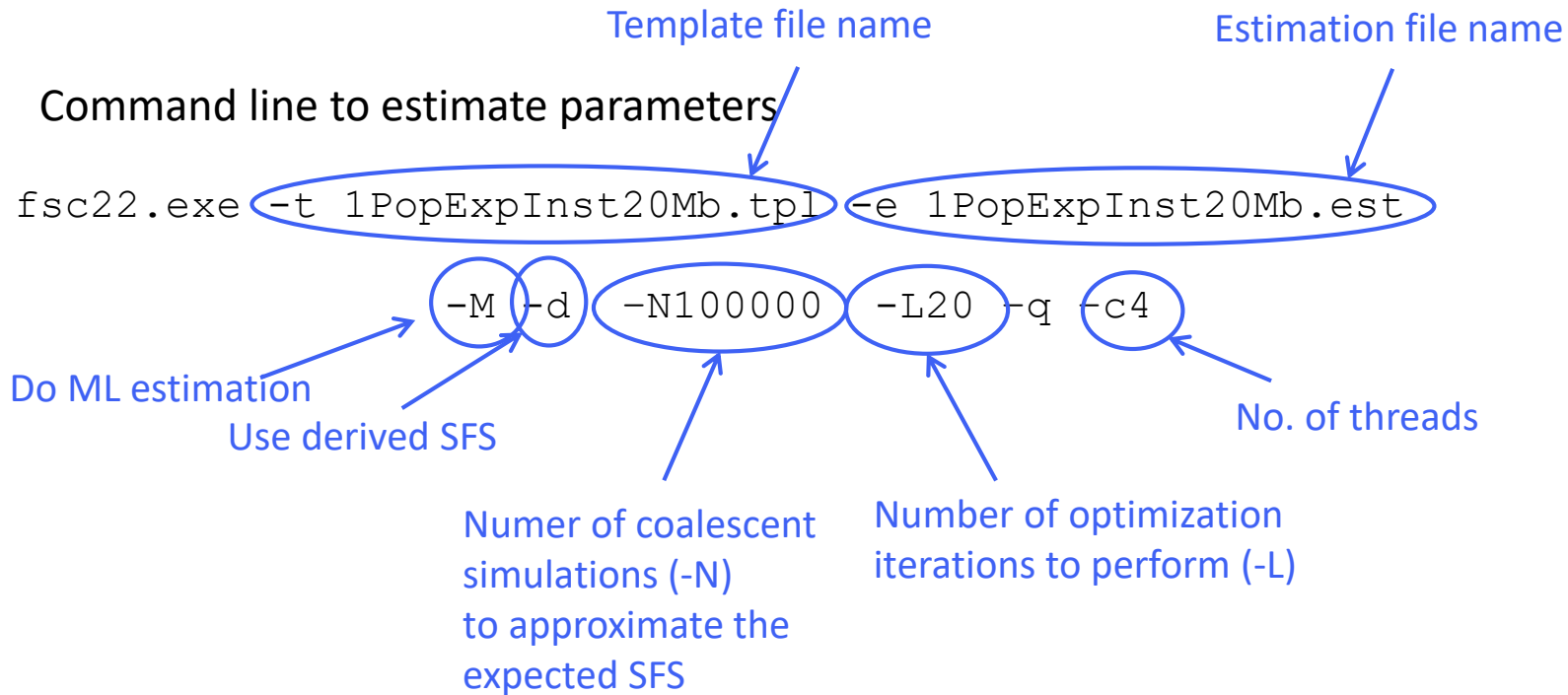
```
//Number of migration matrices : 0 implies no migration between demes
2
//Migration matrix 0
0 0
1e-4 0
//Migration matrix 1: No migration
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
index
2 historical event
1000 0 0 0 1 0 1
5000 1 0 1 0.075 0 1
```



- At generation 5000 in the past, 100% (migrants=1) of lineages migrated from pop1 (source=1) to pop0 (sink=0).
- After the population split, the deme size of the sink population (pop0) is 1500 (new deme size=1500/20000=0.075).
- After the historical event the growth rate of the sink population pop0 is zero.
- After the historical event the migration rate matrix was set to matrix 1, i.e. no migration between populations.



# Launching parameter estimations



Observed SFS file must have the same name as template file and extension  
\_DAFpop0.obs. e.g. `1PopExpInst20Mb_DAFpop0.obs`