Population structure and speciation history of two closely related wild tomato species

Uraiwan Arunyawat

München 2007

Dissertation zur Erlangung des Doktorgrades
der Naturwissenschaften an der Fakultät für Biologie der
Ludwig–Maximilians–Universität München

Population structure and speciation history of two closely related wild tomato species

Uraiwan Arunyawat

aus

Roi-et, Thailand

2007

Erklärung

Diese Dissertation wurde im Sinne von § 12 der Promotionsordnung von Prof. 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, ohne unerlaubte Hilfe angefertigt wurde.

Uraiwan Arunyawat

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

Dissertation eingereicht am: 04.04.2007
Mündliche Prüfung am: 01.06.2007

To my parents,
with so much love

Contents

List of Tables

Note

In this thesis, I present the results from my doctoral project, which I have done for 3.5 years in Munich (October 2003-March 2007) under the supervision of Prof. Wolfgang Stephan and Dr. Thomas Städler. The results from my thesis have contributed to the following manuscripts:

Arunyawat U., W. Stephan, and T. Städler. 2007. Using multilocus sequence data to assess population structure, natural selection and linkage disequilibrium in two closely related wild tomato species. *MBE-manuscript submitted*.

Städler T., U. Arunyawat, and W. Stephan. 2007. Population genomics of speciation in two closely related wild tomatoes (*Solanum* section *Lycopersicon*). *Genetics-manuscript submitted*.

In this thesis, 'we' refers to my collaborators and me. For chapter one, I did all of the analytical work, and for chapter two Thomas Städler contributed to the analytical framework of divergence population genetics. The work in this thesis was funded by the Deutsche Forschungsgemeinschaft through its Priority Program 'Radiations - Origins of Biological Diversity' (SPP-1127), grant Ste 325/5 to Wolfgang Stephan and a Deutsche Akademischer Austausch Dienst (DAAD) fellowship to me.

Summary

Two important aspects were investigated in this dissertation. First, using a multilocus approach to examine the effects of population structure, demography and natural selection on DNA polymorphism in two closely related wild tomatoes, *Solanum peruvianum* and *S. chilense* (*Solanum* section *Lycopersicon*, *Solanaceae*). Sequence data were used for eight unlinked nuclear loci from populations across much of the species' range. Both species exhibit substantial levels of nucleotide variation. The average level of silent nucleotide diversity across all loci in *S. peruvianum* ($\theta_{sil} \approx 2.04\%$) is about 1.3-fold higher than in *S. chilense* ($\theta_{sil} \approx 1.59\%$). One of the loci deviates from neutral expectations, showing a clinal pattern of nucleotide diversity and haplotype structure in *S. chilense*. This geographic pattern is likely caused by an incomplete (ongoing) selective sweep. Both studied wild tomato species exhibit moderate levels of population differentiation (average $F_{st} \approx 0.15$). These estimates of F_{st} may seem surprisingly low in view of the high fragmentation of local populations. It is likely that patterns of population differentiation in our samples reflect the presence of soil seed banks, and historical association mediated by climatic cycles. Interestingly, the pooled sample across different demes exhibits a negative Tajima's D, as possibly a consequence of ancestral population structure. We therefore propose that population structure is one of the most important evolutionary forces to shape patterns of nucleotide diversity within and among populations in these wild tomatoes. Furthermore, intragenic linkage disequilibrium decays very rapidly with physical distance (within a few hundred base pairs), suggesting high recombination rates and effective population sizes in both species. The rapid decline of linkage disequilibrium seems very promising for future association studies with the purpose of mapping functional variation in wild tomatoes.

Second, assessing genealogical footprints of speciation history of wild tomatoes. We present a multilocus sequencing study to assess patterns of polymorphism and divergence in the closely related wild tomato species, *Solanum peruvianum* and *S. chilense*. The dataset comprises seven mapped nuclear loci ($\approx$ 9.3 kb of analyzed sequence across loci) and four local population samples per species that cover much of the species' range (between 80-88 sequenced alleles across both species). Specifically, we employ the analytical framework of divergence population genetics in evaluating the utility of the 'isolation' model of speciation to explain observed patterns of poly-

morphism and divergence. Whereas the isolation model is not rejected by goodness-of-fit criteria established via coalescent simulations, patterns of intragenic linkage disequilibrium provide compelling evidence for historical introgression at two of the seven loci. These results suggest that speciation occurred under residual gene flow, implying natural selection as one of the evolutionary forces driving the divergence of these tomato species. The complexities due to the joint effects of the coalescent process in subdivided populations and the sampling scheme may have conspired to bias the demographic estimates and the scaled time since speciation; there is an obvious need to develop more refined models of divergence that explicitly take population subdivision into account in making historical inferences.

General Introduction

'Nothing in biology makes sense except in the light of evolution'
(Dobzhansky, 1973)

About this thesis and introduction

My thesis involves two aspects. First, quantifying the magnitude and patterns of population structure, natural selection and linkage disequilibrium in two closely related wild tomato species, *Solanum peruvianum* and *S. chilense*, using population-genetic approaches (chapter one). Second, assessing the speciation history of these two taxa within an explicitly population-genetic framework (chapter two). Both chapters are manuscripts for papers to be submitted, which have their required formats and introductions. Both are self-contained and can be read separately. The idea of the following sections is thus to generally introduce the topics dealt within this thesis.

Wild tomatoes studied and their habitats

The tomato clade (*Solanum* section *Lycopersicon*) consists of up to thirteen species, of which the only cultivated species is *S. lycopersicum*. Eleven of them are native to western South America, and two (*S. cheesmanii* and *S. galapagense*) are endemic to the Galapagos Islands (Rick, 1986; Taylor, 1986; Spooner et al., 2005; Peralta et al., 2005). The phylogenetic relationships among the tomato species are not well resolved. However, many analyses revealed two well-defined clades: one corresponding to mating system (self-compatible versus self-incompatible species) and the other, a subset of the first, corresponding to fruit color (red versus green fruits) (Miller and Tanksley, 1990).

Wild tomatoes have become an ideal plant model system for evolutionary analyses because of their recent divergence, the clear phenotypic distinction and the great diversity of mating systems. In this study, we used *S. peruvianum* and *S. chilense* as model taxa to examine the patterns of population structure, demography, and speciation processes in these two closely related self-incompatible species (Figure 1). *S. peruvianum* is distributed along the western side of the Andes from north-central

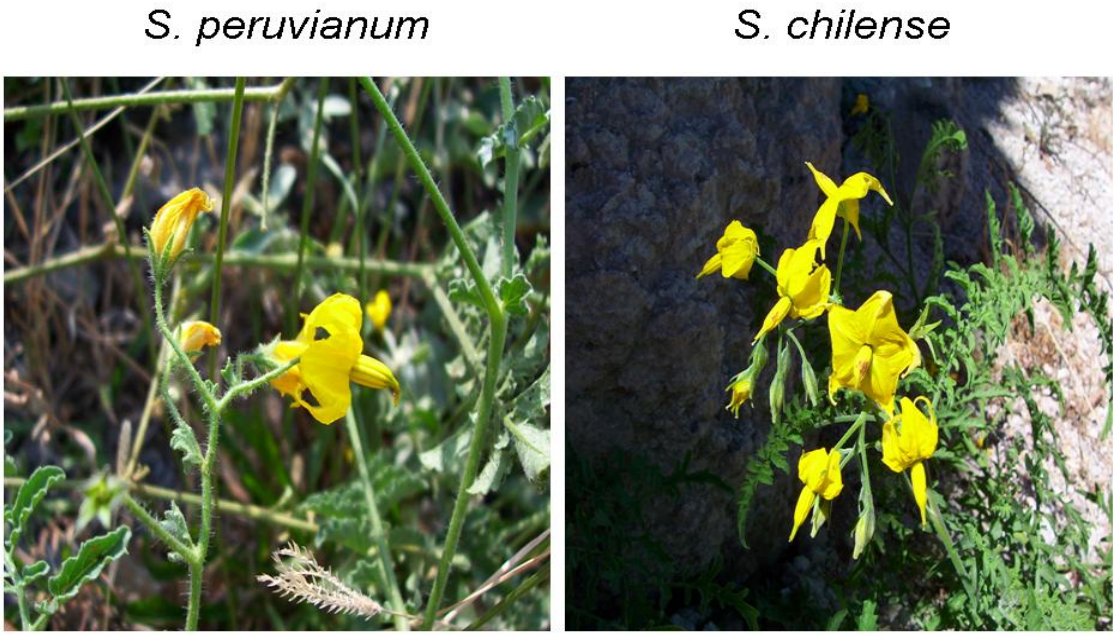

Figure 1: Morphological distinction between *S. peruvianum* and *S. chilense*; *e.g. S. peruvianum* has a curved anther cone but *S. chilense* has a straight anther cone. Leaves of *S. peruvianum* are generally larger than of *S. chilense*. Leaf-shape is more finely subdivided in *S. chilense* ('fern-like') than in *S. peruvianum*. Moreover, *S. peruvianum* has hairy stems, whereas *S. chilense* has smooth stems.

Peru to northern Chile, and *S. chilense* from southern Peru to northern Chile (Figure 2). Both studied taxa grow in a wide diversity of habitats from sea level to the highland up to 3,300 meters (Rick, 1986; Taylor, 1986; see also Figures 3 and 4). Earlier studies, based on single populations in each of five species, suggested that positive directional selection does not have a large effect in the tomato clade; therefore, the analyses of demographic processes and population structure become more significant for understanding patterns of genetic diversity and historical events in wild tomatoes (Städler et al., 2005; Roselius et al., 2005).

Additionally, the differences between the northern and southern populations of *S. peruvianum* have long been recognized. Rick (1986) demonstrated that the northern races were partially crossable among themselves, but showed reduced crossability to southern populations. *S. chilense* is restricted to northern Chile and the three southernmost Peruvian departments (Arequipa, Moquegua, Tacna), where it is

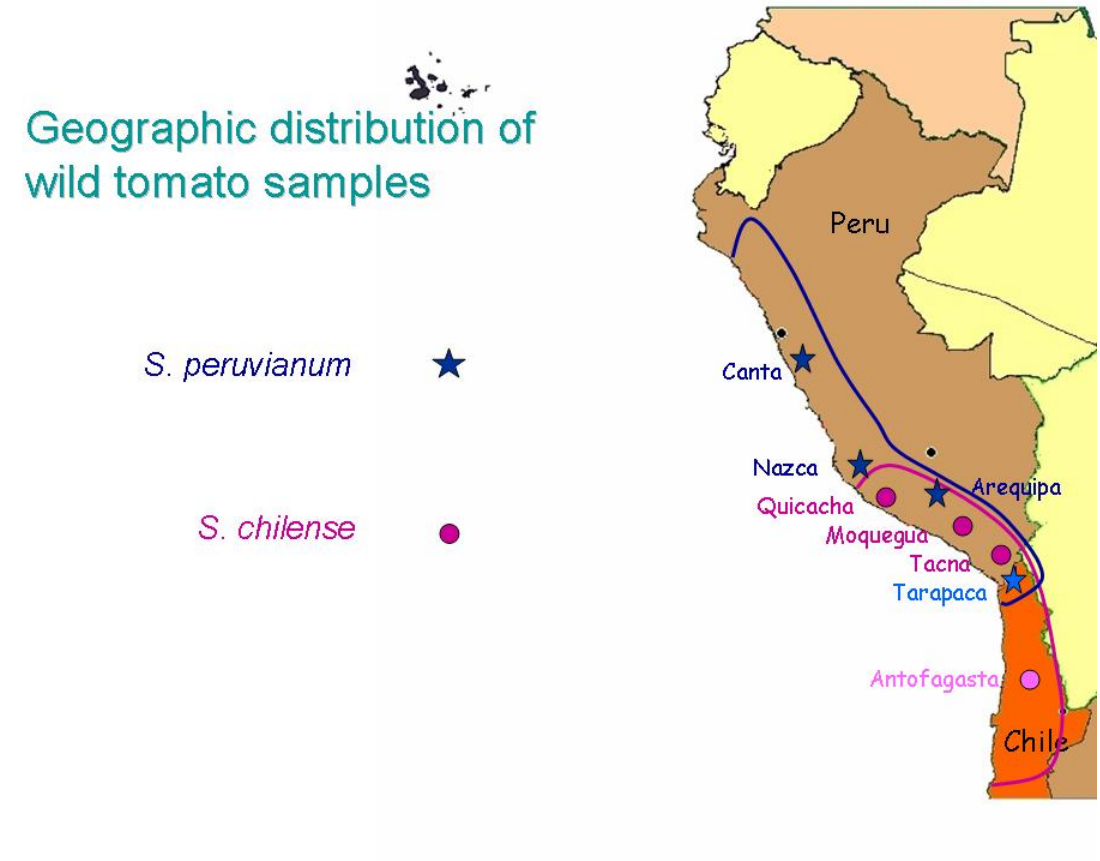

Figure 2: Approximate geographic distribution of wild tomato samples. From north to south, the blue line indicates the geographic range of *S. peruvianum* and the purple line shows the geographic range of *S. chilense*. Note that there is a region of sympatry of both species (overlap between the blue and purple lines). *S. peruvianum* samples are indicated with blue stars and *S. chilense* samples with purple circles. Note that the light blue star and purple circle indicate the old samples from the previous study (Städler et al., 2005).

broadly sympatric with the widely distributed *S. peruvianum* (see Figure 2). Interestingly, only *S. peruvianum* from the far-northern part of its geographic range can be hybridized with *S. chilense*, whereas populations in regional sympatry appear to be genetically isolated (Rick and Lamm, 1955; Rick, 1979). Therefore, it is of great interest to elucidate the details of the speciation process in the two closely related wild tomatoes and the divergence patterns of these morphologically distinct species using a population genetic approach.

Figure 3: Natural habitats of *S. peruvianum*. Both habitats are at similar elevations of ≈ 2,100 meters, however, plenty of plants were found in the Canta habitat, whereas only few plants along the road in the Nazca habitat (at the time of collection, May 2004).

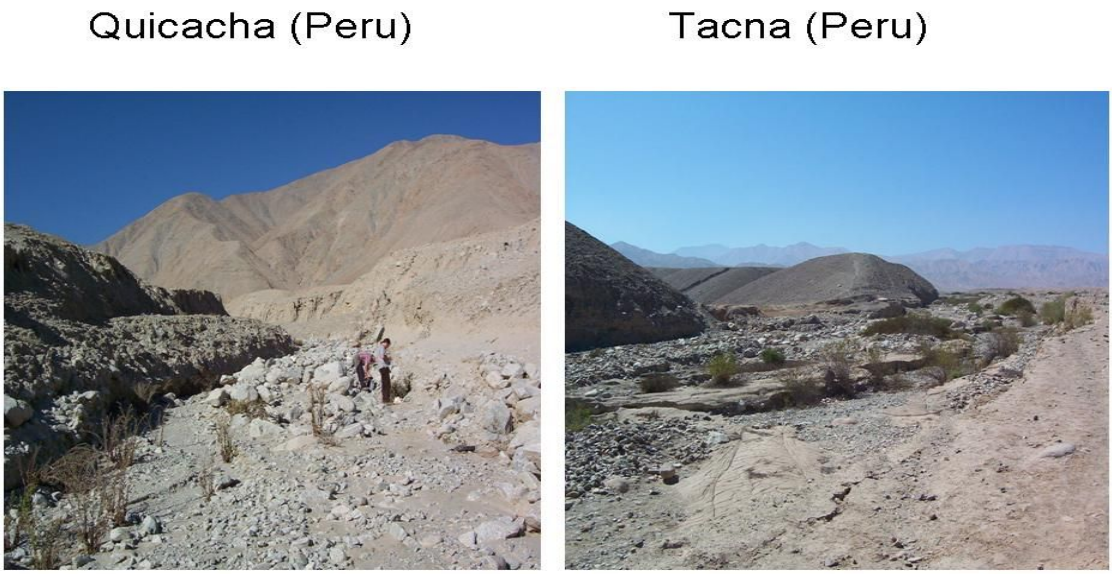

Figure 4: Natural habitats of *S. chilense*. Both habitats are arid deserts; the Quicacha habitat is at an elevation of ≈ 1,800 meters (some *S. peruvianum* plants co-occur in sympatry). The Tacna collection site is at an elevation of ≈ 1,260 meters.

Neutral theory and neutrality tests

In the late 1960s, Motoo Kimura first proposed the neutral theory of molecular evolution. This neutral theory contends that most evolutionary change at the molecular level is driven by genetic drift rather than natural selection. Since then, testing the neutral hypothesis has been one of the main objectives of molecular population genetics. The neutral theory has been used as a null model against which specific occurrences of selection may be detected. For Single Nucleotide Polymorphism (SNP) data, one of the most popular tests is Tajima's D statistic (Tajima, 1989). The D statistic is based on the fact that under the standard neutral model, estimates of Watterson's θ_w (based on the number of segregating sites) and of the average pairwise number of nucleotide differences (π) are identical. This test measures the skew in the frequency spectrum; a negative D value indicates an excess of rare polymorphisms and a positive D suggests an excess of intermediate frequency polymorphisms. There are several similar approaches based on slightly different statistics, *e.g.* Fu and Li's D test (1993) and Fay and Wu's H test (2000). An alternative test based on the joint analysis of interspecific divergence and intraspecific polymorphism is the Hudson-Kreitman-Aguadé test (Hudson et al., 1987). This HKA test assesses whether levels of within- and between-population DNA variation are positively correlated, as predicted by the neutral mutation hypothesis.

Plant molecular population genetics approaches

Population geneticists spend most of their time focusing on two aspects: describing the nature of genetic variation of populations (*i.e.* patterns and magnitudes of polymorphisms as well as their frequency distribution in populations), and exploring the evolutionary forces acting on populations. Generally, patterns of genetic diversity within and among populations are influenced both by evolutionary processes that affect the entire genome, such as demographic history and population structure, and by processes that act at individual genes such as natural selection. A multilocus approach is a powerful way to disentangle the effects of different evolutionary forces on DNA variation. This approach has been used for several well-studied plant species, *e.g.* species of *Arabidopsis* (Wright et al., 2003; Ramos-Onsins et al., 2004; Schmid et al., 2005; Nordborg et al., 2005), and maize (Tenaillon et al., 2004; Wright et al., 2005).

The era of empirical molecular population genetics in plants, based on sequence data, began less than 20 years ago. Since the first publication of sequence diversity in plants (Shattuck-Eidens et al., 1990), many additional comprehensive studies of plant nucleotide diversity have been published *e.g.* (Gaut and Clegg, 1993; Miyashita et al., 1996; Purugganan and Suddith, 1998; Savolainen et al., 2000; Tiffin and Gaut, 2001; Olsen and Purugganan, 2002; Morrell et al., 2003; Wright et al., 2005; Liu and Burke, 2006). Most of these plant population genetics studies have been focused on detecting the signature of positive selection and/or examining demographic history. However, demographic processes have been poorly addressed in plant studies, in part because an assessment of demography requires large multilocus data sets (Wright and Gaut, 2005).

Moreover, very few empirical plant population genetics studies have analyzed and compared sequence diversity among local populations, as most plant studies are based on 'species-wide' samples that use single individuals from many locations (Wright and Gaut, 2005). It is important to study the magnitude and consequences of population structure, as well as to include explicit sampling of local populations. There are two main reasons for this, (i) 'real' population sampling will provide insights into demographic factors, which will facilitate understanding of the evolutionary process as well as the design of association studies, (ii) sampling of local populations may provide additional insights into the nature and strength of selection. For example in chapter one, analyzing a few genuine population samples enabled us to discover and interpret the clinal pattern of variability at locus CT208 as signatures of an ongoing selective sweep in *S. chilense*. Sampling only a single population or, alternatively, single individuals from many locations across the species range, might have entirely missed this signature.

Population structure

Populations of organisms are often substructured. Therefore, the issue of population structure have recieved much attention in population genetics studies. In the presence of population structure, several factors including levels of gene flow among populations and the number of demes are expected to contribute to levels of nucleotide diversity within and among populations, and consequently influence species-wide levels of variation (Whitlock and Barton, 1997; Wakeley and Aliacar, 2001; Laporte and Charlesworth, 2002). There is considerable evidence that population structure shapes

patterns of genetic variation in many plant species, *e.g.* in *A. thaliana* (Sharbel et al., 2000; Schmid et al., 2006), *A. lyrata* (Wright et al., 2003; Clauss and Mitchell-Olds, 2006), *Populus tremula* (Ingvarsson, 2005), and *Silene tatarica* (Tero et al., 2003).

F_{st} is an estimate of population differentiation measuring the differentiation of subpopulations relative to the total population. F_{st} can be used as a statistic to summarized patterns of differentation between populations. In chapter one, we show that both insect-pollinated wild tomatoes exhibit moderate levels of F_{st} estimates, which are broadly comparable to the estimates in outcrossing plant species (based on both allozyme and nucleotide data). The estimates of F_{st} in both studied taxa may seem surprisingly low in view of the high fragmentation of local populations. It is likely that patterns of population differentiation in our samples reflect the presence of soil seed banks, and historical association mediated by climatic cycles.

Linkage Disequilibrium (LD)

LD is the nonrandom association of alleles/SNPs at different loci. The terms linkage and LD are often confused. Linkage refers to the correlated inheritance of loci through the physical connection on a chromosome, whereas LD refers to the correlation between polymorphisms (*e.g.* SNPs) that is caused mainly by the history of mutation and recombination. While there are a variety of statistics to measure LD, the two most common LD statistics are r^2 and $D^{'}$. Both methods reflect different aspects of LD and perform differently under various conditions. Generally, r^2 summarizes both recombinational and mutational history, whereas $D^{'}$ measures only recombinational history (reviewed by Flint-Garcia et al., 2003). Two methods are widely used to visualize the extent of LD between polymorphic sites; the first is a scatter plot of r^2 values versus physical distance, which is effective for visualizing the rate at which LD declines (as used in chapter one). Alternatively, LD matrices are used to visualize the linear arrangement of LD between polymorphic sites.

LD plays an important role in association analysis, which recently emerged as a powerful method to identify Quantitative Trait Loci (QTL) in plants. Therefore, a detailed understanding of the extent and patterns of LD within a given target species will facilitate the choice of appropriate methodology for association mapping. In chapter one, we show that LD decays rapidly in both wild tomato species, reflecting high rates of recombination as well as high effective population sizes. Moreover, the fast decay of LD looks very promising for association mapping of functional variation

in wild tomatoes.

Speciation

Speciation is the evolutionary process by which new biological species arise. One of many ways to 'classify' modes of speciation is based on the extent to which populations are geographically isolated from one another. Four types of this speciation are; allopatric (due to geographic isolation), peripatric (due to mostly geographic isolation), parapatric (due to little geographic isolation), and sympatric speciation (due to non-geographic isolation) (Mayr, 1942; Ridley, 2003). In chapter two, we have assessed the speciation history in wild tomatoes, in particular testing the following methods.

Isolation speciation model

Wakeley and Hey (1997) provide a simple model of allopatric speciation which makes use of the genealogical information contained in sequence data. The model assumes that an ancestral population splits into two descendent species at some point in the past, and that there was no gene flow between the diverging species (Figure 5). This model assumes that effective population size is constant within species, but can change at the time of isolation. The input data are the counts of four mutually exclusive types of polymorphisms (Figure 6). We need sequence data from several independent loci in order to reduce genealogical correlations among sites. The parameter estimates should yield information about demographic and temporal aspects of speciation event in the tomato genus.

LD test of historical gene flow

Machado et al. (2002) introduced a test of gene flow based on patterns of LD among specific types of segregating sites, *i.e.* using a subset of total intragenic LD (Figure 7; see chapter two for details). The LD test of historical gene flow (gene flow that occurred after initial species divergence) among the diverging species may allow us to contrast pairs of populations where both are from outside the other species' range (allopatric comparison), and pairs of populations from regions of sympatry (see Städler et al., 2005). If historical gene flow in fact characterized the divergence between *S. chilense* and *S. peruvianum*, we would expect to see more evidence of it (based on the LD test statistic) in sympatric than in allopatric comparisons, unless

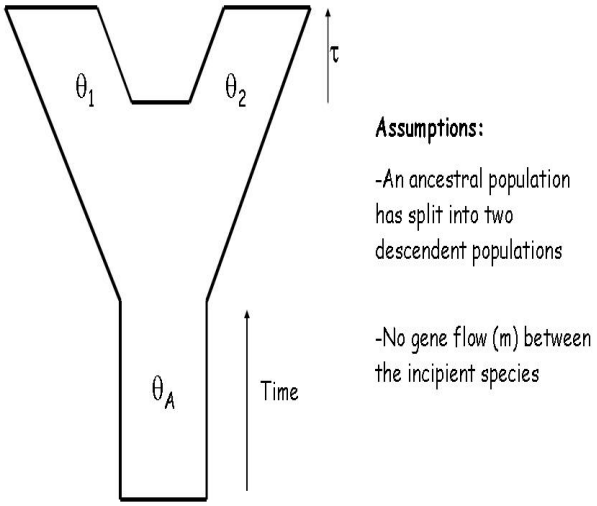

Figure 5: The isolation model of allopatric speciation assumes that an ancestral population splits into two descendent species at some point in the past, and that there was no gene flow between the diverging species. The model yields estimates of the four parameters θ_A, θ_1, θ_2 and τ, where θ_A, θ_1, θ_2 denote the population mutation parameters of the ancestral species and two extant species, respectively. The time since the species split is τ which is equal to 2ut, where u denotes the mutation rate and t is the number of generations since speciation (Wakeley and Hey, 1997).

high within-species gene flow has homogenized haplotypes (and thus patterns of LD among SNPs) in both species after their genetic differentiation. Evidence of historical gene flow would imply that the speciation history of wild tomatoes did not proceed in a strictly allopatric fashion, and natural selection would have to be invoked as one of the forces underlying historical species divergence. In addition, the earlier study of Städler et al. (2005) based on single populations found limited evidence for historical introgression from *S. chilense* into *S. peruvianum*. In the present study, we therefore predict that using more samples per species will yield more power for testing of historical gene flow.

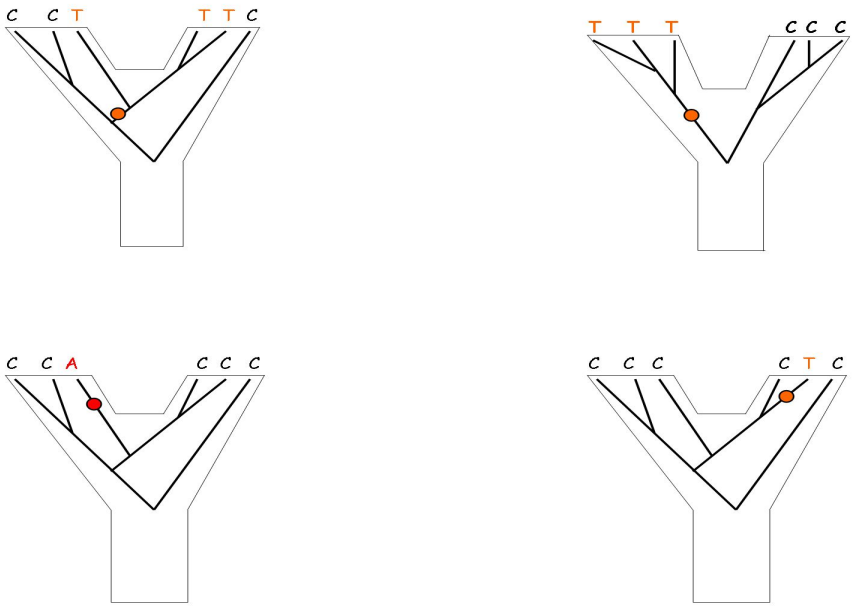

Figure 6: Hypothetical gene genealogies with examples of four specific classes of segregating sites. The upper left diagram presents shared polymorphic site; upper right, fixed difference between the species; lower left, exclusive polymorphism in species 1; and lower right, exclusive polymorphism in species 2. The red dots indicate mutation events.

Scope of the thesis

In chapter one, a multilocus approach was used to examine the patterns and magnitudes of population structure, demography and natural selection in two closely related wild tomato species, using sequence data for eight unlinked nuclear loci from populations across much of the species' range. We address the following basic questions:

1) What are the levels and patterns of nucleotide diversity in wild tomatoes?
2) Is there any evidence for positive selection in the studied species?
3) What are the characteristic levels of population structure in wild tomatoes?
4) How fast is the decline of LD with physical distance in these wild tomatoes?

In comparison to the previous studies in wild tomatoes (Baudry et al., 2001; Städler et al., 2005; Roselius et al., 2005), which were based on only single populations per species, this study allows for more generality by adding three additional populations, broadly covering most of the species' range. Our assessment of population structure allowed us to understand several patterns of variability in wild

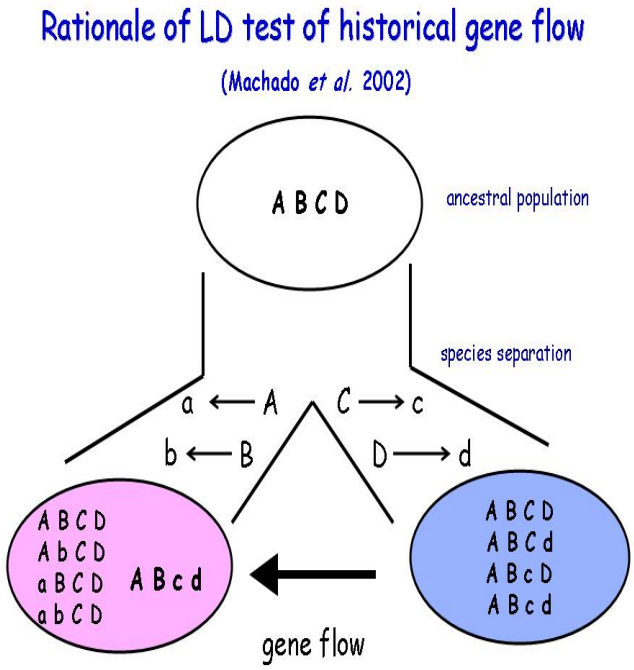

Figure 7: Rationale of the LD test of historical gene flow (Machado et al., 2002). Each letter represents a SNP (Capital letters represent the ancestral state, and small letters the derived state). Each population experiences mutation and recombination events so that multiple haplotypes are generated in each population. Assuming there is gene flow between the diverging species, LD among pairs of shared polymorphisms in the recipient species should be positive (*e.g.* C-D and/or c-d should be overrepresented), and LD among pairs of sites where one member is a shared and the other an exclusive polymorphism should be negative (*e.g.* A-d and/or B-c should be overrepresented).

tomatoes. One of our most interesting findings is that Tajima's D statistic is higher in the samples of the individual populations than in the pooled sample, where the pooled samples show a significant excess of low-frequency polymorphisms. We argue that this may also be a consequence of ancestral population structure (see Discussion). We thus propose here that population structure is one of the most important evolutionary forces to shape patterns of nucleotide diversity in both species. Moreover, the rapid decline of LD with physical distance seems promising for future association studies with the purpose of mapping functional variation in wild tomatoes.

In chapter two, we investigate speciation processes in the two closely related wild tomato species, applying the analytical framework of divergence population genetics. More specifically, we test the suitability of the isolation speciation model (divergence without gene flow) for the *S. peruvianum* – *S. chilense* divergence, using multilocus sequence data from seven 'neutral' nuclear loci. We also test for historical gene flow by analyzing patterns of intragenic LD, and compare between sympatric and allopatric population pairs. In this study, we cannot reject the isolation model based on overall goodness-of-fit criteria, however, patterns of LD are indicative of historical gene flow between *S. peruvianum* and *S. chilense*.

List of Abbreviations

bp	Base pair
D_{FL}	Fu and Li's D test
D_{ss}	LD among pairs of shared polymorphisms
D_{sx}	LD among pairs of sites where one member is a shared and the other an exclusive polymorphism
D_T	Tajima's D test
F_s	Fu's F test
H	Fay and Wu's H test
kb	Kilobases
LD	Linkage disequilibrium
n	Number of sample sizes
N_e	Effective population size
r	Physical recombination rate
r^2	Correlation coefficient for a pair of biallelic sites
SNPs	Single nucleotide polymorphisms
S	Number of segregating sites
S_f	Number of fixed differences
S_s	Number of shared polymorphisms
S_{x1}	Number of exclusive polymorphisms in species 1
S_{x2}	Number of exclusive polymorphisms in species 2
WH	Wakeley and Hey Isolation model
μ	Mutation rate
θ	Nucleotide diversity based on number of segregating sites
π	Nucleotide diversity based on the average pairwise differences between sequences
ρ	Population recombination rate (Hudson, 2001)
γ	Population recombination rate (Hey, 1997)

Chapter 1

Using multilocus data to assess population structure, natural selection and linkage disequilibrium in wild tomatoes

1 Introduction

Understanding the evolutionary forces that shape patterns of nucleotide diversity within and among populations and recently diverged species is an important aim of population genetics. Generally, patterns of genetic diversity within and among populations are influenced both by evolutionary processes that affect the entire genome, such as demographic history and population structure, and by processes that act at individual genes such as natural selection. A multilocus approach is a powerful way to disentangle the effects of different evolutionary forces on DNA variation. This approach has been used for several well-studied species of *Drosophila*, *e.g.* (Glinka et al., 2003; Das et al., 2004; Ometto et al., 2005), *Arabidopsis* (Wright et al., 2003; Ramos-Onsins et al., 2004; Schmid et al., 2005; Nordborg et al., 2005), and maize (Tenaillon et al., 2004; Wright et al., 2005).

In the presence of population structure, several factors including levels of gene flow among populations and the number of demes are expected to contribute to levels of nucleotide diversity within and among populations, and consequently influence species-wide levels of variation (Whitlock and Barton, 1997; Wakeley and Aliacar, 2001; Laporte and Charlesworth, 2002). There is considerable evidence that population structure shapes patterns of genetic variation in many plant species, *e.g.* in *A. thaliana* (Sharbel et al., 2000; Schmid et al., 2006), *A. lyrata* (Wright et al., 2003;

Clauss and Mitchell-Olds, 2006), *Populus tremula* (Ingvarsson, 2005), *Silene tatarica* (Tero et al., 2003), and *Pinus densata* (Ma et al., 2006). The effects of population structure might also influence the magnitude and pattern of linkage disequilibrium (LD) (*e.g.* Ostrowski et al. (2006)), and the presence of population structure can also lead to the detection of spurious associations in association studies (*e.g.* Helgason et al. (2005)).

In principle, several factors can lead to an increase in LD, for instance population structure, low recombination rate, natural and artificial selection, inbreeding and small effective population size. However, other factors such as outcrossing, high recombination rate and large effective population size may propel a decay of LD (reviewed by Gupta et al. (2005)). Since the availability of genome-wide sequences and/or single nucleotide polymorphism (SNP) maps, LD mapping has been used extensively in animal and plant systems, as well as to dissect the molecular bases of human diseases (Flint-Garcia et al., 2003; Rafalski and Morgante, 2004). Therefore, a detailed understanding of the extent and patterns of LD within a given target species will facilitate the choice of appropriate methodology for association mapping.

Even though LD has received much attention recently, more studies are needed to investigate patterns of LD as well as factors that influence LD. In plant genomics only a few well-studied species such as *A. thaliana*, *P. tremula*, rice, maize, barley and sunflower have been characterized for the extent and decay of LD with physical distance (Lin et al., 2001; Remington et al., 2001; Tenaillon et al., 2001; Nordborg et al., 2002; Garris et al., 2003; Kraakman et al., 2004; Ingvarsson, 2005; Liu and Burke, 2006).

Wild tomatoes (*Solanum* section *Lycopersicon*) have become a suitable plant model system for evolutionary analyses because of their recent divergence, the clear phenotypic distinction and the great diversity of mating systems. Wild tomatoes are native to western South America, with two endemic species in the Galapagos Islands (Rick, 1986; Taylor, 1986; Spooner et al., 2005; Peralta et al., 2005). Our earlier studies, based on single populations in five species, suggested that nucleotide diversity in wild tomatoes is influenced by mating system, among-locus variation in neutral mutation rate and/or selective constraints among loci, while evidence for positive selection was scarce (Städler et al., 2005; Roselius et al., 2005). In other words, if positive directional selection does not have a large effect in the tomato clade then the analyses of demographic processes and population structure become more

significant for understanding patterns of genetic diversity and historical events in wild tomatoes.

In this study, we adopt a multilocus approach to examine the effects of population structure and demographic history on nucleotide variability in wild tomatoes. We also characterize the (intragenic) decay of LD using multiple natural populations across much of the species range of the self-incompatible, closely related species (*S. peruvianum* and *S. chilense*). We ask specifically:

1) What are the patterns of nucleotide diversity in wild tomatoes?

2) Do wild tomato populations show genetic differentiation?

3) How fast is the decline of LD with physical distance in these two closely related wild tomato species?

Using polymorphism data for eight unlinked nuclear loci from four natural populations per species, we found substantial levels of nucleotide polymorphism and modest levels of population differentiation in both species. More interestingly, the presence of population structure (as well as sampling design) may have facilitated the ability to discover a clinal pattern of nucleotide variation at CT208, a probable signature of an ongoing selective sweep in *S. chilense*. Furthermore, LD decays very rapidly, reflecting fairly high rates of recombination at all loci as well as high effective population size.

2 Materials and Methods

2.1 Population sampling

Wild tomatoes have been taxonomically reassigned to the genus *Solanum* section *Lycopersicon* (Spooner et al., 1993; Olmstead et al., 1999; Spooner et al., 2005). For this study, we adopted the new nomenclature and chose two self-incompatible tomato species. *S. peruvianum* is distributed along the western side of the Andes from north-central Peru to northern Chile, and *S. chilense* from southern Peru to northern Chile (cf. Fig. 1, Städler et al. (2005)). Both studied wild tomatoes are patchily distributed species, and grow in a wide diversity of habitats from sea level to the highland up to 3,300 meters (Rick, 1986; Taylor, 1986). In their native habitats, it appears that *S. peruvianum* is the most widespread and highly subdivided species (Rick, 1986). Moreover, it shows the greatest level of nucleotide polymorphism, even on a local population scale (Baudry et al., 2001; Städler et al., 2005; Roselius et al., 2005). In order to adequately sample the geographic ranges of the species, three new populations of each species were collected in Peru by T. Städler and T. Marczewski (May, 2004). Voucher specimens have been deposited at USM (Lima, Peru) and MSB (Munich, Germany). In addition, we included one population of each species (Tarapaca for *S. peruvianum* and Antofagasta for *S. chilense*) from an earlier analysis (Städler et al., 2005). The population samples and geographic locations are given in Table 1.1.

Table 1.1: Geographical location of the populations analyzed

Species	Population	Location	Coordinates (latitude, longitude)	Abbreviated population
S. peruvianum	Tarapaca	Northern Chile	18°33'S, 70°09'W	TAR
	Arequipa	Southern Peru	16°27'S, 71°42'W	ARE
	Nazca	Southern Peru	14°51'S, 74°44'W	NAZ
	Canta	Central Peru	11°32'S, 76°42'W	CAN
S. chilense	Antofagasta	Northern Chile	22°14'S, 68°23'W	ANT
	Tacna	Southern Peru	17°53'S, 70°08'W	TAC
	Moquegua	Southern Peru	17°04'S, 70°52'W	MOQ
	Quicacha	Southern Peru	15°38'S, 73°48'W	QUI

2.2 Loci studied

For each population, we analyzed at least five individuals (10 alleles) of eight unlinked nuclear loci (CT093, CT208, CT251, CT066, CT166, CT179, CT198 and CT268) that represent a subset of those studied earlier (Baudry et al., 2001; Städler et al., 2005; Roselius et al., 2005). The loci encompass regions of low to high recombination as estimated by Stephan and Langley (1998). The characterization of our loci are showned in Table 1.2.

Table 1.2: Characterization of eight unlinked analyzed loci

Locus	Chromosome	Putative encoded protein	R_N [a]
CT093	V	S-adenosylmethionine decarboxylase proenzyme	0
CT208	IX	Alcohol dehydrogenase class III	0
CT251	II	At5g37260gene	0.46
CT066	X	Arginine decarboxylase	0.93
CT166	II	Ferredoxin-NADP reductase	1.61
CT179	III	Tonoplast instrinsic protein	1.97
CT198	IX	Submergence induced protein 2-like (SIP)	2.10
CT268	I	Receptor-like protein kinase	2.33

[a]Recombination rate x 10^{-8} ,Stephan and Langley (1998)

2.3 DNA amplification and sequencing

We isolated genomic DNA from dried leaves of mature plants using the DNeasy Plant Mini Kit (Qiagen GmbH, Hilden, Germany). Polymerase Chain Reaction (PCR) primers were designed based on the published cDNA or genomic DNA sequences from *S. lycopersicum*, which are available from the Tomato Gene Index at The Institute for Genomic Research (TIGR; http://www.tigr.org/tdb/lgi/). PCR primers and conditions are deposited at http://www.zi.biologie.uni-muenchen.de/evol/. PCR products were sequenced directly from both strands on an ABI3730 DNA analyzer (Applied Biosystems, Foster city, CA). Direct sequencing was also used to confirm polymorphic sites in heterozygotes. Since it is essential to resolve haplotypes, we designed haplotype-specific sequencing primers based on heterozygous nucleotides or indels as previously described (Städler et al., 2005). Briefly, we exploited putative

or confirmed SNPs to anchor the 3′-end of sequencing primers that were intended to resolve the heterogeneous PCR products. This approach enabled us to verify SNPs (and indel variation) and establish haplotype phase based on overlapping information supported by multiple primer pairs. In this study, haplotype phase was thus completely resolved for all new sequences. Sequences were edited and aligned using the Sequencher program (Gene Codes, Ann Arbor, MI) and adjusted manually afterwards. Locus CT208 was resequenced in the TAR population, revealing eight alleles (instead of 10 as in a previous study; Baudry et al. (2001)). Moreover, we designed new primers for CT166 and CT208, for which shorter PCR products were sequenced than previously (Baudry et al., 2001; Roselius et al., 2005).

2.4 Estimation of nucleotide diversity and neutrality tests

We estimated levels of nucleotide diversity for all sites and silent sites (using non-coding and synonymous sites), calculating Watterson's estimator (Watterson, 1975)($\theta_w = 4N_e\mu$ where N_e denotes the effective population size and μ the mutation rate per site and generation) and π the average number of pairwise differences between sequences in a population (Nei, 1987). We tested the deviation from neutrality by using Tajima's D statistic (Tajima, 1989). This test is based on the fact that under the standard neutral model, estimates of θ_w (based on the number of segregating sites) and of the average number of nucleotide differences (π) are identical. The test measures the skew in the frequency spectrum; a negative D value indicates an excess of rare polymorphisms and a positive D suggests an excess of intermediate frequency polymorphisms. Tajima's test is conservative for testing departures from neutral equilibrium conditions, in particular under the assumption of no recombination. We also employed Fu and Li's D (Fu and Li, 1993) and Fay and Wu's H (Fay and Wu, 2000) statistics. The H statistic measures the differences between the average number of nucleotide differences and the estimator θ_H, which is based on the frequency of derived variants. A significantly negative H value indicates an excess of high-frequency derived variants, which may be indicative of positive selection. The significance of H was evaluated by 10,000 coalescent simulations, using the observed number of segregating sites and no recombination. All standard analyses were performed in DnaSP version 4.0 (Rozas et al., 2003).

In addition, we used the multilocus HKA statistics (Hudson et al., 1987) to assess the ratio of polymorphism within species to the divergence between species, as implemented in the program HKA (http://lifesci.rutgers.edu/ heylab/). Significance of Tajima's D and Fu and Li's D statistics were also assessed using 10,000 coalescent simulations in HKA program. In the tests mentioned above, we used sequences from *S. ochranthum* or *S. lycopersicoides* from Roselius et al. (2005) as outgroup species, except for CT208 for which we obtained a new sequence from *S. ochranthum*.

2.5 Tests of population differentiation

We calculated the F_{st} statistics using the Analysis of Molecular Variance approach (AMOVA), as implemented in Arlequin 3.1 (Excoffier et al., 2005), to quantify population differentiation between all pairwise comparisons within species. F_{st} is an estimate of population differentiation measuring the differentiation of subpopulations relative to the total population. Its significance was assessed using permutation tests (10,000 permutations). We also estimated F_{st} values based on the method of Hudson et al. (1992) using DnaSP program. Both methods yielded comparable results (data not shown).

2.6 Analyses of recombination and intragenic LD

We estimated the minimum number of recombination events (R_m) using the four-gamete test of Hudson and Kaplan (1985) and the population recombination parameter (ρ), where $\rho = 4N_e$c and c the recombination fraction between sites, using Hudson's composite likelihood method (Hudson, 2001). This method is based on pairwise LD between sites, as implemented in the LDhat 2.0 package (McVean et al., 2002). Moreover, we calculated the degree of LD in terms of the Z_{ns} statistic (Kelly, 1997), which is the average of squared allele-frequency correlations (r^2) (Hill and Robertson, 1968) over all pairwise comparisons.

We investigated the decay of LD over physical pairwise distance following the methods of Remington et al. (2001). The estimates of LD were calculated by using (r^2) between pairs of polymorphic sites. The expected decay of LD was modeled as:

$$E(r^2) = \left[\frac{10+\rho}{(2+\rho)(11+\rho)}\right]\left[1+\frac{(3+\rho)(12+12\rho+\rho^2)}{n(2+\rho)(11+\rho)}\right]$$

where n denotes the number of sequences (Hill and Weir, 1988). We fitted this equation to the data using the *R* statistical package (http://www.r-project.org/). The nonlinear regression yields a least-squares estimate of ρ per base pair; this estimate may not be precise and unrealistic due to several factors, *e.g.* the nonindependence between linked sites and nonequilibrium populations. Nonetheless, this model is still useful to characterize the rate of decay of LD, *e.g.* (Remington et al., 2001; Brown et al., 2004; Ingvarsson, 2005; Liu and Burke, 2006). These analyses were done for each locus separately, both within populations and for the combined data set, and for all eight loci together. Observed sites with multiple hits were excluded in the recombination and LD analyses, and all singletons were removed in the LD analyses.

3 Results

3.1 Patterns of nucleotide diversity and tests of neutrality

We sequenced eight unlinked nuclear loci in three populations of each species, with a total concatenated length of > 10 kb per 'allele'. We also added one population of each species from an earlier study (Städler et al., 2005). Therefore, in this study we evaluated in total four populations and about 40-46 alleles per locus per species. The total length (including indels) of the individual loci ranges from 778 bp to 1887 bp. Most of the loci contained both coding and noncoding sites (introns and/or flanking regions), except for CT066 and CT268 which contribute only coding sites (Table 1.3).

Table 1.3: Length of eight loci analyzed

				Coding	
Locus	Chromosome	Total[a]	Noncoding	Synonymous	Nonsynonymous
CT093	V	1389	359	248.3	780.7
CT208	IX	1069	621	107.8	339.2
CT251	II	1672	348	318.0	1005.0
CT066	X	1346	0	331.8	1012.2
CT166	II	1265	823	91.1	349.9
CT179	III	899	318	153.1	425.9
CT198	IX	693	359	76.2	256.8
CT268	I	1881	0	435.1	1445.9

[a]excluding sites with gaps from the total alignment, based on *S. peruvianum* (n = 40-44 alleles)

We quantified polymorphism by using Watterson's θ_w and π for all sites (θ_{all}, π_{all}) and for silent sites (θ_{sil}, π_{sil}). For each locus and each population, the estimates of nucleotide diversity are given in Table 1.4. Across individual loci, the levels of variation (θ_{all}) varies from 0.15 - 2.92% in *S. peruvianum* and 0.04 -1.86% within *S. chilense*. CT198 appears to be the most polymorphic locus, whereas CT093 shows the least polymorphism in both species.

We also examined the weighted average levels of nucleotide diversity across all eight loci for each population, as presented in Table 1.4. Most of the loci and populations generally exhibit substantial levels of polymorphism. In *S. peruvianum*, we found that the northernmost population (CAN) shows the highest level of variability which is a bit higher than that of the southernmost population (TAR), whereas the ARE population exhibits the lowest level of polymorphism. In *S. chilense*, the three new populations (TAC, MOQ and QUI) display comparable levels of sequence diversity, which is twice as higher as in the ANT population.

We used Tajima's D and Fu and Li's D statistics to test for deviations from neutrality. Based on these statistics, most of our populations do not show significant departures from neutral expectations. The only two exceptions are the ARE and ANT populations. The ARE population shows significant deviations from the neutral model at four loci, where three loci (CT093, CT179 and CT268) exhibit positive Tajima's D values, while only CT066 exhibits a significantly negative value (Table 1.5 and 1.6). The ANT population also shows significantly positive Tajima's D values at several loci as reported in our previous study (Roselius et al., 2005). The multilocus neutrality tests, based on all sites, are reported in Table 1.4. Multilocus Tajima's D values are slightly and consistently negative in *S. peruvianum* (except for the ARE population), whereas in *S. chilense* the Tajima's D values are close to zero (except for the ANT population which shows a very significantly positive value). Moreover, we found that the Fu and Li's D statistic exhibits similar patterns as Tajima's D. However, the estimates of Fu and Li's D indicate a departure from neutrality for the CAN population (D_{FL} = -0.91, P = 0.02), as well as for the ANT population in *S. chilense* (D_{FL} = 1.25, $P < 0.001$). This negative D_{FL} value suggests an excess of polymorphisms on external branches of the genealogy (*i.e.*, singletons) of the CAN sample, whereas the positive value for the ANT population indicates an excess of polymorphisms on internal branches (*i.e.*, intermediate-frequency variants).

In addition, we averaged the levels of silent polymorphism across all populations in *S. peruvianum*(θ_{sil} = 2.18% , π_{sil} = 2.04%), which is about 1.3- to 1.5-fold higher than in *S. chilense* (θ_{sil} = 1.50% , π_{sil} = 1.59%). Concordant patterns are found for the estimates using all sites. We also calculated levels of nucleotide variation using the combined sample (treated as a single population) in both species (Table 1.4). We found an interesting pattern, in that the θ estimates are consistently higher than the mean θ's for individual populations, whereas the π estimates are less affected by

Table 1.4: Summary of nucleotide polymorphism and multilocus neutrality tests

Population	θ_{sil} [a]	π_{sil} [a]	θ_{all} [a]	π_{all} [a]	D_T [b]	D_{FL} [b]	HKA [b] (P-valuves)
S. peruvianum							
TAR	2.38	2.26	1.26	1.20	−0.22	−0.49	0.96
ARE	1.39	1.45	0.72	0.78	0.27	0.28	0.67
NAZ	2.17	2.04	1.17	1.13	−0.19	−0.40	0.87
CAN	2.79	2.39	1.44	1.28	−0.54	−0.91*	0.98
All [c]	3.44	2.40	1.81	1.29	−1.06***	−1.69***	0.96
S. chilense							
ANT	0.83	1.06	0.43	0.55	1.44***	1.25***	0.02*
TAC	1.64	1.63	0.93	0.92	−0.09	−0.05	0.99
MOQ	1.86	1.82	1.03	1.01	−0.28	0.04	0.97
QUI	1.66	1.81	0.92	1.01	0.27	0.43	0.98
All [c]	2.25	2.01	1.27	1.10	−0.53	−1.11**	0.56

*P<0.05, **P<0.01, ***P<0.001

[a] weighted average across eight loci, presented in percentage per site
[b] Tajima's D(1989), Fu and Li's D (1993) and HKA (1987)
[c] analyses of the combined sampled (treated as a single population)

pooling the samples. Hence, we obtained a highly significantly negative multilocus Tajima's D for the combined sample of *S. peruvianum* sequences (D_T = -1.06, $P <$ 0.001). In *S. chilense*, however, this effect is less pronounced. Using the combined sample, Tajima's D is negative but only marginally significant (D_T = -0.55, P = 0.07). Similarly, for Fu and Li's D statistic, both species exhibit significantly negative values when calculated using the combined sample, suggesting an excess of singletons in the samples. We will discuss this result below. It appears to be a consequence of the underlying population structure of the two tomato species *S. peruvianum* and *chilense*, but is difficult to explain by the standard models of population subdivision (such as the island model).

Furthermore, we applied Fay and Wu's H test using *S. ochranthum* or *S. lycopersicoides* as outgroups (based upon availability; Roselius et al. (2005)). At locus CT208, two *S. chilense* populations (TAC and MOQ) show significantly negative H

values (H = -15.12, P = 0.02 for TAC, and H = -21.33, $P < 0.001$ for MOQ) that indicate an excess of high-frequency derived variants (Fay and Wu, 2000). Figure 1.1 summarizes patterns of nucleotide variation at locus CT208 across the *S. chilense* populations. The northernmost QUI population exhibits more SNPs that belong to the ancestral varaint group than SNPs of the MOQ and TAC populations. We identified the alleles which are similar to the outgroup as the 'ancestral' haplotype group and the other alleles as the 'derived' haplotype group. We found that the MOQ population shows a very high frequency of derived variants, in that only one allele belongs to the 'ancestral' haplotype group. Likewise, for the TAC population only two alleles are of the 'ancestral' group. In addition, the 'derived' group has fixed in the southernmost ANT population, which is clearly indicative of a selective sweep. Furthermore, in *S. peruvianum* a significantly negative H value was found at locus CT066 for the ARE population (Table 1.5).

Additionally, the multilocus HKA test (Hudson et al., 1987) was used to assess departures from the neutral model. We observed no evidence for any significant deviations in *S. peruvianum*. In *S. chilense*, none of the new populations show significant deviations from the neutral model, whereas the 'old' ANT population clearly departs from the neutral model (P = 0.02), as shown in Table 1.4. A statistically significant HKA test was also found in the previous study, which was based on 14 loci (Roselius et al., 2005).

Due to the relatively few polymorphisms and/or haplotype structure for the ARE population (*S. peruvianum*) at the majority of loci, and particularly for the ANT population (*S. chilense*) (Roselius et al., 2005), these samples may therefore be regarded as 'outliers'. Because of those patterns in both populations (which do not appear to be 'typical' species-wide patterns), we excluded both populations from further analyses of population differentiation, recombination and intragenic linkage disequilibrium. Note that if we exclude the ARE and ANT populations, we obtain higher average levels of diversity in both species. Moreover, average Tajima's D values become more negative in *S. peruvianum* (-0.31) and are approximately zero in *S. chilense* (-0.03).

3.2 Levels of population differentiation

Table 1.7 shows estimates of F_{st} and the permutation tests of population differentiation across three populations for each species at eight loci, using an AMOVA

Table 1.5: Summary of nucleotide polymorphism in *S. peruvianum*

Locus	Population	Length[a]	N[b]	S[c]	θ_{all}[d]	π_{all}[d]	D_{all}[e]	FW-H[f]
CT093	TAR	1390	10	23	0.59	0.57	-0.14	3.02
	ARE	1393	10	9	0.23	0.32	1.68*	-0.80
	NAZ	1392	12	24	0.57	0.49	-0.60	-3.94
	CAN	1393	12	31	0.74	0.57	-1.00	2.45
	All	1389	44	61	1.01	0.53	−1.69*	-1.40
CT208	TAR	1087	8	41	1.46	1.39	-0.23	3.86
	ARE	1090	10	24	0.78	0.65	-0.77	3.56
	NAZ	1087	12	47	1.43	1.26	-0.54	4.39
	CAN	1085	12	43	1.31	1.17	-0.51	0.76
	All	1069	42	89	1.94	1.32	-1.15	-7.72
CT251	TAR	1678	10	70	1.48	1.43	-0.14	3.56
	ARE	1713	10	37	0.76	0.82	0.37	0.89
	NAZ	1701	12	55	1.07	1.14	0.28	3.00
	CAN	1702	10	70	1.45	1.35	-0.35	6.40
	All	1672	42	132	1.84	1.37	-0.92	1.86
CT066	TAR	1346	10	40	1.05	0.98	-0.31	2.49
	ARE	1346	10	10	0.26	0.15	−1.92**	−7.11**
	NAZ	1346	12	25	0.62	0.71	0.28	-4.45
	CAN	1346	12	43	1.06	0.84	-0.93	2.64
	All	1346	44	66	1.13	0.95	-0.56	-1.88
CT166	TAR	1298	8	42	1.25	1.13	-0.51	5.07
	ARE	1330	10	30	0.80	0.78	-0.11	2.93
	NAZ	1299	12	45	1.15	0.88	-1.07	2.09
	CAN	1316	12	75	1.89	1.58	-0.75	9.27
	All	1265	42	121	2.22	1.20	−1.68*	6.41
CT179	TAR	958	10	29	1.07	1.07	0.00	-7.47
	ARE	915	9	34	1.37	1.79	1.54*	-5.97
	NAZ	915	12	49	1.77	1.62	-0.37	-1.03
	CAN	911	12	56	2.04	1.86	-0.40	0.70
	All	1389	43	99	2.55	1.93	-0.88	-3.88
CT198	TAR	760	10	62	2.88	2.92	0.07	8.00
	ARE	705	10	32	1.60	1.58	-0.08	-2.58
	NAZ	770	10	50	2.29	2.27	-0.05	2.84
	CAN	757	10	57	2.66	2.63	-0.05	8.44
	All	693	40	97	3.29	2.57	-0.78	4.34
CT268	TAR	1884	10	56	1.05	0.94	-0.51	1.96
	ARE	1884	10	34	0.64	0.82	1.41*	-2.93
	NAZ	1881	12	70	1.23	1.27	0.15	7.18
	CAN	1884	12	68	1.19	1.10	-0.35	6.58
	All	1881	44	132	1.61	1.26	-0.80	3.40

All = Analyses of combined sample (treated as a single population); $^{*}P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$; NA= Not Available

[a] excluding indels
[b] Number of alleles sequenced
[c] Segregating sites
[d] in percentage per site
[e] Tajima's D(1989)
[f] Fay & Wu's H(2000), using *S. ochranthum* or *S. lycopersicoides* as an outgroup species

Table 1.6: Summary of nucleotide polymorphism in *S. chilense*

Locus	Population	Length[a]	N[b]	S[c]	θ_{all}[d]	π_{all}[d]	D_{all}[e]	FW-H[f]
CT093	ANT	1393	10	8	0.20	0.27	1.38	-0.53
	TAC	1393	12	18	0.43	0.49	0.61	-0.24
	MOQ	1393	10	25	0.63	0.53	-0.79	2.13
	QUI	1393	13	17	0.39	0.27	-1.38	-4.45
	All	1393	45	45	0.74	0.51	-1.07	-0.19
CT208	ANT	1111	10	1	0.03	0.04	0.82	NA
	TAC	1094	12	31	0.94	0.75	-0.92	-15.12^{*}
	MOQ	1077	10	29	0.95	0.57	-1.93^{**}	-21.33^{***}
	QUI	1078	14	35	1.02	1.31	1.22	5.54
	All	824	46	45	1.24	1.04	-0.57	-5.06
CT251	ANT	1726	10	6	0.12	0.16	1.32	-1.07
	TAC	1716	12	66	1.27	1.26	-0.02	-5.45
	MOQ	1716	10	52	1.07	1.16	0.40	-4.53
	QUI	1693	14	72	1.34	1.59	0.86	-8.09
	All	1690	46	109	1.47	1.37	-0.25	-10.89
CT066	ANT	1346	10	18	0.47	0.74	2.66	0.00
	TAC	1346	12	33	0.81	0.90	0.50	3.73
	MOQ	1346	10	37	0.97	1.00	0.14	4.00
	QUI	1346	12	33	0.81	0.89	0.46	2.82
	All	1346	44	64	1.09	0.89	-0.66	4.25
CT166	ANT	1307	10	33	0.89	0.88	-0.08	-2.40
	TAC	1323	12	44	1.10	1.09	-0.04	4.79
	MOQ	1317	10	58	1.56	1.64	0.26	8.80
	QUI	1314	12	28	0.70	0.71	0.03	-6.73
	All	1279	44	79	1.42	1.38	-0.10	5.12
CT179	ANT	916	10	23	0.89	1.19	1.62^{*}	0.27
	TAC	919	10	31	1.19	1.30	0.43	-6.40
	MOQ	919	10	38	1.46	1.51	0.17	-4.27
	QUI	915	10	27	1.04	1.04	-0.01	2.49
	All	907	40	75	2.05	1.56	-0.71	-6.35
CT198	ANT	772	10	8	0.37	0.56	2.23^{**}	0.62
	TAC	772	10	20	0.92	0.65	-1.36	-3.47
	MOQ	772	10	20	0.92	0.83	-0.46	-2.67
	QUI	772	10	37	1.69	1.83	0.48	3.64
	All	772	40	54	1.64	1.15	-1.08	-3.41
CT268	ANT	1884	10	29	0.54	0.72	1.54^{*}	-8.80
	TAC	1884	12	48	0.84	0.86	0.10	-2.42
	MOQ	1884	10	46	0.86	0.85	-0.04	-9.07
	QUI	1884	14	43	0.72	0.80	0.48	-0.04
	All	1884	46	88	1.06	1.08	0.07	-0.86

All = Analyses of combined sample (treated as a single population); $^{*}P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$; NA= Not Available

[a] excluding indels
[b] Number of alleles sequenced
[c] Segregating sites
[d] in percentage per site
[e] Tajima's D(1989)
[f] Fay & Wu's H(2000), using *S. ochranthum* or *S. lycopersicoides* as an outgroup species

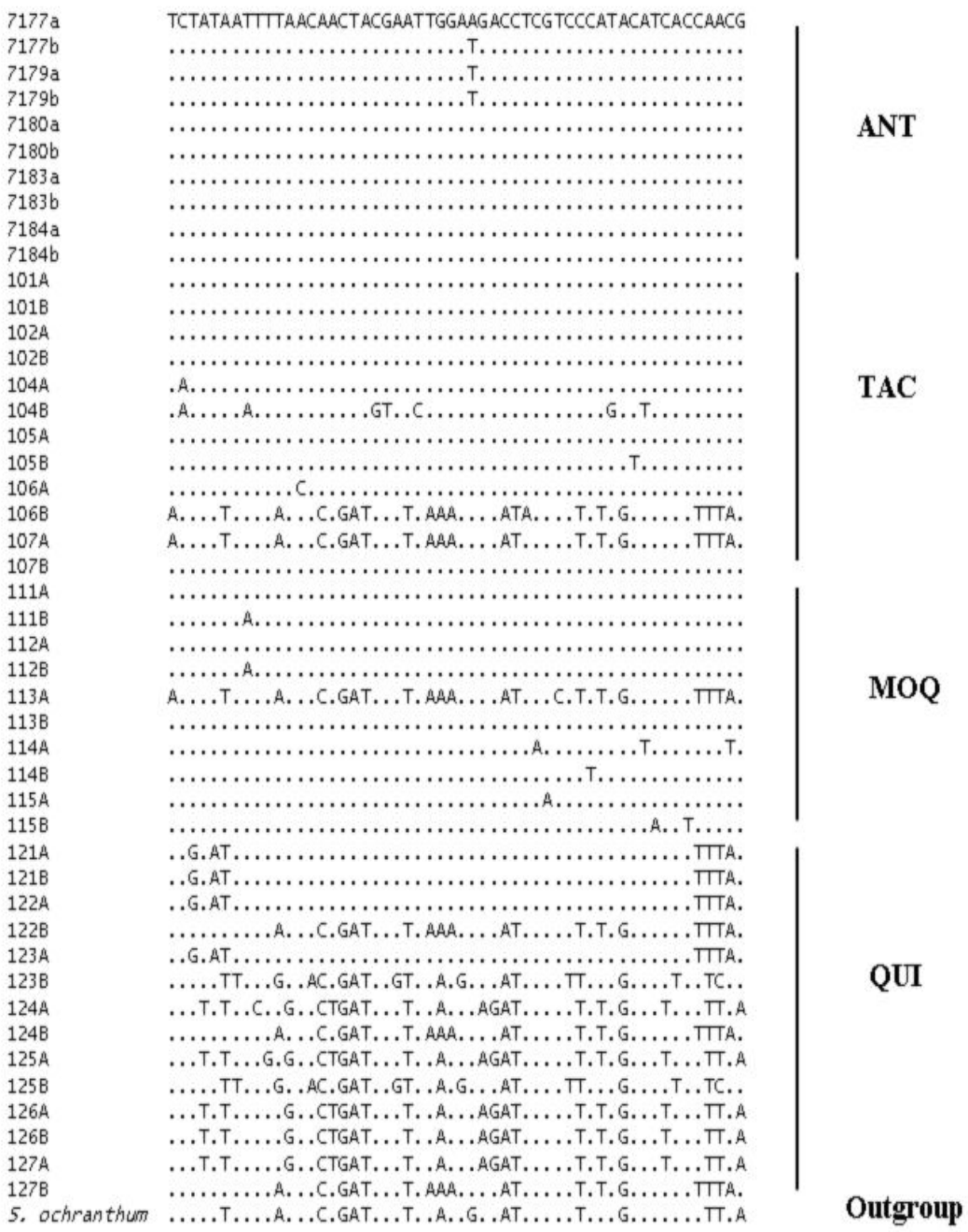

Figure 1.1: Clinal pattern of nucleotide variability (SNPs) at locus CT208 in *S. chilense*. From up to bottom, samples are ordered from the south (ANT) to the north (QUI) of the geographic range. Sample identifiers are given in the left column. The Polymorphic sites shown here are distributed over the entire (sequenced) locus. Nucleotide positions at which the outgroup (*S. ochranthum*) differs from all *S. chilense* sequences were eliminated for visual clarity.

approach (Excoffier et al., 1992). In *S. peruvianum*, we observed significant genetic differentiation at all eight loci, with F_{st} ranging from 0.089 (CT166) to 0.234 (CT066). The average F_{st} estimate is 0.145 for *S. peruvianum*, which is comparable to the value for *S. chilense* ($F_{st} = 0.147$). Six loci show evidence of significant population differentiation in *S. chilense*, whereas two loci (CT251 and CT066) are not significantly differentiated. Notably, locus CT066 exhibits the highest level of population differentiation in *S. peruvianum* but shows the lowest level of differentiation in *S. chilense*.

Table 1.7: Population differentiation at eight loci

	S. peruvianum		*S. chilense*	
Locus	F_{st}[a]	P-value[b]	F_{st}[a]	P-value[b]
CT093	0.131	0.004^{**}	0.225	0.003^{**}
CT208	0.176	$<0.001^{***}$	0.208	0.017^{*}
CT251	0.124	$<0.001^{**}$	0.087	0.087^{ns}
CT066	0.234	$<0.001^{***}$	<0.001	0.399^{ns}
CT166	0.089	0.009^{**}	0.179	0.009^{**}
CT179	0.204	$<0.001^{***}$	0.131	0.012^{*}
CT198	0.101	0.008^{**}	0.210	0.002^{**}
CT268	0.128	$<0.001^{***}$	0.204	$<0.001^{***}$
Average	**0.145**	$<0.001^{***}$	**0.147**	$<0.001^{***}$

^{ns} not significant, $^{*}P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$

[a] F_{st} statistic based on AMOVA approach (Excoffier et al., 1992)
[b] evaluated by permutation tests

In addition, we calculated among-population pairwise estimates of F_{st} across eight loci, as summarized in Table 1.8. Interestingly, despite the greatest geographic distance between them, the northernmost *S. peruvianum* population (CAN) is less differentiated from the southernmost population TAR ($F_{st} = 0.089$) than from the geographically intermediate NAZ population ($F_{st} = 0.134$). In *S. chilense*, the estimate of population differentiation between TAC and MOQ is very low and not significant ($F_{st} = 0.008$). In contrast, the other population pairs (TAC-QUI and MOQ-QUI) show substantial levels of differentiation ($F_{st} \approx 0.21$). Thus, unlike within *S. peruvianum*, there is at least a tendency that levels of population differentiation correlate with geographic distance.

A hierarchical AMOVA approach was used to additionally quantify differentiation between the species. We observed 30.3% of the total variation among species and only 10.1% among populations within species, while 59.6% of the total variation was found within populations. Thus, the fixation index among species is 0.303, and permutation tests are highly significant ($P < 0.001$), suggesting a considerable level of differentiation between *S. peruvianum* and *S. chilense*, which clearly exceeds the level of differentiation between subpopulations within species.

Table 1.8: Average pairwise estimates of F_{st} across eight loci

S. peruvianum				*S. chilense*			
Population	TAR	NAZ	CAN	Population	TAC	MOQ	QUI
TAR	-	0.197^{***}	0.090^{***}	TAC	-	0.008^{ns}	0.218^{***}
NAZ		-	0.134^{***}	MOQ		-	0.209^{***}
CAN			-	QUI			-

^{ns} not significant, $^{***}P< 0.001$ (evaluated by permutation tests)

3.3 Recombination and intragenic LD

Using the four-gamete test to infer the minimum number of recombination events (R_m) in all three populations for each species, we found very diverse estimates across loci, varying from 0 to 13 for individual populations (Table 1.9). These R_m estimates are higher for the analyses of the combined samples in both species (with n = 30-36 alleles per species). Overall, the estimated minimum number of recombination events is largely consistent with the levels of physical recombination rate for each locus, except for CT251; this locus exhibits high Rm estimates in both species, in contrast to its low recombination rate estimated by Stephan and Langley (1998).

We also estimated the population recombination parameter (ρ) at each locus using the composite-likelihood approach of Hudson (2001). We found that ρ ranges from 0 to 0.102 per site in *S. peruvianum* and from 0 to 0.053 in *S. chilense* (Table 1.9). We computed the species-wide weighted average of ρ in *S. peruvianum* across the three populations ($\rho \approx 0.0389$), which is about three-fold higher than in *S. chilense* ($\rho \approx 0.0123$). In order to have more alleles and power for the analysis, we combined two populations per species and treated them as a single sample. Because both the TAR-

CAN (*S. peruvianum*) and TAC-MOQ (*S. chilense*) pairs show relatively low levels of population differentiation, these were chosen for a combined analysis (Table 1.8). For the estimates based on two relatively undifferentiated populations, the weighted average ρ across the eight loci is ≈ 0.0348 in *S. peruvianum*, while it increases to ≈ 0.0238 in *S. chilense* (data not shown). The ratio of these ρ estimates is about 1.5, which is close to the ratio of the θ estimates between both species (see Table 1.4).

Furthermore, we estimated the degree of LD using the Z_{ns} statistic (Table 1.9). Overall, the Z_{ns} values mirror the ρ estimates in that low recombination rates correspond to high levels of LD. Figure 1.2 illustrates the decline of LD with physical distance, using pooled data of all eight loci for the nonlinear regression model. The expected value of (r^2) decays to negligible levels (*i.e.*, < 0.05) within < 150 base pairs for the combined sample in *S. peruvianum* and within < 750 base pairs in *S. chilense*, while LD extends about two to four times larger distances within individual populations. It should be noted that the extended LD within individual populations is possibly due to the smaller sample size and/or low polymorphism levels within populations. Moreover, we present the decay of LD with physical distance for each locus separately, as shown in Figures 1.3 and 1.4. In general, the decay of LD is relatively fast at all loci in both species, whereas CT208 in *S. chilense* exhibits somewhat higher levels of LD than the other loci. This is certainly caused by the unusual pattern of SNP and haplotype structure at this locus (see Figure 1.1). Finally, it does not seem that the decay of LD is much faster at loci with higher recombination rates.

Table 1.9: Summary of recombination parameters and Z_{ns}

	S. peruvianum					*S. chilense*				
Locus	Population	Hap[a]	R_m [b]	ρ[c]	Z_{ns} [d]	Population	Hap[a]	R_m [b]	ρ[c]	Z_{ns} [d]
CT093	TAR	0.80	3	0.0134	0.242	TAC	0.67	2	0.0021	0.373
	NAZ	0.83	2	0.0042	0.371	MOQ	1.00	1	0.0049	0.353
	CAN	1.00	3	0.0233	0.187	QUI	0.38	1	0.0014	0.653
	All	0.88	6	0.0119	0.105	**All**	0.63	4	0.0053	0.173
CT208	TAR	0.75	2	0	0.745	TAC	0.58	0	0	0.913
	NAZ	0.50	1	0	0.408	MOQ	0.70	0	0	NA
	CAN	0.58	1	0	0.552	QUI	0.43	1	0	0.367
	All	0.59	4	0.0017	0.161	**All**	0.53	2	0.0025	0.264
CT251	TAR	0.80	9	0.0051	0.383	TAC	1.00	9	0.0096	0.354
	NAZ	0.83	5	0.0147	0.227	MOQ	0.90	6	0.0074	0.356
	CAN	1.00	7	0.0431	0.197	QUI	0.50	4	0.0023	0.489
	All	0.88	15	0.0305	0.092	**All**	0.78	14	0.0081	0.182
CT066	TAR	0.80	3	0.0104	0.281	TAC	0.67	3	0.0037	0.346
	NAZ	0.58	0	0.0007	0.453	MOQ	0.90	0	0.0015	0.474
	CAN	0.92	5	0.0431	0.145	QUI	0.83	2	0.0007	0.486
	All	0.76	9	0.0240	0.091	**All**	0.76	4	0.0030	0.179
CT166	TAR	0.75	3	0.0015	0.475	TAC	0.75	6	0.0095	0.258
	NAZ	0.83	5	0.0140	0.238	MOQ	0.80	3	0.0037	0.379
	CAN	0.83	4	0.0051	0.264	QUI	0.58	1	0.0015	0.707
	All	0.81	10	0.0126	0.097	**All**	0.71	8	0.0130	0.141
CT179	TAR	0.70	4	0.0112	0.284	TAC	1.00	5	0.0386	0.258
	NAZ	1.00	10	0.1015	0.137	MOQ	0.80	6	0.0162	0.271
	CAN	0.92	8	0.0548	0.156	QUI	0.50	1	0.0010	0.581
	All	0.88	14	0.0965	0.057	**All**	0.77	12	0.0294	0.123
CT198	TAR	0.70	7	0.0103	0.342	TAC	0.70	1	0.0051	0.333
	NAZ	0.90	5	0.0411	0.223	MOQ	0.70	1	0.0026	0.659
	CAN	0.90	3	0.0154	0.294	QUI	0.60	5	0.0026	0.358
	All	0.83	15	0.0602	0.088	**All**	0.63	9	0.0201	0.183
CT268	TAR	1.00	13	0.0307	0.246	TAC	0.83	13	0.0530	0.168
	NAZ	1.00	10	0.0424	0.173	MOQ	0.80	8	0.0127	0.241
	CAN	1.00	12	0.0771	0.119	QUI	0.57	4	0.0042	0.374
	All	1.00	24	0.0795	0.054	**All**	0.72	17	0.0215	0.096

All = Analyses of combined sample, based on three populations (treated as a single population)
NA= Not Available

[a]Haplotype fraction
[b]minimum number of recombination events (Hudson and Kaplan 1985)
[c]population recombination rate (Hudson 2001)
[d]average of squared allele-frequency correlation (r^2) over all pairwise comparisons (Kelly 1997)

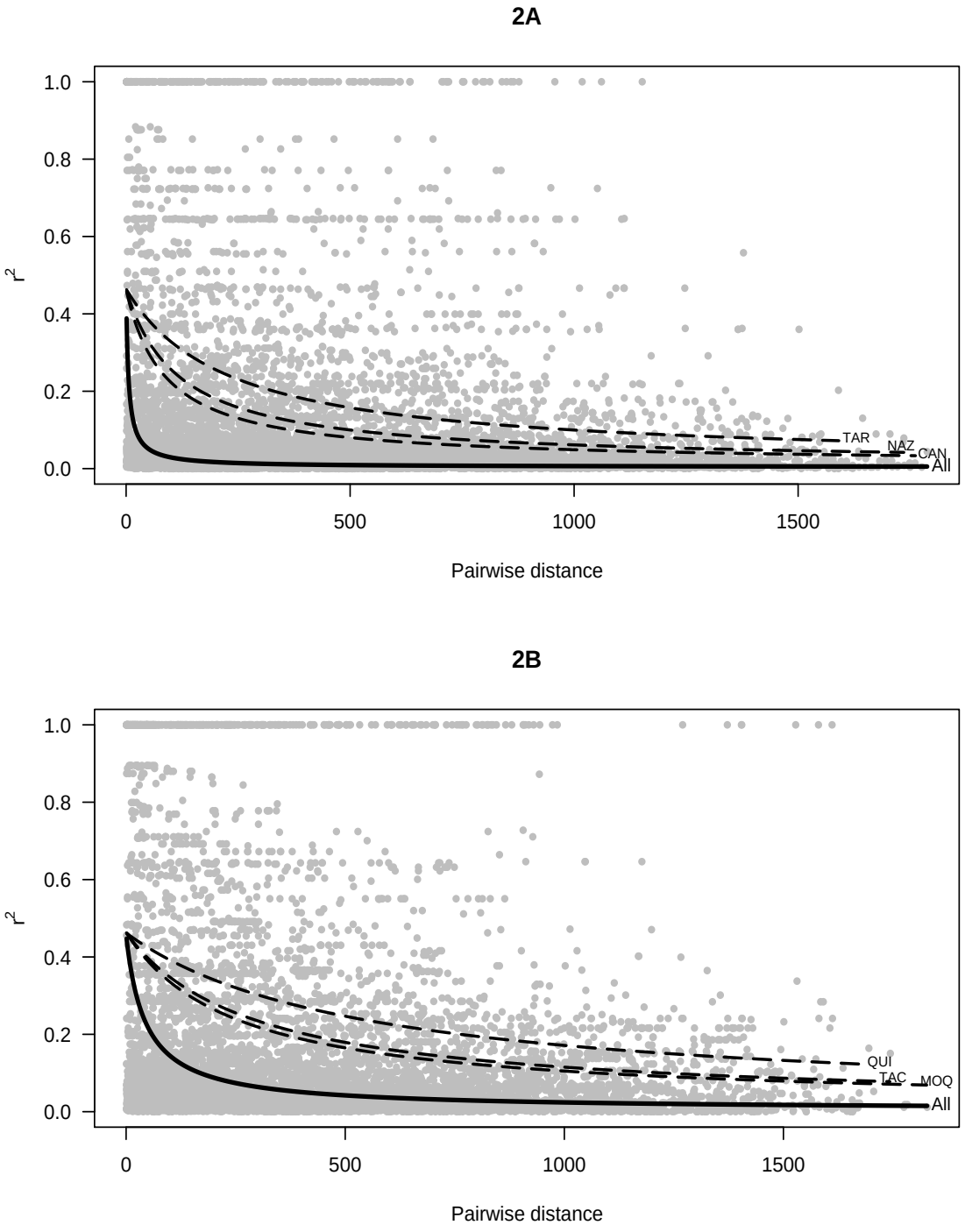

Figure 1.2: Plots of the square allele frequencies (r^2) versus pairwise distance in base pairs between polymorphic sites across eight loci in; *S. peruvianum* (2A), and *S. chilense* (2B). The solid lines depict the expected decline in linkage disequilibrium for all combined samples, and broken lines for individual populations. All lines are based on nonlinear regressions of r^2 against distance, using Equation of Hill and Weir (1988).

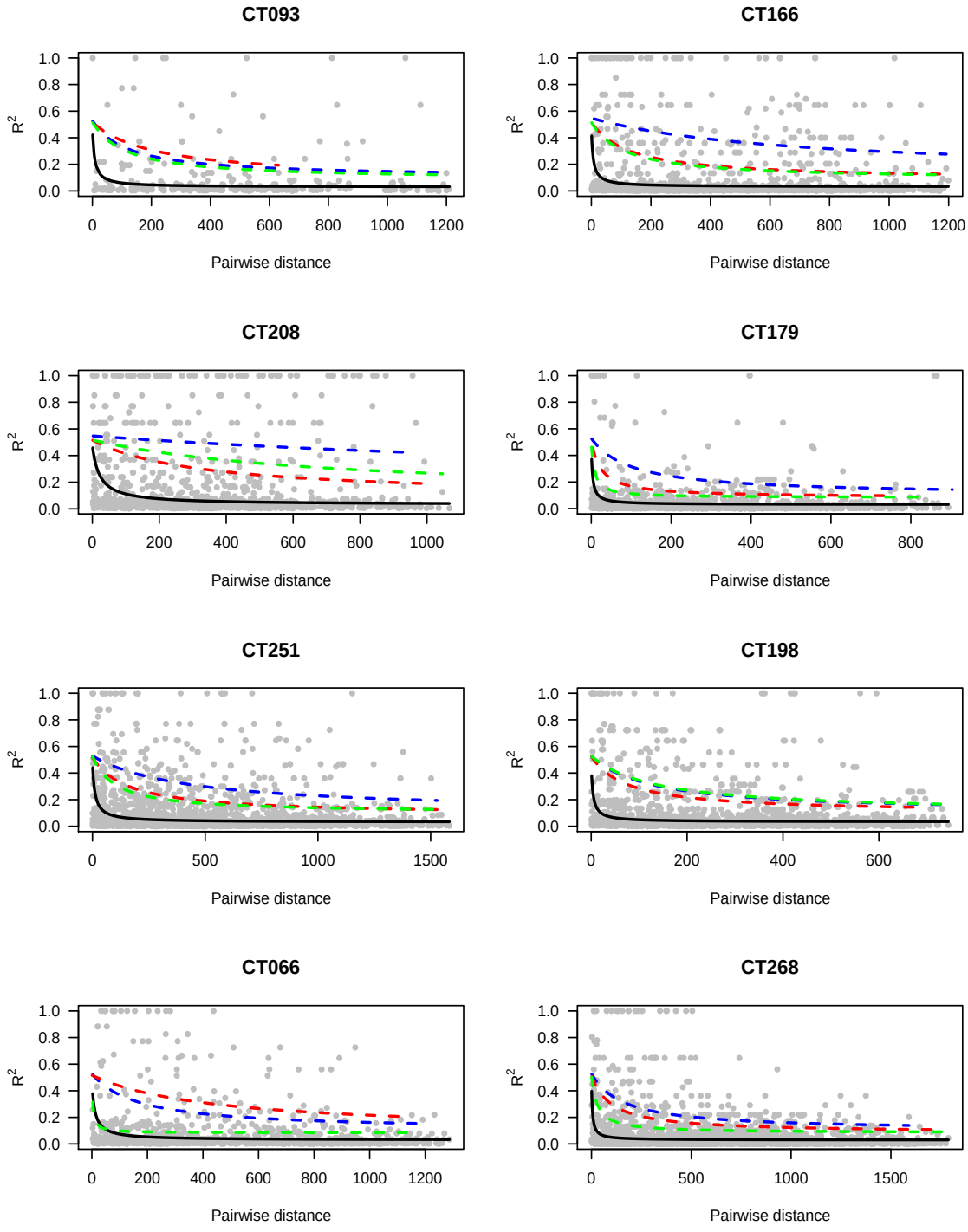

Figure 1.3: Plots of the square allele frequencies (R^2) versus pairwise distance in base pairs between polymorphic sites in *S. peruvianum*. The black thick lines depict the expected decline in linkage disequilibrium for all combined samples and broken lines for individual populations; TAR (blue), NAZ (red), and CAN (green). All lines are based on nonlinear regressions of r^2 against distance, using Equation of Hill and Weir (1988). The loci are presented from low to high physical recombination rates (Stephan and Langley, 1998), ordering from top to bottom and then left to right.

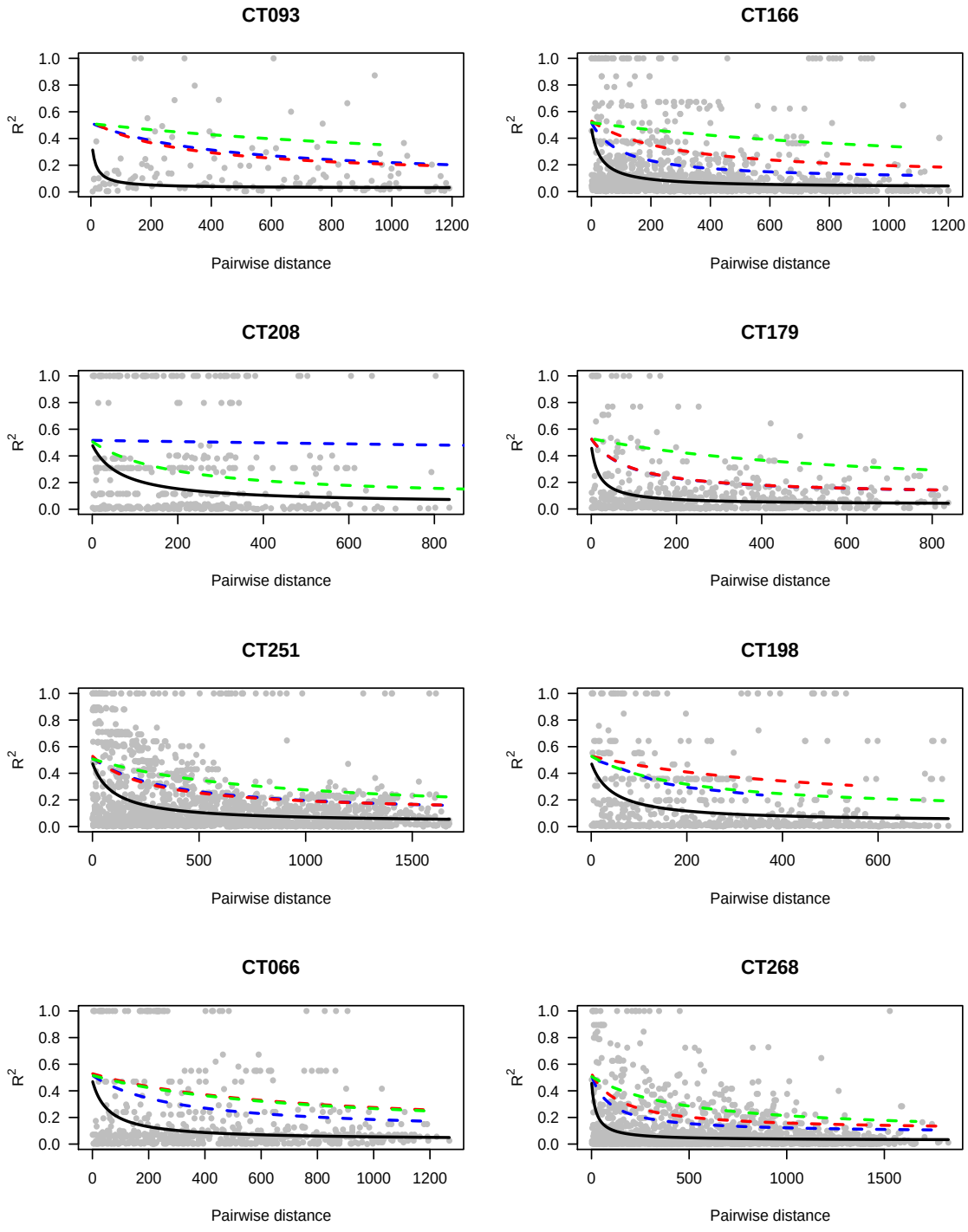

Figure 1.4: Plots of the square allele frequencies (R^2) versus pairwise distance in base pairs between polymorphic sites in *S. chilense*. The black thick lines depict the expected decline in linkage disequilibrium for all combined samples and broken lines for individual populations; TAC (blue), MOQ (red), and QUI (green). All lines are based on nonlinear regressions of r^2 against distance, using Equation of Hill and Weir (1988). The loci are presented from low to high physical recombination rates (Stephan and Langley, 1998), ordering from top to bottom and then left to right.

4 Discussion

4.1 Levels and patterns of nucleotide diversity

Both studied wild tomato species exhibit substantial levels of nucleotide variation. The level of silent polymorphism (π_{sil}) across eight loci is 0.0204 in *S. peruvianum* and 0.0159 in *S. chilense*. This average value is somewhat higher in *S. peruvianum* than the estimates in several other outcrossing angiosperms, whereas the level of diversity in *S. chilense* is comparable to other outcrossing taxa, *e.g. Arabidopsis lyrata* ($\approx$ 0.0140; Wright et al. (2003)), *A. halleri* ($\approx$ 0.0150; Ramos-Onsins et al. (2004)), maize ($\approx$ 0.0120; Tiffin and Gaut (2001)), and *Populus tremula* ($\approx$ 0.0160; Ingvarsson (2005)). However, the species-wide level of silent diversity in *S. peruvianum* is slightly lower than that obtained for wild sunflower ($\approx$ 0.0234; Liu and Burke (2006)).

When there is population structure, the number of demes is an important factor that might influence the species-wide level of variation (Whitlock and Barton, 1997; Pannell and Charlesworth, 1999; Wakeley and Aliacar, 2001; Laporte and Charlesworth, 2002). Given that *S. peruvianum* appears to be more patchily distributed than *S. chilense* (Rick 1986; T.S., personal observation), it is not surprising to find higher levels of nucleotide polymorphism in *S. peruvianum* than in *S. chilense*. However, other factors might also contribute to this difference between both species, such as demographic processes (*e.g.* population bottlenecks and expansion since species divergence). Some populations (*e.g.* ARE and ANT) are probably not under equilibrium conditions, a situation that might be mediated by local population bottlenecks and/or extinction-recolonization processes.

In general, the level of variation in selfing populations should be lower than in outcrossing taxa because the effective population size is expected to be halved in completely selfing populations (Nordborg and Donnelly, 1997). Certainly, the relevant estimates in outcrossing species are much higher than species-wide estimates in inbreeding taxa such as *A. thaliana* ($\pi_{sil} \approx 0.0083$; Schmid et al. (2005)), wild barley ($\pi_{all} \approx 0.0075$; Morrell et al. (2005)), as well as a local population estimate for soybean ($\pi_{sil} \approx 0.0015$; Zhu et al. (2003)).

4.2 Evidence of an ongoing selective sweep in *S. chilense*

Locus CT208 in *S. chilense* features an intriguing geographic pattern of nucleotide diversity, in that levels of nucleotide variation gradually diminish from north to south, with essentially no variation in the southernmost sample. Moreover, the northernmost population exhibits many haplotypes distinguished by ancestral variation (SNPs), as inferred by outgroup comparison (*S. ochranthum*), whereas the southernmost sample is fixed for a derived haplotype. In other words, the frequency of derived variants increases from north to south, with significantly negative H values for the MOQ and TAC populations, where the derived haplotype group has nearly fixed (9:1 and 10:2 respectively, cf. Figure 1.1). This clinal pattern of nucleotide variation is one of the first such instances in plants that we are aware of. One example is the recent study of European Aspen that showed clinal variation of four SNPs, suggestive of an adaptive response in *phyB2* to local photoperiodic conditions (Ingvarsson et al., 2006). However, the clinal pattern of variation seen in our study appears to be different, in that many SNPs spread out over the entire locus CT208, and mainly in noncoding regions. Additionally, the clinal pattern of nucleotide variation is quite different from a 'classical' expected pattern under a selective sweep, where most of the neutral variation linked to the selected locus is lost. Generally, a selective sweep scenario can be detected by a reduction in nucleotide diversity *e.g.* (Maynard Smith and Haigh, 1974; Kim and Stephan, 2002; Beisswanger et al., 2006; Kane and Rieseberg, 2007).

Two alternative possibilities to explain this geographic pattern of variation among populations might be (i) a mutational origin of the derived haplotype and its subsequent spread due to positive selection, and (ii) introgression of the selected haplotype from an unidentified donor species (or very divergent population). The first scenario seems unrealistic because divergence appears to be too high for the selected haplotype group to have arisen by mutation within *S. chilense*. The second possibility amounts to introgression of an adaptive haplotype and its subsequent spread. Morjan and Rieseberg (2004) suggested that even very low migration rates might be sufficient for the spread of advantageous alleles. One putative example of the introgression of an adaptive haplotype is the recent study on the brain-size gene *microcephalin* in humans (Evans et al., 2006). This study marshaled evidence for introgression of an allele from an archaic *Homo* lineage into modern humans and its subsequent rise to high frequency under positive selection. In our case, two potential sources of origin

for the selected haplotype are the tomato relatives *S. lycopersicoides* and *S. sitiens*, as both species occur within the geographic range of *S. chilense*. However, these two species are characterized by very strong reproductive isolation from *S. chilense* (Rick, 1979). Clearly, it would be useful to obtain sequences from both species as outgroup comparisons.

Santiago and Caballero (2005) showed that a selective sweep in a structured population can lead to an increase in neutral variation in particular subpopulations. After the selective sweep, variation at linked neutral sites is expected to be reduced in the population where the new variant first appeared, as predicted by the classical sweep scenario. However, the effect in the other populations, where the mutation is introduced by gene flow, can be different depending on the level of genetic differentiation among populations. The model of a selective sweep in two subpopulations (Santiago and Caballero, 2005) can plausibly be used to explain the level and pattern of nucleotide diversity at locus CT208, where levels of variation are different among populations. The explanation for the geographic pattern seen in CT208 might be the presence of a soil seed bank, which probably retards the loss of genetic diversity within populations as well as differentiation among them. A seed bank is a likely contributing factor prolonging the time needed for local adaptation and/or selective sweeps (in addition to population subdivision *per se*).

4.3 Population structure and its consequences

We obtained levels of F_{st} estimates in both species that indicate moderate population structure. Levels of genetic differentiation in wild tomatoes, which are insect-pollinated herbaceous perennials, are close to the mean level found in a metaanalysis of other comparable outcrossing plant species, based on allozyme data (Hamrick and Godt, 1996). Furthermore, based on nucleotide data, levels of F_{st} estimates in both studied wild tomatoes are broadly comparable to the estimates in *A. lyrata* (Wright et al., 2003), as well as in the wind-pollinated tree species Populus tremula (Ingvarsson, 2005). Life-history traits such as the breeding system, life forms, geographic ranges and seed dispersal mechanisms tend to be associated with different levels of genetic diversity within and differentiation among populations (Gottlieb, 1977; Hamrick and Godt, 1989, 1996). Moreover, Hamrick and Godt (1996) reported that life form and breeding system in particular had highly significant influences on genetic diversity and its distribution. More explicitly, outcrossing species have significantly less

genetic diversity distributed among subpopulations, regardless of their other traits. However, the genetic diversity maintained by a species is not only a function of its life history traits but also heavily depends on the ecological and evolutionary history of the species (Hamrick and Godt, 1989, 1996), see above and companion study; (Städler et al., MS).

Generally, the pollen and/or seed dispersal potential of plants should affect their ability to maintain local genetic diversity. Hamrick and Godt (1996) showed that outcrossing species with limited pollen and/or seed dispersal tend to have greater genetic differentiation among populations than species with more potential for gene movement. In addition, it is most likely that the tall stature and comparatively low population densities of trees should result in larger dispersal distances for seeds and pollen than would be expected to occur in herbaceous species (*e.g.* wild tomatoes), whose individuals are generally shorter and found in more dense stands. Hence, pollen and/or seed dispersal under equilibrium conditions may not be sufficient to explain the patterns of population differentiation in our samples, especially the low differentiation between the northernmost and southernmost populations in *S. peruvianum*, despite the greatest geographic distance and the large differences in habitats between them. An alternative explanation for patterns of differentiation in our samples rests on the likely presence of soil seed banks, and historical association of tomato populations mediated by climatic cycles (as alternative to equilibrium gene flow). Soil seed banks probably play an important role in maintaining the large genetic diversity in wild tomatoes (Roselius et al., 2005). The presence of seed banks can have a major impact on effective population size and consequently the maintenance of genetic diversity in plants (Levin, 1990; Nunney, 2002).

Given that the El Niño Southern Oscillation (ENSO) is a key phenomenon for weather patterns in the tropical Pacific Ocean (Devries, 1987; Tudhope et al., 2001; Tudhope and Collins, 2003), these cyclical events are expected to affect seed germination and other aspects of plant ecology and evolution over large regions of coastal western South America. Gutiérrez and Meserve (2003) suggested that the ENSO influences plant establishment and the replenishment of soil seed banks in many plant taxa. It thus seems likely that the ENSO has a major effect on fluctuations in the number and size of tomato populations, and consequently influences patterns of population structure over time. More specifically, the effects of ENSO might result in high (transient) levels of connectivity across the species range, perhaps via 'bridging'

populations, or might influence species-wide or regional expansions and contractions.

One of our most interesting findings is that Tajima's D statistic is higher in the samples of the individual populations than in the pooled sample, where the pooled samples show a significant excess of low-frequency polymorphisms. We argue that this may also be a consequence of the population structure in both wild tomato species. At first sight, this seems perplexing, as standard models of population structure, such as the island model, cannot explain this observation. Indeed, the latter model predicts that the pooling of samples from different populations should increase the proportion of intermediate- to high-frequency polymorphisms (Tajima, 1989; Pannell, 2003). The average structure of genealogies of our polymorphic ('typical') samples is characterized by deep internal branch (probably reflecting subdivision in the ancestral species) with the final coalescence events occurring in the ancestral species. Two aspects of our data lead to this conclusion, (i) the high within-deme diversity compared to species-wide levels (*i.e.* relatively low F_{st}), and (ii) the absence of fixed nucleotide differences between populations and even between the two species (Städler et al., 2005). Thus, pooling the population samples does not result in additional internal branches of the genealogy that would account for higher proportions of intermediate-frequency polymorphism. Rather, pooling of samples yields an excess of singletons, as most of the singletons within local populations still remain singletons in the pooled sample.

In addition, more negative D_T values for pooled samples were found in some previous studies, without offering a general explanation *e.g.* (Ingvarsson, 2005; Liu and Burke, 2006). Our study therefore highlights the importance of sampling strategies, as most studies do not include 'real' population samples (*e.g.* using single individuals per deme). When sampling single individuals from many locations and pooling them as species-wide samples, one may obtain negative D_T values, the interpretation of which, however, may not be straightforward. Hence, the biological factors underlying these patterns need to be explored theoretically.

4.4 Recombination and the decay of linkage disequilibrium

Intragenic LD in wild tomatoes decays rapidly to very low levels ($r^2 < 0.05$) within a few hundred base pairs in both species. This is true even at the loci with very low estimates of physical recombination rate (*i.e.* CT093 and CT208; Stephan and Langley (1998)), suggesting considerable levels of recombination at all loci, as also shown by appreciable R_m estimates in both species. Hence, the estimates of

physical recombination rate at some loci might possibly be underestimated. Moreover, high levels of haplotype diversity in both species also imply sufficient levels of recombination, which consequently result in the rapid decline of LD in both species.

In general, the decay of LD with distance varies greatly in different species, such as within 250 kb in *Arabidopsis thaliana* (Nordborg et al., 2002, 2005), from 0.2-1.5 kb in maize (Remington et al., 2001; Tenaillon et al., 2001), less than 1 kb in *Drosophila melanogaster* (Long et al., 1998), and from 5-80 kb in humans (Reich et al., 2001). Overall, the rapid decay of LD in wild tomatoes is comparable to that previously documented in several outcrossing plant species. For example, LD decays to about 50% within 2 kb in loblolly pine (Brown et al., 2004), to negligible levels within 250 bp in European Aspen (Ingvarsson, 2005), within a few hundred base pairs in Norway Spruce (Heuertz et al., 2006), and within 200 bp in wild sunflower (Liu and Burke, 2006).

In plants particularly, different mating systems and recombination rate are important factors that affect the decay of LD with distance. The relationship between recombination and mating system can increase or decrease levels of LD. Nordborg and Donnelly (1997) suggested that effective recombination rate is related to the degree of selfing because recombination is less effective in selfing populations where individuals are more likely to be homozygous than in outcrossing taxa. Therefore, LD will be more extensive in selfing than in outcrossing populations. Indeed, LD in outcrossing species decays much faster than under a primarily selfing mating system, *e.g.* significant LD extends to 100 kb in rice (Garris et al., 2003), to >50 kb in soybean (Zhu et al., 2003), and to about 70 kb in potato (outcrossing species but usually vegetatively propagated; Simko et al. (2006)).

The difference in LD observed across species is a result of the interplay of many factors, *e.g.* recombination rate, mating system, selection, effective population size, and population structure (reviewed by Rafalski and Morgante (2004)). Among these factors, high recombination rate, obligate outcrossing and high N_e in concert provide the most plausible explanation for the rapid decline of LD in our samples, whereas the more extensive LD seen at locus CT208 in *S. chilense* is easily explained by the unusual haplotype structure within and among populations (incomplete selective sweep).

Under equilibrium conditions, it is possible to calculate estimates of the effective population size based on the population parameters ρ and θ, where $\rho = 4N_e$r and $\theta = 4N_e\ \mu$ (Li, 1997). In this study, we estimated the effective population size of wild tomatoes on the basis of ρ, using the average physical recombination rate per generation based on six loci (r ≈ 1.56 x 10^{-8}, excluding CT093 and CT208 because the recombination rates appear to be underestimated, see above). Using the estimated mean ρ based on two relatively undifferentiated populations per species (see Results), we obtain estimates of $N_e \approx 6.63$ x 10^5 for *S. peruvianum* and ≈ 4.88 x 10^5 for *S. chilense*. These N_e estimates are somewhat lower than the N_e estimates from our previous study, which were based on θ estimates from single populations per species (Roselius et al., 2005).

5 Conclusion

In comparison to the previous studies in wild tomatoes by Baudry et al. (2001), Städler et al. (2005), and Roselius et al. (2005), which were based on only single populations per species, our study allows for more generality by adding three additional populations, broadly covering most of the species' range. We propose that population structure is one of the most important evolutionary forces that shaped patterns of nucleotide diversity within and among populations in these wild tomatoes. Given that wild tomatoes are subdivided species, our assessment of population structure allowed us to understand several patterns of variability in wild tomatoes. First, analyzing a few genuine population samples enabled us to discover and interpret the clinal pattern of variability at CT208 as (likely) signatures of an ongoing selective sweep in *S. chilense*. Sampling only a single population or, alternatively, single individuals from many demes across the species range, might have entirely missed this signature. Second, our unexpected finding that pooling of samples from different local populations led to an excess of low-frequency variants, which contradicts predictions of the standard models of population subdivision, suggests that the population structure of both tomato species (in particular *S. peruvianum*) is complex and probably influenced by subdivision of the ancestral species. Third, the rapid decay of LD in both species is very useful for high-resolution mapping in association studies given that appropriate candidate genes are chosen. For this purpose, however, a high-density marker screening would be needed.

Chapter 2

Population genomics of speciation in two closely related wild tomatoes (*Solanum* section *Lycopersicon*)

1 Introduction

The biological and geographic determinants of species divergence have long been contentious, and it is now increasingly appreciated that patterns of genetic variation and differentiation may provide valuable insights into the evolutionary processes shaping this divergence. The importance of geographic isolation in facilitating evolutionary divergence as a consequence of mutation and genetic drift (or additionally, adaptive differentiation) was recognized early, and the process of allopatric speciation is uncontroversial on theoretical grounds (Mayr, 1963; Losos and Glor, 2003; Coyne and Orr, 2004). If residual gene flow characterized the divergence of incipient species, however, modes other than strict allopatric speciation must be invoked, and these invariably require natural selection as one of the factors underlying species divergence. In addition to the putatively rare cases of sympatric speciation (*e.g.*, Savolainen et al. (2006)), divergence under residual gene flow may proceed in parapatry, *i.e.*, geographically adjacent populations may be subject to directional selection that incidentally confers reproductive isolation (Endler, 1977; Turelli et al., 2001; Gavrilets, 2003). Another scenario is an initial period of divergence in allopatry followed by secondary contact allowing gene flow and thus direct selection for stronger interspecific barriers (reinforcement of reproductive isolation; Rice and Hostert (1993); Coyne and Orr (2004); Hoskin et al. (2005)). Some researchers posit that interspecific hybridization

and postdivergence gene flow following secondary contact may promote novel advantageous gene combinations in populations of mixed ancestry, perhaps contributing to adaptive divergence and speciation *e.g.* (Arnold, 1997; Seehausen, 2004; Rieseberg et al., 2004; Mallet, 2005).

Multilocus gene sequences collected within and among closely related species contain a wealth of historical-demographic information and are particularly informative when considered in the framework of genealogical (coalescence) models *e.g.*, (Tajima, 1989; Hudson, 1990; Rosenberg and Nordborg, 2002). As an extension of population genetic procedures to the species level, the analytical framework of divergence population genetics (DPG) encompasses coalescent-based models to infer historical attributes of lineage divergence from a common ancestor, and to assess the utility of simple speciation models (Hey and Kliman, 1993; Wakeley and Hey, 1997; Wang et al., 1997; Machado et al., 2002; Hey and Machado, 2003; Hey and Nielsen, 2004). The DPG approach accommodates the stochastic nature of lineage sorting (Edwards and Beerli, 2000; Hudson and Turelli, 2003) and thus the (gradually decreasing) segregation of shared ancestral polymorphism in the descendent species, as these become more differentiated through genetic drift and the accumulation of new mutations.

The 'isolation' model of speciation (Wakeley and Hey, 1997; 'WH model') assumes divergence in isolation without subsequent gene flow, and as such is an explicit model of allopatric speciation. The WH model makes quantitative predictions regarding patterns of nucleotide diversity across multiple loci, and sequence data obtained from recently diverged taxa can provide scaled estimates of population-size changes and the timing of speciation, as well as probe for signatures of postdivergence gene flow (Wakeley and Hey, 1997; Wang et al., 1997; Kliman et al., 2000; Machado et al., 2002; Broughton and Harrison, 2003). More recently, bidirectional gene flow following initial species divergence has been incorporated as additional model parameters in the isolation-with-migration (IM) model (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004), but a notable restriction of its current implementation is the assumption of nonrecombining data within loci. Despite this limitation, the IM model appears to enjoy increasing popularity *e.g.*, (Dolman and Moritz, 2006; Kronforst et al., 2006).

The currently available coalescent-based speciation models (WH, IM) further assume panmixia within both extant and ancestral species, an assumption that is rarely tested or even discussed in empirical applications of these models. We suspect

that not only DPG studies in plants but also those in many animal groups are at risk of using inappropriate models of divergence, given the likely importance of population subdivision in many taxa; similar concerns have been raised in the context of statistical phylogeography (Knowles and Carstens, 2007). In assessing possible biases in the WH parameter estimates due to population subdivision and sampling, we will return to this issue in the Discussion.

Multilocus genealogical studies of speciation scenarios are still very limited for plants (Ramos-Onsins et al., 2004; Städler et al., 2005; Zhang and Ge, 2007). Building on our pilot study that was limited to single populations per species (Städler et al., 2005), the current paper provides an in-depth assessment of the divergence process between two closely related wild tomato species, *Solanum peruvianum* and *S. chilense* (*Solanum* section *Lycopersicon*, *Solanaceae*). Previous studies of *Lycopersicon* using a variety of molecular markers have generally found low levels of differentiation among species *e.g.*, (Miller and Tanksley, 1990; Baudry et al., 2001; Peralta and Spooner, 2001), implying a fairly recent divergence of the tomato clade. In particular, our multilocus study of three self-incompatible species demonstrated variation between species pairs in the proportion of loci showing some/many fixed interspecific differences, *vs.* those with appreciable numbers of shared polymorphisms (Städler et al., 2005). These differential signals of genealogical divergence highlight the suitability of wild tomatoes as a plant speciation model under the DPG framework; they also imply widespread incomplete lineage sorting.

Based on multiple population samples per species, we used seven effectively unlinked nuclear loci to estimate population parameters and scaled divergence times between these outcrossing tomato species. Additionally, we sought evidence for divergence under residual gene flow by applying a test based on patterns of intragenic linkage disequilibrium (LD). While we cannot reject the isolation model based on overall goodness-of-fit criteria, patterns of LD are indicative of historical gene flow between the diverging lineages.

2 Materials and Methods

2.1 Study system and sampling

For this in-depth study, we chose two of the previously used self-incompatible wild tomato species. In contrast to our exploratory study of three taxa (Städler et al., 2005), we adopt the current taxonomic treatment of tomatoes as a section within the large genus Solanum (*e.g.*, Spooner et al. (1993); Olmstead et al. (1999)). Our study species are the widely distributed *S. peruvianum* and the southernmost tomato species, *S. chilense*. Native to western South America, the two morphologically differentiated species have partly overlapping ranges in the arid coastal regions of southern Peru and northern Chile, west of the continental divide (Rick and Lamm, 1955; Rick, 1979, 1986; Taylor, 1986); see Figure 1 in Städler et al. (2005). Systematists have recently proposed to recognize four species in what traditionally was regarded as the polymorphic *S. peruvianum* (Peralta et al., 2005; Spooner et al., 2005). According to their proposition, our sampling of three natural populations in central and southern Peru (see below) would encompass both the new entity *S. corneliomuelleri* and *S. peruvianum* sensu stricto. However, there appear to be neither molecular data nor crossing results that would validate the proposed split of *S. corneliomuelleri* from *S. peruvianum* s. str.; we thus treat all of our new samples as *S. peruvianum*. There is no published evidence for interspecific hybridization between *S. chilense* and *S. peruvianum* in their natural habitats, in concordance with the strong reproductive barriers uncovered in experimental crossing studies (Rick and Lamm, 1955; Rick, 1979, 1986).

For each of the two study species, three new population samples were collected in southern and central Peru in May, 2004 (TS and T. Marczewski); population nomenclature, geographical locations and altitude are summarized in Table 2.1. With the exception of the Canta population (*S. peruvianum*), all new samples are from regions of sympatry with the other species, even though this may not be true at a local scale. The Canta population, however, is far north of the *S. chilense* species range. Five to seven plants were collected per population, and altitude and geographic coordinates were determined by GPS. We sampled approximately 3-5 gram of fresh leaf tissue per plant and stored it in plastic bags with silica gel until our return to Munich. Voucher specimens have been deposited at USM (Lima, Peru) and MSB (Munich, Germany). Our exploratory study used one accession (equivalent to a population sample) of each species, obtained from the Tomato Genetics Resource Center at U.C. Davis (Städler

et al., 2005). These accessions were both from northern Chile (*S. chilense*: accession LA2884, Antofagasta, five plants; *S. peruvianum*: LA2744, Tarapaca, five plants).

Table 2.1: Geographical origin of the sampled populations

Species/ population	Population identifier	Department/ region	Latitude/ longitude	Elevation (m)
S. peruvianum				
Tarapaca	TAR	Tarapaca	18°33'S,70°09'W	400
Arequipa	ARE	Arequipa	16°27'S,71°42'W	2180
Nazca	NAZ	Ica	14°51'S,74°44'W	2140
Canta	CAN	Lima	11°32'S,76°42'W	2050
S. chilense				
Antofagasta	ANT	Antofagasta	22°14'S,68°23'W	2900
Tacna	TAC	Tacna	17°53'S,70°08'W	1260
Moquegua	MOQ	Moquegua	17°04'S,70°52'W	2450
Quicacha	QUI	Arequipa	15°38'S,73°48'W	1830

Within each species, the populations are listed from south to north, and are generally named after a nearby town or city. All samples except TAR and ANT are from Peru; TAR and ANT originate from northern Chile (see Figure 1 in Städler et al. 2005). With few exception, we sequenced 10 to 12 alleles per population (five or six plants).

2.2 Choice of marker loci

Linkage maps are available for all 12 tomato chromosomes, based mainly on interspecific crosses between *S. lycopersicum* and *S. pennellii* (Tanksley et al. (1992); http://www.sgn.cornell.edu). Sequence information is available for many of the mapped cDNA markers (Ganal et al., 1998), which have been integrated into longer 'tentative contigs' in the Tomato Gene Index at the Institute for Genomic Research (http://www.tigr.org/tdb/lgi). For this study focusing on multiple populations per species, we chose a subset of the loci previously used in our initial surveys (Baudry et al., 2001; Städler et al., 2005; Roselius et al., 2005); CT066, CT093, CT166, CT179, CT198, CT208, CT251 and CT268 are eight anonymous, single-copy cDNA markers previously mapped by Tanksley et al. (1992). Given that it was impossible to se-

quence so many samples at all the previous loci, this reduced set of genes was chosen primarily because it yielded very similar proportions among the isolation model parameters as when using the full set of genes (Städler et al., 2005).

However, our companion study focusing on tests of neutrality, population subdivision and linkage disequilibrium found clear evidence of non-neutral evolution at locus CT208 (Arunyawat et al., MS). Because it is paramount to avoid the inclusion of loci under positive selection in testing the isolation model, we decided to base our WH simulations on the seven remaining loci that show no obvious departures from neutral expectations. Sequencing and haplotype determination: Genomic DNA was extracted from dried leaf tissue using the DNeasy plant mini kit (Qiagen GmbH, Hilden, Germany). PCR conditions followed those of our previous studies (Städler et al., 2005; Roselius et al., 2005); they as well as all primer information can be accessed at http://www.zi.biologie.uni-muenchen.de/evol/Downloads.html. For locus CT166, we designed new PCR primers and amplified a shorter fragment (about 1,300 bp) compared to the original studies. Sequencing was performed on an ABI 3730 DNA analyzer (Applied Biosystems, Foster City, CA). Distinct haplotypes within heterozygous individuals were resolved by applying a suite of haplotype-specific sequencing primers. In most cases, we exploited putative or confirmed SNPs to anchor the $3'$-end of sequencing primers that were intended to resolve the heterogeneous PCR products. This approach enabled us to verify SNPs (and indel variation) and establish haplotype phase based on overlapping information supported by multiple primer pairs. Sequence alignments were initially done either in Sequencher (Gene Codes, Ann Arbor, MI) or in Sequence Navigator (Applied Biosystems, Darmstadt, Germany) and adjusted manually in MacClade (Maddison and Maddison, 1992).

2.3 Estimation of polymorphism, frequency spectrum and haplotype structure

Standard population genetic analyses of the sequence data were performed using the program packages DnaSP, version 4.0 (Rozas et al., 2003) and SITES (Hey and Wakeley, 1997). As a measure of intraspecific polymorphism, we calculated Watterson (1975)'s estimator θ_W of the population mutation parameter, as well as the estimator π (average pairwise sequence divergence within samples; Nei (1987)). In this study, we restrict our attention to SNPs (excluding all insertion-deletion polymorphism) and report all estimates of nucleotide diversity as per-site values. We tested

for departures from neutral equilibrium expectations by applying the intraspecific, standard Tajima (1989)'s D test, which can extract signatures of changes in effective population size based on the frequency spectrum of segregating mutations. Significance of the population- specific multilocus D statistics was assessed with 10,000 coalescent simulations, as implemented in the HKA program. The observed number of distinct haplotypes per sample was evaluated with Fu (1997)'s F_s test statistic, although there is no straightforward way to obtain the probability of multilocus (mean) F_s values.

2.4 Testing the isolation speciation model

The multilocus data were fitted to the WH isolation model (Wakeley and Hey, 1997; Wang et al., 1997). This simple model of allopatric speciation assumes that an ancestral, panmictic species characterized by the population mutation parameter θ_A gave rise to two extant species at time τ in the past ($\tau = 2$ut, where t is the number of generations since speciation). The extant species are characterized by the mutation parameters θ_1 and θ_2 , respectively. Hence, effective population size is assumed to be constant within species, but is allowed to change at the time of speciation. The model further assumes the neutrality of segregating variants and no gene flow (introgression) subsequent to the initial species divergence. Our current data encompass multiple populations and – if treated as a single sample – were not expected to match the model assumption of panmixia within species. In fact, our companion paper demonstrates significant population subdivision in both species, as expected for insect-pollinated angiosperm taxa (Arunyawat et al., MS). In order to extract biologically meaningful WH parameter estimates, data from single population pairs were sequentially fitted to the isolation model, *i.e.*, without pooling populations within species. Pooling was done, however, for two populations per species showing the least amount of genetic differentiation, as discussed below.

As shown by Wakeley and Hey (1997), the expectations of the observable quantities S_{x1}, S_{x2}, S_s and S_f (exclusive polymorphisms for species 1 and 2, shared polymorphisms and fixed differences, respectively) are functions of the four model parameters. Hence, by equating observations with expectations, their moment-based algorithm yields the parameter estimates from multilocus data. For each pair of interspecific populations (*e.g.*, TAR-MOQ, CAN-TAC, *etc.*, see Table 2.1) and each of the seven loci, we calculated the number of observations for each of the four site categories

(excluding sites with observable multiple hits), and simultaneously estimated the population recombination parameter γ , using the program SITES (Hey and Wakeley, 1997). For each pairwise comparison, we ran 10,000 coalescent simulations using a modified WH program (Wakeley and Hey, 1997; Wang et al., 1997; Städler et al., 2005).

To avoid biasing the recombination rate downwards in the WH simulations, we let γ be 'unknown' for the *S. chilense* samples, except in a few cases where γ for the *S. peruvianum* sample was estimated to be zero but the *S. chilense* sample had a $\gamma > 0$; in such cases we used the *S. chilense* γ and let the *S. peruvianum* γ be 'unknown'. As implemented in WH, the relative magnitudes of the model parameters θ_1 and θ_2 and θ_A determine the level of recombination for 'unknown' entries, such that $\gamma_2 = \gamma_1 \times (\theta_2/\theta_1)$ and $\gamma_A = \gamma_1 \times (\theta_A/\theta_1)$, where γ_1 in our case is the previously estimated locus-specific recombination parameter for the *S. peruvianum* sample. This has the effect of imposing a fixed ratio of γ 's in a given interspecific comparison, but allows for variation in levels of recombination across loci.

2.5 Linkage disequilibrium test of gene flow

Machado et al. (2002) introduced a test of gene flow based on patterns of linkage disequilibrium (LD) among specific classes of segregating sites, *i.e.*, using a subset of total intragenic LD. Under a scenario of gene flow, LD among pairs of shared polymorphisms (average = D_{ss}) in the recipient species should tend to be positive (*i.e.*, preponderance of ancestral-ancestral and/or derived-derived SNP associations), and LD among pairs of sites where one member is a shared and the other an exclusive polymorphism (average = D_{sx}) should tend to be negative (*i.e.*, preponderance of ancestral-derived SNP associations). Both expected effects can be seen as a consequence of insufficient time for recombination to erode LD (given introgression has occurred after initial species separation) compared to the situation where shared polymorphisms represent truly ancestral mutations, *i.e.*, those preceding speciation.

Unlike for our initial study (Städler et al., 2005), the availability of outgroup sequences from species outside the tomato clade allowed us to use the LD test statistic proposed by Machado et al. (2002). Here, we use sequence data generated from either *S. lycopersicoides* (CT093, CT268) or *S. ochranthum* (all other loci; Roselius et al. (2005)) to polarize LD. It may be expected that polarized LD has greater power to detect historical introgression than our previous, unpolarized LD test (Städler et al.,

2005), although this has not been formally evaluated. Using the LD measure D' (=D/D_{max}, possible range from -1 to +1), the observed values of the test statistic x (D_{ss} - D_{sx}; Machado et al. (2002); computed in the SITES program) were confronted with expectations generated by the same set of WH simulations that was used to test the quality of fit of the isolation model. For these analyses, only loci with at least four pairs of sites in each of the above categories were used, which excludes loci with observed S_s values < 4.

3 Results

3.1 Single nucleotide polymorphism and tests of neutrality

Levels of nucleotide variation at eight nuclear loci and tests of neutrality have been summarized in our companion paper (Arunyawat et al., MS). For our purpose of testing simple models of speciation, we utilized the seven loci that appear to evolve neutrally and re-calculated population-specific mean diversity levels and multilocus Tajima's D and Fu's F_s statistics (Table 2.2). Levels of overall and silent polymorphism (using only noncoding and synonymous sites) are consistently higher for *S. peruvianum* than for *S. chilense* populations, except for the ARE sample (see below). Average θ_{sil} estimates for 'typical' *S. peruvianum* populations range from 2.18-2.94% whereas the range of θ_{sil} estimates for 'typical' *S. chilense* populations is 1.68-1.93% (Table 2.2).

The multilocus Tajima (1989)'s D statistic is slightly, but not significantly, negative in the three *S. peruvianum* samples TAR, NAZ and CAN, while estimates of D are close to zero in the three *S. chilense* samples TAC, MOQ and QUI (Table 2.2). Clearly, the new sequence data obtained for three natural populations of *S. chilense* indicate that our first sample (ANT; Städler et al. (2005)) is uncharacteristic in terms of polymorphism, frequency spectrum and haplotype structure (Table 2.2). Similarly, the *S. peruvianum* ARE sample departs from the other populations by exhibiting lower polymorphism, a slight excess of intermediate-frequency polymorphism (positive D), and a slight deficit of distinct haplotypes (positive F_s; Table 2.2). It is likely that these two 'outlier' samples (ANT, ARE) reflect local or regional bottlenecks and/or greater degrees of isolation from neighboring demes (*i.e.*, lower proportions of immigrants via gene flow), compared to the bulk of the species' range. Because we are interested in demographic estimates reflecting species-wide patterns subsequent to their recent divergence, we decided to restrict our tests of the isolation model to the three apparently 'typical' population samples per species.

The estimates of Tajima's (1989) D statistic presented in Table 2.2 are based on all classes of polymorphism, *i.e.*, including substitutions at noncoding, synonymous and nonsynonymous sites. Although there is clear evidence that levels of nonsynonymous polymorphism are markedly lower than levels of silent polymorphism (compare θ_{all} and θ_{sil} in Table 2.2), as expected under strong purifying selection, D estimates based on only silent sites are very similar (data not shown). This suggests that

Table 2.2: Levels of nucleotide polymorphism in eight population samples

Species/ Population	Length (bp)	#SNPs	θ_{all} (%)	π_{all} (%)	θ_{all} (%)	π_{all} (%)	D_T	F_s
S. peruvianum								
TAR	9,314	322	1.24	1.18	2.43	2.31	-0.22	0.64
ARE	9,286	186	0.71	0.79	1.43	1.55	0.41	1.67
NAZ	9,304	318	1.14	1.11	2.18	2.08	-0.14	-1.16
CAN	9,309	400	1.45	1.30	2.94	2.52	-0.55	-1.98
S. chilense								
ANT	9,344	125	0.47	0.61	0.98	1.25	**1.53**	7.38
TAC	9,353	260	0.93	0.94	1.69	1.73	0.03	-0.70
MOQ	9,347	276	1.04	1.05	1.93	2.00	-0.05	-0.47
QUI	9,317	257	0.91	0.98	1.68	1.80	0.13	2.72

Length refers to the total number of base pairs across the seven loci, where all indels have been exclided; #SNPs reports the total number of segregating sites per sample, some of which were excluded for the WH analyses due to multiple hits. The estimators of nucleotide diversity (θ and π) are weighted means across the seven loci and are given for all sites ('all'; mean = 1,333 bp/locus) and for silent sites ('sil'; mean = 574 bp/locus). Tajima's D and Fu's F_s statistics are based on all SNPs and are given as the unweighted means among loci; the multilocus D is significant at $P<0.001$ for ANT (boldface). Populations TAR and ANT (from Städler et al. (2005)) are included here for comparison.

the *segregating* nonsynonymous variants are not under strong negative selection but rather behave in a nearly-neutral manner. Because there was no obvious evidence of non-neutrality in the data comprising these seven loci, we felt justified in using the entire SNP data for testing the isolation model of speciation.

3.2 Polymorphic site categories in interspecific population contrasts

The among-locus distributions of four categories of polymorphic sites, namely exclusive polymorphisms in species 1 and 2 (S_{x1} and S_{x2}, respectively), polymorphisms that are shared among species (S_s), and fixed differences between the species (S_f) represent summary statistics of the data and collectively yield the WH isola-

Table 2.3: Distribution of polymorphic sites in the interspecific contrast of the most variable population samples (Canta-Moquegua)

Locus	Length(bp)	S_{x1}	S_{x2}	S_s	S_f	$\gamma_{peru}(\%)$
CT066	1,332	32	27	9	0	2.05
CT093	1,392	24	19	6	0	2.07
CT166	1,279	56	41	11	0	1.61
CT179	882	36	20	9	0	4.79
CT198	753	45	9	8	0	2.44
CT251	1,684	47	28	21	2	3.61
CT268	1,880	51	31	13	1	3.91
Total	9,202	291	175	77	3	–

S_{x1}, number of exclusive polymorphisms in the *S. peruvianum* sample (Canta); S_{x2}, number of exclusive polymorphisms in the *S. chilense* sample (Moquegua); S_s, number of shared polymorphisms; S_f, number of fixed diferences (Wakeley and Hey, 1997); multiple hits were eliminated from the dataset. The recombination estimator γ refers to the Canta sample (see text) and is given per site x 10^{-2}.

tion model parameters (Wakeley and Hey, 1997). When the three chosen population samples per species were treated as combined (pooled) samples, there were no fixed differences between these two species at any of the loci (data not shown). However, under such conditions the WH approach cannot yield reliable estimates of the model parameters, as the method requires that a range of possible genealogies is represented in the data (Wakeley and Hey, 1997).

As a first approximation to the model assumption of panmixia, we contrasted all nine combinations of interspecific pairs of populations (for limitations and caveats, see Discussion). Table 2.3 presents the distribution of polymorphic sites for the population pair characterized by the highest within-population level of polymorphism. This locus-by-locus tabulation clearly shows two features of the multilocus data that are found in all interspecific population comparisons: all loci exhibit multiple shared polymorphisms while there are only very few fixed interspecific differences at one or two of the loci (Table 2.3). Table 2.4 summarizes the polymorphic site counts for all nine comparisons, and these data emphasize the ubiquity of shared polymorphisms and the paucity of fixed differences (S_f between two and seven for these interspecific comparisons based on single populations per species).

Table 2.4: Summary of polymorphic site counts in nine interspecific population contrasts (*S. peruvianum-S. chilense*)

Population contrast	Length(bp)	S_{x1}	S_{x2}	S_s	S_f
TAR-TAC	9,186	232	174	74	6
TAR-MOQ	9,178	230	187	74	5
TAR-QUI	9,191	241	182	64	7
NAZ-TAC	9,227	236	175	68	2
NAZ-MOQ	9,208	236	190	67	2
NAZ-QUI	9,195	224	165	78	4
CAN-TAC	9,208	296	160	75	3
CAN-MOQ	9,202	291	175	77	3
CAN-QUI	9,206	291	160	81	4

The four site categories S_{x1}, S_{x2}, S_s, S_f, (see Table 2.3 and text for definitions) have been summarized over all seven loci; multiple hits were eliminated from the dataset.

3.3 Parameter estimates and model fitting

Table 2.5 presents WH parameter estimates for all nine pairwise (interspecific) population comparisons. Across these nine comparisons, the estimates of ancestral population size (θ_A) are roughly comparable to those for the *S. chilense* samples (θ_2), whereas the effective size of *S. peruvianum* (θ_1) is estimated to be larger. The WH parameter estimates have broad confidence intervals (Table 2.5). The means of simulated parameter values are generally very close to the point estimates, except that the simulated θ's for *S. peruvianum* samples are often somewhat higher than the corresponding point estimates (data not shown). A notable result is that the T estimates ($= \tau/\theta_1$) imply very recent divergence from a common ancestor, in the range of 0.11-0.20 units ago (where time is measured in units of $2N_e$*S. peruvianum* generations).

Coalescent simulations implementing the estimated recombination levels did not reject the simple isolation model, as neither the wwh (Wang et al., 1997) nor the χ^2 (Kliman et al., 2000) test statistics approach significance in any of the pairwise interspecific comparisons (Table 2.5). In particular, the generally high values of P_{wwh} reflect the absence of even moderate (let alone drastic) differences in the proportions of S_s and S_f among loci (Table 2.3), while the fairly low P_{χ^2} value for the NAZ-QUI

comparison (Table 2.5) reflects additional deviations due to variable levels of exclusive polymorphisms among loci for these two samples. Summarizing these contrasts between observed data and expectations under the simple WH isolation model, it is evident that our data fit the model surprisingly well; this is mainly due to the fairly even distribution of the polymorphic site classes across loci.

We also ran these analyses on three pooled datasets, which were motivated as follows. Our companion study found that genetic differentiation among the *S. chilense* populations Tacna and Moquegua is very low ($F_{st} < 0.01$; Arunyawat et al. (MS)). In addition, the most polymorphic *S. peruvianum* populations (Canta and Tarapaca; Table 2.2) exhibit the least divergence within that species ($F_{st} \approx 0.09$). Hence, we generated a pooled TAC&MOQ sample and contrasted it with either of the above *S. peruvianum* populations as well as with a pooled TAR&CAN sample. In these interspecific contrasts, the number of fixed differences across all seven loci is between two and five; pooling other (or more) samples eliminates fixed differences entirely, as was observed for the species-wide dataset.

Table 2.5 also presents the WH estimates and goodness-of-fit statistics for these pooled contrasts. As was observed for the single-population comparisons, the demographic estimates imply a roughly stable population size for *S. chilense* compared to the ancestral species, while *S. peruvianum* appears to have undergone a population-size expansion. Likewise, coalescent simulations do not reject the simple isolation model, again reflecting the fairly even among-locus distributions of the polymorphic site classes (data not shown). Compared to the single-population contrasts, variances for the population mutation parameters of the extant taxa (θ_1 and θ_2) tend to be lower (Table 2.5), probably reflecting additional power due to the higher number of sequenced alleles/SNPs.

3.4 Linkage disequilibrium test of historical gene flow

We additionally tested for historical gene flow by analyzing patterns of intragenic LD, using information that remains untapped by the multilocus goodness-of-fit test that is often the only assessment of the suitability of the isolation model (via coalescent simulations). Table 2.6 presents empirical data and simulation results using the LD test statistic x, which ought to exhibit more positive values under a scenario of (historical) gene flow. Five out of six mean observed x values are significantly elevated, with stronger signals seen in the most inclusive sample contrast (TAR&CAN

Table 2.5: WH parameter estimates and isolation model fitting

Pop 1 (*S. peruvianum*)	Pop 2 (*S. chilense*)	θ_1	θ_2	θ_A	τ	T	P_{χ^2}	P_{wwh}
TAR	TAC	121.6	68.9	101.8	22.8	0.187	0.742	0.878
		65.0-298.3	40.7-116.8	58.4-150.3	15.1-30.6	0.08-0.31		
TAR	MOQ	127.5	87.8	96.8	24.9	0.195	0.521	0.903
		70.8.-275.5	51.8-153.4	55.1-145.9	17.0-32.8	0.09-0.32		
TAR	QUI	129.2	71.1	95.4	25.8	0.200	0.911	0.775
		71.4-296.4	44.9-116.8	52.5-144.6	18.1-33.8	0.08-0.32		
NAZ	TAC	150.9	90.2	73.6	23.5	0.156	0.364	0.563
		82.9-257.8	54.5-128.6	45.6-116.4	17.6-30.4	0.09-0.27		
NAZ	MOQ	144.9	113.8	72.7	25.8	0.178	0.483	0.774
		81.3-237.4	65.2-173.0	43.4-120.5	19.0-33.5	0.11-0.32		
NAZ	QUI	108.7	66.5	95.7	19.9	0.184	0.094	0.293
		56.9-220.1	37.6-107.5	56.7-144.5	12.9-26.7	0.09-0.30		
CAN	TAC	222.9	72.9	87.1	24.6	0.110	0.933	0.649
		110.2-591.4	45.5-106.3	51.3-136.1	18.1-31.6	0.04-0.21		
CAN	MOQ	198.7	89.5	88.6	26.6	0.134	0.898	0.722
		108.3-413.4	55.8-135.4	52.5-137.2	19.6-33.9	0.06-0.24		
CAN	QUI	186.4	65.3	99.2	23.2	0.124	0.395	0.439
		94.9-508.0	40.1-100.3	57.4-149.0	16.4-30.1	0.04-0.22		
TAR	TAC&MOQ	115.4	91.3	102.5	26.6	0.231	0.521	0.853
		68.7-218.3	62.0-135.4	60.8-151.5	19.7-34.1	0.12-0.36		
CAN	TAC&MOQ	191.7	97.5	86.4	28.1	0.147	0.882	0.732
		111.9-313.5	68.6-125.7	55.7-132.3	22.4-35.2	0.09-0.26		
TAR&CAN	TAC&MOQ	203.7	85.8	99.2	27.9	0.137	0.755	0.789
		130.6-316.1	60.8-108.5	61.7-150.6	22.5-35.4	0.09-0.22		

For each of the 12 interspecific comparisons, the data (cf. Table 2.4) were fitted to the WH isolation model. θ_1, population mutation parameter for population 1 ($= 4N_e\mu$, estimated over all sites); θ_2, population mutation parameter for population 2; θ_A, population mutation parameter for the ancestral species; τ, scaled time parameter ($=2\mu t$). T ($= \tau/\theta$) is the estimated time of species divergence in units of $2N_1$ generations, where N_1 is the effective size of population 1. Below the primary parameter estimates, 95% confidence intervals are shown, determined by 10,000 coalescent simulations. The P-values for both the wwh (Wang et al. 1997) and χ^2 test statistics are the proportions of simulated values $\geq$ the observed values. The bottom three comparisons contrast pooled samples; within-species pooling was conditioned on (still) observing fixed differences in the multilocus data and low population differentiation (see text)

vs. TAC&MOQ) and the allopatric contrast CAN *vs.* TAC&MOQ. Interestingly, locus CT066 exhibits significantly high x values for all samples, and locus CT166 shows significant or nearly significant x values for many of the contrasts; there is a tendency for this signal to be stronger in *S. peruvianum*. Finally, there are two marginally significant x values for locus CT179 in *S. chilense*, whereas there is no LD signal of gene flow at any other locus (Table 2.6).

It should be noted that the three pooled-sample contrasts examined in Table 2.6 do not represent independent datasets; the LD test partly rests on a relational framework, in that shared polymorphisms can only be identified in reference to a second sample (population, species, *etc.*), while exclusive polymorphisms are unique to particular populations or more inclusive samples. In addition to the three pooled-sample comparisons, we also subjected the individual population contrasts to the LD test of (historical) introgression. Figure 2.1 plots the P-values for individual loci in both species, expressed as medians of the individual population contrasts. Consistent with the results for the pooled-sample comparisons, median P-values for loci CT066 and CT166 are between 0.030 and 0.124, reflecting (often) significant or marginally significant values of the LD test statistic x in many single-population contrasts. These results suggest bidirectional interspecific gene flow following initial species divergence, whereas there is no compelling evidence for such a scenario at the other loci (Figure 2.1; but see Discussion).

Table 2.6: Linkage disequilibrium test of historical gene flow

Sample A (*peru*) Sample B (*chil*)	CT066	CT093	CT166	CT179	CT198	CT251	CT268	Observed mean x	Simulated means x
A: TAR&CAN *vs.*	0.751	0.203	0.396	0.179	0.272	0.152	0.123	0.297	0.125
	0.016	0.263	**0.068**	0.340	0.160	0.283	0.402	**0.022**	
B: TAC&MOQ	0.673	0.239	0.244	0.470	0.151	-0.055	0.000	0.246	0.072
	0.023	0.167	0.122	**0.068**	0.248	0.858	0.682	**0.034**	
A: TAR *vs.*	0.650	0.268	0.428	0.224	0.071	0.025	0.090	0.251	0.086
	0.031	0.182	0.092	0.263	0.417	0.605	0.351	**0.049**	
B: TAC&MOQ	0.662	-0.110	0.339	0.192	0.089	-0.056	0.109	0.175	0.086
	0.028	0.806	0.127	0.291	0.367	0.868	0.308	0.142	
A: CAN *vs.*	0.863	0.000	0.619	0.136	0.370	0.100	0.019	0.301	0.108
	0.014	0.665	**0.030**	0.351	0.124	0.399	0.714	**0.027**	
B: TAC&MOQ	0.615	0.299	0.482	0.470	0.192	-0.070	-0.075	0.273	0.081
	0.032	0.151	**0.049**	**0.053**	0.246	0.871	0.864	**0.029**	

For each of the three pooled-sample contrasts (*S. peruvianum* vs. *S. chilense*), the observed LD test statistic x is given in the first line (for each locus and each of six samples), with the proportion of successful coalescent simulations $\geq$ the observed values of x (P-values) immediately below; P-values <0.07 are highlighted in boldface. The last two columns summarize the observed mean x (with P-values below) and the simulated mean x (assuming no interspecific gene flow) over all loci (see text).

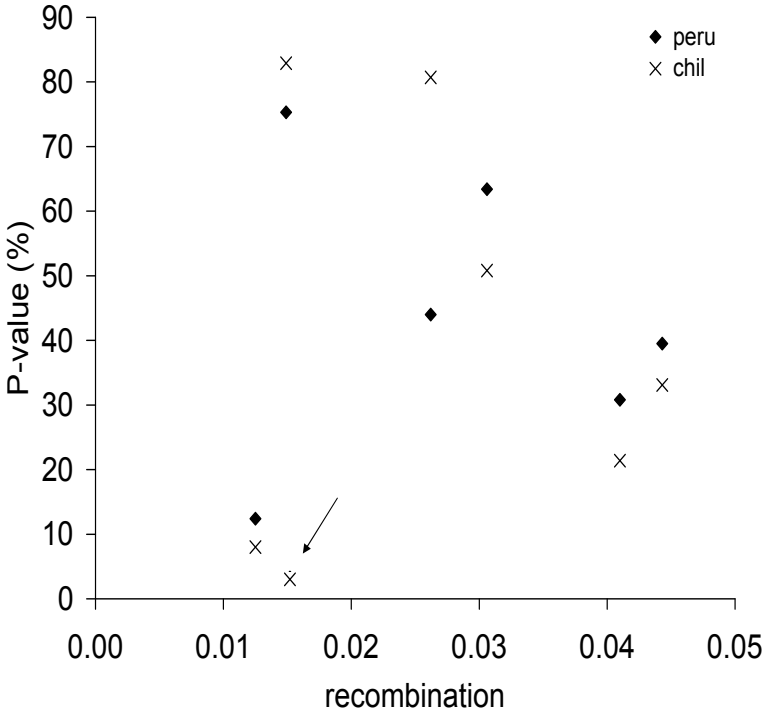

Figure 2.1: Scatter plot of average locus-specific rates of recombination and median P-values for single-population contrasts (LD test of historical gene flow). Recombination is expressed as γ per site, and the locus-specific values are the means obtained for the *S. peruvianum* populations (*i.e.*, both species are plotted with the *S. peruvianum* recombination estimates). The P-values are locus-specific medians obtained from up to nine interspecific population contrasts. The arrow highlights a 'hidden' P-value of 0.033 for locus CT066 in *S. peruvianum*. Note that this lower left area of the plot contains the data for loci CT066 and CT166 (cf. Table 2.6).

4 Discussion

The principal limitation of our previous study of speciation scenarios in wild tomatoes was the availability of only single populations per species, leading to uncertainty about the generality of our initial findings (Städler et al., 2005). Here we have shown that consistent demographic estimates under the WH isolation model (Wakeley and Hey, 1997; Wang et al., 1997) are obtained when 'typical' populations of both *S. peruvianum* and *S. chilense* are chosen to portray their respective genealogical histories, where 'typical' refers to samples exhibiting broadly comparable levels and partitioning of nucleotide polymorphism across multiple loci. For example, our original *S. chilense* sample (Antofagasta) can now be interpreted as reflecting a local or regional bottleneck (based on unusually low levels of polymorphism and strong haplotype structure) that is certainly not characteristic of the species-wide demographic history following speciation. Consequently, relying on such samples would result in misleading WH parameter estimates, as was anticipated in our previous study (Städler et al., 2005). Whereas population-size contraction in *S. chilense* compared to the ancestral species was inferred in that study, our current demographic estimates based on three other populations concur that historical effective population size has remained fairly constant, while there are consistent signatures of larger effective population size in *S. peruvianum* (Tables 2.2 - 2.5).

4.1 Consequences of population subdivision and sampling scheme

We caution that the demographic estimates summarized in Table 2.5 may not adequately capture the actual historical demography of these recently diverged species. Coalescent models for subdivided species are a useful guide to interpret the demographic inferences and to identify possible biases of our analyses under the WH framework. Wakeley (1999) introduced the distinction between the 'scattering' and the 'collecting' phase of the coalescent process in a subdivided population (island model). The brief scattering phase traces the genealogy of a local sample until all remaining lineages (looking backward in time) are located in different demes, and it is characterized by local coalescent events and migration events to different demes. The timescale of the ensuing, much longer collecting phase depends on the rate of migration between demes and the number and size of demes in the total population,

in that ancestral lineages can only coalesce when they occupy the same deme (Wakeley, 1999, 2001; Wakeley and Aliacar, 2001). Given that we have sampled multiple gene copies per population, we have to consider effects of the scattering phase and the potential non-exchangeability of the sampled sequences.

Our 'typical' local samples do contain much of the species-wide diversity, as reflected in the high local estimates of nucleotide diversity (Table 2.2) and moderate estimates of population differentiation ($F_{st} \approx 0.15$; Arunyawat et al. (MS)). This implies that local coalescent events during the scattering phase are fairly rare, or in other words, that the number of ancestral lineages at the beginning of the collecting phase is fairly close to its maximal value (the size n of the local sample). In this regard, some of our local samples almost approximate a species-wide sample where single sequences are obtained from a larger number of demes. However, we have noted that pooling of more than two populations per species eliminates all fixed interspecific differences, while single-population comparisons and selected pooled-sample contrasts show between two and seven fixed differences (Table 2.4, and Results). We thus 'see' the expected effects of underestimating the within-species diversity using single-deme samples, as well as underestimating divergence between species when considering species-wide samples (Wakeley, 2000, 2003; Ingvarsson, 2004); both are consequences of the distribution of diversity within and between local demes under population subdivision and the (potentially) higher species-wide effective population size under restricted migration (Slatkin, 1987; Strobeck, 1987; Whitlock and Barton, 1997; Pannell, 2003). A related aspect involves the dependence of widely used basic statistics evaluating the site frequency spectrum (such as Tajima's D) and the expected number of distinct haplotypes (such as FU's F_s) on random mating (*i.e.*, absence of substructure) within populations from which samples are drawn. Analyzing patterns of diversity under a scenario of spatial (range) expansion in a two-dimensional stepping-stone model, Ray et al. (2003) found that levels of inter-deme migration may have substantial effects on patterns of diversity and the site frequency spectrum. In particular, samples taken from single demes may completely miss the signatures of species-wide spatial expansion, in that D and F_s fail to exhibit deviations from neutral equilibrium expectations under low-to-moderate levels of gene flow. These effects have been extended to the infinite-island model (Excoffier, 2004) and appear to be exacerbated under temporal and/or spatial environmental heterogeneity (Wegmann et al., 2006).

Because the ability to quantitatively distinguish between a pure demographic expansion and a range expansion in a subdivided population hinges on future theoretical work, we cannot rule out a scenario of range expansion (with concomitant increase in species-wide population size) of one or both species following their recent divergence. Although our WH analyses yield demographic parameter estimates for the extant species that are consistent with the site frequency spectra for individual population samples (Tables 2.2, 2.5), they may be biased due to the complexities of the coalescent process in subdivided populations (discussed above), in concert with the sampling scheme and unrealistic assumptions of the isolation model. Likewise, the estimated divergence time ($\approx 0.28\ N_e$ generations ago for the most inclusive interspecific contrast; Table 2.5) may be upwardly biased because the full interspecific multilocus sequence comparison revealed the absence of fixed differences. This lack of fixed differences implies very recent divergence but also may reflect the effects of population structure, as accruing fixed differences is expected to take longer under subdivision compared to divergence under panmixia within sister species (Wakeley, 2000). Given these complexities, there clearly is a need for more realistic models of speciation that explicitly take population subdivision into account in extracting signals of species' demographic history from multilocus sequence data.

4.2 Fit of the isolation model

Despite the fact that the genealogical histories of our samples violate the model's assumption of within-species panmixia, we found that the data fit the simple isolation model quite well, which is primarily a consequence of observing similar proportions of exclusive polymorphisms, shared polymorphisms and fixed differences across loci (Tables 2.4, 2.5). An evaluation of multilocus studies using the WH approach suggests that only very recent or current introgression may allow formal rejection of the isolation model, making this a very conservative test. For example, several multilocus DPG studies were able to reject the isolation model due to large differences in patterns of shared polymorphisms and fixed differences among loci, a signature attributed to recent interspecific gene flow at some (but not all) regions of the genome (Machado et al., 2002; Besansky et al., 2003). Such among-locus signatures are consistent with what is known about the genetic architecture of incomplete postzygotic reproductive isolation, and with inferences drawn from numerous studies on natural hybrid zones (Barton and Hewitt, 1985; Rieseberg et al., 1999; Coyne and Orr, 2004).

4.3 Sources of shared polymorphisms and signatures of historical gene flow

In addition to reflecting truly ancestral mutations as envisaged under the isolation model, shared polymorphisms between recently diverged taxa can arise through introgression subsequent to species divergence, a biologically plausible process under parapatric speciation or upon secondary contact after some divergence in allopatry. One approach that may be informative about the historical source of mutations yielding shared polymorphisms among species uses the proportion of shared polymorphisms among all segregating sites. For the pooled *S. chilense* sample TAC&MOQ, this proportion is 80/348 (= 0.230) in the contrast with the sympatric *S. peruvianum* sample TAR, whereas it is 83/336 (= 0.247) when compared to the allopatric CAN sample. Inspection of Table 2.4 reveals that these proportions are generally higher in single-population comparisons. Similar patterns hold when the focal species is *S. peruvianum*: in the contrast with the *S. chilense* sample TAC&MOQ, this proportion is 80/302 (= 0.265) for the sympatric *S. peruvianum* sample TAR, whereas it is 83/363 (= 0.229) for the allopatric CAN sample. This latter, somewhat lower proportion, however, is unlikely to be due to consequences of allopatry but instead reflects the higher level of (exclusive) polymorphism in CAN (Table 2.4). Overall, these patterns indicate that shared polymorphisms tend to be geographically widespread in both species (whereas many exclusive polymorphisms are geographically restricted and overall rare), consistent with them reflecting ancestral mutations. These signatures appear to make a scenario of very recent introgression in areas of current sympatry unlikely, which is in agreement with the strong postzygotic barriers revealed in crossing experiments (Rick and Lamm, 1955; Rick, 1979, 1986; Städler et al., 2005); and see below.

A second approach for probing the historical genesis of shared polymorphisms among recently diverged species is implemented in the LD-based test of gene flow (Machado et al., 2002). Using coalescent simulations, we found significantly elevated mean values of the LD test statistic x in five out of six pooled-population contrasts (Table 2.6), as well as in several single-population comparisons (data not shown). More importantly, two loci exhibit consistently high x values across comparisons, reflecting stronger LD for a subset of intragenic LD than expected if all shared polymorphisms were truly ancestral mutations (Figure 2.1). We argue that, in general, nonsignificant mean values of the LD test statistic are less informative than the data

for single loci, because a significantly high x at one or more loci might be offset by low values at other loci, conceivably yielding a nonsignificant mean x. Under this rationale, there appears to be strong evidence for bidirectional (historical) gene flow at loci CT066 and CT166. These two loci were among those suggestive of interspecific gene flow in our initial study using single populations per species (Städler et al., 2005). Importantly, this genealogical signal is not restricted to regions of current sympatry, as best seen in the contrasts involving the allopatric CAN population (Table 2.6). The S_s/S_{total} ratios discussed above are in excellent agreement with this geographically dispersed signature of historical interspecific gene flow.

Figure 2.1 displays the relationship between LD-based signatures of gene flow and estimates of the population recombination parameter γ across loci. As a group, the two loci with low P-values (*i.e.*, high observed values of the test statistic x) are characterized by an average level of recombination that is about 50% of that for the five other loci, although there is a lot of scatter for this latter group. Very similar proportions hold when recombination is estimated with Hudson's composite-likelihood method, as implemented in the package LDhat (Hudson, 2001; McVean et al., 2002) (data not shown). Hypothetically, if interspecific gene flow ceased simultaneously across the genome at some point in time, we would expect that the LD-based signal for gene flow is maintained longer in regions of low recombination. For a range of recombination levels as represented across the studied loci (Figure 2.1), we might then expect a positive correlation between P-values and levels of recombination. However, many studies concur that the initial build-up of reproductive isolation is likely to involve a few loci or genomic regions (with those regions being protected from introgression), while most of the genome remains permeable to interspecific gene flow for much longer (Rieseberg et al., 1999; Osada and Wu, 2005; Turner et al., 2005; Patterson et al., 2006; Machado et al., 2007). Our LD-based results are thus fully compatible with expectations for regions of low recombination (given that post-divergence introgression has actually occurred), and they are consistent with a suite of observations on the nature of postzygotic reproductive barriers and their consequences for genomic divergence.

Our evidence for post-divergence gene flow between these taxa implies that *some* of the polymorphisms shared between *S. peruvianum* and *S. chilense* do not represent genuine ancestral mutations. Unlike some other large-scale DPG studies that found sharing of entire haplotypes between species with partially overlapping ranges (Machado et al., 2002; Besansky et al., 2003; Ramos-Onsins et al., 2004), our evidence for introgression is much more subtle and simultaneously more difficult to uncover. The implications of having found evidence for a divergence-with-gene-flow model of speciation are, perhaps, also more interesting than for species that continue to exchange genes. In the wild tomato taxa under study, speciation has resulted in genomes that appear to be fully isolated from each other despite the extremely low level of molecular divergence (indicative of recent speciation), as discussed in the next section.

4.4 Implications of patterns of postzygotic reproductive isolation

Most accessions of *S. peruvianum* that have been tested are isolated from *S. chilense* by strong intrinsic postzygotic incompatibilities, in that usually high incidences ($\approx 97\%$) of embryonic breakdown were observed in experimental crosses (Rick and Lamm, 1955; Rick, 1979, 1986). These results have been obtained under non-competitive interspecific pollination, and it is thus unknown if post-pollination or other prezygotic barriers contribute to reproductive isolation between these species in sympatry. Our previous study identified intriguing patterns of postzygotic incompatibility in terms of both geography and developmental failure, as synthesized from data and descriptions in the original studies (cited above; Städler et al. (2005)). Briefly, the postzygotic barrier is strongest in regions of sympatry and further south, where only *S. chilense* occurs. Experimental crosses using *S. peruvianum* s.l. accessions from northern Peru, however, yielded partially fertile F_1 hybrids in appreciable frequencies (Rick and Lamm, 1955; Rick, 1986). Most of these northern Peruvian accessions have been proposed to constitute the novel species *S. arcanum* (Peralta et al., 2005; Spooner et al., 2005), but the striking feature of the strongest reproductive barriers concurrent with sympatry and low molecular divergence (*i.e.*, sister species) remains. These combined biological facets are indicative of reinforcement of reproductive isolation that in the case of postzygotic barriers to interspecific gene flow might be mediated by direct selection for hybrid inviability (Grant, 1966; Coyne,

1974; Wallace, 1988; Coyne and Orr, 2004).

Both the described hybrid embryonic breakdown and the fact that crosses between the cultivated tomato (*S. lycopersicum*) as the maternal parent and both *S. peruvianum* and *S. chilense* result in viable hybrids only under embryo culture, implicate endosperm-embryo imbalances in the usual failure of such crosses (Rick and Lamm, 1955; Rick, 1986; Städler et al., 2005). From a mechanistic point of view, the seemingly very rapid build-up of postzygotic barriers between *S. peruvianum* and *S. chilense* might have been caused by changes in any of a number of genes instrumental in endosperm function. Embryo inviability (hybrid seed failure) is a common interspecific barrier in angiosperms, and there is substantial evidence that such barriers involve imbalances in endosperm-embryo interactions early in seed development (Cooper and Brink, 1942; Lester and Kang, 1998; Bushell et al., 2003; Gutiérrez and Meserve, 2003; Gehring et al., 2004).

These provisional interpretations account for the rapidity (and possibly very local genomic signature) of the evolution of near-complete postzygotic isolation and imply the involvement of natural selection as one of the forces governing species divergence. As with all studies using a DPG approach, the inference of natural selection in the history of species divergence rests on evidence for non-allopatric speciation (*i.e.*, evidence for post-divergence gene flow) and thus is necessarily an indirect one. Complementary approaches will be required to demonstrate (past) natural selection on genes or genomic regions directly involved in important trait differences or reproductive incompatibility among species, as discussed elsewhere (Städler et al., 2005; Noor and Feder, 2006).

5 Conclusion

Our extensive study of the two closely related wild tomato species *S. peruvianum* and *S. chilense* has uncovered evidence for species divergence under residual gene flow. The geographically dispersed signature of post-divergence gene flow (*i.e.*, not restricted to regions of current sympatry) is consistent with historical introgression and subsequent spread through much of the species' ranges, either via range expansions or intraspecific gene flow. More generally, historical introgression points to a parapatric mode of speciation or at least requires a period of secondary contact, during which natural selection was instrumental in completing reproductive isolation. The demographic estimates under the WH isolation model imply population (or range) expansion for *S. peruvianum* and an effective size for *S. chilense* similar to that of the ancestral species. From the outset, however, the WH assumption of random mating within species was known to be violated for our dataset, as it is for any such study in subdivided species regardless of the sampling scheme. Because of the complexities of the coalescent process in subdivided populations, the WH estimates of both historical demography and the time since speciation may be biased. To fully exploit the genealogical information contained in our sequence data, one would need an explicit model allowing for population subdivision in both the ancestral and the extant species, with interspecific gene flow subsequent to divergence included as parameters to be estimated from the data.

Bibliography

Arnold, M. L. (1997). *Natural Hybridization and Evolution*. Oxford University Press, New York.

Arunyawat, U., Stephan, W., and Städler, T. (MS). Using multilocus sequence data to assess population structure, natural selection and linkage disequilibrium in wild tomatoes. *manuscript*.

Barton, N. H. and Hewitt, G. M. (1985). Analysis of hybrid zones. *Annu Rev Ecol Syst*, 16:113–148.

Baudry, E., Kerdelhué, C., Innan, H., and Stephan, W. (2001). Species and recombination effects on DNA variability in the tomato genus. *Genetics*, 158:1725–1735.

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

Besansky, N. J., Krzywinski, J., Lehmann, T., Simard, F., Kern, M., Mukabayire, O., Fontenille, D., Touré, Y., and Sagnon, N. (2003). Semipermeable species boundaries between *Anopheles gambiae* and *Anopheles arabiensis*: evidence from multilocus DNA sequence variation. *Proc Natl Acad Sci U S A*, 100:10818–10823.

Broughton, R. E. and Harrison, R. G. (2003). Nuclear gene genealogies reveal historical, demographic and selective factors associated with speciation in field crickets. *Genetics*, 163:1389–1401.

Brown, G. R., Gill, G. P., Kuntz, R. J., Langley, C. H., and Neale, D. B. (2004). Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc Natl Acad Sci U S A*, 101:15255–15260.

Bushell, C., Spielman, M., and Scott, R. J. (2003). The basis of natural and artificial postzygotic hybridization barriers in *Arabidopsis* species. *Plant Cell*, 15:1430–1442.

Clauss, M. J. and Mitchell-Olds, T. (2006). Population genetic structure of *Arabidopsis lyrata* in europe. *Mol Ecol*, 15:2753–2766.

Cooper, D. C. and Brink, R. A. (1942). The endosperm as a barrier to interspecific hybridization in flowering plants. *Science*, 95:75–76.

Coyne, J. A. (1974). The evolutionary origin of hybrid inviability. *Evolution*, 28:505–506.

Coyne, J. A. and Orr, H. A. (2004). *Speciation*. Sinauer Associates, Sunderland, MA.

Das, A., Mohanty, S., and Stephan, W. (2004). Inferring the population structure and demography of *Drosophila ananassae* from multilocus data. *Genetics*, 168:1975–1985.

Devries, T. J. (1987). A review of geological evidence for ancient El Niño activity in Peru. *J Geophys Res*, 92:14471–14479.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*, 35:125–129.

Dolman, G. and Moritz, C. (2006). A multilocus perspective on refugial isolation and divergence in rainforest skinks (*Carlia*). *Evolution*, 60:573–582.

Edwards, S. V. and Beerli, P. (2000). Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, 54:1839–1854.

Endler, J. A. (1977). *Geographic Variation, Speciation, and Clines*. Princeton University Prss, Princeton NJ.

Evans, P. D., Mekel-Bobrov, N., Vallender, E. J., Hudson, R. R., and Lahn, B. T. (2006). Evidence that the adaptive allele of the brain size gene *microcephalin* introgressed into *Homo sapiens* from an archaic *Homo* lineage. *Proc Natl Acad Sci U S A*, 103:18178–18183.

Excoffier, L. (2004). Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Mol Ecol*, 13:853–864.

Excoffier, L., Laval, G., and 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., and 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.

Fay, J. C. and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155:1405–1413.

Flint-Garcia, S. A., Thornsberry, J. M., and Buckler, E. S. (2003). Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol*, 54:357–374.

Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147:915–925.

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

Ganal, M. W., Czihal, R., Hannappel, U., Kloos, D. U., Polley, A., and Ling, H. Q. (1998). Sequencing of cDNA clones from the genetic map of tomato (*Lycopersicon esculentum*). *Genome Res*, 8:842–847.

Garris, A. J., McCouch, S. R., and Kresovich, S. (2003). Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.). *Genetics*, 165:759–769.

Gaut, B. S. and Clegg, M. T. (1993). Molecular evolution of the *Adh1* locus in the genus zea. *Proc Natl Acad Sci U S A*, 90.

Gavrilets, S. (2003). Perspective: models of speciation: what have we learned in 40 years? *Evolution*, 57:2197–2215.

Gehring, M., Choi, Y., and Fischer, R. L. (2004). Imprinting and seed development. *Plant Cell*, 16 Suppl:203–213.

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

Gottlieb, L. D. (1977). Electrophoretic evidence and plant systematics. *Annls Missouri Bot Gard*, 64:161–180.

Grant, V. (1966). The selective origin of incompatibility barriers in the plant genus *Gilia*. *Am Nat*, 100:99–118.

Gupta, P. K., Rustgi, S., and Kulwal, P. L. (2005). Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol Biol*, 57:461–485.

Gutiérrez, J. R. and Meserve, P. L. (2003). El Niño effects on soil seed bank dynamics in north-central Chile. *Oecologia*, 134:511–517.

Hamrick, J. L. and Godt, M. J. W. (1989). Allozyme diversity in plant species. In *Plant population genetics, breeding and germplasm resources*, pages 43–63, Sunderland, MA. Sinauer Associates.

Hamrick, J. L. and Godt, M. J. W. (1996). Effects of life history traits on genetic diversity in plant species. *Phil Trans R Soc Lond B*, 351:1291–1298.

Helgason, A., Yngvadóttir, B., Hrafnkelsson, B., Gulcher, J., and Stefánsson, K. (2005). An icelandic example of the impact of population structure on association studies. *Nat Genet*, 37:90–95.

Heuertz, M., De Paoli, E., Källman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M., and Gyllenstrand, N. (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of norway spruce (*Picea abies* (L.) Karst). *Genetics*, 174:2095–2105.

Hey, J. and Kliman, R. M. (1993). Population genetics and phylogenetics of dna sequence variation at multiple loci within the *Drosophila melanogaster* species complex. *Mol Biol Evol*, 10:804–822.

Hey, J. and Machado, C. A. (2003). The study of structured populations–new hope for a difficult and divided science. *Nat Rev Genet*, 4:535–543.

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

Hey, J. and Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics*, 145:833–846.

Hill, W. G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet*, 38:226–231.

Hill, W. G. and Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theor Popul Biol*, 33:54–78.

Hoskin, C. J., Higgie, M., McDonald, K. R., and Moritz, C. (2005). Reinforcement drives rapid allopatric speciation. *Nature*, 437:1353–1356.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surv Evol Biol*, 7:1–44.

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

Hudson, R. R. and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–164.

Hudson, R. R., Kreitman, M., and Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, 116:153–159.

Hudson, R. R., Slatkin, M., and Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132:583–589.

Hudson, R. R. and Turelli, M. (2003). Stochasticity overrules the "three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. *Evolution*, 57:182–190.

Ingvarsson, P. K. (2004). Population subdivision and the Hudson-Kreitman-Aguadé test: testing for deviations from the neutral model in organelle genomes. *Genet Res*, 83:31–39.

Ingvarsson, P. K. (2005). Nucleotide polymorphism and linkage disequilibrium within and among natural populations of european aspen (*Populus tremula* L., *Salicaceae*). *Genetics*, 169:945–953.

Ingvarsson, P. K., García, M. V., Hall, D., Luquez, V., and Jansson, S. (2006). Clinal variation in *phyB2*, a candidate gene for day-length-induced growth cessation and bud set, across a latitudinal gradient in european aspen (*Populus tremula*). *Genetics*, 172:1845–1853.

Kane, N. C. and Rieseberg, L. H. (2007). Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, *Helianthus annuus*. *Genetics*.

Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics*, 146:1197–1206.

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

Kliman, R. M., Andolfatto, P., Coyne, J. A., Depaulis, F., Kreitman, M., Berry, A. J., McCarter, J., Wakeley, J., and Hey, J. (2000). The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics*, 156:1913–1931.

Knowles, L. L. and Carstens, B. C. (2007). Estimating a geographically explicit model of population divergence. *Evolution*, 61:477–493.

Kraakman, A. T., Niks, R. E., Van den Berg, P. M., Stam, P., and Van Eeuwijk, F. A. (2004). Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics*, 168:435–446.

Kronforst, M. R., Young, L. G., Blume, L. M., and Gilbert, L. E. (2006). Multilocus analyses of admixture and introgression among hybridizing *Heliconius* butterflies. *Evolution*, 60:1254–1268.

Laporte, V. and Charlesworth, B. (2002). Effective population size and population subdivision in demographically structured populations. *Genetics*, 162:501–519.

Lester, R. N. and Kang, J. H. (1998). Embryo and endosperm function and failure in *Solanum* species and hybrids. *Ann Bot*, 82:445–453.

Levin, D. A. (1990). The seed bank as a source of genetic novelty in plants. *Am Nat*, 135:563–572.

Li, W. H. (1997). *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Lin, J. Z., Brown, A. H., and Clegg, M. T. (2001). Heterogeneous geographic patterns of nucleotide sequence diversity between two alcohol dehydrogenase genes in wild barley (*Hordeum vulgare* subspecies *spontaneum*). *Proc Natl Acad Sci U S A*, 98:531–536.

Liu, A. and Burke, J. M. (2006). Patterns of nucleotide diversity in wild and cultivated sunflower. *Genetics*, 173:321–330.

Long, A. D., Lyman, R. F., Langley, C. H., and Mackay, T. F. (1998). Two sites in the delta gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. *Genetics*, 149:999–1017.

Losos, J. B. and Glor, R. E. (2003). Phylogenetic comparative methods and the geography of speciation. *Trends Ecol Evol*, 18:220–227.

Ma, X. F., Szmidt, A. E., and Wang, X. R. (2006). Genetic structure and evolutionary history of a diploid hybrid pine *Pinus densata* inferred from the nucleotide variation at seven gene loci. *Mol Biol Evol*, 23:807–816.

Machado, C. A., Haselkorn, T. S., and Noor, M. A. (2007). Evaluation of the genomic extent of effects of fixed inversion differences on intraspecific variation and interspecific gene flow in *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, 175:1289–1306.

Machado, C. A., Kliman, R. M., Markert, J. A., and Hey, J. (2002). Inferring the history of speciation from multilocus `DNA` sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Mol Biol Evol*, 19:472–488.

Maddison, W. P. and Maddison, D. R. (1992). *Macclade3: Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Sunderland, MA.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends Ecol Evol*, 20:229–237.

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

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

Mayr, E. (1963). *Animal Species and Evolution*. Harvard University Press, Cambridge, MA.

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

Miller, J. C. and Tanksley, S. D. (1990). `RFLP` analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. *Theor Appl Genet*, 80:437–448.

Miyashita, N. T., Innan, H., and Terauchi, R. (1996). Intra- and interspecific variation of the alcohol dehydrogenase locus region in wild plants *Arabis gemmifera* and *Arabidopsis thaliana*. *Mol Biol Evol*, 13:433–436.

Morjan, C. L. and Rieseberg, L. H. (2004). How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Mol Ecol*, 13:1341–1356.

Morrell, P. L., Lundy, K. E., and Clegg, M. T. (2003). Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite migration. *Proc Natl Acad Sci U S A*, 100:10812–10817.

Morrell, P. L., Toleno, D. M., Lundy, K. E., and Clegg, M. T. (2005). Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. *Proc Natl Acad Sci U S A*, 102:2442–2447.

Nei, M. (1987). *Molecular Evolutianry Genetics*. Columbia University Press, New York.

Nielsen, R. and Wakeley, J. (2001). Distinguishing migration from isolation: a `Markov` chain `Monte Carlo` approach. *Genetics*, 158:885–896.

Noor, M. A. and Feder, J. L. (2006). Speciation genetics: evolving approaches. *Nat Rev Genet*, 7:851–861.

Nordborg, M., Borevitz, J. O., Bergelson, J., Berry, C. C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J. N., Noyes, T., Oefner, P. J., Stahl, E. A., and Weigel, D.

(2002). The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet*, 30:190–193.

Nordborg, M. and Donnelly, P. (1997). The coalescent process with selfing. *Genetics*, 146:1185–1195.

Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N. A., Shah, C., Wall, J. D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M., and Bergelson, J. (2005). The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol*, 3:1289–1299.

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

Olmstead, R. G., Sweere, J. A., Spangler, R. E., Bohs, L., and Palmer, J. D. (1999). Phylogeny and provisional classification of the *Solanaceae* based on chloroplast `DNA`. In *Solanaceae IV, Advances in Biology and Utilization*, pages 111–137, Kew. Royal Botanical Gardens.

Olsen, K. M. and Purugganan, M. D. (2002). Molecular evidence on the origin and evolution of glutinous rice. *Genetics*, 162:941–950.

Ometto, L., Glinka, S., De Lorenzo, D., and Stephan, W. (2005). Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of `DNA` variation. *Mol Biol Evol*, 22:2119–2130.

Osada, N. and Wu, C. I. (2005). Inferring the mode of speciation from genomic data: a study of the great apes. *Genetics*, 169:259–264.

Ostrowski, M. F., David, J., Santoni, S., McKhann, H., Reboud, X., Le Corre, V., Camilleri, C., Brunel, D., Bouchez, D., Faure, B., and Bataillon, T. (2006). Evidence for a large-scale population structure among accessions of *Arabidopsis thaliana*: possible causes and consequences for the distribution of linkage disequilibrium. *Mol Ecol*, 15:1507–1517.

Pannell, J. R. (2003). Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution*, 57:949–961.

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

Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441:1103–1108.

Peralta, I. E., Knapp, S., and Spooner, D. M. (2005). New species of wild tomatoes (*Solanum* section *Lycopersicon*: Solanaceae) from northern Peru. *Systematic Botany*, 30:424–434.

Peralta, I. E. and Spooner, D. M. (2001). Granule-bound starch systhase (*GBSSI*) gene phylogeny of wild tomatoes (*Solanum* L. section *Lycopersicon* [Mill.] Wettst. subsection *Lycopersicon). American Journal of Botany*, 88:1888–1902.

Purugganan, M. D. and Suddith, J. I. (1998). Molecular population genetics of the *Arabidopsis* cauliflower regulatory gene: nonneutral evolution and naturally occurring variation in floral homeotic function. *Proc Natl Acad Sci U S A*, 95:8130–8134.

Rafalski, A. and Morgante, M. (2004). Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends Genet*, 20:103–111.

Ramos-Onsins, S. E., Stranger, B. E., Mitchell-Olds, T., and Aguadé, M. (2004). Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. *Genetics*, 166:373–388.

Ray, N., Currat, M., and Excoffier, L. (2003). Intra-deme molecular diversity in spatially expanding populations. *Mol Biol Evol*, 20:76–86.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, 411:199–204.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., Kresovich, S., Goodman, M. M., and Buckler, E. S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci U S A*, 98:11479–11484.

Rice, W. R. and Hostert, E. E. (1993). Laboratory experiments on speciation: what have we learned in forty years? *Evolution*, 47:1637–1653.

Rick, C. M. (1979). Biosystematic studies in *Lycopersicon* and closely related species of *Solanum*. In *The Biology and Taxonomy of the Solanaceae, edited by Hawkes, J. G., Lester, R. N. and Skelding, A. D.*, pages 667–678, New York. Academic Press.

Rick, C. M. (1986). Reproductive isolation in the *Lycopersicon peruvianum* complex. In *Solanaceae: Biology and Systematics, edited by D'Arcy, W. G.*, pages 477–495, New York. Columbia University Press.

Rick, C. M. and Lamm, R. (1955). Biosystematic studies on the status of *Lycopersicon chilense*. *Am J Bot*, 42:663–675.

Ridley, M. (2003). *Evolution*. Blackwell Publishing, Oxford.

Rieseberg, L. H., Church, S. A., and Morjan, C. L. (2004). Intregation of populations and differentiation of species. *New Phytol*, 161:59–69.

Rieseberg, L. H., Whitton, J., and Gardner, K. (1999). Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, 152:713–727.

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

Rosenberg, N. A. and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet*, 3:380–390.

Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X., and Rozas, R. (2003). Dna`SP`, `DNA` polymorphism analyses by the coalescent and other methods. *Bioinformatics*, 19:2496–2497.

Santiago, E. and Caballero, A. (2005). Variation after a selective sweep in a subdivided population. *Genetics*, 169:475–483.

Savolainen, O., Langley, C. H., Lazzaro, B. P., and Fréville, H. (2000). Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. *Mol Biol Evol*, 17:645–655.

Savolainen, V., Anstett, M. C., Lexer, C., Hutton, I., Clarkson, J. J., Norup, M. V., Powell, M. P., Springate, D., Salamin, N., and Baker, W. J. (2006). Sympatric speciation in palms on an oceanic island. *Nature*, 441:210–213.

Schmid, K. J., Ramos-Onsins, S., Ringys-Beckstein, H., Weisshaar, B., and Mitchell-Olds, T. (2005). A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics*, 169:1601–1615.

Schmid, K. J., Törjék, O., Meyer, R., Schmuths, H., Hoffmann, M. H., and Altmann, T. (2006). Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers. *Theor Appl Genet*, 112:1104–1114.

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends Ecol Evol*, 19:198–207.

Sharbel, T. F., Haubold, B., and Mitchell-Olds, T. (2000). Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of europe. *Mol Ecol*, 9:2109–2118.

Shattuck-Eidens, D. M., Bell, R. N., Neuhausen, S. L., and Helentjaris, T. (1990). Dna sequence variation within maize and melon: observations from polymerase chain reaction amplification and direct sequencing. *Genetics*, 126:207–217.

Simko, I., Haynes, K. G., and Jones, R. W. (2006). Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. *Genetics*, 173:2237–2245.

Slatkin, M. (1987). The average number of sites separating DNA sequences drawn from a subdivided population. *Theor Popul Biol*, 32:42–49.

Spooner, D. M., Anderson, G. J., and Janson, 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., and Knapp, S. (2005). Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum* L. section *Lycopersicon*(Mill.) Wettst.]. *Taxon*, 54(1):43–61.

Städler, T., Arunyawat, U., and Stephan, W. (MS). Population genomics of speciation in two closely related wild tomatoes (*Solanum* section *Lycopersicon*). *manuscript*.

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

Stephan, W. and Langley, C. H. (1998). DNA polymorphism in *Lycopersicon* and crossing-over per physical length. *Genetics*, 150:1585–1593.

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.

Tanksley, S. D., Ganal, M. W., Prince, J. P., de Vicente, M. C., Bonierbale, M. W., Broun, P., Fulton, T. M., Giovannoni, J. J., Grandillo, S., and Martin, G. B. (1992). High density molecular linkage maps of the tomato and potato genomes. *Genetics*, 132:1141–1160.

Taylor, I. B. (1986). Biosystematics of the tomato. In *The Tomato Crop: A scientific Basis for improvement, edited by Atherton, J. G. and Rudich, J.*, pages 1–34, London. Chapman & Hall.

Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., and Gaut, B. S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc Natl Acad Sci U S A*, 98:9161–9166.

Tenaillon, M. I., U'Ren, J., Tenaillon, O., and Gaut, B. S. (2004). Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol Biol Evol*, 21:1214–1225.

Tero, N., Aspi, J., Siikamäki, P., Jäkäläniemi, A., and Tuomi, J. (2003). Genetic structure and gene flow in a metapopulation of an endangered plant species, *Silene tatarica*. *Mol Ecol*, 12:2073–2085.

Tiffin, P. and Gaut, B. S. (2001). Sequence diversity in the tetraploid *Zea perennis* and the closely related diploid *Z. diploperennis*: insights from four nuclear loci. *Genetics*, 158:401–412.

Tudhope, A. W., Chilcott, C. P., McCulloch, M. T., Cook, E. R., Chappell, J., Ellam, R. M., Lea, D. W., Lough, J. M., and Shimmield, G. B. (2001). Variability in the El Niño-southern oscillation through a glacial-interglacial cycle. *Science*, 291:1511–1517.

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

Turelli, M., Barton, N. H., and Coyne, J. A. (2001). Theory and speciation. *Trends Ecol Evol*, 16:330–343.

Turner, T. L., Hahn, M. W., and Nuzhdin, S. V. (2005). Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol*, 3:e285.

Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics*, 153:1863–1871.

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

Wakeley, J. (2001). The coalescent in an island model of population subdivision with variation among demes. *Theor Popul Biol*, 59:133–144.

Wakeley, J. (2003). Polymorphism and divergence for island-model species. *Genetics*, 163:411–420.

Wakeley, J. and Aliacar, N. (2001). Gene genealogies in a metapopulation. *Genetics*, 159:893–905.

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

Wallace, B. (1988). Selection for the inviability of sterile hybrids. *J Hered*, 79:204–210.

Wang, R. L., Wakeley, J., and Hey, J. (1997). Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics*, 147:1091–1106.

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

Wegmann, D., Currat, M., and Excoffier, L. (2006). Molecular diversity after a range expansion in heterogeneous environments. *Genetics*, 174:2009–2020.

Whitlock, M. C. and Barton, N. H. (1997). The effective size of a subdivided population. *Genetics*, 146:427–441.

Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., and Gaut, B. S. (2005). The effects of artificial selection on the maize genome. *Science*, 308:1310–1314.

Wright, S. I. and Gaut, B. S. (2005). Molecular population genetics and the search for adaptive evolution in plants. *Mol Biol Evol*, 22:506–519.

Wright, S. I., Lauga, B., and Charlesworth, D. (2003). Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. *Mol Ecol*, 12:1247–1263.

Zhang, L. B. and Ge, S. (2007). Multilocus analysis of nucleotide variation and speciation in *Oryza officinalis* and its close relatives. *Mol Biol Evol*, 24:769–783.

Zhu, Y. L., Song, Q. J., Hyten, D. L., Van Tassell, C. P., Matukumalli, L. K., Grimm, D. R., Hyatt, S. M., Fickus, E. W., Young, N. D., and Cregan, P. B. (2003). Single-nucleotide polymorphisms in soybean. *Genetics*, 163:1123–1134.

Appendix A

Primers

Primers used for PCR-reactions are as following:
F — Forward primer
R — Reverse primer
Tm – Annealing temperature

Locus CT093

CT093F 5' CTCCCCTCGGCTACAGCATT 3'
CT093R 5' AGCAGCCCTTCAGAACGGACT 3'
Tm = 55-56°C

Locus CT208

CT208F 5' CTATGGAGTTATATTTTCACCACA 3'
CT208R 5' ACTTTTGAGAGGACATCAATTT 3'
Tm = 54°C

Locus CT251

CT251F 5' TCTCTTCATCCAGTTATCCG 3'
CT251R 5' CAAGGAAGTATCGAGTCCGA 3'
Tm = 50-53°C

Locus CT066

This locus had to be amplified as two separate pieces
CT066F-a 5' CGCTGTCCCTCTTACCACCC 3'
CT066R-a 5' AATTGCTCTGCCACTTTCGCTAC 3'
$Tm = 57°C$
CT066F-b 5' TATTCTGAGTTAGTCCGCCTTGG 3'
CT066R-b 5' ATGATAGGTGCGAACAGGGTC 3'
$Tm = 55°C$

Locus CT166

CT166F 5' TGGAGCAGAGGTCAAGATTAC 3'
CT166R 5' CATTCCATTGCTCTGCCTTC 3'
$Tm = 56°C$

Locus CT179

CT179F 5' CGAATTCATCTCCACACTCA 3'
CT179R 5' TAAGACCAGCCAAACTACCAC 3'
$Tm = 54°C$

Locus CT198

CT198F 5' TGACAAACTACCGAATTACGA 3'
CT198R 5' GGTGATTTATTTAGTGCCACA 3'
$Tm = 54°C$

Locus CT268

CT268F 5' CTATGGAGTTATATTTTCACCACA 3'
CT268R 5' ACTTTTGAGAGGACATCAATTT 3'
$Tm = 56°C$

Appendix B

Protocols

DNA Extraction

The DNeasy Plant Mini Kit was used to extract DNA from dried tomato leaves, following protocol in the Handbook 01/2004 (Pages 18-21) with minor changes as following.

Cell Lysis

1) transfer 10-15 mg of silica dried leaves into a 1.5 ml reaction tube.
2) incubate the tube in liquid nitrogen and then pulverize plant material with a sterile plastic pistil.
3) add 450 μl of buffer AP1 into the tube and vortex vigorously until the plant material is suspended.
4) add 4 μl of RNase A (100 mg/ml) into the reaction tube.
5) incubate the mixture in a water bath for 30-35 minutes at 65 °C, and mix gently by inverting tube every 10 minutes.
6) add 130 μl of buffer AP2 to the lysate mix, then incubate for 5 minutes on ice.

Cell-debris removement

1) centrifuge the lysate at 13,000 rpm for 5 minutes.
2) transfer the supernatant to QIAshredder Mini Spin Column.
3) centrifuge at 7000 rpm for 1-2 minutes.
4) remove the lysate by adding 1.5 volumes of Buffer AP3/E and mix by pipetting.

DNA cleaning

1) transfer 650 μl of the mixture to a DNeasy Mini Spin Column placed on a 2 ml collection tube.
2) centrifuge at 7000 rpm for 1-2 minutes and then discard flow-through and collection tube.
3) place DNeasy Mini Spin Column in a new collection tube and clean twice by each time adding 500 μl of buffer AW to the column.
4) centrifuge at 7000 rpm for 1-2 minutes and discard flow-through.
5) centrifuge again for 2 minutes at 13,000 rpm to dry the membrane.
6) dry the column at 37 °C for 15 minutes to remove residual ethanol.

DNA elution

1) transfer the dried column into a new 1.5 ml tube.
2) add 50 ml of buffer AE directly into the DNeasy membrane.
3) incubate for 5 minutes at room temperature.
4) centrifuge for 2 minute at 7000 rpm.
5) keep the tube containing approximately 50 μl genomic DNA.
6) repeat steps 1-5, until this step two tubes of 50 μl dilute genomic DNA are obtained.
7) store DNA at 4 °C (or at -20 °C for long-term storage).

PCR Reactions

All PCR reactions were performed in an end volume of 30 μl.

Tag-polymorrase reaction

PCR mixture

19.05 μl	ddH_2O
3.0 μl	10x Buffer
3.0 μl	dNTPs
1.2 μl	MgCl_2 (50 mM)
1.5 μl	Primer F (10 μM)
1.5 μl	Primer R (10 μM)

0.15 μl Taq-Polymerase
0.6 μl Template DNA

PCR cycle Program

1) initial denaturation at 94 °C, 4 minutes
2) 30 amplification cycles:

    Denaturation at 94 °C, 45 seconds
    Annealing at locus-specific Tm (see appendix A), 1.5 minute
    Extension at 72 °C, 3 minutes

3) final extension at 72 °C for 3 minutes
4) hold at 4 °C

Phusion reaction

PCR mixture

19.6 μl ddH_2O
6.0 μl 5x HF-Buffer
0.6 μl dNTP mix
1.5 μl Primer F (10 μM)
1.5 μl Primer R (10 μM)
0.3 μl Phusion-Polymerase
0.5 μl Template DNA

PCR cycle Program

1) initial denaturation at 98 °C, 30 seconds
2) 30 amplification cycles:

    denaturation at 98 °C, 5 seconds
    annealing at locus-specific Tm (see appendix A), 20 seconds
    extension at 72 °C, 1 minute

3) final extension at 72 °C for 8 minutes
4) hold at 4 °C

PCR Clean-up

PCR reactions were cleaned using ExoSAP (Cleveland, USA)
1) add 2 μl of ExoSAP into PCR product (30 μl)
2) run ExoSAP profile in PCR machine:

37 °C for 30 minutes
80 °C for 15 minutes

Sequencing

Sequencing were perform on a MegaBace 1000 (Amersham Pharmacia Biotech) and an ABI 3730 DNA Analyzer (Applied Biosystems). The reaction were run separately for forward and reverse primers for the final volume of 10 μl per reaction.

MegaBace 1000

Using the DYEnamic ET Terminator Cycle Sequencing Kit protocol (Amersham Biosciences, UK).

Sequencing reaction

1.5 μl PCR product
2.0 μl Primer (2 μM)
4.0 μl Sequencing mix
2.5 μl ddH_2O

Sequencing program

1) 30 amplification cycles of each:

Denaturation at 95 °C, 20 seconds
Annealing at 50 °C (depends on Tm), 15 seconds
Extension at 60 °C, 60 seconds

2) Hold at 4 °C

sequencing clean-up for MegaBace

Sequencing clean-up was done using Ethanol DNA precipitation. Cleaning was performed directly on the 96-well plate.
1) add 10 μl of ddH_2O and 2 μl of Sodium acetate/EDTA into each well.
2) add 80 μl of ethanol (98%) into each well.
3) cover the plate with aluminium foil and vortex.
4) centrifuge the plate for 45 minutes at 3000 rpm.
5) discard the supernatant.
6) wash two times with 150 μl of ethanol (70%).
7) dry the plate for 2 hours at room temperature (until all ethanol evaporate).
8) Before running on the sequencer add 15 μl of ddH_2O and vortex shortly to elute the DNA pellet.

ABI 3730

Using the ABI BigDye Teminator v1.1 sequencing kit (Amersham Biosciences, UK).

Sequencing reaction

2.0 μl	DNA from PCR product
2.0 μl	Primer (10μlM)
1.0 μl	5x Buffer
2.0 μl	Sequencing mix
3.0 μl	ddH_2O

Sequencing program

1) initial denaturation at 96 °C, 1 minute
2) 50 amplification cycles of each:

    Denaturation at 96 °C, 20 seconds
    Annealing at 50-55 °C (depends on Tm), 15 seconds
    Extension at 60 °C, 4 minutes

3) Hold at 4 °C

Curriculum Vitae

Name:	Uraiwan (Ann) Arunyawat
Date of Birth:	08.09.1973
Nationality:	Thai
Office Address:	Department of Genetics, Faculty of Science,
	Kasetsart University, Bangkok 10900, Thailand
Office Phone :	+66 (0)2 5625444 or 5625555
E-mail address:	arunyawat@zi.biologie.uni-muenchen.de
	fsciuwa@ku.ac.th

Education

10/2003 - 03/2007	Ph.D. (Plant Molecular Population Genetics)
	University of Munich (LMU), Germany
10/2000 - 11/2001	Certificate (Plant Genetics Resources)
	IPK-Gatersleben, Germany
06/1995 - 10/1997	M.Sc. (Plant Molecular Biology)
	Chiang Mai University, Thailand
06/1991 - 04/1995	B.Sc. (Agriculture)
	Khon Kaen University, Thailand

Fellowships and Awards

04/2003 - 03/2007	DAAD fellowship	LMU Munich, Germany
10/2000 - 11/2001	DSE fellowship	IPK-Gatersleben, Germany
03/1994 - 06/1994	IAAS exchange student	Helgeland, Norway

Research and Academic experiences

2003 - 2007:	Population structure and speciation processes in wild tomatoes
2002 - 2003:	Teaching instructor, Department of Genetics, KU, Thailand
2000 - 2001:	Searching for *Nr* resistance markers in Lettuce by using AFLPs
1998 - 2000:	Teaching instructor, Department of Genetics, KU, Thailand
1995 - 1997:	Random Amplified Polymorphic DNA Technique for Genetic Analysis of *Curcuma spp*

Publications and Presentations

Arunyawat U., W. Stephan, and T. Städler. 2007. Using multilocus sequence data to assess population structure, natural selection and linkage disequilibrium in two closely related wild tomato species. *MBE-manuscript in review.*

Städler T., **U. Arunyawat**, and W. Stephan. 2007. Population genomics of speciation in two closely related wild tomatoes (*Solanum* section *Lycopersicon*). *Genetics-manuscript in review.*

Arunyawat U., W. Stephan, and T. Städler. 2007. Poulation Structure and linkage disequilibrium in wild tomatoes (*Solanum peruvianum* and *S. chilense*). *In* Abstract book of the 20^{th} Annual conference of the Plant Population Biology, 17-19 May 2007, Basel, Switzerland. *(Oral presentation).*

Arunyawat U., W. Stephan, and T. Städler. 2006. Nucleotide polymorphism and population structure in two closely related wild tomato species (*Solanum peruvianum* and *S. chilense*). *In* Abstract book of the 10^{th} Evolutionary Biology meeting, 20-22 September 2006, Marseilles, France. *(Oral presentation).*

Arunyawat U., K. Roselius, G. Feldmaier-Fuchs, W. Stephan, and T. Städler. 2005. Assesssing the speciation history of two closely related wild tomatoes (*Solanum peruvianum* and *S. chilense*). *In* Abstract book of the XVII International Botanical Congress, 17-23 July 2005, Vienna, Austria. *(Poster presentation).*

Acknowledgements

I am extremely grateful to my supervisor, Prof. Wolfgang Stephan, for introducing me into the world of molecular population genetics and giving the opportunity to do my doctoral research in his group. His endless enthusiasm for research encourages me during various phases of my study.

I am deeply indebted to my co-supervisor, Dr. Thomas Städler, for his guidance in both professional and personal aspects. His perfectionist in work helps me to improve my working skills, especially during the last period of completing this thesis. Thank you very much.

I would like to thank Prof. Susanne Renner and Dr. Laura Rose for their invaluable suggestions during my committee meeting. I am most grateful to 'plant people', Kestin Roselius, Tobias Marczewski, Lukasz Grzeskowiak, and Carlos Merino for several valuable discussions, in particular Lukasz and Carlos for reading part of my manuscript. I am thankful to Sarah Peter for her friendship and moral support. I must also thank Andy and Angelika to introduce me using LaTeX. Of course, I am very thankful to the entire Munich group for the good working atmosphere and the companionship that they provided during my stay in Munich.

My big thanks go to Traudl Feldmaier-Fuchs for excellent and expert laboratory assistance, particularly for her patience with my German "You are my best German teacher I ever have", Vielen Dank!!. My thank extends to Katrin Kümpfbeck, Anne Wilken, Hilde Lainer, and Anica Vrljic for all their help. Special thanks to Veronica and her child 'Mara' for their friendship –'Muchas gracias y Pelota'.

Huge thanks to Apple, Goonpawa Jamornmarn, for her unconditional support. Thanks for being my best friend and taking care of me.

A big hug for my family – Mom, Dad, Aey and Oun – I thank you all for your love and unlimited support. Your love has helped me to combat my occasional despair of completing a PhD. Finally, our dream comes true!!