Efficient Methods for the Estimation of Demographic Parameters from Population Genetic Data – with an Application to Wild Tomatoes

Lisha Naduvilezhath

München 2012

The Jaatha logo on the title page was designed by Brintha Antony.

Efficient Methods for the Estimation of Demographic Parameters from Population Genetic Data – with an Application to Wild Tomatoes

Lisha Naduvilezhath

Dissertation
an der Fakultät für Biologie der
Ludwig-Maximilians-Universität
München

vorgelegt von
Lisha Naduvilezhath
aus Palai, Indien

München, den 13.06.2012

Erstgutachter: Prof. Dr. Dirk Metzler
Zweitgutachter: Prof. Dr. Wolfgang Stephan

Tag der Abgabe: 13.06.2012
Mündliche Prüfung am: 18.07.2012

ERKLÄRUNG

Diese Dissertation wurde im Sinne von §12 der Promotionsordnung von Prof. Dirk Metzler betreut. Ich erkläre hiermit, dass die Dissertation nicht einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich nicht anderweitig einer Doktorprüfung unterzogen habe.

EIDESSTATTLICHE VERSICHERUNG

Ich versichere ferner hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbstständig und ohne unerlaubte Hilfe angefertigt worden ist.

München, den 13.06.2012 _____

Lisha Naduvilezhath

DECLARATION OF CONTRIBUTIONS AS A CO-AUTHOR

In this thesis, I present the results of my doctoral studies conducted from January 2009 to June 2012. The results are presented in three chapters, all of which are the product of collaborations with other scientists. The work of this doctoral thesis has resulted in two publications. They constitute Chapters 1 and 2 of this dissertation and are supplemented by appendices A and B. Chapter 3 (with appendix C) is a manuscript which is to be submitted.

Chapter 1: The study was designed by Dirk Metzler, Laura Rose, and me. Dirk Metzler drafted the first version of Jaatha which I further refined. All the simulations and analyses were performed by me. With the assistance of Dirk Metzler and Laura Rose I wrote the manuscript. This chapter has been published in *Molecular Ecology*.

Chapter 2: Aurélien Tellier, Peter Pfaffelhuber, Bernhard Haubold, Thomas Städler, Wolfgang Stephan, and Dirk Metzler conceived and designed the experiments. I assisted on the design and analyses of the experiment in which Jaatha and $\partial a \partial i$ were compared. Aurélien Tellier, Peter Pfaffelhuber, Dirk Metzler, and I performed the experiments. Aurélien Tellier and Dirk Metzler analyzed the data. Aurélien Tellier, Peter Pfaffelhuber, Bernhard Haubold, Laura Rose, Thomas Städler, Wolfgang Stephan, Dirk Metzler, and I wrote the paper. Peter Pfaffelhuber, Bernhard Haubold, Dirk Metzler, and I designed the software used in analysis. This chapter has been published in *PLoS ONE*.

Chapter 3: I designed and implemented the new version of Jaatha with guidance from Dirk Metzler. I conducted all the simulation studies with the new Jaatha version. Florian Ruhland assisted me with the runs of the first version of Jaatha. I performed the analyses and wrote the manuscript with Dirk Metzler, Laura Rose, and Paul Staab. Paul Staab and I created the R package. The manuscript is about to be submitted.

Lisha Naduvilezhath	Prof. Dr. Dirk Metzler

Table of Contents

List of Figures

List of Tables

To my family

Summary

We developed an algorithm called Jaatha, Joint site frequency spectrum associated approximation of the ancestry. Jaatha estimates parameters for a user-defined demographic model of two recently diverged populations. As input it requires single nucleotide polymorphism (SNP) data and an outgroup sequence. Jaatha is designed as a fast composite-likelihood method that approximates the likelihood function. For simplification, SNPs are assumed to be unlinked and the sequence data is reduced to a set of summary statistics instead of taking the full-data. Even when the assumption of independent SNPs is violated, we demonstrate that Jaatha estimates parameters with the same accuracy as other methods. Under certain conditions it is superior, especially if divergence occurred very recently.

The demand for a new method to estimate demographic parameters was motivated by a data set of *Solanum chilense* and *S. peruvianum*. The recently diverged wild tomato species live in diverse habitats in South America: While *S. peruvianum* prefers more mesic environments, *S. chilense* can be found on extremely dry and saline soils and at high altitudes of up to 4,000 meters. An important question is which genomic regions are involved for adaptation to such harsh conditions. These regions are also of economical value for plant breeders because traits from wild relatives such as disease resistance have been successfully crossed into the cultivated tomato *S. lycopersicum*, which has resulted in a yield increase of up to five fold. Previous studies pointed out that demography has to be accounted for, when searching for these regions at the molecular level. As an example we analyzed the demographic history of *S. chilense* and *S. peruvianum*, the precondition to detecting selection.

Since the great majority of available methods for demography estimation were not suitable for the *Solanum* data, the goal of this dissertation was to develop a method that could deal with the challenges posed by the data: signs of population expansion, gene flow between species and recent divergence, presence of high number of sites with multiple mutations, and a high within-locus recombination rate. When Jaatha was applied to the *Solanum* data set, an approximate divergence time of the two species ($95\,\%$ bootstrap

confidence interval) of 2.0 $(0.9, 3.9)$ million years was obtained when an average generation time of three years was assumed[1]. The divergence time coincides with a change in climate in the Central Andes around 3-5 million years ago which is known to have created a unique environment for rapid plant diversification. Furthermore, our results indicate the present-day effective (breeding) population size of *S. chilense* to be $\approx 73{,}000$. The current effective population size of *S. peruvianum* is at least four times larger than that of *S. chilense* and this species has expanded after speciation. With a simulation-based likelihood ratio test approach, we find significant evidence for gene flow following the split, even when multiple mutations per site are taken into account.

We provide insights into the demographic history of *S. chilense* and *S. peruvianum*, and furthermore conclude that summary statistic based methods such as Jaatha are a promising alternative to resource-demanding full-data methods, especially with the advent of novel sequencing technologies. The composite-likelihood estimators implemented in Jaatha are consistent, *i.e.* the approximation will improve with more examined genetic regions. Next generation sequencing strategies have recently made it possible to obtain whole-genome data sets of different species. Since Jaatha has been shown to be a robust and flexible method that can easily adapted to other scenarios, we anticipate that it will be widely applied and extended. A promising step for future research is to jointly estimate selection and demographic parameters.

[1]The exact generation time is difficult to determine because seed germination and fecundity are affected by the climatic patterns typical to the region of occurrence.

Zusammenfassung

Wir haben einen Algorithmus namens Jaatha entwickelt (Joint site frequency spectrum associated approximation of the ancestry), der populationsgenetische Parameter für ein benutzerdefiniertes Modell zweier nah verwandter Arten schätzt. Als Eingabe benötigt Jaatha Daten in Form von Single-Nucleotide-Polymorphismen (SNPs) und eine Außengruppensequenz. Jaatha ist eine schnelle Composite-Likelihood-basierte Methode, d.h. die zuoptimierende Likelihoodfunktion wird vereinfacht, indem die SNPs als ungekoppelt angenommen werden. Außerdem werden die Sequenzdaten zu Statistiken zusammengefasst. Wir zeigen, dass Jaatha mit der gleichen Präzision Parameter schätzt wie andere Methoden, selbst wenn die Annahme unabhängiger SNPs verletzt ist. In bestimmten Fällen liefert Jaatha genauere Parameterschätzungen, etwa, wenn die Divergenz der beiden Populationen erst vor kurzem stattgefunden hat.

Als Beispieldatensatz für die neue Methode dienten die SNP-Daten zweier nah verwandter Wildtomatenarten, *Solanum chilense* und *S. peruvianum*. Ihre Habitatpräferenzen sind zum Teil unterschiedlich: Während *S. peruvianum* eher mesische Regionen bevorzugt, wächst *S. chilense* auch auf sehr trockenen und salzhaltigen Böden. Welche genomischen Regionen zur Anpassung an solche extremen Umweltbedingungen beitragen, ist eine wichtige Fragestellung mit Anwendungsmöglichkeiten in der Pflanzenzüchtung. Da Demographie und Selektion ähnliche Muster in den Sequenzdaten erzeugen können, müssen demographische Effekte in Selektionsanalysen berücksichtigt werden.

Die meisten der vorhandenen Methoden zur Demographieschätzung lassen sich jedoch nicht auf unseren Beispieldatensatz anwenden. Daher war es Ziel dieser Arbeit eine Methode zu entwickeln, die sowohl Populationswachstum, als auch Genfluss, multiple Mutationen an einer Position, kurze Divergenzzeiten und Rekombination innerhalb eines Gens berücksichtigen kann. Für *S. chilense* und *S. peruvianum* berechnete Jaatha eine ungefähre Divergenzzeit von 2.0 Millionen Jahren bei einer Generationszeit von drei Jahren. Das bedeutet, die Divergenz fällt in die Zeit nach einer Klimaänderung in den Zentralanden vor ca. 3-5 Millionen Jahren, die günstige Bedingungen für rasche Artbil-

dung schuf. Für die effektiven Populationsgrößen von *S. chilense* und *S. peruvianum* wurden Werte von $\approx 72,000$ und $\approx 288,000$ Individuen geschätzt, wobei die Population von *S. peruvianum* seit der Divergenz gewachsen ist. Des Weiteren konnten wir mit einem Likelihood-Quotienten-Test zeigen, dass es signifikante Anzeichen für Genfluss zwischen den Arten gibt.

Mit unseren Ergebnissen konnten wir nicht nur die Demographie entschlüsseln, sondern auch zeigen, dass schnelle Composite-Likelihood-basierte Methoden eine vielversprechende Alternative zu langsamen exakten Methoden darstellen. Dies gilt vor allem, wenn eine große Anzahl an genomischen Regionen untersucht wird, was durch neue Sequenziertechnologien immer häufiger der Fall ist. Da sich Jaatha durch Robustheit und Flexibilität auszeichnet, kann es einfach an andere Szenarien und Arten angepasst werden. Deshalb erwarten wir, dass Jaatha für viele zukünftige Studien von großem Nutzen sein wird.

General Introduction

Observing similarities and differences in traits in different individuals, such as height (Yang *et al.*, 2010) or language (Barbieri *et al.*, 2012) in humans, may help to elucidate their relatedness and for example identify parents and their offspring. In population genetics the same idea is used: Based on observed differences in deoxyribonucleic acid (DNA) sequences, *e.g.* single nucleotide polymorphisms (SNPs) or *segregating sites*, the (evolutionary) history of a sample is inferred – in fact of its entire population.

In the early twentieth century, still a few decades before the structure of DNA was proposed (Watson and Crick, 1953), scientists started describing changes in frequencies of mutations over time and exploring several factors contributing to these changes (*e.g.* Fisher, 1930; Wright, 1931; Haldane, 1932)[2]. The randomness acting on the frequency of a mutation due to the finite size of the population is termed *genetic drift*. The smaller the population is, the stronger are the effects of drift. Another force that can alter the frequency of a polymorphism is natural selection. *Positive selection* on a certain mutation may lead to an increase in its frequency because it conveys a benefit for the individual carrying it. If positive selection is strong enough the mutation and surrounding regions will eventually fix in the population, such that sequence variation around that position will be low in the population (*selective sweep*, Smith and Haigh, 1974). The main factor acting against the spread of the mutation especially in its initial stages is genetic drift.

When searching genomes for traces of selection (*e.g.* for selective sweeps as in Parsch *et al.*, 2001; Beisswanger *et al.*, 2006; Andersen *et al.*, 2012) we run into the following problem: selection can leave similar signatures on DNA as demographic changes (Robertson, 1975; Andolfatto and Przeworski, 2000; Teshima *et al.*, 2006; Siol *et al.*, 2010). For example, a decrease in the population size leads to a reduction in the observed number of differences in the population, similar to a selective sweep. One commonly used approach is to look for selection after or simultaneously accounting for *demography* (e.g. in

[2]Only in 1943 first evidence arose that the genetic information inherited to the next generation was coded in DNA (Avery *et al.*, 1944).

pearl millet Clotault *et al.*, 2012; Williamson *et al.*, 2005). The idea is that demography should affect the entire genome while selection should act locally on specific parts (Wakeley, 2008[3]). A *demographic model* is a mathematically-tractable idealization of processes that the population experienced in the past. It may include estimates of effective population sizes, different numbers of populations, population size changes, the times and severities of these, rates of gene flow[4] between species, and possible population splitting (divergence) events. The concept of *effective population size*, N_e, was first introduced by Wright and can be understood as the number of "breeding" individuals (Wright, 1931[5]) or, more formally, as the number of individuals in an idealized constant-sized population that would experience the same amount of genetic drift (reviewed in Charlesworth, 2009). The effective population size is usually smaller than the consensus size. However, there are several definitions of N_e depending on the aspect under consideration (for definitions see Durrett, 2008). N_e is usually part of a compound parameter, $\theta = 4N_e\mu$, where μ is the mutation rate per generation per locus and N_e is the diploid size of the population (sometimes also per site or with a coefficient of 2 instead of 4). In theoretical population genetics, often the limits $N_e \to \infty$ or $\mu \to 0$ are considered.

Besides being indispensable for identifying the action of non-neutral forces, modeling demography can complement archaeological findings (*e.g.* Xing *et al.*, 2010) and give valuable insights to conservation biologists (Fernández *et al.*, 2005; Lin *et al.*, 2012). The major goal of population genetics is to understand the interactions between selection and genetic drift, population structure, varying mutation and recombination rates, and gene flow. How these factors contribute to the birth of new species, a process termed *speciation*, and adaptation to new environments is of particular interest (Jones *et al.*, 2012; Andrew *et al.*, 2012) and has been debated intensively (*e.g.* Fitzpatrick *et al.*, 2009).

Speciation With and Without Gene Flow

During a time of geographical isolation of two populations, through *e.g.* rise of a mountain, the two populations can accumulate genetic differences which can lead to genetic incompatibility (*allopatric speciation*). This process can be characterized as a *snowball effect* (Orr and Turelli, 2001; Städler *et al.*, 2012). In the *isolation-with-migration* model it is assumed that two species diverged from each other and may share genetic material

[3] p. 128
[4] *Gene flow* or *gene migration* describes the exchange of genetic material between populations.
[5] pp. 110-111

by gene flow through individuals *migrating* from one population to the other (Hey and Nielsen, 2004). Gene flow counteracts this genetic isolation of the species and homogenizes the genetic differences that have been accumulating. Although only small levels of gene flow are necessary to prevent speciation (Wright, 1931), there is a growing number of examples proposing that speciation can happen in the presence of gene flow (*sympatric speciation*) or introgression[6]: in sticklebacks (Jones *et al.*, 2012), in butterflies (The Heliconius Genome Consortium, 2012), in sunflowers (Strasburg and Rieseberg, 2008), in water fleas (Cristescu *et al.*, 2012). Other examples are reviewed in Arnold (2004), Smadja and Butlin (2011), or Marie Curie SPECIATION Network (2012).

Introgression can confer adaptations to new environments as has been demonstrated in the plant system *Iris fulva* and *I. hexagona* (Arnold and Bennett, 1993). The authors demonstrated that the more genetic markers of the shade tolerant *I. fulva* the hybrids contained, the better the hybrids could adapt to lower levels of light. Another consequence of increased gene flow is that it can lead to the extinction of species through the formation of hybrid swarms (Rhymer and Simberloff, 1996). Such a reversed speciation event caused by anthropogenic eutrophication has been reported in the whitefishes (Vonlanthen *et al.*, 2012).

The main challenge when accessing levels of gene flow is to separate it from ancestral variation (Hey, 2006). Additionally, violation of model assumptions can cause false signals of gene flow (Becquet and Przeworski, 2009). Furthermore, simulation studies have shown that estimating the timing of migration precisely is difficult as well (Becquet and Przeworski, 2009; Strasburg and Rieseberg, 2011). Sousa *et al.* (2011) explain these simulation results in terms of the probability of genealogies[7]. They demonstrate that two genealogies with distinct migration timings can have the same probability under a model with migration (*posterior probability*).

The Wild Tomatoes *Solanum chilense* and *S. peruvianum*

The *Solanum* species complex has been under intensive investigation because it encompasses various species which are of agricultural importance, *e.g.* potato *Solanum tuberosum*, eggplant *S. melanogena*, and tomato *S. lycopersicum*. Recently the tomato draft genome of the cultivated tomato *S. lycopersicum* cultivar 'Heinz 1706' as well of its wild relative *S. pimpinellifolium* has been reported to be complete (The Tomato Genome Con-

[6] *Introgression* describes the exchange of genetic material between a hybrid and its parental species (Arnold, 2004).

[7] A *genealogy* describes the ancestral process of the samples.

sortium, 2012). For plant breeders the wild tomatoes are a precious source of advantageous traits such as pathogen resistance. Since the 1940's breeders have successfully incorporated useful traits into the cultivated tomato and have thus increased the yield 4 to 5-fold (Rick and Chetelat, 1995). In this project we investigated the demography of two other wild relatives of the domesticated tomato, *S. chilense* and *S. peruvianum*.

Close to the border of Chile and Peru, including the South American Atacama desert, one of the driest areas on earth (Navarro-González *et al.*, 2003), the two wild tomato species *S. chilense* and *S. peruvianum* co-occur (Chetelat *et al.*, 2009; Städler *et al.*, 2005). *S. peruvianum's* habitat range extends further into the north and *S. chilense* further into the south. The species are phenotypically and genetically similar and appear to be recently diverged (Rick *et al.*, 1979; Städler *et al.*, 2005, 2008). Ecologically they are differentiated: *S. peruvianum* prefers more mesic environments of lower elevations up to 2,500 m, while *S. chilense* is a generalist and inhabits all elevations up to 4,000 m, dry, and wet habitats (Chetelat *et al.*, 2009). These characteristics make *S. chilense* and *S. peruvianum* an ideal system to study selection on pathogen resistance or drought tolerance (Xia *et al.*, 2010; Rose *et al.*, 2011).

Previous results indicated that *S. chilense* and *S. peruvianum* have large effective population sizes, high recombination rates (Arunyawat *et al.*, 2007), and a divergence time of ≤ 0.55 million years when analyzed with a model without gene flow (Städler *et al.*, 2008), although there were indications of gene flow after the split of the species (Städler *et al.*, 2008) and population expansion (Arunyawat *et al.*, 2007). Further in the data set of Arunyawat *et al.* (2007) and Städler *et al.* (2008), which we used in this study, several positions are hit by multiple mutations ($> 7\%$ of SNPs).

Methods for Inferring Demography

To estimate *demographic parameters* from polymorphisms in selectively neutral sequences, several samples from a population are needed. Major advances in the last 60 years in computer science, biology, and related fields from the description of the DNA structure in 1953 by Watson and Crick (Watson and Crick, 1953), through the development of sequencing techniques such as Sanger sequencing (Sanger *et al.*, 1977), next generation sequencing (Brenner *et al.*, 2000), up to the new pyrosequencing methods (Margulies *et al.*, 2005) have now made it possible to compare several genomes with each other. Whole genome data sets are being reported for several species, including human (1000 Genomes Project Consortium, 2010), mouse (Keane *et al.*, 2011), *Drosophila* (Begun *et al.*, 2007),

Arabidopsis (Cao *et al.*, 2011), or *Eschericia coli* (Lukjancenko *et al.*, 2010).

With these great number of genomic regions at hand, approximate methods, in particular the ones which use simplifications of the data (*e.g.* composite-likelihood methods which are introduced later), are promising. Although full-data methods approximate the likelihood more precisely, they are usually restricted to small amounts of data (Beerli and Felsenstein, 2001).

After briefly describing sequence evolution models, I will give an overview of methods for demography estimation starting from the ones that consider the full data set to methods that simplify the data by using summary statistics (explained latter; for a comprehensive overview on general population genetic software see Excoffier and Heckel, 2006). The latter class encompasses diffusion-approximation-based approaches, composite likelihood methods, and a group of methods called approximate Bayesian computation (ABC). The programs will be introduced in the light of being useful for the demography analysis of *S. chilense* and *S. peruvianum* described above (an overview is given in Tab. 1).

Table 1: **Overview of main methods for the estimation of demographic parameters and their features.**

	Software	Data summary	Sequence evolution[a]	Recent speciation	Gene flow	Recombination	Reference
single population	Bayesian skyline plots	full	FS	✗	✗	✗	Drummond *et al.*, 2005
	Msvar	full	SMM	✗	✗	✗	Storz and Beaumont, 2002
n-island	GENETREE	full	IS	✗	✓	✗	Bahlo and Griffiths, 2000
	MIGRATE-N	full	FS/IS	✗	✓	✓	Beerli, 2006
two-population speciation on full-data	IM, IMa, IMa2, Choi	full	FS/IS	✓	✓	✗	Choi and Hey, 2011
	LAMARC	full	FS/IS	✗	✓	✓	Kuhner, 2006
	CoalHMM	full	FS/IS	✓	✗	✓	Mailund *et al.*, 2011
	BATWING	full	FS/IS	✓	✗	✗	Wilson *et al.*, 2003
two-population speciation on SS	ABC[b]	SS	IS	✓	✓/✗	✓	Beaumont and Balding, 2004
	MIMAR	SS	IS	✓	✓	✓	Becquet and Przeworski, 2007
	Garrigan	JSFS	IS	✓	✗	∞	Garrigan, 2009
	$\partial a \partial i$ [c]	JSFS[d]	IS	✓	✓	∞	Gutenkunst *et al.*, 2009
	Chen	JSFS[d]	IS	✓	✗	∞	Chen *et al.*, 2007

[a] FS = finite-sites model, IS = infinite-sites model, SMM = stepwise mutation model for microsatellites
[b] ABC methods can also be used for demographic models other than two-population speciation models.
[c] $\partial a \partial i$ can be used for up to three populations.
[d] solve diffusion approximation numerically

Sequence Evolution Models for DNA Data

Two types of sequence evolution models are distinguished, the *infinite-sites model* (ISM) and *finite-sites models* (FSM). The ISM assumes that each mutation affects a new site (Kimura, 1969; Watterson, 1975). Thus all mutations the sequence has ever experienced are visible in an ISM data set. In contrast, FSM allow for multiple mutations (also called recurrent mutations), including back mutations to the ancestral nucleotide on a finite number of sites. Substitution models of different complexities described by different numbers of parameter are available, ranging from one (Jukes-Cantor model of Jukes and Cantor, 1969) to ten (general time reversible model, GTR, of Tavaré, 1986). Due to the finite length of the sequences analyzed in biological studies FSMs should be more appropriate but only the ISM allows for analytical solutions. Methods which analyze the full data can in general handle FSMs, while most methods based on summarizing the data assume the ISM.

Mutation rate variation between sites within a sequence can be modeled with Γ distributions (parameterized by the shape parameter α; the scale parameter is set to $1/\alpha$, Yang, 1996). Neglecting the variation in mutation rate, or back and multiple mutations can have dramatic effects on confidence intervals and parameter estimation since it can lead to similar patterns in the DNA as population expansion and should thus be accounted for (Lundstrom *et al.*, 1992; Aris-Brosou and Excoffier, 1996; Schneider and Excoffier, 1999).

Full-data Likelihood Methods

The *likelihood* L of a set of parameters $\boldsymbol{\lambda}$, *i.e.* the probability of the data $\mathcal{D}$ given $\boldsymbol{\lambda}$, is defined as

$$L(\boldsymbol{\lambda}) = P(\mathcal{D}|\boldsymbol{\lambda}) = \int_{\mathcal{G}} P(\mathcal{D}|\mathcal{G})P(\mathcal{G}|\boldsymbol{\lambda})d\mathcal{G}, \tag{1}$$

where $\mathcal{G}$ is a genealogy which includes the branching pattern and the branch lengths (coalescent times, Hey, 2006). To evaluate the first term of the product $P(\mathcal{D}|\mathcal{G})$, which describes the probability of the data given a genealogies, an adequate sequence evolution model needs to be defined; for the second term $P(\mathcal{G}|\boldsymbol{\lambda})$, which is the probability for the genealogy given the parameter values, a model for genealogies is needed such as the coalescent model[8] (Hey, 2006). Since $\mathcal{G}$ is an unknown nuisance variable, we have to

[8]The *coalescent model* is the stochastic limit process for many population models (*e.g.* Wright-Fisher model) which traces back the ancestry of samples to the most recent common ancestor (Wakeley, 2008,

integrate over it (Felsenstein, 1988; Hey, 2006). Since the space of all possible genealogies is not feasible to sample in reasonable time, approximative methods are needed. Two such approaches are available: *Importance Sampling* (*e.g.* Griffiths and Tavaré, 1995) and *Markov chain Monte Carlo* (Metropolis *et al.*, 1953; Hastings, 1970; Wakeley, 2008[9]). (A third, less common method with a hidden Markov model (CoalHMM) will be explained at the end of this section.)

Importance Sampling and MCMC approximate the likelihood with randomly simulated genealogies that are most likely to have produced the observed data $\mathcal{D}$ (Wakeley, 2008[9]). Both methods can also be combined as in Griffiths and Tavaré (1994). The random sampling of a sampling space and subsequent averaging is termed *Monte Carlo integration* and can be included into both frameworks (*e.g.* Hey and Nielsen, 2007). The difference between the two approaches is that the sampled genealogies in the Importance Sampling method are in most cases uncorrelated while in the MCMC case the genealogies are correlated (Wakeley, 2008[9]). For Importance Sampling, a probability function has to be defined (*proposal distribution*) from which genealogies that may have produced the data $\mathcal{D}$ will be sampled (Wakeley, 2008[10]). It can be difficult to define such a proposal distribution, but has been successfully done for the ISM by Griffiths and colleagues and implemented in GENETREE (Bahlo and Griffiths, 2000). All the other methods described in this section (except CoalHMM) make use of the computationally intensive MCMC method. In the MCMC method, a *Markov chain*[11] is designed that will have its stationary distribution at $P(G|\boldsymbol{\lambda})$ and from which genealogies will be sampled (Wakeley, 2008[12]). The disadvantage of MCMC is that a large number of steps until the Markov chain reaches its equilibrium (*burn-in*) have to be discarded and, due to the correlation in successive genealogies, only the ones sampled in intervals are kept. However, the main advantage of this method is that it is flexible regarding the mutation model.

For estimating past effective population sizes of a single population under an MCMC framework Msvar (Storz and Beaumont, 2002) and "skyline plots" (*e.g.* Drummond *et al.*, 2005) are available. Msvar is often applied in evolutionary or conservation biology because it analyzes *microsatellite*[13] data (*e.g.* Elmer *et al.*, 2009; Lin *et al.*, 2012). An ad-

p. 59). It was first described mathematically by Kingman (Kingman, 1982).

[9] pp. 252-270

[10] pp. 270-283

[11] A *Markov chain* is a stochastic, discrete-time process on a (discrete-)state space. The chain possesses the *Markov property*, *i.e.* its next state only depends on the current state. (Wakeley, 2008, pp. 134-135)

[12] pp. 283-289

[13] *Microsatellites* are repetitive sequences with a motif length of up to six bases (Goldstein and Schlötterer, 1999, pp. 1-2).

vantage of Bayesian skyline plots is that besides visualizing fluctuations in the ancestral population size with their uncertainties, they can estimate parameters of substitution models. Bayesian skyline plots are implemented in the software package Beast (Drummond *et al.*, 2005).

GENETREE (Bahlo and Griffiths, 2000; Nath and Griffiths, 1996; Griffiths and Tavaré, 1994) and MIGRATE-N (Beerli and Felsenstein, 1999, 2001; Beerli, 2006) assume a demographic model with n subpopulations (n-island model) and are useful to estimate uneven migration rates and changes in subpopulation size. GENETREE assumes an ISM and can therefore give estimates for the age of mutations and for the time to the most recent common ancestor of all samples (Bahlo and Griffiths, 2000). MIGRATE-N estimates parameters including recombination rate from microsatellite or sequence data in a maximum likelihood or Bayesian framework (Beerli, 2006). The sequence data can be analyzed with ISM or FSM and can incorporate mutation rate heterogeneity as well.

For two diverging species a prominent group of methods developed by Hey and colleagues under the isolation-with-migration model include IM (Hey and Nielsen, 2004) (for locus specific migration rates see Won and Hey, 2005), IMa (Hey and Nielsen, 2007), IMa2 (Hey, 2010), and the recent modification of IMa2, a method by Choi and Hey (2011) including population assignment. In Hey and Nielsen (2007) an analytical result is exploited thus making IM**a** faster than IM but population size changes are no longer included in the model. Hey (2010) extended the framework to incorporate up to ten populations and in Choi and Hey (2011) to jointly estimate demography and population assignment. A limitation of these methods is that they neglect within-locus recombination, although it is important to consider *e.g.* when analyzing population genetic data in viruses (McVean *et al.*, 2002), humans (Jeffreys and May, 2004), *Drosophila melanogaster* (Kliman and Hey, 1993), or in the wild tomatoes, *S. chilense* and *S. peruvianum* (Arunyawat *et al.*, 2007; Naduvilezhath *et al.*, 2011, supplement). These species contain regions in which the recombination rate is of the same order of magnitude as the mutation rate or even higher and thus should not be ignored (McVean *et al.*, 2002; Arunyawat *et al.*, 2007). However, in a simulation study Strasburg and Rieseberg (2010) reduced the sequences to non-recombining blocks which eliminated most of the bias and improved the results.

Another method to calculate the likelihood from full data is implemented in LAMARC (Kuhner, 2006). It additionally estimates the recombination rate and can perform maximum-likelihood as well as Bayesian analyses (explanation of *Bayesian* follows latter). A main assumption of the method was that for the last $4N_e$ generations the population structure has been stable, such that it was not adequate for recently diverged species (Kuhner,

2006). Only very recently Kuhner and colleagues announced a new LAMARC version which now also enables multiple populations to diverge recently from a common ancestor in the Bayesian framework (LAMARC Team, 2012).

BATWING is a program that can be applied to DNA, SNP, or microsatellite data of one or more populations to estimate population histories, including population growth, which is assumed to have started in all populations simultaneously (Wilson *et al.*, 2003). But it does neither allow for gene flow nor within-locus recombination. Mailund *et al.* (2011) developed a hidden Markov model approach, CoalHMM, which models two neighboring sites along the alignment of two genomes to calculate the divergence times and recombination rates. Although the approach is limited to two genomes, it offers the possibility to include different mutation models (*e.g.* GTR model, Mailund *et al.*, 2011). But gene flow is not modeled in this approach.

The methods of Hey, Kuhner and others have the great advantage that they approximate the likelihood for the estimated parameters by taking the full data into account, and are shown to converge to the true parameter values if they run for a long time (Beerli and Felsenstein, 2001). Generally, these programs have run times of several weeks, sometimes even months until they converge. The implementations of these methods also need an adept programmer and a considerable amount of time. Further, the underlying demographic model is strictly defined and does not allow for much alteration to incorporate for example known biological information as bottlenecks during a certain time period. GENETREE and MIGRATE-N are further limited by the number of subpopulations and parameters they can handle and the results are sensitive to the choice of starting values of the Markov chain (Beerli and Felsenstein, 2001). Due to these limitations summary-statistics-based methods have gained an enormous increase in interest over the past decade.

Ways of Summarizing Sequence Data

The two-dimensional joint site frequency spectrum (JSFS) J summarizes homologous sequences from two populations, P_1 and P_2. It was first introduced by Li and Stephan (2006) and contains the counts (or frequencies) of a derived allele at a site in the sequence alignment in both populations. To determine whether a site is derived an *outgroup sequence* to the ones analyzed is needed. Any site that is different from the outgroup is regarded as *derived*. If $J[a, b] = j_{ab} = 5$, five positions in the examined data set are found for which the derived type is present in exactly a samples of P_1 and in b samples of P_2. How parameters of a demographic model effect the JSFS is shown in Figure 1. Numeri-

cal solutions for the expected JSFS using diffusion theory[14] have been provided by Chen *et al.* (2007) and Gutenkunst *et al.* (2009). Recently, Chen (2012) provided an analytical solution (based on coalescent theory) for a two-species JSFS without migration for small numbers of samples.

In the following, *summary statistics* (SS) are summarizations of the observable variation in the genetic data. On the JSFS a set of SS $\mathbf{S} = (S_1, \ldots, S_n)$ can be defined, where $S_i(J) = \sum_{(a,b)\in A_i} j_{ab}$ and $A_1, \ldots, A_n$ is a partition of $A = \{0, \ldots, m_1\} \times \{0, \ldots, m_2\} \setminus \{(0,0), (m_1, m_2)\}$ with m_i being the sample size of P_i. Wakeley and Hey distinguished four SS ($n = 4$), containing shared, fixed, and exclusive polymorphic sites in P_1 and in P_2 (Wakeley and Hey, 1997). Modifications of these were specified in Leman *et al.* (2005) and Becquet and Przeworski (2007). Other commonly used SS are Tajima's D (Tajima, 1989), Fu and Li's D and F (Fu and Li, 1993), F_{ST} (Wright, 1943), and the expected heterozygosity H_e (Nei, 1978).

Summary-statistics based Likelihood Methods

A large group of methods based on SS are approximate Bayesian computation (ABC) methods based on the seminal works of Tavaré *et al.* (1997), Pritchard *et al.* (1999), and Beaumont *et al.* (2002). ABC methods base their inferences on estimating the posterior distribution $P(\boldsymbol{\lambda}|\mathcal{D})$ (Beaumont and Rannala, 2004; Shoemaker *et al.*, 1999) by using

$$P(\boldsymbol{\lambda}|\mathcal{D}) = \frac{P(\boldsymbol{\lambda})P(\mathcal{D}|\boldsymbol{\lambda})}{P(\mathcal{D})}, \tag{2}$$

which is named after Reverend Thomas Bayes (1702-1761), the *Bayes' rule* (*e.g.* Wakeley, 2008[15]). $P(\boldsymbol{\lambda})$ is called the *prior distribution* and represents a strength of ABC methods. Into the prior, (biological) knowledge about the parameter values can be included. $P(\mathcal{D}|\boldsymbol{\lambda})$ is the likelihood which we have encountered before (eqn. 1). Since for many non-standard demographic models no analytical expression for the likelihood is available, ABC simulations are used instead. $P(\mathcal{D})$ is the probability of the data, independently of the parameters. Tavaré *et al.* (1997) replaced the full data $\mathcal{D}$ with a single SS, which was extended by Weiss and von Haeseler (1998) to multiple SS. In Pritchard *et al.* (1999) the

[14] In the case of estimating the JSFS *diffusion approximations* model the evolution of changes in allele frequencies through discrete generations under the assumptions of large population sizes, only small changes in allele frequencies, and independence of SNPs. A advantage of this approach is that selection can be included into the model as well (*e.g.* Gutenkunst *et al.*, 2009, supplement).

[15] p. 27

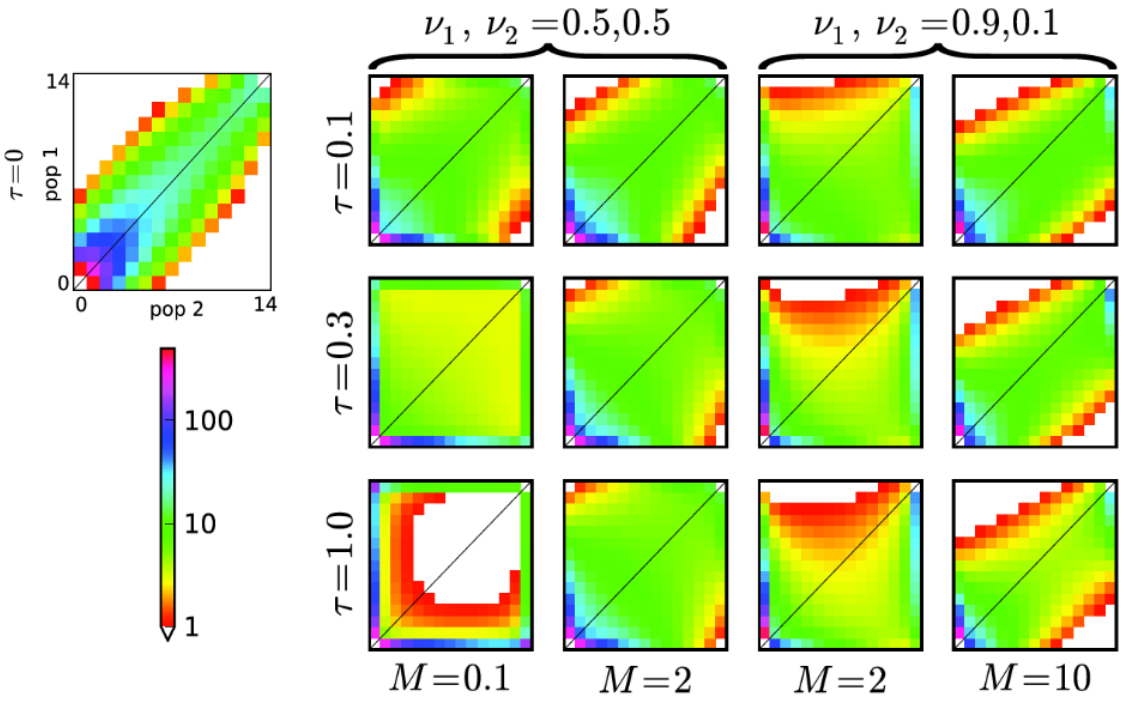


Figure 1: **Various demographic parameters influence the JSFS.** The number of sites in the JSFS of two diverging species is colored according to the scale for different population sizes, divergence time τ, and migration rate M. The demographic model assumes that an ancestral population of size N_A split into two populations $2N_A\tau$ generations ago into sizes $\nu_1 N_A$ and $\nu_2 N_A$. Both populations have not experienced any size change after the split. The JSFS is calculated with $\theta = 1000$ and the diffusion approximation of Gutenkunst *et al.* (2009) which implies that the SNPs are assumed to be independent. Every generation $M/(2N_A)$ individuals are replaced by migrants from the other population. With increasing M and decreasing τ the populations have more shared polymorphisms along the diagonal, while decreasing M and increasing τ yields more fixed differences along the axes of the matrices between the populations. This figure is a slightly modified version of Figure 1 in Gutenkunst *et al.* (2009).

following rejection-based method including Monte Carlo integration was implemented:

1. For the observed data $\mathcal{D}$ calculate SS.
2. Choose a tolerance level δ.
3. Simulate $\boldsymbol{\lambda}'$ from the prior $P(\boldsymbol{\lambda})$.
4. With the model and $\boldsymbol{\lambda}'$ simulate data $\mathcal{D}'$.
5. Calculate SS$'$ for $\mathcal{D}'$.
6. If the Euclidean distance between SS$'$ and SS is $\leq \delta$ accept $\boldsymbol{\lambda}'$.
7. Repeat steps 3-6 until a certain number of acceptances have occurred.

With this algorithm actually $P(\boldsymbol{\lambda} \mid ||SS' - SS|| \leq \delta)$ is sampled instead of $P(\boldsymbol{\lambda}|\mathcal{D})$. In the last decade many improvements of ABC have been developed, just to mention a few: local regression adjustments of SS (Beaumont *et al.*, 2002), identifying loci under selection (Beaumont and Balding, 2004), mathematical transformations (partial least squares, PLS) to choose informative SS (Wegmann *et al.*, 2009), general linear model adjustments of SS (Leuenberger and Wegmann, 2010). User friendly implementations have for example been implemented in DIY ABC (Cornuet *et al.*, 2008) or PopABC (Lopes *et al.*, 2009). The main advantage of ABC methods is that they can also be applied to demographic scenarios (with various number of populations) that do not conform to the models used in full-data methods, such as evidence for past population size reductions due to *e.g.* an ice age (Hamilton *et al.*, 2005). Furthermore, indications about parameter ranges for the mutation or recombination rates can be set in the prior.

Another possibility is to numerically solve the expected JSFS with diffusion approximations (Chen *et al.*, 2007; Gutenkunst *et al.*, 2009). In Chen *et al.* (2007) a single sample from two populations is analyzed allowing for population growth. Besides a bigger sample size, Gutenkunst *et al.* (2009) includes gene flow and up to three populations. Gutenkunst's results are implemented in the software *∂a∂i* . The demographic model can be specified according to the needs but the diffusions approximations are based on the ISM.

MIMAR is a method developed by Becquet and Przeworski based on four summary statistics (SS) of the JSFS (Becquet and Przeworski, 2007). The MCMC method is especially intended for recently diverged species, includes variable or fixed recombination rates for each locus, and estimates gene flow. Thus, this method sounds ideal for the *S. chilense* and *S. peruvianum* data analysis. But as we noted in Tellier *et al.* (2011) the Markov chains seems to have difficulty to converge properly.

A similar and fast method which uses the JSFS and a composite likelihood with MCMC is implemented in Garrigan (2009). The *composite-likelihood* methods approximate the likelihood by assuming that the recombination rate between segregating sites is infinitely large, thus making the sites independent of each other (Kim and Stephan, 2000; Hudson, 2001; McVean *et al.*, 2002). This method might be especially useful if the recombination rates in the examined regions are high. Importantly, Wiuf (2006) showed that composite-likelihood estimators are consistent for many demographic models including the ones we are interested in. *Consistency* means the larger the number of examined regions becomes the closer the estimator is to the true value[16]. Garrigan computes the likelihood for the parameter under the assumptions of an ISM by simulating 10^5 data sets to calculate the expected JSFS and compares it to the observed JSFS. The demographic model includes population growth in one population but no gene flow between the diverging populations.

Scope of This Dissertation

In the previous sections I elaborated on the features of currently available methodology for demography estimation. However, to my knowledge there exists no suitable single method that can cope with all the challenges posed by the *S. chilense* and *S. peruvianum* data (Tab. 1): signs of population expansion, gene flow between species, and recent divergence, presence of high number of sites with multiple mutations and a high within-locus recombination rate.

To fill this gap, the aim of this dissertation was to develop a method that could handle all these challenges to be later applied to the *S. chilense* and *S. peruvianum* data, a premise for future research on selection. Further, many characteristics of the *Solanum* data set are likely to be important in other speciation processes, *e.g.* during range expansions into new habitats. Thus, a method that is also applicable to other species will help to determine factors contributing to speciation (*e.g.* gene flow).

In Chapter 1, we thus developed a composite-likelihood method: Jaatha (JSFS associated approximation of the ancestry, also Malayalam for "past"). The algorithm of Jaatha which can simultaneously estimate four demographic parameters is explained. Its performance is compared to the full-data likelihood method IM (Hey and Nielsen, 2004) and to another composite-likelihood method $\partial a \partial i$ (Gutenkunst *et al.*, 2009). The tomato data is analyzed

[16]It has to be noted that not all estimators are consistent: Tajima proved that the average number of pairwise differences π is inconsistent to estimate θ (Tajima, 1983).

using an ISM. The results are published in

Jaatha: a Fast Composite-likelihood Approach to Estimate Demographic Parameters
Lisha Naduvilezhath, Laura E. Rose, Dirk Metzler
Molecular Ecology (2011) 20, 2709-2723.

In Chapter 2, Jaatha is compared to PopABC (Lopes *et al.*, 2009), MIMAR (Becquet and Przeworski, 2007), and a likelihood method introduced in the article. Comparisons of different SS with differing numbers of loci are a also conducted. It has resulted in the publication

Estimating Parameters of Speciation Models Based on Refined Summaries of the Joint Site-Frequency Spectrum
Aurélien Tellier, Peter Pfaffelhuber, Bernhard Haubold, Lisha Naduvilezhath,
Laura E. Rose, Thomas Städler, Wolfgang Stephan, Dirk Metzler
PLoS ONE (2011) 6(5): e18155.

In Chapter 3, we extend Jaatha to analyze more than four demographic parameters. On simulated data sets with FSMs we attempt to estimate parameters under an ISM as well as an FSM. We also apply the new Jaatha version to the *Solanum* data using ISM and FSM. The results are in preparation to be submitted.

Chapter 1

Jaatha: a Fast Composite-likelihood Approach to Estimate Demographic Parameters

Lisha Naduvilezhath, Laura E. Rose, Dirk Metzler
Molecular Ecology (2011) 20, 2709-2723.

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

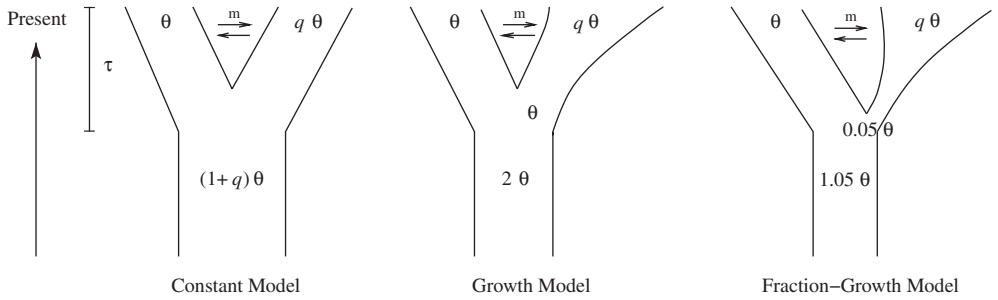


Fig. 1 The different demographic models for populations P_1 and P_2 used for the simulation study where θ = population mutation parameter for P_1, m = migration rate scaled by $4N_1$ generations, q = size ratio between P_2 and P_1 and τ = divergence time measured in $4N_1$ generations.

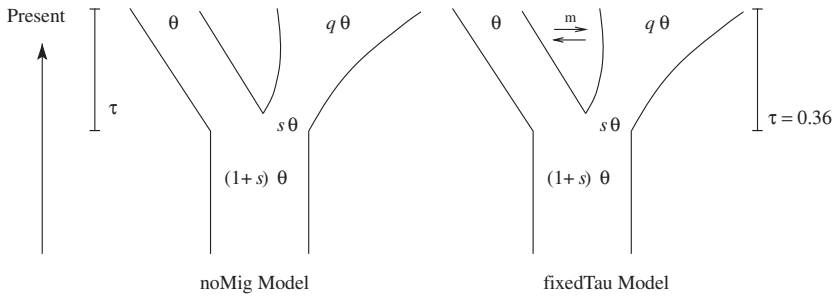


Fig. 2 Additional models applied to the tomato data, where s = the initial size ratio of P_2 and P_1 immediately after the split. The other parameters are defined as in Fig. 1.

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

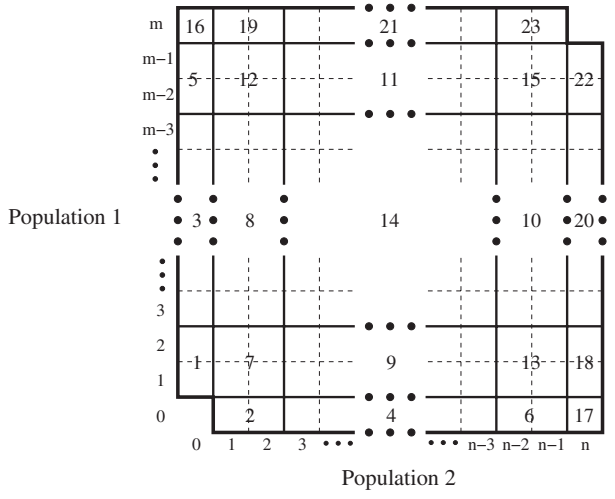


Fig. 3 Partition of domain of the joint site frequency spectrum (JSFS) for two populations where m and n denote the number of sampled alleles per locus of each population. Entries of the JSFS are summed up to result in 23 summary statistics.

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 $\{x: a < x \leq b\}$. 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 $\left[\exp\left(2\cdot\widehat{\phi}-\overline{\phi^*}-1.96\cdot\sigma(\phi^*)\right),\right.$ $\left.\exp\left(2\cdot\widehat{\phi}-\overline{\phi^*}+1.96\cdot\sigma(\phi^*\right)\right]$, 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

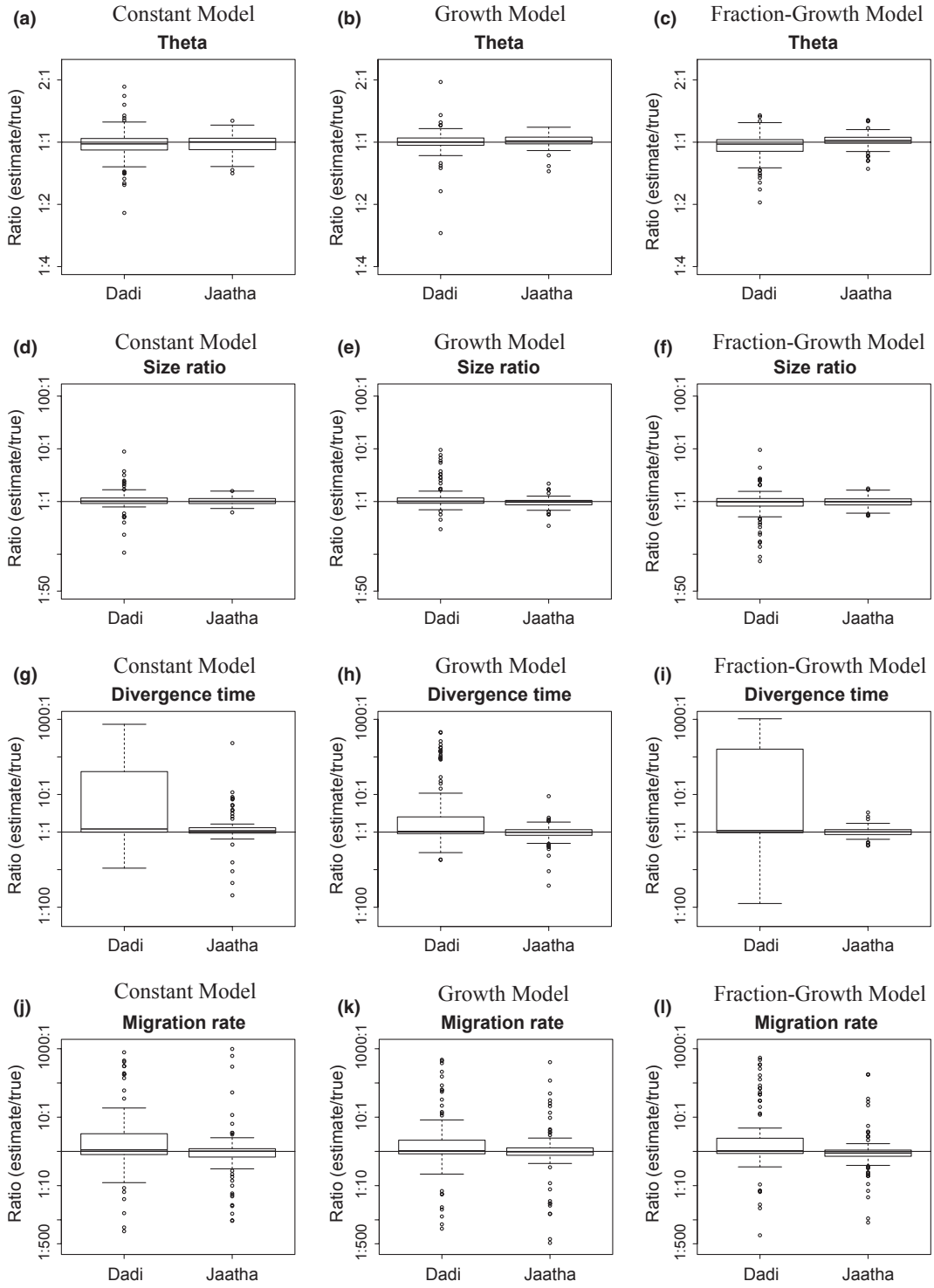


Fig. 4 Ratio of estimated to true values by *∂a∂i* and Jaatha of four parameters across models and methods for `100-loci scenario`. The 100 simulated data sets were generated without intralocus recombination.

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

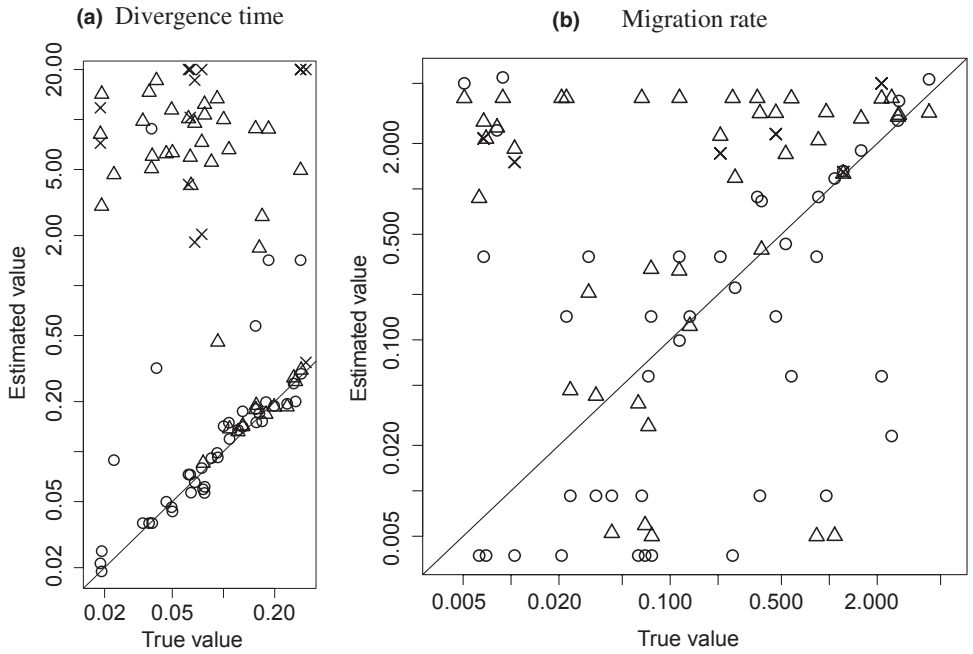


Fig. 5 The values estimated by Jaatha (○), IM (× for ESS > 100) and ∂a∂i (Δ) of (a) divergence time and (b) migration plotted against true values for the `100-loci scenario` of the *Constant Model* where true $\tau < 0.3$.

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

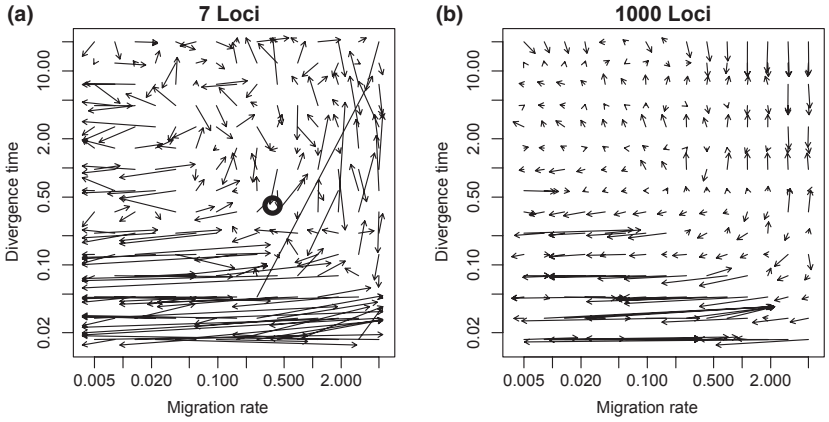


Fig. 6 Arrow plots of divergence time and migration for (a) 7 loci and (b) 1000 loci assuming the *Growth Model* with 45 samples per species and symmetric migration rates using Jaatha. The true values for the migration rate and divergence time are at the tails, and the estimated values are at the heads 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.

Chapter 2

Estimating Parameters of Speciation Models Based on Refined Summaries of the Joint Site-Frequency Spectrum

Aurelien Tellier, Peter Pfaffelhuber, Bernhard Haubold, **Lisha Naduvilezhath**, Laura E. Rose, Thomas Städler, Wolfgang Stephan, Dirk Metzler
PLoS ONE (2011) 6(5): e18155.

OPEN ACCESS Freely available online

PLoS one

Estimating Parameters of Speciation Models Based on Refined Summaries of the Joint Site-Frequency Spectrum

Aurélien Tellier[1]*, Peter Pfaffelhuber[2], Bernhard Haubold[3], Lisha Naduvilezhath[1], Laura E. Rose[1], Thomas Städler[4], Wolfgang Stephan[1], Dirk Metzler[1]

1 Department of Biology II, Section of Evolutionary Biology, LMU University of Munich, Planegg-Martinsried, Germany, **2** Faculty of Mathematics and Physics, University of Freiburg, Freiburg, Germany, **3** Department of Evolutionary Genetics, Max-Planck-Institute for Evolutionary Biology, Plön, Germany, **4** Institute of Integrative Biology, Plant Ecological Genetics, ETH Zurich, Zurich, Switzerland

Abstract

Understanding the processes and conditions under which populations diverge to give rise to distinct species is a central question in evolutionary biology. Since recently diverged populations have high levels of shared polymorphisms, it is challenging to distinguish between recent divergence with no (or very low) inter-population gene flow and older splitting events with subsequent gene flow. Recently published methods to infer speciation parameters under the isolation-migration framework are based on summarizing polymorphism data at multiple loci in two species using the joint site-frequency spectrum (JSFS). We have developed two improvements of these methods based on a more extensive use of the JSFS classes of polymorphisms for species with high intra-locus recombination rates. First, using a likelihood based method, we demonstrate that taking into account low-frequency polymorphisms shared between species significantly improves the joint estimation of the divergence time and gene flow between species. Second, we introduce a local linear regression algorithm that considerably reduces the computational time and allows for the estimation of unequal rates of gene flow between species. We also investigate which summary statistics from the JSFS allow the greatest estimation accuracy for divergence time and migration rates for low (around 10) and high (around 100) numbers of loci. Focusing on cases with low numbers of loci and high intra-locus recombination rates we show that our methods for the estimation of divergence time and migration rates are more precise than existing approaches.

Citation: Tellier A, Pfaffelhuber P, Haubold B, Naduvilezhath L, Rose LE, et al. (2011) Estimating Parameters of Speciation Models Based on Refined Summaries of the Joint Site-Frequency Spectrum. PLoS ONE 6(5): e18155. doi:10.1371/journal.pone.0018155

Editor: John J. Welch, University of Cambridge, United Kingdom

Received November 12, 2010; **Accepted** February 27, 2011; **Published** May 26, 2011

Funding: This work was supported by grants of the DFG Forschergruppe 1078, "Natural selection in structured populations", to D.M., P.P., and L.E.R.; DFG grants STE 325/9 and STE 325/13 to W.S.; Swiss National Science Foundation grant 31003A_130702 to T.S.; and Volkswagen Foundation grant I/82752 to A.T.. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: tellier@biologie.uni-muenchen.de

Introduction

Understanding speciation processes is crucial in numerous fields including conservation biology, ecology, host-parasite co-evolution and human evolution [1]. According to the "biological species concept", a species is defined as a group of interbreeding individuals that are reproductively isolated from other taxa [2]. Under this framework, the study of the speciation process focuses on the conditions leading to the emergence of reproductive isolation [3].

Allopatric population divergence is the classical scenario for isolation between populations [2]. In this model, two populations diverge in complete geographic isolation from one another. A second scenario considers divergence with continuing gene flow between populations, for example when species ranges abut (parapatry) or overlap following secondary contact, allowing for introgression. The latter model has been suggested to describe speciation events between human populations and ape species or sub-species [4], *Drosophila* species [5], and the wild tomato species *Solanum peruvianum* and *S. chilense* [6]. Key theoretical predictions have been generated to distinguish parapatric and allopatric population divergence based on genomic data [5,7]. These show that under the model of parapatric separation greater variation in divergence time is expected across the genome compared to an allopatric model [5]. In other words, the variance of shared polymorphisms between populations can be used to distinguish between recent divergence without gene flow and an older split characterized by high levels of subsequent gene flow between populations [7]. However, to reliably use these variances for parameter estimation, data sets with large numbers of sequences are needed, which is a practical constraint in studies of many non-model organisms [8].

The most widely used general model of population divergence is the "isolation-migration" model [5]. This model has six parameters, assuming two populations are used: the splitting time, the effective population size of each extant population and of the ancestral population, and the rates of gene flow. Bayesian Markov-Chain Monte-Carlo (MCMC) methods to sample from the posterior distribution of the parameters given the full sequence data are implemented in the program IM and its successors IMa and IMa2 [5,9,10,11]. Since the development and application of these methods to different species, a surprising number of cases indicate that speciation can occur in the presence of continual gene flow between incipient species [12]. However, existing implementations of these methods are limited to certain types of input data. For example, IM, IMa and IMa2 require that

haplotypes are known and that there is no intra-locus recombination. This second assumption is particularly problematic in species in which the ratio of recombination to mutation rates is high, including *Drosophila melanogaster* [13] and wild tomato species [14,15,16]. In these species, recombination cannot be ignored since sequenced genomic fragments have experienced one or more recombination events [17]. In practice, researchers have excluded segments or haplotypes with evidence of recombination for inference of parameters using this method. This ostensible "solution" has two disadvantages. First, it introduces bias into parameter estimation because genealogies of samples without recombination tend to be shorter [5,18]. Specifically, divergence time and current population sizes are shown to be overestimated, and ancestral population size is underestimated [18]. Second, for studies with few sequenced loci, the amount of data available for inference is significantly reduced, contributing to higher variances in parameter estimates.

Other methods rely on summary statistics such as the joint site-frequency spectrum (JSFS) [19], which is an array S of dimension $(n_1+1)\times(n_2+1)-2$ where entry $S_{i,j}$ is the number of polymorphic sites for which the derived state is found i times in the sample from population 1 and j times in the sample from population 2. For example, $S_{2,3}=10$ if 10 polymorphisms are found as doubletons in population 1 and as tripletons in population 2. For parameter estimation, Wakeley and Hey [19] summarized the JSFS by a vector $W=(W_1,W_2,W_3,W_4)$ containing the number of private polymorphisms in species 1 and 2, respectively (W_1, W_2), fixed differences between species (W_3), and shared ancestral polymorphisms (W_4). Examples of JSFS expectation values are shown in Fig. 1 for various combinations of parameter values. Methods using summaries are aimed to be computationally faster than maximum-likelihood and Bayesian full-data methods while being reasonably accurate, especially when many independent loci are used [20]. The method MIMAR (MCMC estimation of the isolation-migration model allowing for recombination [4]) uses a variant of the Wakeley-Hey summary statistics W. Approximate Bayesian Computation (ABC) methods were also developed to estimate parameters of the isolation-migration model from summary statistics such as the amount of private polymorphisms and diversity per population and in the pooled sample (popABC [21]). A great advantage of ABC methods is that they can be implemented in a few days or weeks whereas the implementation of full-likelihood methods or Bayesian full-data MCMC algorithms may take months or years, though to check the quality of the summary statistics in the ABC might require additional time consuming simulations. More recently, Gutenkunst *et al.* [22] developed the method ∂a∂i, which takes into account the entire JSFS. Note that in ∂a∂i, all sites are considered to be independent, and the JSFS is calculated for all sites and not per locus contrary to other methods [22]. In this composite likelihood approach, the expectation values of the full JSFS are numerically computed using diffusion approximations.

The present study was motivated by research on non-model organisms, including, for example, two recently diverged species of wild tomatoes (*S. peruvianum* and *S. chilense*). Not only do these species appear to have recently diverged but gene flow may be ongoing [6,23]. The programs IM, IMa and IMa2 cannot be used due to high levels of intra-locus recombination. Furthermore, given the low number of genes sampled (7 to 13 in this case) methods based on the data summary W have limited power to distinguish between divergence in isolation and divergence with continuing gene flow. Since we wished to determine whether these two species split recently with no or negligible levels of gene flow, or split less recently, but diverged in the presence of gene flow, we realized that previously described methods were not adequate.

Our first aim is to show as a proof of concept that refining the summary of the JSFS to more classes results in improved estimates of divergence time and gene flow. For this purpose we decompose the class W_4 (shared polymorphisms) of the JSFS into further classes for singletons and doubletons shared between species (see Fig. 1). The rationale behind this new decomposition is that if gene flow between species has been low, as expected if the two species are distinct, there should be (i) few incidences of shared polymorphisms compared to the number of private polymorphisms per species [19], and (ii) recent migrants lead to an excess of low-frequency shared polymorphisms (singletons and doubletons) whose frequency over time is affected by drift. We observe in Figure 1 that indeed private polymorphism is in large excess compared to shared polymorphisms. However, under the assumption of constant gene flow [5], small variations in the low

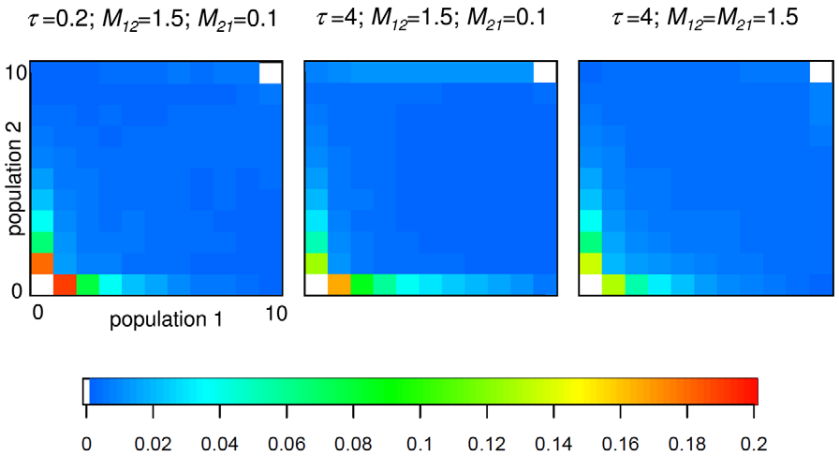


Figure 1. Three examples of joint site-frequency spectra for an Isolation-Migration model. An ancestral population of size $\theta_A=5$ splits into two incipient populations ($\theta_1=\theta_2=5$) at time $\tau=0.2$ or 4 in the past. 10 individuals are sampled from the two current populations and sequenced at 1,000 independent loci of 1,000 bp each. Intra-locus recombination occurs at a rate $\rho=0.02$. The color legend indicates the proportion of polymorphisms in a given JSFS class. Migration rate from population 1 to 2 (M_{12}), from population 2 to 1 (M_{21}) and split time (τ) are indicated for each panel.
doi:10.1371/journal.pone.0018155.g001

frequencies of shared polymorphism are indicative of the strength and symmetry of gene flow (Fig. 1). In the case of symmetric migration rates (gene flow from species 1 to species 2 equals that from species 2 to species 1) there is a symmetrical amount of shared low frequency polymorphism (singletons, doubletons) in both species (Fig. 1, third panel). On the other hand, if migration from species 1 to species 2 is high (and the opposite migration rate is low, Fig. 1 first and second panel) there is a higher proportion of shared polymorphism at low frequency in species 2, and a deficit of shared polymorphism at low frequency in species 1 (Fig. 1). We use the information from these differences in the amount of shared low frequency polymorphism in either species to estimate divergence time and gene flow using a simple likelihood ratio calculation method based on Hey and Nielsen [5]. We show that methods with more complex decompositions of W perform better than MIMAR.

The second aim is to develop a computationally efficient method designed for species with high levels of recombination (on the order of the mutation rate), which decreases the correlation across polymorphic sites. We neglect these dependencies and employ a composite likelihood approach based on a Poisson point process approximation of the JSFS, which significantly reduces the run time of the simulations. The parameter estimations are realized by local log-linear regression analysis. We demonstrate that this leads to a quantitative improvement of the use of the Wakeley-Hey summary statistics, because it allows the estimation of unequal directional gene flow between populations. Furthermore, computation time is much reduced compared to other methods. We show that our method is faster and gives more accurate estimates of divergence times and rates of gene flow than MIMAR, popABC, and ∂a∂i. However, for very recent divergence times ($<0.1\ N_e$ generations) all methods overestimate divergence time and gene flow, although our more complex summary of the JSFS seems to be more robust than other methods. Importantly, we show that our composite likelihood methods based on the assumption of genealogically independent SNPs are also more accurate than previous methods when estimating parameters at low recombination rates. As a practical conclusion for the use of JSFS statistics, we apply our composite likelihood method to determine which JSFS decompositions yield the highest accuracy for estimating divergence and gene flow parameters. We provide this comparison for the case where 7 loci (approximately 300 to 400 SNPs as found in studies in wild tomato species [14,23,24]) or 100 sequenced loci (as available for some model organisms such as Drosophilids or primates [8]) are available.

Methods

1. General model

We consider a neutral IM model in which an ancestral population splits into two populations that may exchange migrants. It is assumed that n_1 and n_2 alleles are sampled in the two populations and sequenced for a number of independently evolving loci (all loci have the same n_1 and n_2). Following Wakeley and Hey [19], μ is the average mutation rate across loci and can be used to estimate the effective population sizes of the three populations (N_A, N_1, N_2) if the scaled mutation rates $\theta_A = 4N_A\mu$, $\theta_1 = 4N_1\mu$ and $\theta_2 = 4N_2\mu$ can be estimated from the data. Note that as in Wakeley and Hey [19], τ is the estimated time of species divergence (in units of $2N_1$ generations). The two migration rates m_{12} and m_{21} are defined as follows: m_{12} is the fraction of population 2 that is replaced by migrants from population 1 each generation, and *vice versa* for m_{21}. The migration parameter is rescaled as twice the number of individuals in a population replaced by migrants (backward in time) with $M_{21} = 4N_1m_{21}$ and $M_{12} = 4N_2m_{12}$. In the current version, this model assumes that each locus is located on an autosome and follows the infinite-site mutation model with reciprocal recombination [25]. The coalescent simulations use Hudson's ms program [26]. Similar to Becquet and Przeworski [4], our model allows for intralocus recombination but not for gene conversion. The population recombination rate per base pair per generation is c. This value is assumed to be constant and known within a given locus and across all loci, *i.e.* we do not allow for variable recombination rates in the genome.

Following the description of the IM model by Hey and Nielsen [5], the posterior distribution of the parameters $\Theta = (\theta_A, \theta_1, \theta_2, \tau, M_{12}, M_{21}, c)$ is

$$\pi(\Theta|\Omega) \propto p(\Omega|\Theta)p(\Theta). \quad (1)$$

where Ω is the data, $p(\Omega \mid \Theta)$ is the likelihood of the vector of parameter values, Θ, and $p(\Theta)$ is its prior probability.

The full JSFS can be used to compare nucleotide sequence data of derived alleles from n_1 sequences from population 1 to n_2 sequences from population 2 [19]. It is assumed that an outgroup sequence is available and can be used to determine which allele is derived. Each derived allele is assigned to one cell of the JSFS depending on its frequency in the population. Note that i and j take integer values between 0 and n_1 and 0 and n_2, respectively. Wakeley and Hey [19] and Hey and Nielsen [5] used summary statistics for parameter inference in the isolation-migration model. Formally, they are written as

$$\begin{aligned} &W_1 = \sum_{1 \le i \le n_1 - 1} \left(S_{i,0} + S_{i,n_2}\right);\ W_2 = \sum_{1 \le j \le n_2 - 1} \left(S_{0,j} + S_{n_1,j}\right); \\ &W_3 = S_{0,n_2} + S_{n_1,0};\ W_4 = \sum_{1 \le i \le n_1 - 1} \sum_{1 \le j \le n_2 - 1} S_{i,j}. \end{aligned} \quad (2)$$

Note that in MIMAR, Becquet and Przeworski [4] make use of an outgroup sequence to derive a slightly different set of four summary statistics for the frequencies of a derived allele:

$$\begin{aligned} &W'_1 = \sum_{1 \le i \le n_1 - 1} S_{i,0};\ W_2{}' = \sum_{1 \le j \le n_2 - 1} S_{0,j}; \\ &W_3{}' = S_{0,n_2} + S_{n_1,0};\ W_4{}' = \sum_{1 \le i \le n_1} \sum_{1 \le j \le n_2} S_{i,j}. \end{aligned}$$

We demonstrate that using additional classes of the JSFS allows us to utilize more information than these original approaches, and improves the estimation of Θ. We present two methods that differ a) in the summary statistics used, *i.e.* different classes of the JSFS are used as summary statistics, and b) in the estimation procedure used to calculate the parameter values. To investigate the benefit of various sets of summary statistics for the joint estimation of divergence time and gene flow, we focus on estimating $\Theta = (\tau, M_{12}, M_{21})$ assuming that θ_A, θ_1, θ_2, and c are known.

2. Maximum likelihood method

Our first approach is based on the maximum likelihood inference of the set of parameters $\Theta = (\tau, M_{12}, M_{21})$ [4,7]. The data summaries are defined as a vector of four summary statistics extracted from the JSFS: D, D', D'', D^*. Our simplest summary of the JSFS, D, is a vector of 7 values (D_k, $k = 1,\ldots,7$) expanding the four classes W_k ($k = 1,..,4$) in Eq. 2. Additional classes relative to the

Wakeley-Hey set are created by splitting each class of private polymorphisms to each species (W_1 and W_2) and the fixed differences class (W_3), by distinguishing whether the derived allele is fixed or absent in the other species. This results in the following relation between Eq. 2 and elements of D: $W_1 = D_1+D_6$, $W_2 = D_2+D_7$, $W_3 = D_3+D_4$ and $W_4 = D_5$ (Appendix S1). The other vectors of summary statistics (D', D'', D^*) have more elements, 12 for D' and D'' and 23 for D^*, because singletons and doubletons in each population are included as new classes of shared polymorphism (see Appendix S1 for details). Compared to Nielsen and Wakeley [7] and Becquet and Przeworski [4], the class of shared polymorphisms between populations W_4 (Eq. 2) is further divided. The amount of information taken into account from the JSFS increases from D to D^*, as shared low frequency and private polymorphisms are counted as separate elements of the summary statistics vector.

Following Eq. 1, the likelihood $L_D(\Theta) = p(D \mid \Theta)$ of the parameter combination Θ, for the given data summaries D (or similarly for D', D'', D^*) is an integral over all genealogies G (or Ancestral Recombination Graphs, ARG) [9,27] as

$$L_D(\Theta|D) = p(D|\Theta) = \int_G p(D|\Theta,G)p(G|\Theta)dG. \quad (3)$$

The branch lengths of G are scaled in units of $2N_1$ generations. Since the probability of the sequence data depends only on G and the mutation rate, we get:

$$p(D|\Theta) = \int_G p(D|\theta_1,G)p(dG|\theta_2/\theta_1,\theta_A/\theta_1,\tau,M_{12},M_{21},c).$$

Thus, the likelihood $p(D \mid \Theta)$ can be approximated for each locus by generating a set of I genealogies G_m, $m \in \{1,...,I\}$, using Hudson's ms [26] as

$$p(D|\Theta) \approx \frac{1}{I}\sum_{m=1}^{I} p(D|\theta_1,G_m). \quad (4)$$

In Eq. 4, $p(D \mid \theta_1,G_m)$ can be computed explicitly. The number $S_{i,j}$ of polymorphic sites of frequency i in population 1 and j in population 2 is Poisson distributed with mean $L_{i,j}\theta_1/2$, where $L_{i,j}$ is the total length of ARG branches leading to i sequences in the first and j sequences in the second sample. Conditional on the genealogies, the probabilities of observing each element D_k of the vector D are independent. The likelihood of the data for a given locus is approximated by

$$p(D|\Theta) \approx \sum_{m=1}^{I} \frac{1}{I} \prod_{k=1}^{K} p(D_k|\theta_1,G_m). \quad (5)$$

Note that for the vector D, $K = 7$, but for D' and D'', $K = 12$, and for D^*, $K = 23$.

A modified version of Hudson's ms is used to calculate the likelihood values for each simulated genealogy, and 10,000 genealogies were randomly drawn for each parameter combination. In the following, the maximum-likelihood methods based on these summaries are called D_1 (using vector D), D_2 (using vector D'), D_3 (using vector D'') and D_4 (using vector D^*).

Since this method is not yet optimized for speed, the distribution of likelihood values is simply computed for values of Θ, *i.e.* τ, M_{12} and M_{21}, within a defined range. The maximum likelihood parameter values are obtained by local regression analysis using the locfit function available in the statistical software R (locfit package; [28]).

3. Composite likelihood method

Our second method is a variant of the method Jaatha, which is implemented as R code available from http://evol.bio.lmu.de/_statgen/software/jaatha. This method is computationally efficient because it takes advantage of the high recombination rate observed in *Drosophila* [13] and in some outcrossing plant species, including wild tomatoes [16]. This allows us to simplify the computation by treating the sites within and between loci as if they were independent. A further advance of this method is the improvement in estimation of rates of gene flow between populations, for example when migration rates are unequal.

Briefly, the method comprises three steps. First, summary statistics, *i.e.* classes of the JSFS, are calculated by coalescent simulations over the range of the three parameters to be estimated. Second, the three-dimensional parameter space is subdivided into $8\times8\times8$ blocks. In each block, a log-linear regression (generalized linear model of Poisson type [29]) is fitted to the simulated data to describe for each of the JSFS classes how the expected number of mutations in this class depends on the N_p parameters. Third, the composite likelihood of each block, given the observed values of JSFS summaries, is approximated using the fitted local log-linear regressions, and parameter estimates are obtained within the region with the highest likelihood. Note that the composite likelihood method is equivalent to the fitting of a multivariate Poisson distribution [30] to the summary statistics as a function of the genetic model parameters.

The parameters, τ, M_{12}, and M_{21}, of the isolation-migration model are estimated. Using Hudson's ms as coalescent simulator, we calculate summary statistics from the JSFS for numerous points on a grid in the parameter space (in this case a three-dimensional space). In the initial version of Jaatha, the JSFS is split into 23 elements constituting the vector $\check{D}_k$, $k \in \{1,...,23\}$. The vector $\check{D}$ is similar to D^* mentioned above as it considers classes of shared polymorphisms that are singletons or doubletons in both populations ($\check{D}_6$ in Appendix S1). However, $\check{D}$ differs from D^* through the addition of classes of shared polymorphism with nearly fixed frequencies (such as $n_1 - 1$, $n_1 - 2$, $n_2 - 1$, $n_2 - 2$). We give a detailed description of $\check{D}$ in Appendix S1. In practice, simulations considered 40 different values for each parameter, and for each of the $40\times40\times40 = 64{,}000$ parameter combinations, 10 coalescent simulations were performed and the vector $\check{D}$ of summary statistics was stored.

Next, the three-dimensional space of parameters was divided into sub-regions of size N_R for all three parameters. Each region contained $N_R{}^3$ points characterized by the set of summary statistics $\mathcal{J}$. In practice, we chose $N_R = 5$, *i.e.* we subdivided the parameter space into $8\times8\times8$ blocks each of which contained $5\times5\times5$ different parameter combinations used in the simulation step. For each block and for each of the 23 summary statistics a log-linear Poisson regression model with the three parameters (τ, M_{12}, and M_{21}) as explanatory variables was fitted to the simulated data from $5\times5\times5\times10 = 1{,}250$ simulations (generalized linear model of Poisson type; [29]. For $x = 1,\ldots,5$; $y = 1,..,5$ and $z = 1,..,5$ let τ_x, $M_{12,y}$ and $M_{21,z}$ be the parameter values in a certain block. Then, x is an affine transformation of $\log(\tau_x)$ and the same holds for y with $\log(M_{12,y})$ and z with $\log(M_{21,z})$. Fitting the log-linear Poisson model for a certain block b and a certain summary $\mathcal{J}_k$ requires the estimation of coefficients ($\alpha_{1,k}$, $\alpha_{2,k}$, $\alpha_{3,k}$, $\alpha_{4,k}$) such that the following equation holds for the expected value $d_{k,x,y,z}$ of $\check{D}_k$

$$\log(d_{k,x,y,z}) = \alpha_{1,k}x + \alpha_{2,k}y + \alpha_{3,k}z + \alpha_{4,k}. \quad (6)$$

or, equivalently,

$$d_{k,x,y,z} = \tau_x^{\beta_{1,k}} \cdot M_{12,y}^{\beta_{2,k}} \cdot M_{21,z}^{\beta_{3,k}} \cdot \beta_{4,k},$$

where parameter values of 0 are replaced by small positive values and $\beta_{i,k}$ is a transformation of $\alpha_{i,k}$. Given any parameter values τ, M_{12}, and M_{21} in the range of block k, the observed values of the summary statistic $\check{D}_k$ are assumed to be Poisson distributed with expected value $\tau_x^{\beta_{1,k}} \cdot M_{12,y}^{\beta_{2,k}} \cdot M_{21,z}^{\beta_{3,k}} \cdot \beta_{4,k}$. If $d_{1,\varphi}$, $d_{2,\varphi}$, …, $d_{23,\varphi}$ are the expected values of the 23 summary statistics for a certain combination $\varphi = (\tau, M_{12}, M_{21})$ of parameter values and $F = (F_1, \ldots, F_{23})$ are the observed values, then the Poisson model likelihood of φ is

$$L_F(\varphi) = \Pi \frac{d_{i,\varphi}^{F_i}}{F_i!} \cdot e^{-d_{i,\varphi}}.$$

Note that Eq. 6 uses the logarithm of the parameter values to increase the resolution at low values, *i.e.* recent divergence time and low gene flow.

The first two steps are carried out independently of the observed data, and the most time-consuming part of the method is to fit regression models that describe how the expectation values of the summary statistics depend on the model parameters in the simulated data. The results of these steps can be reused to analyze data with similar sample sizes and parameter ranges. We have tried four different strategies for parameter estimation (called J_1, J_2, J_3 and J_4):

- J_1. Only the $8\times8\times8 = 512$ parameter combinations in the centers of the blocks are considered. Compute the Poisson model likelihood of each block center using the log-linear regression model. Output the block center with the highest value.
- J_2. Output a weighted mean of the block centers. The weights are the Poisson model likelihoods as computed in J_1.
- J_3. For each block, start in the block center and numerically optimize the Poisson model likelihood within the block. Output the highest value that is found in any of the blocks.
- J_4. Start an optimization in each block center. Allow the optimization search paths to change between the blocks. Near the block boundaries mixtures of the log-linear regression models fitted to the neighboring blocks are used to estimate the expected values of the summary statistics.

On a standard desktop computer, strategies J_1 and J_2 only take a few seconds, strategy J_3 takes less than five minutes and strategy J_4 takes 10 to 15 minutes for one data set. This requires that the log-linear model fitting has been performed in advance. Note that this step does not depend on the data. The fitting procedure takes about three to four days and the stored results can be re-used for data sets with the same sample sizes n_1 and n_2.

4. Power analysis

i) Analysis for various JSFS coarsenings. We conducted a power analysis to compare the different coarsenings of the JSFS for estimating divergence time and detecting post-divergence gene flow. Sets of sampled loci were simulated under the IM model using Hudson's ms. We defined the simulated values of the model parameter as τ_{sim}, $M_{12\text{-}sim}$, and $M_{21\text{-}sim}$. Then using the JSFS obtained for each set of simulation, we estimated the three parameters of the model (τ_{est}, $M_{12\text{-}est}$, and $M_{21\text{-}est}$) using our maximum likelihood methods (D_1–D_4) and the composite method (J_1–J_4). For comparison, estimations were also computed using the MCMC-likelihood program MIMAR [4]. To make the methods comparable, MIMAR, D_{1-4} and J_{1-4} have identical fixed values for population sizes and recombination rate (θ_A, θ_1, θ_2, and c) when estimating divergence and migration. The model underlying our simulation study is motivated by research on sequence variation in genes from non-model organisms for which few loci (here 7, each of length 1,000 bp) are available in two closely related species. However, our methods can also be applied to species for which numerous sequenced loci are available. In this case, the accuracy of the parameter estimates increases (see Fig. S11).

We evaluated how the different coarsenings of the JSFS affect the accuracy of parameter estimates compared to MIMAR. For these analyses we fixed a recent divergence time to $\tau = 0.1$ but varied the migration rates (M_{12}, M_{21}) from very low ($M_{12} = M_{21} = 0.5$) to intermediate ($M_{12} = M_{21} = 2$). Moreover, we investigated how other parameters of the model influence the accuracy of each method. Based on population sizes observed in wild tomatoes [14,23], the mutation parameters θ_A, θ_1, θ_2 are assumed to be equal ($\theta_A = \theta_1 = \theta_2$), taking a value of 5 or 10. Similarly, the recombination rate $\rho = 4N_1c$ takes values of 5 (low c), 10 (intermediate c) or 20 (approximating high recombination). For each set of parameter values, 20 datasets of 7 loci were generated and analyzed using our maximum likelihood methods (D_{1-4}), the composite method (J_{1-4}) and MIMAR. MIMAR was run twice with two and 10 million steps of burn-in, the outputs being calculated based on 100,000 or 500,000 steps, respectively. Convergence to maximum likelihood values was assessed by a high rate of accepted steps, as recommended (over 10%; [4,31]). The results of this analysis are shown in Figures 2 and 3 (and Figs. S1 and S2, Tables S1 and S2, Appendix S1).

ii) Analysis of robustness and speed. The second accuracy analysis deals with testing the robustness and speed of the composite method (J_1–J_4) by comparing performance with that obtained with MIMAR [4], the ABC implementation popABC [21], and the program ∂a∂i [22]. We generated 100 simulated data sets for a wide range of parameter values chosen at random. The divergence time was set from very recent ($\tau = 0.01$) to ancient ($\tau = 9$), migration rates were unequal ($M_{12} \neq M_{21}$) each ranging from very low ($M = 0.01$) to high ($M = 9$). The mutation parameters $\theta_A = \theta_1 = \theta_2$ and the scaled recombination rate $\rho = 4N_1c$ were chosen at random between 5 and 20 per locus. The uniform priors for divergence time and migration rates are identical for our composite method (J_1–J_4), MIMAR, and popABC, and are defined as $0.01 < \tau < 10$ and $0.01 < M_{12}, M_{21} < 10$. Note that all methods have identical fixed values for the population sizes and recombination rate (θ_A, θ_1, θ_2, and c).

We used popABC to generate 300,000 simulations for each of the 100 data sets assuming fixed values of ρ and $\theta_A = \theta_1 = \theta_2$ for seven independent loci. The rejection and regression steps of the ABC were performed using the ABCreg code [32], with estimates of τ, M_{12} and M_{21} calculated as the mode of the best 3,000 (1%) simulations. Tests with popABC using all 22 possible summary statistics did not lead to reliable estimates. ABC methods can lack statistical power to estimate parameters when the number of summary statistics is too large [33,34], because too few simulated datasets are close enough to the observed data, and the regression part of the ABC procedure cannot be realized. Therefore, we used fewer summary statistics. A first set of estimations are conducted

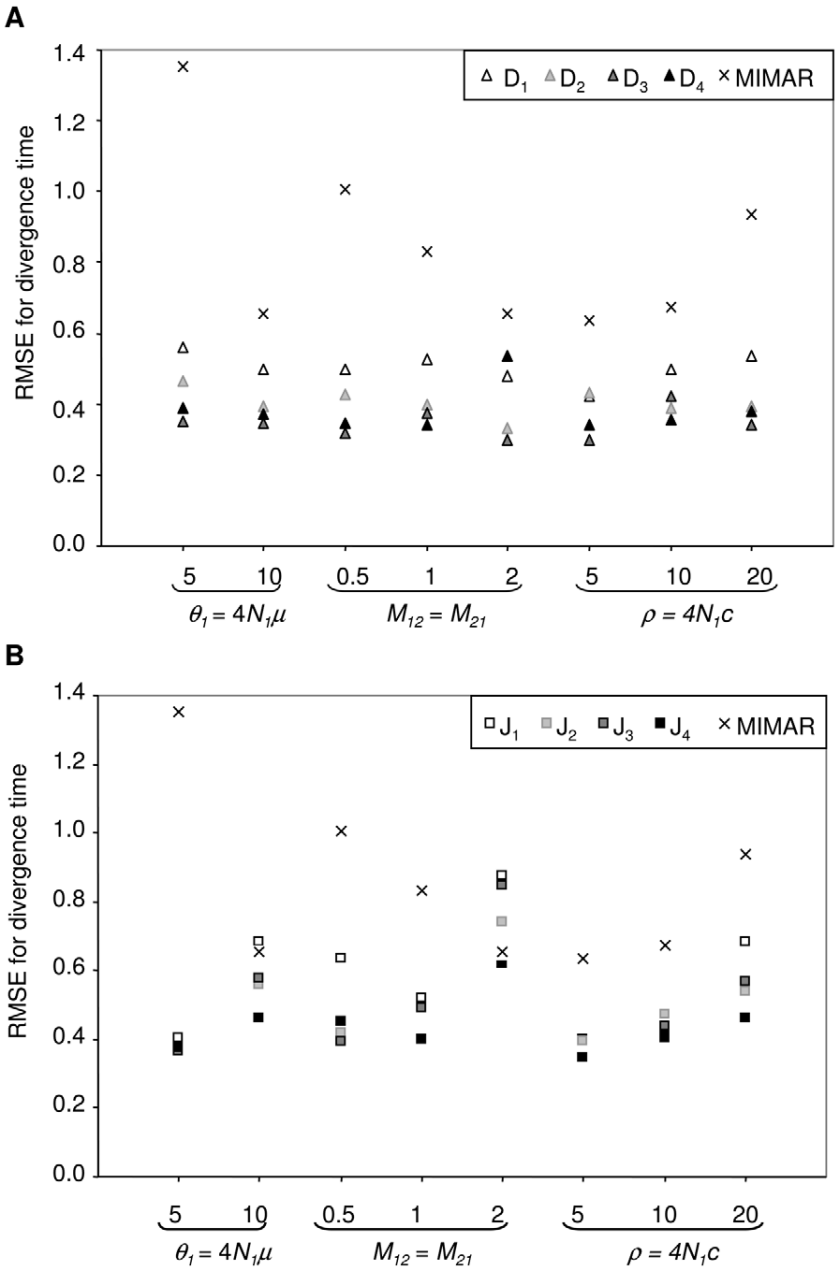


Figure 2. RMSE for the estimate of divergence time (τ) as a function of the population mutation rate (θ), values of simulated migration rate (M_{12}= M_{21}) and population recombination rate (ρ). The RMSE is computed across 140 datasets with divergence time fixed at τ = 0.1. (a) For the four maximum likelihood methods (D_1–D_4) and MIMAR, (b) for the four composite-likelihood methods (J_1–J_4) and MIMAR.
doi:10.1371/journal.pone.0018155.g002

based on six statistics from popABC closely related to the JSFS, *i.e.* for each species: the mean mutation frequency spectrum, an estimate of F_{ST} based on segregating sites, and the number of private segregating sites [21]. A second set of estimations with 11 summary statistics was constructed by adding the number of segregating sites per species and for both species pooled, and the frequency of private polymorphisms. Finally, a third set of estimations with 14 statistics additionally comprised the number of different haplotypes in each species and for the pooled samples [21]. These 100 identical data sets were also analyzed using the ∂a∂i program [22]. However, we were unable to obtain reasonable parameter estimates from MIMAR. In fact, despite using 10 to 20 million burn-in steps, convergence to a maximum likelihood value for τ, M_{12} and M_{21} (fixing ρ and $\theta_A = \theta_1 = \theta_2$) could not be obtained after more than 4 weeks of running. This is probably due to the wide range of priors for τ, M_{12} and M_{21} extending over several orders of magnitude (C. Becquet pers. comm.).

iii) Finding the best summary statistics. We looked for the best set of summary statistics, *i.e.* coarsenings D, D', D'', D^* or $\check{D}$ of the JSFS, to be used for parameter estimation with our fast composite likelihood method. We ran methods J_{1-4} with these 5 different vectors of summary statistics and compared estimates

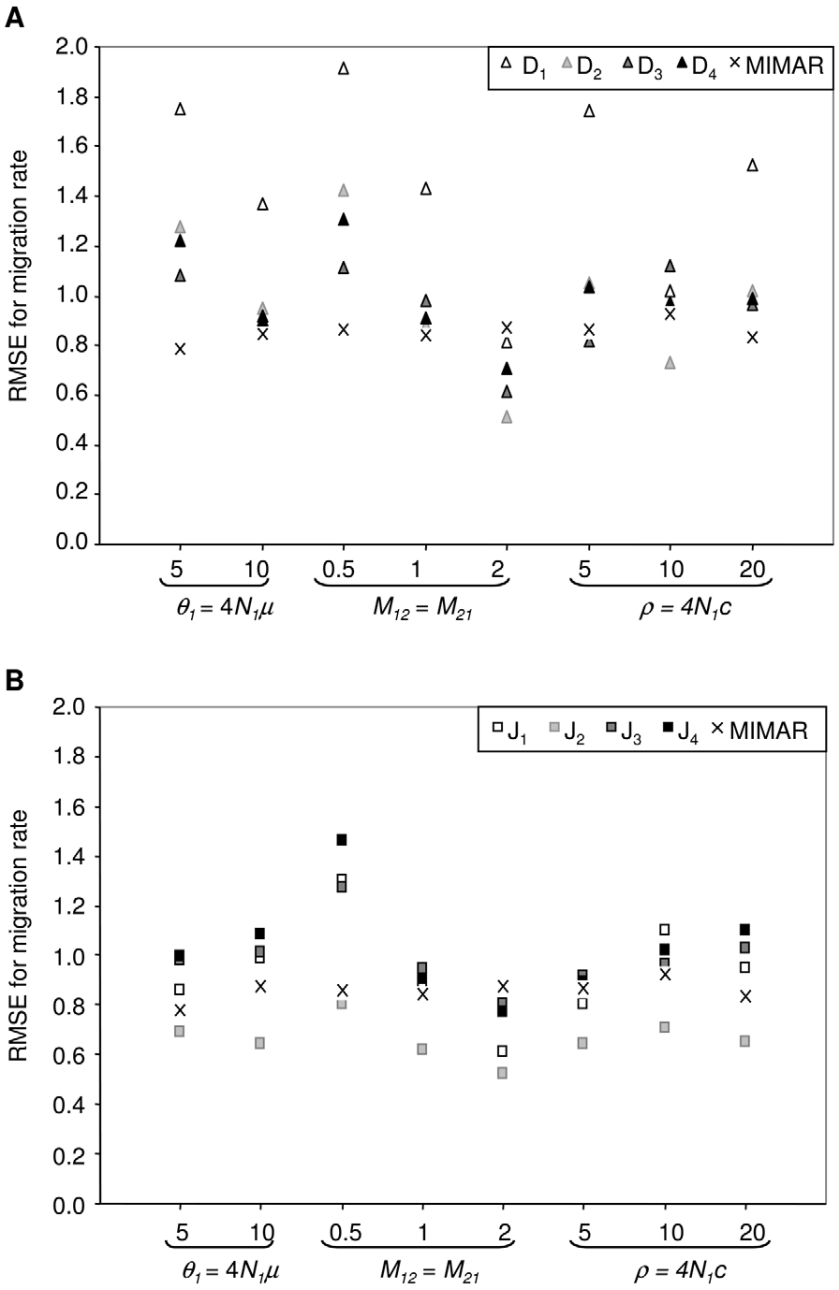


Figure 3. RMSE for the estimate of migration rate ($M_{12}=M_{21}$) as a function of the population mutation rate (θ), values of simulated migration rate ($M_{12}=M_{21}$) and population recombination rate (ρ). The RMSE is computed across 140 datasets with fixed divergence time at $\tau = 0.1$. (a) for the four maximum likelihood methods (D_1–D_4) and MIMAR, (b) for the four composite-likelihood methods (J_1–J_4) and MIMAR.
doi:10.1371/journal.pone.0018155.g003

with those obtained running methods J_{1-4} with the Wakeley-Hey vector of statistics (W). We analyzed the 100 simulated data sets of 7 loci (each of length 1,000 bp) with randomly chosen parameter values as described above. In addition, we performed a second analysis with simulated data sets of 100 independent loci of 1,000 bp each with parameters values in the same range as above ($0.01<\tau<9$, $M_{12}\neq M_{21}$ and $0.01<M<9$, $\theta_A=\theta_1=\theta_2$ and $\rho=4N_1c$ chosen at random between 5 and 20 per locus). The results of this analysis are shown in Figure 5 (and Figs. S6, S7, S8, S9, S10, Appendix S1).

iv) Statistical treatment. The results are presented in the format commonly used for power analyses. We report the mean of the estimate for each parameter value and three other statistics (see for example [35,36]). The relative error (RE) is the relative difference between the estimated parameter value and the true parameter value that was used to simulate the data. For example, for the divergence time (τ), the relative error is RE_τ:

$$RE_\tau = \frac{\tau_{est} - \tau_{sim}}{\tau_{sim}}.$$

The root mean square error (RMSE) is the square root of the average squared difference (over n_{sim} data sets) between the estimated value and the simulated value divided by the simulated value, and similarly, for τ:

$$RMSE_{\tau} = \sqrt{\frac{1}{n_{sim}} \sum \left(\frac{\tau_{est} - \tau_{sim}}{\tau_{sim}} \right)^2}.$$

The Factor 2 (F_2) is the proportion of data sets for which the estimated value (of τ or M) is at least half and at most twice the simulated value. Analyses of variance statistics were computed using the *glm* function, and multiple mean comparisons are based on Tukey's HSD test (confirmed by a Bonferroni test), as implemented in the R software ([28]; see Appendix S1, Tables S1 and S2 for details). We also analyzed the coverage of the methods, which is defined as the probability that the true parameter values are within the estimated 95% confidence range for τ and M. A possible approach to construct confidence ranges is based on the χ^2-approximation for the distribution of log-likelihood ratios. In the case of two parameters, the confidence range consists of all parameter combinations for which the natural logarithm of the ratio of the maximum likelihood and the likelihood of the candidate values is smaller than 2.99 [37]. Coverage analyses were performed for this type of confidence range for the composite likelihood and the maximum likelihood methods, and for the credibility ranges reported by MIMAR based on 140 datasets of 7 loci (each 1,000 bp).

Results

1. General results

All methods (maximum likelihood, composite likelihood, MIMAR, popABC, and ∂a∂i) showed variation in estimates of divergence time and, in particular, migration rates (Figs. 2, 3, 4 and Tables 1, 2). However, our methods showed the smallest relative error and RMSE for divergence time, resulting in good power to detect recent divergence ($\tau = 0.1$; Figs. 2 and 3, Fig. S1). MIMAR significantly underestimated migration rates and overestimated divergence time compared to other methods (Figs. 2 and 3; Figs. S1 and S2).

Over a large range of divergence times, from very recent ($\tau = 0.01$) to very old ($\tau = 9$), large overestimations were not common (relative error >10; Tables 1 and 2). However, migration rates were consistently overestimated by the composite likelihood methods, ∂a∂i, and popABC (*i.e.* relative error of 10 to 950; Table 1). Our methods J_{1-4} perform better than popABC and ∂a∂i in estimating both the divergence time and migration rates (Tables 1 and 2), and estimates of migration are always more accurate for high divergence times ($\tau > 0.5$) than for recent population splits ($\tau < 0.5$; Figs. S8 and S9).

An interesting, though expected, pattern is found when divergence time is fixed to a recent split, e.g. $\tau = 0.1$. For our eight methods and MIMAR, a positive correlation is found between the relative error in estimates of divergence time and migration rates (Fig. S2). This means that when a given method over- or underestimates the divergence time, it also over- or underestimates the migration rate.

The estimates of divergence time and migration rates are only slightly affected by other population parameters, such as the mutation rate (θ) and the recombination rate (ρ). In fact, the relative error of the divergence time depends only on the method chosen and the population mutation rate. A significant interaction between method and θ is analyzed further by calculating the RMSE, in order to find which method performs better for a given value of θ (Fig. 2 and 3, Table S1). For all methods, the relative error of migration rates decreases when gene flow between populations increases (Fig. 3, Table S2).

2. Estimating divergence time

Our maximum likelihood methods D_3 and D_4 and composite-likelihood methods J_2 and J_3 perform better in estimating

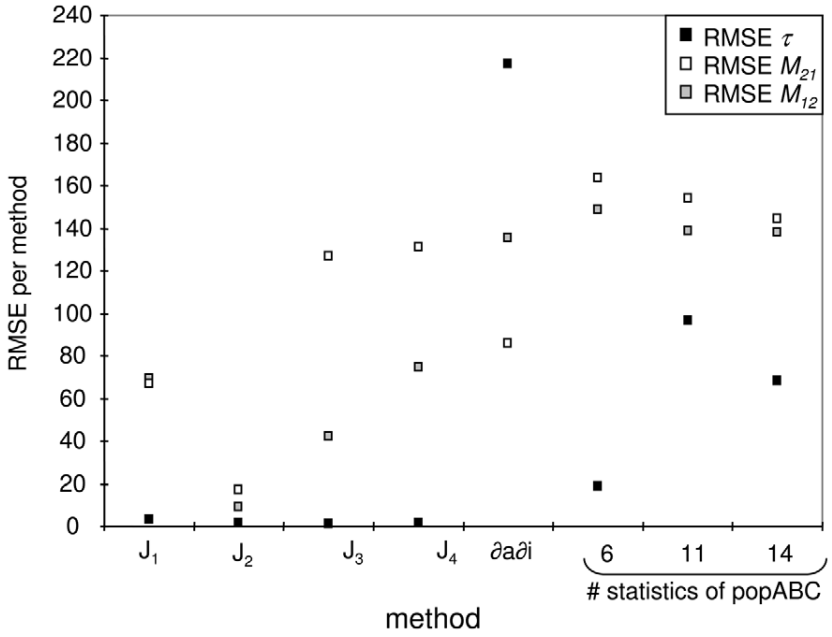


Figure 4. Comparison of RMSE for estimates of the divergence time and migration rates ($M_{12} \neq M_{21}$) between methods. Results are shown for the four composite-likelihood methods (J_1–J_4), ∂a∂i, and for popABC with 6, 11 and 14 summary statistics (computed across 100 datasets). doi:10.1371/journal.pone.0018155.g004

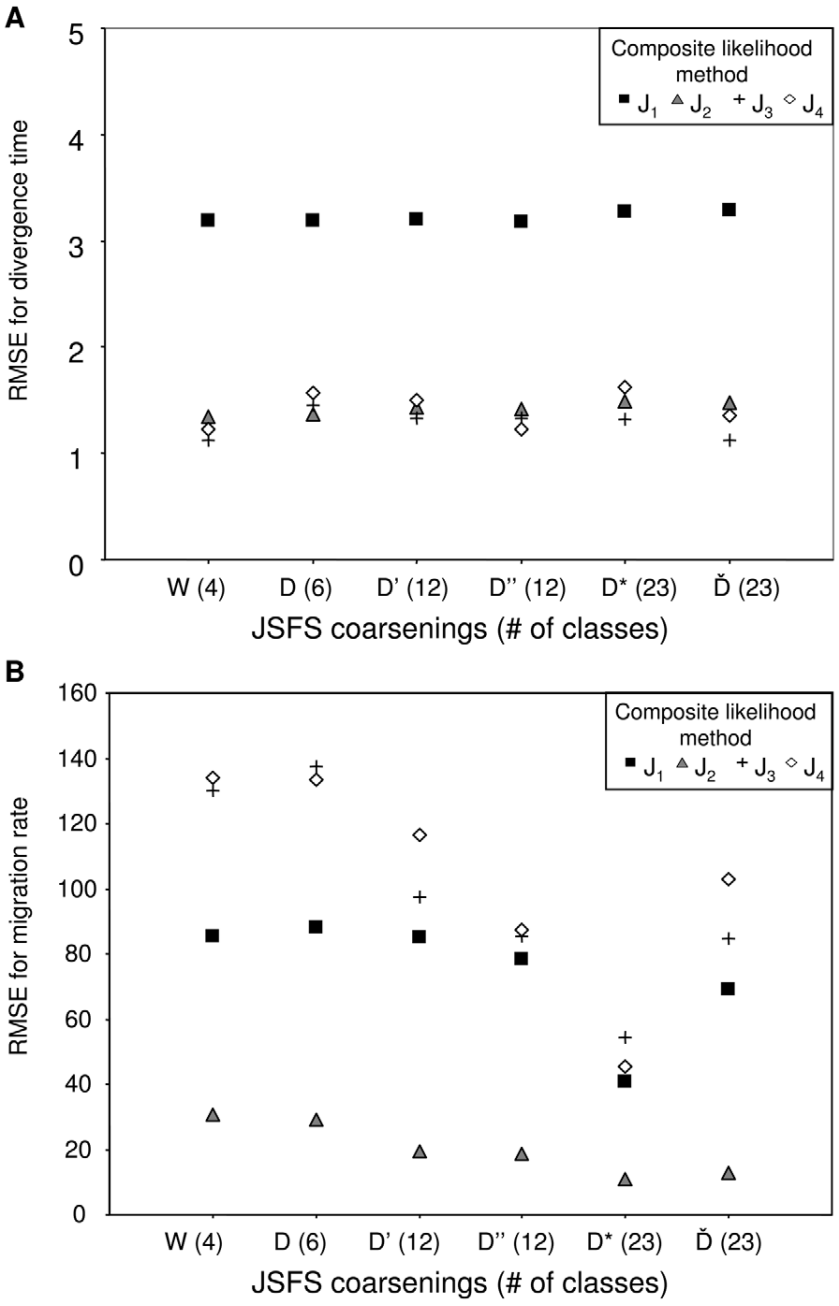


Figure 5. Power analysis of the various JSFS coarsenings to estimate divergence time and migration rates for 100 datasets of 7 loci. RMSE are computed for estimates of (a) the divergence time, and (b) migration rates ($M_{12} \neq M_{21}$) for the four composite-likelihood methods (J_1–J_4) based on six vectors of summary statistics. The vector *W* is defined by the four Wakeley-Hey classes from Eq. 2, and other vectors *D*, *D′*, *D″*, *D** and *Ď* are refined decompositions of the JSFS with higher numbers of classes.
doi:10.1371/journal.pone.0018155.g005

divergence time than other methods (MIMAR, D_1, D_2, J_1, J_4; see the lower RMSE in Fig. 2; Fig. S1). MIMAR shows increased accuracy in estimating τ as the migration rate (M) increases, reflecting the dependence between these parameter estimates (Fig. 3). This means that estimates of divergence time are improved by increasing the number of segregating sites, *i.e.* increasing θ (Fig. 2, Figs. S3 and S4). On the other hand, our methods do not show this trend (Figs. S6 and S7). On the contrary, the RMSE for divergence time increases as a function of θ for methods $D_{1–4}$ (ANOVA in Table S1). According to the RMSE and Factor 2 values, our methods D_2, D_3, D_4 and J_2, J_4 are the most accurate for estimating recent divergence time (Fig. 2, Figs. S3 and S4).

3. Estimating gene flow

Estimates of migration rates are generally less accurate than those of divergence time. The maximum likelihood methods $D_{1–4}$ show greater variance in estimates than the composite methods $J_{1–4}$ and MIMAR. However, MIMAR always underestimates the migration rate (Fig. 3). This consistent underestimation of

Table 1. Relative error for estimates of divergence time with our composite likelihood methods, ∂a∂i, and popABC for 100 randomized datasets of 7 loci.

	Composite methods				∂a∂i	popABC		
	J_1	J_2	J_3	J_4		6 summary statistics	11 summary statistics	14 summary statistics
Minimum	−0.959	−0.953	−0.958	−0.959	−0.693	−0.875	−0.998	−0.998
Quartile 25%	−0.074	−0.157	−0.083	−0.094	0.107	0.569	−0.040	−0.770
Median	0.217	0.121	0.166	0.172	2.685	2.562	2.825	1.105
Quartile 75%	0.653	0.523	0.564	0.439	99.504	8.646	11.045	6.764
Maximum	30.404	11.894	7.001	8.59	957.562	139.88	775.128	578.51
Mean	0.747	0.434	0.454	0.498	96.953	8.146	23.170	15.635

doi:10.1371/journal.pone.0018155.t001

migration rates by MIMAR results in small RMSE values because as the estimated migration rate goes to zero, the relative error, by definition, goes to −1 (Fig. S1b). Underestimation of migration rates by MIMAR is also revealed by the small Factor 2 values (Fig. S4). However, the lowest RMSE values are obtained for method J_2 (Fig. 3b). All four composite methods show consistently low RMSE values at the three recombination rates tested (ρ in Fig. 3b). Maximum likelihood methods are more accurate for estimating migration rates when the true migration rate is large (lower RMSE and higher Factor 2, Fig. 3a). Overall, our eight methods estimate gene flow better when the rates are high.

4. Robustness and comparisons of methods

Our maximum likelihood methods are not sensitive to recombination, while MIMAR shows higher RMSE values in estimates of divergence time as recombination increases (Figs. 3a and 4a). Likewise, the RMSE increases for estimates of divergence time using our composite likelihood methods as ρ increases (although not significantly based on the ANOVA analysis; Table S1). Sensitivity to recombination is not found for estimates of migration rate (Fig. S7).

The ∂a∂i method tends to overestimate divergence time compared to other methods (Table 1). Relative error for estimates of very recent divergence times ($\tau<0.1$) is high, although the median of the relative error rates is similar to results of popABC (Table 1). Compared to popABC, ∂a∂i is more accurate in estimating migration rates, demonstrating the statistical power gained by considering the maximum amount of information from the JSFS (Table 2, Fig. S5). However, the overall performance of ∂a∂i in estimating divergence time and migration rates is worse than that of our composite-likelihood methods (higher RMSE in Fig. 4, Fig. S5).

5. Advantage of using more than four JSFS based summary statistics and more loci

We demonstrate the benefit of using more than four statistics of the JSFS for estimating divergence time and migration rates. Methods relying on relatively few classes within the JSFS such as MIMAR and our maximum likelihood method D_1 (with only 7 classes of the JSFS) tend to over- or underestimate divergence time and migration rates more often than the other maximum likelihood methods (D_{2-4}; Figs. 1, 2 and 3). In fact, RMSE values for divergence time are higher for D_1 and MIMAR compared to D_{2-4} (Fig. 2a), and higher for migration rate under D_1 compared to D_{2-4} (Fig. 3a). Second, estimates from composite-likelihood methods show RMSE values that are several orders of magnitude lower for divergence time than those obtained with popABC, which relies on very limited information from the JSFS (Fig. 4). Running popABC with six statistics was the most accurate method to estimate divergence time, compared to using more statistics (11 and 14; Fig. 4). Third, JSFS-based summary statistics provide more accurate estimates (*i.e.* lower RMSE and higher Factor 2) of unequal migration rates between populations ($M_{12}\neq M_{21}$) than do popABC statistics (Fig. 4, Tables 1 and 2).

Finally, our comparison of the different JSFS coarsenings using the composite likelihood method shows that estimates of migration rates are more accurate when considering the vectors D'', D^* or $\check{D}$

Table 2. Relative error for estimates of the migration rate from population 1 to 2 (M_{12}) with composite likelihood methods, ∂a∂i, and popABC for 100 randomized datasets of 7 loci.

	Composite methods				∂a∂i	popABC		
	J_1	J_2	J_3	J_4		6 summary statistics	11 summary statistics	14 summary statistics
Minimum	−0.996	−0.983	−0.998	−0.996	−0.968	−0.910	−0.989	−0.990
Quartile 25%	−0.509	−0.504	−0.56	−0.565	−0.072	−0.163	−0.797	−0.855
Median	−0.084	−0.07	−0.031	−0.101	0.371	9.175	−0.016	−0.201
Quartile 75%	0.801	0.69	0.464	0.499	3.883	57.874	20.738	17.902
Maximum	660.63	61.39	418.4	510.6	951.11	729.420	959.534	959.534
Mean	11.633	2.07	5.41	14.44	37.406	67.407	40.28	39.919

doi:10.1371/journal.pone.0018155.t002

compared to vectors W, D and D' (Fig. 5b). The vectors D'', D^* and $\check{D}$ contain 12 or 23 summary statistics from the JSFS, whereas W and D have only four and six. Note, however, that the RMSE for estimating divergence time is not affected by the choice of summary statistics (Fig. 5a). For datasets with seven loci, the composite likelihood method J_2 performs better for all coarsenings of the JSFS, as shown by the dramatic decrease of the RMSE for migration rates in Figure 5b. For datasets with 100 loci, estimates of divergence time and especially migration rates are improved compared to the seven loci case (RMSE values in Fig. S11 and Fig. 5). However, note that for 100 loci, the best estimates of migration rates are obtained with our composite likelihood methods J_{3-4} using coarsenings with 23 statistics (D^* or $\check{D}$, Figure S11b).

Discussion

There is growing interest in speciation models and the estimation of the parameters of these models from DNA sequence data. To perform such statistical inferences requires the use of efficient sets of summary statistics to apply to the increasing amount of sequence data [34]. Recent theoretical studies have focused on examining the biases in estimating parameters of the isolation-migration model [5,9] when some key assumptions are violated, such as constant levels of post-divergence gene flow, the absence of population structure, and no migration from an unsampled species [4,18]. Following the approach pioneered by the authors of the MIMAR software, we developed methods to tackle two limitations of existing estimation procedures: the pervasive problem of intra-locus recombination and the often limited number of loci sequenced (around 10) and individuals sampled. These two factors typically represent severe limitations for studying recent speciation in non-model species, such as wild tomatoes [6,23].

The JSFS is a summary of polymorphism data that contains information about the parameters of the isolation-migration model [5,7,19]: the divergence time (τ), the population sizes of the two extant populations (θ_1 and θ_2), the ancestral population size (θ_A), and the migration rates between populations (M_{12} and M_{21}). The likelihood methods of Nielsen and Wakeley [7] and Becquet and Przeworski [4] use four classes of the JSFS to estimate parameters. In addition to these four classes, our coarsenings D', D'', D^* and $\check{D}$ take low-frequency polymorphisms that are shared between populations into account. We show that this provides a significant improvement for estimating the divergence time and gene flow between populations under recent divergence and across a range of intra-locus recombination rates.

Reliable estimates of migration rate and divergence time are linked to variances in the four classes of the JSFS [4,7]. Thus, data sets with many sequences are needed [8]. When only a few loci are sampled, estimates of divergence time and gene flow are correlated [5]. Our novel sets of JSFS-based summary statistics allow to improve the joint estimates of these two parameters, especially when only a small number of loci and SNPs are sampled. In other words, when the information content of the data is limited, one should avoid using a small part of the JSFS and a few summary statistics, because too much information is disregarded (see Fig. 1). Especially in the case of recent divergence, our methods are more accurate than previous ones to disentangle migration from divergence by considering more summary statistics for low-frequency shared polymorphisms. Indeed, if gene flow occurs between diverging species, the rate of gene flow should be low, and this would be reflected by a higher number of shared low-frequency polymorphisms. The use of a more complex summary of the JSFS thus enhances the accuracy of joint parameter estimates of the IM model for any number of sampled loci (for example 7 or 100). Note that in our examples, the simulated 7 loci contain approximately 350 SNPs to emulate date sets obtained from *Drosophila* and wild tomatoes [14,23,24]. This number of SNPs in combination with high recombination rates explains the improvement of statistical accuracy shown by our methods compared to previous ones, except for very recent divergence (where all methods fail).

Our results show in addition that the coverage of the maximum likelihood methods (varying from 64 to 86%) is higher than that of the composite likelihood methods (50%) and MIMAR (around 10%). These results indicate that the MIMAR runs may have converged on local optima and confirm that the chi-square approximation for confidence intervals is applicable to our composite likelihood method [37]. However, even for our maximum-likelihood method, coverage stays below the target value of 95%. We thus advocate that general approaches like parametric bootstrapping would have to be applied for hypothesis testing and to compute confidence intervals in our newly proposed estimation methods [38].

A second quantitative improvement is achieved by developing a simulation-based composite likelihood method that considerably reduces the time of computation compared to MIMAR and our maximum likelihood methods. These methods, as well as full likelihood procedures such as IM [5], require extensive search of the parameter space, which is very time-consuming. Typically, our maximum likelihood methods and MIMAR must run for three to four weeks for a single data set on a standard desktop computer. On a similar machine, popABC can be run for three to four days to generate a table of 300,000 simulations. The rejection and regression steps are then instantaneous. Our composite-likelihood methods require three to four days to generate the JSFS grid of parameter combinations. However, an advantage is that this grid can be used for multiple analyses with the same type of model and identical sample sizes. Note also that our priors can be used for any number of loci, so that the runtime of our composite-likelihood methods does not scale with the number of loci. ABC methods (*e.g.*, popABC) can also re-use a given simulated parameter space if the data sets to be analyzed have identical prior distributions.

Our methods J_{2-4} (with coarsenings D^* or $\check{D}$) provide the most accurate estimates of migration rate. The assumption of independence of sites does not affect the power of these methods over a range of recombination rates (∂a∂i shows a similar behavior). This indicates that methods which take intra-locus recombination into account are also valid when rates of recombination are low [4]. However, the converse is not true. Methods based on the full likelihood analysis of haplotypic data which assume no intra-locus recombination [5,9] are biased if recombination is present [4,18,31]. Another advantage of our composite-likelihood method is that unequal rates of gene flow between diverging species can be estimated (as does ∂a∂i, [22]). Unequal migration rates introduce an asymmetry in the JSFS between the expected numbers of shared low-frequency polymorphisms in each species [22]. Thus, unequal rates of gene flow between species can only be estimated by using a more complex summary of the JSFS than the four Wakeley-Hey summary statistics included in our W vector (W_1, W_2, W_3, W_4).

Estimates of divergence time and migration rates with the ABC method clearly suffer from large overestimates (relative error >50). For popABC extreme overestimates of the divergence time occur when the true value is very low ($\tau<0.1$ in Tables 1 and 2, Fig. S10), independent of the migration rate. Similarly, M_{12} (or M_{21}) is biased under low migration (M_{12} or $M_{21}<0.1$), independent of the

divergence time. In contrast, when using the composite likelihood methods ($J_{1–4}$), large relative errors are observed for estimates of the migration rate M_{12} if the true migration rate is low ($M_{12}<0.1$) and the divergence time is very recent ($\tau<0.1$, Figs. S8 and S9). This means that the summary statistics (whether all 22 or a subset) used in the ABC framework of popABC are not sufficiently sensitive to obtain precise joint estimates of gene flow and divergence time. Furthermore, note as well that popABC does not incorporate an outgroup, which might also explain the reduced information contained in the summary statistics.

We also notice that inaccurate estimation of parameters with popABC following the regression is due to wide posterior distributions. The mode of the posterior estimated by ABCreg [32] was always contained in the posterior calculated by the rejection algorithm in popABC (also based on the best 1% of the simulations; [21]). However, when posterior distributions have wide 95% credibility intervals, the mode computed after the regression step overestimates the true value, especially for migration rates. Wide posterior distributions for divergence time and migration rate estimates occurred when either of these parameters was small (recent divergence $\tau<1$ or small migration $M<0.1$). Estimates obtained with 14 summary statistics are more accurate than those obtained with 11, although they differ only by the inclusion of haplotype diversity in each population and over pooled populations (Fig. 4). This highlights the fact that information contained in haplotype structure helps to disentangle the effects of migration and divergence on genetic diversity. We suggest that an ABC method using more classes of the JSFS such as our vectors D^* or $\check{D}$ (in addition to haplotype diversity), would show better inference of recent divergence times and gene flow, and might be robust over a range of recombination rates.

Finally, we find less accurate estimates of divergence time and gene flow with ∂a∂i than with our composite likelihood methods ($J_{1–4}$; Fig. 4). This is surprising since ∂a∂i is also a composite likelihood approach, in which the expected values of the full JSFS are computed numerically via a diffusion approximation [22]. This method overestimates divergence time, especially for very recent divergence events ($\tau<0.1$), but estimations of migration rate are in line with results from our composite methods and popABC (Table 1 and 2). In other words, when only a few loci are sampled and divergence is recent, the amount of information contained in the JSFS appears to limit the precision of the inferred gene flow parameters. We suggest that our composite-likelihood method based on local regression is more robust to the violation of the assumption that all SNPs are independent than are methods based on diffusion approximations. This would explain the lower accuracy of ∂a∂i compared to our methods. Details of the behavior of ∂a∂i when estimating parameters are, however, beyond the scope of this paper.

In conclusion, we have shown that existing statistical methods to infer speciation parameters in the isolation-migration framework based on the JSFS are improved by more extensive partitioning of the JSFS classes. We have developed a composite-likelihood method that allows to distinguish the signatures of young divergence from those of older divergence time but with recurrent gene flow between populations; these methods are particularly suitable for species with intra-locus recombination and a limited amount of data (less than 20 loci). When analyzing data from two or more diverging populations or species, it should be kept in mind that departures from the stringent model assumptions [5,12,19], such as drawing inference from coding sequences or introns with different selection regimes between species [24], may bias estimates of divergence time, gene flow, and population sizes [18,31].

Supporting Information

Appendix S1 Supplementary information.
(PDF)

Figure S1 Relative error for estimates of (a) the divergence time (τ) and (b) the migration rate ($M = M_{12} = M_{21}$), for the maximum likelihood methods (D_1–D_4), MIMAR and the composite-likelihood methods (J_1–J_4). Relative error is calculated as $(\tau_{est}-\tau_{sim})/\tau_{sim}$ where τ_{est} is the estimated value and τ_{sim} is the simulated value. Groups with significant differences between means following multiple comparisons (Tukey HSD test at 0.05) are indicated by letters for each method (group *a* for the smallest mean). Values that are more than 1.5 times the nearest interquartile range (25% or 75%) are displayed as diamonds, those more than 3 times are displayed as stars.
(TIF)

Figure S2 Analysis of regression between errors in estimates of migration rate ($M_{12} = M_{21}$) and divergence time τ for the 9 methods tested. (a) $D_{1–4}$ for the maximum likelihood methods, (b) $J_{1–4}$ for the composite likelihood methods and (c) for MIMAR. Positive (negative) relative error indicates over (under)-estimation of the parameter. Regression coefficients and p-values are calculated using the *lm* function in the R software. P-values indicate the significance of the test whether the slope of the linear regression is zero.
(TIF)

Figure S3 Factor 2 as a percentage of the estimates of divergence time (τ) in the range $\tau_{sim}/2<\tau_{est}<\tau_{sim}\times 2$ as a function of the population mutation rates (θ), values of simulated migration rates ($M_{12} = M_{21}$) and population recombination rates (ρ). The Factor 2 (F_2) is the proportion of data sets for which the estimated value (of τ or M) is at least half and at most twice the simulated value: (a) for the four maximum likelihood methods (D_1–D_4) and MIMAR, (b) for the four composite-likelihood methods (J_1–J_4) and MIMAR.
(TIF)

Figure S4 Factor 2 as a percentage of the estimates of migration rate ($M = M_{12} = M_{21}$) in the range $M_{sim}/2<M_{est}<M_{sim}\times 2$ as a function of the population mutation rate (θ), values of simulated migration rates ($M_{12} = M_{21}$) and population recombination rates (ρ). (a) For the four maximum likelihood methods (D_1–D_4) and MIMAR, (b) for the four composite-likelihood methods (J_1–J_4) and MIMAR.
(TIF)

Figure S5 Factor 2 for estimates of the divergence time and migration rates (M_{12}, M_{21}) for the four composite-likelihood methods (J_1–J_4), ∂a∂i and for popABC with 6, 11 and 14 summary statistics (computed over 100 datasets).
(TIF)

Figure S6 Distribution of relative error for (a) divergence time and for (b) migration rate depending on the population mutation rate (θ) for composite-likelihood method J_4. For clarity, only relative errors lower than 15 are shown in (b).
(TIF)

Figure S7 Distribution of the relative error of (a) divergence time and of (b) migration rate depending on the population recombination rate (ρ) for composite-likelihood method J_4. For clarity, only relative errors lower than 15 are shown in (b).
(TIF)

Figure S8 Relative error for estimation of migration rate depending on the simulated value of the migration rate (M_{12} in blue and M_{21} in red) for composite method J_2. (a) For simulated divergence times less than 0.5, and (b) for simulated divergence times greater than 1. Note the difference in scale of the y-axes between (a) and (b).
(TIF)

Figure S9 Relative error in the estimation of the migration rate (M_{12} in blue and M_{21} in red) depending on the simulated value of the migration rate for composite likelihood method J_4. (a) For simulated divergence times smaller than 0.5, and (b) for simulated divergence times greater than 1. Note the difference in scale of the y-axes between (a) and (b).
(TIF)

Figure S10 Relative error in the estimation of migration rate depending on the simulated value of the migration rate (M_{12} in blue and M_{21} in red) for popABC estimates with 6 summary statistics. (a) For simulated divergence times smaller than 0.5, and (b) for simulated divergence times greater than 1.
(TIF)

Figure S11 Power analysis of the various JSFS coarsenings to estimate divergence time and migration rates for 100 datasets of 100 loci. RMSE are computed for estimates of the (a) divergence time (τ) and (b) migration rates ($M_{12} \neq M_{21}$) for the four composite-likelihood methods (J_1–J4) based on six vectors of summary statistics with different numbers elements. The vector W is defined by the Wakeley-Hey 4 classes from Eq. 2, and other vectors D, D', D'', D^* and $\check{D}$ are refined decompositions of the JSFS with higher number of classes.
(TIF)

Table S1 ANOVA table of analysis of error in the estimation of divergence times (τ).
(PDF)

Table S2 ANOVA table of analysis of error in the estimation of migration rates ($M_{12} = M_{21}$).
(PDF)

Acknowledgments

We are grateful to Céline Becquet and Ryan Gutenkunst for help with the MIMAR and ∂a∂i simulations, respectively.

Author Contributions

Conceived and designed the experiments: AT PP BH TS WS DM. Performed the experiments: AT PP LN DM. Analyzed the data: AT DM. Wrote the paper: AT PP BH LER LN TS WS DM. Designed the software used in analysis: PP BH DM.

References

1. Hey J (2006) On the failure of modern species concepts. Trends Ecol Evol 21: 447–450.
2. Mayr E (1963) Animal species and evolution. Cambridge, MA, USA: The Belknap Press.
3. Coyne JA, Orr HA (2004) Speciation. Sunderland, MA: Sinauer Associates.
4. Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. Genome Res 17: 1505–1519.
5. 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.
6. 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.
7. Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: A Markov chain Monte Carlo approach. Genetics 158: 885–896.
8. Wang Y, Hey J (2010) Estimating divergence parameters with small samples from a large number of loci. Genetics 184: 363–379.
9. Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci USA 104: 2785–2790.
10. Won YJ, Hey J (2005) Divergence population genetics of chimpanzees. Mol Biol Evol 22: 297–307.
11. Hey J (2010) Isolation with migration models for more than two populations. Mol Biol Evol 27: 905–920.
12. Hey J (2006) Recent advances in assessing gene flow between diverging populations and species. Curr Opin Genet Dev 16: 592–596.
13. Andolfatto P, Wall JD (2003) Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. Genetics 165: 1289–1305.
14. Arunyawat U, Stephan W, Städler T (2007) Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. Mol Biol Evol 24: 2310–2322.
15. Roselius K, Stephan W, Städler T (2005) The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. Genetics 171: 753–763.
16. Stephan W, Langley CH (1998) DNA polymorphism in Lycopersicon and crossing-over per physical length. Genetics 150: 1585–1593.
17. Nordborg M, Tavare S (2002) Linkage disequilibrium: what history has to tell us. Trends Genet 18: 83–90.
18. Strasburg JL, Rieseberg LH (2010) How Robust Are "Isolation with Migration" Analyses to Violations of the IM Model? A Simulation Study. Mol Biol Evol 27: 297–310.
19. Wakeley J, Hey J (1997) Estimating ancestral population parameters. Genetics 145: 847–855.
20. Garrigan D (2009) Composite likelihood estimation of demographic parameters. BMC Genet 10: 72.
21. Lopes JS, Balding D, Beaumont MA (2009) PopABC: a program to infer historical demographic parameters. Bioinformatics 25: 2747–2749.
22. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet 5: e1000695.
23. Städler T, Arunyawat U, Stephan W (2008) Population genetics of speciation in two closely related wild tomatoes (Solanum section Lycopersicon). Genetics 178: 339–350.
24. Tellier A, Fischer I, Merino C, Xia H, Camus-Kulandaivelu L, et al. (2011) Fitness effects of derived deleterious mutations in four closely related wild tomato species with spatial structure. Heredity *in press: advance online publication Jan 2011, doi:2010.1038/hdy.2010.2175.*
25. Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893–903.
26. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.
27. Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet 22: 521–565.
28. R Development Core Team (2005) R: A language and environment for statistical computing. Vienna, Austria: R foundation for Statistical Computing.
29. McCullagh P, Nelder JA (1989) Generalized linear models 2nd edition. London, UK: Chapman and Hall/CRC.
30. Karlis D, Meligkotsidou L (2005) Multivariate Poisson regression with covariance structure. Stat Comp 15: 255–265.
31. Becquet C, Przeworski M (2009) Learning about modes of speciation by computational approaches. Evolution 63: 2547–2562.
32. Thornton KR (2009) Automating approximate Bayesian computation by local linear regression. BMC Genet 10: 35.
33. Joyce P, Marjoram P (2008) Approximately sufficient statistics and Bayesian computation. Stat Appl Genet Mol Biol 7: article 26.
34. Nunes MA, Balding DJ (2010) On optimal selection of summary statistics for Approximate Bayesian Computation. Stat Appl Genet Mol Biol 9: article number: 34.
35. Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM, et al. (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics 24: 2713–2719.
36. Jensen JD, Thornton KR, Aquadro CF (2008) Inferring selection in partially sequenced regions. Mol Biol Evol 25: 438–446.
37. Pawitan Y (2001) In all likelihood: Statistical modelling and inference using likelihood. Oxford, UK: Oxford University Press. 525 p.
38. Efron B (1985) Bootstrap confidence intervals for a class of parametric problems. Biometrika 72: 45–58.

Chapter 3

Distinguishing Gene Flow from Effects of Violating Infinite-Sites Model Assumptions with an Application to *Solanum chilense* and *S. peruvianum*

to be submitted

Lisha Naduvilezhath, Paul R. Staab, Laura E. Rose, Dirk Metzler

Abstract

With the advent of "next generation" sequencing technologies, large data sets of several thousand loci from multiple individuals across populations are now available. These sequence data sets allow us to study more complex models, *e.g.* include finite-sites models (FSM), and reliably estimate demographic parameters using fast composite-likelihood methods. We conducted an intensive simulation study to evaluate the effects of neglecting the assumption of infinite sites and concluded that with increasing population mutation rates θ, divergence times and migration rates were severely overestimated, whereas θ itself was underestimated. Here we present a new and fast version of the composite-likelihood method Jaatha which can jointly estimate more than four demographic parameters also under FSMs. With simulated data we show that Jaatha can estimate FSM parameters such as the mutation rate heterogeneity accurately if enough loci are available. We applied the method with an FSM to estimate divergence time parameters of *Solanum chilense* and *S. peruvianum*. A likelihood ratio testing approach uncovered a significant evidence for gene flow following the divergence of both species $\approx 1.28N_e$ generations ago, where N_e is the effective population size of *S. chilense*.

3.1 Introduction

In recent years a great number of reports on whole genome data sets have followed the advent of new sequencing technologies (*e.g.* pyrosequencing, Margulies *et al.*, 2005). Examples are the onset of the human 1000 genomes project (1000 Genomes Project Consortium, 2010) and the 1001 genomes project of *Arabidopsis thaliana* (Weigel and Mott, 2009; Cao *et al.*, 2011). Though less impressive in number of genomes, sequenced whole-genome data is available from several other organisms, *e.g.* from the fruit fly *Drosophila* (Begun *et al.*, 2007), mouse *Mus musculus* (Keane *et al.*, 2011), and *Escherichia coli* (Lukjancenko *et al.*, 2010).

These vast amounts of data enable us to estimate parameters of complex models with great precision (Lascoux and Petit, 2010; Keinan and Clark, 2012). These models accommodate the biological information relevant to the study organism to shed light on evolutionary processes, such as speciation (The Heliconius Genome Consortium, 2012). Detailed models might be needed for modeling demography as the first step to inferring natural selection (*e.g.* Clotault *et al.*, 2012). The necessity to account for demography first was pointed out due to its "selection-mimicking" effects on genetic variability (Robertson,

1975; Andolfatto and Przeworski, 2000; Teshima *et al.*, 2006; Siol *et al.*, 2010).

For estimation of parameters of species divergence in the isolation-with-migration framework (Hey and Nielsen, 2004) various approaches have been implemented; including Markov chain Monte Carlo methods such as IM and its further developments (Hey and Nielsen, 2004, 2007; Hey, 2010; Choi and Hey, 2011, including population assignment of samples), LAMARC (Kuhner, 2006), and MIMAR (Becquet and Przeworski, 2007). A hidden Markov model was introduced by Mailund *et al.* (2011) to estimate the divergence time and recombination rates along an alignment of two genomes but without gene flow. More variable in the underlying demographic model are approaches like the diffusion approach ∂a∂i (Gutenkunst *et al.*, 2009), the composite-likelihood method of Garrigan (2009), and approximate Bayesian computation (ABC) methods (*e.g.* Beaumont *et al.*, 2002; Beaumont and Balding, 2004; Bazin *et al.*, 2010).

Full-data methods IM (Hey and Nielsen, 2004) or LAMARC (Kuhner, 2006) have the advantage to include finite sites mutation models (FSM) into the estimation. Well-known examples of FSMs from simple to more complex models are Jukes Cantor (JC), Kimura-2-parameter, Felsenstein 81, Hasegawa Kishino Yano (HKY), and general time reversible (GTR) model (Jukes and Cantor, 1969; Kimura, 1980; Felsenstein, 1981; Hasegawa *et al.*, 1985; Tavaré, 1986). The models differ in the number of parameters: In the JC model all mutation rates from one base to another are the same (1 parameter), while in the GTR model ten parameters are specified, four for the equilibrium base frequencies and six describing the symmetrical mutation rate from one base to another for each of the $\binom{4}{2}$ pairs of distinct bases. In the HKY model, base frequencies, as well as transition and transversion rates are specified (6 parameters).

Some methods (such as ∂a∂i by Gutenkunst *et al.*, 2009 or the analytical framework of Chen, 2012) assume an infinite-sites mutation model (ISM) which makes mathematical predictions not only easier but sometimes only possible (Kimura, 1969; Watterson, 1975). Under an ISM it is assumed that all mutations that have occurred along the sequences since the most recent common ancestor of the sample affect a new site and all mutations are therefore visible in the data set. The challenge is how to treat sites in which multiple hits can be detected and consequently clearly violate ISM. These columns represent sites that have been hit by at least two mutations so that the ISM is violated. A common approach is to exclude these sites from further analysis. This procedure may be reasonable if few positions show multiple hits; but ideally one should consider a model of sequence evolution under an FSM. Furthermore back mutations will not be visible at all in the data set, but should also be considered for an accurate estimate of the population

mutation parameter θ. Desai and Plotkin (2008) concluded that when the population mutation parameter per site exceeds 0.05 neglecting back mutations and multiple mutations (in the following termed *neglecting finite sites*) will introduce a bias which might lead to too many false positives in testing for selection. For the special case of molecular hyperdiversity, *i.e.* when pairwise differences at synonymous sites exceeds 0.05, as *e.g.* in the nematode *Caenorhabditis* sp. 5, Cutter *et al.* (2012) have described the effects of FSM and sampling schemes on various population genetic summary statistics. They conclude that different sampling schemes should be considered in the analysis.

Another important aspect of generating finite-sites data is the mutation rate heterogeneity between sites. Rogers and Harpending (1992) showed that based on the shape of the distribution of the number of sequence polymorphisms, it is possible to estimate the timing and extent of population expansion. They applied this approach to study human mitochondrial data but assumed infinite sites. Subsequently several authors noted that the ISM assumption is not met in the case of mitochondrial data (*e.g.* Lundstrom *et al.*, 1992; Aris-Brosou and Excoffier, 1996; Schneider and Excoffier, 1999). Furthermore, Aris-Brosou and Excoffier (1996) observed that mutation rate heterogeneity affect the number of segregating sites in a similar way as a recent population expansion. Using an ISM instead of an FSM with mutation rate heterogeneity (modeled by a Γ distribution with shape parameter α and denoted by "+Γ", as *e.g.* HKY+Γ) can lead to deviations in parameter estimation up to 20% in a simple expansion model and can have a severe effect on the estimation of (bootstrap) confidence intervals (Schneider and Excoffier, 1999). In *A. thaliana*, models including variable mutation rates fitted better than models without (François *et al.*, 2008). In this case variation in mutation rates was modeled by assigning each of the n loci a locus-specific mutation rate μ_i and the mutation rate hyperparameter $\mu = \sum_{i=1}^{n} \mu_i$ was estimated. Recently Martincorena *et al.* (2012) investigated how the mutation rate varies along various *E. coli* genomes. They concluded that the mutation rate changes non-randomly with lower mutation rates on genes that are either under positive selection or are highly expressed and suggested that the mutation rate itself is a parameter optimized by evolution.

In Naduvilezhath *et al.* (2011), we introduced the composite-likelihood method, Jaatha, which is a method to estimate demographic parameters specifically those of a speciation model of recently diverged species from single nucleotide polymorphisms (SNPs). Although Jaatha is flexible regarding the demographic model, simulating the entire parameter space a priori is only possible with a maximum of four parameters. For more complex demographic scenarios, estimating four parameters might be too limiting. Here

we present a new and version of Jaatha which can now estimate any reasonable number of parameters of a user-defined speciation model of recently diverged species. The main modification to the previous version is that after an initial coarse search, the program simulates more thoroughly parameter space that are important for the observed data set.

With simulated data, we investigate the effects of assuming the ISM although the data is generated under an FSM and find that it can lead to an overestimation of the divergence time and the migration rates, and to an underestimation of θ. We show that parameter estimations improve considerably when applying Jaatha with a finite-site sequence simulator (such as `seq-gen` by Rambaut and Grassly, 1997) and that FSM parameters such as the mutation rate heterogeneity can be estimated precisely.

In the seven loci of the wild tomatoes *Solanum chilense* and *S. peruvianum* analyzed in Naduvilezhath *et al.* (2011), 7.3% of the polymorphic sites (70 positions) showed 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. This high number of affected sites suggests that we should take back mutations and double hits into account when analyzing the *Solanum* data. Although strong species barriers exist between these species and hybrids do not form (R. Chetelat, personal communication), our previous analysis of the seven reference loci yielded significant non-zero migration rates for all models (Naduvilezhath *et al.*, 2011). Here we explore two explanations for observing this apparent gene flow between *S. chilense* and *S. peruvianum*: 1. The detection of migration could be due to a gradual loss of gene flow after the split such that at present no hybridization is possible. This idea was modeled in the "Decreasing Migration“ model (description below). 2. The non-zero estimation of the migration rate might be an artifact of neglecting FSM.

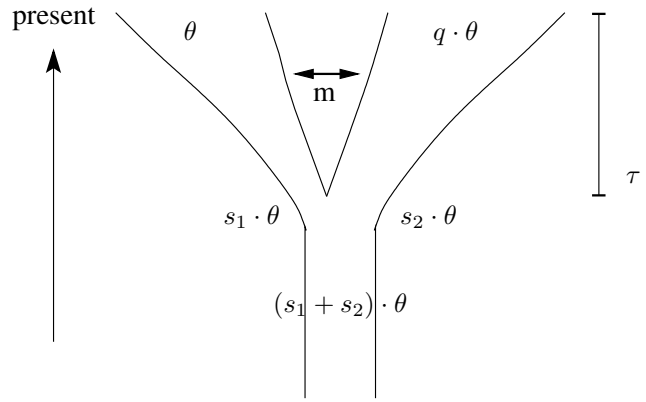


Figure 3.1: **Basic demographic model.** In this speciation model, a single ancestral population splits into two populations P_1 and P_2. In the following all size ratios are relative to N_1, $\theta = 4N_1\mu$ and μ is the mutation rate per generation per locus. P_1 grows exponentially after the split from the size ratio s_1 to its present size and shrinks if $s_1 > 1$. P_2 starts immediately after the split with a size ratio of s_2 and grows or shrinks exponentially to reach the present day size ratio of q. Besides the size ratios q, s_1, and s_2 between the two populations, the model is parameterized by the population mutation rate θ, the divergence time τ, and the symmetric migration rate m. The last three parameters are scaled with $4N_1$ following the parameterization in Hudson's `ms` program (Hudson, 2002).

3.2 Models and Methods

3.2.1 Demographic Models

In the basic model (Fig. 3.1) from which all other models (except "Decreasing Migration" model) are derived, the ancestral population splits $\tau \cdot 4N_1$ generations before present into populations P_1 and P_2, where $\theta = 4N_1\mu$ with N_i denoting the present day effective population size of P_i and μ the mutation rate per locus per generation. Both populations can encounter a size change: P_1 from size ratio s_1 to its present day size and P_2 from size ratio s_2 to its present day size q, where s_1, s_2, and q are size ratios relative to N_1, *e.g.* $q = \frac{N_2}{N_1}$. When $s_1 = 1$ (or $s_2 = q$) no size change occurs in P_1 (P_2). A symmetric migration rate m between the species is assumed, which is scaled with $4N_1$, such that on average each generation $m/4 = \frac{m}{4N_1} \cdot N_1$ individuals replace inhabitants of the other population. The `ms` command line and the parameter ranges from which the values where chosen can be found in Section C.1.1.

In the following we will introduce demographic models that will be later referred to. The fixed and estimated values will be shown in the results table again. With the model "θ/τ" two parameters were estimated, θ and τ. Data sets were simulated under this model

with the following parameters of the basic model fixed to values previously estimated for the tomato data (Naduvilezhath *et al.*, 2011): $s_1 = 1$, $s_2 = 0.3$, $m = 0.5$, $q = 4.5$, and $\rho = 20$ which is the population recombination rate per locus also scaled with $4N_e$.

The models "Constant" and the "Fraction-Growth" contained four parameters to estimate: θ, divergence time τ, present day population size ratio q, and migration rate m, with $s_1 = 1$ fixed. In the model "Constant" we fixed $s_2 = q$, thus disabling population size change in P_2 after the split. In the model "Fraction-Growth" s_2 was fixed to 0.05 and we allowed for population size change.

The following models were fitted to the tomato data: In the model "FixedS2" four *main* parameters are estimated θ, q, τ, and m. The parameters s_1 and s_2 are fixed to 1 and 0.3, respectively, implying size change in P_2 only. The model "NoMig" differs from the model "FixedS2" only in that m is not estimated but kept fixed at 0. In the model "SingleGrowMig" the parameter s_2 is estimated additionally to the ones described for the model "FixedS2", thus allowing for a size change in P_2. In the model "BothGrowNoMig" the migration rate m is set to 0, and s_1 included into the parameter space compared to "SingleGrowMig". In the model "BothGrowMig" two parameters are estimated in addition to the four main ones, s_1 and s_2.

The model "Decreasing Migration" is different from the basic model in that the migration rate m between both populations decreases in two steps from m to zero (Fig. 3.2). The time span with gene flow following the split of both populations is denoted with τ_m, the one without τ_0. At time $\tau_0 + \frac{1}{2}\tau_m$ before present the migration rate is set to half of its value, and at time τ_0 to 0. The `ms` command line is given in Section C.1.2.

3.2.2 Parameters of Finite Site Models

For finite-sites simulations we used the HKY+Γ model (Hasegawa *et al.*, 1985), which is parameterized by the base frequencies, the Γ-shape parameter α, and a transition-transversion ratio κ. The mutation rate heterogeneity between sites in a locus is commonly modeled with the the shape parameter α of a Γ distribution (scale parameter β is fixed to $1/\alpha$ Yang, 1996). The lower α gets the more the mutation rate varies between sites. The transition-transversion ratio κ is defined such that it equals 0.5 if both occur with the same rate. Since we also need an outgroup sequence in Jaatha to determine if a site is derived or not, we simulated outgroup sequences which diverged $T \cdot \tau$ from the ancestor of the two populations.

We now define the "*Solanum* configuration" as follows: The nucleotide frequencies

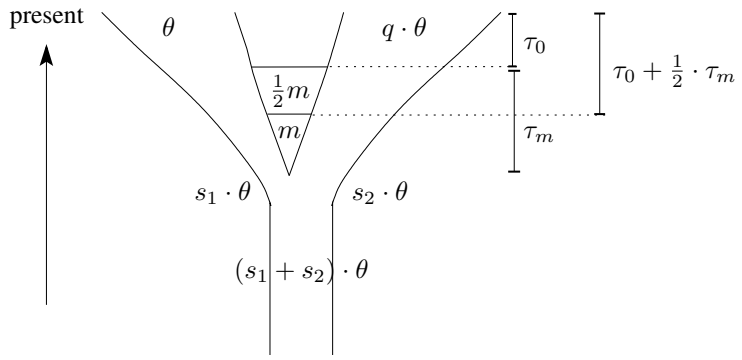


Figure 3.2: **"Decreasing Migration" model.** Up to seven parameters of this model were estimated: population mutation rate θ, divergence time τ, size ratio between the present day population sizes q, starting size of P_1 and P_2 relative to N_1 immediately after the split s_1 and s_2, symmetric migration rate m following the split, and two times, τ_0 and τ_m. Characterizing the migration behavior from the past to the present, directly after the split during the time span τ_m there was symmetrical gene flow between the two populations with rate m and $0.5 \cdot \tau_m$ later with $0.5 \cdot m$. During the most recent time span τ_0 there was no migration between the populations. All population sizes are again relative to that of P_1.

$p(.)$ are set to those observed in the *Solanum chilense* and *S. peruvianum* data set described later: $p(adenine) = 0.26$, $p(cytosine) = 0.20$, $p(guanine) = 0.22$, and $p(thymine) = 0.32$. $\kappa = 2$ for the simulations based on 1.6 which is observed when comparing the tomato data to the outgroup sequence, but likely to be higher in reality because of invisible back mutations. The divergence time factor T is set to 2. If the Γ-shape parameter α was not estimated , it was set to 0.7[1] (for parameter range see Sec. C.1.3).

3.2.3 New Jaatha Version

The aim of Jaatha is to estimate a set of n parameters of a speciation model of two species P_1 and P_2 from a SNP data set D. We summarize the data set D with a set of summary statistics (SS) on the two dimensional joint site frequency spectrum (JSFS) J. The JSFS counts the number of single nucleotide polymorphisms (SNPs) in D for which the derived allele occurs in each population, *e.g.* $J[a, b] = j_{ab} = 5$ means that there are 5 positions in

[1] 0.07 is the average of the values across loci suggested by `Modeltest 3.7` (Posada and Crandall, 1998), with values for the different *Solanum* loci ranging from 0.46 to 1.09. `Modeltest` is typically applied on phylogenetic data sets to test between a fixed set of sequence evolution models assuming no gene flow. However, Städler *et al.* (2008) and Naduvilezhath *et al.* (2011) find indications of gene flow in the *Solanum* data.

D at which the derived allele is found in exactly a individuals of P_1 and in b individuals of P_2. On the JSFS we define a set of SS $\mathbf{S} = (S_1, \ldots, S_{n_{SS}})$, where $S_i(J) = \sum_{(a,b)\in A_i} j_{ab}$ and $A_1, \ldots, A_{n_{SS}}$ is a partition of A and $A = \{0, \ldots, m_1\} \times \{0, \ldots, m_2\} \setminus \{(0,0), (m_1, m_2)\}$ with m_i being the sample size of P_i. A description of the A_i we used with $n_{SS} = 23$ can be found in Naduvilezhath *et al.* (2011). This set of SS serves here as a default.

Since Jaatha draws parameter values from the log-scaled parameter space, the following parameter values are specified on log scale, as the set of true parameter values $\mathbf{p} = (p_1, \ldots, p_n)$ that we want to estimate. Jaatha is a composite-likelihood method which means that the likelihood is approximated by assuming unlinked SNPs (Kim and Stephan, 2000; Hudson, 2001; McVean *et al.*, 2002). Further we assume that the SS are $S_i \sim Pois(\widehat{\lambda}_i(\hat{\mathbf{p}}))$ such that we can calculate the composite likelihood for a parameter combination $\hat{\mathbf{p}}$ by

$$\begin{aligned} L_{s_1,\ldots,s_{n_{SS}}}(\hat{\mathbf{p}}) &= P(S_1 = s_1, \ldots, S_{n_{SS}} = s_{n_{SS}} | \mathbf{p} = \hat{\mathbf{p}}) \\ &\approx \prod_{i=1}^{n_{SS}} P(S_i = s_i | \mathbf{p} = \hat{\mathbf{p}}) = \prod_{i=1}^{n_{SS}} \frac{\widehat{\lambda}_i(\hat{\mathbf{p}})^{s_i} \cdot e^{-\widehat{\lambda}_i(\hat{\mathbf{p}})}}{s_i!}, \end{aligned} \tag{3.1}$$

where $\widehat{\lambda}_i(\hat{\mathbf{p}})$ is our estimate for the expected value $\mathbb{E}S_i$. For the calculation of $\widehat{\lambda}_i(\hat{\mathbf{p}})$ we first simulate data sets in a specific parameter space $\mathcal{B}$ for which we calculate the SS of a simulated data set $\widehat{\mathbf{S}}$ and then fit to each of the $\widehat{S}_i$ a Poisson generalized linear model (GLM) with log-link using the *glm()* function in R (R Development Core Team, 2009). These GLMs describe how the simulated $\widehat{S}_i$ depend on the log-scaled parameters $\hat{\mathbf{p}}$ in $\mathcal{B}$. The parameter values $\widehat{\mathbf{p}}$ in $\mathcal{B}$ that maximizes the approximate Poisson probability $L_{s_1,\ldots,s_{n_{SS}}}(\hat{\mathbf{p}})$ (eqn. 3.1) of $\mathbf{S}$ are determined with the *optim()* function in R using the optimization procedure of Byrd *et al.* (1995) and the middle of $\mathcal{B}$ as the starting point.

The new version of Jaatha consists of an initial and a refined search: First we find good starting positions by simulating very coarsely across the entire parameter space. Taking a set of best starting points chosen by their score $\mathcal{Z}$ we can then conduct a more thorough search to fit GLMs to the simulated $\widehat{\mathbf{S}}$ in smaller regions of the parameter space and thus improving the fit of the GLMs and consequently the likelihood approximations. $e^{\mathcal{Z}}$ is proportional to the likelihood in equation 3.1. In the following paragraphs we will give a more detailed explanation of the two phases and the variables that can be specified by the user.

1. Initial Search: Finding good starting positions

First we divide the parameter space into equally-sized blocks by dividing each parameter range $[min_{p_i}, max_{p_i}]$ into δ intervals such that we obtain δ^n blocks with min_p and max_p being the minimum and maximum of the parameter range for parameter p_i and $i \in [1, n]$. Within each block we simulate s_{ini} data sets of n_{loc} loci with, on the log scale uniformly (in the following simply uniformly drawn) drawn parameter values within each block. To ensure a better sampling of the edges, we additionally simulate data sets for all corner points of each parameter block. For all data sets we calculate $\widehat{\mathbf{S}}$ and fit GLMs to them. With these GLMs within each block we can find the parameter combination that maximizes the score of the observed SS. Each of the δ^n blocks provides a single best parameter combination. Out of this list, n_{RP} starting positions (default $n_{RP} = 10$) points with the highest score $\mathcal{Z}$ $\{\tilde{\mathbf{p}}_1, \ldots, \tilde{\mathbf{p}}_{n_{RP}}\}$ are selected to run the in-depth-search on.

2. Refined Search: Finding n_{RP} best point estimates
   For $b \in \{1, \ldots, n_{RP}\}$ do:

   (a) Assembling a list $\mathcal{L}$ of best parameter estimates starting from $\tilde{\mathbf{p}}_b$:
   Around $\tilde{\mathbf{p}}_b$ we perform a *Jaatha step* to obtain $\tilde{\mathbf{p}}'_b$: First we define a block $\mathcal{B}_{\tilde{\mathbf{p}}_b} = [\tilde{\mathbf{p}}_b - \mathbf{r}, \tilde{\mathbf{p}}_b + \mathbf{r}]$, where $r_i = r$ for all $i \in 1, \ldots, n$. r is set by the user. Within this block $\mathcal{B}_{\tilde{\mathbf{p}}_b}$ we simulate s_{main} data sets of n_{loc} loci with uniformly chosen parameters from within this block (corner points in addition), calculate $\widehat{\mathbf{S}}$, fit GLMs as described above, and estimate a new optimal parameter combination $\tilde{\mathbf{p}}'_b$. Then around $\tilde{\mathbf{p}}'_b$ we run a Jaatha step to find $\tilde{\mathbf{p}}''_b$. For the GLM fitting to find $\tilde{\mathbf{p}}''_b$ we only reuse simulations of previous blocks if $\tilde{\mathbf{p}}'_b$ falls within the block, otherwise the simulations are deleted from memory. Especially for the FSM runs this was necessary to reduce the amount of memory usage. This procedure is iterated and the search stops when the score of the new parameter combination failed to change over the last t_{stop} steps by at least ϵ in a single step. The maximum number of steps can be specified as another stopping criterion (t_{max}) which was necessary in particular when ϵ was small such that the score did not seem to converge. Throughout this phase we keep a list ($\mathcal{L}$) of n_B parameter combinations with the highest scores.
   To avoid being trapped in local maxima there is an option to weigh simulations of previous blocks with $w \in [0, 1]$. Each time simulations of a block are kept, we multiply the weight of these simulations in the GLM fitting by w, such that if $w = 1$ all simulation results have the same contribution in each step.

   (b) Evaluation of the parameter estimates in $\mathcal{L}$:

After phase 2 (a) has finished, the parameter combinations which were stored in $\mathcal{L}$ will be used to simulate s_{final} independent simulations for each of them to calculate the (composite-)likelihood of each parameter combination (using eqn. 3.1) with

$$\widehat{\lambda}_i(\mathbf{p}.) = \sum_{j=1}^{s_{final}} S_{i,j},$$

where $p. \in \mathcal{L}$ and $S_{i,j}$ is the i-th SS of the j-th simulation. The parameter combination with the highest likelihood will then be reported as the result for b.

Since we start the detailed search for each of the n_{RP} refine points, Jaatha will report n_{RP} parameter combinations in total. The Jaatha results in the following always represent the parameter combination with the overall highest likelihood.

Another option that can be set by the user is ext_θ, which specifies whether θ is excluded from the parameter range from which the random values are chosen for the simulations. If this option is set, θ is fixed to the value of 5 for the simulations, which reduces the dimension of block $\mathcal{B}$ by one while the other parameters are calculated as described above. θ is then estimated separately of the other parameters as in Naduvilezhath *et al.* (2011) with

$$\widehat{\theta}_{\mathcal{B}} = \frac{\sum_{i=0}^{n_{SS}} s_i}{\sum_{j=0}^{n_{SS}} \widehat{\lambda}_i(\hat{\mathbf{p}}_{\mathcal{B}})/(5 \cdot n_{loc})}, \tag{3.2}$$

where the parameter combination of the block center of $\mathcal{B}$ is denoted by $\hat{\mathbf{p}}_{\mathcal{B}}$. This approach however, is only reasonable when ISM assumptions are met.

An implementation of the algorithm can be downloaded as an R package (R Development Core Team, 2009) from `http://evol.bio.lmu.de/_statgen/software/jaatha/`.

Optimization of Jaatha Settings

To test the influence of six different Jaatha options (δ, s_{ini}, s_{main}, r, ϵ, and w) on the accuracy and the run time, we conducted an analysis in which the two parameters θ and τ of the "θ/τ" model were estimated. The data sets consisted of 100 loci simulated under an ISM with 25 samples per population. Four values each of τ and of θ were chosen on a uniform grid from the log-transformed parameter range described in Section C.1.1. For

each of the above mentioned settings three values were tested: $\delta \in\{2, 3, 4\}$, $s_{ini} \in\{100, 200, 300\}$, $s_{main} \in\{200, 400, 600\}$, $r \in\{0.05, 0.1, 0.2\}$, $\epsilon \in\{0.5, 1, 2\}$, and $w \in\{0.7, 0.9, 1\}$. Each of the 729 ($= 3^6$) program-setting combinations were tested on 16 data sets (one for each θ-τ combination) such that in total 11,664 runs were evaluated. The other Jaatha settings were kept fixed at $n_{SS} = 23$, $t_{stop} = 5$, $n_{loc} = 70$, $n_B = 10$, $ext_\theta = true$, $s_{final} = 200$, $t_{max} = 200$, and $n_{RP} = 10$. The accuracy was measured for each parameter $p \in \{\theta, \tau\}$ in terms of the root mean squared error (RMSE) between the simulated p_{true} and estimated value p_{est}:

$$RMSE(p) = \frac{\sqrt{(p_{est} - p_{true})^2}}{p_{true}} \tag{3.3}$$

Decreasing Migration Model

As an example of a model with seven parameters, we assessed the accuracy of the parameter estimation in the "Decreasing Migration" model (Fig. 3.2). We simulated 100 ISM data sets of 200 loci with uniformly chosen parameter values on the log-scaled parameter range (for details of parameter ranges and `ms` command see App. C.1.2). To be able to assess the uncertainty of the estimates when applied to *S. chilense* and *S. peruvianum*, we also simulated 100 data sets with seven loci with a HKY model with the "*Solanum* configuration" (for definition see Sect. 3.2.1). On these data sets we applied Jaatha assuming an ISM, therefore neglecting the fact that the data were generated under an FSM. The simulated data sets were analyzed with Jaatha setting `J1` (Tab. C.1) and the *Solanum* loci described later with the same setting except $n_{RP} = 16$.

3.2.4 Applications of Jaatha under Finite Sites Models

Effects of Infinite Sites Violations

To assess the quality of the estimations and to determine which parameters were most affected if we neglect back mutations and double mutations and analyze the data under infinite-sites (IS) assumptions, we conducted the following simulation study: With `ms` (Hudson, 2002) we constructed genealogies based on 100 loci with a sample size of 45 per population under the "Constant" and "Fraction-Growth" model. Along these gene trees we used `seq-gen` (Rambaut and Grassly, 1997) under a HKY + Γ model with the *Solanum* configuration (Sect. 3.2.1), with transition-transversion ratio κ and the outgroup divergence time T variable (Hasegawa *et al.*, 1985). We tested three values of κ: 1, 2, and

5.

For the simulation study, five values for the Γ-shape parameter α were chosen: 0.2, 0.3, 0.5, 0.7, and 1[2]. Hence in total we simulated data under 15 HKY models (3 values of κ, 5 of α). To account for possible variation in the sequences for each genealogy, five sequences were simulated (repetitions). The value of θ for the simulated data sets ranged from 1.25 to 125 per locus (0.001-0.1 per site; for other parameter ranges see Sect. C.1.1). Three different values for T were chosen (1.5, 3, and 6) to see if they had an impact on the results. Jaatha defines a nucleotide to be derived when it is different from the outgroup sequence, independently of which nucleotide was present.

In total, we analyzed 27000 data sets with four methods of Jaatha 0.2 (described in Tellier *et al.*, 2011) under the assumption of an ISM to estimate four parameters. Thus we carried out $1.08 \cdot 10^5$ Jaatha runs (= 15 HKY models, 3 values of T, 2 demographic models, 100 data sets, 3 repetitions, 4 Jaatha methods). This large number of runs was only feasible because we applied Jaatha 0.2, which allows reusing the results of the training phase. In Figures 3.5, C.2, and C.3 the average over the repetitions are plotted.

Using a Finite Sites Sequence Evolution Simulator in Jaatha

To estimate parameters under an FSM with Jaatha we simulated data with `ms` in conjunction with `seq-gen` in the initial and refined search phase (Hudson, 2002; Rambaut and Grassly, 1997).

Choice of Summary Statistics We define seven additional SS, which are supposed to be sensitive for recurrent mutations and evaluate whether including them improves the accuracy ($n_{SS} = 30$). They are not part of the JSFS partition, but may be more sensitive to FSM. We defined them as the number of positions which contained

S_{24}**:** three base types in population P_1 or three base types in population P_2
S_{25}**:** four base types in population P_1 or four base types in population P_2
S_{26}**:** transitions within one population and transversion to outgroup
S_{27}**:** transitions in both populations and transversion to outgroup
S_{28}**:** transversions within one population and transition to outgroup
S_{29}**:** transversions in both populations and transition to outgroup

[2]The estimated α found in the literature for tracheophyte genes ranges between 0.18 and 0.78 and for κ between 2.6 and 5.3 (Soltis *et al.*, 2002).

SS_{30}**:** a base present in at least 95% of the samples in one population and in the other population in at most 5% of the samples

The summary statistics SS_{24}-SS_{29} should contain information about the divergence of the two species and SS_{30} about recent migration events.

To compare the performance of the 23 original SS $SS_1, \ldots, SS_{23}$ with the extended set $SS_1, \ldots, SS_{30}$ and to decide whether to set the option ext_θ, we simulated 25 genealogies with 100 loci each under the "FixedS2" model (Sec. 3.2.1) and $T = 2$. Sequences of 1 kb in length and with two repetitions were generated using the HKY + Γ model with the *Solanum* base frequencies (as in 3.2.4), $\kappa = 3$, and $\alpha = 0.7$. The four parameters to estimate were θ, q, m, and τ. The initial search phase however, was only conducted with $n_{SS} = 23$ and for the refined search the same starting points were chosen for the run with $n_{SS} = 23$ and $n_{SS} = 30$. For the Jaatha application settings `J2` and `J3` with the appropriate n_{SS} were used (Tab. C.1).

Estimating Mutation Rate Heterogeneity For three values of κ (1,2,5), ten sequence files each were simulated with 100 loci and 25 samples per population under the "FixedS2+Γ" model with the *Solanum* base frequencies and $T = 2$ (Jaatha settings `J4` in Tab. C.1). Parameter values for θ, q, τ, m, and α were uniformly drawn from the log-scaled parameter range given in Section C.1.3. We fixed the values of κ in Jaatha to the true κ values with which the data sets were simulated because we believe this estimate can be calculated from the data sets. Only 30 datasets were used in this analysis because including α estimation increases the run time of the sequence simulator. For the Jaatha runs, the previously described 30 SS were used ($n_{SS} = 30$). For a comparison, we also estimated parameters with the ISM with similar settings (`J5` and `J6` in Tab. C.1).

3.2.5 *Solanum* Data Set

S. chilense and *S. peruvianum* are diploid perennial plants that inhabit the Western Coast of South America. All *S. chilense* and *S. peruvianum* analyses were performed on the 7 loci of average gap-free length of 1250 bp (954 SNPs) from the species-wide samples with average sample/allele sizes of 44 for *S. chilense* and 43 for *S. peruvianum* (Arunyawat *et al.*, 2007; Städler *et al.*, 2008). The outgroup for all loci was *S. ochranthum* which diverged from the ancestor approximately 5.8 to 13.6 million years ago (L. Rose, unpublished data). In all FSM-estimations with the *Solanum* loci, we set the base frequencies to the ones observed in the tomato loci (as specified in Sect. 3.2.4) and T and

κ to 2 ("*Solanum* configuration"). In the various analyses if α was not estimated, it was set to 0.7. For the simulations during Jaatha runs the sample sizes were set to the average value across loci for each population.

Is Migration Rate Significantly Different from Zero?

To test whether migration rate was significantly different from zero, we followed a likelihood ratio testing approach with null model having no gene flow (as mentioned by Hey, 2006). For this we calculated the (composite) log likelihood-ratio (ℓLR), *i.e.* ℓLR= $log(\frac{L(''FixedS2“)}{L(''NoMig“)}) = log(L(''FixedS2“)) - log(L(''NoMig“))$, where L is the likelihood of the specified model. This yielded a ℓLR of ≈ 14 for the *Solanum* data. Since the models are not nested we could not apply a χ^2 approximation to calculate p-values but instead used a simulation procedure (Naduvilezhath *et al.*, 2011). We tested how often we would observe such a high or higher ℓLR if the data were simulated under the assumption of no gene flow.

We simulated 50 sequence files with the best "NoMig“ parameter estimates for the *Solanum* loci under the "*Solanum* configuration“ where sample sizes 44 and 43, a recombination rate per locus of 25, and sequence length of 1250 bp. These data sets were then analyzed, as we evaluated the *Solanum* data, under the "FixedS2“ and "NoMig“ model and the ℓLR of the best parameter estimates calculated. The Jaatha setting for these analyses were the same as used for the *Solanum* data (`J7` for the "FixedS2“ model and `J8` for the "NoMig“ model) but with $n_{RP} = 10$ for the "FixedS2“ model.

In Naduvilezhath *et al.* (2011) we also performed a likelihood ratio test comparing two FSMs, which showed significant evidence for gene flow. The difference of the analysis conducted here to the previously used FSM model was that α and κ were not fixed but estimated from the data as well.

Confidence Intervals

For the best fitting model "FixedS2+Γ“ we constructed bias-corrected bootstrap confidence intervals as described by Efron and Tibshirani (1993). We simulated 100 bootstrap data sets of 7 loci with the recombination rate ρ per locus per $4N_1$ generations of 5 which was the lowest ρ value estimated for the tomato loci (Naduvilezhath *et al.*, 2011, Suppl.). Increasing ρ will make the confidence intervals narrower because the data will be more unlinked and thus decrease the variances of the SS. Therefore our test is conservative. The other simulation details were set as in the composite-likelihood ratio test (Sec. 3.2.5).

In Naduvilezhath *et al.* (2011) we had shown with a 7 loci meta-bootstrap analysis that bootstrap confidence intervals have a reasonable coverage probability. To reduce the run time we fixed α to 2.5 which is the *Solanum* estimate under this model.

3.3 Results

3.3.1 New Jaatha Version

Optimization of Jaatha Settings

We simulated datasets of 100 loci under the two-parameter "θ/τ" demographic model to quantify the effects of different Jaatha settings on the accuracy of the parameter estimates and the run time. Decreasing the size of the parameter range from which random samples were chosen for the simulations (r) had the biggest influence on precision of the estimates of τ (Fig. 3.3(a)) but run time increased from an average of 30 minutes (r=2) to 40 minutes (r=0.05, CPU time on a single kernel of a Quad-Core AMD Opteron with 2.7 GHz). The same effect was also visible from a simple linear model in which the run time was the response variable and the different settings the explanatory ones. Decreasing the threshold for the stopping criterion (ϵ) also had a small positive effect, which is hardly visible in Figure 3.3(b). The number of simulations $s_{ini} \in [100, 300]$ and $s_{main} \in [200, 600]$ with Jaatha showed almost no effect on run time and, surprisingly, on the accuracy in the explored ranges (Fig. 3.3(c) and 3.3(d)). Nevertheless other Jaatha runs show that increasing the number of simulations helps when starting from different starting points to converge in a couple of final best parameter estimates (data not shown). And thus increasing the number of simulations especially in the refined search (s_{main}) is wise. But our results suggest that satisfactory results are obtained with the tested number of simulations. The RMSE of τ increased drastically when the true divergence time $\tau = 20$ (Fig. C.1). The effects on the estimation of θ were similar to the ones described above for τ though lower in RMSE value.

In the two-parameter scenario, decreasing the weights (w) of old simulation blocks or dividing the parameter space into more starting blocks (δ) influenced neither the RMSE of the estimates, nor the run time.

Hence if a fast but accurate search is to be conducted, the following setting values might be a good start: $s_{ini} = 100$, $s_{main} = 200$, $r = 0.05$ or even smaller, $\delta = 2$ or 3 depending on the dimension of the parameter range, $\epsilon = 2$. Based on these results, settings of the Jaatha analyses were determined.

The user should keep in mind that this analysis is based on three values for each setting and on a single model for estimating two parameters, but experience has shown that the results are useful for other scenarios as well. However, we point out that including additional parameters adds an extra dimension to the parameter space and advise to choose Jaatha setting values after runs on simulated data.

Decreasing Migration Model

In Figure 3.4 Jaatha estimates of the 7 parameters with 7 (simulated with FSM) and 200 (simulated with ISM) loci are shown. In the 7 loci case, a large uncertainty is associated with all parameter estimates. Surprisingly, the value of τ_0 is never estimated to be greater than ≈ 0.7. The corresponding τ_m on the other hand are mostly overestimated to the upper limit of the parameter range of τ_m, such that the estimate of the divergence time $\tau_0 + \tau_m$ is quite accurate, even with 7 loci (Fig. 3.4(f)). For the 200 loci case, the more recent time τ_0 can be better estimated than τ_m. No obvious connection to migration rate could be seen. In the cases in which τ_0 is not accurately estimated, it negatively correlates with τ_m (Fig. 3.4(i)). If enough loci are available, the parameters θ, q, and τ can be estimated confidently, and with a bit more fluctuation in accuracy, τ_0 and the starting sizes after the split of both populations, s_1 and s_2 can also be estimated. Even with 200 loci, migration rate estimates seem to be uniformly distributed across the parameter range, while in data sets with 7 loci migration rate has a slightly bigger tendency to be overestimated.

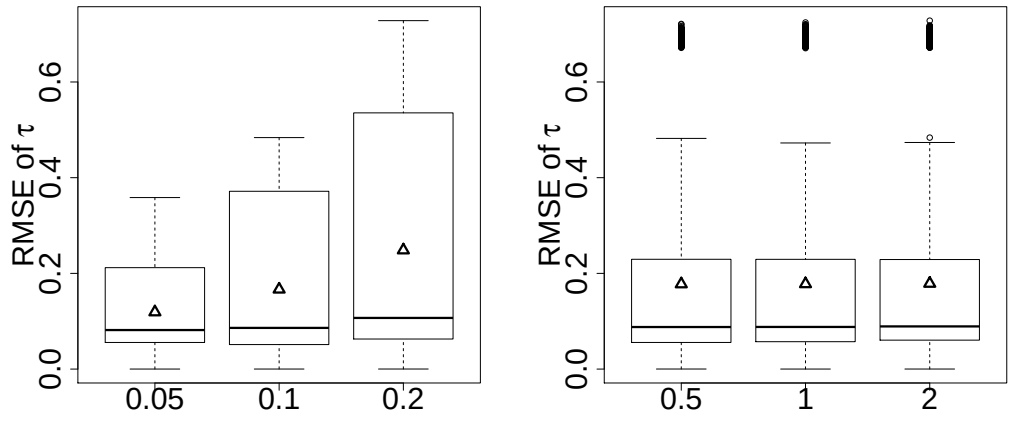
3.3.2 Applications of Jaatha under Finite Sites Models

Violations of the Infinite-Sites Model cause Overestimation of Divergence Time and Migration Rates

To quantify the effect of neglecting finite sites we simulated data under a HKY+Γ model with varying parameter values and analyzed the data under an ISM, thus neglecting back and multiple mutations. As expected, with increasing values of the true population mutation rate, θ was increasingly underestimated (Fig. 3.5(a)). The size ratio q was the least sensitive to increasing values of θ (Fig. 3.5(b)), although for high true values of θ, it was overestimated up to 50% in the "Fraction-Growth" model. The two most affected parameters were the divergence time τ and the migration rate m (Fig. 3.5(c) and 3.5(d)), which were overestimated by up to three orders of magnitude. With increasing values of θ, the number of back mutations increases as well, which leads to a higher variance in all esti-

(a) Half block size r

(b) Stopping criterion ϵ



(c) Number of simulations s_{ini}

(d) Number of simulations s_{main}

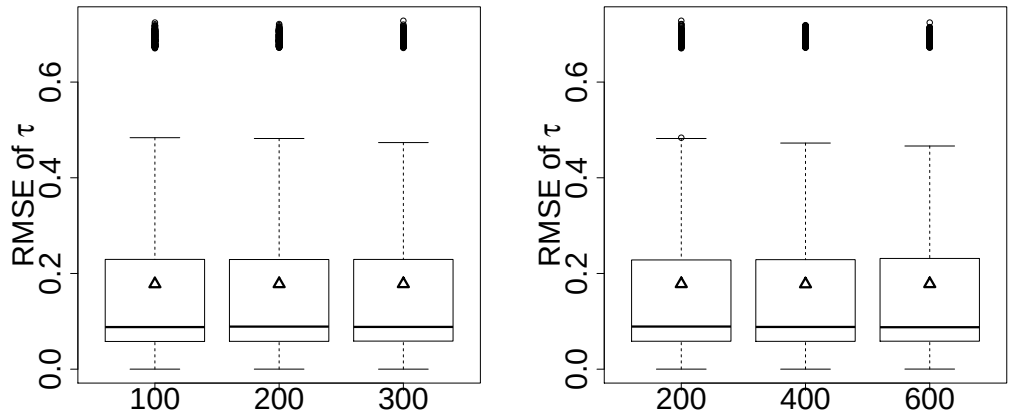


Figure 3.3: **Influence of Jaatha settings on RMSE of divergence time τ.** The mean RMSE is depicted as $\triangle$. Decreasing the size of the parameter range of the simulation area (r) increases the precision of τ estimation. The same influence on accuracy, though much less pronounced, has decreasing the score difference for the stopping criterion of the refined search (ϵ). Decreasing r or ϵ increases run time (data not shown). However, increasing the number of simulations in the initial (s_{ini}) or the refined search (s_{main}) had little influence on the accuracy.

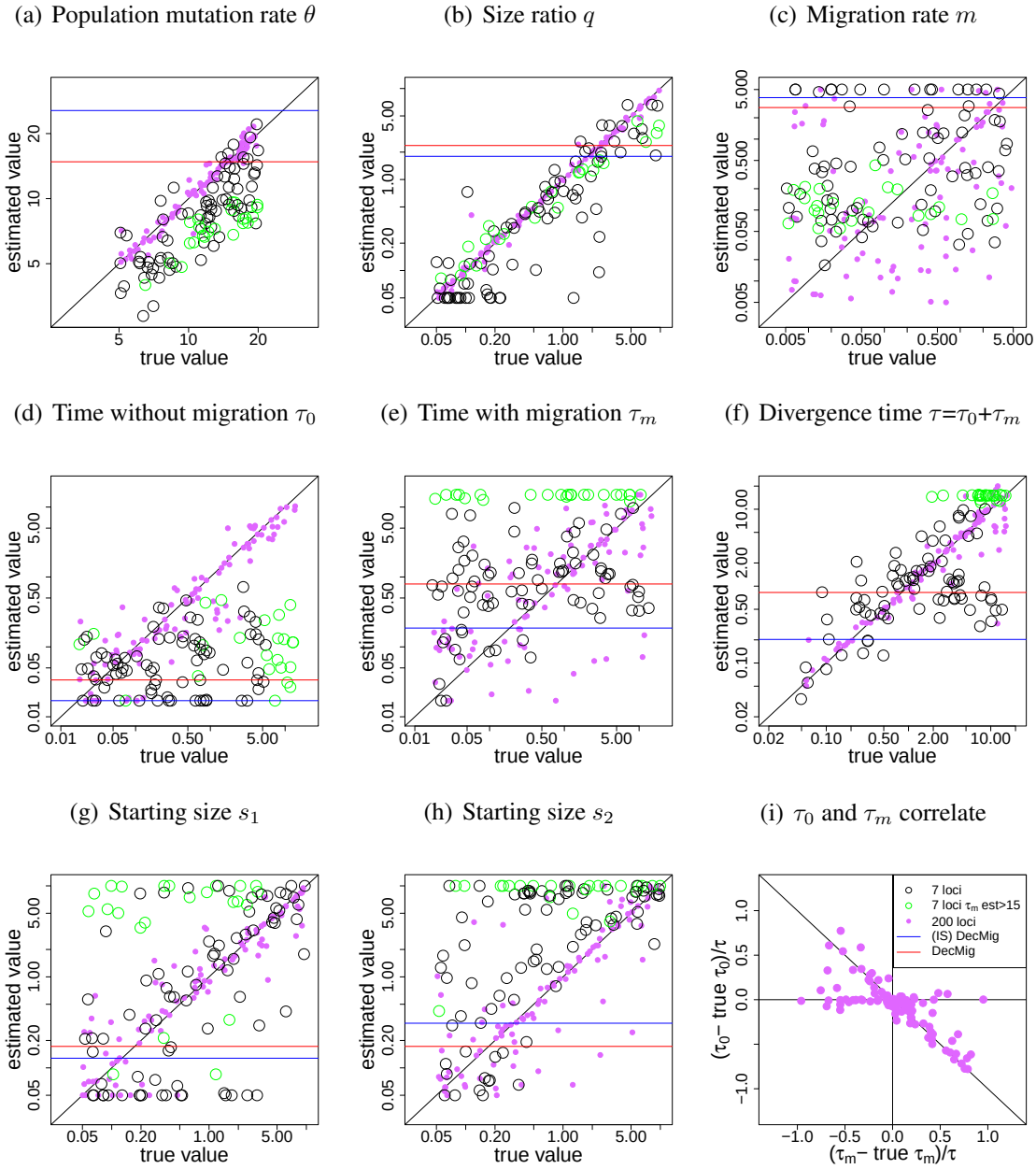


Figure 3.4: **Parameter estimation under the "Decreasing Migration" model with 7 loci is imprecise but improves with additional loci.** Results with simulated data (7 (○) and 200 (•) loci) and tomato loci (colored lines) with the "Decreasing Migration" model with seven parameters. In the case of 7 loci, when τ_m is estimated to be > 15 (○), parameter estimates are particularly imprecise. (d) Further, τ_0 is never estimated to be greater than ≈ 0.7, a behavior that does not occur when 200 loci are used. (f) The divergence time τ is calculated by $\tau_0 + \tau_m$ and is more precisely estimated than τ_0 and τ_m separately. (i) In the 200 loci case if τ_0 is not calculated correctly the estimates of τ_0 and τ_m correlate negatively such that their sum equals the divergence time τ again.

mations. The misestimation of especially τ and θ (notice the different Y-axes in Fig. 3.5) was particularly severe for low values of α, the mutation rate heterogeneity parameter.

Neglecting back and multiple mutations influenced the misestimations of the parameters in the two demographic models "Fraction-Growth" and "Constant" differently (cp. Figs. 3.5 and C.2). The overestimation of the divergence times was stronger under the "Fraction-Growth" model than under the "Constant" model but the quality of the migration rate estimation did not differ much. Also the size ratio estimate q showed more extreme outliers in the demographic model with population growth than in the one without (cp. Figs. 3.5(b) and C.2(b)). The transition-transversion ratio κ (Fig. C.3), the Jaatha estimation method, and the divergence time factor T had no obvious influence on the estimates.

In general, when θ was above a value of 10 per locus (≈ 0.01 per site), the estimates became much worse compared to the estimates of data sets actually simulated under the correct model used for estimation, the ISM. For data sets with θ estimates above this critical value, we propose that finite sites simulators should be used for the simulation procedure. For data sets with lower mutation rates, bias corrections based on the observed regression lines might be a possibility to obtain results faster, but will be imprecise.

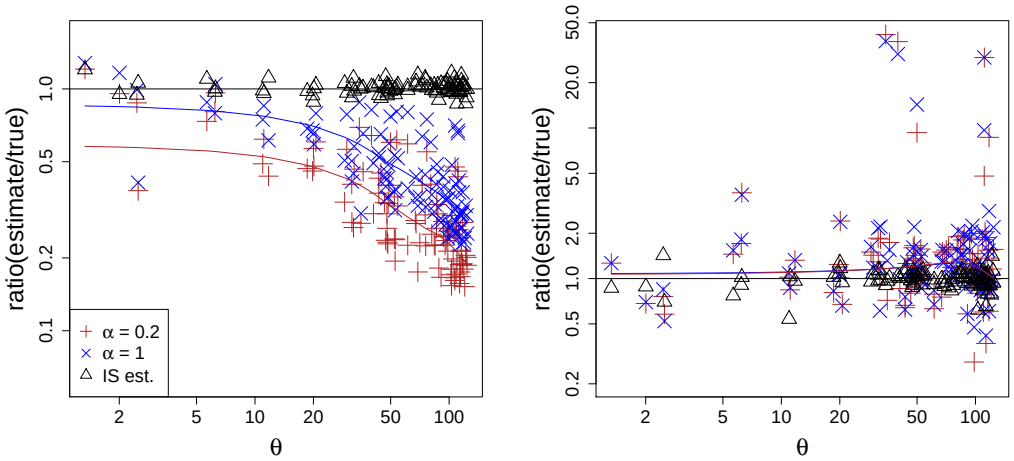
Choice of Summary Statistics

We evaluated whether estimation could be improved by including seven additional summary statistics (SS) that might be more sensitive to FSMs. We also analyzed if, under an FSM, θ could be calculated proportionally to the number of observed segregating sites (ext_θ =TRUE; calculated with eqn. 3.2) or should be estimated as well (ext_θ =FALSE), hence increasing the number of dimensions and the run time. Including θ into the optimization range improved the estimates (cp. in Fig. 3.6 results marked with "ext" and without). Surprisingly, increasing the number of SS increased the precision in θ and q estimates only in the case when θ was calculated externally (ext_θ). There was no improvement in the estimations of divergence time τ or the migration rate m.

Estimating Mutation Rate Heterogeneity

To determine how well Jaatha is able to estimate FSM parameters, especially the Γ shape parameter α, in combination with demographic parameters, we simulated ten data sets each for three different values of the transition-transversion ratio κ (1, 2, 5). If all the 100 loci were generated with the same α, the estimation of α in combination with the

(a) Population mutation rate θ (b) Size ratio q



(c) Divergence time τ (d) Migration rate m

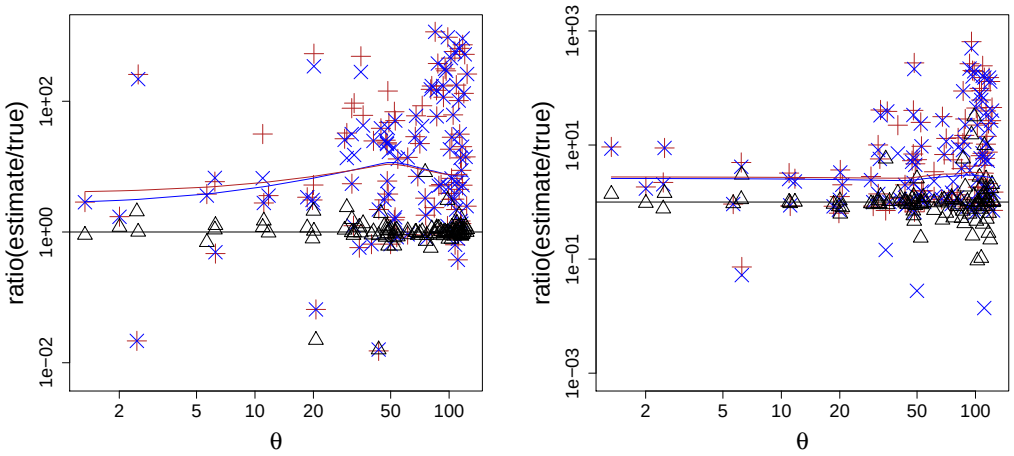


Figure 3.5: **The effect of neglecting finite sites on parameter estimation under the "Fraction-Growth“ model.** The ratio of estimated and true values of θ, q, τ, and m plotted against true θ values under infinite sites assumptions and the "Fraction-Growth“ model. Shown are the data sets simulated with the most extreme α values ($\alpha = 0.2$ and 1), $\kappa = 2$, and $T = 3$. As a comparison, estimates for infinite sites data sets ($\triangle$) are included. The lines plotted are polynomial regression lines fitted to the ratios (with *lowess* function of R). The greatest influence of neglecting finite sites was observed in the estimates of τ and m (notice different scaling of Y-axes).

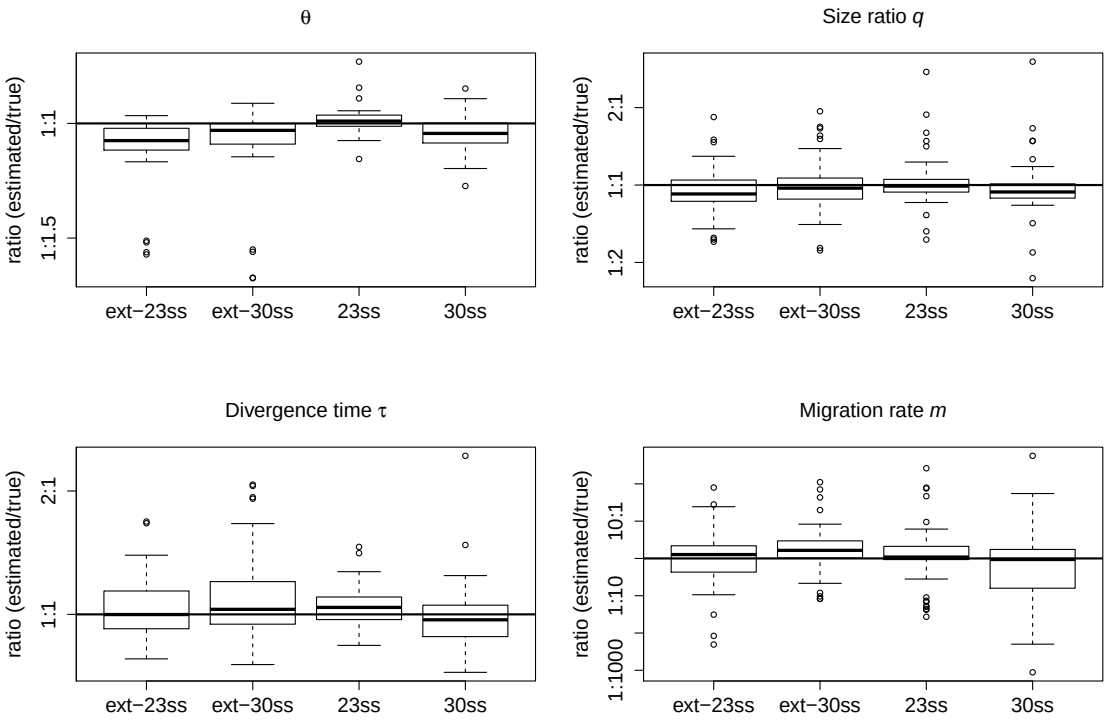


Figure 3.6: **Comparing different numbers of SS and Jaatha setting ext_θ.** Here we compared the usage of 23 and 30 summary statistics (SS) on the same 25 genealogies (each with two sequence simulator runs). Additionally, we assessed the effect of setting the Jaatha option ext_θ, *i.e.* either estimating θ outside of the simulation range with equation 3.2 (marked with "ext") or included θ into Jaatha's optimization range. The best overall results were obtained when including θ into the parameter optimization range and using 23 SS. However, when the option ext_θ was used, including more SS improved the estimates of θ and q.

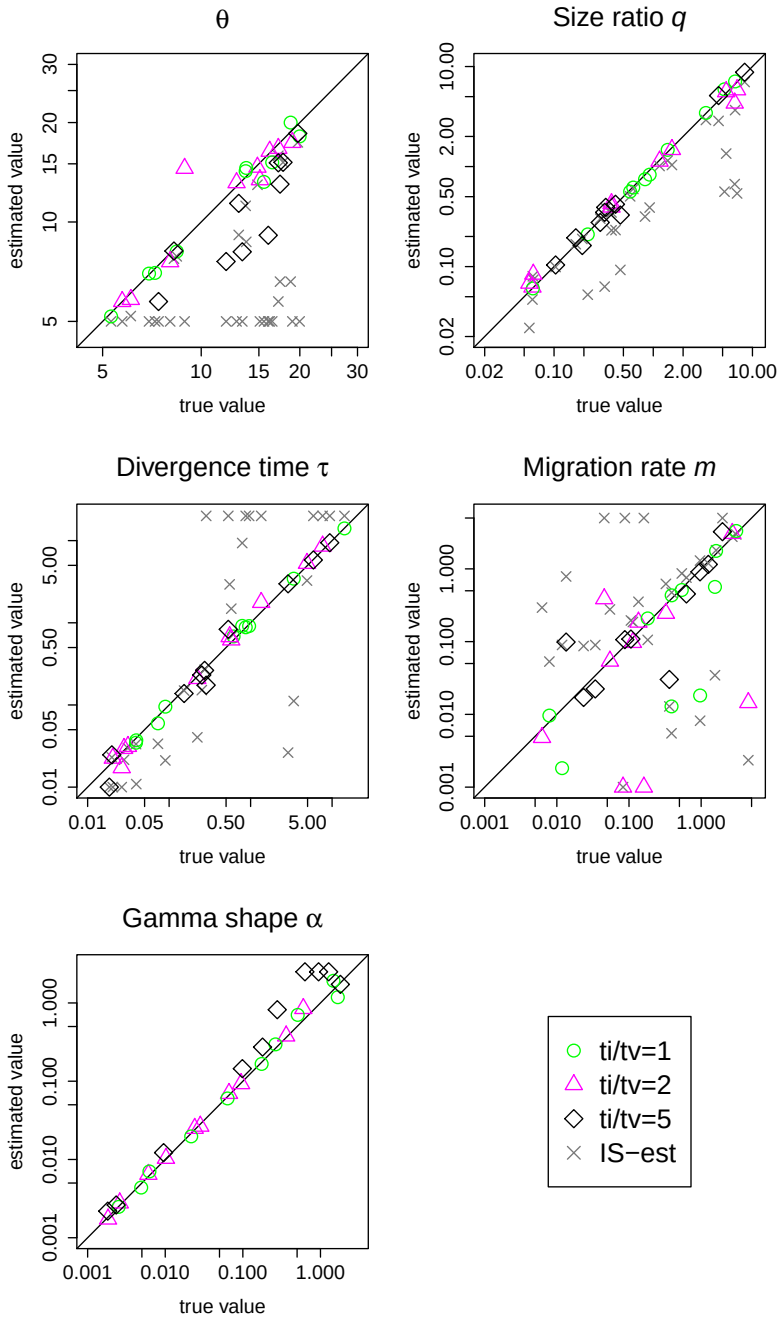


Figure 3.7: **Estimation of Γ shape parameter jointly with demographic parameters.** Here we estimated four demographic parameters and the Γ shape parameter on 30 simulated data sets containing 100 loci, each with 30 summary statistics. For the estimation, the transition-transversion ratios ($\kappa = 1, 2, 5$) were fixed to the true value. Shown are also the estimates of the infinite sites (IS) runs with Jaatha on the same data sets (x). A clear drop in precision of the estimates of all four parameters is observed if an IS model is chosen instead of an finite-sites model.

other parameters was accurate (Fig. 3.7). A great improvement is observed especially in comparison to the results when applying an ISM in Jaatha on the same data sets. However, with the highest κ value of 5, the estimation of θ and of higher values of α got more inaccurate. This suggests further optimizations of the FSM-version of Jaatha are still necessary. For the *Solanum* loci, we estimated $\kappa \approx 1.6$ based on the observed number of transitions and transversions relative to the outgroup and in this κ-range, parameter estimation based on the simulated data sets look robust if enough loci are available. (The ISM runs were run with two Jaatha settings which yielded similar results, therefore only the results with `J5` are shown.)

3.3.3 Application to *Solanum*

Our simulation results in Figure 3.4 indicate that estimating the seven parameters of the "Decreasing Migration model" from only 7 loci is extremely imprecise for all parameters, thus we do not discuss the *Solanum* results for this model further but mention that migration rate is estimated to be extremely high and the time without migration τ_0 to be very recent (Tab. 3.1).

The starting size s_1 of *S. chilense* was estimated below one (0.17), thus indicating population growth in that population after the population split. Could this be an artifact of the extremely high migration rates right after the split in the model or just due to lack of genetic information? To answer this we included three additional models into our analysis: "SingleGrowMig" with population growth in *S. chilense* only and with gene flow between both populations, "BothGrowNoMig" with population growth in both species but no gene flow, and "BothGrowMig" where both populations can grow and there is gene flow. Although the latter two models encompassed more parameters, the "SingleGrowMig" model fitted best to the *Solanum* data. Hence the indication of growth in *S. chilense* is not supported. The two best fitting models in which FSM is applied were the "FixedS2+Γ" and the "SingleGrowMig" model. For the "FixedS2+Γ" with α fixed to 2.5 for the estimation, the following ML estimates [95% bias corrected confidence intervals] were obtained: θ: 13.20 [10.72,16.78], q: 6.05 [3.52,10.51], τ: 0.32 [0.14,0.63], and m: 0.36 [0.11,1.83].

Evidence for Gene Flow even under FSM

We simulated 50 datasets with the "NoMig+Γ" model and analyzed them with both the "FixedS2+Γ" and the "NoMig+Γ" model to see how often Jaatha preferred the true model. Since both models are parameterized by the same number of parameters, this can be done

by computing the log likelihood-ratio (ℓLR) of the models with and without gene flow, *i.e.* the difference of the log composite-likelihood of the gene-flow-model and the log composite-likelihood of the no-gene-flow-model. When the value is positive, the model with migration ("FixedS2+Γ") fitted better and when it is negative, the model without migration fitted better. For the tomato data set, the ℓLR was estimated to be 13.95. The ℓLR of the simulated data sets yielded ratios in the range -7.22 and 14.77 but with only 1 of the 50 data sets preferring the model with gene flow with equal or higher ℓLR than in the tomato data (p-value $= 0.02$). Thus even when allowing for mutation rate heterogeneity by estimating a Γ shape parameter, we still find significant evidence for gene flow between the two species.

Table 3.1: **Estimated parameter values with the seven *S. peruvianum* and *S. chilense* loci.** θ, m, and $\tau_{(.)}$ are scaled with $4N_1$, where N_1 is the effective population size of *S. chilense*. Bold values were fixed for the estimation. In the "+Γ" models α was estimated additionally. * indicates that the value was calculated after the run with τ_0+τ_m. The likelihoods of the ISM estimates are printed in gray because in the tomato data we clearly observe FSM indications and thus the likelihood when using an ISM should be zero. The estimates of τ_0 and τ_m are listed in the lower table. See Tables C.1, C.2, and C.3 for Jaatha settings, additional results with alternative settings, and run times.

Model	θ	q	m	τ	s_1	s_2	α	**# Parameters**	**Likelihood**	**Settings**
NoMig	15.16	10	**0**	0.17	**1**	**0.3**	**0.7**	3	-91.6	`J8`
NoMig+Γ	15.16	10	**0**	0.21	**1**	**0.3**	0.69	4	-83.2	`J7`
FixedS2	14.84	10	0.04	0.22	**1**	**0.3**	**0.7**	4	-81.5	`J9`
(IS) FixedS2	12.50	4.65	0.59	0.39	**1**	**0.3**	-	4	-69.1	`J10`
FixedS2+Γ	13.20	6.05	0.36	0.32	**1**	**0.3**	2.5	5	-69.2	`J7`
SingleGrowMig	13.86	5.67	0.46	0.39	**1**	0.21	**0.7**	5	-69.1	`J11`
SingleGrowMig+Γ	13.20	6.75	0.41	0.29	**1**	0.24	2.5	6	-72.8	`J11`
BothGrowNoMig	15.61	5.13	**0**	0.13	0.42	0.58	**0.7**	5	-94.6	`J11`
BothGrowNoMig+Γ	17.81	3.73	**0**	0.09	0.14	0.19	0.19	6	-294.5	`J11`
BothGrowMig	13.85	4.47	0.75	0.60	0.62	0.03	**0.7**	6	-87.1	`J12`
BothGrowMig+Γ	20	2.41	0.96	0.24	0.10	0.18	1.11	7	-96.8	`J13`
DecMig	14.81	2.36	2.79	0.83*	0.17	0.17	**0.7**	7	-87.1	`J9`
(IS) DecMig	25.67	1.80	3.84	0.20*	0.13	0.31	-	7	-56.3	`J1`

Model	τ_0	τ_m
DecMig	0.03	0.79
(IS) DecMig	0.017	0.19

3.4 Discussion

In this study we introduced a new version of the composite-likelihood method Jaatha, which estimates any number of demographic parameters of a given model from SNP data containing an outgroup sequence. With simulated data we showed that Jaatha gives good results under both finite-sites (FSM) and infinite-sites (ISM) models. In the latter case, it is faster (few hours) than some other methods presently used (*e.g.* ABC approaches as in Sousa *et al.*, 2012 require 10^6 simulations and Jaatha $\approx 5 \cdot 10^4$), such that estimations with a finite-site sequence evolution simulator become feasible.

Many population genetic analyses are based on the ISM assumption (*e.g.* Chen, 2012 or approaches using diffusion approximations like in Gutenkunst *et al.*, 2009). With increasing values of θ, there is a higher likelihood for back and multiple mutations to occur, some of which will not be observed. MCMC approaches as those implemented in LAMARC (Kuhner, 2006) or IM (Hey and Nielsen, 2007) do apply finite-sites models (FSM), however these methods can take several weeks to converge.

Here we investigated the effects of ISM violations on demographic parameter estimation by consider the biologically more realistic scenario of data collected under finite sites scenarios. Our simulations showed that the divergence time and migration rates tend to be overestimated even for moderate values of θ, when the assumption of ISM is not met. While Schneider and Excoffier (1999) showed that departures from an ISM could account for a misestimation of a one-population expansion time of 10 to 20%, we report for a two-population divergence time deviations of more than two orders of magnitude starting from $\theta = 20$ (per locus). If the demographic model included population expansion, the misestimation tended to be larger. Thus, failure to account for back and multiple mutations is particularly severe for populations with high effective population sizes (as it is common in bacteria or the plant kingdom and reviewed in Charlesworth, 2009; Siol *et al.*, 2010) and/or with high mutation rates (high θ values). A possible cause of the overestimation of the migration rates (and consequently of the divergence times) might be the following: If a mutation occurs in one population that would by chance create a pattern like the one in an individual of the other population, this would be interpreted as a migration event by Jaatha. Since Jaatha distinguishes only between the ancestral and derived state, this pattern is easily created. Consequently this behavior also leads to the severe overestimation of the divergence time because a longer divergence time, is needed to account for the differences in both populations.

We have explored a possible solution for dealing with finite-sites (FS) data sets with

Jaatha and show that we are able to estimate mutation rate heterogeneity under several simulation scenarios if enough loci are provided. Simulating the demography using `ms` (Hudson, 2002) and subjecting the simulated sequences to a FS sequence generator such as `seq-gen` provided satisfactory results, especially when θ was included in the optimization range.

θ might be estimated separately from the other parameters, as can be done under the ISM, also under the FSM, if some corrections are applied to θ during the estimation. This will decrease the number of dimensions of the parameter space and the run time. In the following two possibilities are introduced: In equation (1) of McVean *et al.* (2002) a slightly modified version of Watterson's estimator for θ (Watterson, 1975) (per locus) was calculated:

$$\theta_W^* = \frac{1}{a} \ln\left(\frac{L}{L-S}\right) \cdot L, \text{ where } a = \sum_{k=1}^{n-1} \frac{1}{k},$$

where S is the number of segregating sites, L the sequence length, and n the sample size. This can be obtained by computing the first moment of the number of segregating sites S as the product of the sequence length and the probability that a site was hit by at least one mutation along the branches of the genealogy $\left(1 - e^{\left(-\theta_W \sum_{k=1}^{n-1} \frac{1}{k}\right)}\right)$ and then solving for θ_W.

Roychoudhury and Wakeley (2010) proposed an additional adjustment of θ for a few specific scenarios. They proposed an estimate of θ for a K-allele model (in our case K=4 nucleotides) in the special case of so-called *parent-independent mutation*, *i.e.* if the mutation rate from nucleotide b to b' does not depend on b but on b'. Although the result depends on the particular substitution model under consideration, they showed that the number of segregating sites, S, is also a sufficient statistic for an FSM. For example for the simple Jukes Cantor model (Jukes and Cantor, 1969):

$$\theta_W^{**} = \frac{S}{(1 - 1/K) \cdot a}.$$

But for a general mutation model they could only provide an analytical expression in the case of a two-allele model.

In a demographic model with two parameters θ and divergence times τ, large τ values ($\tau \geq 15$) were poorly estimated. Since Jaatha is based on a coalescent simulator (`ms`, Hudson, 2002), if the divergence time is larger than the average time that the two populations need to find their common ancestor Jaatha reaches its limitation. If gene flow is included into the model, greater divergence times could be resolved. Jaatha can be run on

speciation models of two populations with complex demographies. If the number of populations is decreased or increased the summary statistics (SS) might need to be redefined and thus further investigations are needed, though in Tellier *et al.* (2011) we showed for a two-population model the choice of summary statistics (SS) with an ISM did not make much difference.

The choice of summary statistics (SS) for FSM might be a more challenging task. It deserves further consideration because reducing the number of SS would save computation time during the run (*e.g.* with boosting, Lin *et al.*, 2011, or partial least squares (PLS) method, Krämer *et al.*, 2008). We chose 23 SS based on the joint site frequency spectrum (JSFS) plus 7 additional ones which we expected to be more informative about multiple mutations. But in the case of high transition-transversion ratios ($\kappa = 5$) there is still room for further improvement, as might be done with other SS for FSMs.

To decrease the run time of the FSM applications with α estimation there are several possibilities. For example, we have not investigated the option of categorizing the Γ shape (`-g` option in `seq-gen`) as it is commonly done in phylogenetics (Yang, 1996). Alternatives to `seq-gen` which are capable of discriminating between coding and noncoding positions are `indel-Seq-Gen2.0` (Strope *et al.*, 2009) or `SFS_CODE` (Hernandez, 2008). The latter might be a good alternative because in addition to incorporating FSM into complex demographies, it is also able to apply a distribution of selective effects on newly arising mutations, which will be our next step. Siol *et al.* (2010) noted that the JSFS might be especially powerful to detect selection. Furthermore, since composite-likelihood methods require large data sets (Garrigan, 2009; Wiuf, 2006), we believe Jaatha is a powerful tool in this era of next generation sequencing data and look forward to further applications and extensions.

Jaatha was applied to the South American wild tomatoes *Solanum chilense* and *S. peruvianum*. Our estimates for migration are smaller, but still non-zero, compared to our earlier estimates when the finite sites model was not used. Sousa *et al.* (2012) showed in a simulation study under an ABC framework that ABC could distinguish between models with and without migration even with as few as 5 loci. When more loci were available, the confidence in the parameter estimates increased. In light of the results of our LRT and of Sousa *et al.*, we find evidence for speciation in the presence of gene flow in *S. chilense* and *S. peruvianum*, as has been suggested previously (Städler *et al.*, 2008). However, to answer whether gene flow was reduced gradually or not (as modeled in the "Decreasing Migration" model) more sequence data is required.

The size ratio estimate is slightly larger and the divergence time between the two wild

tomato species is more recent (0.32 in units of $4N_1$, N_1 being the effective population size of *S. chilense*) compared to previous estimates. Depending which generation time is chosen, (one or seven years) and a per site mutation rate of $5.1 \cdot 10^{-9}$ (Roselius *et al.*, 2005) divergence time of the two species is either 0.7 million years (My) or 4.6 My. Our analyses suggest that the population structure of *S. chilense* has not changed size since the split (cp. likelihood of model $SingleGrowMig$ of 69.1 and of $BothGrowMig$ of 87.1). Interestingly, the region of the Central Andes where both species co-occur, the Andes underwent a drastic elevation (one third of the present height of the Andes) in the late Tertiary (10 My ago, Jenks, 1975). Around 3-5 My ago, a cooling of the temperatures occurred, leading to the formation of the youngest habitat of the Andes and a unique environment for species radiation (*e.g.* in lupines Smith and Cleef, 1988; Hughes and Eastwood, 2006; Graham, 2009). The timing of the cooling coincides with our divergence estimates of the two species. Therefore environmental changes in the habitat may have allowed for range expansion of the ancestral species and led to the formation of these two distinct present day taxa.

Acknowledgments

We thank Florian Ruland for helping with the management of part of the Jaatha 0.2 runs, Meike Wittmann for valuable discussions and helpful comments on the manuscript, the SuperMuc team for technical support. DFG FOR 1078 for funding to DM and LR and DFG FOR 1078-theoreticians for inspiring questions.

General Discussion

The goal of this dissertation was to develop a method that can estimate demographic parameters for two closely related species and that is able to handle the following challenges: population size changes, recent divergence, gene flow, within-locus recombination, and finite sites data. The implementation of "Jaatha" that is now openly available as an R package on `http://evol.bio.lmu.de/_statgen/software/jaatha/` is presented here. The wild tomato data set of *Solanum chilense* and *S. peruvianum* serve as our application example of Jaatha.

Capabilities of Jaatha

In Chapter 1, we introduced Jaatha, a method to estimate demographic parameters, such as migration rate, population sizes and size changes, and divergence times of two closely related species. As input, Jaatha requires single nucleotide polymorphism (SNP) data and a demographic model for which parameters are to be estimated. Initially for the estimation Jaatha draws parameter values from a user-defined parameter range and simulates data sets, for which summary statistics are calculated. Secondly, generalized linear models (GLMs) that explain how each summary statistics depends on the parameter values are fitted in small areas of the parameter space. These GLMs provide the expectation values for the summary statistics, which are then compared to the summary statistics of the observed data set. The parameter combination that maximizes the composite-likelihood function is finally the point estimate returned by Jaatha.

To assess the uncertainty of the estimates, we calculated bootstrap confidence intervals for each estimate. Since composite-likelihood methods assume unlinked SNPs, an assumption that is not met in general for neighboring SNPs because they tend to share large parts of their genealogy, evaluating the coverage of such bootstrap intervals is important. With an intensive coverage analysis we demonstrated that for all parameters the determined 95 % bootstrap confidence intervals are good approximations, because they

contained the true values in $\approx 95\%$ of cases.

In Chapters 1 and 2, we analyzed Jaatha's performance in comparison to other methods with simulations under different scenarios. We tested the full-data method IM (Hey and Nielsen, 2004, 2007), a likelihood method introduced in Tellier *et al.* (2011), and the summary-statistics based methods *∂a∂i* (Gutenkunst *et al.*, 2009), PopABC (Lopes *et al.*, 2009), and MIMAR (Becquet and Przeworski, 2007). In most cases Jaatha gives comparable results to other methods, and for low divergence times Jaatha outperforms them. This can be explained with the use of summary statistics which are based on the joint site frequency spectrum (JSFS). In particular, for the estimation of divergence times under an infinite-sites model (ISM) the JSFS is powerful: derived sites that are shared between two species arose before the split (or via migration), while polymorphisms that are only present in one of the two populations arose after the split (Chen *et al.*, 2007). The further coarsening of the JSFS in the low-frequencies, as in the default summary statistics set of Jaatha, has further been shown to provide a significant improvement for divergence estimations and migration rates, especially when more loci where available.

Since Jaatha's implementation used in the first two chapters is only feasible for up to four parameters, in Chapter 3, Jaatha was modified to jointly estimate more parameters. This is accomplished with an initial coarse Jaatha search of the entire parameter space, followed by a more thorough search in promising parameter areas, which optimize the composite-likelihood given the observed data.

Now that Jaatha can estimate more parameters, we successfully applied it to estimate additional sequence evolution parameters, such as the mutation rate heterogeneity between sites. The importance of taking finite sites models including mutation rate heterogeneity between sites into consideration has been noted previously (*e.g.* Lundstrom *et al.*, 1992; Aris-Brosou and Excoffier, 1996; Schneider and Excoffier, 1999). In Chapter 3 we quantified the effect of *not* accounting for multiple and back mutations during the estimation and instead assuming an infinite-sites model (ISM). Especially the divergence times and migration rates were overestimated up to three orders of magnitude if the recurrent mutations were neglected and $\theta > 0.016$ per site. With increasing true θ values, θ was underestimated up to fivefold in the tested regions because many multiple and back mutations are not visible in the data sets. The present-day size ratio between the two populations was the least affected parameter because both populations were equally affected by this negligence.

Compared to approximate Bayesian computation (ABC) methods (*e.g.* Beaumont *et al.*, 2002, introduced in General Introduction, p. 11) Jaatha can obtain parameter values under

FSMs because it is based on a smaller number of simulations. While ABC methods use around 10^6 simulations (Sousa *et al.*, 2012), Jaatha needs on average only $4 \cdot 10^4$ which is also why it runs especially fast in ISM cases ($\approx$ two hours for a seven parameter model on a single CPU vs. 1-2 days for ABC methods on several CPUs, Pablo Duchen personal communication).

Although Jaatha is a composite-likelihood method assuming unlinked SNPs, it gives accurate results for simulated data sets in complete linkage, *i.e.* no recombination (Ch. 1). Wiuf (2006) showed that certain composite likelihood estimators, as the ones used in Jaatha, are consistent, which means that the estimator will approach the true value with an increasing number of examined regions. This result was confirmed in our simulation scenarios and in the study of Garrigan (2009). In times of large data sets that are suitable for population genetic analyses *e.g.* from next-generation sequencing technologies, ranging from humans (1000 Genomes Project Consortium, 2010), *Arabidopsis thaliana* (Cao *et al.*, 2011), *Drosophila* (Begun *et al.*, 2007), mouse (Keane *et al.*, 2011), to *Escherichia coli* (Lukjancenko *et al.*, 2010), Jaatha can be used to discriminate between different complex models, with and without gene flow or population expansions.

Since Jaatha is a method that can handle the above mentioned challenges, we envision it to be applied widely. Several characteristics that have been encountered in the *Solanum* data set should also be found in other diverging species. Especially with population genetic data available for several species, Jaatha can be used to confidently estimate parameters of models that appear most likely for the organism of interest. Furthermore, through simulation-based likelihood-ratio testing Jaatha can be used to reject hypotheses of certain modes of speciation, *e.g.* the presence of gene flow.

Limitations of Jaatha

Problems with Jaatha arise if the divergence time is longer than the average time to the common ancestor of both species. In Chapter 3, through simulations we could show that this happens for divergence time larger than $\approx 60N_e$ generations, where N_e is the effective population size of population 1. Thus Jaatha is most useful for species that are recently diverged. The estimates of divergence times need to be considered together with the migration rates because they can have similar influences on the joint site frequency spectrum (JSFS): For example, if divergence time is low, there will be many shared polymorphisms and it will not make a difference on the JSFS how much gene flow is present, such that gene flow in the current version of Jaatha in difficult to assess. Migration esti-

mates improve with increasing divergence times (*i.e.* $> 0.8N_e$) (Ch. 1). This difficulty in estimating migration rates for low divergence times is also encountered with other methods, *e.g.* with IM (Hey and Nielsen, 2007) or $\partial a \partial i$ (Gutenkunst *et al.*, 2009) as shown in Chapter 1.

For almost a decade there has been a heated and extensive debate on Bayesian model choice and hypothesis testing (*e.g.* Knowles and Maddison, 2002; Templeton, 2009; Beaumont *et al.*, 2010; Templeton, 2010a,b; Robert *et al.*, 2011). In the following I will state a few arguments of both fronts and explain if they apply to Jaatha and if there is a possible remedy for it.

Templeton (2009) criticizes that ABC methods assign a fixed set of models relative posterior probabilities, however badly the chosen models may fit to the data. His most recent "attack" on ABC is that the models that are tested against each other during an ABC estimation should be nested, *i.e.* each model should be a special case of the more general one (Templeton, 2010a), or should be mutually exclusive (Templeton, 2010b). As a result, he claims many ABC results are wrong (*e.g.* Templeton, 2010a,b). Although Jaatha is not a Bayesian method, even though we sample the parameter values for the simulations uniformly from the logarithmically scaled parameter range, we apply Jaatha to a fixed set of user-defined models that are sometimes also logically overlapping as in some ABC analyses. But as a measure of fit, the user receives composite (log) likelihoods for the estimated parameter values under each model and no relative fit of the chosen models.

Beaumont *et al.* (2010) pointed out that the likelihood Templeton is computing in his nested clade phylogeographical analysis (NCPA, Templeton, 2004) is not valid and thus should thus not be used for likelihood ratio testing. To assess the relative fit of the models containing gene flow and no gene flow, we do conduct a simulation-based likelihood ratio test (LRT) for the wild tomato data set of *S. chilense* and *S. peruvianum* (Ch. 1 and 3). Jaatha calculates composite likelihoods, assuming that all the SNPs in the data are independent, *i.e.* the composite likelihoods are approximations of the likelihood. Since the two models we tested were characterized by different numbers of parameters we could not apply a χ^2 approximation to compute the p-value. Instead we repeatedly simulated data sets without gene flow and with recombination rates as observed in the data, and analyzed them with the two competing models. Finally the likelihood ratios of the simulated data sets were calculated and compared to the ones of the wild tomato data. The proportion of likelihood ratios of the simulated data sets that was larger or equal to the one of the wild tomatoes was the p-value for the null hypothesis of no gene flow.

Importantly, for the test a realistic recombination rate was chosen based on the observed data and thus, we did not simulate completely unlinked SNPs as would be assumed by Jaatha.

Similarly we applied a LRT to test the null hypothesis of no population expansion in *S. peruvianum*. Furthermore, as demanded by Beaumont *et al.* (2010), in this case we performed a model selection-analysis of Jaatha to examine whether it can correctly discriminate if a data set with just seven loci was coming from a model with or without population expansion. For all the hundred tested cases the p-value was < 0.001, giving strong evidence that Jaatha can satisfactorily distinguish between models with and without expansion. In this extensive simulation study we could not only show that there was significant evidence for population expansion in *S. peruvianum* but, importantly, with LRTs based on composite likelihoods successful model testing can be performed.

Another critique Templeton insists on is to account for the number of parameters in the models ABC is applied on to adjust for the model dimensionality as the Bayesian Information Criterion (BIC Schwarz, 1978) (Templeton, 2010b). As mentioned above Jaatha computes composite likelihoods, hence it would need further investigations if model choice criteria as Akaike Information Criterion (AIC Akaike, 1973) are appropriate to apply to composite likelihoods. However, we have noted this point and report along with likelihood the number of parameters for each model (Ch. 3).

Recently the sensitivity of ABC on the choice of underlying summary statistics was pointed out, *e.g.* by Didelot *et al.* (2011) and Robert *et al.* (2011). These authors suggest when applying ABC to data, the summary statistics should be tested whether they contain enough information to actually distinguish the evaluated models. We support this advice to the user. Also with Jaatha a simulation study with data sets similar in configuration to the observed data should be performed to determine which set of summary statistics is adequate for the scenario of interest. An example of such a model testing-analysis is described above. For parameter estimation with FSMs and low transition transversion rates (ti/tv), we showed that Jaatha's standard set of summary statistics results in accurate estimates. But with larger ti/tv, different summary statistics might be needed. This needs further investigation.

In general, comparisons of full-data and summary statistics based methods provide evidence for the superior performance of full-data methods. Beaumont *et al.* (2002) compared their ABC method to the full-data MCMC method BATWING (Wilson *et al.*, 2003) and observed that in the tested cases, BATWING was superior. They concluded, that when for the tested scenario full-data methods are available, they should be preferred (see

also Hudson, 2002; Wegmann *et al.*, 2009). In our comparison of Jaatha and IM, we also found that IM performed slightly better for the estimation of the population mutation rate θ and the present day size ratio between the two populations (Ch. 1). However, IM did not perform better in estimating low divergence times or migration rates.

For data sets with recombination rates as in the wild tomato data set of *S. chilense* and *S. peruvianum*, IM and its successors have their limitations: Recombination rates as low as 0.005 per bp per $4N_e$ generations resulted in three out of ten cases in which the 90% highest point density (HPD) intervals did not contain the true value for θ (Strasburg and Rieseberg, 2010). When the recombination rate was above 0.02, none of the intervals included the true values (Strasburg and Rieseberg, 2010)[3]. A full-data alternative in cases with recombination might be the new version of LAMARC (LAMARC Team, 2012) which now also accommodates recent divergence. Nonetheless, how the full-data methods can cope with large data sets needs further study (Beerli and Felsenstein, 2001). Especially when many regions are available, we therefore propose that composite-likelihood methods as Jaatha should be considered because of their accuracy and speed.

Results for the Demography of the Wild Tomatoes

The wild tomatoes *S. chilense* and *S. peruvianum* have previously been identified as sister species with a recent divergence time of 0.5 million years if one generation per year was assumed (Städler *et al.*, 2005, 2008). Using the same generation time, mutation rate per locus, and an FSM we get a slightly higher divergence time estimate of 0.7 million years. Previous estimates were smaller because they were obtained with a model not allowing for gene flow. Thus, if gene flow, which homogenizes the differences between populations, is included into the demographic model a longer divergence time is needed to explain the differences. Applying an FSM compared to an ISM only slightly changed the divergence and migration estimates.

Siol *et al.* (2010) hypothesized that in the plant kingdom speciation events often follow a *founder event* that includes a population bottleneck (size decrease) and subsequent population expansion (Mayr, 1942, p. 237). Since we find significant evidence of population expansion in *S. peruvianum* this may point to such an event. Since *S. peruvianum* prefers more mesic habitats, after its divergence from the ancestor it could have expanded its range further into the north starting from where both species co-occur using the "discon-

[3]But the authors demonstrated also that most biases can be avoided when the data is partitioned into apparently nonrecombining blocks.

tinuous north-south migratory pathway for mid-elevation biotas" (Graham, 2009, p. 371). Further into the south one of the most arid areas on earth is found (Navarro-González *et al.*, 2003).

Nakazato and Housworth (2011) have previously noted that the divergence of the wild tomatoes of the Andes are shaped by geography, especially the rise of the Andean mountains in the late Tertiary about 10 million years ago. They investigated the adaptation of *S. pimpinellifolium* and *S. lycopersicum* which inhabit the northern regions of the Andes and conclude that in the study species the complex geography of the Andes and climatic patterns shape their demography (Nakazato *et al.*, 2010; Nakazato and Housworth, 2011). Approximately 3-5 million years ago the temperature in the Andes dropped, which has lead to the formation of one of the most diverse mountain faunas on earth with rapid speciation rates (Smith and Cleef, 1988; Hughes and Eastwood, 2006; Graham, 2009). If we assume an average generation time of three years, our focal species diverged from their common ancestor about 2.0 $(0.9, 3.9)$ ($95\,\%$ bootstrap confidence interval) million years ago. Therefore the speciation event that has lead to *S. chilense* and *S. peruvianum* might have happened as a consequence of the climatic change in the Andes. However, the generation time can not be easily determined and may for example range from one to seven years because seed germination is affected by climatic patterns of the region.

Städler *et al.* (2008) showed previously that there were signs of gene flow between the species. In Chapters 1 and 3 we showed that this evidence was significant under an ISM as well as an FSM, even though only seven loci comprising 954 SNPs were available. Speciation in the presence of gene flow has been reported already in other species, *e.g.* in *Daphnia* (Cristescu *et al.*, 2012). Sousa *et al.* (2012) demonstrated that model choice between a model with and without gene flow could successfully be carried out even with as few as five loci.

Nevertheless, hybrids between the two species have not been observed in the field (L. Rose personal communication with R. Chetelat), suggesting that the hybrids are not viable or do not produce fertile offspring. Thus a demographic model with decreasing gene flow after the split as fitted in Chapter 3 might be a good approach. But as has been stated before the timing of migration events remains difficult to estimate (Becquet and Przeworski, 2009; Strasburg and Rieseberg, 2011), which we also find in Chapter 3, even with 200 loci. A probabilistic explanation of the conclusions of Strasburg and Rieseberg (2011), which were based on coalescent simulations, is given by Sousa *et al.* (2011). They state that two genealogies with different migration timings can have the same posterior distribution. At the end of their commentary, Sousa *et al.* (2011) also provide two solu-

tions how the changes in the migration rate could be modeled: One way could be to use a deterministic function that could model the changes in migration rates after the split. The parameter of the function would then need to be estimated. However, in its current version Jaatha uses the `ms` software (Hudson, 2002), with which a gradual decline could only be incorporated approximately using a step function for the migration rate which we used in Chapter 3. The second approach they suggest is to use different migration rate estimation for each distinct time interval. Although it is now possible to estimate more parameters with Jaatha, the current data set contains only seven loci which does not allow to precisely estimate that many parameters. When in the near future more loci are available, the point estimates will improve. This will allow for estimating parameters with reasonable accuracy in more complex models that could account for population structure (as suggested in Arunyawat *et al.*, 2007) and for different ancestral population sizes (Städler *et al.*, 2008). For the generation of new data, the sequenced genomes of the cultivar *S. lycopersicum* and its wild relative *S. pimpinellifolium* will be advantageous (The Tomato Genome Consortium, 2012).

Future Directions

To optimize Jaatha for the purpose of application to next generation sequencing data, sequencing errors and unphased data should be accounted for. Lynch (2009) showed, for example, that errors in detecting SNPs can have a great influence on demography estimates. Although SNP detection is cheaper and less error prone than microsatellite data (Evans and Cardon, 2004; Kennedy *et al.*, 2003), Jaatha could be extended to accommodate other data types. For instance, microsatellite markers are still widely used in ecological research projects. Consequently, providing a microsatellite version of Jaatha would enlarge the user community.

Recently the impact of sampling schemes under an ISM (Städler *et al.*, 2009) and an FSM (Cutter *et al.*, 2012) have been investigated. Cutter *et al.* (2012) conclude that different sampling schemes produce different evolutionary results and thus species-wide samples as well as population samples should be considered. Different sampling schemes could also be implemented into Jaatha's simulation procedure.

Although in Tellier *et al.* (2011) we showed that Jaatha gave better migration rate estimates than PopABC and MIMAR, there is still potential for improvement. The migration rate cannot be calculated with the same accuracy as *e.g.* the divergence time with most of the examined methods. This suggests at least in the case of $\partial a \partial i$, MIMAR, and

Jaatha that not enough information is contained in the JSFS or the summary statistics used. Gattepaille and Jakobsson (2012) have recently explored taking haplotype information into account, which improved the population assignments of the individuals. The inclusion of haplotype information, as might be done with additional summary statistics on neighboring SNPs, will likely increase the performance of migration rate estimators.

Another slightly different approach to improve the migration rate estimates would be to evaluate the rate of migration by considering the variances in the summary statistics between loci. The idea is that when there is more migration the variance of the summary statistics between different loci should be larger than in the case of smaller migration rates.

A third approach would be to model the variation in contribution of different loci to the total migration rate, as is done in Bull *et al.* (2006) or in MIMAR with recombination rate (Becquet and Przeworski, 2007). The biological motivation behind this is the following: different regions of the genome have different rates of introgression and/or gene flow. For example, Castric *et al.* (2008) found a gene with a five fold higher rate of introgression than its genomic background. This gene controls the pistil self-incompatibility specificity in the two closely related species, *Arabidopsis halleri* and *A. lyrata*. In particular in the initial stages of speciation, as studied with Jaatha, reproductive isolation builds up with a few loci involved which are not receptive for introgression, while the rest of genome still may exchange genes (*e.g.* Städler *et al.*, 2008, and references therein). Such loci however, are likely to be under selection.

Siol *et al.* (2010) hypothesize that the JSFS is particularly useful to detect selection. Thus, a very promising extension of Jaatha would be to include detection of selection, as was done in an ABC framework in Beaumont and Balding (2004). This could for example be done using coalescent simulators that jointly simulate demography and selection, as is done in `msms` (Ewing and Hermisson, 2010) or `SFS_code` (Hernandez, 2008). Different summary statistics might have to be defined that are more sensitive to selection. This may require more complex mutation models that differentiate between synonymous and non-synonymous sites or codon positions.

To extend Jaatha to different numbers of populations, we might need to define another set of summary statistics, which may include more or fewer statistics than in the default set. A good starting point is the application of this default set, in the case of more than two populations, to a multidimensional JSFS, or to the marginal site frequency spectrum, in case of one population. When more than two populations are examined, further summary statistics might also help to increase the accuracy in estimations.

The idea of the method, however, might also be useful for parameter estimation in other research fields. A premise for this is that the scenario can be simulated for different values of the parameters of interest. For instance in the field of systems biology population dynamics are modeled and parameters as the growth rate of a population inferred. Furthermore, ABC has already been applied in this field (Toni *et al.*, 2009; Toni and Stumpf, 2010). Another more distant example could be the estimation of the optimal conformation of a macromolecule (*e.g. proteins* as reviewed in Zimmermann and Hansmann, 2008), which may estimate parameters as the torsional angle. In this case, the energy function needs to be minimized instead of maximizing the composite-likelihood function.

Conclusion

In this dissertation, Jaatha is introduced, which is a robust and flexible method to estimate demographic parameters for a given model of two recently diverged species‘ SNP data. It is a composite-likelihood method that can cope with high recombination rates and finite sites data.

In performance it is in most cases comparable to other currently available methods, but outperforms them when divergence is recent. Additionally, in respect to run times it is much faster than comparable methods. In an extensive simulation study we analyzed the effects on demographic parameters when the assumption of the infinite-sites sequence evolution model is not met. As a result, already for moderate values of the population mutation rate θ, finite sites models should be applied for many biological data sets. Hence Jaatha was modified to include finite sites models. Parameters of finite sites models such as the mutation rate heterogeneity can be accurately estimated, in particular if the transition transversion ratio is low. For the discrimination between competing models likelihood ratio testing has been proven useful. All in all, we conclude that composite-likelihood methods provide a reasonable alternative to full-data methods, especially when many loci are available.

Our motivational data set came from the South American wild tomatoes, *Solanum chilense* and *S. peruvianum* to which Jaatha was applied. We find significant evidence for gene flow between *S. chilense* and *S. peruvianum* and population expansion in *S. peruvianum*. The divergence time of both species (≈ 2 million years ago) follows the sudden uplift of the Andes and a subsequent cooling of the temperature which has previously been shown to have created a unique habitat in the Central Andes.

In its current version Jaatha sheds light on the demographic processes that underly speciation events. Since Jaatha is a reliable method, it sets the stage for a wide range of applications and further extensions, including the detection of selection. Thus in the near future Jaatha might help to identify genomic regions that play a role in speciation and adaptation to new environments.

Bibliography

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

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

Andersen, E. C., Gerke, J. P., Shapiro, J. A., Crissman, J. R., Ghosh, R., Bloom, J. S., Félix, M.-A., and Kruglyak, L. (2012). Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nat Genet*, 44(3):285–290.

Andolfatto, P. and Przeworski, M. (2000). A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics*, 156(1):257–268.

Andrew, R. L., Ostevik, K. L., Ebert, D. P., and Rieseberg, L. H. (2012). Adaptation with gene flow across the landscape in a dune sunflower. *Mol Ecol*, 21(9):2078–2091.

Aris-Brosou, S. and Excoffier, L. (1996). The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Mol Biol Evol*, 13(3):494–504.

Arnold, M. L. (2004). Transfer and origin of adaptations through natural hybridization: were Anderson and Stebbins right? *Plant Cell*, 16(3):562–570.

Arnold, M. L. and Bennett, B. D. (1993). *Natural hybridization in Louisiana irises: genetic variation and ecological determinants*. Hybrid zones and the evolutionay process. Oxford University Press, Oxford.

Arunyawat, U., Stephan, W., and Städler, T. (2007). Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Mol Biol Evol*, 24(10):2310–2322.

Avery, O. T., Macleod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *J Exp Med*, 79(2):137–158.

Bahlo, M. and Griffiths, R. (2000). Inference from gene trees in a subdivided population. *Theor Popul Biol*, 57(2):79–95.

Barbieri, C., Whitten, M., Beyer, K., Schreiber, H., Li, M., and Pakendorf, B. (2012). Contrasting maternal and paternal histories in the linguistic context of Burkina Faso. *Mol Biol Evol*, 29(4):1213–1223.

Bazin, E., Dawson, K. J., and Beaumont, M. A. (2010). Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics*, 185(2):587–602.

Beaumont, M. A. and Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol*, 13(4):969–980.

Beaumont, M. A., Nielsen, R., Robert, C., Hey, J., Gaggiotti, O., Knowles, L., Estoup, A., Panchal, M., Corander, J., Hickerson, M., Sisson, S., Fagundes, N., Chikhi, L., Beerli, P., Vitalis, R., Cornuet, J.-M., Huelsenbeck, J., Foll, M., Yang, Z., Rousset, F., Balding, D., and Excoffier, L. (2010). In defence of model-based inference in phylogeography. *Mol Ecol*, 19(3):436–446.

Beaumont, M. A. and Rannala, B. (2004). The Bayesian revolution in genetics. *Nat Rev Genet*, 5(4):251–261.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Becquet, C. and Przeworski, M. (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Res*, 17(10):1505–1519.

Becquet, C. and Przeworski, M. (2009). Learning about modes of speciation by computational approaches. *Evolution*, 63(10):2547–2562.

Beerli, P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, 22(3):341–345.

Beerli, P. and Felsenstein, J. (1999). Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, 152(2):763–773.

Beerli, P. and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc Natl Acad Sci USA*, 98(8):4563–4568.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., and Langley, C. H. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol*, 5(11):e310.

Beisswanger, S., Stephan, W., and Lorenzo, D. D. (2006). Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. *Genetics*, 172(1):265–274.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J., and Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*, 18(6):630–634.

Bull, V., Beltrán, M., Jiggins, C. D., McMillan, W. O., Bermingham, E., and Mallet, J. (2006). Polyphyly and gene flow between non-sibling *Heliconius* species. *BMC Biol*, 4:11.

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

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet*, 43(10):956–963.

Castric, V., Bechsgaard, J., Schierup, M. H., and Vekemans, X. (2008). Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genet*, 4(8):e1000168.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*, 10(3):195–205.

Chen, H. (2012). The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theor Popul Biol*, 81(2):179–195.

Chen, H., Green, R., Pääbo, S., and Slatkin, M. (2007). The joint allele-frequency spectrum in closely related species. *Genetics*, 177(1):387–398.

Chetelat, R. T., Pertuzé, R. A., Faúndez, L., Graham, E. B., and Jones, C. M. (2009). Distribution, ecology and reproductive biology of wild tomatoes and related nightshades from the Atacama Desert region of northern Chile. *Euphytica*, 167(1):77–93.

Choi, S. C. and Hey, J. (2011). Joint inference of population assignment and demographic history. *Genetics*, 189(2):561–577.

Clotault, J., Thuillet, A.-C., Buiron, M., Mita, S. D., Couderc, M., Haussmann, B. I. G., Mariac, C., and Vigouroux, Y. (2012). Evolutionary history of pearl millet (*Pennisetum glaucum* [l.] r. br.) and selection on flowering genes since its domestication. *Mol Biol Evol*, 29(4):1199–1212.

Cornuet, J. M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J. M., Balding, D. J., Guillemaud, T., and Estoup, A. (2008). Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, 24(23):2713–2719.

Cristescu, M. E., Constantin, A., Bock, D. G., Cáceres, C. E., and Crease, T. J. (2012). Speciation with gene flow and the genetics of habitat transitions. *Mol Ecol*, 21(6):1411–1422.

Cutter, A. D., Wang, G.-X., Ai, H., and Peng, Y. (2012). Influence of finite-sites mutation, population subdivision and sampling schemes on patterns of nucleotide polymorphism for species with molecular hyperdiversity. *Mol Ecol*, 21(6):1345–1359.

Desai, M. M. and Plotkin, J. B. (2008). The polymorphism frequency spectrum of finitely many sites under selection. *Genetics*, 180(4):2175–2191.

Didelot, X., Everitt, R., Johansen, A., and Lawson, D. (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6:49–76.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*, 22(5):1185–1192.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. Springer, New York, 2nd edition.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall, Boca Raton, Florida.

Elmer, K. R., Reggio, C., Wirth, T., Verheyen, E., Salzburger, W., and Meyer, A. (2009). Pleistocene desiccation in East Africa bottlenecked but did not extirpate the adaptive radiation of Lake Victoria haplochromine cichlid fishes. *Proc Natl Acad Sci USA*, 106(32):13404–13409.

Evans, D. M. and Cardon, L. R. (2004). Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *Am J Hum Genet*, 75(4):687–692.

Ewing, G. and Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065.

Excoffier, L. and Heckel, G. (2006). Computer programs for population genetics data analysis: a survival guide. *Nat Rev Genet*, 7(10):745–758.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet*, 22:521–565.

Fernández, J., Villanueva, B., Pong-Wong, R., and Toro, M. A. (2005). Efficiency of the use of pedigree and molecular marker information in conservation programs. *Genetics*, 170(3):1313–1321.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon, Oxford.

Fitzpatrick, B. M., Fordyce, J. A., and Gavrilets, S. (2009). Pattern, process and geographic modes of speciation. *J Evol Biol*, 22(11):2342–2347.

François, O., Blum, M. G. B., Jakobsson, M., and Rosenberg, N. A. (2008). Demographic history of european populations of *Arabidopsis thaliana*. *PLoS Genet*, 4(5):e1000075.

Fu, Y. X. (1995). Statistical properties of segregating sites. *Theoretical Population Biology*, 48(2):172 – 197.

Fu, Y. X. and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709.

Garrigan, D. (2009). Composite likelihood estimation of demographic parameters. *BMC Genet*, 10:72–84.

Gattepaille, L. M. and Jakobsson, M. (2012). Combining markers into haplotypes can improve population structure inference. *Genetics*, 190(1):159–174.

Goldstein, D. and Schlötterer, C. (1999). *Microsatellites: Evolution and Applications.* Oxford University Press, New York.

Graham, A. (2009). The Andes: A geological overview from a biological perspective. *Annals of the Missouri Botanical Garden*, 96(3):371–385.

Griffiths, R. and Tavaré, S. (1995). Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math Biosci*, 127(1):77–98.

Griffiths, R. C. and Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci*, 344(1310):403–410.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*, 5(10):e1000695.

Haldane, J. B. S. (1932). *The causes of evolution*. Longmans, Green & Co., London.

Hamilton, G., Currat, M., Ray, N., Heckel, G., Beaumont, M., and Excoffier, L. (2005). Bayesian estimation of recent migration rates after a spatial expansion. *Genetics*, 170(1):409–417.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, 24(23):2786–2787.

Hey, J. (2006). Recent advances in assessing gene flow between diverging populations and species. *Current Opinion in Genetics & Development*, 16(6):592–596.

Hey, J. (2010). Isolation with migration models for more than two populations. *Mol Biol Evol*, 27(4):905–920.

Hey, J. and 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(2):747–760.

Hey, J. and Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci USA*, 104(8):2785–2790.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, 159(4):1805–1817.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338.

Hughes, C. and Eastwood, R. (2006). Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the Andes. *Proc Natl Acad Sci USA*, 103(27):10334–10339.

Jeffreys, A. J. and May, C. A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet*, 36(2):151–156.

Jenks, W. (1975). *Peru*. The encyclopaedia world of regional geology. Dowden, Hutchinson & Ross, Stroudsburg, Pennsylvania.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., Brady, S. D., Zhang, H., Pollen, A. A., Howes, T., Amemiya, C., Team, B. I. G. S. P. . W.

G. A., Baldwin, J., Bloom, T., Jaffe, D. B., Nicol, R., Wilkinson, J., Lander, E. S., Palma, F. D., Lindblad-Toh, K., and Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, 484(7392):55–61.

Jukes, T. H. and Cantor, C. R. (1969). *Evolution of Protein Molecules*. Mammalian protein metabolism, III. Academic Press, New York.

Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellåker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., Oliver, P. L., McIntyre, R. E., Bhomra, A., Nicod, J., Gan, X., Yuan, W., van der Weyden, L., Steward, C. A., Bala, S., Stalker, J., Mott, R., Durbin, R., Jackson, I. J., Czechanski, A., Guerra-Assunção, J. A., Donahue, L. R., Reinholdt, L. G., Payseur, B. A., Ponting, C. P., Birney, E., Flint, J., and Adams, D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294.

Keinan, A. and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743.

Kennedy, G. C., Matsuzaki, H., Dong, S., min Liu, W., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., Liu, W., Yang, G., Di, X., Ryder, T., He, Z., Surti, U., Phillips, M. S., Boyce-Jacino, M. T., Fodor, S. P. A., and Jones, K. W. (2003). Large-scale genotyping of complex DNA. *Nat Biotechnol*, 21(10):1233–1237.

Kim, Y. and Stephan, W. (2000). Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics*, 155(3):1415–1427.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120.

Kingman, J. (1982). The coalescent. *Stochastic Process. Appl.*, 13(3):235 – 248.

Kliman, R. M. and Hey, J. (1993). Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol Biol Evol*, 10(6):1239–1258.

Knowles, L. L. and Maddison, W. P. (2002). Statistical phylogeography. *Mol Ecol*, 11(12):2623–2635.

Krämer, N., Boulesteix, A.-L., and Tutz, G. (2008). Penalized partial least squares with applications to B-Spline transformations and functional data. *Chemometrics and Intelligent Laboratory Systems*, 94:60 – 69.

Kuhner, M. K. (2006). LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, 22:768–770.

LAMARC Team (29 May 2012). LAMARC - Likelihood Analysis with Metropolis Algorithm using Random Coalescence. http://evolution.genetics.washington.edu/lamarc/index.html. accessed: 4 June 2012.

Lascoux, M. and Petit, R. J. (2010). The 'New Wave' in plant demographic inference: more loci and more individuals. *Mol Ecol*, 19(6):1075–1078.

Leman, S., Chen, Y., Stajich, J., Noor, M., and Uyenoyama, M. (2005). Likelihoods from summary statistics: recent divergence between species. *Genetics*, 171(3):1419–1436.

Leuenberger, C. and Wegmann, D. (2010). Bayesian computation and model selection without likelihoods. *Genetics*, 184(1):243–252.

Li, H. and Stephan, W. (2006). Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet*, 2(10):e166.

Lin, K., Li, H., Schlötterer, C., and Futschik, A. (2011). Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. *Genetics*, 187(1):229–244.

Lin, L.-H., Qu, Y.-F., Li, H., Zhou, K.-Y., and Ji, X. (2012). Genetic structure and demographic history should inform conservation: Chinese cobras currently treated as homogenous show population divergence. *PLoS ONE*, 7(4):e36334.

Lopes, J. S., Balding, D., and Beaumont, M. A. (2009). PopABC: a program to infer historical demographic parameters. *Bioinformatics*, 25(20):2747–2749.

Lukjancenko, O., Wassenaar, T. M., and Ussery, D. W. (2010). Comparison of 61 sequenced *Escherichia coli* genomes. *Microb Ecol*, 60(4):708–720.

Lundstrom, R., Tavaré, S., and Ward, R. H. (1992). Modeling the evolution of the human mitochondrial genome. *Math Biosci*, 112(2):319–335.

Lynch, M. (2009). Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, 182(1):295–301.

Mailund, T., Dutheil, J. Y., Hobolth, A., Lunter, G., and Schierup, M. H. (2011). Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden markov model. *PLoS Genet*, 7(3):e1001319.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

Marie Curie SPECIATION Network, Butlin, R., Debelle, A., Kerth, C., Snook, R. R., Beukeboom, L. W., Cajas, R. F. C., Diao, W., Maan, M. E., Paolucci, S., Weissing, F. J., van de Zande, L., Hoikkala, A., Geuverink, E., Jennings, J., Kankare, M., Knott, K. E., Tyukmaeva, V. I., Zoumadakis, C., Ritchie, M. G., Barker, D., Immonen, E., Kirkpatrick, M., Noor, M., Garcia, C. M., Schmitt, T., and Schilthuizen, M. (2012). What do we need to know about speciation? *Trends Ecol Evol*, 27(1):27–39.

Martincorena, I., Seshasayee, A. S. N., and Luscombe, N. M. (2012). Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature*, 485(7396):95–98.

Mayr, E. (1942). *Systematics and the Origin of Species*. Columbia University Press, New York.

McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160(3):1231–1241.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Naduvilezhath, L., Rose, L. E., and Metzler, D. (2011). Jaatha: a fast composite-likelihood approach to estimate demographic parameters. *Mol Ecol*, 20(13):2709–2723.

Nakazato, T. and Housworth, E. A. (2011). Spatial genetics of wild tomato species reveals roles of the Andean geography on demographic history. *Am. J. Bot.*, 98(1):88–98.

Nakazato, T., Warren, D. L., and Moyle, L. C. (2010). Ecological and geographic modes of species divergence in wild tomatoes. *Am J Bot*, 97(4):680–693.

Nath, H. B. and Griffiths, R. C. (1996). Estimation in an island model using simulation. *Theor Popul Biol*, 50(3):227–253.

Navarro-González, R., Rainey, F. A., Molina, P., Bagaley, D. R., Hollen, B. J., de la Rosa, J., Small, A. M., Quinn, R. C., Grunthaner, F. J., Cáceres, L., Gomez-Silva, B., and McKay, C. P. (2003). Mars-like soils in the Atacama desert, Chile, and the dry limit of microbial life. *Science*, 302(5647):1018–1021.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 89(3):583–590.

Orr, H. A. and Turelli, M. (2001). The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. *Evolution*, 55(6):1085–1094.

Parsch, J., Meiklejohn, C. D., and Hartl, D. L. (2001). Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. *Genetics*, 159(2):647–657.

Posada, D. and Crandall, K. A. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics*, 14(9):817–818.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol*, 16(12):1791–1798.

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.

Rambaut, A. and Grassly, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci*, 13(3):235–238.

Rhymer, J. M. and Simberloff, D. (1996). Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, 27:83–109.

Rick, C. M. and Chetelat, R. T. (1995). Utilization of related wild species for tomato improvement. In *Acta Horticulturae*, number 412, pages 21–38. International Society Horticultural Science.

Rick, C. M., Fobes, J. F., and Tanksley, S. D. (1979). Evolution of mating systems in *Lycopersicon hirsutum* as deduced from genetic variation in electrophoretic and morphological characters. *Plant Systematics and Evolution*, 132:279–298. 10.1007/BF00982390.

Robert, C. P., Cornuet, J.-M., Marin, J.-M., and Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proc Natl Acad Sci USA*, 108(37):15112–15117.

Robertson, A. (1975). Remarks on the Lewontin-Krakauer test . *Genetics*, 80(2):396.

Rogers, A. R. and Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol*, 9(3):552–569.

Rose, L. E., Grzeskowiak, L., Hörger, A. C., Groth, M., and Stephan, W. (2011). Targets of selection in a disease resistance network in wild tomatoes. *Mol Plant Pathol*, 12(9):921–927.

Roselius, K., Stephan, W., and Städler, T. (2005). The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics*, 171(2):753–763.

Roychoudhury, A. and Wakeley, J. (2010). Sufficiency of the number of segregating sites in the limit under finite-sites mutation. *Theor Popul Biol*, 78(2):118–122.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74(12):5463–5467.

Schneider, S. and Excoffier, L. (1999). Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics*, 152(3):1079–1089.

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, 6:461–464.

Shoemaker, J. S., Painter, I. S., and Weir, B. S. (1999). Bayesian statistics in genetics: a guide for the uninitiated. *Trends in Genetics*, 15(9):354 – 358.

Siol, M., Wright, S. I., and Barrett, S. C. H. (2010). The population genomics of plant adaptation. *New Phytol*, 188(2):313–332.

Smadja, C. M. and Butlin, R. K. (2011). A framework for comparing processes of speciation in the presence of gene flow. *Mol Ecol*, 20(24):5123–5140.

Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet Res*, 23(1):23–35.

Smith, J. M. B. and Cleef, A. M. (1988). Composition and origins of the world's tropicalpine floras. *J of Biogeography*, 15(4):631–645.

Soltis, P. S., Soltis, D. E., Savolainen, V., Crane, P. R., and Barraclough, T. G. (2002). Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. *Proc Natl Acad Sci USA*, 99(7):4430–4435.

Sousa, V. C., Beaumont, M. A., Fernandes, P., Coelho, M. M., and Chikhi, L. (2012). Population divergence with or without admixture: selecting models using an ABC approach. *Heredity*, 108(5):521–530.

Sousa, V. C., Grelaud, A., and Hey, J. (2011). On the nonidentifiability of migration time estimates in isolation with migration models. *Mol Ecol*, 20(19):3956–3962.

Städler, T., Arunyawat, U., and Stephan, W. (2008). Population genetics of speciation in two closely related wild tomatoes (*Solanum* section *Lycopersicon*). *Genetics*, 178(1):339–350.

Städler, T., Florez-Rueda, A. M., and Paris, M. (2012). Testing for "snowballing" hybrid incompatibilities in *Solanum*: impact of ancestral polymorphism and divergence estimates. *Mol Biol Evol*, 29(1):31–34.

Städler, T., Haubold, B., Merino, C., Stephan, W., and Pfaffelhuber, P. (2009). The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics*, 182(1):205–216.

Städler, T., Roselius, K., and Stephan, W. (2005). Genealogical footprints of speciation processes in wild tomatoes: demography and evidence for historical gene flow. *Evolution*, 59(6):1268–1279.

Storz, J. F. and Beaumont, M. A. (2002). Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution*, 56(1):154–166.

Strasburg, J. L. and Rieseberg, L. H. (2008). Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris*–large effective population sizes and rates of long-term gene flow. *Evolution*, 62(8):1936–1950.

Strasburg, J. L. and Rieseberg, L. H. (2010). How robust are "Isolation with Migration" analyses to violations of the IM model? A simulation study. *Mol Biol Evol*, 27(2):297–310.

Strasburg, J. L. and Rieseberg, L. H. (2011). Interpreting the estimated timing of migration events between hybridizing species. *Mol Ecol*, 20(11):2353–2366.

Strope, C. L., Abel, K., Scott, S. D., and Moriyama, E. N. (2009). Biological sequence simulation for testing complex evolutionary hypotheses: indel-seq-gen version 2.0. *Mol Biol Evol*, 26(11):2581–2593.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595.

Tavaré, S. (1986). *Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences*, volume 17, pages 57–86. Amer Mathematical Society.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518.

Tellier, A., Pfaffelhuber, P., Haubold, B., Naduvilezhath, L., Rose, L. E., Städler, T., Stephan, W., and Metzler, D. (2011). Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum. *PLoS One*, 6(5):e18155.

Templeton, A. (2004). *A maximum likelihood framework for cross validation of phylogeographic hypotheses*. Evolutionary Theory and Processes: Modern Horizon. Kluwer Academic Publishers, Dordrecht. pp. 209–230.

Templeton, A. R. (2009). Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation. *Mol Ecol*, 18(2):319–331.

Templeton, A. R. (2010a). Correcting approximate Bayesian computation. *Trends Ecol Evol*, 25(9):488–489; author reply 490–491.

Templeton, A. R. (2010b). Reply to Berger *et al.*: Improving ABC. *Proc Natl Acad Sci USA*, 107(41):E158.

Teshima, K., Coop, G., and Przeworski, M. (2006). How reliable are empirical genomic scans for selective sweeps? *Genome Res*, 16(6):702–712.

The Heliconius Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, advanced online publication.

The Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485:635–641.

Toni, T. and Stumpf, M. P. H. (2010). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface*, 6(31):187–202.

Vonlanthen, P., Bittner, D., Hudson, A. G., Young, K. A., Müller, R., Lundsgaard-Hansen, B., Roy, D., Piazza, S. D., Largiader, C. R., and Seehausen, O. (2012). Eutrophication causes speciation reversal in whitefish adaptive radiations. *Nature*, 482(7385):357–362.

Wakeley, J. (2008). *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, Colorado.

Wakeley, J. and Hey, J. (1997). Estimating ancestral population parameters. *Genetics*, 145(3):847–855.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7(2):256–276.

Wegmann, D., Leuenberger, C., and Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182(4):1207–1218.

Weigel, D. and Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol*, 10(5):107.

Weiss, G. and von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics*, 149(3):1539–1546.

Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C. D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci USA*, 102(22):7882–7887.

Wilson, I. J., Weale, M. E., and Balding, D. J. (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Stat. Soc.: Series A (Statistics in Society)*, 166(2):155–188.

Wiuf, C. (2006). Consistency of estimators of population scaled parameters using composite likelihood. *J Math Biol*, 53(5):821–841.

Won, Y. and Hey, J. (2005). Divergence population genetics of chimpanzees. *Mol Biol Evol*, 22(2):297–307.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16(2):97–159.

Wright, S. (1943). Isolation by distance. *Genetics*, 28(2):114–138.

Xia, H., Camus-Kulandaivelu, L., Stephan, W., Tellier, A., and Zhang, Z. (2010). Nucleotide diversity patterns of local adaptation at drought-related candidate genes in wild tomatoes. *Mol Ecol*, 19(19):4144–4154.

Xing, J., Watkins, W. S., Hu, Y., Huff, C. D., Sabo, A., Muzny, D. M., Bamshad, M. J., Gibbs, R. A., Jorde, L. B., and Yu, F. (2010). Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biol*, 11(11):R113.

Yang (1996). Maximum-likelihood models for combined analyses of multiple sequence data. *J Mol Evol*, 42(5):587–596.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565–569.

Zimmermann, O. and Hansmann, U. H. E. (2008). Understanding protein folding: small proteins in silico. *Biochim Biophys Acta*, 1784(1):252–258.

Appendix A

Online Supplementary Material for Naduvilezhath *et al.*, 2011

A.1 Parameter ranges and command lines

The parameters written in italics are drawn on a logarithmic scale from different priors except the parameters θ and ρ which are uniformly drawn.
divergence time τ: $[20 \cdot (\frac{5}{6})^{39}, \ldots, 20 \cdot (\frac{5}{6})^{0}]$ with resulting range: $[0.016, \ldots, 20]$
migration rate m: $[5 \cdot (\frac{5}{6})^{39}, \ldots, 5 \cdot (\frac{5}{6})^{0}]$ with resulting range: $[0.004, \ldots, 5]$
size ratio q: $[10 \cdot (\frac{87}{100})^{39}, \ldots, 10 \cdot (\frac{87}{100})^{0}]$ with resulting range: $[0.044, \ldots, 10]$
population-scaled mutation rate θ *(per locus)*: $[5, \ldots, 20]$ (0.005-0.02 per bp)
recombination rate ρ *(per locus)*: $[5, \ldots, 20]$ (0.005-0.02 per bp)
In the training phase the number of loci is always seven, $\theta = 5$, and the migration rate is symmetric.
Each demographic model has been simulated under three scenarios:
`7-Loci scenario` 7 loci, asymmetric migration rate, and recombination
`100-Loci scenario` 100 loci, symmetric migration rate, and **no** recombination
`1000-Loci scenario` 1000 loci, symmetric migration rate, and recombination

Constant Model: `ms 50` *numLoci* `-t` θ `-r` ρ `1000 -I 2 25 25 -m 1 2` *m12* `-m 2 1` *m21* `-n 2` q `-eN` τ `1+`q `-ej` τ `2 1`
Growth Model: `ms 50` *numLoci* `-t` θ `-r` ρ `1000 -I 2 25 25 -m 1 2` *m12* `-m 2 1` *m21* `-n 2` q `-eN` τ `2 -ej` τ `2 1 -g 2 log(`q`)/`τ
Fraction-Growth Model: `ms 50` *numLoci* `-t` θ `-r` ρ `1000 -I 2 25 25 -m`

`1 2` *m12* `-m 2 1` *m21* `-n 2` *q* `-eN` τ `1.05 -ej` τ `2 1 -g 2 log(`*q*`/0.05)/`τ

A.2 Supplemental Figures and Tables

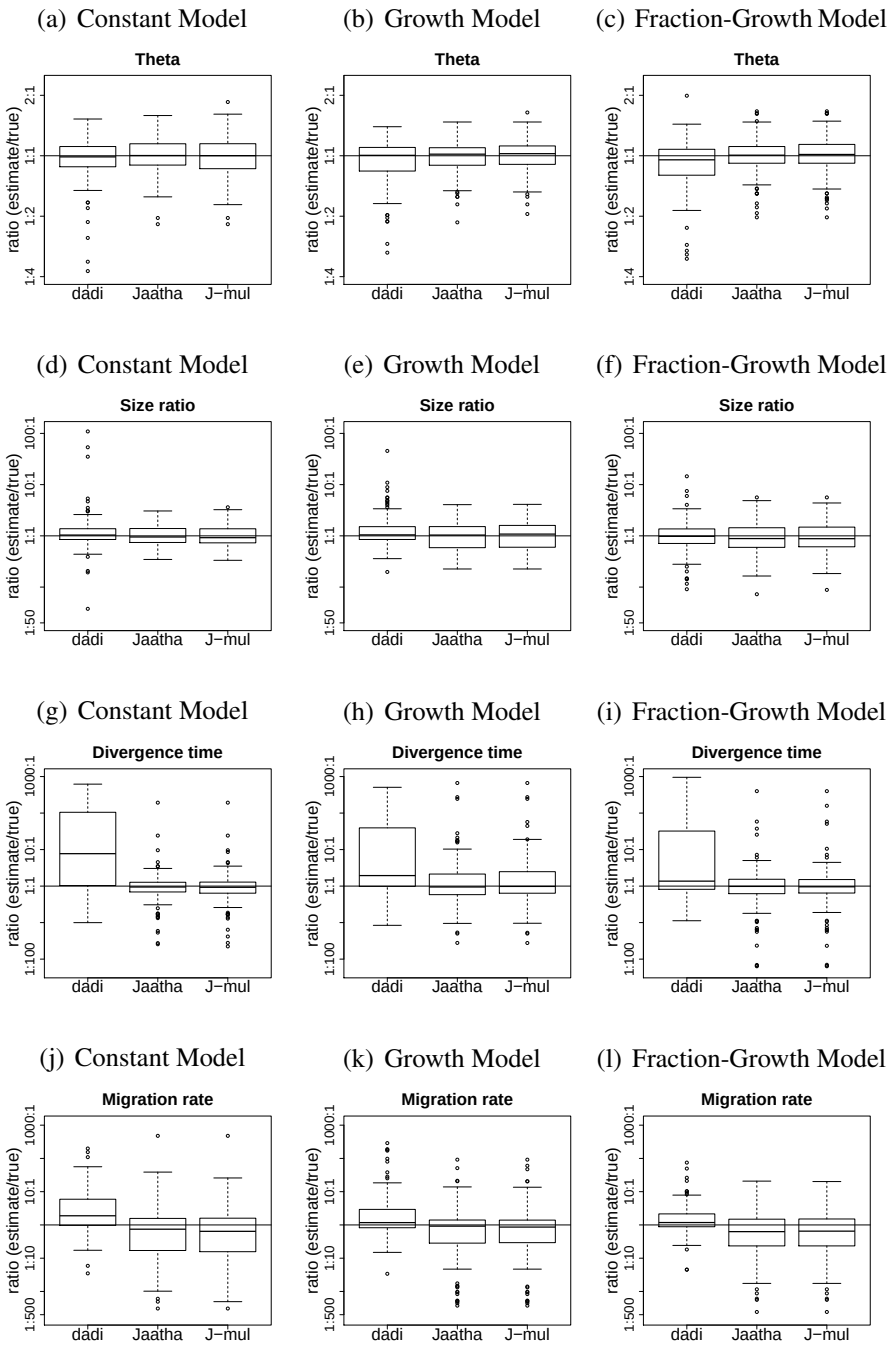


Figure A.1: Ratio of estimated to true values by $\partial a \partial i$, Jaatha, and Jaatha with the (composite) likelihood estimation based on a multinomial model (J-mul) of four parameters across models and methods for `7-loci scenario`.

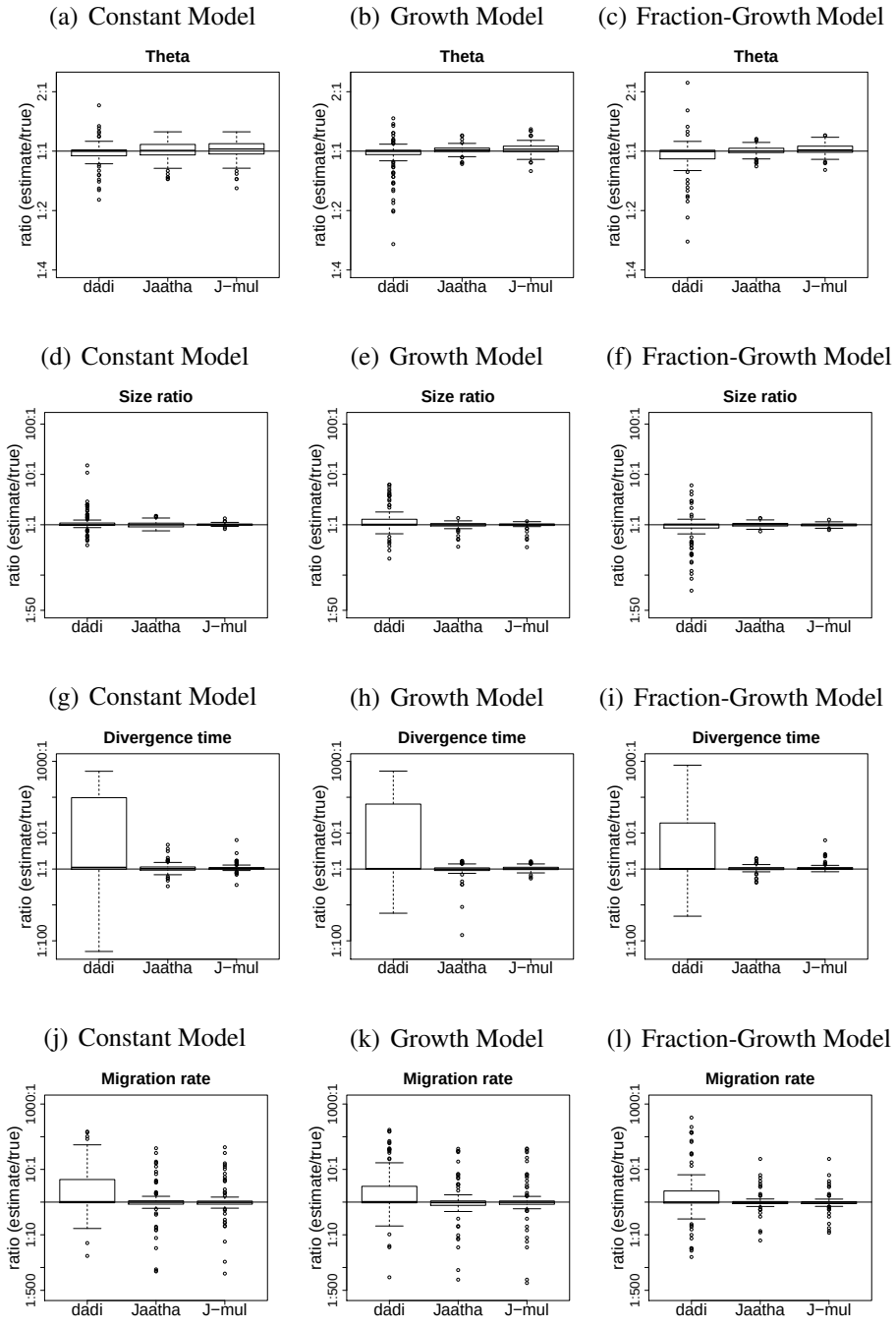


Figure A.2: Ratio of estimated to true values by $\partial a \partial i$, Jaatha, and Jaatha with the (composite) likelihood estimation based on a multinomial model (J-mul) of four parameters across models and methods for `1000-loci scenario`.

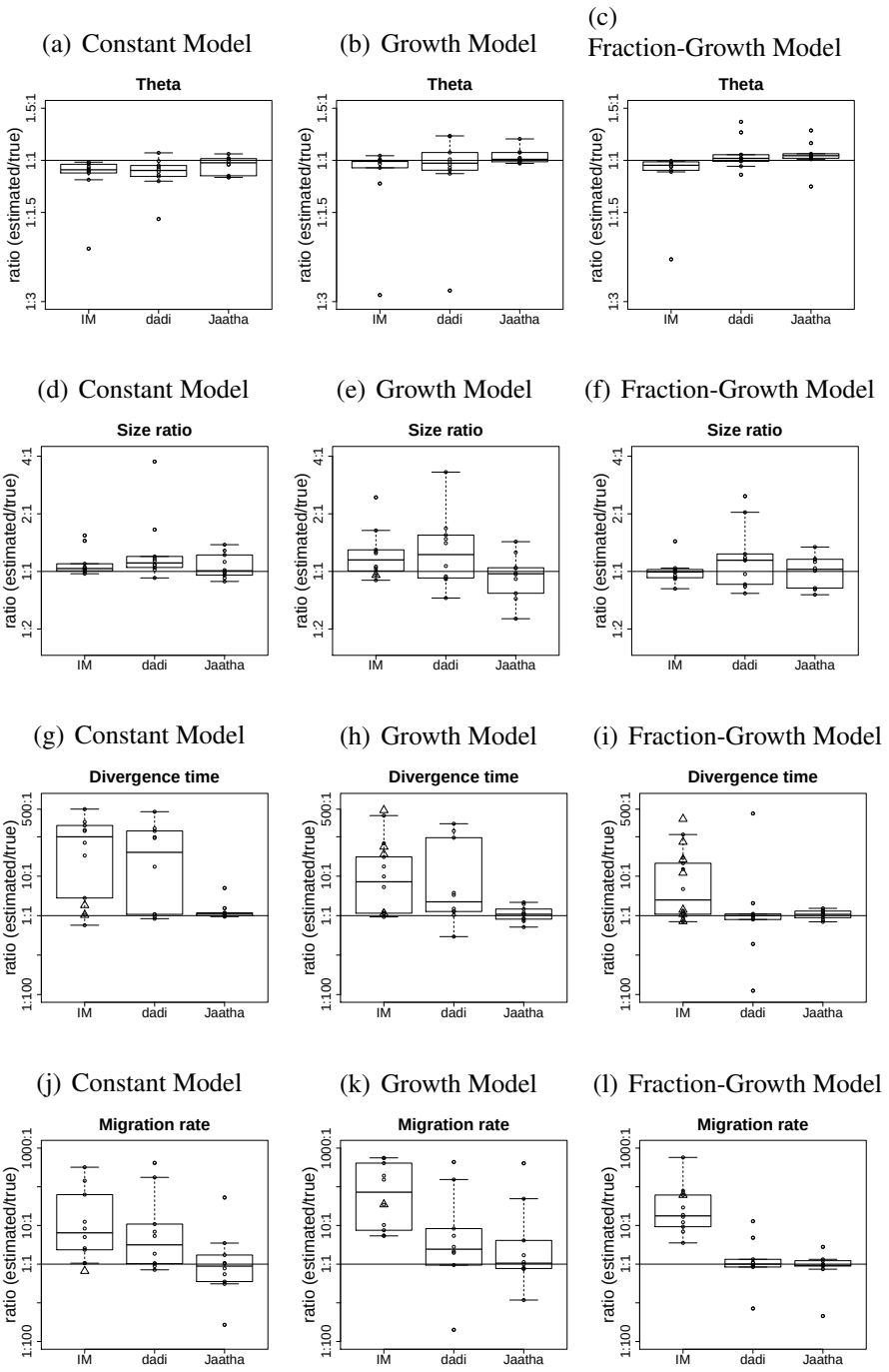


Figure A.3: Ratio of estimated to true values of four parameters across models and methods for `100-loci scenario` (no recombination). IM results with ESS $<$100 are not included in the boxplots but drawn in additionally ($\triangle$). Results for $\partial a \partial i$ and Jaatha with the same 10 simulated datasets for *Constant*, *Growth*, and *Fraction-Growth Model* are shown.

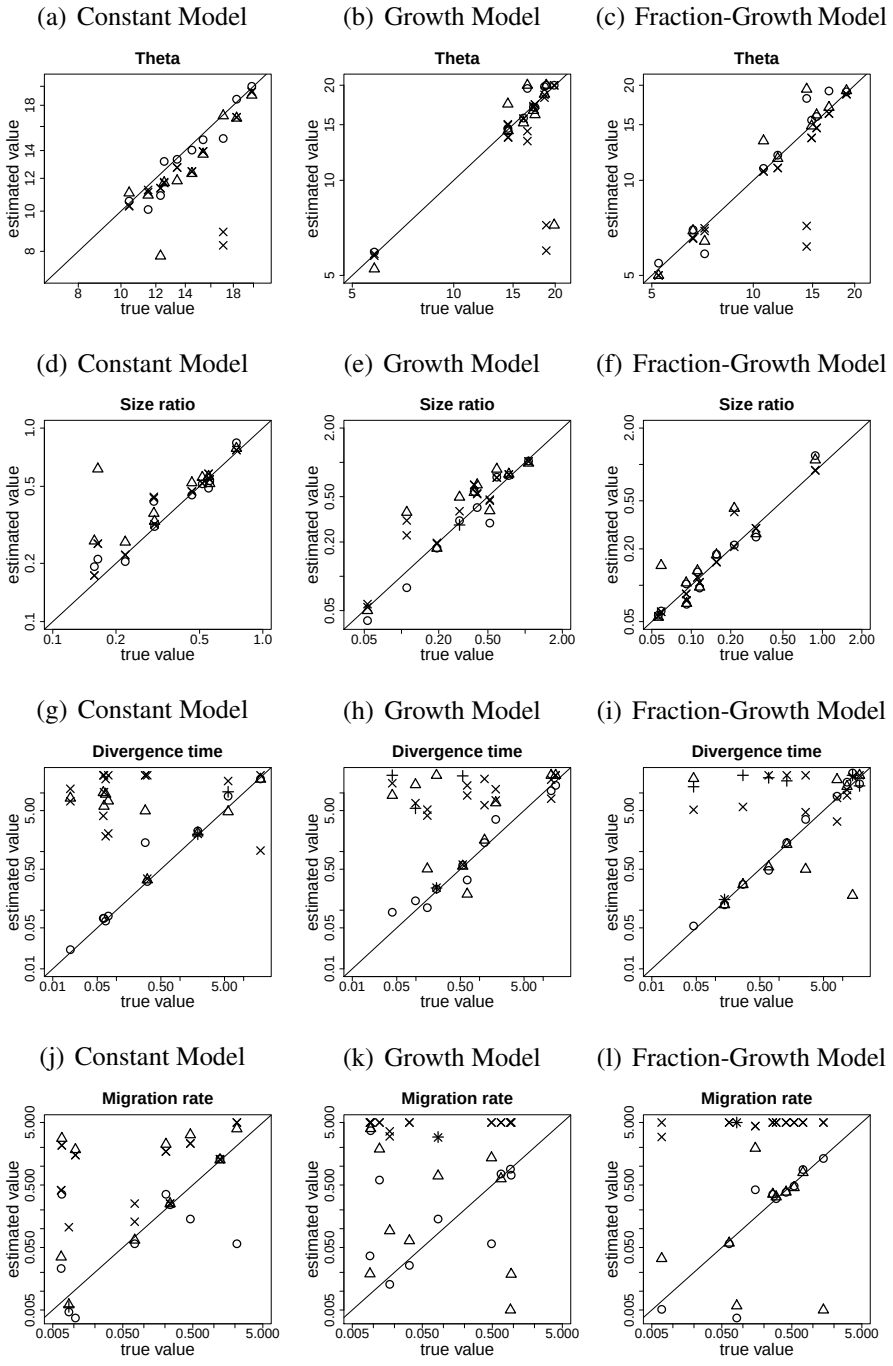


Figure A.4: Estimations of the four parameters using the three methods: Jaatha (○), $\partial a \partial i$ (△), and IM (x for ESS >100; + for ESS < 100 of that variable). These methods were applied to 10 simulated datasets each with 100 loci, without intralocus recombination. Shown are the estimations assuming three different underlying demographic models.

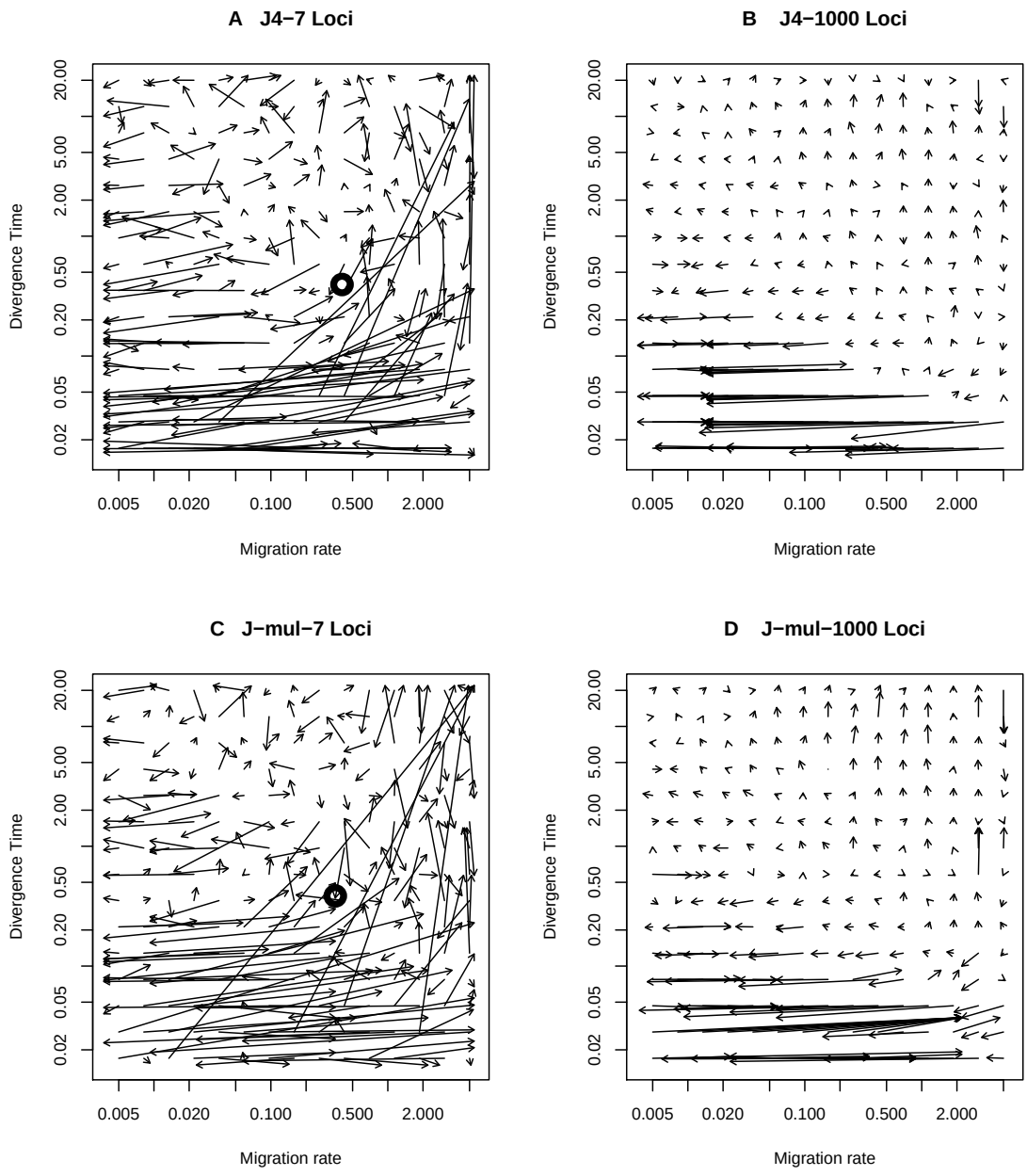


Figure A.5: Arrow plots of divergence time and migration for 7 and 1000 loci under the *Growth Model* with 45 samples per species and symmetric migration rates with J_4 (**A** and **B**, as in Tellier *et al.*) and Jaatha using a multinomial approximation (J-mul) for the composite likelihood (**C** and **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.

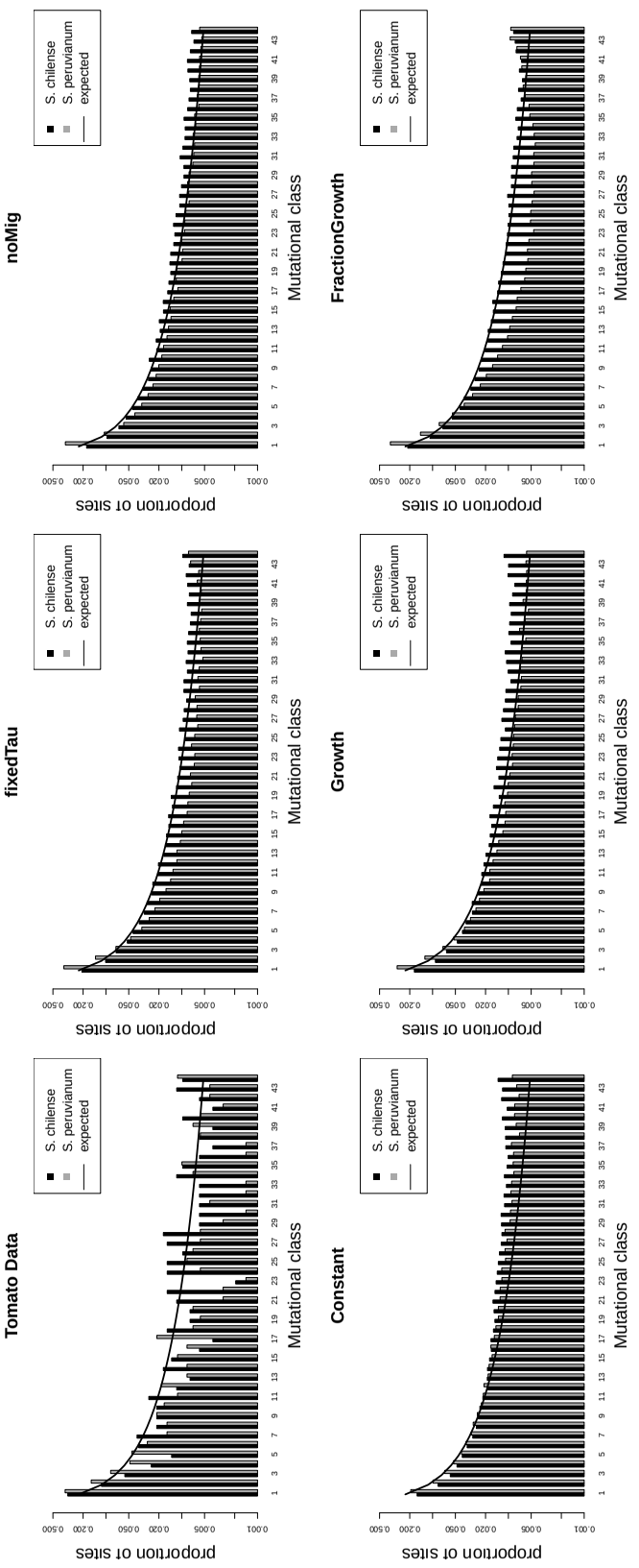


Figure A.6: The marginal site frequency spectra (SFS) for the tomato data and the average of 100 simulated data sets with each 7 loci for the tested five models *fixedTau*, *noMig*, *Constant*, *Growth*, and *FractionGrowth*. The line represents the expected SFS of the neutral Wright-Fisher Model of constant size without migration (Fu, 1995).

Table A.1: **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 $\theta_{site} = 0.01$.

locus	ρ per locus
CT066	4
CT093	8
CT166	13
CT179	21
CT198	18
CT251	15
CT268	35

Table A.2: **Estimates for parameters of models fitted to tomato data** - 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* 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-likelihoods (bottom rows) are calculated using the Poisson model and indicate that the *fixedTau Model* fits best while the *Constant Model* is the worst.

Parameter	Estimation Method	*Constant*	*Growth*	*Fraction-Growth*	*noMig*	*fixedTau*
$\widehat{\theta}_1$	J_1	9.41	10.26	12.56	13.34	13.74
		(6.98-12.63)	(8.37-13.14)	(9.51-16.33)	(10.60-17.17)	(11.50-19.48)
	J_2	9.33	10.19	12.56	13.34	13.34
		(6.99-12.94)	(8.39-12.74)	(9.58-16.03)	(10.70-17.31)	(11.05-18.21)
	J_4	10.08	9.68	12.56	13.46	13.08
		(7.87-13.86)	(7.83-12.38)	(9.68-16.14)	(10.22-17.65)	(10.79-17.23)
	multinomial	9.41	10.30	12.56	13.34	13.74
		(7.11-13.19)	(8.21-13.40)	(9.18-17.30)	(10.29-17.52)	(10.73-17.57)
$\widehat{q}$	J_1	1.88	3.77	3.77	7.57	3.77
		(1.18-2.94)	(2.10-6.23)	(2.34-5.65)	(5.16-13.57)	(2.30-5.88)
	J_2	1.88	3.93	3.77	7.57	4.11
		(1.18-2.88)	(2.47-6.20)	(2.49-5.54)	(5.63-12.30)	(2.71-6.81)
	J_4	1.66	4.55	4.26	8.66	4.64
		(1.08-2.40)	(2.80-7.30)	(2.72-6.16)	(5.44-14.77)	(2.90-7.13)
	multinomial	1.74	4.64	4.24	8.66	4.79

		(1.12-2.61)	(2.81-7.32)	(2.72-6.11)	(5.39-15.29)	(2.91-7.29)
$\widehat{m}$	J_1	0.23	0.56	0.56	*0*	0.56
		(0.03-3.56)	(0.12-5.24)	(0.27-1.11)		(0.23-1.54)
	J_2	0.34	0.42	0.56	*0*	0.56
		(0.09-2.89)	(0.11-2.91)	(0.29-1.06)		(0.24-1.49)
	J_4	0.56	0.41	0.75	*0*	0.56
		(0.14-4.05)	(0.14-1.81)	(0.40-1.27)		(0.23-1.07)
	multinomial	0.36	0.36	0.74	*0*	0.57
		(0.05-6.55)	(0.10-2.07)	(0.39-1.26)		(0.27-1.26)
$\widehat{\tau}$	J_1	0.36	0.36	0.90	0.15	*0.36*
		(0.08-1.06)	(0.08-1.07)	(0.35-2.11)	(0.11-0.19)	
	J_2	0.51	0.36	0.90	0.15	*0.36*
		(0.07-1.86)	(0.10-0.89)	(0.37-1.97)	(0.11-0.19)	
	J_4	0.49	0.40	0.78	0.14	*0.36*
		(0.04-2.78)	(0.12-0.97)	(0.38-1.60)	(0.10-0.23)	
	multinomial	0.40	0.39	0.77	0.14	*0.36*
		(0.07-1.51)	(0.11-0.96)	(0.37-1.57)	(0.09-0.22)	
$\widehat{s}$	J_1	-	-	-	0.47	0.23
					(0.04-0.41)	(0.18-1.02)
	J_2	-	-	-	0.47	0.25
					(0.05-0.44)	(0.18-0.96)
	J_4	-	-	-	0.43	0.27
					(0.07-0.73)	(0.17-0.98)

	multinomial	-	-	-	0.42 (0.17-0.97)	0.28 (0.09-0.99)
log-likelihood	J_1	-188.01	-123.45	-101.58	-133.06	-96.02
	J_2	-189.51	-119.70	-101.58	-133.06	-93.96
	J_4	-183.82	-118.15	-97.47	-132.66	-92.85
	multinomial	-186.23	-120.12	-97.54	-132.92	-95.62

Appendix B

Online Supplementary Material to Tellier *et al.*, 2011

Tellier et al.

ONLINE APPENDIX

Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum

Aurélien TELLIER, Peter PFAFFELHUBER, Bernhard HAUBOLD, Lisha NADUVILEZHATH, Laura E. ROSE, Thomas STÄDLER, Wolfgang STEPHAN, and Dirk METZLER

Content

Appendix 1: Four classes of the Joint-Frequency Spectrum for the Wakeley-Hey model.

The Joint site-Frequency Spectrum (JSFS) compares SNP data from n_1 samples from population 1 to n_2 samples from population 2. The JSFS is calculated as an array of dimension $(n_1+1) \times (n_2+1)$. A cell at row i and column j contains the number of polymorphic sites $S_{i,j}$ which are found i times in population 1 and j times in population 2. For example, $S_{2,3} = 10$ if 10 polymorphisms are found as doubletons in population 1 and tripletons in population 2. Four summary statistics are relevant for isolation-migration parameter inference: private polymorphisms in species 1 and 2, respectively (W_1, W_2), fixed differences between species (W_3), and shared ancestral polymorphisms (W_4) (Wakeley and Hey 1997).

$$W_1 = \sum_{1 \le i \le n_1 - 1} S_{i,0} + S_{i,n_2} \; ; \; W_2 = \sum_{1 \le j \le n_2 - 1} S_{0,j} + S_{n_1,j} \; ; \; W_3 = S_{0,n_2} + S_{n_1,0} \; ; \; W_4 = \sum_{1 \le i \le n_1 - 1} \sum_{1 \le j \le n_2 - 1} S_{i,j}$$

We represent the JSFS graphically, as well as the four different classes $W_{1\text{-}4}$ as follows:

<table>
<tr><td></td><td>0</td><td>1</td><td>2</td><td>3</td><td>…</td><td>n_2-3</td><td>n_2-2</td><td>n_2-1</td><td>n_2</td></tr>
<tr><td>0</td><td>X</td><td colspan="6">W_2</td><td>W_3</td></tr>
<tr><td>1</td><td rowspan="6">W_1</td><td colspan="6" rowspan="6">W_4</td><td rowspan="6">W_1</td></tr>
<tr><td>3</td></tr>
<tr><td>…</td></tr>
<tr><td>n_1-3</td></tr>
<tr><td>n_1-2</td></tr>
<tr><td>n_1-1</td></tr>
<tr><td>n_1</td><td>W_3</td><td colspan="6">W_2</td><td>X</td></tr>
</table>

Appendix 2: Summaries of the Joint-Frequency Spectrum for the Maximum-Likelihood method

We have tested four different sets of summary statistics derived from the JSFS. The four vectors of summary statistics are described below as D, D', D'', D^*.

Formally, the 7 values of vector *D* are written as:

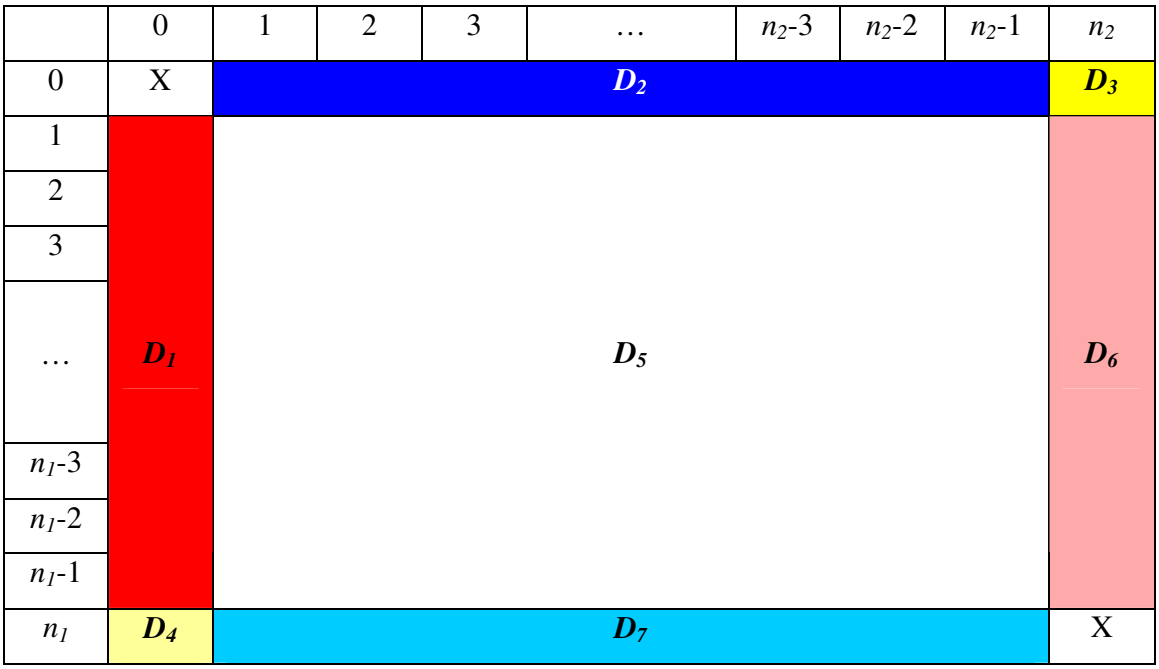
$$D_1 = \sum_{1 \le i \le n_1 - 1} S_{i,0}\ ;\ D_2 = \sum_{1 \le j \le n_2 - 1} S_{0,j}\ ;$$

$$D_3 = S_{0,n_2}\ ;\ D_4 = S_{n_1,0}\ ;$$

$$D_5 = \sum_{1 \le i \le n_1 - 1} \sum_{1 \le j \le n_2 - 1} S_{i,j}\ ;$$

$$D_6 = \sum_{1 \le i \le n_1 - 1} S_{i,n_2}\ ;\ D_7 = \sum_{1 \le j \le n_2 - 1} S_{n_1,j}$$

In relation to Eq. 2, $W_1=D_1+D_6$, $W_2=D_2+D_7$, $W_3=D_3+D_4$ and $W_4=D_5$. In other words, $W_1=D_1+D_6$ means that we separate polymorphic SNPs in population 1 which are not found in population 2 from those that are fixed in population 2 (i.e. polarizing the private polymorphism using information from an outgroup). As above, we represent graphically the JSFS and the classes of vector *D* as follows:



The second decomposition D_k' (k=1,..,12) is based on extracting the number of singletons from various classes of D.

$$D'_1 = S_{0,1}\,;\; D'_2 = \sum_{2\le i\le n_1-1} S_{i,0}\,;\; D'_3 = S_{1,0}\,;\; D'_4 = \sum_{2\le j\le n_2-1} S_{0,j}\,;$$

$$D'_5 = S_{0,n_2}\,;\; D'_6 = S_{n_1,0}\,;\; D'_7 = S_{1,1}\,;\; D'_8 = S_{1,n_2}\,;\; D'_9 = S_{n_1,1}$$

$$D'_{10} = \sum_{2\le i\le n_1-1}\sum_{2\le j\le n_2-1} S_{i,j}\,;\; D'_{11} = \sum_{2\le i\le n_1-1} S_{i,n_2}\,;\; D'_{12} = \sum_{2\le j\le n_2-1} S_{n_1,j}$$

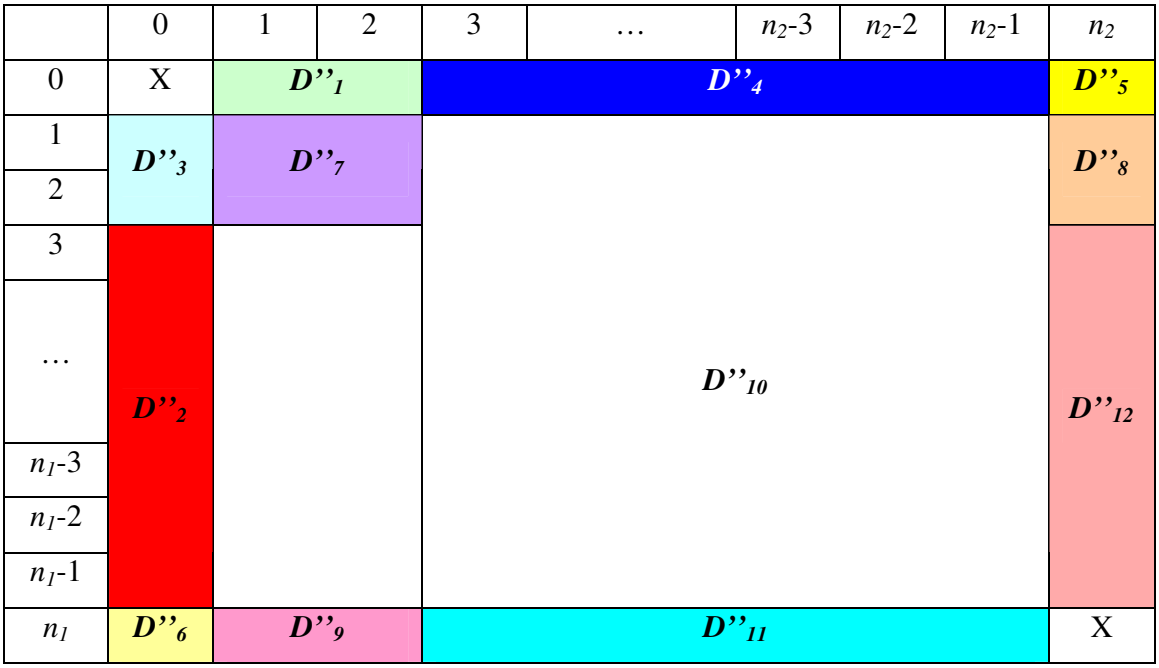
	0	1	2	3	…	n_2-3	n_2-2	n_2-1	n_2
0	X	***D'1***	***D'4***						***D'5***
1	***D'3***	***D'7***							***D'8***
2	***D'2***								***D'11***
3									
…			***D'10***						
n_1-3									
n_1-2									
n_1-1									
n_1	***D'6***	***D'9***	***D'12***						X

The third decomposition D_k''(k=1,..,12) contains low frequency polymorphism defined by singletons and doubletons from various classes of D.

$D_1'' = S_{0,1} + S_{0,2}$; $D_2'' = \sum_{3 \le i \le n_1 - 1} S_{i,0}$; $D_3'' = S_{1,0} + S_{2,0}$; $D_4'' = \sum_{3 \le j \le n_2 - 1} S_{0,j}$;

$D_5'' = S_{0,n_2}$; $D_6'' = S_{n_1,0}$; $D_7'' = S_{1,1} + S_{1,2} + S_{2,1} + S_{2,2}$; $D_8'' = S_{1,n_2} + S_{2,n_2}$; $D_9'' = S_{n_1,1} + S_{n_1,2}$

$D_{10}'' = \sum_{3 \le i \le n_1 - 1} \sum_{3 \le j \le n_2 - 1} S_{i,j}$; $D_{11}'' = \sum_{3 \le i \le n_1 - 1} S_{i,n_2}$; $D_{12}'' = \sum_{3 \le j \le n_2 - 1} S_{n_1,j}$



The fourth decomposition, the vector D_k^* (k=1,…,23), countains singletons and doubletons in separate classes.

$D_1^* = S_{0,1}$; $D_2^* = S_{0,2}$; $D_3^* = \sum_{3 \le j \le n_2 - 1} S_{0,j}$; $D_4^* = \sum_{3 \le i \le n_1 - 1} S_{i,0}$;

$D_5^* = S_{1,0}$; $D_6^* = S_{1,1}$; $D_7^* = S_{1,2}$; $D_8^* = S_{2,0}$; $D_9^* = S_{2,1}$; $D_{10}^* = S_{2,2}$;

$D_{11}^* = S_{0,n_2}$; $D_{12}^* = S_{1,n_2}$; $D_{13}^* = S_{2,n_2}$ $D_{14}^* = S_{n_1,0}$; $D_{15}^* = S_{n_1,1}$; $D_{16}^* = S_{n_1,2}$;

$D_{17}^* = \sum_{3 \le j \le n_2 - 1} S_{1,j}$; $D_{18}^* = \sum_{3 \le j \le n_2 - 1} S_{2,j}$; $D_{19}^* = \sum_{3 \le i \le n_1 - 1} S_{i,1}$; $D_{20}^* = \sum_{3 \le i \le n_1 - 1} S_{i,2}$;

$D_{21}^* = \sum_{3 \le i \le n_1 - 1} \sum_{3 \le j \le n_2 - 1} S_{i,j}$; $D_{22}^* = \sum_{3 \le i \le n_1 - 1} S_{i,n_2}$; $D_{23}^* = \sum_{3 \le j \le n_2 - 1} S_{n_1,j}$

<table>
<tr><td></td><td>0</td><td>1</td><td>2</td><td>3</td><td>…</td><td>n_2-3</td><td>n_2-2</td><td>n_2-1</td><td>n_2</td></tr>
<tr><td>0</td><td>X</td><td>D^*_1</td><td>D^*_2</td><td colspan="4">D^*_3</td><td>D^*_{11}</td></tr>
<tr><td>1</td><td>D^*_5</td><td>D^*_6</td><td>D^*_7</td><td colspan="4">D^*_{17}</td><td>D^*_{12}</td></tr>
<tr><td>2</td><td>D^*_8</td><td>D^*_9</td><td>D^*_{10}</td><td colspan="4">D^*_{18}</td><td>D^*_{13}</td></tr>
<tr><td>3</td><td rowspan="5">D^*_4</td><td rowspan="5">D^*_{19}</td><td rowspan="5">D^*_{20}</td><td colspan="4" rowspan="5">D^*_{21}</td><td rowspan="5">D^*_{22}</td></tr>
<tr><td>…</td></tr>
<tr><td>n_1-3</td></tr>
<tr><td>n_1-2</td></tr>
<tr><td>n_1-1</td></tr>
<tr><td>n_1</td><td>D^*_{14}</td><td>D^*_{15}</td><td>D^*_{16}</td><td colspan="4">D^*_{23}</td><td>X</td></tr>
</table>

Appendix 3: Summaries of the Joint-Frequency Spectrum for the Composite-Likelihood analysis method

Here we consider singletons, doubletons, and polymorphic sites with high frequencies n_1 -1 and n_1 -2 in population 1 or n_2 -1 and n_2 -2 in population 2 ($\breve{D}_k$, k=1,…,23) separately.

$$\breve{D}_1 = S_{1,0} + S_{2,0};\ \breve{D}_2 = \sum_{3\le i\le n_1-3} S_{i,0};\ \breve{D}_3 = S_{n_1-2,0} + S_{n_1-1,0};\ \breve{D}_4 = S_{n_1,0};$$

$$\breve{D}_5 = S_{0,1} + S_{0,2};\ \breve{D}_6 = S_{1,1} + S_{2,1} + S_{1,2} + S_{2,2};\ \breve{D}_7 = \sum_{3\le i\le n_1-3} S_{i,1} + \sum_{3\le i\le n_1-3} S_{i,2};$$

$$\breve{D}_8 = S_{n_1-2,1} + S_{n_1-1,1} + S_{n_1-2,2} + S_{n_1-1,2};\ \breve{D}_9 = S_{n_1,1} + S_{n_1,2};\ \breve{D}_{10} = \sum_{3\le j\le n_2-3} S_{0,j};$$

$$\breve{D}_{11} = \sum_{3\le j\le n_2-3} S_{1,j} + \sum_{3\le j\le n_2-3} S_{2,j};\ \breve{D}_{12} = \sum_{3\le i\le n_1-3}\sum_{3\le j\le n_2-3} S_{i,j};\ \breve{D}_{13} = \sum_{3\le j\le n_2-3} S_{n_1-2,j} + \sum_{3\le j\le n_2-3} S_{n_1-1,j};$$

$$\breve{D}_{14} = \sum_{3\le j\le n_2-3} S_{n_1,j};\ \breve{D}_{15} = S_{0,n_2-2} + S_{0,n_2-1};\ \breve{D}_{16} = S_{1,n_2-2} + S_{1,n_2-1} + S_{2,n_2-2} + S_{2,n_2-1};$$

$$\breve{D}_{17} = \sum_{3\le i\le n_1-3} S_{i,n_2-2} + \sum_{3\le i\le n_1-3} S_{i,n_2-1};\ \breve{D}_{18} = S_{n_1-2,n_2-2} + S_{n_1-2,n_2-1} + S_{n_1-1,n_2-2} + S_{n_1-1,n_2-1};$$

$$\breve{D}_{19} = S_{n_1,n_2-2} + S_{n_1,n_2-1};\ \breve{D}_{20} = S_{0,n_2};\ \breve{D}_{21} = S_{1,n_2} + S_{2,n_2};\ \breve{D}_{22} = \sum_{3\le i\le n_1-3} S_{i,n_2};\ \breve{D}_{23} = S_{n_1-2,n_2} + S_{n_1-1,n_2}$$

For a simple model with equal mutation rates in the two populations ($\theta_1 = \theta_2$) and equal gene flow rates ($M_{12} = M_{21}$), we verify by coalescent simulations of the Wakeley-Hey model that the JSFS shows an axis of symmetry along the diagonal (0,0) to (n_1,n_2). In this case, symmetry also appears among elements of the J vector, namely: $\breve{D}_1$=$\breve{D}_5$, $\breve{D}_2$=$\breve{D}_{10}$, $\breve{D}_3$=$\breve{D}_{15}$, $\breve{D}_4$=$\breve{D}_{20}$, $\breve{D}_7$=$\breve{D}_{11}$, $\breve{D}_8$=$\breve{D}_{16}$, $\breve{D}_9$=$\breve{D}_{21}$, $\breve{D}_{13}$=$\breve{D}_{17}$, $\breve{D}_{14}$=$\breve{D}_{22}$, $\breve{D}_{19}$=$\breve{D}_{23}$. This means that the number of sites with a mutation fixed in population 1 and absent in population 2 ($\breve{D}_4$) is equal to the number of mutations fixed in population 2 and absent in population 1 ($\breve{D}_{20}$).

<table>
<tr><td></td><td>0</td><td colspan="2">1</td><td>2</td><td>3</td><td>…</td><td>n_2-3</td><td>n_2-2</td><td>n_2-1</td><td>n_2</td></tr>
<tr><td>0</td><td>X</td><td colspan="2">$\check{D}_5$</td><td colspan="3">$\check{D}_{10}$</td><td colspan="2">$\check{D}_{15}$</td><td>$\check{D}_{20}$</td></tr>
<tr><td>1</td><td rowspan="2">$\check{D}_1$</td><td colspan="2" rowspan="2">$\check{D}_6$</td><td colspan="3" rowspan="2">$\check{D}_{11}$</td><td colspan="2" rowspan="2">$\check{D}_{16}$</td><td rowspan="2">$\check{D}_{21}$</td></tr>
<tr><td>2</td></tr>
<tr><td>3</td><td rowspan="3">$\check{D}_2$</td><td colspan="2" rowspan="3">$\check{D}_7$</td><td colspan="3" rowspan="3">$\check{D}_{12}$</td><td colspan="2" rowspan="3">$\check{D}_{17}$</td><td rowspan="3">$\check{D}_{22}$</td></tr>
<tr><td>…</td></tr>
<tr><td>n_1-3</td></tr>
<tr><td>n_1-2</td><td rowspan="2">$\check{D}_3$</td><td colspan="2" rowspan="2">$\check{D}_8$</td><td colspan="3" rowspan="2">$\check{D}_{13}$</td><td colspan="2" rowspan="2">$\check{D}_{18}$</td><td rowspan="2">$\check{D}_{23}$</td></tr>
<tr><td>n_1-1</td></tr>
<tr><td>n_1</td><td>$\check{D}_4$</td><td colspan="2">$\check{D}_9$</td><td colspan="3">$\check{D}_{14}$</td><td colspan="2">$\check{D}_{19}$</td><td>X</td></tr>
</table>

Appendix 4: Relative of error in estimates of divergence times and error in migration rates depending on the method.

Here we present the results of the power analysis for sets of 7 loci with 20 replicates (results in text in Figures 1-2), for given values of θ and ρ.

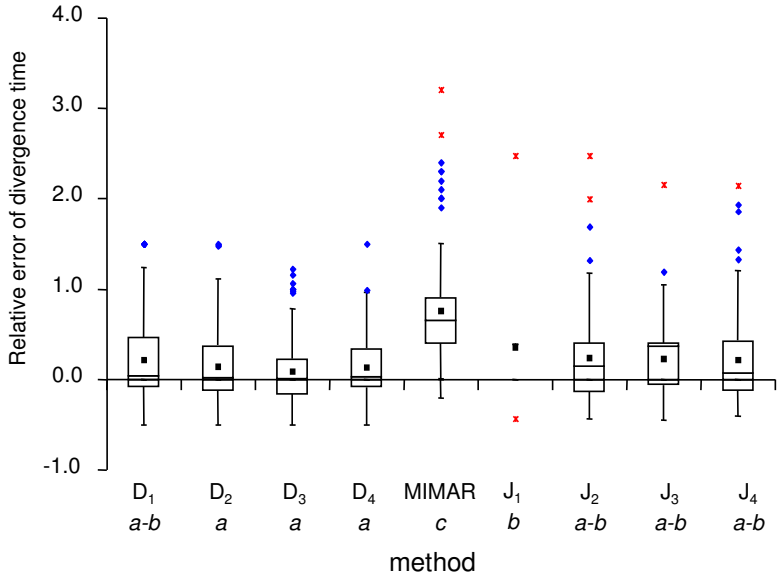


Figure S1a: Relative error for estimates of the divergence time (τ) for the maximum likelihood methods (D_1-D_4), MIMAR and the composite-likelihood methods (J_1-J_4). Relative error is calculated as $(\tau_{est} - \tau_{sim})/\tau_{sim}$ where τ_{est} is the estimated value and τ_{sim} is the simulated value. Groups with significant differences between means following multiple comparisons (Tukey HSD test at 5%) are indicated by letters for each method (group *a* for the smallest mean). Values that are more than 1.5 times the nearest interquartile range (25% or 75%) are displayed as diamonds, those more than 3 times are displayed as stars.

Note that only three fixed values of the divergence time were estimated using method J_1. The value of $\tau = 0.13954$ (relative error = 0.39541, black rectangle in Figure S1a) was the most frequently estimated value with 111 occurrences over the 140 datasets using method J_1.

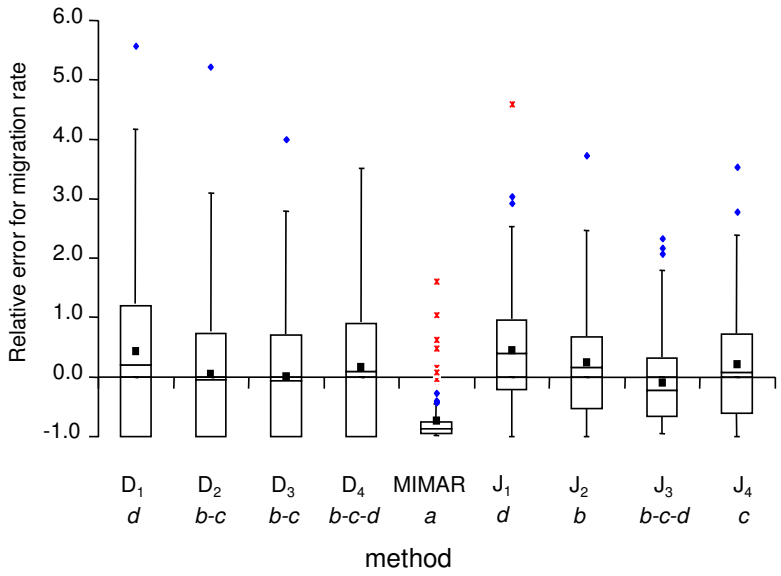
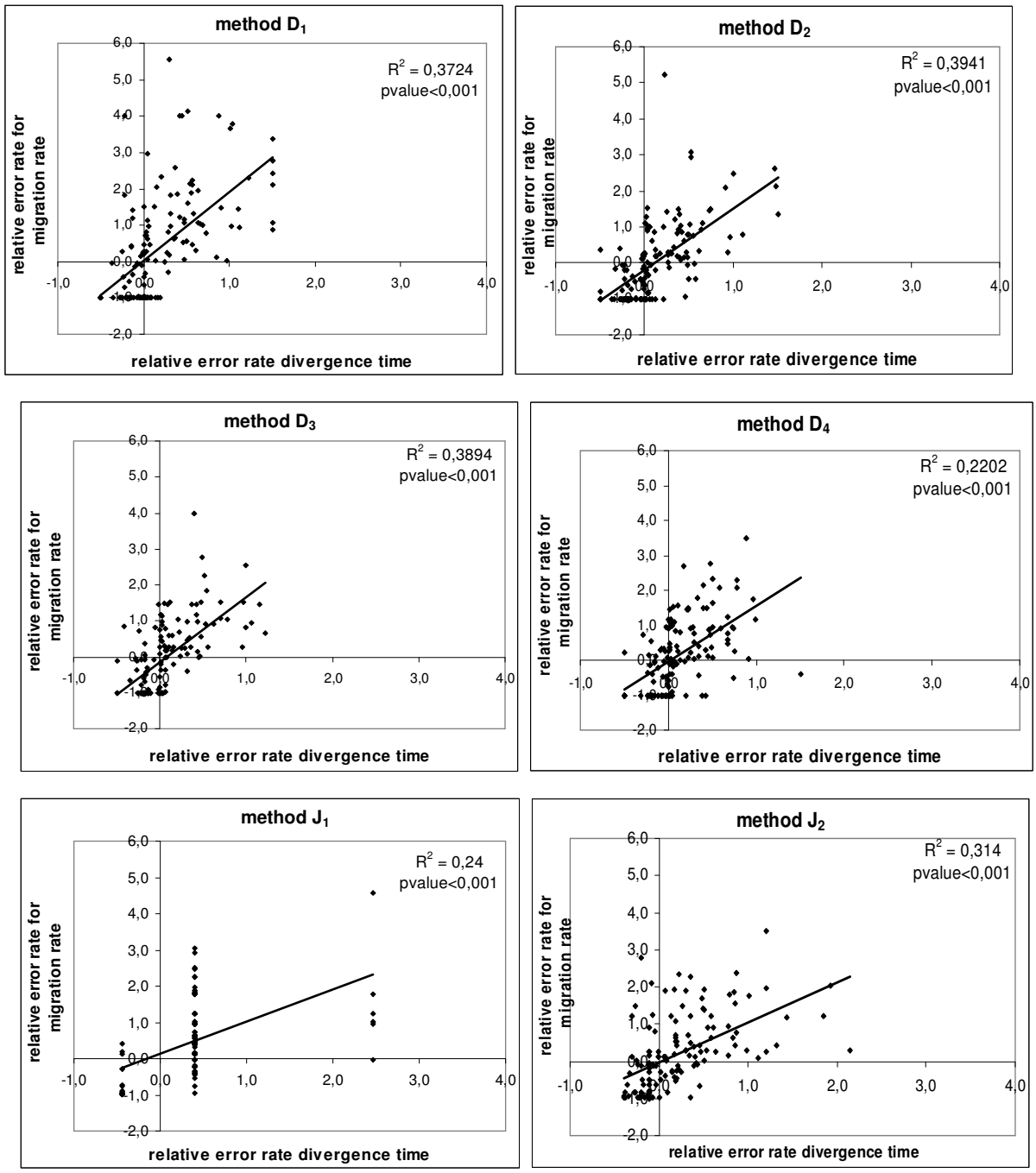
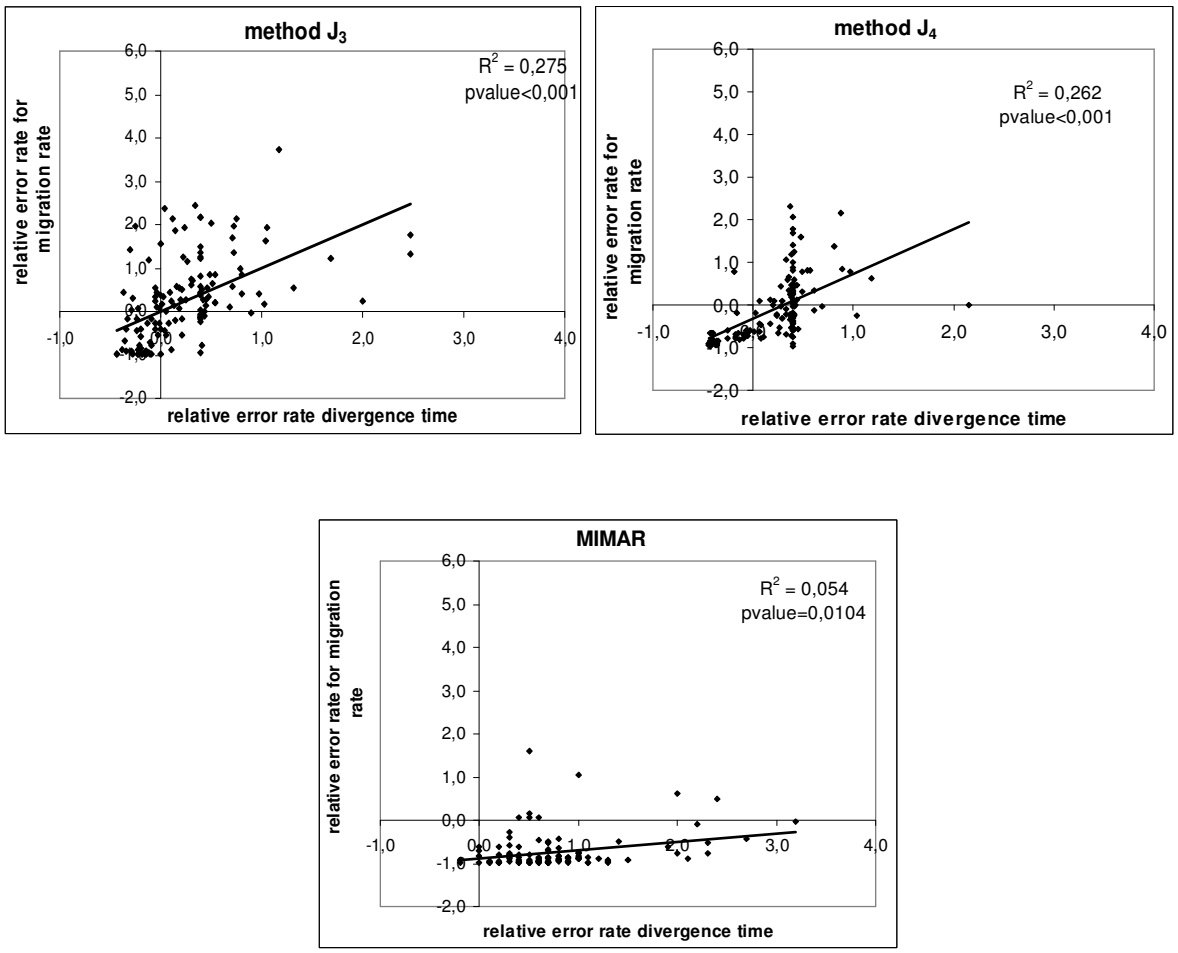


Figure S1b: Relative error for estimates of the migration rate ($M = M_{12} = M_{21}$) for the maximum likelihood methods (D_1-D_4), MIMAR and the composite-likelihood methods (J_1-J_4). Relative error is calculated as $(M_{est} - M_{sim}) / M_{sim}$, where M_{est} is the estimated value and M_{sim} is the simulated value. Groups with significant differences between means following multiple comparisons (Tukey HSD test at 5%) are indicated by letters for each method (group *a* for the smallest mean). Values that are more than 1.5 times the nearest interquartile range (25% or 75%) are displayed as diamonds, those more than 3 times are displayed as stars.

Figure S2: Analysis of regression between errors in estimates of migration rate (M_{12}=M_{21}) and divergence time τ for the 9 methods tested. $D_{1\text{-}4}$ for the maximum likelihood methods, $J_{1\text{-}4}$ for the regression methods and MIMAR. Positive (negative) relative error indicates over (under)-estimation of the parameter. Regression coefficients and p-values are calculated using the *lm* function in the R software. P-values indicate the significance of the test whether the slope of the linear regression is zero.





For all nine methods, positive correlations are found between the relative bias in estimates of divergence time and migration rates. This means that when a method over (under)-estimates the divergence time, it also over (under)-estimates the migration rate.

Appendix 5: Analysis of variance for error in estimates of divergence times and error in migration rates depending on the method and other parameters.

The analysis of variance was performed using the *glm* function, and multiple mean comparisons are based on Tukey's HSD test (confirmed by Bonferroni test) as implemented in the R software (R DEVELOPMENT CORE TEAM 2005). Groups of significance for the multiple comparison tests are shown on Figure S1a. In the *glm* function we use the option: family = Gaussian. We considered all possible two way and three way interaction terms between the different parameters (Method, θ, ρ, M), and sequentially remove non-significant interactions. P-values for single parameters and the interaction term Method*θ are similar to those of Table S1 when only those four terms are considered in the ANOVA formula. Here we show the significance (or non-significance) of interesting interactions for the behavior of the different methods (Method= D_1-D_4, MIMAR, J_1-J_4).

Table S1: ANOVA table of analysis of error in the estimation of divergence times (τ).

	Df	Sum of squares	Mean Square	F value	p-value
Method	8	39.656	4.957	24.566	<0.0001***
θ (population mutation rate)	1	1.897	1.897	9.4	<0.01**
ρ (recombination rate)	1	0.126	0.126	0.626	0.429
M (migration rate)	1	0.431	0.431	2.137	0.144
Method*θ	8	6.993	0.874	4.332	<0.0001***
Method*ρ	8	0.748	0.094	0.463	0.882
Method*M	8	2.656	0.332	1.645	0.107
θ * M	1	0.261	0.261	1.292	0.256
Method * θ * M	8	0.916	0.115	0.567	0.805
Residuals	1195	294.811	0.247		

The analysis of variance was performed using the *glm* function, and multiple mean comparisons are based on Tukey's HSD test (confirmed by Bonferroni test) as implemented in the R software (R DEVELOPMENT CORE TEAM 2005). Groups of significance for the multiple comparison tests are shown on Figure S1b. In the *glm* function we use the option: family = Gaussian. We considered all possible two way and three way interaction terms between the different parameters (Method, θ, ρ, M), and sequentially removed non-significant interactions. P-values for single parameters and the interaction term θ*M are similar to those of Table S2 when only those four terms are considered in the ANOVA formula. Here we show the significance (or non-significance) of interesting interactions for the behavior of the different methods (Method= D_1-D_4, MIMAR, J_1-J_4).

Table S2: ANOVA table of analysis of error in the estimation of migration rates (M_{12}=M_{21}).

	Df	Sum of squares	Mean Square	F value	p-value
Method	8	127.58	15.948	16.093	<0.0001***
θ (population mutation rate)	1	0.54	0.54	0.540	0.463
ρ (recombination rate)	1	2.18	2.18	2.205	0.138
M (migration rate)	1	10.72	10.72	10.822	<0.01**
Method*θ	8	7.40	0.925	0.934	0.487
Method*ρ	8	9.60	1.2	1.21	0.288
Method*M	8	5.51	0.689	0.696	0.696
θ * M	1	4.72	4.72	4.763	0.029*
Method * θ * M	8	4.48	0.56	0.565	0.807
Residuals	1195	1356.9	1.136		

Figure S3: Factor 2 as a percentage of the estimates of divergence time (τ) in the range $\tau_{sim}/2 < \tau_{est} < \tau_{sim}\times 2$ as a function of the population mutation rates (θ), values of simulated migration rates ($M_{12}=M_{21}$) and population recombination rates (ρ).The Factor 2 (F_2) is the proportion of data sets for which the estimated value (of τ or M) is at least half and at most twice the simulated value

a) for the four maximum likelihood methods (D_1-D_4) and MIMAR, b) for the four composite-likelihood methods (J_1-J_4) and MIMAR.

Figure S3a

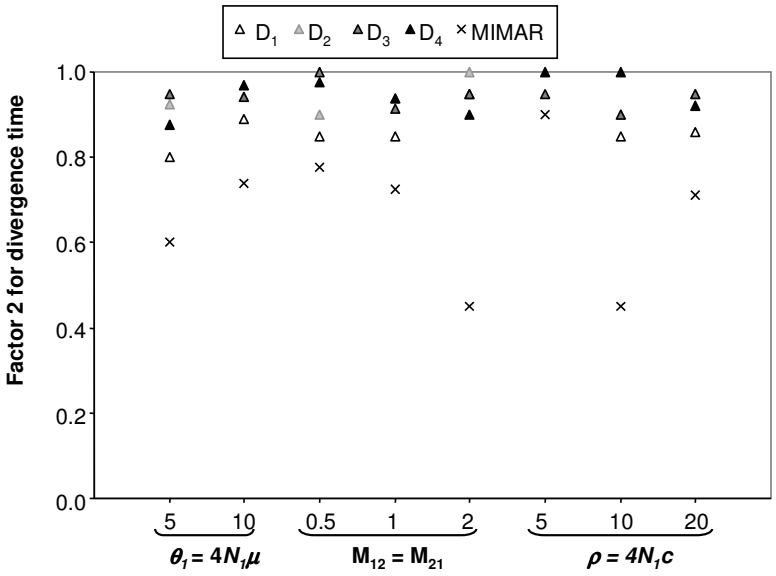


Figure S3b

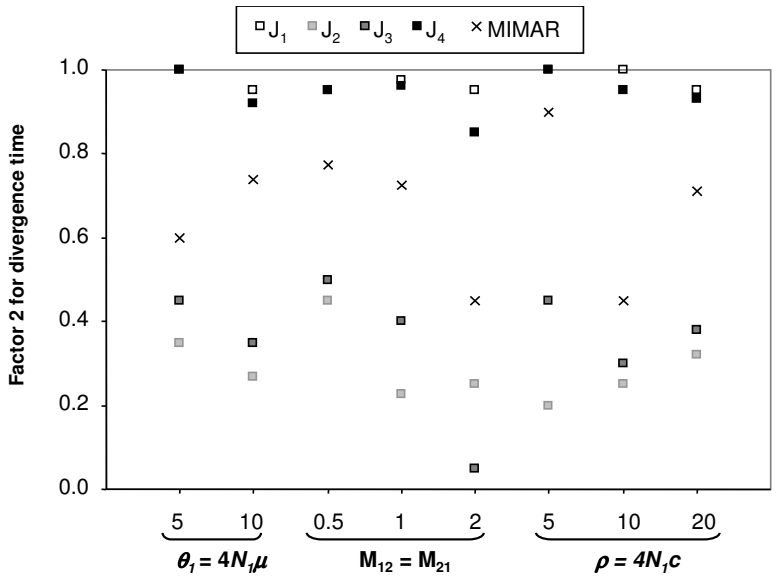


Figure S4: Factor 2 as a percentage of the estimates of migration rate ($M=M_{12}=M_{21}$) in the range $M_{sim}/2 < M_{est} < M_{sim} \times 2$ as a function of the population mutation rate (θ), values of simulated migration rates ($M_{12}=M_{21}$) and population recombination rates (ρ). a) for the four maximum likelihood methods (D_1-D_4) and MIMAR, b) for the four composite-likelihood methods (J_1-J_4) and MIMAR.

Figure S4a

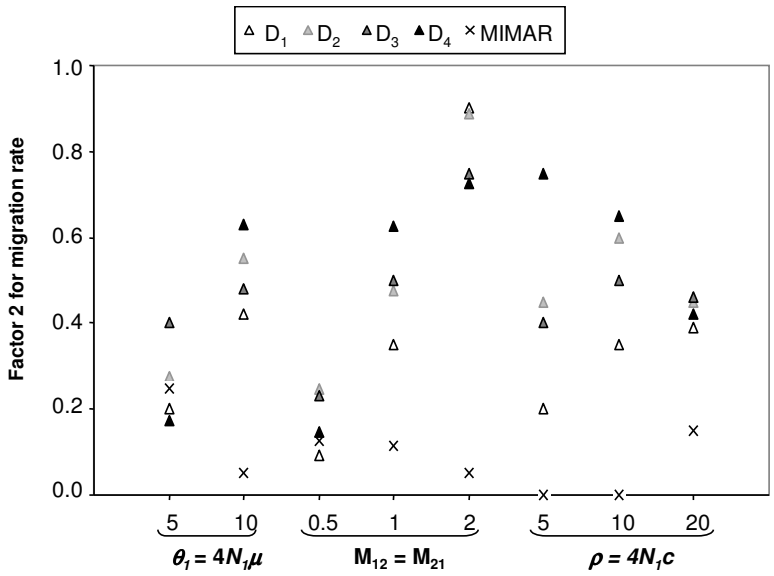
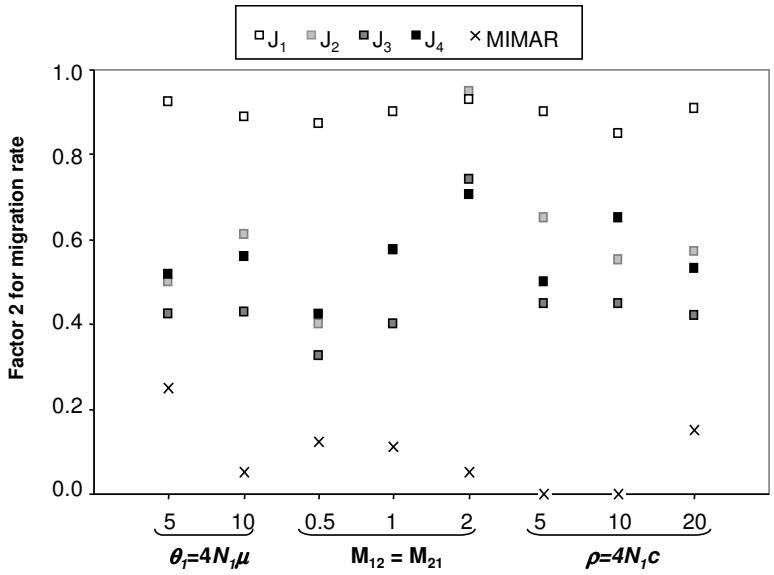


Figure S4b



Appendix 6: Results of the 100 datasets analysis: Factor 2, error in estimates of divergence times and errors in migration rates depending on the method and other parameters.

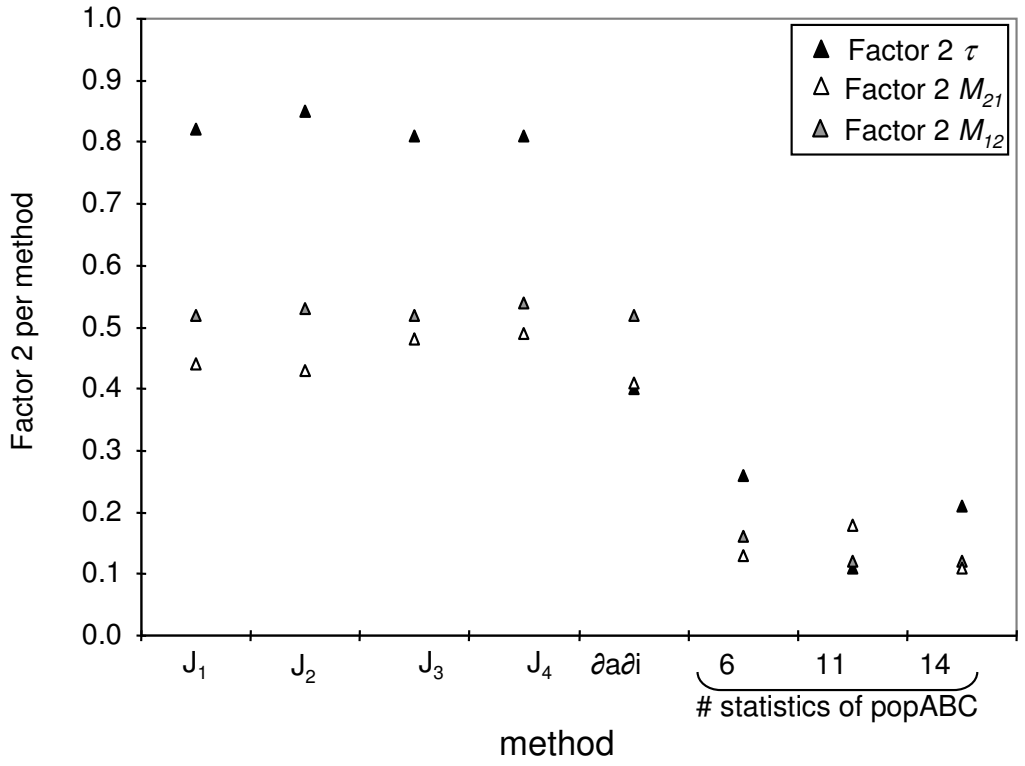


Figure S5: Factor 2 for estimates of the divergence time and migration rates (M_{12}, M_{21}) for the four composite-likelihood methods (J_1-J4), ∂a∂i and for popABC with 6, 11 and 14 summary statistics (computed over 100 datasets).

Figure S6: Distribution of relative error for divergence time (a) and for migration rate (b) depending on the population mutation rate (θ) for composite-likelihood method J_4.

Figure S6a:

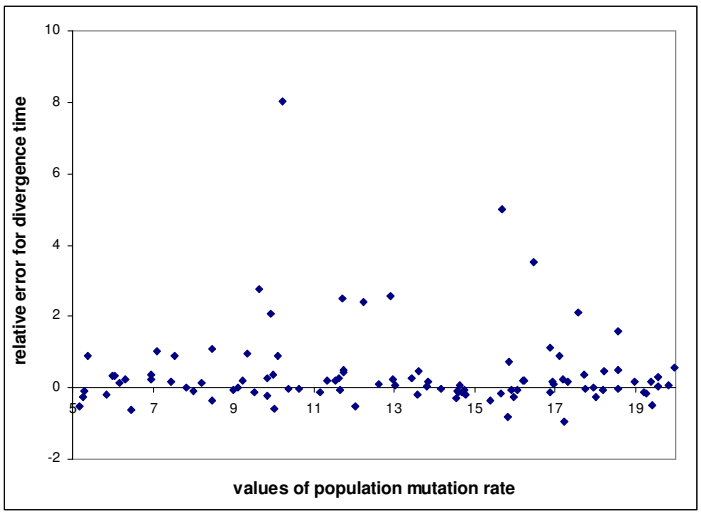
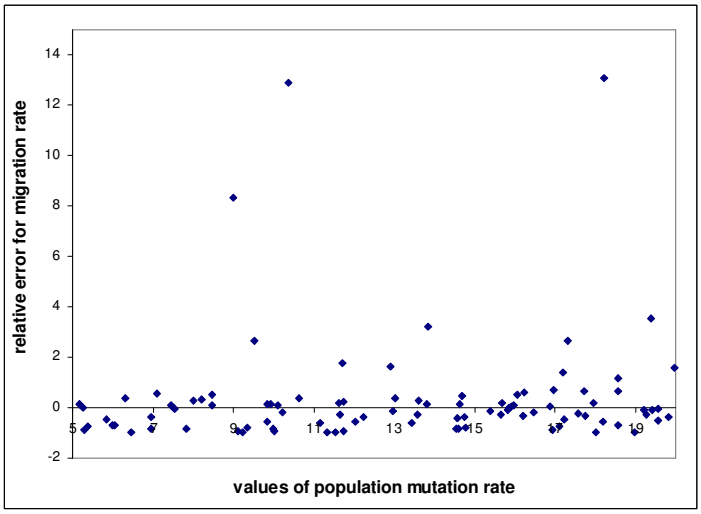


Figure S6b:



For clarity, only relative errors lower than 15 are shown.

Figure S7: Distribution of the relative error of divergence time (a) and of migration rate (b) depending on the population recombination rate (ρ) for composite-likelihood method J_4.

Figure S7a:

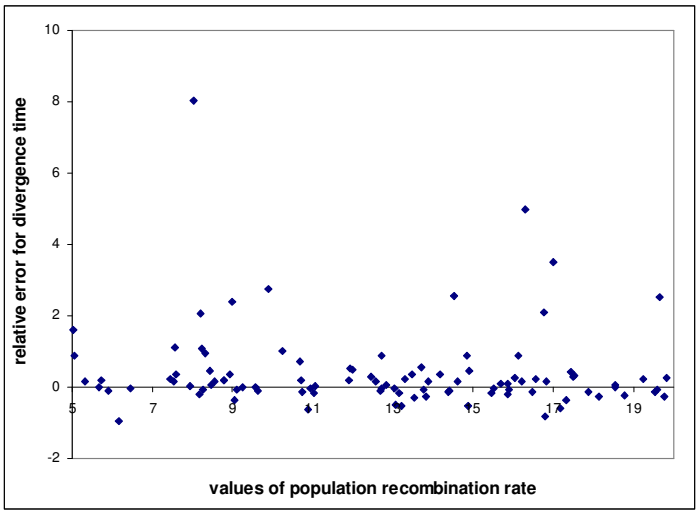
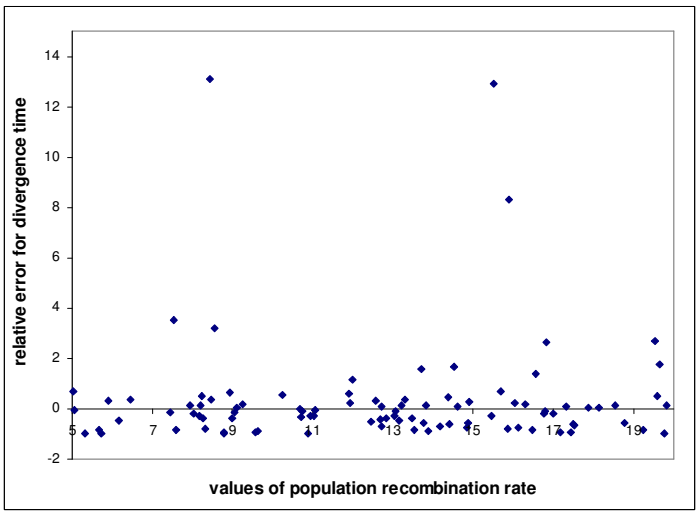


Figure S7b:



For clarity, only relative errors lower than 15 are shown.

Figures S6 and S7 highlight the absence of any clear correlation between error in estimating WH parameters and the population size (θ) or the recombination rate (ρ). These conclusions are valid for all composite-likelihood methods and popABC results.

Figure S8: Relative error for estimation of migration rate depending on the simulated value of the migration rate (M_{12} in blue and M_{21} in red) for composite method J_2. a) for simulated divergence times less than 0.5, and b) for simulated divergence times greater than 1. Note the difference in scale of the y-axes between a and b.

Figure S8a

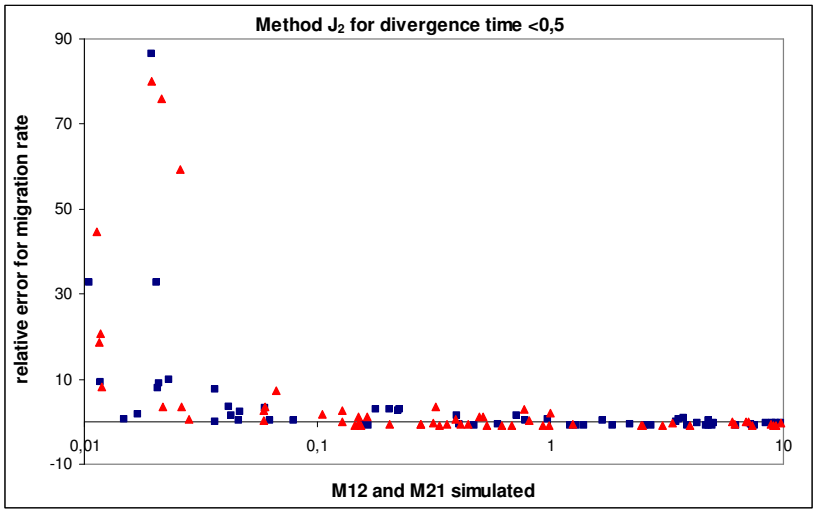


Figure S8b

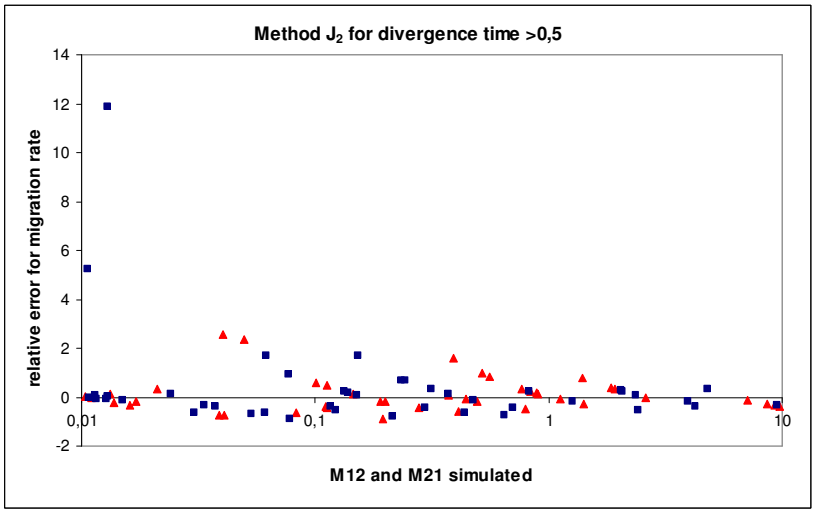


Figure S9: Relative error in the estimation of the migration rate (M_{12} in blue and M_{21} in red) depending on the simulated value of the migration rate for regression method J_4. a) for simulated divergence times smaller than 0.5, and b) for simulated divergence times greater than 1. Note the difference in scale of the y-axes between a and b.

Figure S9a

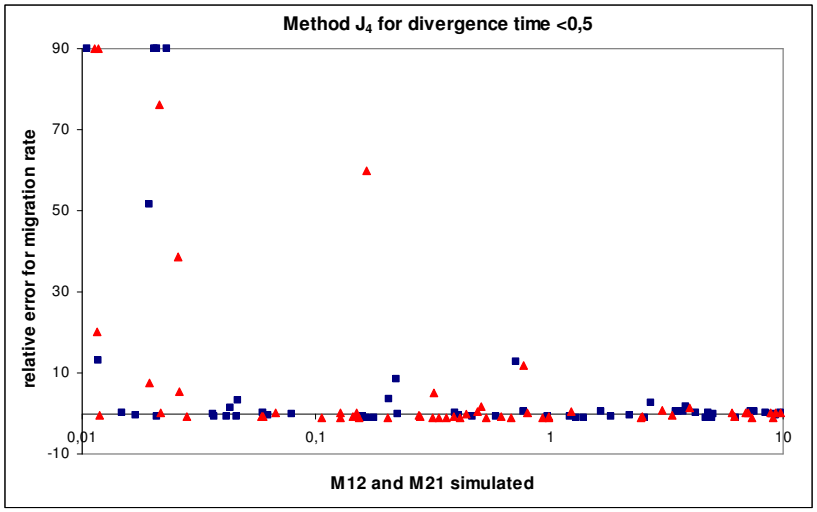


Figure S9b

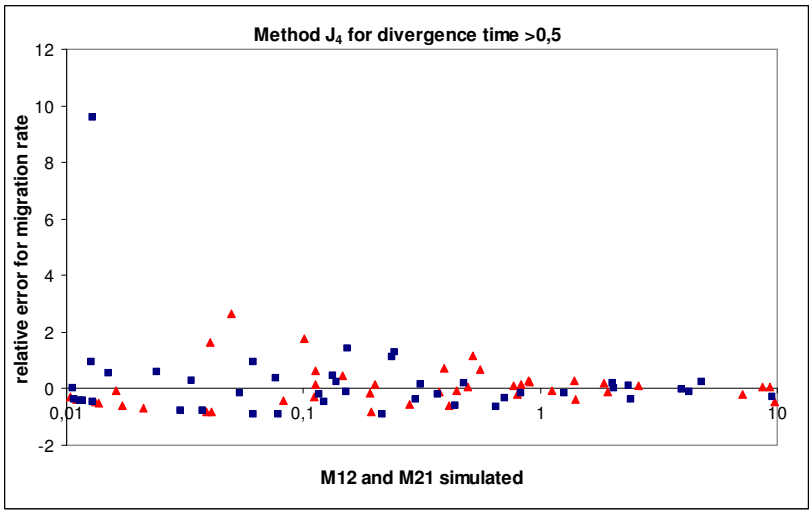


Figure S10: Relative error in the estimation of migration rate depending on the simulated value of the migration rate (M_{12} in blue and M_{21} in red) for popABC estimates with 6 summary statistics. a) for simulated divergence times smaller than 0.5, and b) for simulated divergence times greater than 1.

Figure S10a

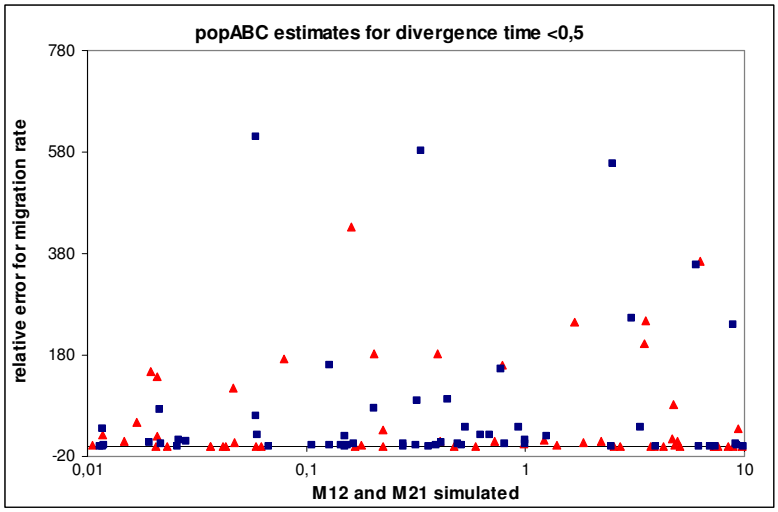
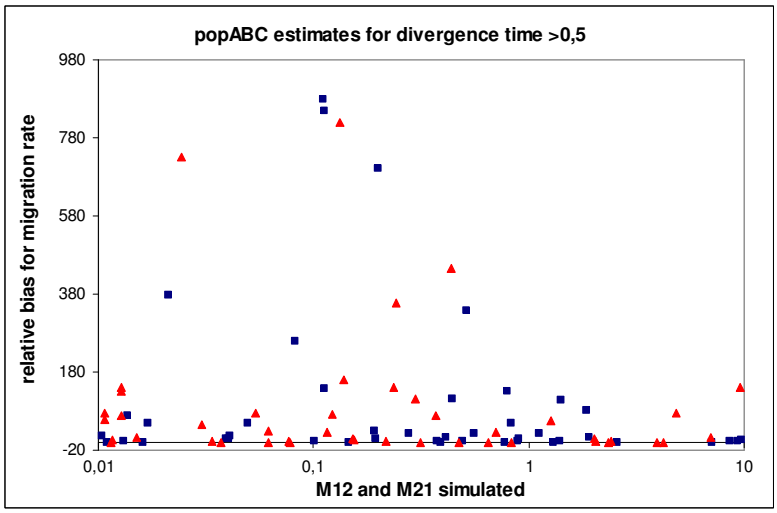


Figure S10b



Figures S7 and S8 show that estimates of migration rates are less accurate for recent divergence times (τ<0.5)(difference in scale of the y-axes in Figures S8a, S9a and S8b and S9b). Moreover, with composite methods J_2 and J_4, high migration rates can be better estimated (have little relative error) even with recent divergence (<0.5)(Figure S8a, S9a). However, we do not find the same trend for popABC (Figure S10), showing the inaccuracy of estimating migration rates with this method independent of divergence.

Appendix 7: Results of the 100 datasets analysis for 100 loci: RMSE for estimates of divergence times and RMSE for migration rates depending on the method and JSFS coarsenings.

Figure S11: Power analysis of the various JSFS coarsenings to estimate divergence time and migration rates for 100 datasets of 100 loci. RMSE are computed for estimates of the (A) divergence time and (B) migration rates ($M_{12} \neq M_{21}$) for the four composite-likelihood methods (J_1-J4) based on six vectors of summary statistics with different numbers elements. The vector W is defined by the Wakeley-Hey 4 classes from Eq. 2, and other vectors D, D', D'', D^* and $\check{D}$ are refined decompositions of the JSFS with higher number of classes.

Figure S11a

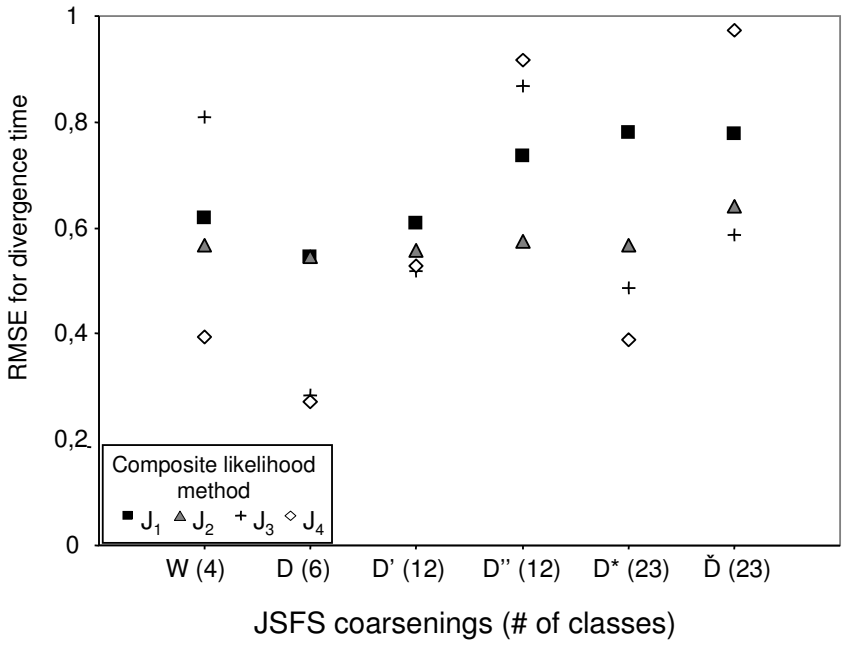
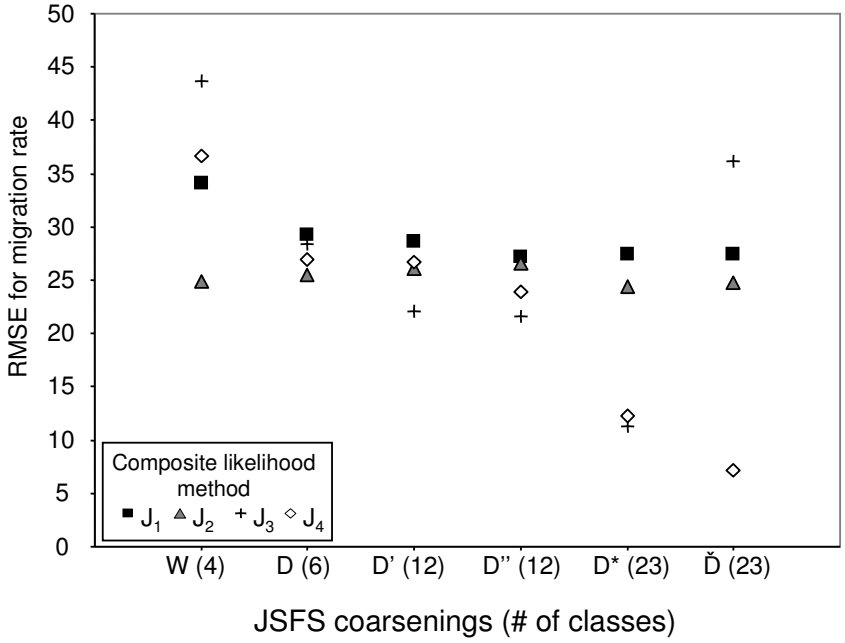


Figure S11b


RMSE for migration rate
50
45
40
35
30
25
20
15
10
5
0
Composite likelihood method
J1 J2 J3 J4
W (4)
D (6)
D' (12)
D'' (12)
D* (23)
Ď (23)
JSFS coarsenings (# of classes)

Appendix C

Supplement to Naduvilezhath *et al.*, *in prep.*

C.1 Parameter Ranges and Command Lines for Demographic Models

C.1.1 Basic model

The data sets were simulated with $nLoci$ loci. The population mutation parameter θ (per locus) and the recombination rate ρ (per locus) were drawn uniformly from the given parameter ranges. The other parameters were chosen uniformly from the following ranges after log transformation:

population-scaled mutation rate $\theta \in [5, 20]$
recombination rate $\rho \in [5, 20]$
size ratios q**,** s_1 **and** $s_2 \in [0.05, 10]$
migration rate $m \in [0.005, 5]$
divergence time $\tau \in [0.017, 20]$

ms command line (Hudson, 2002):

ms 50 $nLoci$ -t θ -r ρ 1000 -I 2 25 25 -m 1 2 m -m 2 1 m -n 2 q -eN τ (s_1+s_2) -ej τ 2 1 -g 1 $\frac{log(\frac{1}{s1})}{\tau}$ - g 2 $\frac{log(\frac{q}{s2})}{\tau}$

C.1.2 Decreasing Migration Model

For the "Decreasing Migration“ model, the parameter values for θ, m, ρ, q, s_1, s_2 were chosen as in the basic model (Sect. C.1.1). The data sets were simulated with $nLoci$ loci with the following two additional parameters drawn uniformly from the parameter ranges after log transformation:

times τ_m and τ_0 $\in [0.017, 15]$

The divergence time τ is the sum of the time with (τ_m) and without (τ_0) gene flow.

ms command line (Hudson, 2002):

ms 50 $nLoci$ -t θ -r ρ 1000 -I 2 25 25 -m 1 2 0 -m 2 1 0
-em τ_0 1 2 $(0.5 \cdot m)$ -em τ_0 2 1 $(0.5 \cdot m)$ -em $(0.5 \cdot \tau_m + \tau_0)$ 1 2 m
-em $(0.5 \cdot \tau_m + \tau_0)$ 2 1 m -n 2 q -eN τ $(s1+s_2)$ -ej τ 2 1 -g 1 $\frac{log(\frac{1}{s_1})}{\tau}$
-g 2 $\frac{log(\frac{q}{s_2})}{\tau}$

C.1.3 Finite-Sites Models

All parameters were chosen as described in the case of the ”Basic Model“ (Sect. C.1.1) with one additional parameter uniformly drawn on the logarithmic scale after log transformation:

Γ-shape parameter α $\in [0.001, 2.5]$

The ms and seq-gen command lines for a HKY model are shown for the ”Basic Model“, where L is the sequence length being simulated, T is the factor of the divergence time to the outgroup, κ is the transition transversion ratio, and α the Γ-shape parameter. The base frequencies following the -f option were always set to the values observed in the tomato loci. The output of ms is a file called ”treeFile“ which serves as an input for seq-gen.

ms (Hudson, 2002) and seq-gen (Rambaut and Grassly, 1997) command line:

ms (50+1) $nLoci$ -r ρ L -I 3 25 25 1 -m 1 2 m -m 2 1 m -n 2 q
-eN τ (s_1+s_2) -ej $T \star \tau$ 3 1 -ej τ 2 1 -g 1 $\frac{log(\frac{1}{s_1})}{\tau}$ -g 2 $\frac{log(\frac{q}{s_2})}{\tau}$ -T | tail
-n +4 | grep -v // > treeFile
seq-gen -mHKY -l L -s $\frac{\theta}{L}$ -p (L+1) -t κ -f 0.26 0.20 0.22 0.32
-a α < treeFile

C.2 Additional Tables and Figures

Table C.1: **Jaatha settings** used for the different analyses. The columns stand for the following settings in Jaatha (for more detailed explanation see Sect. 3.2.3): δ- number of (#) intervals each of the n dimensions is divided into (results in δ^n start blocks), s_{ini}-# simulations per block in the initial search, s_{main}-# simulations per block in the refined search, r-half side length of the blocks in refined search, ϵ-score difference required for stopping, n_{RP}-# best start points, n_{SS}-# summary statistics, ext_θ-`true`: θ is calculated outside of the block with equation 3.2 during refined search and `-false`: θ is calculated like the other parameters, M_{ini}-mutation model for initial search, M_{main}-mutation model for refined search, s_{final}-# simulations for the calculation of likelihoods, t_{max}-maximum # steps during refined search. Weight w was always kept at 0.9, # steps t_{stop} in which there was no score change of at least ϵ at 5, # simulated loci n_{loc} for the GLM fittings at 70, and # best parameter combinations n_B kept in each $\mathcal{L}$ list at 10.

Reference	δ	s_{ini}	s_{main}	r	ϵ	n_{RP}	n_{SS}	ext_θ	M_{ini}	M_{main}	s_{final}	t_{max}
J1	3	200	200	0.05	1	10/16	23	TRUE	IS	IS	200	200
J2	3	100	200	0.05	2	4	23/30	TRUE	FS	FS	100	200
J3	3	300	200	0.05	2	10	23/30	FALSE	FS	FS	100	170
J4	2	300	200	0.1	2	10	23	FALSE	FS	FS	100	170
J5	2	300	200	0.1	2	8	23	FALSE	IS	IS	100	170
J6	3	300	400	0.1	2	10	23	FALSE	IS	IS	100	170
J7	3	300	200	0.1	2	16/10	23	FALSE	FS	FS	200	170
J8	3	300	200	0.1	2	9	23	FALSE	IS	FS	200	170
J9	3	300	200	0.05	2	16	23	FALSE	IS	FS	100	170
J10	3	300	200	0.05	2	16	23	FALSE	IS	IS	200	170
J11	3	300	200	0.1	2	16	23	FALSE	IS	FS	200	170
J12	3	200	300	0.05	1	16	23	FALSE	IS	FS	300	170
J13	2	400	300	0.1	2	16	23	FALSE	FS	FS	300	170
J14	3	200	500	0.05	0.5	40	23	TRUE	IS	IS	200	200
J15	2	300	200	0.1	2	16	23	FALSE	FS	FS	100	170
J16	3	400	200	0.1	2	16	23	FALSE	IS	FS	300	170

Table C.2: **Estimated parameter values for the seven *S. peruvianum* and *S. chilense* loci with alternative settings** under three different models. θ, m and τ are scaled with $4N_1$, where N_1 is the effective population size of *S. chilense*. Bold values are fixed parameter values for the estimation. The corresponding Jaatha settings can be found in Table C.1.

Model	θ	q	m	τ	s_1	s_2	α	# **Parameters**	**Likelihood**	**Settings**
(IS) FixedS2	12.41	4.98	0.59	0.38	**1**	0.29	-	5	-65.47	`J14`
FixedS2+Γ	12.44	7.07	0.35	0.26	**1**	**0.3**	2.5	5	-76.2	`J15`
BothGrowMig	13.25	4.23	0.73	0.57	0.41	0.05	**0.7**	6	-97.7	`J16`

Table C.3: **Jaatha settings and run times** for *Solanum* analyses. The CPU time on a single processor Quad-Core AMD Opteron kernel is reported.

Model	**Settings**	**Run time** [h]
NoMig	`J8`	35
NoMig+Γ	`J7`	109
FixedS2	`J9`	83
(IS) FixedS2	`J10`	2
FixedS2+Γ	`J7`	186
SingleGrowMig	`J11`	84
SingleGrowMig+Γ	`J7`	386
BothGrowNoMig	`J11`	35
BothGrowNoMig+Γ	`J7`	322
BothGrowMig	`J12`	105
BothGrowMig	`J16`	72
BothGrowMig+Γ	`J13`	194
DecMig	`J9`	101
(IS) DecMig	`J1`	19

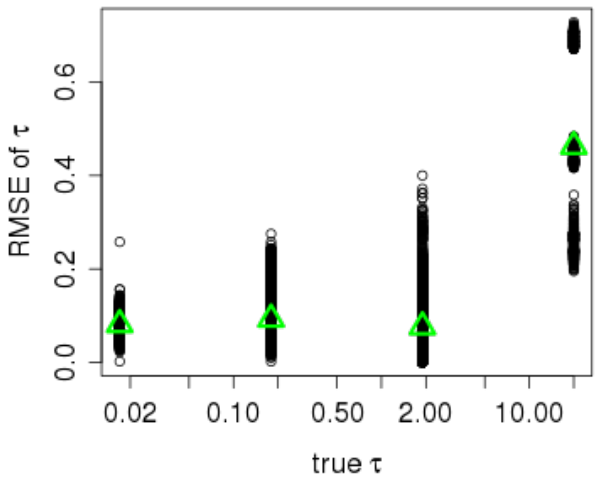
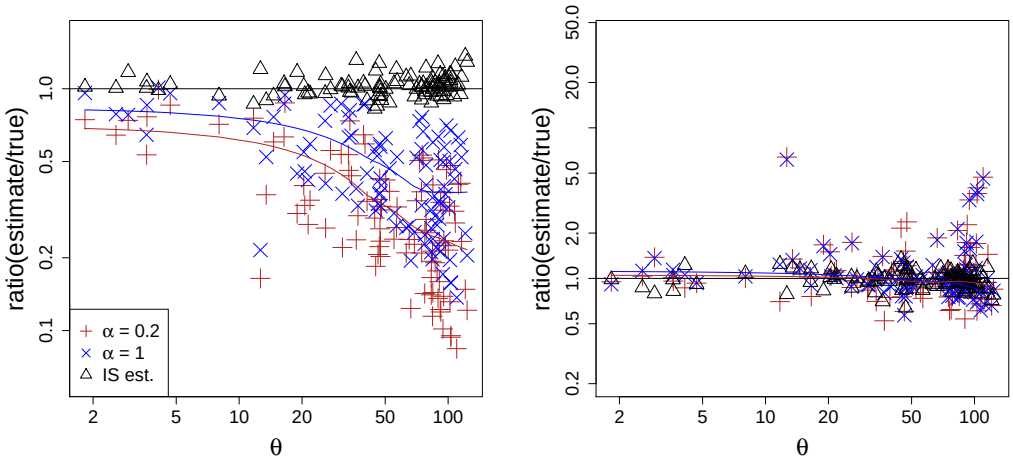


Figure C.1: **Jaatha becomes imprecise when estimating large divergence times** ($\tau = 20$)**.** The true value of the divergence time τ is plotted against the RMSE of τ ($\circ$). The average value is shown as $\triangle$. If τ gets larger Jaatha has trouble finding the correct value of τ.

(a) Population mutation rate θ

(b) Size ratio q



(c) Divergence time τ

(d) Migration rate m

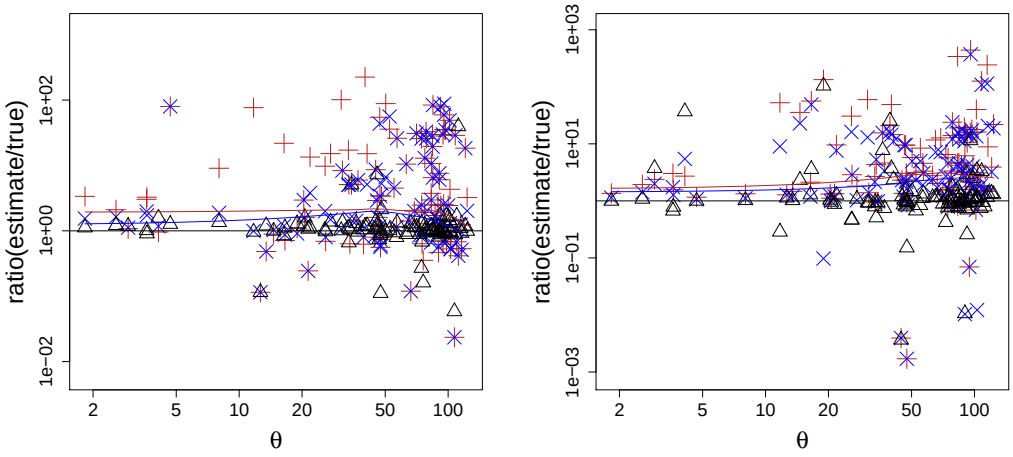
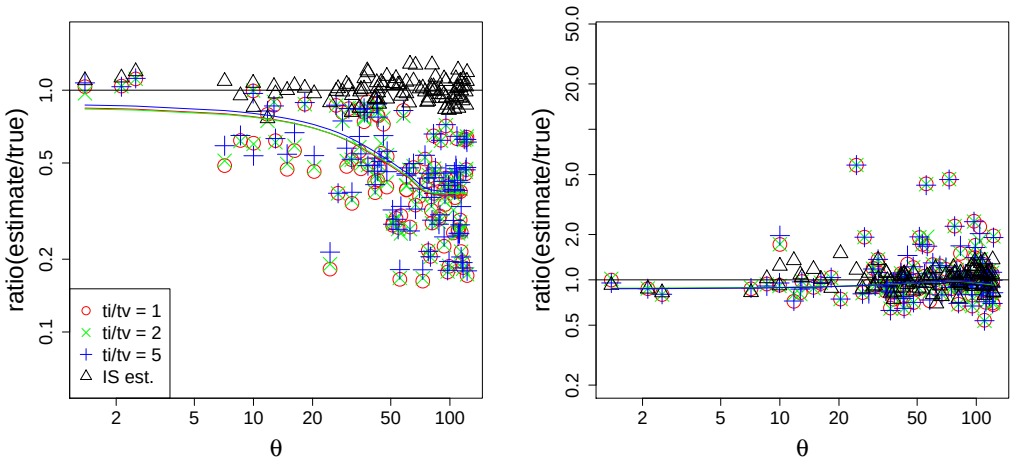


Figure C.2: **The effect of neglecting finite sites on parameter estimation under the "Constant“ model.** The ratio of estimated and true values of θ, q, τ, and m plotted against true θ values under infinite-sites assumptions and the "Constant“ model. Shown are the data sets simulated with the most extreme α values ($\alpha = 0.2$ and 1), $\kappa = 2$, $T = 3$. As a comparison, estimates for infinite sites data sets ($\triangle$) are included. The lines plotted are polynomial regression lines fitted to the ratios (with *lowess* function of R).

(a) Population mutation rate θ

(b) Size ratio q



(c) Divergence time τ

(d) Migration rate m

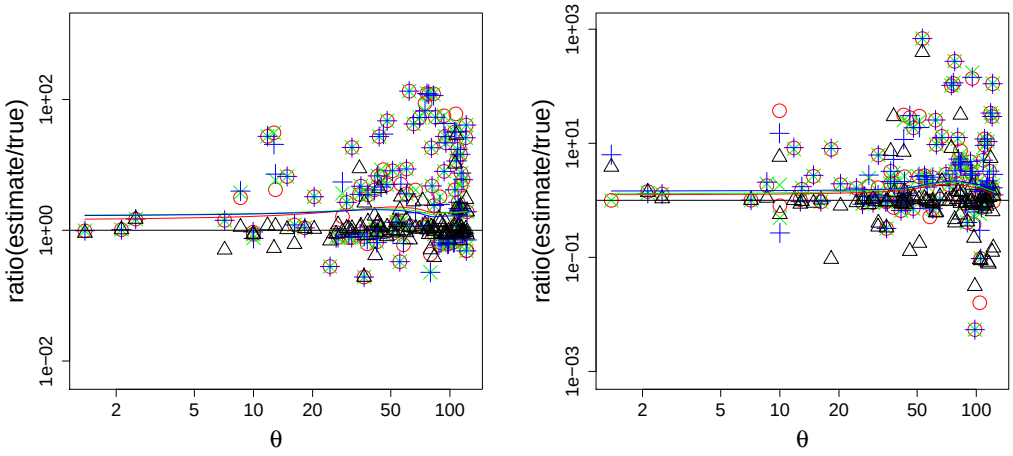


Figure C.3: **Different transition-transversion ratios have almost no influence on the estimations.** The ratio of estimated and true values of θ, q, τ, and m plotted against the true θ values under infinite-sites assumptions for three different values of κ (1, 2, and 5). The data were simulated with a finite sites model with $\alpha = 1$ and $T = 6$ under the "Constant“ model. As a comparison, estimates for infinite sites data sets (△) are included. The lines plotted are polynomial regression lines fitted to the ratios (with *lowess* function of R).

Acknowledgements

I had the great privilege to be supervised by two highly motivated and great advisors, Prof. Dirk Metzler and Prof. Laura Rose. First of all my sincere gratitude goes to Prof. Dirk Metzler who was open for questions approximately 20 hours per day, if not in person then through email. Thanks for all the great explanations, ideas, patiences, and motivation. He introduced me into the topic of population genetics in my first bioinformatics lecture at the Goethe University in Frankfurt in 2005. By that time I was sure that population genetics was the one field I would never go into because it sounded too much like magic. But 'never say never' ... and I was taught better ...

Thank you also to Prof. Laura Rose through whom I had my first (theoretical) contact with the wild tomatoes. It is always enlightening and exciting to talk to her. Thanks also for all your patience with my English and words of encouragement.

I am also grateful to the committee members Prof. Wolfgang Stephan, Prof. Susanne Renner, Prof. Ute Vothknecht, Prof. John Parsch, Prof. Anne-Laure Boulesteix who took the time and effort to join my PhD committee, also throughout the years.

Thanks also to the participants of the theory group of the DFG Forschergruppe 1078. The meetings were inspiring and directing. Also a big thanks to my co-authors of the manuscript. I enjoyed working with you!

For technical support on Linux and the cluster, I would like to say thanks to Werner Schultheiss, Dietfried Molter, and the SuperMuc suppport team.

A tribute to my office mates: I would like to thank my dear friend Meike Wittmann who was not only the best person to discuss ideas with or to ask the silliest questions to but also supported me a lot and encouraged me not to give up. I feel very honored to have shared an office with such a knowledgeable though still young person! To Matthias Birkner who seemed to me like a mathematical encyclopedia. You explained Durrett to me in such a way that I understood it, and importantly, still remember it. Martin Hutzenthaler and Meike thanks for all the soccer games and always funny conversations! I will miss them ... I am happy to know that such a competent person as Paul Staab will continue

the work on Jaatha. Thanks to you for all you input. My first room mates were Stefan Laurent and Nicolas Svetec who made me feel very welcomed in Munich.

A big thanks to Pavlos Pavlidis who convinced me to use Ubuntu and was always open for discussions on any topic! Always with a smile on his face, he never got tired fixing my computer problems.

The time I was allowed to spent in such a nice research environment was great and the people in the whole evolutionary biology group were the ones that made it special, especially thanks to Miriam Linnenbrink (who is the sunshine in person), Lena Müller, Ingrid Kroiss, Claus Kemkemer (who made the lab feel connected), Winfried Hense (for endless philosophical discussions), Anna Vrljic (for providing us indirectly with oxygen and making our office more comfortable).

Finally, my gratitude extends to my friends and family: Brintha Antony, Susanne Franssen, Iris Vargas Jentzsch, Cornelia Pokalyuk, Saumya Jacob, Gaby Olbricht and Luise Stöberl, my big family scattered through the world! Special thanks to the Vellaramkalayils' from Groß-Gerau, Mainz, and Stuttgart and Stegers' from Darmstadt. Through my whole life you made me aware how blessed I am to have such a caring family. I could not imagine my life without you!

My parents Abraham and Leelamma and my sister Asha who have always supported me with love in all possible ways throughout my studies and kept my back free so I didn't have to worry about anything. Thanks also to my invaluable husband, Maneesh Mathew, whose patience and continuous encouragement made this possible. Without all of you I wouldn't be where I am today!

This is for my *great* (both meanings apply here) family.