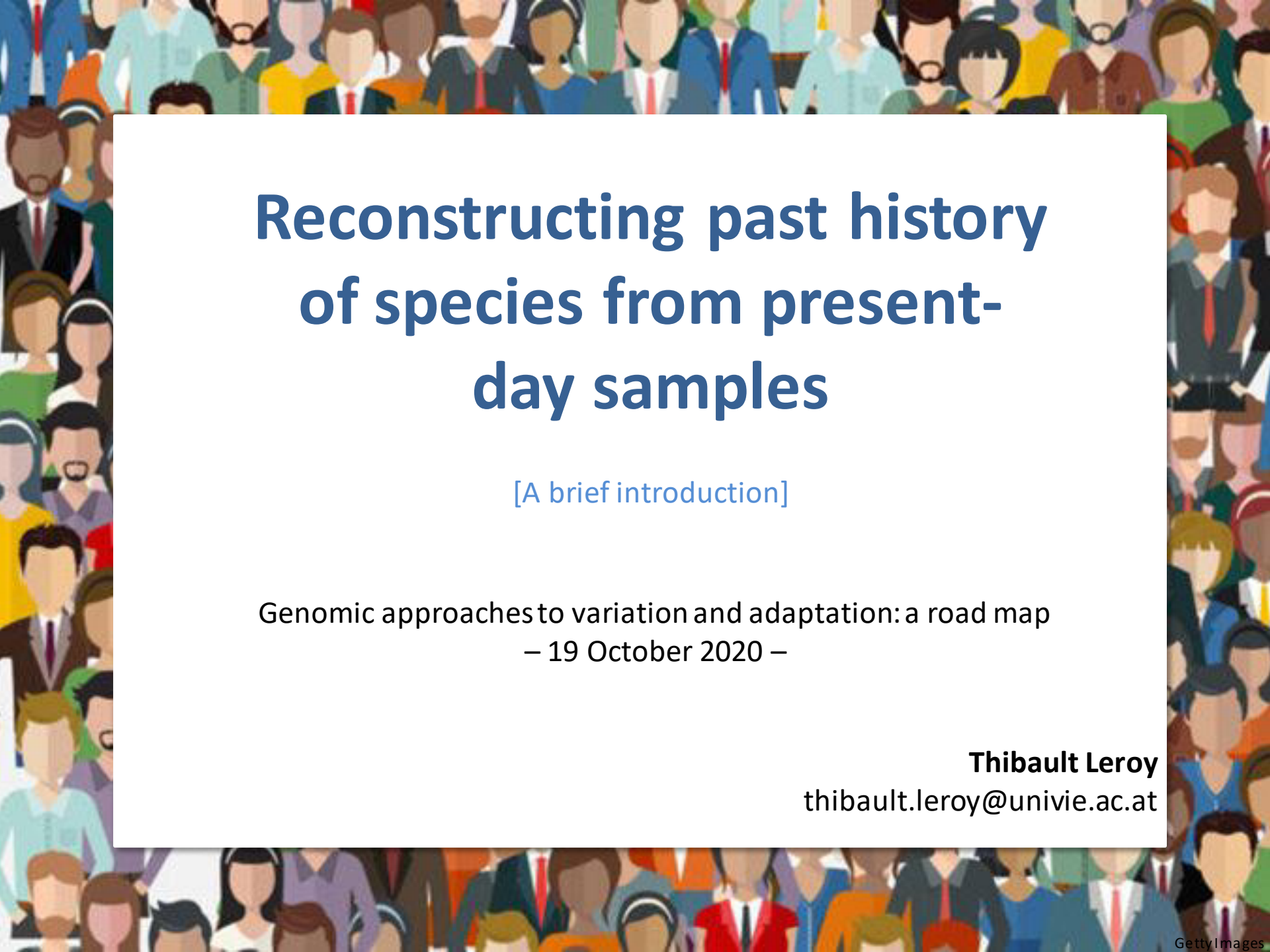


# Reconstructing past history of species from present-day samples

[A brief introduction]

Genomic approaches to variation and adaptation: a road map
– 19 October 2020 –

**Thibault Leroy**
thibault.leroy@univie.ac.at

## Genome(s): a wealth of information

- Nuclear vs. Organelle genomes (mtDNA and cpDNA)
- Coding and non-coding regions
- Genetic variation (SNPs, indels, transposable elements, larger structural variants)

Genetic diversity is highly variable among the tree of life!

θ (diploid species) = 4Neμ !


*Leffler et al. Plos Biol 2012*

## Effective population size (*Ne*)

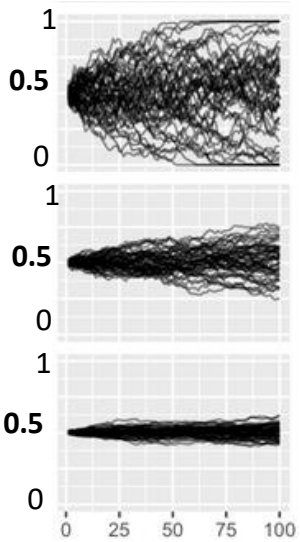
= the number of individuals in a Wright-Fisher model (i.e. the size of an idealized population) that would produce the same amount of genetic drift as in the real population

*Ne* is therefore a key parameter in population genetics!

*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 (*Ne*)

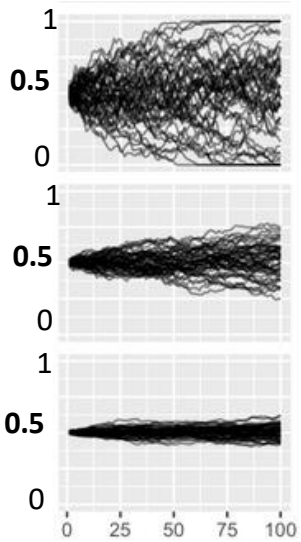
= the number of individuals in a Wright-Fisher model (i.e. the size of an idealized population) that would produce the same amount of genetic drift as in the real population

*Ne* is therefore a key parameter in population genetics!

*Sewall Wright*

*Ronald A. Fisher*

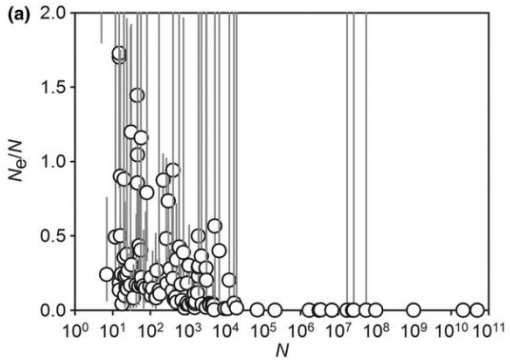
*WF model:*
- *non-overlapping generations*
- *no selection*
- *no mutation*
- *no migration*
- *random mating*




WF: a model for the allele frequency dynamics

Effective population sizes << census population sizes (i.e. the number of individuals in the population)

*Ne*/*Nc* ratios:




Crucial parameters for ecology & evolution…

But both *Ne* and *Nc* are difficult to estimates precisely (grey line = 95% CI for the ratio!)

*Palstra & Fraser, 2012 Ecol Evol*

## Genome(s): a wealth of information

- Nuclear vs. Organelle genomes (mtDNA and cpDNA)
- Coding and non-coding regions
- Genetic variation (SNPs, indels, transposable elements, larger structural variants)

Genetic diversity is highly variable among the tree of life!

$\theta$ (diploid species) = $4N_e\mu$ !

One SNP in every 1,000 nucleotides on average, which means there are roughly 3 million SNPs (*i.e.* heterozygous sites) in your own genome



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


*Leffler et al. Plos Biol 2012*

## Genome(s): a wealth of information

~8x10$^{-2}$


*Leffler et al. Plos Biol 2012*

- Nuclear vs. Organelle genomes (mtDNA and cpDNA)
- Coding and non-coding regions
- Genetic variation (SNPs, indels, transposable elements, larger structural variants)

Genetic diversity is highly variable among the tree of life!

θ (diploid species) = 4Neμ !



One SNP in every 1,000 nucleotides on average, which means there are roughly 3 million SNPs (*i.e.* heterozygous sites) in your own genome

~1x10$^{-3}$

## Genome(s): a wealth of information

- Nuclear vs. Organelle genomes (mtDNA and cpDNA)
- Coding and non-coding regions
- Genetic variation (SNPs, indels, transposable elements, larger structural variants)

Genetic diversity is highly variable among the tree of life!

θ (diploid species) = 4Neμ !

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



Eurasian lynx
~1.0x10$^{-4}$

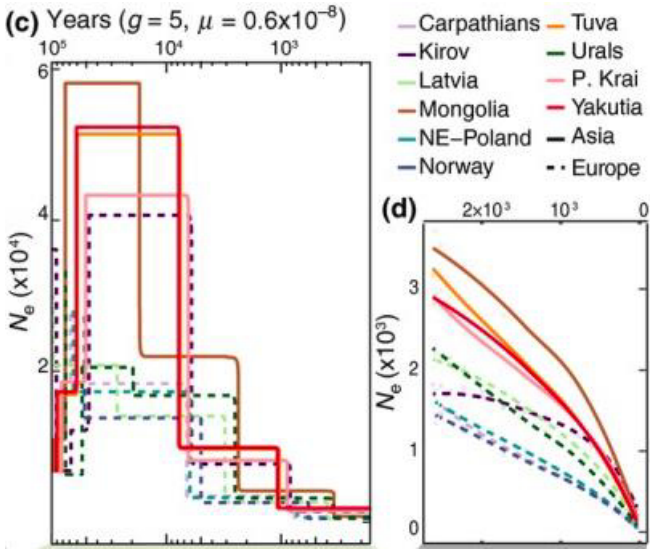

*Leffler et al. Plos Biol 2012*

## Genome(s): a wealth of information

- Nuclear vs. Organelle genomes (mtDNA and cpDNA)
- Coding and non-coding regions
- Genetic variation (SNPs, indels, transposable elements, larger structural variants)

Genetic diversity is highly variable among the tree of life!

θ (diploid species) = 4Neμ !

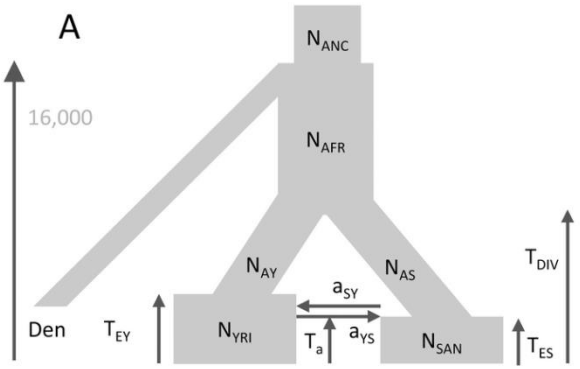



Eurasian lynx ~1.0x10$^{-4}$


*Leffler et al. Plos Biol 2012*

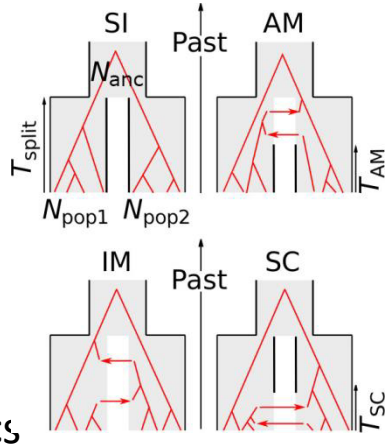
## What can we (try to) infer?

- (Historical changes in) effective population sizes
- Inferences of population splits
- Periods of isolation (allopatry) vs. periods of gene flow




*Excoffier et al. 2013 Plos Genetics*

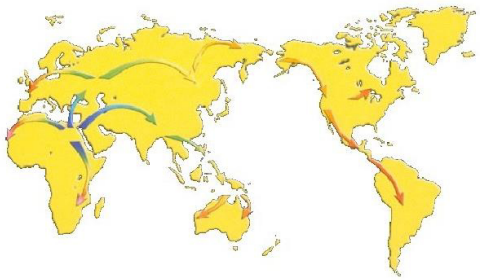
SI=Strict Isolation, AM=Ancient Migration, IM=Isolation-with-Migration (=island model), SC= Secondary Contact



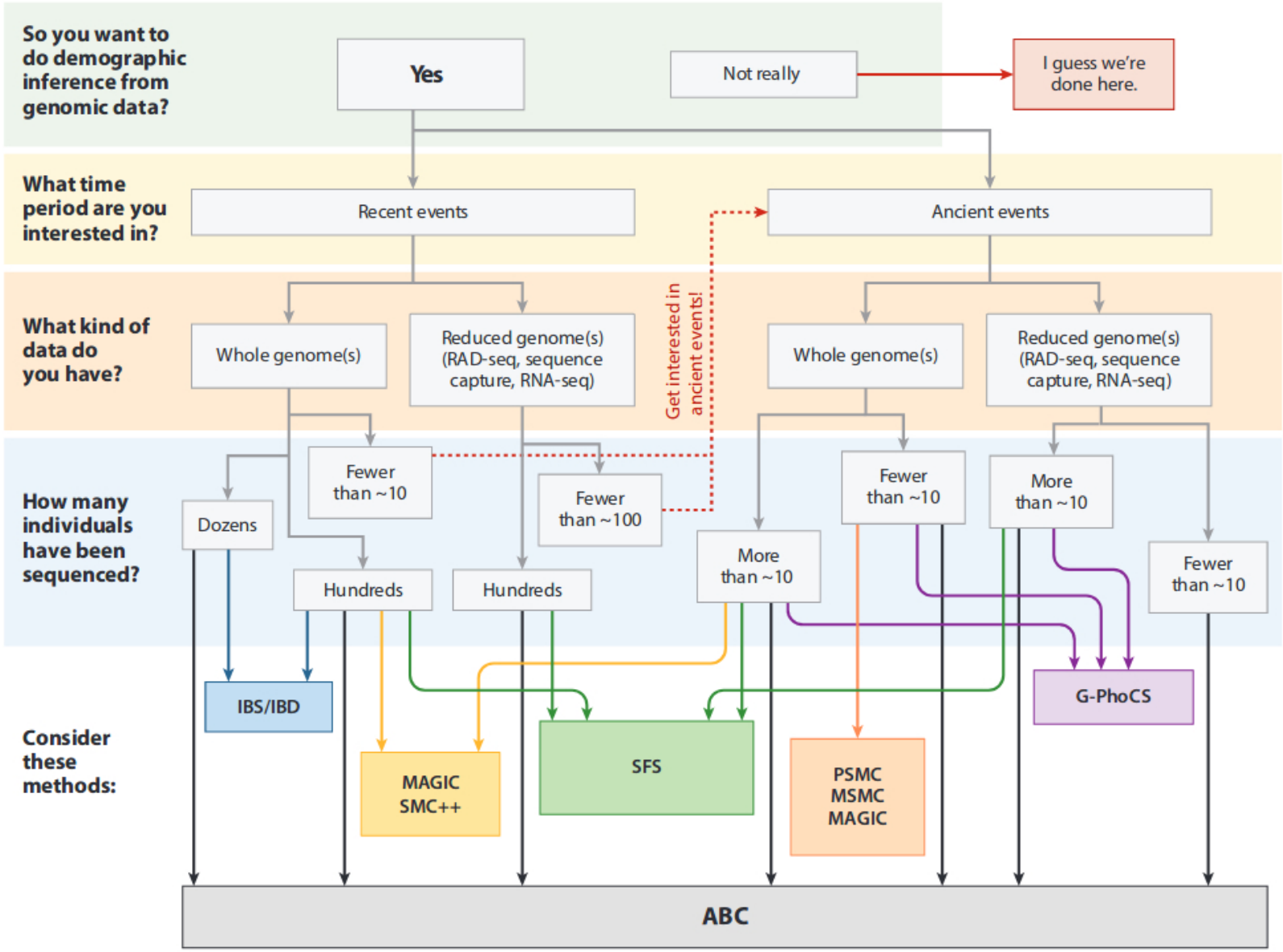

*Leroy et al. 2017 New Phytologist*

## Why it is so important?

- Explain present-day empirical patterns, demographic dynamics
- Past source of adaptive (or maladaptive) variation
- Long-term climatic oscillations (the past, the future?)
- Human impact: before / after Holocene
- Arguments to fight against manipulation & obscurantism



# Many methods available, but depend on your genomic data and research questions!

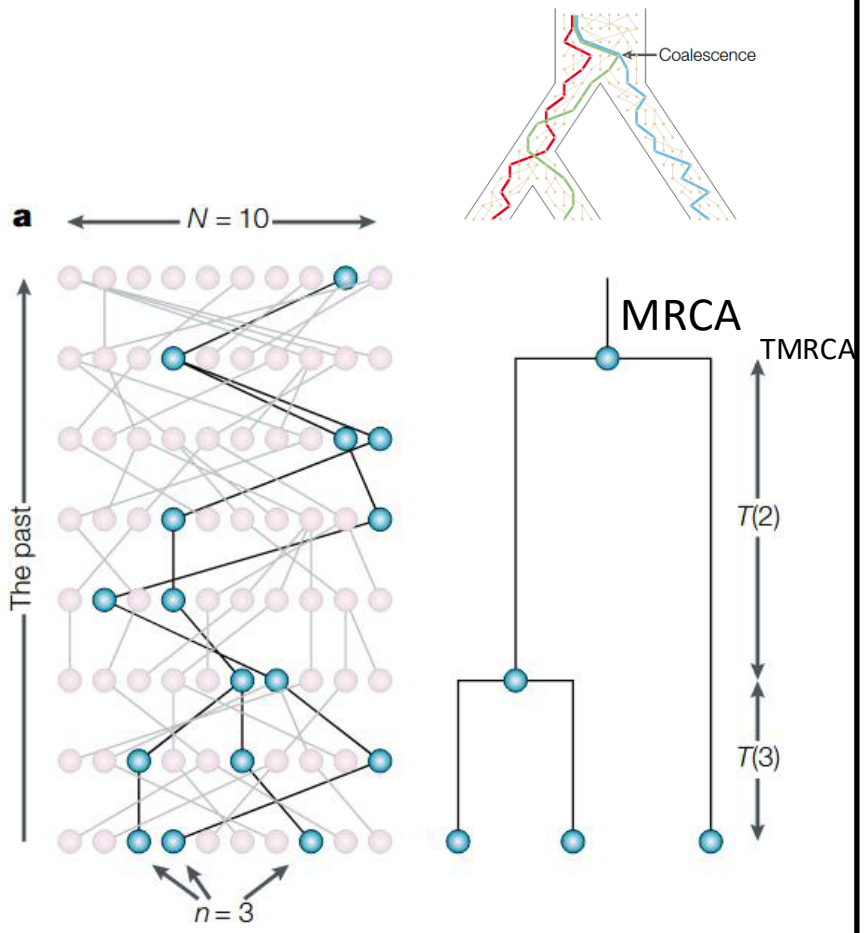



'Recent events' here: <1000 generations ago

*Beichman et al. 2018 Nat Rev Ecol Evol Syst*

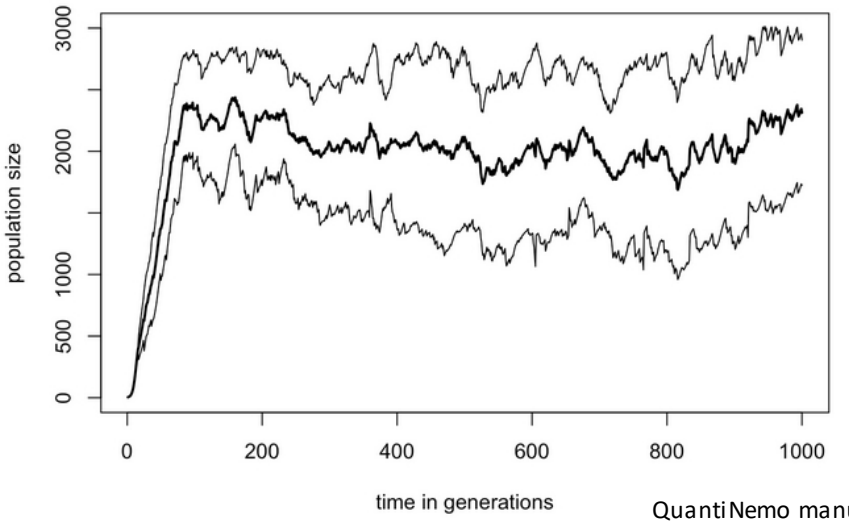
Demographic reconstruction rely on simulation-based inferences

## Backward in time simulations
(i.e. coalescence)




Rosenberg & Nordborg 2002 *Nature Rev Genet*

## Forward in time simulations




QuantiNemo manual

Demogenetic simulators:

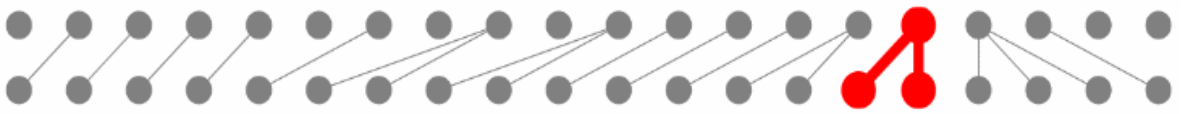
- Nemo/QuantiNemo
- SLiM
- …

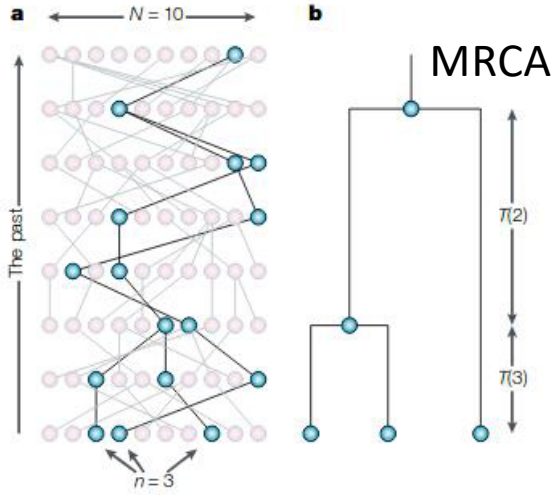
# Coalescent theory

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 $\frac{1}{2Ne}$

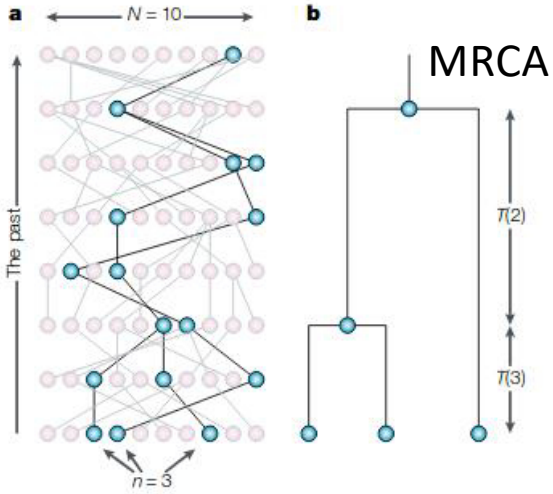
⟶ A direct relationship between time and Ne

# Coalescent theory

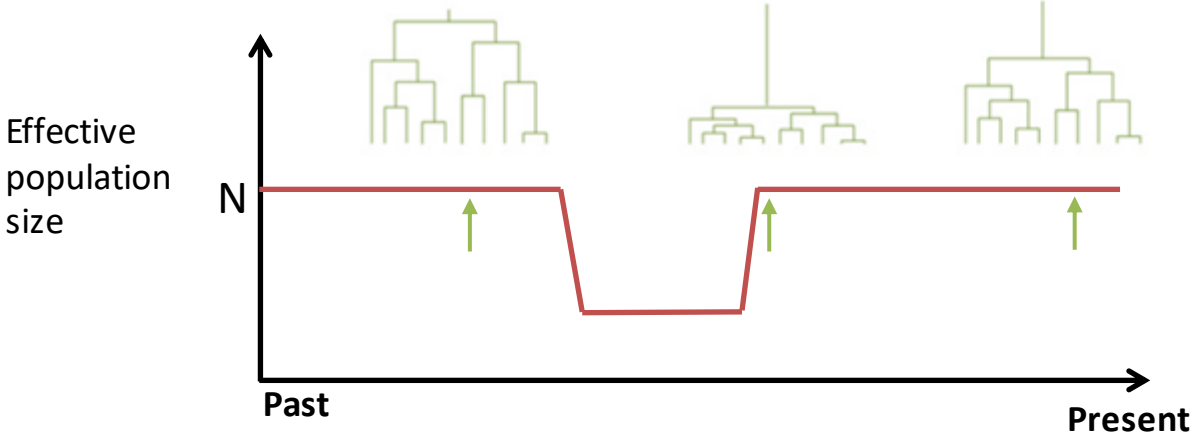
A stochastic process that describes how population genetic processes determine the shape of the genealogy of sampled gene sequences

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 variable population size" coalescent model (Griffiths & Tavaré, 1994; Donnelly & Tavaré, 1995)

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

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

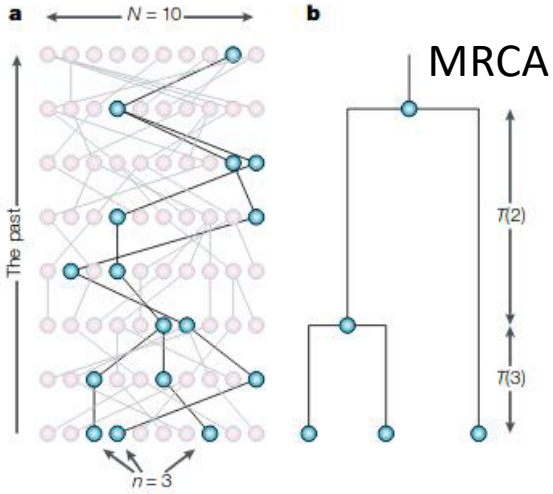
*For example, if we select a few people at random from a small, isolated village, they are likely to share an ancestor in recent generations.*

# Coalescent theory

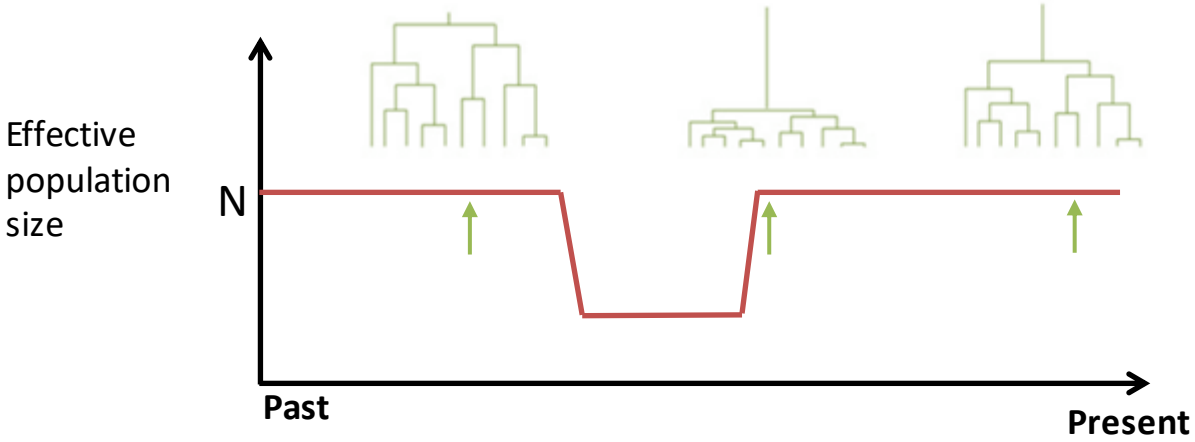
A stochastic process that describes how population genetic processes determine the shape of the genealogy of sampled gene sequences

n individuals sampled from a population of:

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



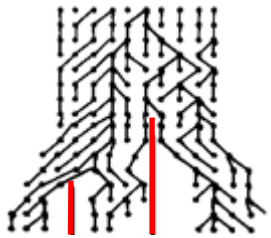

e.g. Rosenberg & Nordborg, 2002




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

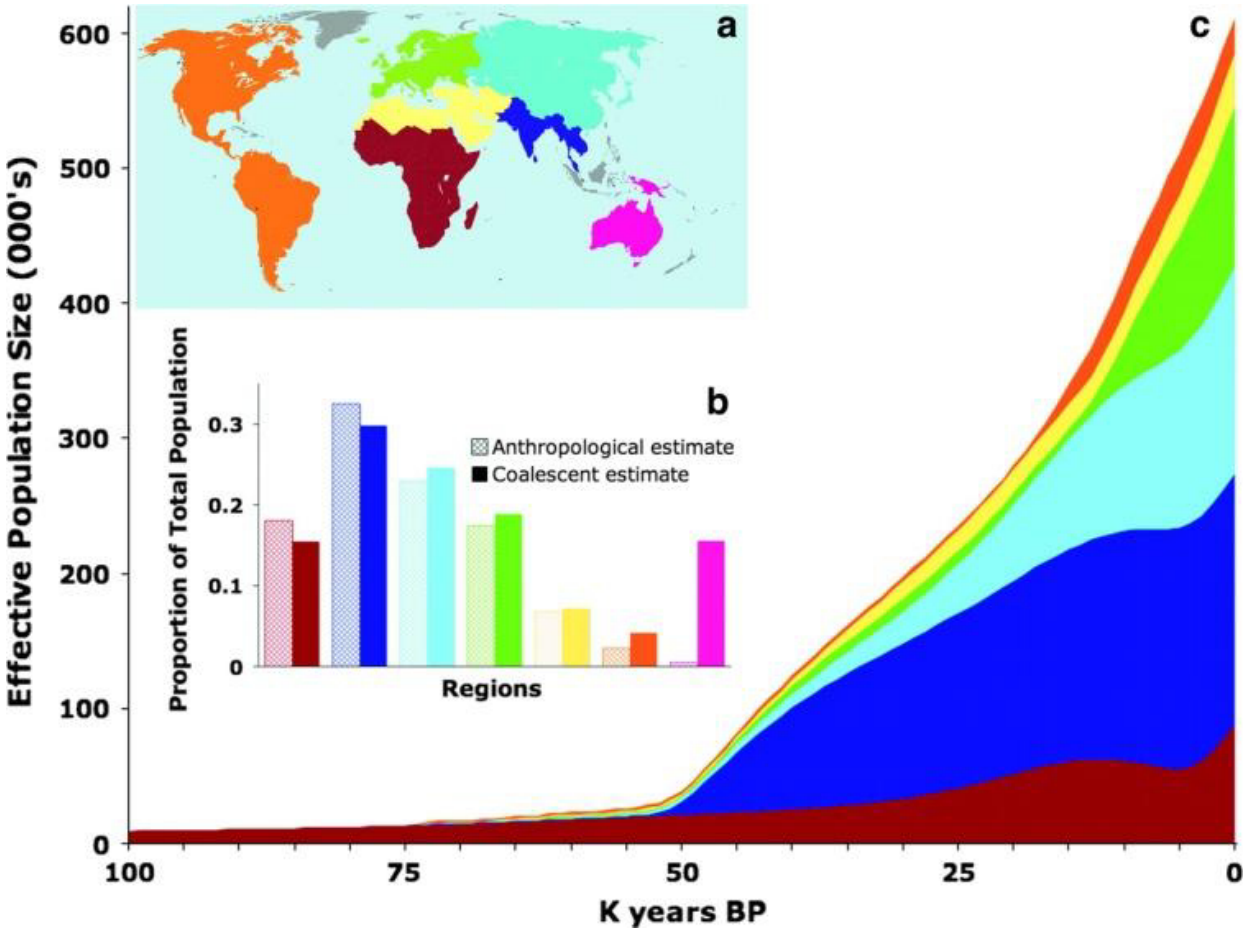
→ Maximum likelihood estimates of parameters (pop expansions/bottlenecks)

Coalescent-based inference methods with population subdivision (e.g. Bahlo and Griffiths 2000; Beerli and Felsenstein 2001)



**Application of "The variable population size" coalescent**

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




(Atkinson et al. 2008 MBE)

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

# Multiple loci / genomes

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

→ But full likelihood methods are not applicable to genome-scale datasets because of two significant limitations:
1) they do not scale well in the number of loci being analyzed
2) they 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)

=> PSMC, …

→ Composite likelihood (dadi, MOMI, fastsimcoal, …)

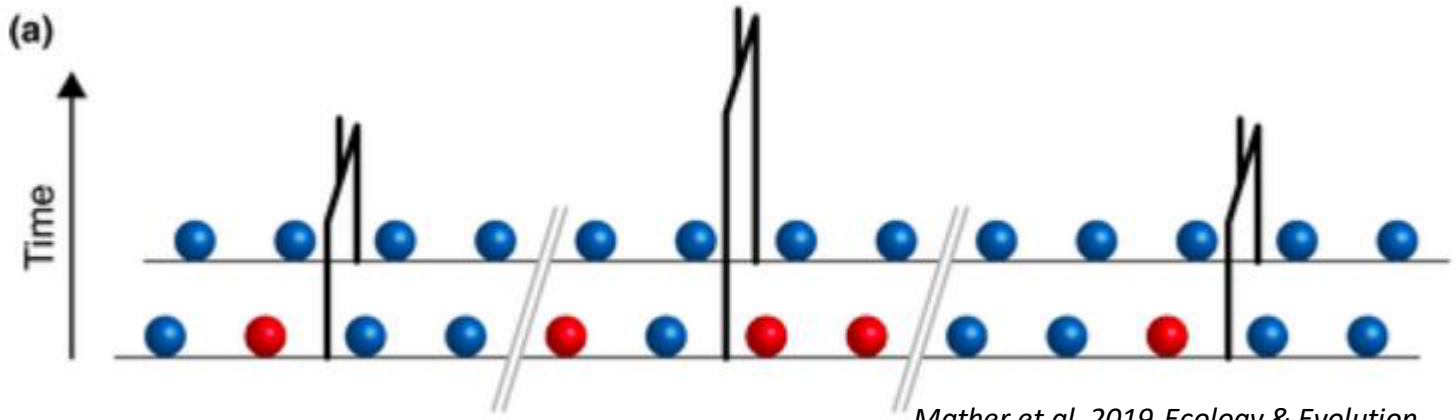
→ Approximate Bayesian computation

## PSMC-like methods: basics

### Pairwise Sequentially Markovian Coalescent (PSMC)

Identification of historical recombination events + Local time to the most recent common ancestor (TMRCA) on the basis of the local density of heterozygotes

**TMRCA between the two alleles carried by an individual**



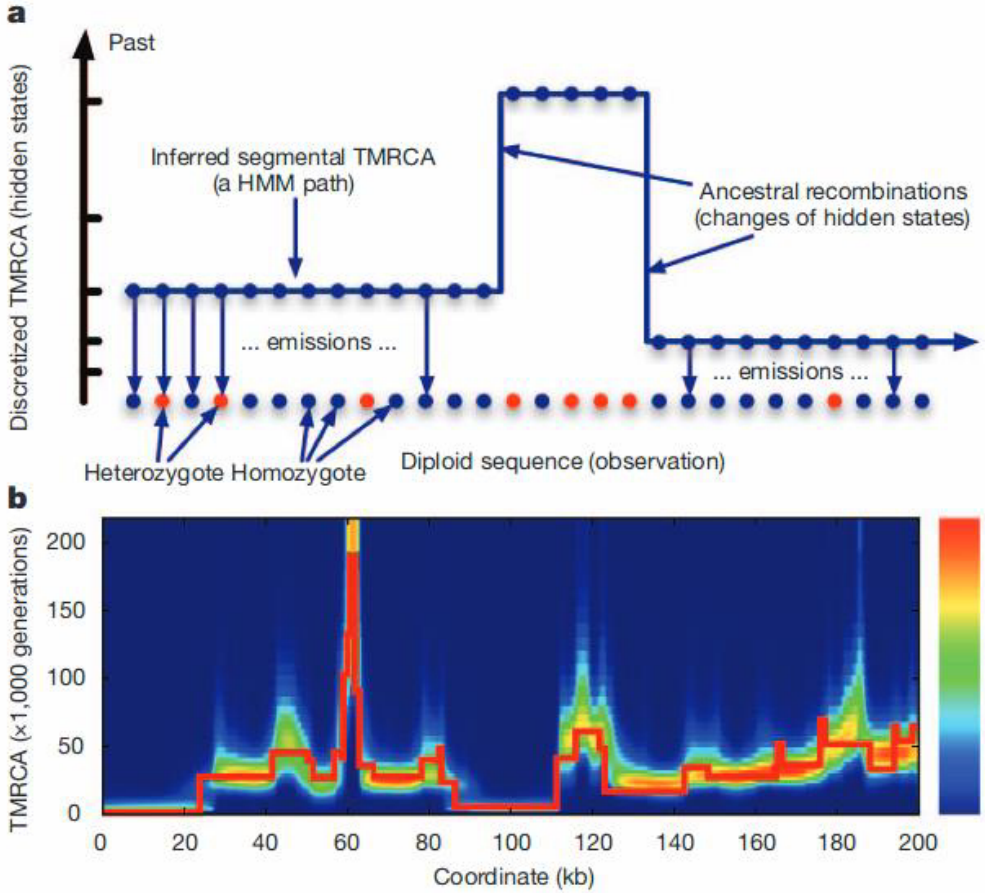

*Mather et al. 2019 Ecology & Evolution*

# PSMC-like methods: basics

## Pairwise Sequentially Markovian Coalescent (PSMC)

Identification of historical recombination events + Local time to the most recent common ancestor (TMRCA) on the basis of the local density of heterozygotes

**TMRCA between the two alleles carried by an individual**



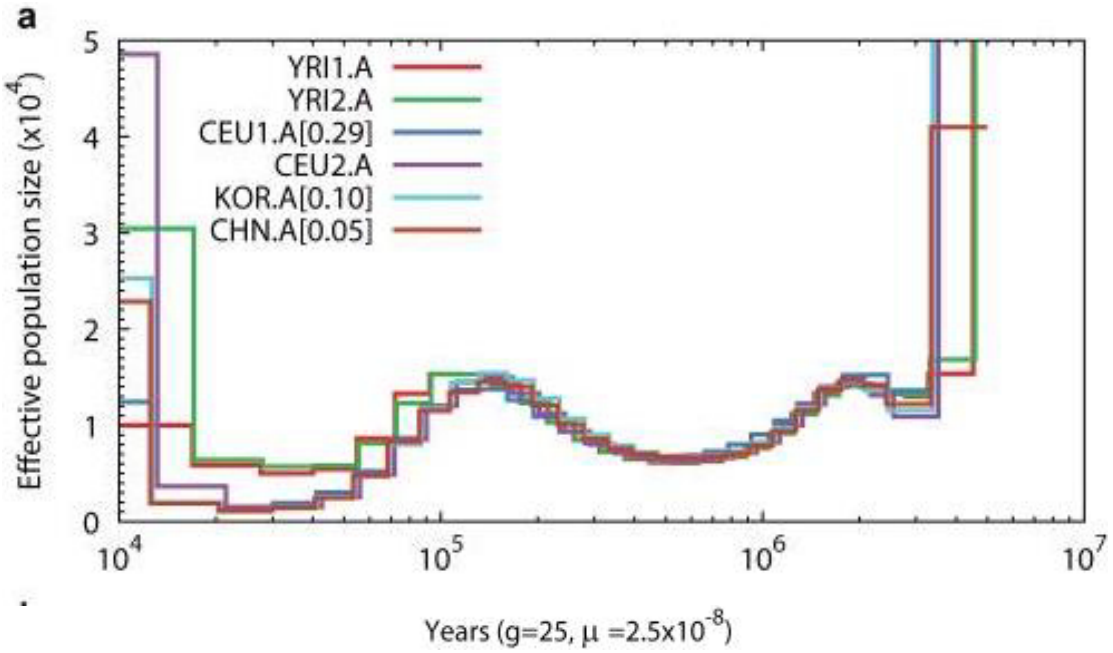

Estimating TMRCA of the two alleles at each locus is used to create a TMRCA distribution across the genome. **And since the rate of coalescent events is inversely proportional to *Ne*, PSMC identifies periods of *Ne* change.** For example, when many loci coalesce at the same time, it is a sign of small *Ne* at that time.

*Li & Durbin, 2011 Nature*

## PSMC-like methods: basics

### Pairwise Sequentially Markovian Coalescent (PSMC)

Identification of historical recombination events + Local time to the most recent common ancestor (TMRCA) on the basis of the local density of heterozygotes



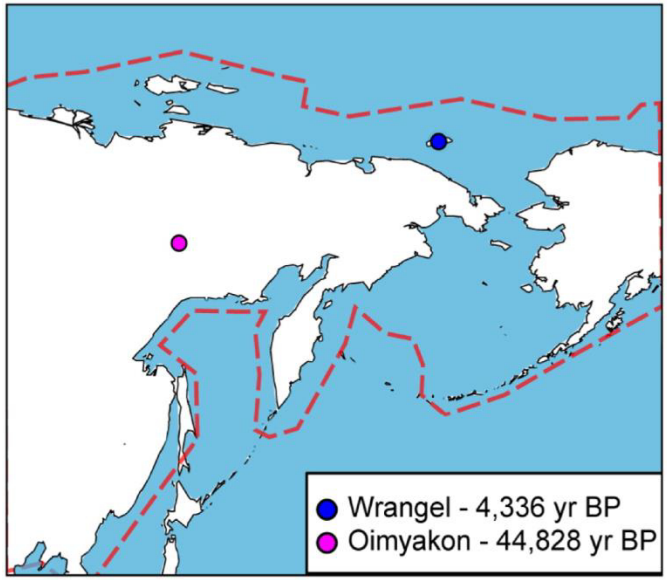

'Skyline plots'

*Li & Durbin, 2011 Nature*

# PSMC-like methods: basics

## Reconstruction of the demographic history of an extinct species (ancient DNA)

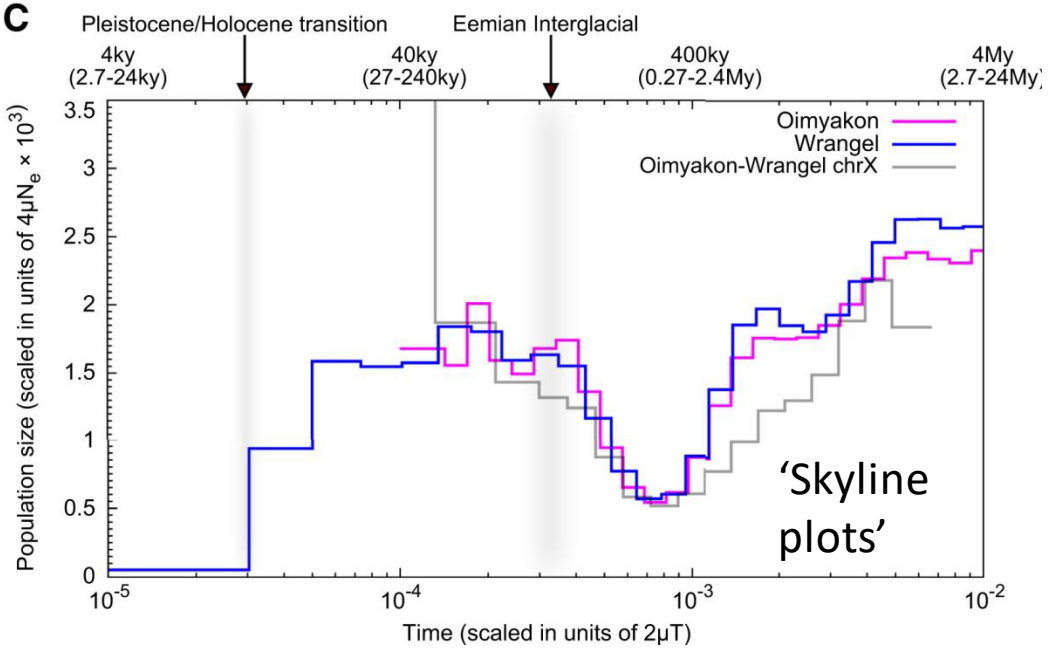
**A**




**B**

| Sample | $^{14}$C date ± error (years) | Median calibrated date (years) | # raw reads ($\times 10^6$) | Average coverage | Average read length (bp) |
|---|---|---|---|---|---|
| Wrangel | 3,905 ± 47 | 4,336 | 1,262 | 17.1 | 69 |
| Oimyakon | 41,300 ± 900 | 44,828 | 1,401 | 11.2 | 55 |

**C**




*Woolly mammoth, Rouffignac Cave, France*

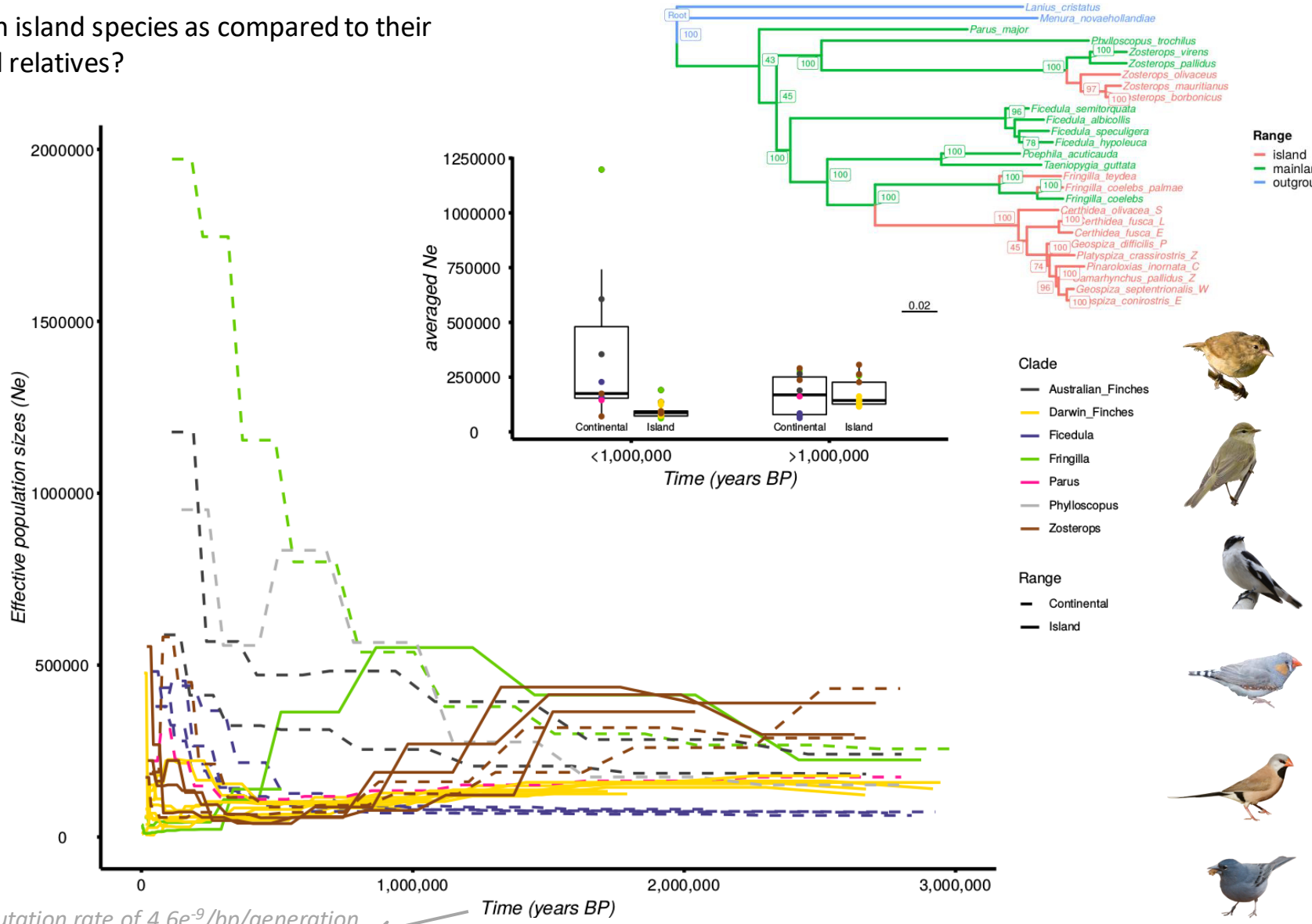
Coalescence times -> real times
(require estimates of the mutation rates and generation time)

*Palkopoulou et al. 2015, Current Biol*

# PSMC-like methods: basics

## Investigating the demo. of many species (« phylogenetically-oriented » sequencing projects)

Lower *Ne* in island species as compared to their continental relatives?



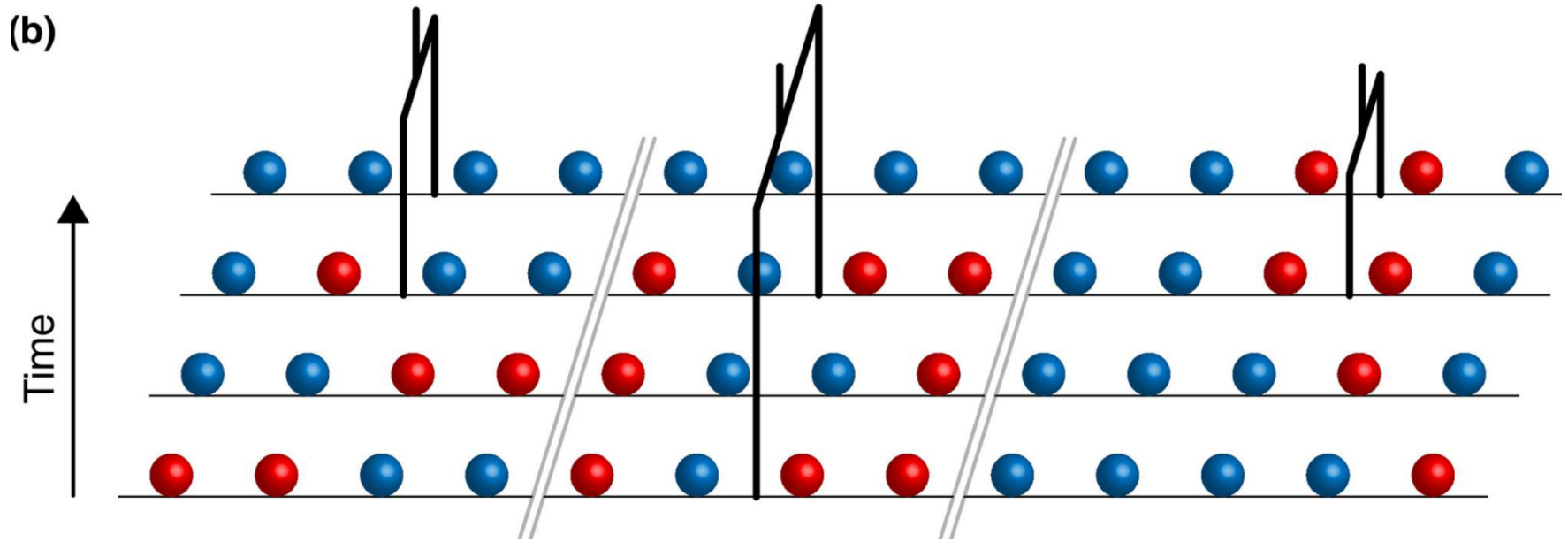

*Assume a mutation rate of $4.6e^{-9}$/bp/generation and a generation time of 2 years*

*Leroy et al. 2020, in revision*

## PSMC-like methods: basics

### Multiple Sequentially Markovian Coalescent (MSMC)

Identification of historical recombination events + Local time to the first event of coalescence among all samples



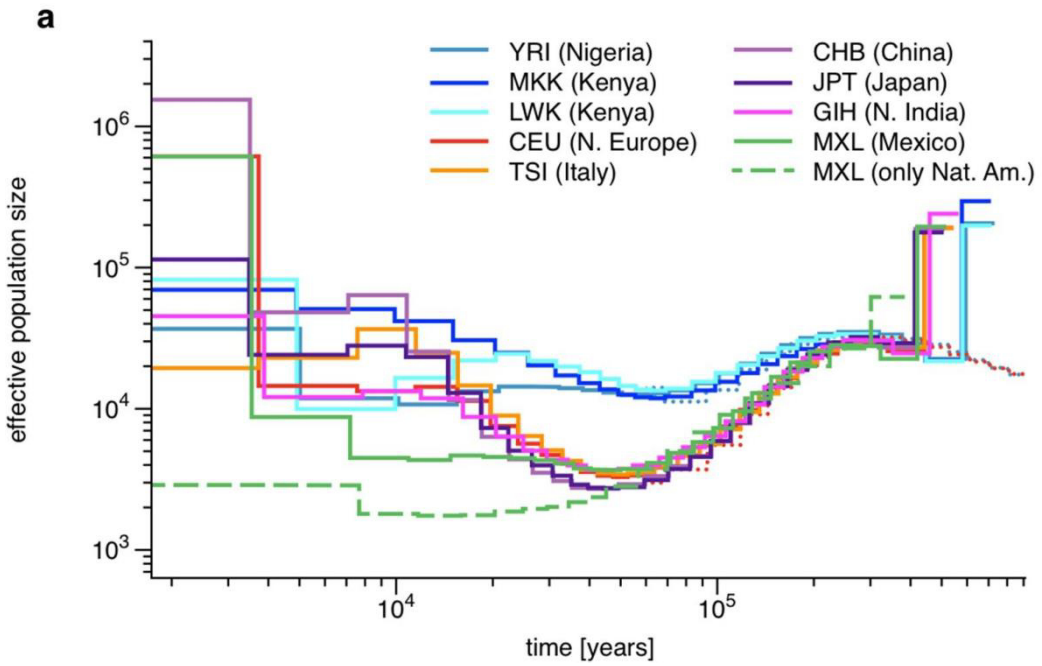

"When multiple genomes are available, MSMC has better power to resolve recent changes in effective population size, because adding alleles increases the chance that there will be a coalescence event in the recent past. "

*Mather et al. 2019 Ecology & Evolution*

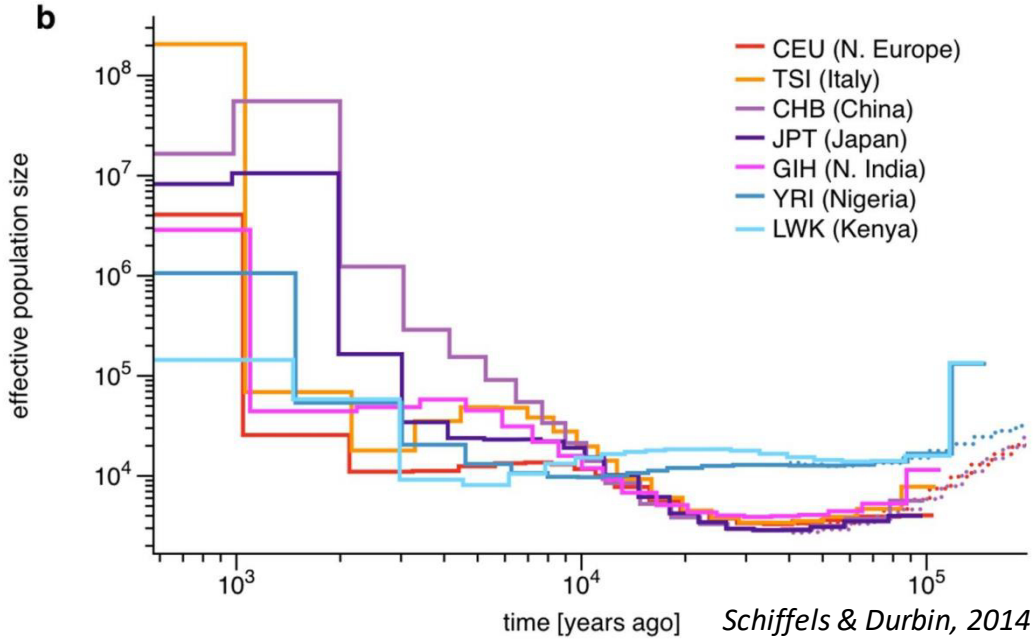
## PSMC-like methods: basics

**Multiple Sequentially Markovian Coalescent (MSMC)**

2 individuals per pop




8 individuals per pop




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

*Schiffels & Durbin, 2014 Nature Genetics*

## PSMC-like methods: basics

### Multiple Sequentially Markovian Coalescent (MSMC)

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

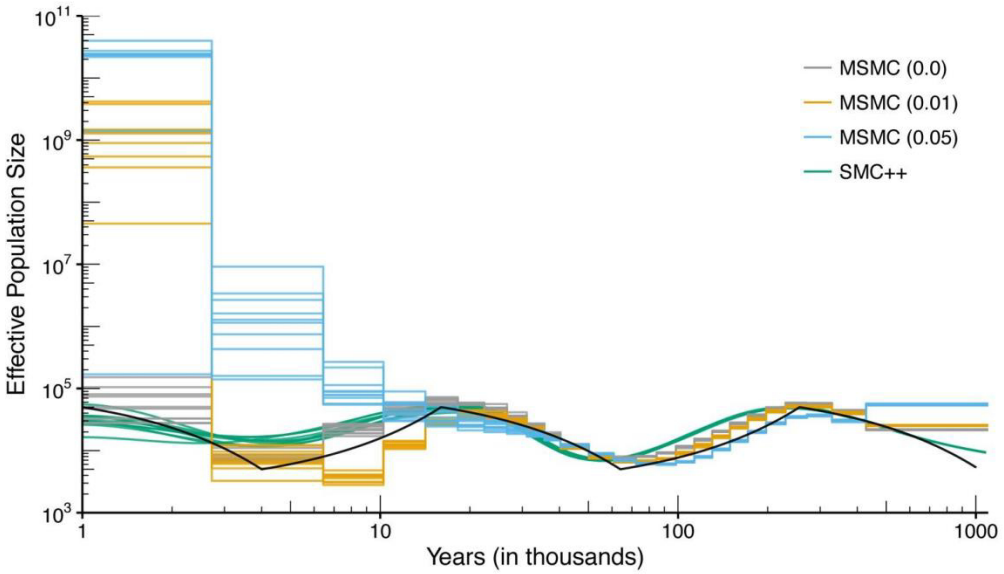
| | | | | | | |
|---|---|---|---|---|---|---|
| A | T<br>G | G | C<br>G | A<br>G | G | |

*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…

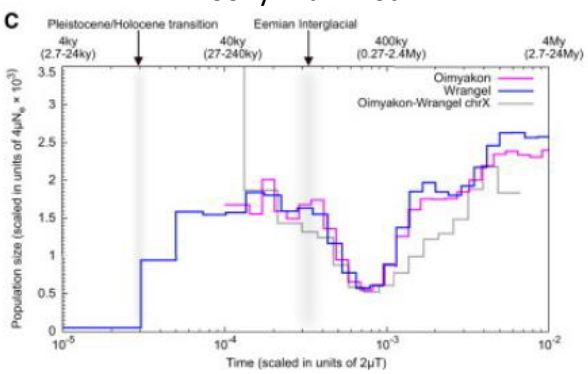



Some other methods using unphased data are becoming popular to overpass this problem (e.g. smc++,PopSizeABC) but requires tens of samples…

*Terhorst et al. 2016 Nature Genetics; Boitard et al. 2016 Plos Genetics*

## PSMC-like methods: pros & cons

Woolly Mammoth



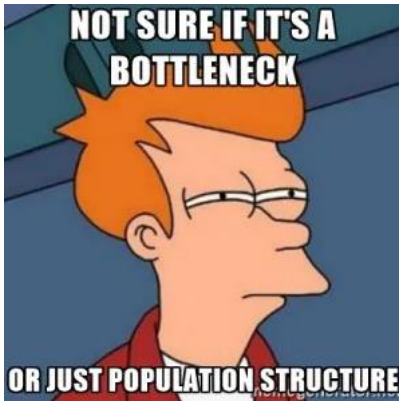

*Palkopoulou et al. 2015, Current Biol*

**Advantages:**

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

**Limitations:**

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




*moleculareecologist.com*

- PSMC estimates for recent times (<10kyrs) are rarely accurate
- Sensitive to the quality of the genome assembly (e.g. N%, scaffold length)
- Sensitive to the quality of the data
  To make reliable demographic, Nadachowska-Brzyska et al. (2016) suggested filters :
  e.g. A mean genome-wide coverage of at least 18X, no more than 25% of missing data, …
- Problem of rescaling to real time for non-model species (incorrect mutation rates or generation times)
- Doesn't recover sudden changes in Ne or very ancient changes

## Composite likelihood methods: basics

We have previously seen that full-likelihood methods are not well adapted to nuclear genome data because of the volume of data and the recombination events.

One way to solve the problem was to simplify this genetic information to summary statistics describing the dataset and to compute the likelihood only based on these summary statistics (*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

## Composite likelihood methods: basics

We have previously seen that full-likelihood methods are not well adapted to nuclear genome data because of the volume of data and the recombination events.

One way to solve the problem was to simplify this genetic information to summary statistics describing the dataset and to compute the likelihood only based on these summary statistics (*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

**Composite likelihood methods: basics**

We have previously seen that full-likelihood methods are not well adapted to nuclear genome data because of the volume of data and the recombination events.

One way to solve the problem was to simplify this genetic information to summary statistics describing the dataset and to compute the likelihood only based on these summary statistics (*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 |

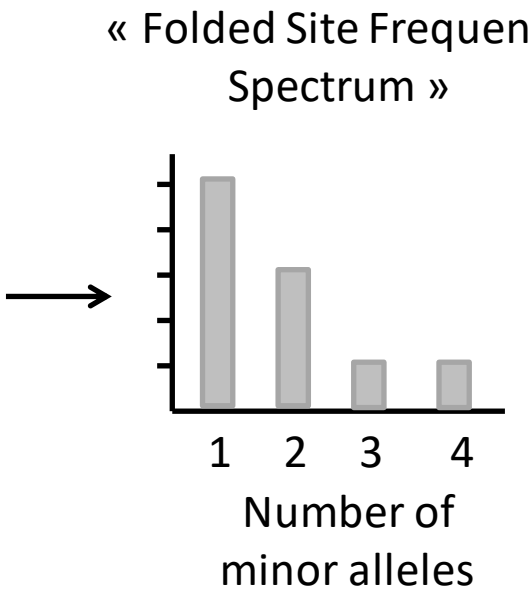
**Composite likelihood methods: basics**

We have previously seen that full-likelihood methods are not well adapted to nuclear genome data because of the volume of data and the recombination events.

One way to solve the problem was to simplify this genetic information to summary statistics describing the dataset and to compute the likelihood only based on these summary statistics (*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 |
|---|---|---|---|---|---|---|---|---|---|---|



**Composite likelihood methods: basics**

We have previously seen that full-likelihood methods are not well adapted to nuclear genome data because of the volume of data and the recombination events.

One way to solve the problem was to simplify this genetic information to summary statistics describing the dataset and to compute the likelihood only based on these summary statistics (*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)

**Composite likelihood methods: basics**

We have previously seen that full-likelihood methods are not well adapted to nuclear genome data because of the volume of data and the recombination events.

One way to solve the problem was to simplify this genetic information to summary statistics describing the dataset and to compute the likelihood only based on these summary statistics (*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)

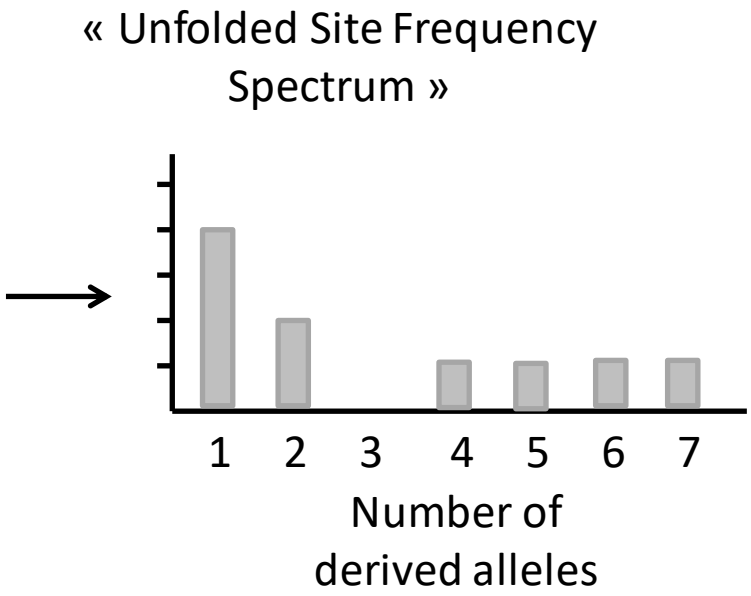
**Composite likelihood methods: basics**

We have previously seen that full-likelihood methods are not well adapted to nuclear genome data because of the volume of data and the recombination events.

One way to solve the problem was to simplify this genetic information to summary statistics describing the dataset and to compute the likelihood only based on these summary statistics (*i.e.* a « composite » likelihood)

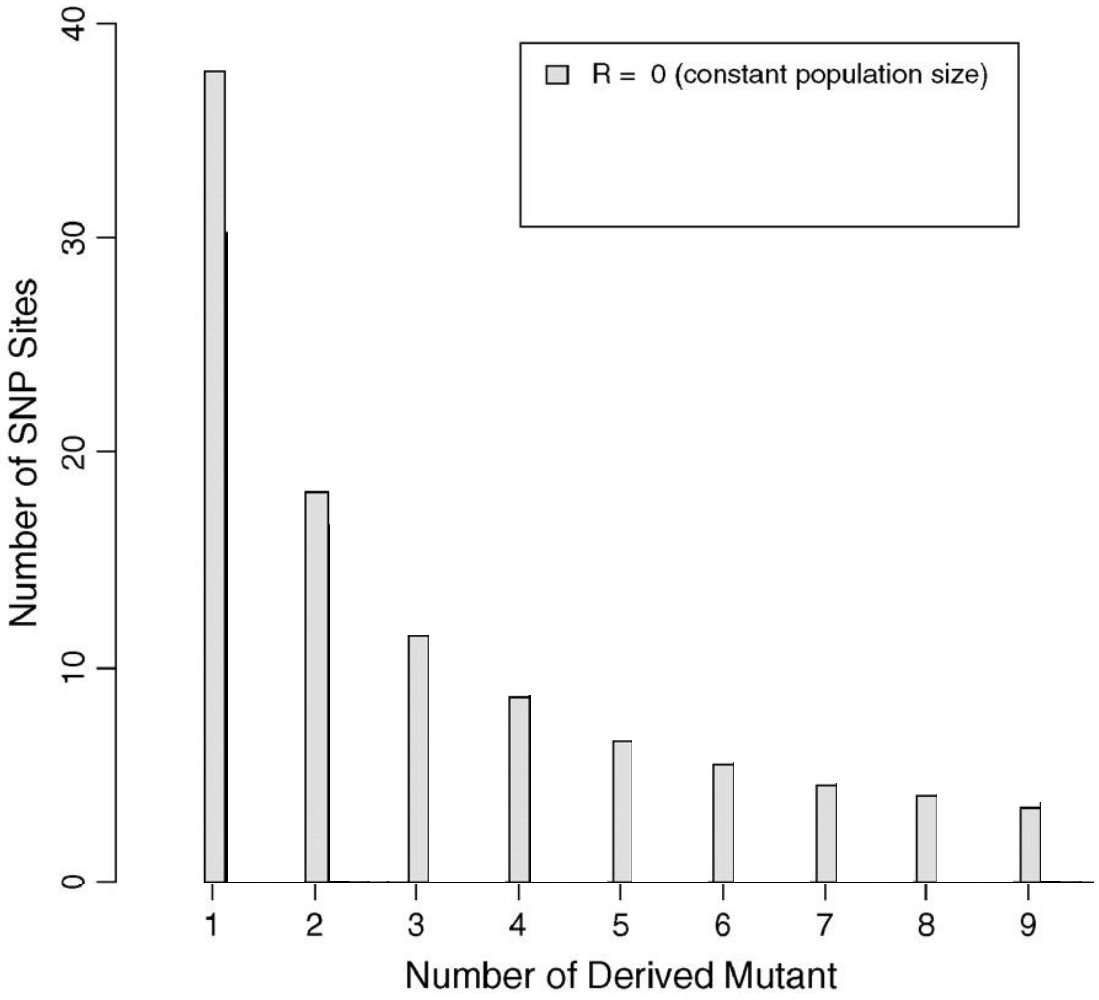
```
Anc-A1:  A..G..G..T..T..C..A..A..C..G
Anc-A2:  A..G..G..T..T..C..A..A..C..G
Ind1-A1: A..C..G..A..T..G..A..A..T..G
Ind1-A2: A..G..G..A..T..G..T..A..C..G
Ind2-A1: T..G..G..T..T..C..T..A..T..G
Ind2-A2: A..G..G..A..G..C..T..A..C..G
Ind3-A1: A..G..A..A..T..G..T..A..T..G
Ind3-A2: A..C..G..A..T..C..T..G..T..A
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 |



# Composite likelihood methods: basics

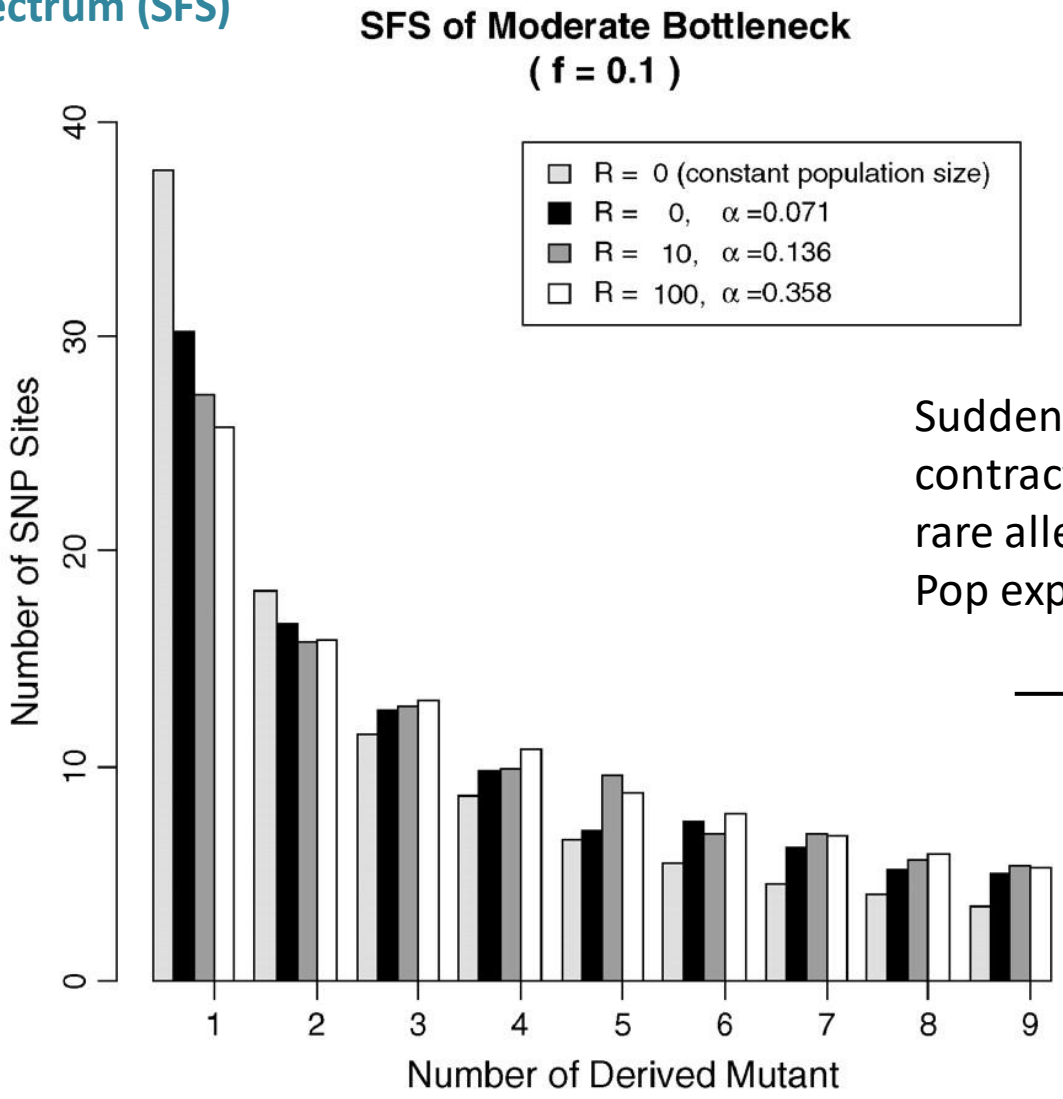
## Site Frequency Spectrum (SFS)




Zhu & Bustamante, 2005

## Composite likelihood methods: basics

### Site Frequency Spectrum (SFS)



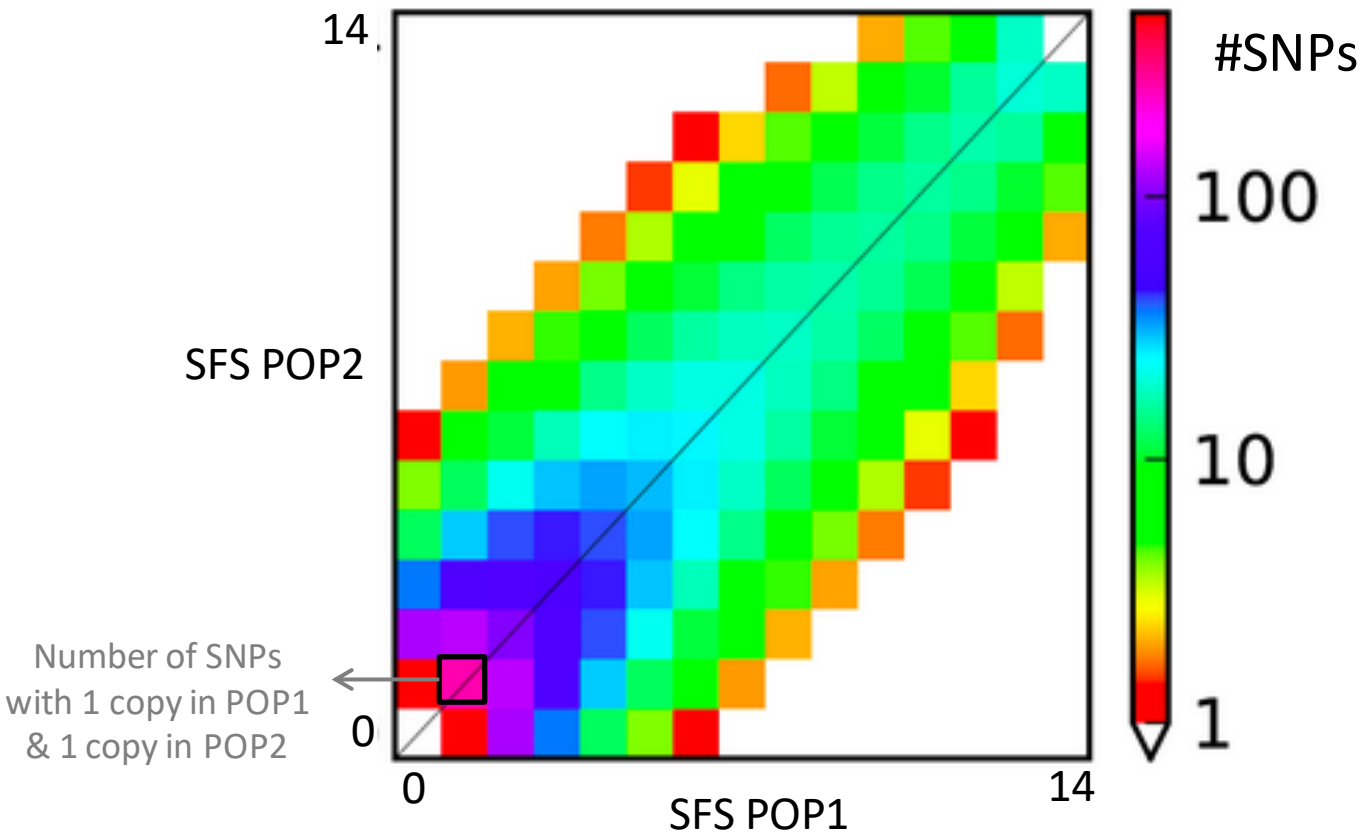

Sudden population contraction -> deficit of rare alleles
Pop expansion -> excess

→ SFS are therefore informative about *Ne* changes

Zhu & Bustamante, 2005

# Composite likelihood methods: basics

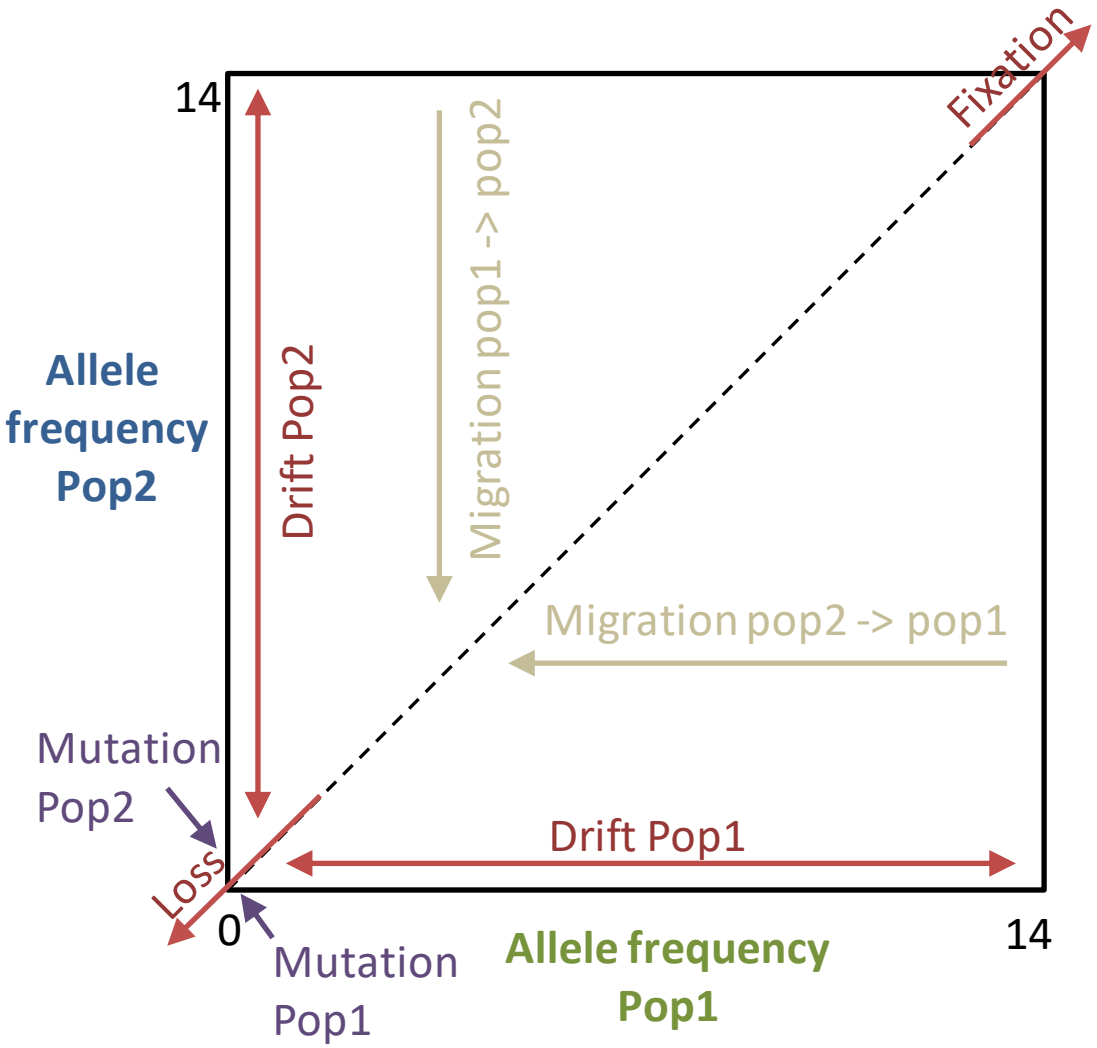
## Joint Site Frequency Spectrum (2D-SFS)




Adapted from Gutenkunst et al. 2009 Plos Genet

# Composite likelihood methods: basics

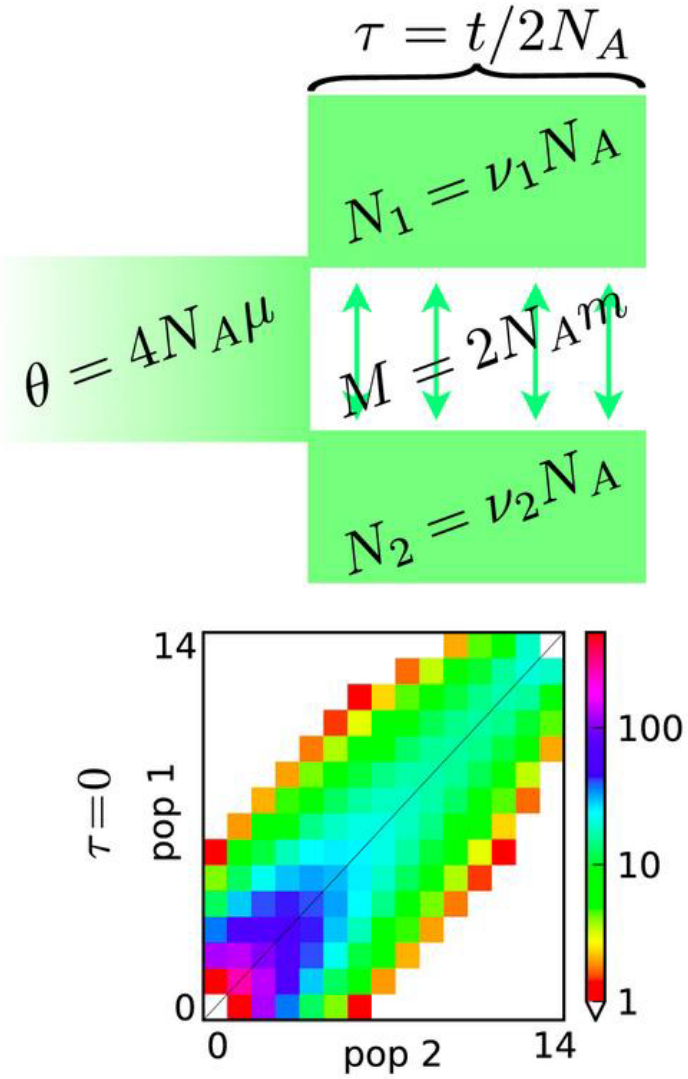
## Joint Site Frequency Spectrum (2D-SFS)

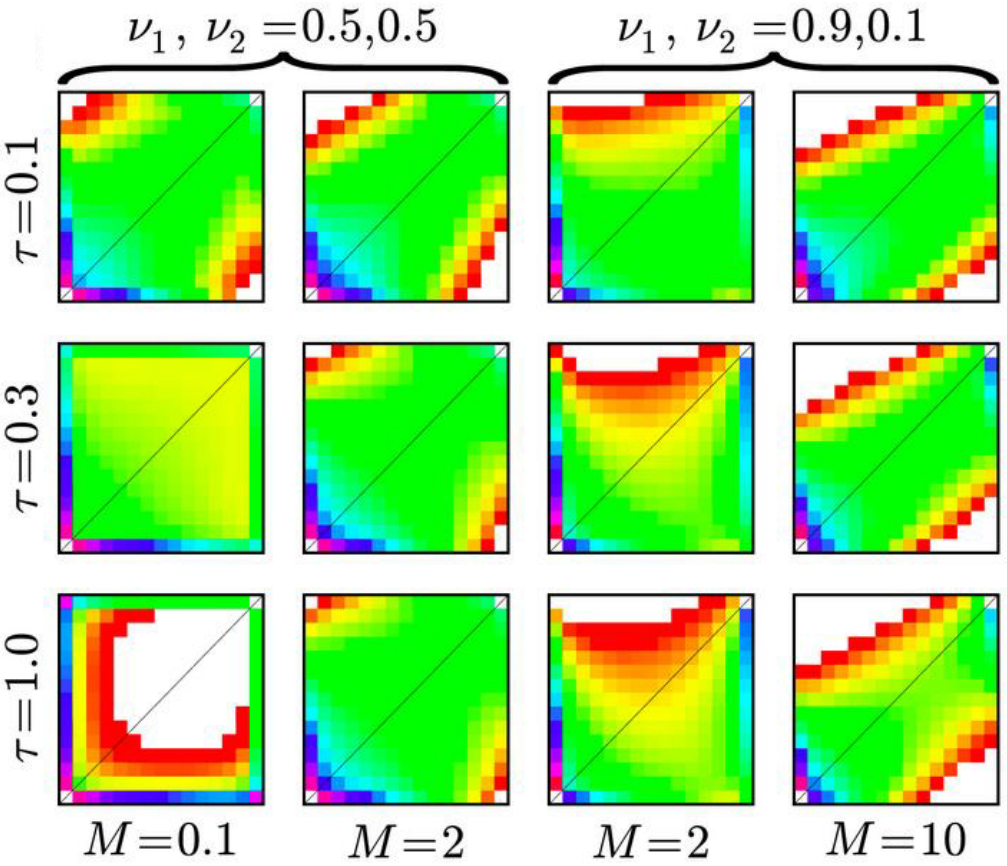



Adapted from Gutenkunst et al. 2009 Plos Genet

**Composite likelihood methods: basics**

## Simulated 2D-SFS



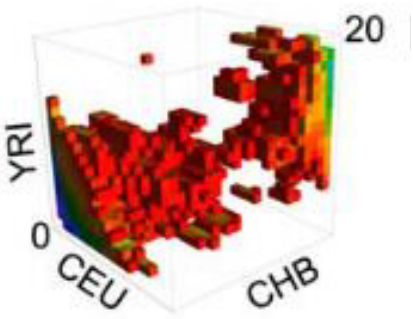

Gutenkunst et al. 2009 Plos Genet

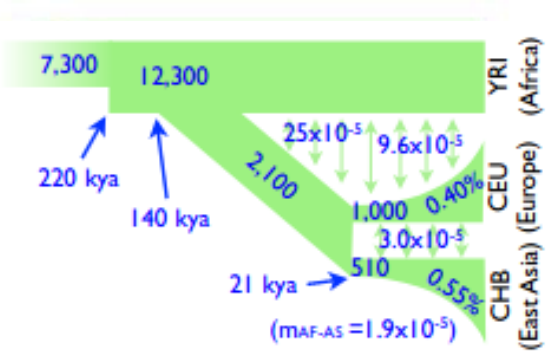
## Composite likelihood methods: basics

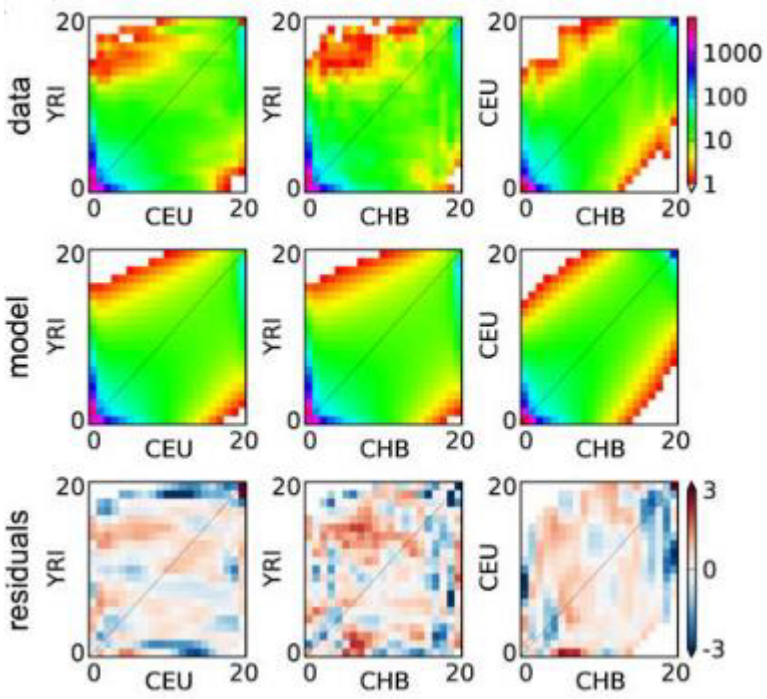
### 3-dimensional SFS

The implementation (∂a∂i program) is quite flexible. It was initially able to handle up to three simultaneous populations

Data: 1000 genomes project (human)




YRI: Yoruba, Nigeria (AFR)
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)

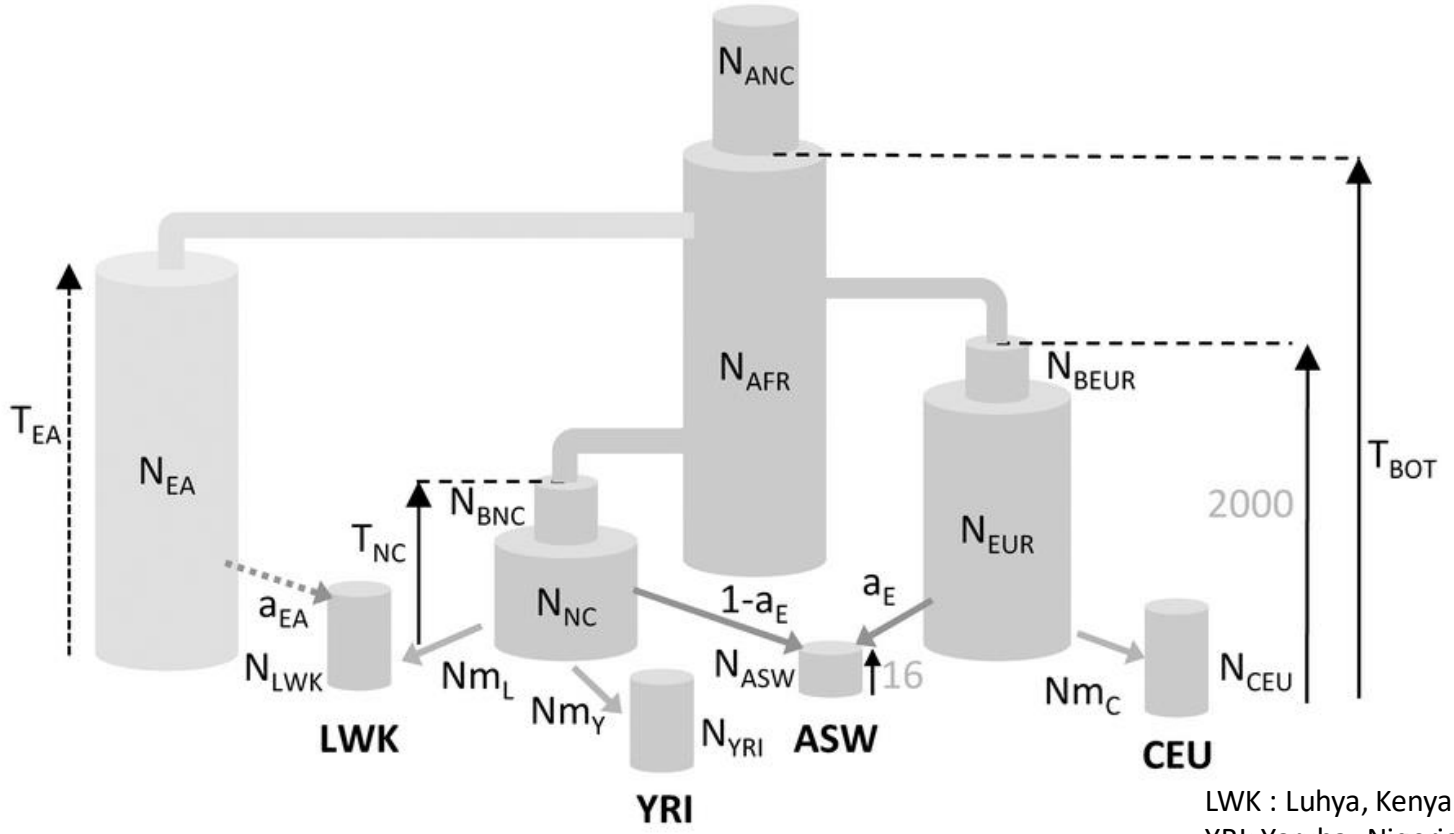
*Gutenkunst et al. 2009 Plos Genetics, 2010 Nature Precedings*

## Composite likelihood methods: basics

### Multiple pairwise joint SFS

fastsimcoal2 is another very popular tool

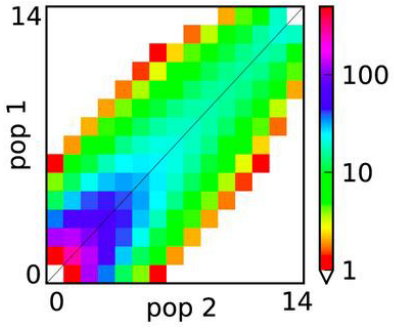
Using a multiple paiwise joint SFS strategy, fastsimcoal2 is (in theory) be able to infer demography of an arbitrary number of populations




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 methods: 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:**

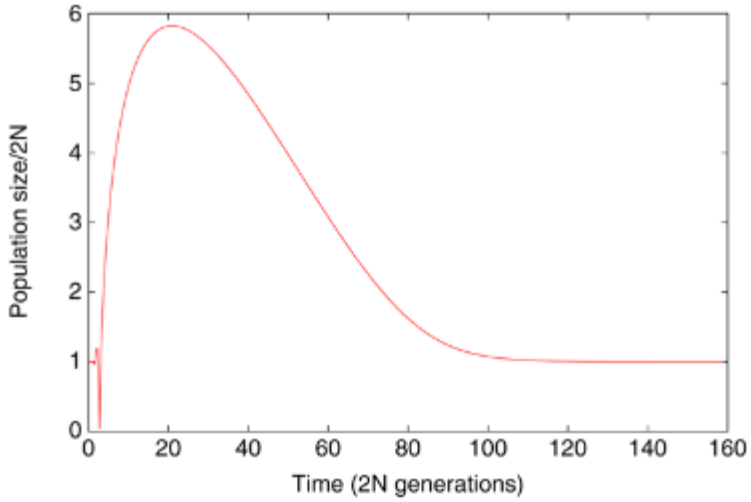
Computational issues

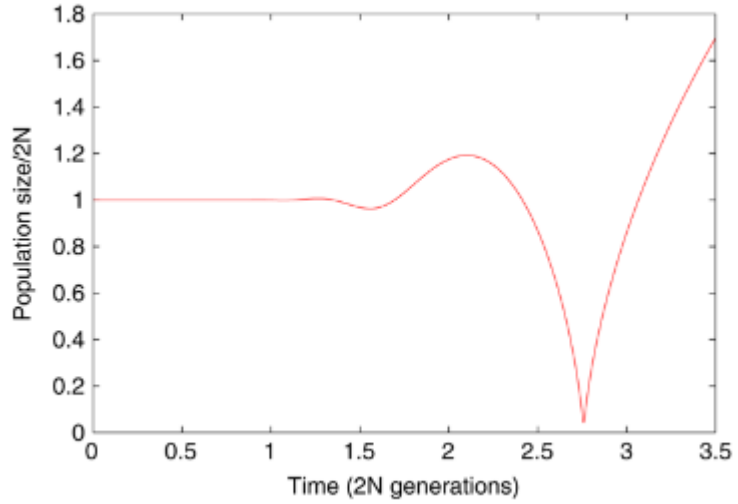
- Convergence problems are possible

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 model-based methods!)
- Correct parameter estimates are challenging
- limitations on how informative allelic spectra can be

## Two demographic histories with the same spectrum as a constant size populations




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

## Approximate Bayesian Computation in practice

Likelihood-free demographic inferences

**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, FST between pairs of populations, 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)

**2/ Demographic models**

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/ Posterior distributions of the parameters**

*Leroy et al. 2014 IGE*

# Approximate Bayesian Computation in practice

Likelihood-free demographic inferences

**The rationale:**

**Real Data**

100 loci x 1kb

↓

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

**Simulations**

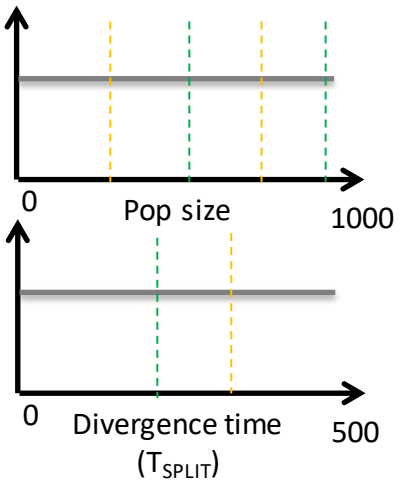
Model 1

(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

…

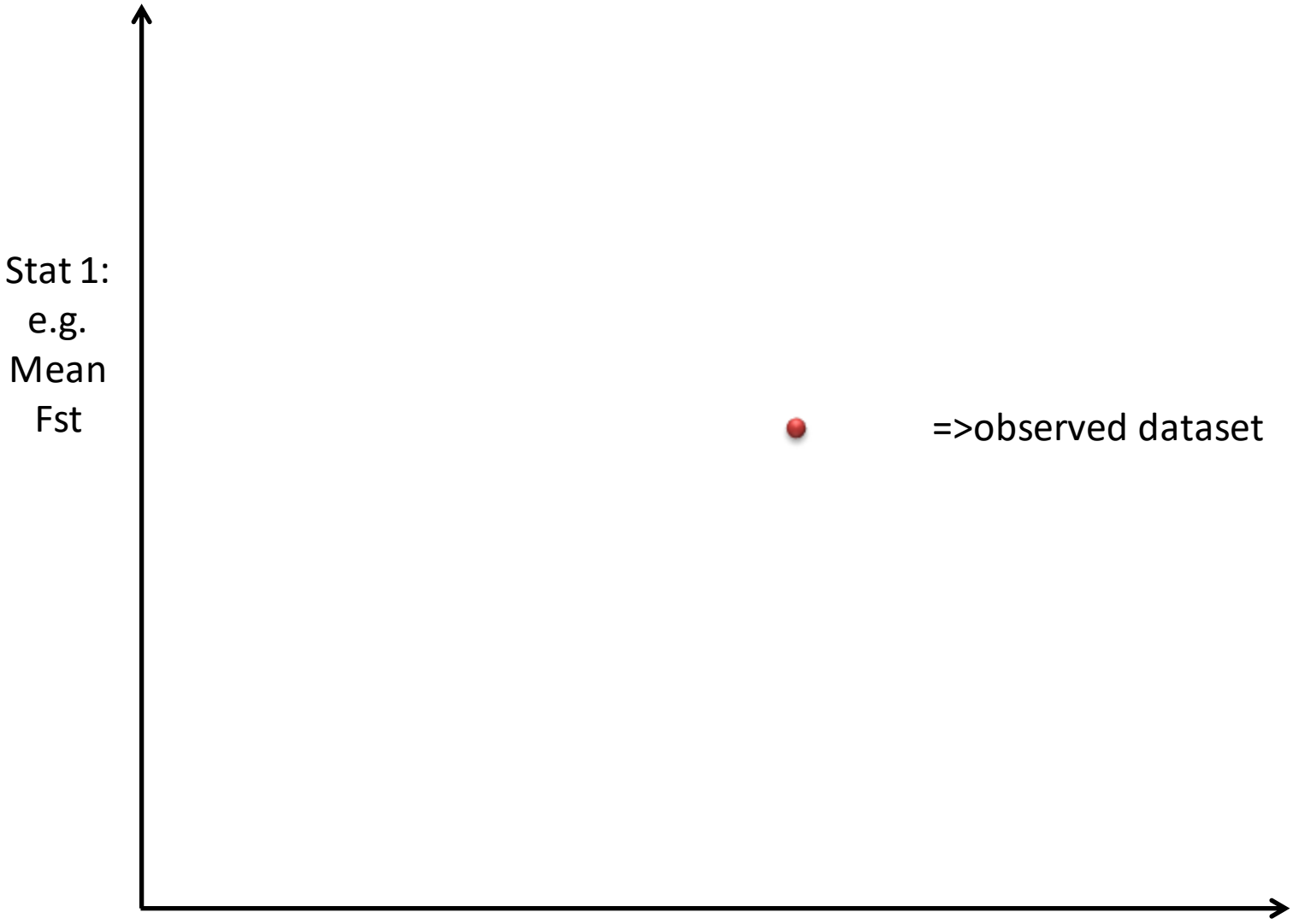
Same summary statistics

**Simulating large numbers of datasets** under several hypothesized evolutionary scenarios

Data generated by simulation are then **reduced to summary statistics**

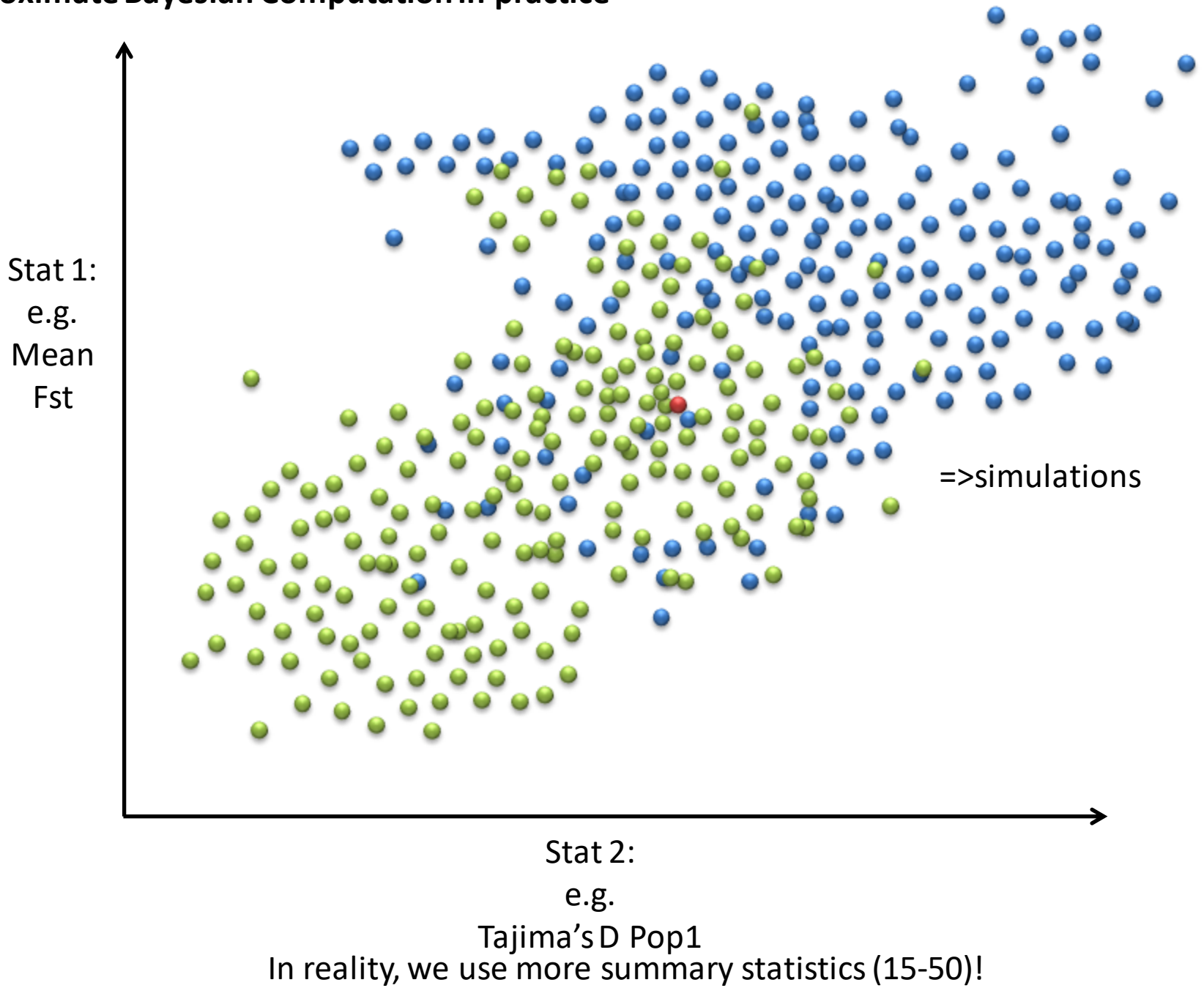
**Compute the Euclidian distance** between the simulated and the observed summary statistics

**Approximate Bayesian Computation in practice**



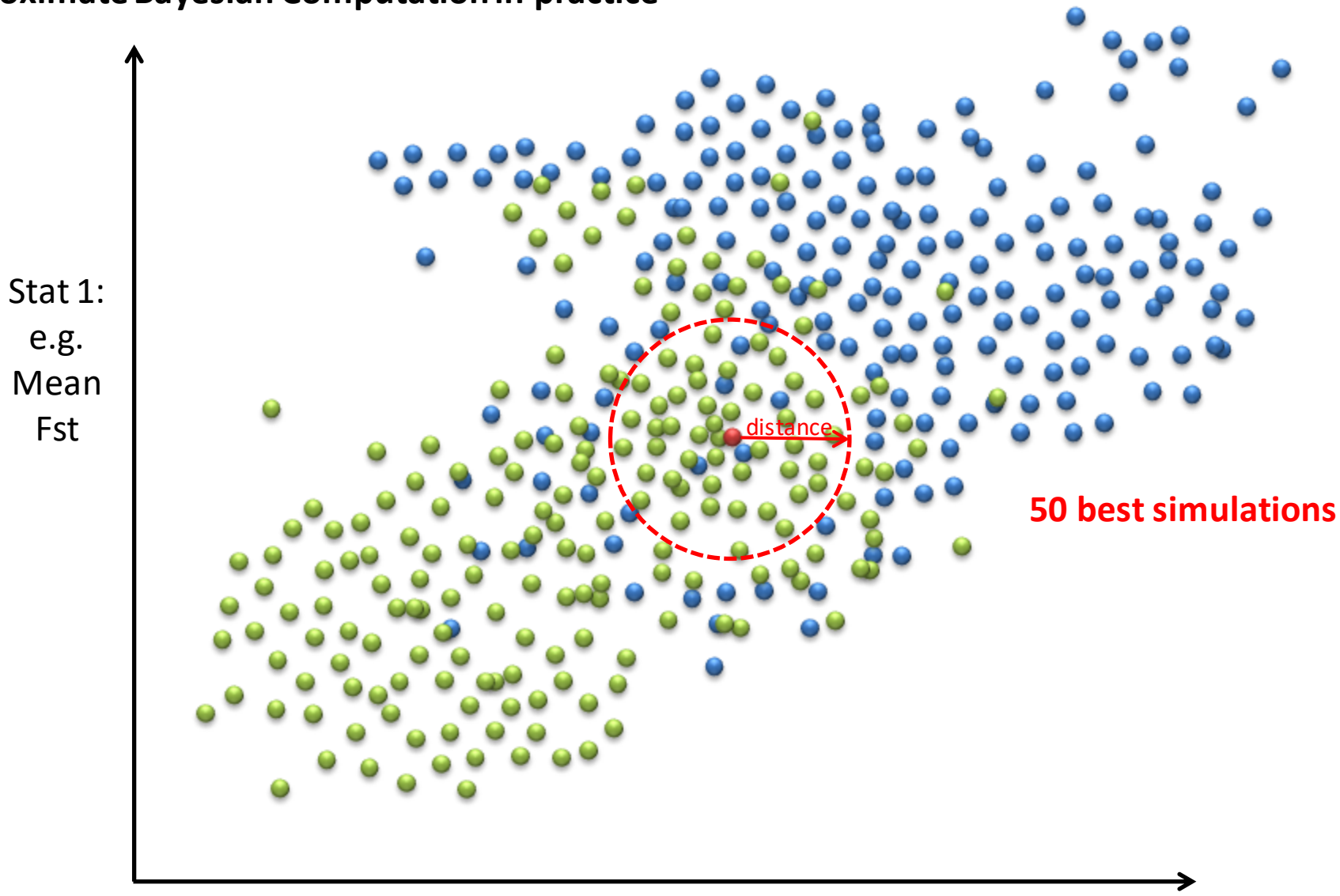

In reality, we use more summary statistics (15-50)!

**Approximate Bayesian Computation in practice**



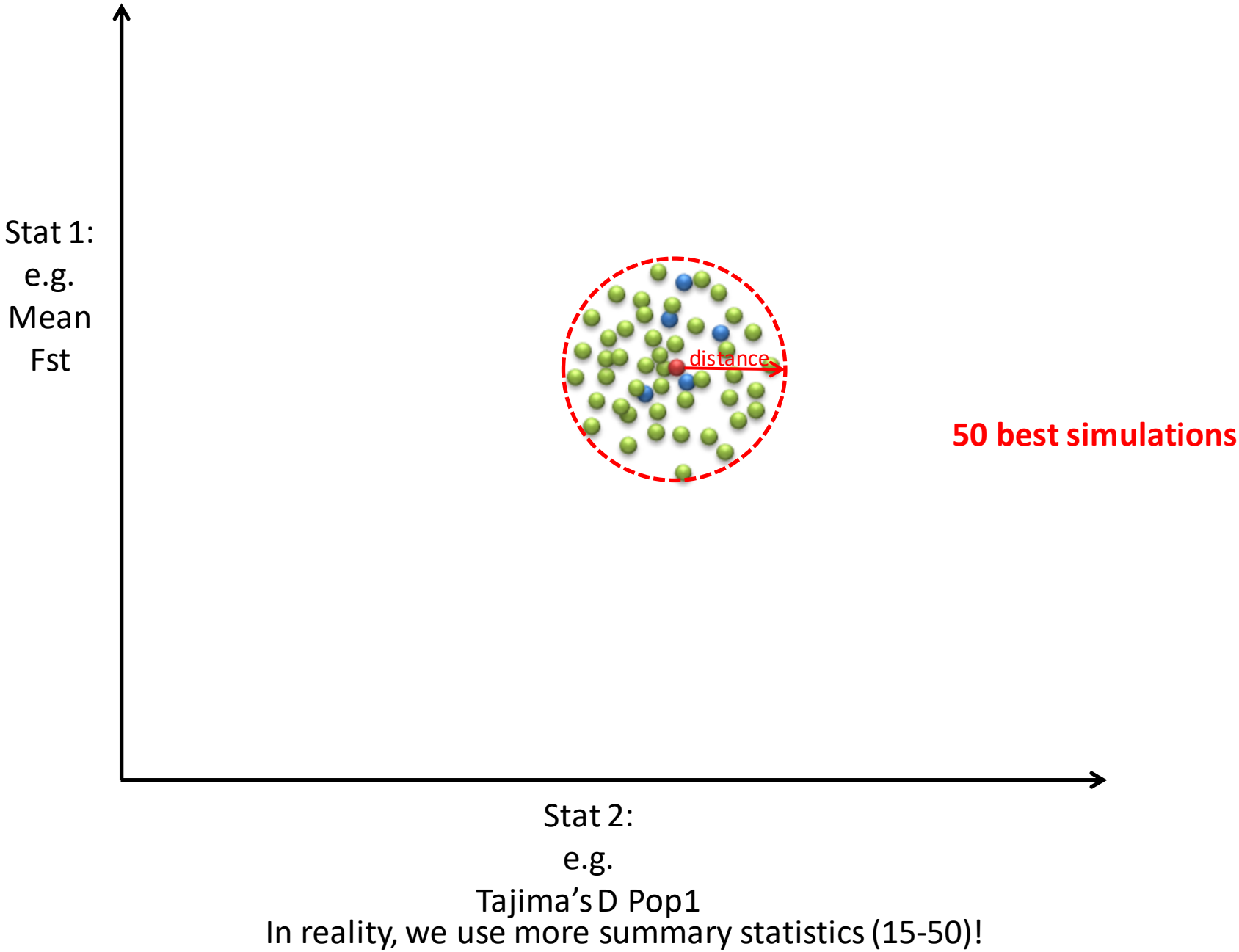

In reality, we use more summary statistics (15-50)!

## Approximate Bayesian Computation in practice



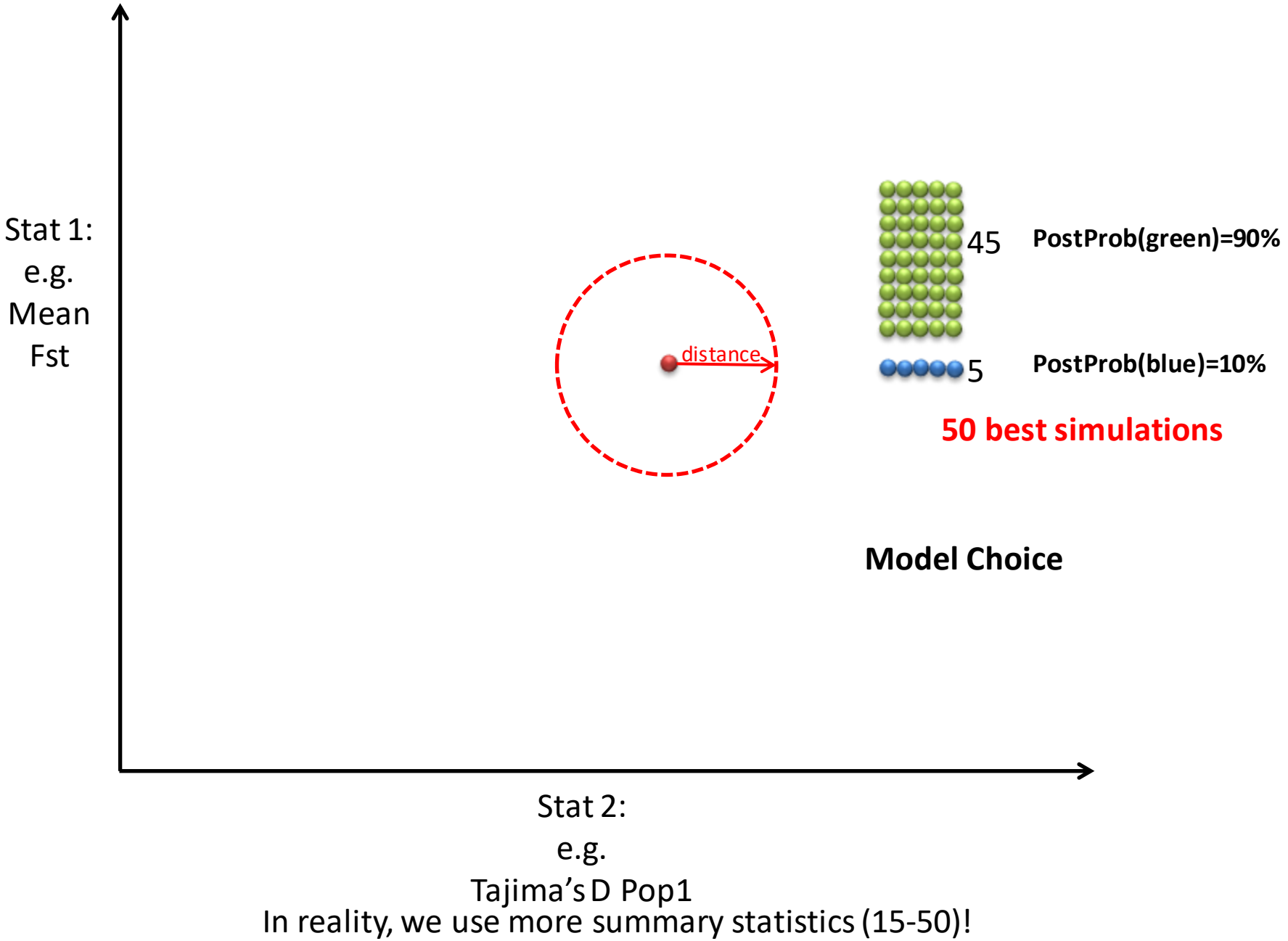

In reality, we use more summary statistics (15-50)!

**Approximate Bayesian Computation in practice**



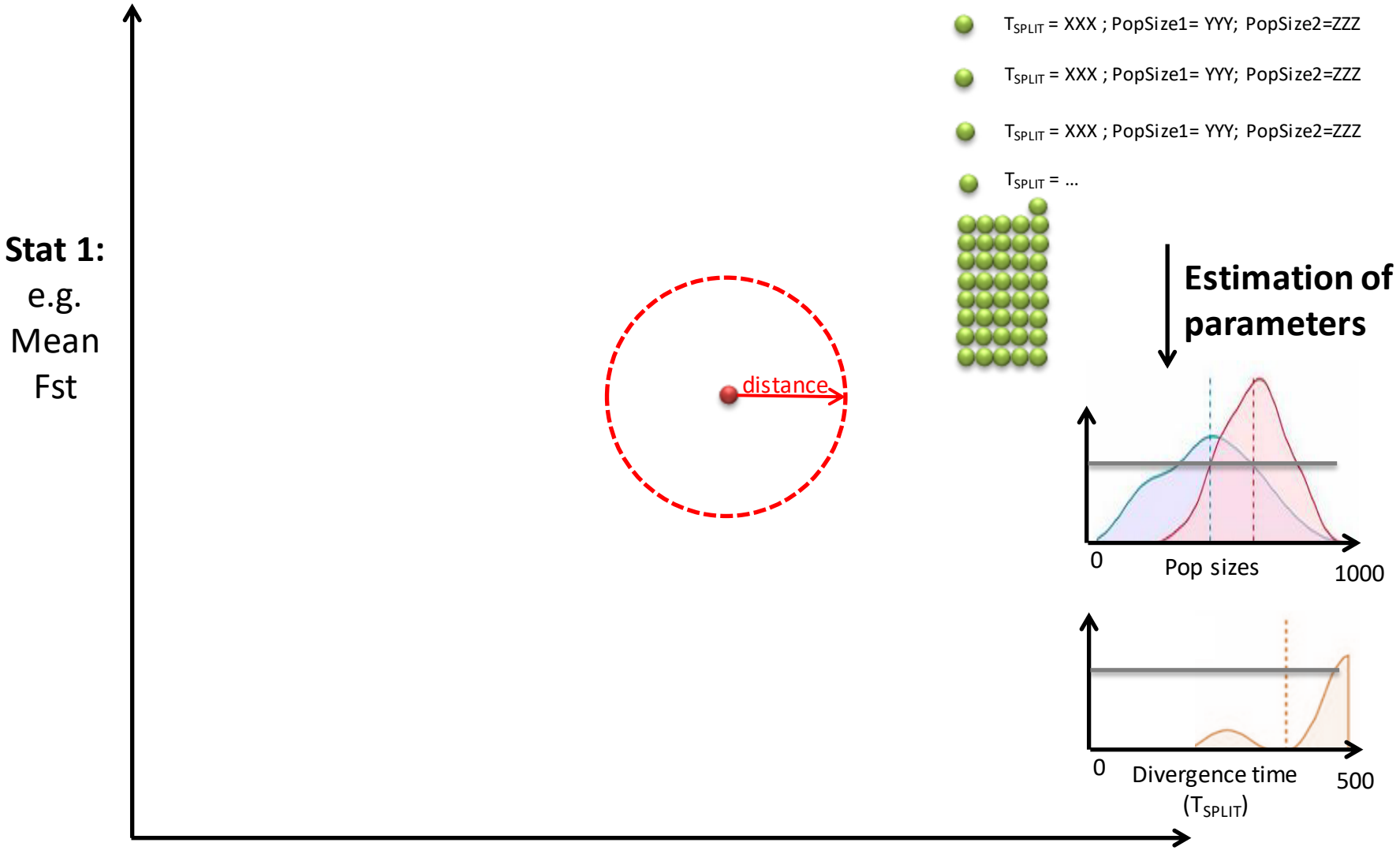

In reality, we use more summary statistics (15-50)!

# Approximate Bayesian Computation in practice



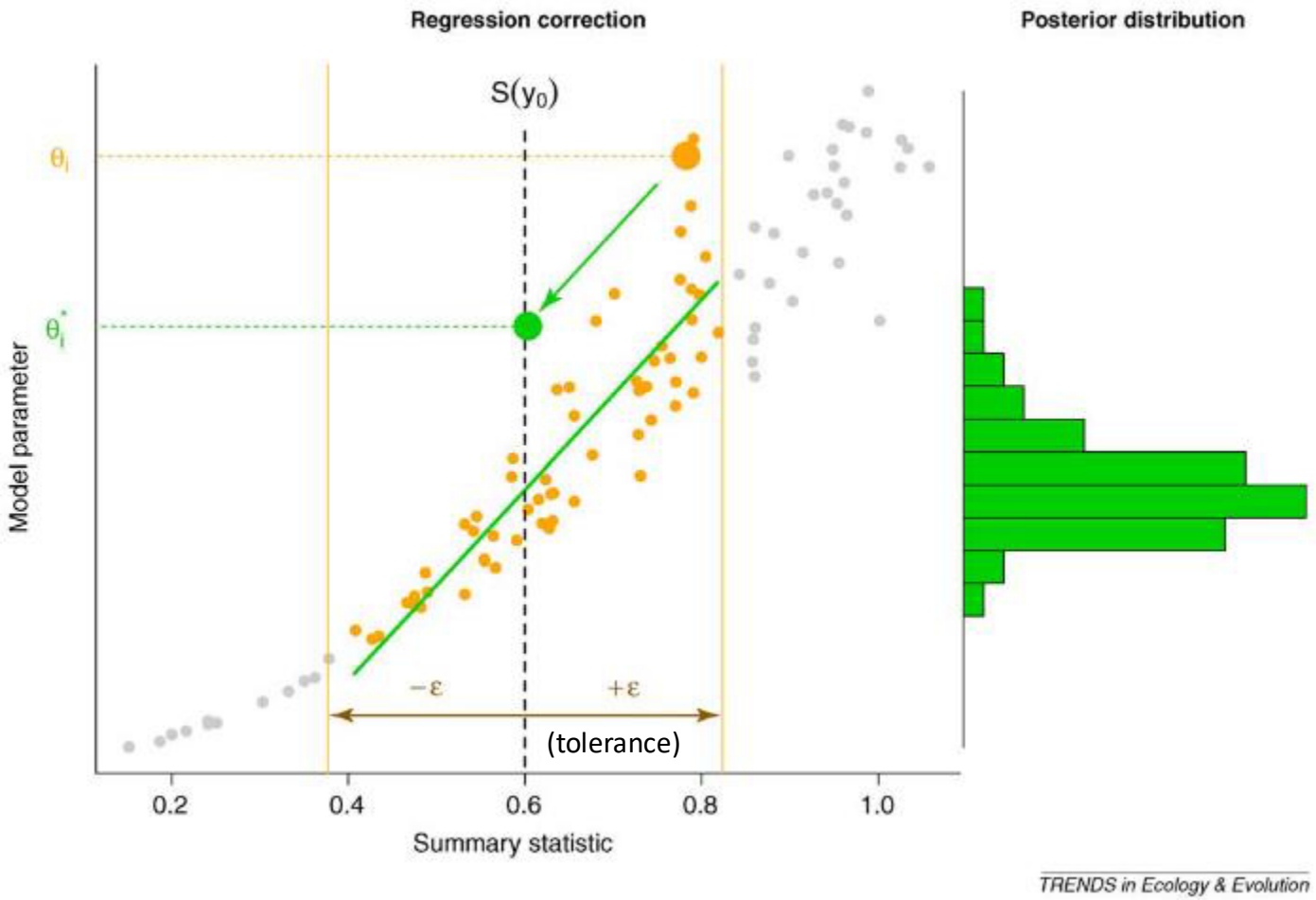

In reality, we use more summary statistics (15-50)!

# Approximate Bayesian Computation in practice




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

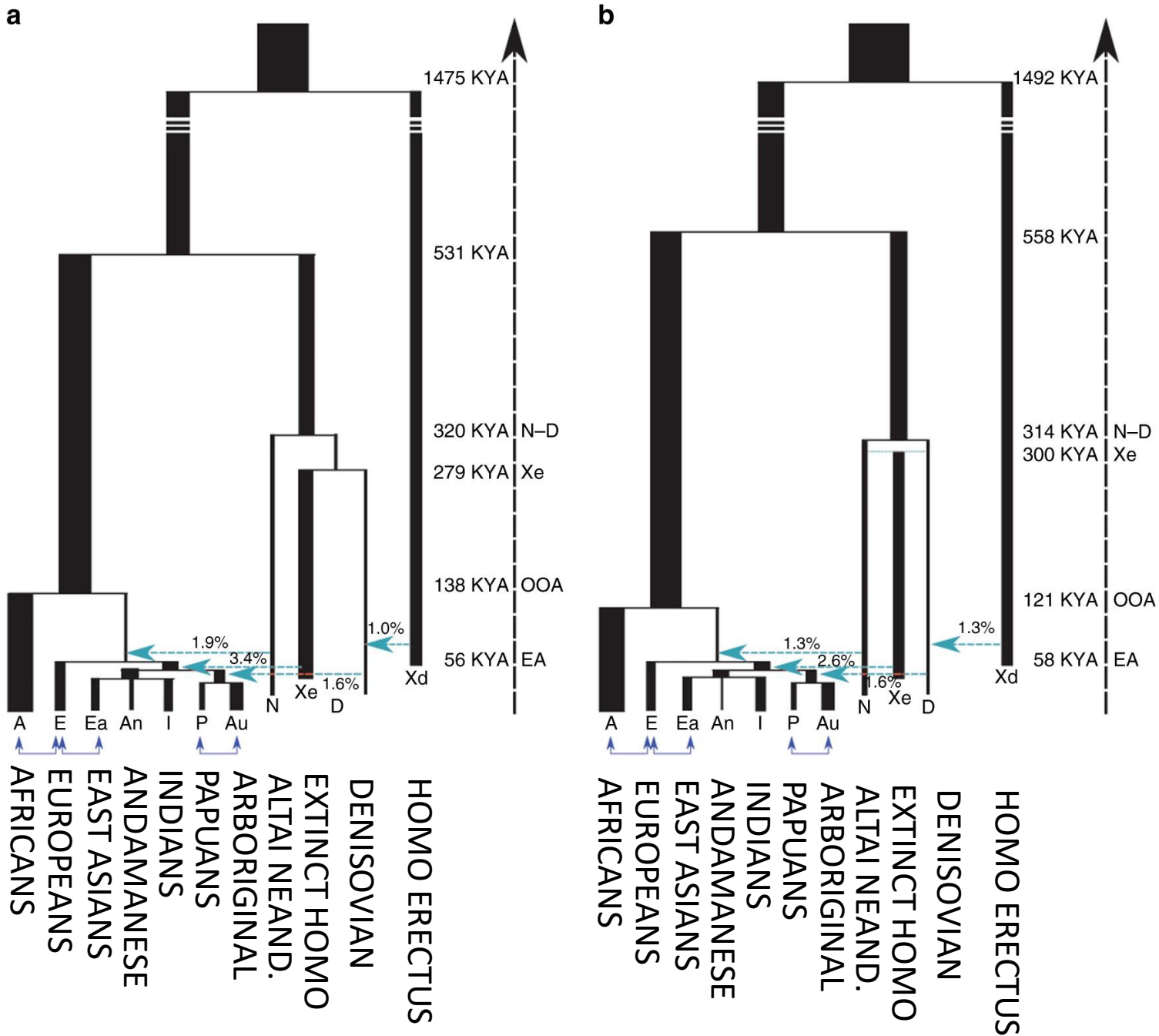
# Approximate Bayesian Computation in practice




In reality, it is slightly more complex, most 'standard ABC algorithms' also perform a local regression adjustement (linear or not) before to generate the posterior distribution

Rejection/regression methods remain popular, but more and more recent ABC algorithms now use machine learning tools (to reduce the computational burden)

# Application of an ABC strategy (with a deep learning method)

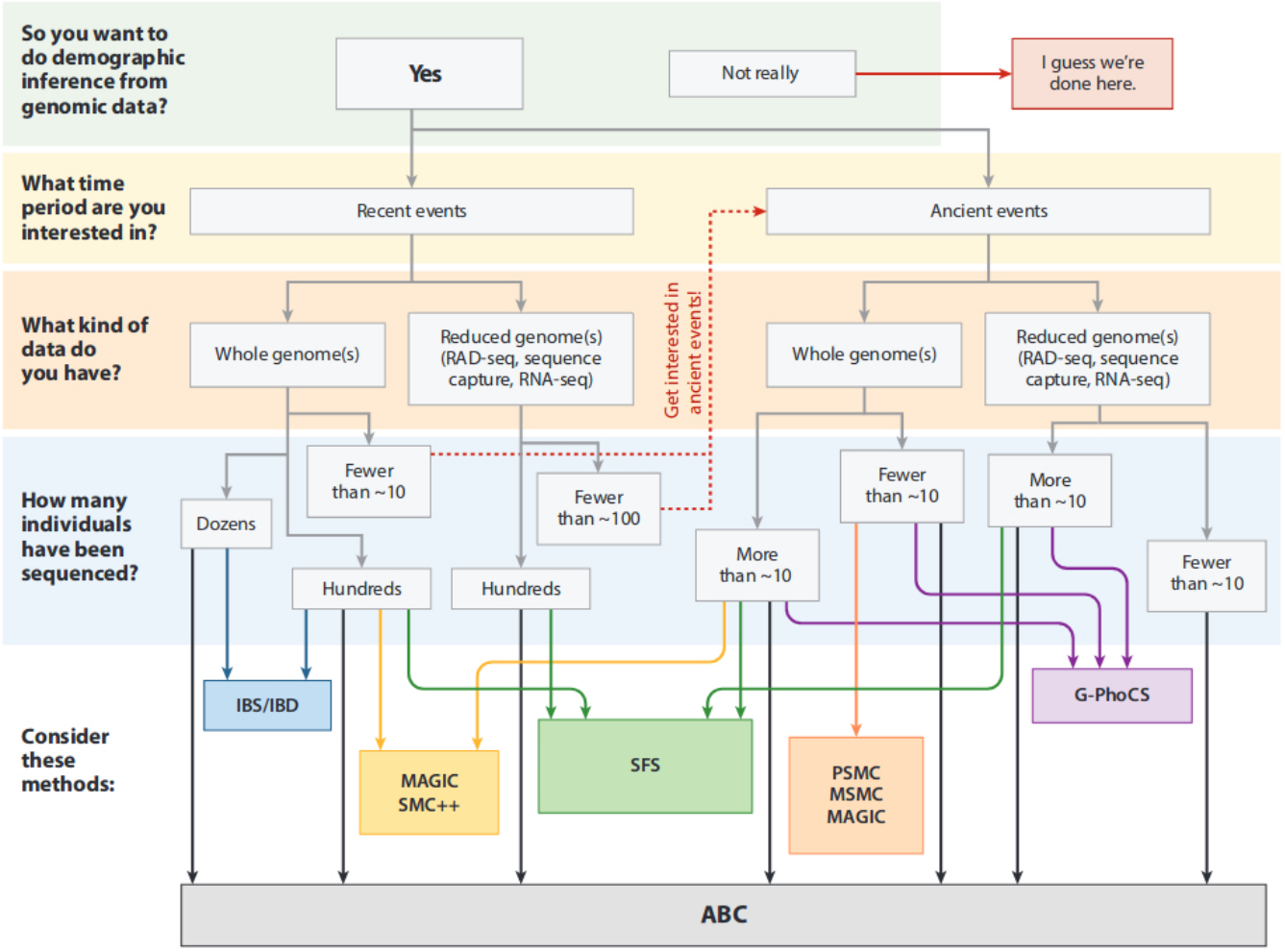



*Mondal et al. 2019 Nature Communications*

# Approximate Bayesian Computations: pros & cons

## Advantages:

- Flexible framework




'Recent events' here: <1000 generations ago

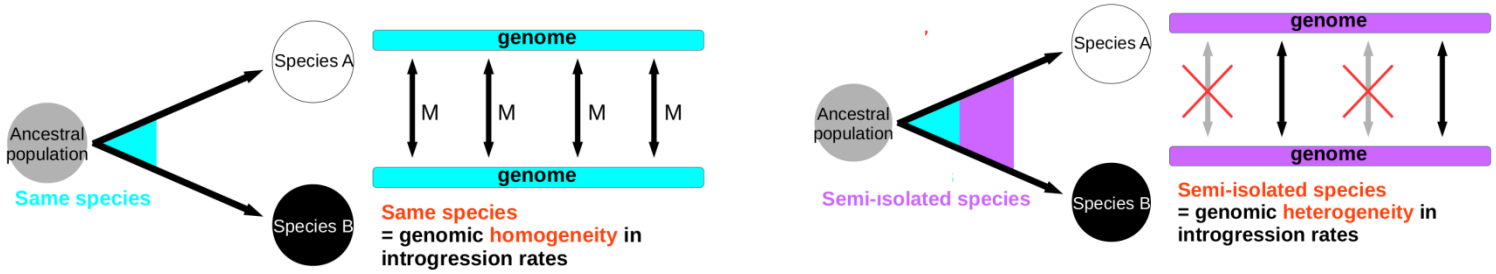
*Beichman et al. 2018 Nat Rev Ecol Evol Syst*

# Approximate Bayesian Computations: pros & cons

**Advantages:**

- Flexible framework

Incl. possibility to modelize genome-wide heterogeneity in migration rates or effective population sizes




*Modified from Camille Roux*

- Likelihood-free: no convergence issues
- Both model choice and estimation of parameters are easy to do
- Relatively straigtforward statistical model checking

**Limitations:**

- Considerable computational load (still true, but now becoming less and less the case)
- Human time (too)
- Risk of not including the true model (as for any model-based methods!)

## Demographic modeling tools: powerful, but…

**Never forget that models intentionally simplify reality!**

The exact evolutionary history is much more complex than what we can infer!

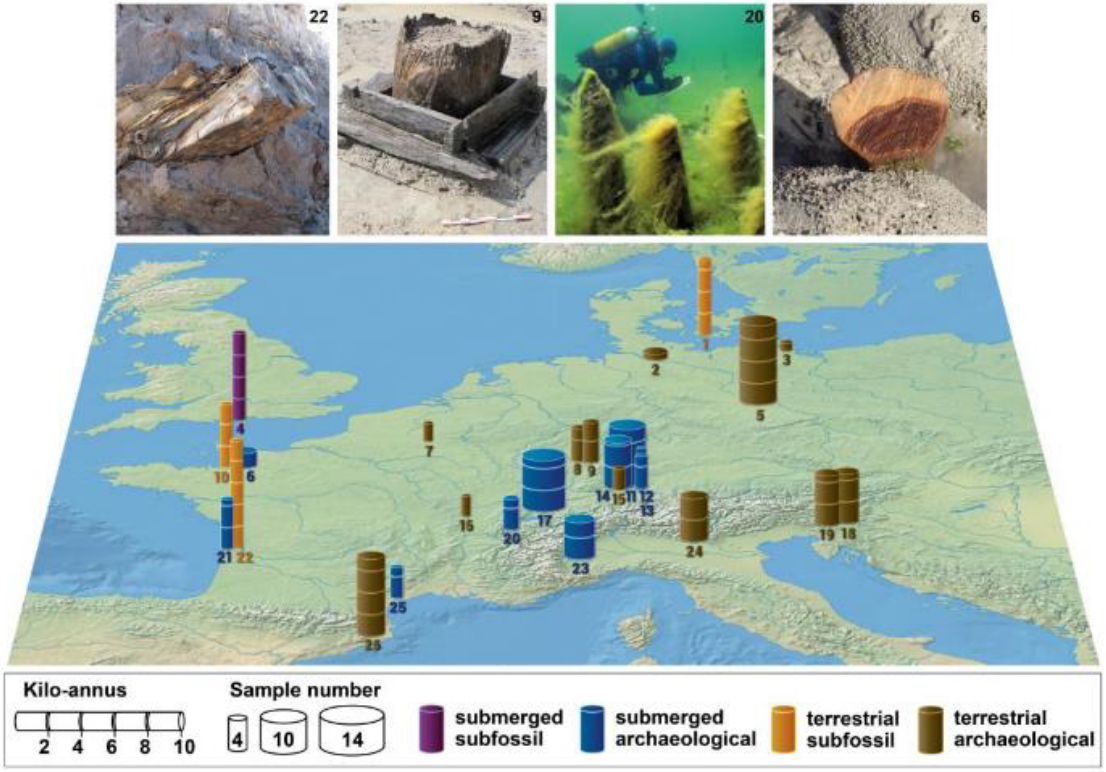
> **"…all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind…"**
> **— George Box —**

To always keep in mind:
1- We assume that the summary statistics (e.g. SFS, Fst…) capture the past demographic events.
2- Reducing a dataset to few summary statistics = considerable loss of genomic information!

Fossil records, when available, are extremely important to increase the confidence in the results! Not mutually exclusive, both are crucial!

**The future**

The availability of large collection of present-day samples for a large range of species (including plants) and also ancient DNA samples in the future will probably allow us to infer evolutionary history to a much greater extent




*Wagner et al. 2018 Mol Ecol*

Methods are evolving very fast => one of the most dynamic fields in population genomics

⟶ Soon: joint inferences of demography and selection?

⟶ Next course: selection