EMBO Practical course
Population genomics Background and tools

Gene-genealogy methods for demography & Approximate Bayesian Computation - ABC

May 25, 2017 – Napoli

Mathias Currat

Department of Genetics and Evolution – Anthropology Unit
University of Geneva, Switzerland

UNIVERSITÉ
DE GENÈVE

Unité d'Anthropologie
Département GENEV

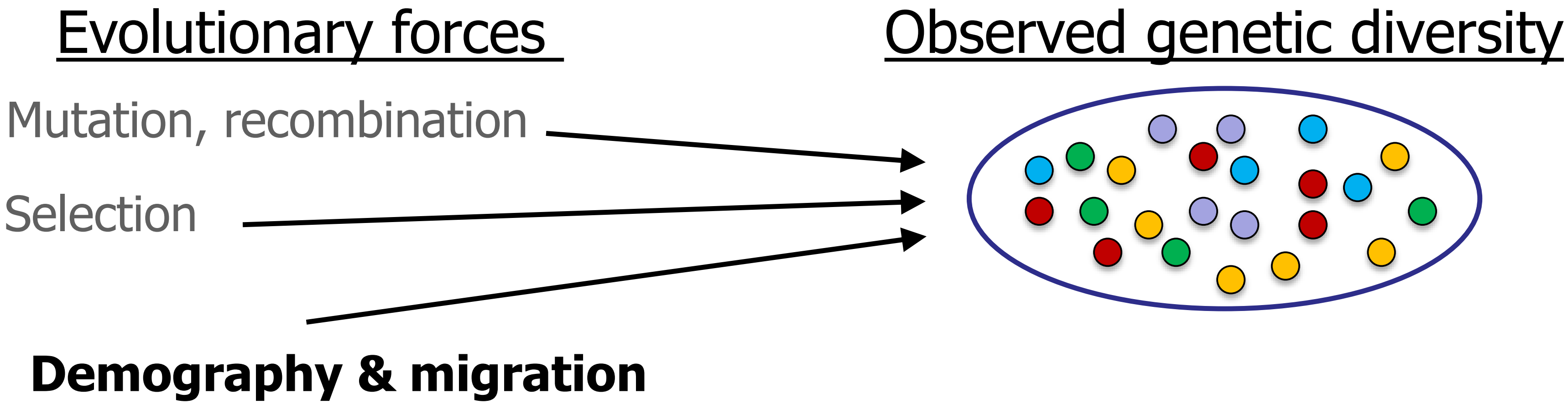
EMBO
excellence in life sciences

Outline

1. Genetic Diversity and Population Demography
2. Demographic Reconstruction
3. Coalescent Simulations
4. Approximate Bayesian Computation (ABC)
5. Practicals

1. Genetic Diversity and Population Demography

Effect of demography on genetic diversity



Evolutionary forces

Observed genetic diversity

Mutation, recombination

Selection

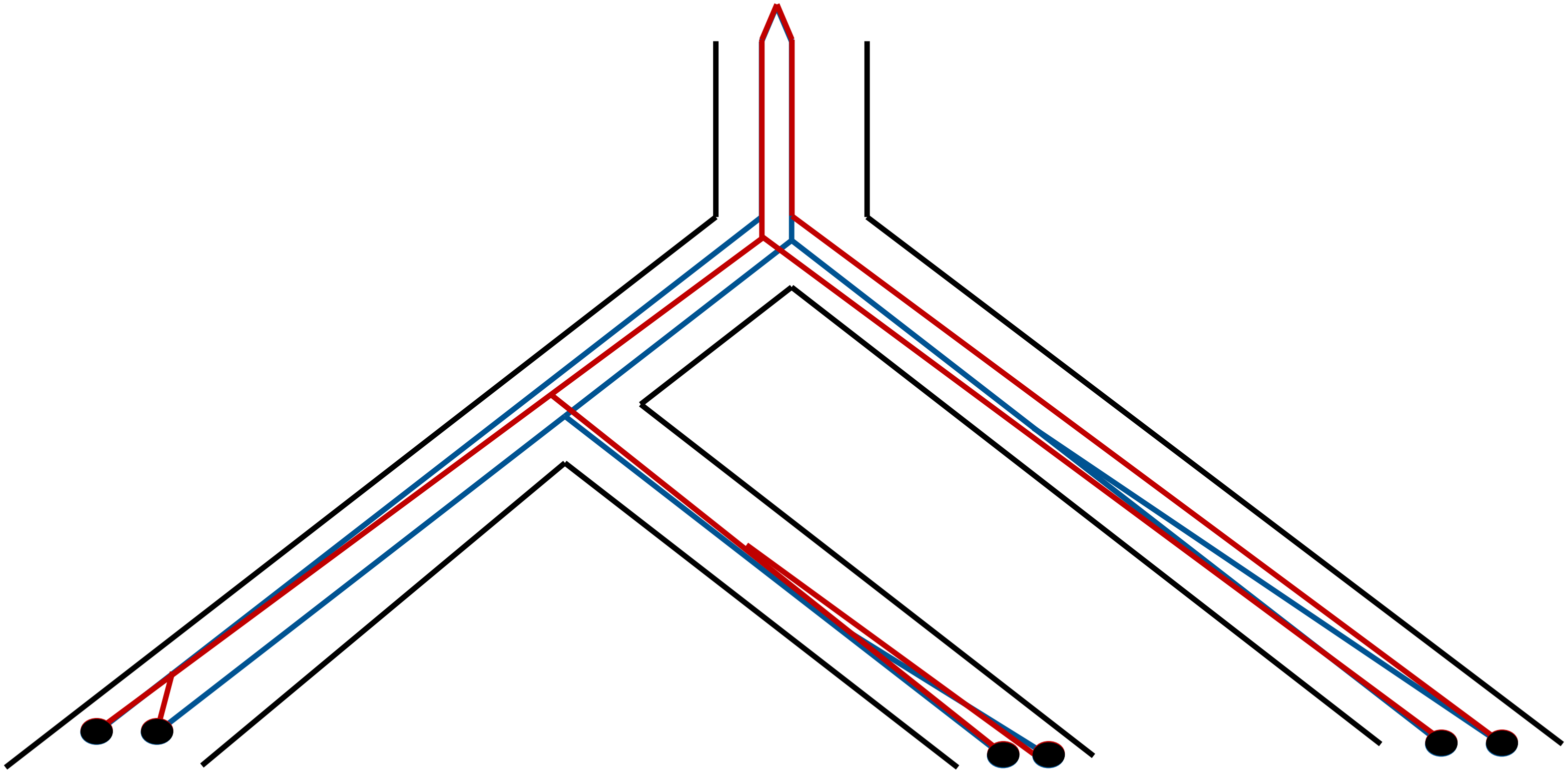
Demography & migration

- Low population size → More genetic drift
- Large population size → Less genetic drift
- Few migrations among populations → High genetic differentiation
- Many migrations among populations → Genetic homogenisation
- Temporal dynamics (growth, bottleneck, etc…) → …
- Spatial dynamics (population expansion or contraction) → …

It is possible to make inferences on population demography from genetic data using appropriate tools

Course example: coalescent simulations and ABC

Gene genealogy ≠ Population genealogy

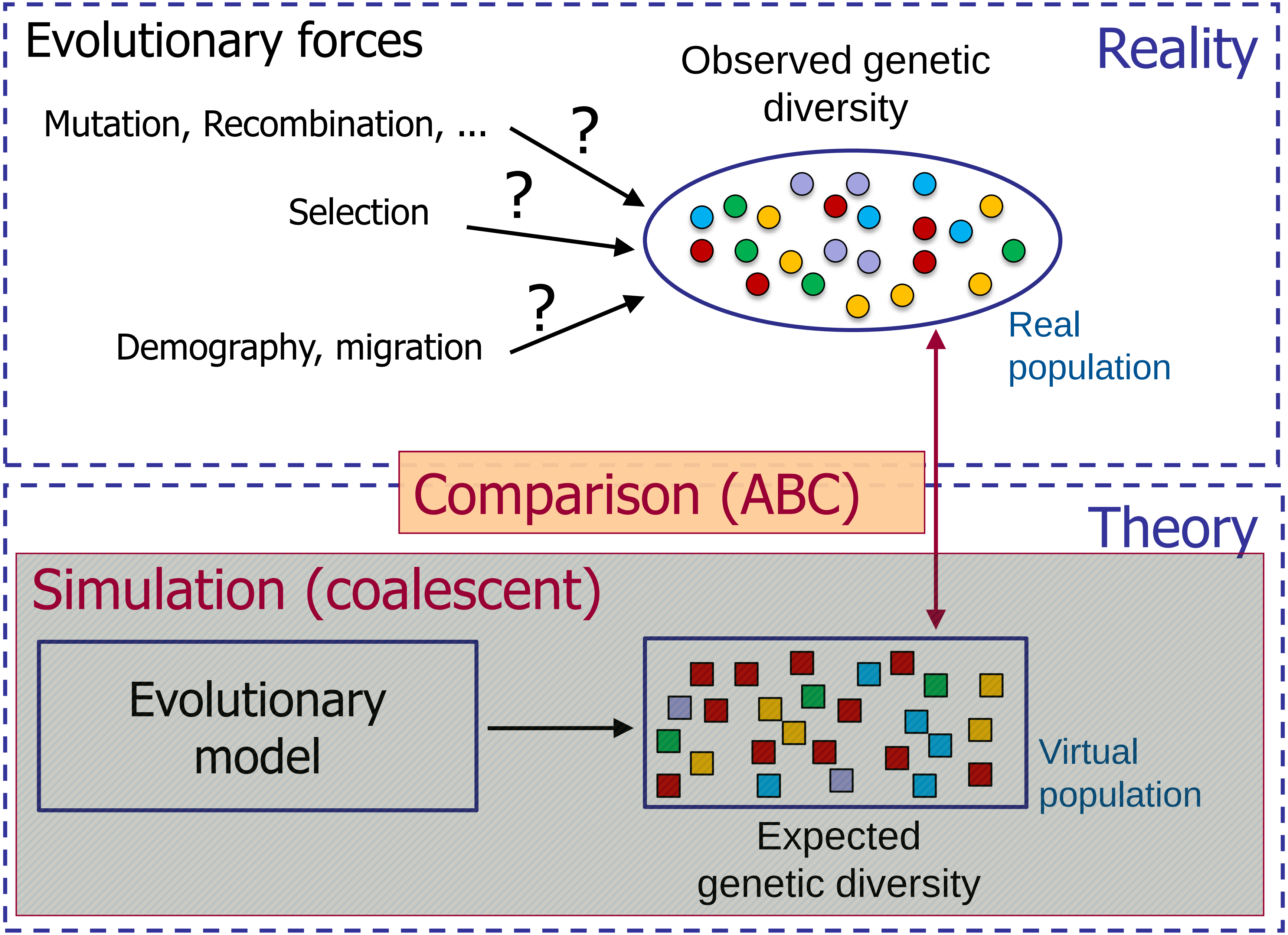


The reconstruction of population demographic history requires to overlap the information from a maximum of genetic loci (portions of DNA).

→ Demography affects the whole genome while selection affects a limited number of loci

2. Demographic reconstruction from genetic/genomic data

Main principles



Reality

Evolutionary forces

Mutation, Recombination, …

?

Selection

?

Demography, migration

?

Observed genetic diversity

Real population

Comparison (ABC)

Theory

Simulation (coalescent)

Evolutionary model

Virtual population

Expected genetic diversity

Modeling/Simulation part

- Many genetic simulation resourcess available, choose carefully the most adapted to your question.

  A (non-exhaustive) list:
  https://popmodels.cancercontrol.cancer.gov/gsr/packages/

- Two main kinds of genetic simulation approaches:
  1. Forward-in-time: i.e. Wright-Fisher (cf Andrew Clark Lecture)
  2. Coalescent: i.e. Fastsimcoal

A model is not a reproduction of the reality but a simplified theoritical representation of the main processes and elements that one wants to better understand

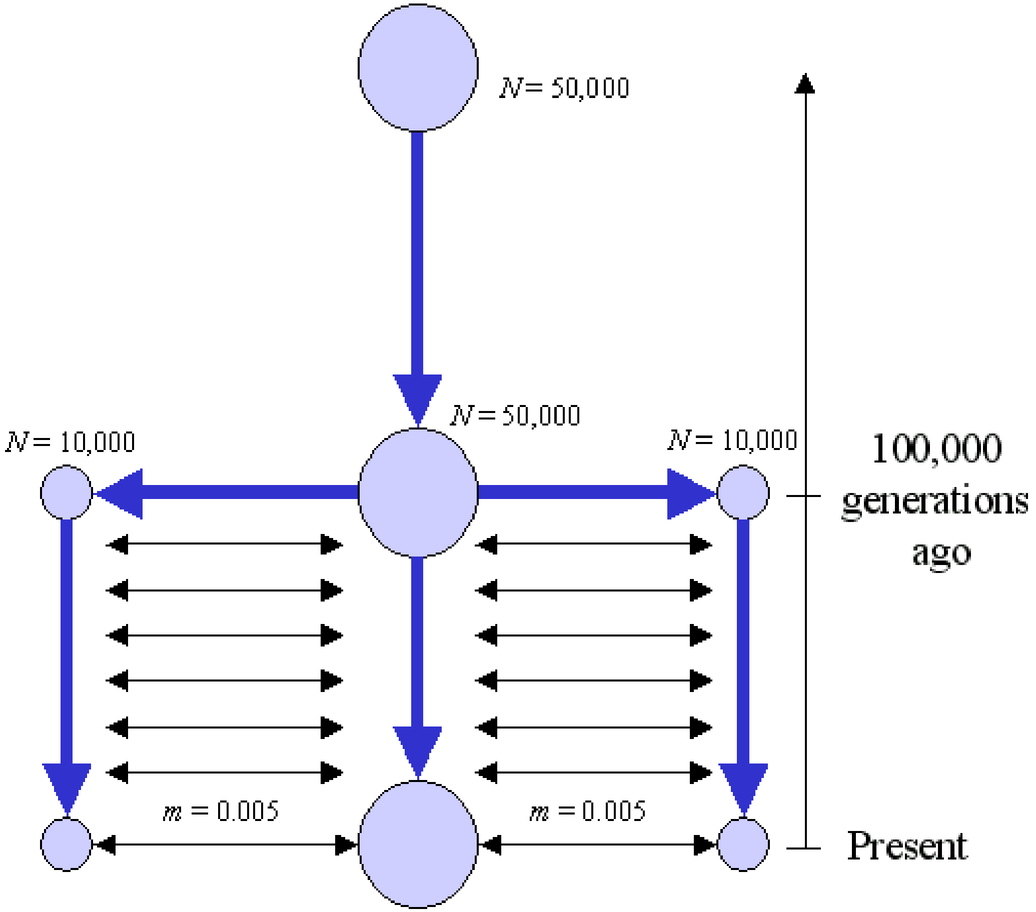
3 – Coalescent simulation

fastSimcoal2: example of demographic scenario

Example of input file

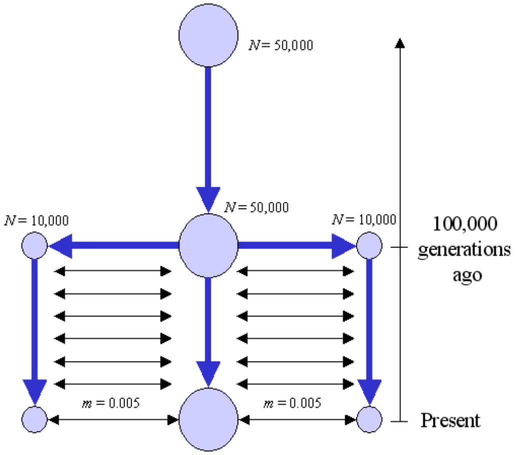
```
3 samples to simulate
//Deme sizes (haploid number of genes)
10000
50000
10000
//Sample sizes
2
0
3
//Growth rates
0
0
0
//Number of migration matrices
2
//Migration rates matrix 0 :
0.000 0.005 0.000
0.005 0.000 0.005
0.000 0.005 0.000
//Migration rates matrix 1 :
0 0 0
0 0 0
0 0 0
//Historical event: time, source, sink, migrants, new
deme size, new growth rate, new migration matrix
2 historical events
100000 0 1 1 1 0 1
100000 2 1 1 1 0 1
```

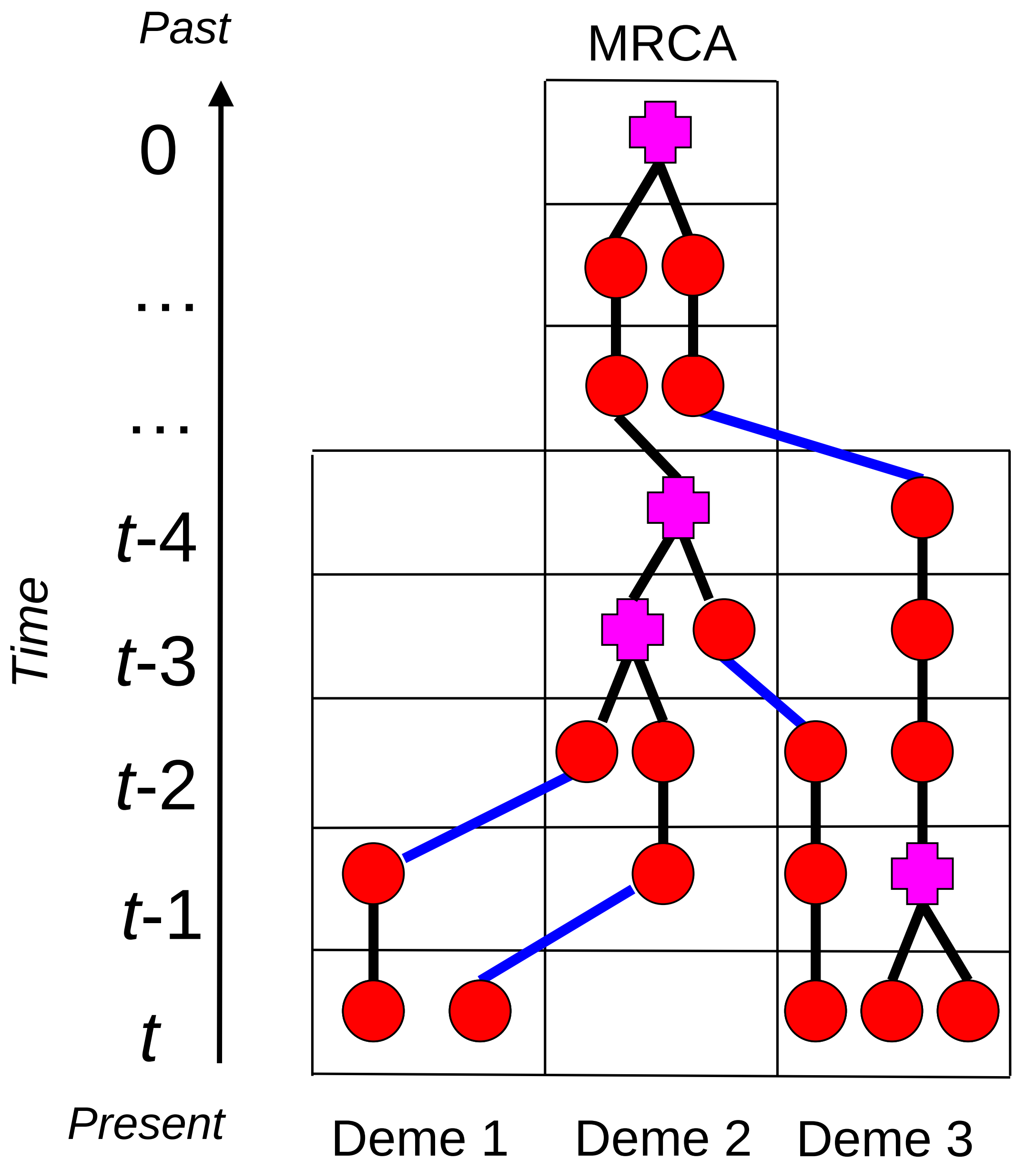
fastSimcoal2: A general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography.



Fastsimcoal:Excoffier et al, PLoS genetics 2013
http://cmpg.unibe.ch/software/fastsimcoal2/

Coalescent implementation





At each generation, 2 kinds of events are possible

- Migration

with **Prob$_m$ = *m*/*N*** —

where *m* = migration rate
N = deme (population) size

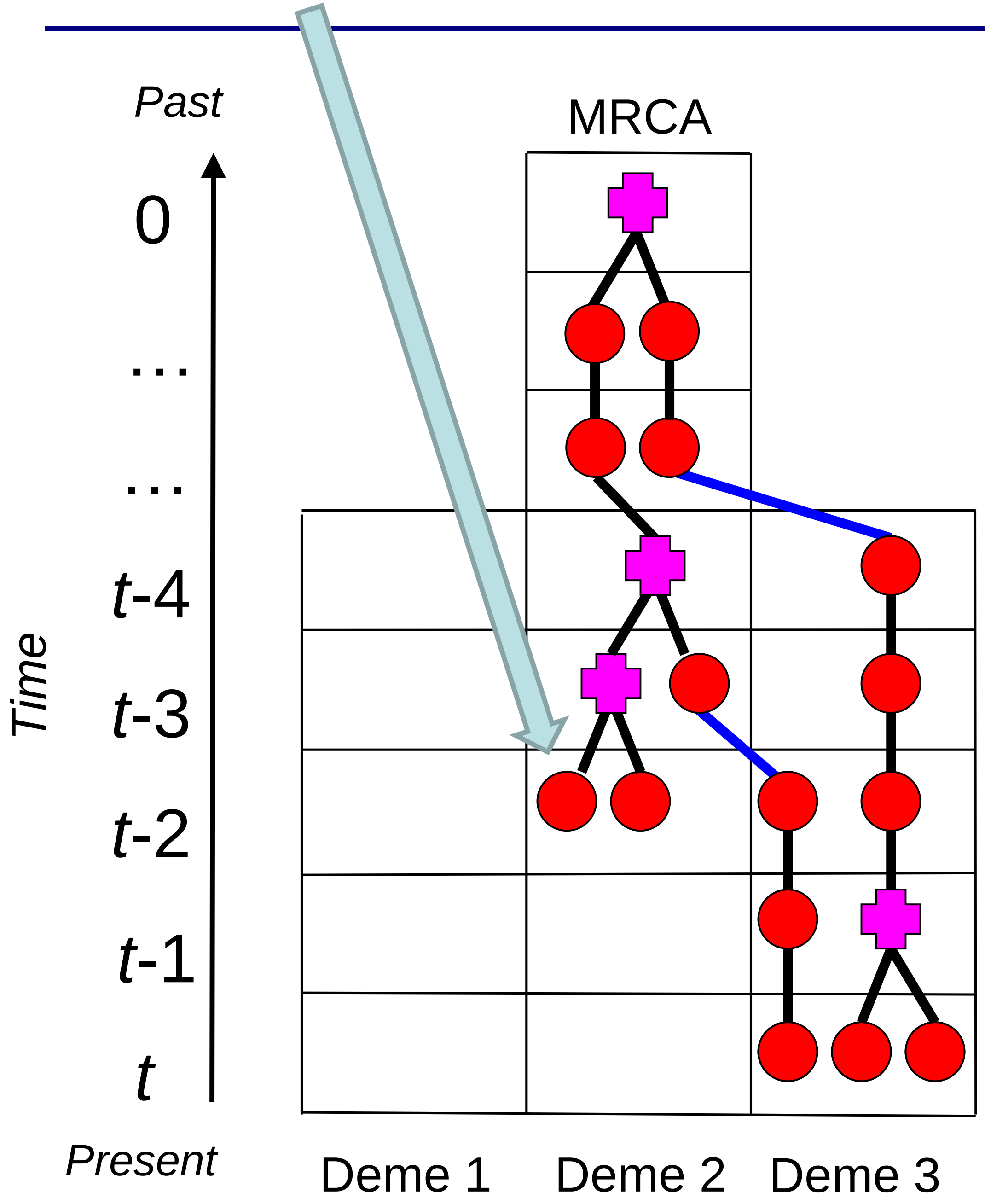
- Coalescent event ✚

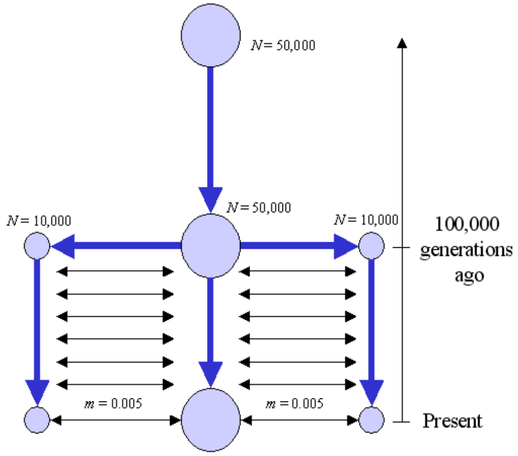
with **Prob$_c$ = *n*(*n*-1)/*N***

where *n* = gene number

● = neutral gene

Ancient DNA !





At each generation, 2 kinds of events are possible

- Migration

with $\mathbf{Prob_m = \mathit{m/N}}$ ▬

where m = migration rate
N = deme (population) size

- Coalescent event ✚

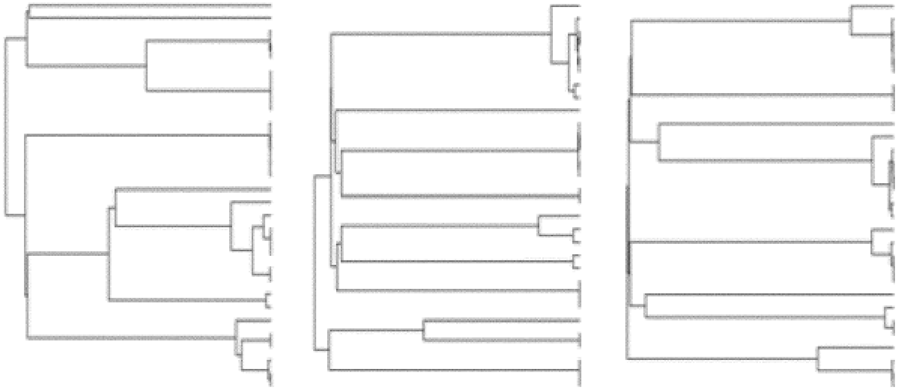
with $\mathbf{Prob_c = \mathit{n}(\mathit{n}-1)/\mathit{N}}$

where n = gene number

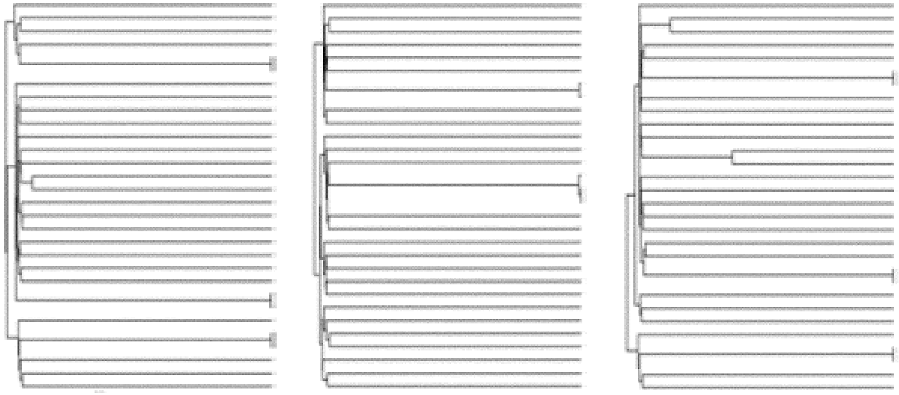
● = neutral gene

A stochastic process

Small size
Expanding
population

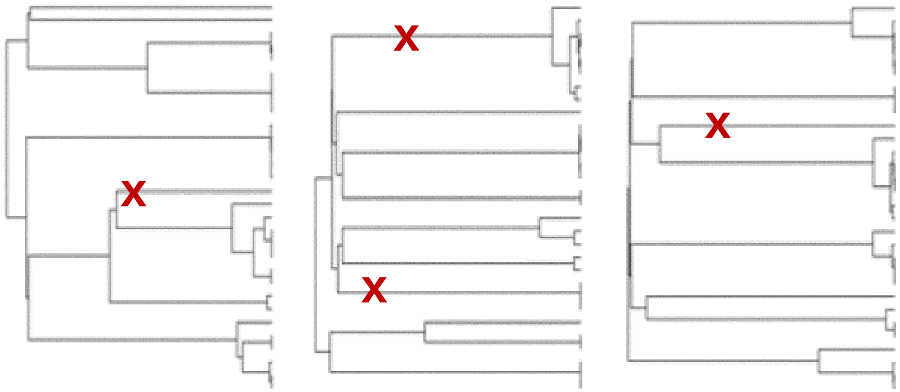


Large size
Expanding
Population



Simulation of genetic diversity

Small size
Expanding
population



DNA, STR, SNP

ATTATCGATATAT

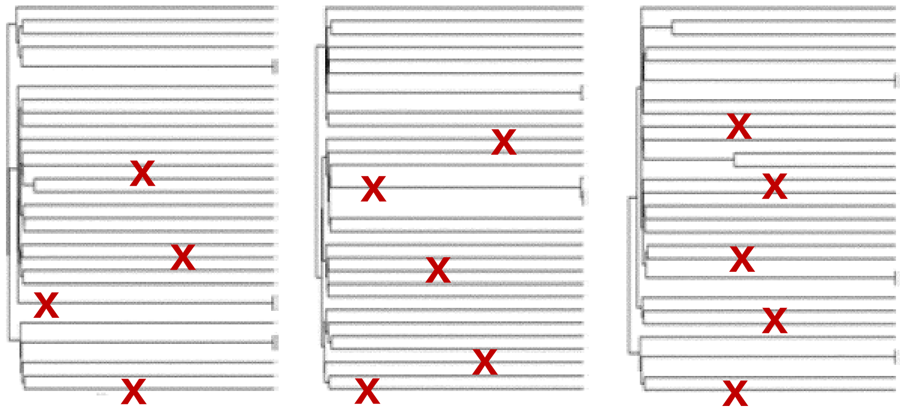
….

ATAATCGATATAT

….

….

Large size
Expanding
Population



….

….

….

….

μ = mutation rate **X = mutation**

Arlsumstat: computation of summary statistics

Arlsumstat is a Linux version of Arlequin 3.5 which compute summary statistics from arlequin projects in a very efficient way, specifically designed for ABC.

Excoffier & Lischer, Mol Ecol Res 2010
http://cmpg.unibe.ch/software/arlequin35/

Executable name:

arlsumstat3522_64bit

Input data file: *.arp

Input settings files:

arl_run.ars, ssdefs.txt

Associated Script:

LaunchArlSumStatModified.sh

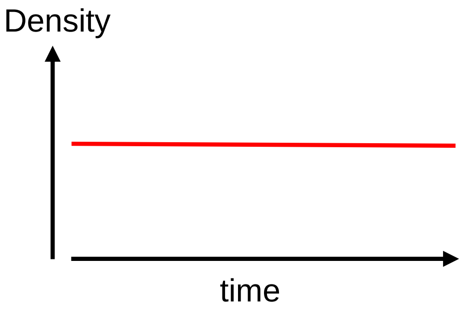
```
[Profile]
            Title="A series of simulated samples"
            NbSamples=1
            GenotypicData=1
            GameticPhase=0
            RecessiveData=0
            DataType=DNA
            LocusSeparator=NONE
            MissingData='?‘

[Data]
            [[Samples]]
                        SampleName="Sample 1"
                        SampleSize=25
                        SampleData= {
1_1         1           TATTCTAATTCAGCTTCTGAACGTAAGG
                        TAGTAGTCTGCATAGCGGCGTTGTGCGA
1_2         1           TAGTCGTCTGCGTATTGGGGTTGTGCAG
                        TAGTCGTCTGCGTATTGGGGTTGTGCAG
1_3         1           TATGCTAATTCAGCTTCTGATCGTAAGG
                        TAGTCGTCTGCATAGTGGCGTTGTGCGA
1_4         1           AATGCTAATTCAGCTTCTGATCGTAAGG
                        TAGTCGTCTGCATAGTGGCGTTGTGCGA
1_5         1           TATGCTAATTCAGCTTCTGATCGTAAGG
                        TATTCTAATTCAGCTTCTGAACGTAAGG
```

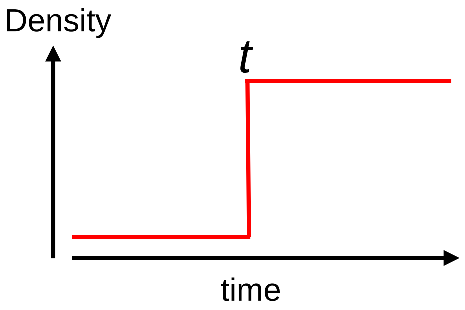
Translation of demography to genetics

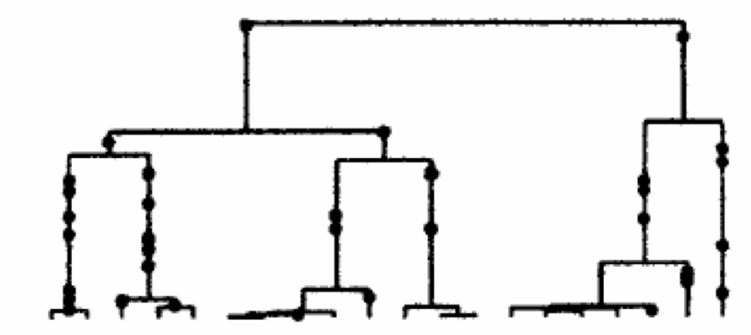
Population with constant size

Population after a demographic increase

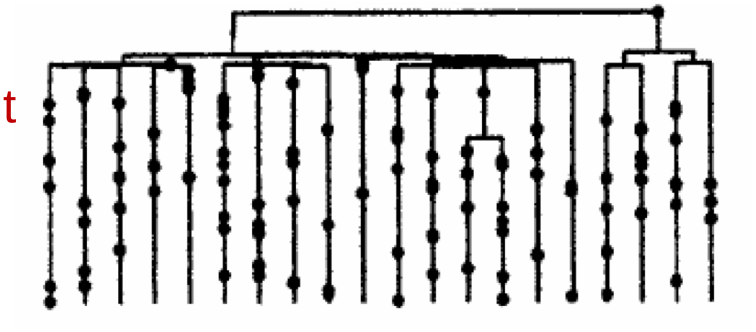


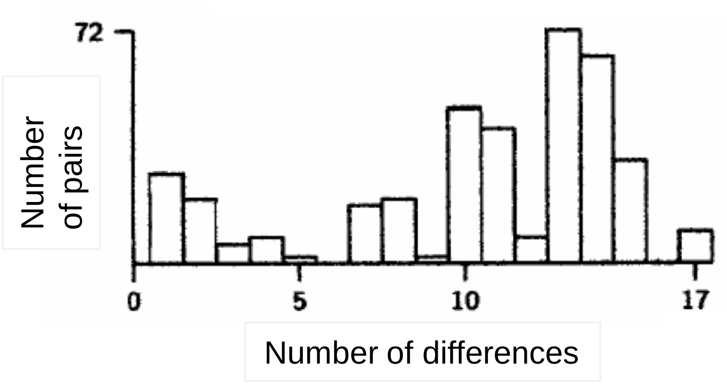
Demographic scenarios



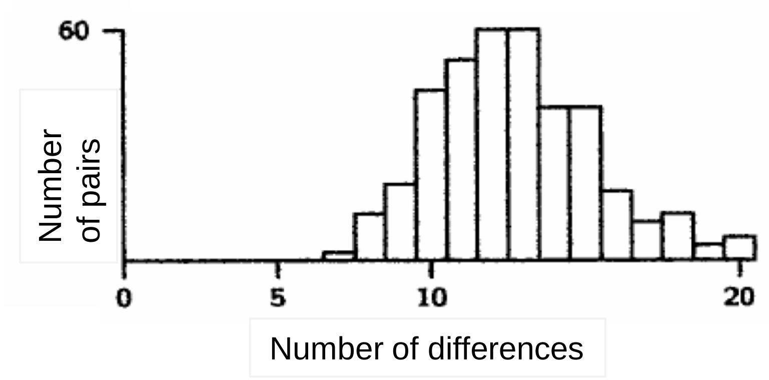


Coalescent trees



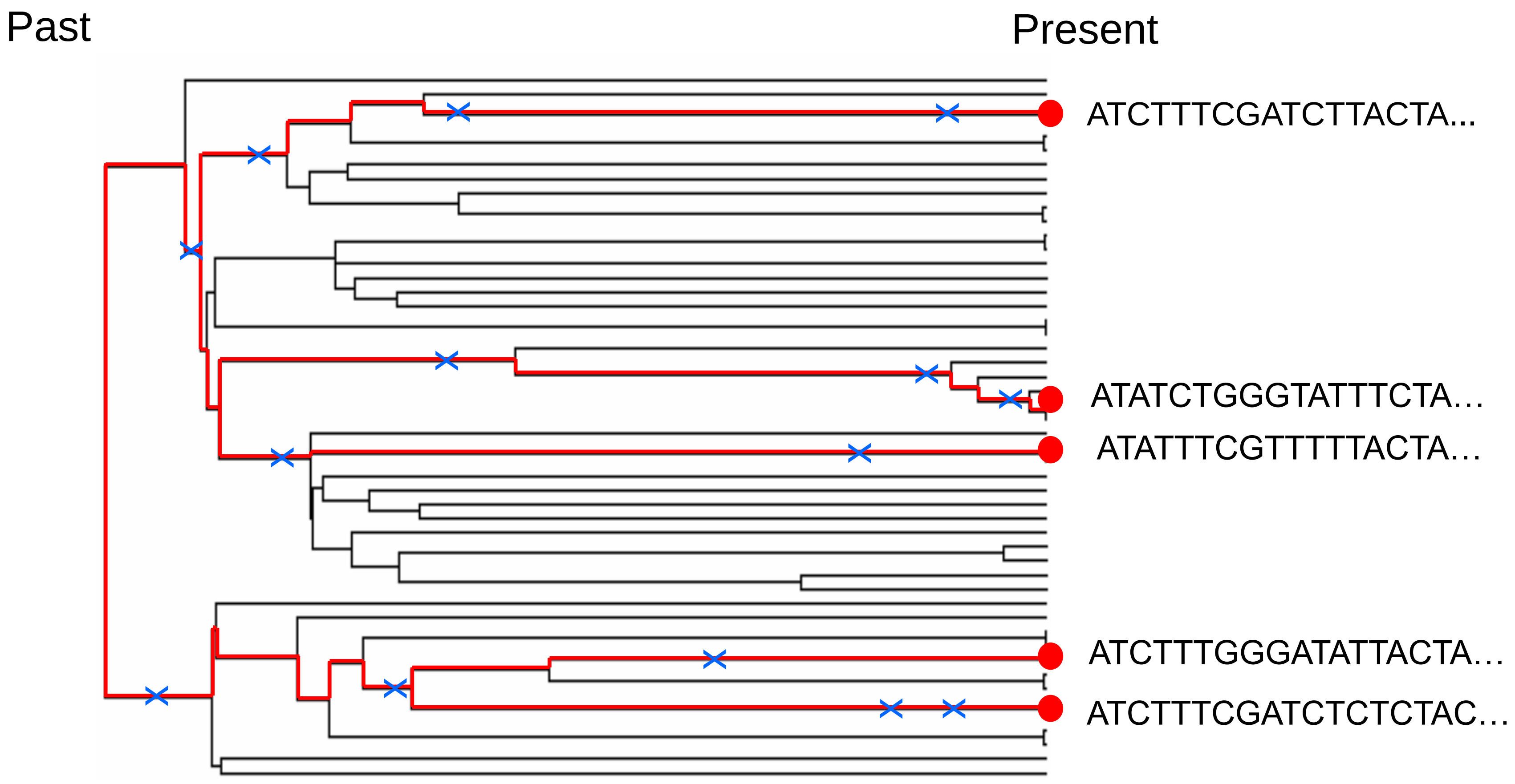


Summary statistics



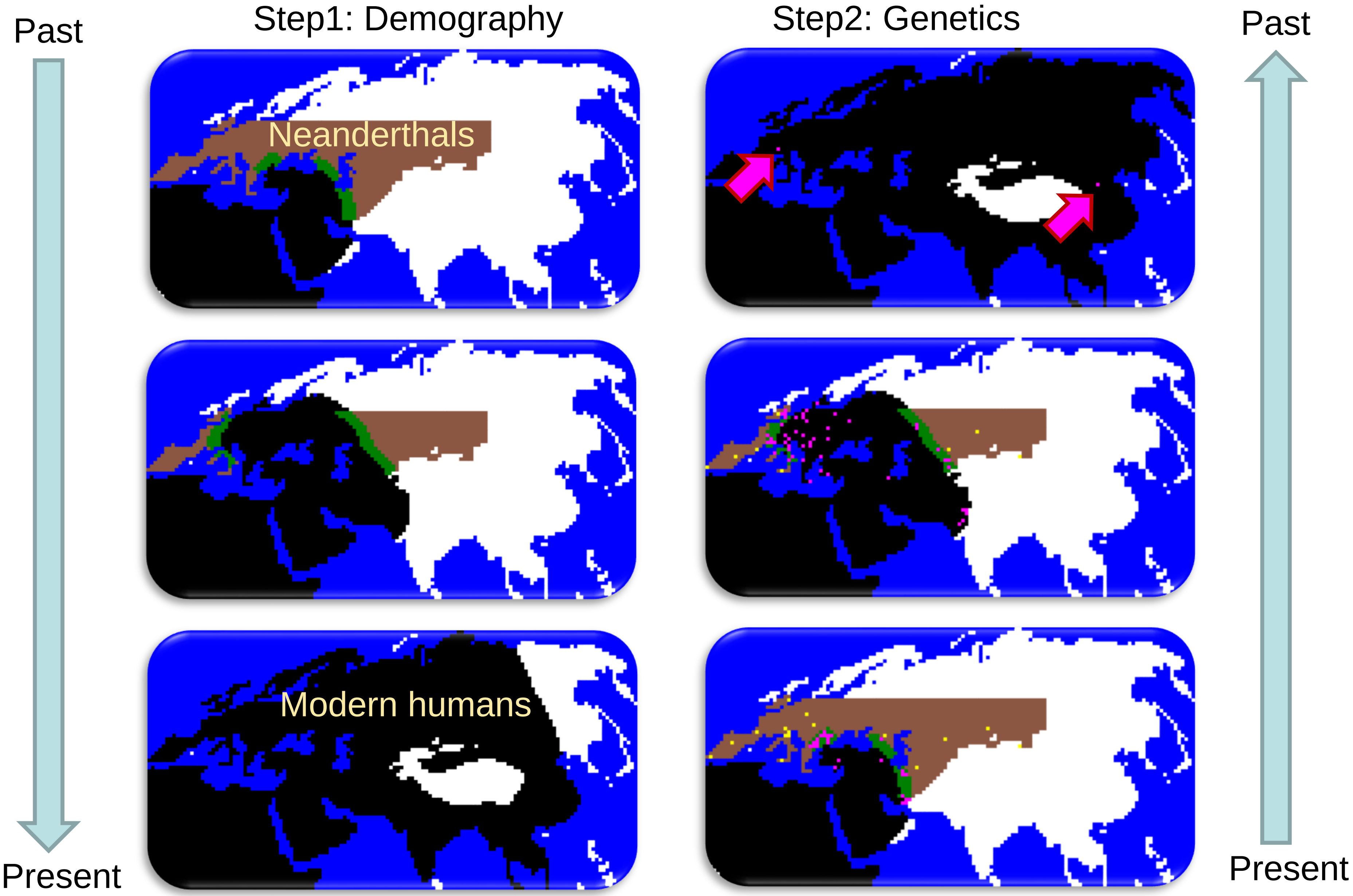
Modified from Harpending et al (1998)

Advantage of the coalescent approach



Simulation of only the sampled genes ● and their ancestors, not the whole population → huge gain in computational time !

Spatially-explicit coalescent simulation



Step1: Demography

Step2: Genetics

Past

Present

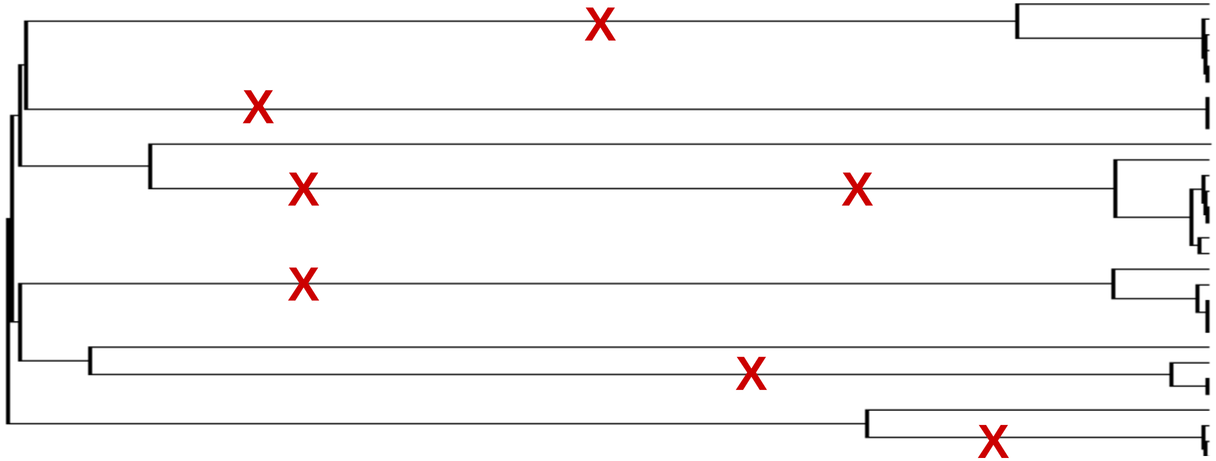
Past

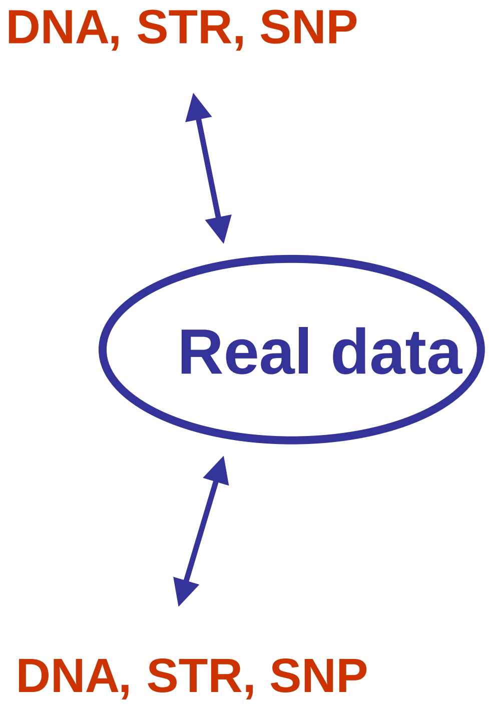
Present

Currat & Excoffier, PNAS 2011

Comparison between simulated and empirical data

Scenario A

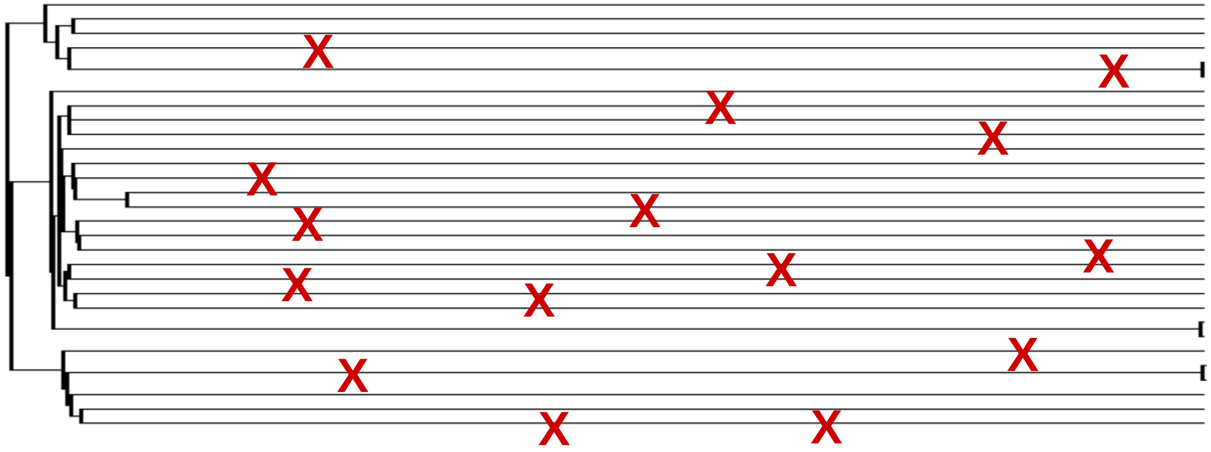



DNA, STR, SNP

Real data

DNA, STR, SNP

Scenario B



X = mutation

4 – Approximate Bayesian Computation

(ABC)

ABC main principles

D → Data (genetic/genomic)

M → Model (evolutionary scenario)

θ → Model Parameter (demographic/biological/…)

$$P(\theta|D) \propto f_M(D|\theta)\ P(\theta)$$

Posterior distribution

Probability distribution of θ knowing *D*

Likelihood function

Probability distribution *D* given θ, based on model *M*

Prior distribution

Probability distribution of θ before knowing *D*

Problem: the computation of the likelihood function may be very costly or even impossible for complex models.

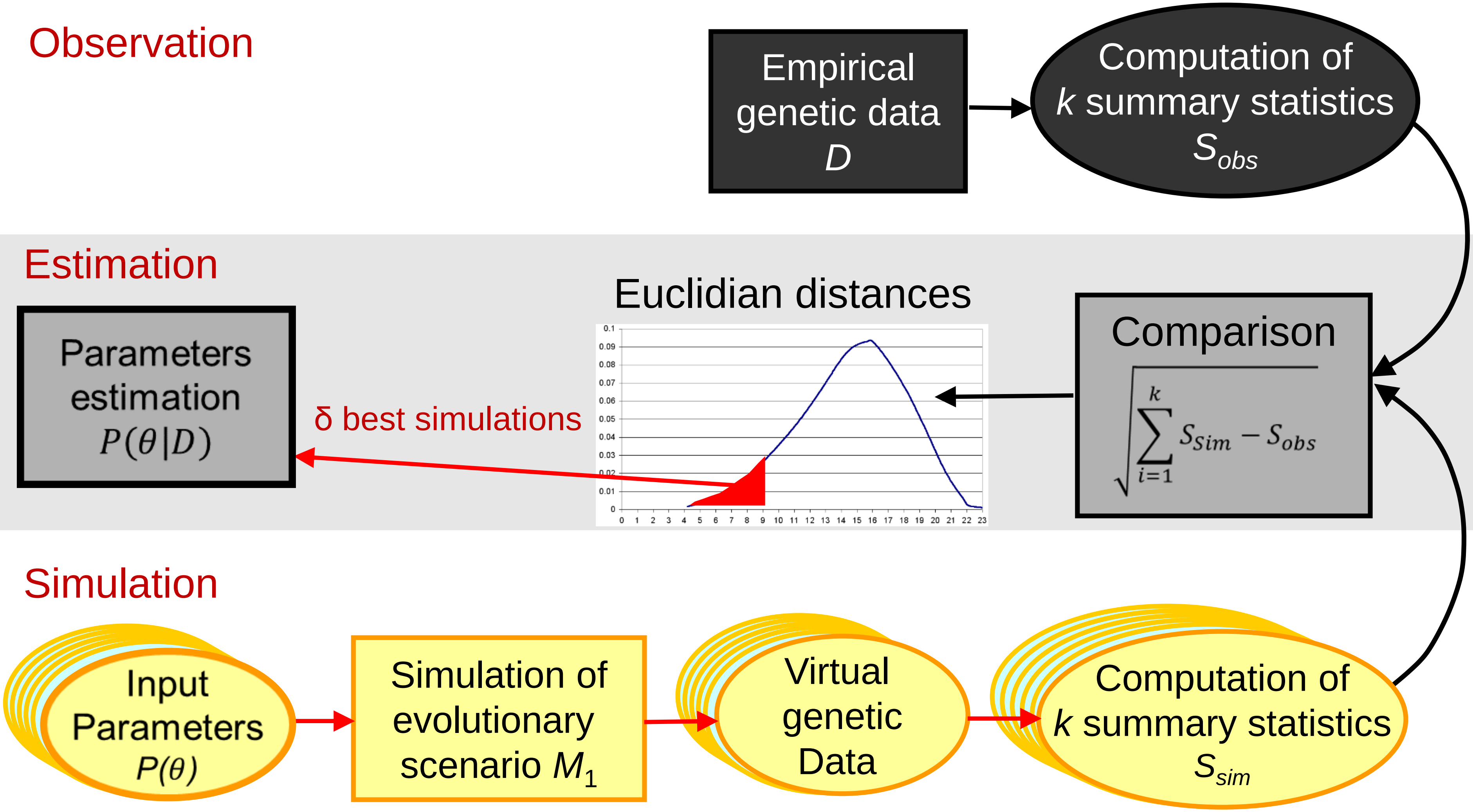
The ABC approach has been designed to bypass the computation of the likelihood function by approximating it using stochastic simulation of the model.

Tavaré *et al*, Genetics (1997), Beaumont *et al*, Genetics (2002)

Many recent developements and several packages to run ABC (DiyABC, PopABC, Abc R package, etc…)

For the practicals, you will use ABCtoolbox, Wegmann *et al*, Bioinformatics 2010

Parameter estimation through ABC



Observation

Empirical genetic data *D*

Computation of *k* summary statistics S_{obs}

Estimation

Euclidian distances

Comparison

$$\sqrt{\sum_{i=1}^{k} S_{Sim} - S_{obs}}$$

δ best simulations

Parameters estimation $P(\theta|D)$

Simulation

Input Parameters $P(\theta)$

Simulation of evolutionary scenario M_1

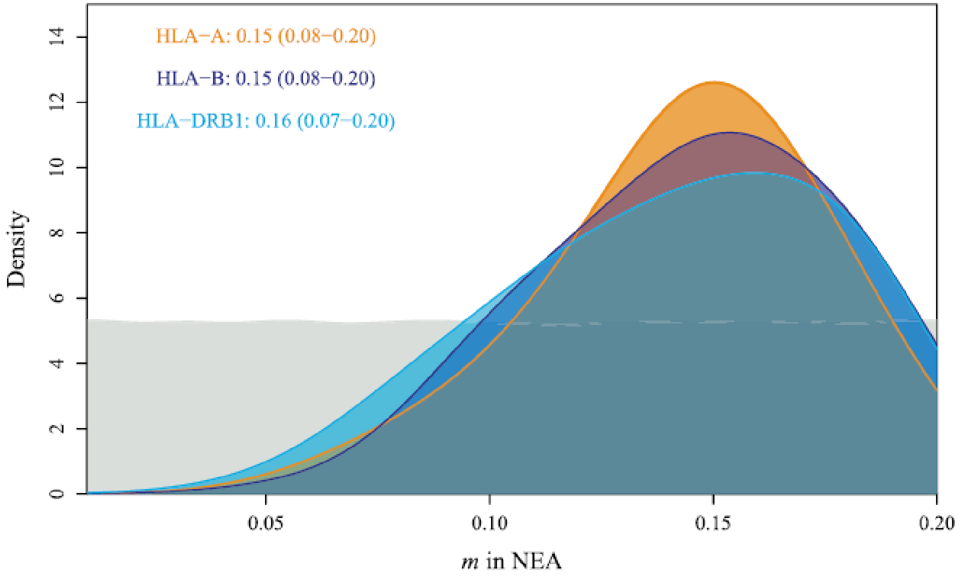
Virtual genetic Data

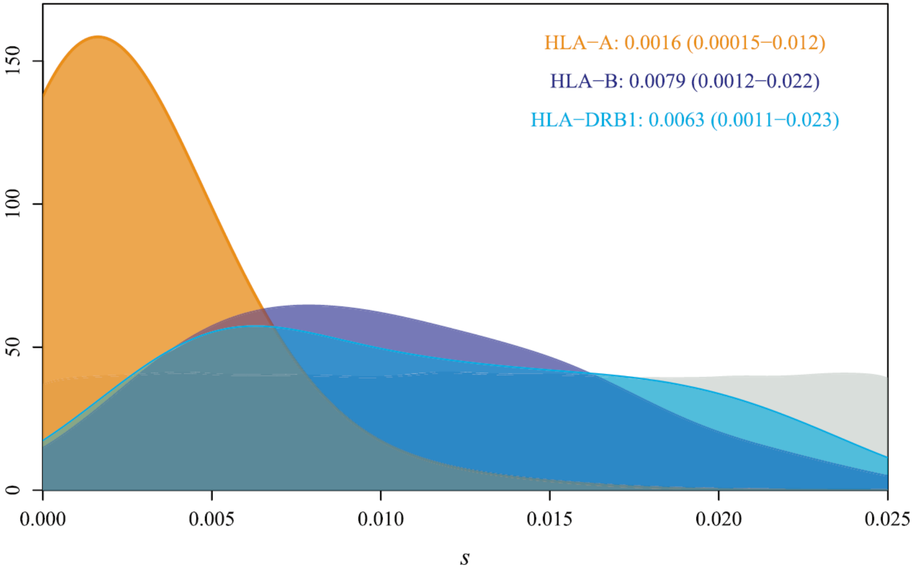
Computation of *k* summary statistics S_{sim}

Examples of parameter estimation outputs

Prior and posterior distributions

Di *et al.* BMC Evolutionary Biology (2015)





Point estimates and confidence intervals

Alves *et al.* Mol. Biol. Evol. (2016)

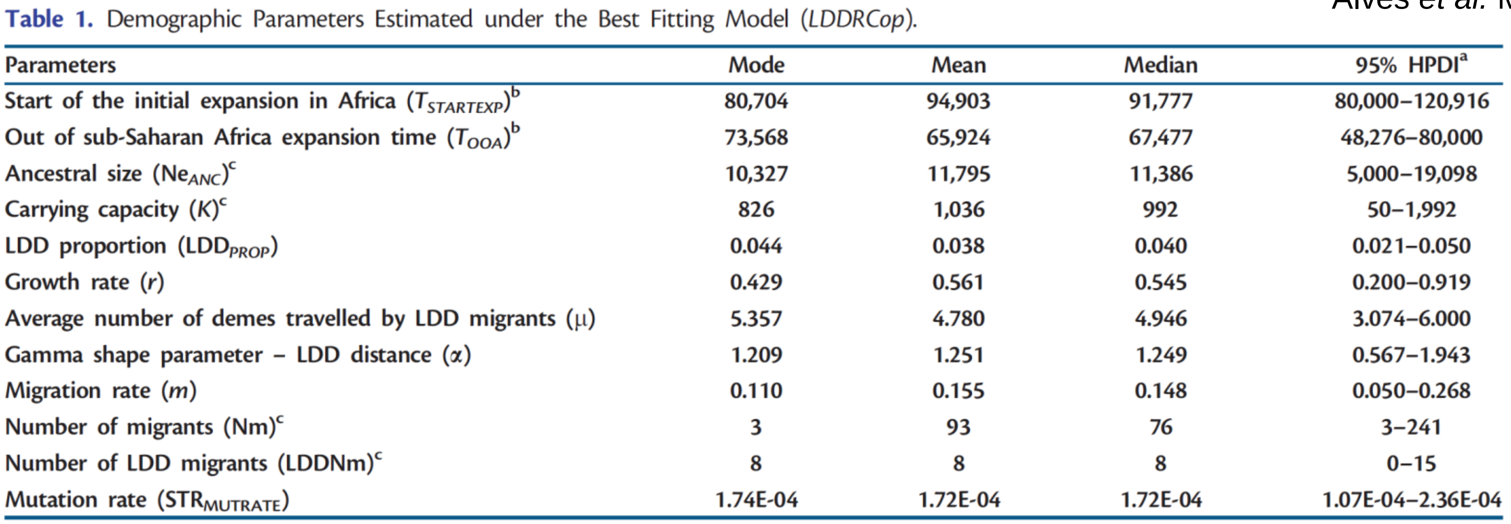
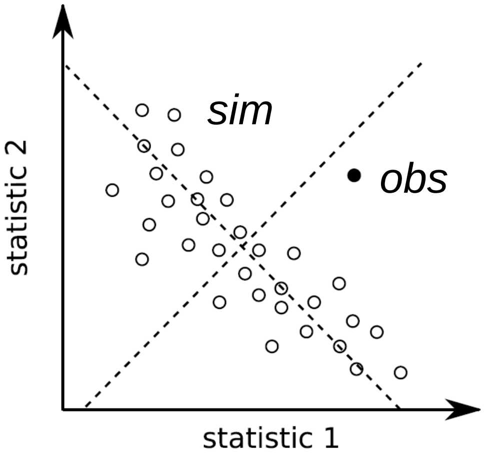


Table 1. Demographic Parameters Estimated under the Best Fitting Model (*LDDRCop*).

Parameters	Mode	Mean	Median	95% HPDI[a]
Start of the initial expansion in Africa ($T_{STARTEXP}$)[b]	80,704	94,903	91,777	80,000–120,916
Out of sub-Saharan Africa expansion time (T_{OOA})[b]	73,568	65,924	67,477	48,276–80,000
Ancestral size (Ne_{ANC})[c]	10,327	11,795	11,386	5,000–19,098
Carrying capacity (K)[c]	826	1,036	992	50–1,992
LDD proportion (LDD_{PROP})	0.044	0.038	0.040	0.021–0.050
Growth rate (r)	0.429	0.561	0.545	0.200–0.919
Average number of demes travelled by LDD migrants (μ)	5.357	4.780	4.946	3.074–6.000
Gamma shape parameter – LDD distance (α)	1.209	1.251	1.249	0.567–1.943
Migration rate (m)	0.110	0.155	0.148	0.050–0.268
Number of migrants (Nm)[c]	3	93	76	3–241
Number of LDD migrants (LDDNm)[c]	8	8	8	0–15
Mutation rate ($STR_{MUTRATE}$)	1.74E-04	1.72E-04	1.72E-04	1.07E-04–2.36E-04

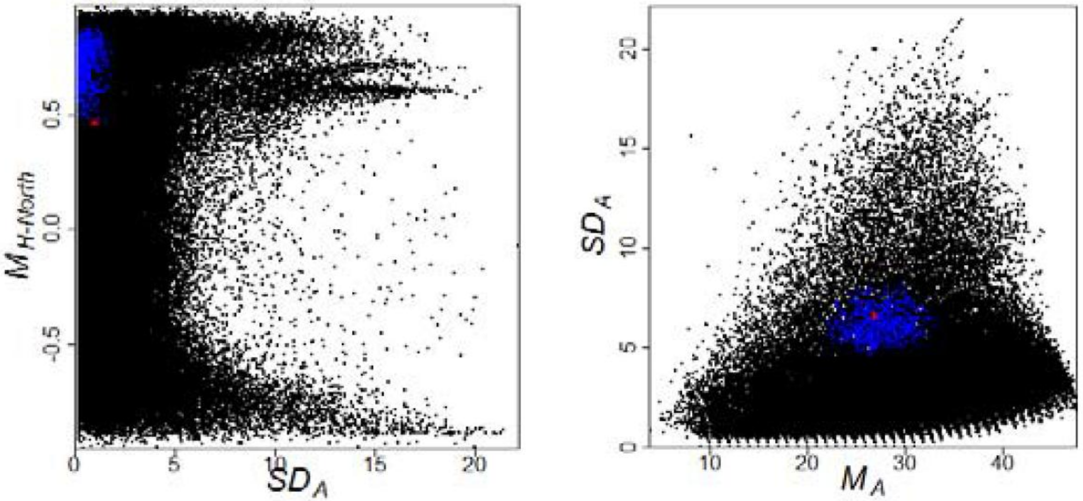
Validation techniques: model fit

Is the model plausible ? Is it capable to reproduced adequately empirical statistics ?



ABCtoolbox user manual

Visual inspection of 2D joint densities for each pair of statistics



Di *et al.* BMC Evolutionary Biology (2015)

ABCtoolbox provides model fit statistics:

Marginal p-value

Tukey p-value.

→ Low p-values indicate poor fit.

Validation techniques: accuracy of estimates

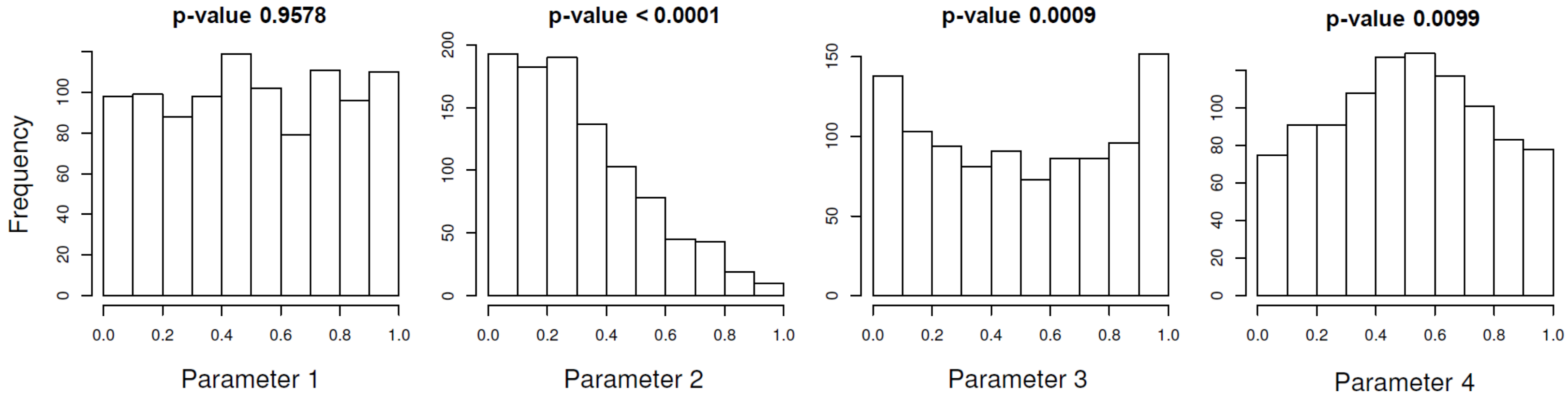
How accurate is the estimation of a parameter ?

The **cross-validation** procedure repeats the estimation with the output of one simulation considered as empirical values (pseudo-observed data).

TRUE	Estimated			
Pop. Size	**Pop. Size Mode**	**Pop. Size Mean**	**Pop. Size Quantile**	**Pop. Size HDI**
10070	11987	16920	0	0.75
14386	23494	24055	0.067487	0.749736
46270	29248	31159	0.874571	0.868895
11806	10070	14996	0.001913	0.105752
24072	17741	20153	0.666673	0.689085

Checking for **biased posteriors**

Kolmogorov-Smirnov test of quantile distribution against an uniform distribution.



unbiased

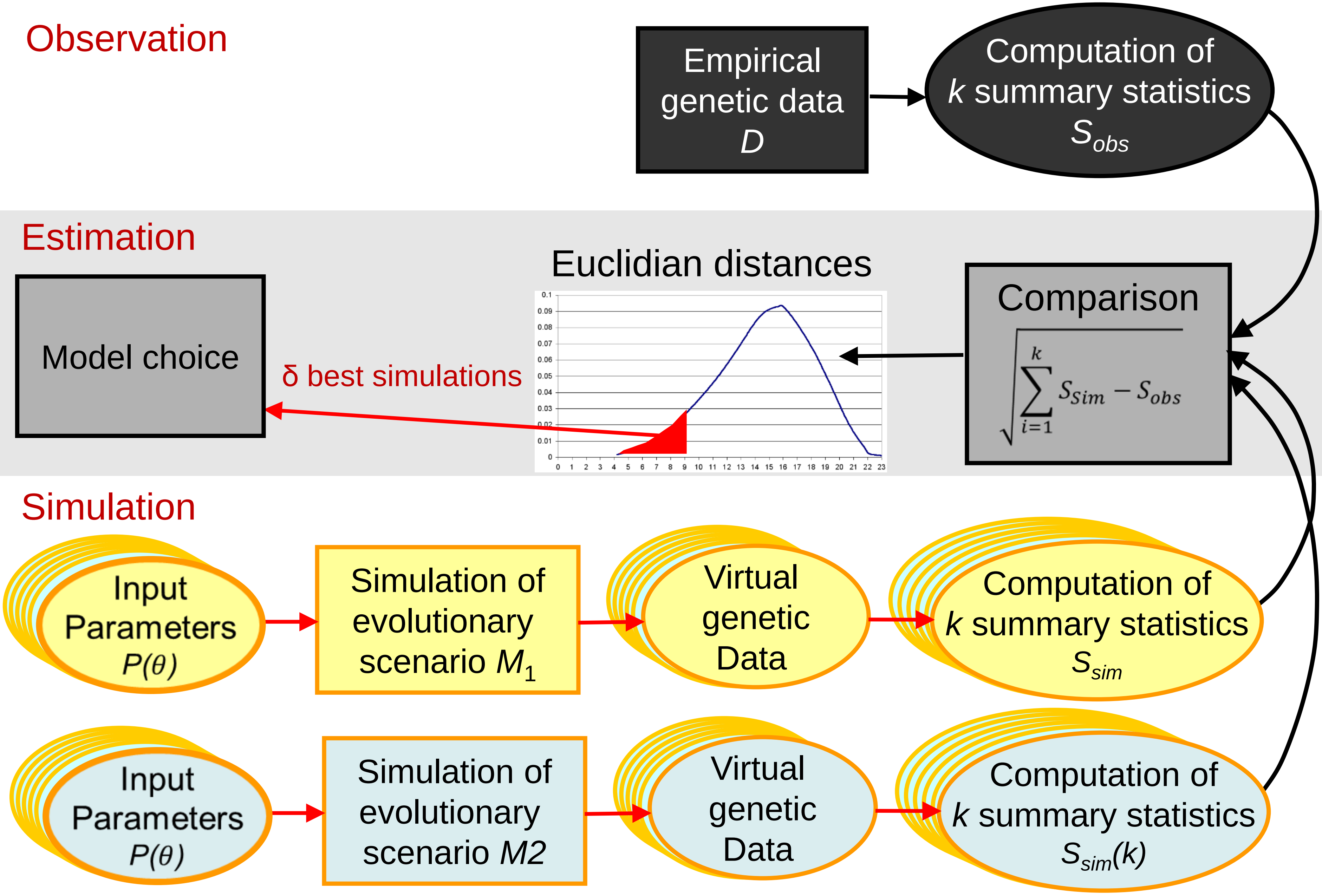
Biased → too large values

Biased → too narrow prior

Biased → too large prior

ABCtoolbox user manual

Model choice through ABC



Observation

Empirical genetic data *D*

Computation of *k* summary statistics S_{obs}

Estimation

Model choice

δ best simulations

Euclidian distances

Comparison

$$\sqrt{\sum_{i=1}^{k} S_{Sim} - S_{obs}}$$

Simulation

Input Parameters $P(\theta)$

Simulation of evolutionary scenario M_1

Virtual genetic Data

Computation of *k* summary statistics S_{sim}

Input Parameters $P(\theta)$

Simulation of evolutionary scenario *M2*

Virtual genetic Data

Computation of *k* summary statistics $S_{sim}(k)$

Examples of model choice outputs

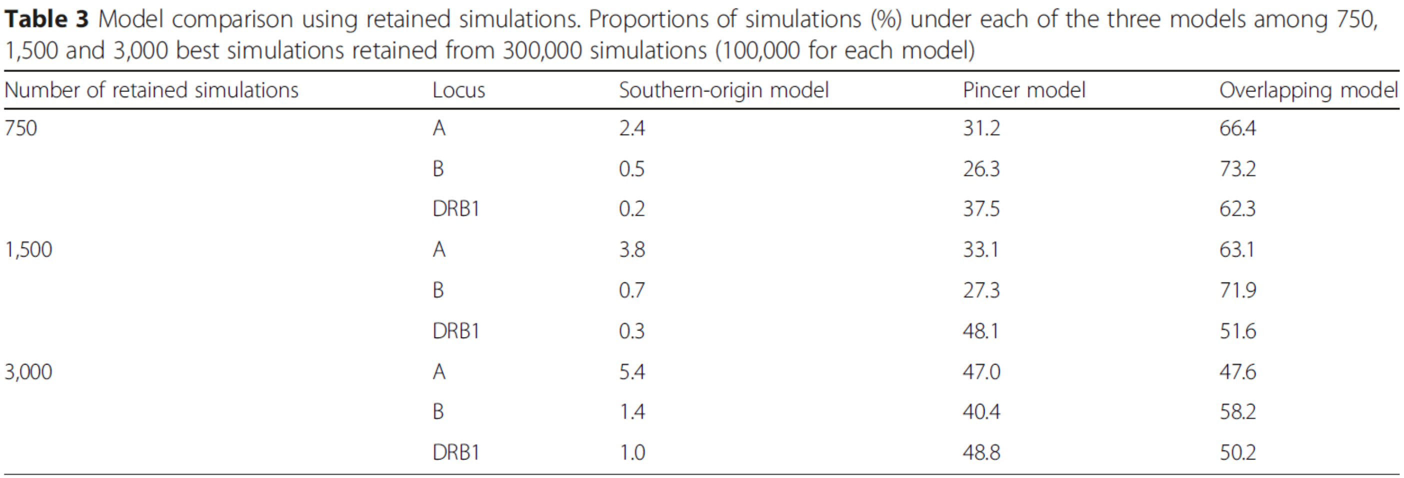
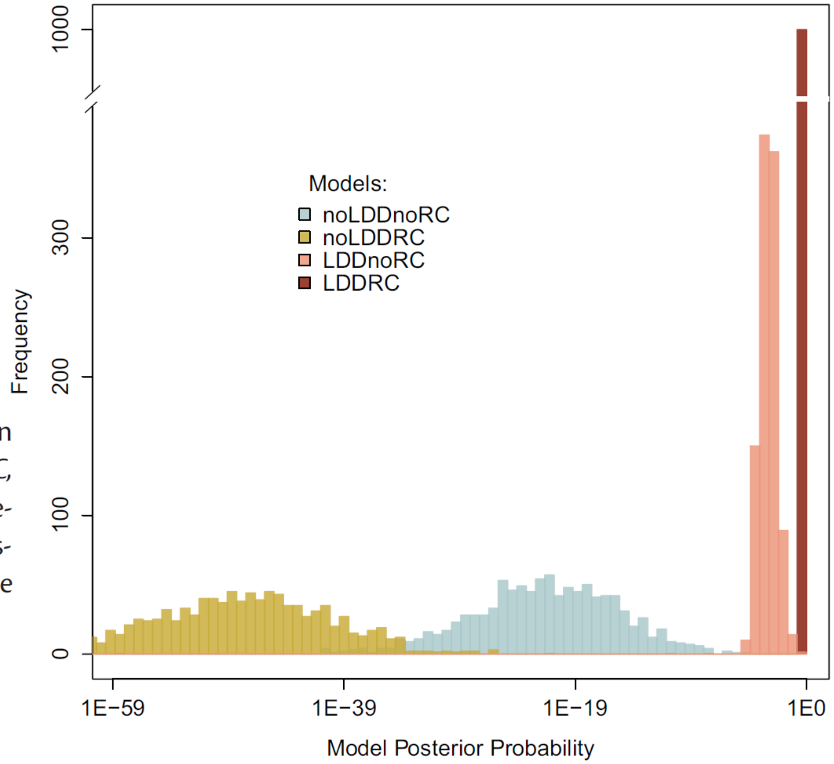


Table 3 Model comparison using retained simulations. Proportions of simulations (%) under each of the three models among 750, 1,500 and 3,000 best simulations retained from 300,000 simulations (100,000 for each model)

Number of retained simulations	Locus	Southern-origin model	Pincer model	Overlapping model
750	A	2.4	31.2	66.4
	B	0.5	26.3	73.2
	DRB1	0.2	37.5	62.3
1,500	A	3.8	33.1	63.1
	B	0.7	27.3	71.9
	DRB1	0.3	48.1	51.6
3,000	A	5.4	47.0	47.6
	B	1.4	40.4	58.2
	DRB1	1.0	48.8	50.2

Di *et al.* BMC Evolutionary Biology (2015)



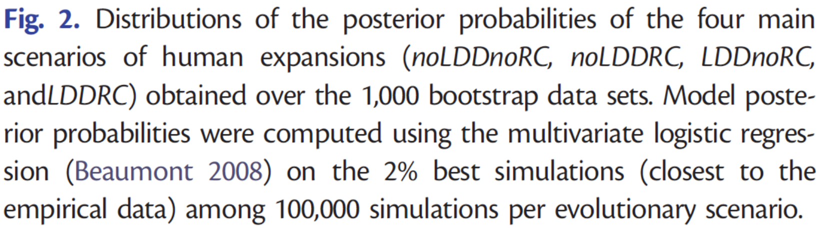
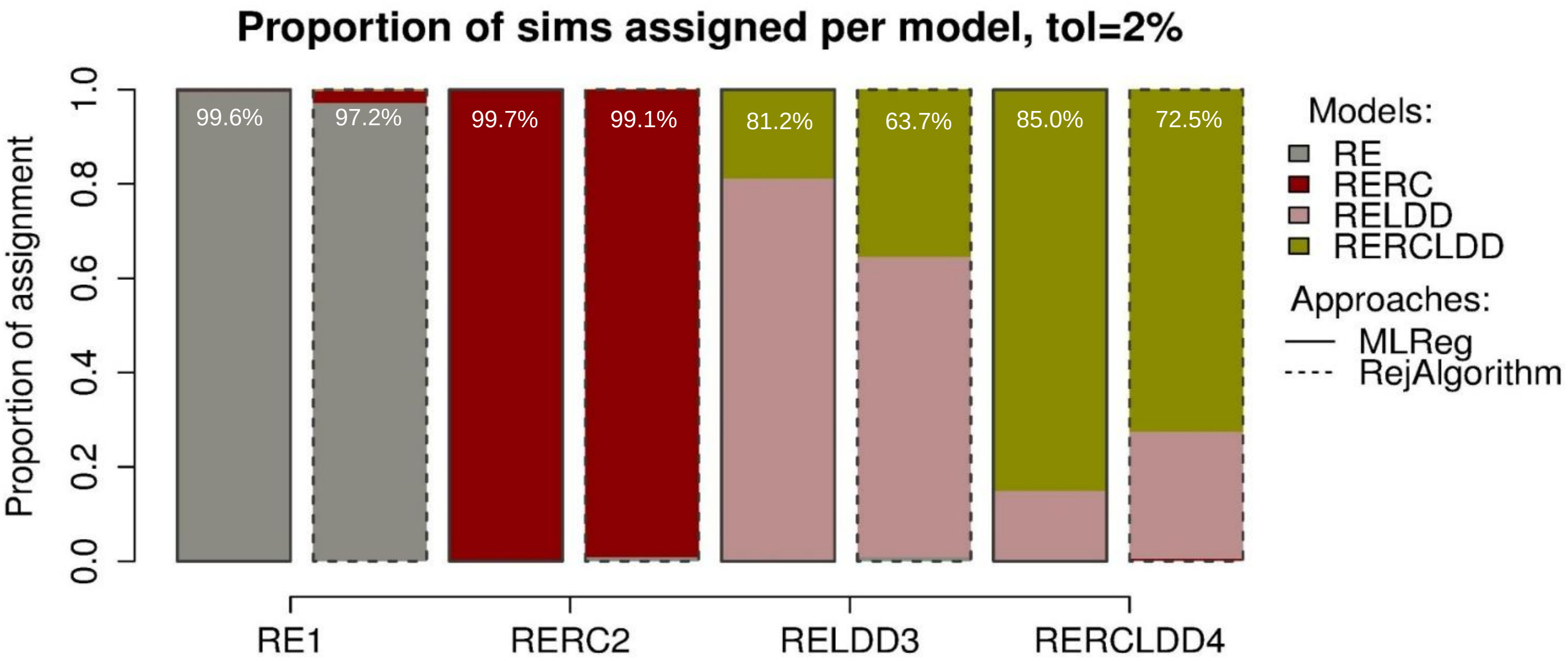


Fig. 2. Distributions of the posterior probabilities of the four main scenarios of human expansions (*noLDDnoRC, noLDDRC, LDDnoRC,* and*LDDRC*) obtained over the 1,000 bootstrap data sets. Model posterior probabilities were computed using the multivariate logistic regression (Beaumont 2008) on the 2% best simulations (closest to the empirical data) among 100,000 simulations per evolutionary scenario.

Alves *et al.* Mol. Biol. Evol. (2016)

Validation techniques: cross-validation procedure



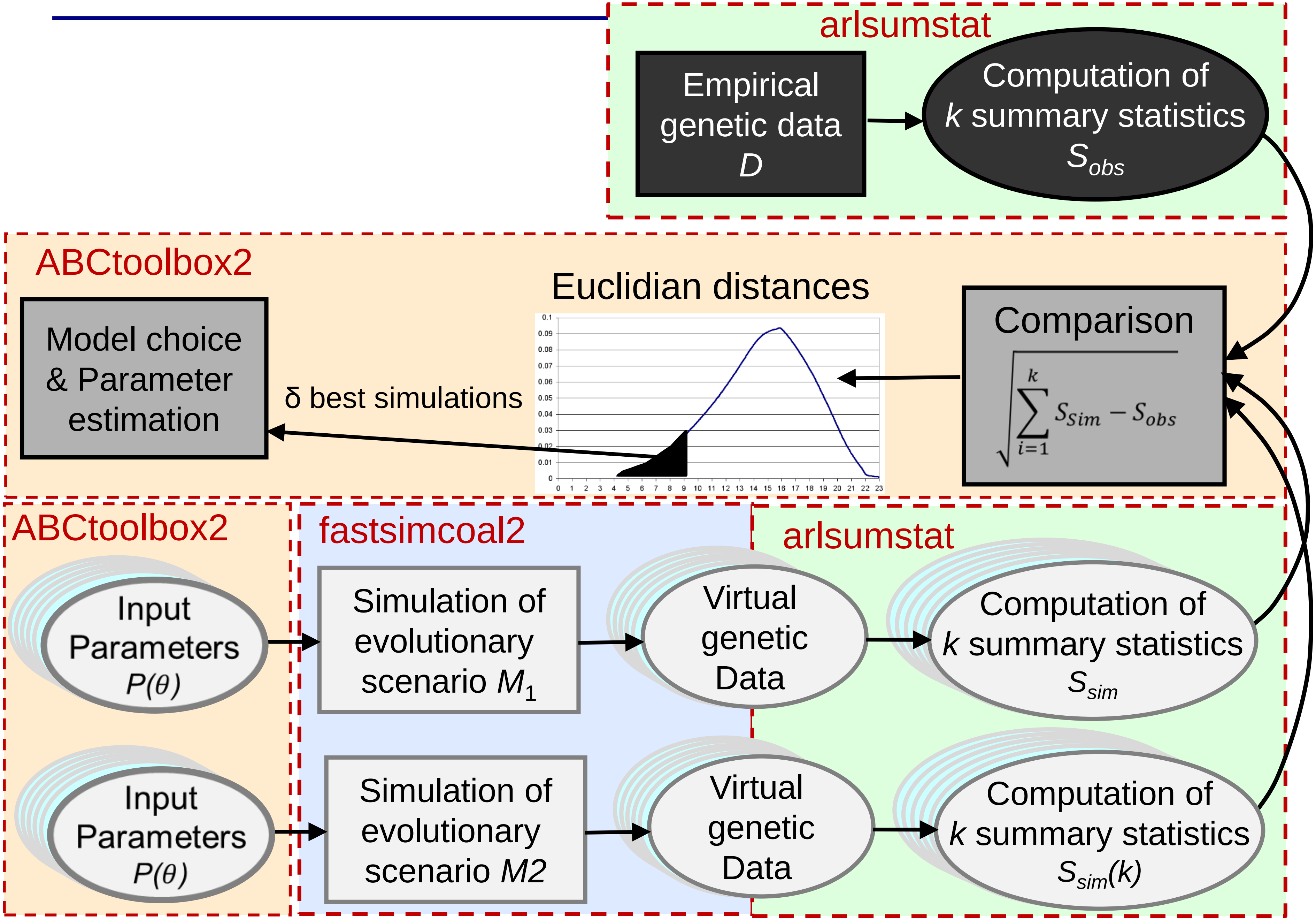
Alves *et al.* Mol. Biol. Evol. (2016)

Practical difficulties

1. Choice of the prior distribution(s)
   - Distribution shape and parameters (uniform, log uniform, normal, etc…)
2. Design of the model(s)
   - Reproduce the main elements but avoid unnecessary complexity
   - Model's output sufficiently different to be distinguished
3. Choice of the summary statistics
   - Enough to capture the main the characteristics of the model and have sufficient power for the estimation
   - Not too many to avoid incorporating random noise or distorting the estimation procedure
4. Choice of the number of simulations to perform
   - Enough simulations to explore the parameter space
5. Choice of the tolerance/retained parameter
   - Start between 1% and 5% and check that the results are robust acrosse different values
6. Validation of the method
   - Check the capability of the model to reproduce real data and the accuracy of parameter estimation

5. Practicals

Practicals



arlsumstat

Empirical genetic data *D*

Computation of *k* summary statistics S_{obs}

ABCtoolbox2

Model choice & Parameter estimation

δ best simulations

Euclidian distances

Comparison

$$\sqrt{\sum_{i=1}^{k} S_{Sim} - S_{obs}}$$

ABCtoolbox2

Input Parameters *P(θ)*

Input Parameters *P(θ)*

fastsimcoal2

Simulation of evolutionary scenario M_1

Simulation of evolutionary scenario *M2*

Virtual genetic Data

Virtual genetic Data

arlsumstat

Computation of *k* summary statistics S_{sim}

Computation of *k* summary statistics $S_{sim}(k)$

Practicals

STEP 1: SIMULATION OF DEMOGRAPHIC SCENARIO (fastsimcoal)

STEP 2: COMPUTATION OF SUMMARY STATISTICS (Arlsumstat)

STEP 3: USE A PARAMETER PRIOR DISTRIBUTION (ABCtoolbox)

STEP 4: GENERATE ABC SIMULATION DATASETS

(**OPTIONAL STEP 5**: GENERATE A NEW DATASET WITH TWO PARAMETERS)

STEP 6: MODEL CHOICE WITH ABC

STEP 7: PARAMETER ESTIMATION WITH ABC

(**OPTIONAL STEP 8**: EXPLORE AN ADDITIONAL SCENARIO)

Practical: Model choice

EMBO_modelchoicemodelFit.txt

dataSet	model1_marginalDensityPValue	model2_marginalDensityPValue	model3_marginalDensityPValue	model1_marginalDensity	model2_marginalDensity	model3_marginalDensity	model1_TukeydepthPValue	model2_TukeydepthPValue	model3_TukeydepthPValue	model1_TukeyDepth	model2_TukeyDepth	model3_TukeyDepth	model1_BayesFactor	model2_BayesFactor	model3_BayesFactor	model1_posteriorProbability	model2_posteriorProbability	model3_posteriorProbability	chosenModel
0	0.64	0.02	0	13.3006	0.000935	2.05E-17	0.67	0	0	0.171717	0	0	14228.7	7.03E-05	1.54E-18	0.99993	7.03E-05	1.54E-18	1

	Stationary-Small	Stationary-Big	Growing
	model1	model2	model3
Marginal Density PValue	0.64	0.02	0.00
Marginal Density	13.30	0.00	0.00
Tukeydepth PValue	0.67	0.00	0.00
Tukey Depth	0.17	0.00	0.00
Bayes Factor	14228.70	0.00	0.00
Posterior Probability	**1.00**	0.00	0.00

Model 1 - Stationary population of small effective size - is clearly the best supported model. In fact this is the only one able to reproduce the (pseudo) observed data

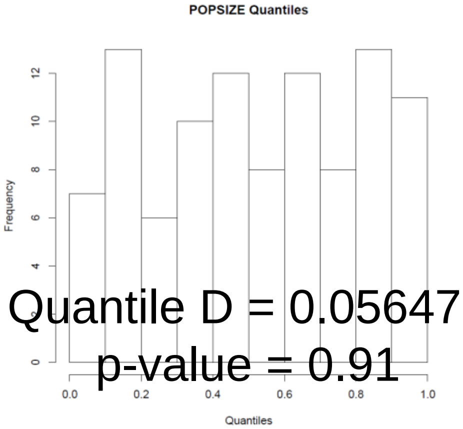
Practical: Parameter estimation

EMBO_Stationary-Small-Panmictic-Populationmodel0_MarginalPosteriorCharacteristics.txt

dataSet	POPSIZE_mode	POPSIZE_mean	POPSIZE_median	POPSIZE_q50_lower	POPSIZE_q50_upper	POPSIZE_q90_lower	POPSIZE_q90_upper	POPSIZE_q95_lower	POPSIZE_q95_upper	POPSIZE_q99_lower	POPSIZE_q99_upper	POPSIZE_HDI50_lower	POPSIZE_HDI50_upper	POPSIZE_HDI90_lower	POPSIZE_HDI90_upper	POPSIZE_HDI95_lower	POPSIZE_HDI95_upper	POPSIZE_HDI99_lower	POPSIZE_HDI99_upper
0	6200.55	6660.97	6365.78	4720.82	8326.55	2738.71	11721.1	2310.27	12687.1	1658.27	14526.5	4173.69	7634.03	2091.22	10702.5	1900.09	12087.7	1326.7	13711.3

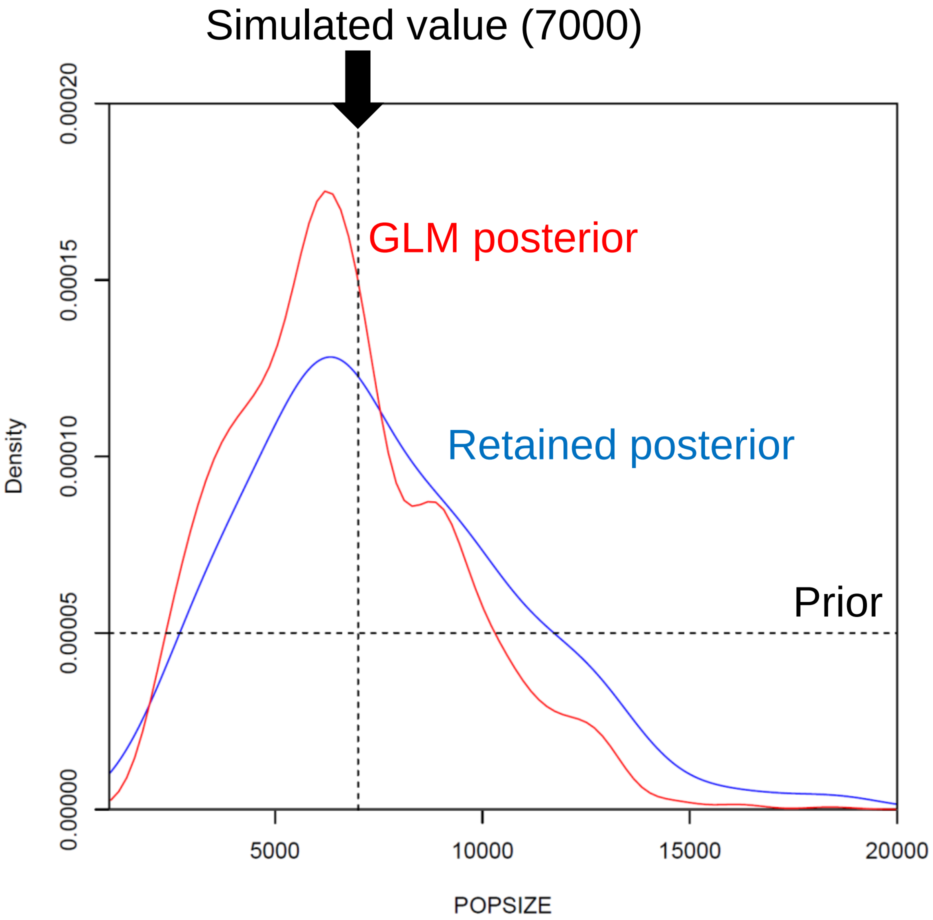
Population size				
Mode	Mean	Median	HDI95 lower	HDI95 upper
6201	6661	6365	1900	12088

Validation:



Quantile D = 0.056477
p-value = 0.91

Sim number	Marginal Density	POPSIZE	POPSIZE mode	POPSIZE mean
7	23.9526	7312	6391.68	8095.51
15	24.1679	7427	6391.68	8058.82
21	24.0918	4288	6391.68	7214.8
…	…	…	…	…


Simulated value (7000)
GLM posterior
Retained posterior
Prior
Density
POPSIZE