

# On the inference of spatial structure from population genetics data using the Tess program

Gilles Guillot

Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo, P.O. Box 1066, Blindern 0316, Oslo Norway. Also at Gothenburg Stochastic Centre, Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden and Department of Applied Mathematics and Computer Science, INRA/Agro-ParisTech, Paris, France

Received on March 6 2009; revised on April 6 2009; accepted on April 16 2009

Associate Editor: Alex Bateman

## ABSTRACT

**Motivation:** In a series of recent papers, Tess, a computer program based on the concept of hidden Markov random field, has been proposed to infer the number and locations of panmictic population units from the genotypes and spatial locations of these individuals. The method seems to be of broad appeal as it is conceptually much simpler than other competing methods and it has been reported by its authors to be fast and accurate. However, this methodology is not grounded in a formal statistical inference method and seems to rely to a large extent on arbitrary choices regarding the parameters used. The present article is an investigation of the accuracy of this method and an attempt to assess whether recent results reported on the basis of this method are genuine features of the genetic process or artefacts of the method.

**Method:** I analyse simulated data consisting of populations at Hardy-Weinberg and linkage equilibrium and also data simulated under a scenario of isolation-by-distance at mutation-migration-drift equilibrium. *Arabidopsis thaliana* data previously analysed with this method are also reconsidered

**Results:** 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 highly inaccurate results; 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 whose number and features depend on the parameters used; Results previously reported about the *A. thaliana* using Tess seem to a large extent to be artefacts of the statistical methodology used.

The findings go beyond population clustering models and can be an help to design more efficient algorithms based on graphs.

**Availability:** The data analysed in the present paper are available from <http://folk.uio.no/gillesg/Bioinformatics-HMRF>

**Contact:** gilles.guillot@bio.uio.no

## 1 BACKGROUND

### 1.1 Spatial population genetics

The last ten 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 managements 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 three thousand citations in the least ten 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 organisation 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 organisation 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 (François *et al.*, 2006; Corander *et al.*, 2008). Examples of applications can be found

in (Coulon *et al.*, 2006; Fontaine *et al.*, 2007; Sacks *et al.*, 2008; Hannelius *et al.*, 2008; Joseph *et al.*, 2008; Galarza *et al.*, 2009).

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

A recent series of papers (Chen *et al.*, 2007a; Wang *et al.*, 2007; François *et al.*, 2008) 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 summarised as follows: the log-probability of an individual to 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, Chapter 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 (Green, 1995; Celeux and Soromenho, 1996; Richardson and Green, 1997; Stephens, 2000; Cappé *et al.*, 2003). See also (Robert and Casella, 2004, chapter 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 (Guyon, 1991; Gelman and Meng, 1998; Green and Richardson, 2002; Hurn *et al.*, 2003; Møller *et al.*, 2006; McGrory *et al.*, 2007). 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, however a formal statistical inference scheme is possible (Grelaud *et al.*, 2009).

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 e.g. (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 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 2 in the present paper for the location of sampling sites in the 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 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 launch MCMC simulations that attempt to 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 likelihood while François *et al.* (2008) recommend to select the run that achieves the largest Deviance Information Criterion (DIC).

## 1.3 Goal of the present paper

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.

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). Then I re-analyse 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 re-consider the *Arabidopsis thaliana* data analysed by François *et al.* (2008).

## 2 MODEL AND ALGORITHM

The data at hand are assumed to consist of the sampling locations  $s = (s_i)_{i=1,\dots,n}$  and genotypes  $z = (z_{il})_{i=1,\dots,n;l=1,\dots,L}$  at  $L$  loci of  $n$  individuals. The genotype  $z_{il}$  at locus  $l$  is denoted by  $\{\alpha_{il}, \beta_{il}\}$  for diploid organisms, and by  $\alpha_{il}$  for haploid organisms.

### 2.1 Model underlying the Tess program

**2.1.1 Cluster membership model** The cluster membership of individuals are denoted by  $c = (c_i)_{i=1,\dots,n}$ . 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:

$$\pi(c_i = k | c_{\partial i}) \propto \exp \left\{ \psi \sum_{j \in \partial i} I_{c_j = k} \right\} \quad (1)$$

where  $\partial i$  denotes the set of neighbours of  $i$  in  $G$  and  $I_{c_j = k}$  is the indicator function of the event  $\{c_j = k\}$ .

**2.1.2 Genotype model** The frequencies of alleles are assumed to vary across populations. The frequency of allele  $j$  at locus  $l$  in population  $k$  is denoted by  $f_{klj}$  for  $j = 1, \dots, J_l$ . It is assumed

that Hardy-Weinberg equilibrium 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  $f_{klj}$ . For diploid organisms, this can be written:

$$\pi(z_{1l}, \dots, z_{nl} | f_{kl}, c) = \prod_{i=1}^n f_{kl\alpha_i} f_{kl\beta_i} (2 - \delta_{\alpha_i}^{\beta_i}) \quad (2)$$

where the factor  $(2 - \delta_{\alpha_i}^{\beta_i})$  accounts for the fact that  $\{\alpha_{il}, \beta_{il}\}$  is an un-ordered set.

For haploid organisms, the assumption of Hardy-Weinberg equilibrium can be written:

$$\pi(z_{1l}, \dots, z_{nl} | f_{kl}, c) = \prod_{i=1}^n f_{kl\alpha_i} \quad (3)$$

Linkage equilibrium is assumed between loci and the likelihood can be written as

$$\pi(z | f, c) = \prod_{l=1}^L \pi(z_{1l}, \dots, z_{nl} | f_{kl}, c) \quad (4)$$

The model is completed by assuming a Dirichlet distribution of the allele frequencies with independence across loci and populations.

## 2.2 Algorithm underlying the Tess program

**2.2.1 MCMC Transition kernel** In the Tess program, the vector of parameters involved is  $\theta = (c, f, \psi, K)$ . The iterative algorithm let  $\psi, K$  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  $c$ .

**2.2.2 Selecting MCMC runs** Let us denote by  $(\theta^t)_{t=1, \dots, T}$  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  $K \in \{1, \dots, K_{\max}\}$ , Chen *et al.* (2007a) estimate  $(\psi, K)$  as the value that achieves the largest within-run average likelihood:

$$\widehat{\psi, K} = \underset{\psi, K}{\operatorname{Argmax}} \frac{1}{T} \sum_{t=1}^T L(\theta^t) \quad (5)$$

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

$$\widehat{\psi, K} = \underset{\psi, K}{\operatorname{Argmax}} \frac{1}{T} \sum_{t=1}^T DIC(\theta^t) \quad (6)$$

the maximum being taken over the various MCMC runs available.

## 3 RE-ANALYSIS OF SIMULATED DATA

### 3.1 Five-island data

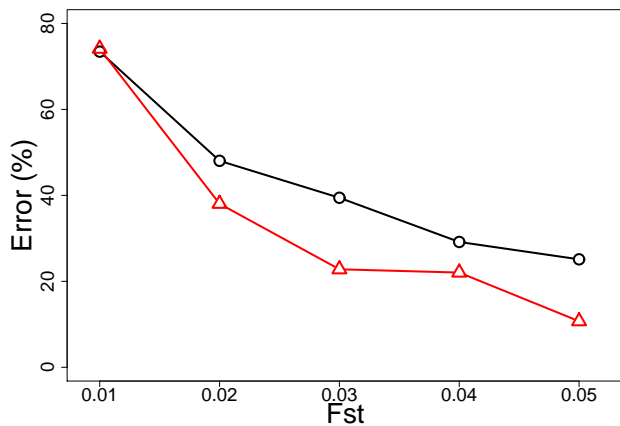
**3.1.1 Material and method** I re-analysed the simulated datasets considered in Chen *et al.* (2007a) which Wang *et al.* (2007) and François *et al.* (2008) cite as their unique reference to support

the method regarding the choice of the interaction parameter  $\psi$  and the maximum number of populations  $K_{\max}$ . The whole set of data analysed there consists of fifty independent datasets, each dataset mimicking the presence of five populations of hundred individuals. Each dataset is characterised 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 spatialised 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 sub-sampled 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 Hardy-Weinberg equilibrium for each sub-population in the overall original dataset. In Chen *et al.* (2007a), the maximum number of population  $K_{\max}$  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  $K_{\max}$  and the interaction parameter  $\psi$ . I analysed these data setting  $K_{\max}$  to 10 and making runs with eight different values for  $\psi$ , namely  $\psi \in \{0.15, 0.3, 0.6, 0.7, 0.8, 0.9, 1.2, 2.4\}$ . Following the recommendations found in Chen *et al.* (2007a) and Chen *et al.* (2007b), 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 10000 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 Deviance Information Criterion.

The criterions considered to assess the performances of the Tess program were the accuracy in estimating the true 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 `dataEN.txt`) and denoted hereafter by  $\hat{K}_{\text{off}}$ . 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  $\hat{K}_{\text{eff}}$ . 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 been done using Tess 1.01 and Tess 1.1 and it gave strictly similar results.

**3.1.2 Results** Formal statistical testing on each population did not lead to reject the assumption of within-population HWLE and between-population differentiation. The performances of the Tess software were surprisingly poor. For the five-island data, the number of populations is always overestimated and the proportion of individuals incorrectly assigned is large. See Figure 1 and Table A of Supporting Material for details. Setting the interaction parameter  $\psi$  to the value 0.6 recommended in Chen *et al.* (2007a) and Chen



**Fig. 1.** 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 ten MCMC runs. Black circles: average error for the best of ten runs in terms of average likelihood. Red triangles: average error for the best of ten runs in terms of deviance information criterion. This figure corresponds to Figure 1 in Chen *et al.* (2007a).

*et al.* (2007b) improves notably the results in terms of assignment but the the number of populations is still largely overestimated (Table B of supporting material). However, considering 0.6 as a golden number suitable for all conditions is not appropriate. Indeed, analysing datasets consisting of a single populations with  $\psi = 0.6$  leads to a systematic overestimation of  $K$  (Table C of supporting material) and to poor accuracy in terms of assignments.

**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) regarding the accuracy of Tess should therefore 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 chose  $\psi$  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 obtained 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 Gauffre *et al.* (2008) on simulated as well as on empirical data.

## 3.2 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 analyse 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 ten 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 25000 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 step-wise 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  $20 \times 20$  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  $\{0.15, 0.3, 0.6, 1.2\}$ . For each dataset and for each value of  $\psi$ , I made 10 runs of 2000 burn-in iterations followed by 10000 additional iterations used as MCMC output. As recommended by François *et al.* (2008) I took the run achieving the lowest Deviance Information Criterion to estimate cluster memberships. In order to use the same inference model as François *et al.* (2008), I used the admixture model. I made computations with Tess 1.2.

**3.2.2 Results** The analysis of these data with Tess leads to identify highly distinct patterns depending on the value of the interaction parameter prescribed. For each of the ten datasets considered, the lowest DIC was achieved by the runs with the lowest value allowed for the interaction parameter  $\psi$  and corresponded to runs where the estimated number of clusters was always 10 (the maximum number allowed under MCMC computations). In these runs, the ten clusters inferred did not display any kind of interpretable spatial pattern. See Figure A of Supporting Material for examples from one dataset. For larger values of  $\psi$ , Tess inferred between 1 and 10 clusters with little consistency between the "best runs" obtained from different values of  $\psi$ .

**3.2.3 Discussion** The performances of the Tess program on genuine IBD data had never been assessed previously. Although Chen *et al.* (2007a) claimed that the Tess program is suitable for analysing this kind of data, it remains unclear whether it makes sense to analyse IBD data with a model assuming the presence of populations at HWLE (even if individuals are allowed

to have admixed ancestries). The present analysis of simulated IBD data suggests that despite the absence of spatially organised and genetically admixed HWLE populations, Tess has a tendency to infer the presence of several HWLE clusters. This feature is not surprising as similar results have been reported earlier regarding other clustering programs (Frantz *et al.*, 2009; Guillot and Santos, 2009; Schwartz and McKelvey, 2008). The poor performances of a model that does not correspond to the on-going demographic/genetic is presumably exacerbated by the use of a poor inference method. It is not clear if under certain values of the interaction parameter  $\psi$ , the inferred cluster memberships might display a spatial pattern that reflects the spatial variations of allele frequencies. But again, in front of real data where the "truth" is not known, one would face anyway the question of selecting the value of  $\psi$ .

#### 4 RE-ANALYSIS OF *ARABIDOPSIS THALIANA* DATA

François *et al.* (2008) applied statistical modelling of spatial population genetic structure to a dataset of *Arabidopsis thaliana* in Europe. This study reported evidence of the presence of three populations displaying a clear spatial pattern which was interpreted as the signature of a past colonisation process from East to West. These results contrasted with earlier findings (Nordborg *et al.*, 2005) and might therefore be attributed to the specific model and parameter setting used in François *et al.* (2008).

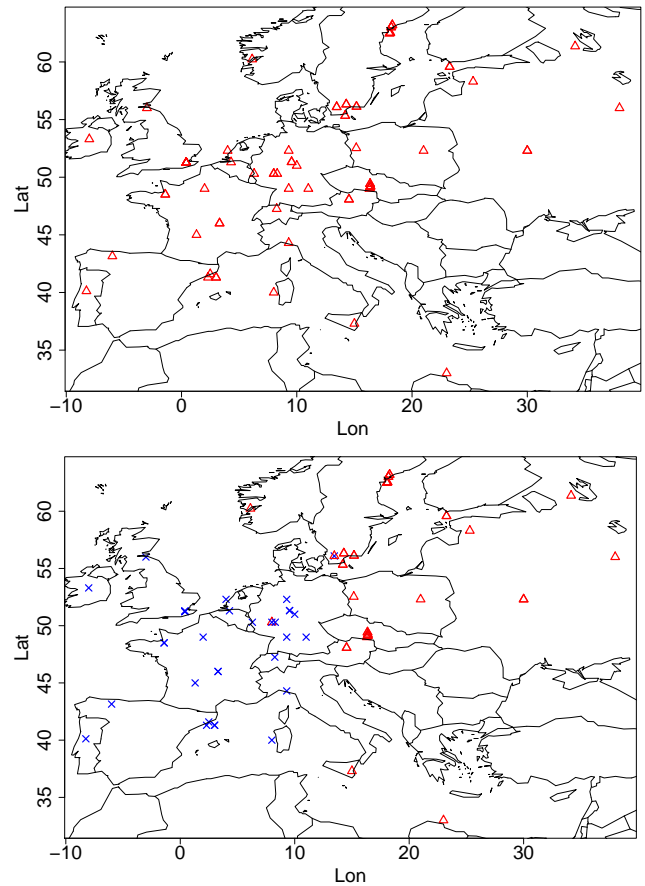
In contrast with the study carried out by Chen *et al.* (2007a), François *et al.* (2008) do not use the plain Delaunay graph generated by the set of sampling sites. They follow a suggestion of Chen *et al.* (2007b) and edit the Delaunay graph by adding and removing edges. The information given in François *et al.* (2008) about the graph used is given as a map (Figure S1 in François *et al.* (2008)) that according to 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 the introduction of the present paper, the graph used seems to involve many arbitrary choices. Therefore, I used Tess with the plain Delaunay graph. I used the admixture model and I set  $K_{\max}$  to 5. I investigated the output of Tess for values of the interaction parameter  $\psi$  in  $\{0.15, 0.6\}$ .

I made runs of 50000 iterations (including 20000 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  $\text{Lon}=20^\circ$ . 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 re-analysis 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 CONCLUSION

The main conclusions of the present study are as follows: (i) Using the Tess program under the no-admixture model to analyse data



**Fig. 2.** Spatial pattern of population membership inferred by Tess on the *Arabidopsis thaliana* data using the plain Delaunay graph. Top  $\psi = 0.15$ ,  $K_{\max} = 5$ , Tess does not infer any structure. Bottom:  $\psi = 0.6$ ,  $K_{\max} = 5$ , Tess infers the presence of two clusters. The run achieving the smallest DIC is the one in the top panel.

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 artifact 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 prior 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.

## FUNDING

This work was carried out within the CEES/NFR project INTEGRATE and also partially supported by French Agence Nationale de la Recherche grant No NT05-4-42230.

## ACKNOWLEDGEMENT

I am grateful to Magnus Nordborg and Mattias Jakobsson for providing the *Arabidopsis thaliana* data, to Raphael Leblois for providing the simulated data under the program IBDSim 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.

## REFERENCES

Balding, D. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**, 781–791.

Cappé, O., Robert, C., and Rydén, T. (2003). Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society, series B*, **65**(3), 679–700.

Cavalli-Sforza, L. and Edwards, A. (1967). Phylogenetic analysis. models and estimation procedures. *American Journal of Human Genetics*, **19**(3), 233–257.

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*.

Chen, C., Durand, E., Forbes, F., and François, O. (2007a). Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, **7**(5), 747–756.

Chen, C., Durand, E., and François, O. (2007b). *Tess reference manual*. <http://membres-timc.imag.fr/Olivier.Francois/tess.html>.

Corander, J., Sirén, J., and Arjas, E. (2008). Bayesian spatial modeling of genetic population structure. *Computational Statistics*, **23**(1), 111–129.

Coulon, A., Guillot, G., Cosson, J., Angibault, J., Aulagnier, S., Cargnelutti, B., Galan, M., and Hewison, A. (2006). Genetics structure is influenced by landscape features. Empirical evidence from a roe deer population. *Molecular Ecology*, **15**, 1669–1679.

Dobzhansky, T. and Wright, S. (1941). Genetics of natural populations. v. relations between mutation rate and accumulation of lethals in populations of drosophila pseudoobscura. *Genetics*, **26**(1), 23–51.

Excoffier, L. and Heckel, G. (2006). Computer programs for population genetics data analysis: a survival guide. *Nature Review Genetics*, **7**, 745–758.

Fontaine, M., Baird, S., Piry, S., Ray, N., Tolley, K., Duke, S., Birkun, A., Ferreira, M., Jauniaux, T., Llavona, A., Östürk, B., Östürk, A., Ridoux, V., Rogan, E., Sequeira, M., Siebert, U., Vikingson, G., Bouquegneau, J., and Michaux, J. (2007). Rise of oceanographic barriers in continuous populations of a cetacean: the genetic structure of harbour porpoises in old world waters. *BMC Biology*, **5**(30).

François, O., Ancelet, S., and Guillot, G. (2006). Bayesian clustering using hidden Markov random fields. *Genetics*, **174**, 805–816.

François, O., Blum, M., Jakobsson, M., and Rosenberg, N. (2008). Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genetics*, **4**(5), e1000075.

Frantz, A., Cellina, S., Krier, A., Schley, L., and Burke, T. (2009). Using spatial bayesian methods to define management units in a continuous population of wild boar (*sus scrofa*): clusters or isolation-by-distance? *Journal of Applied Ecology*, pages doi: 10.1111/j.1365-2664.2008.01606.x.

Galarza, J., Carreras-Carbonell, J., Macpherson, E., Pascual, M., Roques, S., Turner, G., and Ricod, C. (2009). The influence of oceanographic fronts and early-life-history traits on connectivity among littoral fish species. *Proceedings of the National Academy of Sciences*, **106**(5), 1473–1478.

Gauffre, B., Estoup, A., Bretagnolle, V., and Cosson, J. (2008). Spatial genetic structure of small rodent in a heterogeneous landscape. *Molecular Ecology*, **17**, 4616–4629.

Gelman, A. and Meng, X. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, **13**, 163–185.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**(4), 711–732.

Green, P. and Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, **97**(460), 1055–1070.

Grelaud, A., Robert, C. P., Marin, J., Rodolphe, F., and Tally, J. F. (2009). ABC methods for model choice in Gibbs random fields. *Arxiv:0807.2767v2*.

Guillot, G. (2008). Inference of structure in subdivided populations at low levels of genetic differentiation. The correlated allele frequencies model revisited. *Bionformatics*, **24**, 2222–2228.

Guillot, G. and Santos, F. (2009). A computer program to simulate multilocus genotype data with spatially auto-correlated allele frequencies. *Molecular Ecology Resources*, pages doi: 10.1111/j.1755-0998.2008.02496.x.

Guillot, G., Estoup, A., Mortier, F., and Cosson, J. (2005). A spatial statistical model for landscape genetics. *Genetics*, **170**(3), 1261–1280.

Guttorp, P. (1995). *Stochastic modelling of scientific data*. Chapman & Hall.

Guyon, X. (1991). *Random fields on a network*. Springer Verlag.

Hannellius, U., Salmela, E., Lappalainen, T., Guillot, G., Lindgren, C., von Döbeln, U., Lahermo, P., and Kere, J. (2008). Population substructure in Finland and Sweden revealed by a small number of unlinked autosomal SNPs. *BMC Genetics*, **9**(54).

Hurn, M., Husby, O., and Rue, H. (2003). *Spatial Statistics and Computational Methods*, chapter A tutorial in image analysis, pages 87–141. Lecture Notes in Statistics. Springer Verlag.

Joseph, L., Dolman, G., Donnellan, S., Saint, K., Berg, M., and Bennett, A. (2008). Where and when does a ring start and end? testing the ring-species hypothesis in a species complex of australian parrots. *Proceedings of the Royal Society of London, series B*, **275**(1650), 2431–2440.

Lao, O., Lu, T., Nothnagel, M., Junge, O., Freitag-Wol, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L., Comas, D., Holmlund, G., Kouvatzi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., Gether, U., Werge, T., Rivadeneira, F., Hofman, A., Uitterlinden, A., Gieger, C., Wichmann, H., Rütger, A., Schreiber, S., Becker, C., Nürnberg, P., Nelson, M., Krawczak, M., and Kayser, M. (2008). Correlation between genetic and geographic structure in Europe. *Current Biology*, **18**(16), 1241–1248.

Latch, E., Dharmarajan, G., Glaubitz, J., and Rhodes, O. J. (2006). Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, **7**, 295–302.

Leblois, R., Estoup, A., and Rousset, F. (2009). IBDSim: a computer program to simulate genotypic data under isolation by distance. *Molecular Ecology Resources*, **9**, 107–109.

Malécot, G. (1948). *Les mathématiques de l'hérédité*. Masson.

Marin, J. and Robert, C. (2007). *Bayesian Core. A Practical Approach to Computational Bayesian Statistics*. Springer-Verlag.

- McGrory, C. A., Titterton, D. M., and Reeves, A. N. (2007). Variational Bayes for estimating the parameters of a hidden potts model. *Statistics and Computing*, pages doi: 10.1007/s11222-008-9095-6.
- Møller, J., Pettitt, A., Reeves, R., and Berthelsen, K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, **93**(2), 451–458.
- Nielsen, R. (2001). Statistical tests of neutrality at the age of genomics. *Heredity*, **86**, 641–647.
- Nordborg, M., Hu, T., Ishino, Y., Jhaveri, J., Toomajian, C., and et al. (2005). The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology*, **3**(7), e196.
- Novembre, J. and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646 – 649.
- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A., Auton, A. Indap, A., King, K., Bergman, S., Nelson, M., Stephens, M., and Bustamante, C. (2008). Genes mirror geography within Europe. *Nature*, **456**, 98–101.
- Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
- Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, series B*, **59**(4), 731–792.
- Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer-Verlag, second edition.
- Rousset, F. (2007). Genepop'007: a complete re-implementation of the Genepop software for windows and linux. *Molecular Ecology Notes*, **8**(1), 103–106.
- Sacks, B., Bannasch, D. L., Chomel, B. B., and Ernst, H. (2008). Coyotes demonstrate how habitat specialization by individuals of a generalist species can diversify populations in a heterogeneous ecoregion. *Molecular Biology and Evolution*, **25**(7), 1354–1395.
- Schwartz, M. and McKelvey, K. (2008). Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. *Conservation Genetics*, pages doi 10.1007/s10592-008-9622-1.
- Sisson, S. A. and Chan, Y. V. (2005). Trans-dimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association*, **100**, 1077–1089.
- Stephens, M. (2000). Bayesian analysis of mixtures with an unknown number of components - an alternative to reversible jump methods. *The Annals of Statistics*, **28**.
- Wang, S., Lewis, C., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M., Molina, J., Gallo, C., Mazzotti, G., Poletti, G., Hill, K., Hurtado, A., Labuda, D., Klitz, W., Barrantes, R., Cira-Bortolini, M., Salzano, F., Petzl-Erler, M., Tsuneto, L., Llop, E., Rothhammer, F., Excoffier, L., Feldman, M., Rosenberg, N., and Ruiz-Linares, A. (2007). Genetic variation and population structure in native Americans. *PLoS Genetics*, **3**(11), e185.
- Weir, B. and Cockerham, C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, **38**(6), 1358–1370.
- Wright, S. (1943). Isolation by distance. *Genetics*, **28**, 114–138.