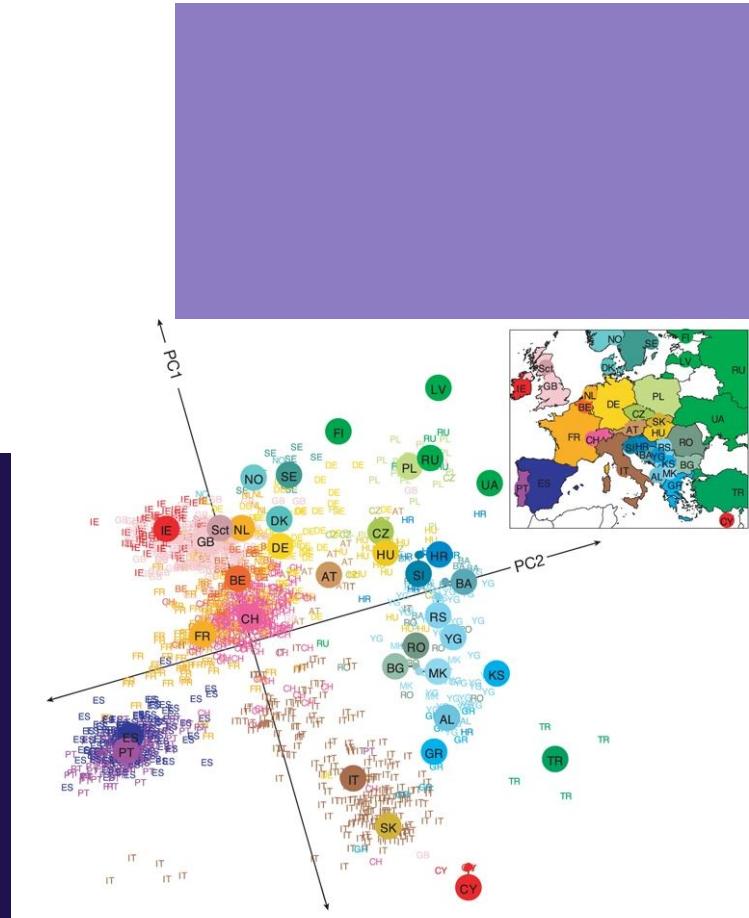


(Intro to) Demographic modeling methods

28/05/2025

Physalia course

Thibault Leroy, Yann Bourgeois



Goals for today's lecture

- Understand the importance of coalescence rates in demographic modeling
- Reconstruct past population sizes from one or a few genomes
- Introduce the interest of more complex methods based on likelihood & ABC algorithms
- Identify which method should be preferred regarding the dataset/research question
- Introduce pros and cons of each method

Current genetic diversity = long-term processes and past history

Genetic diversity is highly variable among the tree of life!

π = the average number of nucleotide differences per site between pairs of sequences

1:AAATACCA**A**ACAC
2:AAATACC**C**ATCAAC
3:AAATACC**T**CAAG
4:AAATACC**T**CAAC
5:AAATACC**T**CGAC

Between two sets of human chromosomes, one SNP in every 1,000 nucleotides on average (human genome size ~3.1 Gb, which means that your own genome contains roughly 3 million heterozygous sites)



$\sim 1 \times 10^{-3}$



Current genetic diversity = long-term processes and past history

Genetic diversity is highly variable among the tree of life!

Between two sets of chromosomes, one SNP in every 12.5 nucleotides on average, which means 80 heterozygous sites per Kb in a single diploid individual
(genome size: 180 Mb => 14.4 million hz sites)



Ciona savignyi
 $\pi \sim 8 \times 10^{-2}$

Between two sets of human chromosomes, one SNP in every 1,000 nucleotides on average (human genome size ~3.1 Gb, which means that your own genome contains roughly 3 million heterozygous sites)



$\sim 1 \times 10^{-3}$



Current genetic diversity = long-term processes and past history

Can we understand why some species are more genetically diverse than some others?



Leffler et al. Plos Biol 2012

Current genetic diversity = long-term processes and past history

Can we understand why some species are more genetically diverse than some others?

At mutation-drift equilibrium = assuming constant population sizes

$$\theta = 2 * \text{ploidy} * N_e * \mu$$

$$(\text{i.e. } \theta \text{ (diploid species)} = 4N_e \mu)$$

At equilibrium $\theta = \pi$



Current genetic diversity = long-term processes and past history

Can we understand why some species are more genetically diverse than some others?

At mutation-drift equilibrium =
assuming constant population sizes

$$\theta = 2 * \text{ploidy} * N_e * \mu$$

(i.e. θ (diploid species) = $4N_e \mu$)
At equilibrium $\theta = \pi$

How to explain this low present-day diversity? Is it linked to the past history of the species?



Lynx lynx
 $\sim 2.0 \times 10^{-4}$

Lynx pardinus (Iberian)
 $\sim 1.0 \times 10^{-4}$



Current genetic diversity = long-term processes and past history

Can we understand why some species are more genetically diverse than some others?

π = the average number of nucleotide differences per site between pairs of sequences

1:AAATACCA**A**CAAC
 2:AAATACC**C**TAAC
 3:AAATACC**T**CAAG
 4:AAATACC**T**CAAC
 5:AAATACC**T**CGAC

Deviations from mutation-drift equilibrium
 (Genome-wide deviations from Tajima's D = 0)

$$\theta = \# \text{ polym. sites/harmonic num.}$$

$$= \sum_{i=1}^{n-1} \frac{1}{i}$$

$$\theta = 3 / 2.083 = 1.44$$

(0.11 per bp)

$$\text{Tajima's D} = \frac{\pi - \theta}{\sqrt{\text{var}(\pi - \theta)}}$$

At mutation-drift equilibrium:
 Tajima's D~0

D>0: Deficit of rare alleles = population contraction

D<0: Excess of rare alleles = population expansion



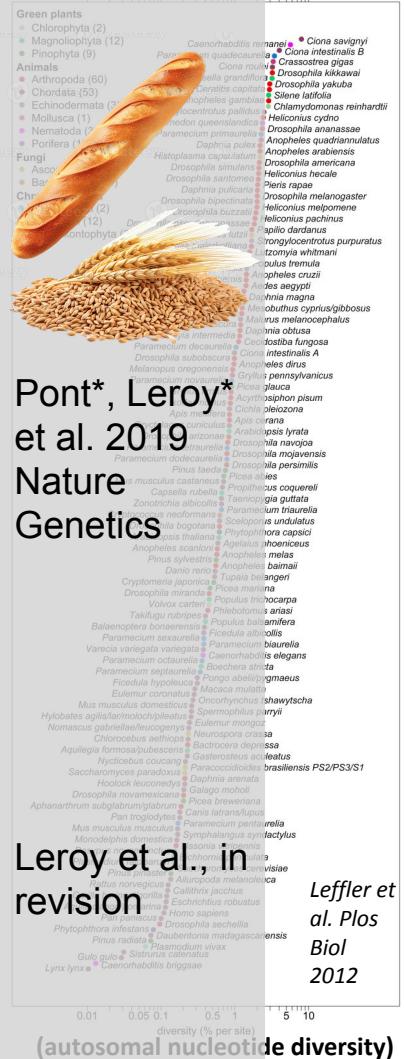
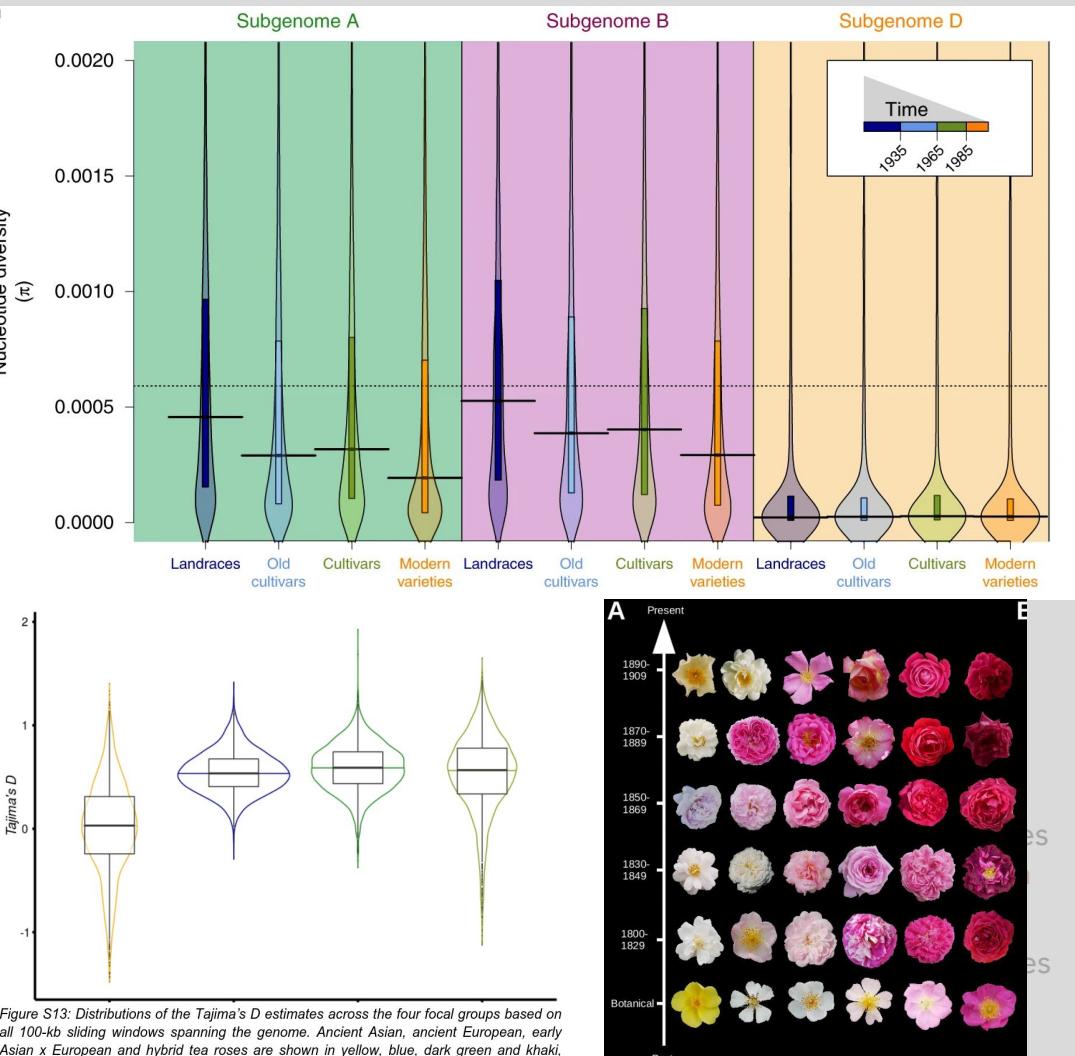
Simple but often very meaningful summary statistics!

Deviations from mutation-drift equilibrium
(Genome-wide deviations from Tajima's D = 0)

$$\theta = \# \text{ pos}$$

$$\theta = 3,000,000 (0.11)$$

Figure S13: Distributions of the Tajima's D estimates across the four focal groups based on all 100-kb sliding windows spanning the genome. Ancient Asian, ancient European, early Asian x European and hybrid tea roses are shown in yellow, blue, dark green and khaki,



Pont*, Leroy* et al. 2019
Nature Genetics

Leroy et al., in revision
Leffler et al. Plos Biol 2012

Current genetic diversity = long-term processes and past history

Can we understand why some species are more genetically diverse than some others?

π = the average number of nucleotide differences per site between pairs of sequences

1:
AAATACCA**A**CAAC
TTCATGTTTGATG

Deviations from mutation-drift equilibrium
(Genomic deviation Tajima's)

How can we (try to) precisely reconstruct the evolutionary history of a given species?

→ Demographic modelling

$$\theta = \frac{1}{\sum_{i=1}^{n-1} \frac{1}{i}} = 3 / 2.083 = 1.44 \text{ (0.11 per bp)}$$

At mutation-drift equilibrium:
Tajima's D~0

alleles
= population contraction

D<0: Excess of rare alleles
= population expansion



Leffler et al. Plos Biol 2012

All (demographic) models are wrong, but can still be informative!

Demographic modeling approaches require openness to the fact that, by definition, a model intentionally simplifies reality!

“...all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind ...” – George Box –

Effective population size: a crucial parameter

Census population size (N_c) : the number of individuals in a population that you can observe



Effective population size: a crucial parameter

Census population size (N_c) : the number of individuals in a population that you can observe

≠

Effective population size (N_e): the number of individuals in a Wright-Fisher model (*i.e.* the size of an idealized population) that would produce the same amount of genetic drift as in the real population



N_e captures the effects of the genetic drift and is therefore a key parameter in population genetics!

Effective population size: a crucial parameter

Effective population size (N_e): the number of individuals in a **Wright-Fisher model** (*i.e.* the size of an idealized population) that would produce the same amount of genetic drift as in the real population

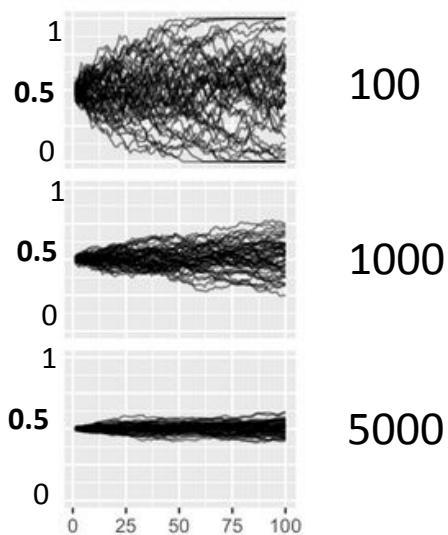


Sewall Wright Ronald A. Fisher



WF model:

- non-overlapping generations
- no selection
- no mutation
- no migration
- random mating

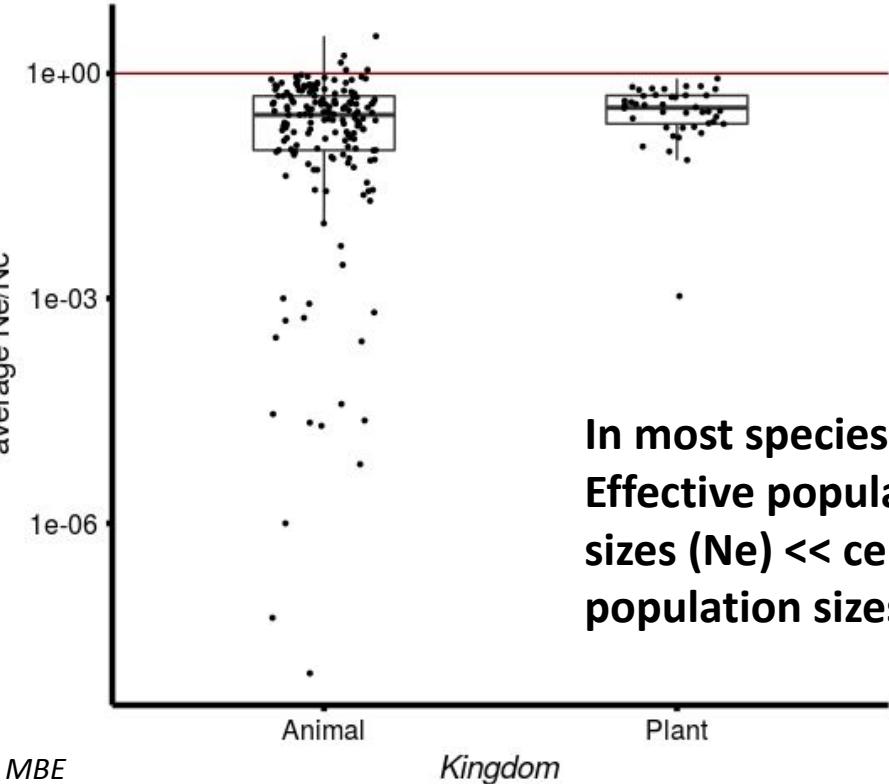
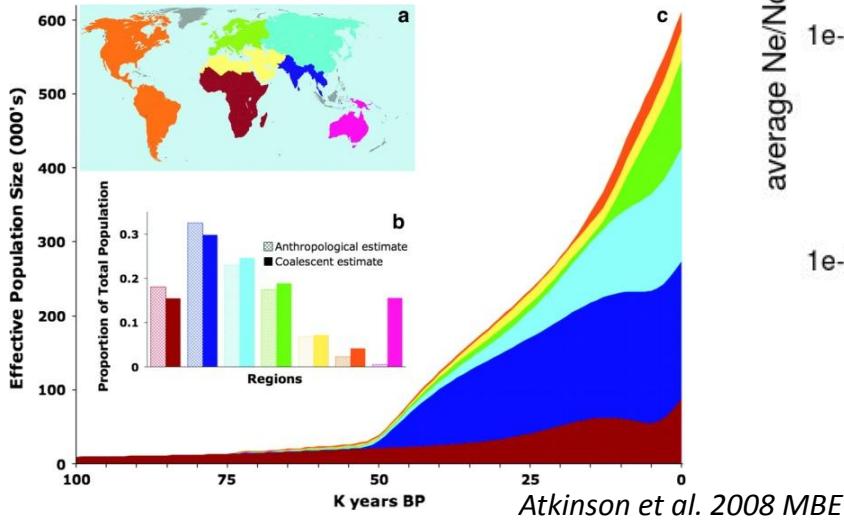


WF: a model for the
allele frequency
dynamics

Effective population size: a crucial parameter



Census size (N_c) $\sim 8.2 * 10^9$ (2025)
Effective size (N_e): $\sim 6.2 * 10^5$?
 $N_e/N_c = 7.6 * 10^{-5}$



Data from Hoban et al. 2020 BiolCons

Introducing the coalescent theory through N_e

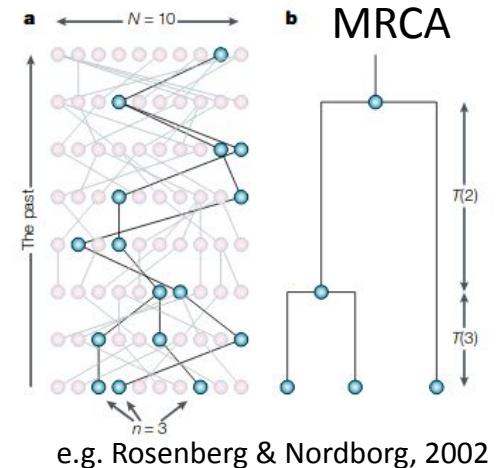
Coalescence: a stochastic process that describes how population genetic processes determine the shape of the genealogy of sampled gene sequences

Introducing the coalescent theory through N_e

Coalescence: a stochastic process that describes how population genetic processes determine the shape of the genealogy of sampled gene sequences

n individuals sampled from a population of:

- Size N (constant & large, well-mixed population)
- New (neutral) mutations
- No selection, no subdivision, no migration

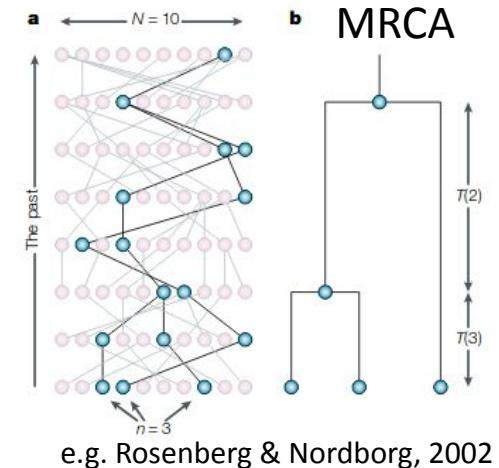
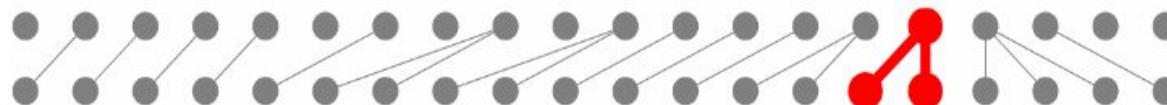


Introducing the coalescent theory through N_e

Coalescence: a stochastic process that describes how population genetic processes determine the shape of the genealogy of sampled gene sequences

n individuals sampled from a population of:

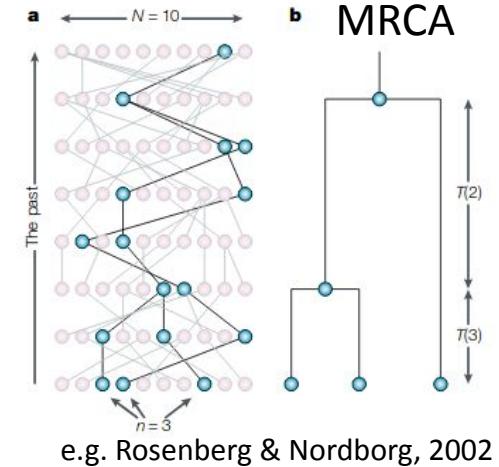
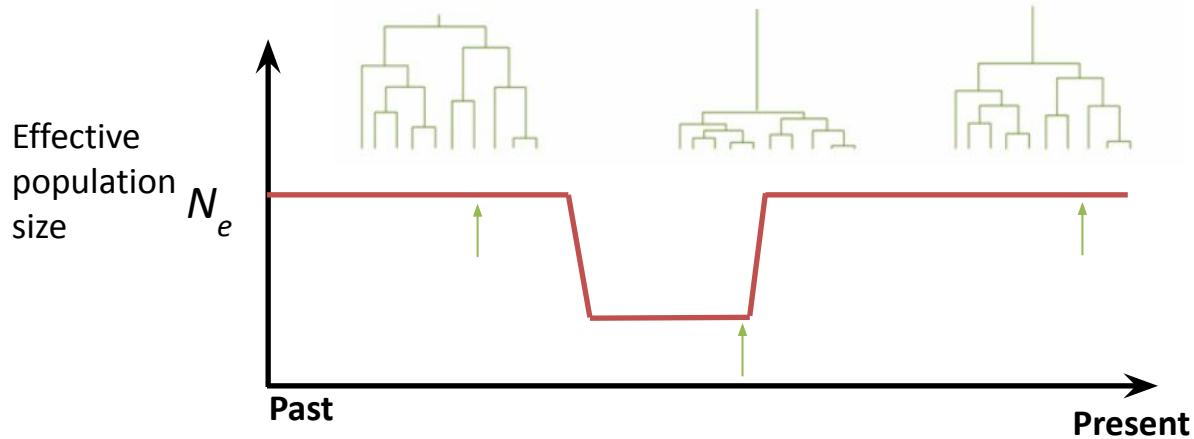
- Size N (constant & large, well-mixed population)
- New (neutral) mutations
- No selection, no subdivision, no migration



The probability of 2 alleles in generation t coalesce in $t-1$ is $\frac{1}{2Ne}$

→ A direct relationship between time and N_e

Introducing the coalescent theory through N_e



*The rates of coalescence are informative about population size because
coalescence events are more likely to occur when the population is small.*

*For example, if we select a few people at random from a small, isolated village,
they are likely to share an ancestor in recent generations.*

Population decline: shorter coalescence times (“shorter branches”)

Population expansion: longer coalescence times (“longer branches”)

Full likelihood inferences and limitations

“The variable population size” coalescent model (Griffiths & Tavaré, 1994;
Donnelly & Tavaré, 1995)

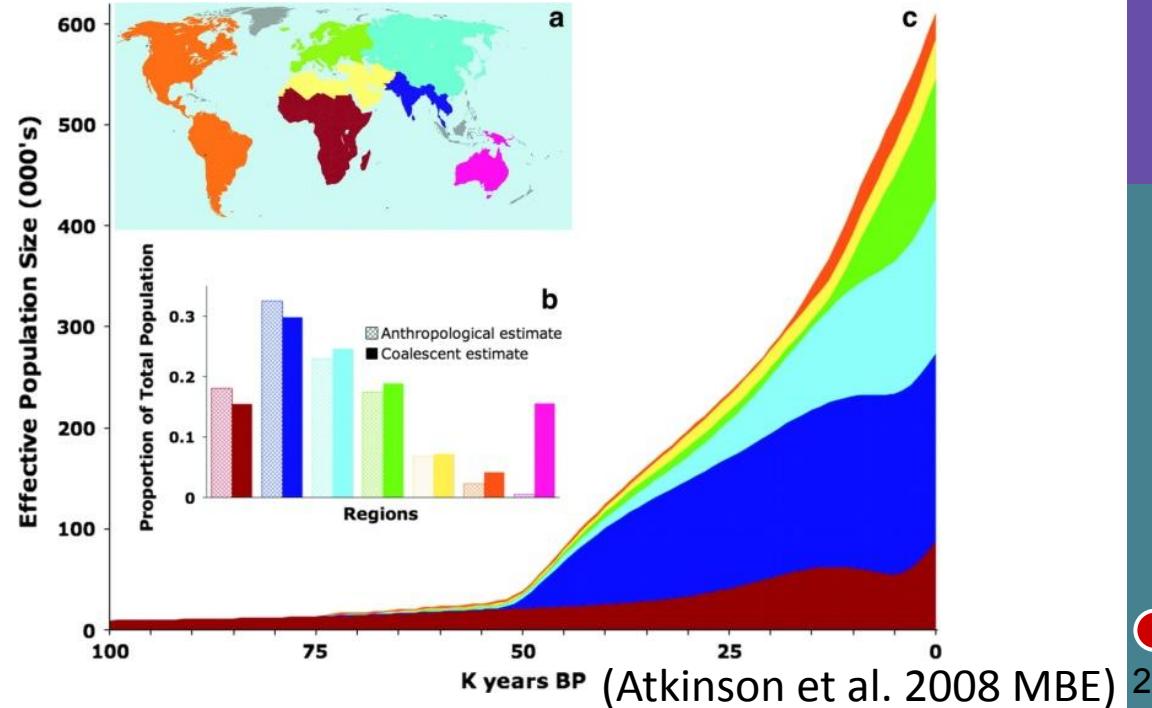
→ Maximum likelihood estimates of parameters (pop expansions/bottlenecks)

Full likelihood computation for one locus, *e.g.* mtDNA

**While sometimes informative,
statistical resolution of inferences
from only one locus (here, the
mtDNA) is generally poor...**

Why? Power diminishes rapidly as we move further back in time, primarily because there are few independent lineages that explore such deep time depths. For example, in humans, mitochondrial DNA provides no information beyond approximately 200,000 years ago, when all humans trace back to a common maternal ancestor.

→ Need more loci,
complete genomes?



Approximations to overcome scalability challenges

Ideally, we would like to estimate the full likelihood of observing all these variants along the genome

- Full likelihood methods are however not applicable to genome-scale datasets (yet) because of two main limitations:
 - 1) Methods do not scale well with the number of loci being analyzed
 - 2) Methods are not well suited for handling recombination (modeling genomic linkage is particularly challenging)

We need to find a way to **approximate** this...

- Approximating the coalescent with recombination

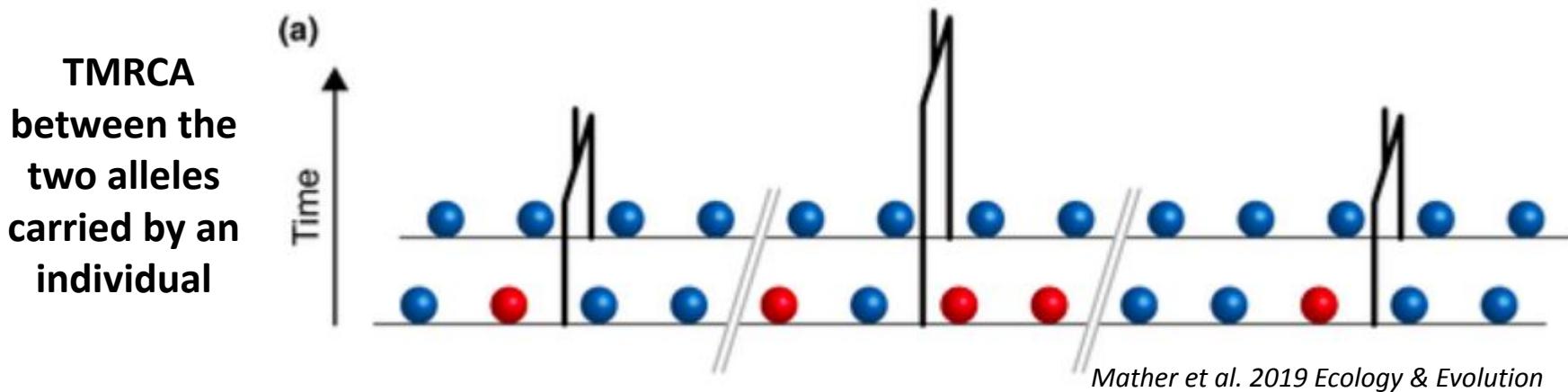
e.g. McVean & Cardin, 2005

SMC (sequential Markov coalescent)

Many demographic inference methods are based on the SMC (or SMC') approximation:
=> PSMC, MSMC, SMC++, ...

Pairwise Sequentially Markovian Coalescent (PSMC)

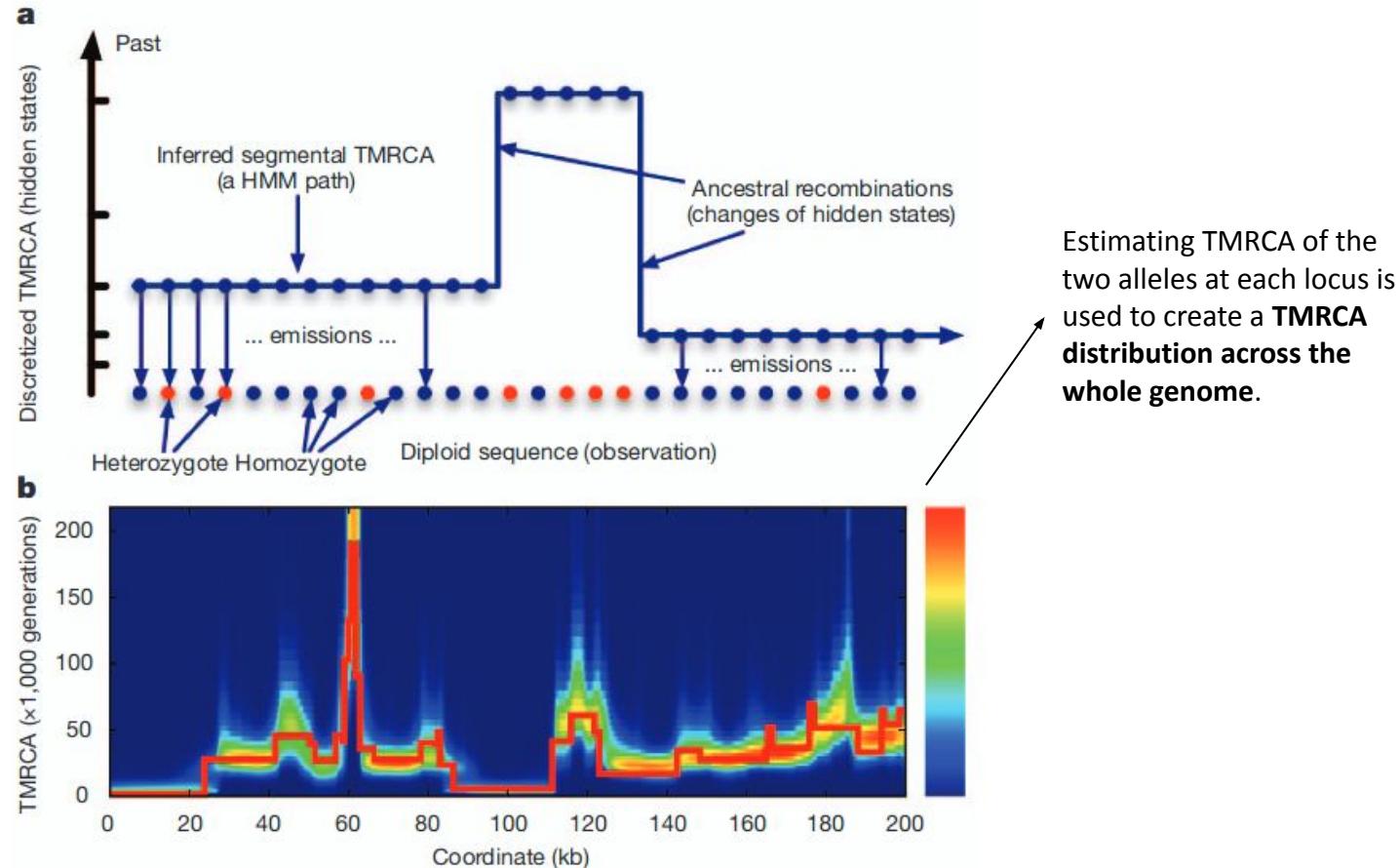
PSMC (Li & Durbin, 2011): local time to the most recent common ancestor (TMRCA) on the basis of the local density of heterozygotes in short genomic blocks



Despite being a remarkably simple likelihood model for analyzing the pattern of genetic mutations in a single diploid individual, a decade of empirical applications has demonstrated the unexpectedly high power of PSMC to infer the population/species history of that individual.

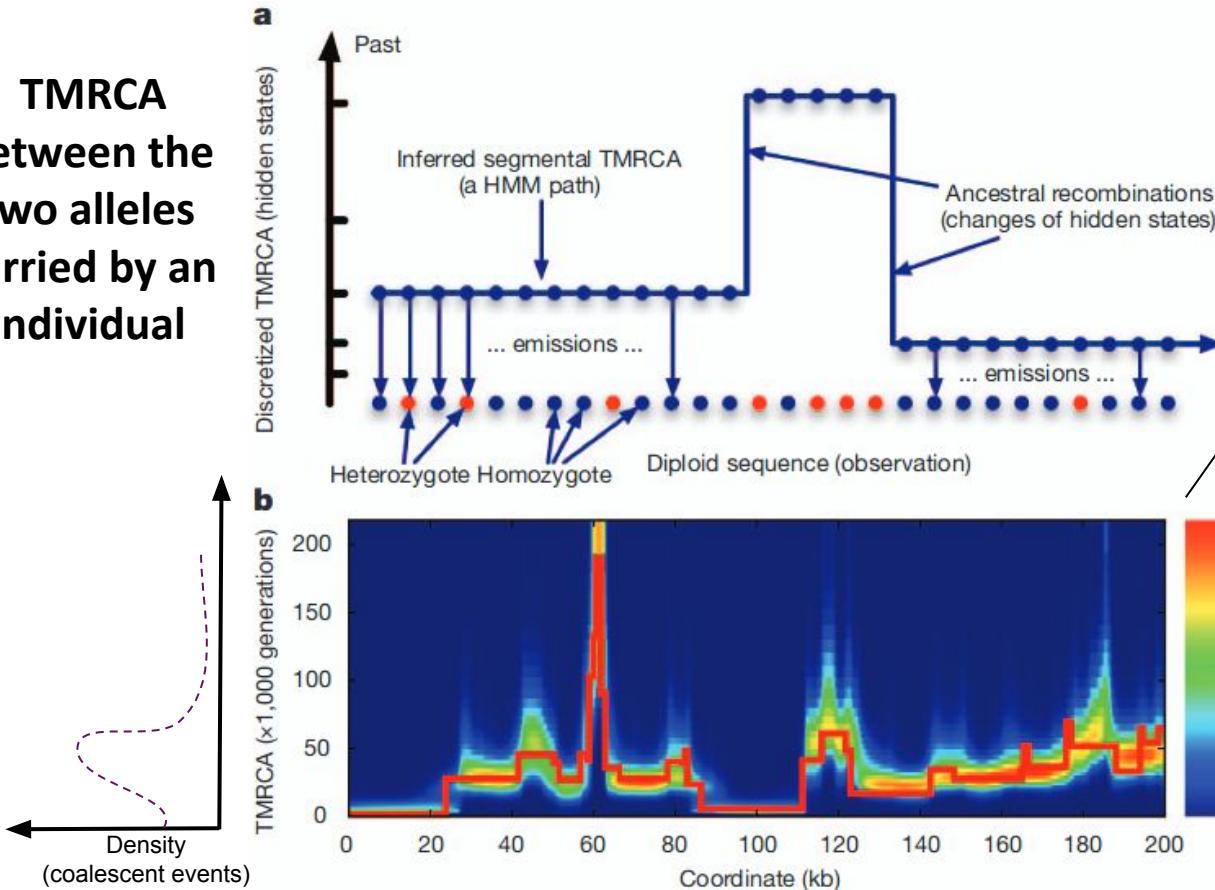
Pairwise Sequentially Markovian Coalescent (PSMC)

TMRCA
between the
two alleles
carried by an
individual



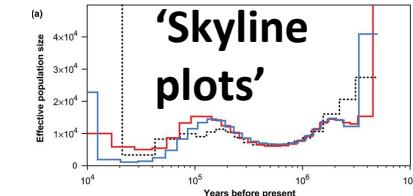
Pairwise Sequentially Markovian Coalescent (PSMC)

TMRCA
between the
two alleles
carried by an
individual



Estimating TMRCA of the two alleles at each locus is used to create a **TMRCA distribution across the whole genome**.

Since the rate of coalescent events is inversely proportional to N_e , PSMC identifies periods of N_e changes. For example, **when many loci coalesce at the same time, it is (assumed to be) a sign of small N_e at that time**.



Pairwise Sequentially Markovian Coalescent (PSMC)

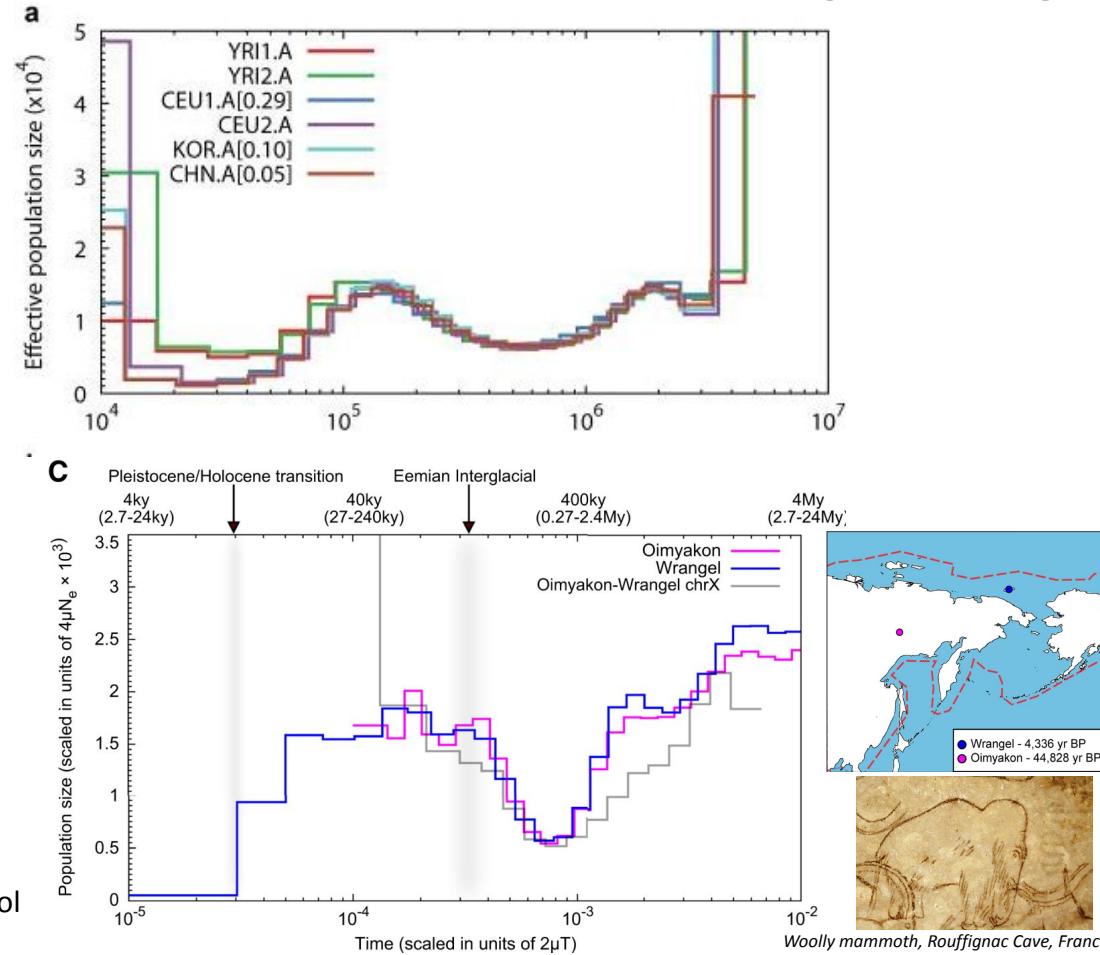
Example of applications:



(e.g. Li & Durbin, 2011 Nature)

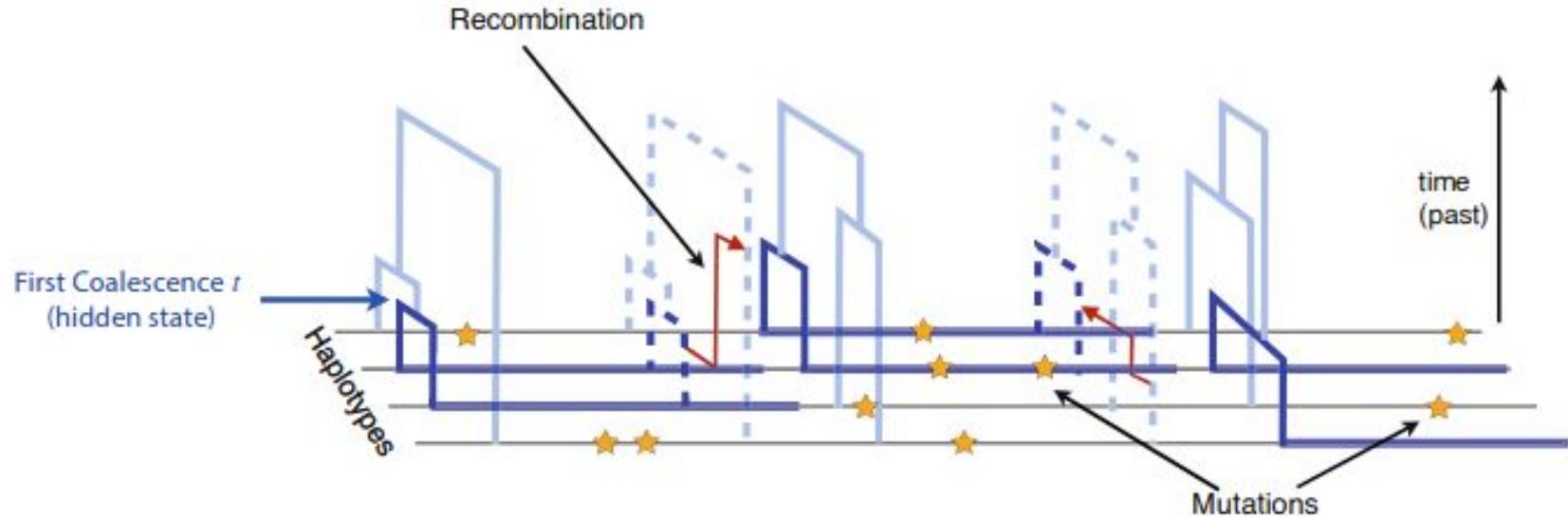


Palkopoulou et al. 2015, Current Biol



Multiple Sequentially Markovian Coalescent (MSMC)

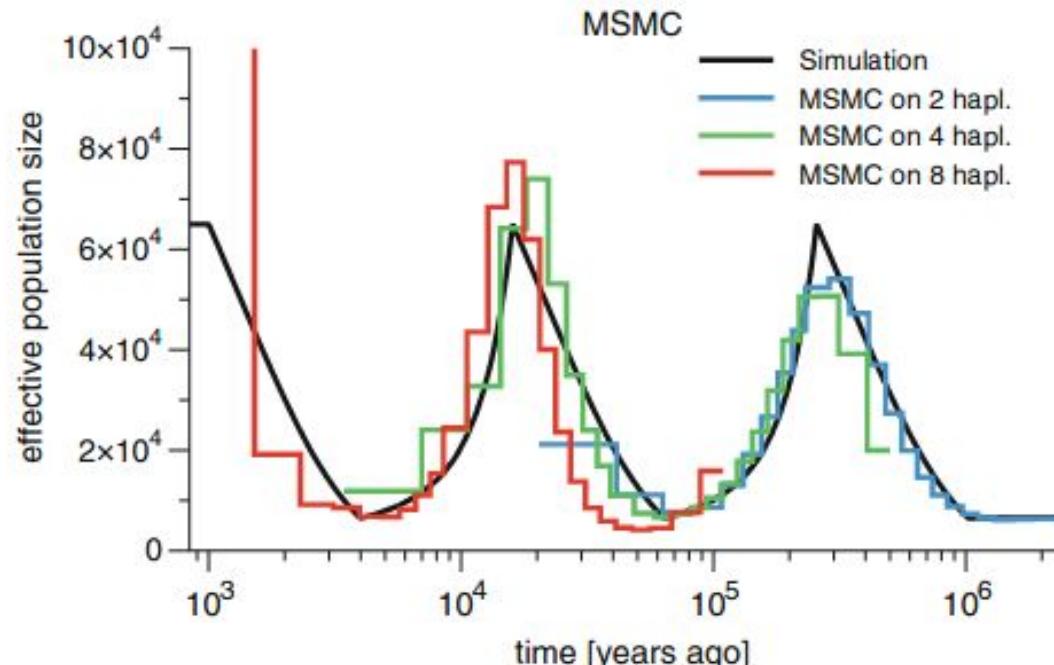
MSMC has better power than PSMC to resolve recent changes in effective population size, because of its use of >2 sequences. Adding alleles increase the probability for two of the copies to coalesce in the recent past.



Multiple Sequentially Markovian Coalescent (MSMC)

MSMC has better power than PSMC to resolve recent changes in effective population size, because of its use of >2 sequences. Adding alleles increase the probability for two of the copies to coalesce in the recent past.

**More recent estimates
with more individuals**
(because of more recent
first events of coalescence)



SMC++ development to overcome MSMC limitations

A main issue with MSMC is that this method requires phased genomes (or at least with unphased data the MSMC estimation accuracy is low)

Genotypes

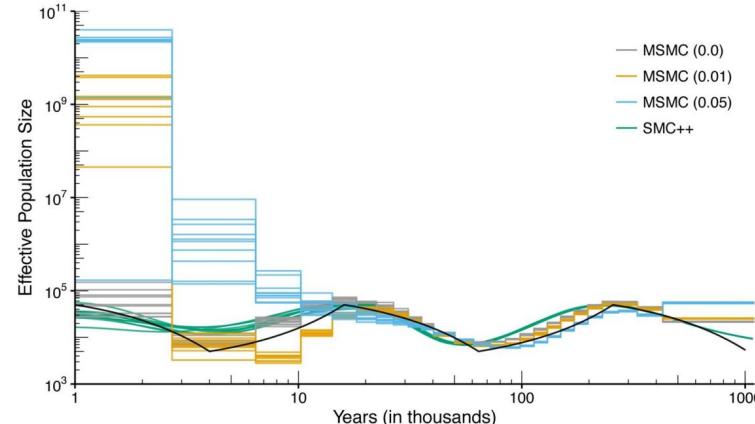
A	T	C	A	G
G	G	G	G	

vs.

Haplotypes

>all1
ATGCGG
>all2
AGGGAG

Computational haplotype phasing (*i.e.* identify the alleles that are co-located on the same chromosome) represents a hard task to achieve...



Some other methods using unphased data are becoming popular to overcome this problem (*e.g.* SMC++, PopSizeABC, ...) but requires tens of whole-genome sequences.

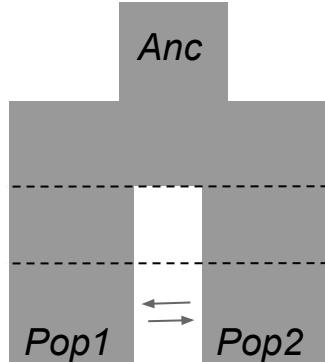
PSMC-like methods: pros and cons

Advantages:

- Rapid, simple, extremely popular
- Only one individual needed (for PSMC at least)
=> 'Genome papers' + Phylogenomic-oriented papers+ aDNA

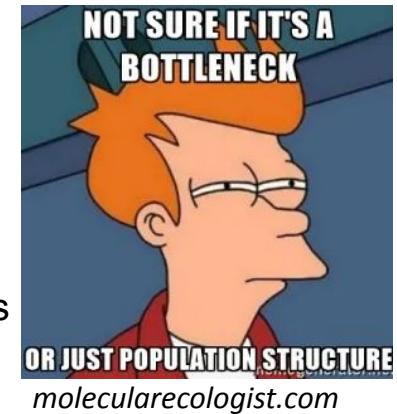
Limitations (1/2):

- Simplistic approach (assumes a panmictic population, *i.e.* drift-only)
=> change in N_e in a PSMC plot can be actually caused by other factors
e.g. population structure



This past history induced population structure at many loci

Contemporary population structure associated with this past history will be artificially considered as a period of low N_e



PSMC-like methods: pros and cons

Limitations (2/2):

- PSMC estimates for recent times (<10kyrs) are rarely accurate (more sequences needed to use the other PSMC-like methods (MSMC, SMC++, ...)!)
- Sensitive to the quality of the genome assembly and sequencing data (coverage)
Nadachowska-Brzyska et al. (2016) suggested filters :
e.g. A mean coverage of at least 18X, <25% of missing data, ...
- Doesn't recover sudden changes in N_e or very ancient changes ($>5\text{-}10 N_e$ generations)
- Problem of rescaling coalescent times to real time for non-model species
(incorrect mutation rates or knowledge about generation times)

Site Frequency Spectrum (SFS)-based approaches

Full-likelihood methods = not adapted to WGS because of big data & recombination events

One way to circumvent this problem is to use summary statistics to describe the dataset
e.g. compute the likelihood only based on the SFS (*i.e.* a “composite” likelihood)

Ind1-A1: A..C..G..A..T..G..A..A..T..G

Ind1-A2: A..G..G..A..T..G..T..A..C..G

Ind2-A1: T..G..G..T..T..C..T..A..T..G

Ind2-A2: A..G..G..A..G..C..T..A..C..G

Ind3-A1: A..G..A..A..T..G..T..A..T..G

Ind3-A2: A..C..G..A..T..C..T..G..T..A

Ind4-A1: A..G..G..T..G..G..T..A..T..G

Ind4-A2: A..G..G..T..T..C..T..A..T..G

Site Frequency Spectrum (SFS)-based approaches

Full-likelihood methods = not adapted to WGS because of big data & recombination events

One way to circumvent this problem is to use summary statistics to describe the dataset
e.g. compute the likelihood only based on the SFS (*i.e.* a “composite” likelihood)

Ind1-A1: A..C..G..A..T..G..A..A..T..G

Ind1-A2: A..G..G..A..T..G..T..A..C..G

Ind2-A1: T..G..G..T..T..C..T..A..T..G

Ind2-A2: A..G..G..A..G..C..T..A..C..G

Ind3-A1: A..G..A..A..T..G..T..A..T..G

Ind3-A2: A..C..G..A..T..C..T..G..T..A

Ind4-A1: A..G..G..T..G..G..T..A..T..G

Ind4-A2: A..G..G..T..T..C..T..A..T..G

Site Frequency Spectrum (SFS)-based approaches

Full-likelihood methods = not adapted to WGS because of big data & recombination events

One way to circumvent this problem is to use summary statistics to describe the dataset
e.g. compute the likelihood only based on the SFS (*i.e.* a “composite” likelihood)

Ind1-A1: A..**C**..G..A..T..**G**..**A**..A..T..G

Ind1-A2: A..G..G..A..T..**G**..T..A..**C**..G

Ind2-A1: **T**..G..G..**T**..T..C..T..A..T..G

Ind2-A2: A..G..G..A..**G**..C..T..A..**C**..G

Ind3-A1: A..G..**A**..A..T..**G**..T..A..T..G

Ind3-A2: A..**C**..G..A..T..C..T..**G**..T..**A**

Ind4-A1: A..G..G..**T**..**G**..**G**..T..A..T..G

Ind4-A2: A..G..G..**T**..T..C..T..A..T..G

Minor allele	1	2	1	3	2	4	1	1	2	1
frequency	/	/	/	/	/	/	/	/	/	/
(MAF)	8	8	8	8	8	8	8	8	8	8

Site Frequency Spectrum (SFS)-based approaches

Full-likelihood methods = not adapted to WGS because of big data & recombination events

One way to circumvent this problem is to use summary statistics to describe the dataset
e.g. compute the likelihood only based on the SFS (*i.e.* a “composite” likelihood)

Ind1-A1: A..C..G..A..T..G..A..A..T..G

Ind1-A2: A..G..G..A..T..G..T..A..C..G

Ind2-A1: T..G..G..T..T..C..T..A..T..G

Ind2-A2: A..G..G..A..G..C..T..A..C..G

Ind3-A1: A..G..A..A..T..G..T..A..T..G

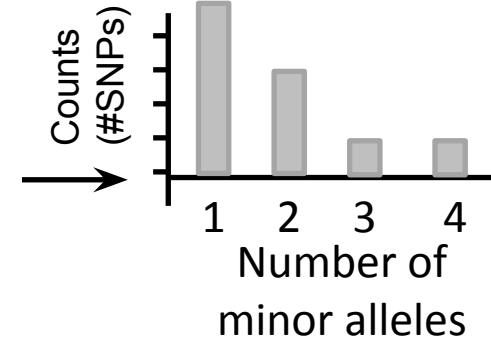
Ind3-A2: A..C..G..A..T..C..T..G..T..A

Ind4-A1: A..G..G..T..G..G..T..A..T..G

Ind4-A2: A..G..G..T..T..C..T..A..T..G

Minor allele frequency (MAF)	1	2	1	3	2	4	1	1	2	1
	/	/	/	/	/	/	/	/	/	/
8	8	8	8	8	8	8	8	8	8	8

« Folded 1D-Site Frequency Spectrum » (folded 1D-SFS)



Site Frequency Spectrum (SFS)-based approaches

Full-likelihood methods = not adapted to WGS because of big data & recombination events

One way to circumvent this problem is to use summary statistics to describe the dataset

e.g. compute the likelihood only based on the SFS (*i.e.* a “composite” likelihood)

Anc-A1: A..G..G..T..T..C..A..A..C..G

Anc-A2: A..G..G..T..T..C..A..A..C..G

Ind1-A1: A..C..G..A..T..G..A..A..T..G

Ind1-A2: A..G..G..A..T..G..T..A..C..G

Ind2-A1: T..G..G..T..T..C..T..A..T..G

Ind2-A2: A..G..G..A..G..C..T..A..C..G

Ind3-A1: A..G..A..A..T..G..T..A..T..G

Ind3-A2: A..C..G..A..T..C..T..G..T..A

Ind4-A1: A..G..G..T..G..G..T..A..T..G

Ind4-A2: A..G..G..T..T..C..T..A..T..G

Site Frequency Spectrum (SFS)-based approaches

Full-likelihood methods = not adapted to WGS because of big data & recombination events

One way to circumvent this problem is to use summary statistics to describe the dataset
e.g. compute the likelihood only based on the SFS (*i.e.* a “composite” likelihood)

Anc-A1: A..G..G..T..T..C..A..A..C..G

Anc-A2: A..G..G..T..T..C..A..A..C..G

Ind1-A1: A..C..G..A..T..G..A..A..T..G

Ind1-A2: A..G..G..A..T..G..T..A..C..G

Ind2-A1: T..G..G..T..T..C..T..A..T..G

Ind2-A2: A..G..G..A..G..C..T..A..C..G

Ind3-A1: A..G..A..A..T..G..T..A..T..G

Ind3-A2: A..C..G..A..T..C..T..G..T..A

Ind4-A1: A..G..G..T..G..G..T..A..T..G

Ind4-A2: A..G..G..T..T..C..T..A..T..G

Derived allele frequency (DAF)	1 /	2 /	1 /	5 /	2 /	4 /	7 /	1 /	6 /	1 /
	8	8	8	8	8	8	8	8	8	8

Site Frequency Spectrum (SFS)-based approaches

Full-likelihood methods = not adapted to WGS because of big data & recombination events

One way to circumvent this problem is to use summary statistics to describe the dataset
e.g. compute the likelihood only based on the SFS (*i.e.* a “composite” likelihood)

Anc-A1: A..G..G..T..T..C..A..A..C..G

Anc-A2: A..G..G..T..T..C..A..A..C..G

Ind1-A1: A..C..G..A..T..G..A..A..T..G

Ind1-A2: A..G..G..A..T..G..T..A..C..G

Ind2-A1: T..G..G..T..T..C..T..A..T..G «Unfolded 1D-Site Frequency

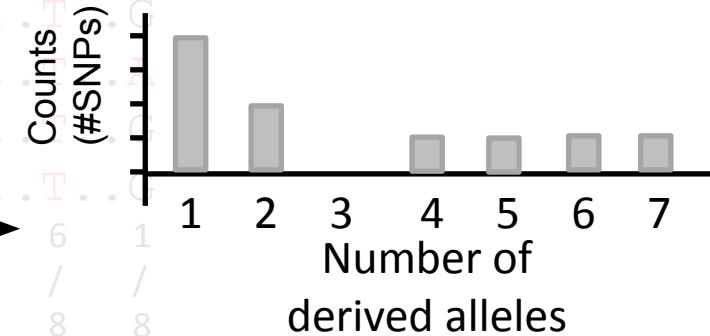
Ind2-A2: A..G..G..A..G..C..T..A..C..G Spectrum » (unfolded 1D-SFS)

Ind3-A1: A..G..A..A..T..G..T..A..T..G

Ind3-A2: A..C..G..A..T..C..T..G..T..G

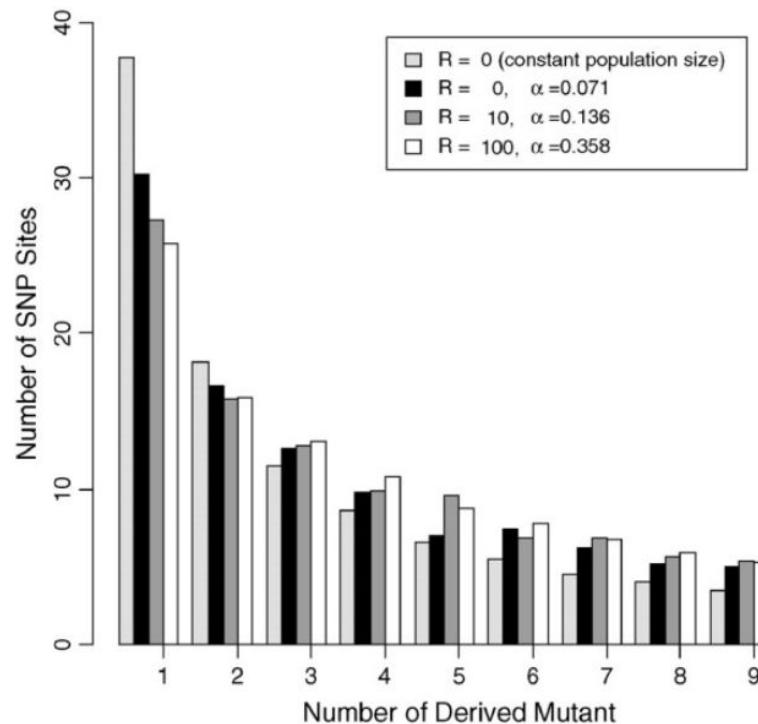
Ind4-A1: A..G..G..T..G..G..T..A..T..G

Ind4-A2: A..G..G..T..T..C..T..A..T..G



Derived allele frequency (DAF)	1 / 8	2 / 8	1 / 8	5 / 8	2 / 8	4 / 8	7 / 8	→	6 / 8	1 / 8	1 / 8	1 / 8	1 / 8	1 / 8	1 / 8
--------------------------------	-------	-------	-------	-------	-------	-------	-------	---	-------	-------	-------	-------	-------	-------	-------

Site Frequency Spectrum (SFS)-based approaches



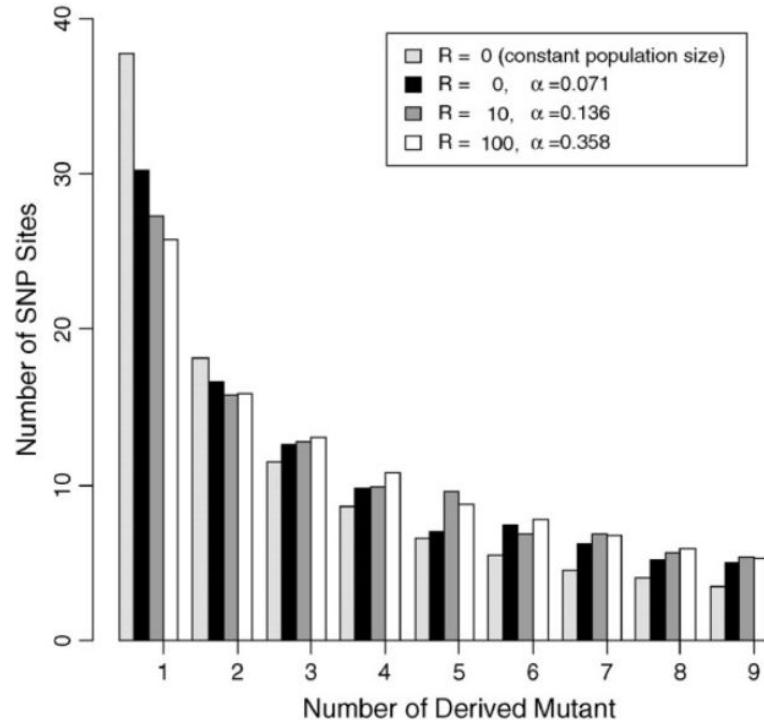
Sudden population contraction (bottleneck)
-> deficit of rare alleles

//Pop expansion -> excess

→ SFS are therefore informative about N_e changes

Zhu & Bustamante, 2005

Site Frequency Spectrum (SFS)-based approaches



Sudden population contraction (bottleneck)
-> deficit of rare alleles

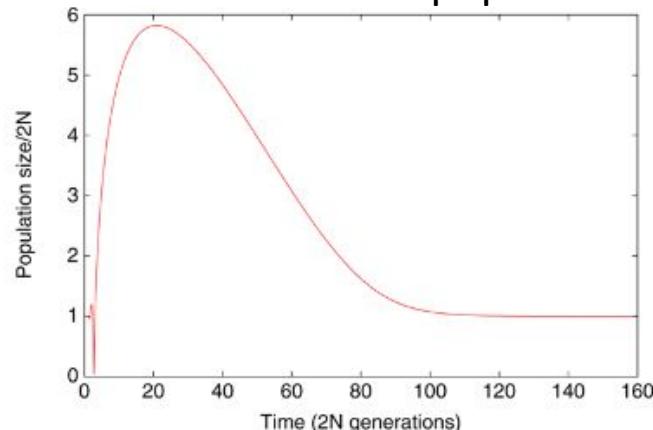
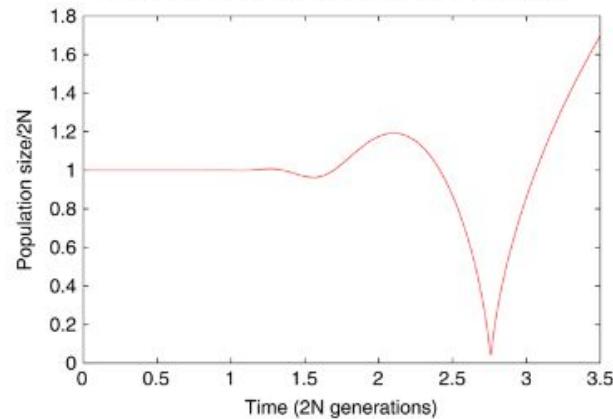
//Pop expansion -> excess

→ SFS are therefore informative about N_e changes

Site Frequency Spectrum (SFS)-based approaches

However, a single-dimensional SFS provides only limited information !

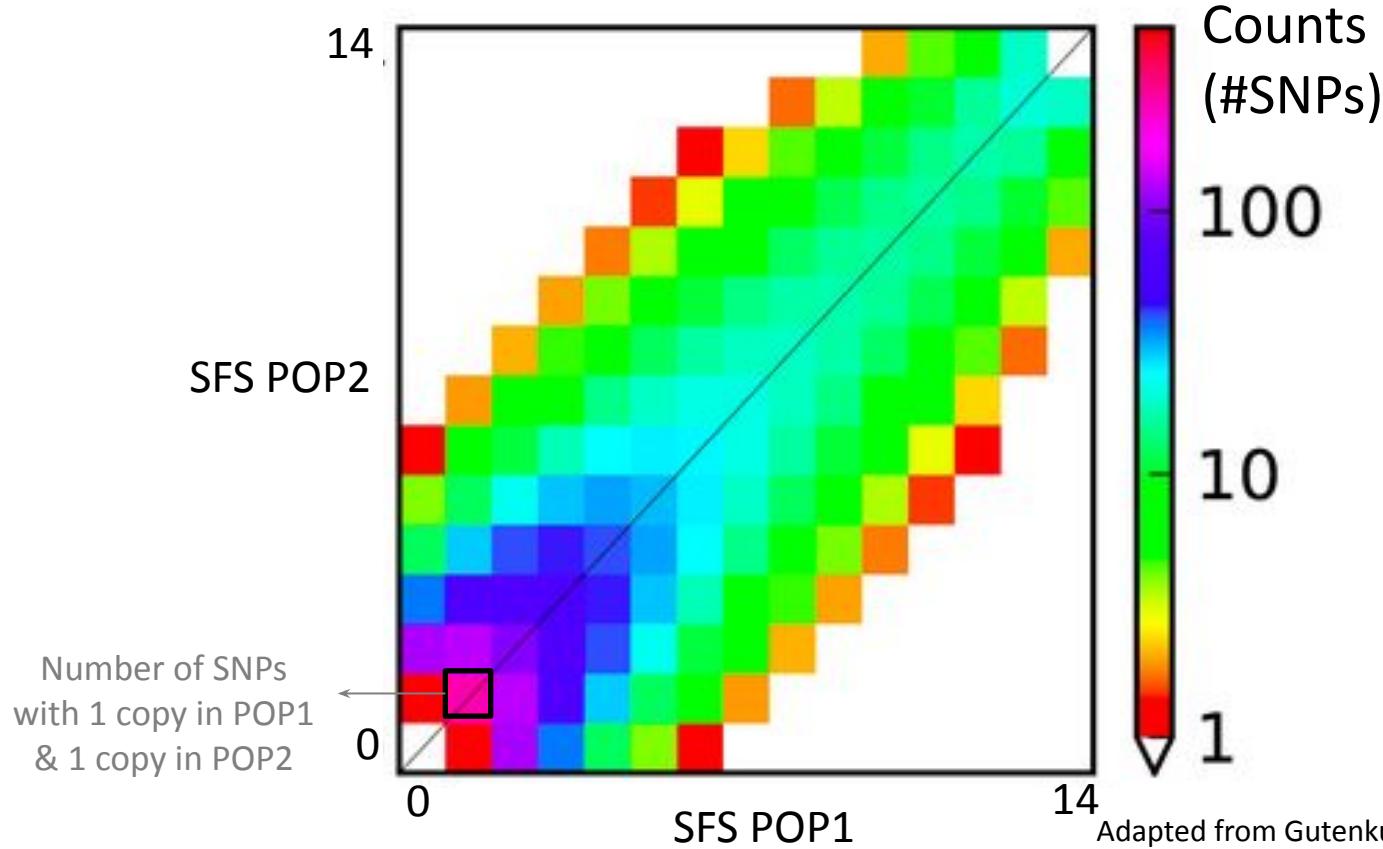
Here, two demographic histories with the same spectrum as a constant size populations!



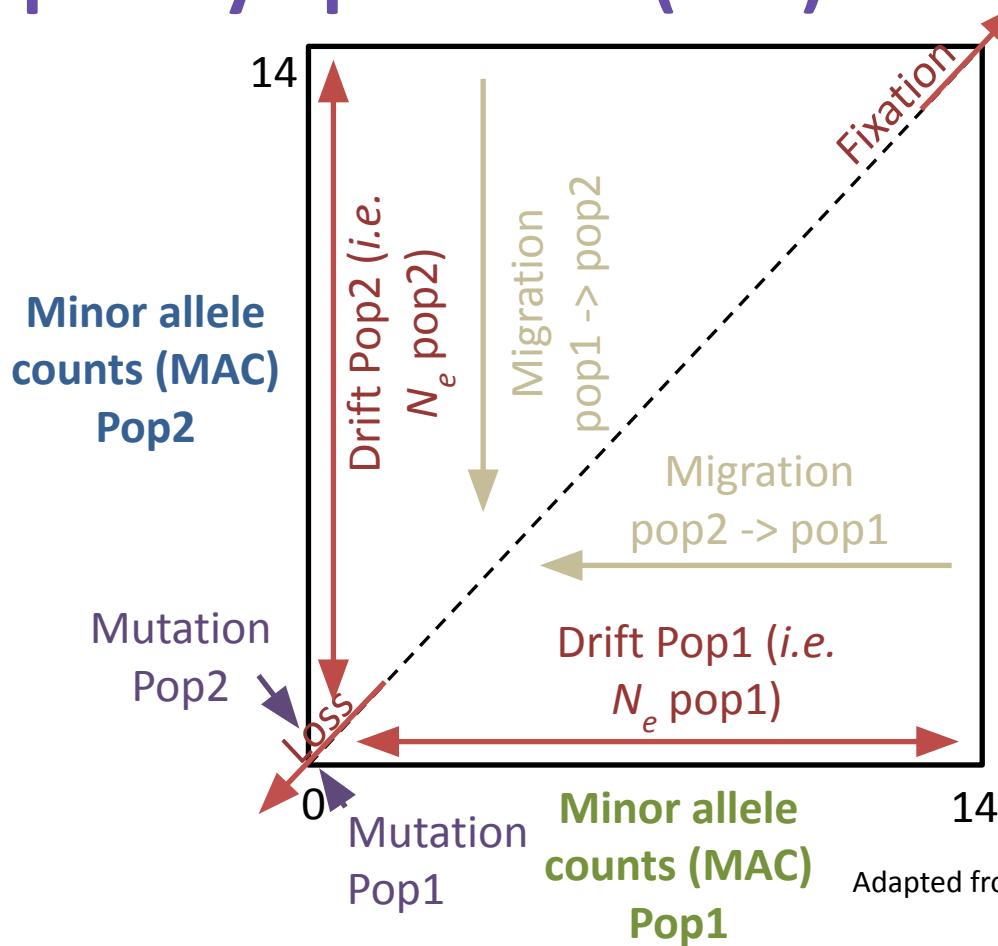
Myers et al. 2008 “Can one learn history from the allelic spectrum?”

-> 2-dimensional site frequency spectrum (2D-SFS)
(use of polymorphism data from 2 populations)

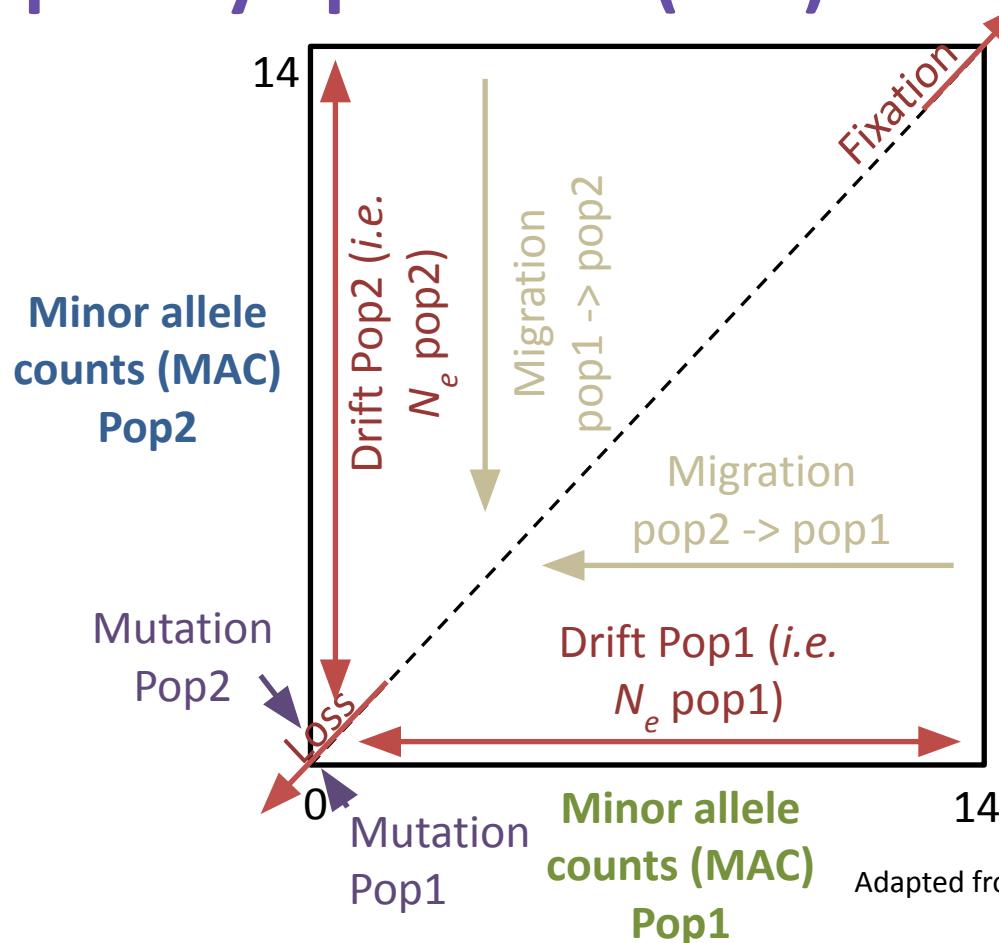
Site Frequency Spectrum (SFS)-based approaches



Site Frequency Spectrum (SFS)-based approaches



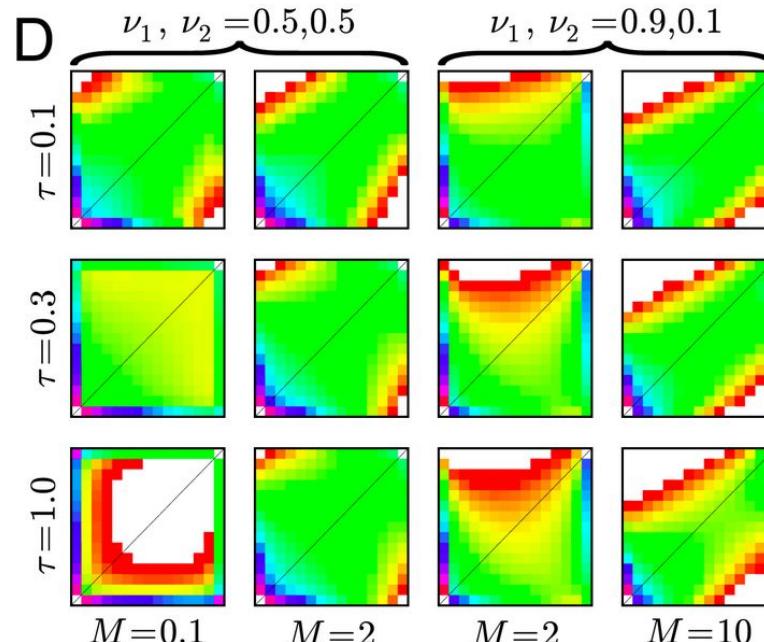
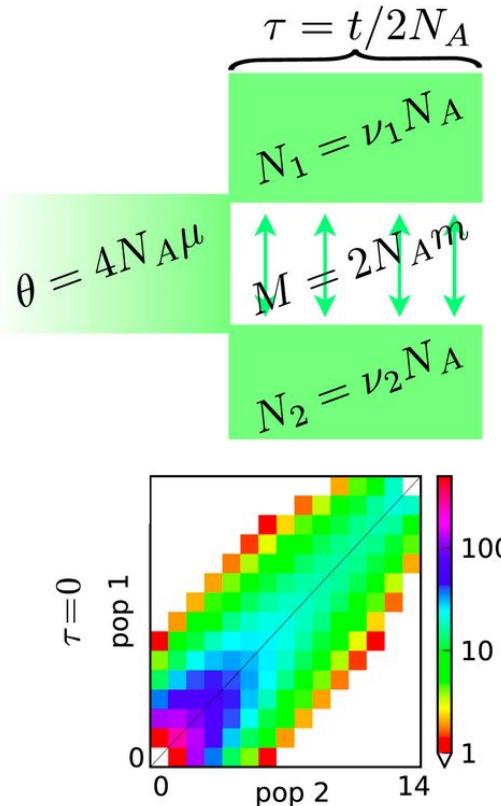
Site Frequency Spectrum (SFS)-based approaches



Recombination is
not taken into
account!
-> All sites are
assumed to be
independent

Adapted from Gutenkunst et al. 2009 Plos Genet

Composite likelihood approach : $\partial a \partial i$

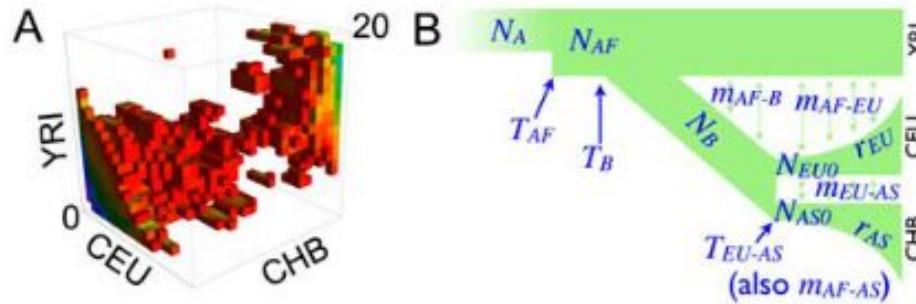


Composite likelihood approach : $\partial\text{a}\partial\text{i}$

3-dimensional SFS

The implementation ($\partial\text{a}\partial\text{i}$ program) is quite flexible. It was initially able to handle up to three simultaneous populations

Data: 1000 genomes project (human)



CEU: US with Northern or Western European Ancestry (EUR)

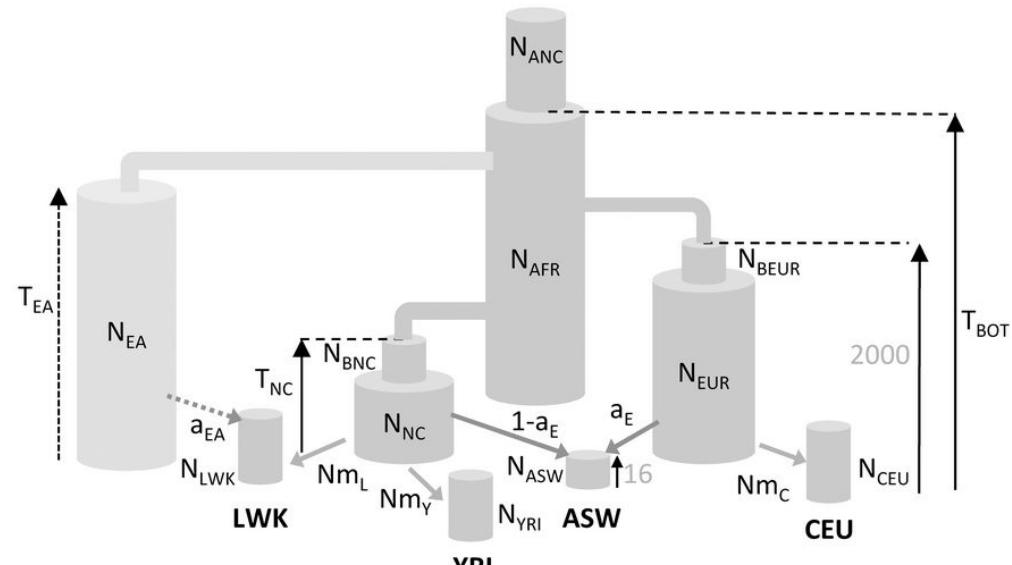
CHB: Han Chinese, Beijing, China (EAS)

Improvements to extend the $\partial\text{a}\partial\text{i}$ strategy to 4-populations (MULTIPOP, Lukic & Hey 2012 Genetics) & 5-populations (*Moments*, Jouganous et al. 2017 Genetics)

Composite likelihood approach : fastsimcoal

fastsimcoal2 is another very popular tool

Using a multiple pairwise joint SFS strategy, fastsimcoal2 is (in theory) be able to infer demography of an arbitrary number of populations



Excoffier *et al.* 2013 Plos Genetics

LWK : Luhya, Kenya (AFR)

YRI: Yoruba, Nigeria (AFR)

ASW: African Ancestry in SW USA (AFR)

CEU: Northern & Western European

Ancestry USA (EUR)

Composite likelihood approach : pros & cons

Advantages:

Computationally efficient :

- accuracy of the inferences increase with the number of SNPs, without increasing the computational load
- Several order of magnitude faster than ABC (even more for full-likelihood methods)
- Can be used to infer complex scenarios

Limitations:

Computational issues

- Convergence problems are possible (computation of the likelihood)

Biological problems

- All sites are assumed to be independent
- Assume that the 2D-SFS is correct (can be an issue if only few individuals were sequenced or in case of low coverage data)
- Risk of not including the true model (as for any other model-based approaches!)
- Correct parameter estimates are challenging
- Limitations on how informative allelic spectra can be

An intro to approximate Bayesian computation...

Likelihood-free demographic inferences

- No need for an explicit likelihood function
- No convergence issues associated with the computation of the likelihood
- High flexibility in model complexity

...

An intro to approximate Bayesian computation...

The rationale:

1/ Observed dataset

Ind1-A1: ATCCACATGCA...

Ind1-A2: ATCGACATGCA...

Ind2-A1: TTTCGACATGCT...

Ind2-A2: ATCGACATGCA...

Ind3-A1: ATCGACATACA...

Ind3-A2: ATCCACATGCA...

Ind4-A1: ATCGACATGCT...

Ind4-A2: ATCGACATGCT...



Summary statistics (e.g. mean number of alleles, F_{ST} between pairs of populations, nucleotide diversity, Tajima's D, SFS, etc...)



The choice and the number of summary statistics to use for the ABC analysis are crucial (e.g. Beaumont et al 2002)

An intro to approximate Bayesian computation...

The rationale:

1/ Observed dataset

Ind1-A1: ATCCACATGCA...

Ind1-A2: ATCGACATGCA...

Ind2-A1: TTCGACATGCT...

Ind2-A2: ATCGACATGCA...



Summary statistics



The choice and the number of summary statistics to use for the ABC analysis are crucial (e.g. Beaumont et al 2002)

2/ Demographic models & simulations

A set of candidate models are hypothesized and simulations are performed using a coalescent sampler (e.g. ms)



Same summary statistics are computed for all simulated datasets

3/ Model Choice

Euclidian distance between summary statistics from the observed dataset and simulations



Identify the « best model »: simulations with the closest summary statistics

4/ Estimation of parameters under the best model



An intro to approximate Bayesian computation...

The rationale:

Real Data

100 loci x 1kb

Summary statistics of pop genomics (or SFS) e.g. Fst, Tajima's D...

Simulations

Model 1

(e.g. 1 million multilocus simulations,
i.e. 1 million with 100 loci x 1kb)

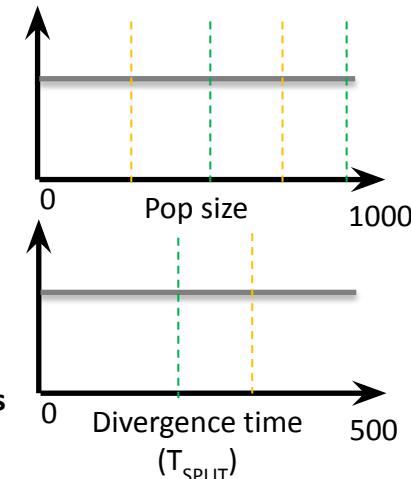
For each simulation, we repeatedly sample a parameter value from **prior** distribution

e.g.
POP SIZE1: uniform[0-1000]
POP SIZE2: uniform[0-1000]
TSPLIT : uniform[0-500]

e.g.
Simul1:
PopSize1=763
PopSize2=261
Tsplit = 330
Simul2:
PopSize1=493
PopSize2=921
Tsplit = 234
... x i simulations

Same summary statistics than for the real data

Model 2



Same summary statistics

ABC in practice (very simplified)...

1/ Observed dataset

Stat 1:
e.g. Mean
Fst



=>observed
dataset

Stat 2

e.g. Tajima's D Pop1

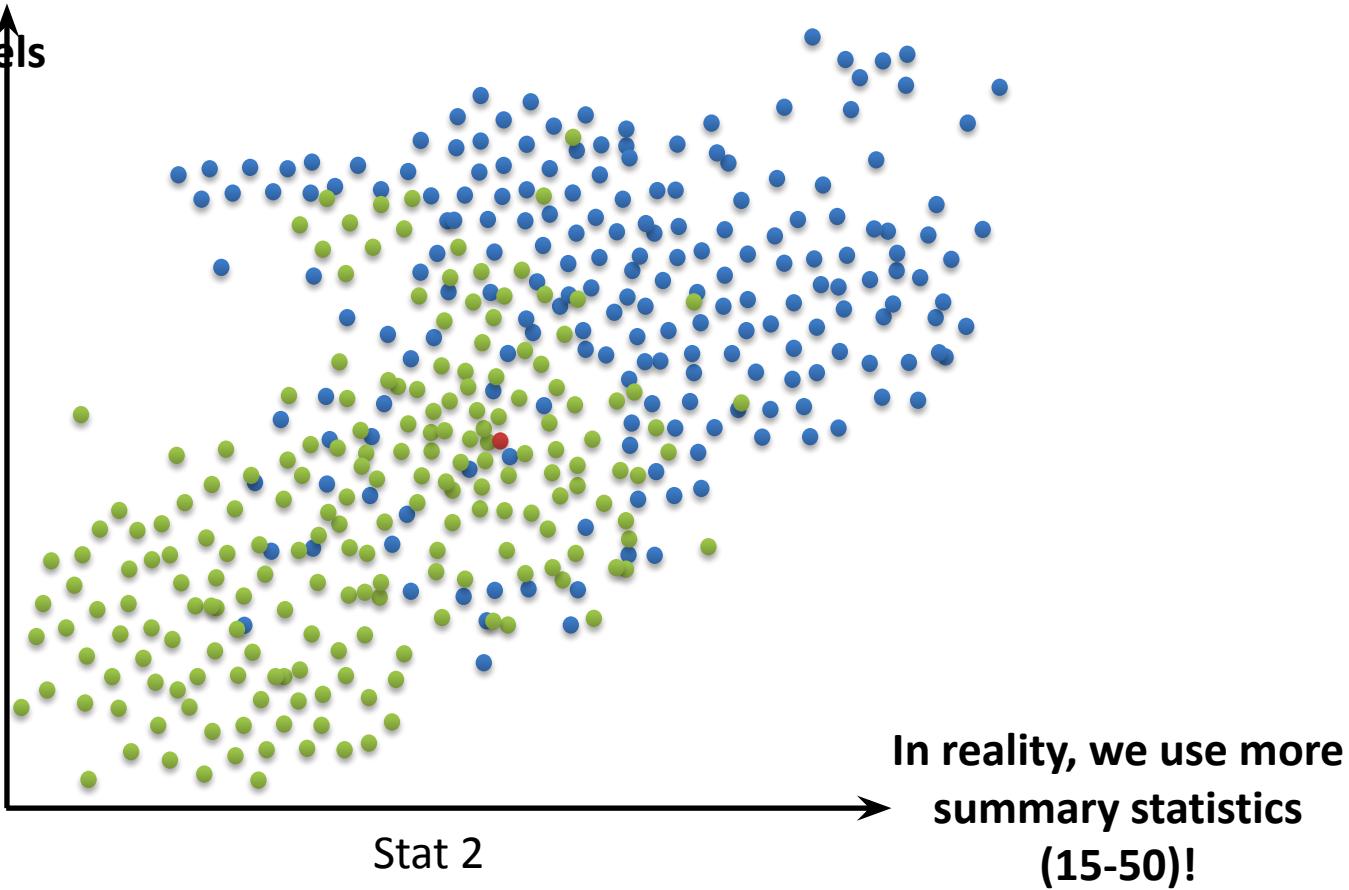
In reality, we use more
summary statistics
(15-50)!

ABC in practice (very simplified)...

2/ Demographic models
& simulations

Stat 1:
e.g. Mean
 F_{ST}

Stat 2
e.g. Tajima's D Pop1



ABC in practice (very simplified)...

3/ Model Choice

Stat 1:
e.g. Mean
Fst

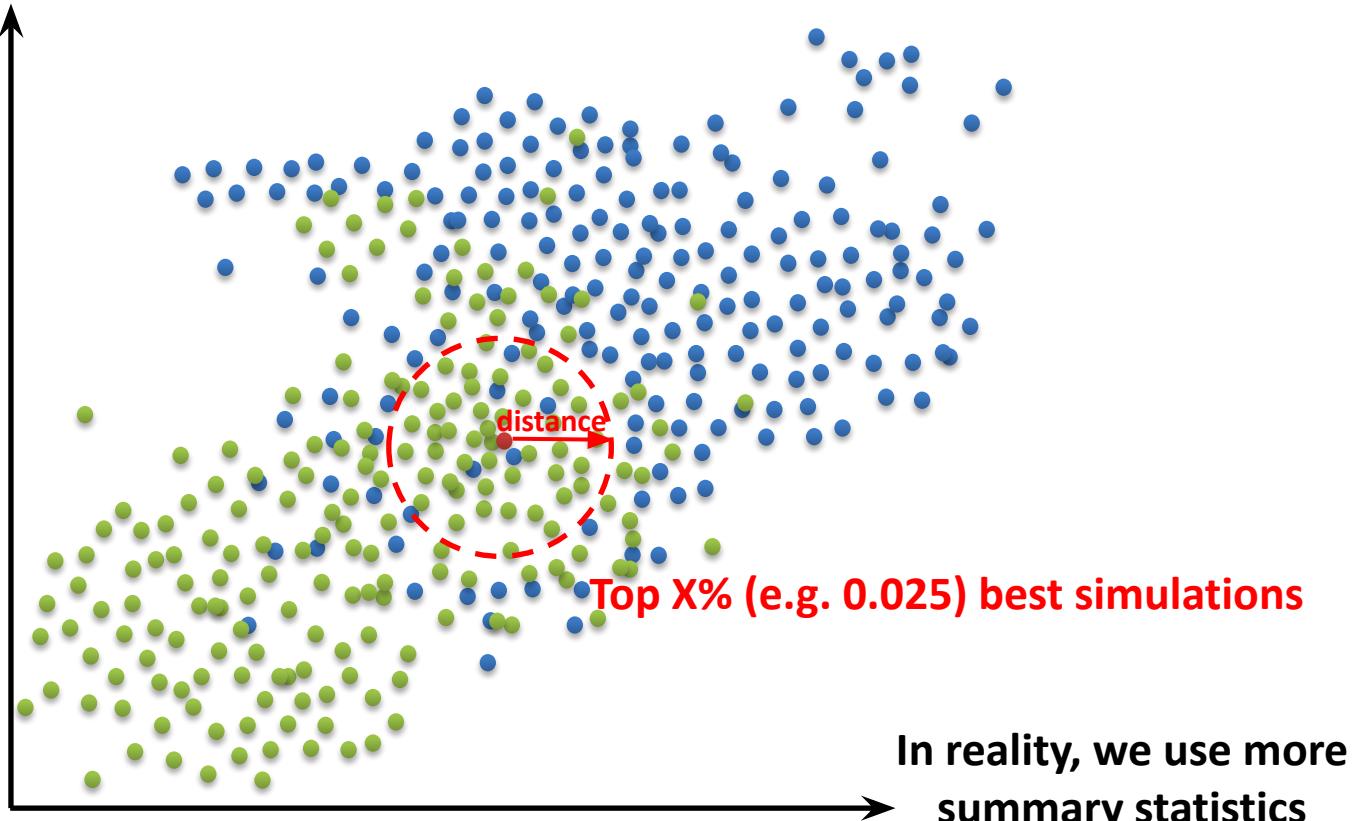
e.g. Tajima's D Pop1

Stat 2

Top X% (e.g. 0.025) best simulations

In reality, we use more
summary statistics
(15-50)!

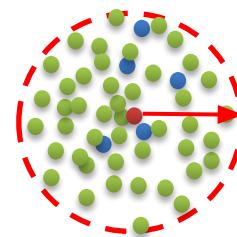
distance



ABC in practice (very simplified)...

3/ Model Choice

Stat 1:
e.g. Mean
Fst



Top X% (e.g. 0.025) best simulations

e.g. Tajima's D Pop1

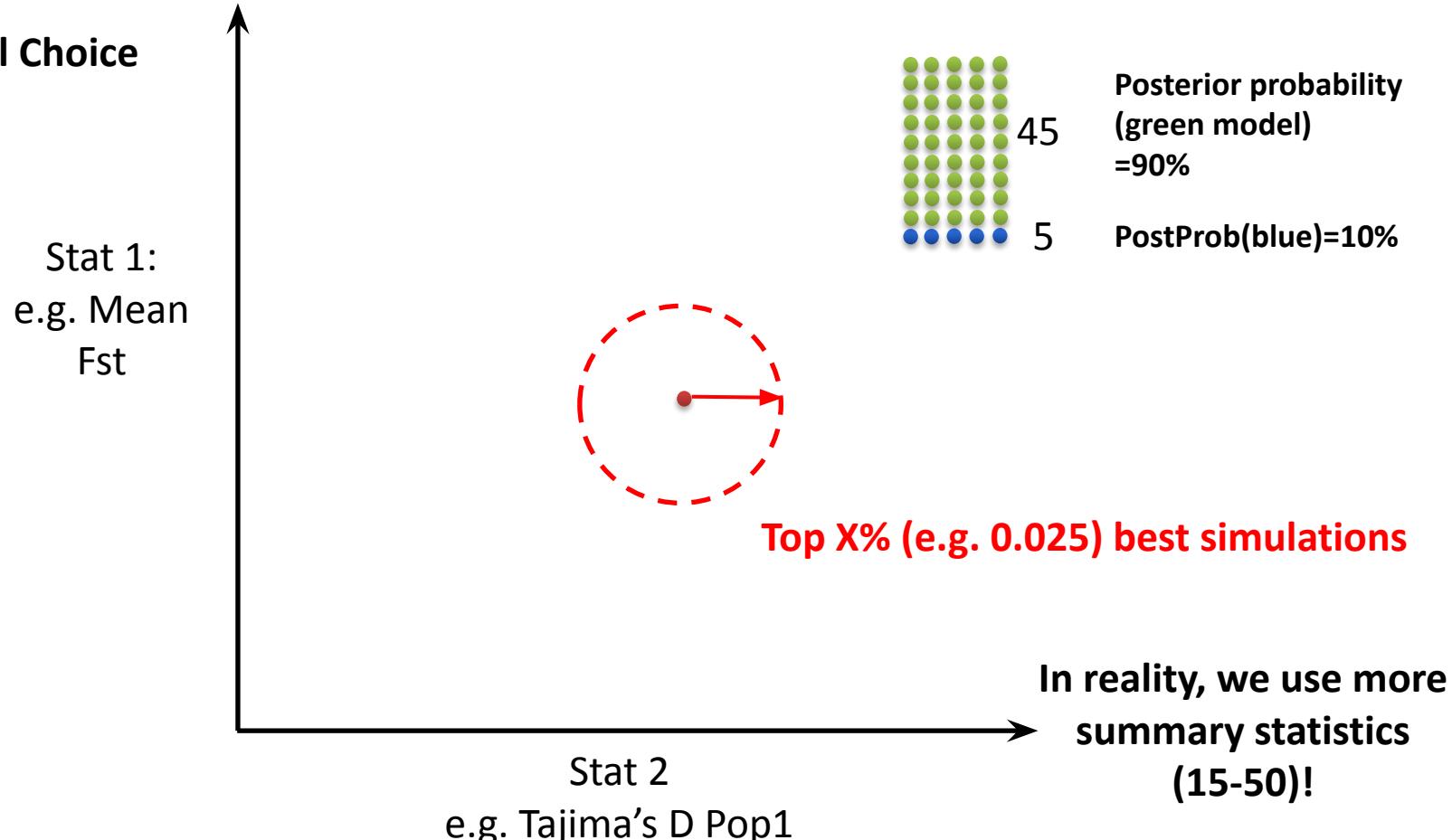
Stat 2

In reality, we use more
summary statistics
(15-50)!

ABC in practice (very simplified)...

3/ Model Choice

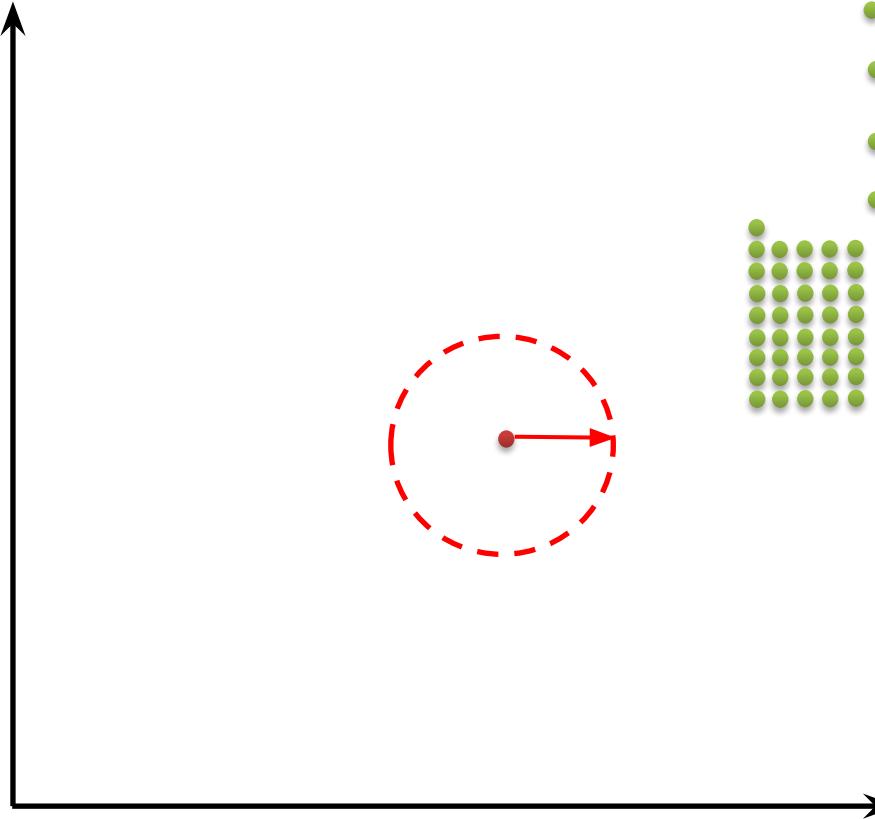
Stat 1:
e.g. Mean
Fst



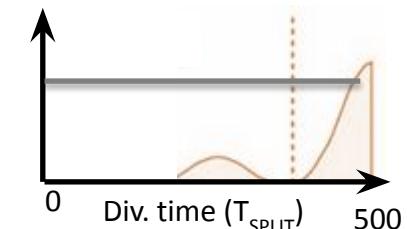
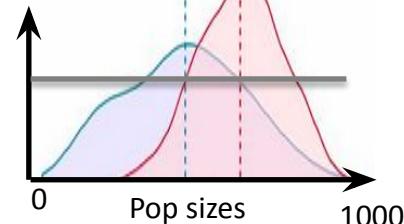
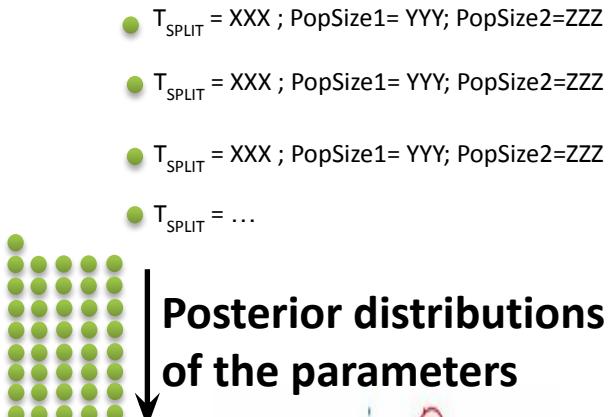
ABC in practice (very simplified)...

4/ Estimation of parameters under the best model

Stat 1:
e.g. Mean Fst

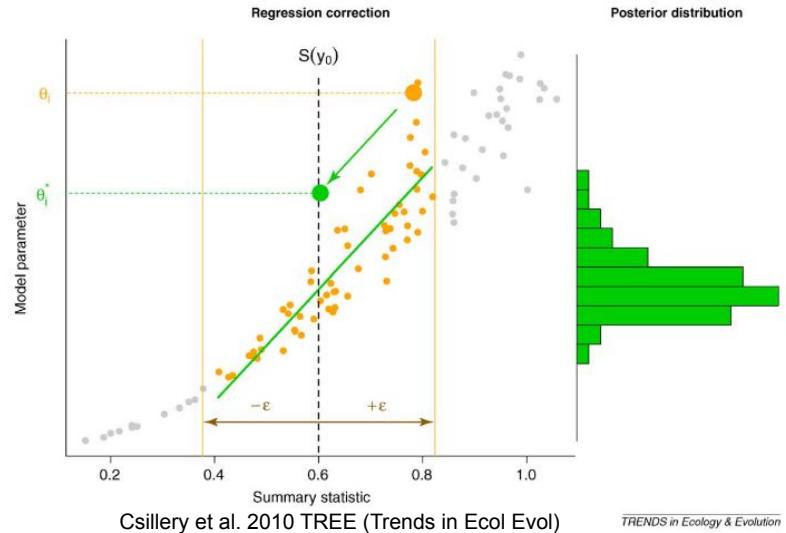


e.g. Tajima's D Pop1



ABC nowadays (very complex)...

In reality, ‘standard’ ABC algorithms use more complex strategies, potentially including a local regression adjustment (linear or not) before to generate the posterior distribution

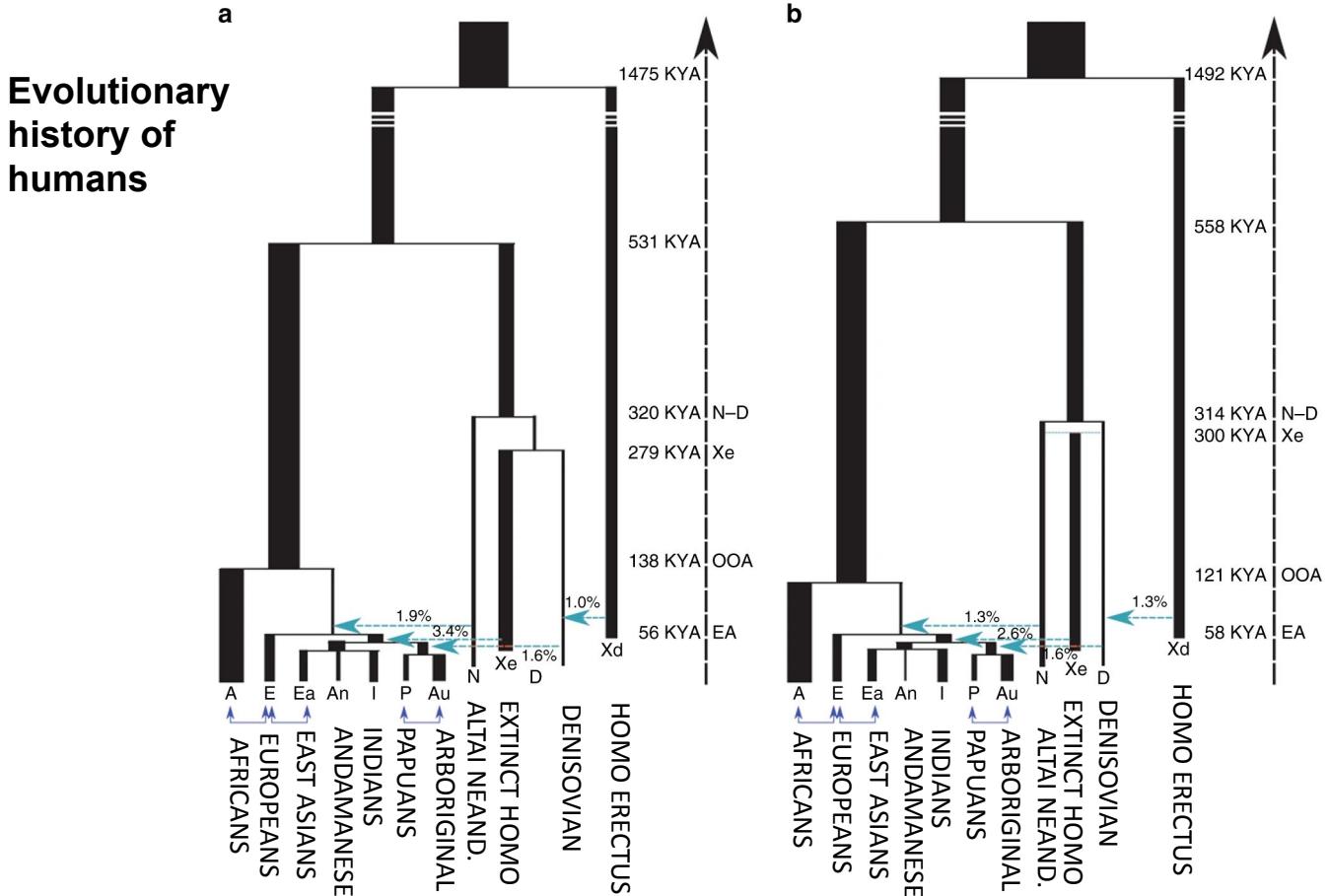


These traditional rejection/regression methods remain popular

but more and more recent ABC algorithms use complex machine learning tools (e.g. neural network (NN) or regression random forest (RF) parameter estimation in relying on the machine-learning tool to automate the inclusion of summary statistics in ABC algorithms)

-> methods learn patterns and relationships between parameter values and summary statistics across many simulations: enhancing efficiency, robustness and parameters estimation accuracy (however this comes at the cost of an increasing ‘black box’ nature)

Examples of applications



ABC with deep learning to infer extremely complex demographic scenarios involving 'ghost' populations

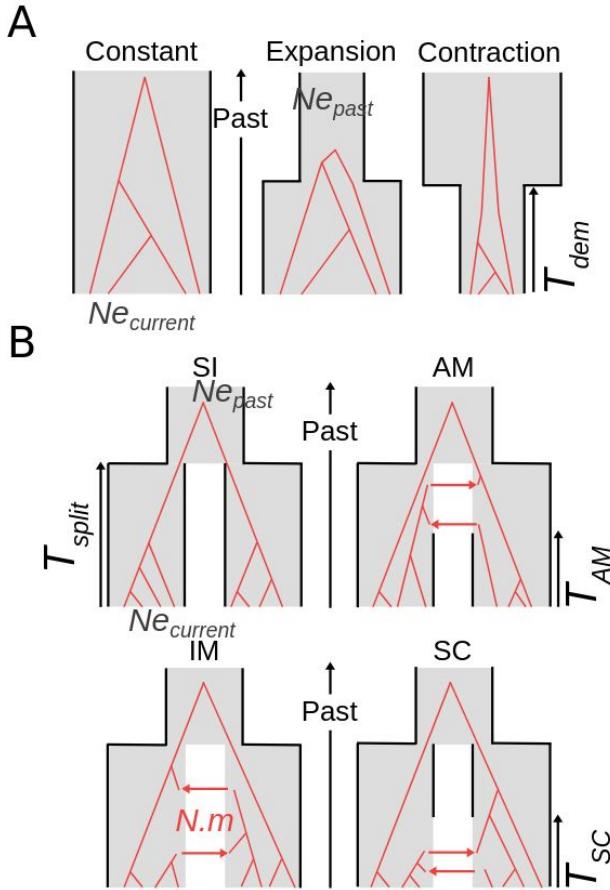
Mondal et al.
2019 Nature
Communications
(among many!)

Examples of applications

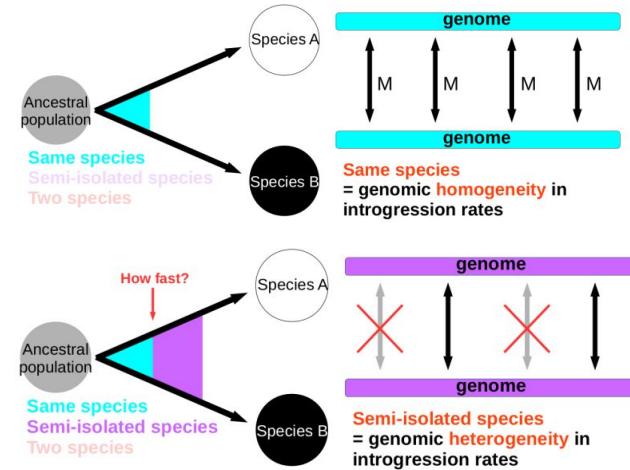
Simpler approaches:
demographic simulations in
single or pairs of populations
(DILS, cf
Camille Roux et al.)

Physalia
“Model-based
demographic
inference from
population
genomics”

Fraïsse et al. 2021
Mol Ecol Res



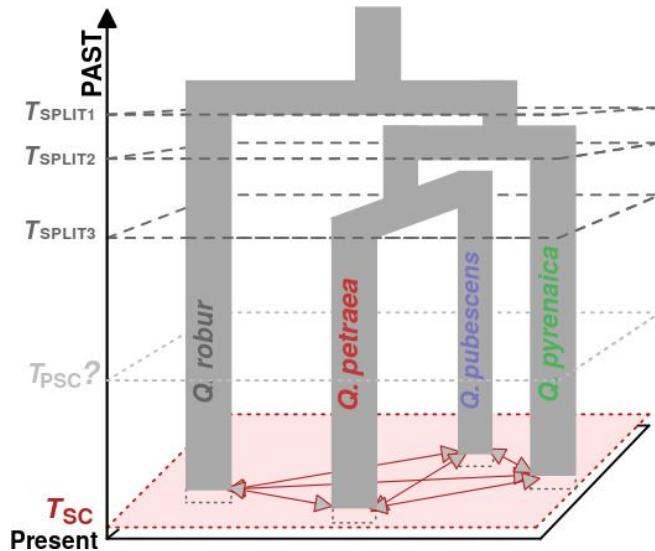
Although the number of populations is very limited, DILS accounts for many confounding factors in demographic modelling: presence of linked selection (background selection + selective sweeps), as well as the presence of barriers to gene flow



Examples of applications

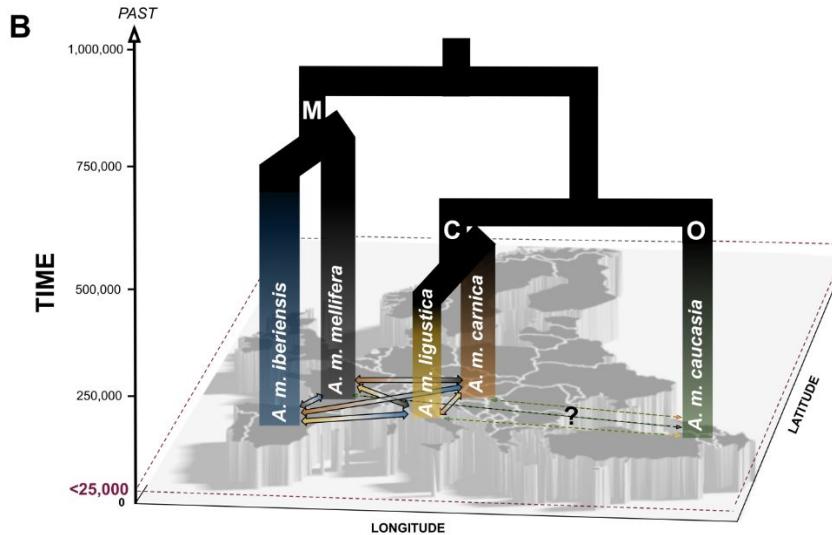
Inferences based on all pairs of pops/subspecies/species to built a general hypothetical model

Evidence for post-glacial secondary contacts in European white oaks (*Quercus*)



Leroy et al. 2017; 2020; *New Phytologist*

Evidence for post-glacial secondary contacts in Eurasian honey bees (*Apis mellifera*)

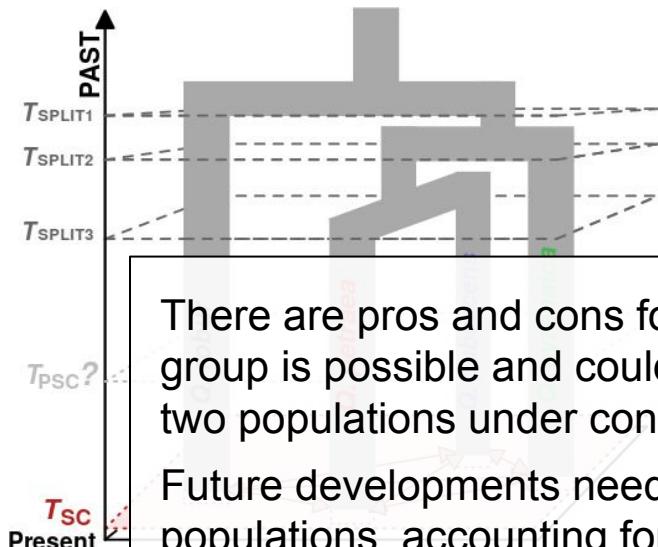


Leroy et al. 2024 *Mol Biol Evol*

Examples of applications

Inferences based on all pairs of pops/subspecies/species to built a general hypothetical model

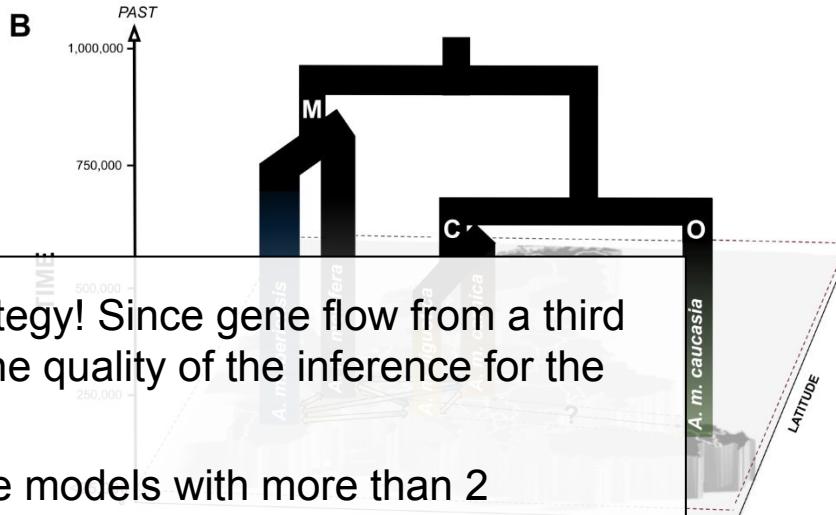
Evidence for post-glacial secondary contacts in European white oaks (*Quercus*)



There are pros and cons for this strategy! Since gene flow from a third group is possible and could impact the quality of the inference for the two populations under consideration.

Future developments needed to have models with more than 2 populations, accounting for linked selection and barriers to gene flow!

Evidence for post-glacial secondary contacts in Eurasian honey bees (*Apis mellifera*)



Approximate Bayesian Computation : pros & cons

Advantages:

- Likelihood-free
- Flexible framework

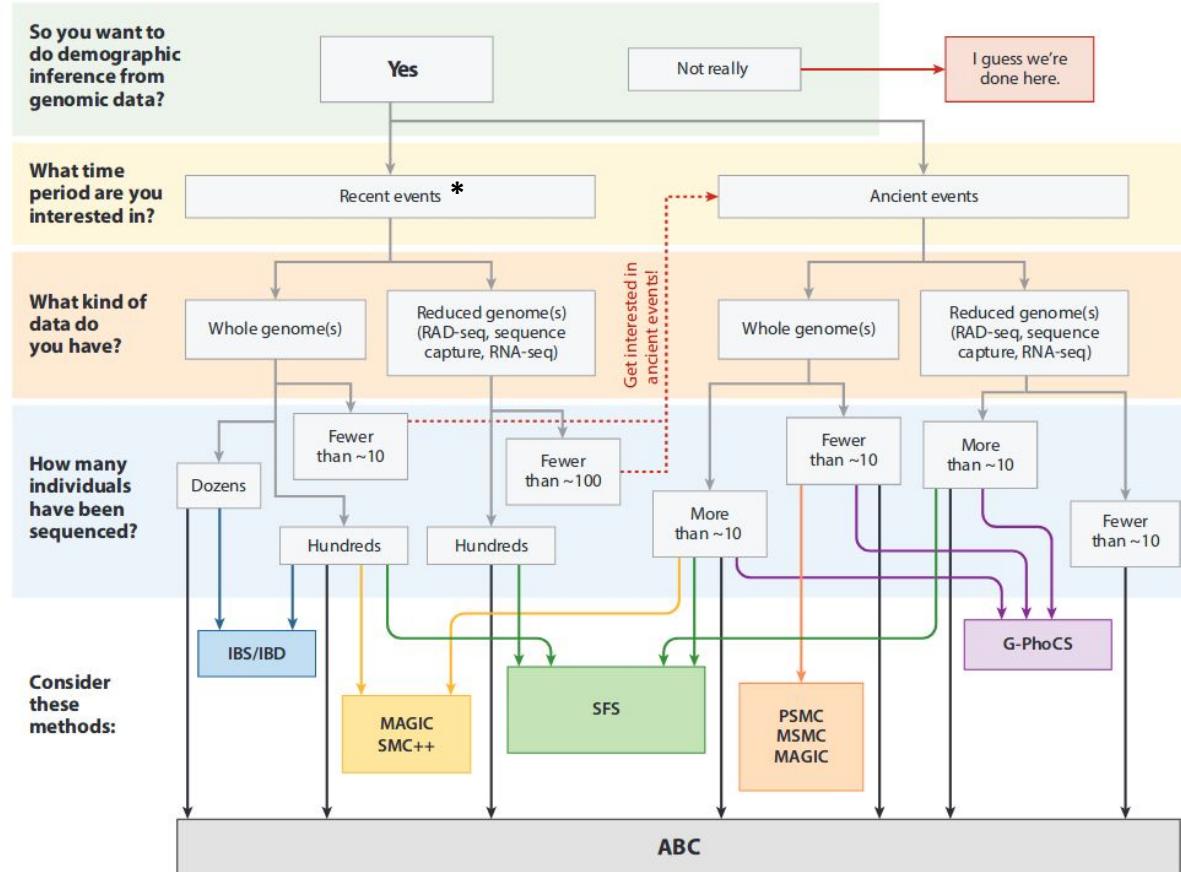
e.g. including possibility to modelize complex genome-wide processes such as heterogeneity in migration rates or effective population sizes (-> DILS)

- Both model choice and estimation of parameters are relatively straightforward
- Convenient statistical model checking based on the SFS / summary statistics

Limitations:

- Considerable computational load (still true, but now becoming less and less the case)
- Human time (especially for newbies, ABC often considered as highly complex)
- Risk of not including the true model (as for any model-based methods!)

Demographic methods: which one to prefer?



To finish, back to the low diversity of lynx...

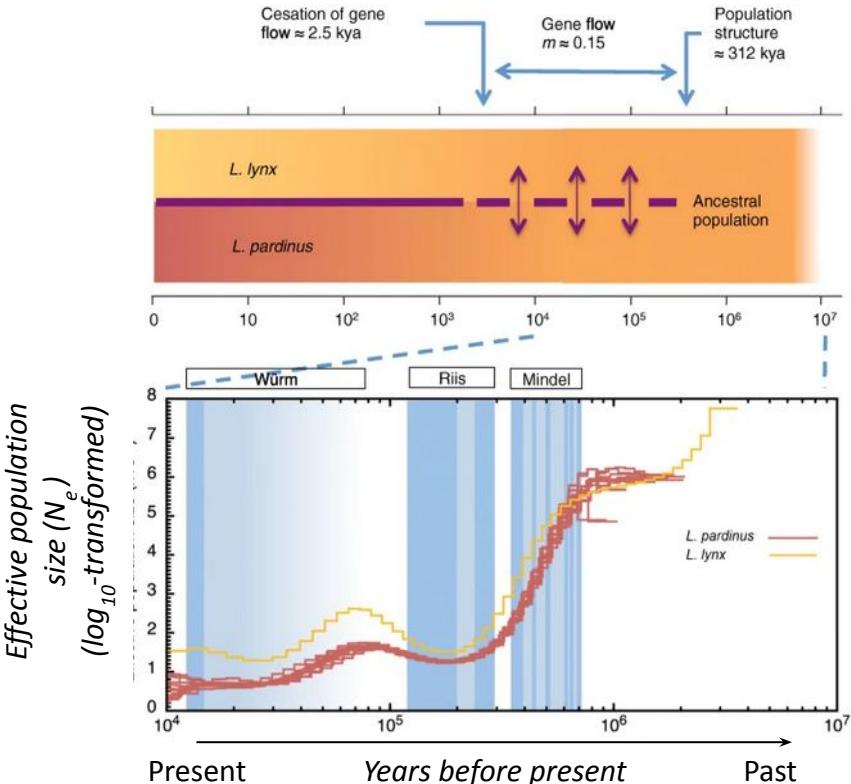
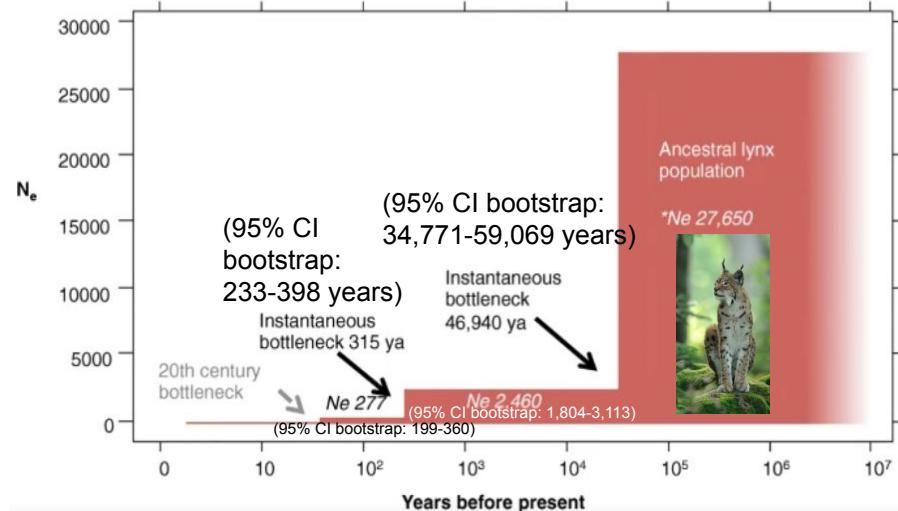


Table S20. Models considered in dadi. For each model we give the number of parameters (K), the AIC score and the AIC differences with respect to the one with the minimum AIC (Δi).

Model	ln(L)	K	AIC	Δi
Two instantaneous changes	-3116.59	4	6241.18	0.00
One exponential change followed by an instantaneous change	-3292.73	4	6593.46	-352.28
One exponential change	-4583.89	2	9171.79	-2930.60
One instantaneous change	-4881.06	2	9766.12	-3524.94



To finish, back to the low diversity of lynx...

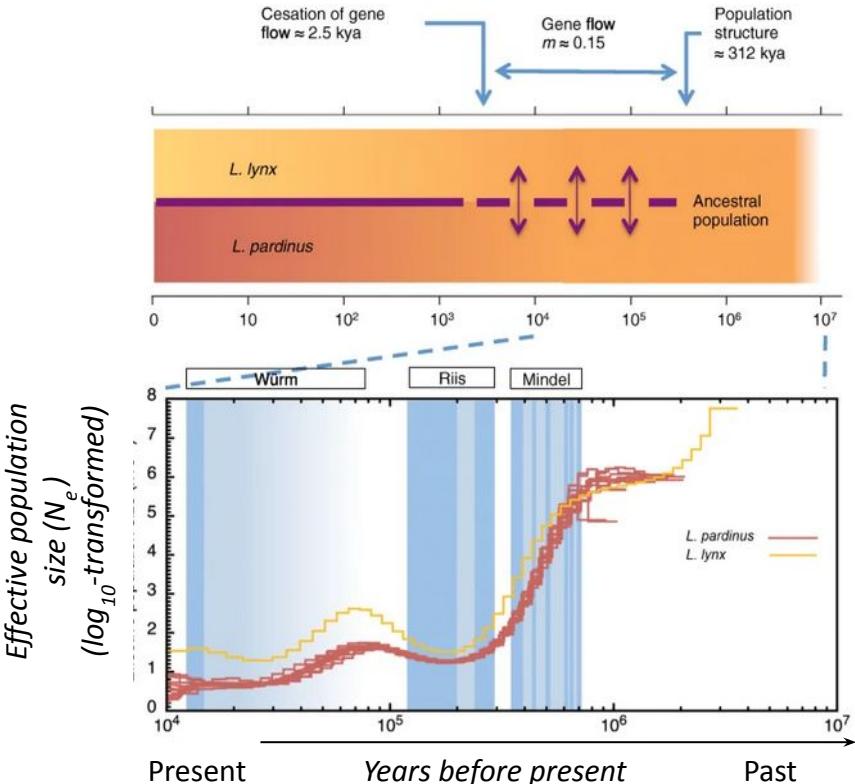
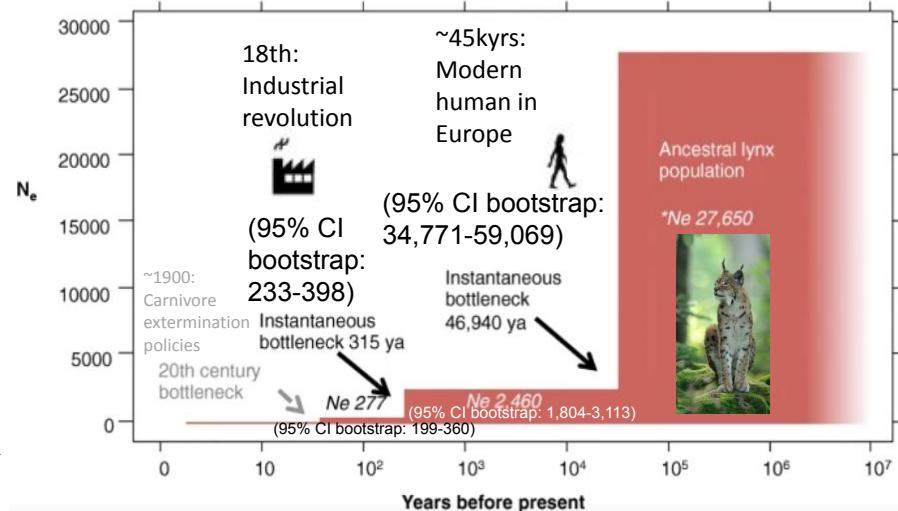


Table S20. Models considered in dadi. For each model we give the number of parameters (K), the AIC score and the AIC differences with respect to the one with the minimum AIC (Δi).

Model	$\ln(L)$	K	AIC	Δi
Two instantaneous changes	-3116.59	4	6241.18	0.00
One exponential change followed by an instantaneous change	-3292.73	4	6593.46	-352.28
One exponential change	-4583.89	2	9171.79	-2930.60
One instantaneous change	-4881.06	2	9766.12	-3524.94



To finish, back to the low diversity of lynx...

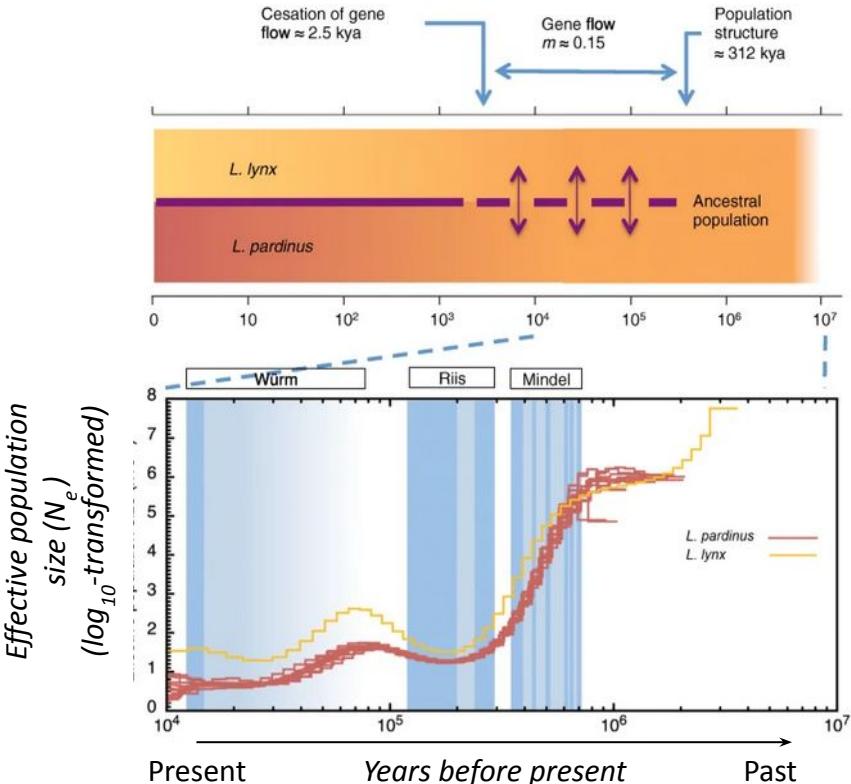
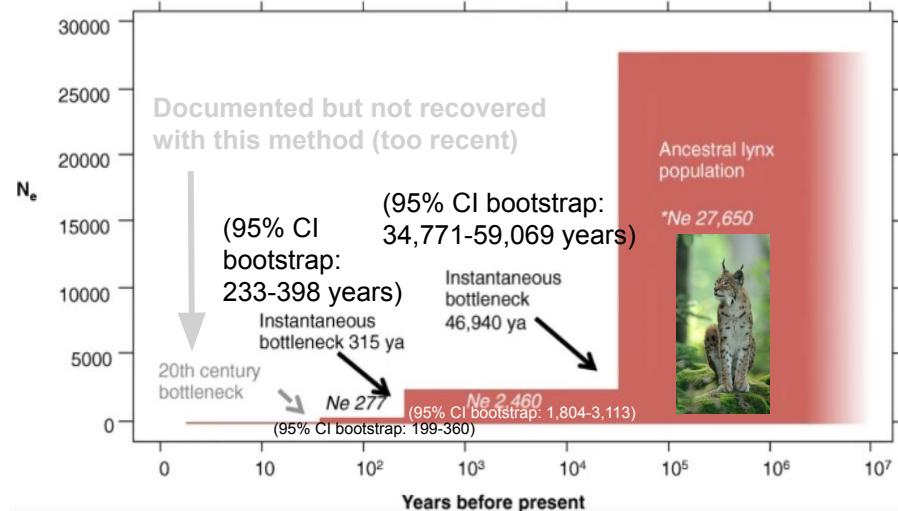


Table S20. Models considered in dadi. For each model we give the number of parameters (K), the AIC score and the AIC differences with respect to the one with the minimum AIC (Δi).

Model	ln(L)	K	AIC	Δi
Two instantaneous changes	-3116.59	4	6241.18	0.00
One exponential change followed by an instantaneous change	-3292.73	4	6593.46	-352.28
One exponential change	-4583.89	2	9171.79	-2930.60
One instantaneous change	-4881.06	2	9766.12	-3524.94



To finish, back to the low diversity of lynx... + LD-based demographic modelling

When recombination is not considered, the **mutation rate becomes the sole free parameter, guiding the timing of the coalescence process** (Hudson, 1990).

-> **Time is needed for mutations to accumulate, low resolution in the recent past.**

General idea of LD-based demographic inferences (GONE):

Inferring recent demographic history of a population (within the past 100-200 generations) **from the observed spectrum of linkage disequilibrium (LD) of pairs of loci over a wide range of recombination rates**

Recent Demographic History Inferred by High-Resolution Analysis of Linkage Disequilibrium

Enrique Santiago,^{*1} Irene Novo,² Antonio F. Pardiñas,³ María Saura,⁴ Jinliang Wang,⁵ and Armando Caballero²

¹Departamento de Biología Funcional, Facultad de Biología, Universidad de Oviedo, Oviedo, Spain

²Centro de Investigación Marína, Departamento de Bioquímica, Genética e Immunología, Edificio CC Experimentais, Campus de Vigo, Universidade de Vigo, Vigo, Spain

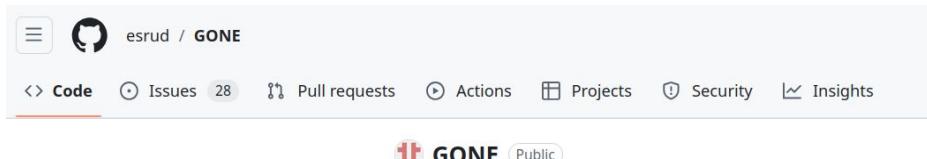
³MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, United Kingdom

⁴Departamento de Mejora Genética Animal, INIA, Madrid, Spain

⁵Institute of Zoology, Zoological Society of London, London, United Kingdom

*Corresponding author: E-mail: esr@uniovi.es.

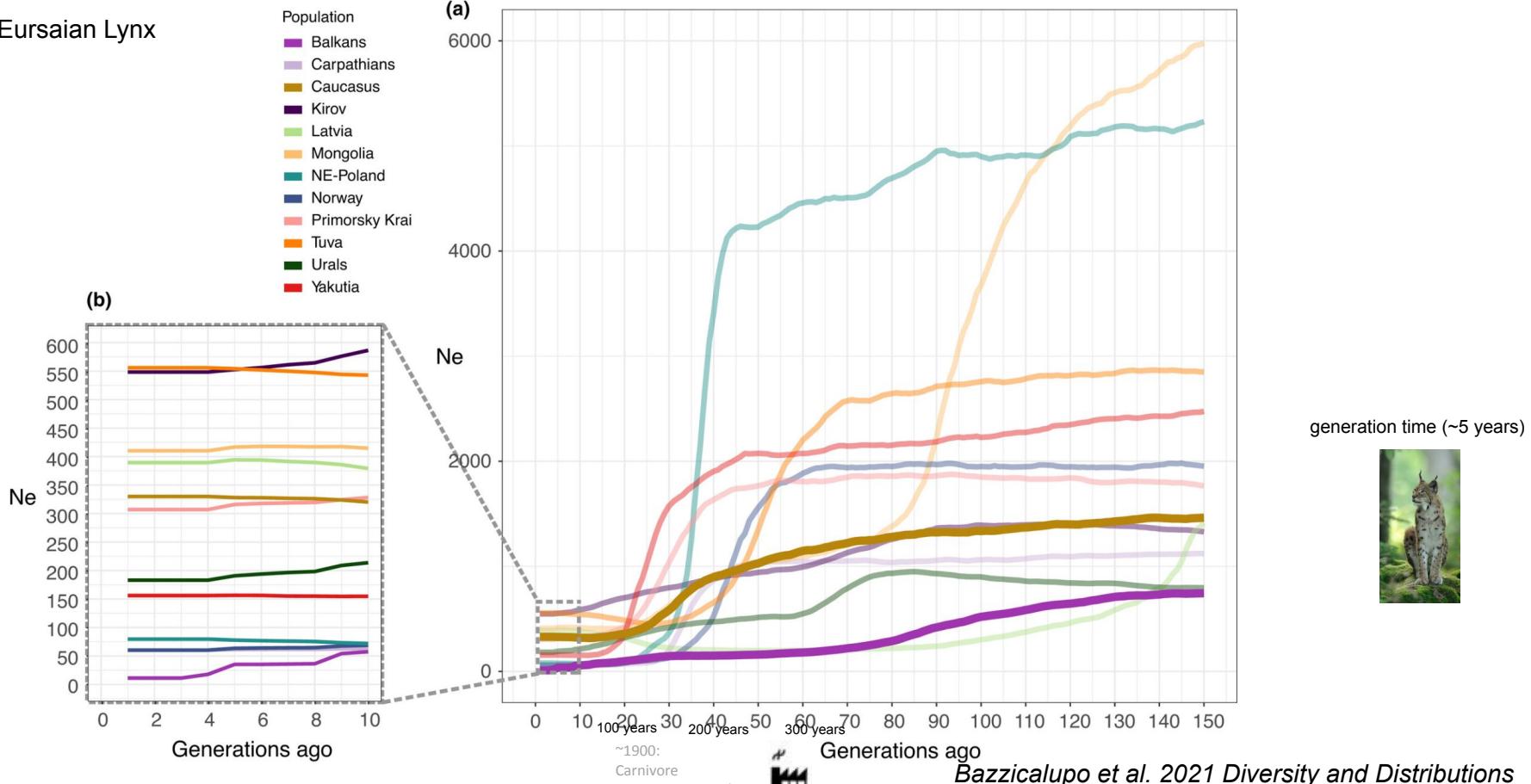
Associate editor: Yuseob Kim



- If you are interested by evolution of N_e over recent time (<100 generations), LD-based modelling approaches could be of high interest (see the practical for more details)

To finish, back to the low diversity of lynx...

Eurasian Lynx



Question 1

Which of the following can explain present-day genetic variation within a species?

- A - Recombination rate
- B - Effective population size
- C - Past demography including changes in population size and introgression
- D - Mutation rate
- E - How biased are the pipelines used to estimate it !?!

Question 1

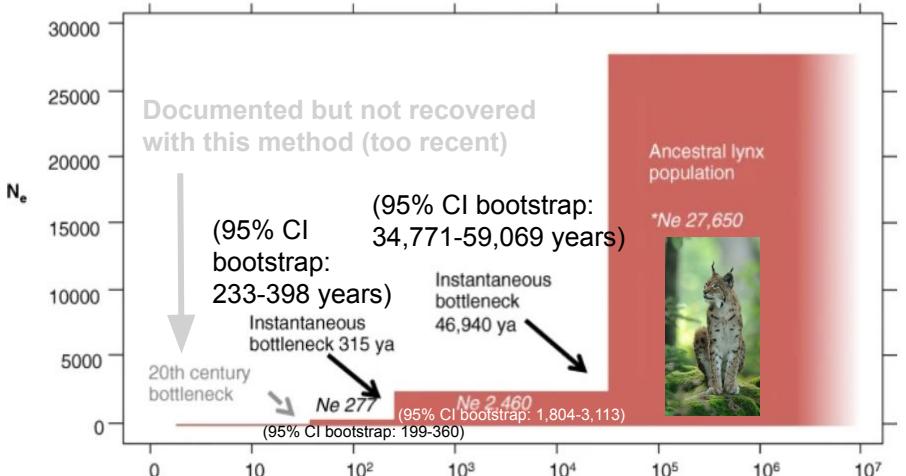
Which of the following can explain present-day genetic variation within a species?

- A - Recombination rate
- B - Effective population size
- C - Past demography including changes in population size and introgression
- D - Mutation rate
- E - How biased are the pipelines used to estimate it !?!

At equilibrium $\Theta = \pi$

$$\Theta = 2 * \text{ploidy} * N_e * \mu$$

(i.e. Θ (diploid species) = $4N_e\mu$)



Question 2

Which of the following statements about effective population size (N_e) are accurate?

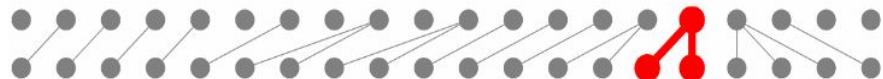
- A - It tells us how likely a population is to grow in size next generation
- B - It determines how strongly genetic drift acts in a population
- C - It's equal to the number of visible individuals in a wildlife survey
- D - It increases with the number of chromosomes in a species
- E - It affects the probability that two gene copies coalesce in the previous generation

Question 2

Which of the following statements about effective population size (N_e) are accurate?

- A - It tells us how likely a population is to grow in size next generation
- B - It determines how strongly genetic drift acts in a population
- C - It's equal to the number of visible individuals in a wildlife survey
- D - It increases with the number of chromosomes in a species
- E - It affects the probability that two gene copies coalesce in the previous generation

Effective population size (N_e): the number of individuals in a Wright-Fisher model (*i.e.* the size of an idealized population) that **would produce the same amount of genetic drift as in the real population**



The probability of 2 alleles in generation t coalesce in $t-1$ is $\frac{1}{2N_e}$
→ A direct relationship between time and N_e

Question 3

Which of the following parameters can methods like PSMC, MSMC, and SMC++ infer from genomic data?

A - Historical changes in effective population size

B - Timing of population divergence

C - Ploidy level of the organism

D - Changes in mutation rates over time

E - Your future salary trajectory

Question 3

Which of the following parameters can methods like PSMC, MSMC, and SMC++ infer from genomic data?

A - Historical changes in effective population size

B - Timing of population divergence

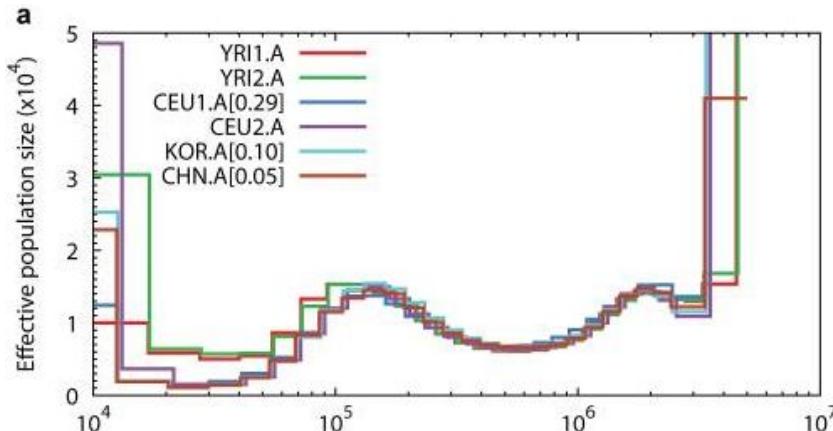
C - Ploidy level of the organism

D - Changes in mutation rates over time

E - Your future salary trajectory



(e.g. Li & Durbin, 2011 Nature)



Question 4

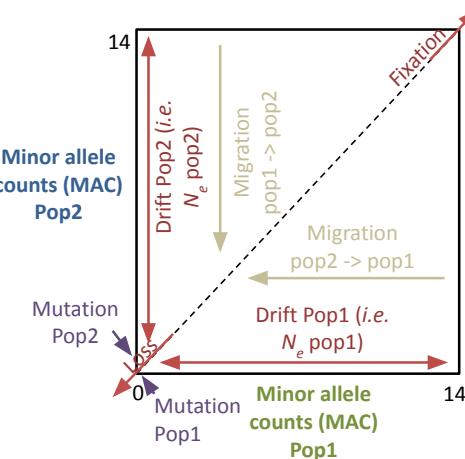
Which of the following statements best describes composite likelihood approaches ?

- A - They calculate exact likelihood across whole genomes
- B - They approximate the likelihood by estimating likelihood from the site frequency spectrum
- C - They rely only on very simple summary statistics such as Tajima's D
- D - They require unfolded spectra to produce accurate results
- E - They account for linkage disequilibrium to simplify computations

Question 4

Which of the following statements best describes composite likelihood approaches ?

- A - They calculate exact likelihood across whole genomes
- B - They approximate the likelihood by estimating likelihood from the site frequency spectrum
- C - They rely only on very simple summary statistics such as Tajima's D
- D - They require unfolded spectra to produce accurate results
- E - They account for linkage disequilibrium to simplify computations



Question 5

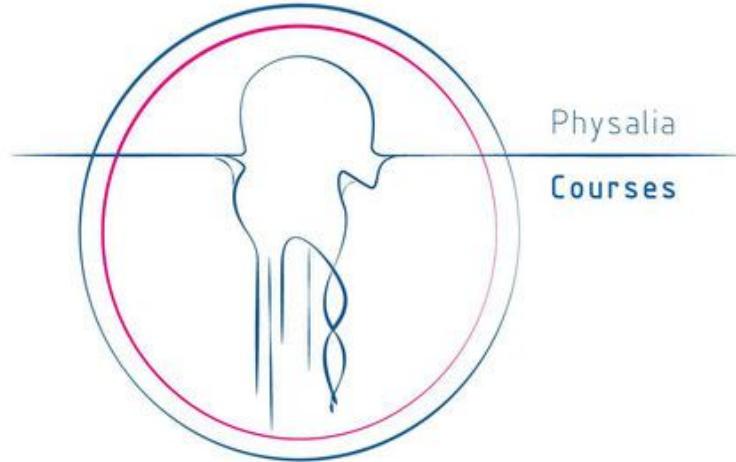
Which of the following statements best describes a characteristic of ABC?

- A - It computes exact likelihoods for complex demographic models using raw sequence data
- B - It is only applicable when data are available from multiple populations
- C - ABC compares the summary statistics of the simulated data to those of the observed data
- D - The Euclidean distance in ABC measures the genetic distance between individuals across populations
- E - It assumes that most model parameters are already known

Question 5

Which of the following statements best describes a characteristic of ABC?

- A - It computes exact likelihoods for complex demographic models using raw sequence data
- B - It is only applicable when data are available from multiple populations
- C - ABC compares the summary statistics of the simulated data to those of the observed data
- D - The Euclidean distance in ABC measures the genetic distance between individuals across populations
- E - It assumes that most model parameters are already known



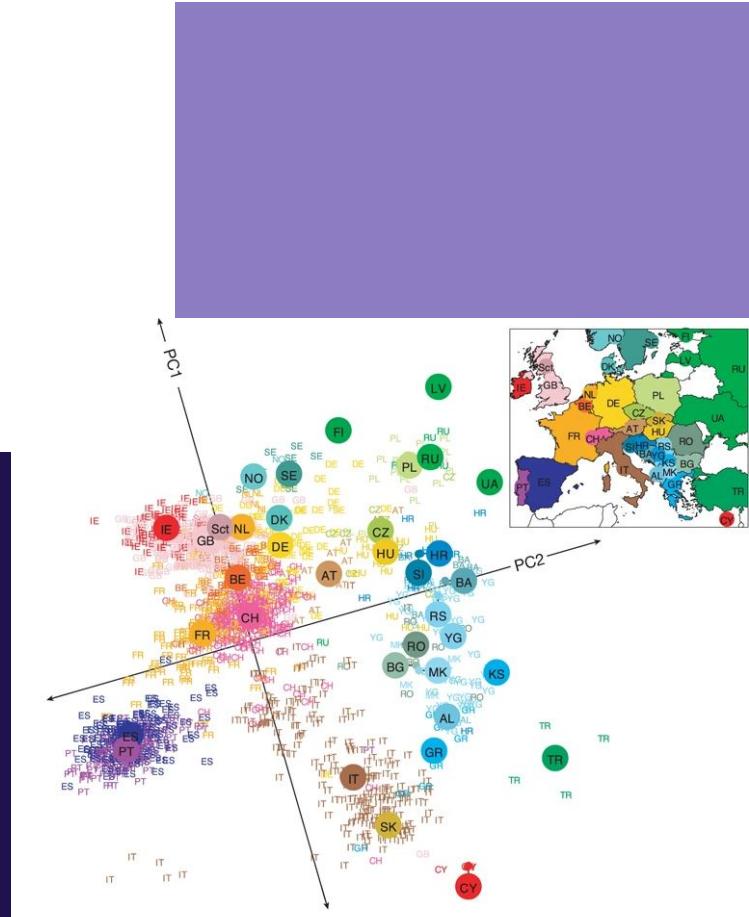
(Intro to) Demographic modeling methods

Recap Practical

28/05/2025

Physalia course

Thibault Leroy, Yann Bourgeois



Working with compressed vcf files

zmore Flowers_et_al_2019.SNPs.trnq.vcf.gz

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=hiDP,Description="DP > 2200">
##FILTER=<ID=hiFS,Description="FS > 60.0">
##FILTER=<ID=hiSOR,Description="SOR > 3.0">
##FILTER=<ID=indel6,Description="Overlaps a user-input mask">
##FILTER=<ID=loAN,Description="AN < 178">
##FILTER=<ID=loBaseQRankSum,Description="BaseQRankSum < -8.0">
##FILTER=<ID=loDP,Description="DP < 800">
##FILTER=<ID=loMQ,Description="MQ < 40.0">
##FILTER=<ID=loMQRankSum,Description="MQRankSum < -3.0">
##FILTER=<ID=loQD,Description="QD < 8.0">
##FILTER=<ID=loReadPosRankSum,Description="ReadPosRankSum < -1.5">
##FILTER=<ID=repmask,Description="Overlaps a user-input mask">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF block">
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how the alternate alleles are phased in relation to each other across the sample">
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample (but not across samples) identifies a unique haplotype phase across all variants in the sample">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description="Unconditional reference genotype confidence, encoded as a phred quality -10*log10 p(genotype call is true)">
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistics which comprise the Fisher's Exact Test to detect strand bias.">
##GATKCommandLine.GenotypeGVCFs=<ID=GenotypeGVCFs,Version=3.7-0-gcfedb67,Date="Fri Aug 04 15:34:38 EDT 2017",Epoch=1501875278990,CommandLineOptions=null,read_filter=null,disable_read_filter=null,intervals=null,excludeIntervals=null,interval_set_rule=UNION,interval_merging=ALL,interval_padding=0,reduce_padding=0>
--More--
```

Press "q" to leave

Working with compressed vcf files

zmore Flowers_et_al_2019.SNPs.tronq.vcf.gz | grep "#CHROM"

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Abel	Abouman	Ajwa
Alig	Amir_haj	Aseel	Atlantica_CAP1_POPMAL1	Atlantica_CAP50_BOA1					Aziza	Azraq_azraq	Barmel
Began	Besser_haloo	Biddajaj	Boufkouss_Rarass	BouslKhine	Braim	Canariensis_93115			Canariensis_93116		
Canariensis_93121	Canariensis_DP6A	Canariensis_JBMPL_P3		Canariensis_JBMPL_P9					Chichi	Dajwani	
Dedhi	Deglet_noor	Dibbas	Ebrahimi	Ewent_ayob	Fagous		Fard4	Faslee	Gajar	Halawy	
Hamria	Hawawiri	Hayany	Helwa	Hilali	Hiri	Horra	Jao	Jihl	Kabkab	Kamla	Karbali
Kashoowari	Khadrawy	Khalte	Khastawi		Khenezi		Khisab	Kuproo	Lulu	Maktoumi	Manjoura
Mazafati	Medjool	Nagal	Naquel_khuh	Nebeit_seif	Otaquin		Piavom		Rabee	Raslatmar	
Reclinata_DP18	Rhars	Rothan	SaidiSamany	Shagri	Silani	Sultana		Sylvestris_P59			
Sylvestris_RIV_2248_PL_F	Sylvestris_RIV_2249_PL_M	Sylvestris_RIV_2256_PL_F	Sylvestris_RIV_7394_PI_M								
Sylvestris_RIV_7395_PL_M	Sylvestris_SYL87_JCP_651	Tagiat	Theophrasti_02a						Theophrasti_05a		
Theophrasti_A1	Theophrasti_A5	Theophrasti_B1	Theophrasti_B3						Theophrasti_B5		
Theophrasti_C1	Theophrasti_C4	Theophrasti_D1	Theophrasti_D3						Theophrasti_D5		
Theophrasti_E1	Theophrasti_E2	Theophrasti_F1	Theophrasti_F2								
Theophrastis_GOLK_001_91020	Theophrastis_THE83_91051	Thory	Um_al_blaliz	Um_al_hamam					Zagloul		
Zahidi											

zmore Flowers_et_al_2019.SNPs.tronq.vcf.gz | grep "#CHROM" | awk '{print NF-9}'

105 individuals in the dataset

Working with compressed vcf files

```
zmore Flowers_et_al_2019.SNPs.tronq.vcf.gz | grep "PASS" | wc -l  
419628
```

```
zmore Flowers_et_al_2019.SNPs.tronq.vcf.gz | grep -v "#" | grep "PASS" | wc -l  
419628
```

```
zmore Flowers_et_al_2019.SNPs.tronq.vcf.gz | awk '$7 == "PASS" {print $0}' | wc -l  
419628
```

OPTIONAL: Can you estimate a variant density (# high quality SNPs / length of the first scaffold on the VCF)?

```
zmore Flowers_et_al_2019.SNPs.tronq.vcf.gz | grep -v "#" | head -1  
NW_008246507.1 6801 . C A 2072.27 PASS (continued...)
```

```
zmore Flowers_et_al_2019.SNPs.tronq.vcf.gz | grep "NW_008246507.1" | tail -1  
NW_008246507.1 4532558 . A G 20425.49 PASS (continued...)
```

```
zmore Flowers_et_al_2019.SNPs.tronq.vcf.gz | grep "NW_008246507.1" | awk '$7 == "PASS" {print $0}' | wc -l  
142392
```

$142392 / 4532558 = 0.0314 \rightarrow 1/0.0314$ On average, **one variant every 31.8 bases** in the vcf (1st scaff)
Limits: rough estimate, it doesn't take into account the some regions of the scaff completely uncovered

Step 1: identify non-admixed individuals

head -3 samples_Qvalues.txt

IID	Species	Cluster_outgroups	Cluster_Pdac	Cluster_Ptheo	Note
Canariensis_93115	P._canariensis	0.99998	1E-05	1E-05	'Pure'
Canariensis_93116	P._canariensis	0.99998	1E-05	1E-05	'Pure'

```
grep "P._dactylifera_ME" samples_Qvalues.txt | awk '$4 > 0.90 {print $0}' > samples_Qvalues.PdactyliferaME.txt
```

head -2 samples_Qvalues.PdactyliferaME.txt

Abouman	P._dactylifera_ME	1E-05	0.99998	1E-05	'Pure'
Ajwa	P._dactylifera_ME	1E-05	0.99998	1E-05	'Pure'

```
grep "P._dactylifera_NAfr" samples_Qvalues.txt | awk '$4 > 0.90 {print $0}' > samples_Qvalues.PdactyliferaNAfr.txt  
grep "P._theophrasti" samples_Qvalues.txt | awk '$5 > 0.90 {print $0}' > samples_Qvalues.Ptheophrasti.txt
```

awk '{print \$1}' samples_Qvalues.PdactyliferaME.txt > samples_Qvalues.PdactyliferaME.list

head -2 samples_Qvalues.PdactyliferaME.list

Abouman

Ajwa

awk '{print \$1}' samples_Qvalues.PdactyliferaNAfr.txt > samples_Qvalues.PdactyliferaNAfr.list

awk '{print \$1}' samples_Qvalues.Ptheophrasti.txt > samples_Qvalues.Ptheophrasti.list

Step 1: identify non-admixed individuals

How many individuals are identified as "relatively pure" in each population or species based on this criterion?

```
less samples_Qvalues.PdactyliferaME.list | wc -l  
45
```

```
less samples_Qvalues.PdactyliferaNAfr.list | wc -l  
9
```

```
less samples_Qvalues.Ptheophrasti.list | wc -l  
16
```

...

```
wc -l samples_Qvalues.PdactyliferaME.list  
45 samples_Qvalues.PdactyliferaME.list  
...
```

```
for i in *list; do wc -l $i; done  
45 samples_Qvalues.PdactyliferaME.list  
9 samples_Qvalues.PdactyliferaNAfr.list  
16 samples_Qvalues.Ptheophrasti.list
```

Remember to:

- infer population structure (PCA, Structure, ...)
- remove individuals with footprints of recent admixture
- remove individuals with very recent family relationships (if any)

... before to start inferring past demography!

To recap, here:

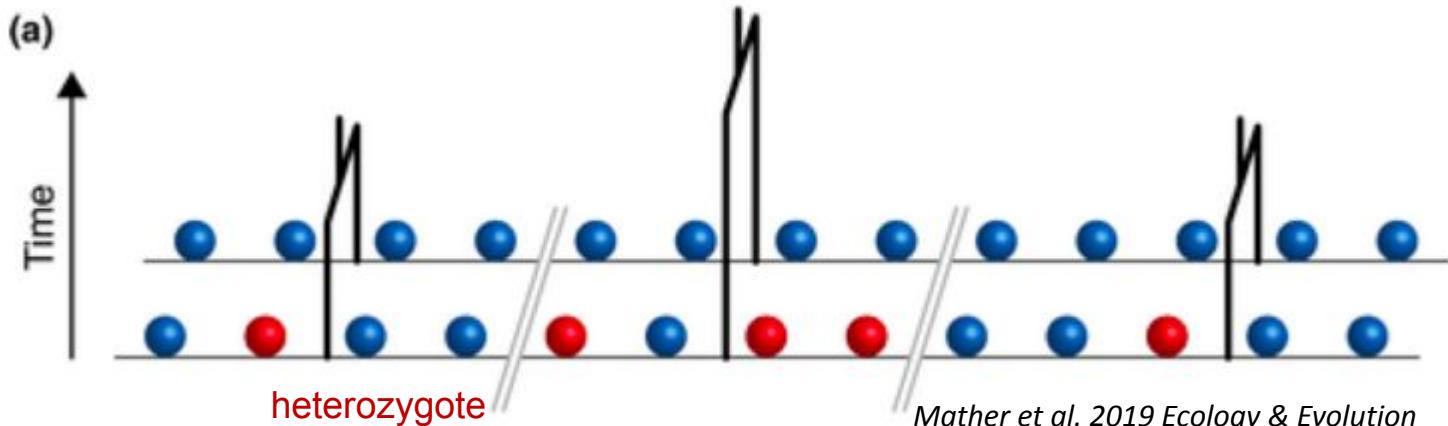
45 "pure" individuals from *Phoenix dactylifera* (Middle East)

9 "pure" individuals from *P. dactylifera* (North Africa)

16 "pure" individuals from *P. theophrastii*

Step 2: Perform inferences of N_e with SMC++

TMRCA
between the
two alleles
carried by an
individual



Before

```
zmore Flowers_et_al_2019.SNPs.tronq.vcf.gz | grep "NW_008246507.1" | awk '$7 == "PASS" {print $0}' | wc -l  
142392
```

After

```
less ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.tronq.vcf | grep "NW_008246507.1" | awk '$7 ==  
"PASS" {print $0}' | wc -l  
142392
```

No changes! What does it mean?

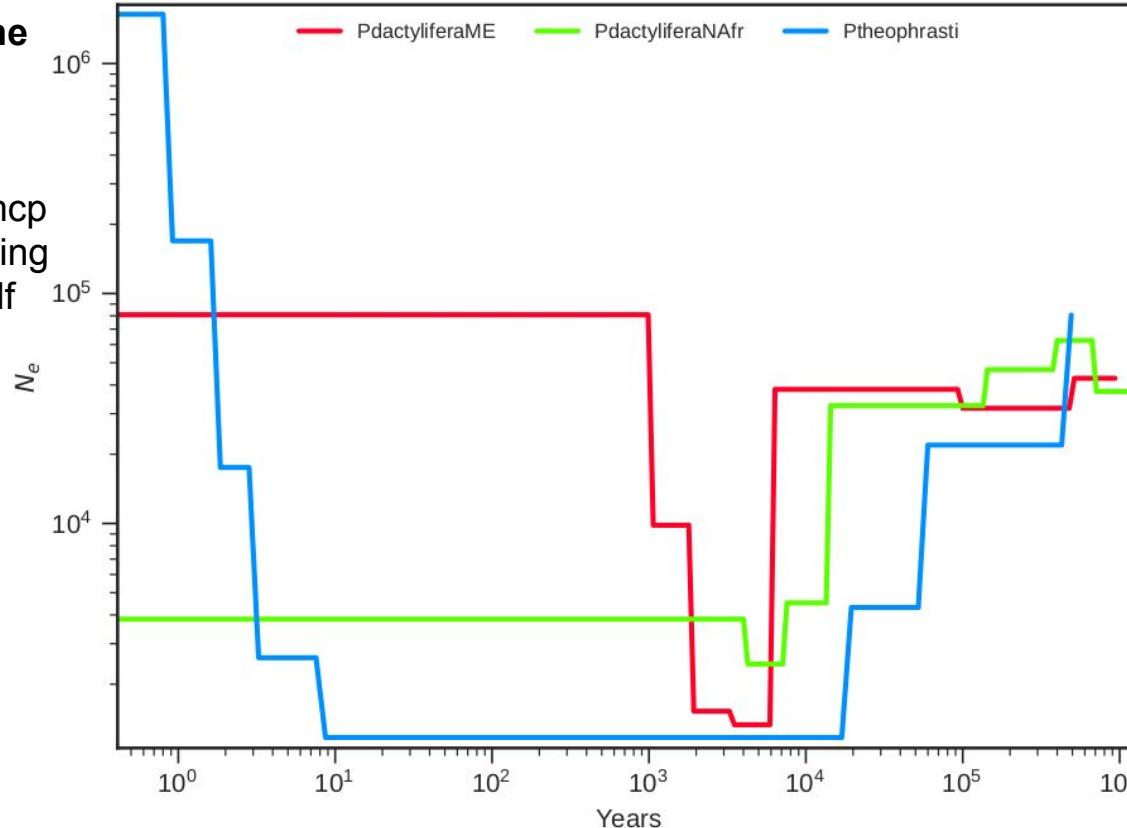
Flowers et al. only deposited variants corresponding to biallelic SNPs!

Step 2: Perform inferences of N_e with SMC++

```
smc++ plot -g 10 -c physalia_smcpp_plot_missingcutoff500.pdf PdactyliferaMe/model.final.json  
PdactyliferaNAfr/model.final.json Ptheophrasti/model.final.json
```

output of the
smc++ plot
function

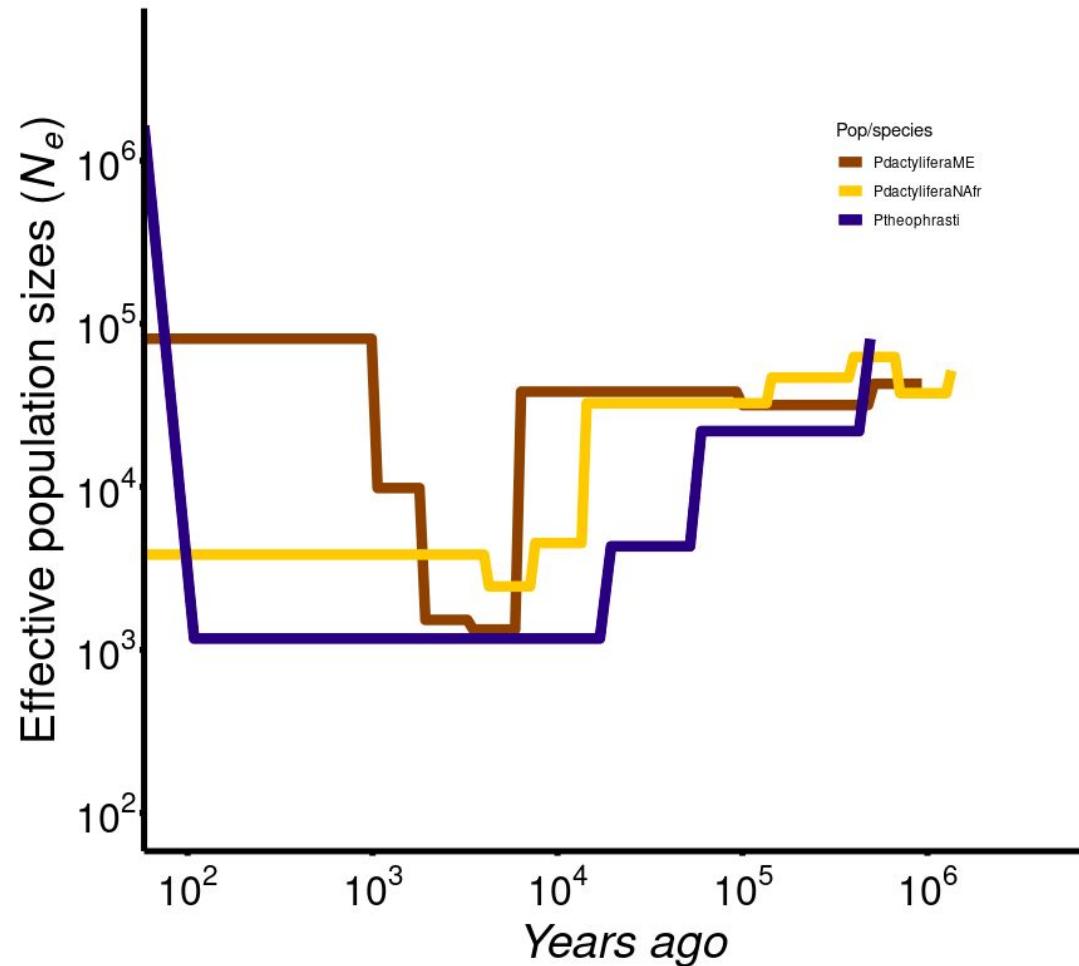
physalia_smcp
p_plot_missing
cutoff500.pdf



Step 2: Perform inferences of N_e with SMC++

```
inferredNe500=read.csv("physalia_smcpp_plot.csv",header=TRUE,sep=",")  
colnames(inferredNe500)<-c("Population","Years","Ne","plot-type","plot-num")  
My_colors <-  
c("PdactyliferaME"="chocolate4","PdactyliferaNAfr"="goldenrod1","Ptheophrasti"="navyblue")  
  
ggplot(inferredNe500, aes(x=Years, y=Ne)) +  
  geom_line(aes(colour=Population),size=3)+  
  scale_colour_manual("Pop/species",values=My_colors)+  
  xlab("Years ago")+ ylab(expression(paste("Effective population sizes (", italic(N[e]), ")")))+  
  scale_x_log10(limits=c(100,5000000),breaks = scales::trans_breaks("log10", function(x) 10^x),labels =  
  scales::trans_format("log10", scales::math_format(10^.x)))+  
  scale_y_log10(limits=c(100,5000000),breaks = scales::trans_breaks("log10", function(x) 10^x),labels =  
  scales::trans_format("log10", scales::math_format(10^.x)))+  
  theme_bw() +  
  theme(plot.margin = margin(1,3,1.5,1.2, "cm")) +  
  theme(legend.position="none", panel.border = element_blank(), panel.grid.major = element_blank(),  
  panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")) +  
  theme(axis.line = element_line(colour = 'black', size = 1.75), axis.ticks = element_line(colour = 'black',  
  size = 1.75),  
  axis.text.x = element_text(colour="black",size=24,angle=0,hjust=.5,vjust=.5,face="plain"),  
  axis.text.y = element_text(colour="black",size=24,angle=0,hjust=.5,vjust=.5,face="plain"),  
  axis.title.x = element_text(colour="black",size=28,angle=0,hjust=.5,vjust=.2,face="italic"),  
  axis.title.y = element_text(colour="black",size=28,angle=90,hjust=.5,vjust=.5,face="italic"))
```

Step 2: Perform inferences of N_e with SMC++



Step 2: Perform inferences of N_e with SMC++

Optional 1

```
cd ~/Day3_demography/data-datepalm/Res_SMCpp/res_SMCpp_missing50kb/
```

```
smc++ plot -g 10 -c .../physalia_smcpp_plot_missingcutoff50000.pdf PdactyliferaMe/model.final.json  
PdactyliferaNAfr/model.final.json Ptheophrasti/model.final.json  
= same command than before but from the res_SMCpp_missing50kb/
```

ls

PdactyliferaMe

physalia_smcpp_plot_missingcutoff50000.csv

physalia_smcpp_plot_missingcutoff500.csv_FOR_IMPATIENT_PEOPLE_ONLY.csv

Ptheophrasti

Rplot_ggplot2_smcpp_missingcutoff500.pdf

PdactyliferaNAfr

physalia_smcpp_plot_missingcutoff50000.pdf

physalia_smcpp_plot_missingcutoff500.pdf

res_SMCpp_missing50kb

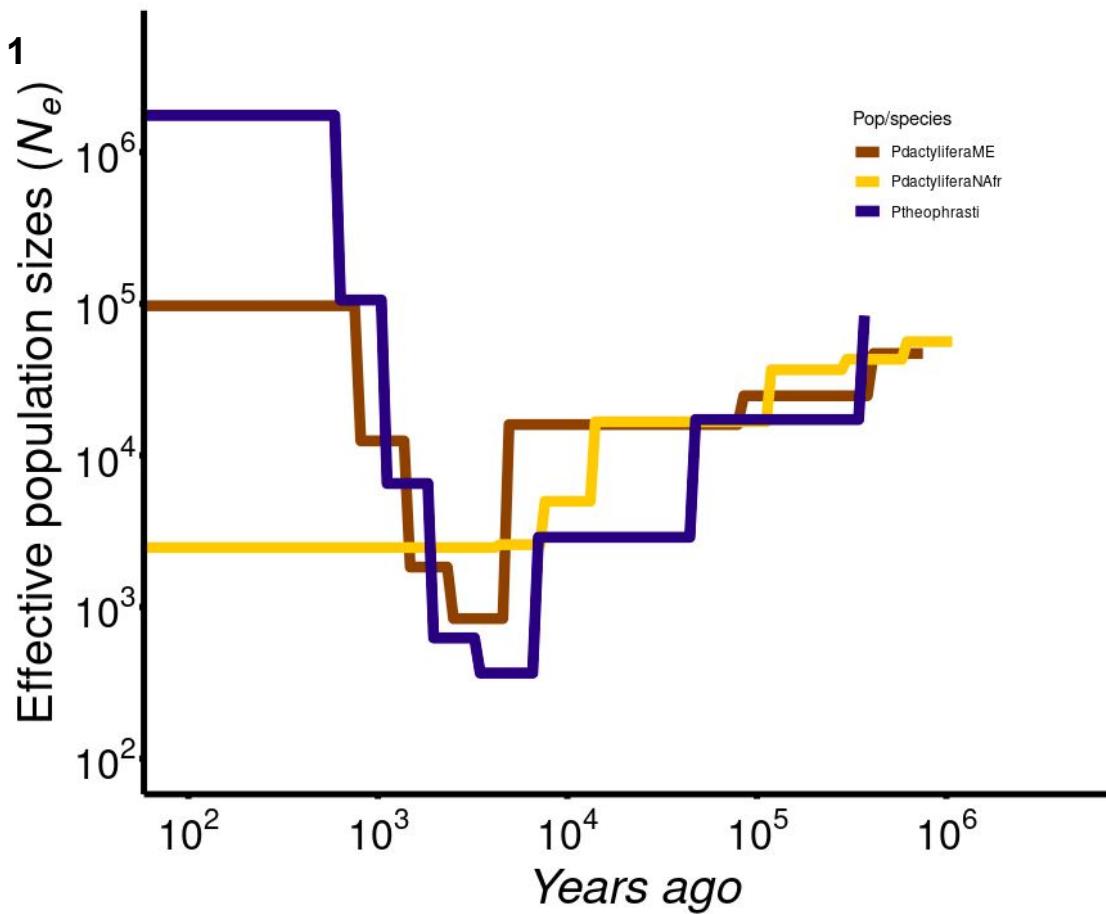
script_generate_smcpp.R

Step 2: Perform inferences of N_e with SMC++

```
inferredNe50kb=read.csv("physalia_smcpp_plot_missing50kb.csv",header=TRUE,sep=",")  
colnames(inferredNe50kb)<-c("Population","Years","Ne","plot-type","plot-num")  
  
ggplot(inferredNe50kb, aes(x=Years, y=Ne)) +  
  geom_line(aes(colour=Population),size=3)+  
  scale_colour_manual("Pop/species",values=My_colors)+  
  xlab("Years ago")+ ylab(expression(paste("Effective population sizes (", italic(N[e]), ")")))+  
  scale_x_log10(limits=c(100,5000000),breaks = scales::trans_breaks("log10", function(x) 10^x),labels =  
  scales::trans_format("log10", scales::math_format(10^.x)))+  
  scale_y_log10(limits=c(100,5000000),breaks = scales::trans_breaks("log10", function(x) 10^x),labels =  
  scales::trans_format("log10", scales::math_format(10^.x)))+  
  theme_bw() +  
  theme(plot.margin = margin(1,3,1.5,1.2, "cm")) +  
  theme(theme(legend.position=c(0.8, 0.8), panel.border = element_blank(), panel.grid.major =  
  element_blank(), panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")) +  
  theme(axis.line = element_line(colour = 'black', size = 1.75), axis.ticks = element_line(colour = 'black', size =  
  1.75),  
  axis.text.x = element_text(colour="black",size=24,angle=0,hjust=.5,vjust=.5,face="plain"),  
  axis.text.y = element_text(colour="black",size=24,angle=0,hjust=.5,vjust=.5,face="plain"),  
  axis.title.x = element_text(colour="black",size=28,angle=0,hjust=.5,vjust=.2,face="italic"),  
  axis.title.y = element_text(colour="black",size=28,angle=90,hjust=.5,vjust=.5,face="italic"))
```

Step 2: Perform inferences of N_e with SMC++

Optional 1



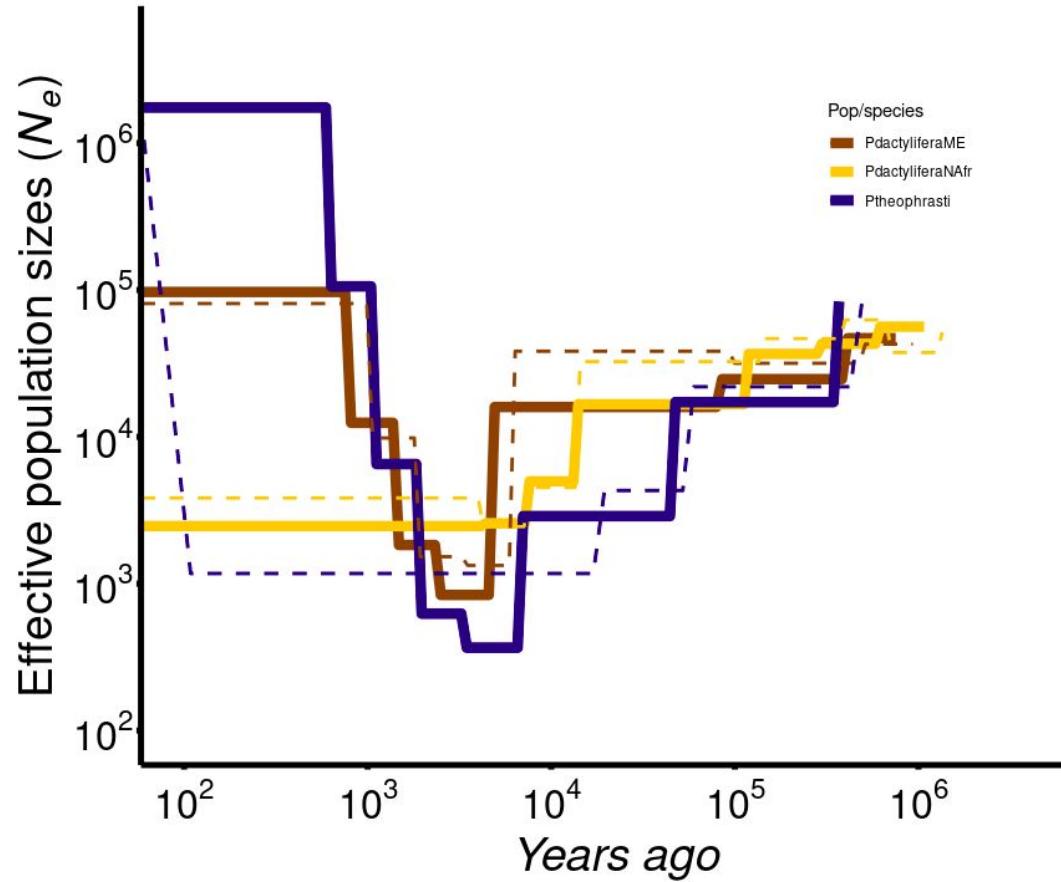
Step 2: Perform inferences of N_e with SMC++

Optional 2

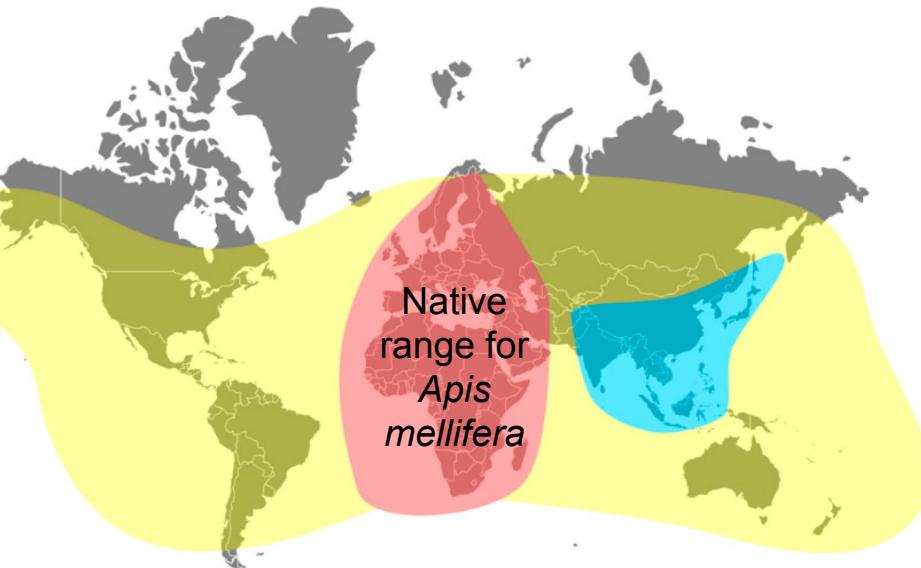
```
ggplot() +  
  geom_line(data=inferredNe500,aes(x=Years, y=Ne,colour=Population),size=1,lty=2)+  
  geom_line(data=inferredNe50kb,aes(x=Years, y=Ne,colour=Population),size=3)+  
  scale_colour_manual("Pop/species",values=My_colors)+  
  xlab("Years ago")+ ylab(expression(paste("Effective population sizes (",italic(N[e]),")")))+  
  scale_x_log10(limits=c(100,5000000),breaks = scales::trans_breaks("log10", function(x) 10^x),labels =  
  scales::trans_format("log10", scales::math_format(10^.x)))+  
  scale_y_log10(limits=c(100,5000000),breaks = scales::trans_breaks("log10", function(x) 10^x),labels =  
  scales::trans_format("log10", scales::math_format(10^.x)))+  
  theme_bw()+  
  theme(plot.margin = margin(1,3,1.5,1.2, "cm"))+  
  theme(legend.position=c(0.8, 0.8),panel.border = element_blank(), panel.grid.major = element_blank(),  
  panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))+  
  theme(axis.line = element_line(colour = 'black', size = 1.75), axis.ticks = element_line(colour = 'black', size =  
  1.75),  
  axis.text.x = element_text(colour="black",size=24,angle=0,hjust=.5,vjust=.5,face="plain"),  
  axis.text.y = element_text(colour="black",size=24,angle=0,hjust=.5,vjust=.5,face="plain"),  
  axis.title.x = element_text(colour="black",size=28,angle=0,hjust=.5,vjust=.2,face="italic"),  
  axis.title.y = element_text(colour="black",size=28,angle=90,hjust=.5,vjust=.5,face="italic"))
```

Step 2: Perform inferences of N_e with SMC++

Optional 2



Apis mellifera (Western honey bee)

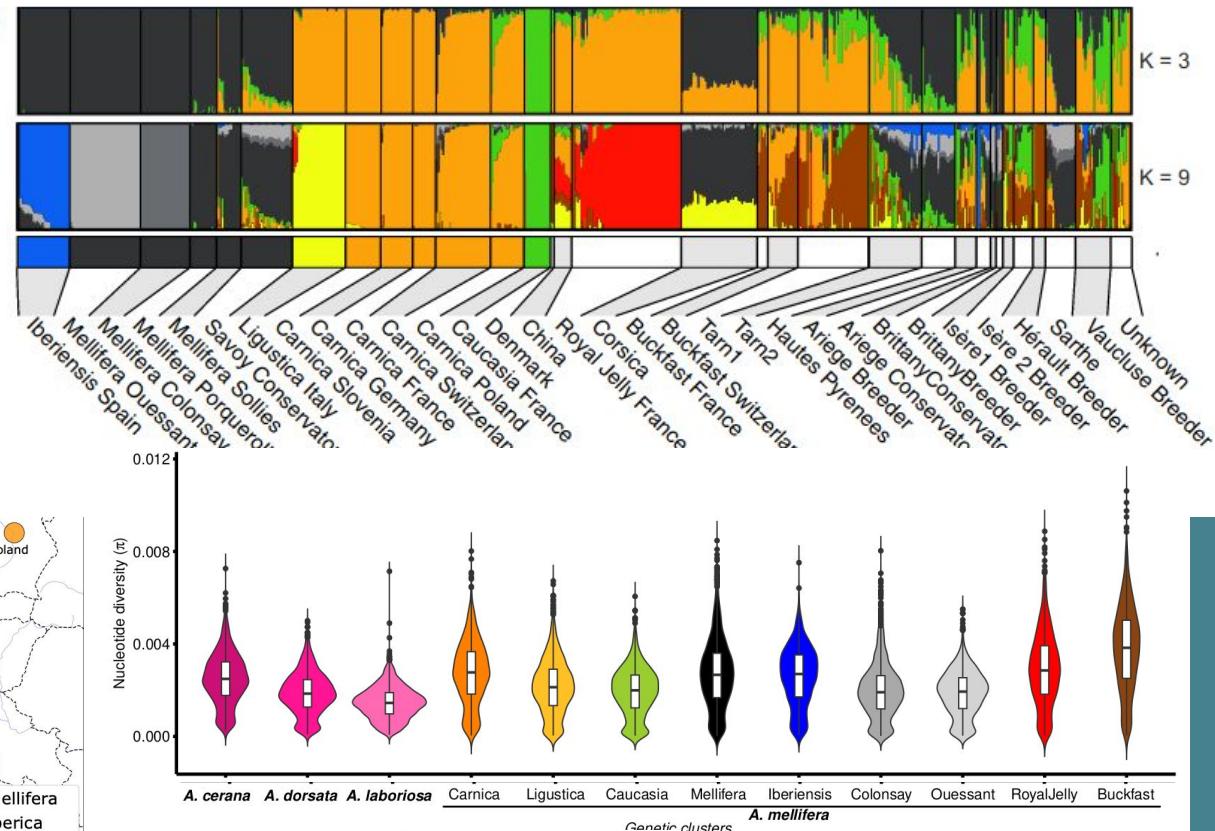
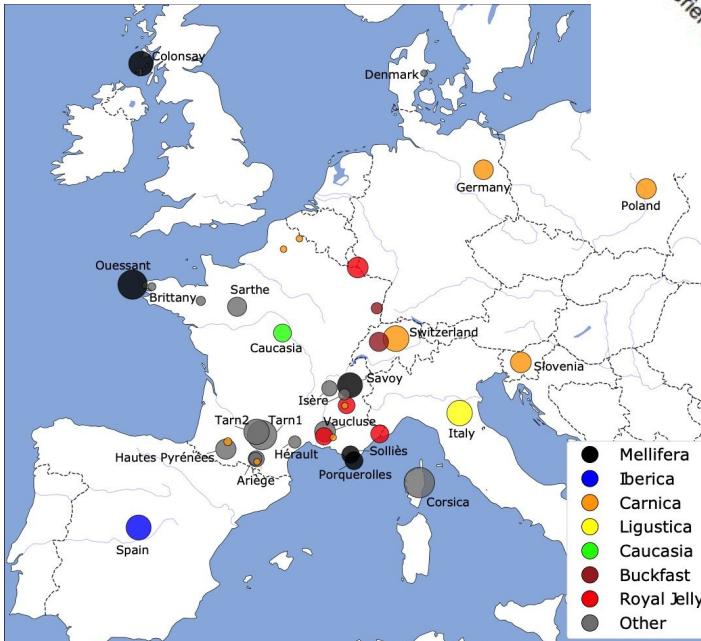


Beaurepaire et al. 2020



Complex population structure and haplotype patterns i (c) Western European honey bee from sequencing a large haploid drones

David Wragg¹ | Sonia E. Eynard¹ | Benjamin Basso² | Kamila Canale-Tabe
 Emmanuelle Labarthe¹ | Olivier Boucher³ | Kaspar Bienefeld⁴ |
 Małgorzata Bieńkowska⁵ | Cecilia Costa⁶ | Aleš Gregorc⁷ |
 Per Kryger⁸ | Melanie Parejo⁹ | M. Alice Pinto¹⁰ | Jean-Pierre Bidanel¹¹ |
 Bertrand Servin¹ | Yves Le Conte¹² | Alain Vignal¹



Inferring Long-Term and Short-Term Determinants of
Genetic Diversity in Honey Bees: Beekeeping Impact
and Conservation Strategies

Thibault Leroy^{1,4} Pierre Fau¹, Benjamin Basso², Sonia Eynard¹, David Wragg^{1,3},
 Alain Vignal¹

¹GenPhyS, Université de Toulouse, INRAE, ENVT, Castanet Tolosan 31326 France

²Abellès et Environnement, INRAE, Avignon, France

³Beebytes Analytics CIC, Roslin Innovation Centre, Easter Bush Campus, Midlothian, UK

⁴Université de Toulouse, INRAE, ENVT, Castanet Tolosan 31326 France

Leroy et al. 2024
Molecular Biology
and Evolution

Step 3: Perform inferences of recent N_e

Do you notice anything particularly unique about honey bees in the parameter file for GONE?

```
#####
```

PHASE=1 ### Phase = 0 (pseudohaploids), 1 (known phase), 2 (unknown phase)

cMMb=25 ### CentiMorgans per Megabase (if distance is not available in map file).

DIST=1 ### none (0), Haldane correction (1) or Kosambi correction (2)

NGEN=2000 ### Number of generations for which linkage data is obtained in bins

NBIN=400 ### Number of bins (e.g. if 400, each bin includes NGEN/NBIN = 2000/400 = 5 generations)

MAF=0.0 ### Minor allele frequency (0-1) (recommended 0)

ZERO=1 ### 0: Remove SNPs with zeroes (1: allow for them)

maxNCHROM=-99 ### Maximum number of chromosomes to be analysed (-99 = all chromosomes; maximum number is 200)

maxNSNP=50000 ### Maximum approx number of SNPs per chromosomes to be analysed (maximum number is 50000)

hc=0.05 ### Maximum value of c analysed (recommended 0.05; maximum is 0.5)

REPS=40 ### Number of replicates to run GONE (recommended 40)

#threads=-99 ### Number of threads (if -99 it uses all possible processors)

threads=1

Letter

Exceptionally high levels of recombination across the honey bee genome

Martin Beye,^{1,9} Irene Gattermeier,¹ Martin Hasselmann,¹ Tanja Gempe,¹ Morten Schioett,¹ John F. Baines,² David Schlipalius,³ Florence Mougel,⁴ Christine Emore,³ Olav Rueppell,⁵ Anu Sirviö,⁶ Ernesto Guzmán-Novoa,⁷ Greg Hunt,³ Udo Lüscher,¹ Michael S. Koenig,⁸

RESEARCH ARTICLE

Extreme Recombination Frequencies Shape Genome Variation and Evolution in the Honeybee, *Apis mellifera*

Andreas Wallberg¹, Sylvain Glémén^{2,3}, Matthew T. Webster^{1*}

Step 3: Perform inferences of recent N_e

less Output_Ne_seqapipop870_Ouessant_500kSNPs.withheader_FOR_IMPATIENT_PERSON_ONLY.txt

Ne averages over 40 independent estimates.

Generation Geometric_mean

1	66.7276
2	66.7276
3	66.7276
4	66.7276
5	74.7291

head -1 Output_Ne_seqapipop870_Ouessant_500kSNPs.withheader_FOR_IMPATIENT_PERSON_ONLY.txt

Ne averages over 40 independent estimates

grep -v "independent"

Output_Ne_seqapipop870_Ouessant_500kSNPs.withheader_FOR_IMPATIENT_PERSON_ONLY.txt >
Output_Ne_Ouessant.txt

Generation Geometric_mean

1	66.7276
2	66.7276
3	66.7276

Step 3: Perform inferences of recent N_e

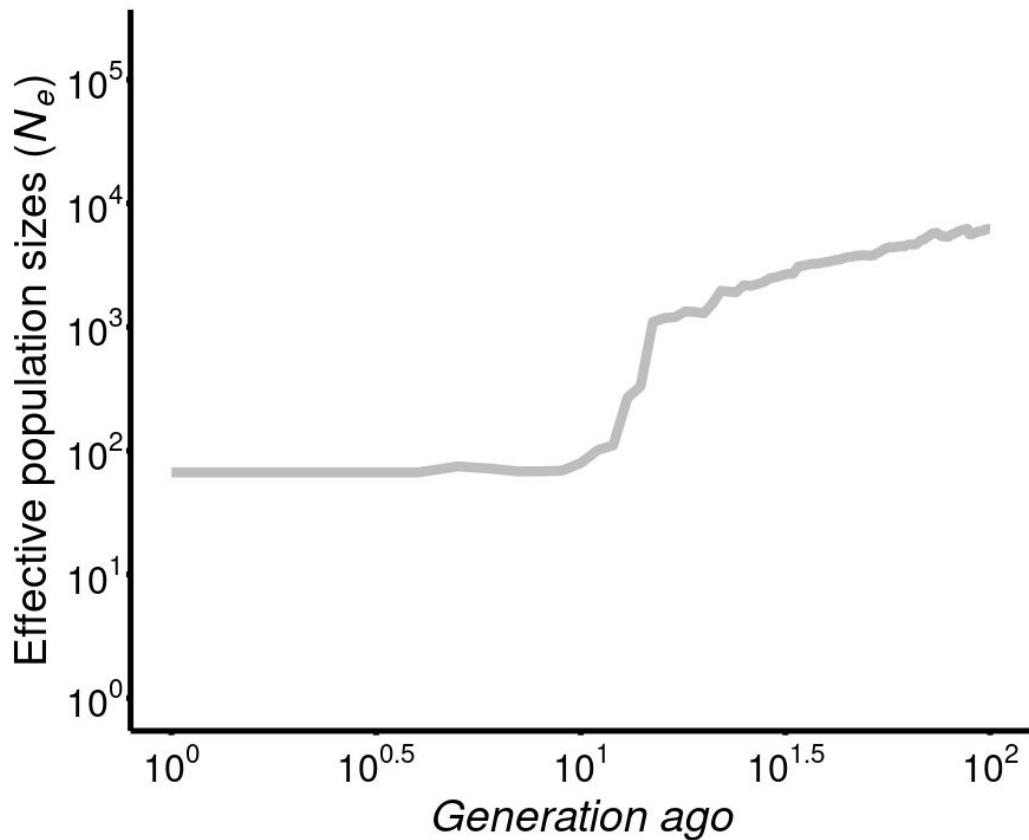
Rscript script_GONE_ggplot2.R

```
library(ggplot2)

setwd("~/Day3_demography/data-honeybees/recentNe_GONE")
GONEinferredNe=read.table("Output_Ne_Ouessant.txt",header=TRUE,sep="\t")

pdf(file="Rplot_ggplot2_Ouessant_GONE.pdf",width=10,height=10)
ggplot(GONEinferredNe, aes(x=Generation, y=Geometric_mean)) +
  geom_line(colour="grey",size=3) +
  xlab("Generation ago") + ylab(expression(paste("Effective population sizes (", italic(N[e]), ")")))) +
  scale_x_log10(limits=c(1,100),breaks = scales::trans_breaks("log10", function(x) 10^x),labels =
  scales::trans_format("log10", scales::math_format(10^.x)))+
  scale_y_log10(limits=c(1,200000),breaks = scales::trans_breaks("log10", function(x) 10^x),labels =
  scales::trans_format("log10", scales::math_format(10^.x)))+
  theme_bw()+
  theme(plot.margin = margin(1,3,1.5,1.2, "cm"))+
  theme(legend.position="none",panel.border = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))+
  theme(axis.line = element_line(colour = 'black', size = 1.75), axis.ticks = element_line(colour = 'black', size = 1.75),
  axis.text.x = element_text(colour="black",size=24,angle=0,hjust=.5,vjust=.5,face="plain"),
  axis.text.y = element_text(colour="black",size=24,angle=0,hjust=.5,vjust=.5,face="plain"),
  axis.title.x = element_text(colour="black",size=28,angle=0,hjust=.5,vjust=.2,face="italic"),
  axis.title.y = element_text(colour="black",size=28,angle=90,hjust=.5,vjust=.5,face="italic"))
dev.off()
```

Step 3: Perform inferences of recent N_e

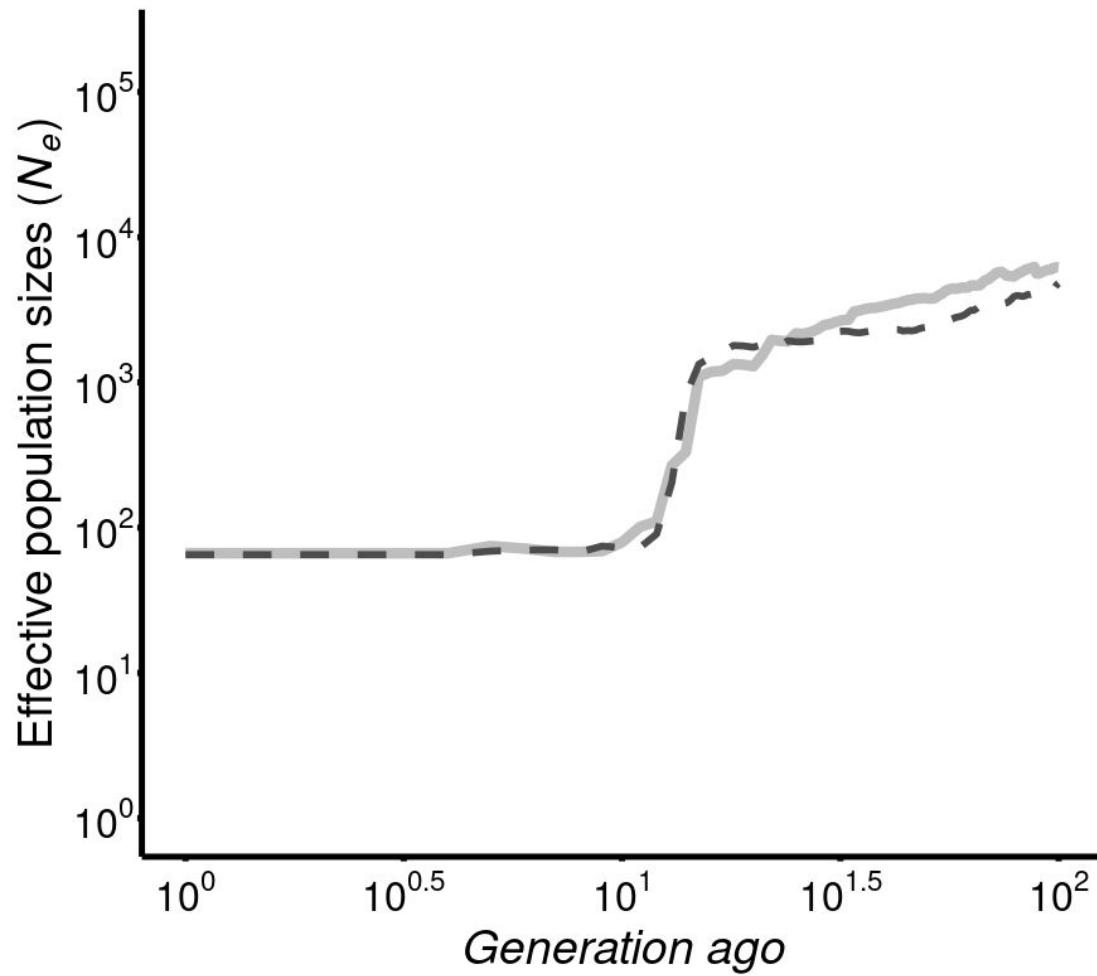


Step 3: Perform inferences of recent N_e

In case you have performed a second inference with GONE to compare you could have added a second line to have a look at the difference ;)

```
ggplot() +  
  geom_line(data=GONEinferredNe, aes(x=Generation, y=Geometric_mean), colour="grey", size=3) +  
  geom_line(data=GONEinferredNe2, aes(x=Generation, y=Geometric_mean), colour="grey30", size=2, lty=2) +  
  xlab("Generation ago") + ylab(expression(paste("Effective population sizes (", italic(N[e]), ")"))) +  
  scale_x_log10(limits=c(1,100), breaks = scales::trans_breaks("log10", function(x) 10^x), labels = scales::trans_format("log10",  
  scales::math_format(10^.x))) +  
  scale_y_log10(limits=c(1,200000), breaks = scales::trans_breaks("log10", function(x) 10^x), labels =  
  scales::trans_format("log10", scales::math_format(10^.x))) +  
  theme_bw() +  
  theme(plot.margin = margin(1,3,1.5,1.2, "cm")) +  
  theme(legend.position="none", panel.border = element_blank(), panel.grid.major = element_blank(), panel.grid.minor =  
  element_blank(), axis.line = element_line(colour = "black")) +  
  theme(axis.line = element_line(colour = 'black', size = 1.75), axis.ticks = element_line(colour = 'black', size = 1.75),  
  axis.text.x = element_text(colour="black", size=24, angle=0, hjust=.5, vjust=.5, face="plain"),  
  axis.text.y = element_text(colour="black", size=24, angle=0, hjust=.5, vjust=.5, face="plain"),  
  axis.title.x = element_text(colour="black", size=28, angle=0, hjust=.5, vjust=.2, face="italic"),  
  axis.title.y = element_text(colour="black", size=28, angle=90, hjust=.5, vjust=.5, face="italic"))
```

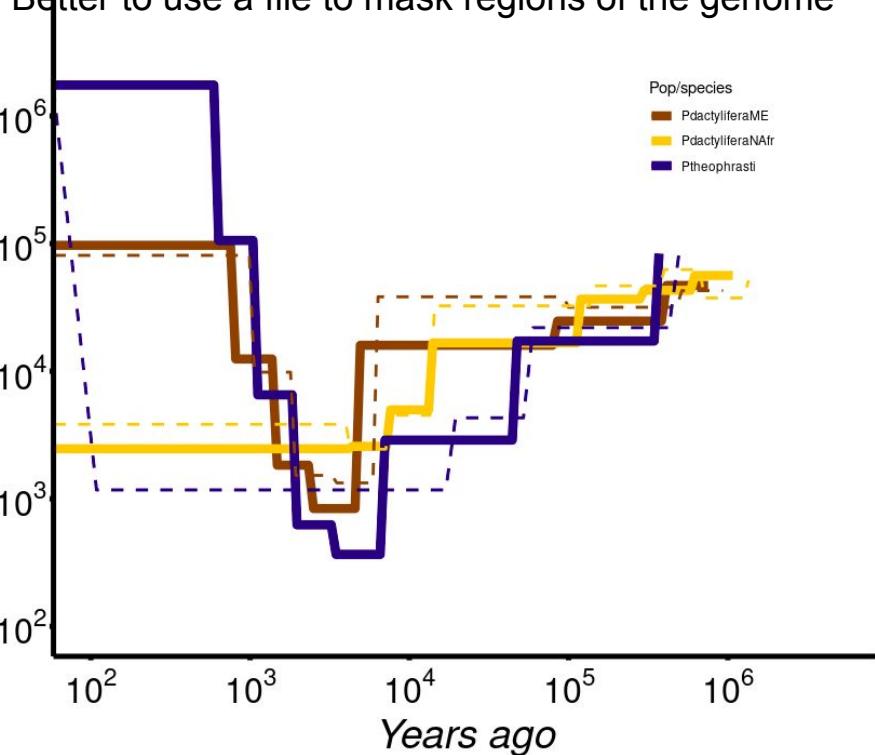
Step 3: Perform inferences of recent N_e



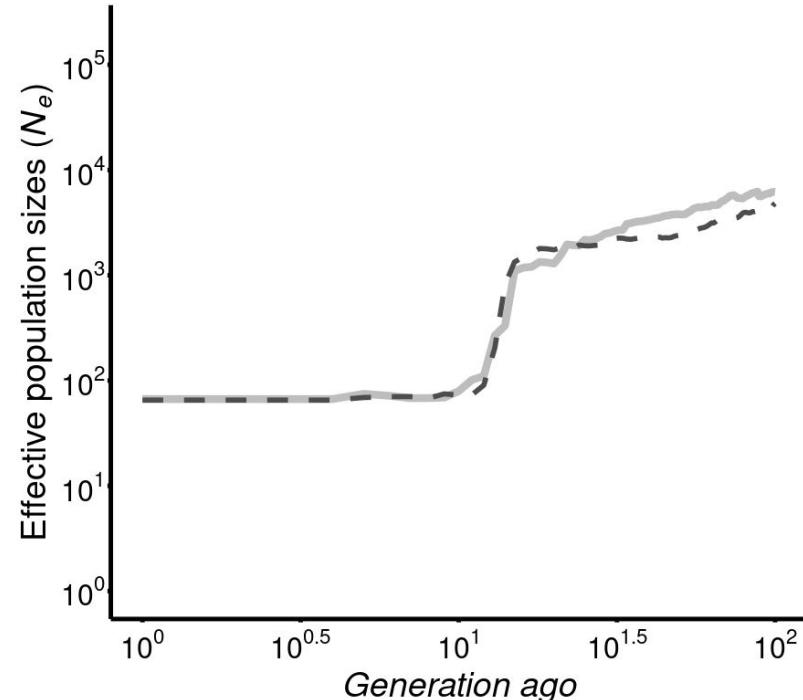
To summarize up to now

Coalescent-based approaches are good for the ancient past, but can be quite inaccurate for recent past

Better to use a file to mask regions of the genome



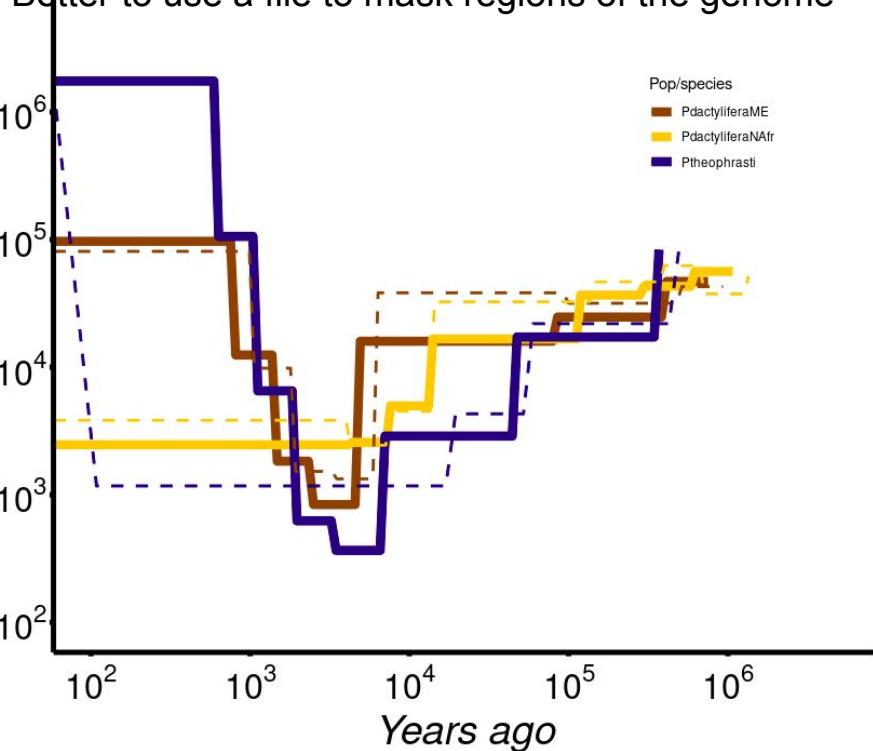
LD-based approaches are good for the recent past, but can be quite inaccurate for (a bit more) ancient past (note that here we stopped at 100 generations only)



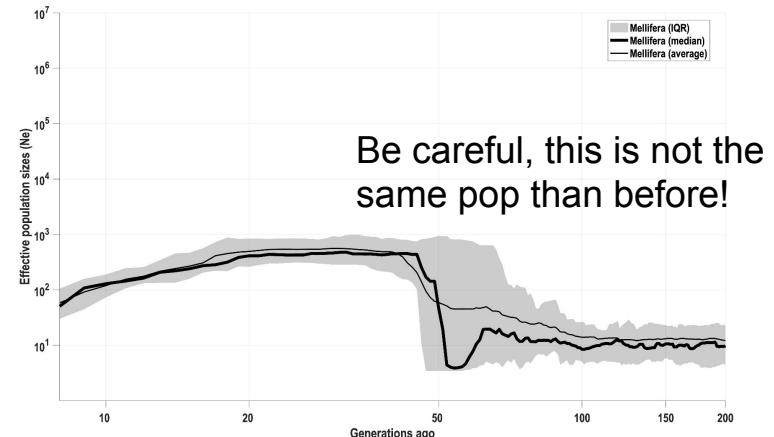
To summarize up to now

Coalescent-based approaches are good for the ancient past, but can be quite inaccurate for recent past

Better to use a file to mask regions of the genome

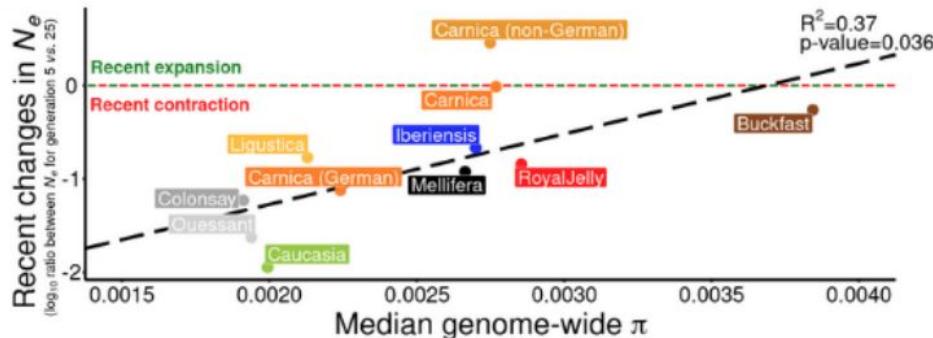
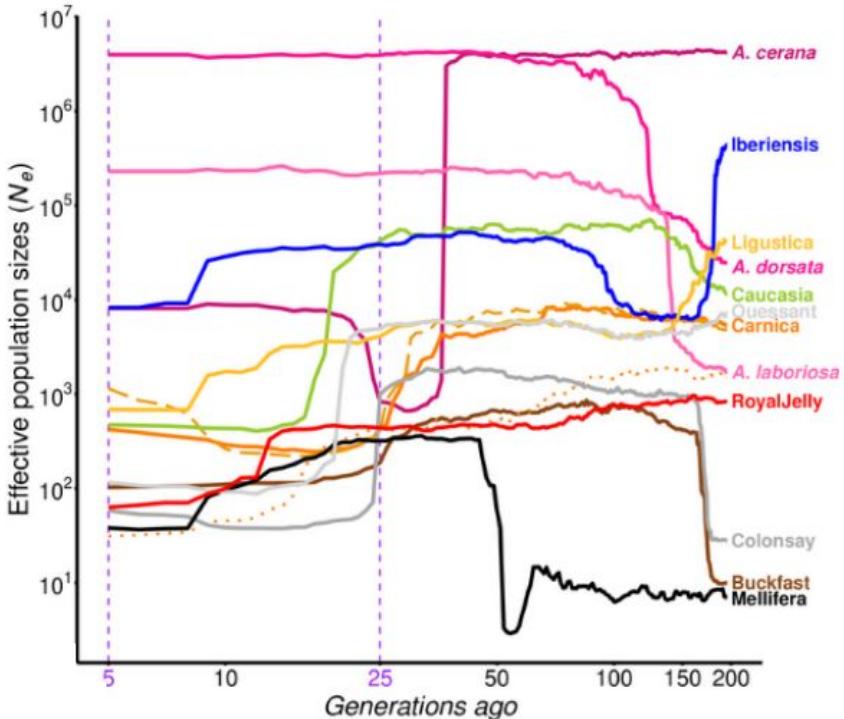


LD-based approaches are good for the recent past, but can be quite inaccurate for (a bit more) ancient past (note that here we stopped at 100 generations only)



What can we learn biologically from this recent LD pattern ?

-> Human (beekeeping) impacts



The observed variable levels of diversity is explained by the variation of the intensity of the bottlenecks among the populations

Pretty clear and strong signal for a “semi-domesticated species”

DILS

Input file:

less

Sequences_Mellifera_Carnica_outLaboriosa_randomlyselected_200_genomic_regions.head200.fasta

>NC037638_1:20100001-20200000|Mellifera|POR1|Allele1

TAAATTTAAAAATGAAATCGCACTAACACTATAACAATTTATTTACTTCGTCACTCTCTTACTANGC
AACTTATTCTACCGTTGATATTCACTATTGTAATTAAACAAAAAAACTCCATTCAATCTATTTATTTAC
AATTCGATAATATTCGATATGGTATAATAAGTAGATCGATCAAATATATTCTAACACTCACGTTTGAT
TTCTATTNTNTNTNTNTNGAAAAGTANATTATCAATANATGAANCNATACNTCNTNGAA
NTATTGNTATTAANTACATAGATTNTATTAAATCCCAATTAATNATCAATNTATGAANATNNAAATGTTG
CAATTTGATAGANGCTCGGAAAAAACTCGATGAGAGCTATATTCTTACTAATAATGATAACGTTTAA
CTGATGCTATTTAACNGATNTNNNACAAACACCAGTTAATTATGGTAGGTAGATAACATTCCA
TCTTNGCAATGTT(...)

>NC037638_1:20100001-20200000|Mellifera|POR10|Allele1

TAAATTTAAAAATGAAATCGCACTAACACTATAACAATTTATTTACTTCGTCACTCTCTTACTAGGC
AACTTATTCTACCGTTGATATTCACTATTGTAATTAAACAAAAAAACTCCATTCAATCTATTTATTTAC
AATTCGATAATATTCGATATGGTATAATAAGTAGATCGATCAAATATATTCTAACACTCACGTTTGAT
TTCTATTNTNTCTTATTGTTATCGAAAAGTAAATTATCAATACATGAACCGATATACTTCCTCGAA
ATATTGCTATTAATTACATAGATTNTATTAAATCCCAATTAATNATCAATNTATGAANATNNAAATGTTG
CAATTTGATAGANGCTCGGAAAAAACTCGATGAGAGCTATATTCTTACTAATAATGATAACGTTTAA
CTGATGCTATTTAACNGATNTNNNACAAACACCAGTTAATTATGGTAGGTAGATAACATTCCA
TCTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN(...)

Step 4: Perform relatively simple ABC demographic modelling under DILS

The screenshot shows the DILS web application interface for Step 4. On the left, a sidebar menu includes: ABC, Upload data, Data filtering, Populations/species, Prior distributions, Run ABC (selected), Results visualization, and Information.

Number of ABC analysis to run

1 to 5 ABC analyses can be performed from the same input file

1

DILS runs freely on a computer server. To avoid saturating it, we have limited the number of analyses carried out at a given time and for a given input file to **5**. Beyond 5, you will have to upload it again whenever you want. **The selected number of analysis cannot be modified** once the choices had been checked/validated

Input file upload

Browse... Sequences_Ligustica_Caucasian_outLaboriosa_20win.fasta
Upload complete

Input file format

Information extracted from the uploaded file

General information List of individuals List of populations or species

List of loci

nSpecies	nIndividuals	nLoci
3	50	20

Email address

address to receive the download link of the results

thibault.leroy@inrae.fr

This address will only be used for 2 things:
1) send the results of DILS to the user
2) contact users on the day when a collaborative meta-analysis will be considered

Genomic regions

coding non coding

Please check/validate your choices

Step 4: Perform relatively simple ABC demographic modelling under DILS

menu DILS



Welcome

ABC

Upload data

Data filtering

Populations/species

Prior distributions

Run ABC

Results visualization

Information

Maximum proportion of missing data (N, gaps, ...)



Minimum sequence length per gene

30

Minimum number of sequences per gene and per population/species

12

max_N_tolerated

-Float between 0.0 and 1.0.

-Defines the maximum proportion of N in the sequence of a gene beyond which this sequence is not considered.

Example

Please check/valid your choices

Step 4: Perform relatively simple ABC demographic modelling under DILS

menu DILS



Welcome

ABC

Upload data

Data filtering

Populations/species

Prior distributions

Run ABC

Results visualization

Information

Mutation and recombination

Mutation rate

0,000000003

Ratio r/μ

0,1

Time of demographic change

min

100

max

1000000

Population size

min

100

max

1000000

Please check/valid your choices

Step 4: Perform relatively simple ABC demographic modelling under DILS

menu DILS ≡

-  Welcome
-  ABC
-  Upload data
-  Data filtering
-  Populations/species
-  Prior distributions
-  Run ABC
-  Results visualization
-  Information


CHECKED


CHECKED


NON CHECKED


CHECKED

Step 4: Perform relatively simple ABC demographic modelling under DILS

The screenshot shows the DILS web application interface. On the left, a sidebar menu includes options like Welcome, ABC (which is selected), Upload data, Data filtering, Populations/species, Prior distributions, Run ABC (highlighted in yellow), Results visualization, and Information. The main area displays four green status boxes: 'ABC' (CHECKED), 'Upload data' (CHECKED), 'Data filtering' (CHECKED), and 'Populations/species' (CHECKED). Below these are two sections: 'Information summary' and 'Run ABC'. The 'Run ABC' section contains the message: 'No longer possible to use it directly from the cluster in Lyon!'.

But.. pretty easy to deploy on your own computer/computing cluster

See DILS Manual on github:

<https://github.com/dils-popgen/dils/blob/master/manual.pdf>

1. clone git repository

```
cd $HOME  
git clone https://github.com/popgenomics/DILS_web
```

2. move into the DILS repertory

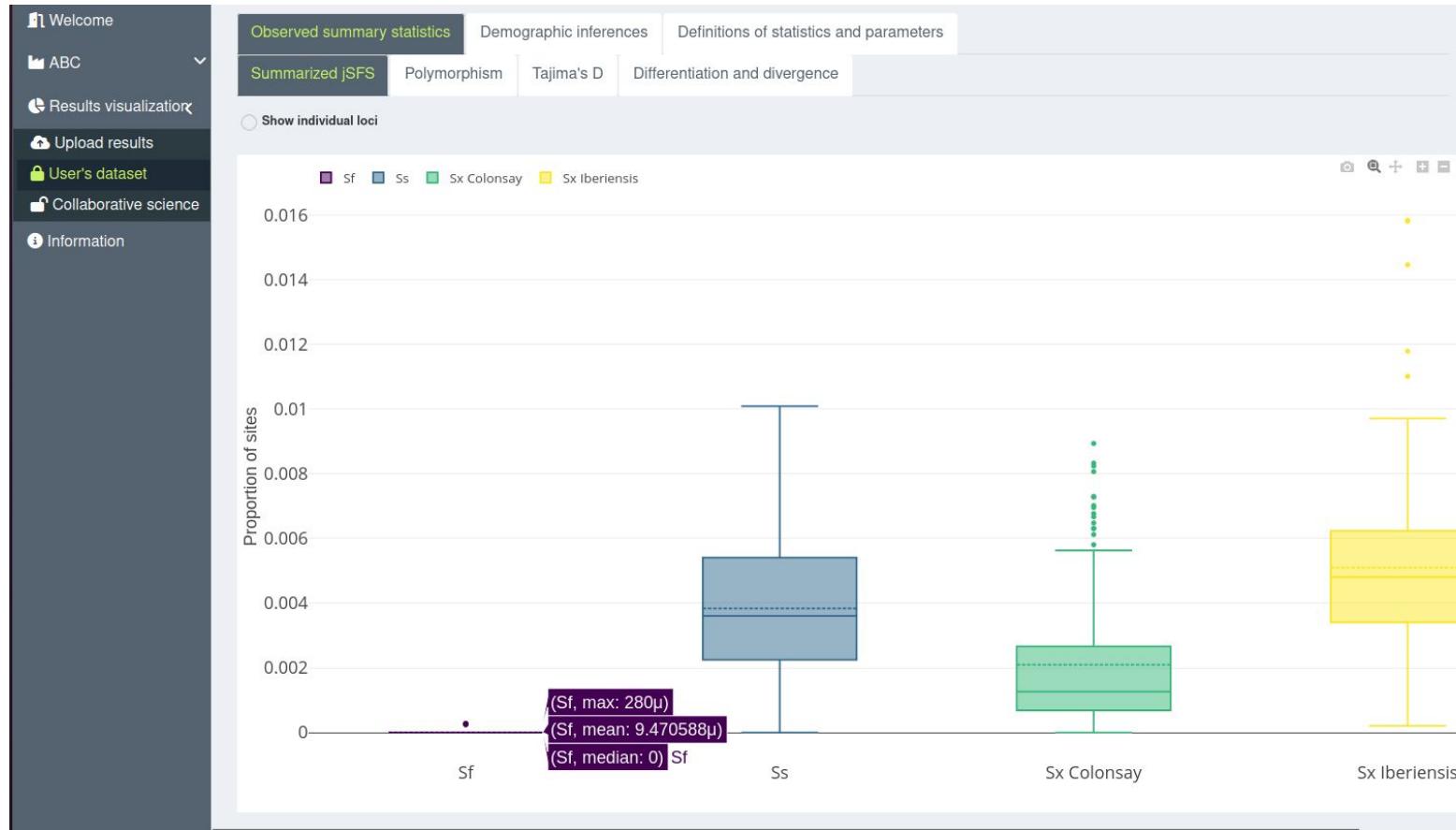
```
cd $HOME/DILS_web
```

3. build singularity image (could take ≈ 20minutes)

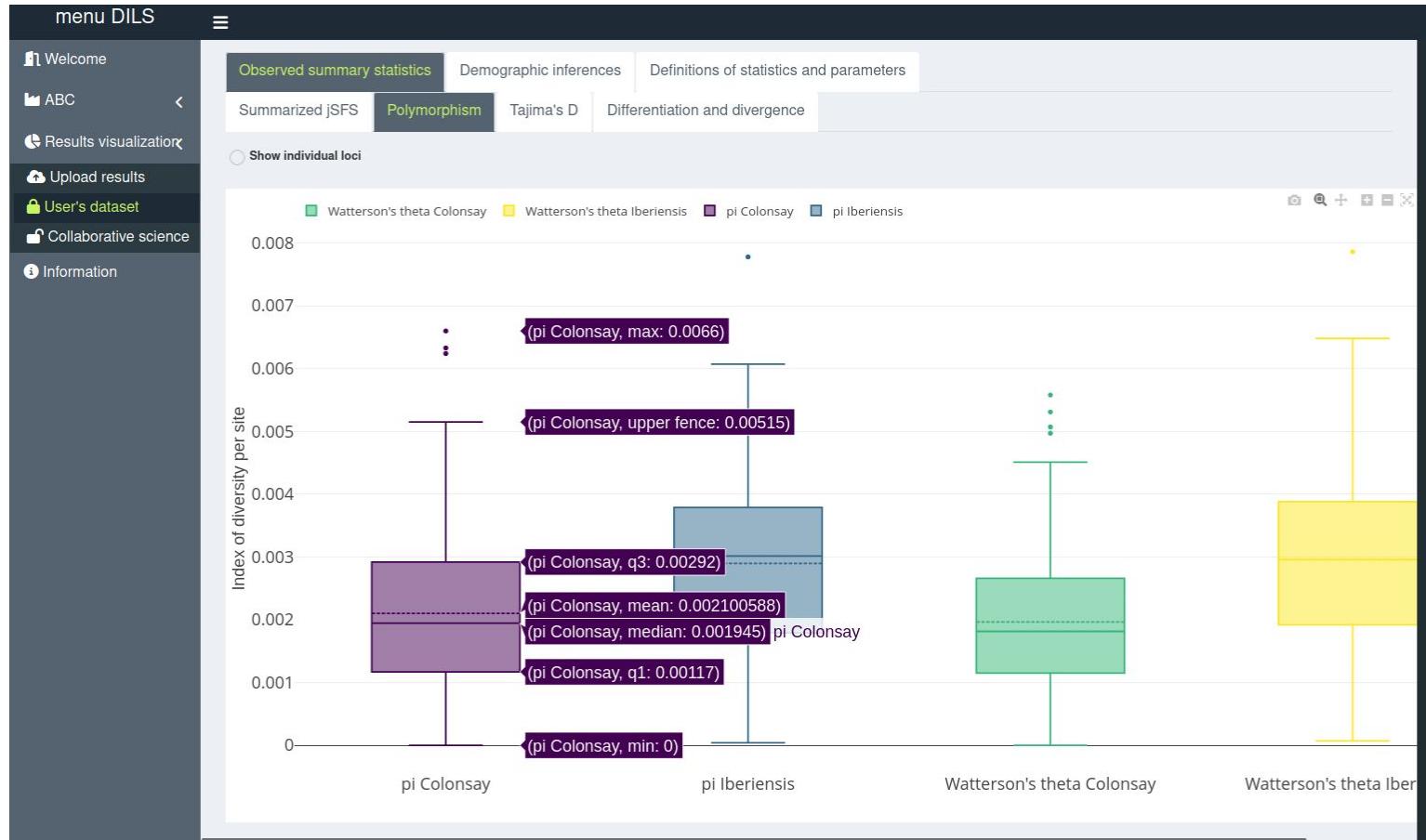
```
sudo singularity build DILS.sif DILS.def
```

Step 5: Explore results of inferences under DILS

5.1 Observed summary statistics



Step 5: Explore results of inferences under DILS



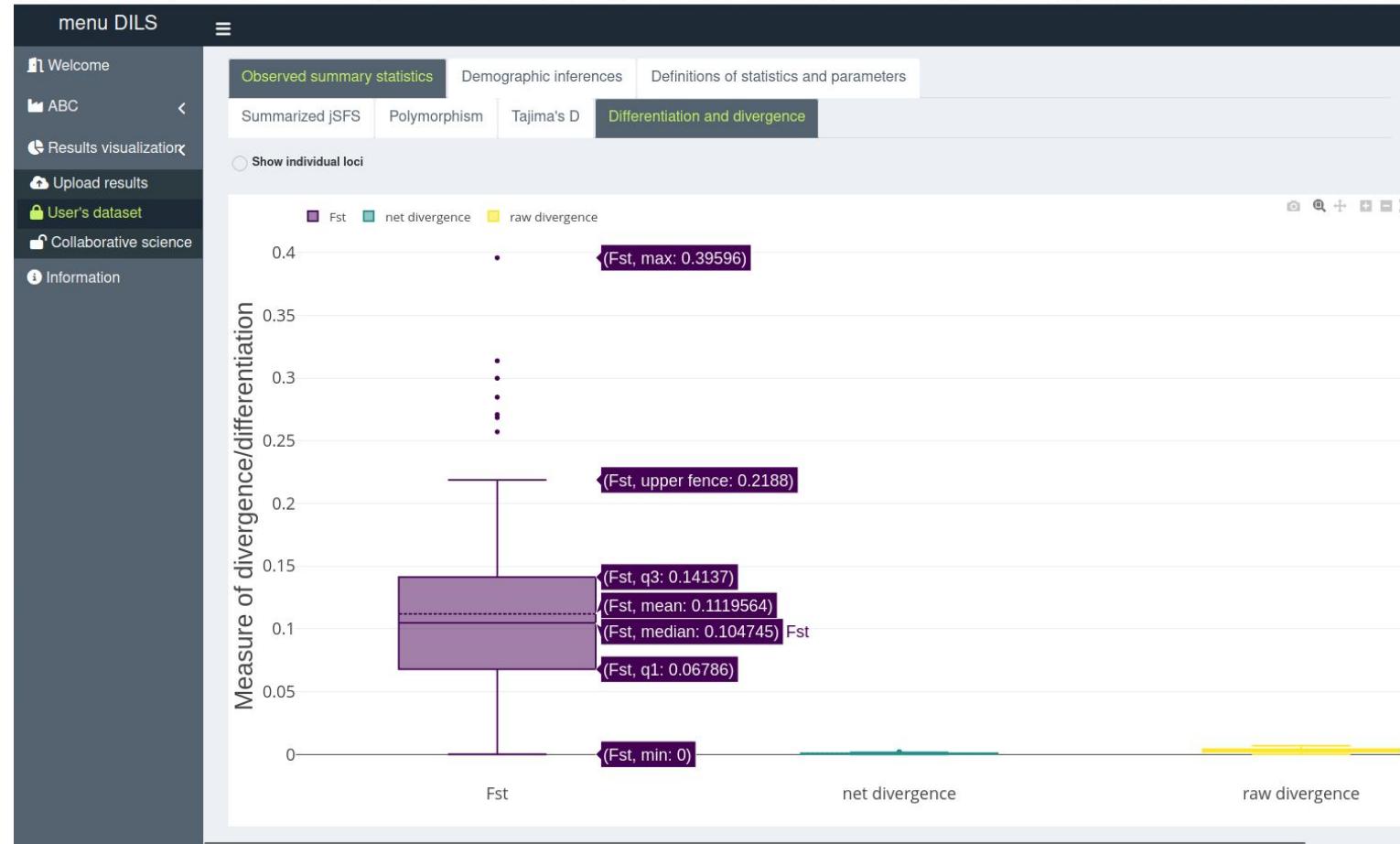
Step 5: Explore results of inferences under DILS

5.1 Observed summary statistics



Step 5: Explore results of inferences under DILS

5.1 Observed summary statistics



Step 5: Explore results of inferences under DILS

5.2 Model choices

menu DILS ≡

Welcome ABC Results visualization Upload results User's dataset Collaborative science Information

Observed summary statistics Demographic inferences Definitions of statistics and parameters

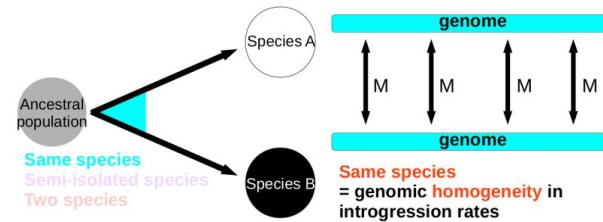
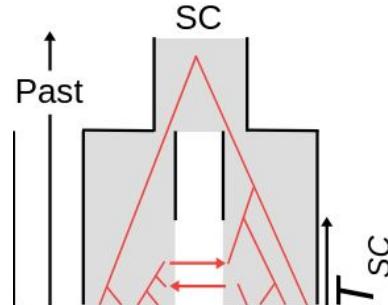
Multilocus model comparison Locus specific model comparison Estimated parameters Goodness-of-fit test

MIGRATION VERSUS ISOLATION
best model = migration
post. proba = 0.82907

IM VERSUS SC
best model = SC
post. proba = 0.74767

N-HOMO VERSUS N-HETERO
best model = Nhomo
post. proba = 0.74047

M-HOMO VERSUS M-HETERO
best model = Mhomo
post. proba = 0.81253



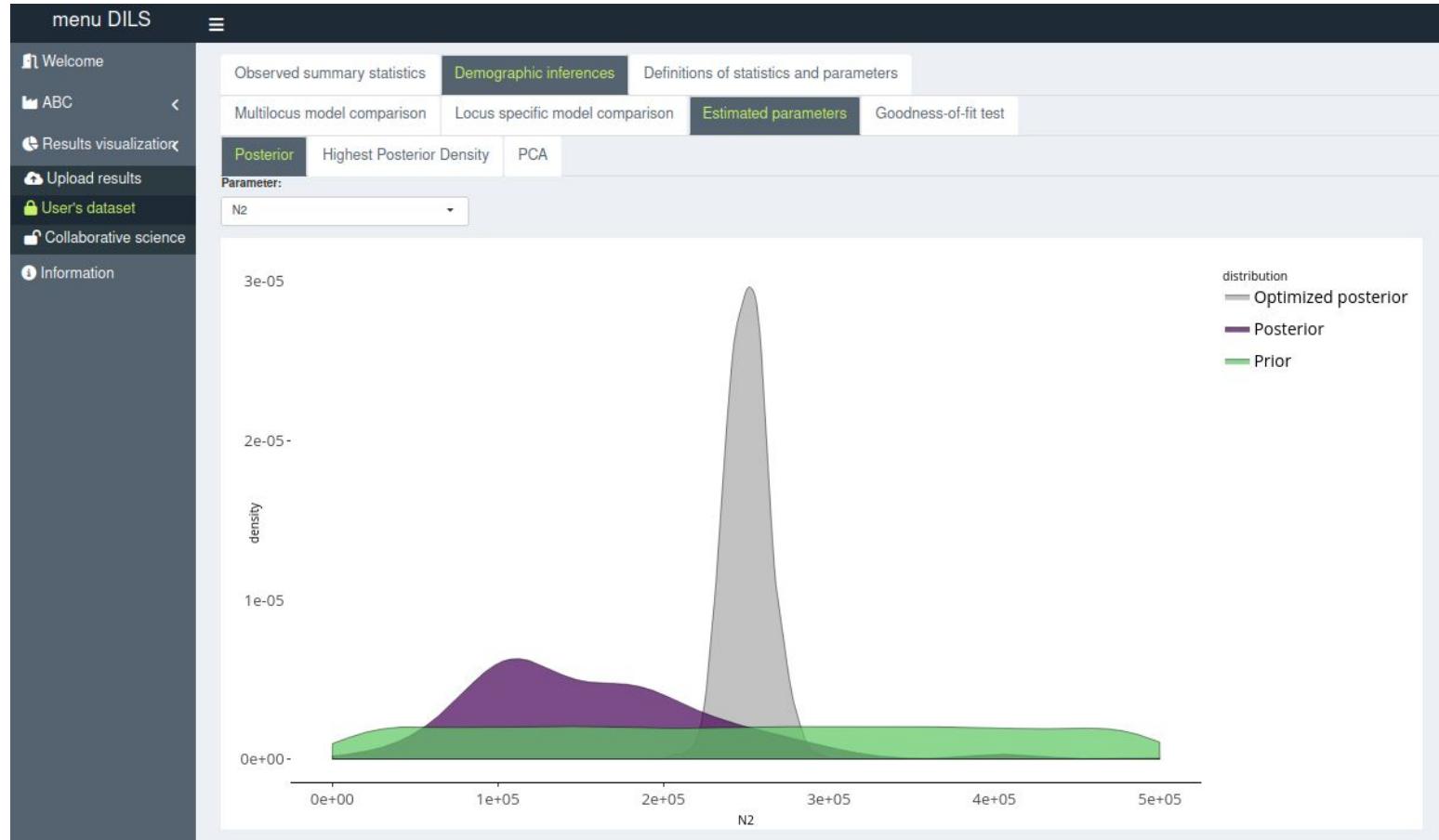
Step 5: Explore results of inferences under DILS

5.2 Model choices



Step 5: Explore results of inferences under DILS

5.3 Estimated parameters8



Step 5: Explore results of inferences under DILS

5.3 Estimated parameters

menu DILS



Welcome

ABC



Results visualization

Upload results

User's dataset

Collaborative science

Information

Observed summary statistics

Demographic inferences

Definitions of statistics and parameters

Multilocus model comparison

Locus specific model comparison

Estimated parameters

Goodness-of-fit test

Posterior

Highest Posterior Density

PCA

prediction method

neural network

ABC - neural network

Parameter	HPD 0.025	HPD median (posterior)	HPD 0.975
N1	20 626	44 785	87 034
N2	50 697	145 966	306 538
Na	48 928	98 871	135 360
Tsplit	55 134	176 831	403 102
Tsc	-2 071	21 098	71 633
M12	2.88284	5.80824	8.63876
M21	5.62817	8.21084	9.62832

Step 5: Explore results of inferences under DILS

5.4 Model checking



Does the PCA suggest that simulations (or a subset of them) reproduce the observed summary statistics?

According to you, why might some gray points (not used for the posterior) be closer to the yellow dot than some purple points in the PCA?

Step 5: Explore results of inferences under DILS

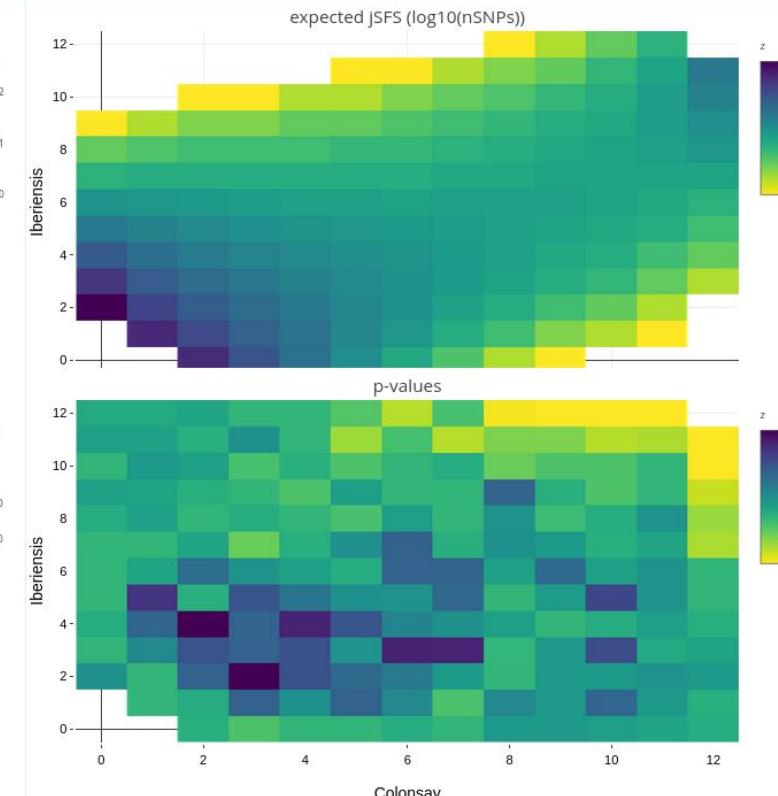
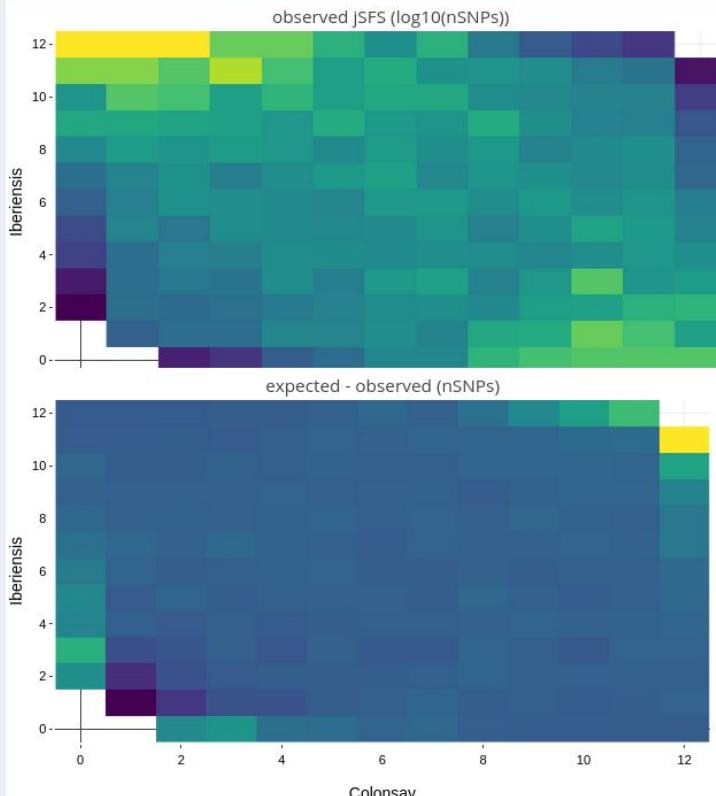
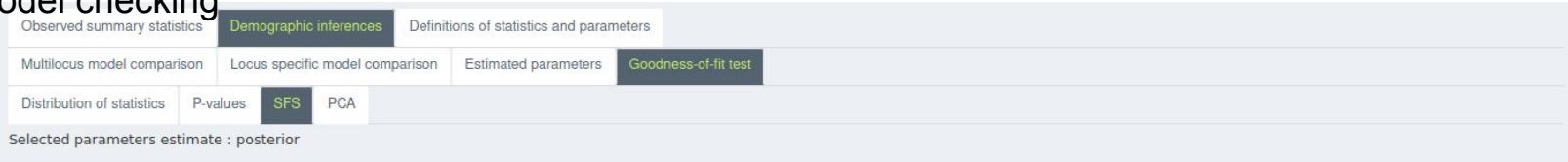
5.4 Model checking

	stats	mean_exp	mean_obs	pvals_fdr_corrected
1	bialsites_avg	27.5426	44.05882	0.0026
2	bialsites_std	13.40905	21.66556	0
3	sf_avg	0.00002	0.00001	0.13752
4	sf_std	0.00004	0.00005	0.17648
5	sxA_avg	0.00131	0.0021	0.06651
6	sxA_std	0.00102	0.00209	0.0273
7	sxB_avg	0.00283	0.0051	0.04534
8	sxB_std	0.00185	0.00256	0.12146
9	ss_avg	0.00295	0.00384	0.14349
10	ss_std	0.00202	0.00228	0.27243
11	successive_ss_avg	2.81251	2.62353	0.3822
12	successive_ss_std	2.4242	1.9065	0.1876
13	pIA_avg	0.00138	0.0021	0.0286
14	pIA_std	0.00089	0.00127	0.0294
15	pIB_avg	0.00183	0.0029	0
16	pIB_std	0.001	0.00142	0.02779
17	pearson_r_pi	0.72069	0.63213	0.19009
18	thetaA_avg	0.00141	0.00197	0.0572
19	thetaA_std	0.00083	0.00115	0.0286
20	thetaB_avg	0.00191	0.00296	0.0273
21	thetaB_std	0.00096	0.00137	0.0286
22	pearson_r_theta	0.74393	0.60166	0.12146
23	DtsjA_avg	-0.09708	0.29495	0.04534
24	DtsjA_std	0.9019	0.80635	0.19009
25	DtsjB_avg	-0.20887	-0.17624	0.43126
26	DtsjB_std	0.71777	0.51259	0.01463
27	divAB_avg	0.00188	0.0031	0.0300
28	divAB_std	0.00104	0.0015	0.13752
29	netdivAB_avg	0.00028	0.0006	0.13752
30	netdivAB_std	0.00036	0.00042	0.2733
31	FST_avg	0.07732	0.11196	0.19953
32	FST_std	0.07946	0.06275	0.31157
33	pearson_r_divAB_netDivAB	0.51824	0.74446	0.14502

How many summary statistics are well captured versus not well captured for the ABC analysis you have selected?

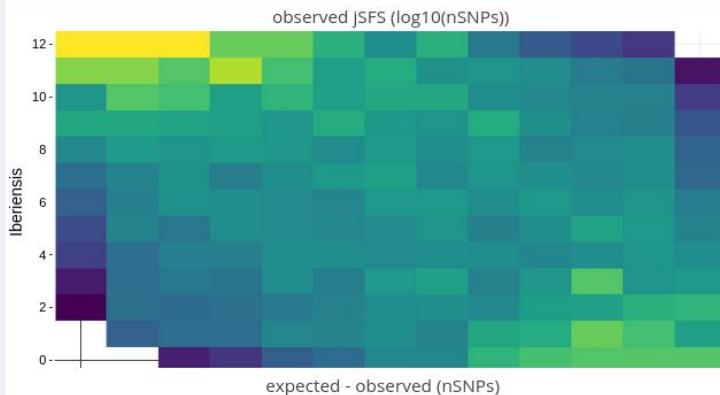
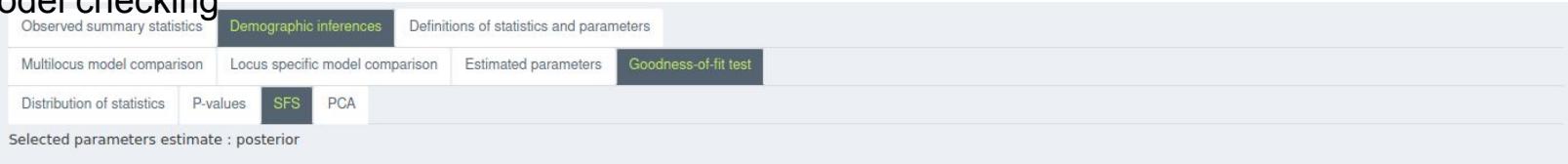
Step 5: Explore results of inferences under DILS

5.4 Model checking



Step 5: Explore results of inferences under DILS

5.4 Model checking



Understand the SFS. Why is there missing information in certain corners of the SFS (e.g. $x=0$ & $y=0$)?
-> **It is no longer polymorphic! We only work on polymorphic sites!**

How well does the jSFS align with the simulations? **Correct** Are many bins in the SFS associated with significant p-values? **More or less yes**
Overall confidence: limited

Overall, you have probably seen relatively good fits to the data

"...all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind ..." — George Box —