From theory to application: DIYABC, a user-friendly program to infer complex population histories using Approximate Bayesian Computation

Cornuet J-M, Santos F, Robert CP, 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.

The software DIYABC and the companion paper are both freely available at http://www1.montpellier.inra.fr/CBGP/diyabc.

ABC day in Paris – 26/06/09

General context

- Likelihood + MCMC (+ IS) → difficult for complex situations.

- Approximate Bayesian Computation (e.g. Beaumont et al. 2002) allows to make inferences on complex problems.

- In its current state, the ABC approach remains inaccessible to most biologists because there is not yet a simple software solution.

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

➢ Only microsatellite data, no gene flow between populations

Program:

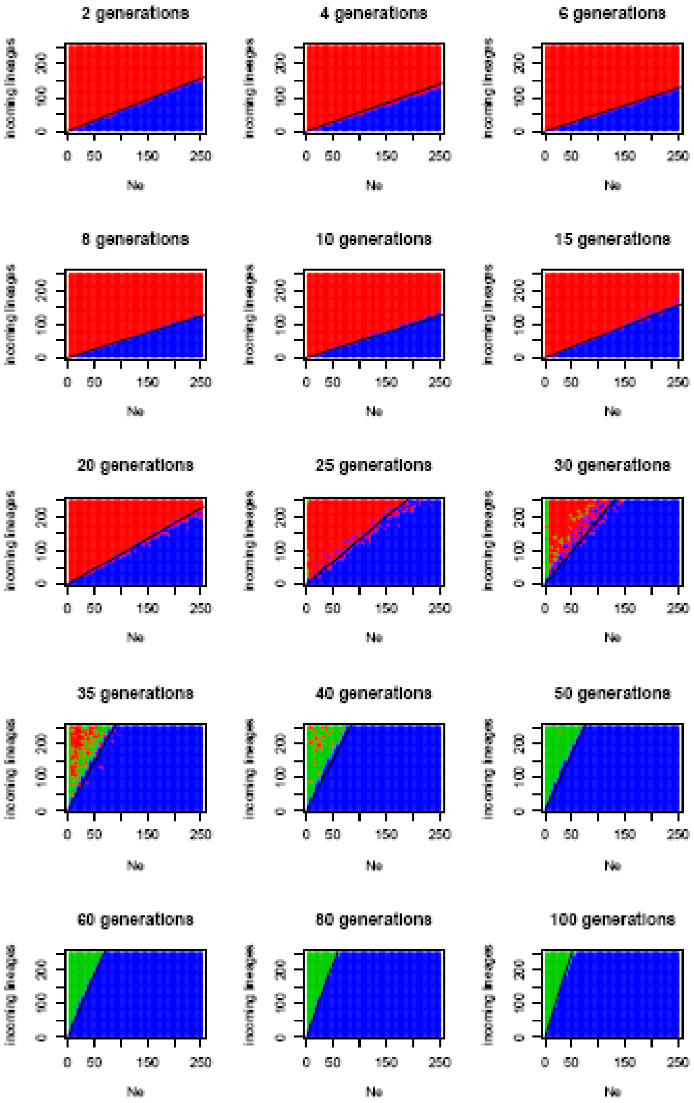
- written in Delphi
- running under a 32-bit Windows operating system (e.g. Windows XP)
- multi-processor
- user-friendly graphical interface

Two coalescence algorithms:

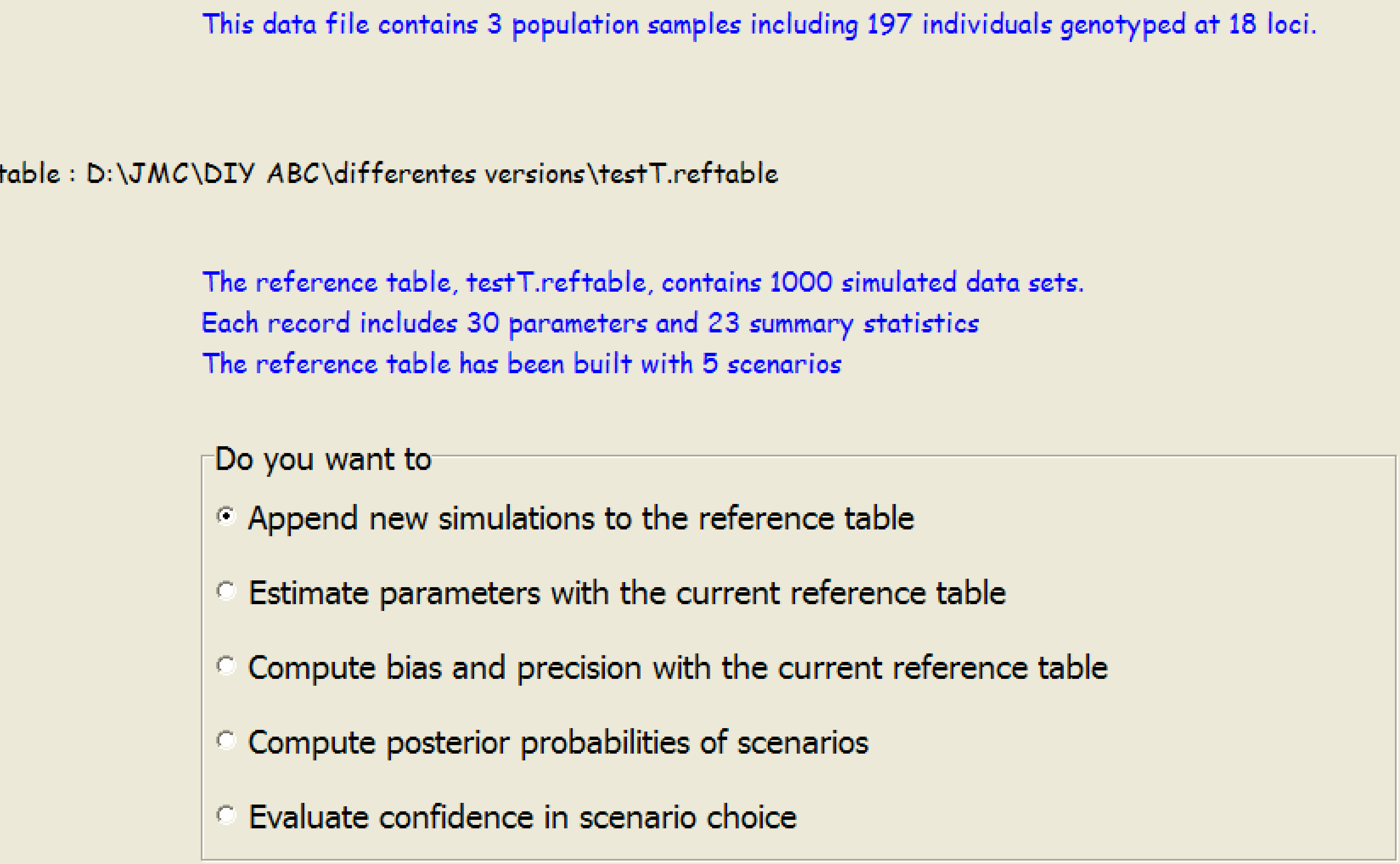
- Continuous time (CT)
- Generation by generation (GbG)

Rule optimizing computation speed and limiting bias in coalescence rate

if $(1 < g \leq 30)$ do CT if $n_{el}/N_e < 0.0031g^2 - 0.053g + 0.7197$ else do GbG
if $(30 < g \leq 100)$ do CT if $n_{el}/N_e < 0.033g + 1.7$ else do GbG
if $(100 < g)$ do CT if $n_{el}/N_e < 5$ else do GbG



Graphs indicate in green the area of the plane for which the generation by generation (GbG) algorithm is faster than the continuous time (CT) algorithm, in red the area for which the CT algorithm produces significantly (5%) less coalescences than the GbG algorithm and in blue the area for which the CT algorithm produces the same number of coalescences than the GbG algorithm (with tolerance=5%) and is faster. Limits between areas are almost linear. The black line (intercept=0) has a slope taken as $0.0031g^2 - 0.053g + 0.7197$ for $g \leq 30$, $0.033g + 1.7$ for $30 < g \leq 100$ and 5 when $100 < g$, g being the duration of the coalescence module in number of generations. N_e is the diploid effective population size.



This data file contains 3 population samples including 197 individuals genotyped at 18 loci.

table : D:\JMC\DIY ABC\differentes versions\testT.reftable

The reference table, testT.reftable, contains 1000 simulated data sets.
Each record includes 30 parameters and 23 summary statistics
The reference table has been built with 5 scenarios

Do you want to

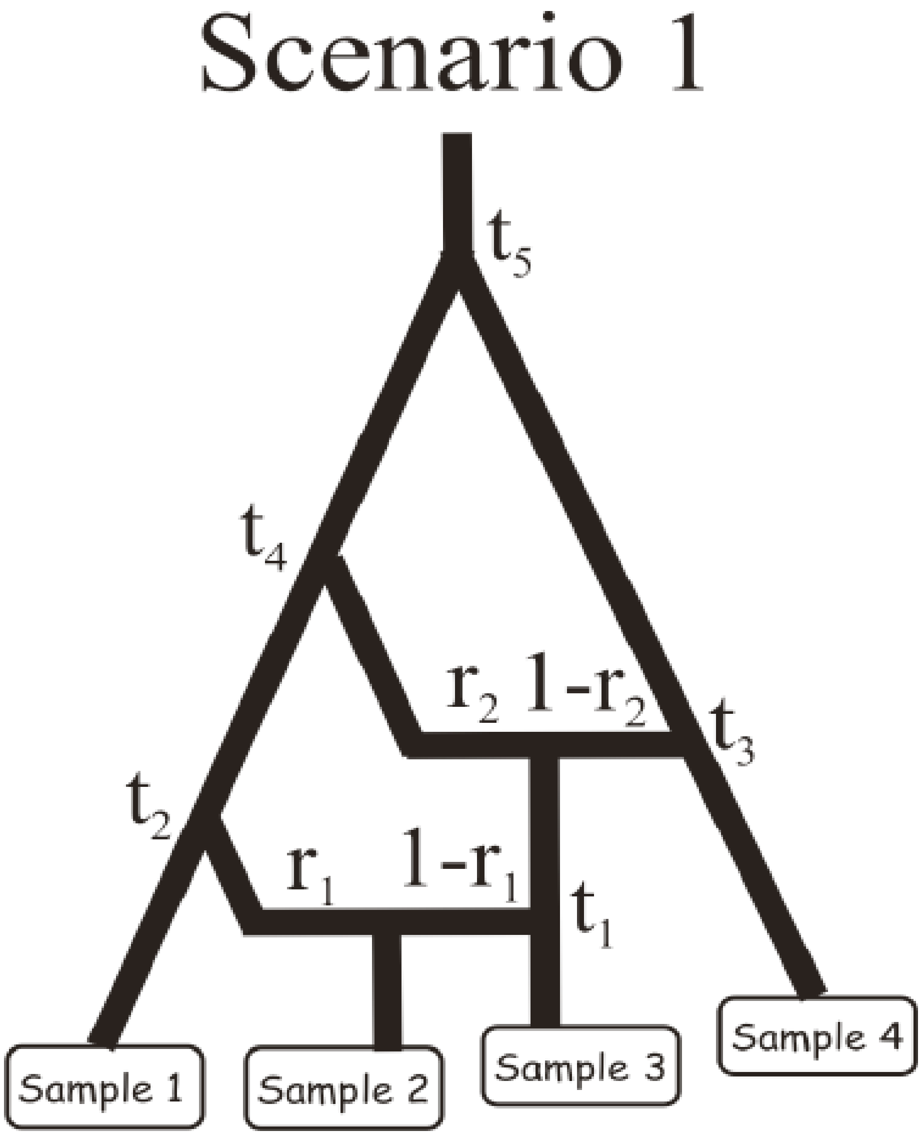
● Append new simulations to the reference table

○ Estimate parameters with the current reference table

○ Compute bias and precision with the current reference table

○ Compute posterior probabilities of scenarios

○ Evaluate confidence in scenario choice

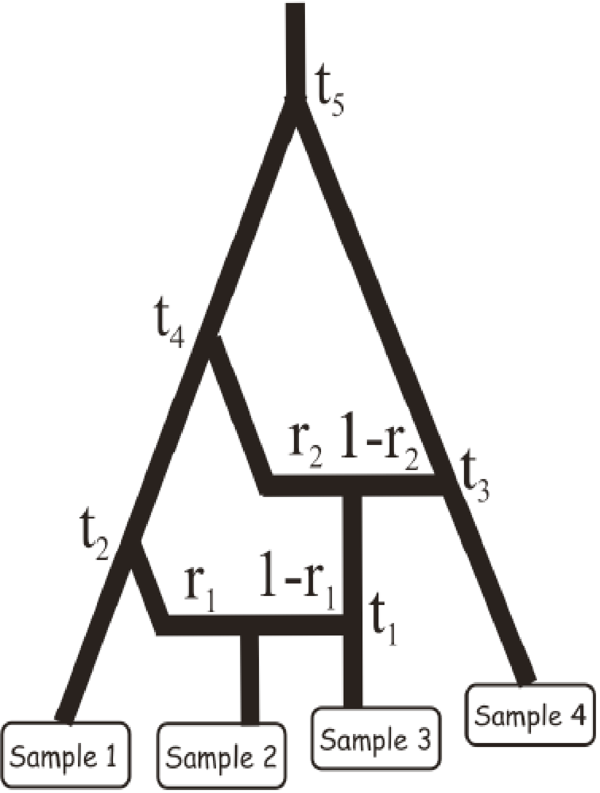


ONE simulated test data set

- 10 loci
- 30 diploid ind / sample

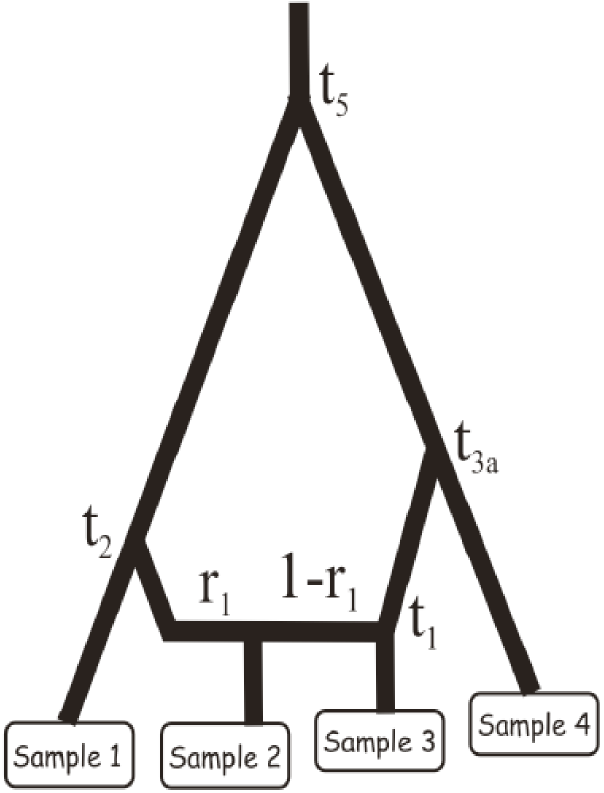
Parameter
N (1,000)
r_1 (0,6)
r_2 (0,4)
t_1 (10)
t_2 (500)
t_3 (10,000)
t_4 (20,000)
t_5 (200,000)
$\bar{\mu}$ (0.0005)
P (0.22)

Scenario 1



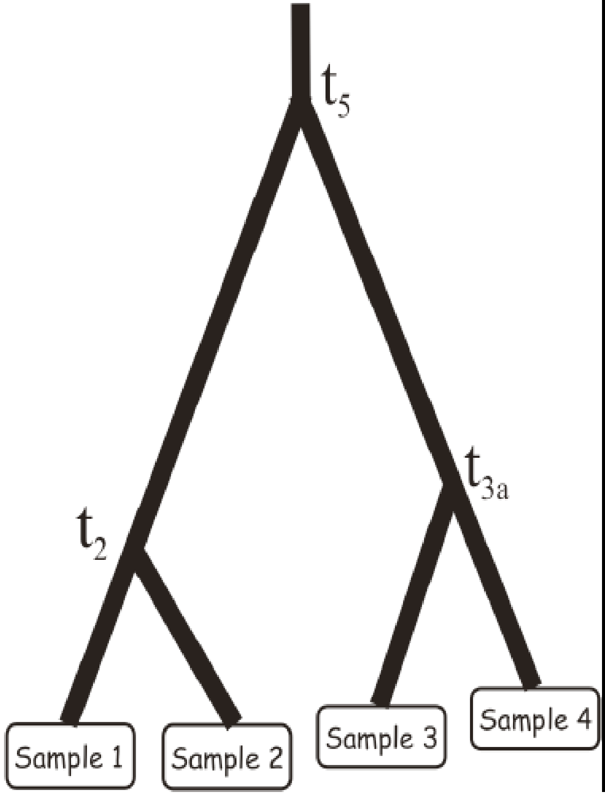
2 admixture events

Scenario 2

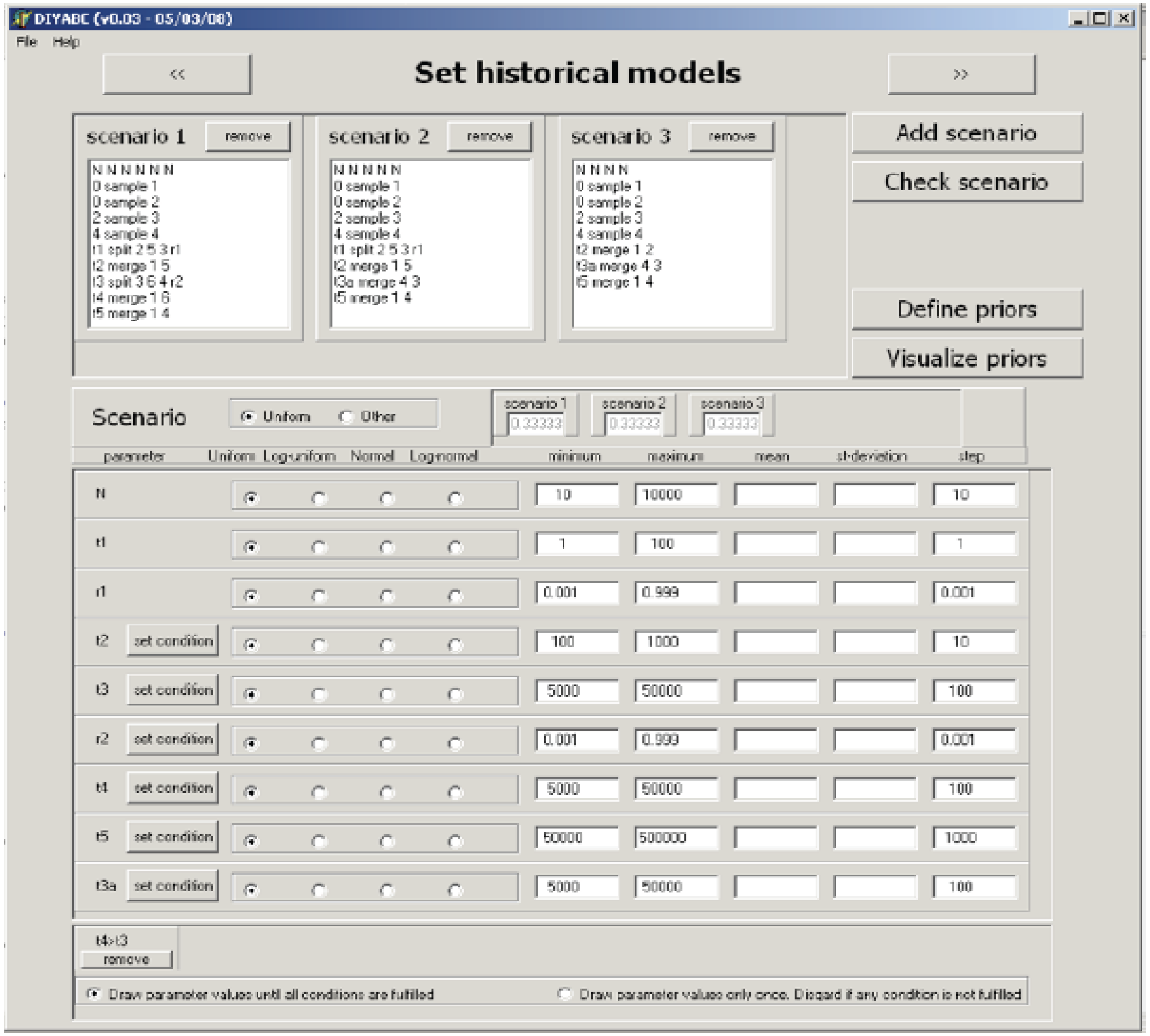


1 admixture event

Scenario 3



0 admixture events



Set historical models

scenario 1

```
N N N N N N
0 sample 1
0 sample 2
2 sample 3
4 sample 4
t1 split 2 5 3 r1
t2 merge 1 5
t3 split 3 6 4 r2
t4 merge 1 6
t5 merge 1 4
```

scenario 2

```
N N N N N
0 sample 1
0 sample 2
2 sample 3
4 sample 4
t1 split 2 5 3 r1
t2 merge 1 5
t3a merge 4 3
t5 merge 1 4
```

scenario 3

```
N N N N
0 sample 1
0 sample 2
2 sample 3
4 sample 4
t2 merge 1 2
t3a merge 4 3
t5 merge 1 4
```

Scenario ● Uniform ○ Other

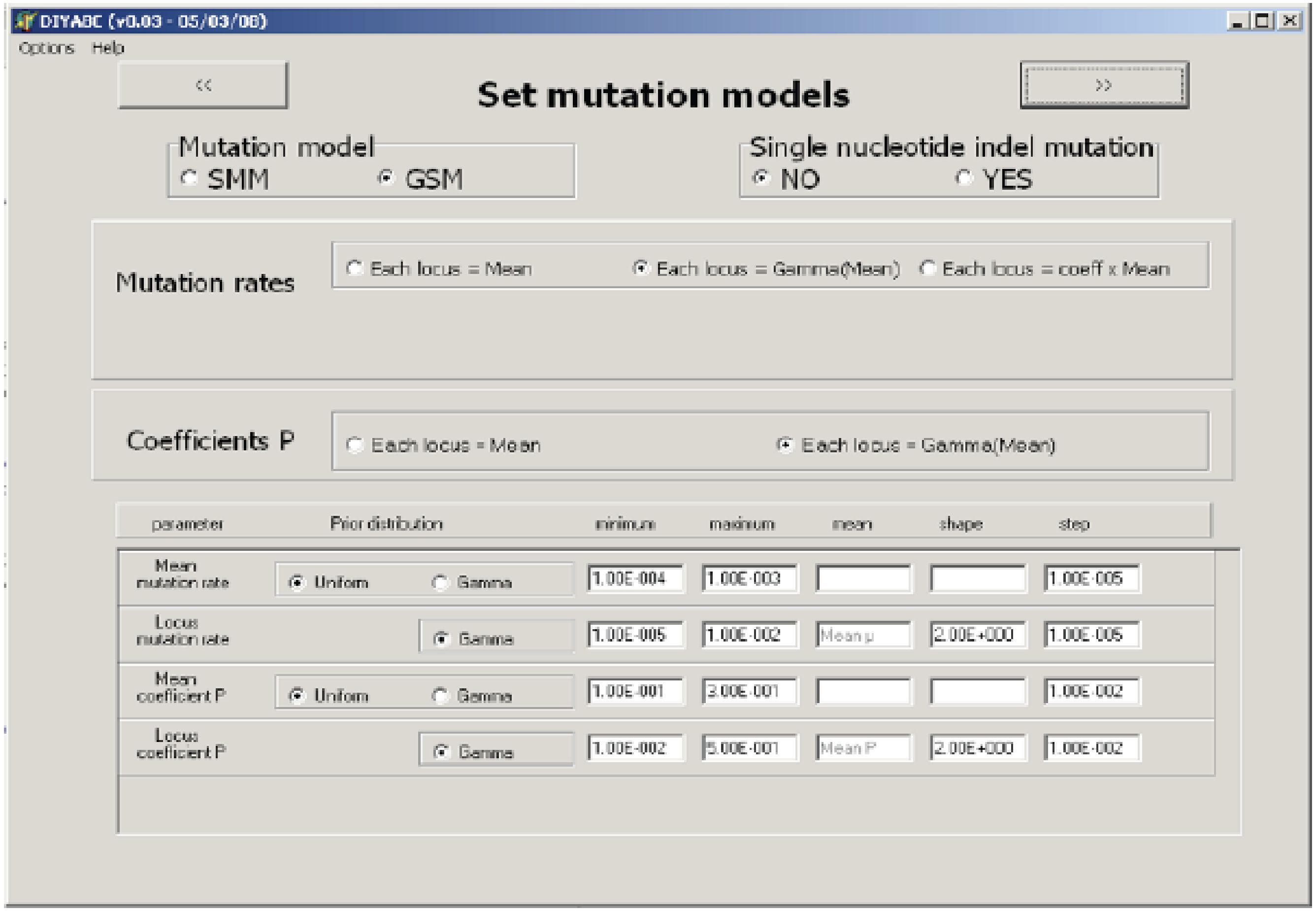
scenario 1	scenario 2	scenario 3
0.33333	0.33333	0.33333

parameter		Uniform	Log-uniform	Normal	Log-normal	minimum	maximum	mean	st-deviation	step
N		●	○	○	○	10	10000			10
t1		●	○	○	○	1	100			1
r1		●	○	○	○	0.001	0.999			0.001
t2	set condition	●	○	○	○	100	1000			10
t3	set condition	●	○	○	○	5000	50000			100
r2	set condition	●	○	○	○	0.001	0.999			0.001
t4	set condition	●	○	○	○	5000	50000			100
t5	set condition	●	○	○	○	50000	500000			1000
t3a	set condition	●	○	○	○	5000	50000			100

t4>t3

● Draw parameter values until all conditions are fulfilled

○ Draw parameter values only once. Discard if any condition is not fulfilled



Set mutation models

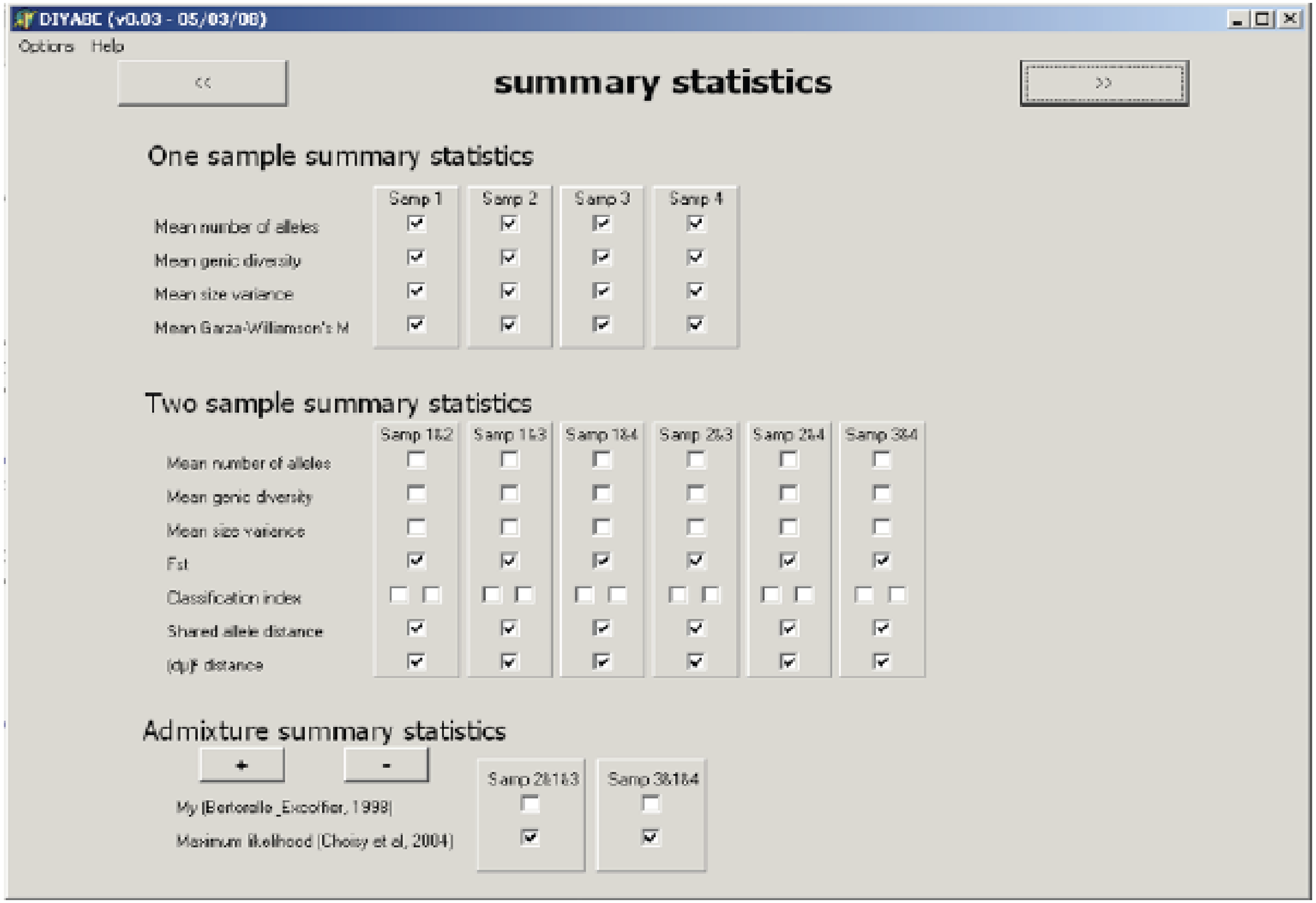
Mutation model: ○ SMM ● GSM

Single nucleotide indel mutation: ● NO ○ YES

Mutation rates: ○ Each locus = Mean ● Each locus = Gamma(Mean) ○ Each locus = coeff x Mean

Coefficients P: ○ Each locus = Mean ● Each locus = Gamma(Mean)

parameter	Prior distribution	minimum	maximum	mean	shape	step
Mean mutation rate	● Uniform ○ Gamma	1.00E-004	1.00E-003			1.00E-005
Locus mutation rate	● Gamma	1.00E-005	1.00E-002	Mean μ	2.00E+000	1.00E-005
Mean coeficient P	● Uniform ○ Gamma	1.00E-001	3.00E-001			1.00E-002
Locus coeficient P	● Gamma	1.00E-002	5.00E-001	Mean P	2.00E+000	1.00E-002



summary statistics

One sample summary statistics

	Samp 1	Samp 2	Samp 3	Samp 4
Mean number of alleles	☑	☑	☑	☑
Mean genic diversity	☑	☑	☑	☑
Mean size variance	☑	☑	☑	☑
Mean Garza-Williamson's M	☑	☑	☑	☑

Two sample summary statistics

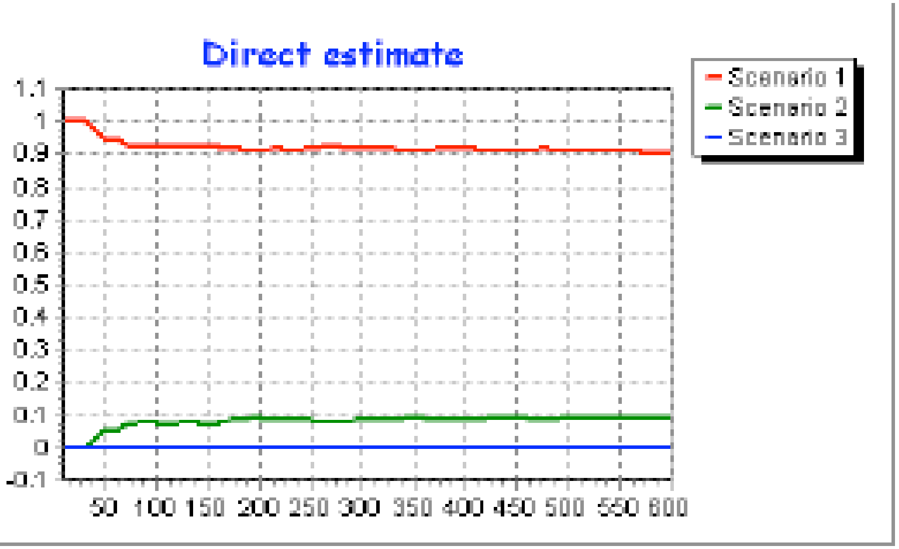
	Samp 1&2	Samp 1&3	Samp 1&4	Samp 2&3	Samp 2&4	Samp 3&4
Mean number of alleles	□	□	□	□	□	□
Mean genic diversity	□	□	□	□	□	□
Mean size variance	□	□	□	□	□	□
Fst	☑	☑	☑	☑	☑	☑
Classification index	□ □	□ □	□ □	□ □	□ □	□ □
Shared allele distance	☑	☑	☑	☑	☑	☑
(dμ)² distance	☑	☑	☑	☑	☑	☑

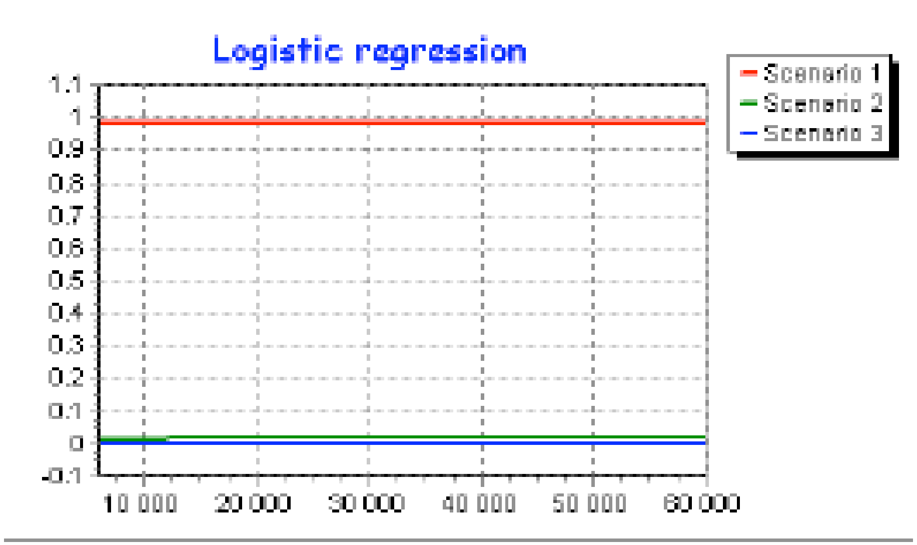
Admixture summary statistics

	Samp 2&1&3	Samp 3&1&4
My (Bertorelle_Excoffier, 1998)	□	□
Maximum likelihood (Choisy et al, 2004)	☑	☑

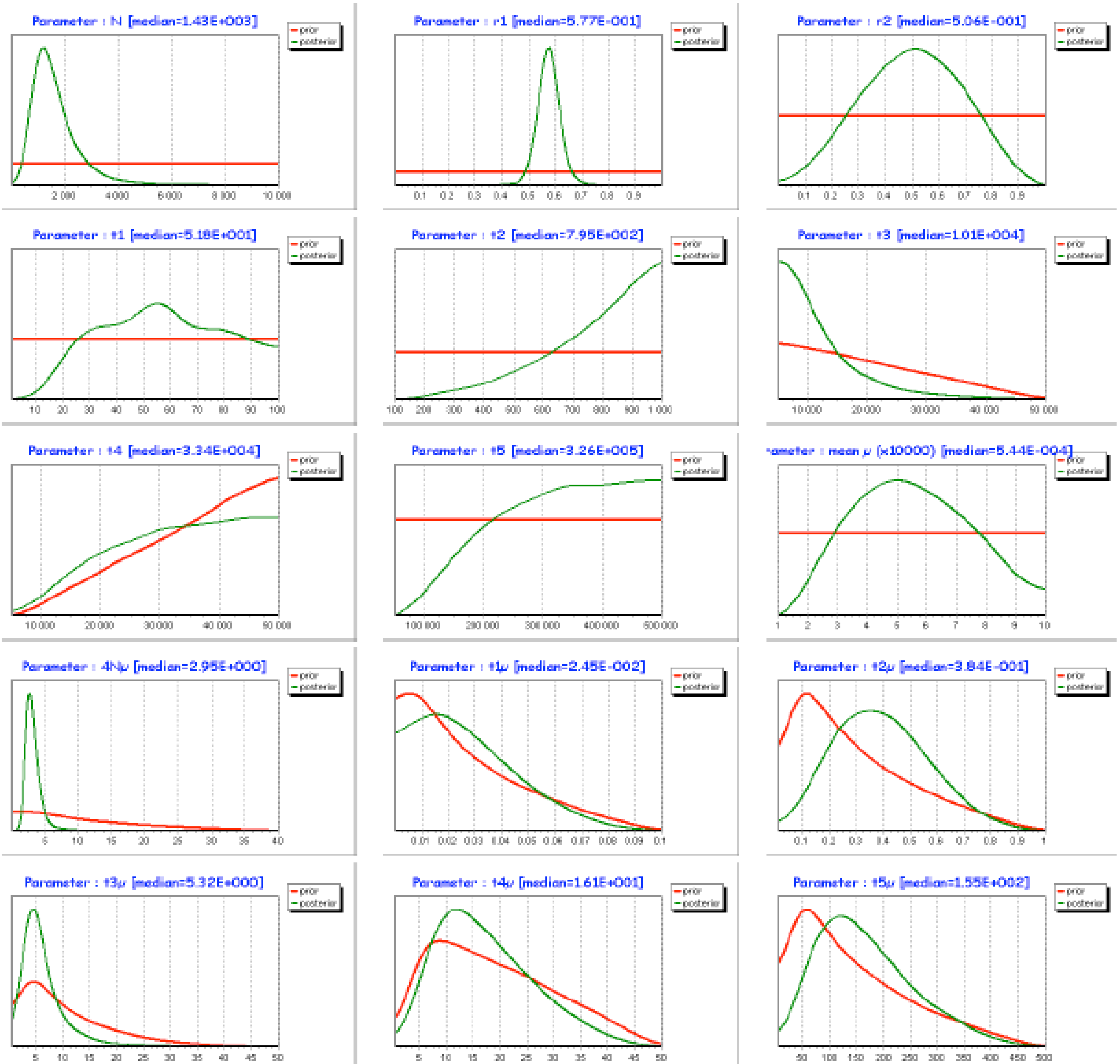
(Relative) posterior probabilities of scenarios 1, 2 and 3

→ Reference table = 3 x 10^6 data sets

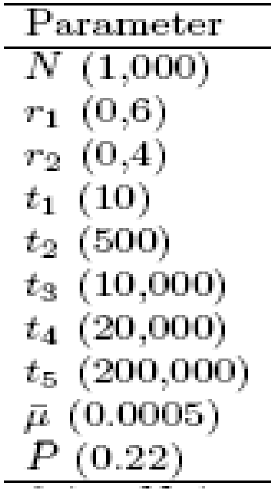




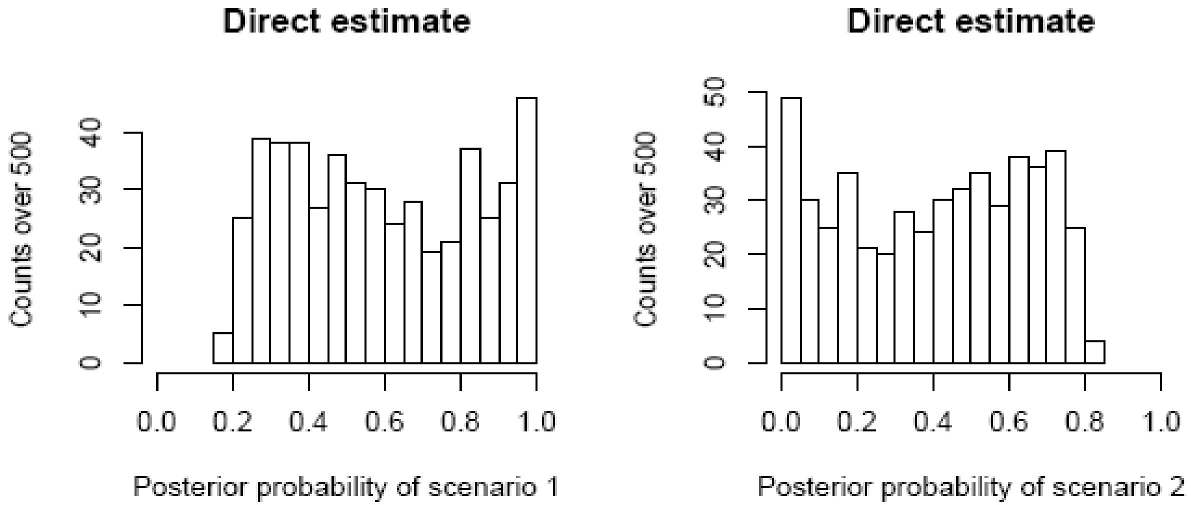
Estimation of posterior distributions under scenario 1



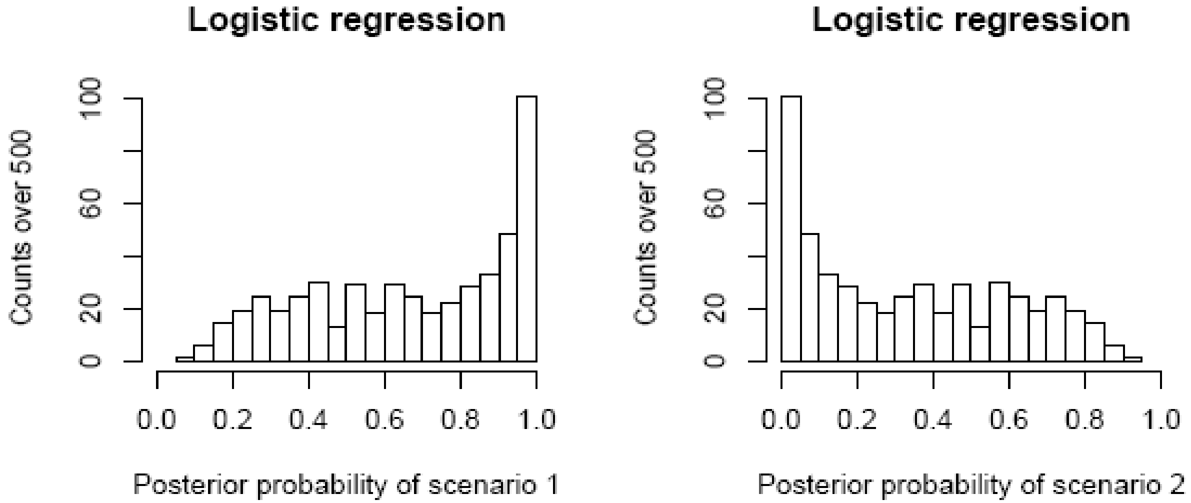
Power to discriminate scenarios: 500 test data sets simulated under scenario 1 → type I error



Parameter
N (1,000)
r_1 (0,6)
r_2 (0,4)
t_1 (10)
t_2 (500)
t_3 (10,000)
t_4 (20,000)
t_5 (200,000)
$\bar{\mu}$ (0.0005)
P (0.22)



Type I error = 0.414



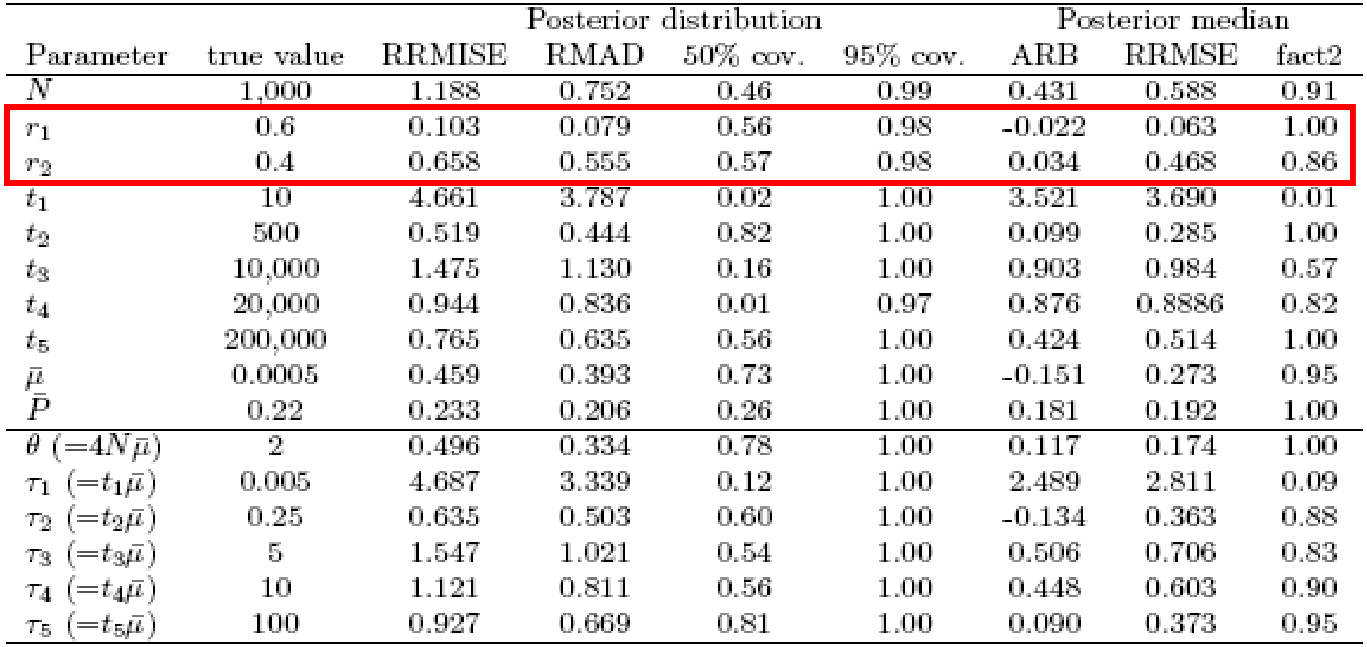
Type I error = 0.300

Power to discriminate scenarios: 500 test data sets simulated under scenario 2 and 3 → type II error

- Direct estimate: scenario 2 = 0.014 and scenario 3 = 0.000

- Logistic regression: scenario 2 = 0.020 and scenario 3 = 0.000

Accuracy to estimate parameters under scenario 1 (500 simulated test data sets)



		Posterior distribution				Posterior median		
Parameter	true value	RRMISE	RMAD	50% cov.	95% cov.	ARB	RRMSE	fact2
N	1,000	1.188	0.752	0.46	0.99	0.431	0.588	0.91
r_1	0.6	0.103	0.079	0.56	0.98	-0.022	0.063	1.00
r_2	0.4	0.658	0.555	0.57	0.98	0.034	0.468	0.86
t_1	10	4.661	3.787	0.02	1.00	3.521	3.690	0.01
t_2	500	0.519	0.444	0.82	1.00	0.099	0.285	1.00
t_3	10,000	1.475	1.130	0.16	1.00	0.903	0.984	0.57
t_4	20,000	0.944	0.836	0.01	0.97	0.876	0.8886	0.82
t_5	200,000	0.765	0.635	0.56	1.00	0.424	0.514	1.00
$\bar{\mu}$	0.0005	0.459	0.393	0.73	1.00	-0.151	0.273	0.95
$\bar{P}$	0.22	0.233	0.206	0.26	1.00	0.181	0.192	1.00
θ $(=4N\bar{\mu})$	2	0.496	0.334	0.78	1.00	0.117	0.174	1.00
τ_1 $(=t_1\bar{\mu})$	0.005	4.687	3.339	0.12	1.00	2.489	2.811	0.09
τ_2 $(=t_2\bar{\mu})$	0.25	0.635	0.503	0.60	1.00	-0.134	0.363	0.88
τ_3 $(=t_3\bar{\mu})$	5	1.547	1.021	0.54	1.00	0.506	0.706	0.83
τ_4 $(=t_4\bar{\mu})$	10	1.121	0.811	0.56	1.00	0.448	0.603	0.90
τ_5 $(=t_5\bar{\mu})$	100	0.927	0.669	0.81	1.00	0.090	0.373	0.95

Example of inferences on a complex population history: 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.

© Survival

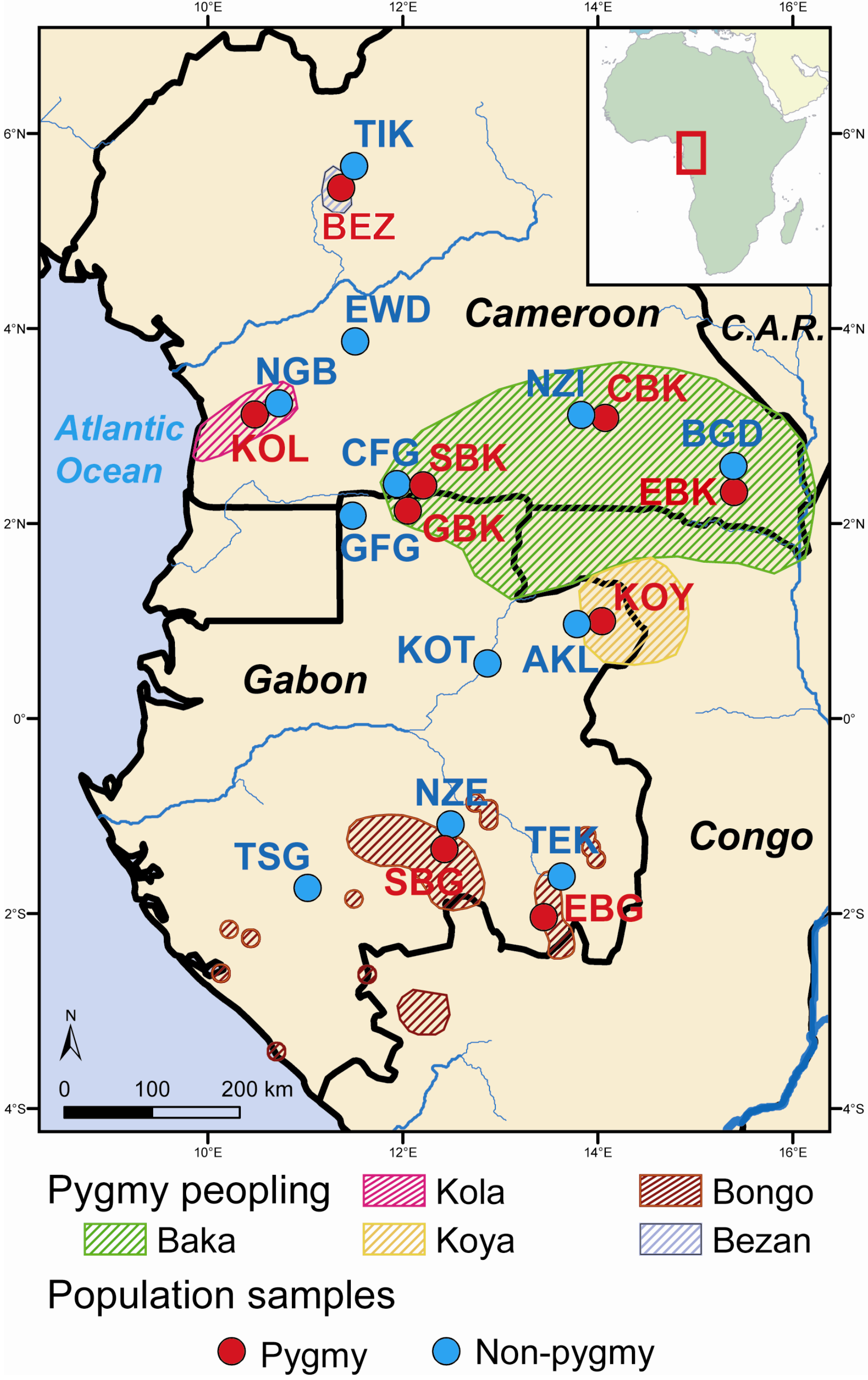
- 604 individuals

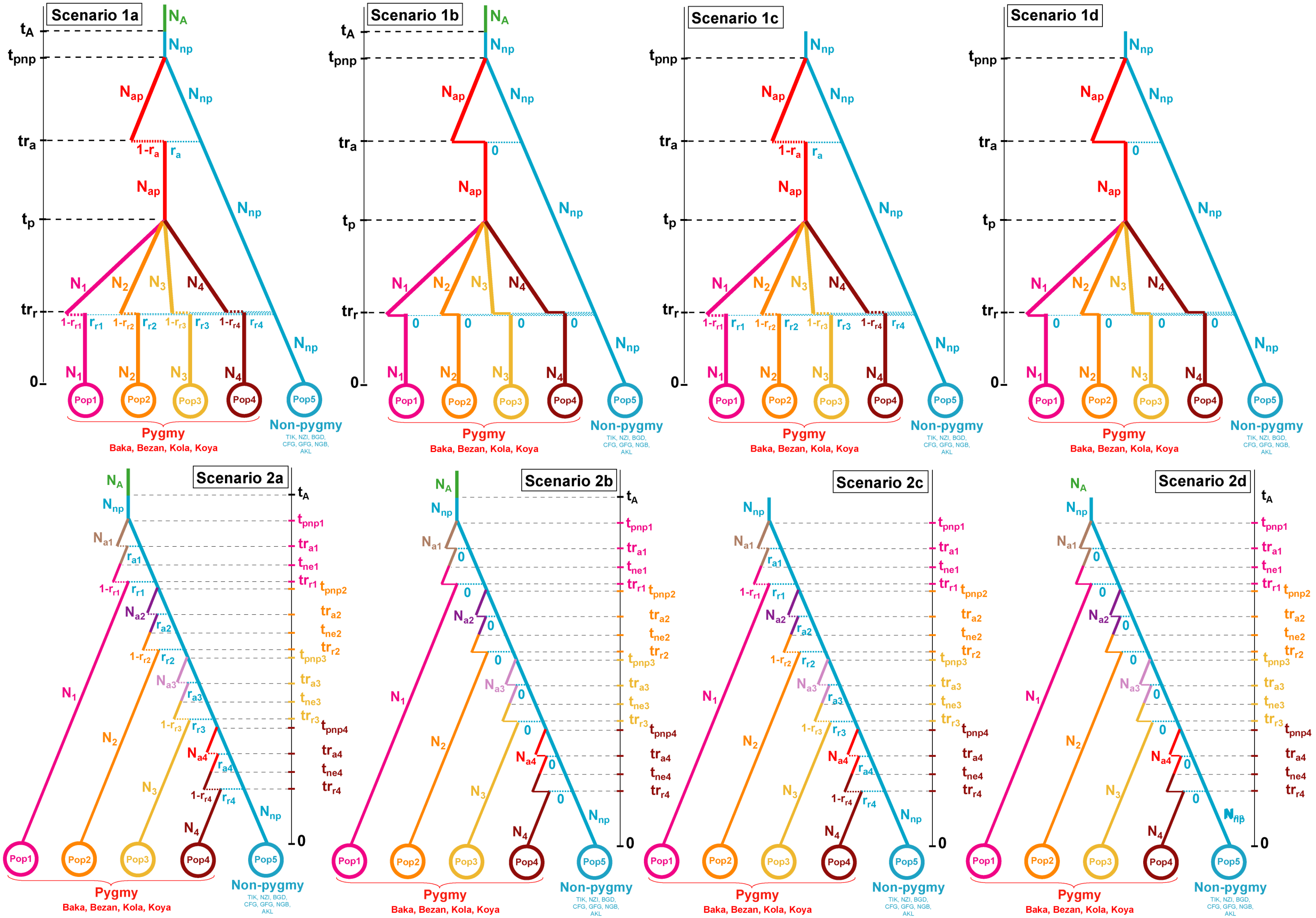
- 12 non pygmy and nine neighbouring pygmy populations

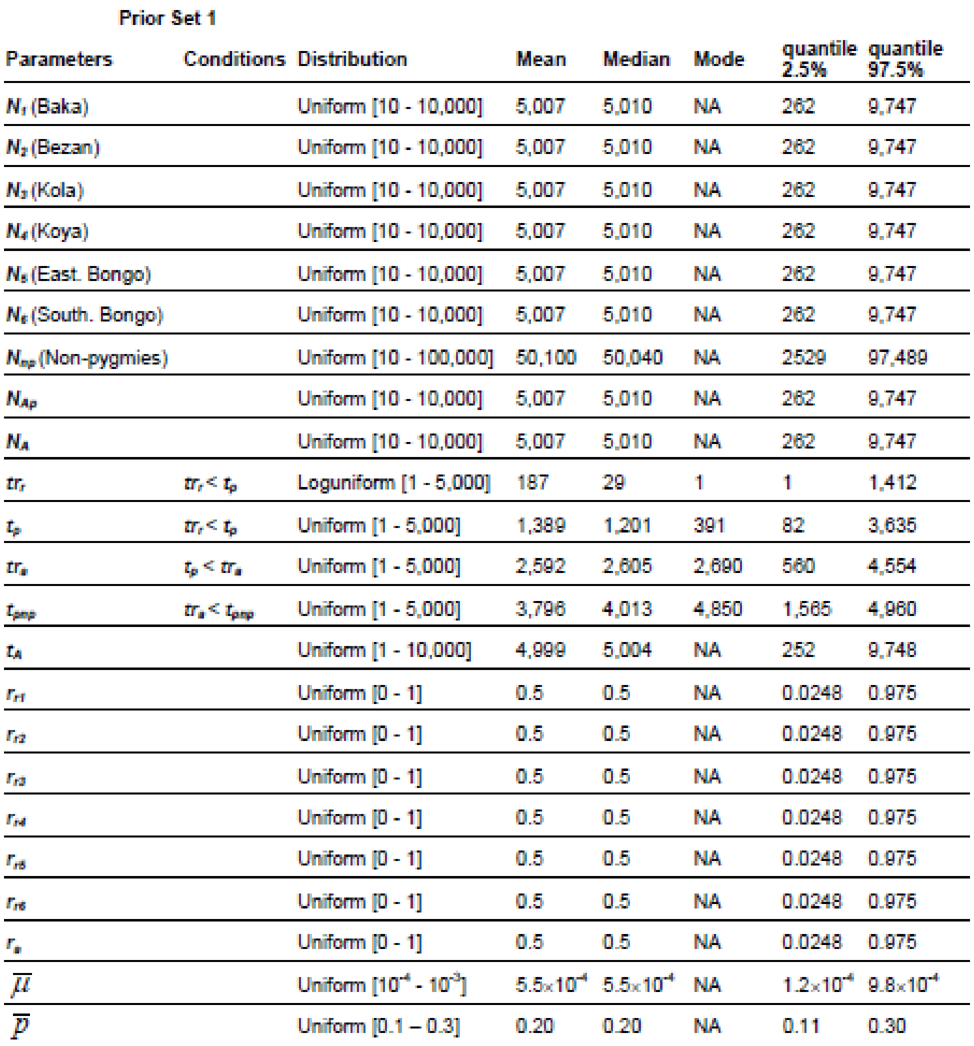
- 28 microsatellite loci

→No genetic structure between non pygmy populations

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




Scenario 1a
N_A
N_{np}
t_A
t_{pnp}
N_{ap}
N_{np}
tr_a
$1-r_a$
r_a
N_{ap}
t_p
N_{np}
N_1
N_2
N_3
N_4
tr_r
$1-r_{r1}$
r_{r1}
$1-r_{r2}$
r_{r2}
$1-r_{r3}$
r_{r3}
$1-r_{r4}$
r_{r4}
N_{np}
0
N_1
N_2
N_3
N_4
Pop1
Pop2
Pop3
Pop4
Pop5
Pygmy
Baka, Bezan, Kola, Koya
Non-pygmy
TIK, NZI, BGD, CFG, GFG, NGB, AKL
Scenario 1b
N_A
N_{np}
t_A
t_{pnp}
N_{ap}
N_{np}
tr_a
0
N_{ap}
t_p
N_{np}
N_1
N_2
N_3
N_4
tr_r
0
0
0
0
N_{np}
0
N_1
N_2
N_3
N_4
Pop1
Pop2
Pop3
Pop4
Pop5
Pygmy
Baka, Bezan, Kola, Koya
Non-pygmy
TIK, NZI, BGD, CFG, GFG, NGB, AKL
Scenario 1c
N_{np}
t_{pnp}
N_{ap}
N_{np}
tr_a
$1-r_a$
r_a
N_{ap}
t_p
N_{np}
N_1
N_2
N_3
N_4
tr_r
$1-r_{r1}$
r_{r1}
$1-r_{r2}$
r_{r2}
$1-r_{r3}$
r_{r3}
$1-r_{r4}$
r_{r4}
N_{np}
0
N_1
N_2
N_3
N_4
Pop1
Pop2
Pop3
Pop4
Pop5
Pygmy
Baka, Bezan, Kola, Koya
Non-pygmy
TIK, NZI, BGD, CFG, GFG, NGB, AKL
Scenario 1d
N_{np}
t_{pnp}
N_{ap}
N_{np}
tr_a
0
N_{ap}
t_p
N_{np}
N_1
N_2
N_3
N_4
tr_r
0
0
0
0
N_{np}
0
N_1
N_2
N_3
N_4
Pop1
Pop2
Pop3
Pop4
Pop5
Pygmy
Baka, Bezan, Kola, Koya
Non-pygmy
TIK, NZI, BGD, CFG, GFG, NGB, AKL
Scenario 2a
N_A
N_{np}
t_A
t_{pnp1}
N_{a1}
tr_{a1}
r_{a1}
t_{ne1}
tr_{r1}
$1-r_{r1}$
r_{r1}
t_{pnp2}
N_{a2}
tr_{a2}
r_{a2}
t_{ne2}
tr_{r2}
$1-r_{r2}$
r_{r2}
t_{pnp3}
N_1
N_{a3}
tr_{a3}
r_{a3}
t_{ne3}
tr_{r3}
$1-r_{r3}$
r_{r3}
t_{pnp4}
N_2
N_{a4}
tr_{a4}
r_{a4}
t_{ne4}
tr_{r4}
$1-r_{r4}$
r_{r4}
N_3
N_{np}
N_4
0
Pop1
Pop2
Pop3
Pop4
Pop5
Pygmy
Baka, Bezan, Kola, Koya
Non-pygmy
TIK, NZI, BGD, CFG, GFG, NGB, AKL
Scenario 2b
N_A
N_{np}
t_A
t_{pnp1}
N_{a1}
tr_{a1}
0
t_{ne1}
tr_{r1}
t_{pnp2}
0
N_{a2}
tr_{a2}
0
t_{ne2}
tr_{r2}
t_{pnp3}
0
N_1
N_{a3}
tr_{a3}
0
t_{ne3}
tr_{r3}
t_{pnp4}
0
N_2
N_{a4}
tr_{a4}
0
t_{ne4}
tr_{r4}
0
N_3
N_{np}
N_4
0
Pop1
Pop2
Pop3
Pop4
Pop5
Pygmy
Baka, Bezan, Kola, Koya
Non-pygmy
TIK, NZI, BGD, CFG, GFG, NGB, AKL
Scenario 2c
N_{np}
t_{pnp1}
N_{a1}
tr_{a1}
r_{a1}
t_{ne1}
tr_{r1}
$1-r_{r1}$
r_{r1}
t_{pnp2}
N_{a2}
tr_{a2}
r_{a2}
t_{ne2}
tr_{r2}
$1-r_{r2}$
r_{r2}
t_{pnp3}
N_1
N_{a3}
tr_{a3}
r_{a3}
t_{ne3}
tr_{r3}
$1-r_{r3}$
r_{r3}
t_{pnp4}
N_2
N_{a4}
tr_{a4}
r_{a4}
t_{ne4}
tr_{r4}
$1-r_{r4}$
r_{r4}
N_3
N_{np}
N_4
0
Pop1
Pop2
Pop3
Pop4
Pop5
Pygmy
Baka, Bezan, Kola, Koya
Non-pygmy
TIK, NZI, BGD, CFG, GFG, NGB, AKL
Scenario 2d
N_A
N_{np}
t_A
t_{pnp1}
N_{a1}
tr_{a1}
0
t_{ne1}
tr_{r1}
t_{pnp2}
0
N_{a2}
tr_{a2}
0
t_{ne2}
tr_{r2}
t_{pnp3}
0
N_1
N_{a3}
tr_{a3}
0
t_{ne3}
tr_{r3}
t_{pnp4}
0
N_2
N_{a4}
tr_{a4}
0
t_{ne4}
tr_{r4}
0
N_3
N_{np}
N_4
0
Pop1
Pop2
Pop3
Pop4
Pop5
Pygmy
Baka, Bezan, Kola, Koya
Non-pygmy
TIK, NZI, BGD, CFG, GFG, NGB, AKL

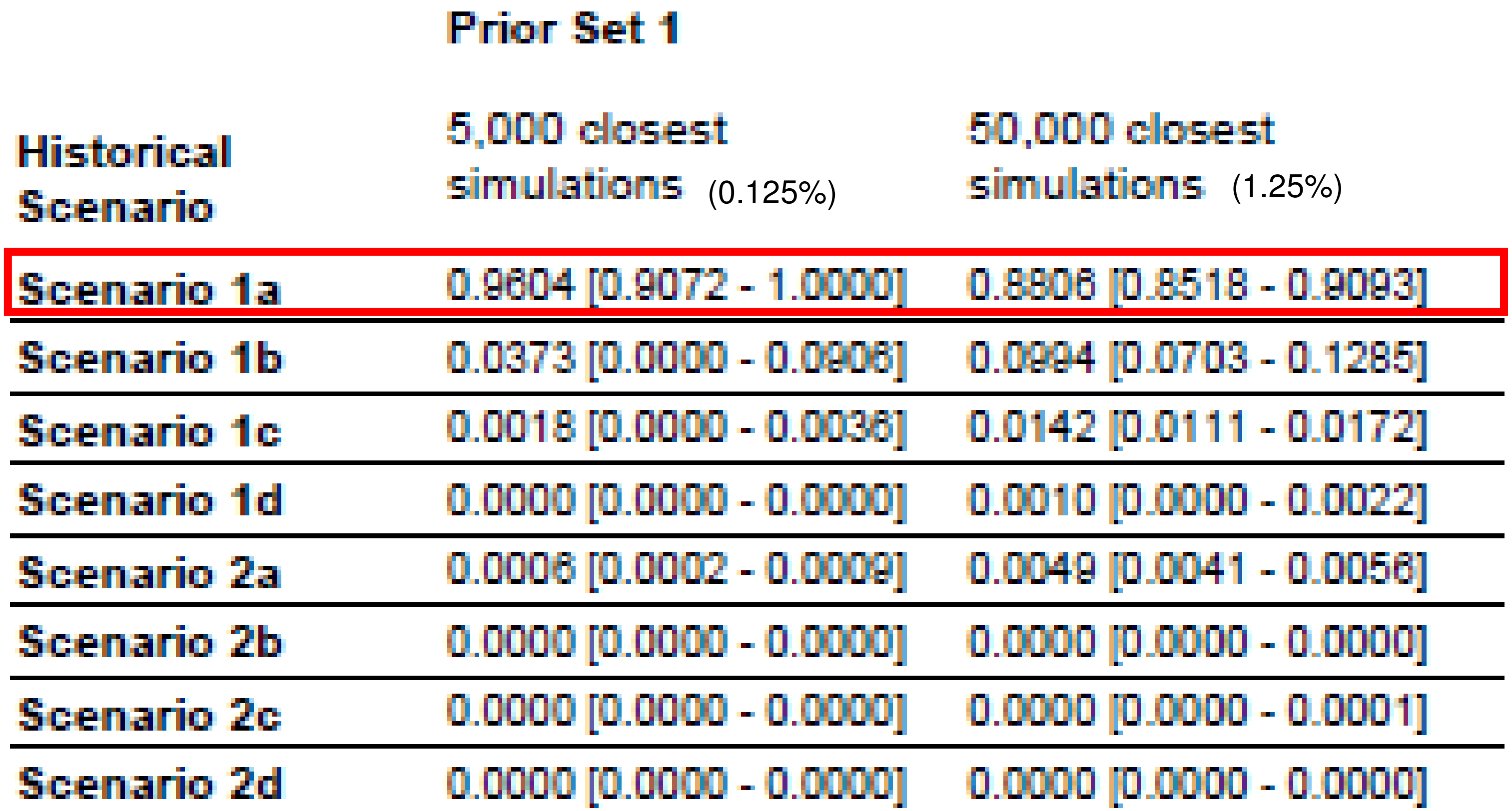


Prior Set 1

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

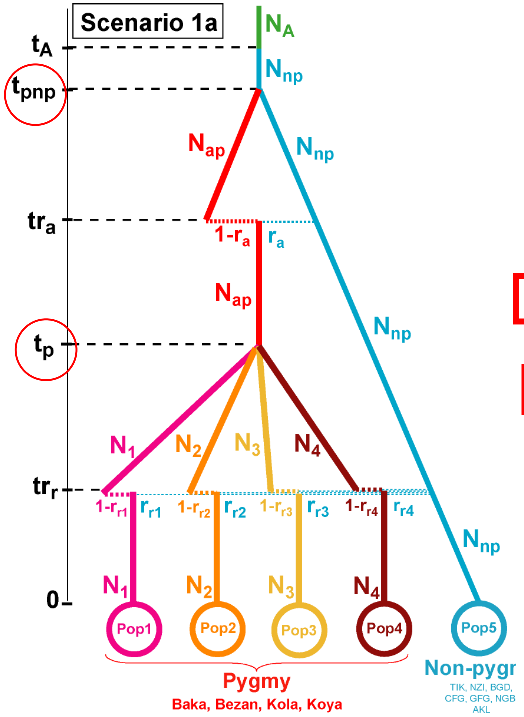
→ 500,000 simulations per scenario (total: 4 M)

Relative posterior probabilities for each scenario



Historical Scenario	Prior Set 1 5,000 closest simulations (0.125%)	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	0.0006 [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]

Estimation of parameters under scenario 1a



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
$\overline{\mu}$	0.00024	0.00021	0.00016	0.00011	0.00056
$\overline{p}$	0.11	0.11	0.10	0.10	0.16

Power study: 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]

Main perspectives

- DNA sequence data, SNP, AFLP
- Gene flow between populations
- Reproduction systems (autofecondation, clonality)
- Selection