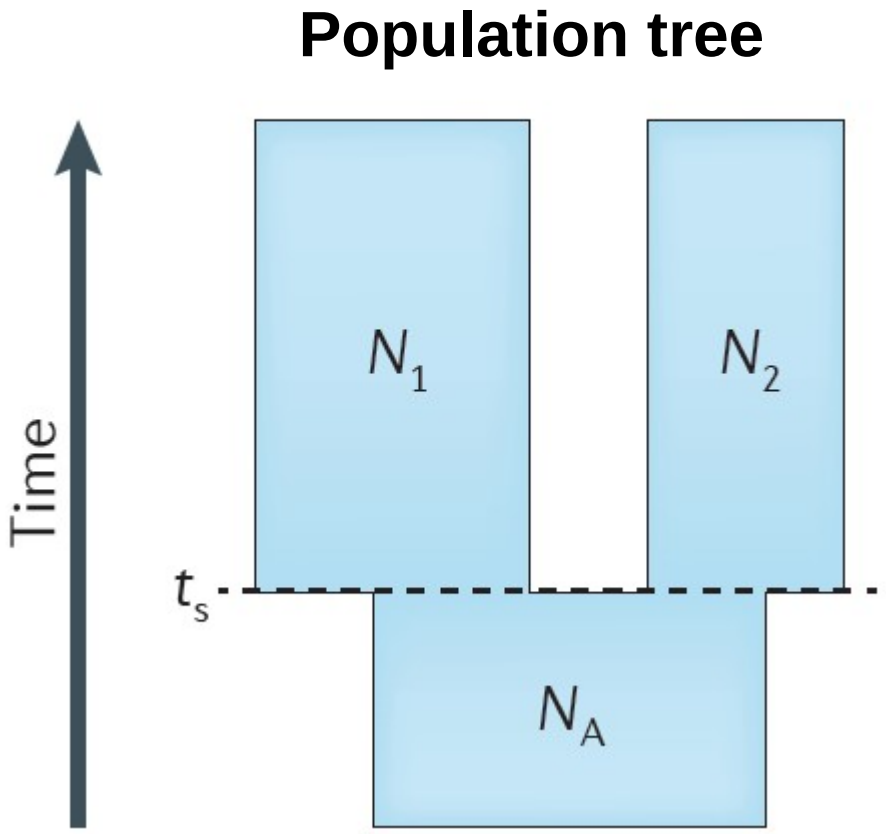
Intro to estimation of demographic parameters based on the Site Frequency Spectrum (SFS)

Vitor Sousa

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

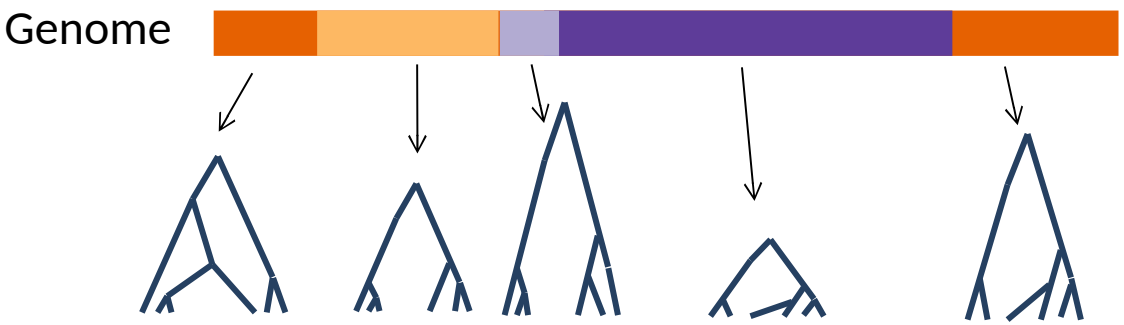
Because of recombination, different regions of the genome can have different gene trees

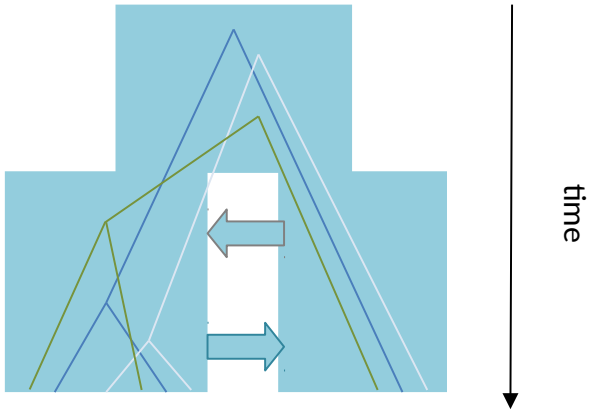
- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

Coalescent theory describes the relationship between gene trees and parameters of a population tree

Reconstructing the demographic history from genomic data

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



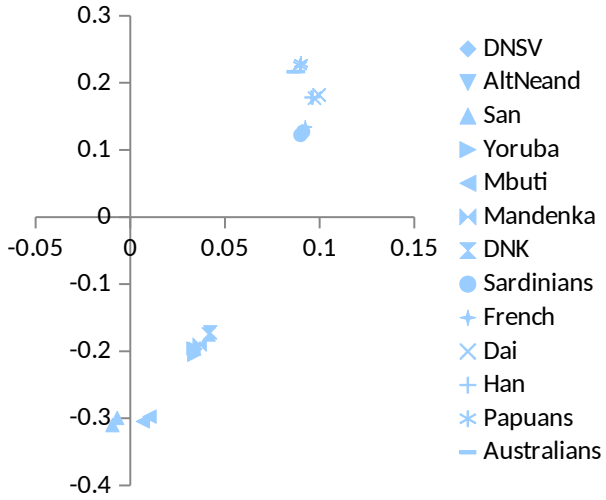


- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

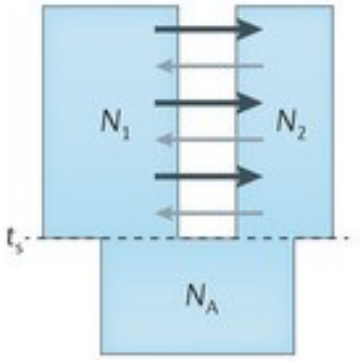
Coalescent theory describes the relationship between gene trees and parameters of a population tree

On the need of models to interpret population genetic data

«Model-free» methods
e.g. PCA



Model-based
methods



Ind. 1 ATACCG
Ind. 2 ATTCGG
Ind. 3 ATACCG

Genomic data

Evolutionary Processes:

- Demography
- Selection
- Mutation
- Recombination

Objectives

- Definition of site frequency spectrum (SFS)
- Intuition about how the SFS relates to demographic history
- How to perform parameter estimation using fastsimcoal2

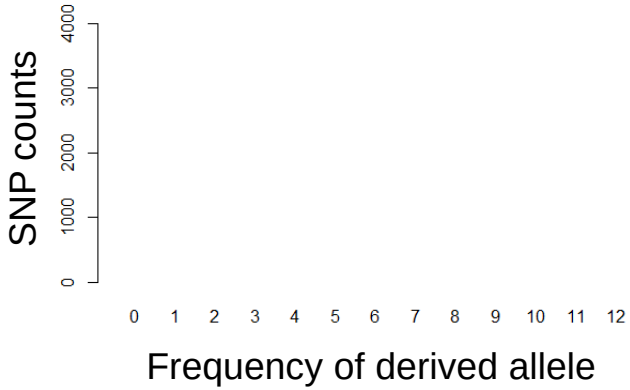
Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- At a single population – 1D SFS

Data

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

Observed SFS



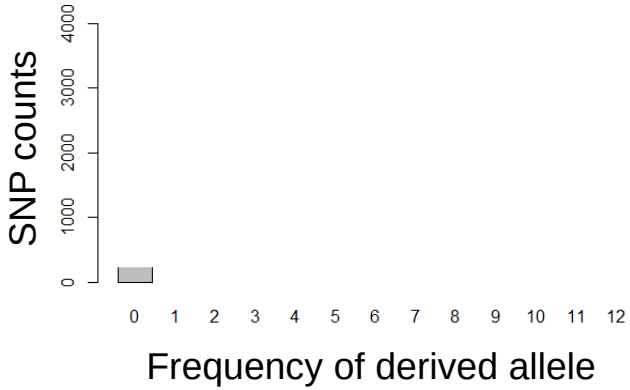
Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- At a single population – 1D SFS

Data

Observed SFS

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

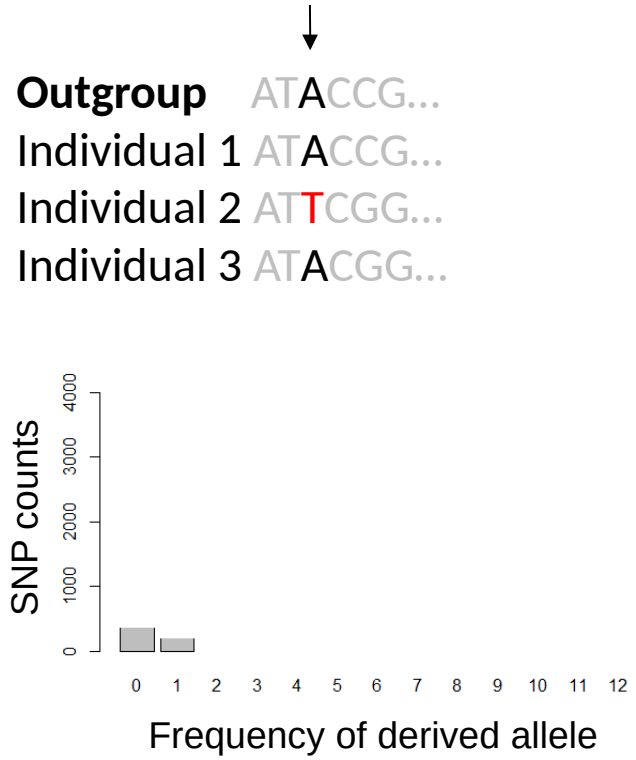


Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- At a single population – 1D SFS

Data

Observed SFS



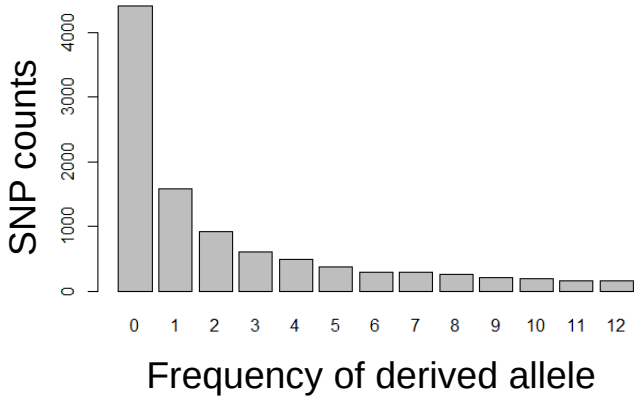
Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- At a single population – 1D SFS

Data

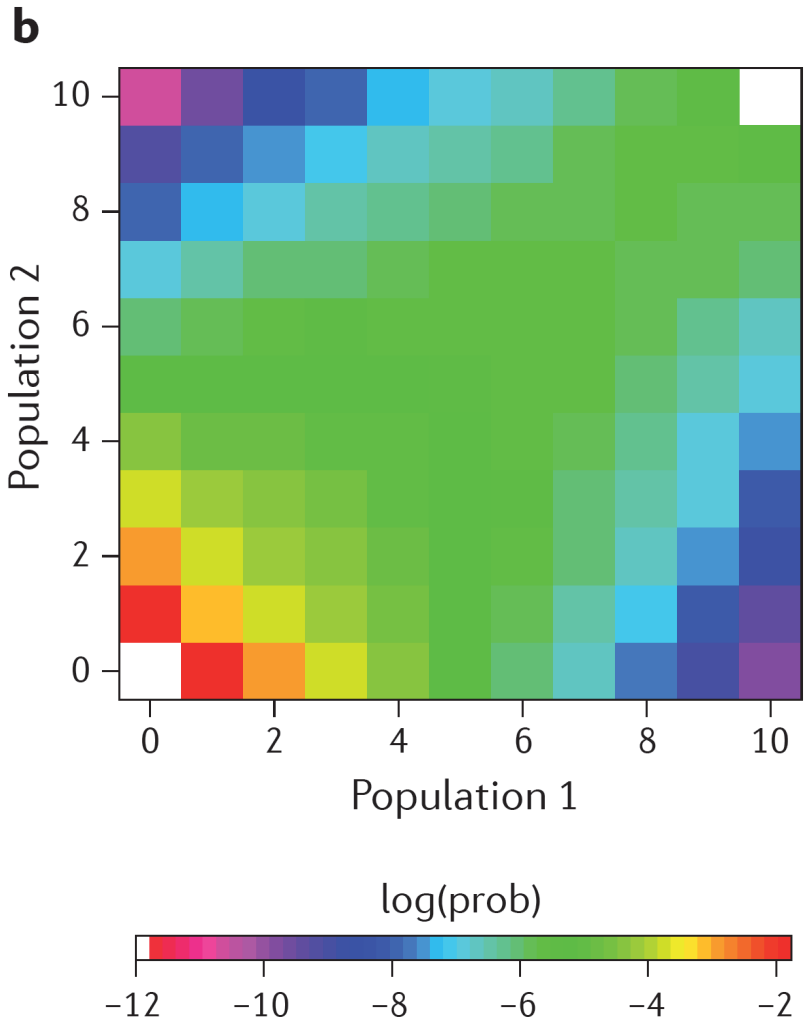
Outgroup ATACCG…
Individual 1 ATACCG…
Individual 2 ATTCGG…
Individual 3 ATACGG…

Observed SFS



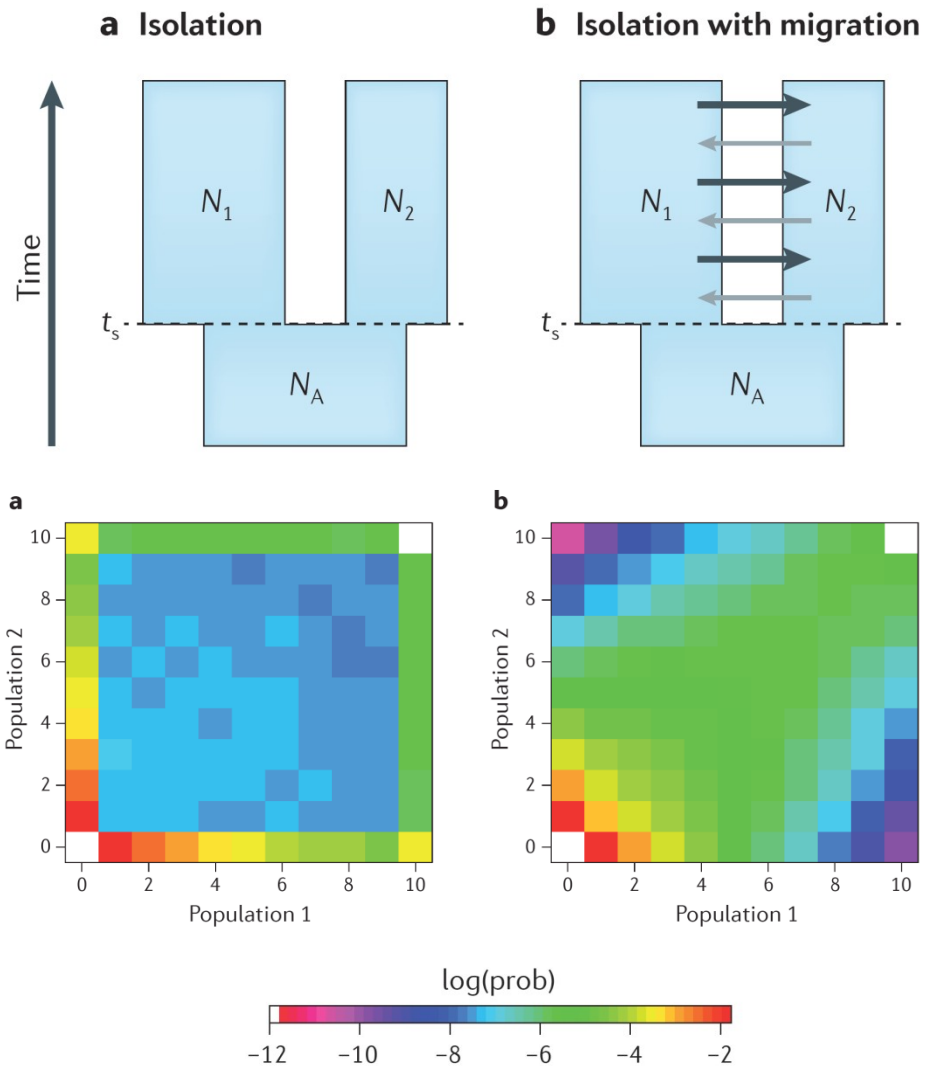
Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- At a single population – 1D SFS
- At multiple populations – 2D SFS



Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- At a single population – 1D SFS
- At multiple populations – 2D SFS
- **The SFS contains information about the demographic history of population**

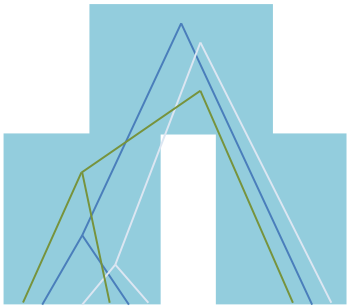


Sousa and Hey (2013) Nat. Rev. Genet.

Inferring the demographic history from the SFS

Genomic Data

Model

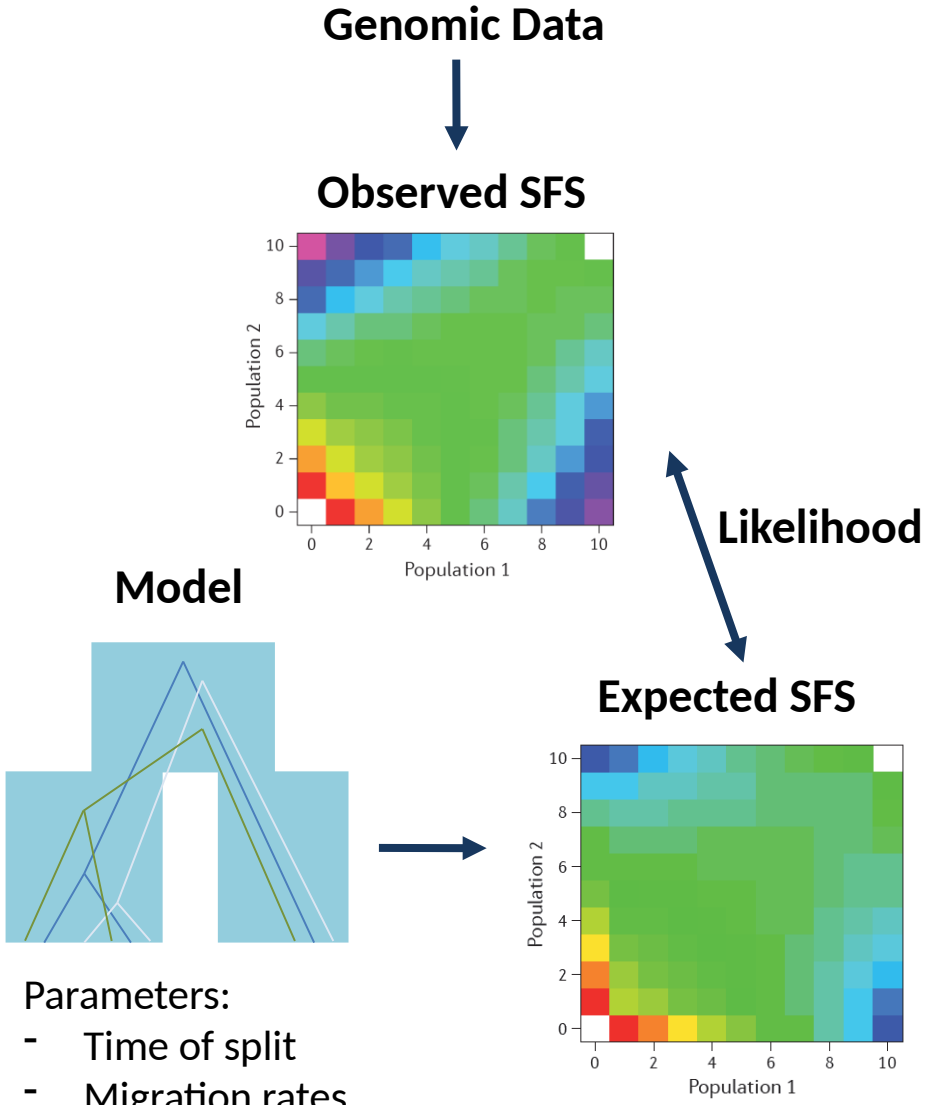


Parameters:

- 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
- Fastsimcoal2 finds the expected SFS under a model with coalescent simulations

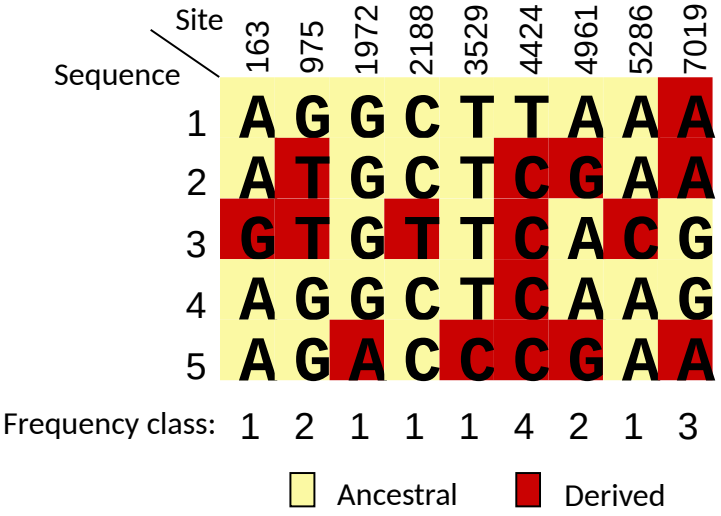


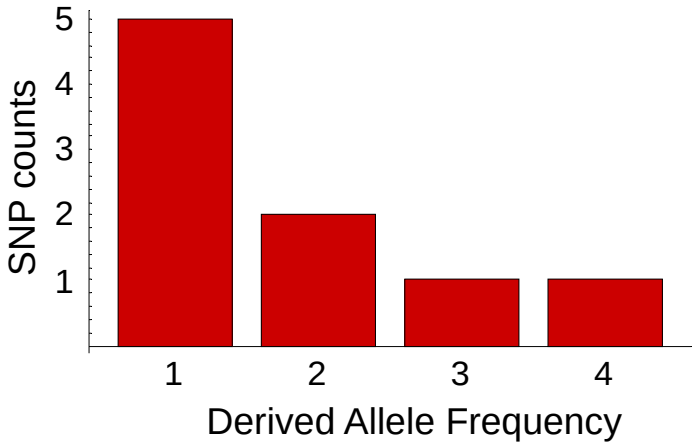
Excoffier et al. (2013) *PloS Genetics*

Global patterns of polymorphism

Site frequency spectrum (SFS)

One can summarize the pattern of polymorphism within a population by looking at the **distribution of derived allele (mutations) frequencies**





The **site frequency spectrum** is also influenced by the past demography of a population, but it does not use information about linkage between sites.

It is best suited for the study of many unlinked (or recombining) DNA sequences.

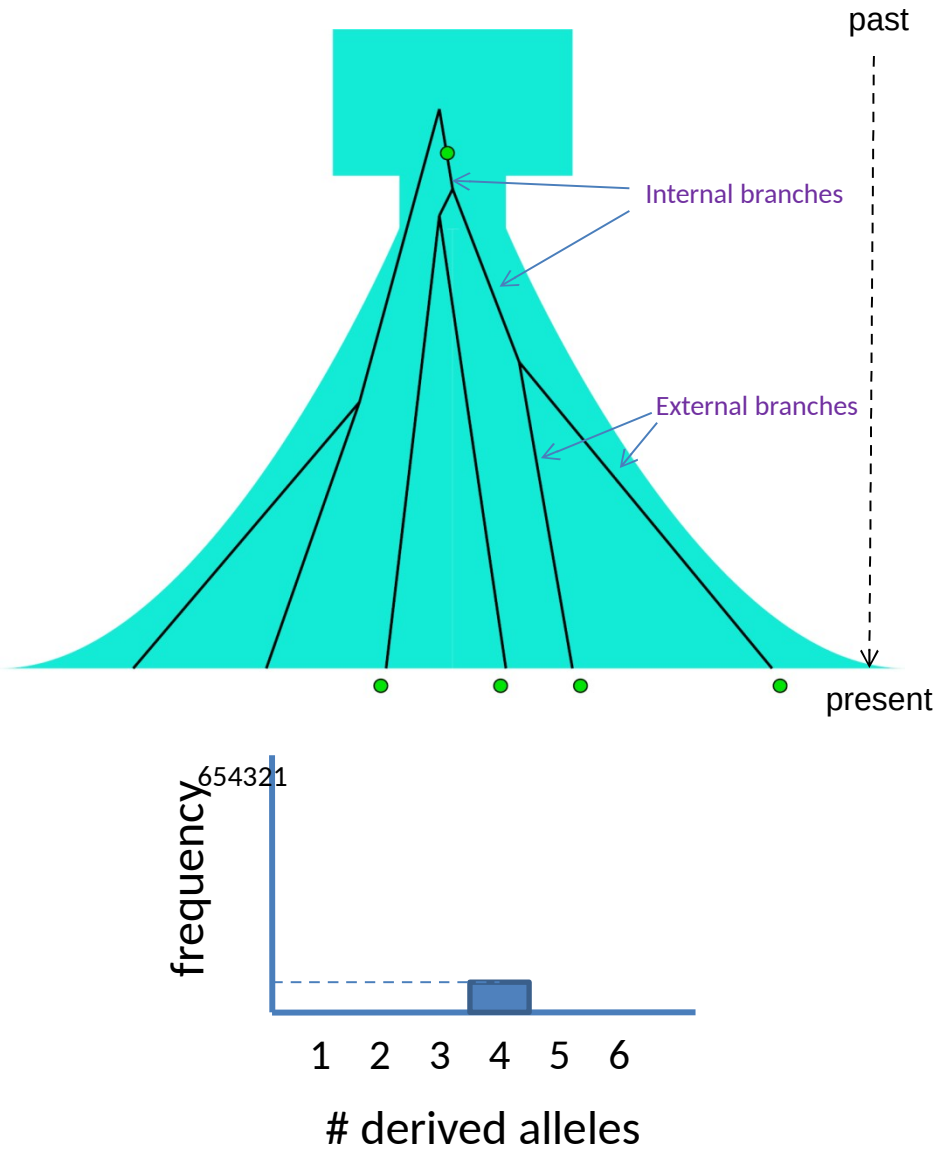
In a stationary population, the expected SFS relative frequencies are given by

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

Fu and Li, 1993

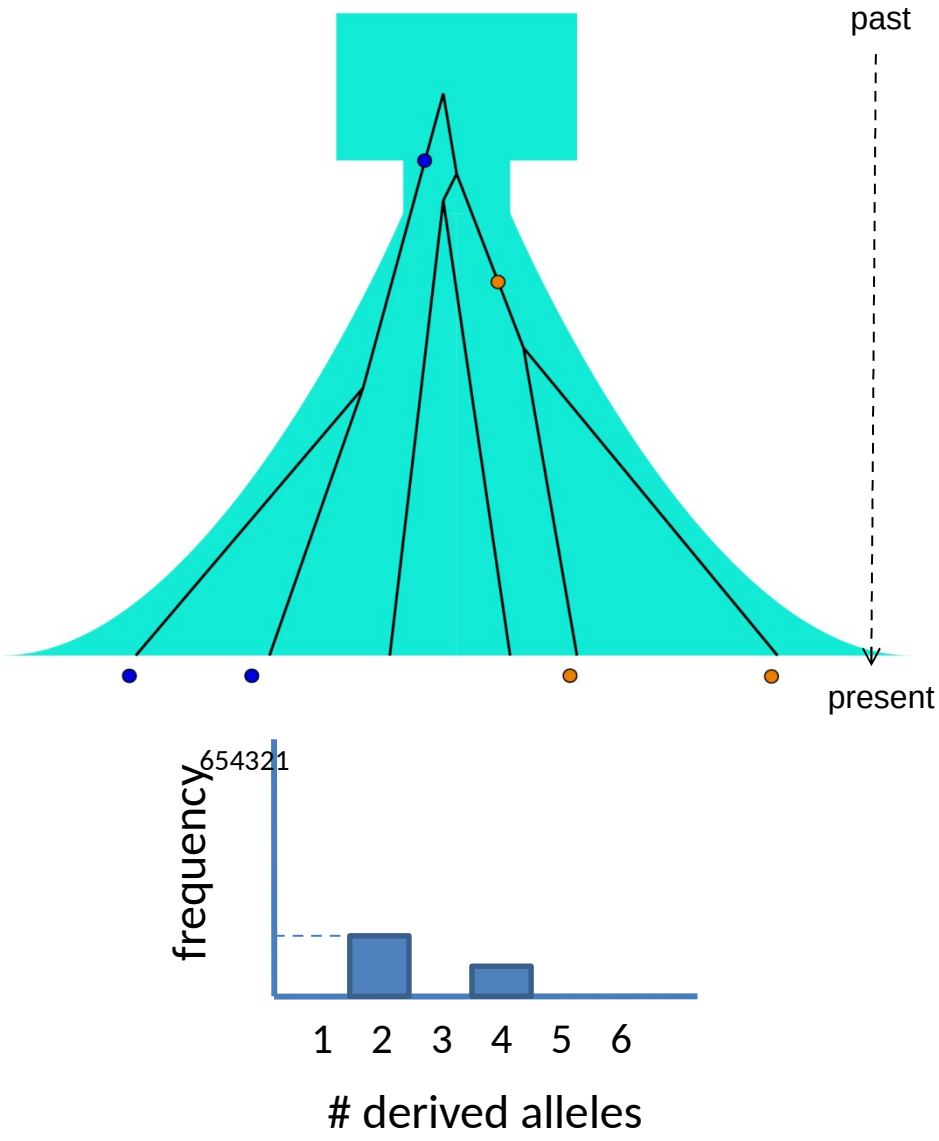
Demography affects the SFS

- 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



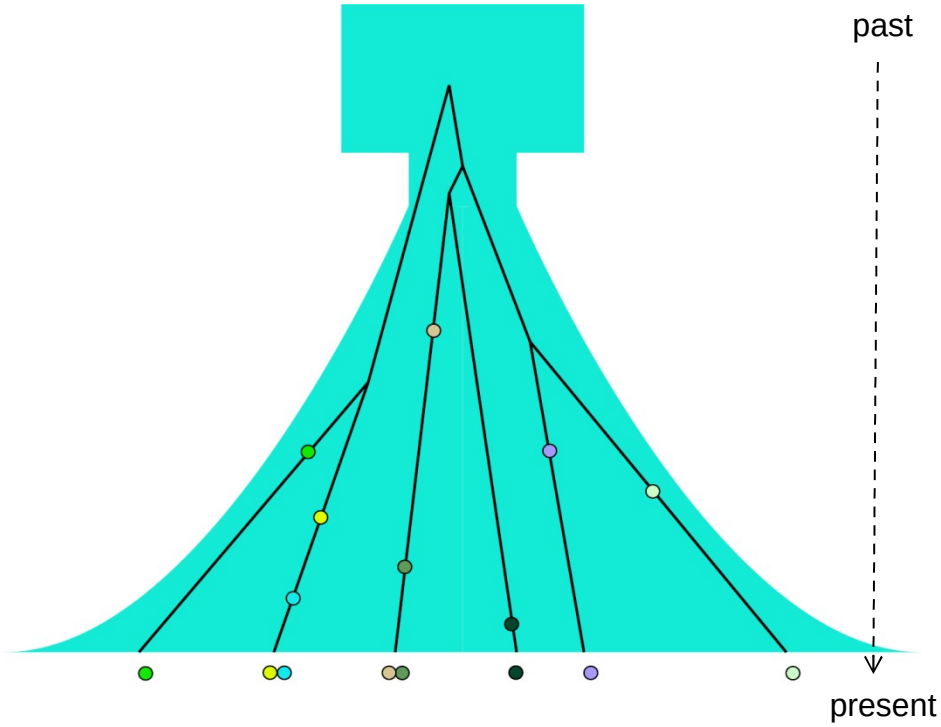
Demography affects the SFS

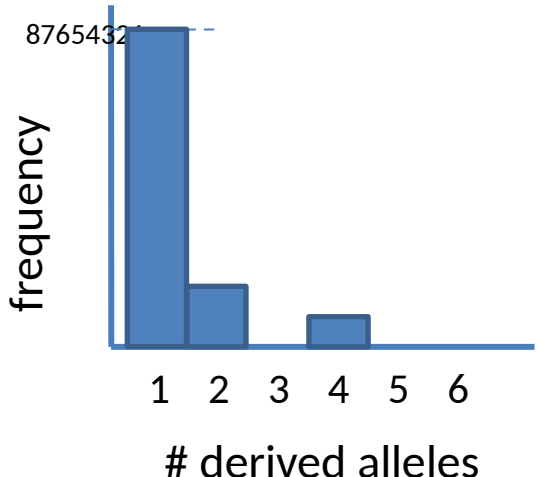
- A recent population growth following a bottleneck leads to gene trees with long external branches
- Very few mutations in the internal branches
- Most mutations in long external branches are only found in one lineage, resulting in an excess of singletons



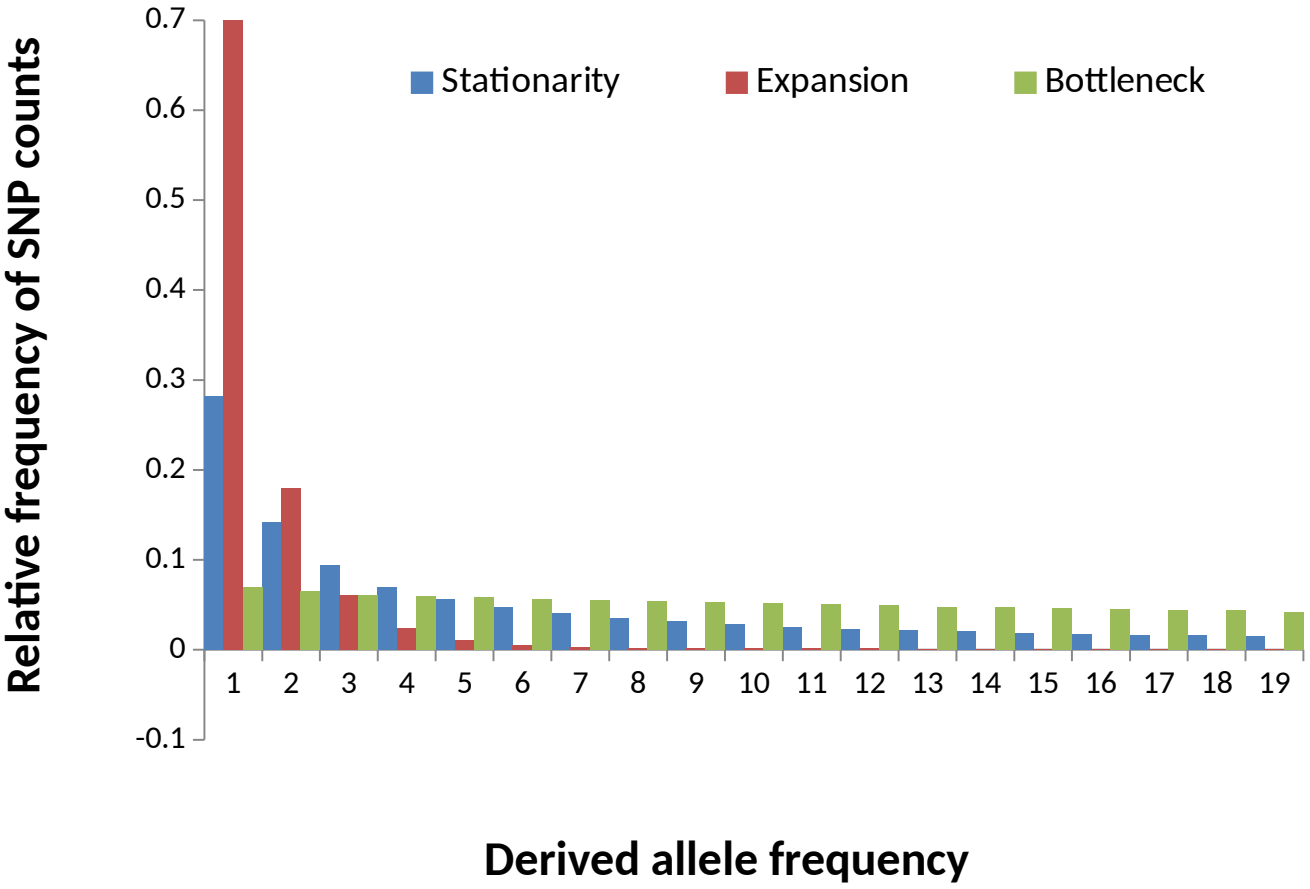
Demography affects the SFS

- 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

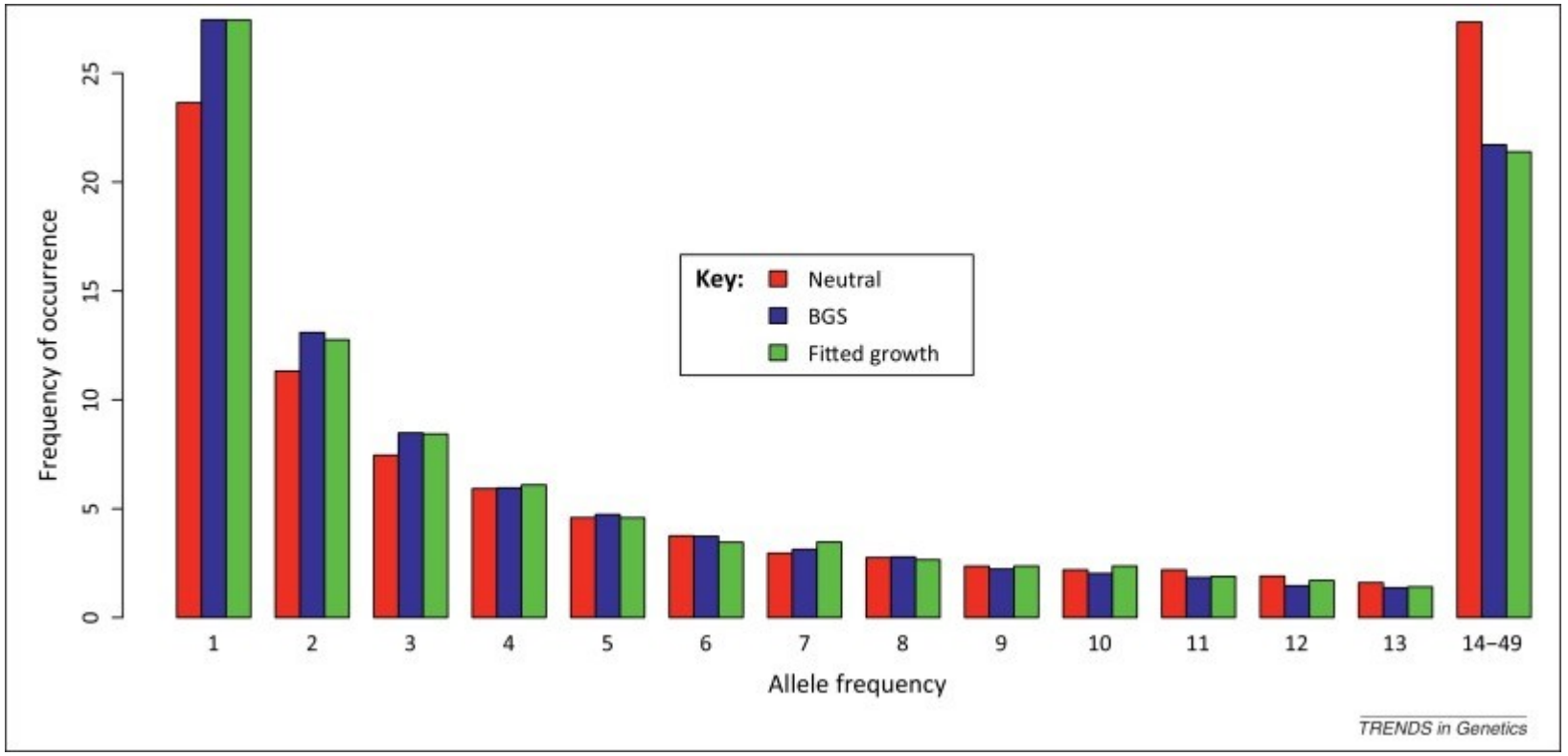




SFS depends on past demography



Natural selection also affects the SFS

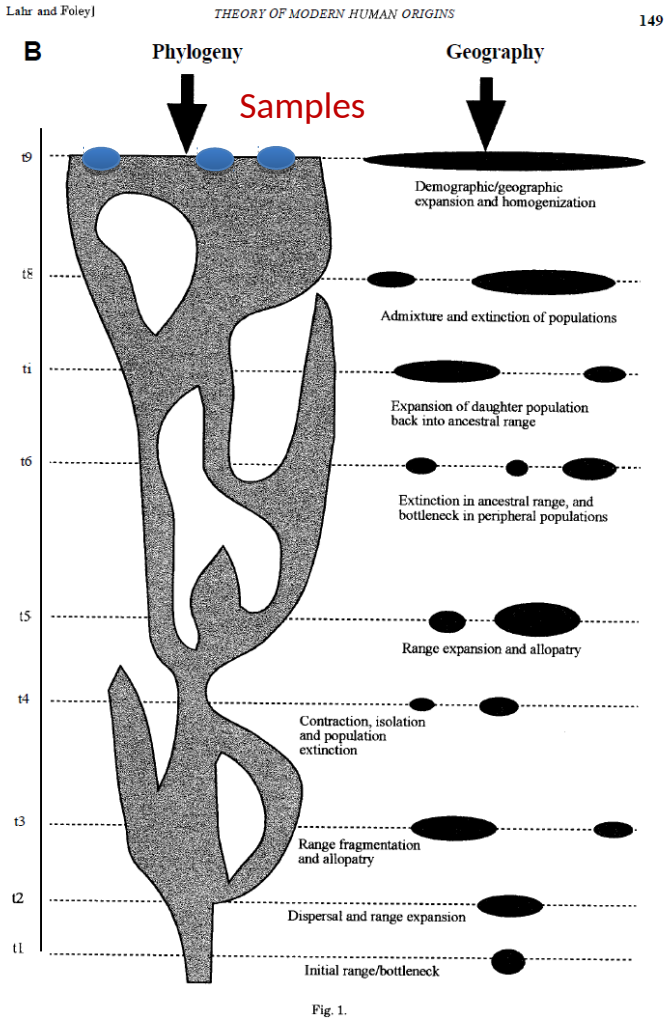


Background selection (BGS) leads to patterns similar to population expansion.

Bank et al (2014) Trends in Genetics

Framework for demographic inference

"Truth"



Lahr and Foley, 1994

Samples genetic analyses

Envisioned demographic scenarios

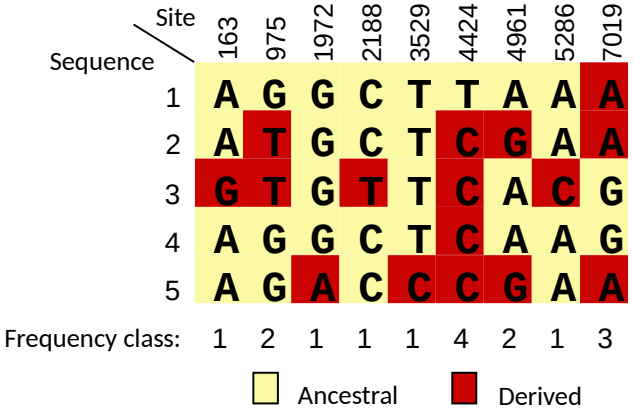
Models

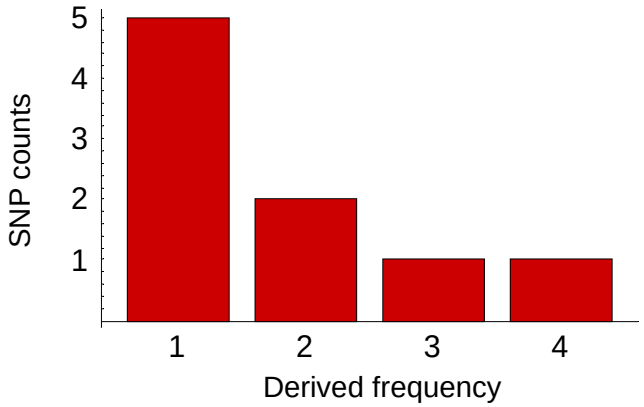
Estimation of demographic parameters

Composite likelihood approaches

Program $\partial a \partial i$: Diffusion Approximation for Demographic Inference
http://code.google.com/p/dadi/

- Uses the site frequency spectrum as data

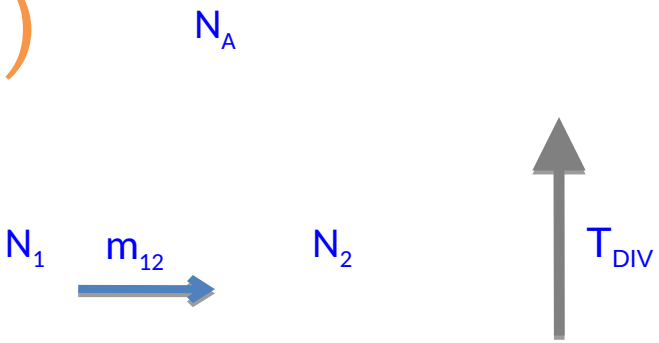




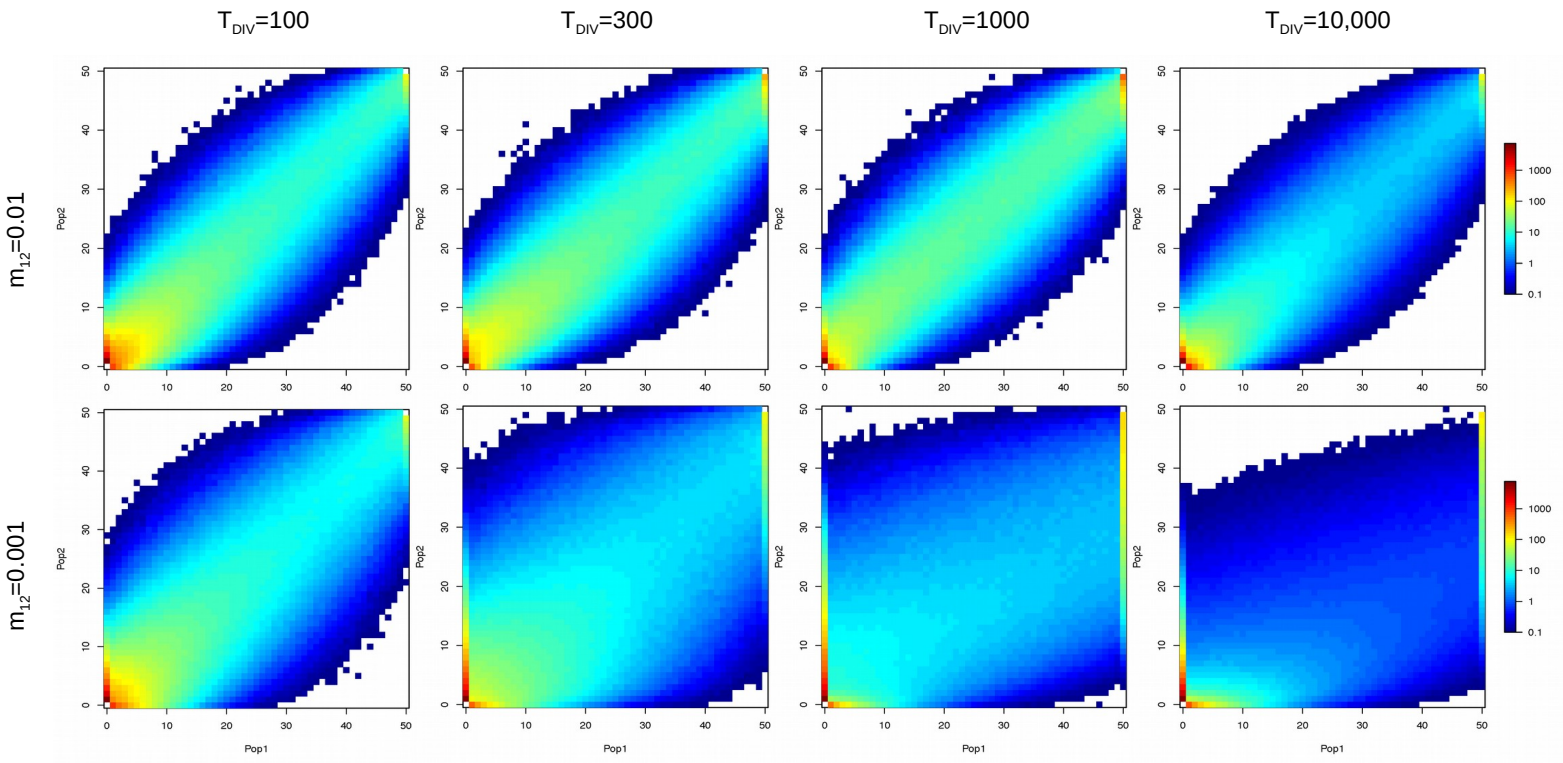
- Composite likelihood as it assumes that sites are unlinked
- Obtain likelihood by an efficient and flexible diffusion approximation (forward in time)
- Computation time does not depends on data size
- Can be applied to quite complex models, up to three populations
- Fast for simple models, but slower for more complex models due to sub-optimal routine for parameter optimization
- Obtain CI's by bootstrap approach
- Model validation is difficult (rarely done)

Joint SFS (2D-SFS)

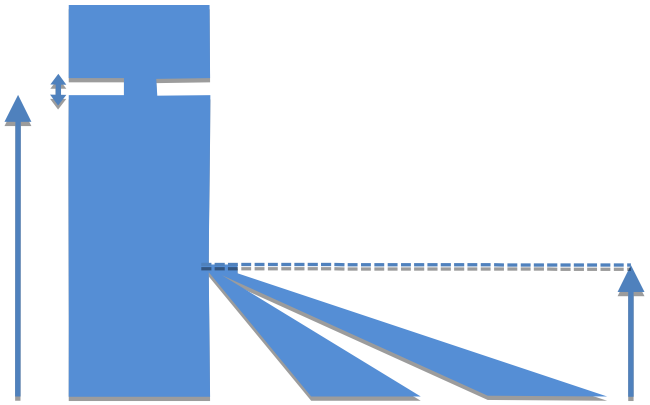
The joint SFS (2D-SFS) is a matrix where the entries are the frequencies of a derived allele with *i* copies in the first population and *j* copies in the second population

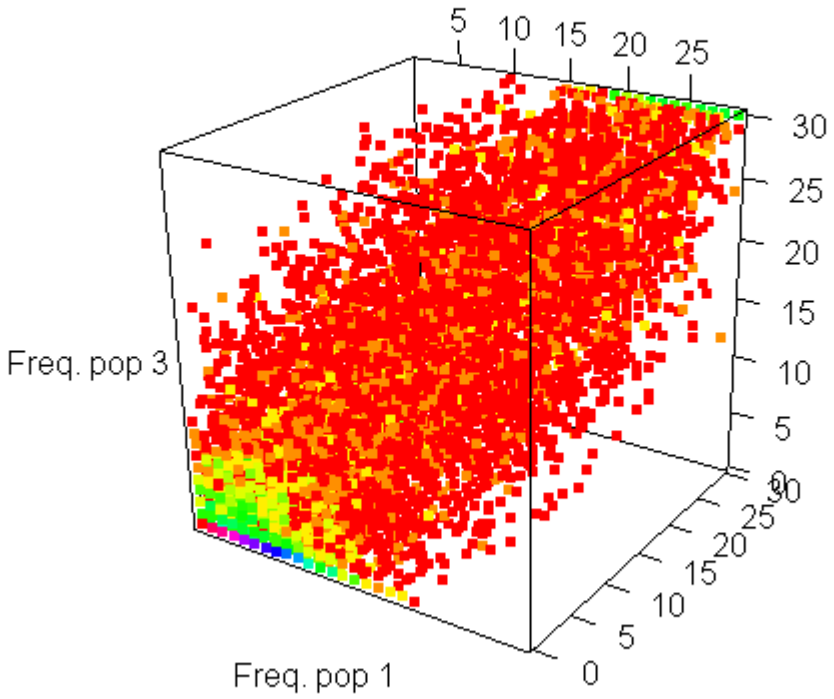


Model of Isolation with migration (IM)



Multidimensional SFS





Multidimensional SFS carries a lot of information on demography

Problems with estimation of demographic parameters from SFS

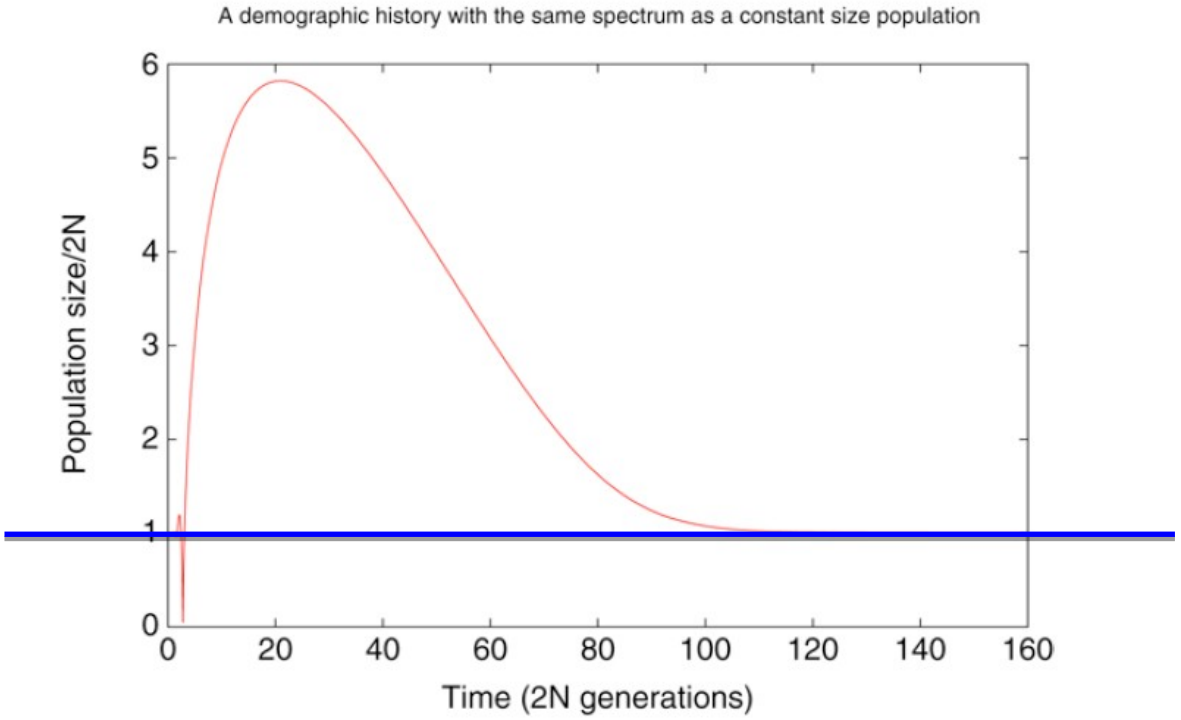
Can one learn history from the allelic spectrum?

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



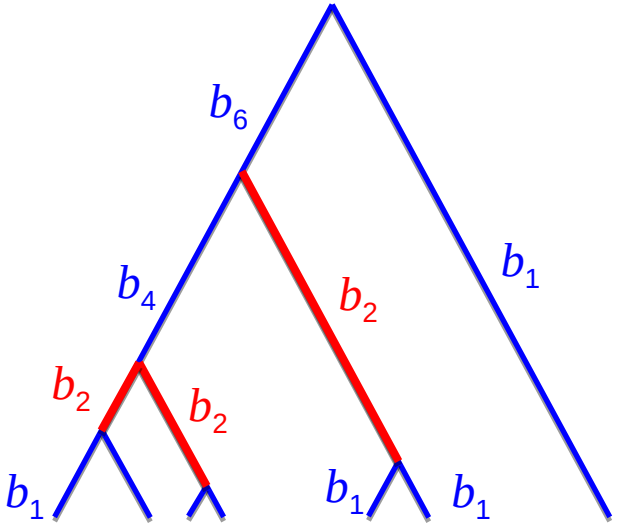
Estimating the SFS from the coalescent

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

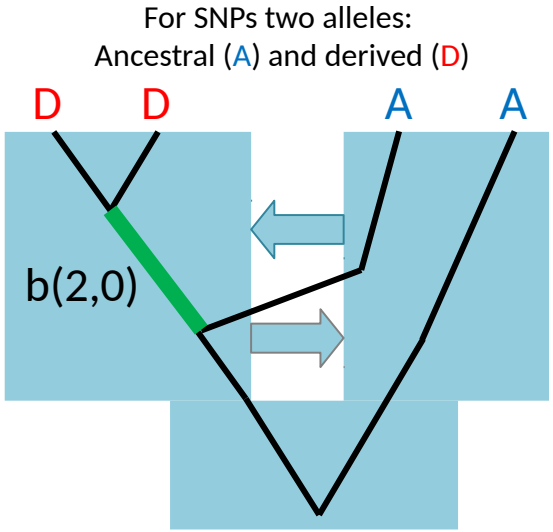


Gene trees and expected SFS
A two populations case

The expected SFS can be obtained given a gene tree under a given model

AFS(x,y)=E[b(x,y)]/E[Tlength]

- E[.] means expected value
- b(x,y) sum of branch length of branches leading to (x,y) configuration
- Tlength – total branch length of the tree



2D-SFS

Pop2 \ Pop 1	0	1	2
0			b(2,0)/Tl
1			
2			

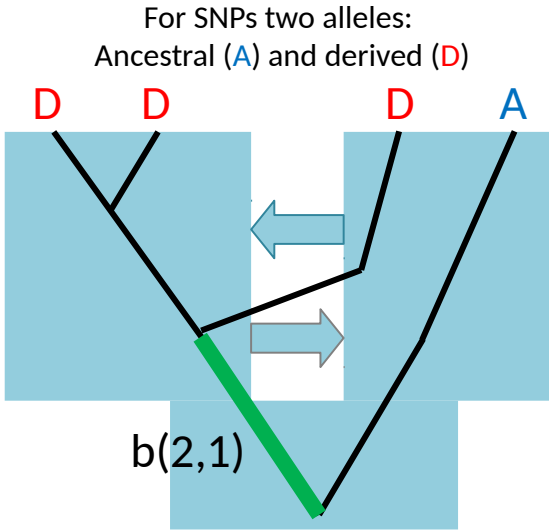
Nielsen (2000) Genetics

Gene trees and expected SFS
A two populations case

The expected SFS can be obtained given a gene tree under a given model

AFS(x,y)=E[b(x,y)]/E[Tlength]

- E[.] means expected value
- b(x,y) sum of branch length of branches leading to (x,y) configuration
- Tlength – total branch length of the tree



2D-SFS

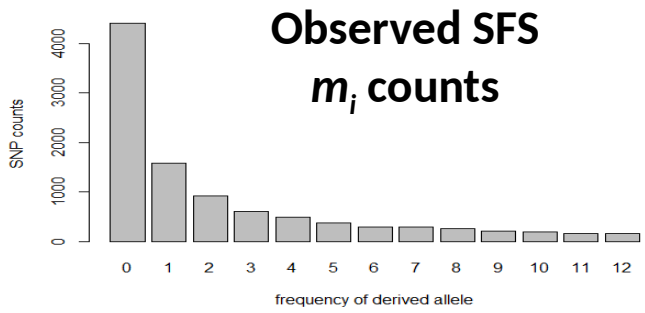
Pop2 \ Pop 1	0	1	2
0			b(2,0)/Tl
1			b(2,1)/Tl
2			

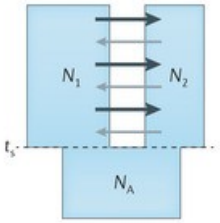
Nielsen (2000) Genetics

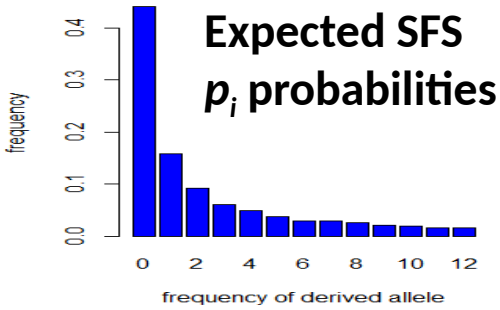
Composite likelihood

Assuming all sites are independent, given S polymorphic sites (SNPs) out of L sites (Adams and Hudson, 2004)

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







EstLhood

Estimating the SFS and likelihoods with 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

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
- CI 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
- In the current setting, CL ignores information on linkage disequilibrium (recombination) between sites, and cannot be used to estimate recombination dependent parameters

Advantages of SFS for parameter inference

- Accuracy of estimates increases with data size, but computing time does not
- Can be used in scenarios as complex as ABC
- Very fast estimation as compared to ABC to the same amount of data

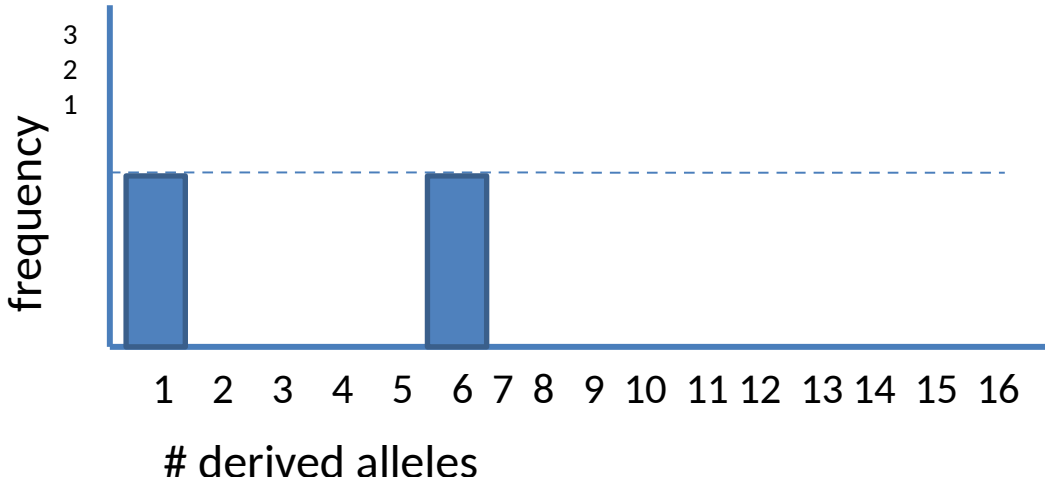
Potential problems

- 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 to correctly estimate SFS (likely to miss singletons or show errors)

Estimating SFS from observed data

- How to deal with missing data?

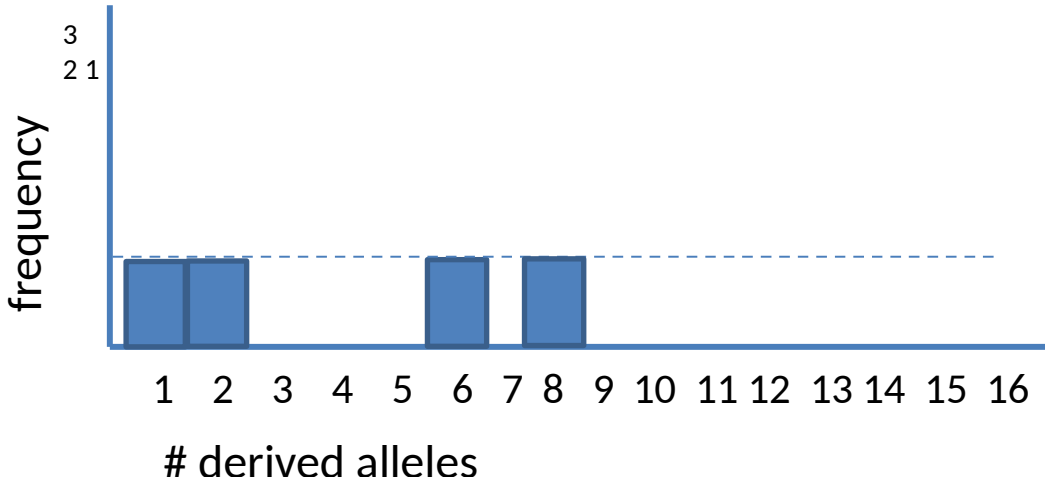
	Freq. derived	Sample size	Rel. freq
SNP1	1	16	1/16
SNP2	6	12	1/2
SNP3	1	12	1/12
SNP4	6	16	3/8



Estimating SFS from observed data

- How to deal with missing data?

	Freq. derived	Sample size	Rel. freq
SNP1	1	16	1/16
SNP2	6	12	1/2
SNP3	1	8	1/12
SNP4	6	16	3/8

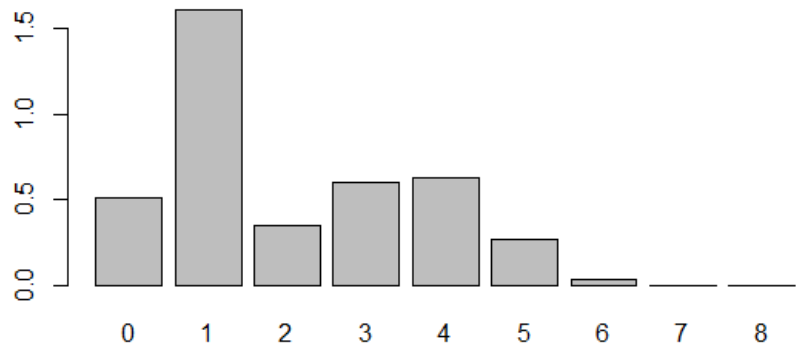


Estimating SFS from observed data

- How to deal with missing data?

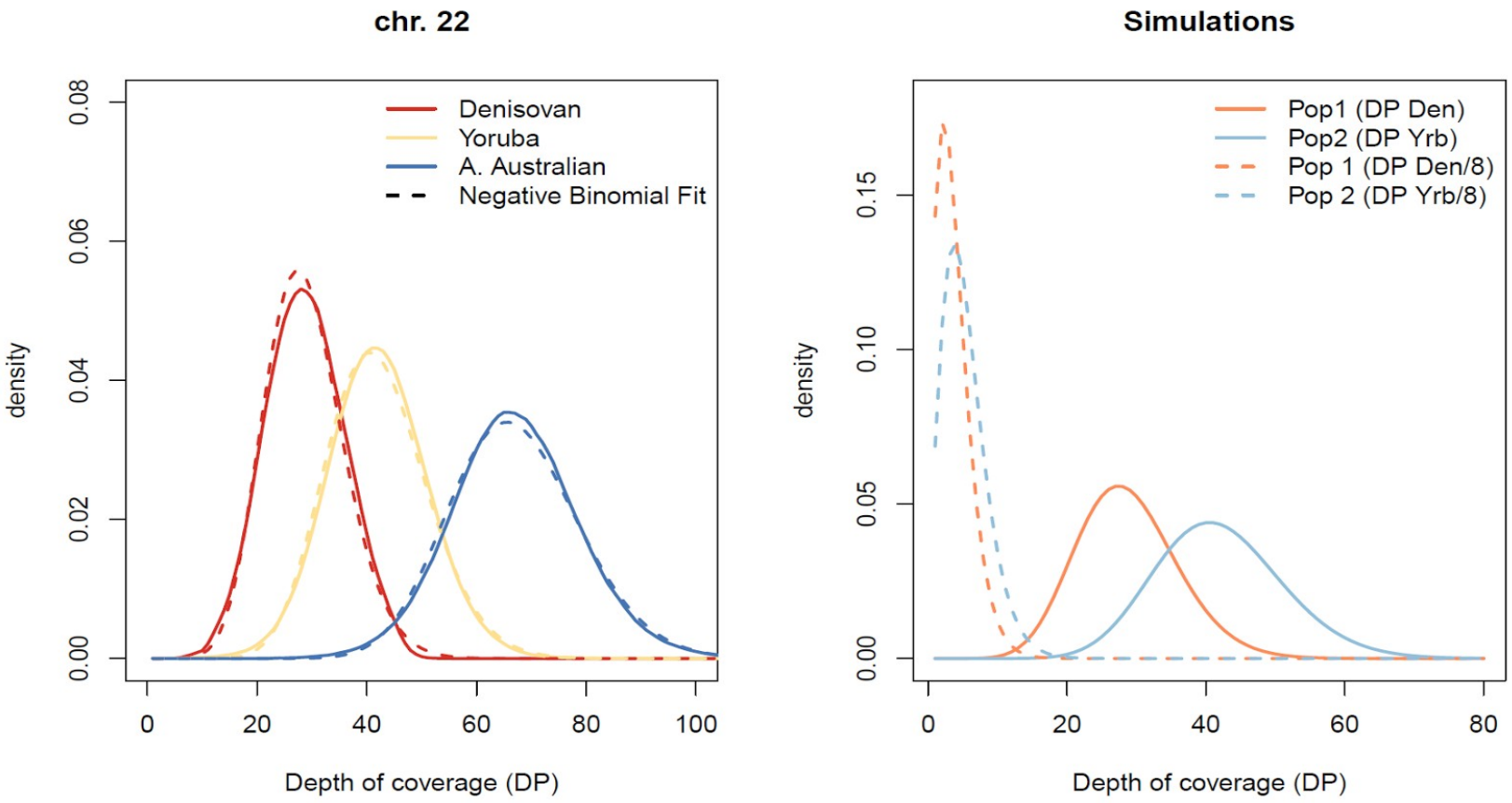
- Solution:
  - Find minimimum sample size
  - Resample without replacement

	Freq. derived	Sample size	Rel. freq
SNP1	1	16	1/16
SNP2	6	12	1/2
SNP3	1	8	1/12
SNP4	6	16	3/8



Gavel et al. (2014) PNAS

Effect of depth of coverage on SFS



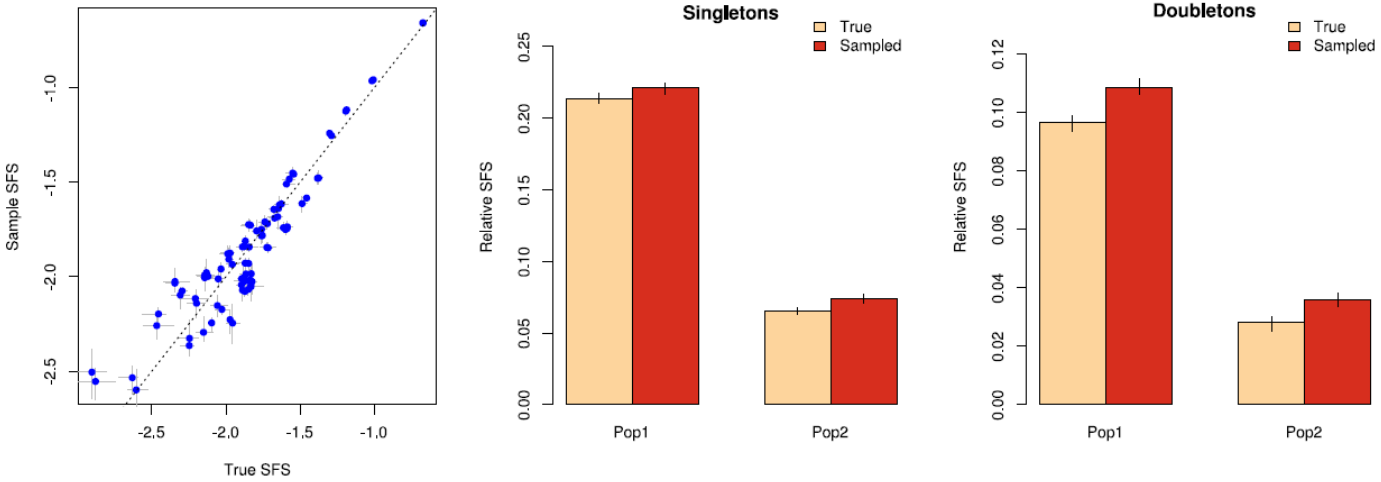
- Compared 2D SFS based on depth of coverage of observed data (mean larger than >20x), with a distribution 8 times smaller.

Malaspinas et al. (2016) Nature

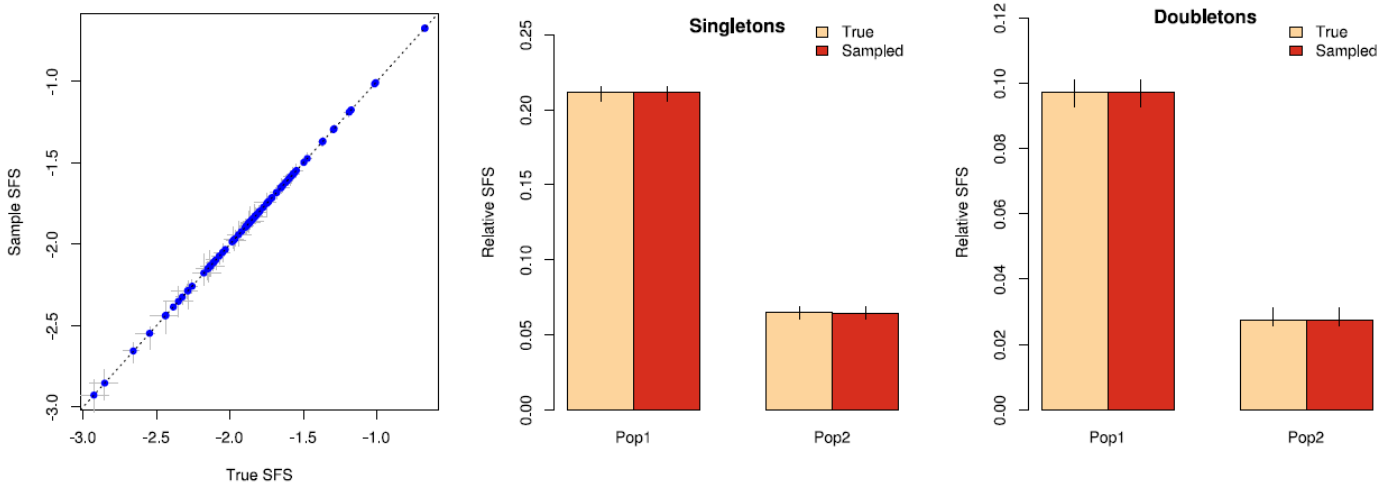
Effect of depth of coverage on SFS

Using low depth of coverage and allowing for missing data can lead to incorrect estimation of observed SFS

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



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



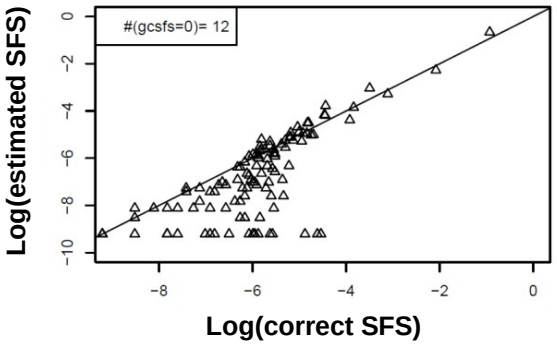
Malaspinas et al. (2016) Nature

Estimating SFS from observed data for NGS data accounting for genotype uncertainty

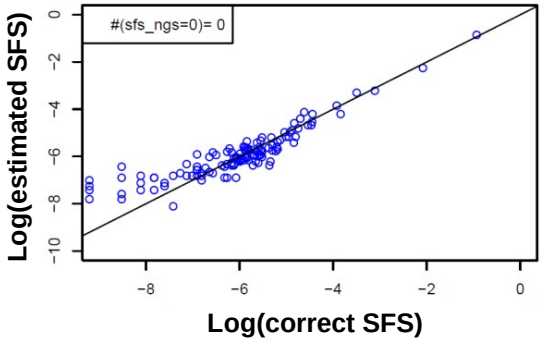
- For NGS data we can account for genotype uncertainty with the method implemented in the ANGSD and ngsTools programs
- INPUT: BEAGLE files

SFS with genotype uncertainty

- Integrating genotype uncertainty approximates well the true SFS, and better that based on filtered SNPs
- 2D SFS estimated using ngsTools (Fumagalli et al. 2014)



Filtering , keeping sites with missing data



ngsTools Genotype uncertainty and missing data

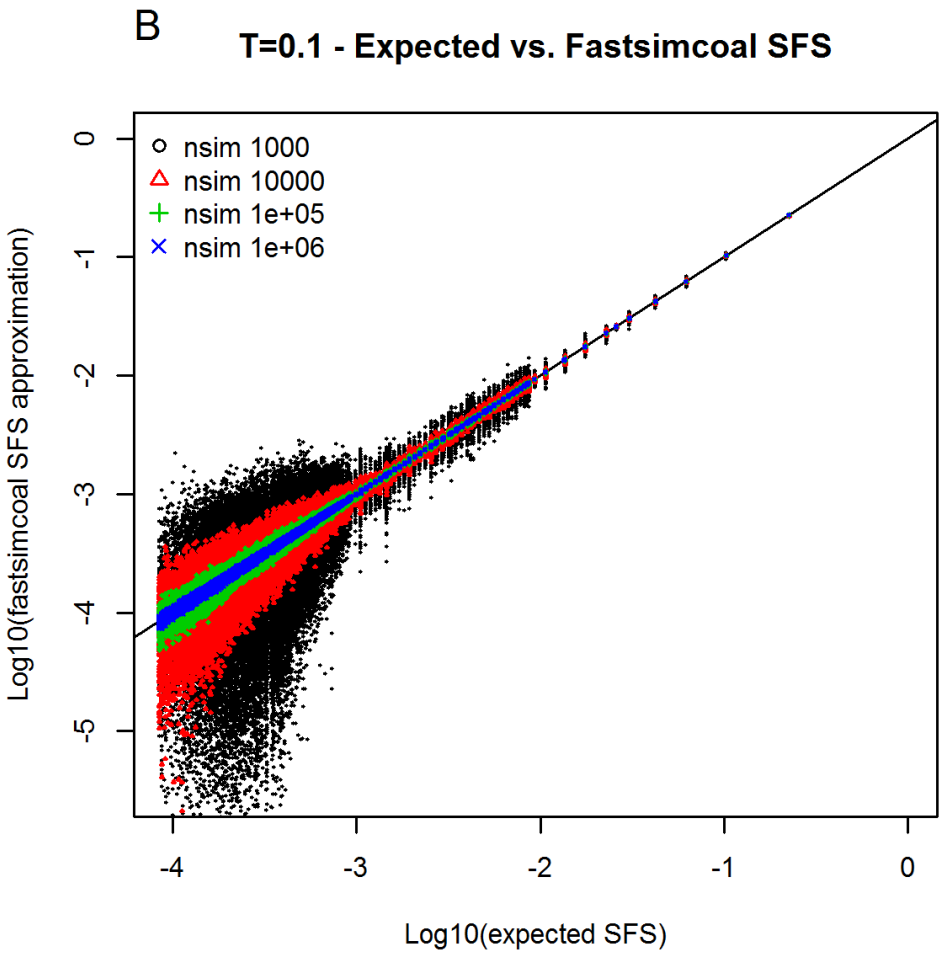
fastsimcoal2 program

- Fastsimcoal2 can estimate parameters from coalescent simulations
- Uses a conditional expectation (CEM) maximization algorithm to find maxCL parameters
- Large number of sims per point (>50000)
- Relatively fast and can explore wide and unbounded parameter ranges, as compared to dadi
- Can handle more than 3 populations
- For more than 4 populations, uses a composite composite likelihood

$CL_{1234...}=CL_{12}\times CL_{13}\times CL_{14}\times CL_{23}\times...)$

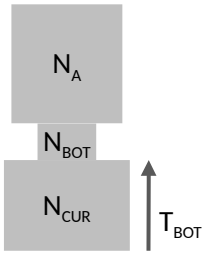
Expected SFS and the coalescent

- Fastsimcoal2 approximates the expected SFS by coalescent simulations
- Increasing the number of simulations improves the approximation of the expected SFS

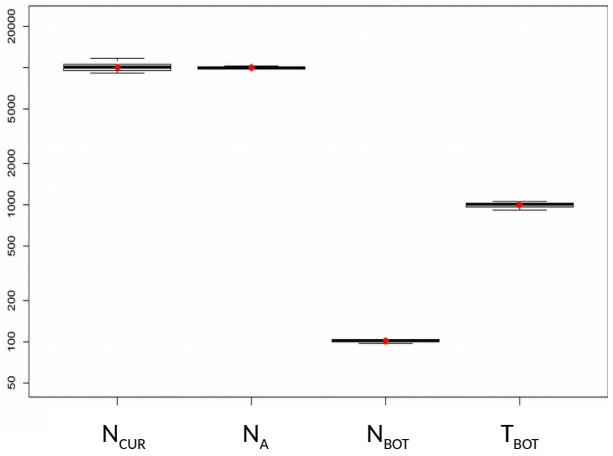


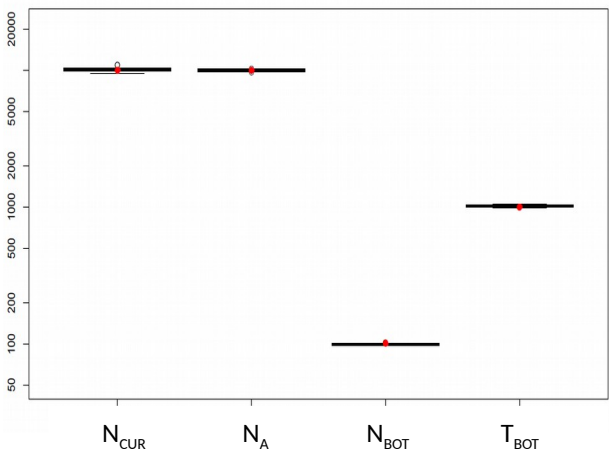
Excoffier et al. (2013) PloS Genetics; Chen (2012) TPB

Comparisons of approaches

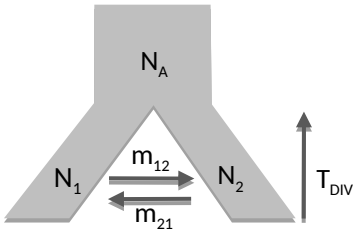


Simulation of 20 Mb data



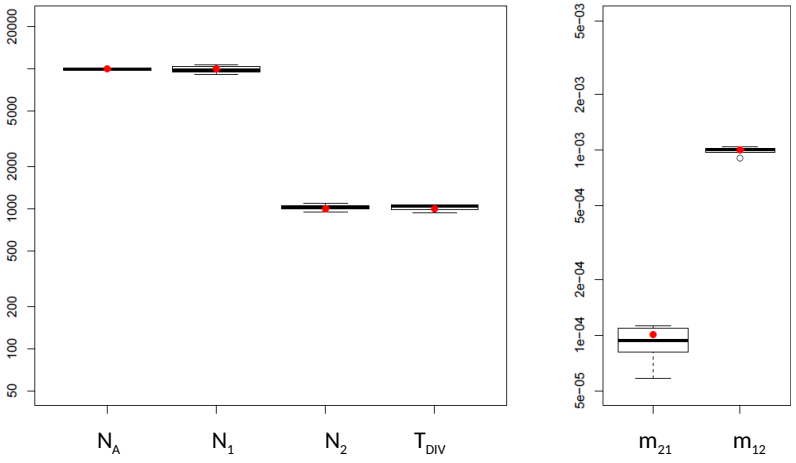


Comparisons of approaches

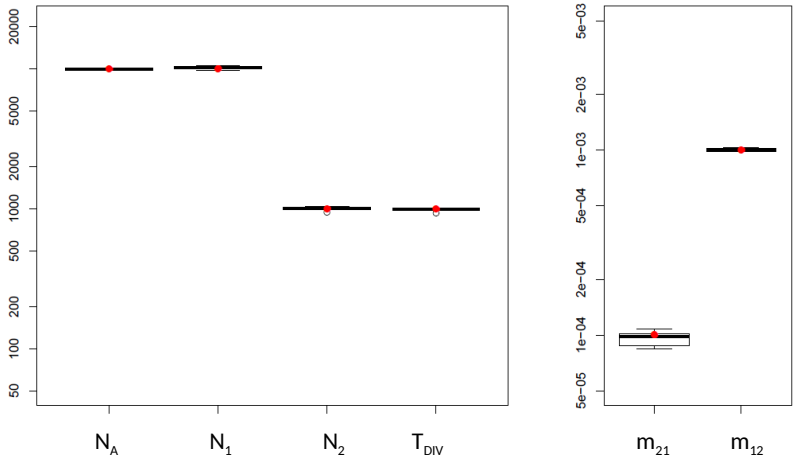


Simulation of 20 Mb data

fastsimcoal2



∂a∂i



FASTSIMCOAL2 INPUT FILES

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

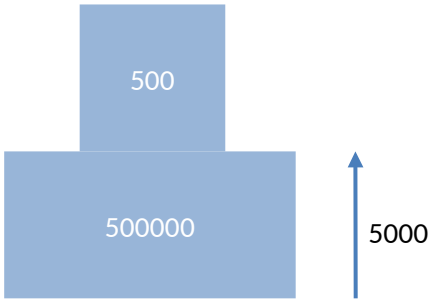
	d0_0	d0_1	d0_2	d0_3	d0_4	d0_5
d1_0	19985747	8350	1628	360	62	8
d1_1	966	0	0	0	0	0
d1_2	479	0	0	0	0	0
d1_3	328	0	0	0	0	0
d1_4	249	0	0	0	0	0
d1_5	1760	13	18	13	19	0

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	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

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 PAR files

Number of samples to simulate

Deme sizes (2N)

Sample sizes

Growth rates

Migration matrices

Historical events

No. of independent loci to simulate

No. of data blocks to simulate

Definition of genetic data type to simulate

2PopDivMigr10Loci.par

```
//Parameters for the coalescence simulation program : fsimcoal2.exe
2 samples to simulate :
//Population effective sizes (number of genes)
20000
1000
//Samples sizes and samples age
5
5
//Growth rates: negative growth implies population expansion
0
0
//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 0.005 0 1
//Number of independent loci [chromosome]
10 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 recomb. and mut. rates and optional
parameters
DNA 1000 0 2.5e-8 0.33
```

Here we simulate 10 recombining segments of 1000 bp DNA, in two populations of sizes 20000 and 1000 having diverged 5000 generations ago from a small population of size 100

TPL files

TPL are like PAR files, but the actual parameter values are replaced by 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

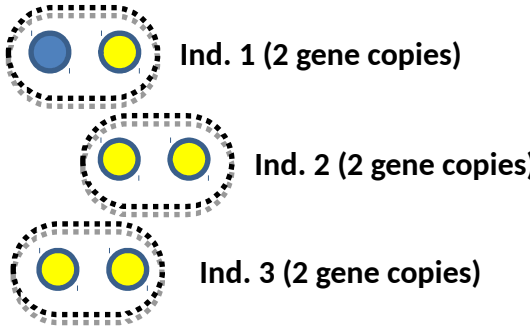
2PopDivMigr10Loci.par

```
//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

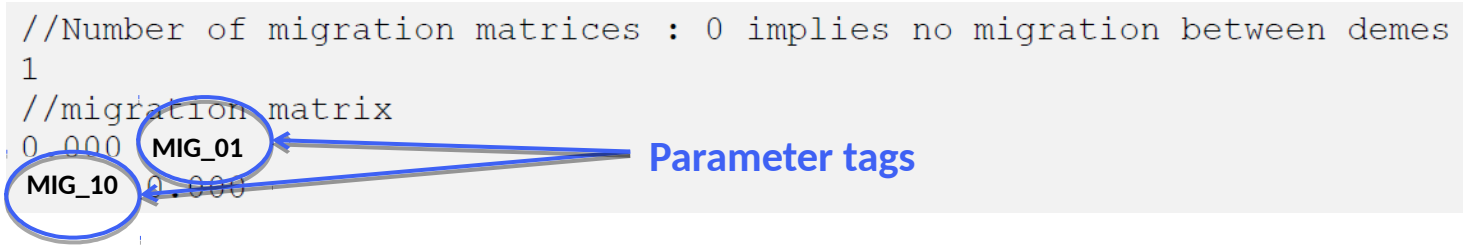
Population effective sizes are given in number of gene copies. For a diploid species with N=500 individuals, this corresponds to a 2N=1000 gene copies, as each individual carries two gene copies at any given site.

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



TPL files

MIGRATION



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.

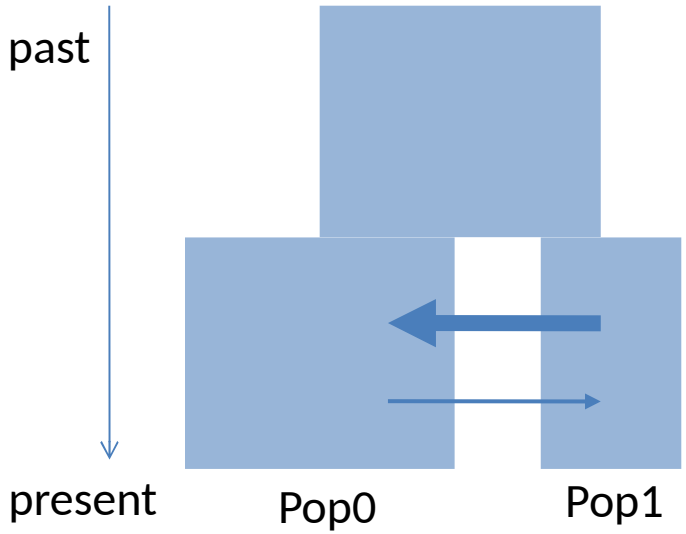
1PopStationary10Loci.par

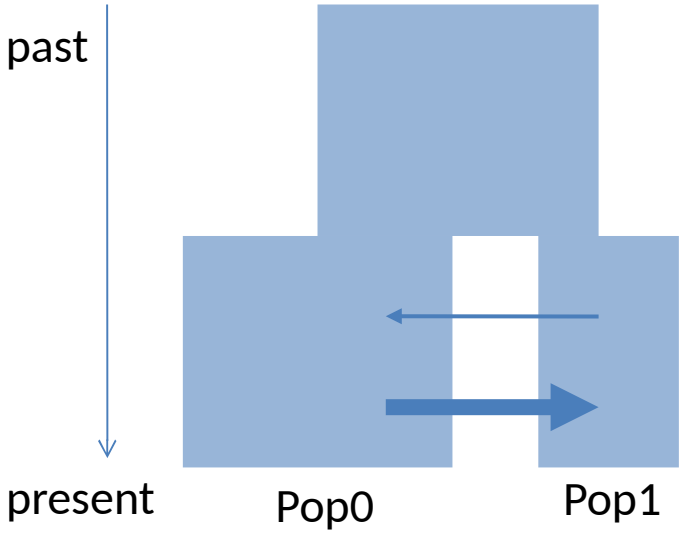
```
//Number of migration matrices : 0 implies no migration between demes
0
```

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
```



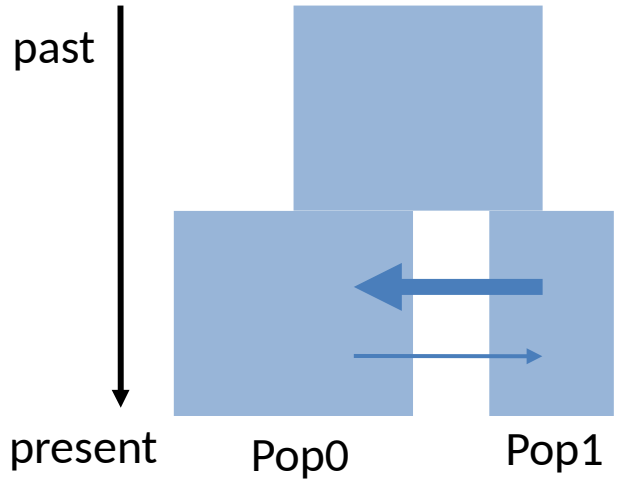


A note on looking backward in time

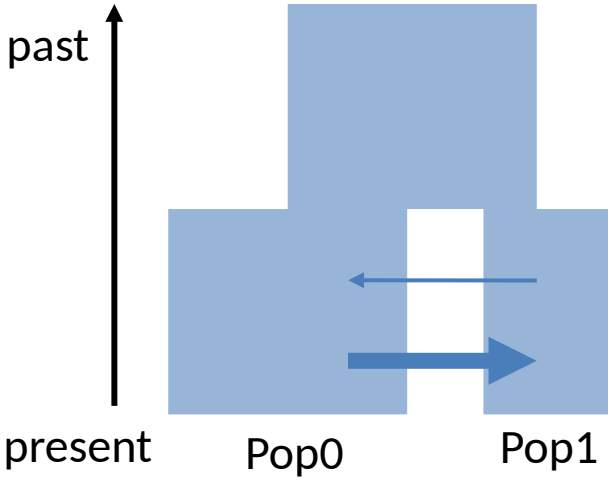
Assuming that we look forward in time and that the size of the arrows are proportion to the migration rate, to what model does the following migration matrix corresponds to?

```
//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!

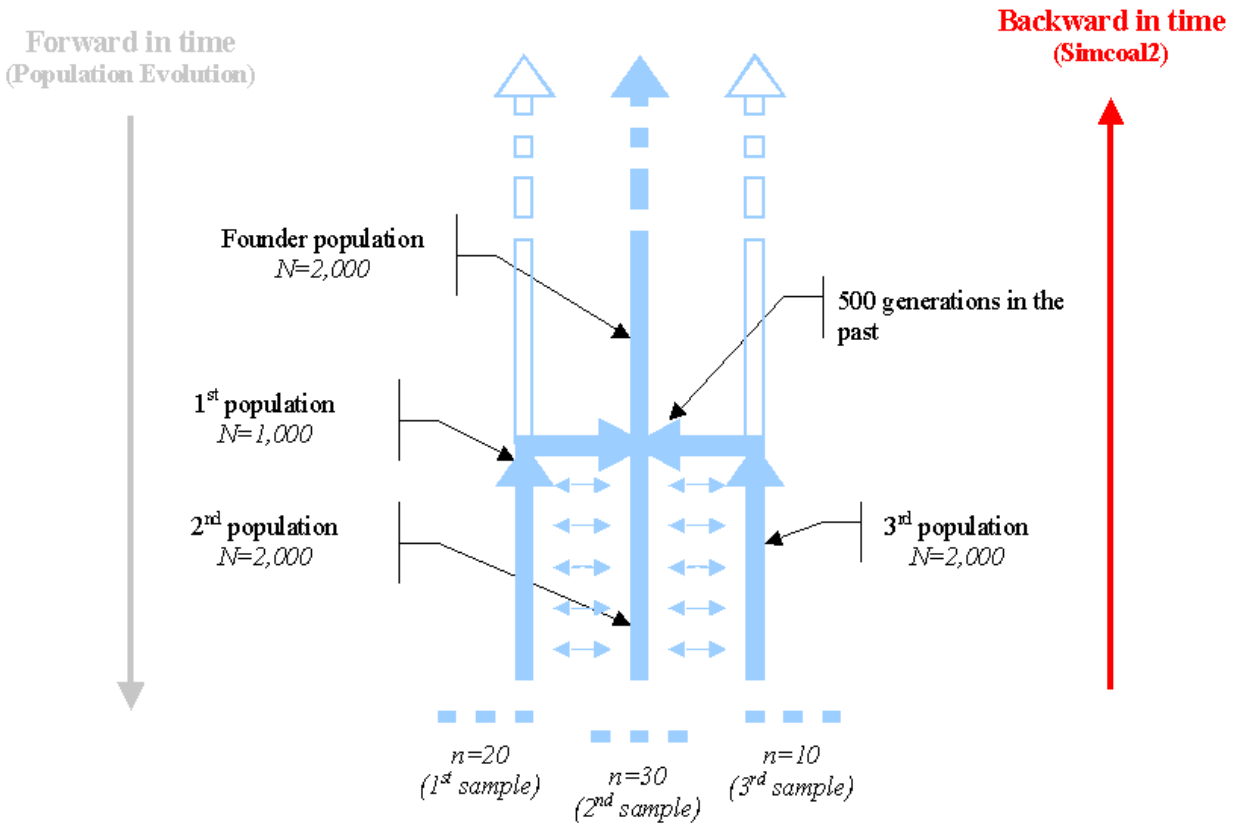
Historical events (backward in time)

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.



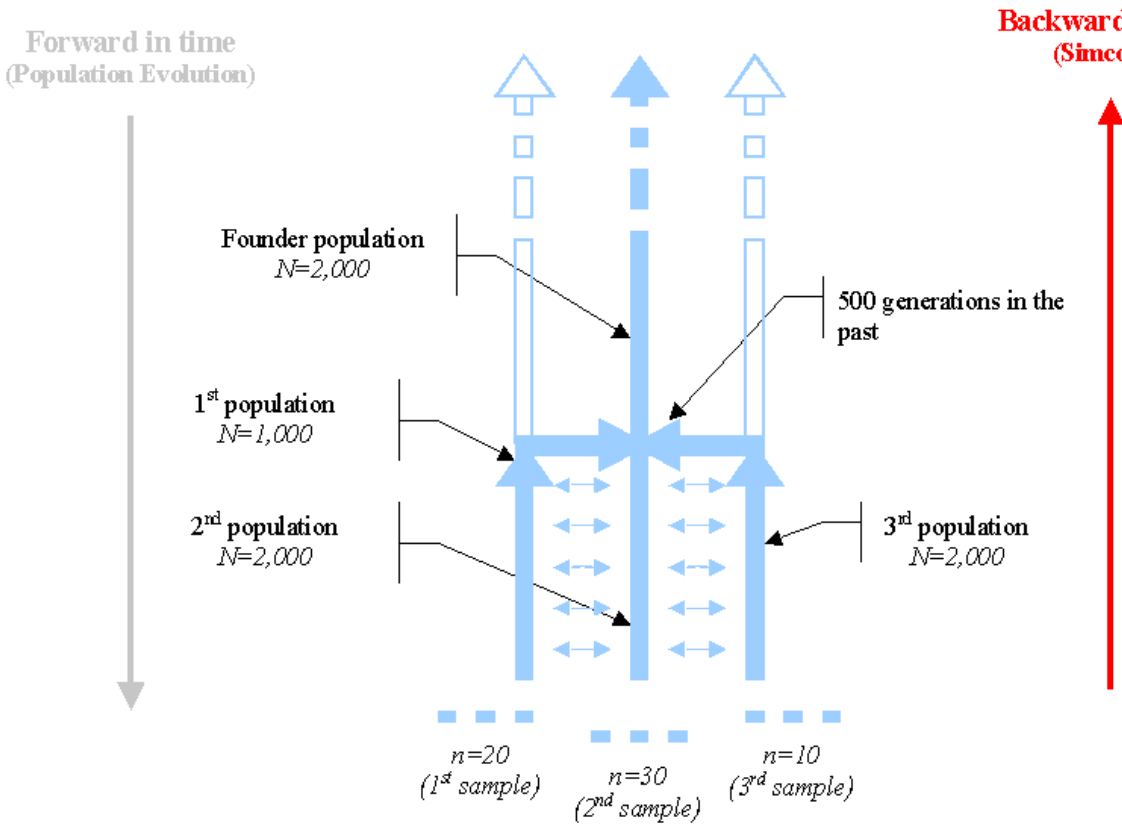
Historical events (backward in time)

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% 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. N2=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

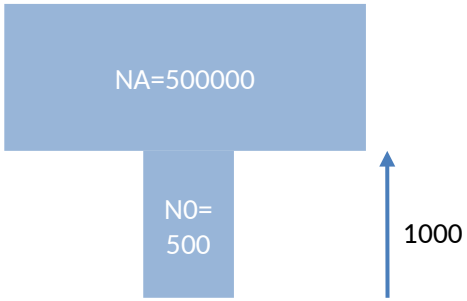
Change the size of a given population

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 N0=500 diploids at time zero, it implies that NA=500000 diploids.
- The migration matrix valid after the event is the migration rate 0. Since it is not defined it imples no migration.

Recent instantaneous demographic contraction



1PopContrInst10loci.par

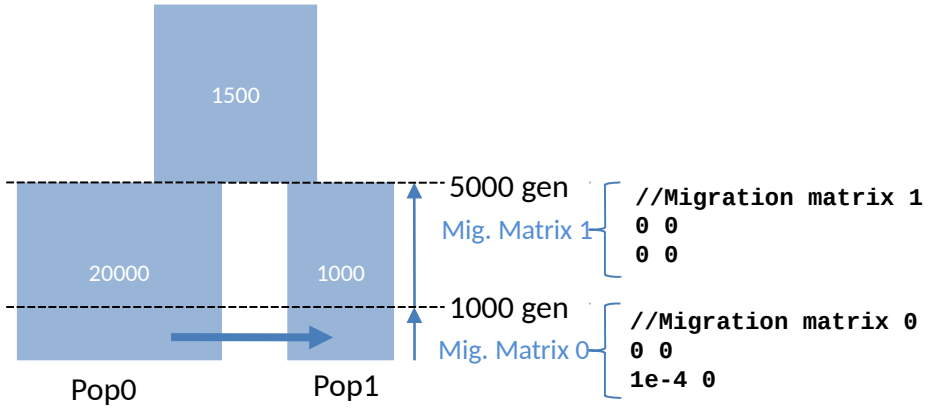
Historical events in fastsimcoal2

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 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.



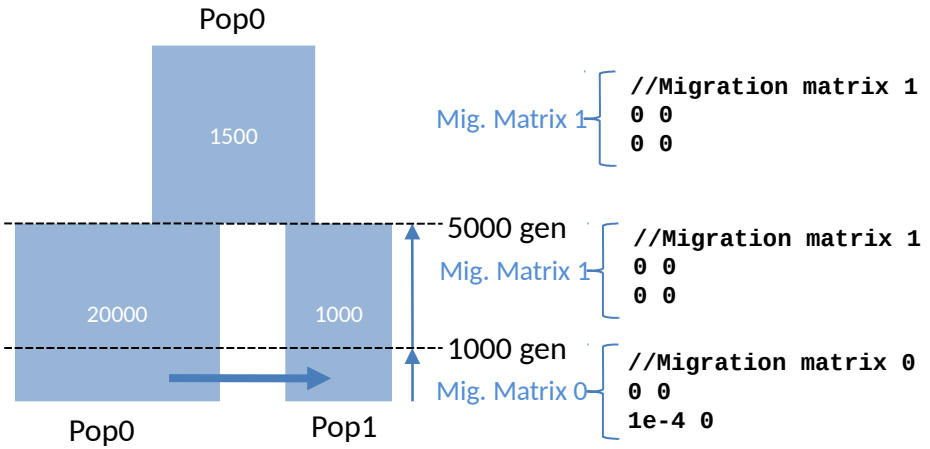
Historical events in fastsimcoal2

Population split (merge populations going backwards in time)

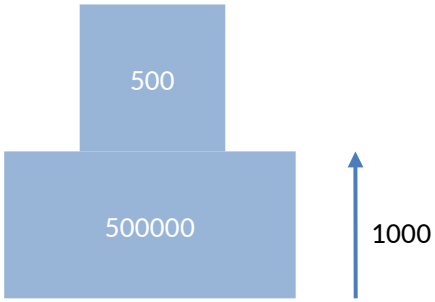
2PopDivMigr10Loci.par

```
//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



Command line to estimate parameters

```
fsc22.exe -t 1PopExpInst20Mb.tpl -e 1PopExpInst20Mb.est -M0.001 -d -n100000 -N100000 -l5 -L20 -q -c4
```

Template file name

Estimation file name

No. of threads

Do ML estimation
And stop criterion

Use derived SFS

Minimum (-n) and maximum (-N) number of simulations to estimate the expected SFS

Minimum (-l) and maximum (-L) number of iterations to perform

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

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 (see later)
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

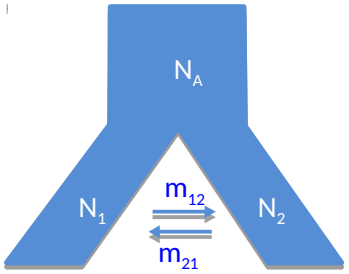
Protocol for parameter estimation

- **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
- Define the **demographic model**
- **Estimate the parameters** – repeat 50-100 runs, and selecting the run with maximum likelihood
- **Bootstrap** to obtain confidence intervals for each parameter – bootstrap 10-100 datasets, by repeating a few runs for each dataset

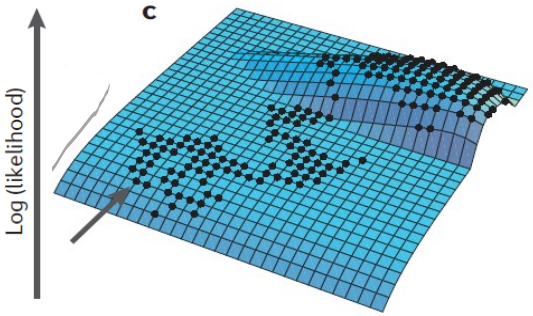
Protocol for model comparison based on AIC

- Get the observed SFS (see previous slide)
- 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 critera) for each model
- Select the model with minimum AIC
- This is only correct for datasets with unlinked (independent) SNPs

Other Full likelihood approaches



- E.g. IMa (Isolation with Migration)
  - Uses a MCMC algorithm to explore the likelihood surface and get posterior distributions of parameters
  - Accurate, but quite slow and restricted to few loci (dozens to 100's)
  - Needs some tuning and multiple runs for checking for convergence
  - Model validation is tedious, rarely done
  - Restricted to this particular model, but now extended to more than 2 populations (slower)



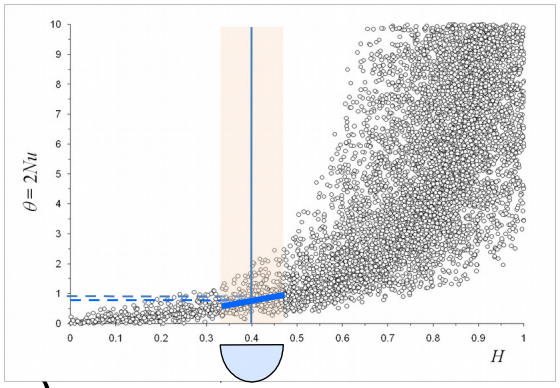
Hey and Nielsen (2007); Hey (2010)

Approximate Bayesian Computations (ABC)

- Data are replaced by summary statistics

$$\Pr(\theta \mid \text{Data}) \rightarrow \Pr(\theta \mid S)$$

- Simulations are used to produce joint distributions of S and parameters, and simulations leading to statistics close to observed ones are retained for the estimations



✓Obtains posterior distributions

✓Applicable to complex models (sometimes too complex)

✓Model validation is easy, routinely done

✓Easy to parallelize on a cluster

✘Less powerful than full likelihood methods

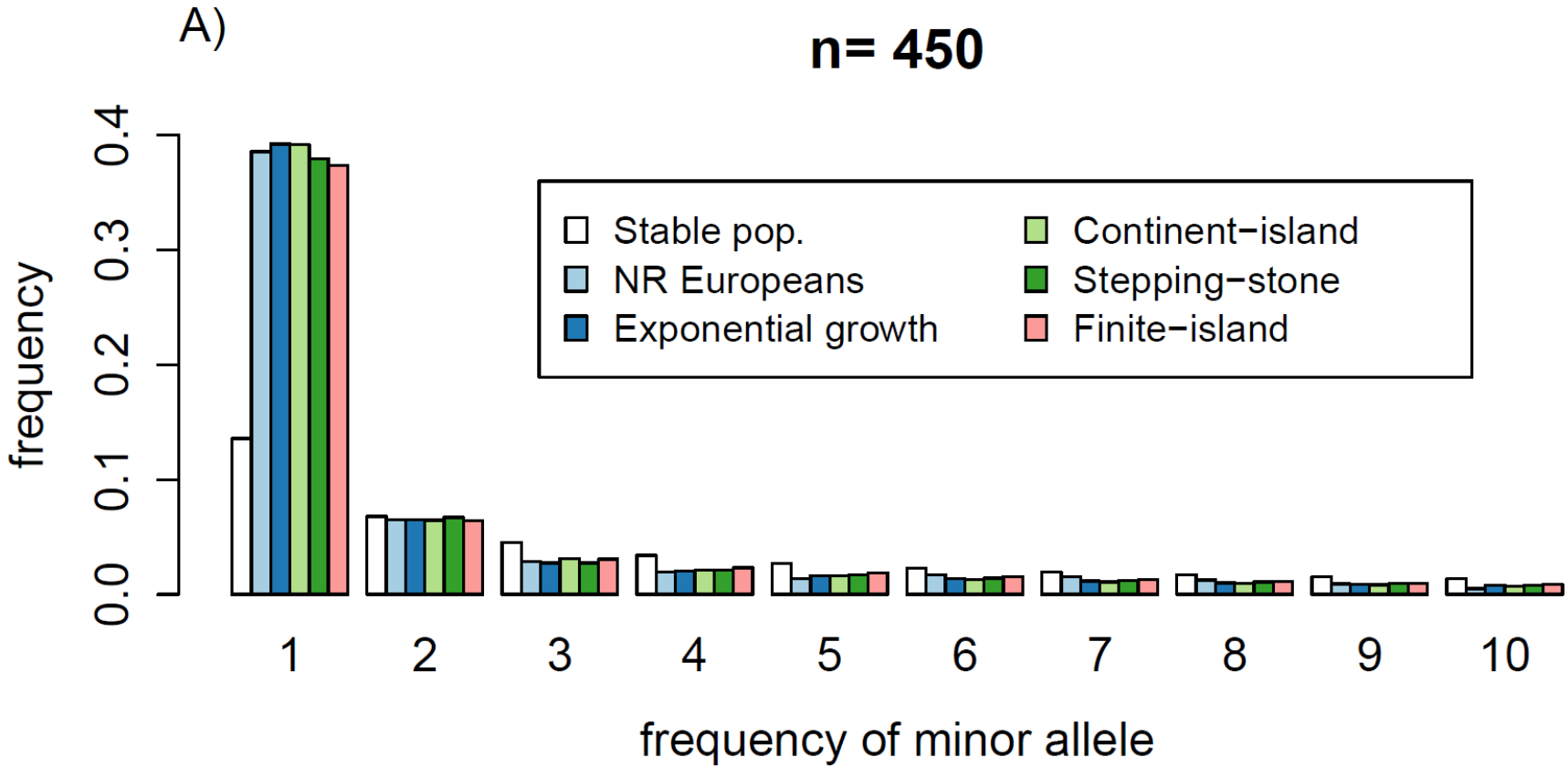
✘Choice of priors needs to be carefully done

✘Choice of statistics is difficult, tuning phase is necessary

✘Computationally still demanding

Population structure also affects the SFS

- Different structured population models can reproduce the excess of singletons observed in European human populations (NR Europeans)



Sousa et al. (2014) Curr. Opin. Gen. Dev.

Sampling also affects the SFS

- Continent-island model (100 islands) varying the sampling scheme:
  - 1 deme sampled (450 diploids)
  - 20 demes sampled evenly (45 genes copies per deme)
  - 20 demes sampled unevenly (431 diploids from 1 deme, 1 diploid from each remaining demes)
- **Different conclusions would be reached if sampling was ignored**, although data was generated under exactly the same model