Jaatha: A Fast Composite Likelihood Approach to Estimate Demographic Parameters

Lisha Naduvilezhath, Laura E. Rose, Dirk Metzler

October 30, 2010

Corresponding Author: Lisha Naduvilezhath
LMU Biocenter
Department Biology II
Grosshadernerstrasse 2
82152 Planegg
Germany
+49 (0)89 2180-74152 (ph.)
+49 (0)89 2180-74104 (fax)
Lisha@bio.lmu.de

Running Head: Estimation of demographic parameters

Key Words: composite likelihood method, demography, recent divergence, wild tomatoes
(Solanum chilense, S. peruvianum)
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 which 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, though 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 peruvianum* and *S. chilense*. Our results indicate that these species likely diverged between $1.44 \cdot N$ and $3.6 \cdot N$ 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, *Solanum peruvianum* likely experienced a population expansion following speciation.
Introduction

The availability of more and more affordable genome technologies have allowed scientists to venture outwards from the classical model systems and to begin answering questions about evolutionary genetics and trait evolution in non-model 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, Nielson, 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 which 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 IM/IMa against moderate violations of this and other assumptions was examined in a recent simulation study (Strasburg & Rieseberg, 2010). 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. The underlying equilibrium assumptions for the population structure make this method inappropriate for recent speciation events (Kuhner, 2006). To analyze datasets with a high amount of intralocus recombination from recently diverged species, Becquet & Przeworski (2007) introduced an MCMC method (MIMAR) which 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 which can require several weeks to converge.

Gutenkunst et al. (2009) implemented a promising diffusion approximation in $\partial a \partial i$ which is considerably faster than the methods described above and can be used for various demographic scenarios of up to three populations. In this composite-likelihood method which assumes unlinked SNPs (see also Hudson (2001); Kim & Stephan (2000); McVean et al. (2002)) the data is 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 summary statistics of Wakeley & Hey (1997) can be computed from the JSFS. 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. 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 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 dataset. 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 “JSFS associated approximation of the ancestry”, 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 datasets each. Due to the fast runtime and great flexibility of the underlying demographic cases we chose $\partial a \partial i$ for comparing the results with our program. To compare Jaatha to the full-likelihood method IM we applied the programs to simulated datasets 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 peruvianum* and *S. chilense*. 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 (Städler *et al.*, 2008). Although the isolation model of speciation could not be rejected (Wakeley & Hey, 1997), Städler *et al.* (2008) found some evidence for post-divergence introgression using the LD-based method of Machado *et al.* (2002). Due to the recency of divergence and high amount of within-locus recombination in these species, this dataset 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 $\mu$ 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 $\theta_1$ and three additional parameters.

In our simulation studies described below we assess the accuracy of Jaatha’s estimations for the parameters $\theta_1$, the population size ratio $q = \frac{N_2}{N_1} = \frac{\theta_2}{\theta_1}$, the divergence time $\tau$, 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 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. peruvianum* and *S. chilense* (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. Since the current version of Jaatha is restricted to estimating four parameters including $\theta_1$ 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 $\tau$ to 0.36. This is the estimate of $\tau$ with the models *Constant* and *Growth*.

### 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 $S = (S_1, \ldots, S_{23})$ 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.

Jaatha provides four different parameter optimization procedures for the estimation phase, which differ in their run-time ($\leq 0.5$ sec, $\leq 1$ sec, $\leq 15$ sec, $\leq 20$ min on a modern desktop PC). The training phase takes up to 5 days on a modern desktop PC, but its results can be reused for datasets 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 $A = \{0, \ldots, m\} \times \{0, \ldots, n\} \setminus \{(0, 0), (m, n)\}$. 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 Figure 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 partitions are also possible and are analyzed by Tellier et al. (2010).

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 \(\theta'_1, \tau', m',\) and \(q'.\) For fixed values \(\tau', m',\) and \(q',\) we estimate \(\theta_1\) by \(\theta'_1 \cdot y/y'.\) Thus, we separate the estimation of \(\theta_1\) 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 \(P = [\tau_{\text{min}}, \tau_{\text{max}}] \times [m_{\text{min}}, m_{\text{max}}] \times [q_{\text{min}}, q_{\text{max}}].\) 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

\[
d : P \rightarrow [1, 40] \times [1, 40] \times [1, 40] \\
(\tau, m, q) \mapsto (d_\tau, d_m, d_q) = \left(\log z_\tau (\tau/\tau_{\text{max}}) + 1, \log z_m (m/m_{\text{max}}) + 1, \log z_q (q/q_{\text{max}}) + 1\right),
\]

where \(z_p = \sqrt[p]{p_{\text{min}}/p_{\text{max}}}\) for each parameter \(p \in \{\tau, m, q\}.\) The inverse transformations are given by \(p = p_{\text{max}} \cdot z_p^{-1}.\) 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 \((\tau, m, q)\) corresponding to grid points, Jaatha calls the program ms (Hudson, 2002) to simulate 10 independent datasets 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.

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 8 intervals \([1, 5.5], [5.5, 10.5], [10.5, 15.5], \ldots, [35.5, 40],\) where \((a, b]\)
denotes the interval \(x : a < x \leq b.\) We chose these grid and bin sizes because they provide a reasonable compromise between accuracy and run-time. Each of the \(8 \times 8 \times 8 = 512\) 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 \(d_\tau, d_m\) 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 of the range \( (a_p - 3, b_p + 3] \), such that we obtain four different weights \( \frac{1}{2}, \frac{1}{4}, \frac{1}{8} \).

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

\[
\hat{\beta}_{0,i} + \hat{\beta}_{\tau,i} \cdot d_{\tau} + \hat{\beta}_{m,i} \cdot d_{m} + \hat{\beta}_{q,i} \cdot d_{q} = \ln(\lambda_i),
\]

where \( \lambda_i \) is the expected value of \( S_i \), which is assumed to be Poisson distributed. The dependence of \( \lambda_i \) on the original parameters \( \tau, m, \) 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 \texttt{glm()} to fit the weighted Poisson GLMs (R Development Core Team, 2009).

**Estimation phase:** For the estimation of \( \theta \) let \( s_1, \ldots, s_{23} \) be the values of the 23 summary statistics observed in the given dataset, \( b \) 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 center of \( b \).

One simulated dataset of the training phase consists of 7 loci with \( \theta_1 = 5 \) per locus, so we estimate \( \theta_1 \) for bin \( b \) by

\[
\hat{\theta}_b = \frac{\sum_{i=0}^{23} s_i}{\sum_{j=0}^{23} s_j^{(b)}/35},
\]

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

The composite likelihood of a parameter combination \((\tau, m, q)\) 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 \( \tau, m, \) and \( q \) and \( \theta = \hat{\theta}_b \), where \((d_{\tau}, d_{m}, d_{q}) \in b \). 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 $\tau, m$ and $q$ and then optimize the resulting approximate composite likelihood function

$$L_{s_1, \ldots, s_{23}}(\tau, m, s) = \prod_{i=1}^{23} \frac{\hat{\lambda}_i(\tau, m, q)^{s_i} \cdot e^{-\hat{\lambda}_i(\tau, m, q)}}{s_i!}.$$ 

Here, $\hat{\lambda}_i(\tau, m, q)$ is our estimation for $\mathbb{E}S_i$ in terms of $\tau, m, q$, and, implicitly the corresponding $\hat{\theta}_b$. Jaatha provides the following four procedures to optimize $L_{s_1, \ldots, s_{23}}(\tau, m, s)$, which offer different trade-offs between efficiency and accuracy.

**Bin Center Method** ($\text{Cent}$): We compute the composite likelihoods of the parameter combinations that correspond to the centers of the 512 bins in the log-scaled parameter space. The parameter combination of the highest composite likelihood among these 512 is returned.

**Weighted Mean of Centers Method** ($\text{wCent}$): In the log-scaled parameter space we compute the weighted mean of the bin centers using the composite likelihoods as weights. The result is the parameter combination corresponding to this center of mass.

**Within-Bin Optimization** ($\text{wBin}$): In each bin center we start a maximization procedure for the composite likelihood within the bin. This is realized with the `optim` function of R and the optimization procedure of Byrd et al. (1995).

**Interpolation Between Bins** ($\text{IbBin}$): In each bin center we start an optimization procedure that – in contrast to the Within-Bin method – may leave its bin. To allow for smooth transitions between the bins we interpolate the estimates of $\lambda_i$ near the bound-
aries of the bins. Let the log-transformed parameter values \((d_\tau, d_m, d_q) = (x, y, z)\) be in a bin \(b\) with center \((x_b, y_b, z_b)\) and let \((x', y_b, z_b), (x_b, y', z_b)\) and \((x_b, y_b, z')\) be the centers of the neighboring bins \(b_x, b_y\) and \(b_z\) that are closest to \((x, y, z)\). Let \(\lambda_{i,b}, \lambda_{i,x}, \lambda_{i,y}\) and \(\lambda_{i,z}\) be the expectation values of a summary statistic \(S_i\) predicted for \((x, y, z)\) by the Poisson GLMs of the four bins. We interpolate between the bins by setting \(\hat{\lambda}_i\) to a weighted mean \(w_b \cdot \lambda_{i,b} + w_x \cdot \lambda_{i,x} + w_y \cdot \lambda_{i,y} + w_z \cdot \lambda_{i,z}\), where the weights are computed by the following procedure. Initially, all weights are set to \(1/4\). Then, for each \(p \in \{x, y, z\}\) and the corresponding \(p_b \in \{x_b, y_b, z_b\}\) and \(p' \in \{x', y', z'\}\), the weights \(w_b\) and \(w_p\) are multiplied by \(c - c \cdot ((p - p_b) / \Delta)^2\) and \(c - c \cdot ((p - p') / \Delta)^2\), respectively, where \(\Delta = 5\) is the distance between neighboring bin centers and the coefficient \(c\) keeps the sum of the weights equal to \(1\).

An R script (R Development Core Team, 2009) containing an implementation of Jaatha is freely available from the web-site

http://evol.bio.lmu.de/_statgen/software/jaatha/.

**Comparison of Jaatha, IM and \(\partial a \partial i\)**

We compare the accuracy of parameter estimations for \(\theta, \tau, m,\) and \(q\) by Jaatha, IM (Hey & Nielsen, 2004), and \(\partial a \partial i\) (Gutenkunst et al., 2009). A simulation study to compare a variant of Jaatha to MIMAR (Becquet & Przeworski, 2007) and PopABC (Lopes et al., 2009) has been performed by Tellier et al. (2010). We applied the three programs Jaatha, IM and \(\partial a \partial i\) to datasets that we simulated with Hudson’s ms software for three different demographic models, each with three scenarios described below. These scenarios differ in their number of loci, type of migration (asymmetric or symmetric), and amount of recombination. For each scenario and population, 100 datasets were simulated with 25 sequences sampled from each population. The parameter ranges and the underlying demographic models were as described above (for ms commands as well as parameter ranges, see supplementary
information).

7-loci scenario 100 datasets were simulated with seven loci, asymmetric migration between populations, and a within locus recombination rate chosen randomly between 5 and 20 per locus.

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

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

Since 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. Since IM has a high demand for computer run-time, this simulation study was limited to 10 datasets. We restricted the run-time to five weeks per IM run. To assess convergence we performed two independent runs for each of the 10 datasets. $\partial a/\partial i$ was run on all three demographic models and all three simulation scenarios. The underlying demographic models and parameter ranges (except for $\theta$) were precisely specified for $\partial a/\partial i$ 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 Dataset

For the two wild tomato species $S. peruvianum$ and $S. chilense$, sequences of seven loci between 0.8 to 1.9 kb in size were available (Städler et al., 2008). Since 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. For the simulations in Jaatha’s training phase we sampled 45 sequences per species, matching the average
number of samples available in the tomato dataset. We fit all five models specified above to the tomato data and compared the Poisson-model maximum-likelihood (ML) values for the models.

To assess the uncertainty of the parameter estimates for the tomato data, we used a parametric bootstrap approach. For each combination of model and estimation method, we simulated 1000 bootstrap replicates using the respective ML estimates (Table 1). Each replicate, simulated using the ms program, contained 7 loci with a fixed rate of recombination from 45 samples per population. Due to time constraints, the slower estimation method IbBin was not used. A normal approximation of the log-transformed bootstrap results was used to derive the biased-corrected intervals. The bias-correction was calculated in the following way: 

$$2 \cdot \hat{\theta} - \bar{\theta} \pm 1.96 \cdot \sigma(\bar{\theta}),$$

where \(\hat{\theta}\) is the maximum-likelihood value, \(\bar{\theta}\) is the mean and \(\sigma(\bar{\theta})\) the standard deviation of the bootstrap results.

Our analyses indicated non-zero migration following 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, since the likelihoods were computed for the Poisson model which neglects linkage between the polymorphic sites. For this reason and since the models are not nested, we could not apply \(\chi^2\)-approximations to compute \(p\)-values. Instead we used another parametric bootstrap approach. Using the ML parameter estimates from noMig Model for each estimation method assuming no migration (values from column 6 of Table 1) we simulated 1000 datasets using Hudson’s ms. We then analyzed the simulated datasets 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 based on the noMig Model and the maximum composite likelihood based on the models allowing for migration. We compared these likelihood ratios to the corresponding likelihood ratios from the analysis of the tomato dataset. The fraction of bootstrap datasets with a likelihood ratio
equal to or higher than the original likelihood ratio is then a $p$-value for the null hypothesis of no gene flow after the split.

**Results**

*Comparison of accuracy of parameter estimation by Jaatha, $\partial a/\partial i$, and IM*

We evaluated the performance of Jaatha’s four estimation variants and compared them to two other methods: $\partial a/\partial i$, a composite likelihood approach and IM, a full-likelihood approach. For the parameter estimates of $\theta$ and $q$, Jaatha and $\partial a/\partial i$ have similar accuracy (Fig. 4).

Jaatha estimates divergence times reliably, especially when divergence times are so low that other methods fail, i.e. $\tau < 0.3$ (Fig. 4, Fig. 5A). For datasets with low divergence times, $\partial a/\partial i$ systematically estimates the most extreme $\tau$ and $m$, which explains the large variances of these two estimates by $\partial a/\partial i$ in Fig. 4, in Supplemental Fig. S3, and in Supplemental Fig. S4. Migration rate estimates are similar between Jaatha and $\partial a/\partial i$, although $\partial a/\partial i$ has a slight tendency to overestimate migration when divergence is recent (i.e. for low $\tau$; Fig. 4, Fig. 5B). The accuracy of $\partial a/\partial i$ improves as $\tau$ increases.

To compare our method to IM, we analyzed simulated datasets of 100 loci with no intralocus recombination. Due to the computational demands of IM, this analysis was restricted to ten datasets. For the IM analyses, we executed two independent runs of each dataset and evaluated their convergence using the effective sampling size (ESS). The numbers of non-converging 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 $\theta$ and $q$ more accurately than $\partial a/\partial i$ and Jaatha, however IM tends to overestimate the divergence time and migration rate (Supplemental Fig. S1, Supplemental Fig. S2).

As the number of sampled loci increases, our method gets more accurate, with the *IbBin* method showing the greatest improvement (Fig. 4, Supplemental Fig. S3, Supple-
mental Fig. S4). The more thorough optimization methods, wBin and IbBin, are superior when many loci are available (i.e. > 100). For datasets with few loci, the very fast optimization methods, Cent and wCent, are as accurate as the more thorough procedures.

Application to tomato data

For the two wild tomato species S. peruvianum and S. chilense, sequences of seven loci between 0.8 to 1.9 kb in size were available (Städler et al., 2008). These loci are not known to be targets of selection. The point estimates for the different parameters of models and estimation methods are shown in Table 1.

As our simulation studies with seven loci indicate, point estimates should be taken with caution (Supplemental Fig. S1). Consistent results across all models are that S. peruvianum has experienced a size expansion (i.e. q > 1) and is currently larger than S. chilense (at least twice the size). All models also require non-zero estimates of migration to explain the high amount of shared polymorphism between the two species. In the model which assumes no migration, extremely short divergence times are required to offset the lack of on-going migration (i.e. less than half of the divergence time as in the other models).

To evaluate the credibility of the tomato estimates we examined arrow plots with simulated data (Fig. 6). At the tail of each of these arrows are the values of m and τ which were used for the simulation of the dataset and which were to be estimated. The arrows point to the parameter combination of m and τ which were estimated by Jaatha. Therefore, the shorter the arrow, the better the estimation. These arrow plots are from analyses of 100 simulated datasets using the IbBin estimation method for both 7 and 1000 loci under the Growth Model with asymmetric migration. Arrows parallel to the migration rate axis indicate precise τ estimates with imprecise estimates for m (Fig. 6). These are frequent for τ < 0.05. With 1000 loci, divergence times are also difficult to estimate when τ and m are high. The estimates for the tomato data are located in a region of long arrows indicating

15
low certainty in parameter estimates in this range.

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 favored 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 row 1 of Table 2 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 to the distribution of log likelihood-ratios from the corresponding bootstrap datasets ($p < 0.006$ for Growth and $p < 0.001$ for Fraction-Growth Model, see also Supplemental Table S1).

**Discussion**

In this paper 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 $\partial a\partial i$) and, for low divergence times (e.g. 0.017-0.15 measured in $4N_1$ generations), Jaatha even outperformed other programs. Furthermore, although our method is based on the assumption of the independence of sites, its accuracy is not compromised when used on datasets of sufficiently many unlinked loci with limited or no within-locus recombination (Supplemental Figure S1). 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 estimation methods, Cent, wCent, wBin, and IbBin, which differ only in the final optimization step. When only few loci are available for analysis, the optimal method is wBin, providing a good compromise between run time
(<15 sec) and accuracy. For datasets with more loci, IbBin gives the best results and should be the method of choice.

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, since our simulation studies indicate that analyses based only on a limited number of loci (e.g. seven or fewer) prove to be challenging for accurate parameter estimation, we consider our parameter estimates from 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 splitting time between these two species is either 730,000 years if we assume one generation per year or $\sim$ 5.1 million years, if we assume a generation every 7 years. According to the best fitting model, the effective population size of S. chilense is $\sim$72,000. All models indicate that S. peruvianum is larger than S. chilense, consistent with the conclusions in Städler et al. (2005). Our estimated size ratio between these two species ranges from 1.66 to 8.66, 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 $\tau$ are needed to explain the present day differences in population sizes between these species, since 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.
All models estimate non-zero migration rates, indicating that some gene flow was likely following the initial divergence between these species. Our parametric bootstrap analysis indicates that a model with population growth in *S. peruvianum* as well as post divergence migration is needed for good fit of the model to the tomato data. Therefore, we could reject the hypothesis that no migration occurred after the split of these two species. This was a surprising finding, since 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).

Since our simulation studies show a remarkable improvement in accuracy when the number of loci is increased, we aim to develop and analyze a much larger dataset for this pair of tomato species (Supplemental Figure S4). 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 dataset is the sampling regime as discussed in 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 datasets, and this assumption is known to be violated in the dataset 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 $\partial a \partial i$ 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 $\theta_1$ 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 $\theta_1$. 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 which 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); Leuenberger & Wegmann (2010); Wegmann et al. (2009)). 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 datasets 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” datasets do not contain sufficient information. Thus, it is questionable whether one should try to estimate more than four parameters from such datasets and whether it is worthwhile to apply sophisticated and runtime-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 datasets 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.

**Acknowledgments**

We thank Ryan Gutenkunst for helping to run \(\partial a \partial i\) and the DFG Forschergruppe FOR1078, especially Peter Pfaffelhuber and Joachim Hermisson for fruitful discussions. This work
was supported by the German Research Foundation (DFG) grant ME 3134/3-1 to LR and DM.

References


List of Figures

Fig. 1: The different demographic models for populations $P_1$ and $P_2$ used for the simulation study where $\theta = \text{population mutation parameter}$, $m = \text{number of migrating individuals relative to } P_1 \text{ in } 4N_1 \text{ generations}$, $q = \text{size ratio between } P_2 \text{ and } P_1$, and $\tau = \text{the divergence time measured in } 4N_1 \text{ generations}$.

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

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

Fig. 4: Ratio of estimated to true values by $\partial a \partial i$ and Jaatha ($C=\text{Cent}$, $wC=w\text{Cent}$, $B=w\text{Bin}$, $IbB=Ib\text{Bin}$) of four parameters across models and methods for 100-loci scenario.

Fig. 5: The values estimated by Jaatha $w\text{Bin}$ ($\circ$), IM ($\times$ for ESS $>100$), and $\partial a \partial i$ ($\triangle$) of A) divergence time and B) migration plotted against true values for the 100-loci scenario of the Constant Model where true $\tau < 0.3$.

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 using $Ib\text{Bin}$ method. 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 $\tau$ is estimated precisely but $m$ is not. The large
open circle is the estimated value for the tomato data under this model using the method IbBin.
Figure 1:
Figure 2:
Figure 3:
Figure 4:
A Divergence Time

B Migration Rate

Figure 5:
Figure 6:
Table 1: **Estimates for parameters of models fitted to tomato data** - Estimates for the parameters ($\theta_1$ per locus, $q$ size ratio between *S. chilense* and *S. peruvianum*, $m$ symmetric migration rate, $\tau$ divergence time) using the Cent, wCent, wBin, and IbBin methods. In parentheses are the 95% confidence intervals estimated using a parametric bootstrap approach. The log-likelihoods (bottom rows) indicate that the fixedTau Model fits best while the Constant Model is the worst.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimation Method</th>
<th>Constant</th>
<th>Growth</th>
<th>Fraction-Growth</th>
<th>noMig</th>
<th>fixedTau</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>Cent</td>
<td>9.41</td>
<td>10.26</td>
<td>12.56</td>
<td>13.34</td>
<td>13.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(7.74-10.98)</td>
<td>(8.58-12.02)</td>
<td>(9.76-14.96)</td>
<td>(11.04-15.55)</td>
<td>(12.41-18.60)</td>
</tr>
<tr>
<td></td>
<td>wCent</td>
<td>9.33</td>
<td>10.19</td>
<td>12.56</td>
<td>13.34</td>
<td>13.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(7.57-11.03)</td>
<td>(8.57-11.72)</td>
<td>(9.53-15.41)</td>
<td>(11.21-15.24)</td>
<td>(11.73-17.00)</td>
</tr>
<tr>
<td></td>
<td>wBin</td>
<td>9.41</td>
<td>10.30</td>
<td>12.56</td>
<td>13.34</td>
<td>12.22</td>
</tr>
<tr>
<td></td>
<td>(7.69-11.12)</td>
<td>(8.62-11.68)</td>
<td>(9.72-14.73)</td>
<td>(10.47-16.05)</td>
<td>(9.70-14.64)</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>IbBin</td>
<td>10.08</td>
<td>9.68</td>
<td>12.56</td>
<td>13.46</td>
<td>13.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(8.77-12.42)</td>
<td>(7.76-11.18)</td>
<td>(9.87-14.89)</td>
<td>(10.61-16.88)</td>
<td>(11.23-16.56)</td>
<td></td>
</tr>
<tr>
<td>Cent</td>
<td>1.88</td>
<td>3.77</td>
<td>3.77</td>
<td>7.57</td>
<td>3.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.49-2.33)</td>
<td>(2.49-5.48)</td>
<td>(2.77-4.96)</td>
<td>(5.41-12.13)</td>
<td>(2.55-5.23)</td>
<td></td>
</tr>
<tr>
<td>wCent</td>
<td>1.88</td>
<td>3.93</td>
<td>3.77</td>
<td>7.57</td>
<td>4.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.45-2.39)</td>
<td>(2.77-5.74)</td>
<td>(2.73-5.10)</td>
<td>(5.76-11.54)</td>
<td>(3.35-5.74)</td>
<td></td>
</tr>
<tr>
<td>q</td>
<td>1.83</td>
<td>4.24</td>
<td>4.29</td>
<td>8.67</td>
<td>4.94</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.41-2.31)</td>
<td>(2.44-5.72)</td>
<td>(3.09-6.63)</td>
<td>(5.68-15.47)</td>
<td>(3.52-7.44)</td>
<td></td>
</tr>
<tr>
<td>wBin</td>
<td>1.66</td>
<td>4.55</td>
<td>4.26</td>
<td>8.66</td>
<td>4.64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.20-1.96)</td>
<td>(3.26-7.29)</td>
<td>(3.10-6.51)</td>
<td>(5.90-15.95)</td>
<td>(3.12-6.62)</td>
<td></td>
</tr>
<tr>
<td>IbBin</td>
<td>0.23</td>
<td>0.56</td>
<td>0.56</td>
<td>0</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.02-0.86)</td>
<td>(0.21-2.15)</td>
<td>(0.33-0.91)</td>
<td>(0.29-1.28)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>0.34</td>
<td>0.42</td>
<td>0.56</td>
<td>0</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.07-1.30)</td>
<td>(0.12-1.27)</td>
<td>(0.19-0.93)</td>
<td>(0.30-1.21)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wBin</td>
<td>0.36</td>
<td>0.36</td>
<td>0.73</td>
<td>0</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cent</td>
<td>wCent</td>
<td>wBin</td>
<td>IbBin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>-----------</td>
<td>-------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\tau)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.10-1.42)</td>
<td>(0.10-2.31)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.05-1.52)</td>
<td>(0.12-0.82)</td>
<td>(0.44-1.70)</td>
<td>(0.13-0.73)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.36</td>
<td>0.36</td>
<td>0.90</td>
<td>0.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>0.15</td>
<td>0.36</td>
<td>0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\tau)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.07-2.23)</td>
<td>(0.07-3.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.14-0.92)</td>
<td>(0.41-1.25)</td>
<td>(0.11-0.21)</td>
<td>(0.08-3.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.47</td>
<td>0.40</td>
<td>0.78</td>
<td>0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.23</td>
<td>0.25</td>
<td>0.33</td>
<td>0.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.23-0.81)</td>
<td>(0.23-0.76)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.05-0.28)</td>
<td>(0.06-0.33)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.25-0.76)</td>
<td>(0.41-1.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.05-0.28)</td>
<td>(0.07-2.23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Cent</td>
<td>wCent</td>
<td>wBin</td>
<td>IbBin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-188.01</td>
<td>-123.45</td>
<td>-101.58</td>
<td>-133.06</td>
<td>-96.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-189.51</td>
<td>-119.70</td>
<td>-101.58</td>
<td>-133.06</td>
<td>-93.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-189.51</td>
<td>-119.70</td>
<td>-101.58</td>
<td>-133.06</td>
<td>-93.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-183.82</td>
<td>-118.15</td>
<td>-97.47</td>
<td>-132.66</td>
<td>-92.85</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(0.20-0.78)</th>
<th>(0.13-0.80)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log-likelihood</td>
<td>(0.17-0.72)</td>
<td>(0.10-0.54)</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Model compared to <em>noMig</em></th>
<th>Estimation method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cent</td>
</tr>
<tr>
<td><em>Constant</em></td>
<td>-54.94</td>
</tr>
<tr>
<td><em>Fraction-Growth</em></td>
<td>31.48</td>
</tr>
</tbody>
</table>