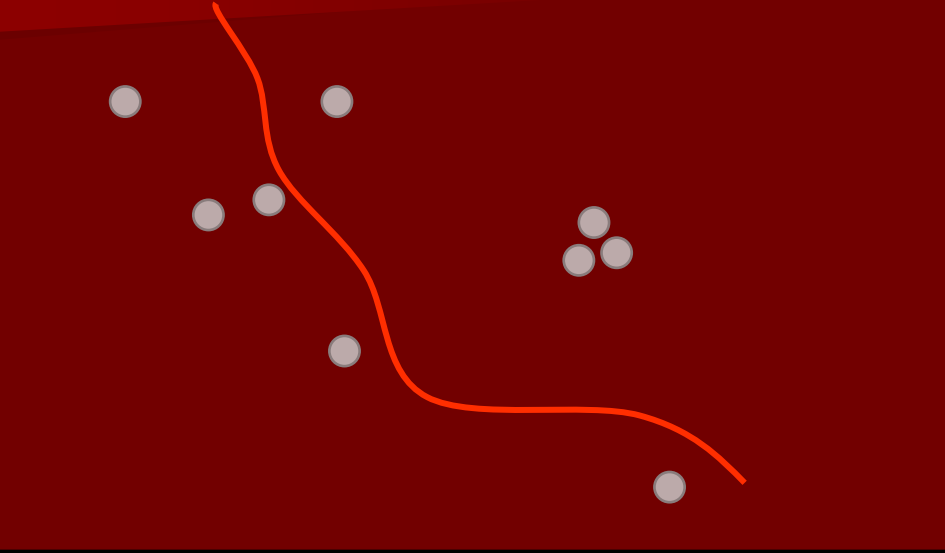
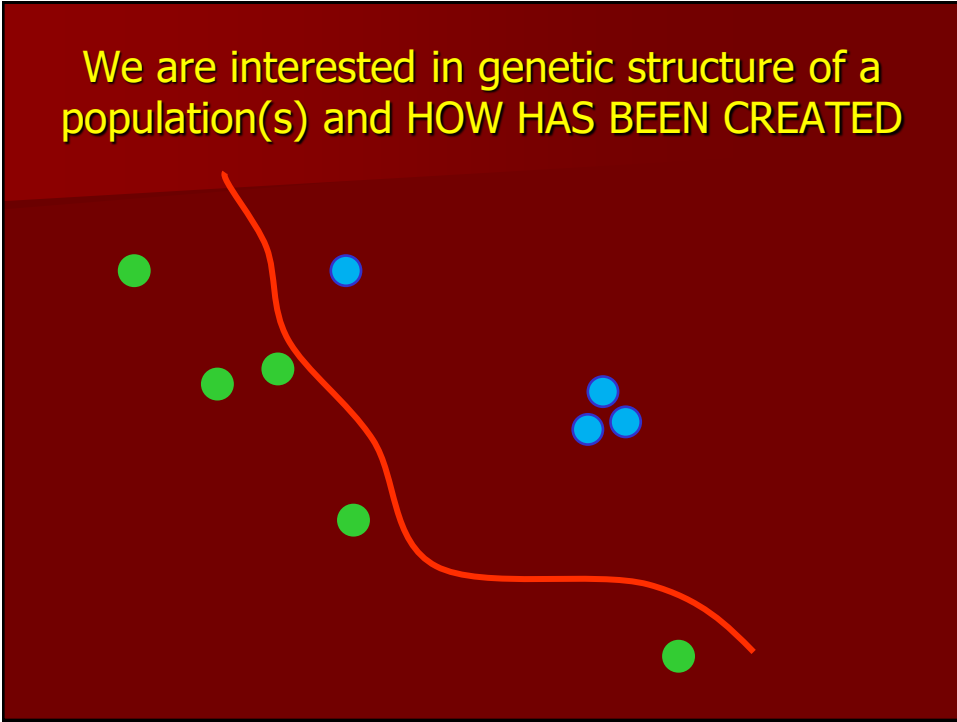


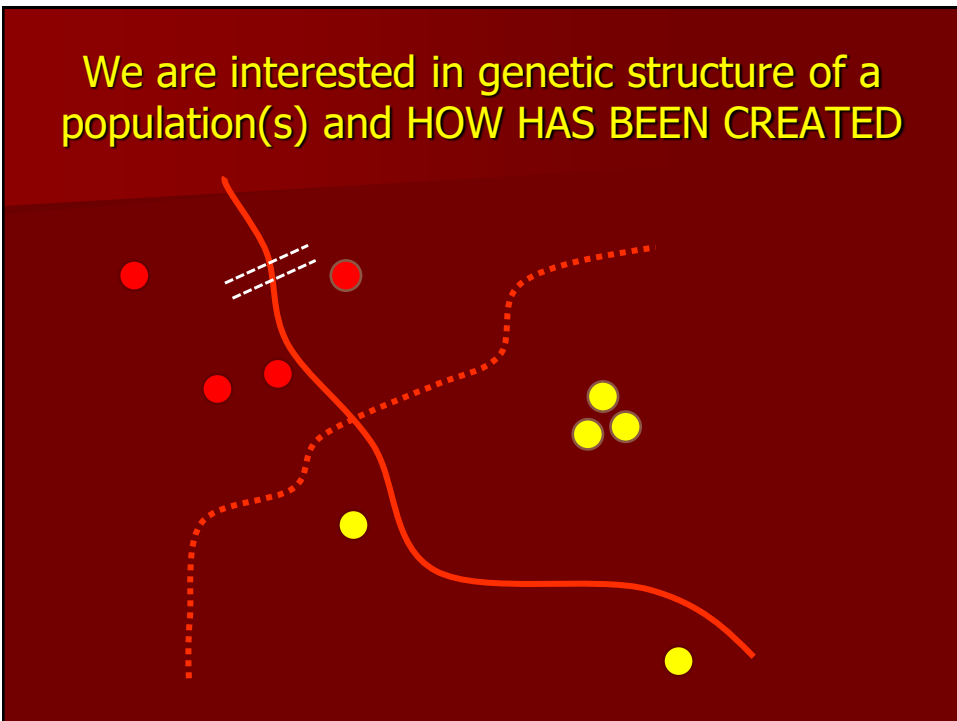
We are interested in genetic structure of a population(s) and HOW HAS BEEN CREATED

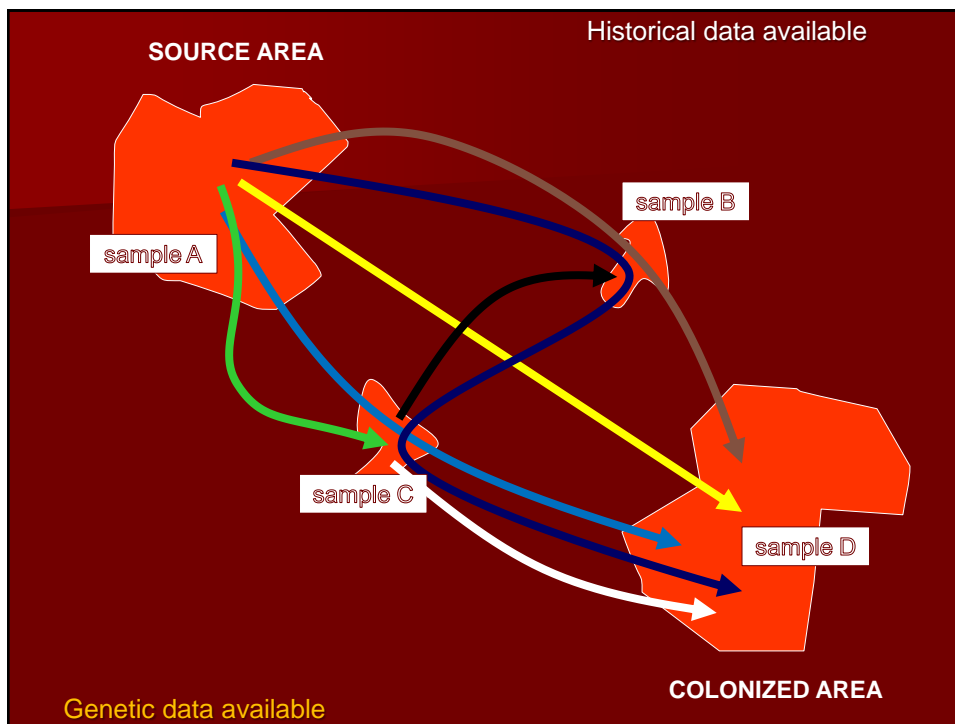


We are interested in genetic structure of a population(s) and HOW HAS BEEN CREATED



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





## Population history (& genetic data)

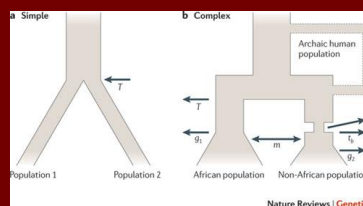
- Past evolutionary and demographic processes have left traces in the genetic variation – analyzing them we attempt to reconstruct **evolutionary history of populations**
- Studying population history = modelling
  - **Selection of the most appropriate model** (evolutionary scenario)
  - **Estimation 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)

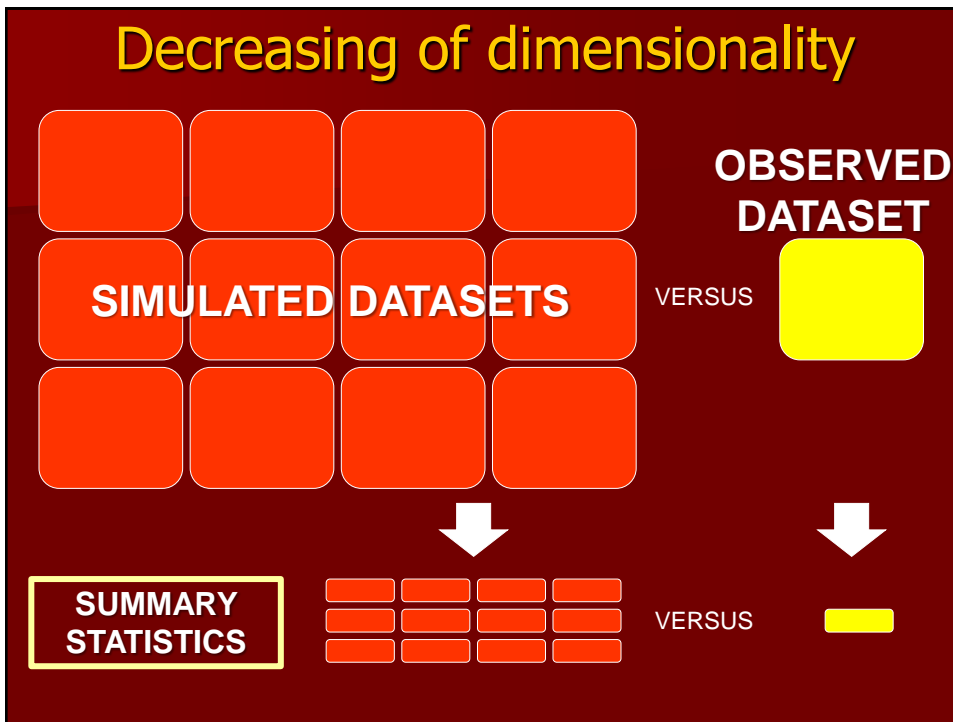
and hence **models much more realistic**



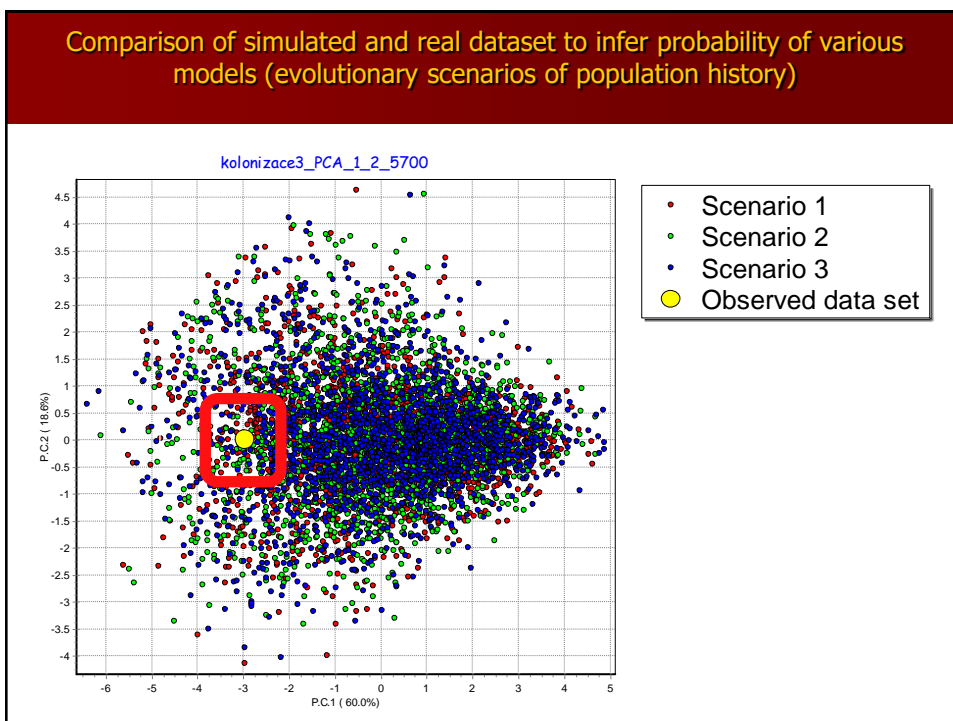
## Approximate Bayesian Computation

- 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**
- According to the most supported model we can **ESTIMATE VALUES** of its parameters – **parameter estimation**

## Decreasing of dimensionality



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



Copyright © 2002 by the Genetics Society of America

## Approximate Bayesian Computation in Population Genetics

Mark A. Beaumont,<sup>a,1</sup> Wenyang Zhang<sup>b</sup> and David J. Balding<sup>c</sup>

<sup>a</sup>*School of Animal and Microbial Sciences, The University of Reading, Whiteknights, Reading RG6 6AJ, United Kingdom,*  
<sup>b</sup>*Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent CT2 7NF, United Kingdom and*  
<sup>c</sup>*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.

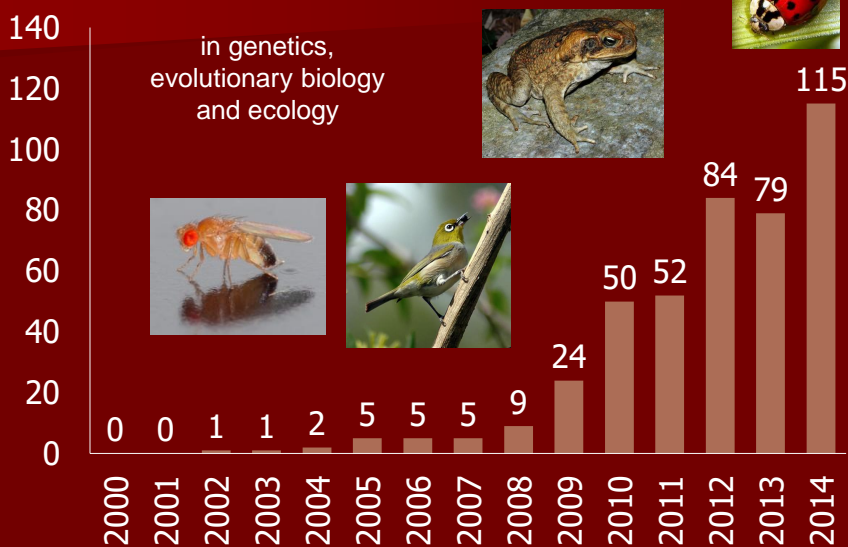
## NEW APPROACH

## Approximate Bayesian Computation (ABC)

Beaumont et al. 2002, Genetics

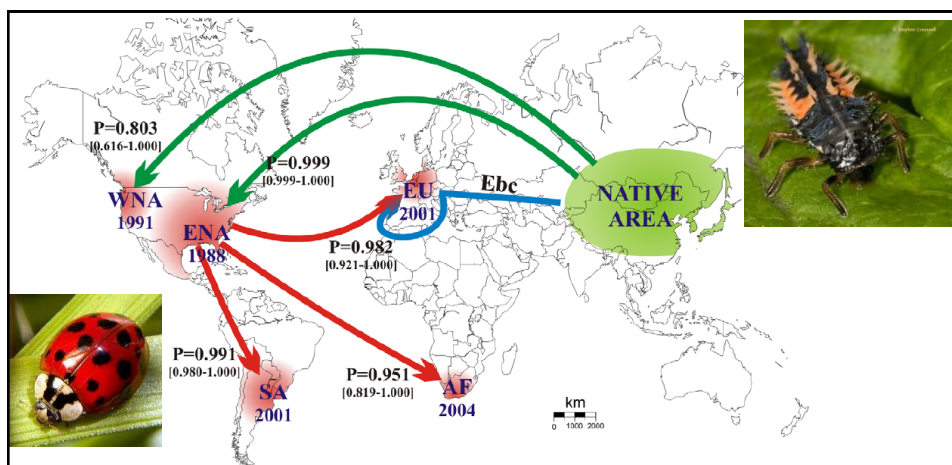
- estimations of parameters
- useful for model choice among various scenarios applied on the same data
- **the likelihood criterion is replaced by a similarity criterion between simulated & observed datasets**
- measured by a distance between **summary statistics** computed on both datasets

## "Approximate Bayesian Computation" topic in Web of Science



## ABC approach used successfully for description of recent invasion scenarios

- Estoup & Clegg 2003, Molecular Ecology: *Zosterops lateralis*, Pacific islands
- Estoup et al. 2004, Evolution: *Bufo marinus*, Australia
- Pascual et al. 2007, Molecular Ecology: *Drosophila subobscura*, invasion over Atlantic ocean



- Lombaert et al. 2010, PLoS ONE: *Harmonia axyridis*, invasion through the Atlantic and subsequently to the whole world



# Software

**Table 3.** Software incorporating ABC.

Software	Keywords and Features	Reference
DIY-ABC	Software for fit of genetic data to complex situations. Comparison of competing models. Parameter estimation. Computation of bias and precision measures for a given model and known parameters values.	[53]
ABC R package	Several ABC algorithms for performing parameter estimation and model selection. Nonlinear heteroscedastic regression methods for ABC. Cross-validation tool.	[54]
ABC-SysBio	Python package. Parameter inference and model selection for dynamical systems. Combines ABC rejection sampler, ABC SMC for parameter inference, and ABC SMC for model selection. Compatible with models written in Systems Biology Markup Language (SBML). Deterministic and stochastic models.	[55]
ABCtoolbox	Open source programs for various ABC algorithms including rejection sampling, MCMC without likelihood, a particle-based sampler, and ABC-GLM. Compatibility with most simulation and summary statistics computation programs.	[56]
msBayes	Open source software package consisting of several C and R programs that are run with a Perl "front-end." Hierarchical coalescent models. Population genetic data from multiple co-distributed species.	[57]
PopABC	Software package for inference of the pattern of demographic divergence. Coalescent simulation. Bayesian model choice.	[58]
ONeSAMP	Web-based program to estimate the effective population size from a sample of microsatellite genotypes. Estimates of effective population size, together with 95% credible limits.	[59]
ABC4F	Software for estimation of F-statistics for dominant data.	[60]
2BAD	Two-event Bayesian Admixture. Software allowing up to two independent admixture events with up to three parental populations. Estimation of several parameters (admixture, effective sizes, etc.). Comparison of pairs of admixture models.	[61]

doi:10.1371/journal.pcbi.1002803.t003

Sunnåker et al. 2013, PLOS Computational Biology

## SOFTWARE

### Methods in Ecology and Evolution



*Methods in Ecology and Evolution* 2013, 4, 684–687

doi: 10.1111/2041-210X.12050

#### APPLICATION

### EasyABC: performing efficient approximate Bayesian computation sampling schemes using R

Franck Jabot\*, Thierry Faure and Nicolas Dumoulin

*Irstea, UR LISC Laboratoire d'ingénierie des systèmes complexes, 24 avenue des Landais – BP 20085, Aubière F-63172, France*

abc – Csilléry et al. 2011, *Methods in Ecology and Evolution*

Bez GUI:

**SimCoal** – simulator + ABC regression – Anderson et al. 2005

**msBayes** – simulator + ABC regression – Hickerson et al. 2007

S GUI:

**ONeSAMP** – ABC rejection – jen jedna Wright-Fisher populace – Tallmon et al. 2004

**popABC** – ABC rejection – Lopes et al. 2009



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

Jean-Marie Cornuet<sup>1</sup>, Virginie Ravigné<sup>2</sup>, Arnaud Estoup<sup>1,3\*</sup> *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<sup>1</sup>, Pierre Pudlo<sup>1,2,3</sup>, Julien Veysier<sup>1,3,4</sup>, Alexandre Delne-Garcia<sup>1,3</sup>, Mathieu Gautier<sup>1,3</sup>, Raphaël Leblois<sup>1,3</sup>, Jean-Michel Marin<sup>2,3</sup>, and Arnaud Estoup<sup>1,3\*</sup>  
<sup>1</sup>Inra, UMR1062 Cbgp, Montpellier, France, <sup>2</sup>Université Montpellier 2, UMR CNRS 5149, I3M, Montpellier, France, <sup>3</sup>Institut de Biologie Computationnelle (IBC), 95 rue de la Galéra, 34095 Montpellier, France, <sup>4</sup>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 population history using the ABC approach

(Cornuet et al. 2008, 2010, 2014)

## DIYABC

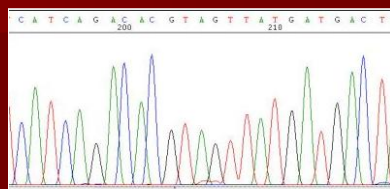


# Genetic data

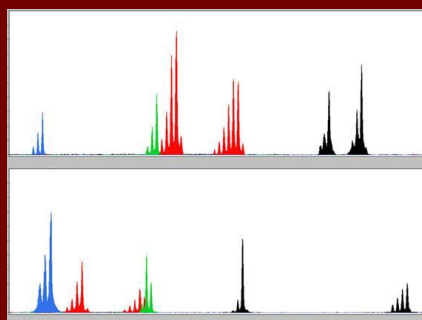


- Sequences

- SNPs



- Genotypes



## ABC works in 3 steps

### 1. SIMULATION STEP:

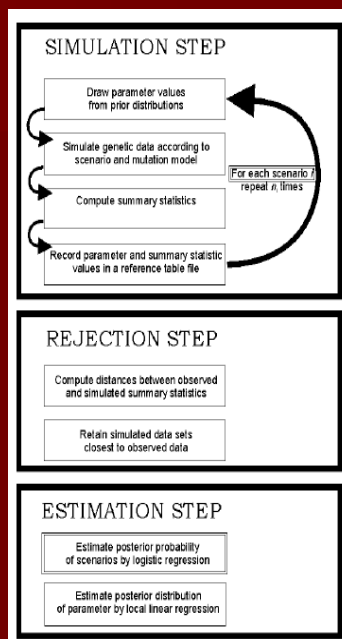
a very large **reference table** is produced and recorded

prior parameter distributions  
scenario  
mutation model

summary statistics (e.g. number of alleles, expected heterozygosity, *f<sub>st</sub>*)

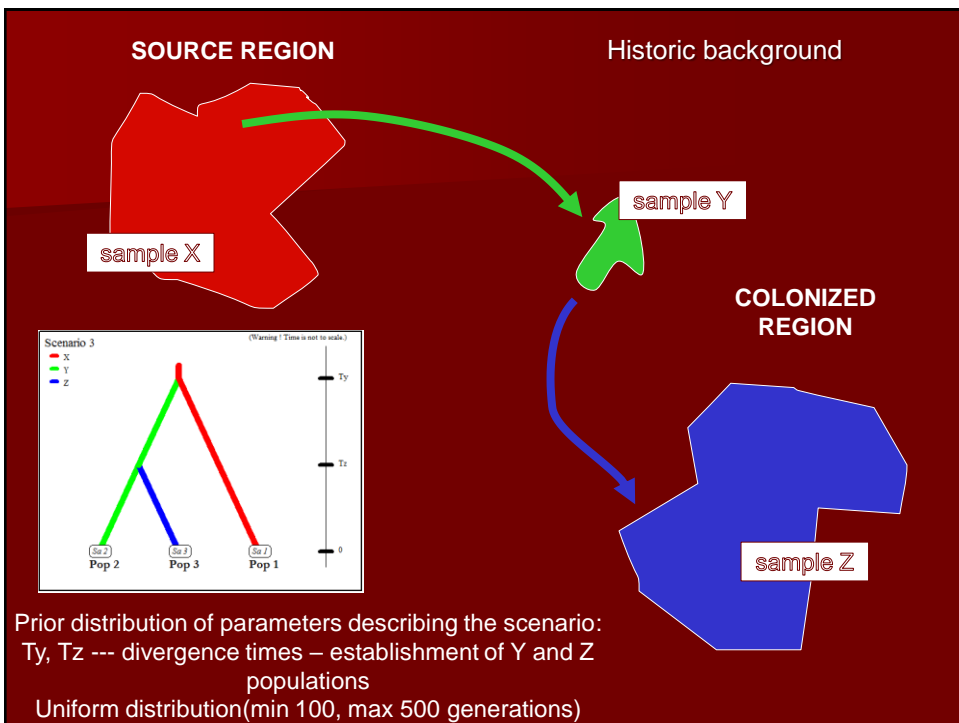
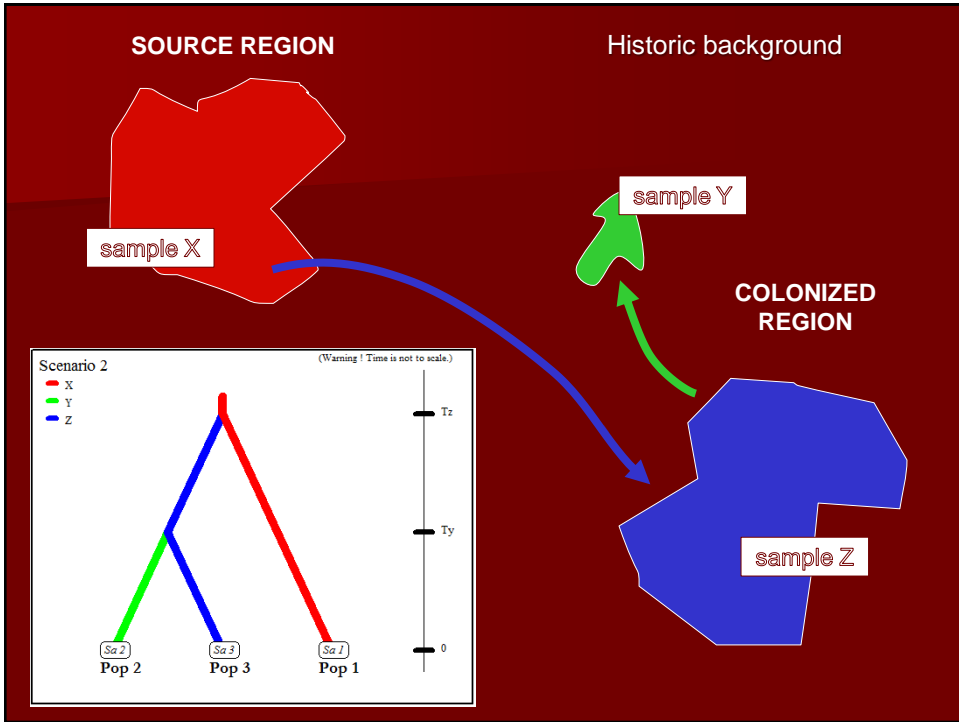
the most time-consuming step

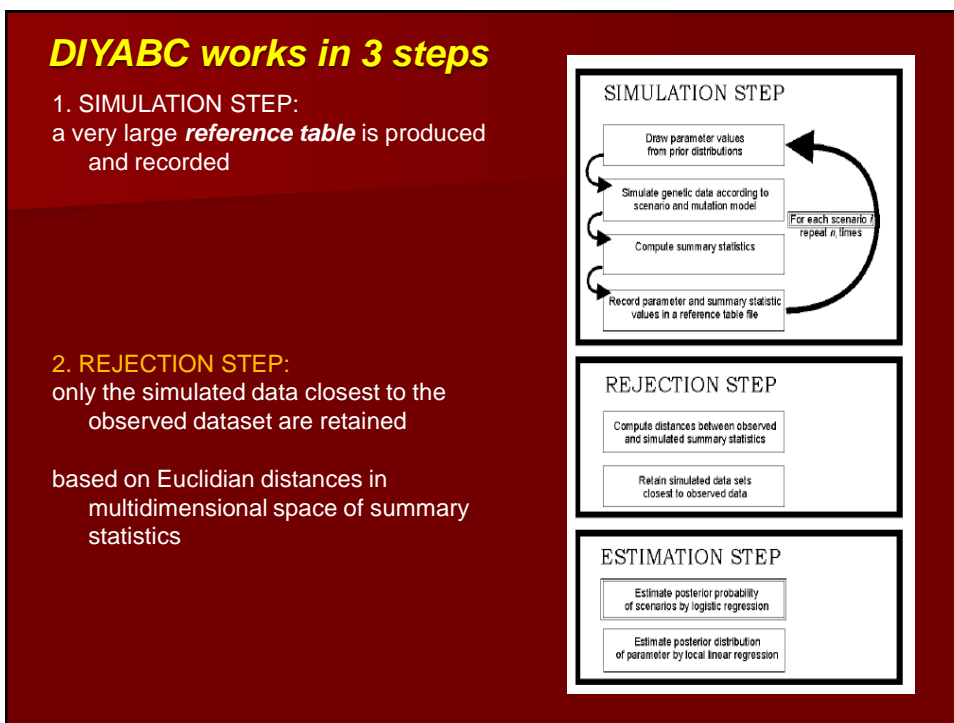
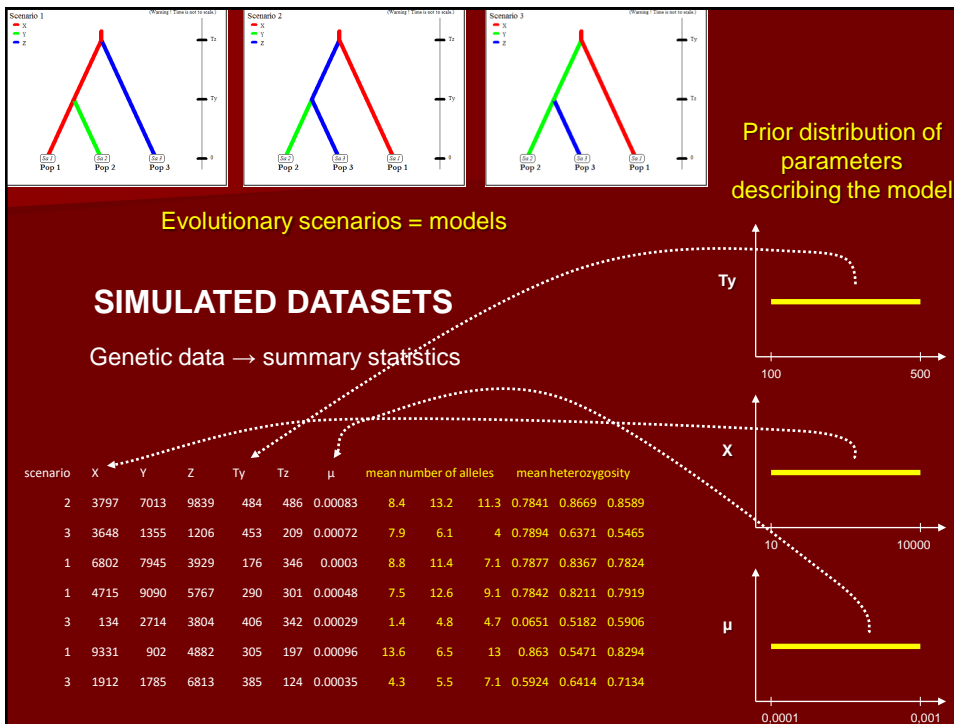
based on the genealogical tree of sampled genes and coalescent theory



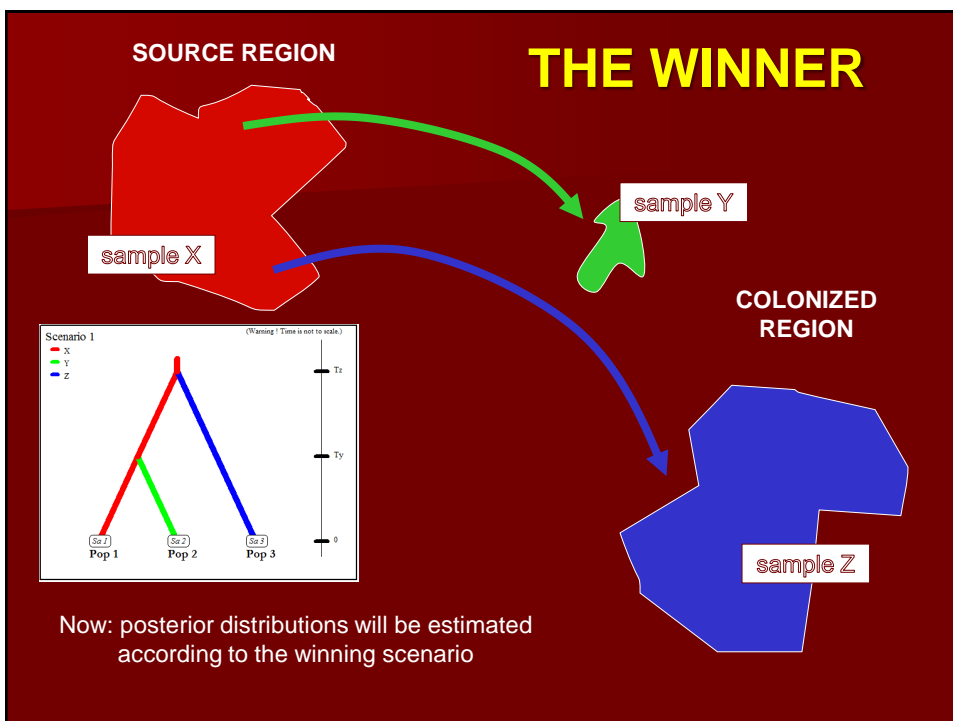
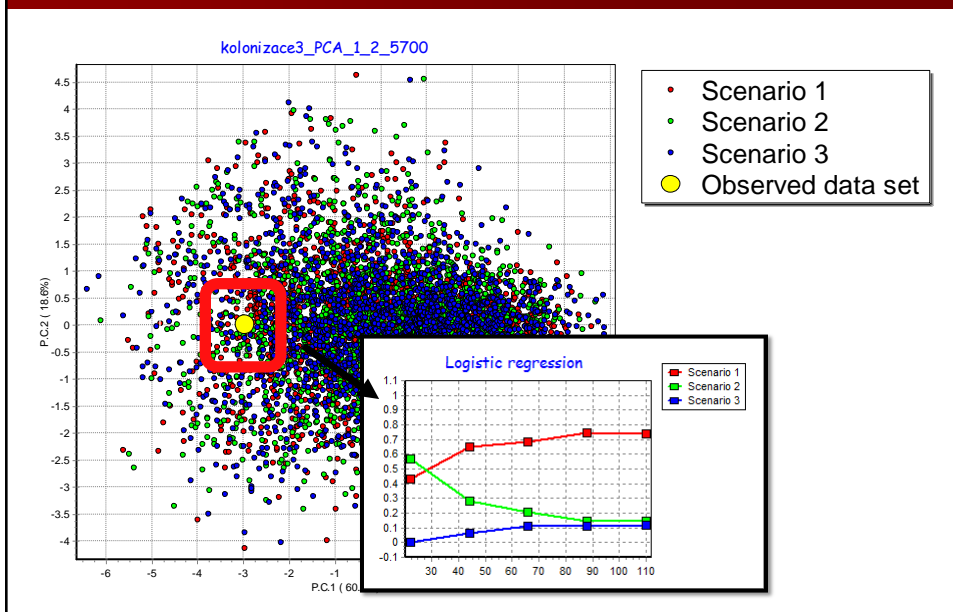
Cornuet et al. 2008, Bioinformatics







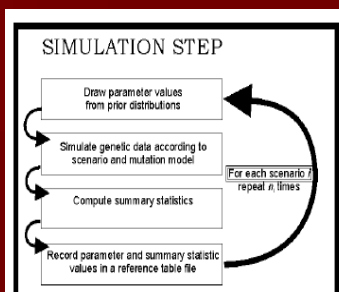
Comparison of our observed dataset with simulated ones and inferring posterior distributions of scenarios



## DIYABC works in 3 steps

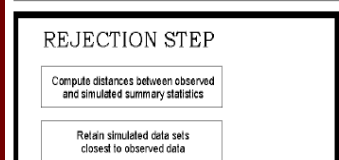
### 1. SIMULATION STEP:

a very large **reference table** is produced and recorded



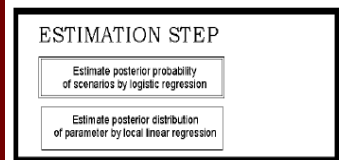
### 2. REJECTION STEP:

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



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

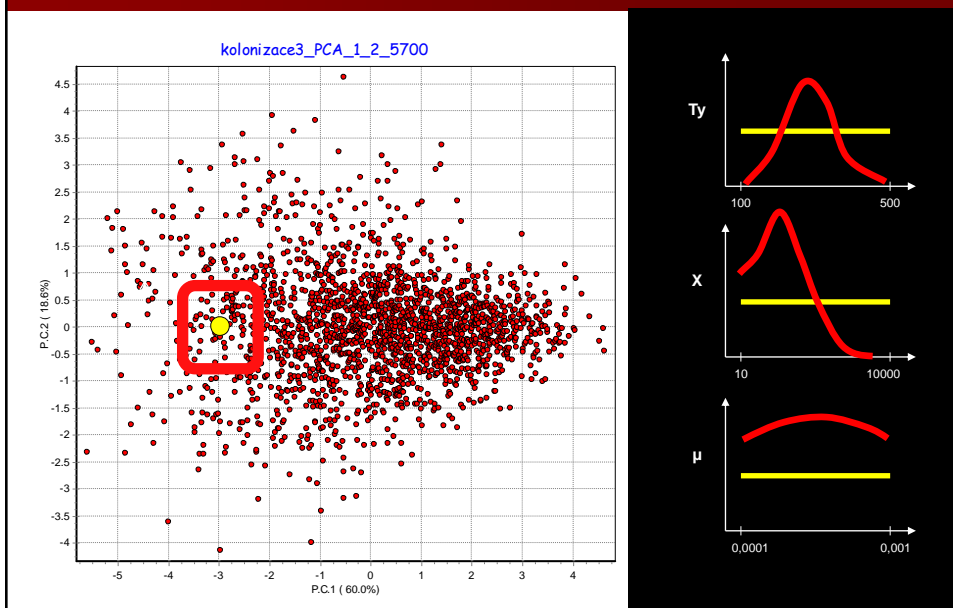




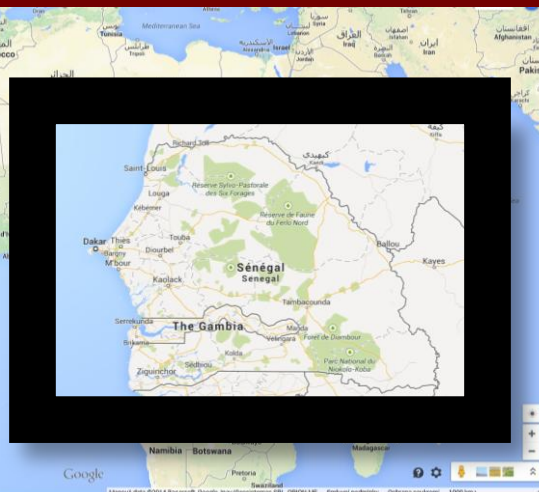
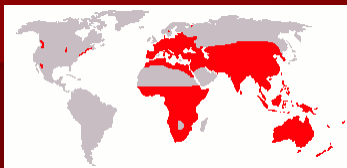


Photo: Jaroslav Červený

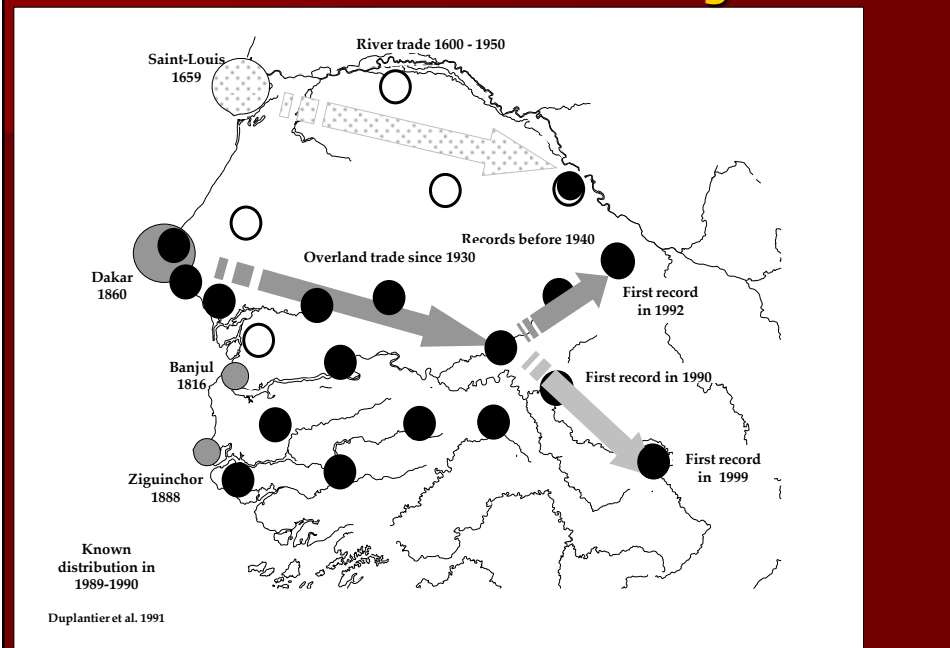
## Black rat (*Rattus rattus*) invasion in Senegal

Konečný et al. 2013,  
Molecular Ecology

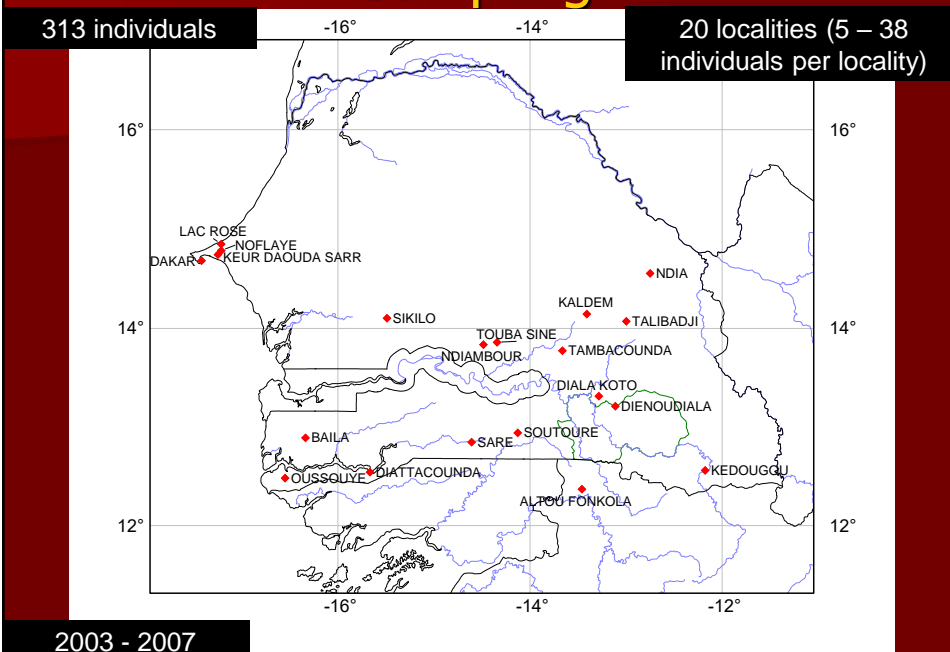
## *Rattus rattus* distribution



# Historic background

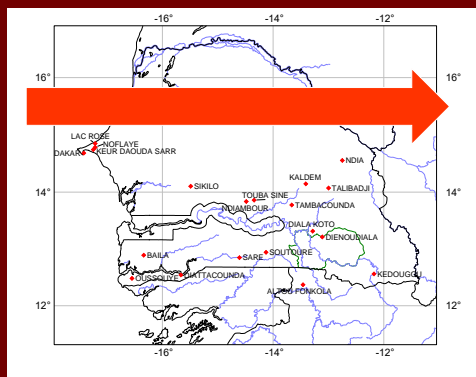


# Sampling

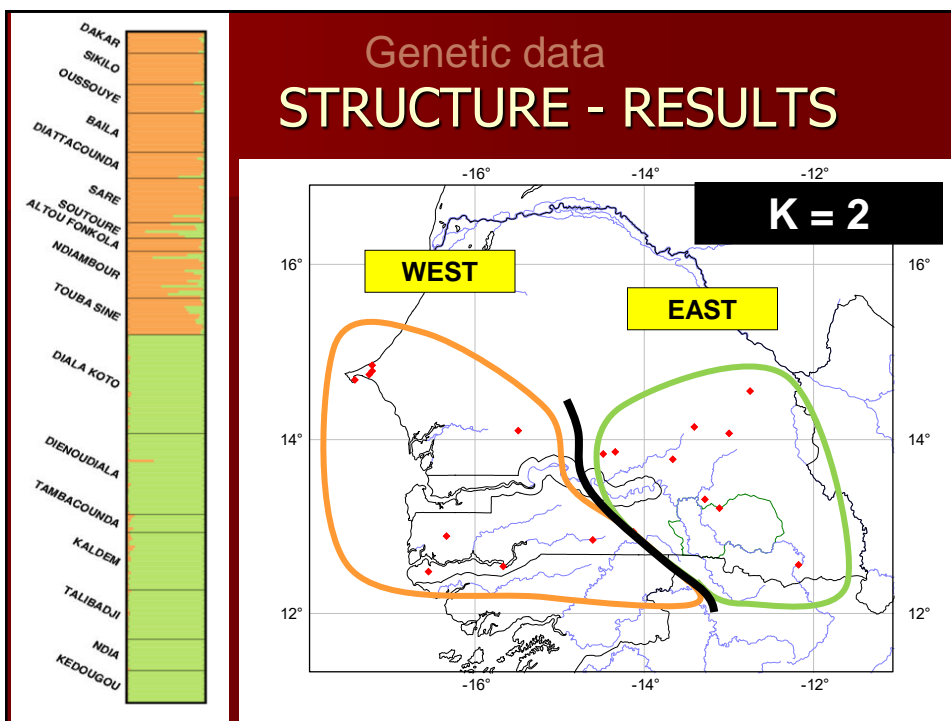


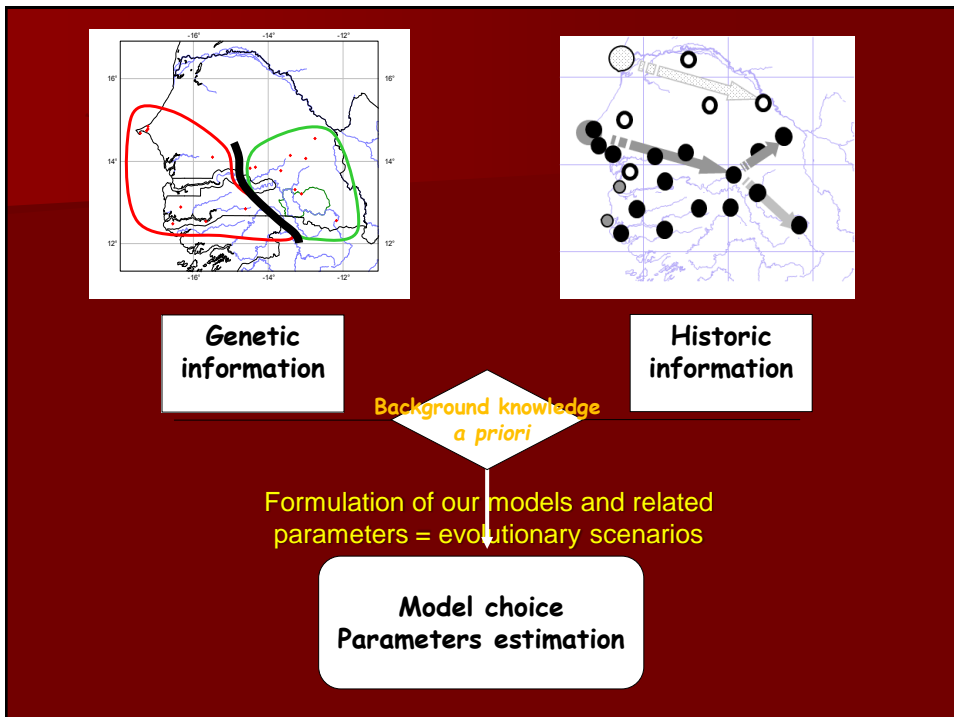
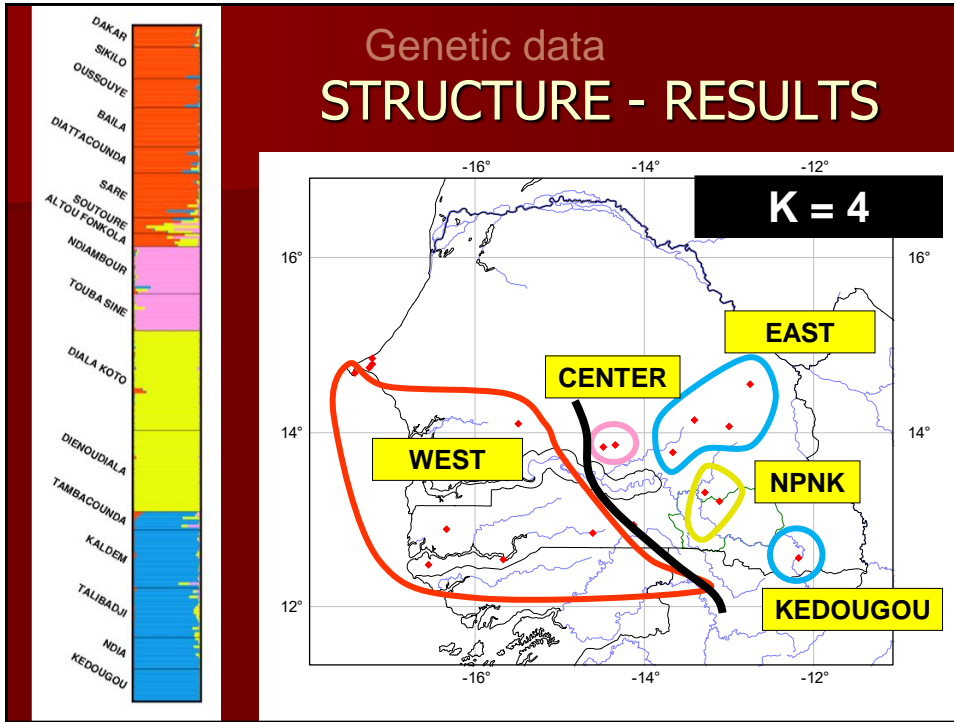
## Genetic data

- 14 microsatellites (9 – 22 alleles, mean: 14.14)
- mean allelic richness – 3.06 (range 1.87 – 4.71)
- mean expected heterozygosity – 0.538 (range 0.323 – 0.762)

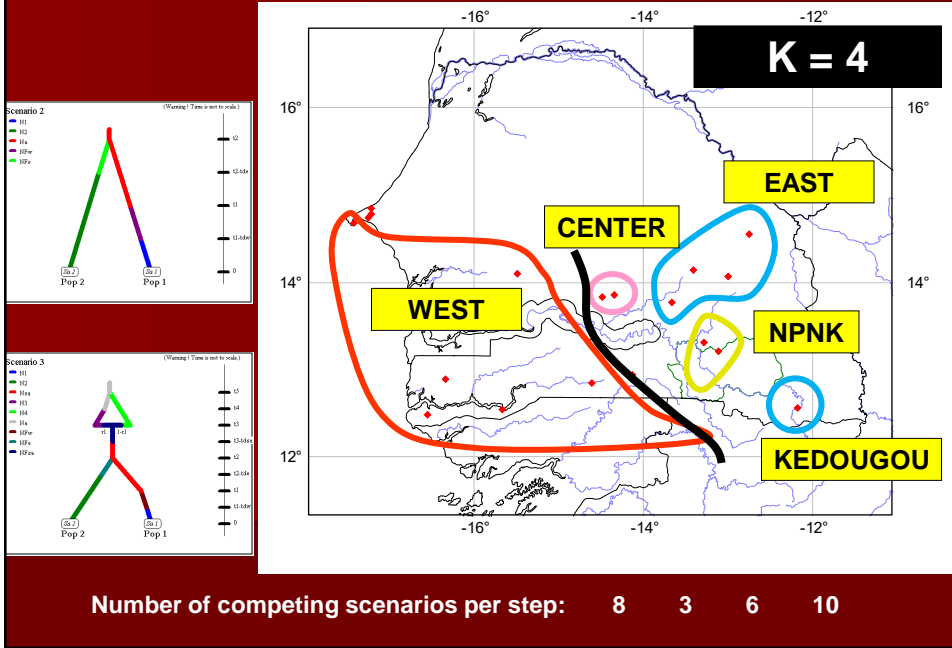


both allelic richness and heterozygosity **decreased** with longitude

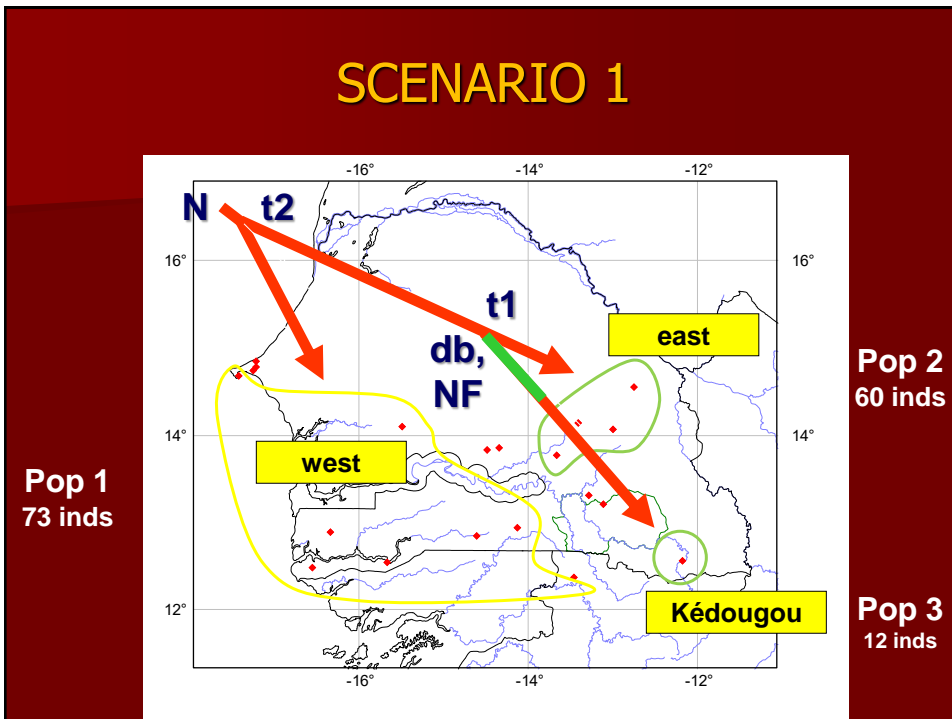




# ABC analysis in four steps – four questions

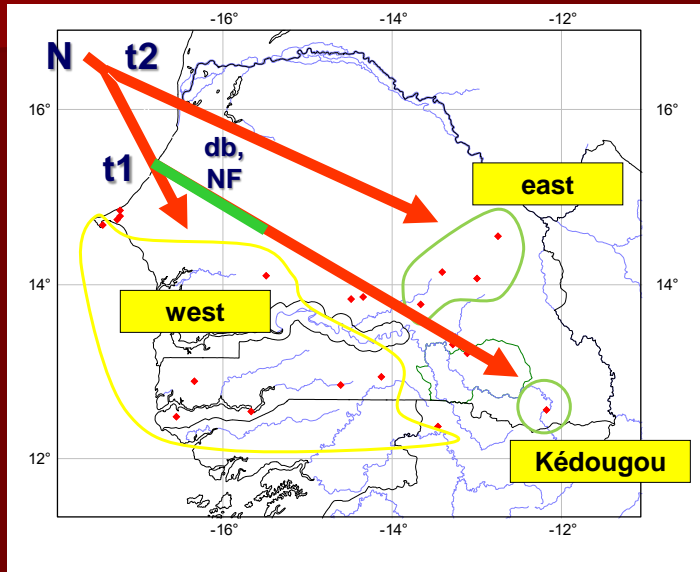


# SCENARIO 1



## SCENARIO 2

Pop 1  
73 inds

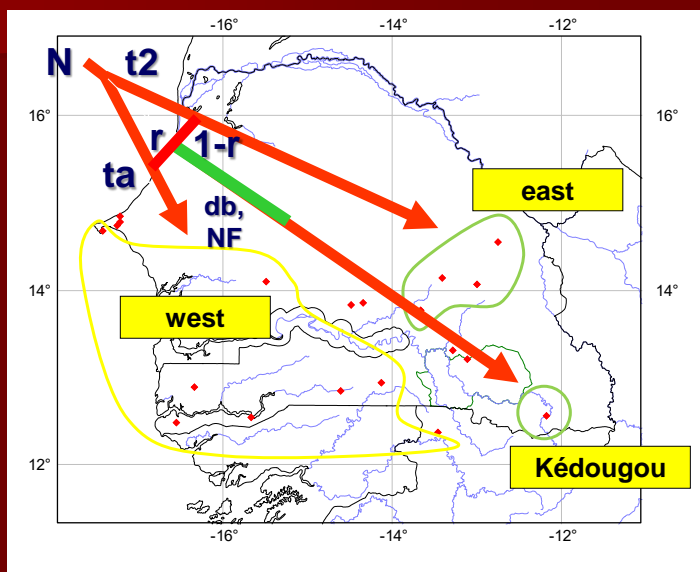


Pop 2  
60 inds

Pop 3  
12 inds

## SCENARIO 3

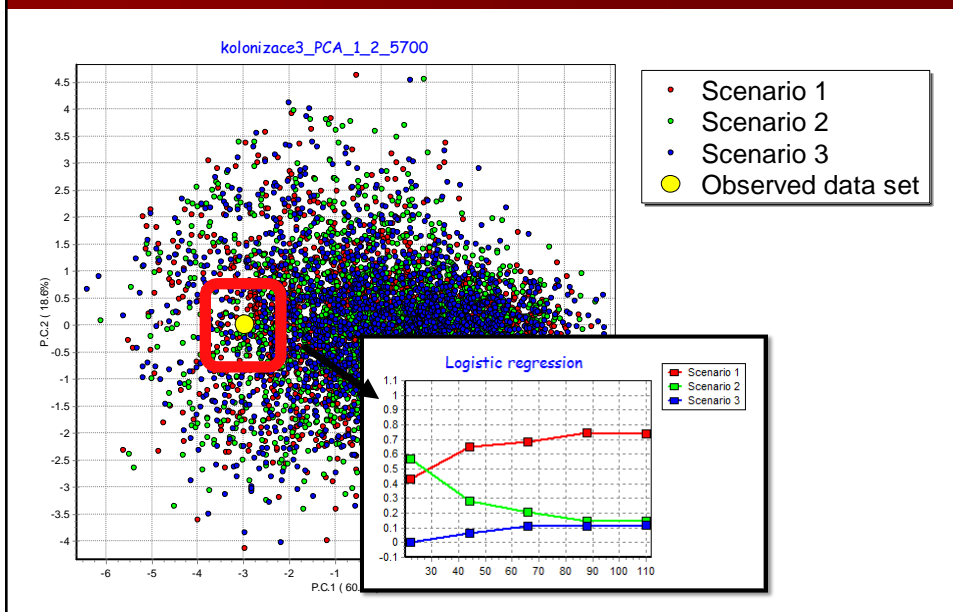
Pop 1  
73 inds



Pop 2  
60 inds

Pop 3  
12 inds

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



THE WINNER IS... SCENARIO 1

