Physalia

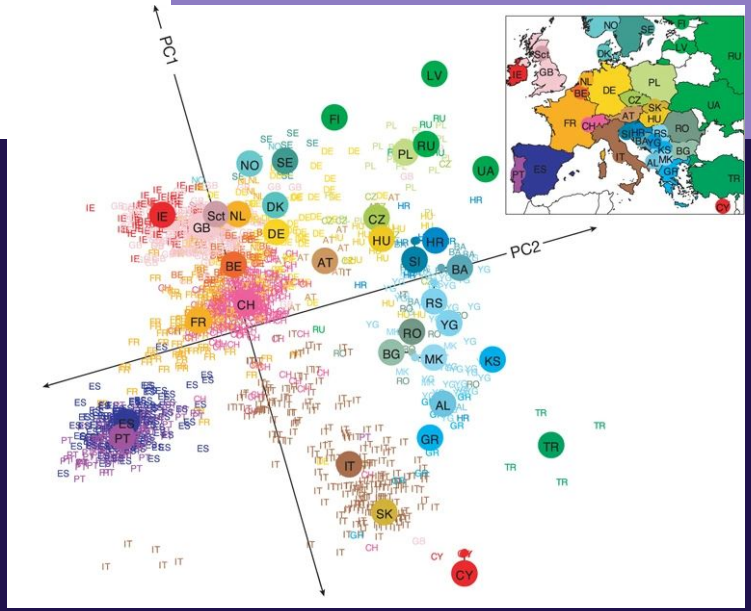
Courses

(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:AAATACCAACAAC
2:AAATACCATCAAC
3:AAATACCATCAAG
4:AAATACCATCAAC
5:AAATACCATCGAC

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)

~1x10^{-3}

Leffler et al. Plos Biol 2012

(autosomal nucleotide diversity)

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)

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

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?

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***

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

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

At equilibrium $\theta = \pi$

Leffler et al. Plos Biol 2012

(autosomal nucleotide diversity)

Current genetic diversity = long-term processes and past history

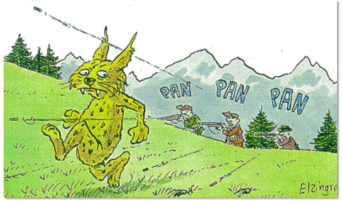
Can we understand why some species are more genetically diverse than some others?

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
~2.0x10^{-4}

Lynx pardinus (Iberian)
~1.0x10^{-4}

Leffler et al. Plos Biol 2012

(autosomal nucleotide diversity)

Current genetic diversity = long-term processes and past history

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

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

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

1:AAATACCAACAAC
2:AAATACCATCAAC
3:AAATACCATCAAG
4:AAATACCATCAAC
5:AAATACCATCGAC

θ = # polym. sites/harmonic num.

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

θ = 3 / 2.083 = 1.44
(0.11 per bp)

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

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

D>0: Deficit of rare alleles = population contraction

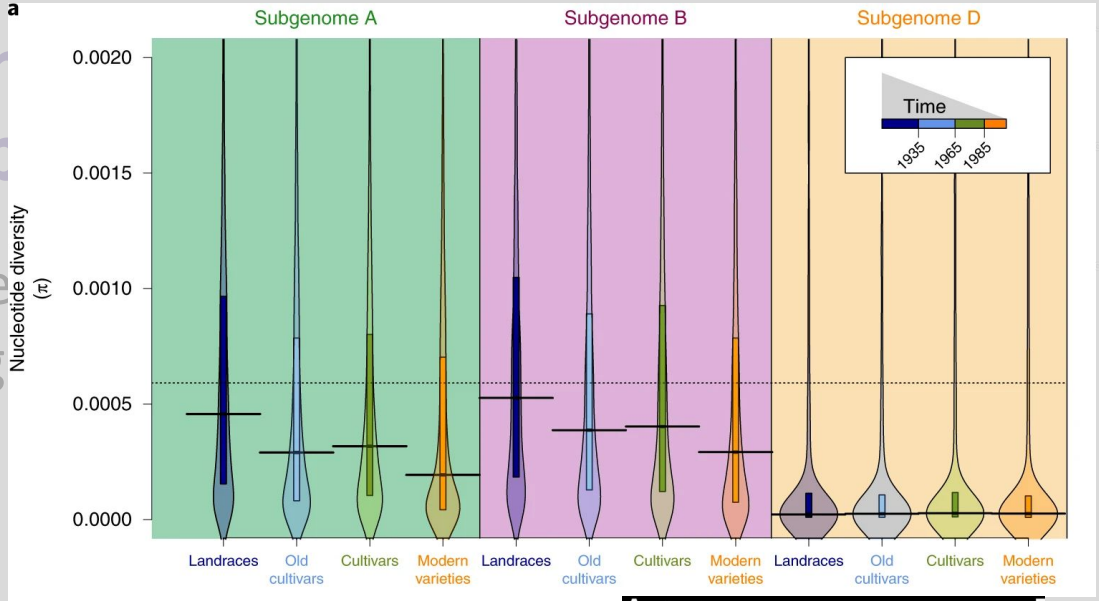
D<0: Excess of rare alleles = population expansion

Leffler et al. Plos Biol 2012

(autosomal nucleotide diversity)

Simple but often very meaningful summary statistics!

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



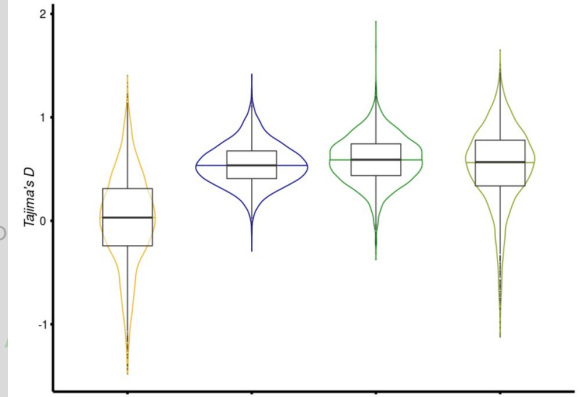
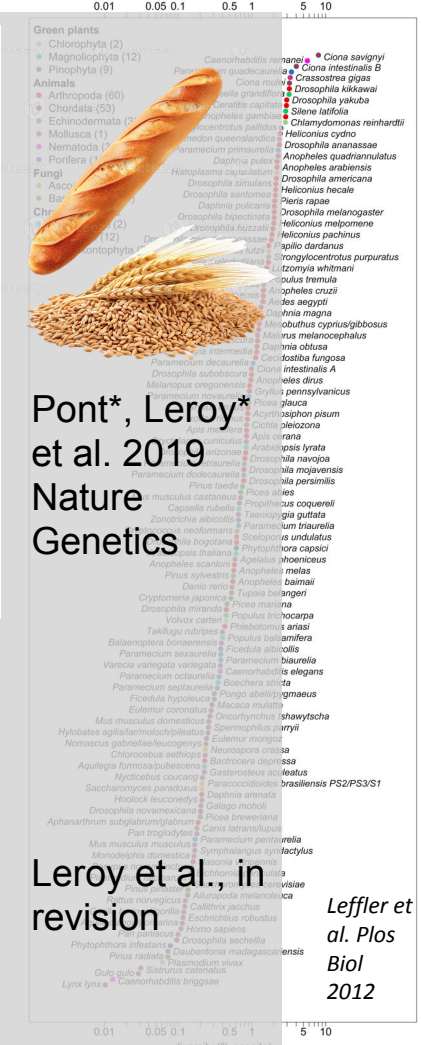


Figure S13: Distributions of the Tajima's D estimates across the four focal groups based on all 100-kb sliding windows spanning the genome. Ancient Asian, ancient European, early Asian x European and hybrid tea roses are shown in yellow, blue, dark green and khaki,



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?

Deviations from mutation-drift equilibri
(Genom
deviatio
Tajima's

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

1:AAATACCAACAAC

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

→ Demographic modelling

lleles
= population contraction

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

θ = 3 / 2.083 = 1.44
(0.11 per bp)

Tajima's D~0

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

D<0: Excess of rare alleles
= population expansion

Leffler et al. Plos Biol 2012

(autosomal nucleotide diversity)

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 (*Nc*) : the number of individuals in a population that you can observe

Effective population size: a crucial parameter

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

≠

Effective population size (*Ne*): 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

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

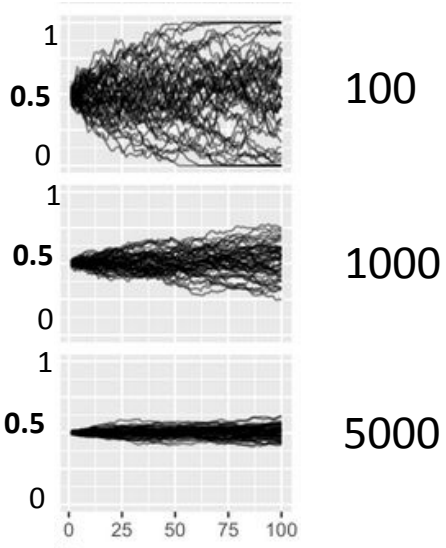
Effective population size: a crucial parameter

Effective population size (*Ne*): 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*



100

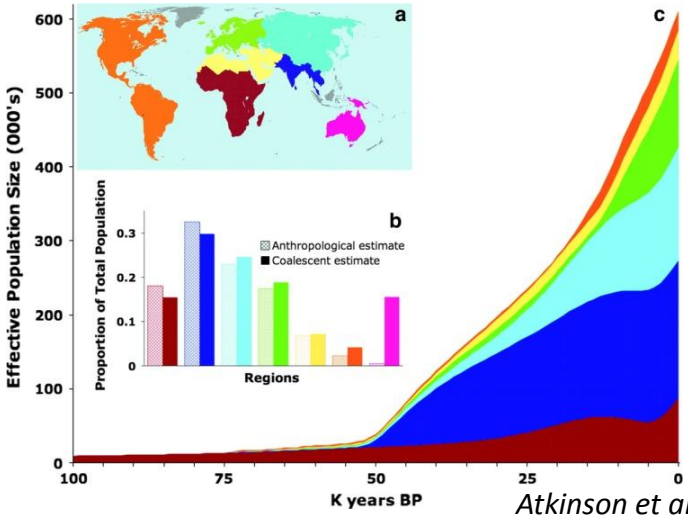
1000

5000

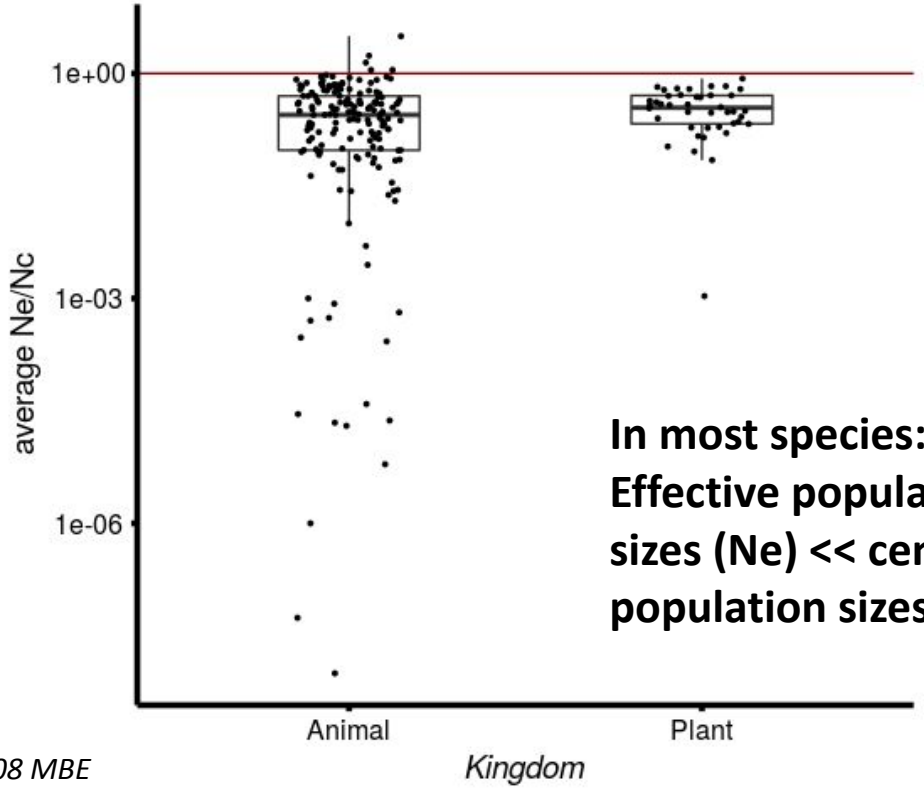
WF: a model for the allele frequency dynamics

Effective population size: a crucial parameter

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



Atkinson et al. 2008 MBE



In most species: Effective population sizes (Ne) << census population sizes (Nc)

Data from Hoban et al. 2020 BiolCons

Introducing the coalescent theory through N_e

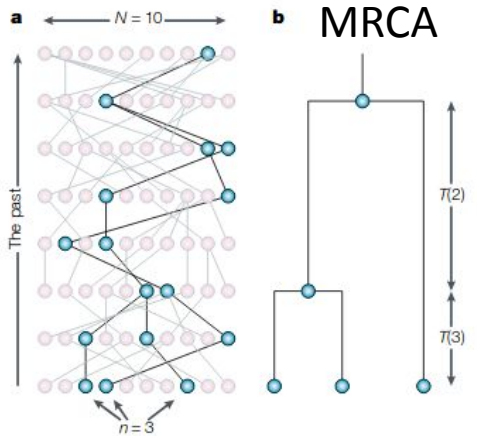
Coalescence: a stochastic process that describes how population genetic processes determine the shape of the genealogy of sampled gene sequences

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



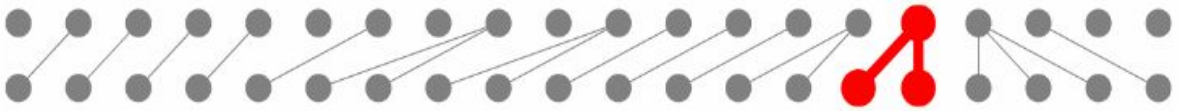
e.g. Rosenberg & Nordborg, 2002

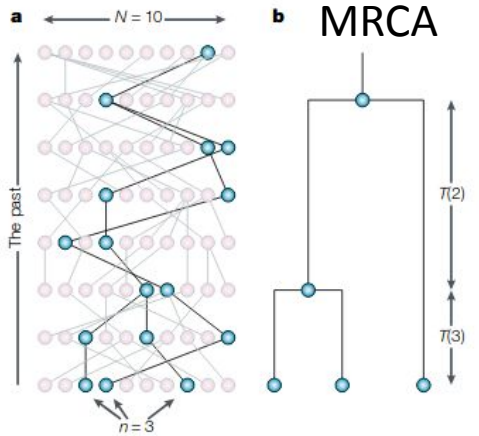
Introducing the coalescent theory through N_e

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



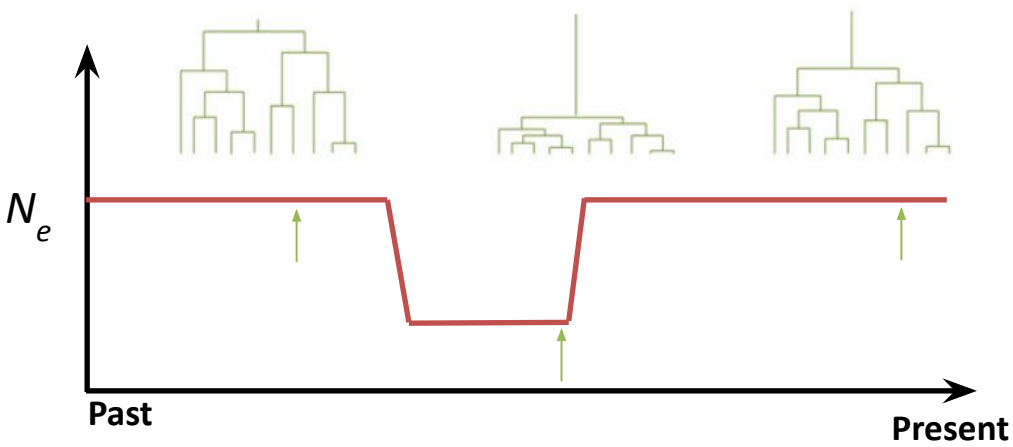


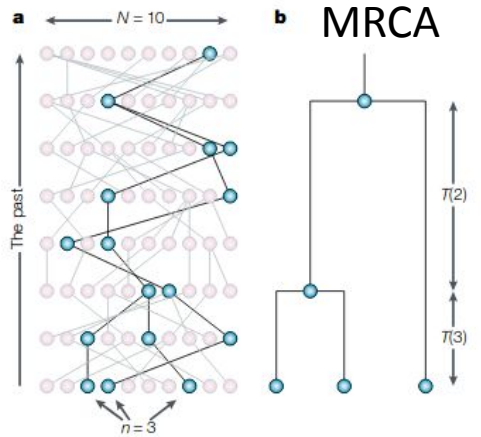
e.g. Rosenberg & Nordborg, 2002

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





e.g. Rosenberg & Nordborg, 2002

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)

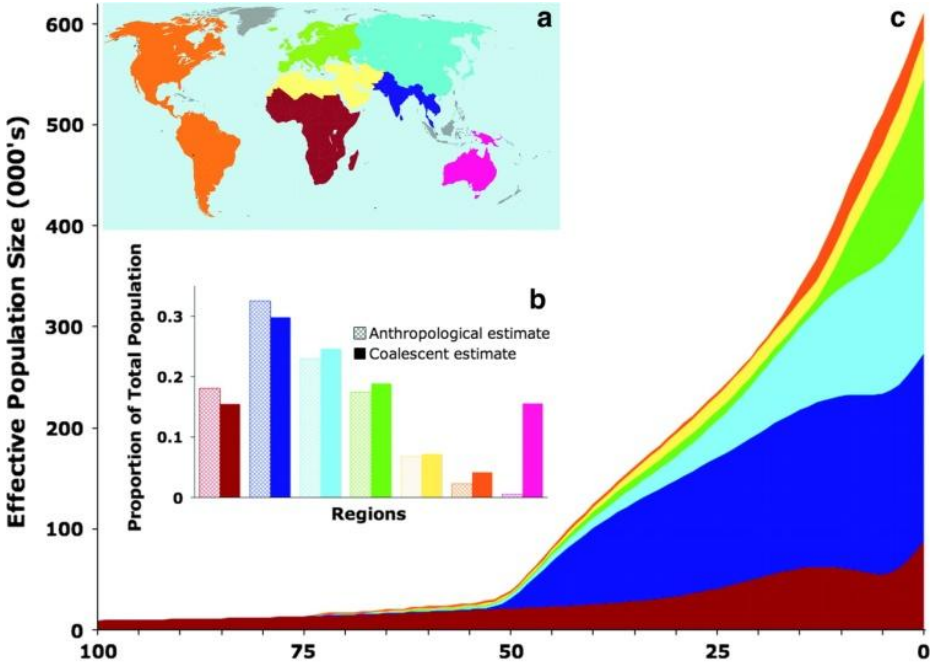
⟶ Maximum likelihood estimates of parameters (pop expansions/bottlenecks)

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?



(Atkinson et al. 2008 MBE)

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

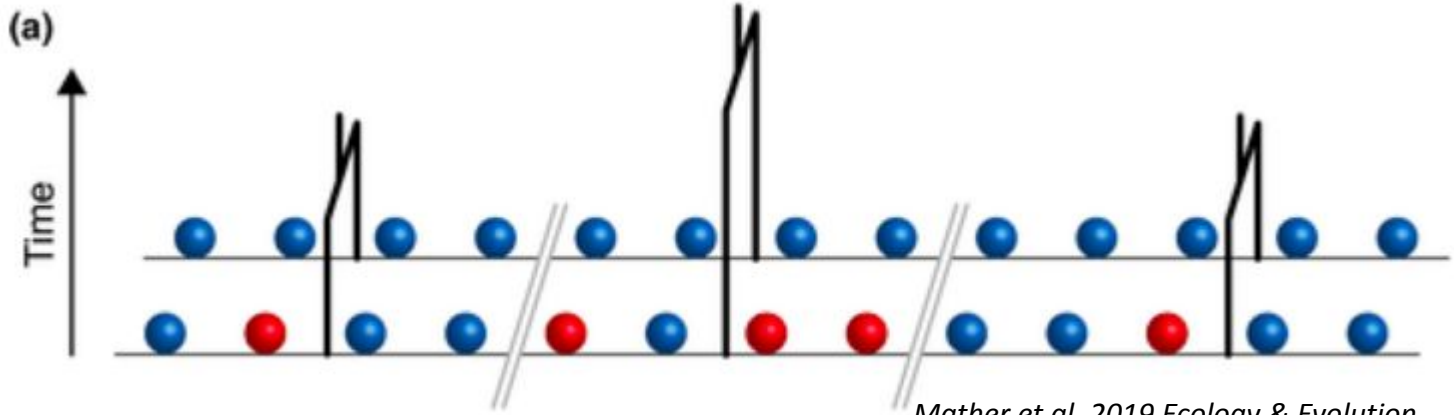
e.g. McVean & Cardin, 2005
SMC (sequential Markov coalescent)

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

TMRCA between the two alleles carried by an individual

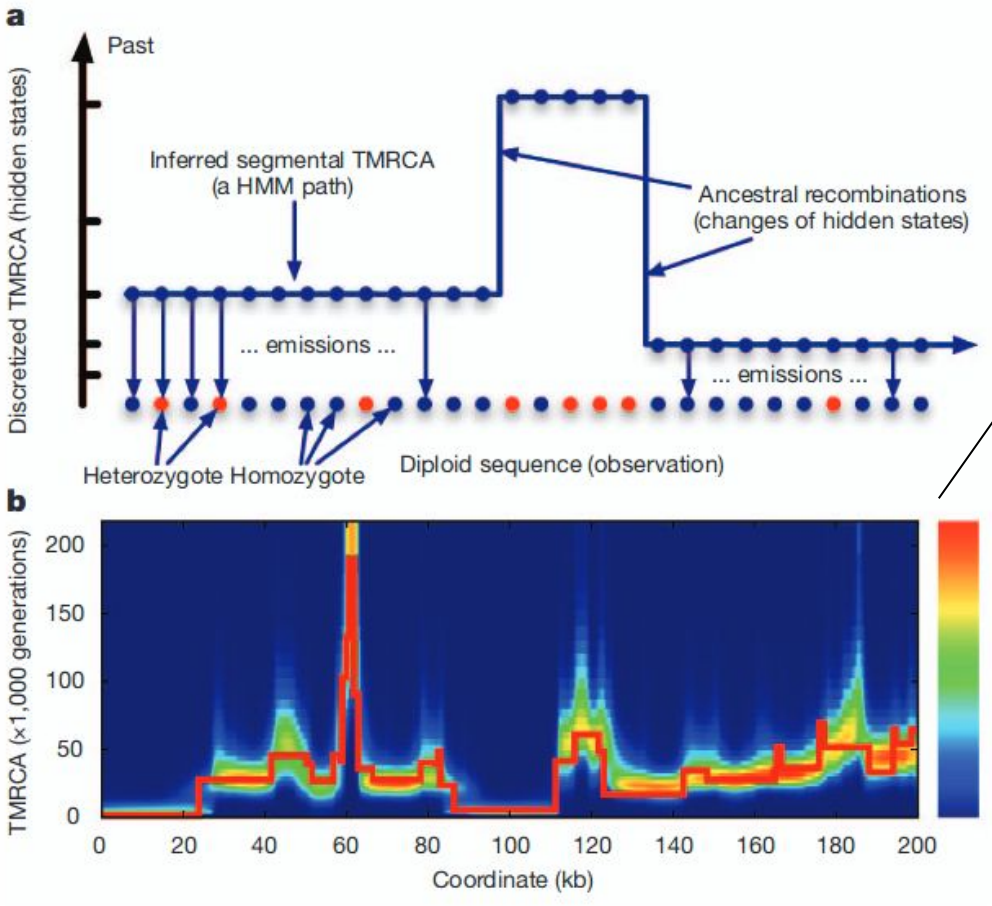


Mather et al. 2019 Ecology & Evolution

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)

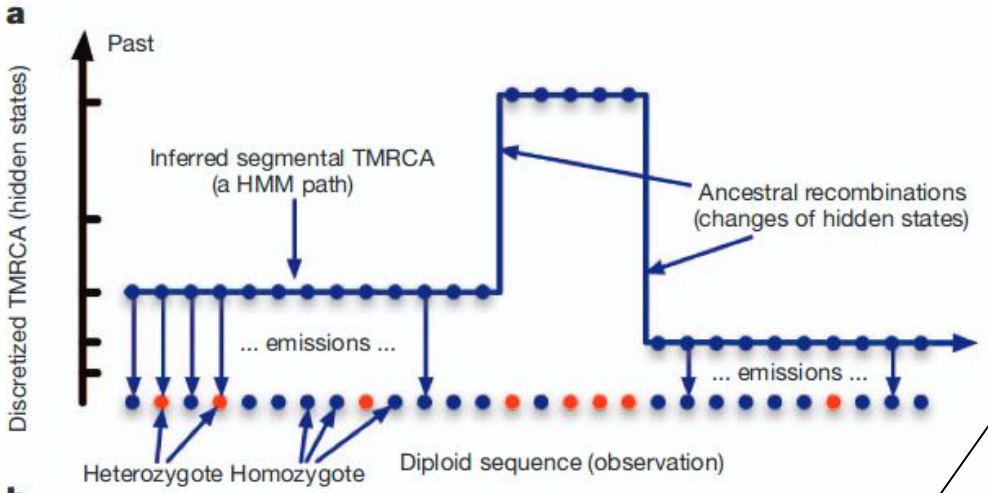
TMRCA between the two alleles carried by an individual

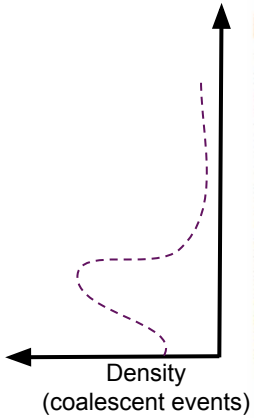


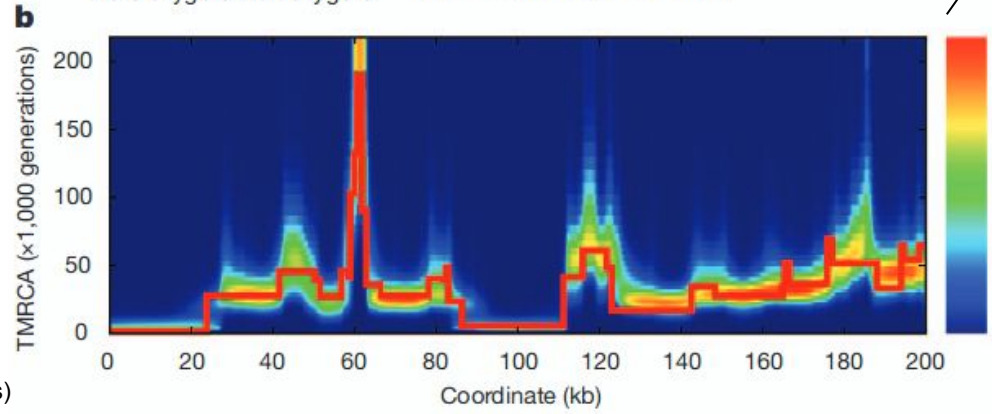
Estimating TMRCA of the two alleles at each locus is used to create a **TMRCA distribution across the whole genome**.

Pairwise Sequentially Markovian Coalescent (PSMC)

TMRCA between the two alleles carried by an individual

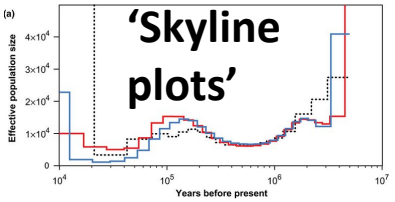






Estimating TMRCA of the two alleles at each locus is used to create a **TMRCA distribution across the whole genome**.

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.**

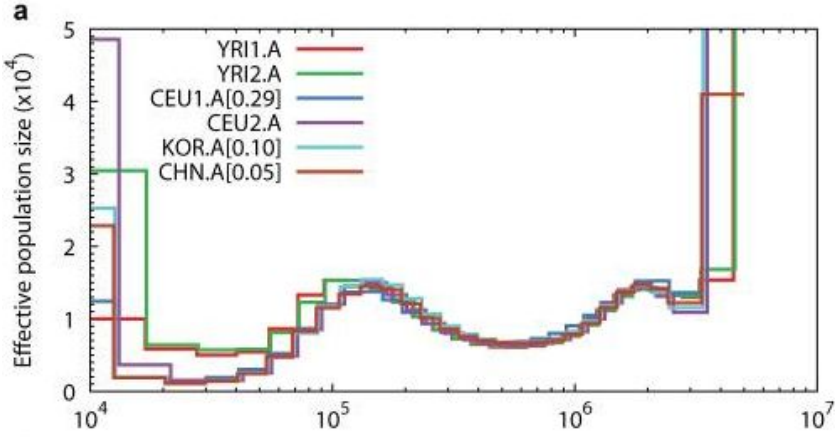


'Skyline plots'

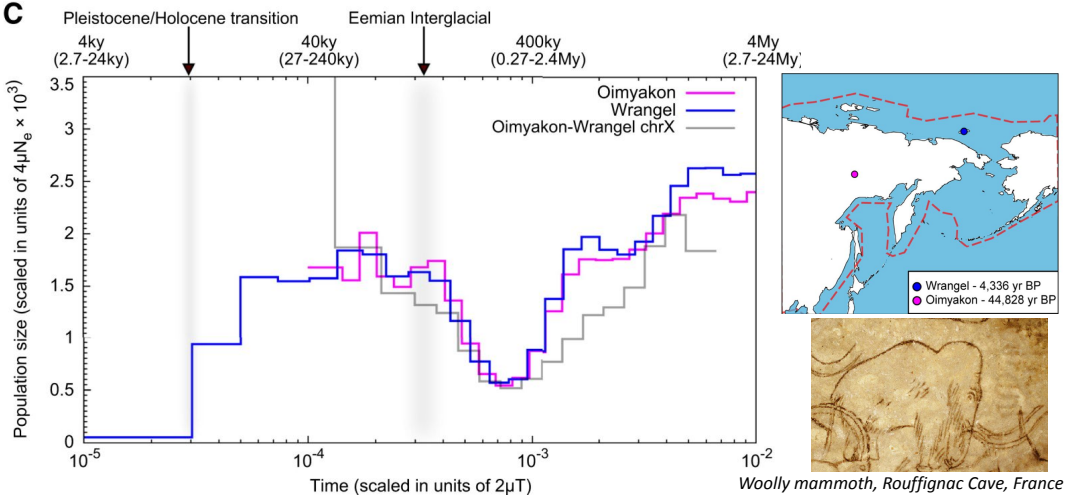
Pairwise Sequentially Markovian Coalescent (PSMC)

Example of applications:

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



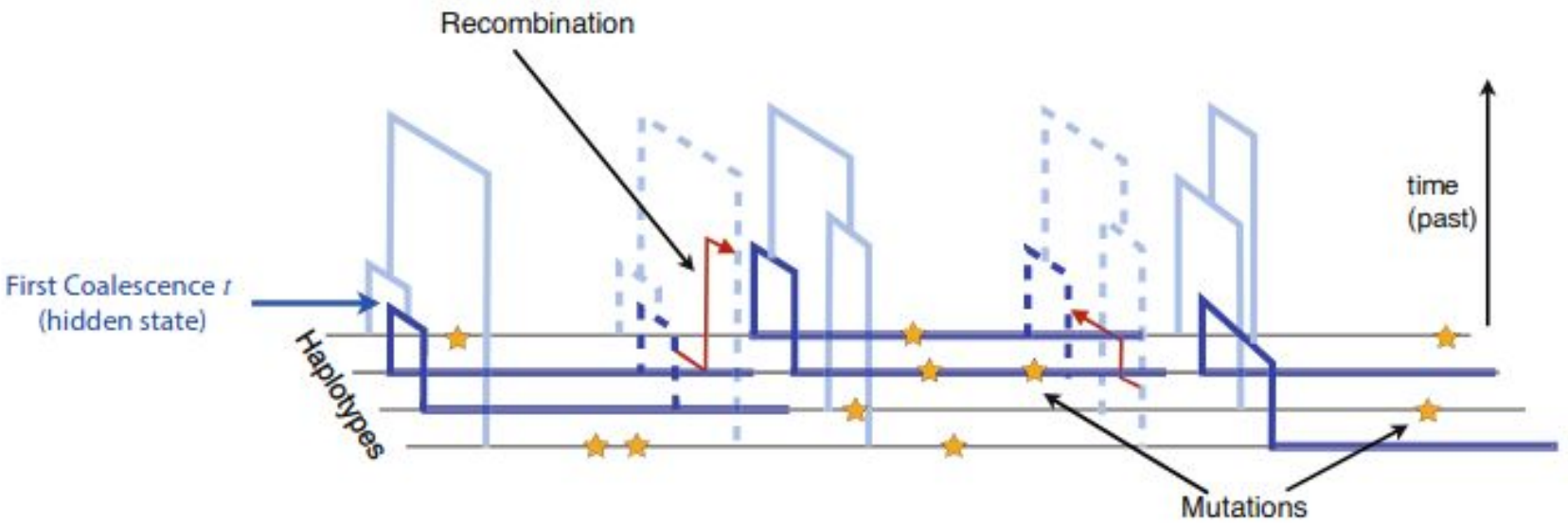
Palkopoulou et al. 2015, Current Biol



Woolly mammoth, Rouffignac Cave, France

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.

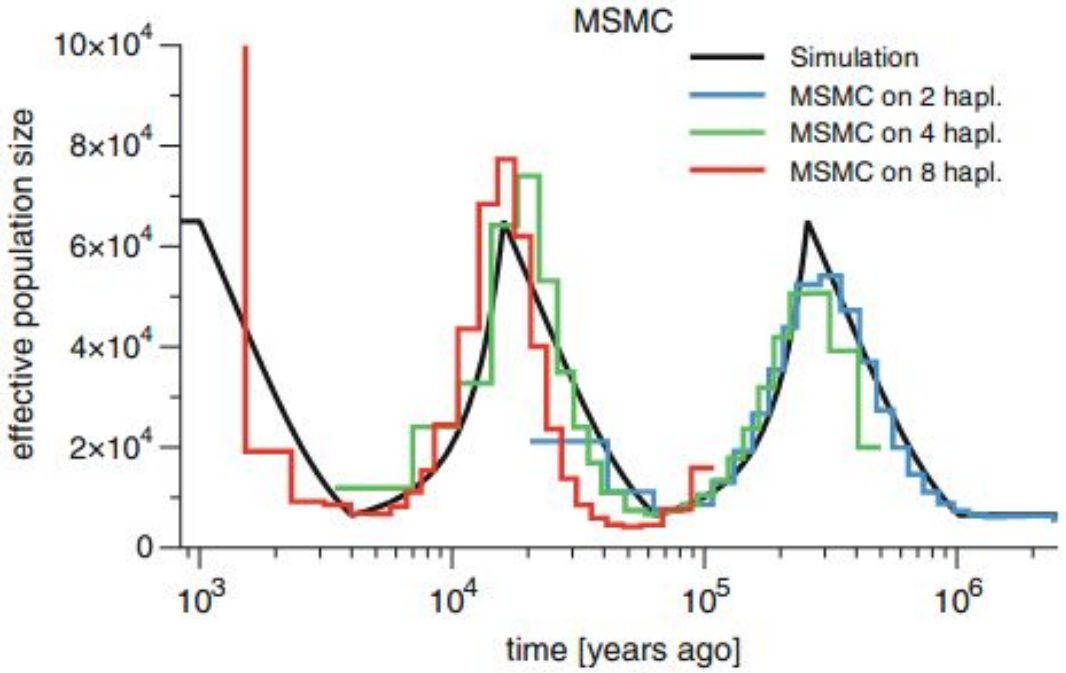


Mather et al. 2019 Ecology & Evolution; Schiffels & Wang, 2021 Springer Protocols

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)



Schiffels & Wang, 2021 Springer Protocols

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

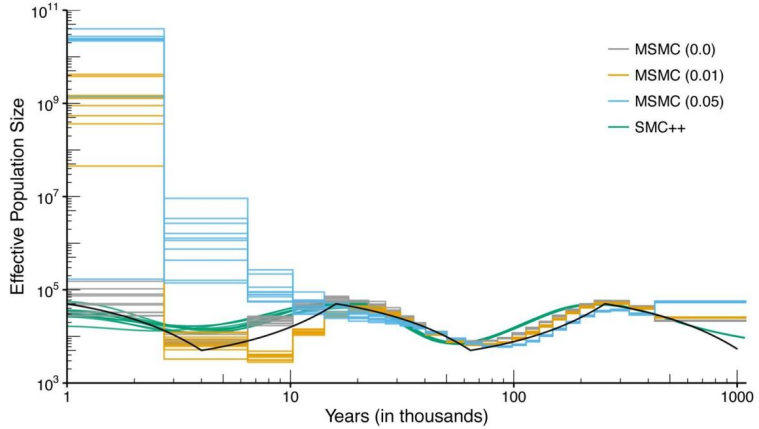
A	T G	G	C G	A G	G

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.

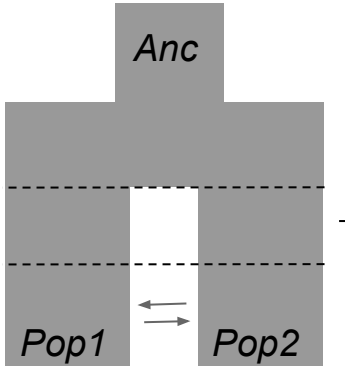
PSMC-like methods: pros and cons

Advantages:

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

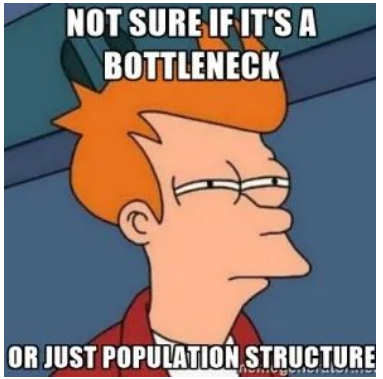
Limitations (1/2):

- Simplistic approach (assumes a panmictic population, *i.e.* drift-only)
  => change in *Ne* 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 artifactually considered as a period of low N_e



molecularecologist.com

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 (>5-10 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 frequency (MAF)	1/8	2/8	1/8	3/8	2/8	4/8	**1/8**	1/8	2/8	1/8

Site Frequency Spectrum (SFS)-based approaches

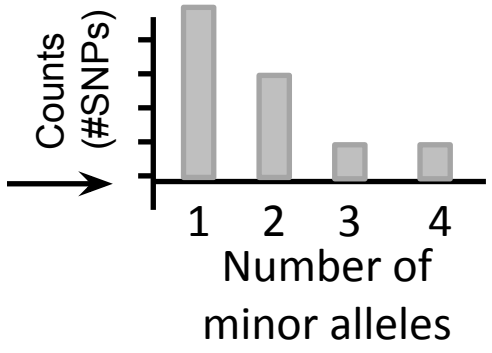
Full-likelihood methods = not adapted to WGS because of big data & recombination events

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)	1/8	2/8	1/8	3/8	2/8	4/8	**1/8**	1/8	2/8	1/8

« 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)	1/8	2/8	1/8	5/8	2/8	4/8	7/8	1/8	6/8	1/8

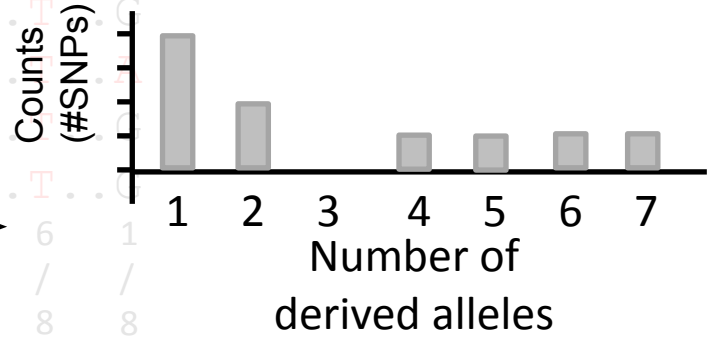
Site Frequency Spectrum (SFS)-based approaches

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..C
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..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
```

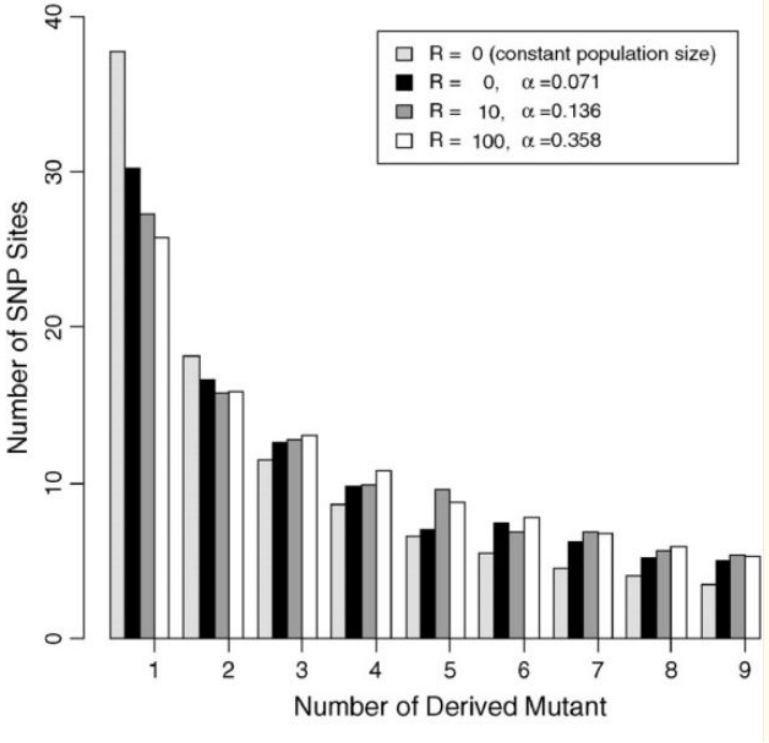
« Unfolded 1D-Site Frequency Spectrum » (unfolded 1D-SFS)



Derived allele frequency (DAF)

1/8 2/8 1/8 5/8 2/8 4/8 7/8 1/8 6/8 1/8

Site Frequency Spectrum (SFS)-based approaches



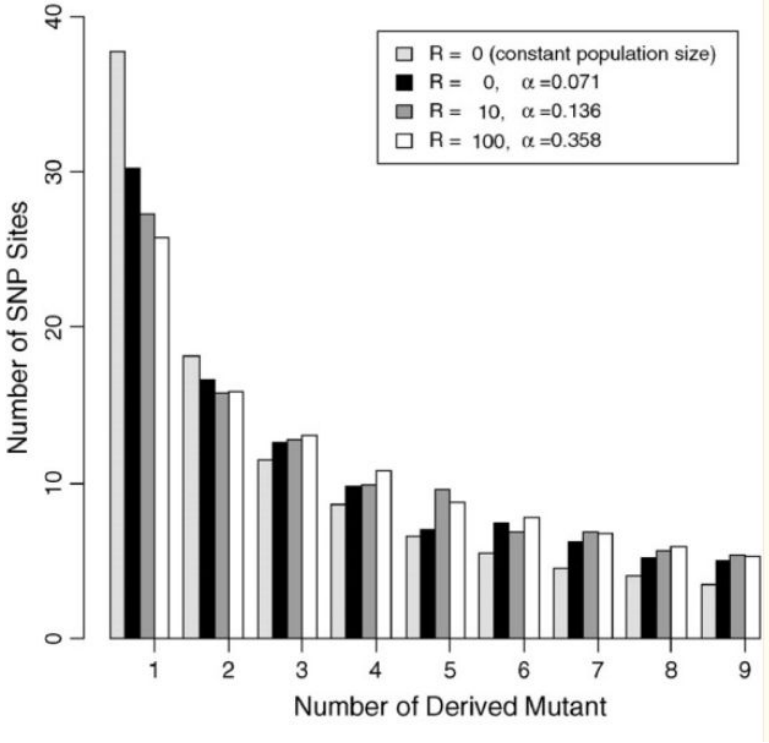
Zhu & Bustamante, 2005

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

//Pop expansion -> excess

⟶ SFS are therefore informative about *Ne* changes

Site Frequency Spectrum (SFS)-based approaches



Zhu & Bustamante, 2005

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

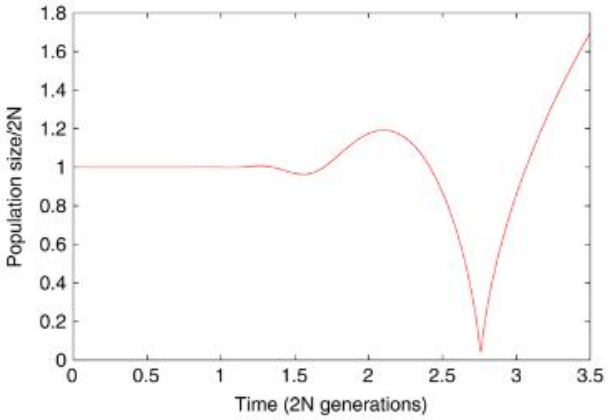
//Pop expansion -> excess

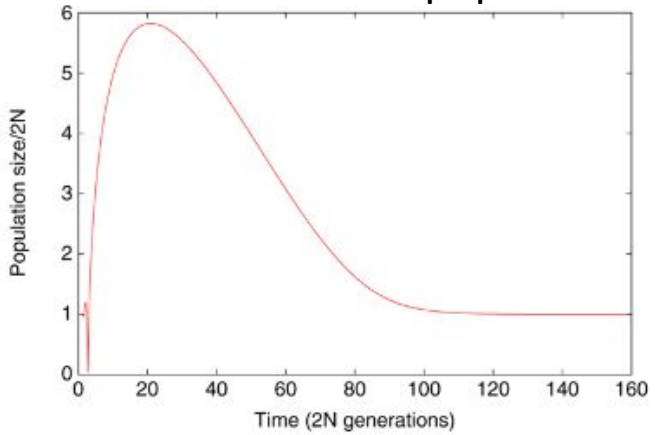
⟶ SFS are therefore informative about *Ne* changes

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!

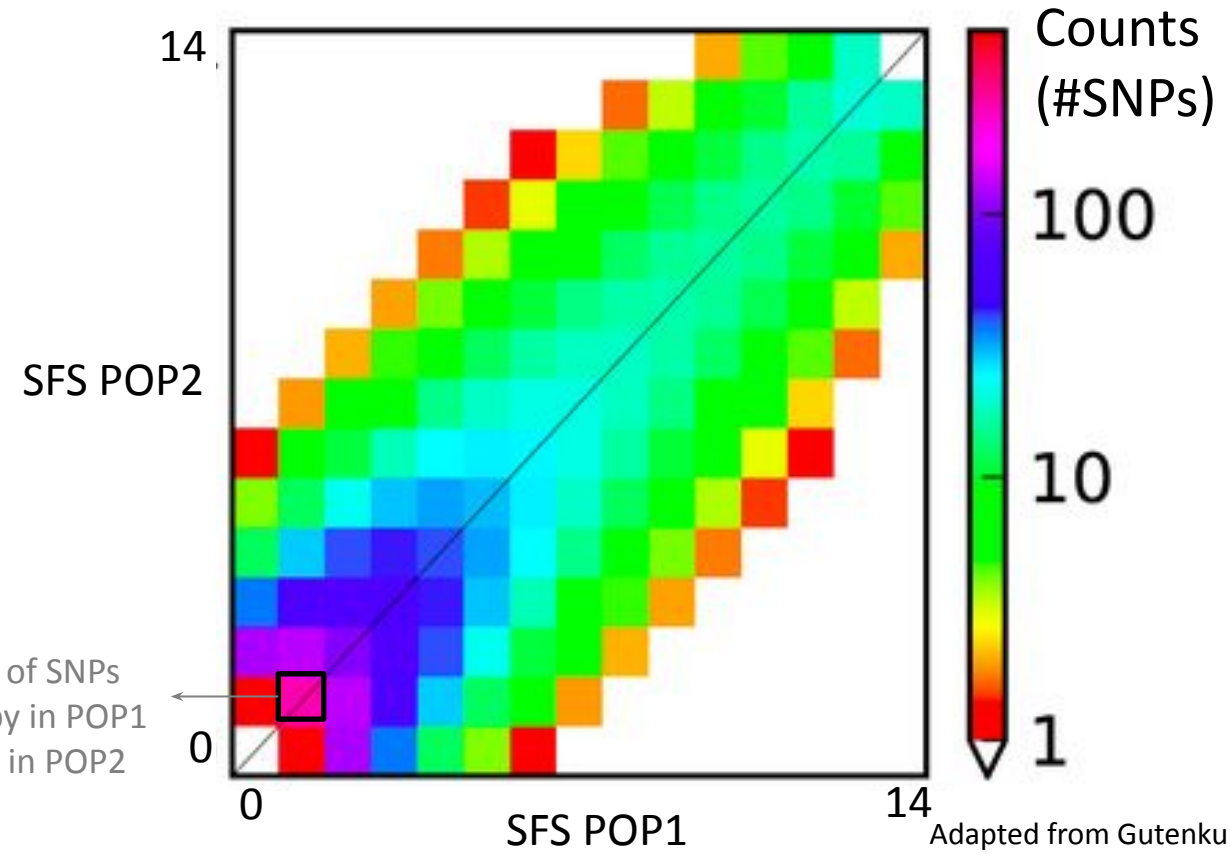




Myers et al. 2008 "Can one learn history from the allelic spectrum?"

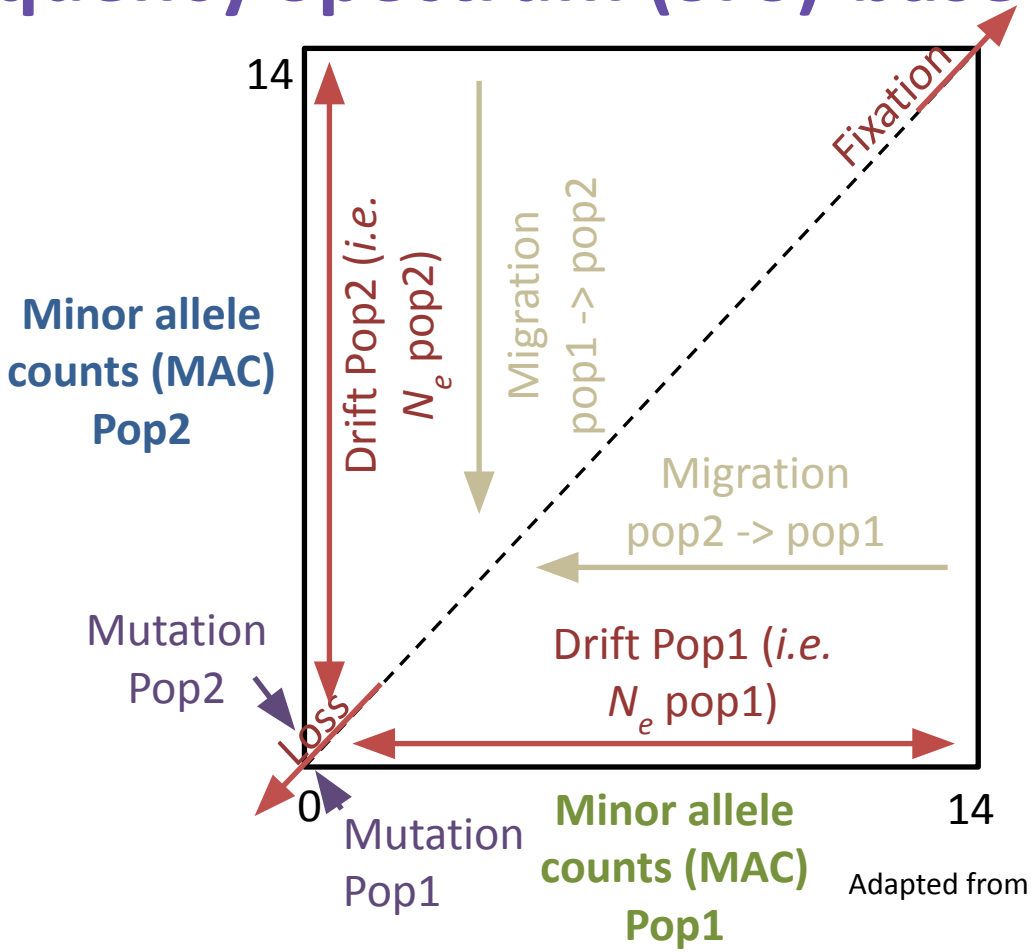
-> 2-dimensional site frequency spectrum (2D-SFS)
(use of polymorphism data from 2 populations)

Site Frequency Spectrum (SFS)-based approaches



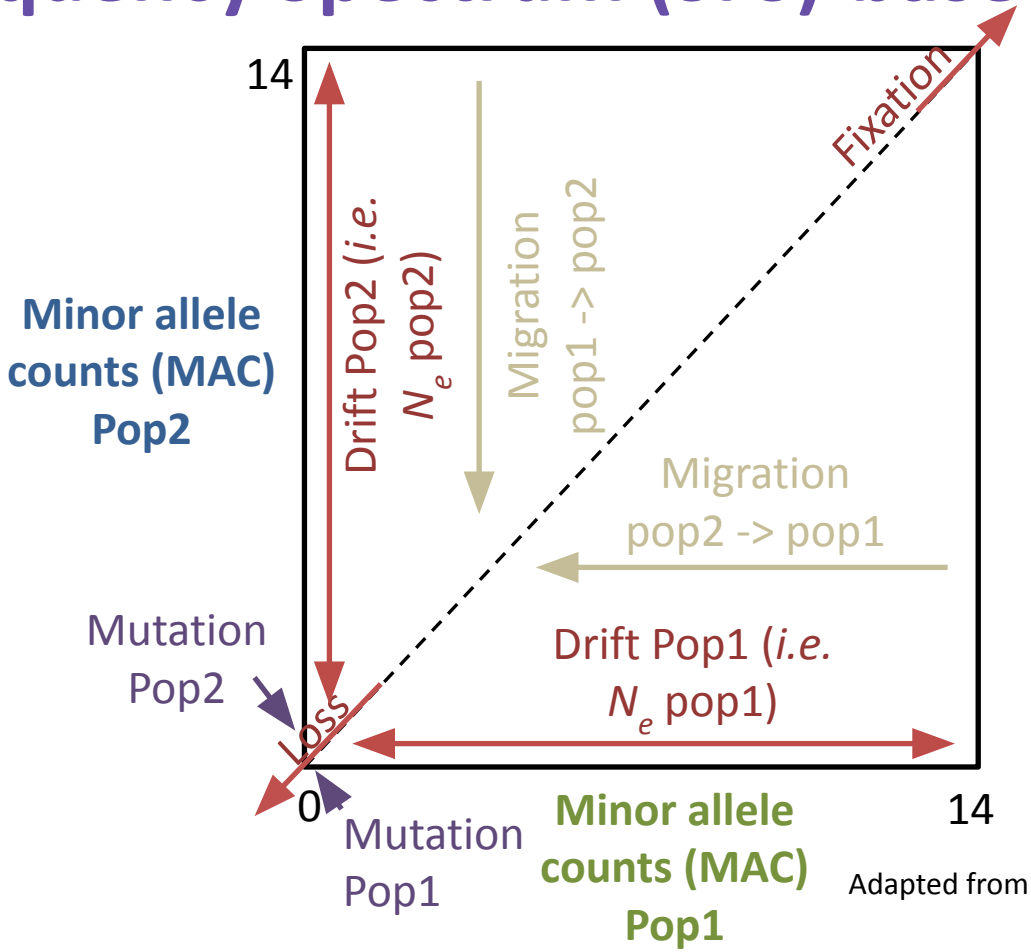
Adapted from Gutenkunst et al. 2009 Plos Genet

Site Frequency Spectrum (SFS)-based approaches



Adapted from Gutenkunst et al. 2009 Plos Genet

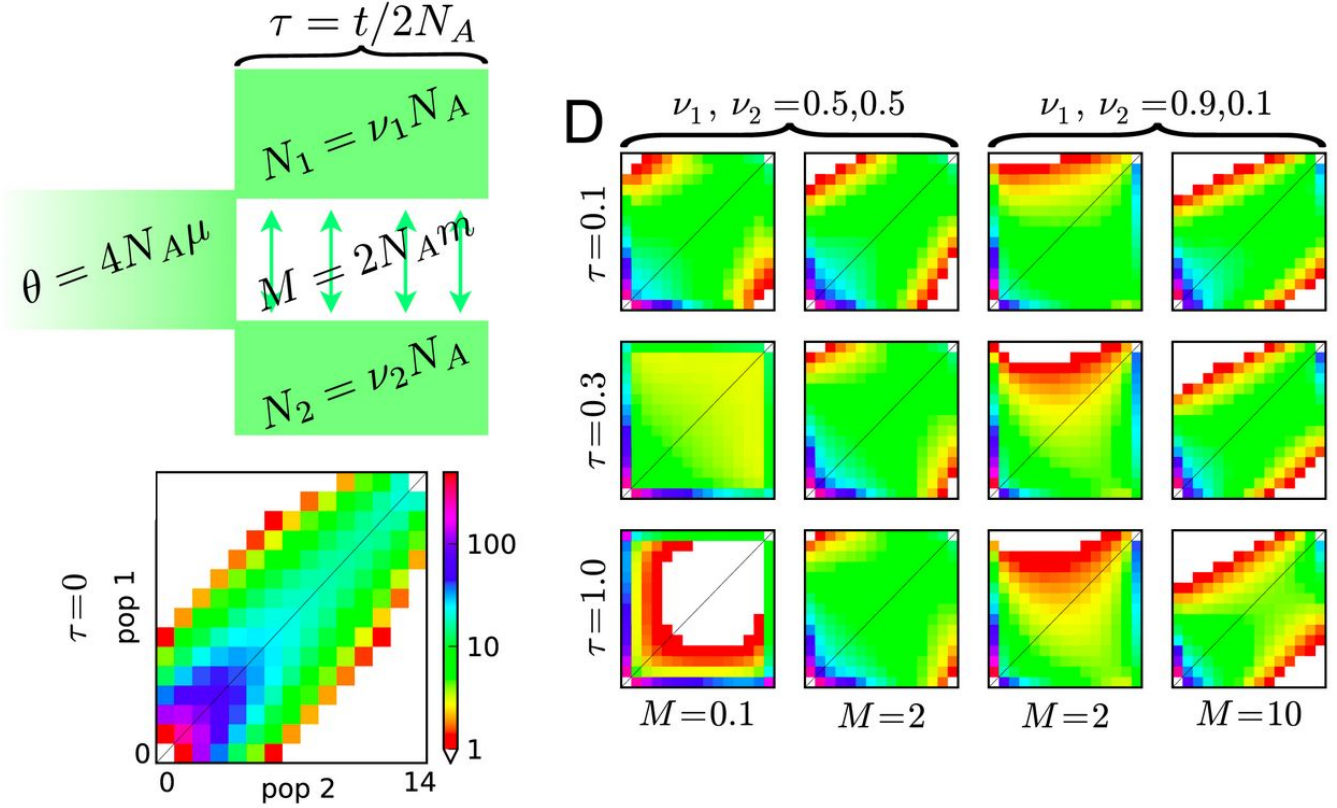
Site Frequency Spectrum (SFS)-based approaches



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

Adapted from Gutenkunst et al. 2009 Plos Genet

Composite likelihood approach : ∂a∂i



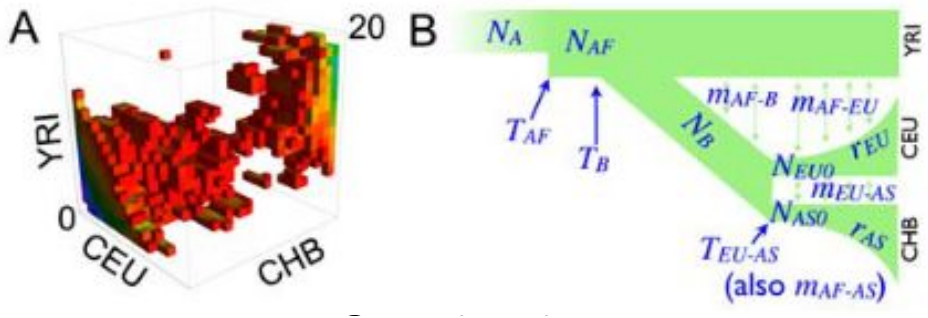
Gutenkunst et al. 2009 Plos Genet

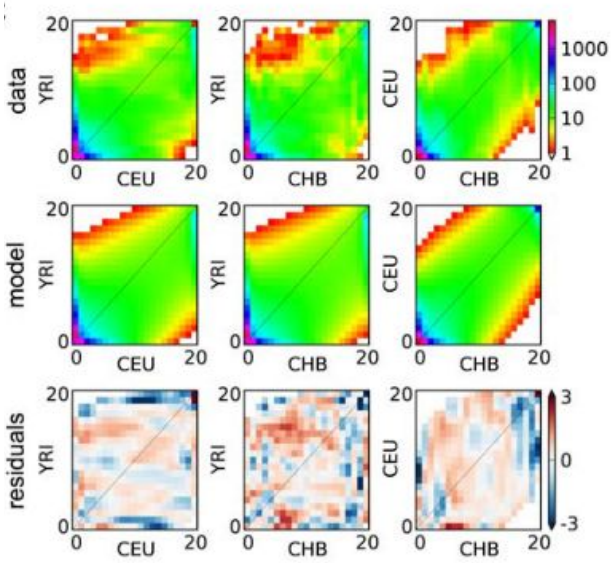
Composite likelihood approach : ∂a∂i

3-dimensional SFS

The implementation (∂a∂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)

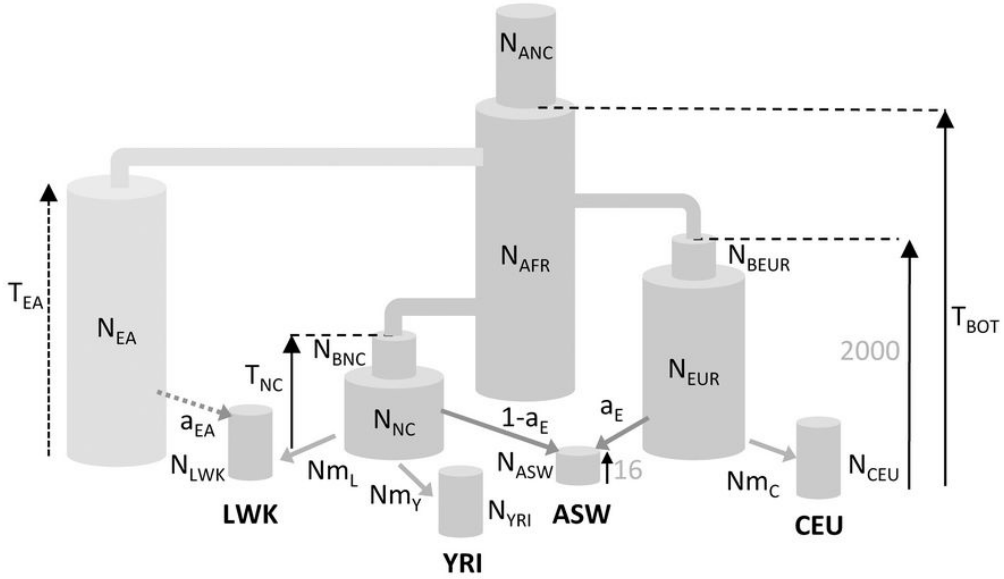
CHB: Han Chinese, Beijing, China (EAS)

Improvements to extend the ∂a∂i strategy to 4-populations (MULTIPOP, Lukic & Hey 2012 Genetics) & 5-populations (*Moments*, Jouganous et al. 2017 Genetics)

Gutenkunst et al. 2009 Plos Genetics, 2010 Nature Precedings

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 isssue 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: TTCGACATGCT…
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.

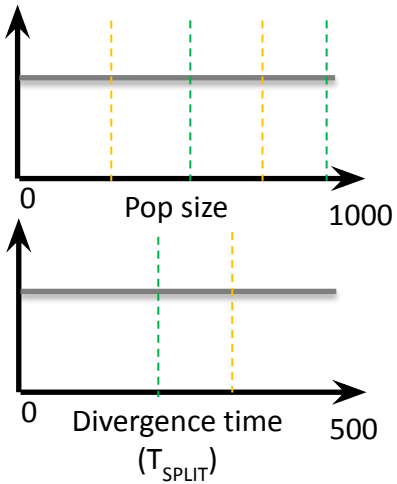
POP SIZE1: uniform[0-1000]
POP SIZE2: uniform[0-1000]
TSPLIT : uniform[0-500]

e.g.

Simul1:
PopSize1=763
PopSize2=261
Tsplit = 330

Simul2:
PopSize1=493
PopSize2=921
Tsplit = 234

… x i simulations



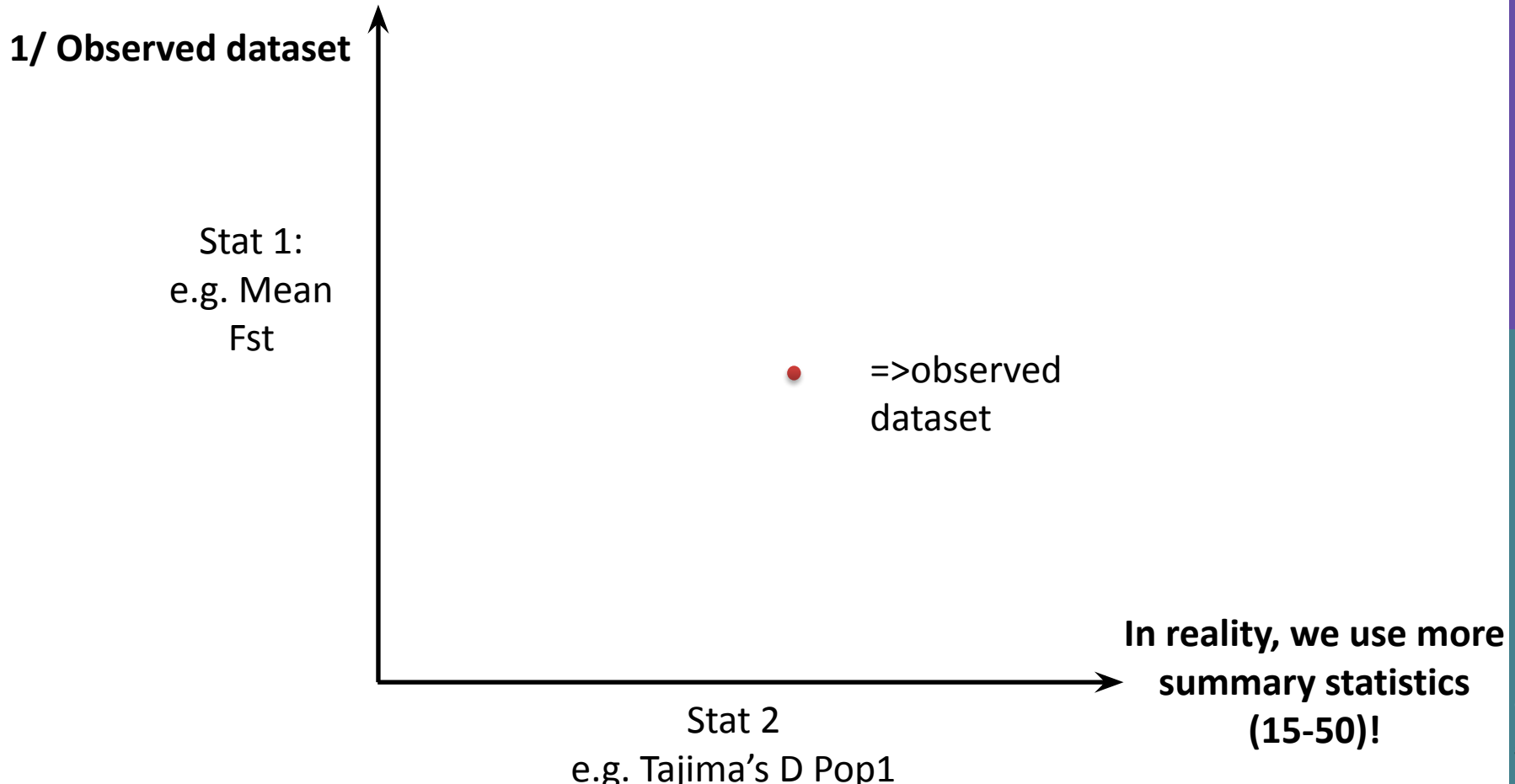
Same summary statistics than for the real data

Model 2

…

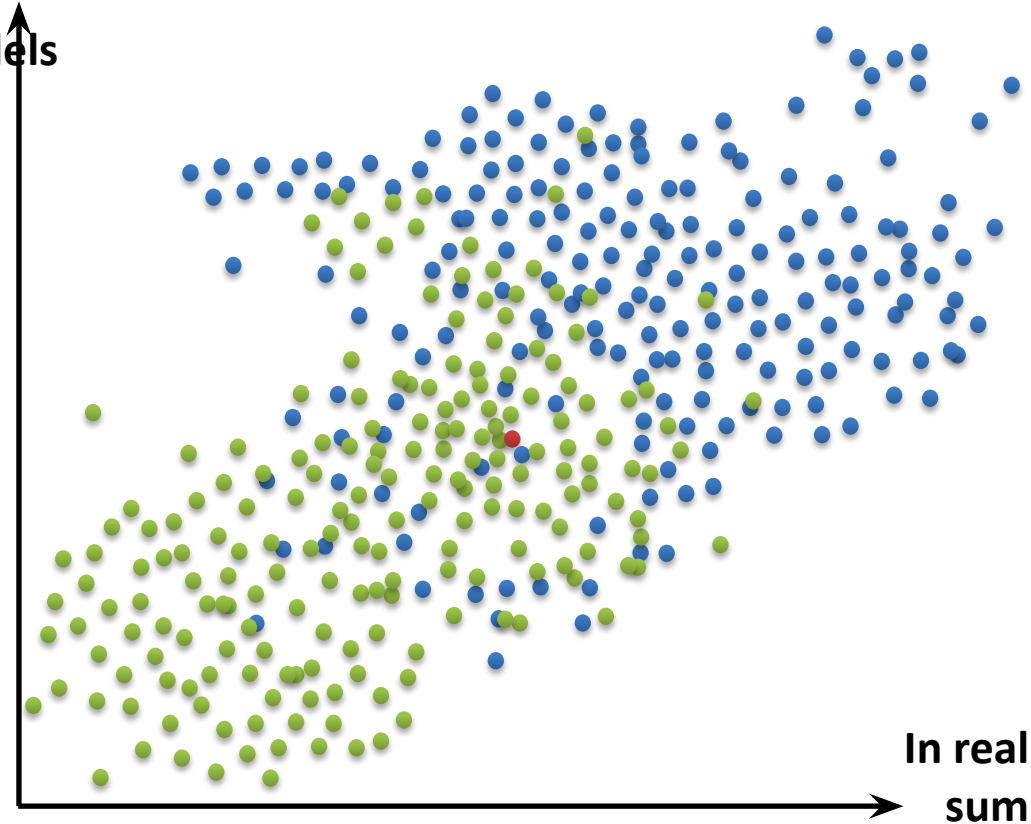
Same summary statistics

ABC in practice (very simplified)…



ABC in practice (very simplified)…

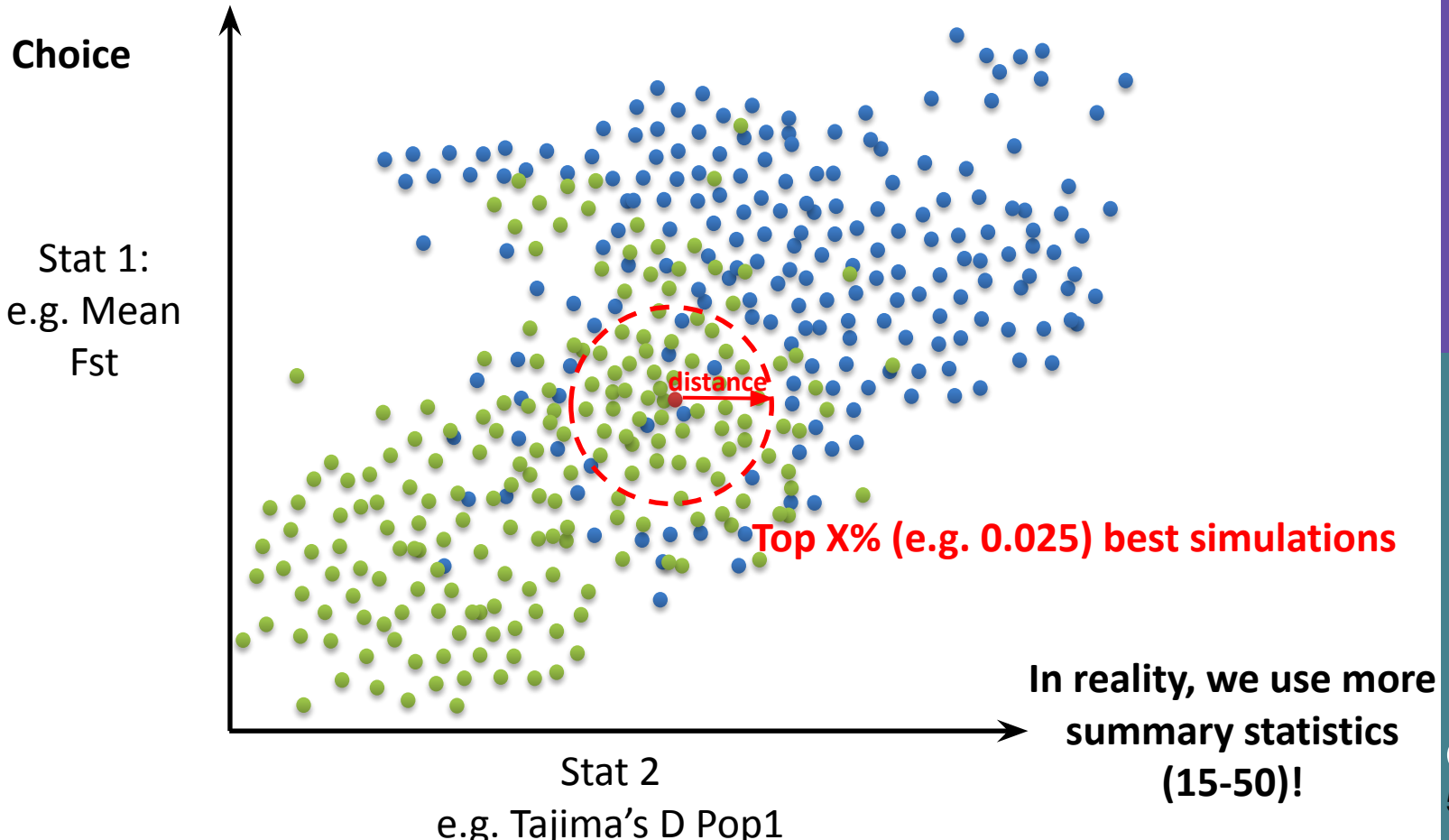
2/ Demographic models & simulations



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

ABC in practice (very simplified)…

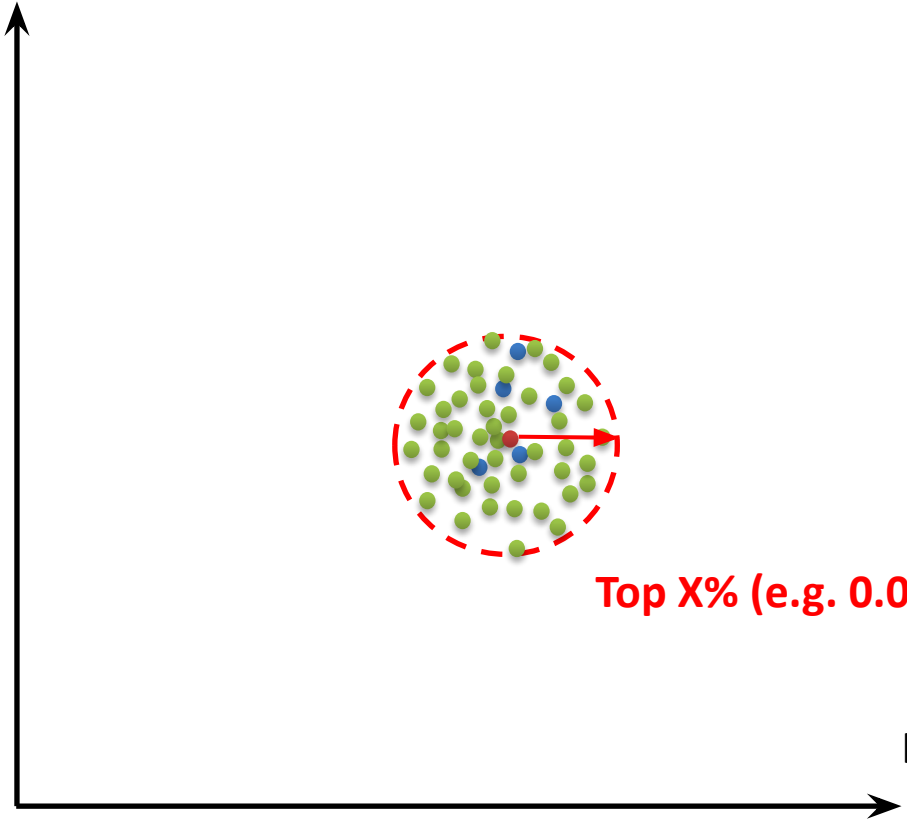
3/ Model Choice



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

ABC in practice (very simplified)…

3/ Model Choice

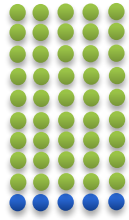


Top X% (e.g. 0.025) best simulations

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

ABC in practice (very simplified)…

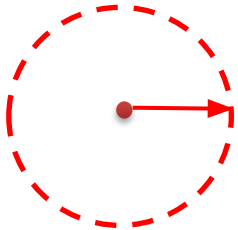
3/ Model Choice



Posterior probability (green model) =90%

PostProb(blue)=10%

Stat 1:
e.g. Mean Fst



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

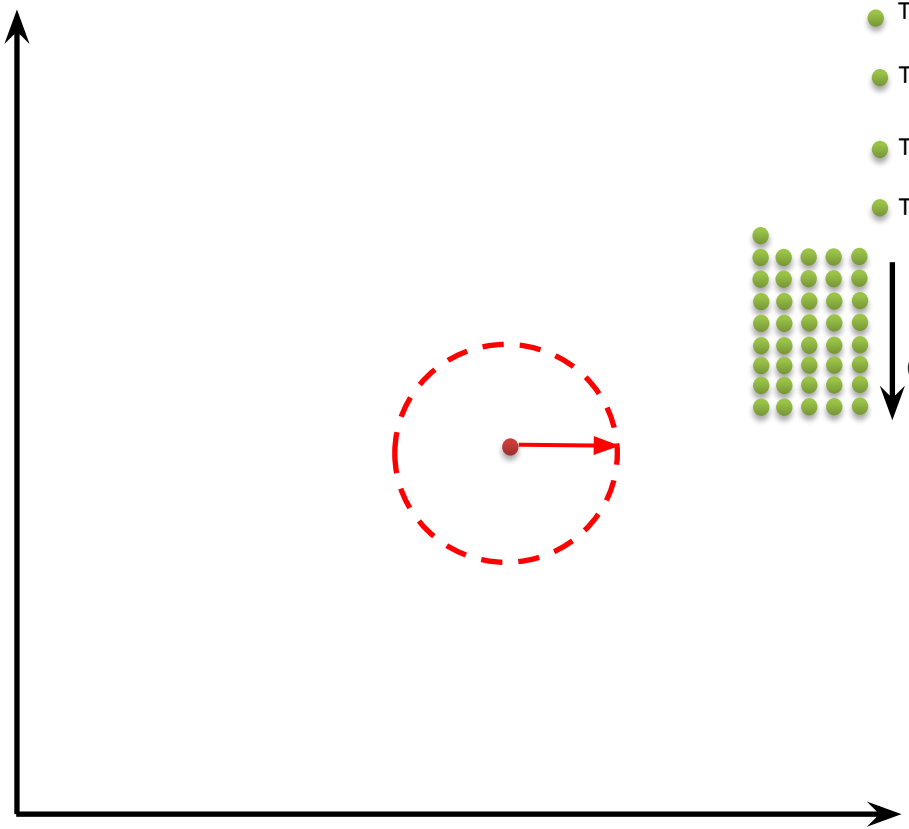
Stat 2
e.g. Tajima's D Pop1

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

ABC in practice (very simplified)…

4/ Estimation of parameters under the best model

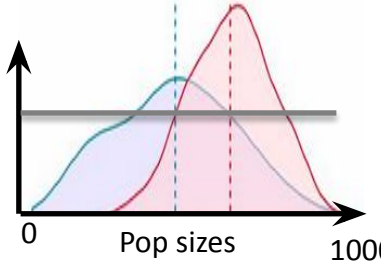
Stat 1:
e.g. Mean Fst

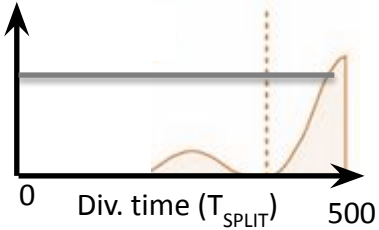


Stat 2
e.g. Tajima's D Pop1

- T_{SPLIT} = XXX ; PopSize1= YYY; PopSize2=ZZZ
- T_{SPLIT} = XXX ; PopSize1= YYY; PopSize2=ZZZ
- T_{SPLIT} = XXX ; PopSize1= YYY; PopSize2=ZZZ
- T_{SPLIT} = …

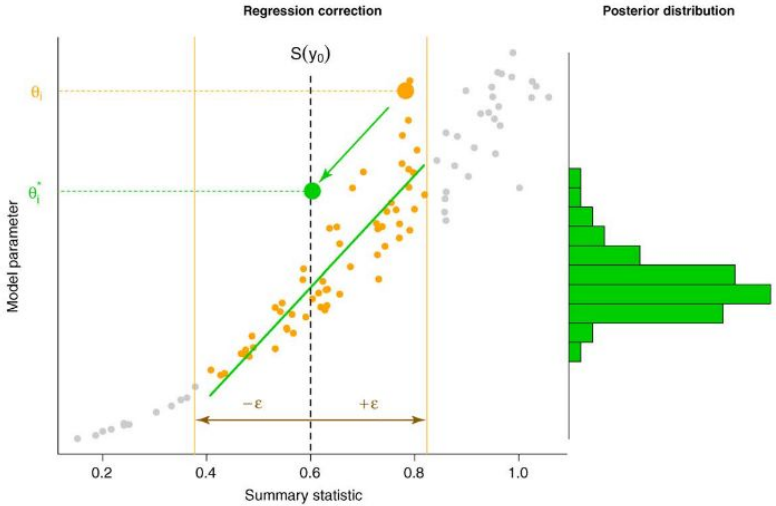
Posterior distributions of the parameters





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



Csillery et al. 2010 TREE (Trends in Ecol Evol)

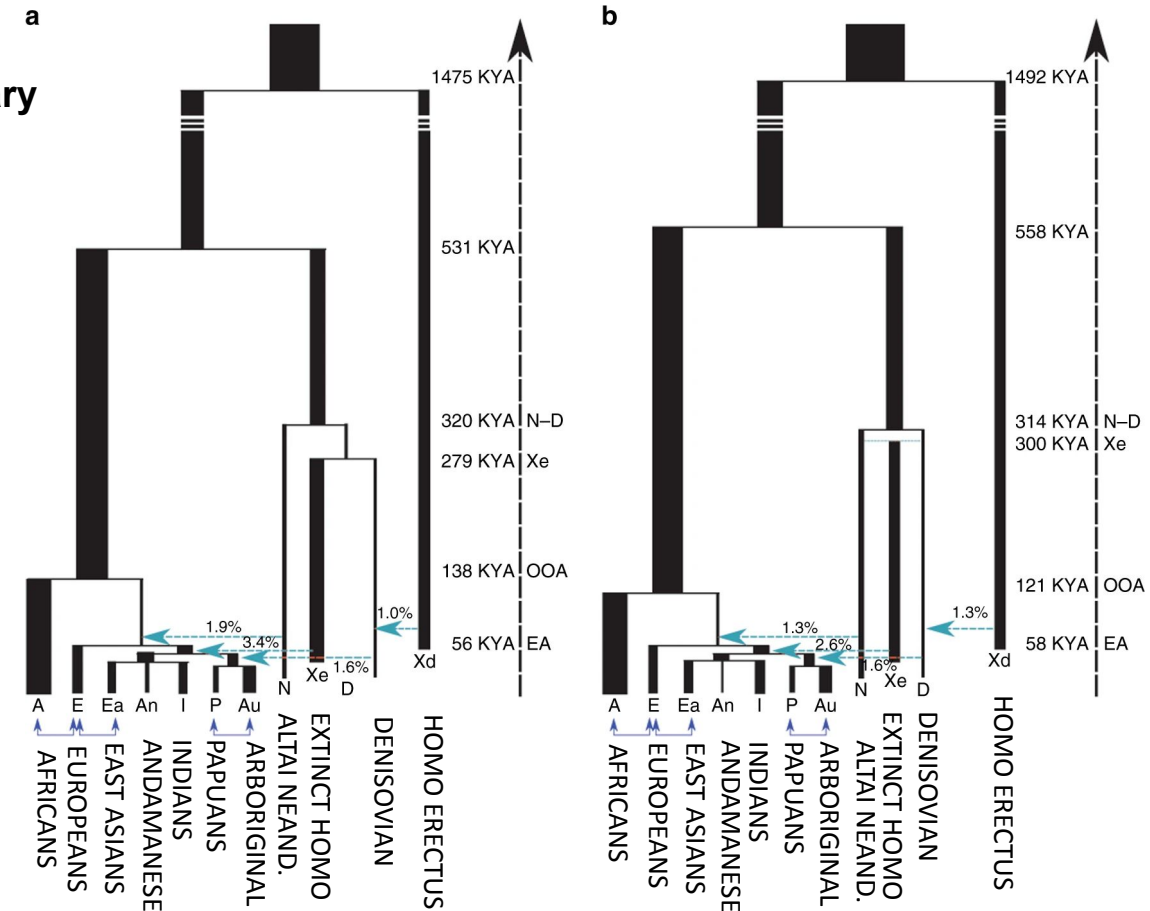
These traditional rejection/regression methods remain popular

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

Evolutionary history of humans



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

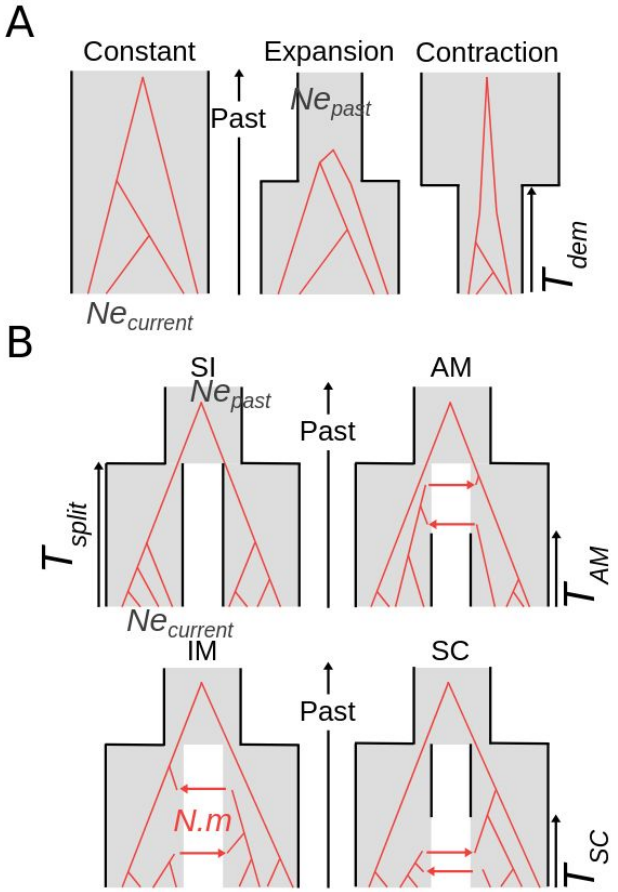
Mondal et al. 2019 Nature Communications (among many!)

Examples of applications

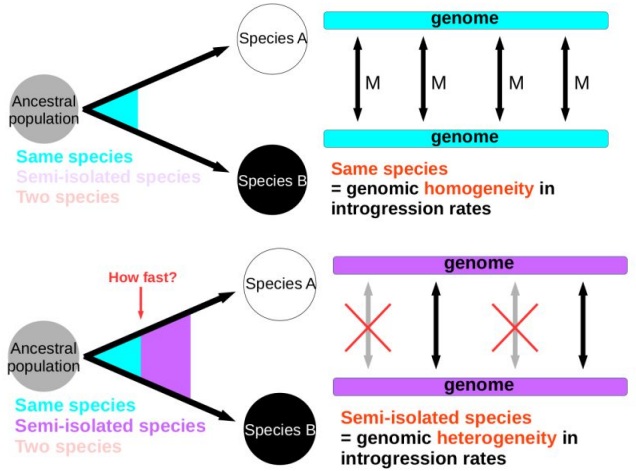
Simpler approaches: demographic simulations in single or pairs of populations (DILS, cf Camille Roux et al.)

Physalia "Model-based demographic inference from population genomics"

Fraïsse et al. 2021 Mol Ecol Res



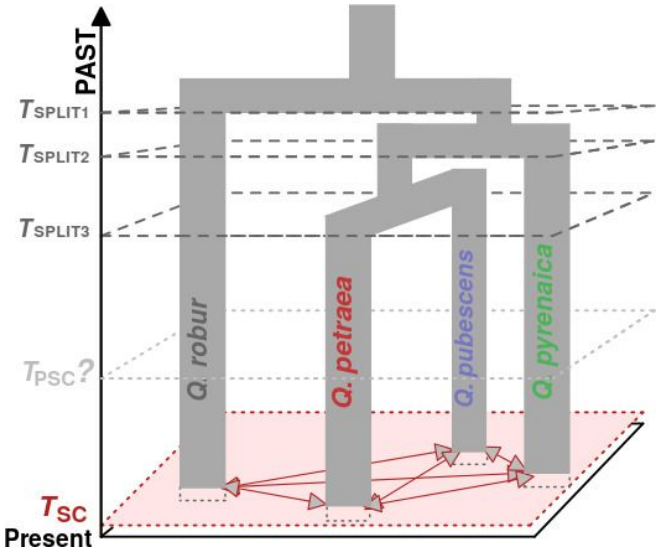
Although the number of populations is very limited, DILS accounts for many confounding factors in demographic modelling: presence of linked selection (background selection + selective sweeps), as well as the presence of barriers to gene flow



Examples of applications

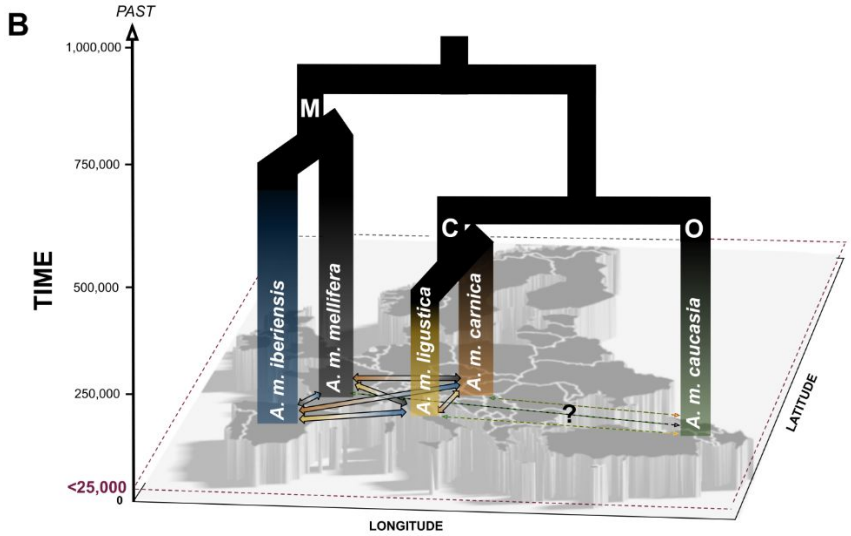
Inferences based on all pairs of pops/subspecies/species to built a general hypothetical model

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*)

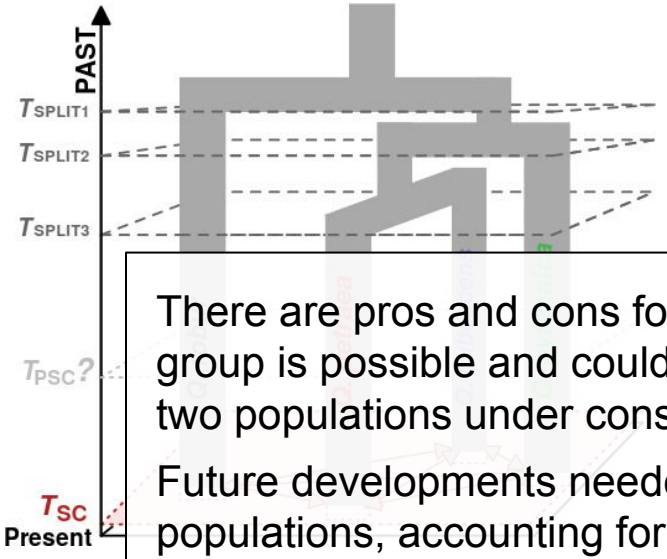


Leroy et al. 2024 Mol Biol Evol

Examples of applications

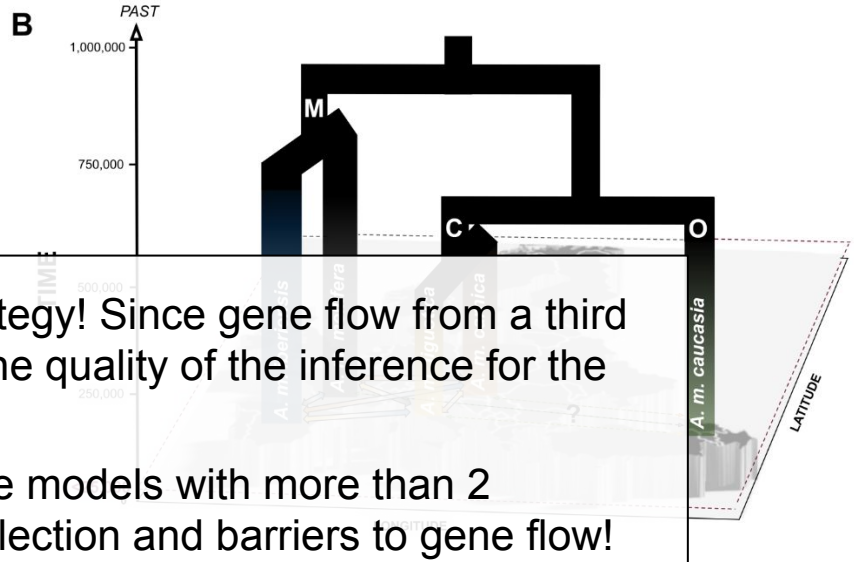
Inferences based on all pairs of pops/subspecies/species to built a general hypothetical model

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



Leroy et al. 2017; 2020; New Phytologist

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



Leroy et al. 2024 bioRxiv

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!

Approximate Bayesian Computation : pros & cons

Advantages:

- Likelihood-free
- Flexible framework

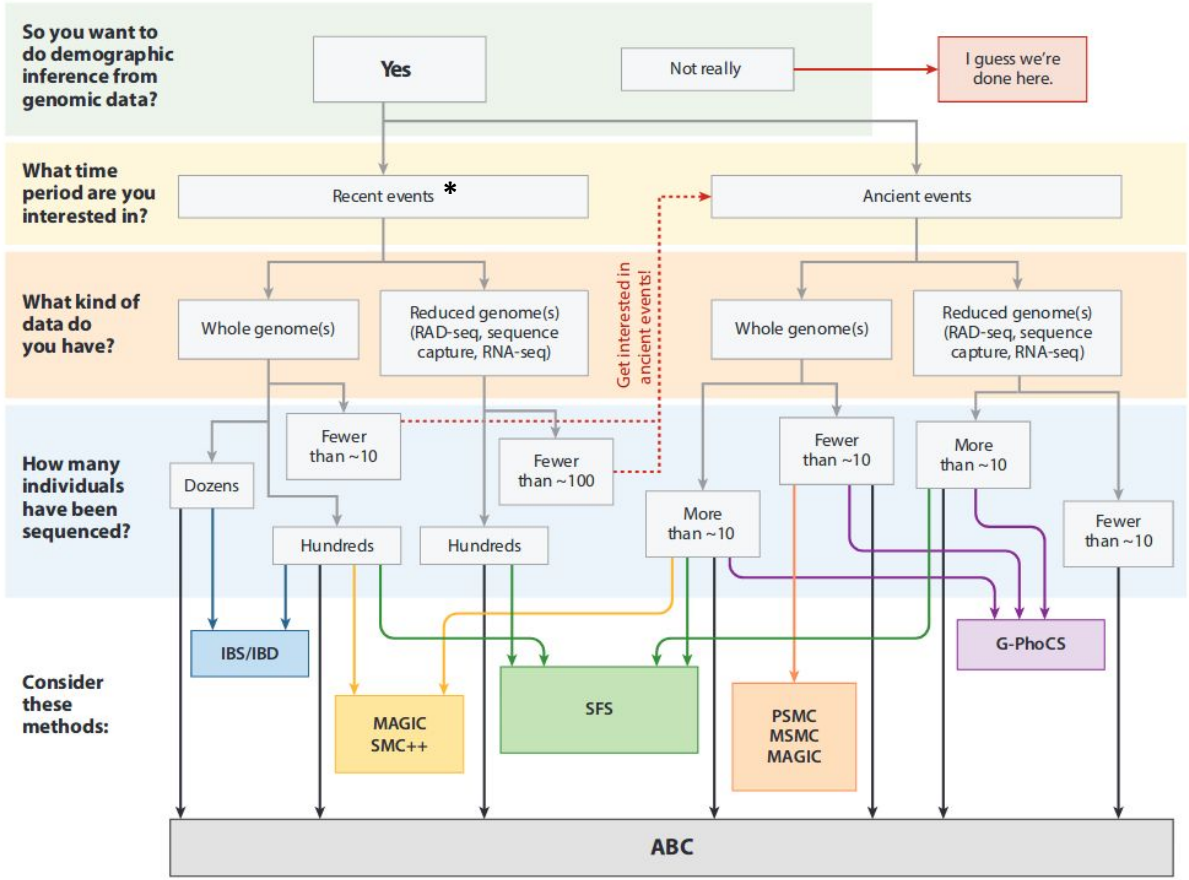
e.g. including possibility to modelize complex genome-wide processes such as heterogeneity in migration rates or effective population sizes (-> DILS)

- 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?



* 'Recent events' here: <1000 generations ago

Beichman et al. 2018 Nat Rev Ecol Evol Syst

To finish, back to the low diversity of lynx…

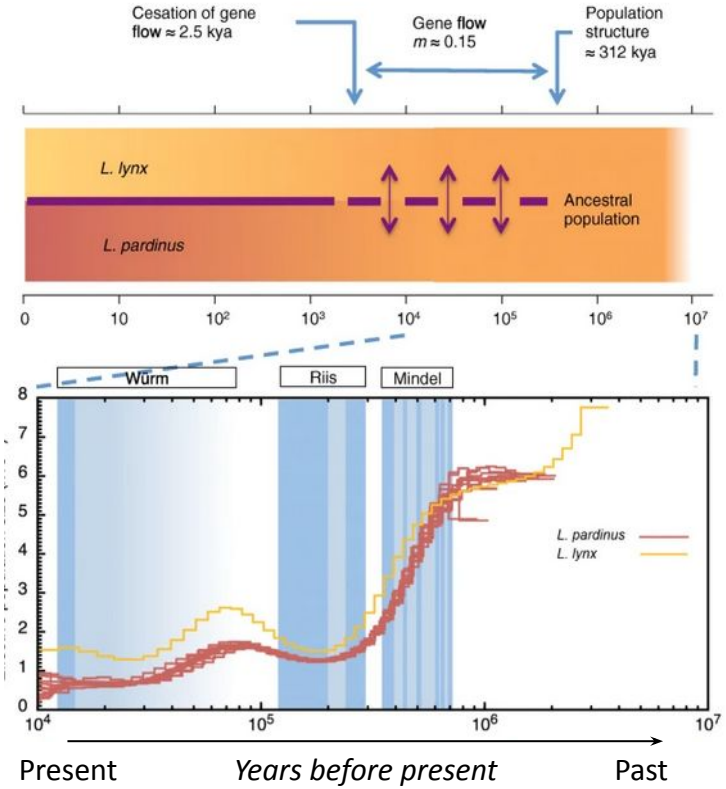
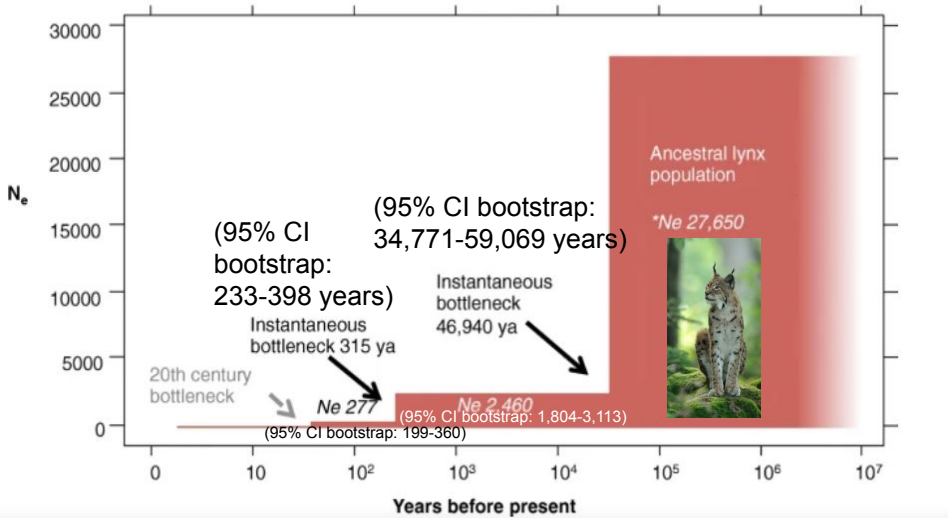


Table S20. Models considered in dadi. For each model we give the number of parameters (K), the AIC score and the AIC differences with respect to the one with the mínimum AIC (Δi).

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



Abascal et al. 2016 Genome Biology

To finish, back to the low diversity of lynx…

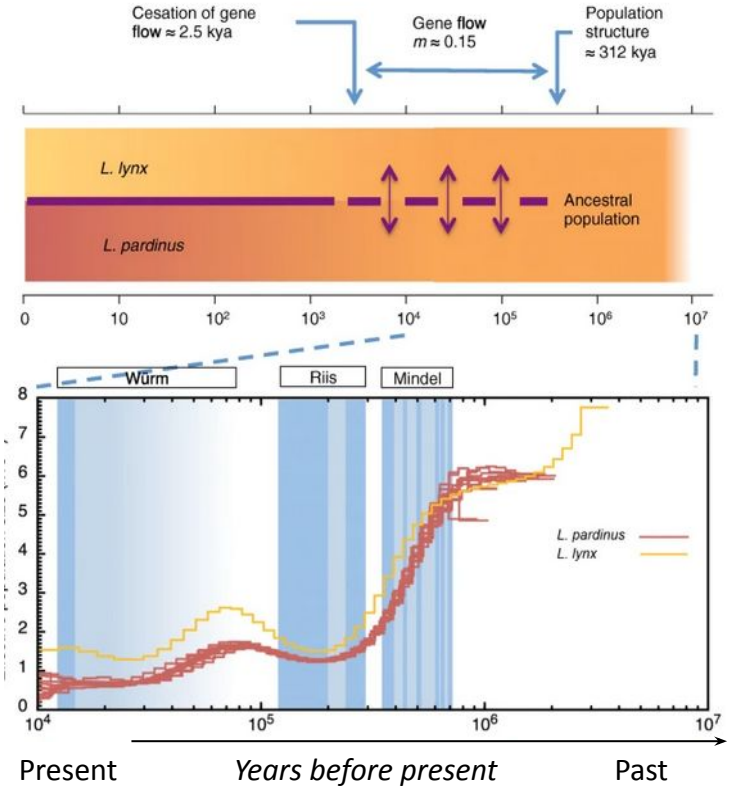
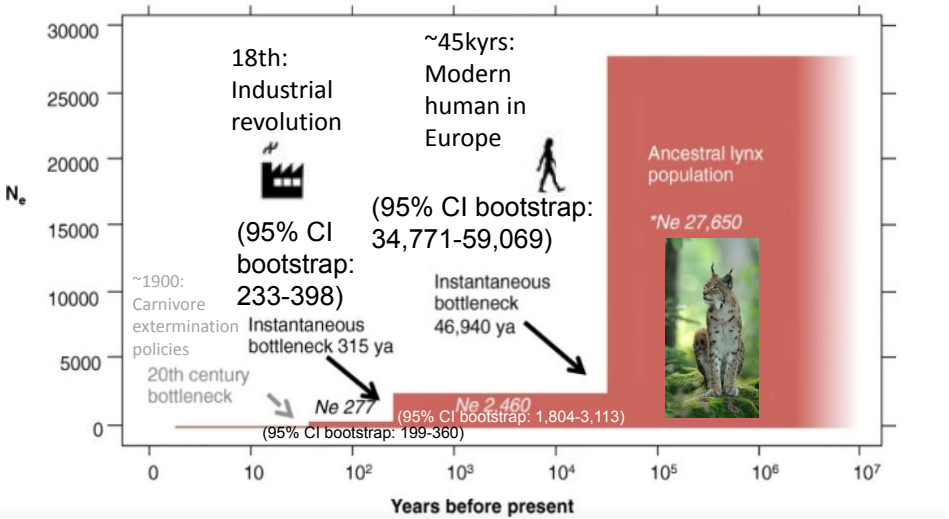


Table S20. Models considered in dadi. For each model we give the number of parameters (K), the AIC score and the AIC differences with respect to the one with the mínimum AIC (Δi).

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



Abascal et al. 2016 Genome Biology

To finish, back to the low diversity of lynx…

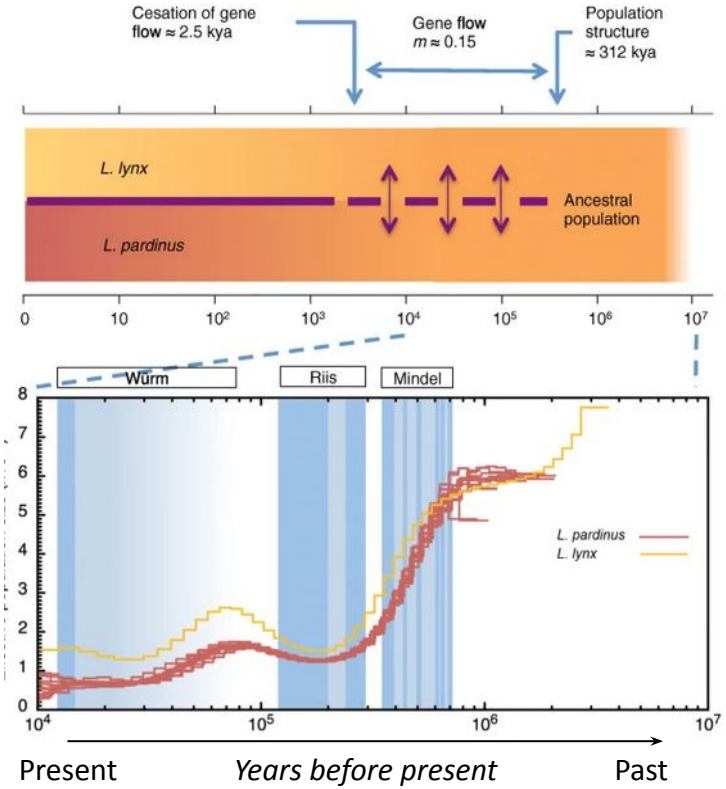
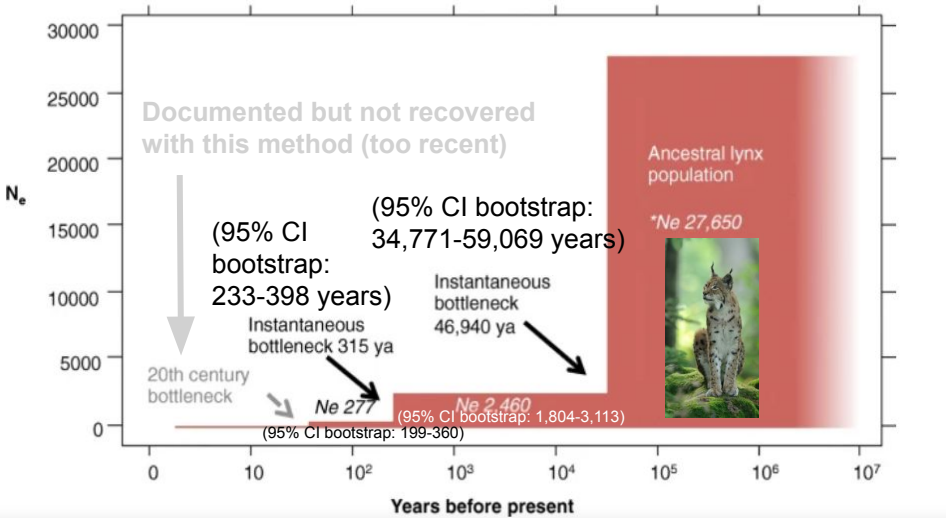


Table S20. Models considered in dadi. For each model we give the number of parameters (K), the AIC score and the AIC differences with respect to the one with the minimum AIC (Δi).

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



Abascal et al. 2016 Genome Biology

To finish, back to the low diversity of lynx…

+ LD-based demographic modelling

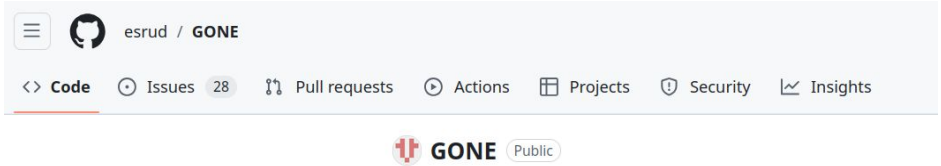
When recombination is not considered, the **mutation rate becomes the sole free parameter, guiding the timing of the coalescence process** (Hudson, 1990).
-> **Time is needed for mutations to accumulate, low resolution in the recent past.**

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 Caballero[2]

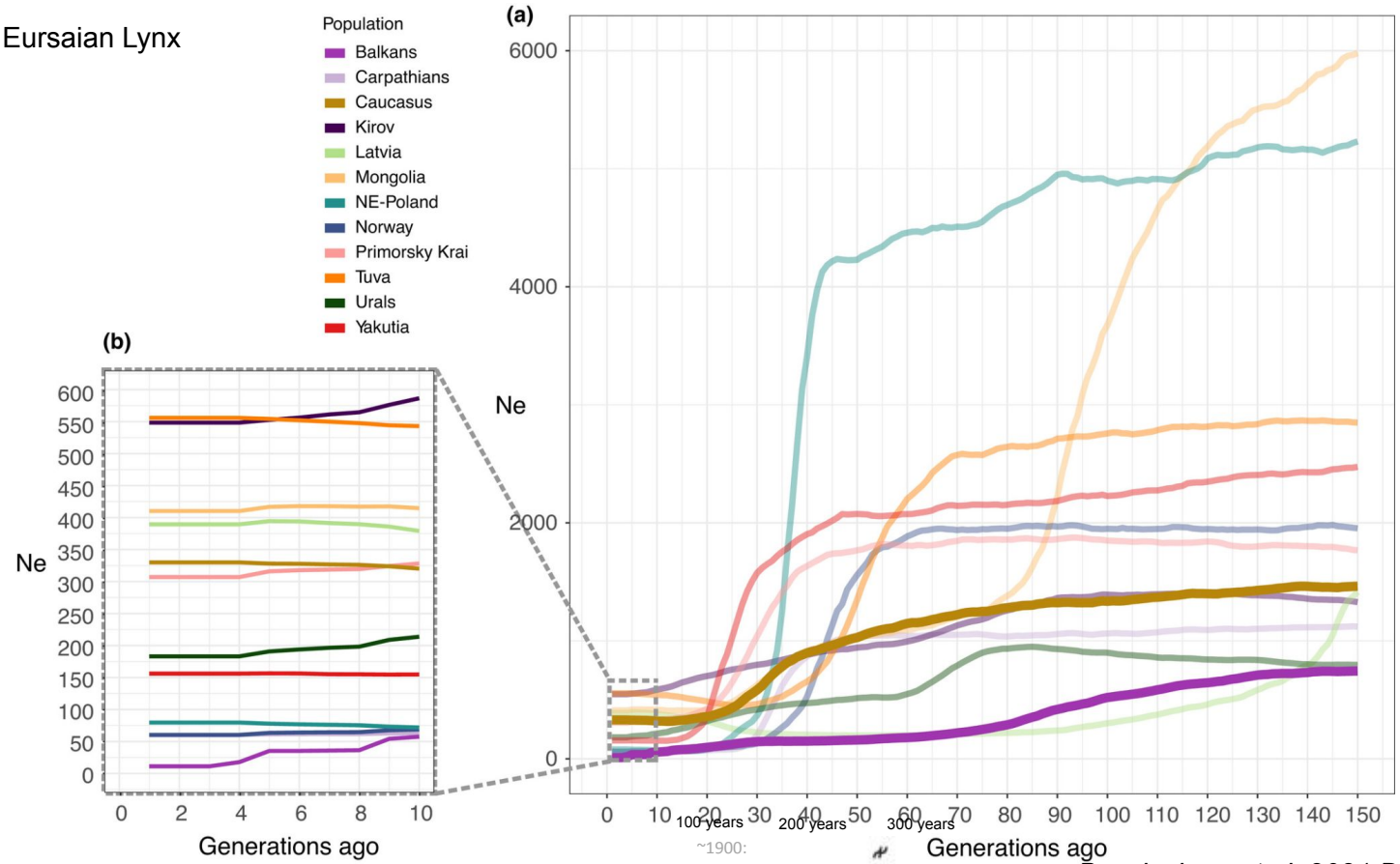
[1]Departamento de Biología Funcional, Facultad de Biología, Universidad de Oviedo, Oviedo, Spain
[2]Centro de Investigación Mariña, Departamento de Bioquímica, Genética e Inmunología, Edificio CC Experimentais, Campus de Vigo, Universidade de Vigo, Vigo, Spain
[3]MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, United Kingdom
[4]Departamento de Mejora Genética Animal, INIA, Madrid, Spain
[5]Institute 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…

Eursaian Lynx



generation time (~5 years)

Bazzicalupo et al. 2021 Diversity and Distributions

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

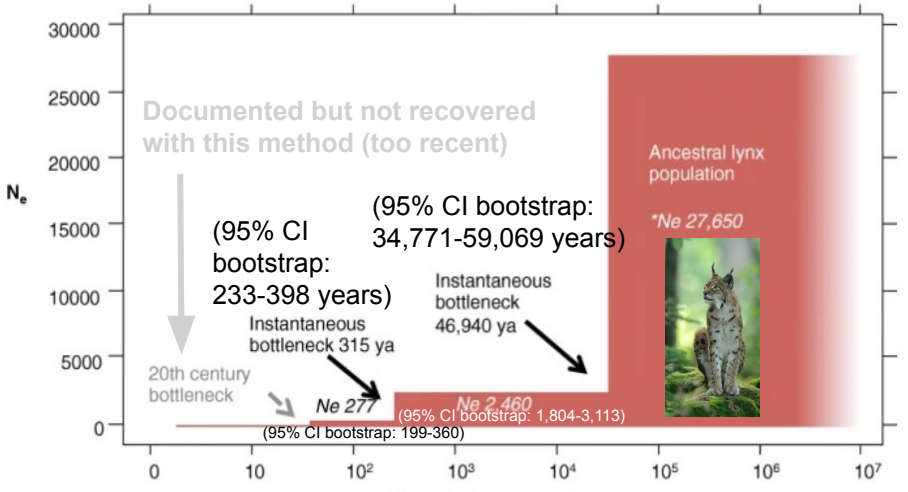
D - Mutation rate

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

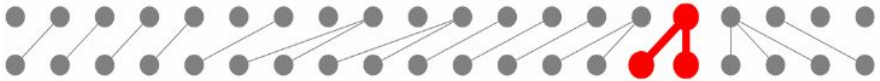
B - It determines how strongly genetic drift acts in a population

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 (*Ne*): 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 $\frac{1}{2Ne}$

⟶ 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

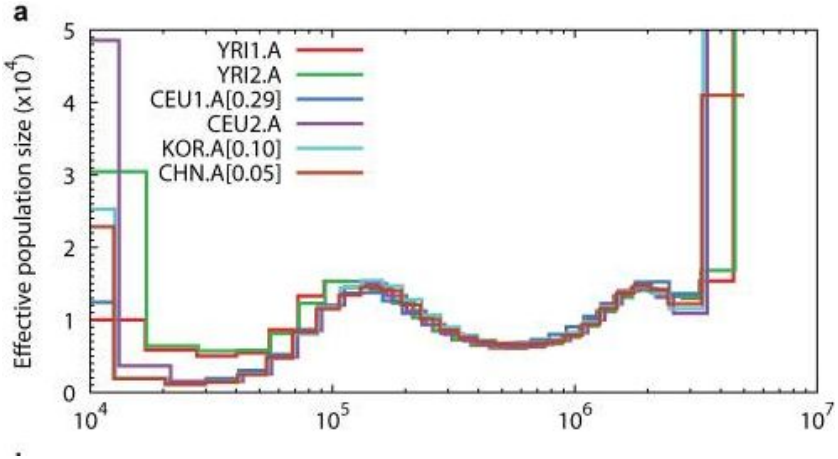
B - Timing of population divergence

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 ?

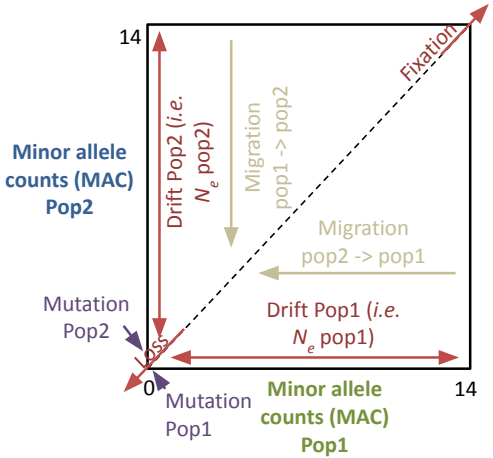
A - They calculate exact likelihood across whole genomes

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