Population Genetic Approaches to Speciation of Wild Tomatoes with Special Reference to *Solanum habrochaites* and *S. arcanum*

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

Carlos Gonzalo Merino Méndez
Aus Lima

ERKLÄRUNG

Diese Dissertation wurde im Sinne von §12 der Promotionsordnung von Herrn Professor Dr. Wolfgang 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 hiermit ehrenwörtlich, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt wurde.

München, 24.11.2009

Carlos Gonzalo Merino Méndez

1. Gutachter: Prof. Dr. Wolfgang Stephan
2. Gutachter: Prof. Dr. John Parsch

Dissertation eingereicht: 24.11.2009

To my parents,
for their support.

Summary

This thesis entails the results of three research projects. These have focused on the influence of diversity, demography and structure in the divergence (i.e. the speciation process) of four wild tomato species.

In the first project, using coalescent simulations, we studied the impact of three different sampling schemes on patterns of neutral diversity in structured populations. Specifically, we evaluated two summary statistics based on the site frequency spectrum (Tajima's D and Fu and Li's D) as a function of migration rate, demographic history of the entire metapopulation and the sampling scheme. Using simulations, we demonstrate strong effects of the sampling scheme on Tajima's D and Fu and Li's D statistics, particularly under species-wide expansions. Under such scenarios, the effects of spatial sampling may persist up to very high levels of gene flow (*Nm* > 25). This suggests that validating the assumption of panmixia is crucial if robust demographic inferences are to be made from local or pooled samples.

For the second project, we investigated how selection acts in four species of wild tomatoes (*S. habrochaites*, *S. arcanum*, *S. peruvianum*, and *S. chilense*) using sequence data from eight housekeeping genes. Our analysis quantified the number of adaptive and deleterious mutations, and the distribution of fitness effects of new mutations (its mean and variance) taking into account the demography of the species. We found no evidence for adaptive mutations but very strong purifying selection in coding regions of the four species. More interestingly, the four species exhibit different strength of purifying selection in non-coding regions (introns). Taking into account the results from the first project, we also highlighted the utility of analyzing pooled samples and local samples from a metapopulation in order to measure selection and the distribution of fitness effects.

Finally, the third project deals with the estimation of nucleotide diversity and population structure in *S. habrochaites* and *S. arcanum*. We also compared

these results to those of *S*. *peruvianum* and *S*. *chilense*. We found that *S*. *arcanum* and *S*. *habrochaites* present lower diversity levels than *S*. *peruvianum* and *S*. *chilense*. Our neutrality tests have not revealed any particular pattern, leading us to conclude that the loci sequenced for the present study have not evolved under strong positive selection, although they show a distinctive pattern of purifying selection (second project). We also tested the demography of all four species and found a strong expansion after a bottleneck in the recent past for *S*. *peruvianum* and a similar statistically significant pattern for *S*. *arcanum*, even though the signal seemed weaker in this case. Additionally, we found moderate levels of population sub-structure in these species, similar to previous results found in *S*. *peruvianum* and *S*. *chilense*. Still, regardless of the levels of population structure, we found at least two (Rupe and San Juan from *S*. *arcanum*) populations collected in the field that could actually be considered as a single deme. We also expanded these population structure analyses to gain insight into the phylogenetic relations between the four species in order to contribute to the taxonomical treatment of the *Solanum* section *Lycopersicon* from a population genetics perspective. Thus, we found a clear differentiation between *S. arcanum* and *S. peruvianum* based on all polymorphic sites.

Introduction

Divergence population genetics (DPG) is concerned with how speciation events occur. To draw conclusions on the history of populations and species, DPG uses molecular data through a population genetics approach (Avise 1989). The idea that species might evolve, *i.e.* that current species have arisen from previous ones, was already under discussion at the time Darwin published his evolutionary theory (Darwin 1859). However, he was the first to offer a mechanism for this to take place. According to Darwin's theory, individuals more suited to the environment are more likely to survive and more likely to reproduce. This slow process results in populations that adapt to the environment over time, and ultimately form new species. The underlying genetic basis for the adaptive trait does not arise because of the environment; the genetic variant pre-exists in the population and becomes subsequently selected because it provides the bearer of that variant a fitness advantage.

In this thesis, we focus specifically on a group of phylogenetically closely related species (Figure 1) of the wild tomato species complex and apply the DPG approach. Wild tomatoes inhabit a vast range of climatic (*e.g.* 50mm to over 4m annual precipitation), biogeographical and environmental habitats. They are natural to North-west South America ranging from Ecuador to Chile along the western Andes. Additionally, one species is endemic to the Galapagos Islands. They are indeed found in temperate deserts as well as wet tropical rainforests, and each species seems to have a particular geographical distribution within this environmental diversity (Rick 1973; Rick et al. 1978; Rick et al. 1979; Taylor 1986; Peralta & Spooner 2001). Classical common-garden studies document the presence and magnitude of a variety of morphological and physiological differences within and between species (reviewed in Taylor 1986). Considering the geographical and geological structure (*e.g.* sea-level to 3000m elevation) characteristic of their natural range due to a dynamic recent geological history (Young et al. 2002) and the unique environment of islands, it is sound to say abiotic ecological conditions play a critical role in these species' phenotypic evolution and speciation. Many researchers have indeed identified traits that are putative adaptive responses to species' habitats in wild tomatoes (*e.g.*, Rick 1973; Rick et al. 1976; Patterson et al. 1978; Rick et al. 1978; Vallejos 1979; Bloom et al. 2001; Nakazato et al. 2008).

In the following sections of this introduction, I will give a brief overview of the theoretical framework underlying the present study (*i.e.* Coalescence Theory). Then I will introduce some concepts related to how new species originate (*i.e.* Speciation). Afterwards, the effects of Spatial Structure on Populations will be described. Finally, the system we are using (*i.e.* Wild Tomatoes) will be introduced and the Scope of this Thesis will be presented.

1. Coalescence Theory

Polymorphisms in DNA occur due to mutation. All copies of a specific DNA site in a species are related to each other and descend from a common ancestor. This process can be visualized through a genealogical tree. Polymorphism at such site is due to mutations occurring on the branches of the genealogical tree and the frequency of each sequence variant (allele) is equal to the proportion of branches that inherits the allele. Therefore, the history of the descent of lineages (*i.e.* the process that gave rise to the tree) as well as the mutational history are reflected in the pattern of polymorphism. For example, if we would sequence a DNA region from a number of randomly chosen individuals and find no polymorphism, we could conclude the region is under purifying selection. Another option would be to consider that the individuals are unusually closely related. To make a decision we could make assumptions about the process that gave rise to the data (see more about this example below).

1.1. Why use the Coalescent?

One can use a phylogenetic method to determine the pattern of descent between the different species, *i.e.* the process of species evolution also known as speciation (Mayr 1942), which is a tree-like process. Since DNA sequence evolution recapitulates the evolution at the phenotypic level, nowadays it is only needed to analyze DNA sequence to draw conclusions about the evolution of species. Only a single DNA sequence from each species would be needed to build a gene tree, since typically the gene tree is assumed to represent the species tree.

However, this approach is incomplete in studying recent speciation events, since the diversity within species is ignored. For example, the same approach cannot answer questions about the demographic scenarios under which species have evolved. In such case, we might find ourselves dealing with migration between populations or with a population history that is not tree-like. Furthermore, using different genes might produce different trees (Rosenberg & Nordborg 2002). Researchers have thus to consider a different strategy to infer the genealogy of populations from estimated trees. The tree then becomes of no interest itself, but a way to estimate the parameters of the random genealogical process that has given rise to each possible sampled sequence. Therefore, in order to look into the demographic history of a species researchers use a genealogical approach also called the “Coalescent approach”, rather than a phylogenetic one.

Going back to the example at the beginning of this section, using the coalescent we could simulate many random repetitions of the evolutionary process. If the fraction of the random genealogical and mutational histories that could give rise to the observed data is small, we could conclude that our assumptions cannot explain the pattern. Hence, the interpretation of data depends on the genealogy of the sequences, which is unknown. Because of this, we treat the genealogy as random as well as we do mutation. Just as mutations occur differently across runs of evolution (Luria & Delbrück 1943), if evolution were repeated, samples from different ‘runs’ of evolution would have different genealogical trees.

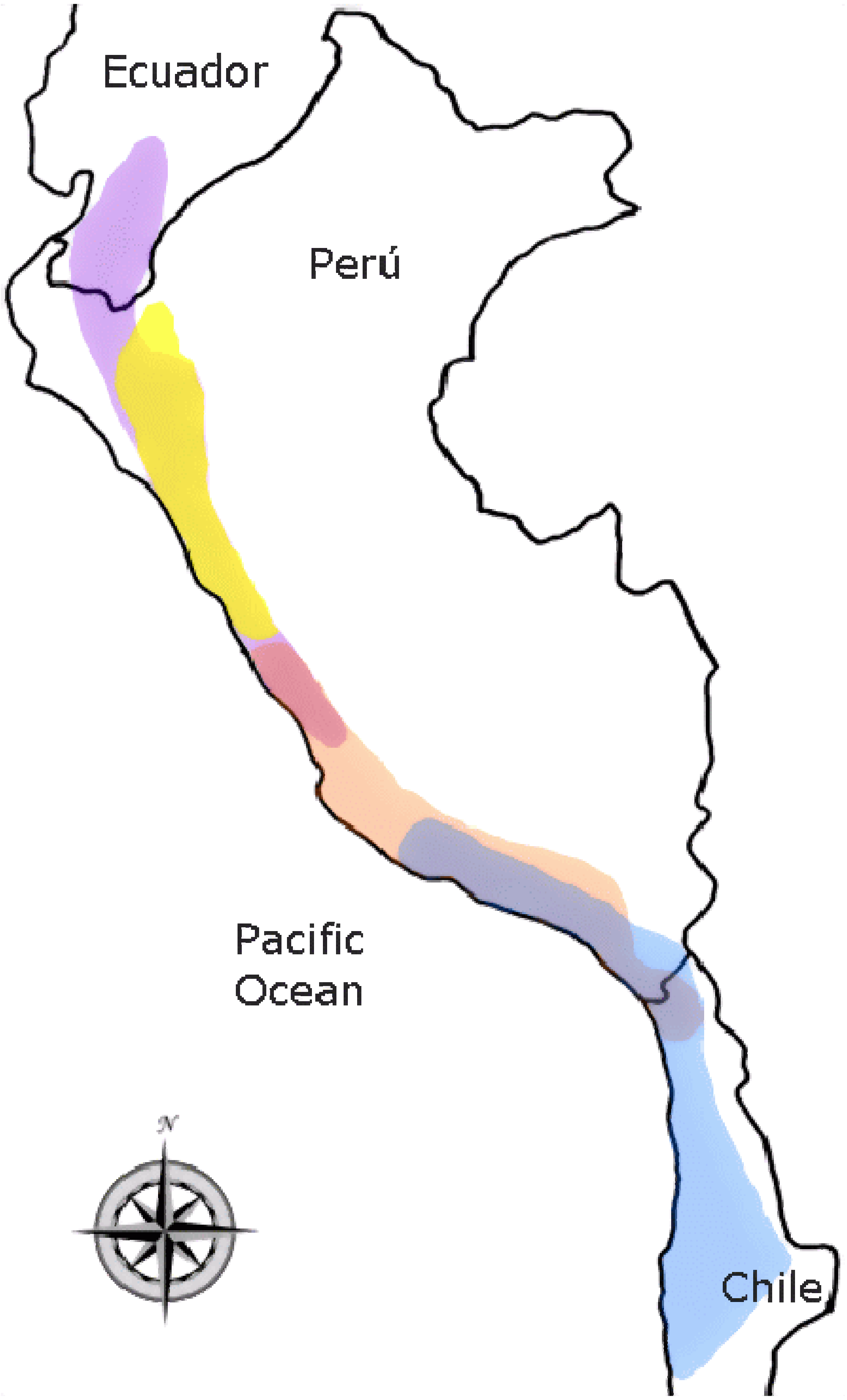


Figure 1 Map of West South America showing the distribution of four wild tomato species: *S. habrochaites* (purple), *S. arcanum* (yellow), *S. peruvianum sensu stricto* (light salmon) and *S. chilense* (light blue). Note the overlap in distribution.

1.2. The Standard Neutral Model

To be able to answer such questions using a genealogical approach, a simple population model has been first proposed. Two authors (Wright 1931; Fisher 1930) independently described what has come to be the standard simple population model. The Wright-Fisher model (hereafter WF model or standard neutral model) assumes a single population of constant size that has persisted for a very long time (for mathematical purposes, an infinite amount of time). Additionally, the model also assumes that individuals are mating randomly within the population (panmixia). Other two conditions of the model are non-overlapping generations (*i.e.* individuals die after reproducing) and a random number of offspring (which follows a Poisson distribution). Under these circumstances, the population is well represented by its effective population size (N_e). The effective population size (Wright 1938) is defined as the size of an ideal population (such as that described by the WF model), that undergoes the same amount of genetic drift as the real population under consideration.

1.3. Deviations from Neutrality

Real observed populations from which we sample DNA of various individuals, almost never meet the conditions of the Wright-Fisher model. The evolutionary forces shaping the observed genetic diversity are of course natural selection acting on a population, but also demography, *i.e.* random population size changes. However, one can distinguish these forces based on genome wide polymorphism data. In fact, demographic factors act independently from natural selection and can shape the genetic variation across the entire genome by affecting the effective population size of the population or species. Such processes are presumed to increase of decrease nucleotidic diversity in all loci at the same rate. Some generic examples of demographic factors include events that randomly change population sizes and species localities such as bottlenecks, habitat fragmentation and range expansion. These changes lead to random loss or gain in the spatial distribution of taxa, resulting in increasing or decreasing numbers of individuals in the population. Natural selection on the other hand, acts either by eliminating deleterious alleles, fixating advantageous ones or maintaining polymorphisms when each allele is advantageous under different conditions. This, of course, takes place throughout many generations in which individuals with better-adapted alleles have higher fitness. Therefore, natural selection acts on each gene individually, instead of acting on the whole genome.

1.4. Defining the Coalescent

Thus, the coalescent is a genealogical approach, which comes as a natural extension of the classic population-genetics theory and models such as that discussed above (Nordborg 2001). It was discovered independently by several authors (Malecot 1973; Malecot 1975; among others) in the 1970's although the definite treatment came from (Kingman 1982). In the coalescent, sampled lineages, randomly "pick" their parents in the previous generation. Every time two lineages pick the same parent, they are said to coalesce. This is assumed to happen in the simplest case in the absence of selection, and eventually all sampled lineages coalesce into one single lineage called the most recent common ancestor (MRCA) of the sample. The time it takes for all sampled lineages to coalesce depends on the size of the population (the bigger the size of the population, the longer it takes). This means the coalescent can handle different demographic scenarios. For example, the coalescence can be modeled to fit a range expansion with subsequent subdivision, in which the time to the MRCA would be longer than if the population had recently undergone a bottleneck.

1.5. The Frequency Spectrum

As it is computationally difficult to keep full track of entire sequences, data are usually summarized by essential information derived from the simulated coalescent tree. Commonly used statistics include Tajima's D, the population mutational parameter (θ), the effective population size (N_e), the frequency of each polymorphic site in the sample, among others. The distribution of the frequencies of all polymorphic sites in a sample is thus described as the number of singletons, doubletons, etc., in a sample. This distribution is formally called the Site Frequency Distribution or Frequency Spectrum (FS) when applied to a single population or species and, the Joint Frequency Spectrum (JFS) when applied to two species or populations (Figure 2), taking into account the frequency of each allele in each population. The shape of such distribution provides qualitative information on the processes involved in the history of the sample by comparing to the expected FS under the classic WF model (Hey & Machado 2003; Braverman et al. 1995; Fu & Li 1993). Demographic processes affect the FS. For example, past population growth (*e.g.* range expansions) results in an excess of low-frequency alleles when compared with neutral expectations (Nei et al. 1975). Tajima (1989) confirmed this conclusion and examined other effects of such demographic processes on the FS.

1.6. The Effect of Recombination

Besides mutation, another factor to take into account in the coalescent is recombination. Recombination allows two physically linked genes to exchange the alleles to which they are connected. This in turn allows linked sites to have different genealogical trees. Thus, a sample of recombining sequences can be seen as a “walk through tree space” (Wiuf & Hein 1999). As one “walks” from one end to the other of a sequence, one finds himself in different trees. The trees, however, only change gradually. This happens because each recombination event only affects a subset of the branches of the tree, changing only the topology of this section of the tree. Recombination, thus, has profound effects on the coalescent (Figure 3), in that the coalescent with recombination does not generate a random tree, but rather a random graph (Nordborg & Tavare 2002; Hudson 1983; Griffiths & Marjoram 1996). This complication is, however, readily incorporated into the model. The extent to which the histories of different sites are correlated depends on the recombinational distance between them, which is a function of the frequency with which recombination occurs between the two of them. The further away two loci are located, the highest the recombinational rate between them. As recombination approaches infinity, the genealogies of such loci are conditionally independent, given the historical demography of the group under consideration. This is particularly important in outbreeding plant species, where reduced recombination rates are not observed along the whole genome but only in certain regions, such as around centromeres.

Apart from mutation and recombination, many other factors can be included in the model (Nordborg 2001). Some phenomena, such as variation in reproductive success, age structure, seed banks (Kaj et al. 2001) and skewed sex ratios, change only the rate of coalescence (Rosenberg & Nordborg 2002). Other factors such as population structure or fluctuation in population size, however, also change the shape of genealogical trees. The only factor that causes real, although not insurmountable (Kaplan et al. 1988; Neuhauser & Krone 1997; Slatkin 2001), difficulties is selection. By definition, under selection, some genotypes reproduce more than others, which means that, going back through time, lineages do not randomly pick parents. In the current study, we do not use coalescence methods with selection.

1.7. The Coalescent and Summary Statistics

Summary statistics are the result of functions applied to a data set, which represents much of the information in the data. For a set of DNA sequences, one commonly used summary statistic is S, which represents the number of variable sites in the sample. Summary statistic methods make no use of the genealogy that underlies a data set. They begin, not with an evolutionary tree, but by summarizing some aspect of the data. However, the coalescent can be used to calculate summary statistics. The advantage of this is that summary statistics are often easier to use to fit models to data than would be the case with the data itself. For example, the difference between the average number of pairwise differences in a DNA sequence sample and the total number of observed mutations that is predicted by the basic coalescent model is just in the neutrality test based on Tajima's D statistic (Tajima 1989). Thus, the coalescent can be used to design statistical tests of models of evolution. It is often possible to show mathematically how departures from the standard model, such as those caused by population structure or selection (Nordborg 2001), affect the test statistic, which makes it possible to interpret the observed deviation.

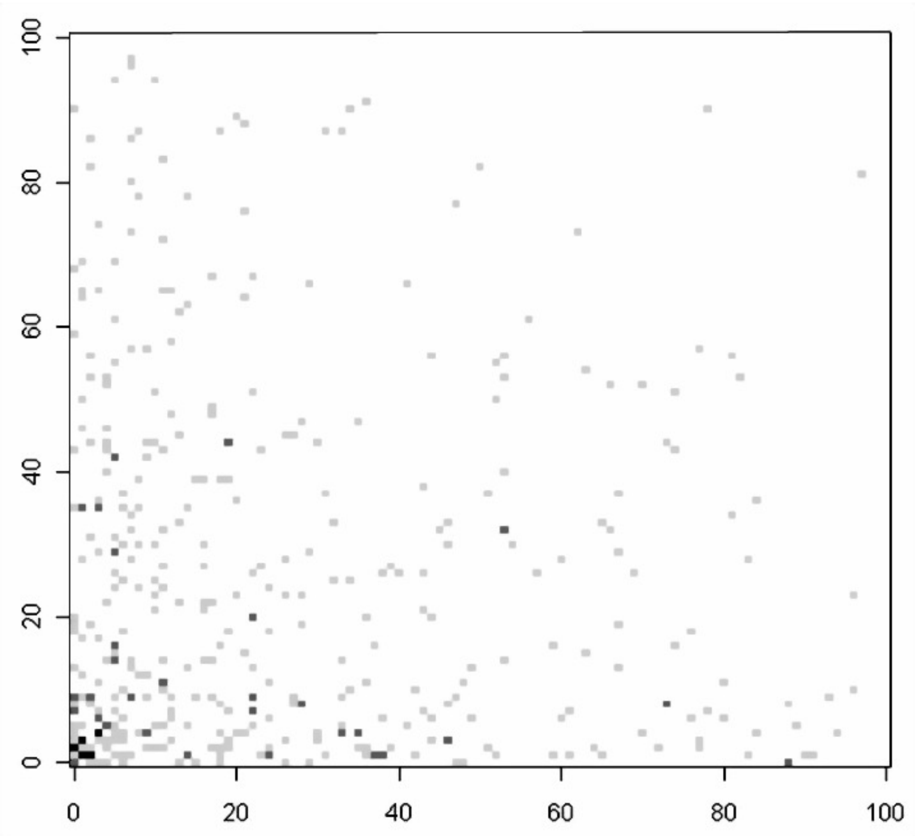


Figure 2 Hypothetical Joint Frequency Spectrum. The X-axis represents the relative frequency of any given allele in population 1. The Y-axis represents the relative frequency of the same allele in population 2. The intensity of the square represents the number of alleles at a given frequency in populations 1 and 2.

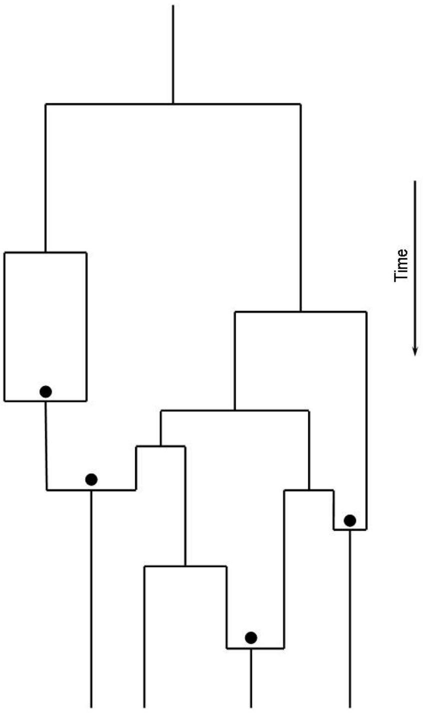


Figure 3 Ancestral recombination graph. Recombination events are represented by closed circles. The topology of the tree is thus affected by presenting less external branches than expected without recombination. Adapted from Griffiths & Marjoram (1996).

1.8. The Coalescent and Likelihood

In most cases, to calculate model parameters, researchers use coalescent simulations going backwards into time from the present (when the sample is collected). Therefore, it is as a simulation tool that the coalescent is most widely used (Hudson 1990; Nordborg 2001). Additionally, samples that are simulated under various models can be combined with data to test hypotheses. The canonical approach to do this, developed mainly by Hudson and colleagues (Hudson 1990; Kreitman 2000; Nielsen 2001), can be described as follows. If a pattern of polymorphism in a data set is found, one might want to know whether, for example, it is the result of a neutral process or if it is the result of selection. To do this one could simulate several possible datasets under the same null model (which in this case should not include selection) and calculate summary statistics from them. One can then compare the distribution that is obtained from the simulated data, to the values that are obtained from one's sample. If one's sample values are very rarely seen in our simulations, the null model can be rejected. If the null hypothesis (such as neutral evolution) is rejected, one can propose an alternative hypothesis to explain the observed pattern, such as selection or a more complex demographic history of the populations.

A well-known example of using coalescent simulations to test a hypothesis is provided by Takahata et al. (2001). The hypothesis tested in their study was the multiregional model of the origin of humans (Mountain & Cavalli-Sforza 1994). The approach they used was to simulate data to be compared with data from ten human loci. Their model included migration between three subpopulations — African, European and Asian — followed by their divergence. They simulated many genealogies and for each, they determined the position of the MRCA. In the empirical data, nine out of ten loci had an African ancestor (and the tenth locus had only a single polymorphic site). This pattern was found to be highly unlikely under the investigated multiregional model, unless the African population size was much larger than the Asian and the European. Therefore, they rejected the multiregional model and favored an out-of-Africa origin for humans.

Coalescent simulations can also be useful in study design, for example, to determine the number of loci, or the sampling scheme that need to be used for testing a given hypothesis. Additionally, simulated samples help to evaluate the performance of new statistical tests (Wakeley 1996). This is especially true when methods are developed before the appropriate data to which they can be applied are available. In this case, coalescent simulations conveniently provide data sets on which new methods and their statistical power can be tested. This approach can be valuable whether or not the proposed tests are based on coalescent theory (Pritchard et al. 2000). Finally, one of the most remarkable aspects of the coalescent is that it allows full likelihood analysis of evolutionary models (Stephens et al. 2001). In theory, all one needs to do is evaluate the likelihood equation (Felsenstein, 1988) for our data and for our favorite models:

$$L = \sum_{G} P(D \mid G, \mu) P(G, \alpha)$$

where *L* is the likelihood (the probability of the data given the parameters), *D* is the data (typically DNA sequences), *μ* is the collection of parameters in the mutation model and α is the collection of parameters (such as population sizes and migration rates) for the population process. The tree or genealogy, G, is a so-called nuisance parameter, which we remove by averaging the likelihood over all possible values (Hey & Nielsen 2007). In the event that features of G (or G itself) are of interest, it is more natural to treat them as random variables than as parameters to be estimated (Donelly and Tavare, 1995).

Unfortunately, this is not easy in practice, because summing over all possible genealogies turns out to be exceedingly difficult. A promising alternative is to use approximate methods based on summary statistics (Tavare et al. 1997; Weiss & von Haeseler 1998; Pritchard et al. 1999; (Wall 2000). Instead of numerically evaluating the equation above, the full data set is replaced by a set of summary statistics. Data are then simulated using various parameter values, and each simulation is accepted or rejected according to how good the summary statistics of the simulated data are compared to the real data. For a given set of parameter values, the likelihood is then approximated by the proportion of simulations that are accepted. This summary-statistic approach has great potential for the inference of parameters under models for which complete evaluation of the equation is intractable. Important for its success is the extent to which values of summary statistics capture the information in full data sets. Careful choices of such statistics will be needed for likelihood computations in complex models. Typically, the use of such summary statistics is at the heart of the ABC (Approximate Bayesian Computation) methods.

1.9. The Coalescent in our Study

In the present study, we are first interested in the neutral process of evolution and how they shape the diversity observed. The second part of the thesis will deal with the speciation process between the various species of the wild tomato clade. As a starting point, the isolation model (Wakeley & Hey 1997) will be used to estimate historical population parameters. Using the observed polymorphism, four population parameters will be estimated: θ for *S. arcanum*, θ for *S. habrochaites*, θ for the ancestral population, and the scaled speciation time τ (see *e.g.*, (Stadler et al. 2008). To analyze this model we will use coalescent simulations (Hudson 1990). Thus, we will be able to see what the model can tell us about speciation of the species studied. We will also be able to assess the utility of the model in explaining the observed patterns of polymorphism.

Given the fact that previous studies in our group (Arunyawat et al. 2007; Stadler et al. 2008) have provided us with sequence data for other two wild tomato species, we will analyze four species (*i.e. S. peruvianum*, *S. chilense*, *S. arcanum* and *S. habrochaites*). Thanks to this, many pairwise comparisons can be done. These will be carried out using the original isolation model. Thus, we will be able to evaluate several speciation scenarios by comparing different pairs of species (*e.g.* sympatric vs. allopatric). Additionally, we will be able to test the hypothesis of speciation between *S. arcanum* and *S. peruvianum* which used to be regarded as a single species taxon (see below). The large amount of data made available by this project and U. Arunyawat's doctoral thesis combined with the statistical tools developed in our department, will allow us to get a better idea of the population genetics of wild tomatoes.

2. Speciation

2.1. The Concept of Species

Before discussing the speciation process, we shall define the concept of species. Even though it was Darwin who advanced the concept of the evolution of species (intrinsically related to that of speciation), the concept of species was already being debated. In his seminal work, Darwin (1859) himself wrote about the difficulties associated with the definition of species. It took almost a century until Mayr (1942) introduced a new use of the concept when it came to defining a species. Instead of focusing on trying to find a definition of the concept of species that could satisfy all researchers and be used in all cases, Mayr elevated several different approaches to species identification to the level of concept.

While the general concept of species is hard to define, the definition of species in plants has its own intricacies (Rieseberg & Willis 2007). Some botanists even doubt the existence of plant species because of the evidence for interspecific hybrids (Arnold 1997); and the impossibility to readily assort phenotypic variation into discrete categories in some plant groups (Mishler & Donoghue 1982). Moreover, some had claimed that populations are the most inclusive reproductive unit instead of species, because gene flow between populations of the same species can be very low. Recent works seem to contradict these claims. Rieseberg et al. (2006) used morphometric data from more than 200 plant genera to demonstrate that discrete clusters of morphologically similar individuals occur within most sexual plant lineages and that these clusters correspond closely to groups with significant post-pollination reproductive isolation. Furthermore, the same study concludes that interspecific hybridization is not the primary cause of poorly defined species boundaries. Additionally, molecular population genetics studies (Morjan & Rieseberg 2004) imply that earlier direct estimates of migration rates were too low and that actual migration rates do not differ, on average, from those of animals. What is more, both theoretical (Whitlock 2003) and empirical work (McDaniel & Shaw 2005) further indicates that even in species with low gene flow, populations may evolve in concert through the spread of advantageous alleles.

2.2. Reproductive Isolation

Thus, for the present work, we use the biological species concept (*i.e.* considering a species a set of actually or potentially interbreeding populations capable of creating fertile offspring). It is clear from this definition that different species cannot freely cross and produce fertile progeny. This is due to the existence of multiple reproductive barriers isolating them. Reproductive isolation is mediated by two kinds of reproductive barriers: pre-zygotic and post-zygotic. Pre-zygotic barriers can be further classified in pre-pollination and post-pollination. Pre-pollination barriers (called pre-mating barriers in animals) limit the transfer of gametes (*i.e.* pollen) from one species to the other (Kaneshiro 1980). Some pre-pollination barriers (mechanical, ecogeographic and temporal) are also found in animal species, while pollinator isolation is exclusively found in plants. As for post-pollination barriers, typical examples include con-specific pollen precedence (the advantage of con-specific pollen over non con-specific pollen in fertilizing eggs) and gametic incompatibilities (failure of non con-specific pollen to fertilize eggs). Finally, examples of post-zygotic barriers are hybrid inviability, hybrid sterility and hybrid breakdown (sterility or reduction of fertility in subsequent generations).

It is of particular importance to estimate the relative contribution of different reproductive barriers in limiting gene flow among contemporary populations and to determine the order and speed with which they arose. All else being equal, pre-zygotic barriers will contribute more to isolation than post-zygotic barriers (Ramsey et al. 2003). Among the former, ecogeographic, pollinator and mating system isolation are the most important among plants. For example, Rieseberg et al. (2006) used artificial crosses to test the contribution of pre- and post-zygotic barriers to isolation between two species. The production of hybrid seeds (overcoming pre-zygotic barriers) and fertility of first generation hybrids (overcoming post-zygotic barriers) were measured. However, since the effect of pre-zygotic barriers forcefully precedes the effect of post-zygotic barriers, reduced hybrid seed production is found to contribute 75% of the total isolation caused by both barriers.

As for the speed and the order in which barriers arise, there is not much known yet. Although individual reproductive barriers can arise rapidly, most plant species remain separated by numerous barriers. This implies that complete speciation typically requires many thousands of generations. The main exceptions to this are hybrid and polyploid speciation. Fully isolated polyploid species may arise in one or two generations, and diploid or homoploid hybrid species may achieve isolation in as few as 60 generations (Ungerer et al. 1998) Other genetic studies on post-pollination barriers have focused on self-incompatibility (SI) mechanisms. SI mechanisms prevent the pollen of an individual from pollinating its own eggs in some hermaphroditic species. Early observations have shown that SI species are less compatible in interspecific crosses than self-compatible (SC) species. Thus, SI may also contribute to interspecific incompatibilities, as confirmed in a study by (Bernacchi & Tanksley 1997). In this study using a cross between *Solanum lycopersicum*, and *S. habrochaites,* a SI locus was found to co-localize with a detected inter-specific incompatibility QTL and that crosses between SC species fail after transformation with a SI gene from a SI species. Another study (Igic & Kohn 2001) concludes that diversification of genes that contribute to SI appears to result from frequency-dependent selection. Interestingly, other plant reproductive proteins appear to be under positive selection as well, including candidates for species-specific recognition between pollen and stigma.

As for post-zygotic barriers, chromosomal rearrangements may be another cause of them. However, according to population genetics theory, strongly underdominant chromosomal rearrangements (those that reduce the fitness of heterozygotes) cannot fix in a population because of their negative effect on fitness, unless the population were small and inbred. Weakly underdominant rearrangements are more easily established but contribute little to reproductive isolation. In contrast with underdominant chromosomal rearrangements, the Bateson-Dobzhansky-Muller (BDM) model accounts for the accumulation of interspecific incompatibilities in genes without loss of fitness (*i.e.* divergence without initial isolation). According to this model, in the course of time, geographically isolated or neighboring populations may accumulate distinct mutations. These mutations are compatible with the ancestral genotype but are incompatible between populations. Generally, these incompatibilities involve two or more loci, although it is theoretically possible that they result from accumulation of independent mutations in a single locus. The effect of these incompatibilities can be hybrid inviability or hybrid weakness. The latter is often manifested as necrosis in developing seedlings or adult plant tissue, similar to the phenotype of hypersensibility responses to pathogen attacks (Bomblies & Weigel 2007). These observations suggest that hybrid weakness may also result from the accumulation of mutations in pathogen resistance genes, which diverge in different populations in response to selection pressure exerted by pathogens.

While the BDM model describes how natural selection can play only an indirect role in the evolution of reproductive barriers, by bringing about trait changes that inadvertently prevent gene flow between diverging populations, the concept of speciation by reinforcement is the opposite: under reinforcement, natural selection directly favors the evolution of barriers to mating between incipient species. This model, like the BDM model, imagines two highly diverged populations that have accumulated some degree of genetic incompatibility while isolated from each other. Nonetheless, because genetic differentiation between the groups is incomplete, when they co-occur in sympatry, less fit hybrids can be formed. Natural selection will thus act directly to ‘reinforce’ the partial isolation between two groups by favoring traits that reduce inter-type mating. Although the frequency of speciation by reinforcement continues to be debated, there is recent solid theoretical as well as empirical support for this mode of speciation in a few well-described cases (Servedio & Noor 2003).

2.3. Speciation and the Wakeley-Hey Model

Reproductive isolation is thus not the cause of speciation but a consequence; resulting from divergence between species, which in turn is the result of either diversifying selection, genetic drift or a combination of both. The evolution of reproductive isolation based on several or many loci is what (Mayr 1963) refers to “as gradual speciation”. Considering the geographic setting in which it can take place it can be classified in three modes of speciation: allopatric speciation, parapatric speciation and sympatric speciation. Allopatric speciation is the evolution of reproductive isolation between populations separated by a geographic barrier, which prevents interspecific gene flow. It can be further divided into allopatric speciation by vicariance and peripatric speciation. In allopatric speciation by vicariance, a physical barrier divides a widespread species into two populations which diverge through time. Peripatric speciation is also known as founder’s effect speciation, by which a localized colony diverges from an ancestral species which remains almost unchanged. In opposition to allopatric speciation, sympatric speciation is the evolution of reproductive barriers between subsets of a single initially randomly mating population. An intermediate mode of speciation is parapatric speciation, in which neighboring populations of a widespread species, between which there is limited gene flow; diverge by adaptation to different environments. Studying these modes of speciation and their consequences on the observable genetic diversity is at the heart of evolutionary biology.

A model often used to look into divergence between populations or species is the Wakeley-Hey isolation model (Wakeley & Hey 1997), which allows estimating historical population parameters. The underlying model (Figure 4) includes one ancestral population or species which at some point in the past divided in two extant populations or species. As populations diverge from each other, new mutations particular to each of the new populations arise. However, there is also a certain amount of shared polymorphism reminiscent of the pre-divergence period. Comparing polymorphism between the two extant species, segregating sites can be classified into four mutually exclusive categories:

1) sites that share polymorphisms in both species (S_s),

2) sites monomorphic in one but polymorphic in the second one (S_1),

3) sites monomorphic in the second species but polymorphic in the first one (S_2), and

4) sites showing fixed differences between species(S_f).

The best way to directly observe these polymorphisms is using the JFS (see above). Wakeley and Hey (1997) calculated the expectations for each of these four categories of polymorphism given the set of parameters summarizing the divergence process, namely the time from divergence (τ) and the population mutational parameter (θ) for both extant populations as well as the ancestral one (θ_1, θ_2, θ_a). Wakeley and Hey derive thus a statistical method to infer those parameters using the JFS with sequence data from the two extant species or populations. This model can also be used when there is gene flow between the diverging populations. Here, the question arises of how can divergence occur with gene flow between populations? The general answer to the puzzle is that divergence can happen at some genes, even if there is gene flow for other genes. Such a process occurs as follows: F1 hybrids between species carry a full set of genes from each species, but F2 backcross hybrids do not. Thus, it is possible for some genes to pass between species through F2 backcross hybrids depending on which genes they carry.

Unfortunately, to make things more complicated, when one population separates into two, genetic variation can still be shared for some period of time even in the absence of gene flow (Avise 1975; Pamilo & Nei 1988; Hey 1991). Genealogies are more likely to coalesce within species if divergence times are longer (Wakeley & Hey 1997). If the sizes of both populations are large, and gene trees are deep within populations, then genealogies and genetic variation might be shared at some genes for very long periods of time, possibly even after the populations have diverged and become reproductively isolated species. This means that it is possible to sequence a copy of a gene from one species and find that it is more similar to a gene from a closely related species than it is to another copy of the gene from the same species. This will happen simply by chance if the populations separated only recently, even if there is no gene flow. Thus, an important challenge is to determine whether or not genetic variation that is shared by both populations is simply a remnant of variation in the common ancestor or if it is due to gene flow after the population started to separate.

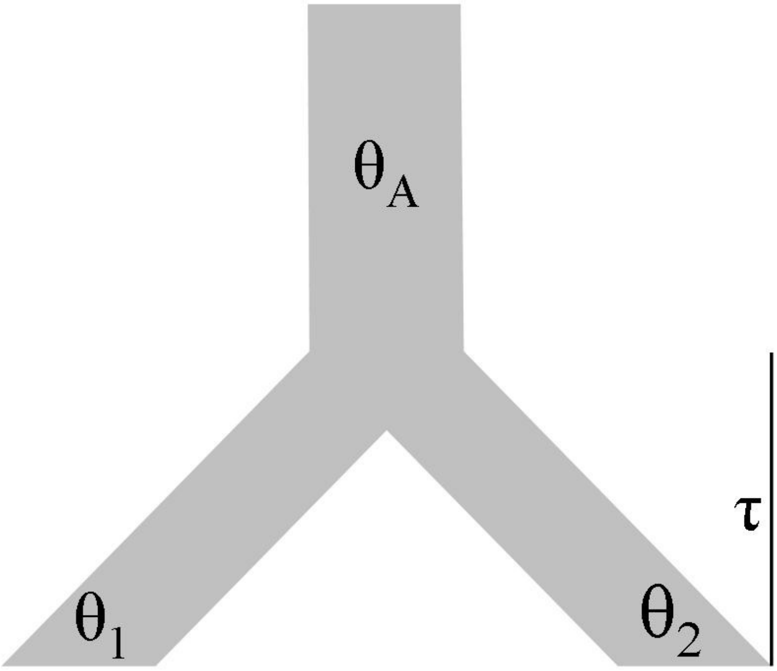


Figure 4 The Wakeley-Hey isolation model. θA, θ1 and θ2 are the population mutational parameter for the ancestral population and for extant populations 1 and 2, respectively. τ is the time from divergence.

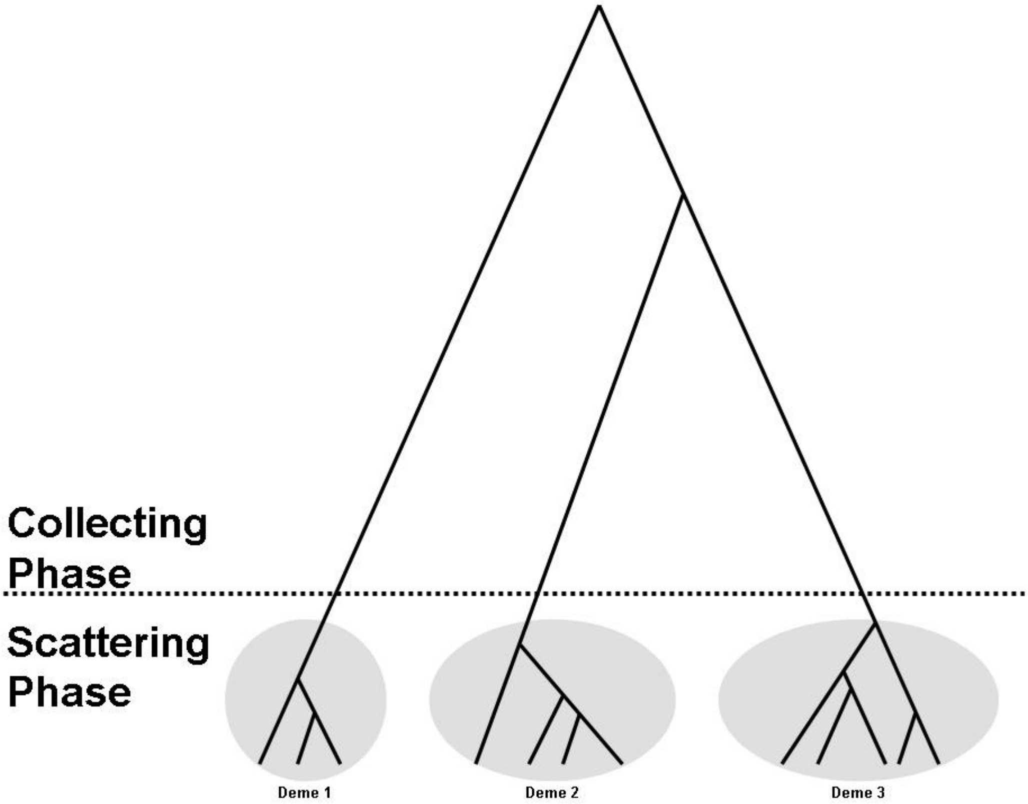


Figure 5 A hypothetical genealogy with three demes showing the scattering phase (characterized by local coalescent events) and collecting phase (where only one lineage is present in each deme).

To handle this problem, one looks then at different genes between the two populations studied. A history of divergence with gene flow is generally indicated if there is variance among genes for divergence, such that the variation in divergence among the different genes is greater than expected under a model without gene flow (Wakeley & Hey 1997). Although gene flow is considered to be a homogenizing process that prevents divergence, very recent results in insects (Turner et al. 2005; Shaw 2002; Emelianov et al. 2004) and the Hawaiian silversword (Lawton-Rauh et al. 2007a; Lawton-Rauh et al. 2007b), indicate that some gene flow continues to occur despite ecological and physiological speciation among taxa. These findings seem to suggest that gene flow is a common feature of the early stages of the divergence process. These results using population genetic methods contradict early views that gene flow is rare or non-existent between populations that continue to diverge and become species (Mayr 1963). However, it is maybe too soon to appreciate just how frequently divergence and speciation happen with, and without, gene flow (Hey 2006).

3. Spatial Structure of Populations

While we might not be sure just how often speciation or divergence occurs under gene flow, it is generally considered that by far the most likely, and most explicable, form of speciation occurs when populations diverge from each other while separated by an external barrier to gene flow, such as simple physical distance. It is generally recognized that most, if not all, species are found in fragmented habitats with populations separated by physical distance rather than as a unique panmictic population (Olivieri et al. 1990). This is described usually as a metapopulation, *i.e.* the network of demes or populations connected with each other by migration of individuals or gametes. When isolated from each other, populations accumulate genetic changes, which can be adaptive or not as genetic differences can also accumulate purely through random sampling processes (genetic drift). One way or the other, spatially isolated populations in a given species can present distinct patterns of polymorphism, *i.e.* population structure. Spatial structure of populations has important consequences on neutral as well as selected genetic diversity (Tellier et al. 2005) as selection and random process differ from that of a single panmictic population.

3.1. Models of Population Structure

There are several models used to theoretically study the influence of spatial structure within a studied species. The island model of migration (Wright 1940) considers a mainland population with migration to one or more island populations. Island models can vary widely in numbers of populations, sizes of populations and rates of gene flow. They are useful for understanding the effects of small population size and limited gene flow on rates of genetic drift and levels of divergence between island populations. Another prevalent model is the Stepping-stone model (Kimura & Weiss 1964). Unlike the island model, this one specifically includes a spatial element, only allowing adjacent populations to exchange genes with each other. Stepping-stone models can be one-dimensional, two- or three-dimensional. The isolation by distance model (Malecot 1969; Wright 1943) is a stepping-stone model taken to the extreme. In this model, every individual is restricted in the distance that its genes can travel (usually a short distance on average but see Wingen et al. 2007). If a population is evenly distributed over a landscape and movement per generation covers a short distance on average, then individuals are much more closely related to nearby individuals than to distant individuals. Finally, since population structure can vary both spatially and temporally (Hedrick 2006), the metapopulation model (Wade & McCauley 1988; Slatkin 1977), takes into account not only structure in space but also in time, *i.e.* the founding and extinction of entire populations. In this model a metapopulation is composed of several demes which can change in size, divide in new demes or disappear as time goes by. Note that metapopulation structure with limited gene flow increases the effective size of the metapopulation compared to a single panmictic population with identical number of individuals (Laporte & Charlesworth 2002). However, extinctions and recolonizations are shown to decrease the effective size of the metapopulation by increasing the probability of identity by descent (Wang & Caballero 1999).

3.2. The Effects of Structure

The aforementioned spatial and temporal shifts in population structure lead to changes in the probability of gene flow among populations. Such changes in gene flow impact the rate of random versus non-random association among polymorphisms (linkage disequilibrium, LD). Because population structure can have such a strong effect on the patterns of polymorphism and LD, research has focused on distinguishing demography from natural selection. For example, surveys of nucleotide diversity in the wild ancestor of maize, *Zea mays ssp. parviglumis,* have revealed significant population genetic structure (Moeller et al. 2007). This influenced observed patterns of nucleotide polymorphism depending strongly on the geographic region from which subpopulations were sampled, probably due to the demographic history of subpopulations in those regions. Overall, these results suggest that explicitly accounting for population structure may be important for identifying loci that have been targets of selection.

Furthermore, the partitioning of sequence variation among subpopulations (*i.e.* structure), can have important effects on statistical tests of the neutral equilibrium (NE) model based on a single panmictic unit, even when the majority of polymorphism is harbored within populations (Moeller et al. 2007). In particular, species-wide samples (*i.e.* collected throughout the whole species range) may often contain an excess of rare variants because multiple subpopulations each contain singleton polymorphisms (Hammer et al. 2003). Similar to the patterns identified in *Z. mays ssp. parviglumis* (Moeller et al. 2007), molecular population genetic studies in humans have suggested that sampling across multiple subpopulations influences the estimates of the extent and pattern of nucleotide polymorphism (Ptak & Przeworski 2002). Thus, when nucleotide variation is structured among subpopulations, the sampling strategy clearly influences the estimation of population genetic parameters and inferences about natural selection and demographic history (Stadler et al. 2009). In short, disentangling the effects of demographic history from those of positive selection under population structure is much harder than previously assumed.

To better understand the effect of the sampling strategy on the estimation of population parameters, it is necessary to look into the coalescent process of a metapopulation. When the number of demes in the metapopulation is large, the genealogy of a sample includes two phases, called the scattering phase and the collecting phase (Wakeley 1999). The scattering phase comprises the very recent history of the sample, during which coalescence occurs mainly locally in each deme. Recent migration events are shown by migrants coalescing in their deme of origin during this phase. At the end of the scattering phase, *i.e.* the start of the second, and much longer, collecting phase, one finds every lineage is present only in one deme. Migration events then move lineages from deme to deme, until a pair of lineages fall into the same deme, at which time they can coalesce. Note that the collecting phase can be characterized as a Kingman coalescent type of process (Wakeley 2000).

The ability to break the genealogy into these two parts, and to consider them separately, depends only on the number of demes being large. When this is true, the time spent in the scattering phase can be ignored because the collecting phase dominates the history (Wakeley 1998). The parameters that determine the pattern of genetic variation in a sample are, in addition to the size of each deme and the number of demes, the rates of migration and extinction/recolonization and the founding-propagule sizes for each sampled deme (Wakeley & Aliacar 2001). One of the objectives of this thesis is to study how the sampling strategy reveals the pattern of genetic variation detected in a metapopulation. For example, in comparison to species-wide samples, local samples (and to a lesser extent, pooled samples) are influenced by the scattering phase of the coalescent process, resulting in shorter external branches in proportion to the whole coalescence tree and hence lower proportions of singletons within each deme. This has consequences for using inference methods to detect selection. For example, the sampling scheme impacts Fu and Li's D more than it does Tajima's D.

Now that the theoretical background and the basic key evolutionary questions are presented, I will describe our study system of the wild tomato data, and then describe in more details the precise questions addressed in this thesis.

4. Wild Tomatoes

Solanum Sect. *Lycopersicon* is a relatively small monophyletic clade within the large and diverse *Solanaceae* family (D'Arcy et al. 1979). It consists of 13 closely related species or subspecies including the domesticated tomato, *Solanum lycopersicum* (formerly *L. esculentum*) (Spooner et al. 2005; Peralta et al. 2005). Most of these species are limited in distribution to a small area in western Peru, Chile, and Ecuador (Rick 1976). Only *Solanum lycopersicum var. esculentum*, the domesticated tomato, and *S. lycopersicum var. cerasiforme*, its small-fruited feral putative congener, are found outside this narrow range, being common throughout many parts of the world, especially in Mesoamerica and the Caribbean (Rick 1976). Historical and linguistic studies suggest that the cultivated tomato was most likely selected from wild forms of cerasiforme (Jenkins 1948; Rick 1976); however, phylogenetic/diversity studies based on isozymes and DNA polymorphism have not clarified this issue (Rick & Fobes 1975; Rick et al. 1974; Williams & St. Clair 1993; Miller & Tanksley 1990).

All members of the clade are closely related diploids (2n =24) (Nesbitt & Tanksley 2002; Peralta & Spooner 2001; Rick et al. 1979) that share a high degree of genomic synteny (Chetelat & Ji 2007). Rick (1963) tested through cross experiments whether there were any interbreeding barriers within the *S. peruvianum* complex. His results showed that they are to some degree intercrossable (Rick 1979). Breeding systems vary from allogamous self-incompatible to facultative allogamous and self-compatible to autogamous and self-compatible (Rick 1963; Rick 1979; Rick 1986). The self-incompatibility system in tomatoes is gametophytic and controlled by a single, multiallelic S locus (Tanksley & Loaiza-Figueroa 1985).

4.1. Isolation and Incompatibility in Tomatoes

The *Solanum sect. Lycopersicon* group contains species with both allopatric and sympatric components to their distribution ranges (T. Nakazato, D. Warren, and L. C. Moyle, unpublished data). Nonetheless, there are very few reports of natural hybridization among wild tomatoes (Taylor 1986), despite decades of field collections and observations. This observation suggests that non-ecological barriers are also potentially important in isolating wild tomato species. Within the *Solanaceae* in general, some of the strongest post mating reproductive isolating barriers appear to occur at the stage of pollen–pistil incompatibility (De Nettancourt 2001; Hancock et al. 2003; McCormick 1998), and interspecific crosses suggest pollen–pistil interactions could act to isolate wild *Solanum sect. Lycopersicon* species (*e.g.*, Hogenboom 1972). Several studies have examined the genetic basis of pollen and seed sterility between the cultivated tomato and each of several wild congeners (*e.g.*, *S. habrochaites*: Moyle & Graham 2005; *S. pennellii*: Moyle & Nakazato 2008). Overall, these studies indicate that individual hybrid incompatibility QTL appear to be recessive (or at most additive), that a relatively modest number of QTL underlie hybrid incompatibility, and that there are roughly comparable numbers of pollen and seed sterility QTL (*i.e.*, within the same order of magnitude) (Moyle & Nakazato 2008).

BDM factors also can also play a role in causing hybrid weakness or inviability in this clade. Hybrid weakness is often manifested as necrosis in developing seedlings or adult plant tissue, similar to the phenotype of pathogen attacks (Bomblies & Weigel 2007). Tomato lines with resistance gene (Cf-2) from *S. pimpinellifolium* exhibit autonecrosis of mature leaves, but no autonecrosis was observed when complementary gene (RC3) from *S. pimpinellifolium* was also introduced (Krüger et al. 2002). These observations imply that hybrid weakness may result from changes in pathogen resistance genes, which diverge in response to selection pressure exerted by pathogens.

4.2. Taxonomical Treatment

There has been a lot of debate about the taxonomy of tomatoes since Linnaeus (1753) treated them as part of the genus *Solanum*, alongside potatoes. A year later Miller (1754) recognized tomatoes as a separate genus (*Lycopersicon*). The latter has been the predominantly accepted view up to the present. Recently, molecular data have been used to analyze phylogenetic relationships between the former genus *Lycopersicon* and genus *Solanum* (Olmstead et al. 1999; Bohs & Olmstead 1999). Based on this evidence and their own data, Spooner et al. (2005; 1993) have proposed that species formerly belonging to *Lycopersicon* be included in genus *Solanum sect. Lycopersicon*. This new taxonomical treatment coincided with the breakup of *L. peruvianum* into four *Solanum* species (Peralta et al. 2006). The description of a new wild tomato species on the Galapagos Islands (Darwin et al. 2003) led to the breakup of *L. cheesmanii* into *S. cheesmaniae* and *S. galapagense*. The following table summarizes the changes in taxonomic treatment:

Lycopersicon Name	*Solanum* Classification
L. cheesmanii	*S. cheesmaniae* *S. galapagense*
L. chilense	*S. chilense*
L. chmielewskii	*S. chmielewskii*
L. esculentum	*S. lycopersicum*
L. hirsutum	*S. habrochaites*
L. parviflorum	*S. neorickii*
L. pennellii	*S. pennellii*
L. peruvianum	*S. arcanum* *S. corneliomuelleri* *S. huaylasense* *S. peruvianum*
L. pimpinellifolium	*S. pimpinellifolium*

Though much evidence supports the inclusion of former genus *Lycopersicon* into *Solanum*, the division of *S. peruvianum* into four species is arguable. So far, a new formal description has only been published for *S. arcanum* and *S. huaylasense* (Peralta et al. 2005). Additionally, in a classic study, Rick (1963) concluded that it is most reasonable to classify accessions putatively belonging to *S. corneliomuelleri* as part of *S. peruvianum*.

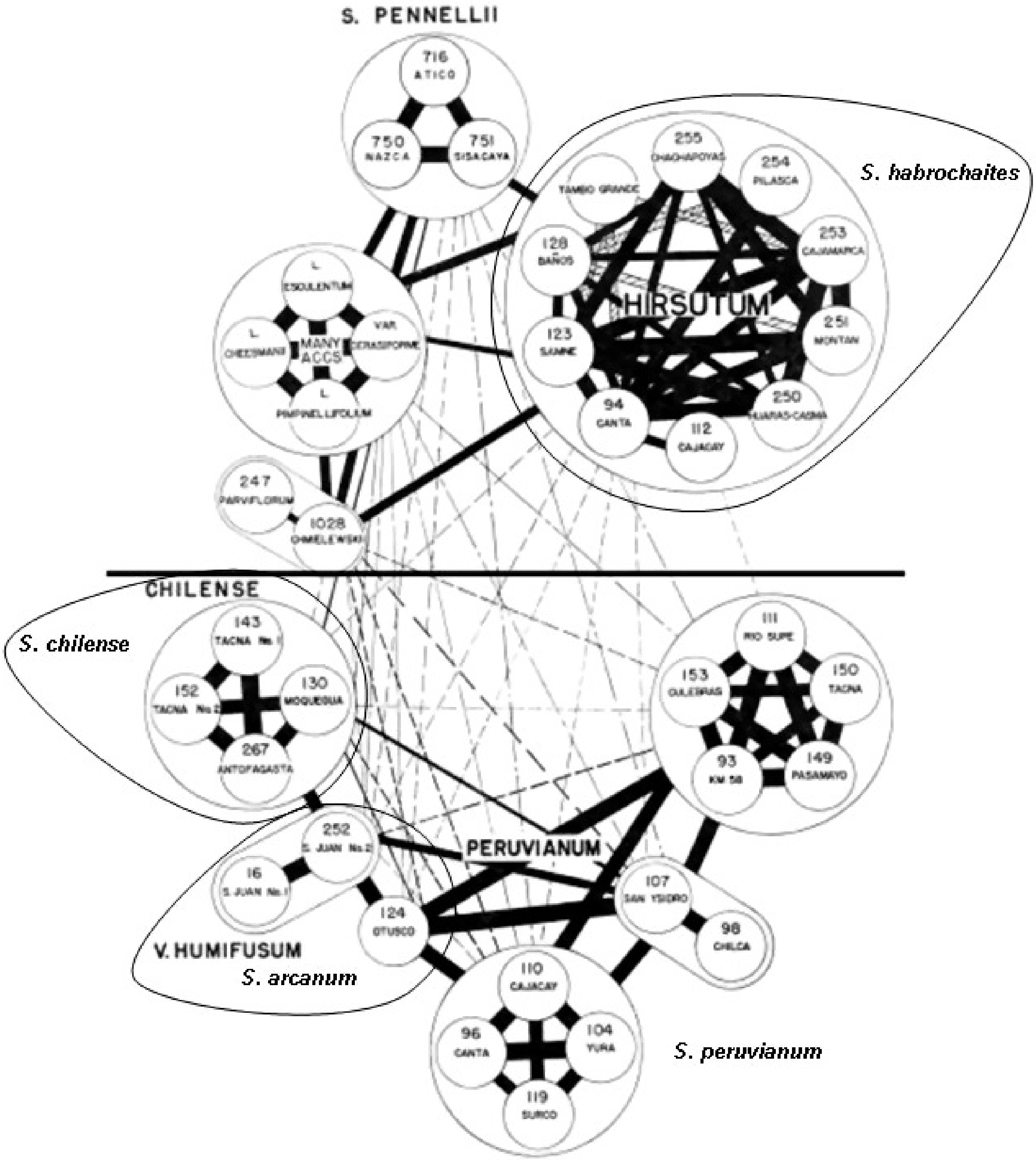


Figure 6 Rick's (1979) polygon scheme showing crossability results; the width of connecting bands indicates the amount of seed produced by crosses, and dashed lines indicate crosses that failed to produce hybrids. Adapted from Rick (1979). Names in italics correspond to the taxonomical treatment used in our study and are only shown for species used in our study. In the case of *S. peruvianum*, the graph shows populations presently assigned to *S. peruvianum* sensu stricto as well as *S. peruvianum sensu lato*. Therefore, we don't use an oval to indicate this species (as we do with *S. arcanum*, *S. chilense* and *S. habrochaites*) to avoid misunderstandings.

5. Scope of this thesis

5.1. Previous Research

Past studies from our group focused on the effect of the mating system and recombination on single nucleotide polymorphism (Roselius et al. 2005) as well as testing the isolation model of speciation, particularly the assumption of divergence without gene flow (Stadler et al. 2005). Thus, it was concluded that the evolution of *Solanum sect. Lycopersicon* has been dominated by demographic processes. A high effective population size, despite small census size was also noticed, suggesting the presence of soil seed banks and extensive population structure (Roselius et al. 2005). Weak evidence of post-divergence gene flow from *S. chilense* to *S. peruvianum* was also found (Stadler et al. 2005), leading to more detailed studies of these two species. Arunyawat *et al.* (2007) discovered that population structure was an important demographic factor shaping the patterns of nucleotide diversity within and among populations in wild tomatoes. In addition, results from Arunyawat *et al.* (2007) contradict the breakup of *Solanum peruvianum* in the latest taxonomical treatment. According to their analyses, populations from Canta (Peru) putatively belonging to *S. corneliomuelleri* were the least differentiated when compared to *S. peruvianum sensu stricto*. Finally, Stadler et al. (2008) confirmed that divergence between *S. peruvianum* and *S. chilense* took place under residual gene flow in both directions either under a parapatric mode of speciation or through a period of secondary contact. This study also detected a population (or range) expansion for *S. peruvianum* and estimated the effective population size for *S. chilense* as similar to that of the ancestral species from which both diverged. For the present study, we analyze the sequence of the same loci as Arunyawat *et al.* (2007) in two new species: *S. arcanum* and *S. habrochaites* (see below for details).

Solanum sect. Lycopersicon is ideal for integrating genomic tools and approaches into ecological and evolutionary research. Wild species within *Lycopersicon* span broad morphological, physiological, life history, mating system, and biochemical variation, and are separated by substantial, but incomplete post-mating reproductive barriers, making this an ideal system for genetic analyses. This is matched by many logistical advantages, including extensive historical occurrence records for all species in the group and publicly available germplasm for hundreds of known wild accessions (Moyle 2008). Due to their recent origin, the clear phenotypic distinction between species, their diversity of mating systems and the well-known genetics of the cultivated tomato (Stadler et al. 2005), wild tomato species are a good speciation model.

5.2. Objectives

The aim of this thesis is to study the population history and speciation process in closely related plant species, using wild tomato as a model system. We are interested in studying population structure and range expansion in the four wild tomato species (*S. peruvianum*, *S. chilense*, *S. habrochaites* and *S. arcanum*). The study uses a set of eight DNA loci obtained in the four species and various coalescent tools. The study is composed of three projects, each addressing specific questions:

As explained in this introduction, the sampling strategy is of prime importance when doing statistical inference of evolutionary processes. Our work is concerned with estimates of species-wide and population specific demographic history based on summary statistics. Particularly, it becomes important to know the joint impact of population subdivision and the sampling scheme on the site frequency spectrum in populations which are not at demographic equilibrium. The first project deals with a theoretical work assessing the effect of the sampling strategy on summary statistics and parameter estimation of species wide expansion.

The second project looks for signatures of selection in eight reference loci. These eight loci have been used in previous studies from our group. Results so far, have not shown any significant evidence of strong selection, and thus these loci have been considered as a good reference for studying neutral evolutionary processes. The aim of the second chapter is to confirm that in the four species studied by our group so far, these loci can still be regarded as reference loci, *i.e.* to measure to which degree their evolution is neutral.

As mentioned in this introduction, previous studies have demonstrated the existence of moderate population structure in wild tomatoes, and possible existence of seed banks and range expansion. In the third project, we reveal if such pattern also exists in other species of the tomato complex, *i.e. S. habrochaites* and *S. arcanum*. This involves first, to define precisely what a population is. Thus, we investigate if each sampled populations is an interbreeding community, and its degree of isolation from other populations. Then, we study the population and species history for each species.

As highlighted in this introduction, wild tomato species are a complex of closely related species in which regular taxonomical changes occur. This is due to insufficient definition of species' ecological habitats and a lack of species definition from a population genetics point of view. Previous results from our group seem indeed hard to reconcile with the new suggested taxonomical treatment. A first specific question also addressed in the third project is thus whether population genetics data support the subdivision of former *S. peruvianum sensu lato* into *S. arcanum* and *S. peruvianum sensu stricto*.

Materials and Methods

This thesis covers three research projects. Therefore, this section and that describing the results are divided in three sub-sections. The first section of this chapter describes the methods used for the first research project of this thesis. These included modeling and simulating data to test our hypothesis. The second and third projects analyze molecular data. Therefore, they share the same plant material and sequencing methodologies which will only be described for the second project.

First Project

1. Coalescent simulations under two models of population structure

All patterns of sequence diversity were generated using the coalescent simulation software ms (Hudson 2002) to model the following evolutionary scenario. At the time of sampling, the population consists of I demes, each containing N_0 diploid individuals. For the first set of simulations, the subdivided population is at equilibrium with constant population size. Along every line mutations accumulate at rate μ, and $\theta = 4N_0 \mu$. We chose to simulate mainly with a fixed value of $\theta = 1$ (for simulations implementing 100 demes) rather than with a fixed number of segregating sites S or choosing different θ's for different simulation scenarios to obtain realistic values of S and/or π. Simulating with a fixed S has been shown to yield inaccurate results under non-equilibrium demography (Ramos-Onsins & Rozas 2002), and the critical values for Tajima's D and Fu and Li's D depend not too strongly on θ. Moreover, the effects we describe in our results are orders of magnitude larger than what could be generated by choosing a different θ or fixing S.

The demes exchange (haploid) migrants under either an island model or a two-dimensional stepping-stone model at rate m, and we consider a broad range of gene flow: $0.1 \leq 4N_0m \leq 100$. Under the island model an ancestral line in deme j switches its location to deme i at rate $m/(I - 1)$. Under the stepping-stone model, we assume that $l = a^2$; *i.e.*, the population is arranged in a square lattice of a x a demes and we assume periodic boundary conditions. This means that the migration rate is m/4 if $i = (i_1, i_2)$ and $j = (j_1, j_2)$ are neighboring demes and 0 otherwise. Here, i and j are neighbors if $|(i_1 - j_1)\text{mod } a| + |(i_2 - j_2)\text{mod } a| = 1$. In other words, an individual at location (a, i_2) can migrate to $(1, i_2)$ and one at (i_1, a) can migrate to $(i_1, 1)$ and vice versa. We modified ms to be able to efficiently sample sequences from randomly chosen demes rather than fixed demes. Specifically, in each iteration the modified version of ms shuffles the entries in the sample configuration array inconfig at the beginning of the function segtre_mig (Hudson 2002). The C code of this program is available from http://guanine.evolbio.mpg.de/sampling.

2. Implementing range expansions under population subdivision

For an additional set of coalescent simulations, we assume that the structure in the population was created some time τ in the past (τ is measured in units of $4N_0$ generations), *i.e.*, before time τ the population was panmictic and of size N_A. This scheme ought to be plausible under range expansions, *e.g.*, as exemplified by migration out of Africa by both humans and Drosophila melanogaster and subsequent colonization of expansive areas, or temperate-zone populations expanding from glacial refugia, or following a speciation event such as that inferred for the two wild tomato species *S. peruvianum* and *S. chilense* (Stadler et al. 2005; Stadler et al. 2008). Moreover, this is essentially a generalized “isolation with migration” (IM) model of divergence with a large number of extant demes (Hey & Nielsen 2004; Nielsen & Wakeley 2001; Wilkinson-Herbots 1998). Looking forward in time, at time τ the ancestral population splits into l demes of equal size and in equal proportions. The “expansion factor” for the total population at time τ is thus given by

$$\beta = l \times N_0/N_A$$

Note that a value of $\beta = 1$ implies constant population size in the sense that a panmictic population at time τ in the past split into l demes each of size N_0 (= N_A/l), without changing the total census size of the entire, now subdivided population. This particular scenario can be seen as a form of ‘‘range fragmentation,’’ albeit without decline in total population size.

3. Sampling schemes and descriptors of diversity and differentiation

Simulated samples of total size n (20 in our numerical examples) from the structured population were implemented as local, pooled, or scattered. Local samples contain n sequences from a single island; *i.e.*, only one arbitrarily chosen deme is sampled. Pooled samples contain several lines each from several demes (we take five lines from each of four demes in our simulations). Scattered samples encompass single sequences from each of n different demes, *i.e.*, only one sequence per sampled deme.

A commonly used statistic to quantify population structure from patterns of diversity within and among local populations is F_{ST} (*e.g.*, Hudson et al. 1992). In particular, if the number of demes is large and the population is in equilibrium,

$$E[F_{ST}] = 1/(1 + 4N_0m)$$

(*e.g.*, Wright 1951), and thus migration rates can, in principle, be estimated from observed values of F_{ST} (but see Whitlock & McCauley 1999 for numerous caveats and Jost 2008 for a more fundamental critique of F_{ST}-based estimates of differentiation and gene flow). In our simulations, F_{ST} can be computed only under the ‘‘pooled’’ sampling scheme. We used the formula $F_{ST} = 1 - \pi_w/\pi_b$, where π_b is the ‘‘average’’ number of differences for pairs of sequences taken from different demes, and π_w is the average number of differences for pairs sampled within demes. The average here means only the average over all pairs of sequences, and not over all simulation runs. Thus, for every simulation run, we recorded exactly one F_{ST} value. For the stepping-stone model, we used the same computations; *i.e.*, F_{ST} does not take isolation by distance into account.

To describe sequence diversity patterns, we focus on the site frequency spectrum, as summarized by statistics such as the widely used Tajima’s D

(Tajima 1989) and Fu and Li's D (Fu and Li 1993). For clarity, we denote these distinct D statistics as D_T and D_{FL}, respectively. Using these particular summary statistics enables us to perform power analyses to reject the standard neutral model. Moreover, we chose to include D_{FL} because the singleton class appeared to be the major reason for the lower/ more negative D_T values in pooled vs. local samples of wild tomatoes (Arunyawat et al. 2007). The statistical package R was used to drive our modified version of ms and to compute these statistics from its output; the corresponding R scripts are available at http://guanine. evolbio.mpg.de/sampling.

Second Project

1. Plant Samples and Sequencing

Samples were collected in Peru by T. Staedler, T. Marczewski and C. Merino in two separate collection trips (2004 and 2006). Four to five populations per species were used: Ancash, Canta, Otuzco, Contumaza, and Lajas for *S. habrochaites*; and Otuzco, Rupe, San Juan, and Cochabamba for *S. arcanum*. From each population six individuals were collected except for Otuzco (seven individuals for *S. habrochaites* and 8 for *S. arcanum*), Ancash (four individuals) and San Juan (five individuals). The population samples and geographic locations are summarized in Table 1. Voucher specimens have been deposited at the herbarium of the Universidad Nacional Mayor de San Marcos (USM, Lima, Peru).

Population	Coordinates	Climate	Census Population Size
Ancash	09°31'S, 7°53'W	dry-mesic	n.a.
Canta	11°31'S, 76°41'W	mesic	++
Cochabamba	06°29'S, 78°54'W	dry-mesic	++
Contumaza	07°22'S, 78°48'W	dry-mesic	++++
Lajas	06°33'S, 78°46'W	dry-mesic	++
Otuzco	07°56'S, 78°36'W	mesic	+++
Rupe	07°17'S, 78°49'W	dry	+
San Juan	07°17'S, 78°33'W	dry-mesic	++

Table 1 Collection sites and description. Plus symbols (+) are used to express approximate relative population sizes.

Collected leaves were dehydrated in sealable plastic bags using silica gel. When the collection trip ended, they were transported to the lab where genomic DNA was isolated using the DNeasy Plant Mini Kit (Qiagen GmbH, Hilden, Germany). At least five individuals (10 alleles) per population were sequenced for each of eight unlinked reference loci used in previous studies (CT093, CT208, CT251, CT066, CT166, CT179, CT198, and CT268; Arunyawat et al. 2007; Roselius et al. 2005; Stadler et al. 2008; Stadler et al. 2005). These loci correspond to anonymous, single-copy cDNA markers originally mapped by (Tanksley et al. 1992). Putative functions are proposed for all these loci (see Table 2; modified from Roselius et al. 2005). The loci were not chosen, however, on account of their function but based on the fact that they are located in regions of certain recombination rates as estimated by Stephan & Langley (1998), based on recombination nodes (RN), following the work of Sherman & Stack (1995).

Locus	Chromosome	Length (bp)	Putative encoded protein	RN
CT066	10	1346	Arginine decarboxylase	0.93
CT093	5	1415	S-adenosylmethionine decarboxylase proenzyme	0
CT166	2	2673	Ferredoxin-NADP reductase	1.61
CT179	3	995	Tonoplast intrinsic protein D-type	1.97
CT198	9	779	Submergence induced protein 2-like	2.1
CT208[a]	9	1767	Alcohol dehydrogenase, class III	0
CT251[a]	2	1779	At5g37260 gene	0.46
CT268[a]	1	1887	Receptor-like protein kinase	2.33

Table 2 Chromosome location, putative function, and recombination rate (RN) of sequenced loci. Locus designations refer to particular EST sequences that have been integrated into longer "tentative contigs" in the TIGR Plant Transcript Assemblies (http://plantta.jcvi.org/cgi-bin/plantta_release.pl). The length per locus is given across the total alignment of all five tomato species (without outgroup), including indels.

[a] From Baudry et al. (2001).

Each locus was initially sequenced with the same PCR primers that were used for amplification of each locus. PCR primers and conditions are deposited at http://www.zi.biologie.unimuenchen.de/evol/Downloads.html. To resolve haplotype phase, *i.e.* to confirm linkage between each pair of SNP alleles in each individual, two independent strategies were used. On the one hand, allele-specific primers anchoring in polymorphic sites were used as previously described (Stadler et al. 2005). To this end, primers were either designed based on direct sequencing of PCR product, or taken from previous studies

(Arunyawat et al. 2007). In this way, we have used overlapping sequences from different primers to confirm the presence of SNPs as well as establish the phase between each of them. On the other hand, due to technical difficulty (*e.g.* SNPs being located in a region unsuitable for primer design) we have also used cloning. In this approach, PCR products were inserted into a plasmid and cloned. After screening colonies by PCR with the original PCR primers to confirm presence of the target sequence, the plasmid was extracted and sequenced using plasmid primers.

2. Statistical tests of neutrality

Two types of sampling schemes are used. The pooled sample refers as the combined sequences from all populations for a given species, and the population sample refers to a separate analysis of the four to five populations per species (Stadler et al. 2009). We calculate K_a/K_s and π_a/π_s ratios for synonymous and non-synonymous sites. The McDonald-Kreitman test (McDonald & Kreitman 1991), as well as the calculation of the proportion of adaptive substitutions α (Smith & Eyre-Walker 2002; Bierne & Eyre-Walker 2004) are applied to pooled samples. Both tests are based on a comparison of the divergence between two species, taking into account the rate of synonymous and non-synonymous substitutions. *S. lycopersicoides* was used as an outgroup for CT093 and CT268, *S. ochranthum* was used for all other loci. All statistical analyses are applied using DnaSP v. 5.0 (Librado & Rozas 2009) and SITES (Hey Lab, Department of Genetics, Rutgers University). α is computed using the DoFE software (Smith & Eyre-Walker 2002; Bierne & Eyre-Walker 2004).

3. Site frequency spectrum and purifying selection

We calculate a simplified version of the SFS comprising three categories (Fay et al. 2001). The minor allele at each polymorphic SNP is called rare if its frequency is below 5%, intermediate if the frequency is higher than 5% and lower than 20%, and common if its frequency is higher than 20%. These categories of SNP frequency are calculated for the pooled sample (40 to 62 sequences per species). For population samples, however, as we have only 12 sequences per population, two classes of polymorphic sites are used: low frequency (f<20%) grouping singletons and doubletons, and common frequency (f>20%).

When classifying the SNPs we used the outgroup specified above to determine the frequency of the derived state. We allowed for multiple hits and calculated the frequency of all derived states. If one site is polymorphic within the species and different from the outgroup, it is also considered as a multiple hit.

The dataset is partitioned into three categories of sites: synonymous (S), non-synonymous (NS) and non-coding (NC) polymorphic sites. For each category, the two or three classes of the simplified SFS are computed. Under simplified assumptions, NS and NC sites fall into three classes: neutral, slightly deleterious and strongly deleterious (Fay et al., 2001). Neutral NS or NC sites are responsible for all common SNPs in the SFS and a proportion of rare and intermediate SNP classes. Slightly deleterious mutations account for the excess of rare frequency polymorphism, as well as a small fraction of the intermediate frequency SNPs. Strongly deleterious mutations are assumed here to rarely rise to detectable frequency (Fay et al. 2001). The amount of non-synonymous and non-coding SNPs in each class is compared to that observed for the synonymous ones. The synonymous sites are assumed to be neutral and their SFS is thus only determined by past demographic events and metapopulation structure. Under purifying selection, an excess of rare-frequency polymorphisms in NS or NC sites in comparison to the amount of rare-frequency polymorphisms in synonymous sites is thus expected. On the other hand, similar frequencies of common SNPs are expected for the various classes reflecting the neutral evolution of common SNPs (Fay et al. 2001).

The amount of polymorphism S*, NS* and NC* are computed for each frequency class, rare, intermediate, and common, where * denotes the ratio of the number of SNPs per total number of sites (Fay et al. 2001). The proportion of non-synonymous sites NS* is calculated as 1 minus the number of fourfold degenerate sites divided by the total number of coding sites using DnaSP v 5.0

(Librado & Rozas 2009). We calculated NS*, S*, and NC* for all polymorphic sites and for only private polymorphisms for each species. Calculations were made using either Microsoft Office Excel 2003 or R scripts (R Development Core Team 2005).

4. Purifying selection on shared polymorphisms among species

The distribution of ancestral deleterious mutations shared among species is computed using the pooled sample. The aim is to determine if metapopulation structure in the ancestral and incipient species favors the maintenance of deleterious mutations. The SFS is computed for the shared polymorphic S*, NS*, and NC* sites for each of the six possible pairwise comparisons of species. Correlations are analyzed for the amount of shared polymorphism in S* sites with the amount of NS* and NC* sites for rare, intermediate, and common alleles. The amount of shared polymorphic sites S* between species is supposed to be inversely proportional to the divergence time between species (following results in Hey & Wakeley 1997), and depends also on the population size of the two species, levels of introgression between them (Hey & Wakeley 1997), and the metapopulation structure of each species. Positive correlations between S* and NS* (and NC*) are expected as pairs of species with short divergence (high S*) would also exhibit higher rates of shared NS* or NC*. Correlating the shared S* and shared NS* (NC*) allows us to control for different mutation rate and divergence time between pairs of species.

Linear regressions are calculated between S* and NC* and NS* found for all pairwise species comparisons. If NC and NS sites evolve neutrally, one expects a regression line with equation S*=NC* (or S*=NS*). The linear regression analysis is performed using the lm command of the statistics software R (R Development Core Team 2005).

5. Quantifying purifying selection and demography for each species

For each species we concatenated all loci using all polymorphism data (shared and private among species) and calculated the distribution of fitness effects

(DFE) for the NS and NC sites by comparing with synonymous sites (data not shown). This is realized using the maximum-likelihood method by (Eyre-Walker & Keightley 2009), which infers demographic, DFE parameters and α simultaneously using all information in the SFS. As previous work reveals evidence for species expansion in *S. peruvianum* and *S. chilense* (Arunyawat et al. 2007), we need to take into account demographic expansion to estimate the DFE and α. This method is available on Keightley's web-server (http://homepages.ed.ac.uk/eang33/).

The demographic model is a simple one-step population size change from N_1 ancestral population size to N_2, the present effective population size, assumed to be at equilibrium between mutation, selection and drift. The population expansion ($N_1 < N_2$) or contraction ($N_1 > N_2$) occurs t generations ago. Each deleterious mutation has a different fitness coefficient s, which is assumed to be drawn from a gamma distribution with shape parameter b and mean parameter $N_2E(s)$ (Keightley & Eyre-Walker 2007). It is also assumed that there is a class of neutral sites at which mutant alleles have no effect on fitness. For diploid organisms, the fitness of the wild-type, heterozygote mutant, and homozygote mutant genotypes are 1, 1-s/2, and 1-s respectively (Keightley & Eyre-Walker 2007). In a second step, α, the rate of positively selected mutations, is estimated for coding regions (Eyre-Walker & Keightley 2009). The demography and the DFE parameters are thus used to predict the expected number of substitutions due to deleterious mutations. The difference between this expected number and the observed number of substitutions gives an estimate of α (Eyre-Walker & Keightley 2009). The number of polymorphisms and substitutions for NS, S and NC sites are calculated using *S. lycopersicoides* as an outgroup for CT093 and CT268, and *S. ochranthum* for all other loci. The parameters estimated are thus N_2, t, E(s), b, and f_0 which is the proportion of sites that have never experienced a mutation (invariant sites) assuming the ancestral population size N_1=100. α is estimated for coding regions and is compared to the results obtained with the method of Bierne & Eyre-Walker (2004; data not shown).

The model assumes that all sites are unlinked, which is in agreement with previous studies revealing that *S. peruvianum* and *S. chilense* show high recombination rates (Stephan & Langley 1998; Arunyawat et al. 2007). The model further assumes that all sites have the same mutation rate and no multiple hits occur. We correct for multiple hits by calculating the number of substitutions and polymorphisms using the DnaSP conservative criteria (Nei & Gojobori 1986). The demographic parameters (N_2, t) are estimated using the site frequency spectrum for synonymous sites and non-coding sites. This dataset has the maximum number of polymorphic sites, thus the highest statistical power to infer parameters. Note that demographic and DFE parameters are calculated for the pooled sample of each species, taken here as representative of the whole species.

6. Purifying selection and metapopulation structure

Our objective here is to investigate for each species the effect of metapopulation structure on the strength of selection against the deleterious mutations. We test if the strength of selection at the population level is identical to the strength of selection estimated for the whole species using the DFE (see above).

In a metapopulation with low migration and weak to strong purifying selection, an excess of low-frequency alleles (f<20%) private to each population is expected when comparing sites under selection and synonymous sites. This results in higher population differentiation (F_{ST}) for NS (NC) sites compared to S sites between populations. This occurs because purifying selection prevents deleterious alleles to rise to high frequency at which they are likely to migrate among demes (Fay et al., 2001; Whitlock, 2003). Synonymous sites are used here to reflect the demography and the metapopulation structure of each species.

Genetic differentiation between populations of a given species is estimated using F_{ST} per polymorphic SNP given as output from the BayeScan program by

(Foll & Gaggiotti 2008). We compare the distribution of F_{ST} values for all polymorphic SNPs for each species for each type of site (S, NS, NC). Due to the non-normal distribution of F_{ST} values, non-parametric statistical tests are used to compare the F_{ST} distribution for the different types of sites. The effect of the type of site (NC, S or NS) on distribution of F_{ST} is evaluated using a one-way Kruskal-Wallis test. If the effect of the type of site is significant (at 5%), pairwise Wilcoxon tests determine which type of site has higher F_{ST} values (R Development Core Team 2005).

We use then the program BayeScan by Foll & Gaggiotti (2008) to detect outlier SNPs that deviate from the expected distribution of F_{ST} values. In the approach developed by (Foll & Gaggiotti 2008) the posterior probability that a locus is under selection is estimated assuming an island model of metapopulation. A locus may or may not be under selection. For both models, a Bayes Factor (BF) is calculated, which indicates the model that fits the data best. The program BayeScan calculates the F_{ST} for every SNP, and estimates the posterior probability of a SNP to be under the effect of selection.

Finally, to test the possibility of heterogeneous levels of purifying selection in space, we calculate NS*/S* and NC*/S* ratios for each population of each species, using only the private polymorphic sites for that population. This allows us to study how selection acts within a population, taking into account only the scattering phase of the metapopulation coalescent (Wakeley & Aliacar 2001). Note that these ratios are very weakly biased by demography as single populations show only a weak signature of the species-wide demography (Stadler et al. 2009).

Third Project

1. Estimation of Nucleotide Diversity and Neutrality Tests

Levels of nucleotide diversity are estimated using θ (Watterson 1975). The average number of pairwise differences, π (Nei 1987), is also estimated. We also calculate π_{between} by performing pairwise comparisons between sequences coming from different collection sites (among populations) and different species (among species). Thus, we obtain a better assessment of the species-wide nucleotide diversity within each species and an unbiased estimation of nucleotide divergence between species, respectively. We then test for deviations from neutrality using Tajima's D statistic (Tajima 1989). Negative values of Tajima's D indicate an excess of low-frequency polymorphisms, while positive values indicate an excess of intermediate-frequency polymorphisms, both of which could be explained by either selection or demography acting on the tested locus. In order to further test if deviations from neutrality are due to selection or demography, we applied the R_2 test (Ramos-Onsins & Rozas 2002) using multilocus concatenated data (of silent polymorphism). This test detects population growth in the recent past, which Tajima's D is unable to differentiate from selection. Significance of this test is assessed by coalescent simulations using population parameters from the data ($\alpha \leq 0.05$). Values under the lower critical value are considered significant and indicative of a population expansion in the recent past. We performed the test on *S. arcanum* and *S. habrochaites* as well as previously published data (Arunyawat et al. 2007) of *S. peruvianum* and *S. chilense* to offer a reference framework. Since locus CT198 presents a mutant allele with a premature stop codon, to avoid bias in the summary statistics, we remove alleles carrying the stop codon mutation in the calculation of all summary statistics for this locus. All estimations are performed as implemented in the DnaSP software (Librado & Rozas 2009).

2. Population Structure

In order to assess whether the distribution of haplotypes might follow a geographic pattern, haplotype networks are built. To this end, we use Sneato software (The McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center). This software generates a minimum spanning tree depicting relations among biological sequences, such as DNA, using Prim’s algorithm. This allows us to visualize the distance between haplotypes (as mutational steps) and the number of copies of each haplotype in a minimum spanning network.

We also estimated the extent of population differentiation calculating F_{ST} (Hudson et al. 1992) performing pairwise comparisons between populations of the same species as well as between populations from two different species. F_{ST} estimation is performed using DnaSP software, which estimates the number of effective migrants among populations (Hudson et al. 1992). The objective is to investigate whether divergence between populations is greater than diversity within populations. In addition, we also quantify levels of differentiation within and between species in a hierarchical analysis of molecular variance (AMOVA, Excoffier et al. 1992) as implemented in Arlequin (Excoffier et al. 2005). This analysis allows us to test the hierarchical structure we have defined in our dataset, *i.e.* that several populations belong to a single species.

Another way of analyzing the difference between the geographic and genetic definitions of populations is to use Principal Coordinates Analysis (PCoA). This method is suitable to analyze high-dimensional data, *i.e.* data with a high number of variables. To do this, a dissimilarity matrix (*i.e.* a Euclidian distance matrix) is calculated from the data. Then, eigenvectors and Eigen values are calculated. The eigenvectors are used as axes on which to project the data points. Since Eigen values measure the amount of data variance explained by the eigenvectors, the Eigen values are used to rank the eigenvectors and select which ones to use for the graphic representation of the dataset, *i.e.* which ones explain the biggest part of the variance in the data. The analysis is applied to multilocus genotypic unphased data as implemented in DARwin (Perrier & Jacquemoud-Collet 2006).

Additionally, we used the software STRUCTURE (Pritchard et al. 2000) to infer the presence of distinct populations, *i.e.* single panmictic units, in our data and compare these to the geographic definition of populations from which individuals were collected. For such simulations we used a running set-up similar to that used by (Evanno et al. 2005): an admixture model with K putative populations from were the admixed populations draw their alleles, allele frequencies independent among populations and 10^6 iterations for the burn-in period as well as for the parameter estimation. Thus, our null hypothesis is that our data points aggregate into groups that resemble the populations from where samples were taken. Otherwise, finding more population units than sampled would indicate that a cryptic reproductive structure occurs within each population. On the contrary, finding less panmictic units than populations sampled would indicate high rates of gene flow or too recent divergence between the sampled populations.

In order to assess which value of K better fits our data, we follow the guidelines by (Evanno et al. 2005). Since STRUCTURE estimates the likelihood of each K value used for the simulations, we select K generating the maximum likelihood.

3. Divergence between Species

The latest revised taxonomical treatment of wild tomatoes led to the breakup of *L. peruvianum* into four *Solanum* species (Peralta et al. 2006), two of which are *S. peruvianum* and *S. arcanum* (the others being *S. huaylasense* and *S. corneliomuelleri*). In order to evaluate evidence supporting the split of *S. arcanum* from *S. peruvianum*, two methods already described (see above) are used specifically on data for these two species.

Although the STRUCTURE model assumes that loci are independent within populations (which is not the case for sequence data given that STRUCTURE considers each nucleotide as a locus), STRUCTURE will perform well as long as recombination is significantly high so that LD does not dominate the regions under analysis. LD can also affect the analysis when data are not phased, *i.e.* one does not know the full haplotype of individuals. This is the case in our dataset, as for the various loci we cannot assign sequences from different loci to a given chromosome and sequences are arbitrarily assigned as “a” or “b” at each locus. This leads STRUCTURE to interpret each “a” allele at a different locus as linked with each other. To investigate whether this might have a strong effect on the results of our simulations, we perform a second set of random simulations with the same parameters, to check for reproducibility of the results. The second set of estimations is performed on a “randomized” set, where alleles were randomly renamed as “a” or “b”, creating random phase in the data.

Additionally, we use π_{between} to estimate the divergence between species. For this, we follow the same procedure formerly used by (Arunyawat et al. 2007), *i.e.* we include all pairwise comparisons of sequences from different species and exclude pairwise comparisons within the same species. We include in our comparisons both *S. chilense* and *S. habrochaites* to offer a reference framework. Thus, we calculate π_{between} for silent, synonymous, and all sites for all 6 possible pairwise comparisons among our 4 species and use these data to calculate the net number of nucleotide differences (Wakeley 2000).

Results

Following the format of the Materials and Methods section, the Results section is also divided in three sub-sections, named First Project, Second Project, and Third Project.

First Project

Our coalescent simulations yield results for levels of nucleotide diversity; summary statistics based on the site frequency spectrum; and population differentiation under both equilibrium and non-equilibrium demographic history, all obtained for three different sampling schemes: local, pooled, and scattered. All our findings are consistent between the island model and the stepping-stone model, and in this section we focus on quantitative results obtained under stepping-stone spatial structure; results for the island model are not presented here.

1. The site frequency spectrum under population subdivision

The simplest demographic scenario we analyzed is an equilibrium population subdivided into 100 demes; *i.e.*, we first focus on the effects of population structure per se without any past changes of population size. As expected, characteristics of the site frequency spectra depend strongly on the sampling scheme. For the stepping-stone model of population structure, Figure 7 shows the simulation results under various levels of gene flow. For a migration rate of $4N_0m = 10$, local samples produce values of Tajima's D_T (Fu and Li's D_{FL}) that are significantly different from values expected under the standard neutral model (two-tailed test, $P < 0.05$) in 16% (39%) of all cases, while scattered samples give significant results in only 6% (7%) of all simulations. In particular, we see that local samples generate values for both statistics that are higher than expected for samples from panmictic populations, reflecting a site frequency distribution skewed toward intermediate-frequency mutations; this result mirrors the recent work of (De & Durrett 2007).

For migration rates (in units of $4N_0m$) between ~2 and ~50, pooled samples exhibit site frequency spectra that are broadly intermediate between those of local and scattered samples. The differences in sample genealogies (as reflected in estimates of D_T and D_{FL}) gradually diminish with higher levels of gene flow, but some differences among sampling schemes are still apparent at fairly high migration rates (*e.g.*, $4N_0m > 50$ for D_T and $4N_0m$ ~100 for D_{FL}; Figure 7). Importantly, pooling data from several subpopulations does not generate negative values of D_T or D_{FL} without an expansion of the total population. These observations also hold for the island model, albeit with smaller discrepancies between the summary statistics for scattered samples and those for the two other sampling schemes (*i.e.*, a less pronounced skew toward intermediate frequency mutations for both pooled and local samples; data not shown). For lower levels of gene flow ($\sim 4N_0m < 2$), the site frequency spectra of local samples gradually shift towards the standard neutral model's expectations, while those of pooled samples yield increasingly positive values of both D_T and D_{FL} under decreasing levels of migration. We checked our simulations of this equilibrium model against analytical results showing that local samples ought to be invariant for the level of nucleotide diversity, π, irrespective of the level of symmetrical interdeme migration in an island model (Strobeck 1987; Slatkin 1987). For both models of population structure, we found approximately invariant mean p-values for local samples over the entire range of simulated migration rates (data not shown).

2. The impact of population/range expansions on the site frequency spectrum

Next, we considered scenarios of (range) expansions under concomitant establishment of subdivision of the total species range, as described in the materials and methods section. In particular, we simulated a single ancestral population that at time τ before the present experienced a fragmentation into I demes, where the total number of individuals across the subdivided population could vary from N_A (the number of individuals in the single ancestral population, in which case the expansion factor $\beta = 1$) to 100 x N_A (equivalent to $\beta = 100$).

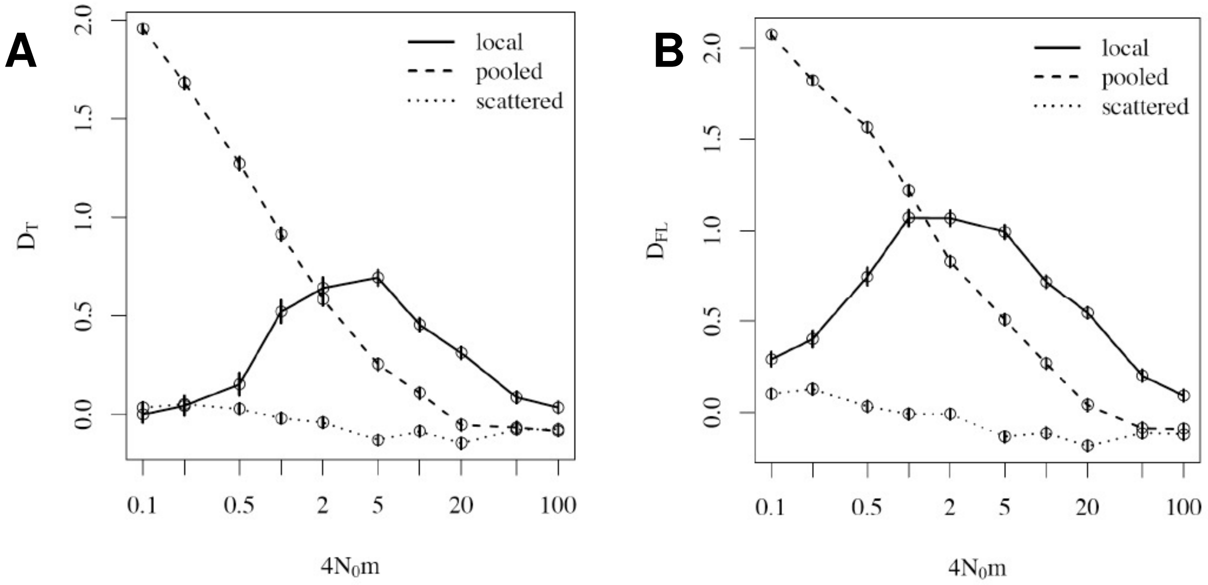


Figure 7 Averages of Tajima's D_T (A) and Fu and Li's D_{FL} (B) under three sampling schemes as a function of levels of gene flow. We simulated an equilibrium stepping-stone model with I = 100 islands; the simulations were carried out without recombination. Every plotted point is based on 1000 independently generated data sets. Standard errors of the means are indicated by vertical lines.

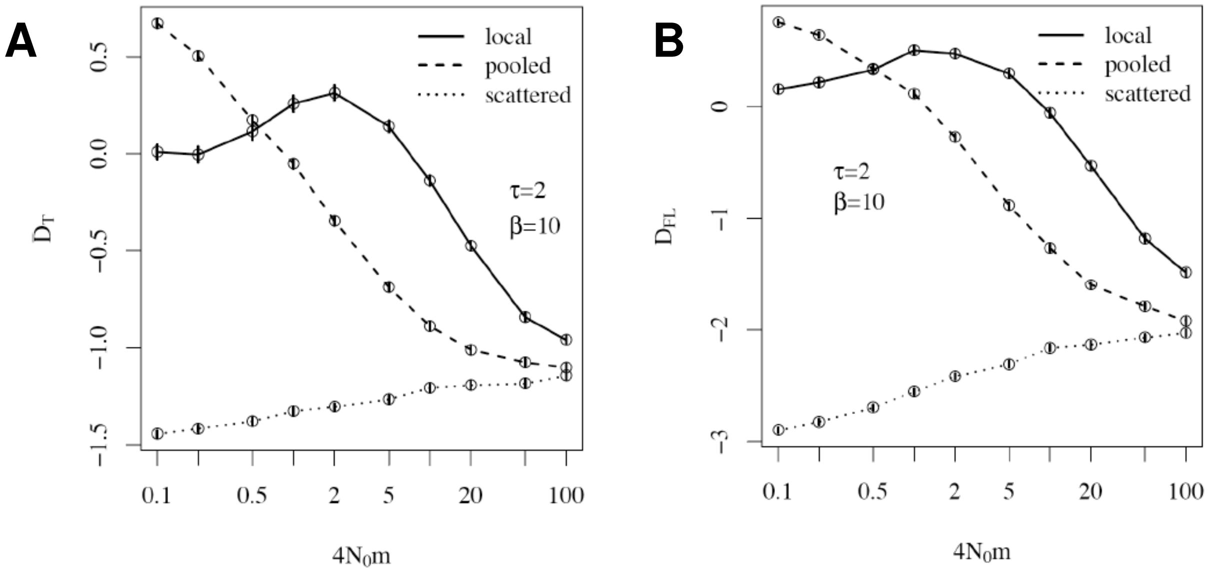


Figure 8 Averages of Tajima's D_T (A) and Fu and Li's D_{FL} (B) under species-wide expansion (β = 10, τ = 2) as a function of migration rates between demes. We simulated a stepping-stone model with 100 demes without recombination. Every plotted point is based on 1000 independently generated data sets; error bars represent standard errors.

Under a steppingstone model with a 10-fold population expansion $8N_0$ generations in the past, local samples still exhibit values of D_T and D_{FL} that would be expected under the neutral standard model conditions, as long as gene flow is fairly low (Figure 8); these findings are consistent with simulation results by Ray et al. (2003).

Scattered samples, however, contain a clear signal of the species-wide expansion at any level of gene flow. The reason is that the coalescent of scattered samples almost behaves like a neutral one with a population size proportional to the number of demes. Hence, this coalescent picks up the signal of expansion as in the panmictic case. The same should hold for any sampling scheme under sufficiently high migration, but Figure 8 shows that even under high levels of gene flow ($4N_0m = 100$), the site frequency spectrum of local samples is still different from that of pooled and scattered samples. Moreover, the simulation results plotted in Figure 8 were obtained under a 10-fold expansion, and we may expect discrepancies between local and scattered samples to extend to even higher migration rates with higher expansion factors (see Figure 9). Again, D_T and D_{FL} values obtained for pooled samples are intermediate between those of local and scattered samples except for very low migration rates (~4N0m < 0.5), similar to the case of equilibrium subdivided populations (Figure 7 and Figure 8). These simulation results imply that both local and pooled samples may be expected to underestimate the extent of any species-wide (range) expansion to various degrees, depending on levels of gene flow connecting the demes and the age and magnitude of the expansion. The latter aspect, however, is influenced by our choice of simulating an instantaneous expansion followed by a period of constant population size until the time of sampling; an exponential expansion scheme until the present would yield higher detectability of expansion from local and pooled samples.

We point out that quantitative details of our simulation results for the three sampling schemes depend to some extent on our choice of sampling 20 sequences distributed over one (“local”), four (pooled), and 20 demes (“scattered”), respectively.

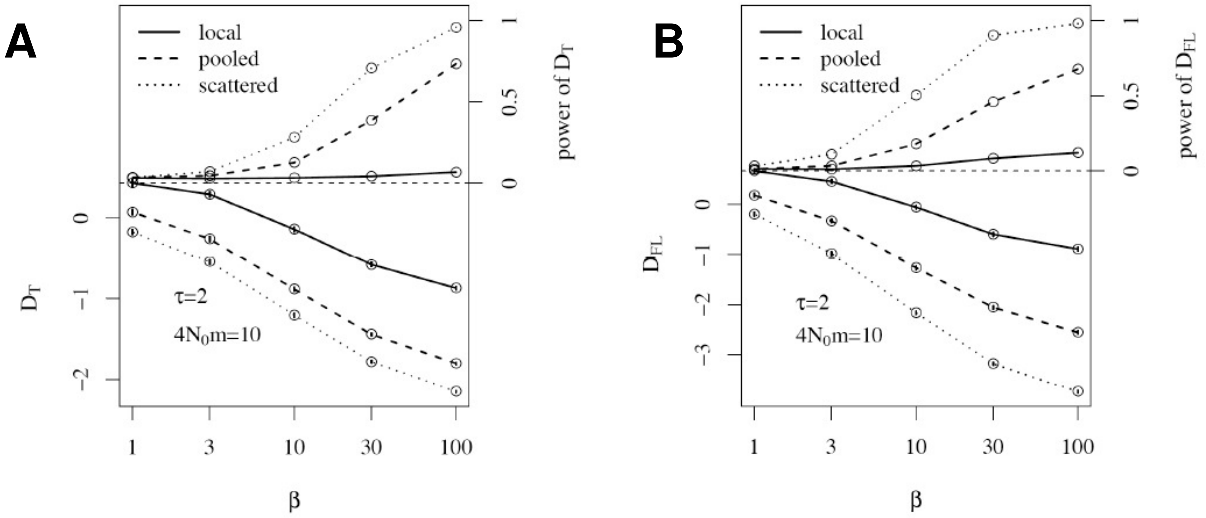


Figure 9 Averages of Tajima's D_T (A) and Fu and Li's D_{FL} (B) as functions of the expansion factor β (with fixed migration rate and time of expansion), with the power of these statistics evaluated in the top part of each plot (right Y-axis; power was assessed at a level of P = 0.05). The same simulation scheme as in Figure 8 was used.

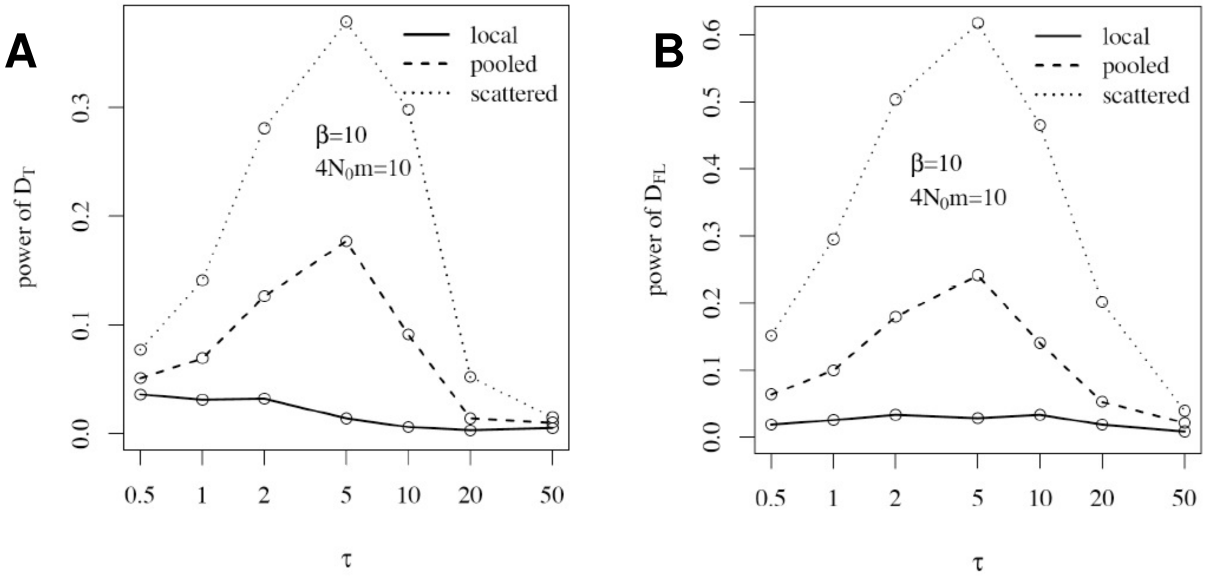


Figure 10 The power of Tajima's D_T (A) and Fu and Li's D_{FL} (B) as a function of the expansion time τ (with fixed migration rate and expansion factor). The same simulation scheme as in Figure 8 was used; power was assessed at a level of P = 0.05.

Generally speaking, decreasing the number of sequences sampled per deme and/or increasing the number of demes that sequences are sampled from shifts the site frequency spectrum more toward that of scattered samples, reflecting the diminished impact of the scattering phase of the coalescent process on such samples. The exact numerical composition of local and pooled samples also affects the migration rate at which the expected D_T and D_{FL} values for pooled samples drop below those for local samples (see Figure 7 and Figure 8).

3. Power of D_T and D_{FL} to detect expansion under different sampling schemes and expansion times

Illustrated by an intermediate level of interdeme migration ($4N_0m = 10$), we assessed the power of the test statistics D_T and D_{FL} under a range of expansion factors and times of expansion. For the three sampling schemes, Figure 9 summarizes the power of D_T and D_{FL} assuming an expansion time of $t = 8N_0$ generations ago. The low power of local samples to detect significant departures from the standard neutral model regardless of the magnitude of the species-wide expansion is striking; qualitatively, this is consistent with results by Ray et al. (2003). Even if the species-wide expansion was 100-fold, local samples deviate from the standard neutral model expectations in only 8% (for D_T) and 12% (for D_{FL}) of all cases under these conditions. In sharp contrast, scattered samples deviate from standard neutral expectations in 98% (99%) of all cases for $\beta = 100$.

Next, we illustrate the effect of the timing of the expansion for a fixed 10-fold expansion and a migration rate of $4N_0m = 10$. Under these conditions, expansion times in the approximate range $1 < \tau < 15$ can be detected in principle, but again with striking differences in power exhibited by local vs. scattered samples (Figure 10). The shape of the curves in Figure 10 can be explained intuitively: if τ is very small (*i.e.*, establishment of population structure was very recent), samples appear to be drawn from a panmictic population of constant size. On the other hand, if τ is very large, samples appear to be drawn from an equilibrium subdivided population and hence the expansion cannot be detected. Under the island model, all results are qualitatively the same but with even smaller power to reject stable population size for local samples (at most 10% less power; data not shown). As these power assessments are based on simulated sample genealogies of single loci, the actual power available with empirical multilocus data would be correspondingly higher.

All simulation results appear to depend only weakly on the number of demes, as long as I >> n. However, an increase in the number of demes carries important connotations, as the genealogical signatures of past expansions remain detectable for longer time periods than with fewer demes. For example, simulating with a constant ratio $\tau/I = 0.02$ (*i.e.*, equivalent to $\tau = 2$ for $I = 100$, $\tau = 10$ for $I = 500$, etc.) under otherwise equal demographic conditions, we obtained fairly constant but sampling-specific estimates of D_T and D_{FL} (data not shown). One interpretation is that the effective population size is proportional to I but depends on the sampling scheme. These observations imply approximately equal detectability of expansions (for a given sampling scheme) over a large range of I and thus the dependency of "relevant" τ -values on I. This latter effect is a direct consequence of lengthening the collecting phase of the coalescent process with increasing numbers of demes in the total population.

Second Project

1. K_a/K_s ratios and McDonald-Kreitmen tests

Taking divergence between species (or to an outgroup) into account, the K_a/K_s ratios are lower than one for all species and all loci (data not shown). Loci CT166 and CT208 contain zero or few non-synonymous SNPs, indicating that they are under strong purifying selection. The McDonald-Kreitman test does not show significant departure from neutrality, except for two marginally significant cases (data not shown). At locus CT066 in *S. habrochaites*, a higher silent diversity than expected is observed (Fishers exact test, $P < 0.05$). A higher non-synonymous diversity than expected is detected for CT198 in *S. arcanum* (Fishers exact test, $P < 0.05$). The latter may hint at positive selection.

2. Pooling effect and purifying selection

The pooling of several population samples shows an excess of low-frequency (derived) variants at synonymous polymorphic sites (Table 3). This is because the pooled sample for synonymous sites is primarily affected by species-wide demographic events such as population size expansion (Stadler et al. 2009). Very negative Tajima's D_T values at synonymous (S) sites for *S. peruvianum* and *S. arcanum* indicate strong expansion as opposed to *S. chilense* and *S. habrochaites* (Table 3; see average across loci).

With the exception of *S. arcanum*, D_T values of all sites are smaller than those of synonymous sites for most loci (and for the average across loci), whereas silent sites exhibit values that are on average close to those of synonymous sites or only slightly smaller (Table 3). Thus comparison of S and silent sites (S+NC) indicates that non-coding regions exhibit a slight excess of low-frequency polymorphisms at the species level (except in *S. arcanum*).

Species	Sites	CT066	CT093	CT166	CT179	CT198	CT208	CT251	CT268	Average across Loci[a]
S. peruvianum	*synonymous*	-0.485	-1.436	**-1.879***	-0.659	-0.205	-0.981	-0.392	-0.527	-0.726
	silent	-0.485	-1.680	-1.674	-0.655	-0.792	-1.129	-0.781	-0.527	-0.954
	all sites	-0.561	-1.688	-1.683	-0.755	-0.783	-1.149	-0.938	-0.797	-1.052
S. chilense	*synonymous*	-0.293	-0.49	-0.560	0.010	-0.858	-1.480	0.746	0.808	0.056
	silent	-0.293	-0.858	0.006	-0.634	-1.099	-0.410	0.097	0.808	-0.163
	all sites	-0.661	-1.400	-0.103	-0.714	-1.075	-0.410	-0.246	0.065	-0.497
S. habrochaites	*synonymous*	-0.225	-1.360	-1.164	-0.241	-0.147	0.210	0.133	0.214	-0.256
	silent	-0.225	-1.623	-1.444	-1.198	0.202	-0.654	0.231	0.214	-0.423
	all sites	-0.389	-1.773	-1.480	-1.184	0.19	-0.654	-0.624	-0.140	-0.675
S. arcanum	*synonymous*	0.190	-1.717	-1.041	0.502	-1.44	-0.938	-0.753	-1.286	-0.812
	silent	0.190	-1.718	-0.240	0.347	-0.268	-1.775	-0.894	-1.286	-0.795
	all sites	0.032	-1.820	-0.240	0.237	-0.207	-1.775	-0.964	-1.108	-0.814

Table 3 Values of Tajima's DT per locus and per species for the pooled samples. SNPs are grouped by categories: synonymous, silent sites (non-coding and synonymous), and all sites. *na:* non applicable (absence of coding region or non-synonymous sites). * significance level of *P*<0.05 [a] Weighted average across 8 loci.

All polymorphic sites (S+NC+NS) show a smaller D_T than silent sites, indicating that purifying selection acts on coding regions by keeping NS deleterious mutations at low frequency (except in *S. arcanum*). Few comparisons, however, do not follow this trend because some loci (CT066 and CT268) have only coding region or have only one NS site (CT208 in *S. peruvianum*).

3. Ancestral polymorphism and deleterious mutations

The distribution of polymorphism between pairs of species indicate that most of the shared polymorphisms are high-frequency variants (f > 20%) for all types of sites (S*, NS* and NC*). This is reflected in the scale of the y and x axes of Figure 11b (common polymorphisms) having a ten fold difference compared to that of Figure 11a (rare and intermediate polymorphisms). Not surprisingly, very few rare shared NS sites are found in species comparisons because since the time of divergence, selection keeps deleterious mutations at low frequency, and they are subject to elimination by drift in one or both species. The five other regressions remaining for intermediate and common NS sites, as well as NC sites show regression slopes significantly different from 1 (P <0.001) (Figure 11). Purifying selection is thus acting in coding regions on NS mutations and in non-coding regions.

Similar regression slopes are found for NS* as a function of S* for intermediate and common frequencies, indicating that non-synonymous SNPs shared among species are neutral (or very mildly deleterious; Figure 11). The slope of the regression for NS* as function of S* is 0.137 for intermediate and 0.132 for common-frequency SNPs, indicating that approximately 87% of the shared non-synonymous SNPs between species are under purifying selection.

Selection against deleterious mutations in coding regions is stronger than in non-coding regions as the comparisons of NC* and S* have higher regression slopes than NS* and S* (Figure 11). The rationale being that for a given shared amount of S* between two species higher amounts of shared NC* than NS* are found because purifying selection is weaker at NC sites than NS sites. Approximately 52% of the non-coding sites can be estimated as being under purifying selection (one minus the slope of the common NC* regression). There is an excess of NC* shared polymorphisms for the comparison *S. peruvianum*-*S. arcanum* (Figure 11a). When this point is removed, the regression equation for NC*and S* for intermediate-frequency sites becomes y=0.578x-0.067 (R^2 = 0.942), whose slope is then not different from the other regressions slopes for NC* and S* (Figure 11). Such a high amount of NC sites shared between *S. peruvianum* and *S. arcanum* (Figure 11a) at intermediate polymorphism frequency could indicate recent introgression between these two species at one or several introns.

4. Distribution of fitness effects

Estimates of the demographic parameters reveal that two out of four species have undergone expansion. The strongest expansion is found for *S. peruvianum* with a 10-fold expansion (Table 4, N_2/N_1=10), in accordance with the negative D_T at synonymous sites (Table 3). *S. arcanum* exhibits a 3-fold and very recent expansion. *S. habrochaites* shows evidence for a very recent expansion (barely negative D_T in Table 3). However, *S. chilense* shows a 10-fold expansion that occurred much further in the past (3.5 times older) than for *S. peruvianum* (Table 4). The signature of expansion in this species is thus not detectable by D_T (Table 3).

One minus the ratio NC*/S* indicate the percentages of sites under selection in non-coding regions: 64, 29, 30, and 35% for *S. peruvianum*, *S. chilense*, *S. habrochaites*, and *S. arcanum*, respectively. For non-coding sites, the DFE has a negative mean for all species ($-N_eE(s)$), confirming that negative purifying selection is the main force acting on the evolution of these intronic sequences. The strength of purifying selection is weak in *S. habrochaites*, *S. chilense*, and *S. peruvianum* [$-N_eE(s)$ is between 2 and 340], which is several orders of magnitude lower than in *S. arcanum* (Table 4). The value of $-N_eE(s)$ given for *S. arcanum* being very large, we do not think that such measure is realistic and meaningful. However, from the distribution of fitness effect of mutations, it is clear that new mutations in *S. arcanum* are either neutral or very strongly deleterious, indicative of a huge variance of the DFE. The shape of the distribution of fitness effects for non-coding sites is a negative exponential for all species, meaning that most of the mutations have very mildly deleterious effects (0> N_eE(s)> −1) (Table 4). However, the exponential DFE in *S. arcanum* exhibits a majority of very mildly deleterious mutations (58%) but also 36% of strongly deleterious fitness effects (–N_eE(s)>100). Note that *S. peruvianum* and *S. chilense* have a very similar DFE that is substantially different from that of *S. arcanum*.

		S. peruvianum	*S. chilense*	*S. habrochaites*	*S. arcanum*
N_2/N_1		10	10	8.79	3.07
t/N_1		8.27	26.83	0.031	2.21
−N_eE(s) (mean effective selection intensity)		337	79.5	2.73	**nqr*
b (shape of gamma distribution)		0.072	0.079	0.448	0.02
f_0 (proportion of invariants sites)		0.61	0.81	0.84	0.8
LogL (log likelihood)		-2593	-1877	-1139	-1865
Proportion of mutants in −N_eE(s) range	0 – 1 (very mildly deleterious)	0.565	0.603	0.478	0.582
	1 – 10	0.102	0.12	0.462	0.027
	10 – 100	0.119	0.139	0.06	0.029
	>100 (very strongly deleterious)	0.214	0.138	0	0.362

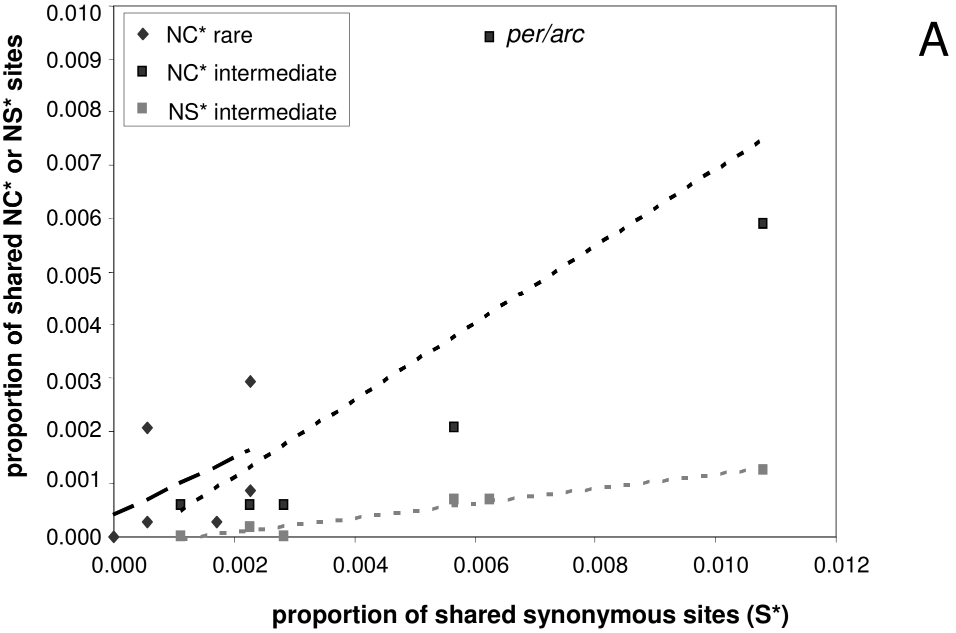
Table 4 Analysis of the demographic and DFE parameters for the non-coding sites (Keightley & Eyre-Walker 2007). The ratio of current and ancestral effective population size (N2/N1) and the ratio of time of expansion (t/N1). The proportion of mutants within certain range of selection coefficients is given as intervals of˜ NeE(s). Coefficients of˜ NeE(s) between 0 and 1 represent very mildly deleterious mutations, and for >100 mutations with very strong deleterious effects on fitness. **nqr*: non quantitatively relevant. The mean N_eE(s) given has a value greater than 10^9 which is not quantitatively relevant as a population genetics parameter.

In coding regions, the strength of purifying selection is more homogeneous among species and very similar between *S. peruvianum*, *S. habrochaites*, and *S. arcanum*. The percentage of sites under purifying selection for coding regions are 90, 87, 91, and 89%, respectively, for *S. peruvianum*, *S. chilense*, *S. habrochaites*, and *S. arcanum*. The selection coefficients –N_eE(s) for NS sites are several orders of magnitude larger than for NC sites except for *S. arcanum* (Table 5). Most new NS mutations have highly deleterious effects on fitness (Table 5). However, *S. chilense* show selection coefficients [N_eE(s)= –318] that are several orders of magnitude weaker than for the other three species [N_eE(s)>2×10^4]. Note that *S. arcanum* has experienced a similar strength of selection (and DFE) as *S. peruvianum* despite having a smaller effective population size. Note that in general, the log-likelihood values of the parameter estimates are lower for *S. peruvianum* than for other species.

		S. peruvianum	*S. chilense*	*S. habrochaites*	*S. arcanum*
$-N_e$E(s) (mean effective selection intensity)		18.5×10^5	318	2.19×10^4	1.65×10^5
b (shape of gamma distribution)		0.138	0.38	0.164	0.157
f_0 (proportion of invariants sites)		0.62	0.63	0.92	0.78
LogL (log likelihood)		-2246	-1890	-1073	-1446
Proportion of mutants in $-N_e$E(s) range	0 – 1 (very mildly deleterious)	0.109	0.087	0.155	0.121
	1 – 10	0.041	0.122	0.071	0.053
	10 – 100	0.057	0.277	0.104	0.076
	>100 (very strongly deleterious)	0.793	0.514	0.67	0.75

Table 5 Analysis of the DFE parameters for the non-synonymous sites (Keightley & Eyre-Walker 2007).

The proportion of adaptively driven substitutions (α) is negative or very close to zero in *S. arcanum*, and *S. habrochaites* similar to values found by the method of Bierne & Eyre-Walker (2004; data not shown). Discrepancies are found in estimates of α for *S. peruvianum* and *S. chilense* between the two methods (data not shown). Note, however, that all confidence intervals obtained by maximum likelihood with the DoFE software of Bierne & Eyre-Walker (2004) comprise α values found by the method of (Eyre-Walker & Keightley 2009), and are centered around zero. This indicates that these eight housekeeping genes do not show evidence for strong positive selection.



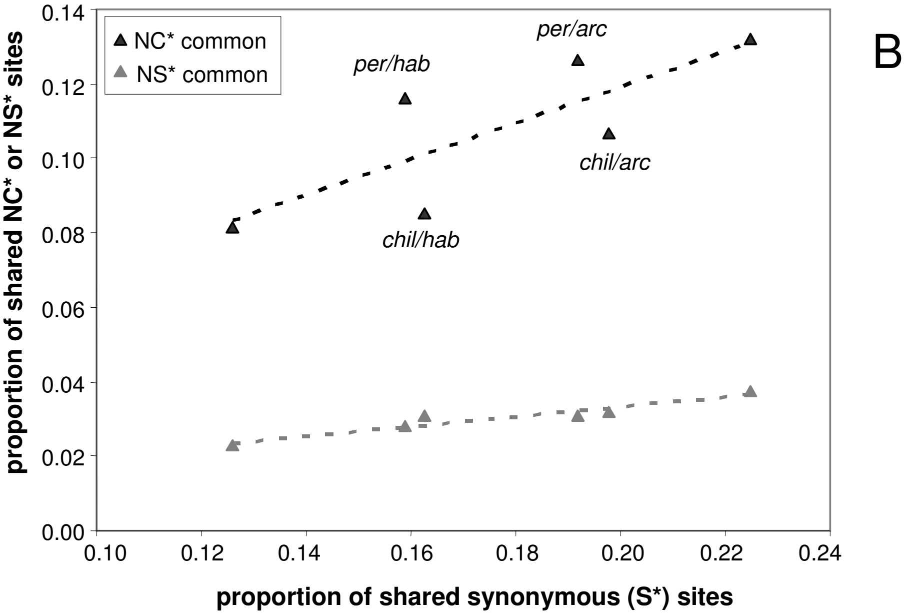
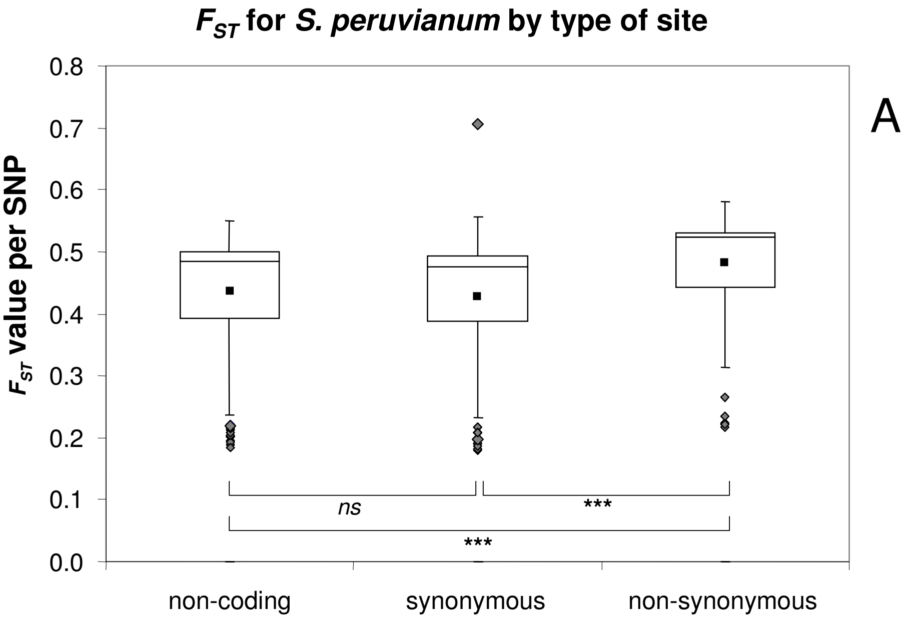
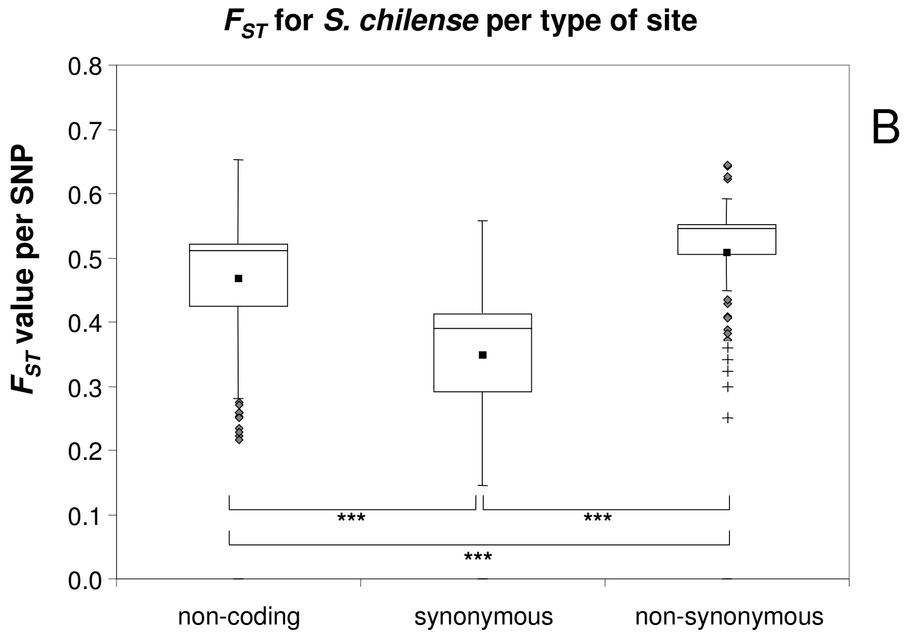


Figure 11 Amount of shared polymorphic sites for the six pairwise comparisons between species. The three frequency classes are rare (SNP frequency <5%, diamonds), intermediate (5%<SNP frequency<20%, rectangle), and common (SNP frequency>20%, triangle). The proportion of non-coding polymorphic sites (NC*) is in black, and non-synonymous sites (NS*) in grey. Regression equations and R2 are indicated. A. The rare NC* as a function of S* (y=0.534x+0.04; R2 =0.19) indicated by the black long dashed line. The intermediate NC* as a function of S* (y=0.725x-0.03; R2 =0.36) is denoted by the black short dashed line, and the intermediate NS* as a function of S* (y=0.137x-0.02; R2 =0.94) by the grey short dashed line. The pairwise comparison between *S. peruvianum* and *S. arcanum* is highlighted for intermediate NC* (per/arc). B. The common NC* as a function of S* (y=0.481x+2.28; R2 =0.55) is denoted by the black short dashed line, and the common NS* as a function of S* (y=0.132x+0.69; R2 =0.88) by the grey short dashed line. Pairwise comparisons between *S. chilense* (or *S. peruvianum*) with *S. arcanum* and *S. habrochaites* are shown respectively as chil/arc and chil/hab (and per/arc and per/hab).



F_{ST} for *S. peruvianum* by type of site
A
F_{ST} value per SNP
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
ns
***
***
non-coding
synonymous
non-synonymous



F_{ST} for *S. chilense* per type of site
B
F_{ST} value per SNP
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
***
***
***
non-coding
synonymous
non-synonymous

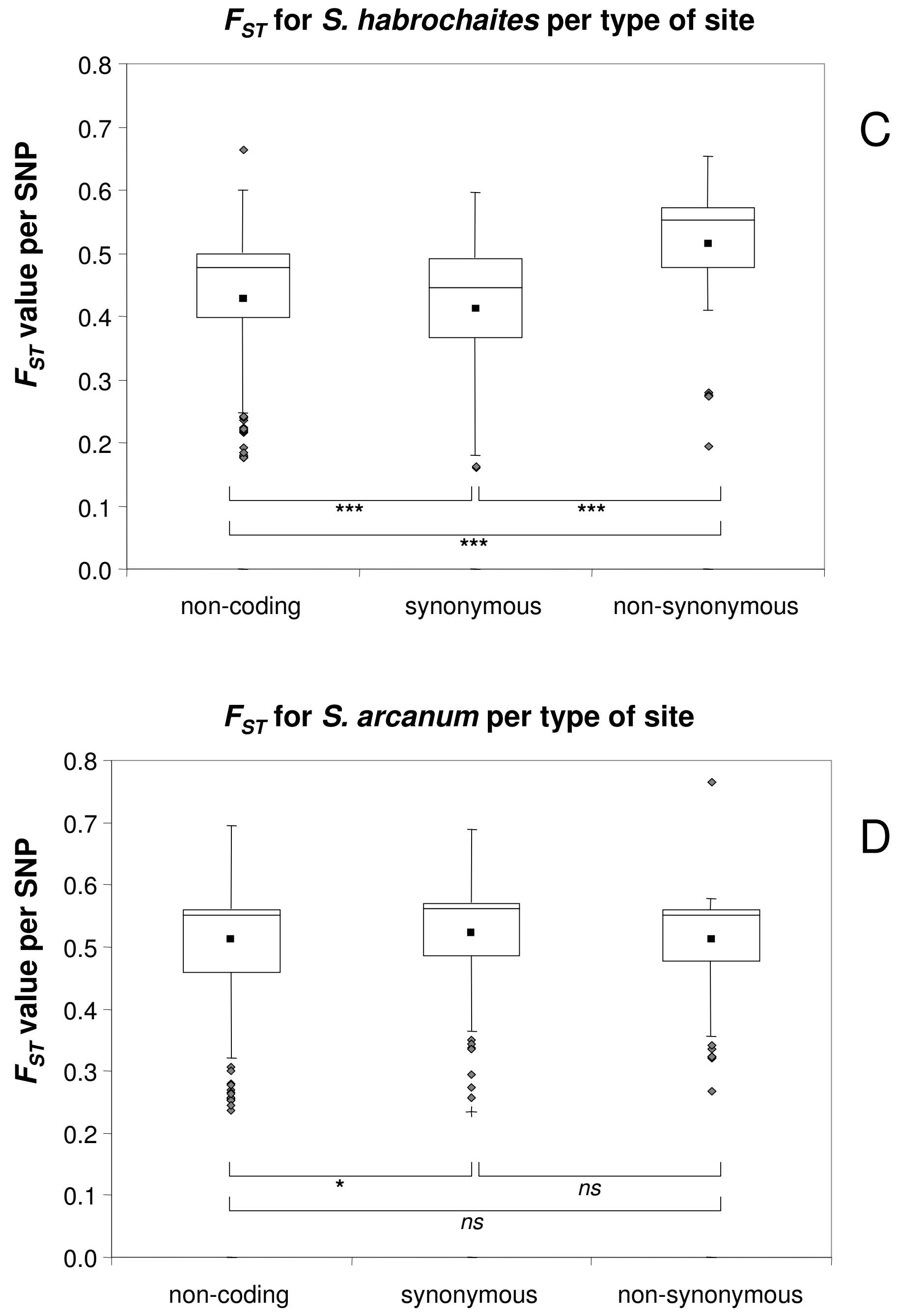


Figure 12 Boxplot of FST distribution for each polymorphic site (SNP) for non-coding, synonymous, and non-synonymous sites. FST values that are more than 1.5 times the interquartile range from the nearest quartile are displayed as diamonds. For more than 3 times the interquartile range, they are displayed as crosses. The mean of the distribution is indicated by a full black rectangle.

The three pairwise Wilcoxon tests to determine which type of site have higher FST values are indicated. The P-values are corrected (divided by 3 following a Bonferroni correction) and significance levels are indicated as follows: * P-value <0.1; *** P-value <0.001; ns =non-significant. A for *S. peruvianum*, B for *S. chilense*, C for *S. habrochaites*, and D for *S. arcanum*.

5. Purifying selection and metapopulation structure

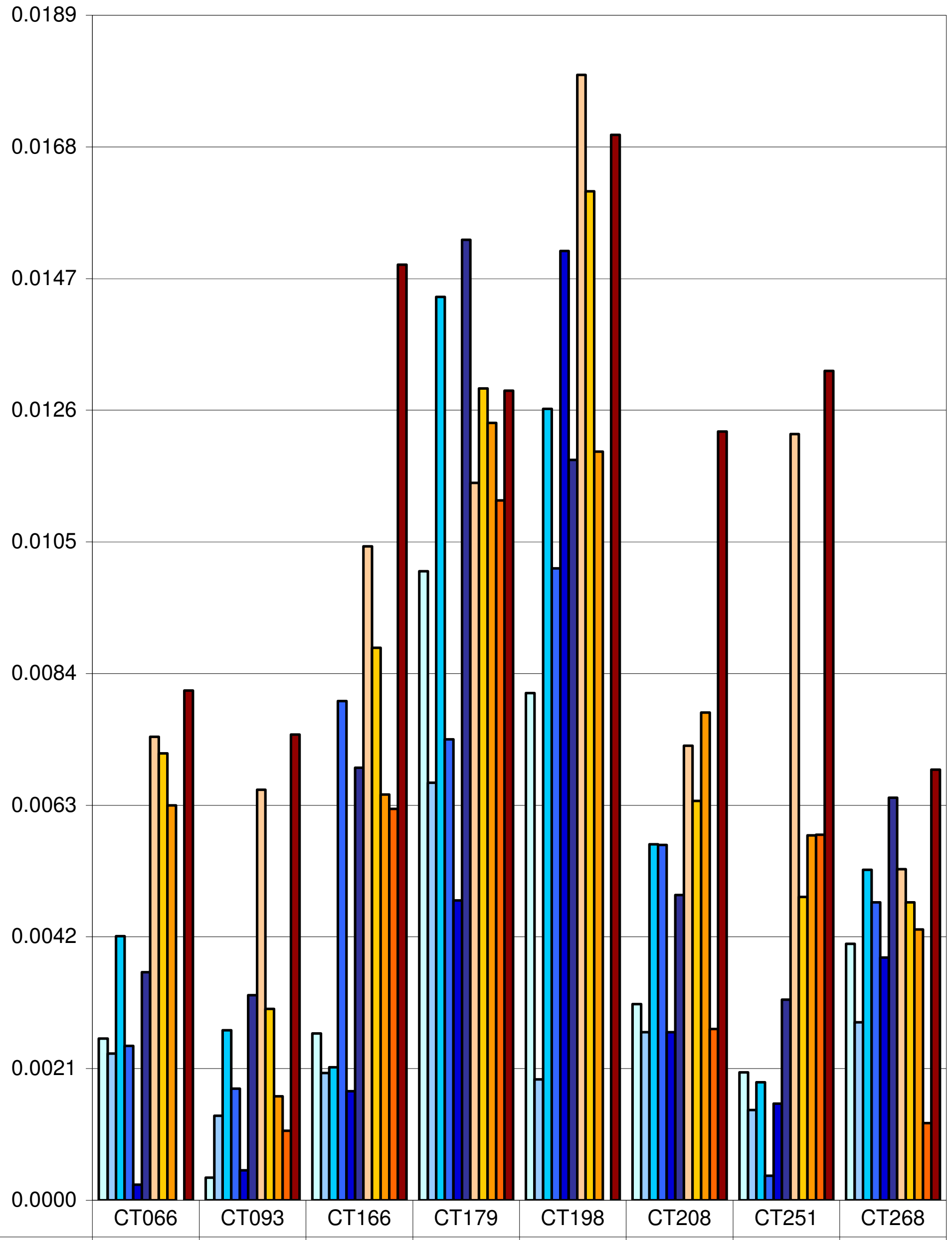
Differences in F_{ST} between synonymous and NC or NS sites are due to the effect of purifying selection increasing the number of low frequency polymorphisms, which will be found as private polymorphism in one population, and increase differentiation among populations (F_{ST}). For all species the distribution of F_{ST} depends the types of site (NC, S or NS) with highly significant P-values ($P<0.001$), *S. arcanum* having the highest P-value ($P=0.00027$; Figure 12). However, not all pairwise comparisons show differences between the distributions of the types of sites in the four species. In *S. peruvianum*, F_{ST} values for synonymous sites are lower than for non-synonymous but not different from non-coding sites (Figure 12a). In *S. chilense* and *S. habrochaites*, all pairwise comparisons show higher F_{ST} distributions for NC and NS sites compared to S sites, with NC sites being intermediate between low F_{ST} at S sites and high at NS sites (Figure 12b, c). This indicates that in these species there is an excess of private polymorphism (thus high F_{ST}) at NC and NS sites compared to synonymous sites. Finally, in *S. arcanum* no significant difference could be detected between NS and S sites (Figure 12d). NS and NC sites have a lower mean F_{ST} value compared to synonymous sites only in *S. arcanum*. Note that the power of these non-parametric tests is low, especially due to the non-normal (highly skewed) distributions of F_{ST} values. Using BayeScan on polymorphic SNPs with frequency higher than 5% (in the species) reveals a clear pattern for differences in the F_{st} distribution between types of sites in *S. chilense*. Such a pattern was not observed in any of the three other species, where all F_{ST} values where similar for S, NC and NS sites (data not shown).

The selection pressure acting on NS or NS sites can be observed at the population level when calculating the ratios NS*/S* and NC*/S* for private polymorphisms for each population. Low ratios would indicate high levels of purifying selection, and high ratios (> 1) would show weak selection or relaxed selective constrains. Species with higher effective population size (*S. peruvianum* and *S. chilense*) do not show strong variation of these ratios among populations (data not shown). This tends to indicate that levels of purifying selection are homogeneous among populations in these two species. However, in *S. arcanum* and *S. habrochaites* variable ratios are observed for several different populations, mainly showing a noticeable excess (up to six times) of NC or NS sites compared to the number of S sites. Whether this reflects variability in the strength of selection among populations, or random effects of drift cannot be assessed here due to small sampling sizes.

Third Project

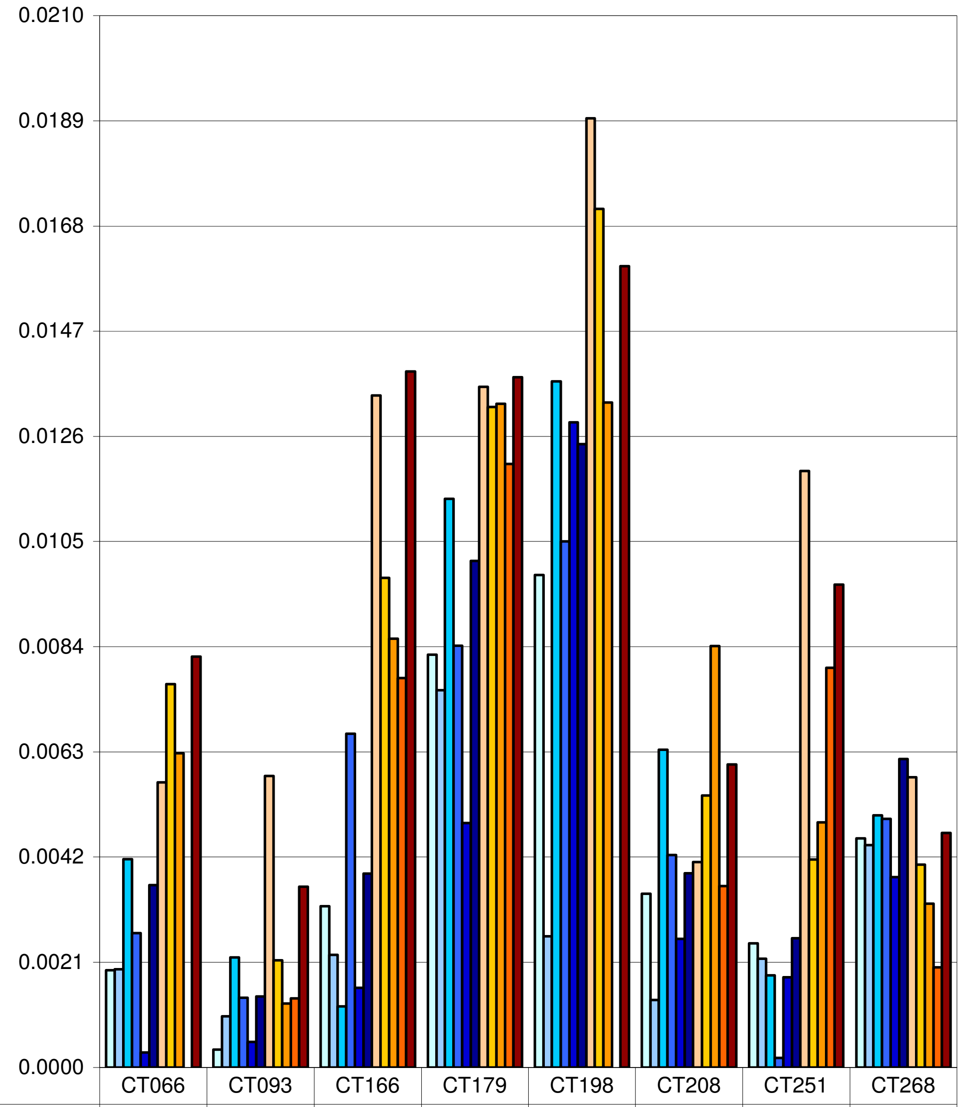
1. Estimation of Nucleotide Diversity and Neutrality Tests

Diversity levels measured through Watterson's θ (Watterson 1975; Figure 13) are similar to those found in *S. peruvianum* and *S. chilense* (Stadler et al. 2008). Similarly, π (Nei 1987; Figure 14) shows levels of diversity per locus corresponding to those estimated through θ_w. Note that the Cochabamba population of *S. arcanum* exhibits no polymorphism at two loci (CT066 and CT198). Also noteworthy, θ_W values for the pooled sample (*i.e.*, all populations from a species analyzed together as a single population) in *S. arcanum* show higher values than any single populations from the species. *S. habrochaites* shows a similar, although less evident, tendency. This indicates that private alleles are found in each population due to limited gene flow among demes. By pooling populations of the species, high amount of private polymorphisms will contribute to increase significantly θ_{W}. This is not the case for π, because pairwise comparisons take into account the frequency of alleles. In short, by pooling all populations from a species, an excess of low-frequency polymorphism is created.



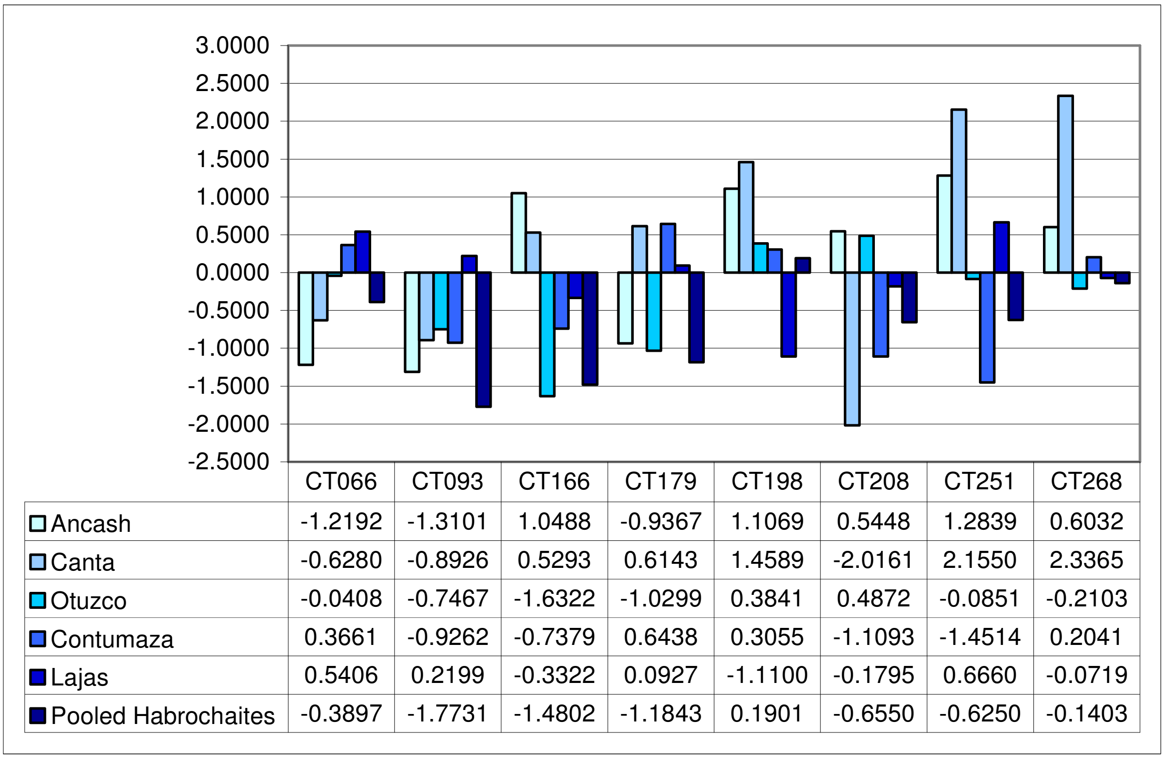
	CT066	CT093	CT166	CT179	CT198	CT208	CT251	CT268
■ Ancash	0.0026	0.0004	0.0027	0.0100	0.0081	0.0031	0.0020	0.0041
■ Canta	0.0023	0.0014	0.0020	0.0067	0.0019	0.0027	0.0014	0.0028
■ Otuzco	0.0042	0.0027	0.0021	0.0144	0.0126	0.0057	0.0019	0.0053
■ Contumaza	0.0025	0.0018	0.0080	0.0074	0.0101	0.0057	0.0004	0.0048
■ Lajas	0.0003	0.0005	0.0017	0.0048	0.0151	0.0027	0.0015	0.0039
■ Pooled Habrochaites	0.0036	0.0033	0.0069	0.0153	0.0118	0.0049	0.0032	0.0064
■ Otuzco	0.0074	0.0066	0.0104	0.0114	0.0180	0.0073	0.0122	0.0053
■ Rupe	0.0071	0.0031	0.0088	0.0130	0.0161	0.0064	0.0048	0.0048
■ San Juan	0.0063	0.0017	0.0065	0.0124	0.0119	0.0078	0.0058	0.0043
■ Cochababmba	0.0000	0.0011	0.0062	0.0112	0.0000	0.0027	0.0058	0.0012
■ Pooled Arcanum	0.0081	0.0074	0.0149	0.0129	0.0170	0.0123	0.0132	0.0069

Figure 13 Watterson's Theta values per population and locus for *S. habrochaites* and *S. arcanum*. Habrochaites and Arcanum indicate the pooled sample (see text) of all populations of each species.



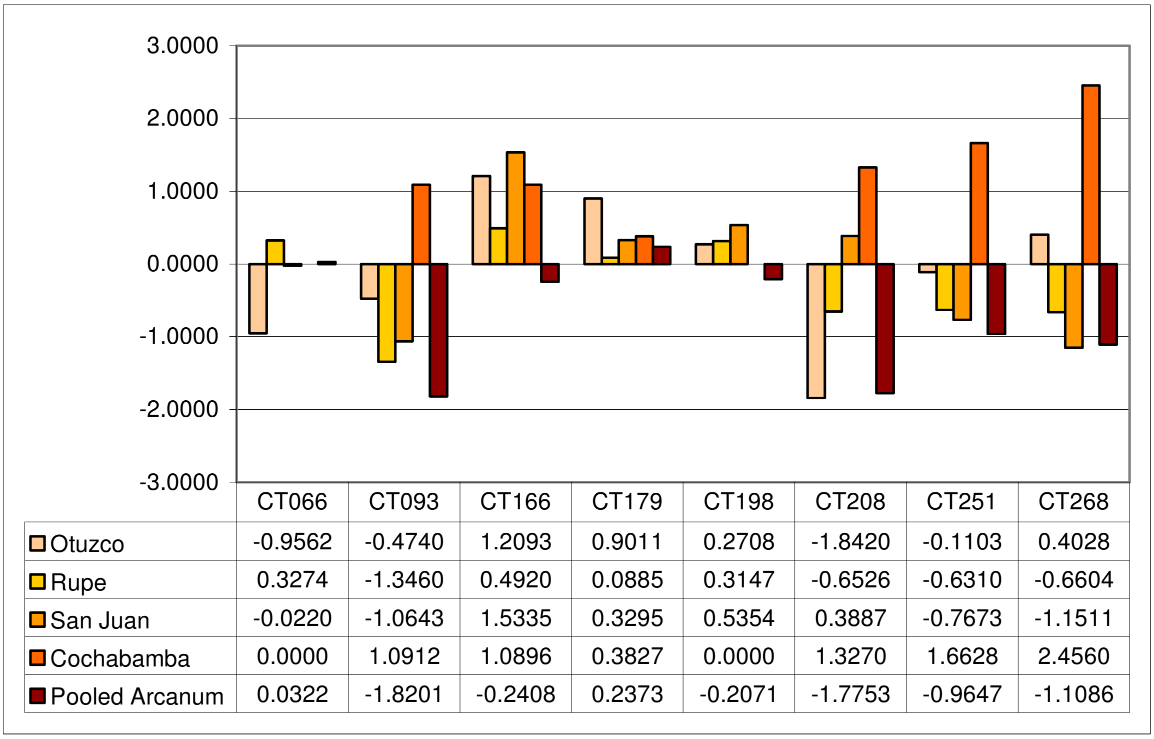
	CT066	CT093	CT166	CT179	CT198	CT208	CT251	CT268
□ Ancash	0.0019	0.0004	0.0032	0.0082	0.0098	0.0035	0.0025	0.0046
■ Canta	0.0020	0.0010	0.0023	0.0075	0.0026	0.0014	0.0022	0.0044
■ Otuzco	0.0042	0.0022	0.0012	0.0114	0.0137	0.0063	0.0018	0.0050
■ Contumaza	0.0027	0.0014	0.0067	0.0084	0.0105	0.0042	0.0002	0.0050
■ Lajas	0.0003	0.0005	0.0016	0.0049	0.0129	0.0026	0.0018	0.0038
■ Pooled Habrochaites	0.0036	0.0014	0.0039	0.0101	0.0125	0.0039	0.0026	0.0062
■ Otuzco	0.0057	0.0058	0.0134	0.0136	0.0190	0.0041	0.0119	0.0058
■ Rupe	0.0077	0.0021	0.0098	0.0132	0.0171	0.0054	0.0042	0.0041
■ San Juan	0.0063	0.0013	0.0086	0.0133	0.0133	0.0084	0.0049	0.0033
■ Cochabamba	0.0000	0.0014	0.0078	0.0121	0.0000	0.0036	0.0080	0.0020
■ Pooled Arcanum	0.0082	0.0036	0.0139	0.0138	0.0160	0.0061	0.0096	0.0047

Figure 14 Average number of pairwise differences per population and locus for *S. habrochaites* and *S. arcanum*. Habrochaites and Arcanum indicate the pooled sample (see text) of all populations of a species.



	CT066	CT093	CT166	CT179	CT198	CT208	CT251	CT268
■ Ancash	-1.2192	-1.3101	1.0488	-0.9367	1.1069	0.5448	1.2839	0.6032
■ Canta	-0.6280	-0.8926	0.5293	0.6143	1.4589	-2.0161	2.1550	2.3365
■ Otuzco	-0.0408	-0.7467	-1.6322	-1.0299	0.3841	0.4872	-0.0851	-0.2103
■ Contumaza	0.3661	-0.9262	-0.7379	0.6438	0.3055	-1.1093	-1.4514	0.2041
■ Lajas	0.5406	0.2199	-0.3322	0.0927	-1.1100	-0.1795	0.6660	-0.0719
■ Pooled Habrochaites	-0.3897	-1.7731	-1.4802	-1.1843	0.1901	-0.6550	-0.6250	-0.1403

Figure 15 Tajima's D values per population for *S. habrochaites* across all loci analyzed. 'Pooled Habrochaites' refers to the pooled sample (see text).



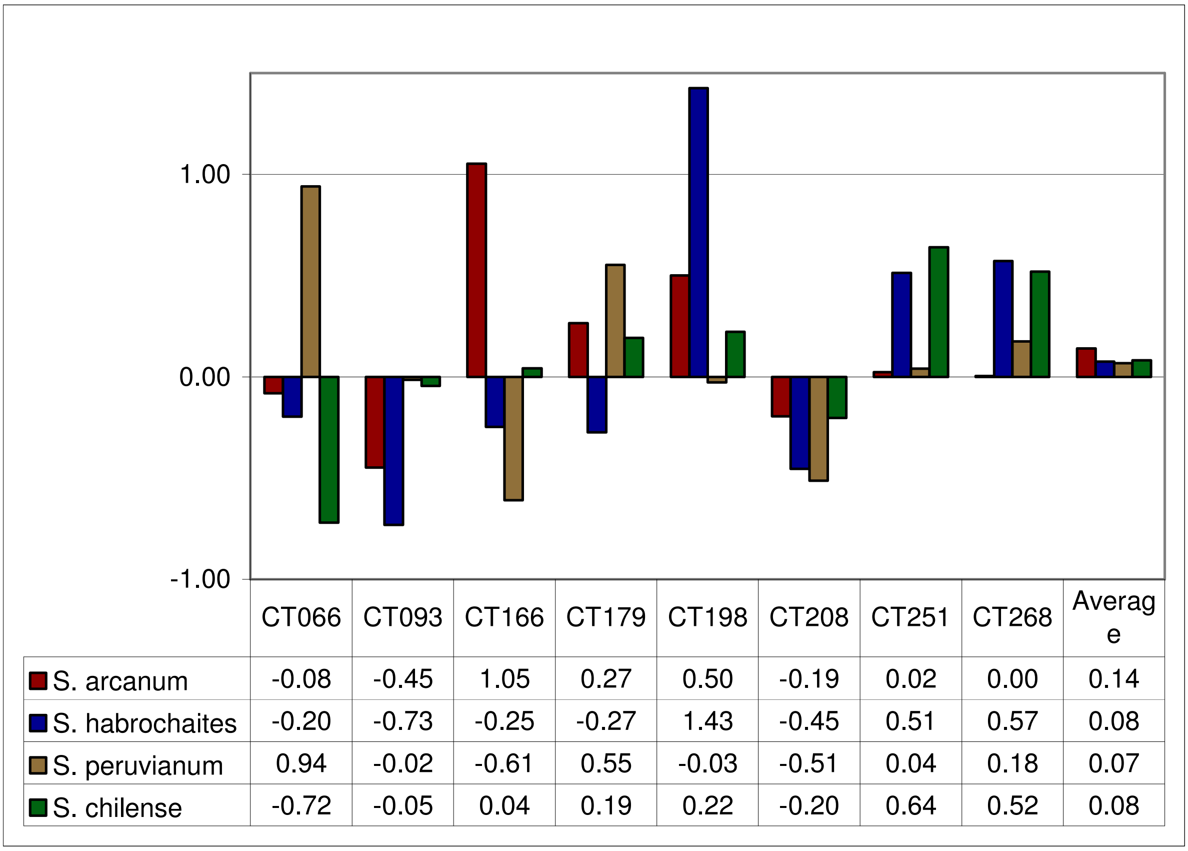
	CT066	CT093	CT166	CT179	CT198	CT208	CT251	CT268
■ Otuzco	-0.9562	-0.4740	1.2093	0.9011	0.2708	-1.8420	-0.1103	0.4028
■ Rupe	0.3274	-1.3460	0.4920	0.0885	0.3147	-0.6526	-0.6310	-0.6604
■ San Juan	-0.0220	-1.0643	1.5335	0.3295	0.5354	0.3887	-0.7673	-1.1511
■ Cochabamba	0.0000	1.0912	1.0896	0.3827	0.0000	1.3270	1.6628	2.4560
■ Pooled Arcanum	0.0322	-1.8201	-0.2408	0.2373	-0.2071	-1.7753	-0.9647	-1.1086

Figure 16 Tajima's D values per population for *S. arcanum* across all loci analyzed. 'Pooled Arcanum' refers to the pooled sample (see text).

Differences between loci are visible. Locus CT198 presents the highest degree of diversity, while locus CT093 shows the lowest. Diversity levels are largely affected by the locus and to a lesser degree by the species. This reflects different selective pressures per locus, *i.e.* loci under selective constraints would show less diversity than those under more relaxed conditions. Since there is a clearly stronger locus effect on diversity (than, for example, a species effect), we can hypothesize that this reflects differences in selective pressures between loci.

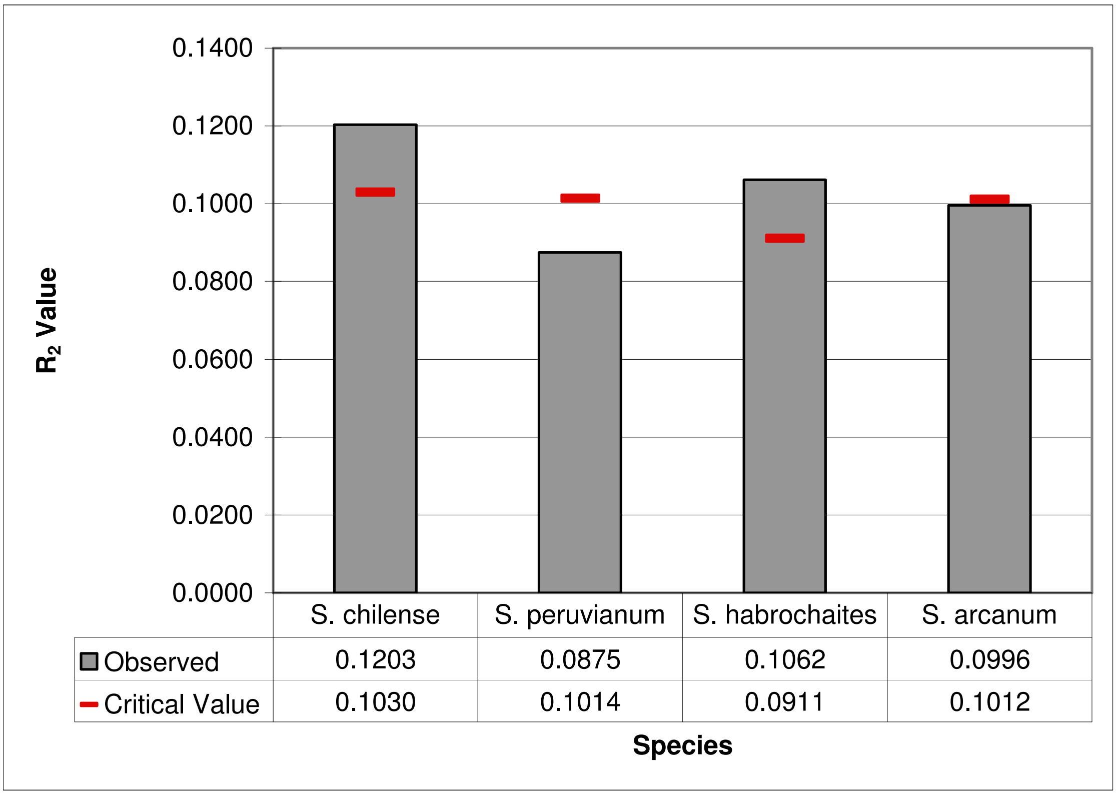
Tajima's D (Tajima 1989) test results are shown per population (Figure 15 and Figure 16) and per species (
Figure **17**). Most values do not significantly deviate from the standard neutral model, exceptions being the Canta population from *S. habrochaites* (for loci CT208, CT251 and CT268); the Otuzco population from *S. arcanum* (for locus CT208). Another population with a particular pattern is the Cochabamba population from *S. arcanum*. This population shows positive (although not significant, except in the case of CT268) values for all loci except CT066 and CT198, where no polymorphism is found. There is no particular locus exhibiting recurrent evidence of non-neutral values. Again, the pooled sample is informative (see above). For *S. arcanum*, a significantly negative value of Tajima's D is found for locus CT093. The distribution of the values for *S. arcanum* and *S. habrochaites* is in agreement with that found for *S. chilense* and *S. peruvianum* in previous studies (Arunyawat et al. 2007).

More intriguing yet is locus CT198. This locus presents a mutant allele with a premature stop codon at position 245 (located in the last third of the second exon, see Figure 19). The mutant allele presents much less polymorphism than "wild type" alleles although this locus present the highest amount of nucleotide diversity from all loci analyzed (also in *S. peruvianum* and *S. chilense*). However, coalescent simulations used to test the number of haplotypes (Nei 1987) showed no departure from neutral expectations (data not shown).



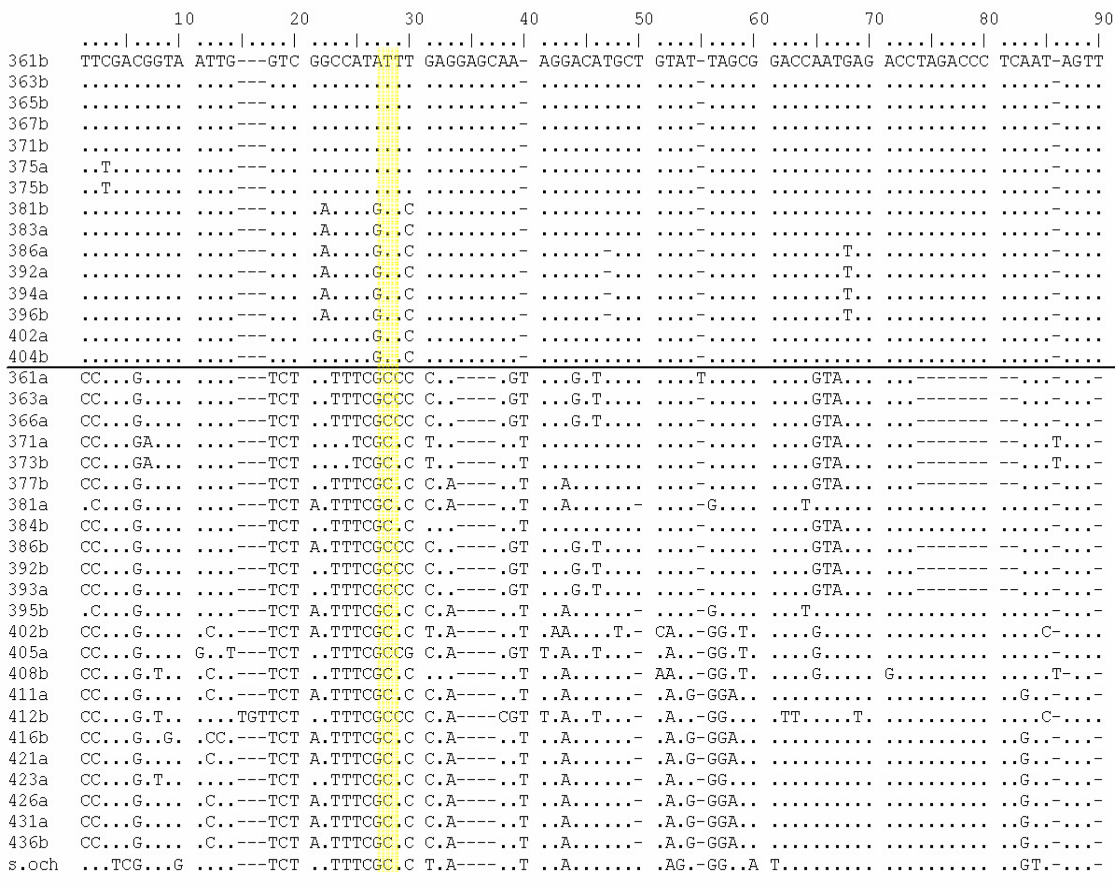
	CT066	CT093	CT166	CT179	CT198	CT208	CT251	CT268	Averag e
🟥 S. arcanum	-0.08	-0.45	1.05	0.27	0.50	-0.19	0.02	0.00	0.14
🟦 S. habrochaites	-0.20	-0.73	-0.25	-0.27	1.43	-0.45	0.51	0.57	0.08
🟫 S. peruvianum	0.94	-0.02	-0.61	0.55	-0.03	-0.51	0.04	0.18	0.07
🟩 S. chilense	-0.72	-0.05	0.04	0.19	0.22	-0.20	0.64	0.52	0.08

Figure **17** Tajima’s D values per locus averaged across all populations in each species.



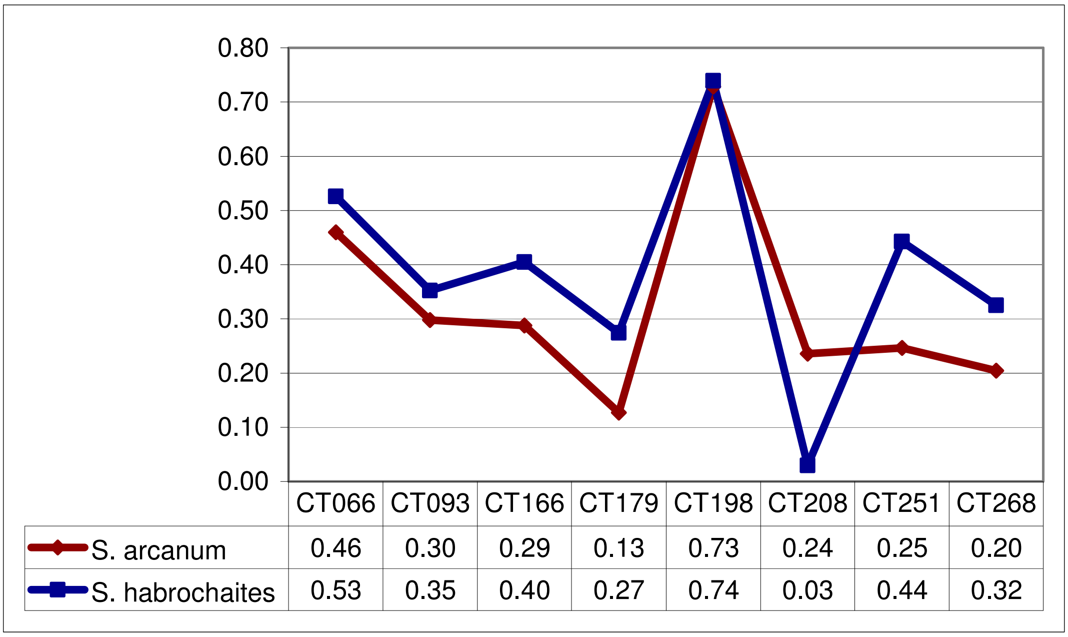
	S. chilense	S. peruvianum	S. habrochaites	S. arcanum
■ Observed	0.1203	0.0875	0.1062	0.0996
▬ Critical Value	0.1030	0.1014	0.0911	0.1012

Figure 18 Results of the R_2 test for population expansion. The red lines show the lower limit of the 95% confidence interval evaluated using coalescent simulations. Values under this limit are considered significant.



```
                 10         20         30         40         50         60         70         80         90
         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
361b     TTCGACGGTA ATTG---GTC GGCCATATTT GAGGAGCAA- AGGACATGCT GTAT-TAGCG GACCAATGAG ACCTAGACCC TCAAT-AGTT
363b     .......... ....---... .......... .........- .......... ....-..... .......... .......... .....-....
365b     .......... ....---... .......... .........- .......... ....-..... .......... .......... .....-....
367b     .......... ....---... .......... .........- .......... ....-..... .......... .......... .....-....
371b     .......... ....---... .......... .........- .......... ....-..... .......... .......... .....-....
375a     ..T....... ....---... .......... .........- .......... ....-..... .......... .......... .....-....
375b     ..T....... ....---... .......... .........- .......... ....-..... .......... .......... .....-....
381b     .......... ....---... .A....G..C .........- .......... ....-..... .......... .......... .....-....
383a     .......... ....---... .A....G..C .........- .......... ....-..... .......... .......... .....-....
386a     .......... ....---... .A....G..C .........- ......-... ....-..... .......T.. .......... .....-....
392a     .......... ....---... .A....G..C .........- ......-... ....-..... .......T.. .......... .....-....
394a     .......... ....---... .A....G..C .........- ......-... ....-..... .......T.. .......... .....-....
396b     .......... ....---... .A....G..C .........- ......-... ....-..... .......T.. .......... .....-....
402a     .......... ....---... ......G..C .........- .......... ....-..... .......... .......... .....-....
404b     .......... ....---... ......G..C .........- .......... ....-..... .......... .......... .....-....
361a     CC...G.... ....---TCT ..TTTCGCCC C..----.GT ...G.T.... ....T..... ....GTA... ...------- --...-...-
363a     CC...G.... ....---TCT ..TTTCGCCC C..----.GT ...G.T.... ....-..... ....GTA... ...------- --...-...-
366a     CC...G.... ....---TCT ..TTTCGCCC C..----.GT ...G.T.... ....-..... ....GTA... ...------- --...-...-
371a     CC...GA... ....---TCT ....TCGC.C T..----..T .......... ....-..... ....GTA... ...------- --...T...-
373b     CC...GA... ....---TCT ....TCGC.C T..----..T .......... ....-..... ....GTA... ...------- --...T...-
377b     CC...G.... ....---TCT ..TTTCGC.C C.A----..T ..A....... ....-..... ....GTA... ...------- --...-...-
381a     .C...G.... ....---TCT A.TTTCGC.C C.A----..T ..A......- ....-G.... ...T...... .......... .....-...-
384b     CC...G.... ....---TCT ..TTTCGC.C ...----..T .......... ....-..... ....GTA... .......... .....-...-
386b     CC...G.... ....---TCT A.TTTCGCCC C..----.GT ...G.T.... ....-..... ....GTA... ...------- --...-...-
392b     CC...G.... ....---TCT ..TTTCGCCC C..----.GT ...G.T.... ....-..... ....GTA... ...------- --...-...-
393a     CC...G.... ....---TCT ..TTTCGCCC C..----.GT ...G.T.... ....-..... ....GTA... ...------- --...-...-
395b     .C...G.... ....---TCT A.TTTCGC.C C.A----..T ..A......- ....-G.... ...T...... .......... .....-...-
402b     CC...G.... .C..---TCT A.TTTCGC.C T.A----..T .AA....T.- CA..-GG.T. ....G..... .......... ....C-....
405a     CC...G.... G..T---TCT ..TTTCGCCG C.A----.GT T.A..T...- .A..-GG.T. ....G..... .......... .....-....
408b     CC...G.T.. .C..---TCT ..TTTCGC.C ...----..T ..A......- AA..-GG.T. ....G..... G......... .....T-..-
411a     CC...G.... .C..---TCT A.TTTCGC.C C.A----..T ..A......- .A.G-GGA.. .......... .......... ..G..-...-
412b     CC...G.T.. ....TGTTCT ..TTTCGCCC C.A----CGT T.A..T...- .A..-GG... .TT.....T. .......... ....C-....
416b     CC...G..G. .CC.---TCT A.TTTCGC.C C.A----..T ..A......- .A.G-GGA.. .......... .......... ..G..-...-
421a     CC...G.... .C..---TCT A.TTTCGC.C C.A----..T ..A......- .A.G-GGA.. .......... .......... ..G..-...-
423a     CC...G.T.. ....---TCT ..TTTCGC.C C.A----..T ..A......- .A..-GG... .......... .......... ..G..-...-
426a     CC...G.... .C..---TCT A.TTTCGC.C C.A----..T ..A......- .A.G-GGA.. .......... .......... ..G..-...-
431a     CC...G.... .C..---TCT A.TTTCGC.C C.A----..T ..A......- .A.G-GGA.. .......... .......... ..G..-...-
436b     CC...G.... .C..---TCT A.TTTCGC.C C.A----..T ..A......- .A.G-GGA.. .......... .......... ..G..-...-
s.och    ...TCG...G ....---TCT ..TTTCGC.C T.A----..T ..A....... .AG.-GG..A T......... .......... ..GT.-...-
```

Figure 19 Image showing part of the alignment of polymorphic sites of a few alleles. Above the horizontal line is the mutant allele presenting the premature stop codon (highlighted in yellow). Note the lack of polymorphism in the mutant allele compared to the “wild type”.



	CT066	CT093	CT166	CT179	CT198	CT208	CT251	CT268
S. arcanum	0.46	0.30	0.29	0.13	0.73	0.24	0.25	0.20
S. habrochaites	0.53	0.35	0.40	0.27	0.74	0.03	0.44	0.32

Figure 20 AMOVA F_{ST} values per species and locus.

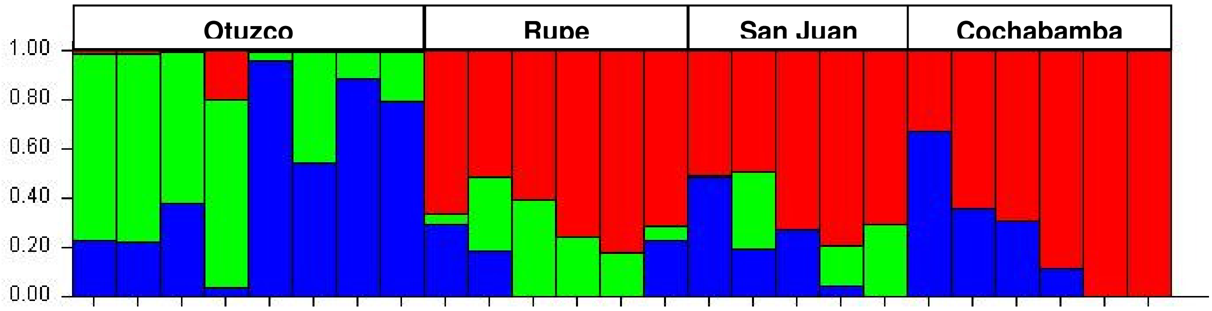


Figure 21 Bar plot for *S. arcanum* produced by STRUCTURE with K = 3, which obtained the maximum likelihood value.

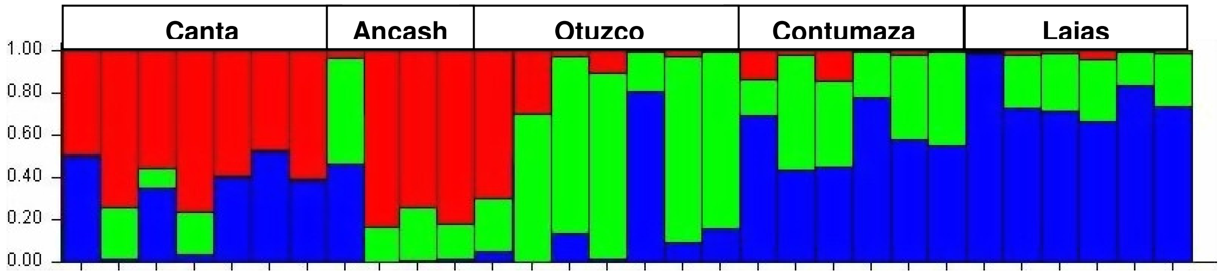


Figure 22 Bar plot for *S. habrochaites* produced by STRUCTURE with K = 3, which obtained the maximum likelihood value.

	Rupe	San Juan	Cochabamba
Otuzco	0.20	0.23	0.33
Rupe		0.02	0.24
San Juan			0.26

Table 6 Pairwise F_{ST} averaged across all loci for *S. arcanum*.

	Canta	Otuzco	Contumaza	Lajas
Ancash	0.14	0.25	0.25	0.33
Canta		0.29	0.25	0.26
Otuzco			0.15	0.25
Contumaza				0.14

Table 7 Pairwise F_{ST} averaged across all loci for *S. habrochaites*.

The mutant allele is found in all populations of *S. habrochaites* except Ancash and in only one population of *S. arcanum*, namely Otuzco. It is worth mentioning that Otuzco is the only collection site where both species were collected simultaneously since their sampling populations are sympatric at this location. Furthermore, the mutant allele is never present in homozygous state, which would be in agreement with the mutation's rendering the allele non-functional. This, however, would be in contradiction with the high frequency in which it is found (14-50% in each population, 35% species-wide).

Interestingly, locus CT198 presents the highest level of population structure as determined by AMOVA F_{ST} (Figure 20). This might indicate very low levels of gene flow between populations at this locus. This, in turn, might explain how the mutant allele (potentially deleterious) may have been preserved within populations due to smaller effective population sizes, making genetic drift stronger than purifying selection. However, average F_{ST} estimates of all within-species pairwise population comparisons, including and excluding the mutant alleles (data not shown), produced very low values for this locus. Still, the F_{ST} value of the comparison between wild type and mutant alleles was very high (0.78), clearly implying that mutant and wild type alleles are highly differentiated from each other. Additionally, locus CT198 presents positive Tajima's D values in every single population and a non-significant value for R_2 (data not shown). Noteworthy, haplotype networks for this locus show that mutant alleles are more closely related to *S. habrochaites* than to *S. arcanum*. Also, they show between one and three mutational steps between all versions of the allele, while wild type alleles show up to 17 mutational steps between them. Taken together, these results point to balancing selection and introgression of this allele from another species, but we need to take into account metapopulation dynamics before further assessing this locus.

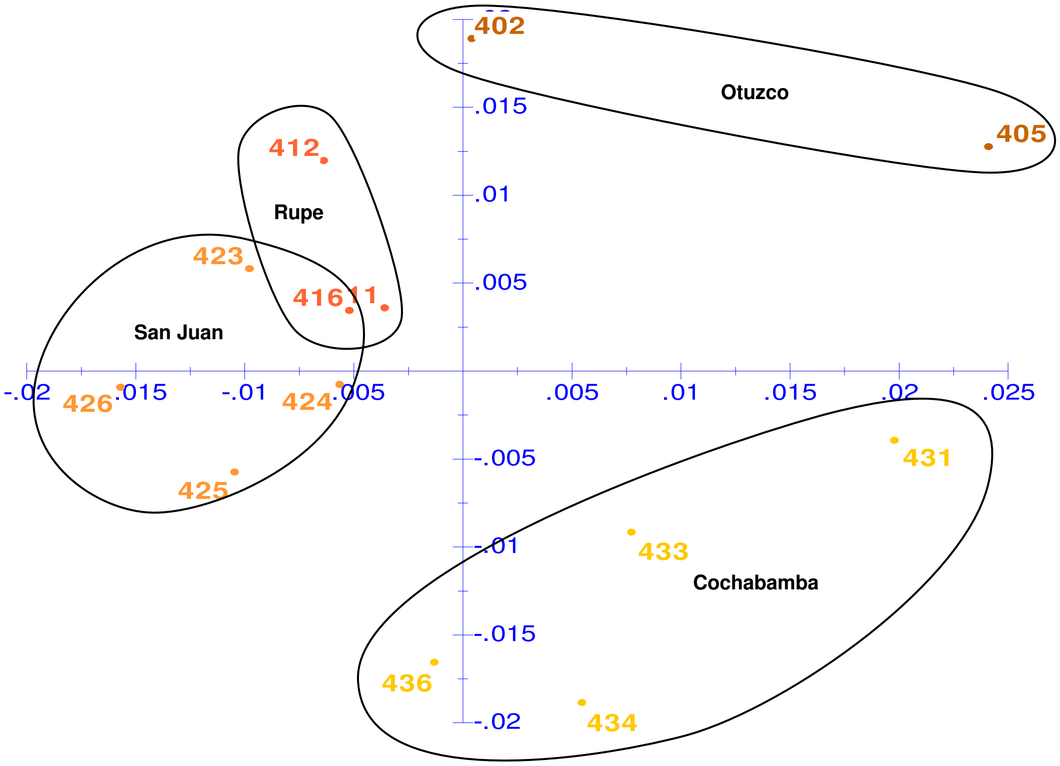


Figure 23 PCoA for *S. arcanum*. Colors indicate population of collection: Otuzco, Rupe, San Juan and Cochabamba. The Y- and X-axes explain 28% and 25% of the variability of the data, respectively.

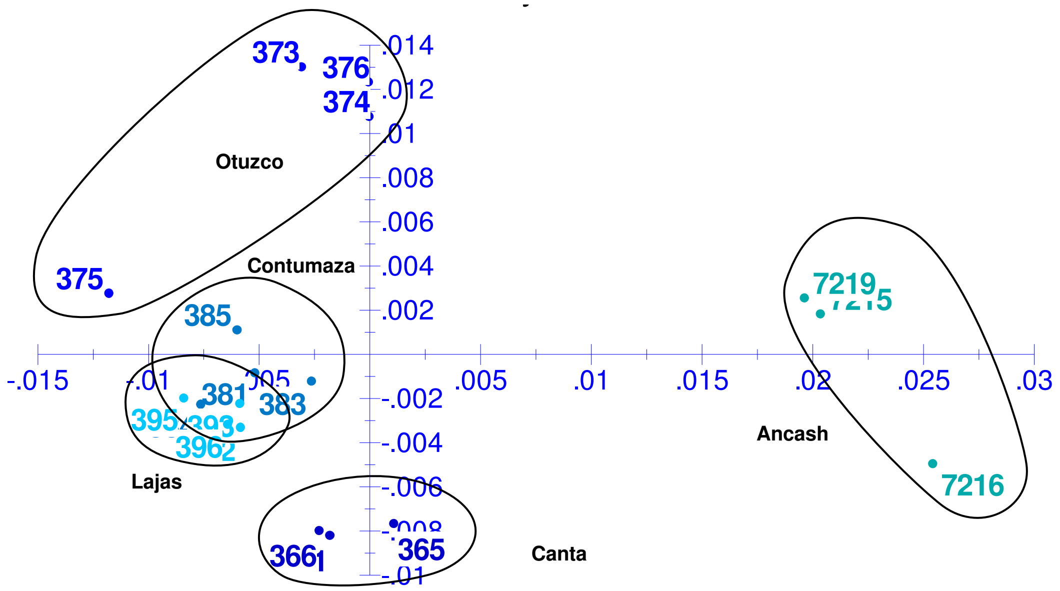


Figure 24 PCoA for *S. habrochaites*. Colors indicate population of collection: Canta, Otuzco, Contumaza, Lajas and Ancash. The Y- and X-axes explain 40% and 15% of the variability of the data, respectively.

2. Species expansion

R_2 statistic (Ramos-Onsins & Rozas 2002) is used to assess whether the species studied underwent a recent population size expansion. According to these results (Figure 18), *S. peruvianum* has undergone a recent population expansion. *S. chilense* and *S. habrochaites*, show no signature of this phenomenon. *S. arcanum* shows a significant value ($p \leq 0.05$), although not as evident as that of *S. peruvianum*. Therefore, one can argue that the expansion in *S. arcanum* was weaker or it was not as recent as that of *S. peruvianum*.

3. Population Structure in *S. arcanum* and *S. habrochaites*

Haplotype networks fail to show any pattern in the distribution of haplotypes among populations (data not shown). This indicates a large amount of recombination as well as the presence of shared polymorphism between populations and species.

Pairwise F_{ST} (Hudson & Slatkin 1992), shows population differentiation within species (Table 6, Table 7). Most remarkably, the Rupe and San Juan populations of *S. arcanum* show very little differentiation, indicating these populations could be considered as one interbreeding unit. This may be expected since these two populations are only separated by a distance of 28 km, while all other populations sampled in this species are separated by distances ranging between 68 km to 161 km. We also evaluated an isolation by distance pattern by estimating the correlation of pairwise F_{ST} to geographic distance (data not shown). The correlation was found significant for *S. arcanum* with a correlation coefficient of $r^2=0.36$. This was not the case for *S. habrochaites*. The caveat to this result is that one would need a much more extensive sampling in order to be able to draw conclusions about isolation by distance in this species.

AMOVA F_{ST} (Excoffier et al. 1992), shows significant ($p \leq 0.01$) levels of population differentiation (Figure 20) for *S. arcanum* at all loci. Similarly, *S.*

habrochaites shows significant values for all except one locus (CT208). Taken together, these results suggest that (as for diversity) population differentiation is largely locus-specific. Additionally, it is clear that, in general, *S. arcanum* has lower levels of population differentiation than *S. habrochaites* (except for one locus, namely CT208).

While our AMOVA F_{ST} values are in general correlated with the average of pairwise population comparisons per locus (data not shown), this is not the case for locus CT198. In this case, AMOVA F_{ST} shows the highest values for both species when compared to all other loci. However, F_{ST} averages for this locus fall into the distribution of all loci.

STRUCTURE analysis for each species shows that the maximum likelihood is reached for K = 3 for both species. This is surprising as we have sampled 4 to 5 different populations. Thus, a physically distinct population in the field is not necessarily a panmicticly reproductive unit in genetic terms. However, previous results from our lab (Arunyawat et al. 2007) also showed moderate amounts of population structure in *S. peruvianum* and *S. chilense*. This means that our results are in accordance with previous findings on metapopulation structure for wild tomatoes. For this value of K the bar plots show very different results for each species.

In the case of *S. arcanum* (Figure 21), the Otuzco cluster is composed mainly of alleles coming from two different putative populations. The Cochabamba cluster is made up exclusively of two populations, both present in Otuzco (one of them only present in one individual). Rupe and San Juan are a mixture of all three putative populations, where the one present only in one individual in Otuzco dominates all other populations. This result confirms what we found through F_{ST}: that Rupe and San Juan are genetically very similar to each other and may be considered as one population. Also, it is clear that Otuzco is the most dissimilar population when compared to the other three.

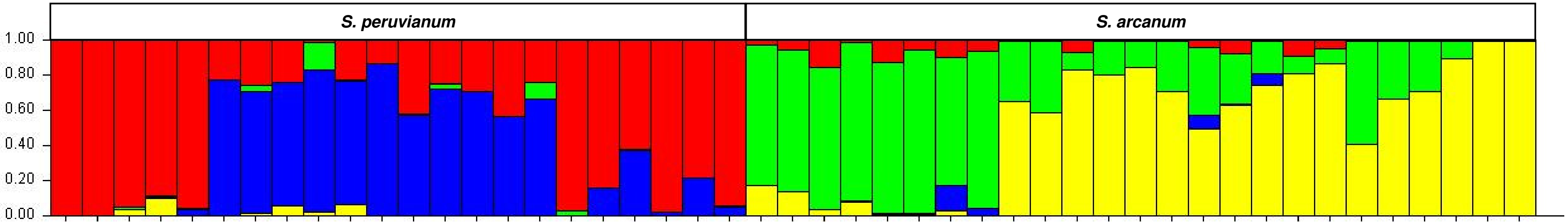


Figure 25 Bar plot for *S. arcanum* and *S. peruvianum* produced by STRUCTURE with K = 4.

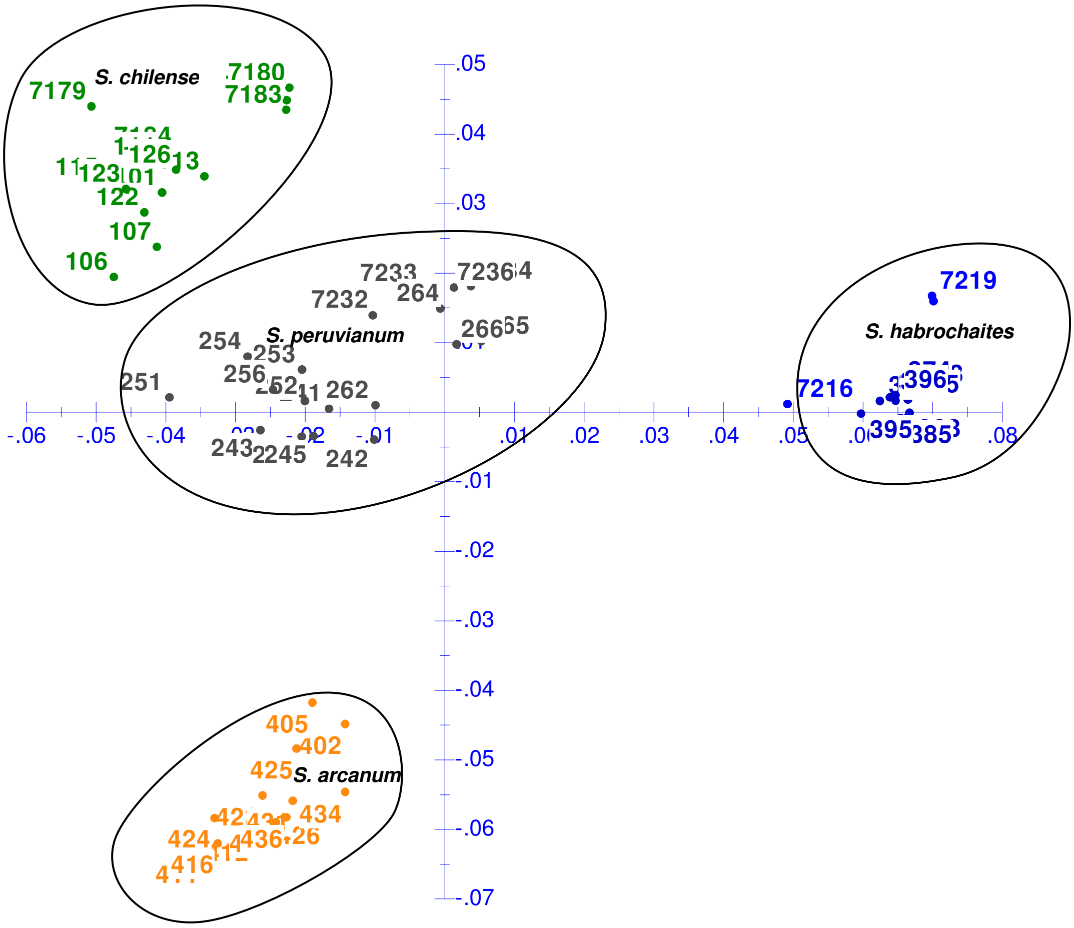


Figure 26 PCoA using all polymorphic sites of *S. peruvianum* (gray), *S. chilense* (green), *S. habrochaites* (blue) and *S. arcanum* (orange). The Y- and X- axes explain 34% and 19% of the variability of the data, respectively.

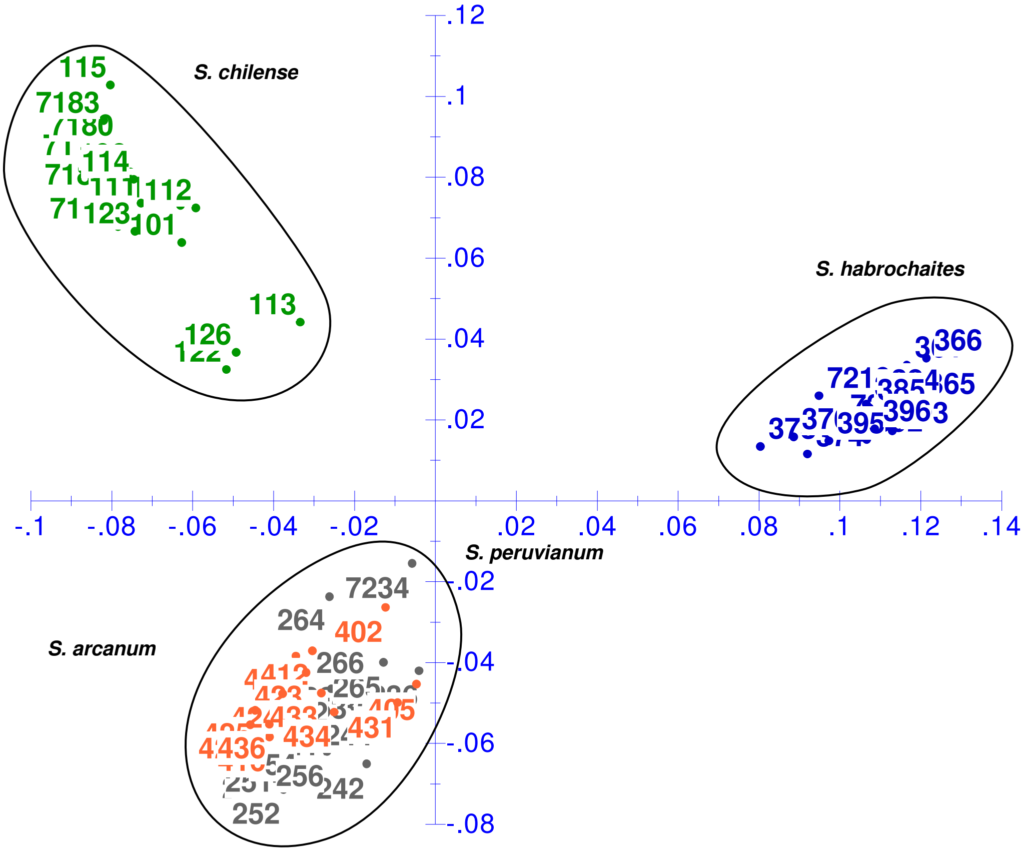


Figure 27 PCoA using non-synonymous sites of *S. peruvianum* (gray), *S. chilense* (green), *S. habrochaites* (blue) and *S. arcanum* (orange). The Y- and X- axes explain 25% and 14% of the variability of the data, respectively.

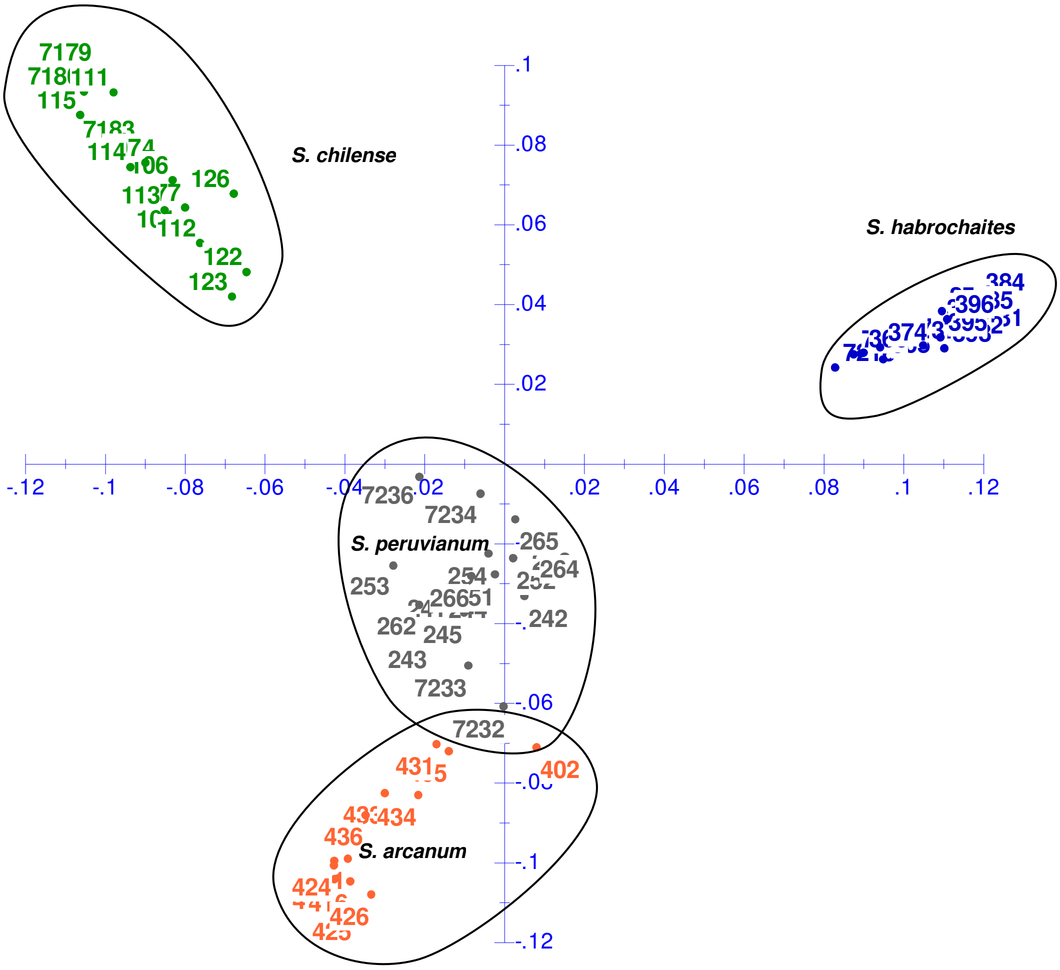


Figure 28 PCoA using synonymous sites of *S. peruvianum* (gray), *S. chilense* (green), *S. habrochaites* (blue) and *S. arcanum* (orange). The Y- and X- axes explain 22% and 15% of the variability of the data, respectively.

For *S. habrochaites*, STRUCTURE results (Figure 22) show that the Ancash cluster is the least similar to all others (composed almost entirely of a single putative population). All other clusters, except Canta, are made up of the same 2 putative populations. Thus, with the exception of Ancash and Canta (which share the same dominant putative population), a North-South gradient is observed, where the northernmost Lajas is almost completely made up of a single putative population, while the proportion of this is gradually replaced by the predominant putative population of Otuzco (southern-most in this group). This phenomenon might be explained by higher gene flow between the Otuzco, Cochabamba and Lajas populations, since they are located closer together (between 65 to 150 km away from each other) than Ancash and Canta (190 to 590 km).

Using PCoA, we can visually assess the degree of similarity between individuals and compare it with the populations from which they are collected. PCoA confirms the results from STRUCTURE. In the case of *S. arcanum* (Figure 23), it clusters together Rupe and San Juan populations, which we have shown to be very similar to each other (see above). A North-South pattern can be observed on the Y-axis. As for *S. habrochaites* (Figure 24), the results are more complex. The Ancash population is clearly separated from all others, thus confirming that this population is most differentiated from all others. Otuzco, Contumaza and Lajas, are distributed along the Y-axis in a gradient that matches the one shown by STRUCTURE between Lajas and Otuzco.

4. Divergence between Species

PCoA results vary greatly depending on whether all sites or a subset of sites (synonymous, non-synonymous or silent) are used. Using all polymorphic sites (Figure 26), PCoA shows a very clear differentiation between *S. peruvianum* and *S. arcanum*. What's more, the closest species to *S. peruvianum* is *S. chilense*, according to this analysis. However, by performing the same analysis using only synonymous sites (Figure 28) *S. peruvianum* clusters closely with *S. arcanum*. Furthermore, by using non-synonymous sites to perform the analysis (Figure 27) PCoA cannot discriminate between *S. peruvianum* and *S. arcanum*.

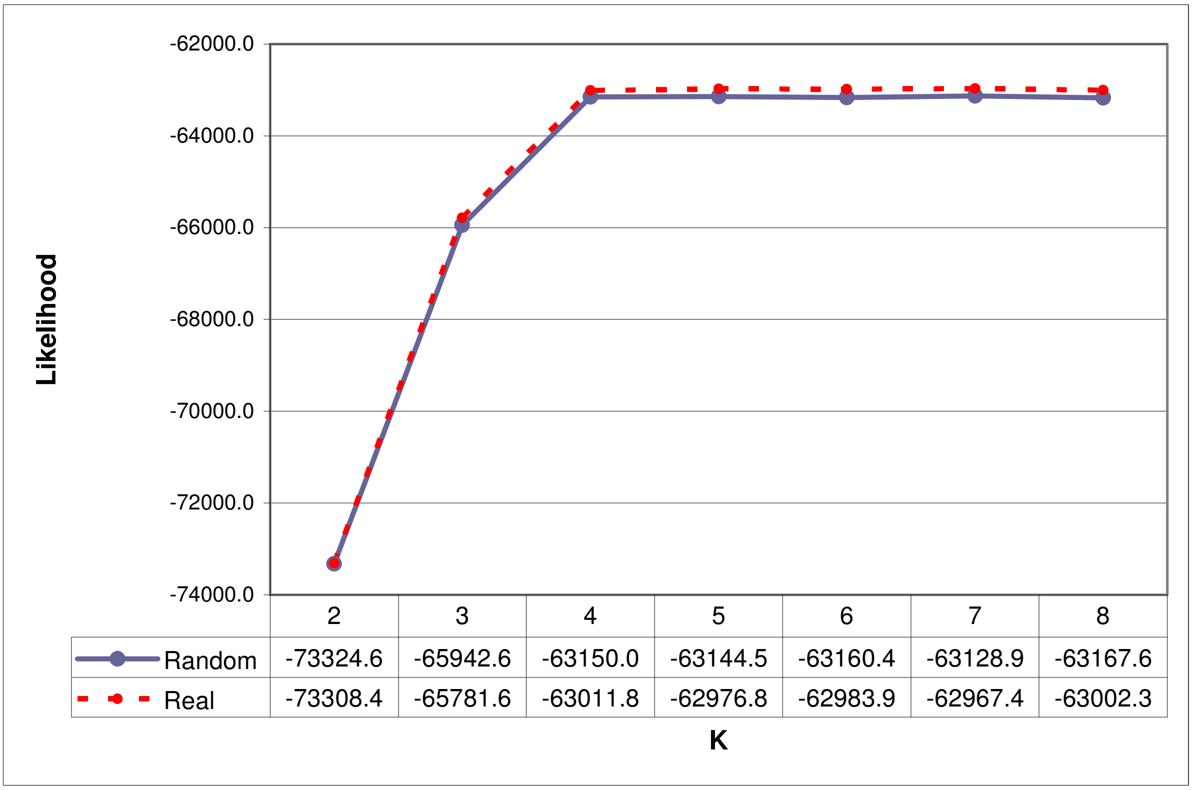
This clearly indicates that non-coding regions within the sequenced loci are most similar between *S. chilense* and *S. peruvianum*, but quite divergent between *S. peruvianum* and *S. arcanum*. Arguably, one could assume that *S. peruvianum* and *S. arcanum* have evolved under the same selective pressures. This would mean that the difference between the plots obtained by using synonymous and non-synonymous sites might be explained as selection acting directly on non-synonymous sites and indirectly on synonymous sites linked to the former.

STRUCTURE results for *S. peruvianum* and *S. arcanum* combined (Figure 25) show a maximum likelihood value at K=4. This means that the alleles present in *S. peruvianum* and *S. arcanum* are best explained as coming from four putative populations. These results (obtained using all polymorphic sites) confirm the clear differentiation between these two species since *S. peruvianum* is mostly comprised of two population which are different from those that mainly compose *S. arcanum* (Figure 25, red and blue in *S. peruvianum*, and green and yellow in *S. arcanum*). Furthermore, if one compares this graphic to that based exclusively on populations of *S. arcanum* (Figure 21), it is clear that the two putative populations present in Figure 25 refer to the two clusters visible in the former (*i.e.* the cluster for Otuzco and the cluster for Rupe, San Juan and Cochabamba).

We also performed the same analysis using the whole dataset (*i.e.* all four species: *S. peruvianum*, *S. chilense*, *S. habrochaites*, and *S. arcanum*). Likelihood values were very similar for K≥4 (Figure 29). However, all bar-plots (Figure 30), regardless of the value of K (as long as K≥4) show 4 major putative populations composing the dataset, which exactly match the 4 species that compose the dataset. Furthermore, repeating the analysis using the randomized dataset (see Materials and Methods for details), reproduces the same exact results. Thus, we confirm that having the phase of sequence data, *i.e.* the labeling of alleles as “a” or “b”, does not have any effect on the results of STRUCTURE. These results also confirm the divergence between *S. peruvianum* and *S. arcanum*, which appear as two distinct species. However note that for K=4 or K=8, some hints for introgression (or ancestral polymorphism) can be found between the two species, *e.g.* in Figure 30 A: yellow bars characteristic of *S. peruvianum* are found in *S. arcanum*.

We get very different results if we look into π_{between} (Figure 31). In this case, we get overall lower values using all sites and the highest values using synonymous sites. Still, it is worth mentioning that the *S. arcanum*-*S. peruvianum* comparison has the lowest values in all cases except when using synonymous sites to perform the analysis.

The contrasting difference between π_{between} results and those obtained using PCoA and STRUCTURE, might lie in the fact that π_{between} is based on pairwise comparison, meaning it takes into account the frequency in which different alleles appear in each population. This can be interpreted as synonymous sites having a higher amount of shared polymorphism but in different frequencies within each *S. peruvianum* and *S. arcanum*. This is an important point, since gene flow would not only produce shared polymorphism but would also homogenize its frequency.



K	2	3	4	5	6	7	8
Random	-73324.6	-65942.6	-63150.0	-63144.5	-63160.4	-63128.9	-63167.6
Real	-73308.4	-65781.6	-63011.8	-62976.8	-62983.9	-62967.4	-63002.3

Figure 29 Likelihood values for the 4 species dataset (results in Figure 30). The likelihood of each K is computed for the real data and the randomized dataset to explore the effect of having unphased data.

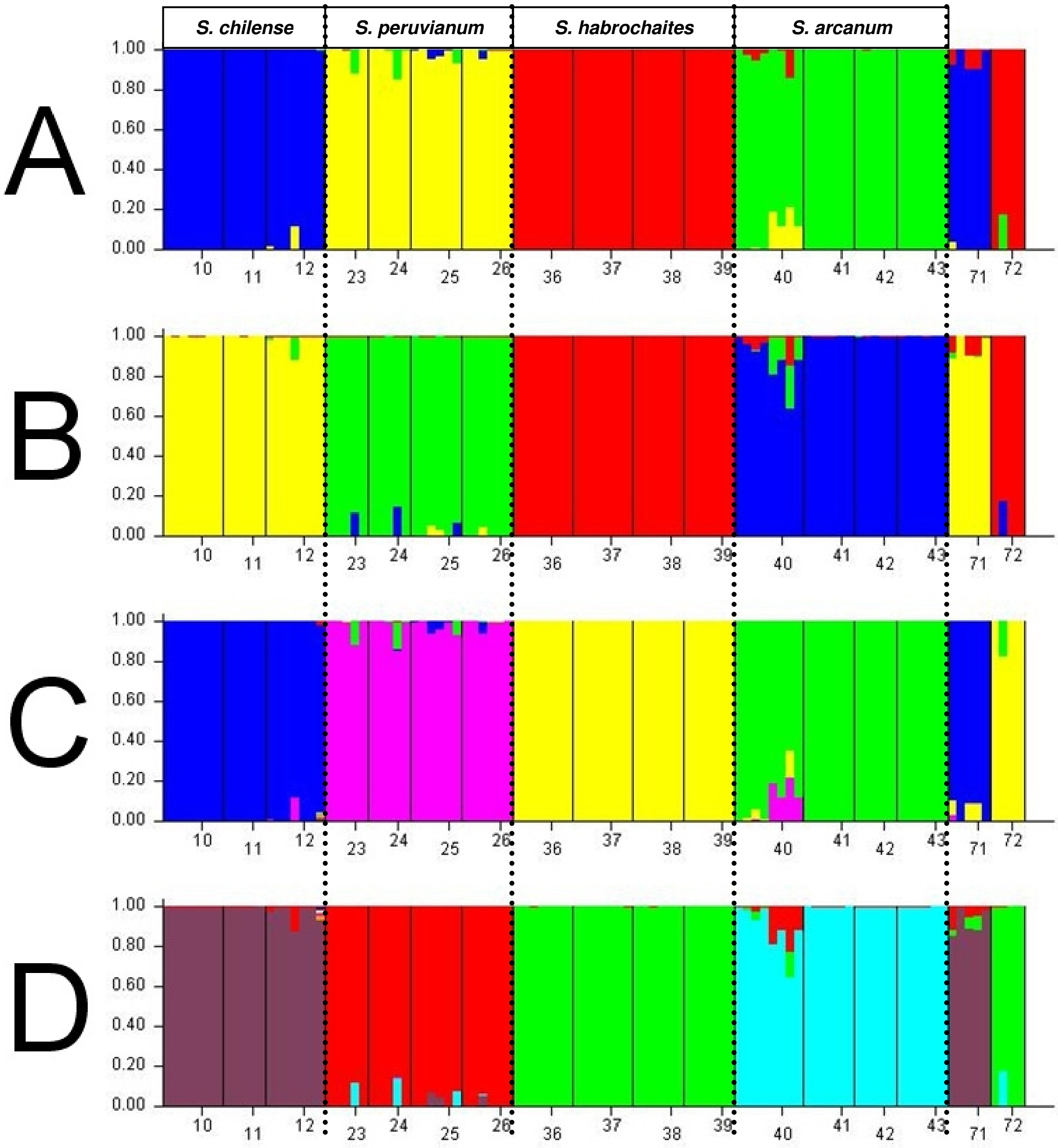
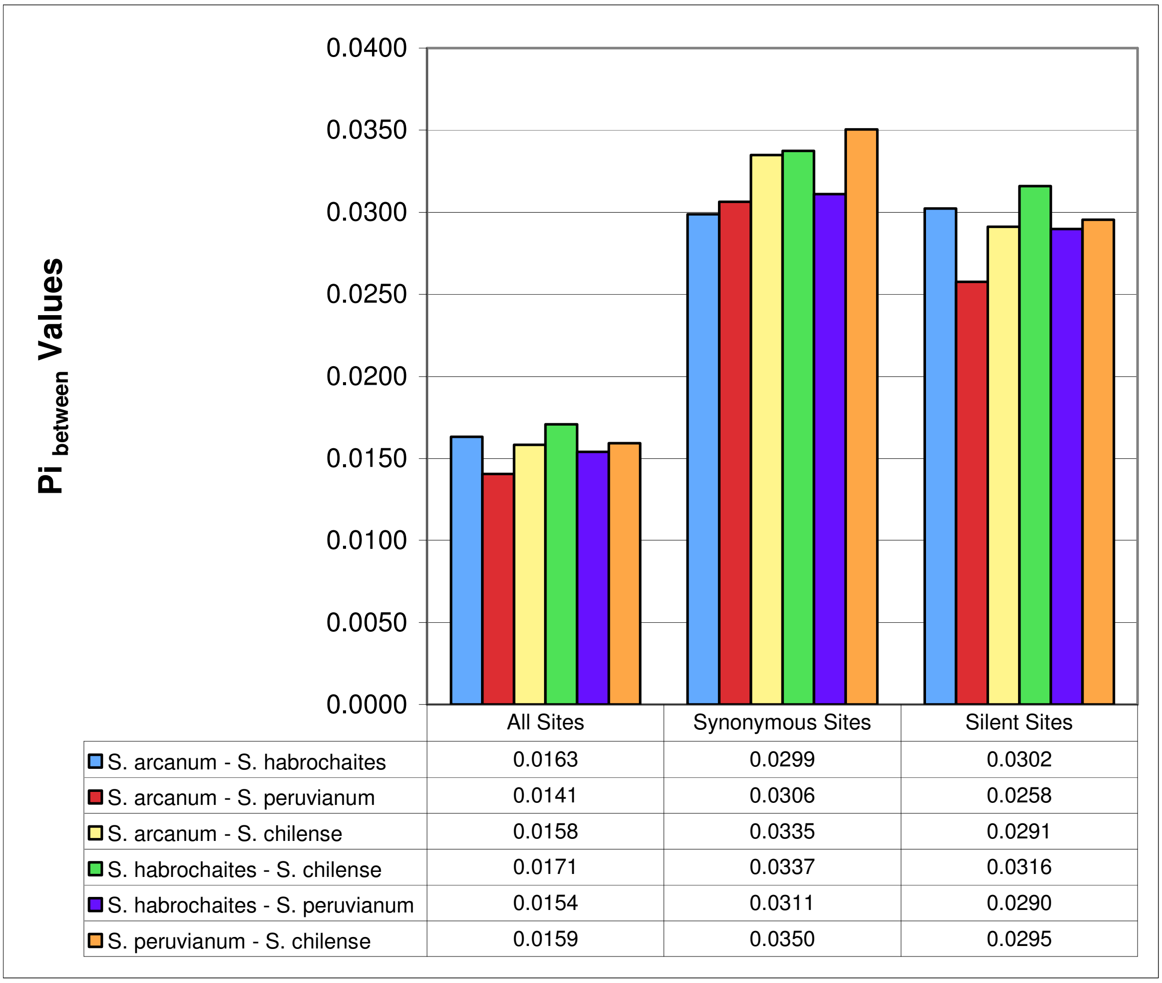


Figure 30 Barplots of STRUCTURE analyses performed on all 4 species. Numbers denote species: 10-12 and 71, *S. chilense*; 23-26, *S. peruvianum*; 36-39 and 72, *S. habrochaites*; and 40-43, *S. arcanum*. Note that although STRUCTURE software orders populations in numerical order (*i.e.* populations 71 and 72 are lumped at the right end of the graphic), they still clearly show the same features of the species to which they belong. Letters denote dataset and value of K: A, real data analyzed with K=4; B, randomized data (see text) analyzed with K=4; C, real data analyzed with K=8; and D, randomized data analyzed with K=8. Notice that all graphs have a single interpretation, *i.e.* that each species cluster is formed of a single putative ancestral population, with very few alleles coming from a different putative ancestral population.



	All Sites	Synonymous Sites	Silent Sites
🟦 S. arcanum - S. habrochaites	0.0163	0.0299	0.0302
🟥 S. arcanum - S. peruvianum	0.0141	0.0306	0.0258
🟨 S. arcanum - S. chilense	0.0158	0.0335	0.0291
🟩 S. habrochaites - S. chilense	0.0171	0.0337	0.0316
🟪 S. habrochaites - S. peruvianum	0.0154	0.0311	0.0290
🟧 S. peruvianum - S. chilense	0.0159	0.0350	0.0295

Figure 31 $\pi_{between}$ values per site calculated for all possible pairwise comparisons among the four species analyzed: *S. peruvianum*, *S. chilense*, *S. habrochaites* and *S. arcanum*.

Discussion

In this study, we have focused our observations on diversity, demography and structure. In the first, theoretical, project we used coalescent simulations to assess the effect of sampling on summary statistics. For the other two, empirical, projects we used sequenced data of eight EST-based reference loci from four wild tomato species (*S. peruvianum*, *S. chilense*, *S. habrochaites* and *S. arcanum*). For the second project, we used the distribution of fitness effects to estimate the amount of selection on these loci. For the third project, we have estimated diversity levels, performed neutrality tests and examined levels of population structure. Furthermore, using population genetics analysis we assess the current taxonomical treatment of *S. peruvianum*. Thus, we have gained insight into the evolution of recently divergent species taking into account metapopulation dynamics and its interplay with a species' demography and their influence on the speciation processes.

In the first project, using coalescent simulations, we studied the impact of three different sampling schemes on patterns of neutral diversity in structured populations. The three sampling schemes were: a scattered or species-wide sample (each sequence coming from a different deme), a local sample (all sequences coming from the same deme) and a pooled sample (equal number of sequences coming from different demes). Specifically, we evaluated two summary statistics based on the site frequency spectrum (Tajima's D and Fu and Li's D) as a function of migration rate, demographic history of the entire metapopulation and the sampling scheme. Using simulations implementing both finite-island and two-dimensional stepping-stone spatial structure models, we demonstrate strong effects of the sampling scheme on Tajima's D and Fu and Li's D statistics, particularly under species-wide expansions. Local samples (and to a lesser extent, pooled samples) are influenced by local, rapid coalescence events in the underlying coalescent process and hence show lower proportions of singletons. Under species-wide expansion scenarios, these effects of spatial sampling may persist up to very high levels of gene flow ($Nm >$ 25), implying that local samples cannot be regarded as being drawn from a panmictic population. This suggests that validating the assumption of panmixia is crucial if robust demographic inferences are to be made from local or pooled samples.

For the second project, we investigated how selection acts in the aforementioned species of wild tomatoes using sequence data from eight housekeeping genes. Our analysis quantified the number of adaptive and deleterious mutations, and the distribution of fitness effects of new mutations (its mean and variance) taking into account the demography of the species. We found no evidence for adaptive mutations and very strong purifying selection in coding regions of the four species. More interestingly, the four species exhibit different strength of purifying selection in non-coding regions (introns). This suggests that closely related species with similar genetic background, but occurring in different environments differ in the mean and variance of deleterious fitness effects. We also showed that variable selection in introns is also found among populations of a given species, suggesting local difference in the strength of purifying selection. Taking into account the results from the first project, we also highlighted the utility of analyzing pooled samples and local samples from a metapopulation in order to measure selection and the distribution of fitness effects.

Finally, we estimated nucleotide diversity and population structure in *S. habrochaites* and *S. arcanum* and compared these results to those of *S. peruvianum* and *S. chilense* (Arunyawat et al. 2007) for the third project. First, we found that *S. arcanum* and *S. habrochaites* present lower diversity levels than *S. peruvianum* and *S. chilense*. Our neutrality tests have not revealed any particular pattern, leading us to conclude that the loci sequenced for the present study have not evolved under strong positive selection, although they show a distinctive pattern of purifying selection (second project). We also tested the demography of all four species and found a strong expansion after a bottleneck in the recent past for *S. peruvianum* and a similar statistically significant pattern for *S. arcanum*, even though the signal seemed weaker in this case. Additionally, we found moderate levels of population sub-structure in these species, similar to previous results found in *S. peruvianum* and *S. chilense*. Still, regardless of the levels of population structure, we found at least two (Rupe and San Juan from *S*. *arcanum*) populations collected in the field that could actually be considered as a single deme.

We also expanded these population structure analyses to gain insight into the phylogenetic relations between the four species in order to contribute to the taxonomical treatment of the *Solanum* section *Lycopersicon* from a population genetics perspective. Thus, we found a clear differentiation between *S. arcanum* and *S. peruvianum* based on all polymorphic sites. However, this pattern disappears if we perform similar analyses using only non-synonymous sites. This is an important point since different inferences about speciation may be concluded depending on whether coding or non-coding regions of the genome are used for such analyses.

In this final chapter, I will discuss the demography of *S*. *arcanum* and *S*. *peruvianum*, the metapopulation structure of *S*. *arcanum* and *S*. *habrochaites*, the phylogenetic relations between *S*. *arcanum* and *S*. *peruvianum* and finally the discovery of a mutant allele in intermediate frequency at one locus in *S*. *habrochaites* and *S*. *arcanum*.

1. Population structure of *S. arcanum* and *S. habrochaites*

Structured populations are difficult to analyze. First, there is the problem of sampling from a structured population, which can bias estimation of population parameters (Stadler et al. 2009). As the results from the first project show, different sampling schemes from a structured population even under high levels of gene flow can significantly affect summary statistics estimation. This has to do with the fact that sampling within demes results in an overrepresentation of the scattering phase of the coalescent (*i.e.* the polymorphism that occurs within each deme). Second, population structure can have a direct effect on the effective population size as well as divergence levels between species. Structured populations present private polymorphism in each deme following two processes. First, due to random drift and low level of gene flow, some neutral mutations are private to demes reflecting the scattering phase of metapopulation coalescence (Pannell & Charlesworth 1999; Wakeley & Aliacar 2001). Second, due to small deme effective population sizes (making genetic drift stronger), purifying selection cannot eliminate weakly deleterious mutations effectively, and thus such mutations are found private to a deme at low frequency (Fay et al. 2001; Whitlock 2003). This phenomenon affects divergence, since it creates a higher species-wide diversity because of the excess of low-frequency polymorphism private to each deme. These effects have to be taken into account when interpreting divergence between species with known metapopulation structure.

There are several ways of estimating population structure. The most extensively used is F_{ST}, a fixation index that measures the genetic differentiation among populations and is often expressed as the proportion of genetic diversity due to allele frequency differences among populations. This method has been criticized (Charlesworth 1998; Neigel 2002) but continues to be used as the standard measure of genetic differentiation and hence to asses population structure and levels of gene flow. An alternative strategy is to use π_{between} as a measure of divergence between populations. It is the average number of pairwise differences between sequences belonging to different populations. Hence, it estimates the differentiation between populations and thus, assesses the levels of population structure, which has the advantage of not being sensible to high levels of diversity and directly comparing sequences, instead of allele frequencies. In our study, we use both as well as an analysis of molecular variance (AMOVA) analog of F_{ST}, plus PCoA and the software STRUCTURE (see Materials and Methods section for details). PCoA allows one to identify key components of population differentiation without assuming an evolutionary model (McVean 2009). Our PCoA results closely resemble those generated by STRUCTURE software simulations, which in turn also agree with our estimates of pairwise F_{ST}.

As reported for *S. peruvianum* and *S. chilense* (Arunyawat et al. 2007), we found moderate levels of population structure for *S. habrochaites* and *S. arcanum*. We have also found that populations sampled in the field do not necessarily correspond to isolated interbreeding units, even when the species studied shows moderate levels of population structure. This is the case for two populations of *S. arcanum* (Rupe and San Juan) that show a pairwise F_{ST} value close to zero, similar patterns of allelic frequencies according to STRUCTURE and cluster together in PCoA. A similar, although more complex, phenomenon occurs in *S. habrochaites* where three populations (Otuzco, Contumaza and Lajas) show a gradient in the proportion of ancestry of three putative ancestral populations as depicted by STRUCTURE analysis. Lajas and Contumaza overlap in PCoA with Otuzco closest to them. Furthermore, the distribution of their allele frequencies as a gradient (in STRUCTURE) and their positioning along the Y-axis in PCoA match their geographical distribution along the North-South axis. This is also observed in *S. habrochaites* where populations are distributed along the Y-axis following their North-South geographical distribution. This is not surprising as it has been analytically shown that the spatial arrangement of populations on the first two principal components mimics the structure of the migration matrix (Novembre & Stephens 2008), which can be considered an approximation to the geographic distribution of the samples (although differences in sample sizes produce a distortion of the migration-space). Furthermore, a very similar method (Principal Components Analysis) has been previously used to find a close correspondence between genetic and geographic distances (Novembre et al. 2008).

Still, PCoA results have to be interpreted with care, as the structure depicted by it can be the result of several different demographic processes. For this reason, we looked into the possibility that our data would fit an isolation by distance model by estimating the correlation of pairwise F_{ST} to geographic distance (data not shown). The correlation was found significant ($p<<0.01$) for *S. arcanum* with a correlation coefficient of $R^2=0.36$. This was not the case for *S. habrochaites*. Our sampling, however, may not be extensive enough to test for such model. Furthermore, the distances between populations are not uniform between populations for *S. habrochaites* (Ancash and Canta are much further away from all other populations). This, in addition to the small amount of populations sampled, may be responsible for the lack of correlation between population differentiation and distance in this species. However, one can speculate about the spatial structure of *S. habrochaites*. An island model is not expected to fit our data. Under this model, we would expect equal rates of gene flow between all analyzed populations regardless of the geographic distance between them. Under such scenario, we would expect STRUCTURE analysis to show equal proportions of the putative ancestral populations that are present in the other populations of the species. This pattern is not observed in our data. A better fitting model would be the stepping stone model. The main difference between this and that of isolation by distance lies in the property that the stepping stone model restricts gene flow to shorter distances which would translate into very strong differentiation between our sampling populations on opposite edges of the sampling range. In our case, STRUCTURE found that such pairs of populations still share common ancestry from at least one of the putative ancestral populations modeled by the software.

Another interesting question is whether our data fits information on weakly defined morphotypes in *S. arcanum* as defined in the most recent taxonomical treatment of these species (Peralta et al. 2008). These morphotypes are similar to the races of *S. peruvianum* reported by Rick (1986) according to his inter-fertility experiments (Figure 6). The recent taxonomical treatment that we follow here, re-groups these races based on morphological data. Cleary, genetic and morphologic data are not easy to integrate without some discrepancies which researchers have tried to reconcile in this latest new taxonomical treatment. According to this treatment (Peralta et al. 2008), *S. arcanum* is an extremely variable species, comprising four weakly defined morphotypes with discrete geographic ranges. The complex overlapping variability, especially in leaf morphology, prevents the recognition of these as formal taxa. These morphotypes are:

1) Marañón, which grows in the Río Marañón Valley, includes the Chamaya-Cuvita and the Marañón races that Rick (1986) recognized as closely related based on their inter-fertility in experimental crosses.
2) Humifusum, growing in Pacific drainages.
3) Chotano, growing in the Río Chota Valley near Yamaluc in the Department of Cajamarca. According to the authors of the current taxonomical treatment, the 'humifusum' and 'Chotano' races also appear to be closely related based on data from inter-fertility experiments (Rick

1986) and molecular analyses (Peralta & Spooner 2001), and differ mainly in pubescence.

4) Lomas, growing at the Lomas of Cerro Campana and Virú. These populations are incredibly variable from year to year; specimens collected in El Niño years have very large leaves, while those collected in drier seasons have smaller, more pubescent leaves with fewer leaflets with less lobed margins. The ‘Lomas’ populations are also quite variable in other characters, such as the inflorescence branching pattern.

Using the geographic ranges given in the current taxonomical treatment (Peralta et al. 2008) as reference to compare the location of our sampled populations, we can assume that most populations of *S. arcanum* (except for Cochabamba) belong to the Humifusum morphotype, while Cochabamba most likely belongs to the Chotano morphotype. As expected, PCoA plots the Cochabamba population most distantly from the other three populations of the species. STRUCTURE analysis shows that one of the putative ancestral populations is completely missing in Cochabamba (Chotano), while it is present in small proportion in both Rupe and San Juan, and in an even larger proportion in Otuzco. This result is in agreement with the existence of incomplete reproductive barriers between this races/morphotypes, since apparently a whole gene pool has been prevented from introgressing into the Cochabamba population.

2. Species wide expansion in *S. peruvianum* and *S. arcanum*

As selection and demography can produce similar patterns in DNA polymorphism, neutrality tests showing deviations from the standard neutral model cannot directly be interpreted in terms of selection acting on a locus. Therefore, it is important to know the demography of a sample before drawing conclusions on whether or not it evolved under selection. Here, we have gained some insight into the demography of four species of wild tomatoes. Using a conservative approach (the use of “pooled samples” –see Materials and Methods section–), it is clear that demography has played an important role in the evolution of this group of species. Due to the conservative nature of our estimates, we cannot discard the possibility that *S. chilense* and *S. habrochaite*s have both undergone recent population expansion after a bottleneck. Furthermore, in the first project we re-examined the demography estimated in a previous study (Arunyawat et al. 2007). Comparing Tajima's D estimates averaged across populations to those obtained from pooled samples, to results from simulations, we were able to determine expansion factors for both *S. peruvianum* (≥40-fold) and *S. chilense* (≤4-fold). These results are hard to reconcile with those obtained from the distribution of fitness effects which estimate expansion factors as 10-fold for *S. peruvianum* with the expansion occurring recently, and 10-fold in *S. chilense* with a much older expansion. A population with an expansion that took place long ago might have reached equilibrium and thus, might not show traces of the expansion (*i.e.*, no negative Tajima's D). The discrepancy between the different methods for estimating expansion suggests that the method by Keightley & Eyre-Walker (2007) is not reliable for estimating precisely past demographic events. It is also clear from our theoretical work using simulations (on the first project) that given Tajima's D values are not sufficient to allow robust estimation of the expansion factor. As already mentioned (see above), our findings using the R_2 test fail to detect expansion at *S. chilense* and *S. habrochaites*, though data from previous works (Arunyawat et al. 2007; Stadler et al. 2009) as well as our results in the second project indicate that these species actually underwent an expansion. Still, this test at least detects that the other two species analyzed (*S. arcanum* and *S. peruvianum*) have undergone a recent expansion after a bottleneck, even though this expansion might be underestimated due to our use of pooled samples.

In conclusion, as reported in the second project of this thesis (Stadler et al. 2009), we find that pooling together samples from different populations produces a much lower Tajima's D value, than the average of the same populations. Negative Tajima's D values for the pooled sample show an artificial excess of low-frequency polymorphism due to private alleles present at low frequency in each of the populations, which would point to lower levels of gene flow among populations than would be expected under weak population structure. This would also have a confounding effect on the detection of purifying selection, since the latter would produce a similar signature. It is always the case that trying to distinguish between selection and demography poses a problem that can only be solved by using multilocus data to establish the demography of a species prior to trying to detect selection. This is one of the clear advantages of Eyre-Walker and Keightley's distribution of fitness effects estimation, which uses multilocus data to model a population expansion (Eyre-Walker et al. 2006; Keightley & Eyre-Walker 2007; Eyre-Walker & Keightley 2009).

Taken together, these findings highlights the importance of adequately planning the sampling for a study, according to the aims of the study and taking into account the effects the sampling might have on the population parameter estimation, before drawing conclusions from the results. Therefore, in order to obtain an appropriate estimate of the demography and avoid the effect of pooling populations to obtain a species-wide sample on summary statistics, it would make sense to collect only an allele per population from as many populations as possible from the distribution range of the species. This, however, is often not possible, since sample collection is a time-consuming activity requiring a certain level of expertise in the identification of a species, and extensive time traveling to find natural populations of the species. In the case of wild tomatoes, it is possible to use accessions from collections maintained at research facilities (*e.g.* TGRC University of Davis, USA).

3. Relation between *S. arcanum* and *S. peruvianum*

While the general concept of species is hard to define, several different approaches to species identification have been elevated to the level of species concept. From the point of view of population genetics, reproductive isolation (encompassing divergence, population differentiation and gene flow) is a key element in the definition of a species (following the so-called biological concept of species). However, for recently diverged species, such as wild tomatoes, where reproductive barriers are still incomplete, it may prove hard to apply such a concept. Therefore, systematists rely more on the morphological concept of species, in an attempt to grasp the underlying phylogenetic relations among such species. The irony lies in the fact that divergence population genetics provides a framework for the study of the speciation of such species which ultimately would help solve the phylogenetic relations between them. With this motivation in mind, I try to contribute to the taxonomic treatment of this group of species using a population genetics approach.

The latest taxonomic treatment split *S. peruvianum sensu lato* in four species, namely *S. arcanum*, *S. corneliomuelleri*, *S. huaylasense* and *S. peruvianum* sensu stricto (Peralta et al. 2008; Peralta et al. 2006; Peralta & Spooner 2005; Peralta et al. 2005; Peralta & Spooner 2001; Spooner et al. 2005). Therefore, a key question arising from our dataset is whether *S. arcanum* and *S. peruvianum* are more closely related to each other than to *S. chilense* and *S. habrochaites*, which would suggest that they actually constitute a single species. To this end, we analyze levels of divergence and population differentiation between these species using a suite of different methods.

We analyzed all populations of the four species together. This analysis would fit a hierarchical island model (Excoffier et al. 2009) in which demes exchange more migrants within species than between species. Since Structure identifies groups of individuals corresponding to the uppermost hierarchical level, we would expect to estimate *K*= 4 clusters by Maximum Likelihood. Analysis of the four species together using STRUCTURE software showed almost identical likelihood values for $K \geq 4$. At first glance, this may sound confusing. However, it has been reported (Pritchard et al. 2007; Evanno et al. 2005) that in most cases, once the real *K* is reached, its likelihood at larger values of *K* either reaches a plateau or continues increasing slightly. According to Pritchard et al. (on the accompanying documentation of the software), their estimation of *K* generally works well in datasets with a small number of discrete populations (between two and four). However, deviations from the STRUCTURE model such as isolation by distance or inbreeding may explain that the value of the model choice criterion continues to increase with increasing *K*. Therefore, we can pick *K=4* as the real number of populations present in the data. Additionally, results from this analysis showed some degree of gene flow between species, since each of them presented a small percentage of the other species putative ancestral population gene pool in their own.

PCoA showed very different pictures of the relations between the *S. peruvianum* and *S. arcanum* when using different subsets of the data. It has been shown (McVean 2009) that for SNP data the projection of samples onto the principal component axes can be obtained directly from considering the average coalescent times between pairs of genomes. Thus, PCoA projections can be interpreted in terms of underlying processes, including migration, geographical isolation, and admixture. Furthermore, the same study demonstrates a link between PCA and Wright's F_{ST}. Therefore, the differences in the distance between the relative positions of *S. arcanum* and *S. peruvianum* can be interpreted as differences in the coalescent time between coding and non-coding regions. This, still, is not easy to interpret. A sliding window analysis of the divergence between species (data not shown) failed to discover bias in a specific locus that would explain the differences between coding and non-coding regions. Furthermore, non-synonymous sites seem to be driving this lack of divergence between *S. arcanum* and *S. peruvianum*, while the synonymous sites are dragged into this pattern through their linkage to the non-synonymous sites, since using only synonymous sites to perform PCoA shows a pattern intermediate between that of non-coding regions and non-synonymous sites (data not shown). Taken together, these results could be interpreted as supporting a hypothesis of non-coding regions evolving under strong divergent selection. It has been shown that non-coding regions can in fact present signatures of selection (Andolfatto 2005; Asthana et al. 2007; Boyko et al. 2008). In a review, Servedio & Noor (2003) state that reinforcement includes some form of selection against interspecific mating. Traditionally this encompasses intrinsic genetic incompatibilities. Since coding regions cannot diverge more than purifying selection would allow, it would be plausible to consider non-coding regions to be under divergent selection to ensure incompatibility between closely related species in partial sympatry. To test whether the results from the PCoA are significant, it would be interesting to run STRUCTURE on the same subsets of the data that were used for this analysis.

4. A non-functional mutant allele under positive selection?

Locus CT198 bears high homology to genes encoding aci-reductone-dioxygenase proteins, common to bacteria, plants and animals. The aci-reductone-dioxygenase (ARD) family is involved in the methionine salvage pathway. Abiotic stresses in plants result in elevated levels of ethylene and polyamine. S-adenosylmethionine (SAM) is involved in the biosynthesis of polyamine and ethylene (Ravanel et al. 1998). The enzymatic reactions with SAM in ethylene and polyamine synthesis produce a byproduct, 5′-methylthioadenosine (MTA) that can be recycled to methionine. This methionine salvage pathway is an ubiquitous biochemical pathway that maintains methionine levels, regenerates SAM, and eliminates MTA, thus allowing a high rate of ethylene and polyamine biosynthesis even when the pool of free methionine is small (Bleecker & Kende 2000; Ravanel et al. 1998; Schlenk 1983).

We found a stop codon in the coding region. Given the importance of this pathway, one expects these genes to be under purifying selection. The high frequency of the mutant allele, and the small amount of polymorphism associated with the allele suggest an on-going sweep. However, the fact that the mutant allele is found only in heterozygote state (indicative of the potentially deleterious effect of the mutation) is quite suggestive of balancing selection (heterozygote advantage). Furthermore, given that the mutation found at this locus renders the allele non-functional (which would have a deleterious effect on the fitness of the individual) one is forced to ask why the allele is still present in the population. There are two possible scenarios: the first considers that we are looking at a functional locus, in which case selection should be acting on the locus; the other scenario would be to assume we are looking at a pseudogene, in which case the presence of the allele might be explained by demography, considering a recent origin for the mutant allele (as to account for the lack of polymorphism in it).

If locus CT198 is a single-copy functional gene, then tomatoes need to have an advantage from having a non-functional allele for this locus. It has been reported (Yeh et al. 2001) that a protein exhibiting homology to submergence-induced protein 2A (the protein putatively encoded by CT198) is capable of supporting hepatitis C virus replication in an otherwise non-permissive cell line. If the protein encoded by this locus has a different function (besides that described as part of the methionine salvage pathway) which would allow infection by a pathogen, one can theorize that like in the case of sickle cell anemia, having a deleterious allele can provide heterozygotes with the advantage of partial immunity to a pathogen. However, we cannot exclude at this point the possibility that CT198 is a multiple-copy gene and that at least a copy of the gene is non-functional and therefore not under selection. Still, that would not explain why there is so little polymorphism in the mutant allele. The lack of polymorphism in the mutant allele could only be explained (in this case) by a recent origin (or introgression) of the gene or being under strong purifying selection. In the case of a recent origin, we could explain that the lack of polymorphism is due to an ongoing sweep which would increase the frequency of this allele faster than recombination or mutation can introduce polymorphism around it. Purifying selection seems irreconcilable with a non-functional mutant allele in intermediate frequency. In order to determine whether we are in the presence of a paralog or mutant allele, we suggest sequencing the flanking regions of locus CT198. Comparison of the sequence of the flanking regions between the mutant allele and the wild type allele, would allow us to estimate the divergence between them. If the mutant and the wild type alleles belong to a single-copy gene, divergence between the flanking regions should be comparable between mutant and wild type alleles as within them. On the other hand, in the case of paralogs, we would expect to find more divergence between mutant and wild type alleles than within them.

Conclusions

Although, in general, wild tomatoes and their relatives live in dry to very dry habitats (*e.g.*, the Atacama Desert), they also occur in a great range of other habitats. *Solanum habrochaites* occurs in cloud forest habitats to elevations of 3600m, but is also found in coastal areas and in dry forests on the western Andean slopes. *S. arcanum* inhabits dry habitats, occurring in the inter-Andean valleys subject to severe rain shadows (*e.g.*, in the Valley of the Río Marañón), while *S. chilense* inhabits the extremely dry high-elevation deserts of the western Andean slope. Furthermore, several species are found in the unique lomas habitat along the Pacific coast of Peru and northern Chile (*S. chilense*, *S. habrochaites, S. peruvianum*, and some populations of *S. arcanum*).

Given this wide range of habitats and the importance of abiotic conditions for plant growth, local adaptation for stress is expected between populations. Thus, it is likely that not only signatures of local adaptation would be found in their genomes at genes responsible for water stress tolerance (Dr. Camus-Kulandaivelu, personal communication), but also differences in the strength and variance of purifying selection. Our study found differences in the strength of purifying selection between species as well as differences in the distribution of new mutations' deleterious fitness effects within species, which might be correlated with more stressful or variable environments.

Quantifying levels of selection and adaptation in recently diverged species requires taking into account the speciation process, the influence of demography and their spatial structure. A theoretical possibility for the analysis of such datasets from multiple species and populations is to integrate them in an Approximate Bayesian Computation framework coalescent simulation of speciation models. This would allow testing of complicated speciation models with metapopulation structure. However, such models are computationally very intensive and not developed yet.

Here we have established the utility of wild tomatoes as a model for studying population structure dynamics and the speciation process in recently diverged species. Our approach (divergence population genetics) has therefore proven useful in disentangling the phylogenetic relationship within this group and understanding the basis for species differentiation. A big gap remains though, in integrating these results with morphological data regarding these species.

References

Andolfatto, P., 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437, 1149-1152.

Arnold, M., 1997. *Natural Hybridization and Evolution*, Oxford: Oxford Univ. Press.

Arunyawat, U., Stephan, W. & Stadler, T., 2007. Using Multilocus Sequence Data to Assess Population Structure, Natural Selection, and Linkage Disequilibrium in Wild Tomatoes. *Molecular Biology and Evolution*, 24, 2310-2322.

Asthana, S. et al., 2007. Widely distributed noncoding purifying selection in the human genome. *Proceedings of the National Academy of Sciences*, 104(30), 12410-12415. Available at: http://www.pnas.org/content/104/30/12410.full.

Avise, J., 1989. Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution*, 43, 1192-1208.

Avise, J., 1975. Systematic value of electrophoretic data. *Systematic Zoology*, 23, 465-481.

Baudry, E. et al., 2001. Species and Recombination Effects on DNA Variability in the Tomato Genus. *Genetics*, 158, 1725–1735.

Bernacchi, D. & Tanksley, S.D., 1997. An interspecific backcross of Lycopersicon esculentum × L. hirsutum: Linkage analysis and a QTL study of sexual compatibility factors and floral traits. *Genetics*, 147, 861-877.

Bierne, N. & Eyre-Walker, A., 2004. The genomic rate of adaptive amino acid substitution in Drosophila. *Molecular Biology and Evolution*, 21, 1350-1360.

Bleecker, A. & Kende, H., 2000. Ethylene: a gaseous signal molecule in plants. *Annual Review of Cell and Developmental Biology*, 16, 1-18.

Bloom, A. et al., 2001. Genetics and water relations of the chilling response in wild and cultivated tomatoes. In *Ecological Society of America Annual Meeting*. Madison, WI.

Bohs, L. & Olmstead, R., 1999. Solanum phylogeny inferred from chloroplast DNA sequence data. In M. Nee et al. *Solanaceae IV, Advances in Biology and Utilization*. Kew: Royal Botanical Gardens, pp. 97-110.

Bomblies, K. & Weigel, D., 2007. Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nature Reviews Genetics*, 8, 382-93.

Boyko, A.R. et al., 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics*, 4, e1000083.

Braverman, J. et al., 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, 140, 783-796.

Charlesworth, B., 1998. Measures of Divergence Between Populations and the Effect of Forces that Reduce Variability. *Molecular Biology and Evolution*, 15, 538-543.

Chetelat, R. & Ji, Y., 2007. Cytogenetics and evolution. In M. Razdan & A. Mattoo *Genetic improvement of Solanaceous crops*. Enfield, NH.: Science Publishers, pp. 77-112.

D'Arcy, W. et al., 1979. The classification of the Solanaceae. In *The biology and taxonomy of the Solanaceae.* London: Academic Press for the Linnean Society, p. 3–47.

Darwin, C., 1859. *On the Origin of Species by Means of Natural Selection*, Murray.

Darwin, S., Knapp, S. & Peralta, I.E., 2003. Tomatoes in the Galapagos Islands: morphology of native and introduced species of Solanum section Lycopersicon (Solanaceae). *Systematics and Biodiversity*, 1, 29–54.

De Nettancourt, D., 2001. *Incompatibility and Incongruity in Wild and Cultivated Plants*, Berlin: Springer.

De, A. & Durrett, R., 2007. Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum. *Genetics*, 176, 969–981.

Emelianov, I., Marec, F. & Mallet, J., 2004. Genomic evidence for divergence with gene flow in host races of the larch budmoth. *Proc Biol Sci*, 271, 97-105.

Evanno, G., Regnaut, S. & Goudet, J., 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology*, 14, 2611-20.

Excoffier, L., Hofer, T. & Foll, M., 2009. Detecting loci under selection in a hierarchically structured population. *Heredity*, 103(4), 285-98.

Excoffier, L., Laval, G. & Schneider, S., 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, 1, 47-50.

Excoffier, L., Smouse, P.E. & Quattro, J.M., 1992. Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. *Genetics*, 131, 479-491.

Eyre-Walker, A. & Keightley, P.D., 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular biology and evolution*, 26, 2097-108.

Eyre-Walker, A., Woolfit, M. & Phelps, T., 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, 173, 891-900.

Fay, J.C., Wyckoff, G.J. & Wu, C., 2001. Positive and Negative Selection on the Human Genome. *Genetics*, 158, 1227–1234.

Fisher, R., 1930. The distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh*, 50, 205-220.

Foll, M. & Gaggiotti, O., 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, 180, 977-993.

Fu, Y. & Li, W., 1993. Statistical Tests of Neutrality of Mutations. *Genetics*, 133, 693-709.

Griffiths, R. & Marjoram, P., 1996. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, 3, 479–502.

Hammer, M. et al., 2003. Human population structure and its effects on sampling Y chromosome sequence variation. *Genetics*, 164, 1495–1509.

Hancock, C. et al., 2003. The S-locus and unilateral incompatibility. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 358, 1133.

Hedrick, P.W., 2006. Genetic polymorphism in heterogeneous environments: the age of genomics. *Annual Review of Ecology, Evolution and Systematics*, 37, 67-93.

Hey, J. & Machado, C.A., 2003. The study of structured populations — new hope for a difficult and divided science. *Nature Reviews Genetics*, 4(7), 535-543.

Hey, J. & Nielsen, R., 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, 104, 2785-2790.

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

Hey, J. & Wakeley, J., 1997. A coalescent estimator of the population recombination rate. *Genetics*, 145, 833.

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

Hey, J., 1991. The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. *Genetics*, 128, 831-840.

Hogenboom, N.G., 1972. Breaking breeding barriers in Lycopersicon.1. The genus Lycopersicon, its breeding barriers and the importance of breaking these barriers. *Euphytica*, 21, 221-227.

Hudson, R.R. & Slatkin, M., 1992. Levels of Gene Flow From DNA Sequence Data. *Genetics*, 132, 583-589.

Hudson, R.R., Boos, D. & Kaplan N.L., 1992. A Statistical Test For Detecting Geographic Subdivision. *Molecular Biology And Evolution*, 9, 138-151.

Hudson, R.R., 1990. Gene genealogies and the coalescent process. In D. Futuyma & J. Antonovics *Oxford Surveys in Evolutionary Biology Vol. 7.* Oxford: Oxford University Press, p. 1–43.

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

Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23, 183-201.

Igic, B. & Kohn, J.R., 2001. Evolutionary relationships among self-incompatibility RNases. *Proceedings of the National Academy of Sciences*, 98, 13167–13171.

Jenkins, J., 1948. The origin of the cultivated tomato. *Economic Botany*, 2, 379-392.

Jost, L., 2008. GST and its relatives do not measure differentiation. *Molecular Ecology*, 17, 4015–4026.

Kaj, I., Krone, S. & Lascoux, M., 2001. Coalescent theory for seed bank models. *Journal of Applied Probability*, 38, 285-300.

Kaneshiro, K.Y., 1980. Sexual Isolation, Speciation and the Direction of Evolution. *Evolution*, 34, 437-444.

Kaplan, N., Darden, T. & Hudson, R.R., 1988. Coalescent process in models with selection. *Genetics*, 120, 819–829.

Keightley, P.D. & Eyre-Walker, A., 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, 177, 2251-61.

Kimura, M. & Weiss, G., 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49, 561–576.

Kingman, J., 1982. The coalescent. *Stochastic Process. Appl.*, 13, 235–248.

Kreitman, M., 2000. Methods to detect selection in populations with applications to humans. *Annu. Rev. Genomics Hum. Genet.*, 1, 539–559.

Krüger, J. et al., 2002. A tomato cysteine protease required for Cf-2-dependent disease resistance and suppression of autonecrosis. *Science*, 296, 744-747.

Laporte, V. & Charlesworth, B., 2002. Effective Population Size and Population Subdivision in Demographically Structured Populations. *Genetics*, 162, 501-519.

Lawton-Rauh, A., Friar, E. & Remington, D., 2007. Collective evolution processes and the tempo of lineage divergence in the Hawaiian silversword alliance adaptive radiation (Heliantheae, Asteraceae). *Molecular Ecology*, 16, 3993-3994.

Lawton-Rauh, A., R.H., R. & M.D., P., 2007. Diversity and divergence patterns in regulatory genes suggest differential gene flow in recently derived species of the Hawaiian silversword alliance adaptive radiation (Asteraceae). *Molecular Ecology*, 16, 3995-4013.

Librado, P. & Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25, 1451.

Linnaeus, C., 1753. *Species plantarum*, Stockholm: Holmiae.

Luria, S. & Delbrück, M., 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28, 491–511.

Malecot, G., 1973. Genetique des populations diploıdes naturelles dans le cas d'un seul locus. III. Parente, mutations et migration. *Ann. Genet. Sel. Anim.*, 5, 333–361.

Malecot, G., 1975. Heterozygosity and relationship in regularly sub-divided populations. *Theoretical Population Biology*, 8, 212–241.

Malecot, G., 1969. *The Mathematics of Heredity*, San Francisco: W. H. Freeman.

Mayr, E., 1963. *Animal species and evolution*, The Belknap press of Harvard university press.

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

McCormick, S., 1998. Self-incompatibility and other pollen-pistil interactions. *Current Opinion in Plant Biology*, 1, 18–25.

McDaniel, S.F. & Shaw, A.J., 2005. Selective sweeps and intercontinental migration in the cosmopolitan moss Ceratodon purpureus (Hedw.) Brid. *Molecular Ecology*, 14, 1121-1132.

McDonald, J. & Kreitman, M., 1991. Adaptive Protein Evolution At The Adh Locus In Drosophila. *Nature*, 351, 652-654.

McVean, G., 2009. A Genealogical Interpretation of Principal Components Analysis. *PLoS Genetics*, 5, e1000686.

Miller, J. & Tanksley, S.D., 1990. RFLP analysis of phylogenetic relationships and genetic variation in the genus Lycopersicon. *Theoretical and Applied Genetics*, 80, 437-448.

Mishler, B. & Donoghue, M., 1982. Species Concepts - A Case For Pluralism. *Systematic Zoology*, 31, 491.

Moeller, D., Tenaillon, M. & Tiffin, P., 2007. Population Structure and Its Effects on Patterns of Nucleotide Polymorphism in Teosinte (Zea mays ssp. parviglumis). *Genetics*, 176(3), 1799-1809.

Morjan, C. & Rieseberg, L.H., 2004. How species evolve collectively: implications of gene ow and selection for the spread of advantageous alleles. *Molecular Ecology*, 13, 1341-1356.

Mountain, J. & Cavalli-Sforza, L., 1994. Inference Of Human-Evolution Through Cladistic-Analysis Of Nuclear-Dna Restriction Polymorphisms. *Proceedings Of The National Academy Of Sciences*, 91, 6515-6519.

Moyle, L.C. & Graham, E., 2004. Genetics of hybrid incompatibility between Lycopersicon esculentum and L. hirsutum. *Genetics*, 169, 355.

Moyle, L.C. & Nakazato, T., 2008. Comparative genetics of hybrid incompatibility: sterility in two Solanum species crosses. *Genetics*, 179, 1437.

Moyle, L.C., 2008. Ecological and evolutionary genomics in the wild tomatoes (solanum sect. lycopersicon). *Evolution*, 62, 2995-3013.

Nakazato, T., Bogonovich, M. & Moyle, L.C., 2008. Environmental Factors Predict Adaptive Phenotypic Differentiation Within And Between Two Wild Andean Tomatoes. *Evolution*, 62, 774-792.

Nei, M. & Gojobori, T., 1986. Simple methods for estimating the numbers ofsynonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3, 418-426.

Nei, M., Maruyama, T. & Chakraborty, R., 1975. The bottleneck effect and genetic variability in populations. *Evolution*, 29, 1-10.

Nei, M., 1987. *Molecular Evolutionary Genetics* 2nd., New York: Columbia Univ. Press.

Neigel, J.E., 2002. Is FST obsolete? *Conservation Genetics*, 3, 167-173.

Nesbitt, T.C. & Tanksley, S.D., 2002. Comparative Sequencing in the Genus Lycopersicon: Implications for the Evolution of Fruit Size in the Domestication of Cultivated Tomatoes. *Genetics*, 162, 365–379.

Neuhauser, C. & Krone, S., 1997. The genealogy of samples in models with selection. *Genetics*, 145, 519–534.

Nielsen, R. & Wakeley, J., 2001. Distinguishing Migration From Isolation: A Markov Chain Monte Carlo Approach. *Genetics*, 158, 885–896.

Nielsen, R., 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity*, 86, 641-647.

Nordborg, M. & Tavare, S., 2002. Linkage disequilibrium: What history has to tell us. *Trends in Genetics*, 18, 83–90.

Nordborg, M., 2001. No Title. In D. Balding, M. Bishop, & C. Cannings *Handbook of Statistical Genetics*. Chichester, UK: John Wiley & Sons, p. 179–212.

Novembre, J. & Stephens, M., 2008. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40, 646-649.

Novembre, J. et al., 2008. Genes mirror geography within Europe. *Nature*, 456, 98-101.

Olivieri, I., Couvet, D. & Gouyon, P., 1990. The genetics of transient populations: research at the metapopulation level. *Trends in Ecology & Evolution*, 5, 207–210.

Olmstead, R. et al., 1999. Phylogeny and provisional classification of the Solanaceae based on chloroplast DNA. In M. Nee et al. *Solanaceae IV, Advances in Biology and Utilization*. Kew: Royal Botanical Gardens, pp. 111-137.

Pamilo, P. & Nei, M., 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5, 568-583.

Pannell, J.R. & Charlesworth, B., 1999. Neutral genetic diversity in a metapopulation with recurrent local extinction and recolonization. *Evolution*, 53, 664-676.

Patterson, B.D., Paull, R. & Smillie, R.M., 1978. Chilling resistance in Lycopersicon hirsutum Humb. and Bonpl. a wild tomato with a wide altitudinal distribution. *Plant Physiology*, 5, 609-617.

Peralta, I.E. & Spooner, D.M., 2001. Granule-Bound Starch Synthase (Gbssi) Gene Phylogenyof Wild Tomatoes (Solanum L. Section Lycopersicon [Mill.] Wettst. Subsection Lycopersicon). *American Journal of Botany*, 88, 1888-1902.

Peralta, I.E. & Spooner, D.M., 2005. Morphological Characterization and Relationships of Wild Tomatoes (Solanum L. sect. Lycopersicon). In *Monographs In Systematic Botany*. Missouri Botanical Garden, pp. 227-257.

Peralta, I.E., Knapp, S. & Spooner, D.M., 2005. New Species of Wild Tomatoes (Solanum Section Lycopersicon: Solanaceae) from Northern Peru. *Systematic Botany*, 30, 424-434.

Peralta, I.E., Knapp, S. & Spooner, D.M., 2006. Nomenclature For Wild And Cultivated Tomatoes. *Report of the Tomato Genetics Cooperative*, 56, 6-12.

Peralta, I.E., Knapp, S. & Spooner, D.M., 2006. Nomenclature for wild and cultivated tomatoes. *Tomato Genetics Cooperative Report*, 56, 6-12.

Peralta, I.E., Spooner, D.M. & Knapp, S., 2008. Taxonomy of Wild Tomatoes and their Relatives (Solanum sect. Lycopersicoides, sect. Juglandifolia, sect. Lycopersicon; Solanaceae). In C. Anderson *Systematic Botany Monographs Volume 84*. The American Society of Plant Taxonomists.

Perrier, X. & Jacquemoud-Collet, J., 2006. DARwin software. Available at: http://darwin.cirad.fr/darwin.

Pritchard, J.K. et al., 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16, 1791–1798.

Pritchard, J.K., Stephens, M. & Donnelly, P., 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155, 945–959.

Pritchard, J.K., Stephens, M. & Donnelly, P., 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155, 945.

Pritchard, J.K., Wen, X. & Falush, D., 2007. *Documentation for structure software: Version 2.2*,

Ptak, S. & Przeworski, M., 2002. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends in Genetics*, 18, 559–563.

R Development Core Team, 2005. *Computing RFfS*, Vienna, Austria.

Ramos-Onsins, S.E. & Rozas, J., 2002. Statistical Properties of New Neutrality Tests Against Population Growth. *Molecular Biology and Evolution*, 19(12), 2092-2100.

Ramsey, J., Bradshaw, H. & Schemske, D., 2003. Components of reproductive isolation between the monkeyflowers Mimulus lewisii and M. cardinalis (Phrymaceae). *Evolution*, 57, 1520-1534.

Ravanel, S. et al., 1998. The specific features of methionine biosynthesis and metabolism in plants. *Proceedings of the National Academy of Sciences*, 95, 7805-7812.

Ray, N., Currat, M. & Excoffier, L., 2003. Intra-deme molecular diversity in spatially expanding populations. *Molecular Biology and Evolution*, 20, 76–86.

Rick, C.M. & Fobes, J., 1975. Allozyme variation in the cultivated tomato and closely related species. *Bulletin of the Torrey Botanical Club*, 102, 376–384.

Rick, C.M. et al., 1976. Genetic and biosystematic studies on two new sibling species of Lycopersicon from interandean Peru. *Theoretical and Applied Genetics*, 47, 55–68.

Rick, C.M., Fobes, J.F. & Tanksley, S.D., 1979. Evolution of mating systems in Lycopersicon hirsutum as deduced from genetic variation in electrophoretic and morphological characters. *Plant Systematics and Evolution*, 132, 279-298.

Rick, C.M., Holle, M. & Thorp, R., 1978. Rates of cross-pollination inLycopersicon pimpinellifolium: Impact of genetic variation in floral characters. *Plant Systematics and Evolution*, 129, 31–44.

Rick, C.M., Zobel, R. & Fobes, J., 1974. Four peroxidase loci in red-fruited tomato species: genetics and geographic distribution. *Proceedings of the National Academy of Sciences*, 71, 835–836.

Rick, C.M., 1963. Barriers to Interbreeding in Lycopersicon peruvianum. *Evolution*, 17, 216-232.

Rick, C.M., 1963. Barriers to interbreeding in Lycopersicon peruvianum. *Evolution*, 17, 216-232.

Rick, C.M., 1979. Biosystematic studies in Lycopersicon and closely related species in Solanum. In J. Hawkes, R. Lester, & A. Skelding *The biology and taxonomy of the Solanaceae.* London: Academic Press for the Linnean Society, p. 667–678.

Rick, C.M., 1976. Natural variability in wild species of Lycopersicon and its bearing on tomato breeding. *Genetica Agraria*, 30, 249-259.

Rick, C.M., 1973. Potential genetic resources in tomato species: clues from observations in native habitats. In A. Srb *Genes, Enzymes, and Populations*. New York, NY: Plenum Press, p. 255.

Rick, C.M., 1986. Reproductive isolation in the Lycopersicon peruvianum complex. In W. D'arcy *Solanaceae, biology and systematics*. New York: Columbia University Press.

Rieseberg, L.H. & Willis, J., 2007. Plant Speciation. *Science*, 317, 910-914.

Rieseberg, L.H., Wood, T. & Baack, E., 2006. The nature of plant species. *Nature*, 440, 524-527.

Roselius, K., Stephan, W. & Stadler, T., 2005. The Relationship of Nucleotide Polymorphism, Recombination Rate and Selection in Wild Tomato Species. *Genetics*, 171, 753-763.

Rosenberg, N.A. & Nordborg, M., 2002. Genealogical Trees, Coalescent Theory And The Analysis Of Genetic Polymorphisms. *Nature Reviews Genetics*, 3, 380-390.

Schlenk, F., 1983. Methylthioadenosine. *Adv. Enzymol.*, 54, 195-265.

Servedio, M.R. & Noor, M.A., 2003. The Role Of Reinforcement In Speciation: Theory And Data. *Annual Review of Ecology, Evolution and Systematics*, 34, 339-364.

Shaw, K., 2002. Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proceedings of the National Academy of Sciences*, 99, 16122–16127.

Sherman, J.D. & Stack, S.M., 1995. Two-Dimensional Spreads of Synaptonemal Complexes from Solanaceous Plants. VI. High-Resolution Recombination Nodule Map for Tomato (Lycopersicon esculentum). *Genetics*, 141, 683-708.

Slatkin, M., 1977. Gene flow and genetic drift in a species subject to frequent local extinction. *Theoretical Population Biology*, 12, 253–262.

Slatkin, M., 2001. Simulating genealogies of selected alleles in a population of variable size. *Genet. Res.*, 78, 49–57.

Slatkin, M., 1987. The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology*, 32, 42–49.

Smith, N.G. & Eyre-Walker, A., 2002. Adaptive protein evolution in Drosophila. *Nature*, 415, 1022-1024.

Spooner, D.M., Anderson, G.J. & Jansen, R.K., 1993. Chloroplast DNA evidence for the interrelationships of tomatoes, potatoes and pepinos (Solanaceae). *American Journal of Botany*, 80, 676-688.

Spooner, D.M., Peralta, I.E. & 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.

Stadler, T. et al., 2009. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics*, 182, 205-16.

Stadler, T., Arunyawat, U. & Stephan, W., 2008. Population Genetics of Speciation in Two Closely Related Wild Tomatoes (Solanum Section Lycopersicon). *Genetics*, 178, 339-350.

Stadler, 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.

Stephan, W. & Langley, C.H., 1998. DNA Polymorphism in Lycopersicon and Crossing-Over per Physical Length. *Genetics*, 150, 1585–1593.

Stephens, M., Smith, N. & Donnelly, P., 2001. A new statistical method for haplotype reconstruction from population data. *American Journal Of Human Genetics*, 68, 978-989.

Strobeck, C., 1987. Average Number of Nucleotide Differences in a Sample From a Single Subpopulation: A Test for Population Subdivision. *Genetics*, 117, 149-153.

Tajima, F., 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, 123, 585-595.

Takahata, N., Lee, S. & Satta Y., 2001. Testing multiregionality of modern human origins. *Molecular Biology and Evolution*, 18, 172-183.

Tanksley, S.D. & Loaiza-Figueroa, F., 1985. Gametophytic Self-Incompatibility is Controlled by a Single Major Locus on Chromosome-1 in Lycopersicon peruvianum. *Proceedings of the National Academy of Sciences*, 82, 5093-5096.

Tanksley, S.D. et al., 1992. High Density Molecular Linkage Maps of the Tomato and Potato Genomes. *Genetics*, 132, 1141-1 160.

Tavare, S. et al., 1997. Inferring coalescence times from DNA sequence data. *Genetics*, 145, 505-518.

Taylor, I.B., 1986. Biosystematics of the tomato. In J. G. Atherton & J. Rudich *The tomato crop: a scientific basis for improvement*. London: Chapman and Hall, pp. 1-34.

Tellier, A., Villaréal, L. & Giraud, T., 2005. Maintenance of sex-linked deleterious alleles by selfing and group selection in metapopulations of the phytopathogenic fungus Microbotryum violaceum. *The American Naturalist*, 165, 577-89.

Turner, T., Hahn, M. & Nuzhdin, S., 2005. Genomic islands of speciation in Anopheles gambiae. *PLoS Biology*, 3, e285.

Ungerer, M. et al., 1998. Rapid hybrid speciation in wild sunflowers. *Proceedings Of The National Academy Of Sciences*, 95, 11757-11762.

Vallejos, C.E., 1979. Genetic diversity of plants for response to low temperatures and its potential use in crop plants. In J. Lyons *Evolution*. p. 565.

Wade, M. & McCauley, D., 1988. Extinction and recolonization: their effects on the genetic differentiation of local populations. *Evolution*, 42, 995–1005.

Wakeley, J. & Aliacar, N., 2001. Gene Genealogies in a Metapopulation. *Genetics*, 159, 893–905.

Wakeley, J. & Hey, J., 1997. Estimating Ancestral Population Parameters. *Genetics*, 145, 847-855.

Wakeley, J., 1996. Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical Population Biology*, 49, 369-386.

Wakeley, J., 1999. Nonequilibrium Migration in Human History. *Genetics*, 153, 1863-1871.

Wakeley, J., 1998. Segregating sites in Wright's island model. *Theoretical Population Biology*, 53, 166–175.

Wakeley, J., 2000. The effects of subdivision on the genetic divergence of populations and species. *Evolution*, 54, 1092-1101.

Wall, J., 2000. A comparison of estimators of the population recombination rate. *Molecular Biology and Evolution*, 17, 156–163.

Wang, J. & Caballero, A., 1999. Developments in predicting the effective size of subdivided populations. *Heredity*, 82, 212-226.

Watterson, G., 1975. On the Number of Segregating Sites in Genetical Models without Recombination. *Theoretical Population Biology*, 7, 256-276.

Weiss, G. & von Haeseler, A., 1998. Inference of population history using a likelihood approach. *Genetics*, 149, 1539–1546.

Whitlock, M. & McCauley, D., 1999. Indirect measures of gene flow and migration: FST doesn't equal 1/(4Nm+1). *Heredity*, 82, 117–125.

Whitlock, M., 2003. Fixation probability and time in subdivided populations. *Genetics*, 164, 767-779.

Wilkinson-Herbots, H., 1998. Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology*, 37, 535–585.

Williams, C. & St. Clair, D., 1993. Phenetic relationships and levels of variability detected by restriction fragment length polymorphism and random amplified polymorphic DNA analysis of cultivated and wild accessions of Lycopersicon esculentum. *Genome*, 36, 619–630.

Wingen, L., Brown, J. & Shaw, M., 2007. The population genetic structure of clonal organisms generated by exponentially bounded and fat-Tailed dispersal. *Genetics*, 177, 435-448.

Wiuf, C. & Hein, J., 1999. Recombination as a point process along sequences. *Theoretical Population Biology*, 55, 248–259.

Wright, S., 1931. Evolution in Mendelian populations. *Genetics*, 16, 97-159.

Wright, S., 1943. Isolation by distance. *Genetics*, 28, 114–138.

Wright, S., 1938. The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences*, 24, 253-259.

Wright, S., 1951. The genetical structure of populations. *Annals of Eugenics*, 15, 323–354.

Wright, S., 1940. The statistical consequences of mendelian heredity in relation to speciation. In J. Huxley *The New Systematics*. Oxford University Press, pp. 161-183.

Yeh, C. et al., 2001. Identification of a hepatic factor capable of supporting hepatitis C virus replication in a nonpermissive cell line. *The Journal of Virology*, 75, 11017-11024.

Young, K. et al., 2002. Plant evolution and endemism in Andean South America: An introduction. *The Botanical Review*, 68, 4–21.

Curriculum Vitae

Personal Data

Name: Carlos Gonzalo Merino Méndez

Date and Place of Birth: June 16th, 1978 in Lima, Peru

Nationality: Peruvian

Professional Training

3/1997 – 12/2002	Biological Sciences Faculty San Marcos Public Head University – UNMSM (Lima, Peru) Bachelor’s Degree in Biology
9/2006	Biological Sciences Faculty San Marcos Public Head University – UNMSM (Lima, Peru) Licenciatura (Diploma) in Biology
7/2006 – Present	Biological Sciences Faculty Ludwig Maximilian University – LMU (Munich, Germany) PhD (Doktorarbeit) in Biology

Work Experience

9/1999 – 9/2001	Biochemistry and Nutrition Research Centre (CIBN) San Marcos Public Head University – UNMSM (Lima, Peru) Lab Assistant
1/2002 – 07/2006	Language Centre – Pontifical Peruvian Catholic University – PUCP (Lima, Peru) Part-Time English Teacher
5/2003 – 8/2003	International Potato Centre – CIP (Lima, Peru) Internship
9/2003 – 06/2006	International Potato Centre – CIP (Lima, Peru) Research conducive to a Thesis at UNMSM Grant Holder

Additional Training

4/1998 Antonio Raymondi Biological Sciences Research Institute (ICBAR) San Marcos Public Head University (Lima, Peru)
7th ICBAR's Scientific Reunion

11/1998 Peruvian Professional School of Biologists (Trujillo, Peru)
1st National Biotechnology and Bioengineering Congress

11/1999 Ricardo Palma Private University (Lima, Peru)
Molecular Biology International Seminar

9/2001 Iberian American Cellular Biology Society & Peruvian Professional School of Biologists
7th Iberian American Cellular Biology Congress and 1st Peruvian Cellular Biology Congress

Posters and Publications

4/2000 7th ICBAR's Scientific Reunion, Antonio Raymondi Biological Sciences Research Institute (ICBAR) San Marcos Public Head University (Lima, Peru)
“Recuperación y Cuantificación del DNA Atrapado en Matriz de Sílica” (Recovery and Quantification of DNA trapped in a Silica Matrix)

1/2005 XIV Plant & Animal Genome Conference (San Diego, California) “Divergence of microsatellite loci potatoes between wild and cultivated potatoes”

5/2009 Genetics, 182 (1), 205-16, (2009). Städler, T., Haubold, B., Merino, C., Stephan, W., & Pfaffelhuber, P.
“The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations.”

Acknowledgements

First, I would like to thank Prof. Wolfgang Stephan for giving me the opportunity to carry out my doctoral thesis in his group. I would also like to thank Thomas Staedler for developing the research project for this thesis. Additionally, I would like to thank Laura Rose for all the advice, academic and otherwise. Many thanks to four very special people: to Simone and Hilde for the sequencing; and Aurelien and Létizia for the rescuing.

I would also like to thank all my fellow (former and current) PhD students for all the fun, the support, and the friendship. In order of appearance (more or less): to Ann, for the philosophical chats; to Sarah, for everything from where to buy anything and everything, to helping me get up after falling; to Tobi, for introducing me to the project and translating between Spanish and German (not to mention the kindness and the warmth); to Andy, for the laughter; to Angelika, for being an angel; to Christoph Heibl, for the hospitality; to Ana Gutierrez, for being an emergency kit in case of homesickness; to Anja (a.k.a. as Bree), for all the support and fun; to Annegret, for making it much more interesting; to Iris, for the chats in secluded places, for lending me an ear (some times even against her will) and for the Buffy/Angel addiction and so much more; to Robert, for helping me out every time I had to write a script in R, create input files, transform image files format… you get the idea (I don't know how I'm supposed to program if you're not there whenever I get stuck), but most of all for the friendship; to Claus, for making it a fun ride; to Hui, for the shared laughs; to the master students, you guys rock; to Pleuni, for the support on so many things; to Rita, for being so sweet and fun. I'm probably forgetting someone, but, in my defense, I'm still recovering from last night's celebration.

Finally, I have to thank my parents, my sister and the rest of my family for their support and their inspiring me. I wouldn't be here if it wasn't for you.