ABC (Approximate Bayesian Computation) methods to make inference about population history from molecular data: principles and applications

Arnaud Estoup

Centre de Biologie pour la Gestion des Populations (CBGP, INRA)

estoup@supagro.inra.fr

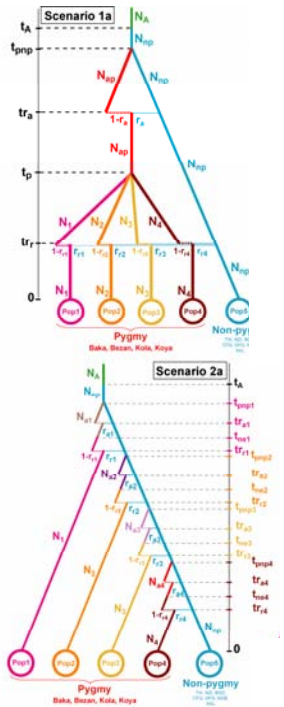
Mathematical and Computational Evolutionary Biology
June 18-22, 2012
Hameau de l'Etoile, Saint Martin de Londres, France

Part 1. - General context -
To make inference about population processes / histories using molecular data : what does it means ?

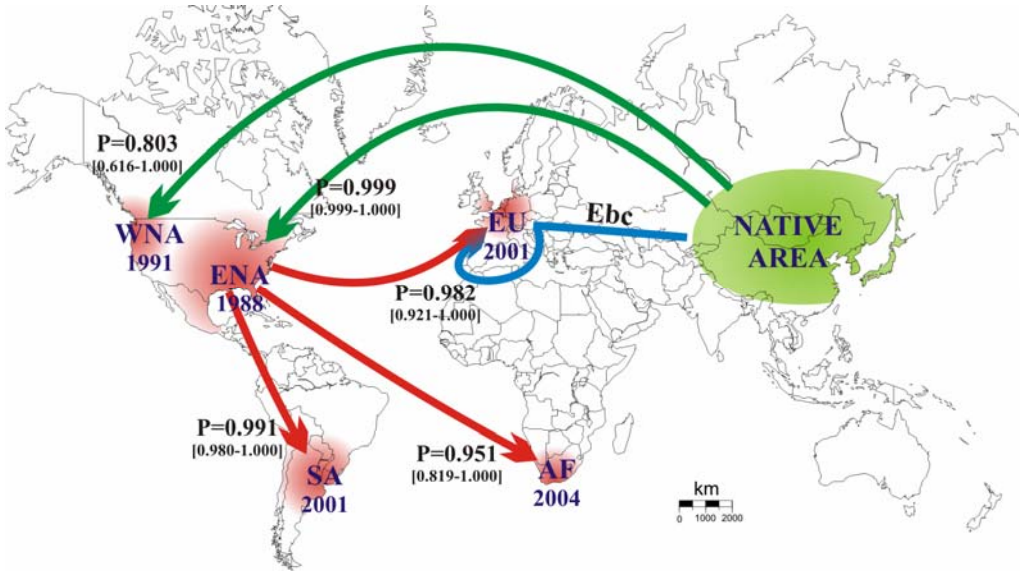
Natural populations have complex demographic histories: their sizes and ranges change over time, leading to fission and fusion processes that leave signatures on their genetic composition. One 'promise' of biology is that molecular data will help us uncover the complex demographic and adaptive processes that have acted on natural populations

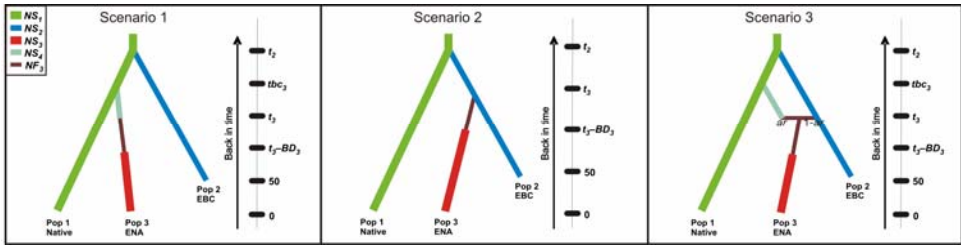
(Czilléry et al. 2010, TREE)

(Non)Independent evolutionary origin of populations adapted to new environments



Routes of introduction of invasive species

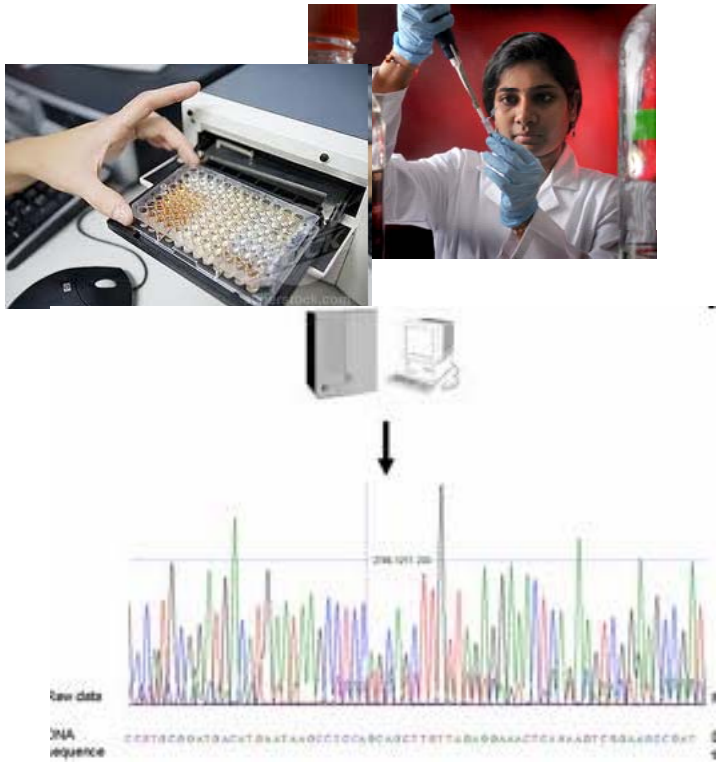




Collect samples
in Natura

Produce molecular data in the lab

(microsatellites, DNA sequences,…)



```
Title line:DIYABC example of microsatellite and mtDNA data file
locus MS6 <A>
locus MS7 <A>
locus COI <M>
POP
BU ,         165200      312312      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
BU ,         200200      312312      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTA
BU ,         149200      306312      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
BU ,         200200      312312      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
BU ,         200200      312312      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTCTGGTAGGAACTGGGATAAGTGTCTTAATTC
BU ,         200200      312312      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
POP
MA ,         200205      312315      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
MA ,         200218      312315      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
MA ,         200205      312312      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
MA ,         200205      312312      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
MA ,         149200      312312      <[CATAAAGATATTGGGACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
MA ,         200200      312315      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
MA ,         149200      312312      <[CATAAAGATATTGGAACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
MA ,         149200      312315      <[CATAAAGATATTGGGACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
MA ,         200200      312315      <[CATAAAGATATTGGGACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
POP
BF ,         200200      312312      <[CATAAAGATATTGGGACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
BF ,         200200      306315      <[CATAAAGATATTGGGACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
BF ,         149200      312312      <[]>
BF ,         200200      312312      <[CATAAAGATATTGGGACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
BF ,         200200      312315      <[CATAAAGATATTGGGACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
BF ,         200200      312312      <[CATAAAGATATTGGGACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
BF ,         200205      312312      <[CATAAAGATATTGGGACTCTATATTTTATTTTTAGAATTTGAGCGGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
BF ,         200205      312312      <[CATAAAGATATTGGGACTCTATATTTTATTTTTAGAATTTGAGCAGGTTTGGTAGGAACTGGGATAAGTGTCTTAATTC
```

Part 2. Methods to make inferences under complex models (scenarios)

→ Approximate Bayesian Computation methods: principles

- Inference: compute likelihood (molecular_data / parameters / model) (e.g. IS if gene tree) and explore parameter space (e.g. MCMC)

- Complex models (scenarios): difficult to compute likelihood + difficult to explore parameter space

Approximate Bayesian Computation (ABC) circumvents these problems → bypass exact likelihood calculations by using (massive) simulations of data sets + summary statistics

Summary statistics are values calculated from the observed and simulated data to represent the maximum amount of information in the simplest possible form.

e.g. nbre alleles, nbre of haplotypes, Heterozygosity, allele size variance, Fst, genetic distance,…

Historical background in population genetics – Key papers

Tavaré S, Balding DJ, Griffiths RC, Donnely P (1997) Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505-518.

Pritchard JK, Seielstad MT, Prez-Lezaum A, Feldman MW (1999) Population growth of human Y chromosomes, a study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**, 1791-1798.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-2035.

"Standard/conventional ABC" → Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-2035.

- Several studies have shown that parameter posterior distributions inferred by ABC are similar to those provided by full-likelihood approaches (see e.g. Beaumont *et al.* 2009; Bortot *et al.* 2007; Leuenberger & Wegmann 2009; Marjoram *et al.* 2003),

- The ABC approach is still in its infancy and continues to evolve, and to be improved.

More and more sophisticated developments on ABC

➢ *Better exploration of the parameter space*

❑ ABC-MCMC (Majoram et al. 2003; Wegmann et al. 2009; Leuenberger & Wegmann 2010)

❑ ABC-PMC (peculiar Monte Carlo schemes→ ABC with Population Monte Carlo; Beaumont et al. 2009)

➢ *Choice of statistics to summarise datasets* (Joyce & Marjoram 2008; Sousa *et al.* 2009; Nunez & Balding 2010) *and how they can be combined* → e.g. extraction of a limited number of orthogonal components (Wegmann *et al.* 2009), non-linear feed-forward neural network (Blum & François, 2009), linear discriminant analysis on statistics for model choice (Estoup *et al.* 2012)

➢ *More efficient conditional density estimation* (Blum & Francois 2009; Wegmann et al. 2010).

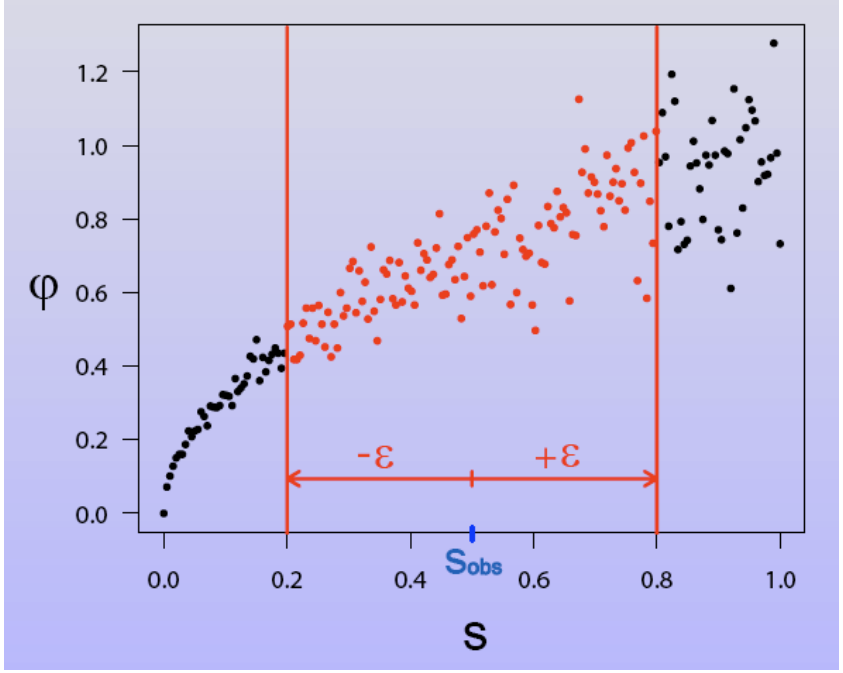
→ Difficult to implement + in practice modest avantages over "Standard ABC" at least for a sufficiently large number of simulations

-GENERAL PRINCIPLE OF ABC –

ABC is intuitively very easy: millions of simulated data sets are produced assuming different parameter values (drawn into priors) and under different models →the simulations that produce genetic variation patterns (simulated summary statistics) close to the observed data (observed summary statistics) are retained and analysed in detail

ABC a la Pritchard et al. (1999) – basic rejection algorithm

➢ CRITERION: If for ALL summary statistics |SS – SO|/SO < ε then record parameter values drawn from prior distributions



ABC a la Beaumont et al. (2002)

- Basic rejection algorithm: acceptation rate decrease (quickly) with nbre of SS and parameters → algorithm inefficient for complex models

- New acceptation / distance criterion to increase the tolerance threshold → multi-statistics Euclidean distance (statistics normalized)

$$\Lambda = \sqrt{\sum_i \left[(SS_i - SO_i) \Big/ ET(SS_i) \right]^2}$$

- Correction step based on regression methods

COESTIMATION OF POSTERIOR DISTRIBUTIONS OF PARAMETERS UNDER A GIVEN SCENARIO (MODEL)

Step 1: Simulate genetic data sets under a scenario (parameter values drawn into priors) = 10^6 itérations (i.e. 10^6 data sets)

→ 10^6 [values of parameters x summary stat x Euclidian distance]

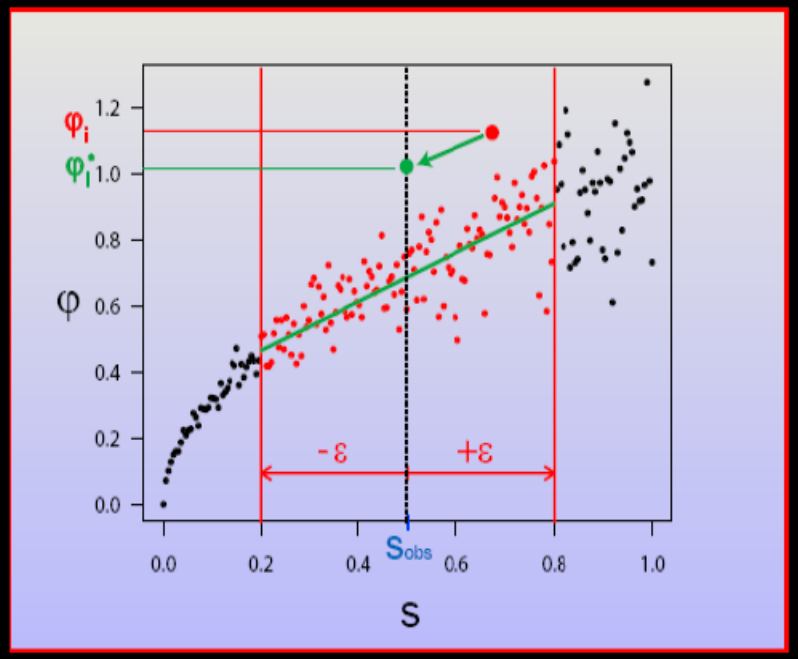
COALESCENCE

Step 2: Sort data sets according to their Euclidian distances

Step 3: Local LINEAR regression (weighted) on the 1% (10000) “best simulations” = those with the smallest Euclidian distances

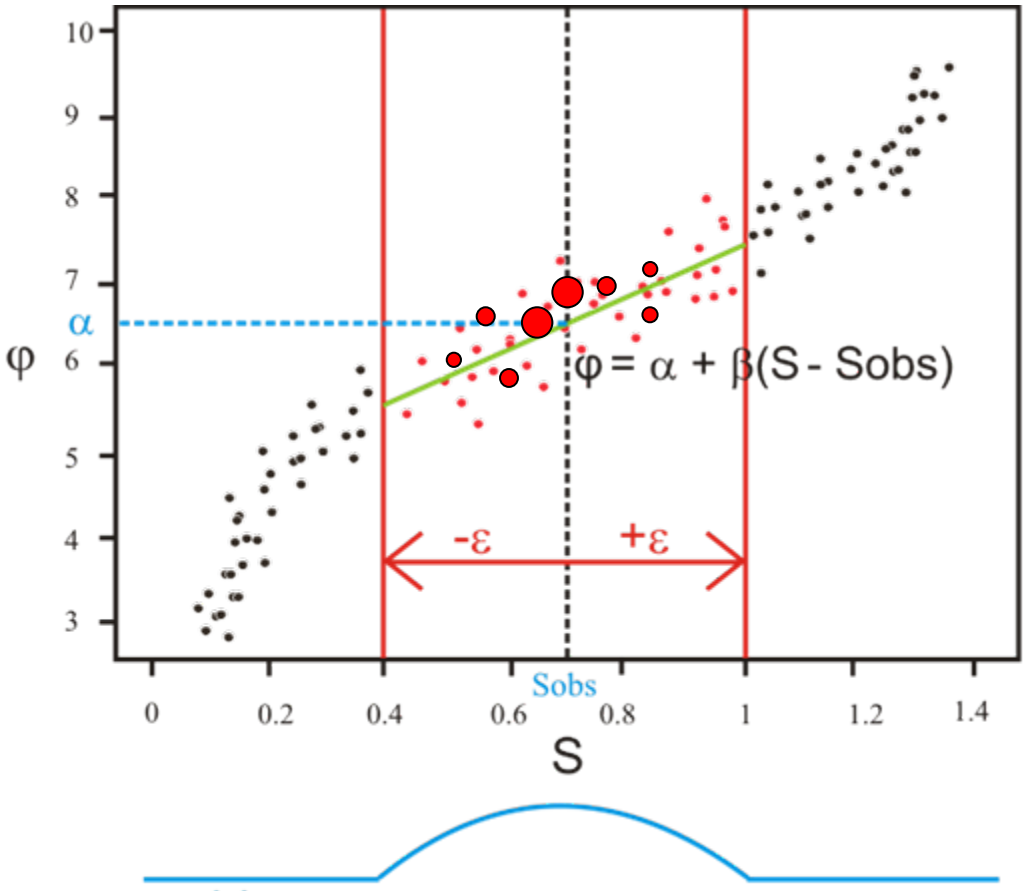
→ correction of accepted parameter values

Beaumont et al (2002) increase the tolerance ε , and correct the posterior distribution by performing linear regression



φ = model parameter, s = summary statistic (here in one dimension)

WEIGHTED local LINEAR regression



Weight

(kernel using Euclidean distance values)

COMPARAISON OF DIFFERENT SCENARIOS (MODELS): estimation of relative probabilities for each scenario *s*

Step1: simulation of genetic data sets (draw from priors)

scenario 1 = 10^6 data sets

scenario 2 = 10^6 data sets

…

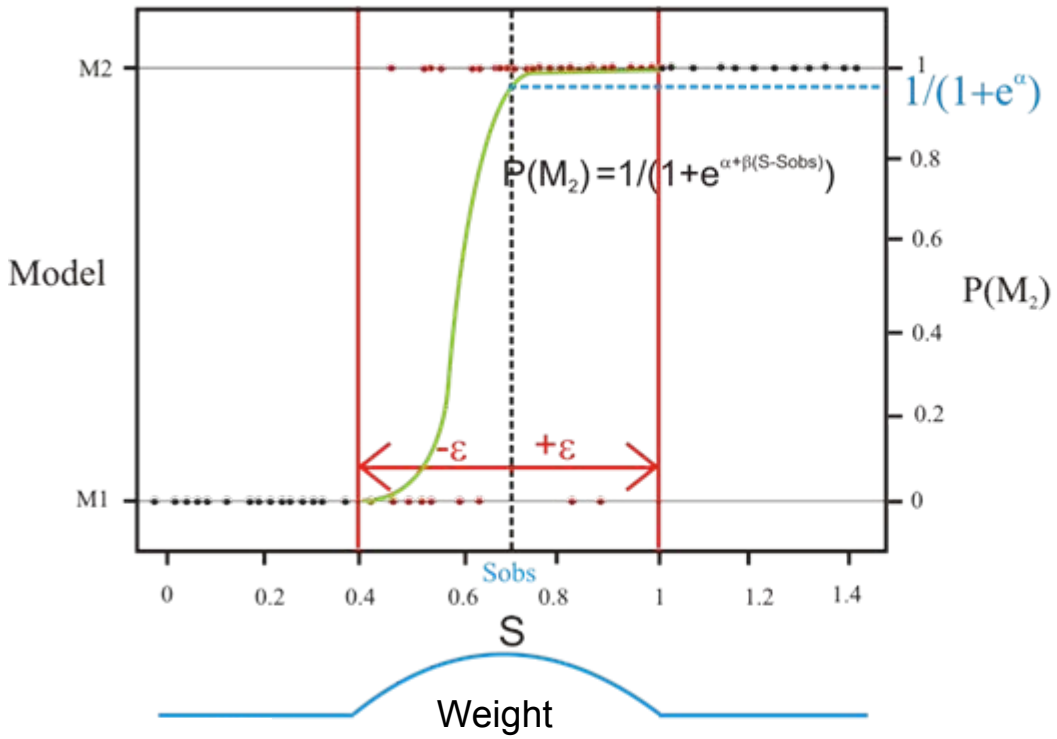
scenario 8 = 10^6 data sets

COALESCENCE

→ $8x10^6$ [scenario identifier x Euclidian distances]

Step 2: sort data sets according to their Euclidian distances

Step 3: The scenario number is considered as a discrete parameter → do Local LOGISTIC regression on the 1% (80000) "best simulations" (i.e. with smallest Euclidian distances) to estimate probabilities of each scenario



$1/(1+e^{\alpha})$ is an estimator of P(M2)

Local LOGISTIC regression can be used
for the estimation of model probability (here, P(M2))

(Beaumont 2008; Fagundes et al. 2007; Cornuet et al. 2008)

ABC = 3 main APPROXIMATIONS

1/ Data = summary statistics (instead of « raw data »)

2/ Selected simulated data sets = in the vicinity of the target = threshold ε

3/ Monte carlo approximation = posterior distributions or probabilities estimated from a sample of values generated by a process with substantial stockasticity

→ If the number of simulations tends to infinity then the approximations 2/ and 3/ do not hold anymore

Part 3. ABC in practice: Not as easy/simple as it seems to be ?

1/ SIMULATE MANY DATA SETS (model(s), priors, summary statistics)

2/ KEEP THE « BEST » SIMULATIONS

3/ MAKE « CORRECTIONS » (regression methods) ON RETAINED SIMULATIONS

Recent reviews on ABC…good for application of ABC

Review

Approximate Bayesian Computation (ABC) in practice

Katalin Csilléry[1], Michael G.B. Blum[1], Oscar E. Gaggiotti[2] and Olivier François[1]

[1] Laboratoire Techniques de l'Ingénierie Médicale et de la Complexité, Centre National de la Recherche Scientifique UMR5525, Université Joseph Fourier, 38706 La Tronche, France
[2] Laboratoire d'Ecologie Alpine, Centre National de la Recherche Scientifique UMR5553, Université Joseph Fourier, 38041 Grenoble, France

MOLECULAR ECOLOGY

Molecular Ecology (2010) 19, 2609–2625 doi: 10.1111/j.1365-294X.2010.04690.x

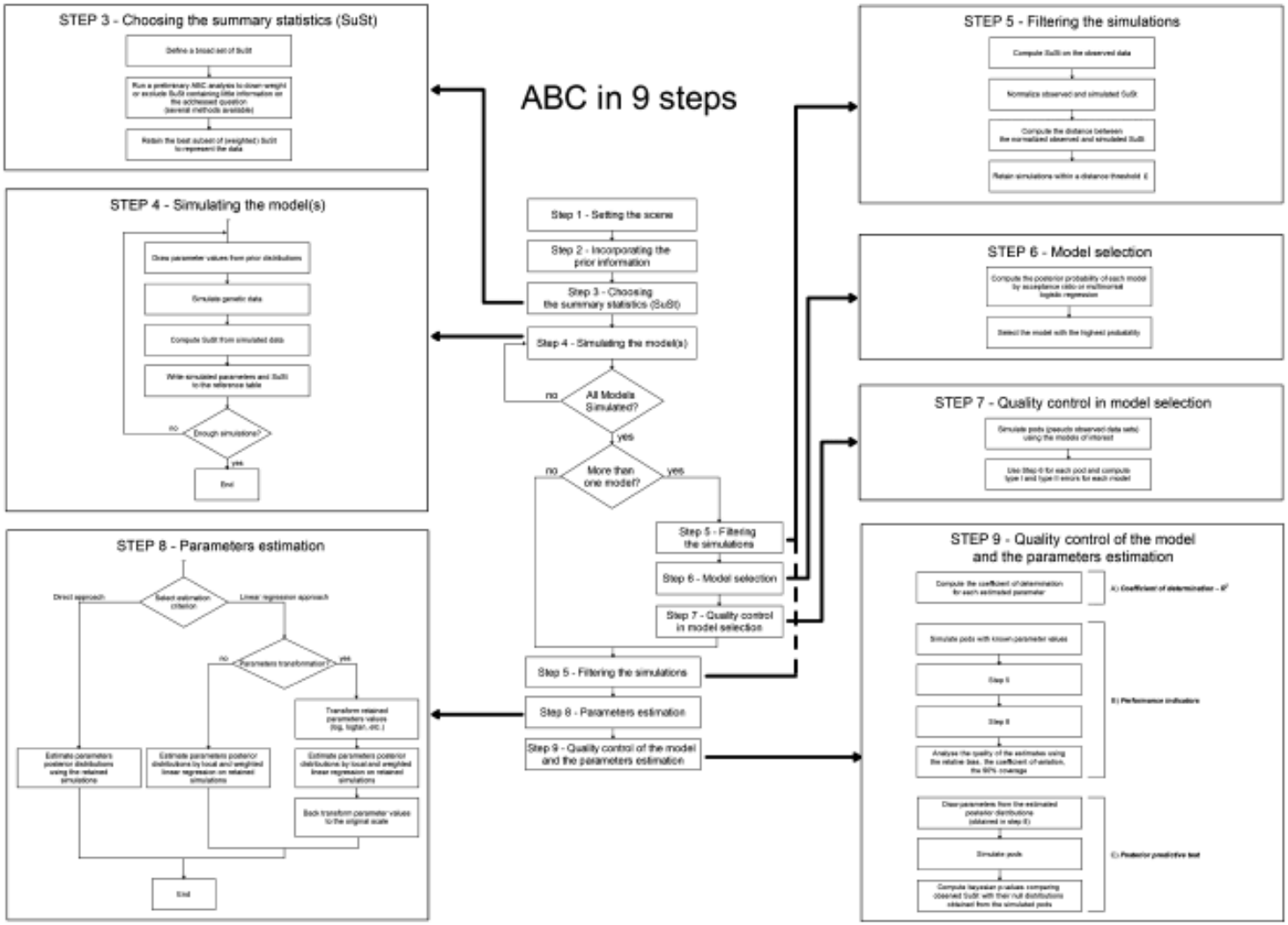
INVITED REVIEW

ABC as a flexible framework to estimate demography over space and time: some cons, many pros

G. BERTORELLE,* A. BENAZZO* and S. MONA*†‡

*Department of Biology and Evolution, University of Ferrara, Via Borsari 46, 44100 Ferrara, Italy, †CMPG, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland, ‡Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Fig. 1 ABC in nine steps.



Bertorelle et al. Mol Ecol 2010

An example of (user-friendly) integrated software for infering population history with ABC: *DIYABC*

Cornuet J-M, Santos F, Robert PC, Marin J-M, Balding DJ, Guillemaud T, Estoup A (2008) Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation. *Bioinformatics*, 24, 2713-2719.

Cornuet JM, Ravigné V, Estoup A (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC bioinformatics*, 11, 401doi:10.1186/1471-2105-11-401.

The software DIYABC (V.0 and V1.) and the companion papers are both freely available at http://www1.montpellier.inra.fr/CBGP/diyabc.

➢In 2008 we wrote (Cornuet et al. 2008) " In its current state, the ABC approach remains inaccessible to most biologists because there is not yet a simple software solution"……less and less true

Table 1 Features of the main online backward coalescent simulators available for ABC analysis. Other software are developed for specific purposes and available upon request to the authors (see e.g. Przeworski 2003)

Name	Type of markers	Demographic model				Recombination	Selection	Serial sampling[1]	Consider explicitly spatial and environmental heterogeneity	ABC integrated	Reference
		One/many populations	Population divergence	Migration	Change in population size						
MS	DNA sequence	Many	Yes	Yes	Yes	Yes	No[2]	No	No	No	Hudson (2002)
Simcoal2	RFLP, STR, DNA sequence, SNP	Many	Yes	Yes	Yes	Yes	No	No	No	No	Laval & Excoffier (2004)
Selsim	STR, DNA sequence	One	No	No	No	Yes	Yes	No	No	No	Spencer & Coop (2004)
SPLATCHE	RFLP, STR, DNA sequence[3]	Many	Yes	Yes	Yes	No[3]	No	No	Yes	No	Currat *et al.* (2004)
Bayesian Serial Simcoal	RFLP, STR, DNA sequence	Many	Yes	Yes	Yes	No	No	Yes	No	No[4]	Anderson *et al.* (2005)
AQUASPLATCHE	RFLP, STR, DNA sequence, SNP	Many	Yes	Yes	Yes	No	No	No	Yes	No	Neuenschwander (2006)
msBayes	DNA sequence	Many[5]	Yes	Yes	Yes	Yes	No	No	No	Yes	Hickerson *et al.* (2007)
DIY ABC	STR, DNA sequence	Many	Yes	No	Yes	No	No	Yes	No	Yes	Cornuet *et al.* (2008)[6]
ONeSAMP	STR	One	No	No	No	No	No	No	No	Yes	Tallmon *et al.* (2008)
PopABC	STR, DNA sequence	Many	Yes	Yes	No	Yes	No	No	No	Yes	Lopes *et al.* (2009)
ABCtoolbox[7]	RFLP, STR, DNA sequence, SNP	Many	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes/No[7]	Wegmann *et al.* (2010)

[1]Samples with different ages can be simulated (relevant if ancient DNA data or time series are available).
[2]MS has been modified by Jensen *et al.* (2008) to include selection.
[3]The new version of SPLATCHE, including recombination and SNP markers, will be available soon (L. Excoffier, pers. comm.).
[4]Data can be by simulated sampling parameter values from prior distributions.
[5]Only a vicariance model can be simulated.
[6]The new version of DIYABC is available online.
[7]ABCtoolbox is a collection of independent command-line programs which facilitate the development of a pipeline to estimate model parameters and compare models; several external simulation programs can be pipelined.

Bertorelle et al. Mol Ecol 2010

DIYABC: Inferences on complex scenarios

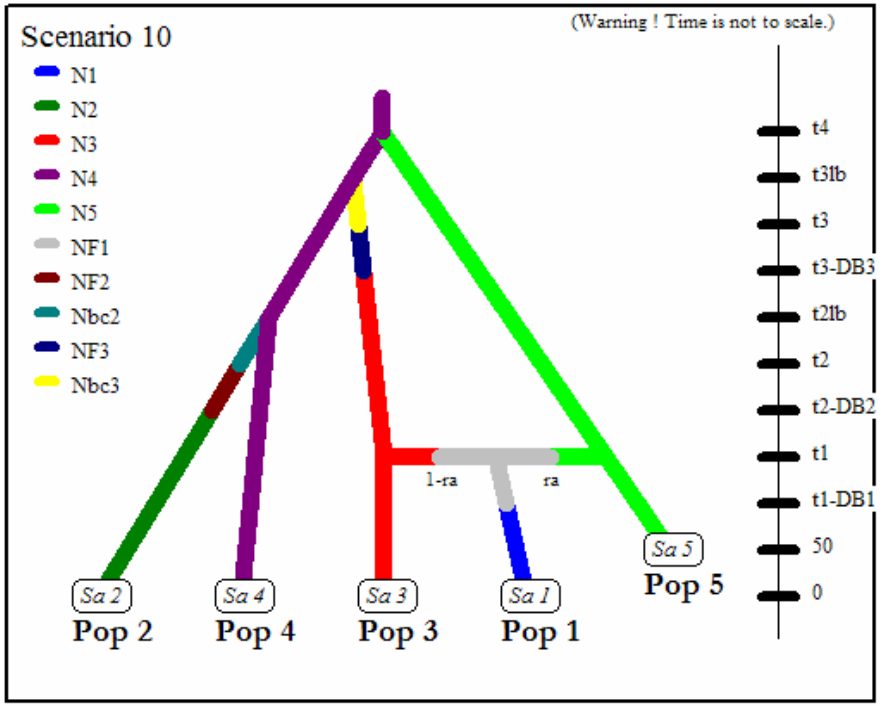
- Historical events = population divergence, admixture, effective size fluctuation
- Large sample sizes (populations, individuals, loci)
- Diploid or haploid individuals
- Different sampling times
- Microsatellite and/or sequence data, no gene flow between populations

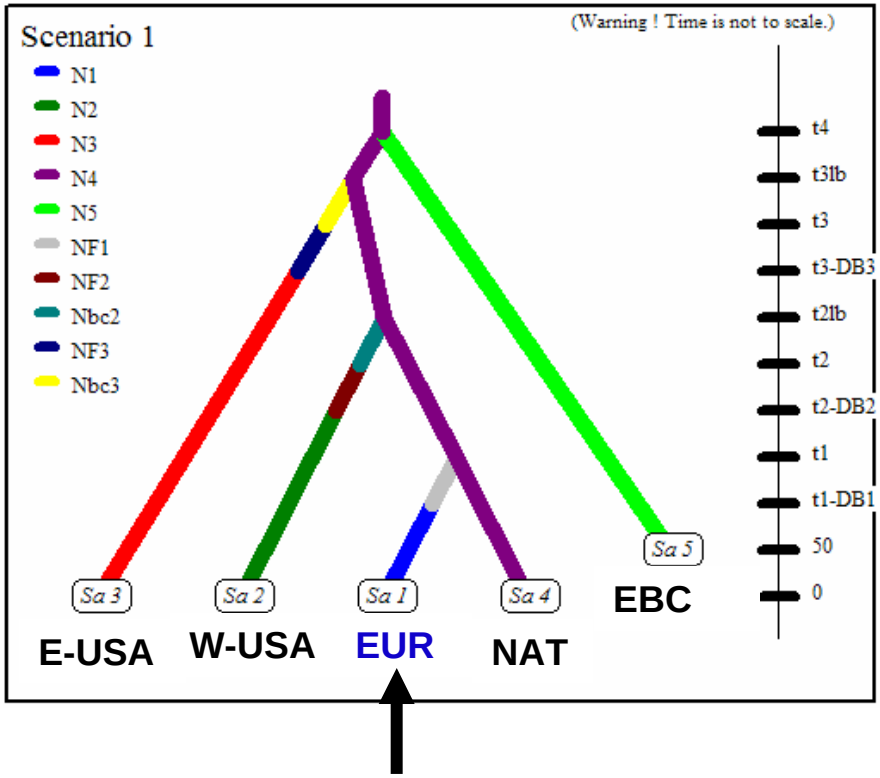
Program:

- writen in Delphi
- running under a 32-bit Windows operating system (XP)
- multi-processor
- user-friendly graphical interface

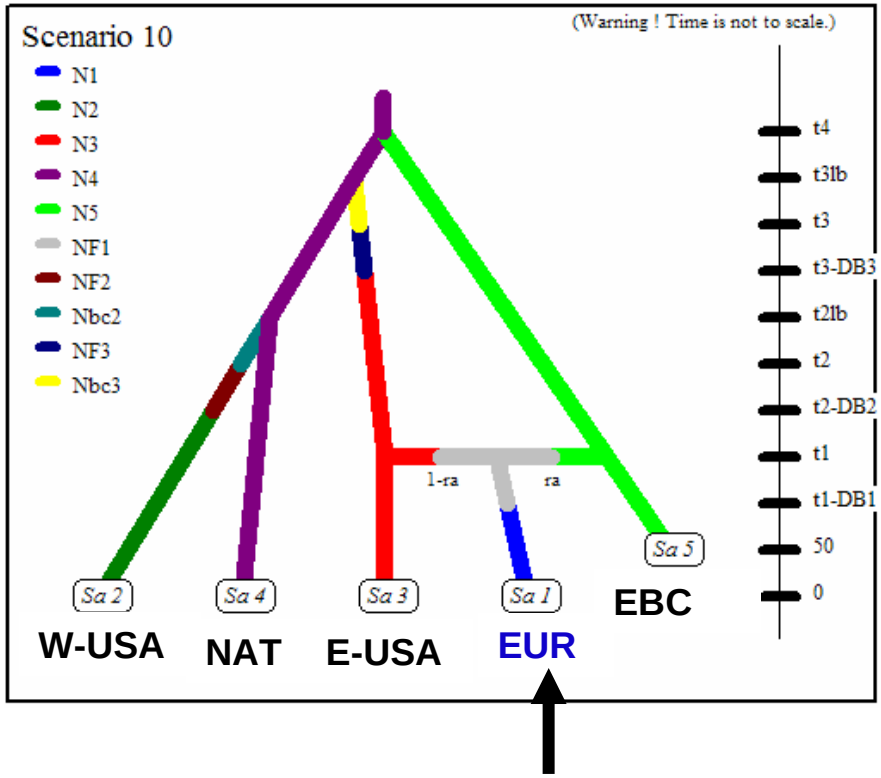
Examples of (complex) scenarios that can be formalised and treated with DIYABC: looking for the source of an invasive population → case of the ladybird *Harmonia axyridis*

N1 N2 N3 N4 N5
0 sample 1
0 sample 2
0 sample 3
0 sample 4
50 sample 5
t1-DB1 VarNe 1 NF1
t1 split 1 5 3 ra
t2-DB2 VarNe 2 NF2
t2 VarNe 2 Nbc2
t2lb merge 4 2
t3-DB3 VarNe 3 NF3
t3 VarNe 3 Nbc3
t3lb merge 4 3
t4 merge 4 5





Origin: native area



Origin: admixture E-USA + EBC

es\DIYABC\ben_plos_prior1_t48_DB30_NF300_size1_range50.reftable

The reference table, ben_plos_prior1_t48_DB30_NF300_size1_range50.reftable, contains 300 simulated data se
Each record includes 8 parameters and 14 summary statistics
The reference table has been built with 4 scenarios

Do you want to

1 (●) Append new simulations to the reference table

5 6 ( ) Estimate posterior distributions of parameters and model checking

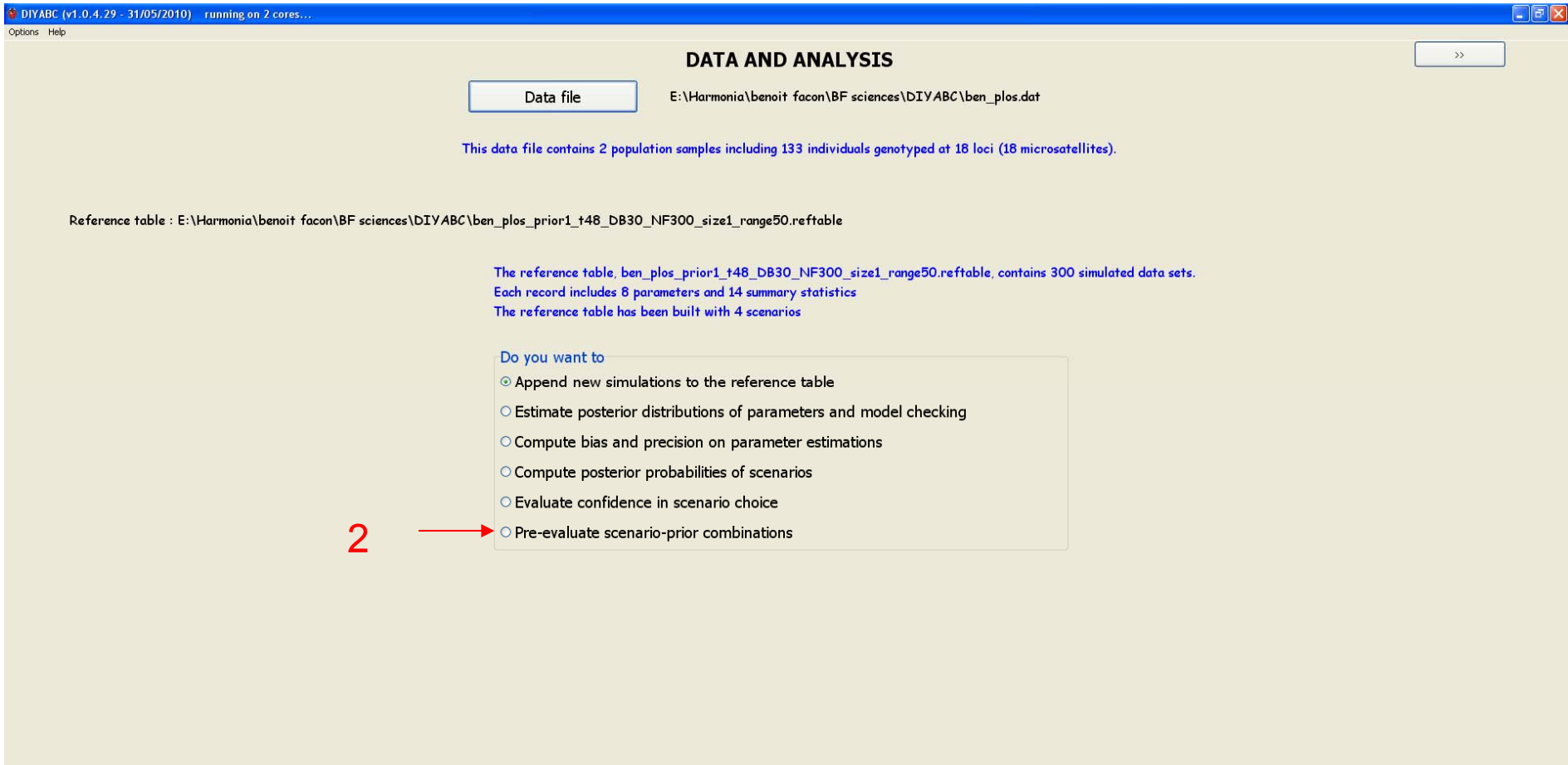
7 ( ) Compute bias and precision on parameter estimations

3 ( ) Compute posterior probabilities of scenarios

4 ( ) Evaluate confidence in scenario choice

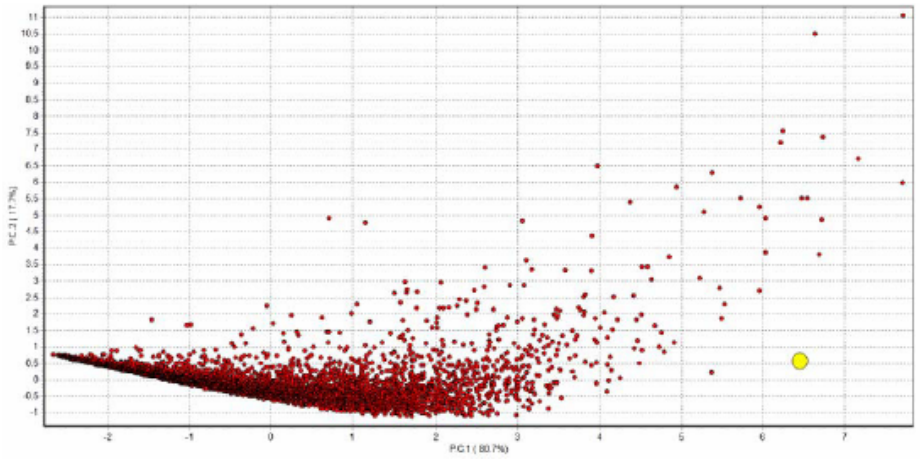
2 ( ) Pre-evaluate scenario-prior combinations

Some useful options of DIYABC



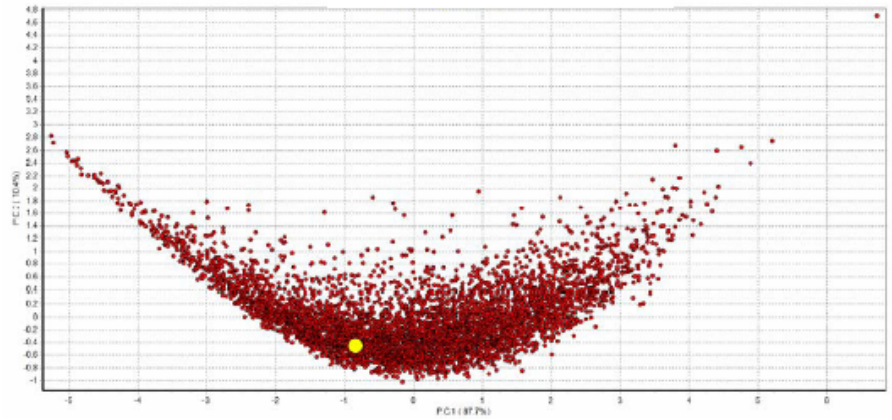
START WITH: Pre-evaluate scenario-prior combination (« prior checking »)

(A) "True" value of $N = 10{,}000$ – "incorrect" prior on N (Uniform[10; 1000]): note that the observed data set (large yellow dot) is positioned outside the cloud of simulated data sets (small dots)



		Probability ($s_{simulated} < s_{observed}$)	
Summary statistics	Observed value	"Incorrect" prior on N	"Correct" prior on N
NAL	9.4000	1.0000 (***)	0.3426
HET	0.7963	1.0000 (***)	0.3349
VAR	34.5601	0.9858 (*)	0.2467

(B) "True" value of $N = 10{,}000$ – "correct" prior on N (Uniform[2000; 20000]): note that the observed data set (large yellow dot) is positioned within the cloud of simulated data sets (small dots)



DATA AND ANALYSIS

Data file

E:\Harmonia\benoit facon\BF sciences\DIYABC\ben_plo

This data file contains 2 population samples including 133 individuals genotyped at 18 loci (1

able : E:\Harmonia\benoit facon\BF sciences\DIYABC\ben_plos_prior1_t48_DB30_NF300_size1_range50.reftable

The reference table, ben_plos_prior1_t48_DB30_NF300_size1_range50.reftable, co
Each record includes 8 parameters and 14 summary statistics
The reference table has been built with 4 scenarios

SOME OTHER USEFUL OPTIONS

Do you want to

- (•) Append new simulations to the reference table
- ( ) Estimate posterior distributions of parameters and model checking
- 7 → ( ) Compute bias and precision on parameter estimations
- ( ) Compute posterior probabilities of scenarios
- 4 → ( ) Evaluate confidence in scenario choice
- ( ) Pre-evaluate scenario-prior combinations

Evaluate confidence in scenario choice (4)

Define a set of compared scenarios (e.g. S1, S2, S3)

Define a target scenario (e.g. S1)

→Type I error on the target scenario (e.g. S1): proportion of cases in which the target scenario is not chosen while it is "true"

Simulation of « pseudo-observed data sets » under the target scenario - drawing parameter values into (prior) distributions

Compute probabilities of each scenario S1, S2, S3 → count p(S1) is not the highest

→ Type II error on the target scenario (e.g. S1): proportion of cases in which the target scenario is chosen while it is "wrong"

Simulation of « pseudo-observed data sets » under another scenario than the target one (e.g. S2) and drawing parameter values into (prior) distributions

Compute probabilities of each scenario S1, S2, S3 → p(S1) > count p(S1) is the highest

Compute bias and precision on parameter estimation (7)

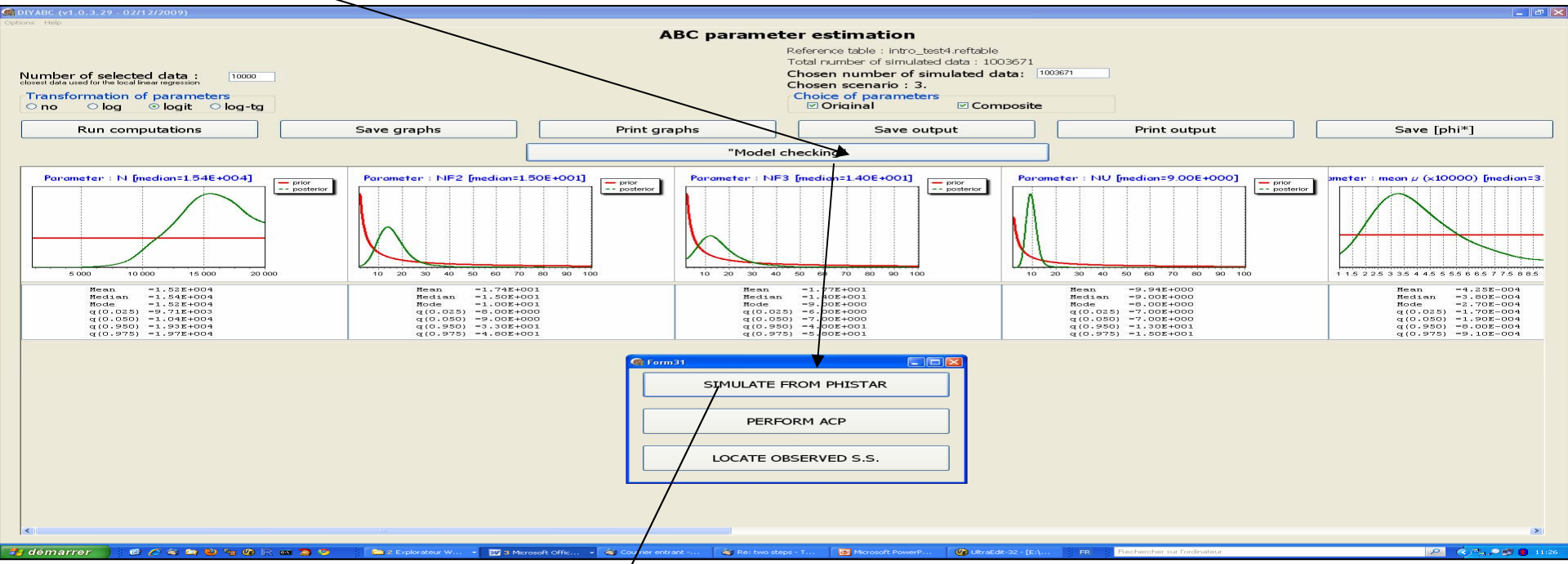
Define a target scenario (e.g. S1)

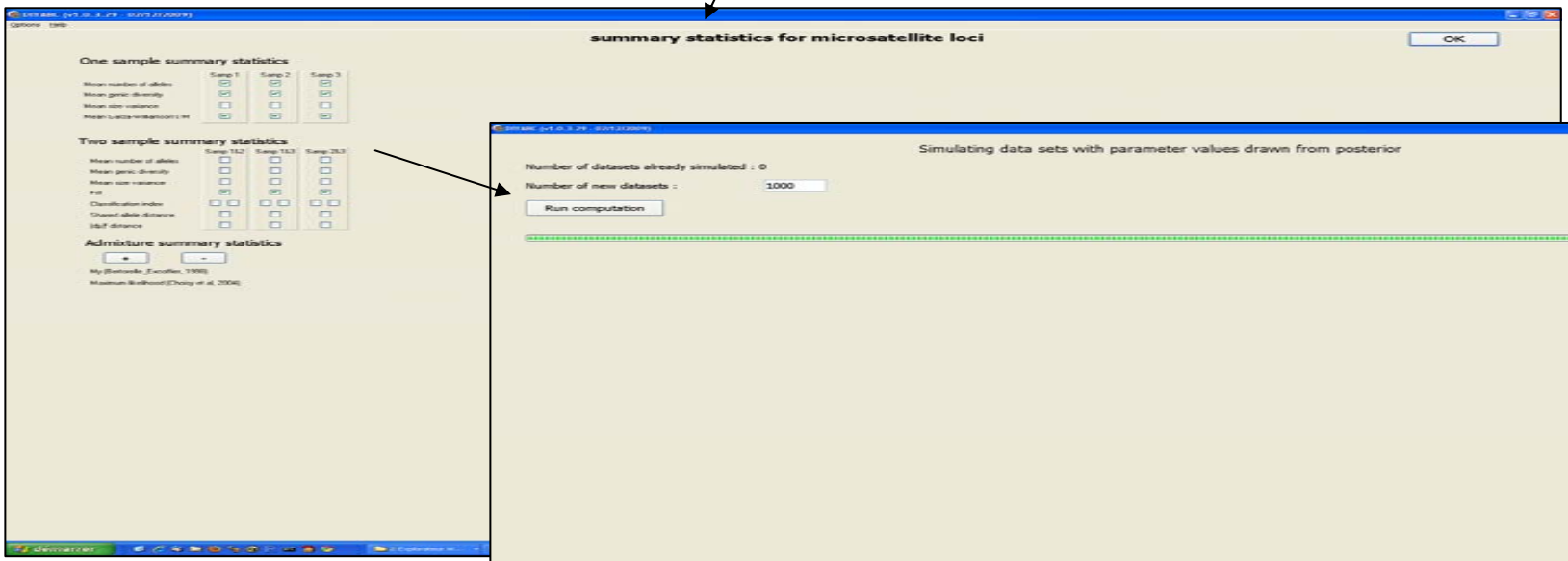
Simulation of « pseudo-observed data sets » under the target scenario - drawing parameter values into (prior) distributions (« true parameter values »)

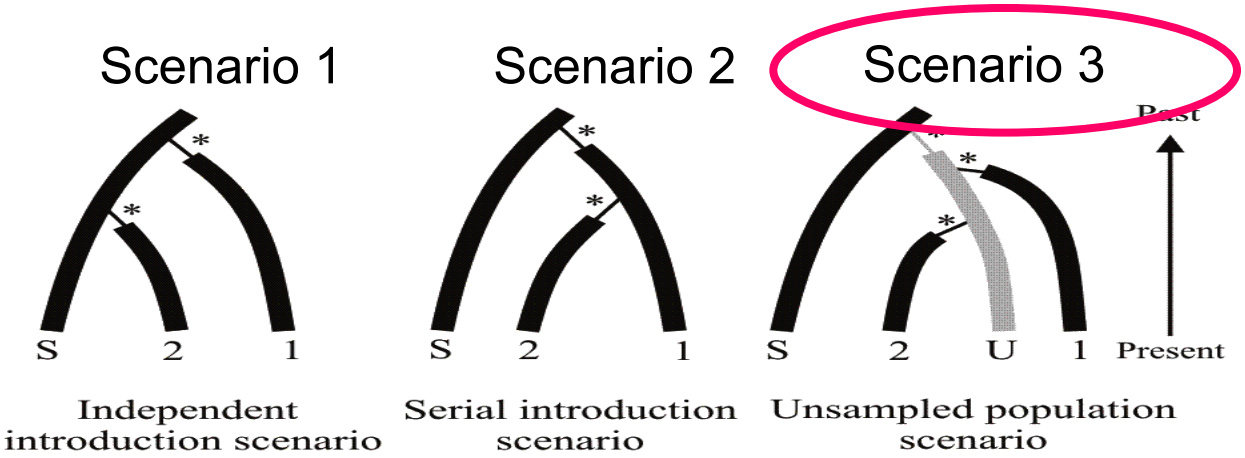
Compute posterior distributions for each parameter under the target scenario

True and estimated parameter values → Bias, RMSE,…for each parameter

FINISH WITH: Model-posterior checking option (“model checking”) = 5 and 6



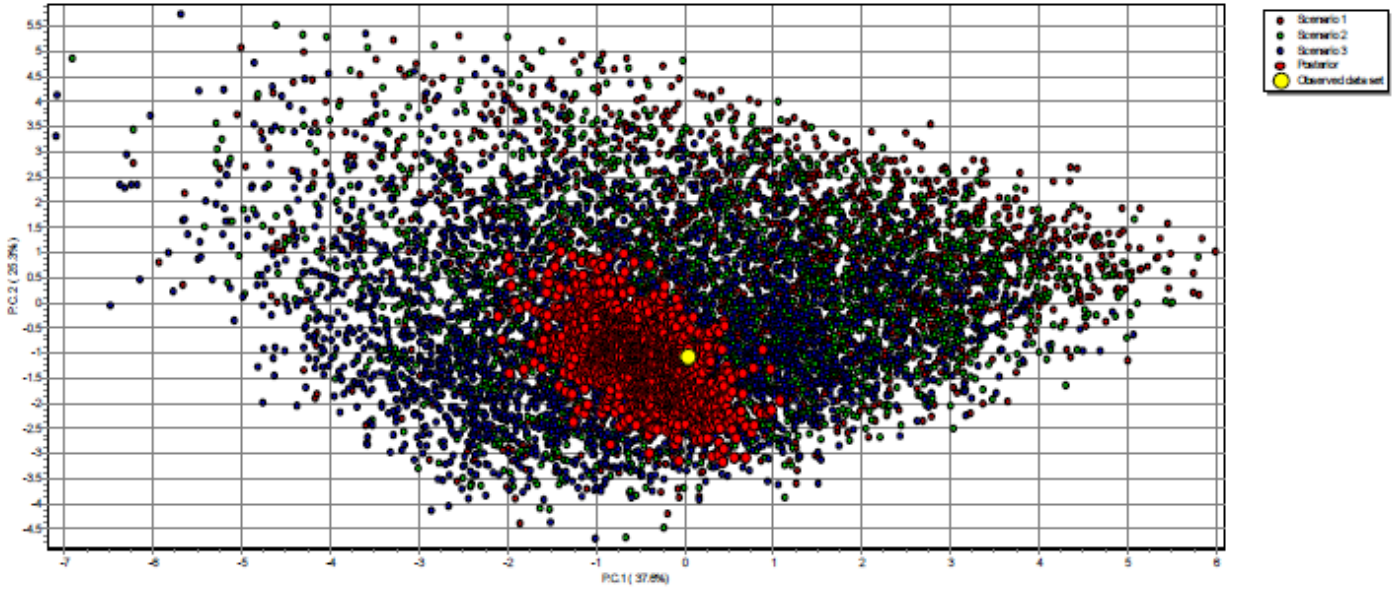




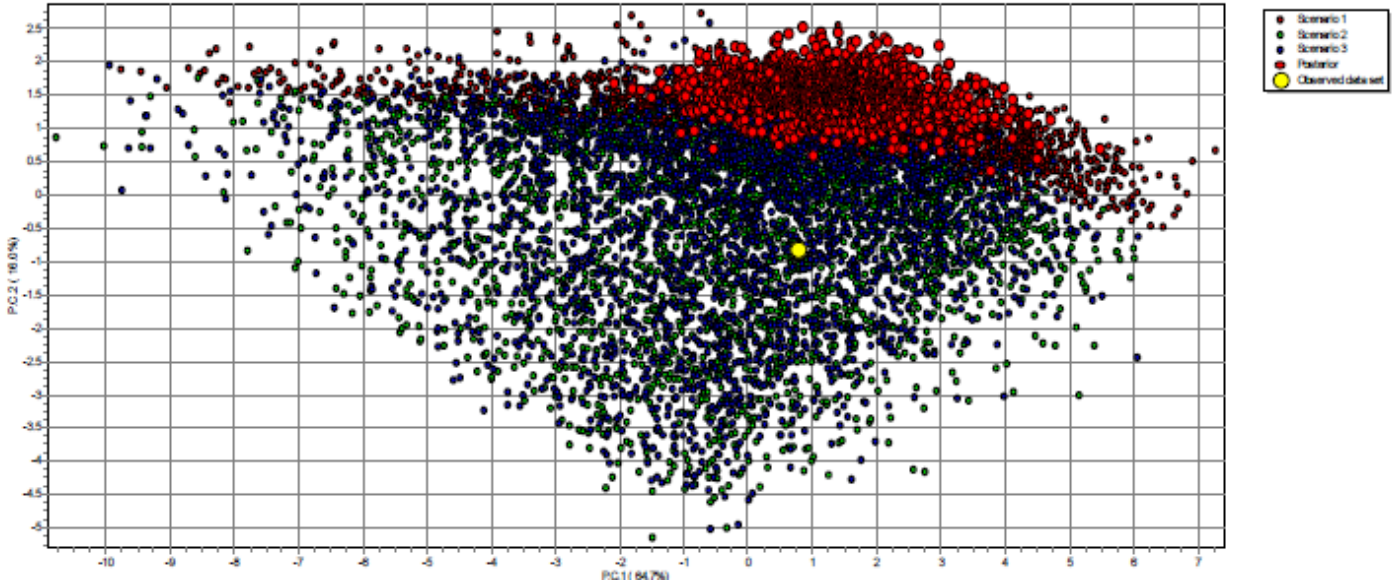
Scenarios to illustrate model-posterior checking

The three presented scenarios are often compared when making ABC inferences on the routes of introduction of invasive species. S is the source population in the native area, and U, the unsampled population in the introduced area that is actually the source of populations 1 and 2 in the scenario 3. The stars indicate the bottleneck events occurring in the first few generations following introductions. We here considered that the dates of first observation were well known so that divergence times could be fixed at 5, 10, 15 and 20 generations for t1, t2, t3 and t4, respectively. The datasets consisted of simulated genotypes at 20 statistically independent microsatellite loci obtained from a sample of diploid individuals collected from the invasive and source populations (30 individuals per population). The "pseudo-observed" test data set analyzed here to illustrate model-posterior checking was simulated under scenario 3 with an effective population size (NS) of 10,000 diploid individuals in all populations except during the bottleneck events corresponding to an effective population size (NFi) of 10 diploid individuals for 5 generations. Prior distributions for ABC analyses (discrimination of scenarios and estimation of posterior distribution of parameters) were as followed: Uniform[1000; 20000] for NS and logUniform[2;100] for NFi. The prior distributions of the microsatellite markers were the same than those described in the legend of Figure 1.

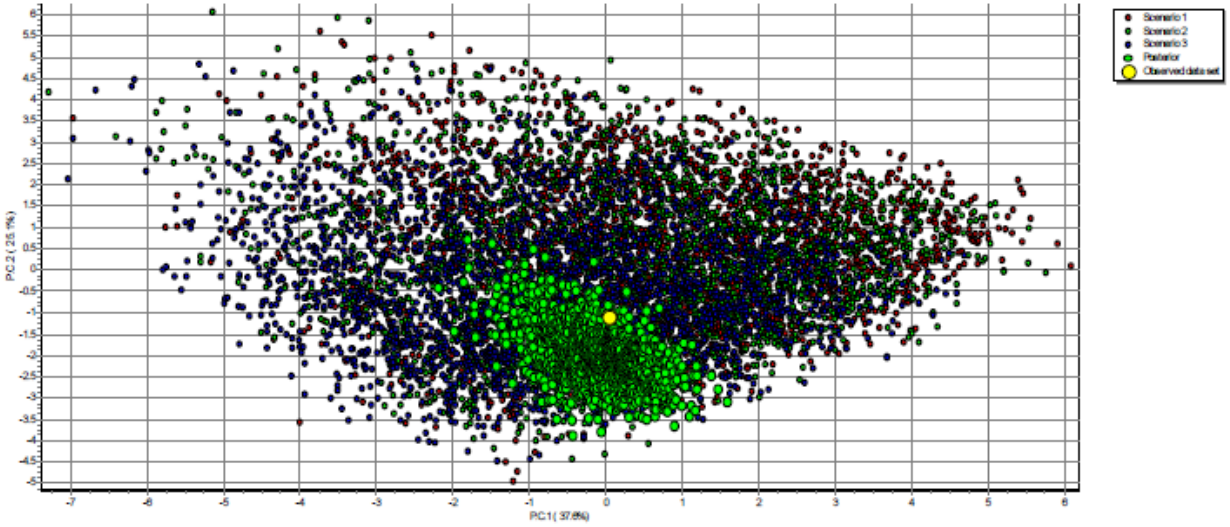
PCA on summary statistics used to estimate parameter posterior distributions under scenario 1



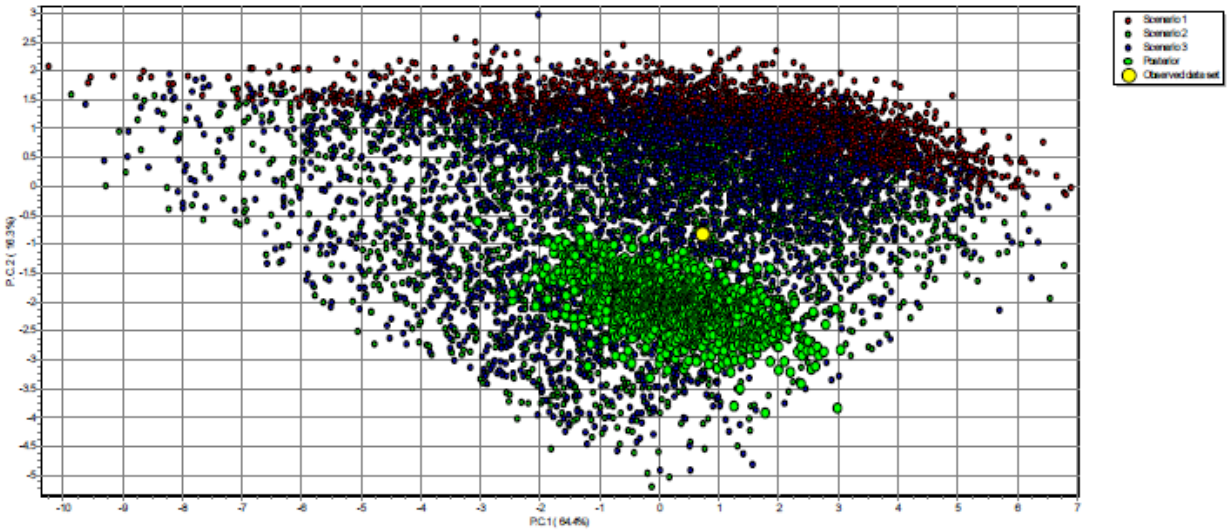
PCA on summary statistics NOT used to estimate parameter posterior distributions under scenario 1



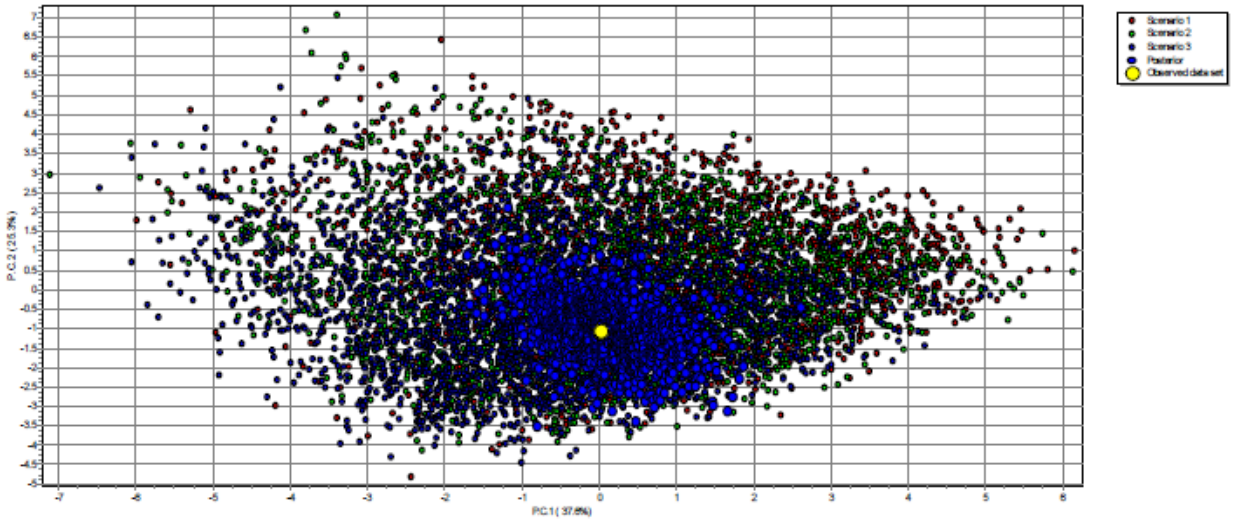
PCA on summary statistics used to estimate parameter posterior distributions under scenario 2



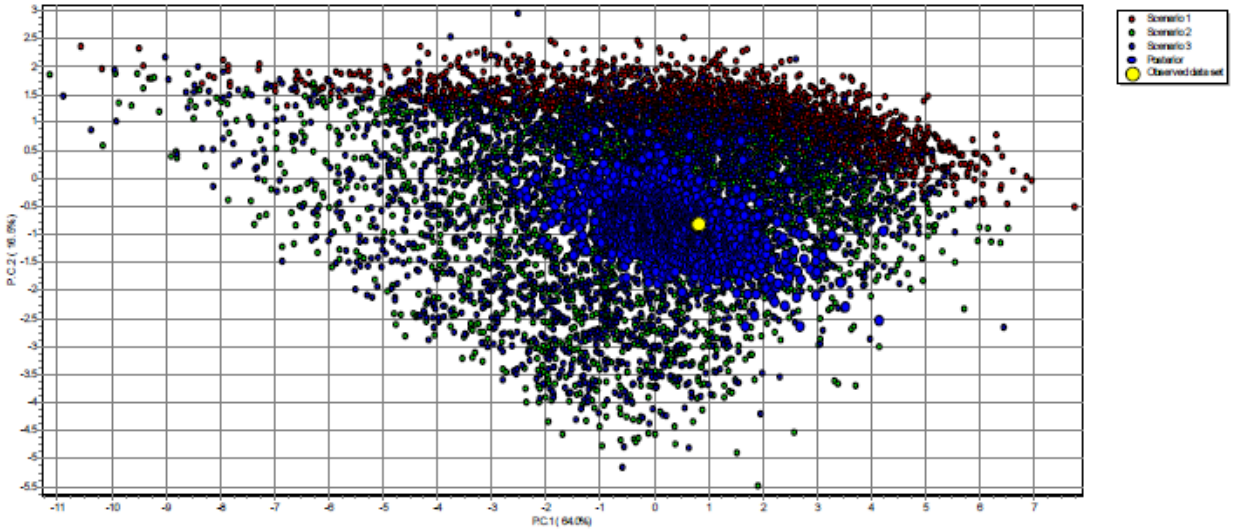
PCA on summary statistics NOT used to estimate parameter posterior distributions under scenario 2



PCA on summary statistics used to estimate parameter posterior distributions under scenario 3



PCA on summary statistics NOT used to estimate parameter posterior distributions under scenario 3



Model-posterior checking for the introduction scenarios 1, 2 and 3

The scenarios 1, 2 and 3 are detailed in Figure 3. The "pseudo-observed" test data set analyzed here was simulated under scenario 3. The probability (simulated<observed) given for each test quantities was computed from 1,000 data tests simulated from the posterior distributions of parameters obtained under a given scenario. Corresponding tail-area probabilities, or p-values, of the test quantities can be easily obtained as probability(simulated<observed) and |1.0 – probability(simulated<observed) | for probability(simulated<observed) ≤ 0.5 and > 0.5, respectively (Gelman et al. 1995). The test quantities correspond to the summary statistics used to compute the posterior distributions of parameters (a) or to other statistics (b). NAL_*i* = mean number of alleles in population *i*, HET_*i* = mean expected heterozygosity in population *i* (Nei 1987), MGW_*i* = mean ratio of the number of alleles over the range of allele sizes (Excoffier et al. 2005), FST_*i_j* = *F*ST value between populations *i* and *j* (Weir and Cockerham 1984), VAR_*i* = mean allelic size variance in population *i*, LIK_*i_j* = mean individual assignment likelihoods of population *i* assigned to population *j* (Pascual et al. 2007), H2P_*i_j* = mean expected heterozygosity pooling samples from populations *i* and *j*, DAS_*i_j* = shared allele distance between populations *i* and *j* (Chakraborty and Jin, 1993). *, **, *** = tail-area probability < 0.05, < 0.01 and < 0.001, respectively.

individual assignment likelihoods of population *i* assigned to population *j* (Pascual et al. 2007), H2P_*i_j* = mean expected heterozygosity pooling samples from populations *i* and *j*, DAS_*i_j* = shared allele distance between populations *i* and *j* (Chakraborty and Jin, 1993). *, **, *** = tail-area probability < 0.05, < 0.01 and < 0.001, respectively.

(a) Summary statistics used to compute parameter posterior distributions

summary statistics	observed value	Probability (simulated<observed) scenario 1		scenario 2		scenario 3
NAL_1	13.6000	0.7275		0.2875		0.6235
NAL_2	3.4000	0.7540		0.9865	(*)	0.4250
NAL_3	3.6500	0.6455		0.4105		0.4765
HET_1	0.8429	0.5625		0.2470		0.4485
HET_2	0.5151	0.4930		0.9890	(*)	0.4335
HET_3	0.5725	0.9125		0.9180		0.8220
MGW_1	0.8242	0.3590		0.7650		0.5230
MGW_2	0.4072	0.3780		0.6710		0.4520
MGW_3	0.4834	0.6110		0.8495		0.7290
FST_1&2	0.2170	0.7880		0.0370	(*)	0.8105
FST_1&3	0.2050	0.6180		0.4605		0.6050
FST_2&3	0.1761	0.0001	(***)	0.9580	(*)	0.6280

(b) Summary statistics NOT used to compute parameter posterior distributions

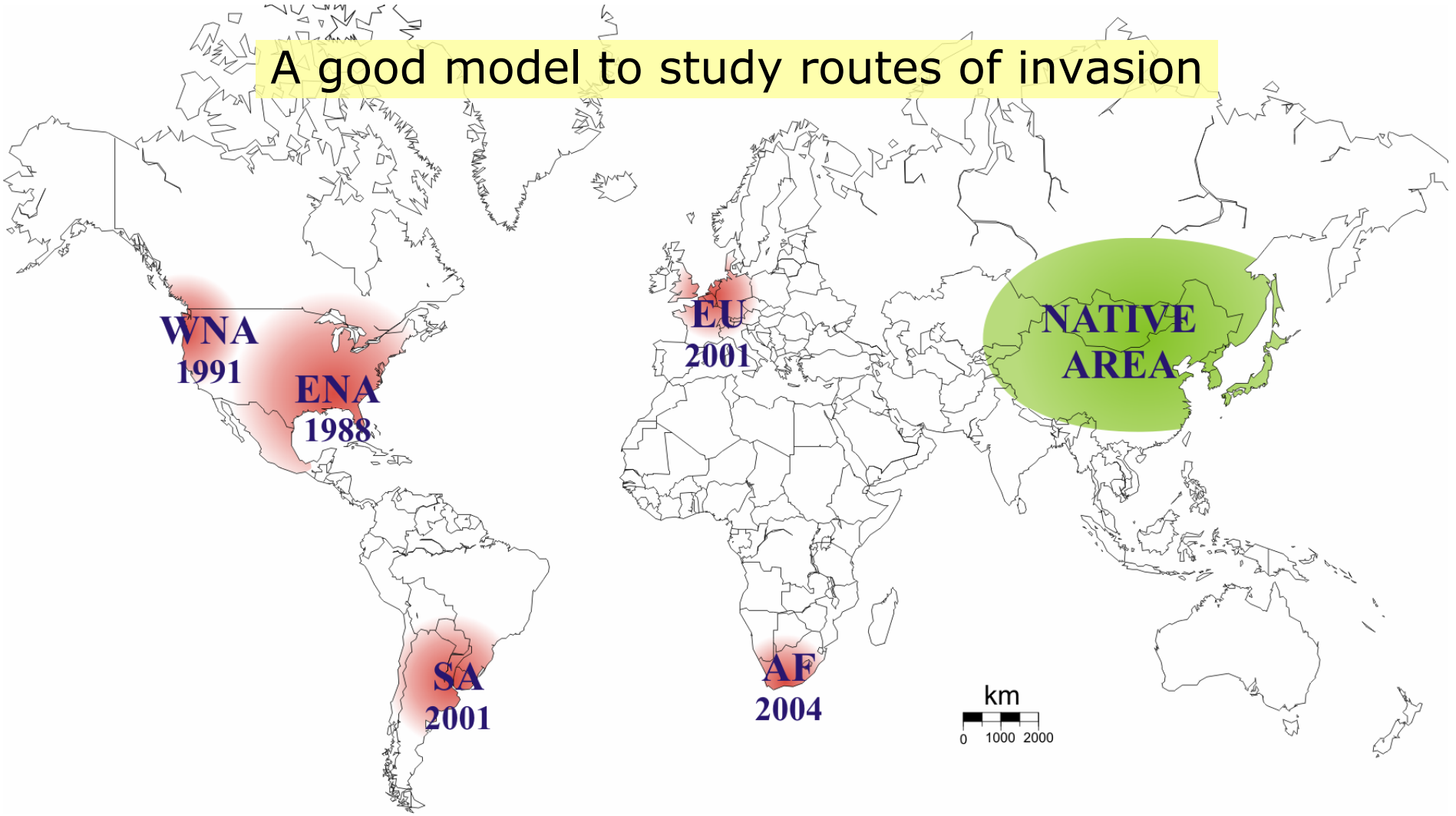
summary statistics	observed value	Probability (simulated<observed) scenario 1		scenario 2		scenario 3
VAR_1	21.7561	0.7475		0.2530		0.6200
VAR_2	9.3385	0.4860		0.3560		0.3590
VAR_3	9.5277	0.5230		0.1790		0.3740
LIK_2&1	38.5648	0.7860		0.4500		0.7240
LIK_2&3	31.7504	0.0000	(***)	1.0000	(***)	0.7160
LIK_3&2	32.1075	0.0000	(***)	0.9850	(*)	0.7830
H2P_1&2	0.7734	0.6565		0.8410		0.6115
H2P_1&3	0.7993	0.9230		0.8230		0.8665
H2P_2&3	0.6020	0.0315	(*)	0.9975	(**)	0.7190
DAS_1&2	0.1329	0.2290		0.4580		0.2635
DAS_1&3	0.1099	0.0550		0.1680		0.0815
DAS_2&3	0.3402	1.0000	(***)	0.0000	(***)	0.2520

Part 4. Inferences about a complex population history – Example 1: worldwide routes of invasion of the ladybird *Harmonia axydis*

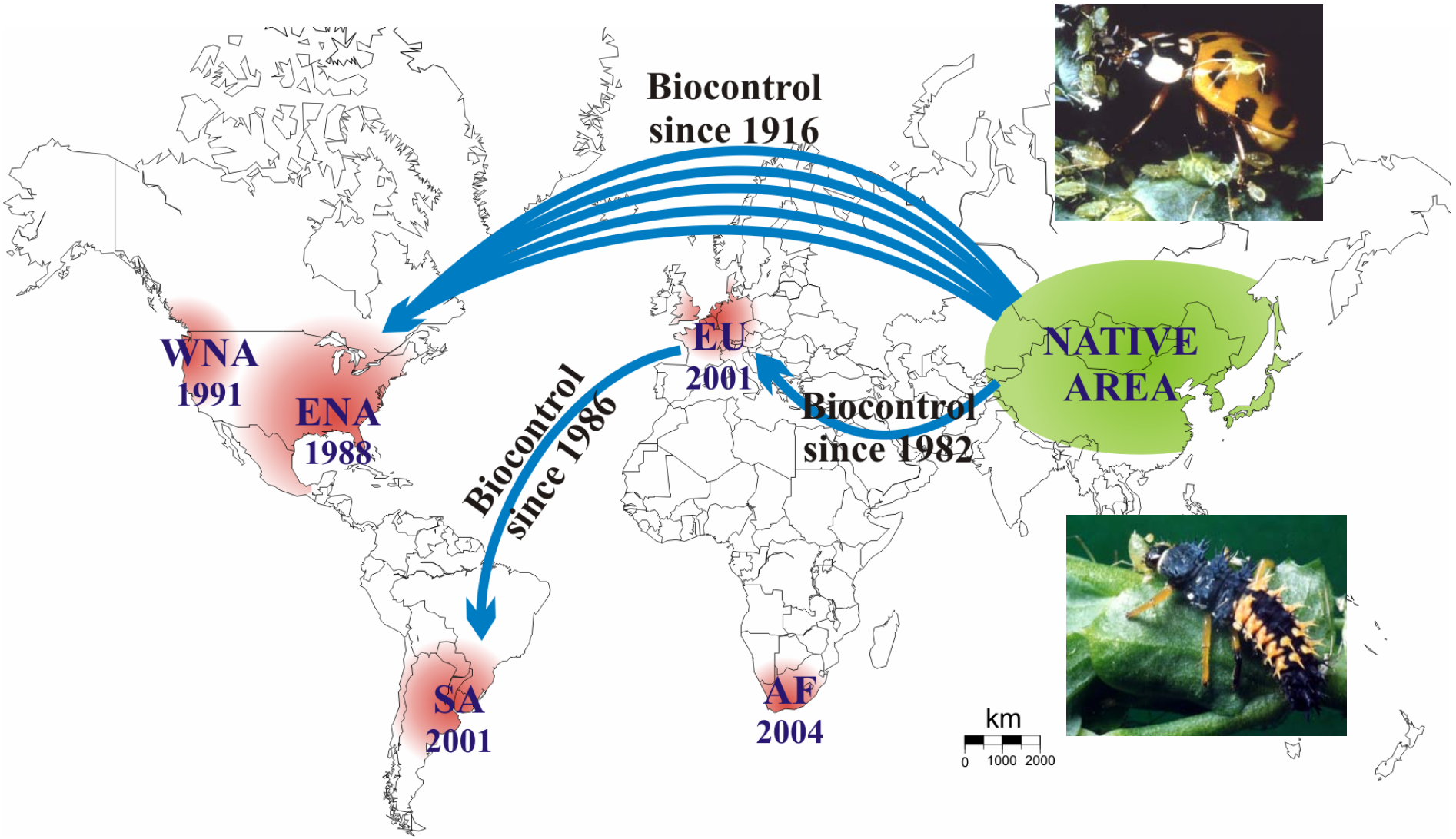
E. Lombaert, T. Guillemaud, J.M. Cornuet, T. Malausa, B. Facon, A. Estoup. 2010. Bridgehead effect in the worldwide invasion of the biocontrol harlequin ladybird. *Plos One*, 5(3) e9743.

Harmonia axyridis

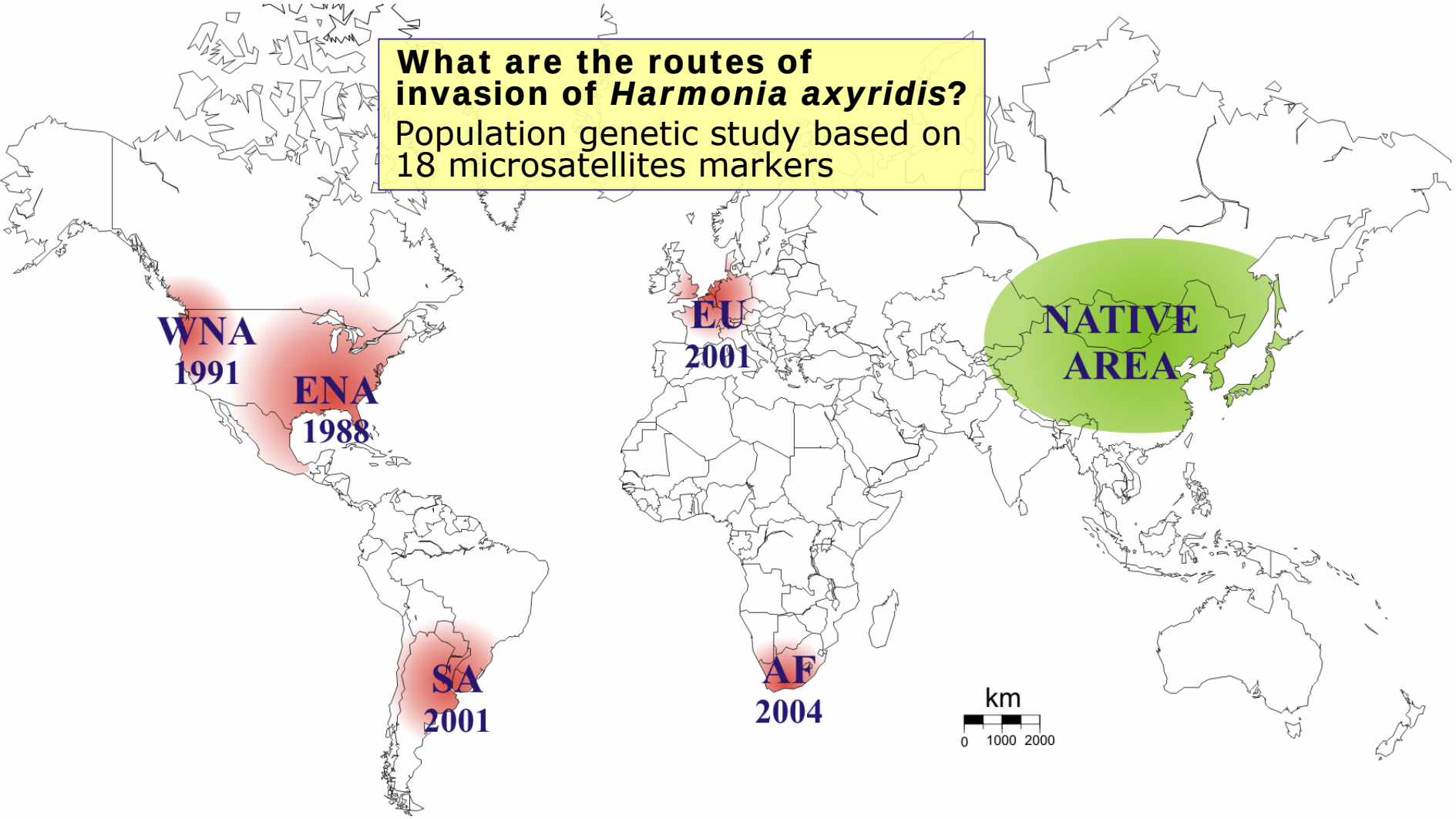
A good model to study routes of invasion



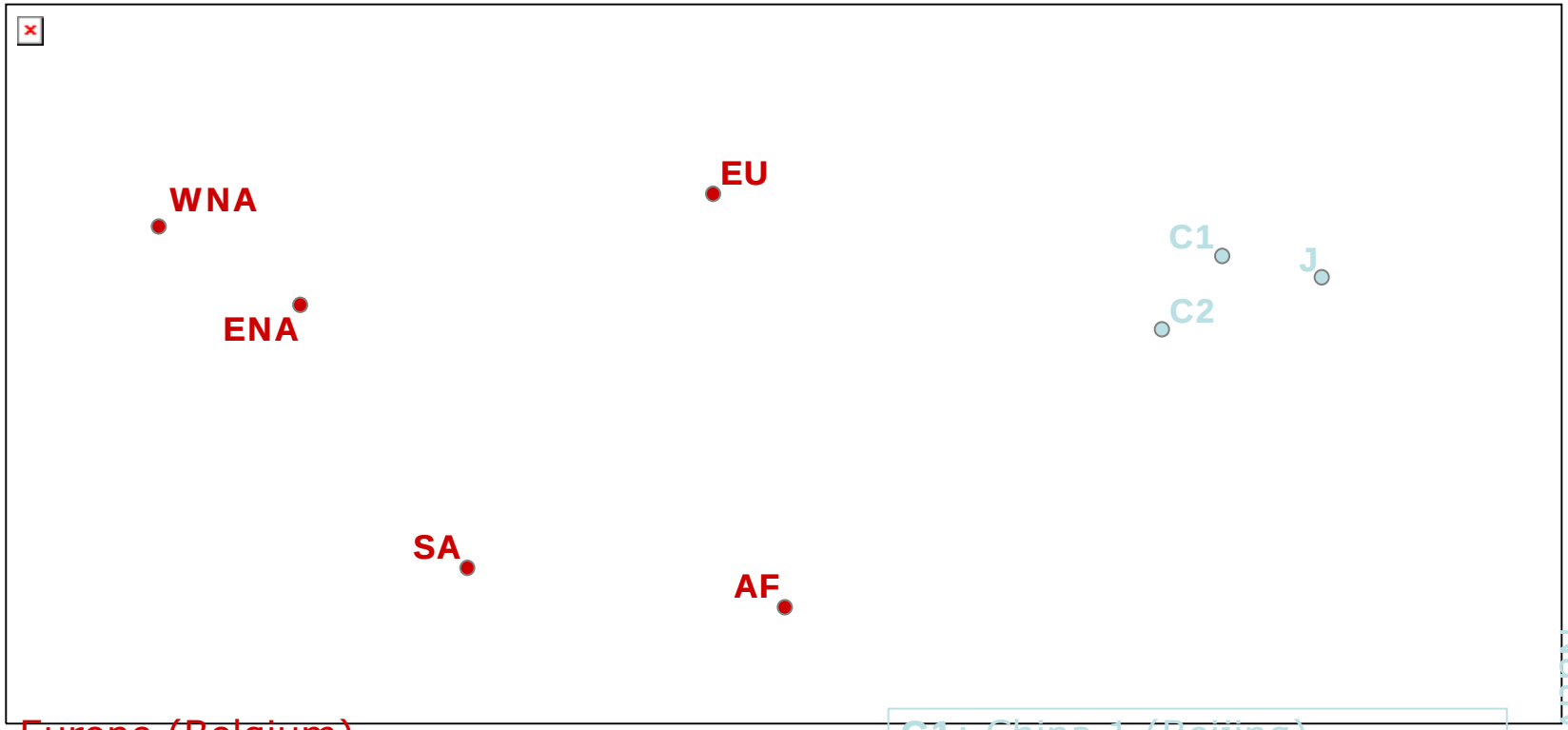
Harmonia axyridis



Harmonia axyridis



Collected samples



EU: Europe (Belgium)
ENA: Eastern North America (Louisiana)
WNA: Western North America (Washington)
SA: South America (Brazil)
AF: Africa (South Africa)

(18 < N < 42)

C1: China 1 (Beijing)
C2: China 3 (Yunnan province)
J: Japan

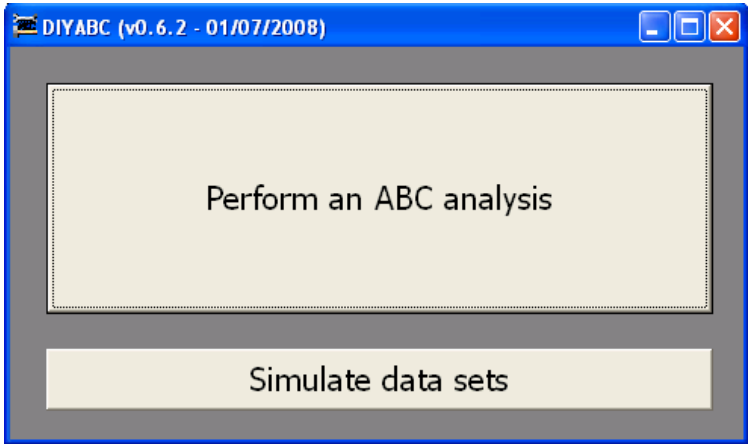
Native (West)

+ EBC: European Biocontrol Strain (INRA 1987)

Recontruction of the routes of invasion

- Approximate Bayesian Computation ("Standard" ABC, Beaumont et al. 2002) using the software DIYABC (Cornuet et al. 2008, 2010):

5 consecutive analyses (cf. dates of first observation)



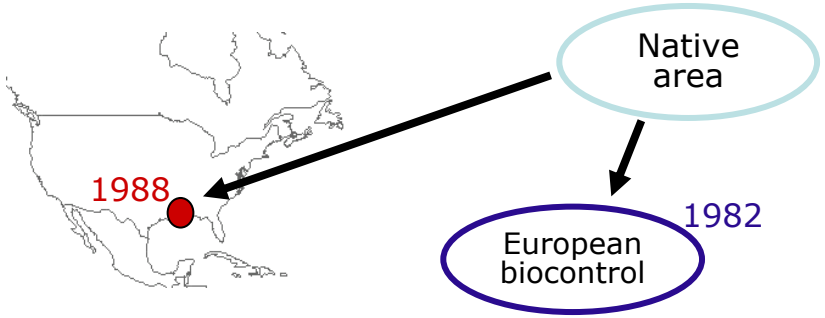
– Step 1: Definition of invasion scenarios.

– Step 2: Simulation of genetic data based on the coalescent model (1M per scenario)
→ summary statistics (Na, He, FST, etc.).

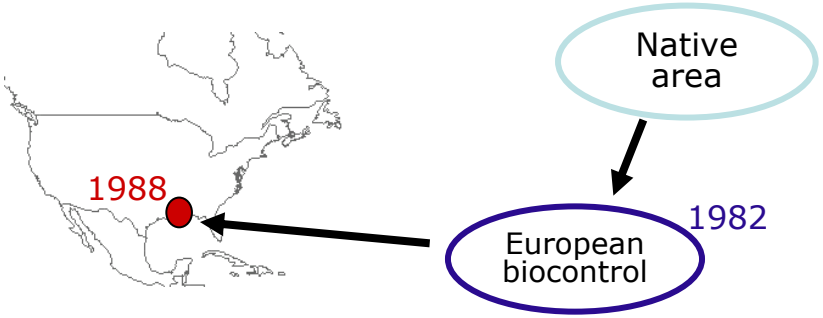
– Step 3: Computation of posterior probabilities of scenarios (logistic regression method: 1% of the best simulations)

Analysis I. East North America (1988)

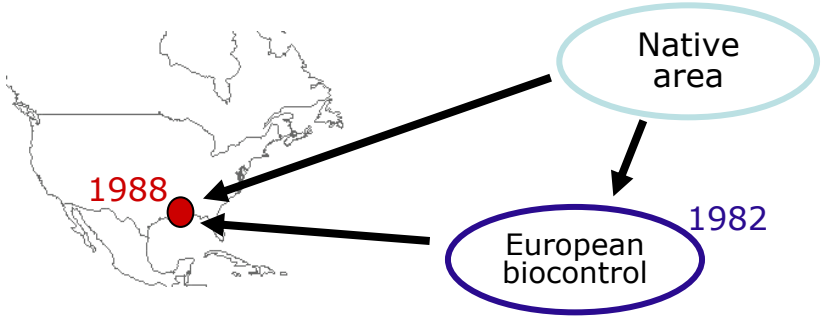
➔ 3 scenarios



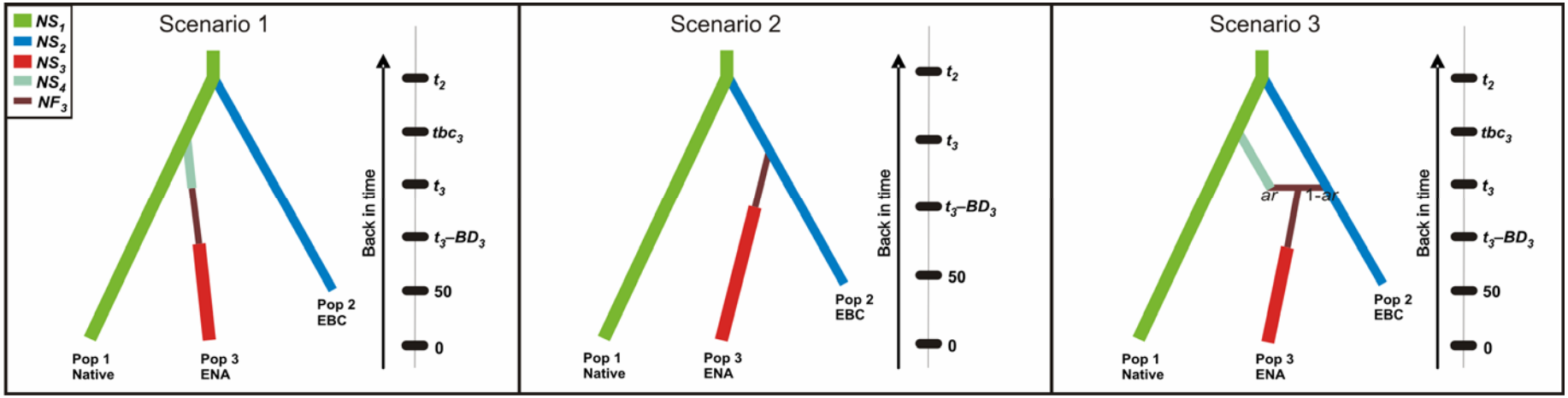
– 1. Native area



– 2. European biocontrol



– 3. Native area + European biocontrol


NS_1
NS_2
NS_3
NS_4
NF_3
Scenario 1
Pop 1
Native
Pop 3
ENA
Pop 2
EBC
Back in time
t_2
tbc_3
t_3
t_3–BD_3
50
0
Scenario 2
Pop 1
Native
Pop 3
ENA
Pop 2
EBC
Back in time
t_2
t_3
t_3–BD_3
50
0
Scenario 3
ar
1-ar
Pop 1
Native
Pop 3
ENA
Pop 2
EBC
Back in time
t_2
tbc_3
t_3
t_3–BD_3
50
0

Prior distributions

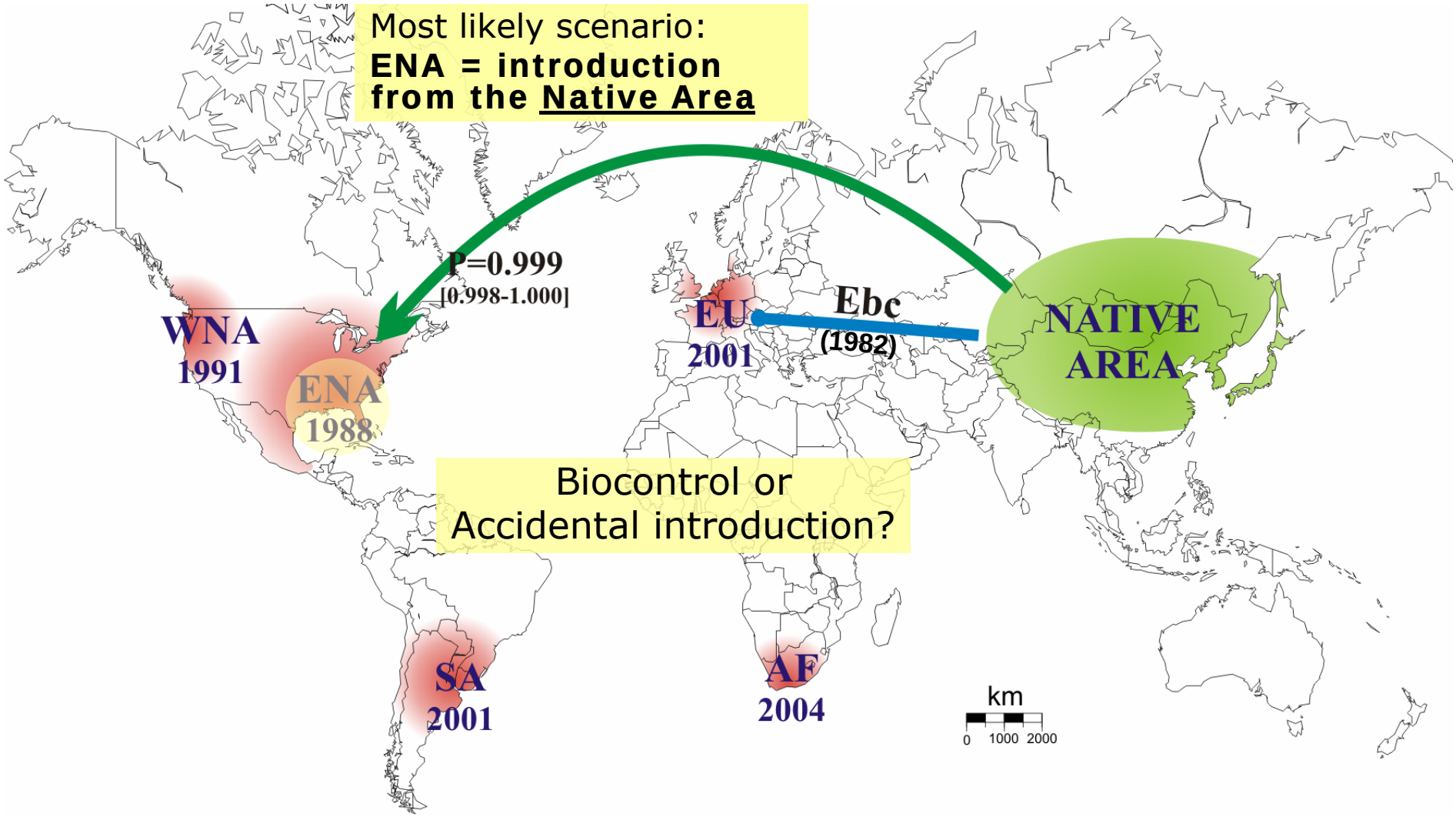
parameters	Prior Set 1 Distribution	Mean	Median	Mode	Quantile 2.5%	Quantile 97.5%	Prior Set 2 Distribution	Mean	Median	Mode	Quantile 2.5%	Quantile 97.5%
NS_i	Uniform [100 – 20,000]	10,056	10,040	NA	640	19,490	Normal (10,000 ; 5,000)	9,993	9,990	9,980	1,640	18,340
NS_k	Uniform [10 – 1,000]	506	508	NA	35	975	Normal (500 ; 250)	502	501	498	86	922
NF_i	Loguniform [2 – 1,000]	162	45	2	2	862	Lognormal (30 ; 30)	136	39	44	2	797
ar	Uniform [0.1 – 0.9]	0.5	0.5	NA	0.12	0.88	Normal (0.5 ; 0.25)	0.5	0.5	0.5	0.15	0.86
t_i	Uniform $[x_i - x_i+5]$	DV	DV	NA	DV	DV	Uniform $[y_i - y_i+5]$	DV	DV	NA	DV	DV
tbc_i	Loguniform $[x_i - 93]$	DV	DV	DV	DV	DV	Loguniform $[y_i - 111]$	DV	DV	DV	DV	DV
BD_i	Uniform [0 – 5]	2.5	2.5	NA	0	5	Uniform [0 – 5]	2.5	2.5	NA	0	5
mean μ	Uniform $[10^{-5} – 10^{-3}]$	$5.0x10^{-4}$	$5.0x10^{-4}$	NA	$3.5x10^{-5}$	$9.8x10^{-4}$	Loguniform $[10^{-5} – 10^{-3}]$	$2.1x10^{-4}$	$1.0x10^{-4}$	$1.0x10^{-5}$	$1.1x10^{-5}$	$8.9x10^{-4}$
mean P	Uniform [0.1 – 0.3]	0.2	0.2	NA	0.10	0.29	Gamma (30 ; 136)	0.22	0.22	0.21	0.15	0.29
mean μSNI	Uniform $[10^{-8} – 10^{-4}]$	$5.0x10^{-5}$	$5.0x10^{-5}$	NA	$2.5x10^{-6}$	$9.7x10^{-5}$	Loguniform $[10^{-8} – 10^{-4}]$	$1.1x10^{-5}$	$1.0x10^{-6}$	$1.0x10^{-8}$	$1.3x10^{-8}$	$7.9x10^{-5}$

Analysis I. East North America (1988)

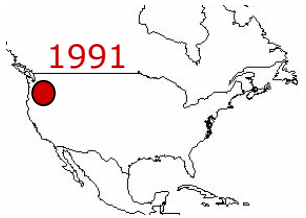
- Results:

Scenarios	Posterior probability (logistic regression)
Native	**0.999**
European biocontrol (Ebc)	0
Native + Ebc	0.001

Analysis I. East North America (1988)



Analysis II. West North America (1991)

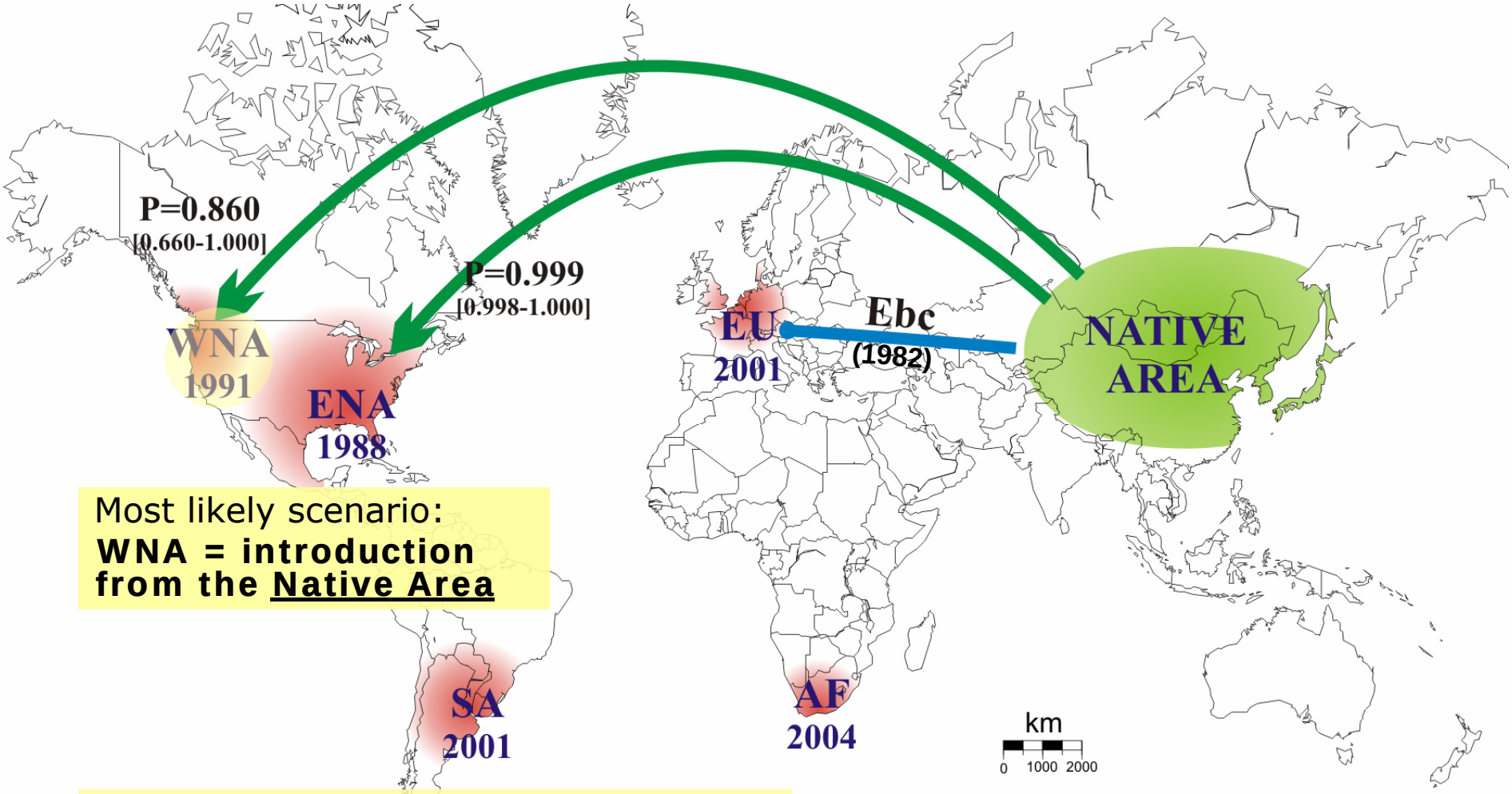


- Single introduction scenarios:
  - Native
  - European biocontrol (Ebc; 1982)
  - East North America (ENA; 1988)

- Admixed introduction scenarios:
  - Native + Ebc
  - Native + ENA
  - Ebc + ENA

➔ 6 scenarios

Analysis II. West North America (1991)



Most likely scenario:
WNA = introduction from the <u>Native Area</u>

➔ **2 independent introductions in North America**

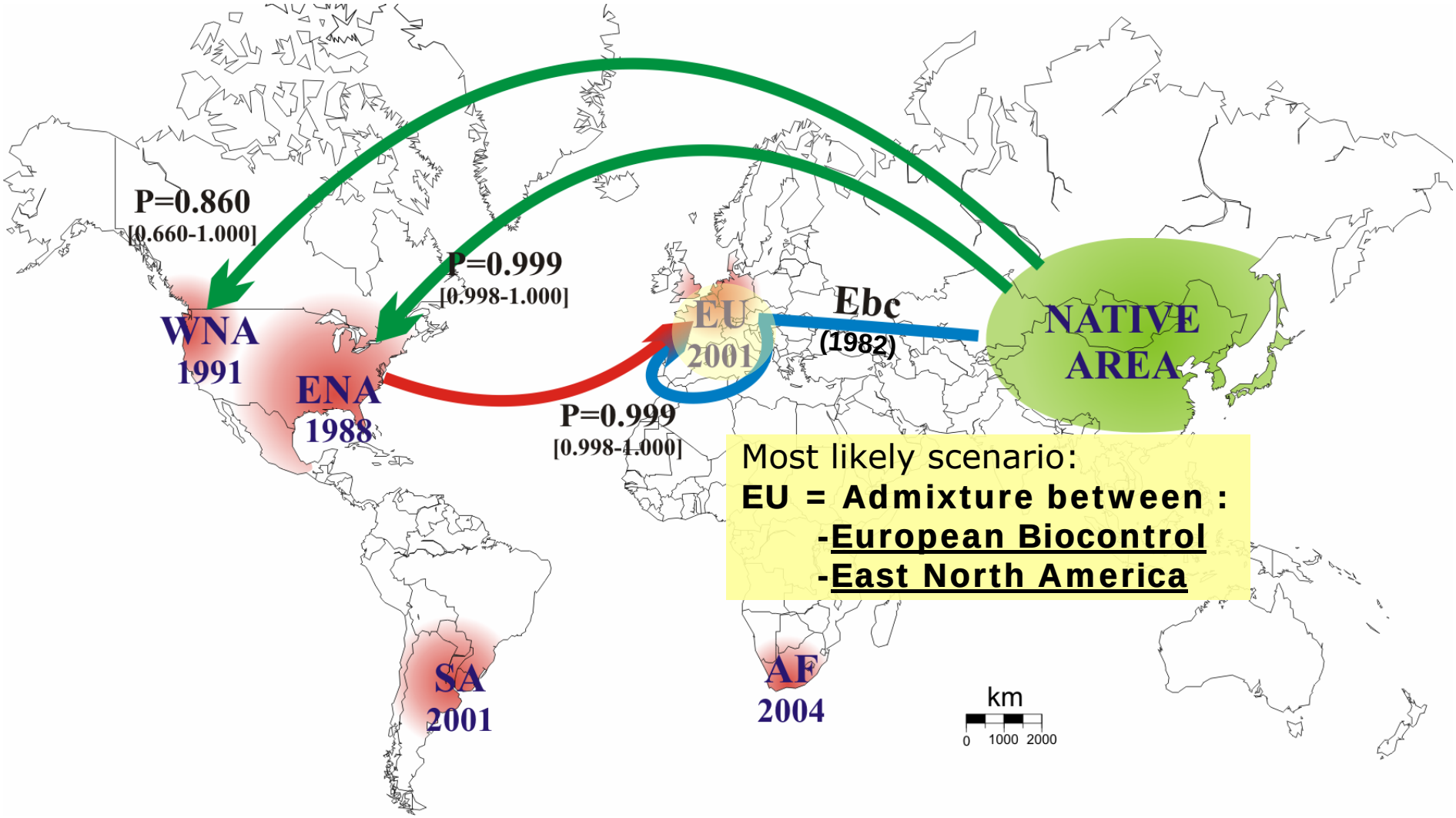
Analysis III. Europe (2001)

- Single introduction scenarios:
  - Native
  - European biocontrol (Ebc; 1982)
  - East North America (ENA; 1988)
  - West North America (WNA; 1991)

- Admixed introduction scenarios:
  - Native + Ebc
  - Native + ENA
  - Native + WNA
  - Ebc + ENA
  - Ebc + WNA
  - ENA + WNA

➔ 10 scenarios

Analysis III. Europe (2001)



Most likely scenario:

EU = Admixture between :

-<u>European Biocontrol</u>

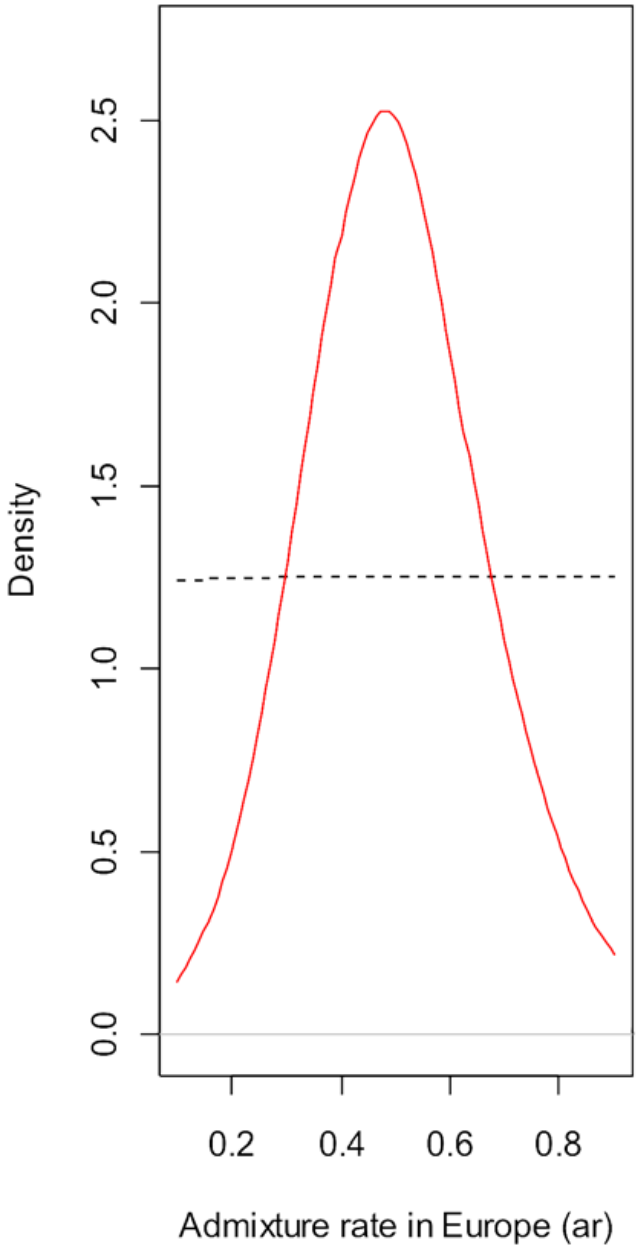
-<u>East North America</u>

ar = Proportion of genes of Ebc origin

Estimation using the local linear regression method under the most likely scenario (1% of best simulations)

Median value of *ar* = 0.43
95% CI: [0.18–0.83]

→ Substantial genetic contribution of European Biocontrol



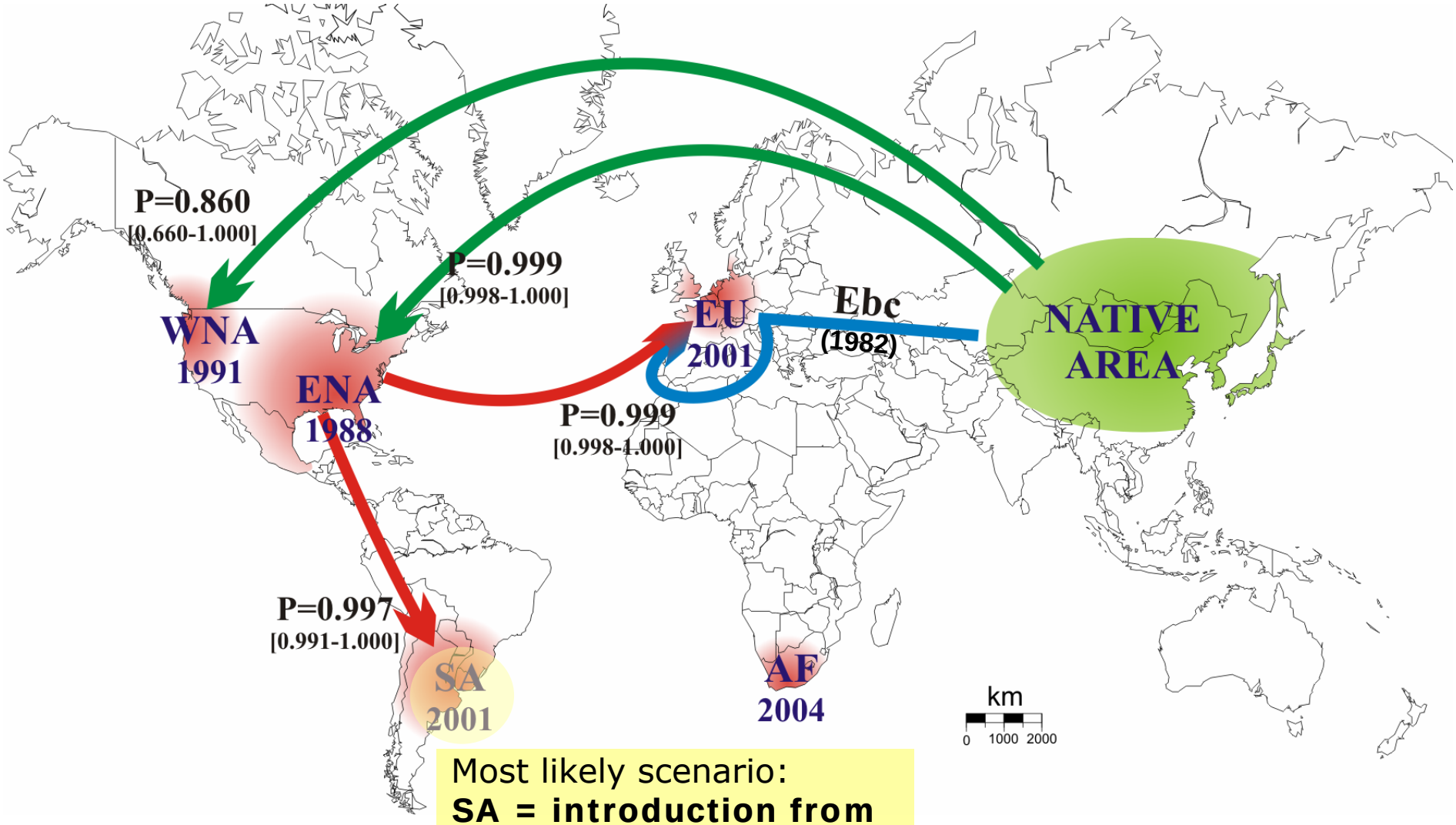
Analysis IV. South America (2001)

- Single introduction scenarios:
  - Native
  - European biocontrol (Ebc; 1982)
  - East North America (ENA; 1988)
  - West North America (WNA; 1991)

- Admixed introduction scenarios:
  - Native + Ebc
  - Native + ENA
  - Native + WNA
  - Ebc + ENA
  - Ebc + WNA
  - ENA + WNA

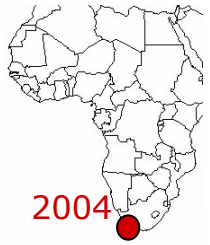
➔ 10 scenarios

Analysis IV. South America (2001)



Most likely scenario:
SA = introduction from East North America

Analysis V. Africa (2004)

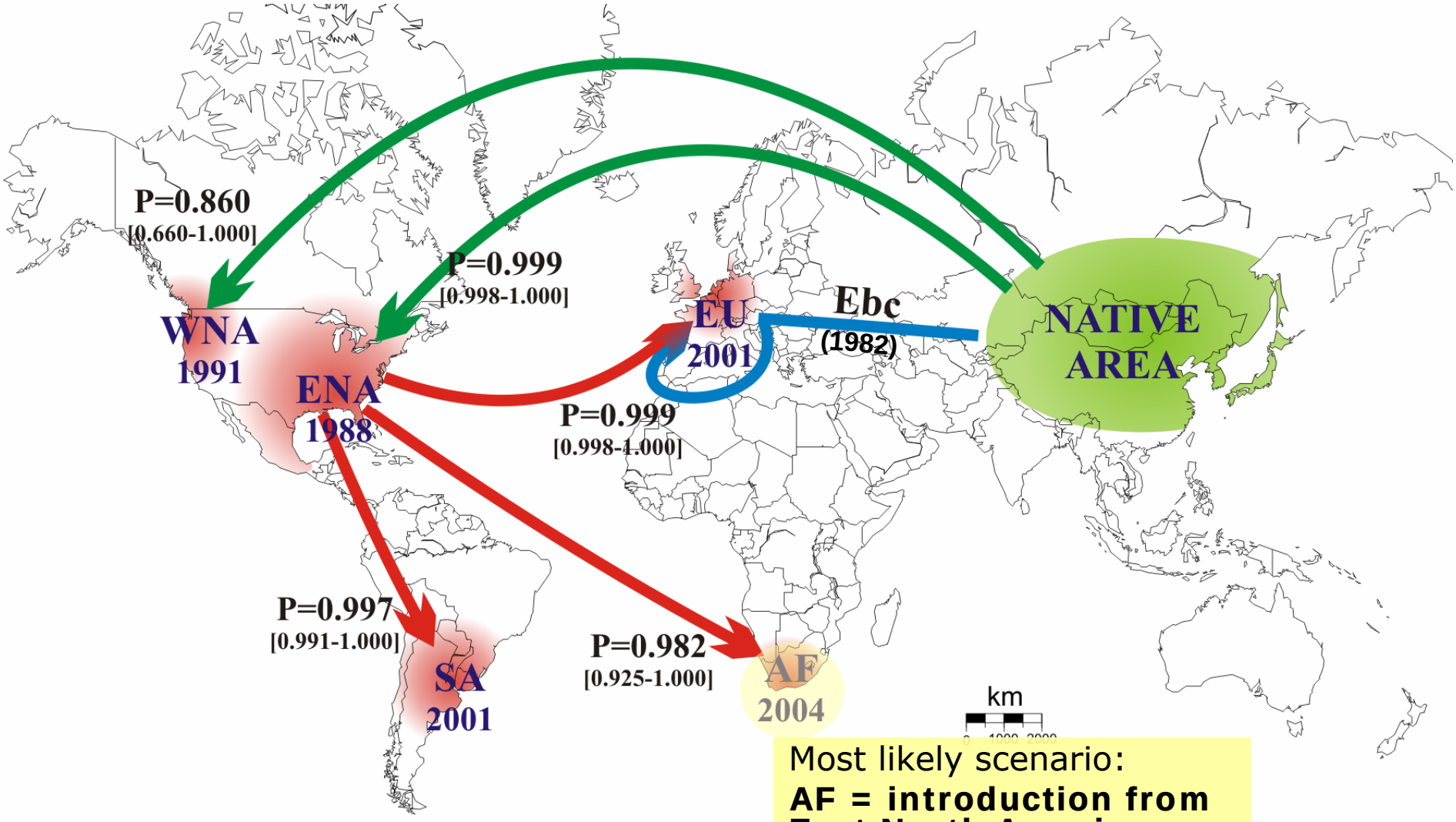


- Single introduction scenarios:
  - Native
  - European Biocontrol (Ebc; 1982)
  - East North America (ENA; 1988)
  - West North America (WNA; 1991)
  - Europe (EU; 2001)
  - South America (SA; 2001)

- Admixed introduction scenarios:
  - Native + Ebc
  - Native + ENA
  - Native + WNA
  - Native + EU
  - Native + SA
  - Ebc + ENA
  - Ebc + WNA
  - Ebc + EU
  - Ebc + SA
  - ENA + WNA
  - ENA + EU
  - ENA + SA
  - WNA + EU
  - WNA + SA
  - EU + SA

➔ 21 scenarios

Analysis V. Africa (2004)



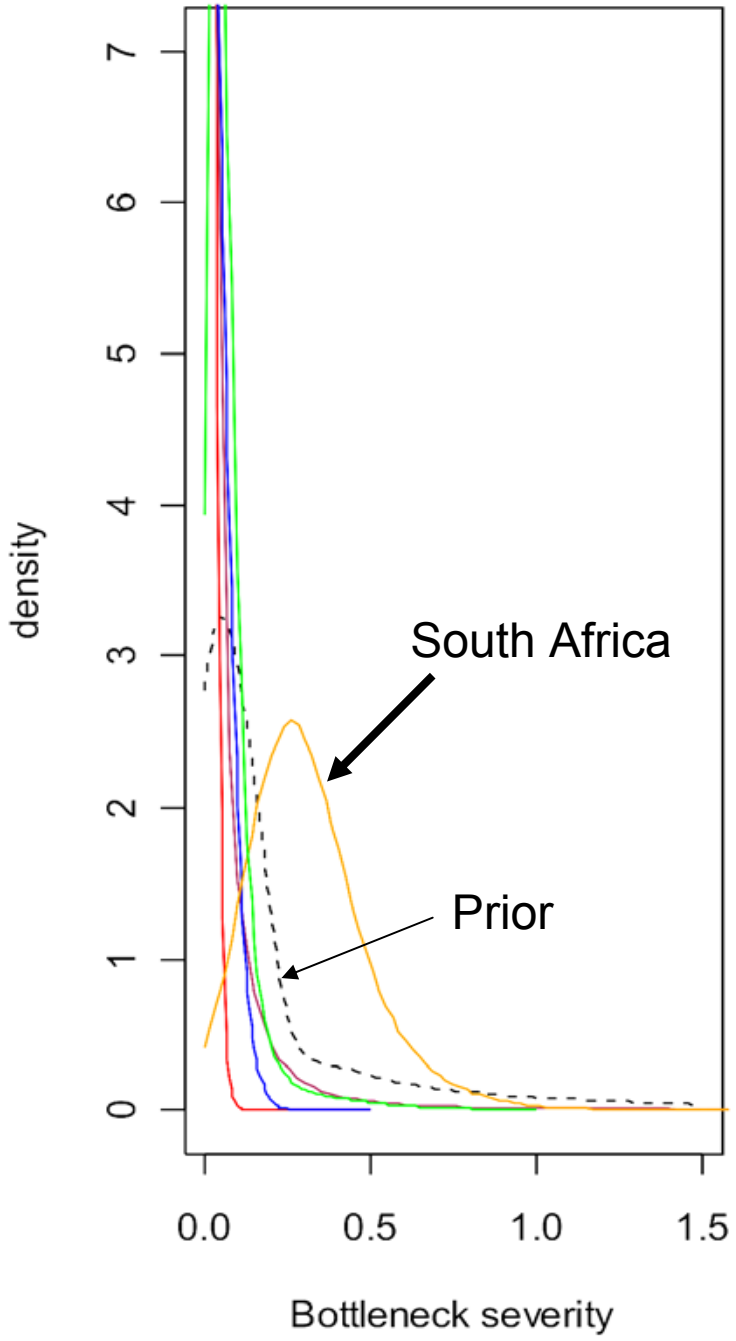
Most likely scenario:

AF = introduction from East North America

Bottleneck severity = DB/NB

Estimation for each invaded area using the local linear regression method (1% of best simulations)

→ Low bottleneck severity except for the invasion of South Africa.



Confidence in scenario selection

- Sensitivity to priors (for each consecutive analysis)

  2 sets of prior tested ➔ similar results

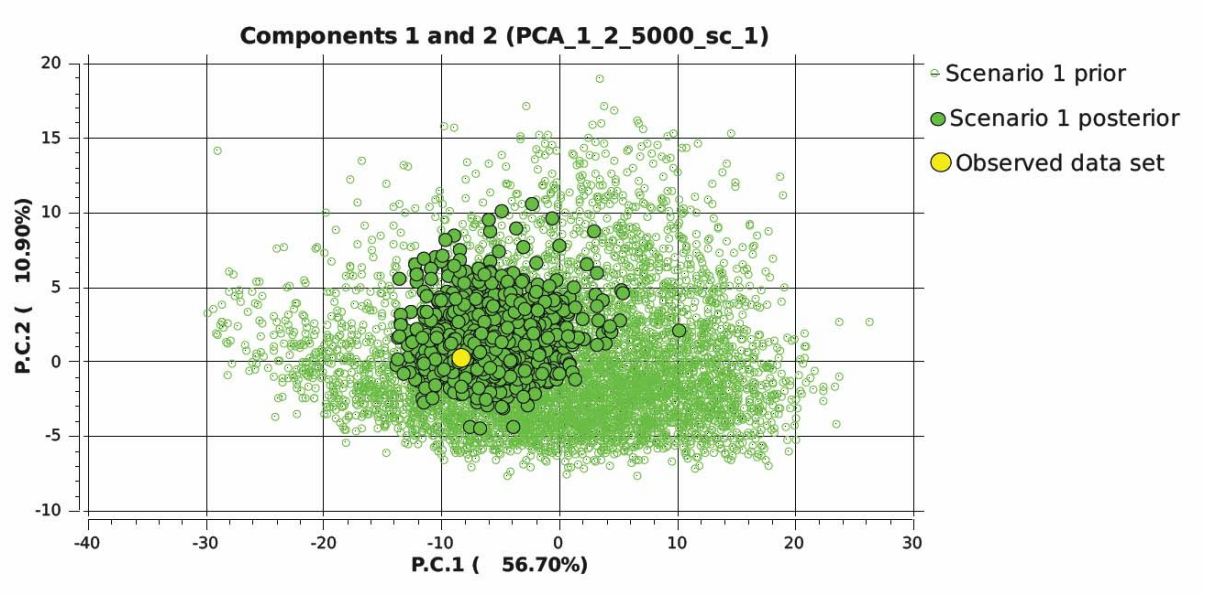
- Type I & II errors (for each consecutive analysis)

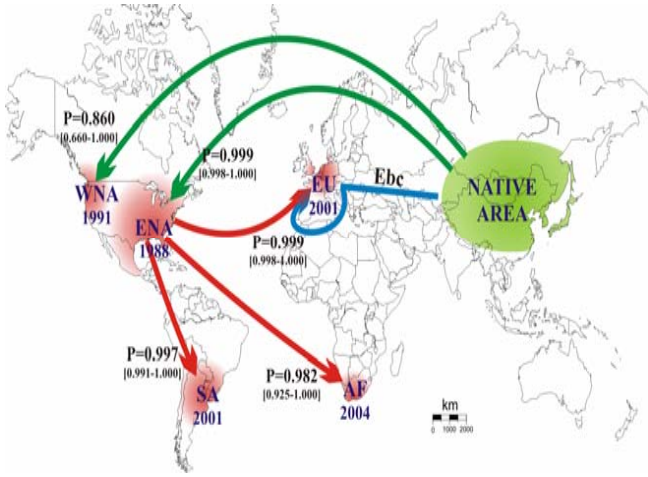
  Computed from simulated data ➔ low errors

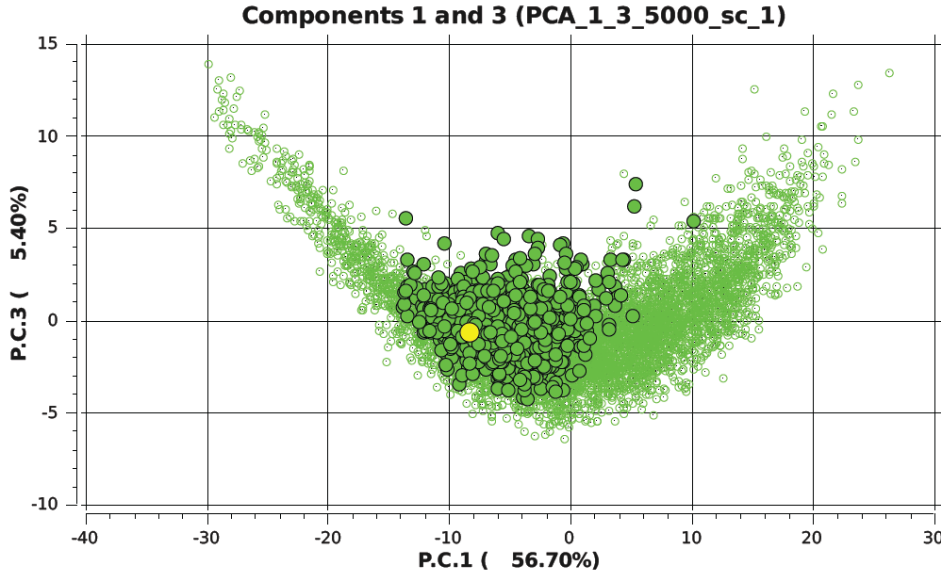
Invaded area	**Number of scenarios**	**Selected scenario**	**Type I error**	**Type II error Mean (min – max)**
East North America	3	Native area	0.10	0.085 (0.05 – 0.12)
West North America	6	Native area	0.11	0.054 (0.01 – 0.13)
Europe	10	Ebc + ENA	0.26	0.008 (0.00 – 0.03)
South America	10	ENA	0.05	0.012 (0.00 – 0.05)
Africa	21	ENA	0.12	0.006 (0.00 – 0.06)

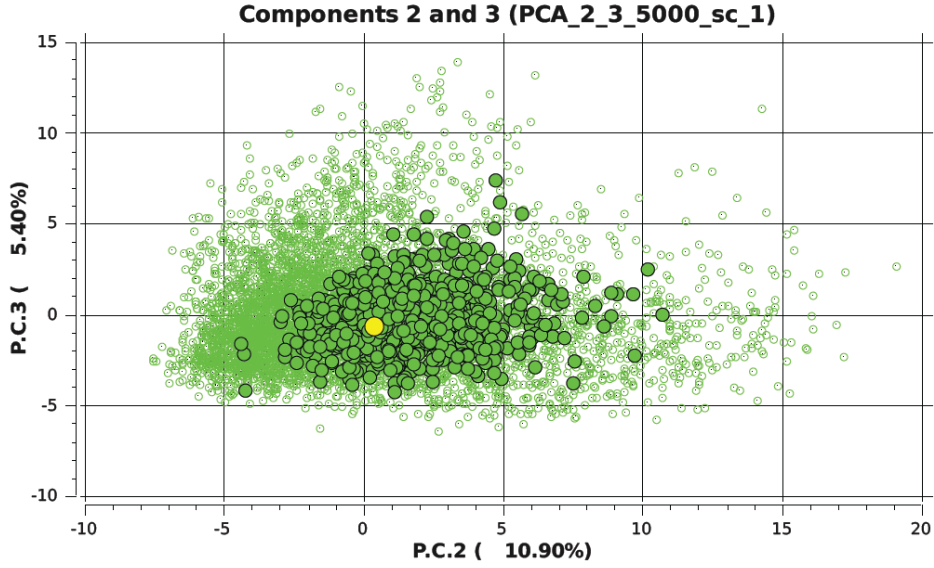
Model checking

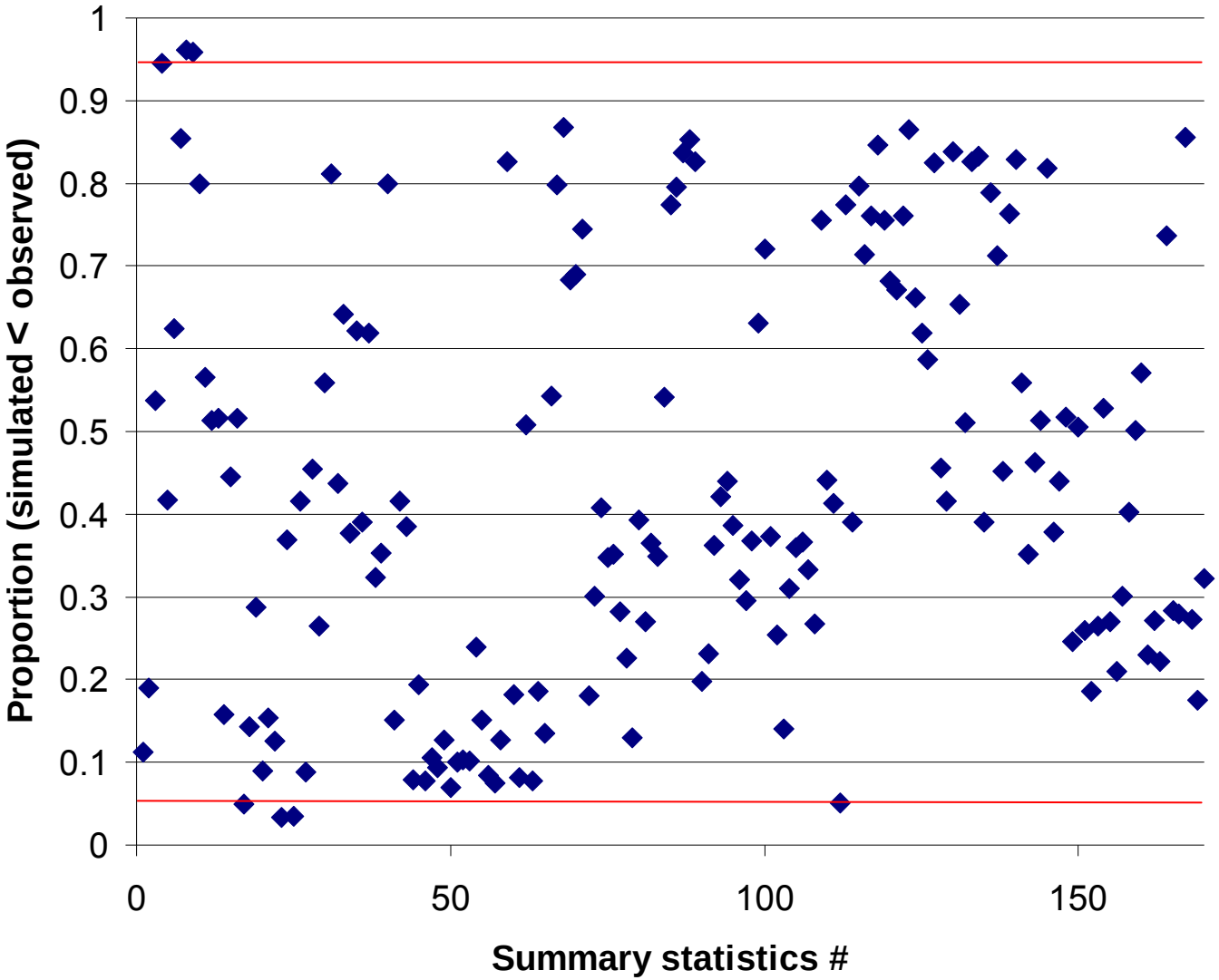
Nsim_ref=1M, Lin_reg=10000, Nsim_pos=10000







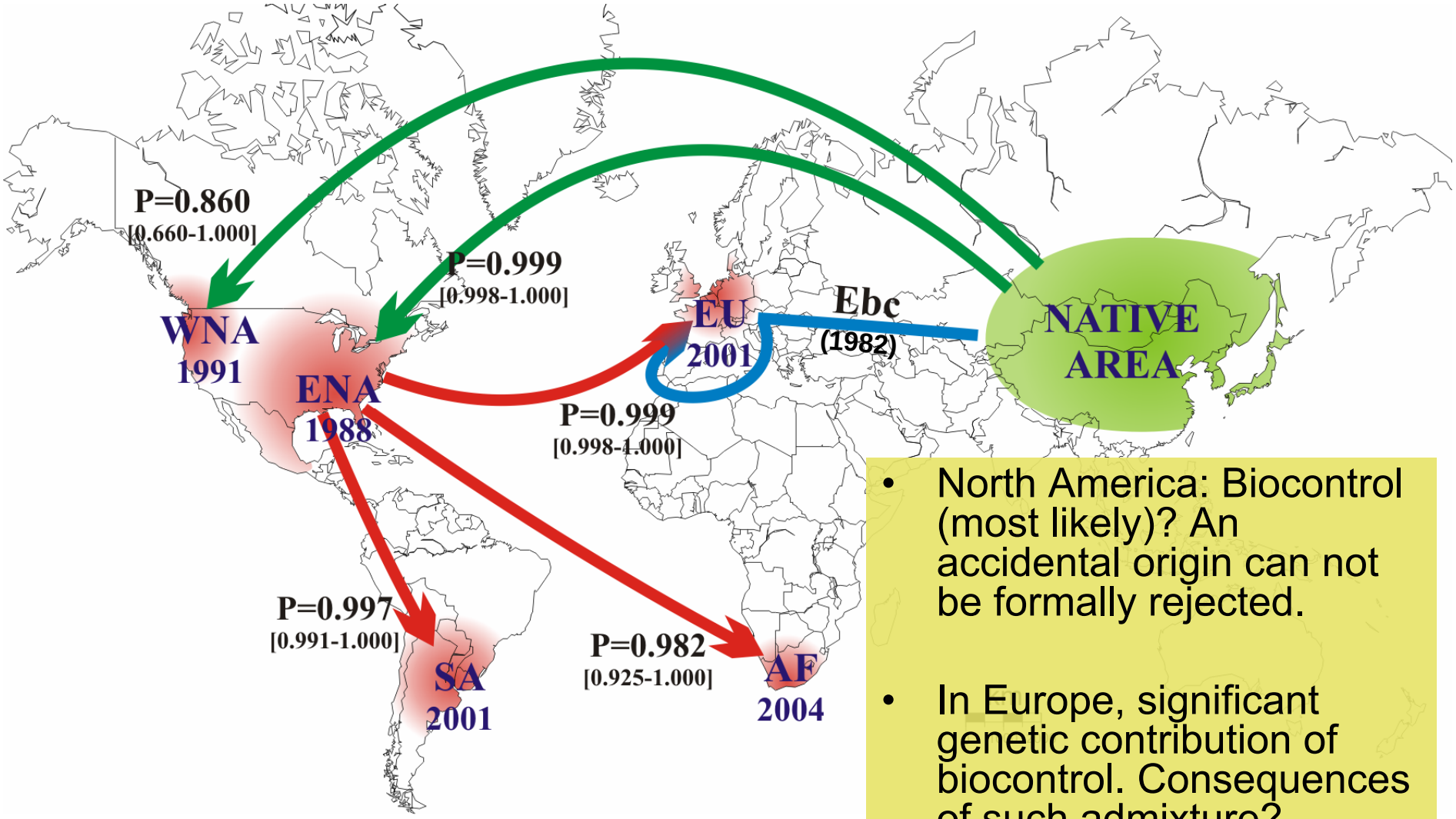




170 Ss: 4.1% within the « 5% tail areas »

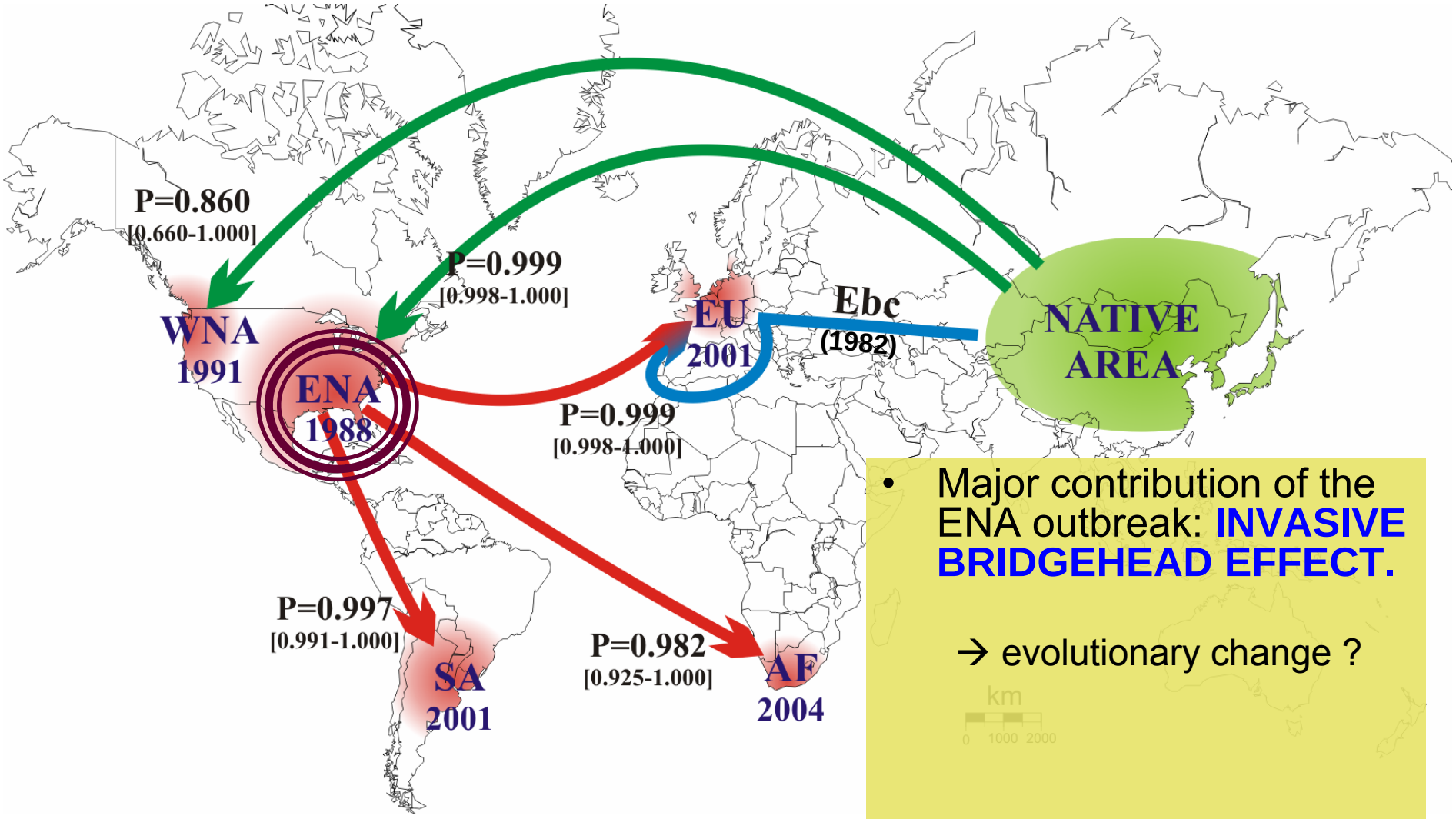
(7 Ss with 1% < p_value < 5 %)

Conclusions



- North America: Biocontrol (most likely)? An accidental origin can not be formally rejected.
- In Europe, significant genetic contribution of biocontrol. Consequences of such admixture?

Conclusions



- Major contribution of the ENA outbreak: **INVASIVE BRIDGEHEAD EFFECT.**

  → evolutionary change ?

Part 5. Inferences about a complex population history – example 2: the case of pygmy populations in Western Africa

Verdu P, Austerlitz F, Estoup A, Vitalis R, Georges M, Théry S, Alain Froment, Lebomin S, Gessain A, Hombert J-M, Van der Veen L, Quintana-Murci L, Bahuchet S, Heyer E (2009) Origins and Genetic Diversity of Pygmy Hunter-Gatherers from Western Central Africa. Current Biology. 19, 312 – 318. http://dx.doi.org/10.1016/j.cub.2008.12.049.

- 604 individuals

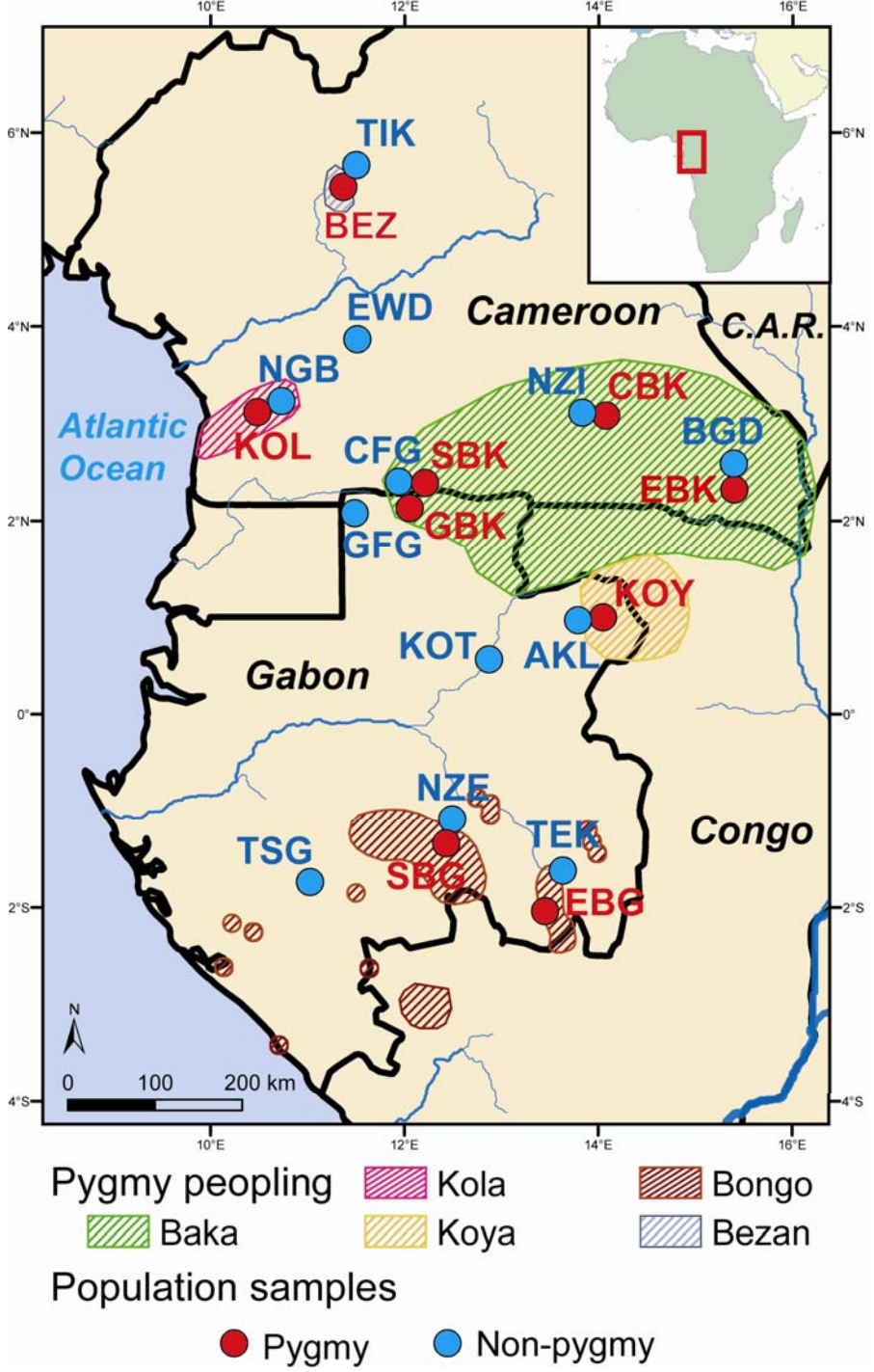
- 12 non pygmy
and 9 neighbouring
pygmy populations

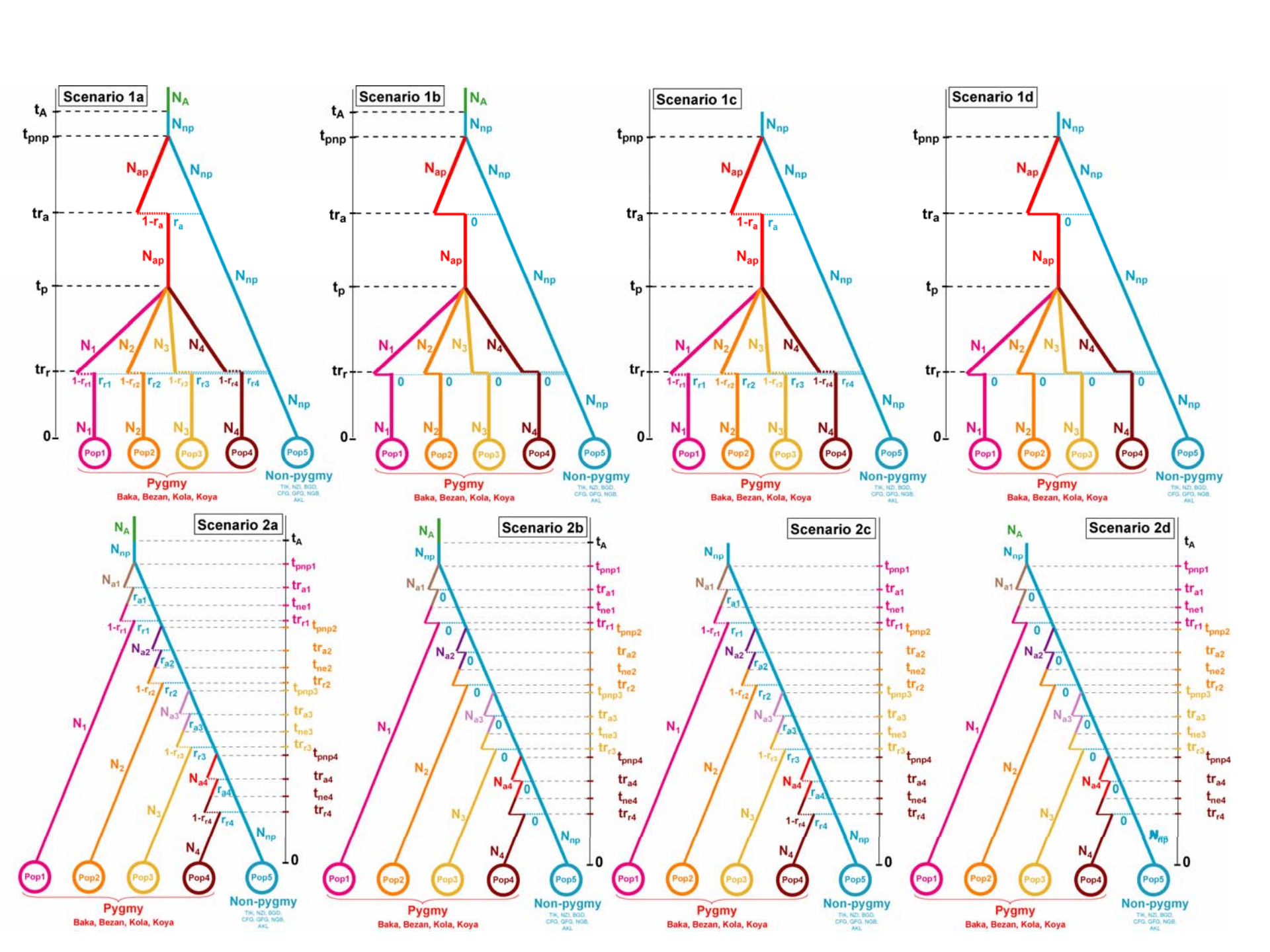
- 28 microsatellite loci

→No genetic structure between non pygmy populations

→Substantial genetic structure between pygmy populations and between pygmy – non pygmy populations

→Substantial socio-culturelles differences between pygmy populations




Scenario 1a
Scenario 1b
Scenario 1c
Scenario 1d
Scenario 2a
Scenario 2b
Scenario 2c
Scenario 2d
Pygmy
Baka, Bezan, Kola, Koya
Non-pygmy
TIK, NZI, BGO, CFG, GFG, NGB, AKL
Pop1
Pop2
Pop3
Pop4
Pop5

35 Summary statistics

Within population genetic variation

Mean NAL: pops 1 2 3 4 5
Mean HET: pops 1 2 3 4 5
Mean VAR: pops 1 2 3 4 5

Between populations genetic variation

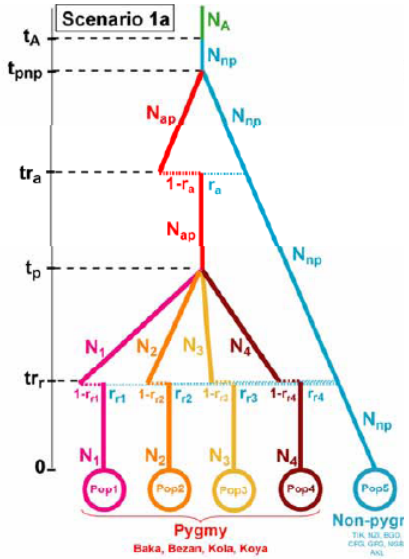
Pairwise FST: 1&2 1&3 1&4 1&5 2&3 2&4 2&5 3&4 3&5 4&5
Pairwise DM2: 1&2 1&3 1&4 1&5 2&3 2&4 2&5 3&4 3&5 4&5

Prior Set 1

Parameters	Conditions	Distribution	Mean	Median	Mode	quantile 2.5%	quantile 97.5%
N_1 (Baka)		Uniform [10 - 10,000]	5,007	5,010	NA	262	9,747
N_2 (Bezan)		Uniform [10 - 10,000]	5,007	5,010	NA	262	9,747
N_3 (Kola)		Uniform [10 - 10,000]	5,007	5,010	NA	262	9,747
N_4 (Koya)		Uniform [10 - 10,000]	5,007	5,010	NA	262	9,747
N_5 (East. Bongo)		Uniform [10 - 10,000]	5,007	5,010	NA	262	9,747
N_6 (South. Bongo)		Uniform [10 - 10,000]	5,007	5,010	NA	262	9,747
N_{np} (Non-pygmies)		Uniform [10 - 100,000]	50,100	50,040	NA	2529	97,489
N_{Ap}		Uniform [10 - 10,000]	5,007	5,010	NA	262	9,747
N_A		Uniform [10 - 10,000]	5,007	5,010	NA	262	9,747
tr_r	$tr_r < t_p$	Loguniform [1 - 5,000]	187	29	1	1	1,412
t_p	$tr_r < t_p$	Uniform [1 - 5,000]	1,389	1,201	391	82	3,635
tr_a	$t_p < tr_a$	Uniform [1 - 5,000]	2,592	2,605	2,690	560	4,554
t_{pnp}	$tr_a < t_{pnp}$	Uniform [1 - 5,000]	3,796	4,013	4,850	1,565	4,960
t_A		Uniform [1 - 10,000]	4,999	5,004	NA	252	9,748
r_{r1}		Uniform [0 - 1]	0.5	0.5	NA	0.0248	0.975
r_{r2}		Uniform [0 - 1]	0.5	0.5	NA	0.0248	0.975
r_{r3}		Uniform [0 - 1]	0.5	0.5	NA	0.0248	0.975
r_{r4}		Uniform [0 - 1]	0.5	0.5	NA	0.0248	0.975
r_{r5}		Uniform [0 - 1]	0.5	0.5	NA	0.0248	0.975
r_{r6}		Uniform [0 - 1]	0.5	0.5	NA	0.0248	0.975
r_a		Uniform [0 - 1]	0.5	0.5	NA	0.0248	0.975
$\overline{\mu}$		Uniform [10^{-4} - 10^{-3}]	5.5×10^{-4}	5.5×10^{-4}	NA	1.2×10^{-4}	9.8×10^{-4}
$\overline{p}$		Uniform [0.1 – 0.3]	0.20	0.20	NA	0.11	0.30

→ 500,000 simulations per scenario (total: 4 M)

Posterior probabilities for each scenario using the logistic regression method



Historical Scenario	Prior Set 1: 5,000 closest simulations (0.125%)	Prior Set 1: 50,000 closest simulations (1.25%)
Scenario 1a	0.9604 [0.9072 - 1.0000]	0.8806 [0.8518 - 0.9093]
Scenario 1b	0.0373 [0.0000 - 0.0906]	0.0994 [0.0703 - 0.1285]
Scenario 1c	0.0018 [0.0000 - 0.0036]	0.0142 [0.0111 - 0.0172]
Scenario 1d	0.0000 [0.0000 - 0.0000]	0.0010 [0.0000 - 0.0022]
Scenario 2a	0.0008 [0.0002 - 0.0009]	0.0049 [0.0041 - 0.0056]
Scenario 2b	0.0000 [0.0000 - 0.0000]	0.0000 [0.0000 - 0.0000]
Scenario 2c	0.0000 [0.0000 - 0.0000]	0.0000 [0.0000 - 0.0001]
Scenario 2d	0.0000 [0.0000 - 0.0000]	0.0000 [0.0000 - 0.0000]

Confidence in scenario choice

100 simulated test datasets for each scenario (parameter values drawn into priors)

→ focal scenario = 1a

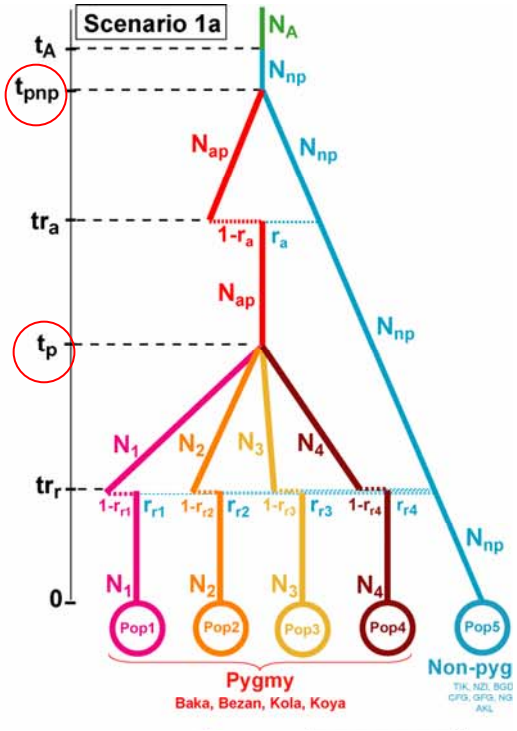
→ Logistic regression

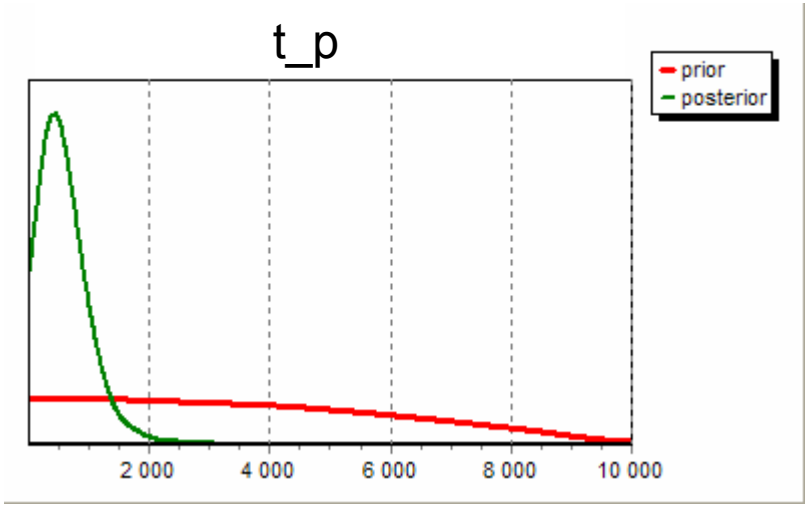
- Type I error rate = 0.26

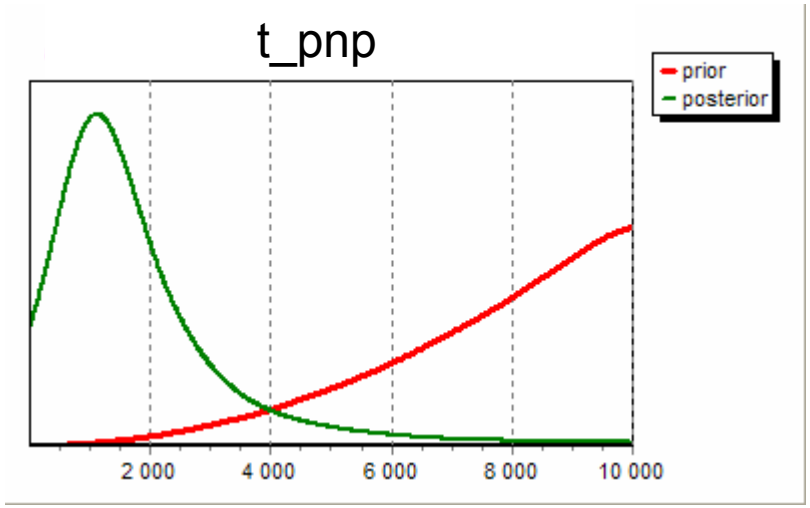
- Type II error rates: mean = 0.046 [min=0.00; max=0.09]

Estimation of parameters under scenario 1a (1/2)

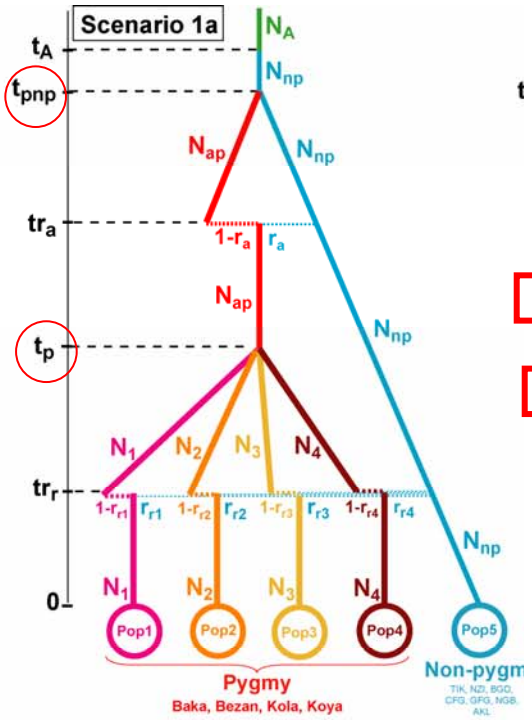
(local linear regression: 1% of best simulations)







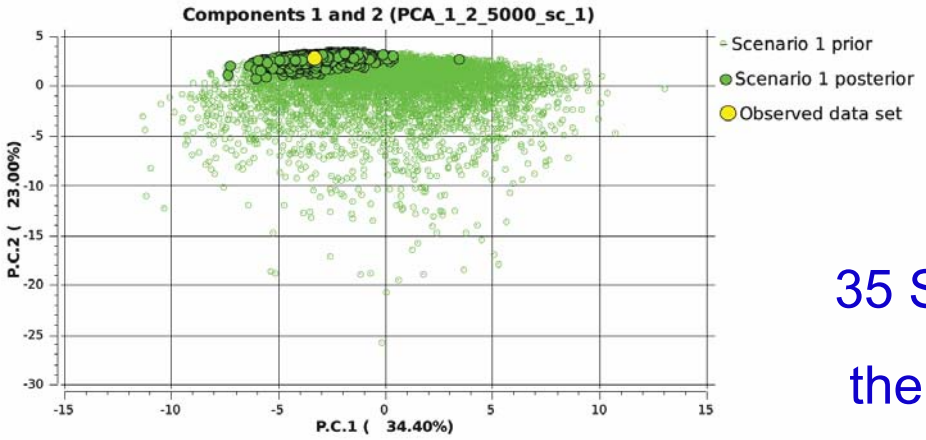
Estimation of parameters under scenario 1a (2/2)



Parameter	mean	median	mode	quantile 2.5%	quantile 97.5%
Original Parameters					
N_1 (Baka)	6,164	6,368	8,137	1,347	9,824
N_2 (Bezan)	5,055	4,840	2,795	790	9,677
N_3 (Kola)	4,486	4,100	3,302	603	9,599
N_4 (Koya)	5,608	5,619	3,197	1,134	9,771
N_{np} (Non-pygmies)	66,265	67,168	77,157	27,926	97,828
N_{ap}	5,901	6,163	8,007	960	9,825
N_A	3,074	2,631	1,071	202	8,404
tr_r	115	67	8	4	485
t_p	364	256	105	29	1,371
tr_a	1,353	1118	771	212	3,749
t_{pnp}	3,101	3170	3,587	921	4,913
t_A	4,217	3,740	2,802	663	9,419
r_{r1}	0.662	0.674	0.696	0.261	0.957
r_{r2}	0.461	0.440	0.416	0.098	0.899
r_{r3}	0.647	0.662	0.672	0.219	0.955
r_{r4}	0.523	0.514	0.465	0.147	0.920
r_a	0.572	0.605	0.927	0.041	0.982
$\overline{\pi}$	0.00024	0.00021	0.00016	0.00011	0.00056
$\overline{p}$	0.11	0.11	0.10	0.10	0.16

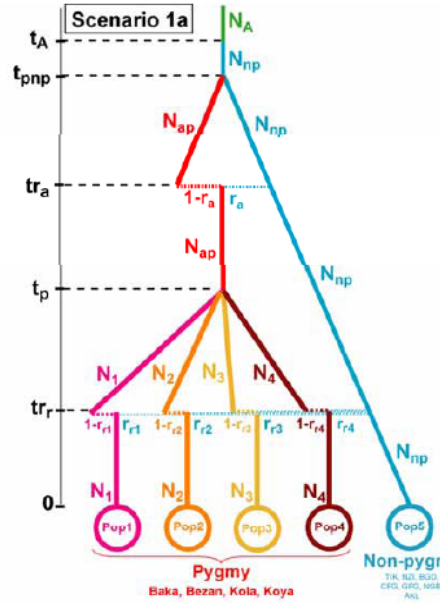
Model checking : stats used for previous inferences

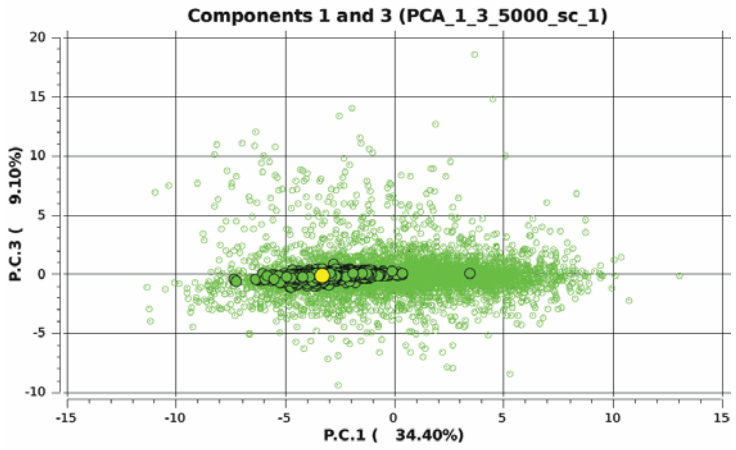
Nsim_ref=1M, Lin_reg=10000, Nsim_pos=10000

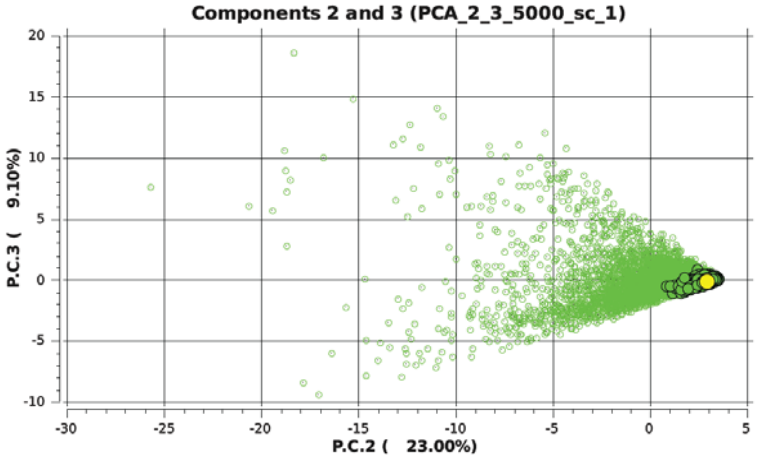


35 Ss: none within

the « tail areas »

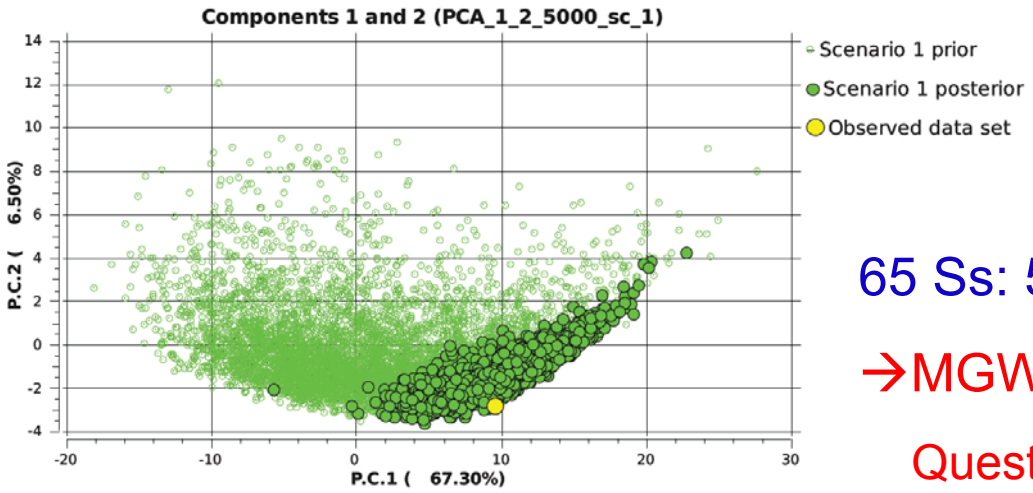






Model checking : stats NOT used for previous inferences

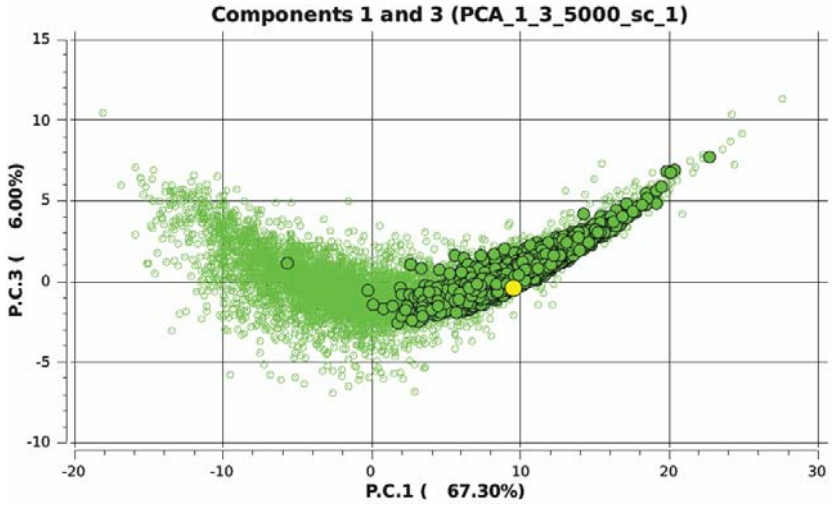
Nsim_ref=1M, Lin_reg=10000, Nsim_pos=10000

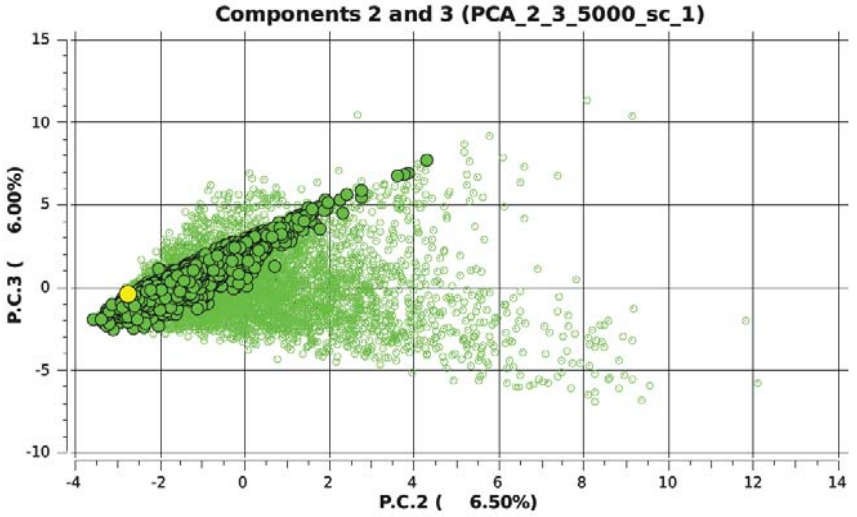


65 Ss: 5 Ss within the « 5% tail areas »

→MGW_obs >> MGW_sim

Question: pbl with demo or mutation model ?

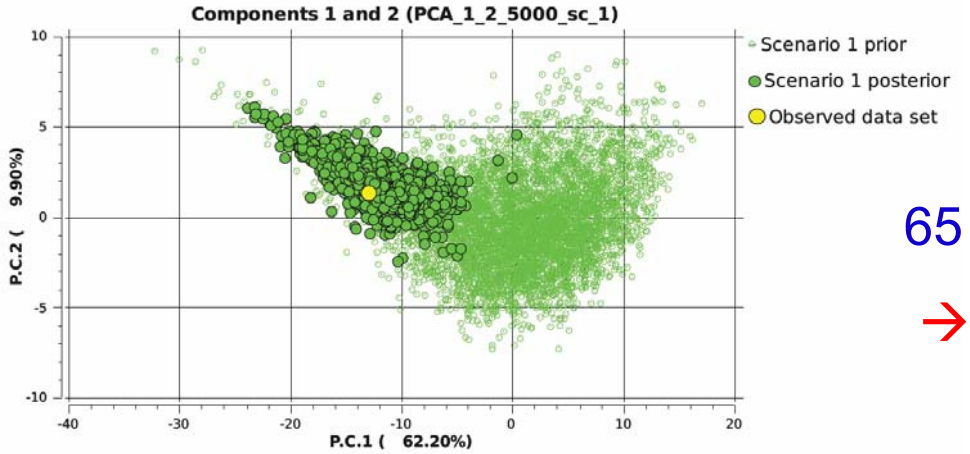




NEW model checking from new simulations assuming a (slightly) different mutation model = possibility of a higher rate of indel mutations

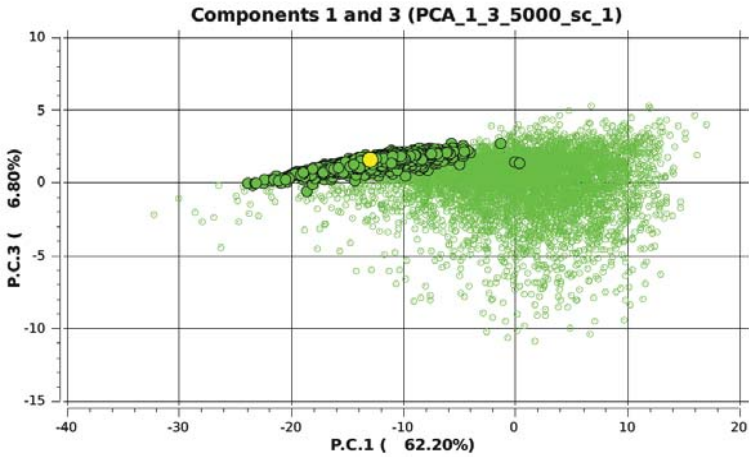
→ Stats NOT used for previous inferences

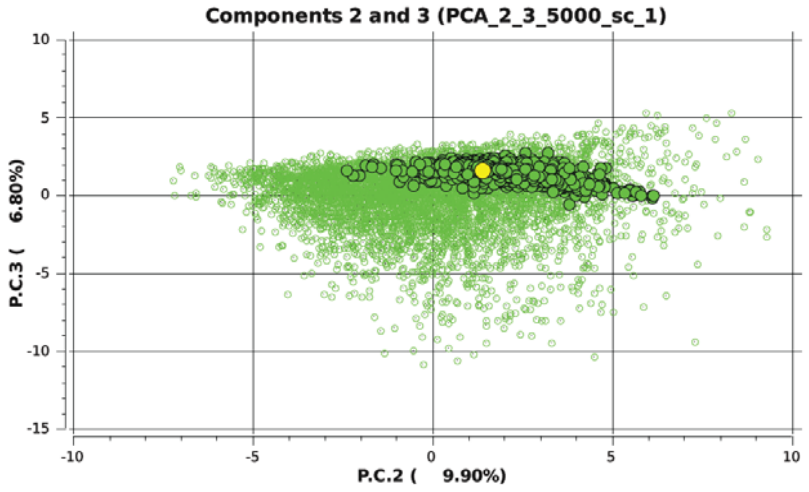
Nsim_ref=1M, Lin_reg=10000, Nsim_pos=10000

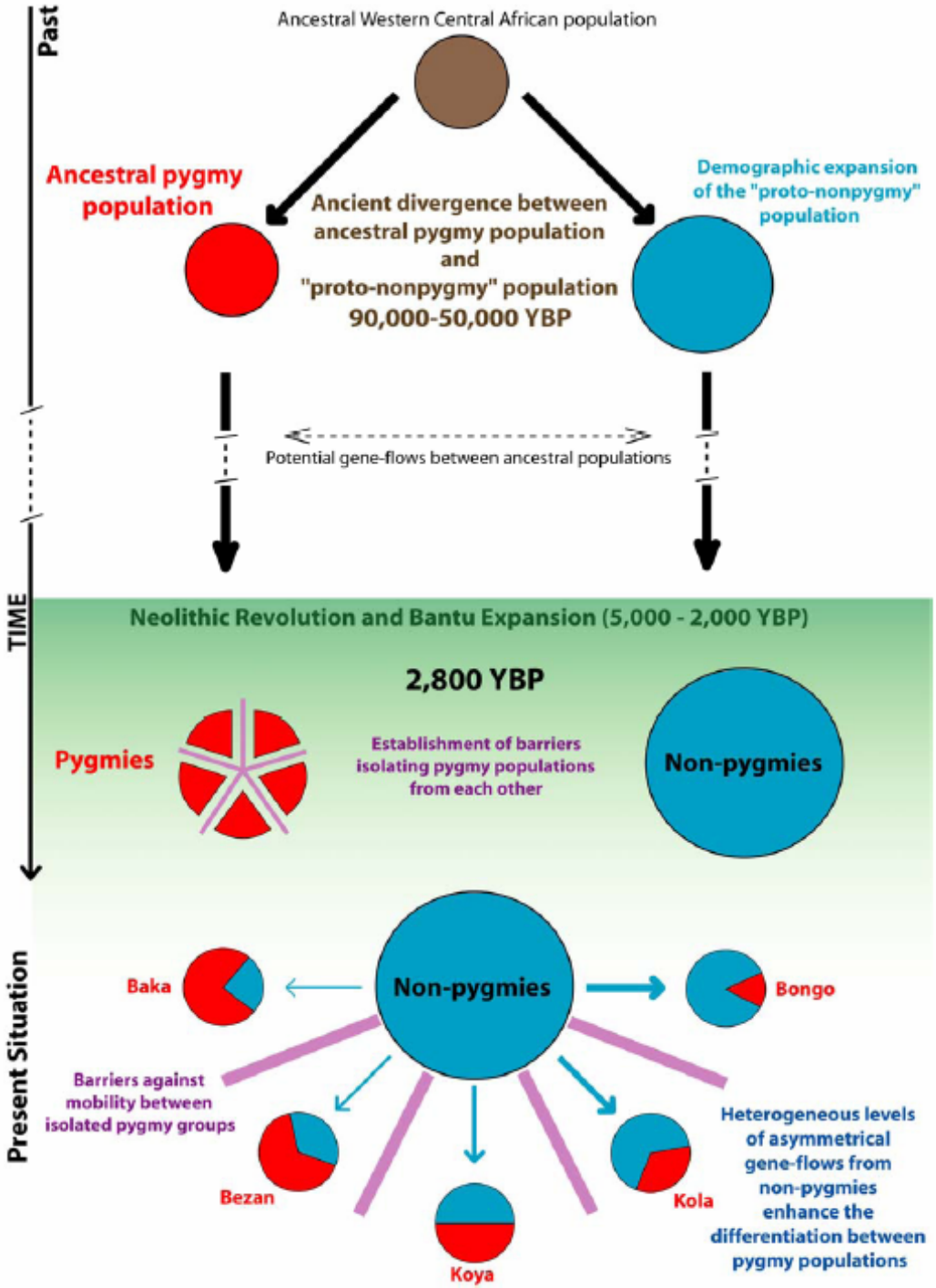


65 Ss: none within the « 5% tail areas »

→ all MGW stats with p_values > 0.10






Past
Ancestral Western Central African population
Ancestral pygmy population
Ancient divergence between ancestral pygmy population and "proto-nonpygmy" population 90,000-50,000 YBP
Demographic expansion of the "proto-nonpygmy" population
Potential gene-flows between ancestral populations
TIME
Neolithic Revolution and Bantu Expansion (5,000 - 2,000 YBP)
2,800 YBP
Pygmies
Establishment of barriers isolating pygmy populations from each other
Non-pygmies
Present Situation
Baka
Non-pygmies
Bongo
Barriers against mobility between isolated pygmy groups
Bezan
Koya
Kola
Heterogeneous levels of asymmetrical gene-flows from non-pygmies enhance the differentiation between pygmy populations

Part 6.
General conclusions

- ABC allows making inferences on complex problems → good option when using a likelihood-based approach is not possible

- More and more accessible to biologists because of the constant development of (more or less) simple/integrated software solutions

- Not as easy as it seems to be
  → needs some practice and reading (formalization of scenarios, prior distributions,…)
  → needs some "validation(s)" through simulations (possible with some softwares; e.g. DIYABC)

- *Does not replace but rather complement* more "traditional" statistical approaches: raw statistics (Fst, heterozygosity,…), NJ-tree, Clustering Bayesian methods (CBM) developed to address questions related to genetic structure (STRUCTURE, BAPS, GENELAND, TESS, GESTE…)
  → CBM = minimize the number of "population units" and hence reduce the number of possible population topologies to compare with ABC

THANK YOU FOR YOUR ATTENTION

Acknowledgements to:

- Jean-Marie Cornuet
- Mark Beaumont
- Julien Veyssier
- Alex Dehne Garcia
- Pierre Pudlo
- Jean-Michel Marin
- Christian Robert
- Laurent Excoffier
- Michael Blum
- David Balding
- Virginie Ravigné
- Eric Lombaert
- Thomas Guillemaud
- Paul Verdu
- *…and many others*

Some key references (non-exhaustive list)

Inferences using « standard ABC »

Pritchard JK, Seielstad MT, Prez-Lezaum A, Feldman MW (1999) Population growth of human Y chromosomes, a study of Y chromosome microsatellites. *Molecular Biology and Evolution* 16, 1791-1798.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* 162, 2025-2035.

Verdu P, Austerlitz F, Estoup A, Vitalis R, Georges M, Thery S, Froment A, Le Bomin S, Gessain A, Hombert JM*, et al.* (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current Biology*, 19, 312-318.

E. Lombaert, T. Guillemaud, J.M. Cornuet, T. Malausa, B. Facon, A. Estoup. 2010. Bridgehead effect in the worldwide invasion of the biocontrol harlequin ladybird. *Plos One*, 5(3) e9743.

*DIYABC (*http://www1.montpellier.inra.fr/CBGP/diyabc)

Cornuet J-M, Santos F, Robert PC, Marin J-M, Balding DJ, Guillemaud T, Estoup A (2008) Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation. *Bioinformatics*, 24, 2713-2719.

Cornuet JM, Ravigné V, Estoup A (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC bioinformatics*, 11, 401doi:10.1186/1471-2105-11-401.

Reviews

Csilléry K, Blum MGB, Gaggiotti O, François O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, 25, 411-417.

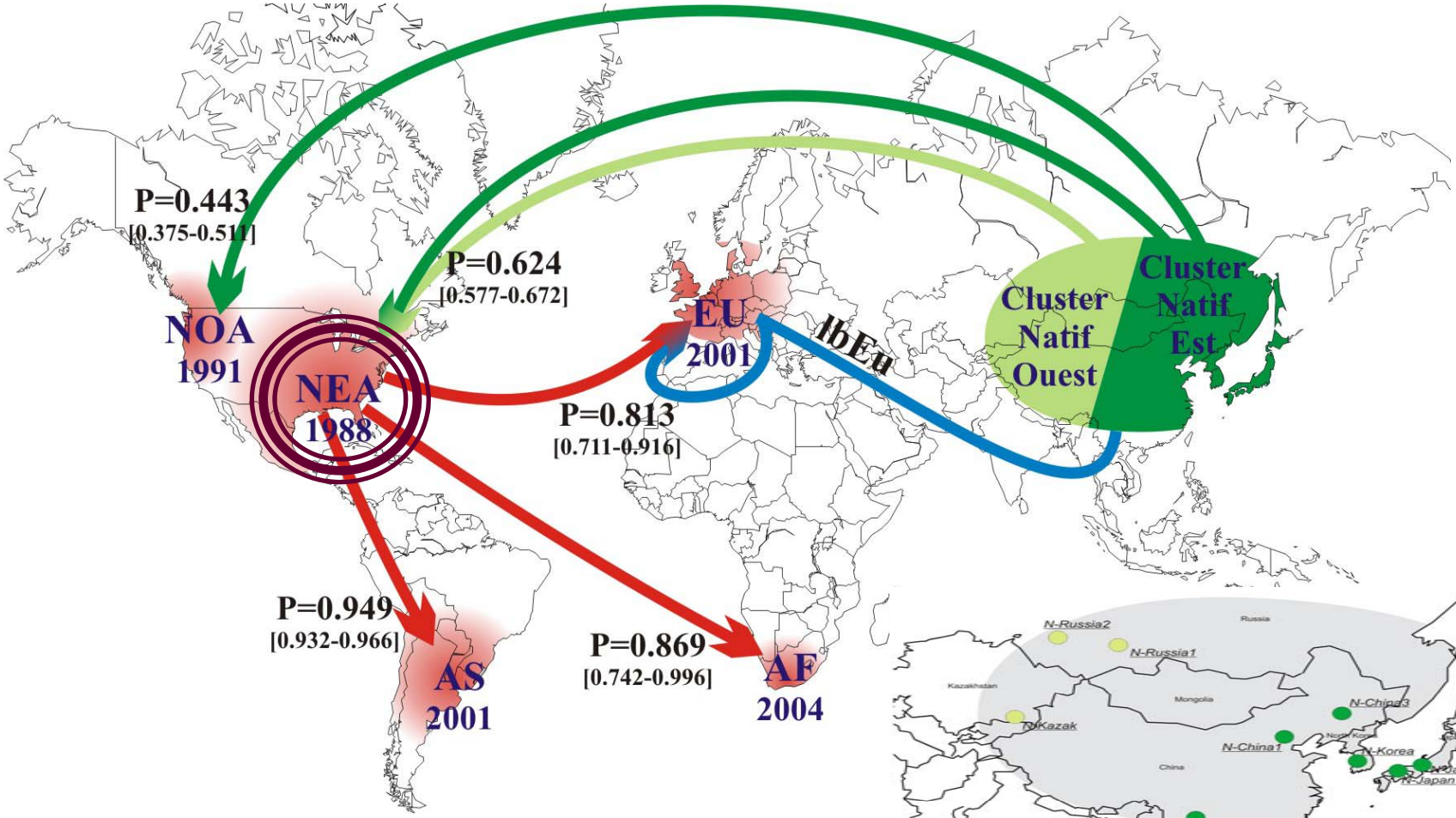
Bertorelle G, Bonazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology*, 19, 2609-2625.

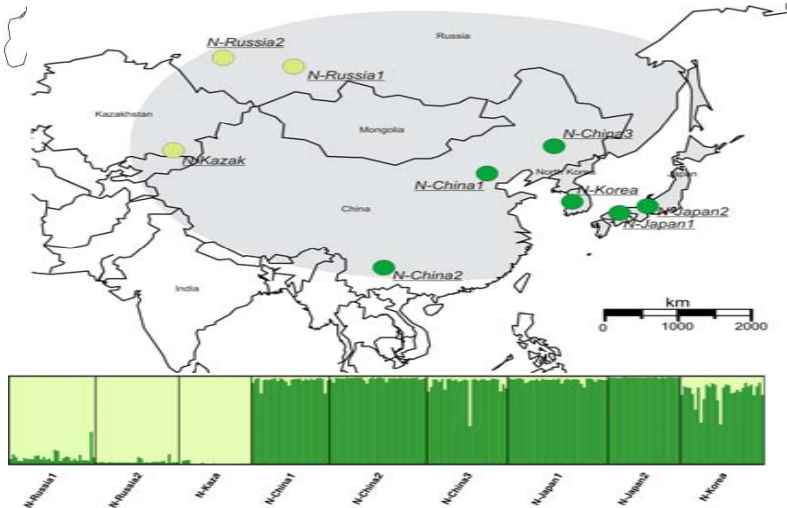
Beaumont M (2010) Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology and Evolutionary Systematics*, 41, 379-406.

Some more results…

A slightly different introduction history for *H. axyridis*: genetic admixture (also) during the first introduction event (cf. bridgehead population)

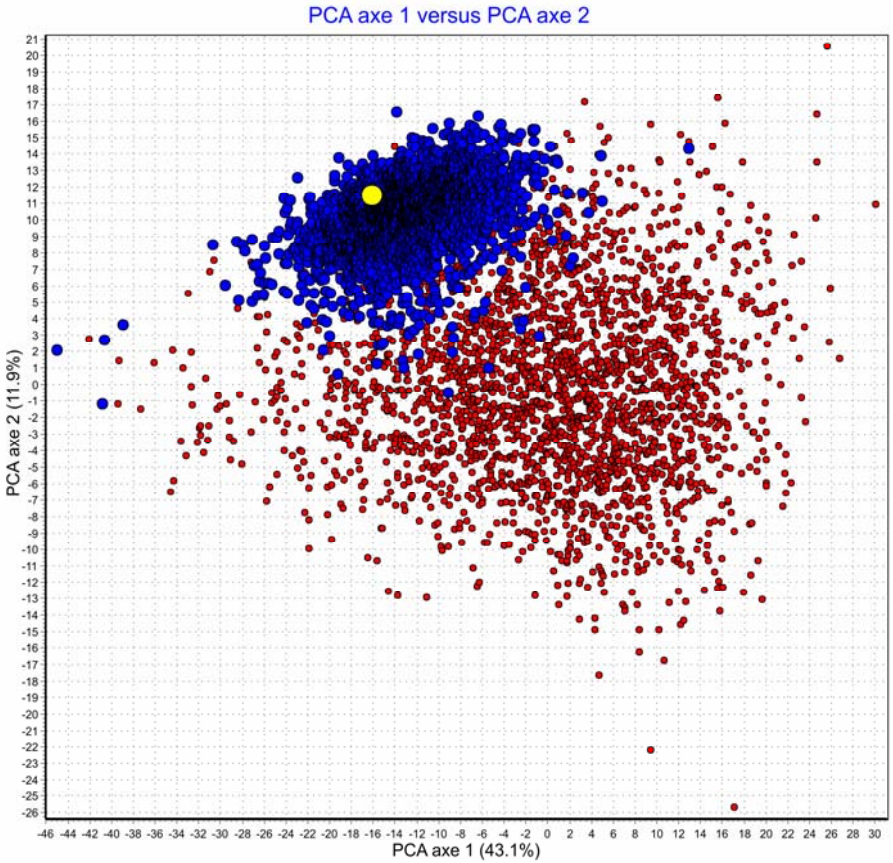
Lombaert et al. (2011) *Molecular Ecology* 20, 4654-4670.

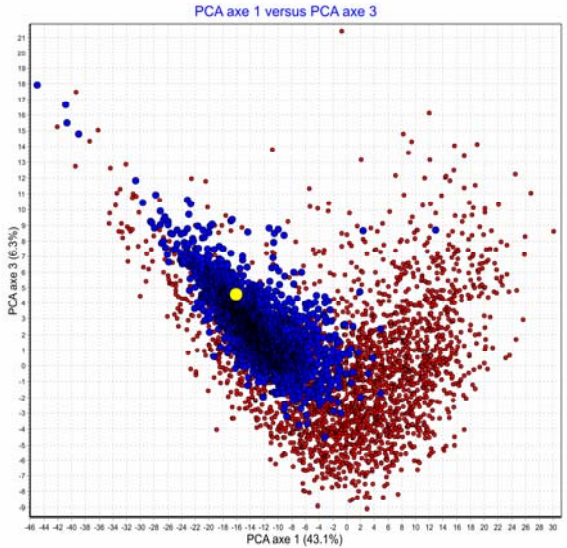


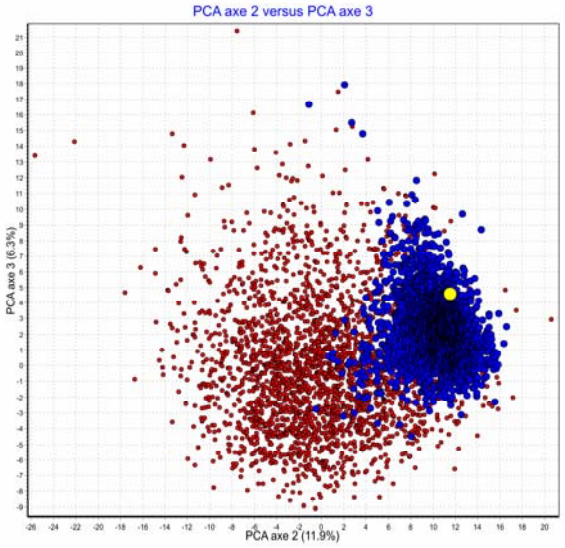


« Better fit »

223 Ss: 2.2 % within the « tail areas »


PCA axe 1 versus PCA axe 2
PCA axe 2 (11.9%)
PCA axe 1 (43.1%)


PCA axe 1 versus PCA axe 3
PCA axe 3 (6.3%)
PCA axe 1 (43.1%)


PCA axe 2 versus PCA axe 3
PCA axe 3 (6.3%)
PCA axe 2 (11.9%)