A computer program to simulate multilocus genotype data with spatially auto-correlated allele frequencies.

Supplementary material.
G. Guillot and F. Santos

September 14, 2008

1 Statistical details on model proposed.

We summarise here our model in terms of a probability density for the parameters and a likelihood for the genotypes given the parameters. They are used explicitly to draw simulations.

1.1 Prior probability distribution of parameters

For $n$ individuals, the set of parameters involved in this model is as follows: $K$ the unknown number of populations, $\lambda$ the rate of the Poisson process generating the polygons of the populations territories, $u_1, ..., u_m$ the locations of the centres of the polygons, $c_1, ..., c_m$ the population membership of each such polygons, $f_{ilj}$ the allele frequency at site $s_i$ for allele $j$ of locus $l$, and $\beta$, the common scale parameter of the hidden Gaussian fields involved to represent the allele frequency surfaces. The specification of the prior itself does not involves directly the frequencies but the Gaussian values, namely $g_{ilj}$ the value of the Gaussian field at site $s_i$ (site of observation of individual $i$) for allele $j$ of locus $l$.

We assume:

- uniform prior on $K$ on $\{1, ..., K_{max}\}$
- uniform prior on $\lambda$ on $[0, \lambda_{max}]$
- Poisson distribution $P(\lambda)$ for $m$
- uniform distribution of $u_1, ..., u_m$ within $D$ conditional on $m$
- independent uniform distribution on $U(\{1, ..., K\})$ of population memberships ($c_1, ..., c_m$)
- uniform distribution of $\beta$ on $[0, \beta_{max}]$
- multivariate normal distribution \( N \left( 0, \Sigma^{(k)}_\beta \right) \) of each vector \( (g^{(k)}_{lj})_{i=1,...,n} \) where \( g^{(k)}_{lj} \) is the set of Gaussian values at locus \( l \) and allele \( j \) for individuals in population \( k \) and \( \Sigma^{(k)}_\beta \) is the corresponding covariance matrix for pairs of observations sites derived from the stable model of equation ??.

Denoting \( \theta = (K, \lambda, m, u, c, \beta, g) \), with \( u = (u_1, ..., u_m) \), \( c = (c_1, ..., c_m) \), and omitting normalising constants, we end-up with a prior probability for this set of parameters given by:

\[
\pi(\theta) \propto \lambda^m \frac{1}{m! \left| D \right|^m} \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{j=1}^{J_l} \frac{1}{\left| \Sigma^{(k)}_\beta \right|} \exp \left[ -\frac{1}{2} g^{(k)}_{lj} \Sigma^{(k)}_\beta^{-1} g^{(k)}_{lj} \right]
\]

(1)

where \( n_k \) is the cardinal of population \( k \).

### 1.2 Likelihood

Assuming that Hardy-Weinberg equilibrium holds locally as explained above, and linkage equilibrium between loci, the likelihood for parameter \( \theta \) and observed genotypes \( z = (z_1, ..., z_n) \) for \( n \) individuals is given by

\[
L(z|\theta) = \prod_{i=1}^{n} f_{id_{\alpha_{il}}} f_{id_{\beta_{il}}} (2 - \delta_{\alpha_{il}}^{3/2})
\]

(2)

where \( z_{il} = \{\alpha_{il}, \beta_{il}\} \), \( f_{id} \) is the frequency obtained from transformation of the Gaussian fields into the Dirichlet fields and \( \delta_{uv} \) is the Kronecker symbol (1 iff \( u = v \), 0 otherwise).

### 2 Details of result section

#### 2.1 Factors affecting the accuracy in case of IBD

In order to get a better insight into the factors that are relevant to the observed decrease in accuracy due to IBD, we have analysed an extra set of simulated files where \( K \) was equal to 2 in all cases (but treated as unknown in the inferences as before). As shown on Fig. 1, there is a clear dependence between ERCA\textsubscript{i} and the scale parameter \( \beta \) and hence with the level of differentiation within IBD population \( D_W \). In particular for the 50 % of files with the highest \( \beta \) (lowest \( D_W \)) values, the number of populations \( K \) is perfectly estimated (\( K = \hat{K} \)), and the average ERCA\textsubscript{i} on these files is about 15%. For datasets with stronger IBD (say \( \beta < 1.5 \)) some errors occur, with larger magnitude as the strength of IBD increases (i.e. \( \beta \) decreases). Interestingly enough, it also appears that there is no marked dependence of ERCA on the level of local differentiation between populations \( D_B \) (Fig. 1). This result indicates that, in the case of strong IBD, our algorithm
fails to retrieve the true structure, whatever the magnitude of the jump of frequencies between populations.

We also launched an extra set of simulations under the IBD model with 200 individuals spatially regularly sampled with $K$ uniform between 1 and 4 but where the number of locus was $L = 100$. The results were strictly similar to those obtained with $L = 10$ (a proportion of runs where $K \neq \hat{K}$ of 28.6, a proportion of runs where $\hat{K} > K$ of 28.4, ERCA$_{i}=12.8$ and ERCA$_{p}=15.7$). This indicates that increasing the number of loci does not improve our inferences (in contrast with the behaviour observed with null allele data).

2.2 Illustration of the effect of the use spatially clumped samples

The loss of accuracy is illustrated in Fig. 2 (second panel from top) which shows the result of an inference on a simulated panmictic data-set with irregularly sampled individuals. The true number of populations and population memberships are perfectly inferred ($K = \hat{K}$, ERCA=0), but the shape of population domains are just seen by the data-set up to the resolution of the spatial sampling, which is coarse in the present case.
Figure 1: Relationships for IBD populations between the scale parameter $\beta$ and three error statistics, according to various levels of local differentiation around the borders of populations as measured by the statistic $D_B$ (see Materials and Methods section for details). The colour codes the local differentiation as three levels: red $D_B < 0.092$, green $0.092 < D_B < 0.097$, blue $D_B > 0.097$ (the three classes being of equal probability under our prior).
Figure 2: Illustration of the effect of spatial sampling and IBD on inferences: (a) true tessellation; (b) panmictic data with spatially irregular sampling; (c) weak IBD ($\beta = 2.5$) with spatially regular sampling; (d) strong IBD ($\beta = 0.5$) with spatially regular sampling; (e) strong IBD with irregular spatial sampling.