TUM School of Life Sciences

Evolution of gene networks involved in adaptation to stressful environments in Solanum chilense

Kai Wei

Vollständiger Abdruck der von der
TUM School of Life Sciences
der Technischen Universität München zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat)
genehmigten Dissertation.

Vorsitzende: Prof. Donna Ankerst, Ph.D.

Prüfende/-r der Dissertation:

1. Prof. Dr. Aurélien Tellier

2. Prof. Dr. John Parsch

Die Dissertation wurde am 02.11.2021 bei der Technischen Universität München
eingereicht und durch die
TUM School of Life Sciences am 13.01.2022 angenommen.

Summary

Local adaptation is one of the main driving forces of species evolution underlying expansion and colonization of new habitats. Understanding the genetic bases of these adaptive processes is crucial to predict a species' evolutionary potential. The wild tomato species *Solanum chilense* (*S. chilense*) is found in habitats from sea level up to 3500 m of altitude surrounding the Atacama desert at the southern edge of the entire tomato clade distribution range. *S. chilense* populations are challenged by multiple environmental stresses, such as drought, salt and low temperature across the different populations. This species is thus an ideal model for the discovery of mechanisms underpinning genomic adaptation. This thesis explores the genetic basis of polygenic adaptation to changing climatic conditions using whole genome as well as comparative transcriptomic data. The main results are as follows.

First, we infer the past demographic history of the species, and search for genes under strong positive selection. We then correlate the demographic history, allele frequencies in space and time, the age of selection events and the reconstructed historical ecological distribution of the species over five main climatic periods spanning 150,000 years. We find evidence for several selective sweeps targeting regulatory networks involved in root hair development in low altitude, and response to photoperiod and vernalization in high altitude populations. These sweeps occur in a concerted fashion in a given regulatory gene network and only at particular time periods, thereby underpinning temporal local adaptation. These genes under positive selection provide subsequently the basis for spatial local adaptation to novel habitats when new ecological niches become available. Second, we identified two conserved networks related to the cell cycle and fundamental metabolic processes responding to drought stress, respectively. These drought-responsive genes mainly originated between the early to middle stages of the tree of life, some genes evolved during the evolution of early flowering plants. Furthermore, few genes are specific to *S. chilense*,

and represent recent evolution and adaptation to drought stress in this species. Most genes underlying transcriptomic response to drought appear evolutionary younger and more variable in their expresison than genes of the transcriptome under non-stress conditions. The network related to metabolic response shows a younger evolutionary age and a stronger gene family expansion than the cell-cycle response network. Stronger positive selection at metabolic genes indicates the rewiring of this network in *S. chilense* and a higher evolutionary potential compared to other more conserved drought stress gene networks.

Our results provide an insight into polygenic adaptation to changing environmental conditions by combining genomic and transcriptomic evolutionary analyses in the wild tomato *S. chilense*. The results of this thesis are a first step to reveal the underlying genetic mechanisms underlying adaptation to arid habitats in plants.

Zusammenfassung

Die lokale Anpassung ist eine der wichtigsten Triebkräfte der Evolution von Arten, die der Ausbreitung und Besiedlung neuer Lebensräume zugrunde liegt. Das Verständnis der genetischen Grundlagen dieser Anpassungsprozesse ist entscheidend für die Vorhersage des evolutionären Potenzials einer Art. Die wilde Tomatenart *Solanum chilense* (*S. chilense*) kommt in Lebensräumen von einer Höhe bis zu 3500 m rund um die Atacama-Wüste am südlichen Rand des gesamten Verbreitungsgebiets dieser Tomatenklade vor. Die Populationen von *S. chilense* werden durch verschiedene Umweltstressfaktoren wie Trockenheit, Salz und niedrige Temperaturen in den verschiedenen Populationen herausgefordert. Diese Art ist daher ein ideales Modell für die Entdeckung von Mechanismen, die der genomischen Anpassung zugrunde liegen. In dieser Arbeit werden die genetischen Grundlagen der polygenen Anpassung an sich ändernde klimatische Bedingungen anhand von Daten des Genoms und aus der Transkriptomik untersucht. Die wichtigsten Ergebnisse sind wie folgt:

Zunächst schließen wir auf die vergangene demografische Geschichte der Art und suchen nach Genen, die einer starken positiven Selektion unterliegen. Anschließend setzen wir die demografische Geschichte, die Allelhäufigkeiten in Raum und Zeit, das Alter der Selektionsereignisse und die rekonstruierte historische ökologische Verteilung der Art über fünf Hauptklimaperioden, die sich über 150.000 Jahre erstrecken, in Beziehung. Wir finden Belege für mehrere Selektionsvorgänge, die sich auf regulatorische Netzwerke auswirken, welche an der Entwicklung der Wurzelhaare in niedrigen Höhenlagen-Populationen und an der Reaktion auf Photoperiode und Vernalisation in hoch gelegenen Populationen beteiligt sind. Diese Selektionsvorgänge treten in einem bestimmten regulatorischen Gennetzwerk und nur in bestimmten Zeiträumen auf, wodurch die zeitliche lokale Anpassung unterstützt wird. Diese Gene, die einer positiven Selektion unterliegen, bilden anschließend die

Grundlage für die räumliche lokale Anpassung an neue Lebensräume, wenn neue ökologische Nischen verfügbar werden. Zweitens haben wir zwei konservierte Netzwerke identifiziert, die mit dem Zellzyklus bzw. mit grundlegenden Stoffwechselprozessen zusammenhängen, die auf Trockenstress reagieren. Diese Gene, die auf Trockenheit reagieren, entstanden hauptsächlich in den frühen bis mittleren Stadien der Phylogenie, andere Gene entwickelten sich während der Evolution der frühen blühenden Pflanzen. Darüber hinaus sind nur wenige Gene spezifisch für *S. chilense* und stehen für die jüngste Evolution und Anpassung an Trockenstress bei dieser Art. Die meisten Gene, die der transkriptomischen Reaktion auf Trockenheit zugrunde liegen, scheinen evolutionär jünger und in ihrer Ausprägung variabler zu sein als die Gene des Transkriptoms unter Nicht-Stressbedingungen. Das Netzwerk, das mit der Stoffwechselreaktion zusammenhängt, zeigt ein jüngeres evolutionäres Alter und eine stärkere Erweiterung der Genfamilien als das Netzwerk der Zellzyklusreaktion. Eine stärkere positive Selektion bei den Stoffwechselgenen deutet auf eine Neuverknüpfung dieses Netzwerks in *S. chilense* und ein höheres evolutionäres Potenzial im Vergleich zu anderen, eher konservierten Gennetzwerken bei Trockenstress hin.

Unsere Ergebnisse geben einen Einblick in die polygene Anpassung an veränderte Umweltbedingungen durch die Kombination von genomischen und transkriptomischen Evolutionsanalysen in der Wildtomate *S. chilense*. Die Ergebnisse dieser Arbeit sind ein erster Schritt zur Aufdeckung der zugrunde liegenden genetischen Mechanismen, die der Anpassung an trockene Lebensräume bei Pflanzen zugrunde liegen.

List of Figures

List of Tables

Supplementary information

All supplementary information can be obtained in my github account (https://github.com/weikai-320722/Schil_30WGS).

Supplementary figures.....Supplementary_figures.pdf

Supplementary tables.....Supplementary_tables.pdf

Supplementary data for chapter 3.....Data S1

Supplementary data for chapter 4.....Data S2

Supplementary script for simulation.....run_scrm.sh

Contents

1 Introduction

1.1 Adaptive evolution

Adaptation is the most basic ability required by organisms to survive and thrive in an environment which is diverse in space and variable in time. Adaptive evolution underlies the formation, expansion and contraction of species and colonization of new habitats. Since Darwin, the mechanism of adaptive evolution has been a core scientific problem of evolutionary biology. In the early studies, biologists mainly study adaptation of organisms to the environment via changes in traits, including morphological and physiological adaptations. With the development of molecular markers and sequencing technologies, especially next-generation sequencing (NGS) technology, researchers are allowed to investigate how natural populations and communities respond to novel environments and to obtain in-depth insights about local adaptation and demographic history directly at the molecular level. It is possible to examine the genetic structure of the whole genome in organisms with high mapping accuracy and resolution, which greatly improves the working efficiency and provides wider and deeper insights. A clade of interest for evolutionary genomics studies is the tomato (*Solanum*) clade which contains many species found originally in various habitats in South and Central-America. Moreover, the complete genomes of cultivated tomato species (*Solanum lycopersicum*) and few wild tomato species (*Solanum lycopersicoides*, *Solanum pennellii* and *Solanum pimpinellifolium*) have been assembled [1-4]. These basic resources are now essential beneficial for the breakthrough to understanding tomato genetics and genetics of adaptation in order to improve the development of new breeds of cultivated tomato.

1.1.1 Genomic adaptation

Natural selection is the process through which populations of living organisms adapt and change to the changes of the environment. [5, 6]. It is one of the main forces driving evolution when species expand their range of habitat and colonize novel habitats [7]. Individuals are naturally variable in a population of species, meaning that

they are always different at the genetic and possibly phenotypic levels. In a neutral setting, this variation has no influence on the fitness of the individuals. Regarding natural selection, this means that some individuals with variation show traits better suited to the external environment than others without this variation [7, 8]. These adaptive traits in some individuals give them some advantages to survive and reproduce to novel habitats. These individuals are more likely to pass the adaptive traits on to their offspring. Over time, these advantageous traits become more common in the population (up to reach fixation) [5]. Through this process of natural selection, favorable traits are transmitted and fixed through generations. Over long period of time, natural selection will also lead to population divergence and speciation, whereby one species gives rise to one or several new and distinctly different species with specific adaptations to novel habitats. This process is a key determinant explaining the diversity of life on Earth [9].

Mutations are changes in the site and structure of the DNA, the molecule in which genes are encoded as well as their regulation. The mutation of genes is an important genetic and evolutionary source within a population, as genes and coding regions underlie the proteome and metabolome of each cell/organ/organism [10, 11]. Mutations can occur randomly for example, when cell replications make an or several errors while copying DNA, or be induced as a result of exposure to some stresses in the environment (harmful chemicals, radiation or extreme climatic conditions). Mutations can be deleterious (harmful to the fitness), neutral, or advantageous (helpful increasing fitness). Therefore, natural selection is one of the basic processes (force) of evolution, along with mutation, migration, and genetic drift [12-14].

In population genetics, natural selection mainly includes three different processes, which cause a change in allele frequency. Positive selection is the process by which new advantageous genetic variants sweep through a population and the main mechanism that Darwin envisioned as giving rise to evolution [15, 16]. Negative selection (also purifying selection) will eliminate deleterious mutations in the

population [17]. Balancing selection occurs when multiple alleles are maintained in a population, which can result in their preservation over long evolutionary periods [18]. A characteristic signature of this long-term balancing selection is an excess number of intermediate frequency polymorphisms near the balanced variant.

In this thesis, we focus on genomic adaptation driven by positive selection. Positive Darwinian selection is a key process driving macro and micro-evolution and underlying local adaptation when species expand their range and colonize novel habitats. Local adaptation encompasses thus the spatial component of adaptation to a new habitat, but also a temporal component when climatic/abiotic conditions vary in time after colonization [19]. The latter is understood here at the scale of hundreds up to thousands of years during which climatic conditions are changing due to glacial/inter-glacial periods or medium to long-term changes in abiotic and biotic environmental conditions (the so-called moving environmental optimum). Population genetics theory investigates the characteristics and polymorphism footprints of local adaptation while separating the two processes: spatial adaptation to a local optimum based on Fisher's geometric model of adaptation [20] and temporal adaptation to a moving optimum [19, 21, 22]. Under both models and large enough population sizes, it is typically expected that bigger steps of adaptation occur first at sites with strong selective coefficients, possibly generating selective sweeps [20-22]. Selective sweeps are polymorphism patterns (footprints) in the genome due to the rapid fixation of advantageous alleles and the associated hitchhiking effect [23, 24]. However, there is so far no complete theory investigating the more realistic scenario combining both local spatial adaptation to newly colonized habitats and further temporal adaptation to temporally variable climatic conditions and the expected patterns of selection.

We identify therefore a number of unanswered theoretical and empirical questions regarding the joint effect of spatial and temporal local adaptation.

How many genes underlie adaptation to new habitats and/or changing habitat in time?

How many genes underpinning temporal or spatial adaptation show strong selection (selective sweeps)?

Are the same genes involved in both adaptive processes?

Do these genes belong to one or few gene pathways/regulatory networks?

What is the timing of adaptation and can selection at a given gene be linked to past changes in climatic conditions?

This lack of knowledge regarding the underpinnings of local adaptation represents a major bottleneck to predict/assess the species potential for adaptability and resilience in the face of the current global climatic change. Indeed, despite numerous studies highlighting genes or even SNPs underlying spatial climatic adaptation [25], future predictions of the strength of selection and change in allele frequencies in response to the predicted climate change trends remain fairly speculative.

To date, numerous studies uncovering the genetic bases of spatial adaptation have been conducted in few model organisms such as humans, invertebrates (*Caenorhabditis elegans*, *Drosophila melanogaster*) or plants (*Arabidopsis spp.*, *Arabis sp.*, *Capsella spp.*). Using genome scan methods based on the selective sweep/hitchhiking model [23, 24], Genome Wide Association studies (GWAs) and functional studies, the crucial role of changes in genes/proteins (coding regions) and/or gene expression and the relevant gene network/pathways has been investigated [26, 27]. Interestingly, most model study systems exhibit spatial local adaptation due to recent selective sweeps being dated less than 25,000 years old [28-30] which are associated with strong bottlenecks during colonization events (e.g. colonization of Europe in humans and Drosophila) or due to recent recolonization after glacial period, bottlenecks in glacial refugia and selfing (in *Arabidopsis thaliana*, *Capsella sp*. and *Arabis sp*.). The observation of recent sweeps, and lack of older selection events, is explained by the lack of power to detect loci under strong positive local and temporal selection due to three factors. First, there is a confounding effect of past demographic history of the populations [31] especially under strong bottlenecks often characterizing the colonization of new habitats [32]. Disentangling the two types

of signatures, neutral demography and selection, is possible by inferring the past demography over the whole genome and search for outlier portions of the genome under selection. However, under strong demographic bottlenecks, it is difficult to carefully control the statistical rate of false positive sweeps and sweep signatures pre-dating the bottleneck are blurred [28, 30, 31]. Second, strong selection at few genes is only the tip of the iceberg, as adaptation likely occurs for traits, which are polygenic with selection acting over many genes. Different number of major genes with selective sweeps signatures can be observed depending on the distribution of selection coefficients, the efficiency of selection (a function of effective population size and recombination rate), the architecture of the traits, place of genes in gene networks/pathways, and gene pleiotropy [33, 34]. While theoretical predictions exist, a general understanding of the genetic architecture of local adaptation is still missing [33]. Third, under temporal and polygenic selection model, strong selection can only occur if population sizes are large enough [31, 33]. Furthermore, theory predicts that selective sweeps can be observed and detected reliably only for ca. 0.1Ne generations after selection, where Ne is the effective population size [24]. As a result, in species undergoing habitat range expansion with strong bottleneck during colonization and life history traits decreasing effective population size (e.g. selfing), we expect 1) only few sweeps to occur under polygenic selection, and 2) only recent sweeps to be detected using polymorphism data. Therefore, in order to reveal the underpinnings of long-term spatial and temporal local adaptation due to strong selection (selective sweeps), the study of alternative species with larger effective population sizes during colonization events is required.

1.1.2 Demographic history and ecological niche

Reconstructing the past history of given species or populations are key to many researches addressing the ecological and evolutionary dynamics of natural populations not only for its own sake, but for disentangling demographic from selective effects [35, 36]. The inference of demographic history refers to identifying a

best model describing population size changes as well as population divergence and mixture events over time, such as divergent time and gene flow (migration) between different populations. Genomic data with numerous variations have important practical benefits for inference of demographic history. To gain insights from genomic data, we need models that describe genetic variation, the Single Nucleotide Polymoprhisms (SNPs), such as the coalescent [37, 38]. Population genetics analysis combined with appropriate dating, can recognize the key factors (such as climatic events in specific period) determining the demographic history of a species. With enough research (genome) resources, this can be done with outstanding detail [39]. Another important role of a precise demographic model is to infer null models for the detection of loci or regions under selection [40]. In identification of loci under positive selection, the neutral simulation can help researcher to obtain an accurate threshold to recognize outliers from genome scan. Therefore, a good demographic inference can significantly improve the results of local adaptation and other downstream analyses.

Another good way to investigation of population dynamics is reconstruction of ecological niche models (ENMs) in different historical periods [41, 42]. It describes both the range of conditions necessary for persistence of the species, and its ecological role in the ecosystem [43]. ENMs are most often used to combined with climatic data to: (i) estimate whether the current habitat is suitable for the survival of a known or unknown species, (ii) evaluate the dynamic changes of suitable habitats over time, and (iii) provide an estimate of the species niche. In our study, ENMs are very suitable to combine demographic model to infer the dynamic changes of species distribution in different historical periods. This helps us gain insight into the direction of habitat expansion *S. chilense* as the climate changes in different climatic periods.

1.1.3 Detection of sweeps under positive selection

When at a locus in the genome a strongly beneficial mutation occurs and spreads to

all individuals of the population, the frequency of this beneficial mutation as well as linked neutral mutations increase. This process is described in a seminal paper by Smith and Haigh [23], for which they coined the term "genetic hitchhiking". This process can remarkably reduce the genetic variation in specific regions and affect the allele frequencies near the sites favored by positive selection. The reduction of genetic diversity in local regions in the genome caused by the fixation of an advantageous allele sweeping by natural selection pressure, is called “selective sweeps”. Based on the above description, further signatures of the selective sweep regions include (1) changes in the site frequency spectrum (SFS) of polymorphisms [15, 44] with a characteristic U-shape SFS, and (2) shifts in level of linkage disequilibrium (LD) in the early and later phase of the fixation process [45, 46]. The selective sweep and hitch-hiking model assume that the population size maintain approximately constant over demographic history and is isolated, and no gene conversion has occurred in the proximity of the beneficial mutation [47].

Although the assumptions of the selective sweep model are relatively strict, some tests have been developed. (1) Methods based on diversity reduction in sweep regions. The most obvious and sustained effect of genetic hitchhiking is the reduction of nucleotide diversity in sweep regions [31]. Some basic statistics of population genetics can provide assessment of genetic diversity directly such as statistics of nucleotide diversity including π and θ_w [48, 49]. It is relatively easy to observe regions of low diversity from these statistics. (2) Detecting sweeps based on shifts of SFS. The studies by Braverman et al. [44] and Fay and Wu [15] suggested that selective sweeps shift SFS toward high-frequency and low-frequency derivative variants (U-shaped SFS). Neutral variants initially associated with beneficial variants increase in allele frequency, while variants not initially associated with beneficial variants decrease in frequency. Some summarized statistics have been widely used in population genetics, such as Tajima’s D and Fay and Wu’s H. The development of the composite-likelihood-ratio (CLR) test is a breakthrough for identification of selective sweeps [24]. It compares the probability of the observed polymorphism data under the

standard neutral model with the probability of observing the data under a model of selective sweep. (3) Identifying selective sweeps based on Linkage Disequilibrium (LD). The third characteristic of a sweep under positive selection refers to the specific LD patterns that occur near beneficial mutations. Against beneficial mutation was fixed in population, the LD levels will raise appeared on both sides of the fixed locus, whereas LD level is reduced between polymorphisms observed on different sides of the fixed locus. This is caused by a single recombination event on the both sides of the selected site allowing multiple polymorphisms to escape the sweep on the same side of the sweep [45, 50]. Therefore, between SNPs on the both sides of sweep will be observed the high level of LD. On the other hand, polymorphisms that reside on different sides of the selected site need a two or more recombination events, thus LD is decreased. Meanwhile, Kim and Nielsen proposed an LD based ω-statistic test to detect selective sweep [45]. A crucial result of population genomics theory is that detection of selection is to be done by comparing outlier regions to the rest of the genome to account for the variance of diversity and SFS due to the past demography. This is because strong positive selection occurs locally in the genome, while the whole genome is affected by genetic drifts and past demographic effects [51, 52].

Despite these methods being widely used in detection of selected sweeps, generating many credible results [53-56], many challenges still need to be faced due to strict assumptions of the underlying model, for example dynamic changes of population size and recombination rate. With the development and progress of machine learning methods in the last decades, a few methods of detecting selected sweeps have been developed [57-59]. Typically, in a machine learning solution, the goal is to accurately detect previously unseen data based on a simulated set of already seen data (the process of training data). A dataset of neutral simulation is critical to obtain a good training dataset. This also requires the researcher to infer the best demographic history.

1.1.4 Evolution of transcriptome

In addition to genome data (sequence), we can also obtain gene expression data, which underlies the phenotype. Distance-based comparative transcriptomics has now been well established to estimate the gene expression difference or distance of orthologous genes between different species. While transcriptome analysis based on distance demand detailed information of transcriptome of at least two species, recent developments of transcriptome indices require only a single transcriptomic information. Here in this thesis, transcriptome indices are used such as the transcriptome age index (TAI) [60] and the transcriptome divergence index (TDI) [61] to understand the evolutionary age or sequence divergence of a gene in combination with its expression level. These integrative methods allow the comparative genomics and transcriptomics to estimate evolutionary process.

The TAI is computed based on phylostratigraphy and expression profile, which doles out each gene to different phylogenetic ages by identification of homologous genes related to other species [62]. Following this definition, genes will be assigned to discrete level of phylogenetic age or nodes along the tree of life, named phylostrata (PS). The evolutionary age of each gene was quantified combining its PS and expression level to obtain a weighted evolutionary age. Finally, weighted ages of all genes are averaged to yield TAI, which is defined as the mean evolutionary age of a transcriptome [60]. A lower value of TAI in a transcriptome describes an older mean evolutionary age, whereas a higher value of TAI denotes a younger mean evolutionary age and implies that evolutionary younger genes are preferentially expressed in the corresponding sample or condition [60, 63]. The TDI represents the mean sequence divergence of a transcriptome [61]. It is quantified by divergence strata (DS), which sort genes into discrete sequence divergence categories based on Ka/Ks ratio (ratio of non-synonymous to synonymous sites). This ratio allows to assess the degree of purifying and positive selection in coding regions. The genes are assigned to different DS and then weighted by their expression level to yield the TDI. A lower value of TDI denotes a more conserved transcriptome profile, whereas a higher TDI value shows a more variable transcriptome [61]. As the tree of life spans

from the cell organism to the current taxonomic level, the TAI integrates both evolutionary ancient and recent signals. In contrast to the TAI, the TDI investigates evolutionary time among related species, depending on chosen species. Hence, TAI and TDI estimate different evolutionary properties on different time scales [64].

1.2 Gene network responses of plant to environmental stress

In many cases, the functional background of genes or their products can be inferred by 'omics' methods, such as large-scale protein-protein interaction studies (interactomics) or studies of gene co-expression networks (transcriptomics). In this context, researchers tend to pay more attention to the regulation mechanism of the complete pathway, as well as restriction or promotion to other pathways, in order to gain an understanding from the perspective of the functional module. Through developments of bioinformatic pipelines, such 'omics' data sets can be unearthed to detect genes in same biological process or pathway [65-67]. To date, large-scale transcriptomic data are adequate for many plant species [68]. Therefore, it may be a suitable ideal to combine comparative genomic with transcriptomic. These multiple 'omics' methods allow for identification of functionally related genes by co-expression analysis, and then estimate evolution of gene expression.

A highly effective way of displaying and investigating co-expressed relationships between genes is by representing them as gene co-expression networks. Another advantage of gene networks is to infer the biological functions of unknown genes. The comparative analysis of co-expression networks shows that some networks are highly conserved in different species [69-71]. These conserved relationships between species have been utilized to predict biological functions for unknown genes in species that are insufficient annotated as compared with certain model species, and some tools were also developed [72-74]. These conserved gene networks are called conserved modules, which reveal gene connections in genome-wide co-expression networks that have similar compositions in terms of gene families, protein domains

(labels) or motifs across multiple species [75]. Many conserved networks can be observed in many fundamental biological processes involved in responded to environmental stresses, for example, protein metabolic, cell cycle, and photosynthesis [69, 76].

Drought is one of the major abiotic stresses that essentially influence the plant development and harvest yields. In the previous decade, worldwide misfortunes in crop production because of drought added up to ~$30 billion [77]. The capacity of a plant to response the water-insufficiency signal and initiate adapting procedures accordingly is characterized as drought response, which is a complex process that proceeds through several strategies: (i) acceleration of growth and reproduction before drought stress to escape stress, (ii) storage of internal water to avoid tissue damage, and (iii) strengthen endurance under drought stress while maintaining the growth process [78]. These strategies involve many processes, such as increasing activity of tissues i.e. root water uptake and closing stomata, activating pathways including phytohormone signaling, antioxidant and metabolite production to regulate osmotic processes [79]. The mechanisms of these response to drought stress have been revealed extensively to construct complex regulatory gene networks, including from the perception of water shortage signaling and physiological and cellular responses. These details of regulatory gene networks have been described in several review papers [80-82].

In addition to drought stress, cold (low temperature, LT) is one of the most harmful environmental stresses. LT constrains plant growth and restricts productively in temperate region. Previous studies indicated that the gene networks involved in cold (low temperature, LT) stress are also conserved in different plant, especially in flowering plants [83, 84]. In addition to affecting the growth and development of plants, LT stress also significantly limits the geographic distribution of plants [83, 85, 86]. Flowering plants can adapt to the LT environment by multiple pathways, such as the amino sugar and nucleotide sugar metabolism pathway, protein export, and lipid

metabolism pathway. Among these, regulation of flowering time is one of most wildly reported ways to adapt to LT in flowering plants [87]. For example, vernalization is the typically programmed physiological process in which prolonged LT-exposure provides competency to flower in plants; widely found in winter and biennial species.

1.3 Study system

The tomato is one of the most important commercial crops and the family consists of one cultivated species (*Solanm lyopersicum L.*) and 12 closely related wild tomato species [88, 89]. The latter constitute important germplasm resources for the improvement of agronomic cultivars [90] and also played a considerable role in basic research on plant breeding systems, genetics, and the evolution of plant species [91-93]. *Solanm lycopersicum* has been planted all over the world and is thus of high economic importance. It has been domesticated from *S. pimpenilifollium* which diverged from other species around 6 million years ago. Originally, however, all tomato species are native to western South America and distributed from Central Ecuador to Chile [94]. Due to a series of sequential bottleneck phenomena during domestication, the cultivated tomato exhibit very limited polymorphism, and a relatively small genome (~750Mb) [1]. By contrast, its wild relatives are characterized by high phenotypic diversity, a larger genome size (1Gb) and have successfully colonized a great variety of natural habitats that differ in both abiotic factors such temperature and drought, and biotic factors such parasite load [89]. Nevertheless, the genetic bases of wild tomatoes extraordinary adaptations are not thoroughly understood.

We focus on the most southerly distributed wild tomato species, *Solanum chilense* which is found in southern Peru and northern Chile in mesic to very arid habitats [93], and the low altitude coastal habitat, and the highland altitudes of the Andean plateau (Figure 1A). Raduski and Igic provided detailed descriptions of the morphology by multiple fieldwork [95]. *S. chilense* is a perennial herbaceous plant

with a woody base that grows upright or creeping in habitats with different environmental conditions. It is usually 0.5 to 1 m tall and 1 m wide when growing in isolation along cliffs or roadsides, but inside deep Andean volcanic canyons it commonly spread more widely due to snowmelt and sparse rain.

1.3.1 Habitats and climatic conditions of *S. chilense*

S. chilense is an outcrossing wild tomato species found in southern Peru and northern Chile in mesic to very arid habitats (Figure 1A) [96]. The ancestral species range is found around the Peruvian and Chilean border mainly along the (mid-altitude) 'pre-cordillera' region (800 - 2000 m altitude) characterized as marginal desert habitat. *S. chilense* colonized independently two different southern, but arid, isolated regions around the Atacama desert at different time points [97, 98]: an early divergence (older than 50,000 ya) with the colonization of coastal habitats (in Lomas formations), and a more recent lineage divergence (less than 20,000 ya) restricted to highland altitudes (above 2,400m) of the Andean plateau. Therefore, the previous studies of population structure divided all populations of *S. chilense* into three differentiation groups: central group (also northern inland of Chile), south-coast group, and south-highland group, using reduced-representation genome sequencing [95, 97, 98]. In addition, there are also a small number of populations distributed in southern Peru and genetically close to the central population.

First, the geographical distribution of populations in the *S. chilense* is extremely discontinuous (Figure 1A), especially between central and south-coast groups, and between south-coast and south-highland groups. South-coastal natural populations instead occupy wide canyons below infrequent mist capturing hill oases called "lomas formations" [99], scattered among the Chilean coastal hills (0-1000 m elevation), and these populations are not continuously distributed along the Pacific Ocean [95]. Second, the span of elevation between different populations is very large from coastal

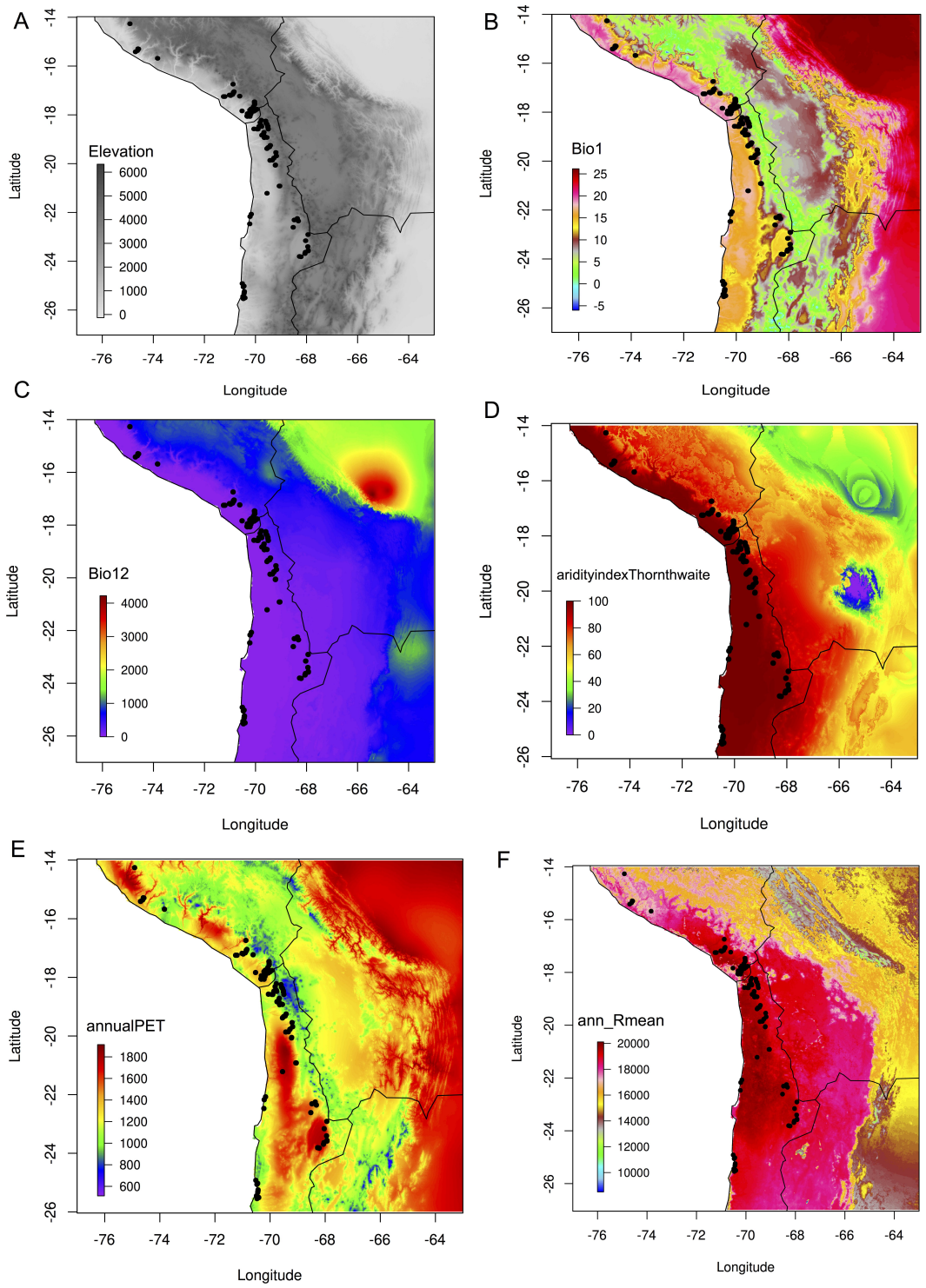

Figure 1. The geographic distribution and climatic conditions of the habitats in *S. chilense.* (A) Elevation. (B) annual mean temperature (Bio1). (C) annual precipitation (Bio12). (D) Thornthwaite aridity index (aridityIndexThornthwaite). (E) annual potential evapotranspiration (annualPET). (F) annual mean solar radiation (ann_Rmean). All climatic variables are described in Data S1E.

to highland/mountain (Figure 1A). The complex central populations occur from the edge of Atacama desert of low altitude (200 m) to Andean cordillera of high altitude (3500 m). Third, the wide and complex geographical distribution also determines very different climatic conditions in their habitats (Figure 1B to 1F). In general, the populations of *S. chilense* are distributing in extremely arid environments (Figure 1B to 1D). Coastal populations expose to a higher temperature than inland populations and almost no precipitation (Figure 1B and 1C). High-altitude populations suffer from the colder environment and slight precipitation (Figure 1B and 1C). In addition, high-altitude populations are also exposed to stronger solar radiation than low-altitude populations (Figure 1F). Overall, the populations of *S. chilense* span multiple types of landforms, resulting in its need to adapt to complex climatic conditions and stresses.

1.3.2 Local adaptation in *S. chilense*

Wide geographical distributions and complex climatic conditions of habitats indicate that *S. chilense* likely had toadapt to changeable habitat environments in the process of population expansion and colonization. Signatures of natural selection (positive or balancing) at few genes involved in abiotic and biotic stress adaptation were found when scanning few candidate genes and analyzing differential expression on transcriptome level [77, 97, 100-103]. From the perspective of climatic conditions of habitats, drought and low-temperature stress are two of the main threats to *S. chilense*. Some drought-related genes exhibit characteristics of local adaptation, such as some genes in ABA (abscisic acid) signaling pathway show nucleotide diversity patterns of local adaptation [104, 105]. In addition, not only in evolution, anatomical, morphological and physiological traits related to leaves and stems also observe significant differences between conditions of optional watering and restricted watering [106]. The combined analysis of population genetic, comparative transcriptomic and plant physiology approaches suggest that genetic adaptations to low temperatures evolved in high-altitude populations of *S. chilense* distributed along Andean altitudinal gradients [107]. The stresses of salt and heat are usually accompanied by drought

stress, especially for low-altitude coastal populations. Some studies indicate that salt and heat-induced changes of transcriptomes and proteomes levels in *S. chilense* and identify a series of related genes [77, 101, 108, 109]. Furthermore, some defense genes are also identified as selective signatures across populations, such as nucleotide binding site, leucine-rich repeat receptors (NLRs) show signatures of adaptation to help colonize to different habitats [98, 110, 111].

In summary, *S. chilense* has revealed many characteristics of local adaptation to complex and extreme environments and identified selective signatures to understand mechanisms of local adaptation. This also fully shows that polygenic adaptation across several gene network and pathways likely occurs in *S. chilense*. But these studies are based on selected candidate genes or transcriptome level and only obtain limited details of local adaptation. Therefore, genome-wide genome scans are expected to insight into a comprehensive understanding of local adaptation in *S. chilense*.

2 Overview

2.1 Motivation

In the present thesis, we obtained full genome sequence data for 30 diploid and highly heterozygous plants from six populations (five plants per population) representing the three main habitats of *Sonalum Chilense*: the central group (area of origin at low to high altitude), south-coastal group, and south-highland group. These habitats differ in their current climatic conditions: the south-highland group is different from the central group, while the south-coast appears as only marginally different from the environment prevailing in the central group. In addition, transcriptome sequencing data from 16 plants were obtained in normal and drought conditions representing the drought population: LA1963. In Chapter 3, the first aim of this study is to study local and temporal adaptation to changing climate using genome sequence data. We (1) perform inference of the past demographic history of the species colonization as well as reconstruct the past climatic history of the species' range; (2) we conduct genome scans for genes under selection in these populations and assign functions and gene network positions to these candidate genes; (3) link climatic and genetic data at candidate genes to highlight the relevance of few key gene regulatory networks (pathways) for local spatial adaptation; and (4) connect past species niche reconstruction and age estimates of the selective sweeps to reveal the history of the temporal local adaptation to ever changing climatic conditions. In chapter 4, I study the evolution of gene networks involved in drought tolerance by combining transcriptome and genome sequence data. These analyses allow me to (1) identify gene regulatory networks correlated with drought response; (2) explore the evolutionary age and sequence divergence of gene networks; and (3) combine population genetics and comparative genomics to understand evolutionary process and trends of gene networks involved in drought tolerance. In the discussion chapter 5, I attempt to obtain a comprehensive understanding of local adaptation and evolution in *S. chilense*.

2.2 Contribution

I am the first author of the two manuscripts constituting the thesis (Chapters 3 and 4). Chapter 3 is currently submitted and Chapter 4 will be submitted in 2022 after adding a section on the new *S. chilense* genome assembly. I have contributed ideas, coded all bioinformatics pipelines, performed all statistical analyses and wrote the first version of both manuscripts. Dr. Silva-Arias has contributed to the demographic model in Chapter 3. The drought stress experiment was performed by Dr. Sharifova who was a guest researcher at TUM. Prof. Tellier and Dr. Silva-Arias have contributed to the discussions and revised the manuscripts of both Chapters 3 and 4.

3 Local and temporal adaptation

3.1 Materials and Methods

3.1.1 Sample collection and sequencing

Plants were grown from individual seeds obtained from the Tomato Genetics Resource Center (TGRC, University of California, Davis, USA; http://tgrc.ucdavis.edu). We attempted to sample plants representing the three main infra-specific lineages of *S. chilense* (i.e. central, south-coast, and south-highland) by accessions spanning the complete known geographic range (Accessions: LA1963, LA3111, LA2931, LA4330, LA2932, LA4107; Figure 1A; Table S1) [97, 98]. Plants were grown in individual pots in glasshouse conditions at the Dürnast plant research facilities of the TUM School of Life Sciences.

Genomic DNA was extracted from five mature plants per accession using the DNA kit from Quiagen according to the manufacturer instructions. Individual samples were sequenced at Eurofins Genomics on an Illumina HiSeq 2500 with standard library size of 300 bp. The 30 *S. chilense* whole genome sequencing data are available on ENA in BioProject PRJEB47577.

3.1.2 Reads mapping, SNP calling and filtering

Quality control of the raw reads was performed with FastQC (v0.11.6) [112]. Base calls with insufficient quality were removed using Trimmomatic (v0.36) using default parameters [113]. Adapters were cut with a seed mismatch of 2 with a palindrome clip threshold of 30 and a simple clip threshold of 10. The trimming was performed with a sliding window approach, cutting once the average quality was below 15 within a window of 4 bases. Low-quality bases from the leading and trailing of reads were also removed. Furthermore, reads with a smaller length than 36 bases were dropped. The trimmed read were mapped to the *Solanum pennellii* reference genome [2] available from Solanaceae Genomics Network (https://solgenomics.net) using

Burrows-Wheeler Alignment tool (v0.7.16) using default settings [114] and afterward sorted with Samtools (v1.5) [115]. Next, the read groups were added, mate information was fixed and technical duplicates were marked with PicardTools (v2.10.6) [116].

Variant calling was performed using the HaplotypeCaller tool of GATK with default parameters for each sample. The individual variant files were then combined into a variant matrix with the GenotypeGVCFs tool. Single nucleotide polimorphisms (SNPs) we selected and then filtered using the GATK VariantFiltration module keeping only the SNPs that met all following criteria: quality (MQ) < 30, rank sum test for mapping qualities (MQRankSum) < -5, quality-by-depth ratio (QD) < 10, strand bias estimated by the symmetric odds ratio test (SOR) >3.0, read position rank sum test (ReadPosRankSum) < -8, phred-scaled p-value using Fischer's Exact test (FS) > 10, and depth coverage (DP) < 3.

Finally, after filtering, the variants were annotated using the snpEff software (v4.3) with default settings [117]. For each SNP, the annotation were identified based on the gene annotation of the *S.pennellii* reference annotation to obtain relationship to coding sequences in the genome and how it may change the coding sequence and affect the gene product.

3.1.3 Population genetics analyses

For all population genetics analyses, we use *S.pennellii* population LA716 as the outgroup to polarize SNPs. The phylogenetic tree was constructed by the maximum likelihood (ML) method using SNPhylo pipeline (v20160204) based on all SNPs [118]. The pipeline has steps such as removing low-quality data and considering linkage disequilibrium. The reliability of each branch was evaluated by bootstrapping with 1,000 replicates. The ML tree was visualized with iTOL (v5.0) [119]. A principal component analysis (PCA) was performed for the whole SNP dataset to seek a

summary of the clustering pattern among sampled genomes using GCTA (v1.91.4) [120]. The inference of population structure was performed using the program ADMIXTURE (v1.3.0) [121]. Nine scenarios (ranging from K = 2 to K = 10) were assessed for genetic clustering.

For each population we obtained statistics with 100 kb sliding windows to estimate their genetic diversity and differentiation. Nucleotide diversity (π), Tajima's D and F_{ST} were calculated with VCFtools (v1.15) [122]. Watterson's estimator of theta (θ_w) and Fay and Wu's H were calculated using ANGSD (v0.921) [123]. The linkage disequilibrium (LD) levels were calculated per population as the genotype correlation coefficient (r^2) between two loci (within the same chromosome) using VCFtools with a maximum distance of 1000 kb. LD decay was then assessed for all pairs of SNPs within 1000 kb distance using PopLDdecay (v3.27) with parameters '-MaxDist 1,000 -Het 0.1 -Miss 0.1' [124].

3.1.4 Inference of demographic history

The demographic analysis was conducted using the Multiple Sequentially Markovian Coalescent method (MSMC2) with phased VCF files [125]. First, The phased data were generated with SHAPEIT (v2.17) [126]. First, the biallelic and no missing SNPs were extracted by VCFtools, then the haplotype graphs were constructed to capture phasing uncertainty with parameters '--thread 10 --prune 100 --burn 100 --main 500 --states 1000 --window 0.5 --rho 0.04 --effective-size N_e'. Here N_e was calculated using a formula: (θw/effective number of sites)/($4\times5.1\times10^{-8}$), then the mean N_e of all populations was calculated for each chromosome. Finally the most likely pairs of haplotypes from the graphs were extracted and converted outputs to phased VCF files. We assumed generation time of 5 years and a rate of 1×10^{-8} mutations per generation to scale time in the result plots. MSMC2 inputs were created according to the following steps: (i) the mask files were generated using genomecov function of bedtools (v2.27.1) [127], that gives information in which regions in the genome could

be called and excluded, and corresponding negative mask files including mappbility information of regions; (ii) the final input files were generated with the script generate_multihetsep.py within the msmc-tools (https://github.com/stschiff/msmc-tools), which merges information from VCF and mask files together. The coalescence rates were estimated by default time segment patterning using MSMC2 (defining 40 hidden states). Besides, the cross-coalescence analysis was performed for each pairwise comparison of genomes between pairs of populations to estimate the population separation history.

The migration rate was estimated with a python script MSMC-IM using the output of MSMC2 as input [128]. It also reports M(t), the cumulative migration probability, evaluated at each time boundaries, which denotes the probability for lineages to have merged by the time t. The mutation rate is 1×10^{-8} as for the analysis with MSMC2. The initial constant effective population size for each population was set as 10^5 calculated from MSMC2 outputs. The pattern of fixed time segments used default values of MSMC2: 1*2+25*1+1*2+1*3. The beta used recommended values: 10^{-8},10^{-6}, which regularize on estimating migration rate and population sizes.

3.1.5 Ensemble niche modelling and temporal distribution projection

We performed an ensemble modeling framework [129] using the BIOMOD package [130, 131] in R (R Core Team 2020) combining models from eight algorithms to account for the uncertainty associated with particular modeling technique. We included the algorithms: 1) generalized linear models – GLM [132]; 2) generalized additive models – GAM [133]; 3) generalized boosting models – GBM [134]; 4) classification tree analysis – CTA [135]; 5) multiple adaptive regression splines – MARS [136]; 6) random forests – RF [137]; 7) artificial neural networks – ANN [138]; and 8) maximum entropy – MAXENT [139].

Occurrence localities include 110 points covering the full extant geographic range of *S. chilense* retrived from the TGRC website and deduplicated keeping one occurrence per raster cell grid using the function *gridSample* of the dismo R package [140]. The pseudo-absence set includes 10,000 sampled within the species distribution area outside 1 km buffer around the presence records. Prevalence was maintained at 0.5, allowing the sum of presence-weights to be equal to the sum of pseudo-absence weights in the model calibration process [141]. Five cross-validation replicates were run in which presence records were randomly divided into training and testing subsets (75% and 25% respectively). For each cross-validation replicate ten runs were performed with different pseudo-absence sets, therefore completing a total of 400 models. To assess the predictive performance of the models, the threshold-independent area under the receiver operating characteristic curve statistic [139] (AUC) and the threshold-dependent true skill statistic [142] (TSS) were calculated. Consensus niche models were obtained using a TSS-weighted average method to account for the predictive power of each fitted model. Models with low predictive power (TSS < 0.7) were discarded.

The environmental predictors include five variables selected from 63 climatic layers available from three public databases (Data S1E). The 19 bioclimatic variables as well as monthly wind speed, water vapor pressure, and solar radiation data were obtained directly from the WorldClim2 database [143]. Data S1E expanded bioclimatic variables that include solar radiation and topographic features were downloaded from ENVIREM dataset [144]. Also, global aridity measures were obtained from the Consultative Group on International Agricultural Research (CGIAR) [145]. Variable selection was implemented using the function *corselect* of the R package fuzzySim [146] by keeping variables with Pearson correlation below 0.75 and variation inflation factor below 10. The bioclimatic layers were cropped from latitude 10° to 27°N and longitude 63°to 80°W with raster R package [147], this geographic extent include the full extant species range. All fitted suitability models were projected to infer the distribution of *S. chilense* suitable habitat under current and past conditions during the

Last Glacial Maximum (LGM; ~21 Kya) downloaded from WorldClim and ENVIREM (Data S1E).

3.1.6 Genome-wide selection scans

We identified selective sweeps using biallelic SNPs by SweeD [148] and OmegaPlus [149]. SweeD (v3.3.1) implements a composite likelihood ratio (CLR) test which detects complete selective sweeps using deviation from neutrality based on Site Frequency Spectrum (SFS). It is based on the SweepFinder algorithm described in Nielsen (2005) [150]. CLR statistic is used to identify regions where the matched the expected SFS generated from a selective sweep based on background SFS. CLR was calculated with default parameters except for using a resolution of 10 kb intervals within each chromosome. OmegaPlus (v2.3.0) detects selective sweeps based on linkage disequilibrium (LD). The ω statistic was computed at 10 kb intervals. We specified a minimum window of 10 kb and a maximum window of 100 kb to be used for computing LD values between SNPs, respectively. The choice of the window length of 10kb for these statistics was based on 1) setting up a minimum average number of SNPs per window as on average we obtained at least 50 SNPs per 10 kb window per population, and 2) the decay of LD which shows r^2 values below 0.3 beyond a 100 kb distance and the slowest decay was always maintained above 0.3 in LA4330. Outlier CLR and ω statistics indicative of a selective sweep are defined by comparison to the genome-wide distribution values. In order to be conservative, we derived here threshold values for each population based on simulations from the estimated demographic history [151].

We used the coalescent simulator SCRM (v1.7.3) [152] to generate 10,000 neutral datasets of 10 Mb based on the demographic history of each population. A mutation rate of 1×10^{-8} was used. An important source of error when performing selection test is the heterogeneity in recombination along the genome [153, 154]. To reduce false positives, we set up simulations with varying recombination rate every

100 kb within each 10 Mb simulated block, because we set a maximum window of 100 kb in genome scan. Precisely, we divided the simulated block of 10 Mb into 100 windows of 100 kb and assign different recombination rates to each window (the SCRM command line of simulation describes in https://github.com/weikai-320722/Schil_30WGS). We used a block of 10 Mb as our analyses revealed that beyond half of this distance the LD is very low ($r^2 < 0.2$). For each of the simulated datasets, both CLR and ω statistics were calculated with the same parameters as above. The maximum value of each statistics was extracted from each simulated dataset, and we thus obtained a distribution of 10,000 maximum values for each statistic. The 95th percentile of this maximum distribution was specified for SweeD and OmegaPlus as the thresholds to identify outliers for selection. We then extracted the overlap regions of the two methods using the genomic coordinates and those regions were regarded as high confident selective sweep regions.

3.1.7 Age of candidate regions under positive selection

The age of hard selective sweeps were also estimated using R package McSwan(v1.1.1) [155]. Similar to SweeD, McSwan compares local frequency spectra (SFS) simulated under neutral and selective demographic models to detect selective sweeps, and it was used to assign selective scans to genomic regions and predict the age of selection events. McSwan facilitates the estimation of sweeps age in non-model organisms because it is not necessary for it to detect sweeps and predict age with high-quality haplotype data. However, this advantage comes at the cost of not jointly estimating the selection coefficient for a particular sweep, so McSwan assumes that the selection strengths in all sweeps are equal.

Coalescent simulations in McSwan are performed using Hudson's ms software [156]. The demographic history must therefore be formatted as a string of switches according to ms conventions. The neutral demographic model was specified using

same parameters with previous simulations with SCRM for SweeD and OmegaPlus. Then the reference table was constructed using the *generate_priors* function using ms-formatted demographic history as defined previously and we simulated neutral and selection SFSs that were each comprised of 2,000 simulations (default recommendation) across sequences 1 Mb in length. The empirical SFSs were generated from scans across the 1 Mb region using the *coalesce* function. For our VCF file, we first convert the VCF to a McSwan-readable file containing per-SNP allele counts by the *convert_VCF* function. Then, the genome scan and estimation of age were pre-formed with the *gscan* function. To precisely determine the boundaries of sweep regions, McSwan iterates its genomic scans over adjacent windows of various lengths and offsets and compares the empirical SFS to the simulated SFS under selection to assign regions as selective sweeps. We set up the iterative scans in sliding windows that ranged from 1000 bp to 100 kb in length and a minimum of 20 SNPs required per window. Each sliding scan was done in 100 overlapping steps (default setting). We then looked for overlap between the regions detected as hard selective sweeps by McSwan to the candidate regions previously detected with SweeD and OmegaPlus.

3.1.8 GO enrichment analysis

Due to the lack of a complete gene function annotation database, we performed a BLASTX against the NCBI database of non-redundant proteins screened for green plants (e-value cutoff was 10^{-6}) and used Blast2GO to assign GO terms for each gene identified in the genome scan analysis [157, 158]. At the same time, we also performed a blast to the *A. thaliana* dataset TAIR10 separately to remove redundant terms [159]. We used, the R package clusterProfiler to perform GO enrichment analysis by generating an annotation database [160]. The false discovery rates (FDR) were calculated to estimate the extent to which genes were enriched in given GO categories. P-values smaller than 0.05 were used as cutoff for a significant level of correlation.

3.1.9 Genetic network construction

For each of the genes enriched in four specific biological processes, we retrieved the interacting gene neighbors using GeneMANIA (v3.5.2) [161]. We generated aggregate interaction networks in GeneMANIA, based on physical interactions, predicted and co-expression. The nodes in the resulting network represent the proteins, and the edges represent the protein-protein interactions. The network edges were weighted by the corresponding data source with the adaptive network weighting method by GeneMANIA. Finally, we performed hierarchical clustering and manually optimized the weighted value cutoff for displaying the gene network, so the gene clusters suggested by the network were similar to the clusters suggested by hierarchical clustering [162]. The functions of networks were extracted form GeneMANIA outputs with FDR cutoff 0.05 (Data S1D).

3.1.10 Genotype–environment association tests for local climate adaptation

To evaluate the relative contribution of the abiotic environment to explaining patterns of genetic variation, we used the Redundancy Analysis (RDA) to associate the SNP set identified within the selective sweep regions with potential environmental selection pressures. RDA analyses were performed with an individual-based approach, using as input allele counts for each locus for each sample (so count of 0,1 or 2). The SNPs containing missing data were removed because RDA requires a complete data frame. RDA was performed using the *rda* function from the vegan package as implemented in R [163], modeling genotypes as a function of predictor variables, and producing constrained axes and representative predictors. All variables were centered and scaled before running genotype-environment association test (GEA). Multi-collinearity between representative predictors was assessed using the variance inflation factor (VIF) and since all predictor variables showed VIF < 20 none were excluded. This may still cause some collinearity, but it is beneficial to find more connections between genotypes and environments. The loadings of the SNPs in the ordination space

determined which SNPs were candidates for being under local adaptation. The SNP loadings were stored as specified in the RDA object. The significance of RDA constrained axes was assessed using the *anova.cca* function and significant axes were then used to identify candidate loci ($P < 0.001$). Candidate loci were identified using 2.5 folds of standard deviation as cutoff (two-tailed p-value = 0.012). To clearly understand the response of genetic variation to climatic variables, the correlations of each candidate SNP with the representative climatic variables were calculated and take the variable with the strongest correlation as the response variable of the outlier SNPs.

3.2 Results

3.2.1 Overall whole-genome sequencing data and variant calling

We sequence whole genomes of 30 heterozygous plants from *Solanum chilense* from six populations (LA3111, LA1963, LA2931, LA2932, LA4107, LA4330) (Figure 2A, Table S1). All reads are aligned to the reference genome assembly of *S. pennellii* and all 30 *S. chilense* individuals show high-quality sequence and mapping scores with more than 97% of mapping paired reads, individual genome coverage ranging between 16 to 24 reads per base, and >70% genome coverage per sample (Data S1A). After SNP calling and stringent filtering, a total of 34,109,217 SNPs are identified across all samples (Table S2) for a genome size estimated approximately to 914Mb [110]. The highest and lower number of SNPs was observed in LA1963 and LA4107, with 27.18 and 21.75 million SNPs, respectively (Table S2). In order to better understand the distribution of SNPs in different populations, we also checked the overlaps of SNPs between different populations (Figure S1). More than 55% (13,904,088) SNPs in each population were observed in all populations and only 4.2 - 7.9% (948,174 - 1,582,572) unique SNPs were detected in a specific population. The two south-coastal populations (LA2932 and LA4107) showed higher number of unique SNPs, especially in the southernmost LA4107 population. Interestingly, in the

pairwise comparison (SNPs that only were observed in two populations), the overlap rate between LA2932 and LA4107 (91,632) resulted higher than others (9,536 - 38,020). Moreover, in other comparison groups (SNPs were detected in three, four, five populations, respectively), two coastal populations always showed the lowest overlap rate with other populations, especially LA4107 (Figure S1).

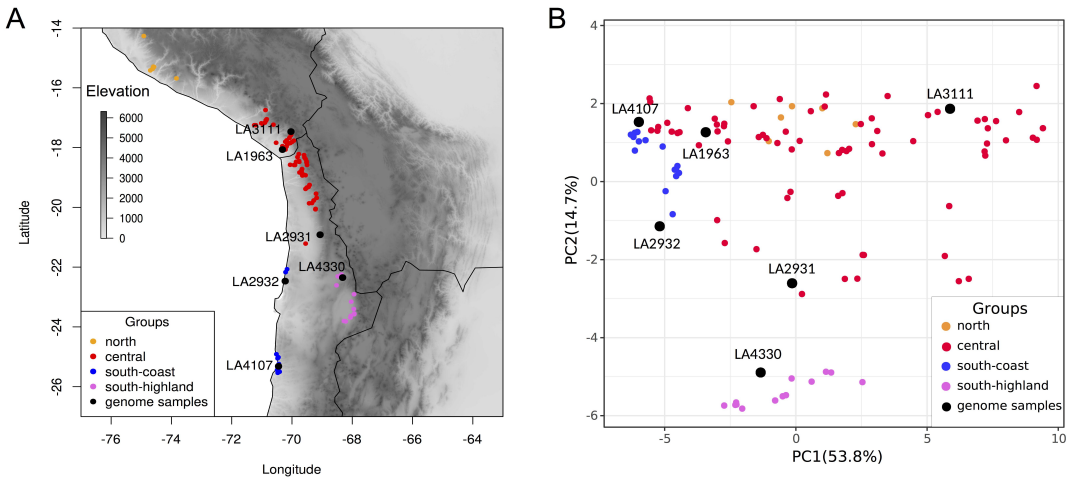

Figure 2. Geographic distribution and climate of sequenced populations of *S. chilense*. (A) Map with distribution of all *S. chilense* populations by the TGRC, the six *S. chilense* populations in this study (black circles), and the four population groups (colorful circles). (B) The PCA of all populations using 63 current climatic variables (Data S1E).

3.2.2 Population structure and statistics of genetic diversity and differentiation

To explore the evolutionary relationship among samples, we conducted phylogenetic tree, principal component analysis (PCA), and admixture analyses (Figure 3). In addition, to assess the impact of geographical distance on genetic differentiation, the correlation between genetic distance (pairwise Nei's distance) and geographical distance was calculated and showed non-significant correlation (Pearson test, r = 0.35, P = 0.2028; Figure 3A; Figure S2A). Phylogenetic and principal component analysis (PCA) analyses support population structuring into three genetic groups (Figure 3B and 3C; confirming the results in [97]): a central group (LA1963, LA3111, LA2931), the south-highland group (LA4330), and the south-coast group (LA2932 and

LA4107). Based on genetic co-ancestry analyses, we partitioned all individuals into known groups by varying the number of presumed ancestral populations (Figure 3D; Figure S2C; K ranged from 2 to 10). The analysis indicated that K=4, is the best supported number of population clusters (Figure S2B): one cluster grouped the three central populations (LA1963, LA3111, LA2931), another only the south-highland population (LA4330), and the two last clusters solely comprised each of the two south-coast populations (LA2932 and LA4107, respectively). The individuals of the population LA2931, the southmost of the central group, displayed small admixed ancestry coefficients (< 5%) with the south-highland group (Figure 3D).

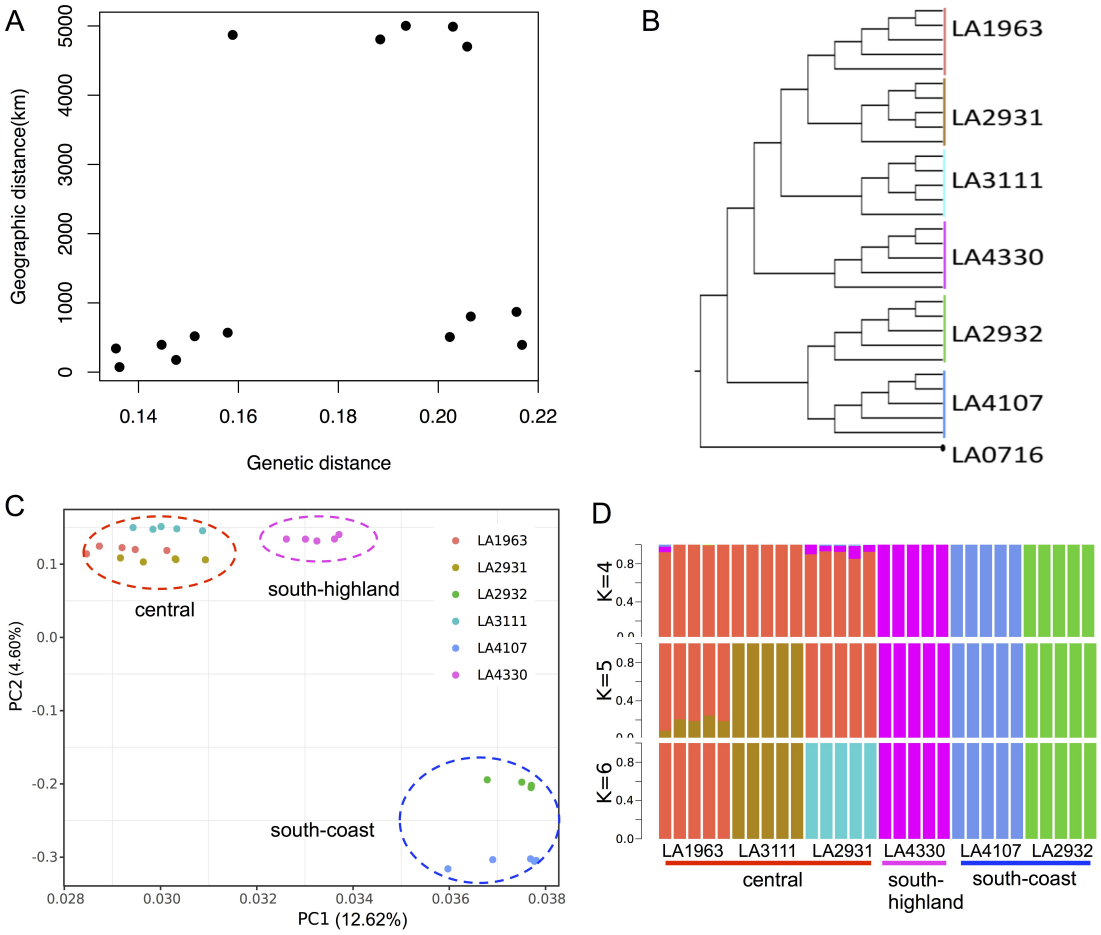

Figure 3. The population structure of 30 individuals. (A) The correlation between genetic distance (Euclidean distance) and geographic distance (Pearson test, r = 0.354, P = 0.281). (B) Maximum likelihood (ML) phylogenetic tree, S. pennellii LA0716 is used as outgroup. (C) PCA and (D) Admixture analysis (optimal K value is 4, Figure S1B).

For each population we obtained statistics of nucleotide diversity (π), Watterson's theta (θ_W) and pairwise differentiation (FST) to estimate their genetic diversity and differentiation (Figure 4; Table S3 and S4). At the genome level, the nucleotide diversity confirmed the trend observed with the SNP number of each population, and π and θ_W showed highly consistent distribution in six populations, respectively (Kendall's $\tau > 0.79$, $P < 0.001$). The central populations exhibited higher genetic diversity than the rest of the populations and lower genetic diversity was observed in the south-coast populations, especially in LA4107 (Figure 4A and 4B; table S3).

Average pairwise F_{ST} was higher between south-coast populations and other populations (Figure 4C; Table S4). The three central populations showed the lowest pairwise F_{ST} values. Highest F_{ST} was observed between two south-coast populations (LA2932 and LA4107) and between LA4107 with the other populations. The south-highland population exhibited intermediate FST with all central populations. These results support a strong isolation between the two south-coast populations as well as among the south-coast populations with the other sampled populations. In contrast, a moderate level of differentiation was found between the south-highland population and the central populations, and low differentiation between the three central populations (Figure 4C; Table S4).

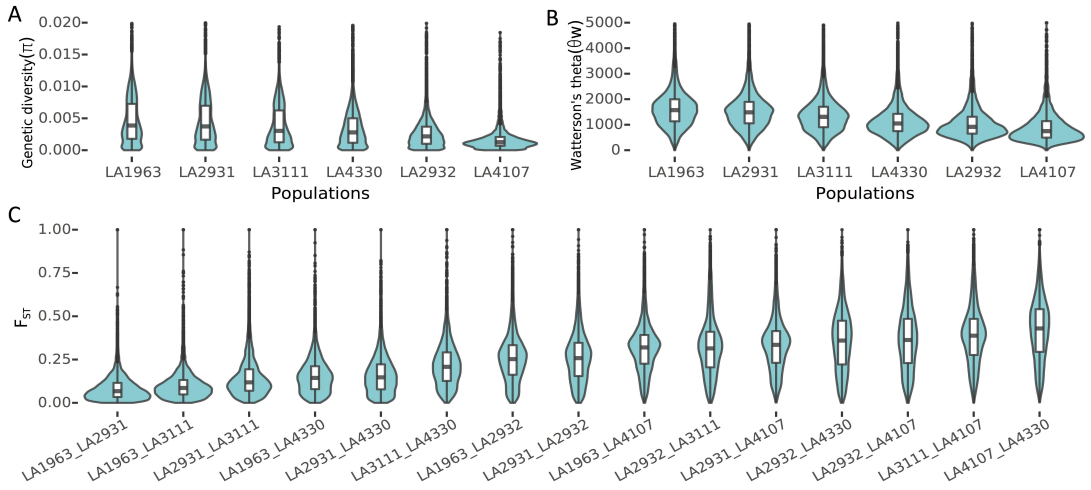

Figure 4. Statistics of nucleotide diversity and differentiation. (A) The distribution of statistics of nucleotide diversity (π). (B) The distribution of statistics of nucleotide diversity (θ_w). (C) The distribution of statistics of differentiation (F_{ST}).

3.2.3 Past demography of *S. chilense* is influenced by colonization events and climatic variations

As indicated by nucleotide diversity and differentiation statistics (Figure 4; Table S3 and S4), we confirm the previous results obtained using a reduced number of genetic markers: *S. chilense* has colonized independently the coastal and highland southern-habitats, and the species area of origin is likely the mid-altitude of the central region [97, 98]. Using multiple sequentially Markovian coalescent analyses, we estimate historical changes in effective population size (N_e, Figure 5A), divergence and potential post-divergence gene flow (Figure 6), and construct a consensus demographic model (Figure 5B; Data S1B). These estimates are compared with the reconstructed past climatic variation highlighting five Marine isotope stages (MIS) climatic periods [164, 165]. The two south-coast populations found in Lomas habitats show early divergence consistent with the admixture analysis (during the Last Inter-Glacial period, MIS5) likely from the lowland area of the central group (LA1963). The colonization of the highland likely occurred later, first in the central group region (LA3111, LA2931) between the last inter-glacial and Last Glacial Maximum (LGM) periods (ca. 75-130 kya, MIS3-4) and then with further colonization of southern

highlands (from 30 kya, MIS1-2, LA4330). All populations show a reduction of diversity matching with the estimated time of the LGM characterized as a cold and dry period and supported by a contraction of the suitable habitats to a narrow strip in lower altitudes, and a subsequent expansion thereafter (Figure 5C). Indeed, the local habitat at the current location of LA2931 and LA4330 was likely unsuitable for the establishment of the most southern highland population until 15 kya (after the LGM, *i.e.* during MIS1-2). The lower genetic diversity of the south populations (and estimated *Ne*) is thus due to the mild bottleneck processes during the southward expansion (Figure 5A; Figure 4; Table S3). Both south-coast populations show consistent signals of long-term history of colonization, subsequent isolation with negligible gene flow, and possible local specialization to sparsely suitable Lomas habitats along the coast (Figure 5B and 5C, Figure 6).

The divergence between the central group populations (during MIS3-4) occurs after the colonization of the coastal habitat (Figure 5B; Figure 6), but before the colonization of the south-highland (LA4330). Moreover, strong post-divergence gene-flow and low differentiation are found in the central group, especially among the pairs LA1963-LA3111 and LA3111-LA2931 (Figure 6), consistent with their geographical and environmental proximity (Figure 2B) and the range contraction during the LGM (MIS2 in Figure 5). The colonization of high-altitude regions in the central group is thus accompanied by high levels of gene flow despite these populations ranging across a large altitudinal gradient (2500m of altitude difference between LA1963 and LA3111 or LA2931). The divergence history results in the south-coast and south-highland populations to be fairly isolated from one another (as separated by the Atacama desert) leading to the suggestion of an incipient speciation process (Figure 5B; Figure 6) [95]. We want to assess if and how many selective sweeps underpin such adaptation to novel habitats (colonization of the coastal and highlands) and to temporally variable climatic conditions (past climatic conditions).

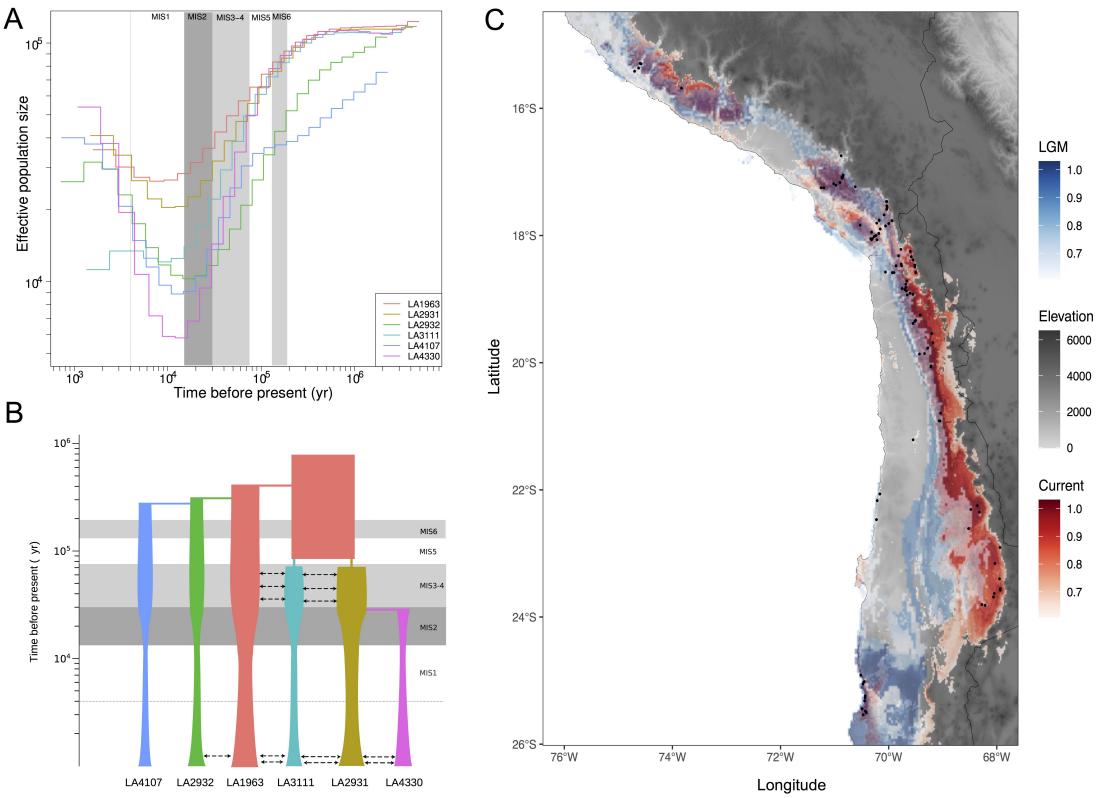

Figure 5. Demographic history and species distribution model of *S. chilense* for current and Last Glacial Maximum (LGM) climate conditions. (A) The estimation of historical patterns of effective population size (Ne) for 10 pairwise genome comparisons per population using the MSMC model. (B) Interpreted demographic scenario for the six samples populations of *S. chilense* including likely estimations of effective population size, divergence times and gene-flow. The width of the boxes represents the relative effective population size, arrows represent the migration between population pairs. Grey background boxes indicate five Marine isotope stages (MIS) climatic periods. (C) Overlay of the reconstruction of the distribution model for *S. chilense* using current climatic variables (red) and LGM past climatic variables (blue). Darker color of the gradient indicates higher suitable habitat for a given climatic period.

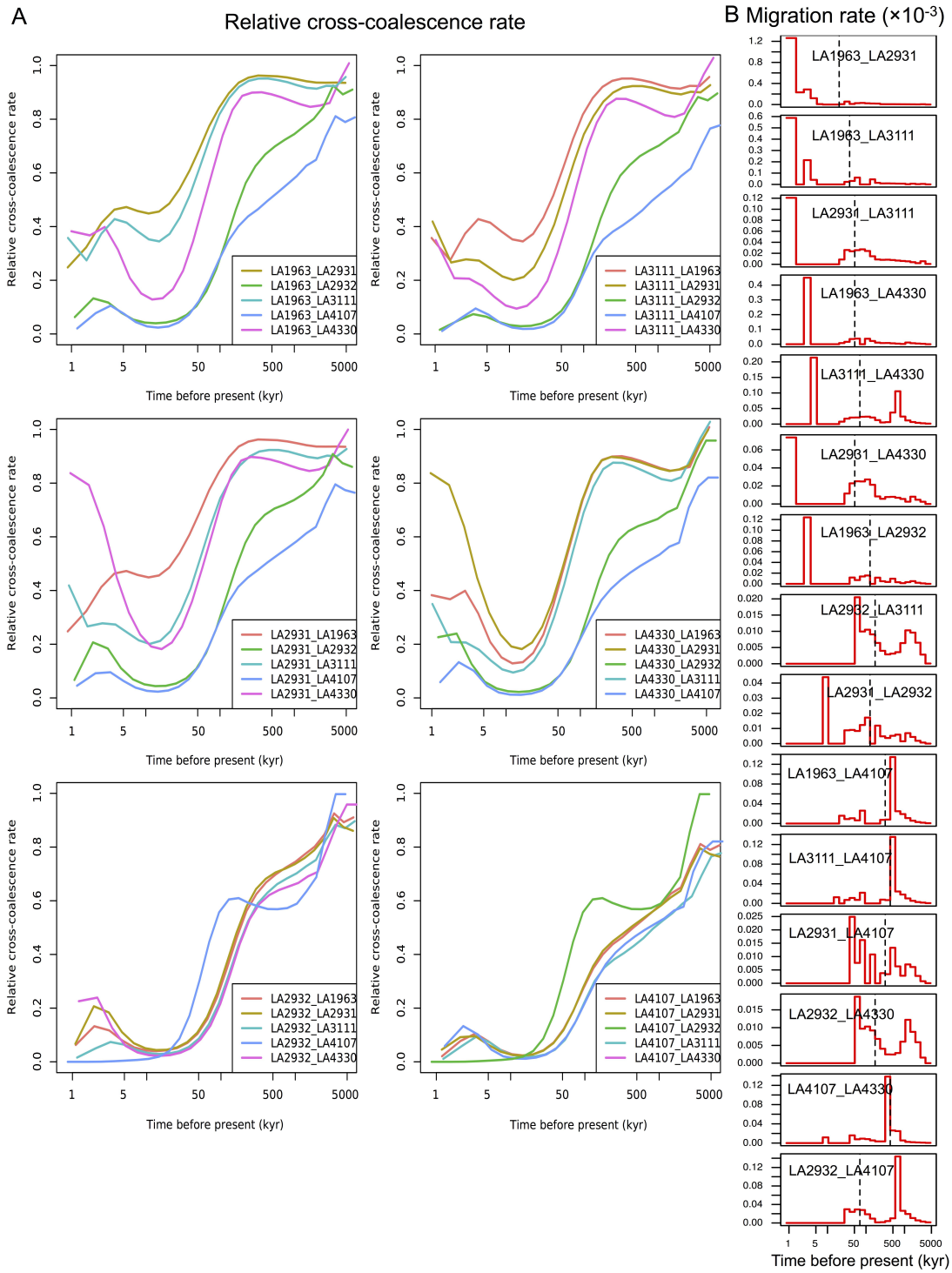

Figure 6. Inference of gene-flow between different populations. (A) Relative cross coalescence rate reveals the age and pace of divergence between different populations. The two populations are well-mixed if the relative cross coalescence rate is 1 and fully separated when the value is 0. (B) Migration rate profiles for pairs of populations. Dashed lines indicate the time of 50% relative cross coalescence rate.

3.2.4 Selective sweeps underpin local adaptation

To detect candidate genes under positive selection we scan the genomes for selective sweep signals using both an SFS-based method (site-frequency spectrum based) SweeD, and an LD-based method (linkage disequilibrium based) OmegaPlus. To disentangle outlier regions putatively under selection from the neutral expected distribution, we set conservative threshold values for the SweeD and OmegaPlus statistics obtained from coalescent simulations with varying rates of recombination based on the results of the demographic inference (Table 1; Figure 7). The overlapping genomic regions under selection between the two methods are then considered as the candidate regions under selection. In total, we find 2,921 regions with SweeD (mean size 212,858bp +/- 3,938) and 13,106 with OmegaPlus (mean size 59,618 bp +/- 521) across all six populations, yielding a total of 520 overlapping regions (mean size 41,082 bp +/- 1,618). Within those regions we find a total of 799 protein-coding candidate genes detected as being likely under positive selection (Data S1C). Among them, the largest number of candidate genes (354) is detected in LA4330 (Table 1), likely because the population 1) has been established recently (Figure 5B), 2) has a lower population size and thus higher amount of linkage disequilibrium so that selective sweeps are easier to detect [166] (Figure 8C), and 3) its habitat is very different from the rest of the species range (Figure 2B). In LA4107, we find 61 candidate genes and about 100 candidate genes were detected in each of the other four populations (Table 1). However, only a few candidate genes are shared among different populations, with the central and south-highland populations sharing a small number of candidate genes, while almost none are shared between the two south-coastal populations (Figure 7C). This lack of common candidate genes among populations is likely due to 1) the high nucleotide diversity and high effective population sizes generating new variants across many genes which are then differentially picked up by selection across different populations, and 2) the relatively old inter-population divergence and timing of local adaptation.

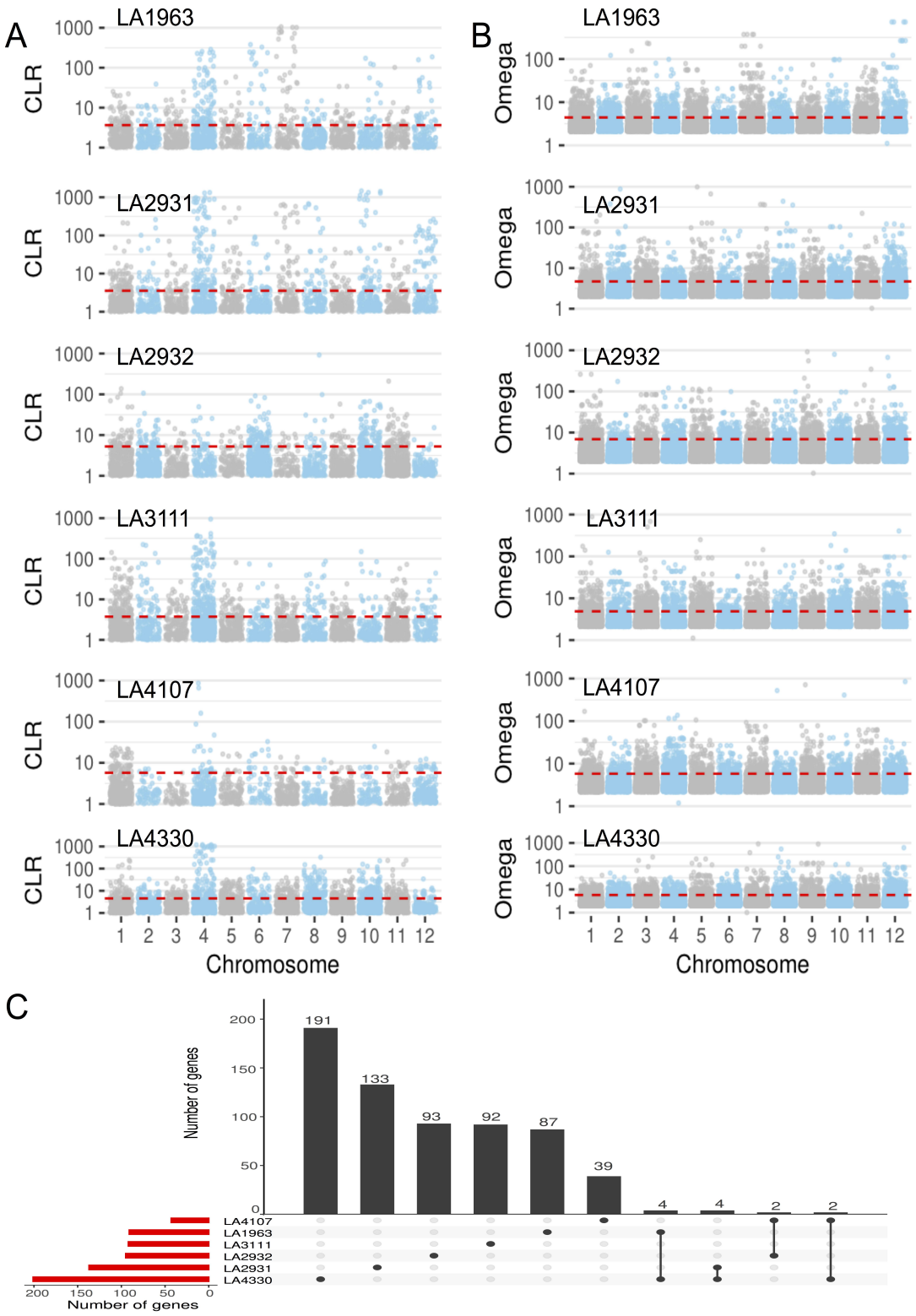

Figure 7. The plot of genome scans among six S. *chilense* populations. (A) The plot of CLR values for six populations using SweeD. (B) The plot of omega values for six populations using OmegaPlus. Dash line denotes cutoff values from neutral simulation. (C) The number of shared candidate genes between different populations. The black points represent shared genes identified in different populations.

Table 1 The summary of genome scans and estimation of sweep age

	Genome Scans				Sweeps age			
Populations	N_{SweeD}	$N_{OmegaPlus}$	$N_{overlaps1}$	N_{genes1}	N_{McSwan}	$N_{overlaps2}$	N_{genes2}	Age_{mean} (kyr)
LA1963	385	2,474	98[b]	86	267	16	14	71±55
LA2931	517	2,268	109	125	355	19	26	62±41
LA2932	374	1,717	46	101	302	15	29	65±48
LA3111	663	2,307	105	107	377	22	22	60±51
LA4107	203	2,047	37	61	194	10	10	63±40
LA4330	779	2,293	125	354	438	39	74	41±30

N_{SweeD}, numbeer of outlier regions from SweeD; $N_{OmegaPlus}$, number of outlier regions from OmegaPlus; $N_{overlaps1}$, number of overlapping regions between SweeD and OmegaPlus; N_{genes1}, number of candidate genes in overlaps1, and all candidate genes show in Data S1C; N_{McSwan}, number of outlier regions from McSwan; $N_{overlaps2}$, number of overlapping regions between McSwan and overlaps1; N_{genes2}, number of genes in overlaps2; Age_{mean}, mean age of overlaps2.

We estimate by simulations the accuracy (statistical power) to detect sweeps due to strong selection under our demographic model to be between 63 and 88% across populations and for each method (Table 2). LA4330 exhibits even relatively high statistical power compared to the other populations (Table 2, Figure 7A-B). This demonstrates that thresholds for sweep detection are stringent and likely minimize the rate of false positives while not detecting all selective sweeps, especially if the selection coefficients are too small (Table 2). Further, only a few candidate genes are shared among different populations, with the central and south-highland populations sharing a small number of candidate genes, while almost none are shared between the two south-coastal populations (Figure 7C). This lack of common candidate genes among populations is likely due to 1) the high effective population sizes (Figure 2) generating new variants across many genes which are then differentially picked up by selection across different populations, and 2) the relatively old inter-population divergence and timing of local adaptation.

Furthermore, we find an overlap between our candidate genes under selective sweep and genes exhibiting signals of positive selection in previous studies in *S. chilense*. It is noteworthy that these previous studies were based on few chosen genes, different plants, different populations and different sample sizes than ours. Among our candidate genes, we indeed find three genes (*JERF*3, *TPP* and *CT*189)

involved in abiotic stress tolerance such as salt, drought or cold [97] as well as three NLRs (nucleotide binding leucine rich repeat, SOLCI006592800, SOLCI001535800, SOLCI005342400) possibly linked to resistance to pathogens [98]. We also find that two of the seven most up-regulated genes under cold conditions in a transcriptomic study of *S. chilense* [107] do appear in our selection scan in high altitude populations: *CBF*3 (Solyc03g026270) in LA2931, and *CBF*1 (Solyc03g026280) in LA4330. These overlapping results indicate that our genome-wide selective sweep scan generalizes the previous studies of selection in *S. chilense* and support the functional relevance of our candidate genes.

Table S5 The estimation of sensitivity to detect selective sweeps using our pipeline with four selection strengths

	s=0.01			*s*=0.1			*s*=0.5			*s*=1		
Populations	SweeD	OmegaPlus	McSwan	SweeD	OmegaPlus	McSwan	SweeD	OmegaPlus	McSwan	SweeD	OmegaPlus	McSwan
LA1963	0	0	0	3.6	9.2	4.9	22.4	31.2	29.8	74.2	81.6	76.4
LA2931	0	0	0	7.9	9.6	10.1	27.3	35.1	30.6	76.3	84.1	77.2
LA2932	0	0	0	4.8	11.8	6.7	19.7	27.4	29.4	68.5	78.3	75.3
LA3111	0	0.3	0	6.8	14.4	7.9	26.8	33.5	30.2	74.1	80.5	78.4
LA4107	0	0	0	2.1	8.5	4.9	17.4	21.7	28.7	63.6	74.8	71.7
LA4330	0	0.7	0	10.3	15.7	11.3	30.1	37.9	34.6	81.7	88.3	79.8

The validation of our pipeline to detect selective sweeps by simulating 1,000 data sets of selective sweeps with four different selective strengths (*s*=0.01, 0.1, 0.5, 1) in six populations, respectively. The numbers denote the percentage of the detected simulated sweeps using different methods.

We then estimate the age, that is the time of the appearance of the selected allele, of 112 selective sweep regions (mean size 22,171 bp +/- 3,240) chosen to conservatively overlap between the three detection methods (McSwan, SweeD and OmegaPlus, Table 1). These regions contain 175 genes and exhibit a mean sweep age of ca. 28,000 years. The ages of sweeps range from as early as 65 kya up to 2.5 kya (Table 2, Figure 8). The highland populations exhibit more recent sweeps (2.5 - 35 kya) than those at the coastal populations (2.5 - 35 kya), consistent with the recent (re)colonization of higher altitudes (Figure 8). The south-coastal populations exhibit

older and large distributions of sweep age consistent with older events of colonization (2.5 - 65 kya). Regarding the key gene networks of relevance for local adaptation highlighted above (root hair, protein lipidation, vernalization and photoperiod), each of them exhibits a narrow range of sweep age values across several populations (Figure 5). The averages of sweep ages observed (Table 1) are perfectly in line with the estimates obtained from the sweep simulations under our demographic model (Table 2). This demonstrates that our statistical power is adequate to estimate sweep ages under the demographic model, and that old sweeps in the highland populations cannot be recovered (even if they occurred) by contrast to the coastal populations.

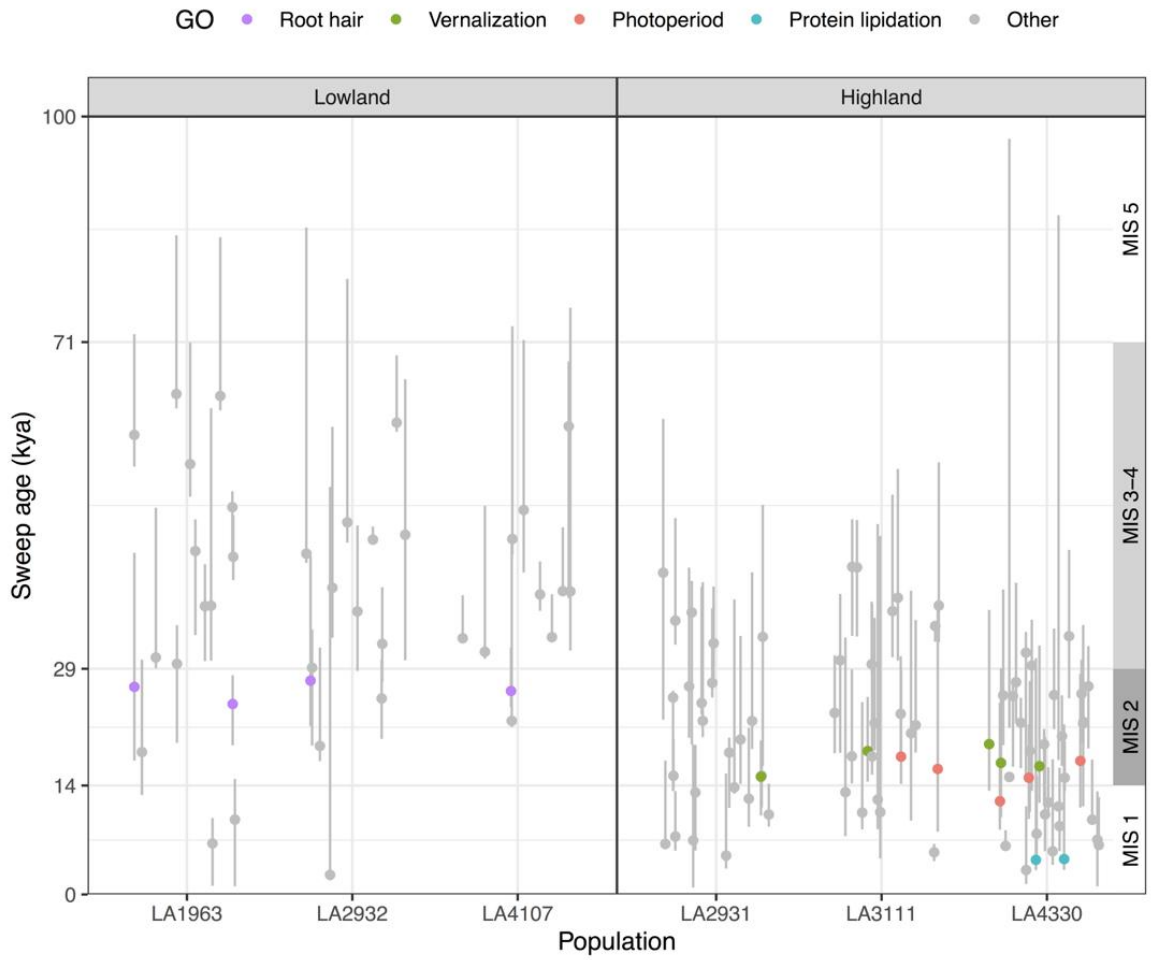

Figure 8. Distribution of sweep age across the five MIS climatic periods.

3.2.5 Statistics and confidence in the genome scans for positive selection

Our candidate regions exhibit typical characteristics of positively selected regions when compared to the genome-wide statistics, that is, lower genetic diversity and high level of nucleotide differentiation (Figure 9, Table S3 and S4). First, genetic diversity (π and θ_w) of almost all candidate regions is smaller than the whole-genome level (Mann-Whitney test, $P < 0.01$, Figure 8A, Table S3), but only so in LA4107 probably due to the generally lower genetic diversity at the genome level in this population. Second, significantly lower and more negative values of both Tajima's D and Fay and Wu's H are calculated in the candidate regions compared to genome-wide values (Mann-Whitney test, $P < 0.01$, Figure 9A; Table S3). Third, the pairwise F_{ST} values between populations are higher for the candidate regions than for the whole-genome patterns, and this difference increases with the increasing degree of differentiation (Figure 9B; Table S4). This also suggests that in the candidate regions there is a close relationship between positive selection and population divergence, namely the hallmark for local adaptation [170]. These comparative analyses suggest that our candidate regions exhibited significant differences with the genomic background levels of diversity and population differentiation, as expected under local adaptation due to positive selection.

As recombination rates can vary across the genome resulting in potential bias in the genome scans using window-based approaches, we use linkage disequilibrium (LD, r^2) as an estimator of the recombination rate in different populations. Our results indicate the LD and LD decay in the central populations to be lower than the south-coast and south-highland populations (Table S3; Figure 9C). While, high LD levels are observed in the candidate regions in different populations, the distribution of LD values follow that of the whole-genome background level (Table S3; Figure 9D). This means that candidate regions show a lower recombination rate than the genome average in our study, as expected as candidate regions usually show low rates of recombination [171, 172]. However, many recent studies have pointed out that if the

recombination rate is not considered in different regions of the genome for genome scans, false positives will occur in low recombination regions [153]. In order to reduce impact of the recombination rate on sweep detection and reduce the rates of false positives, we 1) set a varying recombination rate when generating the neutral simulations (see methods), and 2) choose conservatively to use the maximum SweeD or Omegaplus values from neutral simulations to define high threshold (cutoff) values.

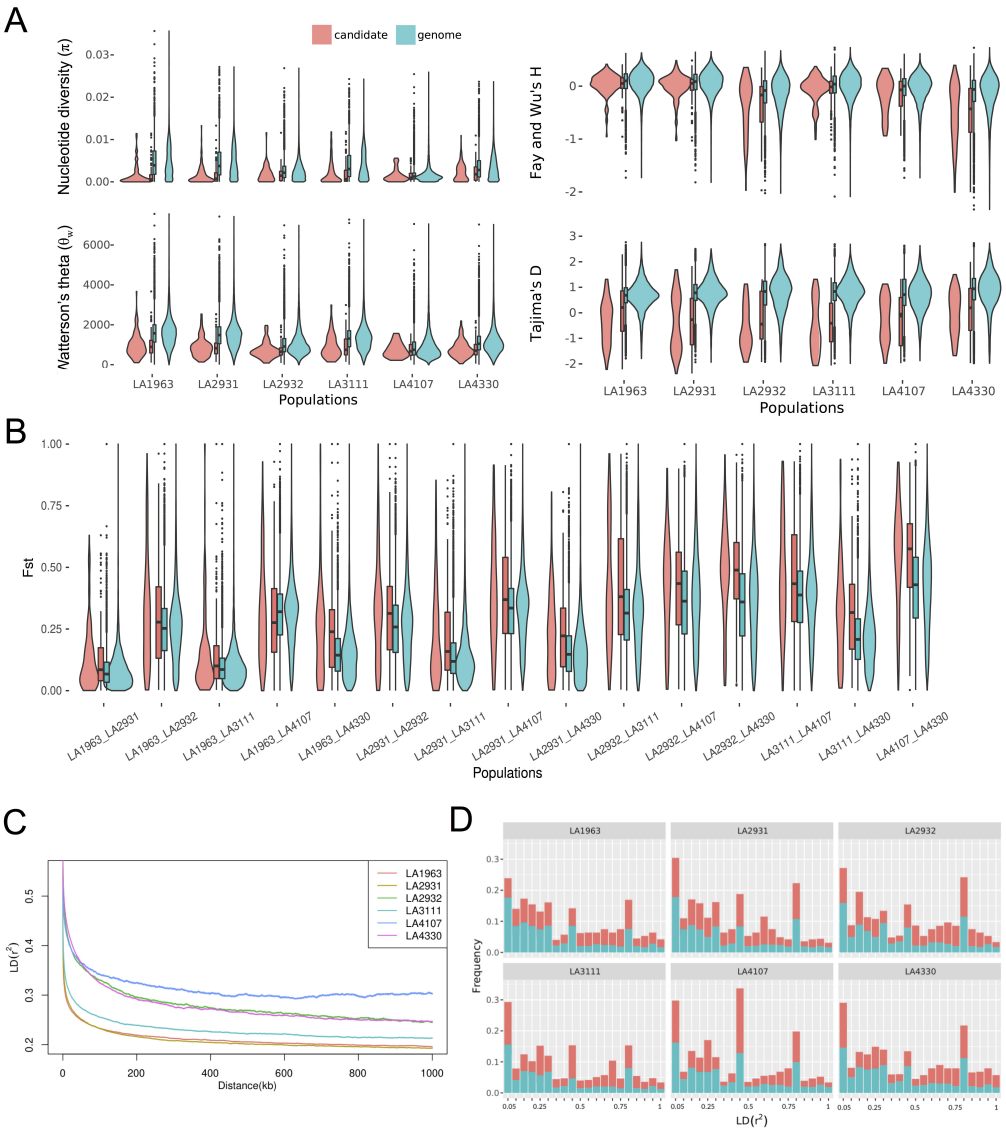

Figure 9. The comparison of statistics between whole-genome and candidate regions by 100 kb sliding windows. (A) The distribution of statistics of genetic diversity (π), Watterson's theta (θ_w), Fay and Wu's H, and Tajima's D. (B) The distribution of F_{ST} of whole-genome and candidate regionss. (C) LD decay against the genetic distance for pairs of linked SNP in 1000 kb distance. (D) The distribution LD of whole-genome (blue) and candidate regions (red) in 100 kb windows. Turquoise denotes whole-genome and red denotes candidate regions.

3.2.6 Gene regulatory networks underlying local adaptation in *S. chilense*

A Gene Ontology (GO) enrichment analysis of the 799 candidate genes reveals significantly different and common biological processes across populations. We find common GO categories to all populations such as basic cell metabolism, immune response, specific organ development and response to external stimuli (Figure 10). Most interesting, however, are four GO categories specifically found in several populations and link to local differences in habitats: (i) root hair cell differentiation functions are enriched in 15 candidate genes, only in the three coastal populations (LA1963, LA2932 and LA4107); (ii) response to circadian rhythm, photoperiodicity and flowering time are enriched in 12 candidate genes in two high-altitude (LA3111 and LA4330) and a south-coast (LA2932) populations; (iii) vernalization response is enriched in eight candidate genes in the three high-altitude populations (LA2931, LA3111, LA4330), and (iv) protein lipidation is enriched in seven candidate genes in the south-highland population (LA4330). Based on the wealth of available data in cultivated tomato and *S. pennellii* (with a well annotated genome) as well as gene ontology in *A. thaliana*, we further study the gene regulatory networks to which the candidate genes belong.

Figure 10. Gene Ontology (GO) analysis in candidate genes under positive selection enriched to the biological process in different populations. The color bar shows gene ratio in candidate genes, the gray boxes are empty and indicate that the biological process is not enriched.

For adaptation to high-altitude conditions, 15 candidate genes are found interconnected in a flowering gene network, which is itself sub-divided into two sub-networks related to flowering, photoperiod and vernalization control pathways (Figure 11A; Data S1D). Photoperiod responsive genes can sense changes in sunlight and affect the circadian rhythm to regulate plant flowering [173, 174], while vernalization genes regulate the flowering and germination through long-term low temperature [87, 175, 176]. These two sub-networks are connected through several key genes, some of which appearing as candidate genes entailing local adaptation in our populations: FL FLOWERING LOCUS C (FLC or AGL25), FLOWERING LOCUS T (FT) and AGAMOUS-LIKE genes (AGL) (Figure 11A and 11B). These key genes are essential regulators acting on the flowering regulation pathway [177-180]. Remarkably, some candidate genes in the recently diverged south-highland population (LA4330) aggregate into an independent network involved in circadian rhythm regulation, connected to the photoperiod network by JUMONJI DOMAIN CONTAINING 5 (JMJD5) also a candidate gene in LA3111 (Figure 11A). In the central-highland population (LA3111) several other candidate genes of the photoperiod network also regulate circadian rhythm and flowering time. Interestingly, the three high-altitude populations (LA3111, LA2931, and LA4330) present candidate genes of the AGAMOUS-LIKE (AGL) gene family in the vernalization network (Figure 11A). We also note that the network of protein lipidation genes appears to be related to the synthesis of fatty acids in the south-highland population (Figure 11D; Data S1D). We speculate here that this latter adaptation may be related to adaptation to lowest-temperature stress of LA4330 (Figure S3A; Data S1E) [181, 182]. Adaptation to high altitude involves the regulation of the flowering, including photoperiod and vernalization pathways, but through different genes in different populations, while cold stress and its consequence (adaptation in lipidation pathway) may be relevant for adaptation to the highest altitudes.

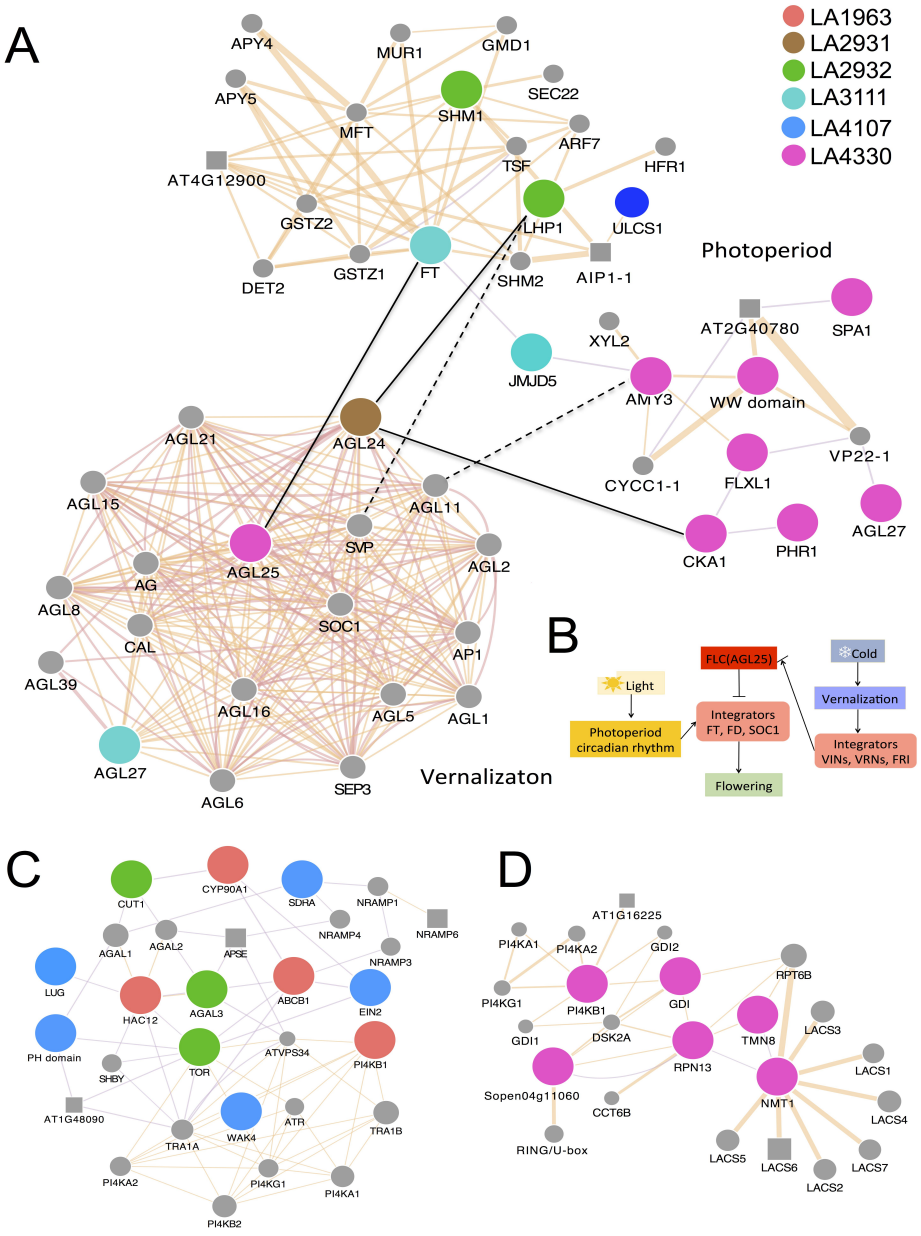

Figure 11. Interaction genetic networks of candidate genes. (A) The network of flowering regulation involved two sub-networks, photoperiod and vernalization pathways, for regulation of flowering. (B) The schematic diagram of flowering regulation involved photoperiod and vernalization is adapted from [87]. "┤" indicates repressive effects on gene expression; "→" indicates promotive effects on gene expression. (C) The network of root development and cell homeostasis. (D) The networks of protein lipidation. Connections represent gene interactions based on physical interactions, informatics predictions and co-expression analyses. Connection thickness is proportional to weighted value of the connected genes. The black lines connected two sub-networks, genes under selection were connected by solid lines and other genes were connected by dashed line. Node colors correspond to genes were detected the different populations in genome scans. Gray circles, not detected in genome scan, but present in *S. chilense*; gray squares: not present in *S. chilense*.

Regarding adaptation to coastal conditions, we find 11 candidate genes related to root development and cellular homeostasis functions clustered in a single network (Figure 11C; Data S1D). We speculate that the drought and water shortage typical of the coastal conditions (Figure 1C and 1D; Data S1E) would promote the differentiation and extension of plant roots [183, 184]. The cell WALL ASSOCIATED LINASE 4 (WAK4), a candidate gene identified LA4107, is shown to act as linker of signal from the cell wall to plasma membrane and thus serve a vital role in lateral root development [185, 186]. It is also possible that the soft soil at the seaside is also beneficial to the extension of root hairs to increase the water absorption area [187, 188]. In addition to root development, we find also genes involved in cell homeostasis (Figure 11C; Data S1D), which would be critical for the coastal drought and salinity conditions to maintain stability of intracellular environment in the coastal habitats [189, 190].

3.2.7 Candidate genes show genotype-environment associations to local climatic conditions

Our candidate loci are hypothesized to be responsible for adaptation to local climatic conditions, so we test for genotype-environment association (GEA) using redundancy analysis (RDA). RDA tests for correlation between allele frequencies and climatic variables and can be used 1) to search more finely for specific selective variants within the genomic sweep regions, and 2) reveal selective gradients defined by multiple environmental variables [163, 191]. We here perform first a “present day” RDA using 144,713 SNPs from all candidate regions and 63 climatic variables representing current (present) conditions for temperature, precipitation, solar radiation, and wind (Data S1E). We find that the two-first RDA axes are significant (ANOVA's $p < 0.001$) and retain most (38% and 21%) of the putative adaptive genetic variance identified in the genome scans in all populations (Figure 12A). Table S5 and S6 summarize outlier SNPs in different RDA models and their correlation with climatic variables. In concordance to the PCAs of both climatic and genomic variation (Figures

2B and 3C), the two main RDA axes cluster the individuals into three groups corresponding to the main geographical regions (central, south-highland, and south-coastal) supporting that those axes synthetize the principal selective pressures for spatial local adaptation along the species distribution (Figure 12A). RDA1 represents the differentiation of the two south-coast populations in correlation with higher precipitation of coldest quarter (Bio19) and annual variation of solar radiation (CV_R). There are 480 and 501 SNPs strongly associated with these two climatic variables, respectively (Table S6). RDA2 summarizes a climatic gradient differentiating the south-highland population manly driven by annual potential evapotranspiration (annualPET) and temperature annual range (Bio7) with 1,184 and 372 strongly associated SNPs, respectively (Table S6).

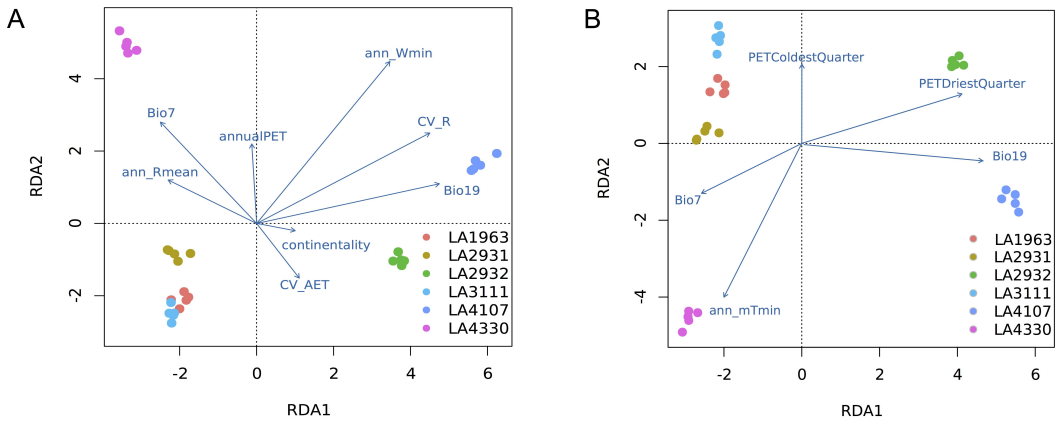

Figure 12. Redundancy analysis (RDA) ordination bi-plots between the climatic variables, populations and the genetic variants in all candidate sweeps. (A) RDA using current climatic variables. (B) RDA using LGM climatic variables. Arrows indicated the direction and magnitude of variables correlated with the populations. The abbreviations of climatic variables are shown in Data S1E.

Further RDA analyses based on gene variants of the GO categories circadian rhythm-photoperiodism, vernalization, root-hair differentiation, and protein lipidation highlight combinations of climatic variables and genetic variants related to local spatial adaptation (Figure 13A, 13C, 13E and 13G). These analysis show that about 40% of the variation are explained by the two main RDA axes. Climatic variables representing temperature variability through the year such as temperature seasonality (Bio4) and temperature annual range (Bio7) are consistently correlated with adaptive

variation of the south-highland population (Figure 12A; Figure 13). A total of 68 SNPs within candidate genes of the population LA4330 are strongly associated to these two variables (Bio4, Bio7) in three of the RDA based on the GO categories (circadian rhythm-photoperiodism, vernalization, and protein lipidation; Data S1F). The RDA based on the root-hair differentiation GO category shows a strong differentiation between lowland and highland populations, with stronger correlation for the southern populations (Figure 13E). This differentiation correlates with atmosphere water vapor availability variables such as annual minimum water vapor pressure (ann_Vmin) and annual actual evapotranspiration (ann_AET) variables for the coastal populations.

Considering that the sweep age estimations indicate that a fraction of the adaptive variation currently observed in *S. chilense* may have originated from the influence of past climate change processes (Figure 8), we implement an “LGM” RDA using 37 climate variables projected to the Last Glacial Maximum conditions (Figure 5B; Figure 13B, 13D, 13F and 13H; Data S1E). This analysis implies the assumption that the species did not shift its geographical range following the late Pleistocene climatic oscillations, but rather underwent a shift in its niche while retaining large portions of its distribution [192, 193]. This process is especially relevant for the highland regions where past distribution projections show no suitable habitat, while population persistence being likely since the demographic reconstruction indicates 1) only mild population size reductions (Figure 5C), and 2) pre-LGM divergence of the highland populations (LA3111 and LA2931 during MIS3-4 climatic periods, Figure 5B). Therefore, we expect that this analysis can uncover additional genomic variation selected in response to temporal climatic changes and underlying the niche expansion towards highly differentiated conditions in the south-highland region (Figure 2B; Figure 5C). RDA analyses using LGM climatic variables capture a smaller proportion of the genetic variability in the first two constrained axes (30%) compared to that using the current climatic variables. We find that some outlier/candidate SNPs appear not to show any correlation with the past climatic variables, but about 30% outlier SNPs are identified in additional genomic regions correlated with past climatic

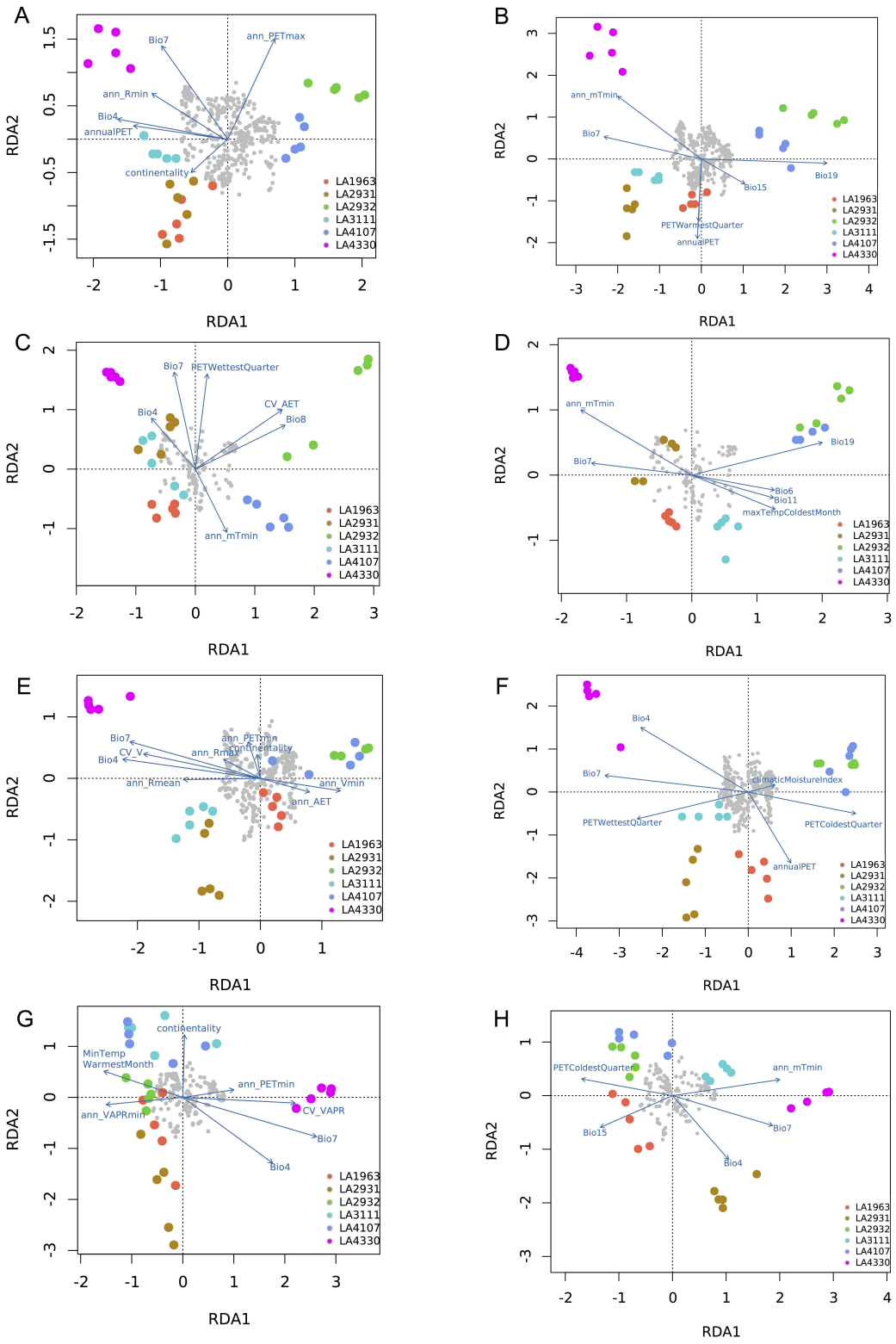

Figure 13. Redundancy analysis (RDA) of SNPs of genes related to four specific GO terms using current (A, C, E, G) and LGM (B, D, F, H) climatic variables. (A) and (B) Circadian rhythm and photoperiodism flowering. (C) and (D) vernalization response. (E) and (F) root hair cell differentiation. (G) and (H) protein lipidation. The color circles denote different populations. Arrows indicated the direction and magnitude of variables. The gray circles denote SNPs.

variables and not with the current variables (Table S5 and S6). The analyses also emphasize the differentiation of central, south-highland, and south-coast populations. Indeed, we find a different RDA for vernalization candidates using the two sets of (present and past) climatic variables. For example, the central populations LA3111 and LA2931 are separated in the past RDA of vernalization using LGM climatic variables indicating warmer climate after LGM may drive gene flow among central populations as seen in the current RDA (Figure 5B; Figure 13C and 13D; Figure 6). The past RDA of LGM climatic variables unveils that high-altitude populations, especially LA4330, have SNPs correlating with temperature (*i.e.* annual mean minimum temperature; ann_mTmin, and temperature annual range; Bio7) whereas coastal populations SNPs do correlate with precipitation and potential evapotranspiration of the coldest and driest seasons (Figure 12B).

3.3 Discussion

By correlating selective sweeps at candidate genes to current and past climatic data, we demonstrate that local spatial adaptation to novel habitats and temporal adaptation to changing conditions are intermingled. Nevertheless, our results show that it is possible to separate the underpinnings of the spatial and temporal adaptation in a species with large effective population size using niche modeling and past climatic reconstruction. As a word of caution, we note that our discussion focuses on four main GO categories which can be reliably associated with physiological traits likely underlying adaptation (though pinpointing the regulatory or coding SNPs under selection was not possible with our sample sizes), while functional information on many candidate genes in Figures 8 and 10 is still lacking to provide a complete picture. Yet, we are confident that our candidate genes under selection are functionally relevant, as demonstrated by the overlap with previous studies focusing on few genes involved in abiotic and biotic stress tolerance and transcriptomic data under cold conditions [97, 98, 107].

For each of the five past climatic periods depicted in Figures 5 and 8, we show different temporally restricted bursts of sweeps for either the south-highland or south-coastal groups of populations. When overlaying specifically the age of the four gene network categories of candidate genes (photoperiod, vernalization, protein lipidation, and root hair differentiation), we find that the sweeps within a given category/gene network are found at the time of a given particular climatic period (Figure 8). This demonstrates that selective sweeps occur following temporal adaptation in response to variable climatic conditions, as suggested under the adaptation to a moving optimum theory [19]. We observe that during the colder and drought climate phases (MIS2 at 30-15kya), the suitable areas of the species decreased to a narrow belt along lower altitude areas (Figure 5C). The persistence of populations at higher altitude would have been possible by niche expansion driven by selection at genes involved in the vernalization gene networks (Figure 8). We also observe a correlation between the age of the sweeps and our “LGM” RDA analysis based on the past climatic data. Therefore, we suggest the photoperiod and protein lipidation adaptations to have occurred in later or after the LGM (Figure 8) and to be instrumental adaptations favoring subsequently the colonization of new habitats in the southern highlands (after the LGM). During this most recent (re)colonization event, sweeps occurred at other genes, including the complete photoperiod gene sub-network in high-altitude populations (Figure 11A). The mentioned adaptation sweeps could have arisen in the ancestral populations from the highland-central region and by gene flow have facilitated the recent expansion of the species' geographic range [194]. In addition, adaptations in the photoperiod and protein lipidation gene networks in the highland populations are adaptations in the later the LGM, we speculate that the populations were already established at high altitude before the LGM (MIS 3-5, Figure 5) likely in the northern part of the range (from the location of LA3111 up to that of LA2931), before a contraction of the species range occurred towards lower altitudes during LGM, and the subsequent divergence of new southern locations concluded 15kya (post-LGM, LA4330). Therefore, we suggest that temporal adaptation does constrain or facilitate spatial local adaptation at later time

steps. The respective contributions of spatial and temporal can be disentangled by comparing the significant genes (and SNPs) found in the past and present RDA analyses.

The *S. chilense* lineage likely originated from coastal habitats in southern Peru. This explains the early divergence and southward colonization process, accompanied by habitat fragmentation and contraction, which yields two highly isolated populations (Figure 5C). The coastal colonization seems to involve fewer sweeps than the adaptation to higher altitudes, for example a burst of selective sweeps in genes related to root anatomical traits during the LGM period (Figures 8, 11C, S6E-F). However, some of the adaptive genomic signals in the coastal populations could be blurred due to the older timing or stronger drift (due to habitat fragmentation along the coast), or be incomplete or partial sweeps which we do not detect with our current scans.

We tentatively suggest some answers to the questions highlighted in the introduction on the basis of our empirical results. We find between 60 and 350 selective sweeps per population, but contrary to our naïve expectations, sweeps have a large variance in age. Indeed, our initial assumption was that most sweeps in the south populations would be recent and occurring during only the recent spatial local adaptation to a new habitat. On the contrary, the resulting picture of local adaptation we observe is that of a complex process involving several discrete adaptive steps, each due to temporally restricted burst of selective sweeps within a given gene network. This adaptive process can be traced back using a combination of sweep age estimates, RDA analyses and past and present climatic models. We suggest that several sweeps do occur concomitantly in a given gene pathway/network at a given time period, possibly as a temporal adaptation to a moving environmental optimum (our climatic periods). Thereafter, these sweeps in a given gene (sub-)network may lay the foundations, in other words, provide exaptation [195, 196], for colonizing new habitats at future climatic periods when new ecological niches/habitats become

available. The enabled future spatial adaptation seems to involve selection at genes in different gene networks (possibly improving upon and fine-tuning the previous phenotypic adaptation). Note that we focus here on selective sweeps resulting from strong positive selection as we cannot assess in our data the occurrence of weaker positive or polygenic selection. It appears that few gene networks seem to be under strong selection at any climatic period, as predicted under polygenic selection model of adaptation to a moving optimum [19, 21, 22, 33]. We speculate further, with the aim to broaden the current theoretical framework, that strong selection underlying temporal adaptation to a moving optimum would occur at one (or few) phenotypic trait, while local spatial adaptation to a new optimum requires strong selection at several traits.

To our knowledge, this is the first study to attempt to disentangle the complex spatial and temporal processes involved in colonization to new highly stressful environments such as hyper-dry high-mountain regions. In the wild tomato *Solanum chilense*, the recruitment of multiple genes belonging to gene networks responsible for regulating the circadian cycle and flowering time is evident. This divergent selection process arises in response to recent climate change processes. The detectability of this polygenic pattern of selection is favored by the recent divergence of the populations involved, persistence of high genetic variability due to mild bottleneck during the expansion process and presence of seed banks, as well as the use of a combination of genome scans and availability of past climatic data. We demonstrate the power to use past climatic reconstruction with scans for selection to detect the underpinnings of local and temporal adaptation, when old selective sweeps can be recovered. Specifically, adaptation in space and time occurs over short time windows during which few genes of a given gene network are strongly selected upon, consistent with a polygenic model of adaptation to a moving optimum.

3.5 Supplementary description of gene network

S. chilense has been proven to be a promising model system to study resistance strategies of plants to survive abiotic stress in extreme environments such as drought, cold, heat and salt tolerance [77, 101, 102, 105, 107, 197-199]. The candidate genes we found in this study enriched biological processes related to above resistance strategies (Figure 10). The population-specific adaptations to local environments we have discovered, mainly related to the adaptation of coastal populations to drought and high-altitude populations to cold. The functions of candidate genes in specific population are consistent with local environments (Figure 11), which are mainly adaptation to changes of temperature in populations distributed along Andean altitudinal gradients (Figure S3A; Data S1E) and drought populations distributed along the coastal of Chile [97, 102, 107].

Flowering is an important fitness trait and there was huge flowering time variation in different conditions for different species, including wild tomato [103, 200]. We revealed that two main regulation pathways for flowering are related to local adaptation, namely photoperiod and vernalization response. They may represent adaptations to solar radiation and low temperature, respectively. The vernalization regulates flowering and appears of interest in the three high-altitude populations. This is consistent with their geographical distribution and the temperature gradually decreasing along the elevation of the Andes (Figure S3A; Data S1E). Some key genes critically involved in flowering, vernalization and photoperiod pathways were detected under positive selection. FLC is a regulatory center of vernalization pathway, which was detected under selection and identified as outlier associated with temperature annual range (Bio7) in GEA. It is regulated by multiple integrator to delay or accelerate flowering in vernalization [201, 202]. High levels of FLC expression repress transcription of the two floral regulatory FT and SUPPRESSOR OF OVEREXPRESSION OF CO 1 (SOC1) (Figure 11B), resulting in delayed flowering [177, 203, 204]. The variants of FLC were detected in coding regions and showed strongest correlation with temperature changes. These variants in coding regions may

then change the activity of FLC to regulate flowering of *S. chilense*. In addition, multiple AGAMOUS-Like genes were detected in different populations and sub-network of vernalization, respectively (Figure 11A). These genes contain two similar domains (MADS-box and K-box) and may play similar function in flowering in different populations. In general, the regulatory pathway of flowering mainly responds to environmental stresses of high-altitude populations, especially in LA3111 and LA4330.

The photoperiod is another main regulatory pathway for flowering. The changes in external light time and strength can change the circadian rhythm of plants to control the flowering time. This is not surprising for high-altitude populations, because they are closer to the sun. But the candidate genes of LA2932 were enriched for circadian rhythm and photoperiodism flowering in our GO analysis, and these genes are also present in the photoperiod sub-network (Figure 11A), such as SERINE HYDROXYMETHYLTRANSFERASE 1 (SHM1). But SHM1 is correlated with temperature (Bio7), this may be an indirect effect because the sunlight also cause temperature changes. The climatic data showed that despite the low altitude of the habitat of LA2932, its value of solar radiation is high and equivalent to south-highland population (Figure S3C; Data S1E). This indicated that some coastal populations also regulate flowering through the photoperiod pathway to possible adaptation to strong solar radiation. In photoperiod pathway, FT, a major downstream integrator of inductive photoperiod, can be activated by multiple upstream genes to promote flowering [205, 206], and ALPHA-AMYLASE-LIKE 3 (AMY3) also was detected, which expression strictly follows the circadian rhythm [207].

Cold condition is a challenge for the south-highland populations. The system of cellular membrane is very important for sensing and resisting cold stress [208]. Some proteins in our networks are integral component of the membrane through the modification of the N-terminal. The GDP DISSOCIATION INHIBITOR family protein (GDI) identified in south-highland population performs a function of protein transport

in the cells, which can transport and help proteins to locate to the membrane [209]. The outlier gene TRANSMEMBRANE 9 SUPERFAMILY MEMBER 8 (TMN8) is a protein location to membrane [210] and showed interaction with GDI in the network (Figure 11D).

Unlike the high-altitude populations, the three coastal populations mainly showed the characteristics of adaptation to drought. Their habitat has experienced the challenges of high temperature and lack of precipitation for a long time. Although studies have shown that coastal populations are greatly affected by the El Niño climate and rainfall is frequent, this seems to be a short-term effect [211-213]. The climatic data indicated their precipitation is much less than the high-altitude populations (Figure S3B; Data S1E). This is also observed in annual actual evapotranspiration (ann_AET), because lack of precipitation led to small actual evapotranspiration (figure S3D; Data S1E). The root hair development is an import strategy for resistance to extreme drought [214-216]. Our results also revealed that SNPs in TOR, a key gene for growth of multiple organs including root, are strongly correlated with actual evapotranspiration, which integrates multiple pathways such as nutrient, energy, hormone, growth factor, and environmental inputs, controlling cell proliferation, growth, and metabolism [184, 217, 218]. Glucose and light/auxin act as upstream signals to precisely regulate the activity of TOR kinase through different pathways to regulate the activity of plant root meristems [183, 184]. And a new study shows the evolution of gravitropism and growth habit regulated by auxins in sand (erect) and rocky (prostrate) land [219]. Interestingly, in *S. chilense* we also found contrasting growth habits, especially between coastal (prostrate) and highland (erect) populations. This is also related to the differentiation and direction of the root system. We also found that some genes are related to cell homeostasis in the same network as for root hair development, suggesting that maintaining cell homeostasis may also be one of the strategies for survival under extreme drought conditions. And cell homeostasis is also necessary for the survival of plants under salt stress [220, 221]. The habitats of the south-coast populations are also characterized by salt stress. The

cytochrome P450 monooxygenase (CYP90A1 or CPD) under selection and found in the network of root development detected in LA1963 (Figure 11C) is an essential gene for homeostasis of brassinosteroids (BRs) and normal growth and development in higher plants [222]. CYP90A1 can maintain BR homeostasis through feedback expressions [223, 224] and showed differential expression under salt stress in *S. chilense* in previous study [77].

4 Evolution of gene networks involved in drought tolerance

4.1 Materials and Methods

4.1.1 Acquisition of transcriptome data and processing of the sequencing reads

Transcriptome sequence data were obtained from Saida Sharifova's experiment of drought stress [151]. Seeds of population LA1963 of *S. chilense* were sown and grown in controlled conditions for 23 days (22℃ day/20℃ night, 16h light/8h dark photoperiod). On the 24th days, all plants were separated into two groups. First group of plants were maintained under normal watering condition, second group of plants were imposed water stress. Then the new leaves and shoot apices were selected for RNA extraction (the details were describe in [151]). Finally, 16 samples were sequenced on Illumina HiSeq4000.

The adapters were removed from raw reads by two consecutive rounds using BBDuk in BBTools v38.90 [225]. Two sets of parameters were used in two rounds respectively: first round 'ktrim=r k=21 mink=11 hdist=2 tpe tbo minlength=21 trimpolya=4'; second round 'ktrim=r k=19 mink=9 hdist=1 tpe tbo minlength=21 trimpolya=4'. Then Low-quality reads were also removed with BBDuk using parameters 'k=31 hdist=1 qtrim=lr trimq=10 maq=12 minlength=21 maxns=5 ziplevel=5'. The clean reads of each sample were mapped to reference genome using BBMap in BBTools. The SAM files were then converted and sorted to BAM files using Samtools v1.11 [115]. The number of reads were mapped to each gene were counted via featureCounts v2.0.1 in each sample [127]. To eliminate the differences between samples, the gene expression level was normalized using the TPM (Transcripts Per Kilobase Million) method (Wagner 2012).

In addition, 5 whole-genome sequence data of population LA1963 were used to perform analysis of population genetics. The processing of the sequencing reads used same pipeline with last chapter, but used new reference of *S. chilense* assembled using population LA3111 by Gustavo Silva.

4.1.2 Relationship analysis of samples

The relationships among samples were performed using TPM values. The correlation coefficient between two samples was calculated to evaluate repeatability between samples using Pearson's test. Principal component analysis (PCA) was performed using the *plotPCA ()* function in DESeq2 R package[226].

4.1.3 Identification of differentially expressed genes

Differential expression analysis of groups among the different conditions and tissues was performed using the DESeq2 R package. The raw read counts from featureCounts [227] were inputted to detect DEGs. The *P*-value ≤ 0.001, the absolute value of log2FoldChange ≥ 1 and a false discovery rate-adjusted *P* ≤ 0.001 were classified as differentially expressed genes.

4.1.4 Weighted gene correlation network analysis

To identify the gene co-expression networks, weighted gene correlation network analysis (WGCNA) was constructed using TPM values to identify specific modules of co-expressed genes associated with drought stress [228]. We first checked for genes and samples with too many missing values using *goodSamplesGenes()* function in WGCNA R package. We removed then the offending genes (the last statement returns 'FALSE'). To construct an approximate scale-free network, a soft thresholding power of 5 was used to calculate adjacency matrix for a signed co-expression network. Topological overlap matrix (TOM) and dynamic-cut tree algorithm were used to extract network modules. We used a minimum module size of 30 genes for the initial network construction and merged similar modules exhibiting >75% similarity. To discover modules of significantly drought-related, module eigengenes were used to calculate correlation with samples.

4.1.5 Identification of transcript factor families and transcript factor binding sites

The protein sequences were obtained applying an open source program GffRead (https://github.com/gpertea/gffread) and were used to identify TF families using online tool PlantTFDB v5.0 [229]. Furthermore, the upstream 2000 bp sequences of the transcription start sites (TSS) were extracted as the gene promoter from the reference genome to detect TFBS. The TFBS dataset of relative species *S. pennellii* was also downloaded from Plant Transcriptional Regulatory Map (PlantRegMap, http://plantregmap.gao-lab.org/) as background of TFBS identification [230]. Then, the TFBS of *S. chilense* was identified using FIMO program in motif-based sequence analysis tools MEME Suit v5.3.2 [231]. The TFBS was extracted with $p < 1e\text{-}5$ and $q < 0.01$.

4.1.6 Construction of phylostratigraphic map

We performed phylostratigraphic analysis based on the following steps. First, the phylostrata (PS) was defined according to the full linkage of *S. chilense* from NCBI taxonomy database. The similar PS was merged and finally 18 PS were generated (Figure 3A). Second, the protein sequences were blast to database of nr (non-redundant) proteins downloaded from NCBI (https://ftp.ncbi.nlm.nih.gov/blast/db/) with a minimum length of 30 amino acids and an E-value below 10^{-6} using blastp v2.9.0 [232]. Third, each gene was assigned to its PS by the following criterion. If no blast hit or only one hit of *S. chilense* with an E-value below 10^{-6} was identified, we assigned the gene to the youngest PS18. When multiple blast hits were identified, we computed lowest common ancestor (LCA) for multiple hits using TaxonKit v0.8.0 [233] and then assigned LCA to specific PS.

4.1.7 Construction of divergence map

We performed divergence stratigraphy analysis to construct sequence divergence map of *S. chilense* using function *divergence_stratigraphy()* of R package 'orthologr' [64] following the steps below. First, the coding sequences for each gene of *S. chilense* and *S. pennellii* (NCBI assembly SPENNV200) were extracted from their reference and annotation file. Second, we identified orthologous gene pairs of both species by choosing the best blast hit for each gene using blastp v2.9.0 [232]. We only considered a gene pair orthologous when the best hit has an E-value below 10^{-6}, the gene pair is considered orthologous; otherwise, it is discarded. Third, codon alignments of the orthologous gene pairs were performed using PAL2NAL [234]. Then, Ka/Ks values of the codon alignments were calculated using Comeron's method [235]. Fourth, all genes were sorted according to Ka/Ks values into discrete deciles, which are called divergence stratum (DS).

4.1.8 Estimation of transcriptome age index and transcriptome divergence index

Domazet-Lošo and Tautz introduced TAI that represents a weighted arithmetic mean of the transcriptome age over all PS using gene expression intensities in different transcriptomes [60]. Analogous to the TAI measure, the TDI was introduced as a measure of average transcriptome divergence and selection pressure in the corresponding sample *[61]*. Here, we calculate TAI and TDI profiles in different samples using *PlotSignature()* function of the 'myTAI' R package and the flat line test was used to estimate statistical significance.

4.1.9 Population genetics analysis

The neucleotide diversity and Tajima's D of population LA963 were calculated using same method with previous analysis (see methods in chapter 3).

4.1.10 Comparative genomic analyses

We identified gene families/clusters between the *S. chilense* and five other plant species, including S. pennellii, S. lycopersicum, S. lycopersicoides, S. tuberosum and Capsicum annuum. The time tree described divergence times among them was directly retrieved from TimeTree database [236]. First, all protein sequences from six species were aligned pairwise using blastp v2.9.0 [232]. Then, the gene families were clustered using mcl v14.137 based on Markov Cluster Algorithm (MCL) method by default parameters [237], and counted number of genes to specific species for each gene family. The input of mcl was generated using mcxload program in mcl tools. CAFE v4.2.1 was used to investigate the dynamic expansion and contraction of gene families by same *lambda* value [238].

4.2 Results

4.2.1 Overall transcriptome and whole-genome sequencing data

We analysed transcriptome and whole-genome sequencing data from 16 and 5 samples of population LA1963, respectively. The 16 transcriptome libraries were constructed and named as follows: leaves under control (or normal) condition (CL-A to D), shoot apices under control condition (CSA-E to H), leaves under drought condition (DL-I to L), shoot apices under drought condition (DSA-M to P). A total of 27,832 genes were identified to express in 16 libraries (Data S2A), of which 1,536 genes were differentially expressed in drought condition and 1,767 genes were differentially expressed in control condition (Data S2A). The SNPs were also identified from whole-genome sequencing data, and a total of 17,589,185 high-quality SNPs passed filtering to perform population genetics analysis.

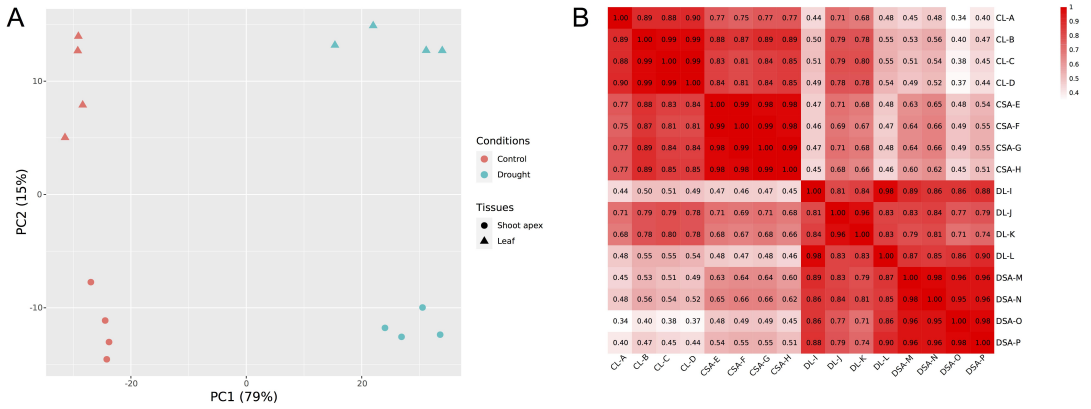

Figure 14. The Relationship of RNA-seq samples. (A) Principal components analysis reveal strong clustering associated with conditions. (B) Heatmap plot o fPearson's correlation coefficient among samples reveal exact drought specificity.

In order to evaluate the relationship between samples, principal components analysis using gene expression profiles revealed strong clustering associated with conditions (Figure 14A). PC1 accounted for 79% and separated the libraries from the two conditions, indicating completely different transcriptomes between drought and control conditions. In addition, samples of different tissues also were separated by PC2 (accounting for 15% of the variance) and suggesting tissue tranascriptome specificity between leaf and shoot apex. Similarly, the libraries transcriptome similarity analysis revealed that the watering condition explains the major differences between treatments (Figure 14B, Pearson's test). Therefore, we focus on comparing the transcriptome of two conditions in this study.

4.2.2 Identification of gene networks to drought tolerance

To obtain reliable gene networks for drought response of *S. chilense*, differential expression analysis and weighted gene co-expression network analysis (WGCNA) were performed. First, three sets of differential expression genes (DEGs) were identified from the three comparison groups (8 control versus 8 drought, each consisting of four leaves and four plant apices) (Figure 15A, Data S2B, fold change > 2, $P < 0.01$). A total of 4,905 DEGs were identified in three comparison groups, of which 2,484 DEGs (1,235 up-regulated expression and 1,249 down-regulated

expressions in drought condition) were shared in all comparison groups (Figure 15B). We deduce that these shared DEGs correspond to functionaly drought responsive network.

Second, 16,181 genes after filtering (see methods) were used in WGCNA, and clustered into 7 co-expression modules named by different colors, with module sizes 183 to 5,364 (Figure 15C, Data S2B). The relationships among modules can be find in Figure S4. Here, we don't directly use DEGs in WGCNA suggested by developer of WGCNA. WGCNA is an unsupervised analysis method, DEGs will lead to a set of correelated genes that will essential from single module and completely invalidate the scale-free topology assumption, so choosing soft thresholding power by scale-free topology fit will fail. Of these modules, the blue module (3,852 genes) shows significantly positive correlation with control condition and negative correlation with drought condition (Figure 15C, Kendall's test, $P < 2.2e-11$). In contrast, the turquoise module (5,364 genes) is significantly positively correlated with drought condition and negatively correlated with control condition (Figure 15C, Kendall's test, $P < 2.34e-13$). In addition, the genes within blue and turquoise modules were observed higher connectivity than other modules ((Kolmogorov-Smirnov test, $P < 2.24e-27$)), indicating a closer co-expression relationship among genes within module related to drought tolerance (Figure 16A).

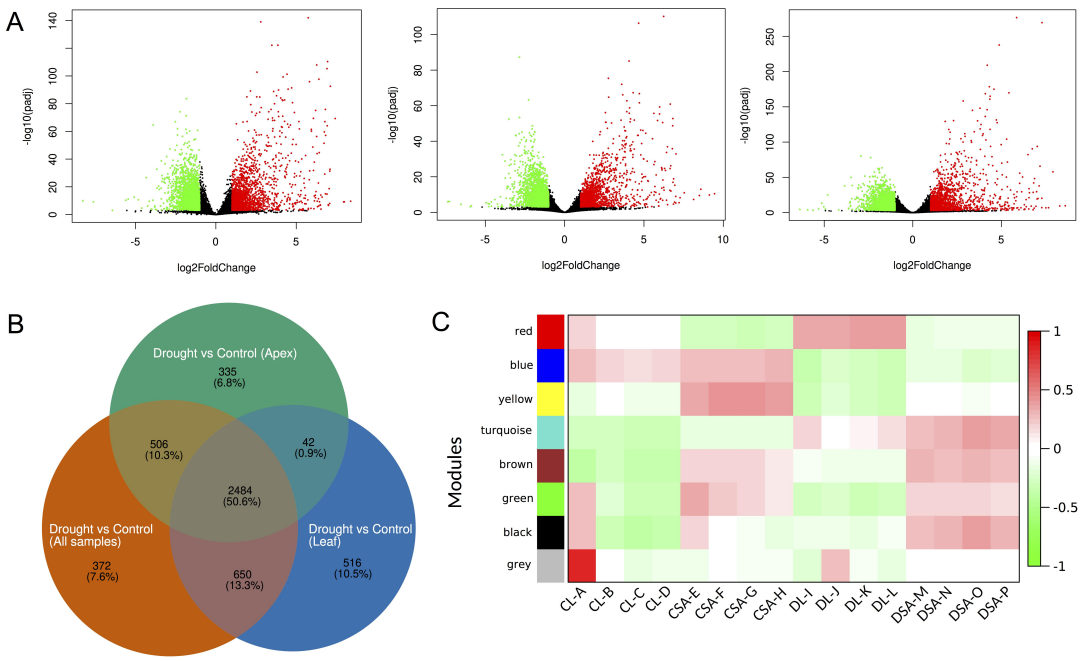

Figure 15. Identification of drought-response networks. (A) DEGs identified from three comparison groups from left to right: 8 control versus 8 drought samples, 4 control leaves versus 4 drought leaves, 4 control shoot apices versus 4 drought shoot apices. Red indicates upregulated genes, and green indicates downregulated genes between control and drought samples. (B) Venn diagram show 2,484 conserved DEGs in three comparison groups. (C) The correlation between samples expression patterns with eight modules.

Table 2 Shared genes between gene sets of DEGs and modules.

Gene set of DEGs	red (474[b])	blue (3,852)	yellow (1,557)	turquoise (5364)	brown (3,666)	green (774)	black (183)
Drought vs Control (all samples: 4,012[a])	6[c]	1,812	113	1,660	90	0	0
Drought vs Control (leaf: 3,692)	49	1,760	112	1,440	200	16	0
Drought vs Control (shoot apex: 3,367)	22	1,487	150	1,457	11	0	5
Drought vs Control (total: 4,905)	56	2,101	192	1,947	225	16	5
Drought vs Control (overlaps: 2,484)	4	1,223	70	1,079	67	0	0

[a]Number of genes in different DEGs sets; [b]number of genes in different co-expressed modules; [c]number of shared genes between DEGs sets and modules.

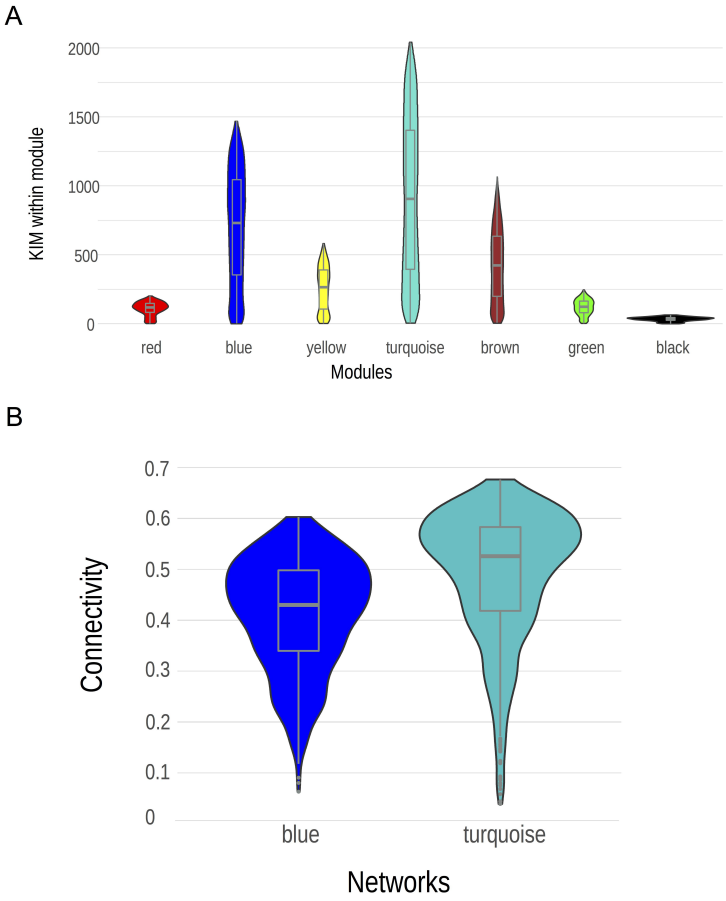

Figure 16. The connectivity of different co-expressed modules. (A) Intramodular connectivity measures (KIM) within module show connection or co-expression of a given gene with respect to the genes of a particular module. (B) The relative connectivity of reconstructive sub-blue and sub-turquoise network. The values were calculated as KIM divided by number of genes within specific module.

We checked also the overlaps between DEGs and modules to confirm that blue and turquoise modules are associated with drought stress in *S. chilense* (Table 2). DEGs shared more genes with blue and turquoise modules than other modules. Almost all shared DEGs (2,302 genes) are overlapped with genes in blue and turquoise modules. This indicates that blue and turquoise modules are two sets of co-expressed drought stress responsive genes. The shared DEGs and module genes were extracted to two subsets from blue and turquoise, 1,223 and 1,079 genes, respectively.

Table 3 Summary of TFs and TFBSs in two networks

	sub-blue		sub-turquoise	
TF family	TF	TFBS	TF	TFBS
BBR-BPC			2	45
bHLH	5[a]	1[b]	9	3
bZIP			8	15
C2H2	1	28		
Dof	1	90	2	361
E2F/DP	2	3		
ERF	2	40	6	158
FAR1			1	1
GATA	1	11	1	30
GRAS	1	39	3	147
HSF			8	1
LBD			1	23
MIKC_MADS			1	32
MYB	3	54	6	37
MYB_related	1	5		
NAC	1	2		
TALE	1	410	3	33
TCP			4	10
Trihelix			4	7

[a]Number of TFs belong to specific TF family; [b]number of genes detected TFBS binding to specific TF in promoter regions.

To construct regulatory relationships among genes, we investigated transcription factors (TFs) and transcription factor binding sites (TFBSs) for the two subsets (Table 3). Then, we extracted genes that have regulatory relationships in Table 3 from two subsets, named as sub-blue (686 genes) and sub-turquoise (948 genes), respectively (Data S2C). The sub-blue and sub-turquoise networks not only show differential expression and co-expression patterns but also can bind with each other to regulatory elements at the sequence level. Finally, the co-expression network was reconstructed using the same steps for genes in sub-blue and sub-turquoise networks. Higher connectivity was observed in sub-turquoise networks (Figure 16B), suggesting a closer regulatory relationship among genes in sub-turquoise than

sub-blue networks. This may be due to more complex TF/TFBS relationships and functions in sub-turquoise networks (Table 3; Table S7).

4.2.3 Functional enrichment analysis of drought-response networks

Although the down-regulated sub-blue is completely different from up-regulated sub-turquoise at the expression level, do they also show functional differences? The functional enrichment revealed that sub-blue network was significantly enriched in biological processes of cell cycle and regulation related processes, including replication and modification of genetic information, ribosome production and assembly, cytoskeleton organization, among others (Figure 17A; Table S7). Conversely, the sub-turquoise network enriched biological processes related to response of physiological and metabolic processes to water shortage and heat, including some metabolic processes, signal pathway, changes of stomata and cuticle, among other processes (Figure 17A; Table S7). The difference of function also determines that genes in two sub-networks work in different cellular compartments. Consistent with the biological process, sub-blue network is mainly enriched in components in the nucleus, including nucleolus, chromosome, nuclear envelope, and is enriched in the ribosome (Figure 17B; Table S8). These cellular components are the center of cell division. The sub-turquoise network is enriched in metabolism related to complexes and membrane structures in the cell. While sub-blue is concentrated in the nucleus, sub-turquoise shows scattered functions in the different cellular compartments. The functional analysis indicated consistency with previous studies of different species that also strong conservation across species in drought-response networks [69, 76]. In addition, the regulation of metabolic and cell cycle processes are the two main strategies for drought response in *S. chilense*. In the following, we refer to sub-blue as the cell-cycle network and sub-turquoise as the metabolic network.

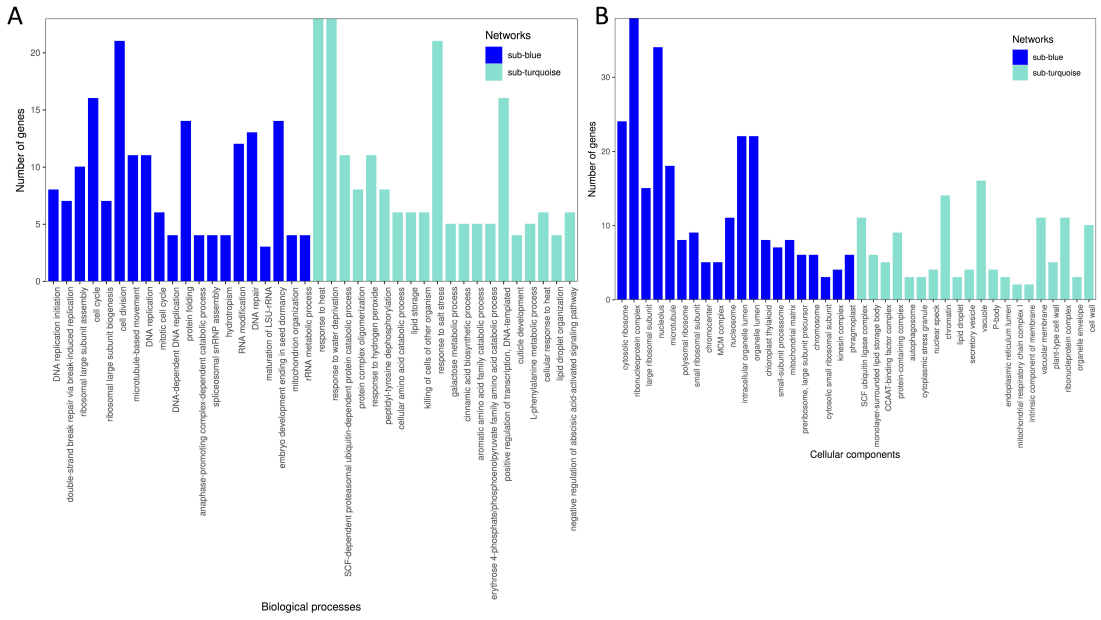

Figure 17. Gene ontology terms enrichment in two drought-response networks, respectively. (A) Top 20 terms of biological process. (B) Top 20 terms of cellular component. The results show two strategies of drought response related to processes of cell cycle and fundamental metabolism in sub-blue and sub-turquoise networks, respectively.

4.2.4 Evolutionary age of drought responsive gene networks in *S. chilense*

Two typical conserved networks were identified as being related to metabolic and cell cycle processes, respectively (Figure 17). Phylostratigraphic analyses can help us to understand the evolutionary period and origin of drought-response genes. We constructed phylostratigraphic maps for cell-cycle and metabolic networks, summarizing the gene emergence in 18 stages of evolution or phylostrata (PS) from PS1 representing the emergence oldest genes (cellular organisms) to PS18 genes originating in *S. chilense*, with no homologue in any other species (Figure 18A and 18B, Data S2D). Most genes in the two analyzed networks were assigned to three main PS, Cellular organisms (PS1), Embryophyta (PS5) and Magnoliopsida (PS8). This suggests that two drought-response gene networks have ancient origin and maintain higher conservation. Additionally, this result indicates that most drought-response pathways could result as exaptations from gene networks formerly involved in core cell process (PS1) and reproductive organ differentiation (PS8).

However, they also show that there is a set of conserved genes specifically related to drought response emerged during the first colonization of plants on land (PS5). Moreover, the cell-cycle network shows older origin with more genes emergence during the PS1-3, while metabolic network present more a larger proportion of genes originated in PS8 (Figure 18A and 18B). Under drought conditions, cell-cycle network genes are down-regulated for almost all PS and up-regulated in the metabolic network (Figure 19).

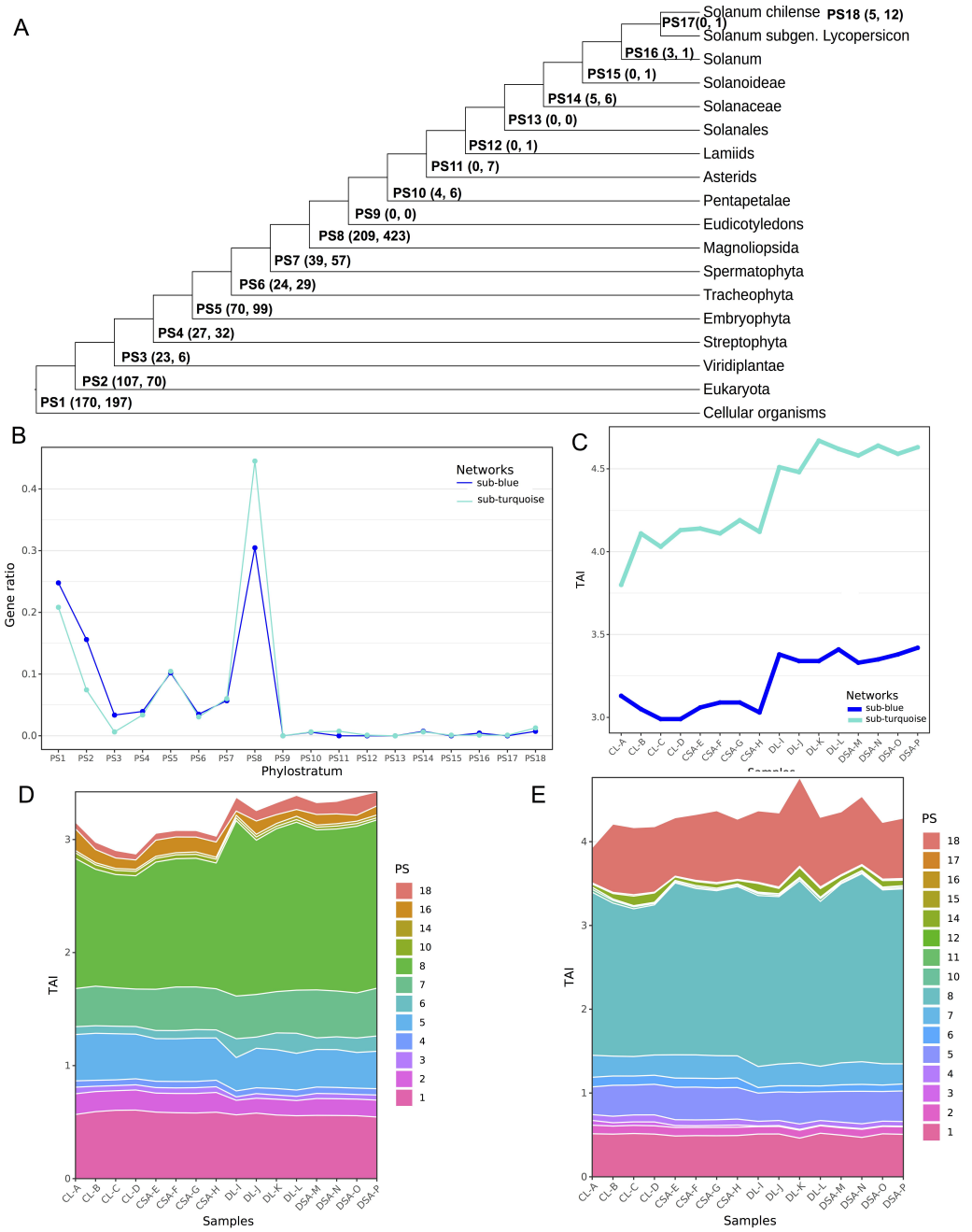

Figure 18. Transcriptome age profiles for two networks. (A) Phylostratigraphic map for two networks.Numbers in parentheses denote the number of genes assigned to each phylostratum (PS) in sub-blue and sub-turquoise network, respectively. (B) Gene ratio in each PS for two networks. (C) TAI profiles of two networks across samples. (D) TAI contributions split according to different PS for sub-blue network. (E) TAI contributions split according to different PS for sub-turquoise network. Sub-blue denotes cell-cycle network and sub-turquoise denotes metabolic network, same blow.

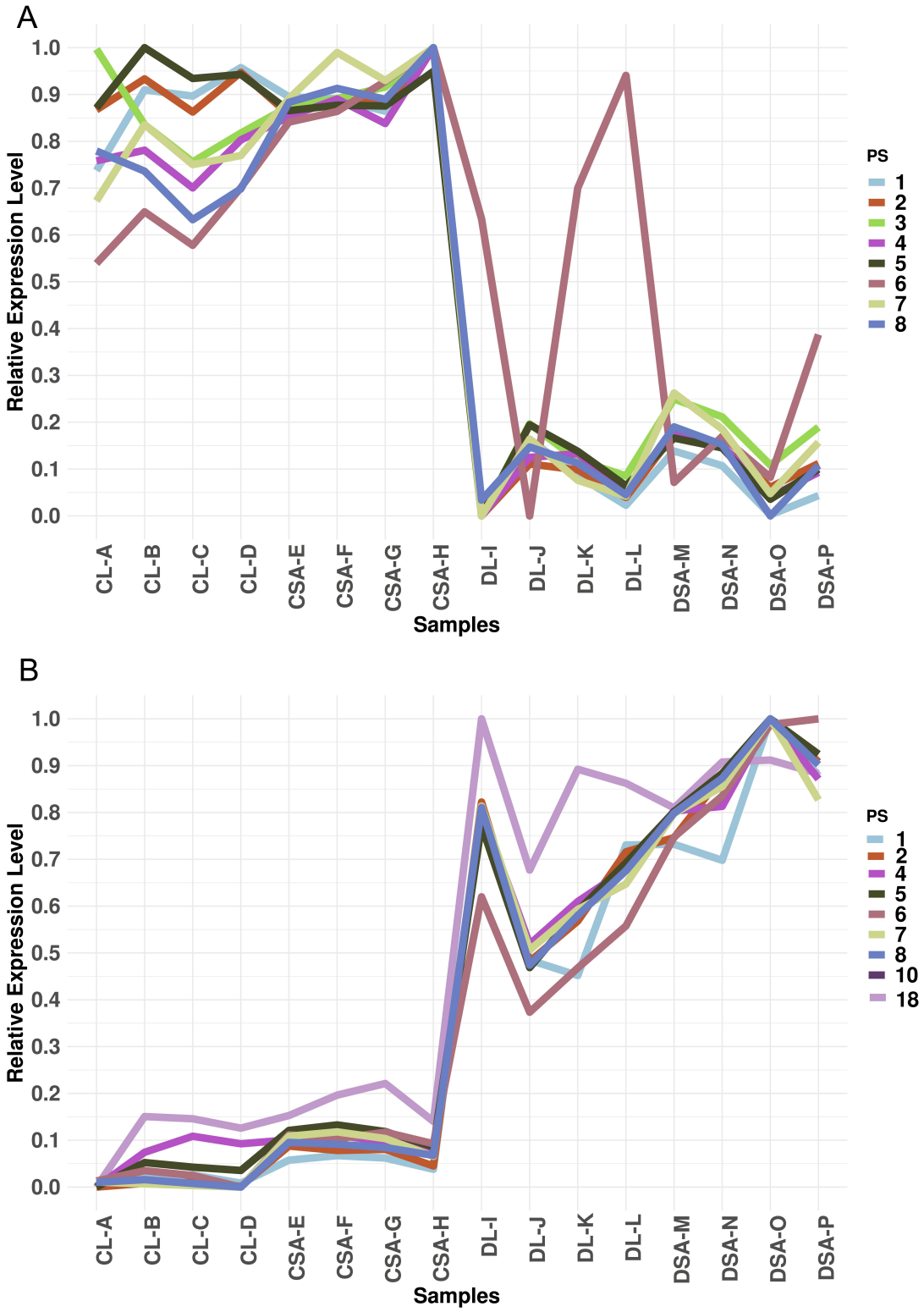

Figure 19. Relative expression level of genes in different PS. (A) down-regulated in sub-blue(cell cycle) network. (B) up-regulated in sub-turquoise (metabolic) network.

To confirm our conjecture, the transcriptome ages of two networks were estimated using TAI. The mean evolutionary ages of the transcriptomes were significantly different among drought and control samples (Figure 18C, flat line test, $P < 0.05$), the TAI profile would be a horizontal line if the same. As we expected, the higher TAI was observed in drought samples, indicating that the drought-response genes exhibit younger transcriptome age in drought samples. Moreover, TAI of the metabolic network is significantly higher than cell-cycle network (Kolmogorov-Smirnov test, $P < 0.01$), suggesting that transcriptome ages of the genes in cell-cycle network are older than the metabolic network.

The contributions of the different PS to the TAI profiles also show notable differences between the two networks analyzed (Figure 18D and 18E). Early divergent genes (PS1 to PS7) are more constantly expressed in all conditions and genes from PS1, PS5 and PS8 are remarkably important in the cell-cycle network, whereas later-emerging genes (PS8 to PS18) contribute increasingly to the differential pattern between control and drought samples, indicating that younger drought-responsive genes are more specific expression under drought stress [60, 63]. The youngest genes PS18 have high contribution in the metabolic network, indicating that younger drought-response genes are more specific expression under drought stress.

4.2.5 Divergence of drought tolerance transcriptome in *S. chilense*

To assess evolutionary patterns in a more recent scale, we calculates the TDI. A total of 10 divergence strata (DS) were constructed based on the sequence divergence between *S. chilense* and *S. pennellii* by computing the Ka/Ks ratio (Figure 20, Data S2D). First, the distributions of the Ka/Ks ratio showed that our drought-responsive genes are under stronger purifying selection pressure, and purifying selection pressure in the cell-cycle network ($Ka/Ks = 0.277 \pm 0.0138$) stronger than in the metabolic network ($Ka/Ks = 0.329 \pm 0.0115$) (Figure 21A; Table 4). Higher TDI values

were observed in transcriptome of drought samples (Figure 21B), suggesting drought-responsive genes have more conserved expression in control samples, whereas more variable transcriptomes are observed in drought samples. This result displays that expression of our drought-responsive genes are more variable and active expression in drought stress. According to the flat line test, we observed that the TDI patterns of two networks deviate significantly from the horizontal line (flat line test, $P < 0.05$), revealing that different selective pressures act on *S. chilense* transcriptomes across conditions. Concordantly to TAI results, we found that transcriptome of the metabolic network is more divergent than sub-blue networks (Figure 21B, Kolmogorov-Smirnov test, $P < 0.01$). However, low TDI in the cell-cycle network, but larger TDI difference between drought and control transcriptomes was found, illustrating that regulation to cell cycle may be an original strategy for drought stress, and tends to be fixed in the transcriptome, because conserved evolution and purifying selection in the recent period. Changes of metabolic processes may be a more rapid strategy to respond to drought stress.

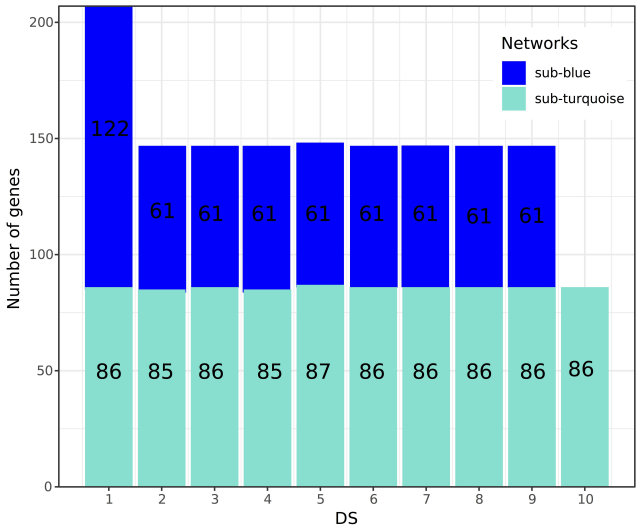

Figure 20. Number of genes in different DS by Ka/Ks ratio from low to high for two networks.

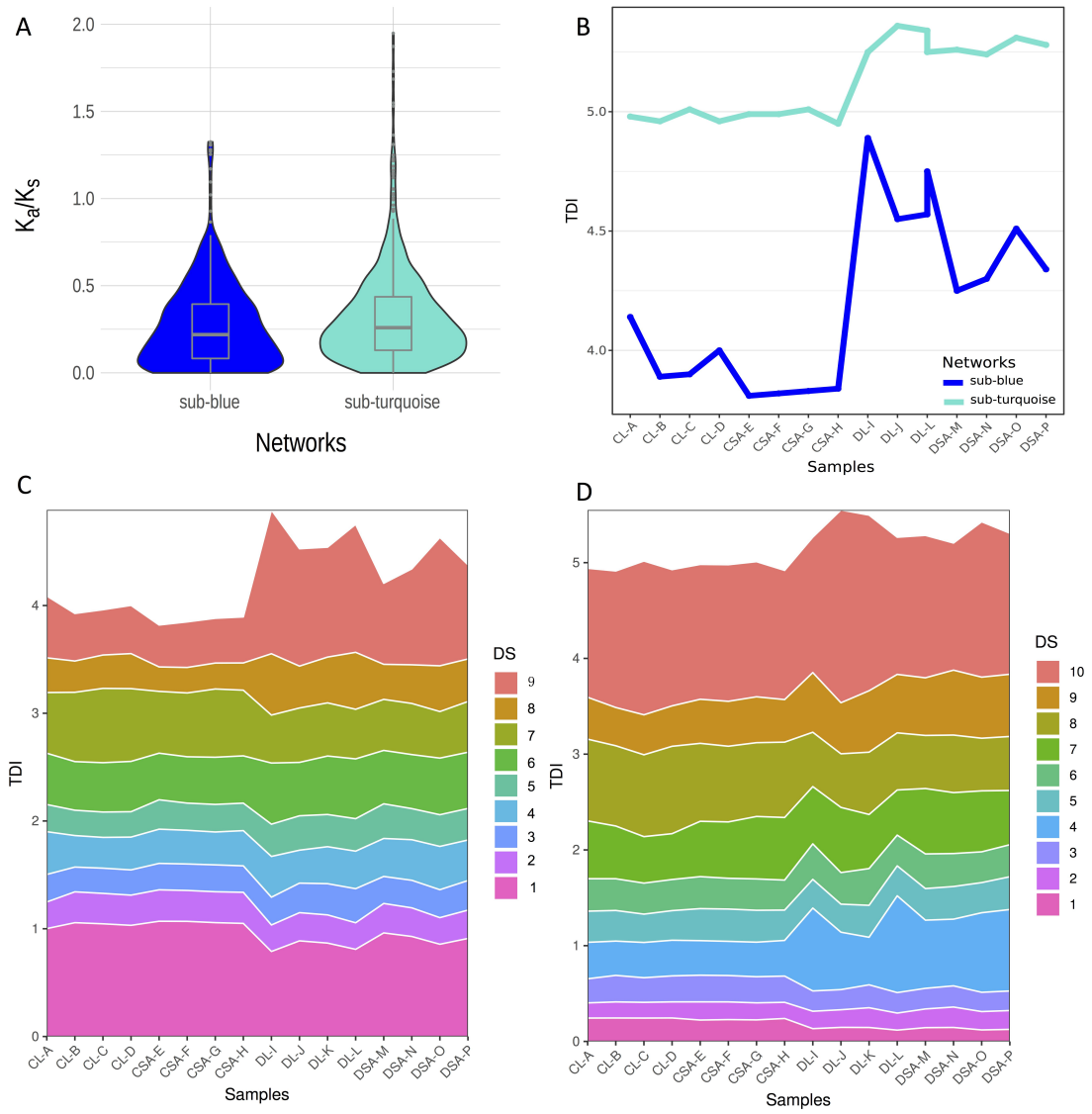

Figure 21. Transcriptome divergence profiles of gene networks. (A) Distribution of Ka/Ks ratio for two networks. (B) TDI profiles of two networks across samples. (C) TDI contributions split according to different divergent stratum (DS) for sub-blue (cell-cycle) network. (D) TDI contributions split according to different DS for sub-turquoise (metabolic) network.

The contributions of low DS (low Ka/Ks in DS1 to DS5) in the cell-cycle network (~ 50%) were larger than the metabolic networks (DS1 to DS5 about 30%), especially in DS1 (lowest Ka/Ks rario, Figure 21C and 21D). This indicates that purifying selection pressure is acting on genes of the cell-cycle network constraining further changes. In contrast to the cell-cycle genes, metabolic networks show about 70% contributions in high DS (higher Ka/Ks ratio in DS6 to DS10), especially in DS10 (highest Ka/Ks ratio), indicating genes in the metabolic network evolve under weaker

purifying selection pressure than cell-cycle network.

The TAI profiles display evolutionary process over a long span of time along the tree of life, while TDI profiles provide an estimation of sequence divergence and selective pressure drafts in two networks. Based on the above analysis, the genes of the cell-cycle network are evolutionary older and higher conserved, whereas genes in the metabolic network are younger and more divergent and have higher evolutionary potential in the future.

4.2.6 Population genetics analysis of drought-response networks

To analyze the highly divergent genes, specially new genes in *S. chilense* lacking homologous genes in other species, we calculated nucleotide diversity (π) and Tajima's D to detect selection signals due to drought adaptation using whole-genome sequencing data. Mean nucleotide diversity (π) of the cell-cycle network genes is higher than the metabolic network (Figure 22A; Table 4), consistent with long-term analysis of evolutionary age. In addition, we found that the π of the promoter regions (upstream 2kb of transcription start site) is significantly higher than the gene regions (Figure 22A; Table 4; Kolmogorov-Smirnov test, $P < 0.01$). This result can explain why few TFs can bind to multiple genes in the regulatory network. TFs are highly conservative, especially in regions of functional domains, but higher polymorphism of TFBSs in the promoter can satisfy complex and diverse regulation to different conditions. We did not find the difference in the promoter regions between two networks (Figure 22A; Table 4). Additional indications of within species selective pressure were obtained with Tajima's D. The metabolic network genes show lower values than cell-cycle network, indicating stronger positive selection pressure on metabolic network than cell-cycle network either in the gene or the promoter regions concordant with ongoing evolution of genes in metabolic network in recent time (Figure 22B; Table 4). however, Tajima's D is not correlated with Ka/Ks ratio (Figure 23A and 23B), displaying selections within *S. chilense* and across *S. chilense* - *S.*

pennellii occur in different periods.

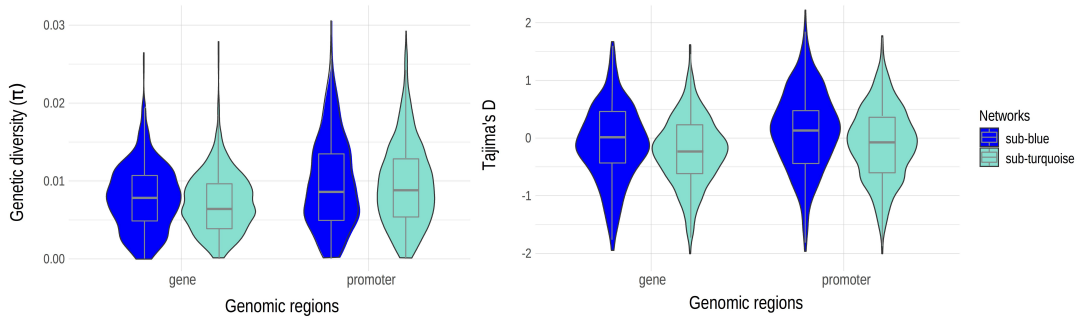

Figure 22. Statistics of population genetics. (A) Nucleotide diversity (π) of genes and promoters in two networks. (B) Tajima's D of genes and promoters in two networks.

Table 4 summary of statistics for two networks

	Tajima's D		Nucleotide diversity (π)		
Network	gene	promoter	gene	promoter	Ka/Ks
sub-blue	-0.00653±0.000178	0.00936±0.000214	0.00786±0.000158	0.00976±0.000221	0.279±0.0138
sub-turquoise	-0.0471±0.00256	-0.0201±0.00379	0.00696±0.000192	0.00973±0.000175	0.319±0.0115

Note: the values denote mean±standard error (SE).

We found a significant correlation between nucleotide diversity and contributions of different DS (Figure 23C and 23D). In the cell-cycle network, the contributions of different DS are significantly positively correlated with nucleotide diversity (Figure 23C). This indicates that DS of high contribution to TDI profiles show high nucleotide diversity. Therefore, sequence divergence of cell-cycle genes occurred in an earlier period. In contrast with the cell-cycle network, a negative correlation is observed between nucleotide diversity and contribution of each DS in the metabolic network (Figure 23D). Hence DS of high contribution shows low nucleotide diversity, and high contributions are observed in DS of high Ka/Ks ratio under positive selection, especially in DS10. Therefore, these genes are more in line with the characteristics of recent positive selection that leads to a more varied transcriptome.

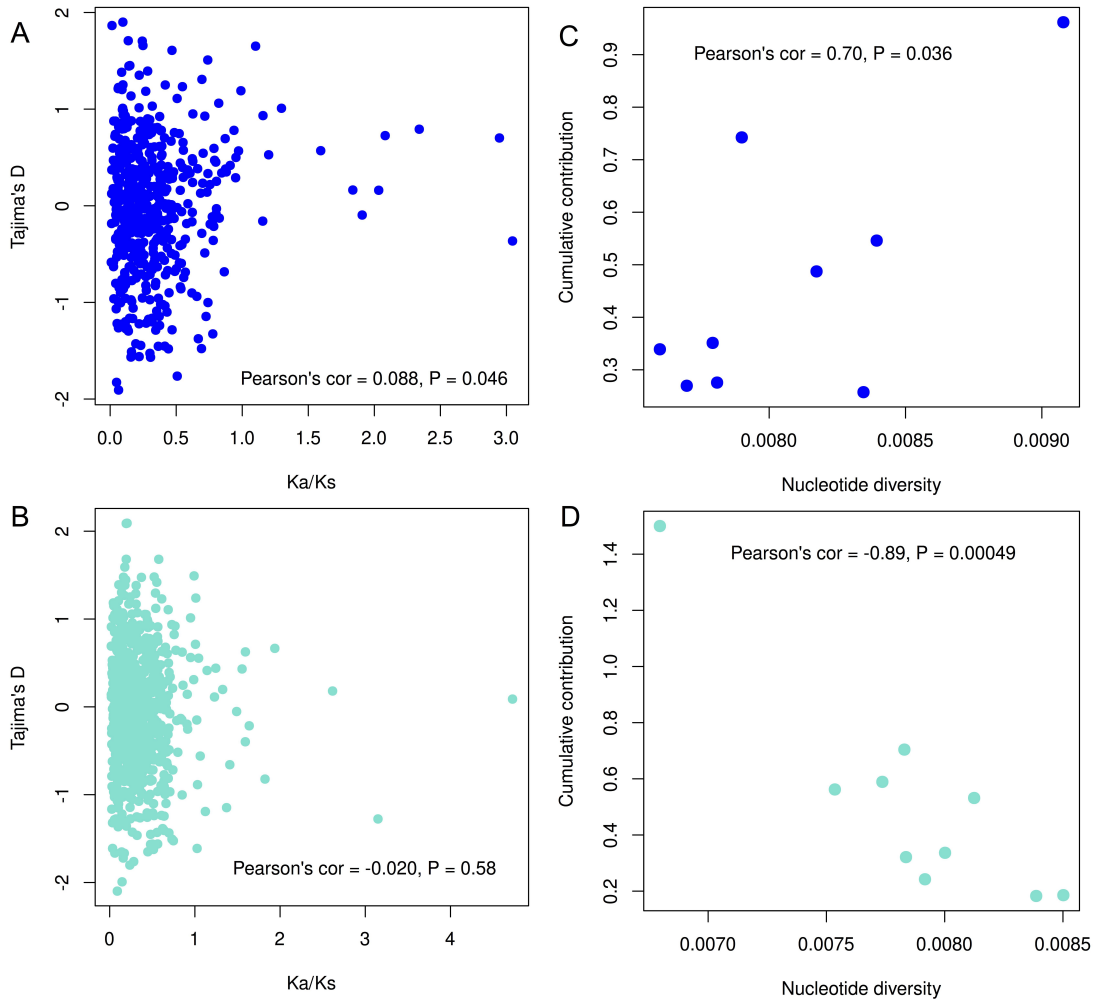

Figure 23. (A) the correlation between Tajima's D and Ka/Ks in sub-blue network. (B) the correlation between Tajima's D and Ka/Ks in sub-turquoise network. (C) the correlation between nucleotide diversity and cumulative contributions of each DS in sub-blue network. (D) the correlation between nucleotide diversity and cumulative contributions of each DS in sub-turquoise network.

The phylotratigraphic analysis identified 12 and 5 unique genes related to drought tolerance in metabolic and cell-cycle networks in *S. chilense,* respectively (Figure 18A; Table 5). Therefore, these genes have not been analyzed in the TDI profiles, because their orthologous genes can't be obtained from *S. pennellii*. In the cell-cycle network, TFBSs of TALE (three amino acids loop extension) homeodomain transcription factors were found in promoters of all 5 new drought-responsive genes, indicating these genes can be regulated via TALE transcription factors. TALE also is the most important transcription factor family in the cell-cycle network (410 genes contain TALE TFBS; Table 3). TALE transcription factor family is involved in the

regulation of growth, differentiation and reproduction, which regulates a complex network [239], indicating these five genes in the cell-cycle network also relate to cell-cycle process. All five genes have negative Tajima's D indicating difference with old genes that are affected by purifying selection, then these new drought-responsive genes related to cell cycle could be mainly subject to positive selection pressure in recent time, and they can increase evolutionary potential of cell-cycle network because old genes have been fixed in the drought transcriptome. The new 12 genes in the metabolic network can bound to multiple transcription factors, due to genes in the metabolic network are involved in diverse processes of fundamental metabolism (Figure 17; Table S7). Below average connectivity was observed in ten genes, indicating also multiple regulations occur in metabolic network. These new genes show positive Tajima's D, suggesting balancing selection occurs to maintain diversified regulation in the metabolic network.

Table 5 Summary of new genes of two networks in *S. chilense*

Networks	Genes	TFBS	π	Tajima's D
sub-turquoise	Schil_g12233	GRAS	0.00864	0.50103
sub-turquoise	Schil_g14997	MYB	0.01342	0.56276
sub-turquoise	Schil_g15233	MYB	0.01743	0.23710
sub-turquoise	Schil_g15353	Dof	0.01541	0.69798
sub-turquoise	Schil_g17548	Dof	0.00363	-0.65064
sub-turquoise	Schil_g17563	GRAS	0.00434	0.19820
sub-turquoise	Schil_g22212	MYB	0.00535	1.05489
sub-turquoise	Schil_g2223	Dof	0.00879	-0.55489
sub-turquoise	Schil_g24009	GRAS	0.00663	0.63914
sub-turquoise	Schil_g25994	ERF	0.00435	1.2801
sub-turquoise	Schil_g32207	GRAS	0.01131	-0.28503
sub-turquoise	Schil_g8855	ERF	0.01737	0.47923
sub-blue	Schil_g1417	TALE	0.00979	-0.10906
sub-blue	Schil_g15646	TALE	0.01104	-0.36042
sub-blue	Schil_g16423	TALE	0.01256	-0.11033
sub-blue	Schil_g22948	TALE	0.00508	-1.67724
sub-blue	Schil_g34254	TALE	0.00490	-0.94322

4.2.7 Drought-responsive genes under positive selection

The overlapped genes between drought-response networks and candidate genes in selective sweeps were observed, 57 and 83 drought-responsive genes in cell-cycle and metabolic networks under positive selection (Data S2E). This indicates that these drought-response genes are important for local adaptation to drought environment condition (Figure 2D). The lower nucleotide diversity (π) and Tajima's D values are observed in positive selection genes in metabolic than cell-cycle networks (Figure 24), indicating that metabolic genes show more significant signature under positive selection and highly evolutionary potential.

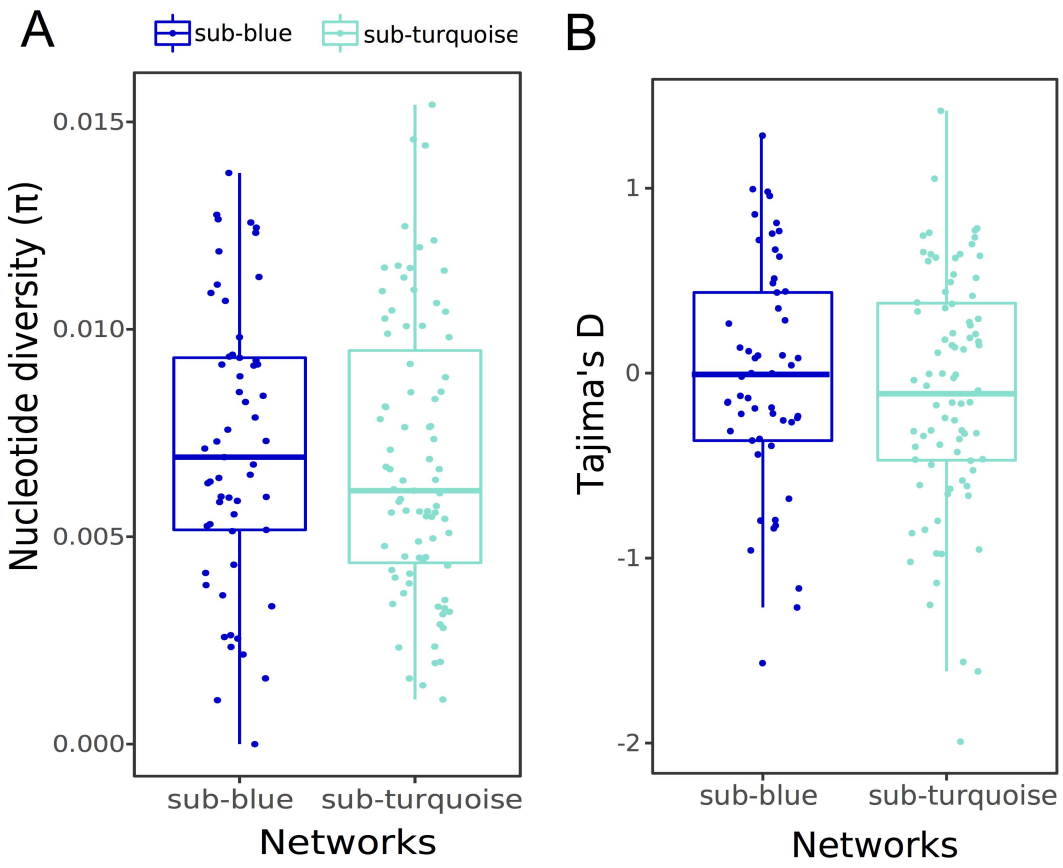

Figure 24. The statistics of drought-responsive genes under selective sweeps. (A) Nucleotide diversity (π). (B) Tajima's D.

4.2.8 Evolution of Gene Family

Defining gene families evolving rapidly among flowering plants has been advantageous in understanding the genomic bases underlying species adaptation [240, 241]. In flowering plants, the expansion or contraction of gene families is an important driver of lineage divergence and local adaptation [242, 243]. We characterized gene families present in the two analyzed drought-response gene networks that underwent significant changes and divergently evolved along different branches (Figure 25). Our results show that 415 and 367 are present in the metabolic and cell-cycle networks out of the 26,367 gene families inferred to be present in the most recent common ancestor of the six studied plant species. Moreover, 152 and 66 (out of 697 and 565) gene families containing drought-responsive genes exhibited significant expansion or contraction ($P < 0.01$) in the metabolic and cell-cycle network, respectively. As we expected, four species (*S.chilense*, *S. pennellii*, *S. lycopersicum*, *S. lycopersicoides*) are observed to have more expanded gene families than contracted in the metabolic network, especially *S.chilense* and *S. pennellii* also show stronger gene family expansion and the most rapidly evolving gene families (Figure 25A). The ratio of gene gain/loss of *S. chilense* is highest, indicating that drought-responsive gene families related to metabolic processes are currently evolving. In contrast, more gene families in the cell-cycle network show contraction and the gene loss rate higher than the gene gain rate in most species, except for *S. pennellii* (Figure 25B) and the number of rapidly evolving gene families related the cell-cycle network is far less than the metabolic network. This analysis also validates our interpretations about the evolutionary age of transcriptome. The drought-responsive genes related to metabolism network show younger evolutionary than cell-cycle genes and have higher current dynamic evolution.

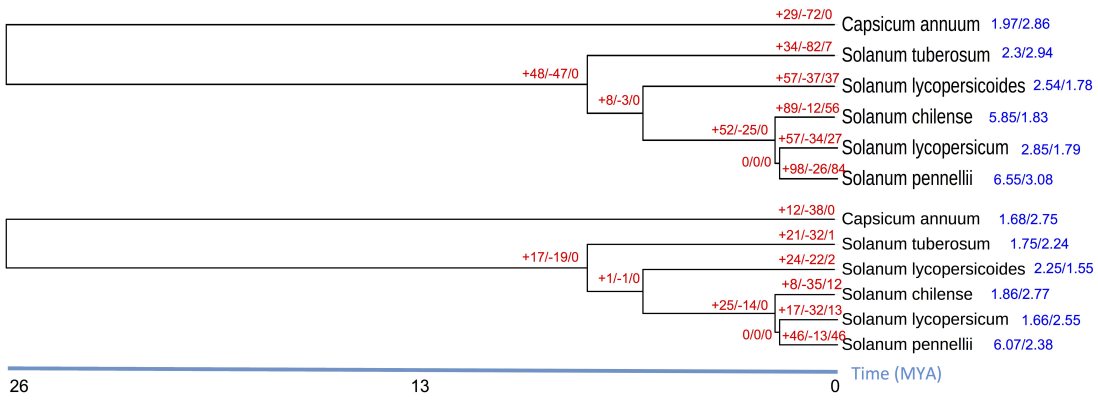

Figure 25. Expansion and contraction of gene families among the six plant species related to the metabolic network (A, sub-turquoise) and the cell-cycle network (B, sub-blue). Red numbers represent the numbers of expanded (+) / contracted (-) / rapidly evolving gene families, blue numbers represent gene gain/loss rates.

4.3 Discussion

In this chapter, we identified two drought response regulatory networks by analyzing overall gene expression profiles of plants growing under control and drought conditions, combining analysis of differential expression, gene co-expression network inference and TF/TFBS prediction. Both networks are typical highly conservative networks involving two clusters of biological processes: fundamental metabolism and cell cycle, which also represents two main strategies for drought response [244, 245]. The organ development is roughly divided into cell proliferation and cell expansion, being water deficit a limiting factor for both of them [246, 247]. Therefore, drought stress reduces the activity of the cell cycle and slows down the growth and development of plants. The down-regulated genes in cell-cycle network also indicate that genes related to cell cycle are suppressed by drought stress to restrict growth of *S.chilense* (Figure 19A). However, our speculations are mainly for the aboveground tissues of *S. chilense*. Our previous study using genome scan reveals that genes related to root hair differentiation under positive selection, and therefore transcriptomes of the root system still need to be fully analyzed. The changes of fundamental metabolism may be a more direct and passive drought-response strategy for drought tolerance. Plant water shortage is first reflected in changes in

metabolic processes, such as accelerating the catabolism of macromolecules to regulate the penetration of tissues to maintain physiological water balance or slowing down metabolism to reduce energy and water consumption [82, 248]. In addition, the signal pathways in the metabolic network are also clearer way to respond to drought stress in plants, for example, the abscisic acid (ABA) signaling pathway in metabolic network is a proven pathway that regulates the response of plants to dehydration and optimizes water utilization [249]. Although these two networks correspond to two different strategies of drought response, they are not isolated, but interact with each other. Water deprivation and heat will first change the metabolic processes and close stomata, then cell cycle will be changed when long-term lack of water, and the increased or decreased cell cycle will change the physiology and metabolism of the plant again [82]. But drought-responsive strategies that regulate the cell cycle may be activated later than metabolism processes, such as glucose metabolism early and rapidly follow the drought stress, whereas the accumulation of amino acids necessary for the cell cycle starts later than glucose metabolism in response to drought [250].

Difference with other genes in other plant originate in early period (PS1 to PS5) in tree of life [64, 251], most drought-responsive genes were found in *S. chilense* originated from the early to middle stages of the tree of life (PS1, PS2, PS5, PS8; Figure 18A and 18B), which are consistent origin with abiotic stress associated genes in *Arabidopsis thaliana*, including water, heat and salt stress related to drought stress [252]. PS8 (flowering plants) is an important period to origin of drought-responsive genes, indicating drought-response networks can be extremely expanded in this period. And the formation and expansion of flowering plants have greatly enriched the drought-response networks [253], and this situation also was observed in maize [254]. Therefore, we suggest that two drought-responsive networks occur in the formation of land species and expand in period of flowering plants formation. We calculated the transcriptome age of drought-responsive genes and found that transcriptomes of drought samples are much younger than those normal samples (Figure 18C). However, the pattern of TAI profiles in this study show also difference with TAI profiles

of developmental stages. TAI profiles in embryogenesis stages exhibit ‘hourglass pattern’ and this is also earliest discovery to apply TAI analysis [60, 63, 64]. Whereas the TAI profiles in the same condition remain stable, and similar with TAI profiles in different tissues [254], that is same transcriptome age in same conditions, and then rapidly increased under drought conditions. We call this TAI pattern as the ‘ladder pattern’, and same pattern is observed in TDI profiles (Figure 21B). This may be a specific pattern in transcriptomes age of comparative differently environmental stress.

Although both networks exhibit similar TAI and TDI patterns, transcriptome age of metabolic network is younger than cell-cycle network. In addition to differences in the origin of genes in two networks, the evolution of the transcriptome expression profile is also a key factor [255-257]. In cell-cycle network, the TAI profile is almost entirely contributed by older phylostrata (PS1 to PS8), but new genes contribute about 20% to TAI profile in metabolic network. This indicates expression of genes in cell-cycle network has been fixed drought transcriptopme experienced adaptive changes in expression to drought environment [256]. Stronger evidence is in the TDI profiles, conserved genes are more contributed for TDI profiles in cell-cycle networks and adaptive changes in expression to drought environment (higher TDI difference between control and drought transcriptomes, Figure 21B). Whereas drought-responsive genes related to fundamental metabolism are unstable in expression when responding to drought stress, and this strategy may trend to respond to the initial stages of water scarcity [258]. Expression of genes in metabolic network is not completely fixed in the drought transcriptome and because these basic metabolic processes are also involved in many other response processes, such as ABA pathway regulates multiple response processes to stresses [259, 260].

Furthermore, more divergent transcriptomes in drought samples illustrate drought-response genes are divergent expression between drought and normal samples (Figure 21A). The drought-response genes related to fundamental metabolism are more divergent than genes related to cell cycle. This indicates that

divergent expression is more active for the metabolic genes than cell cycle genes, and the sequence divergence is also one of the driving forces for new gene origination because it likely facilitates the reshape genome and introduces new features to the genome [261]. This also supports the genes related to fundamental metabolism are younger than cell cycle genes, and there are more new genes in the metabolic than cell cycle networks. However, the difference of transcriptome divergence between drought and normal conditions in the cell-cycle network is much higher than in the metabolic network, and low DS showed higher contribution in the cell-cycle network. This suggests that genes related to cell cycle are higher conserved and suffer stronger purifying selection than genes related to metabolic genes in recent times.

Combining analysis of nucleotide diversity and gene family, we speculate that drought-responsive genes related to cell cycle have sequence divergence and selection occurred in an earlier period, and they suffer strict purify selection and contraction of the gene family. However, genes in the metabolic network show an indication of positive selection in the recent period, and more variable expression in the drought transcriptome. And the sequence divergence is also one of the driving forces for new gene origination because it likely facilitates the reshape genome and introduces new features to the genome [261]. This also supports the genes related to fundamental metabolism are younger than cell cycle genes, and there are more new genes in the metabolic than cell cycle networks.

The gene families in metabolic network are significantly expanded and rapidly evolved, but contracted in cell-cycle network. Therefore, we conjecture drought-responsive genes in the metabolic network manifest greater evolutionary potential in the future. In addition, some new drought-responsive genes are detected in both networks and show different trends of evolution. They are drought genes unique to *S. chilense*, and maybe to arise adaptability to unique habitats. In addition, we found a lot of drought-responsive genes shared with candidate genes under positive sweeps. This indicates that drought stress is an important signature of local

adaptation in *S. chilense.* These genes may be also 'hitchhiking' in the evolution of gene expression. The rest drought-responsive genes also help us to expand the genetic network for drought response.

In this study, we identify two regulatory networks to drought tolerance and estimate their evolution processes. Two networks correspond to two typical drought-response strategies related to the cell cycle and fundamental metabolic processes, respectively. Drought-responsive genes in drought transcriptome are evolutionary younger and more divergent than control transcriptome. And the younger and more variable transcriptome is observed in the network of fundamental metabolism than cell-cycle. Gene families are expanded and contracted in the metabolic and cell-cycle networks, respectively. Therefore, genes in the metabolic network are the preferentially responded to drought. The drought-response strategy of cell cycle originates earlier and has been fixed in drought transcriptome. But some new genes in the cell-cycle network may make up for the lack of evolutionary potential.

5 General discussion and conclusion

5.1 General discussion

5.1.1 Discussion of results

Polygenic adaptation is likely a general feature of adaptation of all species. Several methods to detect such positive selection can well reveal its genetic and evolutionary basis, helping researchers to quickly locate the target regions in the genome related to divergence, colonization and expansion of species along changing environmental conditions [15, 47, 50, 51, 57, 58]. Genome-wide scans have fully exposed positive selection driving local adaptation of populations in different species [16, 54, 55, 262]. This positive selection disproportionately targets specific regulatory regions in the genome, hinting for an important role of gene regulatory networks in evolution [27]. However, these methods of genome scans show also some limitations discussed in recent articles [263-265]. The development of systems biology, especially gene regulatory networks, provides a more intuitive way to visualize how gene networks respond to environmental changes [27, 266-268]. Therefore, this thesis integrates genome scans and gene regulatory networks to obtain a more comprehensive understanding for local adaptation of *S. chilense*. There is some recent advances in population genetics theory which extend positive selection theory into a polygenic selection framework [34, 269]. In a nutshell strong positive selection generating selective sweeps is one occurrence of polygenic models, along with soft or incomplete sweeps which are much more difficult to detect in genomic data. In this thesis it is assumed that our selective sweeps do occur within a polygenic selection framework and represent a subset of the selection for a given adaptation.

The analysis of genome scans associated with climatic data illustrate that an intermingled model of local spatial adaptation to novel habitats and temporal adaptation to changing climatic conditions. Further, reconstruction of niche model using climatic data from different periods allows for separate the underpinnings of the local and temporal adaptation. The age of sweeps spans about 150 kya and sweeps

restricted bursts in each of the five past climatic periods (see chapter 3, Figure 8). Contrary to our expectations, although the south populations were established later in new habitats, especially for south-highland populations, a large number of old sweeps were still are observed in these populations. This also indicates that positive polygenic adaptation is a complex process involving several discrete adaptive steps. In this thesis, we conclude that the characteristics of local adaptation in *S. chilense* includes: (1) the spatial and temporal adaptation proceeding simultaneously, and affecting each other in the same climatic period; (2) selective sweeps may be inherited from the ancestral population, and can be detected for short or long time period, this is also one of the reasons why a large number of older sweeps are identified in south-highland population; (3) the strong gene-flow increases genetic exchange may result in an early pre-adaptation to future climate changes and to new niches/habitats available in future climatic periods; (4) the selective sweeps are repeatable and can be correlated to climatic changes. In other words, when *S. chilense* colonizing new habitats faces climatic conditions to which it has adapted before, adaptation occurs in the same adaptive gene network a further fine tuning.

The extreme drought is the most significant and widespread stress in the habitats of most *S. chilense* populations, especially when compared to other tomato species which have more northern distribution in less arid areas of South-America [93, 96]. Two gene regulatory networks are identified as responding to drought stress. They show different evolutionary ages and transcriptome conservation (Figure 18 and Figure 21). Of the two networks, we find that genes in metabolic network are younger and show variable expression in drought condition and preferentially responded to drought stress. This may indicates that across different evolutionary periods of the tree of life, different strategies to cope with drought stress are adopted [270, 271]. We speculate that over the long-term evolution, drought-response genes in cell-cycle network tend to adaptation in response to long-time drought stress, whereas metabolic regulation are more likely to be an immediate strategy to water shortages [78, 81, 82, 250, 272]. Or these two strategies correspond to an old ancestral

unicellular response to drought stress, while the second metabolic is a response to a physiological stress in a multicellular organism such as flowering plants.

We suggest that drought-responded strategies of cell cycle and metabolism form part of the network of local adaptation, because we observed also many positive selection genes in drought-response networks and show typical signatures of positive selection (Figure 24). But positive selection genes in two networks show different strength of positive selection and evolutionary ages are also different, indicating spatial and temporal adaptation of gene network regulation involved in multiple steps and periods. The positive selection genes in different networks burst out in different periods and show different strength of positive selection. Therefore, we suggest that this phenomenon is caused by a temporal adaptation to a moving environmental optimum. The different strength of positive selection in gene networks at different periods are consistent with polygenic selection model of adaptation to a moving optimum [19, 21, 22, 33].

With understanding of the evolution of organisms, modern molecular evolution studies usually correlate the appearance of genes and gene families with the appearance of biological pathways and morphological characteristics related to local adaptation. Therefore, the gene network structures or the co-option of gene networks to novel habitats provide a new solution for studies of local adaptation [27, 76, 273, 274]. We suggest that gene networks (modules) in this thesis are conserved among different higher plants, and this is one of the basic characteristics of the regulatory network [69-71]. For example, abscisic acid (ABA) pathway in our metabolic network was also widely reported in multiple model plants and crops to respond to drought stress [207, 214, 259, 260]. The many genes (FLC, FT, FD etc. in Figure11) in the flowering regulated network responding to low temperature and circadian rhythm were also reported in many flowering plants, including *Arabidopsis sp.*, tobacco, rice, wheat, etc [177, 178, 275-280]. These genes continue to evolve in addition to maintaining basic core functions leading to duplication of gene modules by neofunctionalization

[75]. Gene regulatory network is composed of many different local sub-networks that relate to different pathways. The multiple single gene duplications (SGDs) is the main form of module duplication followed by genome duplication [75, 281]. A recent study illustrated the conservation and duplication of modules associated with regulation of plant growth and development [75]. The modules found are similar to ours, including the cell-cycle network (chapter 4) and shared some key genes with our list, such as some transcription factors (bZIP and MYB). This indicates that an investigation of local module duplication can give indication about the mechanism of network evolution.

5.1.2 Perspectives

In this work we used two types of resources: genome data and transcriptomics. I will discuss the limitations and perspectives on how to improve the current results and future hypotheses to be tested first based on genomic data, second on transcriptomic data and finally on additional experiments that can be preformed.

First, we obtained selective sweeps from smaller sample size (six representative populations of different geographic regions, with five diploid plants per population). This requires us to have a stringent detection threshold to increase the credibility of the results (just like our results). Therefore, we want to decrease the rate of false positives, at the cost of sacrificing some positive selection sweeps and missing some genes in the gene networks (Figure 11). Therefore, we still look forward to a larger-scale whole-genome sequencing project, preferably including all populations in the TGRC. Then a sample of more representative populations based on population structure analysis can be selected to reduce false negatives, thereby completing the detection of genes under positive selection related to local adaptation. Moreover, the larger sample size also facilitates the analysis of the genotype-environment associations [163]. We have also recently assembled a new complete reference genome of *S. chilense (used in chapter 4)*, which will also help us

obtain more accurate annotation information in the sweep regions. In addition, the habitats of *S. chilense* are mixed with other relative wild tomato species, such as *S. peruvianum*, *S. pennellii* etc. A comparative analysis of genetic bases for local adaptation to same habitat in different wild tomato species would be great and likely meaningful. I do not think for the problem at hand of local adaptation that it is necessary to perform PacBio sequencing. One issue is the accuracy SNP calling in a highly heterozygous genome. As the genetic research on *S. chilense* is still young and high quylity reference genome is still misisng, the high error rate of sequencing has not been resolved in PacBio sequencing and would still need next-generation sequencing to correct SNP calls [282]. However, PacBio and long read sequencing data can help to resolve part of the current genome assembly which are not accurate such as gene duplication or inversions.

Second, the combination of genomics and transcriptomics is still a suitable solution for polygenic adaptation research. We obtained evolution of gene networks responding to drought conditions. But the sequenced samples focus on leaves and shoot apex, and the results from sweeps show gene network related to differentiation of root hair and homeostasis of root cell (Figure 11C). We expect an opposite result from the analysis of root transcriptomic regarding the cell-cycle network, because the aerial tissues of plants stop growing, and the root system tends to accelerate cellular differentiation [216, 283]. In addition, cold stress is also an important adaptive signature in habitat of south-highland populations. The gene networks involved in cold tolerance can also help us increase the diversity of polygenic adaptive networks, however the study of has only limited amount of plants and statistical power to detect all involved genes [107]. It is likely that with the new reference genome being available, more transcriptomic data will be obtained in different conditions in the future, allowing us to reveal more complete picture of local adaptation at the transcriptome level.

From whole-genome and transcriptome sequencing data, we obtain adaptive

gene networks related to complex environmental conditions. An important downstream analysis for plant breeders is identification of key regulatory loci and functional verification by experiments of molecular biology. Therefore, a genome-wide association study (GWAS) would be an effective method to extract key loci related to multiple adaptations of environments and traits. However, this requires the sequencing and phenotyping of a large number of plants, especially in *S. chilense* where the number of SNPs is very large and the linkage disequilibrium very small. Besides, I will also continue the evolutionary analysis of gene networks based theory of network evolution. Genome duplication will be a center for next study to local expansions of gene networks using copy number variation analysis.

5.2 Conclusion

In this thesis, I have done a first step to decipher the genetic bases of local adaptation in the wild tomato species *Solanum chilense*. We have shown first that correlate the demographic history, allele frequencies in space and time, the age of selection events and the reconstructed historical ecological distribution of the species over five main climatic periods spanning 150,000 years. We find evidence for several selective sweeps targeting regulatory networks involved in root hair development in low altitude, and response to photoperiod and vernalization in high altitude populations. Second, based on phenotyping the gene expression response to drought, we find two main response networks regulating cell cycle and metabolic processes. These drought-responsive genes originate from different periods in the tree of life (from ancestral unicellular up to novel genes in this species) and show different evolutionary ages and transcriptomic conservation. In a species with large effective population size and which colonize different habitats around the Atacama desert, we find that adaptation is a mutli-locus process involving many genes and several regulatory networks. Furthermore, in *Solanum chilense* we can recover old ages of selection showing that adaptation occurs by discrete time steps in a given sub-regulatory network. New genes forming a metabolic network of adaptation to drought are also found to have evolved newly only in this species, and these genes show signs of

positive selection. Our results broadly support the polygenic adaptation model and allow us to reveal the underlying genetic mechanisms underlying adaptation to arid habitats in plants.

Acknowledgement

Time always flies so fast that in the blink of an eye, the last four years of my life, when I was a student, will come to a successful conclusion. I would like to take this opportunity to express sincerest thanks and gratitude to all people who gave their time to help me in the production of this dissertation, for their expertise and supports.

Firstly, I would like to thank Prof. Dr. Aurélien Tellier for his expertise, guidance and support throughout as my dissertation's supervisor. He has given up her own time to proof-read my works and made valuable suggestions at all stages of the research. It is not easy to write a dissertation, and my study would not have been possible without his efforts.

I would also like to thank Dr. Gustavo Silva here. She has given a lot help for me in the procedure of this dissertation's completion. His ideas and suggestions also run through the entire PhD project. I am even more grateful that he can always help me solve many scientific problems in the first time. These meaningful discussions have benefited me a lot.

I want to thank Saida for helping to obtain RNA-seq data for this study. Kevin help me to translate the summary into German. My thanks also go to some outstanding colleagues and friends. You have made my life wonderful and I also enjoy being with you.

I would also like to thank Chinese Scholarship Council (CSC) provides me with scholarships for four years.

Finally, I would like to thank my parents, younger sister, and Mire for their support and care for me. It is you who have given me infinite strength.

Bibliography

1. Tomato Genome Consortium x: **The tomato genome sequence provides insights into fleshy fruit evolution.** *Nature* 2012, **485:**635.
2. Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sørensen I, Lichtenstein G: **The genome of the stress-tolerant wild tomato species Solanum pennellii.** *Nature genetics* 2014, **46:**1034-1038.
3. Wang X, Gao L, Jiao C, Stravoravdis S, Hosmani PS, Saha S, Zhang J, Mainiero S, Strickler SR, Catala C, et al: **Genome of Solanum pimpinellifolium provides insights into structural variants during tomato breeding.** *Nature Communications* 2020, **11:**5817.
4. Takei H, Shirasawa K, Kuwabara K, Toyoda A, Matsuzawa Y, Iioka S, Ariizumi T: **De novo genome assembly of two tomato ancestors, Solanum pimpinellifolium and Solanum lycopersicum var. cerasiforme, by long-read sequencing.** *DNA Research* 2021, **28:**dsaa029.
5. Darwin C, Wallace AR: **Evolution by natural selection.** *Evolution by natural selection* 1958.
6. Fisher RA: *The genetical theory of natural selection.* Рипол Класс и к ; 1958.
7. Williams GC: *Adaptation and natural selection.* Princeton university press; 2018.
8. Sober E, Wilson DS: **Adaptation and natural selection revisited.** *Journal of evolutionary biology* 2011, **24:**462-468.
9. Schluter D, Nagel LM: **Parallel speciation by natural selection.** *The American Naturalist* 1995, **146:**292-301.
10. Nei M: **Genetic polymorphism and the role of mutation in evolution.** *Evolution of genes and proteins* 1983, **71:**165-190.
11. Barrett RD, Schluter D: **Adaptation from standing genetic variation.** *Trends in ecology & evolution* 2008, **23:**38-44.
12. Rich S, Bell A, Wilson S: **Genetic drift in small populations of Tribolium.** *Evolution* 1979**:**579-584.
13. Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL: **Genetic drift, selection and the evolution of the mutation rate.** *Nature Reviews Genetics* 2016, **17:**704-714.
14. Santangelo JS, Johnson MT, Ness RW: **Modern spandrels: the roles of genetic drift, gene flow and natural selection in the evolution of parallel clines.** *Proceedings of the Royal Society B: Biological Sciences* 2018, **285:**20180230.
15. Fay JC, Wu C-I: **Hitchhiking under positive Darwinian selection.** *Genetics* 2000, **155:**1405-1413.
16. Zhang J: **Positive Darwinian selection in gene evolution.** In *Darwin's Heritage Today: Proceedings of the Darwin 200 Beijing International Conference: 24–26 October 2009; Beijing*. Citeseer; 2010: 288-309.
17. Loewe L: **Negative selection.** *Nature education* 2008, **1:**59.
18. Hedrick PW: **Balancing selection.** *Current Biology* 2007, **17:**R230-R231.
19. Polechová J, Barton N, Marion G: **Species' range: adaptation in space and time.** *The American Naturalist* 2009, **174:**E186-E204.
20. Chevin LM, Martin G, Lenormand T: **Fisher's model and the genomics of adaptation: restricted pleiotropy, heterogenous mutation, and parallel evolution.** *Evolution: International Journal of Organic Evolution* 2010, **64:**3213-3231.

21. Matuszewski S, Hermisson J, Kopp M: **Fisher's geometric model with a moving optimum.** *Evolution* 2014, **68:**2571-2588.
22. Jain K, Stephan W: **Rapid adaptation of a polygenic trait after a sudden environmental shift.** *Genetics* 2017, **206:**389-406.
23. Smith JM, Haigh J: **The hitch-hiking effect of a favourable gene.** *Genetics Research* 1974, **23:**23-35.
24. Kim Y, Stephan W: **Detecting a local signature of genetic hitchhiking along a recombining chromosome.** *Genetics* 2002, **160:**765-777.
25. Exposito-Alonso M, Burbano HA, Bossdorf O, Nielsen R, Weigel D: **Natural selection on the Arabidopsis thaliana genome in present and future climates.** *Nature* 2019, **573:**126-129.
26. Savolainen O, Lascoux M, Merilä J: **Ecological genomics of local adaptation.** *Nature Reviews Genetics* 2013, **14:**807-820.
27. Fagny M, Austerlitz F: **Polygenic adaptation: Integrating population genetics and gene regulatory networks.** *Trends in Genetics* 2021.
28. Schrider DR, Kern AD: **Soft sweeps are the dominant mode of adaptation in the human genome.** *Molecular biology and evolution* 2017, **34:**1863-1877.
29. Zou Y-P, Hou X-H, Wu Q, Chen J-F, Li Z-W, Han T-S, Niu X-M, Yang L, Xu Y-C, Zhang J: **Adaptation of Arabidopsis thaliana to the Yangtze River basin.** *Genome biology* 2017, **18:**1-11.
30. Garud NR, Messer PW, Petrov DA: **Detection of hard and soft selective sweeps from Drosophila melanogaster population genomic data.** *PLoS Genetics* 2021, **17:**e1009373.
31. Stephan W: **Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation.** *Molecular ecology* 2016, **25:**79-88.
32. Li H, Stephan W: **Inferring the demographic history and rate of adaptive substitution in Drosophila.** *PLoS genetics* 2006, **2:**e166.
33. Jain K, Stephan W: **Modes of rapid polygenic adaptation.** *Molecular biology and evolution* 2017, **34:**3169-3175.
34. Barghi N, Hermisson J, Schlötterer C: **Polygenic adaptation: a unifying framework to understand positive selection.** *Nature Reviews Genetics* 2020, **21:**769-781.
35. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM: **Genomic signatures of positive selection in humans and the limits of outlier approaches.** *Genome research* 2006, **16:**980-989.
36. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG: **Recent and ongoing selection in the human genome.** *Nature Reviews Genetics* 2007, **8:**857-868.
37. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M: **Robust Demographic Inference from Genomic and SNP Data.** *PLoS Genetics* 2013, **9**.
38. Alachiotis N, Pavlidis P: **RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors.** *Communications biology* 2018, **1:**79-79.
39. Navascués M, Leblois R, Burgarella C: **Demographic inference through approximate-Bayesian-computation skyline plots.** *PeerJ* 2017, **5:**e3530.
40. Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, Poss ML, Reed LK, Storfer A, Whitlock MC: **Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions.** *The American Naturalist* 2016, **188:**379-397.
41. Peterson AT, Soberón J, Pearson RG, Anderson RP, Martínez-Meyer E, Nakamura M, Araújo

MB: *Ecological niches and geographic distributions (MPB-49).* Princeton University Press; 2011.

42. Pocheville A: **The ecological niche: history and recent controversies.** In *Handbook of evolutionary thinking in the sciences.* Springer; 2015: 547-586
43. Polechová J, Storch D: **Ecological niche.** *Encyclopedia of ecology* 2008, **2:**1088-1097.
44. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W: **The hitchhiking effect on the site frequency spectrum of DNA polymorphisms.** *Genetics* 1995, **140:**783-796.
45. Kim Y, Nielsen R: **Linkage disequilibrium as a signature of selective sweeps.** *Genetics* 2004, **167:**1513-1524.
46. Stephan W, Song YS, Langley CH: **The hitchhiking effect on linkage disequilibrium between linked neutral loci.** *Genetics* 2006, **172:**2647-2663.
47. Koropoulis A, Alachiotis N, Pavlidis P: **Detecting positive selection in populations using genetic data.** In *Statistical Population Genomics.* Humana, New York, NY; 2020: 87-123
48. Watterson G: **On the number of segregating sites in genetical models without recombination.** *Theoretical population biology* 1975, **7:**256-276.
49. Nei M, Li WH: **Mathematical model for studying genetic variation in terms of restriction endonucleases.** *Proceedings of the National Academy of Sciences of the United States of America* 1979, **76:**5269-5273.
50. Pavlidis P, Alachiotis N: **A survey of methods and tools to detect recent and strong positive selection.** *Journal of Biological Research-Thessaloniki* 2017, **24:**1-17.
51. Stephan W: **Detecting strong positive selection in the genome.** *Molecular ecology resources* 2010, **10:**863-872.
52. Stephan W: **Selective sweeps.** *Genetics* 2019, **211:**5-13.
53. Rubin C-J, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S: **Whole-genome resequencing reveals loci under selection during chicken domestication.** *Nature* 2010, **464:**587-591.
54. Wang M-S, Zhang R-w, Su L-Y, Li Y, Peng M-S, Liu H-Q, Zeng L, Irwin DM, Du J-L, Yao Y-G: **Positive selection rather than relaxation of functional constraint drives the evolution of vision during chicken domestication.** *Cell research* 2016, **26:**556-573.
55. Yuan Y, Zhang Q, Zeng S, Gu L, Si W, Zhang X, Tian D, Yang S, Wang L: **Selective sweep with significant positive selection serves as the driving force for the differentiation of japonica and indica rice cultivars.** *BMC genomics* 2017, **18:**1-13.
56. Zhang Z, Jia Y, Almeida P, Mank JE, van Tuinen M, Wang Q, Jiang Z, Chen Y, Zhan K, Hou S: **Whole-genome resequencing reveals signatures of selection and timing of duck domestication.** *Gigascience* 2018, **7:**giy027.
57. Pavlidis P, Jensen JD, Stephan W: **Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations.** *Genetics* 2010, **185:**907-922.
58. Lin K, Li H, Schlotterer C, Futschik A: **Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics.** *Genetics* 2011, **187:**229-244.
59. Schrider DR, Kern AD: **S/HIC: robust identification of soft and hard sweeps using machine learning.** *PLoS genetics* 2016, **12:**e1005928.
60. Domazet-Lošo T, Tautz D: **A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns.** *Nature* 2010, **468:**815-818.
61. Quint M, Drost H-G, Gabel A, Ullrich KK, Bönn M, Grosse I: **A transcriptomic hourglass in**

plant embryogenesis. *Nature* 2012, **490:**98-101.

62. Domazet-Lošo T, Brajković J, Tautz D: **A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages.** *Trends in Genetics* 2007, **23:**533-539.
63. Piasecka B, Lichocki P, Moretti S, Bergmann S, Robinson-Rechavi M: **The hourglass and the early conservation models—co-existing patterns of developmental constraints in vertebrates.** *PLoS genetics* 2013, **9:**e1003476.
64. Drost H-G, Gabel A, Grosse I, Quint M: **Evidence for active maintenance of phylotranscriptomic hourglass patterns in animal and plant embryogenesis.** *Molecular biology and evolution* 2015, **32:**1221-1231.
65. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A: **A large-scale evaluation of computational protein function prediction.** *Nature methods* 2013, **10:**221-227.
66. Rhee SY, Mutwil M: **Towards revealing the functions of all genes in plants.** *Trends in plant science* 2014, **19:**212-221.
67. Proost S, Mutwil M: **Tools of the trade: studying molecular networks in plants.** *Current Opinion in Plant Biology* 2016, **30:**143-150.
68. Movahedi S, Van Bel M, Heyndrickx KS, Vandepoele K: **Comparative co‐expression analysis in plant biology.** *Plant, cell & environment* 2012, **35:**1787-1798.
69. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *science* 2003, **302:**249-255.
70. Gerstein MB, Rozowsky J, Yan K-K, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ: **Comparative analysis of the transcriptome across distant species.** *Nature* 2014, **512:**445-448.
71. Zarrineh P, Sánchez-Rodríguez A, Hosseinkhan N, Narimani Z, Marchal K, Masoudi-Nejad A: **Genome-scale co-expression network comparison across Escherichia coli and Salmonella enterica serovar Typhimurium reveals significant conservation at the regulon level of local regulators despite their dissimilar lifestyles.** *PLoS One* 2014, **9:**e102871.
72. Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S: **PlaNet: combined sequence and expression comparisons across plant networks derived from seven species.** *The Plant Cell* 2011, **23:**895-910.
73. Tzfadia O, Amar D, Bradbury LM, Wurtzel ET, Shamir R: **The MORPH algorithm: ranking candidate genes for membership in Arabidopsis and tomato pathways.** *The Plant Cell* 2012, **24:**4389-4406.
74. Park CY, Wong AK, Greene CS, Rowland J, Guan Y, Bongo LA, Burdine RD, Troyanskaya OG: **Functional knowledge transfer for high-accuracy prediction of under-studied biological processes.** *PLoS computational biology* 2013, **9:**e1002957.
75. Ruprecht C, Mendrinna A, Tohge T, Sampathkumar A, Klie S, Fernie AR, Nikoloski Z, Persson S, Mutwil M: **FamNet: a framework to identify multiplied modules driving pathway expansion in plants.** *Plant physiology* 2016, **170:**1878-1894.
76. Ficklin SP, Feltus FA: **Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice.** *Plant Physiology* 2011, **156:**1244-1256.
77. Kashyap SP, Prasanna HC, Kumari N, Mishra P, Singh B: **Understanding salt tolerance mechanism using transcriptome profiling and de novo assembly of wild tomato Solanum chilense.** *Scientific reports* 2020, **10:**1-20.
78. Basu S, Ramegowda V, Kumar A, Pereira A: **Plant adaptation to drought stress.**

F1000Research 2016, **5**.

79. Rodrigues J, Inzé D, Nelissen H, Saibo NJ: **Source–sink regulation in crops under water deficit.** *Trends in plant science* 2019, **24:**652-663.
80. Shinozaki K, Yamaguchi-Shinozaki K: **Gene networks involved in drought stress response and tolerance.** *Journal of experimental botany* 2007, **58:**221-227.
81. Shanker AK, Maheswari M, Yadav S, Desai S, Bhanu D, Attal NB, Venkateswarlu B: **Drought stress responses in crops.** *Functional & integrative genomics* 2014, **14:**11-22.
82. Gupta A, Rico-Medina A, Caño-Delgado AI: **The physiology of plant responses to drought.** *Science* 2020, **368:**266-269.
83. Mickelbart MV, Hasegawa PM, Bailey-Serres J: **Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability.** *Nature Reviews Genetics* 2015, **16:**237-251.
84. Ritonga FN, Chen S: **Physiological and molecular mechanism involved in cold stress tolerance in plants.** *Plants* 2020, **9:**560.
85. Guo X, Zhang L, Zhu J, Liu H, Wang A: **Cloning and characterization of SiDHN, a novel dehydrin gene from Saussurea involucrata Kar. et Kir. that enhances cold and drought tolerance in tobacco.** *Plant Science* 2017, **256:**160-169.
86. Shi Y, Ding Y, Yang S: **Molecular regulation of CBF signaling in cold acclimation.** *Trends in plant science* 2018, **23:**623-637.
87. Xu S, Chong K: **Remembering winter through vernalisation.** *Nature Plants* 2018, **4:**997-1009.
88. Peralta I, Knapp S, Spooner D: **The taxonomy of tomatoes: a revision of wild tomatoes (Solanum L. section Lycopersicon (Mill.) Wettst.) and their outgroup relatives (Solanum sections Juglandifolium (Rydb.) Child and Lycopersicoides (Child) Peralta).** *Systematic Botany Monographs* 2007, **84:**1-186.
89. Peralta IE, Spooner DM, Knapp S: **Taxonomy of wild tomatoes and their relatives (Solanum sect. Lycopersicoides, sect. Juglandifolia, sect. Lycopersicon; Solanaceae).** *Systematic botany monographs* 2008, **84**.
90. Rick C, Chetelat R: **Utilization of related wild species for tomato improvement.** In *I International Symposium on Solanacea for Fresh Market 412*. 1995: 21-38.
91. Eshed Y, Zamir D: **An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL.** *Genetics* 1995, **141:**1147-1162.
92. Frary A, Nesbitt TC, Frary A, Grandillo S, Van Der Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB: **fw2. 2: a quantitative trait locus key to the evolution of tomato fruit size.** *Science* 2000, **289:**85-88.
93. Moyle LC, Nakazato T: **Hybrid incompatibility "snowballs" between Solanum species.** *Science* 2010, **329:**1521-1523.
94. Spooner DM, Peralta IE, Knapp S: **Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [Solanum L. section Lycopersicon (Mill.) Wettst.].** *Taxon* 2005, **54:**43-61.
95. Raduski AR, Igić B: **Biosystematic studies on the status of Solanum chilense.** *American Journal of Botany* 2021, **108:**520-537.
96. Nakazato T, Warren DL, Moyle LC: **Ecological and geographic modes of species divergence in wild tomatoes.** *American Journal of Botany* 2010, **97:**680-693.
97. Böndel KB, Lainer H, Nosenko T, Mboup M, Tellier A, Stephan W: **North–south colonization**

associated with local adaptation of the wild tomato species Solanum chilense. *Molecular Biology and Evolution* 2015, **32:**2932-2943.

98. Stam R, Silva‐Arias GA, Tellier A: **Subsets of NLR genes show differential signatures of adaptation during colonization of new habitats.** *New Phytologist* 2019, **224:**367-379.
99. Dillon MO: **The Solanaceae of the lomas formations of coastal Peru and Chile.** *Monographs in Systematic Botany* 2005, **104:**131.
100. Blanchard-Gros R, Martinez J-P, Quinet M: **Comparison of drought and heat resistance strategies among 6 populations of Solanum chilense.** In; *2020*.
101. Zhou S, Sauvé RJ, Liu Z, Reddy S, Bhatti S, Hucko SD, Fish T, Thannhauser TW: **Identification of salt-induced changes in leaf and root proteomes of the wild tomato, Solanum chilense.** *Journal of the American Society for Horticultural Science* 2011, **136:**288-302.
102. Böndel KB, Nosenko T, Stephan W: **Signatures of natural selection in abiotic stress-responsive genes of Solanum chilense.** *Royal Society open science* 2018, **5:**171198-171198.
103. Blanchard-Gros R, Bigot S, Martinez J-P, Lutts S, Guerriero G, Quinet M: **Comparison of Drought and Heat Resistance Strategies among Six Populations of Solanum chilense and Two Cultivars of Solanum lycopersicum.** *Plants* 2021, **10:**1720.
104. Xia H, CAMUS‐KULANDAIVELU L, Stephan W, Tellier A, Zhang Z: **Nucleotide diversity patterns of local adaptation at drought‐related candidate genes in wild tomatoes.** *Molecular Ecology* 2010, **19:**4144-4154.
105. Fischer I, Camus‐Kulandaivelu L, Allal F, Stephan W: **Adaptation to drought in two wild tomato species: the evolution of the Asr gene family.** *New Phytologist* 2011, **190:**1032-1044.
106. Tapia G, Méndez J, Inostroza L: **Different combinations of morpho‐physiological traits are responsible for tolerance to drought in wild tomatoes Solanum chilense and Solanum peruvianum.** *Plant Biology* 2016, **18:**406-416.
107. Nosenko T, Böndel KB, Kumpfmüller G, Stephan W: **Adaptation to low temperatures in the wild tomato species Solanum chilense.** *Molecular Ecology* 2016, **25:**2853-2869.
108. Zhou J, Reddy S, Zhou S, Sauvé RJ, Bhatti S, Fish T, Thannhauser TW: **Effect of heat stress on leaf proteome and enzyme activity in Solanum chilense.** *Plant Stress* 2012, **6:**8-13.
109. Martínez JP, Antúnez A, Araya H, Pertuzé R, Fuentes L, Lizana XC, Lutts S: **Salt stress differently affects growth, water status and antioxidant enzyme activities in Solanum lycopersicum and its wild relative Solanum chilense.** *Australian Journal of Botany* 2014, **62:**359-368.
110. Stam R, Nosenko T, Hörger AC, Stephan W, Seidel M, Kuhn JM, Haberer G, Tellier A: **The de novo reference genome and transcriptome assemblies of the wild tomato species Solanum chilense highlights birth and death of NLR genes between tomato species.** *G3: Genes, Genomes, Genetics* 2019, **9:**3933-3941.
111. Stam R, Silva-Arias GA, Nosenko T, Scheikl D, Hörger AC, Stephan W, Haberer G, Tellier A: **A small subset of NLR genes drives local adaptation to pathogens in wild tomato.** *BioRxiv* 2017**:**210559.
112. Andrews S: **FastQC: a quality control tool for high throughput sequence data.** 2010.
113. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30:**2114-2120.
114. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform.**

bioinformatics 2009, **25:**1754-1760.

115. Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence alignment/map (SAM) format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
116. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics* 2011, **27:**2987-2993.
117. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.** *Fly* 2012, **6:**80-92.
118. Lee T-H, Guo H, Wang X, Kim C, Paterson AH: **SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data.** *BMC genomics* 2014, **15:**162-162.
119. Letunic I, Bork P: **Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation.** *Nucleic acids research* 2021, **49:**W293-W296.
120. Yang J, Lee SH, Goddard ME, Visscher PM: **GCTA: a tool for genome-wide complex trait analysis.** *The American Journal of Human Genetics* 2011, **88:**76-82.
121. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome research* 2009, **19:**1655-1664.
122. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27:**2156-2158.
123. Korneliussen TS, Albrechtsen A, Nielsen R: **ANGSD: analysis of next generation sequencing data.** *BMC bioinformatics* 2014, **15:**1-13.
124. Zhang C, Dong SS, Xu JY, He WM, Yang TL: **PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files.** *Bioinformatics* 2019, **35**.
125. Schiffels S, Durbin R: **Inferring human population size and separation history from multiple genome sequences.** *Nature genetics* 2014, **46:**919-925.
126. Delaneau O, Marchini J, Zagury J-F: **A linear complexity phasing method for thousands of genomes.** *Nature methods* 2012, **9:**179-181.
127. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.
128. Wang K, Mathieson I, O'Connell J, Schiffels S: **Tracking human population structure through time from whole genome sequences.** *PLoS genetics* 2020, **16:**e1008552-e1008552.
129. Araújo MB, New M: **Ensemble forecasting of species distributions.** *Trends in ecology & evolution* 2007, **22:**42-47.
130. Thuiller W, Georges D, Engler R, Breiner F: **biomod2: Ensemble platform for species distribution modeling. R package version 3.1-64.** *Availablt at:* http://CRANR-projectorg/package= *biomod2 (accessed February 2015)* 2014.
131. Thuiller W, Lafourcade B, Engler R, Araújo MB: **BIOMOD–a platform for ensemble forecasting of species distributions.** *Ecography* 2009, **32:**369-373.
132. McCullagh P, Nelder JA: *Generalized linear models.* Routledge; 2019.
133. Hastie TJ, Tibshirani RJ: *Generalized additive models.* Routledge; 2017.
134. De'Ath G: **Boosted trees for ecological modeling and prediction.** *Ecology* 2007, **88:**243-251.
135. Breiman L, Friedman JH, Olshen RA, Stone CJ: *Classification and regression trees.* Routledge;

2017.

136. Leathwick J, Rowe D, Richardson J, Elith J, Hastie T: **Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish.** *Freshwater Biology* 2005, **50:**2034-2052.
137. Breiman L: **Random forests.** *Machine learning* 2001, **45:**5-32.
138. Ripley BD: *Pattern recognition and neural networks.* Cambridge university press; 2007.
139. Phillips SJ, Anderson RP, Schapire RE: **Maximum entropy modeling of species geographic distributions.** *Ecological modelling* 2006, **190:**231-259.
140. Hijmans RJ, Phillips S, Leathwick J, Elith J, Hijmans MRJ: **Package 'dismo'.** *Circles* 2017, **9:**1-68.
141. Barbet‐Massin M, Jiguet F, Albert CH, Thuiller W: **Selecting pseudo‐absences for species distribution models: how, where and how many?** *Methods in ecology and evolution* 2012, **3:**327-338.
142. Allouche O, Tsoar A, Kadmon R: **Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS).** *Journal of applied ecology* 2006, **43:**1223-1232.
143. Fick SE, Hijmans RJ: **WorldClim 2: new 1‐km spatial resolution climate surfaces for global land areas.** *International journal of climatology* 2017, **37:**4302-4315.
144. Title PO, Bemmels JB: **ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling.** *Ecography* 2018, **41:**291-307.
145. Trabucco A, Zomer RJ: **Global aridity index and potential evapotranspiration (ET0) climate database v2.** *CGIAR Consort Spat Inf* 2018.
146. Barbosa AM: **fuzzySim: applying fuzzy logic to binary similarity indices in ecology.** *Methods in Ecology and Evolution* 2015, **6:**853-858.
147. Hijmans RJ, Van Etten J, Cheng J, Mattiuzzi M, Sumner M, Greenberg JA, Lamigueiro OP, Bevan A, Racine EB, Shortridge A: **Package 'raster'.** *R package* 2015, **734**.
148. Pavlidis P, Živković D, Stamatakis A, Alachiotis N: **SweeD: Likelihood-based detection of selective sweeps in thousands of genomes.** *Molecular Biology and Evolution* 2013, **30:**2224-2234.
149. Alachiotis N, Stamatakis A, Pavlidis P: **OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets.** *Bioinformatics* 2012, **28:**2274-2275.
150. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C: **Genomic scans for selective sweeps using SNP data.** *Genome research* 2005, **15:**1566-1575.
151. Sharifova S, Shahmuradov I, Nakayama H, Rowland S, Cheng Z, Zumstain K, Sinha N: **RNA-Seq analysis of drought responsive transcriptome of Solanum Chilense.** *in prep* 2021.
152. Staab PR, Zhu S, Metzler D, Lunter G: **scrm: Efficiently simulating long sequences using the approximated coalescent with recombination.** *Bioinformatics* 2015, **31:**1680-1682.
153. Booker TR, Yeaman S, Whitlock MC: **Variation in recombination rate affects detection of outliers in genome scans under neutrality.** *Molecular Ecology* 2020, **29:**4274-4279.
154. Stevison LS, McGaugh SE: **It's time to stop sweeping recombination rate under the genome scan rug.** Wiley Online Library; 2020.
155. Tournebize R, Poncet V, Jakobsson M, Vigouroux Y, Manel S: **McSwan: A joint site frequency spectrum method to detect and date selective sweeps across multiple population genomes.** *Molecular ecology resources* 2019, **19:**283-295.

156. Hudson RR: **Generating samples under a Wright–Fisher neutral model of genetic variation.** *Bioinformatics* 2002, **18:**337-338.
157. Conesa A, Götz S: **Blast2GO: a comprehensive suite for functional analysis in plant genomics.** *International journal of plant genomics* 2008, **2008**.
158. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21:**3674-3676.
159. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E: **The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome.** *genesis* 2015, **53:**474-485.
160. Yu G, Wang L-G, Han Y, He Q-Y: **clusterProfiler: an R package for comparing biological themes among gene clusters.** *Omics: a journal of integrative biology* 2012, **16:**284-287.
161. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT: **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.** *Nucleic acids research* 2010, **38:**W214-W220.
162. Yuen RKC, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel RV, Whitney J, Deflaux N, Bingham J, Wang Z: **Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder.** *Nature neuroscience* 2017, **20:**602-611.
163. Forester BR, Lasky JR, Wagner HH, Urban DL: **Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations.** *Molecular Ecology* 2018, **27:**2215-2233.
164. Lisiecki LE, Raymo ME: **A Pliocene - Pleistocene stack of 57 globally distributed benthic δ 18O records.** *Paleoceanography* 2005, **20**.
165. Ritter B, Wennrich V, Medialdea A, Brill D, King G, Schneiderwind S, Niemann K, Fernández-Galego E, Diederich J, Rolf C: **Climatic fluctuations in the hyperarid core of the Atacama Desert during the past 215 ka.** *Scientific reports* 2019, **9:**1-13.
166. Koropoulis A, Alachiotis N, Pavlidis P: **Detecting positive selection in populations using genetic data.** In: Humana, New York, NY; 2020: 87-123
167. Tellier A, Laurent SJ, Lainer H, Pavlidis P, Stephan W: **Inference of seed bank parameters in two wild tomato species using ecological and genetic data.** *Proceedings of the National Academy of Sciences* 2011, **108:**17052-17057.
168. Tellier A: **Persistent seed banking as eco - evolutionary determinant of plant nucleotide diversity: novel population genetics insights.** *New Phytologist* 2019, **221:**725-730.
169. Živković D, Tellier A: **All but sleeping? Consequences of soil seed banks on neutral and selective diversity in plant species.** In *Mathematical Modelling in Plant Biology.* Springer; 2018: 195-212
170. Lewontin RC, Krakauer J: **Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms.** *Genetics* 1973, **74:**175-195.
171. Enard D, Messer PW, Petrov DA: **Genome-wide signals of positive selection in human evolution.** *Genome research* 2014, **24:**885-895.
172. Lotterhos KE: **The effect of neutral recombination variation on genome scans for selection.** *G3: Genes, Genomes, Genetics* 2019, **9:**1851-1867.
173. Johansson M, Staiger D: **Time to flower: interplay between photoperiod and the circadian clock.** *Journal of experimental botany* 2015, **66:**719-730.

174. Song YH, Shim JS, Kinmonth-Schultz HA, Imaizumi T: **Photoperiodic flowering: time measurement mechanisms in leaves.** *Annual review of plant biology* 2015, **66:**441-464.
175. Guo X, Liu D, Chong K: **Cold signaling in plants: Insights into mechanisms and regulation.** *Journal of integrative plant biology* 2018, **60:**745-756.
176. Iida H, Mähönen AP: **Growth-mediated sensing of long-term cold in plants.** Nature Publishing Group; 2020.
177. Michaels SD, Himelblau E, Kim SY, Schomburg FM, Amasino RM: **Integration of flowering signals in winter-annual Arabidopsis.** *Plant Physiology* 2005, **137:**149-156.
178. Putterill J, Varkonyi-Gasic E: **FT and florigen long-distance flowering control in plants.** *Current Opinion in Plant Biology* 2016, **33:**77-82.
179. Sheldon CC, Rouse DT, Finnegan EJ, Peacock WJ, Dennis ES: **The molecular basis of vernalization: the central role of FLOWERING LOCUS C (FLC).** *Proceedings of the National Academy of Sciences* 2000, **97:**3753-3758.
180. Turck F, Fornara F, Coupland G: **Regulation and identity of florigen: FLOWERING LOCUS T moves center stage.** *Annu Rev Plant Biol* 2008, **59:**573-594.
181. Jiang H, Zhang X, Chen X, Aramsangtienchai P, Tong Z, Lin H: **Protein lipidation: occurrence, mechanisms, biological functions, and enabling technologies.** *Chemical reviews* 2018, **118:**919-988.
182. Maksimov EG, Mironov KS, Trofimova MS, Nechaeva NL, Todorenko DA, Klementiev KE, Tsoraev GV, Tyutyaev EV, Zorina AA, Feduraev PV: **Membrane fluidity controls redox-regulated cold stress responses in cyanobacteria.** *Photosynthesis research* 2017, **133:**215-223.
183. Li X, Cai W, Liu Y, Li H, Fu L, Liu Z, Xu L, Liu H, Xu T, Xiong Y: **Differential TOR activation and cell proliferation in Arabidopsis root and shoot apexes.** *Proceedings of the National Academy of Sciences* 2017, **114:**2765-2770.
184. Xiong Y, McCormack M, Li L, Hall Q, Xiang C, Sheen J: **Glucose–TOR signalling reprograms the transcriptome and activates meristems.** *Nature* 2013, **496:**181-186.
185. He Z-H, Cheeseman I, He D, Kohorn BD: **A cluster of five cell wall-associated receptor kinase genes, Wak1–5, are expressed in specific organs of Arabidopsis.** *Plant molecular biology* 1999, **39:**1189-1196.
186. Lally D, Ingmire P, Tong H-Y, He Z-H: **Antisense expression of a cell wall–associated protein kinase, WAK4, inhibits cell elongation and alters morphology.** *The Plant Cell* 2001, **13:**1317-1332.
187. Bengough AG, Loades K, McKenzie BM: **Root hairs aid soil penetration by anchoring the root surface to pore walls.** *Journal of Experimental Botany* 2016, **67:**1071-1078.
188. White RG, Kirkegaard JA: **The distribution and abundance of wheat roots in a dense, structured subsoil–implications for water uptake.** *Plant, cell & environment* 2010, **33:**133-148.
189. Forni C, Duca D, Glick BR: **Mechanisms of plant response to salt and drought stress and their alteration by rhizobacteria.** *Plant and Soil* 2017, **410:**335-356.
190. Zhao C, Zhang H, Song C, Zhu J-K, Shabala S: **Mechanisms of plant responses and adaptation to soil salinity.** *The innovation* 2020, **1:**100017.
191. Capblancq T, Luu K, Blum MG, Bazin E: **Evaluation of redundancy analysis to identify signatures of local adaptation.** *Molecular Ecology Resources* 2018, **18:**1223-1233.

192. Jezkova T, Olah ‐ Hemmings V, Riddle BR: **Niche shifting in response to warming climate after the last glacial maximum: inference from genetic data and niche assessments in the chisel ‐ toothed kangaroo rat (D ipodomys microps).** *Global Change Biology* 2011, **17:**3486-3502.

193. Sillero N, Reis M, Vieira C, Vieira J, Morales ‐ Hojas R: **Niche evolution and thermal adaptation in the temperate species Drosophila americana.** *Journal of Evolutionary Biology* 2014, **27:**1549-1561.

194. Kottler EJ, Dickman EE, Sexton JP, Emery NC, Franks SJ: **Draining the swamping hypothesis: little evidence that gene flow reduces fitness at range edges.** *Trends in Ecology & Evolution* 2021.

195. Gould SJ, Vrba ES: **Exaptation—a missing term in the science of form.** *Paleobiology* 1982, **8:**4-15.

196. Larson G, Stephens PA, Tehrani JJ, Layton RH: **Exapting exaptation.** *Trends in ecology & evolution* 2013, **28:**497-498.

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

198. Martinez J-P, Antunez A, Pertuze R, Acosta MDELP, Palma X, Fuentes L, Ayala A, Araya H, Lutts S: **Effects of saline water on water status, yield and fruit quality of wild (Solanum chilense) and domesticated (Solanum lycopersicum var. cerasiforme) tomatoes.** *Experimental Agriculture* 2012, **48:**573-586.

199. Fischer I, Steige KA, Stephan W, Mboup M: **Sequence evolution and expression regulation of stress-responsive genes in natural populations of wild tomato.** *PLoS One* 2013, **8:**e78182-e78182.

200. Li T, Yang X, Yu Y, Si X, Zhai X, Zhang H, Dong W, Gao C, Xu C: **Domestication of wild tomato is accelerated by genome editing.** *Nature biotechnology* 2018, **36:**1160-1163.

201. Bastow R, Mylne JS, Lister C, Lippman Z, Martienssen RA, Dean C: **Vernalization requires epigenetic silencing of FLC by histone methylation.** *Nature* 2004, **427:**164-167.

202. Alexandre CM, Hennig L: **FLC or not FLC: the other side of vernalization.** *Journal of experimental botany* 2008, **59:**1127-1135.

203. Sheldon CC, Jean Finnegan E, Dennis ES, James Peacock W: **Quantitative effects of vernalization on FLC and SOC1 expression.** *The Plant Journal* 2006, **45:**871-883.

204. Liu C, Chen H, Er HL, Soo HM, Kumar PP, Han JH, Liou YC, Yu H: **Direct interaction of AGL24 and SOC1 integrates flowering signals in Arabidopsis.** *Development* 2008, **135:**1481-1491.

205. Abe M, Kobayashi Y, Yamamoto S, Daimon Y, Yamaguchi A, Ikeda Y, Ichinoki H, Notaguchi M, Goto K, Araki T: **FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex.** *Science* 2005, **309:**1052-1056.

206. Wigge PA, Kim MC, Jaeger KE, Busch W, Schmid M, Lohmann JU, Weigel D: **Integration of spatial and temporal information during floral induction in Arabidopsis.** *Science* 2005, **309:**1056-1059.

207. Thalmann M, Pazmino D, Seung D, Horrer D, Nigro A, Meier T, Kölling K, Pfeifhofer HW, Zeeman SC, Santelia D: **Regulation of leaf starch degradation by abscisic acid is important for osmotic stress tolerance in plants.** *The Plant Cell* 2016, **28:**1860-1878.

208. Niu Y, Xiang Y: **An overview of biomembrane functions in plant responses to**

high-temperature stress. *Frontiers in plant science* 2018, **9:**915.

209. Hutagalung AH, Novick PJ: **Role of Rab GTPases in membrane traffic and cell physiology.** *Physiological reviews* 2011, **91:**119-149.
210. Hegelund JN, Jahn TP, Baekgaard L, Palmgren MG, Schjoerring JK: **Transmembrane nine proteins in yeast and Arabidopsis affect cellular metal contents without changing vacuolar morphology.** *Physiologia plantarum* 2010, **140:**355-367.
211. Holmgren M, Scheffer M, Ezcurra E, Gutiérrez JR, Mohren GMJ: **El Niño effects on the dynamics of terrestrial ecosystems.** *Trends in Ecology & Evolution* 2001, **16:**89-94.
212. Gutiérrez JR, Meserve PL: **El Niño effects on soil seed bank dynamics in north-central Chile.** *Oecologia* 2003, **134:**511-517.
213. Chetelat RT, Pertuzé RA, Faúndez L, Graham EB, Jones CM: **Distribution, ecology and reproductive biology of wild tomatoes and related nightshades from the Atacama Desert region of northern Chile.** *Euphytica* 2009, **167:**77-93.
214. Xiong L, Wang R-G, Mao G, Koczan JM: **Identification of drought tolerance determinants by genetic analysis of root response to drought stress and abscisic acid.** *Plant physiology* 2006, **142:**1065-1074.
215. Farooq M, Hussain M, Wahid A, Siddique KHM: **Drought stress in plants: an overview.** *Plant responses to drought stress* 2012**:**1-33.
216. Kim Y, Chung YS, Lee E, Tripathi P, Heo S, Kim K-H: **Root response to drought stress in rice (Oryza sativa L.).** *International journal of molecular sciences* 2020, **21:**1513-1513.
217. Deng K, Dong P, Wang W, Feng L, Xiong F, Wang K, Zhang S, Feng S, Wang B, Zhang J: **The TOR pathway is involved in adventitious root formation in Arabidopsis and potato.** *Frontiers in Plant Science* 2017, **8:**784.
218. Dong Y, Silbermann M, Speiser A, Forieri I, Linster E, Poschet G, Samami AA, Wanatabe M, Sticht C, Teleman AA: **Sulfur availability regulates plant growth via glucose-TOR signaling.** *Nature communications* 2017, **8:**1-10.
219. Wilkinson MJ, Roda F, Walter GM, James ME, Nipper R, Walsh J, Allen SL, North HL, Beveridge CA, Ortiz-Barrientos D: **Adaptive divergence in shoot gravitropism creates hybrid sterility in an Australian wildflower.** *bioRxiv* 2021**:**845354.
220. Golldack D, Li C, Mohan H, Probst N: **Tolerance to drought and salt stress in plants: unraveling the signaling networks.** *Frontiers in plant science* 2014, **5:**151-151.
221. Ryu H, Cho Y-G: **Plant hormones in salt stress tolerance.** *Journal of Plant Biology* 2015, **58:**147-155.
222. Fariduddin Q, Yusuf M, Ahmad I, Ahmad A: **Brassinosteroids and their role in response of plants to abiotic stresses.** *Biologia Plantarum* 2014, **58:**9-17.
223. Tanaka K, Asami T, Yoshida S, Nakamura Y, Matsuo T, Okamoto S: **Brassinosteroid homeostasis in Arabidopsis is ensured by feedback expressions of multiple genes involved in its metabolism.** *Plant physiology* 2005, **138:**1117-1125.
224. Wei Z, Li J: **Regulation of brassinosteroid homeostasis in higher plants.** *Frontiers in Plant Science* 2020, **11:**1480.
225. Bushnell B: **BBTools software package.** *URL* http://sourceforge *net/projects/bbmap* 2014, **578:**579.
226. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome biology* 2014, **15:**1-21.

227. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics* 2014, **30:**923-930.
228. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC bioinformatics* 2008, **9:**1-13.
229. Guo A-Y, Chen X, Gao G, Zhang H, Zhu Q-H, Liu X-C, Zhong Y-F, Gu X, He K, Luo J: **PlantTFDB: a comprehensive plant transcription factor database.** *Nucleic Acids Research* 2007, **36:**D966-D969.
230. Tian F, Yang D-C, Meng Y-Q, Jin J, Gao G: **PlantRegMap: charting functional regulatory maps in plants.** *Nucleic acids research* 2020, **48:**D1104-D1113.
231. Bailey TL, Johnson J, Grant CE, Noble WS: **The MEME suite.** *Nucleic acids research* 2015, **43:**W39-W49.
232. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC bioinformatics* 2009, **10:**1-9.
233. Shen W, Ren H: **TaxonKit: A practical and efficient NCBI taxonomy toolkit.** *Journal of Genetics and Genomics* 2021.
234. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.** *Nucleic acids research* 2006, **34:**W609-W612.
235. Comeron JM: **A method for estimating the numbers of synonymous and nonsynonymous substitutions per site.** *Journal of molecular evolution* 1995, **41:**1152-1159.
236. Hedges SB, Marin J, Suleski M, Paymer M, Kumar S: **Tree of life reveals clock-like speciation and diversification.** *Molecular biology and evolution* 2015, **32:**835-845.
237. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic acids research* 2002, **30:**1575-1584.
238. Han MV, Thomas GW, Lugo-Martinez J, Hahn MW: **Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3.** *Molecular biology and evolution* 2013, **30:**1987-1997.
239. Hamant O, Pautot V: **Plant development: a TALE story.** *Comptes rendus biologies* 2010, **333:**371-381.
240. Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, Taylor CM, Yin Y, Fulton L, Minx P: **The draft genome of the parasitic nematode Trichinella spiralis.** *Nature genetics* 2011, **43:**228-235.
241. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G: **The coffee genome provides insight into the convergent evolution of caffeine biosynthesis.** *science* 2014, **345:**1181-1184.
242. Chen S, Krinsky BH, Long M: **New genes as drivers of phenotypic evolution.** *Nature Reviews Genetics* 2013, **14:**645-660.
243. Xia E-H, Zhang H-B, Sheng J, Li K, Zhang Q-J, Kim C, Zhang Y, Liu Y, Zhu T, Li W, et al: **The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis.** *Molecular Plant* 2017, **10:**866-877.
244. Danilevskaya ON, Yu G, Meng X, Xu J, Stephenson E, Estrada S, Chilakamarri S, Zastrow‐Hayes G, Thatcher S: **Developmental and transcriptional responses of maize to drought stress under field conditions.** *Plant Direct* 2019, **3:**e00129.
245. Farooq M, Wahid A, Kobayashi N, Fujita D, Basra S: **Plant drought stress: effects, mechanisms and management.** *Sustainable agriculture* 2009**:**153-188.

246. Alves AA, Setter TL: **Response of cassava leaf area expansion to water deficit: cell proliferation, cell expansion and delayed development.** *Annals of Botany* 2004, **94:**605-613.
247. Verelst W, Bertolini E, De Bodt S, Vandepoele K, Demeulenaere M, Pè ME, Inzé D: **Molecular and physiological analysis of growth-limiting drought stress in Brachypodium distachyon leaves.** *Molecular Plant* 2013, **6:**311-322.
248. Reddy AR, Chaitanya KV, Vivekanandan M: **Drought-induced responses of photosynthesis and antioxidant metabolism in higher plants.** *Journal of plant physiology* 2004, **161:**1189-1202.
249. Wilkinson S, Davies WJ: **Drought, ozone, ABA and ethylene: new insights from cell to plant to community.** *Plant, cell & environment* 2010, **33:**510-525.
250. Fàbregas N, Fernie AR: **The metabolic response to drought.** *Journal of Experimental Botany* 2019, **70:**1077-1085.
251. Arendsee ZW, Li L, Wurtele ES: **Coming of age: orphan genes in plants.** *Trends in plant science* 2014, **19:**698-708.
252. Mustafin ZS, Zamyatin VI, Konstantinov DK, Doroshkov AV, Lashin SA, Afonnikov DA: **Phylostratigraphic analysis shows the earliest origination of the abiotic stress associated genes in a. Thaliana.** *Genes* 2019, **10:**963.
253. Mencuccini M, Munné-Bosch S: **13 Physiological and Biochemical Processes Related to Ageing and Senescence in Plants.** *The evolution of senescence in the tree of life* 2017**:**257.
254. Wang B, Regulski M, Tseng E, Olson A, Goodwin S, McCombie WR, Ware D: **A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing.** *Genome research* 2018, **28:**921-932.
255. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M: **The evolution of gene expression levels in mammalian organs.** *Nature* 2011, **478:**343-348.
256. Harrison PW, Wright AE, Mank JE: **The evolution of gene expression and the transcriptome–phenotype relationship.** In *Seminars in cell & developmental biology*. Elsevier; 2012: 222-229.
257. Ghanbarian AT, Hurst LD: **Neighboring genes show correlated evolution in gene expression.** *Molecular biology and evolution* 2015, **32:**1748-1766.
258. Dubois M, Inzé D: **Plant growth under suboptimal water conditions: early responses and methods to study them.** *Journal of experimental botany* 2020, **71:**1706-1722.
259. Zhang D-P: *Abscisic acid: metabolism, transport and signaling.* Springer; 2014.
260. Vishwakarma K, Upadhyay N, Kumar N, Yadav G, Singh J, Mishra RK, Kumar V, Verma R, Upadhyay R, Pandey M: **Abscisic acid signaling and abiotic stress tolerance in plants: a review on current knowledge and future prospects.** *Frontiers in plant science* 2017, **8:**161.
261. Wu X, Sharp PA: **Divergent transcription: a driving force for new gene origination?** *Cell* 2013, **155:**990-996.
262. Harris SE, Munshi-South J: **Scans for positive selection reveal candidate genes and local adaptation of.** *bioRxiv* 2016.
263. Weigand H, Leese F: **Detecting signatures of positive selection in non-model species using genomic data.** *Zoological Journal of the Linnean Society* 2018, **184:**528-583.
264. Rymer PD, Umbers KDL, Dudaniec RY, Ahrens CW, Bragg J, Stow A, Dillon S: **The search for loci under selection: trends, biases and progress.** *Molecular Ecology* 2018**:**1342-1356.

265. Pavlidis P, Alachiotis N: **A survey of methods and tools to detect recent and strong positive selection.** *Journal of Biological Research-Thessaloniki* 2017**:**1-17.
266. Todaka D, Shinozaki K, Yamaguchi-Shinozaki K: **Recent advances in the dissection of drought-stress regulatory networks and strategies for development of drought-tolerant transgenic rice plants.** *Frontiers in plant science* 2015, **6:**84.
267. Davidson EH: *The regulatory genome: gene regulatory networks in development and evolution.* Elsevier; 2010.
268. Seki M, Umezawa T, Urano K, Shinozaki K: **Regulatory metabolic networks in drought stress responses.** *Current opinion in plant biology* 2007, **10:**296-302.
269. Rougeux C, Gagnaire PA, Praebel K, Seehausen O, Bernatchez L: **Polygenic selection drives the evolution of convergent transcriptomic landscapes across continents within a Nearctic sister species complex.** *Molecular ecology* 2019, **28:**4388-4403.
270. Nevo E: **Evolution of genome–phenome diversity under environmental stress.** *Proceedings of the National Academy of Sciences* 2001, **98:**6233-6240.
271. Grime JP, Pierce S: *The evolutionary strategies that shape ecosystems.* John Wiley & Sons; 2012.
272. Farooq M, Hussain M, Wahid A, Siddique K: **Drought stress in plants: an overview.** *Plant responses to drought stress* 2012**:**1-33.
273. Ruprecht C, Vaid N, Proost S, Persson S, Mutwil M: **Beyond genomics: studying evolution with gene coexpression networks.** *Trends in Plant Science* 2017, **22:**298-307.
274. Gupta MD, Tsiantis M: **Gene networks and the evolution of plant morphology.** *Current opinion in plant biology* 2018, **45:**82-87.
275. Hall MC, Willis JH: **Divergent selection on flowering time contributes to local adaptation in Mimulus guttatus populations.** *Evolution* 2006, **60:**2466-2477.
276. Wittwer SH, Teubner FG: **The effects of temperature and nitrogen nutrition on flower formation in the tomato.** *American Journal of Botany* 1957**:**125-129.
277. Chen Q, Payyavula RS, Chen L, Zhang J, Zhang C, Turgeon R: **FLOWERING LOCUS T mRNA is synthesized in specialized companion cells in Arabidopsis and Maryland Mammoth tobacco leaf veins.** *Proceedings of the National Academy of Sciences* 2018, **115:**2830-2835.
278. Wingler A: **Interactions between flowering and senescence regulation and the influence of low temperature in Arabidopsis and crop plants.** *Annals of Applied Biology* 2011, **159:**320-338.
279. Song YH, Kubota A, Kwon MS, Covington MF, Lee N, Taagen ER, Cintrón DL, Hwang DY, Akiyama R, Hodge SK: **Molecular basis of flowering under natural long-day conditions in Arabidopsis.** *Nature Plants* 2018, **4:**824-835.
280. Shrestha R, Gómez-Ariza J, Brambilla V, Fornara F: **Molecular control of seasonal flowering in rice, arabidopsis and temperate cereals.** *Annals of botany* 2014, **114:**1445-1458.
281. Blanc G, Wolfe KH: **Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes.** *The plant cell* 2004, **16:**1667-1678.
282. Mahmoud M, Zywicki M, Twardowski T, Karlowski WM: **Efficiency of PacBio long read correction by 2nd generation Illumina sequencing.** *Genomics* 2019, **111:**43-49.
283. Liu C, Zhang X, Zhang K, An H, Hu K, Wen J, Shen J, Ma C, Yi B, Tu J: **Comparative analysis of the Brassica napus root and leaf transcript profiling in response to drought stress.** *International journal of molecular sciences* 2015, **16:**18752-18777.