Likelihood-free inference of demography, mutation rates, and local selection in a Bayesian hierarchical model

Eric Bazin,
UMR BGPI, CIRAD,
Montpellier , France

Kevin J. Dawson
Rothamsted Research
Harpenden, Hertfordshire, UK

Mark A. Beaumont,
School of Biological Sciences,
The University of Reading,
Reading, Berkshire, UK

Introduction and Motivation

- Approximate Bayesian computation (ABC) has proved useful for complex models in which the likelihood function is difficult or expensive to obtain.
- Hierarchical Bayesian models pose a problem for ABC because the number of summary statistics potentially grows with the number of exchangeable units (replicates) at the lower level of the hierarchy.
- This study develops a method for performing hierarchical Bayesian analysis using ABC that is practical for a very large number of summary statistics.

Applications of Hierarchical models in Genetics

- Modelling variability in mutation rate among loci (Yang, 1993).
- Modelling positive selection through a variable substitution model (dN/dS models: Yang and Nielsen, 1998; Wilson and McVean, 2006).
- Modelling local selection through a variable migration rate model (Beaumont and Balding, 2004).
- Modelling assignment of individuals to populations (*Structure*: Pritchard *et al.*, 2000)

Some Notation

We denote by α the vector of hyper-parameters in the model.

For the ith unit/replicate (e.g. locus) there is a vector of observations (X_i), and (unobserved) parameter vectors κ_i and λ_i, giving the matrices:

- $X = (X_1, \ldots, X_L)$
- $\kappa = (\kappa_1, \ldots, \kappa_L)$
- $\lambda = (\lambda_1, \ldots, \lambda_L)$

Here, we treat λ_i as a parameter of interest, and κ_i as a nuisance parameter (*e.g.* genealogy.)

We denote by X_0 the "real" observations, in contrast to simulated observations X.

To avoid too many subscripts, depending on context, X_k, κ_k, λ_k, α_k refers to the kth simulated instance of X, κ, λ, α .

Bayesian hierarchical models

The likelihood function for our model is

$$p\left(X|\kappa,\lambda\right) = \prod_{i=1}^{L} p(X_i|\kappa_i,\lambda_i). \tag{1}$$

with prior

$$p(\alpha,\kappa,\lambda) = \left[\prod_{i=1}^{L} p(\kappa_i,\lambda_i|\alpha)\right] p(\alpha). \tag{2}$$

Because of conditional independence, the posterior distribution (shown here marginal to the nuisance parameter κ), factorises as

$$p(\alpha,\lambda|X) = \left[\prod_{i=1}^{L} p(\lambda_i|X_i,\alpha)\right] p(\alpha|X). \tag{3}$$

Bayesian hierarchical models

Now, focusing attention on a single locus i, the hyper-parameter α and the locus-specific parameter λ_i have the joint posterior density

$$p(\alpha, \lambda_i | X) = p(\lambda_i | X_i, \alpha) p(\alpha | X). \tag{4}$$

This factorisation suggests that we need to use two distinct types of summary statistics in our approximate Bayesian computation:

- *symmetric* summary statistics, which are (symmetric) functions of all the loci together (*e.g.* means, higher moments,...), $S(X) = S(X_1, \ldots, X_L)$;
- *unit-specific* summary statistics, $U(X_i)$.

Bayes sufficiency

Ideally, we want the summary statistic $S(X)$ to satisfy the condition

$$p(\omega|X) = p(\omega|S(X)), \tag{5}$$

at all points ω (in the parameter space), for all priors $p(\omega)$.

In this case, the summary statistic $S(X)$ is *sufficient* in the sense of Kolmogorov (1942) [3]. In other words, the summary statistic $S(X)$ is *Bayes sufficient*.

Marginal sufficiency

Ideally, we want the statistics $S(X)$ and $U(X_i)$ to satisfy the condition

$$p(\alpha, \lambda_i | X) = p(\lambda_i | U(X_i), \alpha) p(\alpha | S(X)), \tag{6}$$

at all points (α, λ_i) (in a section of the parameter space), for the chosen prior (or family of priors). We want this factorisation to hold exactly, or at least as an adequate approximation.

In the terminology of *marginal sufficiency* introduced by Raiffa and Schlaifer (1961) [4] (see also Basu 1977 [1]), this tells us that:

- The summary statistic $S(X)$ is marginally sufficient for the parameter α;
- The summary statistic $(S(X), U(X_i))$ is marginally sufficient for the locus-specific parameter λ_i.

with respect to the chosen prior (or family of priors).

Single step algorithm

For $k = 1$ to $k = N$ iterations:

(i) sample (A_k, K_k, Λ_k) from the prior $p(\kappa, \lambda|\alpha)p(\alpha)$;

(ii) simulate data X_k (at L loci) from $p(X_k|K_k, \Lambda_k)$;

For locus $i = 1$ to $i = L$:

(iii) Compute $(A_k, \Lambda_{k,i}, S(X_k), U(X_{k,i}))$.

(iv) Condition on $S(X) = S(X_0)$ and $U(X_i) = U(X_{0,i})$ using ABC, to obtain a sample of observations $(A_k^*, \Lambda_{k,i}^*)$ from $p(\alpha, \lambda_i|S(X_0), U(X_{0,i}))$.

Practical Issue

Storage space problems.

- We need to store NL multiplied by number of items in $U(X_i)$.
- *E.g.* for 10^3 loci, 10 summary statistics per locus, 10^6 iterations, 8 bytes per number we have 80Gb of storage (as a binary file, or in computer memory).
- Therefore there is a problem with scaling up.

A two step algorithm: more efficient, but only approximate

step 1 Use ABC to obtain a sample of observations A_k^* from

$$p(\alpha|S(X_0)) \approx p(\alpha|X_0).$$

step 2 For locus $i = 1$ to $i = L$, obtain a sample of observations $(A_k^{**}, \Lambda_{k,i}^*)$ from an *approximation* to $p(\lambda_i|X_{0,i}, \alpha)p(\alpha|X_0)$ by doing the following:

(i) resampling A_k^{**} from the observations A_k^* generated in step 1.
(ii) sampling $(K_{k,i}^{**}, \Lambda_{k,i}^{**})$ from the conditional prior $p(\kappa_i, \lambda_i|A_k^{**})$;
(iii) simulating data $X_{k,i}$ (at locus i only) from $p(X_{k,i}|K_{k,i}^{**}, \Lambda_{k,i}^{**})$;
(iv) Marginalise observations to $(A_k^{**}, \Lambda_{k,i}^{**}, U(X_{k,i}))$.
(v) Condition on $U(X_i) = U(X_{0,i})$ using ABC.

Step 1

For $k = 1$ to $k = N$ iterations:

(i) sample (A_k, K_k, Λ_k) from the prior $p(\kappa, \lambda|\alpha)p(\alpha)$;

(ii) simulate data X_k (at L loci) from $p(X_k|K_k, \Lambda_k)$;

(iii) Marginalise by mapping observations $(A_k, K_k, \Lambda_k, X_k)$ to $(A_k, S(X_k)))$.

(iv) Condition on $S(X) = S(X_0)$ using ABC, to obtain a sample of observations A_k^* from

$$p(\alpha|S(X_0)) \approx p(\alpha|X_0),$$

Step 2

For locus $i = 1$ to $i = L$:

- For $k = 1$ to $k = N$ iterations:
  - (i) sample A_k^{**} from $p(\alpha|f(X_0)) \approx p(\alpha|X_0)$ by resampling from the observations A_k^* generated in step 1.
  - (ii) sample $(K_{k,i}^{**}, \Lambda_{k,i}^{**})$ from the conditional prior $p(\kappa_i, \lambda_i|A_k^{**})$;
  - (iii) simulate data $X_{k,i}$ (at locus i only) from $p(X_{k,i}|K_{k,i}^{**}, \Lambda_{k,i}^{**})$;
  - (iv) Marginalise by mapping observations $(A_k^{**}, K_{k,i}^{**}, \Lambda_{k,i}^{**}, X_{k,i})$ to $(A_k^{**}, \Lambda_{k,i}^{**}, U(X_{k,i}))$.

(v) Condition on $U(X_i) = U(X_{0,i})$ using ABC, to obtain a sample of observations $(A_k^{***}, \Lambda_{k,i}^{***})$ from an *approximation* to $p(\lambda_i|X_{0,i}, \alpha)p(\alpha|X_0)$.

Advantages and disadvantages

- *Lower storage requirement*.
  The two step algorithm requires less storage — by a factor of $1/L$ in comparison with one step algorithm.

- *Less writing to disk or to files*.
  The two step algorithm requires less disk-writing, time — by a factor of $1/L$ in comparison with one step algorithm.

- *More simulation*.
  Computational cost of two step algorithm is twice as high as one step algorithm:

The two step algorithm involves an approximation

This approximation is in addition to the approximation involved in conditional density estimation (using some ABC method) on summary statistics rather than on complete data. So, to simplify the explanation of this additional approximation, we will assume that we are performing ABC on complete data.
Now in the two step algorithm we have a sample from

$$p(x_i', \lambda_i|\alpha)p(\alpha|x = X_0),$$

then we condition on $x_i' = X_{0,i}$. This gives us a sample of observations $(A_k^{**}, \Lambda_{k,i}^{**})$ from

$$\frac{p(x_i' = X_{0,i}, \lambda_i|\alpha)p(\alpha|x = X_0)}{p(x_i' = X_{0,i}|x = X_0)}.$$

The approximation

If we modify the two step algorithm so that we sample from $p(\alpha|x_{-i} = X_{0,-i})$ at step 1 (instead of $p(\alpha|x = X_0)$), then we have a sample from

$$p(x_i, \lambda_i|\alpha)p(\alpha|x_{-i} = X_{0,-i}),$$

then we condition on $x_i' = X_{0,i}$. This gives us a sample of observations $(A_k^{**}, \Lambda_{k,i}^{**})$ from

$$\frac{p(x_i = X_{0,i}, \lambda_i|\alpha)p(\alpha|x_{-i} = X_{0,-i})}{p(x_i = X_{0,i}|x_{-i} = X_{0,-i})} = p(\lambda_i, \alpha|x = X_0)$$

instead of

$$\frac{p(x_i' = X_{0,i}, \lambda_i|\alpha)p(\alpha|x = X_0)}{p(x_i' = X_{0,i}|x = X_0)}$$

Further details

$$\frac{p(x_i, \lambda_i|\alpha)p(\alpha|x_{-i})}{p(x_i|x_{-i})}$$

$$= \frac{p(x_i, \lambda_i|\alpha)p(\alpha|x_{-i})}{p(x_i|\alpha)} \cdot \frac{p(x_i|\alpha)p(\alpha|x_{-i})}{p(x_i|x_{-i})}$$

$$= p(\lambda_i|x_i, \alpha)p(\alpha|x)$$

$$= p(\lambda_i, \alpha|x).$$

Justification for the approximation

So, we have a modified algorithm which generates a sample of observations $(A_k^{***}, \Lambda_{k,i}^{***})$ from the *exact* marginal posterior

$$p(\lambda_i | X_{0,i}, \alpha) p(\alpha | X_0).$$

But the only difference between our two step algorithm, and this modified algorithm is how we generated the observations A_k^* at step 1.

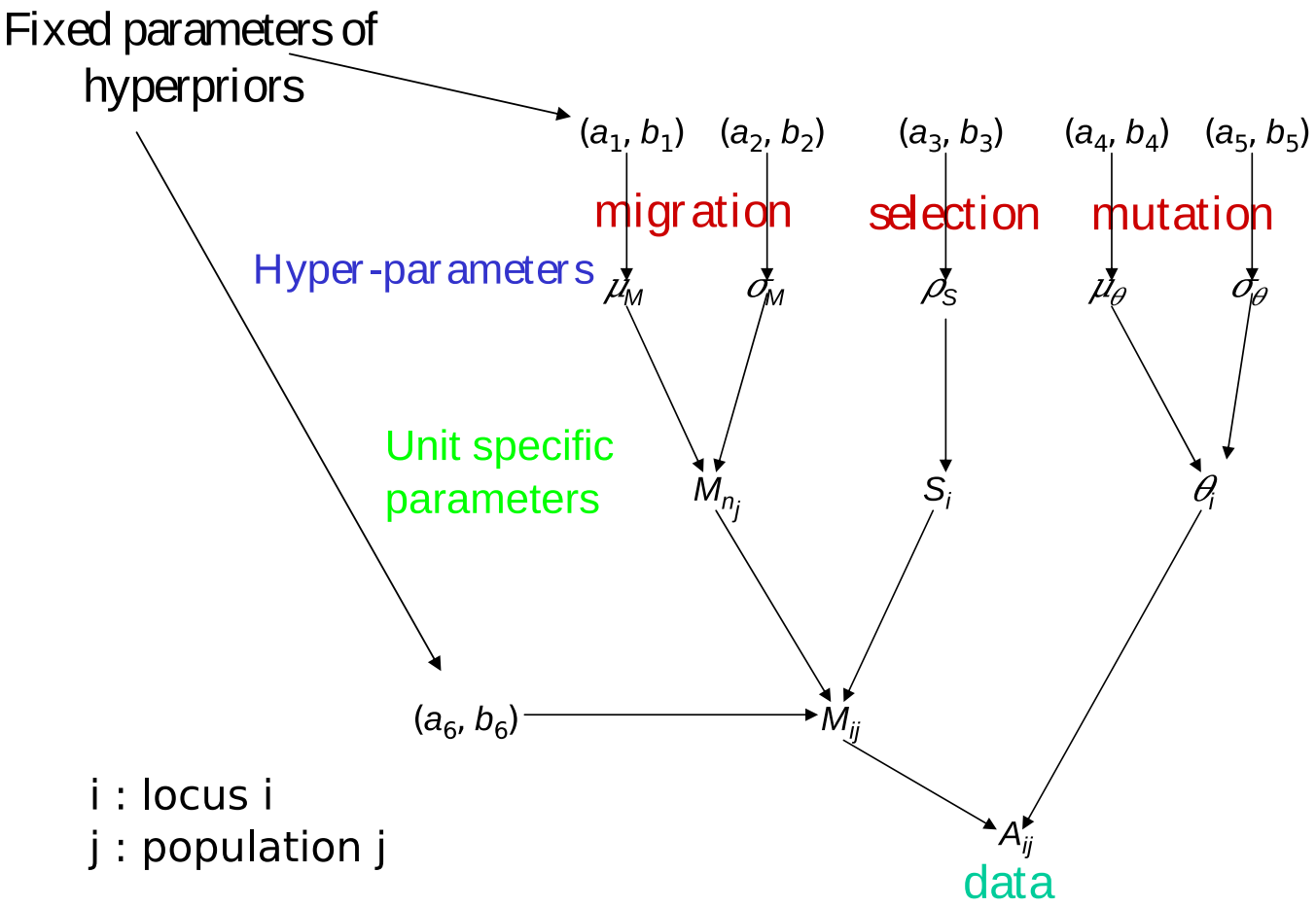
Now, when the number of loci L is large, we expect to have:

$$\begin{aligned} p(\alpha | X_{0,-i}) &\approx p(\alpha | X_{0,-i}, X_{0,i}) \\ &= p(\alpha | X_0). \end{aligned}$$

So, when the number of loci L is large, our two step algorithm differs very little form this modified algorithm. (But we have replaced the data by summary statistics, in the ABC.)

Application to inferring selection

Genetic model



i : locus i
j : population j

Application to inferring selection

Model parameters

Parameter	Description	Prior distribution
μ_M	mean scaled migration rate across populations	$N(a_1, b_1)$
σ_M	standard deviation of scaled migration rate across populations	$N(a_2, b_2)$
ρ_s	probability that a locus is under selection	$\beta(a_3, b_3)$
μ_θ	mean mutation rate across loci	$N(a_4, b_4)$
σ_θ	standard deviation of mutation rate across loci	$N(a_5, b_5)$
θ_i	scaled mutation rate of the ith locus	$Log10\text{-}N(\mu_\theta, \sigma_\theta)$
S_i	indicator that is 0 if the ith locus is neutral and 1 if it is selected	$B(\rho_s)$
M_{ij}	migration rate of the ith locus in population j	...

Application to inferring selection

Approximation

We use the approximation of Petry (1982) that local selection at linked sites gives rise to the same distribution of gene frequencies as a neutral locus with reduced migration rate.

Each locus and each deme has scaled migration rate $M_{ij} = 2Nm_{ij}$, where

$$M_{ij} = \begin{cases} M_{n_j} & \text{if } S_i = 0 \\ M_{s_{ij}} & \text{if } S_i = 1 \end{cases}$$

with

$$M_{n_j} \sim \textit{Log10-N}(\mu_M, \sigma_M)$$

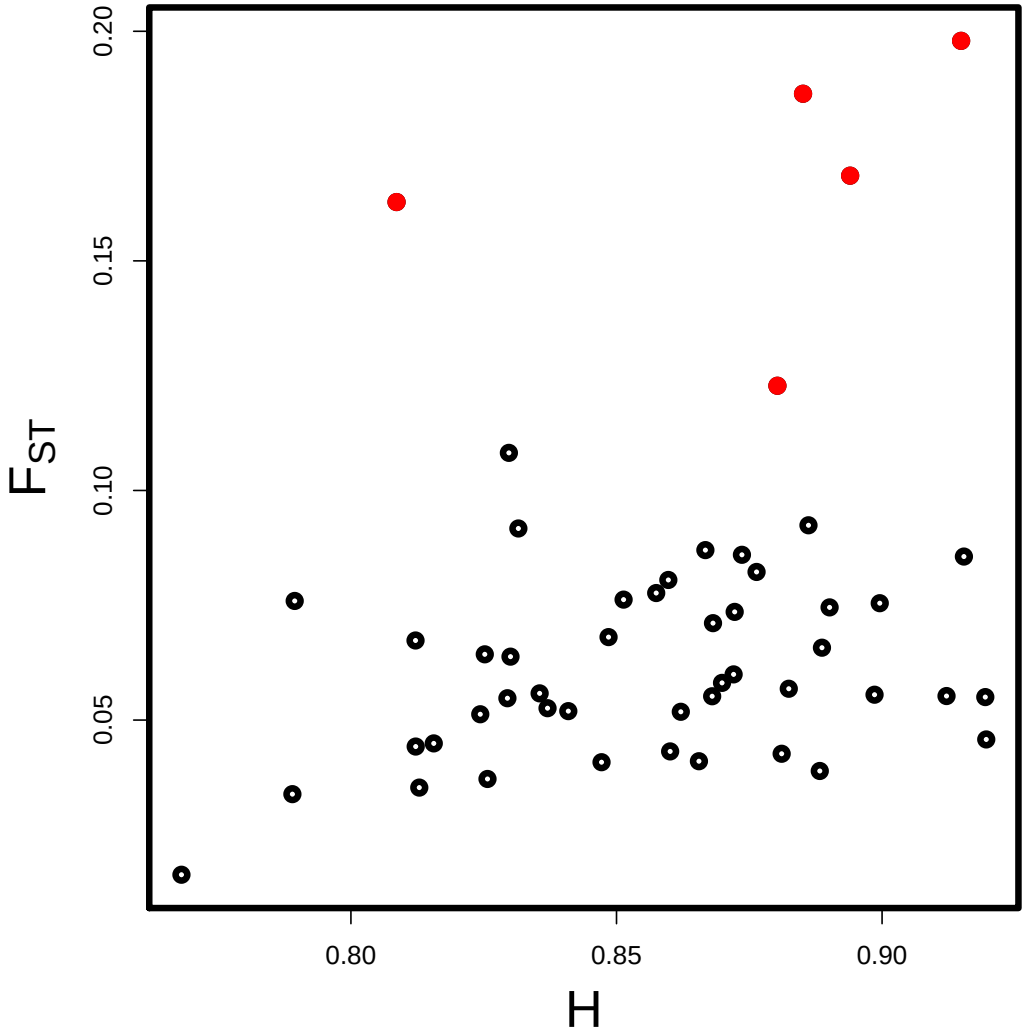
and

$$M_{s_{ij}} \sim \beta(x/M_{n_j}; a_6, b_6)/M_{n_j}$$

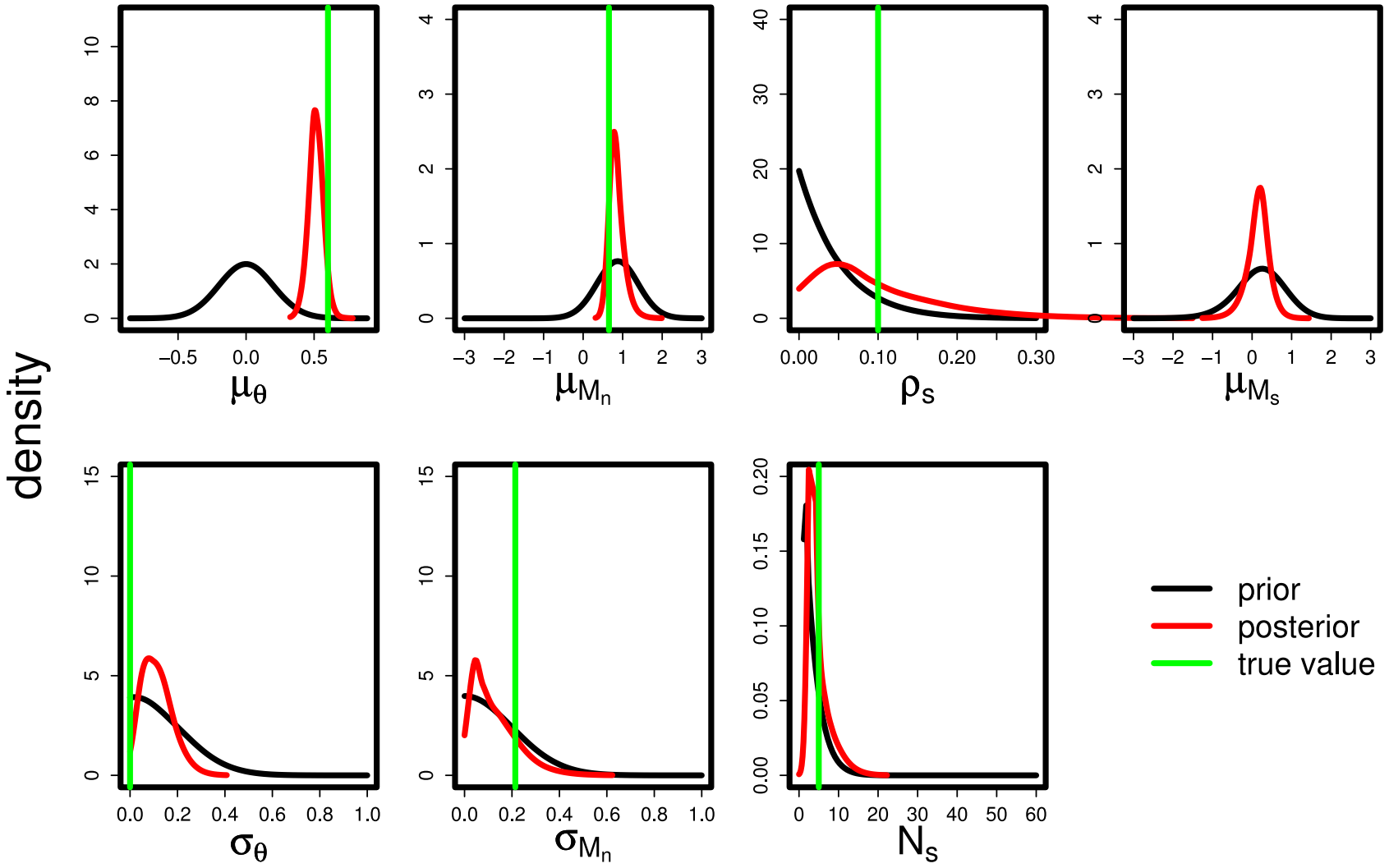
$$\Rightarrow M_{s_{ij}} < M_{n_j}$$

Example

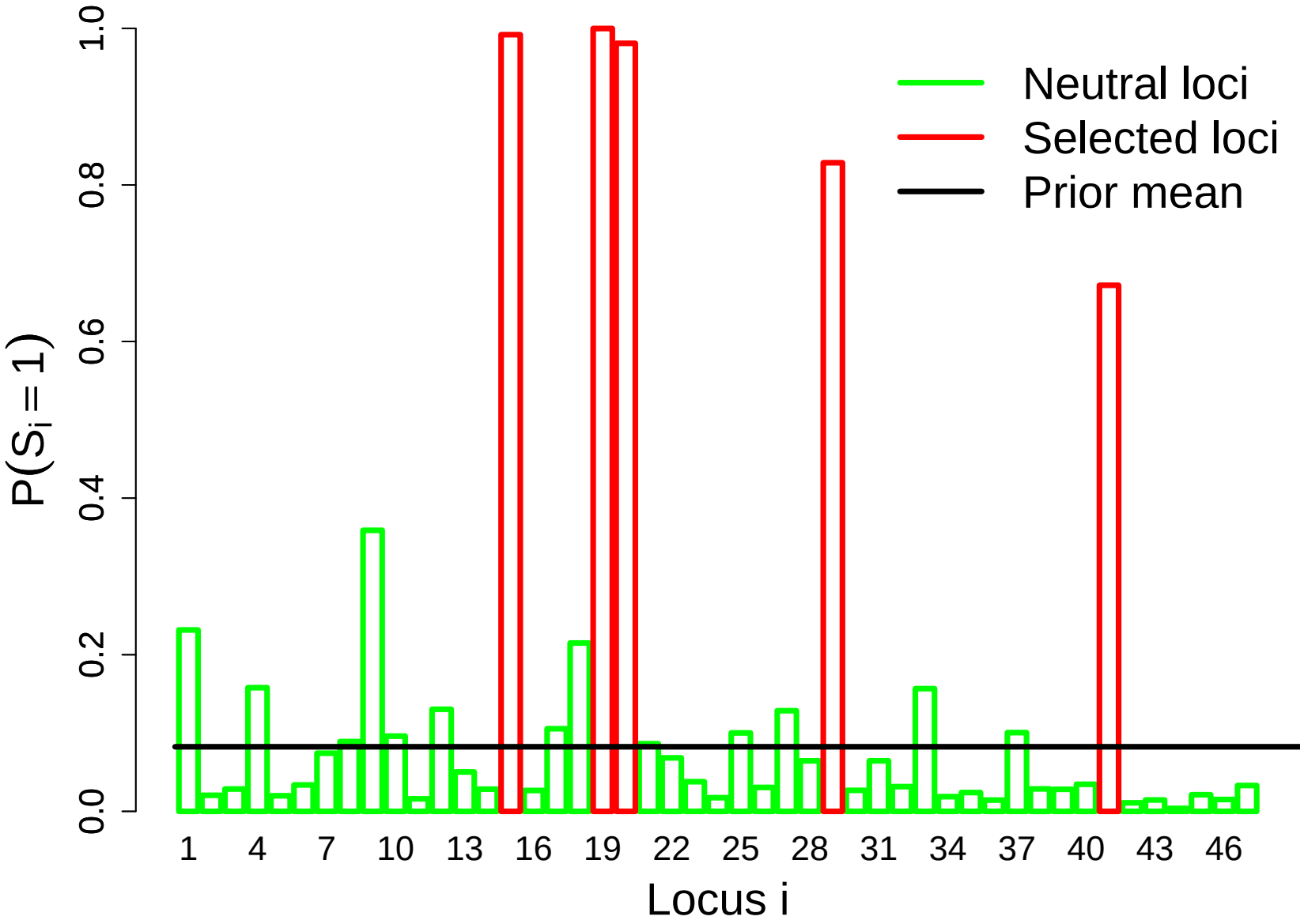
- 600 individuals
- 6 subpopulations
- 50 microsatellites
- 5 under selection



Example

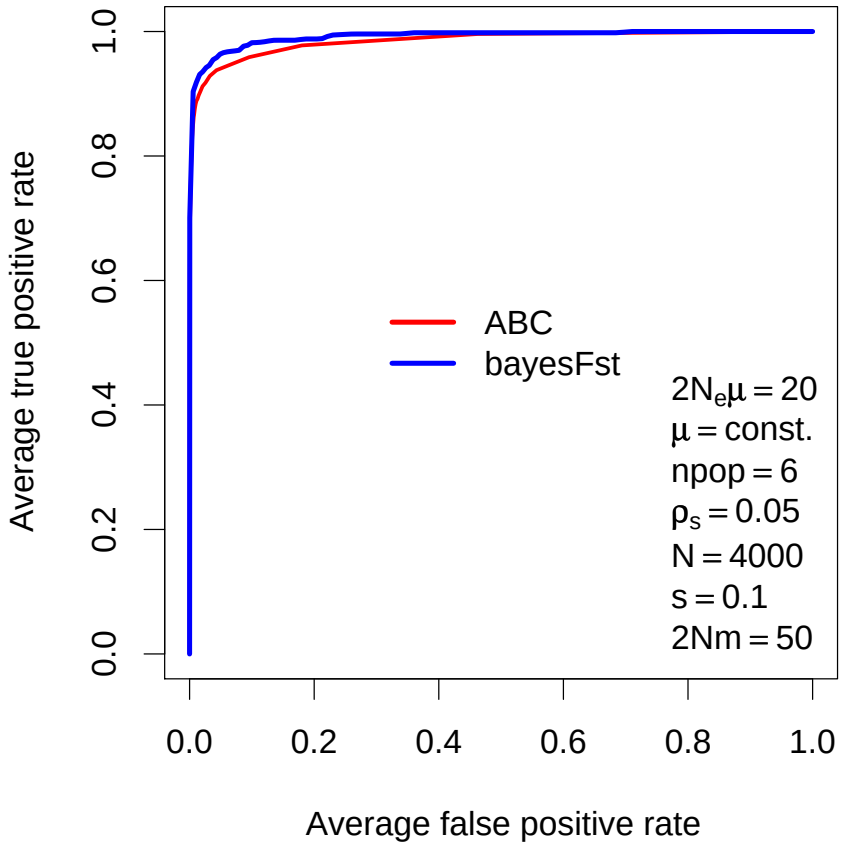

density
μ_θ
μ_{M_n}
ρ_s
μ_{M_s}
σ_θ
σ_{M_n}
N_s
prior
posterior
true value

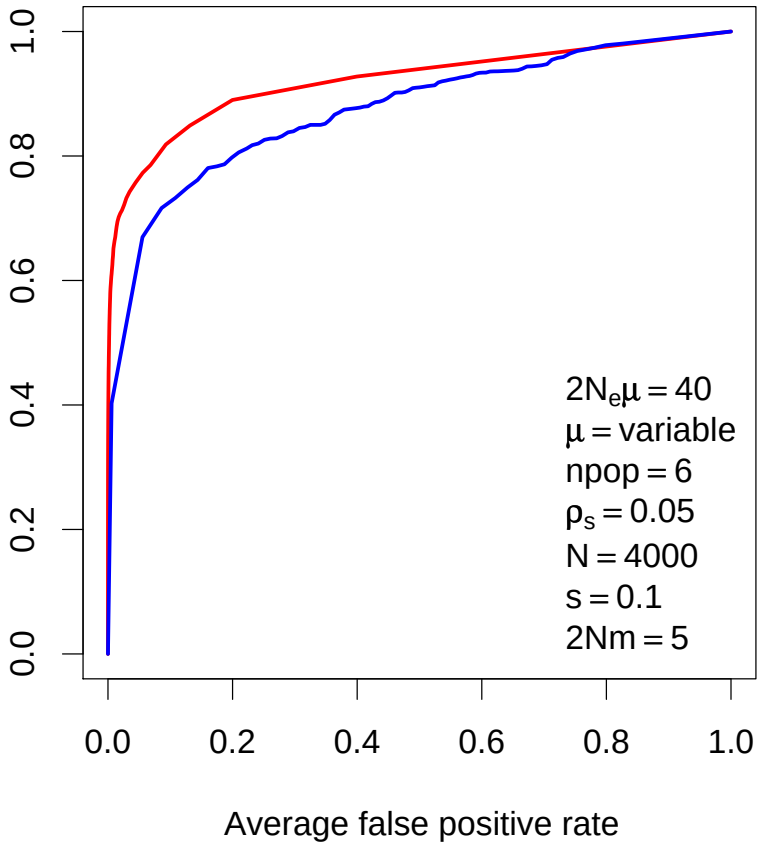
Example


Neutral loci
Selected loci
Prior mean
P(S_i = 1)
0.0
0.2
0.4
0.6
0.8
1.0
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
46
Locus i

ABC vs BayesFST

ROC curves





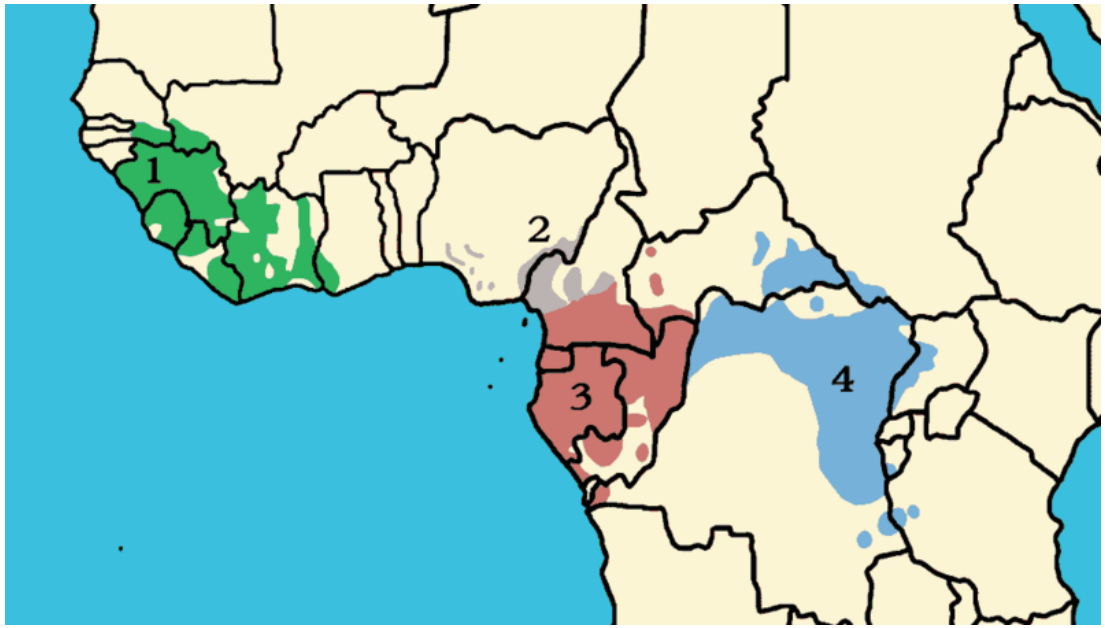
Analysis of Chimpanzee data

OPEN ACCESS Freely available online PLoS GENETICS

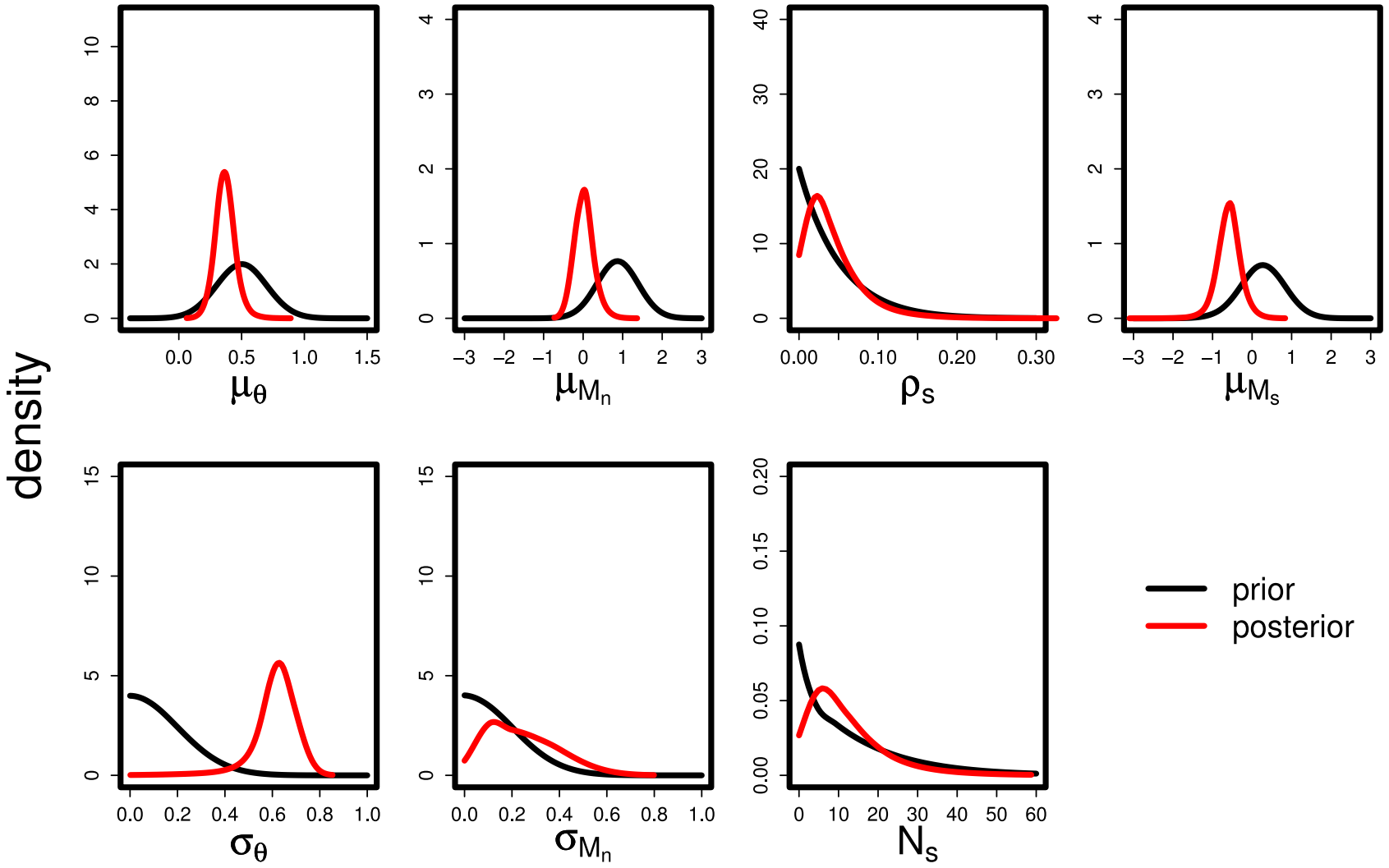
Genetic Structure of Chimpanzee Populations

Celine Becquet[1], Nick Patterson[2], Anne C. Stone[3], Molly Przeworski[1*], David Reich[2,4*]

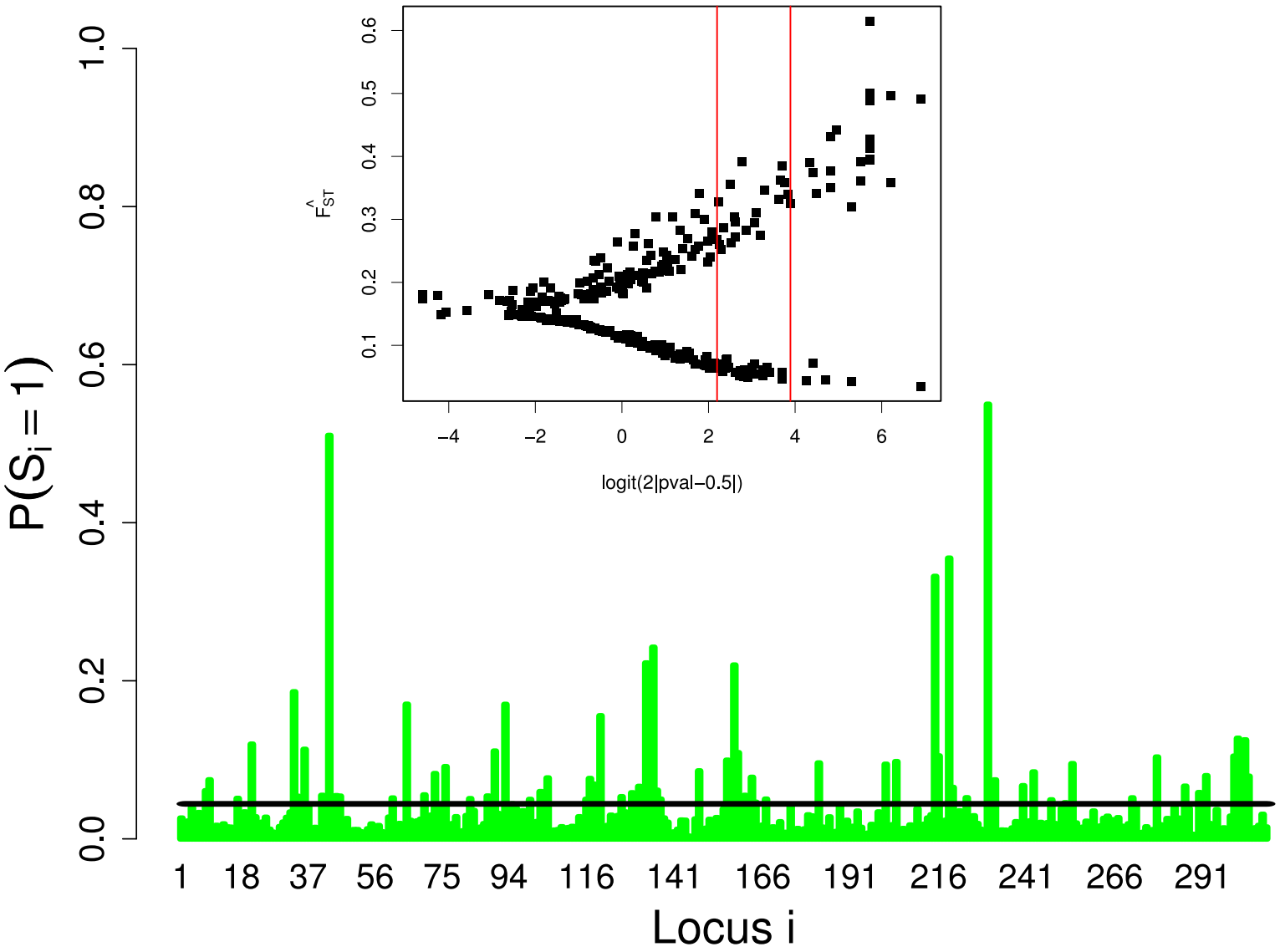
- 64 individuals
- 309 microsatellites
- 3 populations
  - Western
  - Central
  - Eastern



Analysis of Chimpanzee data



Analysis of Chimpanzee data



Software

The **abcselection** software is available from the authors on demand at eric.bazin@cirad.fr

Acknowledgements

This work was supported by a BBSRC grant (reference: BBSB12776) to Mark Beaumont and Kevin Dawson.

Rothamsted Research receives grant-aided support form the BBSRC (the Biotechnology and Biological Sciences Research Council of the United Kingdom).

Bibliography

- Basu D (1977) On the elimination of nuisance parameters. Journal of the American Statistical Association 72(358):355–366
- Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian Computation in Population Genetics. Genetics 162:2025–2035
- Kolmogorov AN (1942) Determination of the centre of dispersion and degree of accuracy for a linited number of observation. Izv. Akad. Nauk, USSR Ser. Mat.6:3–32
- Raiffa H, Schlaifer R (1961) Applied statistical decision theory. Harvard University Press, Cambridge, MA

Summary Statistics

locus-specific summary statistics

For each locus:

- Observed probability of non-identity in state of gene copies between populations, HB (Weir and Cockerham, 1984)
- Weir and Cockerhams estimator of FST
- Log variance in allele length between populations (Slatkin, 1995; Rousset, 1996).
- RST (Slatkin, 1995; Rousset, 1996)
- Variance in W&C FST estimated for individual alleles (microsatellite lengths).
- Proportion of pairwise comparisons between populations in which an allele is observed in at least one of the populations.
- Variance of within-population W&C estimator of FST (Weir and Hill, 2002).
- Variance of within-population RST.

8 summary statistics

Summary Statistics (continued)

symmetric summary statistics

To infer hyperparameters:

- The mean over loci of 8 summary statistics above.
- Variance over loci of 8 summary statistics above.
- Skew over loci of 8 summary statistics above.
- Kurtosis over loci of 8 summary statistics above.
- Covariance over loci of all 28 pairs of summary statistics.

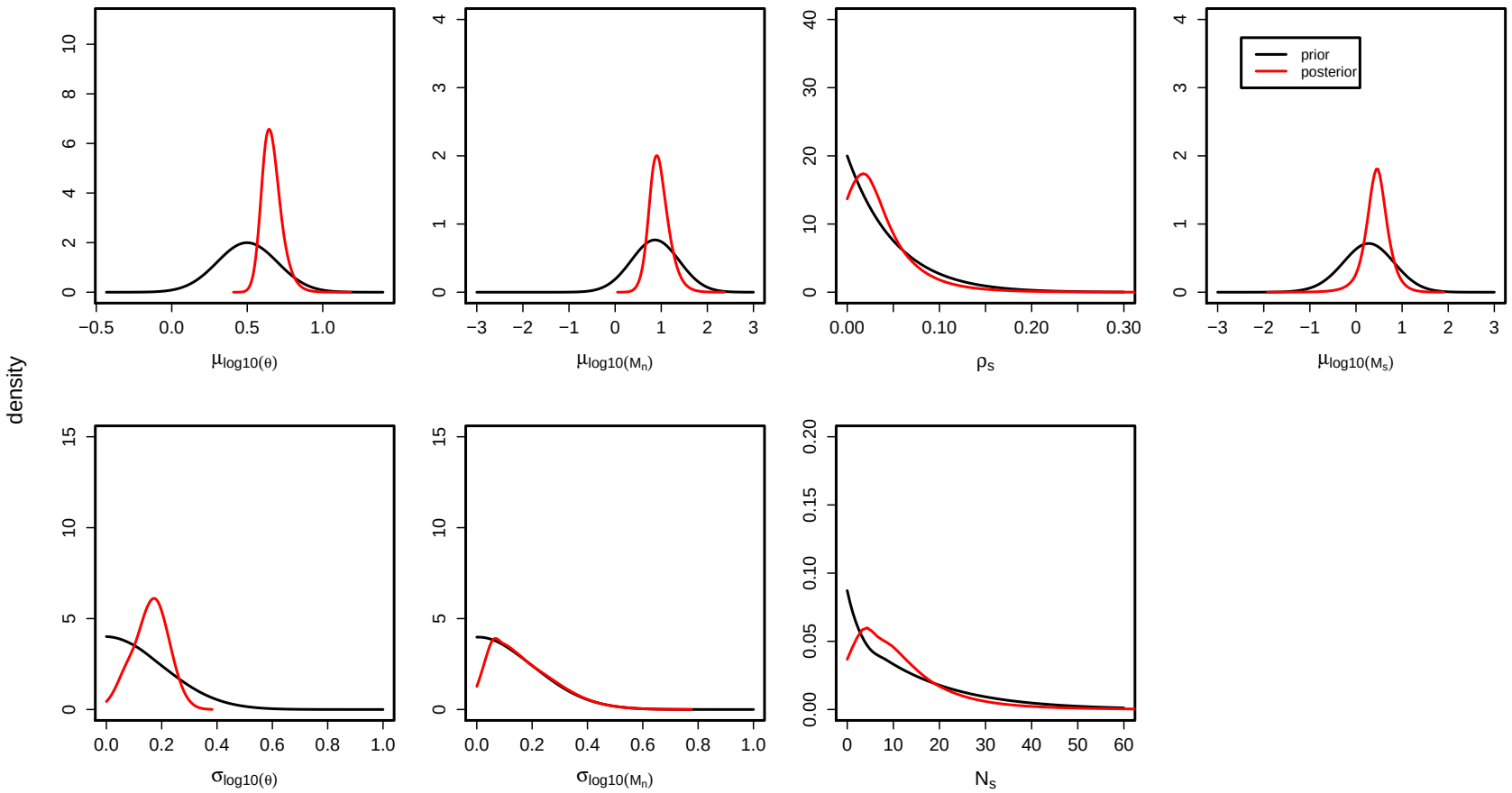
60 summary statistics.

Analysis of European data

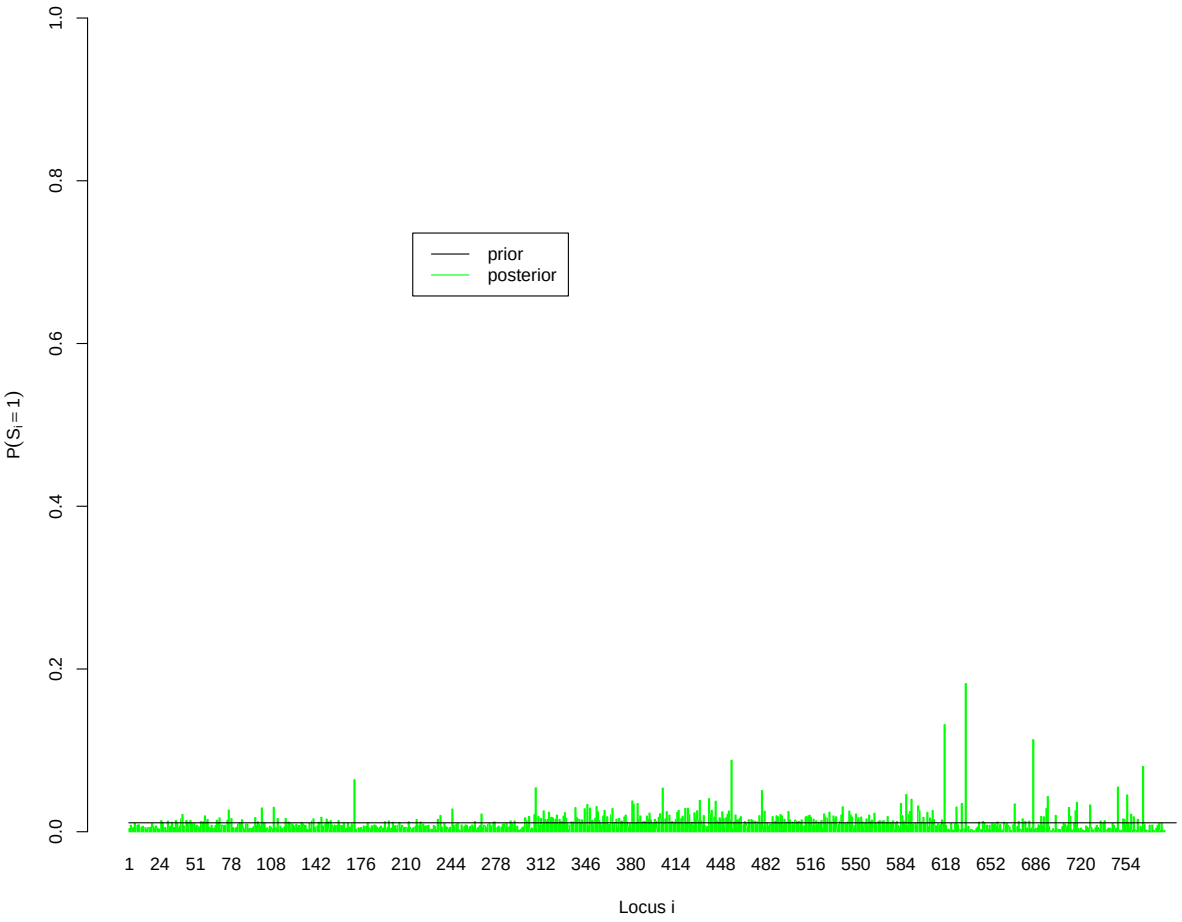
Genetic Structure of Human Populations
Noah A. Rosenberg, *et al.*
Science **298**, 2381 (2002);
DOI: 10.1126/science.1078311

- 160 individuals
- 783 microsatellites
- 8 populations
  - Orcadian
  - Adygei
  - Russian
  - Basque
  - French
  - Italian
  - Sardinian
  - Tuscan

Analysis of European data



Analysis of European data



The model

The likelihood function for our model has the form

$$p\left(X|\kappa,\lambda,\alpha\right) = \prod_{i=1}^{L} p(X_i|\kappa_i,\lambda_i,\alpha), \tag{7}$$

where $X = (X_1, \ldots, X_L)$, $X_i = X_{ij}$, and
$\alpha = (M_{n_1}, \ldots, M_{n_D}, \rho_s, \mu_M, \sigma_M, \mu_\theta, \sigma_\theta)$.
The locus-specific parameters are $(\kappa_i, \lambda_i) = (\theta_i, M_{i1}, \ldots, M_{iD}, S_i)$.
Here we choose to treat the S_i as the parameter of interest, so we define $\kappa_i = (\theta_i, M_{i1}, \ldots, M_{iD})$ and $\lambda_i = S_i$.