MOLECULAR ECOLOGY

Molecular Ecology (2011) **20**, 2709–2723 doi: 10.1111/j.1365-294X.2011.05131.x

Jaatha: a fast composite-likelihood approach to estimate demographic parameters

LISHA NADUVILEZHATH, LAURA E. ROSE and DIRK METZLER
LMU Biocenter, Department Biology II, Grosshadernerstrasse 2, 82152 Planegg, Germany

Abstract

While information about a species' demography is interesting in its own right, it is an absolute necessity for certain types of population genetic analyses. The most widely used methods to infer a species' demographic history do not take intralocus recombination or recent divergence into account, and some methods take several weeks to converge. Here, we present Jaatha, a new composite-likelihood method that does incorporate recent divergence and is also applicable when intralocus recombination rates are high. This new method estimates four demographic parameters. The accuracy of Jaatha is comparable to that of other currently available methods, although it is superior under certain conditions, especially when divergence is very recent. As a proof of concept, we apply this new method to estimate demographic parameters for two closely related wild tomato species, *Solanum chilense* and *S. peruvianum*. Our results indicate that these species likely diverged 1.44·N generations ago, where N is the effective population size of *S. chilense*, and that some introgression between these species continued after the divergence process initiated. Furthermore, *S. peruvianum* likely experienced a population expansion following speciation.

Keywords: composite-likelihood method, demography, recent divergence, wild tomatoes (*Solanum chilense*, *S. peruvianum*)

Received 29 October 2010; accepted 18 April 2011

Introduction

The availability of more and more affordable genome technologies has allowed scientists to venture outwards from the classical model systems and to begin answering questions about evolutionary genetics and trait evolution in nonmodel systems. A first step in many of these evolutionary studies is the description of a species' demography. This is important because some demographic effects can leave similar signatures in the genome as natural selection (Robertson 1975; Andolfatto & Przeworski 2000; Teshima *et al.* 2006).

Here, we focus on the inference of historical demography of two closely related populations or species from neutral loci. We assume that the two populations recently split from a single ancestral population. For this situation, Nielsen, Wakeley and Hey have developed Bayesian MCMC methods to infer parameters including the time since the population split and the migration rates between the populations (Nielsen & Wakeley 2001; Hey & Nielsen 2004). For the case in which no population size change is incorporated, Hey & Nielsen (2007) derived an analytical result that makes the MCMC procedure more efficient. Hey (2010) extended this method to account for up to 10 related populations. Implementations of these methods are available in Jody Hey's programs IM, IMa and IMa2. One limitation of these programs is that they do not allow for intralocus recombination. The robustness of IMa against moderate violations of this and other assumptions was examined in a recent simulation study (Strasburg & Rieseberg 2010). The authors found, for example, that even recombination rates as low as 0.005 per bp per $4N_e$ generations could result in N_e 90% highest point density (HPD) intervals that did not contain the true value used in the simulation (3 of 10 cases). The HPD intervals never included the true value when recombination rates were above 0.02, because recombination events were considered to be mutations.

Correspondence: Lisha Naduvilezhath, Fax: +49 (0)89 2180 74104; E-mail: Lisha@bio.lmu.de

Estimates of divergence time were also biased upwards as recombination increased. Strasburg & Rieseberg (2010) also tested a pragmatic approach, in which they divided the loci into apparently nonrecombining blocks. These blocks were then treated as if they were independent loci and analysed with IMa. This approach is not well reasoned from a theoretical perspective, but in the simulation study of Strasburg & Rieseberg (2010), it removed much of the bias for most parameters.

The software LAMARC (Kuhner 2006) incorporates intralocus recombination using an MCMC method to estimate population genetic parameters in a Bayesian, as well as in a maximum-likelihood framework. Because it assumes a constant population structure, this method is inappropriate for analysing data from two populations that have recently split from a joint ancestral population (Kuhner 2006). To analyse data sets with a high amount of intralocus recombination from recently diverged species, Becquet & Przeworski (2007) introduced an MCMC method (MIMAR) that is based on four summary statistics, similar to those described in Wakeley & Hey (1997). This is in contrast to LAMARC and IM/IMa/IMa2, which employ the likelihood or posterior probability given the full set of sequence data. The major drawback of all of these methods is their rather long run-times that require several weeks to converge.

Gutenkunst *et al.* (2009) implemented a promising diffusion approximation in ∂a∂i, which is considerably faster than the methods described earlier and can be used for various demographic scenarios of up to three populations. In this composite-likelihood method, which assumes unlinked SNPs (see also Kim & Stephan 2000; Hudson 2001; McVean *et al.* 2002), the data are summarized with the full joint site frequency spectrum (JSFS). The JSFS is a matrix of integers ($a_{i,j}$), where $a_{i,j}$ is the number of polymorphic sites where the derived nucleotide type is observed in i sequences of those sampled from species 1 and in j sequences sampled in species 2. The four summary statistics of Wakeley & Hey (1997) can be computed from the JSFS: fixed differences between species, shared polymorphisms, differences that are only polymorphic in species 1, and those that are only polymorphic in species 2. Li & Stephan (2006) showed that it is worthwhile to use more information from the JSFS than these four summary statistics for inference of demographic histories using population genetic data. Other JSFS-based sets of summary statistics are examined by Tellier *et al.* (2011), with the main conclusion that especially further division of the shared polymorphisms results in better estimations of divergence times and migration rates. Garrigan (2009) combines the maximum-likelihood method of Li & Stephan (2006) with a composite-likelihood approach and turns it into a Bayesian (MC)MCMC sampling method to estimate the ratios of population sizes, timing of size changes and population splits. Garrigan (2009) reports a typical run-time of his method of several days for a data set. Li & Stephan (2006) and Garrigan (2009) assume that there is no migration between populations following the split.

Here, we introduce the method Jaatha (abbreviation for '*J*SFS *a*ssociated *a*pproximation of *th*e *a*ncestry', also the Malayalam word for 'past'), which uses JSFS-based summary statistics in a composite-likelihood approach. We perform simulation studies to assess the estimation accuracy of Jaatha for three different demographic models on three different data sets each. Because of the fast run-time and great flexibility of the underlying demographic cases, we chose ∂a∂i for comparing the results with our program. To compare Jaatha with the full-likelihood method IM, we applied the programs to simulated data sets without intralocus recombination.

We apply our new method to estimate demographic parameters based on DNA sequence data from two closely related wild tomato species, *Solanum chilense* and *S. peruvianum*. These species are endemic to the western coast of South America and are closely related to the cultivated tomato. *S. peruvianum* is widespread and often occurs in large stands in central and southern Peru and northern Chile [reviewed in Chetelat *et al.* (2009)]. *S. chilense* has a more restricted range, occurring in northern Chile and southern Peru, and is adapted to exceptionally dry habitats (Chetelat *et al.* 2009). Previous studies support a very recent divergence time between these species with population growth in *S. peruvianum* (Städler *et al.* 2008). Although the 'isolation' model of speciation (Wakeley & Hey 1997) could not be rejected, Städler *et al.* (2008) found some evidence for postdivergence introgression using the LD-based method of Machado *et al.* (2002). Because of the recency of divergence and high amount of within-locus recombination in these species, this data set serves as an appropriate test case for our method.

Methods and models

Demographic models

We assume that autosomal DNA sequences of diploid organisms are sampled from two populations P_1 and P_2 having current effective population sizes N_1 and N_2, respectively. P_1 and P_2 originated $\tau \cdot 4N_1$ generations ago from a common ancestral population P_A of effective size N_A (Wakeley & Hey 1997). Immediately following the split, the effective population size of P_2 was $N_A - N_1$. We denote the mutation rate per locus and per generation by μ and define $\theta_i = 4N_i\mu$ for $i \in \{1,2,A\}$. P_2 may undergo exponential population growth at rate g or

shrinkage (when $g < 0$), whereas P_1 and P_A remain constant in size. We allow for ongoing symmetric migration between P_1 and P_2. Following Hudson (2002), the migration rate m is scaled with $4N_1$. In other words, in each generation, $\frac{m}{4N_1} \cdot N_1 = m/4$ individuals of P_1 and $\frac{m}{4N_1} \cdot N_2$ of P_2 are replaced by migrants from the other population. Assuming the infinite sites model for sequence evolution, Jaatha estimates θ_1 and three additional parameters.

In our simulation studies described below, we assess the accuracy of Jaatha's estimations for the parameters θ_1, the population size ratio $q = \frac{N_2}{N_1} = \frac{\theta_2}{\theta_1}$, the divergence time τ and the migration rate m. The simulations are based on the following three variants of the demographic model (Fig. 1):

Constant Model. The size of population P_2 remains constant following the split, and $\theta_A = \theta_1 + \theta_2$.

Growth Model. The ancestral population splits into two populations of equal size. Thus, $\theta_1 = \frac{1}{2} \cdot \theta_A$ and $\theta_2 = \frac{1}{2}\theta_A \cdot e^{\tau g}$.

Fraction-Growth Model. Immediately following the split, population P_1 is twenty times as large as population P_2. Thus, $\theta_1 = \frac{20}{21} \cdot \theta_A$ and $\theta_2 = \frac{1}{21}\theta_A \cdot e^{\tau g}$. The ms commands (Hudson 2002) to simulate data according to these models are included in the supplementary information.

We consider two additional models for the application to the wild tomato species, *S. chilense* and *S. peruvianum* (Fig. 2). For these two models, we include the initial size ratio s of P_2 and P_1 after the split as an additional parameter. As the current version of Jaatha is restricted to estimating four parameters including θ_1, we had to set one of the remaining parameters to a fixed value. In one case, we set the migration rate to zero (*noMig Model*) and in the other, we set τ to 0.36 (*fixedTau Model*). This is the estimate of τ from the analyses using the *Growth Model*. Changing this value to 0.40 had negligible effect on parameter estimation or the fit of the model to the tomato data (data not shown).

Estimating demographic parameters with Jaatha

The aim of Jaatha is to estimate demographic parameters from SNP data for which ancestral and derived states can be distinguished. Jaatha consists of two phases: a training phase and an estimation phase. In the training phase, Jaatha uses simulated data to learn how the expectation values for 23 summary statistics

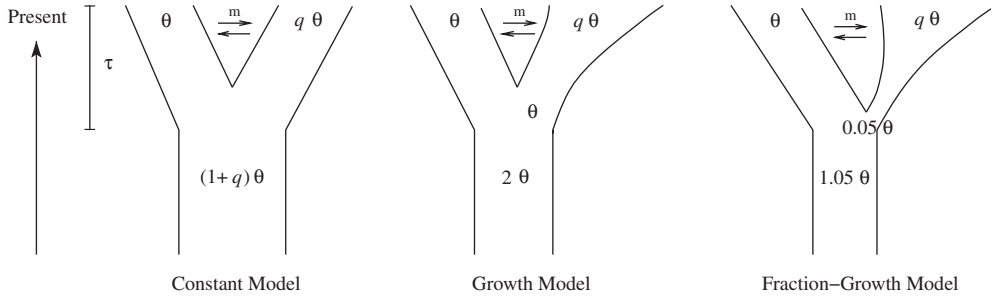

Fig. 1 The different demographic models for populations P_1 and P_2 used for the simulation study where θ = population mutation parameter for P_1, m = migration rate scaled by $4N_1$ generations, q = size ratio between P_2 and P_1 and τ = divergence time measured in $4N_1$ generations.

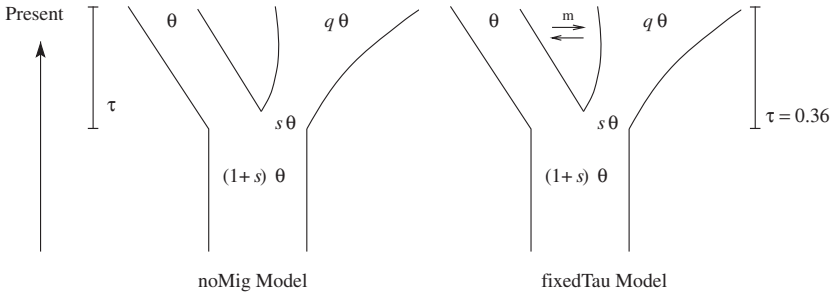

Fig. 2 Additional models applied to the tomato data, where s = the initial size ratio of P_2 and P_1 immediately after the split. The other parameters are defined as in Fig. 1.

$S = (S_1,\ldots,S_{23})$ depend on the model parameters. In the estimation phase, we follow a composite-likelihood approach. That is, we apply maximum-likelihood parameter estimation in a model in which the observed values of $S_1,\ldots,S_{23}$ are independently Poisson distributed. As parameters for the Poisson distributions, we use the results of the training phase. The Poisson approximation corresponds to treating all SNPs as if they were independent. Consequently, sequences from different genomic regions of the same individual can be concatenated before proceeding with Jaatha.

The run-time for the estimation phase of Jaatha is ≤ 15 s. The training phase takes up to 5 days on a modern desktop PC, using a single processor kernel. If more processors kernels are available, it is straightforward to parallelize the training phase. The results of the training phase can be reused for data sets with similar parameter ranges and sample sizes. This is especially advantageous when simulation studies or bootstrap methods are applied to assess estimation accuracy (Efron & Tibshirani 1993).

Joint site frequency spectrum and summary statistics. Our 23 summary statistics $S = (S_1,\ldots,S_{23})$ form a coarsening of the joint site frequency spectrum (JSFS), which is defined as follows: Let m and n be the numbers of sequences sampled from P_1 and P_2, and $A = \{0,\ldots,m\} \times \{0,\ldots,n\}\backslash\{(0,0), (m,n)\}$. The JSFS assigns to each $(a,b) \in A$ the number of polymorphisms $J_{a,b}$ for which the derived state at this position is observed in exactly a sequences sampled from P_1 and b sequences sampled from P_2. We partition A into 23 disjoint subsets $A_1,\ldots,A_{23}$ as shown in Fig. 3 and define each summary statistic S_i by summing up the JSFS within A_i: $S_i = \sum_{(a,b) \in A_i} J_{a,b}$. Other summations of the JSFS are also possible and are compared by Tellier *et al.* (2011).

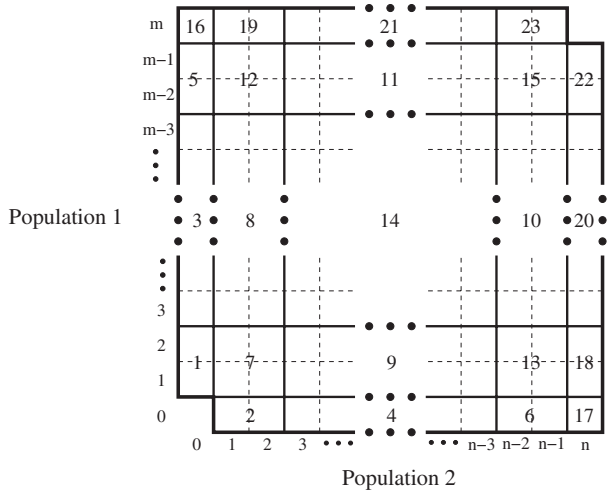

Fig. 3 Partition of domain of the joint site frequency spectrum (JSFS) for two populations where m and n denote the number of sampled alleles per locus of each population. Entries of the JSFS are summed up to result in 23 summary statistics.

Training phase. We use the parameter space of the *Growth Model* as an example to describe the training phase. Let y be the numbers of polymorphisms observed in the data and y' the number of polymorphisms in a simulation with parameter values θ_1', τ', m' and q'. For fixed values τ', m' and q', we estimate θ_1 by $\theta_1' \cdot y/y'$. Thus, we separate the estimation of θ_1 from the estimation of the other parameters. Jaatha generates training data for each parameter combination on a $40 \times 40 \times 40$ grid in the parameter space $\mathscr{P} = [\tau_{\min}, \tau_{\max}] \times [m_{\min}, m_{\max}] \times [q_{\min}, q_{\max}]$. For a higher resolution in the lower parameter ranges, the grid is uniform on the log-scaled parameter space. The log transformation is given by

$$\begin{aligned} d : \mathscr{P} &\to [1,40] \times [1,40] \times [1,40] \\ (\tau, m, q) &\mapsto (d_\tau, d_m, d_q) = (\log_{z_\tau}(\tau/\tau_{\max}) \\ &\quad + 1, \log_{z_m}(m/m_{\max}) + 1, \log_{z_q}(q/q_{\max}) + 1), \end{aligned}$$

where $z_p = \sqrt[39]{\frac{p_{\min}}{p_{\max}}}$ for each parameter $p \in \{\tau,m,q\}$. The inverse transformations are given by $p = p_{\max} \cdot z_p^{d_p - 1}$. The grid consists of all integer triples $(d_\tau,d_m,d_q) \in \{1,2,\ldots,40\}^3 \subset [1,40]^3$ in the log-scaled parameter space. For each of the 64,000 parameter combinations (τ,m,q) corresponding to grid points, Jaatha calls the program ms (Hudson 2002) to simulate 10 independent data sets with 7 loci (1 kb long) and $\theta_1 =$ 5 per locus. The recombination rate is set to 20 with 1000 possible recombination points per locus. Increasing the recombination rate would make the method more precise but would also result in longer run-times of ms.

To fit log-linear generalized linear models (GLMs) of type Poisson to the summary statistics, we divide the log-scaled parameter space into bins. In each dimension, the range [1, 40] is divided into eight intervals [1, 5.5], (5.5, 10.5], (10.5, 15.5], …, (35.5, 40], where $(a, b]$ denotes the interval $\{x: a < x \le b\}$. We chose these grid and bin sizes because they provide a reasonable compromise between accuracy and run-time but they can be changed by the user. Each of the $8^3 = 512$ bins contains 125 ($=5^3$) grid points. For each bin and for each of the 23 summary statistics S_i, we fit a Poisson GLM to the simulated data to estimate how S_i depends on d_τ, d_m and d_q within the range of this bin. For any bin $(a_\tau, b_\tau] \times (a_m,b_m] \times (a_q,b_q]$, we take simulated data from grid points in the range $(a_\tau - 3, b_\tau + 3] \times (a_m - 3, b_m + 3] \times (a_q - 3, b_q + 3]$ into account, whereas in the fitting procedure, we give lower weights to the points outside the bin. This leads to 512 ($=8^3$) parameter combinations at the edges of the parameter space and up to 1331 ($=11^3$)

in the interior. Grid points in the bin are weighted with 1. For the other grid points, the weight is halved for each d_P that lays outside the range $(a_P,b_P]$, such that we obtain four different weights 1, $\frac{1}{2}$, $\frac{1}{4}$ and $\frac{1}{8}$.

The Poisson GLM fits coefficients $\beta_{0,i}$, $\beta_{\tau,i}$, $\beta_{m,i}$ and $\beta_{q,i}$ to the simulated data from the training phase such that

$$\widehat{\beta}_{0,i} + \widehat{\beta}_{\tau,i} \cdot d_\tau + \widehat{\beta}_{m,i} \cdot d_m + \widehat{\beta}_{q,i} \cdot d_q = \ln(\lambda_i),$$

where λ_i is the expected value of S_i, which is assumed to be Poisson distributed. The dependence of λ_i on the original parameters τ, m and q takes the form

$$\lambda_i = \alpha_{0,i} \cdot \tau^{\alpha_{\tau,i}} \cdot m^{\alpha_{m,i}} \cdot q^{\alpha_{q,i}}$$

within each block, where $\alpha_{P,i}$ equals $\beta_{P,i}$ up to a constant factor. Jaatha calls the R function glm() to fit the weighted Poisson GLMs (R Development Core Team 2009).

Estimation phase. For the estimation of θ let $s_1,\ldots,s_{23}$ be the values of the 23 summary statistics observed in the given data set, b be a bin in the log-scaled parameter space, and let $s_1^{(b)},\ldots,s_{23}^{(b)}$ be the Poisson GLM predictions for the summary statistics in the centre of b. One simulated data set of the training phase consists of 7 loci with $\theta_1 = 5$ per locus, so we estimate θ_1 for bin b by

$$\widehat{\theta_b} = \frac{\sum_{i=0}^{23} s_i}{\sum_{j=0}^{23} s_j^{(b)}/35},$$

i.e. Jaatha will always return estimates $(\widehat{\tau},\widehat{m},\widehat{q})$ together with $\widehat{\theta_b}$, where b is the bin that contains $(d_{\widehat{\tau}}, d_{\widehat{m}}, d_{\widehat{q}})$.

The composite-likelihood of a parameter combination (τ,m,q) is the probability that the summary statistics $S_1,\ldots,S_{23}$ take the observed values $s_1,\ldots,s_{23}$, assuming the Poisson model with the parameter values τ, m, q and $\theta = \widehat{\theta_b}$, where $(d_{\widehat{\tau}}, d_{\widehat{m}}, d_{\widehat{q}}) \in b$. In the Poisson model, all sites are assumed to be independent, i.e. unlinked. This corresponds to the heuristic of taking an infinite sites model to the limit of high recombination rates. Thus, S_i is an independent Poisson random variable, and the probability that it takes the values s_i is

$$\Pr(S_1 = s_1, \ldots, S_{23} = s_{23}) = \prod_{i=1}^{23} \frac{\lambda_i^{s_i} \cdot e^{-\lambda_i}}{s_i!},$$

where $\lambda_1 = \mathbb{E}S_1, \ldots, \lambda_{23} = \mathbb{E}S_{23}$ are the expectation values of the summary statistics $S_1,\ldots,S_{23}$. The main idea behind Jaatha is to estimate how $\lambda_1,\ldots,\lambda_{23}$ depend upon τ, m and q and then to maximize the resulting approximate composite-likelihood function

$$L_{s_1,\ldots,s_{23}}(\tau,m,q) \approx \prod_{i=1}^{23} \frac{\widehat{\lambda_i}(\tau,m,q)^{s_i} \cdot e^{-\widehat{\lambda_i}(\tau,m,q)}}{s_i!}. \quad (1)$$

Here, $\widehat{\lambda_i}(\tau,m,q)$ is our estimation for $\mathbb{E}S_i$ in terms of τ, m, q and implicitly the corresponding $\widehat{\theta_b}$. The use of the simple estimator $\widehat{\theta_b}$ saves us one dimension in the optimization procedure, at the cost of some amount of accuracy. As the estimator $\widehat{\theta_b}$ is mainly based on the total number of polymorphisms, using $\widehat{\theta_b}$ in the estimation of the other parameters may have a similar effect as conditioning on the total number of polymorphisms. This suggests that replacing the Poisson-distribution weights in the approximation (eqn 1) by multinomial-distribution weights (as proposed by an anonymous reviewer) may lead to improvements in the approximation accuracy (Sawyer & Hartl 1992; Adams & Hudson 2004). To test this, we have implemented a version of Jaatha, in which we replace approximation (eqn 1) by

$$L_{s_1,\ldots,s_{23}}(\tau,m,q) = \binom{\sum_j s_j}{s_1,\ldots,s_{23}} \cdot \prod_{i=1}^{23} \left(\frac{\lambda_i}{\sum_j \lambda_j}\right)^{s_i}. \quad (2)$$

Jaatha optimizes $L_{s_1,\ldots,s_{23}}(\tau,m,q)$ (or, more precisely, its approximation using formula (1 or 2) within each bin using the *optim* function of R and the optimization procedure of Byrd *et al.* (1995), using the bin centres as starting points. Unless otherwise noted, we use the default Jaatha version with approximation (eqn 1).

Our implementation of Jaatha in R (R Development Core Team 2009) provides three additional variants of the optimization procedure, which combine the procedures J_1, J_2 and J_4 described by Tellier *et al.* (2011) with the estimation of θ described earlier. The R script is freely available from the website http://evol.bio.lmu.de/_statgen/software/jaatha/.

Comparison of Jaatha, IM and ∂a∂i

We compare the accuracy of parameter estimations for θ, τ, m and q by IM (Hey & Nielsen 2004), *∂a∂i* version 1.3.4 (Gutenkunst *et al.* 2009) and a version of Jaatha that uses approximation (eqn 1). A simulation study to compare a variant of Jaatha with MIMAR (Becquet & Przeworski 2007) and PopABC (Lopes *et al.* 2009) has been performed by Tellier *et al.* (2011). We applied the three programs Jaatha, IM and *∂a∂i* to data sets that we simulated with Hudson's ms software for three different demographic models, each with three scenarios described in the following. These scenarios differ in their number of loci, type of migration (asymmetric or symmetric) and amount of recombination. For each scenario and population, 100 data sets were simulated with 25 sequences sampled from each population. The

parameter ranges and the underlying demographic models were as described elsewhere (for ms commands as well as parameter ranges, see Supporting Information).

`7-loci scenario` 100 data sets were simulated with seven loci, asymmetric migration between populations and a within-locus recombination rate ρ chosen randomly between 5 and 20 per locus (0.005–0.02/bp) per $4N_1$ generations where N_1 is the effective population size of P_1, i.e. *S. chilense*.

`100-loci scenario` 100 data sets were simulated with 100 loci, symmetric migration between populations and no within-locus recombination.

`1000-loci scenario` 100 data sets were simulated with 1000 loci, symmetric migration between populations and a within-locus recombination rate chosen randomly between 5 and 20 per locus.

Because IM was designed for data without intralocus recombination, we applied it to the data from the `100-loci scenario` only and reported the HiPt value. To convert the ms outputs to IM inputs, we replaced '0' (ancestral state) with 'A' and '1' (derived state) with 'T'. (Note that θ is defined per locus such that the actual length of the locus does not play a role.) For each IM run, we used one chain without heating. The number of burn-in steps was set to 100 000. As IM has a high demand for computer run-time, this simulation study was limited to 10 data sets per demographic model. We restricted the run-time to five weeks per IM run. To assess convergence we performed two independent runs with different random seeds for each of the 10 data sets.

∂a∂i was run on all three demographic models and all three simulation scenarios. The underlying demographic models and parameter ranges (except for θ) were precisely specified for *∂a∂i* analyses. Note that this is not possible for IM; there, we may neither specify the parameter ranges precisely nor that while P_1 is constant in size, P_2 is not. Parameter estimates that fell outside the ranges were set to the closest value within the range for each method.

Application to tomato data

For the two wild tomato species *S. peruvianum* and *S. chilense*, sequences of 7 loci between 0.8 to 1.9 kb in size were available (Städler *et al.* 2008). Following Städler *et al.* (2008), who found evidence for population expansion only in *S. peruvianum*, we limited the analysis to models with growth in one species. Because this method requires one or more outgroups so that mutations can be classified as either ancestral or derived, we chose *S. ochranthum* and *S. lycopersicoides* as outgroups. We classified a nucleotide as derived when it was different from the outgroup. We followed this rule also for positions with multiple hits. In the tomato loci, 7.34% of the polymorphic sites show three or four different nucleotides across the sampled sequences including the outgroup sequences, and therefore, two or more mutational events must have occurred at these sites. For the simulations in Jaatha's training phase, we sampled 45 sequences per species, matching the average number of samples available in the tomato data set. We fit all five models specified previously to the tomato data and compared the Poisson-model maximum-likelihood (ML) values for the models.

Confidence intervals

To assess the uncertainty of the parameter estimates for the tomato data, we used a parametric bootstrap approach to calculate confidence intervals. For each combination of model and estimation method, we simulated 1000 bootstrap replicates using the respective ML estimates. Each replicate, simulated using the ms program (Hudson 2002), contained 7 loci from 45 samples per population. A normal approximation of the log-transformed bootstrap results was used to derive the bias-corrected intervals $\left[\exp\left(2\cdot\widehat{\phi} - \overline{\phi^*} - 1.96\cdot\sigma(\phi^*)\right), \exp\left(2\cdot\widehat{\phi} - \overline{\phi^*} + 1.96\cdot\sigma(\phi^*\right)\right]$, where $\widehat{\phi}$ is our estimate of the log-scaled parameter ϕ, $\overline{\phi^*}$ is the mean and $\sigma(\phi^*)$ is the standard deviation of the bootstrap results (Efron & Tibshirani 1993). Additionally, we computed bootstrap confidence intervals with BCa correction as described by DiCiccio & Efron (1996). The correction was applied on the logarithmic scale.

The choice of recombination rate used in the bootstrap simulations may affect the width of the confidence intervals. High recombination rates mean that the data are more independent and lead to lower variance in the statistics and to narrower confidence intervals. To be conservative, we used a recombination rate on the low end of the range of plausible values for this parameter. Based on our estimates of recombination rates in *S. chilense* obtained using the LDhat software (Hudson 2001; McVean *et al.* 2002; Table S1, in Supporting Information), and the values reported by Arunyawat *et al.* (2007), we decided to use $\rho = 5$ in the bootstrap simulations.

To validate the coverage of the bootstrap confidence intervals, we performed a metabootstrap analysis. We simulated 1000 data sets under the best-fitting *fixedTau Model* with the estimates for the tomato data ('true values'), used Jaatha on them and computed bootstrap confidence intervals for each of the 1000 resulting estimates, which involved simulating 1000×1000 new data sets. For the recombination rate, we used $\rho = 10$, which is still relatively low compared with

the estimates from the *S. chilense* data (Table S1, in Supporting Information). We counted the bootstrap confidence intervals that contained the 'true value' of the parameter.

Model selection and testing

As our models all have the same number of free parameters, model selection criteria such as AIC or BIC (Akaike 1973; Schwarz 1978) will always favour the one with the highest likelihood. Again, a bootstrap-like simulation strategy can be applied to check whether a model of higher likelihood fits significantly better than the others. For example, our analyses indicated nonzero migration rates after the initial divergence of these species. To determine whether this evidence for introgression was significant, we applied a likelihood-ratio test. These likelihood ratios are actually ratios of *composite*-likelihoods, because the likelihoods were computed for the Poisson model that neglects linkage between the polymorphic sites. For this reason and because the models are not nested, we could not apply χ^2 approximations to compute *P*-values. Instead, we used another simulation-based approach. Using the ML parameter estimates from *noMig Model* assuming no migration (values from column 5 of Table 1), we simulated 1000 data sets with $\rho = 5$ using Hudson's ms. We then analysed the simulated data sets with the *noMig Model* and with the three models *Constant*, *Growth* and *Fraction-Growth* (which allow for migration). We calculated the ratios of the maximum composite-likelihood of the models allowing for migration and the *noMig Model*. We compared these likelihood ratios with the corresponding likelihood ratios from the analysis of the tomato data set. The fraction of simulated data sets with a likelihood ratio equal to or higher than the tomato likelihood ratio is then a *P*-value for the null hypothesis of no gene flow after the split.

We applied a similar likelihood ratio (LR) test to the *fixedTau Model* to test whether the growth of *S. peruvianum* was significant. For this purpose, we modified Jaatha such that the likelihood was optimized only for two parameters, setting the founding size of *S. peruvianum* (s) equal to the present-day population size ratio (q), in the following *constant fixedTau Model* (*cFT*). To assess the power of this test, we simulated 100 data sets with the parameters as estimated for the tomato data in the *fixedTau Model* and $\rho = 10$. Then, we applied Jaatha to the simulated data sets using the *fixedTau Model* as well as the *cFT Model* and calculated the LRs. With each estimate of the *cFT Model* 1000 data sets were generated, analysed with both models, their LR estimated and compared with the original LR. The proportion of LRs that were smaller than the original LR was taken as a *P*-value for the null hypothesis *cFT*. We estimated the power of this test by the fraction of the 100 simulated data sets for which the *P*-value was smaller than 5%.

Results

Comparison of accuracy of parameter estimation by Jaatha, ∂a∂i and IM

We evaluated the performance of Jaatha in comparison with *∂a∂i*, a composite-likelihood approach, and IM, a full-data Bayesian method. For the parameter estimates of θ and q, Jaatha and *∂a∂i* have similar accuracy (Fig. 4). Jaatha estimates divergence times reliably,

Table 1 Estimates for the parameters ($\widehat{\theta_1}$ per locus, $\widehat{q}$ size ratio between *S. peruvianum* and *S. chilense*, $\widehat{m}$ symmetric migration rate, $\widehat{\tau}$ divergence time, $\widehat{s}$ starting size of *S. peruvianum* immediately following the split) using Jaatha. In round parentheses are the 95% bias-corrected confidence intervals estimated using a parametric bootstrap approach. In squared brackets, the 95% bias-corrected and accelerated (BCa) confidence intervals are given. The log likelihoods (bottom rows) indicate that the *fixedTau Model* fits best, while the *Constant Model* is the worst

Parameter	*Constant*	*Growth*	*Fraction-Growth*	*noMig*	*fixedTau*
$\widehat{\theta_1}$	9.41 (7.14–12.59) [6.35–13.92]	10.30 (8.29–13.02) [7.85–12.20]	12.56 (9.61–16.38) [9.29–15.47]	13.34 (10.29–17.35) [9.97–16.60]	12.22 (9.37–15.09) [9.47–15.01]
$\widehat{q}$	1.83 (1.23–2.69) [1.02–2.11]	4.24 (2.58–6.95) [2.39–6.93]	4.29 (2.71–6.38) [2.66–5.93]	8.67 (5.34–15.00) [4.46–10.72]	4.94 (3.28–7.85) [3.25–7.70]
$\widehat{m}$	0.36 (0.06–4.89) [0.004–0.79]	0.36 (0.09–2.34) [0.02–0.71]	0.73 (0.39–1.27) [0.36–1.17]	0	0.55 (0.22–1.03) [0.16–0.96]
$\widehat{\tau}$	0.41 (0.05–1.82) [0.18–3.54]	0.37 (0.11–0.93) [0.17–1.13]	0.79 (0.37–1.63) [0.39–1.76]	0.14 (0.10–0.23) [0.10–0.24]	0.36
$\widehat{s}$	—	—	—	0.44 (0.18–0.98) [0.16–0.90]	0.33 (0.11–1.10) [0.08–0.81]
log-likelihood	−189.51	−119.70	−101.58	−133.06	−93.96

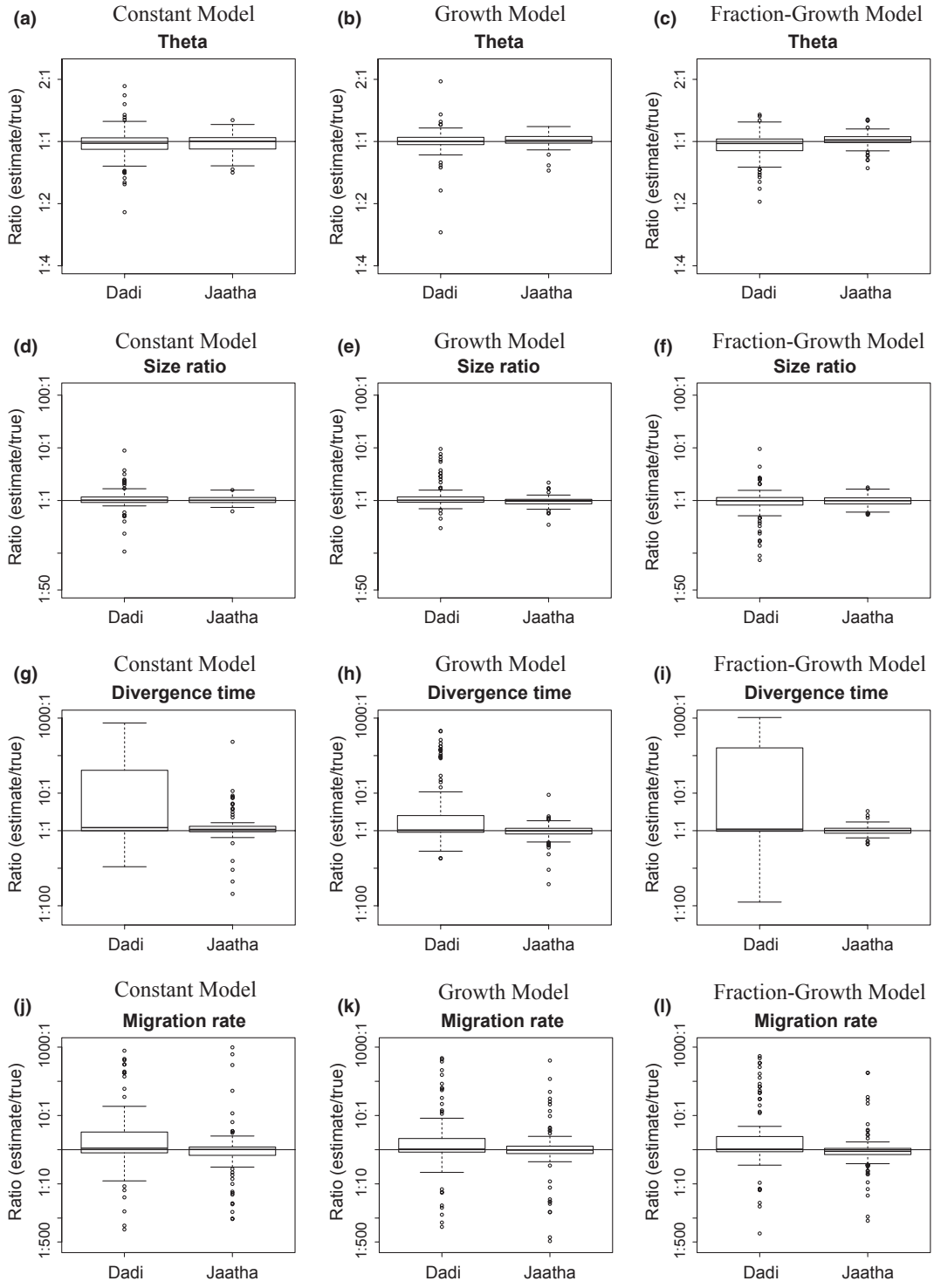

Fig. 4 Ratio of estimated to true values by *∂a∂i* and Jaatha of four parameters across models and methods for `100-loci scenario`. The 100 simulated data sets were generated without intralocus recombination.

especially when divergence times are so low that other methods fail, i.e. $\tau < 0.3$ (Figs 4 and 5a). For data sets with low divergence times, *∂a∂i* systematically estimates the most extreme τ and m, which explains the large variances of these two estimates by *∂a∂i* in Fig. 4, and Figs S1 and S2 (Supporting Information). Migration rate estimates are similar between Jaatha and *∂a∂i*, although *∂a∂i* has a slight tendency to overestimate migration

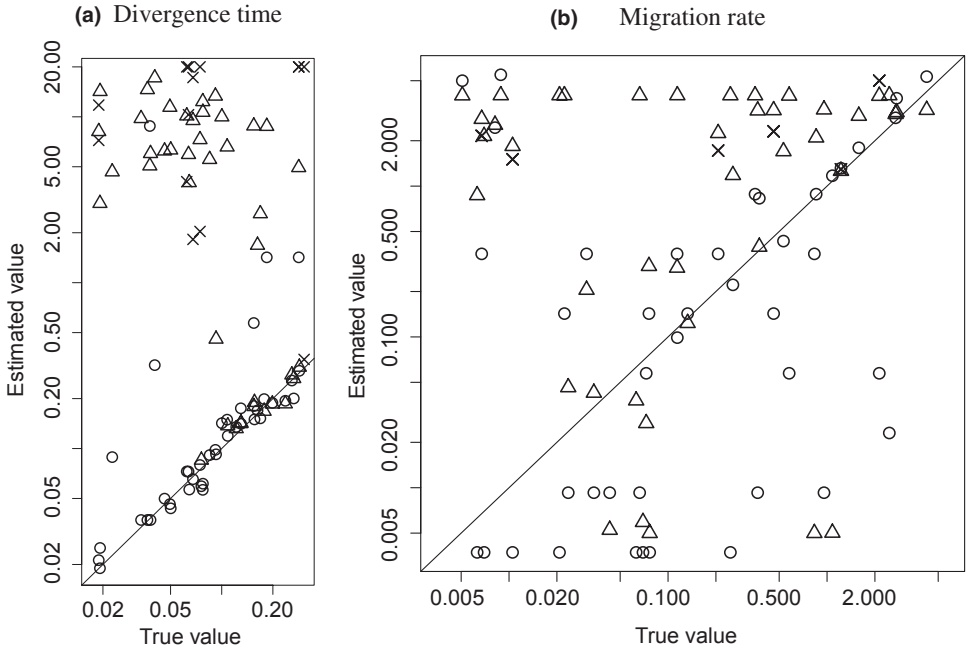

Fig. 5 The values estimated by Jaatha (○), IM (× for ESS > 100) and ∂a∂*i* (Δ) of (a) divergence time and (b) migration plotted against true values for the `100-loci scenario` of the *Constant Model* where true $\tau < 0.3$.

when divergence is recent (i.e. for low τ; Figs 4 and 5B). The accuracy of ∂a∂*i* improves as τ increases.

To compare our method with IM, we analysed simulated data sets of 100 loci with no intralocus recombination. Owing to the computational demands of IM, this analysis was restricted to ten data sets. For the IM analyses, we executed two independent runs of each data set and evaluated their convergence using the effective sampling size (ESS). The numbers of nonconverging runs based on the criterion ESS > 100 were two for the *Constant Model*, four for the *Growth Model* and seven for the *Fraction-Growth Model*. Overall, IM estimates θ and q more accurately than ∂a∂*i* and Jaatha; however, IM tends to overestimate the divergence time and migration rate (Figs S3 and S4, in Supporting Information).

Comparison of different versions of Jaatha

In Tellier *et al.*'s (2011) study, an earlier version of Jaatha with other optimization procedures ($J_1 - J_4$) is examined, where J_3 corresponds to the method described earlier. As the number of sampled loci

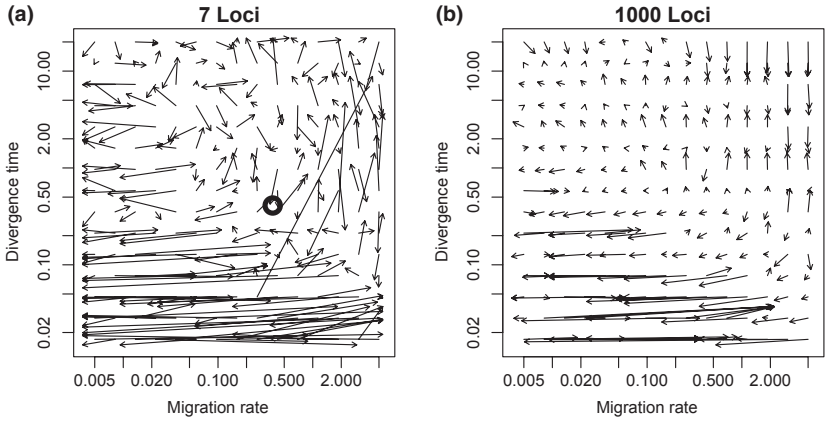

Fig. 6 Arrow plots of divergence time and migration for (a) 7 loci and (b) 1000 loci assuming the *Growth Model* with 45 samples per species and symmetric migration rates using Jaatha. The true values for the migration rate and divergence time are at the tails, and the estimated values are at the heads of each arrow. Short arrows, as in the 1000 loci case, represent accurate estimates. Horizontal arrows indicate that τ is estimated precisely but m is not. The circle is the estimated value for the tomato data under this model.

increases, our method gets more accurate, with the J_4 method showing the greatest improvement (Fig. 6 and Fig. S5A,B, in Supporting Information). These arrow plots are from the analyses of 225 simulated data sets for both 7 and 1000 loci under the *Growth Model* with symmetric migration and ρ uniformly drawn between 5 and 20. For the simulations, we combined 15 different values for the migration rate m with 15 different values for the divergence time τ. For the parameter θ_1 and the population size ratio q, we used the J_4 estimates obtained for the tomato data with the *fixedTau Model* (Table S2, in Supporting Information), $\widehat{\theta_1} = 13.08$ and $\widehat{q} = 4.64$. Jaatha was applied to estimate all four parameters θ_1, τ, m and q (results for θ_1 and q not shown). Each arrow in Fig. 6 and Fig. S5 (Supporting Information) represents the estimation error for one simulated data set. The co-ordinates at the tail of the arrow are the values of m and τ that were used for the simulation. The co-ordinates of the arrowheads are the estimates for m and τ given by Jaatha. Thus, the length of the arrow is a measure for the estimation error. Arrows parallel to the migration rate axis indicate precise τ estimates with imprecise estimates for m (Fig. 6). These are frequent for $\tau < 0.05$. With 1000 loci, divergence times are also difficult to estimate when τ and m are high but this gets better when J_4 or the multinomial model for the likelihood estimation is used (Fig. S5, in Supporting Information). The more thorough optimization methods, J_3 and J_4, are superior when many loci are available (i.e. >100). For data sets with few loci, the very fast optimization methods, J_1 and J_2, are as accurate as the more thorough procedures (data not shown).

The observed differences in accuracy were negligible between the default Jaatha method (J_3) and the variant *J*-mul that uses the multinomial approximation (eqn 2) in the studies of 7 loci (Figs S1 and S5C, in Supporting Information and Fig. 6a). For some simulations with 1000 loci, the *J*-mul estimates for size ratio q and divergence time τ were slightly more accurate than those of J_3 (Figs S2 and S5D, in Supporting Information and Fig. 6B). However, this improvement does not exactly match what can be achieved by using a more thorough numerical optimization procedure (Fig. S5B, in Supporting Information).

Application to tomato data

For the two wild tomato species *S. chilense* and *S. peruvianum*, sequences of seven housekeeping loci between 0.8 and 1.9 kb in size were available (Städler *et al.* 2008). The point estimates for the different parameters of models and estimation methods are shown in Table 1 and Table S2 (Supporting Information). The observed marginal site-frequency spectra (SFS) for the two populations and their expectation values for all models (approximated by averageing over 100 independent simulations) are shown in Fig. S6 (Supporting Information).

Consistent results across all models are that *S. peruvianum* has experienced a size expansion (i.e. $\widehat{q} > 1$) and is currently larger than *S. chilense* (at least 1.7 × the size). All models also require nonzero estimates of migration to explain the high amount of shared polymorphism between the two species. In the model that assumes no migration, extremely short divergence times are required to offset the lack of ongoing migration (i.e. less than half of the divergence time as in the other models).

The estimates for the tomato data are located near a region of long arrows indicating low certainty in parameter estimates in this range (Fig. 6). This underlines the importance of considering confidence intervals for the estimates. In a metabootstrap analysis, we assessed the reliability of the 95% bootstrap confidence intervals given in Table 1. For the parameters θ_1, q and m the coverage was 94%, 94% and 97%, which means that the bootstrap confidence interval is acceptable as approximate 95% confidence intervals. The estimated coverage of the bootstrap confidence intervals was only 92% for s (starting size of *S. peruvianum*). We also computed bias-corrected and accelerated (BCa) bootstrap confidence intervals (Efron & Tibshirani 1993), which takes into account that the variance of an estimator can depend on the true parameter value and applies a correction that is based on the skewness of the bootstrap results. In most cases, using the BCa confidence intervals improved the coverage (θ_1: 95%, q: 93%, m: 95%, s: 94%). The BCa intervals for the tomato data (Table 1, squared brackets) show little difference to the BC intervals in all but three cases, $\widehat{m}$ of *Constant* and *Growth Model* and $\widehat{\tau}$ of the *Constant Model*. Because the bootstrap results are not symmetrically distributed around the mean, in the case of $\widehat{m}$, the BCa intervals are smaller or, in case of $\widehat{\tau}$, larger.

To our surprise, the model having the highest likelihood indicated that gene exchange between the two tomato species continued after their initial divergence. The (composite-) likelihood ratios favoured models with gene flow after the population split (*Growth* and *Fraction-Growth Model*) over the *noMig Model* without gene flow after the split. In fact, the poorest fit to our data is that of the *Constant Model*, which does not incorporate population expansion in *S. peruvianum*. The negative log likelihood-ratios in Table 1 show that this model fits even worse than the *noMig Model*. We confirmed that the models with gene flow *and* growth of *S. peruvianum* fit significantly better than the *noMig Model* by comparing the observed log likelihood-ratio with the

Table 2 Log likelihood-ratios of models with migration to *noMig Model* applied to the tomato data. Positive values indicate that the model with migration is a better fit to the data than one without. In the third column, the ranges of log likelihood-ratios (ℓLR) of 1000 simulated bootstrap replicates are given. In the fourth column (*P*-value), the proportion of bootstrap ℓLR that were bigger than or equal to the corresponding tomato ℓLR are given

Model compared with *noMig*	tomato ℓLR	range of bootstrap ℓLR	*P*-value
Constant	−53.12	[−136, −9]	0.272
Growth	13.31	[−68, 22]	0.003
Fraction-Growth	35.21	[−86, 59]	0.003

distribution of log likelihood-ratios from the corresponding bootstrap data sets ($P < 0.003$ for *Growth* and $P < 0.003$ for *Fraction-Growth Model*, Table 2). We repeated this likelihood-ratio test with six different HKY model parameter settings (Hasegawa *et al.* 1985) with the base frequencies estimated from the tomato data to see whether finite sites models would yield the same results. The finite sites models differed in their transition–transversion (ts/tv) ratio (estimated from the data, values used: either 2 or 3) and the gamma shape parameter α (0.2, 0.3, 0.6). The latter models mutation rate heterogeneity between the sites, with smaller values of α causing more heterogeneity. The recombination rate was set to $\rho = 20$. The incorporation of finite sites did not change the results significantly. The only differences from the earlier analyses were that the *P*-values were slightly larger for the *Growth Model* ($P < 0.004$) and the *P*-values for the *Fraction-Growth Model* were smaller ($P < 0.001$), except for the case where ts/tv ratio = 2 and $\alpha = 0.2$ ($P = 0.07$). However, an α of 0.2 is an extreme value as values of α ranged from 0.46 to 1.09 across loci based on the best-fitting model (GTR + G + I) according to Modeltest (Posada & Crandall 1998).

To examine the power of a Jaatha-based test for population growth, we simulated 100 data sets under the *fixedTau Model* and applied a simulation-based test with the *constant fixedTau Model*, where no population growth was allowed, as the null hypothesis. In all 100 cases, we obtained a significant result ($P < 0.001$), correctly favouring the model including growth over the model without growth. For the tomato data, we obtained a highly significant result as well ($P < 0.001$).

Discussion

In this study, we introduce a new algorithm, Jaatha, for inferring population genetic parameters from DNA sequence data. In most of our simulation studies, Jaatha gave comparable results to other programs (IM and *∂a∂i*) and, for low divergence times (e.g. 0.017–0.15 measured in $4N_1$ generations), Jaatha even outperformed other programs. One possible explanation why *∂a∂i* had difficulty estimating parameters when divergence times are recent may be that the JSFS looks similar to that in the case of high divergence time and high migration rates, and it is therefore difficult to distinguish between these cases (R. Gutenkunst, personal communication). Furthermore, although our method is based on the assumption of the independence of sites, its accuracy is not compromised when used on data sets of sufficiently many unlinked loci with limited or no within-locus recombination (e.g. Fig. 4 and Fig. S3, in Supporting Information). Thus, Jaatha may be a fast and reliable alternative to currently available full-likelihood methods and offers a solution when no suitable full-likelihood method is available.

Jaatha can be run using four different optimization methods, $J_1 - J_4$, where J_3 is described in this manuscript. When only few loci are available for analysis, J_3 provides a good compromise between run-time (< 15 sec) and accuracy. For data sets with more loci, the more precise optimization J_4 gives the best results and should be the method of choice. A variant of Jaatha (*J*-mul) that uses the multinomial approximation (eqn 2) instead of the Poisson approximation (eqn 1) to compute the composite-likelihood is slightly more accurate. This results from the way in which we estimate θ_1. In upcoming versions of Jaatha, we plan to estimate θ_1 in the same way as with other parameters, which means that the exact equation for the composite-likelihood will be analogous to the Poisson approximation (eqn 1).

The current version of Jaatha was intended as a proof of concept for fast and simple parameter estimation procedures in population genetics. However, our application of Jaatha to an analysis of divergence between two closely related wild tomato species shows that Jaatha can be readily applied to draw biologically meaningful conclusions from actual data. However, because our simulation studies indicate that analyses based only on a limited number of loci (e.g. seven or fewer) are challenging for accurate parameter estimation, we consider our parameter estimates for the wild tomato species as preliminary. Based on the best-fitting model (*fixedTau*) and a mutation rate of $5.1 \cdot 10^{-9}$/site/year at silent sites (Roselius *et al.* 2005) and a total length of all loci excluding gaps of 8844 bp (954 SNPs), the split time between these two species is either 730 000 years, if we assume one generation per year, or ~5.1 million years, if we assume a generation every 7 years. The exact generation time of these species is not known. These species are short-lived perennials and have a viable seed bank (R. Chetelat, personal communication).

Seed germination and fecundity are likely affected by El Niño and La Niña cycles, and therefore, two different generation times were considered [see also Roselius *et al.* (2005) and Arunyawat *et al.* (2007)]. According to the best-fitting model, the effective population size of *S. chilense* is ~72 000. All models indicate that *S. peruvianum* is larger than *S. chilense*, although we also allowed for population shrinkage of *S. peruvianum*. These results are consistent with the conclusions made by Städler *et al.* (2005). Our estimated size ratio between these two species ranges from 1.83 to 8.67, including values close to those estimated previously by Städler *et al.* (2005). Our highest values for this size ratio emerge from the model without migration. This model also has the smallest estimated divergence time, which is required to explain the high proportion of shared polymorphism between these species, if migration is excluded. In contrast, from the *Fraction-Growth Model*, in which the population size of *S. peruvianum* is set to 5% of the size of *S. chilense* population at the time of the split, we recover the largest values for divergence times. Higher values of τ are needed to explain the present-day differences in population sizes between these species, because *S. peruvianum* has the larger population size, but was forced in the *Fraction-Growth Model* to be much smaller at the time of the splitting event.

The metabootstrap analysis showed that the coverage of the bias-corrected bootstrap confidence intervals depended on the parameter estimated and is close to the target value of 95% (θ_1: 94%, q: 94%, m: 97%). For the parameter s, the initial population size of *S. peruvianum*, of the *fixedTau Model*, the coverage of the bootstrap confidence intervals was slightly poorer (92%). The bootstrap confidence intervals with BCa correction showed satisfactory coverage for all four parameters (θ_1: 95%, q: 93%, m: 95%, s: 94%).

All models estimate nonzero migration rates, indicating that some gene flow was likely following the initial divergence between these species. With a simulation-based hypothesis test, we showed that there is significant evidence for population growth in *S. peruvianum* ($P < 0.001$) and also for post-divergence migration ($P < 0.003$). The simulation-based approach with multiple finite site models yielded similar significant results. We were surprised to find significant evidence for gene flow after the species split as although contemporary populations of these species are sympatric, no hybrids between these have been reported in the field (R. Chetelat, personal communication). Furthermore, forced hybridizations between these species result in small inviable seeds with underdeveloped embryos and endosperm (Rick & Lamm 1955). One possible explanation for the signature of gene flow following the split is that the accumulation of the present-day hybrid barriers was a gradual process and that some hybridization took place during the early stages of the divergence process. Hybridization likely became less and less common with the acquisition of proper speciation barriers, which are currently in place. The incorporation of haplotype information into Jaatha may allow us to distinguish between hybridization that took place more recently and less recently. We would expect that more recent hybridization would contain recognizable haplotypes brought into the sister species through migration, while recombination would have obliterated shared haplotypes if hybridization occurred early on in the divergence process (Machado *et al.* 2002).

Because our simulation studies show a remarkable improvement in accuracy when the number of loci is increased, we aim to develop and analyse a much larger data set for this pair of tomato species (Fig. 6 and Fig. S2, in Supporting Information). This will serve as a cornerstone for future studies looking at the molecular evolution of genes underlying ecologically relevant traits such as parasite resistance. Another limitation of the current data set is the sampling regime as discussed by Städler *et al.* (2009), in which individuals from four geographically isolated populations per species were studied. Although this is a very good starting point for genetic studies, this is not the preferred sampling scheme for establishing historical demography. Either the species should be sampled on a species-wide level or the structure the sampling scheme introduces (i.e. when local populations are sampled) should be accounted for in the underlying model. Therefore, it will be one of our next steps in the further development of Jaatha to take substructure of the two species into account.

In our simulation studies, we focused on scenarios in which the assumption of infinite sites is met and only four parameters are to be estimated. The assumption of infinite sites is rarely fulfilled in real data sets, and this assumption is known to be violated in the data set from wild tomatoes. However, the current version of Jaatha is only applicable if these two constraints are met, namely infinite sites and estimation of a maximum of four parameters. In this respect, IM and *∂a∂i* are more flexible. Both can be applied for the joint estimation of more than four parameters. Moreover, IM can take into account back-mutations and multiple hits using the HKY model for sequence data (Hasegawa *et al.* 1985) or a stepwise-mutation model for microsatellite data (Kimura & Ohta 1978). Even though the current version of Jaatha estimates four parameters, the optimization step operates on a cube of only three dimensions. This is possible because we apply a method of moments to estimate θ_1 which we can seamlessly combine with the composite ML estimation of the other three parameters because the expectation values of the JSFS are

proportional to θ_1. The latter applies only under infinite sites assumptions. Thus, allowing for finite sites mutation models in Jaatha will expand the search space by at least one dimension.

Future versions of Jaatha will also offer the option to jointly estimate more than four parameters. In this mode, however, it will not be feasible to perform a priori all simulations that are necessary to approximate the composite-likelihood function on a fine grid of parameter combinations. Instead, we will start with a very coarse grid or randomly chosen combinations of parameter values and sample locally from a finer grid as required during the optimization procedure. Of course, the parameter optimization phase of Jaatha will take noticeably longer if more than four parameters are jointly estimated. For a Bayesian version of Jaatha, we plan to build upon ideas from MCMC-ABC (cf. Beaumont *et al*. 2002; Marjoram & Tavare 2006; Wegmann *et al*. 2009; Leuenberger & Wegmann 2010). Jaatha already has in common with ABC methods that the (composite-) likelihood function is not computed but estimated from simulation runs. This makes it very easy to implement changes into the method. Likewise, the choice of summary statistics is of crucial importance. The 23 JSFS-based summary statistics worked well for our purposes but it may be possible to further optimize the set of summary statistics by applying PLS (Wegmann *et al*. 2009) or the method of Joyce & Marjoram (2008) to the JSFS and to haplotype-based statistics.

In our simulation studies, parameter estimates from data sets with a limited number of independent loci (10 or fewer) were quite inaccurate. We conjecture that this is not the result of poor performance of the numerical estimation procedures, but rather because these 'small' data sets do not contain sufficient information. Thus, it is questionable whether one should try to estimate more than four parameters from such data sets and whether it is worthwhile to apply sophisticated and run-time-intensive estimation procedures. In contrast, when data from 100 or 1000 independent loci are available, our simulation studies indicate that simple and fast methods like Jaatha can estimate a limited number of parameters with satisfying accuracy. Full-data methods like IM, which do not rely on summary statistics, are perhaps most useful for data sets with an intermediate number of independent loci. For cases with either very low or very high numbers of independent loci, summary-statistic-based methods like Jaatha may be an alternative to get fast results of reasonable accuracy.

Acknowledgements

We thank Ryan Gutenkunst for helping to run ∂a∂i and the DFG Forschergruppe FOR1078, especially Peter Pfaffelhuber and Joachim Hermisson for fruitful discussions. We also thank Asger Hobolth and an anonymous reviewer for comments, which helped us to substantially improve the manuscript. This work has been supported by the German Research Foundation (DFG) grant ME 3134/3-1 to LR and DM.

References

Adams AM, Hudson RR (2004) Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, **168**, 1699–1712.

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory* (eds Petrov P, Csaki F), pp. 267–281. Akad. Kiado, Budapest.

Andolfatto P, Przeworski M (2000) A genome-wide departure from the standard neutral model in natural populations of Drosophila. *Genetics*, **156**, 257–268.

Arunyawat U, Stephan W, Städler T (2007) Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Molecular Biology and Evolution*, **24**, 2310–2322.

Beaumont M, Zhang W, Balding D (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, **17**, 1505–1519.

Byrd RH, Lu P, Nocedal J, Zhu C (1995) A limited-memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, **16**, 1190–1208.

Chetelat RT, Pertuzé RA, Faúndez L, Graham EB, Jones CM (2009) Distribution, ecology and reproductive biology of wild tomatoes and related nightshades from the Atacama Desert region of northern Chile. *Euphytica*, **167**, 77–93.

DiCiccio T, Efron B (1996) Bootstrap confidence intervals. *Statistical Science*, **11**, 189–228.

Efron B, Tibshirani R (1993) *An Introduction to the Bootstrap*. Chapman & Hall, Boca Raton.

Garrigan D (2009) Composite likelihood estimation of demographic parameters. *BMC Genetics*, **10**, 72.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**, e1000695.

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.

Hey J (2010) Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, **27**, 905–920.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 2785–2790.

Hudson RR (2001) Two-locus sampling distributions and their application. *Genetics*, **159**, 1805–1817.
Hudson RR (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.
Joyce P, Marjoram P (2008) Approximately sufficient statistics and bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, **7**, Article 26.
Kim Y, Stephan W (2000) Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics*, **155**, 1415–1427.
Kimura M, Ohta T (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences of the United States of America*, **75**, 2868–2872.
Kuhner MK (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of pop ulation parameters. *Bioinformatics*, **22**, 768–770.
Leuenberger C, Wegmann D (2010) Bayesian computation and model selection without likelihoods. *Genetics*, **184**, 243–252.
Li H, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in Drosophila. *PLoS Genetics*, **2**, e166.
Lopes JS, Balding D, Beaumont MA (2009) PopABC: a program to infer historical demo graphic parameters. *Bioinformatics*, **25**, 2747–2749.
Machado CA, Kliman RM, Markert JA, Hey J (2002) Inferring the history of speciation from multilocus DNA sequence data: the case of Drosophila pseudoobscura and close relatives. *Molecular Biology and Evolution*, **19**, 472–488.
Marjoram P, Tavare S (2006) Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics*, **7**, 759–770.
McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.
Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.
Posada D, Crandall KA (1998) Modeltest: testing the model of dna substitution. *Bioinformatics*, **14**, 817–818.
R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
Rick C, Lamm R (1955) Biosystematic studies on the status of lycopersicon chilense. *American Journal of Botany*, **42**, 663–675.
Robertson A (1975) Remarks on the Lewontin–Krakauer test. *Genetics*, **80**, 396.
Roselius K, Stephan W, Städler T (2005) The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics*, **171**, 753–763.
Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. *Genetics*, **132**, 1161–1176.
Schwarz G (1978) Estimating the dimensions of a model. *Annals of Statistics*, **6**, 461–464.
Städler T, Arunyawat U, Stephan W (2008) Population genetics of speciation in two closely related wild tomatoes (Solanum section Lycopersicon). *Genetics*, **178**, 339–350.
Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P (2009) The impact of sam pling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics*, **182**, 205–216.
Städler T, Roselius K, Stephan W (2005) Genealogical footprints of speciation processes in wild tomatoes: demography and evidence for historical gene flow. *Evolution*, **59**, 1268–1279.
Strasburg JL, Rieseberg LH (2010) How robust are 'isolation with migration' analyses to violations of the im model? A simulation study. *Molecular Biology and Evolution*, **27**, 297–310.
Tellier A, Pfaffelhuber P, Haubold B *et al.* (2011) Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum. *PLoS One*, (to appear).
Teshima K, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Research*, **16**, 702–712.
Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.
Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computa tion coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.

L.N. is interested in methods to estimate demographic histories and their applications. L.R. studies the molecular evolution of wild tomatoes and coadaptation between plants and microbes. The main focus of D.M.'s research is on the development of model-based methods for the analysis of genetic data.

Data accessibility

For the wild tomato data set of *S. chilense* and *S. peruvianum*, we used sequences from Städler *et al.* (2008). The artificial data sets generated for the simulation study can be provided upon request.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Estimated recombination rates with LDhat for *S. chilense* loci—Recombination rates per locus and per $4N_1$ generations estimated with LDhat (Hudson 2001; McVean *et al.* 2002) using the *S. chilense* sequences and $\theta_{site} = 0.01$.

Table S2 Estimates for parameters of models fitted to tomato data. Estimates for the parameters ($\hat{\theta}_1$ per locus, $\hat{q}$ size ratio between *S. peruvianum* and *S. chilense*, $\hat{m}$ symmetric migration rate, $\hat{\tau}$ divergence time, $\hat{s}$ starting size of *S. peruvianum* right after the split) using the J_1, J_2, J_4, and multinomial estimation methods. In parentheses are the 95% BC-confidence intervals estimated using a parametric bootstrap approach. The log-likelihood (bottom rows) are calculated using the Poisson model and indicate that the *fixedTau Model* fits best while the *Constant Model* is the worst.

Fig. S1 Ratio of estimated to true values by ∂a∂i, Jaatha, and Jaatha with the (composite-) likelihood estimation based on a

multinomial model (J-mul) of four parameters, across models and methods for `7-loci scenario`.

Fig. S2 Ratio of estimated to true values by *∂a∂i*, Jaatha, and Jaatha with the (composite-) likelihood estimation based on a multinomial model (J-mul) of four parameters across models and methods for `1000-loci scenario`.

Fig. S3 Ratio of estimated to true values of four parameters across models and methods for `100-loci scenario` (no recombination). IM results with ESS <100 are not included in the boxplots but drawn in additionally(Δ). Results for ∂a∂i and Jaatha with the same 10 simulated datasets for *Constant*, *Growth*, and *Fraction-Growth Models* are shown.

Fig. S4 Estimations of the four parameters using the three methods: Jaatha (o), ∂a∂i (Δ), and IM (× for ESS > 100; + for ESS < 100 of that variable). These methods were applied to 10 simulated datasets each with 100 loci, without intralocus recombination. Shown are the estimations assuming three different underlying demographic models.

Fig. S5 Arrow plots of divergence time and migration for seven and 1000 loci under the *Growth Model* with 45 samples per species and symmetric migration rates with J_4 (A and B, as in Tellier *et al.)* and Jaatha using a multinomial approximation (J-mul) for the composite-likelihood (C and D). The circle is the estimated value for the tomato data under this model. Each estimation in A and B took on average 15 minutes and in C and D only 15 seconds.

Fig. S6 The marginal site frequency spectra (SFS) for the tomato data and the average of 100 simulated data sets with each seven loci for the tested five models *fixedTau*, *noMig*, *Constant*, *Growth*, and *FractionGrowth*. The line represents the expected SFS of the neutral Wright-Fisher Model of constant size without migration (Fu, 1995).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.