Supplementary material for Bioinformatics Application
Note: “Correcting for ascertainment bias in the inference
of population structure.”

Gilles Guillot 1*, Matthieu Foll 2,3

1 Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo. P.O. Box 1066 Blindern 0316 Oslo Norway. 2 Computational and molecular population genetics lab, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland. 3 Swiss Institute of Bioinformatics.

* Also at Gothenburg Stochastic Centre, Chalmers University of Technology, Gothenburg, Sweden and Department of Applied Mathematics and Computer Science, INRA/Agro-ParisTech, Paris, France.
A DETAILS ON THE MULTINOMIAL-DIRICHLET MODEL WITH CORRELATED ALLELE FREQUENCIES

We consider a dataset consisting of genotypes of $N$ individuals at $L$ unlinked loci. We assume that each individual has a known population of origin among a given set of $K$ populations. Populations exchange genes with a unique and common migrant pool (Balding, 2003), and each population $k$ receives migrants from the pool at rate $\lambda_k$. Each local population is characterised by a set of allele frequencies $\tilde{p}_{klj}$ (frequency of allele $j$ at locus $l$ in population $k$). The migrant gene pool is described through a set of allele frequencies denoted by $p_l = (p_{lj})$. Under this model, allele frequencies $\tilde{p}_{klj} | p, F_{STk}$ follow a Dirichlet distribution with parameter $p_l \times (1 - F_{STk}) / F_{STk}$. The coefficient $F_{STk}$, defined as $1/\lambda_k - 1$, measures how divergent each local population $k$ is from the metapopulation as a whole. Allele frequencies in present time populations are hence assumed to be correlated across populations. We also assume that populations might depart from Hardy-Weinberg equilibrium which is accounted for at the likelihood level through the introduction of inbreeding coefficients.

To complete the specification of the model, we place prior distributions on all parameter as follows:

$$F_{STk} \overset{i.i.d}{\sim} \text{Beta}(s_1, s_2) \quad (I)$$
$$F_{ISk} \overset{i.i.d}{\sim} \text{Beta}(t_1, t_2) \quad (II)$$
$$p_l = (p_{l1}, \ldots, p_{lJ_l}) \overset{i.i.d}{\sim} \text{Dirichlet}(a, \ldots, a) \quad (III)$$
$$\tilde{p}_{klj} | p, F_{STk} \overset{i.i.d}{\sim} \text{Dirichlet}(p \times (1 - F_{STk}) / F_{STk}) \quad (IV)$$

Denoting $\theta = (\phi, \psi)$ where $\phi = (F_{ST}, F_{IS})$ and $\psi = (p, \tilde{p})$, the likelihood is as a product of multinomial distributions over all individuals and loci. The generic terms of the likelihood for genotypes $(\alpha, \alpha)$ and $(\alpha, \beta)$ at locus $l$ are respectively:

$$l(\alpha, \alpha | \theta) = \tilde{p}_{kl\alpha} + F_{ISk}\tilde{p}_{kl\alpha} \quad (VI)$$
$$l(\alpha, \beta | \theta) = 2\tilde{p}_{kl\alpha}\tilde{p}_{kl\beta}(1 - F_{ISk}) \quad (VII)$$

This model is summarised as directed acyclic graph on Fig. 1

![Directed acyclic graph of model considered. Circles denote unknown parameters to be inferred, squares denote data.](image)
B PROBABILITY DISTRIBUTIONS ACCOUNTING FOR THE ASCERTAINMENT PROCESS

B.1 Full joint probability density induced by ascertainment process

There are various ways to discover polymorphic loci. We describe here a simple ascertainment policy. The presentation borrows much from Foll et al. (2008). We assume that there is a fixed set of individuals. For a given locus, these individuals are genotyped. Then it is checked that genotypes at this locus display a minimum level of polymorphism. This minimum level can be for instance that the minority allele is observed at least once in the sample. This minimum level is arbitrary and can vary across experiments. We formally write $n \in A$ if the sample satisfies this condition of the ascertainment policy. If this condition is fulfilled, then the locus will be part of the set of loci analysed, otherwise, this locus is disregarded. The ascertainment policy can be summarised as:

1. sample $\phi$ from $f(\phi)$
2. sample $\psi|\phi$ from $g(\psi|\phi)$
3. sample $n|\phi, \psi$ from $h(n|\phi, \psi)$
4. if $n \notin A$, goto 2

Under this process, $(\phi, \psi, n)$ is sampled from the prior-likelihood model but subject to the condition $n \in A$. The joint probability distribution $\pi_c(\phi, \psi, n)$ is hence proportional to $\pi(\phi, \psi, n) I_{n \in A}$, therefore,

$$\pi_c(\phi, \psi, n) = \frac{\pi(\phi, \psi, n) I_{n \in A}}{g(\psi|\phi) h(n|\phi, \psi) I_{n \in A} d\psi dn} = f(\phi) g(\psi|\phi) h(n|\phi, \psi) I_{n \in A}/K_\phi$$

(B.1)

B.2 Distribution of $\phi$ under the ascertainment process

Integrating $\pi_c(\phi, \psi, n)$ over $\psi$ and $n$, we get

$$\pi_c(\phi) = \int \pi_c(\phi, \psi, n) d\psi dn$$

$$= \int f(\phi) g(\psi|\phi) h(n|\phi, \psi) I_{n \in A}/K_\phi d\psi dn$$

$$= f(\phi)$$

(B.2)

B.3 Distribution of $(\phi, \psi)$ under the ascertainment process

$$\pi_c(\phi, \psi) = \int \pi_c(\phi, \psi, n) dn$$

$$= \int f(\phi) g(\psi|\phi) h(n|\phi, \psi) I_{n \in A}/K_\phi dn$$

$$= f(\phi) g(\psi|\phi) H_{\phi,\psi}/K_\phi$$

(B.3)

where

$$H_{\phi,\psi} = \int h(n|\phi, \psi) I_{n \in A} dn$$

(B.4)

B.4 Distribution of $(\psi|\phi)$ under the ascertainment process

$$\pi_c(\psi|\phi) = \frac{\pi_c(\phi, \psi)}{\pi_c(\phi)}$$

$$= \frac{f(\phi) g(\psi|\phi) H_{\phi,\psi}/K_\phi}{f(\phi)}$$

$$= g(\psi|\phi) H_{\phi,\psi}/K_\phi$$

(B.5)
B.5 Likelihood \( \pi_c(n|\phi, \psi) \) under the ascertainment process

\[
\pi_c(n|\phi, \psi) = \frac{f(\phi) g(\psi|\phi) h(n|\phi, \psi) I_{n \in A}/K_{\phi}}{f(\phi) g(\psi|\phi) H_{\phi, \psi}/K_{\phi}} = h(n|\phi, \psi) I_{n \in A}/H_{\phi, \psi}
\]  

(XIII)

B.6 Distribution of \((\psi, n|\phi)\) under the ascertainment process

\[
\pi_c(\psi, n|\phi) = \frac{1}{K_{\phi}} f(\phi) g(\psi|\phi) h(n|\phi, \psi) I_{n \in A}/f(\phi)
\]

\[
= \frac{1}{K_{\phi}} g(\psi|\phi) h(n|\phi, \psi) I_{n \in A}
\]  

(XIV)

C JUSTIFICATION OF ESTIMATE OF RATIO OF NORMALISING CONSTANTS \( K_{\phi}/K_{\phi^*} \)

To understand what the estimate of ratio of normalising constants should be, it is illuminating to rewrite the joint distribution \( \pi_c \) in Eq. (VIII) as

\[
\pi_c(\phi, \psi, n) = f(\phi) \times
\]

\[
g(\psi|\phi) H_{\phi, \psi}/K_{\phi}
\]

\[
h(n|\phi, \psi) I_{n \in A}/H_{\phi, \psi}
\]

(XV)

(XVI)

(XVII)

where \( H_{\phi, \psi} = \int \pi(n|\psi, \phi) I_{n \in A} dn \). This makes appear the components of the prior distribution and the likelihood, all appropriately corrected for censoring. Obviously, \( H_{\phi, \psi} \) cancels out; but importantly, it reveals that the unknown normalising constant \( K_{\phi} \) arises from the joint contribution of factors (XVI) and (XVII). This situation contrasts with the setting initially described by Møller et al. (2006) and Murray et al. (2006) where the unknown normalising constant was supposed to arise from the likelihood only.

In the present case, this estimate of \( K_{\phi}/K_{\phi^*} \) is therefore \( g(\psi|\phi) h(m|\phi, \nu)/g(\psi|\phi^*) h(m|\phi^*, \nu) \), where \( \nu \) and \( m \) are auxiliary variables sampled jointly from \( g(\nu|\phi^*) h(m|\phi^*, \nu) I_{m \in A}/K_{\phi^*} \). We have indeed the importance sampling identity:

\[
E_{\nu, m} \left[ \frac{g(\nu|\phi) h(m|\phi, \nu)}{g(\nu|\phi^*) h(m|\phi^*, \nu)} \right] = \int g(\nu|\phi) h(m|\phi, \nu)/g(\nu|\phi^*) h(m|\phi^*, \nu) I_{m \in A}/K_{\phi^*} dv dm
\]

(XVIII)

\[
= \int g(\nu|\phi) h(m|\phi, \nu) I_{m \in A}/K_{\phi^*} dv dm
\]

(XIX)

\[
= K_{\phi}/K_{\phi^*}
\]

(XX)
D DETAILS ON THE EXAMPLE OF NUMERICAL IMPLEMENTATION: MIXING OF THE CHAIN

MCMC algorithms for models of population genetics structure have behaviours that varies a lot depending on the particular model used and the size of the dataset (number of individuals \(n\) and number of loci \(L\)). Convergence can become problematic in case of large datasets if the cluster membership variable is unknown. We stress that convergence or mixing issues concern potentially any Monte Carlo inference algorithm in population genetics (and beyond) and are not specific at all to the updating scheme we propose here. In the particular case described in section 4 "Results in a toy example", the cluster membership variable is known and we did not observe any problem of lack of convergence due to multi-modality. The mixing of the chain was also found to be good, a fact also previously reported by Murray et al. (2006). It is illustrated by the trace of a typical run in Fig. 2.

![Example of MCMC simulations with the algorithm proposed for \(F_{ST}\) parameters of two populations. The run consists of 50000 iterations with a thinning of 10 iterations. Red lines depict true values, green lines depict estimated values.](image)

**Fig. 2.** Example of MCMC simulations with the algorithm proposed for \(F_{ST}\) parameters of two populations. The run consists of 50000 iterations with a thinning of 10 iterations. Red lines depict true values, green lines depict estimated values.
E DESCRIPTION OF THE COMPUTER PROGRAM

The computer code developed to carry out simulations in this simplified model is written in C and Fortran and is embedded in an R interface. It is available from folk.uio.no/gillesg/AscB/. This program consists of a self explanatory R script which can be identified through its name (demo.R). The other programs are C and Fortran subroutines internally called by the R script. The Fortran program sub.f is a subroutine that simulates parameters and data (with or without the censoring incurred by the ascertainment process). The C program wrapper.c is used for calling the R random number generator from Fortran. The program demo.R contains R commands for automatically building the shared object file and loading it into R.

REFERENCES