OXFORD UNIVERSITY PRESS | Molecular Biology and Evolution

Copy number variations shape genomic structural diversity underpinning ecological adaptation in the wild tomato *Solanum chilense*

Journal:	*Molecular Biology and Evolution*
Manuscript ID	MBE-23-1089
Manuscript Type:	Discoveries
Date Submitted by the Author:	12-Dec-2023
Complete List of Authors:	Wei, Kai; Technical University of Munich, Department of Life Science Systems Stam, Remco; Christian Albrechts University, Department of Phytopathology and Crop Protection Tellier, Aurélien; Technical University of Munich, Department of Life Science Systems Silva Arias, Gustavo; Technical University of Munich, Department of Life Science Systems; Universidad Nacional de Colombia
Keywords:	wild tomato, Atacama, Copy number variation, Abiotic stress, Flowering time, Adaptation

Copy number variations shape genomic structural diversity underpinning ecological adaptation in the wild tomato *Solanum chilense*

Kai Wei[1]*, Remco Stam[2], Aurélien Tellier[1]*, Gustavo A Silva-Arias[1,3*]

[1]Professorship for Population Genetics, Department of Life Science Systems, School of Life Sciences, Technical University of Munich, Liesel-Beckmann Strasse 2, 85354 Freising, Germany

[2]Department of Phytopathology and Crop Protection, Institute of Phytopathology, Faculty of Agricultural and Nutritional Sciences, Christian Albrechts University, Hermann Rodewald Str 9, 24118, Kiel, Germany

[3]Instituto de Ciencias Naturales, Facultad de Ciencias, Universidad Nacional de Colombia, Sede Bogotá, Av. Carrera 30 # 45-03, 111321, Bogotá, Colombia

*Corresponding authors: Aurélien Tellier: aurelien.tellier@tum.de
Kai Wei: kai.wei@tum.de
Gustavo A. Silva-Arias: gasilvaa@unal.edu.co

Abstract

Copy Number Variations (CNVs) are genomic structural changes constituting genetic diversity and underpinning rapid ecological adaptation. The timing of, and the target genes involved in adaptation through CNVs in the tomato and wild relative lineage still need to be explored at the population level. Therefore, we characterise the CNV landscape of *Solanum chilense*, a wild tomato species, using whole-genome data of 35 individuals from seven populations distributed in contrasting environments. We identify 212,207 CNVs, including 160,926 deletions and 51,281 duplications. We find CNVs for intergenic and coding regions and a higher number of CNVs in diverging populations occupying stressful habitats. CNV and single nucleotide polymorphism analyses concordantly reveal the known species' population structure, underscoring the impact of historical demographic and recent colonisation events on the distribution of CNVs. Furthermore, we identify 3,539 candidate genes with highly divergent CNV profiles across populations. Interestingly, these genes are functionally associated with response to abiotic stimuli and stress and linked to multiple pathways of flowering time regulation. Gene CNV exhibits two evolutionary trends: a contraction with gene loss in central and southern coast populations and an expansion with gene gain in the southern highland group. Environmental association of the CNVs ultimately links the dynamics of gene copy number to six climatic variables. It suggests that natural selection has likely shaped CNV patterns in response to the climatic changes during the recent range expansion of *S. chilense*. Our findings provide insights into the role of CNVs underlying adaptation in marginal populations.

Introduction

Copy number variation (CNV) is the primary type of structural variation (SV) caused by genomic rearrangement, which mainly includes deletion (DEL) and duplication (DUP) events resulting from the loss and gain of DNA segments (Feuk, et al. 2006; Żmieńko, et al. 2014). It is expected that CNV has a more significant impact on gene function because it covers more base-pairs (Shaikh, et al. 2009; Hämälä, et al. 2021) and has a higher per-locus mutation rate than ~~point mutations (single nucleotide polymorphisms, SNPs)~~ (Lupski 2007). CNV is recognised as an essential driver of genomic divergence and local adaptation (Rinker, et al. 2019; Hämälä, et al. 2021; Marszalek-Zenczak, et al. 2023). Genome-wide studies confirm the importance of CNVs as the basis of stress response and yield ~~improvement~~ in multiple plants, such as maise (Springer, et al. 2009), rice (Fuentes, et al. 2019; Qin, et al. 2021), and *Arabidopsis thaliana* (Zmienko, et al. 2020; Marszalek-Zenczak, et al. 2023). However, such studies have been so far conducted in selfing species and/or crops characterised by small population size and domestication bottlenecks (Alonso-Blanco, et al. 2016; Beissinger, et al. 2016; Brumlop, et al. 2019). Therefore, it is difficult in such species to disentangle the effect of random evolutionary processes (genetic drift, chromosomal rearrangements, and demographic history) generating fast and extensive CNVs between populations from the impact of adaptive processes (here positive selection underpinning environmental adaptation). In addition, the dynamic of gene copy number also reflects population history and multiple events, including selection, migration and recombination (Sudmant, et al. 2015; Zhou, et al. 2019; Otto, et al. 2022; Antinucci, et al. 2023; Otto and Wiehe 2023). Indeed, the effective ~~population~~ size (N_e) of populations determines the efficiency of positive and negative selection against genetic drift, as well as the amount of genetic diversity (SNP or CNV) available, thus being a major determinant of the genome architecture (Lynch and Walsh 2007).

The tomato wild relative species *Solanum chilense* is ~~proven to be~~ an excellent model species to study the genetic basis of adaptive evolution ~~when colonising~~ novel habitats (Böndel, et al. 2015; Stam, et al. 2019b; Wei, et al. 2023b). The presence of outcrossing, gene flow, seed banks, and relatively mild bottlenecks during the colonisation of new habitats results in high effective population sizes (*Ne*) as reflected by high nucleotide diversity and high recombination rates, meaning that this species has a high adaptive potential (Arunyawat, et al. 2007; Stam, et al. 2019b; Wei, et al. 2023b). *S. chilense* occurs in southern Peru and northern Chile, from mesic to very arid habitats around the Atacama Desert, and becomes the southern-most distrusted species in the tomato clade (Nakazato, et al. 2010). Moreover, within *S. chilense*, two groups of populations expanded southward during two independent colonisation events (Böndel, et al. 2015; Stam, et al. 2019b; Wei, et al. 2023b): one towards the coastal part of northern Chile (hereafter the southern coast group), and the other towards the high altitudes of the Chilean Andes (southern highland group) (Fig. 1A). The populations currently occurring in the southern coast and southern highland habitats have been shown to exhibit signatures of past positive selection for adaptation to cold, drought, light (photoperiod), heat and biotic stresses (Xia, et al. 2010; Fischer, et al. 2011; Nosenko, et al. 2016; Böndel, et al. 2018; Stam, et al. 2019b; Wei, et al. 2023b). These events of recent positive selection at a cohesive set of genes and drought-responsive gene networks suggest the occurrence of genetic underpinnings to the adaptation of novel habitats during the southward expansion of *S. chilense* populations towards arid areas around the Atacama desert (Wei, et al. 2023a). Previous population genomic studies revealed that these adaptive signatures are based on scans for positive selection solely using single-nucleotide polymorphisms (SNPs). However, whether CNV can also contribute to adaptation to novel habitats in S. chilense and the tomato clade is still unknown.

Reference genomes of several species of the tomato clade, including numerous cultivated tomato varieties, are now sequenced and assembled (Ranjan, et al. 2012; Sato, et al. 2012; Bolger, et al. 2014; Stam, et al. 2019a). Three tomato SV sets have been recently constructed based on a tomato-clade pangenome analysis to investigate the impact of genome rearrangements on gene expression and genomic diversity ~~and provide new genomic resources for the improvement of tomato~~ (Alonge, et al. 2020; Zhou, et al. 2022; Li, et al. 2023). These three studies compare cultivated tomato genomes with that of several wild tomato species, including ~~PacBio and Illumina sequencing from~~ an individual of the *S. chilense* accession LA1969 (belonging to our central group; Fig. 1A). Interestingly, ~~we note that~~ *S. chilense* exhibits the highest number of SV among all wild and cultivated tomato species (Li, et al. 2023). ~~This difference is even more striking when considering that~~ the closer related species *S. peruvianum* and *S. corneliomulleri* show ~~half or fewer~~ SVs ~~than~~ *S. chilense*. All these three species exhibit a similar recent proliferation of transposable elements (Li, et al. 2023). As *S. chilense* has one of the largest genome sizes of the tomato clade and has the highest number of annotated genes, it is crucial to study processes driving ~~gene copy number variation~~ and its relevance for speciation and intraspecific diversification. However, the studies mentioned above ~~focus~~ on the pangenome across species level (wild and cultivated) and an understanding of the role of CNVs in local (ecological) adaptation is still lacking, especially for the adaptation to new arid (southern populations in *S. chilense*) habitats.

We generate whole-genome copy number (CN) profiles for 35 *S. chilense* plants from seven populations (five diploid individuals per population) representing three different geographic habitats: three central (C) populations, two southern highland (SH) populations and two southern coast (SC) populations ~~with different habitats~~ (Fig. 1A; Dataset S1). We first identify ~~candidate~~ genes with highly differentiated CN profiles between populations that are likely candidates of recent positive selection. We then measure the evolutionary trend of CN expansion and contraction across different populations. Finally, we associate the dynamics of gene CN with climatic variables to provide evidence for environmental stresses driving CNV dynamics across populations. Our results suggest that gene CNV contributes to population adaptation to novel habitats in an outcrossing species with a large effective population size and genetic diversity. We illustrate the importance of including an analysis of CN variants to complement genomic scans of recent positive selection based on SNPs.

Results

Summary of CNVs in the genome of *S.chilense*

We identify a total of 212,207 CNVs (160,926 deletions and 51,281 duplications) by aligning each of the 35 whole-genome sequencing datasets (Dataset S1) against a chromosome-level *S. chilense* reference genome (Silva-Arias, et al. 2023) using the combination of four CNV callers (Fig. S1; Dataset S2). We find 73,014 ~~up~~ to 94,621 CNVs per population (Fig. 1B; Table S1) and 31,923 ~~up~~ to 46,579 CNVs per individual (Fig. S1; Table S2). Although the number of deletions in all individuals and populations is much larger than the number of duplications (Fig. 1B; Fig. S1; Table S1 and 2), the size of duplications is larger (39,140 bp +/- 104,577) than deletions and exhibits a skewed distribution (14,052 bp +/- 59,930) (Fig. 1C; Kolmogorov-Smirnov test, *P*=2.2e-16). Deletions are smaller than duplications, as 56% of deletions display a size between 50bp and 1,000bp, against 26% for duplications. We find 37% to 43% of the CNVs to be identified in only one individual for three central populations (Fig. S2; Table S3), while only 12% to 14% of all CNVs are observed in all five individuals of a given population ~~(*i.e.,* CNVs being fixed)~~. Furthermore, the number of CNVs is not homogeneously distributed among populations as more than 20% of CNVs are detected in all five individuals in the southern coast and southern highland populations, especially in the two southern coast populations (25% in SC_LA2932 and 31% in SC_LA4107).

Deletions and duplications are enriched at both ends of the chromosomes (Fig. 1D), consistent with previous studies (Alonge, et al. 2020; Hämälä, et al. 2021; Li, et al. 2023). Although most CNVs (76%

to 79% per population) cover intergenic regions (Fig. 1E; Table S4), about 35% of CNVs are located in genes annotated in the *S. chilense* reference. In addition, 45% and 50% of CNVs across populations overlap with putative regulatory elements 5 kb upstream and 5 kb downstream of genes, respectively. CNVs are typically shaped predominately by transposable elements (Fuentes, et al. 2019; Alonge, et al. 2020), and the annotation reveals that 68% of deletions and 82% of duplications match at least one transposable element annotated in the *S. chilense* genome.

To confirm the validity of our pipeline, which assembles CNV detection from four tools specialised for short-read datasets, we simulated 1,000 deletions and 1,000 duplications with lengths ranging from 50 bp to 1 Mb based on 150 bp short-reads (see supplementary methods). We subsequently detected approximately 90% of simulated CNVs using our pipeline, and the false-positive rate was much smaller than using a single caller (Table S5). Our results, as well as previous claims, indicate that combined multiple callers improve the detection of CNVs and are robust to short-read data (Kosugi, et al. 2019; Mahmoud, et al. 2019; Coutelier, et al. 2022).

CNVs effectively capture ~~the~~ population differentiation

We compare the results of population structure ~~analysis~~ based on genome-wide SNPs and CNVs. The principal component analysis (PCA) based on the genotyped CNV dataset agrees with the clustering patterns from the genome-wide SNP dataset (Fig. 2A; Fig. S3A). We define four genetic subgroups showing strong geographic correspondence. The first principal component (PC1) separates the southern coast populations from inland (central and southern highland) populations, PC2 separates the southern coast subgroup into two clusters (SC_LA2932 and SC_LA4107), and PC3 separates the inland populations into central and southern highland subgroups (Fig. 2A; Fig. S3A). The STRUCTURE analysis confirms this result (Fig. 2B; Fig. S3B with K=4 exhibiting the lowest cross-validation error) and is consistent with the results from the SNP dataset (Fig. S3C; (Wei, et al. 2023b)).

We further explore the differentiation of populations using the V_{ST} statistic, which is analogous to the classically used F_{ST} for SNP data (Redon, et al. 2006). We first compute the V_{ST} values along the whole genome in 10 kb windows of 1 kb step size using two CN quantitative measurements: Control-FREEC (V_{ST}(CN)) and read depth (V_{ST}(RD)) (Table S6). We find a significantly high correlation between these two measures (Pearson's test, *P*=1.06e-07; Fig. S4A). Based on the V_{ST} values, we find similar structure patterns as in previous studies ~~using~~ SNPs (Böndel, et al. 2015; Stam, et al. 2019b; Raduski and Igić 2021; Wei, et al. 2023b), namely the high differentiation between southern coast and inland populations, especially between southern coast and southern highland populations (Table S6). As expected, both V_{ST} estimates (V_{ST}(CN) and V_{ST}(RD)) show a significantly high correlation with F_{ST} (based on SNPs) (Fig. 2C; Fig. S4B; Table S6).

Differentiation of gene CN profiles in different populations

To explore the role of natural selection in shaping CNV frequencies and distribution across populations, we use both V_{ST} measures (V_{ST}(CN) and V_{ST}(RD)) across the 39,245 genes to capture candidate genes under divergent selective pressures by identifying genes with strong CN differentiation across populations (Fig. S5). We perform a permutation test (1,000 times) for each gene using the 35 samples of all seven populations. The candidate genes are identified by considering those that surpass a high differentiation threshold for both V_{ST} measures. The rationale is that high V_{ST} values indicate strong differentiation and possibly adaptive divergence at some CNVs between populations. In total, we obtain 3,539 candidate genes that present CN differentiation across the seven populations (*i.e.,* V_{ST} greater than the ~~maximum~~ 95th percentile of the 1,000 permuted V_{ST} values; Fig. S5; Table S7; Dataset S3) and 2,192 strongly CN-differentiated genes of these belong to the top 99th percentile of the 1,000 permuted V_{ST} values (Fig. S5; Table S7; Dataset S3). In Fig. S6A, we show the distribution of deletions and duplications for these 3,539 candidate genes. Southern highland populations exhibit a pronounced increase in gene gains (duplications) and a minimal reduction in gene loss (deletions), whereas SC populations show a comparatively ~~higher incidence of gene loss~~.

We perform four PCA analyses based on the Control-FREEC–based CN values of 1) all annotated 23,911 genes with CN values (Fig. S6B); 2) the 12,392 genes with V_{ST}(CN)>0 (Fig. S6C); 3) the 3,539 differentiated gene set (Fig. 3A); and 4) the 2,192 strongly differentiated gene set (Fig. S6D). In the PCA based on 23,911 genes with CN values (Fig. S6B), all samples exhibit a cohesive grouping, except SC_LA4107. The southern coast populations separate from the five inland populations (central and southern highland populations) in the second PCA (with V_{ST}(CN)>0; Fig. S6C). This suggests a large difference in the CN range and composition between southern coast and inland populations. Consistent with the PCA based on the genotyped CNVs (and previously on SNP data), PC3 separates the southern highland populations from the central populations when using the differentiated genes CN values (Fig. 3A; Fig. S6D). Note, however, that southern highland populations still show ca. 20% of admixed ancestry coefficients with the central populations (Fig. 2B). These admixture signatures can be interpreted as either gene flow post-colonization of the southern habitats between southern highland and central populations or that the divergence time is very short. Consequently, similar polymorphisms in some parts of the genome are maintained between these populations (Wei, et al. 2023b). These results indicate that the past demographic history of habitat colonisation (and the resulting genetic drift) is an important evolutionary process shaping SNP and CNV frequencies within and between populations of *S. chilense*.

Copy number variation illuminates enriched abiotic stress response pathways in *S. chilense*

We perform functional enrichment analysis of the 3,539 CN-differentiated genes according to GO biological process categories (Dataset S4). We classify these significantly enriched GO categories (*P* < 0.05) into nine groups (Fig. S7A) enriched for 82 (cell wall organisation) to 580 (cellular metabolic process) genes. Interestingly, 400 (11.30%) CN-differentiated genes are enriched ~~in~~ response to stimulus/stress that can be linked to multiple environmental factors, for example response to drought (water deprivation; 14.35% with 60 genes), cold (17.62% with 37 genes), heat (26.43% with 39 genes), red/far red light (15.82% with 65 genes), or ultraviolet (UV; 19.03% with 47 genes) (Fig. 3B; Fig. S7A; Dataset S4). These responsive pathways support multiple sources of evidence ~~of adaptive processes~~ at genes associated with responses to arid conditions along a steep altitudinal gradient in *S. chilense* (Fischer, et al. 2011; Nosenko, et al. 2016; Böndel, et al. 2018; Blanchard-Gros, et al. 2021; Wei, et al. 2023b). For instance, multiple drought- (HSF and DREB3), cold- (FAD7), and light/cold-responsive genes (FT, GI, and FLD) for flowering regulation (Dataset S5). This supports that selection pressure is not only linked to point mutations but is also manifested ~~as~~ CNVs.

We find 227 genes associated with flowering (Fig. S7A; Fig. S7B), an important fitness trait conditioning local adaptation in plant species (Srikanth and Schmid 2011). As a critical part of the transition from vegetative to reproductive growth, flowering is influenced by several environmental conditions. Therefore, divergent flowering times and adaptation along the ecological gradient may be related to differential CN-differentiated genes (Fig. S7C). We find 31 and 36 CN differentiated genes in response to light and cold ~~and~~ involved in flowering regulation (Fig. S7C), of which 25 and 20 genes are linked to photoperiod and vernalisation pathways (Fig. S8). The latter represent two regulatory flowering time pathways ~~by~~ the relative lengths of light-dark periods and low temperature, respectively (Srikanth and Schmid 2011; Gaudinier and Blackman 2020). These genes ~~are increasingly duplicated~~ in southern highland populations (Fig. 3A and B; Table S8; t-test, *P* < 0.05). These include the potential homologs of floral integrator genes FT and FD (Liu, et al. 2008; Srikanth and Schmid 2011; Putterill and Varkonyi-Gasic 2016), putative homologs of CRY2, GI, and ELF3 in the photoperiod pathway (Srikanth and Schmid 2011; Makita, et al. 2021), and a putative homolog of AGL14 in the vernalisation pathway (Hecht, et al. 2005; Pérez-Ruiz, et al. 2015). These candidate genes are well-known flowering time regulators in *A. thaliana* (Dataset S5). Note that these potential candidate genes related to flowering regulation are duplicated only in southern highland populations and either no CNV or only copy loss in central and southern coast populations (Fig. 3A and B; Fig. S8; Table S8; t-test, *P* < 0.05). These findings indicate that gene gains in CN may promote colonisation and adaptation in the southern highland habitats by regulating flowering time via the photoperiod and vernalisation pathways (Wei, et al. 2023b). This ~~genomic~~ finding is consistent with the phenology observed in glasshouse conditions, in which southern highland individuals consistently flower 5-10 days earlier than those from central populations. In addition, other potential flowering regulatory genes in the differentiated gene set are likely involved in flowering regulation via different pathways, namely the putative homologs of the genes FY and FLD (Dataset S5) (Srikanth and Schmid 2011; Cheng, et al. 2017; Bao, et al. 2020) (Srikanth and Schmid 2011; Cheng, et al. 2017; Bao, et al. 2020). The FLD gene shows an increased copy number in all populations (Dataset S5; Fig. S8).

We identified 60 drought-responsive CN-differentiated genes associated with direct responses to water deprivation, encompassing duplicated homologs of ABI4 and AFP1 in the abscisic acid (ABA) pathway, along with a putative WRKY33 transcription factor homolog with varying CNs across populations (Fig. 3B; Dataset S4 and S5). These genes are validated as drought stress-responsive in *A. thaliana* and crops (Xiao, et al. 2021; Liu, et al. 2022; Luo, et al. 2022), including WRKY33 also linked to temperature stress in tomato (Guo, et al. 2022). Furthermore, eleven CN-differentiated genes also belong to the drought-response metabolism co-expression network (module) and demonstrated significantly higher expre under drought compared to well-watered conditions (Fig. S9; t-test, P=2.68e-05), corroborating their role in adaptive responses (Wei, et al. 2023a). Interestingly, the comparable numbers of deletion and duplication genes associated with water deprivation response across all populations (Fig. S7D; Table S8) suggest species-wide adaptation processes in *S. chilense* through alterations in a metabolic network.

Our previous SNP study links root development genes to likely local adaptation processes in coastal populations of *S. chilense* (Wei, et al. 2023b). Accordingly, here we find 73 CN-differentiated genes involved in root development, these showing more CNVs in low-altitude populations (C_LA1963, SC_LA2932, SC_LA4107) than in high-altitude populations (C_LA2931, C_LA3111, SH_LA4117A, SH_LA4330) (Fig. 3E; Table S8; t-test, $P < 0.05$).

Gene expansion and contraction patterns show differences along altitudinal gradients

We reveal that a large number of CN-differentiated genes are potentially involved in response to habitat specialisation. To investigate the CN dynamics of these genes across populations, we perform an analysis of gene CN expansion and contraction across populations based on the population *S. chilense* ultrametric tree (Fig. 4A). The CN of the differentiated genes is expanded (CN gain) in the inland group with a high expansion rate of 1.788. At the same time, it is contracted (CN loss) in the southern coast group with a contraction rate of -0.818 (Table 1). Within the inland group, the southern highland group exhibits an expansion of CN (expansion rate of 0.416). In contrast, the central group shows ~~the~~ number of CN losses (contraction rate of -0.767) three times higher than CN gains (Table 1). This likely indicates that the high rate of CN expansion in the inland group is mainly due to southern highland populations exhibiting high CN gains (Table 1). The two southern highland populations show distinct CN expansion rates of 1.663 (SH_LA4117A) and 1.375 (SH_LA4330). In the central group, although the C_LA1963 and C_LA2931 display a trend of CN contraction, the C_LA3111 exhibits a similar rate of CN expansion (1.037) as the southern highland populations (Table 1). We relate this to a high migration rate between the high-altitude C_LA3111 and southern highland populations and/or the recent divergence of the southern highland group from the central group (Wei, et al. 2023b). In addition, the similar highland habitat environments (Fig. 1A) may also contribute to the same evolutionary trends of CN gain affecting a similar set of genes for C_LA3111 and southern highland populations. Interestingly, the opposite results are observed between the two southern coast populations.

Gene CN appears ~~as contraction~~ in SC_LA2932 (contraction rate of -0.935) while expansion occurred in SC_LA4107 (expansion rate of 0.534; Table 1). This follows our previous observation that the two southern coast populations show a high degree of differentiation, possibly resulting from a long time of evolution in isolation. These results are also consistent with the population structure (Fig. 2) and may reflect the old southernmost colonisation of the coastal habitats and the recent colonisation of the highlands (Stam, et al. 2019b; Wei, et al. 2023b). Considering that the reference genome is assembled from population C_LA3111, which probably does not represent the ancestral state of the species, we also perform the same analysis using gene CN profiles calculated from the reference genome of *S. pennellii*, a drought-adapted wild tomato species. Almost consistent results were observed, except for a decrease in the rate of CN expansion in C_LA3111 (Table S9). This may also be a further hint that the dynamics of gene CN may reflect the evolutionary history of populations. Overall, the copy numbers of these potentially adaptively differentiated genes show an expansion (CN gain) in the two previously elucidated southward colonisation events (Fig. 4B; Fig. S10A) (Stam, et al. 2019b; Wei, et al. 2023b).

We define 155 “rapidly evolving genes” that exhibit significantly higher CN expansion or contraction (Viterbi *P* < 0.05) across the different groups/populations using the reference genome of *S. chilense* (Table 1; Dataset S6). The CN profiles of these rapidly evolving genes also clearly support the population clusters in the PCA, but C_LA3111 appears closer to SH populations than to the other central populations (Fig. S10B and C). The highest number of such rapidly evolving genes are found in the southern highland populations (91 genes), including 71 significant CN expanded genes mainly related to photosynthesis of light reaction, long-day photoperiodism (flowering), response to UV light and cold, and 20 significant CN contracted genes primarily associated with developmental and metabolic processes. We also found 56 rapidly CN-evolving genes in the central populations (Table 1; Dataset S6). Few rapidly evolving genes in C_LA3111 and C_LA2931 with high altitudes (above 2200 m) exhibit a significant trend of CN expansion at genes involved in long-day photoperiodism. This confirms that inland populations at high altitudes may exhibit similar CNV signatures of adaptation as highland populations. Among the 51 rapidly evolving genes in the southern coast populations, 16 genes show ~~exactly~~ opposite CN profiles: a significant contraction in SC_LA2932 versus an expansion in SC_LA4107 (Fig. 4C). These genes include few homologs of photosystem subunits (i.e., *psb*B and *pet*D) mainly involved in photosynthesis (Dataset S5) and may underpin the high genetic differentiation at the CNV level between the two southern coast populations. In addition, the same CN rapidly evolving genes enriched for photosynthesis (light reaction) GO categories are also found in central and southern highland groups (Fig. 4D). These potentially photosynthetic gene families appear to have been contracting (CN loss) in the central group and SC_LA2932 but expanding (CN gain) in the southern highland group and SC_LA4107, suggesting that changes in the photosynthetic pathway are also an important adaptive strategy across the different habitats in *S. chilense*.

CN-differentiated genes are associated with climatic variation along the altitudinal gradient

To further support CNV as the genetic underpinning of adaptive response to abiotic factors, we conduct two genome-environment associations (GEA) analyses between gene CN and 37 climate variables (Dataset S7).

We first implement a redundancy analysis (RDA) to identify climate variables significantly associated with CN-differentiated genes across the seven populations. Three climatic variables are observed to correlate with CN changes in the RDA based on 12,391 genes with V_{ST}(CN) > 0. The first three RDA axes (Permutation test, *P* < 0.001) retain 22.62% of the putative adaptive gene CN variances and only weakly distinguish between inland and southern coast populations (Fig. S11B to D). ~~The gene CN differentiation of~~ 52.11% can be explained by six climate variables (explanatory variables) from five significant RDA axes (Permutation test, *P* < 0.001) based on the 3,539 CN differentiated genes (Fig. 5A; Fig. S11E). These climatic variables are significantly correlated with the profiles of the CN-differentiated genes (Mantel test, *P* < 0.05; Fig. 5B). In concordance with the PCA (Fig. 2A), the two main ordination axes do cluster the seven populations into four groups corresponding to the main geographical habitats (central, southern highland and two southern coast habitats). RDA1 is correlated with the annual temperature range (Bio7) and potential evapotranspiration during the driest period (PETDriestQuarter). This axis represents the differentiation between the southern coast and inland populations (Fig. 5A and B). RDA2 reflects the differentiation between two southern coast populations by mean temperature of the wettest quarter (Bio8). RDA2 also summarises a climatic gradient differentiating the low altitude (C_LA1963) and highland populations, which is mainly driven by solar radiation (ann_Rmean) and potential evapotranspiration (annualPET and PETColdestQuarter) (Fig. 5A and B). These six climatic variables are primarily associated with the colonisation of southern highland and southern coast populations (Fig. 5B). The proportions of gene CN differentiation explained by these six climatic variables range from 0.02 (annualPET) to 0.136 (PETColdestQuarter) (Fig. 5C), in which PETColdestQuarter and PETDriestQuarter (0.121) exhibit the highest importance and correlate with inland and southern coast populations, respectively (Fig. 5A to C). Moreover, temperature changes (Bio7 and Bio8) also explain about 20.8% of the gene CN differentiation (Fig. 5C). Solar radiation (ann_Rmean) is a specific variable correlated with high altitude populations and explains 3.6% of gene CN differentiation (Fig. 5A to C). ~~The~~ consistent RDA model is obtained using the 2,192 strongly CN-differentiated genes (Fig. S11F to H). Finally, as a control for the test, we observe a lack of significant RDA model or associated climate variables (Permutation test, *P* > 0.001) when implemented on the 20,372 genes that are not in the CN-differentiated gene set (Fig. S11A).

We subsequently search for candidate genes (among the 3,539 CN-differentiated genes) that may be associated with the six overrepresented climate variables using latent factor mixed models (LFMM) (Fig. S12A) Campo (Frichot et al. 2013; Caye(Frichot, et al. 2013; Caye, et al. 2019). We identify 312 CN-differentiated genes significantly associated with the six climatic variables (z-test; calibrated $P < 0.01$; Fig. S12A and B; Dataset S8). The PCA based on the CN of these 312 candidate genes displays consistent population clustering in the RDA models (Fig. S13A; Fig. 5A), supporting that the six climate variables reflect gene CN ~~dynamic~~ changes across the species distribution. Among these 312 candidates, we find 217 genes to be significantly associated with three PET climate variables (annualPET, PETDriestQuarter, and PETColdestQuarter), of which 98 genes are shared between the three variables (Fig. S12B). Indeed, PET is the primary variable reflecting the drought status of the habitat. We note that these PET-associated CN-differentiated genes are mainly involved in metabolic and root development processes and are found across all populations (Fig. S13B and C). ~~These~~ physiological processes (ABA signalling pathway, root hair differentiation) are essential responses to drought stress ~~using~~ transcriptome and genome analysis (Wei, et al. 2023a; Wei, et al. 2023b). This confirms that drought tolerance is likely the main environmental pressure driving ~~CNV~~ evolution across *S. chilense* distribution. ~~Further,~~ 69% (34/49) of genes associated with Bio7 are also observed to be correlated with ann_Rmean (Fig. S12B); these genes are mainly duplicated in southern highland populations and lost in southern coast populations (Fig. 5D; Table S10). This likely reflects that cold and high solar radiation are challenging conditions in southern highland populations (Dataset S7). Multiple duplicated genes associated with solar radiation (ann_Rmean) are responsive to UV in high-altitude populations (Fig. 5D), such as (likely) homologs of UV-B receptor ARI12, and DNA repair protein REV1 (Dataset S5) (Tossi, et al. 2019; Thompson and Cortez 2020). In addition, we also find a few CN-differentiated genes, such as putative homologs of CPD (Dataset S5), which relate to pigment (anthocyanins) accumulation and are statistically associated with solar radiation variables.

We finally observe that the number of duplicated genes associated with the six climatic variables in the southern coast and especially southern highland populations is much higher than in the central populations (Fig. S13B). These duplicated genes are involved in response to environments, including light, drought, cold, UV, and carbohydrate (photosynthesis), such as likely homologs of the genes FT, FD, and ABI4 and genes involved in the formation of photosystem subunits (Dataset S5). The number of candidate genes found as deletions is similar in different populations (Fig. S13C). Most lost genes are related to plant growth and development. The GEA analyses confirm the adaptive relevance of gene CN expansion and contraction: (i) the CN-differentiated genes in the central group appear mainly as contraction genes (deletions) while these appear ~~at expansion~~ (duplications) in the southern highland populations; (ii) the adaptive gene CN changes reflect the colonisation of novel habitats at the southern edge of the species distribution; and (iii) the expansion and contraction of gene CN in different populations are the consequences of the response to ~~the~~ different ~~habitat~~ environments.

Discussion

A set of key genomic CNVs are found to be highly correlated with the species colonisation process and environmental variables and thus are likely implicated in the adaptive differentiation between populations, most likely because of their major impact on gene expression (Fuentes, et al. 2019; Rinker, et al. 2019; Alonge, et al. 2020; Hämälä, et al. 2021; Li, et al. 2023). This confirms that CNVs have ubiquitous roles in adaptive processes in ecology and evolution (Żmieńko, et al. 2014; Castagnone-Sereno, et al. 2019; Lauer and Gresham 2019; Mérot, et al. 2020). To better understand the genetic basis behind the fitness effect of CNVs in natural populations, we analyse whole-genome data for 35 *S. chilense* individuals from seven populations, allowing us to identify a genome-wide CNVs dataset. Our CNV calling pipeline resolves hundreds of thousands of CNVs. The number of CNV for each population of *S. chilense* is similar to numbers found in the previous tomato clade panSV-genome study that includes a single sample of *S. chilense* (Li, et al. 2023). CNVs are abundant across all chromosomes and frequently reside within or in close proximity to genes (Fig. 1). Widespread CNVs in the genome exhibit similar performance as SNPs for the inference of population structure and differentiation between populations (Fig. 2; Fig. S3) (Cheeseman, et al. 2016; Fuentes, et al. 2019). Based on the ~~past~~ demographic model we developed previously (Wei, et al. 2023b) as a neutral ~~evolution baseline~~ and the dynamics of CN ~~profile~~ in two southward colonisation events, our results support that most CNV is likely shaped by neutral processes (Silva-Arias, et al. 2023). However, ~~this~~ genome-wide ~~assessment~~ allows us to identify CNV likely related to the adaptive divergence in recently colonised regions in response to abiotic stress.

We identified gene signatures putatively exhibiting footprints of adaptive divergence using CN profiles, and ~~these~~ candidate genes are associated with adaptation to local environments, consistent with genome scans based on SNPs (Wei, et al. 2023b). CN differences of these genes across different populations reflect the neutral and divergent selection process between populations (Fig. 3A; Fig. S6), demonstrating that CNVs must be considered to fully understand how selection shapes genomic structural diversity and local adaptation. Overall, the evolutionary processes generating CNV diversity and divergence follow the historical demography of *S. chilense*, namely two southward independent colonisation events. ~~Genes CN appear~~ expanded in the southernmost SC_LA4107 and southern highland populations, which underwent recent colonisation events and exhibit lower population sizes (Stam, et al. 2019b; Wei, et al. 2023b), while gene CN reveals a trend of contraction in the central and SC_LA2932 populations (close to the species centre of origin). Therefore, we conclude that CN expansion and contraction are not only due to neutral evolutionary processes (past demographic events) but likely reflect and underpin selective events during the two southward colonisation events. Conversely, early established populations exhibit adaptive loss of gene and function processes (Albalat and Cañestro 2016; Helsen, et al. 2020), especially in genes involved in plant growth and development in central populations, or the loss of genes involved in photosynthesis in central and SC_LA2932 populations (Fig. 4D). Changes at photosynthetic gene CN underpin population differentiation between SC_LA2932 (gene loss) and SC_LA4107 (gene gain) ~~representing~~ two different habitats ~~of~~ the southern coast. CN differentiated genes were also enriched in response to multiple abiotic stresses, such as red/far red light, cold, UV, or drought. These response processes can directly affect plant reproduction and growth and regulate flowering regulatory processes (Fig. S7). ~~This further emphasises~~ results based on ~~our SNP study~~ showing that the reproductive cycle, primarily ~~regulating~~ flowering time, may play a key role in adaptation to abiotic stress in *S. chilense* (Wei, et al. 2023b).

~~Flowering regulation involved~~ in response to light (photoperiod) and cold (vernalisation) are key adaptive pathways for *S. chilense* populations to colonise southern habitats ~~based on~~ genome-wide SNPs (Wei, et al. 2023b). Here, we obtain further candidate genes ~~with~~ differentiated gene CN profiles involved in flowering regulatory pathways for response to changes in photoperiod and cold. These genes (putative FT, FD, FLD homologs, etc.) are duplicated in the southern highland populations (Fig. S8). ~~In addition~~, solar radiation is also a challenging condition for plants at high altitudes. Many CN-differentiated genes are indeed enriched ~~in~~ response to UV light (Fig. 3B; Dataset S4), including homologs of genes involved in anthocyanin accumulation. Indeed, the anthocyanin pathway is switched off in cultivated tomato by mutations of splice sites in regulatory genes and anthocyanin-producing tomato varieties have been created by genetic engineering to obtain anthocyanin-rich purple fruits (Gonzali, et al. 2009; Sun, et al. 2020; Gonzali and Perata 2021). These CN-differentiated genes related to the anthocyanin pathway still provide a potential source of natural variation for breeding tomato with anthocyanin. More generally, the large number of gene losses possibly in response to environmental stresses, may indicate that the reduction of the genome size is a powerful evolutionary driver of adaptation (Albalat and Cañestro 2016; Helsen, et al. 2020; Monroe, et al. 2021). Further functional validation will help understand the molecular mechanisms through which CNV drives adaptive evolution in natural populations.

Genome-Environment Association (GEA) analysis ultimately links the dynamics of gene CN to six climatic variables and reveals the population structure of CNVs in connection to four different habitat environments (Fig. 5A and B). These ~~overrepresented~~ climate variables are almost uniformly associated with SNPs in an RDA analysis (Wei, et al. 2023b). ~~These~~ potential CNV-environmental interactions have been observed in *Arabidopsis thaliana* (DeBolt 2010; Zmienko, et al. 2020), *Solanum lycopersicum* (Alonge, et al. 2020), *Theobroma cacao* (Hämälä, et al. 2021), *Oryza sativa* (Fuentes, et al. 2019; Qin, et al. 2021). We are further explicit that CNVs play an essential role in southward colonisation in *S. chilense*. CNVs, especially duplications in southern highland populations exposed to typical high-altitude stresses, show adaptations in genes with functions related to cold, change of photoperiod and solar radiation. The CN profiles of differentiated genes in southern coast populations mainly correlate with drought stress, such as root development, cell homeostasis, or cell wall maintenance. Interestingly, gene CN differentiation related to photosynthesis provides evidence for the genetic underpinning of the adaptive differentiation between SC_LA2932 and SC_LA4107, representing two different coastal habitats (Fig. 1A and 4C). These differentiated genes reveal exactly opposite CN evolutionary trends ~~between them~~ (Dataset S6). Indeed, we see different habitats as SC_LA2932 grows in dry ravines (quebrada) in Lomas formations, whereas SC_LA4107 grows in extremely fine alluvial soil (with even some running water). Moreover, these chloroplast genes are detected in the nuclear genome indicating a widespread event of organellar gene transfer to the nuclear genome in tomato (Pesaresi, et al. 2014; Lichtenstein, et al. 2016; Kim and Lee 2018). These adaptive signatures were not found in previous studies based on genome scans of SNPs (Wei, et al. 2023b). The three central populations display mainly a trend towards gene loss and low correlation with climatic variables (Fig. 5A and B). This is consistent with the fact that GEA analyses based on current climatic data have limited statistical power to detect old adaptive selection signals, whether based on SNPs or CNVs, due to the occurrence of multiple historical confounding events such as genetic drift, migration, and recombination (De Mita, et al. 2013; Manel, et al. 2016). The two central populations (C_LA2931 and C_LA3111) found at high altitudes exhibit few adaptive duplications

~~signatures,~~ but some as possible responses to cold and solar radiation, similar to those observed for the SH populations (Stam, et al. 2019b; Wei, et al. 2023b).

We finally advise that our study likely underestimates the amount and importance of CNVs in *S. chilense* as we do not possess long-read data for all accessions. First, our pipeline to recover CNVs based on short-read data is tested by simulations and is likely conservative, meaning that we probably miss some CNVs. Second, there may be ~~some potential~~ bias in finding footprints of selection when using ~~populations multiplied at~~ the TGRC (UC Davis, USA) as we discussed previously (Wei, et al. 2023b). Though we point out that the use of several selective sweep detection methods conservatively underestimate the amount of (positive) selection signals. The availability of a new reference genome (Silva-Arias, et al. 2023) and ~~few~~ accessions sequenced with long-read (Li, et al. 2023) do open the path to sequence wild accessions with long-read sequencing and a complete assessment of the importance of CNVs at abiotic stress genes in *S. chilense*. Furthermore, the new simulation method to study and infer the neutral and selective processes driving gene duplication and deletion (Otto, et al. 2022; Otto and Wiehe 2023) can be used in the future to refine our conclusions regarding the neutral rates of gene duplication/deletion during the species southward expansion. Despite being conservative regarding the importance of positive selection shaping the CNV diversity in *S. chilense*, our results reinforce the observation that CNV is an important contributor to adaptation across different ecological habitats (Żmieńko, et al. 2014; Rinker, et al. 2019; Hämälä, et al. 2021; Monroe, et al. 2021). The strong selective pressure imposed by the range expansion of *S. chilense* and the need to adapt to novel stressful habitats has shaped the genetic diversity at SNPs and CNVs. In agreement with previous studies, we ~~confirm~~ that natural selection acting ~~through~~ CNVs can reshape the ~~population genome to underpin~~ adaptation (Iskow, et al. 2012; Żmieńko, et al. 2014; Rinker, et al. 2019; Hämälä, et al. 2021).

Materials and Methods

For complete materials and methods, see SI Appendix, Supplementary Information Text.

Sequence Read Processing

We used 35 whole-genome paired-end Illumina data from seven populations of *S. chilense* (five diploid plants for each population) representing four different geographic groups (Fig. 1A). The data are available on European Nucleotide Achieve (ENA) BioProject PRJEB47577. We performed the same pipeline of read processing procedure as in a previous study (Wei, et al. 2023b) including quality trimming, mapping and SNP calling based on the new reference genome of *S. chilense* (Silva-Arias, et al. 2023).

Identification and genotyping of CNVs

We used LUMPY (Layer, et al. 2014), Manta (Chen, et al. 2016), Wham (Kronenberg, et al. 2015) and DELLY (Rausch, et al. 2012) to identify CNVs in the 35 samples. The CNV sets from LUMPY, DELLY, Manta and Wham were merged using SURVIVOR v1.0.7 (Jeffares, et al. 2017). The merged CNV set was inputted to SVTyper v0.7.0 to call genotypes using a Bayesian algorithm (Chiang, et al. 2015).

To assess the sensitivity and accuracy of our pipeline for CNV calling, we simulated 1,000 duplication and 1,000 deletion regions with sizes ranging from 50bp to 1Mb using CNV-Sim v0.9.2 employing the functionality of ART (Huang, et al. 2012), and these simulated reads (150 bp) were used as the input for the same CNV analysis pipelines to identify CNVs.

Population structure analysis

~~The~~ principal component analysis (PCA) was performed using GCTA v1.91.4 (Yang, et al. 2011). ~~The inference of population structure was performed~~ using the program ADMIXTURE v1.3.0 (Alexander, et al. 2009) based on six scenarios (K values from 2 to 7) using SNPs or CNVs.

Quantification of gene copy number

We employed two strategies to quantify gene copy number (CN). First, we used Control-FREEC v11.6 to estimate CN ~~by~~ 10 kb windows with 1 kb step size across the entire genome (Boeva, et al. 2012). We then obtained gene CN from the Control-FREEC outputs, and gene coordinates in the genome. We also used Mosdepth v0.3.2 (Pedersen and Quinlan 2018) to calculate read depth ~~by 1,000 bp~~ sliding windows, and gene read depth was calculated from gene coordinates. We then used median read-depth values of all windows and genes as a normalising factor to obtain the final window and gene CN estimate, respectively, with the formula: CN = (read-depth / median value) × 2.

Identification of candidate genes associated with population differentiation

The V_{ST} and F_{ST} statistics are applied to quantify population differentiation and are computed over 1,000 bp windows. We calculated two V_{ST} estimates based on two different CN quantitative strategies: Control-FREEC and Read Depth. We performed permutation tests (1,000 times) for each gene to extract candidate genes. We then selected candidate genes with V_{ST} values above the 95th and 99th percentile of the permuted V_{ST} distribution for each V_{ST} estimate.

Gene ontology (GO) analysis

We first performed a blast of our genes to the *A. thaliana* dataset TAIR10 (e-value cutoff was 10^{-6}) (Camacho, et al. 2009). We used the R package clusterProfiler to perform GO enrichment analysis using the *A. thaliana* annotation database as the background (Yu, et al. 2012). The Benjamini-Hochberg method was used to calibrate ~~initial *P*~~ values, and calibrated *P* values smaller than 0.05 were used as the cutoff for a significant level to obtain final GO terms.

Expansion and contraction of gene copy number

We computed the expansion and contraction of the 3,359 genes with high CN differentiation between populations. We first constructed a population-based phylogenetic tree using SNPs and TreeMix v1.13 (Pickrell and Pritchard 2012). ~~The~~ ultrametric tree (Fig. 4A) was generated based on *force.ultrametric* function of phytools R package (Revell 2012). We then performed analyses of the expansion and contraction of gene CN using CAFE v4.2.1 Campo (Han(Han, et al. 2013). The branch-specific p-values are obtained by the Viterbi method with the randomly generated likelihood distribution. We set a p-value smaller than 0.05 to detect gene CN with a significant rate of evolution (expansion or contraction) in different groups/populations.

Association analysis between gene copy number and climatic conditions

The environmental data include 37 climatic layers (Dataset S7) obtained from two public databases, WorldClim2 (Fick and Hijmans 2017) and ENVIREM (Title and Bemmels 2018). To evaluate the relative contribution of the abiotic environment to explaining patterns of genetic variation, we first used the Redundancy Analysis (RDA) (Capblancq and Forester 2021) to associate CNs of the 3,539 differentiated genes with climatic variables. RDA was performed using the *rda* function from the vegan package as implemented in R (Forester, et al. 2018). LFMM (latent factor mixed models) is a univariate test (Frichot, et al. 2013; Caye, et al. 2019), which means it builds a model for each gene or SNP and each predictor variable. We then implemented LFMM2 to perform an association test between gene CN and six representative climate variables obtained in RDA, respectively. Benjamini-Hochberg's method was used to calibrate the *P* values with 0.01 as the significant threshold.

Supplementary material

Supplementary data are available online at Molecular Biology and Evolution.

Data Availability

Raw sequence data are available at the European Nucleotide Achieve (ENA) BioProject PRJEB47577. The resource of copy number variation identified in this study and custom scripts for conducting the analyses are available at our Gitlab at the following link: https://gitlab.lrz.de/population_genetics/s_chilense_cnv.

Acknowledgements

KW was funded by the Chinese Scholarship Council. GAS-A was funded by the Technical University of Munich. AT acknowledges funding from DFG (Deutscheforschungsgemeinschaft) Grant Number: 317616126 (TE809/7-1). We thank the Tomato Genetics Resource Center (TGRC) of the University of California, Davis for generously providing us with the seeds of the accession included in this study.

Competing interests

The authors have no conflicts of interest to declare.

Author contributions

KW, GAS-A and AT planned and designed the study. RS and AT obtained the sequencing data. KW performed data analyses. KW wrote the first draft of the manuscript, and RS, GAS-A, and AT edited and improved the manuscript. All authors approved the final manuscript.

References

Albalat R, Cañestro C. 2016. Evolution by gene loss. Nature Reviews Genetics 17:379-391.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome research 19:1655-1664.

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D. 2020. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 182:145-161. e123.

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, Cao J, Chae E, Dezwaan TM, Ding W, et al. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 166:481-491.

Antinucci M, Comas D, Calafell F. 2023. Population history modulates the fitness effects of Copy Number Variation in the Roma. Human Genetics:1-17.

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

Bao S, Hua C, Shen L, Yu H. 2020. New insights into gibberellin signaling in regulating flowering in Arabidopsis. Journal of integrative plant biology 62:118-131.

Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J. 2016. Recent demography drives changes in linked selection across the maize genome. Nature Plants 2:16084.

Blanchard-Gros R, Bigot S, Martinez J-P, Lutts S, Guerriero G, Quinet M. 2021. Comparison of Drought and Heat Resistance Strategies among Six Populations of Solanum chilense and Two Cultivars of Solanum lycopersicum. Plants 10:1720.

Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 28:423-425.

Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sørensen I, Lichtenstein G. 2014. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nature genetics 46:1034-1038.

Böndel KB, Lainer H, Nosenko T, Mboup M, Tellier A, Stephan W. 2015. North–south colonization associated with local adaptation of the wild tomato species Solanum chilense. Molecular biology and evolution 32:2932-2943.

Böndel KB, Nosenko T, Stephan W. 2018. Signatures of natural selection in abiotic stress-responsive genes of Solanum chilense. Royal Society open science 5:171198-171198.

Brumlop S, Weedon O, Link W, Finckh M. 2019. Effective population size (Ne) of organically and conventionally grown composite cross winter wheat populations depending on generation. European Journal of Agronomy 109:125922.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC bioinformatics 10:1-9.

Capblancq T, Forester BR. 2021. Redundancy analysis: A Swiss Army Knife for landscape genomics. Methods in Ecology and Evolution 12:2298-2309.

Castagnone-Sereno P, Mulet K, Danchin EG, Koutsovoulos GD, Karaulic M, Da Rocha M, Bailly-Bechet M, Pratx L, Perfus-Barbeoch L, Abad P. 2019. Gene copy number variations as signatures of adaptive evolution in the parthenogenetic, plant-parasitic nematode Meloidogyne incognita. Molecular Ecology 28:2559-2572.

Caye K, Jumentier B, Lepeule J, François O. 2019. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. Molecular biology and evolution 36:852-860.

Cheeseman IH, Miller B, Tan JC, Tan A, Nair S, Nkhoma SC, De Donato M, Rodulfo H, Dondorp A, Branch OH. 2016. Population structure shapes copy number variation in malaria parasites. Molecular biology and evolution 33:603-620.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32:1220-1222.

Cheng JZ, Zhou YP, Lv TX, Xie CP, Tian CE. 2017. Research progress on the autonomous flowering time pathway in Arabidopsis. Physiol Mol Biol Plants 23:477-485.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nature methods 12:966-968.

Coutelier M, Holtgrewe M, Jäger M, Flöttman R, Mensah MA, Spielmann M, Krawitz P, Horn D, Beule D, Mundlos S. 2022. Combining callers improves the detection of copy number variants from whole-genome sequencing. European Journal of Human Genetics 30:178-186.

De Mita S, Thuillet A-C, Gay L, Ahmadi N, Manel S, Ronfort J, Vigouroux Y. 2013. Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Molecular Ecology 22:1383-1399.

DeBolt S. 2010. Copy Number Variation Shapes Genome Diversity in Arabidopsis Over Immediate Family Generational Scales. Genome biology and evolution 2:441-453.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nature Reviews Genetics 7:85-97.

Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1‐km spatial resolution climate surfaces for global land areas. International journal of climatology 37:4302-4315.

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

Forester BR, Lasky JR, Wagner HH, Urban DL. 2018. Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. Molecular Ecology 27:2215-2233.

Frichot E, Schoville SD, Bouchard G, François O. 2013. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Molecular biology and evolution 30:1687-1699.

Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, Wing RA, McNally KL, Tatarinova T, Grigoriev A. 2019. Structural variants in 3000 rice genomes. Genome research 29:870-880.

Gaudinier A, Blackman BK. 2020. Evolutionary processes from the perspective of flowering time diversity. New Phytologist 225:1883-1898.

Gonzali S, Mazzucato A, Perata P. 2009. Purple as a tomato: towards high anthocyanin tomatoes. Trends in plant science 14:237-241.

Gonzali S, Perata P. 2021. Fruit Colour and Novel Mechanisms of Genetic Regulation of Pigment Production in Tomato Fruits. Horticulturae 7:259.

Guo M, Yang F, Liu C, Zou J, Qi Z, Fotopoulos V, Lu G, Yu J, Zhou J. 2022. A single-nucleotide polymorphism in WRKY33 promoter is associated with the cold sensitivity in cultivated tomato. New Phytologist 236:989-1005.

Hämälä T, Wafula EK, Guiltinan MJ, Ralph PE, dePamphilis CW, Tiffin P. 2021. Genomic structural variants constrain and facilitate adaptation in natural populations of Theobroma cacao, the chocolate tree. Proceedings of the National Academy of Sciences 118:e2102914118.

Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Molecular biology and evolution 30:1987-1997.

Hecht Vr, Foucher F, Ferrándiz C, Macknight R, Navarro C, Morin J, Vardy ME, Ellis N, Beltrán JPo, Rameau C, et al. 2005. Conservation of Arabidopsis Flowering Genes in Model Legumes Plant Physiology 137:1420-1434.

Helsen J, Voordeckers K, Vanderwaeren L, Santermans T, Tsontaki M, Verstrepen KJ, Jelier R. 2020. Gene Loss Predictably Drives Evolutionary Adaptation. Molecular biology and evolution 37:2989-3002.

Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28:593-594.

Iskow RC, Gokcumen O, Lee C. 2012. Exploring the role of copy number variants in human adaptation. Trends in Genetics 28:245-257.

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nature communications 8:14061.

Kim HT, Lee JM. 2018. Organellar genome analysis reveals endosymbiotic gene transfers in tomato. PLoS One 13:e0202279.

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome biology 20:1-18.

Kronenberg ZN, Osborne EJ, Cone KR, Kennedy BJ, Domyan ET, Shapiro MD, Elde NC, Yandell M. 2015. Wham: identifying structural variants of biological consequence. PLoS computational biology 11:e1004572.

Lauer S, Gresham D. 2019. An evolving view of copy number variants. Current genetics 65:1287-1295.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome biology 15:1-19.

Li N, He Q, Wang J, Wang B, Zhao J, Huang S, Yang T, Tang Y, Yang S, Aisimutuola P, et al. 2023. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. Nature genetics.

Lichtenstein G, Conte M, Asis R, Carrari F. 2016. Chloroplast and mitochondrial genomes of tomato. The Tomato Genome:111-137.

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

Liu Z, Hou S, Rodrigues O, Wang P, Luo D, Munemasa S, Lei J, Liu J, Ortiz-Morea FA, Wang X, et al. 2022. Phytocytokine signalling reopens stomata in plant immunity and water loss. Nature 605:332-339.

Luo X, Xu J, Zheng C, Yang Y, Wang L, Zhang R, Ren X, Wei S, Aziz U, Du J, et al. 2022. Abscisic acid inhibits primary root growth by impairing ABI4-mediated cell cycle and auxin biosynthesis. Plant Physiology 191:265-279.

Lupski JR. 2007. Genomic rearrangements and sporadic disease. Nature genetics 39:S43-S47.

Lynch M, Walsh B. 2007. The origins of genome architecture: Sinauer Associates Sunderland, MA.

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. Genome biology 20:246.

Makita Y, Suzuki S, Fushimi K, Shimada S, Suehisa A, Hirata M, Kuriyama T, Kurihara Y, Hamasaki H, Okubo-Kurihara E. 2021. Identification of a dual orange/far-red and blue light photoreceptor from an oceanic green picoplankton. Nature communications 12:3593.

Manel S, Perrier C, Pratlong M, Abi‐Rached L, Paganini J, Pontarotti P, Aurelle D. 2016. Genomic resources and their influence on the detection of the signal of positive selection in genome scans. Molecular Ecology 25:170-184.

Marszalek-Zenczak M, Satyr A, Wojciechowski P, Zenczak M, Sobieszczanska P, Brzezinski K, Iefimenko T, Figlerowicz M, Zmienko A. 2023. Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members. Front Plant Sci 14:1104303.

Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A roadmap for understanding the evolutionary significance of structural genomic variation. Trends in Ecology & Evolution 35:561-572.

Monroe JG, McKay JK, Weigel D, Flood PJ. 2021. The population genomics of adaptive loss of function. Heredity 126:383-395.

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

Nosenko T, Böndel KB, Kumpfmüller G, Stephan W. 2016. Adaptation to low temperatures in the wild tomato species Solanum chilense. Molecular Ecology 25:2853-2869.

Otto M, Wiehe T. 2023. The structured coalescent in the context of gene copy number variation.

Theoretical Population Biology 154:67-78.

Otto M, Zheng Y, Wiehe T. 2022. Recombination, selection, and the evolution of tandem gene arrays. Genetics 221:iyac052.

Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34:867-868.

Pérez-Ruiz Rigoberto V, García-Ponce B, Marsch-Martínez N, Ugartechea-Chirino Y, Villajuana-Bonequi M, de Folter S, Azpeitia E, Dávila-Velderrain J, Cruz-Sánchez D, Garay-Arroyo A, et al. 2015. XAANTAL2 (AGL14) Is an Important Component of the Complex Gene Regulatory Network that Underlies Arabidopsis Shoot Apical Meristem Transitions. Molecular Plant 8:796-813.

Pesaresi P, Mizzotti C, Colombo M, Masiero S. 2014. Genetic regulation and structural changes during tomato fruit development and ripening. Frontiers in plant science 5:124.

Pickrell J, Pritchard J. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. Nature Precedings:1-1.

Putterill J, Varkonyi-Gasic E. 2016. FT and florigen long-distance flowering control in plants. Current opinion in plant biology 33:77-82.

Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, He Q, Ou S, Zhang H, Li X, et al. 2021. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell 184:3542-3558.e3516.

Raduski AR, Igić B. 2021. Biosystematic studies on the status of Solanum chilense. American Journal of Botany 108:520-537.

Ranjan A, Ichihashi Y, Sinha NR. 2012. The tomato genome: implications for plant breeding, genomics and evolution. Genome biology 13:1-8.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333-i339.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3:217-223.

Rinker DC, Specian NK, Zhao S, Gibbons JG. 2019. Polar bear evolution is marked by rapid changes in gene copy number in response to dietary shift. Proceedings of the National Academy of Sciences 116:13446-13451.

Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635-641.

Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, O'Hara R, Casalunovo T, Conlin LK, D'arcy M. 2009. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome research 19:1682-1690.

Silva-Arias GA, Gagnon E, Hembrom S, Fastner A, Khan MR, Stam R, Tellier A. 2023. Contrasting patterns of presence-absence variation of NLRS within <em>S. chilense</em> are mainly shaped by past demographic history. bioRxiv:2023.2010.2013.562278.

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H. 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS genetics 5:e1000734.

Srikanth A, Schmid M. 2011. Regulation of flowering time: all roads lead to Rome. Cellular and Molecular Life Sciences 68:2013-2037.

Stam R, Nosenko T, Hörger AC, Stephan W, Seidel M, Kuhn JM, Haberer G, Tellier A. 2019a. 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 9:3933-3941.

Stam R, Silva-Arias GA, Tellier A. 2019b. Subsets of NLR genes show differential signatures of adaptation during colonization of new habitats. New Phytologist 224:367-379.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M. 2015. Global diversity, population stratification, and selection of human copy-number variation. Science 349:aab3761.

Sun C, Deng L, Du M, Zhao J, Chen Q, Huang T, Jiang H, Li C-B, Li C. 2020. A transcriptional network promotes anthocyanin biosynthesis in tomato flesh. Molecular Plant 13:42-58.

Title PO, Bemmels JB. 2018. ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. Ecography 41:291-307.

Wei K, Sharifova S, Zhao X, Sinha N, Nakayama H, Tellier A, Silva-Arias GA. 2023a. Evolution of two gene networks underlying adaptation to drought stress in the wild tomato Solanum chilense. bioRxiv:2023.2001. 2018.524537.

Wei K, Silva-Arias GA, Tellier A. 2023b. Selective sweeps linked to the colonization of novel habitats and climatic changes in a wild tomato species. New Phytologist 237:1908-1921.

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

Xiao S, Jiang L, Wang C, Ow DW. 2021. Arabidopsis OXS3 family proteins repress ABA signaling through interactions with AFP1 in the regulation of ABI4 expression. Journal of Experimental Botany 72:5721-5734.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics 88:76-82.

Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology 16:284-287.

Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, Cantu D, Gaut BS. 2019. The population genetics of structural variants in grapevine domestication. Nature Plants 5:965-979.

Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, Wu Y, Cheng L, Fang Y, Wu K. 2022. Graph pangenome captures missing heritability and empowers tomato breeding. Nature 606:527-534.

Zmienko A, Marszalek-Zenczak M, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozlowski P, Karlowski WM, Figlerowicz M. 2020. AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis

Genome[OPEN]. The Plant Cell 32:1797-1819.

Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M. 2014. Copy number polymorphism in plant genomes. Theoretical and applied genetics 127:1-18.

Table1. The summary of gene expansion and contraction in different groups/populations based on ~~the~~ phylogenetic ~~and ultrametric~~ tree.

[a]Groups / Populations	Number of CN expanded genes	Number of CN contracted genes	Number of CN gained	Number of CN lost	[b]Rate of average expansion / contraction	[c]Number of rapidly evolving genes
Inland	40	26	167	49	1.788	15 (+13/-2)
C	163	695	355	1,013	-0.767	20 (+5/-15)
SH	527	525	1,143	705	0.416	37 (+32/-5)
SC	48	359	106	439	-0.818	9 (+2/-7)
C_LA1963	137	416	445	728	-0.512	10 (+3/-7)
C_LA2931	212	458	815	878	-0.094	15 (+3/-12)
C_LA3111	364	266	1,068	444	1.037	23 (+6/-15)
SH_LA4117A	813	342	2,574	653	1.663	52 (+38/-14)
SH_LA4330	446	328	1,766	702	1.375	31 (+22/-9)
SC_LA2932	268	846	427	1,514	-0.935	29 (+7/-22)
SC_LA4107	595	640	1,758	1,098	0.534	35 (+25/-10)

The table shows ~~that~~ the expansion and contraction of CN-differentiated genes in different groups / populations based on an ultrametric tree (Fig. 4A). C: central; SH: southern highland; SC: southern coast.

[a]Groups and populations denote the branches in the phylogenetic ~~and ultrametric~~ tree (Fig. 4A).

[b]Rate of average expansion / contraction = (Number of CN gained - Number of CN lost) / (Number of CN expanded genes + Number of CN contracted genes). Positive values indicate CN expansion and negative values indicate CN contraction.

[c]The rapidly evolving genes indicate significant higher CN expansion or contraction (Viterbi P < 0.05) across the different groups/populations. Values outside parentheses represent the total number of the rapidly evolving genes. Positive values in parentheses denote the number of significantly expanded genes and negative values denote the number of significantly contracted genes (see also Dataset S6).

Figures

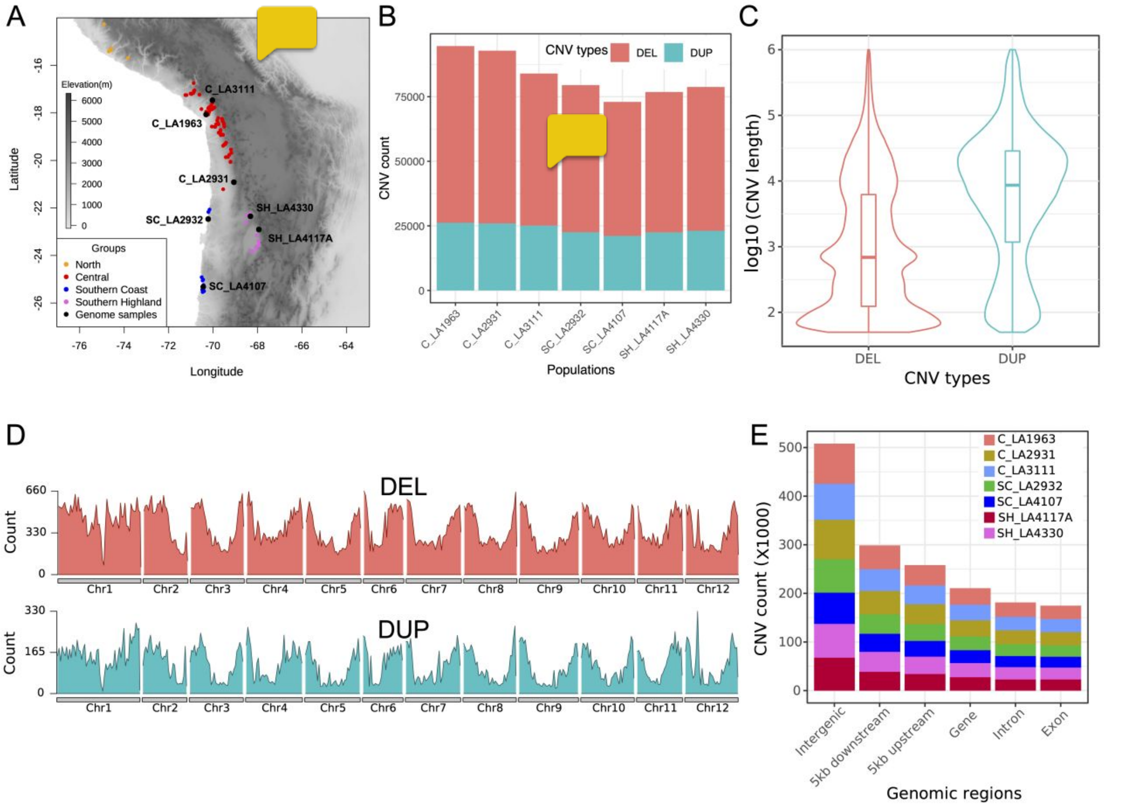

Fig. 1. The summary of the revealed CNVs in the genome of *S. chilense*. (A) Map with the distribution of all S. chilense populations at the Tomato Genetics Resource Center (TGRC), the seven S. chilense populations in this study (black circles), and the four population groups (circles with other colours). C: central; SH: southern highland; SC: southern coast. (B) The number of CNVs merged for five accessions of each population. DEL: deletion; DUP: duplication. (C) The distribution of CNVs size. (D) The number of located CNVs at different genome regions, counted in windows of 1 Mb. (E) The number of CNVs overlapping various genomic features for each population.

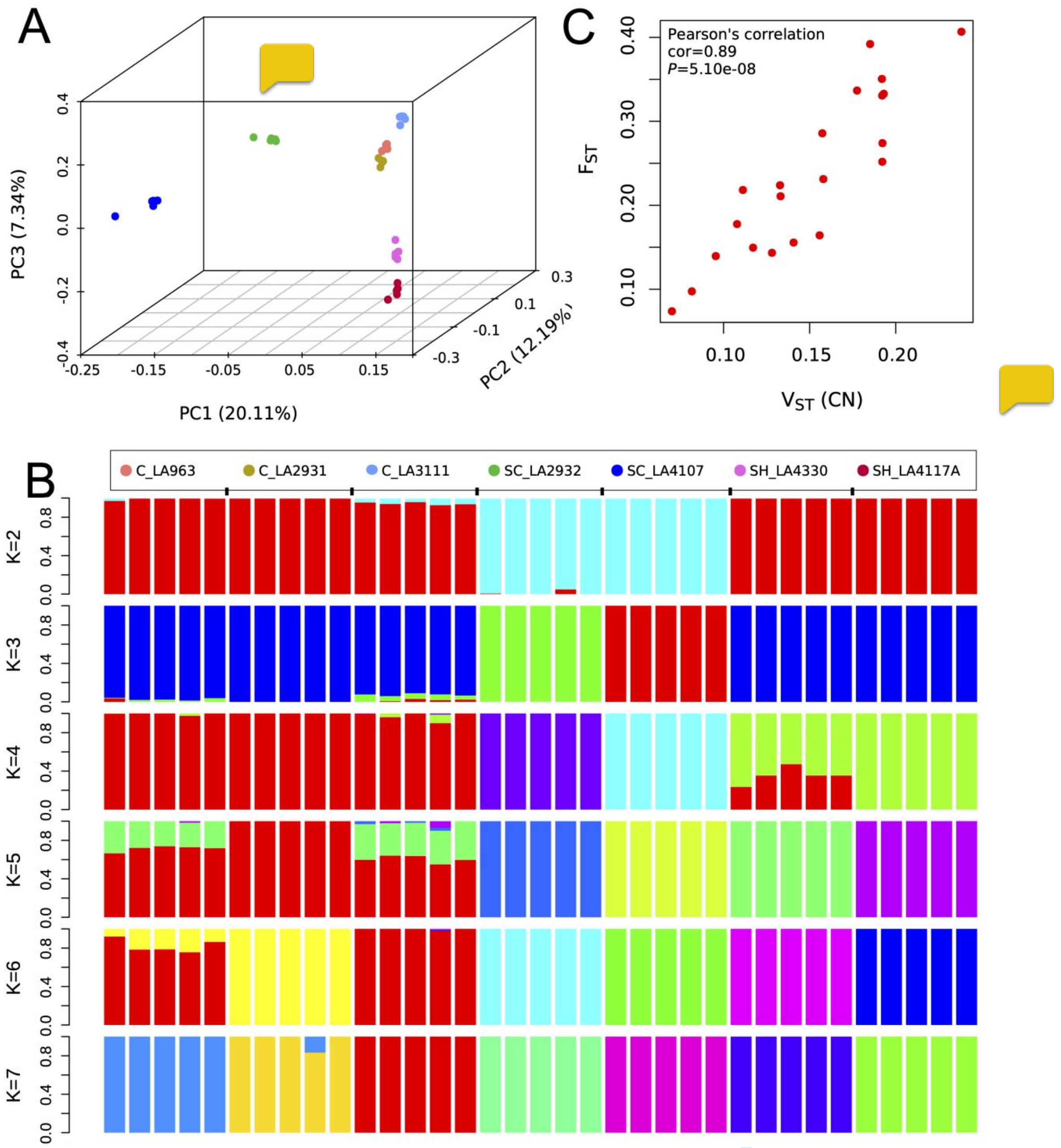

Fig. 2. Structure analysis based on genotyped CNVs. (A) Principal component analysis (PCA) based using genotyped CNVs from 35 *S. chilense* accessions. (B) Structure analysis based on genotyped CNVs and assuming *K* = 2 - 7 subgroups (optimal *K* value is 4; Fig. S3B). C: central; SH: southern highland; SC: southern coast. (C) The correlation between F_{ST} and V_{ST} indicates that CNVs support the known population differentiation.

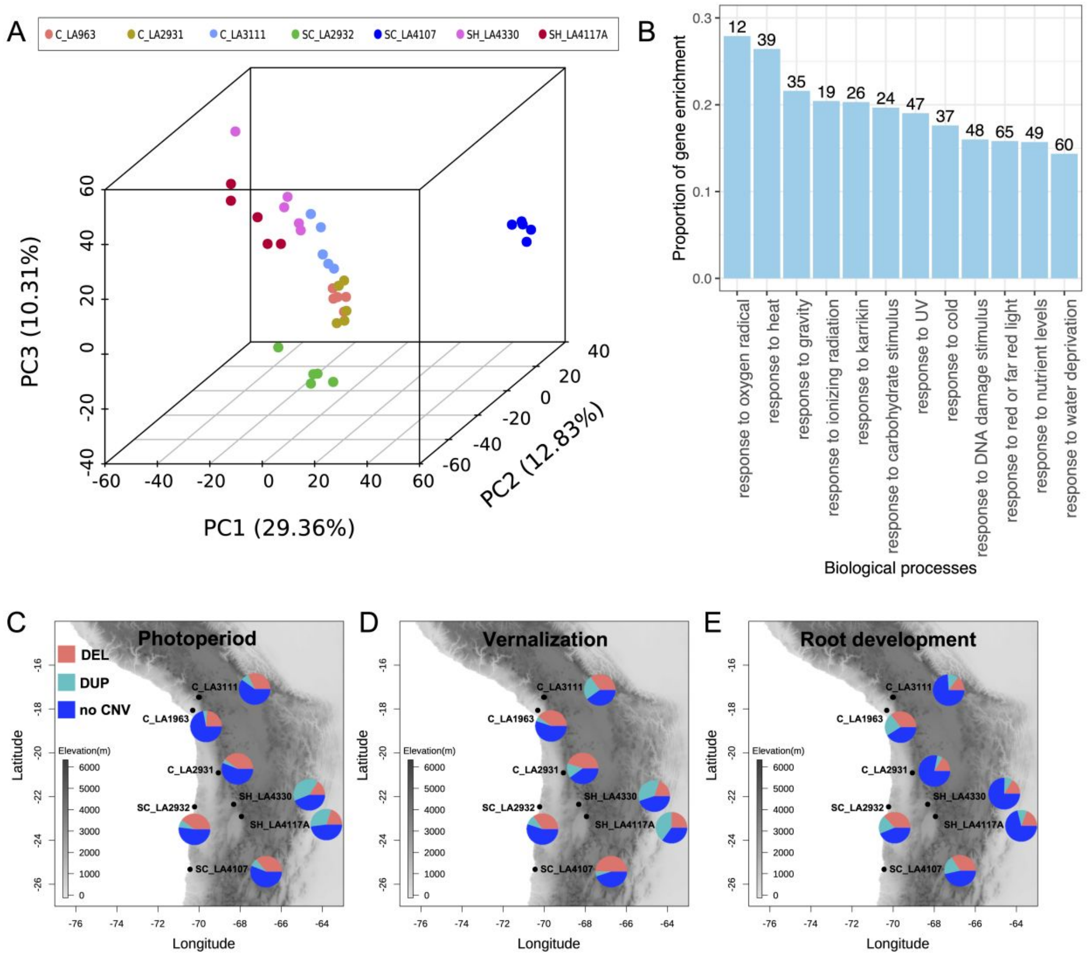

Fig. 3. Genes with differentiated CN profiles among seven populations are linked to response to multiple environmental stimuli. (A) PCA based on the copy number (CN) of 3,539 differentiated genes. C: central; SH: southern highland; SC: southern coast. (B) The proportions of CN-differentiated genes enriched in response to external stimulus/stresses (significantly enriched $P < 0.05$). The ratio of gene enrichment is equal to the number of genes enriched in one GO category divided by the number of background genes in this category. The number on each bar represents the number of genes enriched in that GO category. The CN-differentiated genes involved in photoperiod pathway to regulate flowering time (C), vernalisation pathways to regulate flowering time (D), and root developmental process (E). The pie charts denote the proportions of CN-differentiated genes with deletion (DEL), duplication (DUP) or absence of CNV over all 3,539 genes (see also Table S8).

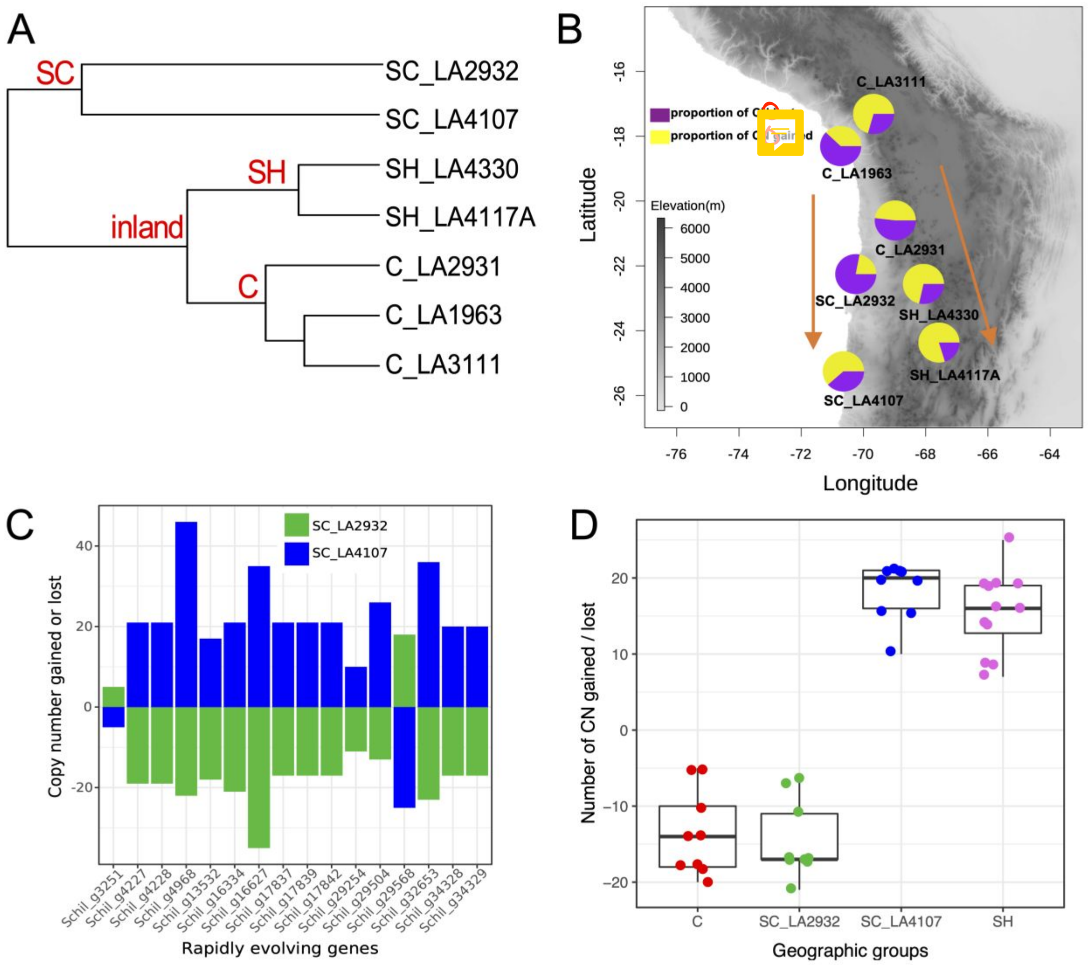

Fig. 4. The expansion and contraction of CN-differentiated genes in different populations using reference the genome of *S. chilense*. (A) The phylogenetic and ultrametric tree is used in gene expansion and contraction analysis (see Table 1). C: central; SH: southern highland; SC: southern coast. (B) The map and pie charts show the dynamics of CN lost and gained in the processes of two southward colonization events, first to the southern coast and second to the southern highland (orange arrows). (C) The number of CN gained (positive values) or lost (negative values) for 16 rapidly evolving genes in two southern coast populations. (D) The number of CN-gained or -lost for rapidly evolving genes related to photosynthesis in different subgroups representing four different habitats.

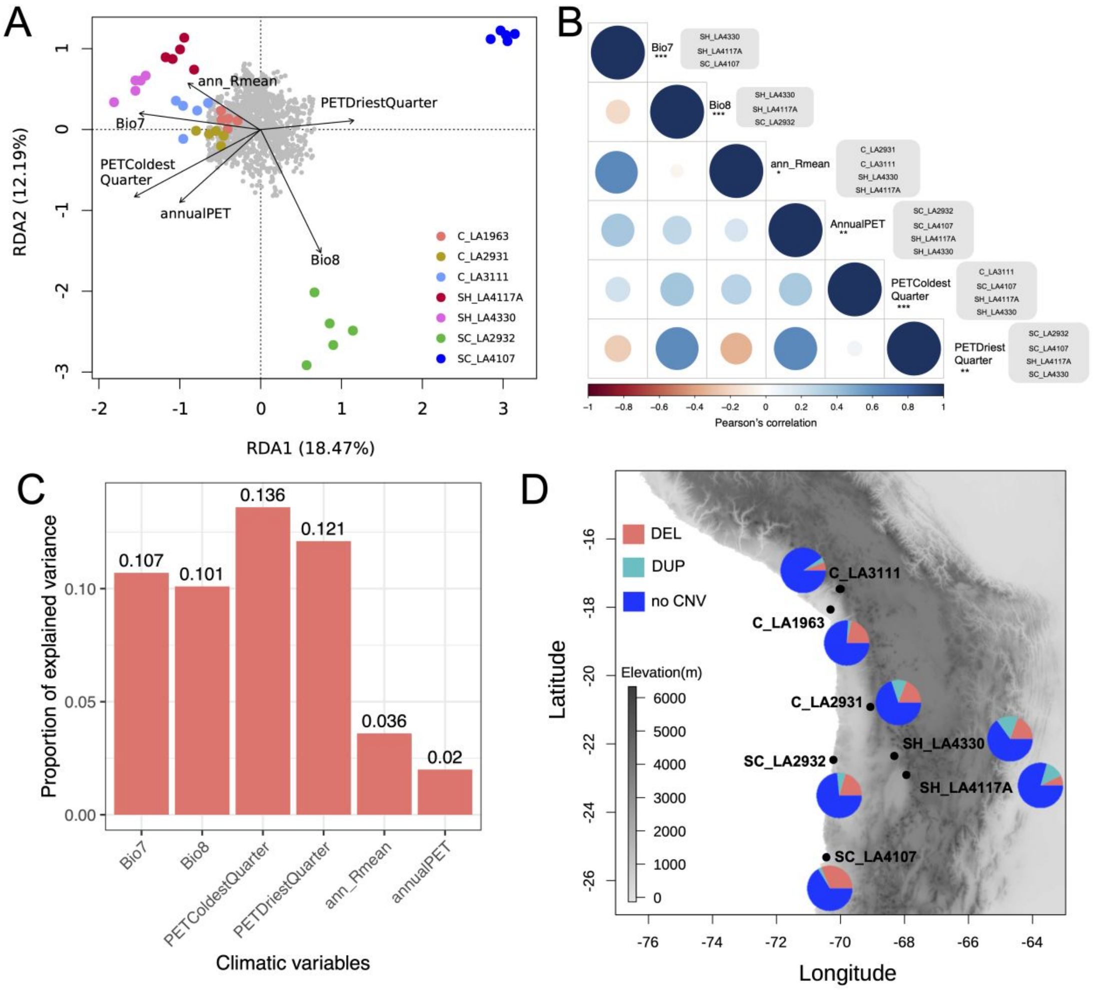

Fig. 5. Genome-Environment Association analysis reveals that CN differentiated genes adapt to different habitat environments. (A) Redundancy analysis (RDA) ordination biplots between the climatic variables (Dataset S7), populations and 3,539 differentiated gene CN. In the RDA, the loading of the climatic variables or the length of the vector indicates the strength of the correlation with the ordination axis. The grey points denote the CN-differentiated genes. C: central; SH: southern highland; SC: southern coast. (B) The correlations between six overrepresented climate variables and populations, respectively. The bubble chart shows correlations between six climate variables. The asterisks (*) indicate the levels of significance of the climate variables for the RDA model (Permutation test; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.0001$). The grey boxes to the right of the climatic variables show the populations significantly associated with that climatic variable (Mantel test, $P < 0.05$). (C) The proportion of explained variance for six overrepresented climate variables in the RDA model. (D) 34 CN-differentiated genes associated with both temperature annual range (Bio7) and solar radiation (ann_Rmean) in seven populations. The pie charts denote the proportions of CN-differentiated genes with deletion (DEL), duplication (DUP) or absence of CNV (see also Table S10).