Characterization of demographic expansions from pairwise comparisons of linked microsatellite haplotypes.

Miguel Navascués, Olivier J Hardy, Concetta Burgarella

Characterization of demographic expansions from pairwise comparisons of linked microsatellite haplotypes

Miguel Navascués[*,†], Olivier J. Hardy[†], Concetta Burgarella[*,‡]

[*] Équipe Éco-évolution mathématique, CNRS UMR 7625 Écologie et Évolution, Université Pierre et Marie Curie and École Normale Supérieure, Paris, France

[†] Service d'Éco-éthologie Évolutive, Université Libre de Bruxelles, Bruxelles, Belgium

[‡] Unidad de Genética Forestal, Departamento de Sistemas y Recursos Forestales, Instituto Nacional para la Investigación Agraria y Alimentaria (CIFOR-INIA), Madrid, Spain

Running head: demographic inference linked microsatellites

Key words: chloroplast microsatellites, Y-chromosome STRs, *Pinus canariensis*, pseudolikelihood, stepwise mutation model

Corresponding author:
Miguel Navascués
Équipe Éco-évolution Mathématique
UMR 7625 Écologie et Évolution
École Normale Supérieure
46 rue d'Ulm
75230 Paris Cedex 05
France
Phone: +33(0)1.44.32.23.41
Fax: +33(0)1.44.32.38.85
Email: m.navascues@gmail.com

Abstract

This work extends the methods of demographic inference based on the distribution of pairwise genetic differences between individuals (mismatch distribution) to the case of linked microsatellite data. Population genetics theory describes the distribution of mutations among a sample of genes under different demographic scenarios. However, the actual number of mutations can rarely be deduced from DNA polymorphisms. The inclusion of mutation models in theoretical predictions can improve the performance of statistical methods. We have developed a maximum pseudolikelihood estimator for the parameters that characterize a demographic expansion for a series of linked loci evolving under a stepwise mutation model. That would correspond to DNA polymorphisms of linked microsatellites (such as those found on the Y-chromosome or the chloroplast genome). The proposed method was evaluated with simulated datasets and with a dataset of chloroplast microsatellites that showed signal for demographic expansion in a previous study. The results show that inclusion of a mutational model in the analysis improves the estimates of the age of expansion in the case of older expansions.

INTRODUCTION

The shape of the genealogy of a random sample of genes (i.e. copies of the same locus) from a population is strongly influenced by the demographic history of the population. For expanding populations the gene genealogy will have a 'star' shape (SLATKIN and HUDSON, 1991), with short internode branches and long terminal branches. Under such genealogies, terminal branches accumulate many mutations producing an 'excessive' number of haplotypes and singletons compared to the expectation for a constant size population. This pattern is exploited in most neutrality test (e.g. TAJIMA, 1989; FU, 1997). In addition, the distribution of pairwise differences between individuals (also known as mismatch distribution) follows a unimodal distribution, in contrast to the ragged patterns that would be found for a sample from a constant size population (SLATKIN and HUDSON, 1991).

In the present work we propose an approach which follows the ideas developed for the study of mismatch distributions of mtDNA for demographic inference (ROGERS and HARPENDING, 1992; ROGERS, 1995; SCHNEIDER and EXCOFFIER, 1999), which are adapted here to linked microsatellites by assuming the stepwise mutation model (KIMURA and OHTA, 1978). The interest of linked microsatellites comes from their application in chloroplast (PROVAN *et al.*, 2001) and mammal Y-chromosome genetic diversity studies (mainly studied in humans, ROEWER *et al.* 1992; WILLUWEIT and ROEWER 2007, but also found in other species, EDWARDS *et al.* 2000; LUO *et al.* 2007). The proposed method is evaluated through simulations and its use is exemplified with a dataset of chloroplast microsatellites for the Canary Island pine (*Pinus canariensis*).

THEORY

Number of mutations between two random gene copies: The classical work by WATTERSON (1975) gives the distribution probability for the number of mutations j occurring between a pair of genes sampled randomly from a population of constant size N:

$$P(j|\theta) = \frac{\theta^j}{(\theta+1)^{j+1}} \tag{1}$$

where $\theta = 2N\mu$ is the effective population size scaled by the mutation rate μ. And LI (1977) obtained the equivalent distribution for the case of a population size change from N_0 to N_1 in a single step t generations ago:

$$P(j|\theta_0, \theta_1, \tau) = P(j|\theta_1) + e^{\left(-\tau\frac{\theta_1+1}{\theta_1}\right)} \times \sum_{j'=0}^{j} \frac{\tau^{j'}}{j'!} \left[P(j-j'|\theta_0) - P(j-j'|\theta_1)\right] \tag{2}$$

where $\theta_0 = 2N_0\mu$, $\theta_1 = 2N_1\mu$, $\tau = 2t\mu$ and $P(j|\theta)$ corresponds to equation 1 for the stationary case.

Fitting equation 2 to the distribution of pairwise genetic differences of a sample of genes has been employed to obtain estimates of the three demographic parameters θ_0, θ_1 and τ (ROGERS and HARPENDING, 1992). However, the use of equation 2 implies that the number of observed genetic differences should correspond to the number of mutations that actually occurred, i.e. an infinite site mutation model (KIMURA, 1969) is being assumed for DNA sequence evolution. This will be an unrealistic model for most datasets in which some amount of multiple hits (i.e. homoplasious mutations) are expected.

Number of differences between two random gene copies: SCHNEIDER and EXCOFFIER (1999) proposed to introduce a mutational model to describe the probabilistic relationship between the number of observed differences, i, and the number of mutations, j, to infer the parameters of a demographic expansion from the distribution of genetic differences between pairs of gene copies. Their strategy consists in using equation 2 and integrate over all possible number of mutations that can produce the observed number of differences:

$$P(i|\theta_0, \theta_1, \tau) = \sum_{j=i}^{\infty} P(j|\theta_0, \theta_1, \tau) P(i|j) \tag{3}$$

where $P(i|j)$ is the probability of observing i differences when j mutations have occurred. Simi-

larly, the stationary state could be described with:

$$P(i|\theta) = \sum_{j=i}^{\infty} P(j|\theta)P(i|j) \tag{4}$$

SCHNEIDER and EXCOFFIER (1999) derived $P(i|j)$ for DNA sequence data with models of nucleotide substitutions with heterogeneous mutation rates across sites. In this work we present the equivalent distribution for the differences between two non-recombining chromosomes typed at several microsatellite loci, assuming a symmetrical stepwise mutation model (KIMURA and OHTA, 1978).

First we will consider the case of a single microsatellite locus. The mutational process, conditioned on the number of mutations j, can be seen as a Markov process of j steps from the state (number of repeats) of one gene to the state of the other gene. Each step might increase or decrease the number of repeats with equal probability. If we defined x as the number of steps in a given direction (let it be the number of steps increasing the number of repeats), x has a binomial distribution, $b\left(x, j, \frac{1}{2}\right) = \binom{j}{x}\left(\frac{1}{2}\right)^j$. The most informative measure of difference available for a pair of microsatellite genes is their absolute difference in number of repeats, δ, which in our Markov process is $\delta = |x - (j - x)|$ and its distribution, from the binomial distribution of x, is:

$$P\left(\delta|j\right) = \begin{cases} \binom{j}{\frac{j}{2}}\left(\frac{1}{2}\right)^j, & \text{if } \delta = 0 \text{ and } j \text{ is even;} \\ \binom{j}{\frac{j+\delta}{2}}\left(\frac{1}{2}\right)^j + \binom{j}{\frac{j-\delta}{2}}\left(\frac{1}{2}\right)^j, & \text{if } \delta \neq 0 \text{ and } (j+\delta) \text{ is even;} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

which is equivalent to equations 22 and 26 from WALSH (2001). Substituting δ for i in equation 3, we can describe the probability distribution of the difference in number of repeats for two microsatellite genes randomly drawn from a population that underwent a demographic expansion.

Now we will consider a non-recombining chromosome (or chromosome fragment) containing L microsatellite loci. We define $\Delta = \{\delta_1, \delta_2, \ldots, \delta_L\}$ as a vector which contains the differences in number of repeats at each locus for a pair of chromosomes. Let k_l equal the number of mutations

at locus l, and let K equal the vector $\{k_1, k_2, \ldots, k_L\}$ with properties: $\sum_{l=1}^{L} k_l = j$; $k_1, \ldots, k_L$ are non-negative integers and $k_l \geq \delta_l$ for any locus l. For each number of mutations j several K vectors can be build, each of them with a probability given by the multinomial distribution:

$$P\left(K = \{k_1, \ldots, k_L\}|j\right) = \begin{cases} \frac{j!}{k_1! \ldots k_L!} p_1^{k_1} \ldots p_L^{k_L}, & \text{if } \sum_{l=1}^{L} k_l = j; \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

where p_l is the probability that a mutation hitting the chromosome hits locus l (i.e. p_l is the ratio between the l locus mutation rate, μ_l, and the global mutation rate for the set of loci, μ_g). Note that, for the multilocus case, θ_0, θ_1 and τ are scaled to global mutation rate μ_g.

Assuming that mutational process is independent among loci, $P(\Delta|K)$ can be calculated as the product of $P\left(\delta_l|k_l\right)$ (from equation 5) for all loci. From $P\left(K|j\right)$ and $P\left(\Delta|K\right)$ it is possible to obtain $P(\Delta|j)$ by integrating over all possible K vectors for j and Δ:

$$P\left(\Delta = \{\delta_1, \ldots, \delta_L\}|j\right) = \sum_{\substack{K=\{k_1,\ldots,k_L\} \\ \sum_{l=1}^{L} k_l = j \\ k_l \geq \delta_l}} P\left(K|j\right) P\left(\Delta|K\right) \quad (7)$$

MATERIALS AND METHODS

Estimation of demographic parameters: In order to obtain estimates $\hat{\theta}_0$, $\hat{\theta}_1$ and $\hat{\tau}$, we fit equation 3 to the distribution of pairwise differences by maximizing the pseudolikelihood of the empirical sample. Let θ_0^*, θ_1^* and τ^* be a combination of possible values for parameters θ_0, θ_1 and τ, we can calculate the likelihood of θ_0^*, θ_1^* and τ^*, for a sample of two chromosomes from equations 3 and 7. For larger samples the pseudolikelihood of θ_0^*, θ_1^* and τ^* is the product of the likelihoods for all possible pairs in the sample. Code in C programming language for the calculation of pseudolikelihood is available from the corresponding author. The combination θ_0^*, θ_1^* and τ^* attaining the highest pseudolikelihood value will be our point estimates $\hat{\theta}_0$, $\hat{\theta}_1$ and $\hat{\tau}$, which we will call *m*aximimum *p*seudolikelihood using a model with *h*omoplasy (MPH) estimates. Parametric boot-

strap confidence intervals around those point estimates can be obtained as described by SCHNEIDER and EXCOFFIER (1999).

Multinomial parameters $(p_1, \ldots, p_L)$ must be provided in order to obtain MPH estimates in the multilocus case. When independent mutation rate estimates are available for every locus, multinomial parameters can be easily calculated by approximating the global mutation rate of the chromosome, μ_g, as the sum of all local mutation rates $\mu_1, \ldots, \mu_L$ (this holds if $\mu_l \ll 1$ for every l). When no mutation rate estimates are available, relative mutation rates can be inferred from allele size variance or homozygosity for loci evolving under a stepwise mutation model (CHAKRABORTY *et al.*, 1997; KIMMEL *et al.*, 1998). Note that the sample from which variance in allele size or homozygosity are estimated does not need to be restricted to the sample from which demographic inferences will be withdrawn, as using additional datasets will increase the precision of those estimates. If relative mutation rate estimates are obtained from the same dataset used for the demographic estimation, the uncertainty in the rates estimates can be taken into account in the parametric bootstrap procedure (see below).

In addition, two other methods to estimate demographic parameters of a population expansion can be proposed for linked microsatellite data. For these we will ignore the problem of recurrent mutation and we will use methods originally described for DNA sequence data evolving under an infinite site model. In order to apply these methods, the Manhattan distance ($D_M = \sum_{l=1}^{L} \delta_l$) will be used to describe differences between two chromosomes (instead of Δ). Assuming that D_M corresponds to the actual number of mutations ($D_M = \hat{\jmath}$), *m*aximum *p*seudolikelihood (MP) estimates can be obtained fitting equation 2 to the distribution of pairwise D_M of the data (this will be equivalent to ROGERS and HARPENDING, 1992, approach). ROGERS (1995) proposed *m*oment based (M) estimates assuming that $\theta_1 \to \infty$ (i.e. θ_1 is very large). For this model $\hat{\theta}_0 = \sqrt{v - m}$ and $\hat{\tau} = m - \hat{\theta}_0$, where m and v are the mean and variance of the number of mutations (D_M in our case) between all pairs of the sample (in practice, if $v < m$ then $\hat{\theta}_0 = 0$ and if $m < \hat{\theta}_0$ then $\hat{\tau} = 0$, as in ROGERS, 1995).

It must be noted that, while these methods (M, MP and MPH) provide estimates for the parame-

ters of a stepwise demographic expansion, they do not provide a formal test for the null hypothesis of a constant neutral demography. Thus, previous to this kind of estimation some evidence of expansion should be obtained, typically through a neutrality test.

Simulations: The accuracy of the three described methods (M, MP and MPH) has been evaluated through coalescent simulations performed with SIMCOAL2 (LAVAL and EXCOFFIER, 2004). Samples of 50 chromosomes were simulated from a population that underwent a step change in its effective population size from θ_0 to θ_1 at time τ before present. Three types of chromosomes (single locus, four loci and eight loci chromosomes, with uniform mutation rate across loci) and ages of the demographic expansion ranging from $\tau = 1$ to $\tau = 9$ were considered. Effective population sizes were between $\theta_0 = 0.01$ and $\theta_0 = 2$, and between $\theta_1 = 1$ and $\theta_1 = 1000$ (always $\theta_0 < \theta_1$). In order to explore the bias and accuracy of the estimates 1000 replicates were run for several scenarios (i.e. combination of parameters: number of loci, age of expansion and effective population sizes before and after expansion). Output of simulations was analysed to obtain estimates $\hat{\theta}_0$, $\hat{\theta}_1$ and $\hat{\tau}$ for the three methods (with the exception of $\hat{\theta}_1$ for the moment method which assumes $\theta_1 \to \infty$).

***Pinus canariensis*, a study case:** *P. canariensis* is an endemic tree of the Canary Islands, whose recent volcanic origin make them an interesting place to study dispersal and colonization processes (JUAN *et al.*, 2000). NAVASCUÉS *et al.* (2006) detected demographic expansion for *P. canariensis* at each island using chloroplast microsatellites and considered these expansions likely linked to the colonization process. However, they considered their estimates for the time of expansion to be biased because of the assumption of a model without homoplasy in the analysis. We have re-analyzed the same data set to obtain estimates for the three methods described above (M, MP and MPH) of the time of expansion for each island. This set comprises 497 individuals from four islands, Tenerife (280 individuals), Gran Canaria (145 individuals), La Palma (48 individuals) and El Hierro (24 individuals), genotyped for six microsatellite loci (Pt15169, Pt30204, Pt71936, Pt87268, Pt26081 and Pt36480; VENDRAMIN *et al.*, 1996). This data has been previously published by GÓMEZ *et al.* (2003), VAXEVANIDOU *et al.* (2006) and NAVASCUÉS *et al.* (2006). Demographic estimates were

not produced for the sample from a fifth island, La Gomera (see NAVASCUÉS *et al.*, 2006). Relative mutation rates among loci were estimated using allele size variance for the whole dataset (including La Gomera sample). For MP estimates confidence intervals were estimated by parametric bootstrap by simulating 1000 samples under an infinite site model and demographic parameters $\hat{\theta}_{0MP}$, $\hat{\theta}_{1MP}$ and $\hat{\tau}_{MP}$ using MS (HUDSON, 2002). Similarly, for MPH confidence intervals parametric bootstrap was performed simulating the evolution of a six microsatellite chromosome, with relative mutation rates as estimated by allele size variance, and demographic parameters $\hat{\theta}_{0MPH}$, $\hat{\theta}_{1MPH}$ and $\hat{\tau}_{MPH}$ using SIMCOAL2 (LAVAL and EXCOFFIER, 2004). There are no reliable mutation rate estimates for chloroplast microsatellites. However, in order to compare $\hat{\tau}$ (in mutation units) with dated geological or biogeographical events, mutation rates in the range 10^{-5}–10^{-4} mutations per generation per locus (and 100 years per generation, PROVAN *et al.*, 1999; NAVASCUÉS *et al.*, 2006) can be used.

RESULTS

Pseudolikelihood profile: Figure 1 shows the pseudolikelihood profiles of the demographic parameters for three example simulated datasets. Pseudolikelihood profiles for τ present narrow bell shapes with the maxima within the proximity of the true value (figure 1a). However, the behavior is quite different for the population size parameters θ_0 and θ_1. For $\theta_1 > 100$ the pseudolikelihood profile often takes an S-shape (figure 1c), and for $\theta_0 < 1$ it often takes an inverted-S-shape (figure 1b). Therefore it is not possible to distinguish very large final population sizes or very small initial population sizes. By increasing the final effective population size (given a fixed τ, note that time is scaled to mutation rate) a threshold is reached from which a coalescent event after the expansion has an extremely low probability and, thus, the shape of the genealogy is the same (terminal branches of length τ) for any θ_1, and the average number of mutations accumulating after the expansion reaches a maximum. By decreasing the initial effective population size another threshold is reached from which a mutation event has an extremely low probability, getting scenarios where no mutations occur before the expansion, regardless of the shape of the genealogy. These

characteristics of the pseudolikelihood profile are helpful to understand the results of the accuracy and bias in the demographic estimates. See Supplemental Material for further discussion about the pseudolikelihood profile.

Bias and accuracy of estimates: The estimation of the effective population sizes before and after the expansion shows a poor performance with biases of some orders of magnitude (figure 2, note logarithmic scale). Only in one case ($\theta_1 = 10$, figure 2b) the estimates are apparently better, but it seems to be for a very narrow range of the parameter values and it still show a tendency for the overestimation of the parameter. A partial explanation of this can be found looking at the shape of the pseudolikelihood profiles of θ_0^* and θ_1^* when they take extreme values. In addition, there seems to be a tendency for pseudolikelihood maxima to be in combinations of extreme values for both θ_0^* and θ_1^* pushing the estimates to have big biases for these two parameters (see, for instance, point estimates for simulations represented in figure 1). Conversely, estimates for the time of expansion are more accurate. Figure 3 presents results for simulations with ages of expansion $\tau = 1$ and $\tau = 7$. Some bias, which increases with the age of expansion, is evident for the two methods that do not account for homoplasy (M and MP). By estimating the number of mutations between two chromosomes by the differences in number of repeats (i.e. D_M) there is a proportion of back and parallel mutations that is missed. Therefore, a younger time of expansion fits well the estimated number of mutation, as it leaves less time for mutations to accumulate. The method that uses a model with homoplasy (MPH) shows a clear improvement over the other two methods, particularly when several linked loci are considered. The reduction of the bias comes with an increase in the variance of the estimator; however, the mean squared error (MSE) for the MPH method was similar or lower than in the other methods (for instance, for simulation of $\tau = 7$ and eight loci, $\widehat{MSE}(\hat{\tau}_M) = 20.23$, $\widehat{MSE}(\hat{\tau}_{MP}) = 10.17$ and $\widehat{MSE}(\hat{\tau}_{MPH}) = 9.15$; and for simulation of $\tau = 1$ and eight loci, $\widehat{MSE}(\hat{\tau}_M) = 0.16$, $\widehat{MSE}(\hat{\tau}_{MP}) = 0.15$ and $\widehat{MSE}(\hat{\tau}_{MPH}) = 0.23$). These results suggest that using a more complex analysis that includes a mutational model for microsatellites can improve the estimates, but only when there has been substantial homoplasious mutation in the sample (i.e. older expansions).

***Pinus canariensis*:** Applying these methods we obtained estimates of the time of expansions (i.e. putative time of colonization) for the *P. canariensis* (table 1). For the MPH method, relative (to total) mutation rates were estimated from allele size variance, giving: $p_{Pt15169} = 0.070$, $p_{Pt30204} = 0.140$, $p_{Pt71936} = 0.530$, $p_{Pt87268} = 0.210$, $p_{Pt26081} = 0.048$ and $p_{Pt36480} = 0.002$. Regardless of the method considered, the geologically older Gran Canaria and Tenerife present older expansions than La Palma and El Hierro in the point estimates. However, taking into account the confidence intervals, the differences in time of expansion among islands are similar in magnitude than the error of the estimates, which makes difficult the interpretation of the results. The major source of error can be attributed to the samples from El Hierro and La Palma as they have a significantly lower sample size. When the analysis takes into account recurrent mutation (by assuming the stepwise mutation model), older times of expansion are obtained, as predicted by NAVASCUÉS *et al.* (2006), except for the island of La Palma. In the light of the results of our simulated data, getting lower estimates in the MPH method that in the MP method is rare but is more frequent for very recent expansion where MPH method can have slightly higher error. Despite the uncertainty around the expansion age estimates and mutation rate of chloroplast microsatellites, the results obtained are roughly congruent with the geological history of the archipelago and with the phylogeography knowledge of the pine specific parasite *Brachyderes rugatus* (see table 1, EMERSON *et al.*, 2000; NAVASCUÉS *et al.*, 2006). However, the analysis seems to gain little from the use of a mutational model due to the young age of the expansions.

DISCUSSION

This work presents an analytical framework to describe the evolution of a set of linked microsatellites between a pair of individuals. WALSH (2001) presented a similar description scaling the process to clock time (i.e. generations). Because of scaling to mutation rate is necessary for demographic inference, a different approach than the one used by WALSH (2001) was used to account for the heterogeneity in mutation rates (i.e. equation 7). As long as relative mutation rates across loci are available, equation 7 can be employed to the analysis of sites evolving under mixed models.

Therefore, it is straightforward to include single nucleotide polymorphisms [for instance, considering them as unique event polymorphisms where $P(i = 0; j = 0) = 1$ and $P(i = 1; j = 1) = 1$] or microsatellites evolving under the geometric step mutation model (following WATKINS, 2007).

Statistical inference from linked microsatellite data can be performed by coalescent MCMC Bayesian (implemented in BATWING software, WILSON and BALDING, 1998; WILSON *et al.*, 2003) and approximate Bayesian (i.e. without likelihoods, PRITCHARD *et al.*, 1999) methods. These are computationally intensive approaches but that are changing the way population genetics analysis is currently done (BEAUMONT and RANNALA, 2004; BEAUMONT, 2004). This applicability of state of the art statistics contrasts to the lack of simple summary statistics specific to linked microsatellites. Using simple statistics is an effective way to describe the genetic diversity of a sample and incorporating mutational models can make them more informative. We demonstrate their usefulness for characterizing demographic expansions, but the same framework could be extended to further statistics, such as neutrality test or estimation of identity by descent probabilities, with linked microsatellites. The development of a neutrality test would be particularly interesting since such a test should be performed before considering the demographic inference under the model of expansion described in this work. A scheme for this could be the estimation of the parameter θ of the constant size model (via BATWING, WILSON and BALDING, 1998, or by maximazing the pseudolikelihood, equation 4), and the use of simulations under the null model to obtain the distribution of some statistic sensitive to demographic expansion, such as the number of haplotypes (FU, 1997; NAVASCUÉS *et al.*, 2006).

The method for demographic inference proposed (MPH) does not substitute the above mentioned coalescent MCMC Bayesian approach (WILSON *et al.*, 2003), which can also be used to characterize demographic expansions; however, it requires less computation time, especially for data sets with large number of individuals. It must be noted that this computation advantage is only attained when a low number of loci is typed. This is because in our approach we integrate over all possible ways to distribute j mutations in l loci (equation 7), which becomes a huge number when j is large and many loci are considered. This increases greatly the computation time, particularly

for highly diverse samples. Nevertheless, most published works on chloroplast or Y-chromosome microsatellites use a number of loci within the limits of computationally 'affordable' (which we suggest around ten, depending on the machine used and the diversity of the sample analysed).

To conclude, the present work extends the methods of demographic inference based in the distribution of pairwise genetic differences to the case of linked microsatellite data. A statistical method to estimate the demographic parameters of a population expansion (i.e. time and magnitude) is developed to be applied on datasets where there is some evidence for demographic expansion (i.e. through a neutrality test). This method assumes that microsatellites follow a stepwise model for mutation to account for homoplasious mutations. This approach improves the estimates of the age of expansion by reducing their bias when the event is relatively old, however little is gained by its application to younger expansion as can be seen both with the simulated and the empirical data analysed.

Acknowledgments: We are grateful to Frantz Depaulis for his support and discussions. Funding for research stays was provided to MN (European Science Foundation ConGen Exchange Grant 1142) and CB (Bourse pour chercheurs étrangers de la Marie de Paris 2007). Comments made by two anonymous referees helped improve a previous version of this paper.

References

BEAUMONT, M., 2004 Recent developments in genetic data analysis: what can they tell us about human demographic history? Heredity **92**: 365–379.

BEAUMONT, M. A., and B. RANNALA, 2004 The Bayesian revolution in genetics. Nature Review Genetics **5**: 251–261.

CHAKRABORTY, R., M. KIMMEL, D. N. STIVERS, L. J. DAVISON, and R. DEKA, 1997 Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proceedings of the National Academy of Sciences of the United States of America **94**: 1041–1046.

EDWARDS, C. J., C. GAILLARD, D. G. BRADLEY, and D. E. MACHUGH, 2000 Y-specific microsatellite polymorphisms in a range of bovid species. Animal Genetics **31**: 127–130.

EMERSON, B., P. OROMÍ, and G. HEWITT, 2000 Colonization and diversification of the species *Brachyderes rugatus* (Coleoptera) on the Canary Islands: evidence from mitochondrial DNA COII gene sequences. Evolution **54**: 911–923.

FU, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147**: 915–925.

GÓMEZ, A., S. GONZÁLEZ-MARTÍNEZ, C. COLLADA, J. CLIMENT, and L. GIL, 2003 Complex population genetic structure in the endemic Canary Island pine revealed using chloroplast microsatellite markers. Theoretical and Applied Genetics **107**: 1123–1131.

HUDSON, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**: 337–338.

JUAN, C., B. C. EMERSON, P. OROMÍ, and G. M. HEWITT, 2000 Colonization and diversification: towards a phylogeographic synthesis for the Canary Islands. Trends in Ecology & Evolution **15**: 104–109.

KIMMEL, M., R. CHAKRABORTY, J. P. KING, M. BAMSHAD, W. S. WATKINS, *et al.*, 1998 Signatures of population expansion in microsatellite repeat data. Genetics **148**: 1921–1930.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**: 893–903.

KIMURA, M., and T. OHTA, 1978 Stepwise mutation model and distribution of allelic frequencies in a finite population. Proceedings of the National Academy of Sciences of the United States of America **75**: 2868–2872.

LAVAL, G., and L. EXCOFFIER, 2004 SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics **20**: 2485–2487.

LI, W.-H., 1977 Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. Genetics **85**: 331–337.

LUO, S.-J., W. E. JOHNSON, V. A. DAVID, M. MENOTTI-RAYMOND, R. STANYON, *et al.*, 2007 Development of Y chromosome intraspecific polymorphic markers in the Felidae. Journal of Heredity **98**: 400–413.

NAVASCUÉS, M., Z. VAXEVANIDOU, S. GONZÁLEZ-MARTÍNEZ, J. CLIMENT, L. GIL, *et al.*, 2006 Chloroplast microsatellites reveal colonization and metapopulation dynamics in the Canary Island pine. Molecular Ecology **15**: 2691–2698.

PRITCHARD, J., M. SEIELSTAD, A. PEREZ-LEZAUN, and M. FELDMAN, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution **16**: 1791–1798.

PROVAN, J., W. POWELL, and P. M. HOLLINGSWORTH, 2001 Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends in Ecology & Evolution **16**: 142–147.

PROVAN, J., N. SORANZO, N. J. WILSON, D. B. GOLDSTEIN, and W. POWELL, 1999 A low mutation rate for chloroplast microsatellites. Genetics **153**: 943–947.

ROEWER, L., J. AMEMANN, N. K. SPURR, K.-H. GRZESCHIK, and J. T. EPPLEN, 1992 Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Human Genetics **89**: 389–394.

ROGERS, A., and H. HARPENDING, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Molecular Biology and Evolution **9**: 552–569.

ROGERS, A. R., 1995 Genetic evidence for a Pleistocene population explosion. Evolution **49**: 608–615.

SCHNEIDER, S., and L. EXCOFFIER, 1999 Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics **152**: 1079–1089.

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**: 555–562.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

VAXEVANIDOU, Z., S. C. GONZALEZ-MARTINEZ, J. CLIMENT, and L. GIL, 2006 Tree populations bordering on extinction: a case study in the endemic Canary Island pine. Biological Conservation **129**: 451–460.

VENDRAMIN, G. G., L. LELLI, P. ROSSI, and M. MORGANTE, 1996 A set of primers for the amplification of 20 chloroplast microsatellites in Pinaceae. Molecular Ecology **5**: 595–598.

WALSH, B., 2001 Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. Genetics **158**: 897–912.

WATKINS, J., 2007 Microsatellite evolution: Markov transition functions for a suite of models. Theoretical Population Biology **71**: 147–159.

WATTERSON, G., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biology **7**: 256–276.

WILLUWEIT, S., and L. ROEWER, 2007 Y chromosome haplotype reference database (YHRD): update. Forensic Science International: Genetics **1**: 83–87.

WILSON, I. J., and D. J. BALDING, 1998 Genealogical inference from microsatellite data. Genetics **150**: 499–510.

WILSON, I. J., M. E. WEALE, and D. J. BALDING, 2003 Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. Journal of the Royal Statistical Society: Series A (Statistics in Society) **166**: 155–188.

Table 1: Age of expansion estimates for *Pinus canariensis* with three methods

Island	$\hat{\tau}_M$	$\hat{\tau}_{MP}$(90%CI)	$\hat{\tau}_{MPH}$(90%CI)	geological age	weevil colonization[b]
Gran Canaria	1.79	1.80(1.24–4.47)	1.98(1.22–2.30)	5.5–3 MYA[a]	>2.56 MYA
Tenerife	1.98	2.38(1.42–2.95)	3.11(1.69–7.14)	3.5–0.2 MYA[a]	1.89–2.56 MYA
La Palma	0.78	0.80(0.44–2.49)	0.49(0.14–0.99)	2 MYA	1.58–2.00 MYA
El Hierro	1.26	1.26(0.43–1.75)	1.61(0.60–2.55)	1 MYA	1.00 MYA

[a] for Gran Canaria and Tenerife the geological age of the last major volcanic period is reported (JUAN *et al.*, 2000).

[b] colonization times estimated for *Pinus canariensis* specific parasite *Brachyderes rugatus* (EMERSON *et al.*, 2000).

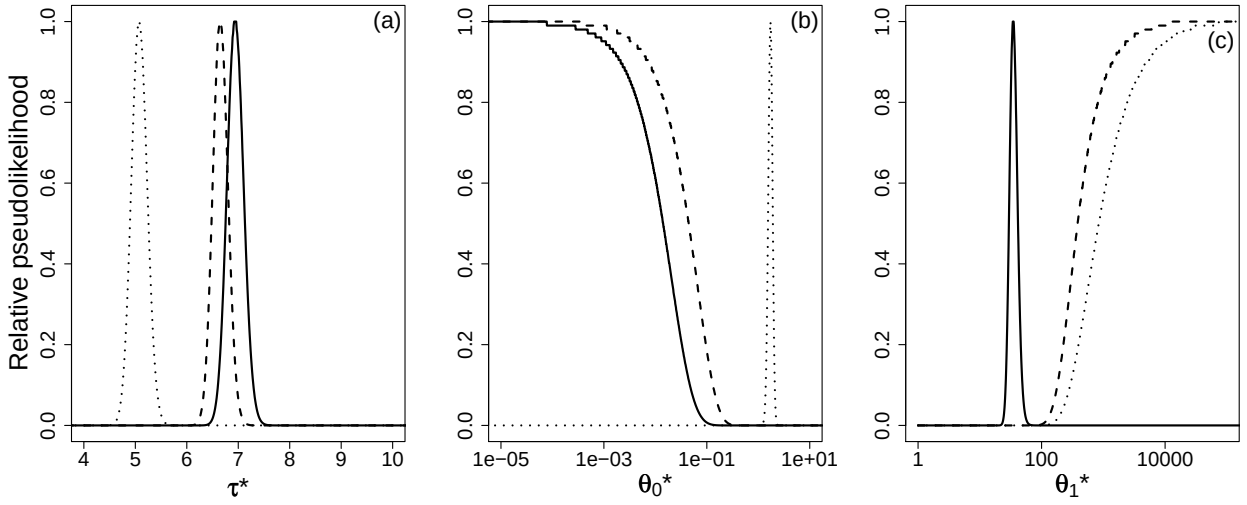

Figure 1: Pseudolikelihood profiles for three simulated dataset. Pseudolikelihood was calculated each time fixing two of the parameters to their true values and letting the third one to vary. Values of pseudolikelihood represented are relative to the maximum value obtained. Data for 50 chromosomes with four microsatellite loci were simulated with parameters: (i) $\tau = 9$, $\theta_0 = 0.01$ and $\theta_1 = 100$ (continuous line), (ii) $\tau = 7$, $\theta_0 = 0.01$ and $\theta_1 = 1000$ (dashed line) and (iii) $\tau = 5$, $\theta_0 = 2$ and $\theta_1 = 1000$ (pointed line). Note that point estimates (maximum psudolikelihood) are, for these three simulations: (i) $\hat{\tau} = 6.7$, $\hat{\theta}_0 = 7.6 \times 10^{-4}$ and $\hat{\theta}_1 = 99835$, (ii) $\hat{\tau} = 6.6$, $\hat{\theta}_0 = 1.4 \times 10^{-4}$ and $\hat{\theta}_1 = 2.1 \times 10^{12}$ and (iii) $\hat{\tau} = 6.7$, $\hat{\theta}_0 = 0.08$ and $\hat{\theta}_1 = 5.8 \times 10^{12}$.

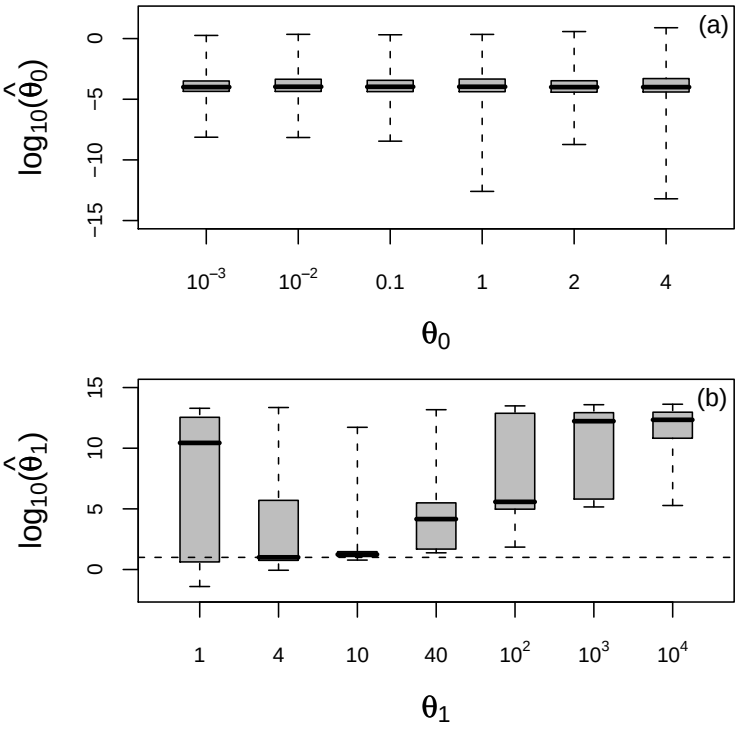

Figure 2: Box-plots showing estimates for the logarithm of population sizes before and after the expansion, $\log_{10}(\hat{\theta}_0)$ and $\log_{10}(\hat{\theta}_1)$, for the MPH method. Datasets were generated by simulation of samples of 50 individuals typed at four linked microsatellite loci. (a) $\log_{10}(\hat{\theta}_0)$ estimates for simulations with demographic parameters $\tau = 7$, $\theta_1 = 10$ and six different values of θ_0. (b) $\log_{10}(\hat{\theta}_1)$ estimates for simulations with demographic parameters $\tau = 7$, $\theta_0 = 0.01$ and seven different values of θ_1. Median is marked with thick black line, box delimits first and third quartiles and whiskers extend to 5% and 95% percentiles. Horizontal dashed line on (b) marks $\theta_1 = 10$ for reference. Distribution for each scenario is build from 1000 replicates.

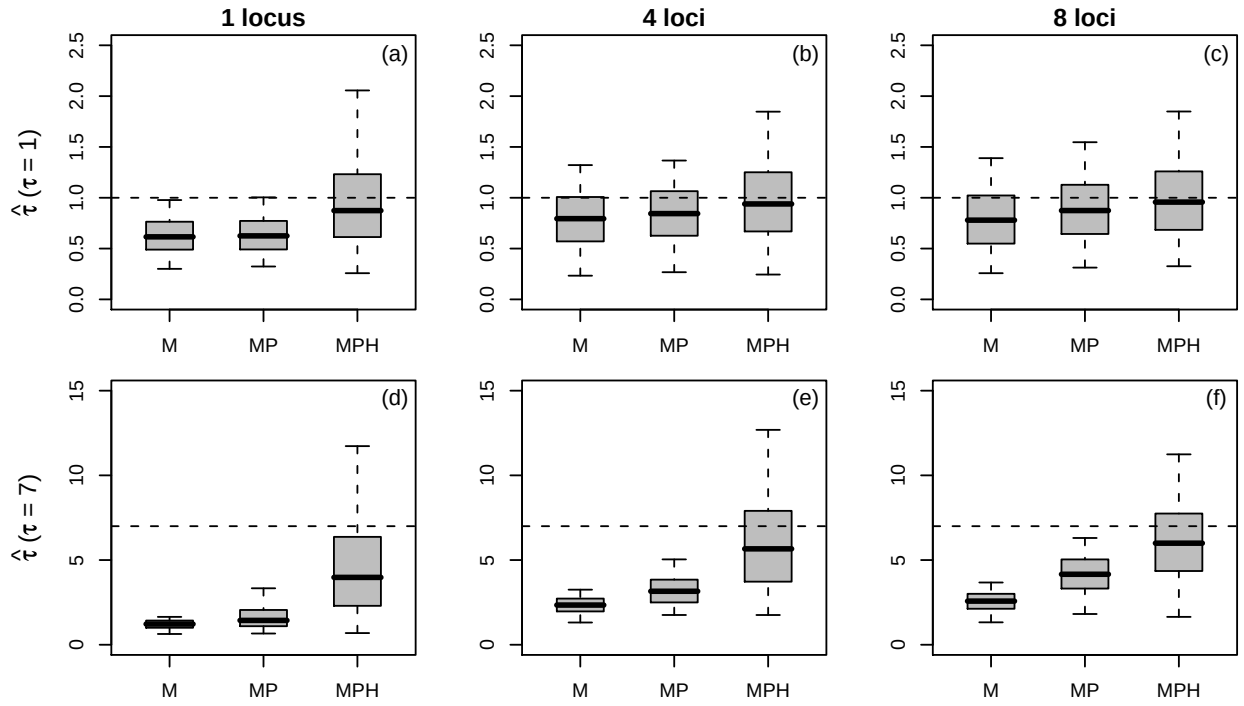

Figure 3: Box-plots showing estimates of time of expansion, $\hat{\tau}$, for the three methods evaluated on six simulation scenarios. The six simulated scenarios differ in the number of loci [one for (a) and (d), four for (b) and (e) and eight for (c) and (f)] and the age of the expansion [$\tau = 1$ for (a), (b) and (c); $\tau = 7$ for (d), (e) and (f)]. Median is marked with thick black line, box delimits first and third quartiles and whiskers extend to 5% and 95% percentiles. Distribution for each scenario is build from 1000 replicates. Horizontal dashed lines mark the true value of parameter τ. M: *m*oment based estimates (ROGERS, 1995); MP: *m*aximum *p*seudolikelihood estimates using a model without homoplasy (equivalent to ROGERS and HARPENDING, 1992) and MPH: *m*aximum *p*seudolikelihood estimates using a model with *h*omoplasy.

Supplemental Material for

Characterization of demographic expansions from pairwise comparisons of linked microsatellite haplotypes

by Miguel Navascués, Olivier J. Hardy and Concetta Burgarella

Pseudolikelihood vs. likelihood

Pseudolikelihood is a statistic used as an approximation of the likelihood by the product of non-independent likelihoods. Thus, one may ask how much this approximation affects the inference. Likelihood and pseudolikelihood surfaces for the different parameters were compared to evaluate the behavior of the pseudolikelihood function under constant size and stepwise expansion models.

The pseudolikelihood profile of the parameter $\theta = 2N\mu$ for a given dataset under the constant population size model can be straightforward calculated from equation 4; the likelihood profile can be estimated through the MCMC algorithm implemented in the software BATWING (WILSON *et al.*, 2003). Although this software is designed within the Bayesian statistical framework, setting an improper prior for θ [uniform(0,∞)] will give a posterior probability estimate proportional to the likelihood. As coalescent scaling in BATWING is set to the mutation rate of a single locus, while our scaling is to the global mutation rate of the chromosome, rescaling can be approximated as $\theta \approx \sum_{l=1}^{L} \theta_{\text{BATWING}}$.

Datasets of 50 chromosomes sampled from a constant size population ($\theta = 10$) typed at four linked microsatellites were simulated with SIMCOAL2 (LAVAL and EXCOFFIER, 2004). The pseudolikelihood profile was calculated from equation 4 in main article and the likelihood profile was obtained from BATWING (burn-in of 20 000 steps plus chain of 200 000 steps) for each simulation. Both pseudolikelihood and likelihood curves presented maximum values around the true value of the parameter θ (figure S1). The main difference between them consists on the narrower shape for the pseudolikelihood curve. While the likelihood curve might be used for the estimation of confidence intervals, the pseudolikelihood profile is inadequate for this purpose as it would give too narrow intervals.

Currently, there is no MCMC implementation of the stepwise demographic expansion for the estimation of likelihood curves (BATWING uses a model of exponential expansion, WILSON *et al.*, 2003). However, it is possible to calculate a true likelihood from our model if independent pairs of chromosomes are sampled from the same demographic history. Two scenarios can be imagined for this sampling scheme. Under the first one, two individuals are sampled from the population and are typed at different independent (i.e. unlinked) loci, each one mutating at the same rate and composed by the same number of linked microsatellite sites. The second scenario would consist on independent sampling ('with replacement') of several pairs of individuals from the same population. Although both scenarios are unrealistic, such sampling scheme allow us to compare the behavior of the pseudolikelihood function with the likelihood.

In order to generate this type of data we simulated 1225 samples of two individuals, typed at four linked microsatellites evolving under the stepwise mutation model and with the same mutation rate, with SIMCOAL2. The demographic scenario was an expansion with $\tau = 7$, $\theta_0 = 0.01$ and $\theta_1 = 10$. Likelihoods for each pair are computed from equation 3 and their product correspond to a true likelihood. In order to obtain an equivalent pseudolikelihood, a sample of 50 individuals (which gives 1225 pairs for the computation of the pseudolikelihood) were also simulated under the same demographic scenario of expansion. Figure S2 presents the obtained curves for four simulations of each type (i.e. 1225 independent pairs for likelihood, 1225 pairs from 50 individuals for the pseudolikelihood). To obtain the curve for each parameter the remaining two parameters were set to their true

value. The shape of the curves are similar for likelihood and pseudolikelihood (see also figure 1 in main article) but likelihood maxima are located closer to the true values of the parameters.

In conclusion, the (maximum) pseudolikelihood approach seems to have an adequate performance for obtaining point estimates. However, the amplitude of the curve correspond to that of a sampling scenario of a larger amount of data, and no confidence intervals should be produced directly from the pseudolikelihood profile.

Pseudolikelihood vs. sum of squared differences

The pseudolikelihood inference approach presented in this work is an extension of the method developed by SCHNEIDER and EXCOFFIER (1999). The main difference between them is the mutational model, since SCHNEIDER and EXCOFFIER (1999) were interested in DNA sequence data while this work targets microsatellites. In both cases a mutational model without homoplasy can be assumed: infinite site model (ISM) for DNA sequence or RFLP data, and the unnamed model which assumes that Manhattan distance (D_M) corresponds to the number of mutation for microsatellites. For both models can be described with equations 1 and 2 in main article as long as appropriate distance metric is employed in the different types of markers.

Another difference is the statistic employed for the fitting of the demographic parameters. SCHNEIDER and EXCOFFIER (1999) obtained the demographic estimates by minimizing the sum of squared differences (SSD) between the observed and expected mismatch distributions. In fact, for a mutational model with no recurrent mutations using any SSD or pseudolikelihood would be largely equivalent and should produce similar results.

We have evaluated our implementation of the MP method (which follows a model without homoplasy) by comparing its results with the approach using SSD as implemented in Arlequin 3.0 (EXCOFFIER *et al.*, 2005). Thousand simulations of 50 individuals typed at four linked microsatellites were run with demographic parameters: $\tau = 7$, $\theta_0 = 0.01$ and $\theta_1 = 10$. For each simulated dataset, MP estimates were obtained as described in the Materials and Methods section of main article. Estimates using SSD were obtained with Arlequin as described in NAVASCUÉS *et al.* (2006). It must be noted that Arlequin is originally designed to compute the statistics under the ISM for DNA sequence or RFLP data. Nevertheless, microsatellite data can be coded in a non standard way (as RFLP, see NAVASCUÉS *et al.*, 2006) so the computation of pairwise differences between haplotypes correspond exactly to D_M, thus, providing a genetic pairwise distribution identical to the one used in the MP estimation.

The results of this comparison are shown in figure S3. Estimates using either of the statistics seem to be congruent in general for τ and for some range of the values of θ_0 and θ_1. The big differences observed must be because the particular characteristics of both implementations. One clear difference is some limits in the values of the parameters θ_0 and θ_1 for Arlequin since, beyond certain threshold, estimates get a fixed value. Other implementation differences affecting the estimates can be due to the precision of calculations or the optimization algorithm.

References

EXCOFFIER, L., G. LAVAL, and S. SCHNEIDER, 2005 Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online **1**: 47–50.

LAVAL, G., and L. EXCOFFIER, 2004 SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics **20**: 2485–2487.

NAVASCUÉS, M., Z. VAXEVANIDOU, S. GONZÁLEZ-MARTÍNEZ, J. CLIMENT, L. GIL, *et al.*, 2006 Chloroplast microsatellites reveal colonization and metapopulation dynamics in the Canary Island pine. Molecular Ecology **15**: 2691–2698.

SCHNEIDER, S., and L. EXCOFFIER, 1999 Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics **152**: 1079–1089.

WILSON, I. J., M. E. WEALE, and D. J. BALDING, 2003 Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. Journal of the Royal Statistical Society: Series A (Statistics in Society) **166**: 155–188.

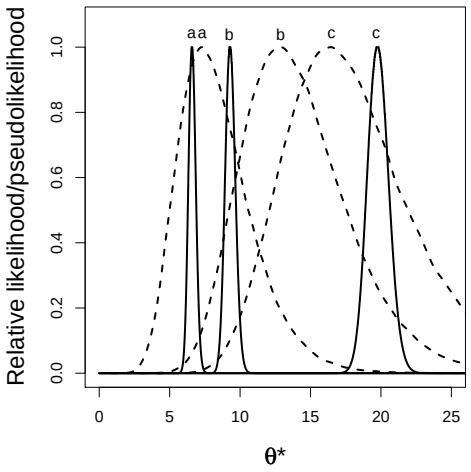

Figure S1: Relative likelihood and pseudolikelihood of the parameter $\theta = 2N\mu$ for three simulations (a, b and c) of a constant size population ($\theta = 10$). A sample of 50 chromosomes was simulated, typed at four linked microsatellites evolving under the stepwise mutation model and with the same mutation rate. Likelihood surfaces (dashed lines) were estimated with BATWING using improper prior on θ [uniform(0,∞)]. Pseudolikelihood (continuous line) was calculated from equation 4. Values represented are relative to the maximum value obtained for each dataset.

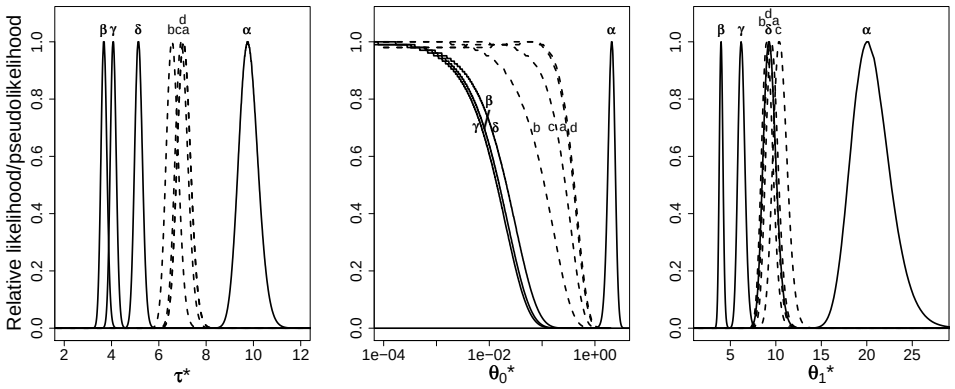

Figure S2: Relative likelihood and pseudolikelihood of the parameter τ, θ_0 and θ_1. Four simulation of 1225 independent pairs (a, b, c and d; dashed lines represent their likelihood profiles) and four simulations of 50 individuals (α, β, γ and δ; continuous lines represent their pseudolikelihood profiles) are represented. To obtain the curve for each parameter the remaining two parameters were set to their true value. Values represented are relative to the maximum value obtained for each dataset. Demographic scenario consisted on a population expansion with parameters $\tau = 7$, $\theta_0 = 0.01$ and $\theta_1 = 10$.

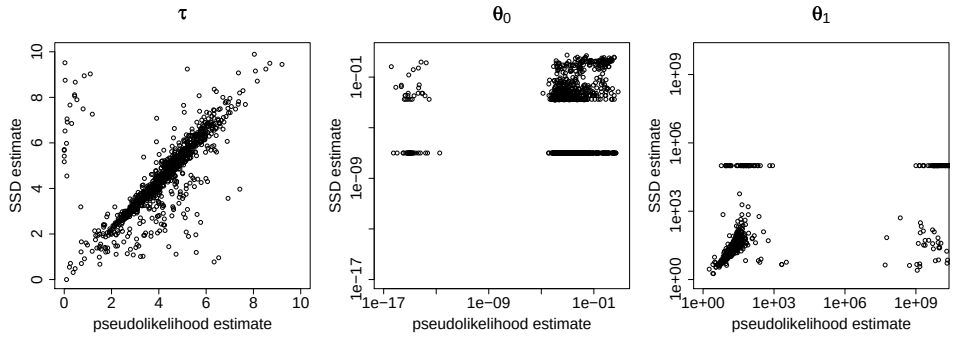

Figure S3: Relationship between point estimates obtained by maximization of the pseudolikelihood or by minimization of the sum of squared differences. Thousand simulations of 50 individuals typed at four linked microsatellites are presented. Demographic parameters were: $\tau = 7$, $\theta_0 = 0.01$ and $\theta_1 = 10$. For each dataset, point estimates of the demographic parameters were obtained by: (i) maximization of pseudolikelihood (MP) and (ii) minimization of sum of squared differences (SSD) between expected and observed mismatch distribution (as implemented in Arlequin 3.0, EXCOFFIER *et al.*, 2005).