MOLECULAR EVOLUTION IN WILD TOMATO SPECIES – WITH EMPHASIS ON LOCAL ADAPTATION TO ABIOTIC STRESS

Iris Fischer

München, 2011

Cover picture by C. Merino

MOLECULAR EVOLUTION IN WILD TOMATO SPECIES – WITH EMPHASIS ON LOCAL ADAPTATION TO ABIOTIC STRESS

Dissertation

Zur Erlangung des Doktorgrades der Naturwissenschaften
der Fakultät für Biologie an der Ludwig-Maximilians-Universität München

vorgelegt von
Iris Fischer
Aus Feldkirchen bei München

München, 2011

Erstgutachter:	Prof. Dr. Wolfgang Stephan
Zweitgutachter:	Prof. Dr. John Parsch
Tag der Abgabe:	12.12.2011
Tag der mündlichen Prüfung:	13.02.2012

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.

Eidesstattliche Versicherung

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

München, den 12.12.2011 _____

Iris Fischer

List of publications

Tellier A, **Fischer I**, Merino C, Xia H, Camus-Kulandaivelu L, Städler T, Stephan W **(2011a)**. Fitness effects of derived deleterious mutations in four closely related wild tomato species with spatial structure. *Heredity* **107**: 189-199.

Fischer I, Camus-Kulandaivelu L, Allal F, Stephan W **(2011)**. Adaptation to drought in two wild tomato species: the evolution of the *Asr* gene family. *New Phytologist* **190**: 1032-1044.

Fischer I, Steige KA, Stephan W, Mboup M. Evolution of regulation of drought-responsive genes in natural populations of wild tomato. *Molecular Ecology*: in preperation.

Declaration of author's contribution

In this thesis I present results form my doctorial research conducted from February 2008 until November 2011 in three chapters. All of them are the results from collaboration with other scientists. The first two chapters are published in peer-reviewed journals; the third chapter is submitted to a journal and is currently under review.

The study presented in the first chapter was designed by A. Tellier, L. Camus-Kulandaivelu, and W. Stephan. The data was provided by T. Städler. I conducted most of the analysis but with help from A. Tellier, C. Merino, and H. Xia. The paper was written by A. Tellier with revision by T. Städler, L. Camus-Kulandaivelu, W. Stephan, and me. The chapter has been published:

> Tellier A, **Fischer I**, Merino C, Xia H, Camus-Kulandaivelu L, Städler T, Stephan W **(2011a)**. Fitness effects of derived deleterious mutations in four closely related wild tomato species with spatial structure. *Heredity* **107**: 189-199.

The study presented in the second chapter was designed by me, L. Camus-Kulandaivelu, and W. Stephan. I generated the data together with L. Camus-Kulandaivelu. I did all the analysis, except the test of gene conversion which I did together with L. Camus-Kulandaivelu, and the ecological niche modelling which was performed by F. Allal. I wrote the manuscript with revisions by L. Camus-Kulandaivelu and W. Stephan. The chapter has been published:

> **Fischer I**, Camus-Kulandaivelu L, Allal F, Stephan W **(2011)**. Adaptation to drought in two wild tomato species: the evolution of the *Asr* gene family. *New Phytologist* **190**: 1032-1044.

The study presented in the third chapter was designed by me, M. K. Mboup, and W. Stephan. I carried out all the expression experiments together with M. K. Mboup and the technician H. Lainer. Sequencing was done by me and K. A. Steige. I performed all the analysis. I wrote the manuscript with revisions by M. K. Mboup, K. A. Steige, and W. Stephan. A manuscript on the findings of this chapter is currently in preperation and will soon be submitted to *Molecular Ecology*:

> **Fischer I**, Steige KA, Stephan W, Mboup M. Evolution of regulation of drought-responsive genes in natural populations of wild tomato.

Abbreviations

aa	amino acid
ABA	abscisic acid
ABRE	ABA responsive element
ARE	anaerobic response element
Aux-RR-core	core of auxin response region
Asr	ABA/water stress/ripening induced
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
bp	base pair
CAN	Canta
cDNA	complimentary DNA
C_q	quantification cycle
DFE	distribution of fitness effects of a new mutation
DNA	deoxyribonucleic acid
DREB	drought responsive element binding
ERE	ethylene responsive element
H_d	haplotype diversity
HSE	heat shock element
kb	kilo base pair
kDa	kilo Dalton
LEA	late embryogenesis abundant protein
LTR	low temperature response
MBS	MYB (transcription factor) binding site
NC	non-coding
NCBI	National Center for Biotechnology Information
NS	non-synonymous
PCR	polymerase chain reaction
qPCR	quantitative (real-time) PCR
QUI	Quicacha
RNA	ribonucleic acid
RT-PCR	reverse transcriptase PCR
S	synonymous

SFS	site frequency spectrum
SNP	single nucleotide polymorphism
TAC	Tacna
TAR	Tarapaca
TGRC	Tomato Genetics Resource Center
UTR	untranslated region
WH model	Wakeley-Hey isolation model

Summary

Understanding the mechanisms of local adaptation of wild species is a central issue in evolutionary biology. DNA sequence data allows investigating the recent demographic history of organisms. Knowledge of this history makes it possible to quantify adaptive and deleterious mutations and to analyze local adaptation at candidate genes taking the demographic context into account. As modulation of gene expression is crucial for an organism's survival during stress conditions, a next step to investigate adaptation is to study the expression profile of candidate genes. Wild species are more valuable systems to investigate local adaptation than model organisms as key issues in ecology and evolution of the later cannot be addressed properly in some cases. Wild tomato species provide several advantages when studying adaptation to abiotic stress: they grow in diverse environments – ranging from mesic to extremely arid conditions – and its genomic information is available from the cultivated relative.

First, we investigated the potential for adaptation and the strength of purifying selection acting at eight housekeeping genes in four closely related wild tomato species (*Solanum arcanum*, *S. chilense*, *S. habrochaites*, *S. peruvianum*) occupying different habitats by analyzing the distribution of fitness effects of a new mutation. There is no evidence for adaptation at these loci, but we detect strong purifying selection acting on the coding regions in all four species. Additionally, we find evidence for negative selection acting on non-coding regions. However, the strength of selection varies among species. Our results suggest that the variance of the distribution of fitness effects differ between closely related species which inhabit different environments.

Second, using a candidate gene approach, we studied the evolution the *Asr* (ABA /water stress/ripening induced) gene family in populations from contrasting environments of *S. chilense* and *S. peruvianum. Asr* genes have been reported to help the plant cope with water-deficit in many ways and are therefore useful candidates to study adaptation to drought stress. The molecular variation in the *Asr* gene family indicates that *Asr1* has evolved under strong purifying selection. Prior reports described evidence for positive selection at *Asr2* – we cannot confirm this hypothesis and argue that patterns of selection discovered previously were caused by demography. *Asr4* shows patterns consistent with local adaptation in a *S. chilense* population that inhabits an extremely dry environment. A new member of the *Asr* family (*Asr5*) was also discovered and seems to exchange genetic material with *Asr3* by gene

conversion. Our results provide a good example for the dynamic nature of gene families in plants, especially of tandemly arrayed genes that are of importance in adaptation.

Third, we investigated the expression profile following cold and drought stress as well as the regulatory regions of *Asr* genes and the dehydrin *pLC30-15*. The latter has been reported to be involved in water and chilling stress response. Populations from different habitats of *S. chilense* and *S. peruvianum* were analyzed. The gene expression of *Asr4* seems to be adaptive to drought conditions. Analysis of the regulatory regions shows a conserved promoter region of *Asr2* and positive selection acting on the downstream region of *pLC30-15*. We provide an example for expression variation in natural populations but also observe plasticity in gene expression. As noise in expression is common in stress responsive genes, we describe this expression plasticity to be advantageous in these stress-responsive genes.

In conclusion, taking the potential distribution of the species into account, it appears that *S. peruvianum* (and *S. habrochaites*) can cope with a great variety of environmental conditions without undergoing local adaptation, whereas *S. chilense* (and *S. arcanum*) seem to undergo local adaptation more frequently. With *Asr4* we identify a gene to be of potential interest for further functional studies and describe wild *Solanum* species to be of great interest as a genetic resource for its cultivated relatives.

Zusammenfassung

Eine der zentralen Aufgaben der Evolutionsbiologie ist es, die Mechanismen zu verstehen, durch welche wilde Arten sich an ihre Umwelt anpassen. Dank DNS-Sequenzdaten kann die demographische Geschichte von Organismen untersucht werden. Fundiertes Wissen dieser Geschichte erlaubt es, im demographischen Kontext adaptive und schädliche Mutationen zu quantifizieren und lokale Anpassung von Kandidatengenen zu analysieren. Da die Regulierung von Genexpression äußerst wichtig für das Überleben eines Organismus während unwirtlichen Bedingungen ist, ist ein nächster Schritt Adaptation zu untersuchen, das Expressionsprofil von Kandidatengenen zu analysieren. Bei der Untersuchung lokaler Anpassung sind wilde Arten nützlicher als Modellorganismen, da Schlüsselfragen der Ökologie und Evolution bei letztern manchmal nicht zureichend beantwortet werden können. Wilde Tomaten bieten einige Vorteile bei der Analyse von Adaptation an abiotischen Stress: Sie wachsen in unterschiedlichen Habitaten – von feuchten bis zu extrem trockenen Gebieten – und ihr domestizierter Verwandter bietet reichlich Information über das Tomatengenom.

Als erstes untersuchten wir das Potential für Adaptation und die Stärke der negativen Selektion an acht Haushaltsgenen in vier nahe verwandten wilden Tomatenarten (*Solanum arcanum*, *S. chilense*, *S. habrochaites*, *S. peruvianum*), die diverse Habitate besiedeln, indem wir die Verteilung der Fitnesseffekte neuer Mutationen analysierten. Es gibt keine Beweise für lokale Anpassung an diesen Genen, aber wir stellen in allen vier Arten fest, dass starke negative Selektion auf die codierenden Regionen wirkte. Außerdem finden wir Hinweise auf negative Selektion in den nicht-kodierenden Regionen, allerdings variiert hier die Stärke der Selektion zwischen den Spezies. Unsere Ergebnisse deuten darauf hin, dass sich die Varianzen der Verteilung der Fitnesseffekte zwischen nahe verwandten Arten, die verschiedene Lebenswelten besiedeln, unterscheiden.

Im zweiten Projekt untersuchten wir die Evolution der Kandidatengenfamilie *Asr* (ABA/water stress/ripening induced) in natürlichen *S. chilense* und *S. peruvianum* Populationen aus unterschiedlichen Habitaten. Einige Publikationen zeigten, dass *Asr* Gene der Pflanze helfen, mit Wasserdefizit umzugehen und sie sind daher geeignete Kandidaten um lokale Anpassung an Trockenheit zu studieren. Die molekulare Variation innerhalb der *Asr* Genfamilie legt nahe, dass *Asr1* starker negativer Selektion ausgesetzt war. Frühere Berichte schilderten Beweise für positive Selektion an *Asr2* – wir können diese Hypothese nicht bestätigen und erörtern, dass in vorherigen Studien positive Selektion mit demographischen Einflüssen verwechselt wurde. *Asr4* zeigt in einer *S. chilense* Population aus einer extrem

trockenen Umgebung ein Muster, das mit lokaler Adaptation übereinstimmt. Wir beschreiben außerdem ein neues *Asr* Gen (*Asr5*), das anhand von Genkonversion genetisches Material mit *Asr3* austauscht. Unsere Ergebnisse sind ein gutes Beispiel für die dynamische Natur von Genfamilien in Pflanzen, besonders von Genen die in tandem arrays liegen, welche eine große Rolle bei lokaler Anpassung spielen.

Im dritten Projekt untersuchten wir das Expressionsprofil nach Kälte- und Trockenheitsstress, sowie die regulatorischen Regionen der *Asr* Gene und des Dehydrins *pLC30-15*. Vom letzteren wurde gezeigt, dass es eine wichtige Rolle in der Reaktion auf Kälte und Wasserdefizit spielt. *Solanum chilense* und *S. peruvianum* Populationen aus unterschiedlichen Habitaten wurden analysiert. Die Genexpression von *Asr4* scheint sich an Trockenheit anzupassen. Die Analyse der regulatorischen Regionen zeigt, dass die Promoter-Region von *Asr2* konserviert ist und das positive Selektion auf die 3'-Region von *pLC30-15* wirkte. Wir zeigen ein Beispiel der Variation von Expression zwischen natürlichen Populationen, aber stellen auch Plastizität von Genexpression fest. Da ein gewisses „Rauschen" der Transkription von stress-induzierten Genen nicht ungewöhnlich ist, erläutern wir, dass Plastizität bei diesen stress-induzierten Genen von Vorteil ist.

Betrachtet man die Verbreitung der wilden Tomatenarten, sieht es so aus als ob *S. peruvianum* (und *S. habrochaites*) mit einer Vielzahl von Umwelteinflüssen zurechtkommen, ohne sich lokal anzupassen, wohingegen *S. chilense* (und *S. arcanum*) des Öfteren lokale Adaptation durchlaufen. Mit *Asr4* bieten wir ein Kandidatengen, das für weitere funktionelle Studien interessant sein dürfte und stellen fest, dass wilde *Solanum* Arten eine wichtige genetische Ressource für kultivierte Arten ist.

Contents

List of Figures

List of Tables

1 General Introduction

1.1 The importance of plant science

"Plant science has never been more important. The growing and increasingly prosperous human population needs abundant safe and nutritious food, shelter, clothes, fibre, and renewable energy, and needs to address the problems generated by climate change, while preserving habitats. These global challenges can only be met in the context of a strong fundamental understanding of plant biology and ecology, and translation of this knowledge into field-based solutions." (Grierson *et al.*, 2011). This observation led to the establishment of a project in which scientists from different fields (academic, commercial, public service) were able to create a list of 100 important questions plant science is facing nowadays (http://www. 100plantsciencequestions.org.uk/index.php). The questions were subdivided in five groups; one of them is "environment and adaptation", highlighting the importance of plants ability to adapt in cultivation and agricultural issues (Grierson *et al.*, 2011). Fundamental questions in this subsection include: How can we test if a trait is adaptive? Can we develop salt/heavy metal/drought tolerant crops without creating invasive plants? Can we develop crops that are more resilient to climate fluctuation without yield loss? To what extent are the stress responses of cultivated plants appropriate for current and future environments? In addition, the project asks key questions concerning society (*e.g.* How can we translate our knowledge of plant science into food security? How can we use plant science to prevent malnutrition?), species interaction (*e.g.* Is it desirable to eliminate all pests and diseases in cultivated plants?), and diversity (*e.g.* How can we ensure the long-term availability of genetic diversity within socio-economically valuable gene pools?). Addressing all these questions requires close cooperation between scientists of various fields in the future (Grierson *et al.*, 2011).

Humans have severe effects on plant ecology and evolution. First, due to climate change invasive species are colonizing new habitats and can out-compete native organisms as they might be better adapted to new environmental conditions (Colautti & Barrett, 2010). Second, human settlement, agriculture, and forest clearance cause habitat fragmentation which in turn leads to reduced diversity due to several factors, including population bottlenecks, reduced gene flow, or inbreeding. This severely reduces the ability of plants to adapt to new and/or changing environments (Willi *et al.*, 2006). Anthropogenic climate change leads to more rapid environmental shifts than geological climate change and understanding plant adaptation is of

great significance facing emerging global problems (Anderson *et al.*, 2011). This is especially true for key ecological traits such as drought tolerance (Anderson *et al.*, 2011).

1.2. Studying local adaptation in plants

Investigating the fitness effects of new mutations is a useful way to understand the potential and speed by which species adapt to various environments. A key predictor for the adaptive potential of a population is the distribution of fitness effects of new mutations (DFE), which denotes the probability of a mutation having a given fitness effect. A new mutation can either increase the fitness, but it can also have negative fitness effects due to accumulation of (slightly) deleterious mutations (Eyre-Walker & Keightley, 2007). Purifying selection acts against deleterious mutations and measuring the strength of purifying selection is especially relevant for developing conservation strategies for species with small population sizes. Two parameters can be employed to determine the strength of purifying selection: the mean $E(s)$ and the variance $V(s)$ of the distribution of the selection coefficient s of new mutations (Eyre-Walker & Keightley, 2007). It has been shown that $E(s)$ differs between species (Martin & Lenormand, 2006a). Within species, however, $E(s)$ is fixed whereas $V(s)$ varies across different habitats and is higher in more stressful environments (Martin & Lenormand, 2006b). This suggests that recently diverged species might show the same DFE means but differ in the variance of selection coefficients depending on the habitat they occupy.

Adaptation is characterized as the movement of a population towards a phenotype that leads to the highest fitness in a particular environment (Fisher, 1930). While the genetic basis of adaptation in natural populations remains widely unknown (Orr, 2005) many approaches have been developed to detect adaptation in several organisms employing information on DNA variation. Recently, these methods have frequently been applied to plant model systems (Siol *et al.*, 2010). One commonly used method to detect adaptation at the DNA level is to identify regions of low diversity that are linked to a selected gene – known as the hitchhiking effect (Maynard Smith & Haigh, 1974). Such an approach has been used successfully in many model organisms including *Drosophila* (Glinka *et al.*, 2006; Beisswanger & Stephan, 2008) and the common sunflower *Helianthus annuus* (Kane & Rieseberg, 2008). Another way to study adaptation is the "candidate gene" approach, which has the advantage of revealing the strength and type of selection that has acted on particular genes. With this approach, genes that have been identified in previous experiments are chosen for investigating signatures of adaptation at the DNA level. As this method does not require a sequenced genome, the

candidate gene approach has been applied frequently to several plant species where whole genome data is only available for few model organisms. Over the last years, candidate genes were successfully studied in various plant species: genes related to drought and salt tolerance in *H. annuus* (Kane & Rieseberg, 2007), immunity genes in wild tomatoes (Rose *et al.*, 2007) and *Zea mays* (Moeller & Tiffin, 2008), drought-responsive genes in maritime pine (*Pinus pinaster*; Eveno *et al.*, 2008), genes related to cold tolerance in *Arabidopsis thaliana* (Zhen & Ungerer, 2008), cold hardiness-related genes in costal Douglas fir (*Pseudotsuga menziesii* var. *Menziesii*; Eckert *et al.*, 2009), protease inhibitor genes in poplar (*Populus balsamifera*; Neiman *et al.*, 2009), cold-related genes in *Pinus sylvestris* (Wachowiak *et al.*, 2009), and genes involved in adaptation to serpentine soils in *A. lyrata* (Turner *et al.*, 2010). Although the candidate gene approach is used frequently, it is important to note that the distinction between local (selective) and genome-wide (demographic) effects is not always unambiguous. This may be the case in populations with a large effective population size, *e.g.* in *Drosophila* or bacteria (Charlesworth & Eyre-Walker, 2006; Ellegren, 2009) and/or when recurrent positive selection occurs. Demographic scenarios can also mimic selective events on a local scale (Thornton *et al.*, 2007).

Another way to investigate the evolution of candidate loci is to study their expression profile. Modulation of gene expression is crucial for an organism's survival as environmental changes require a fast and specific response. Investigating differences in gene regulation between populations from contrasting environments is not only essential for understanding local adaptation but is also the first step to test the feasibility for downstream experiments, *e.g.* using transgenic organisms. Microarrays allow analyzing the transcriptome of species, but so far studies on gene expression in natural populations are limited to few species like *Boechera holboellii*, a close relative of *A. thaliana* (Knight *et al.*, 2006), the arthropod *Orchesella cincta* (Roelofs *et al.*, 2009), the snail species *Littorina saxatilis* (Martínez-Fernández *et al.*, 2010), fishes (Larsen *et al.*, 2011), or *Drosophila melanogaster* (Hutter *et al.*, 2008; Müller *et al.*, 2011).

1.3. Candidate genes and the significance of gene families in adaptation to abiotic stress

Plants are sessile during most of their life cycle and therefore experience strong selective pressure to adapt to changing environmental conditions in their habitat (*e.g.* precipitation,

temperature). Drought and cold stress are the major abiotic constraints that terrestrial plants are facing and they have been shown to have adverse effects on the plant growth and crop production (Yáñez *et al.*, 2009). Both drought and cold tolerance are complex traits but it has been shown that similar genes are expressed during both types of stress (Shinozaki & Yamaguchi-Shinozaki, 2000). A 7000 cDNA micro-array experiment showed that the expression of more than 250 genes was induced in *A. thaliana* after a drought stress treatment (Seki *et al.*, 2002). In a similar experiment in *A. thaliana*, 4% of all transcripts showed responsiveness to low temperature (Fowler & Thomashow, 2002). Drought and cold stress lead to accumulation of the phytohormone abscisic acid (ABA) and it has been demonstrated that application of ABA mimics stress conditions (Mahajan & Tuteja, 2005). ABA plays an important role in the plant's response to osmotic stress: It fine-tunes stomatal closure (Jones & Mansfield, 1970), it enhances expression of stress-related genes (Bray, 2004), and fosters root growth in long-term drought conditions (Saab *et al.*, 1990). Shinozaki & Yamaguchi-Shinozaki (2000) suggested that cold and drought stress signals and ABA share common elements and are cross talking in their signalling pathways. Therefore, studying genes that are involved in the ABA pathway are good candidates to investigate adaptation.

Late embryogenesis abundant (LEA) proteins are induced by ABA and were shown to accumulate in vegetative organs during dehydration and low temperature stress (Ingram & Bartels, 1996; Bray, 1997). This suggests a protective role during water-limiting and chilling conditions. Members of the LEA protein family can be found all over the plant kingdom: in angiosperms, gymnosperms (Shinozaki & Yamaguchi-Shinozaki, 1996; Bray, 1997), bryophytes (Proctor *et al.*, 2007), and algae (Tanaka *et al.*, 2004). They belong to the group of the hydrophilins which are characterized by a high glycin content and high hydrophilicity (Garay-Arroyo *et al.*, 2000). The LEA proteins are subdivided into seven groups based on their amino acid sequences (Battaglia *et al.*, 2008). Here, we analyze two types of LEA proteins: Dehydrins, which belong to Group 2, and ASRs, which belong to Group 7 (Battaglia *et al.*, 2008). Some dehydrins have been shown to have cryoprotective functions, while others have been found to prevent inactivation of enzymes during dehydration (Reyes *et al.*, 2005), but their functional role still remains speculative. In *S. tuberosum* (potato) and *S. sogarandinum* an increased level of dehydrins could be correlated with cold tolerance in tubers and stems (Rorat *et al.*, 2006). Additionally, dehydrins were induced after drought stress in apical parts (Rorat *et al.*, 2006). The drought- and ABA-inducible dehydrin used here was described in *S. chilense* and denoted *pLC30-15* (Chen *et al.*, 1993). The *pLC30-15* gene has been subject to a previous population genetic study which showed that diversifying

selection acted on its coding region in a wild tomato population from a dry environment (Xia *et al.*, 2010).

My thesis is mostly focused, however, on the members of the *Asr* (ABA/water stress/ripening induced) gene family. As the name suggests, *Asr* genes have been shown to be induced by application of ABA, abiotic stress (drought, cold, salinity), and during ripening (Iusem *et al.*, 1993; Rossi & Iusem, 1994; Amitai-Zeigerson *et al.*, 1995; Schneider *et al.*, 1997; Vaidyanathan *et al.*, 1999). *Asr* genes encode small (approx. 13 kDa) highly-charged proteins and transcripts were first discovered in tomato (Iusem *et al.*, 1993; Rossi & Iusem, 1994). Frankel *et al.* (2006) found four copies on chromosome IV that lie in a tandem array and describe an insertion of 186 amino acids (aa), containing 10 imperfect repeats, present in the ASR4 protein, but absent in the other ASR proteins. *Asr*-like genes are found across the entire plant kingdom (*e.g.* pummelo - Canel *et al.* (1995); rice - Vaidyanathan *et al.* (1999); pine - Padmanabhan *et al.* (1997), and ginkgo - Shen *et al.* (2005)). There is variation in copy number between species, ranging from one in grape (Cakir *et al.*, 2003) to six in maize and rice (Frankel *et al.*, 2006). Notably, the *Asr* gene family is absent in *Arabidopsis* (Carrari *et al.*, 2004). The *Asr* genes seem to exhibit a particularly high duplication activity in tomatoes, since *Asr3* cannot be found in other Solanaceae (Frankel *et al.,* 2006). Several functions of *Asr* genes have been pointed out to help the plant deal with drought stress. In the cytoplasm, the unstructured ASR1 monomers act as chaperons, possibly to prevent proteins from losing their structure during desiccation (Konrad & Bar-Zvi, 2008). With an increasing zinc level in the cell, ASR proteins are located in the nucleus where ASR1 forms homodimers (Maskin *et al.*, 2007) with a zinc dependent DNA-binding activity (Kalifa *et al.*, 2004a). Additionally, it was discovered in grape (*Vitis vinifera*) that ASR proteins form heterodimers with DREB (drought response element binding) proteins (Saumonneau *et al.*, 2008). This DNA binding activity could stabilize the DNA during stress conditions but was also associated with the modulation of sugar transport activity (Carrari *et al.*, 2004; Frankel *et al.*, 2007; Maskin *et al.*, 2008). The last observation places the *Asr* genes at a key position given the interaction of sugar and ABA pathways discovered in seed developmental processes (Finkelstein & Gibson, 2001) and stress signalling (León & Sheen, 2003). In that context, two studies revealed patterns of positive selection at *Asr2* in populations of wild tomato species that dwell in dry environments by a phylogenetic (Frankel *et al.*, 2003) and a population genetic approach (Giombini *et al.*, 2009). Expression analyses in several species suggest high plasticity in relative *Asr* expression. In pine (*P. taeda*) and rice (*Oryza sativa*), expression patterns vary depending on the gene copy (Padmanabhan *et al.*, 1997; Philippe *et al.*, 2010). Studies in pine,

potato, and *Ginkgo biloba* indicate an organ-specific induction of *Asr* (Padmanabhan *et al.*, 1997; Schneider *et al.*, 1997; Shen *et al.*, 2005) and differences in expression between several stresses have been described in potato, rice, and ginkgo (Schneider *et al.*, 1997; Vaidyanathan *et al.*, 1999; Shen *et al.*, 2005). Using semi-quantitative RT-PCR, it was shown that *Asr1* and *Asr2* are induced in leaves and that *Asr2* is induced and *Asr3* is down-regulated in roots of cultivated tomato (Maskin *et al.*, 2001). Analyzing different accessions of wild tomato using Northern Blots, Frankel *et al.* (2006) demonstrated that *Asr1* and *Asr4* are up-regulated in leaves of plants from humid environments. All these findings make *pLC30-15* and mostly *Asr* genes interesting candidates to study local adaptation on the gene expression level.

Duplicated genes are an important source of adaptation. This is especially the case in plants where a large fraction of diversity is caused by gene duplication and subsequent adaptive specialization of paralogous gene copies (Flagel & Wendel, 2009). Duplicates of transcription factors in *A. thaliana* are preferentially retained after polyploidy (Paterson *et al.*, 2006) and for MADS-box transcription factors, it has been suggested that gene duplication followed by increased opportunity for novel gene interactions played an important role in early angiosperm diversification (Shan *et al.*, 2009). Genes involved in stress response are often tandemly duplicated which makes these arrayed genes interesting for studying adaptation to drought and cold (Maere *et al.*, 2005; Mondragon-Palomino & Gaut, 2005; Rizzon *et al.*, 2006; Hanada *et al.*, 2008). Analysis of expression data of *Arabidopsis* indicates that a gain (or loss) of stress responsiveness is more common in tandemly duplicated genes than in non-tandem duplicates (Zou *et al.*, 2009). Moreover, gene families likely to be involved in lineage specific adaptive evolution are mainly generated by tandem duplication (Hanada *et al.*, 2008). According to the classical model on gene duplication (Ohno, 1970), selective constraints remain on one copy after a gene duplication event, whereas the other copy can accumulate mutations. The most common fate of this latter copy is a loss of function. In rare cases, however, a mutation can be advantageous in a specific environment leading to a neofunctionalization of one copy (Beisswanger & Stephan, 2008) or subfunctionalization of both copies (Hughes, 1994). Concerted evolution, in which two copies of a gene do not evolve independently, can also be observed. The most common mechanism causing this phenomenon is gene conversion, whereby two copies exchange short tracts of DNA in a "copy-and-paste" manner. This will both decrease the sequence variation between copies and increase the genetic variation within the gene family by creating new haplotypes (Takuno *et al.*, 2008). New (chimeric) haplotypes can be advantageous in genes that experience diversifying selection, since it increases the genetic diversity (Takuno *et al.*, 2008;

Innan, 2009). Therefore, it is possible that the members of a gene family exhibit very different evolutionary histories.

1.4. *Solanum* species as non-model organisms to investigate local adaptation

In the past, most studies on plant evolution and adaptation were conducted using model organisms such as *A. thaliana*, *O. sativa*, or *Z. mays* for which whole genome data is available. For those plants, an environmental context is not clear or cultivation caused reduced diversity due to bottlenecks and artificial selection. To understand local adaptation, however, plants from natural environments – in which they evolved – are required (Anderson *et al.*, 2011). This is why more and more scientists investigate non-model organisms over the past years (Song & Mitchell-Olds, 2011). As non-model organisms are mostly lacking sequenced genomes, it is reasonable to study wild relatives of model organisms (Song & Mitchell-Olds, 2011). This has successfully been done in relatives of *e.g. A. thaliana* (Riihimäki *et al.*, 2005; Knight *et al.*, 2006; Turner *et al.*, 2010; Leinonen *et al.*, 2011), sunflower (Kane & Rieseberg, 2007; Kane & Rieseberg, 2008), rice (Grillo *et al.*, 2009), and tomato (Moyle, 2008).

The plant family Solanaceae ("nightshades") is cosmopolitan and its members inhabit a broad variety of habitats showing great diversity, both morphologically and genetically. It contains various economically important species, *e.g.* bell peppers and chilis (*Capsicum annuum*), ornamental plants such as *Petunia*, and also tobacco (*Nicotiana tabacum*). *Solanum* is the largest of 90 genera in the Solanaceae family and one of the largest genera among angiosperms as it contains approximately 1,400 species (Planetary Biodiversity Inventory *Solanum* Project; http://www.nhm.ac.uk/solanaceaesource/). The Solanoideae subfamily (which includes *Solanum*) is a monophyletic group with a chromosome number based on x = 12 (Olmstead & Palmer, 1992; Olmstead & Sweere, 1994). The genus contains very important food plants such as potato (*S. tuberosum*), tomato (*S. lycopersicum*), and eggplant (*S. melongena*). Centres of diversity of *Solanum* species cluster in the Southern Hemisphere, most importantly in South America (Edmonds & Chweya, 1997). The hyperdiverse nature of this genus, which represents almost 1 % of the angiosperm flora on Earth (Whalen & Caruso, 1983), makes it an extraordinary system to investigate its use for humans (Knapp *et al.*, 2004). In recent years, the use of this group as a genetic resource for cultivated *Solanum* species has been investigated. Since most cultivated plants lose a lot of genetic variation during the domestication process, they are extremely susceptible to all kinds of biotic and abiotic

stresses. For example, the spread of pathogens in green houses or fields is a major problem in agriculture. As wild relatives of cultivated species usually show higher tolerance to environmental factors, they serve as genetic resources for plant breeding.

Wild tomatoes are an interesting plant species to study evolutionary biology for several reasons, including the availability of cultivated tomato genomic resources, the recent divergence of the *Solanum* species, their clear phenotypic distinction (Peralta *et al.*, 2008), and the diversity of mating systems (Spooner *et al.*, 2005; Moyle, 2008). For a long time, tomatoes constituted the genus *Lycopersicon*, but recent taxonomic revision suggested grouping them in the genus *Solanum* (section *Lycopersicon*) together with potato and the eggplant (Spooner *et al.*, 1993; Peralta & Spooner, 2001). Most *Solanum* sect. *Lycopersicon* species are native to western South America (Ecuador, Peru, and Chile), along the western and eastern Andean slopes, but with two endemic species on the Galapagos islands (Spooner *et al.*, 2005). According to the latest taxonomical classification, tomatoes consist of 12 wild species and their cultivated relative, *S. lycopersicum* (Spooner *et al.*, 2005).

In this thesis, I investigated four wild tomato species that show differences in their ecological habitats and features: *S. arcanum*, *S. chilense*, *S. habrochaites*, and *S. peruvianum* (with the main focus on the sister species *S. chilense* and *S. peruvianum*). *Solanum habrochaites* occurs from central Ecuador to central Peru (Fig. 1.1) and can dwell in dry costal areas as well as in clouded forests up to 3,600 m (Peralta *et al.*, 2008). *Solanum peruvianum* (*sensu stricto*) is distributed from central Peru to northern Chile (Fig. 1.1) and inhabits a variety of habitats, from coastal deserts to river valleys (Peralta *et al.*, 2008). Furthermore, it may be found at field edges, unlike other *Solanum* species (Chetelat *et al.*, 2009). Recently, *S. peruvianum* (*sensu lato*) was split in four species, including *S. arcanum* (Peralta *et al.*, 2005) which represents the former northern distribution of *S. peruvianum sensu lato* in northern Peru (Fig. 1.1). The species inhabits Andean valleys and rocky slopes and can be subject to rain shadows (Peralta *et al.*, 2008). *Solanum chilense* is distributed from southern Peru to northern Chile (Fig. 1.1) and inhabits arid plains and deserts (Peralta *et al.,* 2008). It also shows a broad range in elevation from sea level up to 3,500 m (Chetelat *et al.*, 2009). The species is known to be robust and drought tolerant and can dwell in hyperarid areas due to its well-developed root system (Moyle, 2008; Peralta *et al.*, 2008). In fact, the potential distribution of *S. chilense* is predicted to be mostly determined by the annual precipitation (Nakazato *et al.*, 2010). Studies of *S. chilense* and *S. peruvianum* revealed population subdivision (Roselius *et al.*, 2005) and apparently population structure has played an important role in the evolution of wild tomatoes (Arunyawat *et al.*, 2007). Other studies

suggested that both species diverged under residual gene flow (Städler *et al.*, 2005; Städler *et al.*, 2008). Taking the difference in range and habitat of these species into account, differences of environmental cues as described by Xia *et al.* (2010) can be expected. This diverse environmental distribution makes wild tomato species an ideal model organism to study local adaptation.

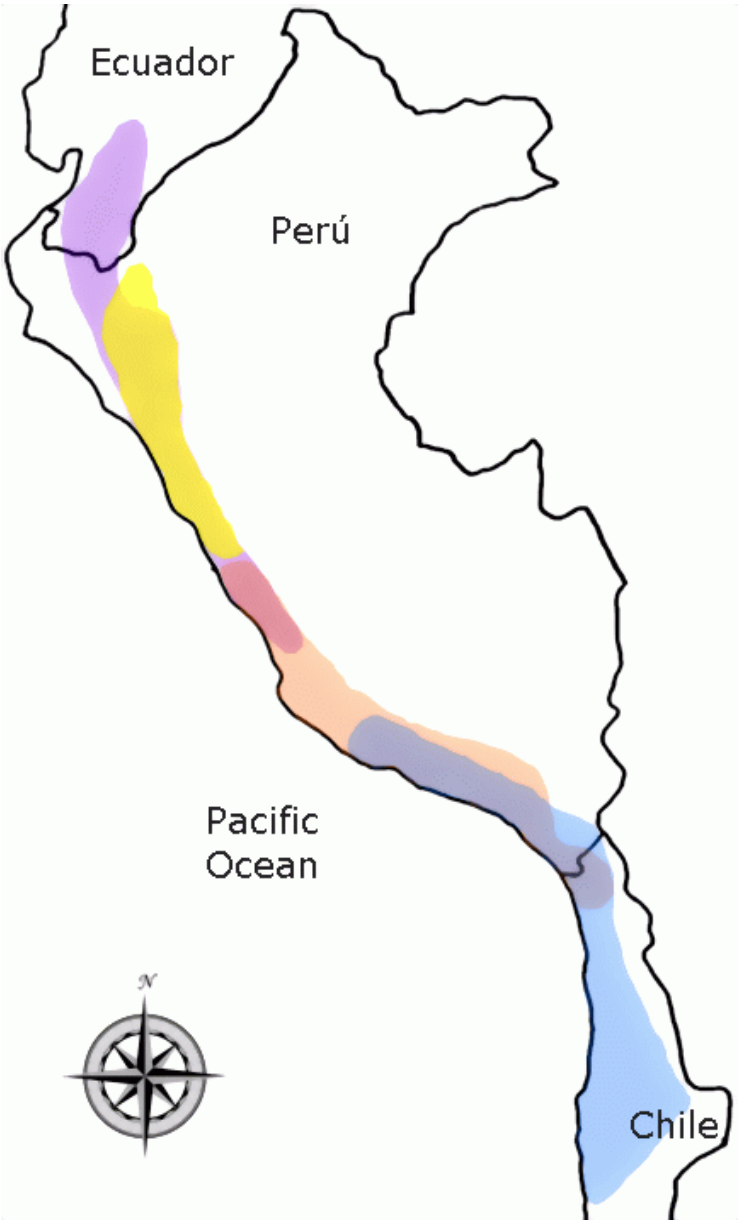

Figure 1.1 Map of West South America showing the distribution of four wild tomato species: *S. habrochaites* (purple), *S. arcanum* (yellow), *S. peruvianum sensu stricto* (orange) and *S. chilense* (light blue). Map by C. Merino and T. Städler.

1.5. The scope of this thesis

The aim of this project was to detect local adaptation at genes involved in stress response in the non-model organism of wild tomato. We achieved this by three projects. In the first project we wanted to qualify the effects (adaptive or deleterious) of new mutations in four wild tomato species (*S. arcanum*, *S. chilense*, *S. habrochaites*, *S. peruvianum*) which inhabit different environments. By doing so, we wanted to gain insight into the potential for adaptation and the strength of purifying selection within these species. We accounted for structure within the *Solanum* species by comparing the pattern observed on synonymous sites (which are evolving under neutrality) to the patterns at non-synonymous and non-coding sites. Our goal for the second project was to investigate the evolutionary forces acting on different members of the *Asr* gene family in the closely related species *S. chilense* and *S. peruvianum*. We employed a population genetics approach to analyze a larger dataset than previous studies (Frankel *et al.*, 2003; Frankel *et al.*, 2006; Giombini *et al.*, 2009) by using several populations from different environments. The fact that demography acts on the whole genome, whereas selection affects only restricted genomic regions, allowed us to detect selection on our candidate genes by comparing them to a set of reference loci previously described by Arunyawat *et al.* (2007) and Städler *et al.* (2008). For the third project, we analyzed the relative expression of *Asr1*, *Asr2*, *Asr4*, and *pLC30-15* in drought and cold stressed *S. chilense* and *S. peruvianum* accessions from contrasting environments to determine differences in gene expression kinetics. These differences can be expression intensity, speed, or variances depending on the type of stress or the gene copy. As *Asr3* and *Asr5* cannot be distinguished at their coding region, they were excluded from this study. Population genetic analysis has provided evidence for local adaptation at *Asr2*, *Asr4*, and *pLC30-15* (Giombini *et al.*, 2009; Xia *et al.*, 2010). We therefore sequenced the regulatory regions of these genes from the same populations in order to investigate the evolutionary forces shaping them. In addition, we wanted to identify conserved *cis*-acting elements. The general aim of my thesis was to gain a better understanding of the evolution of natural populations and their potential to adapt to changing environments.

Heredity (2011) 107, 189–199

www.nature.com/hdy

ORIGINAL ARTICLE

Fitness effects of derived deleterious mutations in four closely related wild tomato species with spatial structure

A Tellier[1], I Fischer[1], C Merino[1], H Xia[1,2], L Camus-Kulandaivelu[1,3], T Städler[4] and W Stephan[1]
[1]*Section of Evolutionary Biology, Department Biology II, University of Munich (LMU), Planegg-Martinsried, Germany;* [2]*College of Horticulture, Northwest A&F University, Shaanxi, China;* [3]*CIRAD, Montpellier, France and* [4]*Institute of Integrative Biology, Plant Ecological Genetics, ETH Zurich, Zurich, Switzerland*

A key issue in evolutionary biology is an improved understanding of the genetic mechanisms by which species adapt to various environments. Using DNA sequence data, it is possible to quantify the number of adaptive and deleterious mutations, and the distribution of fitness effects of new mutations (its mean and variance) by simultaneously taking into account the demography of a given species. We investigated how selection functions at eight housekeeping genes of four closely related, outcrossing species of wild tomatoes that are native to diverse environments in western South America (*Solanum arcanum*, *S. chilense*, *S. habrochaites* and *S. peruvianum*). We found little evidence for adaptive mutations but pervasive evidence for strong purifying selection in coding regions of the four species. In contrast, the strength of purifying selection seems to vary among the four species in non-coding (NC) regions (introns). Using F_{ST}-based measures of fixation in subdivided populations, we suggest that weak purifying selection has affected the NC regions of *S. habrochaites*, *S. chilense* and *S. peruvianum*. In contrast, NC regions in *S. arcanum* show a distribution of fitness effects with mutations being either nearly neutral or very strongly deleterious. These results suggest that closely related species with similar genetic backgrounds but experiencing contrasting environments differ in the variance of deleterious fitness effects.
Heredity (2011) **107,** 189–199; doi:10.1038/hdy.2010.175; published online 19 January 2011

Keywords: natural selection; distribution of fitness effects; population structure

Introduction

Mutations are the raw material of evolution. To understand the nature of quantitative variation, and thus the potential and speed of adaptation of species to various environments, it is important to determine the positive or negative fitness effects of new mutations. The distribution of fitness effects of new mutations (henceforth denoted DFE) specifies the probability of a new mutation having a given fitness effect. Quantifying the DFE is a key predictor of the potential for adaptation of a population due to mutations with positive fitness effects, as well as a predictor of the decrease in fitness following the accumulation of deleterious mutations (Eyre-Walker and Keightley, 2007). Purifying selection against deleterious mutations increases the proportion of low-frequency alleles, and it reduces the effective population size and, thus, the levels of neutral heterozygosity at linked loci or sites (Charlesworth *et al*., 1993). Measuring the strength of purifying selection is particularly important for designing strategies to conserve species with small population sizes, and to understand the appearance and maintenance of low-frequency genetic diseases in humans.

The strength of purifying selection can be measured by the two parameters of the DFE: the mean $E(s)$ and the variance $V(s)$ of the distribution of the selection coefficient s for each new mutation (Eyre-Walker and Keightley, 2007). Using mutation accumulation lines, it has been shown that $E(s)$ differs among species of bacteria and *Drosophila*, as well as between the yeast *Saccharomyces cerevisiae*, the nematode worm *Caenorhabditis elegans*, and the model plant *Arabidopsis thaliana*, taxa characterized by very different genome sizes and genome organization (Martin and Lenormand, 2006b). Moreover, mutations tend to be more deleterious (higher $E(s)$) and less variable (small $V(s)$) in more 'complex' organisms, as defined by their genome size (Martin and Lenormand, 2006b). Interestingly, for a given species, a newly arising mutation will exhibit different fitness effects depending on the environmental (biotic and abiotic) conditions. In other words, the mean mutation effect $E(s)$ is fixed for the species, whereas $V(s)$ may vary across environments. For example, higher variance of the DFE is observed in more stressful environments (Martin and Lenormand, 2006a). These results may also suggest that recently diverged species that occupy different environments might show similar DFE means but different variances of selection coefficients. A key evolutionary question is whether these theoretical expectations are general for all animal

Correspondence: Dr A Tellier, Section of Evolutionary Biology, Department Biology II, University of Munich (LMU), Planegg-Martinsried 82152, Germany.
E-mail: tellier@biologie.uni-muenchen.de
Received 17 September 2010; revised 6 December 2010; accepted 20 December 2010; published online 19 January 2011

and plant species, as well as for coding and non-coding (NC) regions.

More recent methods to measure the DFE rely on using polymorphism data at synonymous (S) and non-synonymous (NS) sites summarized as the site-frequency spectrum, that is, the allele frequency distribution observed in a population sample. Quantifying directional selection (purifying and positive) using the site-frequency spectrum is a powerful approach, but its effectiveness relies on distinguishing the signature of selection in polymorphism data from that of demographic processes. Indeed, similar patterns of genetic diversity and site-frequency spectra, such as an excess of low-frequency polymorphisms, can occur because of the demographic events (population expansion) or because of purifying selection (Eyre-Walker and Keightley, 2007). Methods to measure selection, thus, attempt to estimate the past demography of species, usually based on S sites, and simultaneously or subsequently compute the effects of selection (positive or negative) on NS and NC sites (Eyre-Walker and Keightley, 2007). Such methods have so far been applied mainly to model organisms for which genome-wide polymorphism data are available (humans, *Drosophila*, *S. cerevisiae* and *Arabidopsis*; Wright and Andolfatto, 2008; Eyre-Walker and Keightley, 2009; Keightley and Eyre-Walker, 2010; but see Gossmann *et al*., 2010 and Slotte *et al*., 2010 for recent studies on other plant species).

As a first step, we use simulations to evaluate the robustness of the method of Eyre-Walker and Keightley (2009) for estimating the DFE parameters when only a limited set of 100–300 single-nucleotide polymorphisms (SNPs) and 40–50 sampled alleles are available (for rationale, see Materials and methods). The DFE parameters ($E(s)$ and $V(s)$) can not be accurately estimated with such a low number of SNPs (Keightley and Eyre-Walker, 2010). Using simulated datasets, we show, however, that statistically significant differences can be inferred between the shapes of the DFE for neutral mutations, weakly deleterious mutations and strongly deleterious mutations. We also show that the statistical differences between DFE shapes observed across species are robust under population expansion.

The populations of most (if not all) plant species are spatially sub-structured to some extent. Spatially structured populations with demes connected by migration can lead to patterns of nucleotide diversity dramatically different from those expected in a single panmictic population. For example, the efficacy of positive (negative) selection is affected in spatially structured populations, because drift and migration can counteract the rise (or decrease) in frequency of favorable (unfavorable) alleles. Specifically, the time required for deleterious mutations to be eliminated from a single panmictic population with a given effective population size N_e is shorter than in a structured population with similar N_e. This occurs when gene flow is low, because genetic drift counterbalances the effect of negative selection in demes with small effective size (Whitlock, 2003). Moreover, because selection prevents deleterious mutations from reaching high frequencies, such polymorphisms are mostly private to particular demes rather than shared among subpopulations (Fay *et al*., 2001; Whitlock, 2003). For a given level of migration and mutation rate, both real genetic differentiation (*sensu stricto* Jost, 2008) and the traditional fixation index F_{ST}, thus, ought to be higher at sites under purifying selection compared with sites under neutral evolution, assuming linkage equilibrium between sites (Charlesworth *et al*., 1997). However, for cases in which the effective size per deme, migration among demes and recombination rates are small, the fixation index F_{ST} can be lower under purifying selection compared with neutral evolution (Pamilo *et al*., 1999). This suggests that (i) very strong purging of deleterious mutations in addition to small recombination rate, and (ii) absence of purifying selection can give similar patterns of F_{ST} because migrants possess a fitness advantage (Charlesworth *et al*., 1997; Pamilo *et al*., 1999).

The main empirical objective of this study is to investigate the strength of purifying selection in four closely related wild tomato species: *Solanum peruvianum*, *S. chilense*, *S. habrochaites* and *S. arcanum*. These species are native to western South America, their composite geographic ranges extending from central Ecuador to northern Chile. Collectively, they occupy diverse ecological habitats, with abiotic environments varying from mesic to extremely xeric conditions (Nakazato *et al*., 2010). These species are proposed to exist as structured populations with many demes (over 100) linked by migration (Arunyawat *et al*., 2007; Städler *et al*., 2009; Nakazato *et al*., 2010). We test the prediction that the mean $E(s)$ of the DFE is identical among habitats and species, but that $V(s)$ differs. We quantify and compare the strength of purifying selection acting on coding regions and on NC (intronic) regions for eight nuclear loci with putatively known housekeeping functions. These housekeeping genes are conserved among the species, and we thus expect strong purifying selection acting on their coding sequences. We also assess whether purifying selection acts on the intronic regions.

The four closely related wild tomato species studied here are characterized by fragmented population structure and various degrees of local adaptation to abiotic conditions (Xia *et al*., 2010). Thus, our second objective is to investigate how spatial structure of populations affects the efficacy of purifying selection. We quantify the strength of purifying selection using the distribution of deleterious mutations among populations at the eight studied loci, using F_{ST}-based methods (Foll and Gaggiotti, 2008). Finally, we compare the strength of selection inferred at the species level (that is, via the species-wide DFE) with that based on estimates of the fixation index across each structured population of species.

Materials and methods

Plant material and DNA sequencing

Solanum section *Lycopersicon* consists of 13 nominal species found in a relatively small area in western Peru, Chile and Ecuador and includes the domesticated tomato, *Solanum lycopersicum* (formerly *L. esculentum*; Peralta *et al*., 2008). These species are closely related diploids ($2n = 24$) sharing a high degree of genomic synteny (Ji and Chetelat, 2007). The four studied species, *S. chilense*, *S. peruvianum*, *S. arcanum* and *S. habrochaites* are characterized by spatially structured populations (Arunyawat *et al*., 2007). For this study, new population samples were collected in central and northern Peru by

T Städler and C Merino in September 2006: Canta, Otuzco, Contumaza and Lajas for *S. habrochaites*, and Otuzco, Rupe, San Juan and Cochabamba for *S. arcanum*. The population samples and geographic locations are summarized in Supplementary Table S1. Voucher specimens have been deposited at the herbarium of the Universidad San Marcos (Lima, Peru). Basic population genetic analyses of nucleotide polymorphism within species and divergence among species will be published elsewhere (CM, AT, WS and TS, unpublished data).

For each sampled population, usually five or six diploid individuals (that is, 10 or 12 alleles) were sequenced at eight unlinked nuclear loci that were previously studied in similarly sized samples of *S. chilense* and *S. peruvianum* (CT093, CT208, CT251, CT066, CT166, CT179, CT198 and CT268; Arunyawat *et al*., 2007). These loci are single-copy complementary DNA markers originally mapped by Tanksley *et al*. (1992) in genomic regions with different recombination rates (Stephan and Langley, 1998). The gene products putatively perform key housekeeping functions, and thus purifying selection is suggested to drive their evolution (Supplementary Table S2; Roselius *et al*., 2005). Genomic DNA was extracted from silica-dried tomato leaves using the DNeasy Plant Mini Kit (Qiagen GmbH, Hilden, Germany). PCR primers were the same as developed for our previous studies, and PCR conditions followed those of our previous studies of the same loci in *S. peruvianum* and *S. chilense* (Arunyawat *et al*., 2007); PCR primer information can be accessed at http://evol.bio.lmu.de/downloads/index.html.

PCR amplification was performed with High Fidelity Phusion Polymerase (Finnzymes, Espoo, Finland), and all PCR products were examined with 1% agarose gel electrophoresis. Generally, direct sequencing was performed on PCR products to identify homozygotes and obtain their corresponding sequences. For heterozygotes, a dual approach of both cloning before sequencing and direct sequencing was used to obtain the sequences of both alleles. As before, we developed a series of allele-specific sequencing primers whose 3′-end was anchored on identified SNPs or indels (for details of this approach, see Städler *et al*., 2005). Haplotype phase was thus completely resolved for all sequences. Sequencing reactions were run on an ABI 3730 DNA Analyzer (Applied Biosystems and HITACHI, Foster City, CA, USA). Two alleles were sequenced for each individual, and a total of 39–52 sequences were obtained for each locus × species combination. Contigs of each locus were first built and edited using the Sequencher program (Gene Codes, Ann Arbor, MI, USA) and adjusted manually in MacClade 4 (version 4.06 for OS X, Sinauer Associates, Sunderland, MA, USA). The new sequences for *S. habrochaites* and *S. arcanum* have been deposited in GenBank under accession numbers GU950656–GU951412. In addition, this study also analyzes our previously published sequences sampled from each of four populations in both *S. peruvianum* and *S. chilense*, as well as outgroup sequences from tomato relatives (Baudry *et al*., 2001; Roselius *et al*., 2005; Arunyawat *et al*., 2007). We also included one previously sequenced sample of *S. habrochaites*, which was obtained from the Tomato Genetics Resource Center at UC Davis (http://tgrc.ucdavis.edu; accession LA1775, 'Ancash', see Supplementary Table S1 and Städler *et al*., 2005).

Basic analyses of the sequence data

For each species, we analyzed polymorphic sites per locus and for all concatenated loci together, using the three categories of sites: S, NS and NC or intronic polymorphic sites. We allowed for multiple hits and polarized nucleotide states as ancestral or derived using either *S. ochranthum* or *S. lycopersicoides* as outgroups (depending on availability; Roselius *et al*., 2005; Arunyawat *et al*., 2007). We quantified the strength of purifying selection at the species level by analyzing the pooled samples within species, that is, all four or five populations together, because these should best represent the species-wide diversity (Städler *et al*., 2009).

Tajima's D (D_T; Tajima, 1989) summarizes the site-frequency spectrum. Statistically significant deviations from zero suggest that a locus has not evolved under neutrality, or that past demographic events have affected the site-frequency spectrum. We also conducted tests based on a comparison of the divergence between two species, taking into account the ratios of S and NS substitutions to the ratios of S and NS polymorphisms (K_a/K_s and π_a/π_s ratios), the McDonald and Kreitman (1991) test and the proportion of adaptive substitutions α (Bierne and Eyre-Walker, 2004). Statistical analyses were performed using DnaSP v. 5.0 (Librado and Rozas, 2009) and the program SITES (Hey Lab, Department of Genetics, Rutgers University).

Partitioning polymorphic sites and the site-frequency spectra

We computed the total number of S and NS sites (including segregating and non-segregating) using the approximate method *yn00* of Yang and Nielsen (2000), as implemented in phylogenetics analysis and maximum likelihood. This method takes into account transition/transversion rate bias and base/codon frequency bias. The size of the coding regions, number of S sites and NS sites are found in Supplementary Table S2, using *S. habrochaites* as an example. All models of sequence evolution used here assume that all sites have the same mutation rate and that no multiple hits occur. We thus corrected for multiple hits in our datasets by calculating the number of substitutions and polymorphisms (S, NS and NC) using the DnaSP conservative criteria (Nei and Gojobori, 1986).

We first calculated the simplified version of the site-frequency spectrum comprising three categories developed by Fay *et al*. (2001). The minor allele at each polymorphic SNP is called 'rare' if its frequency is $<5\%$, 'intermediate' if the frequency is between 5% and 20%, and 'common' if its frequency is $>20\%$. These categories of SNP frequencies are calculated for the pooled samples (39–60 sequences, depending on the locus and species) separately for S, NS and NC sites. The proportions of polymorphism S*, NS* and NC* are computed for each frequency class (rare, intermediate and common), where asterisk (*) denotes the ratio of the number of SNPs over the total number of sites (Fay *et al*., 2001). Calculations were made using R scripts (R Development Core Team, 2005).

Under simplified assumptions, NS and NC sites fall into three classes: neutral, slightly deleterious and strongly deleterious (Fay *et al*., 2001). Neutral NS or NC sites are responsible for all common SNPs in the

site-frequency spectrum and a portion of rare and intermediate SNP classes. Slightly deleterious mutations account for the excess of rare-frequency polymorphism, as well as a small fraction of the intermediate-frequency SNPs. Strongly deleterious mutations are assumed to rarely rise to detectable frequency (Fay *et al*., 2001). The proportion of NS and NC SNPs in each class is compared with that observed for the S ones. S sites are assumed to be neutral and their site-frequency spectrum thus only determined by past demographic events and population structure. Under purifying selection, an excess of rare-frequency polymorphisms at NS and NC sites is thus expected compared with the proportion of rare-frequency polymorphisms at S sites. On the other hand, similar frequencies of common SNPs are expected for the various classes, reflecting the effectively neutral evolution of common SNPs. The quantities 1–NS*/S* and 1–NC*/S* thus indicate the proportion of sites under selection for NS sites and NC regions, respectively (Fay *et al*., 2001).

Simulation analyses to estimate the DFE parameters

We aim to estimate the demography, the DFE parameters and α simultaneously for each of the four wild tomato species, using the maximum likelihood method of Eyre-Walker and Keightley (2009). This method is available on PD Keightley's web server (http://homepages.ed.ac.uk/eang33/) and can be summarized as follows. The demographic model is a simple one-step population size change from N_1 (ancestral population size) to $N_2 = N_e$, the present effective population size, assumed to be at equilibrium between mutation, selection and drift. The population expansion ($N_1 < N_2$) or contraction ($N_1 > N_2$) occurred t-generations ago. Each deleterious mutation has a different fitness coefficient s, which is assumed to be drawn from a γ distribution with shape parameter b and mean parameter $N_2E(s) = N_eE(s)$ (Keightley and Eyre-Walker, 2007). It is also assumed that there is a class of neutral sites at which mutant alleles have no effect on fitness. For diploid organisms, the relative fitness of the wild-type genotype, heterozygous mutant and homozygous mutant genotype is 1, 1–s/2 and 1–s, respectively (Eyre-Walker and Keightley, 2009). In a second step, the rate of positively selected mutations α is estimated for coding regions. The demography and the DFE parameters are thus used to predict the expected number of substitutions due to deleterious mutations, and the difference between this expected number and the observed number of substitutions yields an estimate of α (Eyre-Walker and Keightley, 2009). In addition, we use the DoFE software available at A. Eyre-Walker's website (University of Sussex, Brighton, UK) (Bierne and Eyre-Walker, 2004), a method based on the MK test (McDonald and Kreitman, 1991) to estimate α. Unlike Eyre-Walker and Keightley (2009), the DoFE method does not take the demography of each species explicitly into account, and NS mutations segregating within a species are assumed to be neutral.

The accuracy of DFE parameter estimates has been tested so far only for datasets with large numbers of SNPs (>1000; Eyre-Walker and Keightley, 2009). We assessed the accuracy of these parameter estimates for datasets mimicking our available wild tomato sequence data. The software SFS_code available at R. Hernandez website (http://sfscode.sourceforge.net/SFS_CODE) (Hernandez, 2008) was used to simulate a panmictic population with a sample of 20 diploid individuals (that is, 40 chromosomes) sequenced at eight loci of 1250 bp each, using two demographic scenarios: constant population size and a fivefold expansion, the latter occurring $t = 0.8 \times 4\ N_2$ generations ago. The population mutation parameter of the ancestral population θ_1 varies from 0.005 to 0.02, and 50 simulations with a number of S segregating sites between 100 and 200 were retained for analysis. We simulated 50 independent datasets for a given set of DFE parameter values for the NS mutations with increasing deleterious coefficients ($-N_eE(s) = 5$, 50, 500 and 5000), as well as 50 datasets with only neutral mutations. The shape of the deleterious DFE distribution was fixed at $b = 0.1$. This value is conservative for assessments of power, as for strongly leptokurtic shapes the estimates of the mean and variance of the γ distribution are most inaccurate (Keightley and Eyre-Walker, 2010). Similar simulations were run with a more platikurtic γ distribution ($b = 50$).

Strictly speaking, datasets with only neutral mutations do not have a DFE because 100% of the mutations are in the range $0 < -N_es < 1$. However, in our simulations, datasets with relatively low numbers of neutral SNPs may show departures from the neutral site-frequency spectrum, and thus a DFE is estimated by the method of Eyre-Walker and Keightley (2009) with mutations including $-N_es > 1$.

We recorded the means of estimates and the root mean square error of the ratio of current and ancestral effective population size (N_2/N_1), the time of expansion ($t/4N_2$) and the shape of the DFE γ distribution (b). Root mean square error is the square root of the average squared difference (over $n_{sim} = 50$ datasets) between the estimated value and the simulated value, divided by the simulated value. We also recorded the proportions of mutations in different $-N_es$ ranges ($0 < -N_es < 1$; $1 < -N_es < 10$; $10 < -N_es < 100$; $100 < -N_es$). Multiple mean comparisons were performed for the proportions of mutations in different $-N_es$ ranges as a Tukey's honestly significantly different test (confirmed by a Bonferroni test), as implemented in the R software (R Development Core Team, 2005).

DFE and demography estimates for each tomato species

We concatenated the eight loci using all polymorphism data (shared and private among species) for each species, and estimated the ratio of current and ancestral effective population size (N_2/N_1), the time of expansion ($t/4N_2$) and the shape of the DFE γ distribution (b) for the NS and NC sites. Means and standard errors of parameter estimates were obtained using 50 bootstraps for each site type and each species. This method assumes that all sites are unlinked, which is in broad agreement with previous studies revealing that *S. peruvianum* and *S. chilense* exhibit high values of the population recombination parameter (Stephan and Langley, 1998; Arunyawat *et al*., 2007). The number of SNPs used for estimations is shown in Table 1 and Supplementary Table S7.

Purifying selection and population structure

Assuming linkage equilibrium and that polymorphism at S sites is mainly driven by past changes in population size and population structure, purifying selection is expected to increase the within-species fixation index

Table 1 Multilocus values of Tajima's D_T per species for the pooled samples (arithmetic means across eight loci)

Species	*Site category*	*Number of SNPs*	*Average* D_T
S. peruvianum	Synonymous	278	−0.825
	Silent	616	−0.985
	All sites	742	−1.061
S. chilense	Synonymous	189	−0.227
	Silent	402	−0.301
	All sites	518	−0.527
S. habrochaites	Synonymous	84	−0.322
	Silent	202	−0.526
	All sites	260	−0.757
S. arcanum	Synonymous	171	−0.809
	Silent	377	−0.704
	All sites	457	−0.731

Abbreviation: SNP, single-nucleotide polymorphism.
SNPs are grouped by categories: synonymous, silent sites (that is, non-coding and synonymous) and all sites. The number of SNPs for each class is indicated for each species.

F_{ST} at the NS and NC sites, compared with S sites (Charlesworth *et al*., 1997; Whitlock, 2003). This is because negative selection changes the depth and shape of the underlying coalescent, reducing the effective population size with effects analogous to the increased rates of genetic drift in smaller populations (for example, reduced nucleotide diversity, higher F_{ST}). However, very strong purifying selection (markedly reduced nucleotide diversity) and linkage disequilibrium can decrease the within-species fixation index F_{ST} at the NS and NC sites, compared with S sites (Pamilo *et al*., 1999).

We estimated F_{ST} values within species (that is, using all population samples available per species) across all polymorphic SNPs, using the BayeScan program by Foll and Gaggiotti (2008). For each type of site (S, NS and NC), we compared the distribution of F_{ST} values for all polymorphic SNPs within each species. Because of the non-normal distribution of F_{ST} values, non-parametric statistical tests were used to compare the F_{ST} distributions for the different types of sites. The effect of site type (that is, NC, S or NS) on the distribution of F_{ST} was evaluated using a one-way Kruskal–Wallis test. If the effect of site type was significant at the 5% level, we used pairwise Wilcoxon tests to determine which type of site exhibits higher F_{ST} values (R Development Core Team, 2005).

Results

Evidence for purifying selection

The pooled population samples show an excess of low-frequency variants at S sites for *S. peruvianum* and *S. arcanum*, indicating marked species-wide expansions (strongly negative D_T; Table 1). A comparison of S and silent sites (S + NC) indicates that NC regions exhibit a slight excess of low-frequency polymorphisms at the species level, except in *S. arcanum*. Considering all polymorphic sites (S + NC + NS) together yields lower (more negative) D_T values than those for silent sites, indicating that purifying selection acts on coding regions by keeping NS deleterious mutations at low frequency (except in *S. arcanum*; Table 1; Supplementary Table S3). Methods based on divergence show low power to detect signatures of selection at these loci. The K_a/K_s ratios are lower than one for all species and all loci (Supplementary Table S5), and loci CT166 and CT208 contain no or very few NS SNPs. However, the McDonald–Kreitman test does not show significant departures from neutrality, except for two marginally significant loci: CT066 in *S. habrochaites* and CT198 in *S. arcanum* (Supplementary Table S4). These tests of neutrality indicate that the strength of purifying selection varies between loci (Supplementary Tables S3–S5). Note, however, that the signature is identical across species. For example, loci CT208 and CT166 show a similar lack of NS polymorphism, i.e. evidence for strong purifying selection, in all four species.

Values of NS*/S* and NC*/S* indicate the presence of purifying selection acting on these genes. In coding regions, 85–89% of the sites are under purifying selection in all species (Supplementary Table S6). In introns, 24–30% of the sites are under purifying selection in *S. arcanum*, *S. chilense* and *S. habrochaites*, but this proportion reaches 62% for *S. peruvianum* (Supplementary Table S6).

We find no evidence of positive selection acting on these eight genes. Both the methods of Bierne and Eyre-Walker (2004) and Eyre-Walker and Keightley (2009) infer the proportion of adaptively driven substitutions (α) to be negative or not different from zero in all species. Indeed, α values estimated by the method of Eyre-Walker and Keightley (2009) are included in the confidence intervals obtained by the DoFE method of Bierne and Eyre-Walker (2004) and are centered around zero (Supplementary Table S8).

Distribution of fitness effects

Estimates from simulated datasets: On the basis of only 100–200 SNPs, demographic and DFE parameters are mostly overestimated by Eyre-Walker and Keightley's (2009) method. In fact, population expansion is inferred in cases with no expansion and the expansion factor is overestimated in simulated expansion scenarios (Table 2). The root mean square error of estimates of the expansion factor increases with increasing deleterious mutation effects (increasing $-N_2E(s)$; Table 2) and the simulated time of the expansion ($t/4N_2 = 0.8$) was correctly estimated only once for $-N_eE(s) = 5$ (Table 2). Therefore, the method of Eyre-Walker and Keightley (2009) cannot be used to derive meaningful demographic estimates for our wild tomato species using the current datasets.

We find that the mean and the shape of the distribution of fitness effects cannot be retrieved with such a small number of SNPs when assuming a shape $b = 0.1$ of the γ distribution. The estimated means ($-N_eE(s)$) are not informative, as they range from very small to extremely large values for all simulated datasets (Supplementary Table S9). Therefore, the shape b of the DFE is always overestimated, except when high $-N_eE(s)$ values are simulated (500 and 5000 in Table 2). The log likelihood ratios for the fit to the data are higher for simulated datasets without population expansion and with low selection coefficients ($-N_eE(s) = 5$ or 50; Supplementary Table S9). For our low number of SNPs, the method of Eyre-Walker and Keightley (2009) has no statistical power to jointly estimate the shape and the scale of the DFE, because they both influence the mean and the variance of the DFE (Keightley and Eyre-Walker, 2010; PD Keightley, personal communication).

Table 2 Results of power analyses for estimates of demographic and DFE parameters

$-N_eE(s)$	*Expansion*	N_2/N_1	*RMSE* (N_2/N_1)[a]	$t/4N_2$	*RMSE* $(t/4N_2)$[a]	*b*	*RMSE (b)*
0 (Neutral)	None	4.37 (0.5)	—	4.37 (1.57)	—	8.39 (3.84)	28.42
	Fivefold	4.97 (0.1)	3.57	1.76 (0.5)	3.71	17.93 (5.1)	3.71
5	None	6.63 (0.46)	—	0.8 (0.28)	—	20.68 (5.6)	44.72
	Fivefold	7.08 (0.42)	3.6	0.77 (0.10)	0.73	17 (5.17)	0.73
50	None	6.9 (0.47)	—	1.29 (0.32)	—	17.4 (5.13)	40.18
	Fivefold	5.95 (0.41)	3.07	0.53 (0.08)	0.67	6.06 (3.35)	0.67
500	None	7.03 (0.41)	—	0.5 (0.16)	—	7.89 (3.72)	27.43
	Fivefold	7.6 (0.51)	4.47	1.3 (0.08)	0.73	0.13 (0.02)	0.73
5000	None	6.57 (0.46)	—	0.8 (0.26)	—	0.09 (0.01)	0.07
	Fivefold	8.15 (0.45)	4.48	1.38 (0.1)	0.9	0.095 (0.008)	0.9

Abbreviations: DFE, distribution of fitness effect; RMSE, root mean square error.
Estimated parameters are the ratio of current and ancestral effective population size (N_2/N_1), the time of expansion ($t/4N_2$), and the shape of the DFE γ distribution (b). Means of estimates (standard error) and RMSE were calculated using 50 independent simulated datasets with true parameter values: $-N_eE(s)=0$, 5, 50, 500 and 5000, $N_2/N_1=5$, $t/4N_2=0.8$ and $b=0.1$.
[a]The RMSE of the expansion factor or time of expansion was not calculated in simulations with constant population size.

This means that on the basis of the joint estimates of $-N_eE(s)$ and b, one cannot distinguish neutral mutations from purifying selection if the γ distribution is leptokurtic and skewed toward very strongly deleterious mutations ($b=0.1$). The shape of the DFE, b, can be estimated for strong purifying selection ($-N_eE(s)>500$). We observed that estimations of $-N_es$ do not show such extremely large values when the shape of the γ distribution is 50 (data not shown). The mean log likelihood of the estimates is higher than −1300, much higher than that for $b=0.1$, but joint estimates of the shape and the scale are still unreliable (data not shown).

To circumvent the lack of power to jointly estimate the mean and variance of the DFE, we test the accuracy of summarizing the DFE for comparing distributions with different $-N_eE(s)$ and variances. The DFE can be summarized as four values representing the proportions of mutations falling into given ranges of $-N_eE(s)$, as used by Gossmann *et al*. (2010) and Keightley and Eyre-Walker (2010). We show that this summary of the DFE can discriminate between the different purifying selection regimes (Figures 1a and b). The means and standard errors for the estimated proportions of neutral mutations (in the range $0<-N_es<1$) and of very strongly deleterious mutations ($-N_es>100$) are statistically significant between neutral simulations and those with purifying selection (Figure 1). We can thus statistically distinguish between the DFE shape of purifying selection regimes differing by two orders of magnitude: between $-N_es=5$ and $-N_es=500$ or between $-N_es=50$ and $-N_es=5000$ (Figure 1; Tukey honestly significantly different multiple test at $P<0.01$). Note also that there is a larger difference in the proportion of neutral mutations ($0<-N_es<1$) between the regimes of purifying selection under population expansion (Figure 1b versus 1a). Furthermore, combining results in Figures 1a and b, we demonstrate a significant difference between the DFE shape of $-N_es=5$ with a fivefold expansion and $-N_es=500$ without expansion, or between $-N_es=50$ with expansion and $-N_es=5000$ without expansion (pairwise *t*-test, $P<0.01$). Therefore, despite non-accurate estimation of demographic scenarios, we can distinguish weak purifying selection under population expansion from strong purifying selection in a stationary population.

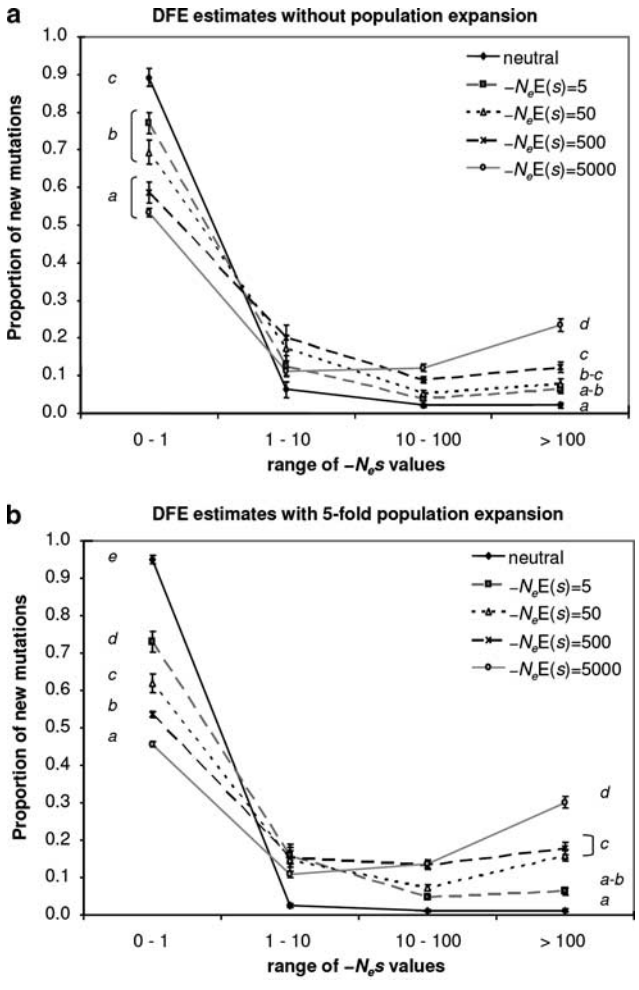

Figure 1 Estimates of the proportions of mutations in different $-N_es$ ranges for simulated datasets with various purifying selection coefficients. (**a**) Under constant population size; (**b**) under a fivefold population expansion, starting in the past at time 0.8 (in units of $4N_2$ generations). Neutral mutations (black solid line with diamonds), $-N_eE(s)=5$ (long dashed gray line with rectangles), $-N_eE(s)=50$ (short dashed black line with triangles), $-N_eE(s)=500$ (long dashed black line with crosses) and $-N_eE(s)=5000$ (gray solid line with circles). Means ± s.e. are shown for 50 independent simulated datasets for each selection coefficient (means sharing a letter are not significantly different at the 0.01 level according to a Tukey's honestly significantly different test). The demographic parameters are $N_2/N_1=5$ and $t/4N_2=0.8$, and the shape of the γ distribution (DFE shape) $b=0.1$.

Estimates for wild tomato species: For NS and NC sites, the DFE has a negative mean for all species ($-N_eE(s)$),

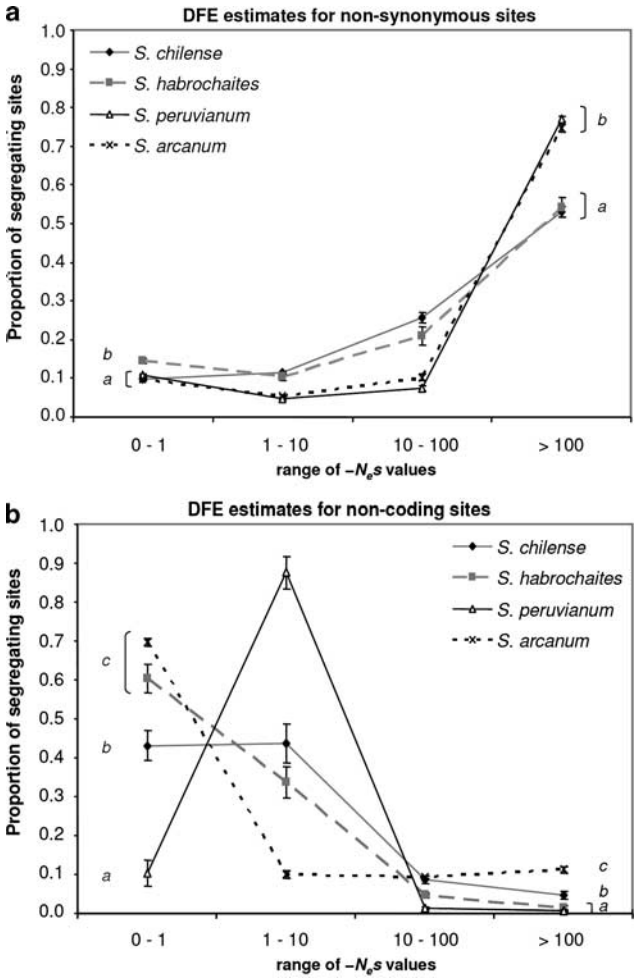

Figure 2 Proportions of mutations in different $-N_e s$ ranges estimated for the four wild tomato species. (**a**) DFE parameters for non-synonymous polymorphisms; (**b**) DFE parameters for non-coding polymorphisms. *S. chilense* (gray solid line with diamonds), *S. habrochaites* (long dashed gray line with rectangles), *S. peruvianum* (black solid line with triangles) and *S. arcanum* (short dashed line with crosses). Means ± s.e. were calculated for each species using 50 bootstraps (means sharing a letter are not significantly different at the 0.01 level according to a Tukey's honestly significantly different test).

confirming that negative purifying selection is the main force driving the molecular evolution of these loci. Estimates of the shape of the γ distribution show that most species have a leptokurtic DFE distribution, with the notable exception of *S. peruvianum* for NC sites (Table 3, estimated $b \approx 82$). All species exhibit a majority of very strongly deleterious mutations at NS sites, with *S. peruvianum* and *S. arcanum* experiencing stronger purifying selection than the two other species (Figure 2a). The estimates of b and $-N_e E(s)$ do not reveal significant interspecific differences for means and variances of the NS DFE (Table 3; Supplementary Table S10).

The shape of the DFE estimated for the NC (intronic) regions indicates weaker purifying selection than for NS sites. This is demonstrated by the higher estimated proportion of neutral mutations ($0 < -N_e E(s) < 1$ in Figure 2b) and lower proportion of strongly deleterious mutations ($-N_e E(s) > 100$ in Figure 2b), compared with NS sites (and the much lower estimated $-N_e E(s)$; Supplementary Table S10). In contrast to NS sites, we did not observe excessively high estimates of $-N_e E(s)$, indicating that the DFE estimation is more accurate for intronic sites. Purifying selection on NC regions, however, seems to vary between species. *S. arcanum* shows the highest variance in the DFE with many neutral and weakly deleterious mutations (70%), as well as the highest proportion of very strongly deleterious mutations (15%; Figure 2b). *S. peruvianum*, on the other hand, exhibits a very narrow DFE around $-N_e E(s) = 5.1$ with $b = 82$ (Table 3; Supplementary Table S10), indicating weak purifying selection acting in intronic regions. *S. chilense* and *S. habrochaites* exhibit shapes and means of the DFE that are intermediate between these extremes (Figure 2b; Table 3). The DFE of *S. peruvianum* is significantly different from that of the other three species in all ranges of $-N_e E(s)$ (Figure 2b). Note that, we report the ratios of population sizes (N_2/N_1) and times of expansion (t/N_2) for each species based on S sites and DFE calculations at the NS and NC sites (Table 3). However, following the results from our power analyses (Table 2), we may not interpret these values as robust estimates of past demographic history for these species.

Purifying selection and population structure

Purifying selection increases the proportion of private low-frequency polymorphisms within demes and results in higher intraspecific F_{ST} values at NC and NS sites, compared with those for S sites (Charlesworth *et al*., 1997). In *S. peruvianum*, F_{ST} values for S sites are lower than for NS sites but not different from those for NC sites (Figure 3a). In *S. chilense* and *S. habrochaites*, all pairwise comparisons show higher F_{ST} values for NC and NS sites compared with S sites, with NC sites being intermediate between lower F_{ST} values at S sites and higher values at NS sites (Figures 3b and c, and Supplementary Information, Section 2). In these three species, the F_{ST}-based results thus agree with the DFE estimates presented above. Purifying selection seems to be very strong in coding regions, and thus F_{ST} for NS sites is higher than for NC (and S) sites. The similar distribution of F_{ST} values for NC and S SNPs in *S. peruvianum* indicates (at best) weak purifying selection on NC sites (Figure 3a). In contrast, *S. arcanum* shows no significant difference between NS and S sites and only a marginally significant difference between S and NC sites (Figure 3d). Furthermore, and contrary to the general expectation outlined above, NS and NC sites show lower mean F_{ST} values in *S. arcanum* than do S sites.

Discussion

The goal of this study was to measure the strength and distribution of purifying selection at eight housekeeping genes in four closely related species of wild tomatoes. Analyzing the species-wide DFE and patterns of F_{ST} within species, we provide evidence for (i) very strong purifying selection acting on NS sites in all species, (ii) weak to strong purifying selection acting on NC regions (introns), (iii) variability in the strength of purifying selection among species in introns and (iv) very low (or non-existent) levels of positive selection on these genes. In the following we discuss how each analysis supports these general conclusions.

Table 3 Estimates of the ratio of current and ancestral effective population size (N_2/N_1), the time of expansion ($t/4N_2$) and the shape of the DFE γ distribution (b) for pooled samples of the four wild tomato species

Species	*Non-synonymous sites*			*Non-coding sites*		
	N_2/N_1	$t/4N_2$	b	N_2/N_1	$t/4N_2$	b
S. peruvianum	4.2 (0.27)	0.17 (0.03)	0.16 (0.01)	6.03 (0.44)	0.37 (0.07)	82 (5.3)
S. chilense	8.12 (0.27)	0.04 (0.04)	0.35 (0.02)	7.7 (0.36)	2.48 (0.52)	29.6 (5.99)
S. habrochaites	1.68 (0.28)	8.3 (1.69)	0.26 (0.03)	4.44 (0.42)	0.88 (0.46)	24.7 (5.85)
S. arcanum	5.77 (0.45)	0.36 (0.05)	0.21 (0.02)	3.85 (0.39)	0.38 (0.1)	0.07 (0.021)

Abbreviation: DFE, distribution of fitness effect.
The means (standard error) of parameter estimates were calculated using 50 bootstraps.

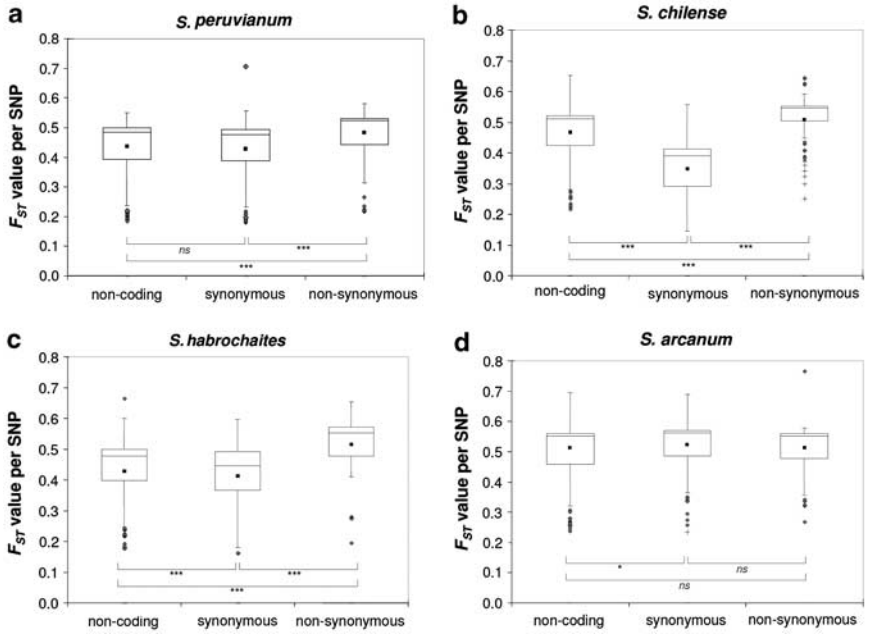

Figure 3 Boxplots of F_{ST} distributions for non-coding, synonymous and non-synonymous polymorphisms. (**a**) *S. peruvianum*, (**b**) *S. chilense*, (**c**) *S. habrochaites*, (**d**) *S. arcanum*. Whiskers extend from 0.25 to 0.75 quartiles. F_{ST} values that are >1.5 times the interquartile range from the nearest quartile are displayed as diamonds, and for >3 times the interquartile range, they are displayed as crosses. The means of distributions are indicated by full black rectangles. Results of pairwise Wilcoxon tests between the three classes of polymorphisms are indicated below the boxplots as follows: *$P<0.1$; ***$P<0.001$; ns, non-significant (P-values are Bonferroni-corrected by the number of pairwise comparisons).

Estimating the strength of purifying selection

We assessed the effect of purifying selection at the species level by using the pooled sample of all populations for each species. On the basis of site-frequency spectrum analysis (DFE estimates and D_T), purifying selection characterized by a negative mean of the DFE generates an excess of low-frequency polymorphisms (Table 1; Fay *et al*., 2001; Whitlock, 2003; Keightley and Eyre-Walker, 2007). This explains why we observe a decrease in D_T values when comparing the different types of sites (S, NC and NS; Table 1). Stronger purifying selection on NS sites compared with introns is corroborated by more negative $N_eE(s)$ estimated for the NS DFE in *S. peruvianum*, *S. chilense* and *S. habrochaites* (Figure 2a versus Figure 2b). Such approaches based on the site-frequency spectrum are more powerful to infer parameters of purifying selection, compared with using ratios of polymorphism to divergence between species as in the McDonald–Kreitman test (McDonald and Kreitman, 1991). However, such methods rely on discriminating the signature of selection from that of demographic scenarios (Fay *et al*., 2001; Keightley and Eyre-Walker, 2007; Eyre-Walker and Keightley, 2009). In our case, past demographic expansions have created an excess of low-frequency polymorphism at the species level (Städler *et al*., 2008, 2009), which has to be taken into account when measuring the DFE (Table 3). There is general agreement on the strength of purifying selection between our DFE estimates and the methodologically independent fixation index analyses for three of the four species (F_{ST} for *S. chilense*, *S. peruvianum* and *S. habrochaites*), although inconsistency is found for *S. peruvianum* between the DFE and F_{ST} results and values of Tajima's D_T in NC regions (Table 1).

Note that both the DFE and F_{ST} analyses assume, perhaps unrealistically, complete linkage equilibrium between sites. A theoretical study showed that linkage disequilibrium changes F_{ST} values at S sites compared with NS sites if the recombination rate is five to ten times lower than the mutation rate (Pamilo *et al*., 1999). Nevertheless, to the extent that purifying selection on NS sites affects the frequency of S polymorphisms due to

linkage disequilibrium (background selection, Charlesworth *et al*., 1997), our analyses underestimate the strength of purifying selection. This is because S sites would then exhibit a site-frequency spectrum biased toward low-frequency variants (lower Tajima's D_T) and higher F_{ST}, compared with a scenario with complete linkage equilibrium among sites.

For *S. arcanum*, however, strong purifying selection is inferred at NS and NC sites from the DFE analysis, but F_{ST} estimates for these site classes are lower than for S sites. This result indicates that only neutral mutations that increase in frequency due to drift are shared among populations, whereas there is a lack of private low-frequency polymorphisms. This pattern is also seen in Tajima's D_T, which is similar for S, silent and all sites (Table 1 and Supplementary Table S3). In other words, the *S. arcanum* site-frequency spectrum does not indicate higher proportions of low-frequency variants at NC and NS sites in excess of that seen at S sites. However, the DFE estimates take into account not only the difference in skew of the site-frequency spectrum between S and NC or NS sites, but also the ratio of segregating sites to potential total sites (NS* or NC*). In fact, when considering the NS*/S* (and NC*/S*) ratios, purifying selection is reflected by a paucity of segregating sites at NS (or NC) sites compared with S sites. Our interpretation, thus, is that purifying selection in *S. arcanum* is very strong, preventing deleterious mutations from segregating in the populations. As the effective population size of this species is smaller than that of *S. peruvianum* and *S. chilense*, one deduces that $-E(s)$ should thus be large in *S. arcanum* (compared with *S. chilense* and *S. peruvianum*). Following results by Pamilo *et al*. (1999), we suggest that a small effective size, very strong purifying selection and potential linkage disequilibrium in this species might account for the lower F_{ST} at NC and NS sites compared with S sites.

Selection in coding regions

At the eight genes studied here, we find high levels of functional constraint as seen by the elevated deleterious effects of NS mutations. Approximately 90% of the NS sites seem to be under selection (Supplementary Table S6). These loci are housekeeping genes, chosen because they are conserved among species (Roselius *et al*., 2005), and thus it is unsurprising that they exhibit high levels of purifying selection in coding regions for maintaining their functions. Strong purifying selection was also found in coding regions of other plant species (Gossmann *et al*., 2010; Slotte *et al*., 2010).

In coding regions, many new mutations are either nearly neutral or very strongly deleterious. Nearly neutral mutations are expected to segregate in populations with frequencies mainly determined by genetic drift, following the nearly neutral theory (Ohta, 1976). Very strongly deleterious mutations are expected to be eliminated, and thus should not reach high frequency. Recent empirical findings provide evidence for such a shape of the DFE; Fudala and Korona (2009) suggested that it is an intrinsic property of mutations in coding regions to be either nearly neutral, due to the redundancy of protein function in biochemical pathways, or nearly lethal if they hit key functions/amino acids in a pathway/protein. Very strongly deleterious mutations may occur at active sites of the protein, and nearly neutral ones at functionally less important sites.

There is no evidence for positive selection acting on these genes in the four wild tomato species (α not different from zero; Supplementary Table S8). Two explanations can be proposed. First, the genes were initially chosen because they are conserved between species and mainly encode housekeeping enzymes. As such they do not represent a random sample of the genome-wide DFE for these species. Second, these species exhibit spatial structuring of populations (for example, Arunyawat *et al*., 2007; Städler *et al*., 2008). Under population structure, mutations responsible for local adaptation are not considered in species-wide measures of fixed adaptive substitutions. Strong spatial structuring is thus suggested to explain the lack of positive selection found by Gossmann *et al*. (2010) in various plant species, in contrast to results in *Capsella grandiflora*, a species with low population subdivision that appears to exhibit high rates of adaptive substitutions ($\alpha = 40\%$; Slotte *et al*., 2010).

Selection in NC regions

We find that purifying selection also appears to impact NC regions (introns) and its strength to be variable among species. Purifying selection is inferred to be strong in *S. arcanum* and slightly weaker in *S. chilense* and *S. habrochaites*. In *S. peruvianum*, the estimated DFE indicates weak purifying selection with a small variance of the DFE, corroborated by non-significant F_{ST} differences between S and NC sites. Natural selection on NC regions was found in several species (for example, human and mouse (Jareborg *et al*., 1999); *Arabidopsis thaliana* (Thomas *et al*., 2007) and poplars (Olson *et al*., 2010)). As intronic regions are known to be of importance in the regulation of expression, splicing and protein synthesis, it is not surprising to find purifying selection on introns, although apparently much weaker than on NS sites (Jareborg *et al*., 1999).

We verified that our results do not depend on one particular locus being subject to strong selection in the NC regions of *S. arcanum*. We thus used PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/), a database to detect described regulatory elements in DNA sequences, and calculated the number of regulatory motifs disrupted by SNPs or indels. *S. arcanum* appears to show fewer mutations and indels disrupting motifs than *S. chilense*, supporting the interpretation that purifying selection is stronger on introns of *S. arcanum* (Supplementary Information, Section 3). However, none of the loci show significantly high numbers of disrupted motifs across this suite of species.

Ecological differences and selection between species

Comparing the four tomato species with regard to the strength of selection also indicates the adaptive potential of a species and selective constraints imposed by the environment. These four species are found in contrasting environments (for example, mesic/dry, high/low altitude and warm/cold) that generate gradients of abiotic stresses (Nakazato *et al*., 2010). We note that the inferred DFE for NS sites is similar for the *S. peruvianum*/*S. arcanum* and the *S. chilense*/*S. habrochaites* pairs (Figure 2a). This significant difference in DFE shape

and/or variance between the pairs of species is not explained, however, by differences in effective population size. Indeed, each pair contains one species with a large N_e (*S. peruvianum* and *S. chilense* with $N_e > 10^6$; Roselius *et al*., 2005; Städler *et al*., 2008) and one species with an $\sim$50% lower N_e estimate (*S. arcanum* and *S. habrochaites*, respectively; CM, AT, WS, TS, unpublished data). As argued above, we are also confident that the significant differences in inferred strength of purifying selection between species (Figure 2) are robust to the bias in demographic estimates. For example, the stronger purifying selection in *S. arcanum* compared with other species is not due to a bias in (under-)estimating its expansion factor.

Recent studies suggest that *S. peruvianum* and *S. habrochaites* exhibit more generalist ecological attributes compared with the two other species. Both species have large geographic ranges and wide niche breadth, encompassing diverse habitats ranging from mesic to dry (Nakazato *et al*., 2010). A screen of candidate genes for abiotic stress response in *S. peruvianum* has found no evidence for molecular signatures of local adaptation (Xia *et al*., 2010). *S. peruvianum* has the second largest geographic range of the four studied species (Nakazato *et al*., 2010), the largest effective population size, and exhibits signatures of a marked demographic expansion (Städler *et al*., 2008). Under such characteristics, geographically varying phenotypic plasticity rather than local adaptation can be promoted (Sultan and Spencer, 2002). We thus propose that lower constraints on the expression of housekeeping genes in *S. peruvianum* correlate with phenotypic plasticity, explaining the apparently relaxed constraints on introns in this species (as for *S. habrochaites*). However, strong functional constraints are evident for the coding regions.

On the other hand, *S. chilense* and *S. arcanum* are found in geographically more restricted areas and characterized by narrower ecological niches (Nakazato *et al*., 2010). *S. chilense* has been described as a specialist species with particular adaptations to extremely dry environments (Nakazato *et al*., 2010; Xia *et al*., 2010). Following our previous argument, we suggest that limited phenotypic plasticity, and thus higher constraints on gene expression of housekeeping genes, may characterize species with narrow ecological distribution (Poot and Lambers, 2008). Compared with *S. peruvianum*, the strong purifying selection inferred for *S. arcanum* indicates that intronic features may have been more conserved in this species. This may suggest that the environment, in which *S. arcanum* is found, exerts some important selective pressures on gene regulation/expression and thus on intronic regions. This could also be due to reduced phenotypic plasticity for gene expression at these genes for a species with quite narrow ecological requirements.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank Hilde Lainer and Simone Lange for excellent technical assistance, Stefan Laurent for valuable discussions, and three anonymous reviewers for their comments. We are also grateful to Peter Keightley for help and advice on using his method. Our field collections in Peru were greatly facilitated by administrative and logistic assistance from Asunción Cano and Gabriel Clostre. Collection and export of tomato samples were made possible through permits issued by the Peruvian 'Instituto Nacional de Recursos Naturales'(INRENA), authorization numbers 099-2006-INRENA-IFFS-DCB and 008493-AG-INRENA. This research has been supported by grants Ste 325/9 and Ste 325/13 from the German Research Foundation to WS, a fellowship from the Chinese Scholarship Council to HX, a fellowship from the Bayerische Eliteförderung to IF, and postdoctoral grant I/82752 from the Volkswagen Foundation to AT.

References

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

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

Bierne N, Eyre-Walker A (2004). The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol Biol Evol* **21**: 1350–1360.

Charlesworth B, Morgan MT, Charlesworth D (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.

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

Eyre-Walker A, Keightley PD (2007). The distribution of fitness effects of new mutations. *Nature Rev Genet* **8**: 610–618.

Eyre-Walker A, Keightley PD (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol* **26**: 2097–2108.

Fay JC, Wyckoff GJ, Wu CI (2001). Positive and negative selection on the human genome. *Genetics* **158**: 1227–1234.

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

Fudala A, Korona R (2009). Low frequency of mutations with strongly deleterious but nonlethal fitness effects. *Evolution* **63**: 2164–2171.

Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV *et al*. (2010). Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol* **27**: 1822–1832.

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

Jareborg N, Birney E, Durbin R (1999). Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res* **9**: 815–824.

Ji Y, Chetelat RT (2007). GISH analysis of meiotic chromosome pairing in *Solanum lycopersicoides* introgression lines of cultivated tomato. *Genome* **50**: 825–833.

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

Keightley PD, Eyre-Walker A (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**: 2251–2261.

Keightley PD, Eyre-Walker A (2010). What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Phil Trans Roy Soc Lond* **365**: 1187–1193.

Librado P, Rozas J (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.
Martin G, Lenormand T (2006a). The fitness effect of mutations across environments: a survey in light of fitness landscape models. *Evolution* **60**: 2413–2427.
Martin G, Lenormand T (2006b). A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. *Evolution* **60**: 893–907.
McDonald JH, Kreitman M (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652–654.
Nakazato T, Warren DL, Moyle LC (2010). Ecological and geographic modes of species divergence in wild tomatoes. *Am J Bot* **97**: 680–693.
Nei M, Gojobori T (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**: 418–426.
Ohta T (1976). Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor Pop Biol* **10**: 254–275.
Olson MS, Robertson AL, Takebayashi N, Silim S, Schroeder WR, Tiffin P (2010). Nucleotide diversity and linkage disequilibrium in balsam poplar (*Populus balsamifera*). *New Phytol* **186**: 526–536.
Pamilo P, Palsson S, Savolainen O (1999). Deleterious mutations can reduce differentiation in small, subdivided populations. *Hereditas* **130**: 257–264.
Peralta IE, Spooner DM, Knapp S (2008). The taxonomy of tomatoes: a revision of wild tomatoes (*Solanum* section *Lycopersicon*) and their outgroup relatives in sections *Juglandifolium* and *Lycopersicoides*. *Syst Bot Monogr* **84**: 1–186.
Poot P, Lambers H (2008). Shallow-soil endemics: adaptive advantages and constraints of a specialized root-system morphology. *New Phytol* **178**: 371–381.
R Development Core Team (2005). *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing (ed.): Vienna, Austria.
Roselius K, Stephan W, Städler T (2005). The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics* **171**: 753–763.
Slotte T, Foxe JP, Hazzouri KM, Wright SI (2010). Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Mol Biol Evol* **27**: 1813–1821.
Städler T, Arunyawat U, Stephan W (2008). Population genetics of speciation in two closely related wild tomatoes (*Solanum* section *Lycopersicon*). *Genetics* **178**: 339–350.
Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P (2009). The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**: 205–216.
Städler T, Roselius K, Stephan W (2005). Genealogical footprints of speciation processes in wild tomatoes: demography and evidence for historical gene flow. *Evolution* **59**: 1268–1279.
Stephan W, Langley CH (1998). DNA polymorphism in Lycopersicon and crossing-over per physical length. *Genetics* **150**: 1585–1593.
Sultan SE, Spencer HG (2002). Metapopulation structure favors plasticity over local adaptation. *Am Nat* **160**: 271–283.
Tajima F (1989). Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
Tanksley SD, Ganal MW, Prince JP, de Vicente MC, Bonierbale MW, Broun P *et al*. (1992). High-density molecular linkage maps of the tomato and potato genomes. *Genetics* **132**: 1141–1160.
Thomas BC, Rapaka L, Lyons E, Pedersen B, Freeling M (2007). *Arabidopsis* intragenomic conserved noncoding sequence. *Proc Natl Acad Sci USA* **104**: 3348–3353.
Whitlock MC (2003). Fixation probability and time in subdivided populations. *Genetics* **164**: 767–779.
Wright SI, Andolfatto P (2008). The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Annu Rev Ecol Evol Syst* **39**: 193–213.
Xia H, Camus-Kulandaivelu L, Stephan W, Tellier A, Zhang Z (2010). Nucleotide diversity patterns of local adaptation at drought-related candidate genes in wild tomatoes. *Mol Ecol* **19**: 4144–4154.
Yang ZH, Nielsen R (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**: 32–43.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)

New Phytologist

Adaptation to drought in two wild tomato species: the evolution of the *Asr* gene family

Iris Fischer[1], Létizia Camus-Kulandaivelu[2], François Allal[2] and Wolfgang Stephan[1]

[1]Section of Evolutionary Biology, Department of Biology II, University of Munich (LMU), Grosshaderner Strasse 2, 82152 Planegg-Martinsried, Germany;

[2]CIRAD, Biological System Department – Research Unit 39 'Genetic Diversity and Breeding of Forest Tree Species', Campus international de Baillarguet TA A-39/C, 34398 Montpellier Cedex 5, France

Author for correspondence:
Iris Fischer
Tel: +49 89 218074161
Email: iris.fischer@bio.lmu.de

Received: *4 October 2010*
Accepted: *3 January 2011*

New Phytologist (2011) **190**: 1032–1044
doi: 10.1111/j.1469-8137.2011.03648.x

Key words: *Asr* (ABA/water stress/ripening induced), gene families, local adaptation, tandem array, wild tomatoes.

Summary

• Wild tomato species are a valuable system in which to study local adaptation to drought: they grow in diverse environments ranging from mesic to extremely arid conditions. Here, we investigate the evolution of members of the *Asr* (ABA/water stress/ripening induced) gene family, which have been reported to be involved in the water stress response.

• We analysed molecular variation in the *Asr* gene family in populations of two closely related species, *Solanum chilense* and *Solanum peruvianum*.

• We concluded that *Asr1* has evolved under strong purifying selection. In contrast to previous reports, we did not detect evidence for positive selection at *Asr2*. However, *Asr4* shows patterns consistent with local adaptation in an *S. chilense* population that lives in an extremely dry environment. We also discovered a new member of the gene family, *Asr5*.

• Our results show that the *Asr* genes constitute a dynamic gene family and provide an excellent example of tandemly arrayed genes that are of importance in adaptation. Taking the potential distribution of the species into account, it appears that *S. peruvianum* can cope with a great variety of environmental conditions without undergoing local adaptation, whereas *S. chilense* undergoes local adaptation more frequently.

Introduction

Adaptation is characterized as the movement of a population towards a phenotype that leads to the highest fitness in a particular environment (Fisher, 1930). The genetic basis of adaptation in natural populations, however, remains broadly unknown (Orr, 2005). Many approaches using DNA variation have been developed to detect adaptation in several organisms. Recently, these methods have been applied to plant model systems (reviewed by Siol *et al.*, 2010). One commonly used method to detect adaptation at the DNA level is to identify regions of low diversity that are linked to a selected gene (known as the hitchhiking effect; Maynard Smith & Haigh, 1974). Such an approach has been used successfully in species including *Drosophila melanogaster* (Glinka *et al.*, 2006; Beisswanger & Stephan, 2008) and sunflower (*Helianthus annuus*; Kane & Rieseberg, 2008). Another way to study adaptation is to use the 'candidate gene' approach, which has the advantage of revealing the strength and type of the selection that has acted on particular genes. With this approach, genes that have been identified in previous experiments are chosen to investigate signatures of adaptation at the DNA level. The candidate gene approach has been applied frequently to several plant species. Examples include: genes related to drought and salt tolerance in sunflower (Kane & Rieseberg, 2007), immunity genes in wild tomatoes (*Solanum* sp.; Rose *et al.*, 2007) and maize (*Zea mays*; Moeller & Tiffin, 2008), drought-responsive genes in maritime pine (*Pinus pinaster*; Eveno *et al.*, 2008), genes related to cold tolerance in *Arabidopsis thaliana* (Zhen & Ungerer, 2008), cold hardiness-related genes in coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*; Eckert *et al.*, 2009), protease inhibitor genes in poplar (*Populus balsamifera*; Neiman *et al.*, 2009), cold-related genes in Scots pine (*Pinus sylvestris*; Wachowiak *et al.*, 2009), and genes involved in adaptation to serpentine soils in *Arabidopsis lyrata* (Turner *et al.*, 2010). Although the candidate gene approach is used frequently, it is important to note that the

distinction between local (selective) and genome-wide (demographic) effects is not always unambiguous. This may be the case in populations with a large effective population size (e.g. *Drosophila* or bacteria; Charlesworth & Eyre-Walker, 2006; Ellegren, 2009) and/or when recurrent positive selection occurs. Thornton *et al.* (2007) showed that some demographic scenarios can also mimic selective events on a local scale.

Duplicated genes are an important source of adaptation, especially in plants, where a large fraction of diversity is the result of gene duplication and subsequent adaptive specialization of paralogous gene copies (Flagel & Wendel, 2009). Paterson *et al.* (2006) found that duplicates of transcription factors in *A. thaliana* are preferentially retained after polyploidy. For the common example of MADS-box transcription factors, it has been suggested that gene duplication followed by increased opportunity for novel gene interactions played an important role in early angiosperm diversification (Shan *et al.*, 2009). Tandemly arrayed genes are of particular interest for studying adaptation to drought, as it has been shown that genes involved in stress responses are often tandemly duplicated (Maere *et al.*, 2005; Mondragon-Palomino & Gaut, 2005; Rizzon *et al.*, 2006; Hanada *et al.*, 2008). In a study of expression data in *A. thaliana*, Zou *et al.* (2009) found that a gain (or loss) of stress responsiveness is more common in tandemly duplicated genes than in nontandem duplicates. Moreover, gene families likely to be involved in lineage-specific adaptive evolution are mainly generated by tandem duplication (Hanada *et al.*, 2008). According to Ohno's (1970) classical model, selective constraints remain on one copy after a gene duplication event, whereas the other copy can accumulate mutations. The most common fate of this latter copy is a loss of function. In rare cases, however, a mutation can be advantageous in a specific environment, leading to a neofunctionalization of one copy (Beisswanger & Stephan, 2008) or subfunctionalization of both copies (Hughes, 1994). Concerted evolution, in which two copies of a gene do not evolve independently, can also be observed. The most common mechanism causing this phenomenon is gene conversion, whereby two copies exchange short tracts of DNA in a 'copy-and-paste' manner. This will both decrease the sequence variation between copies and increase the genetic variation within the gene family by creating new haplotypes (Takuno *et al.*, 2008). New (chimeric) haplotypes can be advantageous in genes that experience diversifying selection, as they increase the genetic diversity (Takuno *et al.*, 2008; Innan, 2009). Therefore, it is possible that the members of a gene family may exhibit very different evolutionary histories.

As plants are sessile, they experience strong selective pressure to adapt to their environment. Drought stress is one of the major abiotic constraints that terrestrial plants face. Because the threat of climate change has become more apparent in recent years, the elucidation of drought adaptation mechanisms in plants is a burning issue (McMichael, 2001); especially in arid and semi-arid areas, where water stress is an important determinant of species distribution. Drought tolerance involves many physiological mechanisms and is thus a complex trait determined by many genes. In a 7000 cDNA micro-array experiment, Seki *et al.* (2002) showed that the expression of > 250 genes was induced in *A. thaliana* after a drought stress treatment.

The genes used in this study are the members of the *Asr* (ABA/water stress/ripening induced) gene family. *Asr* genes encode small (*c.* 13-kDa), highly charged proteins and are grouped in the protein class hydrophilins, which are characterized by a high glycine content and high hydrophilicity. *Asr* transcripts were first discovered in tomato (*Solanum lycopersicum*; Iusem *et al.*, 1993; Rossi & Iusem, 1994), where Frankel *et al.* (2006) described four copies on chromosome IV that lie in a tandem array. In the same study, Frankel *et al.* (2006) described an insertion of 186 amino acids (aa), containing 10 imperfect repeats, that is present in the ASR4 protein, but absent in the other ASR proteins. *Asr*-like genes are found across the entire plant kingdom (e.g. in pummelo (*Citrus maxima*; Canel *et al.*, 1995), rice (*Oryza sativa*; Vaidyanathan *et al.*, 1999), pine (*Pinus taeda*; Padmanabhan *et al.*, 1997), and ginkgo (*Ginkgo biloba*; Shen *et al.*, 2005)) in various copy numbers, ranging from one in grape (*Vitis vinifera*; Cakir *et al.*, 2003) to six in maize and rice (Frankel *et al.*, 2006). Notably, the *Asr* gene family is absent in *A. thaliana* (Carrari *et al.*, 2004). The *Asr* genes seem to exhibit a particularly high duplication rate in tomatoes, as *Asr3* cannot be found in other Solanaceae (Frankel *et al.*, 2006). While *Asr* genes are involved in many physiological processes, such as fruit ripening and seed maturation (Iusem *et al.*, 1993), accumulating functional evidence suggests that this gene family plays a very important role in adaptation to drought in tomato. Several studies indicate that *Asr* genes display a drought-induced expression pattern in tomato (Amitai-Zeigerson *et al.*, 1995; Maskin *et al.*, 2001; Frankel *et al.*, 2006). *Asr* genes have several functions that help the plant deal with drought stress. In the cytoplasm, the unstructured ASR1 monomers act as chaperons, possibly to prevent proteins losing their structure during desiccation (Konrad & Bar-Zvi, 2008). In the nucleus, ASR1 forms homodimers (Maskin *et al.*, 2007) with a zinc-dependent DNA-binding activity (Kalifa *et al.*, 2004) or, as discovered in a grape ASR protein, heterodimers with drought response element binding (DREB) proteins (Saumonneau *et al.*, 2008). This DNA-binding activity is associated with the modulation of sugar transport activity (Carrari *et al.*, 2004; Frankel *et al.*, 2007; Maskin *et al.*, 2008), which places the *Asr* genes at a key position, given the interaction of sugar and ABA pathways discovered in seed developmental processes (reviewed by Finkelstein & Gibson, 2001) and stress signalling (reviewed by Leon & Sheen, 2003). In this context, two

studies revealed patterns of positive selection at *Asr2* in populations of wild tomato species that grow in dry environments by a phylogenetic (Frankel *et al.*, 2003) or a population genetic approach (Giombini *et al.*, 2009).

Tomato is used as a model plant system in evolutionary biology for several reasons, including the availability of cultivated tomato genomic resources, the recent divergence of the *Solanum* species, their clear phenotypic distinction (Peralta *et al.*, 2008), and the diversity of mating systems (Spooner *et al.*, 2005; Moyle, 2008). Recent taxonomic revision suggested grouping tomatoes (until then grouped in the genus *Lycopersicon*) in the genus *Solanum* (section *Lycopersicon*) together with potato (*Solanum tuberosum*) and the eggplant (*Solanum melongena*; Spooner *et al.*, 1993; Peralta & Spooner, 2001). Most *Solanum* sect. *Lycopersicon* species are native to western South America (Ecuador, Peru and Chile), along the western and eastern Andean slopes (Spooner *et al.*, 2005). According to the latest taxonomical classification, tomatoes consist of 12 wild species and their cultivated relative, *Solanum lycopersicum* (Peralta *et al.*, 2008).

This study focuses on two wild tomato species that show differences in their ecological habitats and features: *Solanum chilense* and *Solanum peruvianum*. *Solanum chilense* is distributed from southern Peru to northern Chile and inhabits arid plains and deserts (Peralta *et al.*, 2008). Its distribution also shows a broad range in elevation from sea level up to 3500 m (Chetelat *et al.*, 2009). The species is known to be robust and drought tolerant and can dwell in hyperarid areas as a result of its well-developed root system (Moyle, 2008; Peralta *et al.*, 2008). In fact, the potential distribution of *S. chilense* is predicted to be mostly determined by the annual precipitation (Nakazato *et al.*, 2010). *Solanum peruvianum* is distributed from central Peru to northern Chile and inhabits a variety of habitats, from coastal deserts to river valleys (Peralta *et al.*, 2008). Furthermore, it may be found at field edges, unlike other *Solanum* species (Chetelat *et al.*, 2009). Studies on both species revealed population subdivision (Roselius *et al.*, 2005). It has been suggested that population structure has played an important role in the evolution of wild tomatoes (Arunyawat *et al.*, 2007) and that the species diverged under residual gene flow (Städler *et al.*, 2005, 2008). Taking the difference in range and habitat of the two species into account, one might expect differences of environmental cues, as described by Xia *et al.* (2010). This diverse environmental distribution makes wild tomato species ideal model organisms with which to study local adaptation.

Our goal was to better understand the nature of the evolutionary forces acting on the members of the *Asr* gene family within the tomato clade. We employed a population genetics approach to analyse a larger data set than in previous studies (Frankel *et al.*, 2003, 2006; Giombini *et al.*, 2009) by using several populations from the two closely related species *S. chilense* and *S. peruvianum*, and have therefore more power to detect local adaptation. During this work, we discovered a new member of the *Asr* gene family that has not previously been described. The fact that demography acts on the whole genome, whereas selection affects only restricted genomic regions, allows us to detect selection on our candidate genes by comparing them to a set of reference loci previously described by Arunyawat *et al.* (2007) and Städler *et al.* (2008).

Materials and Methods

Population samples and reference loci

The divergence between the two closely related species *Solanum peruvianum* L. and *Solanum chilense* (Dunal) Reiche occurred < 0.55 million yr ago (Städler *et al.*,

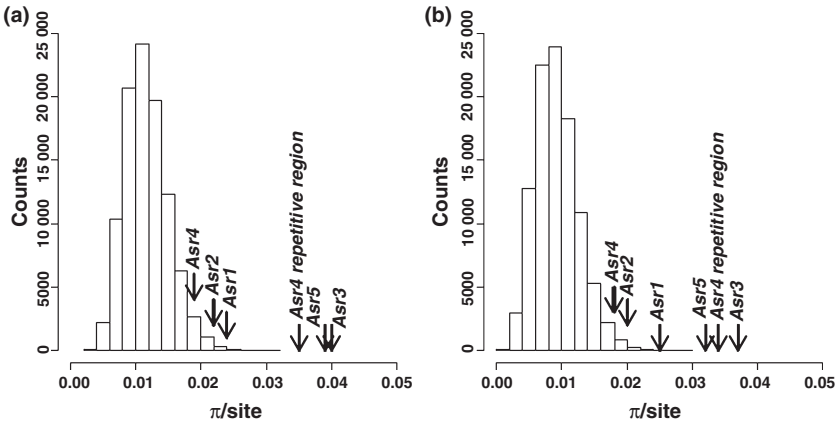

Fig. 1 Distribution of π/site estimated from the reference loci. The distribution was created by 100 000 iterations of coalescent simulations assuming the Wakeley–Hey model (Wakeley & Hey, 1997) for (a) *Solanum peruvianum* and (b) *Solanum chilense*. Input parameters (i.e. θ_1, the Watterson estimator of nucleotide diversity per gene of population 1; θ_2, the Watterson estimator of nucleotide diversity per gene of population 2; θ_A, the Watterson estimator of nucleotide diversity per gene of the ancestral population; τ, time of divergence) were obtained from Städler *et al.* (2008). The observed values for the *Asr* (ABA/water stress/ripening induced) genes are indicated with arrows.

Table 1 Location and habitat characteristics of the sampled populations

Species	Population	Individuals sampled	Location	Latitude, longitude	Annual precipitation (mm)	Precipitation of wettest month (mm)
Solanum chilense	Quicacha	7	Southern Peru	15°38′S, 73°48′W	61	31
	Moquegua	5	Southern Peru	17°04′S, 70°52′W	44	19
	Tacna	7	Southern Peru	17°53′S, 70°08′W	15	5
Solanum peruvianum	Canta	6	Central Peru	11°32′S, 76°42′W	235	77
	Nazca	6	Southern Peru	14°51′S, 74°44′W	103	36
	Tarapaca	5	Northern Chile	18°33′S, 70°09′W	5	1

2008). They are both self-incompatible and relatively drought tolerant compared with other *Solanum* species (Moyle, 2008). *Solanum 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 in Städler *et al.*, 2005). Three populations from climatically different environments were sampled for each species. Five to seven individuals of each population were analysed. The population samples, geographic distributions and habitat characteristics are given in Table 1, and a detailed description of these samples is provided in Baudry *et al.* (2001), Roselius *et al.* (2005), and Arunyawat *et al.* (2007). The Tarapaca sample was obtained from the Tomato Genetics Resource Center at UC Davis, California (accession number LA2744). The other populations were sampled by T. Städler and T. Marczewski in May 2004 (Arunyawat *et al.*, 2007; Städler *et al.*, 2008). *Solanum ochranthum* was used as an outgroup; this sample was provided by L. E. Rose. To distinguish selection from demographic events, the *Asr* genes were compared to seven neutrally evolving reference loci (CT066, CT093, CT166, CT179, CT198, CT251 and CT268), which have been described previously (Arunyawat *et al.*, 2007; Städler *et al.*, 2008).

DNA extraction, PCR amplification and sequencing

Genomic DNA was extracted from tomato leaves using the DNeasy Plant Miniprep Kit (Qiagen GmbH, Hilden, Germany). PCR primers were designed based on information from GenBank (accession number U86130 for *Asr1*, X74907 for *Asr2*, X74908.1 for *Asr3*, DQ058762.1 for *Asr4*, and CU468249 for *Asr5*) using the program NetPrimer (http://www.premierbiosoft.com/netprimer/index.html). To distinguish between *Asr3* and *Asr5*, the sequences were aligned to the *S. lycopersicum* BAC sequence available on EMBL/GenBank (accession number CU468249) using the bl2seq BLAST search. PCR amplification was performed with High Fidelity Phusion Polymerase (Finnzymes, Espoo, Finland). A list of primers and amplification conditions can be found in Supporting Information Table S1. PCR products were confirmed using 1% agarose gel electrophoresis. Direct sequencing of PCR products was performed to identify homozygotes and determine their sequences. For heterozygotes, a cloning approach was used to separate the alleles before sequencing. The Zero Blunt TOPO Cloning Kit (Invitrogen, Carlsbad, CA, USA) was used for cloning and a *Taq* polymerase (Invitrogen) for amplifying the alleles. The DNA was sequenced on an ABI 3730 DNA Analyzer (Applied Biosystems and Hitachi, Foster City, CA, USA). If possible, at least 10 alleles were sequenced for each population using direct sequencing or cloning. Thus, 57 to 65 sequences were obtained for each locus (see Table S2). Sequence chromatograms were manually processed using the SEQUENCHER 4.6 program (Gene Code Corporation, Ann Arbor, MI, USA) or the STADEN v. 4.10 program (http://staden.sourceforge.net/). The sequence alignment of alleles was then constructed using BIOEDIT 7.0.9 (Ibis Biosciences, Carlsbad, CA, USA). Sequence data from this article have been deposited in the EMBL/ GenBank Data Libraries under accession numbers FR727329–FR727636.

Nucleotide diversity analysis

We measured nucleotide diversity with Watterson's (1975) θ_w, the average number of segregating sites per site, and Tajima's (1983) π, the average number of nucleotide differences per site between two sequences, using the DNASP v5 software (Librado & Rozas, 2009). θ_w and π were calculated separately for all sites, nonsynonymous sites, synonymous sites, and noncoding sites for each population. Sliding window analysis of π across the locus was performed in DNASP with a window length of 10 bp and a step size of 1 bp. To determine whether the levels of polymorphism within the species differed significantly from those of the reference loci, coalescent simulations were performed. We ran 100 000 replications to obtain the parameter distribution for the reference loci using MSABC (Pavlidis *et al.*, 2010). The MSABC program applies the approximate Bayesian computation (ABC) methodology to infer demographic events by using summary statistics of gene genealogies that are generated with the MS algorithm (Hudson, 2002). We assumed the Wakeley–Hey 'isolation' model

(WH model; Wakeley & Hey, 1997), in which divergence occurs without subsequent gene flow. Although this assumption might not be valid in wild tomato species, Städler *et al.* (2008) showed that the data for the reference loci fit this simple WH model well. We therefore used the input parameters for the simulations (i.e. θ_1, the Watterson estimator of nucleotide diversity per gene of population 1; θ_2, the Watterson estimator of nucleotide diversity per gene of population 2; θ_A, the Watterson estimator of nucleotide diversity per gene of the ancestral population; and τ, time of divergence) from Städler *et al.* (2008). As the WH model only accommodates two populations, Städler *et al.* (2008) compared two populations in a pairwise manner (comparing one from *S. chilense* with one from *S. peruvianum*). We used the average of these nine comparisons as input parameters. To account for differences in recombination rate between the reference loci, we assumed a recombination rate (R) of 20, which is intermediate (Stephan & Langley, 1998; see also Städler *et al.*, 2008). Graphical visualization of the parameter distributions was performed using the R package (R Development Core Team, 2005). Additionally, the K_a : K_s ratios for each species were calculated in DnaSP, where K_a is the rate of substitutions per nonsynonymous site and K_s the rate of substitutions per synonymous site. To assess whether gene conversion was occurring among copies, the number of shared, fixed and specific polymorphisms between copies was counted (Innan, 2003). Coding, intronic and 5′ untranslated regions (UTR) were analysed separately. A simplified version of the site frequency spectrum was obtained by classifying the polymorphisms in three categories: rare (frequency of minor allele < 5%), intermediate (frequency of minor allele between 5 and 20%), and common (frequency of minor allele between 20 and 50%).

Neutrality tests

We tested for deviations from the standard neutral model using Tajima's (1989) D statistic and Fu & Li's (1993) D (or D^* if no outgroup was available) using DnaSP. A significantly negative value of Tajima's D indicates an excess of rare variants, which is expected under directional selection or population size expansion. A significantly positive Tajima's D value indicates an excess of intermediate-frequency variants, which is expected under balancing selection or under population subdivision. The same is true for Fu and Li's D (and D^*) statistics. Fay & Wu's (2000) H was also calculated, where a significantly negative Fay and Wu's H indicates an overrepresentation of high-frequency derived polymorphisms, which is expected in the case of positive selection. A significantly positive Fay and Wu's H indicates an overrepresentation of intermediate-frequency derived polymorphisms, which is expected under balancing selection. The haplotype test of Depaulis & Veuille (1998) was used to infer haplotype diversity (H_d). All neutrality tests were performed using the option Number of Segregating Sites.

The fixation index F_{st}

The fixation index F_{st} (Hudson *et al.*, 1992; Jost, 2008) was calculated using DnaSP. Assuming there is a similar mutation rate at all reference and candidate genes, F_{st} provides a means to quantify gene flow, where a value close to zero indicates a high migration rate between two populations. The higher the F_{st}, the more the two populations differ genetically. A high F_{st} can therefore be interpreted as evidence for limited gene flow or recent divergence without subsequent gene flow. However, if disruptive selection (i.e. local adaptation) is acting on the candidate genes, F_{st} would be increased compared with the reference loci, because one allele rises to high frequency in one population, but not in the other.

Bioclimatic data and ecological niche modelling

Our knowledge of the distribution of wild tomato species is mostly based on previous sampling schemes and can therefore be biased, as populations in areas that are not accessible might be missed. Using the BIOCLIM (Bioclimate Prediction System) algorithm (Busby, 1986) implemented in the DIVA-GIS software (Hijmans *et al.*, 2001), we predicted the theoretical spatial distribution of *S. chilense* and *S. peruvianum* over South America on a bioclimatic basis. Based on homoclime matching (Booth *et al.*, 1987), the idea of BIOCLIM is to define the climatic rules governing species distributions by creating climatic envelopes from occurrence data, and to identify areas with similar climate profiles (Busby, 1991). Ecological niche modelling of *S. chilense* and *S. peruvianum* was performed using occurrence data derived from the UC Davis database, and all 19 bioclimatic variables available for current conditions (1950–2000) derived from the WorldClim global climate database (Hijmans *et al.*, 2005) at spatial resolutions of 30 arc-second (*c.* 1 × 1 km). The performance of *S. chilense* and *S. peruvianum* distribution models was tested by measuring the area under a relative operating characteristic curve (AUC; Pearce & Ferrier, 2000) using the DIVA-GIS software. A detailed description of this method can be found in Methods S1.

Results

Identification of a new member of the *Asr* gene family in tomatoes

We sequenced the members of the *Asr* gene family in two wild tomato species (*S. chilense* and *S. peruvianum*) and also in *S. ochranthum* as an outgroup. We constructed our primers for the *Asr* genes based on the NCBI (National Center for Biotechnology Information) database, assuming that the

gene family consists of the four members previously described. While sequencing *Asr3*, we discovered a fifth member of the gene family, subsequently called *Asr5*. *Asr3* and *Asr5* have a highly similar coding region but differ greatly in the intronic sequence. As our first set of *Asr3* primers was designed in the coding region, we amplified both genes simultaneously. Therefore, we found four 'alleles' for *Asr3*. We used a BLAST search (bl2seq), where we aligned the four *Asr3* 'alleles' with the *S. lycopersicum* BAC sequence available in NCBI (accession number CU468249). On the basis of the divergent introns, we could then clearly distinguish two *Asr* genes. After designing primers in the flanking region, it was possible to sequence the two genes separately. We annotated the location of the *Asr* genes relative to the *S. lycopersicum* BAC sequence (Fig. S1), where the *Asr* genes lie in tandem, comprising *c.* 40 kb in total. *Asr* genes consist of two exons and an intron of variable length. For all *Asr* genes, the total numbers of sequenced synonymous, nonsynonymous and noncoding sites are given in Table S3.

Gene conversion between *Asr3* and *Asr5*

In general, all *Asr* genes showed a higher level of polymorphism than the reference loci (Table S4). In particular, *Asr3* and *Asr5* exhibited high π and θ_W values compared with the reference loci (Fig. 1, Table S4). The ratio of $\pi_{nonsynonymous}$ to $\pi_{synonymous}$ was also high, and above 1 in some cases (Table S4). A high level of nonsynonymous polymorphism can be explained by several factors. First, the sequences of *Asr3* and *Asr5* could have been mixed up during sequencing. We ruled out this possibility by using specific primers designed in the flanking region. In addition, *Asr3* and *Asr5* could be clearly distinguished when aligning the intronic regions. Secondly, high levels of nonsynonymous polymorphism may be interpreted as a signal of selection. The neutrality tests, however, did not show any evidence of selection at *Asr3* or *Asr5* (Table 2). Another mechanism that could cause this high level of nonsynonymous polymorphism is gene conversion. To investigate this, we counted the number of fixed and shared polymorphisms

Table 2 Results of the neutrality tests

	Asr1	*Asr2*	*Asr3*[b]	*Asr4*	*Asr4* repetitive region	*Asr5*[b]	Reference loci[a] (min, max)[d]
Tajima's *D*							
Solanum chilense[c]	−0.453	−0.832	−0.108	−0.349	−0.668	−0.079	−0.565 (−1.115, −0.070)
Quicacha	−0.136	0.602	1.118	−0.511	−0.989	−0.413	0.132 (−1.377, 0.862)
Moquegua	−0.658	−0.957	−1.147	0.446	0.308	0.148	−0.048 (−0.786, 0.395)
Tacna	0.812	−0.372	0.323	−1.429	−1.445	0.862	0.040 (−1.361, 0.605)
Solanum peruvianum[c]	−0.940	−0.551	−0.781	−1.139	−0.231	−0.946	−1.098 (−1.708, −0.616)
Canta	−0.168	−0.121	−0.758	−0.548	−0.046	−0.625	−0.548 (−1.001, −0.052)
Nazca	0.245	0.267	−0.026	−0.346	0.118	0.507	−0.142 (−1.067, 0.649)
Tarapaca	−0.147	0.356	0.459	0.029	0.393	−0.468	−0.222 (−0.514, 0.069)
Fu and Li's *D*							
S. chilense[c]	−0.989	−0.191	0.082	−1.084	−1.105	−0.849	−1.332 (−3.781, −0.154)
Quicacha	−0.332	−0.121	1.031	0.382	−0.475	−0.897	0.280 (−1.302, 1.325)
Moquegua	−1.333	−0.863	−1.251	0.187	0.184	−0.269	0.109 (−1.225, 0.712)
Tacna	1.191	0.031	0.034	**−2.326***	**−2.218***	0.166	−0.114 (−1.437, 0.513)
S. peruvianum[c]	−0.295	−0.551	−1.752	**−2.635***	−1.403	−1.395	−1.702 (−3.294, −0.614)
Canta	−0.832	−0.590	−0.963	−0.982	−0.302	−0.839	−0.846 (−1.722, −0.123)
Nazca	0.481	1.182	−0.530	−1.411	0.161	0.277	−0.363 (−1.779, 1.397)
Tarapaca	−0.342	0.225	0.222	−0.223	0.107	−0.777	−0.359 (−0.738, 0.660)
Fay and Wu's *H*							
S. chilense[c]	3.319	2.588	–	−5.183	−5.477	–	0.173 (−5.279, 4.724)
Quicacha	5.626	−0.927	–	−8.709	−3.328	–	−1.779 (−6.989, 3.644)
Moquegua	6.333	2.806	–	1.361	2.306	–	−0.749 (−6.222, 8.800)
Tacna	2.330	5.626	–	−5.127	−5.655	–	0.057 (−6.400, 4.788)
S. peruvianum[c]	0.129	4.064	–	1.293	0.904	–	3.240 (−2.146, 7.577)
Canta	4.000	0.444	–	1.867	1.244	–	4.898 (0.349, 9.273)
Nazca	5.511	0.889	–	4.361	2.000	–	1.406 (−1.030, 5.061)
Tarapaca	1.778	0.000	–	0.889	−0.178	–	2.477 (−6.578, 8.000)

[a]Average over seven reference loci.
[b]No outgroup available. Fu and Li's *D** without outgroup was calculated instead.
[c]Pooled over three populations.
[d]Minimum and maximum values of the reference loci are indicated in parentheses.
*$P < 0.05$ (significant results are in bold).

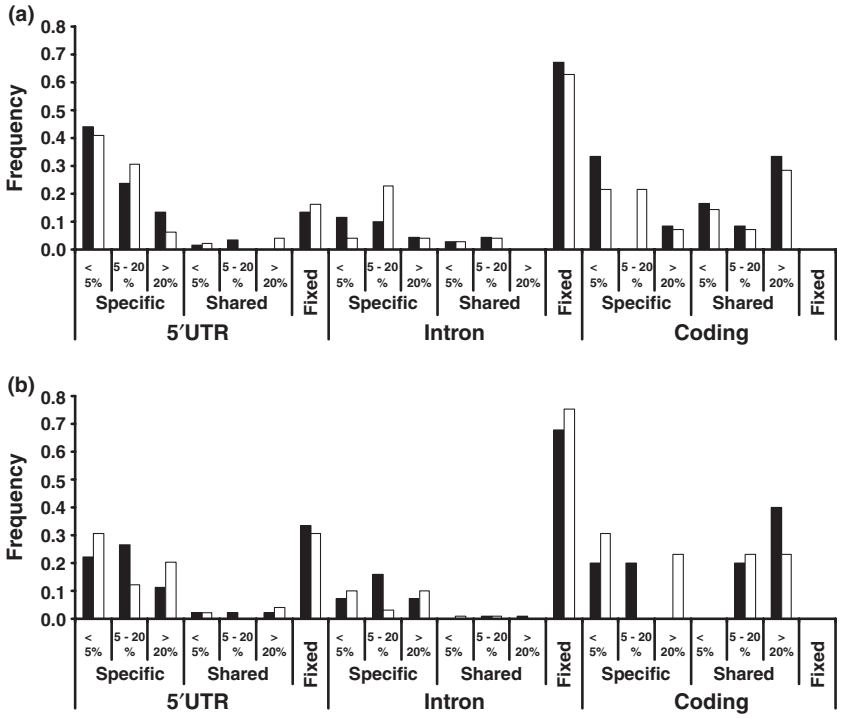

Fig. 2 Numbers of specific, shared, and fixed polymorphisms between *Asr* (ABA/water stress/ripening induced) genes *Asr3* (closed bars) and *Asr5* (open bars). The coding, intronic and 5′ untranslated region (UTR) were analysed separately for (a) *Solanum peruvianum* and (b) *Solanum chilense*. Polymorphism was classified into three categories: frequency of minor allele < 5%, frequency of minor allele between 5 and 20%, and frequency of minor allele > 20%.

Table 3 Haplotype diversity (H_d) of the *Asr* (ABA/water stress/ripening induced) genes and the reference loci

H_d	*Asr1*	*Asr2*	*Asr3*	*Asr4*	*Asr4* repetitive region	*Asr5*	Reference loci[a] (min, max)[c]
Solanum chilense[b]	0.975	0.975	0.977	0.946	0.914	0.985	0.963 (0.929, 0.984)
Quicacha	0.890	0.800	0.821	0.600	0.600	0.891	0.796 (0.538, 0.970)
Moquegua	0.952	0.972	0.944	1.000	0.972	1.000	0.962 (0.911, 1.000)
Tacna	0.956	0.978	0.978	0.982	0.891	1.000	0.955 (0.911, 1.000)
Solanum peruvianum[b]	0.970	0.979	0.982	0.985	0.948	0.981	0.987 (0.966, 1.000)
Canta	0.956	0.933	1.000	0.956	0.956	1.000	0.988 (0.970, 1.000)
Nazca	0.978	0.911	0.891	0.972	0.917	0.893	0.949 (0.773, 1.000)
Tarapaca	0.844	0.956	0.956	0.933	0.933	0.911	0.934 (0.893, 1.000)

[a]Average over seven reference loci.
[b]Pooled over three populations.
[c]Minimum and maximum values of the reference loci are indicated in parentheses.

between *Asr3* and *Asr5*, as well as the polymorphisms specific to one gene. In the coding region, both copies had a high level of shared polymorphism, but no fixed difference (Fig. 2). This is a clear indicator of gene conversion (Innan, 2003). We did not observe this pattern in the 5′ UTR or the intron, where we had a high number of fixed differences, but hardly any shared polymorphisms (Fig. 2).

Evolution of *Asr1* and *Asr2*

Consistent with purifying selection, $\pi_{nonsynonymous}$ was very low at *Asr1* (Table S4). The observed $K_a : K_s$ ratios of 0.006 in *S. chilense* and 0.021 in *S. peruvianum* further support this interpretation. However, we did not find $\pi_{synonymous}$ to be lower at *Asr1* than at the other *Asr* genes (or the reference loci). Additionally, $\pi_{noncoding}$ was not lower than the values observed at the other *Asr* genes, which may explain the high level of polymorphism observed across the whole gene (Table S4).

For *Asr2*, we observed an extremely high value of $\pi_{synonymous}$, especially in *S. peruvianum*, whereas polymorphism was comparable with that of the other *Asr* genes at nonsynonymous and noncoding sites (Table S4). A possible explanation for this pattern could be that *Asr2* is present in multiple copies and our analysis confounded alleles and paralogues. However, we were able to rule this

Table 4 Fixation index F_{st}

F_{st}	*Asr1*	*Asr2*	*Asr3*	*Asr4*	*Asr4* repetitive region	*Asr5*	Reference loci[a] (min, max)[b]
Tacna – Moquegua	0.000	0.026	0.263	0.246	0.335	0.000	0.019 (0.000, 0.044)
Tacna – Quicacha	0.331	0.119	0.195	0.565	0.400	0.449	0.244 (0.006, 0.387)
Moquegua – Quicacha	0.191	0.164	0.311	0.498	0.205	0.469	0.227 (0.025, 0.380)
Tarapaca – Nazca	0.241	0.262	0.115	0.023	0.039	0.118	0.169 (0.042, 0.311)
Tarapaca – Canta	0.155	0.139	0.025	0.039	0.048	0.060	0.067 (0.002, 0.129)
Nazca – Canta	0.194	0.337	0.107	0.007	0.000	0.093	0.135 (0.069, 0.292)

[a]Average over seven reference loci.
[b]Minimum and maximum values of the reference loci are indicated in parentheses.
Asr, ABA/water stress/ripening induced.

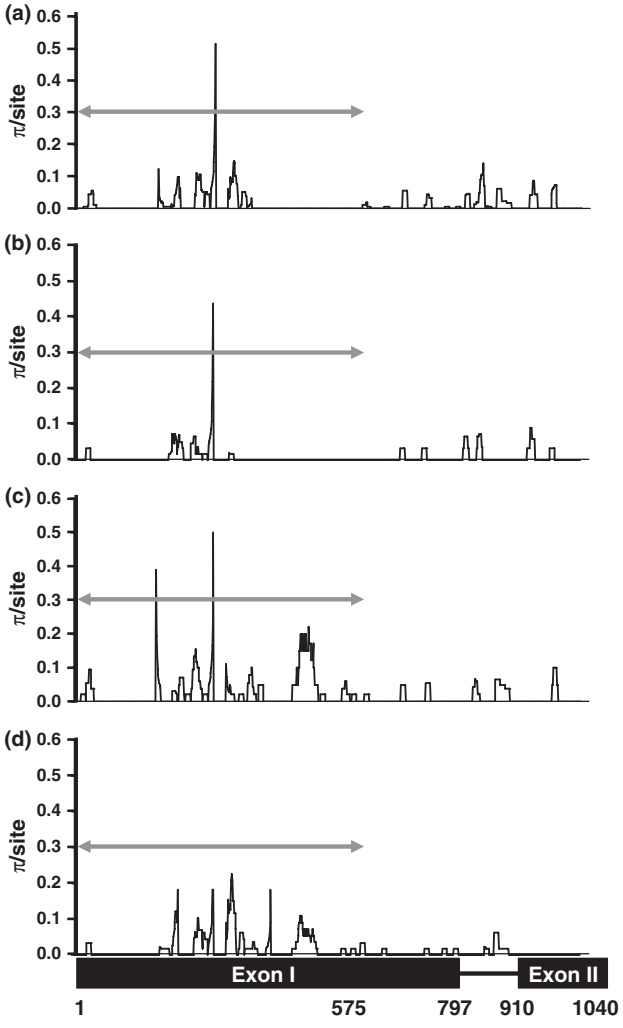

Fig. 3 Sliding window analysis of π/site along the *Asr* (ABA/water stress/ripening induced) gene *Asr4* in *S. chilense*. The analysis was performed with a window length of 10 bp and a step size of 1 bp for (a) *S. chilense* (pooled over three populations), (b) the Quicacha population, (c) the Moquegua population, and (d) the Tacna population. The grey arrows indicate the repetitive region in the first exon; the numbers under the *Asr4* gene indicate the length in bp.

out by re-sequencing *Asr2*. We did not find evidence for adaptation at *Asr2*. Neutrality tests showed no sign of selection (Table 2), nor an unusual haplotype structure (Table 3). Furthermore, we did not observe a higher F_{st} at *Asr2* than at the reference loci (Table 4).

A repetitive region in the first exon of *Asr4*

The beginning of the first exon of *Asr4* contains many insertion/deletion (indel) polymorphisms. The indels always occur in triplets and do not disrupt the open reading frame. The ASR4 proteins in our data set showed insertions of variable length, where the shortest was 50 aa and the longest 147 aa long. We also found imperfect repeats, as described by Frankel *et al.* (2006). For consistency, we refer to this region of the first exon of *Asr4* as the 'repetitive region', although the motifs are not repetitive in all cases. A protein BLAST search revealed no homology to any protein or motif that could clarify the function of this additional domain, but the repeats in the protein alignment showed a recurring SYG motif.

A sliding window analysis revealed that the *Asr4* repetitive region was highly polymorphic, whereas the polymorphism dropped to a low level in the second part of the first exon (Fig. 3). The high π values at the beginning of the first exon in *S. chilense* (Fig. 3a) were mostly caused by the Tacna and Moquegua populations (Fig. 3c + d). In Quicacha and Moquegua (Fig. 3b + c) a peak was observed where π reached a value of 0.5. This was attributable to a single nucleotide polymorphism (SNP) at intermediate frequency, which was present at a low frequency in Tacna (Fig. 3d). Apart from this, Quicacha showed a moderate level of polymorphism in the repetitive region (Fig. 3b). In *S. peruvianum*, we did not detect any difference in polymorphism between the populations (Table S4). Coalescent simulations using MSABC showed that the polymorphism level at the repetitive region was not consistent with π estimated from the reference loci (Fig. 1).

Evidence for adaptation at *Asr4*

Asr4 in *S. chilense* showed strong signatures of adaptation in the Tacna and Quicacha populations. Quicacha displayed a pronounced haplotype structure with a very low H_d (although not outside the reference gene range) and a low number of haplotypes, where one haplotype was predominant. This, together with a negative Fay and Wu's *H* far outside the reference gene range, suggests that positive selection has acted on this haplotype. In Tacna, we also found a predominant haplotype, although the pattern here was not as strong as in Quicacha: Fay and Wu's *H* was negative, but inside the reference gene range (Table 2). This observation is consistent with the very negative Tajima's *D* (outside the reference gene distribution) and significantly negative Fu and Li's *D* values. When the repetitive region was analysed separately, the haplotype structure in Tacna became more pronounced and H_d decreased and fell outside the reference gene range (Table 3). In addition, Fay and Wu's *H* was more negative in Tacna, but less negative in Quicacha (Table 2). In Tacna, Tajima's *D* and Fu and Li's *D* values remained extremely negative.

Signatures of positive selection at *Asr4* were also revealed by pairwise F_{st} analysis. Indeed, in *S. chilense*, F_{st} was elevated at *Asr4* compared with the reference loci and the other *Asr* genes. This pattern was most striking between Tacna and Moquegua. At the reference loci, F_{st} never exceeded 0.05. At *Asr4*, however, F_{st} was 0.246 for the whole gene and 0.335 for the repetitive region (Table 4). This high value was attributable to the fact that the predominant Tacna haplotype was absent in the Moquegua population. This observation is surprising considering the generally high gene flow between these two populations (very low F_{st}). High F_{st} values were also found between Quicacha and the other *S. chilense* populations. As F_{st} at the reference loci was much higher between these populations, this pattern was less striking. The climatic data indicate that the greatest environmental differences occur between the Quicacha and Tacna populations, with Tacna being extremely dry (Table 1). By contrast, we did not observe high F_{st} in *S. peruvianum*; here, F_{st} was lower at *Asr4* than at the reference loci and the other *Asr* genes (Table 4).

Potential geographical distributions of *S. chilense* and *S. peruvianum*

To infer habitat preferences, we predicted the potential geographical distributions of both species by applying the BIOCLIM algorithm to data on the sampling locations of *Solanum* species derived from the Tomato Genetics Resource Center at the UC Davis. In general, the potential fitting area of *S. peruvianum* was broader than that of *S. chilense* (Fig. S2). For *S. peruvianum*, the potential distribution reached the South American east coast (data not shown). It is important to keep in mind that this does not necessarily reflect the actual distribution. *Solanum chilense* was predicted to grow further south, with the best fitting zone at the Atacama Desert (Fig. S2b). By contrast, *S. peruvianum* showed the best fitting zones further north in more mesic environments but also in dry habitats in southern Peru (Fig. S2a). According to AUC calculations, the performance of both species distribution models was good, with AUC = 0.86 for *S. chilense* and AUC = 0.77 for *S. peruvianum*.

Discussion

We searched for signatures of local adaptation in populations of two wild tomato species, *S. chilense* and *S. peruvianum*, in the *Asr* gene family. We used a set of reference loci that were previously studied (Arunyawat *et al.*, 2007) to distinguish between natural selection and demography. While sequencing the four members of the *Asr* gene family previously described, we discovered a new copy, designated *Asr5*, which is highly similar to *Asr3*. We analysed all five copies of the *Asr* gene family and detected patterns consistent with local adaptation in *S. chilense* populations, but not in *S. peruvianum*. We also found that the individual members of the *Asr* gene family showed quite diverse evolutionary histories.

Asr1 exhibited evidence for purifying selection that has been acting on the coding region. The nucleotide diversity was very low in the coding region in both species compared with the reference loci. In addition, *Asr1* showed low divergence from the outgroup, with a $K_a : K_s$ ratio close to zero. The conservation of *Asr1* has been described before (Frankel *et al.*, 2006). In contrast to Frankel *et al.* (2006), however, we did not observe lower synonymous nucleotide diversity at *Asr1* than at the other *Asr* genes. Furthermore, the level of polymorphism at synonymous sites was comparable to that of the reference loci. We found a low level of nonsynonymous polymorphism, such that the protein sequences of ASR1 were almost identical among all studied lines. According to Ohno's (1970) classical model, it can be expected that (at least) one member of the family is under strong evolutionary constraint. One member must be maintained to fulfill the original function of the gene. The fact that *Asr* homologues with a similar protein sequence to tomato *Asr1* can be found throughout the plant kingdom (Frankel *et al.*, 2006) and that *Asr1* is expressed in a housekeeping manner (Maskin *et al.*, 2001; Frankel *et al.*, 2006) suggests that this gene fulfills an important function and that even small alterations in the protein have serious consequences for the plant's fitness.

Previous studies suggest that *Asr2* has adapted to dry environments (Frankel *et al.*, 2003; Giombini *et al.*, 2009). We cannot confirm this hypothesis, as we found no indication for adaptation at *Asr2*, such as patterns of low diversity

within a population and a high fixation index between populations, in our data. In their study, Frankel *et al.* (2003) used a phylogenetic approach to detect adaptation where only a few (four to eight) alleles were analysed for each of the seven *Solanum* species/cultivars investigated. They found a $K_a : K_s$ ratio > 1 in several between-species comparisons and an accelerated rate of evolution in species from dry environments (i.e. *S. chilense* and *S. peruvianum* v. *humifusum*). The Antofagasta sample was again used in the study by Giombini *et al.* (2009), who also found signatures of positive selection in *S. chilense* using a candidate gene approach. However, the Antofagasta sample was found to deviate greatly from equilibrium conditions (Arunyawat *et al.*, 2007). Holliday *et al.* (2010) demonstrated that population bottlenecks can cause skews in the site frequency spectrum that may be confounded with natural selection. In both studies, this was the only sample analysed to infer adaptation in *S. chilense*. It remains possible that local adaptation acted in this population. However, given the results of Arunyawat *et al.* (2007), a demographic effect is more likely to have caused the patterns observed by Frankel *et al.* (2003) and Giombini *et al.* (2009).

In contrast to *Asr2*, we found patterns of local adaptation at the *Asr4* gene in two *S. chilense* populations, Quicacha and Tacna. For Quicacha, the interpretation of the data may be somewhat ambiguous, as in this population the reference loci also showed some indication of population structure (although this was not significant). The evidence for adaptation at Tacna, however, is striking as we did not observe any haplotypic structure at the reference loci in this population. Moreover, F_{st} between Tacna and the other *S. chilense* populations was very high at *Asr4* compared with the reference loci and the other *Asr* genes. An elevated F_{st} is a hallmark of local adaptation, where an allele is favoured in one population but not in others (Beaumont & Balding, 2004; Foll & Gaggiotti, 2008; Riebler *et al.*, 2008). The first part of the first exon seems to play a major role in adaptation and is characterized by a repetitive region with a very high polymorphism level and indels of different length. This repetitive region possesses an SYG motif that has been described to be stress responsive in *Saccharomyces cerevisiae* (Treger & McEntee, 1990). Interestingly, the rice ASR6 protein has an insertion at the N-terminal which displays some sequence similarity to the tomato ASR4 (Frankel *et al.*, 2006), but lacks the SYG motif. The presence of this insertion at an *Asr* gene in a monocot suggests an important function of this region, as it was already present in an ancestral *Asr* gene in early land plants. Alternatively, these domains might also have been gained independently, which would be a remarkable example of convergent evolution. As the zinc-binding sites of the ASR1 protein are located in the first exon (Rom *et al.*, 2006), it may be that variation in this region of ASR4 has an influence on the zinc-binding activity.

Another interesting aspect of this study is the likely occurrence of gene conversion between *Asr3* and *Asr5*. As *Asr3* is exclusive to tomato (Frankel *et al.*, 2006), the *Asr3–Asr5* duplication must have occurred after the tomato cladogenesis. Since only the coding region is the target of gene conversion, it is possible that it is beneficial for the plant to have two similar protein copies. This could be the case in a 'dosage' scenario, in which more proteins of the same kind are favoured. This has been suggested to be the case in stress-responsive genes in variable environments (Kondrashov *et al.*, 2002; Innan & Kondrashov, 2010). Gene conversion can also be advantageous when diversifying selection is acting, as new haplotypes can be created (Innan, 2009). However, detection of signatures of diversifying selection in genes that undergo gene conversion is difficult, because standard neutrality tests cannot be applied (Innan, 2003; Thornton, 2007).

In conclusion, we observed that the *Asr* family is a highly dynamic gene family in which the individual members show quite diverse evolutionary histories. Complex histories of gene families are not uncommon in plants, as adaptive specialization is thought to occur after gene duplication (Flagel & Wendel, 2009). We detected patterns of local adaptation to dry environments at the *Asr4* gene in the *S. chilense* population from Tacna, which has the driest *S. chilense* environment sampled. Interestingly, we did not find any pattern of local adaptation in *S. peruvianum*, although this species inhabits a broader geographical and environmental range than *S. chilense*. This has also been observed in studies of other candidate genes involved in drought tolerance (Arunyawat *et al.*, 2007; Xia *et al.*, 2010). Given the high nucleotide and morphological diversity of *S. peruvianum*, it may be hypothesized that this species can cope with a great variety of environmental conditions, whereas *S. chilense* seems to be undergoing local adaptations more frequently. This observation is supported by the analysis of the potential distribution of *S. peruvianum* and *S. chilense*. *Solanum peruvianum* displays a broader species range than *S. chilense*, covering a large variety of potential habitats. *Solanum chilense*, in contrast, showed the best fitting area at the Atacama Desert, one of the driest regions in the world. These findings are in accordance with a previous study focusing on the potential distribution of several *Solanum* species (Nakazato *et al.*, 2010). A recent study of the *Asr* gene family in rice revealed the same pattern as seen in tomato: the members have different (gene- and species-specific) evolutionary histories and a candidate gene for drought tolerance (rice *Asr3*) could be identified (Philippe *et al.*, 2010). The fact that *Asr3* is a potential candidate for drought adaptation is especially interesting, as it clusters in tandem with rice *Asr4*, whereas the other rice *Asr* genes are distributed over the whole genome (Frankel *et al.*, 2006). This supports recent findings, mainly in *A. thaliana* and rice, that stress-responsive gene families are preferably

clustered in tandem (Maere *et al.*, 2005; Mondragon-Palomino & Gaut, 2005; Rizzon *et al.*, 2006; Hanada *et al.*, 2008; Zou *et al.*, 2009). Our results on *Asr* genes in wild tomato species therefore provide a good example of a tandemly arrayed gene family that is of importance in adaptation to drought.

Acknowledgements

We are grateful to H. Lainer for excellent technical assistance. We thank P. Pavlidis for help with msABC, H. Innan for valuable suggestions on gene conversion analysis, and the Munich Tomato Group, especially A. Tellier, for discussions. We are also grateful to J. Parsch and three anonymous referees for valuable comments that improved the presentation of this paper. This work was funded by grant STE 325/9 from the German Research Foundation to WS and a fellowship from the Graduiertenförderung nach dem Bayerischen Eliteförderungsgesetz to IF.

References

Amitai-Zeigerson H, Scolnik PA, Bar-Zvi D. 1995. Tomato *Asr1* mRNA and protein are transiently expressed following salt stress, osmotic stress and treatment with abscisic acid. *Plant Science* **110**: 205–213.

Arunyawat U, Stephan W, Städler T. 2007. Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Molecular Biology and Evolution* **24**: 2310–2322.

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

Beaumont MA, Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**: 969–980.

Beisswanger S, Stephan W. 2008. Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated *polyhomeotic* genes in *Drosophila*. *Proceedings of the National Academy of Sciences, USA* **105**: 5447–5452.

Booth TH, Nix HA, Hutchinson MF, Busby JR. 1987. Grid Matching – a new method for homoclime analysis. *Agricultural and Forest Meteorology* **39**: 241–255.

Busby JR. 1986. A biogeoclimatic analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia. *Australian Journal of Ecology* **11**: 1–7.

Busby JR. 1991. Bioclim – a bioclimate analysis and prediction system. In: Margules CR, Austin MP, eds. *Nature conservation: cost effective biological surveys and data analysis*. Melbourne, Australia: CSIRO, 64–68.

Cakir B, Agasse A, Gaillard C, Saumonneau A, Delrot S, Atanassova R. 2003. A grape ASR protein involved in sugar and abscisic acid signaling. *Plant Cell* **15**: 2165–2180.

Canel C, Bailey-Serres JN, Roose ML. 1995. Pummelo fruit transcript homologous to ripening-induced genes. *Plant Physiology* **108**: 1323–1324.

Carrari F, Fernie AR, Iusem ND. 2004. Heard it through the grapevine? ABA and sugar cross-talk: the ASR story. *Trends in Plant Science* **9**: 57–59.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. *Molecular Biology and Evolution* **23**: 1348–1356.

Chetelat RT, Pertuze RA, Faundez L, Graham EB, Jones CM. 2009. Distribution, ecology and reproductive biology of wild tomatoes and related nightshades from the Atacama Desert region of northern Chile. *Euphytica* **167**: 77–93.

Depaulis F, Veuille M. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular Biology and Evolution* **15**: 1788–1790.

Eckert AJ, Wegrzyn JL, Pande B, Jermstad KD, Lee JM, Liechty JD, Tearse BR, Krutovsky KV, Neale DB. 2009. Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*). *Genetics* **183**: 289–298.

Ellegren H. 2009. A selection model of molecular evolution incorporating the effective population size. *Evolution* **63**: 301–305.

Eveno E, Collada C, Guevara MA, Leger V, Soto A, Diaz L, Leger P, Gonzalez-Martinez SC, Cervera MT, Plomion C *et al.* 2008. Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Molecular Biology and Evolution* **25**: 417–437.

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

Finkelstein RR, Gibson SI. 2001. ABA and sugar interactions regulating development: cross-talk or voices in a crowd? *Current Opinion in Plant Biology* **5**: 26–32.

Fisher RA. 1930. *The genetical theory of natural selection*. Oxford, UK: Clarendon Press.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist* **183**: 557–564.

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

Frankel N, Carrari F, Hasson E, Iusem ND. 2006. Evolutionary history of the *Asr* gene family. *Gene* **378**: 74–83.

Frankel N, Hasson E, Iusem ND, Rossi MS. 2003. Adaptive evolution of the water stress-induced gene *Asr2* in *Lycopersicon* species dwelling in arid habitats. *Molecular Biology and Evolution* **20**: 1955–1962.

Frankel N, Nunes-Nesi A, Balbo I, Mazuch J, Centeno D, Iusem ND, Fernie AR, Carrari F. 2007. *ci21A/Asr1* expression influences glucose accumulation in potato tubers. *Plant Molecular Biology* **63**: 719–730.

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

Giombini MI, Frankel N, Iusem ND, Hasson E. 2009. Nucleotide polymorphism in the drought responsive gene *Asr2* in wild populations of tomato. *Genetica* **136**: 13–25.

Glinka S, de Lorenzo D, Stephan W. 2006. Evidence of gene conversion associated with a selective sweep in *Drosophila melanogaster*. *Molecular Biology and Evolution* **23**: 1869–1878.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiology* **148**: 993–1003.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**: 1965–1978.

Hijmans RJ, Guarino L, Cruz M, Rojas E. 2001. Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. *Plant Genetic Resources Newsletter* **127**: 15–19.

Holliday JA, Yuen M, Ritland K, Aitken SN. 2010. Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. *Molecular Ecology* **19**: 3857–3864.

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

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

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London, Series B* **256**: 119–124.

Innan H. 2003. The coalescent and infinite-site model of a small multigene family. *Genetics* **163**: 803–810.

Innan H. 2009. Population genetic models of duplicated genes. *Genetica* **137**: 19–37.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* **11**: 97–108.

Iusem ND, Bartholomew DM, Hitz WD, Scolnik PA. 1993. Tomato (*Lycopersicon esculentum*) transcript induced by water deficit and ripening. *Plant Physiology* **102**: 1353–1354.

Jost L. 2008. G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* **17**: 4015–4026.

Kalifa Y, Gilad A, Konrad Z, Zaccai M, Scolnik PA, Bar-Zvi D. 2004. The water- and salt-stress-regulated *Asr1* (abscisic acid stress ripening) gene encodes a zinc-dependent DNA-binding protein. *Biochemical Journal* **381**: 373–378.

Kane NC, Rieseberg LH. 2007. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, *Helianthus annuus*. *Genetics* **175**: 1823–1834.

Kane NC, Rieseberg LH. 2008. Genetics and evolution of weedy *Helianthus annuus* populations: adaptation of an agricultural weed. *Molecular Ecology* **17**: 384–394.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome Biology* **3**: research0008. 1–0008.9.

Konrad Z, Bar-Zvi D. 2008. Synergism between the chaperone-like activity of the stress regulated ASR1 protein and the osmolyte glycine-betaine. *Planta* **227**: 1213–1219.

Leon P, Sheen J. 2003. Sugar and hormone connections. *Trends in Plant Science* **8**: 110–116.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.

Maere S, de Bodt S, Raes J, Casneuf T, van Montagu M, Kuiper M, van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences, USA* **102**: 5454–5459.

Maskin L, Frankel N, Gudesblat G, Demergasso MJ, Pietrasanta LI, Iusem ND. 2007. Dimerization and DNA-binding of ASR1, a small hydrophilic protein abundant in plant tissues suffering from water loss. *Biochemical and Biophysical Research Communications* **352**: 831–835.

Maskin L, Gudesblat GE, Moreno JE, Carrari FO, Frankel N, Sambade A, Rossi M, Iusem ND. 2001. Differential expression of the members of the *Asr* gene family in tomato (*Lycopersicon esculentum*). *Plant Science* **161**: 739–746.

Maskin L, Maldonado S, Iusem ND. 2008. Tomato leaf spatial expression of stress-induced *Asr* genes. *Molecular Biology Reports* **35**: 501–505.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genetical Research* **23**: 23–35.

McMichael AJ. 2001. Impact of climatic and other environmental changes on food production and population health in the coming decades. *Proceedings of the Nutrition Society* **60**: 195–201.

Moeller DA, Tiffin P. 2008. Geographic variation in adaptation at the molecular level: a case study of plant immunity genes. *Evolution* **62**: 3069–3081.

Mondragon-Palomino M, Gaut BS. 2005. Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. *Molecular Biology and Evolution* **22**: 2444–2456.

Moyle LC. 2008. Ecological and evolutionary genomics in the wild tomatoes (*Solanum* sect. Lycopersicon). *Evolution* **62**: 2995–3013.

Nakazato T, Warren DL, Moyle LC. 2010. Ecological and geographic modes of species divergence in wild tomatoes. *American Journal of Botany* **97**: 680–693.

Neiman M, Olson MS, Tiffin P. 2009. Selective histories of poplar protease inhibitors: elevated polymorphism, purifying selection, and positive selection driving divergence of recent duplicates. *New Phytologist* **183**: 740–750.

Ohno S. 1970. *Evolution by gene duplication*. New York, NY, USA: Springer.

Orr HA. 2005. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics* **6**: 119–127.

Padmanabhan V, Dias DMAL, Newton RJ. 1997. Expression analysis of a gene family in loblolly pine (*Pinus taeda* L.) induced by water deficit stress. *Plant Molecular Biology* **35**: 801–807.

Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and *Tetraodon*. *Trends in Genetics* **22**: 597–602.

Pavlidis P, Laurent S, Stephan W. 2010. msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Molecular Ecology Resources* **10**: 723–727.

Pearce J, Ferrier S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling* **133**: 225–245.

Peralta IE, Spooner DM. 2001. Granule-bound starch synthase (GBSSI) gene phylogeny of wild tomatoes (*Solanum* L. section *Lycopersicon* [Mill.] Wettst. subsection *Lycopersicon*). *American Journal of Botany* **88**: 1888–1902.

Peralta IE, Spooner DM, Knapp S. 2008. Taxonomy of wild tomatoes and their relatives (*Solanum* sect. Lycopersicoides, sect. Juglandifolia, sect. *Lycopersicon*; Solanaceae). *Systematic Botany Monographs* **84**: 1–186 + 3 plates.

Philippe R, Courtois B, McNally KL, Mournet P, El-Malki R, Le Paslier MC, Fabre D, Billot C, Brunel D, Glaszmann J *et al.* 2010. Structure, allelic diversity and selection of *Asr* genes, candidate for drought tolerance, in *Oryza sativa* L. and wild relatives. *Theoretical and Applied Genetics* **121**: 769–787.

R Development Core Team. 2005. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Riebler A, Held L, Stephan W. 2008. Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics* **178**: 1817–1829.

Rizzon C, Ponger L, Gaut BS. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Computational Biology* **2**: 989–1000.

Rom S, Gilad A, Kalifa Y, Konrad Z, Karpasas MM, Goldgur Y, Bar-Zvi D. 2006. Mapping the DNA- and zinc-binding domains of ASR1 (abscisic acid stress ripening), an abiotic-stress regulated plant specific protein. *Biochimie* **88**: 621–628.

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

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

Rossi M, Iusem ND. 1994. Tomato (*Lycopersicon esculentum*) genomic clone homologous to a gene encoding an abscisic acid induced protein. *Plant Physiology* **104**: 1073–1074.

Saumonneau A, Agasse A, Bidoyen M, Lallemand M, Cantereau A, Medici A, Laloi M, Atanassova R. 2008. Interaction of grape ASR proteins with a DREB transcription factor in the nucleus. *FEBS Letters* **582**: 3281–3287.

Seki M, Ishida J, Narusaka M, Fujita M, Nanjo T, Umezawa T, Kamiya A, Nakajima M, Enju A, Sakurai T *et al.* 2002. Monitoring the expression pattern of around 7,000 *Arabidopsis* genes under ABA treatments using a full-length cDNA microarray. *Functional & Integrative Genomics* **2**: 282–291.

Shan H, Zahn L, Guindon S, Wall PK, Kong H, Ma H, dePamphilis CW, Leebens-Mack J. 2009. Evolution of plant MADS box

transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. *Molecular Biology and Evolution* **26**: 2229–2244.

Shen G, Pang YZ, Wu WS, Deng ZX, Liu XF, Lin J, Zhao LX, Sun XF, Tang KX. 2005. Molecular cloning, characterization and expression of a novel *Asr* gene from *Ginkgo biloba*. *Plant Physiology and Biochemistry* **43**: 836–843.

Siol M, Wright SI, Barrett SCH. 2010. The population genomics of plant adaptation. *New Phytologist* **188**: 313–332.

Spooner DM, Anderson GJ, Jansen RK. 1993. Chloroplast DNA evidence for the interrelationships of tomatoes, potatoes, and pepinos (Solanaceae). *American Journal of Botany* **80**: 676–688.

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

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

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

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

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

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

Takuno S, Nishio T, Satta Y, Innan H. 2008. Preservation of a pseudogene by gene conversion and diversifying selection. *Genetics* **180**: 517–531.

Thornton KR. 2007. The neutral coalescent process for recent gene duplications and copy-number variants. *Genetics* **177**: 987–1000.

Thornton KR, Jensen JD, Becquet C, Adolfatto P. 2007. Progress and prospects in mapping recent selection in the genome. *Heredity* **98**: 340–348.

Treger JM, McEntee K. 1990. Structure of the DNA damage-inducible gene DDR48 and evidence for its role in mutagenesis in *Saccaromyces cerevisiae*. *Molecular and Cellular Biology* **10**: 3174–3184.

Turner TL, Bourne EC, Wettberg EJ von, Hu TT, Nuzhdin SV. 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics* **42**: 260–263.

Vaidyanathan R, Kuruvilla S, Thomas G. 1999. Characterization and expression pattern of an abscisic acid and osmotic stress responsive gene from rice. *Plant Science* **140**: 21–30.

Wachowiak W, Balk PA, Savolainen O. 2009. Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (*Pinus sylvestris* L.). *Tree Genetics & Genomes* **5**: 117–132.

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

Watterson GA. 1975. Number of segregating sites in genetic models without recombination. *Theoretical Population Biology* **7**: 256–276.

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

Zhen Y, Ungerer MC. 2008. Relaxed selection on the CBF/DREB1 regulatory genes and reduced freezing tolerance in the southern range of *Arabidopsis thaliana*. *Molecular Biology and Evolution* **25**: 2547–2555.

Zou C, Lehti-Shiu MD, Thomashow M, Shiu S. 2009. Evolution of stress regulated gene expression in duplicate genes of *Arabidopsis thaliana*. *PLoS Genetics* **5**: e1000581.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Positions of the *Asr* (ABA/water stress/ripening induced) genes in the gene cluster relative to the BAC sequence of *S. lycopersicum*.

Fig. S2 Potential distribution of *Solanum peruvianum* and *Solanum chilense* estimated from collecting sites over the east coast of South America.

Table S1 Primer sequences and amplification conditions for PCR of the *Asr* (ABA/water stress/ripening induced) genes

Table S2 Numbers of alleles sequenced for each locus

Table S3 Numbers of site categories

Table S4 Nucleotide diversity of the *Asr* (ABA/water stress/ripening induced) genes and the reference loci

Methods S1 Bioclimatic data and ecological niche modeling.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

The evolution of stress-responsive gene regulation in natural populations of wild tomato

Iris Fischer[1], Kim A. Steige[1,2], Wolfgang Stephan[1] and Mamadou Mboup[1]

[1] Section of Evolutionary Biology, Department of Biology II, University of Munich (LMU), Grosshaderner Strasse 2, 82152 Planegg-Martinsried, Germany

[2] Present address: Department of Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Norbyvägen 18 D, 75236 Uppsala, Sweden

Author for correspondence:

Iris Fischer

Phone: +49-89-218074161

Fax: +49-89-218074104

Email: iris.fischer@bio.lmu.de

SUMMARY

Investigating the expression profile of candidate genes is of great importance when studying local adaptation as modulation of gene expression is crucial for an organism's survival during stress conditions. Wild tomato species are a valuable system to study adaptation since they grow in diverse environments facing many abiotic constraints. The expression profile and DNA sequence polymorphism at regulatory regions of genes reported to be involved in water and cold stress response were investigated in populations of two closely related species, *Solanum chilense* and *S. peruvianum.* We show that the gene expression profile of the *Asr* gene (ABA /water stress/ripening induced) *Asr4* is adaptive to drought conditions. Sequence data reveals a conserved promoter region of the *Asr2* gene and the acting of positive selection at the downstream region of the dehydrin *pLC30-15*. This study provides an example for expression variation in natural populations. We observe plasticity in gene expression and find that in this case variability in expression is advantageous, which is common in stress responsive genes. Analysis of regulatory regions provides directions for future functional studies of those genes.

Key words: LEA (late embryogenesis abundant) proteins, *Asr* (ABA /water stress/ripening induced), dehydrin *pLC30-15*, local adaptation, gene expression, wild tomatoes, regulatory elements

INTRODUCTION

Much effort has been made over the last decades to understand the phenomenon of local adaptation, defined as the movement of a population towards a phenotype that leads to the highest fitness in a particular environment (Fisher, 1930). Many approaches using DNA sequence variation to detect adaptation have been developed and applied to plant model systems (Siol *et al.*, 2010). More than 30 years ago, it has already been discovered that protein divergence alone is not sufficient to explain observed phenotypic differences between species. It was therefore proposed that changes in gene regulation could account for many cases of adaptation (Wilson *et al.*, 1974; King & Wilson, 1975). Modulation of gene expression is crucial for an organism's survival as environmental changes require a fast and specific response. Experimental evolution in microorganisms revealed fast expression divergence between strains of *Saccharomyces cerevisiae* (Ferea *et al.*, 1999) and *Escherichia coli* (Cooper *et al.*, 2003) grown in glucose-limited media after 250 and 200,000 generations, respectively. In plants, regulatory changes between domesticated crop species and their wild relatives and their role in adaptation has been described in maize (Doebley *et al.*, 1997; Wang *et al.*, 2005) and rice (Konishi *et al.*, 2006). Therefore, another way to investigate local adaptation is to study the expression profile of genes previously described to be advantageous for the plant's fitness in stress conditions. Microarrays allow analyzing the transcriptome of species, but so far studies on gene expression in natural populations are limited to few species like *Boechera holboellii*, a close relative of *Arabidopsis thaliana* (Knight *et al.*, 2006), the arthropod *Orchesella cincta* (Roelofs *et al.*, 2009), the snail species *Littorina saxatilis* (Martínez-Fernández *et al.*, 2010), fishes (Larsen *et al.*, 2011), and *Drosophila melanogaster* (Hutter *et al.*, 2008; Müller *et al.*, 2011). The question to which degree regulatory divergence are adaptive also remains (Fay & Wittkopp, 2008). Most analyses today are still focused on the evolution of coding sequences but examples for regulatory changes grew over the last few years (Wray, 2007).

Plants are sessile during most of their life cycle and therefore experience strong selective pressure to adapt to changing environmental conditions in their habitat (*e.g.* precipitation, temperature). Drought and cold stress are the major abiotic constraints that terrestrial plants are facing and have been shown to have adverse effects on the plant growth and crop production (Yáñez *et al.*, 2009). Both, drought and cold tolerance are complex traits but it has been shown that similar genes are expressed during both types of stress (Shinozaki & Yamaguchi-Shinozaki, 2000). A microarray experiment showed that the expression of more

than 250 genes was induced in *A. thaliana* after a drought-stress treatment (Seki *et al.*, 2002). In a similar experiment in *A. thaliana*, 4% of all transcripts showed responsiveness to low temperature (Fowler & Thomashow, 2002). Drought and cold lead to accumulation of the phytohormone abscisic acid (ABA), and it has been demonstrated that application of ABA mimics stress conditions (Mahajan & Tuteja, 2005). ABA plays an important role in the plant's response to osmotic stress: It fine-tunes stomatal closure (Jones & Mansfield, 1970), it enhances expression of stress-related genes (Bray, 2004), and fosters root growth in long-term drought conditions (Saab *et al.*, 1990). It was suggested that cold and drought stress signals and ABA share common elements and are cross talking in their signalling pathways (Shinozaki & Yamaguchi-Shinozaki, 2000). Therefore, studying genes that are involved in the ABA pathway are good candidates for investigating adaptation.

Late embryogenesis abundant (LEA) proteins are induced by ABA and were shown to accumulate in vegetative organs during dehydration and low temperature stress (Ingram & Bartels, 1996; Bray, 1997). This suggests a protective role during water-limiting and chilling conditions. Members of the LEA protein family can be found across the entire plant kingdom: in angiosperms, gymnosperms (Shinozaki & Yamaguchi-Shinozaki, 1996; Bray, 1997), bryophytes (Proctor *et al.*, 2007), and algae (Tanaka *et al.*, 2004). They belong to the group of the hydrophilins that are characterized by a high glycin content and high hydrophilicity (Garay-Arroyo *et al.*, 2000). The LEA proteins are subdivided into seven groups based on their amino acid sequences (Battaglia *et al.*, 2008). In this study we analyze two types of LEA proteins: dehydrins, which belong to Group 2, and ASRs, which belong to Group 7 (Battaglia *et al.*, 2008). Some dehydrins have been shown to have cryoprotective functions, while others have been found to prevent inactivation of enzymes during dehydration (Reyes *et al.*, 2005), but their functional role still remains speculative. In *S. tuberosum* and *S. sogarandinum* an increased level of dehydrins could be correlated with cold tolerance in tubers and stems (Rorat *et al.*, 2006). Additionally, dehydrins were induced after drought stress in apical parts (Rorat *et al.*, 2006). The drought- and ABA-inducible dehydrin used in this work was described in *S. chilense* and denoted *pLC30-15* (Chen *et al.*, 1993). A previous study of the *pLC30-15* gene revealed that diversifying selection acted on its coding region in a wild tomato population from a dry environment (Xia *et al.*, 2010).

The other genes used in this study are the members of the *Asr* (ABA/water stress/ripening induced) gene family. As the name suggests, *Asr* genes have been shown to be induced by application of ABA, abiotic stress (drought, cold, salinity), and during ripening (Iusem *et al.*, 1993; Rossi & Iusem, 1994; Amitai-Zeigerson *et al.*, 1995; Schneider *et al.*, 1997;

Vaidyanathan *et al.*, 1999). *Asr* genes have several functions that help the plant deal with stress: as monomers with a chaperon function in the cytoplasm (Konrad & Bar-Zvi, 2008) or as homo- and heterodimers with DREB (drought response element binding) proteins (Maskin *et al.*, 2007; Saumonneau *et al.*, 2008) with DNA-binding activity in the nucleus (Kalifa *et al.*, 2004a). They also serve as transcription factors associated with the modulation of sugar transport activity (Carrari *et al.*, 2004; Frankel *et al.*, 2007; Maskin *et al.*, 2008). Expression analysis in rice (*Oryza sativa*), pine (*Pinus taeda*), *Ginkgo biloba*, and potato (*S. tuberosum*) suggest that relative *Asr* expression differs between copies, organs, and depending on the stress applied (Padmanabhan *et al.*, 1997; Schneider *et al.*, 1997; Shen *et al.*, 2005; Philippe *et al.*, 2010). Using semi-quantitative RT-PCR, it was shown that *Asr1* and *Asr2* are induced in leaves whereas *Asr2* is induced and *Asr3* is down-regulated in roots of cultivated tomato (*S. lycopersicum*; Maskin *et al.*, 2001). Analyzing different accessions of wild tomato using Northern Blots, it was demonstrated that *Asr1* and *Asr4* are up-regulated in leaves of plants from humid environments (Frankel *et al.*, 2006). Also in wild tomato, several studies revealed patterns consistent with local adaptation at *Asr* genes in populations that dwell in dry environments (Frankel *et al.*, 2003; Giombini *et al.*, 2009; Fischer *et al.*, 2011). These findings make *pLC30-15* and *Asr* genes interesting candidates for studying local adaptation on the gene expression level.

The availability of cultivated tomato genomic resources, the recent divergence of the *Solanum* species, and their clear phenotypic distinction (Peralta *et al.*, 2008) make tomato species a popular plant model system that is frequently used to study evolution. Most *Solanum* sect. *Lycopersicon* species are native to western South America (Ecuador, Peru, and Chile), along the western and eastern Andean slopes (Spooner *et al.*, 2005). This study focuses on two recently diverged wild tomato species that show differences in their ecological habitats and features: *S. chilense* and *S. peruvianum*. *Solanum chilense* is distributed from southern Peru to northern Chile and inhabits arid plains and deserts (Peralta *et al.*, 2008). It also shows a broad range in elevation from sea level up to 3,500 m and therefore experiences large temperature gradients during the year (Chetelat *et al.*, 2009). The species is also known to be robust and drought tolerant and can dwell in hyperarid areas (Moyle, 2008; Peralta *et al.*, 2008). In fact, the potential distribution of *S. chilense* was predicted to be mostly determined by the annual precipitation (Nakazato *et al.*, 2010). *Solanum peruvianum* is distributed from central Peru to northern Chile and inhabits a variety of habitats, from coastal deserts to river valleys (Peralta *et al.*, 2008).

Here, we analyze the relative expression of *Asr1*, *Asr2*, *Asr4*, and *pLC30-15* in drought and cold stressed *S. chilense* and *S. peruvianum* accessions from contrasting environments. We were able to determine differences in the gene expression profiles, *i.e.* intensity, speed, or differences depending on the type of stress or the gene copy. Population genetic analysis has provided evidence for local adaptation at *Asr2*, *Asr4*, and *pLC30-15* (Giombini *et al.*, 2009; Xia *et al.*, 2010; Fischer *et al.*, 2011). We therefore sequenced the regulatory regions of these genes from the same populations for which we performed the expression analysis in order to investigate the evolutionary forces shaping them. As the genes were already sequenced previously in the same populations (Xia *et al.*, 2010; Fischer *et al.*, 2011), we can directly compare them to their regulatory regions. In addition, we could identify conserved *cis*-acting elements.

MATERIALS AND METHODS

Gene expression analysis

Plant material, cultivation and replication Seeds of three accessions of each *S. chilense* and *S. peruvianum* were obtained from the Tomato Genetics Resource Center (TGRC) at UC Davies (Table 4.1). As these wild tomato species are outcrossing, they had to be multiplied by cuttings to gain genetically identical replicates. Tomato seeds of the motherplants were treated with 2.7% NaOCl for 20 minutes to foster germination. The tomato seeds were kept on moistened filter paper at room temperature in the dark until they germinated. The tomato seedlings were then transferred to soil and put in the climate chamber at 22°C with a 16h/8h day/night cycle and 70% humidity. The motherplants were grown for approximately three months until they could provide material for 20-25 cuttings. We thus gained three biological replicates for every timepoint. The cuttings were treated with the Neudofix rooting enhancer (Neudorff, Emmerthal, Germany) and transferred to pots containing Vermiculite on top (to ensure ventilation) and soil on the bottom part (to ensure nutrition once the plant grew roots). The cuttings were grown for five weeks under the same conditions as the motherplants until they grew roots and three fresh leaves.

Stress treatment, RNA extraction and cDNA synthesis Drought stress was applied by removing the tomato plants from their pots, carefully rinsing and drying their roots and transferring them into a climate chamber at 22°C. For cold stress, the plants were kept at 4°C. Leaves and roots were quick-freezed in liquid nitrogen at five timepoints: unstressed plants, one hour, three hours, six hours, and 24 hours after stress application. Total RNA was isolated

using the RNeasy Plant Mini Kit (Qiagen GmbH, Hilden, Germany). DNA was removed using an on-column DNaseI digestion protocol. The RNA integrity was assessed by gel electrophoresis. A NanoDrop 1000 Spectrophotometer (Peqlab, Erlangen, Germany) was used to quantify and to asses the quality of the RNA. Only samples with A_{260}/A_{280} and A_{260}/A_{230} values between 1.9 and 2.1 were used for downstream experiments. cDNA was synthesised from 1 µg total RNA with the SuperScriptIII reverse transcriptase and the RNase inhibitor RNaseOUT (both Invitrogen, Carlsbad, CA, USA) using oligo(dT_{20}) primers. The cDNA was treated with RNaseH (New England Biolabs, Ipswich, MA, USA) to remove remaining RNA.

Table 4.1. Location and habitat characteristics of the accessions used for the expression analysis

Accession[a]	Species[c]	Nearby population[b]	Collection site[c]	Coordinates[c]	Altitude[c]	Habitat[c]	Stresses tested
LA1938	*S. chilense*	Quicacha	Quebrada Salsipuedes, Arequipa, Peru	15°41' S / 73°50' W	1400 m	In a large quebrada	drought + cold
LA1967	*S. chilense*	Tacna	Pachia, Tacna, Peru	17°55' S / 70°09' W	1000 m	Extremely dry habitat	drought + cold
LA1969	*S. chilense*	Tacna	Estique Pampa, Tacna, Peru	17°32' S / 70°02' W	3250 m	Near maximum elevation for *S. chilense*	cold
LA2744	*S. peruvianum*	Tarapaca	Sobraya, Tarapaca, Chile	18°33' S / 70°09' W	400 m	Along the margins of cultivated fields and waste places	drought + cold
LA2745	*S. peruvianum*	Tarapaca	Pan de Azucar, Tarapaca, Chile	18°35' S / 69°56' W	600 m	Scattered in cultivated fields and waste places	drought + cold
LA3636	*S. peruvianum*	Canta	Coayllo, Lima, Peru	12°41' S / 76°24' W	NA	NA	drought + cold

[a] Tomato Genetics Resource Center (TGRC) accession number

[b] Nearby populations sampled by T. Städler and T. Marczewski, 2004

[c] According to TGRC database

NA No data available

Quantitative real-time PCR Primers for the quantitative real-time PCR (hence referred to as qPCR) were designed using NetPrimer (http://www.primierbiosoft.com/netprimer) and PrimerBLAST (http://www.ncbi.nlm.nih.gov). All primers used in this project can be found in Supporting Information Table C.S1. As *Asr3* and *Asr5* cannot be distinguished at their coding region (Fischer *et al.*, 2011), they were excluded from this study. qPCR was carried out using

iQ SYBR green on a CFX thermocycler (both BioRad, Hercules, CA, USA). Three technical replicates were analyzed for each sample and the efficiency of the qPCR runs was determined by a 1:10 dilution series. Expression of the target genes was normalized by two constantly expressed reference genes: *CT189*, a 40S ribosomal protein (Roselius *et al.*, 2005) and *TIP4I*, which was shown to be a very stable housekeeping gene in tomato (Expósito-Rodríguez *et al.*, 2008). As the efficiency was around 100% for all runs we applied the $2^{-\Delta\Delta Cq}$ Method (Livak & Schmittgen, 2001) to derive the relative expression quantity from the measured C_q-values. Quality control, reference gene stability, transformation to relative quantity, and normalization was carried out using the program qbase^{PLUS} (Hellemans *et al.*, 2007). We used the two sample Wilcoxon (Mann-Whitney-*U*) test to determine significant differences in relative expression between stressed and control plants using R (R Development Core Team 2005).

Sequence analysis of the *Asr* and *pLC30-15* regulatory regions

Plant material and sequencing We sequenced the promoter region of *Asr2* (*pAsr2*), *Asr4* (*pAsr4*), *pLC30-15* (*5'pLC*), and the 3'UTR of *pLC30-15* (*3'pLC*). Two populations from climatically different environments were sampled for each species (Tacna and Quicacha for *S. chilense*; Tarapaca and Canta for *S. peruvianum*). Five to seven individuals of each population were analyzed (Table C.S2). A detailed description of these samples is provided in Baudry *et al.* (2001), Roselius *et al.* (2005), Städler *et al.* (2005), Arunyawat *et al.* (2007), and Fischer *et al.* (2011). The Tarapaca sample was obtained from the TGRC (accession number LA2744). The other populations were sampled by T. Städler and T. Marczewski in May 2004 (Arunyawat *et al.*, 2007; Städler *et al.*, 2008). *Solanum ochranthum* (TGRC accession LA2682) was used as outgroup. DNA extraction, PCR amplification, cloning, and sequencing was performed as described in Fischer *et al.* (2011). In addition, the *Asr* genes and *pLC30-15* of the motherplants from the expression experiment were sequenced to determine their haplotypes. Sequence data from this article have been deposited in the EMBL/GenBank Data Libraries under accession nos. HE612885-HE613033.

Nucleotide diversity analysis, neutrality tests and haplotype diversity We measured nucleotide diversity by Watterson's θ_w and Tajima's π using the DnaSP v5 software (Librado & Rozas, 2009). θ_w is based on the number of segregating sites (Watterson, 1975), and π on the average number of pairwise nucleotide differences among sequences in a sample (Tajima, 1983). To detect regions of low diversity sliding window analysis of π across the locus was

performed in DnaSP with a window length of 10 bp and a step size of 1 bp. We tested for deviations from the standard neutral model using Tajima's *D* statistic and Fu & Li's *D* in DnaSP. A significantly negative value of Tajima's *D* indicates an excess of rare variants as expected under directional selection or population size expansion (Tajima, 1989). A significantly positive Tajima's *D* value indicates an excess of intermediate-frequency variants as expected under balancing selection or in structured populations (Tajima, 1989). The same is true for Fu and Li's *D* statistics (Fu & Li, 1993). Fay & Wu's *H* was also calculated, where a significantly negative Fay and Wu's *H* indicates an over-representation of high-frequency derived polymorphisms, which suggest that positive selection was acting (Fay & Wu, 2000). A significantly positive Fay and Wu's *H* indicates an over-representation of intermediate-frequency derived polymorphisms, which is the case if balancing selection was acting (Fay & Wu, 2000). The haplotype test of Depaulis & Veuille (1998) was used to analyze whether haplotype diversity (H_d) deviates from the prediction of the standard neutral model. All neutrality tests were performed using the option Number of Segregating Sites in DnaSP.

Motif search on non-coding regions Motifs in the promoter region were searched using the program PlantCARE (Lescot *et al.*, 2002). This program is a database of *cis*-acting regulatory elements and allows *in silico* analysis of promoter sequences.

RESULTS

Faster induction of *Asr4* transcription in a population from a dry environment

We analyzed the expression patterns of *Asr1*, *Asr2*, *Asr4*, and *pLC30-15* from different populations in a time-course experiment after exposing wild tomato plants to cold and drought stress. In general, all genes show a higher induction in *S. chilense* than in *S. peruvianum* and respond more strongly to water deficit than to cold stress (Figs. 4.1-4.4). For the *Asr* genes, the highest transcript level can be observed in the accession from Quicacha, after cold as well as after drought stress. This is especially true for *Asr4* after water deficit stress. The transcript level is significantly higher after 3-24h after stress application compared to the unstressed plant (Fig. 4.1a). No transcription of *Asr4* can be detected after 1h. Interestingly, *Asr4* is already significantly over-expressed in the accession from Tacna after 1h and has its expression peak after 3h (Fig. 4.1a). Previously, evidence for local adaptation in this population was found (Fischer *et al.*, 2011); and indeed, the Tacna accession is homozygous for the predominant haplotype described by Fischer *et al.* (2011). In *S. peruvianum*, the two Tarapaca accessions (both from a very dry environment) differ in their expression patterns.

LA2745 is significantly over-expressed after 1h and shows the peak of *Asr4* expression after 6h (Fig. 4.1b). LA2744, on the other hand, shows a constant increase of *Asr4* transcripts until timepoint 24h (Fig. 4.1b). The Canta accession (from a mesic environment) displays a significant over-expression of *Asr4* after 6h and 24h (Fig. 4.1b).

Asr4 induction after chilling stress is much less intense. All *S. chilense* accessions show a fast induction (1h) and a decrease of *Asr4* transcripts after 6h in the Tacna accessions (Fig. 4.1c). However, the transcript level is higher in the Tacna accession from a high altitude (LA1969) compared to the one from a lower altitude (LA1967). In *S. peruvianum*, *Asr4* induction is much slower: after 3h in Canta and 6h in Tarapaca (Fig. 4.1d).

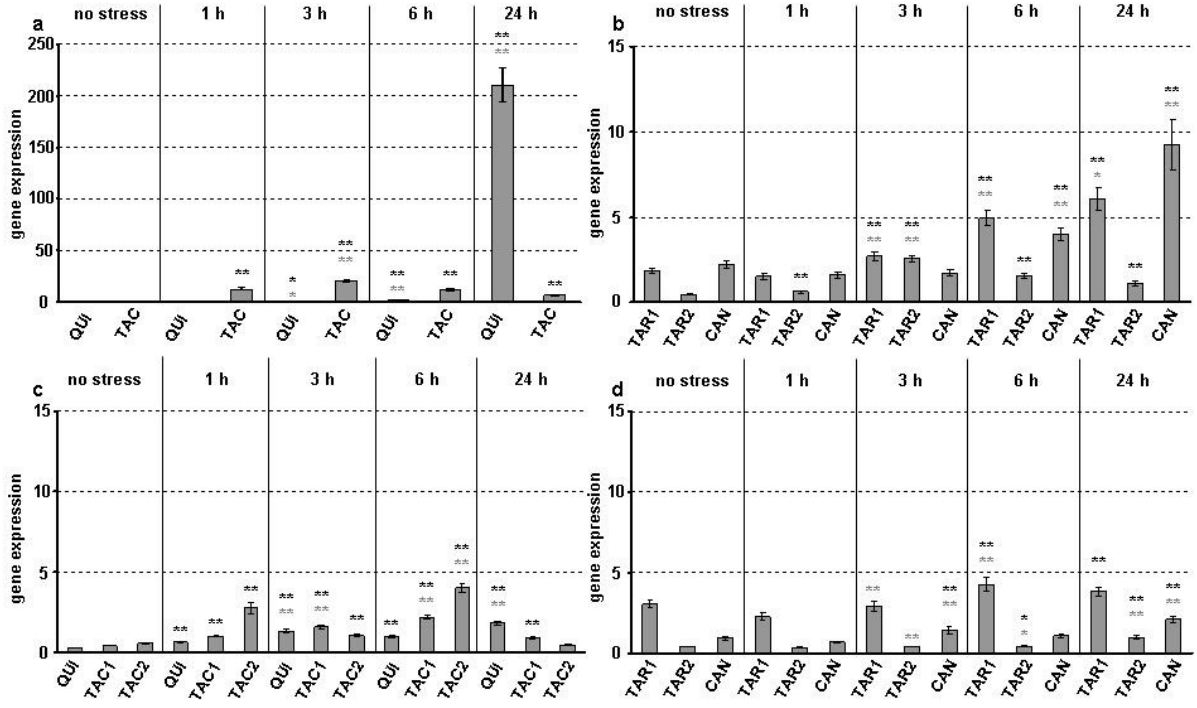

Figure 4.1 Gene expression (displayed relative to the average) of *Asr4* after application of drought and cold stress in *S. chilense* and *S. peruvianum* in unstressed plants, 1 h, 3 h, 6 h, and 24 h after stress application. (a) The following accessions of *S. chilense* were measured after drought stress: LA1938 (QUI) from a dry environment and LA1967 (TAC) from a hyperarid area. (c) The following accessions of *S. chilense* were measured after cold stress: LA1938 (QUI) from a dry environment and LA1967 (TAC1) from a hyperarid area, and LA1969 (TAC2) from a very dry environment and high altitude. The following accessions of *S. peruvianum* were measured after (b) drought and (d) cold stress: LA2744 (TAR1) and LA2745 (TAR2) from a dry environment and LA3636 (CAN) from a humid environment. Vertical lines at bar charts indicate the standard error, black asterisks above bar charts indicate significant over-expression compared to the unstressed control, grey asterisks above bar charts indicate significant over-expression compared to the previous timepoint (* P<0.05; ** P<0.01).

Expression patterns of *Asr1* and *Asr2*

Asr1 is also induced faster in the Tacna accession (1h) compared to the Quicacha accession (3h) after drought stress (Fig. 4.2a). Transcript levels are highest in Quicacha after 24h (Fig. 4.2a). In *S. peruvianum*, *Asr1* is significantly over-expressed after 3h in the Tarapaca accessions, but only after 6h in the Canta accession (Fig. 4.2b). In all accessions, the *Asr1* transcript level keeps increasing until 24h. After cold stress, *Asr1* is induced the fastest and at the highest level in Quicacha (Fig. 4.2c). In the Tacna accession from a cold environment (LA1969), *Asr1* is induced faster than in the other Tacna accession but interestingly at a lower level (Fig. 4.2c). In *S. peruvianum*, *Asr1* is significantly over-expressed after 1-6h but at a much lower level than in Quicacha and at the same level as the other *S. chilense* accessions (Fig. 4.2d).

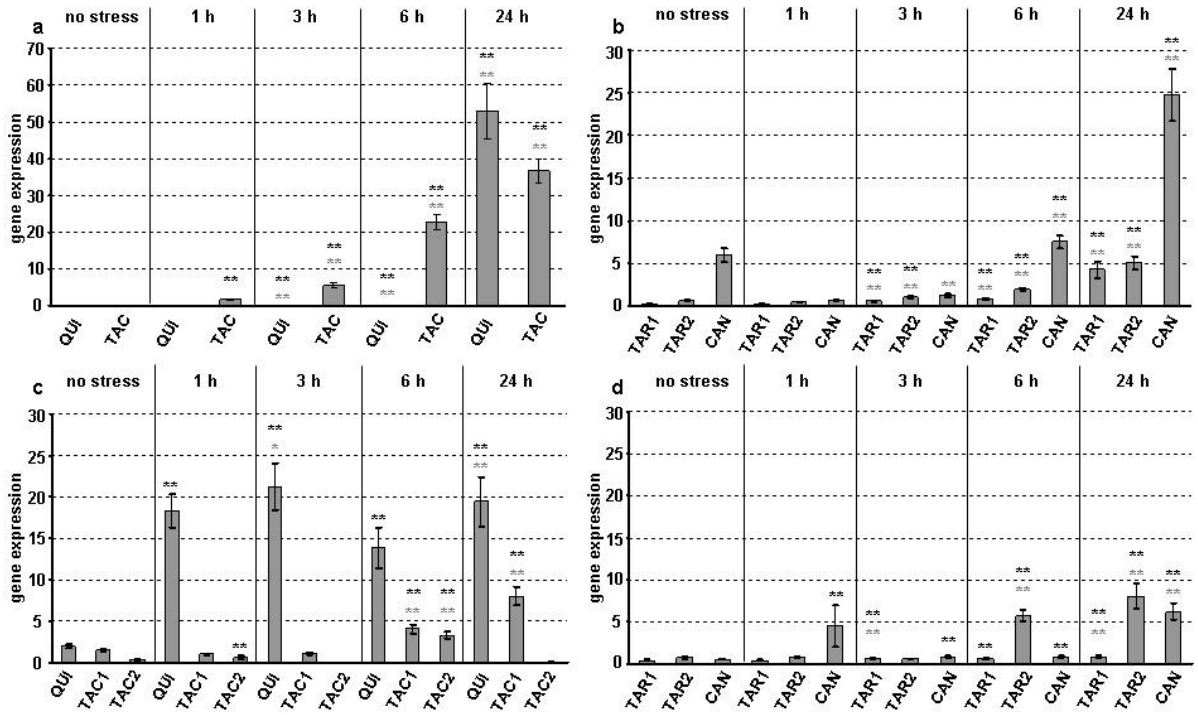

Figure 4.2 Gene expression (displayed relative to the average) of *Asr1* after application of drought and cold stress in *S. chilense* and *S. peruvianum* in unstressed plants, 1 h, 3 h, 6 h, and 24 h after stress application. (For explanation see Fig. 4.1).

Asr2 also shows the highest transcript levels in Quicacha for *S. chilense*, followed by LA2744 for *S. peruvianum* after water deficit stress (Fig. 4.3a+b). After application of drought stress, *Asr2* is significantly over-expressed after 1h in Tacna and 6h in Quicacha (Fig. 4.3a). In *S. peruvianum*, *Asr2* is induced after 1h in one Tarapaca accession (LA2744) and

only after 24h in the other Tarapaca accession (LA2745) and Canta (Fig. 4.3b). In all accessions, transcription levels keep increasing until 24h following drought stress (Fig. 4.3a+b). After application of cold stress, however, LA2745 shows the highest and fastest (1h) induction of *Asr2* in all *S. peruvianum* accessions (Fig. 4.3d). LA2744 and LA3636 are induced more slowly (3h and 24h, respectively) and at a much lower level (Fig. 4.3d). In *S. chilense*, the transcript level is again highest in Quicacha after cold stress. *Asr2* is significantly over-expressed after 1h (LA1938, LA1969) or 3h (LA1967; Fig. 4.3c). In general, induction of *Asr2* is again much lower after cold stress (Fig. 4.3c+d).

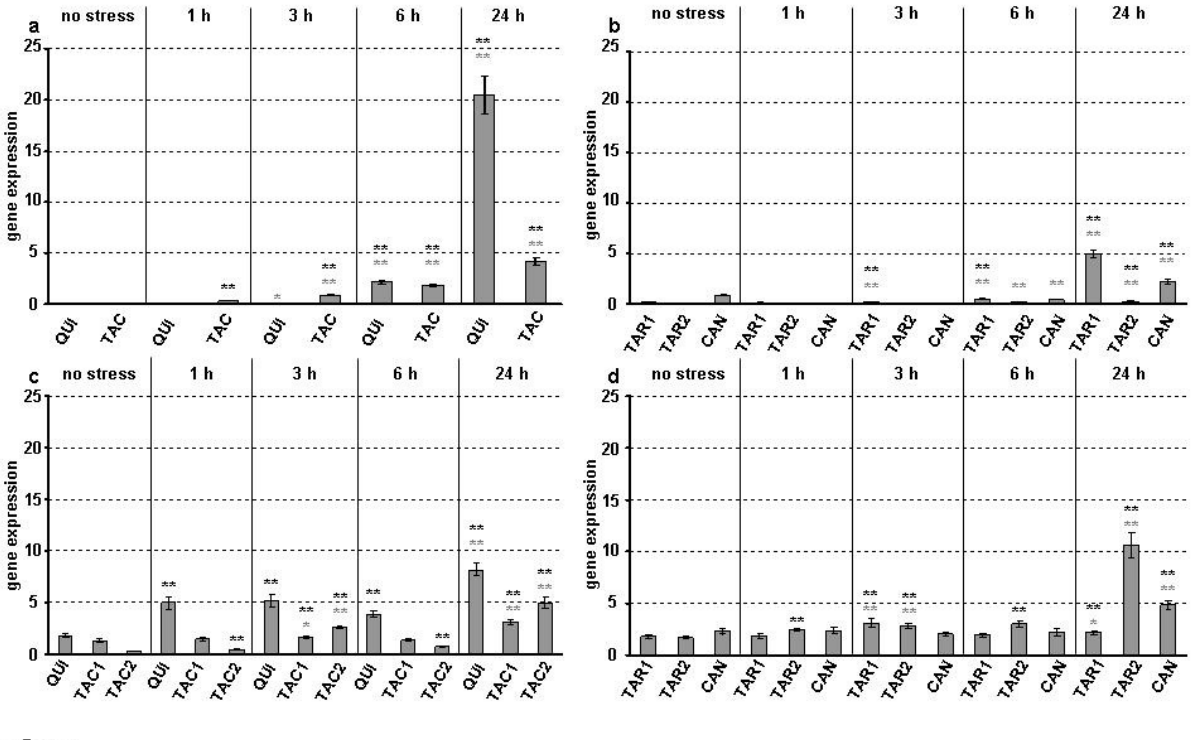

Figure 4.3 Gene expression (displayed relative to the average) of *Asr2* after application of drought and cold stress in *S. chilense* and *S. peruvianum* in unstressed plants, 1 h, 3 h, 6 h, and 24 h after stress application. (For explanation see Fig. 4.1).

High expression levels of *pLC30-15* after drought stress

After application of drought stress, *pLC30-15* is induced fast (after 1h in all accessions) and reaches its peak of transcription after 3-6h – except for Quicacha and Canta where the expression level increases until 24h (Fig. 4.4a+b). Transcript levels are higher in *S. chilense* than in *S. peruvianum* after drought stress, but to a lesser degree than for the *Asr* genes (Fig. 4.4a+b). After cold stress, *pLC30-15* shows a lower transcript level in both species (Fig. 4.4c+d). *pLC30-15* is significantly over-expressed after 1h in all *S. chilense* accessions and

expression keeps increasing until 24h (Fig. 4.4c). In *S. peruvianum*, *pLC30-15* is induced after 1h in LA2744 and 3h in LA3636 and keeps increasing until 24h (Fig. 4.4d). No induction of *pLC30-15* can be detected in accession LA2745 from Tarapaca (Fig 4.4d).

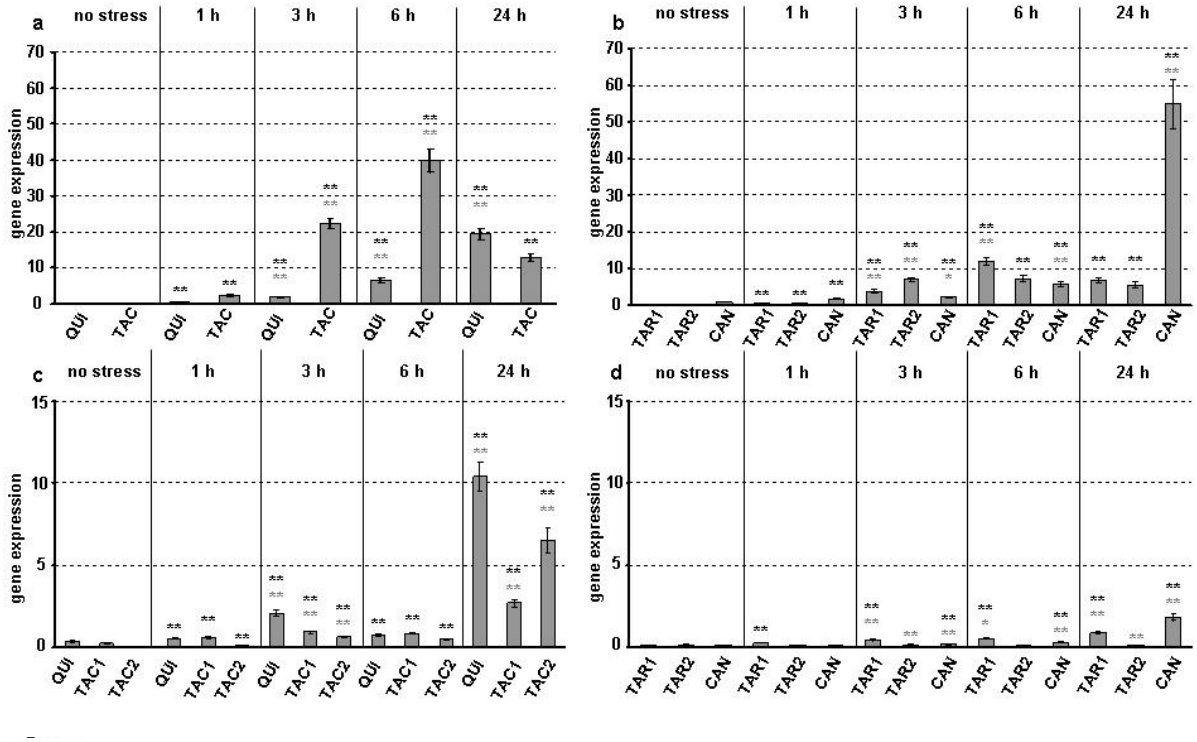

Figure 4.4 Gene expression (displayed relative to the average) of *pLC30-15* after application of drought and cold stress in *S. chilense* and *S. peruvianum* in unstressed plants, 1 h, 3 h, 6 h, and 24 h after stress application. (For explanation see Fig. 4.1).

Conserved *Asr2* promoter region

We sequenced the promoter regions of *Asr2*, *Asr4*, *pLC30-15* (~ 1,500 bp) and the 3'UTR of *pLC30-15* (~ 2,300 bp) to investigate the evolutionary forces acting on regulatory regions of genes involved in stress response. *pAsr2* is highly conserved, especially in Quichacha (Table 4.2). Interestingly, *pAsr2* is even more conserved than the *Asr2* gene. Haplotype diversity is extremely low in the Quicacha population (Table 4.3) and Tajima's *D* and Fu and Li's *D* are significantly negative (Table 4.4), indicating directional selection acting at *pAsr2*. Indeed, we find only two *pAsr2* very similar haplotypes in our dataset, where only one allele represents the second haplotype. This is an extreme case for wild tomato, even given the fact that, according to previous studies, the Quicacha population seems to display a haplotype structure here (Arunyawat *et al.*, 2007; Städler *et al.*, 2008, Fischer *et al.*, 2011). Looking at the

alignment of both species we can identify many regions which are conserved between *S. chilense* and *S. peruvianum* at the *Asr2* promoter region.

Table 4.2 Nucleotide diversity of *pAsr2*, *pAsr4*, *5'pLC*, *3'pLC*, and their corresponding genes

π	*pAsr2*	*Asr2*[b]	*pAsr4*	*Asr4*[b]	*5' pLC*	*pLC*[c]	*3' pLC*
S. chilense[a]	0.016	0.020	0.027	0.018	0.038	0.016	0.027
Quicacha	0.004	0.016	0.022	0.009	0.030	0.014	0.023
Tacna	0.015	0.020	0.026	0.015	0.028	0.012	0.016
S. peruvianum[a]	0.015	0.220	0.032	0.019	0.045	0.015	0.025
Canta	0.016	0.020	0.032	0.019	0.044	0.016	0.027
Tarapaca	0.015	0.022	0.031	0.021	0.043	0.012	0.022
θ_W							
S. chilense[a]	0.014	0.025	0.026	0.020	0.039	0.018	0.028
Quicacha	0.006	0.014	0.021	0.011	0.026	0.010	0.023
Tacna	0.013	0.022	0.026	0.022	0.034	0.013	0.017
S. peruvianum[a]	0.022	0.026	0.038	0.029	0.047	0.021	0.034
Canta	0.020	0.020	0.035	0.021	0.044	0.016	0.031
Tarapaca	0.018	0.021	0.030	0.020	0.041	0.012	0.023

[a] Pooled over two populations

[b] From Fischer *et al.* (2011)

[c] From Xia *et al.* (2010)

The evolution of the *Asr4* and the *pLC30-15* promoter regions

Neither *pAsr4* nor *5'pLC* show reduced nucleotide diversity (Table 4.2). Compared to the *Asr4* gene, where evidence for local adaptation has been shown in a population from a dry environment (Fischer *et al.* 2011), nucleotide diversity is increasing rapidly when moving upstream (Table 4.2). The same is true for *5'pLC*, where signatures of local adaptation have also previously been detected at the *pLC30-15* gene in an *S. chilense* population (Xia *et al.*, 2010). The low haplotype diversity observed at *Asr4* in the Quicacha population increases at *pAsr4* close to 1 (Table 4.3), and no deviation from neutrality can be observed (Table 4.4). The haplotype diversity detected at *pLC30-15* in Quicacha can also be found at *5'pLC*. Here, H_d remains low (Table 4.3) and Fu and Li's *D* remains significantly negative. Additionally, Fay and Wu's *H* becomes very negative in both Quicacha and Tacna (Table 4.4), indicating positive selection.

Positive selection in the *pLC30-15* 3'UTR

As expected at non-coding regions, nucleotide diversity increases when looking at *3'pLC* compared to the *pLC30-15* gene (Table 4.2). However, this is not true in the Tacna population where θ_W and π remain low when moving downstream. The downstream region of *pLC30-15*

is indeed highly conserved. Furthermore, Fay and Wu's *H* becomes extremely negative, indicating positive selection (Table 4.4).

Table 4.3 Haplotype diversity of *pAsr2*, *pAsr4*, *5'pLC*, *3'pLC*, and their corresponding genes

H_d	*pAsr2*	*Asr2* [b]	*pAsr4*	*Asr4* [b]	*5' pLC*	*pLC* [c]	*3' pLC*
S. chilense [a]	0.916	0.975	0.986	0.946	0.959	0.954	0.974
Quicacha	0.286	0.800	0.933	0.600	0.667	0.711	0.889
Tacna	0.933	0.978	0.982	0.982	1.000	0.978	0.982
S. peruvianum [a]	0.965	0.979	0.980	0.985	0.977	0.977	0.965
Canta	0.945	0.933	1.000	0.956	1.000	0.978	0.972
Tarapaca	0.964	0.956	0.905	0.933	0.889	0.822	0.889

[a] Pooled over two populations

[b] From Fischer *et al.* (2011)

[c] From Xia *et al.* (2010)

Table 4.4 Results of the neutrality tests for *pAsr2*, *pAsr4*, *5'pLC*, *3'pLC*, and their corresponding genes

Tajima's *D*	*pAsr2* [d]	*Asr2* [b]	*pAsr4*	*Asr4* [b]	*5' pLC*	*pLC* [c]	*3' pLC*
S. chilense [a]	0.542	-0.832	0.244	-0.349	-0.071	-1.114	-0.158
Quicacha	**-1.688***	0.602	0.356	-0.511	0.887	**2.342***	0.022
Tacna	1.122	-0.372	0.078	-1.429	-0.811	-0.358	-0.499
S. peruvianum [a]	-1.299	-0.551	-0.603	-1.139	-0.207	-0.755	-1.051
Canta	-0.936	-0.121	-0.432	-0.548	0.008	-0.339	-0.700
Tarapaca	-0.783	0.356	0.170	0.029	0.137	0.010	-0.144
Fu and Li's *D*							
S. chilense [a]	0.246	-0.191	-0.391	-1.084	-0.310	-2.085	0.754
Quicacha	**-1.791****	-0.121	0.883	0.382	**1.853****	**1.587****	-0.432
Tacna	0.661	0.031	-0.230	**-2.326***	-1.220	-0.384	0.281
S. peruvianum [a]	-1.678	-0.551	-1.656	**-2.635***	-1.196	-1.463	-2.058
Canta	-1.423	-0.590	-1.179	-0.982	-0.916	-0.673	-1.497
Tarapaca	-0.784	0.225	-0.214	-0.223	0.180	0.119	-0.401
Fay and Wu's *H*							
S. chilense [a]	NA	2.588	3.320	-5.183	2.427	2.244	1.767
Quicacha	NA	-0.927	-3.556	-8.709	-9.190	-1.067	-2.667
Tacna	NA	5.626	3.200	-5.127	-7.727	1.511	-28.673
S. peruvianum [a]	NA	4.064	3.323	1.293	0.754	3.871	4.462
Canta	NA	0.444	3.778	1.867	7.911	0.978	8.806
Tarapaca	NA	0.000	5.143	0.889	-3.167	2.311	1.511

[a] Pooled over two populations

[b] From Fischer *et al.* (2011)

[c] From Xia *et al.* (2010)

[d] no outgroup available. Fu and Li's *D** without outgroup was calculated instead.

NA Not applicable

* $P<0.05$, ** $P<0.01$ (significant results are in bold)

Different types of motifs in the regulatory regions of *Asr2*, *Asr4,* and *pLC30-15*

We analyzed the regulatory regions of *Asr2*, *Asr4*, and *pLC30-15 in silico* in order to identify *cis*-acting elements. As these motifs are short (4-10 bases), some were discovered by chance. Here, we describe only those motifs related to hormone and stress response and organ specific expression, and motifs which lie in conserved regions (i.e. without polymorphism) in the whole species alignment (see Table 4.S4 for additional information of the described motifs). At *pAsr2*, we detected a motif conserved in both species involved in salicylic acid responsiveness (TCA-element). We discovered one motif conserved in *S. chilense* and one conserved in *S. peruvianum* involved in ethylene (ERE) and ABA responsiveness (ABRE), respectively. The ERE and ABRE (conserved in *S. peruvianum*) motifs can also be found at *pAsr4*. In addition, *pAsr4* contains a conserved auxin responsive element (Aux-RR-core). At *5'pLC* we can find five conserved ABRE motifs and a motif involved in methyl jasmonate responsiveness (CGTCA-motif); at *3'pLC* conserved ABRE and CGTCA-motifs are located, as well as an AuxRR-core.

Conserved stress related regulatory elements at *pAsr2* are involved in anaerobic induction (ARE), drought responsiveness (MBS), and general stress and defence response (TC-rich repeats). At *pAsr4*, a conserved stress related motif is ARE. TC-rich repeats and MBS motifs are conserved at *5'pLC* and *3'pLC*. *3'pLC* also contains motifs involved in low-temperature responsiveness (LTR) and heat stress responsiveness (HSE). We also found conserved motifs related to organ specific expression: regulatory elements specific for endosperm expression can be found at all regulatory regions studied (GCN4-motif at *pAsr2* conserved in *S. chilense*, Skn-1-motif at *pAsr4*, *5'pLC*, and *3'pLC*). At *pAsr2* and *pAsr4* we also see the CAT-box, involved in meristem expression.

DISCUSSION

We determined the expression patterns of the stress responsive genes *Asr1*, *Asr2*, *Asr4*, and *pLC30-15* in wild tomato populations from contrasting environments in order to detect local adaptation at the gene expression level. The plants were submitted to drought and cold stress and the transcript levels of the candidate genes were measured using quantitative PCR, a much more sensitive method than the ones previously used for these genes (Maskin *et al.*, 2001; Frankel *et al.*, 2006). We found gene expression of *Asr4* to be adaptive to drought conditions and the gene to be of potential interest for further functional studies. However, we also observe plasticity in *Asr* and *pLC30-15* expression. Analyses of the regulatory regions show a conserved *Asr2* promoter region and positive selection acting at *3'pLC* in a population

from a dry environment. *In silico* analyses of the promoter regions also give valuable insights into potential future downstream experiments.

The general expression pattern of the investigated genes is consistent with the expectation for early response genes. This kind of genes, activated within minutes after stress perception, usually consist of transcription factors which activate the late response genes (Mahajan & Tuteja, 2005). *Asr* genes were already described to be desiccation induced in several plant species including *Solanum chacoense* (Silhavy *et al.*, 1995), *P. taeda* (Padmanabhan *et al.*, 1997), *Lilium longifolium* (Wang *et al.*, 1998), *G. biloba* (Shen *et al.*, 2005), or *O. sativa* (Philippe *et al.*, 2010). These genes, however, have also been shown to be induced by other osmotic stress conditions. An *Asr1* homolog in *S. tuberosum* is induced by cold (Schneider *et al.*, 1997) and other *Asr* homologs are responsive to salt stress in *O. sativa* (Vaidyanathan *et al.*, 1999) and *G. biloba* (Shen *et al.*, 2005). We found that the *Asr* genes and *pLC30-15* are induced much stronger by water deficit than by cold, suggesting that they play a more important role in drought response than in cold response pathways in wild tomato species.

One very interesting finding of this study is that *Asr4* is induced much faster upon water deficit in the Tacna accession, which represents a very dry environment in *S. chilense*, compared to the other *S. chilense* accession from less dry habitats. We also observe a down-regulation of *Asr4* in this accession after 3h. This could explain the findings by Frankel *et al.* (2006), where *Asr4* expression could not be detected after 24h in drought-stressed wild tomato plants from a dry environment. Induction and down-regulation seem to be faster in populations from dry habitats and the transcript was simply not present in a measurable amount after 24h. In the Tacna population, Fischer *et al.* (2011) describe a predominant *Asr4* haplotype which seems to be favoured in this dry habitat but was absent in other *S. chilense* populations. The accession tested here is homozygous for this haplotype. In addition, *Asr4* is more strongly induced in a Tacna accession from a high altitude following cold stress compared to the other *S. chilense* accessions. These results suggest that the *Asr4* regulation is also adaptive – at least in *S. chilense* – and we therefore corroborate its role as a candidate gene for further functional experiments.

Asr1 shows a gradual induction after both stress treatments and no obvious difference between accessions according to their habitat. The findings disagree with the results of Frankel *et al.* (2006) who detected *Asr1* to be expressed only in water-deficit-stressed wild tomato plants from humid environments after 24h. Our data shows *Asr1* to be highly expressed in all accessions after 24h, after drought as well as after cold stress. One

explanation for this discrepancy may be that the method (Northern Blot) used by Frankel *et al.* (2006) was not sensitive enough to detect small amounts of *Asr1* transcripts. Another explanation could be considerable differences of gene expression between accessions of wild tomato, which will be discussed in the next paragraph. It has been demonstrated that *Asr1* is highly conserved (Carrari *et al.*, 2004; Frankel *et al.*, 2006; Fischer *et al.*, 2011), which fulfils Ohno's (1970) prediction on gene families that (at least) one member is subject to purifying selection and maintains the original function. Therefore it can be suggested that not only the *Asr1* gene but also its expression pattern, which could represent that of the ancestral *Asr* gene, is conserved.

Asr2 responds to water deficit more strongly in one *S. peruvianum* accession from the very dry habitat of Tarapaca (LA2744), but not in the other Tarapaca accession (LA2745). This is surprising, as both accessions were sampled in close proximity and under the same environmental conditions. Similar differences between both accessions can also be observed at the expression patterns of *Asr4* and *pLC30-15*. Therefore, it is crucial to keep in mind that differences in gene expression do not necessarily have to be adaptive. It has been demonstrated that stress-related genes have a more variable expression than housekeeping genes (Blake *et al.*, 2006; Maheshri & O'Shea, 2007) and a higher expression divergence of duplicated genes in *A. thaliana* (Ha *et al.*, 2007). Some variability in gene expression was argued to be advantageous as it allows the plant to rapidly respond to environmental changes (Lopez-Maury *et al.*, 2008). It was suggested that noise in gene expression helps the organism to adapt fast to new – not yet experienced – challenges (Furusawa & Kaneko, 2008). Indeed, *Asr* genes have been shown to be very variable in their expression kinetics. They show differences in organ specific expression (Padmanabhan *et al.*, 1997; Maskin *et al.*, 2001), different patterns depending on the gene copy (Frankel *et al.*, 2006; Philippe *et al.*, 2010) or applied stress (Schneider *et al.*, 1997; Shen *et al.*, 2005). Phillippe *et al.* (2010) even described differential expression patterns of rice *Asr4*, *Asr5*, and *Asr6* depending on the cultivar.

Studying *cis*-regulatory elements became of great interest over the last years as phenotypic changes sometimes result from variation in these regions rather than in coding regions (Wray, 2007). A relatively straightforward way to investigate this is to first analyze sequence variation and search for motifs in promoter regions (Proost *et al.*, 2009). Depending on the outcome of these investigations, feasible downstream experiments can be determined. When analyzing the *Asr2* promoter region, it is quite striking how highly conserved it is – even more than the *Asr2* gene itself. This suggests an important function of this region and the

action of purifying selection against changes at *pAsr2*. A study using a *pAsr2*-reporter gene construct already conferred promoter activity and induction by ABA (Rossi *et al.*, 1998). A similar approach could shed light on other factors inducing *Asr2* expression. *In silico* search for potential motifs at *pAsr2* may give hints which factors to investigate. Apart from a conserved ABA responsive element, we found ethylene and salicylic acid responsive elements. Both hormones are known to be involved in plant development and transpiration modulation. Other motifs suggest drought responsiveness, but also general stress response. Additionally, the 3'UTR of *pLC30-15* showed a low nucleotide diversity and patterns consistent with positive selection in the Tacna population. This observation is not necessarily linked to the *pLC30-15* gene but this region could be associated with another locus nearby. As the tomato genome is not yet fully aligned and analyzed it is not possible to make any suggestions about nearby loci. Motif analysis of *3'pLC* shows many conserved stress responsive elements, involved in drought, heat, and low temperature stress. We also found motifs linked to auxin and methyl jasmonate responsiveness – both hormones are involved in wounding stress.

Unlike *pAsr2* and *3'pLC*, *pAsr4* and *5'pLC* are more polymorphic than their corresponding transcribed regions. Interestingly, the patterns consistent with local adaptation at *Asr4* in the Tacna population described by Fischer *et al.* (2011) disappear when moving upstream – neutrality tests show no deviation from a neutral scenario. Also, the haplotype structure observed in Quicacha (Fischer *et al.*, 2011) is not present at *pAsr4*. This shows that the evidence for local adaptation described before are exclusive to the *Asr4* transcribed regions. We stated before that the expression of *Asr4* seems to be adaptive to dry environments. In the motif search we could not identify motifs involved in water deficit response but gene expression could also be modulated by other factors or yet not described motifs. It has also been shown that the promoter region of an *Asr4* homolog in potato is desiccation responsive (Dóczi *et al.*, 2002). Reporter gene constructs can give more insight if this is also the case in tomato. *5'pLC* is more polymorphic than the transcribed region, the haplotype structure and patterns of positive selection detected by Xia *et al.* (2010) in Quicacha remain, suggesting that the forces acting at *pLC30-15* are not limited to the gene. It is possible that these forces might as well be demography rather than selection. Consistent with our findings that *pLC30-15* is involved in drought stress response, the motif search revealed several conserved ABA responsive elements and a drought responsive element at *5'pLC*.

One has to keep in mind that our experiments applied extreme stresses. Especially drought stress is usually much less dramatic in nature. Here, we provide a study where we analyzed expression variation on a population based level in natural populations. So far, such surveys are rare in plant populations coming from their natural habitat. Such studies cannot only shed light on forces shaping natural phenotypic variation. They also provide candidate genes worth of further investigation, *e.g. Asr4* for transgenic experiments. We found *Asr4* to be an interesting candidate, in accordance with a previous study (Fischer *et al.*, 2011). A publication already confirmed an over-expression of maize *Asr1* to increase the drought tolerance in *Zea mays* (Jeanneau *et al.*, 2002). But there are also examples from transgenic plants: Tomato *Asr1* over-expressed in *Nicotiana tabacum* confers higher salt tolerance (Kalifa *et al.*, 2004b) and *L. longifolium Asr* leads to higher drought and salt resistance in *A. thaliana* (Yang *et al.*, 2005). *In silico* analysis of the regulatory regions can narrow down the possible factors which could have an effect on the *Asr* genes and *pLC30-15* expression. Our findings have of course to be confirmed using for example reporter gene constructs. Testing for expression variation and *in silico* analysis of promoter regions are relatively fast first steps towards the improvement of agricultural plant species. Limiting factors of this study are, however, that not all gene expression differences between populations have to be adaptive. There is a natural noise in the responses of stress related genes (Lopez-Maury *et al.*, 2008). Additionally, neutral transcriptome evolution has also been observed (Broadley *et al.*, 2008) but so far a model concerning this theory is lacking. Future studies on natural (plant) populations will certainly provide more insights into transcriptome evolution, also with special respect to local adaptation. Our work provides a good example for local adaptation to drought in a natural wild tomato population and, along with other surveys (Frankel *et al.*, 2003; Giombini *et al.*, 2009; Xia *et al.*, 2010; Fischer *et al.*, 2011), proves *Asr* genes and dehydrins to be promising candidate genes to investigate plant evolution and wild tomato species a valuable genetic resource for genes conferring resistance to abiotic stress.

ACKNOWLEDGEMENTS

We thank H. Lainer for her help in the Lab, S. Voigt for her valuable suggestions with the gene expression analysis, L. Zhang for help with the expression experiments, and the Munich Evolutionary Biology Group for discussion. We are also grateful to L. Camus-Kulandaivelu and M. Linnenbrink for valuable comments that improved the presentation of this paper. This work was funded by grant STE 325/13 from the German Research Foundation to WS and a fellowship from the *Graduiertenförderung nach dem Bayerischen Eliteförderungsgesetz* to IF.

5 General Discussion

For this thesis I studied wild tomato species from natural environments in order to detect patterns consistent with local adaptation on the DNA and gene expression level. The aim of the work presented in the first chapter was to investigate the potential for adaptation and the strength of selection acting at housekeeping genes in four *Solanum* species which dwell in different habitats. We found very strong purifying selection was acting at non-synonymous sites in all analyzed species. Purifying selection was also acting at non-coding sites (introns), but the strength of selection varied between species. However, we found no evidence for positive selection shaping the pattern observed at these loci. The goal of the work presented in the second chapter was to detect signatures of local adaptation in populations of two wild tomato species, *S. chilense* and *S. peruvianum*, at the *Asr* gene family. To distinguish between natural selection and demography we used a set of reference loci studied in the previous chapter. Those loci were already used to infer population structure and speciation in *S. chilense* and *S. peruvianum* (Roselius *et al.*, 2005; Städler *et al.*, 2005; Arunyawat *et al.*, 2007; Städler *et al.*, 2008). While sequencing the four members of the *Asr* gene family previously described we discovered a new copy, designated *Asr5*, which is highly similar to *Asr3*. Indeed, both copies seem to be subject to gene conversion. We analyzed all five copies of the *Asr* gene family and detected patterns consistent with local adaptation in a *S. chilense* population, but not in *S. peruvianum*. The individual members of the *Asr* gene family show quite diverse evolutionary histories, ranging from purifying selection, over gene conversion, to local adaptation. In the third chapter, I present results of a study which aimed to determine the expression pattern of the stress responsive genes *Asr1*, *Asr2*, *Asr4*, and *pLC30-15* from wild tomato populations from contrasting environments. We did this in order to detect local adaptation at the gene expression level. The plants were submitted to drought and cold stress and the transcript level of the candidate genes was measured by qPCR. As this method is more sensitive than the ones previously used to quantify gene expression of *Asr* genes in tomato (Maskin *et al.*, 2001; Frankel *et al.*, 2006), we gained more elaborate insight into the regulation of these genes. We describe gene expression of *Asr4* to be adaptive to drought conditions and the gene to be of potential interest for further functional studies. However, plasticity in *Asr* and *pLC30-15* expression is also evident. Analysis of the regulatory regions shows a conserved *Asr2* promoter region and positive selection acting at the 3'UTR of *pLC30-15* in a population from a dry environment. *In silico* analysis of the promoter regions also gives valuable insight into potential future downstream experiments.

5.1. Fitness effects of derived mutations at housekeeping genes in closely related *Solanum* species

Estimation of the strength of selection in four wild tomato species from different habitats using SFS-based methods revealed purifying selection acted at non-synonymous (NS) and non-coding (NC) sites. This is characterized by an excess of low-frequency derived polymorphism. However, past demographic expansions reported for these species created a similar pattern (Städler *et al.*, 2008; Städler *et al.*, 2009). As SFS-based methods are sensitive to demographic events (Fay *et al.*, 2001; Eyre-Walker & Keightley, 2007) this was taken into account when measuring the DFE. We also quantified the strength of selection using the fixation index F_{st}. Strong purifying selection will increase F_{st} as deleterious mutations cannot rise in frequency and will remain private to one population. In principle, DFE and F_{st} estimates are in agreement with regards to the intensity of selection, except for *S. arcanum* where DFE revealed strong purifying selection at NS and NC sites but F_{st} is lower than for synonymous (S) sites. These results can also be explained by very strong purifying selection in this species. Potential linkage disequilibrium in combination with very strong negative selection can lead to reduced F_{st} if the effective population size is small (Pamilo *et al.*, 1999), which is the case in *S. arcanum*. As all studied loci are housekeeping genes, it is not surprising that purifying selection is acting on their coding region. New mutations in coding regions can either be nearly neutral due to probable redundancy of protein function, or lethal if they hit key functions; recent empirical studies give evidence for this scenario and its role in the shape of the DFE (Fudala & Korona, 2009). Somewhat more unexpected is the fact that purifying selection acted at NC sites, although it varies among species. But evidence for natural selection on NC regions has been reported before, *e.g.* in mouse and human (Jareborg *et al.*, 1999), *A. thaliana* (Thomas *et al.*, 2007), or in balsam poplar (Olson *et al.*, 2010). As also introns have important functions (*e.g.* by expression regulation, correct splicing), purifying selection, however weaker than at NS sites (Jareborg *et al.*, 1999), might not be surprising.

The fact that these genes experienced purifying selection might make it problematic to use them as reference genes in the candidate gene approach applied in Chapter 2. Indeed, neutrally evolving loci would be preferable. However, it is extremely difficult to find loci fulfilling these requirements even in model organisms. The fact that we did not detect patterns consistent with local adaptation at these loci is of course advantageous but one has to keep these results in mind when interpreting the findings of the candidate gene study. In the past,

these reference loci served well to estimate demography and speciation in wild tomato (Roselius *et al.*, 2005; Städler *et al.*, 2005; Arunyawat *et al.*, 2007; Städler *et al.*, 2008; Städler *et al.*, 2009; Tellier *et al.*, 2011b). Employing next generation sequencing techniques will certainly provide more data on wild tomato and loci better suited for estimating past demographic events. Such approaches are currently realized by T. Städler at the ETH in Zurich and in our lab.

5.2. The evolution of *pLC30-15* and members of the *Asr* gene family

Investigating the expression pattern of *pLC30-15* and the *Asr* genes shows consistency with the expectation for early response genes. This kind of genes, activated within minutes after stress perception, usually consist of transcription factors which activate the late response genes (Mahajan & Tuteja, 2005). *Asr* genes were already described to be desiccation induced in several plant species: *e.g.* an *Asr4* homolog in the wild potato *Solanum chacoense* (Silhavy *et al.*, 1995), in roots of the pine species *P. taeda* (Padmanabhan *et al.*, 1997), in desiccated pollen of *Lilium longifolium* (Wang *et al.*, 1998), in several organs of *G. biloba* (Shen *et al.*, 2005), or source and sink leaves of *O. sativa* (Philippe *et al.*, 2010). Those genes, however, have also been shown to be induced by other osmotic stress conditions. An *Asr1* homolog in *S. tuberosum* is induced by cold (Schneider *et al.*, 1997) and other *Asr* homologs are responsive to salt stress in *O. sativa* (Vaidyanathan *et al.*, 1999) and *G. biloba* (Shen *et al.*, 2005). We found that the *Asr* genes and *pLC30-15* are induced much stronger by water deficit than by cold, suggesting that they play a more important role in drought response than in cold response pathways in wild tomato species.

Studying *cis*-regulatory elements became of great interest over the last years as phenotypic changes are suggested to result from variation in these regions rather than in coding regions (Wray, 2007). A relatively cheap way to investigate this is to first analyze sequence variation and search for motifs in promoter regions *in silico* (Proost *et al.*, 2009). Depending on the outcome of these investigations feasibility of downstream experiments can be determined. The promoter region of *pLC30-15* is more polymorphic than the gene, the haplotype structure and patterns of positive selection described in a previous study (Xia *et al.*, 2010) in Quicacha remain, suggesting that the forces acting at the *pLC30-15* gene are not limited to the gene. It is possible that these forces might as well be demography rather than selection. Not surprisingly, the motif search revealed several conserved ABA responsive elements and a drought responsive element at the *pLC30-15* promoter region. Analysis of the

3'UTR of *pLC30-15*, however, showed low nucleotide diversity and patterns consistent with positive selection in the Tacna population. This observation must not necessarily be linked to the *pLC30-15* gene since this region could be associated with another locus nearby. As the tomato genome is not yet fully aligned and analyzed it is not possible to make any suggestions which locus lies nearby to date. Motif analysis of the 3'UTR of *pLC30-15* shows many conserved stress responsive elements, involved in drought, heat, and low temperature stress. We also found motifs linked to auxin and methyl jasmonate responsiveness – both hormones are involved in wounding response. These conserved motifs could regulate a nearby stress-responsive gene (if not *pLC30-15* itself) and future analysis of the tomato genome will undoubtedly provide candidate genes in close proximity.

When analyzing the *Asr* gene family, one interesting aspect is the likely occurrence of gene conversion between *Asr3* and *Asr5*. A previous study describes *Asr3* to be exclusive to tomato (Frankel *et al.*, 2006); therefore the *Asr3-Asr5* duplication most likely occurred after the tomato cladogenesis. In our dataset, only the coding region seems to be target of gene conversion and it is therefore probable that two similar protein copies are beneficial for the plant's fitness. More proteins of the same kind are favoured in a "dosage" scenario which has been suggested to be the case in stress responsive genes in variable environments (Kondrashov *et al.*, 2002; Innan & Kondrashov, 2010). Another case where gene conversion is advantageous is when diversifying selection is acting, since new haplotypes can be created (Innan, 2009). Unfortunately, it is extremely difficult to detect signatures of diversifying selection in genes that undergo gene conversion as standard neutrality tests cannot be applied (Innan, 2003; Thornton *et al.*, 2007).

At the *Asr1* gene, we found evidence for purifying selection that has been acting on the coding region. This conclusion can be drawn as the nucleotide diversity is very low in the coding region in both species, even compared to the reference loci. In addition, *Asr1* shows low divergence to the outgroup, with a K_a/K_s ratio close to zero. *Asr1* has been described before to be very conserved (Frankel *et al.*, 2006). But in contrast to Frankel *et al.* (2006) we did not observe lower synonymous nucleotide diversity at *Asr1* than at the other *Asr* genes. The level of polymorphism at synonymous sites is also comparable with that of the reference loci. But the low level of non-synonymous polymorphism leads to almost identical protein sequences of ASR1 among all studied lines. When investigating the expression pattern, *Asr1* shows a gradual induction after both stress treatments and no obvious difference between populations. The findings disagree with the results of Frankel *et al.* (2006), who detected *Asr1* to be expressed after 24h only in plants from humid environments after application of water-

deficit stress. Our data shows *Asr1* highly expressed in all accessions after 24 h, after drought as well as after cold stress. One explanation for this discrepancy can be that the method (Northern Blot) used by Frankel *et al.* (2006) was not sensitive enough to detect small amounts of *Asr1* transcripts. Another explanation could be considerable differences of gene expression between accessions of wild tomato (discussed in the next paragraph). As *Asr1* is highly conserved (Carrari *et al.*, 2004; Frankel *et al.*, 2006) it fulfils Ohno's (1970) prediction on gene families that (at least) one member is subject to purifying selection and maintains the original function. The fact that *Asr* homologs with a similar protein sequence to tomato *Asr1* can be found over the plant kingdom (Frankel *et al.*, 2006) suggests that this gene has an important function and even small alterations in the protein have serious consequences for the plants' fitness. It has been shown that *Asr1* is expressed in a housekeeping manner in unstressed plants (Maskin *et al.*, 2001; Frankel *et al.*, 2006). Therefore, it can be suggested that not only the *Asr1* gene but also its expression pattern, which could represent the one of the ancestral *Asr* gene, is conserved.

5.3. The significance of *Asr2* as a candidate gene

Asr2 responds to water deficit much stronger in one *S. peruvianum* accession from the very dry habitat of Tarapaca (LA2744), but not in the other Tarapaca accession (LA2745). This is surprising, as both accessions were sampled in close proximity and in the same environmental conditions. Similar differences between both accessions can also be observed at the expression patterns of *Asr4* and *pLC30-15*. It is therefore crucial to keep in mind that differences in gene expression do not necessarily have to be adaptive. It has been demonstrated that stress related genes have noisier expression than housekeeping genes (Blake *et al.*, 2006; Maheshri & O'Shea, 2007) and a higher expression divergence of duplicated genes in *A. thaliana* (Ha *et al.*, 2007). It has been argued that some variability in gene expression is advantageous as it allows the plant to rapidly response to environmental changes (Lopez-Maury *et al.*, 2008). It was suggested that noise in gene expression helps the organism to adapt fast to new – yet not experienced – challenges (Furusawa & Kaneko, 2008). Indeed, *Asr* genes have been shown to be very variable in their expression kinetics. First, they show differences in organ specific expression. For example, an *Asr* homolog water-deficit-stressed in loblolly pine is expressed in roots, but not in stems and needles (Padmanabhan *et al.*, 1997) and tomato *Asr1* is induced upon drought in leaves, but not in roots (Maskin *et al.*,

2001). Second, different patterns depending on the gene copy can also be observed. Unlike *Asr2*, *Asr1* and *Asr4* seem to be induced after desiccation in wild tomato (Frankel *et al.*, 2006) and rice *Asr1* and *Asr2* are up-regulated in water-deficit-stressed source leaves, whereas *Asr3* and *Asr4* are down-regulated (Philippe *et al.*, 2010). Third, different kinds of stress lead to expression differences. An *Asr1* homolog in potato tubers is induced after cold, but not after drought and salt stress (Schneider *et al.*, 1997). In ginkgo roots, *Asr* shows a higher induction following desiccation than after treatment with NaCl (Shen *et al.*, 2005). Philippe *et al.* (2010) even described differential expression patterns of rice *Asr4*, *Asr5*, and *Asr6* depending on the cultivar. Additionally, neutral transcriptome evolution has also been observed (Broadley *et al.*, 2008) but so far a model concerning this theory is lacking. Future studies on natural (plant) populations will certainly provide more insight into transcriptome evolution, also with special respect to local adaptation.

Two previous studies detected patterns consistent with local adaptation at the *Asr2* gene in tomato populations from dry environments (Frankel *et al.*, 2003; Giombini *et al.*, 2009). In our data, we find no indication for adaptation at *Asr2*, such as patterns of low diversity within a population and a high fixation index between populations. Therefore, we cannot confirm this previous hypothesis. Frankel *et al.* (2003) used a phylogenetic approach to detect adaptation where only few (four to eight) alleles were analyzed for each of the seven *Solanum* species/cultivars investigated. They found a K_a/K_s ratio >1 in several between-species comparisons and an accelerated rate of evolution in species from dry environments (*i.e. S. chilense* and *S. peruvianum* v. *humifusum*). They retrieved their samples from the Tomato Genetics Resource Center where the *S. chilense* accession derived from Antofagasta. The Antofagasta sample was again used in the study by Giombini *et al.* (2009) who also found signatures of positive selection in *S. chilense* using a candidate gene approach. However, previous analysis of the Antofagasta sample in our group showed that this population deviates strongly from equilibrium conditions (Arunyawat *et al.*, 2007). Population bottlenecks can cause skews in the site frequency spectrum that may be confounded with natural selection (Holliday *et al.*, 2010). In both studies, Antofagasta was the only sample analyzed to infer adaptation in *S. chilense* and Giombini *et al.* (2009) found the same patterns described at *Asr2* at the reference locus used in their analysis. It remains possible that local adaptation acted in this population. However, given the results of Arunyawat *et al.* (2007), a demographic effect is more likely to have caused the patterns observed by Frankel *et al.* (2003) and Gombini *et al.* (2009).

When analyzing the *Asr2* promoter region, however, it is quite striking how highly conserved it is – even more than the *Asr2* gene itself. This suggests an important function of this region and purifying selection is acting against changes at the promoter of *Asr2*. A study using a *pAsr2*-reporter gene construct already conferred promoter activity and induction by ABA (Rossi *et al.*, 1998). A similar approach could shed light on other factors inducing *Asr2* expression. *In silico* search for potential motifs at the *Asr2* promoter region can give hints which factors to investigate. Apart from a conserved ABA responsive element, we found ethylene and salicylic acid responsive elements and both hormones are known to be involved in plant development and transpiration modulation. Other motifs suggest drought responsiveness, but also general stress response.

5.4. The role of *Asr4* in adaptation to drought

In contrast to the *Asr2* gene, we found patterns of local adaptation at *Asr4* in two *S. chilense* populations, Quicacha and Tacna. The interpretation of the data may be somewhat ambiguous for Quicacha, as in this population the reference loci also show some indication of population structure (although not significant). In Tacna on the other hand, the evidence for adaptation is striking since we do not observe any haplotypic structure at the reference loci in this population. Moreover, the fixation index F_{st} between Tacna and the other *S. chilense* populations is very high compared to the reference loci and the other *Asr* genes – indicating limited gene flow at *Asr4* between Tacna and its neighbouring populations. An elevated F_{st} is a hallmark for local adaptation where an allele is favoured in one population, but not in others (Beaumont & Balding, 2004; Foll & Gaggiotti, 2008; Riebler *et al.*, 2008). Interestingly, the patterns consistent with local adaptation at *Asr4* in the Tacna population disappear when moving upstream – neutrality tests show no deviation from a neutral scenario. Also, the haplotype structure observed in Quicacha at *Asr4* is not present at the promoter region. This shows that the evidence for local adaptation is exclusive to the *Asr4* gene. The first part of the first exon seems to play a major role in adaptation. This region is characterized by a repetitive region with a very high level of polymorphism and indels of different length. This repetitive region possesses a SYG motif that has been described to be stress responsive in *Saccharomyces cerevisiae* (Treger & McEntee, 1990). A similar region was also detected in DS2, an ASR4 homolog in wild potato (Silhavy *et al.*, 1995). The authors describe this part of the protein as highly hydrophilic and discuss sequence similarities with LEA proteins as these also show this hydrophilic character at the N-terminal end. Interestingly, the rice ASR6

protein has an insertion at the N-terminal, which displays some sequence similarity to the tomato ASR4 (Frankel *et al.*, 2006), but lacks the SYG motif. Since this insertion is already present at an *Asr* gene in a monocot, an important function of this region can be inferred as it was already present in an ancestral *Asr* gene in early land plants. Alternatively, these domains might also have been gained independently, which would be a remarkable example of convergent evolution. Since the zinc-binding sites of the ASR1 protein are located in the first exon (Rom *et al.,* 2006), it may be that variation in this region of ASR4 has an influence on the zinc-binding activity.

Interestingly, the gene expression of *Asr4* is induced much faster upon water deficit in the Tacna accession, which represents a very dry environment in *S. chilense*, compared to the other *S. chilense* accession. We also observe a fast down-regulation of *Asr4* in this accession after 3h and a very low transcript level after 24h. These results are consistent with the findings by Frankel *et al.* (2006): In their study, *Asr4* expression could not be detected after 24h in drought-stressed wild tomato plants from a dry environment. Most likely the transcript was not present in a measurable amount after 24h as our data indicates that induction as well as down-regulation seems to be faster in populations from dry habitats. In the Tacna population, we describe a predominant *Asr4* haplotype which seems to be favoured in this dry habitat and was absent in other *S. chilense* populations. Indeed, the accession tested here is homozygous for this predominant haplotype. In addition, *Asr4* is more strongly induced in a Tacna accession from a high altitude following cold stress compared to the other *S. chilense* accessions. Our results suggest that the *Asr4* regulation might also be adaptive and we therefore corroborate its role as a candidate gene for further functional experiments. However, in the motif search we could not identify regulatory elements involved in water-deficit or cold response but gene expression could also be modulated by other factors. It has also been shown that the promoter region of an *Asr4* homolog in potato is desiccation responsive (Dóczi *et al.*, 2002). Again, reporter gene constructs can give more insight if this is also the case in tomato.

5.5. Conclusions and outlook

Wild relatives of crop species have many advantages making these non-model organisms interesting tools for studying plant evolution. First, they provide a first step towards agriculture improvement of their cultivated relatives. Second, a certain amount of genomic information from their domesticated sister species is usually available. And third, they are

sampled from the ecological context they evolved in. We provide a study where we analyzed genetic and expression variation on a population based level in natural populations (Fig. 5.1).

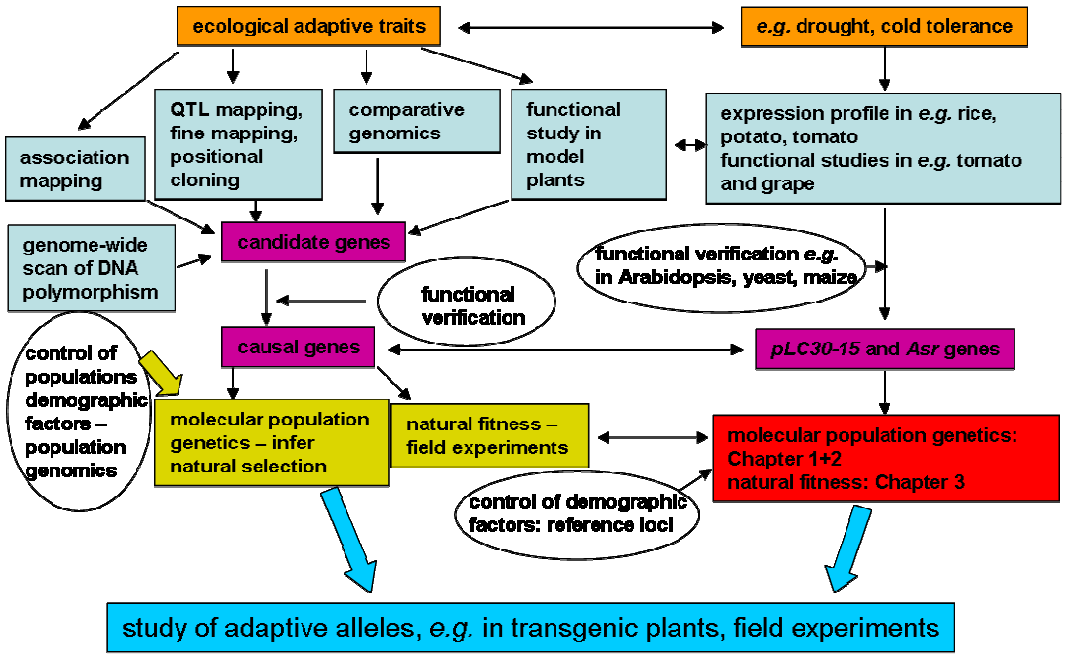

Figure 5.1 Interdisciplinary approaches to evolutionary and ecological genomic studies. Approaches are shown linking variation in phenotypes to variation on genotypes, followed by testing natural selection and fitness in natural environments. The left part of the chart is a general picture; the right part provides specific details from this study. Our contribution is highlighted in red, the further outlook in blue. QTL, quantitative trait locus; *Asr*, abscisic acid/water stress/ripening induced. Modified from Song & Mitchell-Olds (2011).

So far, surveys combining both are rare in plant populations coming from their natural habitat. Such studies cannot only shed light on forces shaping natural phenotypic variation. They also provide candidate genes worth of further investigation, *e.g.* for transgenic experiments. We found *Asr4* to be an interesting candidate, both in the expression and population genetics analysis (Fig. 5.1). A study already confirmed an over-expression of maize *Asr1* to increase the drought tolerance in maize (Jeanneau *et al.*, 2002). But there are also examples from transgenic plants: Tomato *Asr1* over-expressed in *Nicotiana tabacum* confers higher salt tolerance (Kalifa *et al.*, 2004b) and lily *Asr* leads to higher drought and salt resistance in *A. thaliana* (Yang *et al.*, 2005). Wild *Solanum* species served as genetic resource for their cultivated relatives in recent years, however mostly for pathogen resistance traits. *Solanum nigrum* is one such case and shows resistance to races of late blight, which is caused by the oomycete *Phytophthora infestans* – a serious disease in tomato and potato (Campos *et al.*, 2002; Lebecka, 2008). Resistance to *P. infestans* has successfully been transferred from

species of the *S. nigrum* complex to potato (Colon *et al.*, 1993; Zimnoch-Guzowska *et al.*, 2003). In addition, *in silico* analysis of the regulatory regions can narrow down the possible factors which could have an effect on *Asr* and *pLC30-15* expression – findings which have to be confirmed using for example reporter gene constructs. Testing for expression variation and *in silico* analysis of promoter regions are relatively fast first steps towards improvement of agricultural plant species (Proost *et al.*, 2009).

In conclusion, we observed that the *Asr* family is highly dynamic as the individual members show quite diverse evolutionary histories. This observation is not uncommon in plants, since adaptive specialization is thought to occur after gene duplication (Flagel & Wendel, 2009). In rice, a recent study of the *Asr* gene family revealed the same pattern we observe here: the members are experiencing different (gene- and species-specific) evolutionary histories and a candidate gene for drought tolerance (rice *Asr3*) could be identified (Philippe *et al.*, 2010). Interestingly, this potential candidate for drought adaptation clusters in tandem with rice *Asr4*, whereas the other rice *Asr* genes are distributed over the whole genome (Frankel *et al.*, 2006). This supports recent findings, mostly in *Arabidopsis* and rice, that stress responsive gene families are preferably clustered in tandem (Maere *et al.*, 2005; Mondragon-Palomino & Gaut, 2005; Rizzon *et al.*, 2006; Hanada *et al.*, 2008; Zou *et al.*, 2009). With the *Asr* gene family we therefore provide a good example of a tandemly arrayed gene family that is of importance in adaptation to drought. We detected patterns of local adaptation to dry environments at an *Asr* gene in the *S. chilense* population from Tacna, which is the driest *S. chilense* environment sampled. Interestingly, we did not find any evidence for local adaptation in *S. peruvianum*, a species that inhabits a broader geographical and environmental range. This is consistent with the results from the first chapter: constraint is more relaxed on the NC regions (maybe influencing the gene expression) of *S. peruvianum* and *S. habrochaites*. This could lead to a higher phenotypic plasticity which is advantageous for species inhabiting a broad variety of habitats. *Solanum arcanum* and *S. chilense*, on the other hand, occupy narrower niches and sometimes extreme environments. This leads to a higher constraint in plasticity which is in consistency with previous studies (Arunyawat *et al.*, 2007; Xia *et al.*, 2010). As *S. peruvianum* (and *S. habrochaites*) shows high nucleotide and morphological diversity, it may be hypothesized that this species can cope with a great variety of environmental conditions, unlike *S. chilense* (and *S. arcanum*) which seems to be undergoing local adaptations more frequently. In consistence with previous work by Nakazato *et al.* (2010) we find the potential distribution of *S. peruvianum* to be broader than the one of *S. chilense*. Indeed, *S. peruvianum* seems to be covering a large variety of potential habitats.

S. chilense, on the other hand, shows the best fitting area at the Atacama Desert, one of the driest regions in the world.

Bibliography

Amitai-Zeigerson H, Scolnik PA, Bar-Zvi D. 1995. Tomato *Asr1* mRNA and protein are transiently expressed following salt stress, osmotic stress and treatment with abscisic acid. *Plant Science* **110:** 205-213.

Anderson JT, Willis JH, Mitchell-Olds T. 2011. Evolutionary genetics of plant adaptation. *Trends in Genetics* **27:** 258-266.

Arunyawat U, Stephan W, Städler T. 2007. Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Molecular Biology and Evolution* **24:** 2310-2322.

Battaglia M, Olvera-Carrillo Y, Garciarrubio A, Campos F, Covarrubias AA. 2008. The enigmatic LEA proteins and other hydrophilins. *Plant Physiology* **148:** 6-24.

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

Beaumont MA, Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13:** 969-980.

Beisswanger S, Stephan W. 2008. Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated *polyhomeotic* genes in *Drosophila*. *Proceedings of the National Academy of Sciences* **105:** 5447-5452.

Blake WJ, Balazsi G, Kohanski MA, Isaacs FJ, Murphy KF, Kuang Y, Cantor CR, Walt DR, Collins JJ. 2006. Phenotypic consequences of promoter-mediated transcriptional noise. *Molecular Cell* **24:** 853-865.

Bray EA. 1997. Plant responses to water deficit. *Trends in Plant Science* **2:** 48-54.

Bray EA. 2004. Genes commonly regulated by water-deficit stress in *Arabidopsis thaliana*. *Journal of Experimental Botany* **55:** 2331-2341.

Broadley MR, White PJ, Hammond JP, Graham NS, Bowen HC, Emmerson ZF, Fray RG, Iannetta PPM, McNicol JW, May ST. 2008. Evidence of neutral transcriptome evolution in plants. *New Phytologist* **180:** 587-593.

Cakir B, Agasse A, Gaillard C, Saumonneau A, Delrot S, Atanassova R. 2003. A grape ASR protein involved in sugar and abscisic acid signaling. *The Plant Cell* **15:** 2165-2180.

Campos MA, Ribeiro SG, Ridgen DJ, Monte DC, Grossi de Sa MF. 2002. Putative pathogenesis-related genes within *Solanum nigrum* L. var. *americanum* genome: isolation of two genes coding for PR5-like proteins, phylogenetic and sequence analysis. *Physiological and Molecular Plant Pathology* **61:** 205-216.

Canel C, Bailey-Serres JN, Roose ML. 1995. Pummelo fruit transcript homologous to ripening-induced genes. *Plant Physiology* **108:** 1323–1324.

Carrari F, Fernie AR, Iusem ND. 2004. Heard it through the grapevine? ABA and sugar cross-talk: the ASR story. *Trends in Plant Science* **9:** 57-59.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. *Molecular Biology and Evolution* **23:** 1348-1356.

Chen RD, Campeau N, Greer AF, Bellemare G, Tabaeizadeh Z. 1993. Sequence of a novel abscisic acid-induced and drought-induced cDNA from wild tomato (*Lycopersicon chilense*). *Plant Physiology* **103:** 301.

Chetelat RT, Pertuzé RA, Faúndez L, Graham EB, Jones CM. 2009. Distribution, ecology and reproductive biology of wild tomatoes and related nightshades from the Atacama Desert region of northern Chile. *Euphytica* **167:** 77-93.

Colautti RI, Barrett SCH. 2010. Natural selection and genetic constraints on flowering phenology in an invasive plant. *International Journal of Plant Sciences* **171:** 960-971.

Colon LT, Eijlander R, Budding DJ, Pieters MMJ, Hoogendoorn J, Van-Ijzendoorn MT. 1993. Resistance to potato late blight (*Phytophthora infestans* (Mont.) de Bary) in *Solanum nigrum*, *S. villosum* and their sexual hybrids with *S. tuberosum* and *S. demissum*. *Euphytica* **66:** 55-64.

Cooper TF, Rozen DE, Lenski RE. 2003. Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proceedings of the National Academy of Sciences* **100:** 1072-1077.

Depaulis F, Veuille M. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular Biology and Evolution* **15:** 1788-1790.

Dóczi R, Csanaki C, Bánfalvi Z. 2002. Expression and promoter activity of the desiccation-specific *Solanum tuberosum* gene, *StDS2*. *Plant Cell and Environment* **25:** 1197-1203.

Doebley J, Stec A, Hubbard L. 1997. The evolution of apical dominance in maize. *Nature* **386:** 485-488.

Eckert AJ, Wegrzyn JL, Pande B, Jermstad KD, Lee JM, Liechty JD, Tearse BR, Krutovsky KV, Neale DB. 2009. Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*). *Genetics* **183:** 289-298.

Edmonds JM, Chweya JA. (1997). Black nightshades: *Solanum nigrum* L. and related species. In: Heller J, Engels J, Hammer K,eds. *Promoting the conservation and use of underutilized and neglected crops*. Rome, Italy: Institute of Plant Genetics and Crop Plant Research, Gatersleben/International Plant Genetic Resources Institute, 113.

Ellegren H. 2009. A selection model of molecular evolution incorporating the effective population size. *Evolution* **63:** 301-305.

Eveno E, Collada C, Guevara MA, Léger V, Soto A, Díaz L, Léger P, González-Martínez SC, Cervera MT, Plomion C *et al.* 2008. Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Molecular Biology and Evolution* **25:** 417-437.

Expósito-Rodríguez M, Borges AA, Borges-Pérez A, Pérez JA. 2008. Selection of internal control genes for quantitative real-time RT-PCR studies during tomato development process. *BMC Plant Biology* **8:** 131.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8:** 610-618.

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

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* **158:** 1227-1234.

Fay JC, Wittkopp PJ. 2008. Evaluating the role of natural selection in the evolution of gene regulation. *Heredity* **100:** 191-199.

Ferea TL, Botstein D, Brown PO, Rosenzweig RF. 1999. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences* **96:** 9721-9726.

Finkelstein RR, Gibson SI. 2001. ABA and sugar interactions regulating development: cross-talk or voices in a crowd? *Current Opinion in Plant Biology* **5:** 26-32.

Fischer I, Camus-Kulandaivelu L, Allal F, Stephan W. 2011. Adaptation to drought in two wild tomato species: the evolution of the *Asr* gene family. *New Phytologist* **190:** 1032-1044.

Fisher RA. 1930. *The genetical theory of natural selection*, Oxford, UK: Clarendon Press.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist* **183:** 557-564.

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

Fowler S, Thomashow MF. 2002. *Arabidopsis* transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. *Plant Cell* **14:** 1675-1690.

Frankel N, Hasson E, Iusem ND, Rossi MS. 2003. Adaptive evolution of the water stress-induced gene *Asr2* in *Lycopersicon* species dwelling in arid habitats. *Molecular Biology and Evolution* **20:** 1955-1962.

Frankel N, Carrari F, Hasson E, Iusem ND. 2006. Evolutionary history of the *Asr* gene family. *Gene* **378:** 74-83.

Frankel N, Nunes-Nesi A, Balbo I, Mazuch J, Centeno D, Iusem ND, Fernie AR, Carrari F. 2007. *ci21A/Asr1* expression influences glucose accumulation in potato tubers. *Plant Molecular Biology* **63:** 719-730.

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

Fudala A, Korona R. 2009. Low frequency of mutations with strongly deleterious but nonlethal fitness effects. *Evolution* **63:** 2164-2171.

Furusawa C, Kaneko K. 2008. A generic mechanism for adaptive growth rate regulation. *PLoS Computational Biology* **4:** e3.

Garay-Arroyo A, Colmenero-Flores JM, Garciarrubio A, Covarrubias AA. 2000. Highly hydrophilic proteins in prokaryotes and eukaryotes are common during conditions of water deficit. *Journal of Biological Chemistry* **275:** 5668-5674.

Giombini MI, Frankel N, Iusem ND, Hasson E. 2009. Nucleotide polymorphism in the drought responsive gene *Asr2* in wild populations of tomato. *Genetica* **136:** 13-25.

Glinka S, De Lorenzo D, Stephan W. 2006. Evidence of gene conversion associated with a selective sweep in *Drosophila melanogaster*. *Molecular Biology and Evolution* **23:** 1869-1878.

Grierson CS, Barnes SR, Chase MW, Clarke M, Grierson D, Edwards KJ, Jellis GJ, Jones JD, Knapp S, Oldroyd G *et al.* 2011. One hundred important questions facing plant science research. *New Phytologist* **192:** 6-12.

Grillo MA, Li C, Fowlkes AM, Briggeman TM, Zhou A, Schemske DW, Sang T. 2009. Genetic architecture for the adaptive origin of annual wild rice, *Oryza nivara*. *Evolution* **63:** 870-883.

Ha M, Li WH, Chen ZJ. 2007. External factors accelerate expression divergence between duplicate genes. *Trends in Genetics* **23:** 162-166.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiology* **148:** 993-1003.

Hellemans J, Mortier G, De Paepe A, Speleman F, Vandesompele J. 2007. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biology* **8:** R19.

Holliday JA, Yuen M, Ritland K, Aitken SN. 2010. Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. *Molecular Ecology* **19:** 3857-3864.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London, Series B* **256:** 119–124.

Hutter S, Saminadin-Peter SS, Stephan W, Parsch J. 2008. Gene expression variation in african and european populations of *Drosophila melanogaster*. *Genome Biology* **9:** R12.

Ingram J, Bartels D. 1996. The molecular basis of dehydration tolerance in plants. *Annual Review of Plant Physiology and Plant Molecular Biology* **47:** 377-403.

Innan H. 2003. The coalescent and infinite-site model of a small multigene family. *Genetics* **163:** 803-810.

Innan H. 2009. Population genetic models of duplicated genes. *Genetica* **137:** 19-37.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* **11:** 97-108.

Iusem ND, Bartholomew DM, Hitz WD, Scolnik PA. 1993. Tomato (*Lycopersicon esculentum*) transcript induced by water deficit and ripening. *Plant Physiology* **102:** 1353-1354.

Jareborg N, Birney E, Durbin R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Research* **9:** 815-824.

Jeanneau M, Gerentes D, Foueillassar X, Zivy M, Vidal J, Toppan A, Perez P. 2002. Improvement of drought tolerance in maize: towards the functional validation of the *Zm-Asr1* gene and increase of water use efficiency by over-expressing C4–PEPC. *Biochimie* **84:** 1127-1135.

Jones RJ, Mansfield TA. 1970. Suppression of stomatal opening in leaves treated with abscisic acid. *Journal of Experimental Botany* **21:** 714-719.

Kalifa Y, Gilad A, Konrad Z, Zaccai M, Scolnik PA, Bar-Zvi D. 2004a. The water- and salt-stress-regulated *Asr1* (abscisic acid stress ripening) gene encodes a zinc-dependent DNA-binding protein. *Biochemical Journal* **381:** 373-378.

Kalifa Y, Perlson E, Gilad A, Konrad Z, Scolnik PA, Bar-Zvi D. 2004b. Over-expression of the water and salt stress-regulated *Asr1* gene confers an increased salt tolerance. *Plant Cell and Environment* **27:** 1459-1468.

Kane NC, Rieseberg LH. 2007. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, *Helianthus annuus*. *Genetics* **175:** 1823-1834.

Kane NC, Rieseberg LH. 2008. Genetics and evolution of weedy *Helianthus annuus* populations: adaptation of an agricultural weed. *Molecular Ecology* **17:** 384-394.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107-116.

Knapp S, Bohs L, Nee M, Spooner DM. 2004. Solanaceae—A model for linking genomics with biodiversity. *Comparative and Functional Genomics* **5:** 285-291.

Knight CA, Vogel H, Kroymann J, Shumate A, Witsenboer H, Mitchell-Olds T. 2006. Expression profiling and local adaptation of *Boechera holboellii* populations for water use efficiency across a naturally occurring water stress gradient. *Molecular Ecology* **15:** 1229-1237.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome Biology* **3:** research0008.0001–0008.0009.

Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, Yano M. 2006. An SNP caused loss of seed shattering during rice domestication. *Science* **312:** 1392-1396.

Konrad Z, Bar-Zvi D. 2008. Synergism between the chaperone-like activity of the stress regulated ASR1 protein and the osmolyte glycine-betaine. *Planta* **227:** 1213-1219.

Larsen PF, Schulte PM, Nielsen EE. 2011. Gene expression analysis for the identification of selection and local adaptation in fishes. *Journal of Fish Biology* **78:** 1-22.

Lebecka R. 2008. Host–pathogen interaction between *Phytophthora infestans* and *Solanum nigrum*, *S. villosum*, and *S. scabrum*. *European Journal of Plant Pathology* **120:** 233-240.

Leinonen PH, Remington DL, Savolainen O. 2011. Local adaptation, phenotypic differentiation, and hybrid fitness in diverged natural populations of *Arabidopsis lyrata*. *Evolution* **65:** 90-107.

León P, Sheen J. 2003. Sugar and hormone connections. *Trends in Plant Science* **8:** 110-116.

Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S. 2002. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Research* **30:** 325-327.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25:** 1451-1452.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C}{}_{T}$ Method. *Methods* **25:** 402-408.

Lopez-Maury L, Marguerat S, Bähler J. 2008. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics* **9:** 583-593.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences* **102:** 5454-5459.

Mahajan S, Tuteja N. 2005. Cold, salinity and drought stresses: An overview. *Archives of Biochemistry and Biophysics* **444:** 139-158.

Maheshri N, O'Shea EK. 2007. Living with noisy genes: How cells function reliably with inherent variability in gene expression. *Annual Review of Biophysics and Biomolecular Structure* **36:** 413-434.

Martin G, Lenormand T. 2006a. The fitness effect of mutations across environments: A survey in light of fitness landscape models. *Evolution* **60:** 2413-2427.

Martin G, Lenormand T. 2006b. A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. *Evolution* **60:** 893-907.

Martínez-Fernández M, Bernatchez L, Rolán-Alvarez E, Quesada H. 2010. Insights into the role of differential gene expression on the ecological adaptation of the snail *Littorina saxatilis*. *BMC Evolutionary Biology* **10:** 356.

Maskin L, Gudesblat GE, Moreno JE, Carrari FO, Frankel N, Sambade A, Rossi M, Iusem ND. 2001. Differential expression of the members of the *Asr* gene family in tomato (*Lycopersicon esculentum*). *Plant Science* **161:** 739-746.

Maskin L, Frankel N, Gudesblat G, Demergasso MJ, Pietrasanta LI, Iusem ND. 2007. Dimerization and DNA-binding of ASR1, a small hydrophilic protein abundant in plant tissues suffering from water loss. *Biochemical and Biophysical Research Communications* **352:** 831-835.

Maskin L, Maldonado S, Iusem ND. 2008. Tomato leaf spatial expression of stress-induced *Asr* genes. *Molecular Biology Reports* **35:** 501-505.

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

Moeller DA, Tiffin P. 2008. Geographic variation in adaptation at the molecular level: a case study of plant immunity genes. *Evolution* **62:** 3069-3081.

Mondragon-Palomino M, Gaut BS. 2005. Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. *Molecular Biology and Evolution* **22:** 2444-2456.

Moyle LC. 2008. Ecological and evolutionary genomics in the wild tomatoes (*Solanum* Sect. *Lycopersicon*). *Evolution* **62:** 2995-3013.

Müller L, Hutter S, Stamboliyska R, Saminadin-Peter SS, Stephan W, Parsch J. 2011. Population transcriptomics of *Drosophila melanogaster* females. *BMC Genomics* **12:** 81.

Nakazato T, Warren DL, Moyle LC. 2010. Ecological and geographic modes of species divergence in wild tomatoes. *American Journal of Botany* **97:** 680-693.

Neiman M, Olson MS, Tiffin P. 2009. Selective histories of poplar protease inhibitors: elevated polymorphism, purifying selection, and positive selection driving divergence of recent duplicates. *New Phytologist* **183:** 740-750.

Ohno S. 1970. *Evolution by gene duplication*, New York City, NY, USA: Springer.

Olmstead RG, Palmer JD. 1992. A chloroplast DNA phylogeny of the Solanaceae - subfamilial relationships and character evolution. *Annals of the Missouri Botanical Garden* **79:** 346-360.

Olmstead RG, Sweere JA. 1994. Combining data in phylogenetic systematics - an emperical approach using three molecular data sets in the Solanaceae. *Systematic Biology* **43:** 467-481.

Olson MS, Robertson AL, Takebayashi N, Silim S, Schroeder WR, Tiffin P. 2010. Nucleotide diversity and linkage disequilibrium in balsam poplar (*Populus balsamifera*). *New Phytologist* **186:** 526-536.

Orr HA. 2005. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics* **6:** 119-127.

Padmanabhan V, Dias DMAL, Newton RJ. 1997. Expression analysis of a gene family in loblolly pine (*Pinus taeda* L.) induced by water deficit stress. *Plant Molecular Biology* **35:** 801-807.

Pamilo P, Pálsson S, Savolainen O. 1999. Deleterious mutations can reduce differentiation in small, subdivided populations. *Hereditas* **130:** 257-264.

Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and *Tetraodon*. *Trends in Genetics* **22:** 597-602.

Peralta IE, Spooner DM. 2001. Granule-bound starch synthase (GBSSI) gene phylogeny of wild tomatoes (*Solanum* L. section *Lycopersicon* [Mill.] Wettst. subsection *Lycopersicon*). *American Journal of Botany* **88:** 1888–1902.

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

Peralta IE, Spooner DM, Knapp S. 2008. Taxonomy of wild tomatoes and their relatives (*Solanum* sect. *Lycopersicoides*, sect. *Juglandifolia*, sect. *Lycopersicon*; Solanaceae). *Systematic Botany Monographs* **84:** 1–186 + 183 plates.

Philippe R, Courtois B, McNally KL, Mournet P, El-Malki R, Le Paslier MC, Fabre D, Billot C, Brunel D, Glaszmann J-C *et al.* 2010. Structure, allelic diversity and selection of *Asr* genes, candidate for drought tolerance, in *Oryza sativa* L. and wild relatives. *Theoretical and Applied Genetics* **121:** 769-787.

Proctor MCF, Oliver MJ, Wood AJ, Alpert P, Stark LR, Cleavitt NL, Mishler BD. 2007. Desiccation-tolerance in bryophytes: a review. *Bryologist* **110:** 595-621.

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K. 2009. PLAZA: A comparative genomics resource to study gene and genome evolution in plants. *The Plant Cell* **21:** 3718-3731.

R Development Core Team. 2005. *R: a language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Reyes JL, Rodrigo MJ, Colmenero-Flores JM, Gil JV, Garay-Arroyo A, Campos F, Salamini F, Bartels D, Covarrubias AA. 2005. Hydrophilins from distant organisms can protect enzymatic activities from water limitation effects *in vitro*. *Plant, Cell and Environment* **28:** 709-718.

Riebler A, Held L, Stephan W. 2008. Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics* **178:** 1817-1829.

Riihimäki M, Podolsky R, Kuittinen H, Koelewijn H, Savolainen O. 2005. Studying genetics of adaptive variation in model organisms: flowering time variation in *Arabidopsis lyrata*. *Genetica* **123:** 63-74.

Rizzon C, Ponger L, Gaut B. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Computational Biology* **2:** e115.

Roelofs D, Janssens TKS, Timmermans MJTN, Nota B, Marien J, Bochdanovits Z, Ylstra B, Van Straalen NM. 2009. Adaptive differences in gene expression associated with heavy metal tolerance in the soil arthropod *Orchesella cincta*. *Molecular Ecology* **18:** 3227-3239.

Rorat T, Szabala BM, Grygorowicz WJ, Wojtowicz B, Yin Z, Rey P. 2006. Expression of SK3-type dehydrin in transporting organs is associated with cold acclimation in *Solanum* species. *Planta* **224:** 205-221.

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

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

Rossi M, Iusem ND. 1994. Tomato (*Lycopersicon esculentum*) genomic clone homologous to a gene encoding an abscisic acid induced protein. *Plant Physiology* **104:** 1073-1074.

Rossi M, Carrari F, Cabrera-Ponce JL, Vázquez-Rovere C, Herrera-Estrella L, Gudesblat G, Iusem ND. 1998. Analysis of an abscisic acid (ABA)-responsive gene promoter belonging to the *Asr* gene family from tomato in homologous and heterologous systems. *Molecular and General Genetics* **258:** 1-8.

Saab IN, Sharp RI, Pritchard J, Voetberg GS. 1990. Increased endogenous abscisic acid maintains primary root growth and inhibits shoot growth of maize seedlings at low water potentials. *Plant Physiology* **93:** 1329-1336.

Saumonneau A, Agasse A, Bidoyen M-T, Lallemand M, Cantereau A, Medici A, Laloi M, Atanassova R. 2008. Interaction of grape ASR proteins with a DREB transcription factor in the nucleus. *FEBS Letters* **582:** 3281-3287.

Schneider A, Salamini F, Gebhardt C. 1997. Expression patterns and promoter activity of the cold-regulated gene *ci21A* of potato. *Plant Physiology* **113:** 335-345.

Seki M, Ishida J, Narusaka M, Fujita M, Nanjo T, Umezawa T, Kamiya A, Nakajima M, Enju A, Sakurai T *et al.* 2002. Monitoring the expression pattern of around 7,000 *Arabidopsis* genes under ABA treatments using a full-length cDNA microarray. *Functional & Integrative Genomics* **2:** 282-291.

Shan H, Zahn L, Guindon S, Wall PK, Kong H, Ma H, dePamphilis CW, Leebens-Mack J. 2009. Evolution of plant MADS box transcription factors: Evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. *Molecular Biology and Evolution* **26:** 2229-2244.

Shen G, Pang Y, Wu W, Deng Z, Liu X, Lin J, Zhao L, Sun X, Tang K. 2005. Molecular cloning, characterization and expression of a novel gene from *Ginkgo biloba*. *Plant Physiology and Biochemistry* **43:** 836-843.

Shinozaki K, Yamaguchi-Shinozaki K. 1996. Molecular responses to drought and cold stress. *Current Opinion in Biotechnology* **7:** 161-167.

Shinozaki K, Yamaguchi-Shinozaki K. 2000. Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. *Current Opinion in Plant Biology* **3:** 217-223.

Silhavy D, Hutvágner G, Barta E, Bánfalvy Z. 1995. Isolation and characterization of a water-stress-inducible cDNA clone from *Solanum chacoense*. *Plant Molecular Biology* **27:** 587-595.

Siol M, Wright SI, Barrett SCH. 2010. The population genomics of plant adaptation. *New Phytologist* **188:** 313-332.

Song B-H, Mitchell-Olds T. 2011. Evolutionary and ecological genomics of non-model plants. *Journal of Systematics and Evolution* **49:** 17-24.

Spooner DM, Anderson GJ, Jansen RK. 1993. Chloroplast DNA evidence for the interrelationships of tomatoes, potatoes, and pepinos (Solanaceae). *American Journal of Botany* **80:** 676-688.

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

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

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

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

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105:** 437-460.

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

Takuno S, Nishio T, Satta Y, Innan H. 2008. Preservation of a pseudogene by gene conversion and diversifying selection. *Genetics* **180:** 517-531.

Tanaka S, Ikeda K, Miyasaka H. 2004. Isolation of a new member of group 3 late embryogenesis abundant protein gene from a halotorelant green alga by a functional expression screening with cyanobacterial cells. *Fems Microbiology Letters* **236:** 41-45.

Tellier A, Laurent SJY, Lainer H, Pavlidis P, Stephan W. 2011b. Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences* **108:** 17052-17057.

Thomas BC, Rapaka L, Lyons E, Pedersen B, Freeling M. 2007. *Arabidopsis* intragenomic conserved noncolaing sequence. *Proceedings of the National Academy of Sciences* **104:** 3348-3353.

Thornton KR, Jensen JD, Becquet C, Andolfatto P. 2007. Progress and prospects in mapping recent selection in the genome. *Heredity* **98:** 340-348.

Treger JM, McEntee K. 1990. Structure of the DNA damage-inducible gene DDR48 and evidence for its role in mutagenesis in *Saccaromyces cerevisiae*. *Molecular and Cellular Biology* **10:** 3174-3184.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics* **42:** 260-263.

Vaidyanathan R, Kuruvilla S, Thomas G. 1999. Characterization and expression pattern of an abscisic acid and osmotic stress responsive gene from rice. *Plant Science* **140:** 21-30.

Wachowiak W, Balk PA, Savolainen O. 2009. Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (*Pinus sylvestris* L.). *Tree Genetics & Genomes* **5:** 117-132.

Wang C-S, Liau Y-E, Huang J-C, Wu T-D, Su C-C, Lin CH. 1998. Characterization of a desiccation-related protein in Lily pollen during development and stress. *Plant and Cell Physiology* **39:** 1307-1314.

Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, Bomblies K, Lukens L, Doebley JF. 2005. The origin of the naked grains of maize. *Nature* **436:** 714-719.

Watterson GA. 1975. Number of segregating sites in genetic models without recombination. *Theoretical Population Biology* **7:** 256-276.

Whalen MD, Caruso EE. 1983. Phylogeny in *Solanum* sect. *Lasiocarpa* (Solanaceae): Congruence of morphological and molecular data. *Systematic Botany* **8:** 369-380.

Willi Y, Van Buskirk J, Hoffmann AA. (2006). Limits to the adaptive potential of small populations. *Annual Review of Ecology Evolution and Systematics* **37:** 433-458.

Wilson AC, Maxson LR, Sarich VM. 1974. Two types of molecular evolution - evidence from studies of interspecific hybridization. *Proceedings of the National Academy of Sciences* **71:** 2843-2847.

Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nature Reviews Genetics* **8:** 206-216.

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

Yáñez M, Cáceres S, Orellana S, Bastías A, Verdugo I, Ruiz-Lara S, Casaretto JA. 2009. An abiotic stress-responsive bZIP transcription factor from wild and cultivated tomatoes regulates stress-related genes. *Plant Cell Reports* **28:** 1497-1507.

Yang C-Y, Chen Y-C, Jauh GY, Wang C-S. 2005. A lily ASR protein involves abscisic acid signaling and confers drought and salt resistance in Arabidopsis. *Plant Physiology* **139:** 836-846.

Yang ZH, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17:** 32-43.

Zhen Y, Ungerer MC. 2008. Relaxed selection on the *CBF/DREB1* regulatory genes and reduced freezing tolerance in the southern range of *Arabidopsis thaliana*. *Molecular Biology and Evolution* **25:** 2547-2555.

Zimnoch-Guzowska E, Lebecka R, Kryszczuk A, Maciejewska U, Szczerbakowa A, Wielgat B. 2003. Resistance to *Phytophthora infestans* in somatic hybrids of *Solanum nigrum* L. and diploid potato. *Theoretical and Applied Genetics* **107:** 43-48.

Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH. 2009. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. *PLoS Genetics* **5:** e1000581.

Appendix A: Supplementary Online Material Tellier *et al.* (2011)

Section 1: Species-wide analysis

Section 2: Population sample analysis

Section 3: Regulatory motif analysis

Section 1: Species-wide analysis

Table A.S1 Habitat characteristics of the analyzed populations of four *Solanum* species. Where applicable, the TGRC accession numbers are indicated (for descriptions, see Städler *et al*., 2005). All *S. chilense* and *S. peruvianum* populations have been described in Arunyawat *et al*. (2007).

Species	Population	Location	Coordinates (latitude, longitude)	Altitude (m)	Ecological habitat
S. peruvianum	Tarapaca (LA2744)	northern Chile	18°33'S, 70°09'W	400	dry
	Arequipa	southern Peru	16°27'S, 71°42'W	2180	dry
	Nazca	southern Peru	14°51'S, 74°44'W	2140	dry
	Canta	central Peru	11°31'S, 76°41'W	2000-2060	mesic
S. chilense	Antofagasta (LA2884)	northern Chile	22°14'S, 68°23'W	2900	dry
	Tacna	southern Peru	17°53'S, 70°07'W	1270	dry
	Moquegua	southern Peru	17°04'S, 70°52'W	2490	dry
	Quicacha	southern Peru	15°37'S, 73°48'W	1830	dry
S. habrochaites	Canta	central Peru	11°31'S, 76°41'W	2000-2060	mesic
	Ancash (LA1775)	central Peru	09°31'S, 77°53'W	900-1800	mesic
	Otuzco	northern Peru	07°56'S, 78°36'W	2450	mesic
	Contumaza	northern Peru	07°22'S, 78°48'W	2670	mesic
	Lajas	northern Peru	06°33'S, 78°46'W	1880-2090	dry-mesic
S. arcanum	Otuzco	northern Peru	07°56'S, 78°36'W	2350-2550	mesic
	Rupe	northern Peru	07°18'S, 78°48'W	1710-2000	dry-mesic
	San Juan	northern Peru	07°17'S, 78°31'W	1780-2390	mesic
	Cochabamba	northern Peru	06°29'S, 78°54'W	1850-1930	dry-mesic

Table A.S2 Chromosome location, putative function, and sizes of coding and non-coding regions of the studied loci in *S. habrochaites*. The number of sites in each category was estimated with the method of (Yang & Nielsen, 2000) and is based on the alignment of all available sequences in *S. habrochaites* (depending on the locus, 39–60 sequences).

Locus	Chromosome	Putative protein function	Non-coding region	Coding region	
				synonymous	non-synonymous
CT066	10	Arginine decarboxylase	0	361	983
CT093	5	S-adenosylmethionine decarboxylase proenzyme	359	273	755
CT166	2	Ferredoxin-NADP reductase	862	118	322
CT179	3	Tonoplast intrinsic protein D-type	279	177	402
CT198	9	Submergence induced protein 2-like	424	88	245
CT208	9	Alcohol dehydrogenase, class III	660	119	328
CT251	2	At5g37260 gene	376	336	1002
CT268	1	Receptor-like protein kinase	0	396	1488

Table A.S3 Values of Tajima's D_T per locus and per species for the pooled samples. SNPs are grouped by categories: synonymous, silent sites (*i.e.* non-coding and synonymous), and all sites.

Species	Site type	CT066	CT093	CT166	CT179	CT198	CT208	CT251	CT268	average D_T[a]
S. peruvianum	synonymous	-0.486	-1.436	**-1.879***	-0.694	-0.205	-0.982	-0.393	-0.528	-0.825
	silent	-0.486	-1.681	-1.675	-0.803	-0.793	-1.129	-0.781	-0.528	-0.985
	all sites	-0.561	-1.689	-1.684	-0.884	-0.784	-1.149	-0.938	-0.798	-1.061
S. chilense	synonymous	-0.293	-0.187	-0.561	0.010	-0.859	-1.480	0.746	0.809	-0.227
	silent	-0.293	-0.880	0.006	-0.634	-1.099	-0.411	0.098	0.809	-0.301
	all sites	-0.661	-1.068	-0.103	-0.715	-1.076	-0.411	-0.246	0.065	-0.527
S. habrochaites	synonymous	-0.225	-1.360	-1.164	-0.242	-0.147	0.210	0.134	0.215	-0.322
	silent	0.068	-1.623	-1.445	-1.198	0.202	-0.655	0.232	0.215	-0.526
	all sites	-0.390	-1.773	-1.480	-1.184	0.190	-0.655	-0.625	-0.140	-0.757
S. arcanum	synonymous	0.191	-1.717	-1.041	0.503	-1.440	-0.938	-0.754	-1.274	-0.809
	silent	0.191	-1.719	-0.241	0.347	-0.268	-1.775	-0.895	-1.274	-0.704
	all sites	0.032	-1.820	-0.241	0.237	-0.207	-1.775	-0.965	-1.107	-0.731

* significance level of $P < 0.05$

[a] arithmetic mean across eight loci

With the exception of *S. arcanum*, D_T values of all sites are lower than those of synonymous sites for most loci (and for the average across loci), whereas silent sites exhibit values that are on average close to those for synonymous sites or only slightly lower (Table 2.1). A few comparisons, however, do not follow this trend because some loci (CT066 and CT268) have only a coding region or have only one non-synonymous polymorphism (CT208 in *S. peruvianum*).

Table A.S4: McDonald-Kreitman table for the four species (pooled samples for coding regions).

Species	Locus	*Synonymous substitutions* *Ds*	*Synonymous polymorphisms* *Ps*	*Non-Synonymous substitutions* *Dn*	*Non-Synonymous polymorphisms* *Pn*	*P*-value of Fisher's exact test (two-tailed)
S. peruvianum	CT066	20	54	5	12	1.000
	CT093	10	48	4	11	0.467
	CT166	29	124	0	2	1.000
	CT179	18	86	2	6	0.632
	CT198	14	94	2	6	0.597
	CT208	26	90	0	1	1.000
	CT251	28	87	18	45	0.593
	CT268	12	80	14	50	0.190
S. chilense	CT066	24	50	5	13	0.784
	CT093	8	29	4	21	0.747
	CT166	30	76	0	3	0.560
	CT179	24	75	1	3	1.000
	CT198	20	50	3	4	0.667
	CT208	25	57	0	0	*na*
	CT251	27	67	20	42	0.722
	CT268	12	47	16	42	0.393
S. habrochaites	CT066	28	14	6	11	**0.042***
	CT093	16	17	5	3	0.697
	CT166	32	41	0	1	1.000
	CT179	19	56	2	2	0.569
	CT198	18	36	3	2	0.337
	CT208	34	25	0	0	*na*
	CT251	36	16	26	9	0.639
	CT268	18	30	14	26	0.828
S. arcanum	CT066	21	42	5	9	1.000
	CT093	12	32	4	12	1.000
	CT166	31	89	0	0	*na*
	CT179	17	44	2	2	0.574
	CT198	16	57	3	1	**0.044***
	CT208	27	58	0	0	*na*
	CT251	31	58	19	39	0.860
	CT268	16	31	16	30	1.000

na: not applicable (absence of non-synonymous polymorphisms and substitutions)

* significance level of $P < 0.05$

Table A.S5 K_a/K_s ratios for each locus and species.

Species	CT066	CT093	CT166	CT179	CT198	CT208	CT251	CT268
S. peruvianum	0.066	0.228	0.005	0.054	0.243	0.001	0.306	0.246
S. chilense	0.057	0.236	0.004	0.049	0.276	0[a]	0.344	0.270
S. habrochaites	0.073	0.218	0.001	0.065	0.313	0[a]	0.319	0.230
S. arcanum	0.061	0.196	0[a]	0.048	0.338	0[a]	0.348	0.292

[a] absence of non-synonymous polymorphisms/divergence

Taking divergence between species (or to an outgroup) into account, the K_a/K_s ratios are lower than one for all species and all loci. Loci CT166 and CT208 contain zero or few non-synonymous SNPs, indicating that they are under strong purifying selection. The McDonald-Kreitman test does not show significant departures from neutrality, except for two marginally significant cases (Table A.S4). At locus CT066 in *S. habrochaites*, a higher non-synonymous polymorphism than expected is observed (Fisher's exact test, $P < 0.05$). A lower non-synonymous polymorphism than expected is detected for CT198 in *S. arcanum* (Fisher's exact test, $P < 0.05$).

Table A.S6 Summary of DNA polymorphism for all polymorphic sites of each species. S*, NS* and NC* denote the proportions of polymorphic sites for synonymous, non-synonymous and non-coding sites for each frequency class (*i.e.* rare, intermediate, and common). The * denotes the number of SNPs per site, so the proportion of non-synonymous sites NS* is calculated as 1 minus the number of synonymous sites divided by the total number of coding region sites.

Species	Frequency of SNPs	# SNPs S	# SNPs NS	# SNPs NC	NS*	S*	NC*	NS*/S*	NC*/S*
S. peruvianum	*rare* ($f < 5\%$)	135	72	167	0.013	0.073	0.061	0.178	0.836
	intermediate ($5\% \leq f < 20\%$)	65	28	127	0.005	0.035	0.046	0.143	1.314
	common ($f \geq 20\%$)	78	26	44	0.005	0.042	0.016	0.119	0.381
S. chilense	*rare* ($f < 5\%$)	71	72	83	0.013	0.039	0.029	0.333	0.744
	intermediate ($5\% \leq f < 20\%$)	55	17	57	0.003	0.030	0.020	0.100	0.667
	common ($f \geq 20\%$)	63	27	73	0.005	0.034	0.026	0.147	0.765
S. habrochaites	*rare* ($f < 5\%$)	29	29	50	0.005	0.016	0.017	0.313	1.063
	intermediate ($5\% \leq f < 20\%$)	20	18	27	0.003	0.011	0.009	0.273	0.818
	common ($f \geq 20\%$)	35	11	41	0.002	0.019	0.014	0.105	0.737
S. arcanum	*rare* ($f < 5\%$)	76	46	86	0.008	0.041	0.030	0.195	0.732
	intermediate ($5\% \leq f < 20\%$)	47	16	70	0.003	0.026	0.025	0.115	0.962
	common ($f \geq 20\%$)	48	18	50	0.003	0.026	0.018	0.115	0.692

Table A.S7 Summary of the number of S, NS, NC polymorphisms used in the DFE calculations.

Species	# SNPs Synonymous	# SNPs Non-synonymous	# SNPs Non-coding
S. peruvianum	278	126	338
S. chilense	189	116	213
S. habrochaites	84	58	118
S. arcanum	171	80	206

Table A.S8 Eyre-Walker α for the multi-locus datasets as estimated using the DoFE software (Bierne and Eyre-Walker, 2004) with [confidence intervals], compared to estimates obtained with the method by Eyre-Walker and Keightley (2009).

Method	*S. peruvianum*	*S. chilense*	*S. habrochaites*	*S. arcanum*
α (Bierne and Eyre-Walker, 2004)	0.216 [-0.23, 0.5]	0.004 [-0.57, 0.36]	-0.17 [-0.96, 0.3]	0.043 [-0.53, 0.39]
α (Eyre-Walker and Keightley, 2009)	-0.098	0.262	-0.23	-0.14

Table A.S9 Mean of log-likelihood ratios over 50 simulated datasets in the power analysis.

Simulated $-N_eE(s)$	Expansion?	Mean of log-likelihood	Mean of estimated $-N_eE(s)$
0 (neutral)	*no expansion*	-2137	116
	5-fold	-3622	155
5	*no expansion*	-2750	$\mathbf{3 \cdot 10^6}$
	5-fold	-3697	$\mathbf{151 \cdot 10^3}$
50	*no expansion*	-2380	$\mathbf{22 \cdot 10^6}$
	5-fold	-3249	$\mathbf{424 \cdot 10^6}$
500	*no expansion*	-3095	$\mathbf{87 \cdot 10^6}$
	5-fold	-4979	$\mathbf{2 \cdot 10^9}$
5000	*no expansion*	-3522	$\mathbf{38 \cdot 10^9}$
	5-fold	-4525	$\mathbf{55 \cdot 10^{12}}$

Note that the highest log-likelihood ratios are for simulated datasets without population expansion and low selection coefficients ($-N_eE(s)$ = 5 or 50). Estimates of $-N_eE(s)$ show very high values tending to infinity (in bold).

Table A.S10 Mean of the log-likelihood for each species for DFE estimates of NC and NS sites.

	Non-synonymous sites		Non-coding sites	
Species	mean of log-likelihood	mean of estimated $-N_e E(s)$	mean of log-likelihood	mean of estimated $-N_e E(s)$
S. peruvianum	-2351	$2 \cdot 10^{15}$	-3112	5.1
S. chilense	-1935	$7.8 \cdot 10^{14}$	-2401	182
S. habrochaites	-1100	$5 \cdot 10^{12}$	-1473	20.4
S. arcanum	-1562	$2 \cdot 10^{16}$	-2224	205

There is no correlation between the log-likelihood score and the number of segregating sites. Fewer non-synonymous polymorphisms occur, compared to non-coding polymorphisms. However, the estimates of DFE show a higher log-likelihood for non-synonymous sites.

Section 2: Population sample analysis

BayeScan analysis We use the program BayeScan by Foll and Gaggiotti (2008) to detect outlier SNPs that deviate from the expected distribution of F_{st} values. In the approach developed by Foll and Gaggiotti (2008), the posterior probability that a locus is under selection is estimated assuming an island model of the metapopulation. Two selection models are given: a locus is or is not under selection. For both models, a Bayes Factor (BF) will be calculated and this factor indicates which model better fits the data. The program BayeScan calculates the F_{st} for every SNP and estimates the posterior probability of a SNP to be under the effect of selection. As above, to compensate for the low number of non-synonymous polymorphisms, the F_{st} distribution of silent polymorphisms (S + NC sites) is computed. The strength of selection acting on NC sites is thus inferred by comparing the F_{st} distribution at synonymous sites to that at silent sites. Then, the distribution at silent sites (S + NC) is compared to the one at all sites (S + NC + NS) to reveal selection on non-synonymous SNPs. Before running the program, polymorphic sites that are only present at low frequencies < 5% in the pooled samples are removed because they could contribute excessively to high F_{st} values. We chose a sample size (*i.e*. number of iterations used for estimation) of 5000 and a thinning interval (*i.e*. number of iterations between samples) of 20. The total number of iterations is the sample size times the thinning interval (in our case 100,000). Before starting the sampling, we ran ten pilot runs (with a run length of 5000) to adjust the proposal rates.

Figure A.S1a

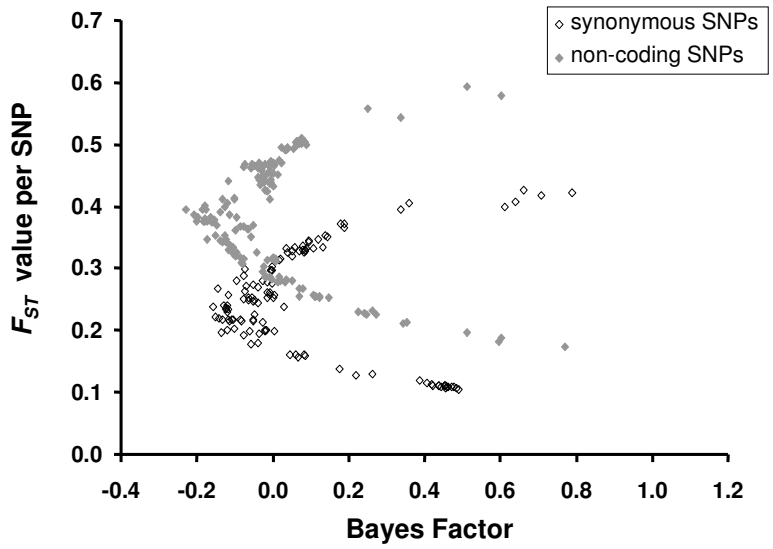

Figure A.S1b

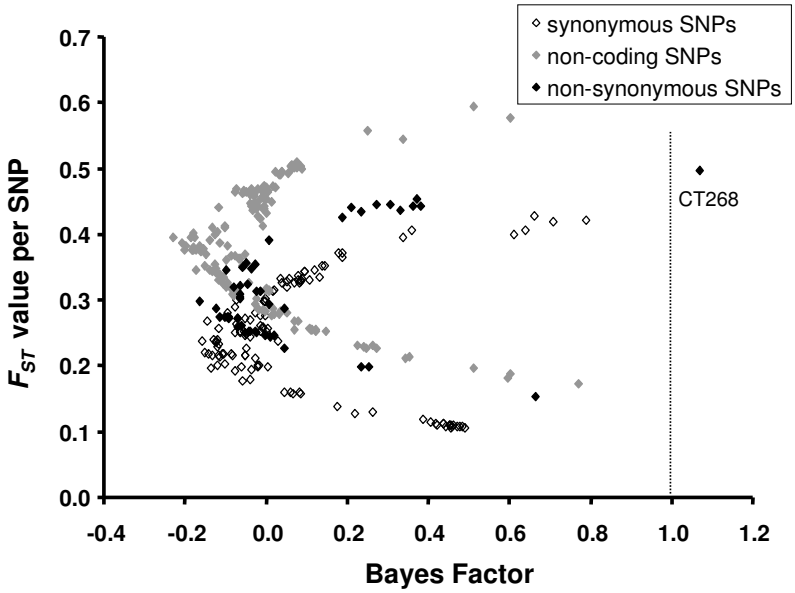

Figure A.S1 F_{st} analysis of individual SNPs from four populations of *S. chilense* with BayeScan (Foll and Gaggiotti, 2008). The Bayes Factor and F_{st} are calculated for each SNP: (a) synonymous (shown as white diamonds) and non-coding (grey diamonds) SNPs, (b) all SNPs, with non-synonymous SNPs shown asblackdiamonds.

In Figure A.S1a, the F_{st} distribution at non-coding SNPs is shown to be higher than that for synonymous polymorphisms. The distribution of F_{st} at NS sites shows an intermediate position between that at synonymous and NC sites (Figure A.S1b). One non-synonymous SNP is shown to exhibit a signal of positive selection (local adaptation) at locus CT268 (Figure A.S1b). The difference between results from all SNPs (Figure 2.3 in text) and those in BayeScan is the exclusion of low-frequency SNPs at the species level (rare at $f < 5\%$). In *S. chilense*, NS sites have a "lower" F_{st} distribution compared to NC sites without these rare SNPs, indicating that many NS polymorphic sites are found only as singletons or doubletons in the species and are private to single populations.

Note that despite the lack of statistical power for using BayeScan with only four or five populations, we show that the intraspecific F_{st} pattern due to purifying selection is observable in *S. chilense* (Figure A.S1). Sites under purifying selection are not, however, expected to be outliers in such an F_{st}-based method (Foll and Gaggiotti, 2008). Indeed, F_{st} increases noticeably with an increasing amount of private intermediate-frequency SNPs.

Section 3: Regulatory motif analysis

We analyzed the non-coding regions of the CT loci using PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/), a database to detect described regulatory elements in DNA sequences. We classify regulatory motifs by length. The rationale is that small regulatory motifs in introns (length < 7 bp) may be found only by luck, although they may not perform a regulatory function. Longer motifs (> 7 bp) might be better indicators of key regulatory elements in these genes.

In CT093, the sequenced non-coding region consists only of the 5' UTR. Not surprisingly, most of the inferred motifs are common *cis*-acting elements. The motifs in the 5' UTR of CT093 seem to be more conserved than the motifs in the non-coding regions of the other loci, which might be expected since it is a promoter region.

The non-coding region analyzed for CT166, which encodes a Ferredoxin-NADP reductase involved in photosynthesis, has four introns. We found light-responsive elements: 3-AF1 binding site, AE box (not in *S. chilense*), Box 4, and GT1 motif. One of these motifs, 3-AF1 binding site, contains SNPs that disrupt the motif in *S. chilense*, but not in *S. arcanum*.

In CT179, encoding a tonoplast intrinsic protein Δ-type involved in water transport, the non-coding region contains two introns. In *S. arcanum* and *S. habrochaites*, we find the MBS motif and a MYB binding site involved in drought-inducibility. The non-coding region features many indels disrupting the motifs. This is especially the case in *S. arcanum* where we cannot observe any conservation of motifs, the MBS motif also being affected by these indels.

The non-coding region of CT198 comprises three introns and part of the 3' UTR. CT198 is described as submergence induced protein 2-like. We found gibberellin-responsive elements (GARE motif – only in *S. habrochaites*; P box – not in *S. peruvianum*) which are exclusive to CT198. Gibberellins are known to be involved in dormancy, germination, sex determination, and flowering; an involvement in the water status of the plant has not been described. The motifs in *S. chilense* are more conserved than in *S. arcanum*, where almost every motif is disrupted by indels and/or SNPs.

At CT208, the non-coding region consists of three introns. No motif related to the fact that CT208 encodes an alcohol dehydrogenase can be found. The non-coding region of *S. arcanum* is more conserved than in *S. chilense*, as there are far fewer disruptions by indels.

The non-coding region of CT251 consists of four introns and the 3' UTR. Compared to the other loci, there are much fewer motifs. CT251 is described as putative At5g37260 gene. In *Arabidopsis thaliana*, this gene encodes a MYB transcription factor. The lack of motifs can

be explained by the fact that CT251 itself encodes a transcription factor. But it is also possible that the non-coding region contains motifs which are not described yet. Again, the *S. arcanum* non-coding region seems to be more conserved and we find more indels in *S. chilense*.

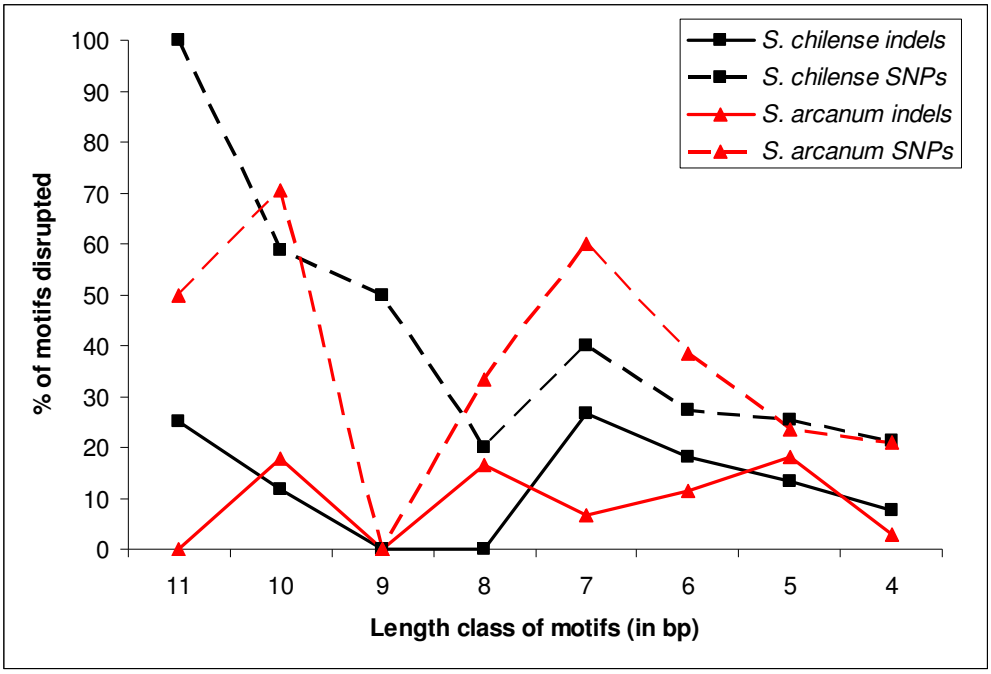

Figure A.S2 Percentages of regulatory motifs disrupted by SNPs (dashed lines) or indels (solid lines) in non-coding (intronic) regions of *S. chilense* (black) and *S. arcanum* (red), shown as a function of the length of the motif.

Appendix B: Supplementary Online Material Fischer *et al.* (2011)

Supporting Information Methods S1

We predicted the potential species distribution of *S. chilense* and *S. peruvianum* based on bioclimatic data. For this, we used the DIVA-GIS software (http://www.diva-gis.org/), applying the BIOCLIM (Bioclimate Prediction System) algorithm. BIOCLIM quantitatively describes the climatic environment inhabited by a species by defining a percentile distribution for each environmental variable considered. The BIOCLIM algorithm then compares the climatic characteristics of a geographic layer with the percentile distributions of variables inferred from species occurrences. With the BIOCLIM Classic option, the program determined six types of areas: "not suitable" areas for which one or more climate variables are outside the 0-100 percentile envelope; areas of "excellent" suitability within the 20-80 percentile envelope for each variable; areas of "very high" suitability within the 10-90 percentile envelope; areas of "high" suitability within the 5-95 percentile envelope; areas of "medium" suitability within the 2.5-97.5 percentile envelope and areas of "low" suitability for which at least one variable is at the boundaries of the 0-100 envelope (*i.e.* between 0-2.5% and 97.5-100%). In our study, we used data of collecting sites from the UC Davis database (http://tgrc.ucdavis.edu/index.aspx). Only accessions where the collection site was clearly defined (latitude, longitude) and which were according to the latest taxonomic definition (*Solanum*) were taken into account. This left us with 141 and 123 accessions for *S. chilense* and *S. peruvianum*, respectively. Ecological niche modeling (ENM) of the species was determined using current conditions (1950–2000) derived from the WorldClim global climate database (http://www.worldclim.org/), considering all 19 bioclimatic variables (at a 30 arc-second resolution, approximately 1km²): (1) annual mean temperature; (2) mean diurnal range, (3) isothermality, (4) temperature seasonality, (5) maximum temperature of warmest month, (6) minimum temperature of coldest month, (7) temperature annual range, (8) mean temperature of wettest quarter, (9) mean temperature of driest quarter, (10) mean temperature of warmest quarter, (11) mean temperature of coldest quarter, (12) annual precipitation, (13) precipitation of wettest month, (14) precipitation of driest month, (15) precipitation seasonality, (16) precipitation of wettest quarter, (17) precipitation of driest quarter, (18) precipitation of warmest quarter, and (19) precipitation of coldest quarter.

The Area Under Curve (AUC), referring to the Receiver Operation Characteristic curve (ROC - Pearce & Ferrier, 2000), is commonly applied to evaluate ecological niche models by

measuring its ability to discriminate locations where the species is present to those where it is absent (Hanley & McNeil, 1982). To calculate AUC, we divided the occurrences for each species into training (selecting 75% of the occurrence data sets randomly) and test points (including the remaining 25%) incorporated by "pseudo-absent" points by random sampling within South America, using DIVA-GIS. AUC values range from 0.5 for models no better than random and 1 indicating that the model can discriminate perfectly between presence and absence records. Models associated with AUC > 0.7 are considered as well performing (Fielding & Bell, 1997).

References

Fielding HA, Bell JF. 1997. *A review of methods for the assessment of prediction errors in conservation presence/absence models*. Cambridge, UK: Cambridge University Press.

Hanley JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating characteristics (ROC) curve. *Radiology* **143:** 29–36.

Pearce J, Ferrier S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling* **133:** 225–245.

Supporting Information – Tables

Table B.S1 Primer sequences and amplification conditions for PCR of the *Asr* genes

Locus	Type of primer	primer sequence 5' → 3'	annealing temperature
Asr1	forward	ATG GAG GAG GAG AAA CAC	54°C
	reverse	CAT CAC ACG AGA TTG GTA A	54°C
Asr2	forward	GTG ATT AGG TAA ACG CAA GG	56°C
	reverse	TTG ATT AGT AGT GGT GGT GG	56°C
Asr3	forward	ACT ATC CCA TAG CGT TG	54°C
	reverse	TTA TCT TGT GTT TGT C	54°C
Asr4	forward	GAA GAC ACC CCC ATT GAG AA	61°C
	reverse	GCT TCT TTC TTC TGA TGA TGT TC	61°C
Asr5	forward	TCC ACT TGA ATT GTT TGA TTA GG	59°C
	reverse	ATG GAA TGC AAA TCC ACC AG	59°C

Table B.S2 Numbers of alleles sequenced for each locus

Species/populations	*Asr1*	*Asr2*	*Asr3*	*Asr4*	*Asr5*
S. chilense	35	34	27	31	31
Quicacha	14	11	8	11	11
Moquegua	7	9	9	9	10
Tacna	14	14	10	11	10
S. peruvianum	30	30	30	29	28
Canta	10	10	9	10	10
Nazca	10	10	11	9	8
Tarapaca	10	10	10	10	10

Table B.S3 Numbers of site categories

Locus	total length[a]	coding[a]	synonymous sites[a,b]	non-synonymous sites[a,b]	non-coding[a]	intron length[a]	length of flanking region[a]
Asr1	1462	282	57.01	224.99	1180	998	182 (3')
Asr2	837	318	59.92	258.08	519	141	378 (5')
Asr3	1163	153	30.4	122.6	1010	632	378 (5')
Asr4	501	396	84.45	311.55	105	105	-
Asr4 repetitive region	108	108	23.38	84.62	-	-	-
Asr5	1357	180	33.95	146.05	1207	833	344 (5')

[a] excluding sites with gaps based on the total alignment

[b] calculated using the Nei & Gojobori (1986) method

Table B.S4 Nucleotide diversity of the *Asr* genes and the reference loci

π	*Asr1*	*Asr2*	*Asr3*	*Asr4*	*Asr4* repetitive region	*Asr5*	reference loci[a]
S. chilense[c]	0.025	0.020	0.037	0.018	0.034	0.032	0.012
Quicacha	0.024	0.016	0.033	0.009	0.015	0.019	0.010
Moquegua	0.024	0.022	0.020	0.023	0.040	0.026	0.011
Tacna	0.018	0.020	0.033	0.015	0.029	0.028	0.010
S. peruvianum[c]	0.024	0.022	0.040	0.019	0.035	0.039	0.014
Canta	0.024	0.020	0.041	0.019	0.036	0.037	0.014
Nazca	0.021	0.015	0.035	0.014	0.019	0.038	0.012
Tarapaca	0.022	0.022	0.039	0.021	0.028	0.035	0.013
θ_W							
S. chilense[c]	0.028	0.025	0.038	0.020	0.042	0.032	0.014
Quicacha	0.025	0.014	0.028	0.011	0.020	0.020	0.010
Moquegua	0.027	0.027	0.026	0.021	0.038	0.024	0.011
Tacna	0.015	0.022	0.031	0.022	0.042	0.024	0.009
S. peruvianum[c]	0.032	0.026	0.051	0.029	0.028	0.052	0.020
Canta	0.025	0.020	0.048	0.021	0.036	0.043	0.016
Nazca	0.020	0.014	0.036	0.015	0.018	0.035	0.012
Tarapaca	0.023	0.021	0.036	0.020	0.026	0.038	0.013
$\pi_{synonymous}$							
S. chilense[c]	0.032	0.047	0.040	0.035	0.058	0.032	0.026
Quicacha	0.019	0.024	0.034	0.019	0.022	0.022	0.022
Moquegua	0.039	0.057	0.016	0.030	0.046	0.010	0.026
Tacna	0.015	0.053	0.035	0.028	0.050	0.010	0.020
S. peruvianum[c]	0.044	0.091	0.041	0.040	0.067	0.049	0.031
Canta	0.046	0.095	0.037	0.049	0.069	0.049	0.030
Nazca	0.045	0.039	0.034	0.031	0.041	0.025	0.026
Tarapaca	0.032	0.087	0.043	0.025	0.034	0.061	0.030
$\pi_{non\text{-}synonymous}$							
S. chilense[c]	0.0007	0.0087	0.0159	0.0118	0.0273	0.0157	0.0031
Quicacha	0.0006	0.0068	0.0193	0.0060	0.0133	0.0066	0.0018
Moquegua	0.0000	0.0087	0.0073	0.0218	0.0382	0.0113	0.0026

Tacna	0.0020	0.0081	0.0089	0.0129	0.0234	0.0123	0.0026
S. peruvianum[c]	0.0025	0.0083	0.0119	0.0124	0.0262	0.0136	0.0033
Canta	0.0000	0.0020	0.0166	0.0108	0.0259	0.0174	0.0034
Nazca	0.0037	0.0008	0.0074	0.0091	0.0122	0.0080	0.0030
Tarapaca	0.0020	0.0153	0.0123	0.0170	0.0265	0.0121	0.0024
$\pi_{non\text{-}coding}$							
S. chilense[c]	0.032	0.024	0.041	0.024	-	0.036	0.013
Quicacha	0.032	0.024	0.036	0.012	-	0.020	0.012
Moquegua	0.031	0.027	0.022	0.020	-	0.029	0.010
Tacna	0.023	0.024	0.037	0.011	-	0.031	0.010
S. peruvianum[c]	0.033	0.021	0.045	0.023	-	0.045	0.016
Canta	0.034	0.020	0.045	0.018	-	0.041	0.017
Nazca	0.024	0.020	0.040	0.014	-	0.045	0.013
Tarapaca	0.028	0.017	0.044	0.033	-	0.044	0.013
π_{non}/π_{syn}[b]							
S. chilense[c]	0.022	0.185	0.398	0.337	0.471	0.491	0.126
Quicacha	0.032	0.283	0.568	0.316	0.605	0.300	0.100
Moquegua	0.000	0.153	0.456	0.727	0.830	1.130	0.121
Tacna	0.133	0.153	0.254	0.461	0.468	1.230	0.135
S. peruvianum[c]	0.057	0.091	0.290	0.310	0.391	0.278	0.114
Canta	0.000	0.021	0.449	0.220	0.375	0.355	0.136
Nazca	0.082	0.021	0.218	0.294	0.298	0.320	0.117
Tarapaca	0.063	0.176	0.286	0.680	0.779	0.198	0.077

[a] average over seven reference loci

[b] $\pi_{non\text{-}synonymous}/\pi_{synonymous}$

[c] pooled over three populations

Reference:

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

Supporting Information – Figures

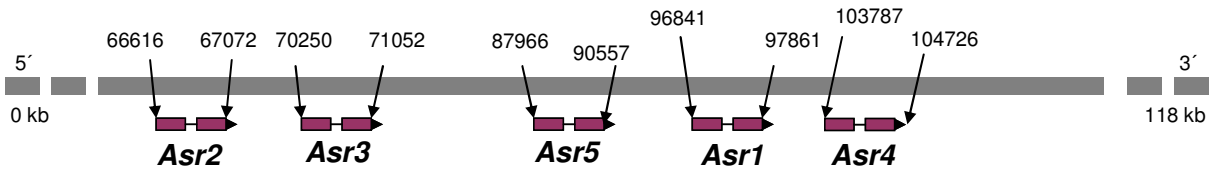

Figure B.S1 Positions of the *Asr* genes (purple) in the gene cluster relative to the BAC sequence of *S. lycopersicum* (grey line) from the NCBI server (accession number CU468249).

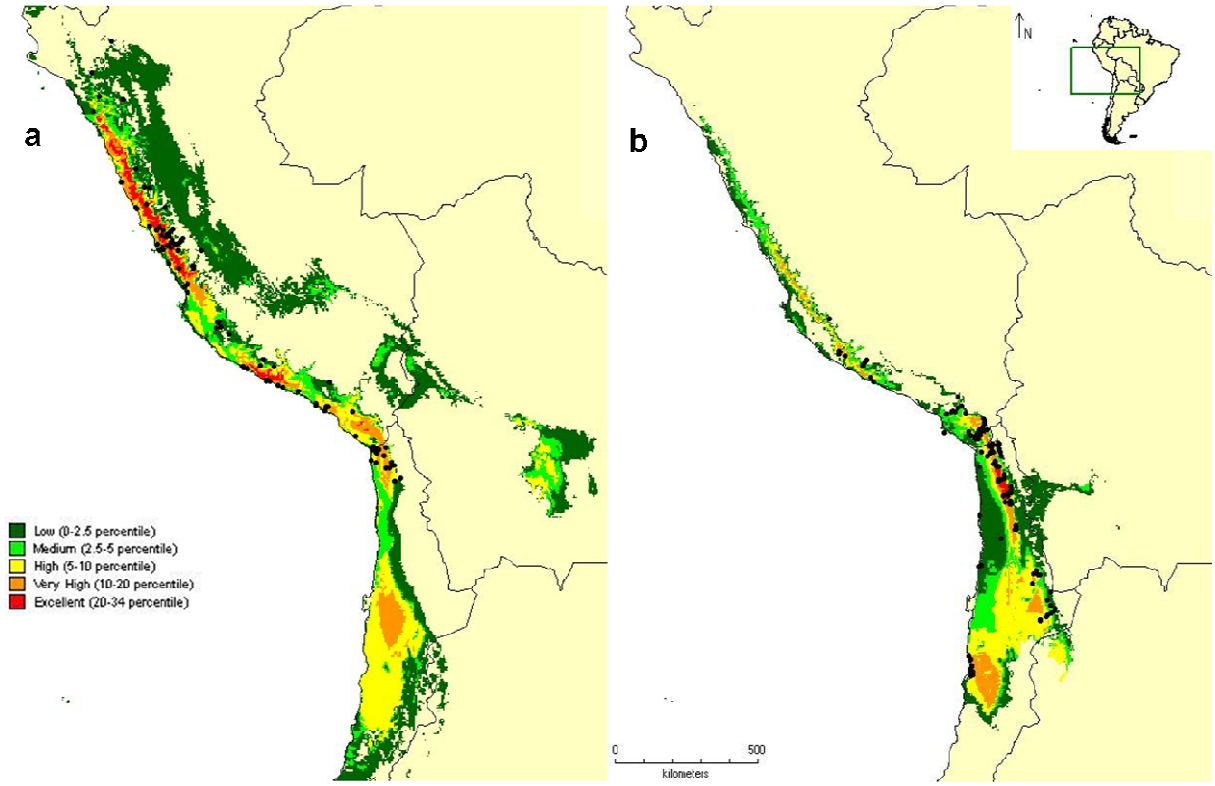

Figure B.S2 Potential distribution of (a) *S. peruvianum* and (b) *S. chilense* estimated from collecting sites (black dots) along the west coast of South America. The potential distribution of *S. peruvianum* is broader and further north (a) than that of *S. chilense* (b). Colours indicate the fitting zones from excellent (red) to marginal (dark green).

Appendix C: Supplementary Online Material Fischer *et al.* (in preperation)

Table C.S1 Primer sequences and amplification conditions for PCR of *pAsr2*, *pAsr4*, and *5'pLC*, *3'pLC*

Locus	Type of primer	primer sequence 5' → 3'	annealing temperature
pAsr2	forward	CTG GTG ACA ACT TCT ATG AGG	60°C
	reverse	AGA TCG CTG TGG TGC TTC	60°C
pAsr4	forward	TAG TAA ACG CGT ATT GAT GTG	61°C
	reverse	CCC AAG TTC TTC AAG ATG C	61°C
5'pLC	forward	CGT CAA ACC GAC CAC CTG	58°C
	reverse	GGG ATG AAC GAA AGA GAG AC	58°C
3'pLC	forward	GAA GTG GAA CAC AAG GAG GAG	58°C
	reverse	CAA GTT GTG ACG CCA TAC C	58°C

Table C.S2 Primer sequences and amplification conditions for the qPCR of the *Asr* genes, *pLC30-15*, and the reference genes

Locus	Type of primer	primer sequence 5' → 3'	annealing temperature
Asr1	forward	CAA ATC GGT AAA CTT GGC A	55°C
	reverse	TGG TGT CCC CCC TCA G	55°C
Asr2	forward	CAC CAT CAC CAT TTG TTC C	55°C
	reverse	CTT CTT TGC CTT GTG TTT CTC	55°C
Asr4	forward	TGG TGG AGG AGT TGG TG	65°C
	reverse	TTG TGT GCA TGC TCT GGA	65°C
pLC30-15	forward	AAG TGG AAC ACA AGG AGG AG	59°C
	reverse	ATC TTC TGT CCA TCC TCT CCA	59°C
CT189	forward	GGG TTC CGA AGA ATC TAT	58°C
	reverse	GTT GAA GAA TGT GGC GTG	58°C
TIP41	forward	ATG GAG TTT TTG AGT CTT CTG C	63°C
	reverse	GCT GCG TTT CTG GCT TAG	63°C

Table C.S3 Numbers of alleles sequenced for each locus

Species/populations	*pAsr2*	*pAsr4*	*5'pLC*	*3'pLC*
S. chilense	13	21	19	20
Quicacha	7	10	7	9
Tacna	6	11	12	11
S. peruvianum	19	16	19	19
Canta	11	9	10	9
Tarapaca	8	7	9	10

Table C.S4 Summary of function and sequences of motifs found at *pAsr2, pAsr4, 5'pLC,* and *3'pLC* using PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/)

Motif	Function	Sequence	found at
ABRE	*cis*-acting element involved in the abscisic acid responsiveness	CGTACGTGCA	*pAsr2, pAsr4, 5'pLC, 3'pLC*
ARE	*cis*-acting regulatory element essential for the anaerobic induction	TGGTTT	*pAsr2, pAsr4*
AuxRR-core	*cis*-acting regulatory element involved in auxin responsiveness	GGTCCAT	*pAsr4, 3'pLC*
CAT-box	*cis*-acting regulatory element related to meristem expression	GCCACT	*pAsr2, pAsr4*
CGTCA-motif	*cis*-acting regulatory element involved in methyl jasmonate responsiveness	CGTCA	*5'pLC, 3'pLC*
ERE	ethylene-responsive element	ATTTCAAA	*pAsr2, pAsr4*
GCN4-motif	*cis*-regulatory element involved in endosperm expression	CAAGCCA	*pAsr2*
HSE	*cis*-acting element involved in heat stress responsiveness	AAAAAATTTC	*3'pLC*
LTR	*cis*-acting element involved in low-temperature responsiveness	CCGAAA	*3'pLC*
MBS	MYB (transcription factor) binding site involved in drought-inducibility	C/TAACTG	*pAsr2, 5'pLC, 3'pLC*
Skn-1-motif	*cis*-acting regulatory element required for endosperm expression	GTCAT	*pAsr4, 5'pLC, 3'pLC*
TC-rich repeats	*cis*-acting element involved in defense and stress responsiveness	ATTTTCTTCA	*pAsr2, 5'pLC, 3'pLC*
TCA-element	*cis*-acting element involved in salicylic acid responsiveness	CCATCTTTTT	*pAsr2*

Appendix D – Primers

Table D.1 Complete list of PCR primers

Locus	Name	Sequence	Tm [C°]
Asr1	Asr1For	ATG GAG GAG GAG AAA CAC	54
	Asr1Rev	CAT CAC ACG AGA TTG GTA A	53
	Asr1F2	AAG GCG GAG GAG GG	50
	Asr1R2	ATT TAT TGT TAT TAT GCA CG	48
Asr2	Asr2For	GTG ATT AGG TAA ACG CAA GG	56
	Asr2Rev	TTG ATT AGT AGT GGT GGT GG	56
	Asr2F2	GTT CCA CCT AAT ATA GAA ATC C	57
	Asr2FSeq	AAC ATC CAC CAG TCA AAA CC	56
	Asr2RSeq	TCT ATC TTG TGC TTG TGT GC	56
Asr3	Asr3_aout1F	ACT ATC CCA TAG CGT TG	54
	Asr3_aout1R	TTA TCT TGT GTT TGT C	54
Asr4	Asr4_aout3R	GAA GAC ACC CCC ATT GAG AA	61
	Asr4_aout3F	CTT GTG TGC ATG CTC TGG AT	61
	Asr4_F1_ex2	GAA CCA AAA CTT CTG AGG ATT AC	60
	Asr4_R1_ex2	CTT TTG CTT CTT TCT TCT GAT G	60
	Asr4_R2_ex2	GCT TCT TTC TTC TGA TGA TGT TC	61
Asr5	Asr5_aout1F	TCC ACT TGA ATT GTT TGA TTA GG	59
	Asr5_aout1R	ATG GAA TGC AAA TCC ACC AG	59
Asr1 promoter	pAsr1For	ATC TCC TTC ACC TCT TTC C	55
	pAsr1Rev	ACA GTG CCA AGT TTA CCG	54
Asr2 promoter	pAsr2For	GCT CTT CTT GTT TTT TCT CC	60
	pAsr2Rev	AAA TGG TGA TGG TGT TGG	60
	pAsr2F2	CTG GTG ACA ACT TCT ATG AGG	60
	pAsr2R2	AGA TCG CTG TGG TGC TTC	60
Asr4 promoter	pAsr4For	GTC CAG GTG TGA TTT CCA AC	61
	pAsr4Rev	CCA CCT TCC TCA TTA CCA TAA C	61
	pAsr4F2	TAG TAA ACG CGT ATT GAT GTG	61
	pAsr4R2	CCC AAG TTC TTC AAG ATG C	61
	pAsr4F3	CAA GTC AAT TGG ATG TCC GTT	61
	pAsr4R3	GGC ACC AAC TCC TCC ACC	61

pLC30-15 5'UTR	pLC_UTRR1	CGT CAA ACC GAC CAC CTG	58
	pLC_UTRF1	GGG ATG AAC GAA AGA GAG AC	58
pLC30-15 3'UTR	pLC_Rev1	GAA GTG GAA CAC AAG GAG GAG	58
	pLC_For1	CAA GTT GTG ACG CCA TAC C	58
tail PCR	rand1	NGT CGA SWG ANA WGA A	NA
	tAsr2_1	AGA TTG GTG AAC TTG GT	NA
	tAsr2_2	AAC ACA AGG CAA AGA AG	NA
	tAsr2_3	CAT TCC ATG AAC ATC ATC	NA
other primers	M13F	GTT GTA AAA CGA CGG CCA	55
	M13R	TCA CAC AGG AAA CAG CTA TGA	55
	oligo(dT20)	TTT TTT TTT TTT TTT TTT TTV	50

Table D.2 Complete list of qPCR primers

Locus	primer	Sequence	Tm [C°]
Asr1	rtAsr1_F2	CAA ATC GGT AAA CTT GGC A	55
	rtAst1_R2	TGG TGT CCC CCC TCA G	55
Asr2	rtAsr2_F3	CAC CAT CAC CAT TTG TTC C	55
	rtAsr2_R3	CTT CTT TGC CTT GTG TTT CTC	55
Asr4	rtAsr4_F2	TGG TGG AGG AGT TGG TG	65
	rtAsr4_R2	TTG TGT GCA TGC TCT GGA	65
pLC	pLCRTR1	AAG TGG AAC ACA AGG AGG AG	59
	pLCqF2	CTG CTT TCG ATT CTT CCT TG	59
CT189	CT189qF	GGG TTC CGA AGA ATC TAT	58
	CT189qR	GTT GAA GAA TGT GGC GTG	58
TIP4I	rtTIP4I_F	ATG GAG TTT TTG AGT CTT CTG C	63
	rtTIP4I_R	GCT GCG TTT CTG GCT TAG	63

Appendix E: Protocols, buffers, and nutrition media

DNA Isolation

DNeasy Plant Mini Kit (*Qiagen, Hilden, Germany*)

1. Grind plant material in a 1.5 ml Eppendorf tube in liquid nitrogen
2. Add 400 µl Buffer AP1 and 4 µl RNase (fridge!) → vortex until no more clumps are visible
3. Lysis: incubate for 15 – 20 minutes at 65°C, mix 2 -3 times by shaking
4. Protein precipitation: add 130 µl Buffer AP2 and mix gently → incubate for 5 min on ice
5. Centrifuge for 5 min at maximum speed (~ 13,200 rpm)
6. Apply lysate to QIAshredder Mini Spin Column placed on a 2 ml collection tube → centrifuge for 2 min at maximum speed
7. Transfer flow-through to a new 1.5 ml Eppendorf tube, do not disturb the pallet
8. Add 1.5 x volume of Buffer AP3/E and mix by pipetting
9. Apply 650 µl of the mixture to a DNeasy Mini Spin Column placed on a 2 ml collection tube → centrifuge for 1 min at 8,000 rpm, discard flow-through (DNA is in column)
10. Repeat the previous step with remaining sample
11. Washing DNA: Place column on a new 2 ml collection tube, add 500 µl Buffer AW → centrifuge for 1 min at 8,000 rpm, discard flow-through
12. Add 500 µl Buffer AW → centrifuge 2 min at maximum speed
13. Elute DNA: Place column on a 1.5 ml Eppendorf tube, put 50 µl of Buffer AE directly onto the membrane → incubate for 5 minutes at room temperature, centrifuge for 1 min at 8,000 rpm
14. Repeat step 13
15. Store DNA at - 20°C

Polymerase Chain Reaction (PCR)

High Fidelity Phusion Polymerase (*Finnzymes, Espoo, Finland*)

12.4 µl	HPLC H_2O (*Sigma-Aldrich, Steinheim, Germany*)
4.0 µl	5X HF Buffer
0.4 µl	dNTPs 10mM (*invitrogen, Carlsbad, CA, USA*)
1.0 µl	Primer Forward 10µm
1.0 µl	Primer Reverse 10 µm
0.2 µl	Phusion polymerase
1.0 µl	genomic DNA

Thermorcycler (*BioRad, Hercules, CA, USA*)

1 cycle	98°C	30 sec
30 cycles	98°C	5 sec
	X* °C	20 sec
	72°C	90 sec
1 cycle	72°C	8 min
	8°C	∞

* see Appendix D for amplification conditions

Touchdown PCR – Thermocycler protocol

1 cycle	98°C	30 sec				
10 cycles	98°C	5 sec	- 1 °C	20 cycles	98°C	5 sec
	Y^+ °C	20 sec	→→→→→		Z^+ °C	20 sec
	72°C	90 sec	after 1 cycle		72°C	90 sec
1 cycle	72°C	8 min				
	8°C	∞				

$^+$ Y = +5 °C optimal T_m, Z = -5 °C optimal T_m according to Appendix D

Tail PCR

12.9 µl	water
4.0 µl	Buffer 5xHF
0.4 µl	dNTPs 10mM
1.0 µl	Random primer
1.0 µl	Primer I – III (move primer from middle towards end of the fragment)
0.2 µl	Phusion polymerase
0.5 µl	DNA (first genomic DNA, then 1:50 dilution of PCR product of previous reaction)

Thermocycler

Reaction I			Reaction II			Reaction III		
1x	98°C	3 min	1x	98°C	60 sec	1x	98°C	15 sec
5x	98°C	25 sec	12x	98°C	25 sec	30x	98°C	15 sec
	62°C	60 sec		64°C	30 sec		37°C	30 sec
	72°C	120 sec		72°C	120 sec		72°C	90 sec
1x	98°C	25 sec		98°C	25 sec	1x	72°C	8 min
	25°C	30 sec		64°C	30 sec		8°C	∞
	72°C	120 sec		72°C	120 sec			
15x	98°C	25 sec		98°C	25 sec			
	67°C	30 sec		44°C	30 sec			
	72°C	120 sec		72°C	120 sec			
	98°C	25 sec	1x	72°C	8 min			
	67°C	30 sec		8°C	∞			
	72°C	120 sec						
	98°C	25 sec						
	44°C	30 sec						
	72°C	120 sec						
1x	72°C	5 min						
	8°C	∞						

Gel electrophoresis

- 1 g agarose (*Serva, Heidelberg, Germany*) dissolved in 100 ml 1 x TAE buffer
- add 2.5 µl ethidium bromide (*Roth, Karlsruhe, Germany*) to 100 ml gel
- 3 µl PCR product + 2 µl Loading Dye were applied on the gel

ExoSAP-IT (*GE Healthcare, Freiburg, Germany*)

- Add 0.5 µl ExoSAP-IT to 10 µl PCR product
- Incubate for 30 min at 37°C
- Heat inactivation of ExoSAP-IT: 15 min at 80°C

Cloning

Ligation

StrataClone Blunt (*Agilent Technologies - Farnell, Oberhaching, Germany*)

2.5 µl H_2O (HPLC)

1.5 µl Salt Solution

0.5 µl Vector Mix

1.5 µl PCR product → incubate 30 min at RT

Transformation

3.0 µl Ligation

25.0 µl *E. coli* Cell Solution (thermo competent) → incubate 20 min on ice

Heatshock:

sec at 42°C

add 250 µl SOB media → incubate 70 min at 37°C shaking at 200 rpm

plate out on TB agar plates containing 40 µl X-Gal (*Peqlab Biotechnologie, Erlangen, Germany*)

→ incubate over night at 37°C

Colony PCR

Taq Polymerase

(*invitrogen, Carlsbad, CA, USA*)

15.4 µl	H_2O (HPLC)	1 cycle	95°C	4 min
2.0 µl	Buffer	30 cycles	95°C	30 sec
0.8 µl	$MgCl_2$		55 °C	45 sec
0.8 µl	dNTPs 10mM		72°C	90 sec
0.4 µl	Primer M13F	1 cycle	72°C	7 min
0.4 µl	Primer M13R		8°C	∞
0.2 µl	Taq polymerase			

Colony picked with a toothpick

Plasmid preparation

QIAprep Miniprep Kit (*Qiagen, Hilden, Germany*)

1. Grow colonies over night in 3 ml TB medium containing Kanamycin
2. Transfer half of the cell culture in a 1.5 ml Eppendorf tube → centrifuge for 1 min at 12,000 rpm, discard supernatant by decanting
3. Repeat previous step with remaining sample and remove all the medium
4. Resuspent pallet in 250 µl buffer P1 (containing RNase – fridge!) → vortex until the pallet is completely resupended
5. Lysis: add 350 µl Buffer P2 and mix gently by inverting the tube 5 – 6 times
6. Protein precipitation: add Buffer N3 and mix gently by inverting the tube 5 – 6 times
7. Centrifuge for 10 min at maximum speed (13,200 rpm)
8. Apply supernatant to a QIAprep Spin column by decanting → centrifuge for 1 min at maximum speed and discard flow-through
9. Wash column with 500 µl Buffer PB → centrifuge for 1 min at maximum speed and discard flow-through
10. Wash column with 750 µl Buffer PE → centrifuge for 1 min at maximum speed and discard flow-through
11. Dry membrane: centrifuge column for 1 min at maximum speed
12. Place column on a 1.5 ml Eppendorf tube
13. Elute DNA: place 30 µl H_2O (70°C) on the membrane and incubate for 5 min at room temperature
14. Centrifuge 1 min at maximum speed

Sequencing

ABI PRISM BigDye Terminator 1.1 (*Applied Biosystems, Carlsbad, CA, USA*)

5.0 µl	H_2O (HPLC)
1.0 µl	SeqBuffer
2.0 µl	SeqMix
1.0 µl	PCR product

Thermocycler

1 cycle	96°C	1 min
40 cycles	96°C	10 sec
	50 °C	15 sec
	60°C	4 min
1 cycle	8°C	∞

RNA extraction

RNeasy Plant Mini Kit (*Qiagen, Hilden, Germany*)

1. Take tissue sample and immediately freeze it in liquid nitrogen! Rinse all equipment with 0.1 M NaOH (*Merck, Whitehouse Station, NJ, USA*), 1 mM EDTA (*Sigma-Aldrich, Steinheim, Germany*), RNAse free water!!! Grind plant tissue.
2. Weight 30 - 70 mg tissue power into 1.5 ml tube. Don't let it thaw!!! Store the remaining powder at -80°C.
3. Add buffer 450 µl RLT buffer containing β-Mercaptoethaonl (*Merck, Whitehouse Station, NJ, USA*) and vortex vigorously. Incubate at 56°C for 3 min.
4. Transfer lysate to (lilac) QIAshredder column placed on a 2 ml collection tube. Spin 2 min at full speed. Transfer supernatant in a new tube without disturbing the pellet.
5. Add 0.5 volume (~ 200 µl) 96% ethanol (*Roth, Karlsruhe, Germany*) and mix by pipetting.
6. Transfer lysate to an (pink) RNeasy spin column placed on a 2 ml collection tube. Spin for 15 sec at 10,000 rpm. Discard flow-through.
7. Add 350 µl RW1 buffer to column (placed on same collection tube). Spin 15 sec at 10,000 rpm. Discard flow-through. When removing the column, make sure the column has no contact with the flow-through!
8. Mix 1x DNase buffer with DNaseI (1 µl in 100 µl buffer) and put 80 µl on the filter.
9. Add 350 µl RW1 buffer to column (placed on same collection tube). Spin 15 sec at 10,000 rpm. Discard flow-through.
10. Add 500 µl RPE buffer to the column. Spin for 15 sec at 10,000 rpm. Discard flow-through.
11. Add 500 µl RPE buffer to the column. Spin for 2 min at 10,000 rpm.
12. Place the column on a new collection tube and spin for 1 min at full speed.
13. Place the column on a 1.5 ml collection tube. Add 30 µl RNase free water to the membrane. Spin for 1 min at 10,000 rpm.
14. Repeat step 11

cDNA synthesis

SuperScriptIII reverse transcriptase (*invitrogen, Carlsbad, CA, USA*)

1. Add 2 µl oligo$(dT)_{20}$ primers (50 µM)
   + X µl RNA (2 µg – or 22 µl RNA)
   + 2 µl 10 mM dNTP's
   + X µl sterile water (add up to 26 µl total)
2. Heat to 65°C for 5 min, incubate on ice for at least 1 min
3. Add to solution:
   8 µl 5X First-Strand buffer
   2 µl DTT (0.1 M)
   2 µl RNaseOUT (*invitrogen, Carlsbad, CA, USA*)
   2 µl SuperScript III RT
4. incubate at 50°C for 60 min
5. inactivate reaction at 70°C for 15 min
6. add 2 µl RNase (*New England Biolabs, Ipswich, MA, USA*) and incubate for 15 min at RT

qPCR

iQ SYBR Green
(*BioRad, Hercules, CA, USA*)

2.0 µl	H_2O (HPLC)
5.0 µl	SYBR green
0.5 µl	Primer Forward 10µm
0.5 µl	Primer Reverse 10 µm
2.0 µl	cDNA

CFX96 Thermorcycler
(*BioRad, Hercules, CA, USA*)

1 cycle	95°C	3 min
40 cycles	95°C	10 sec
	X* °C	30 sec
	→ quantification	

* see Appendix D for amplification conditions

50x TAE buffer

- 242 g TRIS (*Sigma-Aldrich, Steinheim, Germany*)
- 100 ml EDTA [0.5 M] (*Sigma-Aldrich, Steinheim, Germany*)
- 57.1 ml 100 % acetic acid (*Merck, Whitehouse Station, NJ, USA*)
- Ad. 1 l with dist. H_2O

SOB medium

30.7 g SOB medium (*Roth, Karlsruhe, Germany*); Ad 1 l with dist. H_2O

TB (Terrific Broth) medium

- 12g Tryptone (*Roth, Karlsruhe, Germany*)
- 24 g Yeast extract (*Roth, Karlsruhe, Germany*)
- 4 ml 99 % glycerol (*Serva, Heidelberg, Germany*)
- Ad. 900 ml with dist. H_2O

TB agar

- 12g Tryptone (*Roth, Karlsruhe, Germany*)
- 24 g Yeast extract (*Roth, Karlsruhe, Germany*)
- Agar Agar (*Roth, Karlsruhe, Germany*)
- 4 ml 99 % glycerol (*Serva, Heidelberg, Germany*)
- Ad. 900 ml with dist. H_2O

Salt solution

23.1 g KH_2PO_4 [0.17 M] + 125.4 g K_2HPO_4 [0.72 M] (both *Roth, Karlsruhe, Germany*); Ad 1 l with dist. H_2O

Before using TB agar and medium, 100ml salt solution and 1 ml Kanamycin [50 µg/ml] (*Sigma-Aldrich, Steinheim, Germany*) were added.

All solutions and media were autoclaved at 15 psi for 20 minutes.

Acknowledgements

Als erstes möchte ich meinem Supervisor Prof. Stephan für seine Unterstützung, die exzellente Betreuung und für die Chance danken, meine Doktorarbeit in seiner Gruppe durchzuführen. I am also very grateful to Létizia Camus-Kulandaivelu for motivation, fruitful discussion, and her hospitality whenever I visited her in Montpellier. A big Thank you to Mamadou Mboup for discussion and the effort he made to make me a better scientist. And, of course, for the French lessons! I'm also thankful to François Allal and Kim Steige who helped a lot improving our publications. All of you were very supportive and helped me develop my scientific skills.

Ein Riesen-Dankeschön geht natürlich auch an Hilde Lainer, die mich nicht nur in der Laborarbeit, sondern auch in mental düsteren Zeiten unterstützt hat! Vielen lieben Dank auch an Simone Lange, mir nicht nur sehr geduldig die meiste Laborarbeit beigebracht hat (auch wenn mal DNase in der colony PCR war), sondern über die Jahre auch eine sehr gute Freundin geworden ist!

I also owe a lot to the Tomato group. Als erstes ist natürlich Anja zu nennen, mit der ich fachlich diskutieren konnte aber auch privat sehr viel Spaß hatte (unvergessen: Ägypten!). I am also very grateful to Aurélien Tellier for the most valuable discussions (at the institute or at a bar). I also want to thank Laura Rose and Thomas Städler who always answered any tomato-related question I had. A big Thank you also to Carlos Merino for his friendship and discussions (although we rarely had the same opinion), Katharina Böndel for a lot of helpful input, Hui Xia for a fun time in the lab, Astrid Stück for a nice match of paintball, and Lukasz Grzeskowiak for chats in the office.

Thanks a lot also to the remaining Stephan group! Als erstes muss ich Susi Voigt danken, die mir jede blöde Frage über qPCR gerne (?) beantwortet hat – und mit der ich natürlich auch sonst viel Spaß hatte (ich sage nur: Zuckerrüben). Außerdem ein riesiges Dankeschön an Anne Werzner mit der ich jedes Thema (wissenschaftlich oder nicht) diskutieren konnte. Vielen Dank auch an (meinen Leidensgenossen) Robert Piskol, der mir immer mit Rat und Tat zu Seite stand! Ich danke auch Stephan Hutter für den tollen PERL-Kurs. Many thanks also to Francesco Paparazzo, Stefan Laurent, Pavlos Pavlidis, Ricardo Wilches, Meike Wittmann, Nico Svetec, Ann Arunyawat, Pablo Duchén, and Daniel Živković for their support and the fun time!

I am also grateful to the members of the Parsch group. Zuerst danke ich natürlich Claus Kemkemer für seine Freundschaft und dass ich ihn alles fragen konnte, da er ja alles wusste

(und bestimmt auch immer noch weiß). Vielen Dank an Miri Linnenbrink („Ich mag eigentlich keine Chips“), der ich die blödesten Fragen stellen konnte und die mir viel über Mäuse beigebracht hat. Danke auch an Lena Müller (mit dem langweiligen Namen), die als Mit-Doktorandin kurz vor der Abgabe die letzen Wochen mit mir gelitten hat. Ich danke auch Sonja Grath, die mir immer wertvolle Ratschläge gegeben hat und mir immer hilfreich zur Seite stand. Vielen Dank auch an Dr. Winfried Hense für die unterhaltsamen Diskussionen zu allen erdenklichen Themen (v.a. Harry Potter) und das Organisieren der Lotto-Tip-Gemeinschaft. I am also thankful to Sarah Peter, Ana Catalán, John Baines, and Rayna Stamboliyska for their help and support.

Thanks also to John Parsch (not only for proofreading my manuscripts) and Joachim Hermisson – their lectures made me study evolution in the first place! Vielen Dank auch an Lisha Naduvilezhath, die mir bei jeglichen Computer-Problemen zur Seite stand. Großer Dank gebührt auch Gisela Brinkmann für all die Unterstützung beim Sequenzieren (auch wenn es sie oft Nerven gekostet hat), Ana Vrljic für Hilfe in jeder erdenklichen Lebenslage (von Nähen bis Lebensmittelnotstand), Anne Steincke und Hedwig Gebhart, die ich bei Problemen im Labor immer um Rat fragen konnte und Ingrid Kroiß für die Hilfe in allen administrativen Belangen (v.a. dem HiWi-Vertrag)! Danke auch an Dirk Metzler und Martin Hutzenthaler, die mir geduldig Statistik und/oder R erklärt haben. And last but not least a big thank you to Pleuni Pennings who always supported my scientific career in so many ways.

Ich danke außerdem meinen Freunden, dank denen ich in den letzen Jahren (einigermaßen) bei Verstand geblieben bin. Zuletzt möchte ich noch meiner Familie danken, die mich immer unterstützt hat. Vor allem meinen Eltern, die immer für mich da waren, auch wenn es mal finster aussah!