Demographic modeling with fastsimcoal

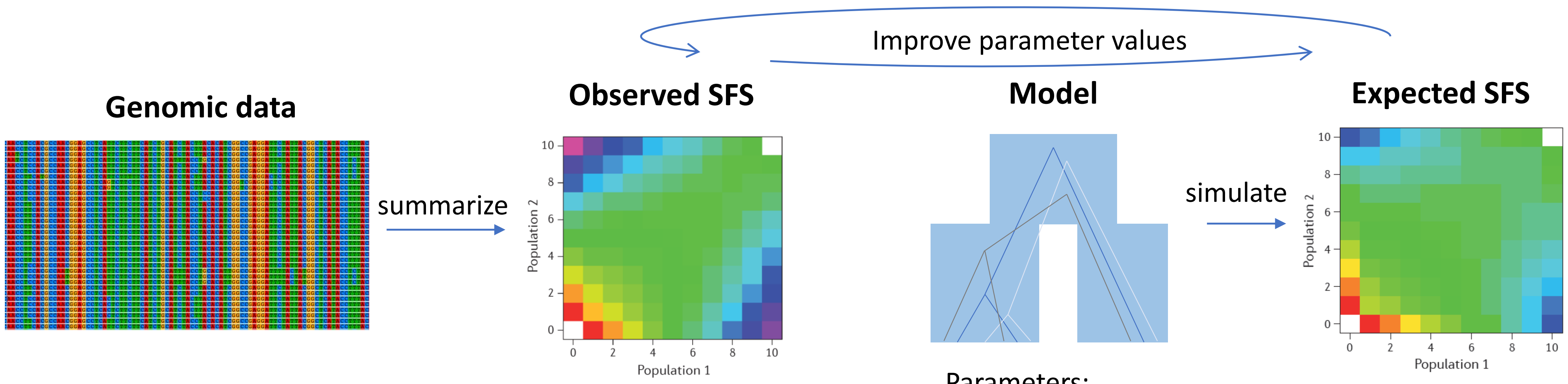
Joana Meier

(some slides are adapted from Vitor Sousa, CE3C, Lisbon, Portugal)

Aims and principle of demographic modeling

Test which of different evolutionary scenarios fits the data best

Estimate model parameters such as strenght of gene flow, divergence time, or population sizes



Compare

Improve parameter values

Genomic data

summarize

Observed SFS

Model

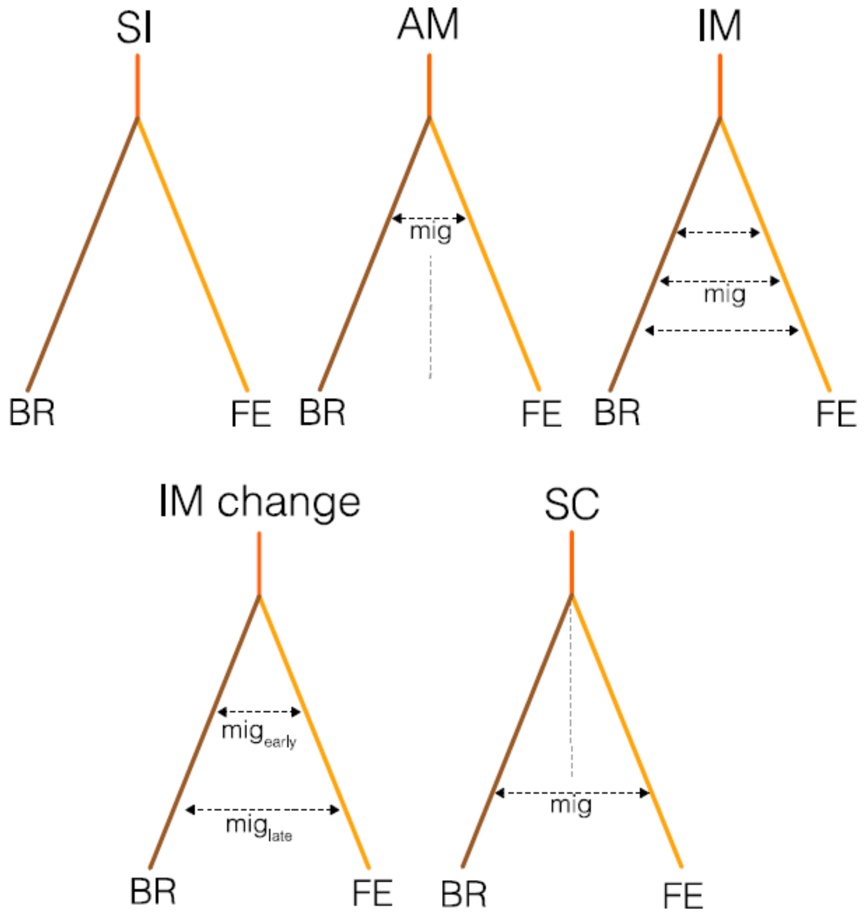
simulate

Expected SFS

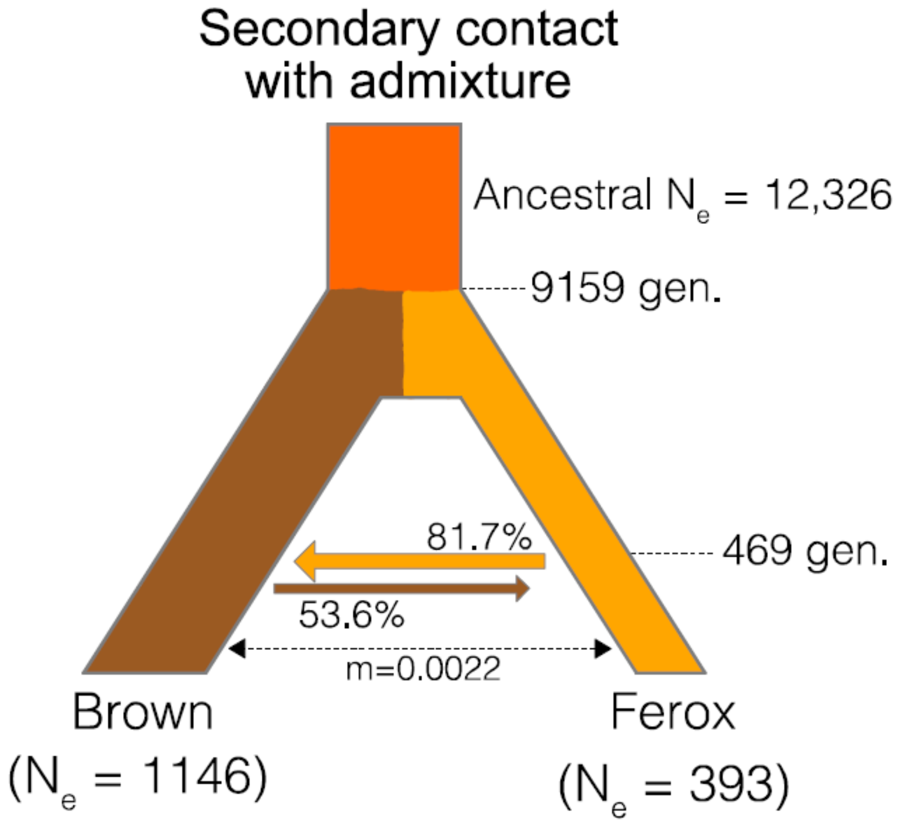
Parameters:

- Time of split
- Migration rates
- Effective sizes

Example 1: Did the rare piscivorous brown trout (ferox) in Scotland evolve in the face of gene flow with normal brown trout or in allopatry?

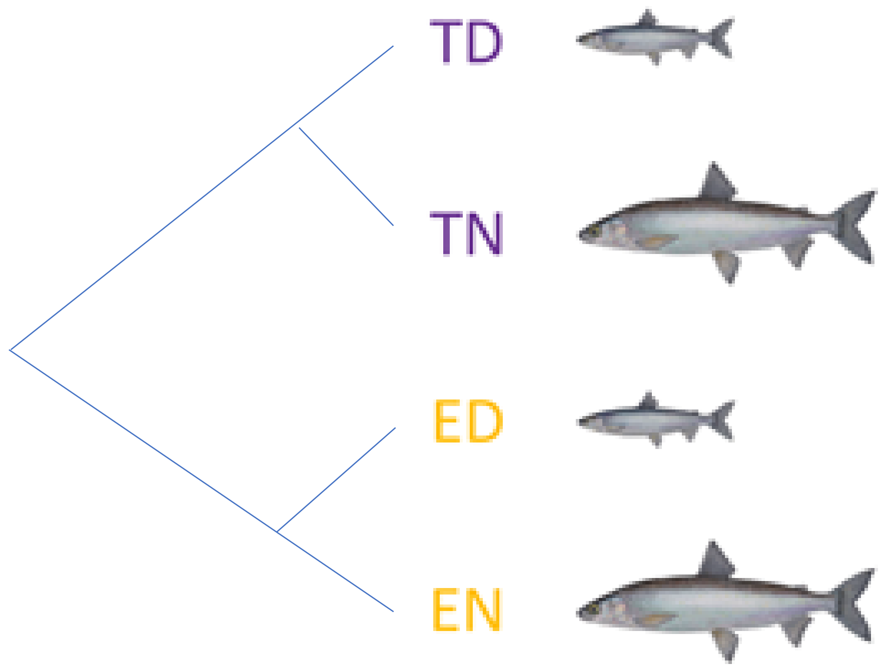


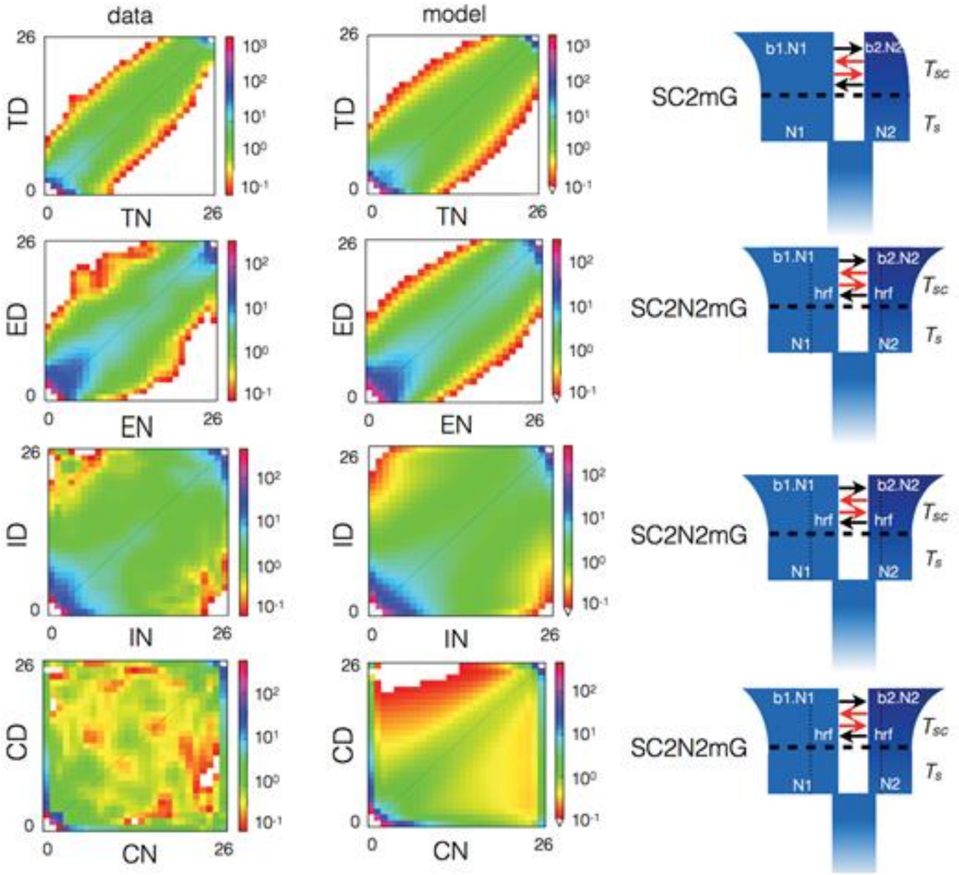
Best model:



Jacobs et al., Genes, 2018

Example 2: Did dwarf limnetic and normal benthic whitefish species evolve in parallel in different North American lakes or do they represent two glacial lineages that came into secondary contact in these lakes?

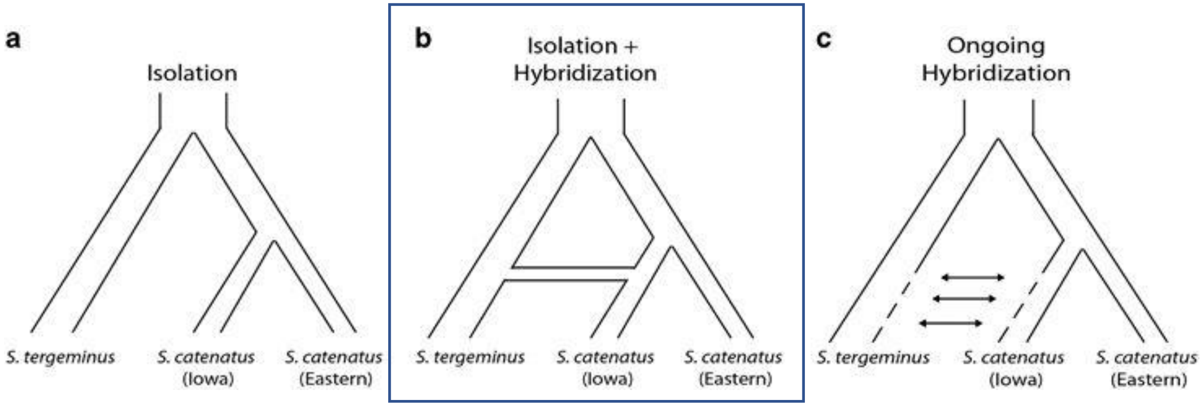




Rougeux et al., 2017, GBE

Example 3: Rattlesnakes and oak tree evolutionary history

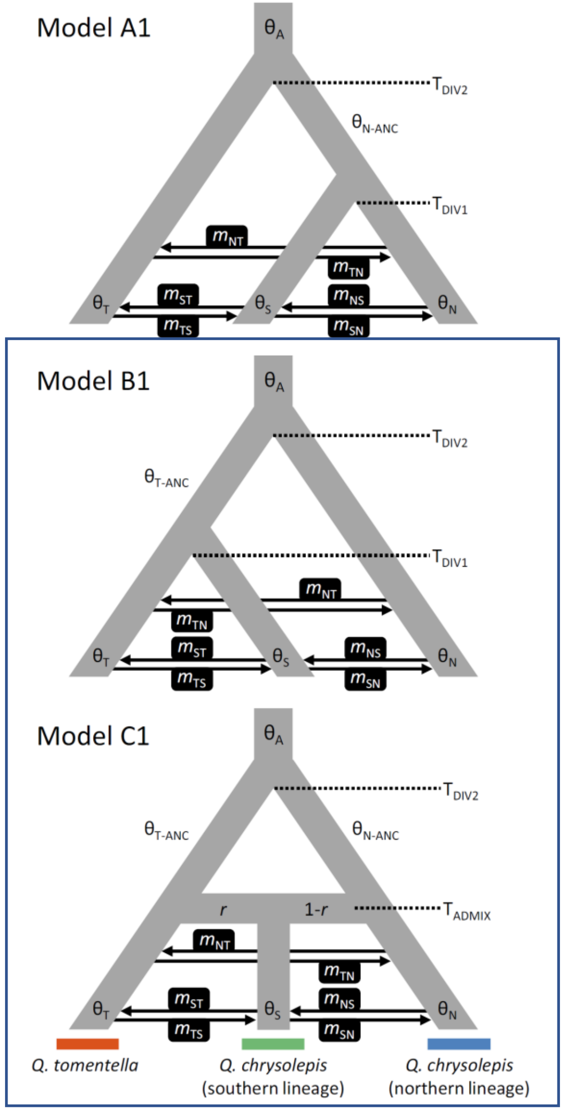
Best model



Sovic et al., 2016, Heredity

2 equally good models:

Ortego et al., 2017,
New Phytologist



“All models are wrong but some are useful”

George Box

Site frequency spectrum (SFS)

Efficient summary of the genome-wide data

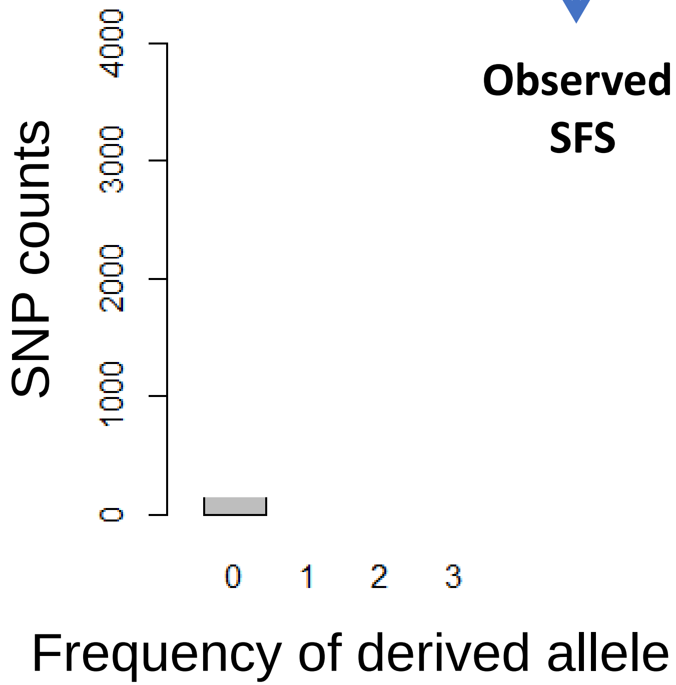
F_{ST}, Tajima’s D, pi, etc are summaries of the SFS

↓

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

Data

Observed SFS



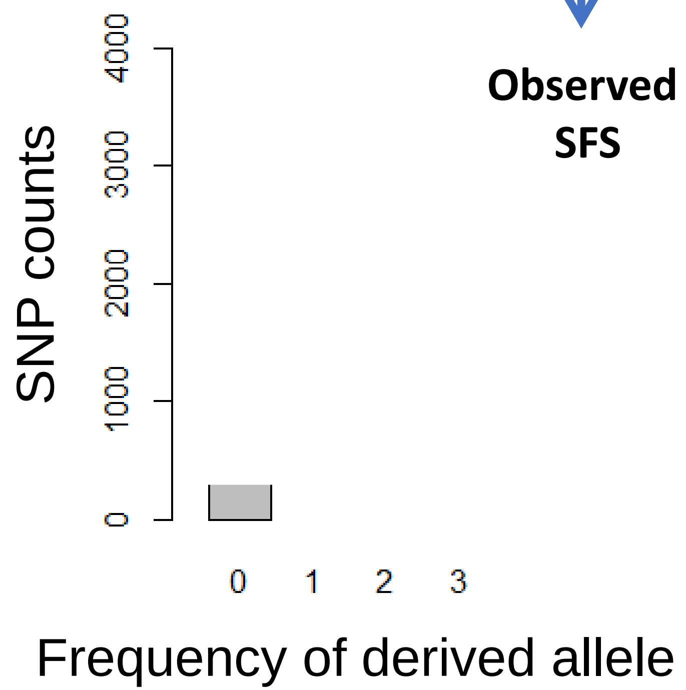
Site frequency spectrum (SFS)

Efficient summary of the genome-wide data

F_{ST}, Tajima's D, pi, etc are summaries of the SFS

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

Data

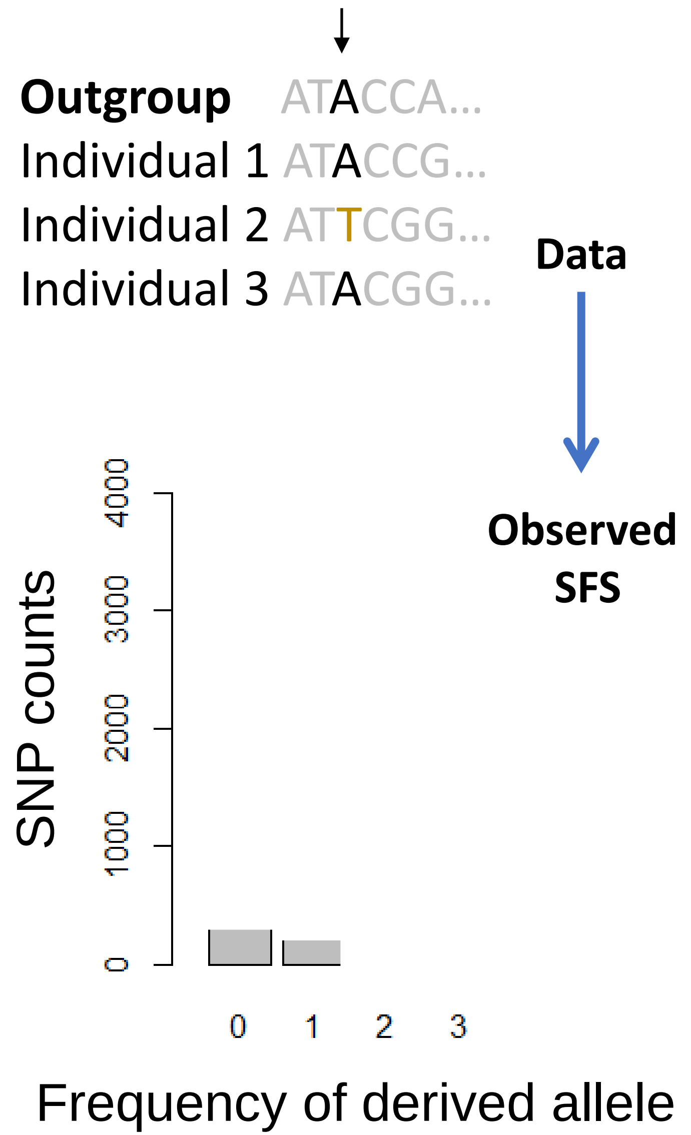


Observed SFS

Site frequency spectrum (SFS)

Efficient summary of the genome-wide data

F_{ST}, Tajima’s D, pi, etc are summaries of the SFS



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

Data

Observed SFS

Site frequency spectrum (SFS)

Efficient summary of the genome-wide data

F_{ST}, Tajima’s D, pi, etc are summaries of the SFS

Each diploid individual provides two haploid sequences

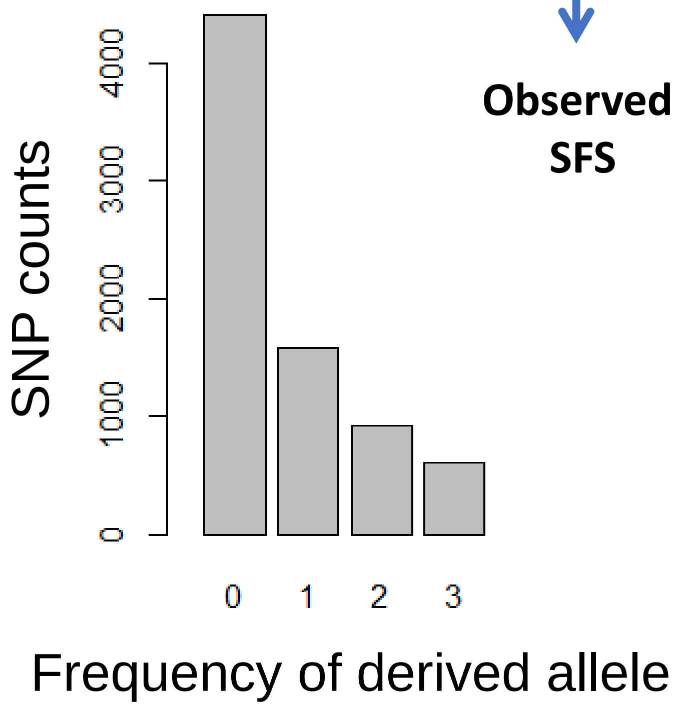
Linkage information is not used -> SNPs are assumed to be independent

As the ancestral state is known, we can infer the derived SFS -> of derived allele frequency (DAF)

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

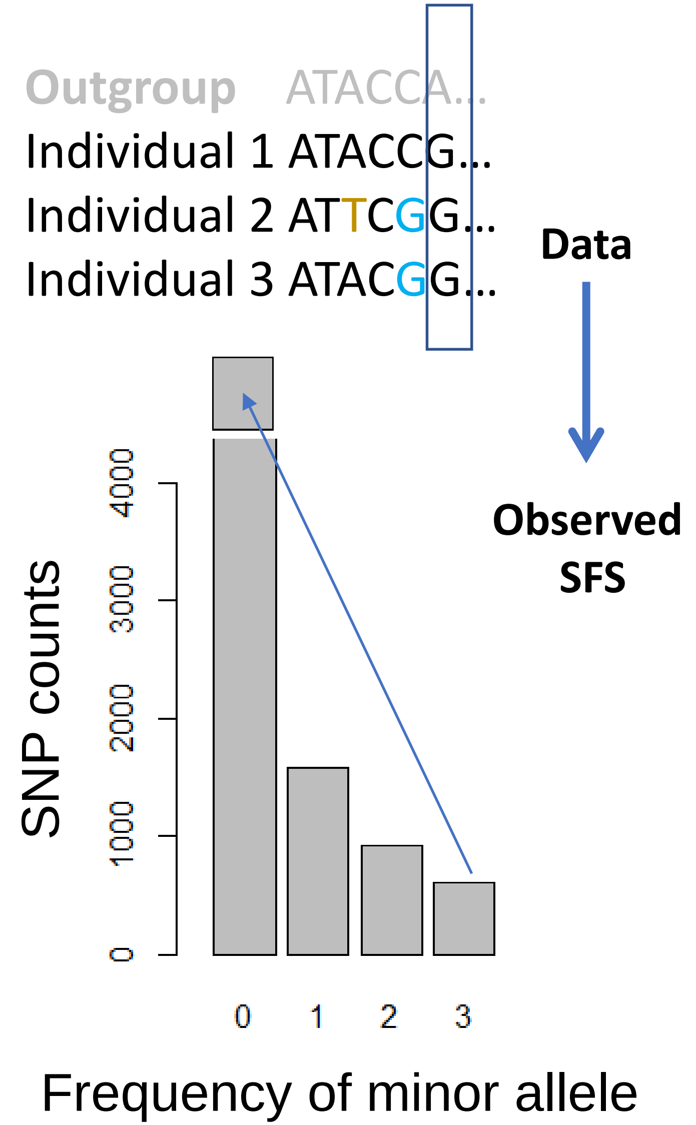
Data

Observed SFS



Site frequency spectrum (SFS)

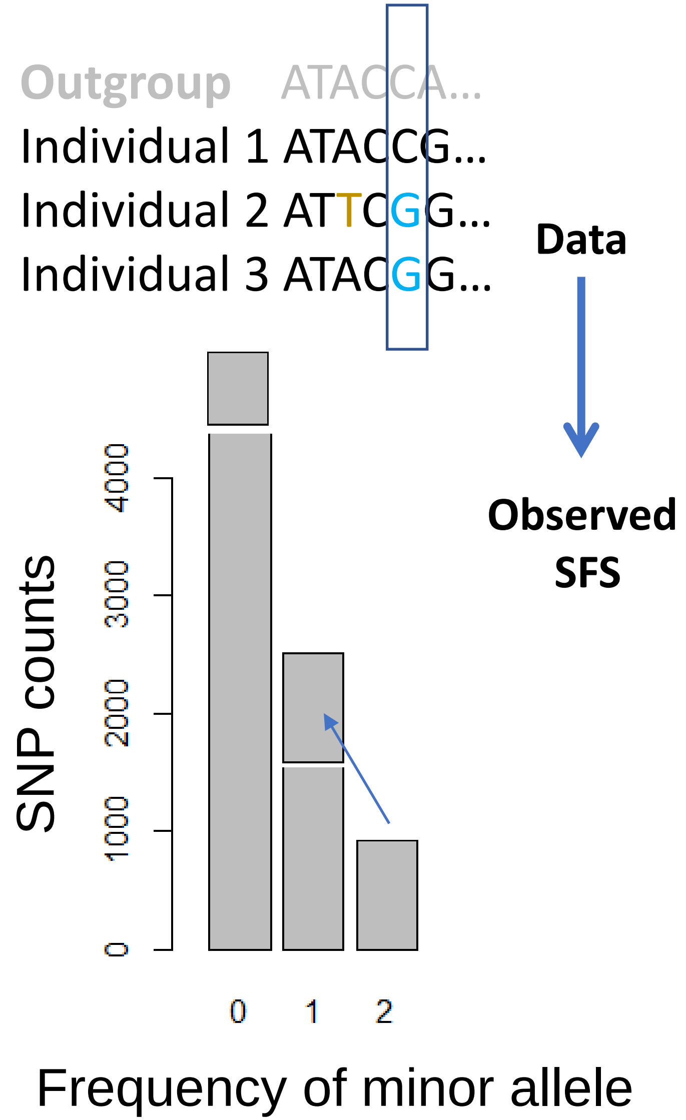
- If the ancestral state is not known, we infer a minor allele frequency spectrum (MAF) or folded SFS



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

Site frequency spectrum (SFS)

- If the ancestral state is not known, we infer a minor allele frequency spectrum (MAF) or folded SFS



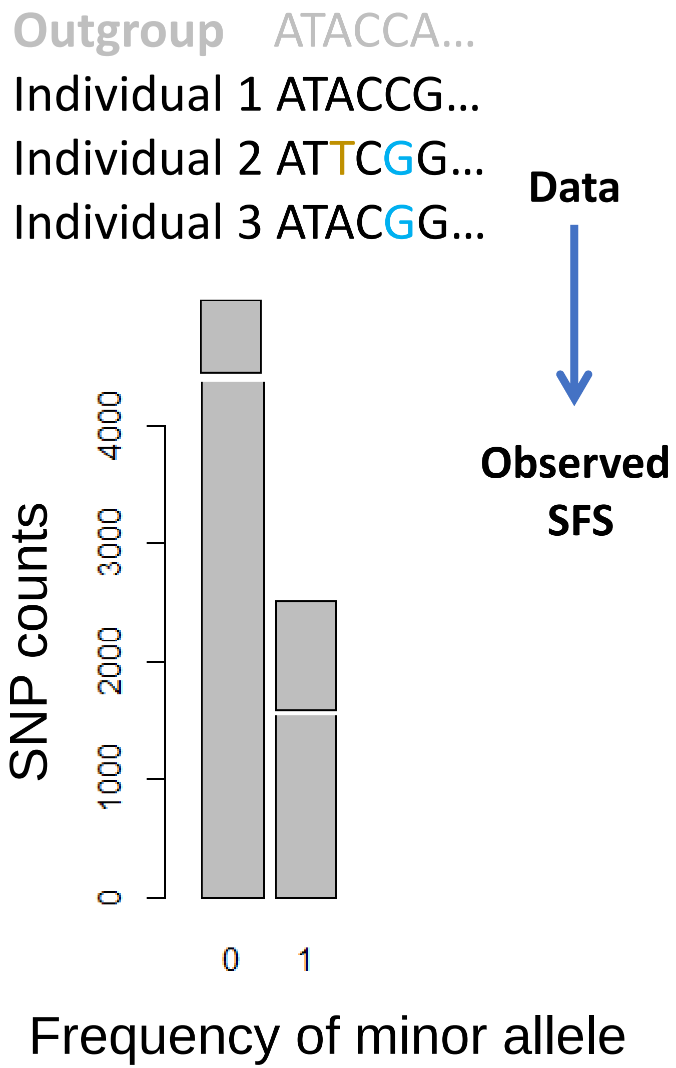
Outgroup ATACCA…
Individual 1 ATACCG…
Individual 2 ATTCGG…
Individual 3 ATACGG…

Data

Observed SFS

Site frequency spectrum (SFS)

- If the ancestral state is not known, we infer a minor allele frequency spectrum (MAF) or folded SFS

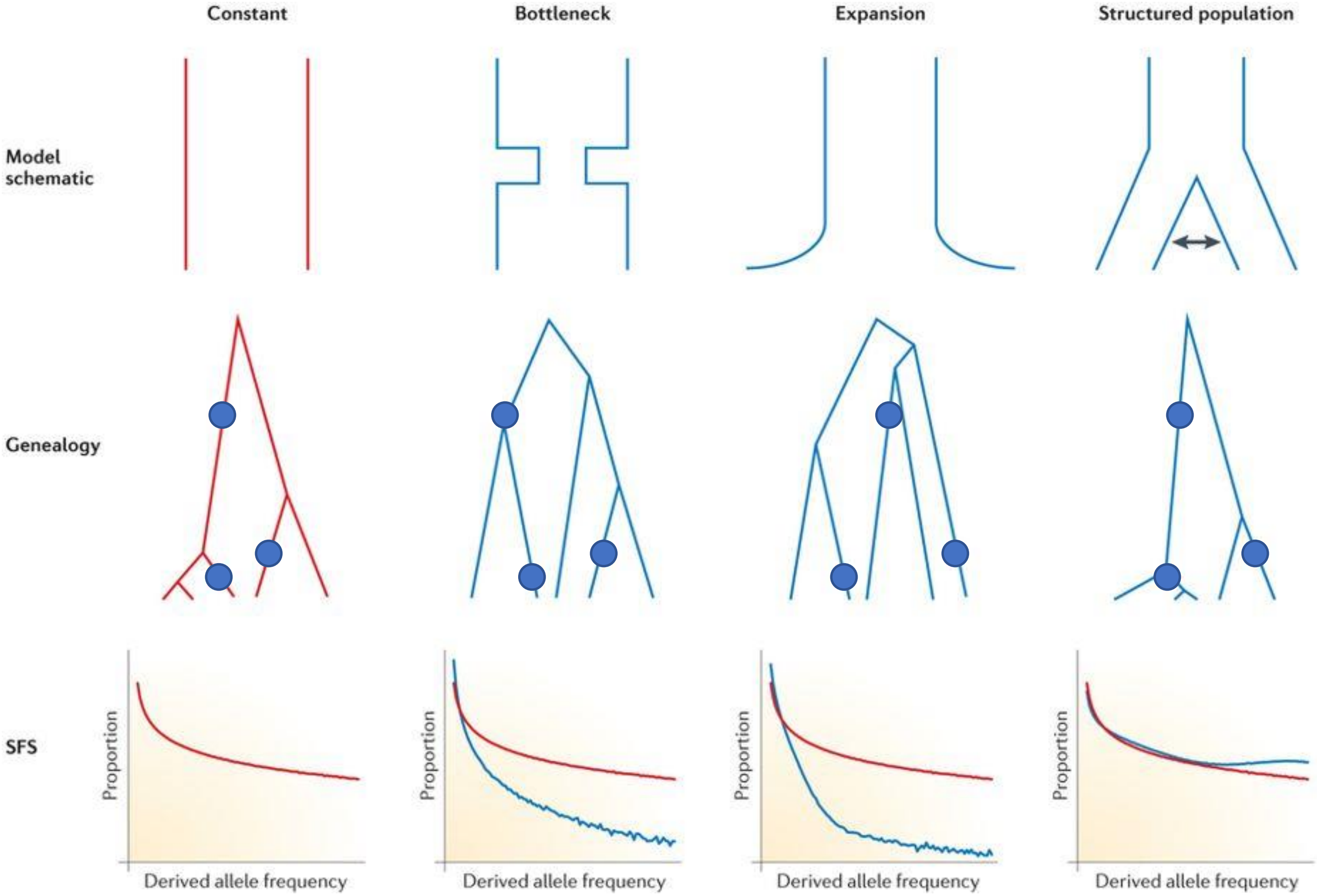


Outgroup ATACCA…
Individual 1 ATACCG…
Individual 2 ATTCGG…
Individual 3 ATACGG…

Data

Observed SFS

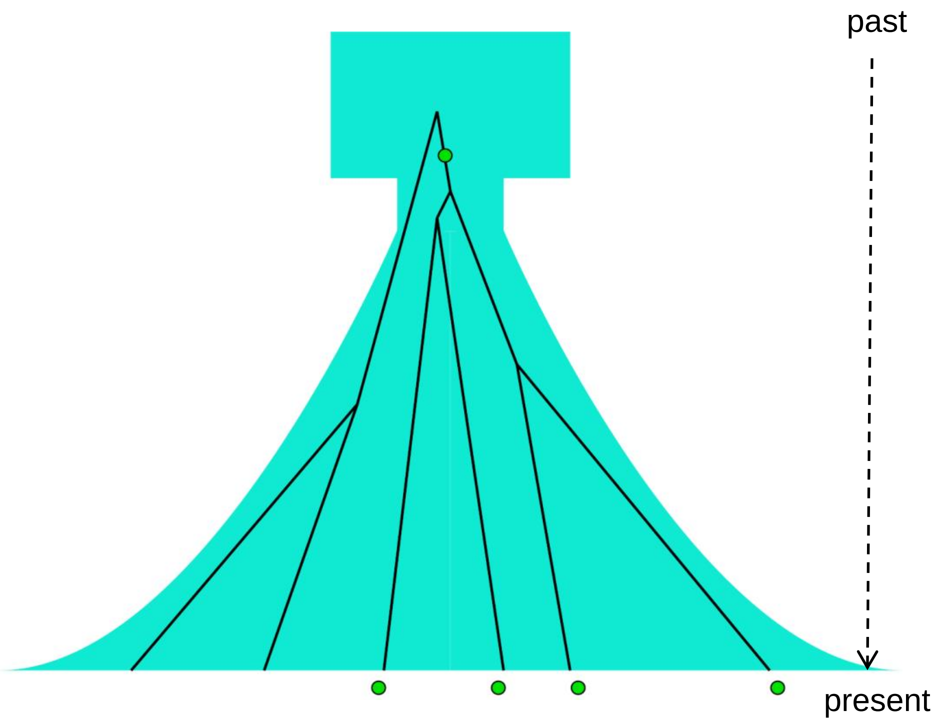
Different demographic scenarios lead to different SFS

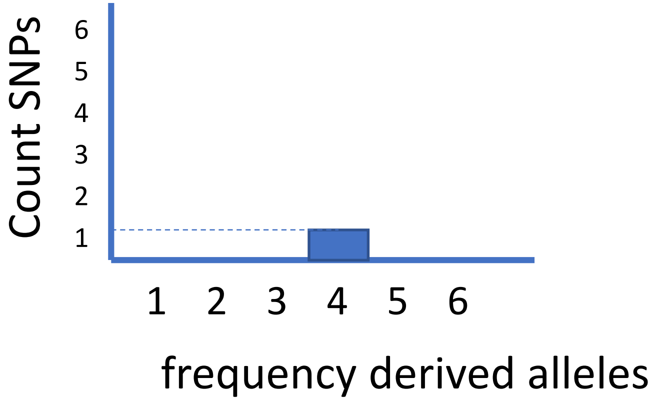


Schraiber & Akey (2015) Nat Rev Genet

Coalescent and the SFS

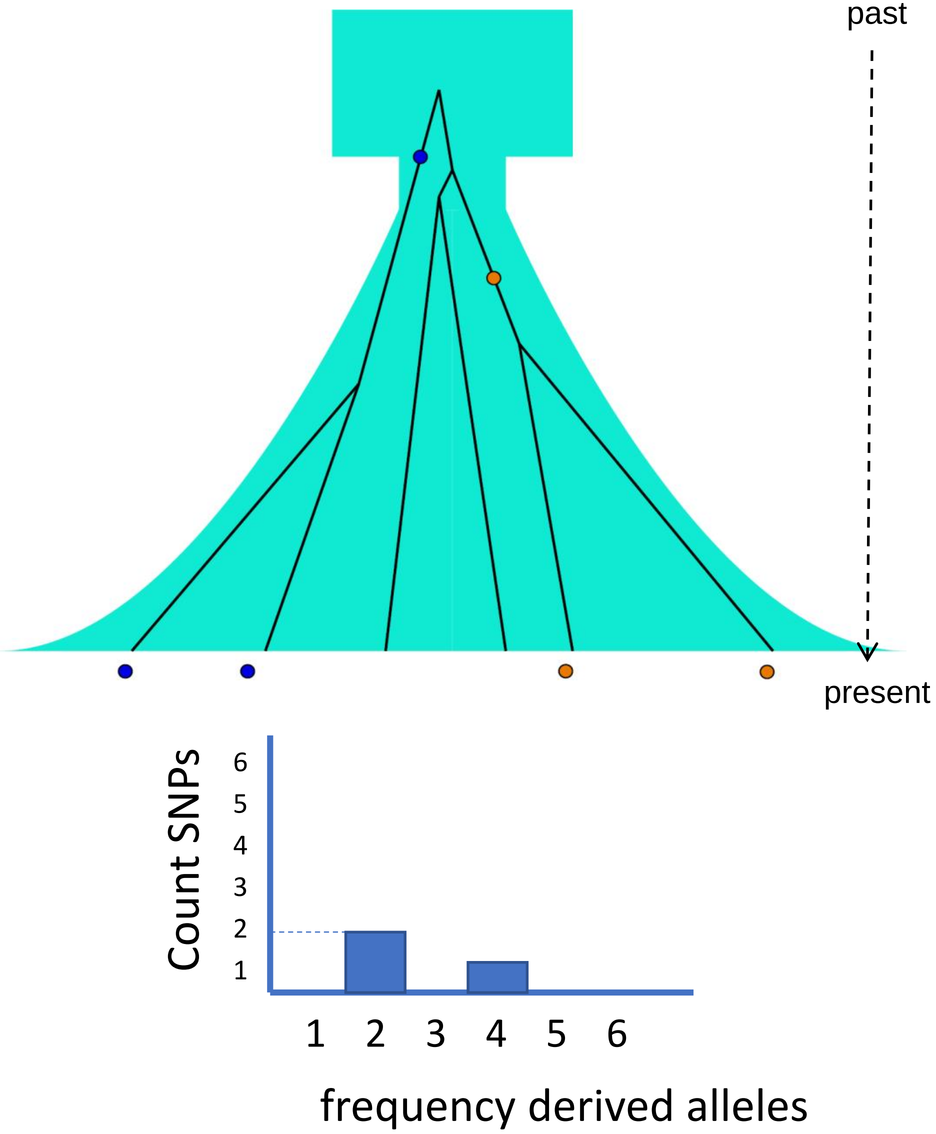
- A recent population growth following a bottleneck leads to gene trees with long external branches
- Very few mutations in the internal branches
- Most mutations in long external branches are only found in one lineage, resulting in an excess of singletons





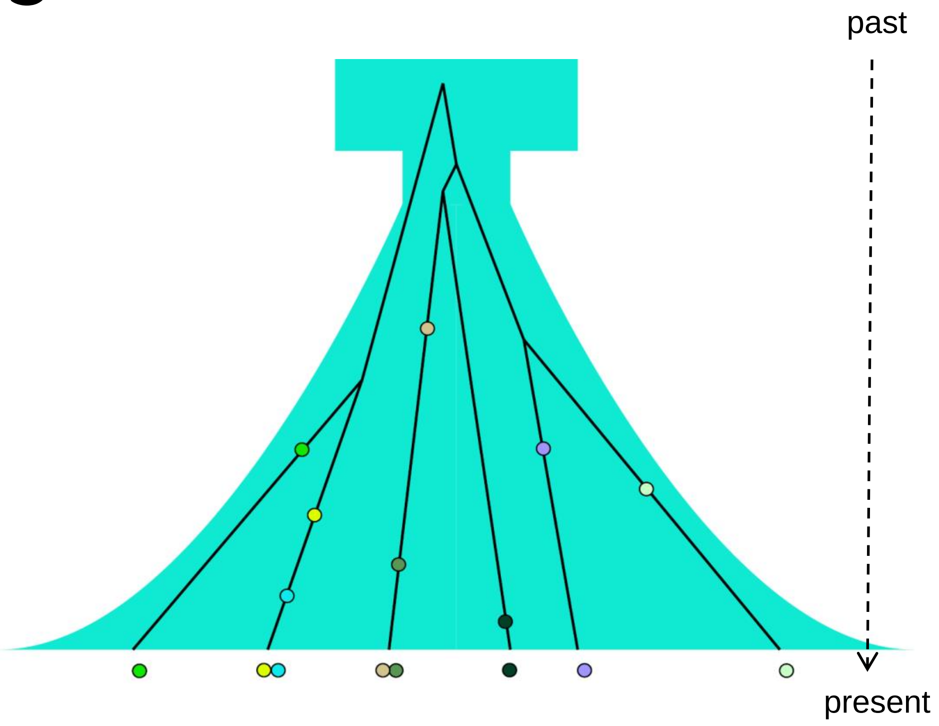
Coalescent and the SFS

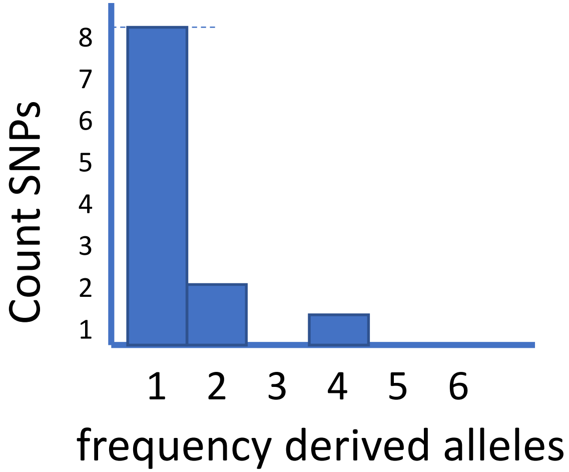
- A recent population growth following a bottleneck leads to gene trees with long external branches
- Very few mutations in the internal branches
- Most mutations in long external branches are only found in one lineage, resulting in an excess of singletons



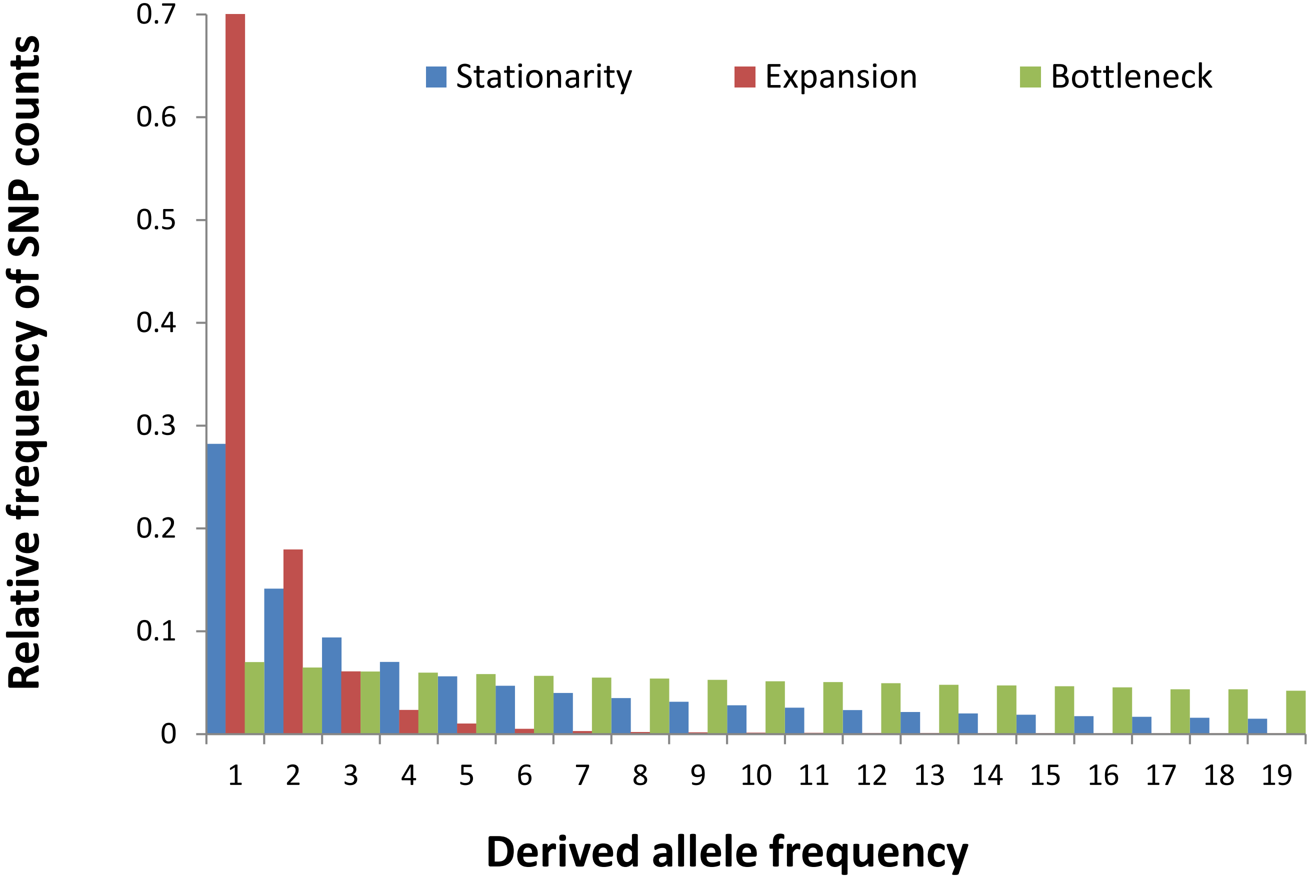
Coalescent and the SFS

- 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



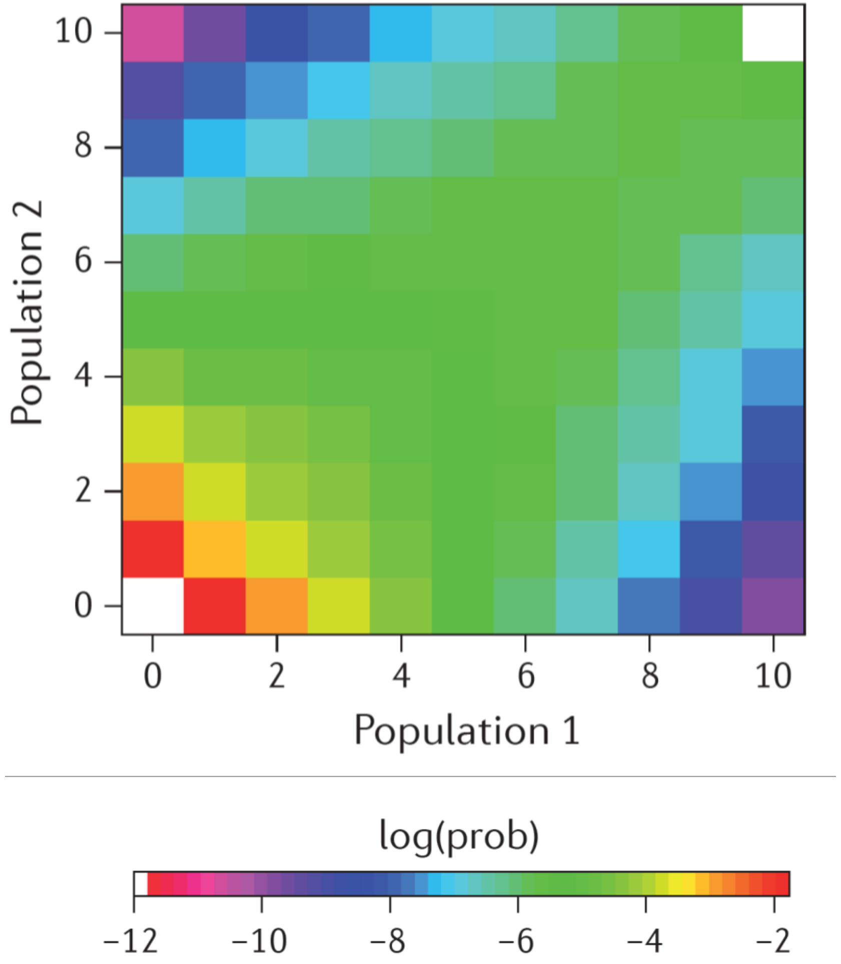


Effects of effect population size changes on the SFS



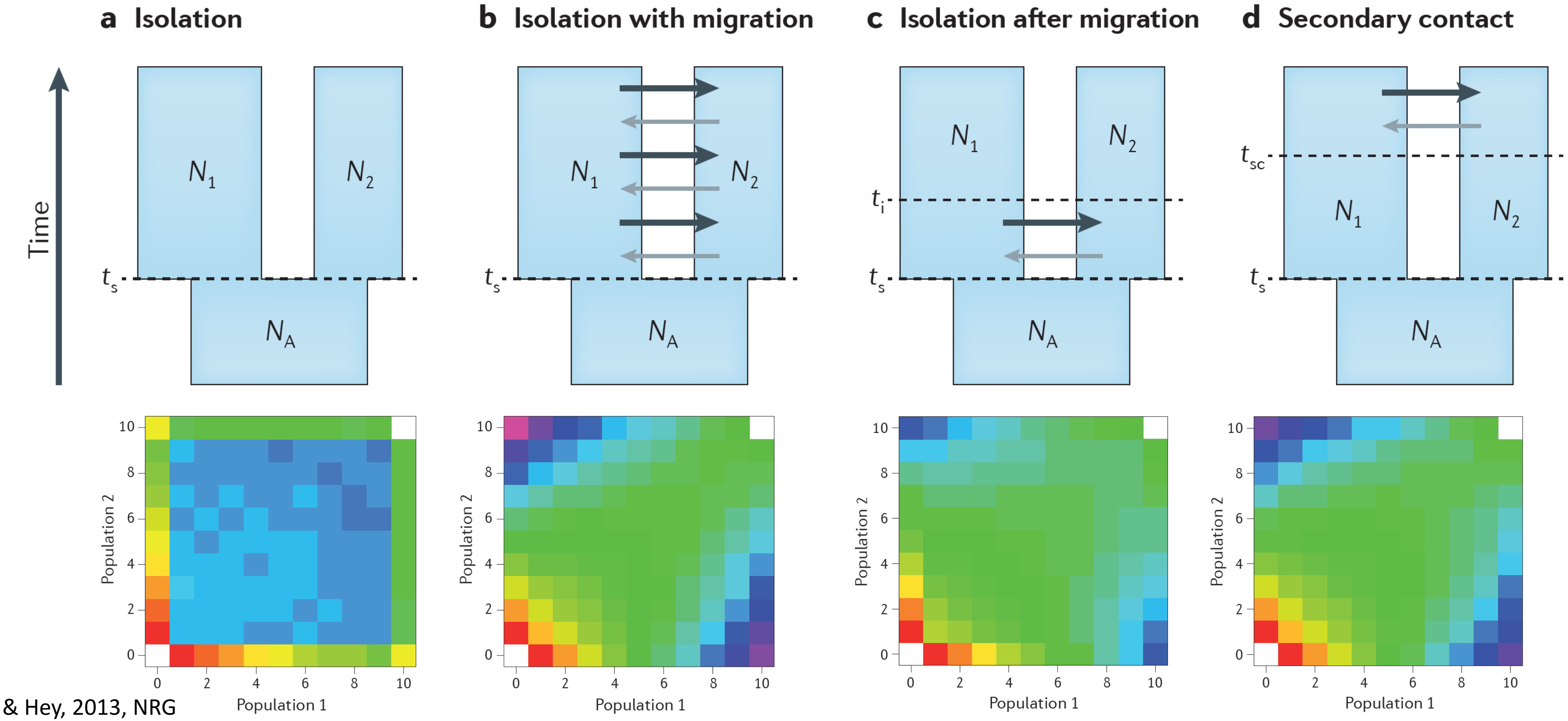
SFS for more than one population

- For 2 populations: 2D SFS containing counts of SNPs with a frequency of the derived or minor allele i in population 1 and j in population 2
- With more populations, the SFS becomes multidimensional or pairwise 2D SFS can be used



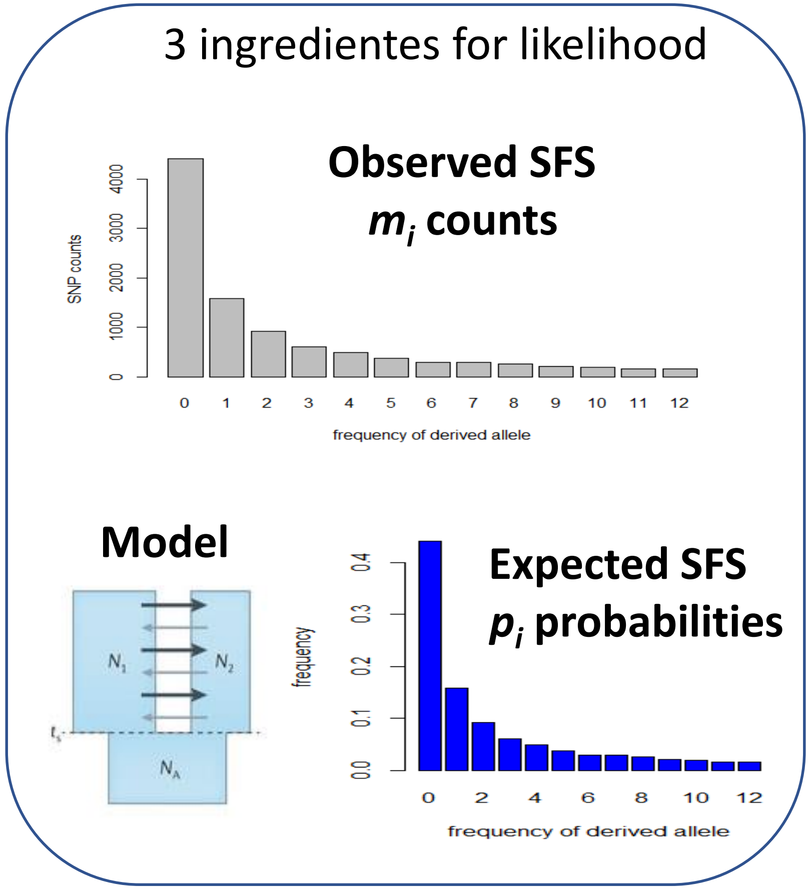
Sousa & Hey, 2013, NRG

Expected SFS under different evolutionary szenarios



Sousa & Hey, 2013, NRG

Composite likelihood



3 ingredientes for likelihood

Observed SFS
***m**$_i$* **counts**

Model

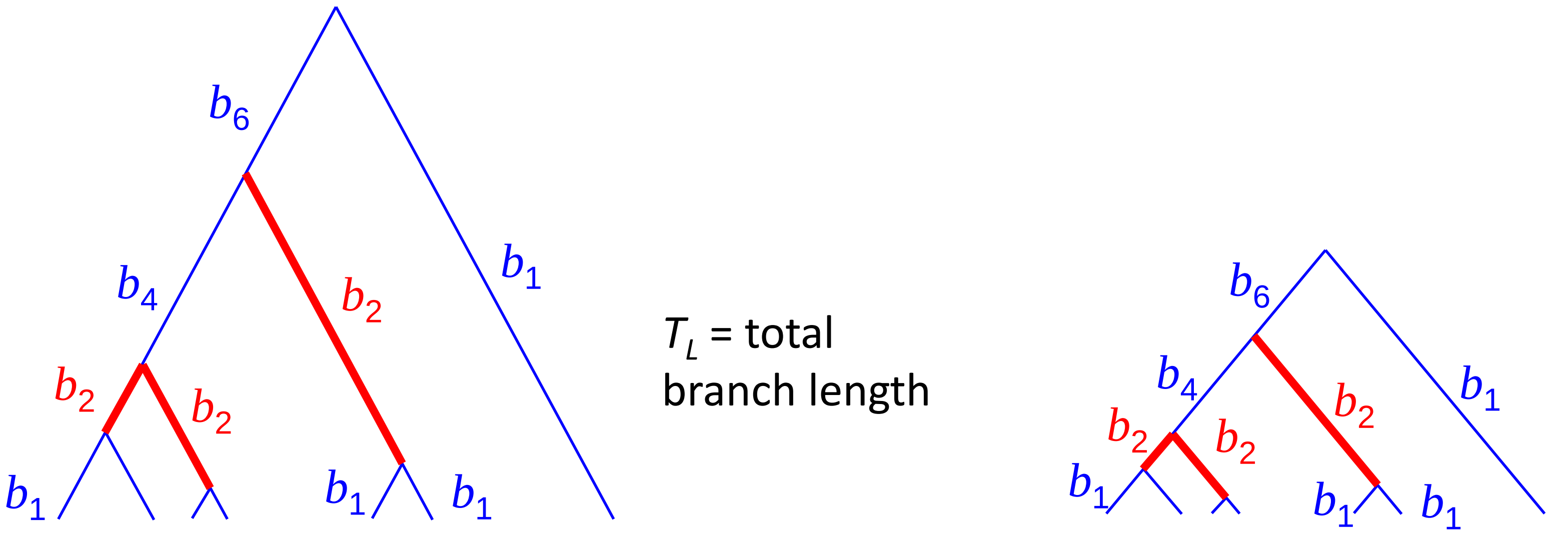
Expected SFS
***p**$_i$* **probabilities**

Given *S* polymorphic sites (SNPs) out of *L* sites (Adams and Hudson, 2004) the composite likelihood is:

$$CL = \Pr(X \mid \theta) \propto \underbrace{P_0^{L-S}}_{\text{probability of no mutation on the tree}} \underbrace{(1-P_0)^{S}}_{\text{probability of at least one mutation in the tree}} \prod_{i=1}^{n-1} \hat{p}_i^{\,m_i}$$

Excoffier et al. (2013) *PloS Genetics*

The exact same SFS can be obtained with a long or short tree



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

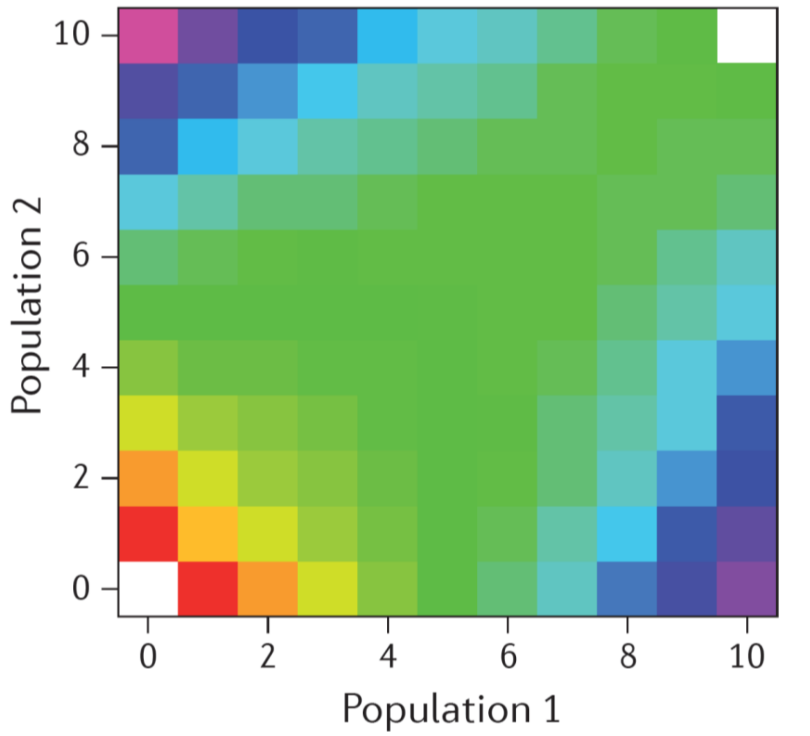
- We need a mutation rate and the number of monomorphic sites to distinguish among the two!
- Or we need to fix some parameters, e.g. the splitting time

fastsimcoal

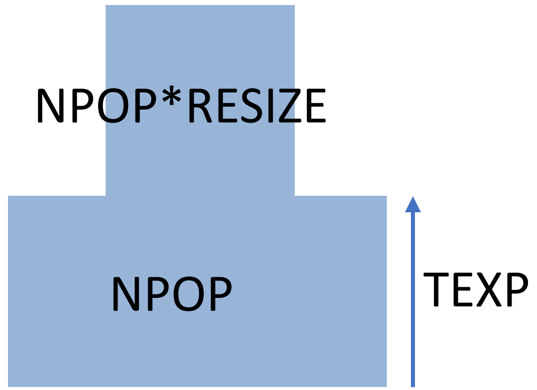
- Fastsimcoal2 can estimate parameters from the SFS using coalescent simulations
- Maximum (composite) likelihood method
- Uses a conditional expectation (CEM) maximization algorithm to find parameter combinations that maximize the likelihood
- It approximate the expected SFS by performing coalescent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

Input files for fastsimcoal

Observed SFS



Model template file



Parameter file

```
NPOP   logunif 1000 100000
TEXP    logunif  500 50000
RESIZE logunif   0.1  100
```

Input files for fastsimcoal2: observed SFS

- 1D, 2D or multidimensional/joint SFS

example_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
```

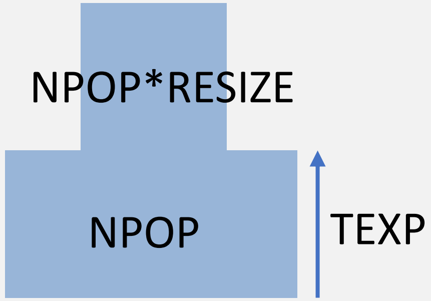
example_jointDAFpop1_0.obs

```
1 observations
      d0_0  d0_1    d0_2    d0_3    d0_4    d0_5
d1_0  1998557   8211    1415    316 55  10
d1_1  1266  101 37  16  5   1
d1_2  611   42  20  8   2   0
d1_3  486   31  12  5   0   0
d1_4  479   15  9   2   3   1
d1_5  1189  46  22  19  18  0
```

Input files for fastsimcoal2: Model template file

example.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
```

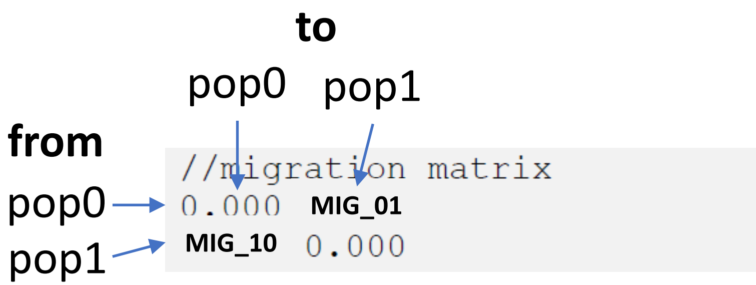


Input files for fastsimcoal2: Estimation file

example.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
```

Input files for fastsimcoal2: Model template file Migration matrices



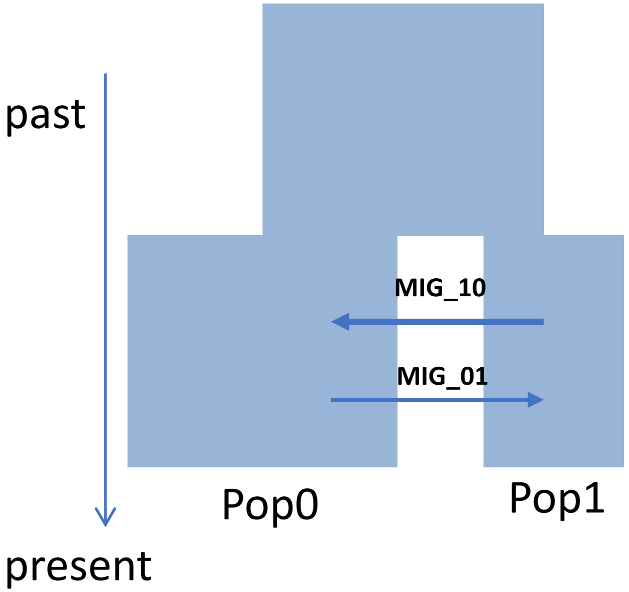
```
//migration matrix
0.000  MIG_01
MIG_10  0.000
```

example2.tpl

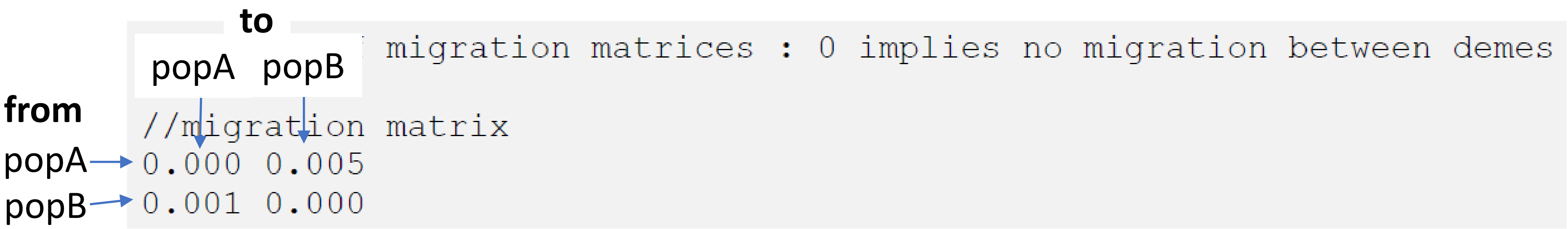
```
//Number of populations (demes or species)
2
//Population effective sizes (number of genes)
NPOP0
NPOP1
//Samples sizes and samples age
10
10
```

Migration is from index in row to index in column **backwards** in time.

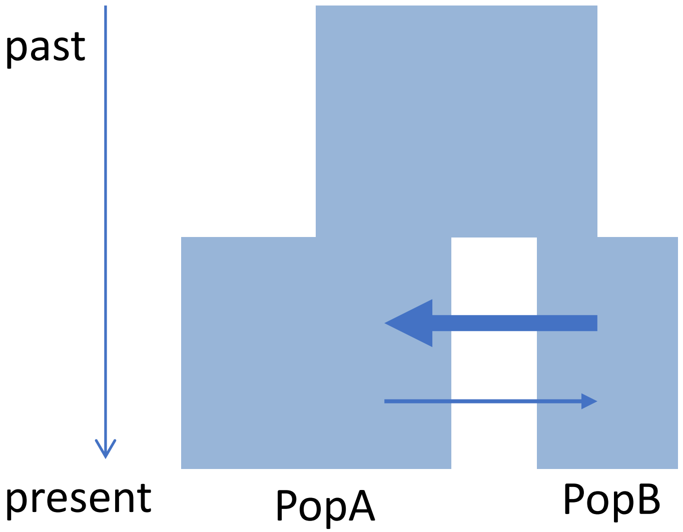
The entry m_{ij} lists the **migration rates backward in time** from population *i* to population *j*. The above-mentioned matrix states that, for each generation backward in time, any gene from population 0 has probability MIG_01 to be sent to population 1, and that a gene from population 1 has a probability MIG_10 to move to population 0.

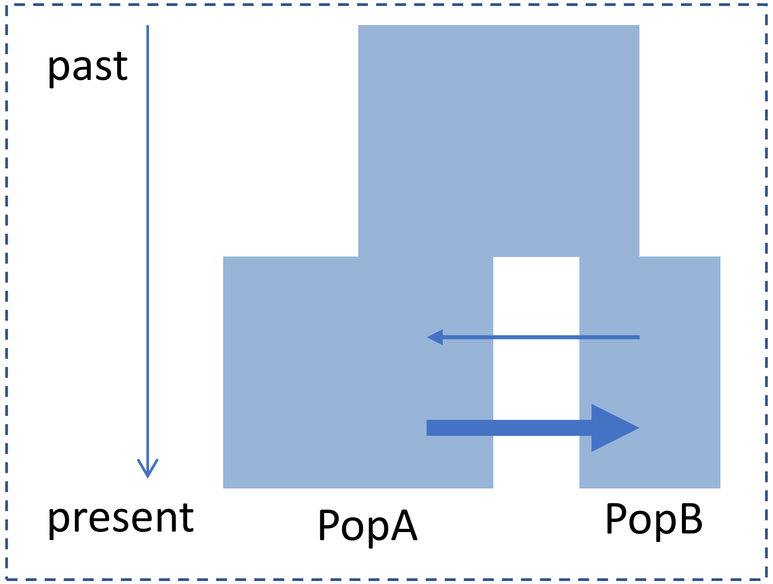


Question: To what model does this migration matrix correspond to?



```
  migration matrices : 0 implies no migration between demes

//migration matrix
0.000 0.005
0.001 0.000
```

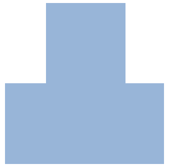




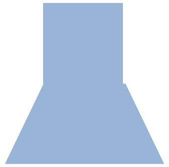
Historical events in fastsimcoal2

//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix index

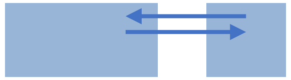
- Change the size of a given population



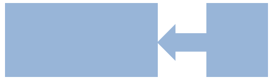
- Change the growth rate of a given population



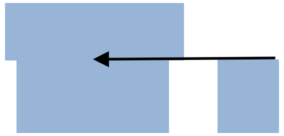
- Change the migration matrix



- Introgression event: Move a fraction of the genes of a given population to another population.



- Fusion of two populations: Move all genes from a population to another population. This would be a population split forward in time.

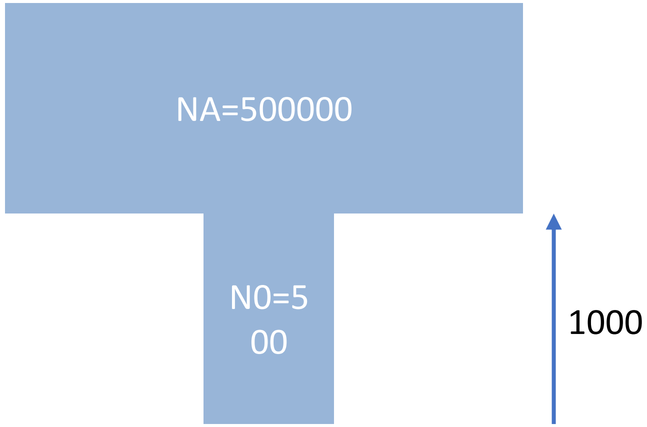


- One or more of these events can occur at the same time
- In the end, all populations must have fused to a single population

Example: Change of population size

```
//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
```

Recent instantaneous demographic contraction



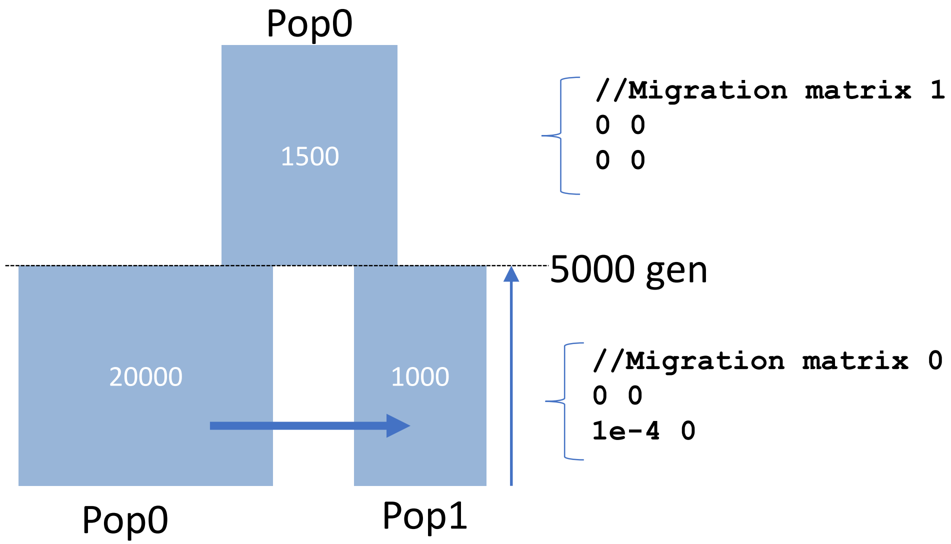
- 1000 generations ago, 0% (migrants=0) of lineages in pop0 (source) migrated to pop0 (sink). This means that 100% of lineages remained in pop0.

- The sink population (pop0) has a size 1000 larger after the event (new size=1000). Given that N0=500 diploids at time zero, it implies that NA=500000 diploids.

- The migration matrix valid after the event is the migration rate 0.

Example: Population split (merge backwards in time)

```
//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
1 historical event
5000 1 0 1 0.075 0 1
```



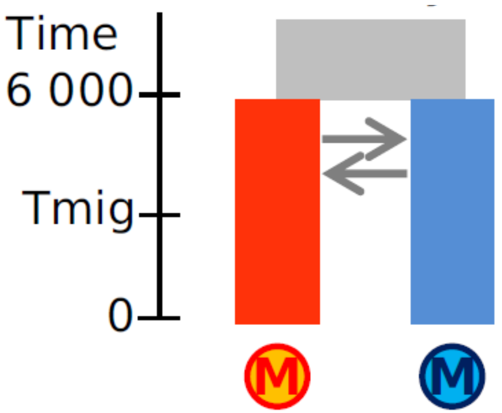
```
//Migration matrix 1
0 0
0 0
```

```
//Migration matrix 0
0 0
1e-4 0
```

- 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.

Now, let’s write our own model

Model with early gene flow (isolation after migration)



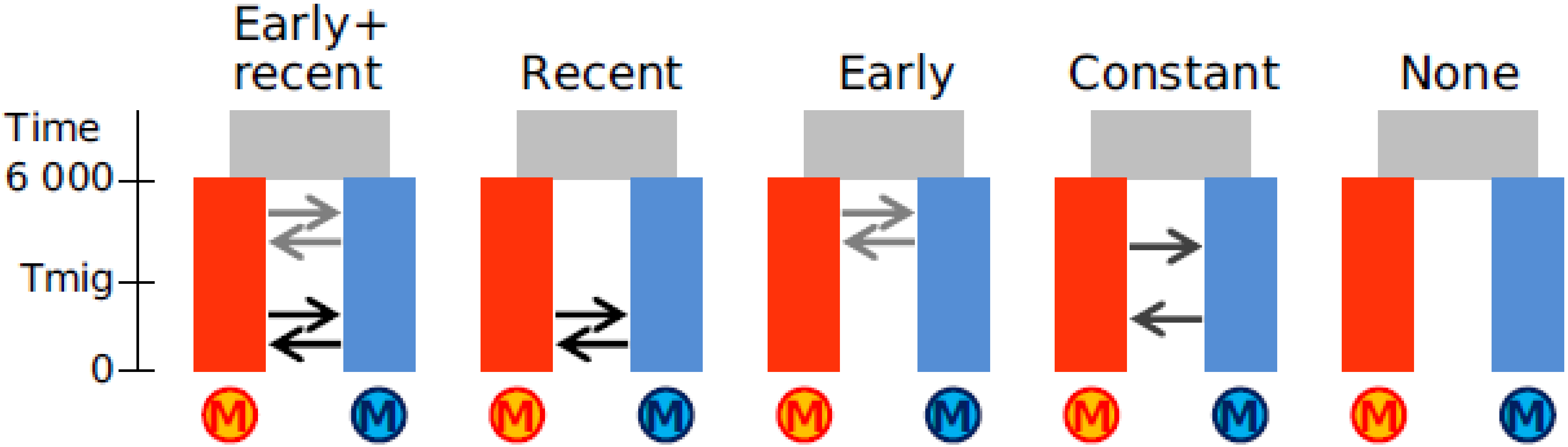
First, we test if a model of speciation with divergence with gene flow and then complete reproductive isolation fits the data well.

We need to produce three input files:

- Observed pairwise SFS: early_geneflow_jointMAFpop1_0.obs
- Model specification: early_geneflow.tpl
- Estimated parameters: early_geneflow.est

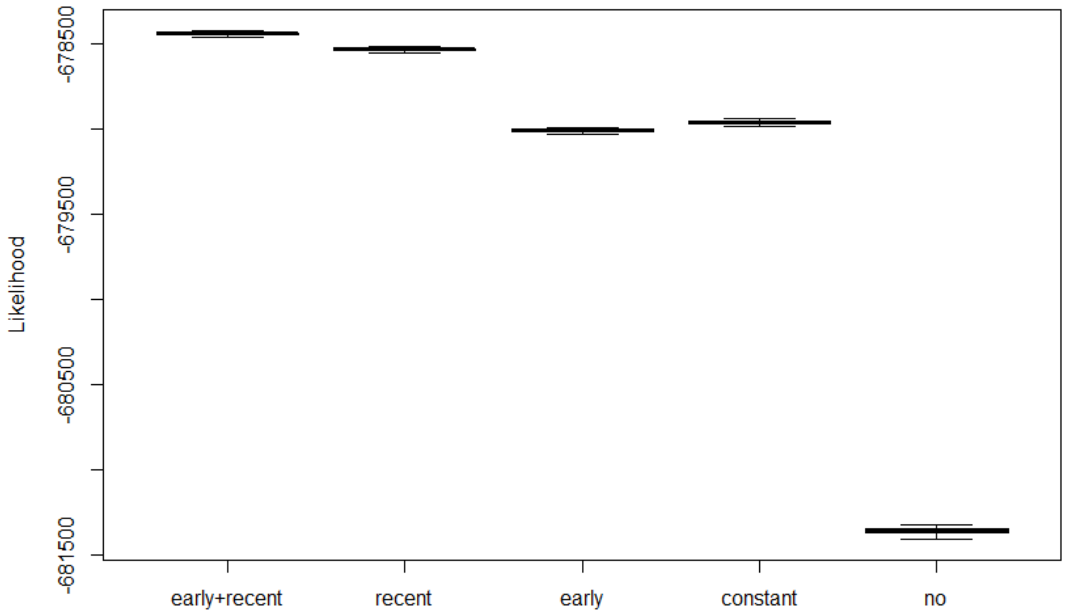
We can modify the example.tpl and example.est files to represent our model. As we do not have a reliable mutation rate, we will fix the divergence time to 6,000 generations.

All models

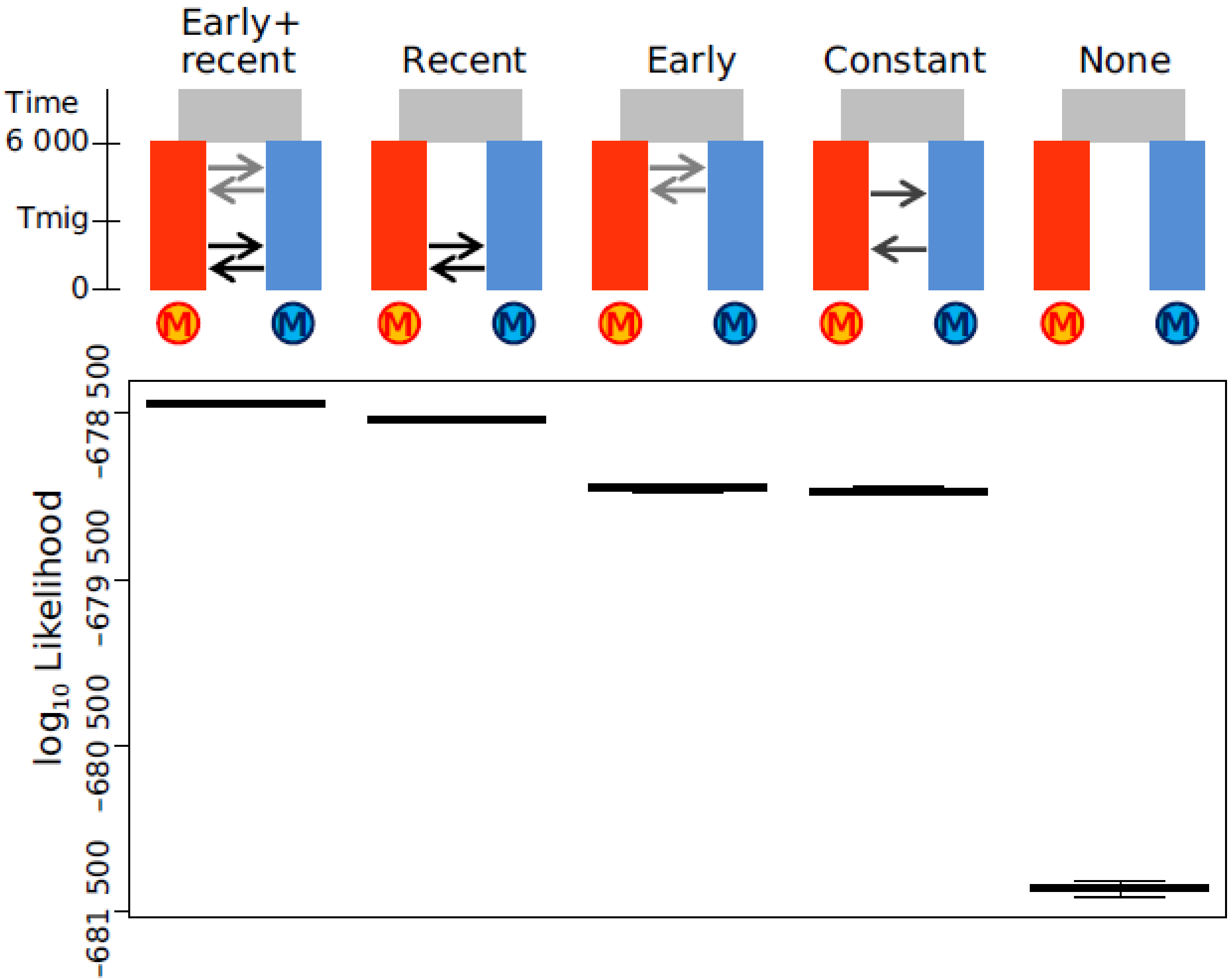


Comparison to published results

Our Results



Meier et al, 2017, MolEcol


Early+ recent
Recent
Early
Constant
None
Time
Tmig
M
M
M
M
M
M
M
M
M
M
log10 Likelihood