Demographic inference Lab activity

Workshop on Genomics 2022

Emiliano Trucchi
Josephine Paris

Demographic inference

Understanding population demography is key in any population genomics analysis

- For understanding effects of mutation-drift equilibrium
- Disentangling the impact of effective population size (N_e)
- Accounting for false positive detection of selection

The following exercises can be applied to population genomics analyses obtained from:

- Whole genome sequencing (including low coverage)
- Restriction-site associated DNA sequencing (RAD-seq)
- SNP panels (high density, but not exome SNP chips!)

NB: Demographic inference is sensitive to missing data, mispolarization of ancestral and derived alleles, not-neutrally evolving genomic regions, "not-neutrally" recombining genomic regions (e.g., transposable and repetitive elements)

Measures of genetic diversity

There are many(!), like:

1) **Nucleotide diversity (π):** average number of single nucleotide differences between haplotypes chosen at random from a sample

2) **Number of segregating sites:** total number of sites that are polymorphic (segregating) in our sample. It can be standardized for the number of individuals sequenced (θ_W)

But today, we will focus on:
The site frequency spectrum: a summary of genome-wide data that describes the number of sites with a given frequency in our sample.

Measures of genetic diversity

The site frequency spectrum: a summary of genome-wide data that describes the number of sites with a given frequency in our sample.

Counts of the number of loci where an allele is found *i* times out of n

We call alleles that are found once in a sample ***singleton***, alleles that are found twice in a sample ***doubleton***, and so on, and this is called the site frequency spectrum.

We often calculate the minor allele frequency spectrum, or the frequency spectrum of derived alleles.

Let’s focus on the SFS

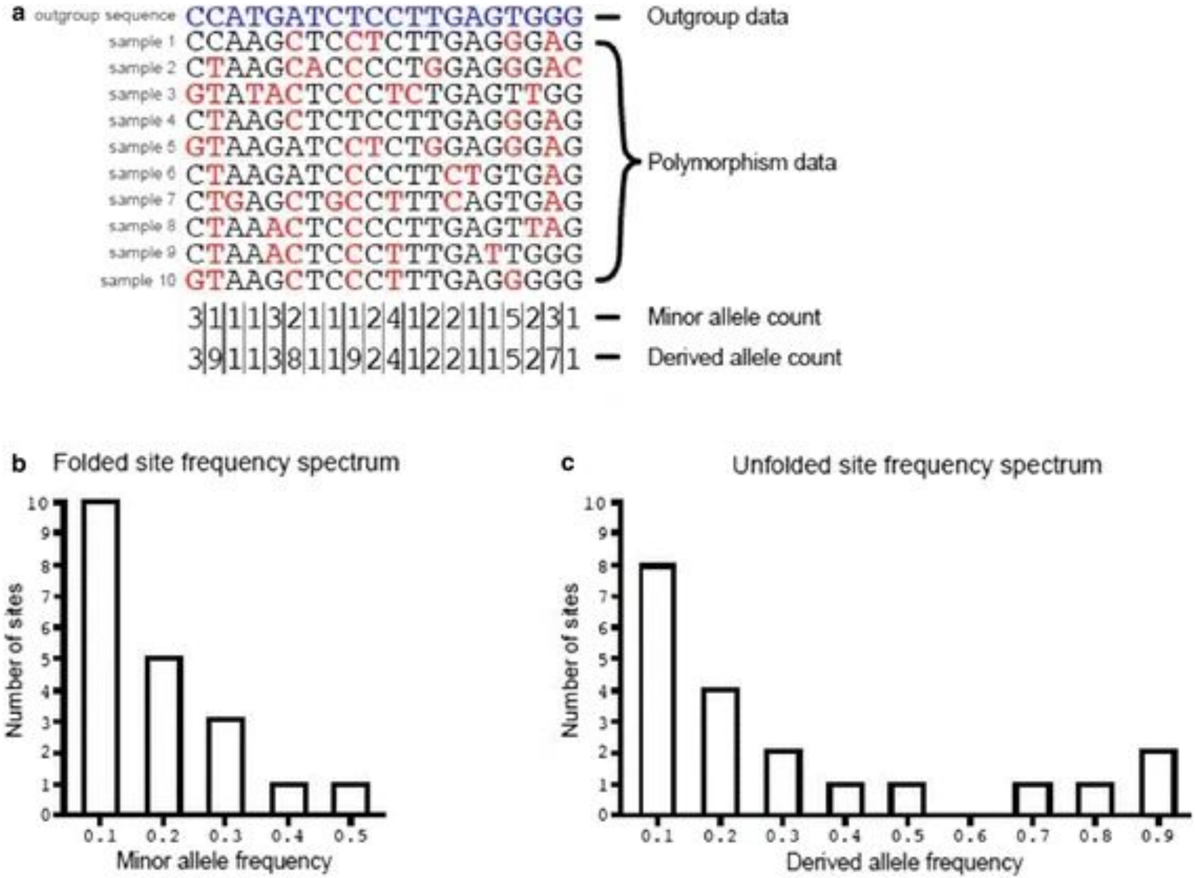
How to make it

Counts of the number of loci where an allele is found *i* times out of n copies

Mmhh, an example ?

The Site Frequency Spectrum

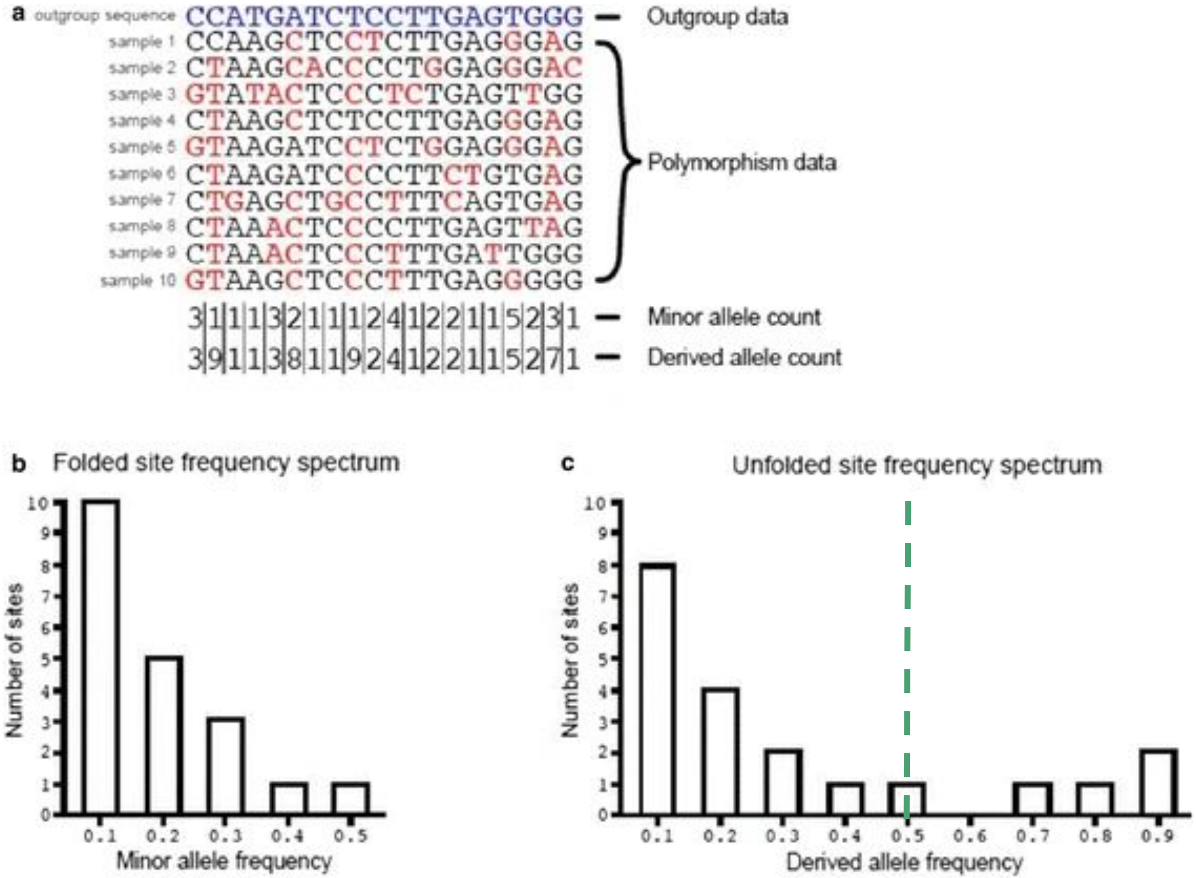
Mmhh, an example...



Booker, 2017. *BMC biology*, *15*(1), 98.

The Site Frequency Spectrum

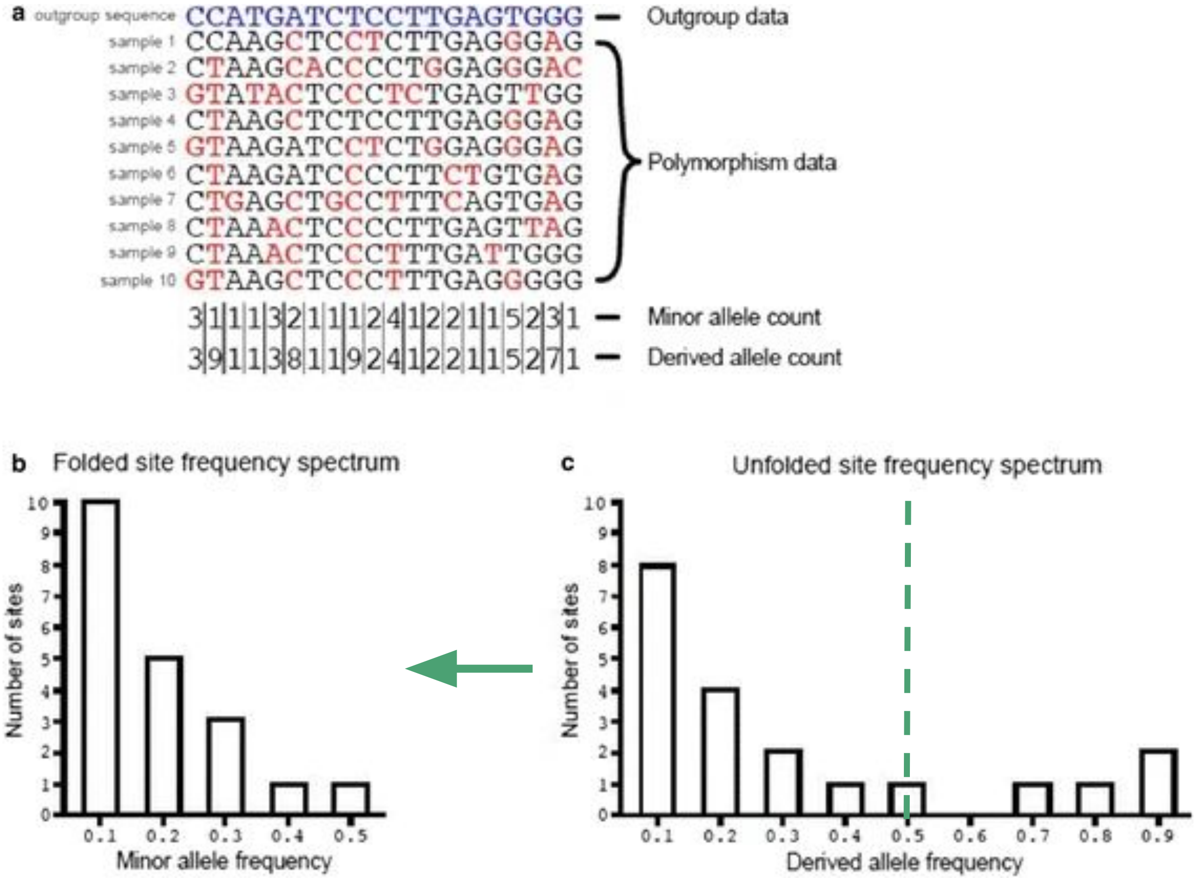
Mmhh, an example...



Booker, 2017. *BMC biology*, *15*(1), 98.

The Site Frequency Spectrum

Mmhh, an example...

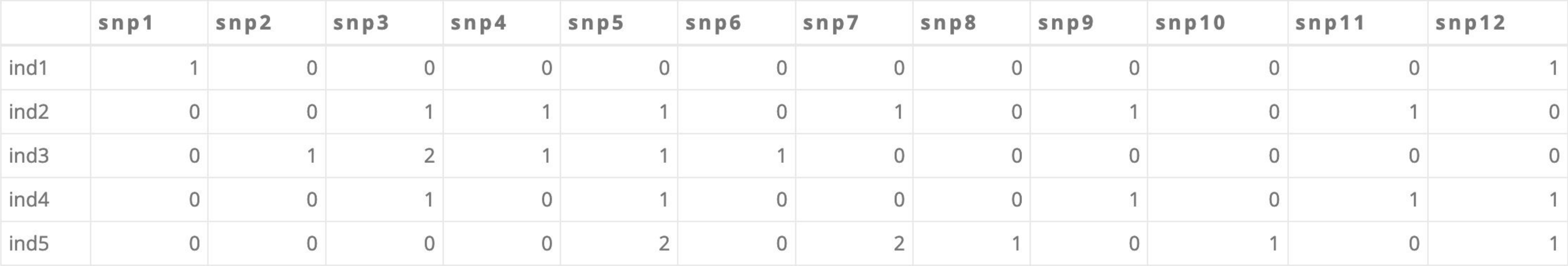


Booker, 2017. *BMC biology*, *15*(1), 98.

The Site Frequency Spectrum

In order to better understand the SFS, we're going to calculate one by hand!

Consider the following genotype data where each entry encodes the genotype as 0 (homozygote for reference allele), 1 (heterozygote), 2 (homozygote for alternative allele). In the table each column corresponds to a site, and each row corresponds to an individual. Not too different from a variant call format (VCF) file.

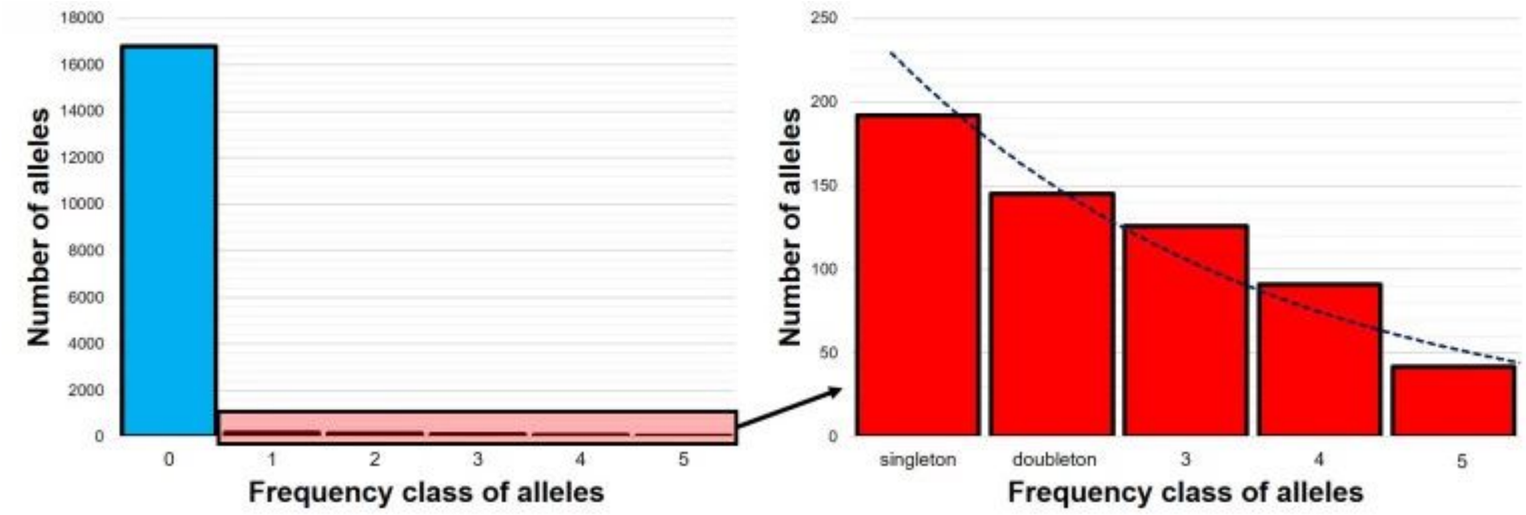


	snp1	snp2	snp3	snp4	snp5	snp6	snp7	snp8	snp9	snp10	snp11	snp12
ind1	1	0	0	0	0	0	0	0	0	0	0	1
ind2	0	0	1	1	1	0	1	0	1	0	1	0
ind3	0	1	2	1	1	1	0	0	0	0	0	0
ind4	0	0	1	0	1	0	0	0	1	0	1	1
ind5	0	0	0	0	2	0	2	1	0	1	0	1

First, you need to compute the absolute allele frequency for each SNP.
Second, you need to count the number of sites with a given allele frequency bin
HINT: think about what your y and x axis will be!

The shape of the SFS depends on the population history

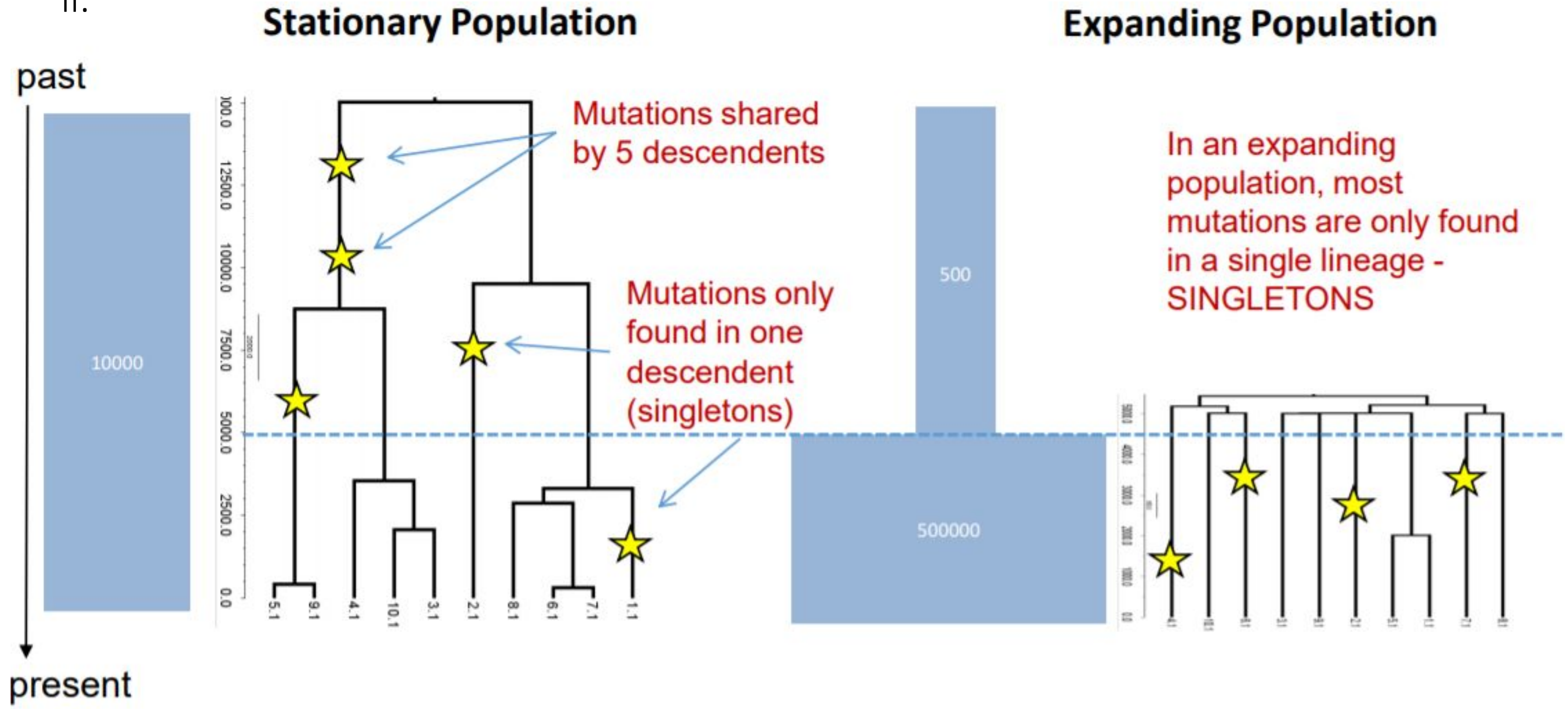
If population size is constant, **$E(S_i) = \theta/i$** (Fu and Li 1993)
Expected exponential distribution with many more rare than common alleles.
We can use the **coalescent theory** to infer the demography of a population. How??



https://theg-cat.com/tag/site-frequency-spectrum/

We expect more unique mutations in expanding population

- ✔ Mutations accumulate along the branches.
- ✔
- ✔ The longer a given branch the more likely it becomes that a mutation have happened on it.



Stationary Population

Expanding Population

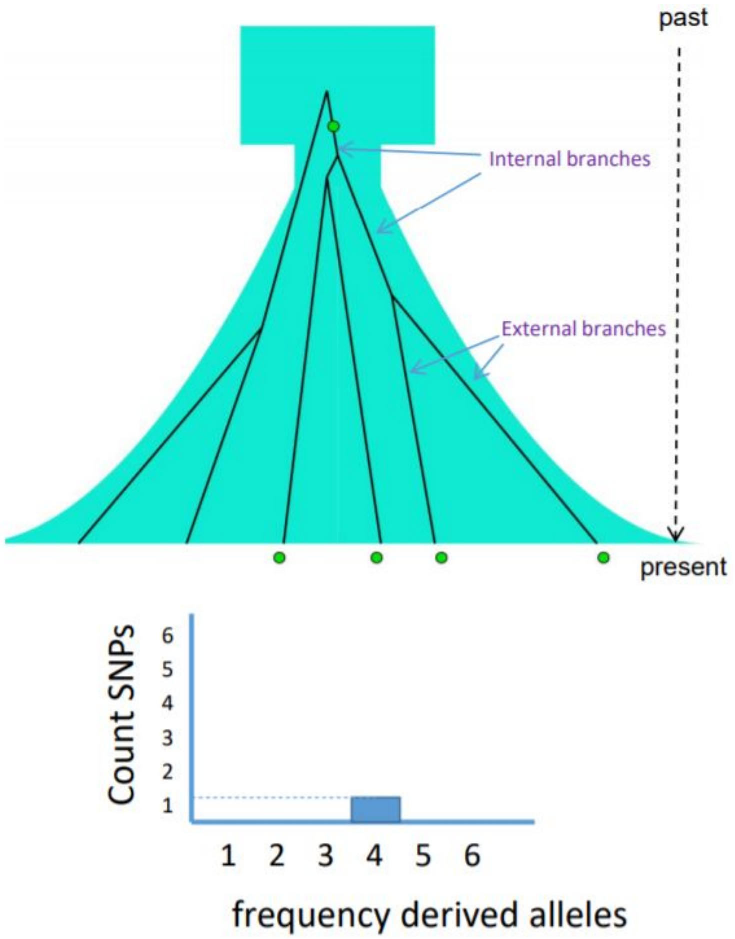
Mutations shared by 5 descendents

Mutations only found in one descendent (singletons)

In an expanding population, most mutations are only found in a single lineage - SINGLETONS

Coalescent and the SFS – Expanding population

- ✔ 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

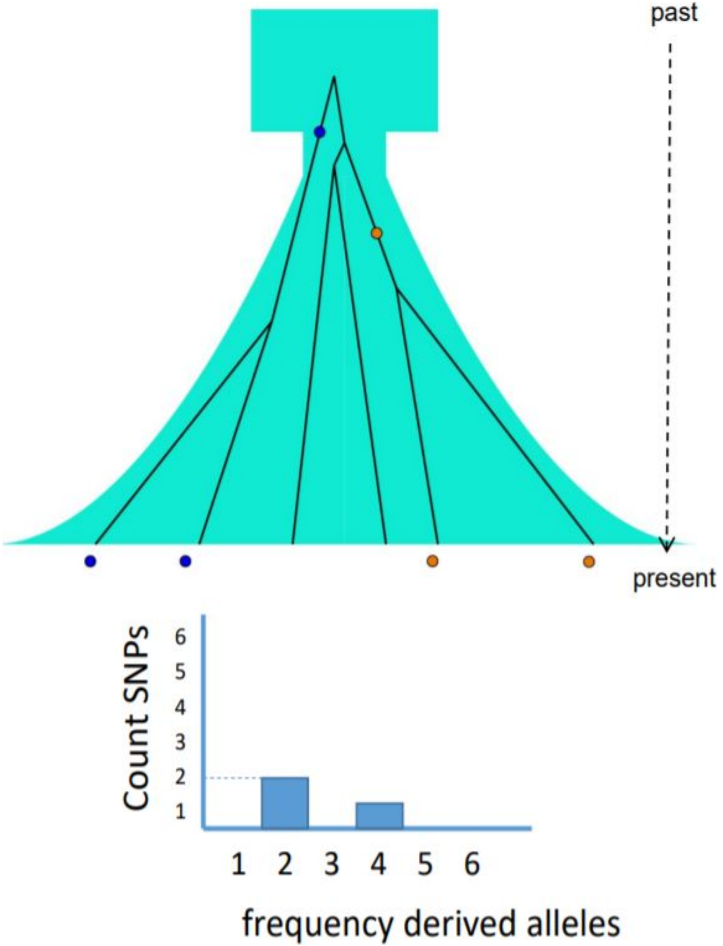


Coalescent and the SFS – Expanding population

✔ A recent population growth following a bottleneck leads to gene trees with long external branches

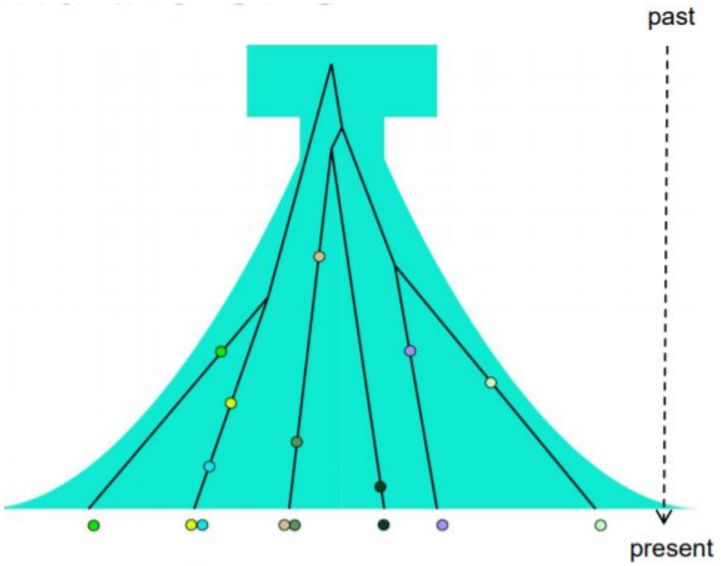
✔ Very few mutations in the internal branches

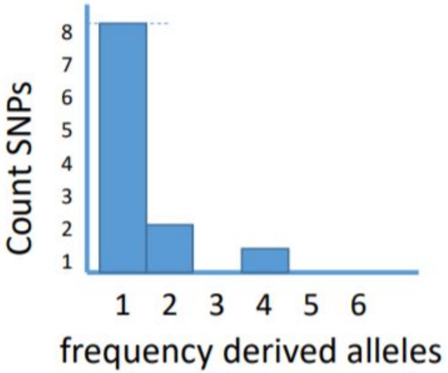
✔ Most mutations in long external branches are only found in one lineage, resulting in an excess of singletons



Coalescent and the SFS – Expanding population

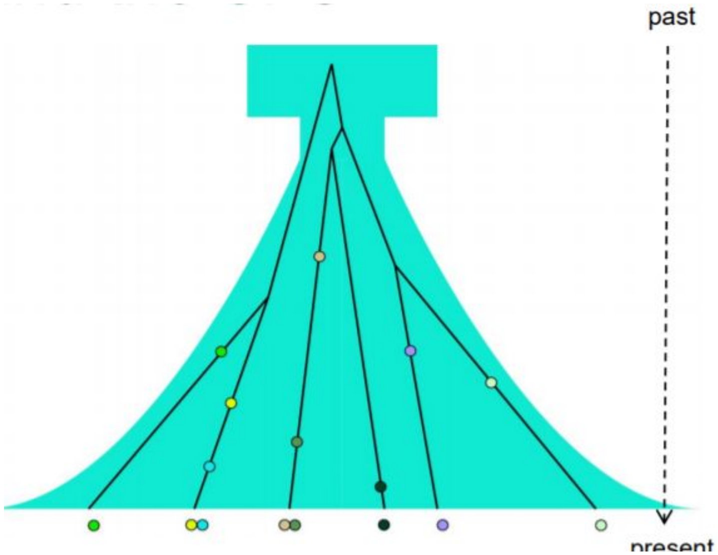
- ✔ A recent population growth following a bottleneck leads to gene trees with long external branches
- ✔ Very few mutations in the internal branches
- ✔ Most mutations in long external branches are only found in one lineage, resulting in an excess of singletons

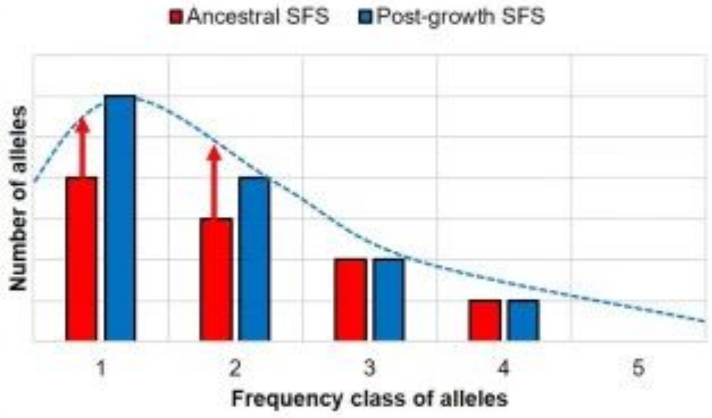




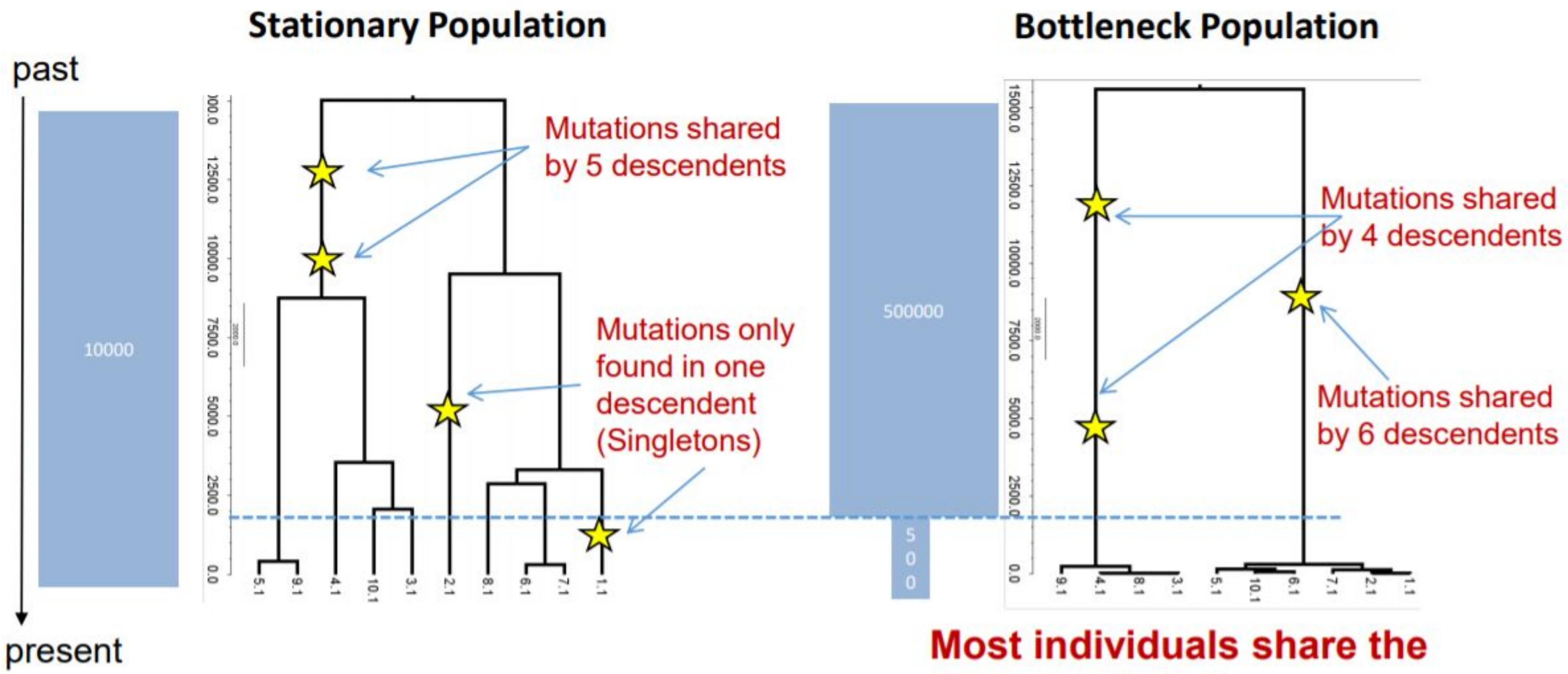
Coalescent and the SFS – Expanding population

- ✔ 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





We expect less diversity in a bottlenecked population


Stationary Population

Bottleneck Population

Mutations shared by 5 descendents

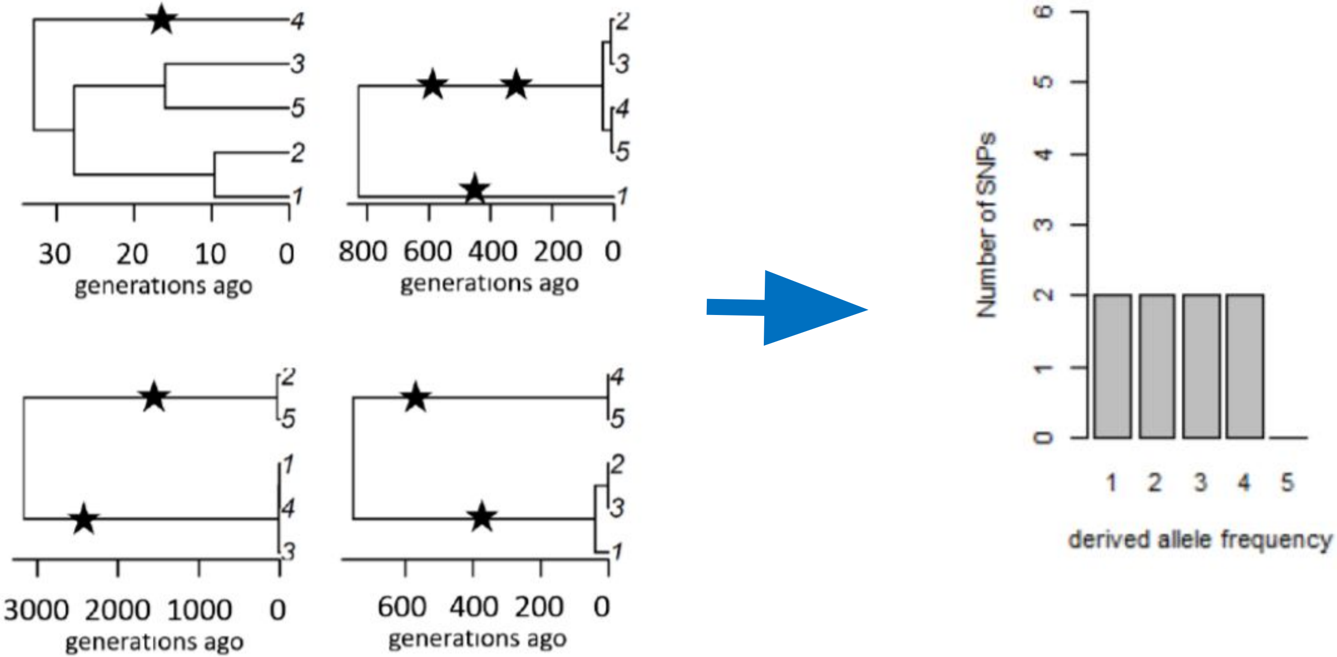
Mutations only found in one descendent (Singletons)

Mutations shared by 4 descendents

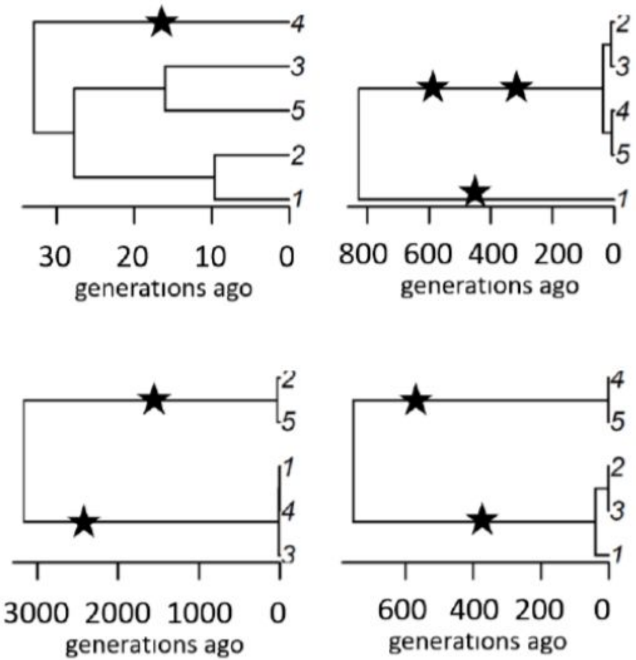
Mutations shared by 6 descendents

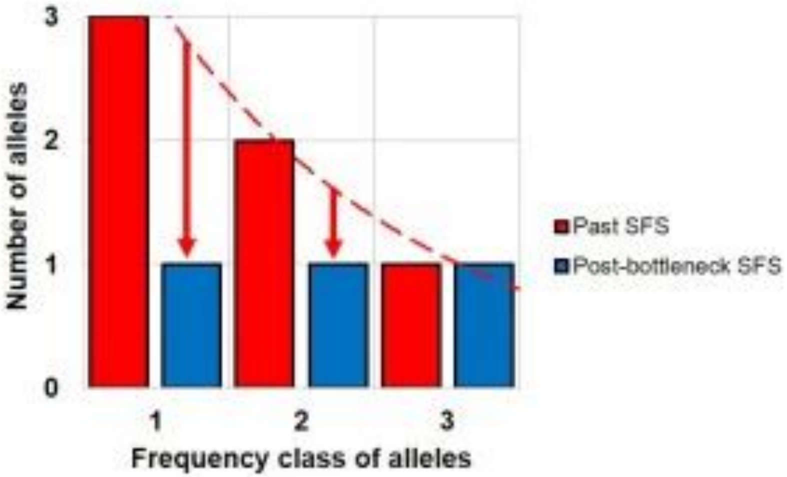
Most individuals share the same mutations

Coalescent and the SFS – Bottlenecked population

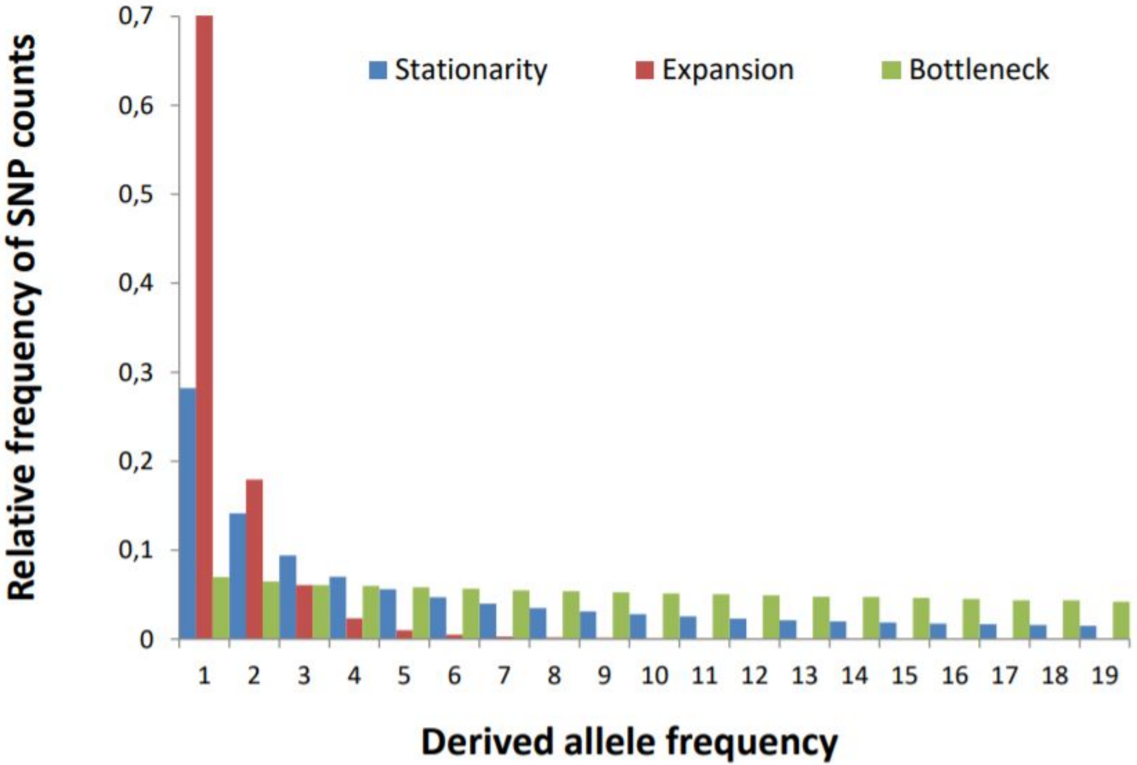


Coalescent and the SFS – Bottlenecked population





SFS depends on past demography



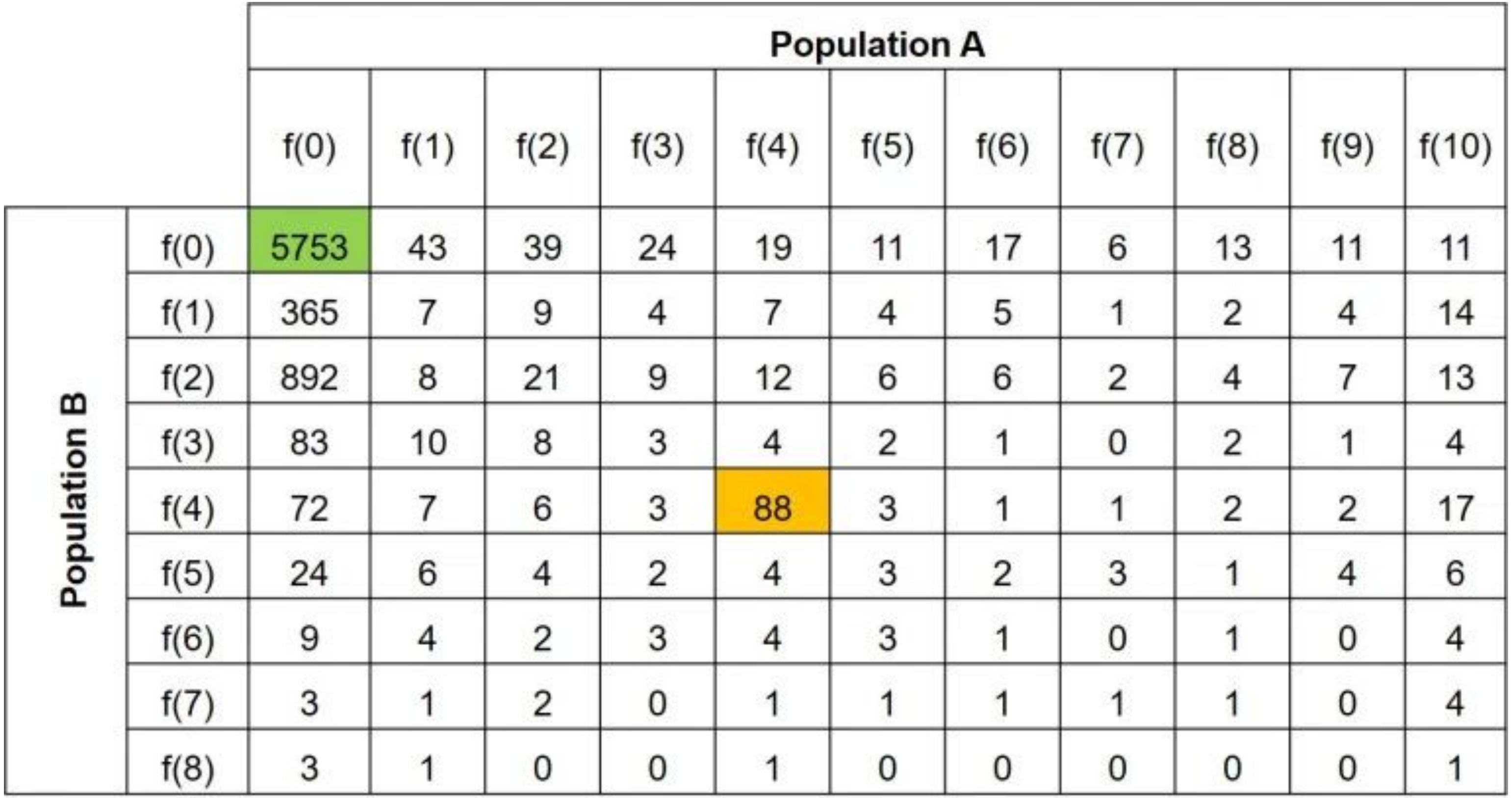
The Site Frequency Spectrum

Folded and unfolded SFS: if the ancestral/derived state is known, a derived (unfolded or polarized) allele frequency (DAF) spectrum can be drawn

Single (1-dimensional) and joint (2...-dimensional) SFS: whether one or more populations are included in the SFS

Further reading: https://theg-cat.com/tag/site-frequency-spectrum/

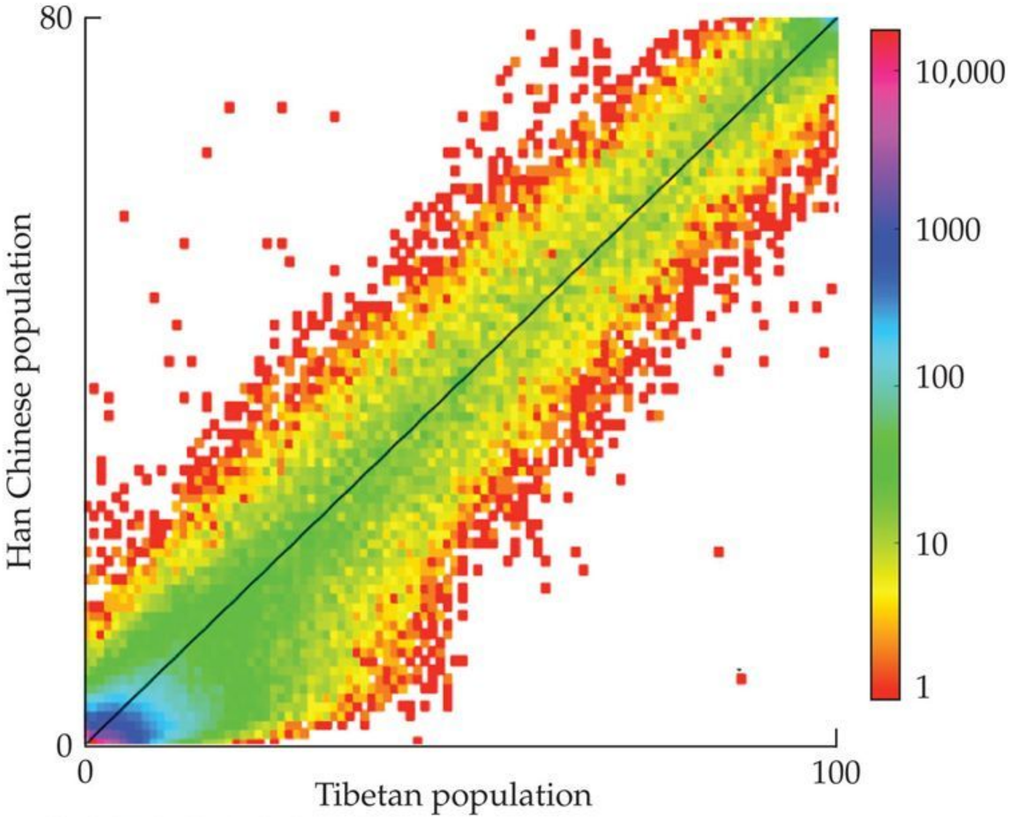
2-D SFS



		Population A										
		f(0)	f(1)	f(2)	f(3)	f(4)	f(5)	f(6)	f(7)	f(8)	f(9)	f(10)
Population B	f(0)	5753	43	39	24	19	11	17	6	13	11	11
	f(1)	365	7	9	4	7	4	5	1	2	4	14
	f(2)	892	8	21	9	12	6	6	2	4	7	13
	f(3)	83	10	8	3	4	2	1	0	2	1	4
	f(4)	72	7	6	3	88	3	1	1	2	2	17
	f(5)	24	6	4	2	4	3	2	3	1	4	6
	f(6)	9	4	2	3	4	3	1	0	1	0	4
	f(7)	3	1	2	0	1	1	1	1	1	0	4
	f(8)	3	1	0	0	1	0	0	0	0	0	1

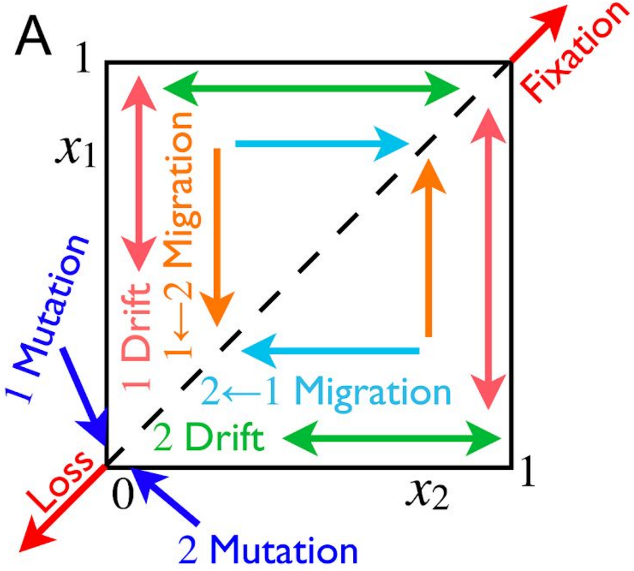
https://theg-cat.com/tag/site-frequency-spectrum/

2-D SFS



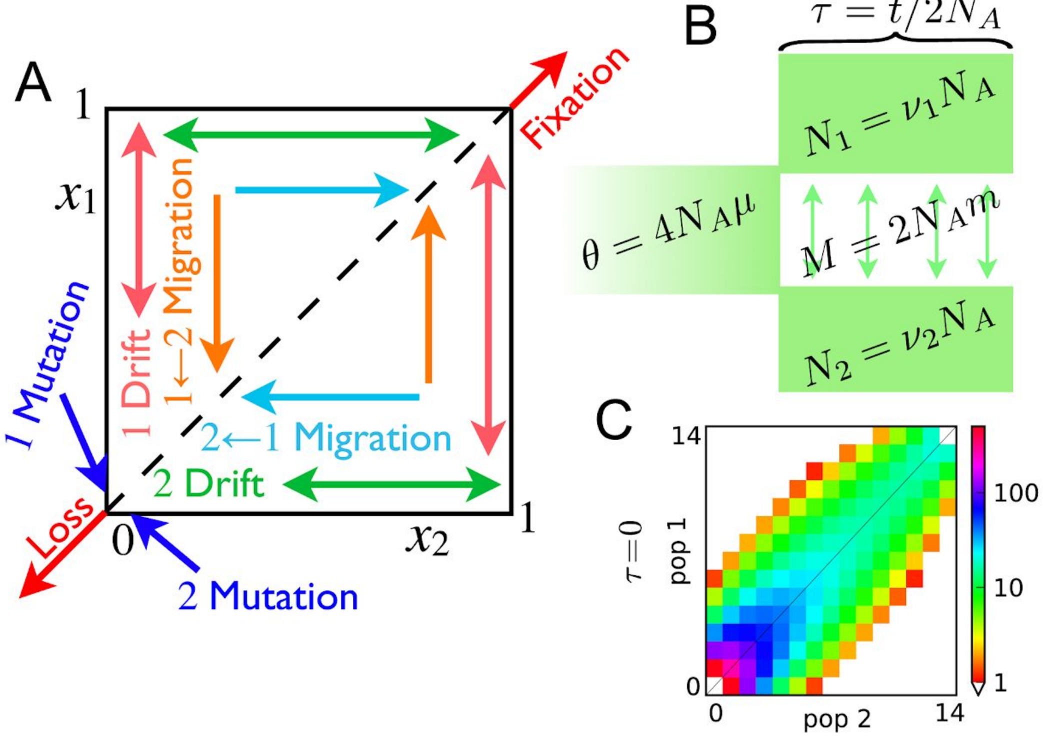
from Nielsen and Slatkin 2005

2-D SFS and demography



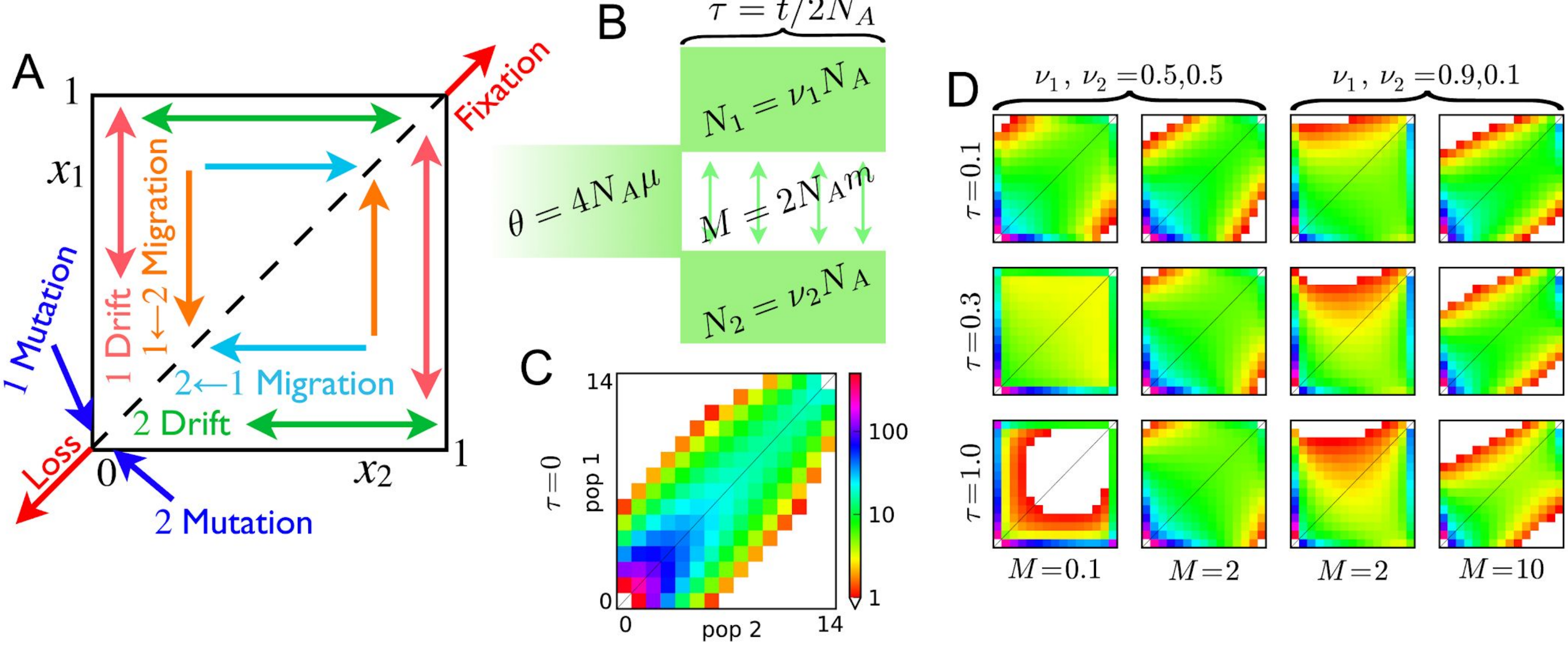
Gutenkunst et al 2009 10.1371/journal.pgen.1000695

2-D SFS and demography



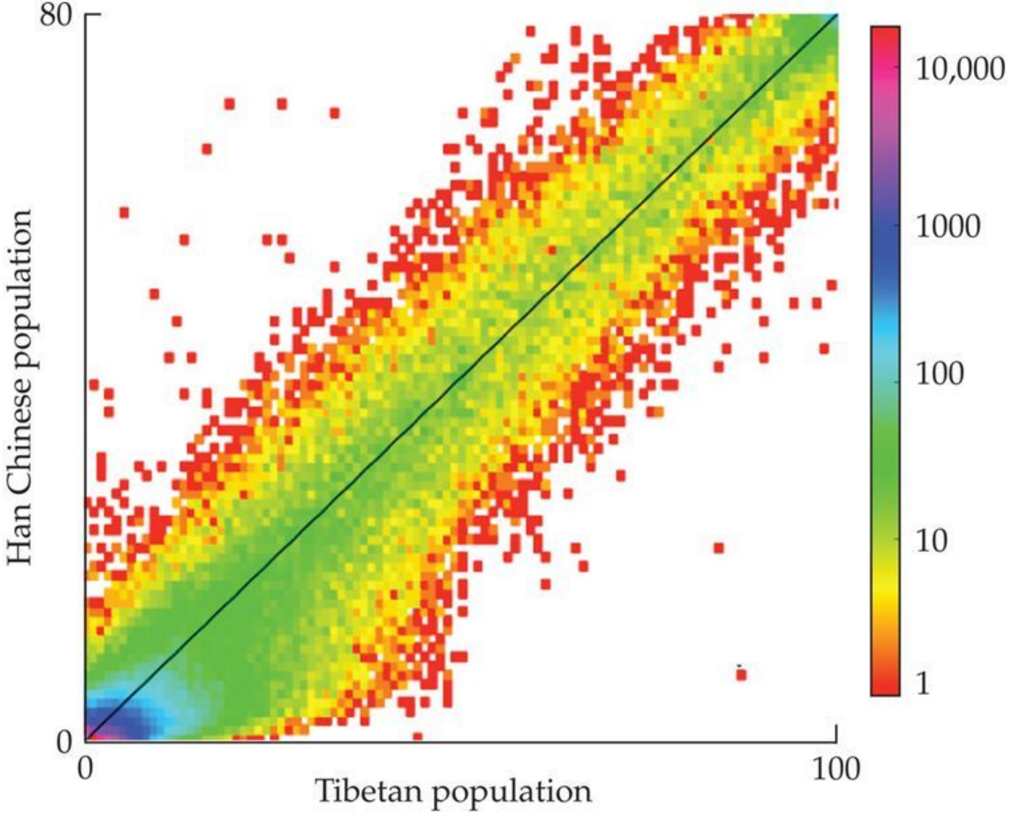
Gutenkunst et al 2009 10.1371/journal.pgen.1000695

2-D SFS and demography



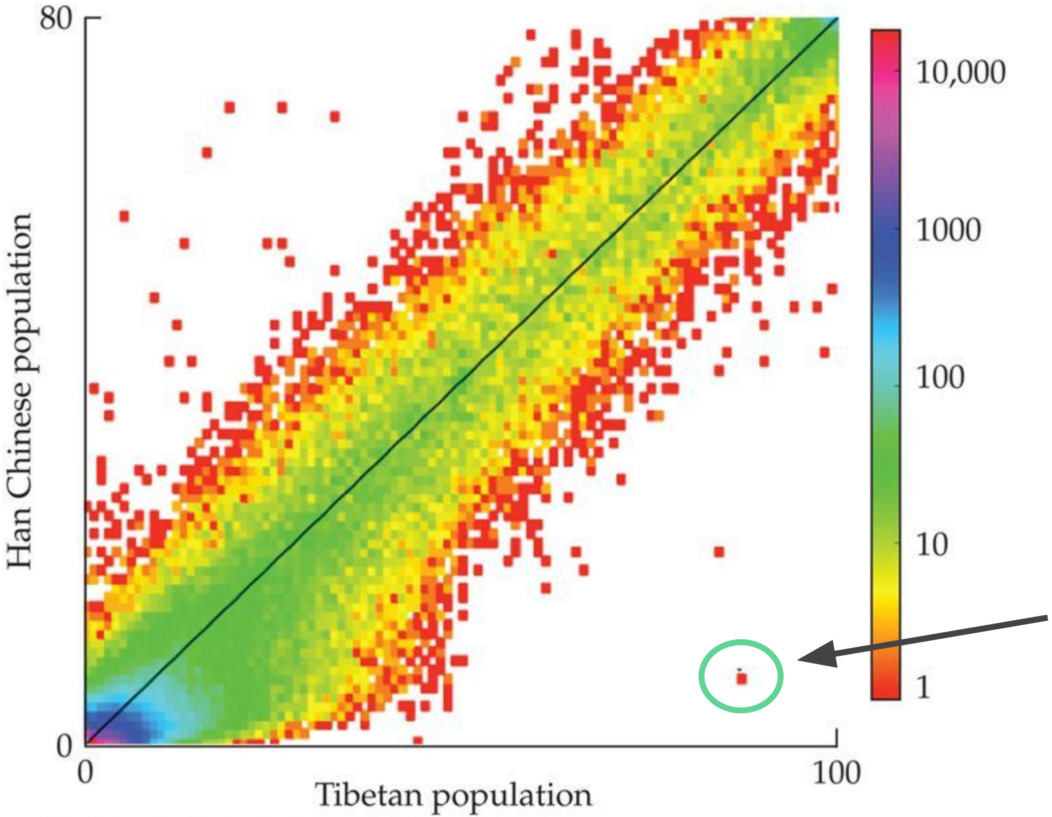
Gutenkunst et al 2009 10.1371/journal.pgen.1000695

2-D SFS and selection



from Nielsen and Slatkin 2005

2-D SFS and selection



EPAS1
hypoxia-inducible
factor-2alpha

from Nielsen and Slatkin 2005

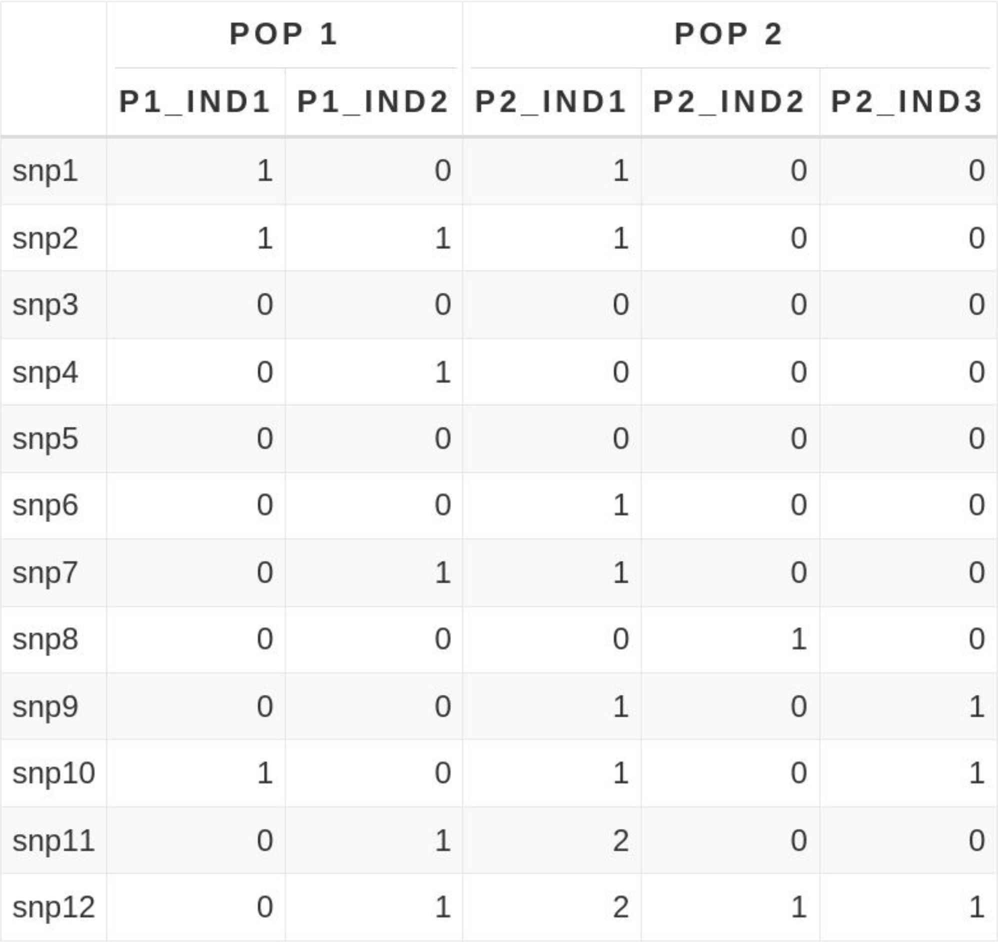
2-D SFS

Let’s try to calculate a 2-D SFS by hand!

Genotype data are the same as before.

First, compute the absolute allele frequency for each SNP in each population. Then, count the number of sites with a given allele frequency in each population.

How many rows and columns does the 2-D SFS have for this dataset? What affects the number of rows and columns? What is the range of possible values for each entry of the matrix?



	POP 1		POP 2		
	P1_IND1	P1_IND2	P2_IND1	P2_IND2	P2_IND3
snp1	1	0	1	0	0
snp2	1	1	1	0	0
snp3	0	0	0	0	0
snp4	0	1	0	0	0
snp5	0	0	0	0	0
snp6	0	0	1	0	0
snp7	0	1	1	0	0
snp8	0	0	0	1	0
snp9	0	0	1	0	1
snp10	1	0	1	0	1
snp11	0	1	2	0	0
snp12	0	1	2	1	1

Demographic inference using SFS as summary stats

0. Access your Amazon EC2 instance
As usual, get your public domain address and login using Guacamole

1. Coalescent-based simulation of SFS under different demographic models using *fastsimcoal2*

You will now simulate genomic data under different demographic models using fastsimcoal2 (http://cmpg.unibe.ch/software/fastsimcoal2/). This software directly outputs the derived site frequency spectrum of each simulated dataset that we will use for comparing the SFS pattern of each demographic model.

Fastsimcoal2

59,029 visits since March

SIB Swiss Institute of Bioinformatics

fastsimcoal2

UNIVERSITÄT BERN

fast sequential markov coalescent simulation of genomic data under complex evolutionary models

- ✔ Can handle very complex evolutionary scenarios including an arbitrary migration matrix between samples, historical events allowing for population resize, population fusion and fission, admixture events, changes in migration matrix, or changes in population growth rates.

- ✔ Different markers, such as DNA sequences, SNPs, STRs (microsatellites) or multi-locus allelic data can be generated under a variety of mutation models

- ✔ Allows to estimate demographic parameters from the (joint) site frequency spectrum (SFS) using simulations to compute the expected SFS and a robust method for the maximization of the composite likelihood.

- ✔ **http://cmpg.unibe.ch/software/fastsimcoal27/**

Demographic inference using SFS as summary stats

0. Access your Amazon EC2 instance
As usual, get your public domain address and login using Guacamole

1. Coalescent-based simulation of SFS under different demographic models using *fastsimcoal2*

You will now simulate genomic data under different demographic models using fastsimcoal2 (http://cmpg.unibe.ch/software/fastsimcoal2/). This software directly outputs the derived site frequency spectrum of each simulated dataset that we will use for comparing the SFS pattern of each demographic model.

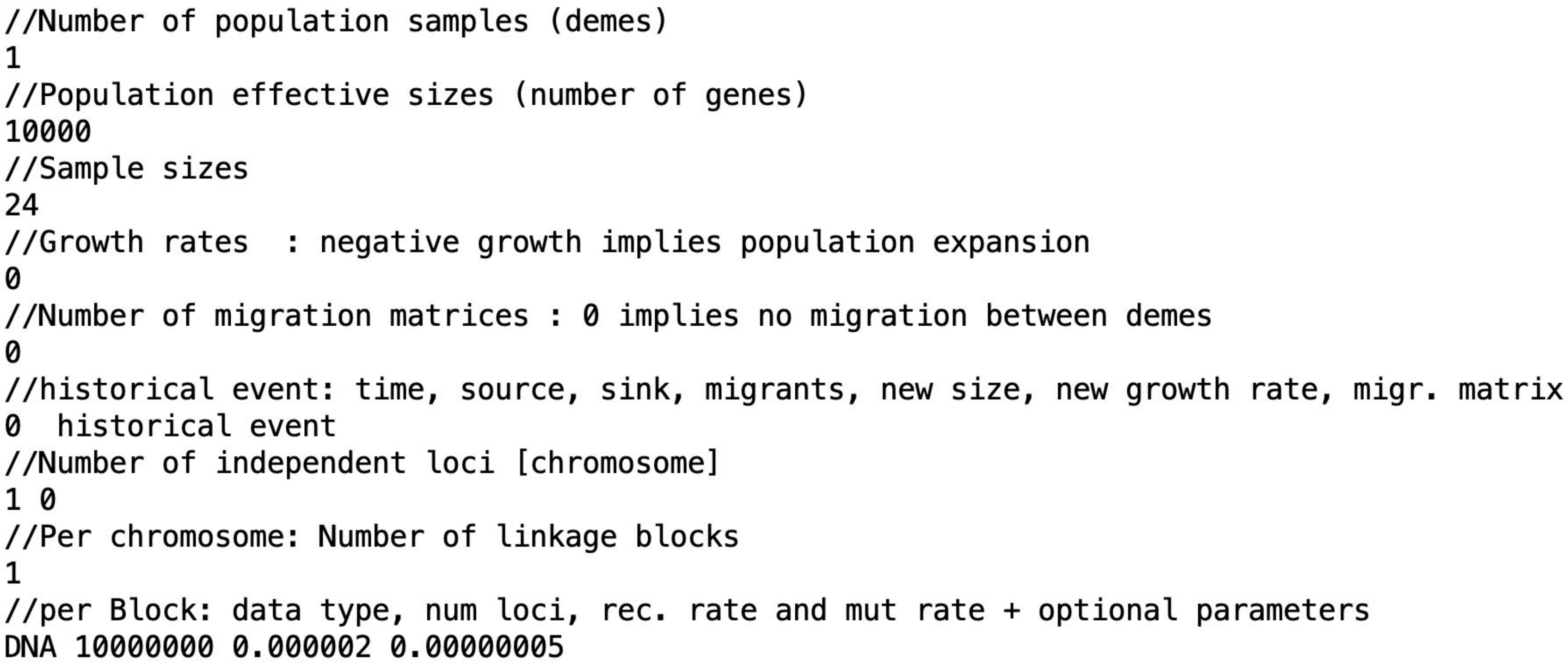
We will compare three simple demographic models:
i) a constant population size
ii) an instantaneous growth
iii) an instantaneous decline occurring 1000 generations in the past.

Open a ***Terminal*** and move to this directory

```
cd ~/workshop_materials/pop_gen/lab_02/fastsimcoal2
```

Demographic inference using SFS as summary stats

constant1pop.par input file: constant population size

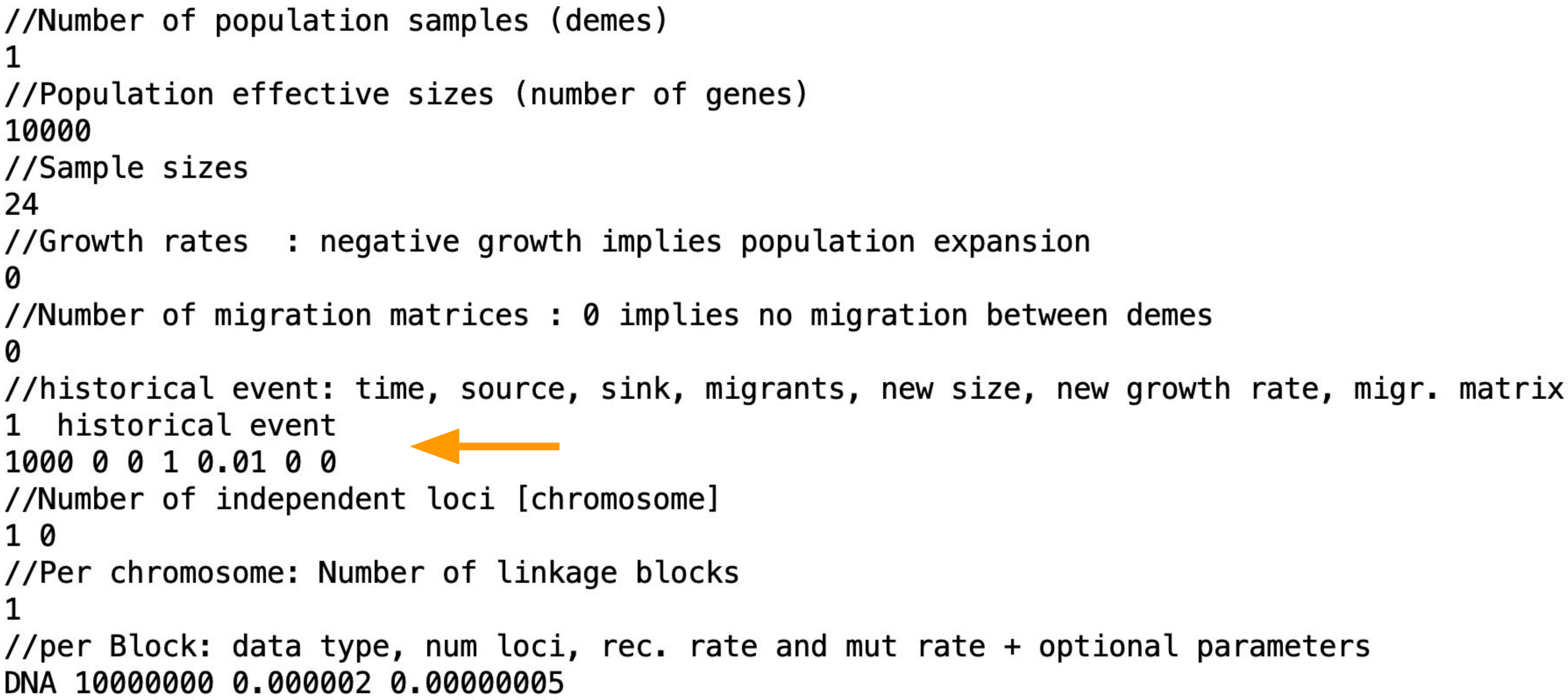


```
//Number of population samples (demes)
1
//Population effective sizes (number of genes)
10000
//Sample sizes
24
//Growth rates  : negative growth implies population expansion
0
//Number of migration matrices : 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new size, new growth rate, migr. matrix
0  historical event
//Number of independent loci [chromosome]
1 0
//Per chromosome: Number of linkage blocks
1
//per Block: data type, num loci, rec. rate and mut rate + optional parameters
DNA 10000000 0.000002 0.00000005
```

Very high just to get enough SNPs in 10Mb

Demographic inference using SFS as summary stats

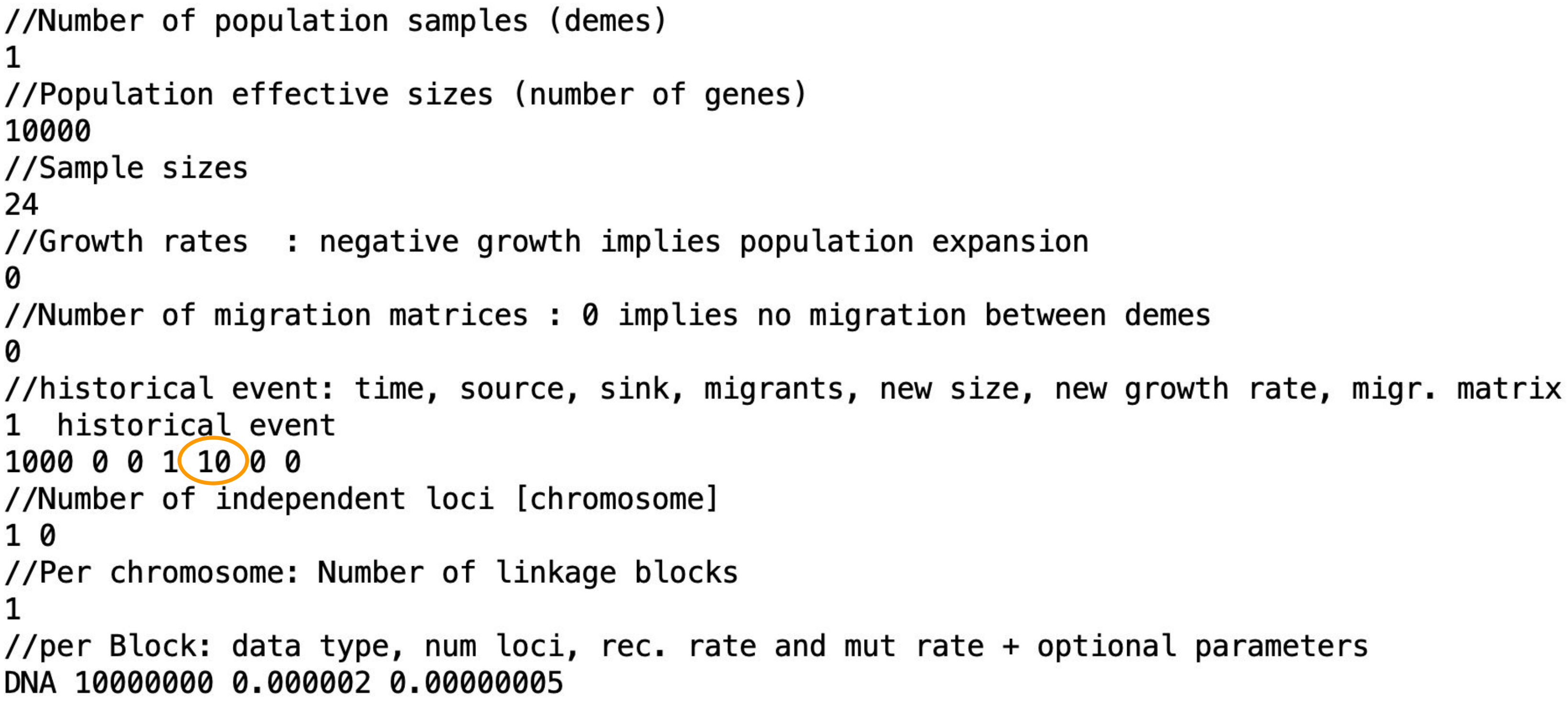
instGrowth1pop.par input file: instantaneous growth population size



```
//Number of population samples (demes)
1
//Population effective sizes (number of genes)
10000
//Sample sizes
24
//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 size, new growth rate, migr. matrix
1  historical event
1000 0 0 1 0.01 0 0
//Number of independent loci [chromosome]
1 0
//Per chromosome: Number of linkage blocks
1
//per Block: data type, num loci, rec. rate and mut rate + optional parameters
DNA 10000000 0.000002 0.00000005
```

Demographic inference using SFS as summary stats

instDecline1pop.par input file: instantaneous decline population size



```
//Number of population samples (demes)
1
//Population effective sizes (number of genes)
10000
//Sample sizes
24
//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 size, new growth rate, migr. matrix
1  historical event
1000 0 0 1 10 0 0
//Number of independent loci [chromosome]
1 0
//Per chromosome: Number of linkage blocks
1
//per Block: data type, num loci, rec. rate and mut rate + optional parameters
DNA 10000000 0.000002 0.00000005
```

Demographic inference using SFS as summary stats

Launch *fastsimcoal2* using the *constant1pop.par* model

```
./fsc26 -i constant1pop.par -n1 -q -s0 -d -k 1000000 -c 4
```

This will take approximately 8 minutes to run …

This command uses several different parameters! These can be viewed by running:

```
./fsc26 -h
```

-n = number of simulations
-q = sets the run to quiet mode. This means minimal messages will be output to console
-s = outputs the DNA as SNP data. The number relates to the maximum number of SNPs, where 0 outputs all of the SNPs
-d = computes derived site frequency spectrum
-k = specifies the number of simulated polymorphic sites kept in memory
-c = number of cores (CPUs)

Demographic inference using SFS as summary stats

You should now see a new folder within the fastsimcoal2 directory called ***constant1pop***.

Inside, look for a file called ***constant1pop_DAFpop0.obs*** and open it with a text editor (*e.g*, ***nano***).

This is the derived allele frequency spectrum from the data simulated under the model of constant population size.

Demographic inference using SFS as summary stats

Let’s plot it!

Remove the first line from the *constant1pop_DAFpop0.obs* file (it says “1 observations”) and also remove any text that says “xx sites with multiple mutations were discarded”

Plot the SFS:

Rscript plot_1Dsfs_constant1pop.R

and open the plot **xpdf constant1Pop_DAF.pdf**

This is your first SFS!

The option -n1 in the fsc command line above tells fsc26 to make 1 simulation.

Let’s compare how the simulations differ. Compare your SFS with your neighbours. Are they the same?

Demographic inference using SFS as summary stats

Now run 1 simulation using the other demographic models provided in the folder as .par files:

instDecline1pop.par

```
./fsc26 -i instDecline1pop.par -n1 -q -s0 -d -k 1000000 -c 4
```

instGrowth1pop.par

```
./fsc26 -i instGrowth1pop.par -n1 -q -s0 -d -k 1000000 -c 4
```

As before, edit the resulting SFS in a text editor by removing the first line and removing any text that says “xx sites with multiple mutations were discarded”

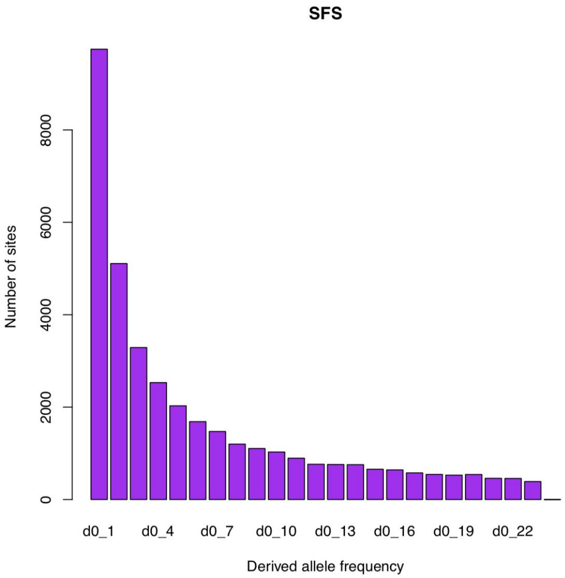
Plot using the relevant R Script

What are the main differences between the SFS?

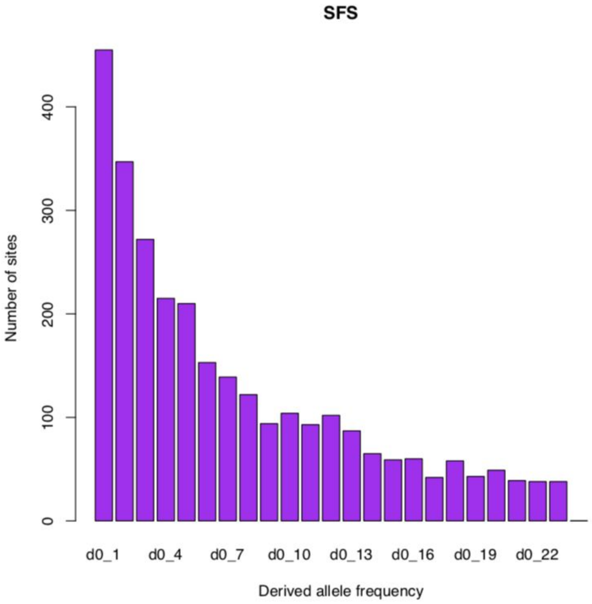
Can you recognize the patterns we talked about before (shifting to the right or to left)? Don’t forget to pay attention to the y axis changes also!

Demographic inference using SFS as summary stats

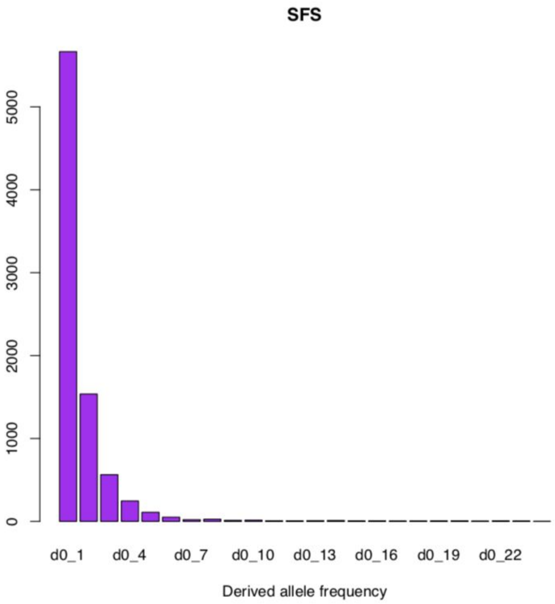
Constant



instDecline



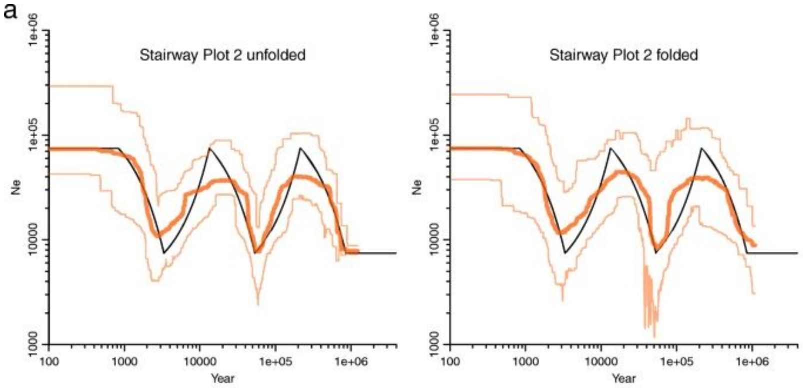
instGrowth



Demographic inference using SFS as summary stats

2. Test a demographic inference from a SFS

Stairway_plot_v2



✔ Cross-platform program package for inferring the demographic histories of populations using SNP frequency spectra.

✔ It is based on a nonparametric method with the capability of handling folded and unfolded SNP frequency spectra (that is, when the ancestral alleles of the SNPs are unknown) of thousands of samples produced with genotyping-by-sequencing technologies;

✔ Particularly suitable for non-model organisms.

✔ **https://github.com/xiaoming-liu/stairway-plot-v2**

Demographic inference using SFS as summary stats

2. Test a demographic inference from a SFS using *stairway_plot*

We are going to run stairwayplot using the last two SFS we generated with *fastsimcoal2* under the constant (*constant1pop.par*) and the instantaneous growth (*instGrowth1pop.par*) models.

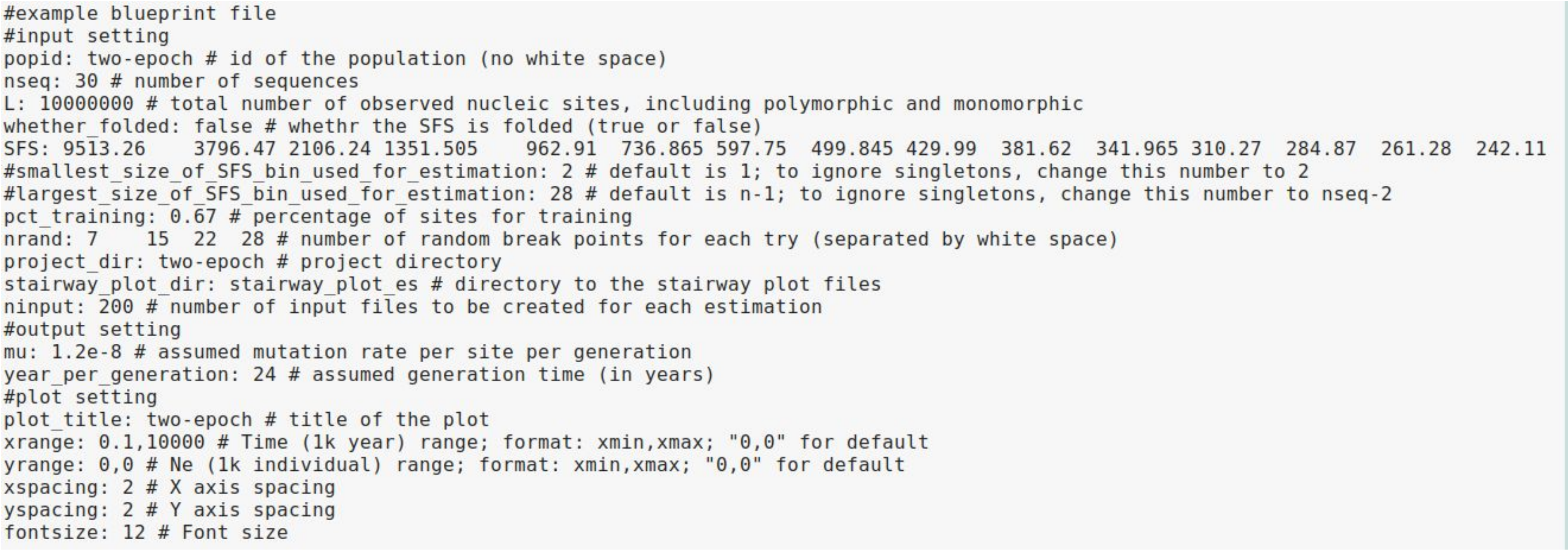
Move into the *stairwayplot* folder.

```
cd ~/workshop_materials/pop_gen/lab_02/stairwayplot
```

Take a look at the example file two-epoch.blueprint

Demographic inference using SFS as summary stats

two-epoch.blueprint



```
#example blueprint file
#input setting
popid: two-epoch # id of the population (no white space)
nseq: 30 # number of sequences
L: 10000000 # total number of observed nucleic sites, including polymorphic and monomorphic
whether_folded: false # whethr the SFS is folded (true or false)
SFS: 9513.26    3796.47 2106.24 1351.505    962.91  736.865 597.75  499.845 429.99  381.62  341.965 310.27  284.87  261.28  242.11
#smallest_size_of_SFS_bin_used_for_estimation: 2 # default is 1; to ignore singletons, change this number to 2
#largest_size_of_SFS_bin_used_for_estimation: 28 # default is n-1; to ignore singletons, change this number to nseq-2
pct_training: 0.67 # percentage of sites for training
nrand: 7    15  22  28 # number of random break points for each try (separated by white space)
project_dir: two-epoch # project directory
stairway_plot_dir: stairway_plot_es # directory to the stairway plot files
ninput: 200 # number of input files to be created for each estimation
#output setting
mu: 1.2e-8 # assumed mutation rate per site per generation
year_per_generation: 24 # assumed generation time (in years)
#plot setting
plot_title: two-epoch # title of the plot
xrange: 0.1,10000 # Time (1k year) range; format: xmin,xmax; "0,0" for default
yrange: 0,0 # Ne (1k individual) range; format: xmin,xmax; "0,0" for default
xspacing: 2 # X axis spacing
yspacing: 2 # Y axis spacing
fontsize: 12 # Font size
```

Demographic inference using SFS as summary stats

2. Test a demographic inference from a SFS using *stairway_plot*

To help with time, we’ve already edited most of the blueprint file to match the simulations in fastsimcoal2 to make one for constantpop and instgrowth.

But some parameters of the blueprint file need to be changed to match those used in the fastsimcoal2 simulations:

- nseq OR sample size
- L OR num. loci
- mu OR mutation rate

You’ll also need to add your simulated SFS to the SFS line in the file (don’t include the monomorphic sites - and remember it’s unfolded)

Demographic inference using SFS as summary stats

Launch *stairway_plot* using the *constant.blueprint* input. Two commands are needed this time:

```
java -cp stairway_plot_es Stairbuilder constant.blueprint
```

```
bash constant.blueprint.sh
```

The software will take a while to complete (20 mins or so). So, you can open another *Terminal*, move to the same folder and start another process using the *instgrowth.blueprint* input.

IMPORTANT: stairway_plot will write the output to the folder specified as *project_dir,* make sure it is different between the two blueprint files.

Once completed you will find the plot with the demographic inference in each *project_dir* folder.

Was *stairwayplot* able to retrieve the two different demographic models?

Demographic inference using SFS as summary stats

3. BONUS Calculate a SFS from a vcf file

We will now use a custom python script to calculate the SFS from real variant data stored in a vcf file. Are we all good with vcf format?

Move to the folder *calc_SFS*

```
cd ~/workshop_materials/pop_gen/lab_02/calc_SFS
```

Launch the python script *vcf2SFS.py* to calculate the SFS for one species (Emperor penguin individuals are specified in the *emp* file) using only loci without missing data (-m 24, there are 24 inds in the vcf)

```
python3 vcf2SFS.py -f NW_008794753.1.filt.biall.recode.vcf -p emp -m 24 -i 1 -s 0
```

This script will output a minor allele SFS as text on the *Terminal* and as a plot. To calculate the SFS for the other species (King penguin -p king) relaunch the script changing the -p option.

Demographic inference using SFS as summary stats

4. BONUS Infer King and Emperor penguin demography

Let’s now try the stairway_plot method with the observed Emperor and King SFS. You need to make two blueprint input files with the two SFS, respectively.

Take into account that the mutation rate for these species is not known. We will use a rather fast mutation rate (1e-7). The generation time is 11 and 16 years for the King and the Emperor, respectively. The total length of this scaffold is 4851924 bp.

We may decide to exclude singletons to be used in the demographic inference as they could be sequencing errors. Which would be the options to be changed in the blueprint files in this case?

These two processes could take some time to run. We can launch them over dinner or overnight and then discuss the results in the next lab.

Actually these data from a single scaffold are not enough and may contain regions that are not neutrally evolving. More data should be added and regions under selection should be removed.