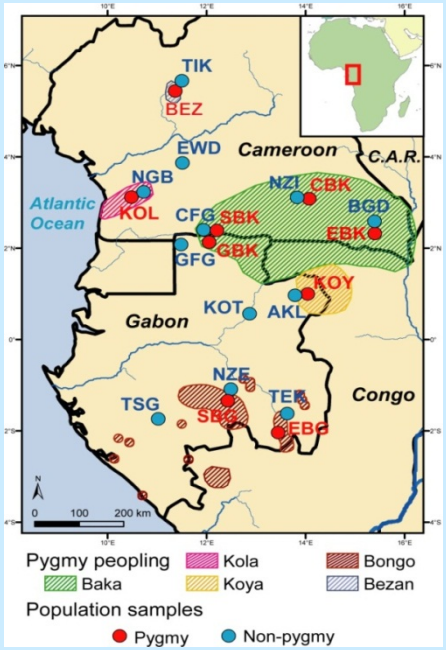
Applications of the coalescent process (ABC, MCMC)

Frédéric Austerlitz

UMR 7206 EcoAnthropologie et Ethnobiologie

http://www.ecoanthropologie.cnrs.fr/spip.php?article519

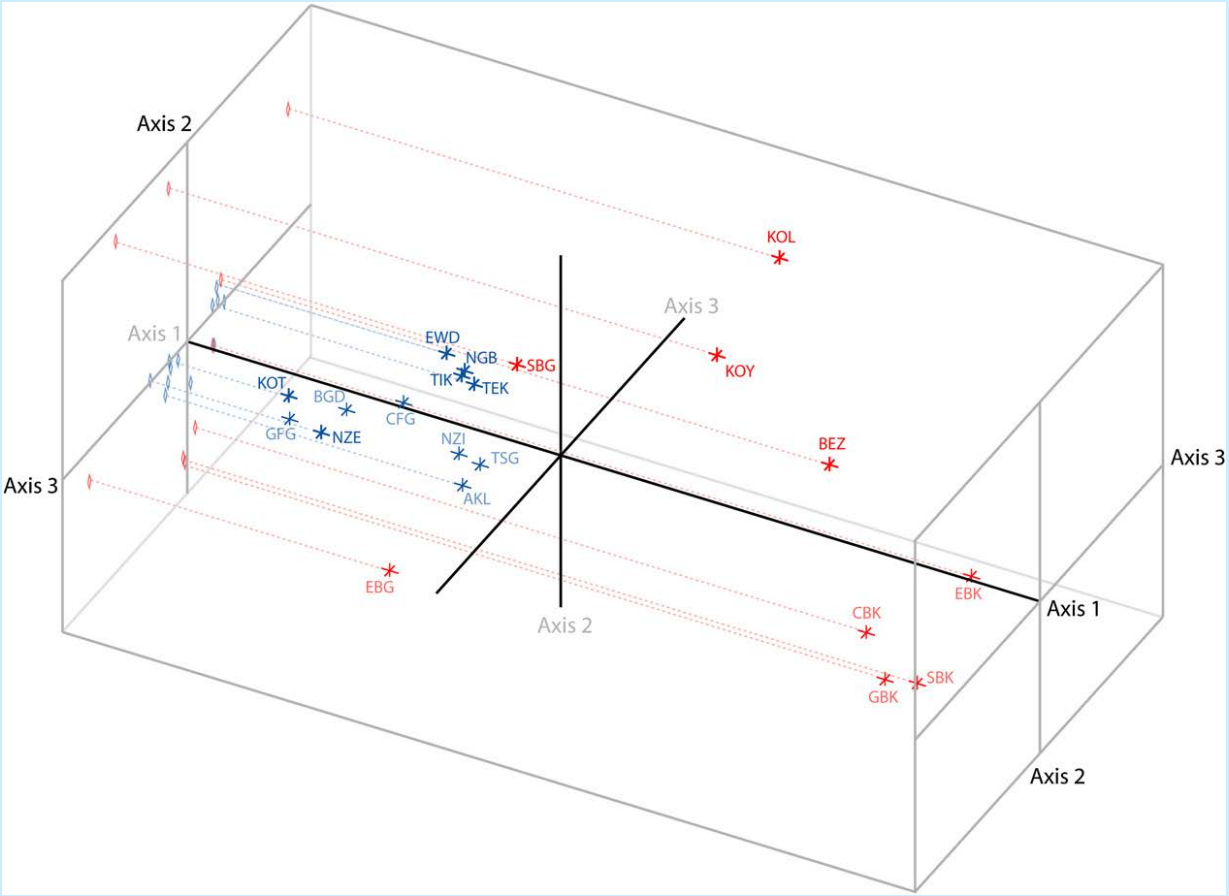
Pygmies and non-Pygmies in Central Africa



- Contrasted lifestyles (traditionally hunter-gatherers vs. farmers).
- Known difference in stature.
- Leave near each other and have complex socio-cultural relations.
- 21 populations: 9 Pygmy populations, 12 non-Pygmy populations
- 28 nuclear microsatellite loci.

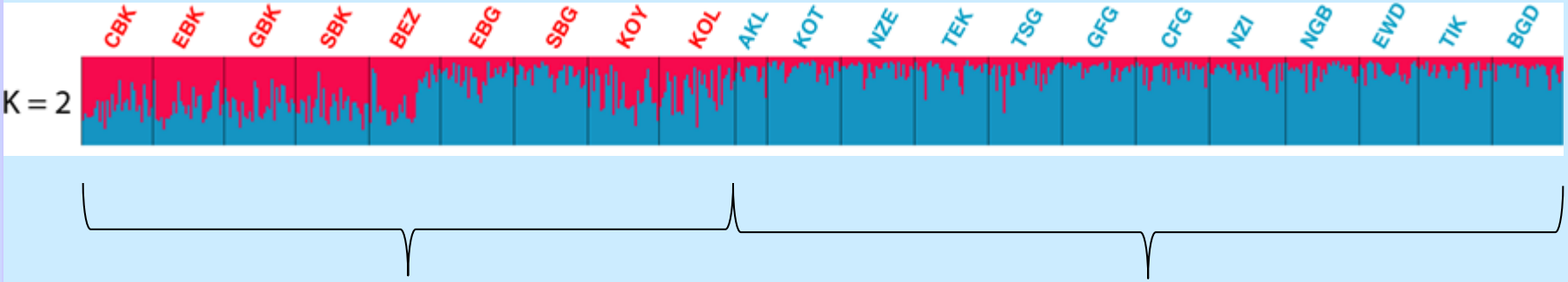
PhD Paul Verdu

PCA based on the pairwise F_{ST} values



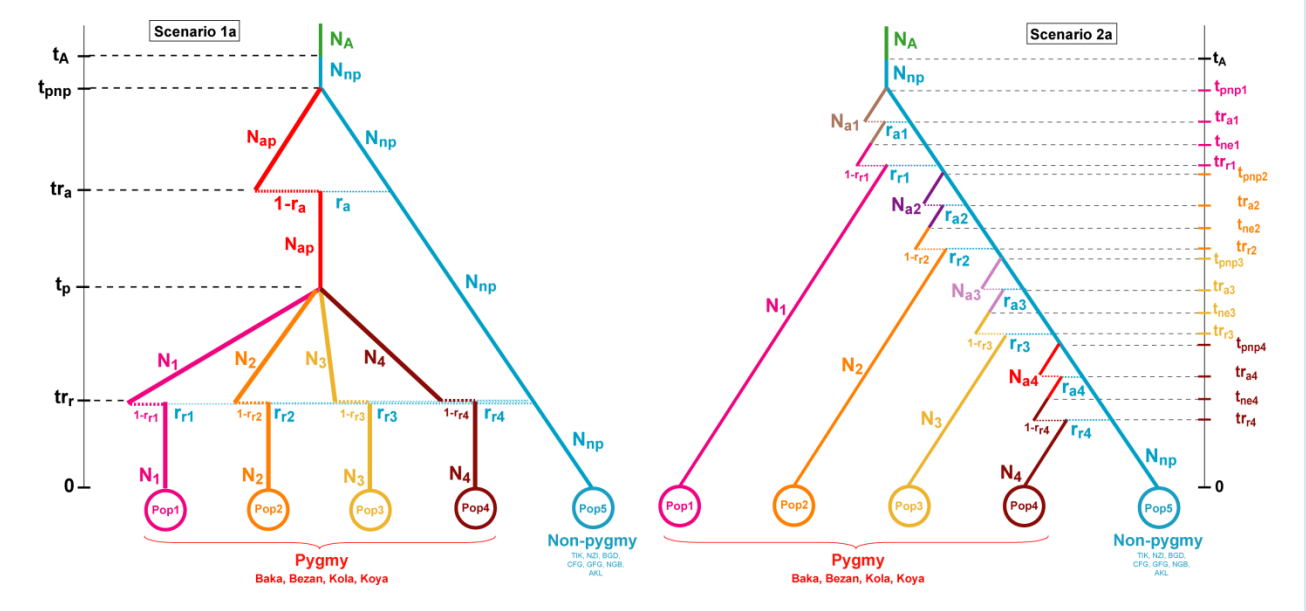
- Pygmy populations clearly differentiated from non-pygmies.
- Pygmy populations more scattered on the graph (more differentiated among each other).

Results with structure



- Pygmies and non pygmies are well separated
- Non-pygmies do not show much signal of introgression.
- Pygmies show variable level of introgression.

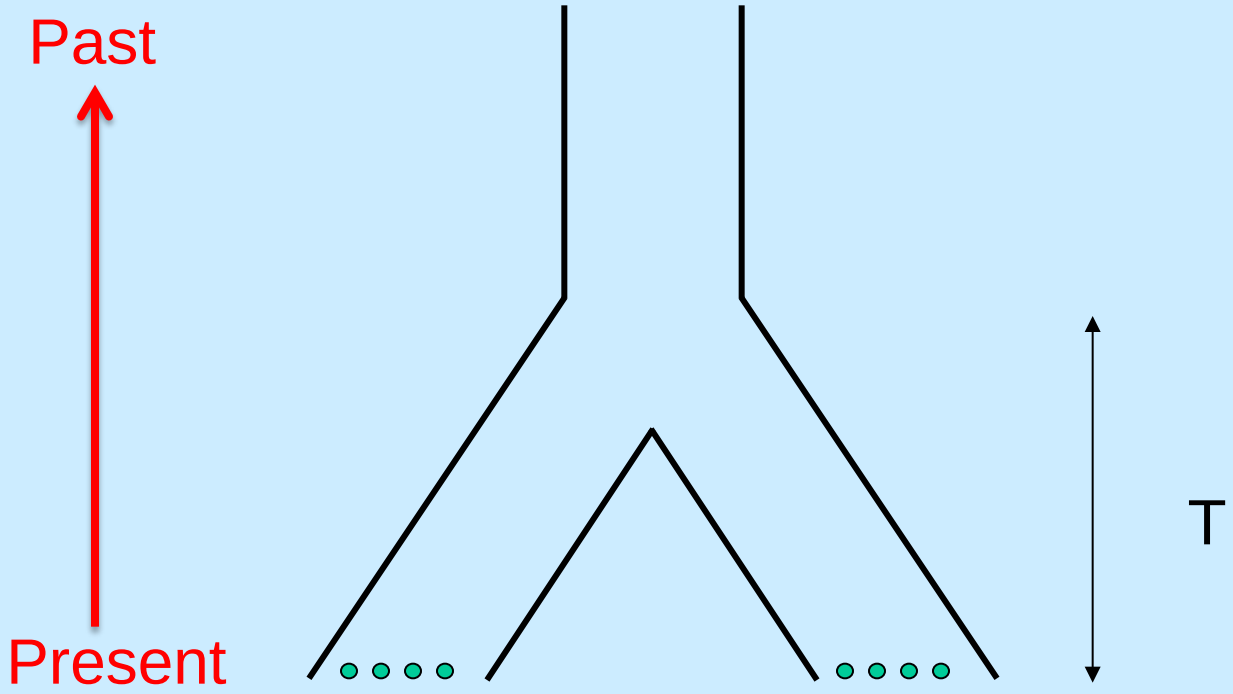
Determine the best scenario with ABC



Common origin Separate origin

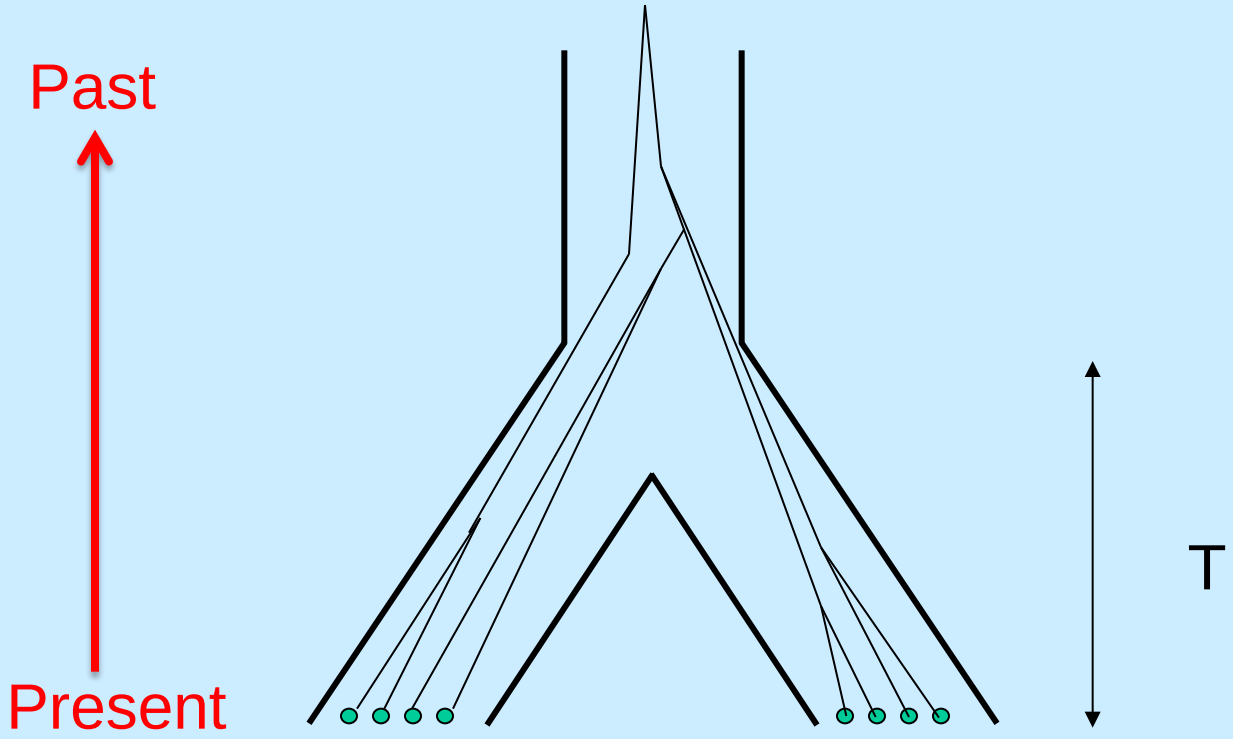
- Comparing two scenarios: common origin of all pygmies vs. separate origin of each pygmy population.
- Among both scenarios, scenarios with or without admixture from non-pygmies into non-pygmies are compared.
- Performed with the software DiY ABC (Cornuet et al, 2008)

Simulation method



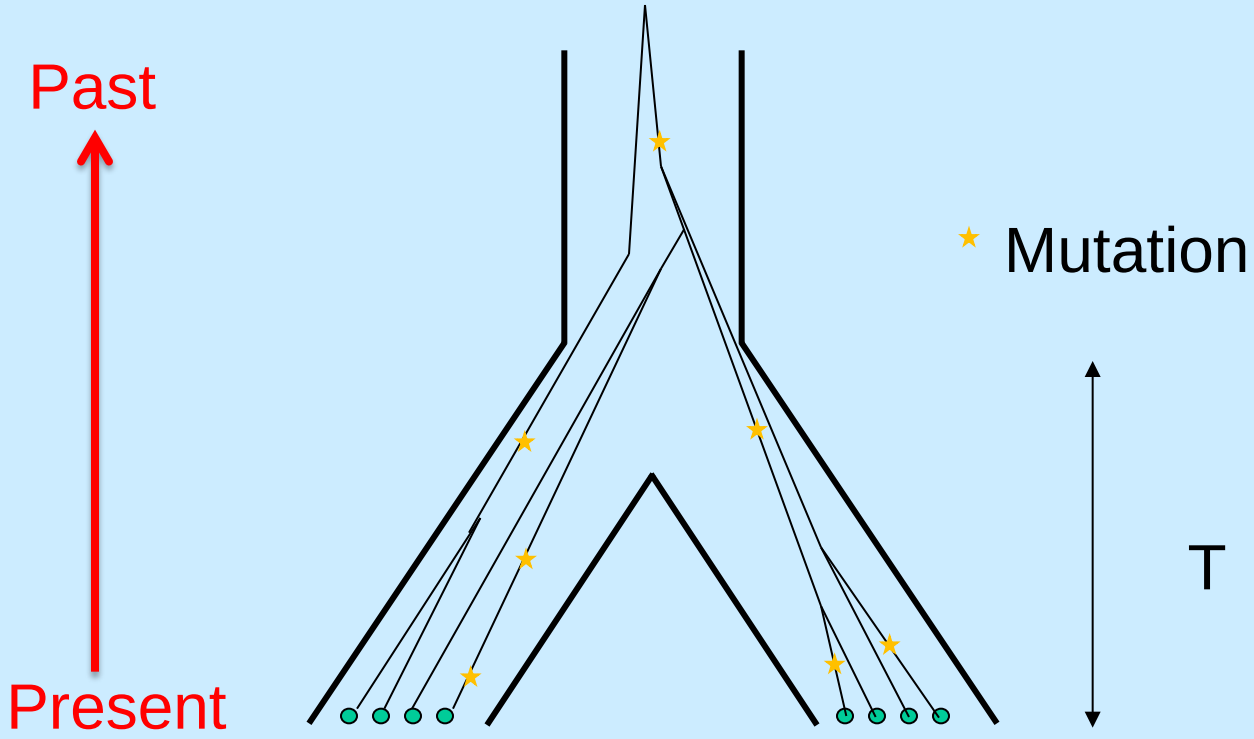
- We consider samples of individuals in the present populations.

Simulation method



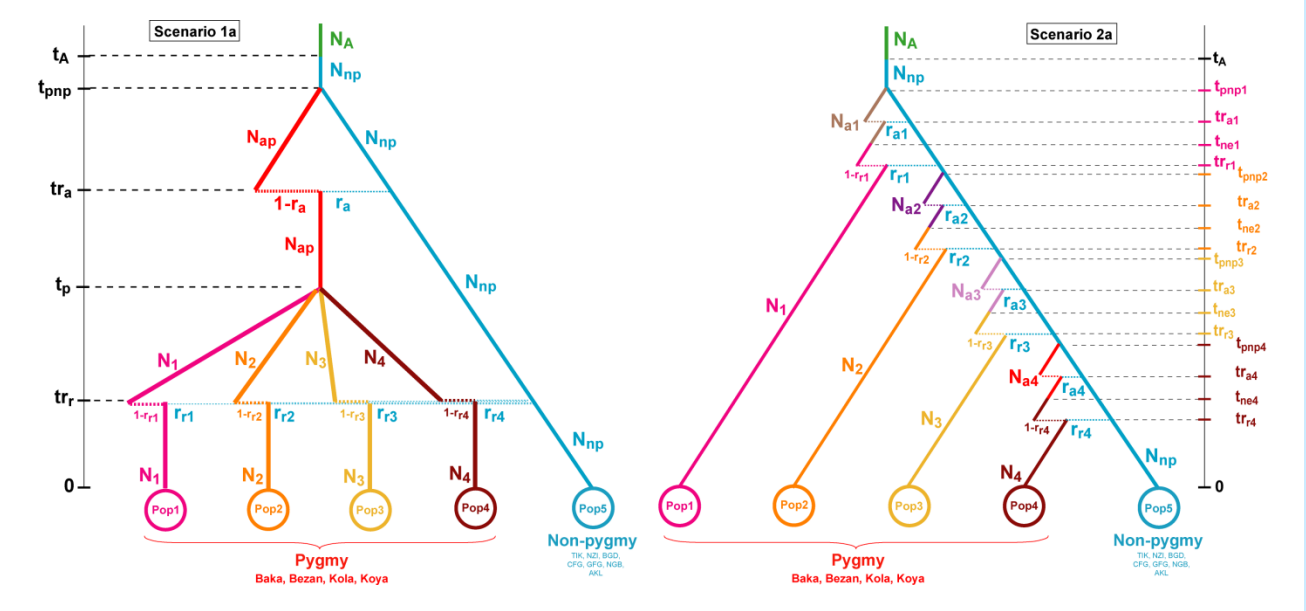
- We consider samples of individuals in the present populations.
- Simulations are performed backward using the coalescent process: generation of a coalescent tree

Simulation method



- We consider samples of individuals in the present populations.
- Simulations are performed backward using the coalescent process: generation of a coalescent tree
- Mutations are placed on the tree according to a given mutation rate (μ)

Determine the best scenario with ABC



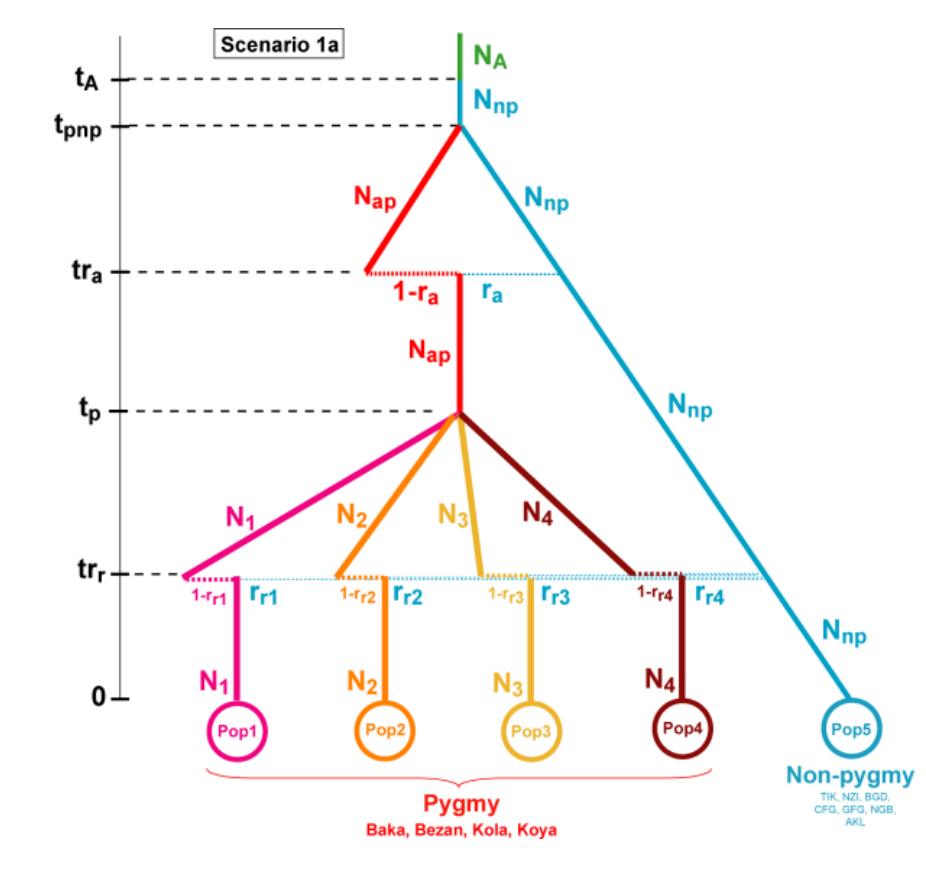
Common origin Separate origin

- We perform millions of simulations under each scenario.
- For each simulation, the size of each population (N_1, N_2, N_3, …), the split times (t_p, t_{pnp}, …) and the admixture rates are drawn in uninformative prior distributions.
- We obtain millions of simulated data sets under each scenario.
- Which scenario has produced simulated data sets closer to reality?

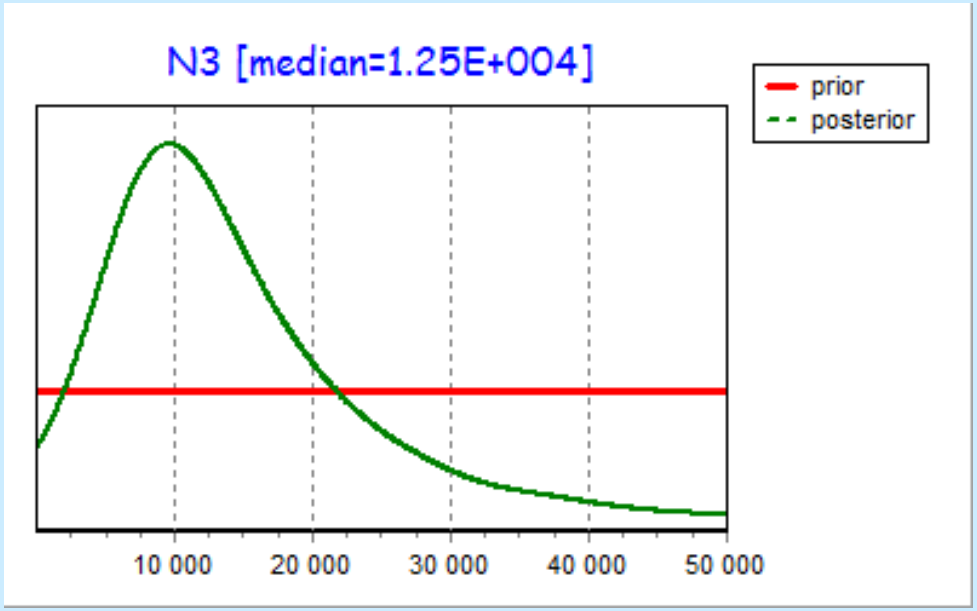
Assessing the proximity between real and simulated data sets

- We compute several summary statistics on the real data set and on each simulated data set.
  - e.g. genetic diversity within population (H_e), among-population differentiation (F_{ST}), …
- We pool all simulated data sets across all scenarios together and we keep a fraction (e.g. 0.1%) of these simulations that are the closest from the real data set according to the summary statistics.
- The scenario that is found in highest proportion among these selected simulations is considered the most likely.

The common origin scenario with admixture wins



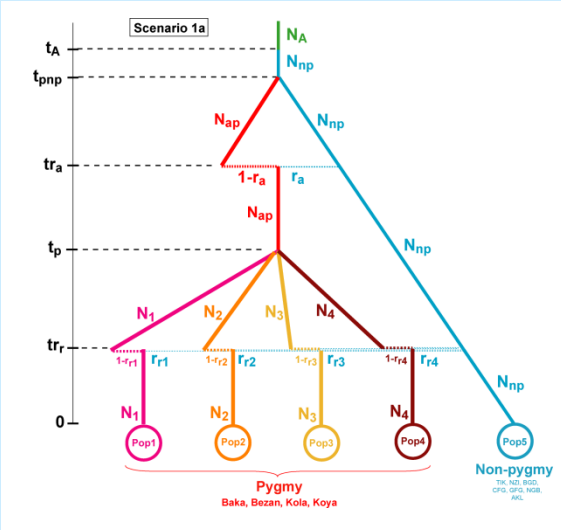
Estimating the parameter values in a given scenario.



- Keeping only the 0.1% best simulations, we can obtain posterior distributions for each parameter.

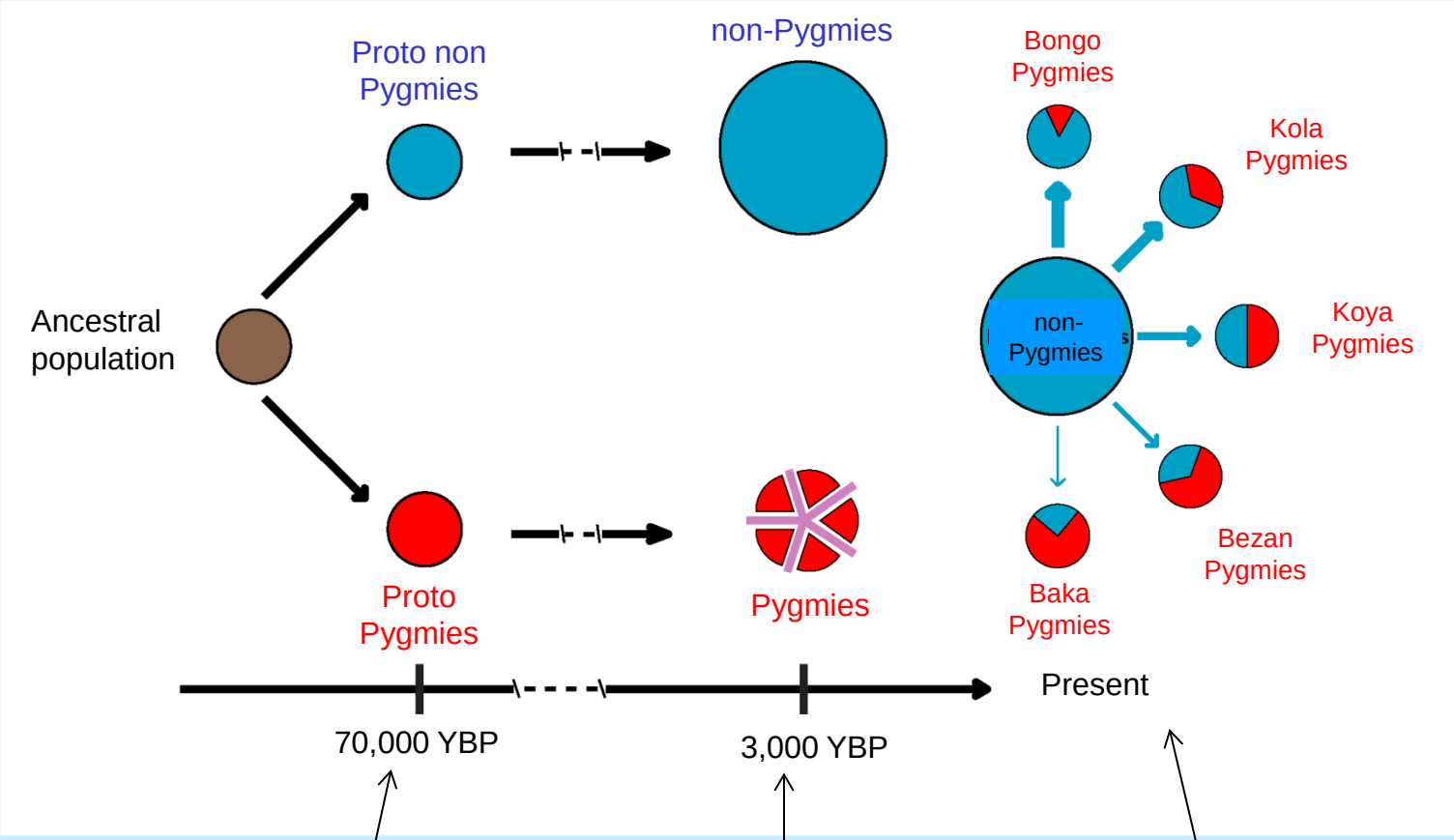
- We can then obtain modal estimates and 95% confidence intervals.

Estimated population sizes



N_1 (Baka)	8,137 (1,347 – 9,824)
N_2 (Bezan)	2,795 (790 – 9,677)
N_3 (Kola)	3,302 (603 – 9,599)
N_4 (Koya)	3,197 (1,134 – 9,771)
N_{np} (Non-pygmies)	77,157 (27,926 – 97,828)
N_{ap} (ancestral pygmy population)	8,007 (960 – 9,825)
N_A (ancestral population)	1,071 (202 – 8,404)

Inferred history



Separation between pygmies and non-pygmies

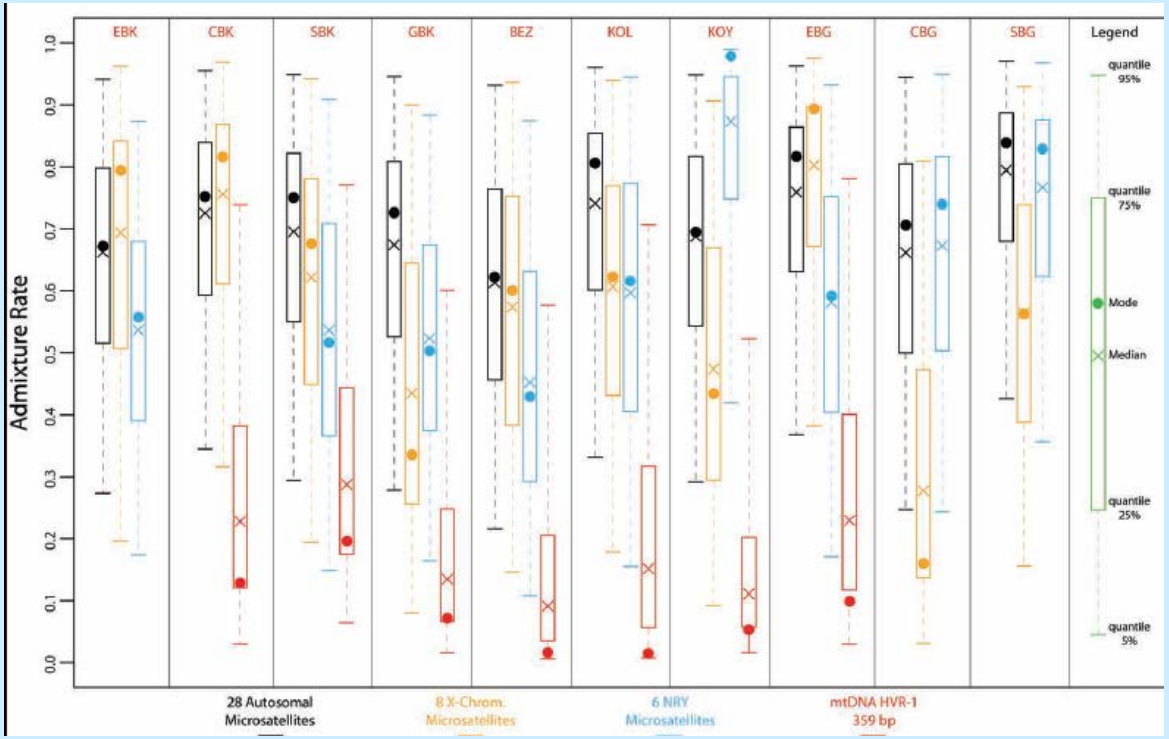
Non-Pygmy expansion Pygmy Splitting

Non-Pygmy admixture into Pygmy groups

Gender specific admixture

- Are the introgression patterns more through the female or the male line?
- We use the information from mitochondrial DNA and Y chromosomes and perform the same ABC estimations.

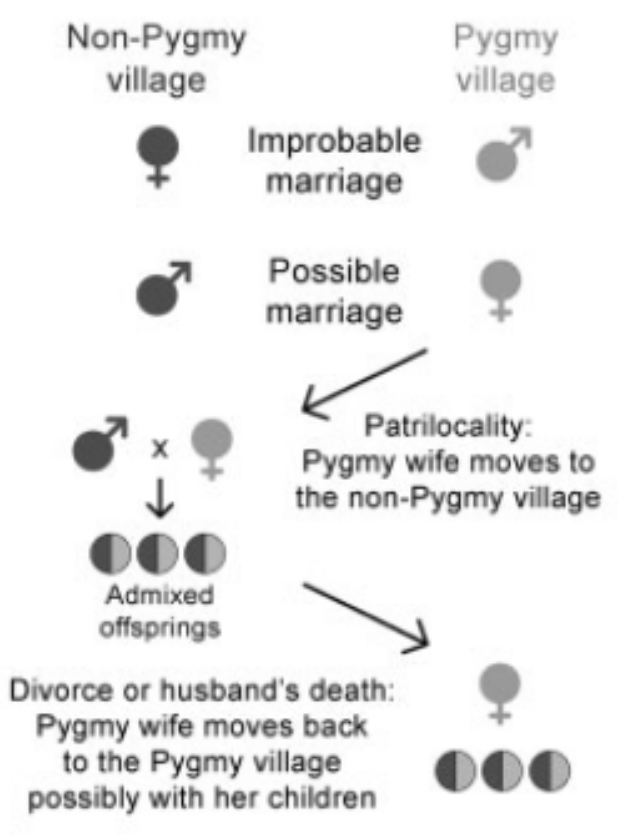
Admixture rate according to the different genetic systems estimated with ABC



- Much more introgression from non-Pygmies into Pygmies through the male line than through the female line.

Conclusion on gender-specific admixture

- We observed much more admixture through the male line than through the female line.
- This can be related to the social constraints between these two groups:
  - Intermarriages occur mostly between Non-Pygmy men and pygmy women.
  - The children of these marriages usually end up in the pygmy communities



History of harbour porpoise in the black sea

Phocoena phocoena

89 individuals genotyped for mitochondrial HV1 and autosomal microsatellite data

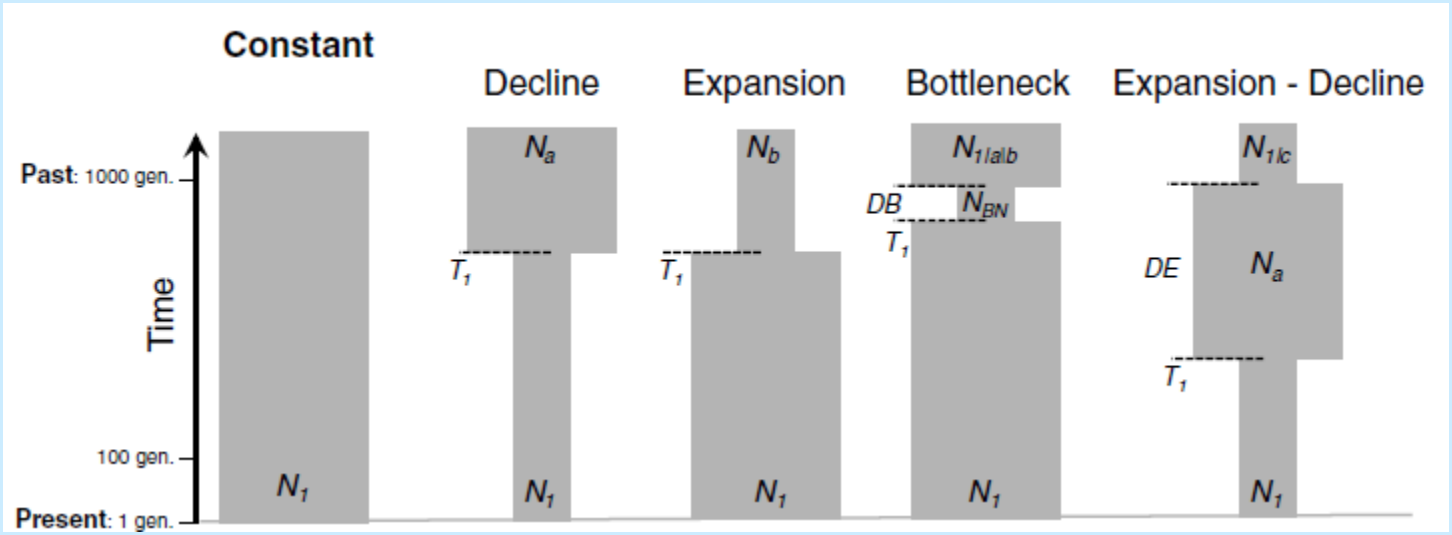
Some historical elements

- Arrived from the Mediterranean sea after its reconnection with black sea ~8000YBP (then disappeared from Mediterranean sea).

- Was severely hunted between the 1960's and 1980's.

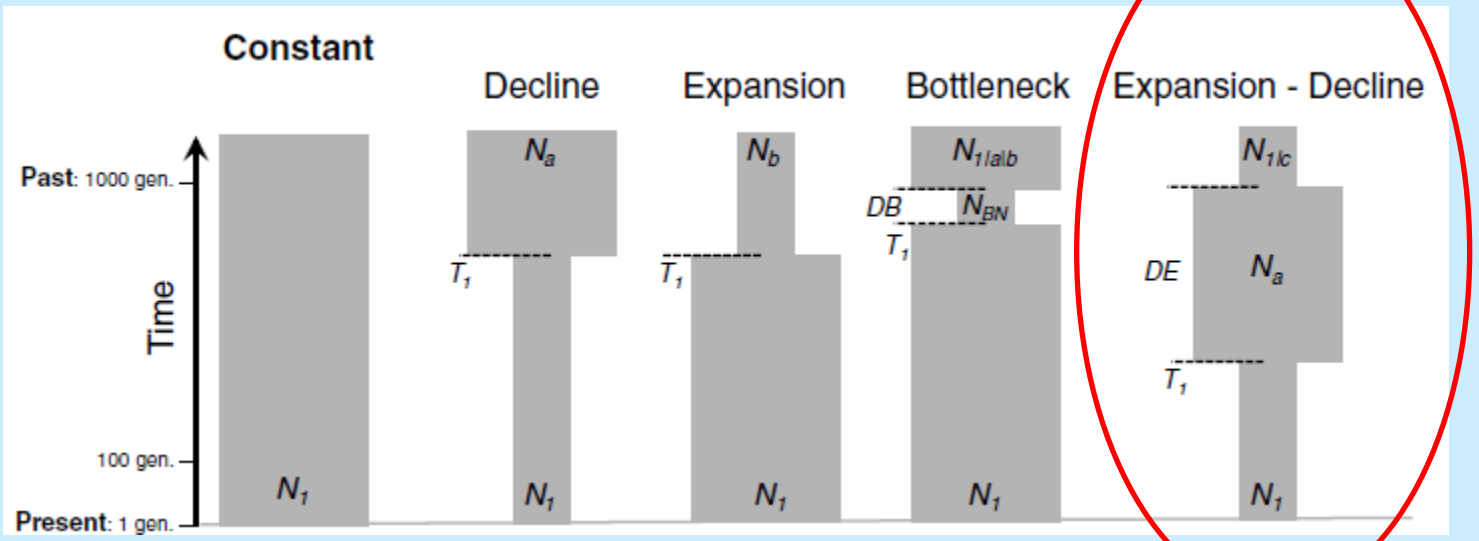
- Can we detect these events from genetic data?
- Can we infer their intensity?

Five possible scenarios



- ABC methods applied on nuclear and mitochondrial DNA data.

Five possible scenarios

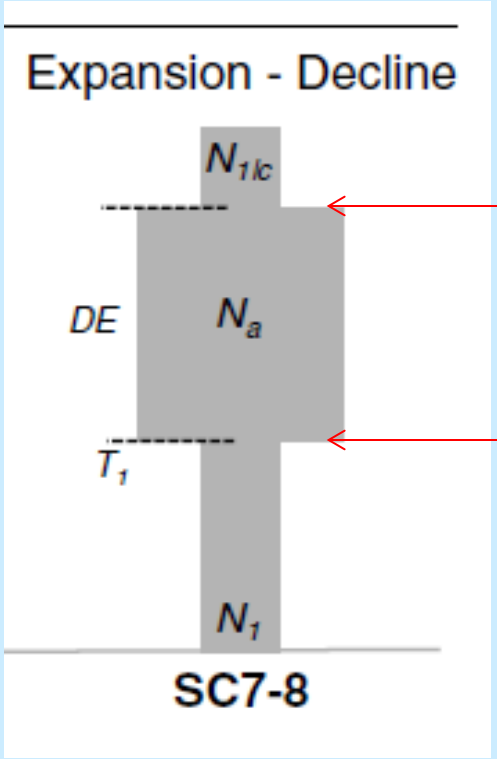


- ABC methods applied on nuclear and mitochondrial DNA data.

- The scenario of expansion-decline appears as the most likely.

Parameters estimated under this scenario

Past



Present

Expansion time 1,499 YPB (1,007 YBP– 4,897 YBP)
Expansion factor: 15 (4–72)

Decline time 6YBP (5 YBP– 48 YBP)
Reduction factor: 90% (84-97%)

- Expansion time: after the reconnection with the Mediterranean sea
- Extremely strong decline clearly related to extensive hunting.

Conclusions

- The flexible ABC framework allows to handle complex scenarios with for instance:
  - Multiple populations with admixture.
  - Several events of population change through time.
  - Different kind of markers.

- Need to be used with precautions (e.g. validation by simulation, control of the realism of the priors…)

Verdu P, *et al*. (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr. Biol.* **19**:312-318.

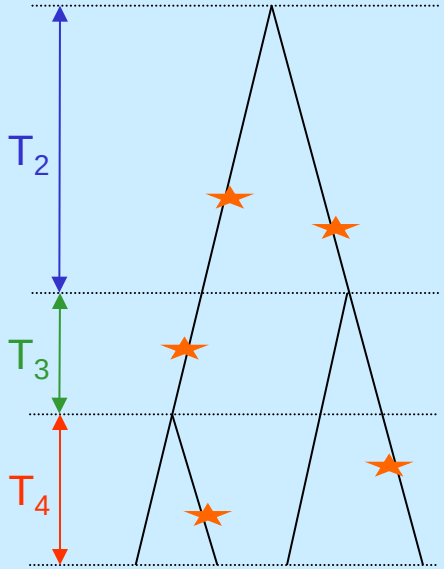
Verdu P, *et al.* (2013) Sociocultural Behavior, Sex-Biased Admixture, and Effective Population Sizes in Central African Pygmies and Non-Pygmies. *Mol. Biol. Evol.* 30:918-937.

Fontaine MC, *et al*. (2012) History of expansion and anthropogenic collapse in a top marine predator of the Black Sea estimated from genetic data. *Proc. Natl. Acad. Sci. USA* **109***:*E2569-E2576.

Likelihood methods

- We could in theory compute the likelihood of the data (D) given the parameters (Θ)

$$L(\Theta) = P(\mathrm{D} \mid \Theta) = \sum_{g \in \mathbf{G}} P(\mathrm{D} \mid g) P(g \mid \Theta)$$



- g is a coalescent tree
- G is the set of all coalescent trees
- P(g | Θ) is obtained with the coalescent theory
- P(D | g) is obtained by assuming a given mutation process on the tree

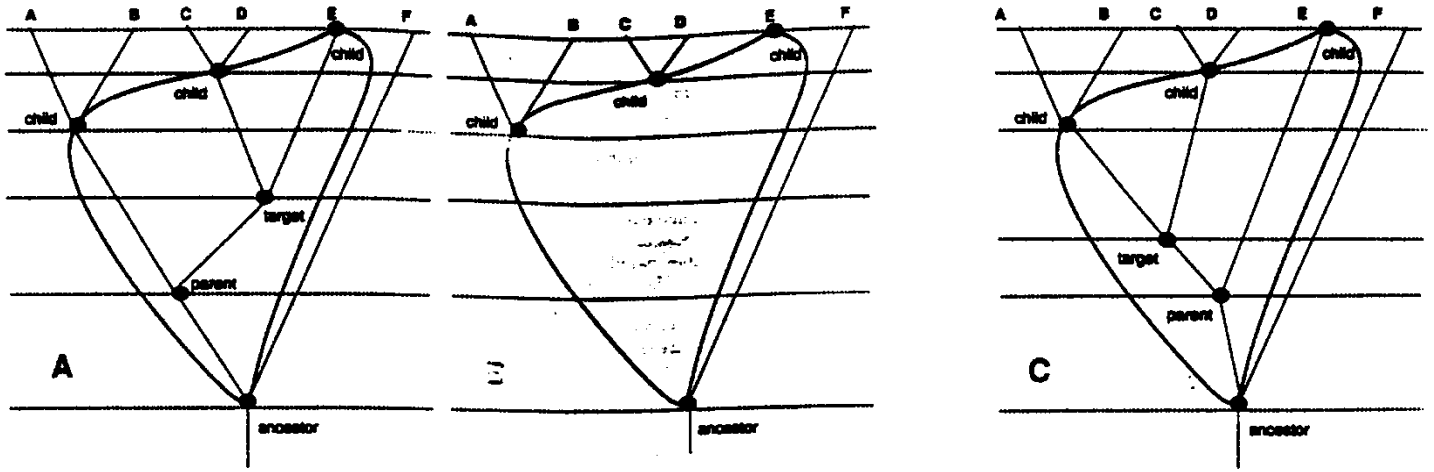
Problem: the total number of trees is very high

Sample size	Number of Possible trees
5	105
10	34459425
15	2.13×10^{14}
20	8.20×10^{21}

- Impossible to consider all tree
- We will explore a subset of possible trees

Monte Carlo Markov Chains (MCMC) (Metropolis-Hastings)

- Random permutations within the tree



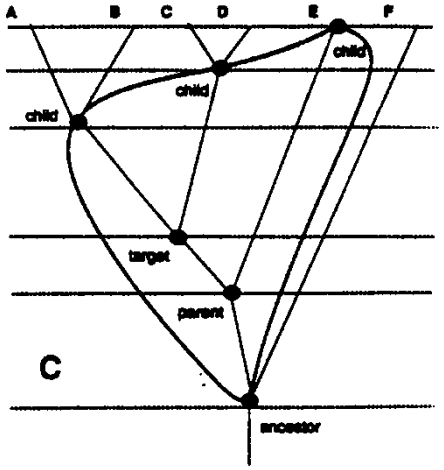
Tree0

Tree1

- The new tree is conserved
  - Always if P(D|Tree1) > P(D|Tree0)
  - With a propabilility *r* = P(D|Tree1) / P(D|Tree0) otherwise
- Converge to the trees that best explains the data (highest likelihood)

Monte Carlo Markov Chains (MCMC) (Metropolis-Hastings)

- Random changes of the parameters
- We start with given set of parameters (Θ_0) and draw a new set of parameters (Θ_1)

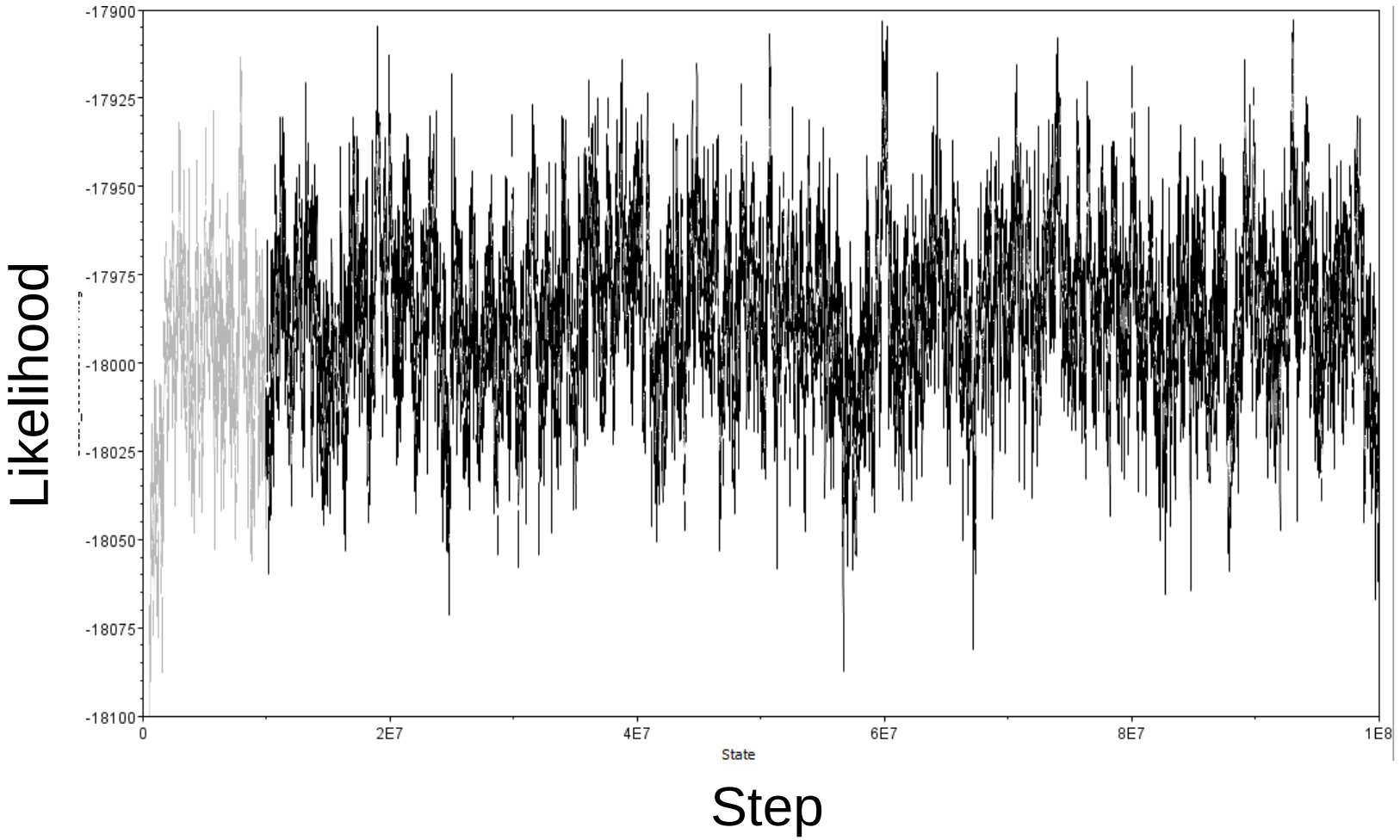


Tree1

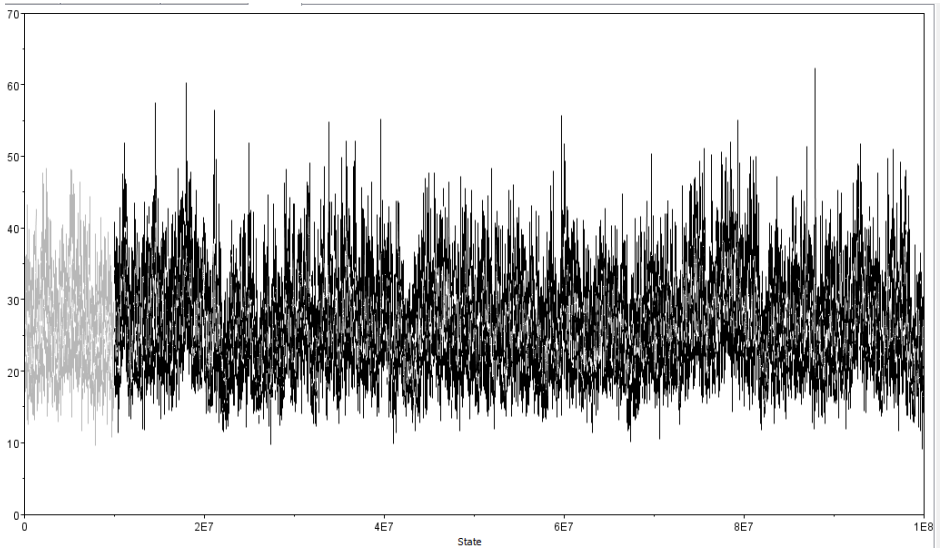
- The new parameters are conserved
  - Always of P(Tree1| Θ_1) > P(Tree1| Θ_0)
  - With a propabilility *r* = P(Tree1| Θ_1) / P(Tree1| Θ_0)

- Converge to the parameters with highest likelihood

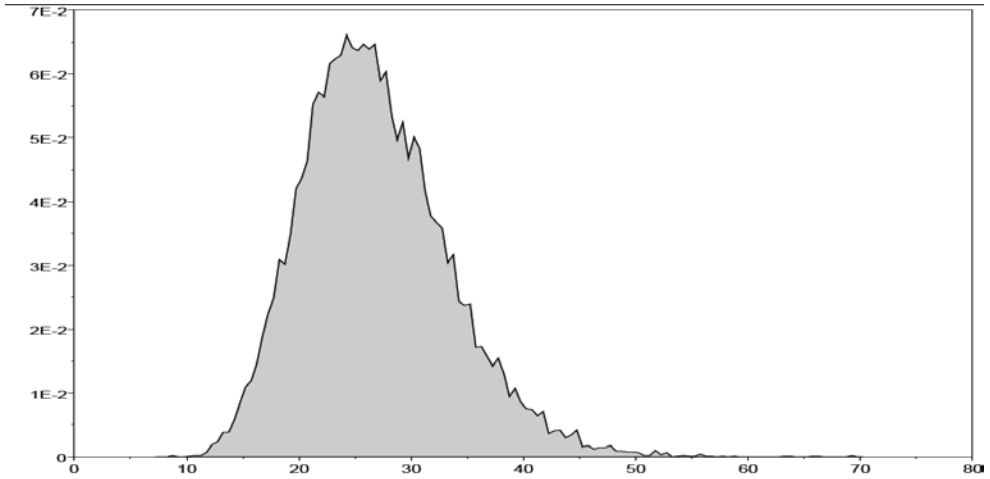
Evolution of the likelihood through the steps



Evolution of a given parameter

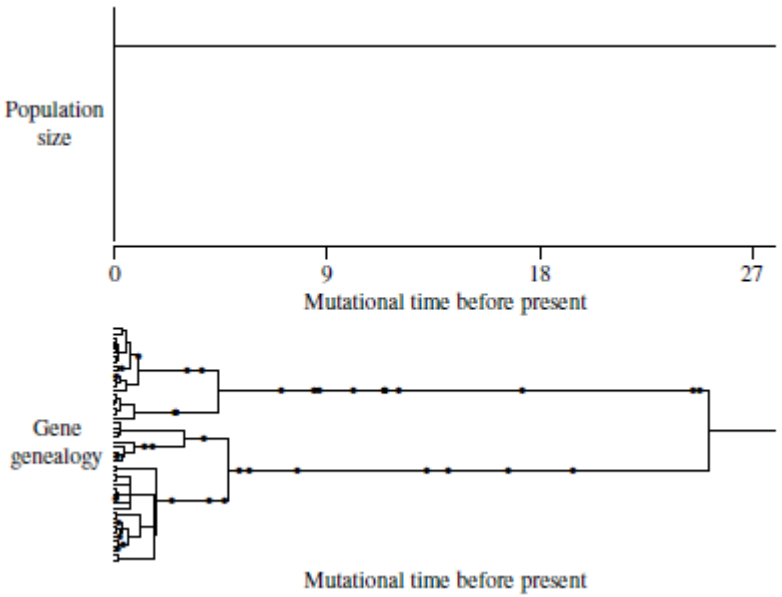


➔ Posterior distribution

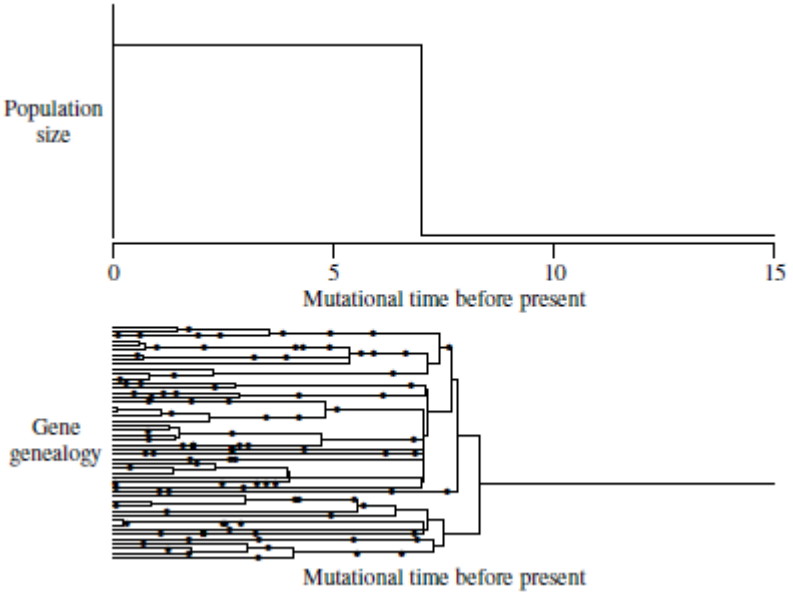


Impact of demography on the coalescent tree

Constant population

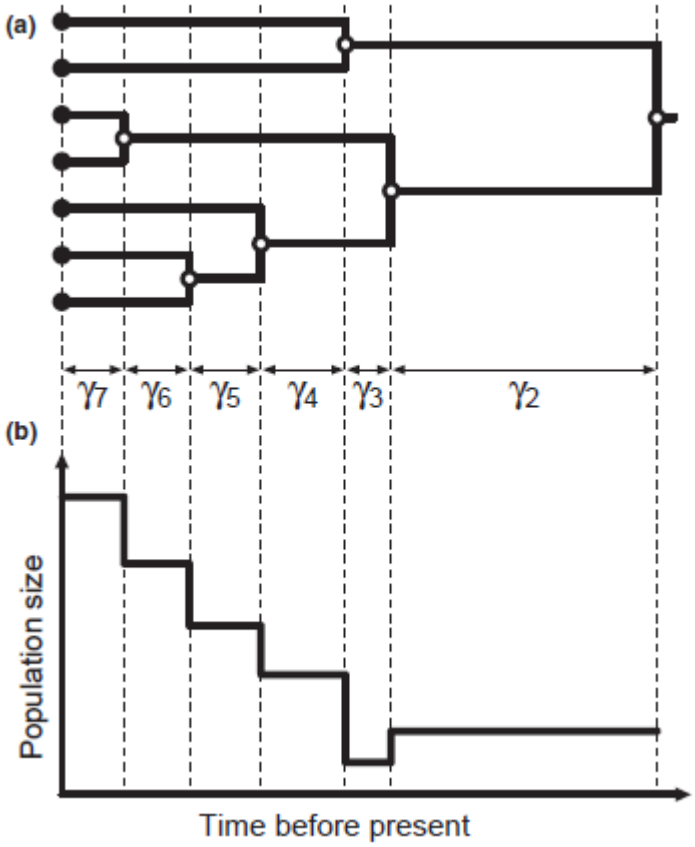


Expanding population



Non-parametric approach

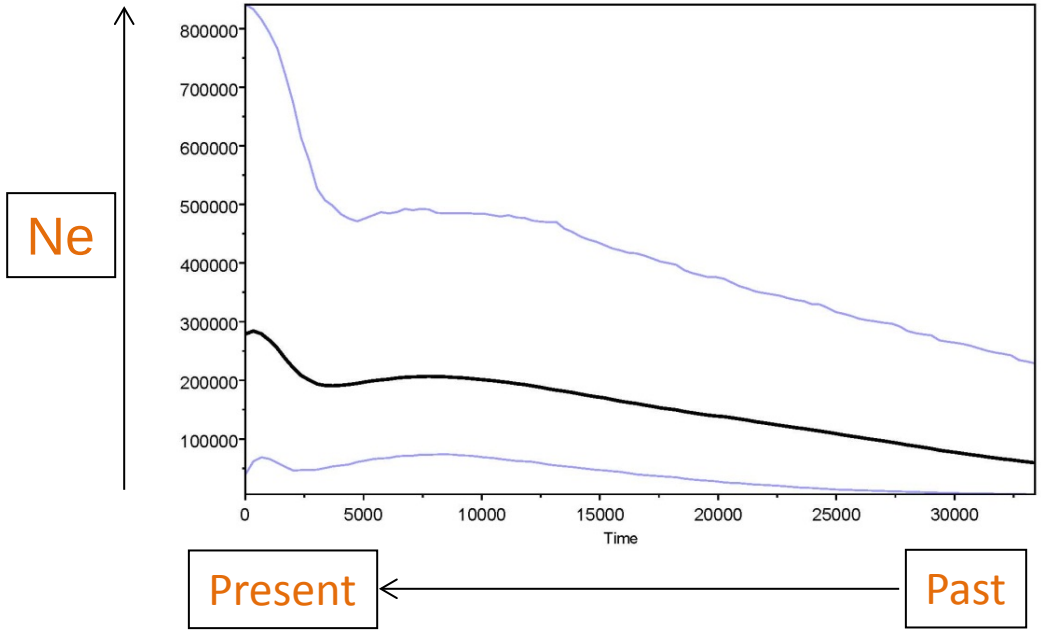
- **Skyline plot:** Uses the time intervals between coalescent serial events.



- Uses the same MCMC approach.

Non-parametric approach

- Permit us to graphically visualize the evolution of population size through time (skyline plot)



[1] Drummond *et al.*, 2005.
[2] Heled et Drummond, 2010.

life-styles among human populations

Before The Neolithic revolution
(12,000 – 5,000 BP)

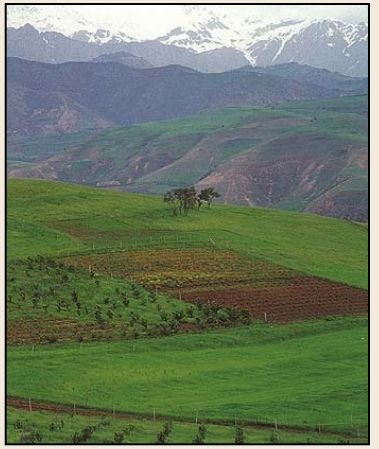
Hunter-gatherers

life-styles among human populations

After

Still hunter-gatherer populations

Nomadic herders

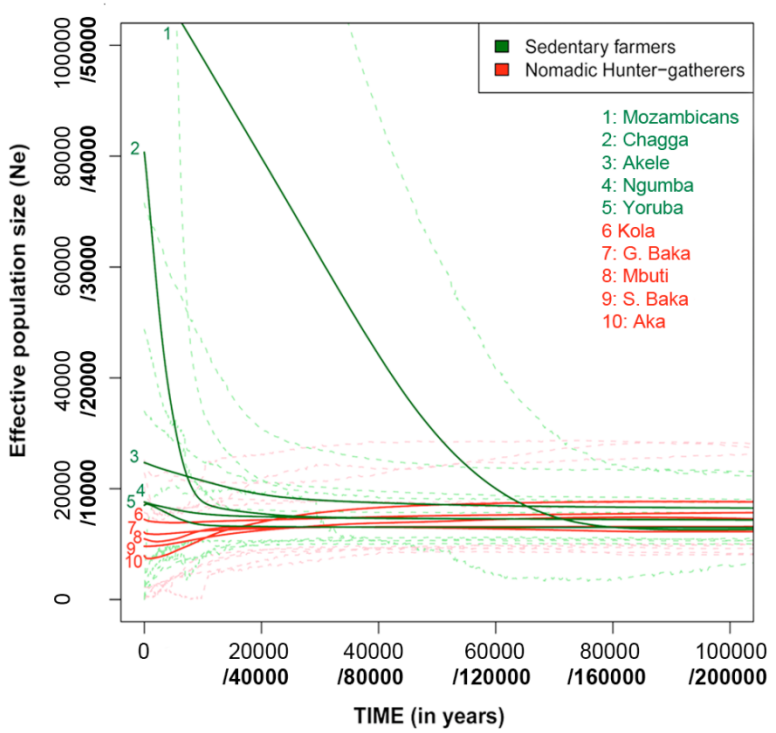


Sedentary farmers

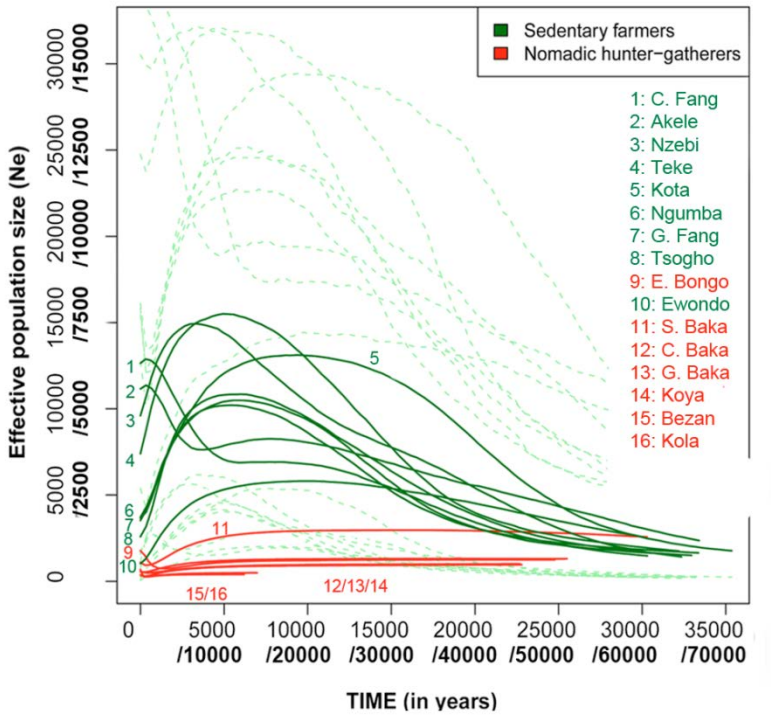
Different demographic patterns?

Results for the non-parametric method (Africa)

autosomes



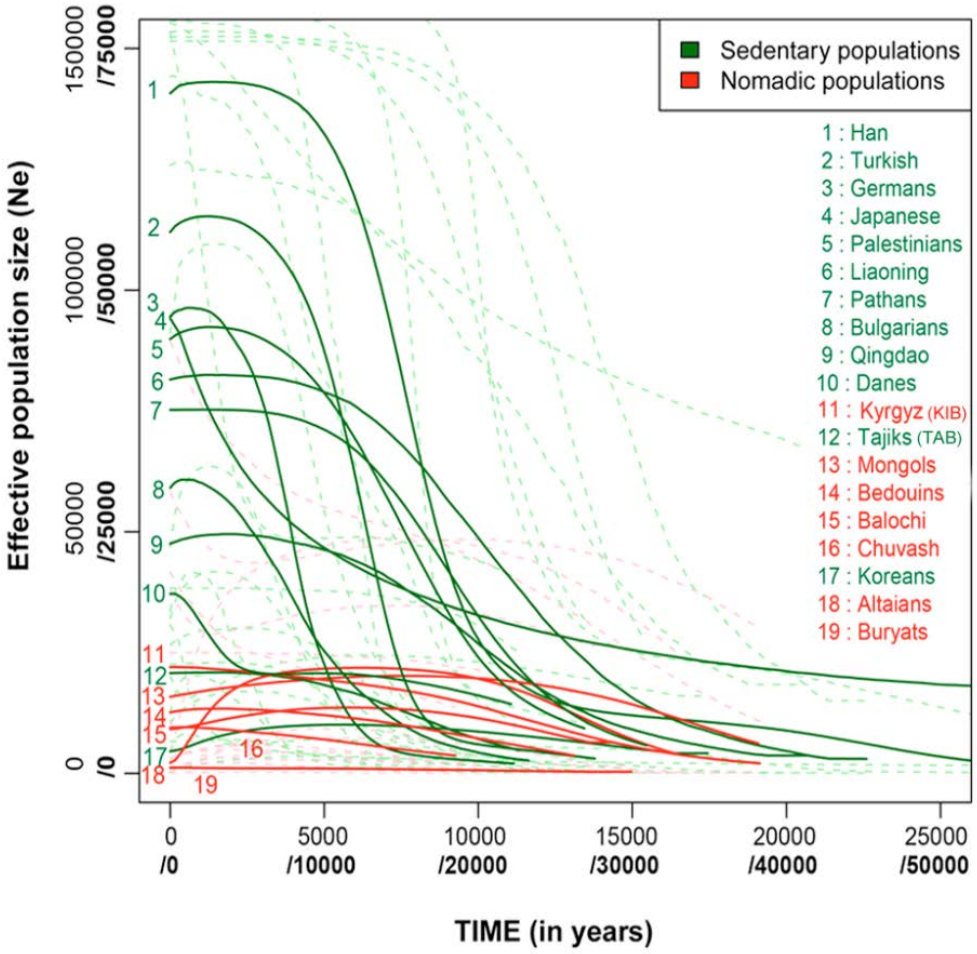
HVS-I



➔ Inferred expansions of current farmers largely predate the emergence of farming about 5,000 YBP !!

Results (Eurasia)

HVS-I

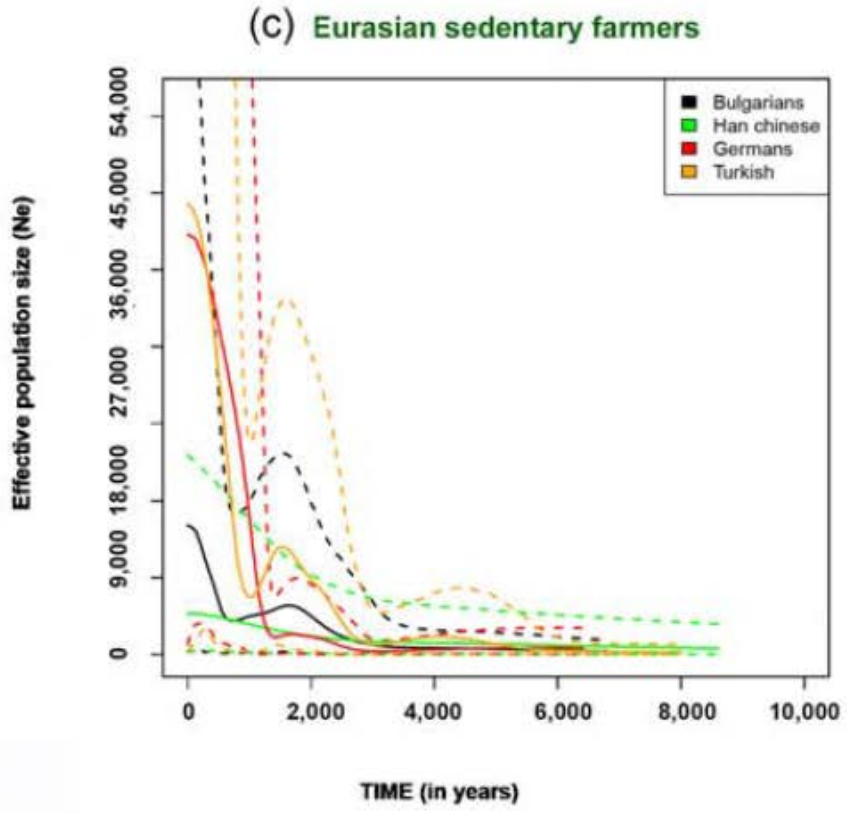


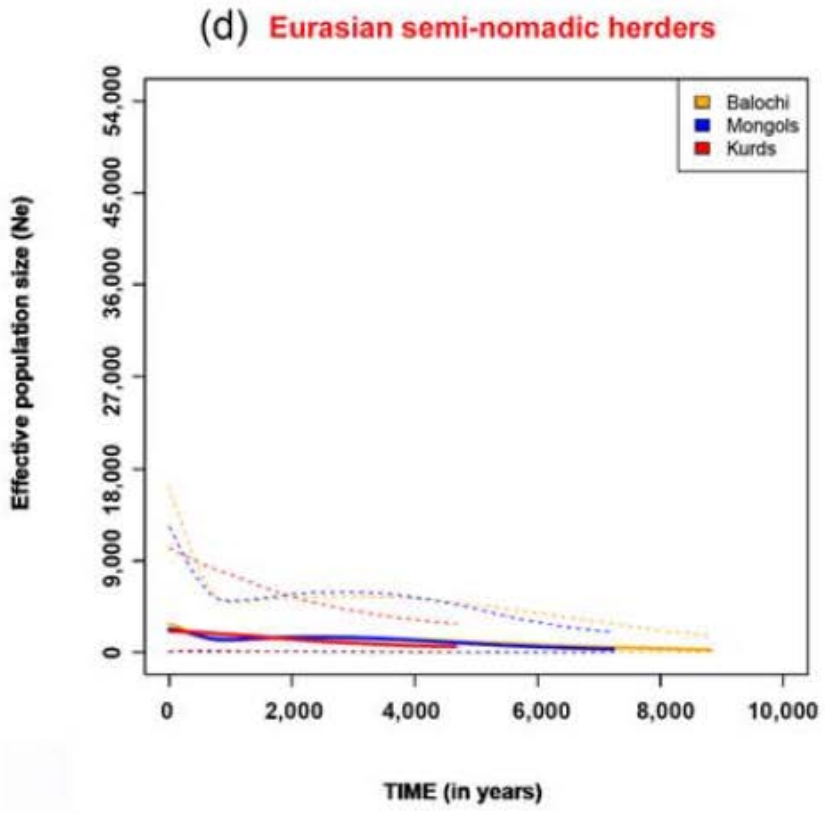
Stronger **Paleolithic** expansions

in farmer than herder populations

Results (Eurasia)

Y chromosome





Impact of life-style on expansion patterns

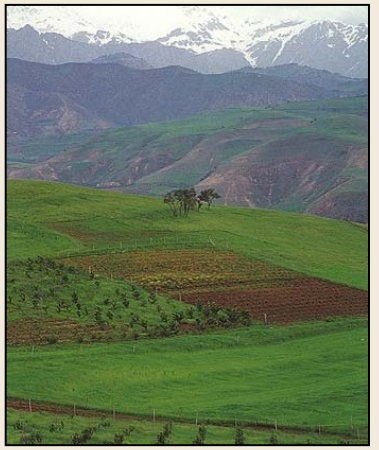
Still hunter-gatherer populations

No expansions

Nomadic herders

Weak expansions

Sedentary farmers

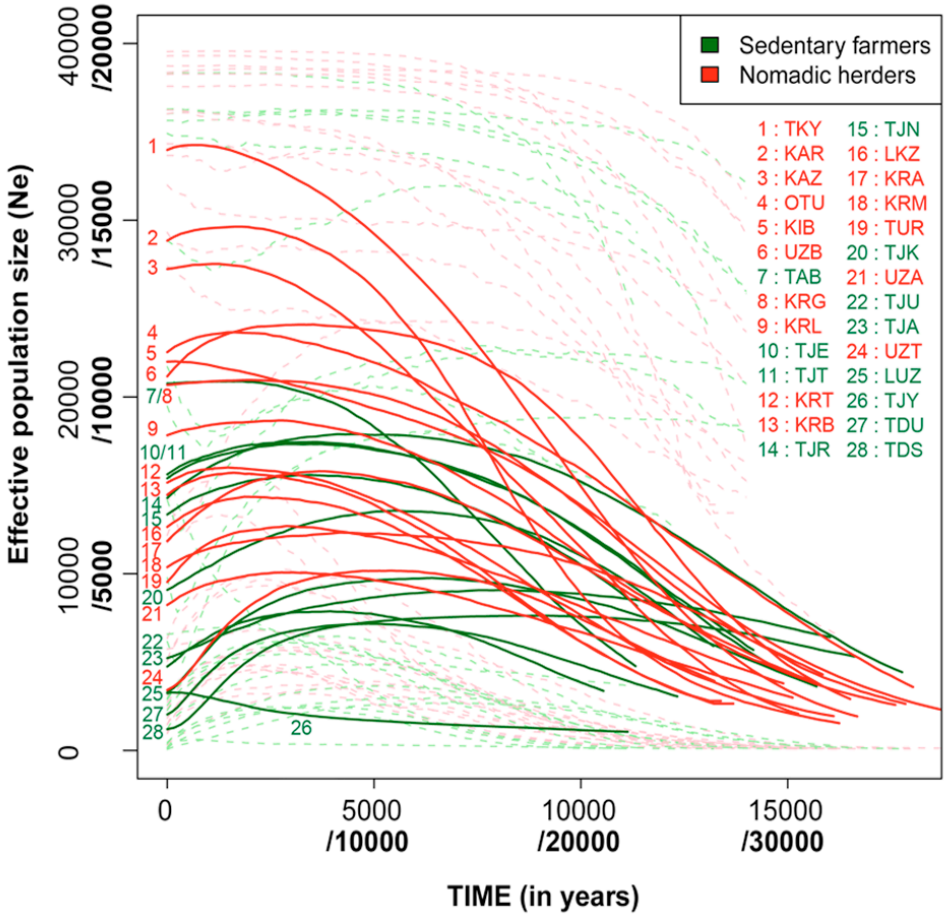


Strong expansions

- Might be linked with the lifestyle of these different groups.
- Most expansions started in the Paleolithic time.
  - Probably connected with technological changes that predated the Neolithic.

Focus on Central Asia

HVS-I

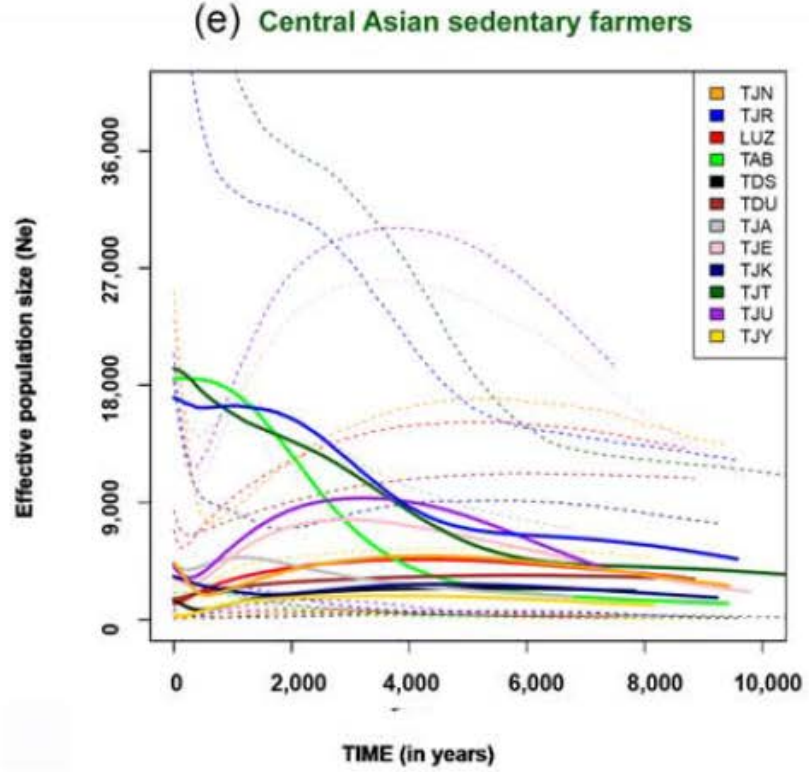


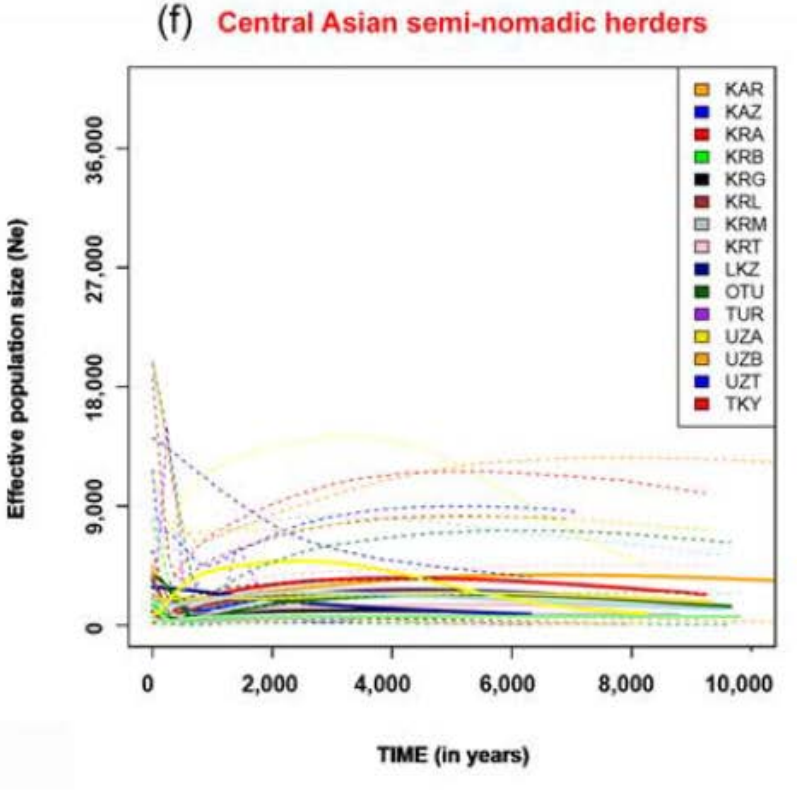
Expansions predating the Neolithic

Tendancy to stronger expansions in nomadic than in sedentary populations?

Focus on Central Asia

Y chromosome





Some local specificities

- Weak expansions Central-Asians farmers
- Can be linked with local environmental specificities: arid climate in Central-Asia
- Differences between men and women, which may be linked to differences in the migration process.

Aimé et al 2013. Molecular Biology and Evolution 30: 2629-2644.

Aimé et al 2014. European Journal of Human Genetics 22: 1201-1207.

Aimé et al 2015. American Journal of Physical Anthropology 157: 217-225

Conclusions on MCMC methods

- Allow to choose between several demographic models.
- Allow to make inferences in models with limited number of parameters.
- Are not suited in situations where there are too many parameters

↳ Development of Approximate Bayesian computation (ABC) methods: not as accurate as MCMC but allows more flexibility.

Perspective: extension to high-throughput genomic data

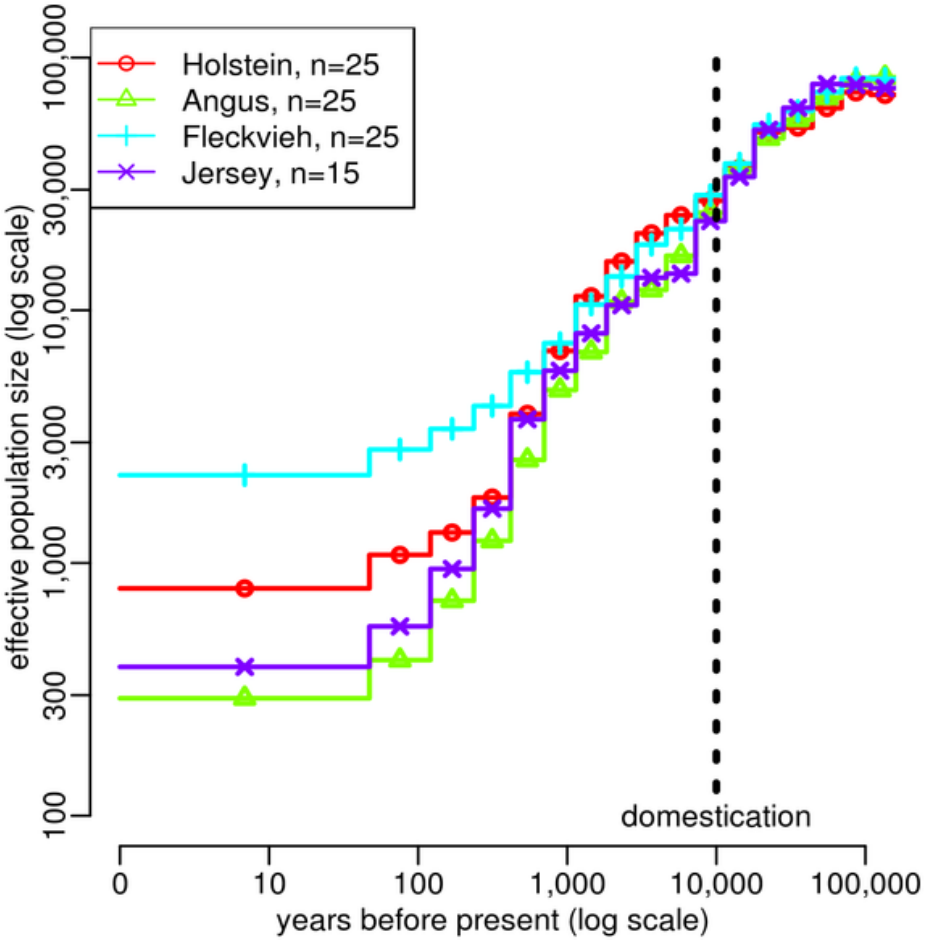
- ABC and MCMC methods are currently extended to DNA chips data and next generation sequencing data (e.g. full genomes).

- Several pitfalls are to be considered:
  - Very large number of data to simulate: this can be solved with specific tools (e.g. sequential Markov coalescent).
  - Ascertainment bias in the DNA chips data.
  - Quality of the sequences.

Perspective: extension to high-throughput genomic data

- These data present however several advantages:
  - Markers at various genetic distances: possibility to use linkage disequilibrium.
  - Very large amount of data: can reduce variance in the estimates if correctly handled.

A skyline plot model with ABC



Boitard S, Rodríguez W, Jay F, Mona S, Austerlitz F (2016) Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. PLoS Genet 12(3): e1005877. doi:10.1371/journal.pgen.1005877

Conclusions

- ABC (and MCMC) methods allow to infer demography on many kinds of genetic data (microsatellites, small sequences, DNA chips, next generation sequences).

- Currently also applied to other kind of data (example linguistic data, PhD thesis of Valentin Thouzeau).