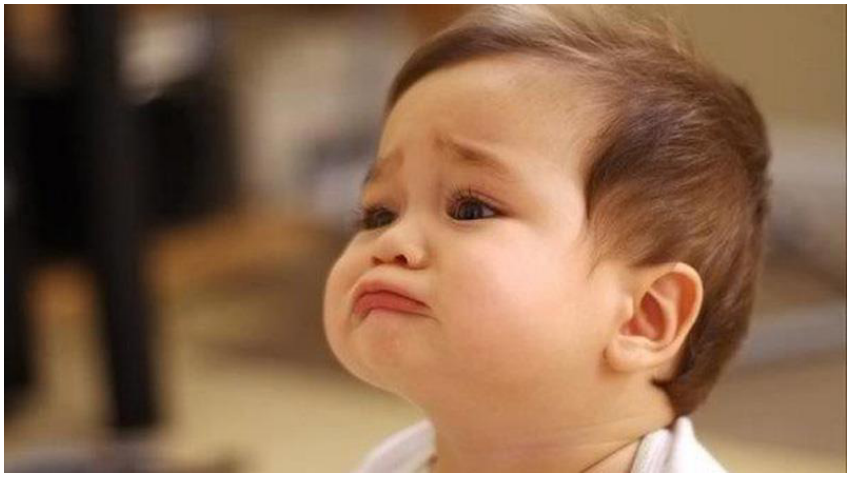
Introduction to demographic modeling

ABC (& beyond) workshop

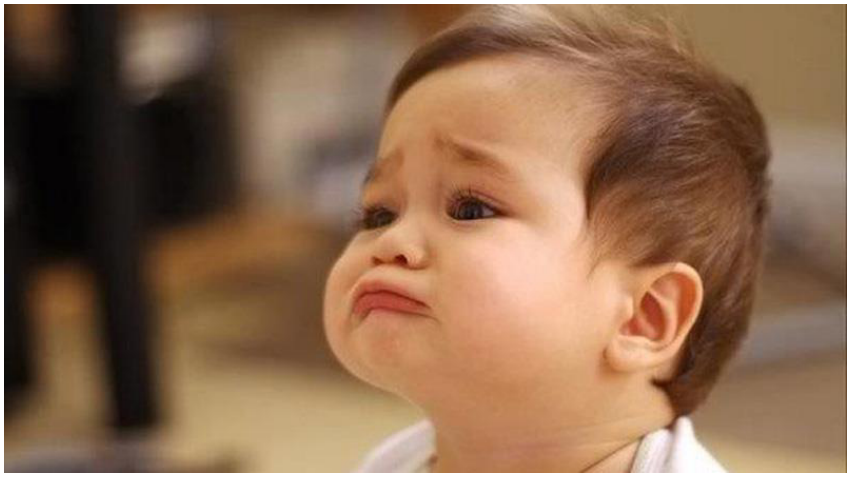
Thibault Leroy, December 6

Cautionary statements



Models intentionally simplify reality

Cautionary statements

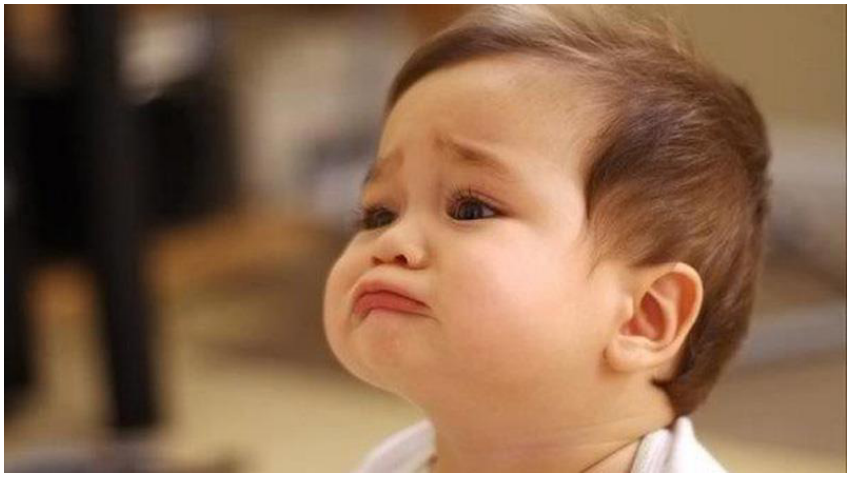


Models intentionally simplify reality
Models intentionally simplify reality
Models intentionally simplify reality
Models intentionally simplify reality
Models intentionally simplify reality
Models intentionally simplify reality
Models intentionally simplify reality

"…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 —

Cautionary statements



Models intentionally simplify reality

- Assumptions are made.
- Approximations are made.

To always keep in mind:

1- We assume that the summary statistics (e.g. SFS, Fst…) capture the past demo. events.

2- Reducing a dataset to few summary statistics = considerable loss of information

- All demographic changes can NOT be detected (be realistic)
- Fossil record >> demographic inferences

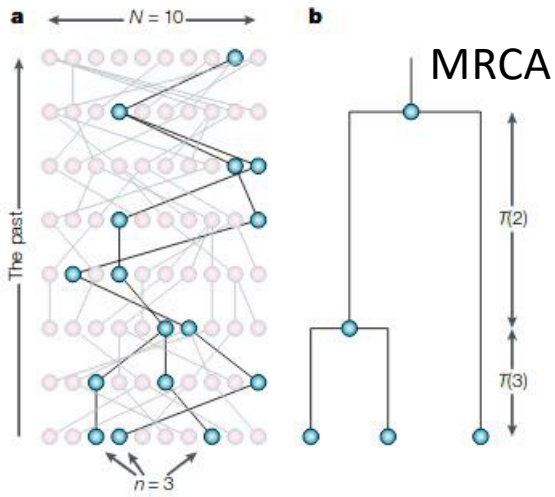
(demographic inferences + fossil record >>>> demographic inferences only)

Coalescent theory (in one slide)

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



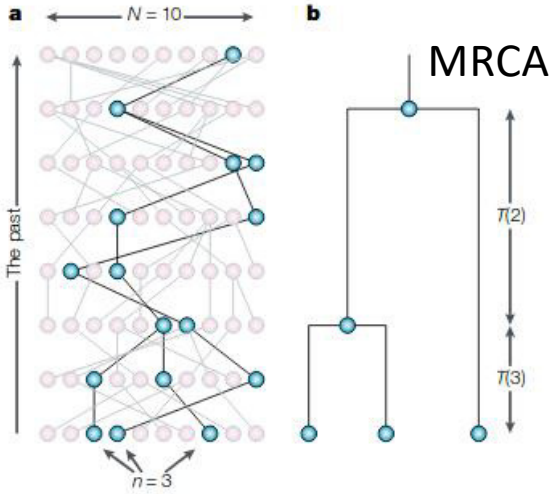
e.g. Rosenberg & Nordborg, 2002

Coalescent theory (in one slide)

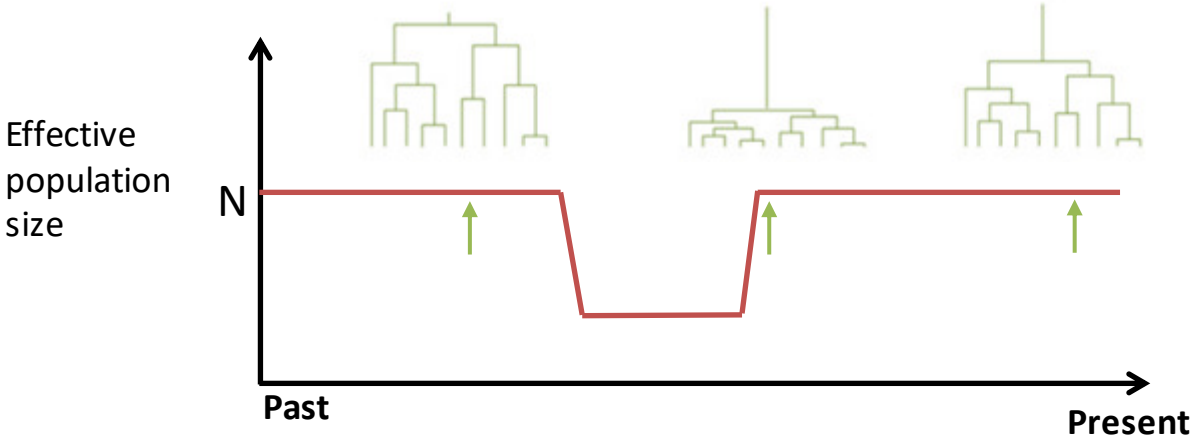
A stochastic process that describes how population genetic processes determine the shape of the genealogy of sampled gene sequences

n individuals sampled from a population of:

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



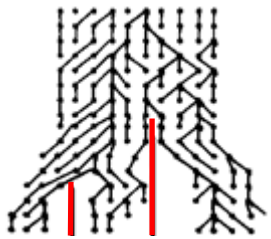
e.g. Rosenberg & Nordborg, 2002



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

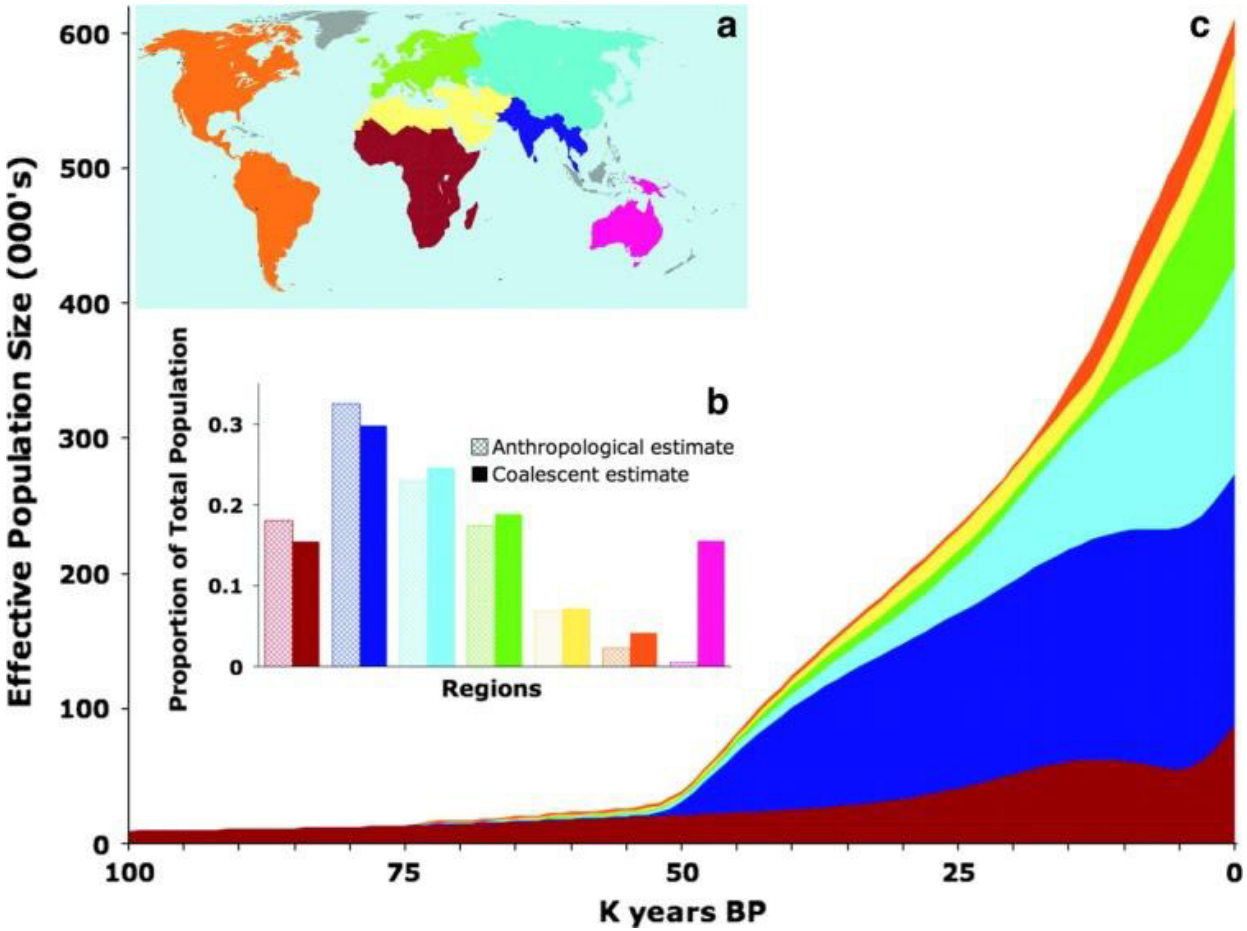
→ Maximum likelihood estimates of parameters (pop expansions/bottlenecks)

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 is generally poor

Multiple loci / genomes

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

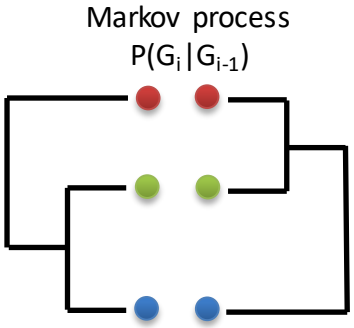
→ But full likelihood methods are not applicable to genome-scale datasets because of two significant limitations:
1) they do not scale well in the number of loci being analyzed
2) they are not well suited for handling recombination (modeling genomic linkage is particularly challenging)

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, …)

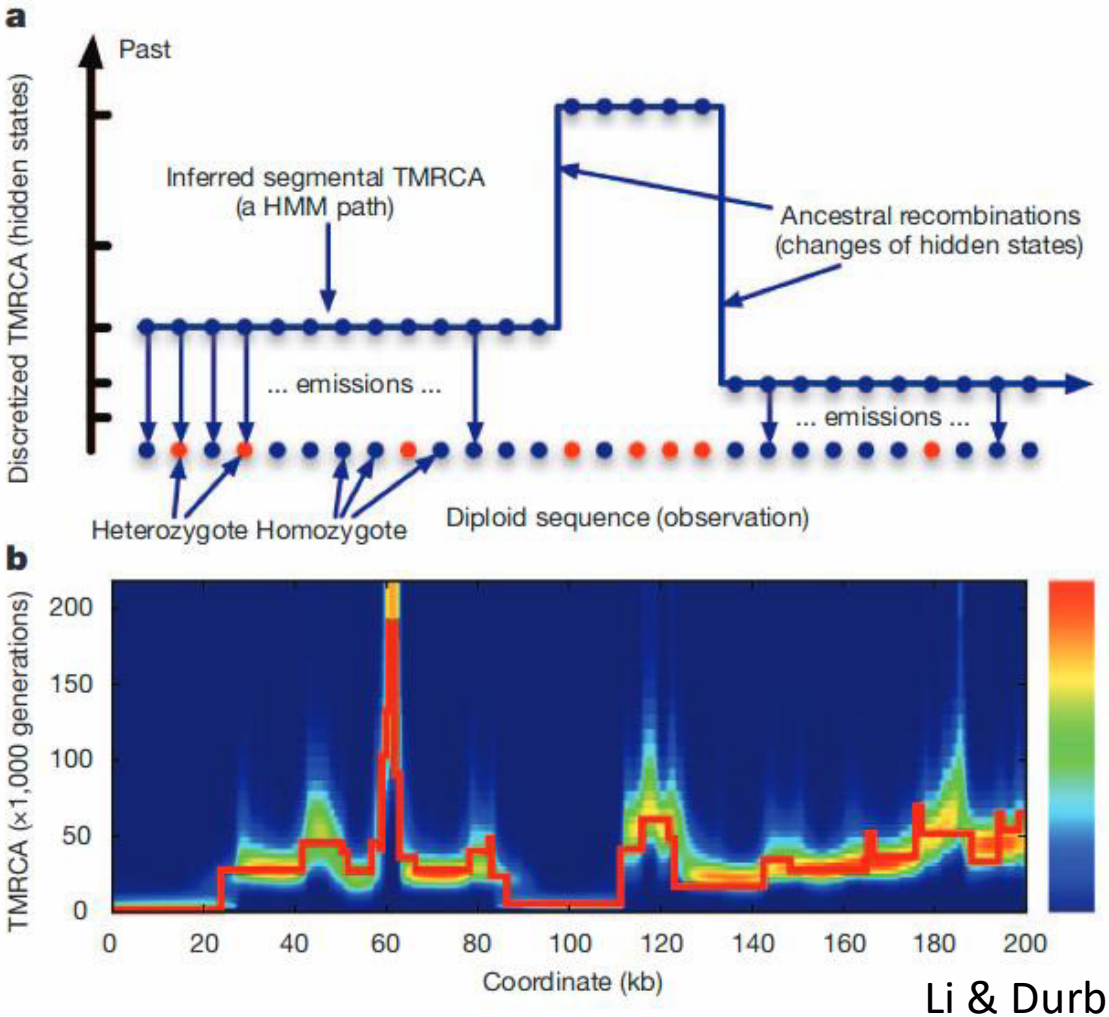
→ Approximate Bayesian computation

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

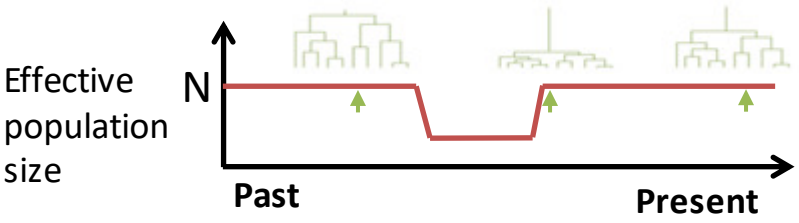


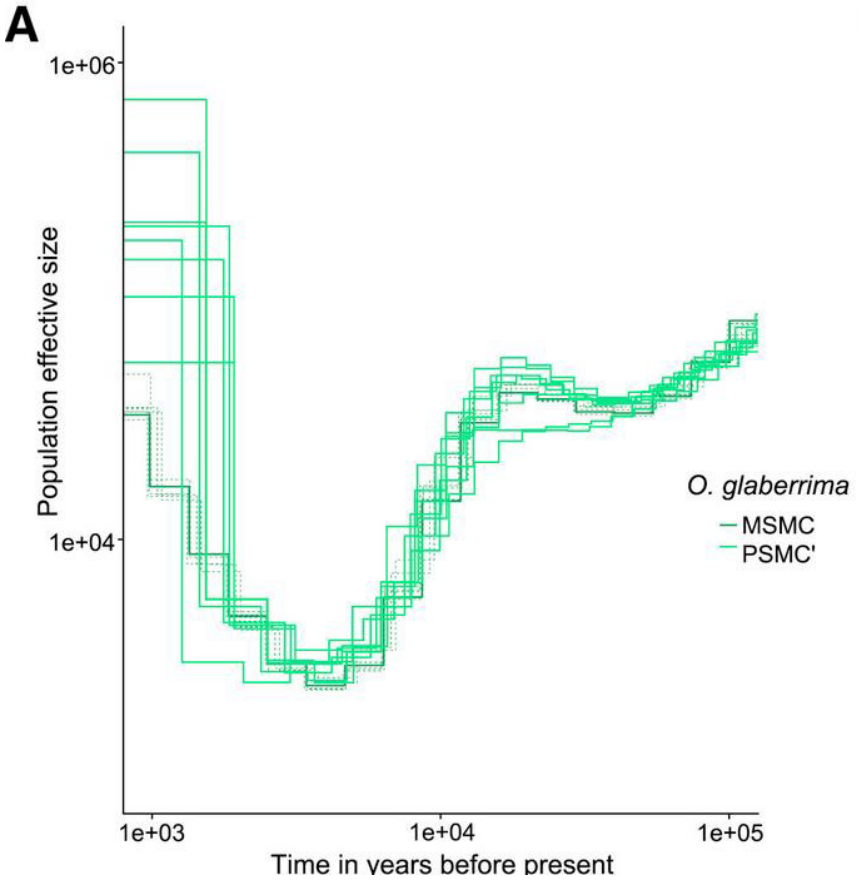
« Skyline plots »

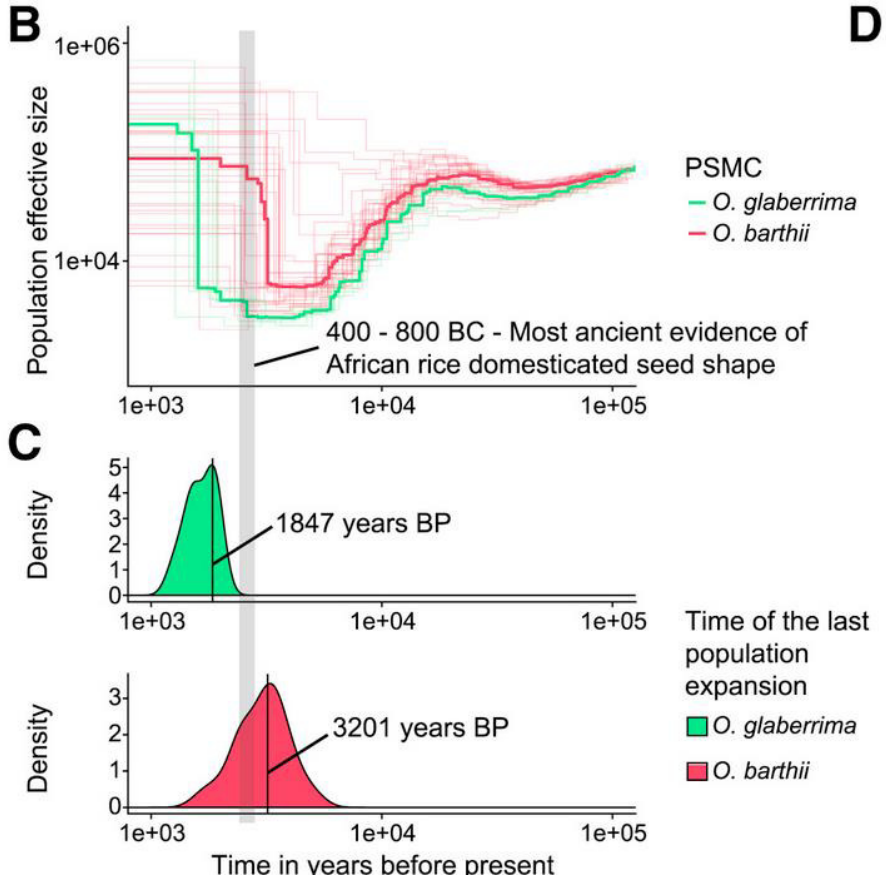
Li & Durbin, 2011 Nature

PSMC-like methods: basics

The rate of the inferred coalescent events at a given time is inversely proportional to *Ne* => identification of periods *of Ne* change





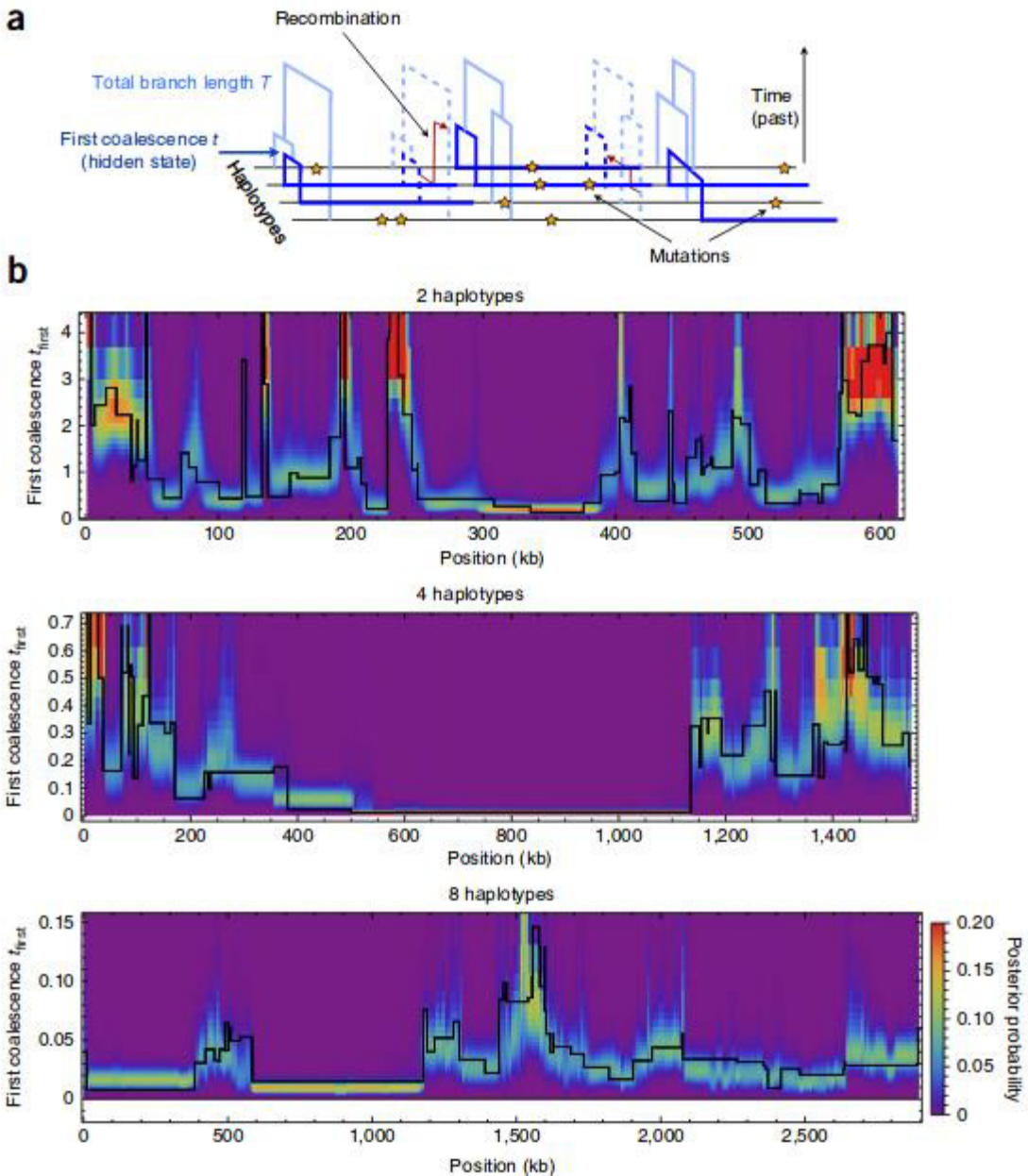


To convert these estimates into real time, all scaled results need to be divided by the mutation rate and multiplied by the generation time.

Cubry et al. 2018 Current Biology

PSMC-like methods: basics

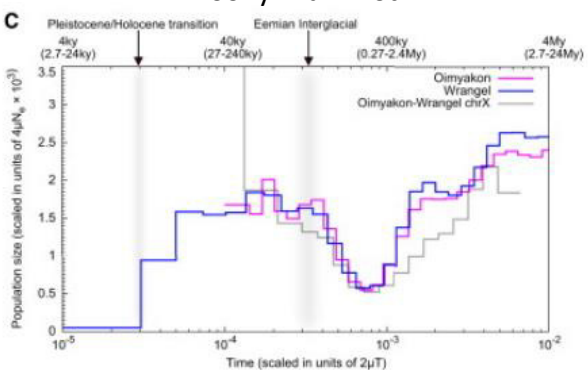
Multiple sequentially Markovian coalescent (MSMC)



Schiffels & Durbin, 2014 Nature genetics

PSMC-like methods: pros & cons

Woolly Mammoth



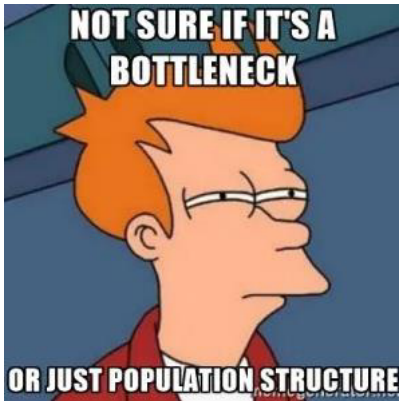
Palkopoulou et al. 2015, Current Biol

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

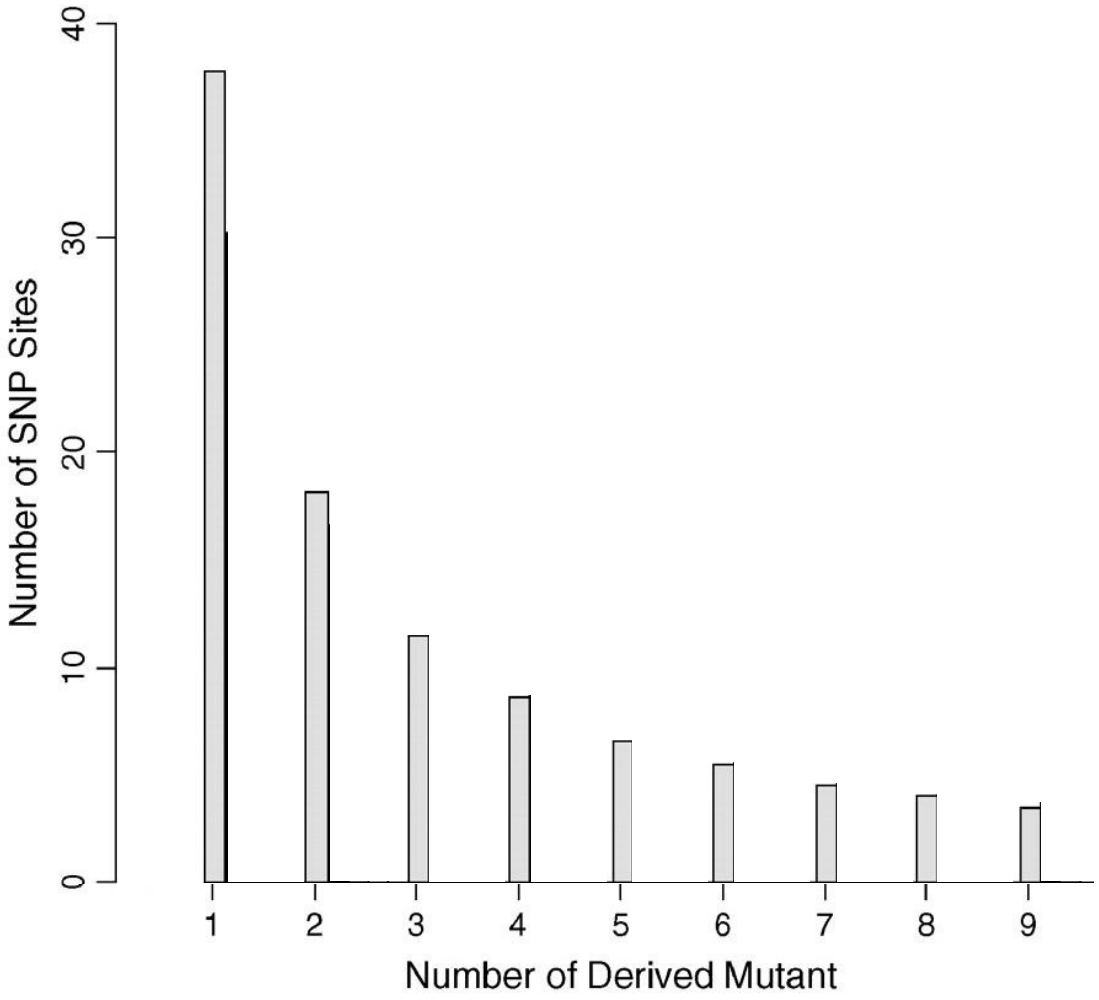


moleculareecologist.com

- Somewhat sensitive to the quality of the genome assembly (N%, scaffold length)
- Somewhat sensitive to the quality of the data
  To make reliable demographic, Nadachowska-Brzyska et al. (2016) suggested filters :
  - A mean genome-wide coverage of at least 18X
  - No more than 25% of missing data
  - A per-site filter of ≥10 reads
- 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

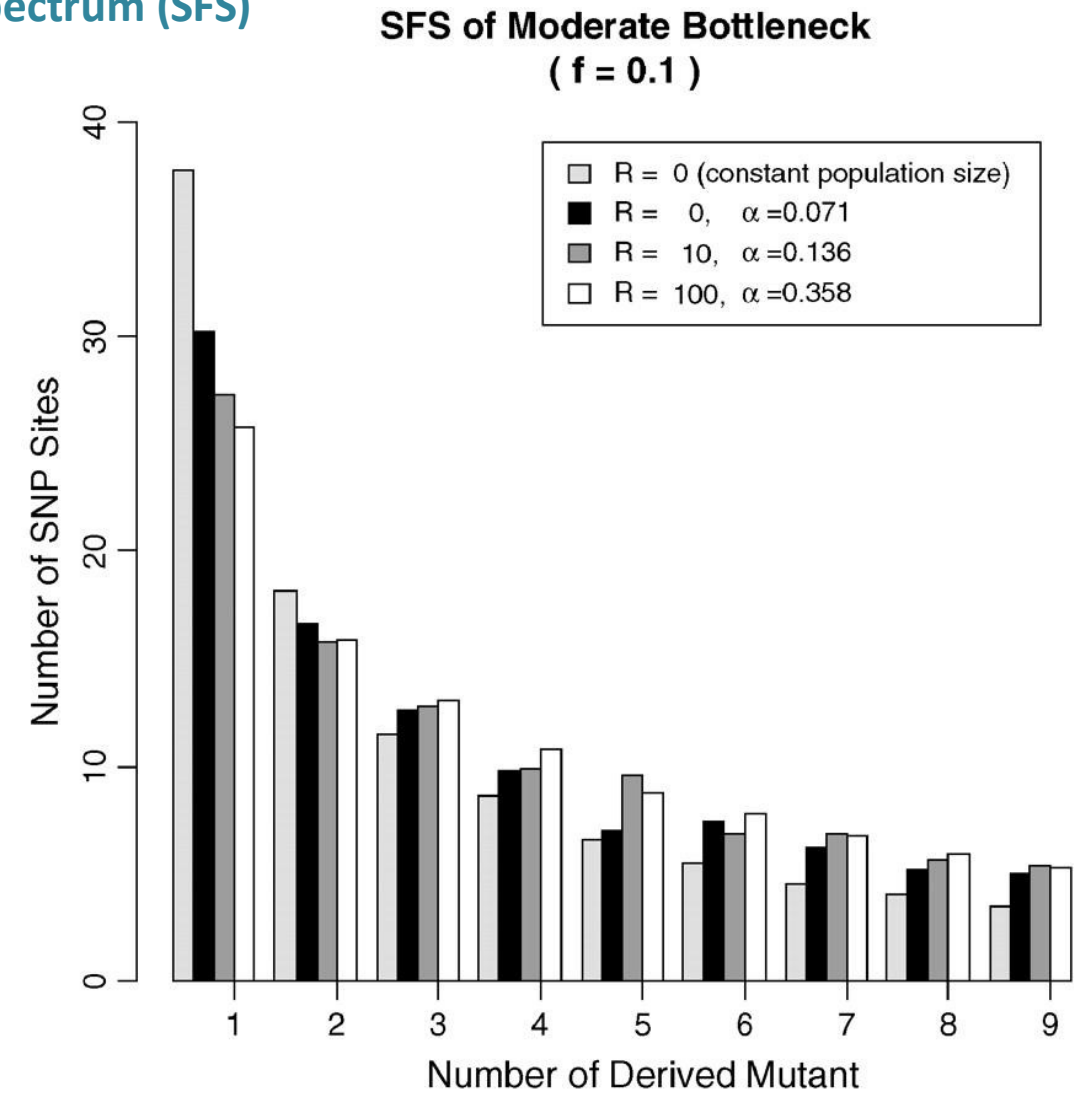
Site Frequency Spectrum (SFS)



Zhu & Bustamante, 2005

Composite likelihood methods: basics

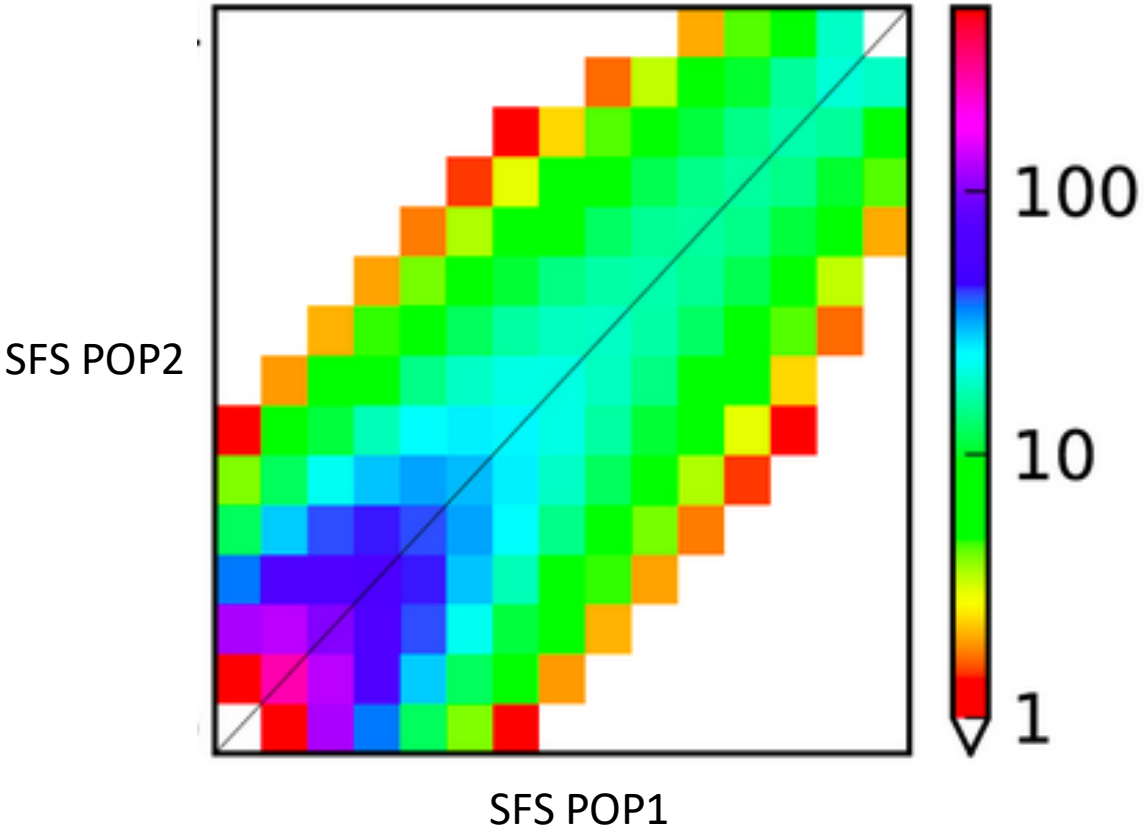
Site Frequency Spectrum (SFS)



Zhu & Bustamante, 2005

Composite likelihood methods: basics

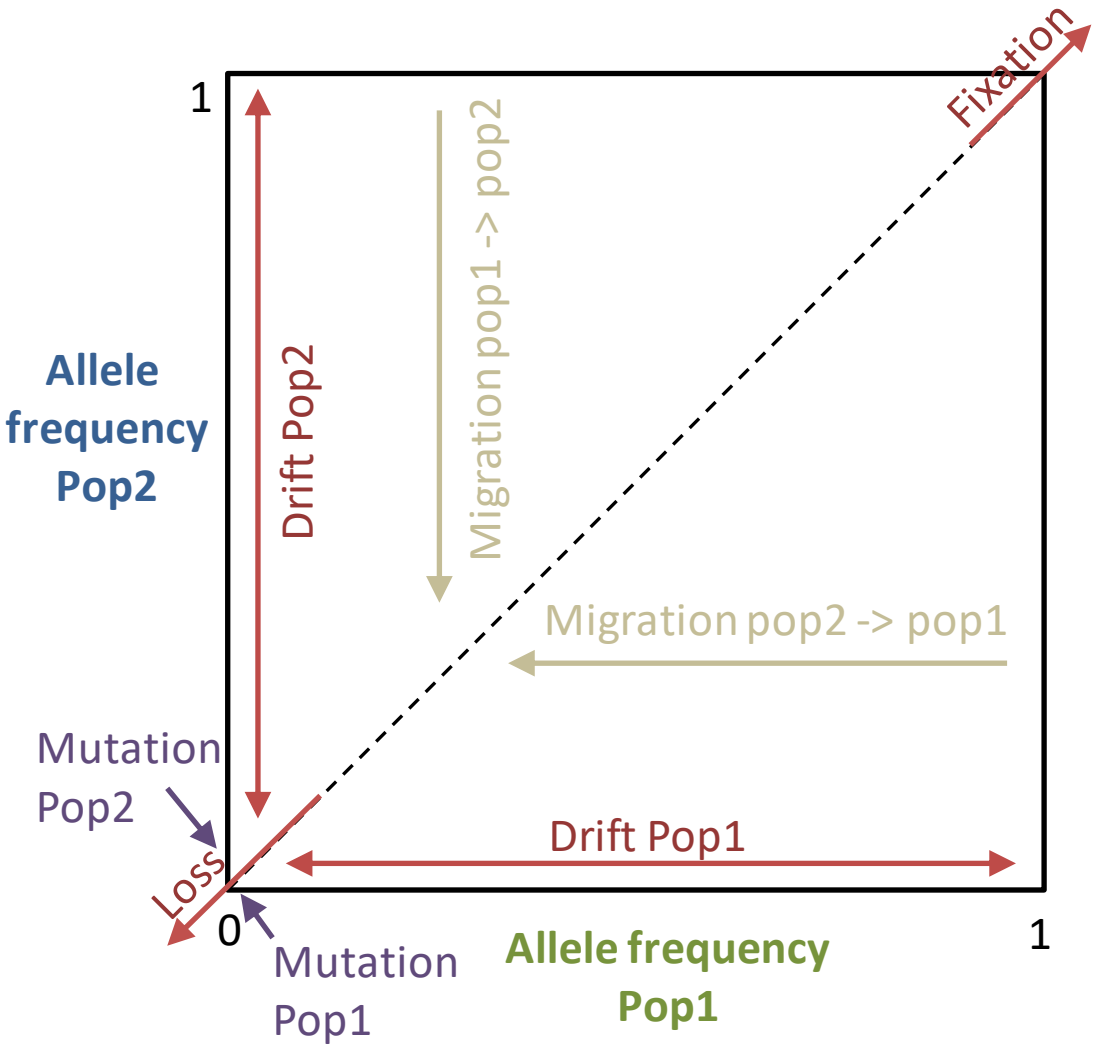
Joint Site Frequency Spectrum (2D-SFS)



Adapted from Gutenkunst et al. 2009 Plos Genet

Composite likelihood methods: basics

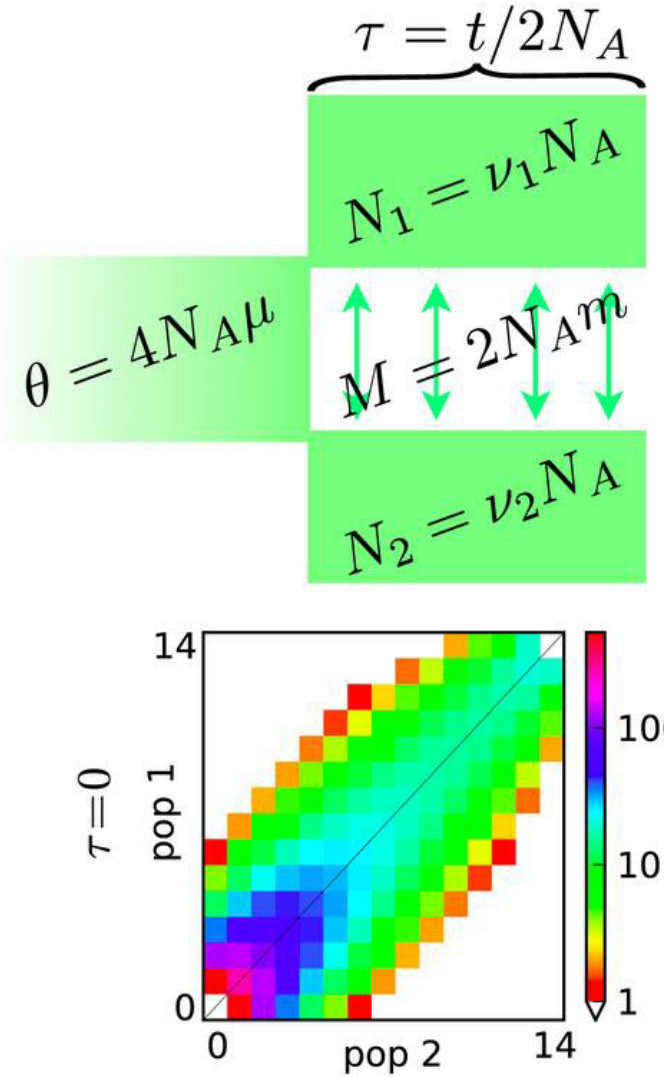
Joint Site Frequency Spectrum (2D-SFS)

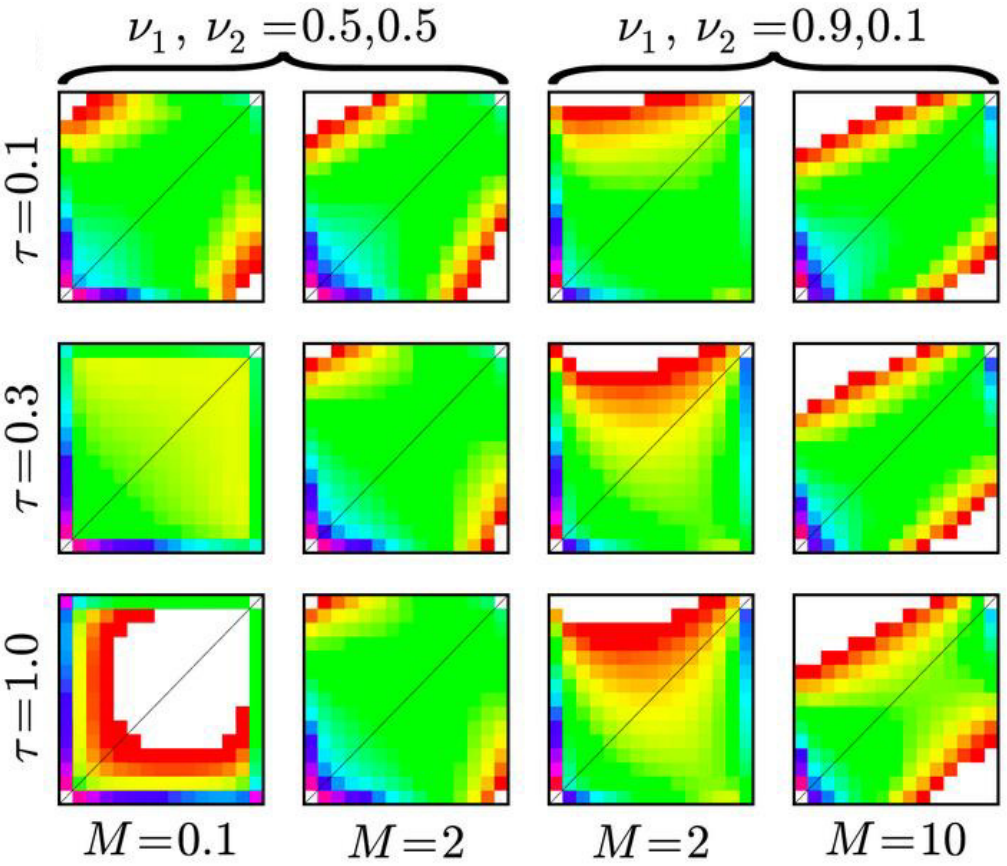


Adapted from Gutenkunst et al. 2009 Plos Genet

Composite likelihood methods: basics

Simulated 2D-SFS





Gutenkunst et al. 2009 Plos Genet

Composite likelihood methods: 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:

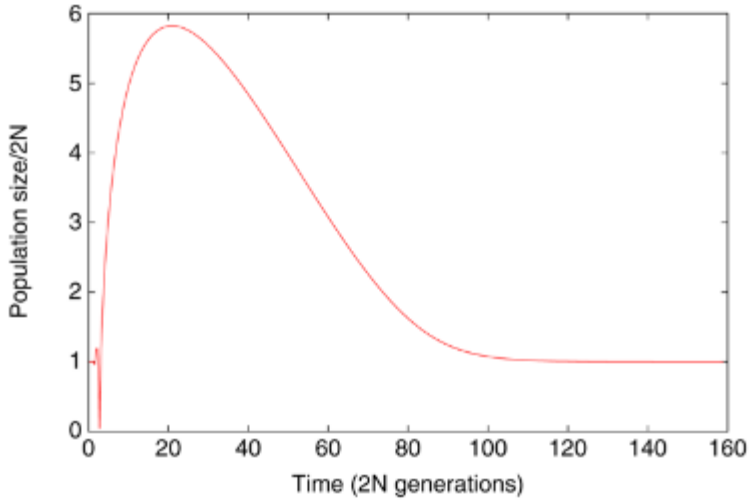
Computational issues

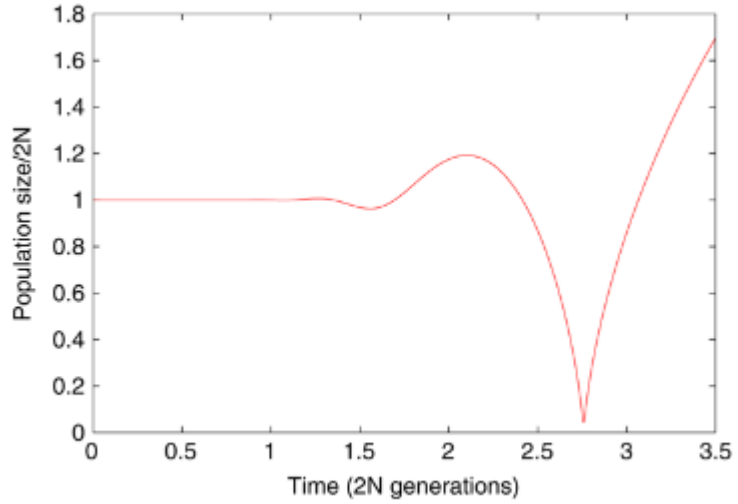
- Convergence problems are possible

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
- 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 (& in 2 slides!)

Likelihood-free demographic inferences

Real Data

100 locus x 1kb

↓

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

Simulations

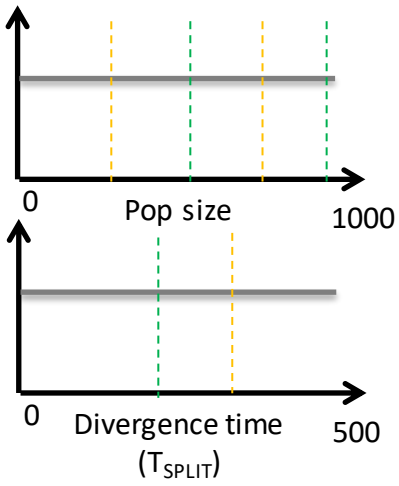
Model 1

(e.g. 1 million multilocus simulations, *i.e.* 1 million – 100 locus 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

…

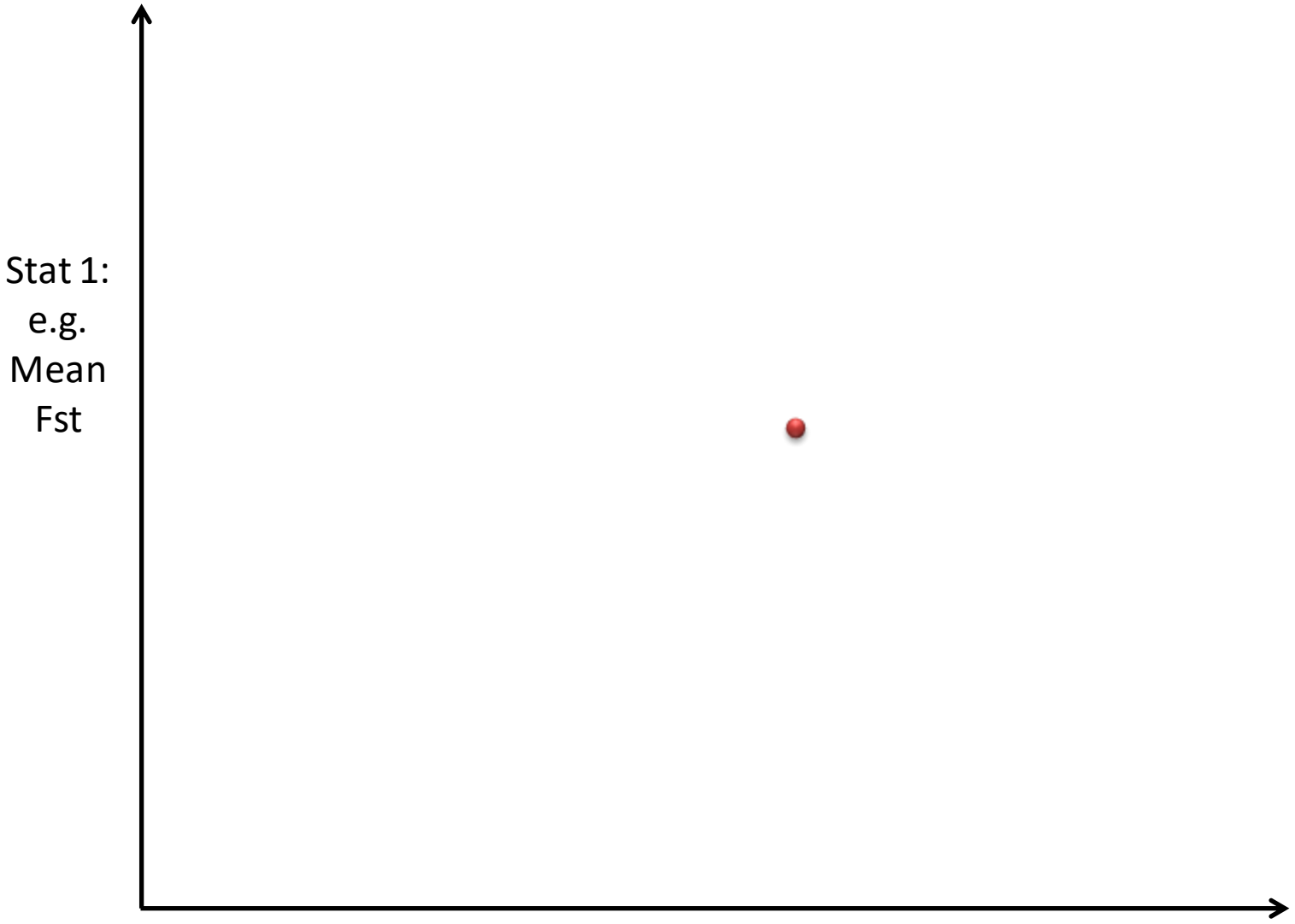
Same summary statistics

Simulating large numbers of datasets under several hypothesized evolutionary scenarios

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

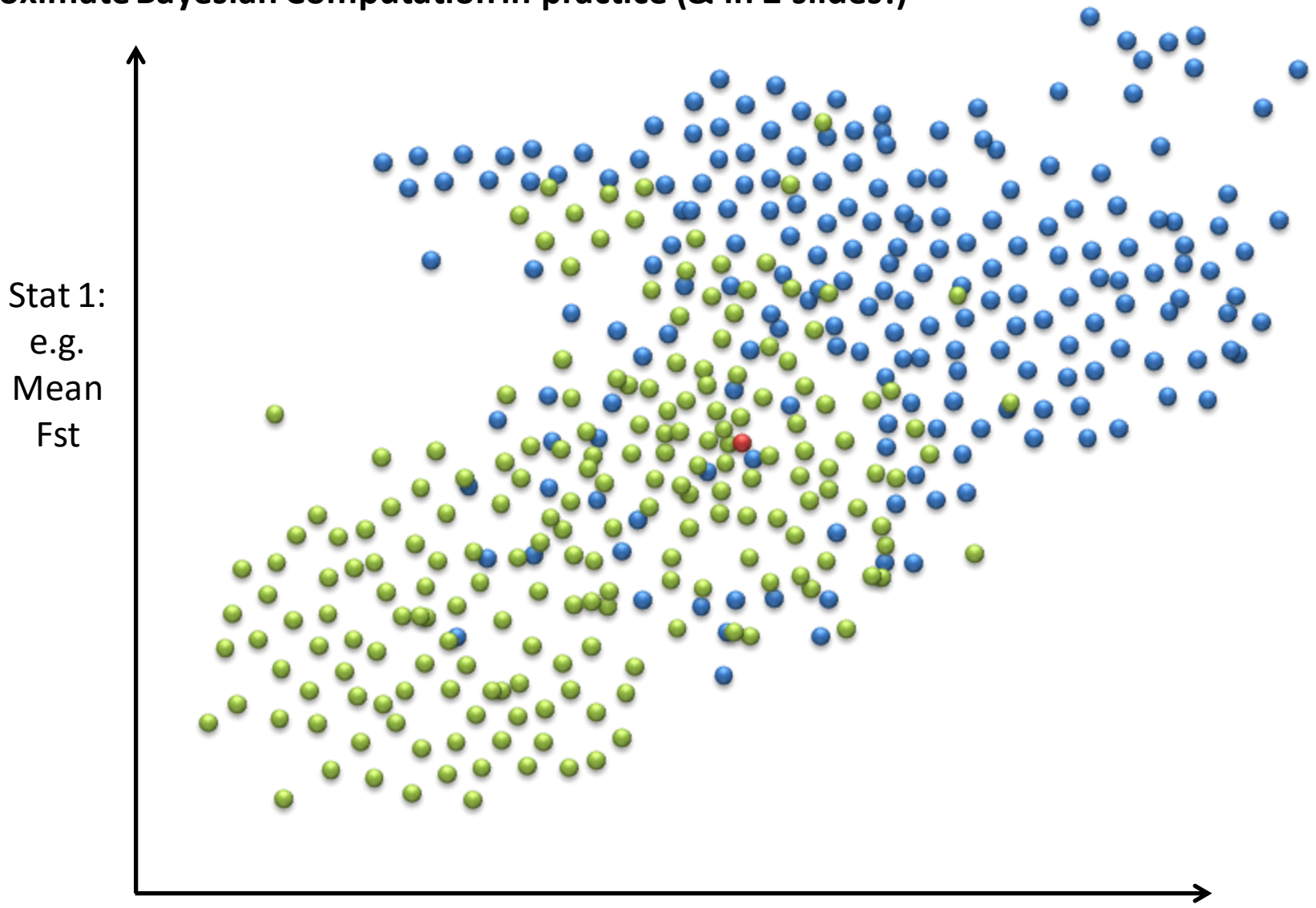
Compute the Euclidian distance between the simulated and the observed summary statistics

Approximate Bayesian Computation in practice (& in 2 slides!)



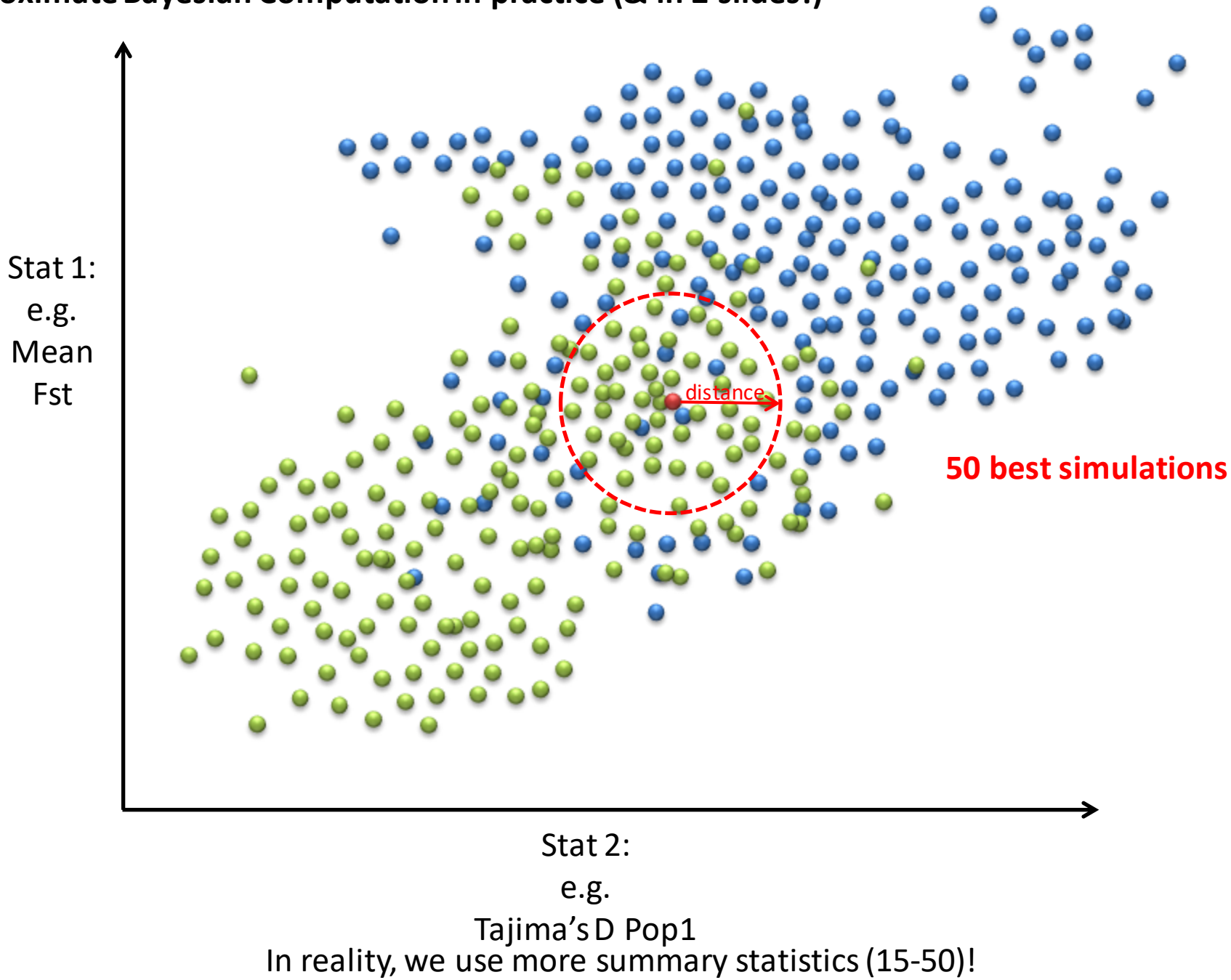
In reality, we use more summary statistics (15-50)!

Approximate Bayesian Computation in practice (& in 2 slides!)



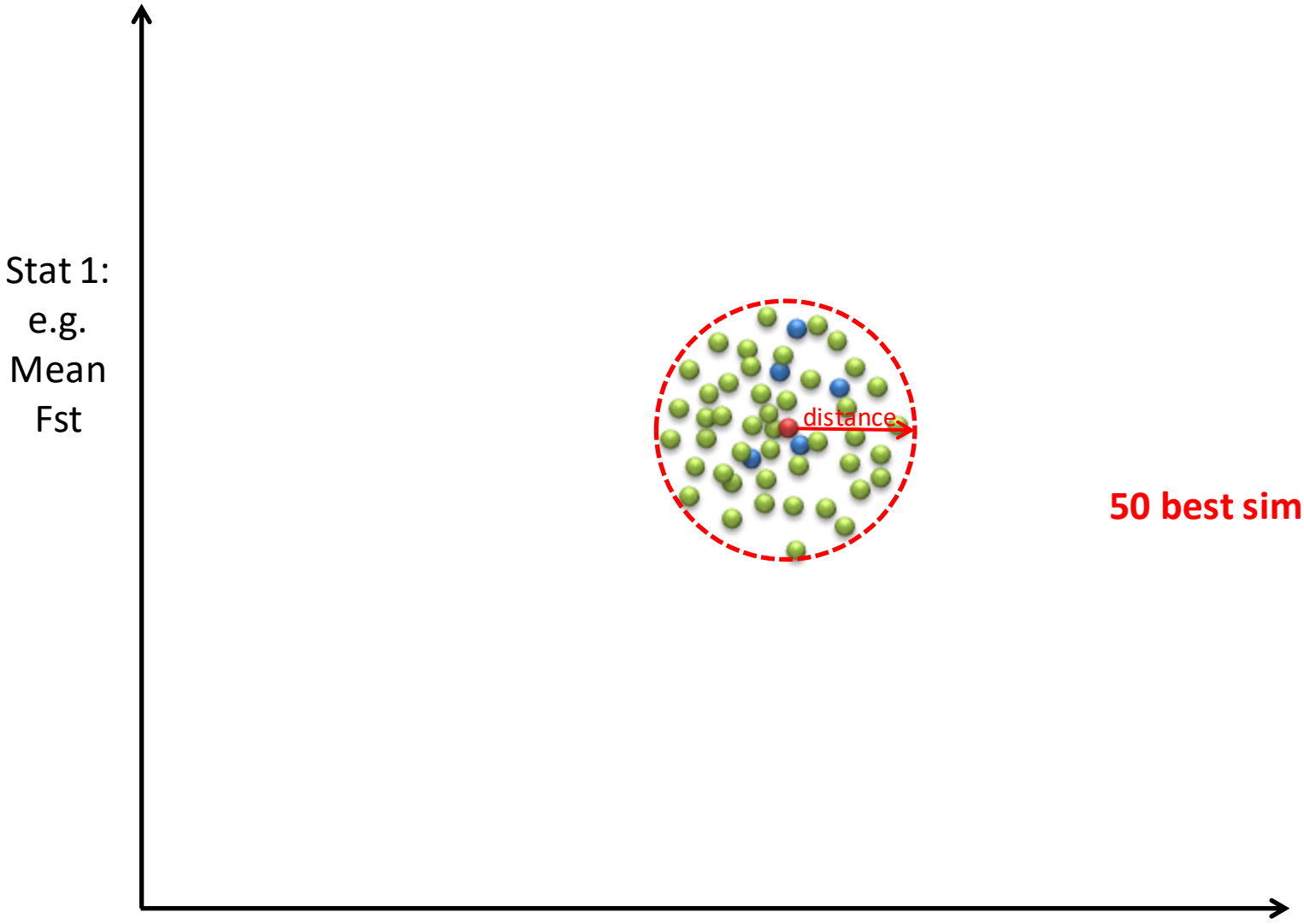
In reality, we use more summary statistics (15-50)!

Approximate Bayesian Computation in practice (& in 2 slides!)



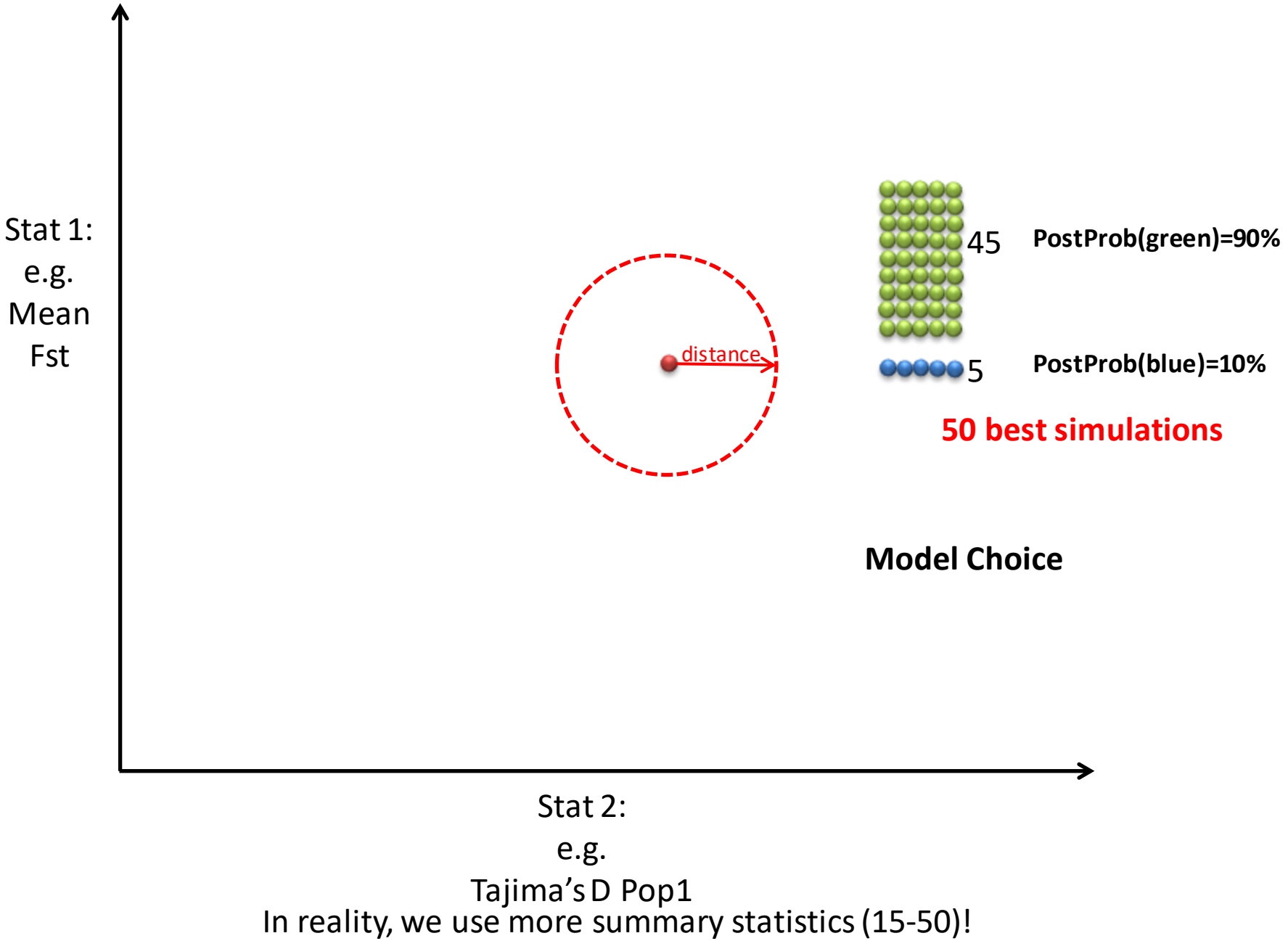
In reality, we use more summary statistics (15-50)!

Approximate Bayesian Computation in practice (& in 2 slides!)



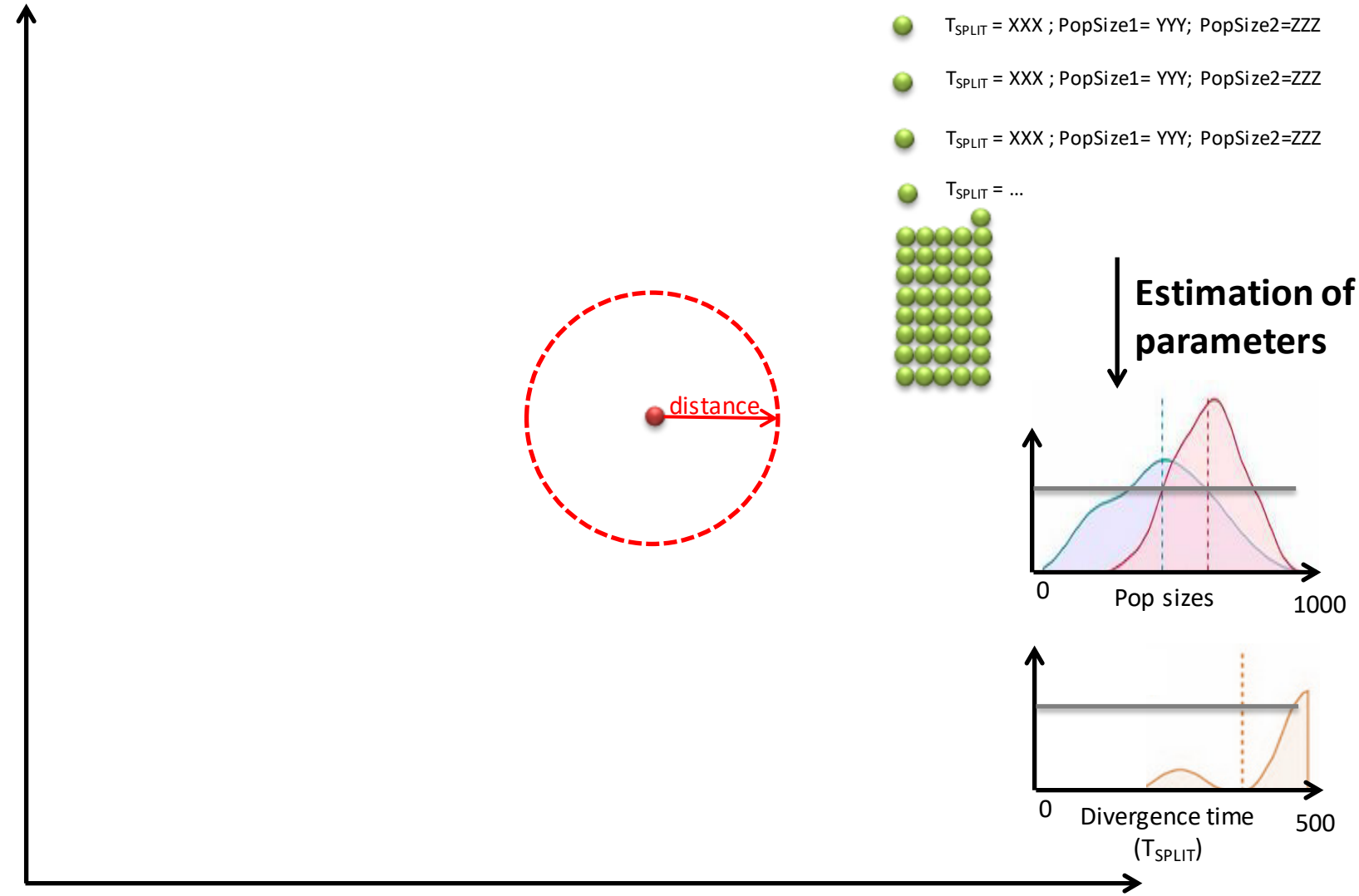
In reality, we use more summary statistics (15-50)!

Approximate Bayesian Computation in practice (& in 2 slides!)



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

Approximate Bayesian Computation in practice (& in 2 slides!)



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

Approximate Bayesian Computations: pros & cons

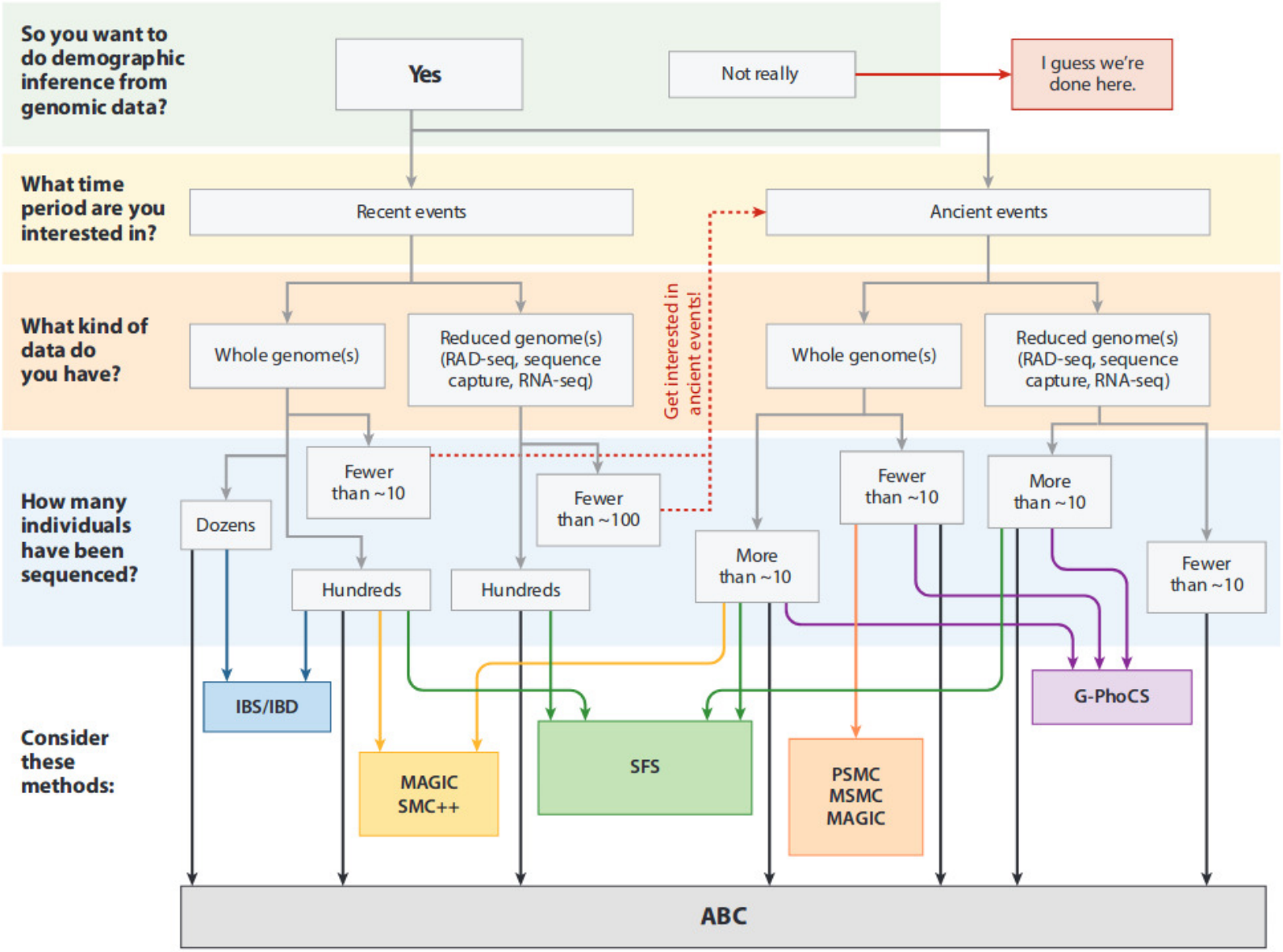
Advantages:

- Flexible framework
- 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 (… at least, before Clark Kent<=> Camille Roux)
- Human time too
- Risk of not including the true model

Summary

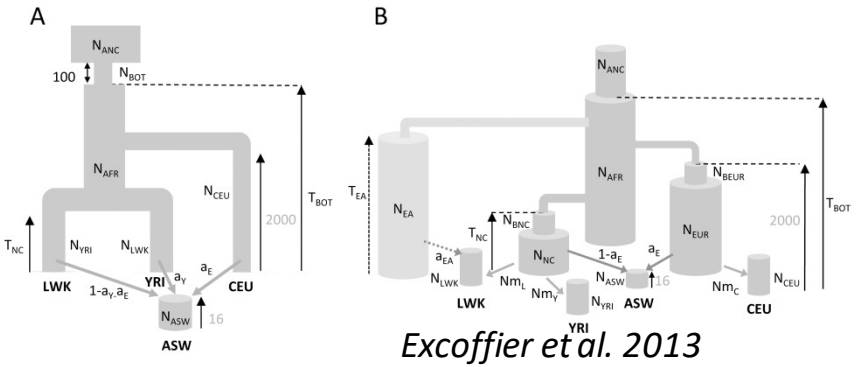


Beichman et al. 2018 Annual Review of Ecology, evolution & systematics

Personal recommendations:

- Inferences based on genetic clusters (panmictic units) rather than sampling units
- Remove first generations or very recent hybrids (e.g. q-values < 0.9)
- Reasonable number of units (4 or 5 pop (or sp) max) [& parameters]
- Use haplotype information, if possible

- Always include models that are <u>biologically impossible</u>
- Test the greatest number of models possible
- But…
  - Be realistic in your hypotheses (N ≠ long-term Ne)
  - Be sequential in your testing
  - Check the robustness of your inferences



Excoffier et al. 2013

- Relative divergence time (e.g. inferred T_{SC}/ inferred T_{SPLIT}) are more accurate than absolute estimates in years (*i.e.* scaling to real time as usually done), due to uncertainties in the mutation rates, generation time...
- Avoid overinterpretation (e.g. focusing on median values of the posterior estimates)
- Remain sceptical (even if you or your PI already envisage a publication in nature)