BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–2

Analysing georeferenced population genetics data with Geneland: a new algorithm to deal with null alleles and a friendly graphical user interface.

Gilles Guillot [1]*,Filipe Santos [2] and Arnaud Estoup [2]

[1]Centre for Ecological and Evolutionary Synthesis Department of Biology, University of Oslo, P.O Box 1066 Blindern, 0316 Oslo Norway. Also INRA Applied Math. Dept., Paris, France.
[2]Centre de Biologie et de Gestion des Populations, INRA / IRD / CIRAD / Montpellier SupAgro, Campus international de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez cedex, France.

Received on March 27 2008; revised on April 9 2008; accepted on April 9 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: We introduce a new algorithm to account for the presence of null alleles in inferences of populations clusters from individual multilocus genetic data. We show by simulations that the presence of null alleles can affect the accuracy of inferences if not properly accounted for and that our algorithm improve signficantly their accuracy.

Availability: This new algorithm is implemented in the program Geneland. It is freely available under GNU public license as an R package on the Comprehensive R Archive Network. It now includes a fully clickable graphical interface. Informations on how to get the software are available on `folk.uio.no/gillesg/Geneland.html`

Supplementary material: Details on the simulation study are available from `folk.uio.no/gillesg/BioInformatics_Geneland`

Contact: gilles.guillot@bio.uio.no

1 INTRODUCTION

Bayesian clustering algorithms have become extremely useful tools to investigate the structure of population genetics data [Excoffier and Henkel, 2006] but the conclusion drawn from the use of such algorithms can be markedly influenced by the presence of genotyping errors [Pompanon et al., 2005]. A well known source of such potential problems is the presence of null alleles arising from variation in the nucleotide sequences of flanking regions that prevent the primer annealing to template DNA during PCR amplification of the microsatellite locus [Dakin and Avise, 2004]. The presence of null alleles results in an excess of homozygous genotypes within a population as compared to the expected proportion under Hardy Weinberg Equilibrium (HWE) and Linkage Equilibrium (LE) [Callen et al., 1993, Paetkau et al., 1995]. While all population genetics clustering softwares are based on HWE and LE within the sought clusters, there is no study to date on the effect of null alleles on the accuracy of inferences with such softwares. In this note, we introduce a new statistical model and an MCMC step to explictly take into account the putative presence of null allele(s) in the analysed data set. We briefly illustrate how the presence of null alleles affects the accuracy of inferences with and without using our null allele filtering scheme.

*to whom correspondence should be addressed

2 METHODS

In case the presence of null alleles is suspected, we introduce a difference between the observed genotypes denoted by $z = (z_{i,l})$ (where the subscript i and l refer to the individual and the locus, respectively) and the true non observed genotypes denoted by $y = (y_{i,l})$. For each locus, we introduce an extra fictional allele denoted by ν_l coding for the putative presence of one or several null alleles for which cumulated frequency has to be estimated. The presence of null alleles is taken into account by estimating jointly y and the other parameters of the model in an MCMC simulation. A generic step updating y visits sequentially the genotype of all the individuals at all loci. If $z_{i,l}$ is an heterozygous genotype, there is no ambiguity and $y_{i,l} = z_{i,l}$. If $z_{i,l}$ consists of a double missing data, there is no ambiguity as the true unobserved genotypes consists necessarily of two null alleles and $y_{i,l} = (\nu_l, \nu_l)$. If $z_{i,l} = (\alpha, \alpha)$ there is an ambiguity. The true genotype could be either genuinely homozygous, $y_{i,l} = (\alpha, \alpha)$, or could be $y_{i,l} = (\alpha, \nu_l)$. We denote by θ the vector of all unknown quatities to be inferred (including y). The conditional probability of a genuine homozygous is

$$\pi[y_{i,l} = (\alpha, \alpha) | z_{i,l} = (\alpha, \alpha), \theta_{-y}] = \frac{f_{kl\alpha}}{f_{kl\alpha} + 2f_{kl\nu_l}}$$

where θ_{-y} denotes the vector of all parameters except y and f_{klj} denotes the allele frequency of allele j at locus l in population k. The full conditionnal probability of a presence of a null alleles is $\pi(y_{i,l} = (\alpha, \nu_l) | z_{i,l} = (\alpha, \alpha), \theta_{-y}) = 1 - \pi(y_{i,l} = (\alpha, \alpha) | z_{i,l} = (\alpha, \alpha), \theta_{-y})$. $y_{i,l}$ is hence sampled randomly according to these two probabilities. The other steps of the Markov chain simulation are similar to those described in [Guillot et al., 2005] except that the likelihood is built on y instead of z.

To assess the benefit of using this extra step, we produced data according to the model implemented for simulations in Geneland. Loosely speaking, it produces spatially organised panmictic populations. For each simulated dataset, we tampered with the genotypes of the initial data sets (i.e. the data sets without null alleles) in a way

Table 1. Accuracy in terms of inference of K (percentage of runs where $\hat{K} \neq K$ and where $\hat{K} > K$), and in terms of individual assignment (Error Rate in Co-Assignment). Note: Null alleles filtered/not filtered refers to the use/non use of the new statistical model and MCMC algorithm developed to take into account the putative presence of null alleles. Each number given has been obtained on 500 independant simulated data sets.

	$\%_{\hat{K}\neq K}$	$\%_{\hat{K}>K}$	ERCA
0 % of null alleles			
Null alleles not filtered	0.6	0	1.91
Null alleles filtered	0.4	0	1.55
2 % of null alleles			
Null alleles not filtered	0.598	0	1.76
Null alleles filtered	0.2	0	2.08
10 % of null alleles			
Null alleles not filtered	1.2	1.2	2.01
Null alleles filtered	0	0	1.65
20 % of null alleles			
Null alleles not filtered	13.8	13.8	1.85
Null alleles filtered	0.6	0	2.11

All values as percentages, see supplementary material for details.

that mimics the presence of null alleles with various frequencies. For each simulated data set, we carried out inference of the number of populations K and individuals population memberships. Details on the simulation study are given as supplementary material.

3 RESULTS AND DISCUSSION

Results are shown in Table 1. We found that inferences with Geneland are robust to the presence of a relatively small proportion of null alleles (i.e. less than 10%; Table 1). However, the presence of null alleles at higher proportions (e.g. 20%) substantially alters the accuracy of inferences on the number of populations with a systematic overestimation of K (Table 1, line 7, $\hat{K}$ tends to be larger than K). We note here that, even in the latter case, the accuracy in individual assignments (ERCA_i) remains good as most of the spurious populations inferred contain very few individuals.

With regards to the use of the new statistical model and MCMC algorithm accounting for null alleles, we found that they efficiently restored the accuracy of inferences (Table 1, line 8). Incidentally, we observed that in case one or several null alleles were simulated, their cumulative frequency at each locus were very accurately estimated (results not shown). Interestingly enough, we observed that the use of our new statistical model and MCMC algorithm did not alter the accuracy of inferences if the data set does not contain null alleles (Table 1, lines 1 and 2). Finally, we found that the use of the extra algorithmic step accounting for null alleles had a negligible effects on computing times (an increase of only a few percents depending on the thinning of the chain).

The resilience of Geneland to the presence of null alleles with frequencies up to 10% is fortunate regarding previous studies, as the presence of null alleles at microsatellite loci has been reported frequently in PCR primer characterisation and in population genetics studies [Dakin and Avise, 2004].This resilience can be explained by the fact that in our simulations (and in real data sets as well), null alleles occur spatially at random so that the spatial locations of individuals carrying null alleles do not display any spatial pattern. Therefore, although the presence of null allele creates an excess of homozygous genotypes, this excess can not be repaired by creating spurious populations while maintaining the geometric constraints in the spatial model on which Geneland is based. In agreement with this, we found that, when using Geneland under the non-spatial model option (making the prior model on population membership similar to that of Structure or BAPS in the non-spatial mode), we found that the inferences became largely unreliable with a systematic overestimation of the number of populations, even for low null allele frequencies; For instance, we obtained $\%_{\hat{K}\neq K}$ = 24.3%, $\%_{\hat{K}>K}$= %22.4, and ERCA_i = 8.03% when analysing the data sets with only 2% of null alleles. For mean frequencies of null alleles larger or equal to 20%, the presence of null alleles becomes an issue even when the spatial model option is used. In this case, the accuracy of inferences is efficiently restored when using the new statistical model and MCMC algorithm we specifically proposed for dealing with null alleles. In practice (i.e. when working with a real data set), the presence of null alleles in the analysed data set may often be suspected but their proportions are unknown. Since we found that the use of an extra algorithmic step accounting for their putative presence restores accuracy of inferences when null alleles are present and does not alter the accuracy of inferences if the data set does not contain null alleles, we recommend to carry out inferences with Geneland with this option which does not increase the computing time significantly.

The present algorithm as well as previously existing functionnalities of Geneland are now available through a graphical user interface (GUI). This GUI is written in Tcl/Tk through the R library tcltk. For the growing community of R users in population genetics (see e.g. the related CRAN task view `cran.r-project.org/web/views/Genetics.html`), this new GUI should prove to be very useful as its allows to use Geneland without any knowledge of the R language.

We thank J.M. Cornuet, J.F. Cosson, M. Fontaine, J.M. Marin, F. Mortier, C.P. Robert and G. Roderick for comment at various stages of this work. This work was financially supported by the French Agence Nationale de la Recherche grant No NT05-4-42230.

REFERENCES

D. F. Callen, A.D. Thompson, Y. Shen, H. A. Phillips, R. I. Richards, and J. C. Mulley. Incidence and origin of 'null' alleles in the (ac)n microsatellite markers. *American Journal of Human Genetics*, 52:922–927, 1993.

E.E Dakin and J.C. Avise. Microsatellite null alleles in parentage analysis. *Heredity*, 93(5):504–509, 2004.

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

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

D. Paetkau, W. Calvert, I. Stirling, and C. Strobeck. Microsatellite analysis of population structure in canadian polar bears. *Molecular Ecology*, 4:347–354, 1995.

F. Pompanon, A. Bonin, E. Bellemain, and P. Taberlet. Genotyping errors: causes, consequences and solutions. *Nature Review Genetics*, 6(11):847–859, 2005.