Supplementary material for:
Using AFLP markers and the Geneland program for the inference of population genetic structure

Gilles Guillot † and Filipe Santos ↑

February 12, 2010

1 Dealing with genotype ambiguity in Bayesian inference of population structure

We consider a data-set consisting of genotypes of \( n \) individuals at \( L \) unlinked loci. We assume that each individual has a known ancestry in one of \( K \) populations denoted by \( c_i \). Each population is characterised by a set of allele frequencies \( f_{klj} \) (frequency of allele \( j \) at locus \( l \) in population \( k \)) assumed to follow a Dirichlet distribution: \( f_{kl} = (f_{kl1}, ..., f_{klL}) \overset{i.i.d.}{\sim} \text{Dirichlet}(1, ..., 1) \). Each cluster is assumed to be at Hardy-Weinberg equilibrium at each locus with linkage equilibrium between loci. Namely, denoting by \( y_i = (y_{i,l})_{l=1}^L \) the true multi-locus genotype of individual \( i \), we have

\[
\begin{align*}
\mathbb{P}(y_{i,l} = a, a | c_i = k, f_k) &= f_{kla}^2 \\
\mathbb{P}(y_{i,l} = a, A | c_i = k, f_k) &= 2f_{kla}f_{kA}
\end{align*}
\]

and

\[
\mathbb{P}(y | c_i = k, f_k) = \prod_{l=1}^L \mathbb{P}(y_{i,l} | c_i = k, f_k)
\]

Following Wasser et al (2004), we introduce the observed genotype \( z_i \), which in the case of AFLP markers differs from the true unobserved genotype \( y_i \). Denoting by \( a \) the recessive allele, we have: \( z_{i,l} = A \) if \( y_{i,l} = (A, A) \) or \( y_{i,l} = (A, a) \) and \( z_{i,l} = a \) if \( y_{i,l} = (a, a) \).

The set of true genotypes is unknown but this uncertainty can be accounted for by Gibbs sampling at each iteration as the full conditional distribution of \( y_{i,l} \) is known as:

\[
\mathbb{P}(y_{i,l} = A, A | z_{i,l} = A, c_i = k, f_k) = \frac{f_{kIA}}{f_{kIA} + 2f_{kla}}
\]

*to whom correspondence should be addressed, Department of Informatics and Mathematical Modelling, Technical University of Denmark, Richard Petersens Plads, Bygning 305, 2800 Lyngby, Denmark;
†Centre de Biologie et de Gestion des Populations, Institut National de la Recherche Agronomique, 34988 Montferrier-sur-Lez, France.
2 Detail of simulations

2.1 Simulation plan

We simulated datasets from the defaults prior-likelihood model assumed in Geneland (assuming spatially and genetically structured populations, see (Guillot et al., 2005; Guillot, 2008; Guillot et al., 2010) for details). Each data-set consisted of $n = 100$ individuals belonging to one of $K = 2$ clusters. We investigated various levels of differentiations between the two clusters reported by the pairwise $F_{ST}$ value. We considered genotypes at bi-allelic loci for dominant then codominant markers. We investigated genotypes at $L = 10, 20, 50, 200, 500$ markers. Each value of $L$ was investigated through a batch of $N = 200$ independent datasets.

2.2 Setting for MCMC computations

For each data-set, we launched a single MCMC run of 200000 iterations with a thinning of 200 and we discarded 100000 iterations for burn-in. We treated the number of clusters as unknown setting the maximum number of clusters to $K_{\text{max}} = 10$. Each MCMC run was post-processed according to the procedure described by Guillot (2008) to get of potential label switching issues.

2.3 Detail of simulation results

[Figure 1 about here.]

[Figure 2 about here.]
References


41 www2.imm.dtu.dk/~gigu/Geneland/Geneland-Doc.pdf.

List of Figures

1 Clustering error (y-axis) as a function of pairwise $F_{ST}$ (x-axis), for datasets simulated with various values of the number of loci $L$ from the default GENELAND prior-likelihood model. Each point represents a data-set consisting of $n = 100$ individuals belonging to $K = 2$ clusters with genotypes at unlinked bi-allelic loci. The colour and shape of the symbol represent the number of clusters inferred ($\circ$: one cluster, $\triangle$: two clusters i.e. correct result, $+$: three clusters, $\times$: four clusters, $\diamond$: five clusters). Left column: dominant markers, right column: codominant markers.

2 Clustering error (y-axis) as a function of pairwise $F_{ST}$ (x-axis), for datasets simulated with various values of the number of loci $L$ from the default GENELAND prior-likelihood model. Each point represents a data-set consisting of $n = 100$ individuals belonging to $K = 2$ clusters with genotypes at unlinked bi-allelic loci. The colour and shape of the symbol represent the number of clusters inferred ($\circ$: one cluster, $\triangle$: two clusters i.e. correct result, $+$: three clusters, $\times$: four clusters, $\diamond$: five clusters). Left column: dominant markers, right column: codominant markers.
Figure 1: Clustering error (y-axis) as a function of pairwise $F_{ST}$ (x-axis), for datasets simulated with various values of the number of loci $L$ from the default GENELAND prior-likelihood model. Each point represents a data-set consisting of $n = 100$ individuals belonging to $K = 2$ clusters with genotypes at unlinked bi-allelic loci. The colour and shape of the symbol represent the number of clusters inferred (○: one cluster, △: two clusters i.e. correct result, +: three clusters, ×: four clusters, ◦: five clusters). Left column: dominant markers, right column: codominant markers.
Figure 2: Clustering error (y-axis) as a function of pairwise $F_{ST}$ (x-axis), for datasets simulated with various values of the number of loci $L$ from the default GENELAND prior-likelihood model. Each point represents a data-set consisting of $n = 100$ individuals belonging to $K = 2$ clusters with genotypes at unlinked bi-allelic loci. The colour and shape of the symbol represent the number of clusters inferred (○: one cluster, △: two clusters i.e. correct result, +: three clusters, ×: four clusters, ◊: five clusters). Left column: dominant markers, right column: codominant markers.