Heredity (2007) 99, 374–380

www.nature.com/hdy

ORIGINAL ARTICLE

Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium

R Kaeuffer[1,2], D Réale[1], DW Coltman[3] and D Pontier[2]
[1]*Canada Research Chair in Behavioural Ecology and GRÉCA, Département des Sciences Biologiques, Université du Québec à Montréal, Montréal, Québec, Canada;* [2]*Laboratoire de Biométrie et Biologie Évolutive, UMR-CNRS 5558, Université de Lyon, université Lyon 1, Villeurbanne cedex, France and* [3]*Departement of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada*

STRUCTURE is the most widely used clustering software to detect population genetic structure. The last version of this software (STRUCTURE 2.1) has been enhanced recently to take into account the occurrence of linkage disequilibrium (LD) caused by admixture between populations. This last version, however, still does not consider the effects of strong background LD caused by genetic drift, and which may cause spurious results. STRUCTURE authors have, therefore, suggested a rough threshold value of the distance (1.0 cM) between two loci below which the pair of loci should not be used. Because of the sensitiveness of LD to demographic events, the distance between loci is not always a good indicator of the strength of LD. In this study, we examine the link between genomic distance and the strength of the correlation between loci (r_{LD}) in a free-ranging population of mouflon (*Ovis aries*), and we present an empirical test of effect of r_{LD} on the clustering results provided by the linkage model in STRUCTURE. We showed that a high r_{LD} value increases the probability of detecting spurious clustering. We propose to use r_{LD} as an index to base a decision on whether or not to use a pair of loci in a clustering analysis.
Heredity (2007) **99,** 374–380; doi:10.1038/sj.hdy.6801010; published online 11 July 2007

Keywords: STRUCTURE software; linkage disequilibrium; population structure; clustering

Introduction

STRUCTURE (Pritchard *et al*., 2000) is the most widely used clustering software applied to detect population genetic structure, with more than 1000 citations for its first version (Pritchard *et al*., 2000) and more than 170 citations for its recent enhanced version (Falush *et al*., 2003) (source: ISI Web of Science database). STRUCTURE generates clusters based on both transient Hardy–Weinberg disequilibrium (HWD) and linkage disequilibrium (LD) caused by admixture between populations. The program works by clustering individuals in groups, where both linkage and HWD are minimized, and therefore, the presence of LD in the data improves clustering results (Falush *et al*., 2003). On the other hand, 'strong' LD or departure from Hardy–Weinberg equilibrium could lead to an overestimation of the number of clusters detected (Falush *et al*., 2003).

STRUCTURE deals with two kinds of LD: the first is mixture LD, which occurs across loci even if they are *unlinked* due to the correlation of allele frequencies 'because individuals with a large component of ancestry in population *k* have an excess of alleles that are common in *k*' (Falush *et al*., 2003). The second is admixture LD, which is 'the correlation that arises between *linked* markers in recently admixed populations' (Pritchard and Wen, 2004). This LD occurs because markers are on the same 'chunk' of chromosome that derives from an ancestral population. The 'admixture model' implemented in the latest version of STRUCTURE (STRUCTURE 2.1; Falush *et al*., 2003) combines admixture LD with map distances between markers to improve clustering results.

Falush *et al*. (2003) defined a third kind of LD: the background LD measured between syntenous loci separated by few cM. Background LD is generated by genetic drift, is expected to be strong at short distance, and can generate spurious clustering (Falush *et al*., 2003). However, the authors have not implemented a way to take that LD into account during clustering yet, and they advise users to avoid too closely linked markers that could be in background LD (Falush *et al*., 2003). Pritchard and Wen (2004) suggested that the distance between markers should not be below 1 cM for humans, although they did not provide empirical evidence in support of this advice. For many species, however, the lack of a genomic map leads to an inability to separate background LD from other types of LD. Furthermore, events other than admixture, such as population bottlenecks (Lynch and Walsh, 1998) or important demographic changes, could also generate strong LD and increase the occurrence of background LD (Jorde, 2000; Peltonen, 2000; Puffenberger *et al*., 1994). In this situation, users should not rely on the genomic map alone to decide

Correspondence: R Kaeuffer, Canada Research Chair in Behavioural Ecology and GRÉCA, Département des Sciences Biologiques, Université du Québec à Montréal, CP 8888 succursale centre-ville, Montréal, Québec, Canada H3C 3P8.
E-mail: kaeuffer.renaud@courrier.uqam.ca
Received 11 October 2006; revised 10 March 2007; accepted 23 March 2007; published online 11 July 2007

which loci to use or not. A measure of the 'strength' of linkage between loci, such as the correlation r_{LD} (Hill and Robertson, 1968) may be therefore more informative than the genomic map to decide which pairs of loci could be used.

Although STRUCTURE is the most widely used program to identify clusters, only few studies have tested its sensitivity to ecological or genetical constraints. For example, Berry *et al*. (2004) and Rosenberg *et al*. (2001) tested the effect of the number of loci on clustering results, whereas Evanno *et al*. (2005) and Waples and Gaggiotti (2006) tested the effects of variation in dispersal rates among populations on the reliability of the number of clusters detected. Furthermore, no study has tested the effect of the strength of LD on clustering results. In this paper, we use knowledge of both the complete history of a mouflon (*Ovis aries*) population and the genomic information about the species to study the potential bias caused by strong LD on clustering with STRUCTURE 2.1. The Kerguelen mouflon population was founded in 1957 by two individuals originating from the Vincennes Zoo (Paris). In 1958, the mouflon started to reproduce, and the population reached the size of 100 individuals at the beginning of the 1970s. The population then grew exponentially, reaching about 700 individuals in 1977, corresponding to a density of about 100 individuals per hectare. Since then, the population has been characterized by cyclical dynamics, fluctuating between 250 and 650 individuals, with winter crashes occurring every 3–5 years (Chapuis *et al*., 1994).

Given the high densities reached during peaks, the low genetic diversity observed (maximum of 4 alleles per loci Kaeuffer *et al*., 2007), the small size of the island and the young age of the population, we do not expect to detect any genetic structure. Furthermore, the strong founder effect in that population has potentially generated strong background LD. This situation favours the possibility of empirically testing for the effect of background linkage on clustering. Finally, we could calculate the distance between syntenous loci based on a genomic map of *Ovis aries* (Maddox *et al*., 2001) and link this distance to the correlation between loci.

In 1993, we sampled 106 individuals that died during the winter crash and for which we knew their exact geographical position. These individuals were genotyped at 22 microsatellite loci: 18 from five linkage groups and seven that were unlinked. We used STRUCTURE 2.1 to estimate the population genetic structure. Using different combinations of loci with variable r_{LD}, we obtained different clustering results and compared them with geographical positions. We then assessed the relationship between the map distance between loci and the strength of LD (that is r_{LD}, a correlation coefficient estimated between pair of loci, Hill and Robertson, 1968), and tested the effects of distance and r_{LD} on the clustering results. We hypothesize that a high r_{LD} value between loci could indicate a potential for the pair of loci to generate spurious clustering with STRUCTURE.

Materials and methods

Population and study site

The population is located on Haute Island, a small island (6.5 km^2) of the Kerguelen archipelago. Kerguelen is a very remote Subantarctic archipelago located in the Southern Indian Ocean (49°20′S, 70°20′E). The climate is Subantarctic, with high precipitation, strong winds and average temperature ranging from 1°C during the winter to 8°C in summer. Rocky landscapes dominate Haute Island, with sparse vegetation cover (about 40%) composed of few endemic species (that is *Azorella selago* and *Agrostis magellanica*) and introduced forage species (that is *Poa* sp. and *Festuca* sp.) (Chapuis *et al*., 1994). The central part of the island is composed by rocky mountains, reaching 321 m high. Mouflons are restricted to shores and low-altitude prairies protected from the dominant winds. In 1993, tissues samples were collected from 106 lambs carcasses from both sexes found on the ground. Positions were recorded using a 125 m^2 grid map of Haute Island.

Genotyping

The samples were kept in 95% ethanol. DNA was extracted using the QIAamp tissue extraction mini kit (Qiagen Inc., Mississauga, Ontario, Canada). PCR amplification was performed at the following 22 ungulate-derived microsatellite loci: ARO28, HEL10, MCM64, MCM152, BM3413, HUJ177, MAF64, MCM527, TGLA13, Ilsts059, TGLA176, RT1, AGLA226, Il2ra, MCM218, NRAMP, OarCP49, TEXAN4, DRBps, INRA26, oMHC1 and TGLA387 (see for details Maddox *et al*., 2001). Reaction conditions (see http://www.thearkdb.org/) were optimized using temperature and $MgCl_2$ gradient PCR. PCR products were analysed with an automatic sequencer (ABI 3730; Applied Biosystems, Foster City, CA, USA) and read using GENEMAPPER 3.5 software (Applied Biosystems).

We characterized variation at each locus, and also tested for departures Hardy–Weinberg equilibrium, using Genepop (Raymond and Rousset, 1995) available at http://www.wbiomed.curtin.edu.au/genepop/genepop_op1.html. Distances between loci were estimated using the SM3 BestPositions, sex averaged on the Arkdb database (http://www.thearkdb.org/) and Maddox *et al*. (2001).

LD

We estimated LD between the 22 loci using the correlation coefficient r_{LD} (Hill and Robertson, 1968). The LD correlation coefficient between all pairs of loci was computed using Linkdos software (Garniergere and Dillmann, 1992) (http://www.wbiomed.curtin.edu.au/genepop/linkdos.html). Then, we used Fisher's exact test available in Genepop (Raymond and Rousset, 1995) (http://wbiomed.curtin.edu.au/genepop/genepop_op2.html) to test if genotypes at one locus are independent from genotypes at other loci. Founder events (Lynch and Walsh, 1998) and population dynamics (Slatkin, 1994) can affect LD. We examined the nature of the relationship between r_{LD} and the distance in cM between two loci in the mouflon population.

Population structure

We used STRUCTURE 2.1 (Pritchard *et al*., 2000; Falush *et al*., 2003) to infer Haute Island population structure. To infer the number of groups, a fully Bayesian process described by Pritchard *et al*. (2000) was run with different values of the number of clusters (K). STRUCTURE would

attribute a probability $\Pr(X \mid K)$ given the data (X), and the $\log \Pr(X \mid K)$ is used to determine the more likely number of clusters (Pritchard *et al*., 2000). However, $\Pr(X \mid K)$ is computationally difficult to estimate, and Pritchard *et al*. (2000, p 948, 958) have proposed an *ad hoc* way to approximate the probability of K given the genotyped data. Pritchard *et al*. (2000, p 950) warned users that the estimated probabilities should be considered as 'a guide to which models are consistent rather than accurate estimates of the posterior probabilities' of K.

STRUCTURE software gives also the assignment probabilities of each individual for each cluster. We used these probabilities to infer the membership of each individual at their most probable groups.

We referred to the sheep genomic map (Maddox *et al*., 2001) for distances between loci for use in the linkage model. The population was young (less than 50 years since the introduction), and we expected to find no population structure or a weak structure. Allele frequencies should probably be correlated between groups, and thus we used the 'correlated allele frequencies' option in the linkage model. The K value that provided the maximum likelihood over the runs was retained as the most probable number of clusters (Pritchard and Wen, 2004). We first ran a series of models with K ranging from 1 to 10, using all loci. We fixed the burn-in period to 500 000 and the running length to 1 000 000 to give consistent results over runs. To verify the consistency of the results, we performed 10 independent runs for each K. Running this first series took us more than 100 h, and we limited the following models to three runs and to a K ranging from 1 to 5. We ran these analyses using three different combinations of selected loci: (1) syntenous loci, (2) non-syntenous loci and (3) syntenous loci separated by large distance (>3 cM). We finally ran STRUCTURE using all the different possible pairs of syntenous loci and estimated the number of clusters for each of these pairs.

We analysed the relationship between the distance separating two syntenous loci and their r_{LD}. Given the non-linear link between the two variables, we fitted a non-linear regression model of the form $y = b_0 + b_1^{\dagger} e^{(-x/th)}$, using the *selfStart* procedure in the package MASS in R (Venables and Ripley, 2002), where y is the r_{LD} and x is the distance between two loci. We used a generalized linear model (logit link function and binomial distribution) to analyse the effect of r_{LD} on the probability of detecting more than one cluster with STRUCTURE. The number of alleles can affect clustering results (Rosenberg *et al*., 2001), and was thus included in the model. Two groups of pairs of syntenous loci (that is with an $r_{LD} < 0.3$ and with an $r_{LD} > 0.5$, respectively) provided completely separated number of estimated clusters (that is 1 or >1 cluster, respectively). To account for the bias caused by the complete separation between these two groups, we ran a bias-robust logistic regression that uses a maximum penalized likelihood estimation (Firth, 1993; Heinze and Schemper, 2002). We used the package *logistf* in R (Ploner *et al*., 2005).

Results

Heterozygosity

The mean number of alleles per locus was 2.5 and ranged from 2 to 4. The average heterozygosity observed is 0.47 ± 0.012 s.e. None of the 22 loci showed any departure from Hardy–Weinberg equilibrium (exact test $P = 0.05$).

LD and the link with distance on the chromosome

Thirty-six of the 231 pairs of loci showed significant LD ($P < 0.05$). The average r_{LD} over the whole set of loci was 0.112 (range: 0.0005–0.878). Of these 36 pairs, eight were from syntenous loci and showed an r_{LD} averaging 0.64, (Table 1). The 28 other pairs in LD were on different chromosomes and were characterized by an average r_{LD} of 0.09. Only 10 pairs of loci, including the eight pairs from syntenous loci were still in significant LD after correcting for multiple testing (Benjamini and Hochberg, 1995). The average r_{LD} of the two non-syntenous pairs in significant LD after correcting for multiple testing was

Table 1 Linkage groups, distance between loci (in cM) and correlations (r_{LD}) between syntenous loci and their associate P-value (in bold significant P-values)

Chromosome	*Linkage group*	*Locus pair*	*Distance (cM)*	r_{LD}	P-*value*
1	1	HUJ177–MAF64	107.3	0.138	0.13
2	2	MCM64–TGLA 13	11.5	0.113	0.068
2	2	MCM64–NRAMP	186.7	0.122	0.226
2	2	MCM64–ARO28	186.7	0.065	0.736
2	2	MCM64–TEXAN4	189.2	0.051	0.775
2	2	TGLA13–NRAMP	175.2	0.092	0.921
2	2	TGLA13–ARO28	175.2	0.047	0.888
2	2	TGLA13–TEXAN4	177.7	0.062	0.994
2	2	NRAMP–ARO28	1	0.878	**<0.0001**
2	2	NRAMP–TEXAN4	2.5	0.851	**<0.0001**
2	2	ARO28–TEXAN4	2.5	0.756	**<0.0001**
5	3	TGLA176–MCM527	105	0.176	**0.004**
13	4	Il2ra–MCM152	42.8	0.034	0.322
13	4	Il2ra–Ilst059	44.4	0.051	0.512
13	4	MCM152–Ilst059	1.6	0.57	**<0.0001**
20	5	oMCH1–DRBps	3	0.6	**<0.0001**
20	5	oMCH1–TGLA387	0.5	0.667	**<0.0001**
20	5	DRBps–TGLA387	2.5	0.701	**<0.0001**

Abbreviation: LD, linkage disequilibrium.

Table 2 Number of clusters inferred by STRUCTURE for different combinations of loci. In bold: most probable number of clusters (highest log Pr(X/K))

K	*All loci*	*Syntenous*	*Non-syntenous*	*Large distance*
1	−3311.2	−1872.0	**−1752.7**	**−2657.9**
2	**−3110.7**	**−1686.1**	−1760.8	−2685.5
3	−3147.5	−1705.1	−1758.4	−2728.2
4	−3254.8	−1740.5	−1760.7	−2704.3
5	−3278.3	−1843.2	−1758.6	−2685.3
6	−3343.2	—	—	—
7	−3414.8	—	—	—
8	−3633.1	—	—	—
9	−3594.0	—	—	—
10	−3415.0	—	—	—

All loci, all the 22 loci used in the study; syntenous, pairs of syntenous loci only; non-syntenous, pairs of non-syntenous only; large distance, pairs of syntenous loci separated by distance $\geqslant$10.5 cM.

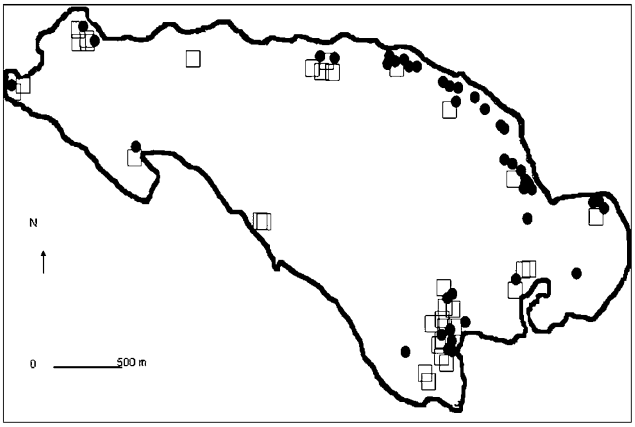

Figure 1 Approximate position and cluster membership of all lambs sampled on Haute Island determined with STRUCTURE software using the whole set of loci (for map readability, individuals sampled from same location were slightly shifted). Circles and squares represent membership of each individual.

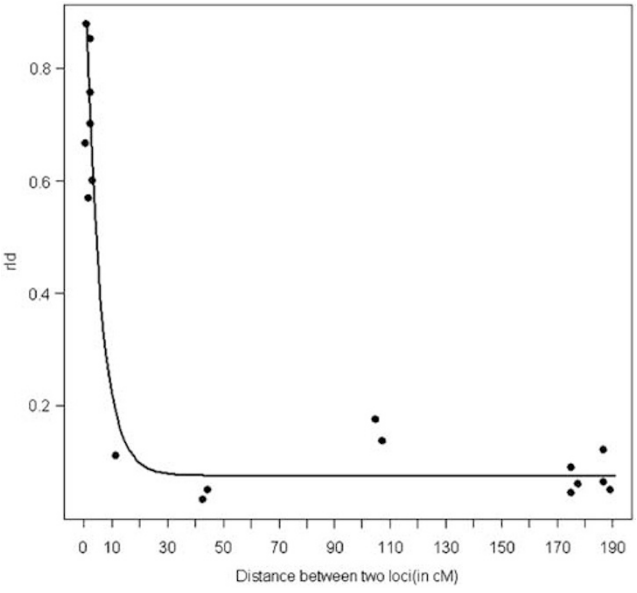

Figure 2 r_{LD} as a function of the distance between two syntenous microsatellite loci (in cM) within a pair.

0.25. Ten other syntenous loci showed a non-significant LD (Table 1).

The correlation coefficient (r_{LD}) between syntenous loci decreased significantly with the distance between them and reached a minimum of 0.08 for distances greater than 30 cM (Figure 2; non-linear regression; best-fitted model parameters: $b_0 \pm$s.e. $= 0.08 \pm 0.03$ ($t = 2.444$; $P = 0.027$), $b_1 \pm$s.e. $= 0.80 \pm 0.09$ ($t = 9.438$; $P < 0.0001$) and $th \pm$s.e. $= 5.26 \pm 1.83$ ($t = 2.873$; $P = 0.012$).

Population structure

Different combinations of loci provided different clustering results (Table 2). Using the whole set of loci, the best fitted model provided two clusters ($K = 2$; Table 2). The estimated membership of each individual in each cluster did not correspond to any obvious geographical structure on the island (Figure 1), suggesting that the number of clusters was overestimated here. Analyses run with combination of syntenous loci alone also provided an estimate of two clusters (Table 2). In contrast, both analyses that used only non-syntenous loci or syntenous loci separated by more that 3 cM provided only one cluster.

The strength of the linkage correlation (that is r_{LD}) between two loci significantly increased the probability of overestimating the number of clusters, in an analysis performed with one pair of syntenous loci at a time (Figure 2). The number of alleles in a pair of syntenous loci and the interaction between r_{LD} and the number of alleles did not affect the probability of overestimating the number of clusters (Figure 2). Eleven out of 18 pairs of loci that estimated one cluster were characterized by a $r_{LD} < 0.12$ (mean $\pm$ s.d. $= 0.09 \pm 0.05$). Four analyses with pairs of loci with a r_{LD} ranging from 0.57 to 0.70 (mean $\pm$ s.d. $= 0.63 \pm 0.06$) provided an estimate of two clusters, and three analyses with pairs of loci with a r_{LD} higher than 0.75 (mean $\pm$ s.d. $= 0.83 \pm 0.06$) estimated three clusters. Finally, analyses using the two pairs non-syntenous loci with a significant LD ($r_{LD} = 0.25$) gave an estimate of one cluster.

Discussion

We detected two subpopulations in the Kerguelen mouflon population using our full dataset and the program STRUCTURE. On one hand, this result could be explained by the fact that groups of related individuals occupied restricted portions of the island.

For example, observations on 18 marked individuals during their first year of life (average number of observation per lamb = 13.6 and range = 10–32) showed that lambs occupied small home ranges (average ± s.e. = 39.0 ± 7.0 Ha) relative to the size of the island (that is 650 Ha). Furthermore, male and female mouflon in Europe are known to be philopatric (Dubois *et al*., 1993, 1995; Petit *et al*., 1997; see also Martins *et al*., 2002) and to reproduce in their natal area (Dubois *et al*., 1995). Coltman *et al*. (2003) have also found weak genetic structure on an insular population of sheep. On the other hand, several lines of evidence suggest this explanation is unlikely. First, the two clusters showed a completely overlapping spatial distribution (Figure 1). Such an overlap of clusters could be explained by frequent dispersal or long movements of related individuals just before the winter crash. This explanation is, however, not consistent with behavioural observations made on the population. All the lambs that died during the winter crash were found within the boundaries of their home range. Furthermore, the population was characterized by a low genetic diversity (maximum of 4 alleles Kaeuffer *et al*., 2007) and a high density relative to other mouflon populations (Boussès and Réale, 1996), which should increase the flow of individuals between areas. Finally, marked males moved easily from one area of the island to another during the rut (Réale *et al*., unpublished), suggesting an unrestricted gene flow within the population. Thus, it is unlikely for the Kerguelen mouflon population to exhibit genetic structure. As mentioned by Pritchard *et al*. (2000), the two clusters found in our analyses do not necessarily have a biological meaning and could be caused by LD between the markers used (Falush *et al*., 2003). To test for the potential bias in STRUCTURE, we used the program Geneland (Guillot *et al*., 2005a) implanted in the R program (R Development Core Team, 2005). This method consists of a Bayesian model implemented in a Markov Chain Monte Carlo that takes into account individual geographical positions, but that does not consider effect of LD on the genetic correlation between populations. Following Guillot *et al*. (2005a, b) recommendations, we first run a model to determine the most probable number of populations. Distribution of posterior probabilities showed a mode at $K = 2$ populations, although the frequency of $K = 1$ was also very high. However, after fixing $K = 2$ and running the model again, all the individuals were assigned to the same population. This result confirms that there is no population genetic structure in the Kerguelen mouflon population (for more discussion on Geneland vs STRUCTURE models see Coulon *et al*., 2006).

In support to the hypothesis that LD can generate spurious number of clusters, we showed that a high r_{LD} between loci strongly affects the probability of detecting more than one cluster in the Kerguelen population (Figure 3). This high r_{LD} (>0.56) was caused by pairs of syntenous loci with distance lower than 3 cM (Table 1 and Figure 2). Our results also indicate that beyond this distance r_{LD} declined dramatically and did not bias the estimation of clusters (Figures 2 and 3). Therefore, a distance higher than 3.0 cM generates one cluster. The impact of such a high r_{LD} between two closely linked loci support the hypothesis that background LD only is responsible for the overestimation of clusters (Falush *et al*., 2003). This is also supported by the fact that the two

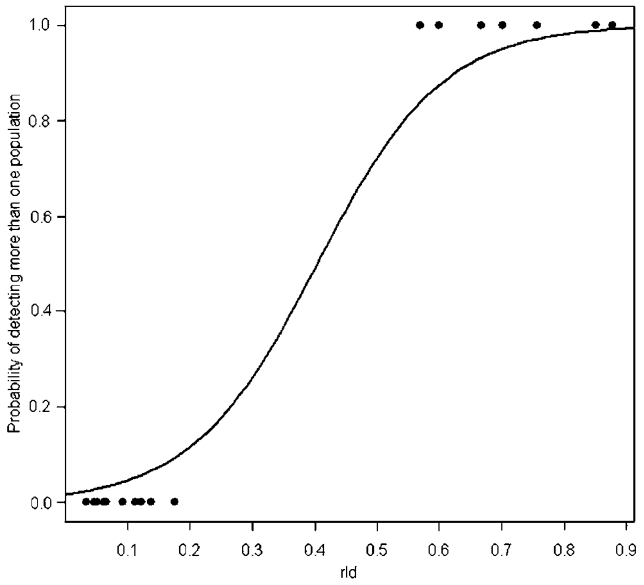

Figure 3 Effect of r_{LD} between two syntenic loci on the overestimation of the number of populations, using the software STRUCTURE and data from the Haute Island population. The model was run using a bias-robust logistic regression (for more information see text) including r_{LD}, number of alleles per locus and their interaction. Only r_{LD} had a significant effect ($\chi^2_1 = 18.5852$, $P < 0.0001$), and the interaction and the number of alleles per locus were not significant ($\chi^2_1 = 0.0004$, $P = 0.984$ and $\chi^2_1 = 0.0015$, $P = 0.9689$, respectively). Number of pairs of syntenous loci = 18.

pairs of unlinked loci that showed a r_{LD} of 0.25 provided only one cluster.

Rosenberg *et al*. (2001) found that increasing the number of loci could improve their clustering results (but see Lecis *et al*., 2006). In our study, analyses using two loci differing in r_{LD} generated different numbers of clusters. This suggests that the strength of the r_{LD} rather than the low number of loci used is responsible for the biased clustering. Moreover, when we used the 22 pairs of loci or only portions of our data set our results indicate that the presence of pairs of loci in strong LD within a sample can generate spurious results (Table 2; comparison of the whole 22 loci vs the non-syntenous loci or loci with large distances), suggesting that STRUCTURE could be sensitive to even a rare pair of loci in strong LD. Our study could not detect an effect of the number of alleles in a pair of loci on the probability of detecting clusters. It should be noted, however, that the allelic diversity in the mouflon population is low and that a higher allelic diversity may have stronger effect.

Our results also indicate that other LD does not bias the clustering results. LD between unlinked loci following a founder event (Lynch and Walsh, 1998) could be confounded with 'mixture' or 'admixture' LD (Falush *et al*., 2003). This could be the case on Haute Island. The population was founded by only two individuals, and some pairs of non-syntenous loci showed stronger LD values than pairs of syntenous loci separated by a large distance. LD obtained from non-syntenous loci or from syntenous loci separated by a large distance, however, did not seem strong enough to generate spurious clustering results (see Table 2).

On the basis of a human study, Pritchard and Wen (2004) advised users against using loci separated by less

than 1 cM. However, in our study, all the loci that were separated by less than 3 cM were characterized by a r_{LD} higher than 0.55 and generated a clustering bias. Furthermore, the selection of loci based on map distance may not be the most efficient approach. First, the strength of LD is not always correlated to physical distance between loci (Jorde *et al*., 1994; Jorde, 1995, 2000). Second genomic maps are still lacking for many species, and the distance between loci is available only for a few species. Finally, the distance *per se* is not responsible for the clustering, but it affects the strength of the linkage that in turn biases the estimation of the number of clusters. LD not only depends on the distance between two loci, but also can be increased by founder events or be decreased by population dynamics (Slatkin, 1994) or number of allele at a given loci (Ott and Rabinowitz, 1997). For all these reasons, we recommend researchers to rely on the r_{LD} to take their decision of using or not using a given pair of loci before running STRUCTURE.

We hope our results will convince researchers that a good knowledge of the LD between loci is an important step before starting analyses in STRUCTURE. As already mentioned by Manel *et al*. (2005) and Waples and Gaggiotti (2006), and despite recommendations made by Pritchard *et al*. (2000) and Falush *et al*. (2003), some authors did not seem to consider LD as a potential issue in the study of population genetic structure. For example, some authors do not report any measure of LD between loci used in STRUCTURE (Kusumo *et al*., 2006). Others used some pairs of closely linked loci (that is about 3 cM in Verardi *et al*., 2006; >0.5 cM in Lecis *et al*., 2006). We do not intend to say that these studies are biased. Results from Lecis *et al*. (2006), for instance, seem consistent to their expectations and the uses of very closely linked loci did not seem to affect their results, potentially because a short distance between two loci may not automatically translate into a strong LD (Peterson *et al*., 1995).

To conclude, our empirical study demonstrates how a strong background LD can lead STRUCTURE to overestimate the number of clusters on a population genetic structure analysis. Therefore, rather than only simply testing for the presence of LD, studies using STRUCTURE should first estimate the r_{LD} between all the pairs of loci before running clustering analyses. This would permit to exclude pairs of loci that could potentially bias the clustering results. For example, in the presence of r_{LD} higher than 0.5 in a sample, one should run two analyses with and without loci with strong r_{LD} and compare the number of clusters detected by STRUCTURE. A different number of clusters may suggest a bias caused by background LD.

Acknowledgements

We thank P Boussès, JL Chapuis, T Micol (TAAF), B Tollu, the Amical des Missions Australes et Polaires Françaises and all fieldworkers who collected mouflon samples and data. A Krupa and A Llewellyn for their help with molecular analyses. We thank RA Nichols and two anonymous reviewers for their comments. This work was supported by the Institut Polaire Français Paul-Emile Victor and the Centre National de la Recherche Scientifique to D Pontier, the Natural Sciences and Engineering Research Council and the Canadian Foundation for Innovation to D Réale, and the Royal Society to D Coltman.

References

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Meth* **57**: 289–300.

Berry O, Tocher MD, Sarre SD (2004). Can assignment tests measure dispersal? *Mol Ecol* **13**: 551–561.

Boussès P, Réale D (1996). Syndrome d'insularité dans une population récente de mouflon (*Ovis musimon*) des îles Kerguelen. *Vie Milieu* **46**: 285–290.

Chapuis JL, Boussès P, Barnaud G (1994). Alien mammals, impact and management in the French Subantarctic islands. *Biol Cons* **67**: 97–104.

Coltman DW, Pilkington JG, Pemberton JM (2003). Fine-scale genetic structure in a free-living ungulate population. *Mol Ecol* **12**: 733–742.

Coulon A, Guillot G, Cosson JF, Angibault JM, Aulagnier S, Cargnelutti B *et al*. (2006). Genetic structure is influenced by landscape features: empirical evidence from a roe deer population. *Mol Ecol* **15**: 1669–1679.

Dubois M, Khazraie K, Guilhem C, Maublanc ML, LePendu Y (1995). Philopatry in mouflon rams during the rutting season: psycho-ethological determinism and functional consequences. *Behav Processes* **35**: 93–100.

Dubois M, Quenette PY, Bideau E, Magnac MP (1993). Seasonal range use by European mouflon rams in medium altitude mountains. *Acta Theriol* **38**: 185–198.

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

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Firth D (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**: 27–38.

Garniergere P, Dillmann C (1992). A computer-program for testing pairwise linkage disequilibria in subdivided populations. *J Hered* **83**: 239.

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

Guillot G, Mortier F, Estoup A (2005b). GENELAND: a computer package for landscape genetics. *Mol Ecol Notes* **5**: 712–715.

Heinze G, Schemper M (2002). A solution to the problem of separation in logistic regression. *Stat Med* **21**: 2409–2419.

Hill WG, Robertson A (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet* **38**: 226–231.

Jorde LB (1995). Linkage disequilibrium as a gene-mapping tool. *Am J Hum Genet* **56**: 11–14.

Jorde LB (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Res* **10**: 1435–1444.

Jorde LB, Watkins WS, Carlson M, Groden J, Albertsen H, Thliveris A *et al*. (1994). Linkage disequilibrium predicts physical distance in the adenomatous polyposis-coli region. *Am J Hum Genet* **54**: 884–898.

Kaeuffer R, Coltman DW, Chapuis JL, Pontier D, Réale D (2007). Unexpected heterozygosity in an island mouflon population founded by a single pair of individuals. *Proc R Soc Lond B Biol Sci* **274**: 527–533.

Kusumo HT, Pfister CA, Wootton JT (2006). Small-scale genetic structure in the sea palm *Postelsia palmaeformis* Ruprecht (Phaeophyceae). *Mar Biol* **149**: 731–742.

Lecis R, Pierpaoli M, Biro ZS, Szemethy L, Ragni B, Vercillo F *et al*. (2006). Bayesian analyses of admixture in wild and domestic cats (*Felis silvestris*) using linked microsatellite loci. *Mol Ecol* **15**: 119–131.

Lynch M, Walsh B (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates Inc.: Sunderland, MA.

Maddox JF, Davies KP, Crawford AM, Hulme DJ, Vaiman D, Cribiu EP *et al*. (2001). An enhanced linkage map of the sheep genome comprising more than 1000 loci. *Genome Res* **11**: 1275–1289.

Manel S, Gaggiotti OE, Waples RS (2005). Assignment methods: matching biological questions techniques with appropriate. *Trends Ecol Evol* **20**: 136–142.

Martins AG, Netto NT, Aulagnier S, Borges A, Dubois M, Vincente L *et al*. (2002). Population subdivision among mouflon sheep (*Ovis gmelini*) ewes and ranging behaviour of rams during the rut. *J Zool* **258**: 27–37.

Ott J, Rabinowitz D (1997). The effect of marker heterozygosity on the power to detect linkage disequilibrium. *Genetics* **147**: 927–930.

Peltonen L (2000). Positional cloning of disease genes: advantages of genetic isolates. *Hum Hered* **50**: 66–75.

Peterson AC, Dirienzo A, Lehesjoki AE, Delachapelle A, Slatkin M, Freimer NB (1995). The distribution of linkage disequilibrium over anonymous genome regions. *Hum Mol Genet* **4**: 887–894.

Petit E, Aulagnier S, Bon R, Dubois M, Crouau-Roy B (1997). Genetic structure of population of the Mediterranean mouflon (*Ovis gmelini*). *J Mammal* **78**: 459–567.

Ploner M, Dunkler D, Southworth H, Heinze G (2005). logistf: Firth's bias reduced logistic regression. R package version 1.03. http://www.meduniwien.ac.at/msi/biometrie/programme/fl/index.html.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Pritchard JK, Wen W (2004). *Documentation for the STRUCTURE software Version 2. Chicago*. http://www.pritch.bsd.uchicago.edu/software/structure2_1.html.

Puffenberger EG, Kauffman ER, Bolk S, Matise TC, Washington SS, Angrist M *et al*. (1994). Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human-chromosome 13q22. *Hum Mol Genet* **3**: 1217–1225.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Raymond M, Rousset F (1995). Genepop (version-1.2) – population-genetics software for exact tests and ecumenicism. *J Hered* **86**: 248–249.

Rosenberg NA, Burke T, Elo K, Feldmann MW, Freidlin PJ, Groenen MAM *et al*. (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* **159**: 699–713.

Slatkin M (1994). Linkage disequilibrium in growing and stable-populations. *Genetics* **137**: 331–336.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, 4th edn. Springer: New York.

Verardi A, Lucchini V, Randi E (2006). Detecting introgressive hybridization between free-ranging domestic dogs and wild wolves (*Canis lupus*) by admixture linkage disequilibrium analysis. *Mol Ecol* **15**: 2845–2855.

Waples RS, Gaggiotti O (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol Ecol* **15**: 1419–1439.