Facing the estimation of effective population size based on molecular markers: comparison of estimators.

B. Jiménez-Mena*,†,‡, E.Verrier*,‡, F. Hospital*,‡

*AgroParisTech, UMR1313 Génétique Animale et Biologie Intégrative, Paris, France,†Bioinformatics Research Center, Aarhus University, Aarhus, Denmark,‡INRA, UMR1313 Génétique Animale et Biologie Intégrative, Jouy-en-Josas, France.

ABSTRACT: We performed a simulation study of several estimators of the effective population size (N_e) ($\widehat{N}_e$based on the rate of decrease in heterozygosity, $\widehat{N}_{eH}$; $\widehat{N}_e$based on thetemporal method,$\widehat{N}_{eT}$, and linkage disequilibrium-based method, $\widehat{N}_{eLD}$). We first focused on$\widehat{N}_{eH}$, which presented an increase in the variability of values over time. The distance from the mean and the median to the true N_e increased over time too. This was caused by the fixation of alleles through time due to genetic drift and the changesin the distribution of allele frequencies. We compared the three estimators of N_e under scenarios of 3 and 20 bi-allelic loci. Increasing the number of loci largely improved the performance of $\widehat{N}_{eT}$ and $\widehat{N}_{eLD}$. We highlight the value of $\widehat{N}_{eT}$ and $\widehat{N}_{eLD}$ when large numbers of bi-allelic loci are available, which is nowadays the case for SNPs markers.

Keywords: Effective population size; Heterozygosity; Genetic drift

Introduction

The effective population size (N_e) is a central parameter in population genetics. It is one of the indicators the most often used to monitor the genetic variability in animal populations under a selection or a conservation program.

$N_e$is defined as the number of individuals in an idealized population that have the same rate of genetic drift as the population under study (Wright, 1931). Many estimators of N_e have been developed, on the basis of pedigree or molecular information. Methods that use molecular markers are becoming very popular due to the decreasing costs of sequencing. However, when estimating the N_e of a population, the values of $\widehat{N}_e$ differ from one method to another.There have been some previous comparisons between the different estimators, but there is not a general and clear consensus about which method to choose under given circumstances.

In this work, we will study in detail one of the estimators ($\widehat{N}_e$ based on the decrease in heterozygosity), and we will make a comparison between this and two other estimators of N_e (based on the temporal and linkage disequilibrium-based method), using different number of loci.

Materials and Methods

Computation of $\widehat{N}_e$ based on the decrease in heterozygosity. This method makes use of the decrease in heterozygosity (*He*) between two different time points. The coefficient of inbreeding (*F*) at generation *t*+1 can be computed as follows:

$$F_{t+1} = \frac{1}{2N_e} + F_t\left(1 - \frac{1}{2N_e}\right) \quad \text{(Eq. 1)}$$

Assuming no mutations, *F*and *He* can be related as:

$$F = 1 - \mathrm{He} \quad \text{(Eq. 2)}$$

We estimated *He* for each locus and averaged it over loci following Nei & Roychoudhury (1974). From equations 1 and 2 we can define Δ*He* as:

$$\Delta\mathrm{He} = \frac{(He_t - He_{t+1})}{He_t} \quad \text{(Eq. 3)}$$

From equations 1 and 3, N_e can be estimated as:

$$\widehat{N}_{eH} = \frac{1}{2\,\Delta\mathrm{He}} \quad \text{(Eq. 4)}$$

Computation of $\widehat{N}_e$based on the temporal method. It is based on the standardized variance in allele frequencies between two different time points ($\hat{V}$). It was first developed in 1971 by Krimbas & Tsakas. We estimated *V* per locus *l* as equation 8 in Waples (1989):

$$\hat{V}_l = \frac{1}{K}\sum_{i=1}^{K} \frac{(x_i - y_i)^2}{0.5\,(x_i + y_i) - (x_i y_i)} \quad \text{(Eq. 5)}$$

K represents the number of alleles, x_i and y_i the frequency of allele *i* at generation *t* and *t*+1, respectively. We averaged $\hat{V}_l$over *L* loci as:

$$\hat{V} = \frac{1}{L}\sum_{l=1}^{L} \hat{V}_l \quad \text{(Eq. 6)}$$

$\widehat{N}_e$ was computed according to equation 12 in Waples (1989); in our case, the sample size is equal to *N*, and the equation was simplified to:

$$\widehat{N}_{eT} = \frac{1}{2\,\hat{V}} \quad \text{(Eq. 7)}$$

Computation of $\widehat{N}_e$ based on linkage disequilibrium method. N_e was estimated using the first of equations 7 of Weir & Hill (1980):

$$\widehat{N}_{eLD} = \frac{1}{3(r^2 - \frac{1}{2N})} \quad \text{(Eq. 8)}$$

where r^2 is the squared correlation coefficient between the pair of loci, averaged for all combinations of loci, and N is the total population size.

Simulations.We modelled a diploid population of N = 1000 individuals, under genetic drift. The population was randomly formed at generation t=0 from a pool of $2N$ different alleles. To form the population at generation t+1, $2N$ genes were randomly chosen from the genes of the population at the previous generation t. To study $\widehat{N}_{eH}$we considered a single locus with $2N$ alleles, and we performed 2000 replicates. For the comparison between the estimators, we used bi-allelic loci and 300 replicates. There was no mutation, selection nor migration. The model assumed discrete generations and constant N. Self-fertilization was allowed. The true N_e was considered to be $N_e = N =$ 1000.N_e was estimated using the information needed for each estimator extracted from the population at generations t and t+1 ($\widehat{N}_{eH}$ and $\widehat{N}_{eT}$), or at generation t ($\widehat{N}_{eLD}$).

Results and Discussion

We studied the evolution of $\widehat{N}_{eH}$under a model of a single multi-allelic locus evolving through time. The variability of values of this estimator increased over time (Figure 1), as well as the variability of *He* (Figure 2). To check whether this variability in $\widehat{N}_{eH}$was coming from the evolution in the number of alleles or the changes in the distribution of allele frequencies, we simulated different scenarios with different starting allele frequencies. We let the population evolve for just one generation to study the effect of the number of alleles and the changes in distribution of allele frequencies separately. The results obtained confirmed that these two factors have an effect on the variation of $\widehat{N}_e$. $\widehat{N}_{eH}$ requires a high number of alleles to converge to the true N_e. As alleles get fixed over time (due to genetic drift), it results in a larger variation of $\widehat{N}_{eH}$.

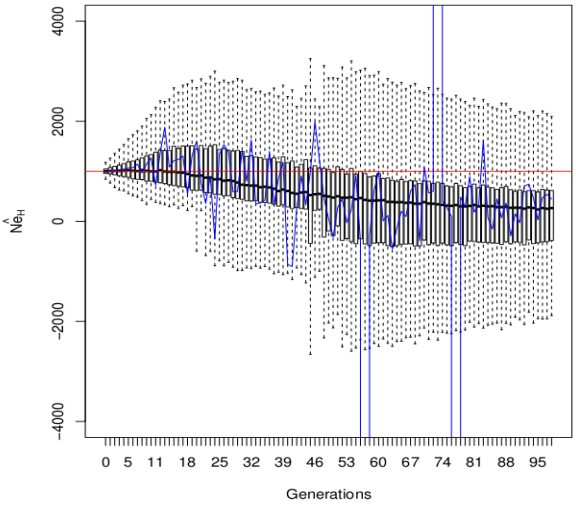

Figure 1: Evolution of $\widehat{N}_{eH}$ over time. The blue, black and red lines represent, respectively, the mean, median and true N_e.

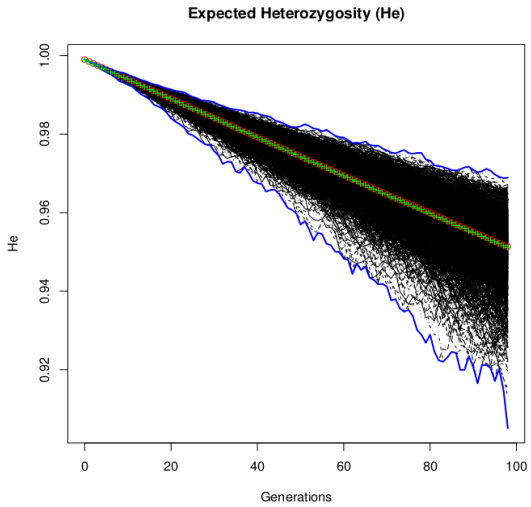

Figure 2: Evolution of *He* over time. One line represents one of the simulations performed. The two blue lines represent the upper and lower limit. The red points represents the mean values of *He* and the green crosses represent the predicted *He*, per generation.

We performed a comparison between the three estimators of N_e ($\widehat{N}_{eH}$, $\widehat{N}_{eT}$ and $\widehat{N}_{eLD}$), studying their evolution through time and using two different numbers of bi-allelic loci (3 and 20 loci). With 3 loci, the estimators had a large variability, which increased through time (Figure 3). As we increased the number of loci, the three estimators improved their performance, which is consistent with previous studies (Leberg, 2005; Palstra & Ruzzante, 2008). With 20loci, $\widehat{N}_{eT}$ and $\widehat{N}_{eLD}$ showed the largestdecrease in variability of $\widehat{N}_e$among the methods and both their mean and median approachedthe true N_e(Figure 4).

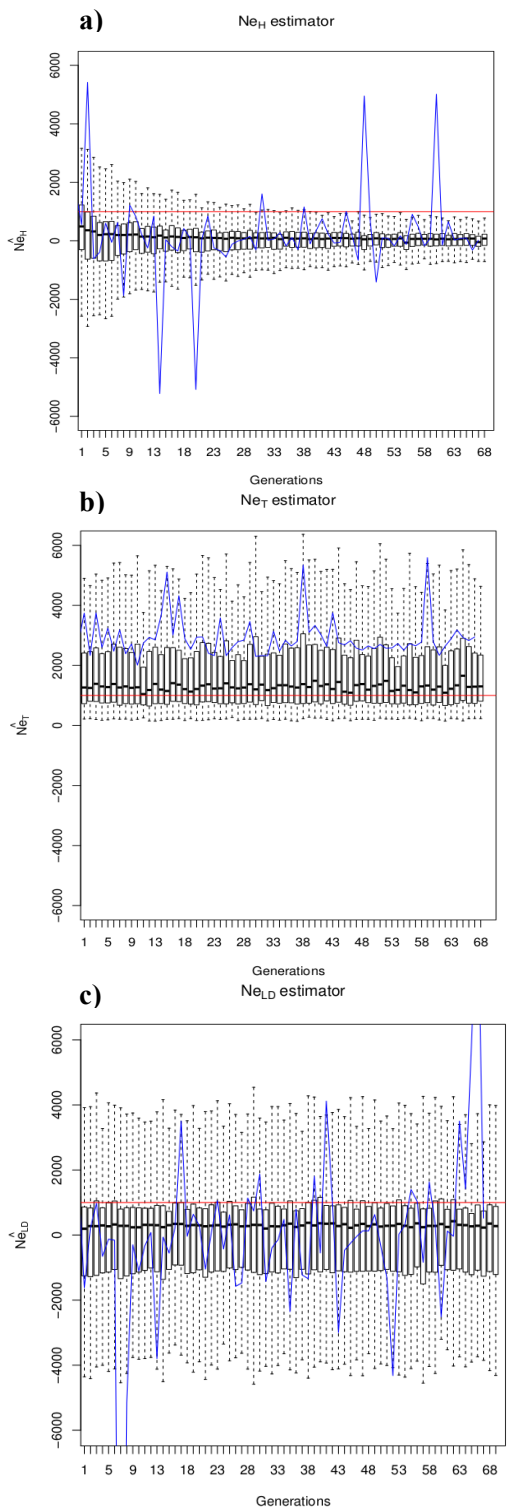

Figure 3: Evolution of the estimators Ne_H (a), Ne_T (b) and Ne_{LD} (c) over time, under the case of 3 bi-allelic loci. The blue, black and red lines represent, respectively, the mean, median and true N_e.

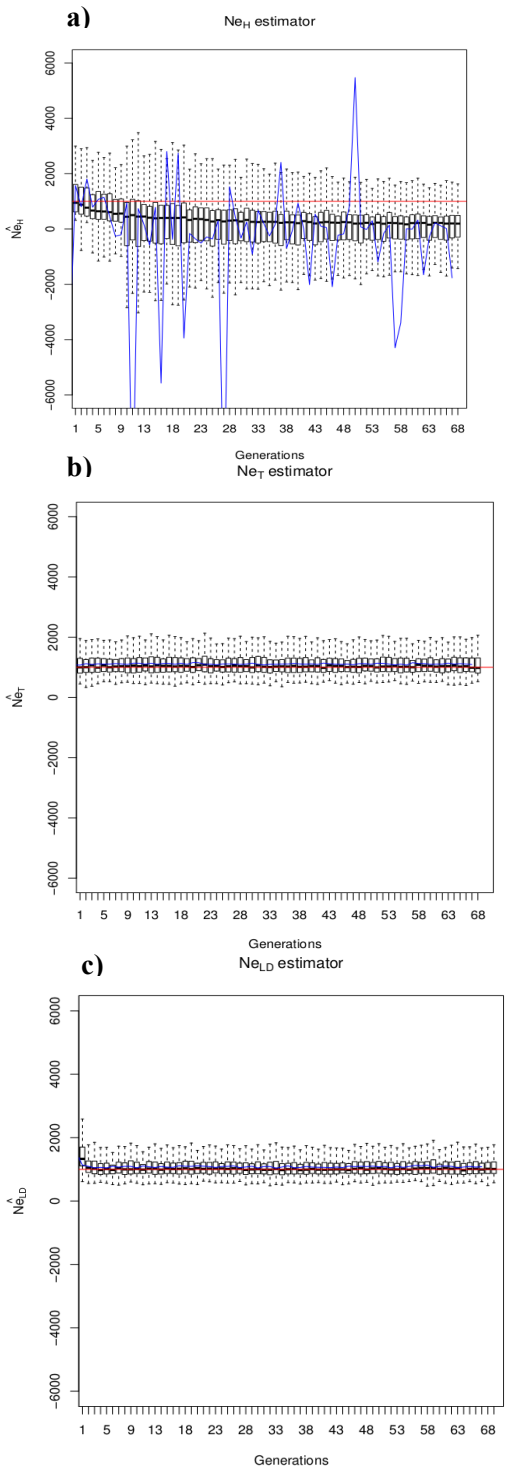

Figure 4: Evolution of the estimators Ne_H (a), Ne_T (b) and Ne_{LD} (c) over time, under the case of 20 bi-allelic loci. The blue, black and red lines represent, respectively, the mean, median and true N_e.

Conclusion

First, we studied the evolution of $\widehat{N}_{eH}$ using a model of a population with a single multi-allelic locus evolving through time. The number of alleles decreased over time due to genetic drift, and this causes a bias in $\widehat{N}_{eH}$. The changes in the distribution of allele frequencies have an effect on this bias too.Secondly, we compared the performance of three estimators of N_e. We showed that all estimators increased their performance when higher numbers of loci were used. $\widehat{N}_{eT}$ and $\widehat{N}_{eLD}$ displayed the most reduced variability of values, and their mean and median remained closer to the true N_e.

Our work suggests that $\widehat{N}_{eT}$ and $\widehat{N}_{eLD}$ perform best when large numbers of loci are used. This is exactly the case of SNPs. This type of marker is becoming very popular nowadays due to its reducing pricing costs. We therefore highlight the use of $\widehat{N}_{eT}$ and $\widehat{N}_{eLD}$ when a large number of SNPs are available. Further research is still needed to provide practical recommendations on the estimators to be used to manage populations, according to their status (selection, conservation…) and the availability and reliability of the required information.

Acknowledgements

The first author benefited from a joint grant from the European Commission and INRA-Animal Genetics division, within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”.

Literature Cited

Krimbas,C.B.&Tsakas,S. (1971).*Evolution,* 25(3):454-460.
Leberg, P. (2005).*J. Wildl. Manag*., 69(4): 1385-1399.
Nei, M. &Roychoudhury, A.K.(1974).*Genetics*, 76: 379-390.
Nei, M. &Tajima, F.(1981).*Genetics*, 98(3): 625-640.
Palstra,F. & Ruzzante, D.E. (2008). *Mol.Ecol.*,17(15):3428-3447
Waples, R. (1989).*Genetics*, 121(2): 379-391.
Weir, B.S. & Hill, W.G. (1980).*Genetics*, 95(2): 477-88.
Wright, F. (1931).*Genetics*, 16(2):97-159.