BIOINFORMATICS LETTER TO THE EDITOR Vol. 25 no. 14 2009, pages 1802–1804
doi:10.1093/bioinformatics/btp337

Genetics and population analysis

Comment on 'On the inference of spatial structure from population genetics data'

Eric Durand[1], Chibiao Chen[1,†] and Olivier François[1,*]

[1]University Joseph Fourier, INP Grenoble, TIMC-IMAG, Faculty of Medicine, 38706 La Tronche, France

Received on April 27, 2009; revised on May 6, 2009; accepted on June 1, 2009

Associate Editor: Alex Bateman

Contact: Olivier.francois@imag.fr

TESS is a Bayesian clustering program for population genetic analyses which assumes K_{max} clusters and computes posterior estimates for membership coefficients or admixture proportions by updating spatially explicit prior distributions (Chen *et al*., 2007; François *et al*., 2006). Version 2.1 of the program was released in January 2009 (Durand *et al*., 2009a, b). The program applies principles of Bayesian computation, a well-defined background with a long tradition in statistics (Gelman *et al*., 2003). It is based on highly validated methods combining Gibbs sampling and Metropolis–Hastings algorithms. The last versions of TESS implement the Deviance Information Criterion (DIC; Spiegelhalter *et al*. 2002) to help users selecting program runs which best describe their data.

Recently, Guillot (2009) reported that inconsistent results could be produced with the version 1.2 of the program. His criticisms focus on three main points: (i) Using TESS to analyze synthetic data simulated with five clusters, the program does not provide correct estimates of the number of clusters and membership coefficients; (ii) TESS infers spurious clusters when it is applied to data generated under isolation-by-distance models; (iii) Previous results obtained for *Arabidopsis thaliana* rely on a particular choice of network, and they are not robust to changes in this latent model. Here, we address these criticisms in turn, and we argue that they are not based on strong evidence.

To choose the number of clusters and provide rational estimates of membership coefficients, the TESS outputs should be compared with each other for a range of values of K_{max}, as pointed out in the TESS manual and in Chen *et al*. (2007). Guillot's simulation study did not comply with this basic rule. It used a single value, $K_{max} = 10$. This improper use of TESS led to incorrect results which can be improved easily. For example, consider a previously analyzed synthetic dataset from Chen *et al*. (2007) (five islands, dataset `Rep05Fst04`). For this example, we performed one run of TESS for each K_{max} ranging from 2 to 10 and using the default values of the other parameters of the program. The membership coefficients displayed in Figure 1 provide unambiguous evidence of five main clusters in the data. The DIC curve decreases sharply and then exhibits a plateau at $K_{max} = 5$, providing additional support for 5 clusters. The mis-classification rate is <3%. These results are replicable for >80% of the five-island examples (those with $F_{ST} \geq 0.03$). For $K \leq 5$, STRUCTURE (Pritchard *et al*., 2000) performed similarly to TESS, and, according to the criterion of Evanno *et al*. (2005), $K = 5$ is also the most supported value for the example data (Fig. 1 B and D).

The second criticism concerns the behavior of the program applied to data simulated under isolation-by-distance models. The correspondence between Bayesian clustering algorithms and principal component analysis (PCA) may give an explanation for the patterns observed by Guillot. Recently, Patterson *et al*. (2006) suggested that PCA could be a mean to estimate the number of clusters. Under a model of recent divergence of K subpopulations, these authors show that the covariance matrix of allele frequencies in Bayesian clustering models has $K-1$ large eigenvalues, while the other remain small. In other words, there are $K-1$ significant axes of variation. Under limited local dispersal, Novembre and Stephens (2008) found that the patterns observed in the first PC maps display artificial sinusoidal shapes. It is then highly probable that these artificial patterns have their counterparts in Bayesian clustering estimates, leading us to observe artificial clusters regardless of the sampling design. The issues discussed by Guillot are not specific to TESS, and they do not mean that Bayesian clustering algorithms are weak methods when they are applied to data exhibiting some form of isolation-by-distance (Durand *et al*., 2009b).

The third criticism addresses the robustness of estimates of admixture proportions to a change in the hidden spatial network of TESS. This was illustrated with a reanalysis of *A.thaliana* data (François *et al*., 2008). We repeated the complete analysis of *A.thaliana* using the default Dirichlet graph and the recent version of TESS. For each K_{max}, we averaged our results over the 10% runs with the lowest values of the DIC using the program CLUMPP (François *et al*., 2008; Jakobsson and Rosenberg, 2007). Note that this important aspect of the method was not implemented by Guillot (see next paragraph). Figure 2 shows that the DIC selects $K_{max} = 4$. Remark that the admixture estimates provide evidence for four almost identical populations for all K_{max} in 4–9. These results are very similar to those obtained by François *et al*. (2008). They also agree with studies that used either the same or distinct data (Beck *et al*., 2008; Nordborg *et al*., 2005; Ostrowski *et al*., 2006). It is unlikely that these clustering results are artifacts of the choice of a particular network.

In addition to an inappropriate method for choosing the number of clusters, Guillot (2009) contains erroneous quotations of our previous works. Section 2.2.2 in Guillot (2009) states that François

*To whom correspondence should be addressed.
†Present address: CyberOptics (Singapore) Pte Ltd, No. 21, Ubi Road 1, #02-01, Singapore 408724.

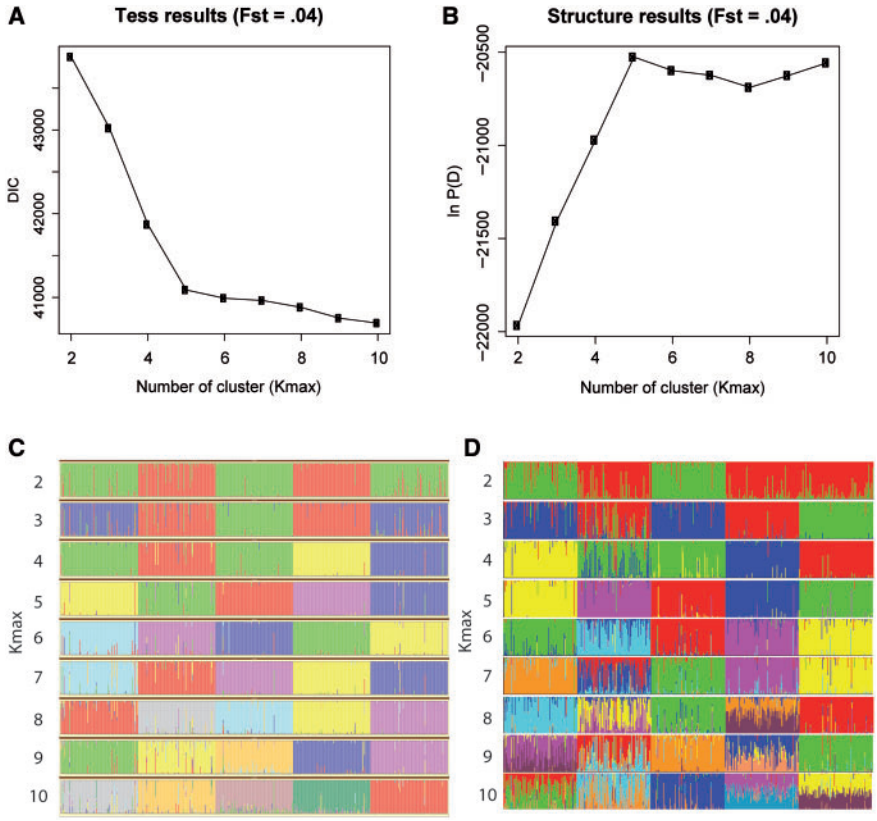

Fig. 1. Analysis of an example dataset (five-islands, $F_{st} = 0.04$). (**A**) DIC for nine TESS runs with K_{max} ranging from 2 to 10. The plateau starting at $K_{max} = 5$ is an indication that the number of clusters is $K = 5$. (**B**) Logarithm of evidence, log P(D), for nine STRUCTURE runs with K ranging from 2 to 10, Evanno *et al*.'s criterion indicates that there are five clusters in the dataset. (**C**) Posterior estimates of cluster membership for TESS. For $K_{max} \geq 5$, five main clusters are visible. (**D**) Posterior estimates of cluster membership for STRUCTURE.

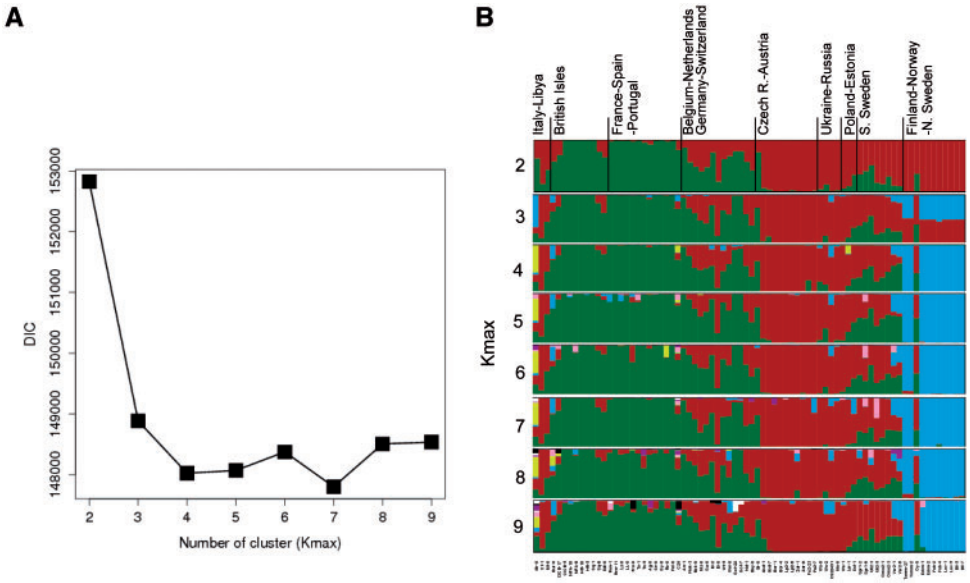

Fig. 2. Analysis of the *A.thaliana* dataset (76 individuals genotyped at 822 loci). (**A**) DIC as a function of K_{max}. Each dot corresponds to the average DIC value computed over the 10 lowest DIC values. (**B**) Posterior estimates of individual admixture coefficients for distinct values of K_{max}. The order of individuals is the same as in Figure 1 of François *et al*. (2008).

et al. 'estimate $[\psi, K]$ as the value that achieves the largest within-run average Deviance Information Criterion' and it gives a mathematical definition for this quantity in Equation (6). There are two mistakes in this sentence. The first one is obvious: 'largest' should be 'smallest' (also at other places in the paper). The second one is more severe. In fact 'within-run average DIC' was not part of the method of François *et al*. (2008). Instead these authors averaged posterior estimates over runs with the lowest values of the DIC. Equation (6) which, according to Guillot, describes François *et al*'s definition for an estimate of ψ and K, is unchangeably wrong. Unlike the likelihood, the DIC is not a function of the parameter, θ, and it is meaningless to compute a within-run average for a constant quantity.

We could question the performances of any Bayesian clustering program applied under the same conditions as in Guillot (2009). For example, we ran the 'without admixture' version of STRUCTURE with $K = 10$ clusters (model M_1) and with $K = 5$ clusters (model M_2) for a distinct five-island example (`Rep02Fst03`). In this example, we based model choice on the weight of evidence of model M_1 against model M_2, computed as the Bayes factor $\mathrm{P}(D|M_1)/\mathrm{P}(D|M_2)$. With $K = 10$, STRUCTURE returned $\log \mathrm{P}(D|M_1) = -21310$, and, with $K = 5$, it returned $\log \mathrm{P}(D|M_2) = -21357$. On Jeffreys' scale, the strength of evidence for $K = 10$ must be very strong (Jeffreys, 1939). As we applied a seemingly valid approach, should we conclude that STRUCTURE is a 'poor inference method'? This would be unwise, as we are ignoring fluctuations. When the values $K = 2, \ldots, 10$ are tested, the $\log \mathrm{P}(D|M)$ curve increases sharply, and, like in Figure 1B, displays a plateau starting at $K = 5$, which is the true value.

Determining the number of clusters in population genetic data is a very difficult issue. We are not aware of any method that could do this in a perfect way, and we do not claim that TESS is the ideal approach. At best, the study by Guillot provide an example of that incorrect use of a statistical criterion can produce erroneous results and lead to misleading conclusions, regardless of the program used.

Conflict of Interest: none declared.

REFERENCES

Beck,J. *et al*. (2008) Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. *Mol. Ecol.*, **17**, 902–915.

Chen,C. *et al*. (2007) Bayesian clustering algorithms ascertaining spatial population structure: A new computer program and a comparison study. *Mol. Ecol. Notes*, **7**, 747–756.

Durand,E. *et al*. (2009a) Tess version 2.1–reference manual. Available at http://membrestimc.imag.fr/Olivier.Francois/tess.html (last accessed date June 2009).

Durand,E. *et al*. (2009b) Spatial inference of admixture proportions and secondary contact zones. *Mol. Biol. Evol.*, doi:10.1093/molbev/msp106.

Evanno,G. *et al*. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.*, **14**, 2611–2620.

François,O. *et al*. (2006) Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics*, **174**, 805–816.

François,O. *et al*. (2008). Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genet.*, **4**, e1000075.

Gelman,A. *et al*. (2003) *Bayesian Data Analysis*, 2nd edn. Chapman & Hall/CRC, New York, USA.

Guillot,G. (2009) On the inference of spatial structure from population genetics data. *Bioinformatics*. in press.

Jakobsson,M. and Rosenberg,N. (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801.

Jeffreys,H. (1939) *Theory of Probability*. Oxford University Press, Oxford, UK.

Nordborg,M. *et al*. (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.*, **3**, e196.

Novembre,J. and Stephens,M. (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genet.*, **40**, 646–649.

Ostrowski,M. *et al*. (2006) Evidence for a large-scale population structure among accessions of *Arabidopsis thaliana*: possible causes and consequences for the distribution of linkage disequilibrium. *Mol. Ecol.*, **15**, 1507–1517.

Patterson,N. *et al*. (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.

Pritchard,J. *et al*. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Spiegelhalter,S.D. *et al*. (2002) Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**, 583–639.