Ortega-Del Vecchyo *et al. BMC Evolutionary Biology* (2017) 17:213
DOI 10.1186/s12862-017-1046-4

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Appropriate homoplasy metrics in linked SSRs to predict an underestimation of demographic expansion times

Diego Ortega-Del Vecchyo[1,2*], Daniel Piñero[1], Lev Jardón-Barbolla[1,4] and Joost van Heerwaarden[1,3*]

Abstract

Background: Homoplasy affects demographic inference estimates. This effect has been recognized and corrective methods have been developed. However, no studies so far have defined what homoplasy metrics best describe the effects on demographic inference, or have attempted to estimate such metrics in real data. Here we study how homoplasy in chloroplast microsatellites (*cpSSR*) affects inference of population expansion time. *cpSSRs* are popular markers for inferring historical demography in plants due to their high mutation rate and limited recombination.

Results: In *cpSSRs*, homoplasy is usually quantified as the probability that two markers or haplotypes that are identical by state are not identical by descent (Homoplasy index, *P*). Here we propose a new measure of multi-locus homoplasy in linked *SSR* called Distance Homoplasy (*DH*), which measures the proportion of pairwise differences not observed due to homoplasy, and we compare it to *P* and its per *cpSSR* locus average, which we call Mean Size Homoplasy (*MSH*). We use simulations and analytical derivations to show that, out of the three homoplasy metrics analyzed, *MSH* and *DH* are more correlated to changes in the population expansion time and to the underestimation of that demographic parameter using *cpSSR*. We perform simulations to show that Approximate Bayesian Computation (*ABC*) can be used to obtain reasonable estimates of *MSH* and *DH*. Finally, we use *ABC* to estimate the expansion time, *MSH* and *DH* from a chloroplast *SSR* dataset in *Pinus caribaea*. To our knowledge, this is the first time that homoplasy has been estimated in population genetic data.

Conclusions: We show that *MSH* and *DH* should be used to quantify how homoplasy affects estimates of population expansion time. We also demonstrate how *ABC* provides a methodology to estimate homoplasy in population genetic data.

Keywords: Homoplasy, Haplotypes, SSRs, Demography

Background

The study of historical demography is important for understanding the ecology and evolution of species. In particular, timing population size changes allows the discussion of past population patterns in the context of historical geological events such as island formation [1] and climate change [2]. One popular source of information to infer past population dynamics is the genealogical signal contained in linked polymorphic markers [3, 4], such as chloroplast microsatellites (*cpSSRs*). As highlighted in recent reviews [5, 6], *cpSSRs* are widely used in plant studies. *cpSSRs* remain popular despite the ascent of genome wide sequencing tools such as Restriction site-associated *DNA* sequencing (*RADseq*) [7], Genotyping-by-sequencing (*GBS*) [8] and targeted sequencing [9] due to two appealing properties: 1) their high mutation rate, ranging from 10^{-6} to 10^{-2} mutations per locus per generation [10], and 2) they can be applied in plant non-model species where few genomic resources have been developed [11].

High mutation rates combined with an approximately step-wise transition between allelic states make *cpSSRs* prone to homoplasious mutations. Homoplasy takes place in a *cpSSR* locus when different alleles at the locus are identical by state but are not identical by descent [12]. Two *cpSSRs* copies of a locus are defined to be

* Correspondence: vdortega@berkeley.edu; jvheerwaarden@wur.nl; joost.vanheerwaarden@wur.nl
[1]Departamento de Ecologia Evolutiva, Instituto de Ecologia, Universidad Nacional Autónoma de México, Mexico City, Mexico
Full list of author information is available at the end of the article

identical by state when they have the same size and are defined as identical by descent when there has not been a mutation since their divergence from a common ancestor. Previous studies have quantified the fraction of the homoplasy, called Molecularly accessible size homoplasy (*MASH*) [12–14] by measuring the differences in the *DNA* sequence of *SSRs* of identical size. Although that approach can reveal a fraction of the homoplasy in *SSRs*, it ignores the homoplastic events due to polymorphisms that lead to *DNA* sequences identical by state but not identical by descent. Therefore, *MASH* does not provide a direct estimate of homoplasy.

The occurrence of homoplasy is an important limitation of *cpSSR* based demographic inference in scenarios of population expansions, causing decreased ability to detect population growth [15] and to systematic underestimation of the expansion time [16]. Although some pseudo-likelihood and Bayesian methods of demographic inference [16, 17] successfully correct for homoplasy, they provide little insight into the relationship between homoplasy and the estimation of demographic parameters, nor do they provide estimates of homoplasy itself. In fact, to our knowledge, no formal analysis of the quantitative relation between homoplasy and the underestimation of the expansion time exists to date. Part of the reason is that the concept of homoplasy was developed to describe the proportion of haplotypes or markers that are identical by descent compared to those that are identical by state, while the problem of erroneous demographic inference is linked to an underestimation of the number of mutations between lineages. To illustrate this, the most common measure of homoplasy, the homoplasy index *P* [12], describes the probability that two *cpSSR* identical by state are not identical by descent. In the case of haplotypes composed of linked *cpSSR*, *P* has been defined as the probability that two haplotypes identical by state are not identical by descent and is dependent on the multi-locus heterozygosity [15]. This is the definition of *P* we will employ here. While simulation studies show that higher values of *P* are associated with an underestimation of the expansion time [15], other studies have found that multi-locus heterozygosity is not particularly sensitive to homoplasy [18], suggesting that *P* may not be the most appropriate measure for describing effects on demographic inference. This motivates the necessity to propose alternative measures of homoplasy that are more directly relevant to demographic inference and that would allow for meaningful quantifications of the effects of homoplasious mutations on the estimation of the expansion time.

In this paper, we propose a new homoplasy metric. We analyze the relationship between three homoplasy metrics, including our proposed metric, and the underestimation of the expansion time. Second, we evaluate the extent to which these homoplasy metrics can be estimated from simulated *cpSSR* data using Approximate Bayesian Computation (*ABC*). Finally, we quantify the level of homoplasy in a real dataset from *Pinus caribaea*, providing an empirical estimate of homoplasy from population genetic data.

Methods

Dataset simulations under a stepwise demographic expansion model

Throughout this study we assume a stepwise demographic expansion model [3]. The model consists on three parameters: $\theta_0 = 2LN_0u$, $\theta_1 = 2LN_1u$ and $\tau = 2Ltu$, where u is the mutation rate per generation at each linked *SSR*, N_0 and N_1 are the effective population sizes before and after the expansion, L is the number of linked *SSR* loci and t is the time in generations since the expansion.

We generated two sets of haplotypes, *hISM* and *hSMM*, in the coalescent simulations under the stepwise demographic expansion model used in this study. We used the same genealogy along with the set of mutations falling in each branch of the genealogy to generate *hISM* and *hSMM* from each coalescent simulation. *hISM* represents a set of linked multi-locus *SSR* haplotypes evolving under the infinite sites model, *ISM* [19] while *hSMM* are a set of linked multi-locus *SSR* haplotypes that evolved under the symmetrical stepwise mutation model, *SMM* [20]. The haplotypes *hISM* are free of homoplasy while the haplotypes *hSMM* can contain homoplasious mutations. These coalescent simulations were performed using a modified a version of the coalescent simulator *msHOT* [21, 22]. The modified version of *msHOT* is available at https://github.com/dortegadelv/HomoplasyMetrics.

Measures of homoplasy

We studied the relationship between homoplasy and the underestimation of the population expansion time τ using three different measures.

The first metric is the commonly used *homoplasy index* (*P*) [12] as used by [15]:

$$P = 1 - \frac{1 - H_{ISM}}{1 - H_{SMM}} = 1 - \frac{F_{ISM}}{F_{SMM}} \tag{1}$$

Where $H_{ISM}$and H_{SMM} are the expected heterozygosities [23] per haplotype estimated in a set of haplotypes containing L linked loci evolving under the infinite sites model (*hISM)* and the stepwise mutation model (*hSMM*), respectively. F_{ISM} and F_{SMM} are the expected homozygosities in the set of haplotypes *hISM* and *hSMM*. Note that F_{SMM} is directly observable from the data, while F_{ISM} is not, in real data of a set of haplotypes *hSMM*.

We also use the per *SSR* locus average of *P*, which we call Mean Size Homoplasy (*MSH*). It estimates the mean reduction in heterozygosity per *SSR* locus. This can be interpreted as the mean homoplasy index *P* per individual loci. It can be expressed as:

$$MSH = 1-\frac{\sum_{i=1}^{L}\frac{1-H^{i}_{ISM}}{1-H^{i}_{SMM}}}{L} = 1-\frac{\sum_{i=1}^{L}\frac{F^{i}_{ISM}}{F^{i}_{SMM}}}{L} \tag{2}$$

Where L is the number of *SSR* in the haplotype. H^{i}_{ISM} and H^{i}_{SMM} are the expected heterozygosities at the i locus in *hISM* and *hSMM*, respectively. F^{i}_{ISM} and F^{i}_{SMM} are the expected homozygosities at the i locus in *hISM* and *hSMM.*

Inference of demographic growth using haplotypes with linked microsatellites is typically based on the distribution of pairwise differences between multi-locus haplotypes, also known as the mismatch distribution, as the shape of this distribution is determined by the time and magnitude of historical population expansions [4]. Based on this, here we present a new metric, distance homoplasy (*DH*), which quantifies the proportion of mutations separating two multi-locus haplotypes that are not observed due to homoplasy. Our rationale for using this measure are studies that use the mode of the distribution of pairwise differences as the basis for estimating τ [4]. Therefore, underestimation of the proportion of pairwise differences should impact the mismatch distribution which in turn should alter the inference of τ. *DH* is expressed as:

$$DH = \frac{\pi_{ISM}-\pi_{SMM}}{\pi_{ISM}} \tag{3}$$

Where π_{SMM} and π_{ISM} are the mean number of differences between two haplotypes using the haplotypes *hSMM* and *hISM*, respectively.

Expected values of π, F^{i} and F in a stepwise demographic expansion model

We derived the expected values for the diversity statistics π, F^{i} and F as a function of the mutation rate u of each linked *SSR*, the number of linked simulated *SSR's* L and the coalescent time T_{ij}, in number of generations, between a pair of haplotypes i and j present in the sample. We use the following equation $E[\lambda] = E\left[E\left[\lambda|T_{ij}\right]\right] = \sum_{x=1}^{t} E\left[\lambda|T_{ij}=x\right]P\left(T_{ij}=x\right)$ where λ stands for any diversity statistic and t is the time in generations since the expansion. T_{ij} is scaled in units of N generations. We explain how to obtain the values of $E[\lambda|\ T_{ij}]$ for every diversity statistic under the *ISM* and *SMM* in the Appendix. The probability distribution of T_{ij} under a stepwise demographic expansion model is equal to:

$$P\left(T_{ij}=x\right) = \begin{cases} {}^{1}\!\big/_{N}\left(e^{-x}\big/_{N}\right) & 0\leq x < t-1 \\ 1-\sum_{x=1}^{t-1} P\left(T_{ij}=x\right) & x=t \end{cases} \tag{4}$$

Where N is the effective population time in the present. The probability distribution of T_{ij} is divided into two phases: 1) After the expansion, the population keeps a constant population size and, therefore, $P\left(T_{ij}=x\right) = {}^{1}\!\big/_{N} e^{-x}\!\big/_{N}$ during that period of time. 2) Before the expansion, all individuals must coalesce quickly at a time very close to the expansion time $T_{ij} = t$ assuming that the population size is very small. To model that effect, we assume that all individuals coalesce exactly at time $T_{ij} = t$ if they have not already coalesced going forward in time.

The equations shown above to estimate the expected value of the diversity statistics are used to obtain estimates of the homoplasy parameters *P, MSH* and *DH*. As an example, following equation (1) the expected value of *P* can be calculated if we know the expected value of the diversity statistics F_{ISM} and F_{SMM}.

$$E[P] = 1-\frac{E[F_{ISM}]}{E[F_{SMM}]} \tag{5}$$

Where:

$$E[F_{SMM}] = \sum_{x=1}^{t} E\left[F_{SMM}|T_{ij}=x\right]P\left(T_{ij}=x\right) \tag{6}$$

$$E[F_{ISM}] = \sum_{x=1}^{t} E\left[F_{ISM}|T_{ij}=x\right]P\left(T_{ij}=x\right) \tag{7}$$

The same approach was also used to obtain the expected values of *MSH* and *DH.* The analytical equations to calculate those expected values are explained in the Appendix.

Simulations to analyze changes in homoplasy measures

We used a simulation framework to test the accuracy of our estimates for the summary statistics π, F and F^{i} under the *ISM* and the *SMM* along with our estimates for the homoplasy values *P, MSH* and *DH*. We used our modified version of the coalescent simulator *msHOT* [21, 22] to generate two different sets of haplotypes (*hSMM* and *hISM*) for each simulated genealogy. The modified version of *msHOT* is available at https://github.com/dortegadelv/HomoplasyMetrics. *msHOT* was used to make simulations under the stepwise demographic expansion model, where we used a value of $\theta_1 = 30$, $\theta_0 = 0.03$ and ten different values of τ {1.5, 3, 4.5, 6, 7.5, 9, 10.5, 12, 13.5, 15}. For each value of τ, we performed 100 simulations of 150 haplotypes with 6 linked *SSRs*. The ten command lines used for those simulations are shown in the Appendix (Command Line 1).

We also did 100 simulations for 9 different numbers of linked *SSRs* in the haplotype, going from $L = 2$ to $L = 10$ to examine how changes in the value of L affect P, *MSH* and *DH*. We simulated 150 haplotypes in each simulation, where the values of the demographic parameters were set to $\theta_1 = 5L$, $\theta_0 = 0.005L$, $\tau = L = 2tuL$, where we kept the parameters t and u fixed to a certain value such that $2tu = 1$. Notice that the divergence time t is kept fixed regardless of the number of linked *SSRs* L in these simulations. The nine command lines used for these simulations are shown in the Appendix (Command Line 2).

Underestimation of expansion time

We quantified the underestimation of expansion time due to homoplasy using a metric called *TS*, which we define as

$$TS = \frac{\widehat{\tau_{ISM}} - \widehat{\tau_{SMM}}}{\widehat{\tau_{ISM}}} \tag{8}$$

Values of the estimated expansion time τ for haplotypes *hISM* and *hSMM*, $\widehat{\tau_{ISM}}$ and $\widehat{\tau_{SMM}}$ respectively, were obtained using the method by Schneider and Excoffier (1999), implemented in the software *Arlequin* [24]. This method infers the parameters θ_0, θ_1 and τ based on the observed distribution of pairwise differences between haplotypes, also called mismatch distribution, and its expectation under a stepwise demographic expansion model [25]. This approach assumes that there is no homoplasy in the sample of haplotypes, therefore any differences between $\widehat{\tau_{ISM}}$ and $\widehat{\tau_{SMM}}$ are due to homoplasious mutations present in *hSMM*. Following [15], to use *Arlequin* for the *hSMM* analysis we coded the *SSRs* as binary data, where the number of repeats were coded with '1' and shorter alleles were coded filling the difference in repeats with '0' .

We simulated 100 replicates of 150 haplotypes with 6 linked *SSRs* for each of 10 different values of τ {1.5, 3, 4.5, 6, 7.5, 9, 10.5, 12, 13.5, 15}. We set a value of θ_1 equal to 30 and 60, which has the same order of magnitude of the value of θ_1 estimated for the *Pinus caribaea* dataset [26] employed in this study *(see Pinus caribaea* dataset*)*, and a value of θ_0 which was 1000 smaller than θ_1 for all simulations. The command lines used for the simulations are shown in the Appendix. Command Line 3 and 4 were used for the simulations done where $\theta_1 = 30$ and $\theta_1 = 60$, respectively.

The value of each homoplasy measure and *TS* was computed for each replicate of each simulation and the relationship of each homoplasy measure with *TS* was analyzed. In the simulations done with a value of $\theta_1 = 30$, we removed 1 out of the 1000 simulations we performed where that simulation was the only that had a *TS* value smaller than -10 (see Additional file 1: Figure S1 for details on the removed simulation).

Estimation of homoplasy and expansion time using ABC

We used another modified version of the program *msHOT* [21, 22], also available at https://github.com/dortegadelv/HomoplasyMetrics, to implement an *ABC* algorithm that estimates the posterior distribution of demographic parameters θ_0, θ_1 and τ and the posterior predictive distribution of the three measures of homoplasy *DH*, *MSH* and *P* (see Appendix for details about the implementation of the *ABC* algorithm). We employed three summary statistics previously used [27] to estimate demographic parameters in a model of population growth: The mean of the variance in the size of the *SSRs* across loci (V), the expected heterozygosity averaged across loci (H) and the number of distinct haplotypes (a). We used the mode of the posterior distribution and posterior predictive distribution as point estimates of the demographic parameters and homoplasy measures, respectively (see Appendix and Additional file 1: Figure S2 for a discussion on why we employed the mode as a point estimate). We also quantified the relative bias and estimated the 50%, 75% and 90% coverage of the demographic parameters and homoplasy measures to ascertain the quality of the point estimates and the inferred posterior distributions (see Appendix). The relative bias is the average difference between the estimated and true value of the parameter divided by its true value [28] . The 50%, 75% and 90% coverage are the proportion of times that the true value is within the 50%, 75% and 90% credible interval.

We compared the real value of the homoplasy measures P, *MSH* and *DH* against the estimated values of the homoplasy measures using our *ABC* approach in 100 simulations of 150 haplotypes with 6 linked *SSRs* with parameters $\theta_1 = 30$ and $\theta_0 = 0.03$ and 10 different values of τ {1.5, 3, 4.5, 6, 7.5, 9, 10.5, 12, 13.5, 15}, where 10 simulations were done for each τ value. The ten command lines used for the simulations are shown in the Appendix (Command Line 5). We also estimated the three homoplasy measures in the simulations we explain in the next paragraph.

We compared the performance of three different methods to infer τ and θ_1 in three different sets of 100 simulations of 150 haplotypes with 6 linked *SSRs* done using three different τ values {3,6,9}, a $\theta_1 = 30$ and a $\theta_0 = 0.03$. The three command lines for these simulations are shown in the Appendix (Command Line 6). One of the three methods we used to estimate τ and θ_1 is our *ABC* approach, and the other two methods use the mismatch distribution to estimate those demographic parameters: 1) One of those methods is the approach taken by [3] as implemented in the software *Arlequin* [24], which assumes that there is no homoplasy in the data (Least Squares approach without taking Homoplasy into account, *LSWH*). 2) The other method we used is a maximum-pseudolikelihood estimator that uses a model where it is assumed that homoplasy can occur in the data [16] (Maximum Pseudolikelihood using a model with Homoplasy, *MPH*). Code for that method was kindly provided by Miguel Navascués.

Estimation of homoplasy and population expansion times in a *Pinus caribaea* dataset

We used a dataset of 7 *SSR* loci from 88 individuals of the species *Pinus caribaea* to estimate τ along with the homoplasy measures *MSH* and *DH*. This dataset is a subset of the data previously published in [26], where an analysis of population structure from four species of *Pinus* subsection *Australes*, including *Pinus caribaea*, identified four different groups (groups I-IV). We took the group containing the largest number of individuals distributed in Central America (group II), and retained only the individuals sampled from Central America in that group (88 out of 93 individuals) for our analysis. A hypothesis of population expansion could not be rejected using information from the mismatch distribution in group II [26], making this group suitable for analysis of expansion. We used *ABC*, *LSWH* and *MPH* to estimate τ in that dataset. The estimations of $\hat{\tau}$ in the three methods used above were later transformed to years using a mutation rate of $5.5\ X\ 10^{-5}$ per *SSR* per generation [29] and a generation time of 42.5 years [26].

We also report the 95% confidence interval of the estimation of $\hat{\tau}$ with *LSWH*, *MPH* and *ABC* using a parametric bootstrap approach as in *Arlequin* [24]. We report the 95% confidence intervals for the *ABC* method instead of the 95% credible intervals to compare the 95% confidence intervals created with *ABC* with those obtained using *LSWH* and *MPH*. For each particular inference method (*LSWH*, *MPH* or *ABC),* the approach involves the simulation of 1000 datasets of 88 individuals with 7 *SSR* using the demographic parameters estimated for the *Pinus caribaea* data using a particular inference method. The value of the parameter $\tau^{\cdot}$ from each of the 1000 datasets was estimated using the inference method under study. Then, for a confidence level of α = 0.05, the approximate limits of the confidence interval were defined as the $\alpha/2$ and $1 - \alpha/2$ percentile values of the 1000 values of $\tau^{\cdot}$. This parametric bootstrap approach was also used to estimate the 95% confidence interval of *MSH* and *DH* using the *ABC* method and 1000 simulations done using the demographic parameters inferred by *ABC*.

Results

Response of different measures of homoplasy to differences in expansion time

First we evaluated the response in the three different measures of homoplasy and their components, π, F^i and F, to changes in expansion time under the demographic stepwise expansion model in haplotypes containing completely linked *SSRs*. We thereby corroborate our theoretical expectations for the different metrics and found that our simulations validate their predictions (Fig. 1a-d).

As can be seen in Fig. 1a-b, the accumulation of homoplasious mutations causes a monotonic increase in the difference between F^i_{ISM} and F^i_{SMM} and between π_{SMM} and π_{ISM} with the expansion time, something that is not observed for the difference between the two measures of haplotype homozygosity F_{ISM} and F_{SMM} (Fig. 1c). This translates to both *MSH* and *DH* increasing steadily with expansion time, while *P* has a parabolic relationship (Fig. 1d) and stays at a constant value close to 0.09 when τ is equal or larger than 8 (Fig. 1d). Therefore, the values of *P* do not seem likely to relate to underestimation of population expansion time, in contrast with *MSH* and *DH.* Additionally, we found that the number of linked *SSRs* in the haplotype does not influence the values of *MSH* and *DH*, but it does change the value of *P* given a fixed divergence time *t* (Additional file 1: Figure S3).

The relation between different measures of homoplasy and underestimation of τ

Simulations of demographic expansion under different values of τ reveal that the standard homoplasy index *P* is not strongly correlated with *TS,* which measures the underestimation of τ due to homoplasy (Fig. 2a, Pearson's ρ = −0.1282, *p*-value = $4.8{*}10^{-5}$). Contrarily, *MSH* and *DH,*have a stronger correlation with an underestimation of τ, where *MSH* has a slightly lower correlation with *TS* (Fig. 2b, ρ = 0.6903, *p*-value $<2.2{*}10^{-16}$) than *DH*, which has the strongest correlation with *TS* of all the homoplasy measures inspected (Fig. 2c, ρ = 0.6989, *p*-value $<2.2{*}10^{-16}$). Correlation between the latter two measures was strong (0.9208), whereas neither of them was strongly correlated with *P* (ρ between *MSH* and *P* = −0.2685; ρ between *DH* and *P* = −0.0977). Simulations with a higher value of $\theta_1 = 60$ (Additional file 1: Figure S4), produced a nearly identical relationship between *DH*, *MSH* and *TS* (ρ between P and *TS* = −0.2617; ρ between *MSH* and *TS* = 0.8673; ρ between *DH* and *TS* = 0.8777), showing that *DH* and *MSH* are robust predictors of an underestimation of τ while *P* is not.

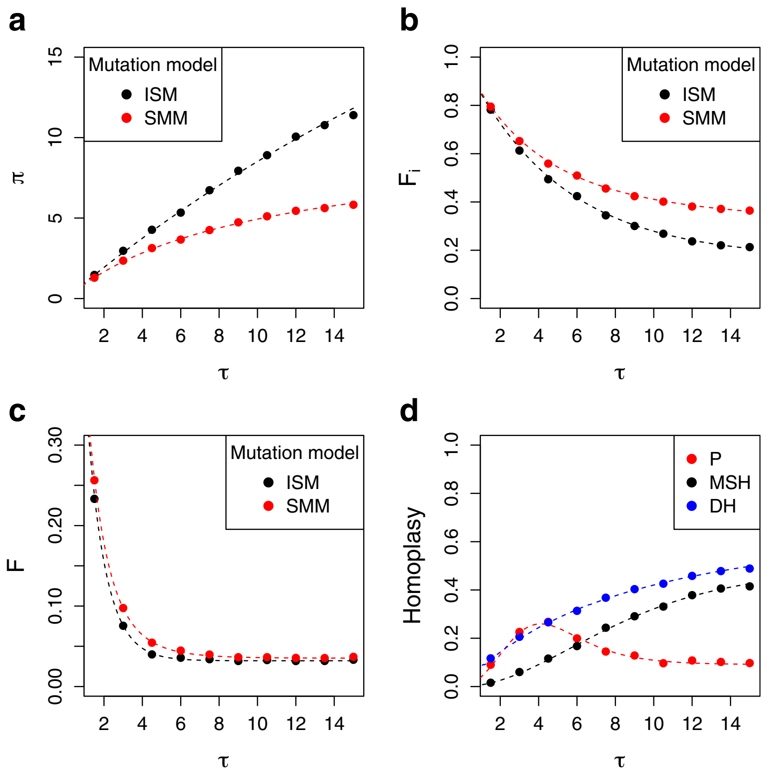

Fig. 1 Homoplasy values in the stepwise demographic expansion model for different values of the expansion time parameter *τ*. The points in each plot are the average values for each statistic across 100 simulations for plots (**a**-**c**), those average values were used to calculate the mean values of the homoplasy index (*P*), mean size homoplasy (*MSH*) and distance homoplasy (*DH*) that are plotted as points in (**d**). The dashed lines are the approximated expected values estimated from our derivations. **a** π_{ISM} and π_{SMM}; **b** F^i_{ISM} and F^i_{SMM}; **c** F_{ISM} and F_{SMM} **d** *P*, *MSH* and *DH*

ABC estimates of homoplasy and expansion time

We used simulated data to evaluate the estimation of homoplasy metrics on an *ABC* framework. We performed linear regressions of the estimated values of the homoplasy measures obtained by our *ABC* approach on their true homoplasy values (Fig. 3) We also measured the relative bias and the correlation between the estimated and true values of the homoplasy measures. On simulations done over a range of τ values, we found that our estimates of *MSH* and *DH* were highly correlated with their real values (r = 0.881 and r = 0.740, respectively) and their relative bias was small (relative bias = −0.040 and 0.030, respectively), indicating that *MSH* and *DH* values are well estimated by our *ABC* approach. On the other hand, the estimates of *P* had a smaller correlation with their true values (r = 0.486) and their relative bias is −0.132, indicating that our *ABC* approach underestimates *P* values by approximately 13.2%. The underestimation can also be seen in Fig. 3. Despite differences in the quality of the point estimates of *P*, *MSH* and *DH*, we found that the 50%, 75% and 95% coverage of the homoplasy measures indicate that the inferred posterior distribution of those measures are well estimated (Additional file 1: Table S1 and Appendix).

We performed more simulations to analyze the performance of the *ABC* approach on simulations done using the same demographic parameters. We evaluated this by creating three sets of simulations done over a single value of τ (τ = 3, 6 or 9). We found that the 50%, 75% and 90% coverage indicate that the posterior distributions of the homoplasy measures are correctly inferred (Additional file 1: Table S2 and Appendix). Second, we found that the average relative bias was small for all homoplasy measures (relative bias = 0.053, 0.043 and 0.068 for *P*, *MSH* and *DH*, respectively; Additional file 1: Table S3). This indicates that, on average, the *ABC* method slightly overestimates the value of the homoplasy measures by approximately 5% on these sets of simulations.

Apart from estimating homoplasy measures, we also used *ABC* to estimate the value of τ. We found that *ABC* and the pseudo-likelihood estimator (*MPH*) perform equally well to obtain estimates of the value of τ, showing that both methods can correct for the effects of homoplasy. As expected, the expansion time is strongly underestimated by the method that does not take homoplasy into account (*LSWH*), particularly for older expansion events where there are higher values of *MSH* and

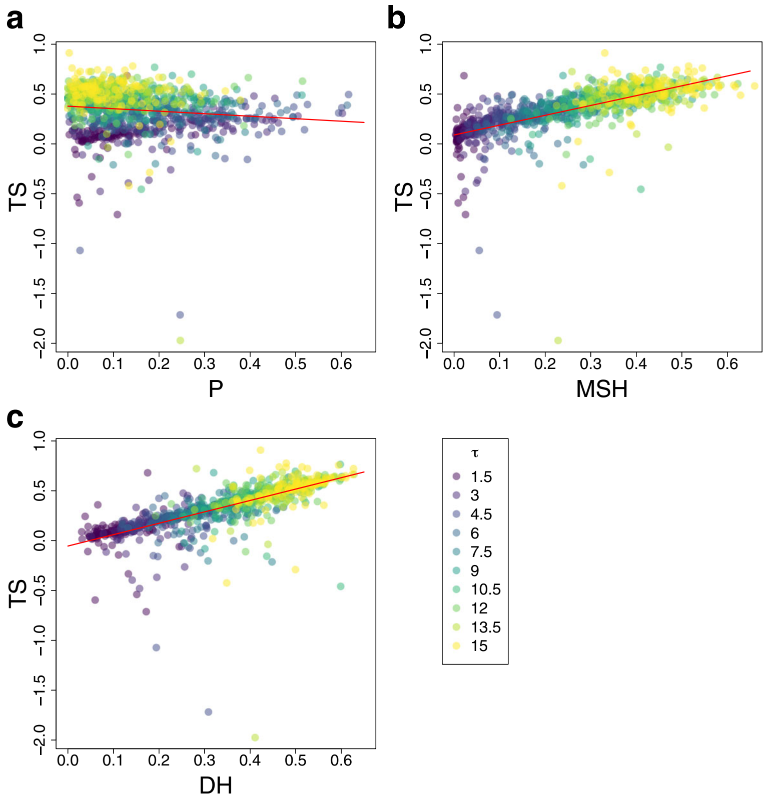

Fig. 2 Linear relationship between *TS* and three measures of homoplasy: **a** *P* (ρ = −0.1282, intercept = 0.3783, slope = −0.2509, *p*-value = 4.83e^{-5}), **b** *MSH* (ρ =0.6903, intercept = 0.0893, slope = 0.9868, *p*-value <2.2e^{-16}) and **c** *DH* (ρ =0.6989, intercept = −0.0541, slope = 1.1424, *p*-value <2.2e^{-16}) in 999 simulations made with the demographic parameters θ_0 = 0.03, θ_1 = 30 and 10 different values of τ

DH (Fig. 4). Additionally, we found that *ABC* gave good estimations of the value of θ_1, compared to *LSWH* and *MPH* which gave overestimations of the actual value of θ_1 (Additional file 1: Figure S5), in line with previous studies done using *LSWH* [3] and *MPH* [16]. It must be pointed out that in *ABC*, as in any Bayesian method, the estimates of the parameters depend on the prior distributions used for the parameters. Prior distributions should contain all the possible demographic parameter values [30] and should not be very wide to avoid low acceptance rates in the *ABC* algorithm (step 5 of the *ABC* algorithm in the Appendix).

Population expansions and homoplasy in data of *Pinus caribaea* populations from central America

The time of expansion for one population of *Pinus caribaea* in Central America was obtained using *LSWH*, *MPH* and *ABC* (Table 1). We found that the only method where homoplasy is not taken into account, *LSWH*, produces lower estimates of τ compared to *ABC* and *MPH*, suggesting that homoplasy may cause the expansion time to be underestimated by approximately 100,000 years in this case. The 95% confidence intervals of τ for all methods is large however (Table 1). *ABC*-based estimates of homoplasy are 0.11 and 0.246 for *MSH* and *DH* respectively, which agrees with the theoretical estimates of 0.106 and 0.269 obtained for those homoplasy measures using equations (24, Appendix) and (17, Appendix) given the demographic parameters estimated using *ABC.*

Discussion and conclusions

Here we propose a homoplasy metric, *DH*, which measures the proportion of pairwise differences that are not observed due to homoplasy. Our theoretical estimates and simulations confirm that the mean number of pairwise differences not counted due to homoplasy increase when population expansion times are older, causing a monotonic increase in the value of *DH* (Fig. 1). We also confirm that *DH* has a strong, linear relationship with underestimation of population expansion times using classical methods of inference based on pairwise differences (Fig. 2), with older expansion times leading to the expected higher *TS* [16] and corresponding increase in *DH*. This in contrast to the standard homoplasy index for multi-locus haplotypes, *P*, which shows no clear relation to *TS* and, starting from a certain τ value, actually decreases as a function of expansion time. The latter can

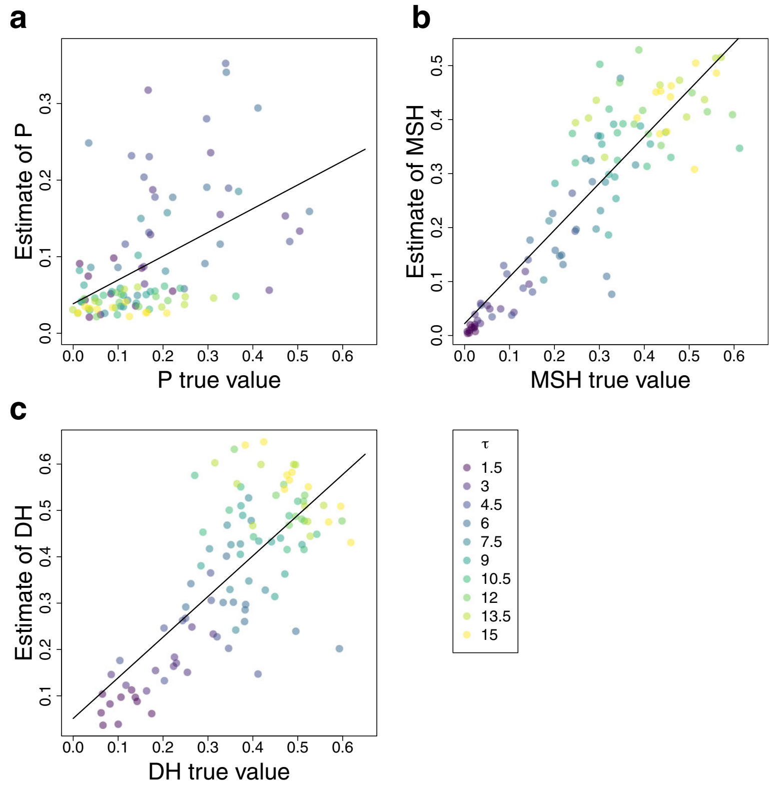

Fig. 3 *ABC* estimates of **a** *P*, **b** *MSH* and **c** *DH* compared with their true values in 100 simulations done with the demographic parameters θ_0 = 0.03, θ_1 = 30 and 10 different values of τ. A linear model was fitted to analyze the relationship between each homoplasy measure true value and their *ABC* estimate of *P* (ρ = 0.4864, intercept = 0.0385, slope = 0.3104, *p*-value = 2.88e^{-7}), *MSH* (ρ = 0.8809, intercept = 0.0220, slope = 0.8667, *p*-value <2.2e^{-16}) and *DH* (ρ = 0.7399, intercept = 0.0512, slope = 0.8771, *p*-value <2.2e^{-16})

be clearly understood from the expected relation between expansion time and homozygosity under the *ISM* and *SSM* model. This shows the value of a homoplasy metric that directly captures the underestimation of the number of mutations [18] when trying to capture effects on demographic inference.

Although conceptually and empirically *DH* most closely relates to the way that homoplasy causes underestimation of expansion time, the average decrease of per-locus heterozygosity, *MSH*, has a rather similar relation to population expansion time (Fig. 1) and also correlates strongly with the underestimation of the population expansion time (Fig. 2). It has been shown that in constant population sizes the value of *MSH* is determined by θ [12], the expected number of mutations between a pair of sequences, so the fact that it also increases with τ is not entirely surprising. Our theoretical estimates indeed confirm that *MSH* increases monotonically with expansion time under the stepwise demographic expansion model (Fig. 1) and also show that

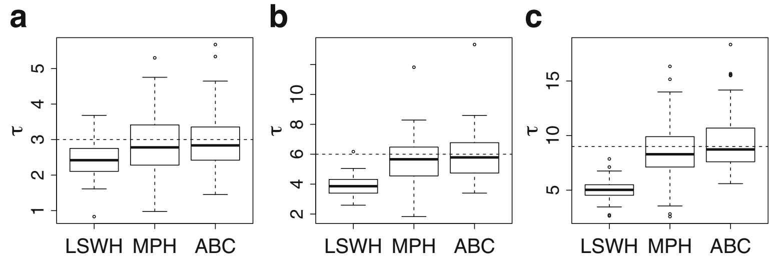

Fig. 4 Estimation of τ using three methods (*LSWH*, *MPH* and *ABC*). The boxplots of the estimation of τ were done on simulations where θ_1 = 30, θ_0 = 0.03 and three different values of τ were used, **a** τ = 3, **b** τ = 6 and **c** τ = 9. 100 simulations were performed for each value of τ. The actual value of τ in each plot is displayed with the dashed line

Table 1 Estimates of homoplasy and times of expansion in the population of *Pinus caribaea* analyzed

$\widehat{\tau_{LSWH}}$	$\widehat{\tau_{MPH}}$	$\widehat{\tau_{ABC}}$	MSH	DH	Theoretical estimate of MSH	Theoretical estimate of DH
4.074 (1.359 – 6.086)	5.994 (1.732 – 10.297)	5.591 (2.645 – 11.005)	0.115 (0.026 – 0.244)	0.246 (0.106 – 0.415)	0.106	0.269

Time of expansion in years (*LSWH*)	Time of expansion in years (*MPH*)	Time of expansion in years (*ABC*)
224,900 (75,000 – 335,900)	330,800 (95,600 – 568,400)	308,600 (146,000 – 607,400)

The estimated values of the time of expansion were obtained using three different methods (*LSWH*, *MPH* and *ABC*). The estimated values of *MSH* and *DH* were obtained using *ABC*. The numbers inside the parentheses denote the upper and lower limits of the 95% confidence interval for the parameter or homoplasy measure. The theoretical estimates of *MSH* and *DH* were estimated using the values of τ and θ_1 obtained using *ABC* and the Eqs. (24) and (17)

there is a relationship between the number of mutations, predicted by older coalescent times due to older population expansions, and *MSH* on the stepwise population expansion model. Additionally, *MSH* and *DH* are not affected by changes in the number of linked *SSRs* analyzed, while *P* does depend on the number of linked *SSRs* studied.

The usefulness of homoplasy metrics such as *DH* and *MSH* depends in part on how well they can be estimated from data. We have shown that we can obtain reasonable average estimates of *MSH* and *DH* using *ABC* [31]. Compared to *MASH*, the *ABC* method we propose is not biased by the fraction of homoplasy unmeasurable by *MASH*. It thereby offers a natural solution to the quantification of homoplasy. Additionally, *ABC* can estimate the posterior distribution of demographic parameters of interest through explicit modeling of *SSR* evolution under different demographic scenarios. The approach proved successful in correcting for bias in the inference of expansion time, with similar performance to *MPH* [16] which also explicitly accounts for homoplasy assuming the *SSR's* evolve according to a *SMM*. One advantage of *ABC*, in addition to allowing for direct estimates of homoplasy, is that more complicated mutational models of *SSR* evolution can easily be incorporated. This could be important, as many *SSR* are known not to evolve in a strictly stepwise manner [32].

Given the potential for erroneous demographic inference when using linked SSR, it is important to obtain such homoplasy estimates from empirical data. With our *ABC* approach, we were able to estimate values of *MSH* and *DH* in a published dataset of *Pinus caribaea*. We found that the underestimation of the expansion time assuming a model that does not take homoplasy into account is of around 80,000 to 100,000 years, a reduction of around 28 to 32% compared to the value estimated with methods that use a more realistic model of *SSR* evolution where homoplasious events are possible. As with the *ABC* approach proposed here, other authors have suggested to use model-based approaches to infer past demographic events using linked *cpSSR* markers in spruces [11]. Since *ABC* simulation based approaches provide an estimate of homoplasy, we believe that *ABC* approaches are useful to quantify the effect of homoplasy on demographic parameters and summary statistics of interest. To our knowledge, this is the first time that homoplasy parameters have been inferred using population genetic data.

Additional file

Additional file 1: Figures S1-S5 and **Tables S1-S3**. (DOCX 3133 kb)

Appendix

Expected values of the parameters π, F^i and F given a certain coalescent time

Given a certain coalescent time T_{ij} between lineages i and j, it is possible to estimate the expected values of a diversity statistic λ, such as the mean number of differences between two haplotypes (π), the expected homozygosities at one i SSR locus (F^i) and the expected homozygosity in a linked *SSR* multilocus haplotype (F) under the *ISM* and the *SMM*. If we can estimate that quantity, we can use this equation:

$$E[\lambda] = \sum\nolimits_{x=1}^{t} E\left[\lambda | T_{ij} = x\right] P\left(T_{ij} = x\right) \tag{9}$$

to obtain the expected value of those diversity statistics given different parameters of a stepwise demographic expansion model. Using those expected values, we can calculate estimates of the homoplasy measures *DH, MSH* and *P.*

We will start by defining the values of the summary statistics π_{ISM} and π_{SMM} given a particular coalescent time. π_{ISM} and π_{SMM} define the value of π in a set of linked multilocus *SSR* haplotypes *hISM* evolving under the infinite sites model *ISM*, and in a set of linked multilocus *SSR* haplotypes *hSMM* that evolved under the symmetrical stepwise mutation model *SMM*, respectively. The expected value of π_{ISM} for L loci given a certain coalescent time T_{ij} and a mutation rate per generation in one loci equal to u is:

$$E[\pi_{ISM}|\ T_{ij}\ ] = \Upsilon = 2LuT_{ij} \tag{10}$$

To obtain the expected value of π_{SMM}, we used the random walk model formulas from [33] to estimate the expected number of differences between a pair of *SSR's* given that x mutations have taken place in the two lineages since their divergence from a common ancestor:

$$\textit{If x is odd: } E[\pi_{SMM}\ |\#mutations = x] = \frac{1}{2^{x-1}}\frac{x+1}{2}\begin{pmatrix} x \\ \frac{x+1}{2} \end{pmatrix} \tag{11}$$

$$\textit{If x is even: } E[\pi_{SMM}\ |\#mutations = x] = \frac{1}{2^{x-2}}\frac{x}{2}\begin{pmatrix} x-1 \\ \frac{x}{2} \end{pmatrix} \tag{12}$$

The number of mutations x that happened between each pair of *SSRs* is distributed as a Poisson random variable with mean $\Upsilon = 2LuT_{ij}$. Therefore, the probability of having an x number of mutations is:

$$P[\#mutations = x] = \frac{\left(\frac{\Upsilon}{L}\right)^{x} e^{-\frac{\Upsilon}{L}}}{x!} \tag{13}$$

Where L is the number of loci. Using those facts, we can obtain the expected number of differences π_{SMM} using the following formula:

$$E\left[\pi_{SMM}|\ T_{ij}\right] = L\sum\nolimits_{x=0}^{\infty} E[\pi_{SMM}\ |\#mutations = x]\ P[\#mutations = x] \tag{14}$$

To be practical, the past sum was carried out until $\sum_{x=0}^{r} P[\#mutations = x]$ was bigger or equal to 0.999 for the smallest possible value of r.

Following (9), (10), and (14), we can obtain the expected values of π_{SMM} and π_{ISM} using:

$$E[\pi_{ISM}] = \sum\nolimits_{x=1}^{t} E\left[\pi_{ISM}|T_{ij} = x\right]P\left(T_{ij} = x\right) \tag{15}$$

$$E[\pi_{SMM}] = \sum\nolimits_{x=1}^{t} E\left[\pi_{SMM}|T_{ij} = x\right]P\left(T_{ij} = x\right) \tag{16}$$

The values of $E[\pi_{SMM}]$ and $E[\pi_{ISM}]$ can be combined to obtain the expected value for the homoplasy parameter *DH*:

$$E[DH] = \frac{E[\pi_{ISM}] - E[\pi_{SMM}]}{E[\pi_{ISM}]} \tag{17}$$

We can also obtain the expected value of *MSH* by estimating the expected values of the homozygosity parameters for each of the i *SSR* loci under the infinite sites model $F^{i}_{ISM} = 1 - H^{i}_{ISM}$ and the stepwise mutation model $F^{i}_{SMM} = 1 - H^{i}_{SMM}$. If we have L SSR's:

$$E\left[F^{i}_{ISM}|\ T_{ij}\right] = (1-u)^{2T_{ij}} = e^{-2T_{ij}u} = e^{-\Upsilon/L} \tag{18}$$

To obtain the value of F^{i}_{SMM} we followed [34] derivations of the expected values of heterozygosity. They defined the probability of having an equal number x of mutations that increase or decrease the number of repeats in a *SSR* given that 2x mutations have occurred:

$$P[\ x\ \textit{mutations that increase repeat number}\ |\#mutations = 2x] = \begin{pmatrix} 2x \\ x \end{pmatrix}\left(\frac{1}{2}\right)^{2x} \tag{19}$$

And given that the probability of having 2x mutations is:

$$P[\#mutations = 2x] = \frac{\left(\frac{\Upsilon}{L}\right)^{2x} e^{-\frac{\Upsilon}{L}}}{(2x)!} \tag{20}$$

Then, summing over all possible values of x, we obtain the value of F_{SMM}:

$$E\left[F^{i}_{SMM}|\ T_{ij}\right] = \sum\nolimits_{x=0}^{\infty} P[\ x\ \textit{mutations that increase repeat number}\ |\#mutations = 2x]P[mutations = 2x] \tag{21}$$

This sum was also done until $\sum_{x=0}^{r} P[\#mutations = x]$ was bigger or equal to 0.999 for the smallest possible value of r. Using (9) we get:

$$E\left[F^{i}_{SMM}\right] = \sum\nolimits_{x=1}^{t} E\left[F^{i}_{SMM}|T_{ij} = x\right]P\left(T_{ij} = x\right) \tag{22}$$

$$E\left[F^{i}_{ISM}\right] = \sum\nolimits_{x=1}^{t} E\left[F^{i}_{ISM}|T_{ij} = x\right]P\left(T_{ij} = x\right) \tag{23}$$

Combining those equations, an analytical equation for the expected value of *MSH* can be obtained:

$$E[MSH] = 1 - \frac{\sum_{i=1}^{L}\frac{E\left[F^{i}_{ISM}\right]}{E\left[F^{i}_{SMM}\right]}}{L} = 1 - \frac{E\left[F^{i}_{ISM}\right]}{E\left[F^{i}_{SMM}\right]} \tag{24}$$

We can also derive a theoretical estimate of homozygosity in the haplotype composed of the L *SSR* loci under the *ISM* and *SMM*. Under the *ISM*, the expected homozygosity value is equal to:

$$E\left[F_{ISM}|\ T_{ij}\right] = (1-Lu)^{2\ T_{ij}} = e^{-2L\ T_{ij}u} = e^{-\Upsilon} \tag{25}$$

To calculate the expected homozygosity under the *SMM* in a haplotype containing a set of linked *SSR's*, we use the following formula:

$$E\left[F_{SMM}\middle|\ T_{ij}\right] = \sum\nolimits_{x=0}^{\infty} P[\textit{two haplotypes are identical by state} \mid \textit{mutations} = x]P[\#\textit{mutations} = x] \tag{26}$$

To calculate the probability that two haplotypes are identical by state given that x mutations happened before they coalesce to a common ancestor it is necessary to: 1) Count all the possible ways in which *x* mutations could happen to make the two haplotypes identical by state and 2) Divide that number between all the possible ways in which *x* mutations could be distributed along the haplotypes. Many of the possible distributions of mutations that could make two haplotypes identical by state have a very low probability of occurring, where most of those cases involve having a very high number of mutations, therefore ignoring those cases does not alter much the value of $E[F_{SMM}]$ calculated. We found that the following formula provides a good approximation to the value of F_{SMM}:

$$\begin{aligned} E\left[F_{SMM}\middle|\ T_{ij}\right] \approx {} & P[\#\textit{mutations} = 0] \\ & + L\sum\nolimits_{x=1}^{L} P[2x \textit{ mutations landed on the same microsatellite and two identical by state haplotypes were produced}] \\ & + \sum\nolimits_{x=2}^{L} \binom{L}{x} P[2 \textit{ mutations landed on x different microsatellites and two identical by state haplotypes were produced}] \end{aligned}$$

The past formula can also be expressed as:

$$\begin{aligned} E\left[F_{SMM}\middle|\ T_{ij}\right] \approx {} & P[\#\textit{mutations} = 0] \\ & + L\sum\nolimits_{x=1}^{L} P[2x \textit{ mutations landed on the same microsatellite} \mid \#\textit{mutations} = 2x]P[\textit{mutations} = 2x] \\ & P[x \textit{ mutations that increase repeat number} \mid \#\textit{mutations} = 2x] + \sum\nolimits_{x=2}^{L} \binom{L}{x} P[2 \textit{ mutations landed on x different microsatellites} \mid \#\textit{mutations} = \\ & 2x]P[\#\textit{mutations} = 2x] \prod_{i=1}^{x} P[1 \textit{ mutation that increase repeat number} \mid \#\textit{mutations} = 2] \end{aligned} \tag{27}$$

Where $P[\#\textit{mutations} = 2x]$ is given by (20), substituting ϒ/L by ϒ $P[x \textit{ mutations that increase repeat number} \mid \# \textit{mutations} = 2x]$ is given by (19) and:

$$P[2x \textit{ mutations landed on the same microsatellite} \mid \#\textit{mutations} = 2x] = \left(\frac{1}{L}\right)^{2x} \tag{28}$$

And, using the multinomial distribution, we can get:

$$P[2 \textit{ mutations landed on x different microsatellites} \mid \#\textit{mutations} = 2x] = \frac{2x!}{\prod_{i=1}^{x} 2!}\left(\frac{1}{L}\right)^{2x} \tag{29}$$

Then, using (9), we get:

$$E[F_{SMM}] = \sum\nolimits_{x=1}^{t} E\left[F_{SMM}\middle|T_{ij} = x\right]P\left(T_{ij} = x\right) \tag{30}$$

$$E[F_{ISM}] = \sum\nolimits_{x=1}^{t} E\left[F_{ISM}\middle|T_{ij} = x\right]P\left(T_{ij} = x\right) \tag{31}$$

Using those results, we can compute an approximate expectation for the value of *P*, which estimates the haplotypic reduction in heterozygosity due to homoplasy, as:

$$E[P] = 1 - \frac{E[F_{ISM}]}{E[F_{SMM}]} \tag{32}$$

ABC algorithm

The input of the algorithm is the value of three summary statistics *S* calculated from a set of linked haplotypes with *L* SSR loci. Those three statistics *S* are *V*, *H* and *a*. We define *V* as the mean of the variance in the size of the *SSRs* across loci, *H* is the expected heterozygosity averaged across loci and *a* is the number of distinct haplotypes. The output of this *ABC* algorithm is a sample of the posterior distribution of the demographic parameters θ_0, θ_1, τ and a sample of the posterior predictive distribution of the homoplasy measures *P*, *MSH* and *DH* [35]. The *ABC* algorithm uses the following steps:

1) Read a set of *H* haplotypes with *L* linked *SSRs* per haplotype and compute their values of *V*, *H* and *a*.
2) Simulate values of θ_0, θ_1 and τ from their respective prior distributions.
3) Simulate a number of *H* haplotypes with *L* linked SSRs with a genealogy created under the coalescent model employing the demographic stepwise expansion model and the above values of θ_0, θ_1 and τ.
4) Estimate *V**, *H**, *a**, *P*, *MSH*, *DH* from the simulated linked *SSRs*.
5) If all of $|V\text{-}V^*|/V$, $|H\text{-}H^*|/H$ and $|a\text{-}a^*|/a$ are less than $\varepsilon = 0.01$, then record the values of θ_0, θ_1, τ, *P*, *MSH* and *DH*.

6) Return to 2) until N values of θ_0, θ_1, τ, P, MSH and DH are obtained (which means that we have obtained N accepted simulations).

The prior distribution used for θ_1 was uniform (0,200). For the reduction in the value of θ_1 to θ_0 we used a prior distribution that followed a uniform(0, 0.01), based on that reduction we obtained the value of θ_0 for each simulation. The prior distribution for τ was dependent on the value $\widetilde{\theta}_1$ sampled from each simulation and was distributed as uniform(0, $2^* \widetilde{\theta}_1$). That prior distribution for τ was motivated by the fact that in a constant population size with a θ value of X, the expected coalescent time multiplied by 2Lu is 2X if the number of samples is large. Therefore, we decided to leave the expected coalescent time in a constant population multiplied by 2Lu as an upper bound for the prior distribution of τ. The same set of prior distributions was used for all the *ABC* analysis reported in this paper.

The N values of θ_0, θ_1, τ are samples from the distribution $P(\Theta| \,|V - V^*|/V < \varepsilon, \;|H - H^*|/H < \varepsilon, |a - a^*|/a < \varepsilon)$ of those parameters if we define Θ as one demographic parameter. On the other hand,the N values of P, MSH and DH recorded are samples from the posterior predictive distribution $P(\mathrm{H}| \,|V - V^*|/V < \varepsilon, \;|H - H^*|/H < \varepsilon, |a - a^*|/a < \varepsilon)$, where H is a homoplasy measure [35]. The importance of choosing an appropriate ε value has been particularly well explained in a recent review [36], where the authors highlight that choosing a small ε value is important to make sure that the distributions $P(\Theta| \,|V - V^*|/V < \varepsilon, \;|H - H^*|/H < \varepsilon, |a - a^*|/a < \varepsilon)$ and $P(\mathrm{H}| \,|V - V^*|/V < \varepsilon, \;|H - H^*|/H < \varepsilon, |a - a^*|/a < \varepsilon)$ converge to the posterior distribution $P(\Theta|\,\mathrm{S})$ and $P(\mathrm{H}|\,\mathrm{S})$, respectively. In our case, we used an $\varepsilon = 0.1$, a value previously used in analysis containing the same summary statistics used here [27]. We evaluated if the *ABC* approach, including the choice of that value of ε, gave accurate point estimates of Θ and H and distributions that converge to the posterior distributions of Θ and H in the next section.

We used the function 'density' from R to build an estimate of the posterior distributions of each of the three parameters and the three homoplasy measures in the model based on the N values recorded of the homoplasy measures and the demographic parameters. The modes of the posterior and posterior predictive distributions were used as estimates of each of the parameters and homoplasy measures, respectively. We chose to use the mode, instead of the median or the mean, as our point estimate of the parameter and homoplasy measures for reasons that will be detailed in the next section.

To test the exactitude of our *ABC* method to estimate demographic parameters and homoplasy measures, we created 3 sets of simulations, where each simulation contained a set of 150 haplotypes composed of 6 linked *SSRs*. One set was used to analyze the change in the estimation of the homoplasy measures at ten different values of τ {1.5, 3, 4.5, 6, 7.5, 9, 10.5, 12, 13.5, 15}, performing 10 simulations for each value of τ. The other three sets were used to analyze the estimation of the homoplasy measures in three specific values of τ {3,6,9}, where we performed 100 simulations for each value of τ. Our *ABC* approach was run in each replicate of each simulation of those four sets until N= 10,000 simulated acceptances were obtained. Due to computational constraints associated with the calculation of the confidence intervals, we used N=1,000 simulated acceptances in the *Pinus caribaea* dataset.

Quality of demographic parameter estimates

We used a set of metrics to measure the quality of our parameter estimates using *ABC*, as recommended in [30]. One of those metrics is the relative bias, which is the average difference between the estimated and true value of the parameter divided by its true value [28]. First, we compared the relative bias when using the median, mode and mean as point estimates of the demographic parameters τ and θ_1, and of the homoplasy measures P, MSH and DH. This comparison was performed in the three different sets of 100 simulations described in the past section where the values of τ were 3, 6 and 9. We found that, on average, using the mode of the posterior distribution as a point estimate provided a smaller relative bias for estimates of τ, *MSH, P* and *DH* (Additional file 1: Table S3). The only parameter where using the mean of the posterior distribution provided a better point estimate was θ_1 (Additional file 1: Table S3). For simplicity and consistency, we decided to use the mode of the posterior distribution for all the point estimates we report. This decision is motivated by the fact that we were more interested in estimating τ and the homoplasy measures P, MSH and DH.

We evaluated our inferred posterior distributions estimating the 50%, 75% and 90% coverage, defined as the fraction of times that the true parameter estimate is inside the 50%, 75% and 90% credible intervals [28]. We define the X% credible interval as the X% highest posterior density interval (HPD). If the credible intervals created using the inferred posterior distributions are correct and have good coverage properties, the true parameter value should be present inside the 50%, 75% and 90% credible intervals with probabilities of 50%, 75% and 90%, respectively. We found that, on average, the 50%, 75% and 90% credible intervals for the demographic parameter τ, and the homoplasy measures P, DH and MSH had good coverage properties. τ and the homoplasy measures had a probability of being inside the credible interval that matched expectations, since the absolute difference between those probabilities and the expected probabilities was smaller than 0.05 (Additional file 1: Table S2). In the case of the parameter θ_1, we found that, on average, the parameter had slightly more conservative and broad intervals, since the true parameter tended to be inside the credible interval more often than what was expected, with the difference between the expected and observed probabilities of θ_1 being in the credible interval was between 0 and 0.1 (Additional file 1: Table S2). This indicates that our *ABC* method, including our selection of prior distributions and ε, is inferring accurate posterior distributions for the homoplasy measures and τ, and slightly biased posterior distributions for the parameter θ_1.

Command lines

```
Command Line 1:
./msHOT 150 100 -t 30 -eN $tau 0.001 -Q -z 6 -seeds
1 2 5
Where $tau took ten different values {0.025, 0.05,
0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25}.
Command Line 2:
./msHOT 150 100 -t $Theta -eN 0.1 0.001 -Q -z $i
-seeds 1 2 $i
Where $i took values going from 2 to 9, and $Theta
was equal to $i times 5.
Command Line 3:
./msHOT 150 100 -t 30 -eN $tau 0.001 -Q -z 6 -seeds
1 2 3
Where $tau took ten different values {0.025, 0.05,
0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25}.
Command Line 4:
./msHOT 150 200 -t 60 -eN $tau 0.001 -Q -z 6 -seeds
1 2 3
Where $tau = {0.0125, 0.025, 0.0375, 0.05, 0.0625,
0.075, 0.0875, 0.1, 0.1125, 0.125}
Command Line 5:
./msHOT 150 10 -t 30 -eN $tau 0.001 -Q -z 6 -seeds
1 2 3
Where $tau took ten different values {0.025, 0.05,
0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25}.
Command Line 6:
./msHOT 150 100 -t 30 -eN $tau 0.001 -Q -z 6 -seeds
2 3 4
Where $tau = {0.05, 0.1, 0.15}.
```

Abbreviations

ABC: Approximate Bayesian Computation; cpSSR: Chloroplast simple sequence repeats; DH: Distance Homoplasy; GBS: Genotyping-by-sequencing; HPD: Highest posterior density interval; ISM: Infinite Sites Model; LSWH: Least Squares approach without taking Homoplasy into account; MASH: Molecular size Homoplasy; MPH: Maximum pseudolikelihood using a model with Homoplasy; MSH: Mean size Homoplasy; P: Homoplasy index; RADseq: Restriction site-associated DNA sequencing; SMM: Stepwise mutation model; SSR: Simple sequence repeats

Acknowledgements

We thank Rodolfo Salas-Lizana, Pablo Vinuesa and Luis Eguiarte for their comments on a first draft of this paper, in DODV's undergraduate dissertation. Joshua Schraiber, Daniele Silvestro and one anonymous reviewer gave helpful comments on this paper. DODV was supported by a SEP-CONACYT scholarship, by an NIH postdoctoral fellowship and a UC-MEXUS CONACYT doctoral fellowship (213627). JVH was supported by a posdoctorate scholarship from the Dirección General de Asuntos del Personal Académico, UNAM.

Funding

DODV was supported by a SEP-CONACYT scholarship, by an NIH postdoctoral fellowship and a UC-MEXUS CONACYT doctoral fellowship (213627). JVH was supported by a posdoctorate scholarship from the Dirección General de Asuntos del Personal Académico, UNAM.

Availability of data and materials

The git repository https://github.com/dortegadelv/HomoplasyMetrics contains data, code and notes to reproduce the results from all the Figures and Tables.

Authors' contributions

DODV and JVH designed the study. DODV performed the simulations. DODV and JVH did the data analysis with suggestions from DP and LJB. DODV and JVH wrote the manuscript with input from DP and LJB. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Departamento de Ecologia Evolutiva, Instituto de Ecologia, Universidad Nacional Autónoma de México, Mexico City, Mexico. [2]Department of Integrative Biology, University of California, Berkeley, USA. [3]Plant Production Systems, Wageningen University, Wageningen, the Netherlands. [4]Centro de Investigaciones Interdisciplinarias en Ciencias y Humanidades, Universidad Nacional Autónoma de México, Mexico City, Mexico.

Received: 5 April 2017 Accepted: 10 August 2017
Published online: 11 September 2017

References

1. Beheregaray LB, Gibbs JP, Havill N, Fritts TH, Powell JR, Caccone A. Giant tortoises are not so slow : rapid diversification and biogeographic consensus in the Galápagos. Proc Natl Acad Sci U S A. 2004;101:6514–9.
2. Galbreath K, Cook J. Genetic consequences of Pleistocene glaciations for the tundra vole (Microtus Oeconomus) in Beringia. Mol Ecol. 2004;13:135–48.
3. Schneider S, Excoffier L. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genet. 1999;152:1079–89.
4. Rogers AR, Harpending H. Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol. 1992;9:552–69.
5. Wheeler GL, Dorman HE, Buchanan A, Challagundla L, Wallace LE. A review of the prevalence, utility, and caveats of using chloroplast simple sequence repeats for studies of plant biology. Appl Plant Sci. 2014;2:1400059.
6. Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, et al. Current trends in microsatellite genotyping. Mol Ecol Resour. 2011;11:591–611.
7. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008;3:1–7.

8. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6:1–10.
9. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007;39:1522–7.
10. Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol. 2002;11:2453–65.
11. Jaramillo-Correa JP, Gérardi S, Beaulieu J, Ledig FT, Bousquet J. Inferring and outlining past population declines with linked microsatellites: a case study in two spruce species. Genet Genomes. 2015;11:9.
12. Estoup A, Jarne P, Cornuet J-M. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol Ecol. 2002;11:1591–604.
13. Culver M, Menotti-raymond MA, Brien SJO. Letter to the editor patterns of size Homoplasy at 10 microsatellite loci in pumas (Puma Concolor). Mol Biol Evol. 2001;18:1151–6.
14. van Oppen MJ, Rico C, Turner GF, Hewitt GM. Extensive homoplasy, nonstepwise mutations, and shared ancestral polymorphism at a complex microsatellite locus in Lake Malawi cichlids. Mol Biol Evol. 2000;17:489–98.
15. Navascués M, Vaxevanidou Z, González-Martínez SC, Climent J, Gil L, Emerson BC. Chloroplast microsatellites reveal colonization and metapopulation dynamics in the Canary Island pine. Mol Ecol. 2006;15:2691–8.
16. Navascués M, Hardy OJ, Burgarella C. Characterization of demographic expansions from pairwise comparisons of linked microsatellite haplotypes. Genet. 2009;181:1013–9.
17. Kuhner MK. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics (Oxford, England) . 2006;22:768–70.
18. Navascués M, Emerson BC. Chloroplast microsatellites: measures of genetic diversity and the effect of homoplasy. Mol Ecol. 2005;14:1333–41.
19. Kimura M. Theoretical Foundation of Population Genetics at the molecular level. Theor Popul Biol. 1971;2:174–208.
20. Ohta T, Kimura M. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res. 1973;22:201–4.
21. Hellenthal G, Stephens M. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics (Oxford, England). 2007;23:520–1.
22. Hudson RR. Generating samples under a Wright-fisher neutral model of genetic variation. Bioinformatics. 2002;18:337–8.
23. Nei M. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics. 1978;89:583–90.
24. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and windows. Mol Ecol Resour. 2010;10:564–7.
25. Li W-H. Distribution of nucleotide differences between two randomly chosen Cistrons in a finite Population. Genet. 1977;85(2):331–7.
26. Jardón-Barbolla L, Delgado-Valerio P, Geada-López G, Vázquez-Lobo A, Piñero D. Phylogeography of Pinus subsection Australes in the Caribbean Basin. Ann Bot. 2011;107:229–41.
27. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol. 1999;16:1791–8.
28. Excoffier L, Estoup A, Cornuet JM. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics. 2005;169:1727–38.
29. Provan J, Soranzo N, Wilson NJ, Goldstein DB, Powell W. A low mutation rate for chloroplast microsatellites. Genet. 1999;153:943–7.
30. Bertorelle G, Benazzo A, Mona S. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol. 2010;19:2609–25.
31. Csilléry K, Blum MGB, Gaggiotti OE, François O. Approximate Bayesian computation (ABC) in practice.Trends Ecol Evol. 2010;25:410–8.
32. Huang Q-Y, Xu F-H, Shen H, Deng H-Y, Liu Y-J, Liu Y-Z, et al. Mutation patterns at dinucleotide microsatellite loci in humans. Am J Hum Genet. 2002;70:625–34.
33. Hizak J, Logozar R. A derivation of the mean absolute distance in one-dimensional random walk. Tech J. 2011;5:10–6.
34. Pritchard JK, Feldman MW. Statistics for microsatellite variation based on coalescence. Theor Popul Biol. 1996;50:325–44.
35. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. CRC Press; Boca Raton, Florida, USA. 2013.
36. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. Approximate Bayesian computation. PLoS Comput Biol. 2013;9:e1002803.