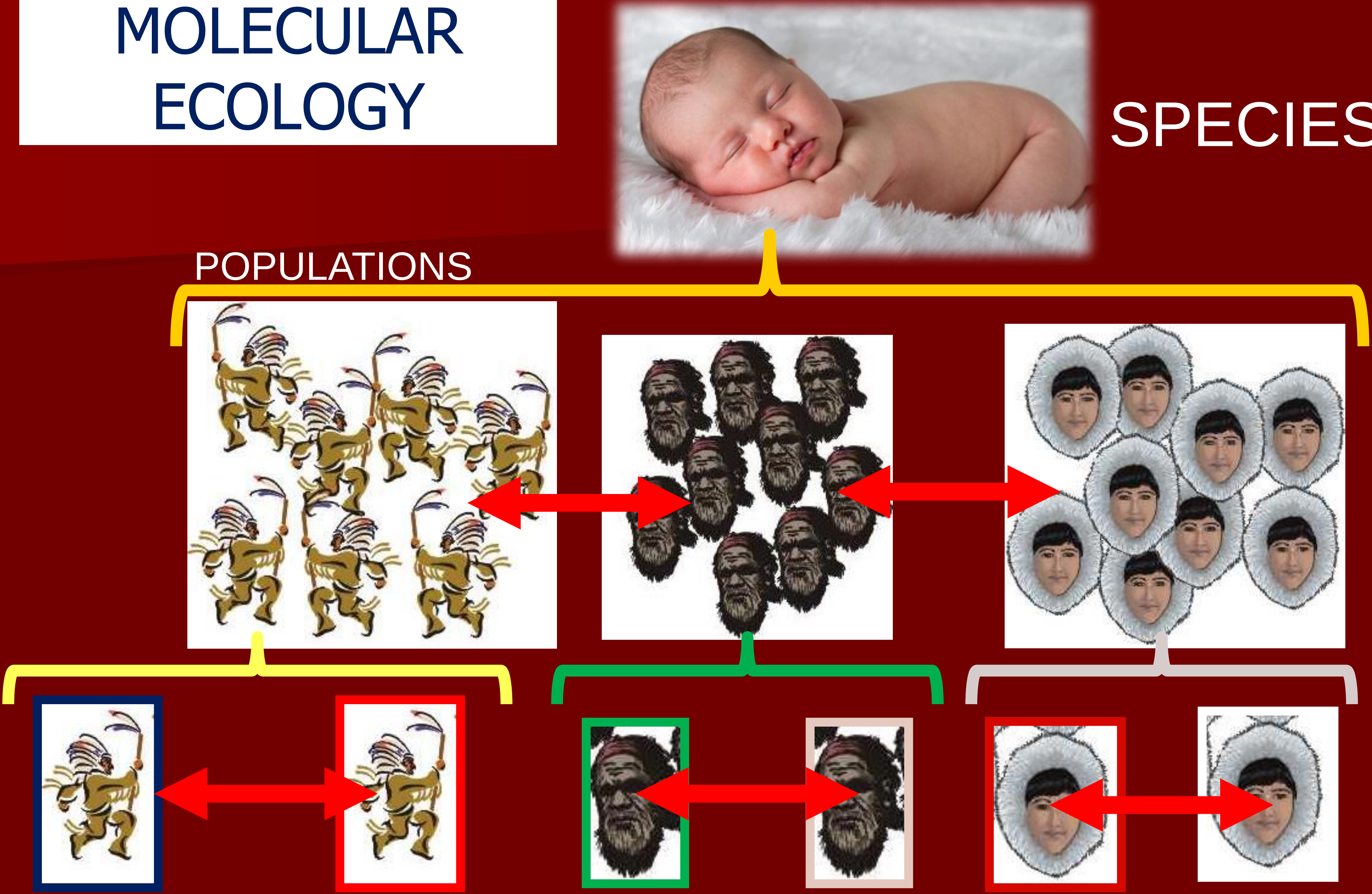
MOLECULAR ECOLOGY



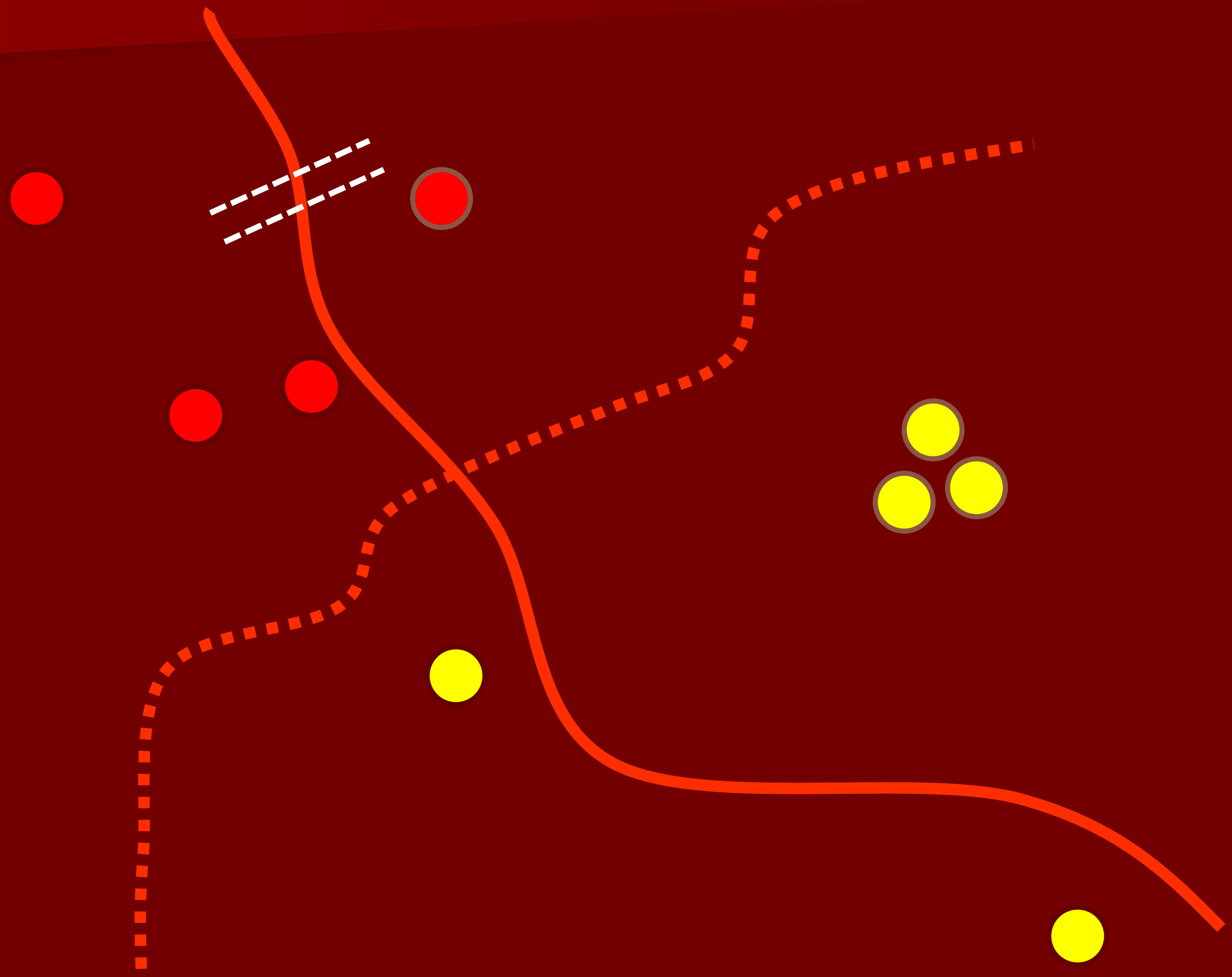
SPECIES

POPULATIONS

SUBPOPULATIONS
(DEMES)

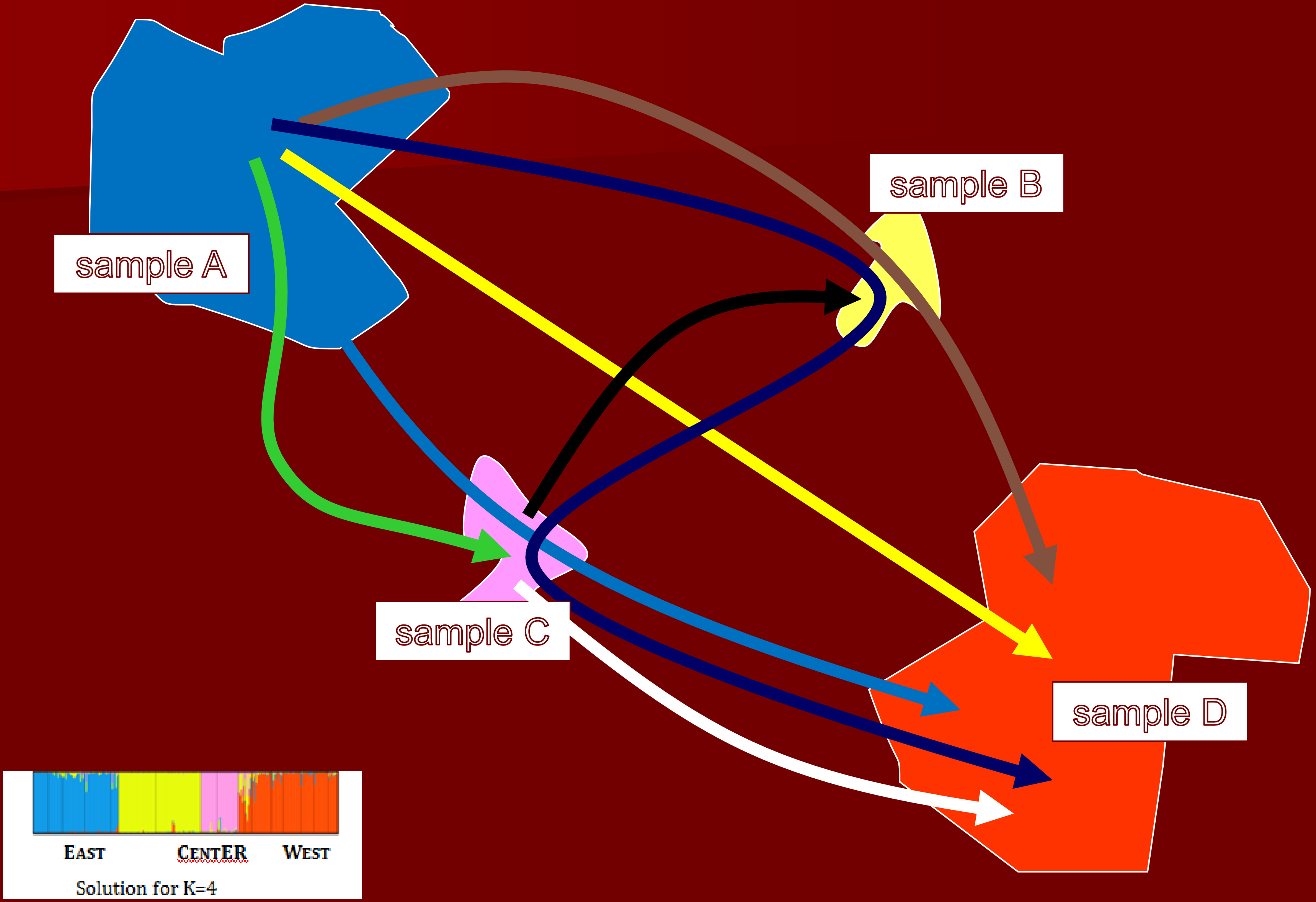
III. POPULATION HISTORY MODELLING

We are interested in genetic structure of a population(s) and HOW HAS BEEN CREATED



Historical data available

SOURCE AREA



COLONIZED AREA

Genetic data available

Population history (& genetic data)

- Past evolutionary and demographic processes have left traces in the genetic variation – analyzing them we attempt to reconstruct evolutionary history of populations

- Studying population history = modelling
  - Selection of the most appropriate model (evolutionary scenario)
  - Estimation of parameters (e.g. time of events, number of founders, duration of bottlenecks, population size, mutation rate)

- Description of recent invasions (invasion genetics)
- Description of older history (phylogeography)

Inferring population history – ABC modelling

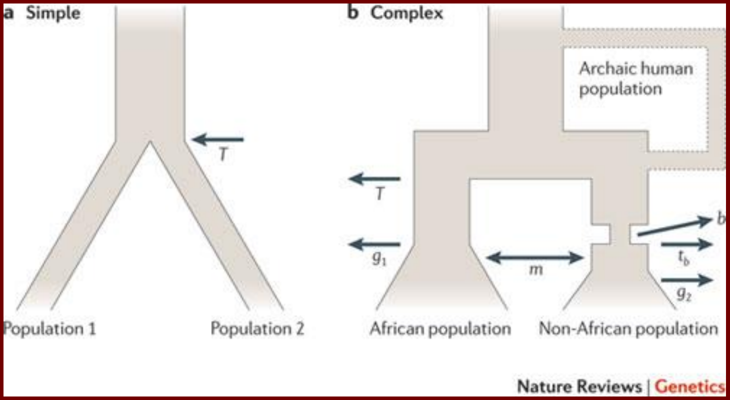
- We have observed data (e.g. microsatellite genotypes)
- We know genetic variation and structure

- We would like to know which demographic processes and how and when have created such an observed data = population evolutionary history

- Why is ABC approach useful in modelling population history?

It allows to deal with much more complex models with many parameters and a lot of complex data (many samples, populations, genetic loci, sequences)

and hence models much more realistic



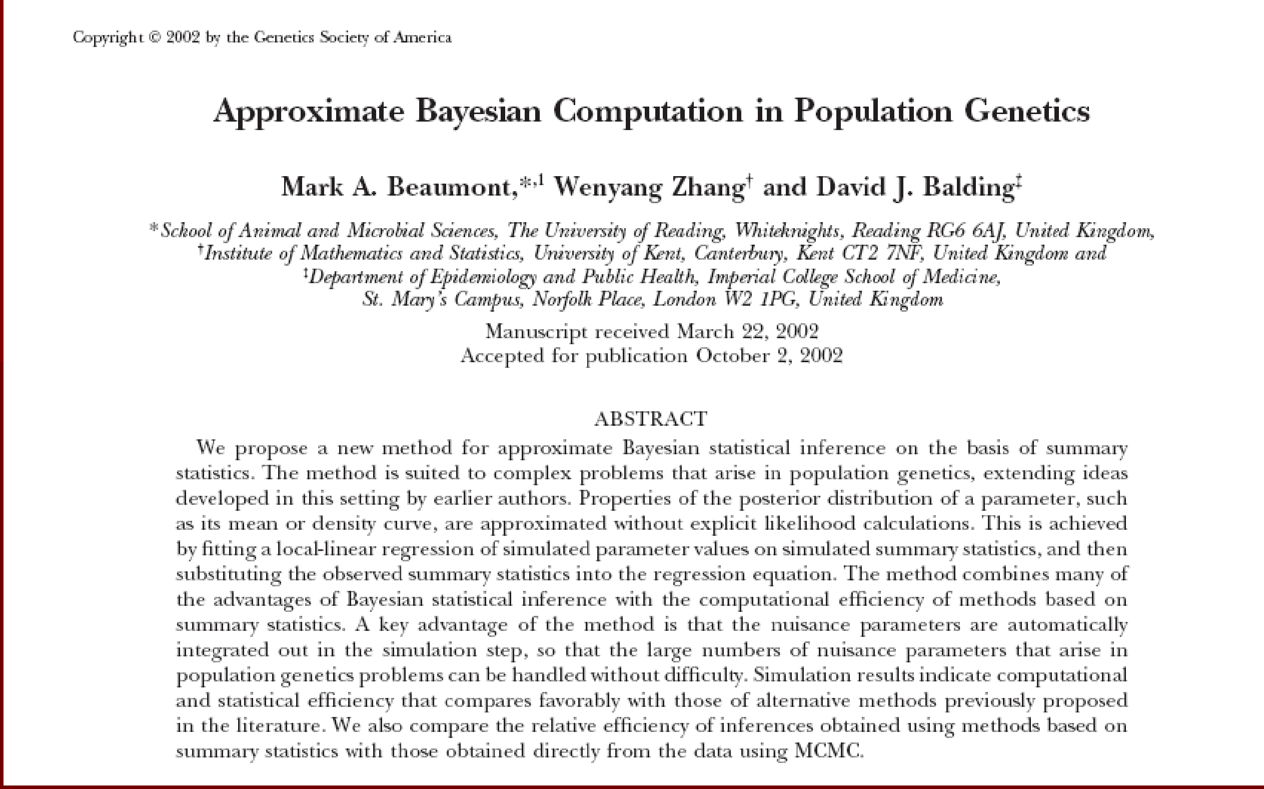
Nature Reviews | Genetics

Approximate Bayesian Computation

- model choice and parameter estimation

- exact LIKELIHOOD function is intractable in complex situations and can be bypassed (approximated) by a SIMILARITY MEASURE between many simulated (under various models) and a single real (observed) data

- data SIMULATION under various models
- COMPARISON of simulated and observed data – model choice
- Acoording to the most supported model we can ESTIMATE VALUES of its parameters – parameter estimation



Approximate Bayesian Computation in Population Genetics

Mark A. Beaumont,*[,1] Wenyang Zhang[†] and David J. Balding[‡]

*School of Animal and Microbial Sciences, The University of Reading, Whiteknights, Reading RG6 6AJ, United Kingdom,
[†]Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent CT2 7NF, United Kingdom and
[‡]Department of Epidemiology and Public Health, Imperial College School of Medicine,
St. Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom

Manuscript received March 22, 2002
Accepted for publication October 2, 2002

ABSTRACT

We propose a new method for approximate Bayesian statistical inference on the basis of summary statistics. The method is suited to complex problems that arise in population genetics, extending ideas developed in this setting by earlier authors. Properties of the posterior distribution of a parameter, such as its mean or density curve, are approximated without explicit likelihood calculations. This is achieved by fitting a local-linear regression of simulated parameter values on simulated summary statistics, and then substituting the observed summary statistics into the regression equation. The method combines many of the advantages of Bayesian statistical inference with the computational efficiency of methods based on summary statistics. A key advantage of the method is that the nuisance parameters are automatically integrated out in the simulation step, so that the large numbers of nuisance parameters that arise in population genetics problems can be handled without difficulty. Simulation results indicate computational and statistical efficiency that compares favorably with those of alternative methods previously proposed in the literature. We also compare the relative efficiency of inferences obtained using methods based on summary statistics with those obtained directly from the data using MCMC.

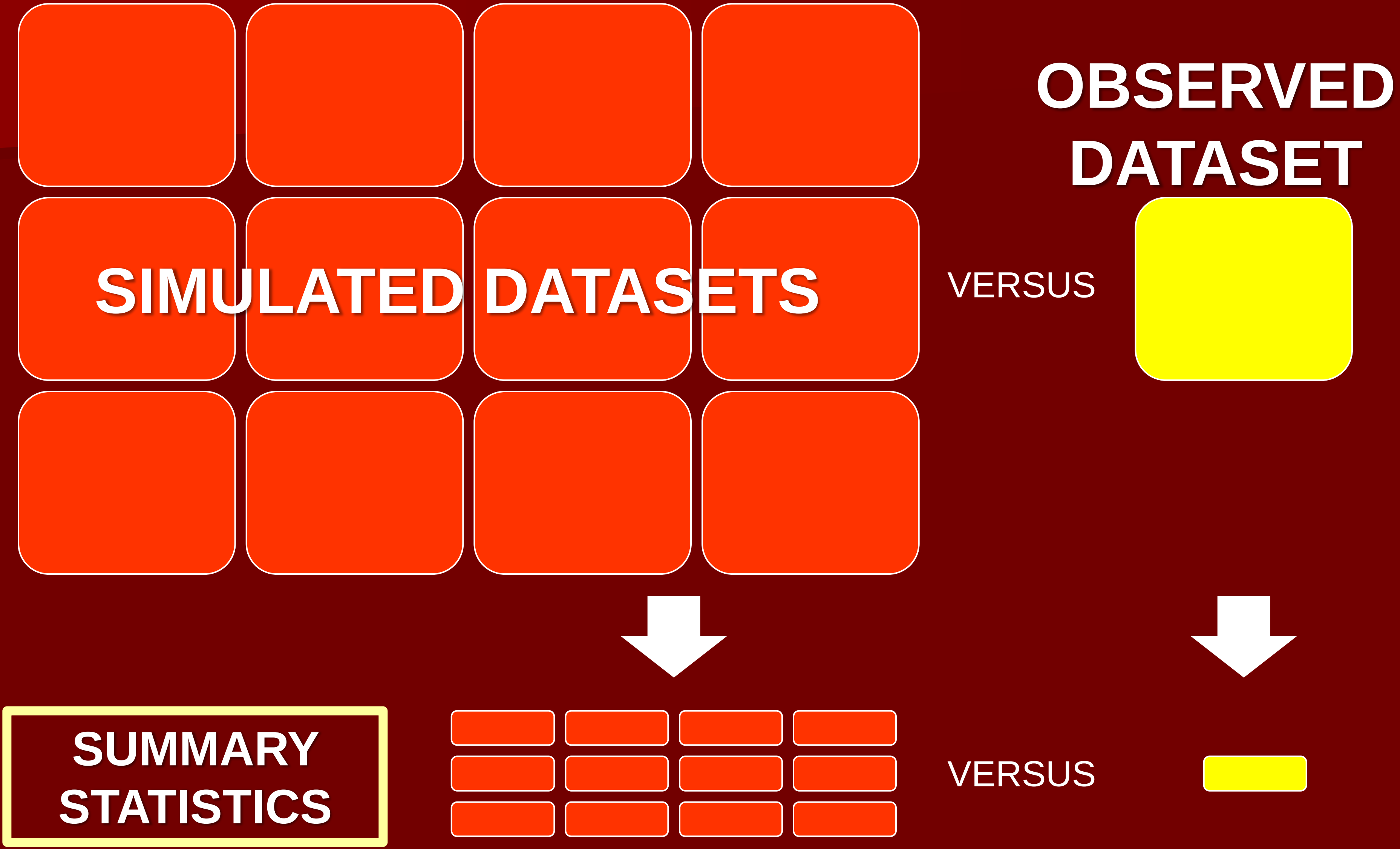
NEW APPROACH

Approximate Bayesian Computation (ABC)

Beaumont et al. 2002, Genetics

- estimations of parameters
- useful for model choice among various scenarios applied on the same data
- **the likelihood criterion is replaced by a similarity criterion between simulated & observed datasets**
- measured by a distance between summary statistics computed on both datasets

Decreasing of dimensionality



SIMULATED DATASETS

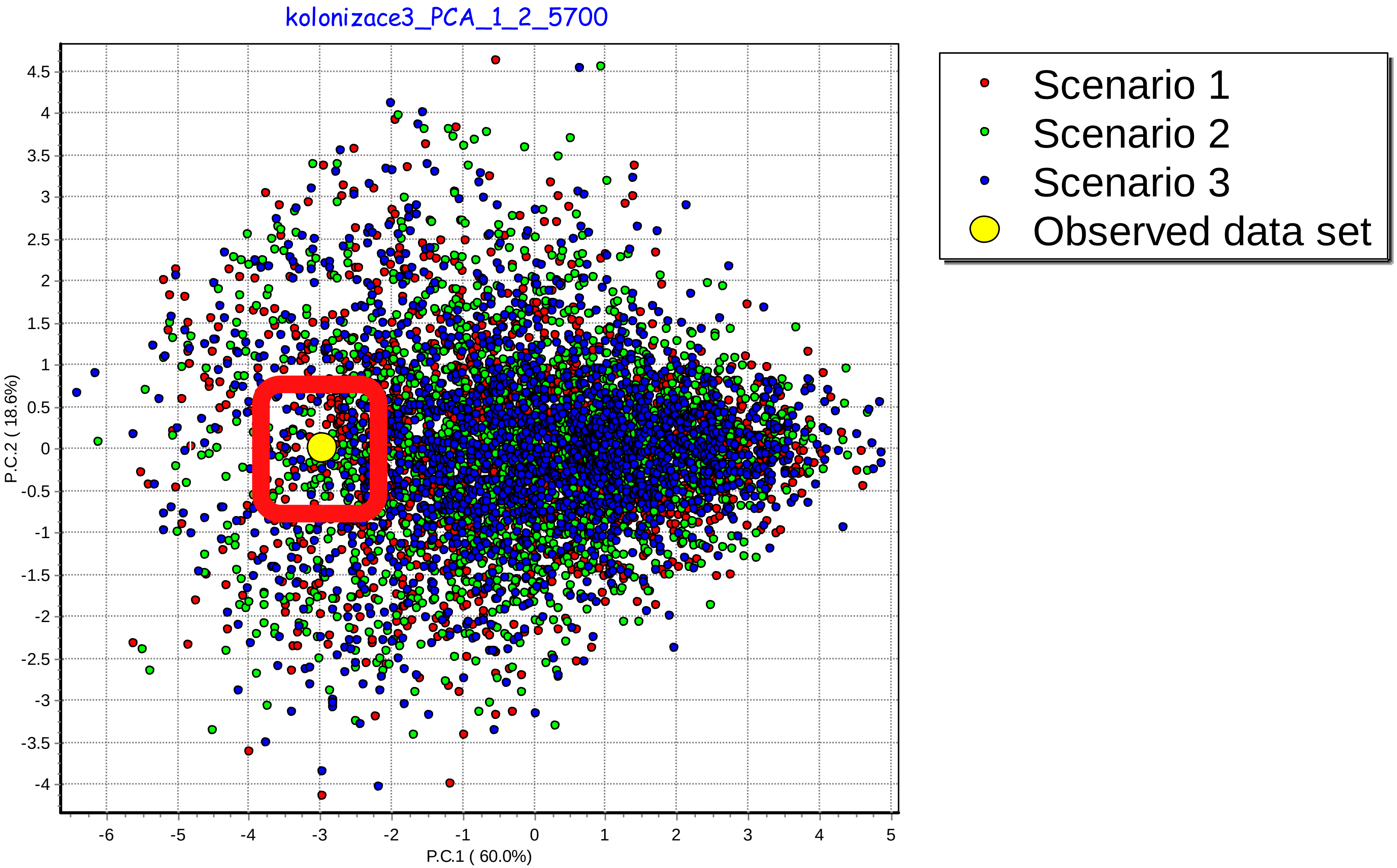
VERSUS

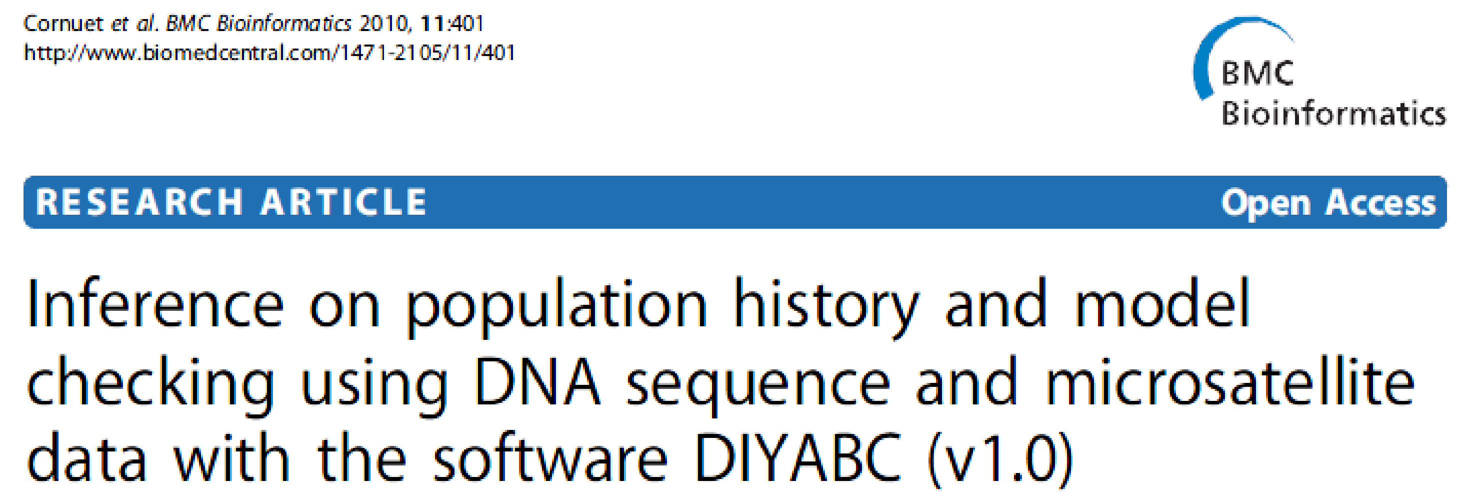
OBSERVED DATASET

SUMMARY STATISTICS

VERSUS

Comparison of simulated and real dataset to infer probability of various models (evolutionary scenarios of population history)





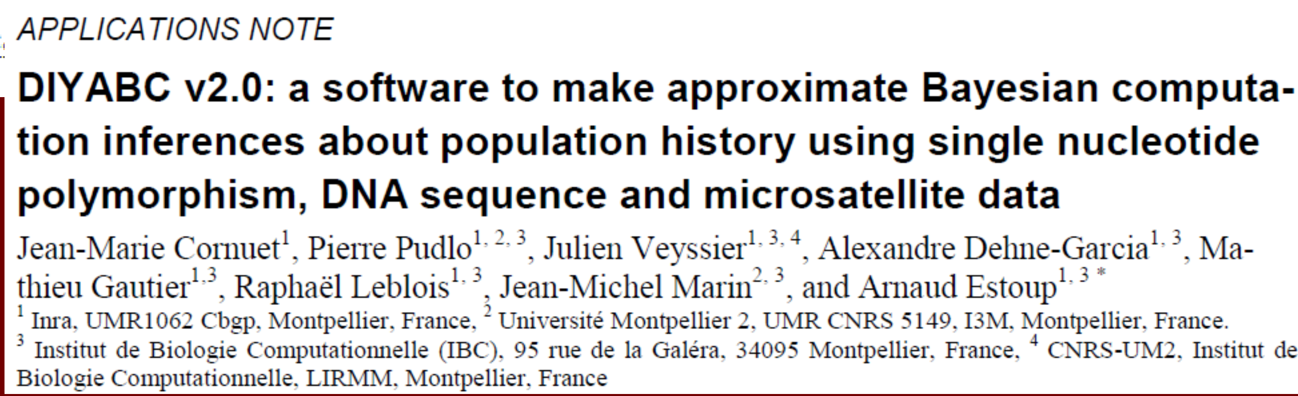
Cornuet *et al. BMC Bioinformatics* 2010, **11**:401
http://www.biomedcentral.com/1471-2105/11/401

BMC Bioinformatics

RESEARCH ARTICLE **Open Access**

Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0)

Jean-Marie Cornuet[1], Virgine Ravigné[2], Arnaud E



APPLICATIONS NOTE

DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data

Jean-Marie Cornuet[1], Pierre Pudlo[1, 2, 3], Julien Veyssier[1, 3, 4], Alexandre Dehne-Garcia[1, 3], Mathieu Gautier[1,3], Raphaël Leblois[1, 3], Jean-Michel Marin[2, 3], and Arnaud Estoup[1, 3 *]

[1] Inra, UMR1062 Cbgp, Montpellier, France, [2] Université Montpellier 2, UMR CNRS 5149, I3M, Montpellier, France.
[3] Institut de Biologie Computationnelle (IBC), 95 rue de la Galéra, 34095 Montpellier, France, [4] CNRS-UM2, Institut de Biologie Computationnelle, LIRMM, Montpellier, France

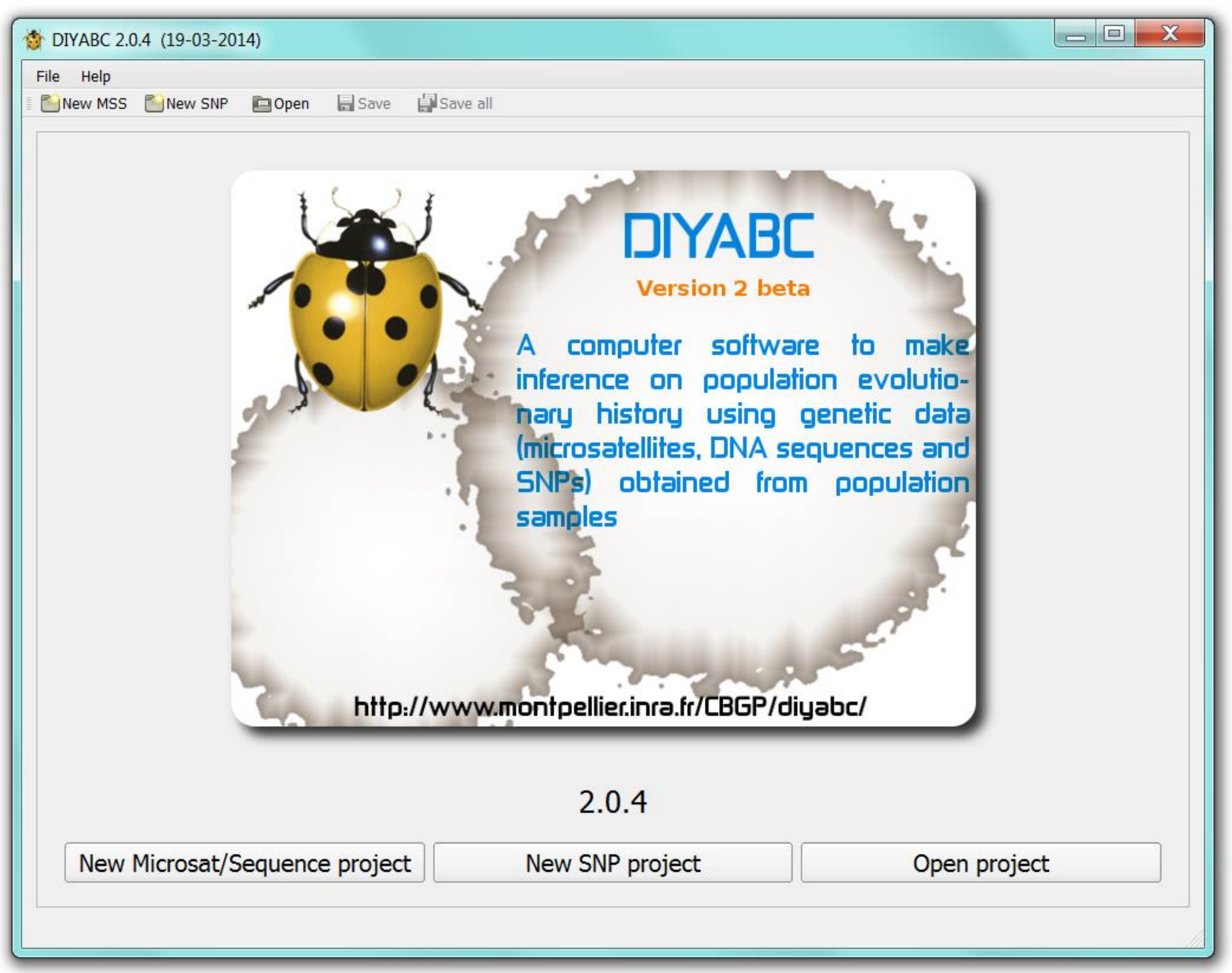
no simple software solution => inaccessible to most biologists

BUT NOW → Do It Yourself: **DIYABC** software

allows to infer populaton history using the ABC approach

(Cornuet et al. 2008, 2010, 2014)

DIYABC



DIYABC

Version 2 beta

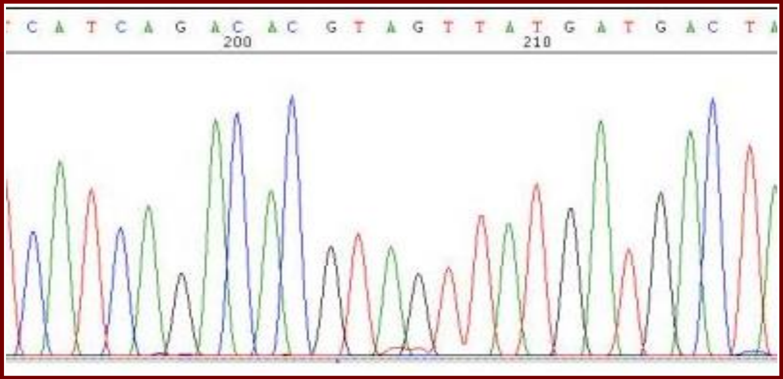
A computer software to make inference on population evolutionary history using genetic data (microsatellites, DNA sequences and SNPs) obtained from population samples

http://www.montpellier.inra.fr/CBGP/diyabc/

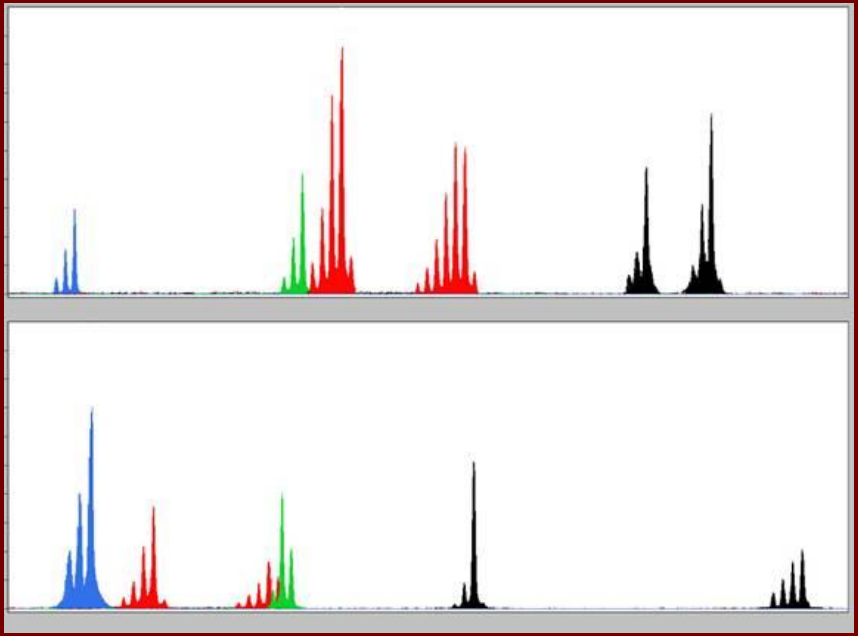
2.0.4

Genetic data

- Sequences
- SNPs

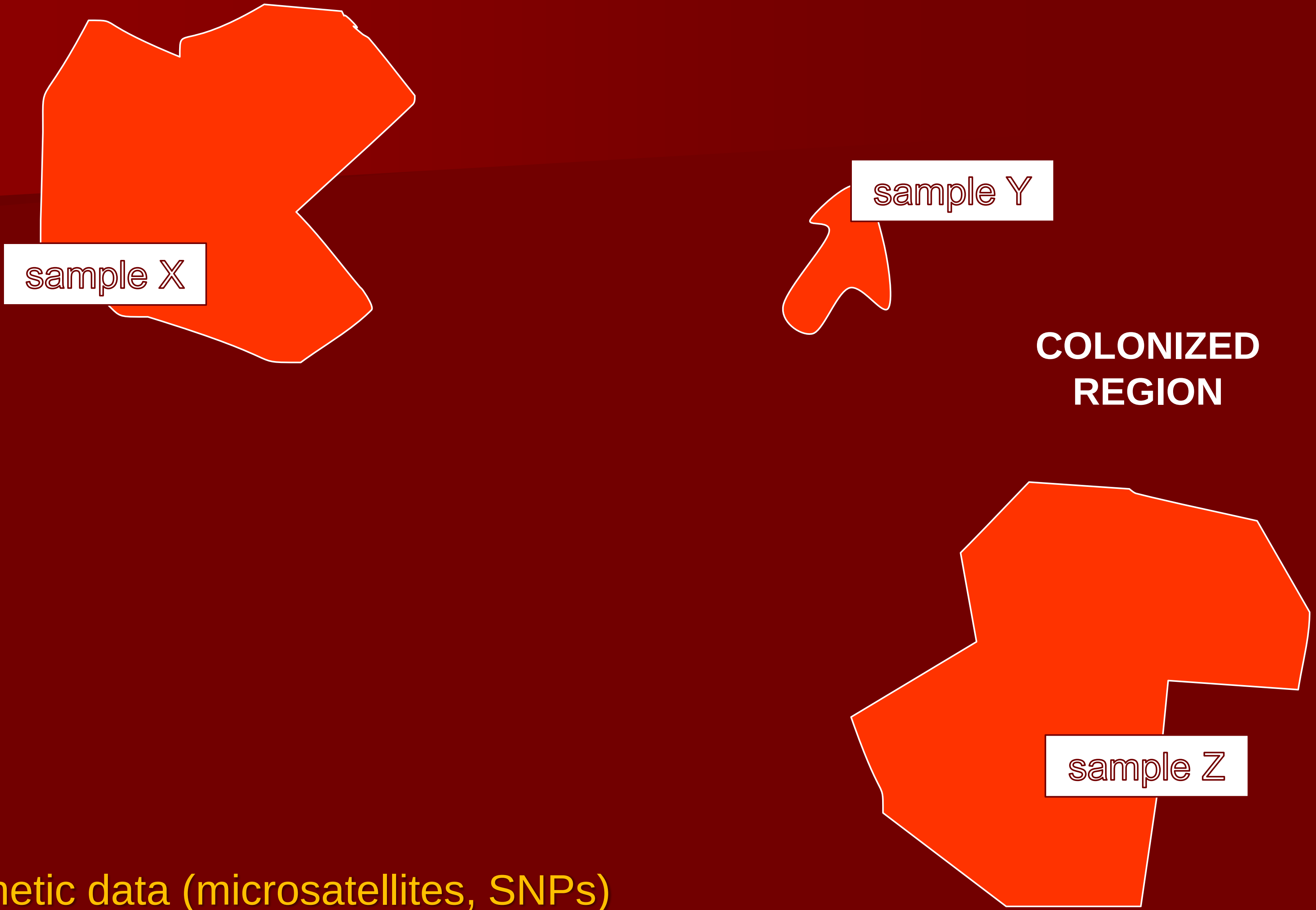


- Genotypes



SOURCE REGION

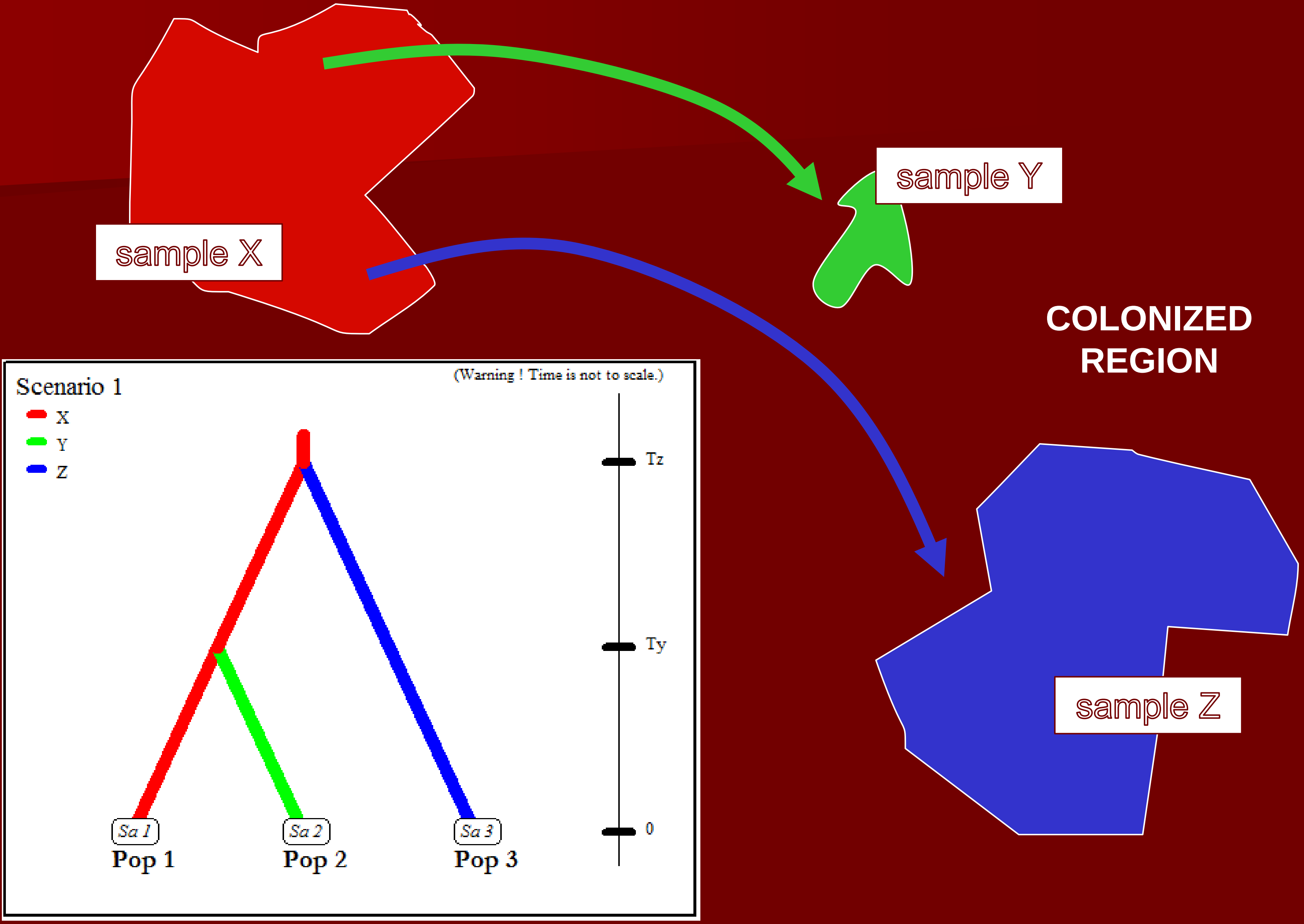
Historic background



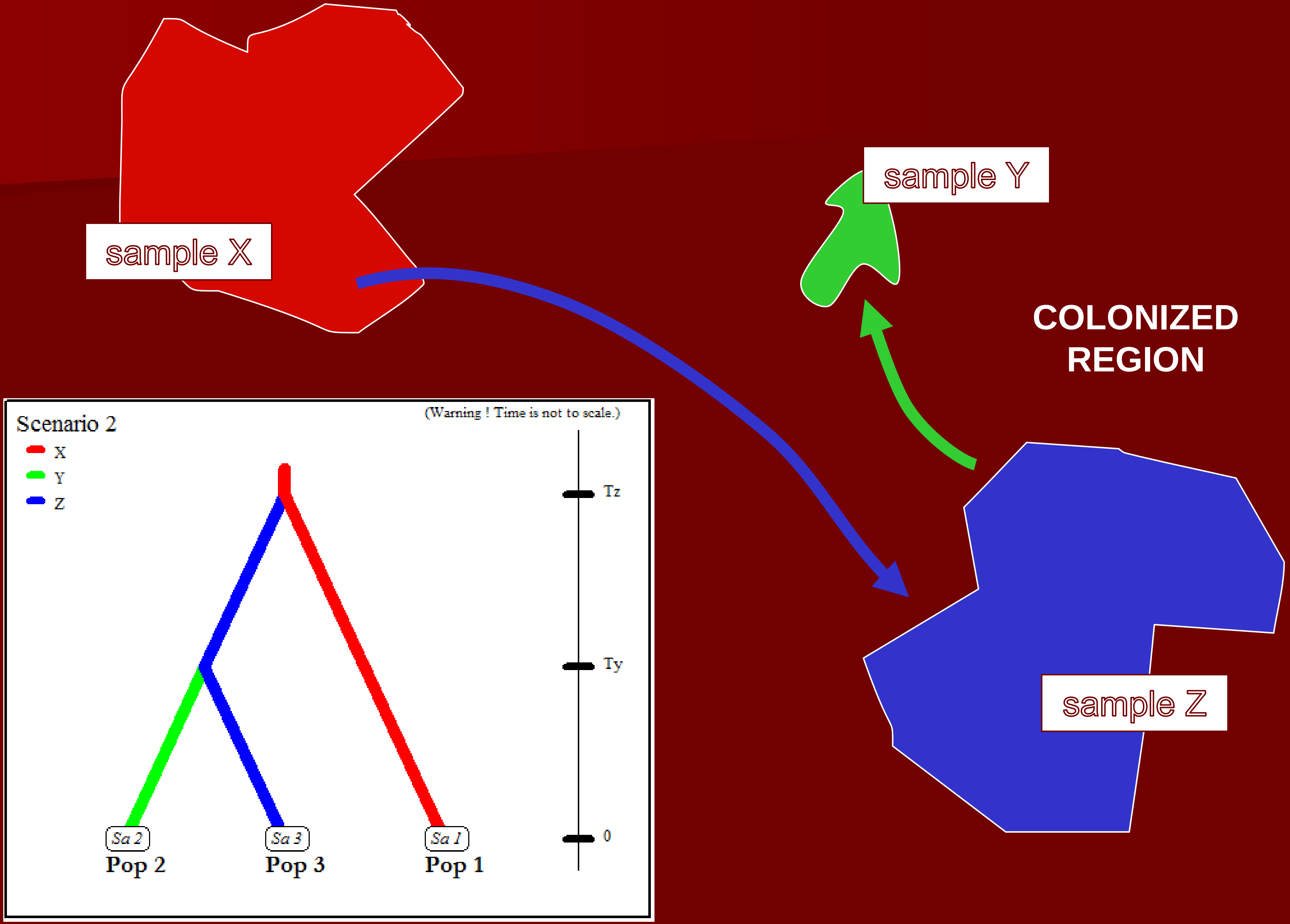
COLONIZED REGION

Genetic data (microsatellites, SNPs)

Historic background



Historic background

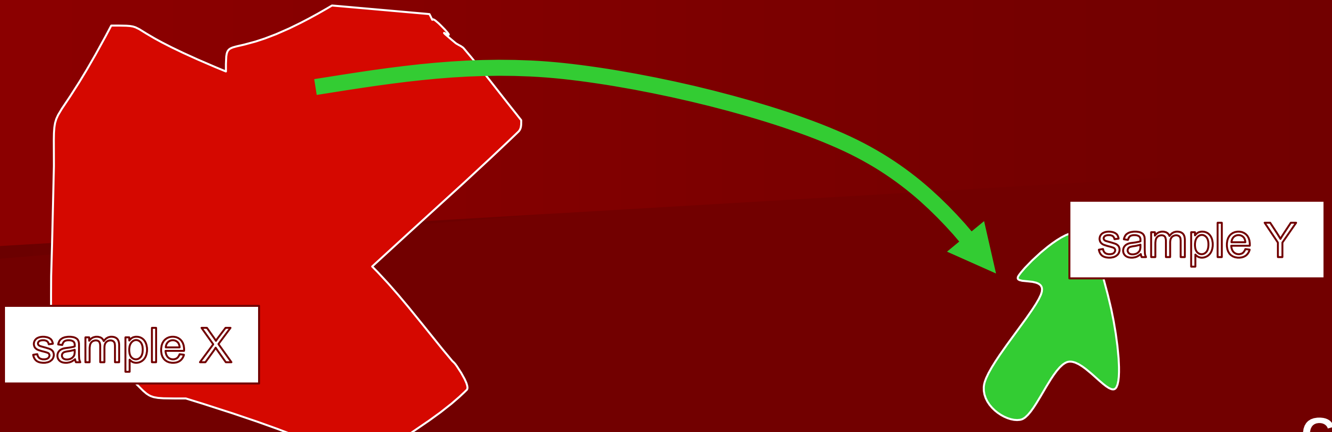


SOURCE REGION

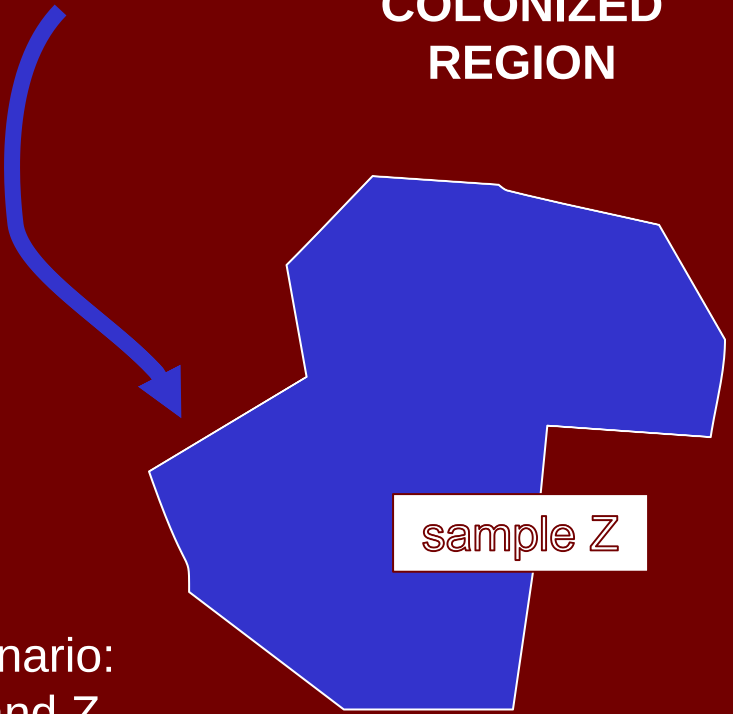
COLONIZED REGION

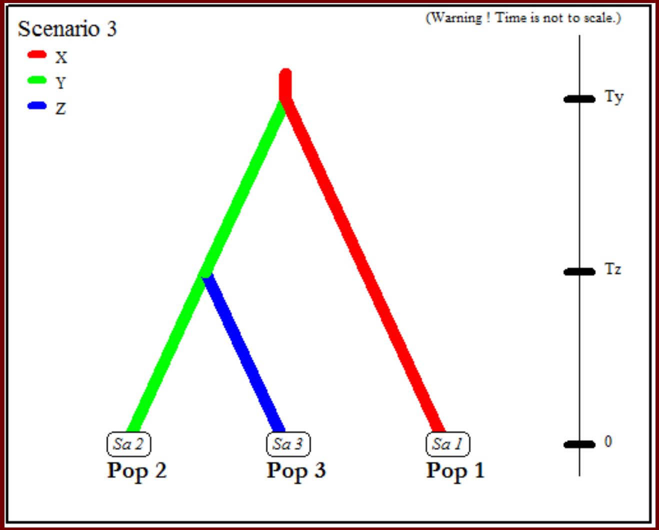
SOURCE REGION

Historic background

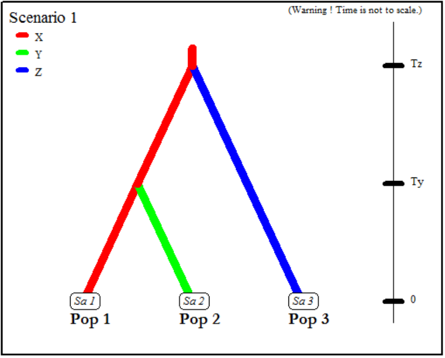


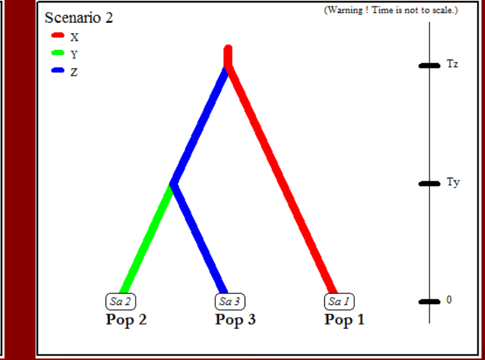
COLONIZED REGION

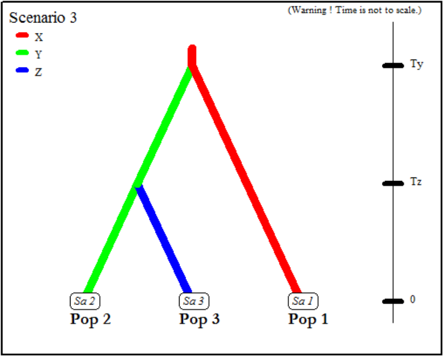




Prior distribution of parameters describing the scenario:
Ty, Tz --- divergence times – establishment of Y and Z populations
Uniform distribution (min 100, max 500 generations)







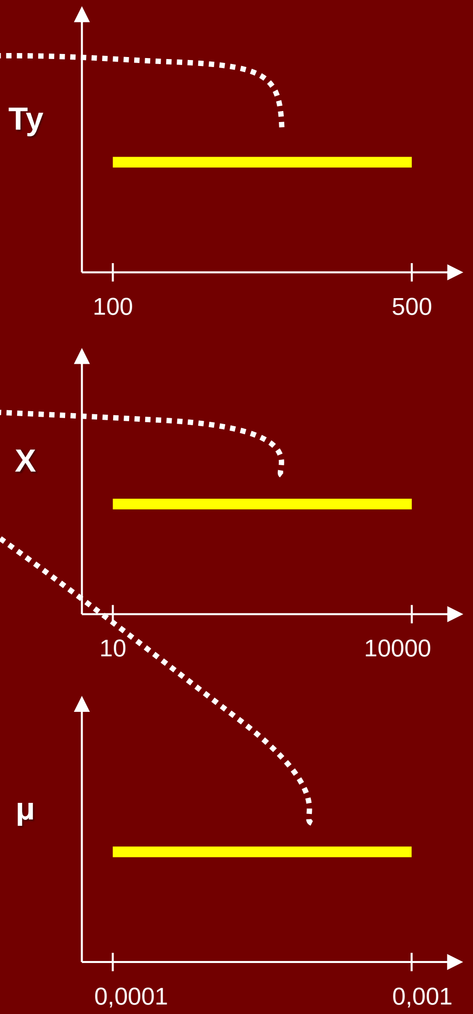
Prior distribution of parameters describing the model

Evolutionary scenarios = models

SIMULATED DATASETS

Genetic data → summary statistics

scenario	X	Y	Z	Ty	Tz	μ	mean number of alleles			mean heterozygosity		
2	3797	7013	9839	484	486	0.00083	8.4	13.2	11.3	0.7841	0.8669	0.8589
3	3648	1355	1206	453	209	0.00072	7.9	6.1	4	0.7894	0.6371	0.5465
1	6802	7945	3929	176	346	0.0003	8.8	11.4	7.1	0.7877	0.8367	0.7824
1	4715	9090	5767	290	301	0.00048	7.5	12.6	9.1	0.7842	0.8211	0.7919
3	134	2714	3804	406	342	0.00029	1.4	4.8	4.7	0.0651	0.5182	0.5906
1	9331	902	4882	305	197	0.00096	13.6	6.5	13	0.863	0.5471	0.8294
3	1912	1785	6813	385	124	0.00035	4.3	5.5	7.1	0.5924	0.6414	0.7134

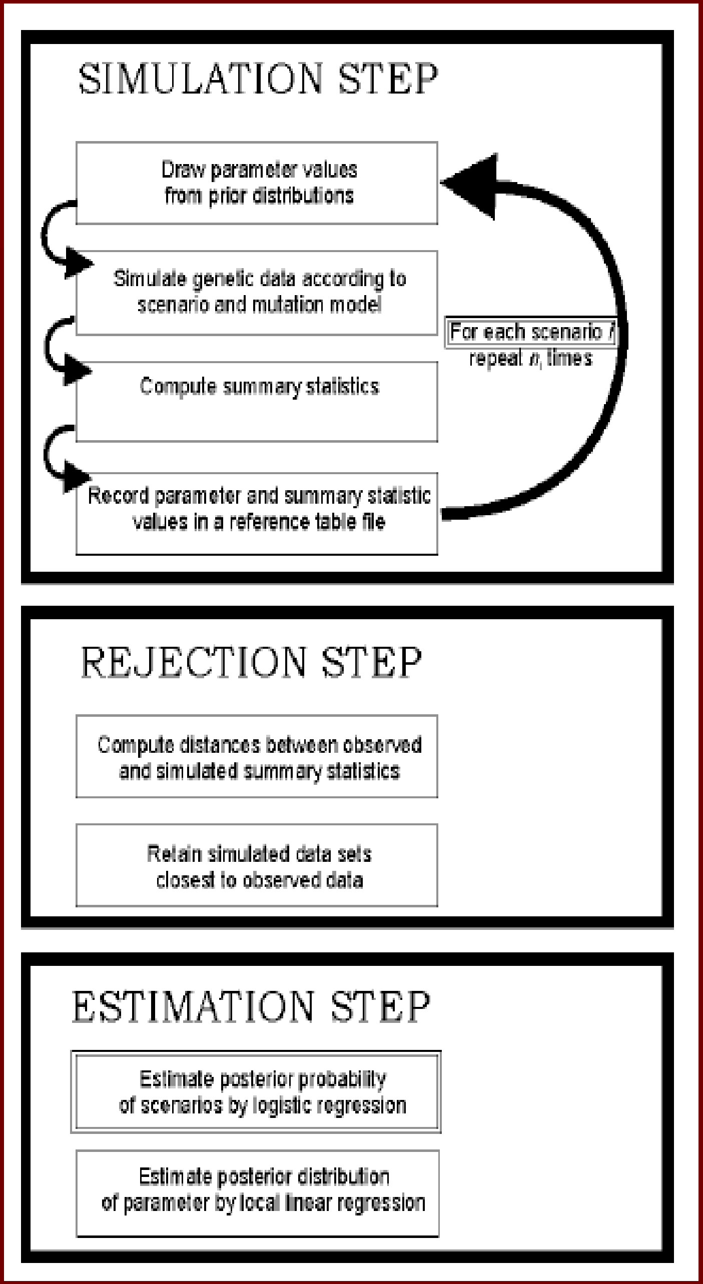


DIYABC works in 3 steps

1. SIMULATION STEP:
a very large ***reference table*** is produced and recorded

2. REJECTION STEP:
only the simulated data closest to the observed dataset are retained

based on Euclidian distances in multidimensional space of summary statistics



SIMULATION STEP

Draw parameter values from prior distributions

Simulate genetic data according to scenario and mutation model

Compute summary statistics

Record parameter and summary statistic values in a reference table file

For each scenario i repeat n_i times

REJECTION STEP

Compute distances between observed and simulated summary statistics

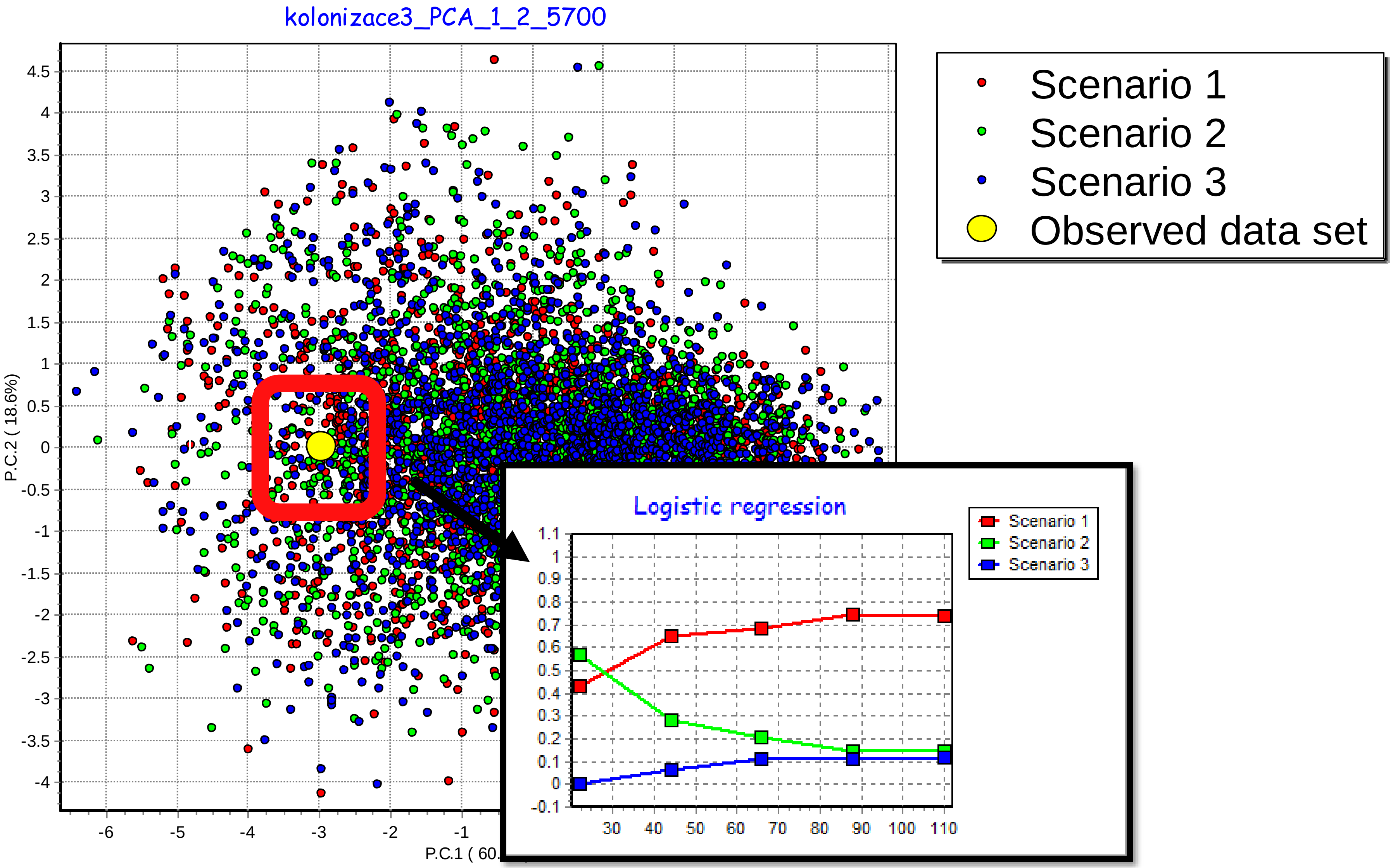
Retain simulated data sets closest to observed data

ESTIMATION STEP

Estimate posterior probability of scenarios by logistic regression

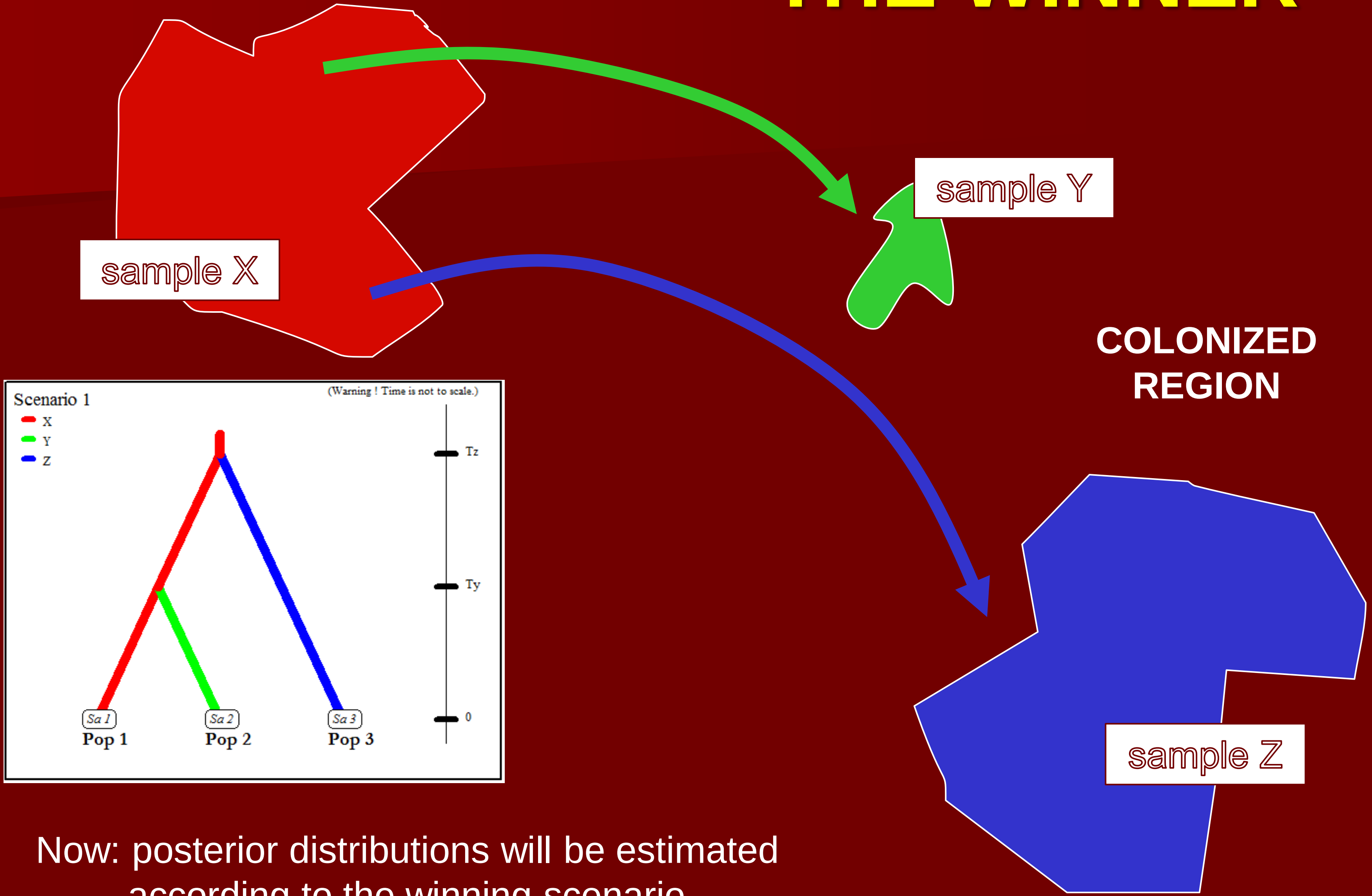
Estimate posterior distribution of parameter by local linear regression

Comparison of our observed dataset with simulated ones and inferring posterior distributions of scenarios



THE WINNER

SOURCE REGION



sample X

sample Y

COLONIZED REGION

sample Z

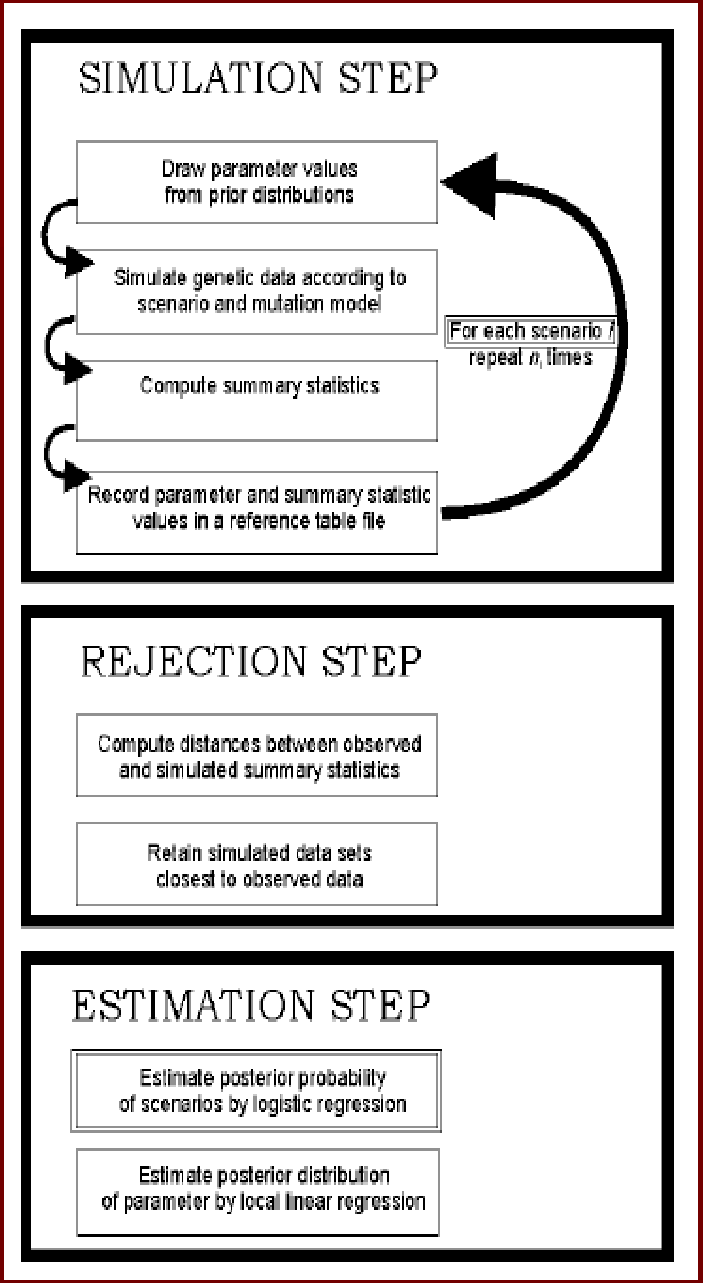
Now: posterior distributions will be estimated according to the winning scenario

DIYABC works in 3 steps

1. SIMULATION STEP:
a very large ***reference table*** is produced and recorded

2. REJECTION STEP:
only the simulated data closest to the observed dataset are retained

3. ESTIMATION STEP:
Estimating posterior distributions of parameters through a local linear regression procedure



SIMULATION STEP

Draw parameter values from prior distributions

Simulate genetic data according to scenario and mutation model

Compute summary statistics

Record parameter and summary statistic values in a reference table file

For each scenario i repeat n_i times

REJECTION STEP

Compute distances between observed and simulated summary statistics

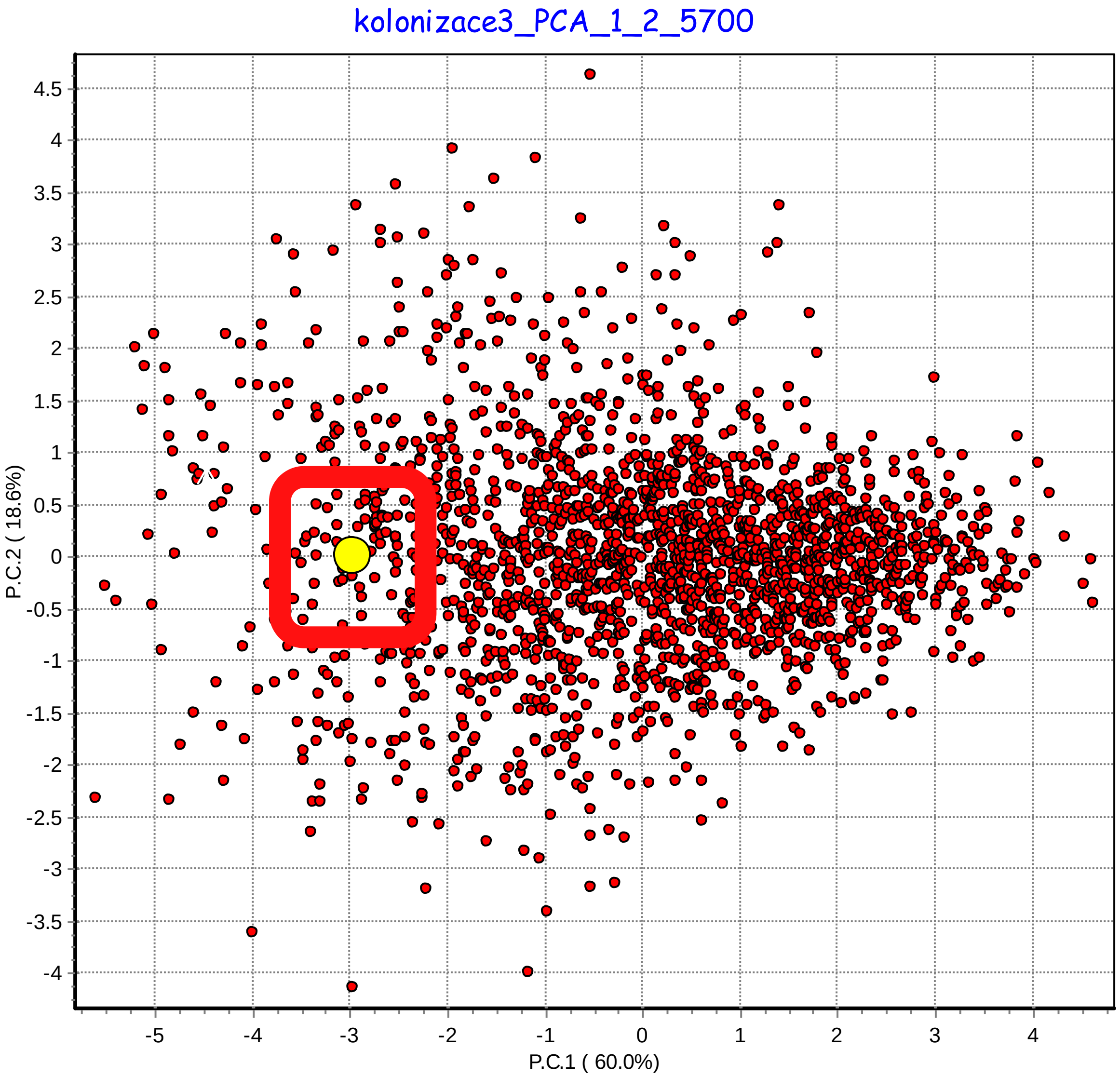
Retain simulated data sets closest to observed data

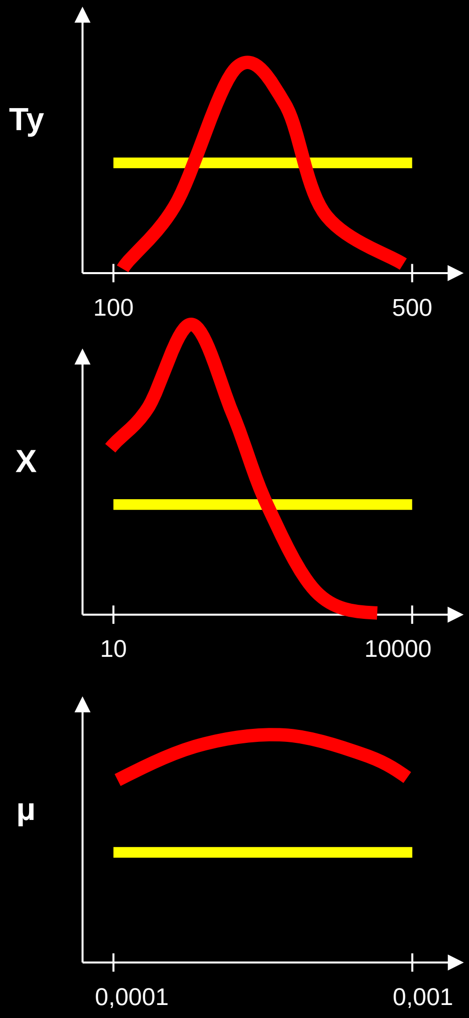
ESTIMATION STEP

Estimate posterior probability of scenarios by logistic regression

Estimate posterior distribution of parameter by local linear regression

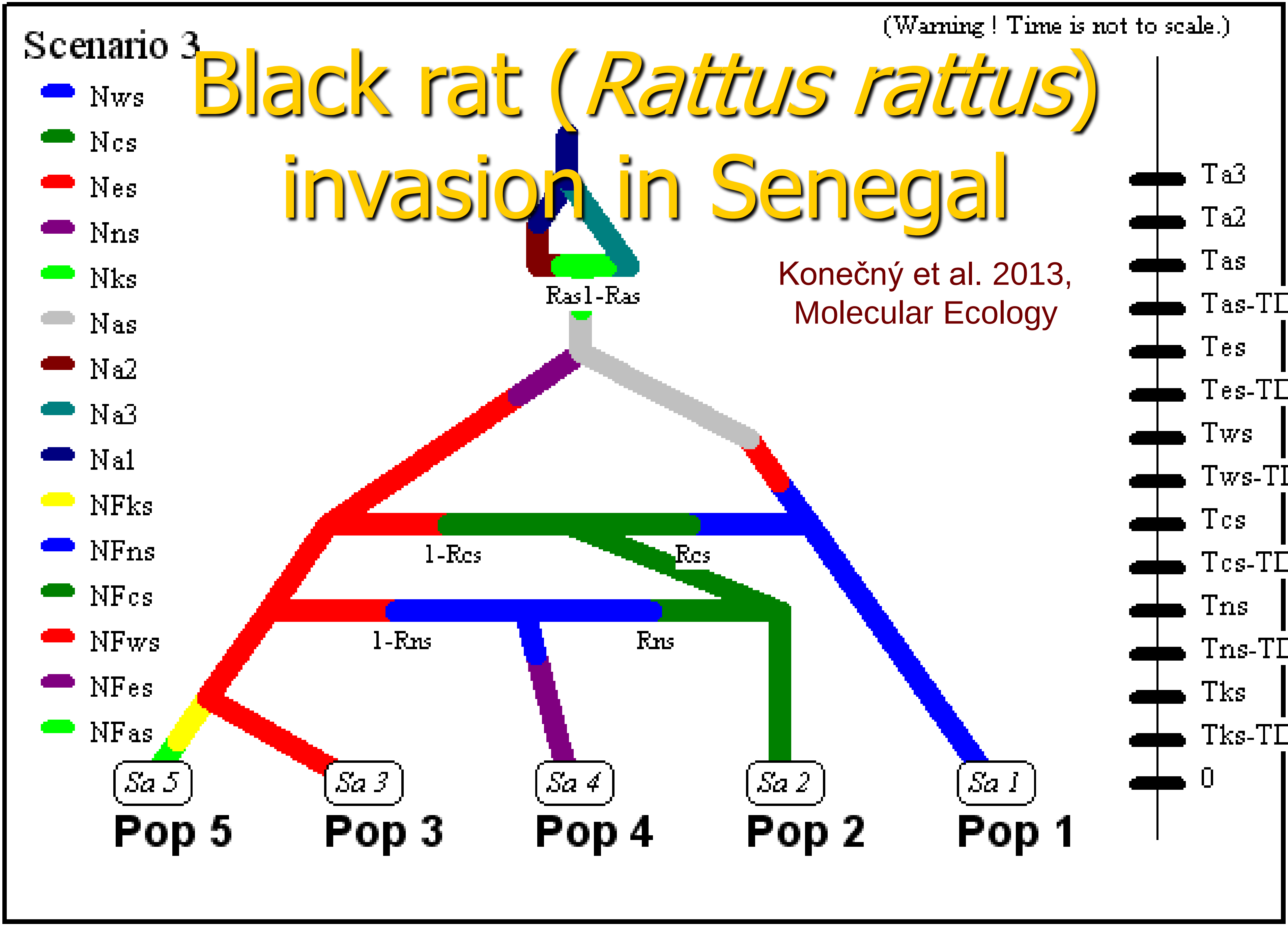
Posterior distributions of parameters are estimated according to the most supported scenario

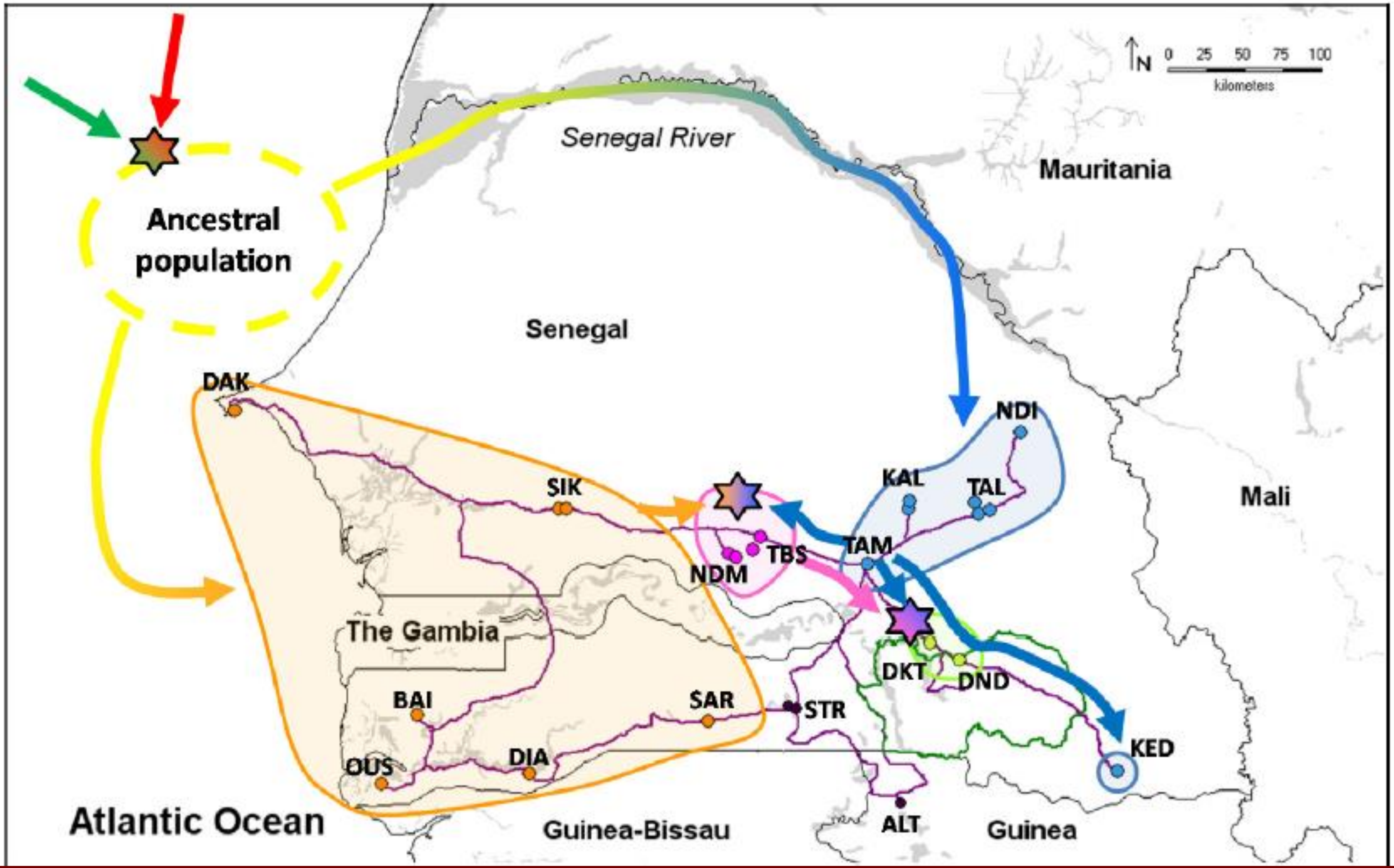




Black rat (*Rattus rattus*) invasion in Senegal

Konečný et al. 2013, Molecular Ecology





Description of *Rattus rattus* spread in Senegal as revealed from ABC

Konečný et al. 2013,
Molecular Ecology