Demographic modeling with fastsimcoal2

Joana Meier

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

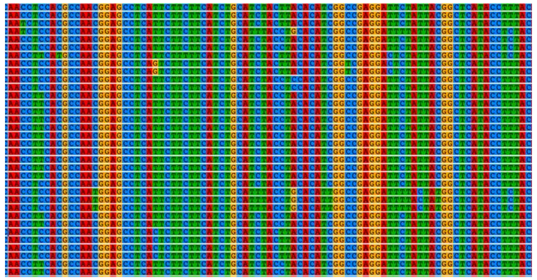
Aims and principle of demographic modeling

Test which of different evolutionary scenarios fits the data best

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

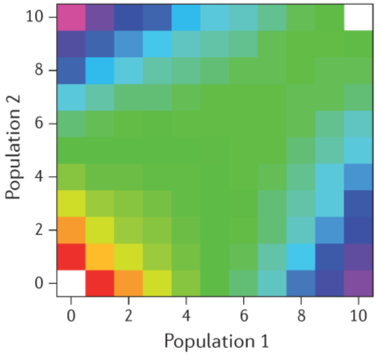
Compare

Genomic data

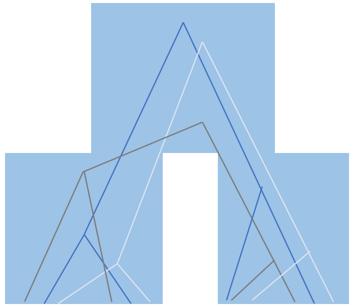


summarize

Observed SFS

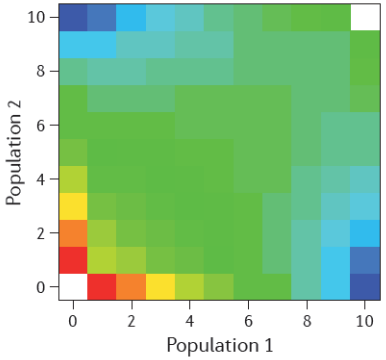


Model



simulate

Expected SFS



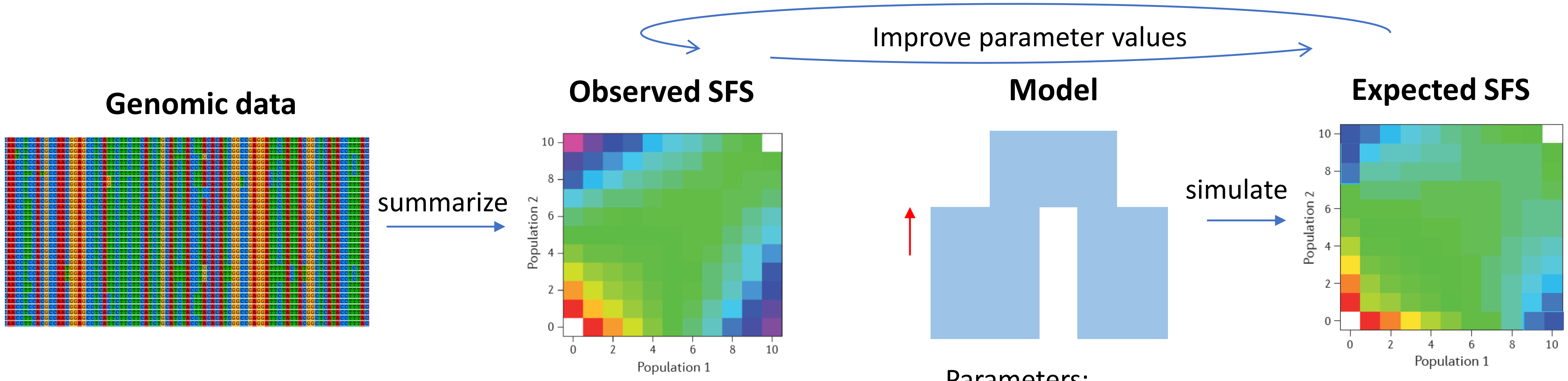
Parameters:

- Time of split
- Migration rates
- Effective sizes

Aims and principle of demographic modeling

Test which of different evolutionary scenarios fits the data best

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



Compare

Improve parameter values

Genomic data

summarize

Observed SFS

Model

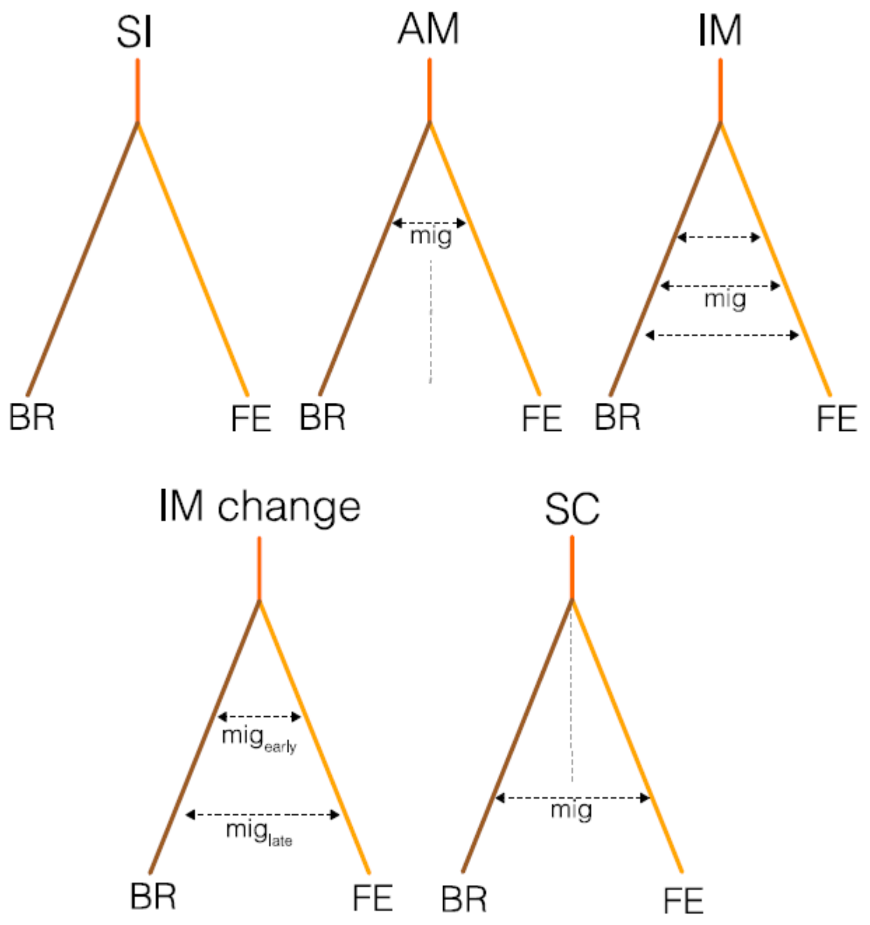
simulate

Expected SFS

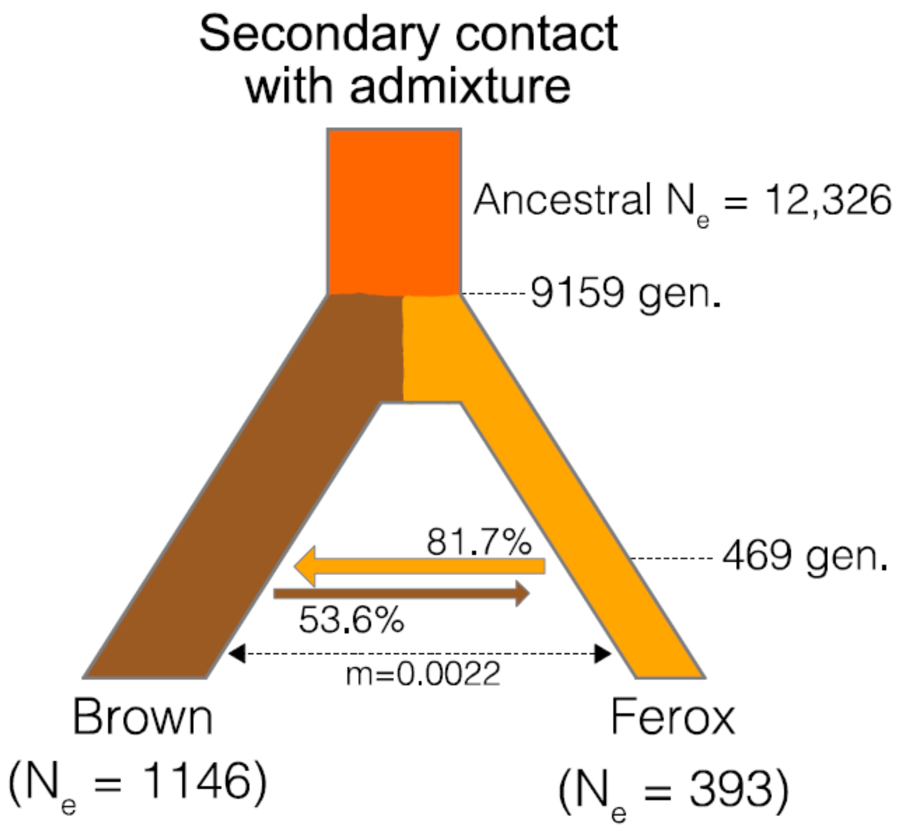
Parameters:

- Time of split
- Migration rates
- Effective sizes

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



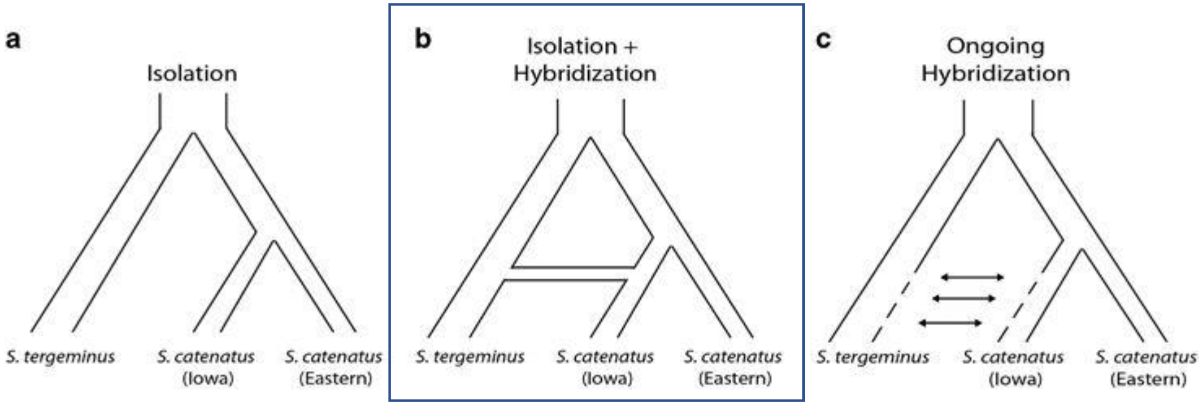
Best model:



Jacobs et al., Genes, 2018

Rattlesnakes and oak tree evolutionary history

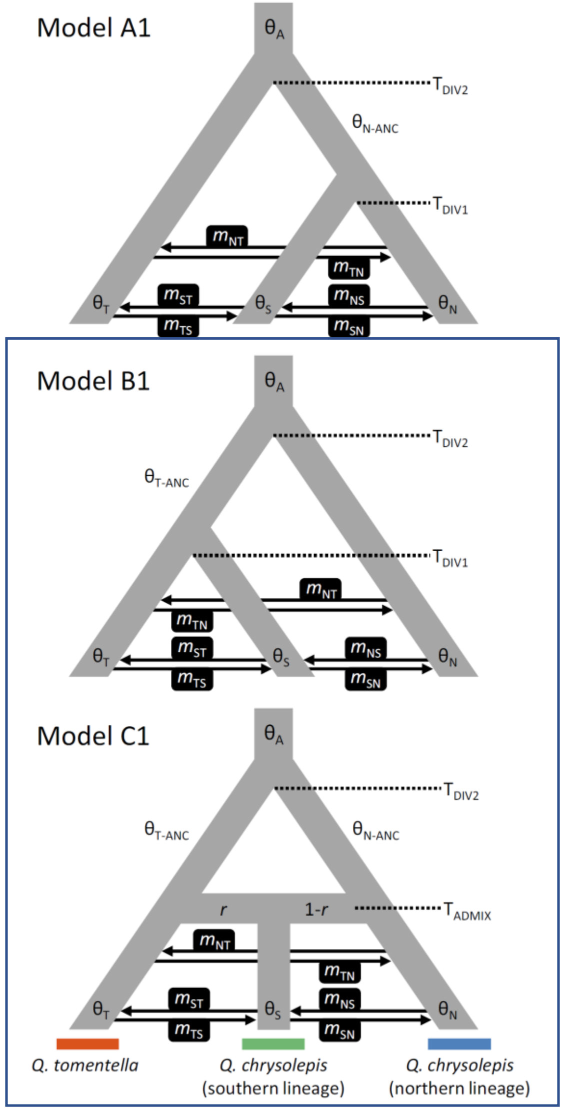
Best model



Sovic et al., 2016, Heredity

2 equally good models:

Ortego et al., 2017,
New Phytologist



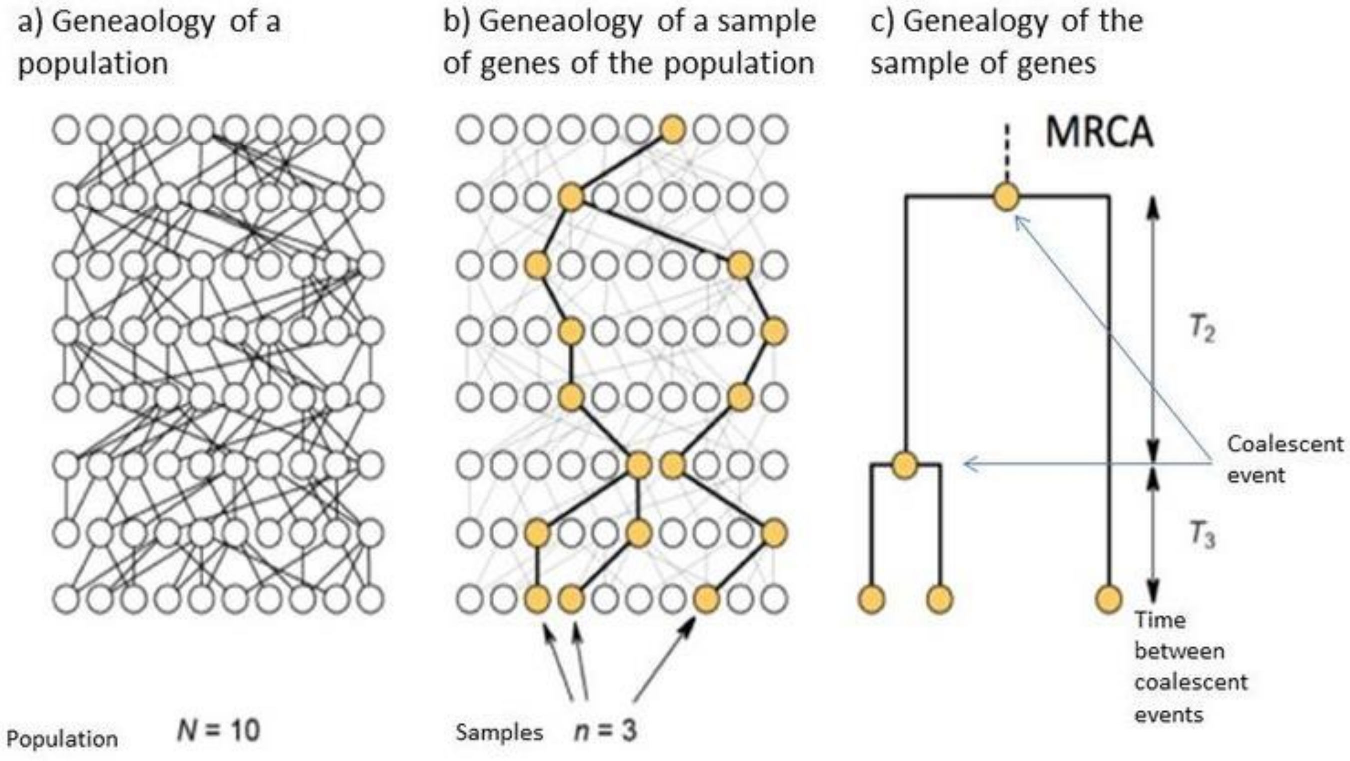
“All models are wrong but some are useful”

George Box

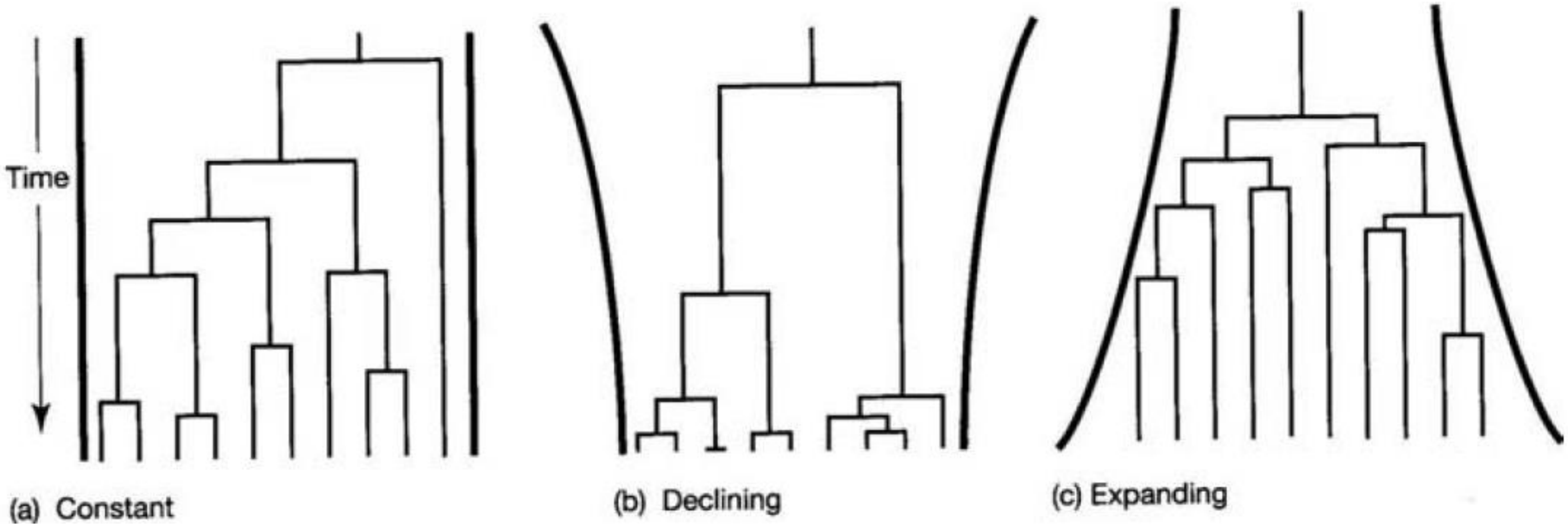
How can we infer the demographic history using sequencing data?

Coalescent theory

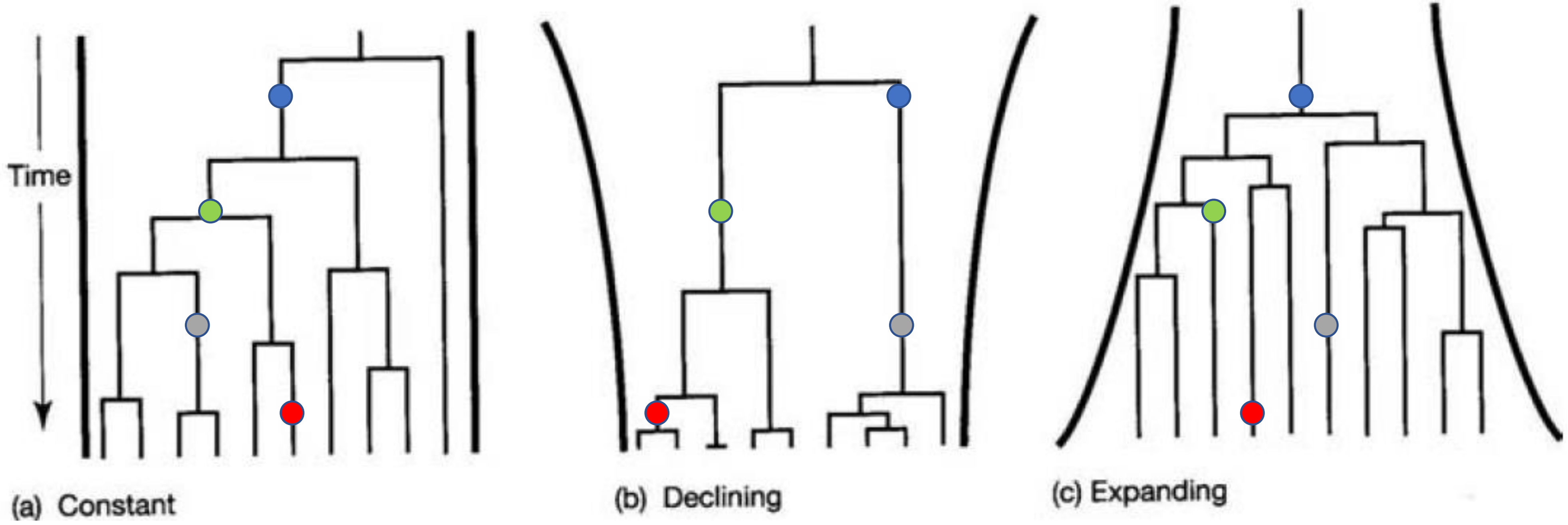
- The coalescent is a model of the ancestral relationships of individuals from an idealized population
- Wright-Fisher population: consists of haploid individuals with non-overlapping generations and random mating. Allele frequencies change randomly due to drift (no selection).


a) Geneaology of a population
b) Geneaology of a sample of genes of the population
c) Genealogy of the sample of genes
MRCA
T_2
Coalescent event
T_3
Time between coalescent events
Population $N = 10$
Samples $n = 3$

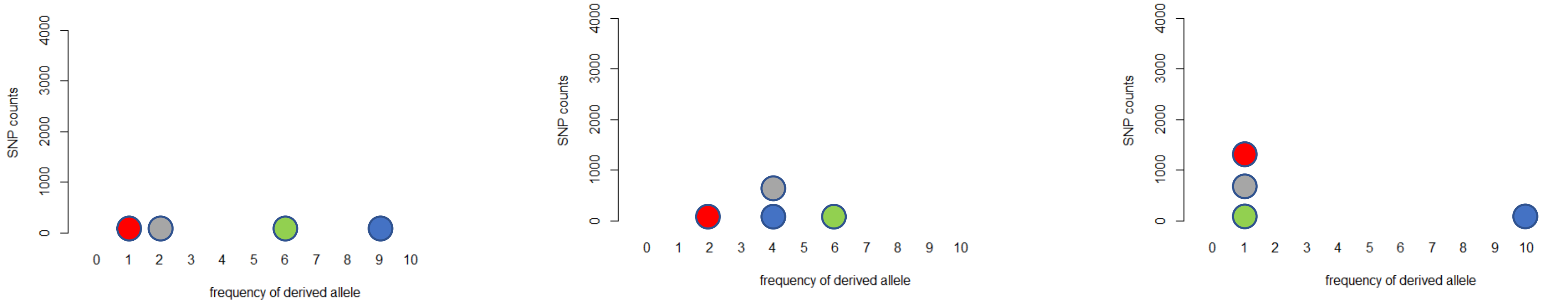
Shape of the genealogy is informative on the population history



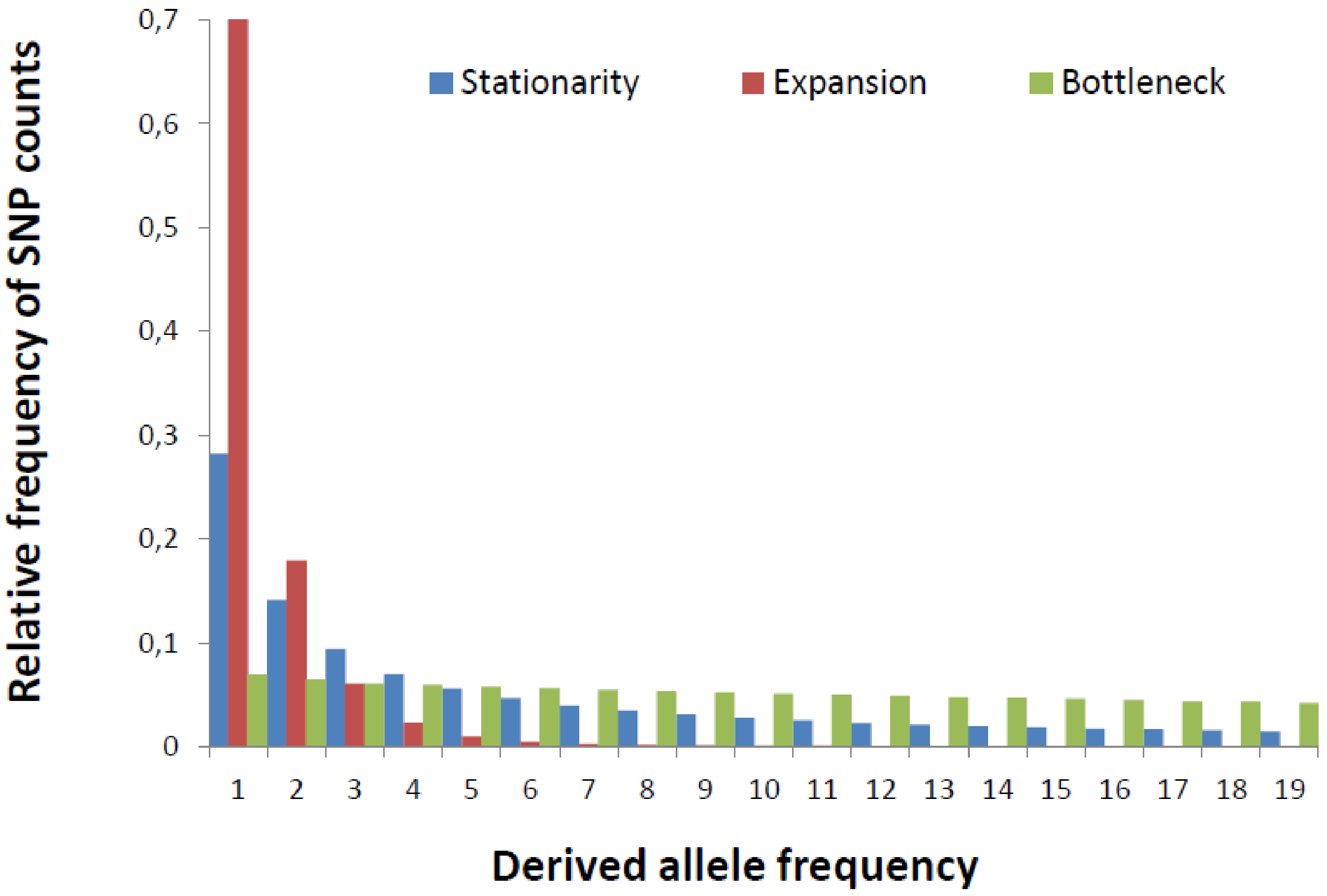
Shape of the genealogy is informative on the population history



Site frequency spectrum (SFS)



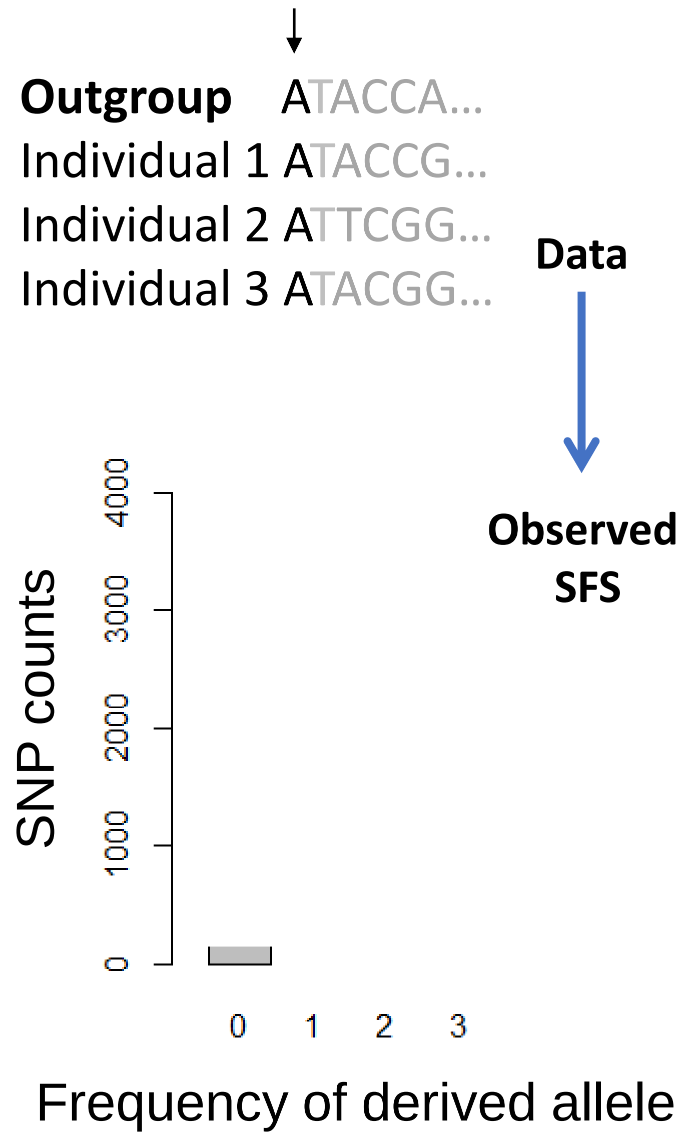
Expected SFS shapes under different demographic histories



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…

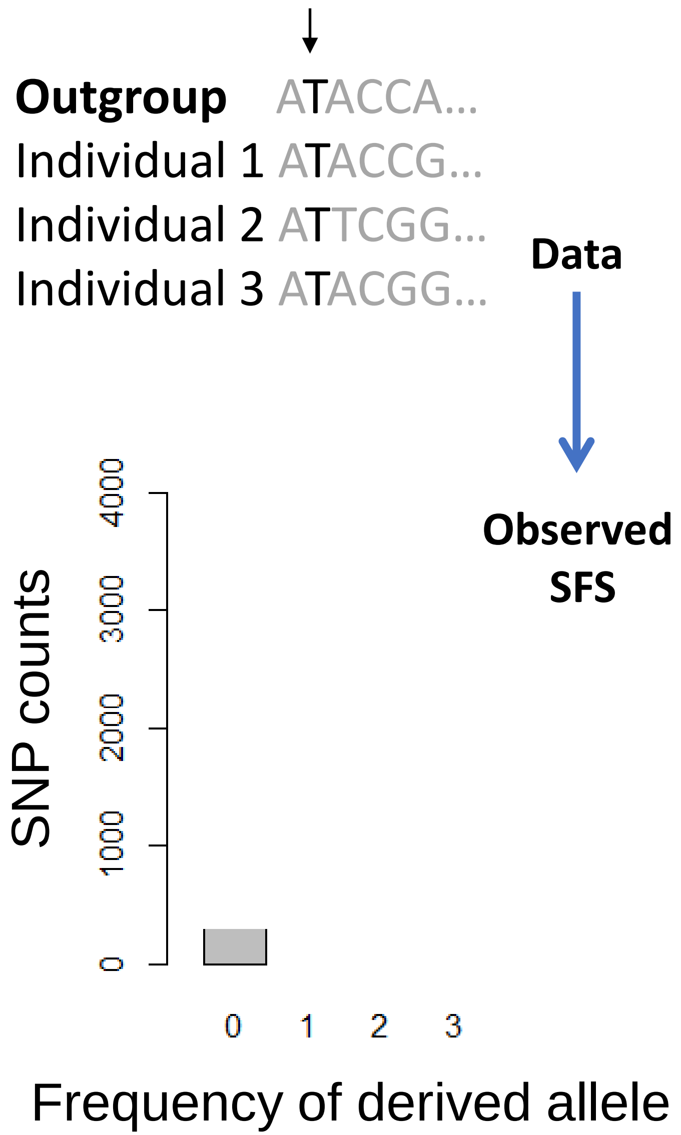
Data

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…

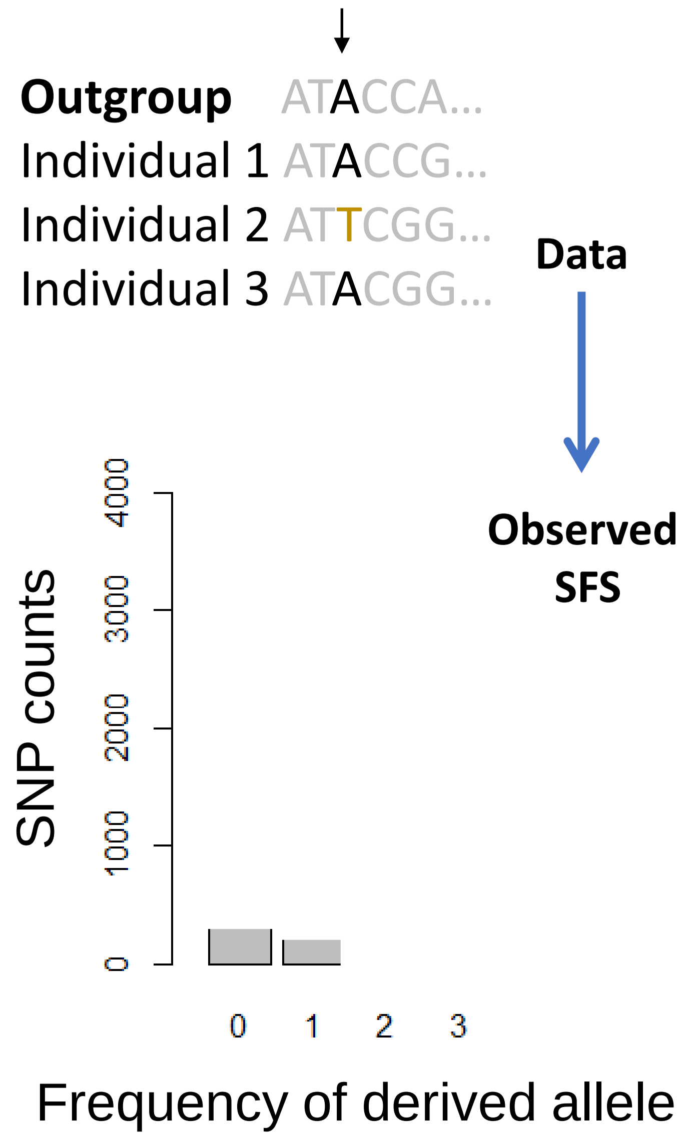
Data

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

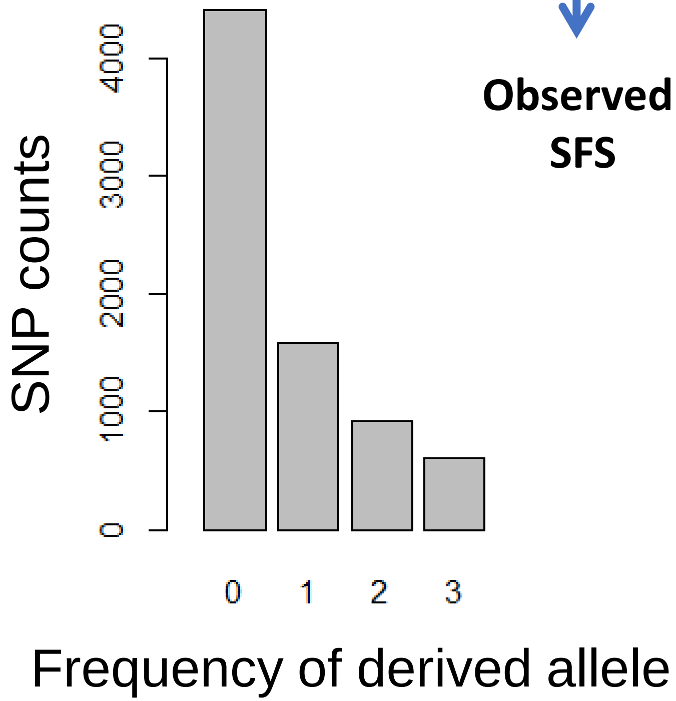
As the ancestral state is known, we can infer the derived SFS -> of derived allele frequency (DAF)

If the ancestral state is not known, we infer the minor allele frequency / folded SFS

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

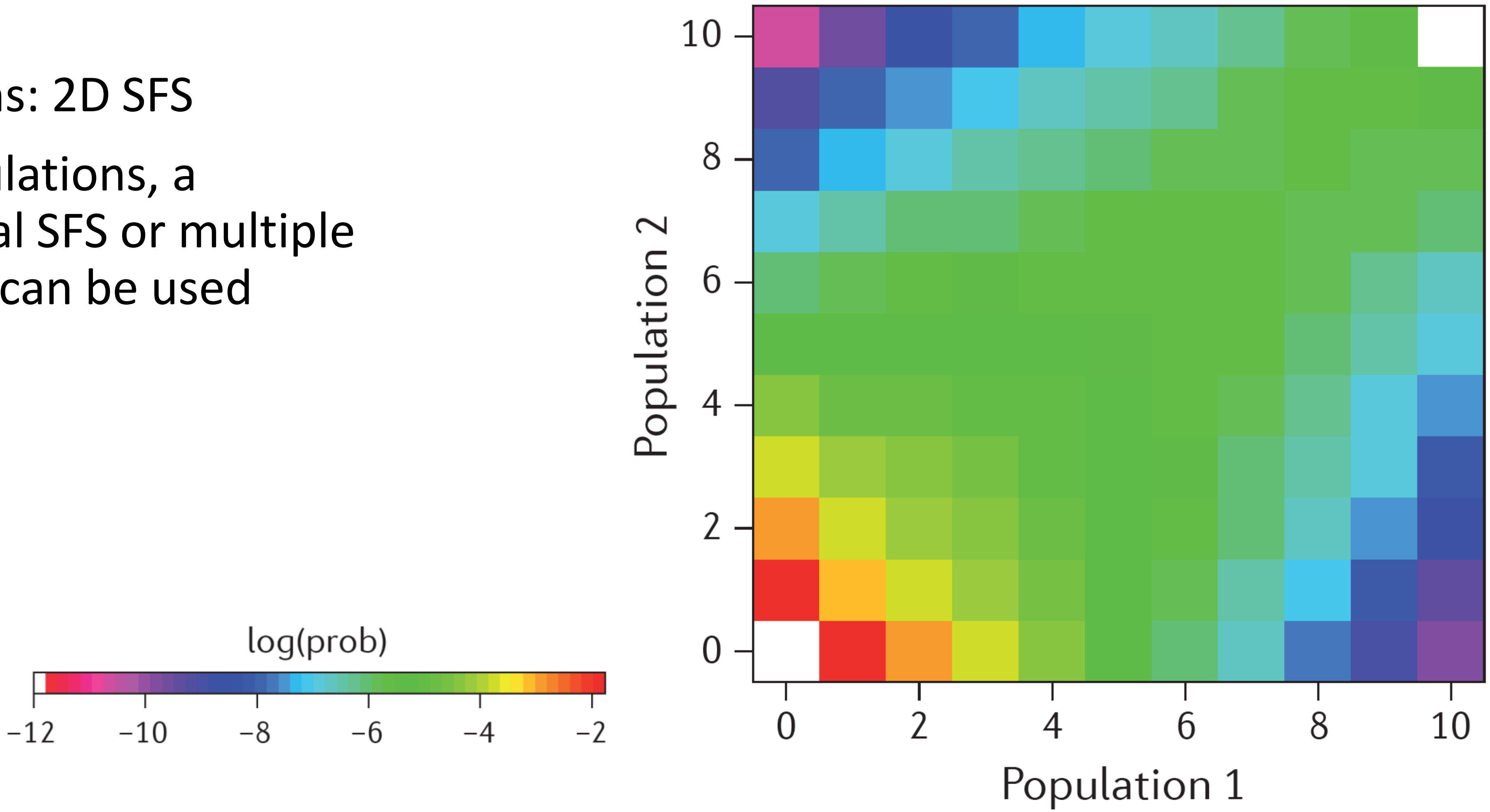
Data

Observed SFS



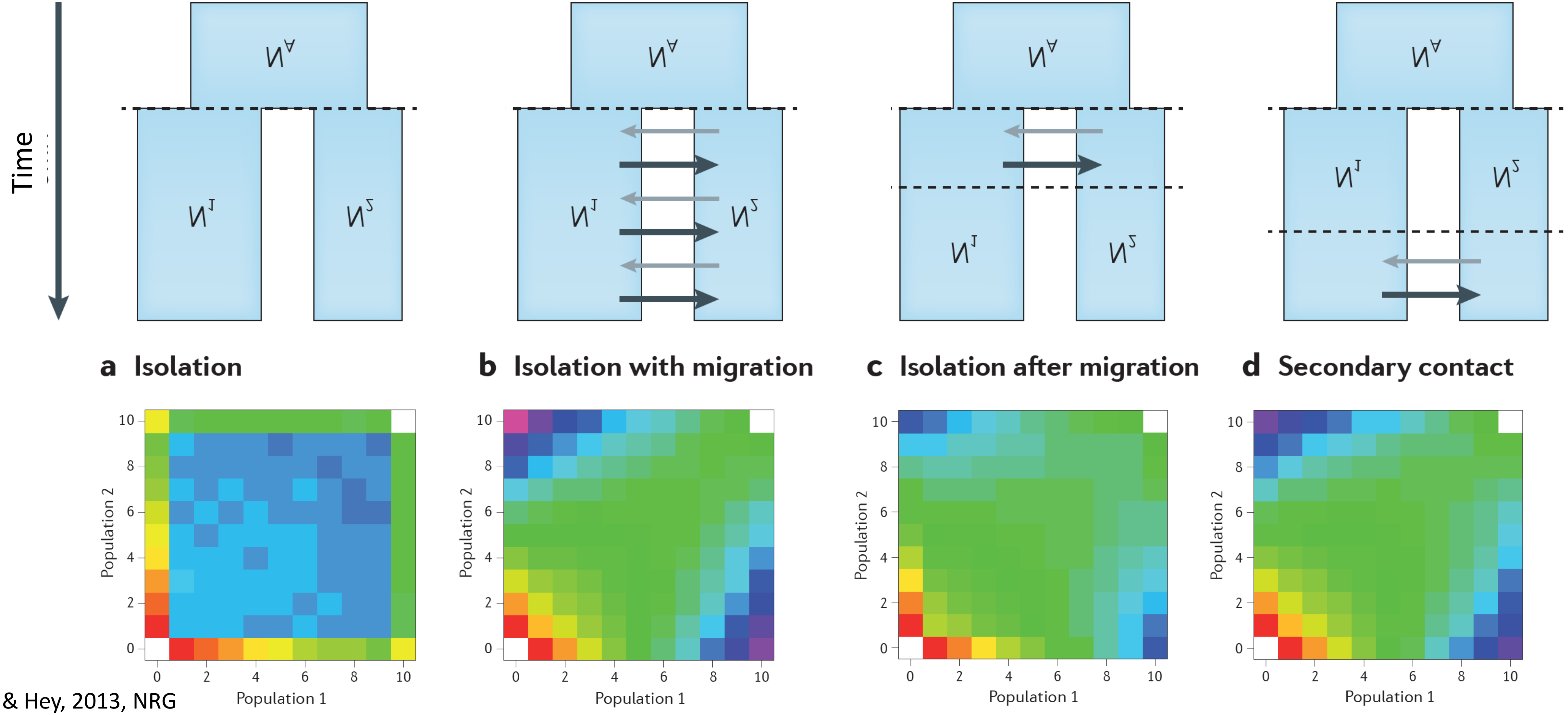
SFS for more than one population

- For 2 populations: 2D SFS
- With more populations, a multidimensional SFS or multiple pairwise 2D SFS can be used



Sousa & Hey, 2013, NRG

Expected SFS under different evolutionary scenarios



a Isolation

b Isolation with migration

c Isolation after migration

d Secondary contact

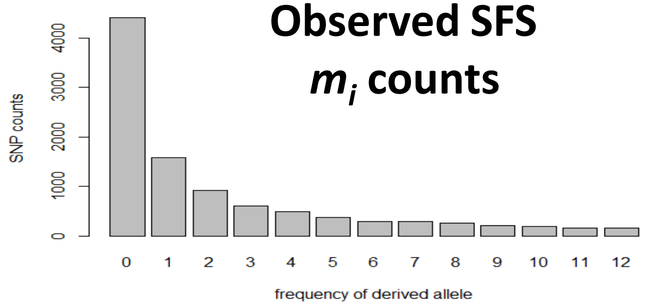
Sousa & Hey, 2013, NRG

Composite likelihood

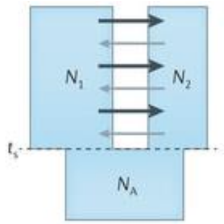
3 ingredientes for likelihood

Observed SFS

***m_i* counts**

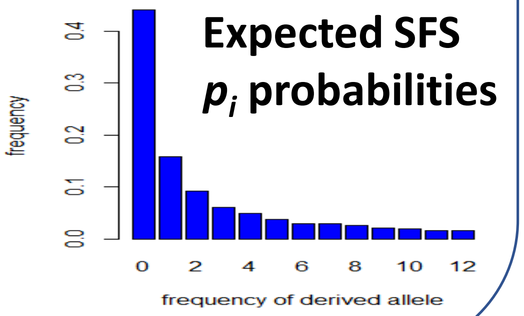


Model



Expected SFS

***p_i* probabilities**

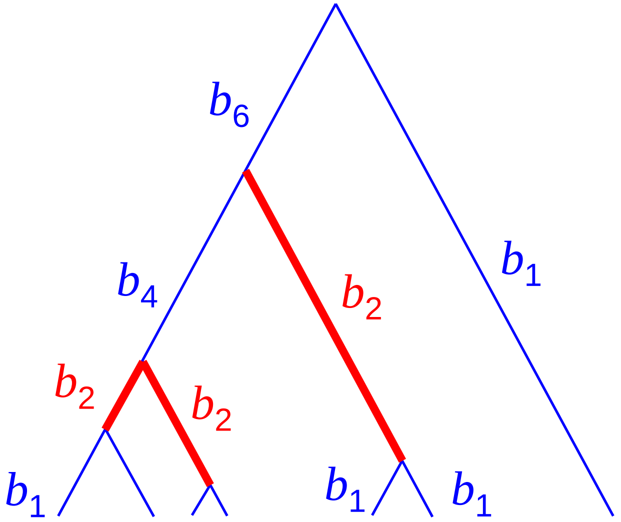


Given *S* polymorphic sites (SNPs) out of *L* sites (Adams and Hudson, 2004) the composite likelihood is:

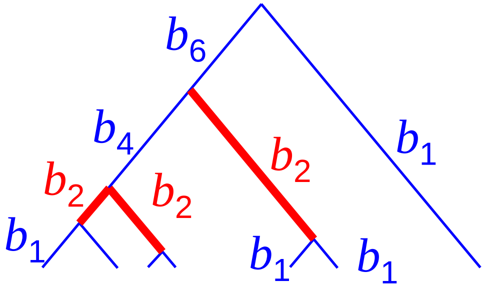
$$CL = \Pr(X \mid \theta) \propto \underbrace{P_0^{L-S}}_{\text{probability of no mutation on the tree}} \underbrace{(1-P_0)^S}_{\text{probability of at least one mutation in the tree}} \prod_{i=1}^{n-1} \hat{p}_i^{m_i}$$

Excoffier et al. (2013) *PloS Genetics*

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



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

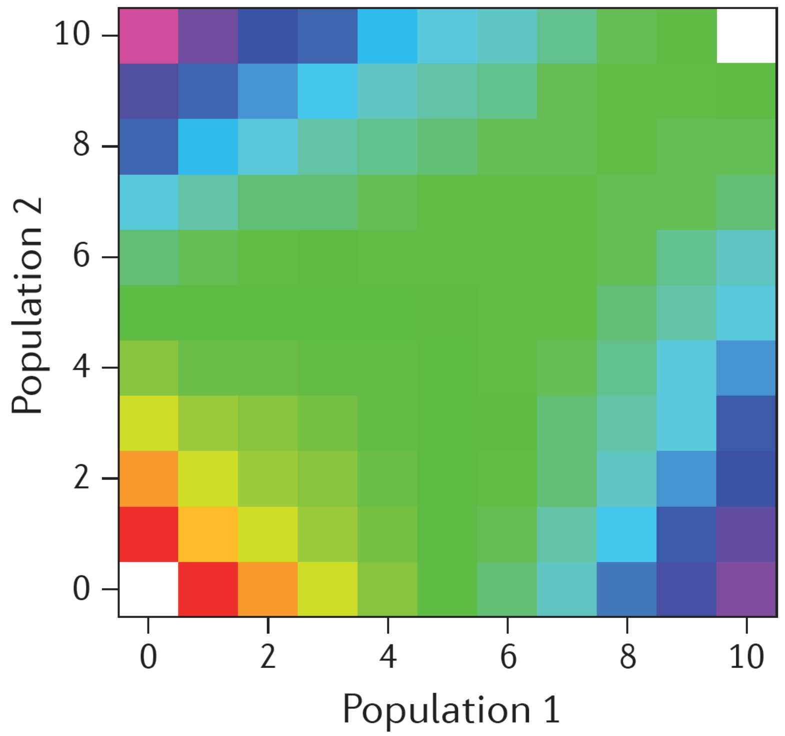
- We need a mutation rate and the number of monomorphic sites to distinguish among the two!
- Or we need to fix some parameters, e.g. the splitting time

fastsimcoal

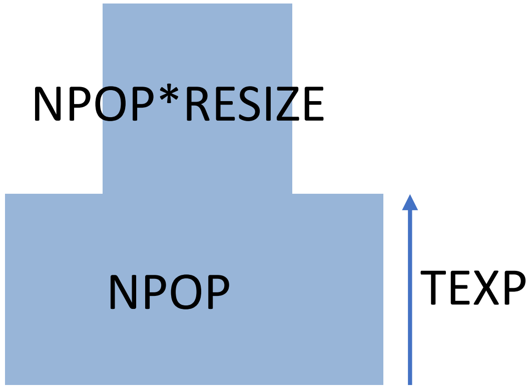
- Fastsimcoal2 can estimate parameters from the SFS using coalescent simulations
- Maximum (composite) likelihood method
- Uses a conditional expectation (CEM) maximization algorithm to find parameter combinations that maximize the likelihood
- It approximates the expected SFS by performing coalescent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

Input files for fastsimcoal

Observed SFS



Model template file



NPOP*RESIZE

NPOP

TEXP

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
```

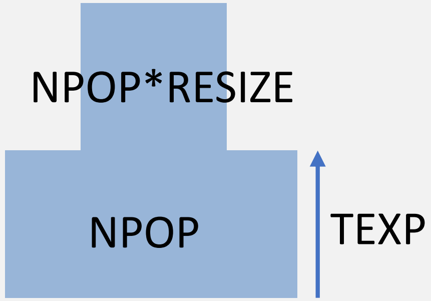
example_jointDAFpop1_0.obs

```
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
```

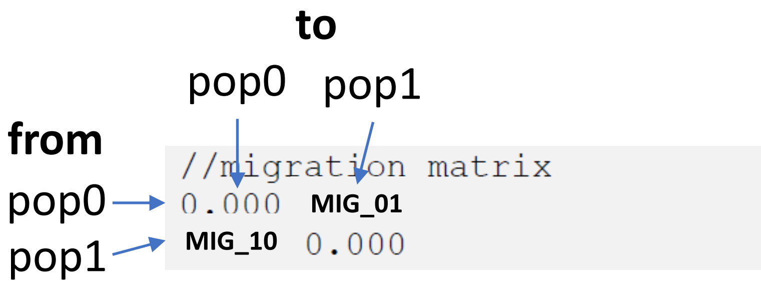


Input files for fastsimcoal2: Estimation file

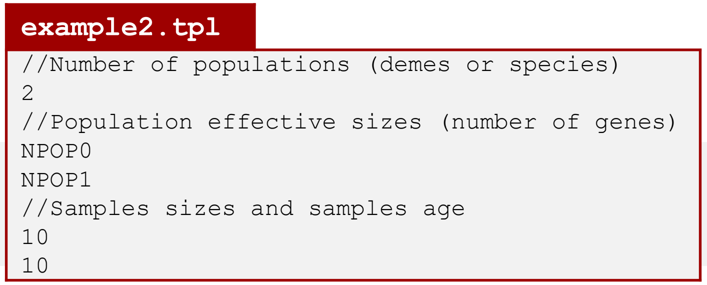
example.est

```
// 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
```

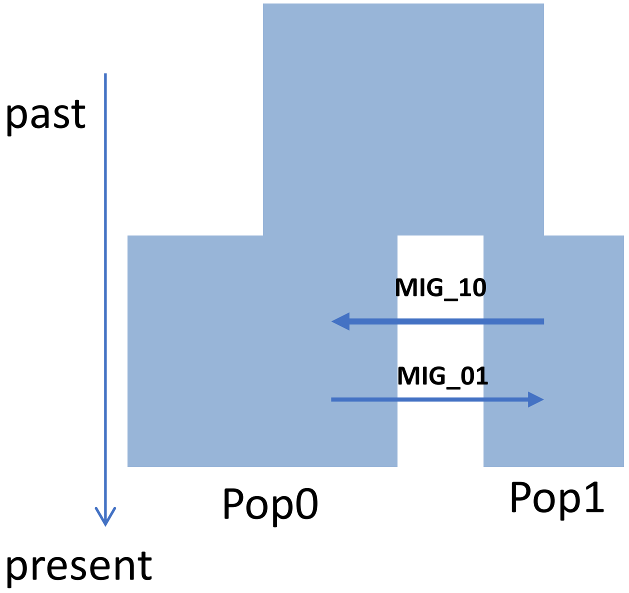


example2.tpl

```
//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.

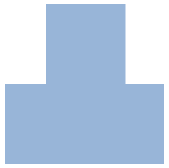
The entry m_{ij} lists the **migration rates backward in time** from population *i* to population *j*. The above-mentioned matrix states that, for each generation backward in time, any gene from population 0 has probability MIG_01 to be sent to population 1, and that a gene from population 1 has a probability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



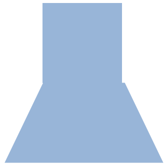
Historical events in fastsimcoal2

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

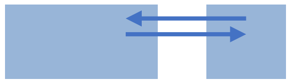
- Change the size of a given population



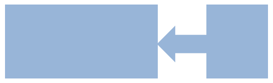
- Change the growth rate of a given population



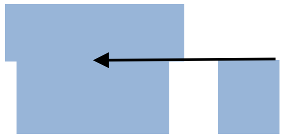
- Change the migration matrix



- 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.

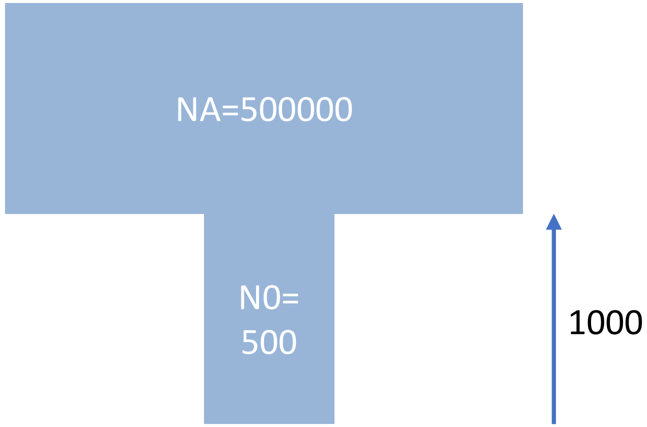


- One or more of these events can occur at the same time
- In the end, all populations must have fused to a single population

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



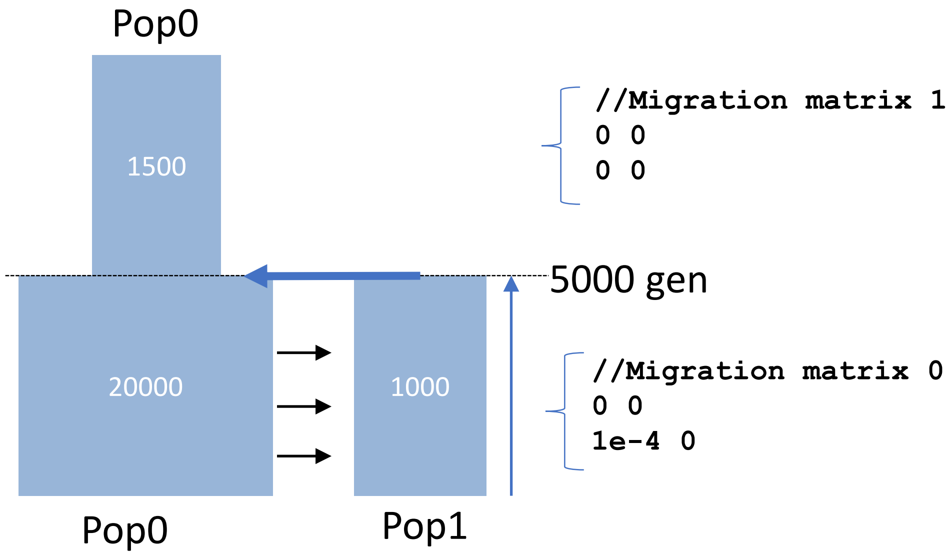
- 1000 generations ago, 0% (migrants=0) of lineages in pop0 (source) migrated to pop0 (sink). This means that 100% of lineages remained in pop0.

- 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.

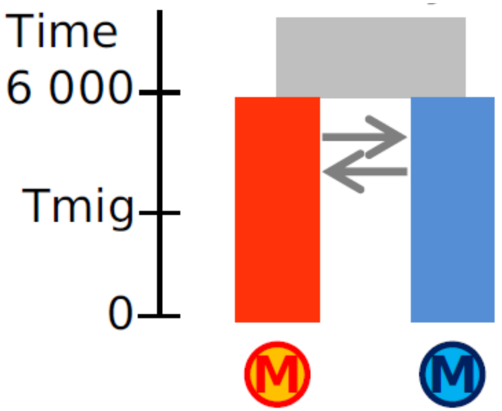
Principle of demographic parameter inference with fastsimcoal2

These operations are done by fastsimcoal2 to estimate parameters

1. Read the *tpl* and *est* files
2. Read the observed SFS (must have same generic name as *tpl* file)
3. Draw random initial values of parameters to be estimated, as defined in *est* file
4. Compute complex parameters function of simple parameters
5. Use the current parameter values to perform coalescent simulations necessary to estimate the expected SFS
6. Compute the likelihood of the parameters using a multinomial distribution
7. For each parameter in turn, use an optimization algorithm to find the parameter value that maximizes the lhood, keeping all other parameters constant
8. Loop step 7 for all parameters
9. Repeat steps 7 and 8 (loops) as many times as specified in the command line
10. Output parameter values with final best associated lhood

Now, let’s write our own model

Model with early gene flow (isolation after migration)



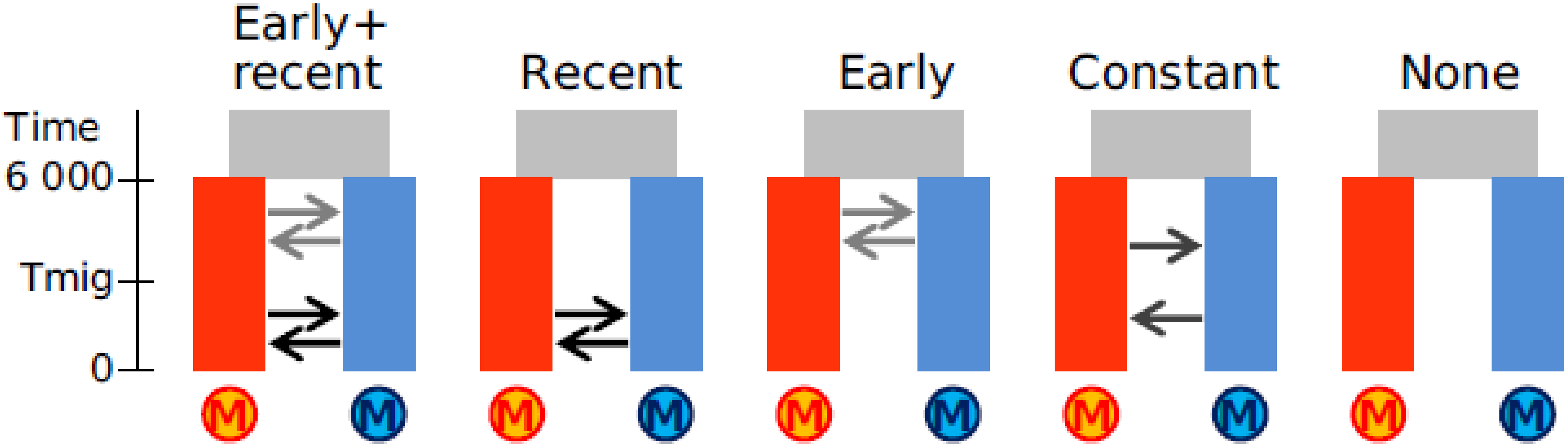
First, we test if a model of speciation with divergence with gene flow and then complete reproductive isolation fits the data well.

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


Early+
recent
Recent
Early
Constant
None
Time
6 000
Tmig
0
M
M
M
M
M
M
M
M
M
M

Steps to infer the best model

- Run each model at least 100 times to find the best run (highest likelihood)

- Compare models to infer the best model
  - AIC
  - Likelihood distributions

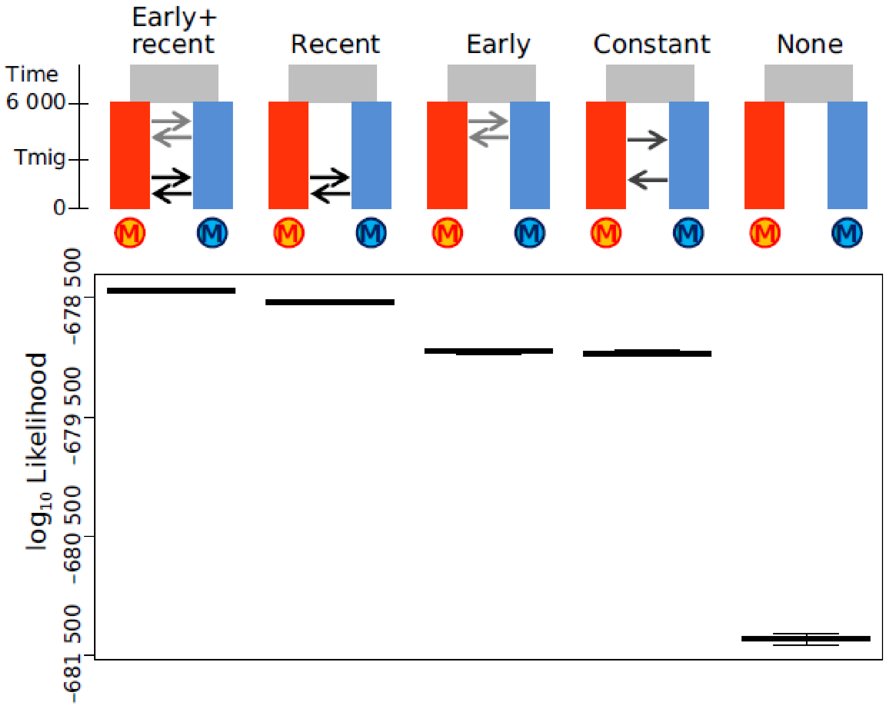
- Best model:
  - Bootstrapping to get confidence intervals for parameter estimates

Model comparison

AIC (Akaike information criterion)

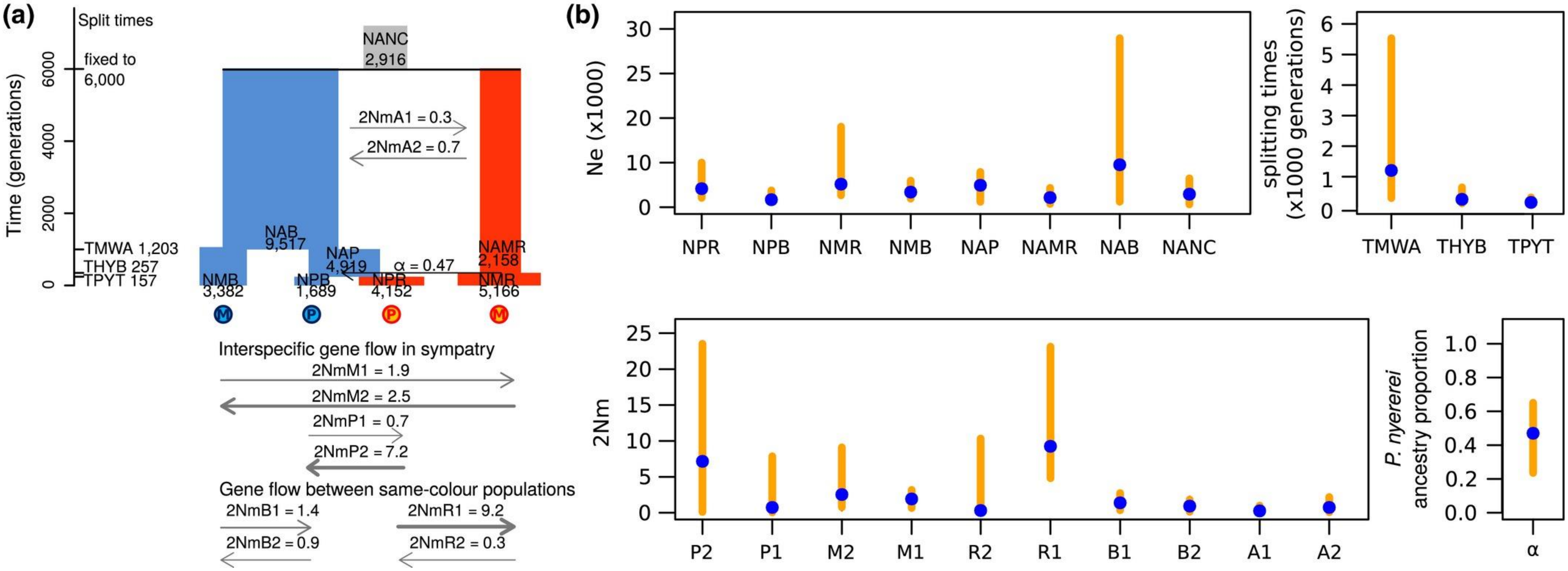
Gene flow model	Two gene flow matrices (recent and early)	Recent gene flow	Early gene flow	Constant gene flow	No gene flow
NMR	5,910	7,070	5,879	7,421	67,998
NMB	5,332	5,541	4,935	6,209	67,629
Nanc	3,637	3,723	3,089	4,785	41,549
Tmig	420	925	1	NA	NA
early mig B->R	4.2x10^{-6}/0.02	NA	3.3x10^{-4}/1.64	NA	NA
early mig R->B	3.3x10^{-4}/1.96	NA	6.8x10^{-4}/4.00	NA	NA
recent mig B->R	9.4x10^{-4}/5.00	4.4x10^{-4}/2.42	NA	2.4x10^{-4}/1.51	NA
recent mig R->B	7.1x10^{-4}/4.22	9.4x10^{-4}/6.66	NA	5.3x10^{-4}/3.91	NA
ΔLhood	601	654	1,059	1,069	3,363
AIC	3,124,032	3,124,274	3,126,138	3,126,179	3,136,742
ΔAIC	-	242	2,106	2,147	12,710

Likelihood distributions



Meier et al, 2017, MolEcol

Confidence intervals for the best parameter estimates


(a)
Split times
fixed to
Time (generations)
NANC
2NmA1 = 0.3
2NmA2 = 0.7
NAB
NAP
α = 0.47
NAMR
TMWA 1,203
THYB 257
TPYT 157
NMB
NPB
NPR
NMR
M
P
P
M
Interspecific gene flow in sympatry
2NmM1 = 1.9
2NmM2 = 2.5
2NmP1 = 0.7
2NmP2 = 7.2
Gene flow between same-colour populations
2NmB1 = 1.4
2NmB2 = 0.9
2NmR1 = 9.2
2NmR2 = 0.3
(b)
Ne (x1000)
NPR
NPB
NMR
NMB
NAP
NAMR
NAB
NANC
splitting times
(x1000 generations)
TMWA
THYB
TPYT
2Nm
P2
P1
M2
M1
R2
R1
B1
B2
A1
A2
P. nyererei
ancestry proportion
α