

Inference of structure in subdivided populations at low levels of genetic differentiation. The correlated allele frequencies model revisited.

Gilles Guillot

Centre for Ecological and Evolutionary Synthesis Department of Biology, University of Oslo, P.O. Box 1066 Blindern, 0316 Oslo Norway and INRA, Applied Mathematics Department, Paris, France.

Received on April 11 2008; revised on June 17 2008; accepted on August 7 2008.

Associate Editor: Dr Alex Bateman

ABSTRACT

Motivation: This paper considers the problem of estimating population genetic subdivision from multilocus genotype data. A model is considered to make use of genotypes and possibly of spatial coordinates of sampled individuals. A particular attention is paid to the case of low genetic differentiation with the help of a previously described Bayesian clustering model where allele frequencies are assumed to be *a priori* correlated. Under this model, various problems of inference are considered, in particular the common and difficult, but still unaddressed, situation where the number of populations is unknown.

Results: A Markov chain Monte-Carlo algorithm and a new post-processing scheme are proposed. It is shown that they significantly improve the accuracy of previously existing algorithms in terms of estimated number of populations and estimated population membership. This is illustrated numerically with data simulated from the prior-likelihood model used in inference and also with data simulated from a Wright-Fisher model. Improvements are also illustrated on a real dataset of eighty-eight wolverines (*Gulo gulo*) genotyped at ten microsatellites loci. The interest of the solutions presented here are not specific to any clustering model and are hence relevant to many settings in population genetics where weakly differentiated populations are assumed or sought.

Availability: The improvements implemented will be made available in version 3.0.0 of the R package Geneland. Informations on how to get and use the software are available from

<http://folk.uio.no/gillesg/Geneland.html>.

Supplementary material:

<http://folk.uio.no/gillesg/CFM/SuppMat.pdf>

Contact: gilles.guillot@bio.uio.no

ability to make inference in models with unprecedented complexity. The simplest model to perform clustering consists in assuming the presence of a known number of populations at Hardy-Weinberg equilibrium and unlinked loci (HWLE). In connection with previous ideas from Smouse et al. [1990], Foreman et al. [1997] and Roeder et al. [1998]. This is the approach taken in the work of Pritchard et al. [2000]. The question of estimating the number of populations K accurately might not be so crucial. Indeed, as pointed out by Balding [2006], “a population is mostly an intellectual construct that imperfectly reflects the reality”. See [Waples and Gaggiotti, 2006] for an ecological perspective and [Robert and Casella, 2004] for a statistical perspective. However, using clustering methods to get an insight about the genetic structure will require in practice an input about the number of sought populations. Hence, the question of estimating K in a formal statistical model has been addressed by Dawson and Belkhir [2001] within a full Bayesian inference scheme, then by Corander et al. [2003], Guillot et al. [2005], Pella and Masuda [2006] and Huelsenbeck and Andolfatto [2007], with different model and/or algorithm variants.

In some of the works mentioned above, it is assumed that allele frequencies are statistically independent across populations. This is a computationally convenient assumption. It might however often reflect the statistical distribution of allele frequencies inaccurately. Indeed, it can be observed on real data, and also under various biological models that allele frequencies tend to display correlation across populations. This can be accounted for by a model where it is assumed that present time populations result from the split of a putative ancestral population into K populations, which then randomly and independently drifted until present time. The likelihood function for this model is known. It is the sampling distribution from the time-dependent solution of the diffusion equation under pure drift, first derived by Kimura and Crow [1956]. The sampling distribution can be obtained using coalescent theory, and has been applied for inference by Nielsen [1998], O’Ryan et al. [1998] and Saccheri et al. [1999]. This model is computationally intensive. To account for correlation of allele frequencies across populations, an alternative and computationally more tractable model described by Nicholson et al. [2002] can be used. This model is based on the Dirichlet distribution. The use of Dirichlet distribution is justified by analogy with the infinite island (or continent-island) model

1 BACKGROUND

Bayesian clustering models have become increasingly popular tools in population genetics to detect, quantify and understand the factors affecting genetic structure and gene flow [Beaumont and Rannala, 2005, Excoffier and Henkel, 2006]. They are also used as a key step in association studies [Marchini et al., 2004, Balding, 2006]. The success of these methods relies on the possibility to integrate various kinds of biological information in the models, and on the

with migration, first obtained by Wright [1931], whose sampling distribution was explicitly obtained from Wright's Dirichlet by Rannala [1996] and Rannala and Hartigan [1996]. Independently it was obtained using arguments based on coalescent theory with migration under an (implicit) continent-island model by Balding and Nichols [1995] (although recognition that the recursion they derived was indeed multinomial-Dirichlet was not made explicit until Balding and Nichols [1997]). Correlation across populations can be accounted for by population specific parameters. This can be viewed as a heuristic, empirical, statistical approximation of the split model.

The correlated allele frequencies model used in this paper has been implemented earlier by Marchini and Cardon [2002], Falush et al. [2003] Guillot et al. [2005], Marchini et al. [2004], Gaggiotti and Foll [2006] and Foll et al. [2008]. The interest of this model is to base inference on a prior more informative than the default model of uncorrelated frequencies and hence get more accurate results, in particular in presence of low differentiation. Resorting to a biologically more explicit model also has the advantage to allow making direct quantitative inference of biological parameters such as drift parameters which might give clues about on-going evolutionary processes not quantifiable with the uncorrelated model.

While this correlated model is biologically appealing, several issues have prevented accurate assessment of its interest as compared to the uncorrelated model, in particular, in the difficult, but common problem, where K is unknown. The goal of the present paper is to describe and address these statistical and computational issues. These issues are described in detail in the next section. A solution in terms of improved MCMC algorithm and post-processing scheme is then described. Its efficiency is illustrated by the analysis of simulated and real datasets. Implications of this new algorithm are discussed in the last section.

2 STATISTICAL AND COMPUTATIONAL ISSUES

2.1 Modeling assumptions

2.1.1 Prior model of population membership: I denote by p a vector parameterizing the population memberships. If population membership is modeled at the individual level, this vector can be simply $p = (c_1, \dots, c_n)$ where $c_i \in \{1, \dots, K\}$ is the individual population membership. In this case, the simplest form of prior that can be placed is an i.i.d prior $\pi(p|K) = 1/K^n$. The vector p can also correspond to the parameters of a spatial model with the aims to include prior information about how populations are spread across space, see e.g. [Guillot et al., 2005, François et al., 2006, Corander et al., 2008]. This includes models prescribed at the individual level such as Markov random fields. For example, in the latter class of models, $p = (\psi, (c_1, \dots, c_n))$ where ψ is the so-called interaction parameter which is usually unknown and hence also has to be inferred [Green and Richardson, 2002, Møller, 2003, Møller et al., 2006, Marin and Robert, 2007]. Spatial models also include partition (or tessellation) model defined not only at the sampled sites but on the continuous spatial domain. A particular, but widely used case, is the colored Poisson-Voronoi tessellation. In this model, p would denote the vector (m, u, c) where m is the number of Poisson events, $u = (u_1, \dots, u_m)$ is the location of these events and (c_1, \dots, c_m) is the colors (coded as integers in $\{1, \dots, K\}$) of the Voronoi cells induced by the Poisson process [Lantuéjoul, 2002, Guillot et al., 2005]. Other population membership models

can be encompassed with this notation, including parametric models (such as product partition models [Hartigan, 1990, Crowley, 1997]) or non-parametric models (such as Dirichlet process prior [Antoniak, 1974, Pella and Masuda, 2006, Gao et al., 2007, Huelsenbeck and Andolfatto, 2007], or more generally species sampling models [Pitman, 1996, Ishwaran and James, 2003]).

Simulations and inferences in the sequel will be made either with the non-spatial model where $\pi(p|K) = 1/K^n$ or with the spatial model based on Poisson-Voronoi tessellation of Guillot et al. [2005]. For both, a flat prior on K ($\pi(K) = 1/K_{max}$ with $K \in \{1, \dots, K_{max}\}$) will be used.

2.1.2 Prior model of allele frequencies: The frequency of allele j at locus l in population k is denoted by f_{klj} . In the uncorrelated allele frequencies model, the vectors $f_{kl.} = (f_{kl1}, \dots, f_{klJ_l})$ are assumed to have distributions independent across loci and populations. This distribution is always assumed to be a Dirichlet distribution. In the correlated frequencies model, frequencies of an ancestral population denoted by f_{Alj} and population specific drift parameters (d_1, \dots, d_K) are introduced. The frequencies of the ancestral population are also assumed to be independently Dirichlet distributed. Present time allele frequencies $f_{kl.}|f_A, d$ are assumed to have a Dirichlet distribution $D(f_{Al1}(1-d_k)/d_k, \dots, f_{AlJ_l}(1-d_k)/d_k)$. The vector of drift parameters (d_1, \dots, d_K) can be interpreted as F_{ST} s. In this model and conditionally on f_A and d , the frequencies are independent across populations, but marginally (integrating out f_A and d) elementary computations show that the correlation of allele frequencies across population is:

$$Cor(f_{klj}, f_{k'l'j}) = 1/(1 + E[d_k] \frac{E[f_{Alj}] - E[f_{Alj}^2]}{E[f_{Alj}^2] - E[f_{Alj}]^2}), \quad (1)$$

see supplementary material for detail. In the most general case, the distribution of the f_{klj} s in the uncorrelated model may depend on population-, locus- and allele-specific parameters α_{klj} . In practice, the α_{klj} are always assumed to be common across populations, loci and alleles, and most often set to one. Similarly, in the correlated model, the f_{Alj} might have locus- and allele-specific parameters but I do not consider this case here. I set it to one as it is most often done in practice, although the effect of this assumption has not been yet thoroughly assessed in the context of clustering (but see [Foll et al., 2008] in another context). In the sequel, I refer to the model where $f_{kl.} \sim D(1, \dots, 1)$ as to the uncorrelated (frequencies) model and to the model where $D(f_{Al1}(1-d_k)/d_k, \dots, f_{AlJ_l}(1-d_k)/d_k)$ with $f_{Alj} \sim D(1, \dots, 1)$ as to the correlated (frequencies) model. Independence is always assumed across loci.

In order to specify fully the model, we need to place a prior on the drift parameters d_k . As this parameter has to lie in $[0, 1]$, it is natural to consider a Beta prior, with independence across populations. A Beta distribution depends on two parameters, their choice will be discussed in detail in the sequel.

The vector of parameters to be inferred is $\theta = (K, p, f)$ in the uncorrelated model and $\theta = (K, p, d, f_A, f)$ in the correlated model. At this point, it appears more clearly that the correlated model can be viewed as a Bayesian and biologically grounded way to make inference under the uncorrelated model with population-, locus- and allele-specific parameters.

2.1.3 Likelihood: The genotype of a (diploid) individual is denoted by $(z_{il}^{(1)}, z_{il}^{(2)})$ $i = 1, \dots, N$ at locus $l = 1, \dots, L$. Assuming Hardy-Weinberg equilibrium and linkage equilibrium within populations, the probability of observing the genotypes given parameters (likelihood) can be written as

$$\pi(z|K, p, f) = \prod_{i=1}^N f_{k_i l z_{il}^{(1)}} f_{k_i l z_{il}^{(2)}} (2 - \delta_{il}) \quad (2)$$

where k_i is the population label of individual i ($k_i \in \{1, \dots, K\}$) and $\delta_{il} = 1$ if $z_{il}^{(1)} = z_{il}^{(2)}$ and 0 otherwise. Note that the form of the model, and in particular its dimensionality at the likelihood level, does not depend on the allele frequencies model.

2.2 Difficulties in the ascertainment of low population differentiation

In practice, it is common to observe a genetic differentiation between two groups of individuals (as measured by the F_{ST} statistic, that appears to be significant at some threshold by a statistical test, while no structure can be inferred by clustering softwares [Waples and Gaggiotti, 2006]. This problem is for instance briefly mentioned in the manual of the Structure software [Pritchard et al., 2007]. It is not at all specific to a particular software, but widely observed among practitioners with various methods. One of the aims of using the correlated frequencies model is to ascertain subtle structures (i.e. low differentiation) potentially undetected when using the uncorrelated frequencies model. This feature of the correlated frequencies model was supported by Rosenberg et al. [2002] on real data. Falush et al. [2003] also reported results from a simulated dataset where the correlated model allowed to detect low differentiation remained undetected by the uncorrelated model. At last, on the basis of simulated data, Guillot et al. [2005] reported results also supporting a higher power of the correlated model. However, their conclusion stood only in the case where K was known, and a detailed and general quantitative comparison is still lacking.

2.3 Inference of the number of populations with a prior model with correlated allele frequencies

The number of populations K is most often unknown in practice, and several strategies have been proposed to estimate it along with other unknown parameters. In their early work, Pritchard et al. [2000] considered K as known, and the estimation of the number of populations K when it is unknown was treated by an approximation of its marginal likelihood from several runs with fixed K , which was described as an expedient by these authors.

Dawson and Belkhir [2001] and Corander et al. [2003] proposed transdimensional Monte-Carlo Markov Chain algorithms to sample on a parameter space that includes the number of populations itself. However, these works dealt only with the uncorrelated frequencies model. Guillot et al. [2005] proposed an algorithm encompassing both the uncorrelated and the correlated model. They reported good results for the inference of K when using the uncorrelated allele frequencies model and poor results using the correlated frequencies model. Guillot et al. [2005] did not exclude the possibility of algorithm weaknesses. But as the correlated model corresponds to the idea that populations tend to be genetically similar, they also

suspected that these poor results might have come from identifiability problems in the model itself, and recommended not to use it for unknown K .

2.4 Ghost populations

The purpose of MCMC simulation in clustering models is to derive an estimation of K and to estimate population membership of individuals to each of the K inferred populations. Toward the first goal, Guillot et al. [2005] proposed to run an MCMC algorithm where K is part of the unknown parameters to be estimated. This first run was used to get an estimate of K . The authors found that this first run was useless to make assignment because of the swap of population labels across the MCMC run. To explain briefly this problem, let us consider a chain reaching a state (after a burn-in period), where some individuals forming a genuine HWLE population are properly grouped together. Because these individuals form a genuine HWLE population, they are likely to remain clustered together along most of the remaining time of the chain. However, because of the split and merge of populations and of the update of population labels across the chain, this genuine HWE group, labeled as population k at a given time of the chain, is likely to be relabeled as population k' ($k \neq k'$), later in MCMC iterations. Hence, trying to estimate population membership from this chain on the basis of the modal population label for each individual, usually gives very poor results [Dawson and Belkhir, 2001]. This problem referred to as "label switching problem" has been reported and discussed in detail in the MCMC literature, see e.g. [Richardson and Green, 1997, Stephens, 1997, Celeux et al., 2000] and [Stephens, 2000]. In the context of population genetics where many variables (markers) are used simultaneously, Pritchard et al. [2000] noted that the label switching occurs marginally on chain with fixed K . This fact is also reported in non-genetic mixture model literature, see e.g. [Jasra and Stephens, 2005] and references therein. It is explained by weak mixing properties of MCMC algorithms in high dimensions that prevent populations to be relabeled. For fixed K with a small number of markers (say less than ten) as it is common in small scale ecological studies, Guillot et al. [2005] observed that label-switching could be frequent and described how it could be sorted by reordering the population according to the frequency of a carefully chosen allele. This proved to work in case allele frequencies differ enough across populations, which is an assumption often violated. Gao et al. [2007] described a similar rule and also reported difficulties in case populations do not differ enough. Whatever the number of markers used, within a model where K is treated as unknown, hence with better mixing properties (as noted e.g. by Jasra and Stephens [2005]), this issue had to be addressed in a better way.

Guillot et al. [2005] proposed to run a second chain with K fixed to the value estimated in the first run and to use this second chain to derive an estimation of all parameters (and mostly to make assignments). They reported good results from simulated data in terms of assignments but observed on a dataset of *Gulo gulo* that the number of non empty populations estimated from this second run could be lower than the number \hat{K} derived from the first run. They referred to the empty inferred clusters as to ghost populations. This problem has been reported subsequently by Coulon et al. [2006], Pilot et al. [2006], Fontaine et al. [2007], Lada et al. [2007], Rowe and Beebe [2007] and Coulon et al. [2008] among others. See also [Excoffier and Henkel, 2006] for a review. It has been to date a challenging

issue. Note that the problem of ghost populations is not inherent to the use of any of the two allele frequencies model, it has been actually disregarded in the case of the correlated model because of the poor results mentioned in the previous sections in case of unknown K .

3 SOLUTIONS

3.1 Improved MCMC moves

An MCMC algorithm can be designed to make inference in the models described in section 2.1 [Guillot et al., 2005]. This algorithm combines different fixed and variable dimension Metropolis-Hastings moves. I improved this basic algorithm according to two aspects:

3.1.1 Joint update of population memberships and allele frequencies In [Pritchard et al., 2000, Falush et al., 2003] and [Guillot et al., 2005], updates of population memberships and frequencies were performed sequentially by Gibbs or Metropolis-Hastings updates of population memberships, then by Gibbs updates of allele frequencies. Proceeding this way, as soon as a population becomes empty along the chain, its allele frequencies are updated on the basis of the information provided by the prior only, (since the Gibbs step samples from the posterior conditioned by no data). I modified the algorithm proposed in [Guillot et al., 2005] in such a way that these updates are performed jointly: a modification of the current state of the vector of population memberships is proposed and new frequencies are proposed from the full conditional distribution. This update is proposed and accepted and globally according to an acceptance ratio given in the supplementary material. Split-merge moves allow to increase or decrease the number of populations. A move of say, split type, involves proposing a new vector of parameters with a new value for K , for the vector of population memberships and for the drift parameters and allele frequencies of the two populations resulting of a split.

3.1.2 Split-merge of populations In [Guillot et al., 2005], the proposal distribution for the drift parameters was taken equal to the prior. This was a poor choice in terms of mixing and I improved it by a standard reversible jump technique. See supplementary material for details.

3.2 Improved MCMC psot-processing scheme

I carried out a detailed analysis of the behavior of the algorithm run with variable K and then of the algorithm run with K fixed at the value estimated from the first run. It reveals that the second chain sometimes “numerically absorbs” certain populations in the sense that a population becoming empty at a given time of the chain has an extremely small probability to become non-empty later in the chain, even though this population corresponds to a genuine HWLE group and was properly identified as such by the first chain. This is clearly a weakness of the fixed K algorithm in terms of mixing and would be solved if estimation of K and assignment could be carried out from a single run with variable K . This idea has been impeded so far by the lack of efficient methods allowing to get rid of the label switching problem.

As a solution to the label switching problem for MCMC in mixture models in fixed dimension (hence for K fixed here), Marin et al. [2005] recently proposed to relabel populations after the MCMC run in such a way that each state of the chain best “looks like” the modal state of the chain (see reference above for details). This strategy can not be used directly in the case of a chain simulated over a parameter space of variable dimension but it can be performed on the subset $(\hat{\theta}^{(t)})_t$ of states where $K = \hat{K}$, see supplementary material for details. Note that the chain restricted to states where $K = \hat{K}$ might still be of varying dimension in particular in the case of a prior model for population membership involving a varying number of components. But the dimension variation at this level does not raise any problem as the relabeling scheme is based only on the set of frequencies.

In the case where \hat{K} is large (say larger than 10), the scheme above becomes numerically heavy and quickly intractable as it involves evaluating $\hat{K}!$ permutations. In this case, as suggested by Marin et al. [2005], I perform assignments from the modal state θ_{priv} among the set of states where $K = \hat{K}$.

I found, as illustrated below, that this strategy was very efficient in the sense that it estimates K accurately, it requires one MCMC run only instead of two and it substantially decreases the risk of having an empty estimated population. I used this scheme to relabel MCMC states within runs but it can also be used straightforwardly to relabel states across runs, a task addressed by the method proposed by Jakobsson and Rosenberg [2007]. Whereas their method is based on a matrix of co-assignments and can only deal with across-run label switching, the present method deals with both problems and is thus more general.

4 ILLUSTRATION OF ALGORITHM IMPROVEMENT

4.1 Analysis of data simulated from the prior-likelihood model

I simulated genotypes for $n = 100$ individuals, at $L = 10, 20, 50, 100$ loci, with $J = 10$ alleles at each locus. Each dataset consisted of $K = 2$ populations. In order to investigate a broad range of situations in terms of levels of differentiation, I simulated the coefficients d_k from a mixture model: $d_k \sim bG + (1 - b)G'$ where b is a Bernoulli variable with probability 0.5, while G and G' have Beta distribution with parameters $(2, 20)$ and $(1, 100)$ respectively. In other words, I simulated datasets with medium-high level of differentiation or with very low levels of differentiation with equal probability 0.5. In the inference, K was treated as unknown and I placed a uniform distribution between 0 and $K_{\text{max}} = 5$. In order to assess the influence of the prior on the drift coefficients d_k , for each dataset I carried out inference with three different priors (involving three different MCMC runs) for each dataset. I considered a Beta(1,100) prior referred to as low differentiation prior, a Beta(2,20) prior referred to as medium differentiation prior, and a (flat) Beta(1,1) prior referred to as flat prior. I also carried out inferences with a fourth run using the uncorrelated frequencies model. I made 100000 MCMC iterations, discarding the first 50000 iterations in the MCMC output analysis. This procedure was repeated $n = 500$ times (each simulated datasets being analyzed in four different ways).

As a first step I simulated data from a non spatial prior. Denoting by $c = (c_1, \dots, c_n)$ the population label of individuals, I placed a prior $\pi(c|K) \propto 1$ giving an equal weight to all clusterings (in particular whatever its spatial configuration). This uninformative prior was used for simulation and inference. In a second step, the same procedure was repeated but with data simulated from the colored Poisson-Voronoi spatial prior described in [Guillot et al., 2005] and I also used this prior to carry out inferences. Each value for L was investigated through $N = 500$ datasets.

The general features observed from this numerical experiment are as follows: (i) the correlated frequency model gives higher accuracy in terms of percentage of individuals correctly assigned, (ii) this higher accuracy is obtained whatever prior placed on the drift coefficients for the inference, (iii) using a prior on d that places more weight on low values (hence assuming lower genetic differentiation) tends to increase accuracies of inferences, (iv) the general higher accuracy in terms of assignments under the correlated model with a prior on d shifted to the left is obtained to the price of a slight tendency to overestimate K , and (v), with a word of caution (because these results are based on computations where the prior assumed in inference is the same as the one used in simulation, which correspond to a best case situation), these simulations show that the use of a spatial prior on a dataset displaying a genuine spatial structure allows an increase in the accuracy of inferences.

The various plots given as supplementary material also suggest the existence of a hard threshold in terms of F_{ST} , beyond which the uncorrelated model becomes blind to any structure because it estimates $\hat{K} = 1$. In contrast, the correlated model seems to be more permissive on K , and it is also able to detect structure at very low differentiation (although with an accuracy decreasing with the number of loci). I also found that MCMC mixing was improved, allowing to avoid staying in states with one or several empty populations.

4.2 Analysis of data simulated from a Wright-Fisher neutral model

I made simulations under a more biologically grounded model, the Wright-Fisher neutral model implemented in the software `ms` [Hudson, 2002]. It assumes the standard coalescent approximation to the Wright-Fisher model with symmetric migration among subpopulations and an infinite-sites model of mutation. I simulated datasets consisting of $n = 100$ individuals belonging to two populations genotypes at twenty independent loci. I considered various scenarios in terms of genetic diversity and amount of gene flows. These aspects are controlled by parameters θ and M defined as $4N_e u$ and $4N_e m$ respectively (where N_e is the effective diploid population size, u is the neutral mutation rate and m is the migration rate per generation). I let the mutation parameter θ vary in $\{0.5, 1, 2, 4\}$ and the migration parameter M vary in $\{0.5, 1, 2, 4, 10, 20, 40\}$. I considered number of loci L equal to 10, 20, 50 and 100. Each combination of θ , M and L was investigated through $N = 100$ datasets. The purpose of this experiment was mostly to investigate the role of the frequency model, hence I did not attempt to produce spatialized data in the simulations. Therefore, inferences were carried out by a non-spatial model. As in the previous section, the simulated datasets were analyzed using the correlated frequencies model with three different priors for the hyper-parameters and also with the uncorrelated frequencies model. The accuracy was measured by the bias

on K and a rate of miss-assignment. Results are given as tables in supplementary material.

The general features are as follows: (i) the accuracy in assignments increases with the number of loci and generally decreases with the migration rate whatever the model used, (ii) the uncorrelated frequency is affected by a downward bias in K whereas the correlated model can be affected by an upward or downward bias depending on θ and M , (iii) the prior on drift affects the accuracy in terms of inference of K (iv) the relative performances for different priors vary with L , θ and M , (v) a prior on drift assuming a medium differentiation gives the best results in the overall, and (vi), for difficult problems both models behave poorly in terms of inference of K , but the correlated model still detects some structure when the uncorrelated become totally blind to any signal.

4.3 Analysis of real data

I re-considered the dataset described by Cegelski et al. [2003], and subsequently analyzed by Guillot et al. [2005] and Corander et al. [2008]. This dataset consists of eighty-eight wolverines (*Gulo Gulo*). They were sampled in western United-States over an area of about 250000 km² and genotyped at ten micro-satellite loci. Cegelski et al. [2003] and Corander et al. [2008] reported the finding of three populations. Using a spatial model and the uncorrelated frequency model, Guillot et al. [2005] reported a mode at 6 for the posterior distribution of K for the best run (in terms of mean posterior density) among 100 runs. Several runs with K fixed at 6 were persistently unable to obtain six non-empty populations, while the best of 100 such fixed K runs corresponded to four non-empty populations. I re-analyzed this dataset with the colored Poisson-Voronoi prior for the spatial component of the model and the correlated frequency model with shape parameters (2, 20) as hyper-prior parameters for the drift coefficients d . I made fifty independent runs of length 500000 discarding the first 100000 iterations (burn-in) in the post-processing. Each runs provided (in a single step) an estimate of K and a map of the estimated populations through the relabeling scheme described above. Thirty five runs gave a mode at six while fifteen runs gave a mode at seven for the posterior distribution of K . Among these fifty runs, three corresponded to a configuration containing one empty population (two runs where $\hat{K} = 7$, one run where $\hat{K} = 6$). All other runs were free from any such ghost population. The best run among the fifty runs corresponded to $\hat{K} = 6$. See supplementary material for the inferred spatial population structure.

Similar analysis of this dataset with MCMC runs of longer length tended to decrease the proportion of runs with output containing ghost populations. Estimated F statistics for the six inferred populations are given in a table in the supplementary material.

The analysis of Guillot et al. [2005] did not give element supporting the existence of six genuine sub-populations versus the presence of model, MCMC or post-processing artifacts. In contrast, the current re-analysis tends to give weight to the assertion that the whole population can be viewed as a set of six groups with a hierarchy of populations with decreased genetic differentiation. The populations described by Cegelski et al. [2003], Corander et al. [2008] and Guillot et al. [2005] match the present structure geographically, hence these new results are consistent with those reported previously but with a finer resolution.

5 DISCUSSION

Although introduced several years ago, implemented for various purposes in many softwares and widely used in practice, a detailed comparison of the correlated and uncorrelated frequency models was lacking. After the introduction of several algorithm modifications improving over the work of [Guillot et al., 2005], I obtain conclusions significantly different from this previous work: the extreme tendency of the correlated model to overestimate K reported in [Guillot et al., 2005] was not inherent to the model but to the algorithm. After algorithm improvement, the present study suggests that using the correlated model with this new MCMC implementation makes sense, even with unknown K . It potentially avoids missing the detection of structure at low level of differentiation.

The analysis of simulated data shows a sensitivity of inferences to the choice of the hyper-prior parameters of the drift coefficients d_k . The choice of the hyper-parameters affect both the accuracy on K and assignments. The optimal hyper-parameters depend on the level of genetic diversity and on the level of differentiation between the sought populations. The diversity is known but it depends on the the locus considered. The differentiation is unknown thus it is not possible to recommend an optimal value in practice. Fortunately, the present simulations suggest that the Beta(2, 20) distribution gives good results on the whole range of situations considered, and overall it performs better than the uncorrelated model. This distributions should be used as a default hyper-prior. This is the approach taken for example by Hannelius et al. [2008] in their study of the Finnish and Swedish populations where thirty-four SNPs were analysed with the the present algorithm. It proved to be powerful to detect a known genetic structure at an extremely low level of differentiation.

Regarding the occurrence of MCMC runs displaying average or modal states where some of the populations are empty (ghost populations), the present work improve substantially on the work of [Guillot et al., 2005] (see supplementary material for detail). However, my experience with the new algorithm shows that ghost populations might still very occasionally occur. In the case of the present algorithm, MCMC of too limited length seem to be responsible for these rare events. In case the issue is not solved by increasing the length of the chain (or this value being already set to the maximum practically feasible value), I still recommend to disregard empty populations and to consider the number of non-empty populations as the correct (and practically meaningful) estimate of K . Note however, that the issue is not specific to the present algorithm. It is actually most likely to occur with any algorithm providing a separate estimate for the number of clusters and for the assignments themselves (in contrast with estimation of K based on the number of non-empty populations). Although this aspect tends to be minimized by many authors, it has been mentioned regarding the softwares released by Corander et al. [2003], François et al. [2006] and Gao et al. [2007].

It would be easy to include the straightforward extra step consisting in automatically removing estimated empty populations at the end of a run (as implemented in certain softwares). Although this could be re-insuring for un-experienced MCMC users, I believe that this extra step would mostly obscure a genuine difficulty in MCMC mixture computations. In particular, it is worth pointing out that if a clustering model prior places a positive probability on the number of empty populations (as most common model do), the posterior probability of this number is also necessarily positive.

Having a detailed look at non-tampered MCMC outputs, including at some intellectually uncomfortable aspects is hence a key toward a good understanding of algorithm and model behavior and of the features of a given dataset. In particular, the occurrence of ghost populations is extremely seldom when data are simulated from the prior-likelihood model used for inference; the presence of ghost population(s) can hence be interpreted as a reliable clue that data depart from modeling assumption.

The improvements proposed not only solve inference issues but also increase the speed of the algorithm as it estimates K as accurately as before at the cost of one MCMC run only instead of two. Further could be gained in terms of computation speed with the uncorrelated model by integrating out over allele frequency. I have not been able to get an analytical expression integrating out over allele frequency for the correlated model. This is unfortunate in terms of algorithm speed. But working on the “complete model” (including allele frequencies) is actually compulsory as long as one is interested in other model developments such as the filtering of null alleles as proposed by Guillot et al. [2008] that requires explicit numerical computations on allele frequencies.

The improvements proposed here will be implemented in the software Geneland. This software has mostly been used to analyze geo-referenced genetic data. However, as stressed above these new improvements do not depend on the prior specified on the population membership variable. All the developments above hold for spatial, as well as for non-spatial, prior models (particular cases for these two classes of models being implemented in Geneland). The proposed algorithm should be useful in the many different situations where low differentiation is encountered. This situation is for example common in association studies, where as pointed out by Marchini et al. [2004], even low differentiation require appropriate treatment to avoid detection of spurious associations.

Acknowledgements: I am grateful to my colleagues J.M Cornuet, A. Estoup, A. le Rouzic, J.M. Marin and C.P. Robert for comments and suggestions at various stages of this work, L. Waits for kindly sharing the wolverine dataset, F. Santos for implementing the new features of Geneland into the graphical user interface and Misoo Ellison and Therese Fosholt-Moe for carefully reading this manuscript. This work was financially supported by Agence Nationale de la Recherche through grant No NT05-4-42230.

REFERENCES

- C.E. Antoniak. Mixtures of Dirichlet process with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.
- D.J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7:781–791, 2006.
- D.J. Balding and R.A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96:3–12, 1995.
- D.J. Balding and R.A. Nichols. Significant genetic correlation among Caucasians at forensic DNA loci. *Heredity*, 78:583–589, 1997.
- M. A. Beaumont and B. Rannala. The Bayesian revolution in genetics. *Nature Review Genetics*, 5:251–261, 2005.
- C.C. Cegelski, L.P. Waits, and J. Anderson. Assessing population structure and gene flow in montana wolverines (*gulo, gulo*) using assignment-based approaches. *Molecular Ecology*, 12:2907–2918, 2003.
- G. Celeux, M. Hurn, and C.P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95 (451):957–970, 2000.

- J. Corander, J. Sirén, and E. Arjas. Bayesian spatial modeling of genetic population structure. *Computational Statistics*, 23(1):111–129, 2008.
- J.C. Corander, P. Waldmann, and M.J. Sillanpää. Bayesian analysis of genetic differentiation between populations. *Genetics*, 163:367–374, 2003.
- A. Coulon, G. Guillot, J.F. Cosson, J.M.A. Angibault, S. Aulagnier, B. Cargnelutti, M. Galan, and A.J.M. Hewison. Genetics structure is influenced by landscape features. Empirical evidence from a roe deer population. *Molecular Ecology*, 15: 1669–1679, 2006.
- A. Coulon, J.W. Fitzpatrick, R. Bowman, B.M. Stith, C.A. Makarewich, L.M. Stenzler, and I.J. Lovette. Congruent population structure inferred from dispersal behavior and intensive genetic surveys of the threatened Florida Scrub-Jay *aphelocoma carulescens*. *Molecular Ecology*, 2008.
- E.M. Crowley. Product partition models for normal means. *Journal of the American Statistical Association*, 92:192–198, 1997.
- K.J. Dawson and K. Belkhir. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, 78:59–77, 2001.
- L. Excoffier and G. Henkel. Computer programs for population genetics data analysis: a survival guide. *Nature Review Genetics*, 7:745–758, 2006.
- D. Falush, M. Stephens, and J.K. Pritchard. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003.
- M. Foll, M.A. Beaumont, and O. Gaggiotti. An approximate Bayesian computation approach to overcome biases that arise when using AFLP markers to study population structure. *Genetics*, 2008.
- M. Fontaine, S.J.E. Baird, S. Piry, N. Ray, K. Tolley, S. Duke, A. Birkun, M. Ferreira, T. Jauniaux, A. Llavona, B. Östürk, A.A. Östürk, V. Ridoux, E. Rogan, M. Sequeira, U. Siebert, G.A. Vikingson, J.M. Bouqueneau, and J.R. Michaux. Rise of oceanographic barriers in continuous populations of a cetacean: the genetic structure of harbour porpoises in old world waters. *BMC Biology*, 5(30), 2007.
- L. Foreman, A. Smith, and I. Evett. Bayesian analysis of DNA profiling data in forensic identification applications. *Journal of the Royal Statistical Society, series A*, 160: 429–469, 1997.
- O. François, S. Ancelet, and G. Guillot. Bayesian clustering using hidden Markov random fields. *Genetics*, 174:805–816, 2006.
- O. Gaggiotti and M. Foll. Identifying the environmental factors that determine the genetic structure of populations. *Genetics*, 174:875–891, 2006.
- H. Gao, S. Williamson, and C.D. Bustamante. A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics*, 176:1635–1651, 2007.
- P.J. Green and S. Richardson. Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1055–1070, 2002.
- G. Guillot, A. Estoup, F. Mortier, and J.F. Cosson. A spatial statistical model for landscape genetics. *Genetics*, 170(3):1261–1280, 2005.
- G. Guillot, F. Santos, and A. Estoup. Analysing georeferenced population genetics data with geneland: a new algorithm to deal with null alleles and a friendly graphical user interface. *Bioinformatics*, 24(11):1406–1407, 2008.
- U. Hannelius, E. Salmela, T. Lappalainen, G. Guillot, C.M. Lindgren, U. von Döbeln, P. Lahermo, and J. Kere. Population substructure in Finland and Sweden revealed by a small number of unlinked autosomal SNPs. *Under revision*, 2008.
- J.A. Hartigan. Partition models. *Communication in Statistics. Theory and methods*, 19(8):2745–2756, 1990.
- R.R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.
- J.P. Huelsenbeck and P. Andolfatto. Inference of population structure under a Dirichlet process model genetics. *Genetics*, 175:1787–1802, 2007.
- H. Ishwaran and L.F. James. Generalized weighted Chinese restaurant process for species sampling mixture models. *Statistical Science*, 13:1211–1235, 2003.
- M. Jakobsson and N.A. Rosenberg. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14):1801–1806, 2007.
- C. C. Jasra, A. Holmes and D. A. Stephens. Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Statistical Sciences*, 20:50–67, 2005.
- M. Kimura and J. Crow. Some genetic problems in natural population. In *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability*, pages 1–22, 1956.
- H. Lada, R.M. Nally, and A.C. Taylor. Distinguishing past from present gene flow along and across a river: the case of the carnivorous marsupial (*antechinus flavipes*) on Southern floodplain. *Conservation Genetics*, 2007. DOI 10.1007/s10592-007-9372-5.
- C. Lantuéjoul. *Geostatistical simulation*. Springer Verlag, 2002.
- J. Marchini and L.R. Cardon. Discussion on statistical modelling and analysis of genetic data. *Journal of the Royal Statistical Society, series B*, 64(4):740–741, 2002.
- J. Marchini, L.R. Cardon, M.S. Phillips, and P. Donnelly. The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5):512–517, 2004.
- J.M. Marin and C.P. Robert. *Bayesian Core. A Practical Approach to Computational Bayesian Statistics*. Springer-Verlag, 2007.
- J.M. Marin, K. Mengersen, and C.P. Robert. *Handbook of Statistics*, volume 25, chapter Bayesian modelling and inference on mixtures of distributions. Elsevier-Sciences, 2005.
- J. Møller, editor. *Spatial Statistics and Computational Methods*, volume 173 of *Lecture Notes in Statistics*. Springer Verlag, 2003.
- J. Møller, A.N. Pettitt, R. Reeves, and K.K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.
- G. Nicholson, A.V. Smith, F. Jónsson, Ó. Gústafsson, K. Stefánsson, and P. Donnelly. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society, series B*, 64(4):695–715, 2002.
- R. Nielsen. Maximum likelihood estimation of population divergence times and population phylogenesis under the infinite sites model. *Evolution*, 53:143–151, 1998.
- C. O’Ryan, M.W. Bruford, M. Beaumont, R.K. Wayne, Cherry M.I., and E.H. Harle. Genetics of fragmented populations of african buffalo (*syncerus caffer*) in south africa. *Animal Conservation*, 1:85–94, 1998.
- J. Pella and M. Masuda. The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fishery and Aquatic Sciences*, 63:576–596, 2006.
- M. Pilot, W. Jedrzejewski, W. Branicki, V.E. Sidorovich, B. Jedrzejewska, K. Stachura, and S.M. Funk. Ecological factors influence population genetic structure of european grey wolves. *Molecular Ecology*, 14:4533–4553, 2006.
- J. Pitman. *Statistics, probability and game theory*, chapter Some developments of the Blackwell-MacQueen urn scheme, pages 245–268. IMS Lecture Notes Monograph series, Hayward California, 1996.
- J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.
- J.K. Pritchard, X. Wen, and D. Falush. *Documentation for Structure software: Version 2.2*. Department of Human Genetic, University of Chicago, Department of Statistics, University of Oxford, 2007.
- B. Rannala. The sampling theory of neutral alleles in an island population of fluctuating size. *Theoretical Population Biology*, 50:91–104, 1996.
- B. Rannala and J.A. Hartigan. Estimating gene flow in island populations. *Genetical Research*, pages 147–158, 1996.
- S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, series B*, 59(4): 731–792, 1997.
- C.P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer-Verlag, second edition, 2004.
- K. Roeder, M. Escobar, J.B. Kadane, and I. Balazs. Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika*, 85:269–287, 1998.
- N. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovskiy, and M.W. Feldman. Genetic structure of human populations. *Science*, 298:2981–2985, 2002.
- G. Rowe and T.J.C. Beebee. Defining population boundaries: use of three Bayesian approaches with the microsatellite data from British natterjack toads (*bufo calamita*). *Molecular Ecology*, 16:795–796, 2007.
- I.J. Saccheri, I.J. Wilson, R.A. Nichols, M.W. Bruford, and P.M. Brakefield. Inbreeding of bottlenecked butterfly populations: Estimation using the likelihood of changes in marker allele frequencies. *Genetics*, 151:1053–1063, 1999.
- P.E. Smouse, R.S. Waples, and J.A. Twoik. A genetic mixture analysis for use with incomplete source population-data. *Canadian Journal of Fishery and Aquatic Sciences*, 47(620-634), 1990.
- M. Stephens. Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, series B*, 62:795–809, 2000.
- M. Stephens. Discussion of the paper by Richardson and Green “On Bayesian analysis of mixtures with an unknown number of components”. *Journal of the Royal Statistical Society, series B*, 59(4):768–769, 1997.
- R.S. Waples and O. Gaggiotti. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, 15:1419–1439, 2006.
- S. Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, 1931.