

SUPPLEMENTARY MATERIAL: DETAIL ON SIMULATION STUDY

The standard model implemented in the package Geneland produces spatially organised panmictic populations. This model assumes that the sampled domain contains K populations at HWE and whose geographical spread is simple in the sense that each population territory can be approximated by the union of a small number of polygonal shaped domains. The number of populations K is treated as unknown and is estimated. The geographical parameters to be inferred consist of the number of polygons, the location of their centre and of their population membership. Although not of direct interest, for the purpose of inference, the allele frequencies in each population are also introduced and inferred jointly with the geographical parameters. See [Guillot et al., 2005] for a more formal description and details.

All the simulations described hereafter were located on the squared spatial domain $[0, 1] \times [0, 1]$. The genotypes were simulated for $N = 200$ individuals at 10 loci with 10 alleles per locus. This corresponds to common features for most actual microsatellite data sets produced in non-model species. The individuals were located regularly (sampled uniformly and independently on $[0, 1] \times [0, 1]$). For each simulation condition, we drew a set of 500 files. The number of populations K were sampled in a uniform distribution on $\{1, 2, 3, 4\}$. Because our aim here is to study spatially structured data, we simulated data sets where the number of polygons m in the hidden tessellation giving the population membership was sampled in a uniform distribution on $\{1, \dots, 15\}$. It is important to note that the number of individuals per population is random, as it depends on the spatial location of borders between populations. In addition, the total number N of individuals sampled is set to 200 whatever the value of K . These two features make our simulation scheme more realistic than the one previously used in [Guillot et al., 2005] where the total number of simulated individuals was proportional to the number of populations. As a reference data set, we draw simulations from Geneland by sampling allele frequencies in different populations from independent Dirichlet distributions. By processing in this way, population genetic differentiation as measured by pairwise F_{ST} has approximately a symmetric empirical distribution with quartiles values equal to 0.04 and 0.16 respectively and a mode at 0.095. It hence spans a broad range of F_{ST} values from weak to marked levels of genetic differentiation.

At each locus, the genotypes of individuals homozygous for a randomly chosen (null) allele become missing data while genotypes heterozygous for the chosen (null) allele become homozygous for the non-chosen allele. It is worth stressing here that by proceeding in this way we consider that null alleles correspond to variation in the nucleotide sequences of flanking regions of molecular markers (e.g. microsatellites) that prevent the primer annealing to template DNA during amplification of the locus by PCR. With regards to microsatellites, other possible causes of null alleles include the preferential amplification of short alleles (due to inconsistent DNA template quality or quantity) or slippage during PCR amplification [Gagneux et al., 1997, Shinde et al., 2003]. These technical problems associated with PCR amplification are not considered here. For each simulated data set without null allele, we created three data sets with null alleles: (i) a data set with a single null allele at two of the ten loci only (hence with an average of 2% of null alleles), (ii) a data set with a single null allele at every locus (and hence with an average of

10% of null alleles), and (iii) a data set with two null alleles at every locus (and hence with an average of 20% of null alleles).

MCMC computations included 50000 iterations with a burnin of 25000 iterations and a thinning of 50 iterations. The minimum and maximum numbers of populations considered in the first run were $K_{min}=1$ and $K_{max}=10$, respectively. Using larger values of K_{max} did not affect the result of the inferences (results not shown). The accuracy in the inference of K was assessed by computing the proportions of runs for which $\hat{K} \neq K$ and $\hat{K} > K$, respectively. The accuracy in terms of inference of population membership was assessed through the error rate in co-assignment defined as:

$$ERCA = \frac{1}{N(N-1)/2} \sum_{i \neq j}^N I_{\{x_{ij} \neq \hat{x}_{ij}\}} \quad (1)$$

where x_{ij} (resp. \hat{x}_{ij}) is the true (resp. estimated) clustering matrix (i.e. $x_{ij} = 1$ whenever individuals i and j belong to the same population, 0 otherwise.) This computes the proportion of pairs of individuals belonging to the same population that were not correctly co-assigned by the inference algorithm. It has the advantage over a statistic referring to single individuals in that it is insensitive to the labelling of populations. The ERCA statistics range between 0 and 1. The way a particular value of ERCA should be interpreted actually depends on the true and inferred numbers of populations as it tends in particular to be lower for large K than for small K . ERCA values will hence not be interpreted as absolute values but will be used for comparing different simulation scenarios.