OPEN ACCESS Freely available online

PLOS ONE

Amount of Information Needed for Model Choice in Approximate Bayesian Computation

Michael Stocks[1]*, Mathieu Siol[2], Martin Lascoux[1], Stéphane De Mita[3,4]

1 Department of Ecology and Genetics, Program in Plant Ecology and Evolution, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden, **2** Institut National de la Recherche Agronomique (INRA), UMR Agroécologie, Dijon, France, **3** Institut National de la Recherche Agronomique (INRA), UMR Interactions Arbres-Microorganismes (IAM), Nancy, France, **4** Institut de Recherche pour le Développement (IRD), UMR Diversité et Adaptation des Plantes Cultivées, Montpellier, France

Abstract

Approximate Bayesian Computation (ABC) has become a popular technique in evolutionary genetics for elucidating population structure and history due to its flexibility. The statistical inference framework has benefited from significant progress in recent years. In population genetics, however, its outcome depends heavily on the amount of information in the dataset, whether that be the level of genetic variation or the number of samples and loci. Here we look at the power to reject a simple constant population size coalescent model in favor of a bottleneck model in datasets of varying quality. Not only is this power dependent on the number of samples and loci, but it also depends strongly on the level of nucleotide diversity in the observed dataset. Whilst overall model choice in an ABC setting is fairly powerful and quite conservative with regard to false positives, detecting weaker bottlenecks is problematic in smaller or less genetically diverse datasets and limits the inferences possible in non-model organism where the amount of information regarding the two models is often limited. Our results show it is important to consider these limitations when performing an ABC analysis and that studies should perform simulations based on the size and nature of the dataset in order to fully assess the power of the study.

Citation: Stocks M, Siol M, Lascoux M, De Mita S (2014) Amount of Information Needed for Model Choice in Approximate Bayesian Computation. PLoS ONE 9(6): e99581. doi:10.1371/journal.pone.0099581

Editor: Magnus Rattray, University of Manchester, United Kingdom

Received January 17, 2014; **Accepted** May 16, 2014; **Published** June 24, 2014

Funding: This work was supported by the European Community's Seventh Framework Programme (FP7/20072013), under grant agreement 211868 (Project Noveltree) and the Eranet Biodiversa LINKTREE and TIPTREE projects. M. Siol and S. De Mita were funded by Agropolis Fondation. The research trip of M. Stocks to IRD was funded by the EBC graduate school on Genomes and Phenotypes at Uppsala University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: mspopgen@gmail.com

Introduction

Central to evolutionary biology and science in general is the need to quantitatively compare models and hypotheses. In population genetics estimating parameters from more complex, biologically realistic models often involves a likelihood function that is difficult to compute. This has led to the development of methods, such as Approximate Bayesian Computation (ABC; [1]), that aim to approximate the likelihood function by simulating under a given model and using summary statistics to capture key aspects of the data in the most informative way (see [2] for an historical overview). Due to the flexibility and efficiency of ABC it is now possible to compare and estimate parameters from a number of complex models, and this has led to the widespread adoption of the method within the population genetics community for assessing and fitting demographic models to molecular data.

Understanding the evolutionary history of a population is an important aspect of studies on natural populations. Aside from giving information about the evolutionary past of organisms, inferring the demographic history and structure of a population is also necessary to understanding the effect of other population genetic processes. For instance, studies aiming to infer signatures of selection at candidate loci or across the genome depend on first knowing the background patterns of genetic variation produced by historical demographic events [3,4]. Methods for estimating demographic histories have therefore become increasingly important, and have fuelled the proliferation of studies using ABC to infer a suitable demographic model.

A typical ABC workflow would consist of a number of steps: i) choose a set of summary statistics describing a given dataset; ii) perform a large number of simulations sampling a pre-supposed distribution of models and model parameters; iii) compute the summary statistics for the simulations; iv) apply a rejection threshold to focus on a region of the parameter space where the relationship between the summary statistics and parameters is assumed to be linear; v) perform either a regression to evaluate model parameters or perform a logistic regression to compare models. There are alternatives to this workflow, but this is the approach most commonly implemented in ABC analyses. The great strength of ABC lies in its flexibility, allowing the user to address a very large set of demographic models.

There are, however, a number of caveats associated with the approximation quality of ABC. These have been well-documented in the literature, but perhaps the most important consideration, and which is inherent to the ABC procedure, is in choosing informative summary statistics [5–7]. The field of population genetics has a long history of summarizing patterns of genetic variation in a way that is sensitive to departures from the standard neutral model. However, the extent to which summary statistics accurately represent the data is hard to evaluate and might be a major limitation to model inference, and particularly model choice [8]. This process has been relatively overlooked in the literature

compared with advances in statistical methods that all somehow assume that data are properly summarized. Furthermore, even if a certain set of summary statistics is informative in accurately estimating parameters from two different demographic models separately, the same set of summary statistics may be uninformative when it comes to comparing these two models with each other [8].

Besides estimating demographic parameters such as population divergence times or migration parameters, model choice is central to many questions in population genetics. The problem of appropriately summarizing the data could be more important for datasets containing low levels of information, either because of an insufficient sampling effort or low levels of variation. Using population genetic simulations, we try to identify what happens when limitations are placed on the amount of information in the data, such as sample size, number of loci and level of genetic diversity. Firstly, this mimics many studies of natural populations where constraints are placed on the amount of data that can be collected. As the use of ABC has increased, so too has it been embraced in non-model organisms where the number of loci and samples are often limited. It is therefore of great interest to understand how a limited dataset impacts the use of ABC model choice. Secondly, it highlights which aspects of an ABC analysis, including the choice of summary statistics, are important in determining the power to reject a null demographic model in favor of a more complex alternative. Placing constraints on the data limits the amount of information available for comparing models, and by doing this we look to tease apart the factors contributing to the power of model choice in ABC.

Here, we use simulations to explore the power of model choice in ABC. In particular, we concentrate on two simple coalescent models commonly used in population genetic studies. The first is a null model of constant effective population size (Standard Neutral Model - SNM), and the second is a simple bottleneck model (BNM) that acts as our alternative model. Bottlenecks are known to occur frequently in natural populations (e.g. [9–12]) and are one of the most commonly investigated demographic models. There is considerable interest in understanding the patterns that bottlenecks leave in genetic data and a lot of work has gone into correctly inferring the parameters of bottleneck models in model species such as *Homo sapiens* and *Drosophila melanogaster* (reviewed in [13]). The model also contains a parameter controlling the severity of the bottleneck, and varying this parameter allows us to investigate the performance of model choice in ABC in more detail. We begin by exploring the relationship between the parameters of the models and a number of summary statistics commonly used in population genetics. Using a subset of these summary statistics, we assess the power to reject the SNM in favor of the BNM whilst varying: 1) the quality of the dataset; 2) the severity of the bottleneck; and 3) the tolerance of the rejection step.

Results

Choice of summary statistics

Figure 1 shows correlation coefficients between different summary statistics and the parameters of the SNM and BNM for the largest dataset with high genetic variance ($n=20$, $l=30$, $\theta=0.005$). The parameter θ is strongly positively correlated with the means of many statistics, such as θ_W, π, θ_H, H_e and S, as well as their quantiles and standard deviations. One exception is the standard deviation of H_e which is strongly negatively correlated with θ. This is in sharp contrast with π which responds positively to an increase in segregating sites. The standard deviation of Tajima's D is negatively correlated with θ, showing that its precision increases with increasing variation. The average value of Fay and Wu's H is independent of θ in the SNM and is slightly positively correlated with θ in the BNM. The variance of H is strongly positively correlated with θ, but the width of the interval between the 5% and 95% quantiles of H increases markedly with increasing θ, although this is due to the use of the non-standardized version of H. Finally, the site frequency spectrum is little affected by the mutation rate, save for a small positive correlation of s1 with the SNM, possibly due to an increased power of detection of rare variants with large values of θ. In contrast, ρ does not have any strong correlation with the means of the statistics. Its effect, however, is visible on the standard deviation, as recombination reduces the variance of the coalescent process. The mean of some statistics, such as Tajima's D, Fay & Wu's H and the site frequency spectrum appear independent of θ, whereas the mean of H_e shows a strong correlation, reflecting an increase in the number of haplotypes with an increase in recombination. Parameters specific to the BNM (N_B and T) are both weakly correlated with most statistics, with the direction of the correlation consistent for both parameters. Both N_B and T are most strongly correlated with Tajima's D ($r_{D,T}=0.299$, $r_{D,N_B}=0.373$), Fay & Wu's H ($r_{H,T}=-0.059$, $r_{H,N_B}=-0.369$) and the low frequency class of the site frequency spectrum ($r_{s1,T}=-0.239$, $r_{s1,N_B}=-0.337$). Among the three classes of the SFS (s1, s2 and s3) the proportion of low frequency variants (s1) is negatively correlated with both N_B and T. Interestingly, more recent and stronger bottlenecks (that is, low values for N_B and T) both result in negative D and an excess of rare variants with a corresponding depletion of high frequency classes. H and its variance also respond to the bottleneck parameters, particularly its severity.

We performed a Principal Component Analysis (PCA) in order to quantify the main features that can be extracted from summary statistics. PCA reduces the dimensions of a set of potentially correlated variables into a smaller set of uncorrelated variables that best explain the variance in the data. Figure S1 shows a PCA of the SNM ($N_B=N_0$, $n=20$, $l=30$, $\theta=0.005$) and the BNM ($N_B=0.1N_0$, $n=20$, $l=30$, $\theta=0.005$) contrasted against the complete parameter space of the BNM ($n=20$, $l=30$). The two models are clearly separated on the first principal component (PC)

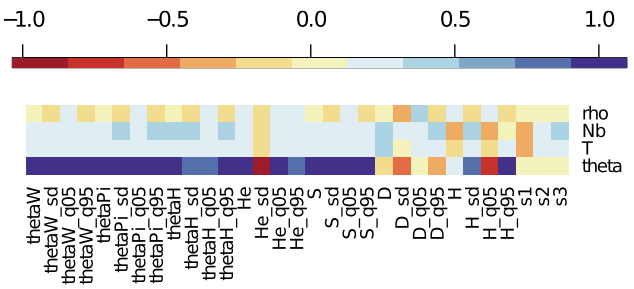

Figure 1. Correlation between parameters and summary statistics. Correlation coefficients between parameters and summary statistics for a SNM (top) and a BNM (bottom) for the larger dataset ($n=20$, $l=30$). The parameters are given by theta and rho in the SNM and by theta, botEnd, botNe and rho in the BNM. The summary statistics give the average, standard deviation (_sd) and the 5 and 95% quantiles (_q5 and _q95) for three common estimates of θ (thetaW, thetaPi, thetaH), haplotypic diversity (He), the number of segregating sites (S) and two tests of neutrality (Tajima's D and Fay & Wu's H). The site frequency spectrum is binned into 3 frequency classes (s1, s2, s3), ranging from low frequency (s1) to high frequency (s3) variants, that represent the proportion of segregating sites that occur in each class.
doi:10.1371/journal.pone.0099581.g001

indicating that there is enough information in the summary statistics to distinguish between the models. There is still a separation between the models on the second and third PCs, but they cannot be distinguished on the fourth PC. Figure 2 shows the first four PCs (PC1, PC2, PC3 and PC4) for the summary statistics calculated under a BNM. Each of the (independent) principal components can be linked to a feature of the genetic polymorphism patterns, and these features can be measured by one or more statistics. For example, PC1 represents the parameter θ as many of the θ estimators cluster together and there is the same pattern of correlation with statistics as observed in Figure 1. The first PC captures most of the signal (94.1%), showing that θ, which controls the amount of genetic variation, is the major parameter shaping the patterns of polymorphism. The second PC captures far less of the signal (3.6%) and likely represents the shape of the site frequency spectrum, as there is differentiation among statistics known to be influenced by the shape of the site frequency spectrum. It is essentially independent from the first PC, and correlated with the three site frequency spectrum categories, the neutrality tests D and H and the quantiles and standard deviation of D (those of H are strongly correlated with θ). The low-frequency variants and Tajima's D contribute most to this PC, although in opposite directions. The third and fourth PCs capture less of the variation (1.3% and 0.4% respectively) but there are still interpretable patterns in the data. In particular, D^{05}, H and H^{05} are clustered separately from the rest of the summary statistics on the third PC and are then separated on the fourth PC. It is possible that these represent derived alleles and that this could be informative for model choice.

Comparison amongst sets of summary statistics

We looked at the effect that a 90% reduction in the effective population size ($N_B = 0.1N$) has on the power to reject the SNM for different sets of summary statistics (Table 1). All sets of summary statistics give good power for large datasets with high nucleotide diversity ($n = 20$, $l = 30$, $\theta = 0.005$), with the proportion of Bayes factors exceeding 3 all being greater than 0.99 ($\Psi_{TPH} = 0.985$; $\Psi_{SFS3} = 0.991$; $\Psi_{T+SFS3} = 0.999$; $\Psi_{SFS5} = 0.997$; $\Psi_{TPH+DH} = 1$). For a large dataset with low genetic diversity ($n = 20$, $l = 30$, $\theta = 0.0015$), three sets of summary statistics are able to reject the SNM with a power greater than 0.9 ($\Psi_{SFS3} = 0.901$; $\Psi_{T+SFS3} = 0.907$; $\Psi_{TPH+DH} = 0.95$), with TPH, in contrast, having very low power ($\Psi_{TPH} = 0.019$). Smaller datasets afford less power to reject the SNM, except when using the TPH+DH set of summary statistics with high nucleotide diversity ($n = 10$, $l = 15$, $\theta = 0.005$), which still allows for a high degree of power in rejecting the SNM ($\Psi_{TPH+DH} = 0.95$).

Table 1 also shows the proportion of times that the SNM is falsely rejected (α). For three of the statistics (TPH, SFS3 and SFS5), the false positive rate is marginally higher in larger datasets, whereas for the TPH+DH set of statistics the pattern is the opposite. In small datasets, false positives are very rare (< 0.01) with a larger θ tending to decrease the rate of false positives. Different sets of summary statistics also result in different α: in particular, SFS3 tends to cause slightly more false positives than other sets of summary statistics. The number of false positives are small so any patterns may be influenced by the underlying variance. However, in none of the categories does the false positive rate reach 5%. While this is encouraging, it is important to note that the false positive rate increases as the Bayes factor used to determine significance is reduced. For large datasets with low genetic variation and the TPH+DH set of summary statistics the false positive rate is 0.166 and 0.397 for Bayes factor cutoffs of 1.5 and 1 respectively. Similarily, for small datasets with high genetic variation the rate is 0.110 and 0.411 for cutoffs of 1.5 and 1 respectively.

In general, the inclusion of statistics that summarize elements of the site frequency spectrum give greater power, which is most clearly reflected in smaller, low diversity datasets ($n = 10$, $l = 15$, $\theta = 0.0015$) by the power difference between TPH and the other three sets of summary statistics ($\Psi_{TPH} = 0$; $\Psi_{SFS3} = 0.306$; $\Psi_{T+SFS3} = 0.214$; $\Psi_{SFS5} = 0.339$; $\Psi_{TPH+DH} = 0.373$). The distri-

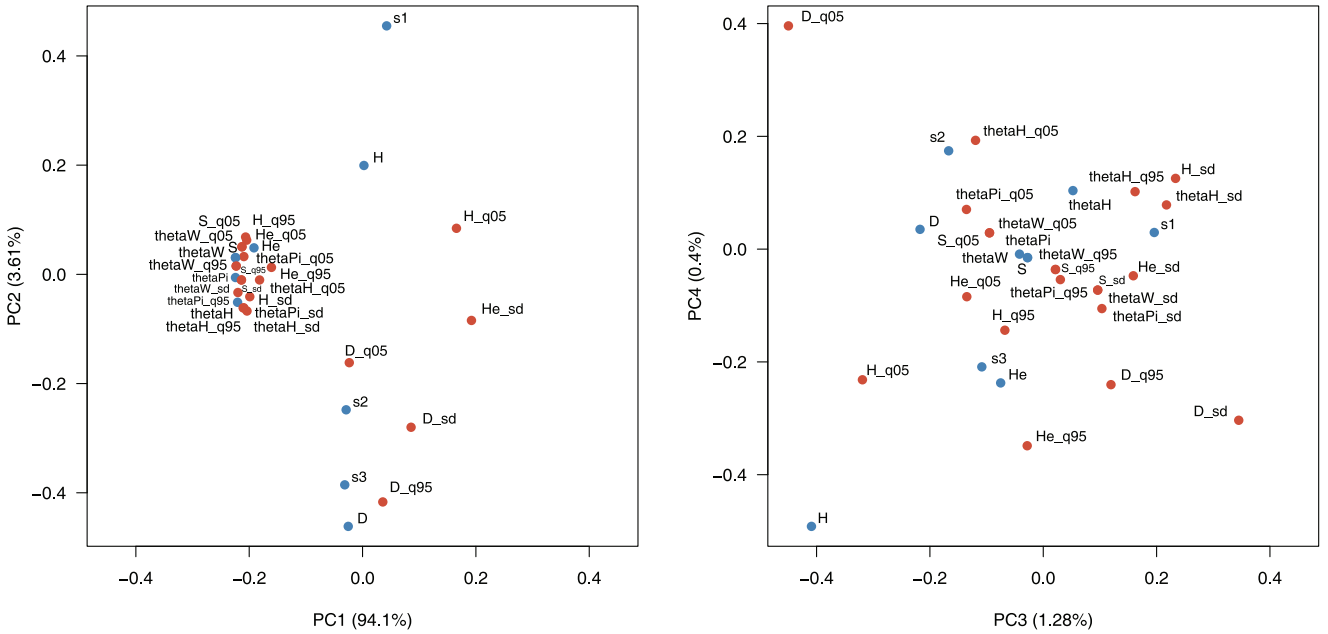

Figure 2. Principal Component Analysis of summary statistics under a bottleneck model. The first four principal components (PCs) of a PCA for summary statistics calculated for 10,000 simulations of a BNM in a large, genetically diverse sample ($n = 20$, $l = 30$, $\theta = 0.005$).
doi:10.1371/journal.pone.0099581.g002

Table 1. Power and false positive rate for different sets of summary statistics.

		$\theta = 0.0015$					$\theta = 0.005$				
		TPH	SFS3	T+SFS3	SFS5	TPH+DH	TPH	SFS3	T+SFS3	SFS5	TPH+DH
$\Psi_{0.1N}$	$n = 20, l = 30$	0.019	0.901	0.738	0.907	0.95	0.985	0.991	0.999	0.997	1
	$n = 10, l = 15$	0	0.306	0.214	0.339	0.373	0.287	0.278	0.482	0.447	0.95
$\alpha_{0.1N}$	$n = 20, l = 30$	0	0.041	0.029	0.028	0.029	0.01	0.01	0.019	0.004	0.008
	$n = 10, l = 15$	0	0.002	0.003	0.009	0.04	0.003	0.001	0.011	0	0.01

The power (Ψ) and false positive rate (α) associated with rejecting a SNM in favor of a BNM for the TPH (θ_W, π, H_e), SFS3 (3 bin relative site frequency spectrum), T+SFS3 (θ_W, 3 bin relative site frequency spectrum), SFS5 (5 bin relative site frequency spectrum) and TPH+DH (θ_W, π, H_e, D, H) sets of summary statistics. $N_B = 0.1N$.

doi:10.1371/journal.pone.0099581.t001

butions of model probabilities (Figure S2) under the SNM and BNM overlap for TPH, even for larger datasets where an increase in samples and loci leads only to a shift in the mean of the distribution so that the models are difficult to separate. However, for sets of summary statistics that incorporate the site frequency spectrum (TPH, SFS5 and TPH+DH) a larger dataset increases the power to distinguish between the two models, and the distribution of model probabilities is skewed.

Somewhat surprisingly, including θ_W with SFS3 decreases the power for datasets with low genetic diversity ($n = 10$, $l = 15$: $\Psi_{SFS3} = 0.306$, $\Psi_{T+SFS3} = 0.214$; $n = 20$, $l = 30$: $\Psi_{SFS3} = 0.901$, $\Psi_{T+SFS3} = 0.738$). To investigate the performance of T+SFS3 in datasets with low and high levels of nucleotide diversity, we extended our analysis to include scenarios with more severe ($N_B = 0.01N$) and less severe ($N_B = 0.2N$) bottlenecks. In genetically diverse datasets there is a clear difference between the value of θ_W in BNMs and SNMs (Figure 3), suggesting that the statistic is informative in choosing between the models. In contrast, the distributions of θ_W for datasets of low nucleotide diversity overlap to some extent, implying that in this case the statistic is not as informative for distinguishing between the two models.

In all scenarios considered, an increase in the severity of the bottleneck increases the power to correctly reject the SNM (Table 2). For all sets of summary statistics, there is excellent power to detect a 90% ($0.1N$) or 99% ($0.01N$) reduction in the effective population size with a large, genetically diverse sample ($n = 20$, $l = 30$, $\theta = 0.005$). The choice of summary statistic becomes increasingly important for a smaller reduction in the population size of 80% as, even for large, genetically diverse samples ($n = 20$, $l = 30$, $\theta = 0.005$), only the TPH+DH set of statistics performs well (0.968). The TPH set of statistics performs poorly with smaller or genetically less diverse datasets, even for strong bottlenecks ($0.01N$), and seems to perform particularly badly in samples with low diversity.

The TPH+DH set of statistics performs better than all other sets of statistics for all dataset types and bottleneck strengths tested. The power to detect population size reductions of 90% or more is greater than or equal to 0.95 for all but the worst datasets, whether that be a smaller dataset with higher genetic diversity ($\Psi_{TPH+DH} = 0.95$; $n = 10$, $l = 15$, $\theta = 0.005$), or a larger dataset with lower genetic diversity ($\Psi_{TPH+DH} = 0.95$; $n = 20$, $l = 30$, $\theta = 0.0015$). For smaller datasets with low genetic diversity, however, the power is still low to reject the SNM ($\Psi_{0.2N} = 0.205$; $\Psi_{0.1N} = 0.373$; $\Psi_{0.01N} = 0.546$). To dissect the performance of TPH+DH, we looked at the value of Tajima's D and Fay & Wu's H as a function of the model probabilities for bottlenecks of varying severity. The value of Tajima's D (Figure 4 and Figure S3) decreases with an increase in the severity of the bottleneck. For small datasets with low genetic variation ($n = 10$, $l = 15$, $\theta = 0.0015$), the overlap in values of Tajima's D is considerable, even between the SNM ($N_B = N$) and the most severe bottleneck model ($N_B = 0.01N$), and this is in line with the low levels of power for this type of dataset ($\Psi_{0.2N} = 0.205$; $\Psi_{0.1N} = 0.373$; $\Psi_{0.01N} = 0.546$). Larger datasets ($n = 20$, $l = 30$) reduce the variation, whilst higher levels of genetic variation produce more negative mean values of Tajima's D (large dataset/low genetic variation: $D_{0.2N} = -0.40$, $D_{0.1N} = -0.55$, $D_{0.01N} = -0.68$; large dataset/high genetic variation: $D_{0.2N} = -0.55$, $D_{0.1N} = -0.75$, $D_{0.01N} = -0.91$). An increase in the severity of the bottleneck causes an overall increase in Fay & Wu's H (Figure S4). There is considerable difference between datasets of low and high nucleotide diversity in the variance of H. There is just a small shift in the mean of H in datasets with low nucleotide diversity ($\theta = 0.0015$) as the severity increases, suggest-

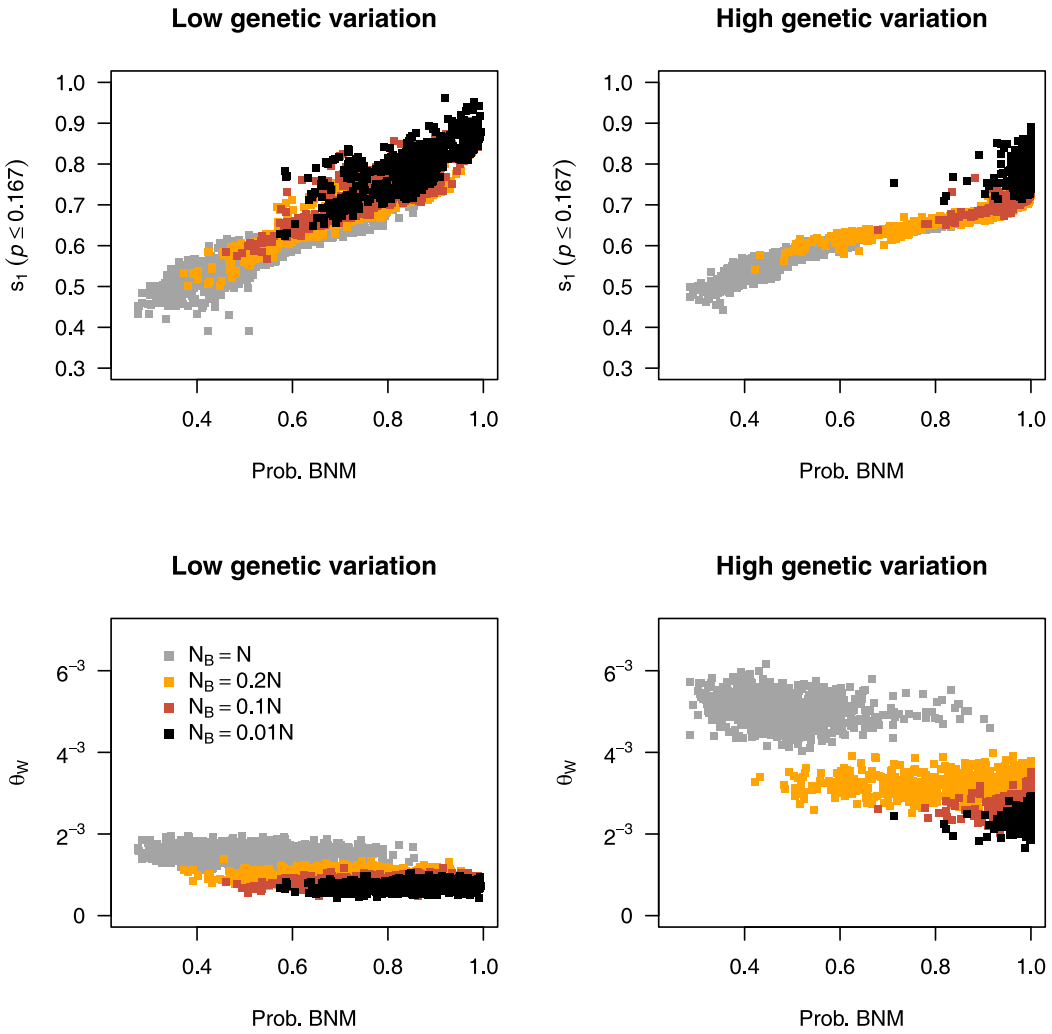

Figure 3. The impact of bottleneck severity and dataset quality on model probabilities and summary statistics. The effect of bottleneck strength on the value of summary statistics and model probabilities for larger datasets (n = 20, l = 30) with low ($\theta = 0.0015$) or high ($\theta = 0.005$) levels of genetic variation. Each point represents the rejection step of an ABC analysis when the T+SFS3 set of statistics is used with a tolerance of 0.001. The y-axis of the top two panels show the values of the first bin of the relative site frequency spectrum s1 (representing rare alleles) and the bottom panels display the value of Watterson's Theta (θ_W). The effective population size during the bottleneck (N^B) is defined relative to the recovered effective population size (N).
doi:10.1371/journal.pone.0099581.g003

ing that H is uninformative. However, in datasets with high nucleotide diversity ($\theta = 0.005$) there is a larger increase in H with increasing bottleneck severity.

For the most informative set of statistics (TPH+DH) we performed additional analyses for bottlenecks of different levels of severity ($N_B = 0.5N$ and $N_B = 0.3N$) to better understand the relationship between bottleneck severity and the power to reject the SNM (Figure 5A). The power to detect weak bottlenecks ($N_B = 0.5N$) is low for each of the datasets tested. Contrastingly, for severe bottlenecks with a 99% reduction in the effective population size, all but the most limited datasets have very high power ($\Psi \geq 0.997$). For a small dataset with low genetic diversity ($n = 10$, $l = 15$, $\theta = 0.0015$), there is low power to reject the SNM for even the most severe bottlenecks ($\Psi_{0.01N} = 0.546$).

Varying the tolerance

We investigated the power of model choice in ABC when the tolerance was varied (Table S1 and Figure S5). Generally, a stricter tolerance leads to a higher level of power, especially for the TPH set of statistics, where the choice of tolerance is very important. For larger datasets with high nucleotide diversity ($n = 20$, $l = 30$, $\theta = 0.005$), choosing a tolerance of 0.001 gives a power of 0.985 to reject the SNM, whereas using a tolerance level of 0.005 would result in a much lower power value of 0.525. Similarly, for large datasets with low nucleotide diversity ($n = 20$, $l = 30$, $\theta = 0.0015$) there is virtually no power to correctly reject the SNM, except when a tolerance of 0.0001 is used, which gives a power value of 0.688. However, a stricter tolerance leads to a higher rate of false positives, with rates reaching a maximum of 1.7% ($n = 20$, $l = 30$, $\theta = 0.005$, tol = 0.0001) and 3.2% ($n = 20$, $l = 30$, $\theta = 0.0015$, tol = 0.0001) for the TPH and SFS5 set of statistics respectively. The choice of tolerance appears to decrease in importance as the summary statistics capture more features of the site frequency spectrum. Tolerance levels of 0.01, 0.005 or 0.001 all give similar levels of power for the SFS5 set of summary statistics, with notable decreases being observed only with the most extreme tolerances (0.1 and 0.0001). It is worth noting however

Table 2. The effect of bottleneck severity on power.

		$\theta = 0.0015$			$\theta = 0.005$		
		$0.2N$	$0.1N$	$0.01N$	$0.2N$	$0.1N$	$0.01N$
$n = 10$, $l = 15$	TPH	0	0	0	0.112	0.287	0.395
	SFS5	0.154	0.339	0.497	0.132	0.447	0.745
	T+SFS3	0.097	0.214	0.33	0.227	0.482	0.732
	TPH+DH	0.205	0.373	0.546	0.614	0.95	0.999
$n = 20$, $l = 30$	TPH	0.002	0.019	0.031	0.707	0.985	0.999
	SFS5	0.602	0.907	0.973	0.789	0.997	1
	T+SFS3	0.434	0.738	0.881	0.835	0.999	0.999
	TPH+DH	0.71	0.95	0.997	0.968	1	1

The power to correctly reject the SNM in favor of BNMs of different strengths, expressed as the relative effective population size during the bottleneck.
doi:10.1371/journal.pone.0099581.t002

that these patterns might just represent random deviations caused by the low number of false positives and higher variance.

Finding the optimal dataset size

In population genetic studies of natural populations there may be limitations on the quality of the dataset available for sampling. It is therefore of interest to ask how many samples or loci need to be obtained in order to have a 95% power to detect a bottleneck. We therefore extended our analyses, using the most informative set of summary statistics (TPH+DH), to include datasets with samples of between 5 and 40 individuals, where the number of loci varied between 1 and 60 and the level of genetic variation was low ($\theta = 0.0015$). Figure 5B shows the power as a function of the product of the number of samples (n) and loci (l) as this can be seen as being proportional to the sequencing cost of a study. The relationship between Ψ and nl is sigmoidal, with the addition of loci and samples increasing the power above 95% when $nl \geq 600$. However, the contributions of the number of samples and loci to the power is not equal, with datasets having more power when $l > n$.

Figure 5 also shows the power to reject the SNM in favor of a BNM for a different number of samples when the number of loci are limited ($l = 15$ or 30; Figure 5C) and for a different number of loci when the number of samples is limited ($n = 10$ or 20; Figure 5D). When the number of loci are limited to 15 the power to reject the SNM remains well below 95%, even for large sample sizes ($\Psi_{n=40} = 0.762$), and the relationship appears to be asymptotic. Increasing the number of loci to 30 greatly increases the power such that sampling more than 20 individuals means that the power is greater than 95%. When the number of samples are limited 95% power is reached when ~ 50 and 30 loci are used for sample sizes of 10 and 20 individuals respectively.

Parameter estimation

To assess the ability of ABC to estimate parameters of a BNM in limited datasets, we performed parameter estimation using the local linear regression method described in [1]. Figure 6 and Table 3 summarize the distribution of the means of the posterior distributions for each replicate under a BNM for large datasets ($n = 20$, $l = 30$). The parameter θ is estimated well by three of the four sets of summary statistics. SFS3, however, gives a very poor estimate of θ (0.00413) in low diversity samples, whilst the estimate is far better in high diversity samples (0.00475). The time of the bottleneck (T) was estimated better in the high diversity samples than in the low diversity samples, with an increase in nucleotide diversity also decreasing the variance of the posterior means. The effective population size during the bottleneck (N_B) is estimated well for TPH+DH in both the low and high diversity datasets. For the rest of the sets of summary statistics, samples with high nucleotide diversity give better results. Table S2 shows the proportion of replicates where the true value lies within the 90%, 50% and 10% credible intervals of the parameter posterior distributions. One notable observation is that, for SFS3, the proportion of replicates in which the true value lies within the credible intervals is surprisingly high. This is particularly striking for high levels of variation and 10% credible intervals where a proportion of 0.578 was found for SFS3, compared to much smaller values for the other summary statistics [7]. It is also interesting to note that, for each parameter for TPH+DH, the proportion of replicates where the true value lies within the credible intervals is better for low nucleotide diversity ($\theta = 0.0015$) than for high nucleotide diversity ($\theta = 0.005$).

Discussion

In this study we chose to address the power of ABC model choice when the amount of data is the limiting factor. In ABC we are challenged with the task of summarizing the data such that these contrasting patterns are captured and there is enough information to distinguish between the two competing models. However, summary statistics that are capable of separating, for example, a population genetic model of constant effective population size from a population expansion model will not be the same as those capable of separating a structured population from an unstructured population [14]. Therefore, the way in which the data can be summarized most informatively will be highly context-dependent [15]. We began by first exploring the behaviour of summary statistics in a bottleneck model, and proceeded by investigating the power that different sets of summary statistics have in separating a bottleneck model from a simple model of constant effective population size.

Choosing summary statistics

A number of studies (e.g. [16,17]) have used correlation coefficients and PCA to guide their choice of summary statistics. We find that when using PCA it was possible to identify categories of summary statistics that are informative for separating the SNM and BNM. On the first PC, statistics strongly correlated with θ are separated from those based on the shape of the site frequency spectrum. This is in agreement with the finding that the inclusion

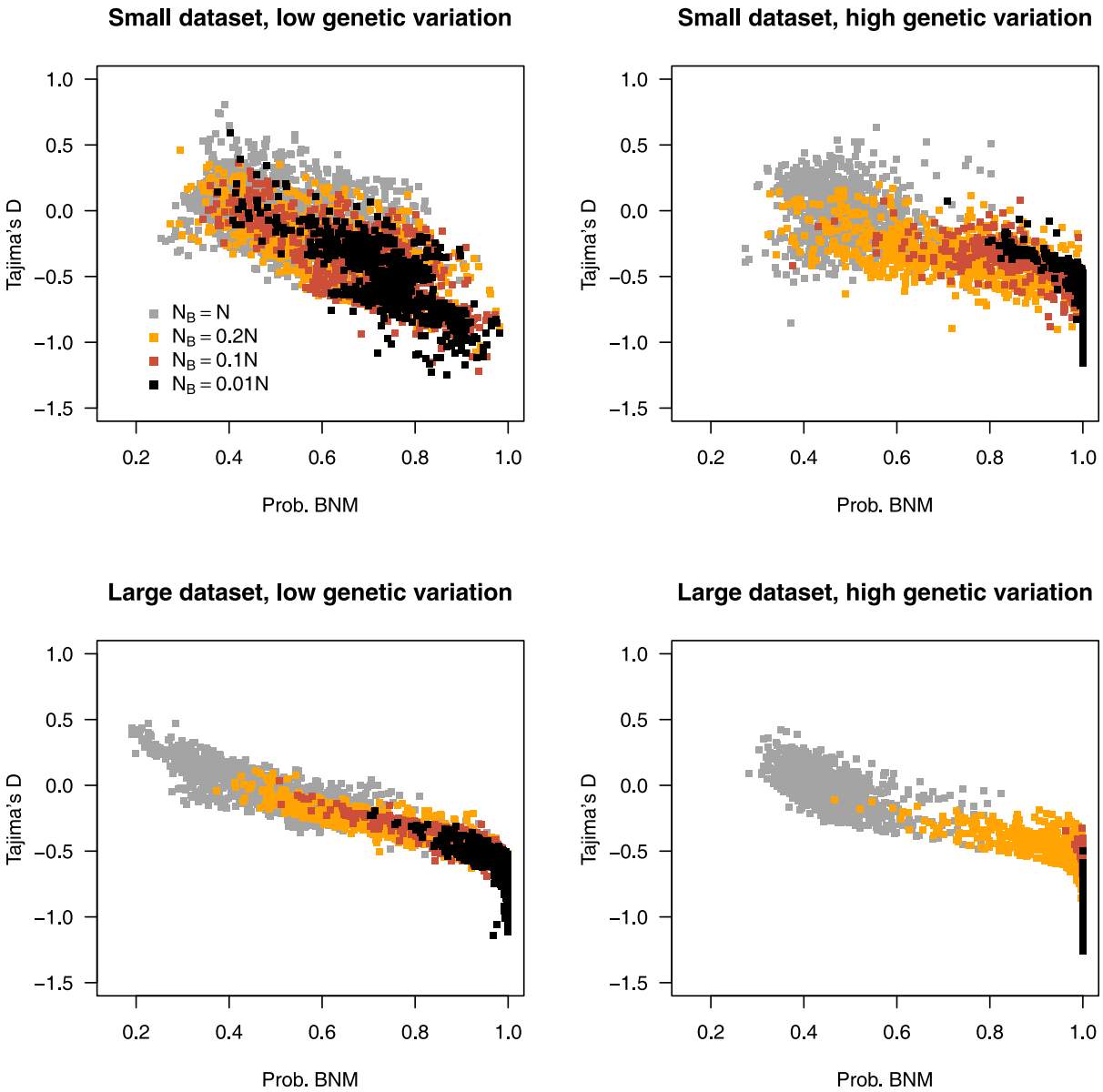

Figure 4. The impact of bottleneck severity and dataset quality on Tajima's D. The effect of bottleneck strength on the value of Tajima's D and model probabilities for both small ($n=10$, $l=15$) and large ($n=20$, $l=30$) datasets with low ($\theta=0.0015$) and high ($\theta=0.005$) genetic variation. Each point represents the rejection step of an ABC analysis when the TPH+DH set of statistics is used with a tolerance of 0.001. The effective population size during the bottleneck (N^B) is defined relative to the recovered effective population size (N).
doi:10.1371/journal.pone.0099581.g004

of SFS-based statistics increases the power to reject the SNM in favor of a BNM. This is in line with population genetic expectations, a population bottleneck affects the average number of nucleotide differences more strongly than the number of segregating sites [18], and is the basis for Tajima's D. Tajima's D proved to be an effective and informative summary of the site frequency spectrum, becoming more negative with an increase in the severity. There may also be great benefit in using the unfolded site frequency spectrum. Only the folded site frequency was tested here, but any additional information given by knowing the derived allele could boost the power.

Additional signals in the third and fourth PCs indicate that derived alleles describe some of the variation, even if these PCs accounted for a small fraction of the overall variation (about 1.7% combined). Combining the signals from both low frequency and high frequency variants proved successful in increasing the power to reject the SNM. For example, of all the summary statistics tested, TPH+DH gave the highest power, and it may be that the combination of Tajima's D and Fay & Wu's H summarizes the site frequency spectrum in an informative way. Fay & Wu's H is often neglected as inferring the derived allele depends on there being a suitable outgroup available, which is not always the case in non-model organisms. In contrast, haplotypic information didn't appear to be overally informative, although this could simply be due to the size of the fragments simulated (750 bp). For the power analysis we assessed only the means of the statistics, but there may be information in the standard deviation and the quantiles of some of the summary statistics. In particular, the standard deviation of

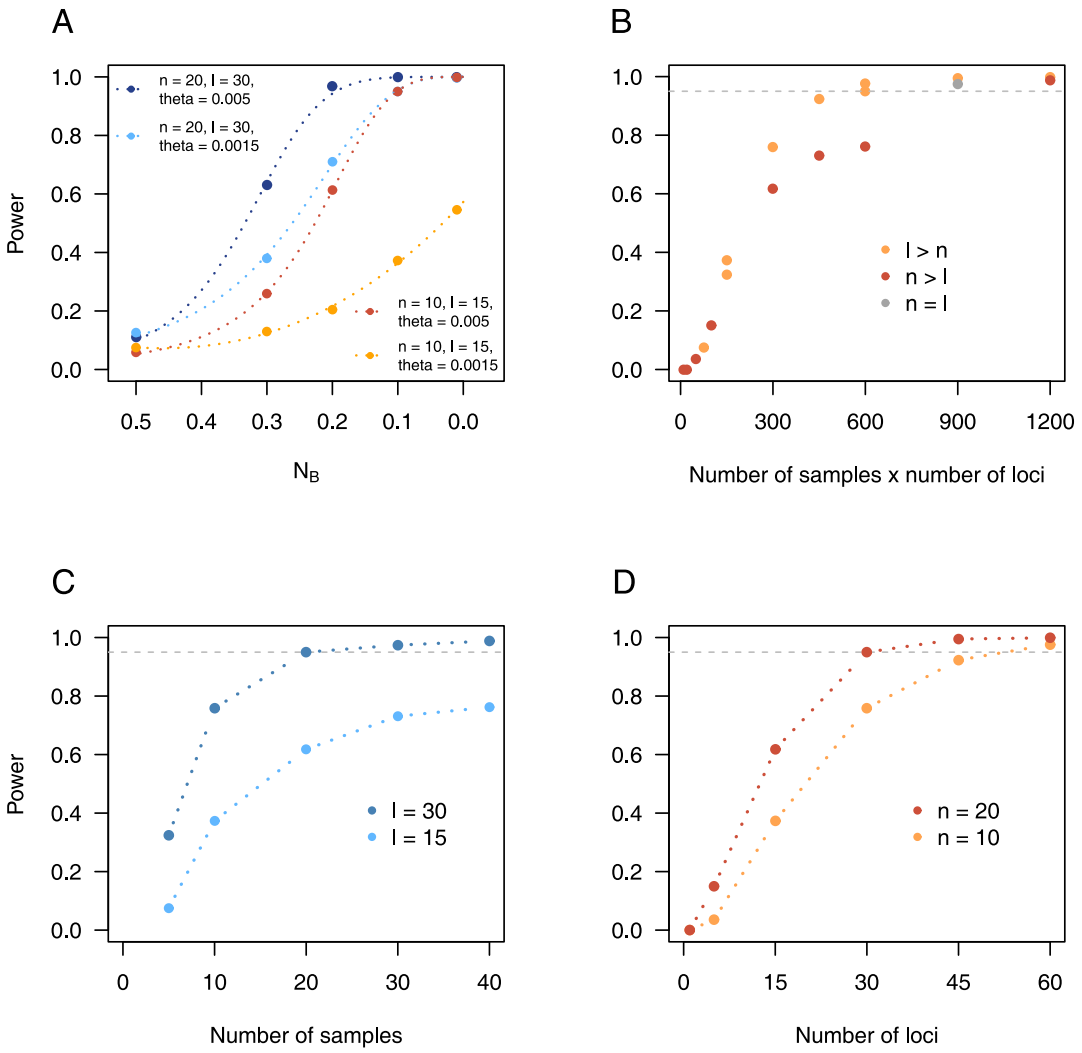

Figure 5. The effect of sample size and the number of loci on power. The power to reject the SNM, for the TPH+DH set of summary statistics, as a function of (A) the bottleneck severity (expressed as the relative effective population size N_B), (B) the product of the number of loci and the number of samples ($N_B = 0.1N$, $\theta = 0.0015$), (C) the number of samples ($N_B = 0.1N$, $\theta = 0.0015$), and (D) the number of loci ($N_B = 0.1N$, $\theta = 0.0015$). In B, C and D the dotted line corresponds to a power of 0.95.
doi:10.1371/journal.pone.0099581.g005

the haplotypic diversity showed a negative correlation with parameters of the BNM model.

One of the key considerations when choosing summary statistics has been in avoiding the curse of dimensionality [1]. As the number of summary statistics increases, so too does the variability in the parameter estimates in the regression step of ABC, leading to poorer parameter estimates. It has been suggested that model choice may be affected less by this problem [19]. While our results generally support this, we do find some limited evidence to the contrary in datasets with low genetic diversity where the power is lower for T+SFS3 compared to the SFS3 set of summary statistics. Accordingly, a number of methods have been established for identifying informative summary statistics in relation to estimating parameters. For example, [20] weight statistics according to the information they give for a parameter of interest, whilst [21] implement a partial least-squares transformation of the summary statistics and [22] apply a machine learning technique (boosting) to find the most informative summary statistics. More recently, research has moved towards identifying summary statistics for model choice. [19] use logistic discriminant analysis to process summary statistics before model choice, [23] weight the summary statistics for model choice after a preliminary regression step and [24] derive conditions under which summary statistics are sufficient for selecting the true model.

While we find PCA to be a highly informative way of summarizing the data, it may not be enough to simply perform PCA and look for patterns in the summary statistics. Some results of our analyses were counter-intuitive, such as the finding that the T+SFS3 set of summary statistics performed worse than SFS3 in datasets with low genetic diversity. This suggests that there can be a complex relationship between some summary statistics and the parameters of the model. This may be the case in the above example where Watterson's θ only adds noise rather than any additional information to the low diversity datasets and could, as discussed above, be due to the curse of dimensionality. However, the SFS3 set of statistics estimates the parameter θ poorly. This may be due to the use of the relative instead of the absolute site frequency spectrum, as information in the absolute SFS regarding θ is lost when it is re-scaled. This draws to attention an interesting problem concerning the choice of summary statistics in ABC. A set

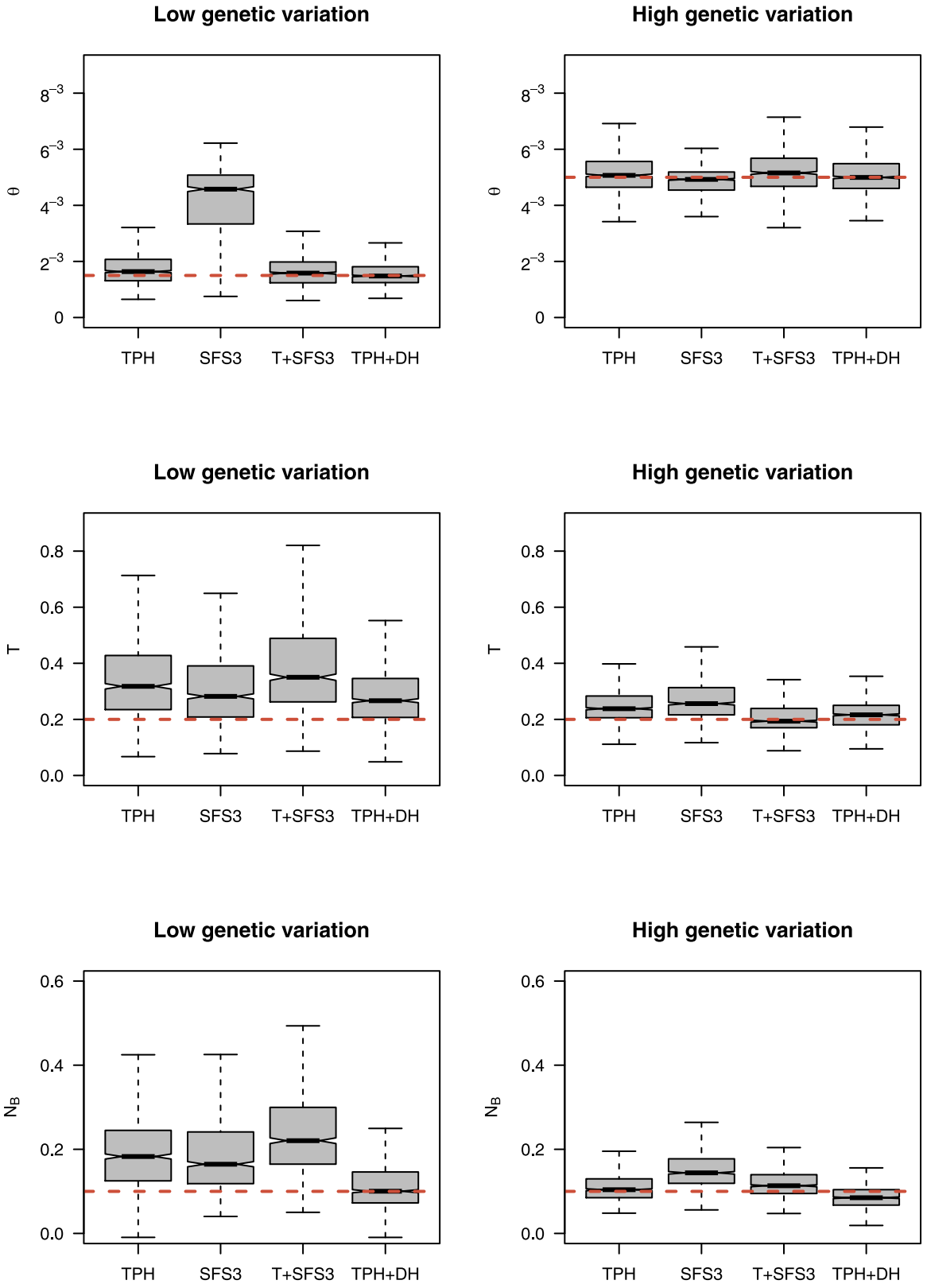

Figure 6. Parameter estimation for different sets of summary statistics. Boxplots showing the distribution of mean values of the 1000 posterior distributions for the replicates for the population scaled mutation rate (θ), the time of the bottleneck (T) and the strength of the bottleneck (N_B) for datasets with low ($\theta = 0.0015$) and high ($\theta = 0.005$) genetic variation. Thick lines denote the median, the boxes extend to the first (25%) and third quartiles (75%) and the whiskers give the minimum and maximum values. The dotted red lines show the true values for the BNM ($n = 20$, $l = 30$).
doi:10.1371/journal.pone.0099581.g006

Table 3. Parameter estimation under a bottleneck model.

	$\theta = 0.0015$			$\theta = 0.005$		
	θ **(0.0015)**	T **(0.2)**	N_B **(0.1)**	θ **(0.005)**	T **(0.2)**	N_B **(0.1)**
TPH	0.00173	0.34206	0.19495	0.00512	0.24841	0.11152
SFS3	0.00413	0.3143	0.18854	0.00475	0.27097	0.15404
T+SFS3	0.00166	0.38112	0.23845	0.00518	0.21601	0.12387
TPH+DH	0.00159	0.28889	0.11915	0.00507	0.21907	0.08665

Mean parameter estimates averaged over all replicates (true values in parentheses). $n = 20$, $l = 30$.
doi:10.1371/journal.pone.0099581.t003

of summary statistics that is informative in distinguishing between two competing models may not be the set of statistics most suitable for estimating parameters under the most probable model. In some cases it may be more suitable to perform model choice using a set of summary statistics known to be informative in separating a class of models, and then to use a second set of summary statistics for estimating parameters known to be informative for the most probable model. Foremost, this emphasizes the importance of performing preliminary analyses of the power afforded by a given set of summary statistics.

Whilst an important aspect to consider, our results suggest that the choice of tolerance is not of overriding importance, although this is dependent on how informative the summary statistics are. This has also been observed in other studies looking at parameter estimation in an ABC framework. [20], for example, found that the tolerance has a relatively minor effect on the estimation of migration rate when they weight summary statistics according to how informative they are. However, [25] noted that, in general a low tolerance was more beneficial for estimating parameters if the number of accepted replicates was sufficiently high. Of more importance is the choice of Bayes factor cutoff. In this study we used a Bayes factor of 3 as the threshold but note that the false positive rate increases with a decreasing Bayes factor cutoff. This suggests that drawing conclusions from analyses where Bayes factors are less than 3 may lead to the inference of an incorrect model.

ABC in non-model organisms

The amount of data available for genetic studies of non-model organisms is often limited, and so it is important that sequencing and sampling efforts are directed towards maximizing the amount of information available for ABC. Specifically, we find that there are a number of factors that govern the power to distinguish between two competing models. In particular, the level of genetic variation is important and this inevitably has consequences on the number of loci and samples required. Here, we considered the power afforded to sequence data, but other types of markers, such as microsatellites, are more variable and would give more information. For our low variation datasets ($\theta = 0.0015$), around 20 individuals and 30 loci would be required in order to have a 95% power to detect a strong bottleneck ($N_B = 0.1N$). In the more variable dataset ($\theta = 0.005$), around 10 individuals and 15 loci would give the same power. In general, and in agreement with expectations of the coalescent [26], we find that sampling more loci rather than individuals is of greater benefit in increasing the power.

It is also important to acknowledge that there is a limit to what one can say with a limited dataset. Our analyses dealt with relatively strong bottlenecks, and these represent quite drastic demographic events. Weaker events will undoubtedly affect genetic data in a subtler way that is harder to detect and therefore requires more data, whether that be more loci, individuals or more variable markers. However, in general we find that the power to detect a bottleneck increases with the severity of the bottleneck. Large samples of loci and individuals are required to detect mild bottlenecks and this is likely to generalize to parameters in other models. Similarly, [27] found that there is lower power to detect a weaker migration rate in an isolation-with-migration model. However, even if bottlenecks with a reduction in the population size of 99% or more can de detected with the correct summary statistics, it is unclear how often bottlenecks of this magnitude occur in natural populations. There are examples (e.g. [9]: $\sim$3% of current size; [28]: $\sim$1.5% of current size), especially in domesticated species (e.g. [29]: $\sim$0.5% of current size), where drastic reductions in the effective population size have been inferred. However, it may be the case that weaker bottlenecks, or more subtle temporal variation in the effective population size, are more frequent but that we simply do not have enough power to detect them with the datasets at hand.

In general, the approach of ABC in summarizing the data into summary statistics is relatively reliable for estimating parameters [8]. Although this seems to be true in most cases, we find that this does not hold in cases where the summary statistics are less informative. This is exemplified by the poor estimate of θ by the SFS3 set of summary statistics. This appears to be resulting from the priors, that are uniformly distributed between 0 and 0.01. When the summary statistics offer no information on the parameters of interest then the expected value of the parameter will be the mean of the prior distribution. In this case, the parameter estimate of θ would approach 0.005 as the summary statistics become less informative, and would explain the poor estimate of θ in low diversity samples (0.00413). A sensitivity analysis might therefore be an important step in determining the influence of the prior over the posterior distribution. This may also be influencing the other parameters. T and N_B appear to be overestimated when the data or summary statistics are insufficient, and so these parameters could tend to the mean of the prior distribution if the dataset or summary statistics are insufficient. There are a number of ways that the estimation of parameters and model choice can be improved. The euclidean distance metric is most commonly used for assessment of the fit of the simulated data to the observed data, but other metrics may provide a better measure. Another common step is to even out the contribution of each of the summary statistics through a normalization step that conforms them to the same standard deviation, and this could lead to improvements in parameter estimation.

To assess the amount of information that the analysis brings, it is strongly advised that the prior and posterior distributions are compared, and this can identify situations where the prior has too

much influence over the analysis. In the present study we have considered two simple models (SNM and BNM) and estimated our power to distinguish those. For both models the data were generated by the models under comparison. In real life, however, we may, for instance, sample individuals from different demes in a structured population. This can have a confounding effect and may lead to the false detection of bottlenecks or effective population size changes [30,31].

Conclusions

Our analysis of the power of ABC model choice in limited datasets suggests that careful consideration of the number of loci and samples is critical when designing a study. Even in scenarios as simplistic as the one examined here, under some conditions, there is simply not enough information contained in the data to confidently separate two distinct models. While ABC in principle allows testing for very complex demographic histories, the amount of information that can be extracted from a given dataset is likely to limit power to make more subtle inferences. However, certain parameters, such as the number of samples, the number of loci and the level of genetic variation, can be used fairly reliably to predict the power of a study to separate different models. What's more, if suitably informative summary statistics are used together with an appropriately large dataset then, in general, ABC model choice is relatively powerful and quite conservative with regard to the false positive rate. Fortunately, with the widespread availability of simulation tools, it is possible to test the probability of detecting a model with a dataset of any given size and level of diversity. Furthermore, the efficiency and flexibility of ABC means that assessing the power of any given dataset is realistic for most studies in non-model organsisms.

Materials and Methods

Demographic models and simulated datasets

We test the power of ABC by comparing two simple population genetic models, each simulated under a coalescent model. The coalescent is a backward in time simulator of a population of gametes that can be subjected to a number of evolutionary forces. The genetic variation in a population of gametes is determined by the mutation rate per generation (μ), and the effective population size (N). The level of genetic variation in a population is then defined as the product of these two parameters: $\theta=4N\mu$. Two coalescent models were considered, the first of which consisted of a population of effective size N (haploid individuals) that remain constant through time (SNM). For the second scenario, a bottleneck model (BNM) was considered where an instantaneous reduction in the population size occurs at 0.2 coalescent time units (τ) in the past (measured in $4N$ generations), and persists for a period of 0.2τ before returning to its original size. For each demographic model, 1000 datasets were simulated whereby the levels of nucleotide diversity (assuming an infinite sites model) and number of samples and loci were varied. For the majority of analyses, we considered a sample size (n) of 10 or 20 individuals (where a population consists of $2N$ gametes), with 15 or 30 loci sequenced (750 bp each in length) and two levels of genetic variation with per base pair scaled mutation rates of $\theta=0.0015$ or $\theta=0.005$ (where $\theta=4N\mu$). For the majority of simulations in the BNM, the relative population size during the bottleneck (N_B) was $0.1N$, although we also varied this parameter ($0.5N$, $0.3N$, $0.2N$, $0.1N$ and $0.01N$) to assess the performance of ABC in detecting bottlenecks of varying severity. The population scaled recombination rate, $\rho=4Nr$, was set to 0.01/bp in each model.

For each of these simulated datasets the mean, standard deviation and 5% and 95% quantiles across loci were calculated for Watterson's θ (θ_W, θ_W^{sd}, θ_W^{05}, θ_W^{95}; [32]), nucleotide diversity (π, π^{sd}, π^{05}, π^{95}; [33]), Fay & Wu's estimate of θ (θ_H, θ_H^{sd}, θ_H^{05}, θ_H^{95}; [34]), haplotype diversity (H_e, H_e^{sd}, H_e^{05}, H_e^{95}; [35]), Tajima's D (D, D^{sd}, D^{05}, D^{95}; [36]), Fay & Wu's non-standardized H (H, H^{sd}, H^{05}, H^{95}; [34]) and the number of segregating sites (S, S^{sd}, S^{05}, S^{95}). The relative site frequency spectrum (SFS) was also summarized by the average proportion of segregating sites that occur within each of three or five evenly sized frequency classes (s1, s2, s3, s4, s5). These represent population genetic statistics that are thought to summarize population sequence data in the most informative way (see for example, [17,25,26]). For the analysis of the power and false positive rate of ABC we combined a number of summary statistics: TPH (θ_W, π, H_e), SFS3 (folded site frequency spectrum in 3 bins), T+SFS3 (θ_W, folded site frequency spectrum in 3 bins), SFS5 (folded site frequency spectrum in 5 bins) and TPH+DH (θ_W, π, H_e, D, H). When analyzing the relationships among summary statistics, the calculation of correlation coefficients and the performance of Principal Component Analysis (PCA) was implemented using the python library NumPy.

Model choice, power and parameter estimation

Joint posterior densities were simulated for each of the two ABC models using 10^6 draws from uniformly distributed priors. The prior bounds for the SNM and BNM were (0, 0.01) and (0, 0.02) for θ and ρ, respectively. Time was measured on a scale of $4N$ generation. For the BNM, the time (T) of the bottleneck (looking backwards in time) was sampled from a prior with bounds (0, 1.5), with the relative population size during the bottleneck (N_B) having bounds of (0, 1). The relative ancestral population size and the duration of the bottleneck were fixed as 1 and 0.2τ respectively. ABC was performed using the python library EggLib [37], which implements the Euclidean distance-based, local linear regression method described in [1]. Model choice was performed using the rejection-based method implemented in EggLib, with model probabilities being defined as the proportion of simulations belonging to each model after the ABC rejection step. Bayes factors were calculated as the ratio of the model probabilities, with a Bayes factor $\geq$ 3 (unless stated otherwise) being considered an acceptable level of significance [38]. Power (Ψ) was defined as the probability of correctly rejecting the SNM and was assessed by calculating the proportion of replicates with a Bayes factor ≥ 3 when the true model was the BNM. False positives were considered as instances where the SNM was falsely rejected and was given by the proportion of replicates with Bayes factors ≥ 3 when the SNM was the correct model. Parameter estimation was carried out using the local linear regression method of [1]. The accuracy of parameter estimation was assessed by comparing the true parameter value with that estimated in ABC using the relative bias, $(\hat{x}-x)/x$, and the relative mean square error, $(\hat{x}-x)^2/x^2$. The tolerance level (ϵ) for both model choice and parameter estimation was fixed at 0.001 unless otherwise stated. Coalescent simulations, ABC analyses and calculation of summary statistics were performed using EggLib. Any additional custom code is provided in the Github repository: https://github.com/mspopgen/Stocks2014a.

Supporting Information

Figure S1 Principal Component Analysis under the SNM and BNM models. The first four principal components (PCs) for summary statistics calculated under the SNM and a BNM ($N_B = N$, $N_B = 0.1N$) and the entire prior parameter space of the BNM. $n = 20$, $l = 30$.
(PDF)

Figure S2 Model probability distributions for different summary statistics. Distribution of model probabilities for the TPH (θ_W, π, H_e), SFS5 (5 bin relative site frequency spectrum) and TPH+DH (θ_W, π, H_e, D, H) sets of summary statistics. $\theta = 0.0015$, $N_B = 0.1N$.
(PDF)

Figure S3 Impact of bottleneck severity on Tajima's D. The effect of bottleneck strength on the value of Tajima's D for both small ($n = 10$, $l = 15$) and large ($n = 20$, $l = 30$) datasets with low ($\theta = 0.0015$) and high ($\theta = 0.005$) genetic variation. Each point represents the rejection step of an ABC analysis when the TPH+DH set of statistics is used with a tolerance of 0.001. The effective population size during the bottleneck (N^B) is defined relative to the recovered effective population size (N).
(PDF)

Figure S4 Impact of bottleneck severity on Fay and Wu's H. The effect of bottleneck strength on the value of Fay & Wu's H and model probability for both small ($n = 10$, $l = 15$) and large ($n = 20$, $l = 30$) datasets with low ($\theta = 0.0015$) and high ($\theta = 0.005$) genetic variation. Each point represents the rejection step of an ABC analysis when the TPH+DH set of statistics is used with a tolerance of 0.001. The effective population size during the bottleneck (N^B) is defined relative to the recovered effective population size (N).
(PDF)

Figure S5 Impact of tolerance. The effect of the tolerance level on model comparison in ABC for datasets with different numbers of samples (n), loci (l) and levels of genetic variation. Different colored lines refer to different sets of summary statistics: TPH and SFS5. $N_B = 0.1N$.
(PDF)

Table S1 Impact of tolerance. The power (Ψ) and false positive rate (α) for different tolerances and sets of summary statistics. $N_B = 0.1N$.
(PDF)

Table S2 Parameter estimates. Proportion of replicates where the true value lies within the 10%, 50% and 90% confidence intervals of the posterior distribution. $N_B = 0.1N$, $n = 20$, $l = 30$.
(PDF)

Acknowledgments

The authors would like to thank two anonymous reviewers for their comments on the manuscript.

Author Contributions

Conceived and designed the experiments: M. Stocks M. Siol ML SD. Performed the experiments: M. Stocks SD. Analyzed the data: M. Stocks M. Siol SD. Wrote the paper: M. Stocks M. Siol ML SD.

References

1. Beaumont MA, Zhang W, Balding DJ (2002) Approximate bayesian computation in population genetics. Genetics 162: 2025–2035.
2. Beaumont MA (2010) Approximate bayesian computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics 41: 379–406.
3. Siol M, Wright SI, Barrett SC (2010) The population genomics of plant adaptation. New Phytologist 188: 313–332.
4. Li J, Li H, Jakobsson M, Li S, Sjödin P, et al. (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Molecular Ecology 21: 28–44.
5. Joyce P, Marjoram P (2008) Approximately suffcient statistics and bayesian computation. Statistical Applications in Genetics and Molecular Biology 7: 1–16.
6. Fearnhead P, Prangle D (2012) Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74: 419–474.
7. Burr T, Skurikhin A (2013) Selecting summary statistics in approximate bayesian computation for calibrating stochastic models. BioMed research international 2013.
8. Robert CP, Cornuet JM, Marin JM, Pillai NS (2011) Lack of confidence in approximate Bayesian computation model choice. Proceedings of the National Academy of Sciences of the United States of America 108: 15112–15117.
9. Thornton K, Andolfatto P (2006) Approximate bayesian inference reveals evidence for a recent, severe bottleneck in a netherlands population of drosophila melanogaster. Genetics 172: 1607–1619.
10. Ness RW,Wright SI, Barrett SC (2010) Mating-system variation, demographic history and patterns of nucleotide diversity in the tristylous plant eichhornia paniculata. Genetics 184: 381–392.
11. Chan YL, Anderson CN, Hadly EA (2006) Bayesian estimation of the timing and severity of a population bottleneck from ancient dna. PLoS Genetics 2: e59.
12. Heuertz M, De Paoli E, Källman T, Larsson H, Jurman I, et al. (2006) Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of norway spruce [*Picea abies* (l.) karst]. Genetics 174: 2095–2105.
13. Gattepaille L, Jakobsson M, Blum MG (2013) Inferring population size changes with sequence and snp data: lessons from human bottlenecks. Heredity 110: 409–419.
14. Blum MG, Nunes M, Prangle D, Sisson S, et al. (2013) A comparative review of dimension reduction methods in approximate bayesian computation. Statistical Science 28: 189–208.
15. Nunes MA, Balding DJ (2010) On optimal selection of summary statistics for approximate bayesian computation. Statistical applications in genetics and molecular biology 9.
16. Hickerson MJ, Dolman G, Moritz C (2006) Comparative phylogepgraphic summary statistics for testing simultaneous vicariance. Molecular Ecology 15: 209–223.
17. Clotault J, Thuillet AC, Buiron M, De Mita S, Couderc M, et al. (2012) Evolutionary history of pearl millet (*Pennisetum glaucum* [L.] R. Br.) and selection on owering genes since its domestication. Molecular Biology & Evolution 29: 1199–1212.
18. Tajima F (1989) The effect of change in population size on DNA polymorphism. Genetics 123: 597–601.
19. Estoup A, Lombaert E, Marin JM, Guillemaud T, Pudlo P, et al. (2012) Estimation of demo-genetic model probabilities with approximate bayesian computation using linear discriminant analysis on summary statistics. Molecular ecology resources 12: 846–855.
20. Hamilton G, Currat M, Ray N, Heckel G, Beaumont M, et al. (2005) Bayesian estimation of recent migration rates after a spatial expansion. Genetics 170: 409–417.
21. Wegmann D, Leuenberger C, Excoffier L (2009) Efficient Approximate Bayesian Computation coupled with Markov Chain Monte Carlo without likelihood. Genetics 182: 1207–1218.
22. Aeschbacher S, Beaumont MA, Futschik A (2012) A novel approach for choosing summary statistics in approximate bayesian computation. Genetics 192: 1027–1047.
23. Prangle D, Fearnhead P, Cox MP, Biggs PJ, French NP (2014) Semi-automatic selection of summary statistics for abc model choice. Statistical applications in genetics and molecular biology 13: 67–82.
24. Marin JM, Pillai NS, Robert CP, Rousseau J (2013) Relevant statistics for bayesian model choice. Journal of the Royal Statistical Society: Series B (Statistical Methodology).
25. Li S, Jakobsson M (2012) Estimating demographic parameters from large-scale population genomic data using Approximate Bayesian Computation. BMC Genetics 13: 22.
26. Wakeley J (2009) Coalescent theory: an introduction, volume 1. Roberts & Company Publishers.
27. Huang W, Takebayashi N, Qi Y, Hickerson MJ (2011) MTML-msBayes: Approximate Bayesian comparative phylogeographic inference from multiple taxa and multiple loci with rate heterogeneity. BMC Bioinformatics 12: 1.

28. Bodare S, Stocks M, Yang J, Lascoux M (2013) Origin and demographic history of the endemic taiwanese spruce (*Picea morrisonicola*). Ecology and Evolution 3: 3320–3333.
29. Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fiedel-Alon A, et al. (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genetics 3: e163.
30. Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA (2010) The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. Genetics 186: 983–995.
31. Peter BM, Wegmann D, Excoffier L (2010) Distinguishing between population bottleneck and population subdivision by a bayesian model choice procedure. Molecular Ecology 19: 4648–4660.
32. Watterson GA (1975) On the number of segregating sites in the genetical models without recombination. Theoret Pop Biol 7: 256–276.
33. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.
34. Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.
35. Nei M (1987) Molecular evolutionary genetics. New York: Columbia University Press.
36. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
37. De Mita S, Siol M (2012) Egglib: processing, analysis and simulation tools for population genetics and genomics. BMC Genetics 13: 27.
38. Jeffreys H (1961) Theory of probability. Oxford: Oxford Univ. Press.