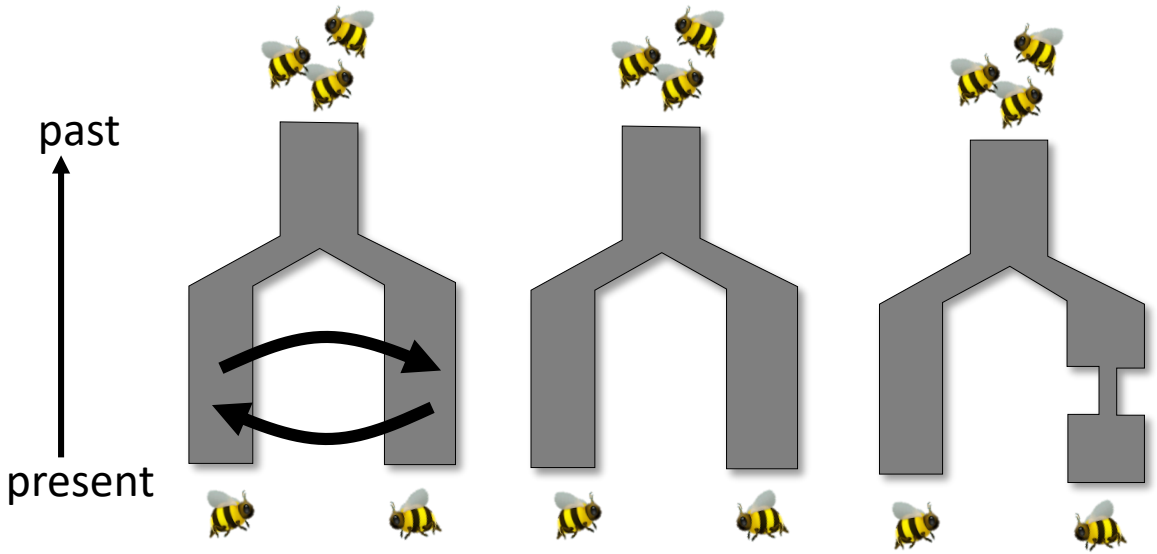
Using modern genetic data to retrace past demographic events

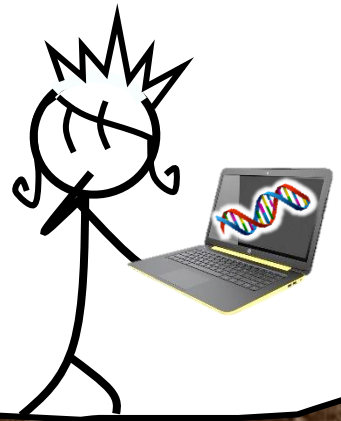
A very short introduction to demographic inferences based on NGS

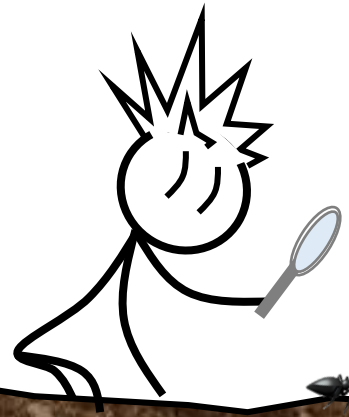
TECHER Maeva, Ecology and Evolution OIST

Workshop @IUSSI2018

Organized by A. Mikheyev and T. Linksvayer

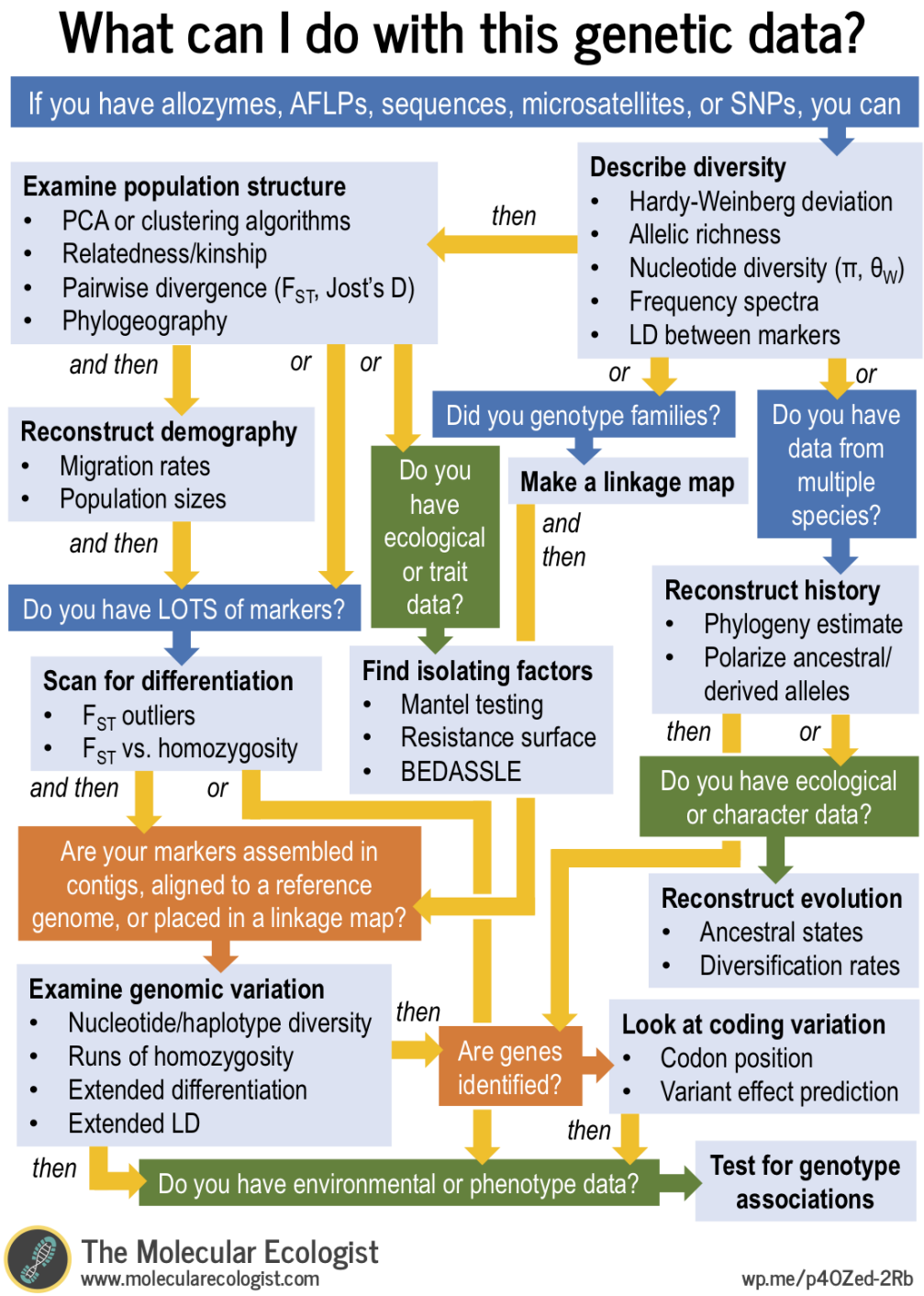




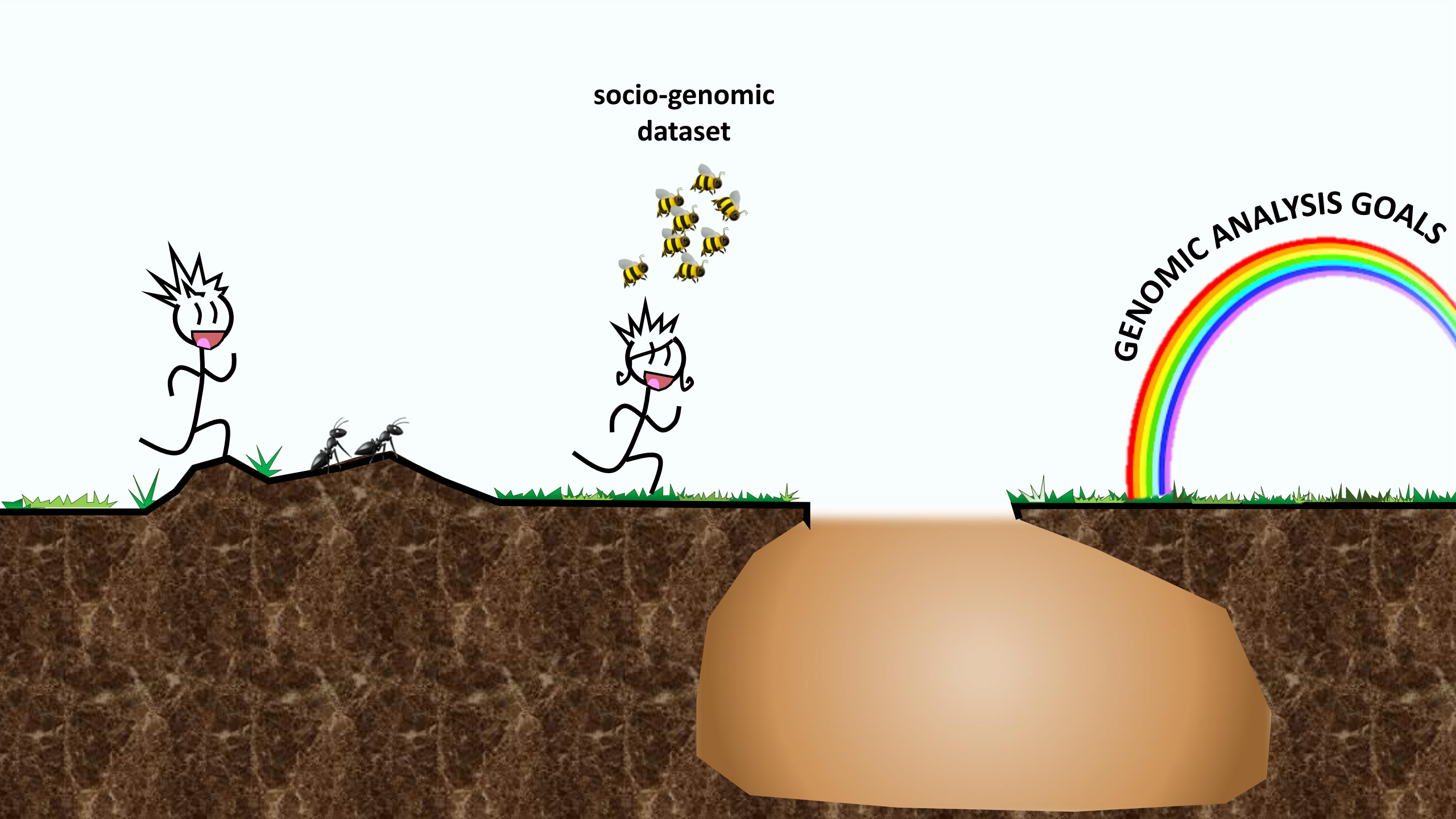


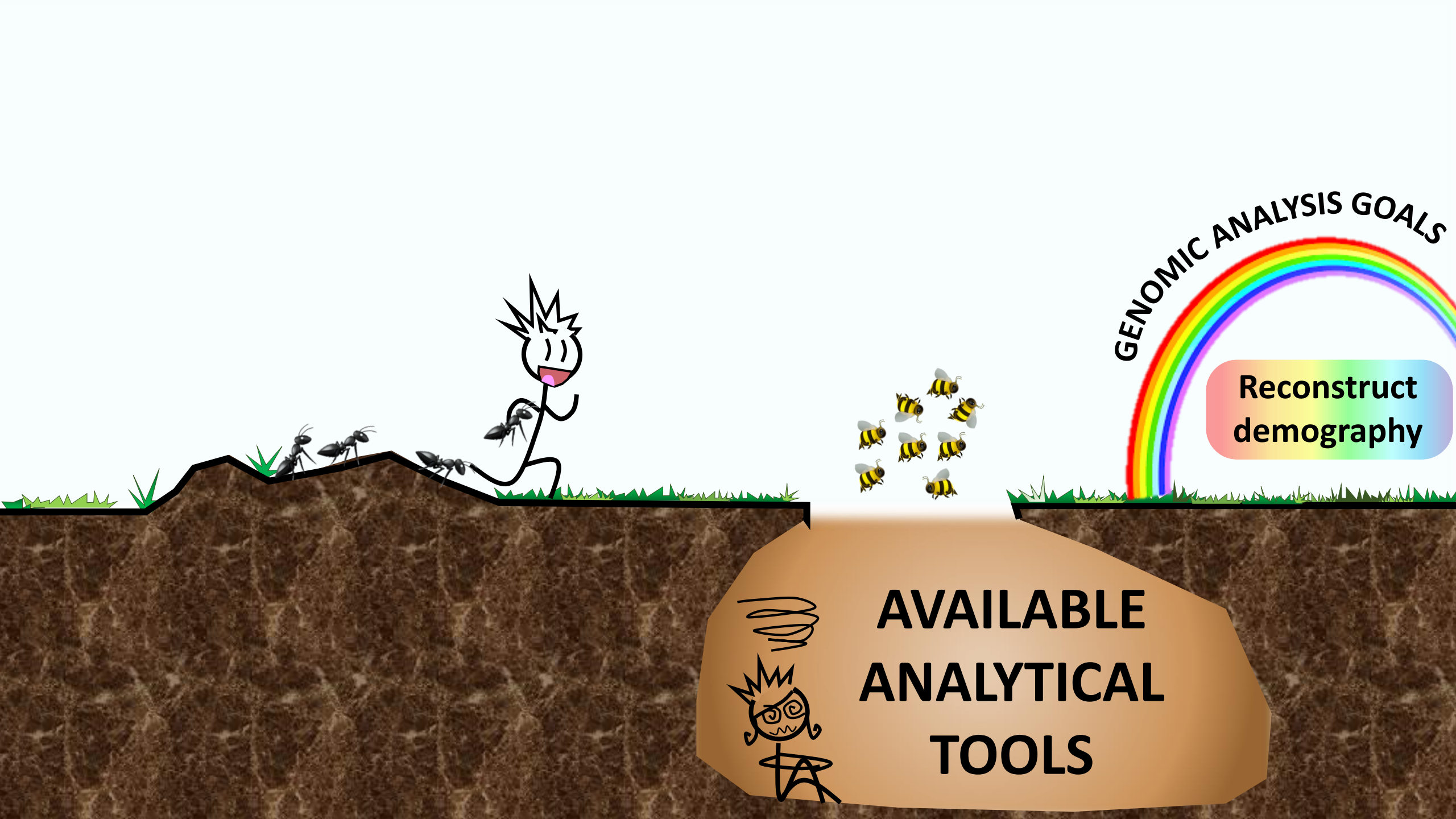
Example with **NGSadmix**

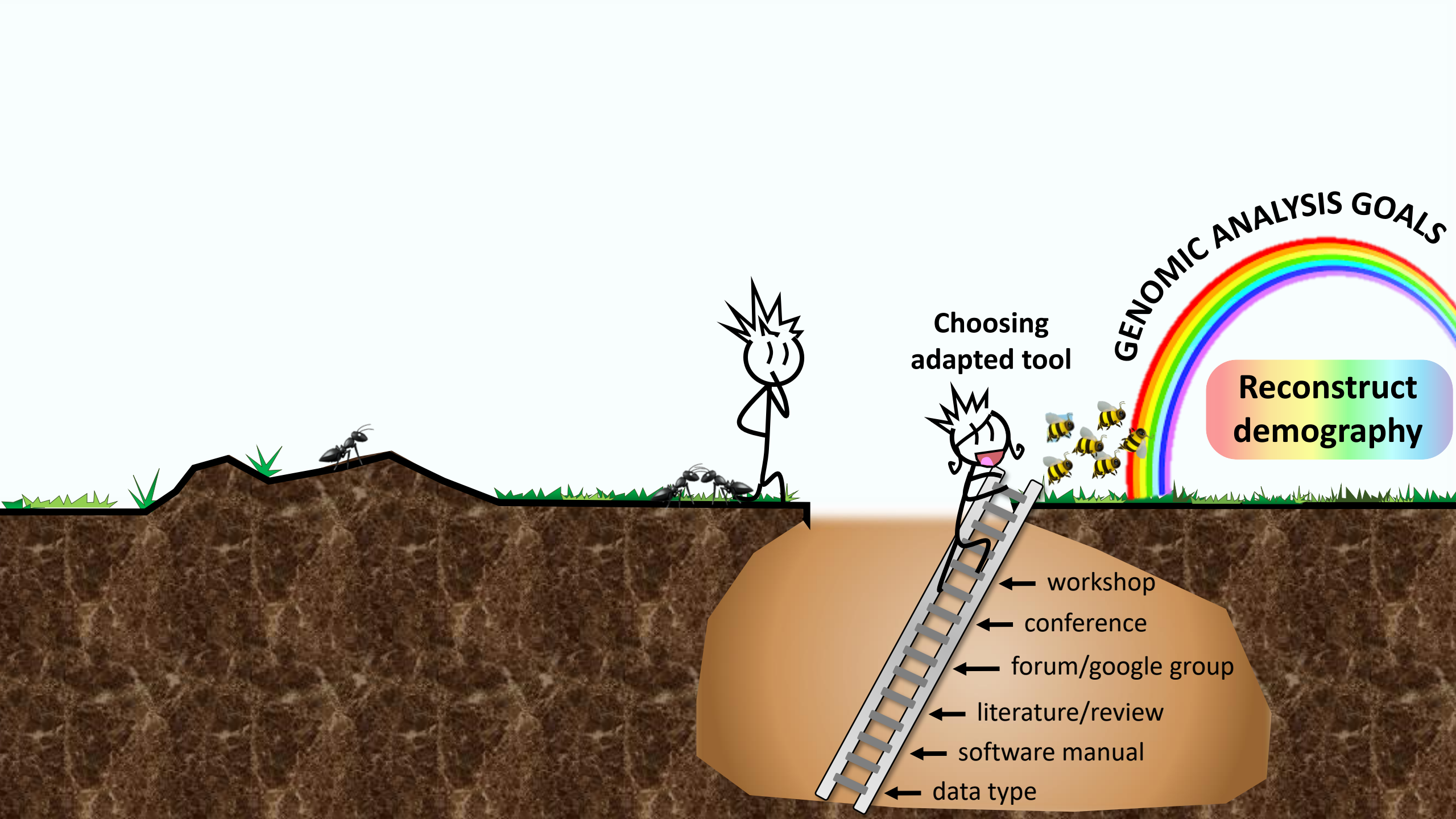
Example using **fastsimcoal2**



Example with **Site Frequency Spectrum (SFS)**

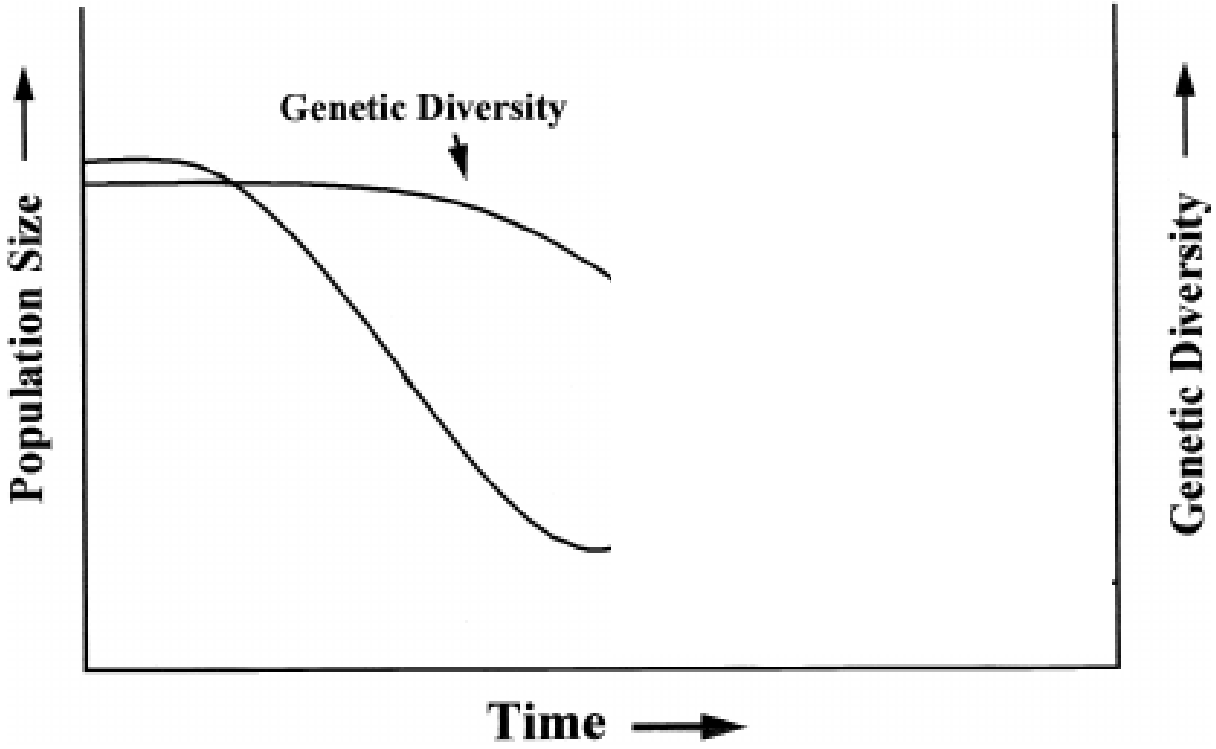

socio-genomic
dataset
GENOMIC ANALYSIS GOALS


GENOMIC ANALYSIS GOALS
Reconstruct demography
AVAILABLE ANALYTICAL TOOLS

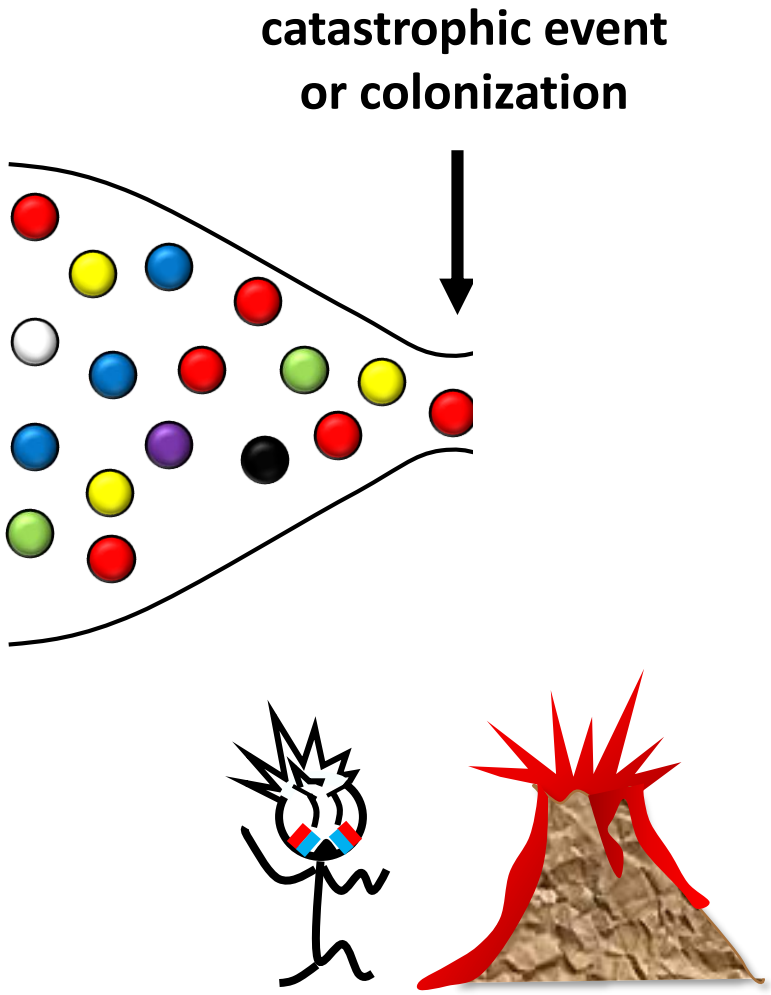

GENOMIC ANALYSIS GOALS
Choosing adapted tool
Reconstruct demography
workshop
conference
forum/google group
literature/review
software manual
data type

Why using demographic inferences?

Past demographic changes leave imprints on current diversity

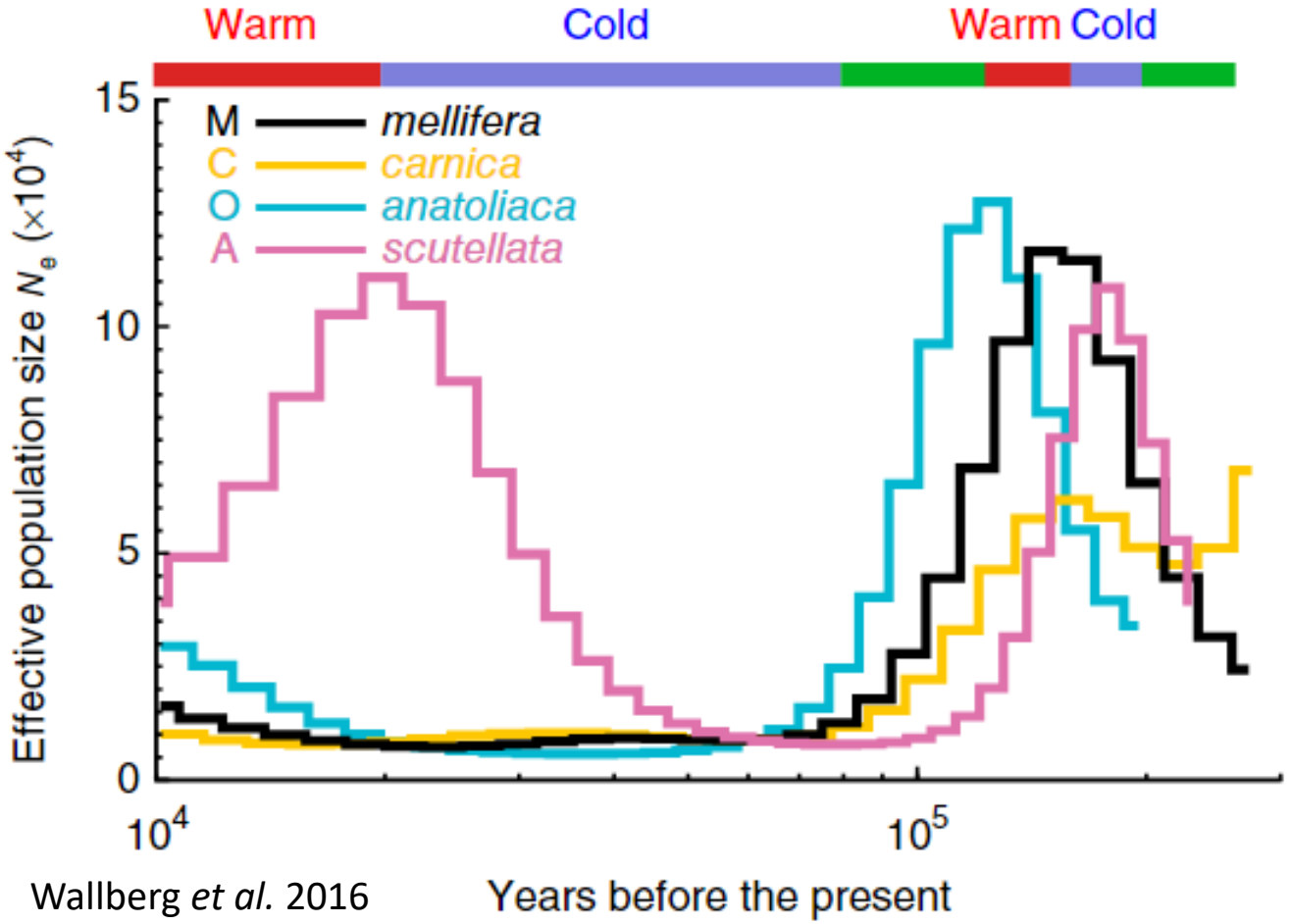


Bickham *et al.* 2000

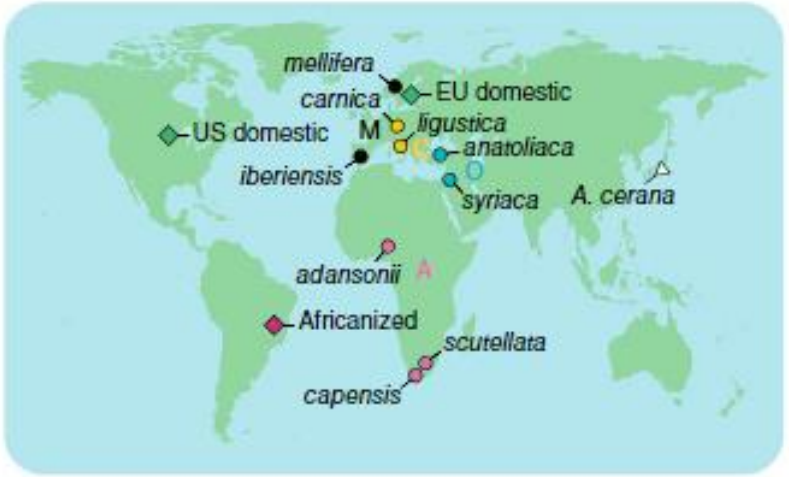


Understanding honey bee evolutionary history with N_e changes

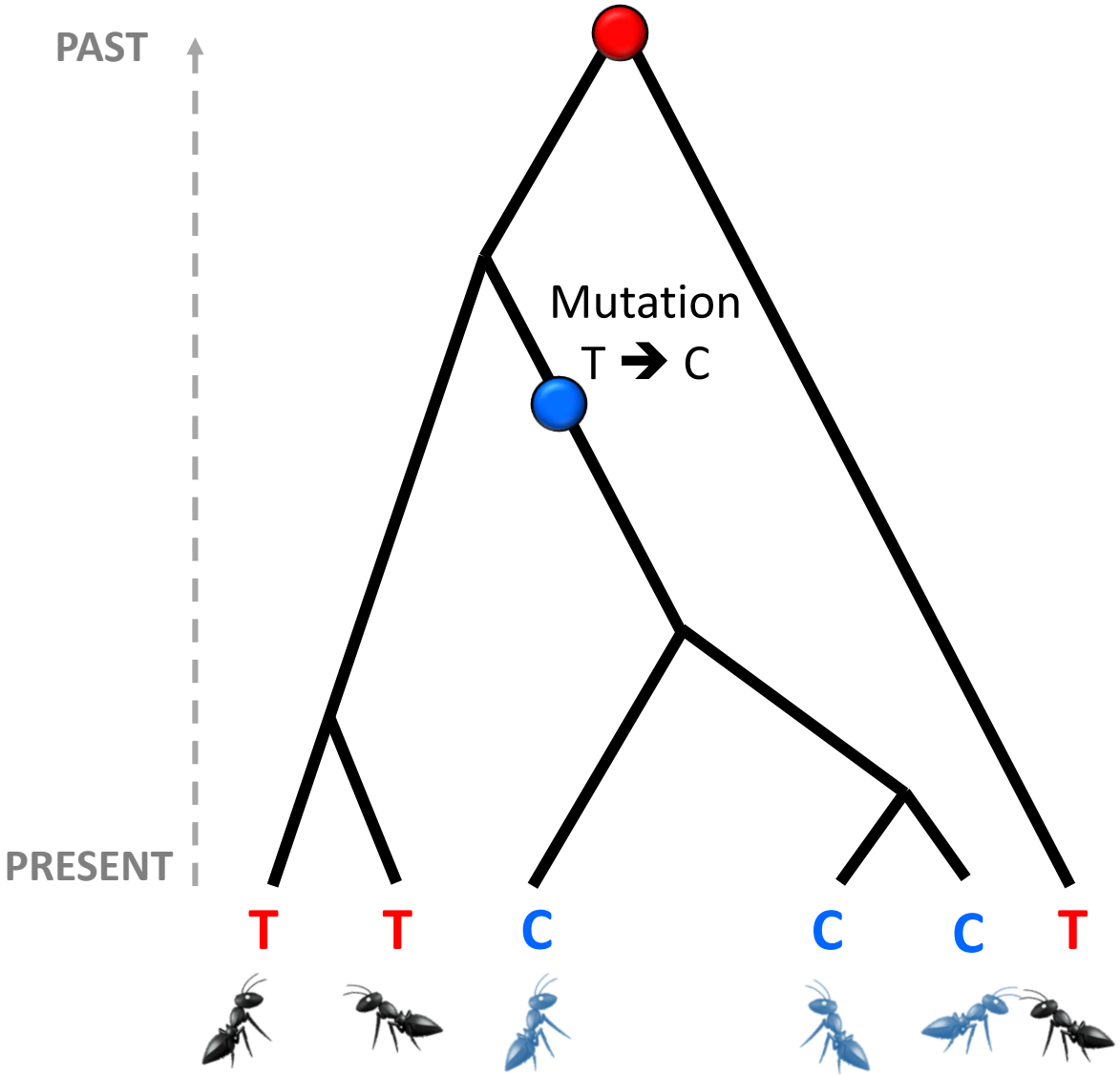
N_e (effective population size) is the idealized population size where all individuals participate in reproduction and actively contribute offspring for the next generation



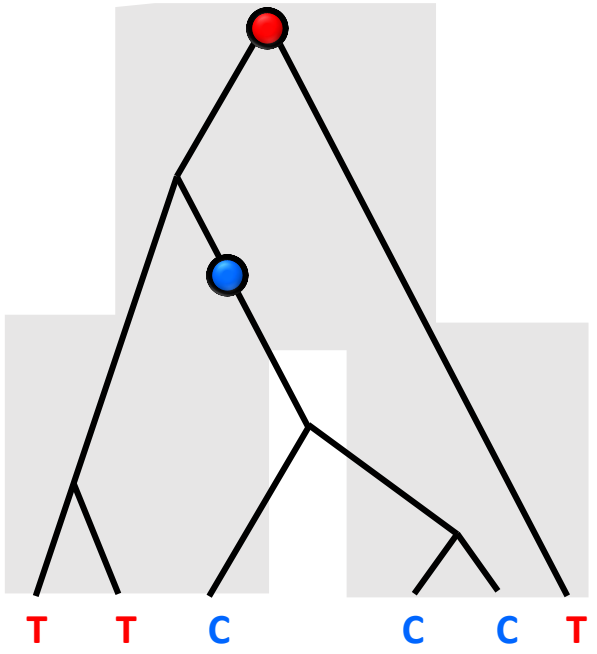
Wallberg *et al.* 2016



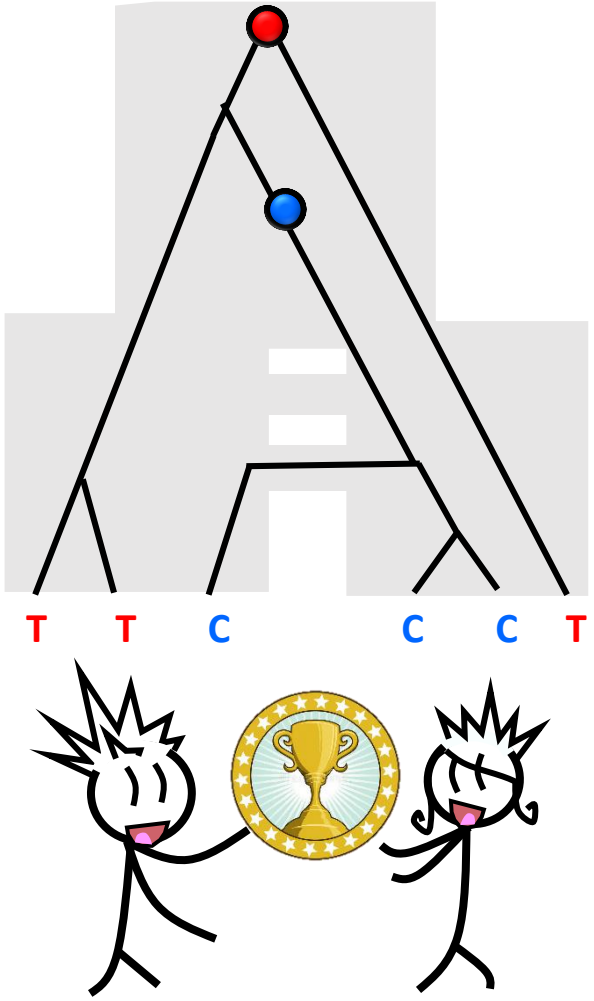
Gene trees informs on divergence between species/populations



Population tree with no migration

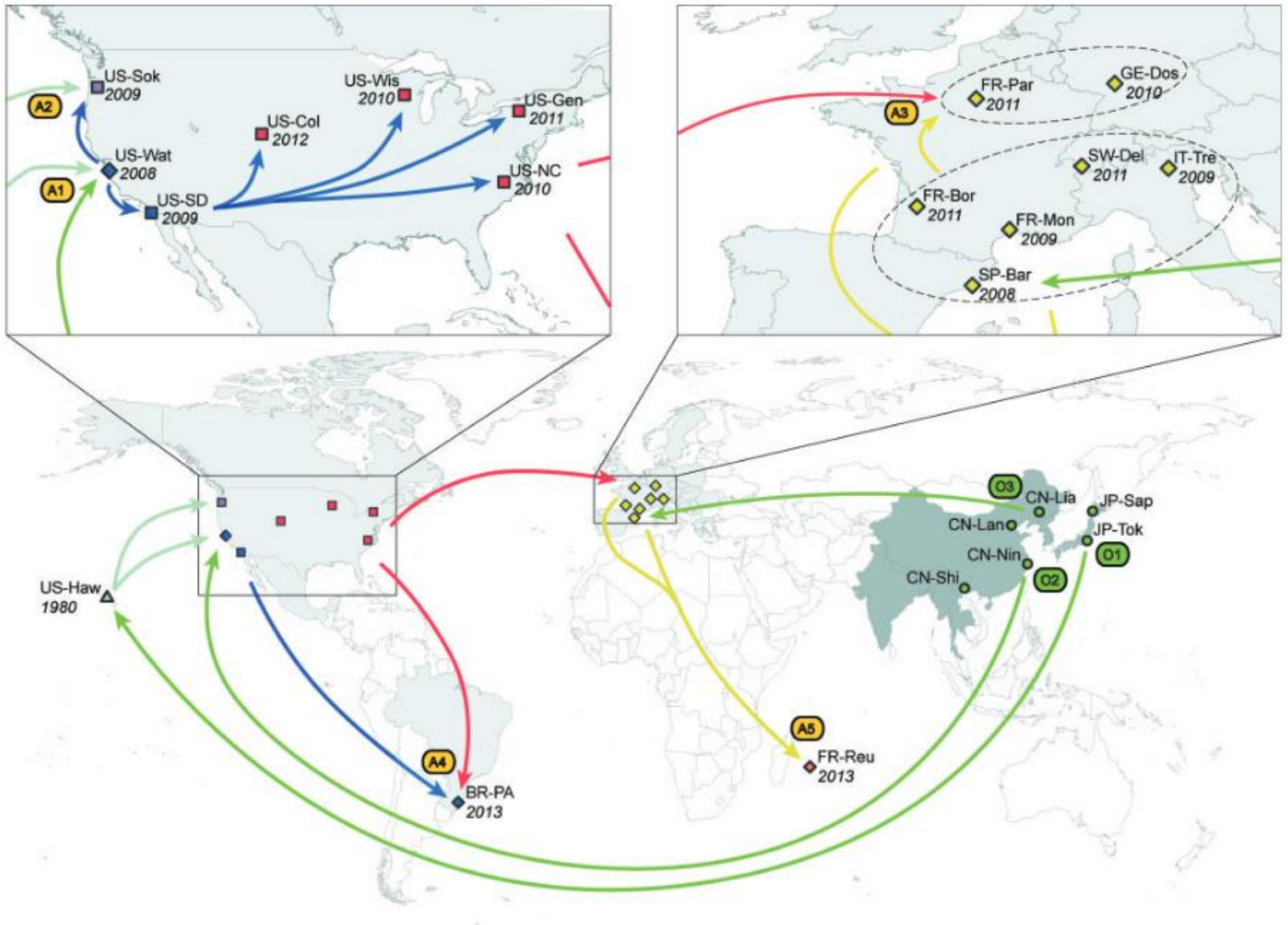


Population tree with migration



adapted from V. Sousa WSPG 2018

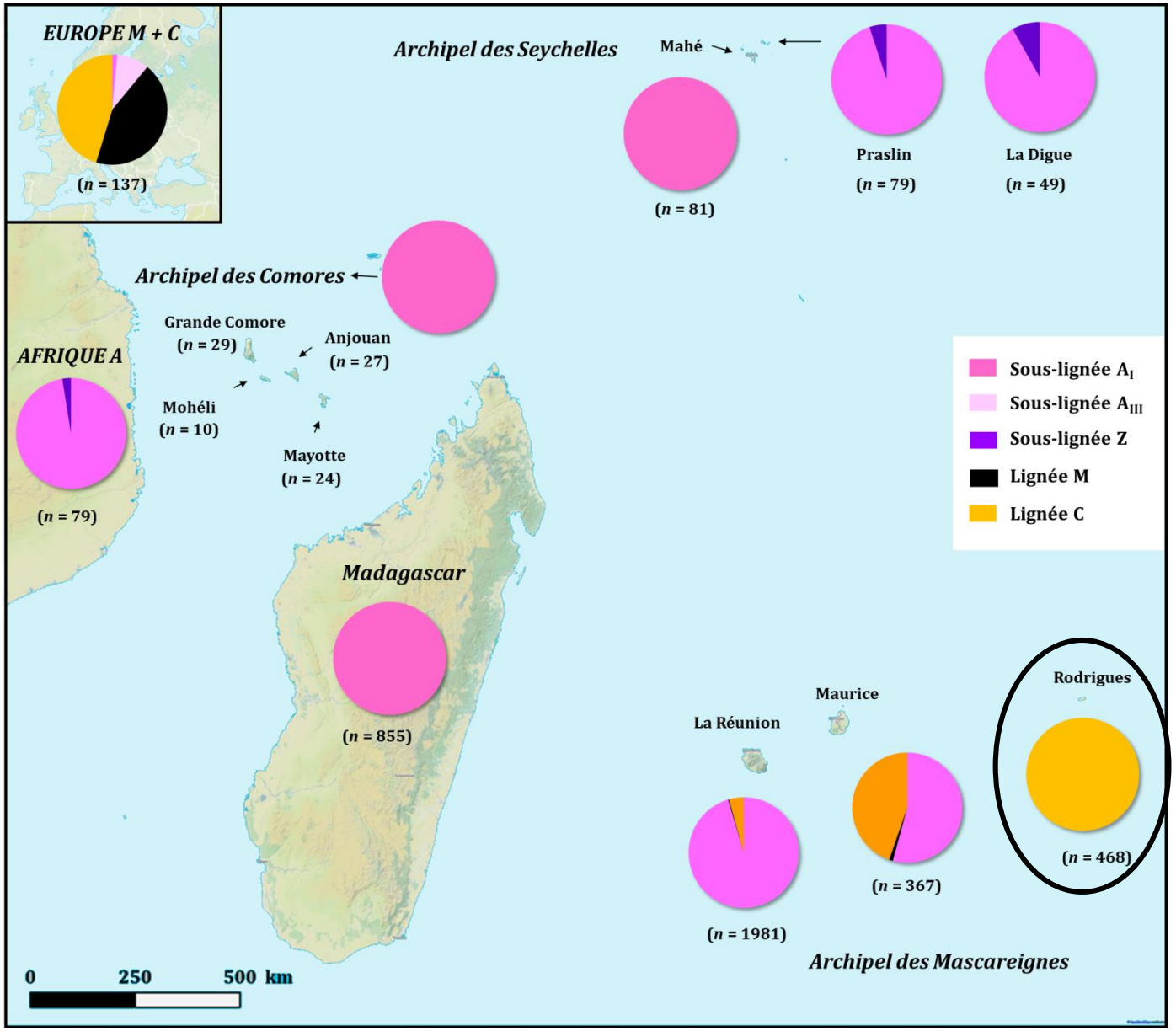
Testing scenarios to retrace complex evolutionary history



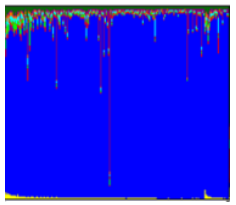
Drosophila suzukii © Theotime Colin

Deciphering the Routes of invasion of Drosophila suzukii by Means of ABC Random Forest (2017) Fraimout *et al.* MBE.

Why did I started using demographic inference?



??



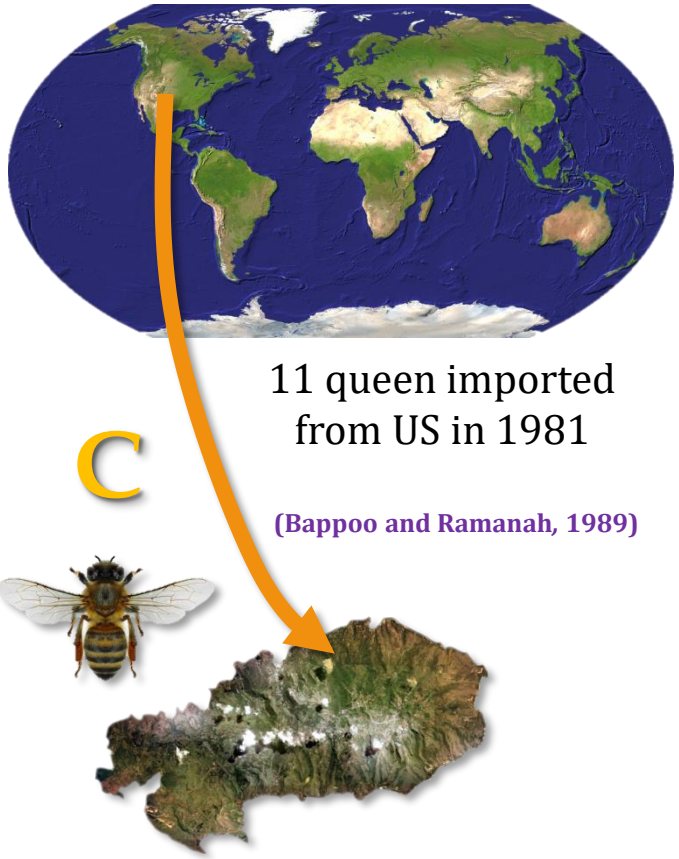
Rodrigues
(*n* = 524)

Techer et al. (2015)

Historical report

USA: 4,5 millions colonies in 1980

(vanEngelsdorp *et al.*, 2008)



11 queen imported from US in 1981

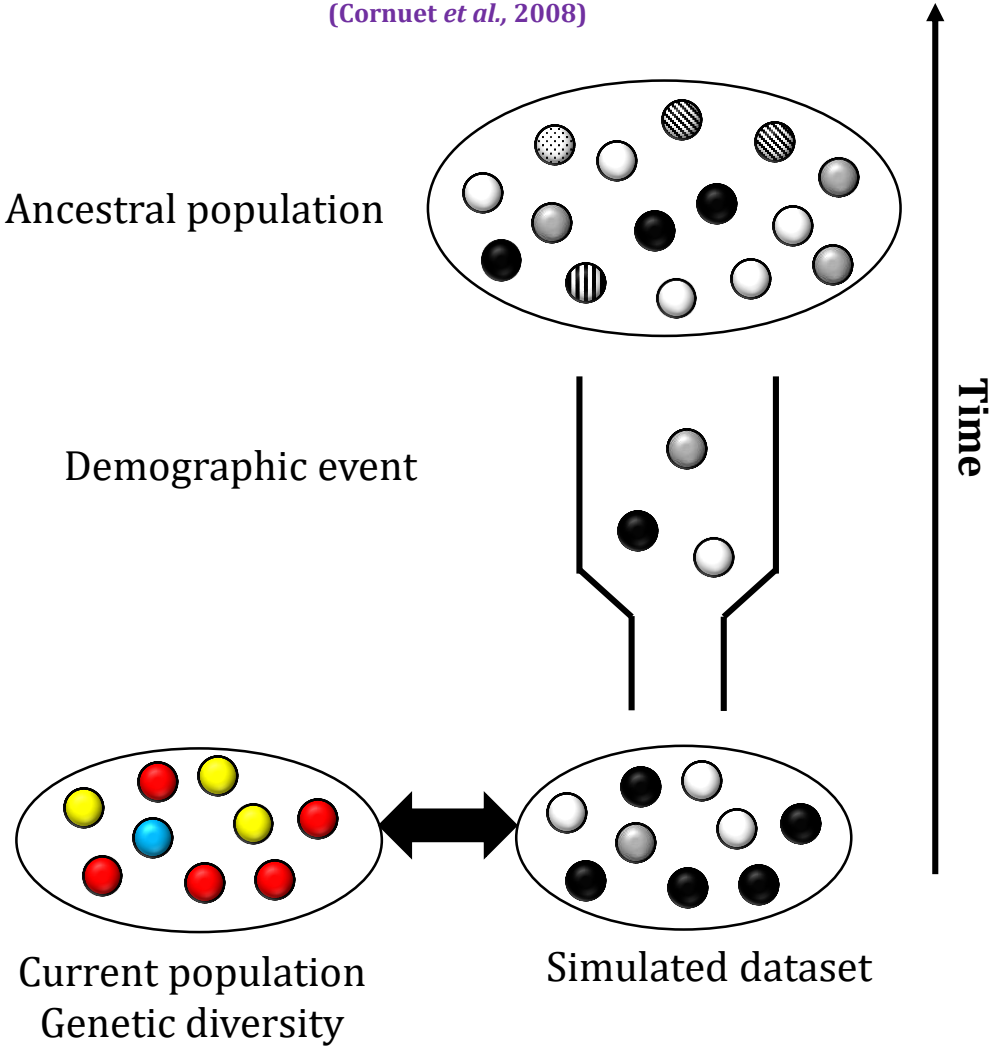
(Bappoo and Ramanah, 1989)

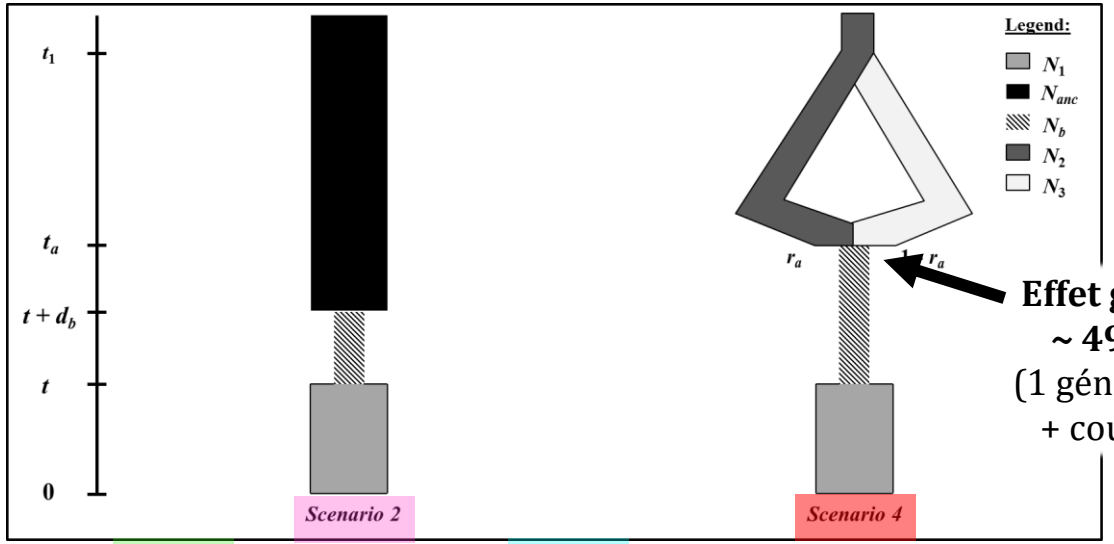
Rodrigues: 2251 colonies in 2010

(Belmin, 2010)

Infering with (DIYABC) (*Approximate Bayesian Computation*)

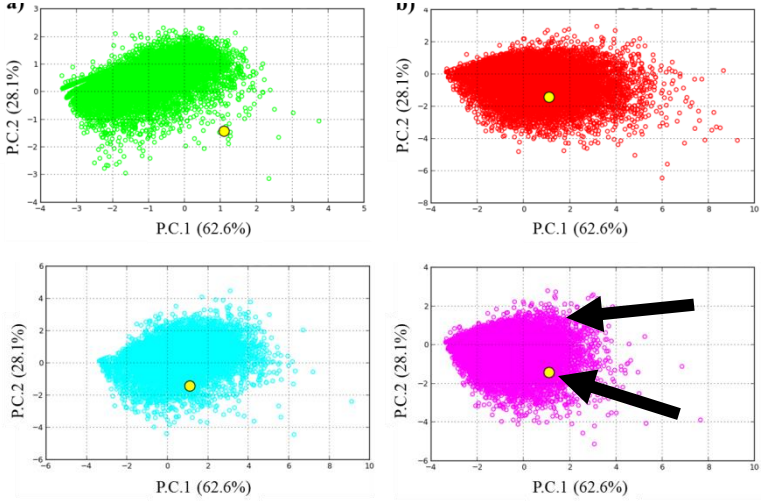
(Cornuet *et al.*, 2008)



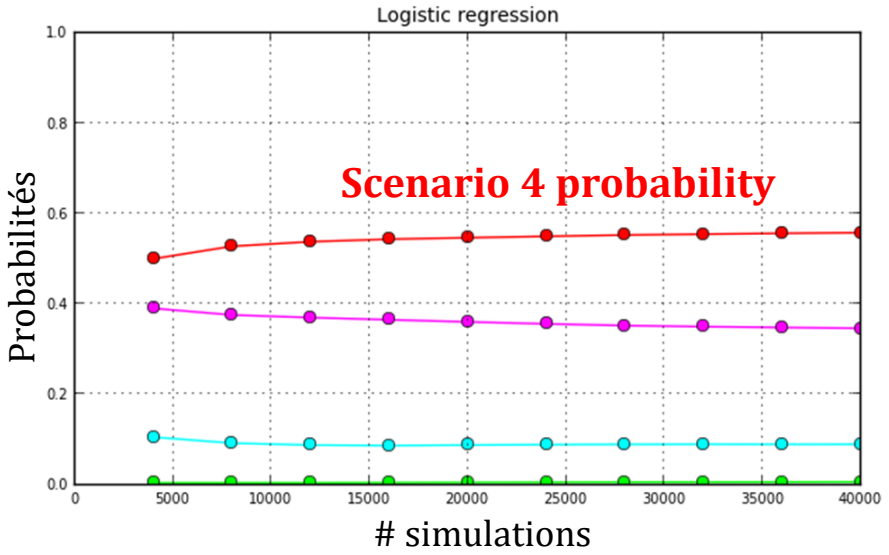


Effet goulot d'étranglement
~ 49 [7- 95] générations
(1 génération tous les 2 ans ou + court en milieu tropical?)

(Excoffier *et al.* 2005)

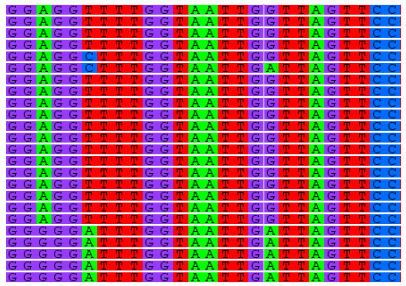


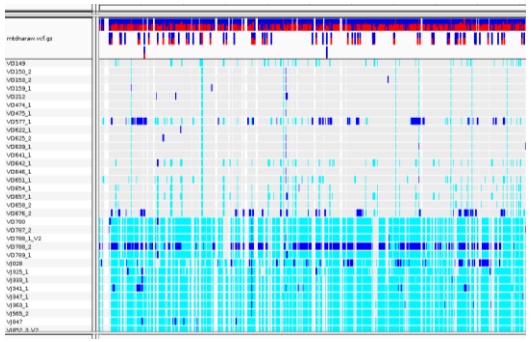
Dataset simulated under each scenario

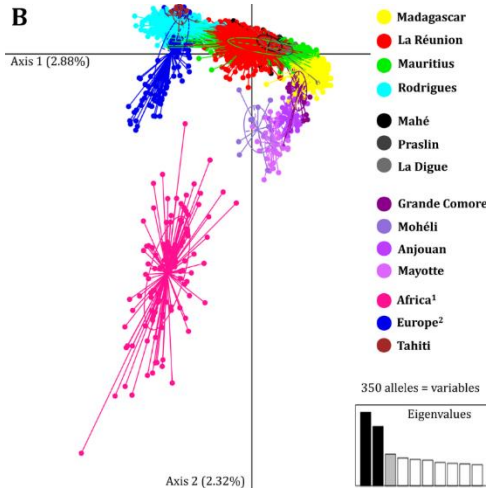


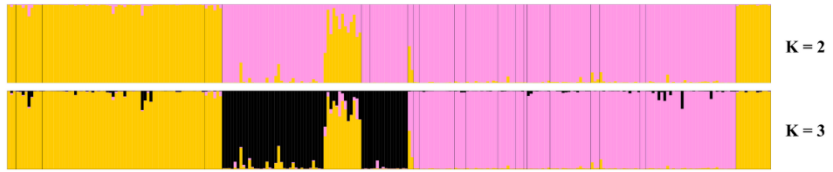
How demographic inference works?

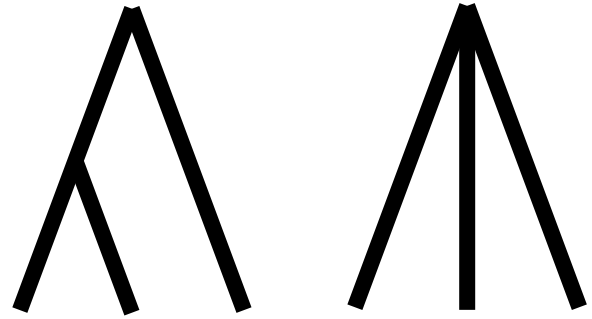
General workflow

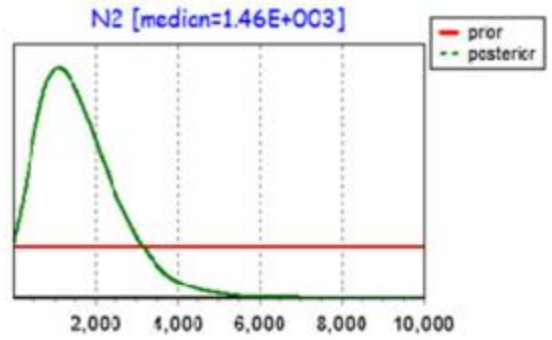










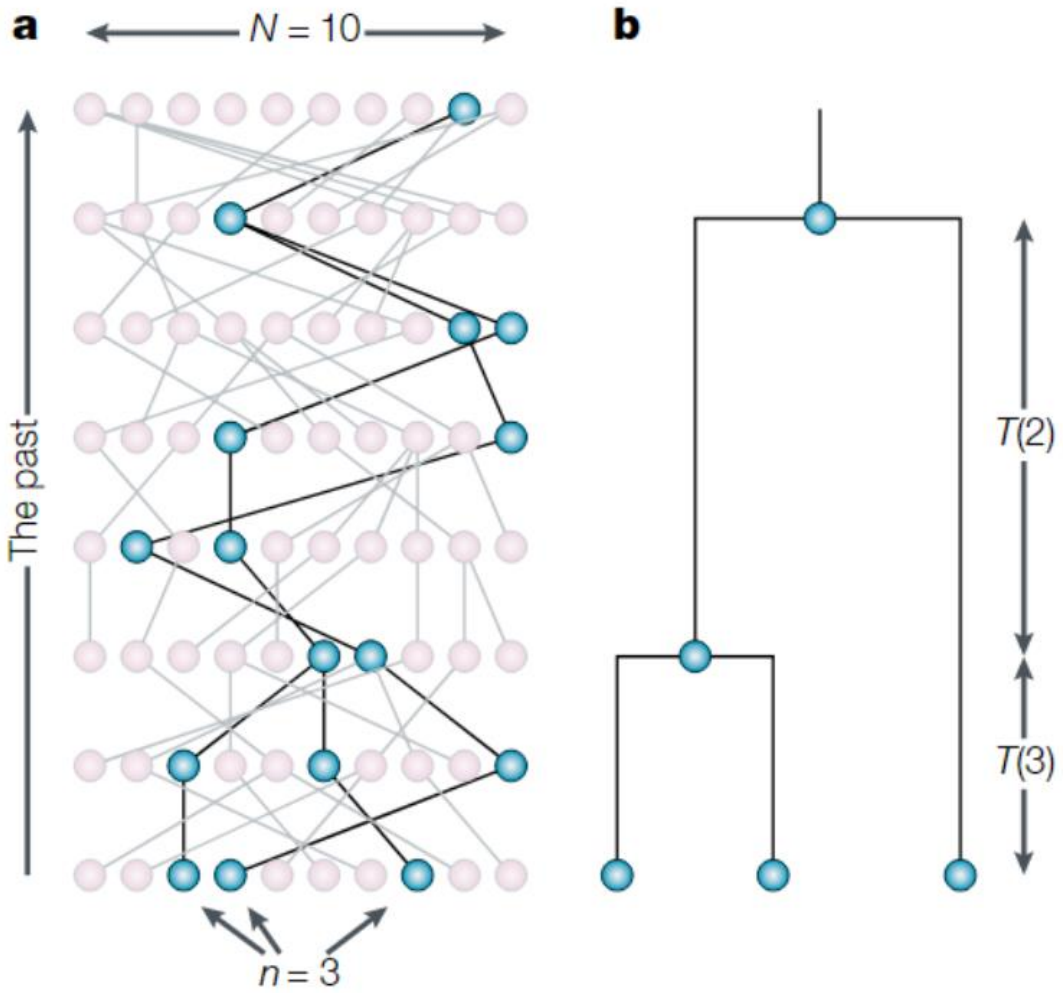


Genomic dataset summary

Model-free or clustering-model

Testing demographic scenarios Estimating parameters

Coalescent theory to simulate genealogies

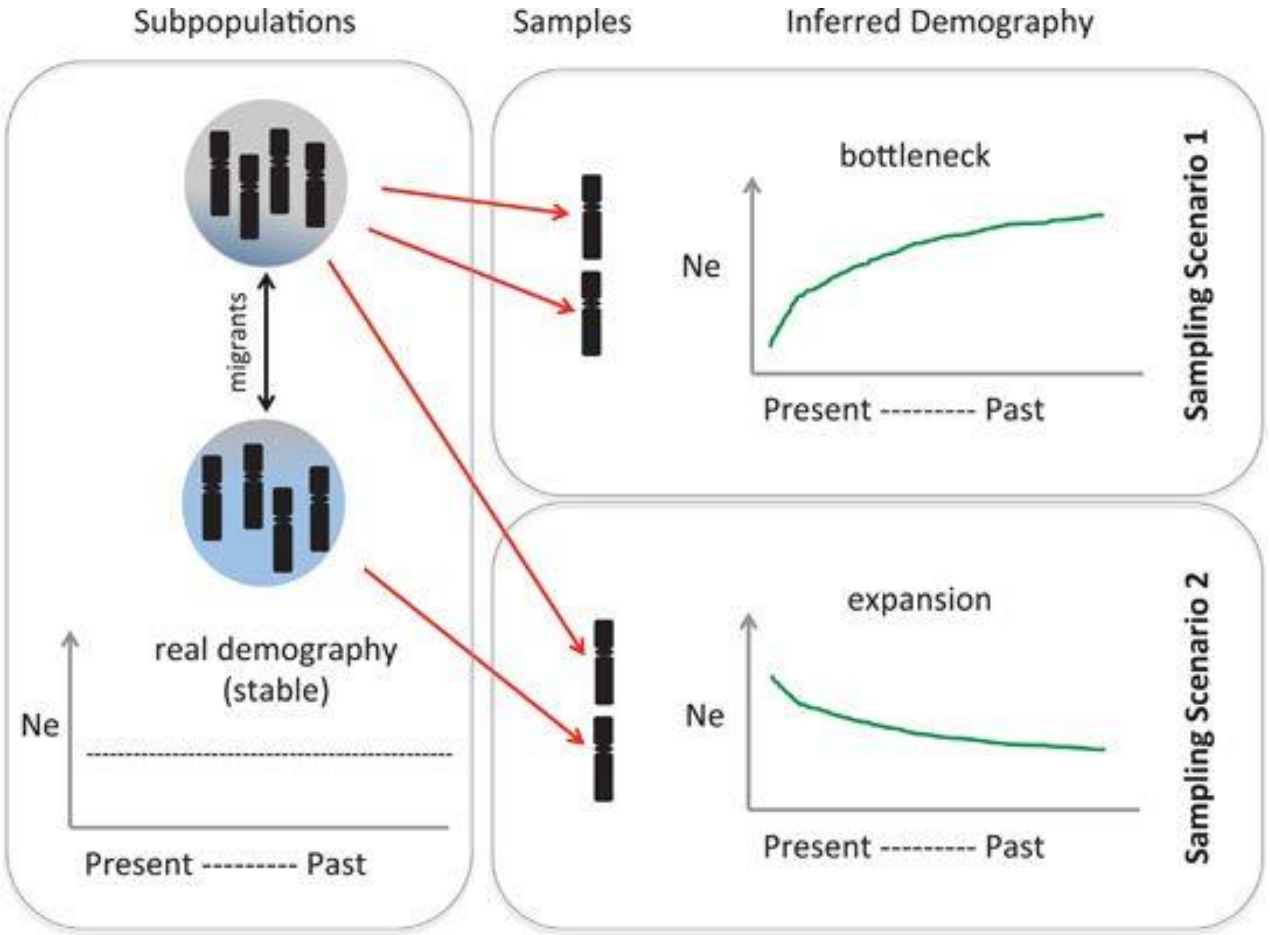


Rosenberg & Nordborg 2002

Retrospective approach that retrace backwards in time the ancestry of sampled genes according to specific demographic scenario

Kingman 1982, Hudson 1983, Tajima 1983

Prior knowledge on your dataset is essential

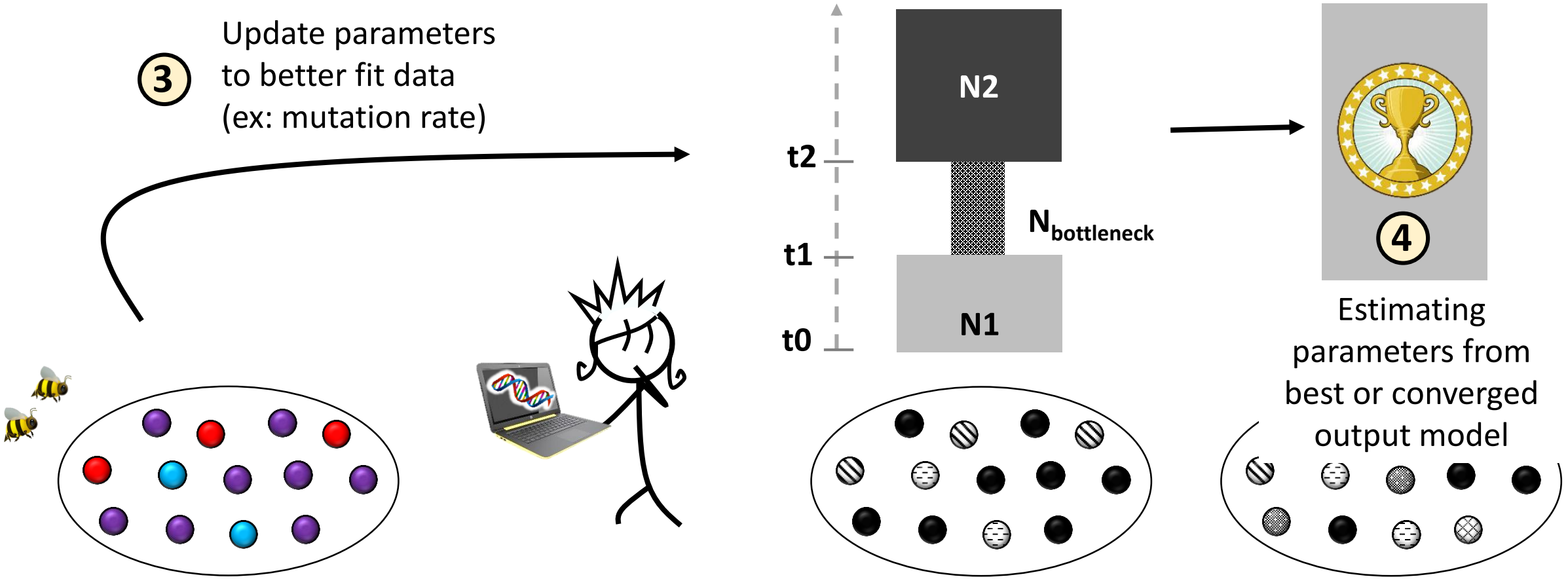


According to Mazet et al. 2016, from Orozco-ter Wengel 2016

Population structure can mimics the same signal as bottleneck

➔ Necessity to assess population structure beforehand with clustering or model-free method (PCA)

How to choose the best model and estimate parameters

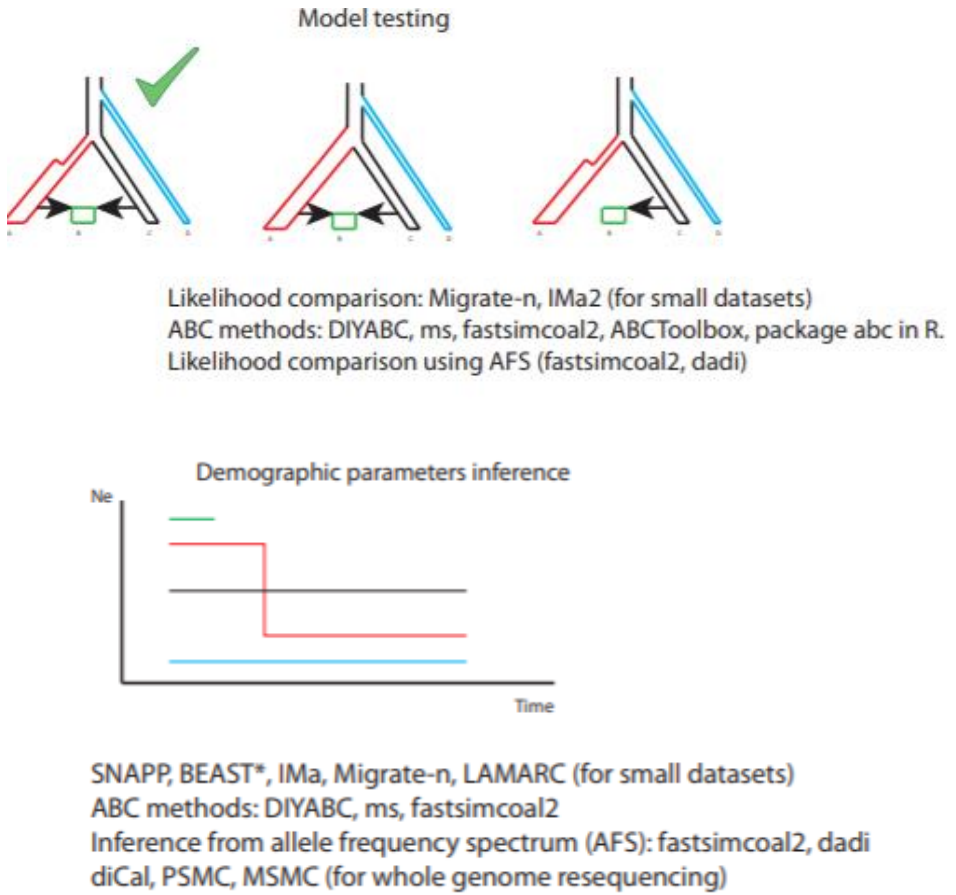


Reviews on different methods to simulate your data

Table 1 | **Software for demographic inferences**

Name	Data type	Inference	Notes	Refs
dadi	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Requires some Python-coding skills; applicable to up to three populations	60
Fastsimcoal2	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Can also be used to simulate data under the SMC	62,63
Treemix	Frequencies of unlinked bi-allelic SNVs	Admixture graph	Highly multimodal likelihood surface and heuristic search; redo inference from many starting points	64
fastNeutrino	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Applicable only to a single population; designed specifically for extremely large sample sizes	65
DoRIS	Lengths of IBD blocks between pairs of individuals	Demographic history	IBD must be inferred (for example, using Beagle or GERMLINE); specification of lower cut-off minimizes false-negative IBD tracts	71,72
IBS tract inference	Lengths of IBS blocks between pairs of individuals	Demographic	IBS can easily be confounded by missing data and/or sequencing errors	76
PSMC	Diploid genotypes from one individual	Demographic history	Best used in MSMC's PSMC mode, which uses the SMC to more accurately model recombination than the original PSMC; applicable to a single population	78
MSMC	Whole genome, phased haplotypes	Demographic history	Requires large amounts of RAM; cross-coalescence rate should not be interpreted as migration rate	82
CoalHMM	Whole genome, phased haplotypes	Demographic history	Multiple applications, including inference of population sizes, migration rates and incomplete lineage sorting	83–87
diCal	Medium-length, phased haplotypes	Demographic history	Uses shorter sequences than MSMC, but can be applied to multiple individuals in complex demographic models; infers explicit population genetic parameters for migration rates	89,92
LAMARC	Short, phased haplotypes	Demographic history	Requires Monte Carlo sampling of coalescent genealogies; very flexible	93
BEAST	Short, phased haplotypes	Species trees, effective population sizes	Used mainly as a method of phylogenetic inference. Can also infer population size history	94
MCMCcoal	Short, phased haplotypes	Divergence times between populations	Now incorporated into the software BPP[131]	95
G-PhoCS	Short, (un)phased haplotypes	Demographic history	Incorporates migration into the MCMCcoal framework. Averages over unphased haplotypes	96
Exact likelihoods using generating functions	Short, phased haplotypes	Demographic history	Implemented in Mathematica; applicable only to specific classes of multi-population models	97,98

Methods and models for unravelling human evolutionary history (2015) Schraiber & Akey Nature Review Genetics

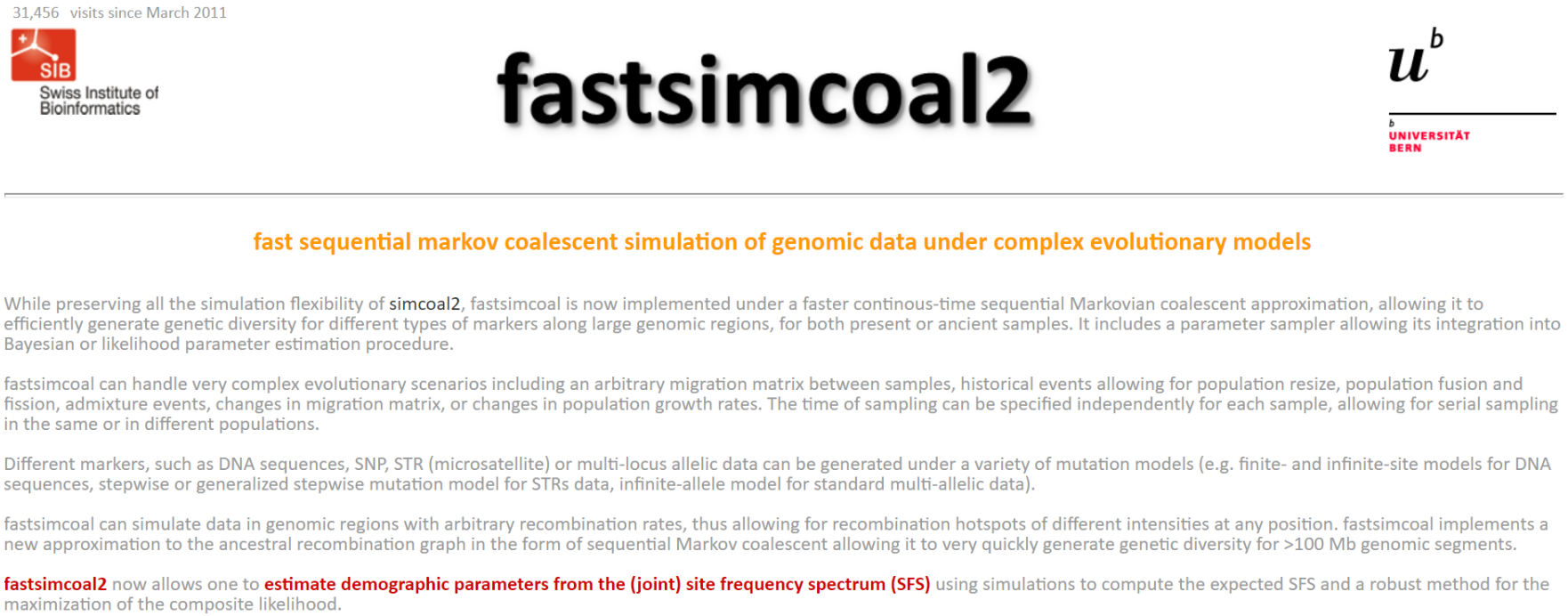


Going down the rabbit hole: a review on how to link genome-wide data with ecology an evolution in natural populations (2017) Bourgeois *et al.* bioRxiV

Time to learn with a practical session using fastsimcoal2

Why fastsimcoal2

The program fastsimcoal2 can handle different data type (Excoffier and Foll 2011, Excoffier et al. 2013) microsatellites, DNA sequences, SNPs, **SFS**

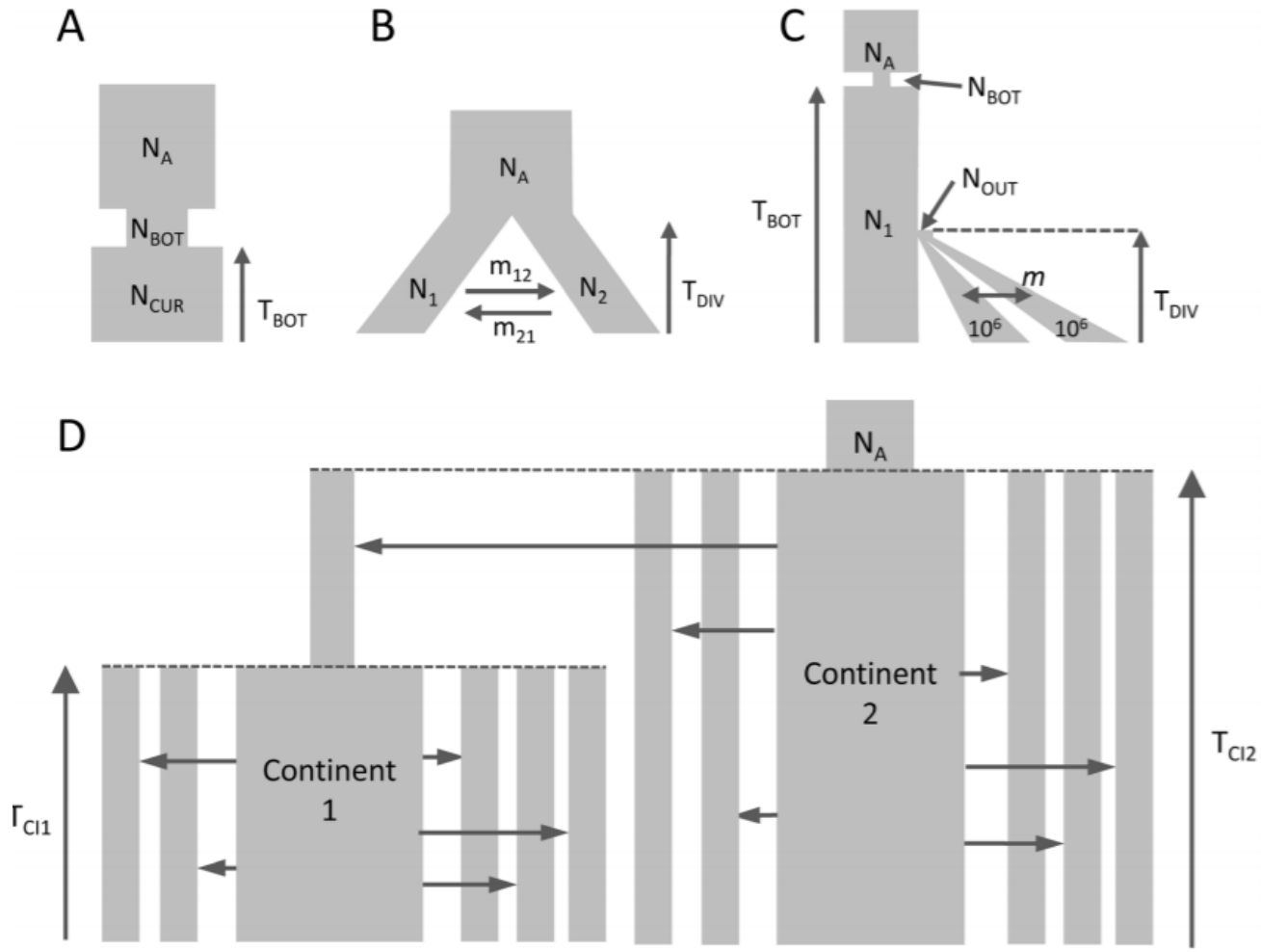


Test best scenario looking at likelihood + estimate parameters using coalescent simulations (compare to dadi which is diffusion-based)

Fastsimcoal2 can handle very complex scenarios

Very flexible but what you should know:

- Estimation will depend on the **mutation rate**
- Have an idea of your organism generation time
- Avoid structure within your population



Workflow for fastsimcoal2 based on SFS (fsc26)

1. Obtaining the observed SFS from your genomic data
2. Design the different demographic scenarios (create .tpl and .est files)
3. Performs simulations and repeat 50-150 independent runs
4. For each scenario select the run with the best likelihood
5. Check and draw your best scenario to see any mistake in your design
6. Calculate the AIC Akaike Information criteria for each model
7. Create pseudo-observed dataset from the best .par file and repeat step 3 to get confidence interval

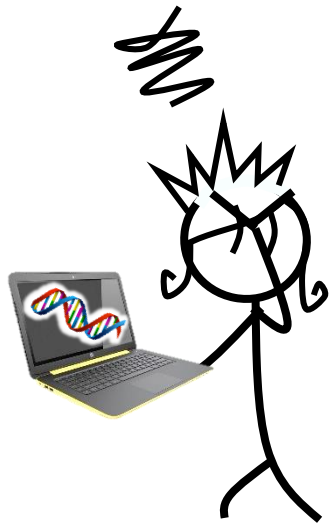
Create your working folder for today

```
cd ~/workshop-iussi2018
ls
```

```
iussi@buxkhjkbs37le4:~/workshop-iussi2018$ ls
demography  README.md
```

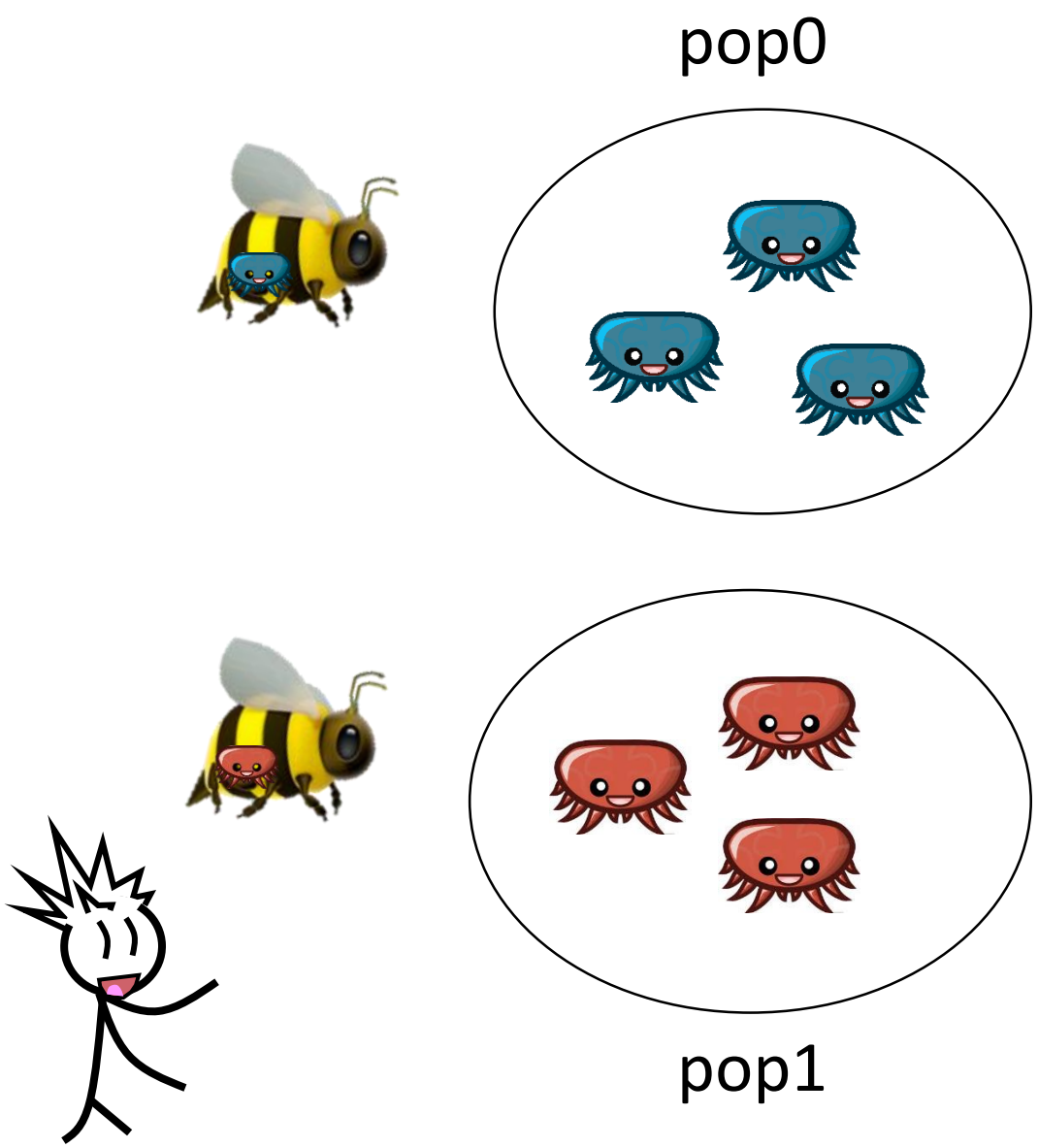
```
cp -r demography yourname
cd yourname
ls
```

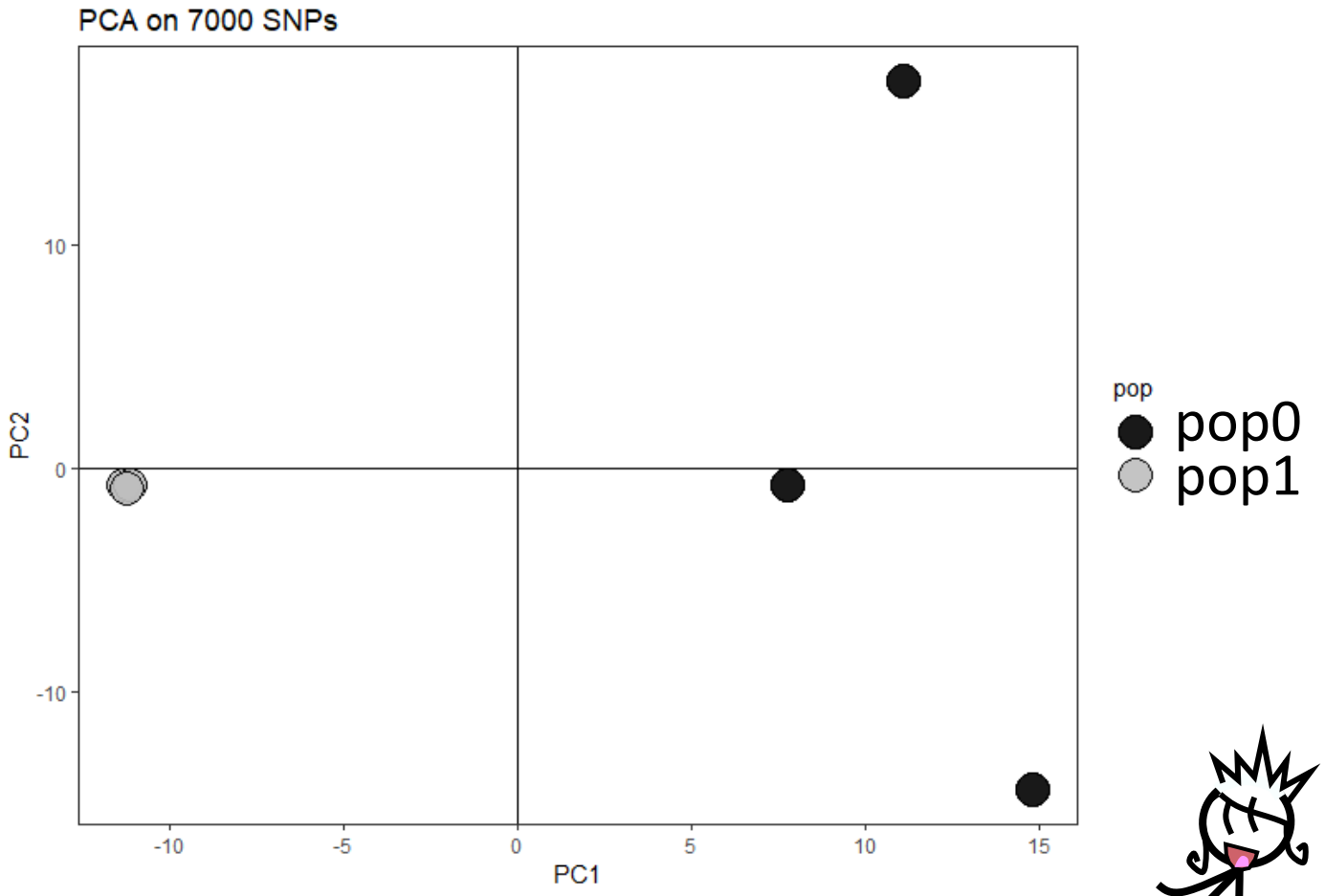
```
iussi@buxkhjkbs37le4:~/workshop-iussi2018$ cp -r demography maeva
iussi@buxkhjkbs37le4:~/workshop-iussi2018$ cd maeva
iussi@buxkhjkbs37le4:~/workshop-iussi2018/maeva$ ls
diyabc-2.1.0-linux32  easySFS  fsc26_linux64  raw
```



Every future step you will do it in your folder to avoid compromising original data for this session

Our data today is based on about 7000 bi-allelic SNPs





Workflow for fastsimcoal2 based on SFS (fsc26)

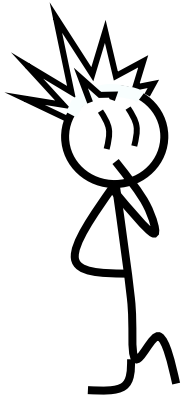
1. Obtaining the observed SFS from your genomic data
2. Design the different demographic scenarios (create .tpl and .est files)
3. Performs simulations and repeat 50-150 independent runs
4. For each scenario select the run with the best likelihood
5. Check and draw your best scenario to see any mistake in your design
6. Calculate the AIC Akaike Information criteria for each model
7. Create pseudo-observed dataset from the best .par file and repeat step 3 to get confidence interval

Site Frequency Spectrum (SFS) or also know as AFS

Help summarize **genome wide information (independent SNPs)** in only a few lines

	SNP1	SNP2	SNP3	SNP4	SNP5
	A	G	T	C	A
	T	G	T	C	G
	A	C	T	C	T
	A	C	A	T	C

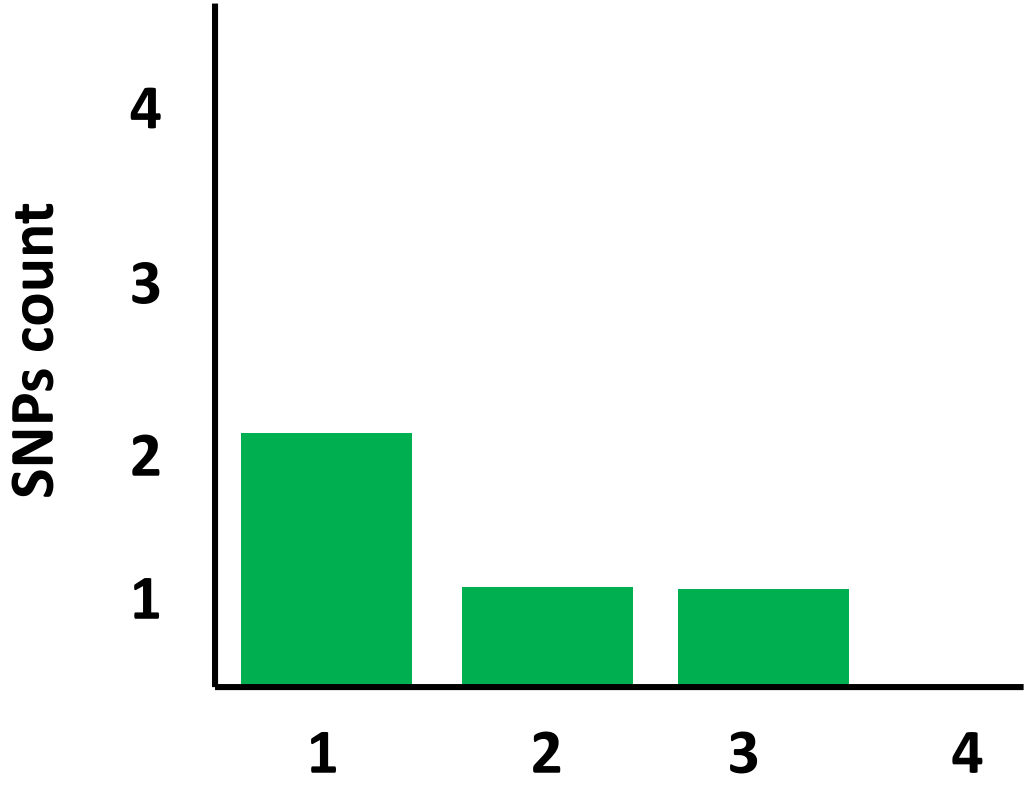
HOW?



Frequency	0	1	2	3
SNP count	0	3	1	1

You can do the same with genotypic data (1d SFS)

	SNP1	SNP2	SNP3	SNP4
	0	1	0	0
	1	0	0	1
	1	0	0	0
	1	0	2	0
	0	0	0	0

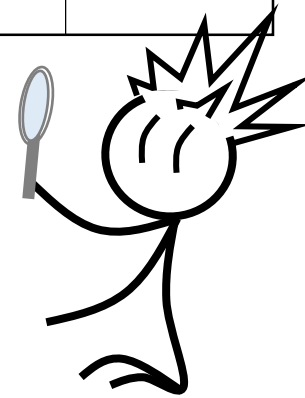


Genotype 0 means **homozygote for reference** allele
Genotype 1 means **heterozygote**
Genotype 2 means **homozygote for alternative** allele

Your turn! Try to calculate the derived SFS from genotype

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12
	0	1	0	0	0	0	2	1	0	0	0	1
	1	0	0	1	1	0	0	0	1	0	0	0
	1	0	0	0	1	1	0	1	0	1	2	1
	0	0	1	0	1	1	0	1	0	0	0	0
	0	0	0	0	0	0	0	1	0	2	0	0

Genotype 0 means **homozygote for reference** allele
Genotype 1 means **heterozygote**
Genotype 2 means **homozygote for alternative** allele

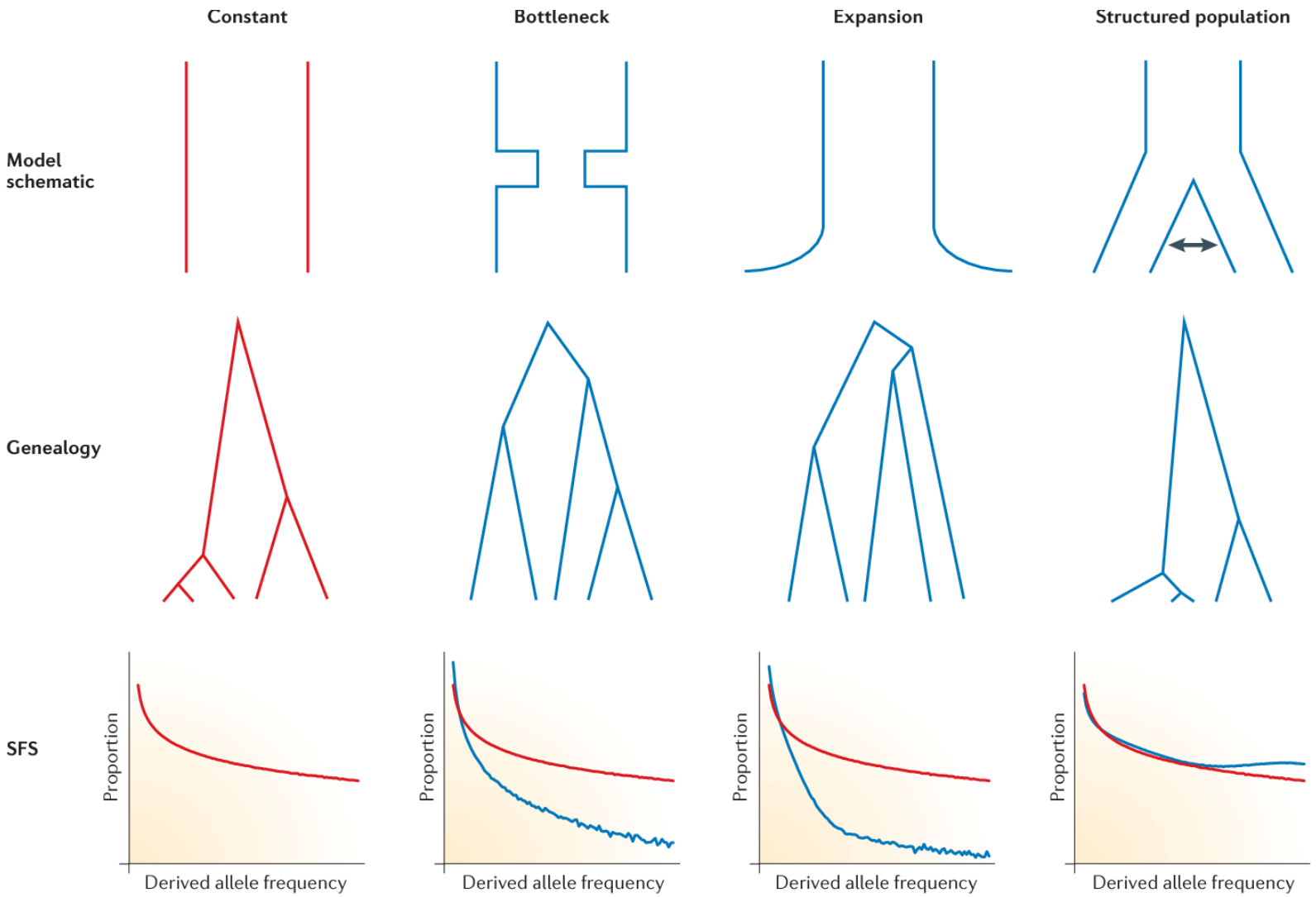


Your turn! Try to calculate the derived SFS from genotype

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12
	0	1	0	0	0	0	2	1	0	0	0	1
	1	0	0	1	1	0	0	0	1	0	0	0
	0	0	0	0	1	1	0	1	0	1	2	1
	0	0	1	0	1	1	0	1	0	0	0	0
	0	0	0	0	0	0	0	1	0	2	0	0

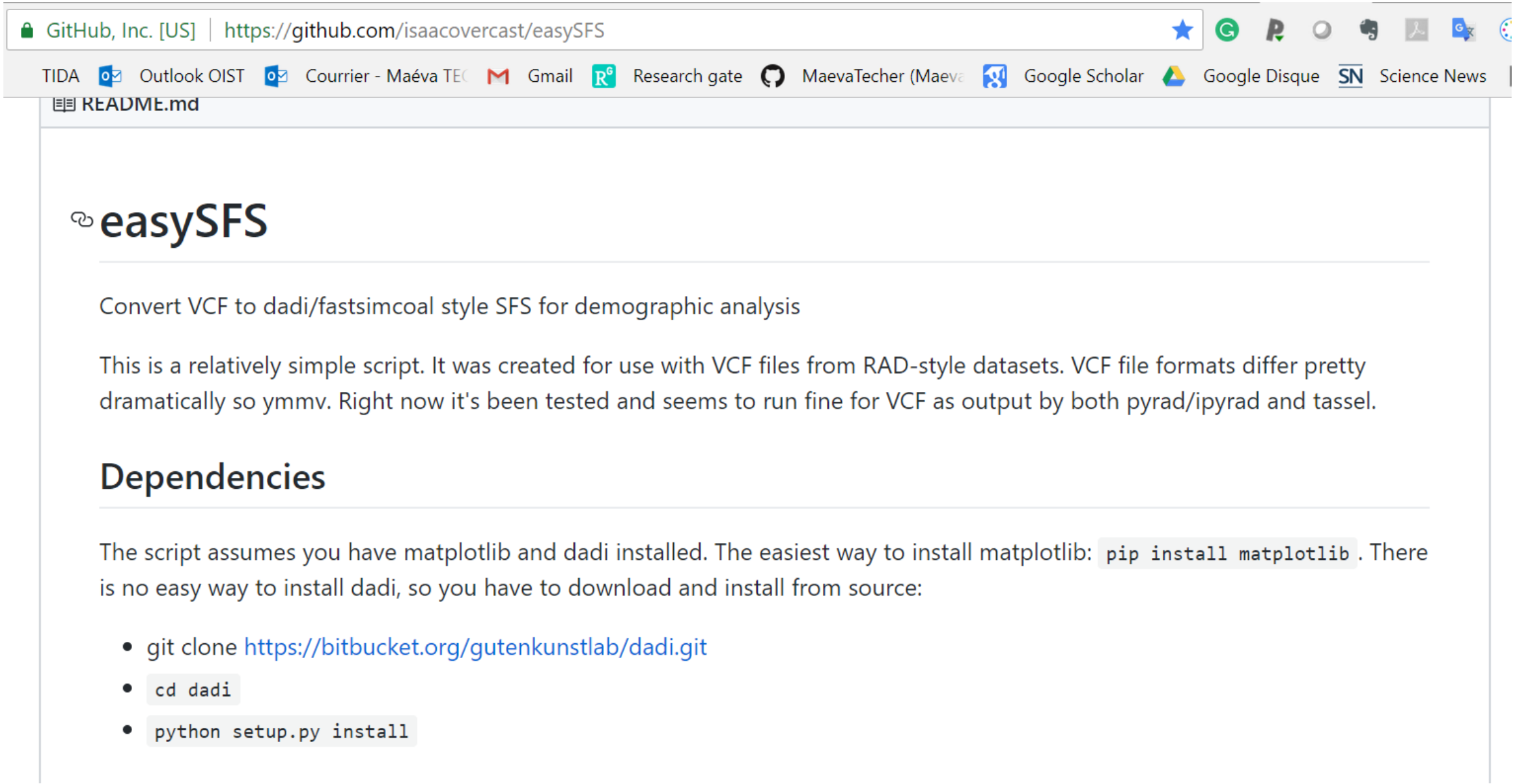
Frequency	1	2	3	4	5	6
SNP count	5	4	2	1	0	0

SFS can reflect the population demographic history



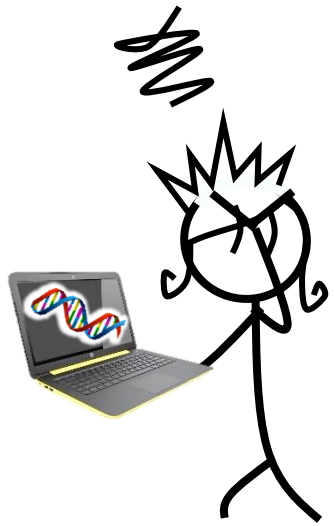
Methods and models for unravelling human evolutionary history (2015) Schraiber & Akey Nature Review Genetics

Several ways to compute your 1d or 2d SFS



Today the SFS was pre-computed from mighty data

```
[maeva-techer@sango10214:/work/MikheyevU/Maeva/workshop]$ /apps/unit/MikheyevU/Maeva/easySFS/easySFS.py
-i mighty.vcf -p dataanonym.txt -a --preview
Processing 2 populations - ['pop0', 'pop1']

    Running preview mode. We will print out the results for # of segregating sites
    for multiple values of projecting down for each population. The dadi
    manual recommends maximizing the # of seg sites for projections, but also
    a balance must be struck between # of seg sites and sample size.

    For each population you should choose the value of the projection that looks
    best and then rerun easySFS with the `--proj` flag.

pop0
(2, 625.0)      (3, 937.0)      (4, 1025.0)     (5, 658.0)


pop1
(2, 5.0)        (3, 7.0)        (4, 8.0)        (5, 5.0)


[maeva-techer@sango10214:/work/MikheyevU/Maeva/workshop]$ /apps/unit/MikheyevU/Maeva/easySFS/easySFS.py
-i mighty.vcf -p dataanonym.txt -a -o /work/MikheyevU/Maeva/workshop/output -f --proj=5,5
Processing 2 populations - ['pop0', 'pop1']
Doing 1D sfs - pop0
Doing 1D sfs - pop1
Doing 2D sfs - ('pop0', 'pop1')
Doing multiSFS for all pops
```



The easySFS output files for fastsimcoal2 are in `yourname/easySFS`

You should have 4 files in it

How do the easySFS output files look like?

Example of Minor Allele Frequency file (1d) where the minor allele is treated reference **(you can also compute DAF with ansgd)**

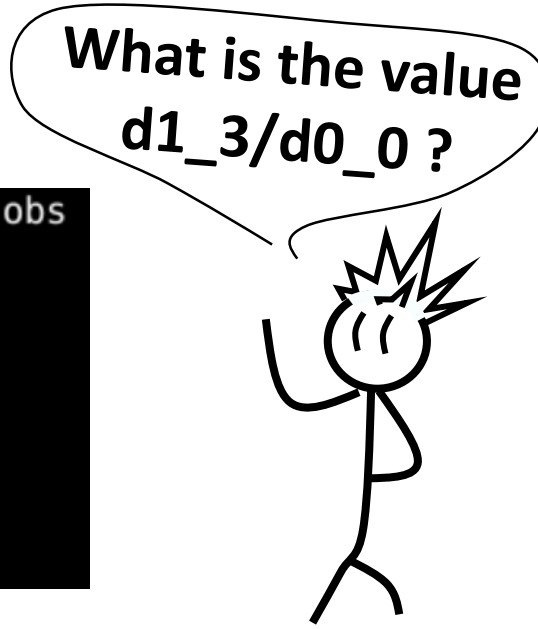
Inspect it by doing `cat pop0_MAF.obs`

```
iussi@buxkhjkbs37le4:~/workshop-iussi2018/demography/easySFS$ cat pop0_MAFpop0.obs
1 observation
d0_0    d0_1    d0_2    d0_3    d0_4    d0_5
5574 219.3333333333332 438.6666666666665 0 0 0
```

Example of joint Minor Allele Frequency file MAF (2d)

`cat mighty_jointMAFpop0_1.obs`

```
iussi@buxkhjkbs37le4:~/workshop-iussi2018/demography/easySFS$ cat mighty_jointMAFpop0_1.obs
1 observation
        d0_0    d0_1    d0_2    d0_3    d0_4    d0_5
d1_0    5044 0 0 0.6666666666666665 0.3333333333333332 60.5
d1_1    143.3333333333333 0 0 0 0 0
d1_2    286.6666666666666 0 0 0 0 0
d1_3    [redacted] 0 0 0 0 0
d1_4    64.66666666666664 0 0 0 0 0
d1_5    60.5 0 0 0 0 0
```



We will use the joint MAF file for two populations

```
cd ..
mkdir 2popdiv
mkdir 2popIM
```

```
cp  easySFS/mighty_jointMAFpop0_1.obs 2popdiv
```

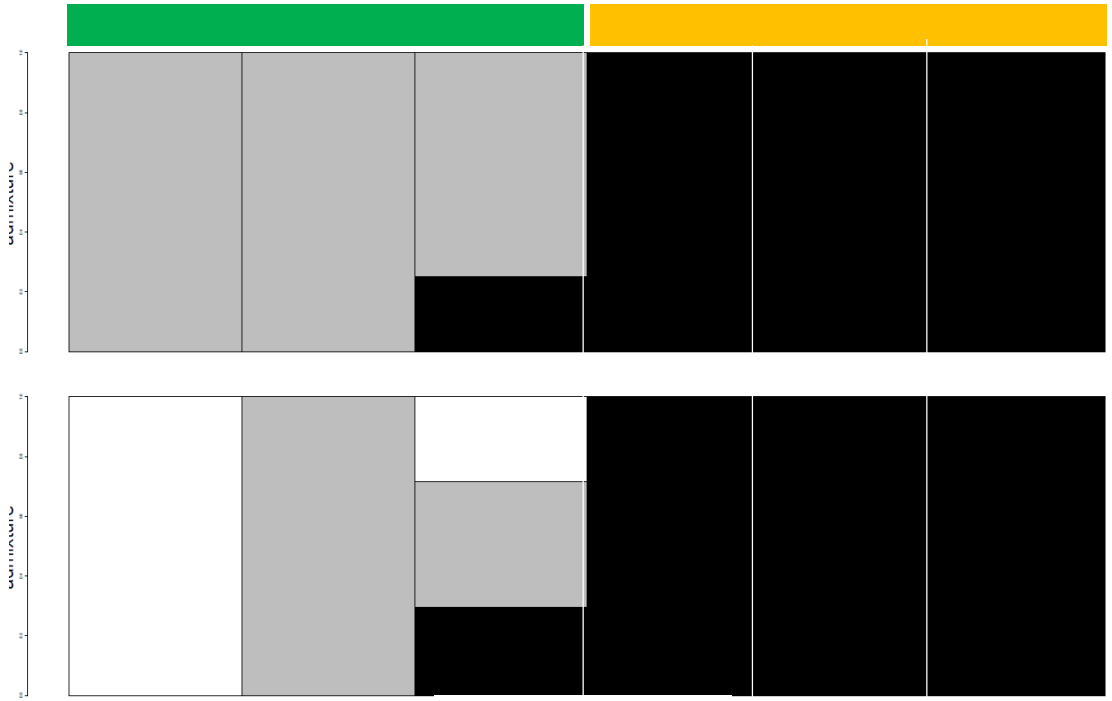
Do the same with the folder IM, this will be your file to work with

Workflow for fastsimcoal2 based on SFS (fsc26)

1. Obtaining the observed SFS from your genomic data
2. Design the different demographic scenarios (create .tpl and .est files)
3. Performs simulations and repeat 50-150 independent runs
4. For each scenario select the run with the best likelihood
5. Check and draw your best scenario to see any mistake in your design
6. Calculate the AIC Akaike Information criteria for each model
7. Create pseudo-observed dataset from the best .par file and repeat step 3 to get confidence interval

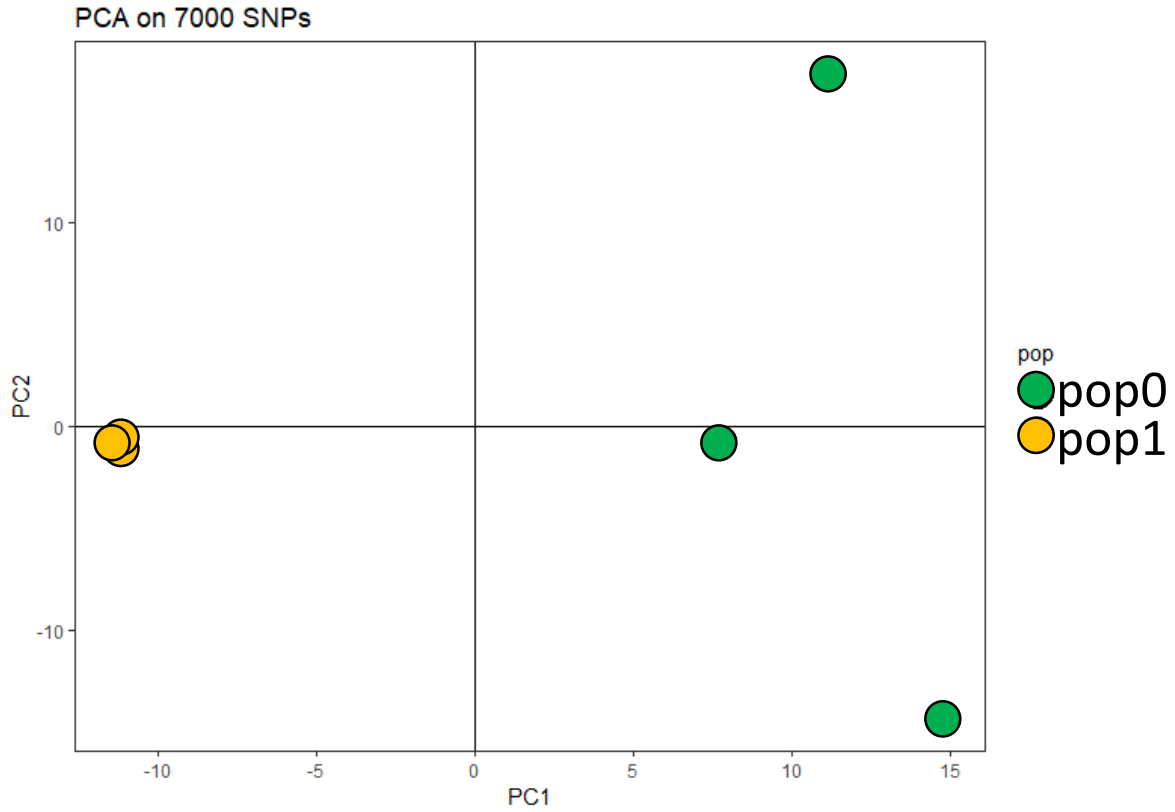
Did our populations follow a strict divergence or IM?

If we had more time we would run NGSAdmix but important thing is that population seems to be admixed



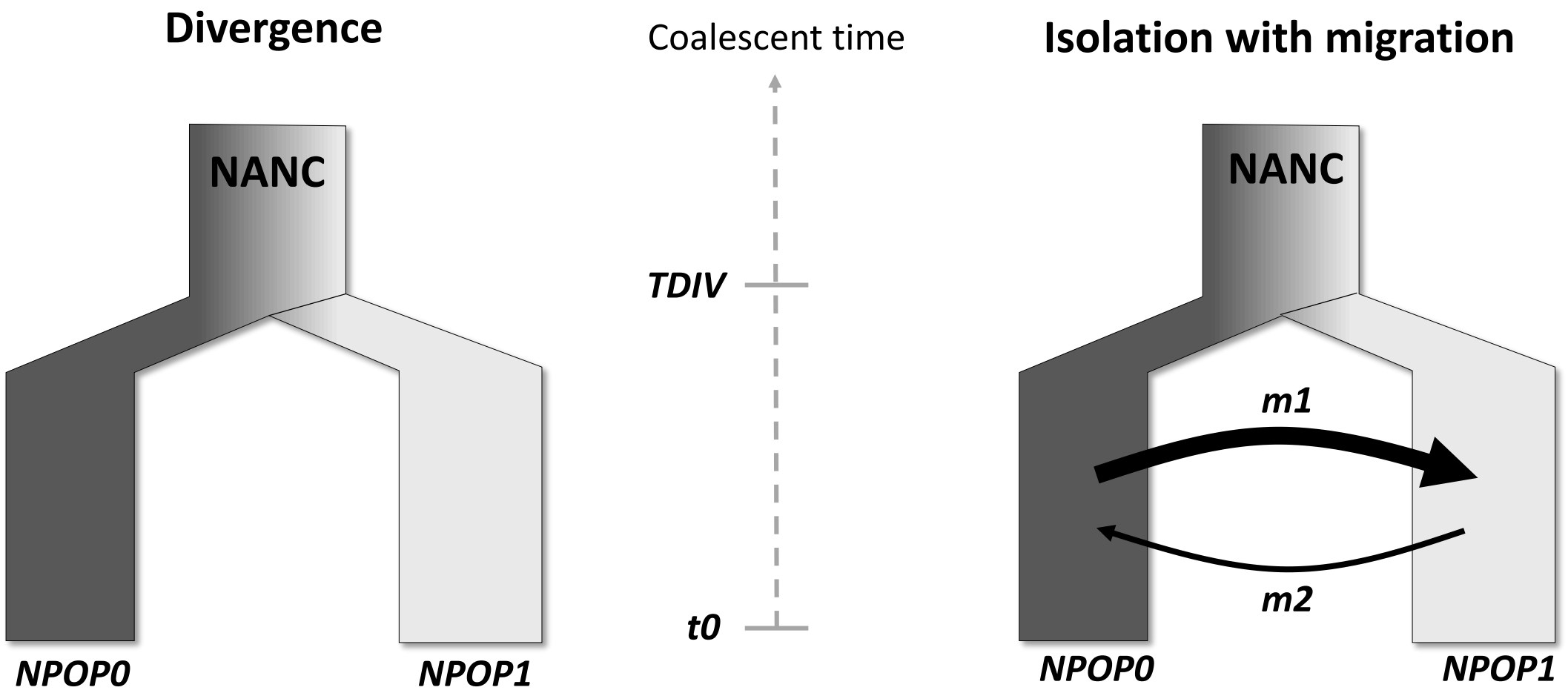
K2

K3




```
NGSadmix -likes mighty.BEAGLE.GL -K 2 -outfiles mightyK2 -minMaf 0.1
```

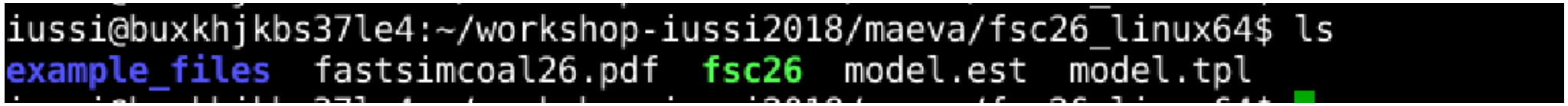
Two scenarios to design



You can set the scenario with a .tpl file

Take a look at model files .tpl and .est from fastsimcoal2 folder

```
cd yourname/fsc26_linux64
ls
```



You will have to edit the file using vim

```
cp model.tpl 2popdiv/2popdiv.tpl
cp model.est 2popdiv/2popdiv.est
vi 2popvd.tpl
:x (save) or :q! (quit without saving)
i (insert text)
```

Editing the template file (example here just to illustrate)

Here our populations NPOP0 and NPOP1

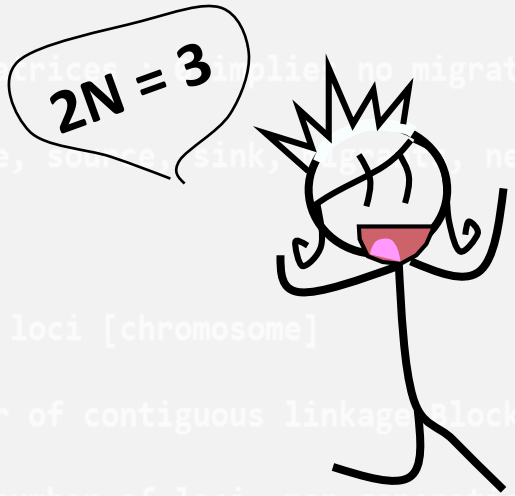
One line each →

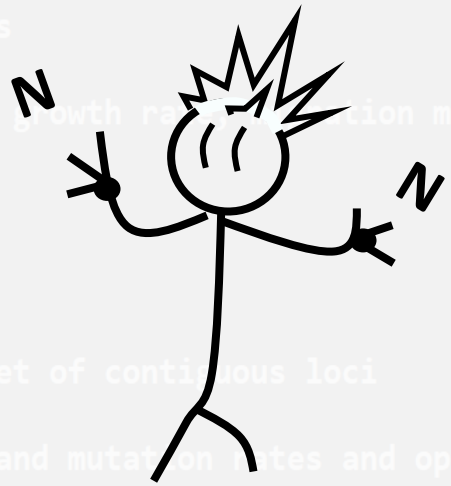
```
[maeva-techer@sango-login2:/work/MikheyevU/Maeva/demography/output/2vdpop-div/realdata]$ cat 2vdpop-di
v72.tpl
//Parameters for the coalescence simulation program : fsimcoal2.exe
2 samples to simulate :
//Population effective sizes (number of genes)
N_VdAc0
N_VdAm1
//Samples sizes and samples age
9
9
//Growth rates  : negative growth implies population expansion
0
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 ind
ex
1  historical event
TDIV 1 0 1 RESIZE0 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 pa
rameters
FREQ  1   0   2.5e-8 OUTEXP
```

Editing the template file (example here just to illustrate)

Sample size is N gene copy so haploid size **We have 3 diploid individuals for POP0 and 3 for POP1**

```
[maeva-techer@sango-login2:/work/MikheyevU/Maeva/demography/output/2vdpop-div/realdata]$ cat 2vdpop-di
v72.tpl
//Parameters for the coalescence simulation program : fsimcoal2.exe
2 samples to simulate :
//Population effective sizes (number of genes)
N_VdAc0
N_VdAm1
//Samples sizes and samples age
9
9
//Growth rates  : negative growth implies population expansion
0
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 ind
ex
1  historical event
TDIV 1 0 1 RESIZE0 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 pa
rameters
FREQ  1   0   2.5e-8 OUTEXP
```





Editing the template file (example here just to illustrate)

Growth rate = 0 means the model do not consider population expansion
We choose this simple assumption here

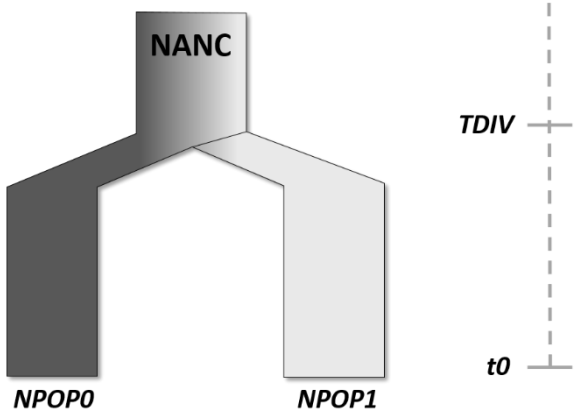
```
[maeva-techer@sango-login2:/work/MikheyevU/Maeva/demography/output/2vdpop-div/realdata]$ cat 2vdpop-di
v72.tpl
//Parameters for the coalescence simulation program : fsimcoal2.exe
2 samples to simulate :
//Population effective sizes (number of genes)
N_VdAc0
N_VdAm1
//Samples sizes and samples age
9
9
//Growth rates  : negative growth implies population expansion
0
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 ind
ex
1  historical event
TDIV 1 0 1 RESIZE0 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 pa
rameters
FREQ  1   0   2.5e-8 OUTEXP
```

Editing the template file (example here just to illustrate)

Here we consider a divergence without migration

```
[maeva-techer@sango-login2:/work/MikheyevU/Maeva/demography/output/2vdpop-div/realdata]$ cat 2vdpop-di
v72.tpl
//Parameters for the coalescence simulation program : fsimcoal2.exe
2 samples to simulate :
//Population effective sizes (number of genes)
N_VdAc0
N_VdAm1
//Samples sizes and samples age
9
9
//Growth rates  : negative growth implies population expansion
0
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 ind
ex
1  historical event
TDIV 1 0 1 RESIZE0 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 pa
rameters
FREQ  1   0   2.5e-8 OUTEXP
```

Editing the template file (example here just to illustrate)




```
[maeva-techer@sango-login2:/work/MikheyevU/Maeva/demography/output/2vdpop-div/realdata]$ cat 2vdpop-di
v72.tpl
//Parameters for the coalescence simulation program : fsimcoal2.exe
2 samples to simulate :
//Population effective sizes (number of genes)
N_VdAc0
N_VdAm1
//Samples sizes and samples age
9
9
//Growth rates  : negative growth implies population expansion
0
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 ind
ex
1  historical event
TDIV 1 0 1 RESIZE0 0 0
```

Historical event →

TIME	SOURCE	SINK	MIGRANTS	DEMESIZE	NEW GROWTH	MATRIX
TDIV	**1**	**0**	**1**	**RESIZE**	**0**	**0**

In the end your .tpl should look like this

```
//Number of population samples (demes)
2 samples to simulate
//Population effective sizes (number of genes)
NPOP0
NPOP1
//Sample sizes
6
6
//Growth rates  : negative growth implies population expansion
0
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
TDIV 1 0 1 RESIZE 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
FREQ 1 0 2.5e-8 OUTEXP
```

Let’s take a look at the .est file (no changes needed)

If you want you can change the range of prior here or transform unif to logunif

You can also set rule to guide the model for example (NPOP0 > NPOP1)

```
iussi@buxkhjkbs37le4:~/workshop-iussi2018/maeva/fsc26_linux64$ cat model.est
// Priors and rules file
// *********************

[PARAMETERS]
//#isInt? #name #dist.#min      #max
//all N are in number of haploid individuals
1  NPOP0     unif     1  100000   output
1  NPOP1     unif     1  100000     output
1  NANC   unif     10   10000   output
1  TDIV  unif        1   200000   output

[RULES]

[COMPLEX PARAMETERS]

0 RESIZE = NANC/NPOP0
```

THE TIME HAS COME, let's simulate!

```
ls yourname/2popdiv
```

```
iussi@buxkhjkbs37le4:~/workshop-iussi2018/maeva/2popdiv$ ls
2popdiv.est  2popdiv.tpl  mighty_jointMAFpop0_1.obs
```

Fastsimcoal2 needs the .obs file to have the same name as .tpl and .est and expect the end to be _jointMAFpop1_0.obs

```
mv mighty_jointMAFpop0_1.obs 2popdiv_jointMAFpop1_0.obs
```

Try to call fastsimcoal2 from this folder

```
../fsc26_linux64/fsc26 -help
```

Take quickly a look at the different options

Workflow for fastsimcoal2 based on SFS (fsc26)

1. Obtaining the observed SFS from your genomic data
2. Design the different demographic scenarios (create .tpl and .est files)
3. Performs simulations and repeat 50-150 independent runs
4. For each scenario select the run with the best likelihood
5. Check and draw your best scenario to see any mistake in your design
6. Calculate the AIC Akaike Information criteria for each model
7. Create pseudo-observed dataset from the best .par file and repeat step 3 to get confidence interval

THE TIME HAS COME, let's simulate!

```
../fsc26_linux64/fsc26
--tplfile 2popdiv.tpl
--estfile 2popdiv.est
-m
--numsims 10000 (normally 1 000 000)
-M
--minnumloops 2 (normally 10)
--numloops 10 (normally 50)
-c 10
```

← Use or compute the Minor Site Frequency spectrum

← Perform parameter estimation by max likelihood from SFS

← Optimization steps

Let’s check the output

You should get after the run a new folder **2popdiv** and a new **2popdiv.par file**

```
iussi@buxkhjkbs37le4:~/workshop-iussi2018/maeva/2popdiv$ ls
2popdiv  2popdiv.est  2popdiv_jointMAFpop1_0.obs  2popdiv.par  2popdiv.tpl  seed.txt
```

Look at the output folder

```
iussi@buxkhjkbs37le4:~/workshop-iussi2018/maeva/2popdiv/2popdiv$ ls
2popdiv_1.simparam  2popdiv.brent_lhoods        2popdiv_maxL.par
2popdiv.bestlhoods  2popdiv_jointMAFpop1_0.txt  2popdiv.pv
```

```
cat 2popdiv.bestlhoods
```

```
iussi@buxkhjkbs37le4:~/workshop-iussi2018/maeva/2popdiv/2popdiv$ cat 2popdiv.bestlhoods
NPOP0   NPOP1   NANC    TDIV    RESIZE  MaxEstLhood     MaxObsLhood
66185   2265    3498    139031  0.0528519       -226.055        -145.007
```

Workflow for fastsimcoal2 based on SFS (fsc26)

1. Obtaining the observed SFS from your genomic data
2. Design the different demographic scenarios (create .tpl and .est files)
3. Performs simulations and repeat 50-150 independent runs
4. For each scenario select the run with the best likelihood
5. Check and draw your best scenario to see any mistake in your design
6. Calculate the AIC Akaike Information criteria for each model
7. Create pseudo-observed dataset from the best .par file and repeat step 3 to get confidence interval

If we have time, repeat this 5 more times and check output

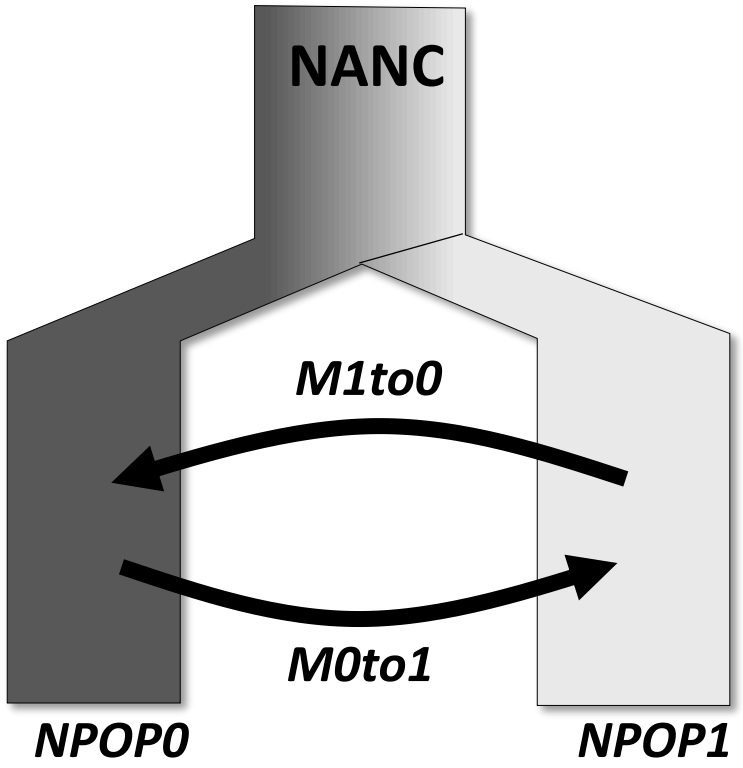
```
for i {1..5}; do cp 2popdiv.tpl 2popdiv”$i”.tpl; done
```

```
iussi@buxkhjkbs37le4:~/workshop-iussi2018/maeva/2popdiv$ ls
2popdiv       2popdiv2.tpl  2popdiv4.tpl  2popdiv.est                 2popdiv.par  seed.txt
2popdiv1.tpl  2popdiv3.tpl  2popdiv5.tpl  2popdiv_jointMAFpop1_0.obs  2popdiv.tpl
```

```
for i {1..5}; do cp 2popdiv.est 2popdiv”$i”.est; done
for i {1..5}; do cp 2popdiv_jointMAFpop1_0.obs
2popdiv”$i”_jointMAFpop1_0.obs; done


../fsc26_linux64/fsc26 --tplfile 2popdiv2.tpl --estfile 2popdiv2.est
-m --numsims 10000 -M –minnumloops 2 --numloops 10 -c 10
```

Choose the one run with the best MaxEstLhood

You build the .tpl and .est file for the 2popIM model now




```
cp 2popdiv.tpl 2popIM.tpl
cp 2popdiv.est 2popIM.est
vi 2popIM.tpl
```

```
//Number of migration matrices : 0 implies no
migration between demes
2
//Migration matrix 0
0 M1to0
M0to1 0
//Migration matrix 1: No migration
0 0
0 0
```

You build the .tpl and .est file for the 2popIM model now

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

[PARAMETERS]
//#isInt? #name   #dist.#min  #max
//all Ns are in number of haploid individuals
1  NPOP0       unif  1   1000000   output
1  NPOP1       unif  1   1000000   output
1  NANC        unif  10   1000000   output
1  TDIV        unif  10   2000000   output
0  N0M1        logunif  1e-2   2000   output
0  N1M0        logunif  1e-2   2000   output
[RULES]

[COMPLEX PARAMETERS]

0  RESIZE    = NANC/NPOP0      output
0  M0to1   = N0M1/NPOP0      output
0  M1to0   = N1M0/NPOP1      output
```

Simulate for the 2popIM ??

```
../fsc26_linux64/fsc26

--tplfile 2popIM.tpl

--estfile 2popIM.est

-m

--numsims 10000 (normally 1 000 000)

-M

--minnumloops 2 (normally 10)

--numloops 10 (normally 50)

-c 10
```

← Use or compute the Minor Site Frequency spectrum

← Perform parameter estimation by max likelihood from SFS

← Optimization steps

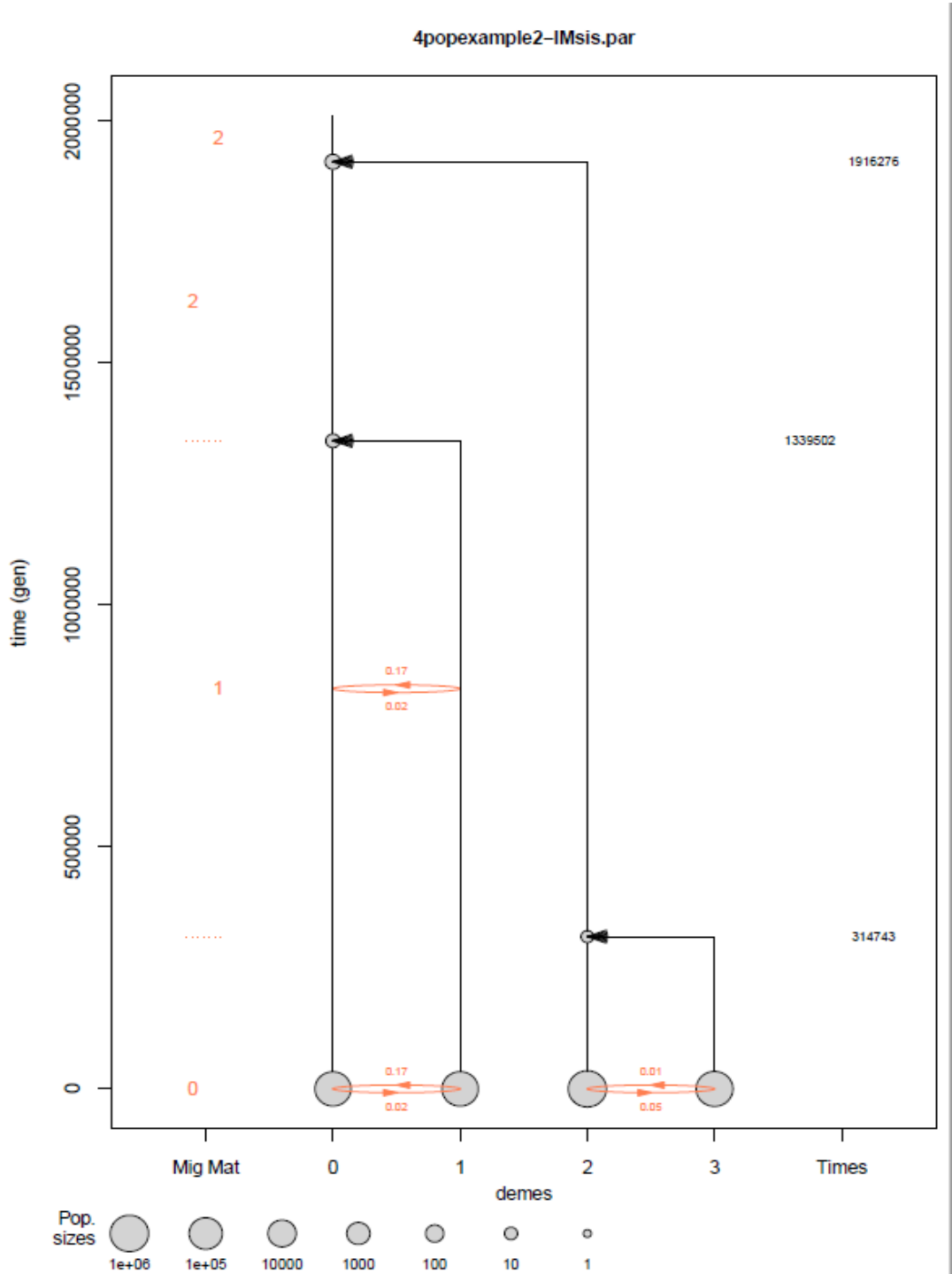
Workflow for fastsimcoal2 based on SFS (fsc26)

1. Obtaining the observed SFS from your genomic data
2. Design the different demographic scenarios (create .tpl and .est files)
3. Performs simulations and repeat 50-150 independent runs
4. For each scenario select the run with the best likelihood
5. Check and draw your best scenario to see any mistake in your design
6. Calculate the AIC Akaike Information criteria for each model
7. Create pseudo-observed dataset from the best .par file and repeat step 3 to get confidence interval

Visualizing the output estimates

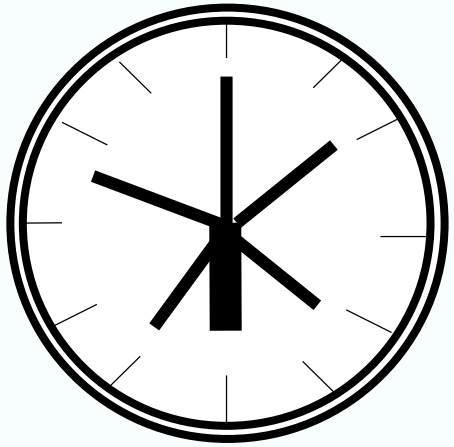
ParFileInterpreter-v6.3.1.r is an R script developed by V. Sousa and L. Excoffier (on fastsimcoal2 site)

You only need to import your SCENARIO_maxL.par and the script will draw the scenario associated with the best estimates for you!



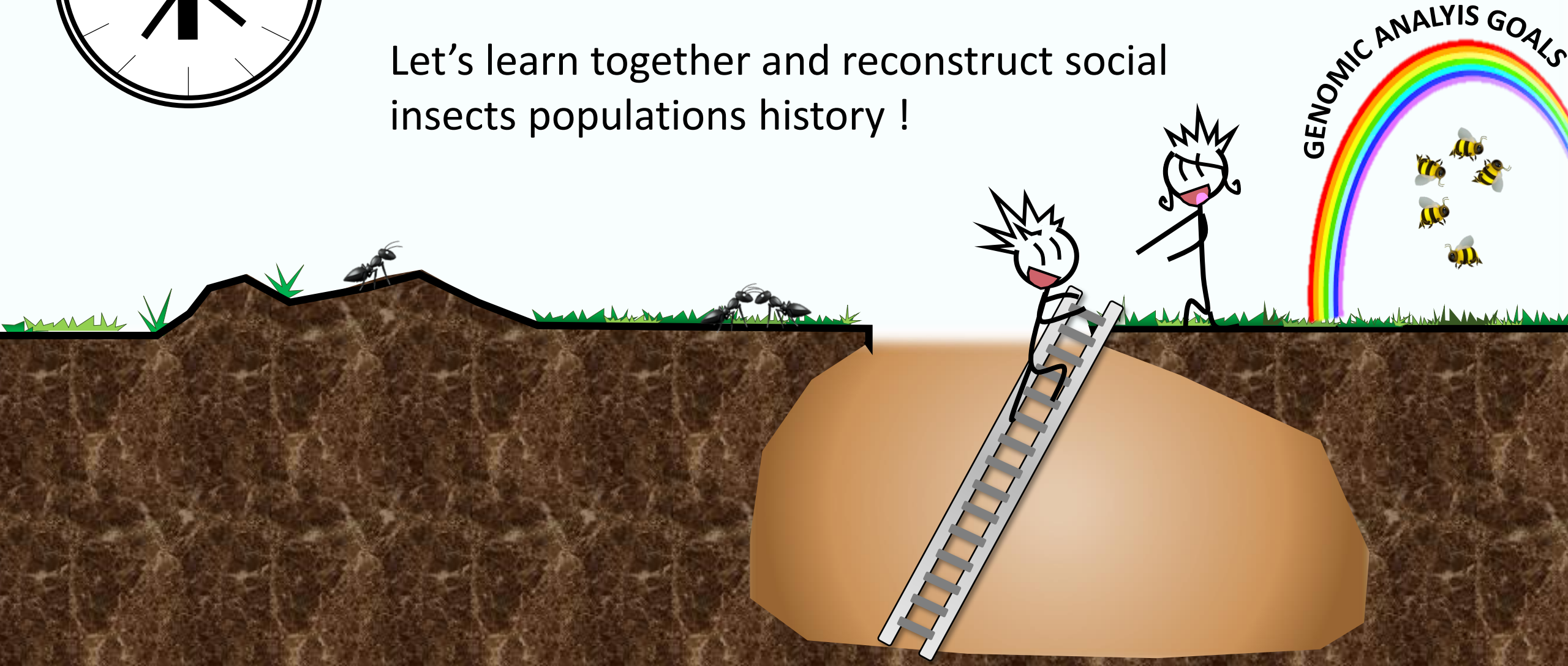
Workflow for fastsimcoal2 based on SFS (fsc26)

1. Obtaining the observed SFS from your genomic data
2. Design the different demographic scenarios (create .tpl and .est files)
3. Performs simulations and repeat 50-150 independent runs
4. For each scenario select the run with the best likelihood
5. Check and draw your best scenario to see any mistake in your design
6. Calculate the AIC Akaike Information criteria for each model
7. Create pseudo-observed dataset from the best .par file and repeat step 3 to get confidence interval

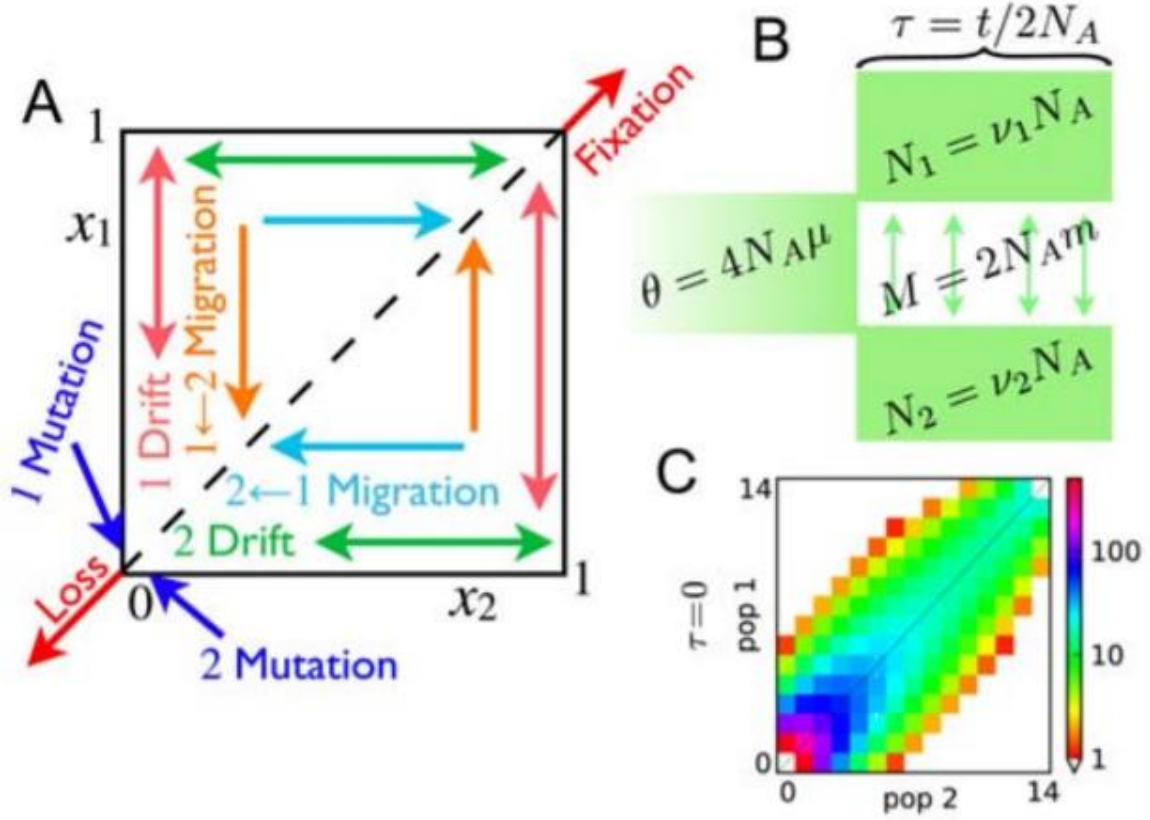


Time's up! So much more to say so don't hesitate to ask questions during the conference or after at maeva.techer@oist.jp

Let's learn together and reconstruct social insects populations history !



SFS can reflect the population demographic history



Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data (2009) Gutenkunst et al. Plos Genet