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, the genetic diversity of the sample presents particular features. First, 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 focus on mismatch distribution to infer demographic processes. The proposed approach 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 their 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 $N$: 
\[ P(j|\theta) = \frac{\theta^j}{(\theta + 1)^{j+1}} \]  

(1)

where \( \theta = 2N\mu \) is the effective population size scaled by the mutation rate \( \mu \). And L\textsuperscript{t} (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:

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

(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 \( \theta_0 \), \( \theta_1 \) and \( \tau \) (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:** To extract demographic information from empirical data, it will be necessary to obtain the distribution of the number of observed genetic differences between two 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 pairwise genetic differences. Their strategy consists in using equation 2 and integrate over all possible number of mutations that can produce the observed number of differences:

\[ P(i|\theta_0, \theta_1, \tau) = \sum_{j=i}^{\infty} P(j|\theta_0, \theta_1, \tau) P(i|j) \]  

(3)
where \( P(i|j) \) is the probability of observing \( i \) differences when \( j \) mutations have occurred. Similarly, the stationary state could be described with:

\[
P(i|\theta) = \sum_{j=i}^{\infty} P(j|\theta)P(i|j)
\]  (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. A symmetrical stepwise mutation model (Kimura and Ohta, 1978) will be assumed for the mutational process of microsatellites.

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, \( b(x, j, \frac{1}{2}) = \binom{j}{x} \left( \frac{1}{2} \right)^{x} \left( \frac{1}{2} \right)^{j-x} \). The most informative measure of difference available for a pair of microsatellite genes is their absolute difference in number of repeats, \( \delta \), which in our Markov process is \( \delta = |x - (j - x)| \) and its distribution, from the binomial distribution of \( x \), is:

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

which is equivalent to equations 22 and 26 from Walsh (2001). Substituting \( \delta \) 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:

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

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, $\mu_l$, and the global mutation rate for the set of loci, $\mu_g$). Note that, for the multilocus case, $\theta_0$, $\theta_1$ and $\tau$ are scaled to global mutation rate $\mu_g$.

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

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

**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 $\theta_0^*, \theta_1^*$ and $\tau^*$ be a combination of possible values for parameters $\theta_0$, $\theta_1$ and $\tau$, we can calculate the likelihood of $\theta_0^*, \theta_1^*$ and $\tau^*$, for a sample of two chromosomes from equations 3 and 7. For larger samples we can calculate the pseudolikelihood of $\theta_0^*, \theta_1^*$ and $\tau^*$ as the product of their 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 $\theta_0^*, \theta_1^*$ and $\tau^*$ attaining the highest pseudolikelihood value will be our point estimates $\hat{\theta}_0$, $\hat{\theta}_1$ and $\hat{\tau}$, which
we will call *maximum pseudolikelihood* using a model with *homoplasy* (MPH) estimates. Parametric bootstrap 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, \(\mu_g\), 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 \((D_M = \sum_{i=1}^{L} \delta_i)\) will be used to describe differences between two chromosomes (instead of \(\Delta\)). Assuming that \(D_M\) correspond to the actual number of mutations \((D_M = j)\), *maximum pseudolikelihood* (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 *moment based* (M) estimates assuming that \(\theta_1 \to \infty\) (i.e. \(\theta_1\) 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 parameters 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 \( \theta_0 \) to \( \theta_1 \) at time \( \tau \) 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.
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 $\tau$ 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 $\theta_0$ and $\theta_1$. 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 $\tau$, 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 $\tau$) for any $\theta_1$, 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 before the expansion). 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 $\theta_0^*$ and $\theta_1^*$ 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 $\theta_0^*$ and $\theta_1^*$ 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 it must be noted that 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, $MSE(\hat{\tau}_M) = 20.23$, $MSE(\hat{\tau}_{MP}) = 10.17$ and $MSE(\hat{\tau}_{MPH}) = 9.15$; and for simulation of $\tau = 1$ and eight loci, $MSE(\hat{\tau}_M) = 0.16$, $MSE(\hat{\tau}_{MP}) = 0.15$ and $MSE(\hat{\tau}_{MPH}) = 0.23$). 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: \( p_{P_{15169}} = 0.070, p_{P_{30204}} = 0.140, p_{P_{171936}} = 0.530, p_{P_{187268}} = 0.210, p_{P_{26081}} = 0.048 \) and \( p_{P_{36480}} = 0.002 \). 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) instead to the mutational time necessary for demographic inference. Because of scaling to mutation rate is used in this work, it was necessarily a different
approach than the one used by WALKSH (2001) 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 contrast 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. In the present work we demonstrate the usefulness of this kind of statistics 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. Particularly interesting would be the development of a neutrality test since such a test should be performed before considering the demographic inference under a model of expansion described in this work. A scheme for this could be the estimation of the parameter $\theta$ of the constant size model (via BATWING, WILSON and BALDING, 1998, or by maximizing the pseudolikelihood, equation 4), then simulate with that estimate 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 can produce some demographic estimates in 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: 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


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.


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

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

\textsuperscript{a} for Gran Canaria and Tenerife the geological age of the last major volcanic period is reported (JUAN et al., 2000).

\textsuperscript{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 $\theta_0$. (b) $\log_{10}(\hat{\theta}_1)$ estimates for simulations with demographic parameters $\tau = 7$, $\theta_0 = 0.01$ and seven different values of $\theta_1$. 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 built 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 $\tau$. 

M: moment based estimates (ROGERS, 1995); MP: maximum pseudolikelihood estimates using a model without homoplasy (equivalent to ROGERS and HARPENDING, 1992) and MPH: maximum pseudolikelihood estimates using a model with homoplasy.