The last 10 years have witnessed a considerable gain of popularity online. Bioinformatics Supplementary data are available at http://folk.uio.no/gillesg/Bioinformatics-HMRF. Availability: The data analysed in the present article are available from http://folk.uio.no/gillesg/Bioinformatics-HMRF. Supplementary information: Supplementary data are available at Bioinformatics online.

1 BACKGROUND

1.1 Spatial population genetics

The last 10 years have witnessed a considerable gain of popularity of clustering models in population genetics. The origin of this interest lies in the fact that many statistical tests in population genetics [for example, for the association between a genotype and a phenotype (Balding, 2006) or for the detection of selection (Nielsen, 2001)] are based on the assumption that the sample arises from a homogeneous population. The detection of non-homogeneity in the form of clusters is therefore often a preliminary step to these analyses. The other crucial aspect of clustering analysis is that detecting and interpreting clusters can give clues about biological processes affecting genetic diversity. It has been extensively used in areas as diverse as demography, epidemiology, ecology, population management and conservation genetics (Excoffier and Heckel, 2006). To have a quantitative idea of the usefulness of clustering models in population genetics, it is perhaps worth mentioning that the articles describing models, algorithms and computer programs to cluster genetic data have received collectively more than 3000 citations in the last 10 years (source ISI web of knowledge).

Many biological processes shaping genetic diversity are mediated by space, an aspect noted in the early work of Dobzhansky and Wright (1941), Wright (1943) and Malécot (1948), and there is increasing evidence that biological processes have a signature in terms of spatial organization of genetic diversity at fine scale. See the work of Lao et al. (2008) and Novembre et al. (2008) for recent examples and implications for association studies. The use of explicitly spatial models for analysing genetic data can be therefore an efficient way to both include information about the sought patterns in the inference step and ease their interpretation. The development of such models has been lately an active area of research. Spatial clustering methods form a particular class of such models. They can be viewed as an attempt to include geographical information in inference schemes about phylogenies (Cavalli-Sforza and Edwards, 1967). The models proposed depend on the biological context but consist essentially of variations around the model pioneered by Pritchard et al. (2000). In its simplest version, this model assumes individuals having pure ancestry in a fixed number of clusters (or populations) at Hardy–Weinberg equilibrium (HWE) with linkage equilibrium (LE) between loci.

Imbedding this model in a spatial framework amounts to prescribing a probabilistic model of the organization of clusters across space. This can be done either through a continuous tessellation (Guillot et al., 2005) or a discrete model based on a graph of spatial neighbourhood (Corander et al., 2008; François et al., 2006). Examples of applications can
It would be therefore natural to carry out inference of parameters on a Markov random fields are often sensitive to the values of choices include the removal of long edges and the removal of some of the edges crossing marine areas.

Then, from this edited graph, Chen et al. (2007a) and François et al. (2008) recommend to select the run that achieves the largest Deviance Information Criterion (DIC). The method described above is not grounded in a formal statistical inference method. However, from simulated data, Chen et al. (2007a) and François et al. (2008) recommend to launch Markov chain Monte Carlo (MCMC) simulations that produce a sample from the posterior distribution of the population memberships. The number of clusters $K$ and the interaction parameter $\psi$ remain fixed along each MCMC run and various values of $K$ and $\psi$ are tried across several runs within limited set of values. The various MCMC runs produce samples with different clustering solutions and Chen et al. (2007a) recommend to select the run that achieves the largest Deviance Information Criterion (DIC).

1.3 Goal of the present article

The method described above is not grounded in a formal statistical inference method. However, from simulated data, Chen et al. (2007a) have reported surprisingly good results. Despite the success of the method, François et al. (2008) suggest some modification (DIC versus likelihood) and report a good robustness of the method to parameter values.

1.2 Inference of spatial population genetics structure with Markov random field models on graphs

A recent series of papers (Chen et al., 2007a; François et al., 2008; Wang et al., 2007) has described a computer program and recommended a statistical methodology to perform inference in a simplified version of the model described in François et al. (2006). The key assumption underlying this model is that population memberships follow a so-called hidden Markov random field model. In informal terms, this can be summarized as follows: the log-probability of an individual belonging to a particular population given the population membership of its closest neighbours is equal (up to proportionality constant) to the number of neighbours belonging to this population [see, e.g. Guttorp (1995, Ch. 4) for details]. The model involves therefore three important quantities: a graph $G$ specifying the set of neighbours of each individual, the proportionality constant $\psi$ in the above-mentioned relationship and the number of populations $K$.

The inference of the number of components $K$ in a statistical model has generated a large volume of literature (Cappe et al., 2003; Celeux and Soromenho, 1996; Green, 1995; Richardson and Green, 1997; Stephens, 2000). See also Robert and Casella (2004, Ch. 11) or Sisson and Chan (2005) for an overview. Likewise, the inference of the interaction parameter $\psi$ in a Markov random field has received a lot of attention (Gelman and Meng, 1998; Green and Richardson, 2002; Guyon, 1991; Hurn et al., 2003; McGrory et al., 2007; Müller et al., 2006). Regarding the neighbourhood structure $G$ when the data are not collected on a grid, an expedient often used consists in assuming the Delaunay graph generated by the sampling sites. The sensitivity to the choice of a particular graph has not been studied;

In a biological context, there is often little prior knowledge about how to choose $K$. Besides, the interpretation of $G$ and $\psi$ is in general challenging. At last, the output of statistical models based on a Markov random fields are often sensitive to the values of the interaction parameters used if this parameter is not inferred [see Marin and Robert (2007, p. 238) for a graphical example]. It would be therefore natural to carry out inference of $G$, $K$ and $\psi$ prior to (or jointly with) the inference of population memberships. However, despite the existence of formal statistical methods to achieve this task, the study by Chen et al. (2007a), François et al. (2008) and to a lesser extent by Wang et al. (2007) base their conclusions on a much simpler method. Briefly, this method consists in starting from the Delaunay graph and editing this graph by adding or removing edges in order to make it biologically more ‘realistic’. The choice of edges to be added or removed involves some kind of arbitrariness. For example, for data with sampling sites spread across Europe, François et al. (2008) disconnect Sicily and Sardinia from the continent, but connect England to France and to the Netherlands. See Figure 1 in the present article for the location of sampling sites in this dataset. Furthermore, these authors disconnect Genoa from Northern sites across the Alps, but let the Western Norway site connected to Southern and Eastern Sweden. Other questionable
Toward this goal, I first recall the model used by Chen et al. (2007a) and François et al. (2008). Then, I reanalyse the simulated data that Chen et al. (2007a) used to advertise their methodology. Since Chen et al. (2007a) and François et al. (2008) discuss the effect of isolation by distance, I analyse some new data simulated according to a scenario of isolation-by-distance. At last, in view of these simulated data I also reconsider the Arabidopsis thaliana data analysed by François et al. (2008).

2 MODEL AND ALGORITHM

The present article is an attempt to assess the value of this method. Toward this goal, I first recall the model used by Chen et al. (2007a) and François et al. (2008). The data at hand are assumed to consist of the sampling locations \( z_{il} \) also on \( G \), at locus \( l \) and \( i \) is denoted by \( \{a_{il}, f_{il}\} \) for diploid organisms, and by \( a_{ilj} \) for haploid organisms.

\section{2.1 Model underlying the Tess program}

\subsection{2.1.1 Cluster membership model}

The cluster membership of individuals are denoted by \( c=(c_i)_{i=1,...,n} \). To account for the fact that spatially close individuals are likely to belong to the same populations, a graph structure denoted by \( G \) is introduced. A default choice for this graph can be the Delaunay graph where two vertices are neighbours if they belong to Voronoi cells that share a common edge. It is assumed that each individual originates from one of \( K \) clusters, and that the vector of cluster memberships follows a \( K \)-state Potts model with parameter \( \psi \) (but whose definition relies also on \( G \), namely:

\[ \pi(c_i=k|c_{-i}) \propto \exp \left( \psi \sum_{j \in \partial i} l_{j,k} \right) \]  

where \( \partial i \) denotes the set of neighbours of \( i \) in \( G \) and \( l_{j,k} \) is the indicator function of the event \( [c_j=k] \).

\subsection{2.1.2 Genotype model}

The frequencies of alleles are assumed to vary across populations. The frequency of allele \( f \) at locus \( l \) in population \( k \) is denoted by \( f_{kl} \). It is assumed that HWE holds in each population for each locus. This amounts to assuming that alleles in each population are sampled independently from a common vector of allele frequencies \( \{f_{kl}\} \). For diploid organisms, this can be written:

\[ \pi(z_{1l},...z_{nl}|f_{kl},c) = \prod_{i=1}^n \prod_{j=1}^n f_{i,kl} f_{j,kl} (2-2f_{i,kl}) \]  

(2)

where the factor \((2-2f_{i,kl})\) accounts for the fact that \( \{a_{il}, f_{il}\} \) is an unordered set.

For haploid organisms, the assumption of HWE can be written:

\[ \pi(z_{1l},...z_{nl}|f_{kl},c) = \prod_{i=1}^n f_{i,kl} \]  

(3)

\section{2.2 Algorithm underlying the Tess program}

\subsection{2.2.1 MCMC transition kernel}

In the Tess program, the vector of parameters involved is \( \theta=(c,f,\psi,K) \). The iterative algorithm leaves \( \psi,K \) constant. It alternates Gibbs updates of the components \( f_{kl} \) of \( f \) (the full conditional being also Dirichlet by standard conjugacy) and Metropolis–Hastings updates of the components of \( c \).

\subsection{2.2.2 Selecting MCMC runs}

Let us denote by \((\hat{\theta})_{T=1,...,T} \) an MCMC run of \( T \) iterations (launched with fixed interaction parameter and fixed number of populations). From several runs with various values of \( \psi \in \Psi \) and \( K \in \{1,...,K_{\text{max}}\} \), Chen et al. (2007a) estimate \((\hat{\psi}, \hat{K})\) as the value that achieves the largest within-run average likelihood:

\[ \hat{\psi}, \hat{K} = \text{Argmax}_{\psi,K} \frac{1}{T} \sum_{t=1}^T L(\theta^t), \]  

(5)

while François et al. (2008) estimate it as the value that achieves the largest within-run average DIC:

\[ \hat{\psi}, \hat{K} = \text{Argmax}_{\psi,K} \frac{1}{T} \sum_{t=1}^T \text{DIC}(\theta^t), \]  

(6)

the maximum being taken over the various MCMC runs available.

3 REANALYSIS OF SIMULATED DATA

\subsection{3.1 Five-island data}

\subsubsection{3.1.1 Material and method}

I reanalysed the simulated datasets considered in Chen et al. (2007a) which Wang et al. (2007) and François et al. (2008) cite as their unique reference to support the method regarding the choice of the interaction parameter \( \psi \) and the maximum number of populations \( K_{\text{max}} \). The whole set of data analysed there consists of 50 independent datasets, each dataset mimicking the presence of five populations of 100 individuals. Each dataset is characterized by a common level of pairwise differentiation measured by \( F_{ST} \) (Weir and Cockerham, 1984) ranging between 0.01 and 0.1 (five datasets for each level). These data were initially produced by Latch et al. (2006) to assess the performances of non-spatial Bayesian clustering softwares and were spatialized on a ring by Chen et al. (2007a), in such a way that each population occupies an approximately circular spatial domain with slight overlap between contiguous populations. To investigate the performances of Tess in the case where the data do not display any structure, I also subsampled each of the datasets described above, taking only 100 individuals belonging to the same population.

I first used Genepop (Rousset, 2007) to test the departure from HWE for each sub-population in the overall original dataset. In Chen et al. (2007a), the maximum number of population \( K_{\text{max}} \) was set to 6, the interaction parameter \( \psi \) was set to 0.6 and the graph was the Delaunay graph. I also used the Delaunay graph, but in contrast with Chen et al. (2007a), I investigated a broader range of values for the maximum number of populations \( K_{\text{max}} \) and the interaction parameter \( \psi \). I analysed these data setting \( K_{\text{max}} \) to 10 and making runs with eight different values for \( \psi \), namely \( \psi \in \{0.15,0.3,0.6,0.7,0.8,0.9,1,2.2\} \). Following the recommendations found in Chen et al. (2007a, b), I launched 10 runs for each value of \( \psi \). Since the data consist of non-admixed
individuals and to be in conditions similar to those of Chen et al. (2007a), I used the non-admixture model. Each run consisted of 2000 burn-in iterations followed by 10 000 additional iterations used as MCMC output. For each dataset, I selected the best run defined either as the one achieving the highest average likelihood or the one achieving the lowest average DIC.

The criteria considered to assess the performances of the Tess software were surprisingly poor. For the five-island data, the estimated number of populations and the accuracy in assigning individuals to their true population of origin. I considered the estimated number of populations "officially" reported by Tess in the text file (generically named dataSN.txt) and denoted hereafter by $K_{off}$. I also considered the estimated number of populations effectively observed when counting the number of distinct populations appearing in the file reporting estimated cluster memberships (generically named dataTR.txt) and denoted hereafter by $K_{eff}$. I computed the proportion of misclassified individuals after permuting populations labels to account for the label switching issue. All computations reported here were done with Tess 1.2. Some investigations have also be done using Tess 1.01 and Tess 1.1 and it gave strictly similar results.

### 3.1.3 Discussion
The validation of the Tess program and methodology on individuals with pure ancestry had been performed in Chen et al. (2007a) by investigating a limited range of values for $K_{max}$ and $\psi$. When analysing real data, the values of $K_{max}$ and $\psi$ achieving the best results are not known and it is natural to investigate a broad range of values. Doing so and following the recommendations of Chen et al. (2007a, b) to select the "best run" can lead to very poor results. The good results reported in Chen et al. (2007a) in several cases seems to be considered as very optimistic, and those reported in the present study as closer to what could be expected in practice when analysing real data. It is not clear in which context the simple method advocated in Chen et al. (2007a) to choose $\psi$ and $K_{max}$ leads to meaningful results in general and it seems that if one is not able to tune these values so as to obtain a known result, Tess infers spurious structure, even though the data follow the model assumed by the clustering algorithm. Note that a tendency of Tess to infer a number of populations much larger than estimates with competing spatial genetics clustering softwares has been also reported in Gaufrère et al. (2008) on simulated as well as on empirical data.

### 3.1 Analysis of data simulated under a scenario of isolation-by-distance at mutation–migration–drift equilibrium

#### 3.2.1 Material and method
Wang et al. (2007) and François et al. (2008) analysed data collected at the continent scale for which it is natural to assume a pattern of isolation-by-distance. Besides, Chen et al. (2007a) described the Tess program as suitable to analyses data displaying clinal variations of allele frequencies showing on a toy example displaying a linear cline (allele frequency increasing linearly from 0 to 1), that estimated cluster memberships displayed also similar pattern. I therefore analysed data simulated according to a model of isolation-by-distance. I considered 10 datasets simulated to mimic the presence of a continuous population at mutation–migration–drift equilibrium under an isotropic dispersal model of isolation-by-distance. The overall population consisted of 25 000 diploid individuals located at the nodes of a $500 \times 500$ grid with absorbing boundary conditions. Each deme consisted of a single individual. Dispersal was modelled through a truncated Pareto dispersal function. The second-order moment of this distribution was set to $\sigma^2 = 10$ and the Kurtosis coefficient was set to 62 that correspond to fairly low level of isolation-by-distance. The simulations were carried out at 10 loci assuming a stepwise mutation model and a constant mutation rate for all loci of $5 \times 10^5$. All simulations were carried out using the program IBDSim (Leblois et al., 2009).

From each such simulation, I extracted a sub-grid of 20 x 20 individuals with contiguous individuals in this smaller grid separated vertically and horizontally by 10 nodes of the original grid. Then I used Tess to make inference under the admixture model. I set $K_{max}=10$ and investigated outputs of Tess for $\psi$ in the range $[0.15, 0.3, 0.6, 1.2]$. For each dataset and for each value of $\psi$, I made 10 runs of 2000 burn-in iterations followed by 10 000 additional

![Fig. 2. Percentage of misclassified individuals achieved by the Tess program for the five-island data. Each point represents an average error (percentage of misclassified individuals after permutations to account for label switching) over five datasets, each dataset being investigated through 10 MCMC runs. Black open circles: average error for the best of 10 runs in terms of average likelihood. Black open triangles: average error for the best of 10 runs in terms of DIC. This figure corresponds to Figure 1 in Chen et al. (2007a).](image-url)
This study reported evidence of the presence of three populations these authors 'reproduces the skeleton of Europe'. This map does not contain enough information to retrieve the exact graph used. Besides, as pointed out in Section 1 of the present article, the graph used seems to involve many arbitrary choices. Therefore, I used Tess with the plain Delaunay graph. I used the admixture model and set $K_{\text{max}}$ to 5. I investigated the output of Tess for values of the interaction parameter $\psi$ in $[0.15, 0.6]$. I made runs of 50,000 iterations (including 20,000 burn-in iterations). For each value of $\psi$, I made a series of 50 such runs and selected the best run according to the DIC criterion. I used Tess 1.2.

For $\psi = 0.15$, Tess does not infer any spatial structure. For $\psi = 0.6$, Tess infers two clusters separated by a North–South line approximately located at Lon = 20°. Among these two patterns, the one consisting of a single cluster achieves the lowest DIC. Following the recommendation of Chen et al. (2007a) regarding $\psi$ in a more objective way does not lead to the inference of any spatial structure. The present reanalysis of the A.thaliana data suggests that the inference of three clusters reported in François et al. (2008) seem to rely to a large extent on the particular graph and interaction parameter used.

5 CONCLUSIONS

The main conclusions of the present study are as follows: (i) using the Tess program under the no-admixture model to analyse data consisting of several genuine HWLE populations with individuals of pure ancestries leads to inaccurate results: overestimation of $K$, high error rate in assignments, inference of spurious populations; (ii) using the Tess program under the admixture model to analyse data consisting of a continuous isolation-by-distance population leads to the inference of spurious HWLE populations and the number and spatial features of these populations depend on the parameters $G, \psi$ and $K$ used; (iii) for certain parameter values, these spatial features are qualitatively similar to those reported on the A.thaliana data in François et al. (2008) although the demographic processes are totally different (isotropic dispersion versus Westward migration); (iv) analysing the A.thaliana data with parameters values different from those used in François et al. (2008) does not lead to the same results and tend to suggest the absence of strong discontinuity of allele frequencies.

It is stressed that the point of the present study is not whether the migration process reported by François et al. (2008) as occurred or not. Rather, it suggests that the inference of spatial patterns of genetic variations should be done with care. In particular, it points out that a given inferred spatial pattern can easily arise as an artefact of a particular spatial model interplaying with a poor statistical inference method. It also brings further weight to the study of Novembre and Stephens (2008) showing that distinct demographic processes can lead to common spatial features. This stresses the need for great care in the interpretation of spatial patterns.

The purpose of this study is not to discourage the use of models based on hidden Markov random fields. There is a number of contexts where this model seems rather well suited. This includes the situations where the habitat of the species is genuinely a network (e.g. hydro-graphic network). In any case, the present study shows that the parameters involved in the model should be inferred within a formal statistical method. If objective knowledge is available about the graph $G$, the parameters to be inferred are the interaction parameter $\psi$, the number of cluster $K$ and the cluster memberships $c$ [the allele frequencies being easily integrated-out under the independent Dirichlet model (Guillot, 2008)]. Green and Richardson (2002) proposed a method to carry out full Bayesian
inference in a closely related model. Although the accuracy of this method has not been thoroughly assessed so far, it is in principle perfectly tailored to carry out clustering in population genetics. Implementing this method in the model considered here seems to be a reasonable objective for future work.

ACKNOWLEDGEMENTS
I am grateful to Magnus Nordborg and Mattias Jakobsson for providing the A.thaliana data, to Raphael Leblois for providing the simulated data under the program BHD Sim and to Arnaud Estoup for many stimulating discussions. Numerical computations have been made possible by the use of the computer cluster Titan of the Norwegian High Performance Computing group. Funding: This work was carried out within the CEE/EN FR project INTEGRATE and also partially supported by French Agence Nationale de la Recherche (grant No NT05-4-42230).

Conflict of Interest: none declared.

REFERENCES
Dobzhansky,T. and Wright,S. (1941) Genetics of natural populations. v. relations between mutation rate and accumulation of lethals in populations of drosophilas pseudoobscura. Genetics, 26, 23–51.

Inference of spatial population structure