Statistical inference

of complex demographic models

in *Drosophila melanogaster* and two wild tomato species

Stefan Laurent

München 2010

Erklärung:

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

Ehrenwörtliche Versicherung:

Ich versichere ferner hiermit ehrenwörtlich, dass die vorgelegte Dissertation von mir selbstständig, ohne unerlaubte Hilfe angefertigt wurde.

München, ...

Stefan Laurent

Dekan: Prof. Dr. Benedikt Grothe

1. Gutachter: Prof. Dr. Wolfgang Stephan

2. Gutachter: Prof. Dr. John Parsch

Dissertation eingereicht am: 2 Dezember 2010

Datum der Disputation: 14 Januar 2011

Declaration of contributions as a co-author

In this dissertation I present the work of my doctoral research from June 2007 to November 2010. It is organized in three chapters. All of them are the results of collaborations with other scientists.

For the work presented in the first chapter, Wolfgang Stephan, Laurent Excoffier and I designed the study. Annegret Werzner, Anne Wilkens and I collected the data. Laurent Excoffier and I conducted the statistical analysis. I developed the computational tools required for the analysis. I did the writing and Wolfgang Stephan and Laurent Excoffier did the revisions. This chapter has been accepted for publication pending minor revisions under the following title:

Laurent SJY, Werzner A, Excoffier L, Stephan W. 2011. Approximate Bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. *Mol Biol Evol* (accepted with minor revisions)

For the second chapter Aurelien Tellier and myself designed the study and conducted the analysis. I developed the computational tools for the analysis. Hilde Lainer collected the data for the species-wide sample. Aurelien Tellier wrote most of the manuscript and I wrote the sections about the ABC analysis. The manuscript was revised by myself and Wolfgang Stephan. A paper is in preparation that contains parts of this chapter.

For the third chapter Pavlos Pavlidis and myself designed the study and did the writing. Pavlos Pavlidis did the programming. This paper has been published under the title:

*Pavlidis P, *Laurent S, Stephan W. 2010. msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Mol Ecol Resour* 10:723−727. (*contributed equally)

Statistical inference

of complex demographic models

in *Drosophila melanogaster* and two wild tomato species

Dissertation

an der Fakultät für Biologie

der Ludwig-Maximilians-Universität

München

vorgelegt von

Stefan Laurent

aus Brüssel

München, den 30.11.2010

À Marie

Acknowledgements

I owe my deepest gratitude to Prof. Wolfgang Stephan for giving me the opportunity to do my doctoral research in the field of population genetics under his supervision. I am also deeply indebted to Prof. Laurent Excoffier from the University of Bern for the inspiring collaboration during my stay in his department.

I would also like to express my gratitude to my co-authors: Dr. Aurelien Tellier, for his important contributions to my thesis and to my scientific formation, as well as for sharing with me his passion for science and soccer analyses on a daily basis, Pavlos Pavlidis, for our fruitful collaboration and his constant readiness to help, and finally Annegret Werzner and Anne Wilkens for helping with the sequencing.

I thank Dr. Stephan Hutter for taking the time to answer my questions at the beginning of my PhD, as well as Dr. Pleuni Pennings for our scientific collaboration and for getting me involved in the organization of the DZG PhD Meeting 2009.

I offer my regards and blessings to all members of the Department of Evolutionary Biology in Munich and of the CMPG lab in Bern, who supported me in any respect during these three years.

Finally, I'd like to warmly thank my parents, Elisa, and my closest family for their strong and continued support.

Summary

The aim of this thesis was to use the genealogical information contained in genetic variation profiles of natural populations to describe the evolution of a particular species.

In the first project we analysed the colonization process that brought *Drosophila melanogaster* from Africa to Asia. Southeast Asian populations of the fruit fly *D. melanogaster* differ from ancestral African and derived European populations by several morphological characteristics. It has been argued that this morphological differentiation could be the result of an early colonization of Southeast Asia that predated the migration of *D. melanogaster* to Europe after the last glacial period (around 10,000 years ago). To investigate the colonization process of Southeast Asia, we collected nucleotide polymorphism data for 200 X-linked and 50 autosomal loci from a population of Malaysia. We analysed this new SNP dataset jointly with already existing data from an African and a European population by employing an Approximate Bayesian Computation (ABC) approach. By contrasting different demographic models of these three populations, we do not find any evidence for an early divergence between the African and the Asian populations. Rather, we show that Asian and European populations of *D. melanogaster* share a non-African most recent common ancestor (MRCA) that existed about 2500 years ago.

The second project of my PhD thesis is an analysis of the importance of seed dormancy at the population level in two wild tomato species. Seed banks, that is, plant seeds remaining in soils for several generations before germination, are of practical importance in conservation biology because they diminish the immediate ecological impact of habitat fragmentation and prevent species extinction. From an evolutionary perspective, seed banks increase the genetic diversity of plant populations and buffer the effect of varying climatic conditions by magnifying the effects of good years and by dampening the effects of bad years. In this study we estimate the germination rates for

two wild tomato species (*Solanum chilense* and *Solanum peruvianum*) found in western South-America in a wide range of habitats by using DNA sequences coupled to a coalescent model in combination with ecological data. We develop an ABC framework to integrate ecological information on above ground population census sizes, in order to estimate seed bank and metapopulation parameters for each species. We provide the first evidence that it is possible to disentangle the effect of the metapopulation structure from that of the seed bank on the effective population size and to obtain accurate estimates of germination rates based on a coalescent model.

The third and last project of this thesis is related to the development of a computational tool that facilitates the analysis of nucleotide polymorphism datasets in an ABC framework. With the availability of whole-genome sequence data, biologists are able to test hypotheses regarding the demography of populations. Furthermore, the advancement of the ABC methodology allows the demographic inference to be performed in a simple framework using summary statistics. We present here msABC, a coalescent-based software that facilitates the simulation of multi-locus data, suitable for an ABC analysis. msABC is based on Hudson's ms algorithm, which is used extensively for simulating neutral demographic histories of populations. The flexibility of the original algorithm has been extended so that sample size may vary among loci, missing data can be incorporated in simulations and calculations, and a multitude of summary statistics for single or multiple populations is generated. The source code of msABC is available at http://bio.lmu.de/~pavlidis/msabc.

Zusammenfassung

Ziel dieser Doktorarbeit war, anhand von genetischen Daten die demographische Evolution von Arten zu untersuchen.

Das erste Projekt der vorliegenden Arbeit beschäftigt sich mit der Analyse des Kolonisierungsprozesses, welcher die Fruchtfliege, *Drosophila melanogaster,* von Afrika nach Asia brachte. Südostasiatische Populationen von *D. melanogaster* unterscheiden sich von afrikanischen und europäischen Populationen durch verschiedene morphologische Merkmale, und es wurde die Hypothese aufgestellt, dass dies die Folge einer sehr alten demographischen Spaltung zwischen asiatischen und afrikanischen Populationen sein könnte. Um den Kolonisierungsprozess zu untersuchen, haben wir innerhalb einer malaysischen Population 200 X-chromosomale und 50 autosomale Fragmente sequenziert. Wir haben diesen neuen Datensatz gemeinsam mit bereits bestehenden Daten aus einer afrikanischen und einer europäischen Population anhand einer ABC Methode (Approximate Bayesian Computation) analysiert. Nachdem wir verschiedene demographische Modelle miteinander vergleichen konnten, steht fest, dass eine frühe Spaltung zwischen afrikanischen und asiatischen Populationen eher unwahrscheinlich ist. Stattdessen zeigen wir, dass asiatische und europäische Populationen einen gemeinsamen Ahnen teilen, der vor 2500 Jahre existierte.

Das zweite Projekt beschäftigt sich mit dem Einfluss der Samenruhe auf die Evolution von zwei wilden Tomatenarten der Gattung *Solanum.* Samenbanken (d.h. Samen, die vor der Keimung mehrere Jahre im Boden verweilen) sind ein wichtiges Merkmal, dass die genetische Vielfältigkeit einer Art bestimmt. Die Identifizierung einer Samenbank hat eine praktische Bedeutung in der Konservationsbiologie, da Samenbanken einen kurzfristigen Schutz gegen die Zerstörung von Lebensräumen bieten. Aus evolutionärer Sicht sind Samenbanken wichtig, um die Effekte von

wechselnden klimatischen Bedingungen zu dämpfen. In dieser Studie schätzen wir die Keimungsrate von Samen aus zwei wilden Tomatenarten (*Solanum peruvianum* und *Solanum chilense*) aus Südamerika anhand von DNS-Sequenzen. Dafür haben wir eine ABC Methode entwickelt, die auf einem Koaleszenz- Modell beruht und die ökologische Informationen über die Anzahl von oberirdischen Pflanzen integriert. Wir zeigen zum ersten Mal, dass es möglich ist, die Effekte von Metapopulationsstruktur und Samenruhe auf die effektive Populationsgröße zu trennen, und daher genaue Schätzungen der Keimungsrate zu erhalten.

Das dritte und letzte Projekt meiner Doktorarbeit besteht aus der Entwicklung eines Computerprogramms, das die Analyse von demographischen Modellen durch ABC-Methoden ermöglicht. Mit der anwachsenden Verfügbarkeit von DNS-Sequenzen befinden sich Biologen jetzt in der Lage, Hypothesen zu testen, die die demographische Vergangenheit von Populationen betreffen. Die neuesten Fortschritte der ABC-Methode ermöglichen es, solche Analysen in einem vereinfachten Rahmen zu führen, und zwar anhand von zusammenfassenden Statistiken. Wir präsentieren in diesem Kapitel msABC, ein neues Koaleszenz-Simulationsprogramm für ABC-Analysen in Populationsgenetik.

Table of Contents

List of Figures

List of Tables

General Introduction

Inferring the history of populations from genetic variation data: The most important information stored in DNA is the genetic instruction for the development and the functioning of all known organisms [*]. However, DNA is also carrying another type of information that reveals itself when we observe the differences between DNA molecules belonging to different individuals. While genetic instructions carry information about the future history of organisms, the differences between the DNA molecules that are found in a given population can be exploited to gain insight into its evolutionary past. It has been a major objective for population geneticists, since the foundation of the field at the beginning of the twentieth century, to understand the relations between the evolutionary process and the frequencies of segregating alleles within a population. Population genetics developed a set of theoretical models that describe and quantify the relations between the forces of the evolutionary process and the patterns of genetic diversity within a population or a network of populations. Since the introduction of polymerase chain reaction (PCR) about 30 years ago, the ease with which biologists are able to collect genetic variation data from natural populations has been ever increasing and the interpretation of this new data at the light of the results of theoretical population genetics already yielded some very interesting insights into the evolutionary past of a series of species including our own, *Homo sapiens.* To start with, analyses of genetic variation data sampled from human populations on a world-wide scale showed that our modern gene pool has recent and predominant African origin and also revealed the major demographic events that occurred during the range expansion of *Homo sapiens* across the world (Prugnolle *et al.* 2005; Fagundes *et al.* 2007). The inferred models of human evolution have been shown to be compatible with the information provided by the available fossil record (Manica *et al.*

* *With the exception of RNA viruses*

2007) and helped to clarify the debate concerning the occurrence of interbreeding between *Homo sapiens* and other members of the *Homo* genus by showing that little or no contribution from gene pools of other hominid species is necessary to explain present patterns of genetic variation in humans. The same kind of datasets have also been used to reveal the genetic structure of human populations, even among closely spaced ones. For example, Novembre *et al.* (2008) found a close correspondence between genetic and geographic distances in European human populations such that DNA of a European individual can be used to infer his geographic origin with a surprising accuracy. Results concerning the structure of human populations are not only of interest from an anthropological point of view but also play a key role in statistical medical genetics where genome-wide arrays of allele frequencies are used to infer mutations that are responsible for genetic diseases (McCarthy *et al.* 2008). The same kind of studies have also been conducted on other species. The results of these analyses revealed the history and the structure of populations in situations where poor fossil record is available, as it is the case for chimpanzee, our closest living relative (Bequet *et al.* 2007), or for pathogenic species for which there is a strong medical interest (Szmaragd and Balloux 2007; Jombart *et al.* 2009; Tanabe *et al.* 2010).

Besides demographic history, there is something even more fascinating that can be found in the patterns of genetic variation observed in natural populations: traces left by past events of natural selection. Locating, timing and quantifying such events is important for the understanding of the adaptive history of life because it reveals the evolutionary challenges that species had to cope with and how evolution modified their genetic instructions in such a way that they could successfully adapt to their ever-changing environment. One of the important selective processes that can be investigated by analyzing genetic variation profiles from wild populations is the fixation dynamic of an advantageous mutation in a population. This process is called a selective sweep and has been the focus of much attention during the last decades since its seminal description by Maynard Smith and

Haigh (1974). When a new mutation significantly increases the fitness of its carrying organism, the frequency of this beneficial mutation is expected to rise until it reaches fixation within the population and eventually within the whole species. Because mutations occur on a molecule composed of many physically linked nucleotides, the effect that selection has on the frequency of the advantageous mutation is transmitted to neighboring loci along the DNA molecule. The range on which the frequencies of linked neutral mutations are affected by the transmitted action of positive selection, depends on the crossing-over rate in this genomic region. If local recombination rates are high, crossing-overs will break down the physical links between loci and the transmission of the effect of positive selection on neighboring loci will be constrained locally. The main consequence of such a selective sweep on genetic variation data is a local reduction in diversity around the selected allele. This particular feature can be searched for in genome-wide scans of genetic polymorphism datasets that have been sequenced in small samples of individuals from wild populations. This approach allows, without prior knowledge about the nature of the adaptations, to identify genes that have been targeted by positive selection within one species and to infer the time and the strength of the selective event (Li and Stephan 2006; Stephan 2010).

Approximate Bayesian computation in population genetics: From these results it appears that molecular data provides biologists with an important amount of information about the demographic and adaptive processes that have affected the evolution of natural populations. However, the current increase of the complexity of molecular datasets and of the models built to address more complex biological questions represent an important challenge for the statistical methods used in these studies. Most statistical approaches that have been developed in the early years of molecular population genetics are model-based and rely on the derivation of likelihood functions (Beaumont and Rannala 2004; Kuhner 2009). Many of these methods, however, are limited by the difficulty of

computing the likelihood functions when the models and datasets are becoming too complex (large numbers of parameters and loci). Early studies have therefore been restricted to the analysis of small datasets using simple theoretical models for which likelihood functions could be derived (Wilson and Balding 1998; Beerli and Felsenstein 2001). It seems that these approaches will probably not be able to keep up with the important amount of data that are already produced by modern high-throughput DNA sequencing technologies (Metzker 2010). New statistical methods had to be developed to bypass the problem of computing exact likelihoods. One of these approaches is approximate Bayesian computation (ABC) (Beaumont *et al.* 2002; Csilléry *et al.* 2010; Beaumont *et al.* 2010), a method that is characterized by the use of summary statistics and of computer simulations.

A long tradition in population genetics consists in using summary statistics of full genetic polymorphism datasets. These summary statistics are numerical values calculated from the full data such as to maximize the information about a specific aspect of the evolutionary process. In molecular population genetics, summary statistics can be grouped into four classes describing: 1. the amount of genetic variation (Watterson 1975; Tajima 1983), 2. the shape of the distribution of frequencies of mutations (Tajima 1989; Fu and Li 1993; Fay and Wu 2000), 3. linkage disequilibrium, that is, the non-random association between alleles on a chromosome (Kelly 1997; Sabeti 2002), and 4. the amount of genetic differentiation between individuals sampled from different populations (Wright 1951; Hudson *et al.* 1992). These summary statistics correlate with parameters such as population sizes, migration rates, time of demographic events (fission or fusions of populations) or selective pressures. Nevertheless, correlations usually heavily depend on the evolutionary models that are considered. Under relatively simple models, analytical relationships between evolutionary parameters and these statistics could be derived (Wright 1951; Wattersson 1975; Zivcovic and Wiehe 2008). However, for more complex models, population geneticists have

developed a set of computational tools to be used for simulating the evolution of genetic information in artificial populations and computing summary statistics on these simulated datasets. The most widely used method to do this is a simulation algorithm that is based on the coalescent process (Hudson 2002; Wakeley 2009), a retrospective model in population genetics that traces all copies from a gene, found in a sample of individuals, to a single ancestral copy known as the most recent common ancestor [*]. The fact that vectors of summary statistics can be easily simulated under a range of complex evolutionary scenarios is one of the key aspects of the ABC estimation procedure in population genetics. This method compares a vector of summary statistics computed on the observed data with those computed on a large numbers of simulated datasets for which the parameters of interest are known. It can therefore be applied when suitable likelihoods are not available or computationally prohibitive and allows population geneticists to use their datasets to make inferences about complex evolutionary histories and to compare their theoretical models to observations made in natural populations. This is done in a Bayesian framework in which the important steps are model building, model fitting and model improvement.

In evolutionary biology, theoretical models tend to be explanatory rather than predictive. The interest of statistical analyses of genetic variation found in natural populations is to uncover the evolutionary history that generated the present data and not to predict what will happen to these populations. The formulation of the statistical models therefore reflects the different hypotheses that evolutionary biologists have about the processes that generated the diversity of their biological systems. In the ABC framework the formulation of scientific hypotheses into abstract statistical models can be done in a flexible way. External information about the biological system, such as mutation rates, recombination rates, dates of demographic events based on fossil records or other ecological informations can be incorporated into the models. This approach increases the

**Coalescent theory studies the statistical properties of the inheritance relationships of gene copies based on different assumptions about the evolutionary history of the population in which the gene copies were sampled.*

information about the parameters of interest that can be extracted from the data. The fit of a model to the data is done by generating many simulations across a large range of parameter values and models, and by retaining only those parameter values that eventually generated simulated datasets almost identical to the observed data. The proportion of retained simulations generated by a given model is then interpreted as the probability that the observed data has been generated by this model given the set of tested models. The retained values of the parameters of the best model are used to construct the posterior probability distributions for the parameters that needed to be quantified. Although this procedure allows the identification of the model that provides the best fit to a dataset among a set of predefined models, it doesn't report any information about the absolute fit of this model to the data. In other words, in an ABC model choice analysis, the model associated with the highest posterior probability is certainly the least bad of all tested models but might well be unable to fully reproduce all the aspects of the complete data. Information about the absolute fit of a model to the data is an important aspect of such statistical analyses because it indicates whether or not new hypotheses need to be constructed. Knowing which aspects of the data are not correctly predicted by a model also gives an idea about how to improve the model. Assessing the absolute fit of a model to the data is usually done by performing predictive simulations, which are the distribution of future observations conditional on the observed data. This is done by sampling numerical values from the posterior probability distribution obtained for the best model and to use them to generate simulated datasets. Observed statistics can be compared with these simulated distributions to identify aspects of the observed dataset that cannot be predicted by the model.

By avoiding the computation of likelihood functions, ABC simplifies the process of comparing evolutionary hypotheses among each other and against the data. It will certainly establish itself as an important tool to analyse the new full-genome datasets generated by new sequencing technologies and help to confront these new data with the important theoretical framework

developed in evolutionary biology since the beginning of the twentieth century.

Outline of the thesis: During my PhD thesis I applied ABC methods to answer current issues in the study of the evolution of *Drosophila melanogaster* and wild tomatoes species from the genus *Solanum*. My worked encompassed the acquisition of new nucleotide polymorphism datasets, the development of computational tools to perform ABC estimations in population genetics, and the statistical analysis of datasets from *Drosophila* and tomatoes in an ABC framework. This manuscript is organized as a cumulative dissertation composed of three chapters.

The first chapter is a study about the origin of Asian populations of the fruit fly *Drosophila melanogaster*. An ongoing debate about the age of the colonization of the Southeast Asian continent by this species could be resolved by showing that Asian flies diverged only recently from their ancestral African populations. In the second chapter Dr. Tellier and I used an ABC approach to investigate the impact of seed dormancy on the evolutionary history of species from the genus *Solanum*. Seed dormancy is a life-history trait that can have a dramatic impact on the genetic diversity of plant populations. It confers an evolutionary potential that wouldn't be suspected if only observable, above-ground individuals would be considered. The third and last chapter of my dissertation is the description of a software written by Pavlos Pavlidis and myself to facilitate coalescent simulations in an ABC framework. It consists of an important modification of the standard coalescent simulator ms (Hudson 2002) that allows the user to set prior distributions on demographic parameters and to summarize the simulated datasets into a series of summary statistics.

Chapter 1

ABC analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia

Laurent SJY, Werzner A, Excoffier L, Stephan W. *Mol Evol Biol.* in press.

Abstract

Southeast Asian populations of the fruit fly *Drosophila melanogaster* differ from ancestral African and derived European populations by several morphological characteristics. It has been argued that this morphological differentiation could be the result of an early colonization of Southeast Asia that predated the migration of *D. melanogaster* to Europe after the last glacial period (around 10,000 years ago). To investigate the colonization process of Southeast Asia, we collected nucleotide polymorphism data for more than 200 X-linked fragments and 50 autosomal loci from a population of Malaysia. We analysed this new SNP dataset jointly with already existing data from an African and a European population by employing an Approximate Bayesian Computation approach. By contrasting different demographic models of these three populations, we do not find any evidence for an early divergence between the African and the Asian populations. Rather, we show that Asian and European populations of *D. melanogaster* share a non-African most recent common ancestor (MRCA) that existed about 2500 years ago.

Introduction

The demographic history of wild populations of the fruit fly *Drosophila melanogaster* has been a subject of investigation for several decades (David and Capy 1988; Baudry *et al.* 2004; Stephan and Li 2007). This is mainly due to the fact that this species is used as a model organism in studies of local adaptation (Ometto *et al*. 2005; Li and Stephan 2006; Pool *et al*. 2006). These studies attempt to detect evidence for past events of positive selection by scanning the genome for specific patterns of genetic variation. In these analyses, the identification of major demographic events that affected the populations in the recent past plays an important role for at least two reasons. First, being able to identify the ancestral and derived populations and obtaining reliable estimates for the times at which the derived populations colonized new habitats is critical in understanding to which environmental conditions these populations had to adapt. Second, studies of local adaptation that are trying to detect genes targeted by positive selection are based on the assumption that the adaptive event required the fixation of beneficial alleles. The rapid increase in frequency of a favorable allele may leave a typical signature in DNA polymorphism data (Maynard Smith and Haigh 1974) that is called a selective sweep. Such signatures can be detected by recently developed methods (Kim and Stephan 2002; Kim and Nielsen 2004; Nielsen 2005; Stephan *et al*. 2006; Jensen *et al*. 2007; Pavlidis *et al.* 2010a). Some of these methods rely on a demographic model that allows them to take into account the fact that neutral demographic forces (such as population size bottlenecks in the recent past) can generate similar signatures as selective sweeps (Barton 1998). Estimation of demographic models can therefore be used to reduce the false positives rate of selective sweeps detection methods and to identify chromosomal regions that are not compatible with neutral demographic scenarios.

Previous studies showed that ancestral populations of *D. melanogaster* live in the African

mainland south of the Saharan desert (Tsacas and Lachaise 1974). Furthermore, historical and morphological lines of evidence indicate that derived populations can be categorized into ancient populations that have colonized the Eurasian continent during prehistoric times, and new populations that have colonized the American and Australian continents during historic times along with recent human migrations (David and Capy 1988). More recently, studies based on a rigorous statistical analysis of genome-wide samples of nucleotide polymorphism data could confirm that *D. melanogaster* originated in sub-Saharan Africa (Li and Stephan 2006). These studies could also show that divergence between African and European populations occurred about 16,000 years ago and that this event was associated with a population size bottleneck (Li and Stephan 2006; Thornton and Andolfatto 2006). In contrast, the timing of the colonization process of the Asian populations has not been identified yet. The first study that analysed Asian populations of *D. melanogaster* contrasted patterns of morphological variation between derived and ancestral populations (David *et al*. 1976). This study showed that Asian populations are characterized by specific morphological and physiological properties such as slower development growth, higher fresh weight, and smaller ovariole numbers than African and European populations. Based on these results the authors proposed the existence of a 'Far Eastern Race' of *D. melanogaster.* One of the hypotheses that has been proposed to explain the morphological divergence was the occurrence of an ancestral divergence between African and Asian populations. This early colonization of Asia would have predated the divergence between African and European flies and occurred before or during the last ice age. In contrast to these results, recent population genetics surveys revealed that non-African populations of *D. melanogaster* share a unique origin (Baudry *et al*. 2004; Schlötterer *et al*. 2006). However, these latter studies do not provide an estimation of the time at which *D. melanogaster* colonized the Asian continent. Furthermore, they do not explicitly model the colonization process.

In this study, we conducted a population genetic analysis of three populations of *D.*

melanogaster from Africa, Europe and Southeast Asia. We sequenced nearly 280 fragments of the X and third chromosome of the Asian population and analysed this new dataset together with existing data that have been collected for the African and European populations (Glinka *et al.* 2003; Ometto *et al.* 2005; Hutter *et al.* 2007). We employed an Approximate Bayesian Computation (ABC) approach (Beaumont *et al.* 2002) to investigate the demographic histories of these three populations. We found that a model in which Asian and European populations share a common non-African ancestor is more likely than a scenario with an independent early colonization of the Southeast Asian region. We show that the divergence between Southeast Asian and European *D. melanogaster* populations occurred about 2500 years ago.

Materials and Methods

Samples: The individuals analysed come from 36 inbred lines sampled from an African population from Zimbabwe, a European population from The Netherlands, and an Asian population from Kuala Lumpur (Glinka *et al.* 2003; Glinka *et al.* 2005).

Collection of DNA sequence data: We sequenced from the Asian sample a subset of the loci that have also been sequenced in the African and European samples (Glinka *et al*. 2003; Ometto *et al.* 2005; Hutter *et al.* 2007). We re-used the same primers to sequence 226 fragments of about 550 bp on the X chromosome and 52 fragments on the third chromosome. Sequences were generated as described in Glinka *et al.* (2003). We aligned the new Asian sequences to the already existing African and European datasets and retained for the demographic analysis only the fragments for which data was available from at least 9 individuals in every population (208 X-linked and 50 autosomal fragments).

Evolutionary scenarios: We analysed two different demographic scenarios to investigate how the Asian population is related to the Africa and European ones. Graphical representations of the two models are given in Figure 1.1 and a description of the prior distributions of the model parameters can be found in Table 1.1. Both models are characterized by the absence of migration between populations and a stepwise expansion in population size of the ancestral African population, as it has been described in Li and Stephan (2006). The first model that we called the Double Colonization Scenario (DCS) describes a demographic history where the Asian and European populations split off independently from the African population (Figure 1.1). This model has been designed to test the hypothesis that the migration of *D. melanogaster* from Africa to Asia predated the colonization of the European continent. The second scenario that we called the Single Colonization Scenario (SCS) describes a situation in which *D. melanogaster* has migrated out of Africa through a single colonization route, in other words, all non-African (cosmopolitan) populations of *D. melanogaster* share a non-African common ancestor (Baudry *et al.* 2004). To model this scenario we relied on previous demographic analyses of the African and European populations. These studies showed that the European population derived from the African one and went through a population size bottleneck (Li and Stephan 2006; Thornton and Andolfatto 2006).

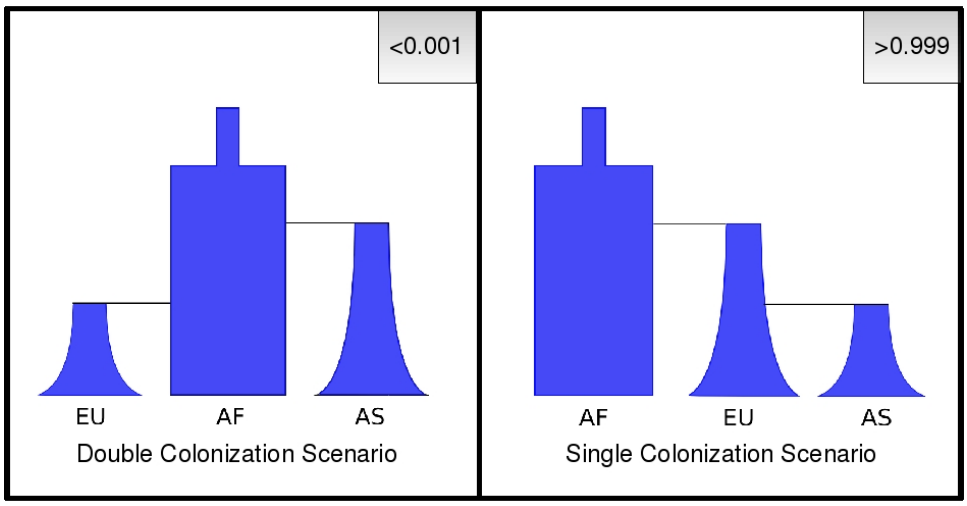


Figure 1.1: Demographic models with associated posterior probability.
AF: Africa, AS: Asia, EU: Europe.

Table 1.1: Prior distributions of the demographic models

Parameter	Prior Distribution			Models
	Min	Max	Distribution	
Sizes				
Current African size	10^5	3×10^7	uniform	All
Current European size	10^4	5×10^6	uniform	All
Current Asian size	10^4	5×10^6	uniform	All
Bottleneck size of the Asian	10	10^5	uniform	All
Bottleneck size of the European	10	10^5	uniform	All
Size of the ancestral population	10^5	2×10^7	uniform	All
Times				
Exit out of Africa	10^2	10^5	uniform	SCS & SCS-2
Divergence between European and Asian population	10^2	Exit out of Africa	uniform	SCS & SCS-2
Exit out of Africa of the Asian population	10^2	10^5	uniform	DCS
Exit out of Africa of the European population	10^2	10^5	uniform	DCS

Sizes are given in effective numbers of individuals (Ne) and times are given in years assuming 10 generations per year. In the coalescent simulations, times were scaled in units of 4Ne generations for the autosomal dataset and in units of 3Ne generations for the X-linked dataset.

We modelled the Asian population assuming that it split off from the European population and underwent a population size bottleneck associated with this founding event (Figure 1.1). Since this model is making the assumption that the European population underwent one bottleneck against two for the Asian population we also investigated an additional model that we called the SCS-2 model, where we inverted the situation and applied one bottleneck to the Asian and two to the European population. This approach allowed the European and Asian populations to experience different levels of genetic drift in our models. It is important to note here that our modelling approach doesn't make any assumption concerning the geographic location of the split between European and Asian populations. (The fact that the Asian population splits off from the European one in the SCS model in Figure 1.1 doesn't mean that this split occurred in Europe.)

Approximate Bayesian inference: To estimate the posterior probabilities of different demographic models and posterior distributions of the parameters of these models, we took an ABC approach. ABC is a computational method in Bayesian statistics that is used in population genetics to perform model-based parameter inference when suitable likelihoods are not available or computationally prohibitive (Pritchard *et al*. 1999; Beaumont *et al*. 2002; Excoffier *et al*. 2005). The method relies on the comparison of a vector of summary statistics computed on the observed data, Δ_{obs}, with those computed on a large number of simulated datasets for which the parameters of interest are known, Δ_{sim}. Here we implemented our ABC algorithm following Excoffier *et al*. (2005).

The algorithm used to estimate the parameters of a model is composed of three steps: a simulation step, a rejection step, and an estimation step. The simulation step consisted in simulating, for every evolutionary scenario, one million datasets that were identical to our observed dataset in terms of numbers of loci and sample sizes. Every evolutionary scenario was defined by a set of parameters (population sizes, age of different demographic events) and every parameter was characterized by a prior distribution (Table 1.1). For each evolutionary scenario we sampled from the prior distribution and used the randomly picked parameter values to perform coalescent-based simulations. The way in which the rejection and the estimation steps have been applied to these simulated datasets differed for the model-choice and the parameter estimation procedures and are described later in the corresponding sections.

To simulate these datasets we incorporated available external information about the local mutation rates (μ) and the local crossing-over rate (r). This allowed us to directly estimate posterior distributions for effective population sizes instead of estimating them for the compound population parameters $\theta = 4 N_e \mu$ and $r = 4 N_e r$ where θ is the coalescent mutation parameter, r the coalescent recombination rate and N_e the effective population size. Mutation rates for every fragment were calculated based on genetic divergence to the sister species *D. simulans* (Kimura 1980) following Li

et al. (1999), that is, by assuming a divergence time between *D. melanogaster* and *D. simulans* of 2.3 My and correcting for pre-speciation divergence. However, divergence-based estimates of mutation rates can potentially be biased by the long-term action of purifying selection on non-coding regions. We therefore took this uncertainty into account by putting a uniform prior distribution on the mutation rate of every simulated fragment centered around the local divergence-based estimate μ_{est} with lower and upper boundaries $\mu_{est}/2$ and $2\mu_{est}$.

Similarly, external information about the local recombination rates was used to generate our simulated datasets. Recombination rates are given as the local rates of crossing-over per site per generation, and were calculated using the *Recombination Rate Estimator* web-based program (Fiston-Lavier *et al.* 2010; see also Hutter and Stephan 2009) available at http://petrov.stanford.edu/RRcalculator.html. However, these estimates have been obtained using a North American population and it is reasonable to think that the real recombination rates in our populations may deviate from these values. To show that the results of our study are robust to this potential deviations, we put a prior distribution on the recombination rate at each locus centered on the Fiston-Lavier *et al.* (2010) estimate r_{est}, with lower and upper boundaries set at $r_{est}/2$ and $2r_{est}$.

The coalescent simulations were performed using a slightly modified version of the program ms that allows to simulate datasets with unequal sample sizes across loci (Hudson 2002; Ross-Ibarra *et al.* 2008). However, to sample parameter values from prior distributions and to compute summary statistics on the simulated data efficiently, we developed our own code. For this we used the GSL C++ library and the libsequence C++ library (Thornton 2003). Simulations were launched on a 64-bit Linux cluster with 510 nodes. We checked for errors in our code by comparing simulation results with similar computations performed by msABC, a coalescent simulator for ABC simulations (Pavlidis *et al.* 2010b).

The model choice procedure: To summarize our datasets for the model choice procedure we computed the following statistics in the three populations: the average and the variance across all fragments of the number of polymorphic sites (S), the average and the variance of Tajima's D (Tajima 1989) and the average Z_{nS} (Kelly 1997). Additionally we computed for all pairs of populations the average distance of Nei, D_A (Nei and Li 1979), and the average proportion of shared polymorphisms between populations, S_S. Monomorphic fragments were removed from the analysis. These statistics have been chosen based on their correlation with important demographic parameters. For example, the number of polymorphic sites is expected to increase with the effective population size and variations of Tajima's D statistic can reflect past fluctuations in population size. The Z_{nS} statistic (Kelly 1997) is a measure of linkage disequilibrium defined as the average r^2 over all pairwise comparisons of polymorphic sites in a sample of sequences, where r^2 is the squared correlation of allelic identity between two loci (Hartl and Clark 1989, pp. 53-54). This statistic is expected to be sensitive to variations in the length of the oldest branches of gene genealogies and can therefore carry information about ancestral population sizes and the severity of population size bottlenecks. The distance of Nei and the proportion of shared polymorphisms are measures of genetic differentiation between our populations and are expected, in a model without migration like ours, to correlate with times of divergence of two populations. All summary statistics presented in this study have been computed using the routines of the C++ library "libsequence" (Thornton 2003).

The posterior probabilities of different demographic models can be estimated on the basis of the Euclidean distance δ between the observed summarized dataset and the simulated summarized datasets of all models. The inference procedure consists in retaining only simulations for which the Euclidean distance is smallest. Pritchard *et al.* (1999) proposed that the posterior probability of a model can be approximated by the proportion of retained simulations under that model, relative to the number of retained simulations under all models. Beaumont (2008) proposed an improvement of

the method that corrects for the fact that retained simulations never exactly match the observed data. The method is based on a weighted multinomial logistic regression procedure, where the response variable is the indicator of the model and the predictor variables are the summary statistics (Fagundes *et al.* 2007; Beaumont 2008).

We computed posterior probabilities for every demographic model, using the 500 simulations associated with smallest Euclidean distance following the method of Beaumont (2008) and applied this procedure to the X-linked and the autosomal dataset independently. Since the same number of datasets have been simulated under each model we computed Bayes factors as the ratio of the posterior probabilities. To investigate if the results of our estimations were stable with regard to the proportion of retained simulations, P_{δ}, we computed posterior probabilities for our three models using several values of P_{δ} ranging from 0.025 to 1%. We also investigated the accuracy of our model choice procedure, following Peter *et al. (*2010). We therefore simulated 1000 pseudo-observed datasets, with the same number of fragments and sample sizes as our autosomal dataset, under each demographic model (DCS and SCS) and computed for each one of them the posterior probability of having been generated under the DCS model and the posterior probability of having been generated under the SCS model. To decide whether a pseudo-observed dataset should be assigned to the DCS or the SCS model we used as an arbitrary threshold value a ratio of posterior probabilities of 10 in favor of one of both models. Precision was measured as the proportion of correctly identified datasets under each model.

Parameters estimation: In ABC estimations, the quality of the analysis generally relies on well-chosen summary statistics (Joyce and Marjoram 2008). This is a sensitive point in the analysis because on the one hand, using a small number of summary statistics may lead to a substantial loss of information compared to the information carried by the full dataset, but on the other hand,

increasing the number of summary statistics can cause two problems. First, summary statistics that are not related to the parameters of the model or that correlate with other summary statistics will be uninformative and will only add noise to the Euclidean distance. Second, correlations between summary statistics will violate the assumption of singularity which is required when performing the locally weighted linear regression for estimating the parameters (Beaumont *et al*. 2002). To overcome this problem Wegmann *et al.* (2009) proposed to reduce the dimensionality of the summarized dataset by performing a partial least-square (PLS) transformation.

The advantage of the PLS transformation is twofold. First, similar to a principal component analysis, it allows extracting a small number of orthogonal components from a matrix composed of a larger number of summary statistics of our dataset. This is leading to a reduction of the uninformative signals of the Euclidean distance and ensures the singularity of the final matrix of summary statistics. Second, in a PLS transformation the dimensionality reduction is coupled with a regression model, and the latent components (i.e. the transformed summary statistics) are constructed to maximize the prediction of the response variable of the regression model (i.e. the parameters of our demographic model) (Boulesteix *et al*. 2007). Although this approach can be applied when estimating the posterior distributions of the parameters of a given demographic model, it cannot be applied for model choice. The reason is that PLS components are constructed independently for every single demographic model whereas our model choice procedure requires that the set of summary statistics remains identical for all compared models.

We summarized our nucleotide polymorphism datasets into 12 summary statistics that carry information about the level of polymorphism, the site frequency spectrum, linkage disequilibrium, and the amount of differentiation between all three populations. All statistics have been computed for every population separately and for the pooled dataset. Summary statistics describing differentiation between populations have been computed for all possible pairwise comparisons. We

used this set of summary statistics to summarize our observed and our simulated data. We constructed the PLS latent components using 10,000 simulated datasets under the best demographic model for the chromosomes X and 3. To do this we employed code available in the ABCtoolbox package (Wegmann *et al.* 2010). Choosing the best number of partial least square components for parameter estimations has been done by investigating the decrease of the root mean square error (RMSE) for every parameter as a function of the number of PLS components. The RMSE indicates the percentage of variation unexplained by the PLS components and is constructed by comparing the simulated parameter values with the ones predicted using a given number of PLS components. We chose the number of components to be used in the parameter estimation procedure such that additional components don't decrease the RMSE of any parameter of the model. The retained PLS components were used to transform the observed and the simulated datasets. The rejection step consisted in computing the Euclidean distance δ between simulated and observed sets of summary statistics and to retain the 5000 simulations closest to the observed data based on their value of δ. Finally, we estimated posterior distributions of the parameters of interest by applying the locally weighted multivariate regression method of Beaumont *et al. (*2002) implemented in the abcEst program (Excoffier *et al*. 2005). We estimated the marginal posterior probability distribution of each demographic parameter using the kernel density estimation method implemented in the R core package and reported the mode and the 95 credibility intervals of these distributions.

To investigate if the results of our estimations were stable with regard to the proportion of retained simulations, P_δ, we re-estimated the marginal posterior distributions of all parameters of the best model using several values of P_δ ranging from 0.01 to 5%. To avoid the posterior distributions to exceed the upper and lower bound of our prior distributions we transformed the data as $z=\log[\tan(1/x))$ as described in Hamilton *et al*. (2005), where x is the original dataset and z is the transformed data.

Posterior predictive simulations: To see how well our best model is able to reproduce the observed data, we performed posterior predictive simulations (Gelman *et al.* 2003; Thornton and Andolfatto 2006). This has been done by sampling parameter values from the probability density functions of the marginal posterior distributions of our best demographic model and by using them to simulate multi-locus summaries of the data. This procedure allowed us to check which aspect of the data could be explained by our model and which aspects might indicate some limitations of our model. We generated 1000 simulated datasets that had the same number of fragments and sample sizes as our X-linked and autosomal datasets and summarized them into the mean and variance of a large number of summary statistics. For every summary statistic we computed the probability that the simulated mean and variance are smaller than the observed ones.

Results

X-linked and autosomal polymorphism patterns in the Asian population: For the X chromosome, we gathered polymorphism data from a total of 226 fragments, spanning 126,154 nucleotide sites (gaps were excluded). 1309 of these sites are polymorphic. Information about this new genome scan is summarized in Table 1.2. On average data could be obtained from 11.6 (of 12) lines. The sequenced fragments show recombination rates ranging between 2.1 and 4.3 x 10^{-8} per bp per generation. The means (SE) of π (Tajima 1983) and θ_W (Watterson 1975) across the X chromosome are 0.0037 (0.0003) and 0.0035 (0.0002), respectively. We found 39 loci with no polymorphism. The average and variance of Tajima's *D* along the X chromosome are 0.17 and 1.28, respectively (Table 1.2). In order to contrast X-linked with autosomal genetic variation we sequenced 52 autosomal loci spanning 28,441 sites from which 323 are polymorphic. Glinka *et al.* (2005) identified four autosomal inversions in the Asian population segregating at frequencies

ranging from 0.04 to 0.24. For this study only lines harboring no inversions were used. The fragments are randomly spread along the third chromosome. On average data could be obtained for 11.6 lines and sequenced fragments show recombination rates between 1.22 and 3.33 x 10^{-8} per bp per generation. The means (SE) of π and θ_W for the autosomal fragments are 0.0039 (0.0005) and 0.0038 (0.0004), respectively, which is slightly higher than for X-linked data. Only four fragments are monomorphic. The average and the variance of Tajima's *D* along chromosome 3 are 0.19 and 0.81, respectively.

Table 1.2: Polymorphism patterns in the Asian population

	X chromosome	**Autosome**
No. of loci	226	52
Average sample size	11.64	11.57
Average length of alignment	567.16	551.04
No. of invariant loci	39	4
Average (SE) θ_w in %	0.35 (0.02)	0.39 (0.0004)
Average (SE) π in %	0.37 (0.03)	0.38 (0.0005)
Average (SE) divergence (*K*) in %	6.8 (0.25)	5.5 (0.54)
Average (SE) θ_w /*K*	0.05 (0.004)	0.086 (0.013)
Average (SE) Tajima *D*	0.17 (0.08)	0.19 (0.12)
Average (SE) Z_{nS}	0.5 (0.02)	0.53 (0.04)

Comparison of the African, European and Asian populations: In order to compare our new Asian data with the data that have previously been collected for the African and European populations (Glinka *et al*. 2003; Ometto *et al*. 2005; Hutter *et al*. 2007), we aligned data from the three populations and analysed the data jointly. Data for all three populations was available for 208 X-linked and 50 autosomal fragments. We computed the number of SNPs *S*, Tajima's *D*, Z_{nS}, and F_{ST} values (Hudson *et al.* 1992) for every population and every fragment and generated distributions of

these statistics across X and chromosome 3. The results of these analyses are summarized in Figure 1.2 and Table 1.3a and 1.3b. Compared to the African and European populations, the Asian population shows the lowest amount of nucleotide diversity (Figure 1.2A) both for X-linked and autosomal data. The distribution of Asian Tajima's *D* values is similar to those observed in the European population although the average *D* is slightly more positive on the Asian X chromosome and the variance slightly smaller on the Asian third chromosome (Figure 1.2B). The Asian population also harbors the highest levels of LD among all three populations both on the X and autosome (Figure 1.2C). The distributions of F_{ST} values between all three pairs of populations indicate that the African and Asian populations are the most differentiated populations, and that the Asian and European populations are the most closely related ones (Figure 1.2D). These analyses show that the Asian population shares several characteristics with the European population: low diversity, large variance in Tajima's *D*, and high LD. This suggests that, similar to the European population (Stephan and Li 2007), the Asian population is also derived, characterized by past population size fluctuations, and might share part of its recent demographic history with the European population. This hypothesis will be further evaluated by the following ABC analysis.

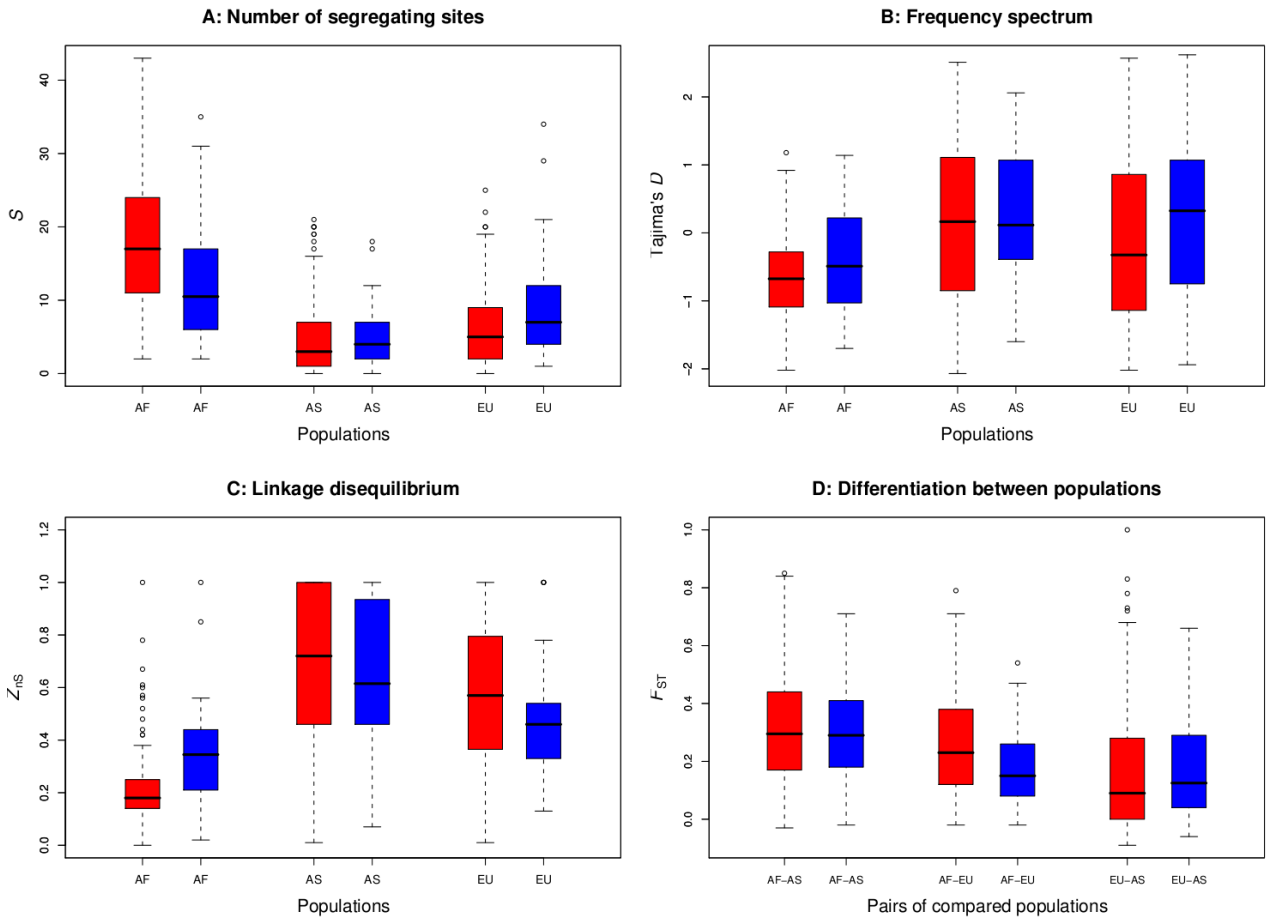


Figure 1.2: Box plots representing the distributions of four summary statistics for all three populations. A) Number of segregating sites, B) site frequency spectrum, C) linkage disequilibrium, and D) differentiation between populations. Results from the analysis of X-linked data are shown in red, while those of autosomal data are shown in blue.

Table 1.3a: Vectors of observed within-population summary statistics for the ABC analysis. These are the observed summary statistics that were used in the ABC analysis

	X				3			
Statistics	Asia	Europe	Africa	pooled	Asia	Europe	Africa	pooled
$\bar{\pi}$	1.67	2.09	5.07	4	1.78	3.08	4.59	3.83
$Var(\pi)$	3.39	4.5	8.44	6.52	2.31	6.97	11.43	7.66
$(\bar{S})$	4.71	6.11	17.42	21.08	5.1	8.84	12.68	17.96
$Var(S)$	24.72	27.2	77.37	99.08	17.85	46.63	77.98	103.59
$\bar{D}$	0.16	-0.11	-0.66	-0.9	0.24	0.15	-0.42	-0.64
$Var(D)$	1.29	1.5	0.37	0.51	0.94	1.21	0.53	0.61
$\bar{K}$	2.8	3.86	9.48	13.97	3.16	4.92	6.22	11.82
$Var(K)$	2.18	3.64	5.26	18.11	1.61	2.81	2.42	14.68
$\bar{Z}_{nS}$	0.53	0.43	0.14	0.11	0.55	0.37	0.23	0.13
$Var(Z_{nS})$	0.1	0.01	0.01	0.01	0.1	0.05	0.02	0.004

Table 1.3b: Vectors of observed summary statistics of population pairs for the ABC analysis. S_S is the proportion of shared polymorphic sites between two populations

	X			**3**		
	AS-AF	**AS-EU**	**EU-AF**	**AS-AF**	**AS-EU**	**EU-AF**
$\bar{D}_a$	0.95	0.27	0.77	0.65	0.3	0.4
$\bar{S}_S$	3.14	3.32	3.99	2.84	3.74	5.12

Statistical evidence for a single colonization event out of Africa: The results of the model choice procedure showed strong evidence in favor of the SCS model. The Bayes factor (Kass and Raftery 1995) was larger than 100 for both X-linked and autosomal datasets (Figure 1.1). The method also revealed that the SCS model, in which the Asian population undergoes two successive bottlenecks, is associated with higher posterior probabilities for both X-linked and autosomal data, when compared to the SCS-2 model in which the Asian population is bottlenecked only once. The posterior probabilities were 0.99 vs 0.01 for the X-chromosome and 0.9996 vs 0.0004 for chromosome 3. When letting P_{δ}, the proportion of retained simulations in the model choice procedure, vary from 0.025 to 1% we observed that the strength of evidence in favor of the SCS model remained decisive for all values of P_{δ} and both chromosomes (Table 1.4). For the comparison between model SCS and SCS-2 the results also remained stable for both chromosomes. Finally, the results of our analysis of the accuracy of our model choice procedure showed that both the SCS and the DCS model could be correctly identified by the method in 97.9 and 96.8% of the cases, respectively. We thus identified the SCS model as our best demographic model and estimated the marginal posterior distribution of each parameter.

Estimation of the demographic parameters: We summarized the X-linked and the autosomal datasets using eight partial least squares components. Using the larger X-linked dataset we find that the effective population size of the ancestral African population is about 1,837,000 individuals (932,000 , 2,531,000) and that the out-of-Africa migration occurred approximately 16,800 years ago (9,400 , 33,500). The colonization of the Southeast Asian continent occurred about 2,500 years ago (700 , 5,200) and the effective size of the Asian founding population was approximately 8,300 individuals (3000 , 77,000).

Table 1.4: Stability of the posterior probabilities of the different demographic models as a function of the proportion of retained simulations in the model choice procedure

	Chromosome-X			Autosome		
Tolerance	**DCS**	**SCS**	**SCS-1**	**DCS**	**SCS**	**SCS-1**
250	3.12×10^{-24}	0.988	0.012	9.55×10^{-289}	1	1.36×10^{-7}
500	$\mathbf{1.95\times10^{-24}}$	**0.989**	**0.011**	$\mathbf{7.82\times10^{-17}}$	**1**	$\mathbf{3.53\times10^{-4}}$
750	2.16×10^{-69}	0.977	0.023	1.18×10^{-14}	0.998	0.002
1000	2.6×10^{-94}	0.966	0.034	2.93×10^{-14}	0.998	0.002
2500	6.67×10^{-117}	0.854	0.146	7.24×10^{-12}	0.997	0.003
5000	2.82×10^{-37}	0.848	0.152	1.01×10^{-10}	0.994	0.006
7500	3.86×10^{-24}	0.598	0.402	1.62×10^{-10}	0.995	0.005
10000	1.03×10^{-22}	0.644	0.356	6.36×10^{-10}	0.991	0.009

The values in bold correspond to the threshold we used for the model choice estimation.

The corresponding estimates based on the autosomal data are comparable except for the estimate of the time of colonization of the Southeast Asian continent that is larger for chromosome 3: 5000 years (1000 , 11,000). Estimates of the demographic parameters for the African and European populations can also be found in Table 1.5 and graphical representations of all marginal posterior distributions for chromosome X and 3 are in Figure 1.3. These estimates do roughly agree with earlier results that were obtained with a maximum likelihood analysis of the joint mutation frequency spectrum of the African and European population (Li and Stephan 2006). When letting P_δ, the proportion of retained simulations in the parameter estimation procedure, vary from 0.01 to 5% we observed that the mode and the credibility intervals of our posteriors distributions remained stable (Figure 1.4).

Table 1.5: Estimates of the demographic parameters

Parameters	Priors	chromosome-X	R^2	Autosome	R^2
Current African population size	$(10^5, 3\times10^7)$	4,786,360 (2,040,701 , 29,208,295)	0.62	3,134,891 (1,371,066 , 28,013,950)	0.55
Current European population size	$(10^4, 5\times10^6)$	1,632,505 (780,907 , 4,870,580)	0.40	878,506 (383,361 , 4,775,964)	0.3
Current Asian population size	$(10^4, 5\times10^6)$	1,632,505 (780,907 , 4,870,580)	0.18	512,748 (143,082 , 4,542,090)	0.12
European bottleneck population size	$(10, 10^5)$	22,066 (14,338 , 81,102)	0.47	32,128 (15,968 , 95,162)	0.48
Asian bottleneck population size	$(10, 10^5)$	8,279 (2,971, 77,482)	0.36	11,862 (4,255 , 81,044)	0.35
Exit out of Africa	$(10, 10^6)$	16,849 (9,392 , 33,452)	0.63	12,843 (7,095 , 31,773)	0.62
Southeast Asia colonization time	$(10^2$, Exit out of Africa)	2,467 (711 , 5,195)	0.72	5,012 (992 , 11,084)	0.68
African expansion time	$(10^2, 4\times10^5)$	25,553 (1,698 , 376,730)	0.02	37,323 (3,636 , 379,212)	0.01
Ancestral African population size	$(10^4, 2\times10^7)$	1,837,229 (931,637 , 2,530,609)	0.83	1,705,328 (609,393 , 2,458,653)	0.82

The time estimations (i.e. modes and credibility intervals) are provided in years assuming 10 generations per year. Population sizes are given in effective numbers of individuals. R2 is the coefficient of determination.

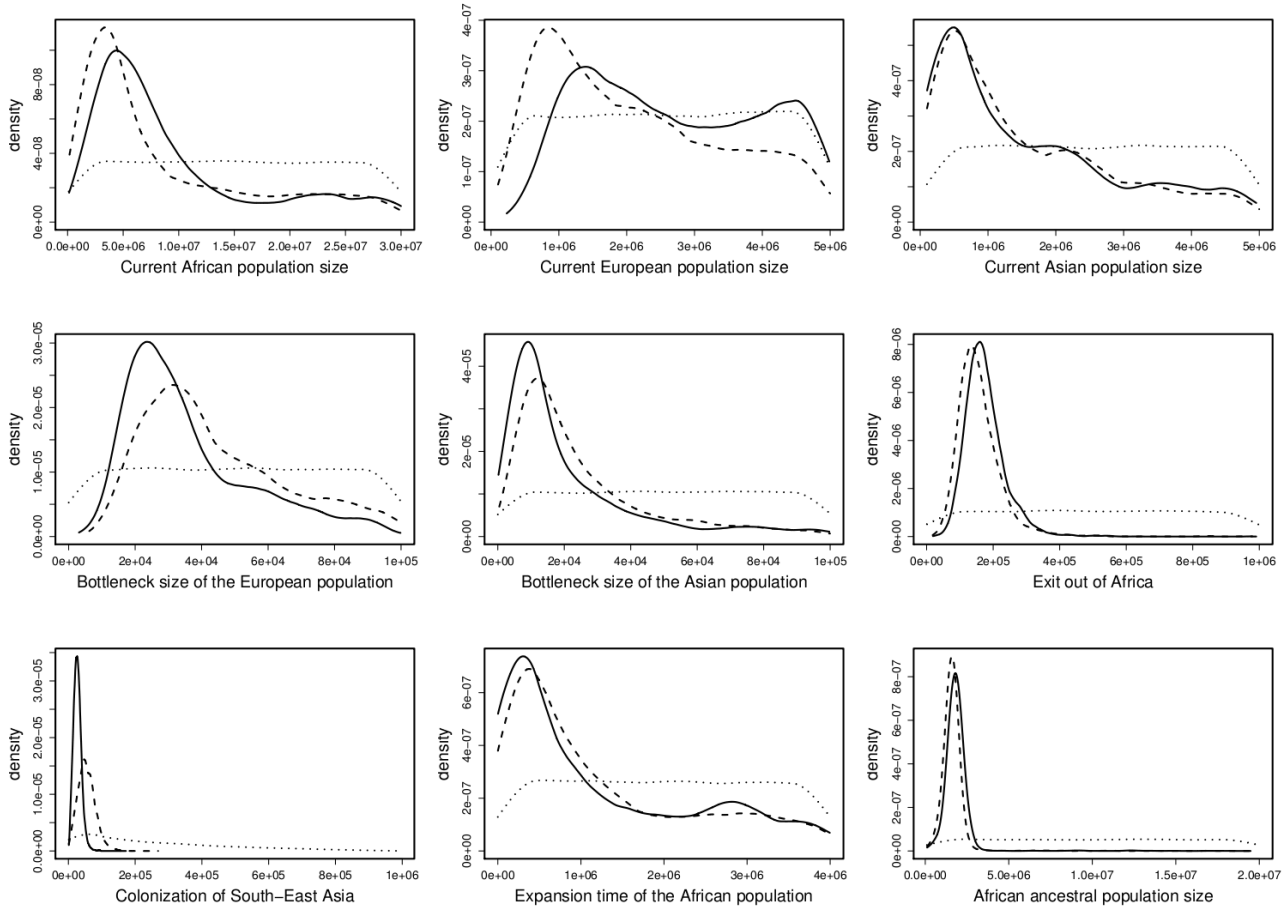


Figure 1.3: Posterior distributions of the parameters of the SCS model inferred by ABC estimation. Dotted lines: Priors; Solid lines: Posteriors for the X-linked dataset; Dashed lines: Posteriors for the autosomal dataset.

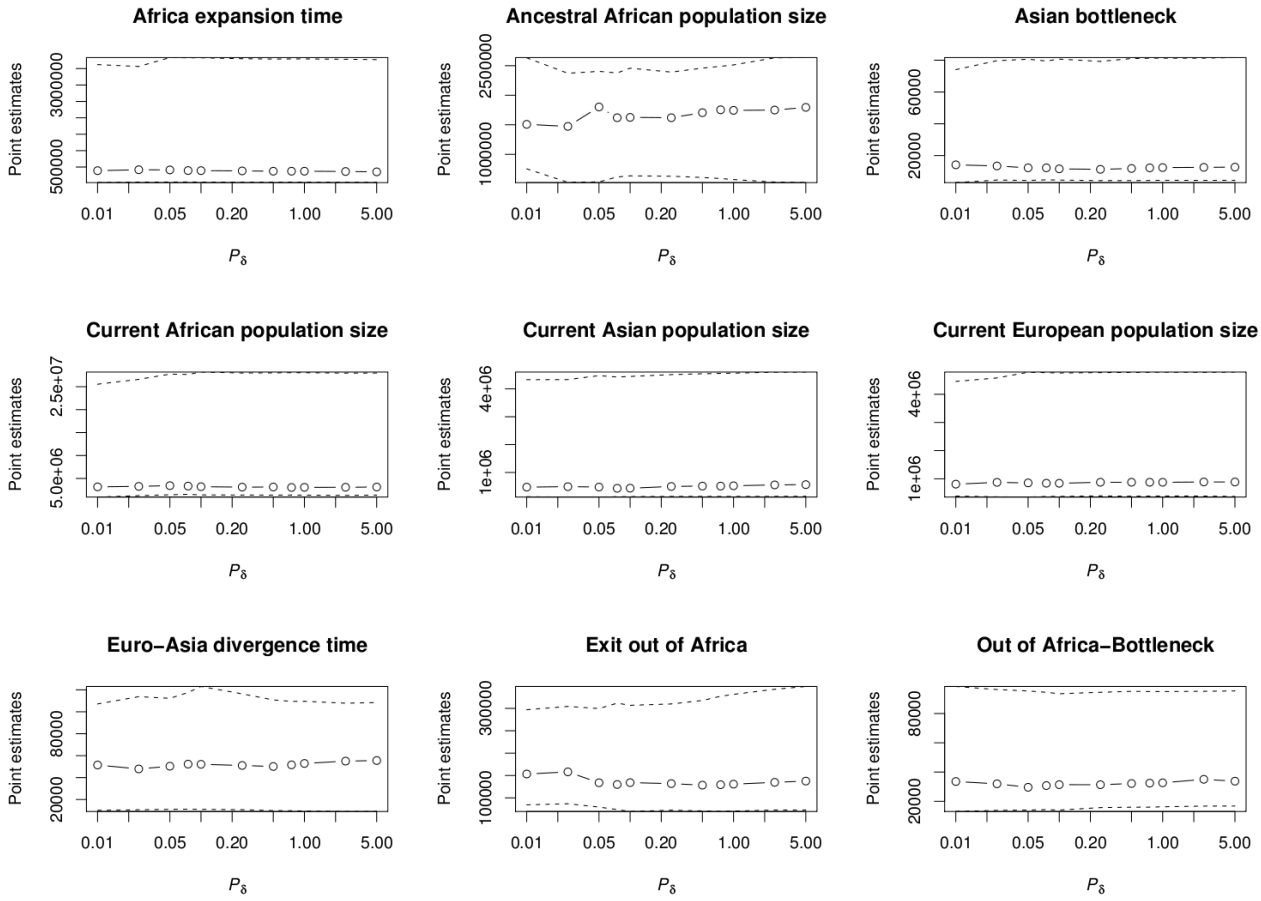


Figure 1.4: Stability of the modes and of the credibility intervals of the posterior distributions of the SCS model, as a function of the proportion of retained simulations in the ABC model choice procedure. Solid lines: Mode; Dashed lines: 95% credibility intervals

Discussion

The results of the performance analysis we carried out on the model choice procedure shows that, for a dataset similar to the one we analyse in this study, the method is able to distinguish between the two demographic scenarios we investigated in this study: the single colonization scenario (SCS), where African and non-African populations have a non-African MRCA, and the double colonization scenario (DCS) where these populations have an African MRCA (Figure 1.1). We also showed that the results of the model choice procedure was stable with regard to variations of the proportion of retained simulations P_δ. These results give credibility to our finding that the SCS provides a better fit to our observed data than the DCS.

The quality of the estimation of each parameter of the SCS can be assessed by the coefficient of determination (R^2). Previous studies have shown that parameters for which R^2 values are smaller than 5-10% are usually difficult to estimate (Neuenschwander *et al.* 2008). For the parameters of our best model, R^2 values are higher than 10% (Table 1.5) except for the parameter representing the time of the African expansion (1% for chromosome X and 2% for chromosome 3). These low values for this parameter could indicate that a simple expansion model is not the best demographic model to explain the patterns of genetic diversity observed in the African population or that the statistics we used to summarize the full dataset (Table 1.3a and 1.3b) did not capture enough information on this parameter. As for the model choice procedure, we showed that our parameter estimates are robust to the proportion of retained simulations P_δ (Figure 1.4).

Migration of human populations towards Southeast Asia started during the last ice period about 55-65 ky ago (Fagundes *et al.* 2007). Due to the current commensal relation between *D. melanogaster* and humans, it could be envisioned that this species had colonized Southeast Asia together with the first human migrants. Such an early colonization scenario has been used once to

explain the strong morphological differentiation of the Asian populations (David *et al.* 1976). However, our present demographic estimations do not support such a scenario. Rather, our results indicate that the Southeast Asian population was founded by flies that diverged from the first non-African populations at a later time. We estimated that the divergence between the African and the derived population occurred about 16,800 ago (9,400 , 33,500) for the chromosome X and about 12,800 years ago (7,100 , 31,800) for chromosome 3. These results are in line with a previous maximum-likelihood analysis of the African and European populations that suggested a date of divergence of 15,800 years (12,000 , 19,000) (Li and Stephan 2006). At this time, rising temperatures following the end of the last ice period could have favored the colonization of northern territories. Early human sedentary settlements during the natufian period (starting 13,000 years ago) in the Fertile Crescent (Bar-Yosef 1998) could already have helped to maintain a relatively large population of *D. melanogaster* in this region.

Our results reject the hypothesis of an ancient migration of *D. melanogaster* to Southeast Asia but they do not provide information about the geographic location of the split between the European and the Southeast Asian populations. Even if it seems likely that migrating *D. melanogaster* populations have entered the region of the Middle East by crossing the Sinai peninsula, additional information, such as genetic polymorphism data from a Middle East population, would be necessary to learn more about the geographic location of the divergence between European and Southeast Asian populations.

Furthermore, we found that the divergence between Asian and European populations occurred about 2500 years ago (700 , 5,200) for the chromosome X and 5000 years ago (990 , 11,000) for chromosome 3. Although favorable climatic conditions were already present before that time, this late colonization of Asia and Europe suggests that *D. melanogaster* depended on the establishment of human populations (and perhaps agriculture) to colonize these new areas. From

these results it appears that a long divergence time, associated with a neutral process, may not explain the observed morphological characteristics of the Asian populations (i.e. smaller number of ovarioles, smaller eggs, and higher fresh weight than the ancestral population), contrary to what has been hypothesized by David *et al*. (1976).

We also estimated the effective size of the founding population that diverged from the ancestral gene pool (22,100 and 32,100 individuals for chromosome X and 3, respectively; Table 1.5) and the size of the founder population that colonized the Southeast Asian region (8,300 and 11,900 individuals for chromosome X and 3, respectively; Table 1.5). These estimates have to be taken with caution since the impact of genetic drift on the history of the derived populations may be much more complex than that produced by a simple population size bottleneck. Recurrent fluctuations in population size (Pool and Nielsen 2008*)* as well as population substructure might also have influenced the way genetic drift shaped the patterns of polymorphism we observe in present populations. Therefore these estimates should be seen as a simple way of summarizing the amount of genetic drift needed to reproduce the skew in the site frequency spectrum we observe in these populations. This model can then be used to correct the rate of false positives of selection detection methods that are based on the site frequency spectrum.

The results of the predictive simulations (Table 1.6) show how well our model is able to reproduce the observed datasets. Unability to reproduce certain summary statistics of the dataset indicates that some aspects of the demographic model are not optimal. The nature of the summary statistics that are not well predicted by our model can also, in some cases, indicate how the model might be improved. Although most summary statistics are correctly predicted by our best model, it fails to account for some aspects of the observed datasets (Table 1.6). Our model predicts higher values for the average, and lower values for the variance of Tajima's *D* compared to observed data on the X- chromosome in the European population. Therefore, it seems that patterns of genetic

variation found on the X-chromosome of European *D. melanogaster* cannot be fully explained by a single population size bottleneck alone. Discrepancies between patterns of genetic variation found at X-linked and autosomal genes in the European population could be due to biased sex-ratios and/or different intensities of selective pressures on the X chromosome (Hutter *et al.* 2007). Also, predicted values of Z_{nS} for the African populations are higher than observed values. This could indicate that this population experienced more complicated size fluctuations than predicted by the expansion model alone or that recombination rates in the African population are higher than what we assume in this study. Interestingly, most aspects of the Asian dataset can be correctly predicted by our model, with the exception of D_A on the third chromosome where predictions are smaller than the observed value.

In conclusion, our study generated a new dataset of X-linked and autosomal nucleotide polymorphism in a Malaysian population of *D. melanogaster*. In addition, we performed a demographic analysis of this data, jointly with one African and one European dataset, by employing an ABC approach. Our results reject the hypothesis that an early migration event could have led to the colonization of Southeast Asia and account for the observed morphological differentiation between African, European and Asian *D. melanogaster*. The statistical model of demographic evolution that we inferred suggests that *D. melanogaster* colonized Eurasia after the Neolithic period, when the rise of agriculture turned small communities of hunter-gatherers into larger sedentary settlements. Since *D. melanogaster* is now a human commensal, it is possible that the adaptive history of this species is mostly characterized by adaptations to post-neolithic human societies rather than by adaptations to the new climatic conditions that the non-African populations encountered. Local adaptation studies that may be based on our demographic analyses will help to answer this question by identifying the genes that were targeted by positive selection during this colonization process.

Table 1.6: Results of the predictive simulations

Statistics	X	3	Statistics	X	3
$(\bar{\pi})_{\text{AF}}$	0.87	0.89	$(\bar{Z}_{\text{nS}})_{\text{EU}}$	0.91	0.62
$Var(\pi)_{\text{AF}}$	0.86	0.84	$Var(Z_{\text{nS}})_{\text{EU}}$	0.01	0.46
$(\bar{S})_{\text{AF}}$	0.85	0.87	$(\bar{\pi})_{\text{AS}}$	0.75	0.69
$Var(S)_{\text{AF}}$	0.77	0.78	$Var(\pi)_{\text{AS}}$	0.90	0.51
$(\bar{D})_{\text{AF}}$	0.21	0.31	$(\bar{S})_{\text{AS}}$	0.80	0.76
$Var(D)_{\text{AF}}$	0.50	0.78	$Var(S)_{\text{AS}}$	0.94	0.62
$(\bar{K})_{\text{AF}}$	0.68	0.68	$(\bar{D})_{\text{AS}}$	0.14	0.25
$Var(K)_{\text{AF}}$	0.16	0.11	$Var(D)_{\text{AS}}$	0.81	0.31
$(\bar{Z}_{\text{nS}})_{\text{AF}}$	0.92	0.96	$(\bar{K})_{\text{AS}}$	0.53	0.63
$Var(Z_{\text{nS}})_{\text{AF}}$	0.93	0.77	$Var(D)_{\text{AS}}$	0.57	0.35
$(\bar{\pi})_{\text{EU}}$	0.70	0.89	$(\bar{Z}_{\text{nS}})_{\text{AS}}$	0.63	0.62
$Var(\pi)_{\text{EU}}$	0.91	0.89	$Var(Z_{\text{nS}})_{\text{AS}}$	0.78	0.76
$(\bar{S})_{\text{EU}}$	0.75	0.93	$(\bar{D}_{\text{a}})_{\text{AF-EU}}$	0.80	0.86
$Var(S)_{\text{EU}}$	0.89	0.87	$(\bar{S}_{S})_{\text{AF-EU}}$	0.75	0.91
$(\bar{D})_{\text{EU}}$	0.10	0.23	$(\bar{D}_{\text{a}})_{\text{AF-AS}}$	0.83	0.97
$Var(D)_{\text{EU}}$	0.99	0.89	$(\bar{S}_{S})_{\text{AF-AS}}$	0.78	0.68
$(\bar{K})_{\text{EU}}$	0.44	0.81	$(\bar{D}_{\text{a}})_{\text{EU-AS}}$	0.87	0.79
$Var(K)_{\text{EU}}$	0.29	0.27	$(\bar{S}_{S})_{\text{EU-AS}}$	0.67	0.68

Values represent the probability that the simulated data is smaller than the observed value

Acknowledgements

We thank the Munich Drosophila group and Matthieu Foll for discussions, two anonymous reviewers for helpful comments on the manuscript, Daniel Wegmann for sharing code, and the Deutsche Forschungsgemeinschaft for funding (grants STE 325/7 and STE 325/12 of the Research Unit 1078). LE was partially supported by a Swiss NSF grant No 3100,126074.

Chapter 2

Using coalescence to investigate seed banks and population structure in wild tomato species

unpublished work

Abstract

The existence of seed banks is of practical importance in conservation biology to diminish the immediate ecological impact of habitat fragmentation and prevent species extinction. From an evolutionary perspective, seed banks increase genetic diversity and are important for buffering the effect of varying climatic conditions. Furthermore, seed dormancy can be seen as a bet-hedging strategy that magnifies the evolutionary effect of good years and dampens the effect of bad years.

In this study we estimate the germination rates for two wild tomato species (*S. chilense* and *S. peruvianum*) found in western South-America in a wide range of habitats using DNA sequences coupled to a coalescent model in combination with ecological data. We use sequences at eight reference loci for a sample of three populations, and a species-wide sample. We develop an Approximate Bayesian Computation framework to integrate ecological information on above ground population census sizes, in order to estimate seed bank and metapopulation parameters for each species. We provide the first evidence that it is possible to disentangle the effect of the metapopulation structure from that of the seed bank on the effective population size, to obtain

accurate estimates of germination rates with coalescent model. *S. chilense* is a specialist species found in dryer areas of Northern-Chile and has an estimated germination rate of 0.14, where *S. peruvianum*, a generalist species found in a wide range of habitats, shows a bigger seed bank due to a significantly smaller germination rate of 0.06. Finally we show that regarding genetic diversity, these species do not experience recent population loss due to human activities.

Introduction

The effective population size (N_e) defines the evolutionary potential of a population because it determines the rate at which adaptive substitutions appear and get fixed (Gossmann *et al.* 2010), as well as the vulnerability to loss of genetic diversity by genetic drift. A key question in plant evolutionary biology, which is also of practical relevance for conservation biology, is to understand how the census size of a population above ground (N_{cs}) is affected by habitat fragmentation and ecological disturbances, and how this process affects in return the effective population size (N_e) (Lande 1988; Espeland and Rice 2010). Habitat loss and fragmentation due to human activities are indeed an acute problem for species conservation, especially in spatially structured plant populations where demes have often small N_e and N_{cs} values. Previous studies have promoted the view that plant species should have lower N_e than N_{cs} (Frankham 1995; Siol *et al*. 2007; Song and Mitchell-Olds 2007; Abe *et al.* 2008; Oddou-Muratorio and Klein 2008). It has been suggested that the N_e of plant species increases with higher metapopulation structure (Amos and Harwood 1998; Charlesworth *et al.* 2003) and longer seed banks (Templeton and Levin 1979; Hairston and Destasio 1988; Nunney 2002; Vitalis *et al.* 2004). However, although most plant species are characterized by metapopulation structure and seed banks simultaneously, the respective contribution of these factors to increasing the effective population size compared to the census size have not been measured so

far on the basis of genetic data. In fact, metapopulation structure is common for many plant species which exist as a spatial collection of numerous demes connected by migration and experiencing recurrent extinction/recolonization events (Pannell and Charlesworth 1999; Wakeley and Aliacar 2001; Pannell 2003).

The first objective of our study is to develop a coalescent model with metapopulation structure and seed bank, and to use this model to perform model-based inference within an Approximate Bayesian Computation (ABC) framework. We combine for the first time ecological data on census sizes with DNA sequence polymorphisms (reflecting the effective size) to estimate the germination rates of two wild tomato species. We introduce here the use of two different spatial samplings of populations for each species, in order to disentangle the effect of the seed bank from that of the metapopulation on N_e. Seed banks, that is, the dormancy of seeds for several generations, are a form of storage of genetic diversity. Theory predicts that lower germination rates increase N_e (Epling *et al.* 1960; Templeton and Levin 1979; Hairston and Destasio 1988; Levin 1990; Ellner and Hairston 1994; Kaj *et al.* 2001; Nunney 2002; Vitalis *et al.* 2004; Siol *et al.* 2007; Honnay *et al.* 2008; Oddou-Muratorio and Klein 2008; Lundemo *et al.* 2009; Ayre *et al.* 2010), although empirical tests of this effect are scarce (Honnay *et al.* 2008; Honnay *et al.* 2009; Lundemo *et al.* 2009). Ecologically, seed banks counter-act habitat fragmentation by buffering against the extinction of small and isolated populations, a phenomenon known as 'temporal rescue effect' (Honnay *et al.* 2008). For example, lower fragmention-driven extinction rates were found for species with long seed banks compared to short-lived seed banks (Stocklin and Fischer 1999). Seed banks evolve as an evolutionary stable strategy in species living in temporally or spatially unpredictable habitats for buffering the effect of varying environment (Evans et al. 2007). It has been suggested theoretically (Cohen 1966; Templeton and Levin 1979; Ellner 1985; Valleriani and Tielborger 2006; Rajon *et al.* 2009) and shown empirically (Evans *et al.* 2007) that adaptation for seed dormancy (variable

germination rates) is a bet-hedging strategy to magnify the evolutionary effect of good years and to dampen the effect of bad years. Bet-hedging is a strategy in which adults release their offspring into several different environments to maximize the chance that some will survive.

Our second objective is to test for the evolution of two different bet-hedging strategies, namely long and short seed banks, in two wild tomato species with different ecological habitats using model-based parameter inference based on DNA polymorphism. We also provide a software (msABC_SB), a modified version of Hudson's ms (Hudson 2002) and msABC (Pavlidis *et al.* 2010b) to perform multi-locus ABC analysis of seed bank and metapopulation parameters.

We study two wild diploid outcrossing tomato species, *Solanum peruvianum* and *S. chilense*, of the family Solanaceae (Peralta *et al.* 2008). These two species have been analysed at the ecological level revealing differences in their habitat (Nakazato *et al.* 2008; Chetelat *et al.* 2009; Nakazato *et al.* 2010): *S. chilense* is a specialist species found in dry to very dry habitats such as the Atacama desert in Northern Chile, and *S. peruvianum* is described as a generalist species found in a wider range of mesic to dry habitats from Central Peru to Northern Chile. These species exist as metapopulations of hundreds of demes, undergoing recurrent extinction/recolonization (Roselius *et al.* 2005; Arunyawat *et al.* 2007; Nakazato *et al.* 2010). It has been hypothesized that these two species differ by their metapopulation structure (number of demes and migration rates) and seed banks because of the geographical structure of the habitat and specific adaptations to different environments (Roselius *et al.* 2005; Arunyawat *et al.* 2007; Nakazato *et al.* 2008; Xia *et al.* 2010).

Materials and Methods

Species: Wild tomatoes form a small monophyletic clade consisting of 13 closely related diploid (2N = 24) species within the family Solanaceae (Peralta *et al.* 2005; Spooner *et al.* 2005; Peralta *et*

al. 2008). All species share a high degree of genomic synteny (Nesbitt and Tanksley 2002) and are intercrossable at various degrees (Rick 1963; Rick *et al.* 1976; Rick 1986). A recent taxonomic revision places the tomato clade within the genus *Solanum* (formerly genus *Lycopersicon*) (Peralta *et al.* 2005; Spooner *et al.* 2005; Peralta *et al.* 2008). In our study we focus on two *Solanum* species: *S. peruvianum* and *S. chilense* as defined in the new taxonomic treatment (Peralta *et al.* 2008). These two species have been studied at the ecological level revealing differences in their habitat (Nakazato *et al.* 2008; Chetelat *et al.* 2009; Nakazato *et al.* 2010). Population genetics studies have focused on revealing the spatial population structure of each species (Arunyawat *et al.* 2007) as well as the degree of genetic divergence and the age of the speciation split (Roselius *et al.* 2005; Städler *et al.* 2005; Städler *et al.* 2008). With regard to the recent adaptive evolution of these species, *S. chilense* has been shown to exhibit positive selection at a gene involved in the ABA pathway supposed to be involved in drought tolerance (Xia *et al.* 2010). *S. peruvianum* has on the other hand been studied for response to biotic stress and coevolution with various parasites such as *Pseudomonas* or the Ascomycete *Cladosporium fulvum*. Evidence for coevolution and balancing selection at resistance genes has been detected in *S. peruvianum* (Rose *et al.* 2005; Rose *et al.* 2007).

Estimates of the deme census sizes: We first estimated the census size of demes for both species, using data from the Tomato Genetics Resource Center (TGRC) (http://tgrc.ucdavis.edu/) at the University of California, Davis, USA. We used the version of the database from March 2010. In total 118 accessions of *S. peruvianum* and 135 of *S. chilense* are reported in the database from sampling in Peru and Chile. This collection is the product of over five decades of field work by C.R. Rick and multiple investigators. They have been considered likely to be a good approximation of current range distributions of these species (Nakazato *et al.* 2010). We extracted the number of plants present in the above-ground population at the moment of sampling for each accession. We obtained

information for 75 *S. peruvianum* and for 107 *S. chilense* accessions. This dataset is defined as the census size of above- ground demes for each species (called N_{cs}). We found that the indicated census sizes of demes varied depending on the investigator and on the size of the populations. Sometimes, only qualitative estimates of census size were given. For example, several demes (around 20 in both species) were referred as being "large", "huge", or "very large", and for populations containing more than 100 plants, the counting is often not precise. For our first calculations, these undefined census sizes were removed from the dataset. Our aim was to obtain the mean of the census size of above-ground populations for each species. For this purpose we assumed that the census size of demes in a metapopulation follows a negative exponential distribution (many demes with small population sizes and few large demes). A negative exponential regression is fitted to the distribution of census sizes for each species using the R software (R Development Core Team 2005). This was realized using the *lm* function in R on the log transform of the distribution of census sizes. The exponential coefficient gives the mean census size per deme for each species.

Ecological data and geographical range of each species: Nakazato *et al.* (2010) have correlated the geographic distribution with ecological habitats of 10 wild tomato species, including *S. peruvianum* and *S. chilense*. This study estimated which environmental variables (mean annual precipitation, mean annual temperature, precipitation seasonality, sun exposure, topsoil pH, effective soil depth) explained each species' distribution. This led the authors to estimate also the range size (in km^2), potential niche breadth, and percentage of niche filling of each species. The range area of *S. peruvianum* was estimated to 80,961 km^2, and the percentage of niche filling as 22.4%. S. chilense shows a smaller range of 62,401 km^2, and the percentage of niche filling was 31.5% (Nakazato *et al.* 2010). We calculated from these results an estimated number of physical (or ecological) demes per species assuming that the total number of observed accessions for *S.*

peruvianum and *S. chilense*, 118 and 135, respectively, fill only 22.4 and 31.5% of their potential niche. We estimated the number of physical demes to be around 526 for *S. peruvianum* and 428 for *S. chilense* (Table 2.1). These numbers of physical (or ecological) demes should not be confounded with the effective number of demes assumed in population genetics model of metapopulation (Wakeley and Aliacar 2001; Charlesworth *et al.* 2003).

Genes sequenced: Seven unlinked nuclear loci are used in this study: CT066, CT093, CT166, CT179, CT198, CT251 and CT268 (Table 2.2). These loci are single-copy cDNA markers originally mapped by Tanksley *et al.* (1992) in genomic regions with different recombination rates (Stephan and Langley 1998). The gene products putatively perform key housekeeping functions, and thus purifying selection is suggested to drive their evolution (Tellier *et al.* 2011). These loci were previously used in population genetic studies of *S. peruvianum* and *S. chilense* (Roselius *et al.* 2005; Städler *et al.* 2005; Arunyawat *et al.* 2007; Städler *et al.* 2008).

Table 2.1: Summary of the key ecological data for the two wild tomato species from the TGRC collection

Species	Total number of accessions in the TGRC database	Number of populations with census size available	% of niche filling[a]	Estimated number of demes in the species range	Estimated mean census size per deme N_{cs}
S. peruvianum	118	75	22.4	526	44 – 84
S. chilense	135	107	31.5	428	33 – 65

[a] data obtained from Nakazato *et al* 2010.

Table 2.2: Chromosome location, putative function, and sizes of coding and non-coding regions of the seven studied loci in *S. peruvianum* and *S.chilense*

Locus	Chromosome	Putative protein function	Non-coding region	Coding region	
				synonymous	non-synonymous
CT066	10	Arginine decarboxylase	0	335	1008
CT093	5	S-adenosylmethionine Decarboxylase proenzyme	359	263	765
CT166	2	Ferredoxin-NADP reductase	823	118	322
CT179	3	Tonoplast intrinsic protein D-type	234	174	404
CT198	9	Submergence induced protein 2-like	359	90	242
CT251	2	At5g37260 gene	348	348	974
CT268	1	Receptor-like protein kinase	0	404	1476

The number of sites in each category was estimated with the method of Yang and Nielsen (2000) and is based on the alignment of sequences for the pooled populations in *S. peruvianum.*

Sampling scheme: The sequence data at the seven reference loci (CT066, CT093, CT166, CT179, CT198, CT251, CT268) obtained by Arunyawat *et al.* (2007) and Städler *et al. (*2008) for three populations of *S. peruvianum* (Tarapaca, Nazca, Canta) and *S. chilense* (Tacna, Moquegua, Quicacha) are referred to as the "local" sample (Table 2.3). We follow Städler *et al.* (2009) in calling the set of sequences from all three populations together the "pooled" sample. We generated a new set of sequences, called the "species-wide" sample, at these same seven loci. It is composed of 14 and 10 accessions from the TGRC database distributed all over the range of *S. peruvianum* and *S. chilense*, respectively, each accession being sequenced for one allele. These three types of samples reflect different parts of the genealogical history of the metapopulation (Städler *et al.* 2009; Chikhi *et al.* 2010). The local sample reflects the past history of the deme, as well as local selective events such as purifying selection (Wakeley and Aliacar 2001; Pannell 2003, Tellier *et al.* 2011). This short coalescent phase within demes is named the scattering phase (Wakeley and Aliacar 2001). The

major part of the metapopulation coalescent tree is not contained in the scattering phase, but in the long collecting phase that describes the coalescences across demes (Wakeley and Aliacar 2001). Species-wide events such as species' size fluctuations are reflected in the collecting phase. It has been recently shown that the best sampling strategy to reveal the collecting phase of a metapopulation is to analyse the species-wide sample, that is, one individual from each of many different demes scattered all over the range of the metapopulation (Städler *et al.* 2009; Chikhi *et al.* 2010). The pooled sample represents an intermediary between the local and the species-wide schemes. Depending on the migration rates and number of demes, it can reflect small or large parts of the collecting phase (Städler *et al.* 2009).

Plant material and sequences for the population sample: The population sample is composed of the sequences previously obtained by Arunyawat *et al.* (2007) and Städler *et al.* (2008). For *S. peruvianum* we used data from two populations (Nazca and Canta) collected by T. Städler in Peru in 2005 and one additional accession from the TGRC database (Tarapaca LA2744) from Chile. For *S. chilense*, we used the three populations (Tacna, Moquegua and Quicacha) collected by T. Städler in Peru in 2005. For each population, five to six individuals were collected, and the seven loci were sequenced, resulting in a total of 10 to 12 sequences per population. Note that we do not used the Arequipa (*S. peruvianum*) and the Antofogasta (*S. chilense*) populations studied by Arunyawat *et al. (*2007) and Städler *et al. (*2008), because it was shown, on the basis of the frequency-spectrum of alleles, that these populations experienced some peculiar demographic events (most likely bottlenecks or admixture) (Arunyawat *et al.* 2007; Städler *et al.* 2008).

Table 2.3: List of the population samples of the two studied Solanum species

Species	Population	Location	Coordinates (latitude, longitude)
S. peruvianum	Tarapaca (LA2744)	Northern Chile	18°33'S, 70°09'W
	Nazca	Southern Peru	14°51'S, 74°44'W
	Canta	Central Peru	11°31'S, 76°41'W
S. chilense	Tacna	Southern Peru	17°53'S, 70°07'W
	Moquegua	Southern Peru	17°04'S, 70°52'W
	Quicacha	Southern Peru	15°37'S, 73°48'W

Where applicable, the TGRC accession numbers are indicated. S. chilense and S. peruvianum populations have been described in Arunyawat *et al*. (2007).

Plant material and sequences for the species-wide sample: We used here the so- called "species-wide sample", for which we selected one plant per 14 accessions of *S. peruvianum* and 10 accessions of *S. chilense* from the TGRC. These populations were chosen to be distributed uniformly over the range of both species (Table 2.4). One allele for each of the seven loci was sequenced per plant of the species-wide sample. Genomic DNA was extracted from tomato leaves using the DNeasy Plant Mini Kit (Qiagen GmbH, Hilden, Germany). PCR primers and PCR conditions followed those of the previous studies of the same loci in *S. peruvianum* and *S. chilense* (Arunyawat *et al*. 2007); PCR primer information can be accessed at http://evol.bio.lmu.de/downloads/index.html. PCR amplification was performed with High Fidelity Phusion Polymerase (Finnzymes, Espoo, Finland), and all PCR products were examined with 1% agarose gel electrophoresis. Generally, direct sequencing was performed on PCR products to identify homozygotes and obtain their corresponding sequences. For heterozygotes, a dual approach of both cloning before sequencing and direct sequencing was used to obtain the sequences of both alleles. We developed a series of allele-specific sequencing primers whose 3'-end was anchored on identified SNPs or indels (for details of this approach, see Städler *et al.* (2005).

The first allele to be present in at least four clones was chosen. Sequencing reactions were run on an ABI 3730 DNA analyser (Applied Biosystems and HITACHI, Foster City, USA). One allele was sequenced for each individual, and a total of 14 (*S. peruvianum*) and 10 (*S. chilense*) sequences were obtained for each locus − species combination. Contigs of each locus were first built and edited using the Sequencher program (Gene Codes, Ann Arbor, USA) and adjusted manually in MacClade 4 (version 4.0 for OS X). These new sequences will be deposited in GenBank (accession numbers XXX).

Table 2.4: List of the species-wide sample with the TGRC accession numbers from the two Solanum species

Species	Population	Location	Coordinates (latitude, longitude)
S. chilense	LA1930	Southern Peru	15°17'S, 74°36'W
	LA1960	Southern Peru	17°05'S, 70°52'W
	LA1958	Southern Peru	17°15'S, 71°15'W
	LA1969	Southern Peru	17°32'S, 70°02'W
	LA3355	Southern Peru	18°03'S, 70°18'W
	LA2778	Northern Chile	18°23'S, 69°33'W
	LA2932	Northern Chile	20°29'S, 70°10'W
	LA2748	Northern Chile	21°12'S, 69°30'W
	LA2750	Northern Chile	22°05'S, 70°12'W
	LA2930	Central Chile	25°24'S, 70°24'W
S. peruvianum	LA0153	Central Peru	09°57'S, 78°13'W
	LA0111	Central Peru	10°48'S, 77°44'W
	LA1616	Central Peru	12°05'S, 76°55'W
	LA1913	Central Peru	14°23'S, 75°12'W
	LA2834	Central Peru	14°46'S, 74°49'W
	LA0446	Southern Peru	15°47'S, 74°23'W
	LA1336	Southern Peru	16°12'S, 73°37'W
	LA1951	Southern Peru	16°25'S, 73°08'W
	LA1333	Southern Peru	16°34'S, 72°38'W
	LA3218	Southern Peru	16°57'S, 72°05'W
	LA1954	Southern Peru	17°01'S, 72°05'W
	LA2964	Southern Peru	17°59'S, 70°50'W
	LA4125	Southern Peru	19°18'S, 69°25'W
	LA2732	Southern Peru	19°24'S, 69°36'W

Population genetics analysis of the sequence data: For both species, summary statistics were computed for each locus and each population, for the pooled sample of three populations and for the species-wide sample. Genetic diversity was summarized as the number of segregating sites (*S*) and the population mutation rate θ_w per locus (Watterson 1975). The site frequency spectrum, was summarized by Tajima's *D* (Tajima 1989). The F_{ST} statistic measures fixation of alleles in subdivided populations and was calculated for the population sample (Hudson *et al.* 1992). In the light of the recent criticisms by Jost (2008), F_{ST} was used as a measure of migration between demes integrating drift, and not as a population differentiation index. All statistics were computed using DnaSP v5.1 for all sites, silent and synonymous sites (Rozas *et al.* 2003). Gaps and multiple hits were excluded in DnaSP. Similar values for the observed data were obtained when using the libsequence C++ library (Thornton 2003).

Population genetics modelling: Compared to the classical Wright-Fisher model of coalescence theory (Kingman 1982), seed banks can be seen as a departure from random mating, since there is separation of individuals into different age classes (overlapping of generations). We used the coalescent model proposed by Kaj et al. (2001) for seed banks. Intuitively, seed banks should slow down the rate of coalescence because of the structure introduced in the pool of ancestors (Templeton and Levin 1979). This is similar to the effect of geographically structured populations on coalescence (Kaj and Lascoux 1999; Wakeley and Aliacar 2001): two lineages will “migrate” among generations before they meet in the same ancestor plant above ground where coalescence can occur. We summarized here the theory on coalescence with seed bank. Kaj et al. (2001) modelled a neutral seed bank with haploid Wright-Fisher type dynamics for a single population with constant size. The

population of plants is composed at each generation of N individuals, with a proportion b_i, $i = 1, \ldots, m$, coming from seeds produced i generations ago. In other words, seeds are allowed to remain in the seed bank for up to m generations. At a given generation, each individual is drawn from a pool of seeds build up during the previous m generations. Each individual is obtained with the probability b_1 from the seeds produced at the previous generation, b_2 from the seeds produced two generations ago, …, and b_m from the seeds produced m generations ago. Kaj *et al.* (2001) have calculated the rate of coalescence in a population to be

$$\beta_1^2 \binom{r}{2} \quad (1)$$

with r being the number of ancestral lineages at any point in time, and β_1 the seed bank rescaling rate

$$\beta_1 = 1 / \sum_{i=1}^{m} i b_i \quad (2)$$

where $\sum_{i=1}^{m} i b_i$ is the expected value of the seed bank age distribution. Similarly, they derive the mutation rate γ along an ancestral line in the coalescent as

$$\gamma = \frac{\beta_1}{2}(b_1\theta_1 + b_2\theta_2 + \ldots + b_m\theta_m) \quad (3)$$

where θ_j is the population mutation rate for individuals produced by seeds of age j. (j=1, …, m) (Kaj *et al.* 2001).

We made here further biological assumptions to derive the rate of coalescence, the mutation and recombination rates per population, as well as to rescale the migration rate among demes in a metapopulation:

1) We assumed that seed germination is a memoryless process modelled as a geometric

process in time. We supposed that the germination rate of a given seed is *b*. Each individual is obtained thus with the probability $b_i = b(1-b)^{i-1}$ from the seeds produced at generation *i*.

For clarity, this means that each individual is obtained with the probability:

- $b_1 = b$ from the seeds produced at the previous generation
- $b_2 = b(1-b)$ from the seeds produced two generations ago, …,
- $b_m = b(1-b)^{m-1}$ from the seeds produced *m* generations ago.

We modified Hudson's ms to introduce the rate of coalescence with seed bank with geometric germination rates (from eq. 2):

$$\beta_1 = 1/\sum_{i=1}^{m} ib(1-b)^{i-1} \quad . \tag{4}$$

Eq. 4 can be approximated as $\beta_1 \simeq \frac{b(1-(1-b)^{m+1})}{1-(1+bm)(1-b)^m}$ if *m* is sufficiently large.

The rate of coalescence implemented in our program is thus $\beta_1^2\left(\frac{r}{2}\right)$ using β_1 from eq. 4.

2) We enforced the condition from (Kaj *et al.* 2001) that the sum of the germination probabilities over *m* generations should be equal to 1. This condition is computed in our program. We calculate the mutation rate under a seed bank model, assuming that the mutation rate does not depend on the age of seeds. The population mutation rate with seed bank is (from eq. 3):

$$\gamma = \frac{\beta_1}{2}\theta(b + b(1-b) + \ldots + b(1-b)^{m-1}) = \frac{\beta_1}{2}\theta \quad , \tag{5}$$

where θ_j is the population mutation rate without seed bank (based on *N*).

It has been suggested that aging of seeds can lead to an increase of the mutation rate, most of the new mutations being deleterious (Levin 1990). However, a recent meta-analysis did not reveal high

levels of genetic diversity accumulating in the soil seed bank (Honnay *et al.* 2008). Reviewing different plant species, the authors did not find evidence for genetic differences between the standing crop and the seed bank. When such differences were found, they are likely to be the result of local selection acting as a filter on the alleles present in the seed bank (Honnay *et al.* 2008). In our study, we analysed only neutral processes. We thus chose to keep the mutation rate constant for any seed age assuming no selection process acts within demes on seeds in the bank.

3) We multiplied the recombination rate per nucleotide r also by β_1 (eq. 4). This is because recombination only occurs in a lineage when a plant is above ground and produces seeds.

4) We rescaled the migration rate (κ) between demes in a metapopulation when there is a seed bank. We assumed here that only pollen migrates between demes, and that this occurs only when plants are above ground. Lineages can thus only migrate between demes by pollen. The migration rate (κ) was thus multiplied as well by β_1 . (Note that if no rescaling is done, seeds and pollen are assumed to migrate.)

5) A key assumption in this model is to assume that every generation, the number of individuals (*N)* in Kaj *et al.* (2001) is equal to N_{cs} in each deme. In other words, each generation above ground and each generation in the seed bank has the same census size equal to N_{cs}. The census size used in our model was calculated above from ecological data. This approximation holds as long as the variation between years in census sizes is reasonable (Nunney 2002).

When there is no seed bank, that is, when all seeds germinate the year after being produced, then $b = 1$. In this case, we verified that equations 4 and 5 define a Wright-Fisher model. Similarly, these equations are in line with findings from Nunney (2002) and Vitalis *et al. (*2004) describing the expected heterozygosity in a population with seed bank. Table 2.5 lists all the parameters of our coalescent models with seed banks.

Parametrization of the demographic models: We modelled the metapopulation as an island model with n_d demes, with migration occurring among demes (Pannell and Charlesworth 1999; Wakeley and Aliacar 2001; Pannell 2003; Wakeley and Takahashi 2004). We did not take into account extinction/recolonization as we did not have data to estimate such parameters. However, we assumed that each species was composed of a set of many demes which allowed us to apply the many-demes model from Wakeley and Aliacar (2001) and Wakeley and Takahashi (2004). The effective migration rate κ is the proportion of individuals in a given deme which come from other demes by migration. In an island model, all demes contribute equally to the migrant pool. Our metapopulation had a current census size of $S_{current} = N_{cs} \times n_d$. To model the demography of each species, we assumed that the $n_d = 200$ demes join into a single panmictic population of size S_{anc} at time t_{event} in the past. The time was measured here in years assuming one generation per year. Three demographic scenarios for the whole metapopulation were considered: 1) an expansion of the species if the ratio $S_{anc} / S_{current} < 1$, 2) constant species size if $S_{anc} / S_{current} = 1$, and 3) a species crashes if $S_{anc} / S_{current} > 1$. We modelled expansion up to 100-fold and crashes up to 10-fold.

Each deme has a seed bank defined by a germination rate b, which is the probability of a seed to germinate at a given generation, and the maximum time m that this seeds can spend in the seed bank (m is fixed to 40). Each deme was assumed to have a census population size of N_{cs} as estimated from ecological data. To compensate for the uncertainty in estimating the census size and the fact that we may underestimate N_{cs}, we assumed large priors for these values between 50 and 1,000. Each mutation happened at rate μ per site per generation.

Table 2.5: List of parameters and compound parameters in the model

	Parameter name	Parameter definition	Range of possible values
Estimated parameters	N_{cs}	Census size of each deme in the metapopulation	[1 - ∞[we use averages calculated from the ecological data
	b	Germination rate	[0 – 1]
	m	Maximum time seeds can spend in the seed bank (in generations)	[1 - ∞[we use 40
	κ	Migration rate between demes (without seed bank rescaling)	[0 – 1]
	n_d	Number of demes in the metapopulation (effective number)	[1 - ∞[we use values ranging from 200 to 1,000
	μ	Mutation rate per nucleotide	we use [10^{-9} – 10^{-8}]
	t_{event}	Time of the population split in generations	
	S_{anc}	Size of the ancestral population	smaller, equal or larger than current metapopulation
Compound parameters (non estimated)	β_1	Rescaling parameters of the seed bank (eq. 4)	
	θ	Population mutation rate without seed bank per deme	$\theta = 4 N_c \mu$
	γ	Population mutation rate with seed banks (eq. 5)	
	r	Local crossing-over rates per nucleotide per generation (obtained for each locus from Stephan and Langley (1998) without seed bank	

Having found the best demographic model for each species, we tested for the necessity of a seed bank for explaining the huge genetic diversity observed in these species. Using the model choice procedure of the ABC (see below), we compared the models with seed bank to two alternative scenarios (list of parameter prior values in Table 2.6a and 2.6b).

First, in a model without seed bank, we tested the hypothesis that the large observed diversity could be a relic of a large ancestral population, submitted to a population crash and habitat fragmentation. In this scenario, an ancestral single pancreatic population (with size up to 25 times $S_{current}$) splits into 200 demes at time t_{event} in the past. We kept here a small census size per deme N_{cs}, the values of which are in a prior distribution between 50 and 1,000 indiviiduals.

Second, we tested a model without seed bank, but in which no prior knowledge was assumed about the census size of populations (N_{cs}). The prior for N_{cs} was fixed between 50 and 25,000, the values used in Städler *et al.* (2009) for the effective size of each deme.

Finally, we compared several models to test for an optimal number of demes per species, assuming the best demographic model per species and a seed bank (b varied here from 0.01 to 1). The number of demes varied between 200 and 1,000 for *S. peruvianum*, and between 200 and 600 for *S. chilense* assuming that the maximum number of demes can be twice as high as the numbers suggested by the ecological data. Based on the work of Nakazato *et al.* (2010), we assumed that the number of ecological demes should be around 500 for *S. peruvianum* and 400 for *S. chilense* (Table 2.1).

Table 2.6a: Summary of prior boundaries of the ABC chosen for each tested model in *S. peruvianum*

Model	Parameters	Min	Max
Seedbank + constant population size			
	N_{cs}	50	1000
	μ	5×10^{-9}	10^{-8}
	b	0.01	0.5
	$\log(\kappa)$	-5	-3
	t_{event}	0	200
Seedbank + expansion			
	N_{cs}	50	1000
	μ	5×10^{-9}	10^{-8}
	b	0.01	0.5
	$\log(\kappa)$	-5	-3
	t_{event}	0	200
	S_{anc} (demes)	2	200
Seedbank + crash			
	N_{cs}	50	1000
	μ	5×10^{-9}	10^{-8}
	b	0.01	0.5
	$\log(\kappa)$	-5	-3
	t_{event}	0	200
	S_{anc} (demes)	200	2000
No seedbak + expansion			
	N_{cs}	50	25000
	μ	5×10^{-9}	10^{-8}
	$\log(\kappa)$	-5	-3
	t_{event}	0	200
	S_{anc} (demes)	2	200
No seedbank + crash			
	N_{cs}	50	1000
	μ	5×10^{-9}	10^{-8}
	$\log(\kappa)$	-5	-3
	t_{event}	0	200
	S_{anc} (demes)	200	5000

Table 2.6b: Summary of prior boundaries of the ABC chosen for each tested model in *S. chilense*

Model	Parameters	Min	Max
Seedbank + constant population size			
	N_{cs}	50	1000
	μ	5×10^{-9}	10^{-8}
	b	0.01	1
	$\log(\kappa)$	-4	-2
	t_{event}	0	200
Seedbank + expansion			
	N_{cs}	50	1000
	μ	5×10^{-9}	10^{-8}
	b	0.01	1
	$\log(\kappa)$	-4	-2
	t_{event}	0	200
	S_{anc} (demes)	2	200
Seedbank + crash			
	N_{cs}	50	1000
	μ	5×10^{-9}	10^{-8}
	b	0.01	1
	$\log(\kappa)$	-4	-2
	t_{event}	0	200
	S_{anc} (demes)	200	2000
No seedbak + expansion			
	N_{cs}	50	25000
	μ	5×10^{-9}	10^{-8}
	$\log(\kappa)$	-4	-2
	t_{event}	0	200
	S_{anc} (demes)	2	200
No seedbank + crash			
	N_{cs}	50	1000
	μ	5×10^{-9}	10^{-8}
	$\log(\kappa)$	-4	-2
	t_{event}	0	200
	S_{anc} (demes)	200	5000

Approximate Bayesian inference: To estimate the posterior probabilities of different demographic models and posterior distributions of the model parameters, we took an ABC approach. The method relies on the comparison of a vector of summary statistics computed on the observed data, Δ_{obs}, with those computed on a large number of simulated datasets for which the parameters of interest are known, Δ_{sim}. Here we implemented our ABC algorithm following Excoffier *et al*. (2005). The algorithm used to estimate the parameters of a model is composed of three steps: a simulation step, a rejection step, and an estimation step. The simulation step consisted in simulating, for every evolutionary scenario, between 400,000 and one million datasets identical to our observed dataset in terms of numbers of loci and sample sizes. Every evolutionary scenario was defined by a set of parameters (germination rate, mutation rate, ancestral population size, age of demographic events (Table 2.5) and every parameter was characterized by a prior distribution (Tables 2.6a and 2.6b). For each evolutionary scenario we sampled from the prior distribution and used the randomly picked parameter values to perform coalescent-based simulations. The way in which the rejection and the estimation steps have been applied to these simulated datasets differed for the model-choice and the parameter-estimation procedures and are described below.

To simulate these datasets we incorporated available external information about the mutation rate (μ) and the local crossing-over rates (r) per nucleotide. The mutation rate was set to 5.1×10^{-9} following Städler *et al.* (2005). Since such divergence-based estimates of the mutation rate can potentially be underestimated due to the long-term effect of purifying selection we modelled this uncertainty. We define a prior distribution on μ with lower and upper boundaries equal to 5×10^{-9} and 10^{-8}. The upper bound of the distribution was based on previous results concerning the intensity of purifying selection acting on this set of genes (Tellier *et al.* 2011). Similarly, external information about the local recombination rates (r) was used to generate our simulated datasets. Recombination

rates are given as the local rates of crossing-over per site per generation, and were taken from Stephan and Langley (1998). The coalescent simulations were performed using our modified version of the program ms as described above (Hudson 2002; Pavlidis *et al.* 2010b). We developed our own code to sample parameter values from prior distributions and to compute summary statistics on the simulated data efficiently. For this we used the GSL C++ library and the libsequence C++ library (Thornton 2003). Simulations were launched on a 64-bit Linux cluster with 510 nodes. Source code is available upon request.

Choice of summary statistics: To summarize our datasets we computed the following statistics for both species: 1) the average value of θ_w across loci and populations (Watterson 1975), 2) the average Tajima's *D* across loci and populations (Tajima 1989), 3) the average F_{ST} value across populations (Hudson *et al.* 1992), 4) θ_{w_SW}, the average θ_w across loci at the species-wide level, and 5) D_{sw}, the average Tajima's *D* across loci at the species-wide level. The rationale behind this choice is as follows.

Statistics computed at the population level are informative on the recent demographic history of the species. For example $\theta_{w,}$ *D* and F_{ST} are expected to correlate with the germination rate and metapopulation structure, changes in population sizes and the migration rates, respectively.

On the other hand, the species-wide statistic D_{sw} is intended to be rather affected by demographic events affecting the metapopulation as a whole (the collecting phase (Wakeley and Aliacar 2001)) such as ancestral population size changes.

The model choice procedure: The posterior probabilities of different demographic models can be estimated on the basis of the Euclidean distance, δ, between the observed summarized dataset and the summarized datasets simulated under all models. The inference procedure consists in retaining

only simulations for which δ is smallest. Pritchard *et al.* (1999) proposed that the posterior probability of a model can be approximated by the proportion of retained simulations under that model, relative to the number of retained simulations under all models. Beaumont (2008) proposed an improvement of the method correcting for the fact that retained simulations never exactly match the observed data. The method is based on a weighted multinomial logistic regression procedure, where the response variable is the indicator of the model and the predictor variables are the summary statistics (Fagundes *et al.* 2007; Beaumont 2008). Here we simulated 400,000 datasets for each of the investigated demographic models and summarized them into the set of summary statistics, $\triangle_{sim}$, described in the previous section. For each demographic model we retained the 500 simulated datasets associated with the smallest δ and applied the method of Beaumont (2008) to estimate posterior probabilities. Bayes factor were calculated as the ratio of these posterior probabilities.

Parameters estimation: To analyse more specifically the demographic models associated with the highest posterior probabilities for *S. peruvianum* and *S. chilense* we estimated the posterior distributions of their parameters. Therefore, for each species, we simulated 1 million datasets under the best model and summarized them into the same set of summary statistics described above. In this case, however, we applied to the simulated and observed sets of statistics a partial least-square (PLS) transformation (Boulesteix *et al.* 2007) that reduces the uninformative signal from the dataset and breaks down the correlations between the different summary statistics. The use of PLS transformation in ABC estimations (Wegmann *et al.* 2009) has been proven to be efficient because of the high dimensionality of datasets and the frequent correlations between summary statistics. Note that correlations between statistics violate the condition of application of the ABC estimation procedure (Beaumont *et al.* 2002). Here we constructed the PLS latent components with 10,000

simulated datasets under the best demographic model using the code available in the ABCtoolbox package (Wegmann *et al.* 2010).

Choosing the best number of partial least square components for parameter estimations has been done by investigating the decrease of the root mean square error (RMSE) for every parameter as a function of the number of PLS components. The RMSE indicates the percentage of variation unexplained by the PLS components and is constructed by comparing the simulated parameter values with the ones predicted using a given number of PLS components (Wegmann *et al.* 2010). We chose the number of components for the parameter estimation procedure such that additional components do not decrease the RMSE of any parameter of the model. The retained PLS components were used to transform the observed and the simulated datasets. The rejection step consisted in computing δ between simulated and observed sets of summary statistics and to retain the 5,000 simulations closest to the observed data based on their value of δ. Finally, we estimated posterior distributions of the parameters by applying the locally weighted multivariate regression method of Beaumont *et al.* (2002) implemented in the abcEst program (Excoffier *et al*. 2005). We estimated the marginal posterior probability distribution of each demographic parameter using the kernel density estimation method implemented in the R core package and reported the mode and the 95% credibility intervals of these distributions. To avoid the posterior distributions to exceed the upper and lower bound of our prior distributions we transformed the data as $z = \log[\tan(1/x)]$ as described in Hamilton *et al.* (2005), where x is the original dataset and z is the transformed data.

Results

Estimates of the deme census sizes: We estimate the mean census size of a deme to be 44 for *S. peruvianum* and 33 for *S. chilense* (Figure 2.1). To test the robustness of these estimations, we assigned different values of census size values to the 20 accessions referred as "large", "huge" or "very large". Assuming values between 200 and 600 for these accessions (probable overestimates), the means of the deme census size were increased to 84 for *S. peruvianum* and 65 for *S. chilense* (Table 2.1). (The R-square coefficients for the exponential regression were lower than above but above 0.9.) This uncertainty on N_{cs} is taken into account when defining the priors in the ABC method.

Population genetics analyses: Table 2.7, 2.8, and 2.9 contain the summary statistics of the genetic polymorphism patterns that we observed in the species-wide, the population and the pooled samples, respectively. In bold are the values of the vector of observed data used for inference in the ABC method. We compared the genetic diversity (θ_w) and shape of the coalescent tree (Tajima's *D*) between the species-wide and the pooled samples. The rationale is that difference in the shape of the coalescent tree between these sampling schemes reflects the degree of population structure and past demographic events (Städler *et al.* 2009; Chikhi *et al.* 2010). The species-wide sample allowed us to study the collecting phase of the whole metapopulation (species) coalescent (Pannell and Charlesworth 1999; Wakeley and Aliacar 2001; Pannell 2003). On the other hand, the pooled sample may reflect only partially the collecting phase of the metapopulation coalescent because of the rapid coalescences occurring within demes.

Interestingly, a theoretical study has shown that in neutral models of metapopulation

(stepping stone or island model), the pooled sample is expected to have a higher value of Tajima's *D* than the species-wide sample (Figures 1 and 2 in Städler *et al.* 2009). We observe this trend at synonymous sites in both species (Figure 2.2).

Note, however, that for silent sites and all sites, the pooled sample shows a lower value of Tajima's *D* than the species-wide sample (except for *S. chilense* at silent sites, Figure 2.2). This effect is due to the action of purifying selection which is stronger on the pooled sample than on the species-wide sample. In fact, purifying selection acting at the population level decreases the effective population size of each deme. Purifying selection increases the amount of drift per deme, and thus results in a higher number of private alleles per deme than under neutral evolution (Charlesworth *et al.* 1993; Charlesworth *et al.* 1997; Whitlock 2003). For the pooled sample, purifying selection creates thus a more negative Tajima's *D* than under neutrality because private deleterious mutations to demes create an excess of low frequency variants in the pooled sample compared to neutrality. To a lesser extent the species wide sample is also affected by purifying selection (smaller variation in Tajima's *D* between the different categories of sites).

In addition to a previous study (Tellier *et al.* 2011), the analysis of the shape of the coalescent tree revealed that purifying selection occurs in each species creating a spurious excess of low-frequency variants in the pooled and the species-wide sample datasets. We decided therefore to use Tajima's *D* values from synonymous sites to estimate the neutral processes governing the evolution of each species and population.

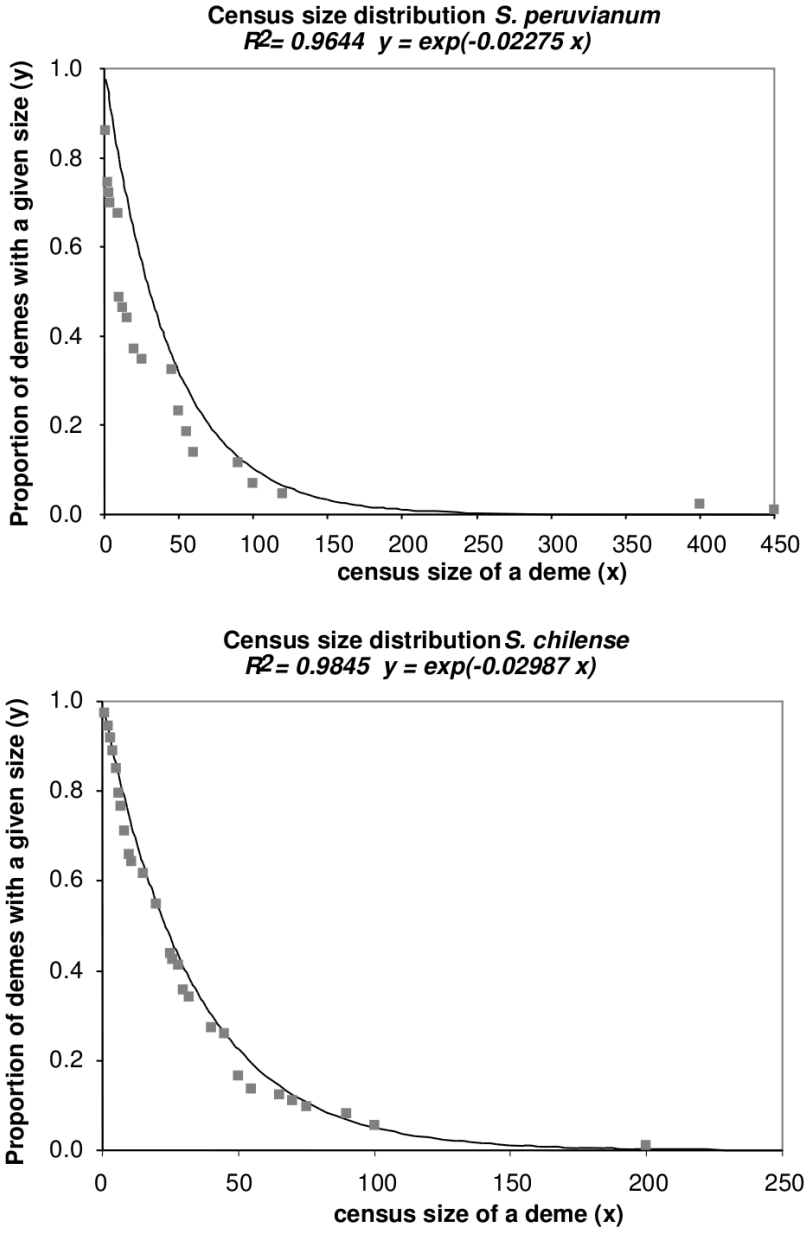


Figure 2.1: Exponential regression for the census size of demes for *S. peruvianum* and *S. chilense*. The coefficient of regression and the equation of the best fitting regression are indicated.

Table 2.7a: Summary statistics at seven loci for the species-wide sample for *S. peruvianum*

Locus	Number of segregating sites S_{sw}	Population mutation rate θ_{W_SW}	Tajima's *D* at all sites D_{SW}	Tajima's *D* at silent sites D_{silent_SW}	Tajima's *D* at synonymous sites D_{syn_SW}
CT066	58	18.24	−1.04	−1.12	−1.12
CT093	39	12.26	−1.62	−1.42	−1.25
CT166	59	18.55	−1.6	−1.55	−1.72
CT179	54	16.98	−0.58	−0.63	−0.82
CT198	59	18.55	−0.61	−0.69	−0.75
CT251	94	29.56	−0.87	−0.79	−0.83
CT268	83	26.1	−0.79	−0.18	−0.18
average across loci[a]	63.71	**20.04**	−1.02	−0.91	**−0.95**

[a] arithmetic average across loci.

Table 2.7b: Summary statistics at seven loci for the species-wide sample for *S. chilense*

Locus	Number of segregating sites S_{sw}	Population mutation rate θ_{W_SW}	Tajima's *D* at all sites D_{SW}	Tajima's *D* at silent sites D_{silent_SW}	Tajima's *D* at synonymous sites D_{syn_SW}
CT066	43	15.2	0.064	0.44	0.44
CT093	21	7.42	−1.26	−1.05	−0.64
CT166	48	16.97	−0.39	−0.32	−0.67
CT179	39	13.79	−0.72	−0.72	−0.24
CT198	25	8.84	−1.02	−1.02	0.02
CT251	24	8.48	−0.34	−0.53	−0.39
CT268	50	17.67	0.005	0.29	0.29
average across loci[a]	35.71	**12.62**	−0.52	−0.41	**−0.17**

[a] arithmetic average across loci.

Table 2.8a: Summary statistics at seven loci for the population samples for *S. peruvianum*

Locus	Population mutation rate for population Tarapaca θ_{W_TAR}	Population mutation rate for population Nazca θ_{W_NAZ}	Population mutation rate for population Canta θ_{W_CAN}	Tajima's *D* at all sites for population Tarapaca D_{TAR}	Tajima's *D* at all sites for population Nazca D_{NAZ}	Tajima's *D* at all sites for population Canta D_{CAN}	Fixation index among populations F_{ST}
CT066	13.43	7.29	13.58	−0.34	0.49	−1.03	0.21
CT093	8.13	7.62	9.93	−0.14	−0.74	−1.07	0.11
CT166	14.27	13.25	22.52	−0.5	−1.16	−0.95	0.07
CT179	6.36	12.58	14.24	−0.1	−0.51	−0.9	0.19
CT198	18.73	14.14	16.97	−0.02	−0.02	−0.23	0.09
CT251	22.98	16.89	22.62	−0.17	0.18	−0.33	0.14
CT268	17.67	22.19	20.53	−0.54	0.04	−0.42	0.12
average across loci[a]	**14.51**	**13.42**	**17.2**	**−0.26**	**−0.25**	**−0.71**	**0.13**

[a] arithmetic average across loci.

Table 2.8b: Summary statistics at seven loci for the population samples for *S. chilense*

Locus	Population mutation rate for population Moquegua θ_{W_MOQ}	Population mutation rate for population Tacna θ_{W_TAC}	Population mutation rate for population Quicacha θ_{W_QUI}	Tajima's *D* at all sites for population Moquegua D_{MOQ}	Tajima's *D* at all sites for population Tacna D_{TAC}	Tajima's *D* at all sites for population Quicacha D_{QUI}	Fixation index among populations F_{ST}
CT066	13.08	10.93	10.93	0.14	0.50	0.46	0.00
CT093	7.42	4.64	9.12	−0.71	0.91	−1.55	0.23
CT166	19.44	13.91	8.61	0.31	−0.05	0.20	0.23
CT179	11.67	9.54	9.19	0.09	0.42	0.07	0.18
CT198	7.07	7.07	13.08	−0.46	−1.36	0.48	0.13
CT251	18.03	21.52	22.33	0.44	−0.02	0.81	0.12
CT268	15.91	15.56	13.21	−0.11	0.05	0.41	0.27
average across loci[a]	**13.23**	**11.88**	**12.35**	**−0.04**	**0.06**	**0.13**	**0.17**

[a] arithmetic average across loci.

Table 2.9a: Summary statistics at seven loci for the pooled sample for *S. peruvianum*

Locus	Number of segregating sites S_{pooled}	Population mutation rate θ_{W_pooled}	Tajima's *D* at all sites D_{pooled}	Tajima's *D* at silent sites D_{silent_pooled}	Tajima's *D* at synonymous sites D_{syn_pooled}
CT066	63	15.41	−0.61	−0.48	−0.61
CT093	60	14.67	−1.71	−1.71	−1.56
CT166	108	26.82	−1.62	−1.61	−1.94
CT179	72	17.61	−1.11	−1.02	−0.87
CT198	88	22.21	−0.73	−0.71	0.02
CT251	122	30.29	−0.88	−0.72	−0.38
CT268	121	29.59	−1.01	−0.79	−1.01
average across loci[a]	90.57	22.37	−1.09	−1.01	**−0.91**

[a] arithmetic average across loci.

Table 2.9b: Summary statistics at seven loci for the pooled sample for *S. chilense*

Locus	Number of segregating sites S_{pooled}	Population mutation rate θ_{W_pooled}	Tajima's *D* at all sites D_{pooled}	Tajima's *D* at silent sites D_{silent_pooled}	Tajima's *D* at synonymous sites D_{syn_pooled}
CT066	64	15.65	−0.74	−0.42	−0.42
CT093	49	11.82	−1.03	−0.87	−0.19
CT166	74	18.1	−0.22	−0.11	−0.34
CT179	61	15.4	−0.71	−0.66	0.003
CT198	51	12.87	−0.98	−1.02	−0.78
CT251	107	25.8	−0.07	0.33	1.023
CT268	84	20.26	−0.12	0.41	0.41
average across loci[a]	70	17.13	−0.55	−0.34	**−0.04**

[a] arithmetic average across loci.

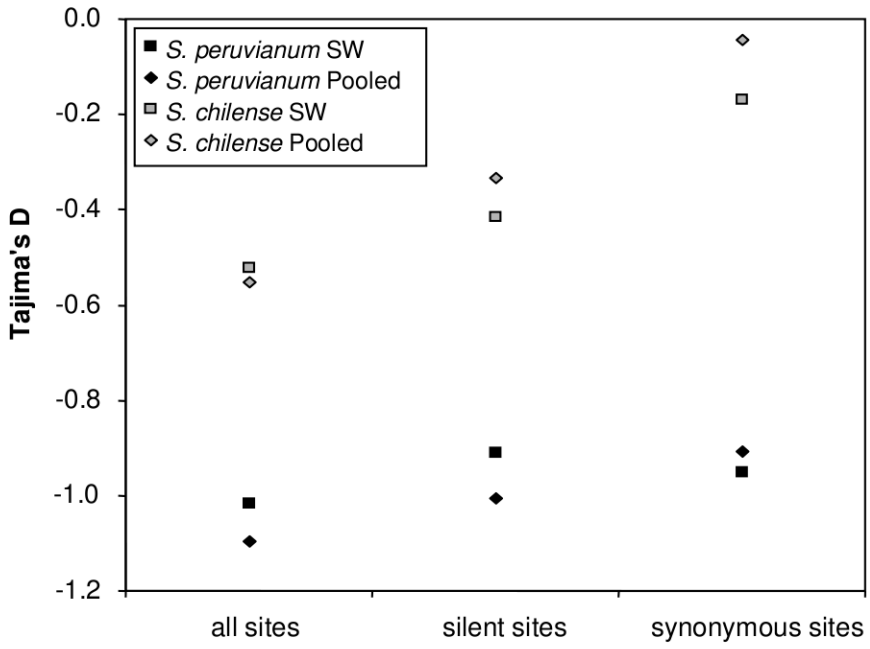


Figure 2.2: Mean Tajima's *D* values across seven loci for all sites, silent, and synonymous sites for both species: S. peruvianum in black, and S. chilense in grey. The rectangles indicate the values of Tajima's D for the species-wide sample, and the diamonds for the pooled sample (pooling of the three populations per species).

Model choice comparisons: Among models with seed bank, a demographic expansion was favored for *S. peruvianum* (Figure 2.3a), and the two models of constant population size and expansion could not be distinguished for *S. chilense* (Figure 2.3b; Bayes factor of 1.09 (Kass and Raftery 1995)). Therefore, we chose to model *S. chilense* with a constant species size in time (though fusion of demes occurs at time t_{event} in the past) because this model has one parameter less than the expansion model. The crash model with seed bank was not favored for either species.

As expected, the species-wide sample data reflects the species demography, as found for the expansion being assessed by the negative Tajima's *D* in *S. peruvianum* (Table 2.7a; Städler *et al.* 2009). However, the barely negative Tajima's *D* in *S. chilense* does not indicate strong species expansion, contrary to expectations from Städler *et al.* (2009). In fact, the expectation of an

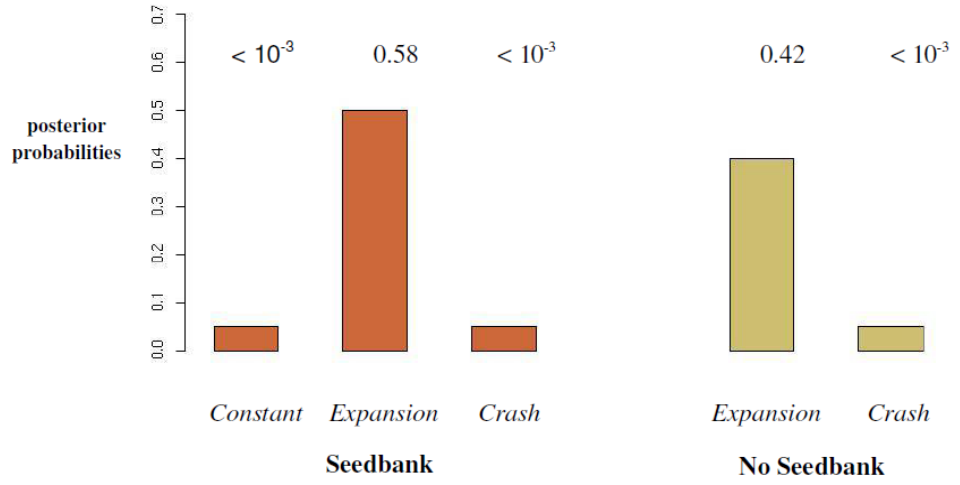
expansion in *S. chilense* was made based on analysis of the pooled sample Tajima's *D* at silent or all sites (Städler *et al.* 2009). Our data on the species-wide sample at synonymous sites indicates, however, that the expansion was small or too old to be detected (Figure 2.2), and that intronic and non-synonymous sites are under strong purifying selection in *S. chilense*.

The first alternative scenario that we compared to the seed bank model was a model without seed bank but with a large ancestral population size. This model was clearly rejected by the model choice procedure in both species (Figure 2.3a and 2.3b). We observed in our simulations that it was possible to generate high genetic diversity (θ_w) in all samples, but that Tajima's *D* values for the species-wide sample and population sample did not fitted our observed data.

The second alternative model was parametrized without a seed bank and with a large prior on $N_{cs} \in$ [50; 25000]. This model cannot be distinguished from the best model with seed bank (Figure 2.3). In *S. peruvianum,* the model with expansion with seed bank is slightly better than the expansion model without seed bank, though not significantly (Figure 2.3a; Bayes factor = 1.38). Similarly in *S. chilense*, the model with constant population size with and without seed bank have similar posterior probabilities (Figure 2.3b; Bayes factor = 1.5). However, when we analysed the posterior distribution of the census size per deme (N_{cs}), in the model without seed bank, we found the mode to be 5.624 [3,520 – 19,670] for *S. peruvianum* and 1,482 [784 – 2,565] for *S. chilense* (in parenthesis the boundaries of the 95% high density probability interval).

We thus estimated that without seed bank, each deme of the metapopulation would present an above-ground census size of around 5,000 for *S. peruvianum* and 1,500 for *S. chilense*, which is significantly outside the range of plausible values collected from the TGRC database (maximum value was 450 in one deme). We conclude that the existence of seed bank is required in both species to explain the large N_e. Note that these results were obtained with large priors on the mutation rate, μ, and the census size of demes, N_{cs}.

Solanum peruvianum



Solanum chilense

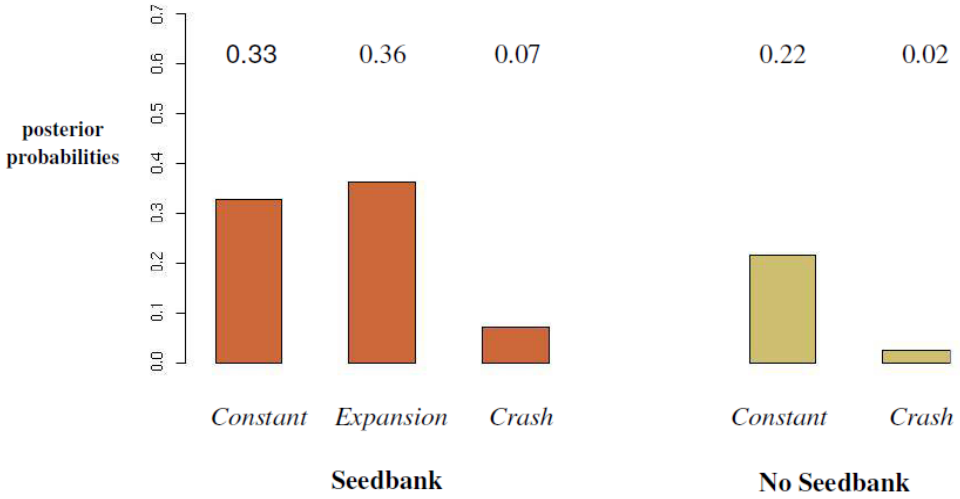


Figure 2.3: Statistical evaluation of alternative evolutionary scenarios. The posterior probabilities for each model are given. A) S. peruvianum: Three types of past demographic event (constant population size, expansion or crash) are tested with seed bank, and two (expansion and crash) without seed bank. B) S. chilense: Three types of past demographic event (constant population size, expansion or crash) are tested with seed bank, and two (constant and crash) without seed bank.

Finally we compared a series of demographic models, similar to the best demographic models for each species, that only differed in their number of effective demes. The aim was to estimate the number of demes (n_d) per species. For *S. chilense* the highest posterior probability was found for a model with 400 demes (Figure 2.4). We assumed that a Bayes factor of less than three in favor of a model does not indicate that this model fits significantly better the observed data. We calculated the Bayes factor for pairs of consecutive numbers of demes, and validated the model for which the Bayes factor did not increased any further (by a factor 3 or more). Using the same procedure we found that *S. peruvianum* was best characterized by a model with 600 demes, though higher numbers were also found showing high probabilities (Figure 2.4). A clear conclusion is that small numbers of demes do not explain the variability and the F_{ST} values found in the observed data. As expected from the known range of both species (Table 2.1), *S. peruvianum* has a higher number of effective demes than *S. chilense*. It is shown theoretically that a higher number of demes or a lower migration rate contribute both equally to increasing N_e of a metapopulation (Wang and Caballero 1999; Wakeley and Aliacar 2001). We show here that with the combination of several summary statistics (θ_w for the three types of sampling schemes and the F_{ST} index) in an ABC framework with meaningful ecological parameters, it is possible to disentangle the effect of the number of demes from that of migration. This is because we estimate the N_e of the metapopulation (in the species-wide sample), but also compute the genetic diversity and fixation index of alleles between demes, values which are dictated by processes shaping the scattering phase of the coalescent in a metapopulation.

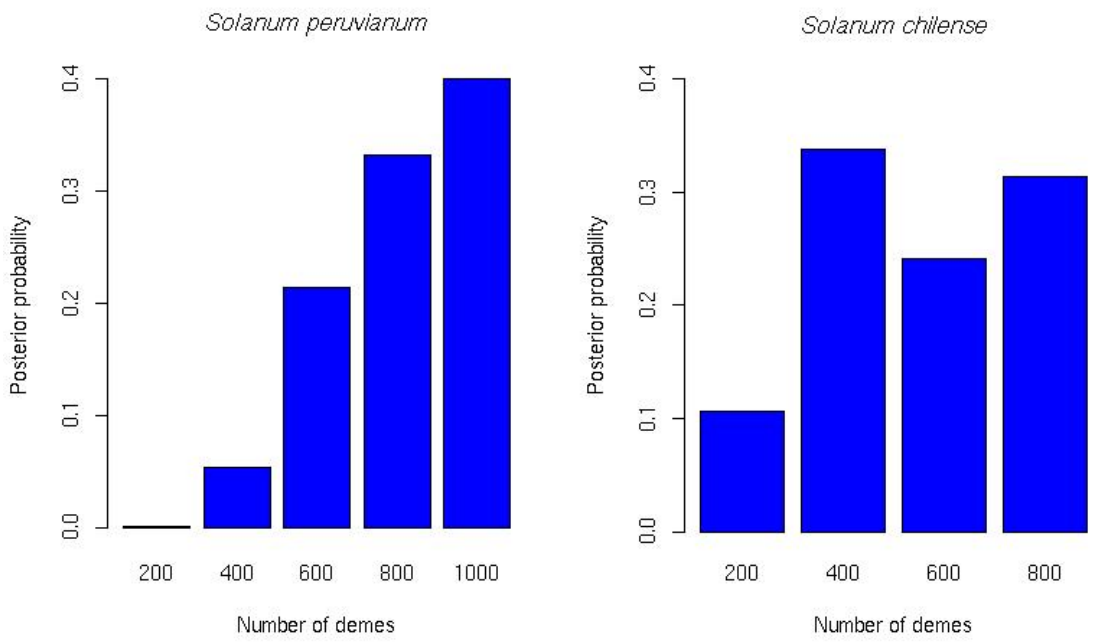


Figure 2.4: Estimate of the effective number of demes per species.

Difference in germination rate between species: Finally, using the model with the optimal number of demes and best demographic scenario for each species, we ran 1,000,000 simulations to estimate its parameters. We compared the posterior estimates for the parameters of the seed bank (*b*), metapopulation structure (κ), and demographic events (time of expansion t_{event}, expansion factor). *S. peruvianum* is estimated to have a longer seed bank, that is, a lower germination rate (*b*), than *S. chilense* (Figure 2.5). On average, seeds would spend 17 years (roughly 1/*b* (Nunney 2002)) in the seed bank for *S. peruvianum* and seven years in *S. chilense* (Table 2.10). We also document the posterior density distributions for the nuisance parameters N_{cs} and μ (Table 2.10, Figure 2.6).

Table 2.10: Summary of the prior and posterior distributions of each parameter

Species	Parameter	Prior		Posterior		
		Lower bound	Upper bound	Mode	HPD 0.025	HPD 0.975
S. peruvianum	Germination rate (b)	0.01	1	0.06	0.01	0.13
	Migration rate (κ)	10^{-4}	10^{-2}	1.15×10^{-3}	5.37×10^{-4}	2.4×10^{-3}
	Time of expansion (t_{event})	0	8×10^{5}	4.66×10^{5}	1.11×10^{5}	7.32×10^{5}
	Expansion ratio	0.01	1	0.118	0.017	0.275
	Census size per deme (N_{cs})	50	200	180.37	80.27	198.09
	Mutation rate (μ)	5×10^{-9}	10^{-8}	8.71×10^{-9}	5.32×10^{-9}	9.93×10^{-9}
S. chilense	Germination rate (b)	0.01	1	0.14	0.04	0.27
	Migration rate (κ)	10^{-4}	10^{-2}	1.66×10^{-3}	4.57×10^{-4}	6.92×10^{-3}
	Time of split (t_{event})	0	8×10^{5}	3.21×10^{5}	2.65×10^{4}	7.81×10^{5}
	Census size per deme (N_{cs})	50	200	180.55	56.77	196.81
	Mutation rate (μ)	5×10^{-9}	10^{-8}	9.28×10^{-9}	5.21×10^{-9}	9.91×10^{-9}

The prior distributions are uniform between the lower and upper bound. The posterior distributions are summarized as the mode and the boundaries of the 95% high probability density interval (HPD 0.025 and HPD 0.975).A rough estimate of divergence time between these species is 550,000 years (Städler *et al*. 2008). The times are calculated assuming a census size per deme of 40 individuals and germination rate of 1/ (0.2^2).

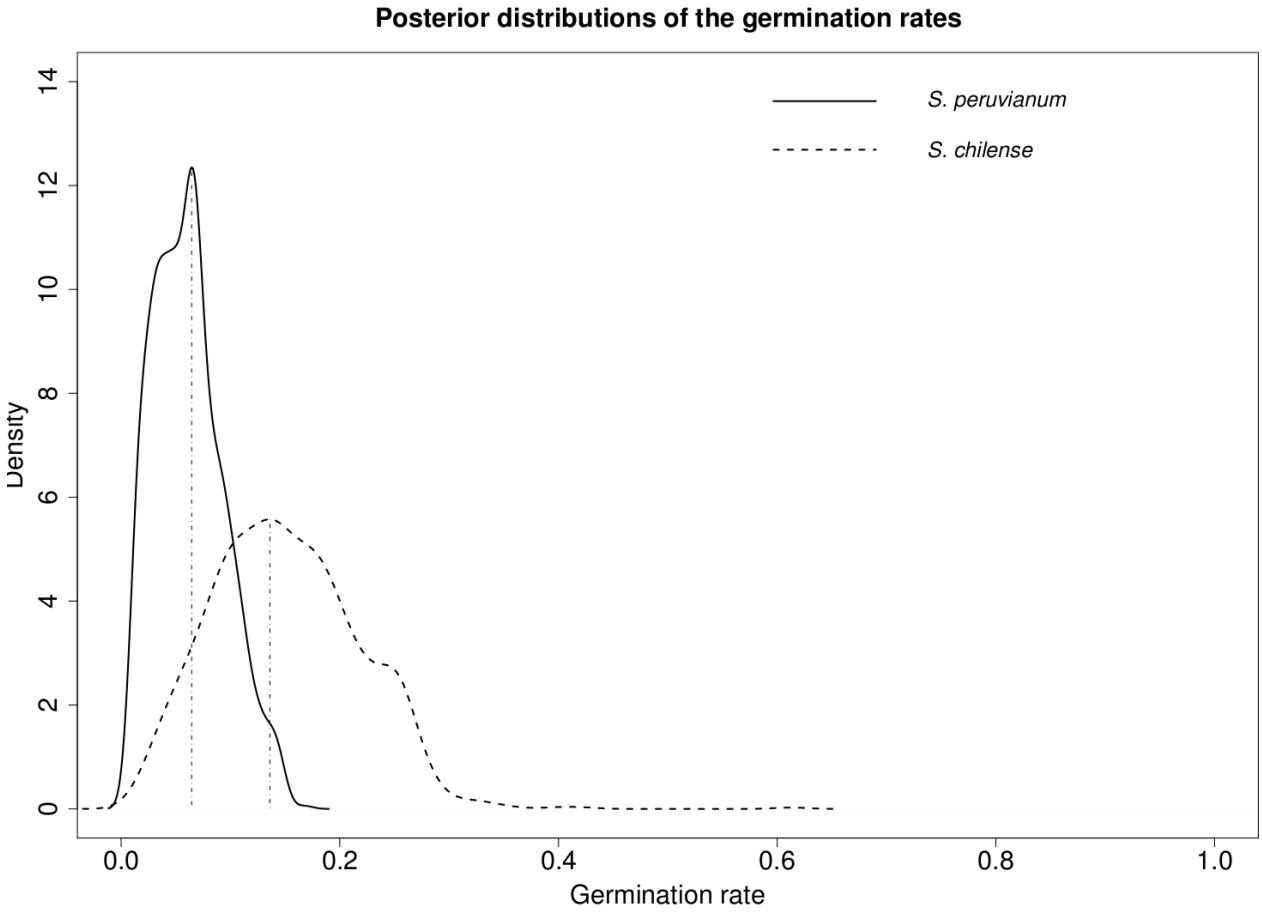


Figure 2.5: Posterior distributions of the germination rate (b) for each species. These curves represent the posterior densities of the germination rate (b) obtained under the best demographic model by the Approximate Bayesian Computation analysis for S. peruvianum (dashed) and S. chilense (dotted).

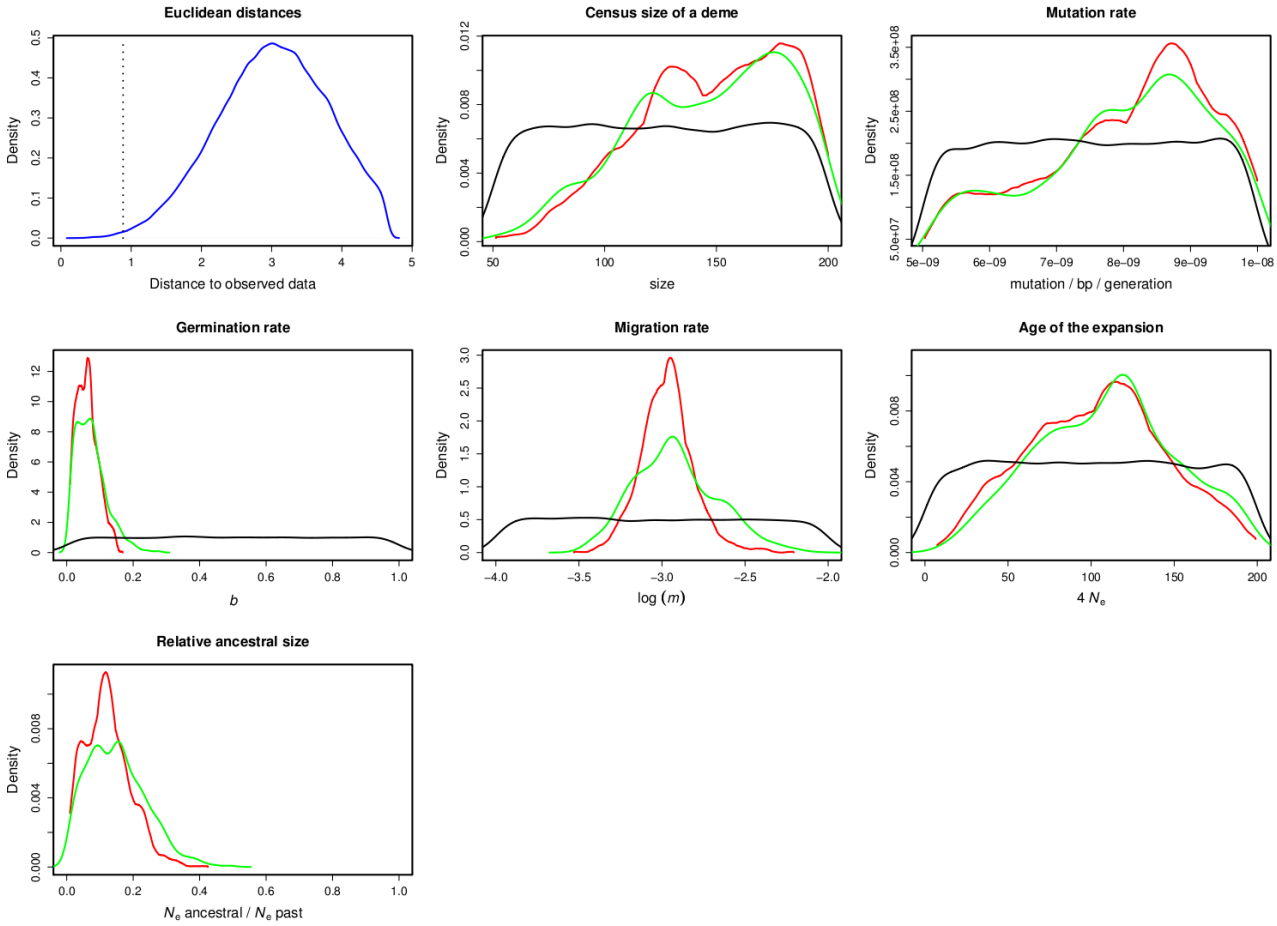


Figure 2.6a: Posterior distributions of the parameters of an island model with demographic expansion for S. peruvianum. Blue line: Distribution of euclidean distances. The dotted line indicates the proportion of retained simulations. Black line: Prior distribution; Green line: Posterior distribution based on the rejection algorithm; Red line: Posterior distribution after the regression adjustment.

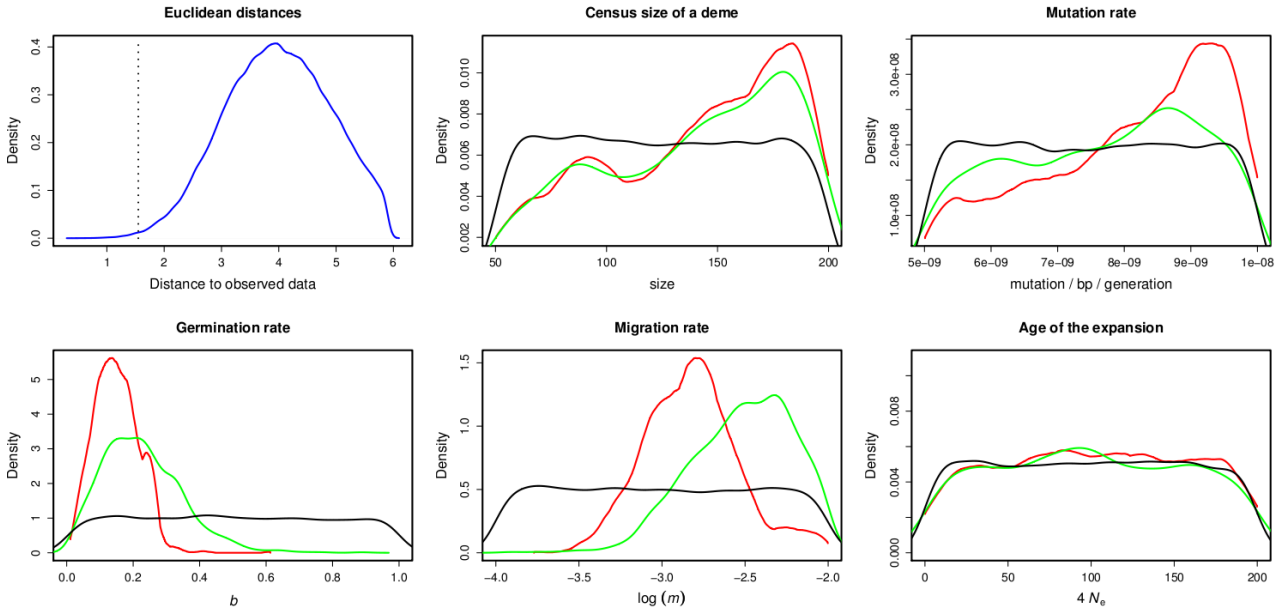


Figure 2.6b: Posterior distributions of the parameters of an island model with demographic expansion for S. chilense. Blue line: Distribution of euclidean distances. The dotted line indicates the proportion of retained simulations. Black line: Prior distribution; Green line: Posterior distribution based on the rejection algorithm; Red line: Posterior distribution after the regression adjustment.

Discussion

We suggest here two hypotheses to explain our results. First, dormancy measured as the germination rate per species (*b*) is a bet-hedging strategy evolving to counter-act the environmental stochasticity over many years (Cohen 1966; Templeton and Levin 1979; Evans and Dennehy 2005). *S. peruvianum* has a longer dormancy, which reflects genetic adaptation to a more unstable environment compared to *S. chilense* (Bentsink et al. 2010) . *S. chilense* is a specialist species found in a small range of dry to very dry habitats in Southern Peru and Northern Chile (Nakazato *et al.* 2008; Nakazato *et al.* 2010; Xia *et al.* 2010). We suggest that although the habitat is unfavourable, the climatic conditions may not vary between years, as it is the case in the habitat of *S.peruvianum*. Unfortunately, the variability of climatic conditions between years was not included in the ecological study of Nakazato *et al.* (2010). On the other hand, *S. peruvianum* is a generalist species found in a wide variety of habitats ranging from coastal plains with mesic environment (lomas) to high altitudes in the Andes and dry habitats (Nakazato *et al.* 2008; Nakazato *et al.* 2010; Xia *et al.* 2010). We thus suggest that longer seed dormancy in this species compared to *S. chilense* can be due to adaptation to habitats that are variable in space or in time. All habitats where *S. peruvianum* is present may thus be more variable in time than the desertic habitats of *S. chilense*. Variability at the spatial level between demes with different environmental characteristics could also promote a longer seed dormancy in *S. peruvianum* as a generalist strategy to colonize (and reduce risks of extinction) in new demes/habitats (Rajon *et al.* 2009).

Second, the difference in germination rates may not reflect different bet-hedging strategies of the two species, but rather environmental influence on germination for each species (Fenner and Thompson 2004; Jurado and Flores 2005). The west coast of South America is affected by the El Nino/Southern Oscillation (ENSO), influencing the size of plant populations. The frequency and

strength of the ENSO before historical times are difficult to establish (Devries 1987; Tudhope and Collins 2003), but it is plausible that the current arid climate has shaped the vegetation of coastal western South America for long period of time (Gregory-Wodzicki 2000; Hartley 2003). We suggest that the germination rates we observe in the two species result from different ENSO occurrences in the location of the two species. Difference in ENSO strength and occurrence may promote the germination of seeds at different time intervals for *S. peruvianum* and *S. chilense*. The two species could also exhibit two different adaptive germination rates, so that seeds germinate in phase with the occurrence of ENSO. Contrary to the bet-hedging hypothesis (Evans and Dennehy 2005), Jurado and Flores (2005) found that dormant species can constitute a large proportion of species in environments with frost, and/or drought. This hypothesis thus states that dormancy is a selective predictable mechanism for the co-occurrence of germination with suitable environmental conditions, as during ENSO events.

Fluctuations in population sizes are likely to be common in wild tomato habitats. The recurrent occurrence of the ENSO phenomenon would affect not only populations of adult plants (Levine *et al.* 2008) but also the replenishment of the seed bank (Gutierrez and Meserve 2003). The germination rate we infer is thus an estimate of the harmonic mean of census population sizes over time, with the buffering effect of the seed bank (Nunney 2002).

Acknowledgements

We thank Hilde Lainer for excellent technical assistance. This research has been supported by grants STE 325/7 and STE 325/12 from the German Research Foundation to SL and WS. AT and PP acknowledge support from the Volkswagen Foundation: post doctoral grant I/82752 to AT, and doctoral grant I/82770 to PP.

Chapter 3

msABC: A modification of Hudson's ms to facilitate multi-locus ABC analysis

Pavlidis P*, Laurent S*, Stephan W. 2010. msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Mol Ecol Resour* 10:723–727 (**contributed equally*)

Abstract

With the availability of whole-genome sequence data biologists are able to test hypotheses regarding the demography of populations. Furthermore, the advancement of the Approximate Bayesian Computation (ABC) methodology allows the demographic inference to be performed in a simple framework using summary statistics. We present here msABC, a coalescent-based software that facilitates the simulation of multi-locus data, suitable for an ABC analysis. msABC is based on Hudson's ms algorithm, which is used extensively for simulating neutral demographic histories of populations. The flexibility of the original algorithm has been extended so that sample size may vary among loci, missing data can be incorporated in simulations and calculations, and a multitude of summary statistics for single or multiple populations is generated. The source code of msABC is available at http://bio.lmu.de/~pavlidis/msabc or upon request from the authors.

Introduction

Along with the increase of population genomic datasets, an important goal is to understand the relationship between patterns of nucleotide polymorphism in natural populations and their evolutionary history. Statistical methods have been employed to estimate demographic parameters using likelihood approaches (Hey and Nielsen 2007) or analysing summary statistics of the data. Among them, ABC benefits from the increase of both available data and computer power (Beaumont *et al*. 2002; Excoffier *et al*. 2005). ABC is applied widely in population genetics studies and usually consists in a two-step procedure. First, simulations are used to sample from the joint distribution of parameters and summary statistics of the simulated data for a given demographic model. Then, a rejection algorithm is applied to retain only values of parameters that generate summary statistics which are similar to the observed values. The retained set of parameter values is then corrected by local linear regression (Beaumont *et al.* 2002) or non-linear regression (Blum and Francois 2010) and considered as an approximation of the posterior distribution. Hudson's (2002) ms is a widely-used coalescent software that generates neutral polymorphism data for a genomic locus sampled from one or more populations undergoing complex demographic scenarios (including past population size changes, merging of populations, and migration). Furthermore, it is computationally efficient for relatively large samples (hundreds or thousands of chromosomes) as well as large genomic segments (tens to a few hundred kilobases). Here, we propose an extension of ms in order to facilitate its usage within ABC and, in particular, to perform the sampling procedure. Our aim is to provide a software that (i) draws parameter values from user-specified prior distributions, (ii) allows to choose from a variety of summary statistics, (iii) can be used for multiple unlinked loci, and (iv) enables the calculation of summary statistics in cases of incomplete information (i.e. missing data). The randomly drawn parameter values are used to perform

coalescent simulations. The simulated data are summarized into a vector of summary statistics and written to a file. This file can then be used to perform the rejection step and the other post-sampling adjustments using linear regression (Beaumont *et al.* 2002; Thornton 2009) or non-linear regression models (Blum and Francois 2010).

Materials and Methods

Generation of data: Currently, ms allows to simulate neutral polymorphism data using a set of constant, user-defined parameter values. Alternatively, employing the "tbs" option, ms permits some of the parameters to be specified from the standard input. However, even in this case the parameter values should be generated *a priori.* This may be tedious when many parameters need to be sampled from one or more distributions. msABC enables the user to specify in the command line the desired sampling distributions (uniform, normal, log-normal, gamma) for the parameters of interest. For each simulated dataset, new parameter values (e.g. the population mutation parameter, θ) are drawn from the specified distribution. In msABC, a dataset may consist of multiple independent loci. The sample size is allowed to vary among loci, similarly to the msnsam program (Ross-Ibarra *et al.* 2008), as is often the case in large genome re-sequencing projects. Furthermore, the simulation of missing information is possible.

Calculation of summary statistics: Following the generation of a dataset, summary statistics are calculated: (i) estimates of variability such as the Watterson's estimator, θ_w (Watterson 1975), or equivalently the number of segregating sites *S*, the average pairwise differences of sequences, π (Tajima 1983), (ii) summaries of the site frequency spectrum (SFS) such as *D* (Tajima 1989) and *H*

(Fay and Wu 2000), and (iii) summaries based on linkage disequilibrium, for example, the average pairwise correlation coefficient Z_{nS} (Kelly 1997). Population differentiation statistics such as F_{ST} (Hudson *et al*. 1992; Slatkin 1993), or pairwise F_{ST} have been implemented for the case of multiple population datasets. Furthermore, fixed differences, shared and private polymorphisms can be calculated between pairs of populations. When datasets are composed of multiple populations, summary statistics i - iii are calculated for each population as well as for the pooled sample. Summary statistics are calculated for each locus, and averages and variances are reported if multiple loci are simulated.

Simulations with incomplete information: Often, in Sanger re-sequencing (e.g. Hutter *et al.* 2007), microchip sequencing (e.g. http://www.dpgp.org/) or high-throughput sequencing projects (e.g. http://www.dpgp.org/), a fraction of data contains missing information, that is, non-identified nucleotides symbolized as 'N'. Missing data affect the values of summary statistics (e.g. they decrease variability), and therefore may bias the demographic inference. In msABC one can simulate missing data by specifying the coordinates (position and sequence) of each 'N' in the alignment. In a simulated dataset, segregating sites coinciding with the position of 'N' in the alignment are replaced by the missing state. The sample size of each site is then updated and the calculation of summary statistics is adapted. Details and examples are provided in the manual (pg. 12).

Code availability: The source code and documentation of msABC is available at http://bio.lmu.de/~pavlidis/msabc or upon request. msABC has been compiled and run on 32-bit Linux machines with the gcc (version 4.2.4) compiler and on 64-bit Linux machines with the gcc (version 4.1.2) compiler.

Results

The sampling process of an ABC analysis may consist of multiple steps. Parameter values are sampled from the prior distribution to simulate polymorphism data. Coalescent simulation programs such as simcoal2 (Laval and Excoffier 2004) and ms (Hudson 2002) have been used extensively for the data generation. Then, the summary statistics are calculated using the simulation results in appropriate software packages [e.g. the libsequence library (Thornton 2003)]. msABC integrates these steps into one software package that efficiently performs the sampling process of the ABC. The benefits from this integration are: (i) it allows researchers without extensive coding skills to estimate demographic models even for complicated scenarios, when the sample size of loci varies, or the dataset includes missing information, and (ii) computations are considerably faster than combining sequentially the steps of the sampling process mentioned in the previous paragraph (Figure 3.1).

Speed measurements: We compared the speed performance of msABC with the combination of the coalescent simulator ms (Hudson 2002) and the libsequence library (Thornton 2003) to calculate summary statistics. msABC out-competes this combination. As illustrated in Figure 3.1, msABC (solid line with circles) is compared with the combination ms-libsequence (dashed line with

crosses). Hudson's ms (dashed line with triangles) is used as lower bound for the time, since it simulates data without calculating summary statistics. Simulations refer to a demographic scenario of two populations with gene flow between them ($4Nm$=0.5, where N is the present day effective population size and m is the fraction of each subpopulation made up of new migrants each generation), with a (global) population contraction to $0.3N$ at time 0.01 (backwards in units of $4N$). A genome of 100 independent loci is simulated 1000 times (θ=10 per locus, ρ=10 per locus). The set of summary statistics consists of θ_w, π, D, H, F_{ST}, shared and fixed polymorphisms, and Z_{nS}. The speed difference is important especially for large (whole-genome) datasets (e.g. the 1001 genome project for *A. thaliana* (Weigel and Mott 2009)), where simulations may require extensive time periods. For example, based on Figure 3.1, simulating a genome of 100 independent loci 10^6 times when the sample size is 500 would require about 92 days on a single computer using ms-libsequence. On the other hand, msABC would require 17 days for the same computations.

Example of parameter estimation: We infer the parameters of a simple demographic model characterized by two diverging populations with recombination, in order to illustrate the usage of msABC. The model consists of three parameters: the population mutation parameter ,θ, which is identical in the present and ancestral populations; the time τ at which the two populations diverged, and the population recombination rate ρ. We used msABC to sample parameter values from uniform priors U(0; 10) and U(0; 1) for θ and τ, respectively, to simulate polymorphism dataset under this demographic model and summarize these datasets into a series of summary statistics. In all simulations, ρ=20 and the simulated datasets consist of 50 loci of 500 bp with sample size n=12. The output of msABC allows to investigate the relations between the parameters θ, τ and, summary statistics of the simulated data. Figure 3.2A illustrates the relation between τ and the amount of

differentiation between the two populations as it is measured by F_{ST}. Strong correlations between parameters and summary statistics indicate that summary statistics can be used to infer values of demographic parameters (Figure 3.2A). Posterior distributions of parameters based on observed summary statistics, the joint distribution of parameters, and simulated summary statistics can be computed from the output of msABC and the rejection/regression analysis (Beaumont *et al*. 2002; Excoffier *et al*. 2005; Thornton 2009). In order to illustrate this estimation procedure, we simulated a dataset by setting θ=5 and τ=0.5 and re-estimate the values of θ and τ using 10^6 simulated datasets. Posterior distributions were estimated by summarizing the data into the mean and the variances of the number of segregating sites, D, Z_{nS}, and F_{ST}. Figure 3.2B illustrates the posterior distribution of the parameter τ.

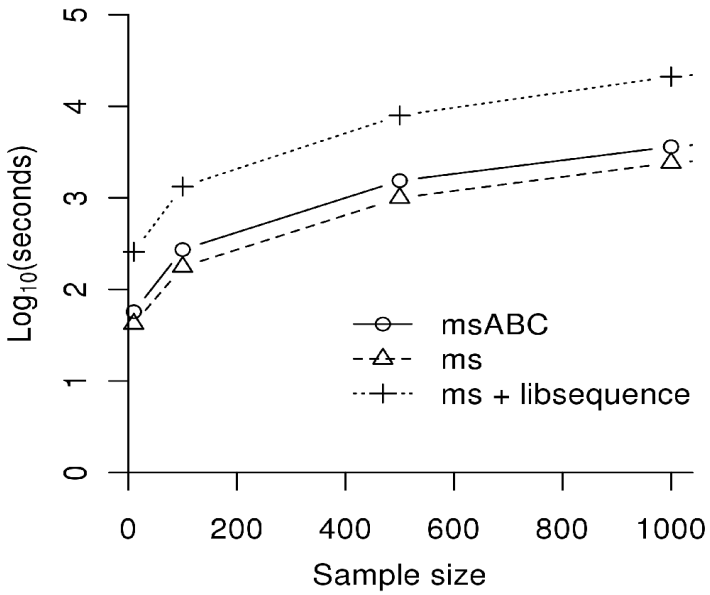
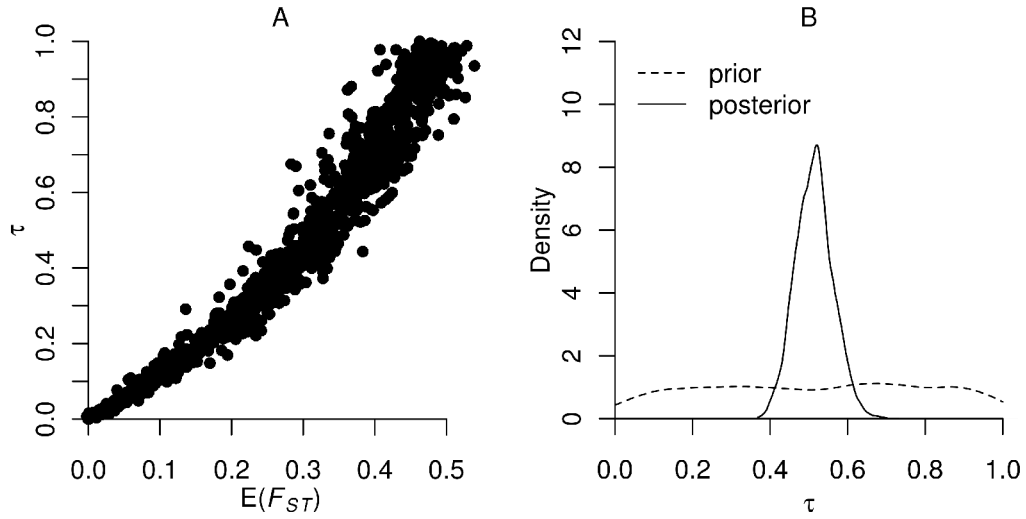


Figure 3.1: Speed comparison (in log10 seconds) when the sample size is between 10 and 1000. msABC is about 6 times faster than the combination ms-libsequence when the sample size equals 1000.

Figure 3.2: Results obtained from msABC. A) Examining the relationship



between the parameter and the summary statistic E(FST). B) Posterior distribution of the parameter , obtained after applying the output of msABC in algorithms that perform the rejection and regression steps of ABC analysis.

Discussion

msABC facilitates the sampling process of an ABC analysis. The command line is similar to the command line of ms, thus shortening the learning curve for a user who is familiar with ms. Although msABC can be used to simulate single loci, most demographic analyses in molecular population genetics are characterized by large datasets composed of several chromosomal fragments scattered along the genomes (Ometto *et al*. 2005; Nordborg *et al*. 2005; Hutter *et al.* 2007). msABC can simulate multi-locus datasets, where each fragment is characterized by its own length, sample size, recombination rate, and mutation rate. msABC provides a collection of commonly used summary statistics that allow to quantify levels of polymorphisms, linkage disequilibrium, population differentiation, and the shape of the frequency spectrum of derived mutations. The

complete list of available summary statistics can be found in the user's manual (see http://bio.lmu.de/~pavlidis/msabc). Furthermore, msABC extends the flexibility of Hudson's ms by allowing variable sample size among fragments and missing data simulation. It allows to analyse datasets that contain missing data by simulating them and then calculating summary statistics. This may be important in demographic inference of large datasets which typically consist of a large amount of incomplete information (e.g. http://www.dpgp.org/).

The speed performance can be important for large datasets. Assuming that simulating data of tens or hundreds of kbs (with typical values of recombination rates) for a sample that consists of hundreds or thousands of individuals may require months of computational time, an improvement of five to six times shortens considerably the time of the inference project. This is especially true if the project is carried out on personal computers instead of cluster machines (Figure 3.1).

Alternative ways to obtain summary statistics from simulated data could be implemented by replicating ms commands with different parameters. In the best case this would require extensive scripting for calculating the priors and summary statistics. However, when missing data are included in the dataset or the sample sizes of loci vary, it would not be possible to perform simulations that match the observed data.

msABC can be used to examine the relationship between parameters and summary statistics (figure 3.2A). This helps to inspect the usability of certain summary statistics in estimating parameters. Summary statistics that are related monotonically to target parameters are expected to be useful for estimating them. Additionally, msABC can be used to obtain the null distributions of a multitude of summary statistics under demographic models.

msABC outputs samples from the joint distribution of parameters and summary statistics under a given demographic model. A follow-up step in the analysis (rejection) retains the closest points to the observed data. The parameter values that have been used to generate those simulations

comprise an approximation of the true posterior distribution of the parameters of interest. An improvement of this approximation has been proposed by Beaumont *et al.* (2002) that corrects for the fact that the accepted simulations never match precisely the observed data (linear-regression). A more sophisticated approach has been suggested by Blum and Francois (2010). msABC does not perform the rejection and regression steps. Algorithms needed to perform these post-sampling steps have been implemented elsewhere [e.g. abcReg by K. Thornton (http://www.molpopgen.org/software/abcreg) or non-linear regression models by Blum and Francois (2010) (http://membres-timc.imag.fr/Michael.Blum/my_publications.html)].

A critical point in ABC refers to the model choice (Pritchard *et al.* 1999; Fagundes *et al.* 2007). Typically, different demographic scenarios are simulated and the scenario with the highest relative posterior probability is then used (Fagundes *et al.* 2007; Francois *et al.* 2008). However, this model does not necessarily provide a good fit to the observed data, since it simply indicates the best model among the tested models (Ratmann *et al.* 2009). Therefore, once the best model and its parameters have been inferred, it is necessary to investigate whether simulations under this model are able to predict the observations (predictive simulations).

Finally, in ABC the set of summary statistics may be crucial. It has been shown that uninformative summary statistics add noise to the distance between simulations and observations (Joyce and Marjoram 2008); thus they should be avoided. Therefore, the smallest set of summary statistics that captures the information carried by the dataset should be used. The choice of summary statistics is an active area of research. Joyce and Marjoram (2008) suggested a scheme for scoring statistics according to whether they improve the inference. Alternatively, Wegmann *et al.* (2009) proposed partial least square regression (Boulesteix and Strimmer 2007) in order to reduce the dimensionality. In Table 3.1 we suggest which summary statistics should be used to infer certain demographic parameters. However, since the information provided by statistics may vary between

demographic scenarios, investigating the relationship between them and demographic parameters under various demographic scenarios of interest is necessary.

Table 3.1: Demographic parameters and population genetics summary statistics that can be used for the inference of parameter values

Demographic parameters	**Summary statistics**
θ (population mutation rate)	θ_w (or S), θ_π
ρ (population recombination rate)	Z_{nS}
time of population size expansion	θ_w (or S), θ_π, D
time of population size contraction	θ_w (or S), θ_π, D, Z_{nS}
magnitude of population size change	θ_w (or S), θ_π, D, Z_{nS}
migration rate (island model)	F_{ST}
τ (time of divergence between two populations)	pairwise F_{ST}

Summary statistics are described in the section Calculation of summary statistics.

Acknowledgements

We would like to thank Dirk Metzler, Stephan Hutter, and Aurelien Tellier (LMU Munich) for useful discussions. This work is supported by grants from the Volkswagen-Foundation (I/82770) to PP and by the DFG (Ste 325/5-3 and 325/12-1) to WS.

General Discussion

The demography of *Drosophila melanogaster*: In our analysis of the demographic history of Asian populations of *D. melanogaster,* we considered that the African population experienced an expansion as proposed by Li and Stephan (2006). After having estimated posterior distributions for the parameters of this model we found, however, that our best model couldn't account for all the aspects of the African dataset. Indeed, ABC model choice analysis even showed that a population size bottleneck was associated with a higher posterior probability than the expansion model (unpublished results). This results are in line with a recent population genetic survey of a large number of African populations (Pool *et al.* 2006) suggesting that the centre of Origin of the species could well be located in Uganda rather than in Zimbabwe as it was assumed in our study. The bottleneck that we detect in our dataset could therefore be the signature of past range expansions from Uganda to Zimbabwe (Excoffier *et al.* 2009). If the centre of origin of *D. melanogaster* is located in Uganda, an under-investigated population, then we might find there populations that are closer to equilibrium than the Zimbabwean populations. These equilibrium populations could greatly improve our ability to detect selective sweeps, since our methods are strongly affected by past demographic events (Pavlidis *et al*. 2010a). A correct understanding of the evolutionary forces that shaped patterns of genetic variation in the African population is also of prime importance because this population plays a central role in the modelling of the history of derived cosmopolitan populations. Therefore, further work is needed to understand the complete demographic history of cosmopolitan populations.

The most important improvement to be done is to incorporate inter-continental migration in our models. The importance of migration has appeared with the discovery that North American

populations carry similar levels of genetic diversity as European populations (Caracristi and Schlötterer 2003). This observation doesn't support the idea that the American population was founded by a single founder event, as we assumed it for the foundation of the European and Asian populations. The reason is that such a scenario is expected to lead to a reduction of genetic polymorphism in the derived population (Ometto *et al* 2005; Li and Stephan 2006). More interestingly, entomological records indicate that *D. melanogaster* was observed for the first time in North America in 1875, in the state of New York (Lintner 1882; Keller 2007). Only 25 years later, the species has been qualified to be „the most commonest species“ all over the United States (Howard 1900). This gives us an idea about the impressive dispersal capacities of *D. melanogaster* and makes it likely that human-mediated worldwide migration events between populations can occur on a regular basis.

Another question that remains to be answered is the interpretation of the observed ratio of X-linked to autosomal diversity (hereafter X/A diversity ratio) in Asian populations. Under the standard neutral model assumptions, the X/A diversity ratio would be 0.75, following the numbers of each chromosome in a mating pair. However, Pool and Nielsen (2008) have shown that the observed reduction of the X/A diversity ratio in the European population could be best explained by recurrent founder events that occurred during the range expansion of *D. melanogaster* out of Africa. We found that the observed X/A diversity ratio (corrected for mutational biases) was 0.63 in the Asian population. Which is smaller than the observed value in the African population (0.87), but larger than the value observed in the European population (0.52). One explanation for this result would be that the recurrent founder events that have been analysed by Pool and Nielsen (2008) could not only represent population size fluctuations associated with range expansion processes but could also reflect the strong impact of seasonality on European populations of *D. melanogaster.* Since Southeast Asian populations are expected to be less exposed to low wintry temperatures, it

could well be that climatic seasonality is responsible for the observed differences in X/A diversity ratios.

Although much effort has already been invested into the analysis of *D. melanogaster's* history (Glinka *et al.* 2003; Ometto *et al.* 2005; Li and Stephan 2006; Thornton and Andolfatto 2006; Pool and Nielsen 2008), it seems that several questions remain unanswered. Answers to these questions might be given soon by the analyses of the new datasets produced by next-generation sequencing technologies (Metzker 2010) and statistical methods allowing the investigation of more complex models (Csilléry *et al.* 2010).

Estimation of the germination rates of two wild tomato species: In the second project we combined ecological and genetic data to investigate the effect of seed dormancy on the molecular evolution of wild tomatoes from the genus *Solanum.* Large population sizes compared to census sizes observed in two wild tomatoes species from western South America are explained, using coalescence theory and ABC analysis, to be due to the existence of seed banks. We show that seed banks increase the effective size of populations, and this effect can be distinguished from the effect of metapopulation structure using a combination of local population and species-wide samples. The ABC method allows us to combine genetic diversity data with ecological information on the number of demes and census size of demes, to perform model-based inference of germination rates for each species. We also show that two species with different ecological habitats and metapopulation structure (different number of demes) exhibit different germination rates. This difference can represent bet-hedging adaptation to environmental stochasticity over space and time.

One critical aspect of the statistical analysis we conducted in this project was the modest amount of genetic information used in the ABC analysis (Tables 2.2). The small number of

independent fragments considered in this study didn't allow us to use the variances of the statistics' distribution across loci in the ABC analysis. Consequently, our ability to identify past fluctuations of population sizes was greatly reduced, as these demographic events have a strong influence on the variance of several summary statistics of polymorphism data. In this study, however, we were not as interested in the species' past history as in identifying a specific life-history trait of these plants: seed dormancy. The main effect of seed dormancy is to increase the level of genetic variability. We believe that the observed average values of diversity that we measured in our samples still contain enough information about the species average level of genetic diversity, to be compared with the expected levels of genetic diversity based on the observed census sizes.

It seems, however, that the population size expansion that we identified in *Solanum peruvianum* has been confirmed by recently developed composite likelihood methods that are expected to make a better use of the full dataset (Lisha Naduvilezhath, personal communication). Issues concerning the amount of information carried by our datasets, about a specific aspect of the investigated model, will be resolved in the future by carrying out performance analyses of our ABC estimation procedure on simulated pseudo-observed datasets.

Another critical aspect that remains to be addressed is the occurrence of very recent demographic events that could have affected wild tomatoes populations. Recent habitat destruction due to urbanization (Aurelien Tellier, personal communication) could be difficult to detect with our coalescent-based inference method. This is due to the fact that very short and recent population size fluctuations might not influence the shape of gene genealogies that can be inferred from genetic data. It remains to be tested how good our method behaves in the presence of such recent events and to assess the potential magnitude of these habitat reductions.

msABC: a coalescent simulator for ABC analysis: The third project's aim was to develop a software that would facilitate the use of ABC estimation procedures in population genetics. Our major concern during the development of this software was to propose a tool that would completely alleviate the necessity for the user to develop his own code at any stage of the ABC analysis. We succeeded in this task by proposing msABC (Pavlidis *et al.* 2010), a software that has been created by performing a series of modifications to the well-established software ms (Hudson 2002). Even if, for an experienced programmer, developing specific tools for a given analysis still remains the optimal solution, the increasing number of msABC users already indicates that it *fulfils* (at least some of) the needs of biologists that are interested in interpreting observed patterns of genetic polymorphism in a model-based framework. Other software or packages implementing ABC methods for population genetics have been published recently (Cornuet *et al.* 2008; Lopes *et al.* 2009; Wegmann *et al.* 2010). Among them the ABC toolbox package generated by Wegmann *et al.* (2010) is proposing, in addition to a very well documented user manual, several interesting programs that can be used to perform the partial-least square transformation (Boulesteix *et al.* 2007) as well as the regression-adjustment of the vectors of retained parameter values (Beaumont *et al.* 2002). It is likely, however, that this list of programs will be extended in the following years, since the development of ABC methods is actually an active field of research (Ratmann *et al.* 2009; Blum and Francois 2010; Bazin *et al.* 2010) and that interesting methodological improvements have already been proposed.

One of these improvements for example deals with the fact that classical ABC methods only provide information about the relative fit of a model to the observed dataset. Different models can be compared against each other, but the model with the highest posterior probability might still provide a poor fit to the observed dataset. To overcome this problem, Ratman *et al.* (2009) proposed ABCμ, an ABC method that incorporates model diagnostics within an ABC framework. ABCμ can

identify which aspects of the dataset cannot be correctly predicted by the model and therefore provides a powerful way to perform model refinement.

Bibliography

Abe T, Wada K, Nakagoshi N. 2008. Extinction threats of a narrowly endemic shrub, *Stachyurus macrocarpus* (Stachyuraceae) in the Ogasawara Islands. *Plant Ecol* 198:169−183.

Amos W, Harwood J. 1998. Factors affecting levels of genetic diversity in natural populations. *Philos Trans R Soc Lond B Biol Sci* 353:177−186.

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

Ayre D, O'Brien E, Ottewell K, Whelan R. 2010. The accumulation of genetic diversity within a canopy-stored seed bank. *Mol Ecol* 19:2640−2650.

Bar-Yosef O. 1998. The Natufian culture in the Levant, threshold to the origins of agriculture. *Evol Anthropol* 6:159−177.

Barton NH. 1998. The effects of hitch-hiking on neutral genealogies. *Genet Res* 72:123−133.

Baudry E, Viginier B, Veuille M. 2004. Non-African populations of *Drosophila melanogaster* have a unique origin. *Mol Biol Evol* 21:1482−1491.

Bazin E, Dawson KJ, Beaumont MA. Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics* 185:587−602.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025−2035.

Beaumont MA, Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* 13:969−980.

Beaumont MA. 2008. Joint determination of topology, divergence time, and immigration in population trees. In: Matsumura S, Forster P, Renfrew C, editors. *Simulations, Genetics and Human Prehistory*.

Beaumont MA. 2010. Approximate Bayesian Computation in Evolution and Ecology. *Annu Rev Ecol Evol S* 41:379−405.

Becquet C, Patterson N, Stone AC, Przeworski M, Reich D. 2007. Genetic structure of chimpanzee populations. *PloS Genet* 3:e66.

Beerli P, 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 U S A* 98:4563−4568.

Bentsink L, Hanson J, Hanhart CJ *et al.* (13 co-authors). 2010. Natural variation for seed dormancy

in Arabidopsis is regulated by additive genetic and molecular pathways. *Proc Natl Acad Sci U S A* 107:4264−4269.

Blum M, Francois O. 2010. Non-linear regression models for Approximate Bayesian Computation. *Stat Comput* 20:63−73.

Boulesteix A, Strimmer K. 2007. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform* 8:32−44.

Caracristi G, Schlötterer C. 2003. Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles. *Mol Biol Evol* 20:792—799.

Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289−1303.

Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet Res* 70:155−174.

Charlesworth B, Charlesworth D, Barton NH. 2003. The effects of genetic and geographic structure on neutral variation. *Annu Rev Ecol Evol S* 34:99−125.

Chetelat RT, Pertuze RA, Faundez 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.

Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA. 2010. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* 186:983−995.

Cohen D. 1966. Optimizing reproduction in a randmoly varying environment. *J Theor Biol* 12:119−129.

Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM, Balding DJ, Guillemaud T, Estoup A. 2008. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* 24:2713−2719.

Csilléry K, Blum MGB, Gaggiotti OE, Francois O. 2010. Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol* 25:410−418.

David J, Bocquet C, Pla E. 1976. New results on the genetic characteristics of the Far East race of *Drosophila melanogaster*. *Genet Res* 28:253−260.

David JR, Capy P. 1988. Genetic variation of *Drosophila melanogaster* natural populations. *Trends Genet* 4:106−111.

Devries TJ. 1987. A review of geological evidence for ancient el-nino activity in peru. *J Geophys Res-Oceans* 92:14471−14479.

Ellner S. 1985. ESS germination strategies in randomly varying environments. I. Logistic-type models. *Theor Popul Biol* 28:50−79.

Ellner S, Hairston NG. 1994. Role of overlapping generations in maintaining genetic variation in a fluctuating environment. *Am Nat* 143:403−417.

Epling C, Lewis H, Ball FM. 1960. The breeding group and seed storage - a study in population dynamics. *Evolution* 14:238−255.

Espeland EK, Rice KJ. 2010. Ecological effects on estimates of effective population size in an annual plant. *Biol Conserv* 143:946−951.

Evans ME, Dennehy JJ. 2005. Germ banking: bet-hedging and variable release from egg and seed dormancy. *Q Rev Biol* 80:431−451.

Evans ME, Ferrière R, Kane MJ, Vanable DL. 2007. Bet hedging via seed banking in desert evening primroses (Oenothera, Onagraceae): demographic evidence from natural populations. *Am Nat* 169:184−194.

Excoffier L, Estoup A, Cornuet J. 2005. Bayesian analysis of an admixture model with mutations

and arbitrarily linked markers. *Genetics* 169:1727−1738.

Excoffier L, Foll M, Petit RJ. 2009. Genetic consequences of range expansions. *Annu Rev Ecol Evol Syst* 40:481−501.

Fagundes NJR, Ray N, Beaumont MA, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A* 104:17614−17619.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405−1413.

Fenner M, Thompson K. 2004. *The Ecology of Seeds*. Cambridge, UK: Cambridge University Press.

Fiston-Lavier S, Singh ND, Lipatov M, Petrov DA. 2010. *Drosophila melanogaster* recombination rate calculator. *Gene* 463:18−20.

Francois O, Blum MGB, Jakobsson M, Rosenberg NA. 2008. Demographic history of European populations of *Arabidopsis thaliana*. *PloS Genet* 4:e1000075.

Frankham R. 1995. Effective population-size adult-population size ratios in wildlife - a review. *Genet Res* 66:95−107.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693−709.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2003. *Bayesian Data Analysis. Ed. 2.* Chapman and Hall/CRC Press, London/New York/Cleveland/Boca Raton, FL.

Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. G*enetics* 165:1269−1278.

Glinka S, Stephan W, Das A. 2005. Homogeneity of common cosmopolitan inversion frequencies in Southeast Asian *Drosophila melanogaster*. *J Genet* 84:173−178.

Gossmann TI, Song B, Windsor AJ, Mitchell-Olds T Dixon CJ, Kapralov MV,Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol* 27:1822−1832.

Gregory-Wodzicki KM. 2000. Uplift history of the Central and Northern Andes: A review. *Geol Soc Am Bull* 112:1091−1105.

Gutierrez JR, Meserve PL. 2003. El Nino effects on soil seed bank dynamics in north-central Chile. *Oecologia* 134:511−517.

Hairston NG, Destasio BT. 1988. Rate of evolution slowed by a dormant propagule pool. *Nature* 336:239−242.

Hamilton G, Stoneking M, Excoffier L. 2005. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proc Natl Acad Sci U S A* 102:7476−7480.

Hartl DL, Clark AG. 1989. *Principles of Population Genetics.* Sinauer Associates, Sunderland, MA.

Hartley AJ. 2003. Andean uplift and climate change. *J Geol Soc London* 160:7−10.

Honnay O, Bossuyt B, Jacquemyn H, Shimono A, Uchiyama K. 2008. Can a seed bank maintain the genetic variation in the above ground plant population? *Oikos* 117:1−5.

Honnay O, Jacquemyn H, Van Looy K, Vandepitte K, Breyne P. 2009. Temporal and spatial genetic variation in a metapopulation of the annual *Erysimum cheiranthoides* on stony river banks. *J Ecol* 97:131−141.

Howard LO. 1900. A contribution to the study of the insect fauna of human excrement. Proc. Wash. Acad. Sci. 2:541−604.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* 132:583−589.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337−338.

Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W. 2007. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. *Genetics* 177:469−480.

Hutter S, Stephan W. 2009. Recombination rates may affect the ratio of X to autosomal noncoding polymorphism in African populations of *Drosophila melanogaster* – Reply. *Genetics* 181:1703.

Jensen JD, Wong A, Aquadro CF. 2007. Approaches for identifying targets of positive selection. *Trends Genet* 23:568−577.

Jombart T, Eggo RM, Dodd P, Balloux F. 2009. Spatiotemporal dynamics in the early stages of the 2009 A/H1N1 influenza pandemic. *Plos Curr* 1:RRN1026.

Jost L. 2008. G_{ST} and its relatives do not measure differentiation. *Mol Ecol* 17:4015−4026.

Joyce P, Marjoram P. 2008. Approximately sufficient statistics and Bayesian computation. *Stat Appl Genet Mol Biol* 7:Article26

Jurado E, Flores J. 2005. Is seed dormancy under environmental control or bound to plant traits? *J Veg Sci* 16:559−564.

Kaj I, Lascoux M. 1999. Probability of identity by descent in metapopulations. *Genetics* 152:1217−1228.

Kaj I, Krone SM, Lascoux M. 2001. Coalescent theory for seed bank models. *J Appl Proba* 38:285−300.

Kass RE, Raftery AE. 1995. Bayes Factors. *J Am Stat Assoc* 90:773−795.

Keller A. 2007. *Drosophila melanogaster*'s history as a human commensal. *Curr Biol* 17:R77−R81

Kelly JK, 1997. A test of neutrality based on interlocus associations. *Genetics* 146:1197−1206.

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160:765−777.

Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167:1513−1524.

Kimure M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16:111−120.

Kingman JFC. 1982. The coalescent. *Stoch Proc Appl* 13:235−248.

Kuhner MK. 2009. Coalescent genealogy samplers: windows into population history. *Trends Ecol Evol* 24:86−93.

Lande R. 1988. Genetics and demography in biological conservation. *Science* 241:1455−1460.

Laval G, Excoffier L. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* 20:2485−2487.

Levin DA. 1990. The seed bank as a source of genetic novelty in plants. *Am Nat* 135:563−572.

Levine JM, McEachern AK, Cowan C. 2008. Rainfall effects on rare annual plants. *J Ecol* 96:795−806.

Lopes JS, Balding D, Beaumont MA. 2009. PopABC: a program to infer historical demographic parameters. *Bioinformatics* 25:2747−2749.

Li YJ, Satta Y, Takahata N. 1999. Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. *Genes Genet Syst* 74:117−127.

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

Lintner J.A. 1882. First Annual Report on the Injurious and Other Insects of the State of New York. (Albany, New York: Weed, Parsons and Co.)

Lundemo S, Falahati-Anbaran M, Stenøien HK. 2009. Seed banks cause elevated generation times and effective population sizes of Arabidopsis thaliana in northern Europe. *Mol Ecol* 18:2798−2811.

Manica A, Amos W, Balloux F, Hanihara T. 2007. The effect of ancient population bottlenecks on human phenotypic variation. *Nature* 448:346−348.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* 23:23−35.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9:356−369.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* 11:31−46.

Nakazato T, Bogonovich M, Moyle LC. 2008. Environmental factors predict adaptive phenotypic differentiation within and between two wild Andean tomatoes. *Evolution* 62:774−792.

Nakazato T, Warren DL, Moyle LC. 2010. Ecological and geographic modes of species divergence in wild tomatoes. *Am J Bot* 97:680−693.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* 76:5269−5273.

Nesbitt TC, Tanksley SD. 2002. Comparative sequencing in the genus Lycopersicon. Implications for the evolution of fruit size in the domestication of cultivated tomatoes. *Genetics* 162:365−379.

Neuenschwander S, Largiadèr CR, Ray N, Currat M, Vonlanthen P, Excoffier L. 2008. Colonization history of the Swiss Rhine basin by the bullhead (Cottus gobio): inference under a Bayesian spatially explicit framework. *Mol Ecol* 17:757−772.

Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet* 39:197−218.

Nordborg M, Hu TT, Ishino Y et al. (23 co-authors). 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *Plos Biol* 3:e196.

Novembre J, Johnson T, Bryc K et al. (12 co-authors). 2008. Genes mirror geography within Europe. *Nature* 456:98−101.

Nunney L. 2002. The effective size of annual plant populations: the interaction of a seed bank with fluctuating population size in maintaining genetic variation. *Am Nat* 160:195−204.

Oddou-Muratorio S, Klein EK. 2008. Comparing direct vs. indirect estimates of gene flow within a population of a scattered tree species. *Mol Ecol* 17:2743−2754.

Ometto L, Glinka S, De Lorenzo D, Stephan W. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Mol Biol Evol* 22:2119−2130.

Pannell JR, Charlesworth B. 1999. Neutral genetic diversity in a metapopulation with recurrent local

extinction and recolonization. *Evolution* 53:664−676.

Pannell JR. Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* 57:949−961.

Pavlidis P, Jensen JD, Stephan W. 2010. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics* 185:907−922.

Pavlidis P, Laurent S, Stephan W. 2010. msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Mol Ecol Resour* 10:723−727.

Peralta IE, Knapp SK, Spooner DM. 2005. New species of wild tomatoes (*Solanum* section *Lycopersicon*: Solanaceae) from Northern Peru. *Syst Bot* 30:424−434.

Peralta IE, Spooner DM, Knapp S. 2008. The taxonomy of tomatoes: a revision of wild tomatoes (*Solanum* section *Lycopersicon*) and their outgroup relatives in sections *Juglandifolium* and *Lycopersicoides*. *Systematic Botany Monographs* 84:1−186.

Peter BM, Wegmann D, Excoffier L. 2009. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Mol Ecol* 19:4648−4660.

Pool JE, Aquadro CF. 2006. History and structure of sub-Saharan populations of *Drosophila melanogaster*. *Genetics* 174:915−929.

Pool JE, Bauer DuMont V, Mueller JL, Aquadro CF. 2006. A scan of molecular variation leads to the narrow localization of a selective sweep affecting both Afrotropical and cosmopolitan populations of *Drosophila melanogaster*. *Genetics* 172:1093−1105.

Pool JE, Nielsen R. 2008. The impact of founder events on chromosomal variability in multiply mating species. *Mol Biol Evol* 25:1728−1736.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* 16:1791−1798.

Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. *Curr Biol* 15:R159—R160.

Rajon E, Venner S, Menu F. 2009. Spatially heterogeneous stochasticity and the adaptive diversification of dormancy. *J Evol Biol* 22:2094−2103.

Ratmann O, Andrieu C, Wiuf C, Richardson S. 2009. Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc Natl Acad Sci U S A* 106:10576−10581.

Rick CM. 1963. Differential Zygotic Lethality in a Tomato Species Hybrid. *Genetics* 48:1497−1507.

Rick CM, Kesicki E, Fobes JF, Holle M. 1976. Genetic and biosystematic studies on two new sibling

species of Lycopersicon from inter-Andean Peru. *Theor Appl Genet* 47:55−68.

Rick CM. 1986. Reproductive isolation in the *Lycopersicon peruvianum* complex. D'arcy, W. G. (Ed.). *Solanaceae: Biology and Systematics; Second International Symposium*, St. Louis, Mo., USA, Aug. 3-6, 1983. Xiii+603p. Columbia University Press: New York, N.Y., USA. Illus.

Rose LE, Langley CH, Bernal AJ, Michelmore RW. 2005. Natural variation in the Pto pathogen resistance gene within species of wild tomato (Lycopersicon). I. Functional analysis of Pto alleles. *Genetics* 171:345−357.

Rose LE, Michelmore RW, Langley CH. 2007. Natural variation in the Pto disease resistance gene within species of wild tomato (Lycopersicon). II. Population genetics of Pto. *Genetics* 175:1307−1319.

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

Ross-Ibarra J, Wright SI, Foxe JP, Kawabe A, DeRose-Wilson L, Gos G, Charlesworth D, Gaut BS. 2008. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. 3:e2411.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism

analyses by the coalescent and other methods. *Boinformatics* 19:2496−2497.

Sabeti PC, Reich DE, Higgins JM *et al.* (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832−837.

Schlötterer C, Neumeier H, Sousa C, Nolte V. 2006. Highly structured Asian *Drosophila melanogaster* populations: a new tool for hitchhiking mapping? *Genetics* 172:287−292.

Siol M, Bonnin I, Olivieri I, Prosperi JM, Ronfort J. 2007. Effective population size associated with self-fertilization: lessons from temporal changes in allele frequencies in the selfing annual *Medicago truncatula*. *J Evol Biol* 20:2349−2360.

Slatkin M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* 47:264−279.

Spooner DM, Peralta IE, Knapp S. 2005. Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum L.* section *Lycopersicon* (Mill.) Wettst.]. *Taxon* 54:43−61.

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.

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 sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* 182:205−216.

Stephan W, Langley CH. 1998. DNA polymorphism in Lycopersicon and crossing-over per physical length. *Genetics* 150:1585−1593.

Stephan W, Song YS, Langley CH. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* 172:2647−2663.

Stephan W, Li H. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity* 98:65−68.

Stephan W. 2010. Genetic hitchhiking versus background selection: the controversy and its implications. *Philos Trans R Soc Lond B Biol Sci* 365:1245−1253.

Stocklin J, Fisher M. 1999. Plants with longer-lived seeds have lower local extinction rates in grassland remnants. *Oecologia* 120:539−543.

Szmaragd C, Balloux F. 2007. The population genomics of hepatitis B virus. *Mol Ecol* 16:4747−4758.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics*

105:437–460.

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

Tanabe K, Mita Toshihiro, Jombart T et al. (21 co-authors). 2010. *Plasmodium falciparum* accompanied the human expansion out of Africa. *Curr Biol* 20:1283–1289.

Tanksley SD, Ganal MW, Prince JP et al. (10 co-authors). 1992. High density molecular linkage maps of the tomato and potato genomes. *Genetics* 132:1141–1160.

Templeton AR, Levin DA. 1979. Evolutionary consequences of seed pools. *Am Nat* 114:232–249.

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327.

Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* 172:1607–1619.

Thornton KR. 2009. Automating approximate Bayesian computation by local linear regression. *BMC Genet* 10:35.

Tsacas L, Lachaise D. 1974. Quatre nouvelles especes de la Cote-d'Ivoire du genre *Drosophila*,

groupe *melanogaster*, et discussion de l'origine du sous-groupe *melanogaster*. *Annls Univ Abidjan E Ecol* 7:193−211.

Tudhope S, Collins M. 2003. Global change: The past and future of El Niño. *Nature* 424:261−262.

Valleriani A, Tielbörger K. 2006. Effect of age on germination of dormant seeds. *Theor Popul Biol* 70:1−9.

Vitalis R, Glémin S, Oliviéri I. 2004. When genes go to sleep: the population genetic consequences of seed dormancy and monocarpic perenniality. *Am Nat* 163:295−311.

Wakeley J, Aliacar N. 2001. Gene genealogies in a metapopulation. *Genetics* 159:893−905.

Wakeley J, Takahashi T. 2004. The many-demes limit for selection and drift in a subdivided population. *Theor Popul Biol* 66:83−91.

Wakeley J. 2009. *Coalescent Theory: An Introduction*. Ben Roberts, Greenwood Village, Colorado.

Wang JL, Caballero A. 1999. Developments in predicting the effective size of subdivided populations. *Heredity* 82:212−226.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7:256−276.

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

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L. 2010. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* 11:116.

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

Whitlock MC. 2003. Fixation probability and time in subdivided populations. *Genetics* 164:767−779.

Wilson IJ, Balding DJ. 1998. Genealogical inference from microsatellite data. *Genetics* 150:499−510.

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

Zivkovic D, Wiehe T. 2008. Second-order moments of segregating sites under variable population size. *Genetics* 180:341−357.

Curriculum Vitae

Stefan Laurent

Rebholzstrasse 8, 81377 München

laurent@bio.lmu.de

Education

2007 – present **University of Munich (LMU)**
PhD, Department of Evolutionary Biology

2004 – 2006 **University of Montpellier II**
Master in Ecology, Biodiversity and Evolution

2003 – 2004 **University of Montpellier II**
Bachelor, Biology of Organisms

2002 – 2003 **University of Tours**
DEUG SV2

2001 – 2002 **University of Perpignan**
DEUG SV1

Work Experience

2006 (August – December)

CIRAD (Montpellier)
Creation of a pedagogical support entitled:
“ Dynamics of diversity in cultivated plants “

2006 (January – July)

Training course CNRS & INRA
Impact of the breeding system on the evolution of genomes:
The case of the Triticeae tribe

2005 (January – June)

Training CNRS
Study of the Hybrid Zone between *Mus musculus musculus* and *Mus musculus domesticus* in the Jutland peninsula

Teaching/Seminars

- Courses and tutorials of Evolutionary Biology at the University of Munich (LMU)
- Organization of the evolutionary biology PhD meeting of the German society for zoology (DZG) in 2009

Technical skills

- Programming languages: C/C++, R, Perl, Shell
- Languages: French (native), German (fluent), English (fluent)

Publications

Laurent SJY, Werzner A, Excoffier L, Stephan W. 2011. Approximate Bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. *Mol Biol Evol* (accepted with minor revisions).

Pavlidis P*, Laurent S*, Stephan W. 2010. msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Mol Ecol Resour* 10:723–727. (**contributed equally*).

Escobar JS, Cenci A, Bolognini J, Haudry A, Laurent S, David J, Glémin S. 2010. *Evolution* 64:2855–2872.

Foitzik S, Bauer S, Laurent S, Pennings PS. 2009. Genetic diversity, population structure and sex-biased dispersal in three co-evolving species. *J Evol Biol* 22:2470–2480.

Presentations

September 2010: *Above-ground plant populations are just the tip of the iceberg: seed banks and metapopulations in wild tomato species*. Mind the Gap Workshop. Max-Planck Institute for Evolutionary Biology. Plön, Germany

June 2009: *Demographic analyses of one ancestral and two derived populations of Drosophila melanogaster.* Evolution Annual Meeting. Moscow, US-ID.

June 2009: *Demographic analyses of one ancestral and two derived populations of Drosophila melanogaster.* SMBE Meeting. Iowa City, US-IA.