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

Abstract

Copy Number Variations (CNVs) are genomic structural variations constituting genetic diversity and underpinning rapid ecological adaptation. The timing and target genes of adaptation by means of CNVs are yet to be explored at the population level in the tomato clade. Therefore, we characterize the CNV landscape of *Solanum chilense*, a wild tomato species, using whole-genome data of 35 individuals from seven populations. We identify 212,207 CNVs, including 160,926 deletions and 51,281 duplications. We find CNVs for both intergenic and coding regions, and a higher number of CNVs in recently diverged populations occupying more recently colonized habitats. Population structure analyses based on CNVs and single nucleotide polymorphisms are in agreement, highlighting that the distribution of CNVs is shaped by past demographic and colonization events. Furthermore, we identify 3,539 candidate genes with highly divergent copy number profiles across populations. These genes are functionally associated with response to abiotic stimuli and stress and linked to multiple pathways of flowering time regulation. Gene copy number variation 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. Genome-environments association ultimately links the dynamics of gene copy number to six climatic variables and suggests that natural selection has likely shaped patterns of CNV in response to the climatic changes during the southward range expansion of *S. chilense*. Our findings provide insights into the role of CNVs underlying ecological adaptation to recently colonized habitats.

Introduction

Copy number variation (CNV) is the main type of structural variation (SV) caused by genomic rearrangement, which mainly includes deletion (DEL) and duplication (DUP) events resulting from loss and gain of DNA segment copy number (CN) (1, 2). It is expected that CNV has a greater impact on gene function because it covers more base-pairs (3, 4) and has a higher per-locus mutation rate than point mutations (single nucleotide polymorphisms, SNPs) (5). CNV is recognized as an important underpinning of genomic divergence and local adaptation (4, 6, 7). Genome-wide studies confirm the importance of CNVs as underpinnings of stress response and yield improvement in multiple plants, such as maize (8), rice (9, 10), and *Arabidopsis thaliana* (7, 11). However, such studies have been so far conducted in selfing species and/or crops characterized by small population size and domestication bottlenecks (12-14). 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 effect of adaptive processes (here positive selection underpinning environmental adaptation). 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 (15).

The crop wild relative tomato species *Solanum chilense* is proven to be an excellent model species to study the genetic basis of adaptive evolution when colonizing novel habitats (16-18). The presence of outcrossing, gene flow, seed banks and relatively mild bottlenecks during the colonization of new habitats results in high effective population sizes (*N*e) as reflected by high nucleotide diversity and high recombination rates, meaning that this species has a high adaptive potential (17-20). *S. chilense* occurs in southern Peru and northern Chile, from mesic to very arid habitats around the Atacama Desert (21), and two groups of populations likely expanded during two distinct southward colonization events (16-18): one towards the coastal part of northern Chile (hereafter the southern coast (SC) group), and the other towards the high altitudes of the Chilean Andes (southern highland (SH) group) (Figure 1A). The populations currently occurring in the SC and SH habitats have been shown to exhibit signatures of past positive selection for adaptation to cold, drought, light (photoperiod), heat and biotic stresses (17, 18, 22-25). These events of recent positive selection at a cohesive set of genes and drought-responsive gene networks suggest the occurrence of genetic underpinnings to adaptation of novel habitats, during the southward expansion of *S. chilense* populations towards arid areas around the Atacama desert (26).

These adaptive signatures revealed by previous population genomic studies are based on scans for positive selection solely using single-nucleotide polymorphisms (SNPs). However, it is still unknown whether CNV can also contribute to adaptation to novel habitats in *S. chilense* and in the tomato clade in general.

Reference genomes of several species of the tomato clade, including numerous cultivated tomato varieties, are now sequenced and assembled (27-30). Two tomato SV sets have been recently constructed based on a tomato-clade pangenome analysis to investigate the impact of genome rearrangements on gene expression, genomic diversity and providing new genetic resource for improvement of tomato (31, 32). These two studies compare cultivated tomato genomes with that of several wild tomato species, including PacBio and Illumina sequencing from the *S. chilense* population LA1969 (belonging to our central group). We note with interest that *S. chilense* exhibits the highest number of SV among all wild and cultivated tomato species. This difference is even more striking when considering that the most related species (*S. peruvianum* and *S. corneliomulleri*) show half or less SVs than *S. chilense*. All three species exhibit a similar recent proliferation of transposable elements (TEs) (32). As *S. chilense* exhibits one of the largest genome sizes of the tomato clade and has the highest number of annotated genes, it is important 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 level across species (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 habitats in *S. chilense*) habitats.

We generate whole-genome CN profiles for 35 *S. chilense* plants from seven populations (five diploid individuals per population) representing four 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 indicative of recent positive selection. We then measure the evolutionary trend of CN expansion and contraction across different populations. We finally, associate the dynamics of gene CN with climatic variables to provide evidence for environment stresses driving CNVs dynamics across populations. Overall, our results suggest that gene CNV contributes to population adaptation to novel habitats in an outcrossing species with large effective population size and large genetic

diversity. We illustrate the importance of including 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 DELs and 51,281 DUPs) by aligning each of the 35 whole-genome sequencing datasets (Dataset S1) against the *S. chilense* reference genome using the combination of four diverse 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). The number of deletions (DELs) in all individuals and populations is much larger than the number of duplications (DUPs) (Fig. 1B and S1; Table S1 and S2). We find 37% to 43% of the CNVs to be identified in only one individual for three C 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 (SC and SH) populations, especially in the SC populations (25% in SC_LA2932 and 31% in SC_LA4107).

Although there are three times as many DELs as DUPs, the size of DUPs (39,140 bp +/- 104,577) is bigger than DELs and exhibiting a skewed distribution (14,052 bp +/- 59,930) (Fig. 1C; Kolmogorov-Smirnov test, P=2.2e-16). The DELs have a smaller size than DUPs as 56% of DELs display a size between 50bp and 1,000bp, against 26% for DUPs. We found that DELs and DUPs are enriched at both ends of the chromosomes (Fig. 1D), consistent with previous studies (4, 31, 32). 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* genome. 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 TEs (9, 31), and the annotation reveals that 68% DELs and 82% DUPs match at least one TE in the *S. chilense* genome. We also confirm that our CNVs are well-called by our pipeline as measured by the 90% validity score obtained for 100 simulated DELs and 100 DUPs (Table S5).

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 CNV dataset using SVtyper (33) is in agreement with the clustering from the genome-wide SNP dataset (Fig. 2A and S3A). We define four genetic (and geographic) subgroups. The first principal component (PC1) separates the SC populations from inland (C and SH) populations, PC2 separates the SC subgroup into two clusters (SC_LA2932 and SC_LA4107), and PC3 separates the inland populations into C and SH subgroups (Fig. 2A and S3A). The STRUCTURE analysis confirms these results (Fig. 2B and S3B with K=4 exhibiting the lowest cross-validation error) and is consistent with the results from the SNP dataset (Fig. S3C).

We further explore the differentiation of populations using the V_{ST} statistic which is analogous to the classically used F_{ST} for SNP data (34). We first compute the V_{ST} values along the full genome in 10 kb windows with a 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 (16-18), namely the high differentiation between SC and inland populations, especially between SC and SH 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 and 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 identify candidate genes defined as outliers for their CN distribution profiles across populations (Fig. S5). We perform first a permutation test (1,000 times) for each gene using the 35 samples of all seven populations. We obtain candidate genes by only considering those above a significant high threshold of differentiation 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 which 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 very 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 exhibit the distribution of DELs and DUPs for these 3,539 candidate genes. SH populations

display the most gene gains (DUPs) and the least gene loss (DELs), while the higher amount of gene loss is found in SC populations.

We performed 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 very differentiated gene set (Fig. S6D). In the PCA based on 23,911 genes with CN values (Fig. S6B), all samples appear to be grouped together except for SC_LA4107. The SC populations separate from the five inland populations (C and SH populations) in the second PCA (with V_{ST}(CN)>0; Fig. S6C). This suggests a large difference in the CN range and composition between SC and inland populations. Consistent with the PCA based on the genotyped CNVs (and previously on SNP data), PC3 separates the SH populations from the C populations when using the differentiated genes CN (Fig. 3A; Fig. S6D). Note, however, that SH populations still show ca. 20% of admixed ancestry coefficients with the C populations (Fig. 2B). These admixture signatures can be interpreted as resulting from gene flow post-colonization of the southern habitats between SH and C populations (18). These results indicate that the past demographic history of habitat colonization (and the resulting amount of genetic drift) is the main evolutionary process shaping SNP and CNV frequencies within and between populations of *S. chilense*.

Genes with CN differentiation are involved in multiple environmental responses

We perform a functional enrichment analysis of the 3,539 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 number of genes ranging from 82 (cell wall organization) to 580 (cellular metabolic process). 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 and S7A; Dataset S4). These findings corroborate the likely adaptive processes at genes underlying response to arid and high/low altitude environments in previous studies conducting scans for positive selection using SNPs in *S. chilense* (16, 18, 23-25, 35).

We find 227 genes associated with flowering (Fig. S7A and S7B), an important fitness trait

conditioning local adaptation in plant species (36). As a key part of the transition from vegetative to reproductive growth, flowering is influenced by a variety of environmental conditions, thus the link to the different 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 vernalization pathways. The latter represent two regulatory pathways of flowering time by the relative lengths of light-dark periods and low temperature, respectively (36, 37). These genes are more often duplicated in SH populations (Fig. 3A and 3B; Table S8; t-test, $P < 0.05$). These include the potential homologs of floral integrator genes FT and FD (36, 38, 39), putative homologs of CRY2, GI, ELF3 in the photoperiod pathway (36, 40), and putative homologs of AGL14 and SEP2 in the vernalization pathway (41, 42). These candidate genes are well known as flowering time regulator in *A. thaliana* and can be found in Dataset S5. Note that these potentially homologs of flowering regulatory genes show either no pattern (no CNV) or only copy loss in C and SC populations (Fig. 3A and 3B; Table S8; t-test, $P < 0.05$). These findings indicate that significant gene CN gains may promote adaptation in the southern highland habitats by affecting the regulation of flowering time via the photoperiod and vernalization pathways. In addition, other potential flowering regulatory genes in the differentiated gene set are likely involved in flowering regulation via the other pathways, namely the putative homologs of the genes FY and FLD (Dataset S5) (36, 43, 44).

We then focus more specifically on drought-responsive genes and find 60 CN differentiated genes linked to direct response to water deprivation (Fig. 3B; Dataset S4). The 60 genes are involved in multiple physiological metabolic responses of plant responses to drought, including the duplicated homologs of ABI4 and AFP1 of the abscisic acid (ABA) pathway, and putative homolog of WRKY33 transcription factor in all populations (with different CNs across populations; Dataset S5) and these genes are confirmed to respond to drought stress in *A. thaliana* and crops (45-47). WRKY33 also is associated with the temperature stress in tomato (48). Eleven genes are also found to belong to the drought-responsive metabolism gene network previous studied (26), these genes showing significantly higher expression in drought than normal conditions (Fig. S8; t-test, P=2.68e-05) based on the data in (26). In addition, the number of DEL and DUP genes which are linked to response to water deprivation are similar across all populations (Fig. S7D; Table S8). This indicates that all *S. chilense* populations exhibit similar drought adaptation via metabolic process.

Previous SNP studies concluded that root development would also be a fitness trait related to drought adaptation in *S. chilense* (18). 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, $P < 0.05$). This is consistent with the previous results of a genome scan for selection using SNP data (18).

CN expansion and contraction of differentiated genes in different populations

We reveal that a large number of CN differentiated genes are potentially involved in response to habitat specialization. To investigate the CN dynamics of these genes across populations, we perform an analysis of gene CN expansion and contraction across populations. Generally, the CN of the differentiated genes is expanded (CN gain) in the inland group with a high expansion rate of 1.788, while it is contracted (CN lost) in the SC group with a contraction rate of -0.818 (Fig. 4A). Within the inland group, the SH group exhibits an expansion of CN (expansion rate of 0.416), while the C group show a CN contraction rate (-0.767) three times higher than CN loss (Fig. 4A). This likely indicates that the high rate of CN expansion in the inland group is mainly due to SH populations exhibiting high CN gain (Fig. 4A). Both SH populations show distinct CN expansion with expansion rates of 1.663 (SH_LA4117A) and 1.375 (SH_LA4330). In the C group, although the C_LA1963 and C_LA2931 display a trend of CN contraction, the C_LA3111 exhibits a higher rate of CN expansion (1.037) similar to the both SH populations (Fig. 4A). We relate this to a high migration rate between the high-altitude C_LA3111 and SH populations, and/or the recent divergence of the SH group from the C group (18). In addition, the similar highland habitat environments (Fig. 1A) may also contribute to the same evolutionary trends of CN gain affecting a similar same set of genes for C_LA3111 and SH populations. Interestingly, the opposite results are observed between two SC 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; Fig. 4A). This follows our previous observation that the two SC groups show a high degree of differentiation possibly resulting from different demographic histories and selective pressures. These results are also consistent with the results of population structure (Fig. 2) and reflect the old southernmost colonization of the coastal habitats and the recent colonization of the highlands (18).

We define 155 “rapidly evolving genes” which exhibit significant CN expansion or contraction (Viterbi $P < 0.05$) across the different groups/populations (Fig. 4A; Dataset S6). The CNs of these 155

rapidly evolving genes also clearly support the population clusters in the PCA, while C_LA3111 appears closer to SH populations than to the other C populations (Fig. 4B). The highest number of such rapidly evolving genes are found in the SH populations (91 genes), including 71 significant CN expansion genes and 20 significant CN contraction genes related to photosynthesis of light reaction, long-day photoperiodism (flowering), response to UV light and cold, and metabolic processes. We obtain 56 genes rapidly evolving CN genes in the C populations (Fig. 4A). Few rapidly evolving genes in C_LA3111 and C_LA2931 with high altitude (> 2200 m) exhibit a significant trend of CN expansion at genes involved in long-day photoperiodism and developmental processes (Dataset S6). 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 SC 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 putative homologs of photosystem subunits (i.e psbB, petD) mainly involved in photosynthesis (Dataset S5) and may possibly underpin the high genetic differentiation at the CNV level between the two SC populations. In addition, the same CN rapidly evolving genes enriched for photosynthesis (light reaction) GO categories are found in C and SH groups (Fig. 4D). These potentially photosynthetic gene families appear to have been contracting (CN loss) in the C group and SC_LA2932, but expanding (CN gain) in the SH group and SC_LA4107, suggesting that changes in the photosynthetic pathway is also an important adaptive strategy across the different habitats in *S. chilense*.

Dynamics of gene CN associated with climatic gradients

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

We first implement a redundancy analysis (RDA) to identify overrepresented climate variables related to gene CN changes across the seven populations. As a control for the test, we first observe a lack of significant RDA model or overrepresented climate variables (Permutation test, $P > 0.001$) when using on the 20,372 genes that are not in the CN differentiated gene set (Fig. S9A). Second, 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 SC populations (Fig. S9B to S9D). The gene

CN differentiation of 52.11% can be explained by six overrepresented climate variables (explanatory variables) from five significant RDA axes (Permutation test, $P < 0.001$) based on the 3,539 CN differentiated genes (Fig. 5A and S9E). These climatic variables are significantly correlated (Mantel test, $P < 0.05$) with the CN profiles of differentiated genes (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 (C, SH and two SC habitats). RDA1 is correlated with the annual range of temperature (Bio7) and potential evapotranspiration during the driest period (PETDriestQuarter) and contributes to the differentiation between the SC and inland populations (Fig. 5A and 5B). RDA2 reflects the differentiation between two SC populations by mean temperature during wettest quarter (Bio8). RDA2 also summarizes a climatic gradient differentiating the inland populations from low altitude (C_LA1963) to highland (SH) and which is mainly driven by high altitude affected solar radiation (ann_Rmean) and potential evapotranspiration (annualPET and PETColdestQuarter) (Fig. 5A and 5B). These six climatic variables are primarily associated with the colonization of SH and SC 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 SC populations, respectively (Fig. 5A to 5C). These variables indicate that *S. chilense* encountered different climatic adaptations during the expansion in SC (drought) and inland (cold) populations. Moreover, changes in temperature (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 explain 3.6% of gene CN differentiation (Fig. 5A to 5C). The consistent RDA model is obtained using the 2,192 very differentiated genes (Fig. S9F to S9H).

CN differentiated genes underlying putative adaptation to climatic variables

We subsequently search for candidate genes (among the 3,539 CN differentiated genes) which may be associated with the six overrepresented climate variables using latent factor mixed models (LFMM) (Fig. S10A) (49, 50). We identify 312 CN differentiated genes which are significantly associated with the six climatic variables (z-test; calibrated $P < 0.01$; Fig. S10A and S10B; Dataset S8). The PCA based on the CN of these 312 candidate genes display consistent population clustering in the RDA models (Fig. S11A and 5A), indicative that the gradients of six climate variables reflect gene CN dynamic changes

across habitat environments of the seven populations. Among these 312 candidates, we find 217 genes to be associated with three PET climate variables (annualPET, PETDriestQuarter and PETColdestQuarter), of which 98 genes are shared between the three variables (Fig. S10B). Indeed, PET is the main measurement 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 found across populations (Fig. S11B and S11C). These physiological processes (ABA signaling pathway, root hair differentiation) have been found to be important responses to drought stress using transcriptome and genome analysis (18, 26). This confirms that drought tolerance is likely is the main driving environmental pressure driving CNV evolution across the different populations. Further, 69% (34/49) of genes associated with Bio7 are also observed to be correlated with ann_Rmean (Fig. S10B), these genes are mainly CN duplicated in SH populations and lost in SC populations (Fig. S5D; Table S9). This likely reflects the fact that not only cold but also high solar radiation is also an adaptive condition in SH 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) (51, 52). In addition, we also find 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 SC and especially SH populations is much higher than in the C populations (Fig. S11B). These duplicated (DUP) 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 formation of photosystem subunits (Dataset S5). The number of candidate genes found as DEL are fairly similar in different populations (Fig. S11C). Most deleted genes are related to plant growth and development. The GEA analyses by RDA confirm thus the adaptive relevance of gene CN expansion and contraction: (i) the differentiated genes in the C group appear mainly as CN contraction genes (DEL) while these appear at CN expansion (DUP) in the SH population; (ii) the adaptive gene CN changes reflect the past demographic history and colonization of novel habitats; (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 colonization 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 (4, 6, 9, 31, 32). Therefore, CNVs have ubiquitous roles in adaptive processes in ecology and evolution (2, 53-55). To better understand the genetic basis behind the fitness effect of CNVs in natural populations, we generate high-quality genome assemblies for 35 *S. chilense* individuals from seven populations, allowing us to identify a genome-wide CNVs dataset. Our pipeline of CNV calling resolves hundreds of thousands of CNVs. The number of CNV for each population of *S. chilense* is similar to numbers found in a previous tomato panSV-genome study for five *S. chilense* samples from one population LA1969 (32). 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 and S3) (9, 56). Based on the past demographic model we developed previously (18) as a neutral evolution baseline, we conclude that CNV is shaped by neutral processes as well as contributes to adaptive divergence between populations.

We identified four times more gene signatures putatively exhibiting footprints of adaptive divergence using CN profiles than when using genome scans based on SNPs (18). The number of CNVs under selection may be slightly inflated as we used only permutation tests to extract adaptive signatures and could not define a stringent simulation based false discovery rate threshold as we did for SNPs (18). CN differences of these genes across different populations reflect the neutral and selective differentiation process between populations (Fig. 3A and S6), demonstrating that CNVs need to be considered to obtain a full understanding of how selection shapes genomic variation. We also suggest that CNVs may be the primary source of adaptive genetic variation before point mutations (SNPs) (57). Overall, the evolutionary processes generating CNV diversity and divergence follows the historical demography of *S. chilense*, namely two southward independent colonization events. Genes CN appear as expansion in the southernmost SC_LA4107 and SH populations which underwent recent colonization events and exhibit lower population sizes (17, 18), while gene CN exhibits a trend of contraction in the C and SC_LA2932 populations (close to the center of origin). Therefore, we conclude that CN expansion or contraction are

not only due to neutral evolutionary processes (past demographic events), but likely reflect and underpin selective events during the two southward colonization events. Conversely, early established populations exhibit the evolutionary trends of adaptive loss of gene and function (58, 59), especially genes involved in plant growth and development in C populations, or the loss of genes involved in photosynthesis in C 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 stress, 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 emphasizes results based on SNP study that reproductive process, especially the regulation of flowering time, may play comprehensive roles for adaptation to multiple environmental conditions in *S. chilense*. In a previous study based on SNPs, although flowering and response to light (photoperiod) and cold (vernalization) have been shown to be key adaptive pathways for *S. chilense* populations to colonize southern habitats (18), we obtain here further important candidate genes with differentiated gene CN profiles involved in both flowering regulatory pathways for response to changes of photoperiod and cold. These genes (putative homologs of FT, FD, FLD, etc.) are duplicated in the SH populations. In addition, solar radiation is also a challenge for plants at high altitude. Many CN differentiated genes are indeed enriched in response to UV light (Figure 3B; Dataset S4), including homologs of genes known to be involved in anthocyanin accumulation. These CN differentiated CN genes related to anthocyanin pathway may provide potential source of natural variation for breeding tomato with anthocyanin (60). Indeed, the anthocyanin pathway is switched off in cultivated tomatoes by mutations in regulatory elements, anthocyanin-producing tomato varieties have been created by genetic engineering to obtain anthocyanin-rich purple fruits (61). More generally the large number of gene losses possibly in response to environmental stresses, may indicate that reduction of the genome size is a powerful evolutionary driver of adaptation (58, 59, 62). Further functional validation will help to understand the molecular mechanisms by which CNV drives adaptive evolution in natural populations.

GEA analysis ultimately links the dynamics of gene CN to six climatic variables and reveals population structure of CNVs in connection to four different habitat environments (Fig. 5A and 5B). These overrepresented climate variables are almost uniformly associated with SNPs in an RDA analysis (18).

These potential CNV–environmental interactions have been observed in *Arabidopsis thaliana* (11, 63), *Solanum lycopersicum* (31), *Theobroma cacao* (4), *Oryza sativa* (9, 10). We further explicit that CNVs play an important role for southward colonization in *S. chilense*. CNVs, especially duplications (DUPs) in SH populations exposed to typical high-altitude stresses show adaptations in gene with functions related to cold, change of photoperiod and solar radiation. The CN profiles of differentiated genes in SC 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. 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 quebrada in loma zone, whereas SC_LA4107 grows in extremely fine alluvial soil (with even some running water). These adaptive signatures were not found in previous studies based on genome scan of SNPs (18). The three C populations display mainly a trend towards gene loss and low correlation with climatic variables (Fig. 5A and 5B). 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 draft, migration, and recombination (64, 65). The two C populations (C_LA2931 and C_LA3111) found at high altitude exhibit few adaptive duplications signatures, but some as possible response to cold and solar radiation, similar to those observed for the SH populations (17, 18).

Our results reinforce the observation that CNV is an important contributor to adaptation across different ecological habitats (4, 6, 7). The strong selective pressure imposed by natural expansion of *S. chilense* and the need to adapt to novel habitats, has shaped the genetic diversity at SNPs as well as CNVs. In agreement with previous studies, we confirm that natural selection acting through CNVs can reshape population genome to underpin adaptation (2, 4, 6, 66).

Materials and Methods

For full materials and methods, see SI Appendix, Supplementary Information Text. Custom scripts for conducting the analyses are available at our Gitlab at the following link: https://gitlab.lrz.de/population_genetics/s_chilense_cnv.

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 (18) including quality trimming, mapping and SNP calling based on the new reference genome of *S. chilense* (available at Sol Genomics Networks: https://solgenomics.net/ftp/genomes/Solanum_chilense/Gustavo/).

Identification and genotyping of CNVs

We used LUMPY (67), Manta (68), Wham (69) and DELLY (70) to identify CNVs in the 35 samples. The CNV sets from LUMPY, DELLY, Manta and Wham were merged using SURIVOR v1.0.7 (71). The merged CNV set was inputted to SVTyper v0.7.0 to call genotypes using a bayesian maximum likelihood algorithm (33).

To assess the sensitivity and accuracy of our pipeline for CNV calling, we simulated 100 DUP and 100 DEL regions with sizes ranging from 50bp to 1Mb using CNV-Sim v0.9.2 utilizing the functionality of ART (72), and these simulated reads were used as 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 (73). The inference of population structure was performed using the program ADMIXTURE v1.3.0 (74) 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 CN. First, we used Control-FREEC v11.6 to estimate CN by 10 kb windows with 1 kb step size across the entire genome (75). We then obtained gene CN from the Control-FREEC outputs and gene coordinates in the genome. We also used Mosdepth v0.3.2 (76) 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 normalizing factor to obtain 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 CN genes to the *A. thaliana* dataset TAIR10 (e-value cutoff was 10-6) (77). We used, the R package clusterProfiler to perform GO enrichment analysis using *A. thaliana* annotation database as the background (78). The Benjamini-Hochberg method was used to calibrate initial *P* values, and calibrated *P* values smaller than 0.05 were used as 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 (79). The ultrametric tree was generated based on *force.ultrametric* function of phytools R package (80). We then performed analyses of expansion and contraction of gene CN using CAFE v4.2.1 (81). The branch-specific p-values are obtained by the Viterbi method with the randomly generated likelihood distribution. We set p-value smaller than 0.05 to detect gene CN with 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 (82) and ENVIREM (83). To evaluate the relative contribution of the abiotic environment to explaining patterns of genetic variation, we first used the Redundancy Analysis (RDA) (84) to associate CNs of the 3,539 differentiated genes with climatic variables. RDA was perform using the *rda* function from the vegan package as implemented in R (85). LFMM (latent factor mixed models) is a univariate test (49, 50), which means that it builds a model for each gene or SNP and each predictor variable. We then implemented LFMM2 to perform association test between gene CN and six representative climate variables obtained in RDA, respectively. The method of Benjamini-Hochberg was used to calibrate the *P* values with 0.01 as the significant threshold.

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.

Reference

1. L. Feuk, A. R. Carson, S. W. Scherer, Structural variation in the human genome. *Nature Reviews Genetics* **7**, 85-97 (2006).
2. A. Żmieńko, A. Samelak, P. Kozłowski, M. Figlerowicz, Copy number polymorphism in plant genomes. *Theoretical and applied genetics* **127**, 1-18 (2014).
3. T. H. Shaikh *et al.*, 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 (2009).
4. T. Hämälä *et al.*, 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 (2021).
5. J. R. Lupski, Genomic rearrangements and sporadic disease. *Nature genetics* **39**, S43-S47 (2007).
6. D. C. Rinker, N. K. Specian, S. Zhao, J. G. Gibbons, 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 (2019).
7. M. Marszalek-Zenczak *et al.*, Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members. *Front Plant Sci* **14**, 1104303 (2023).
8. N. M. Springer *et al.*, Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS genetics* **5**, e1000734 (2009).
9. R. R. Fuentes *et al.*, Structural variants in 3000 rice genomes. *Genome research* **29**, 870-880 (2019).
10. P. Qin *et al.*, Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542-3558.e3516 (2021).
11. A. Zmienko *et al.*, AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome[OPEN]. *The Plant Cell* **32**, 1797-1819 (2020).
12. C. Alonso-Blanco *et al.*, 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481-491 (2016).
13. T. M. Beissinger *et al.*, Recent demography drives changes in linked selection across the maize genome. *Nature Plants* **2**, 16084 (2016).
14. S. Brumlop, O. Weedon, W. Link, M. Finckh, Effective population size (Ne) of organically and conventionally grown composite cross winter wheat populations depending on generation. *European Journal of Agronomy* **109**, 125922 (2019).
15. M. Lynch, B. Walsh, *The origins of genome architecture* (Sinauer Associates Sunderland, MA, 2007), vol. 98.
16. K. B. Böndel *et al.*, North–south colonization associated with local adaptation of the wild tomato species Solanum chilense. *Molecular Biology and Evolution* **32**, 2932-2943 (2015).
17. R. Stam, G. A. Silva-Arias, A. Tellier, Subsets of NLR genes show differential signatures of adaptation during colonization of new habitats. *New Phytologist* **224**, 367-379 (2019).
18. K. Wei, G. A. Silva-Arias, A. Tellier, Selective sweeps linked to the colonization of novel habitats and climatic changes in a wild tomato species. *New Phytologist* **237**, 1908-1921 (2023).
19. U. Arunyawat, W. Stephan, T. Städler, Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Molecular Biology*

and Evolution **24**, 2310-2322 (2007).

20. A. Tellier, S. J. Laurent, H. Lainer, P. Pavlidis, W. Stephan, Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences* **108**, 17052-17057 (2011).
21. T. Nakazato, D. L. Warren, L. C. Moyle, Ecological and geographic modes of species divergence in wild tomatoes. *American Journal of Botany* **97**, 680-693 (2010).
22. H. U. I. Xia, L. Camus-Kulandaivelu, W. Stephan, A. Tellier, Z. Zhang, Nucleotide diversity patterns of local adaptation at drought-related candidate genes in wild tomatoes. *Molecular Ecology* **19**, 4144-4154 (2010).
23. I. Fischer, L. Camus-Kulandaivelu, F. Allal, W. Stephan, Adaptation to drought in two wild tomato species: the evolution of the Asr gene family. *New Phytologist* **190**, 1032-1044 (2011).
24. T. Nosenko, K. B. Böndel, G. Kumpfmüller, W. Stephan, Adaptation to low temperatures in the wild tomato species Solanum chilense. *Molecular Ecology* **25**, 2853-2869 (2016).
25. K. B. Böndel, T. Nosenko, W. Stephan, Signatures of natural selection in abiotic stress-responsive genes of Solanum chilense. *Royal Society open science* **5**, 171198-171198 (2018).
26. K. Wei *et al.*, Evolution of two gene networks underlying adaptation to drought stress in the wild tomato Solanum chilense. *bioRxiv*, 2023.2001. 2018.524537 (2023).
27. A. Ranjan, Y. Ichihashi, N. R. Sinha, The tomato genome: implications for plant breeding, genomics and evolution. *Genome biology* **13**, 1-8 (2012).
28. S. Sato *et al.*, The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641 (2012).
29. A. Bolger *et al.*, The genome of the stress-tolerant wild tomato species Solanum pennellii. *Nature genetics* **46**, 1034-1038 (2014).
30. R. Stam *et al.*, 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 (2019).
31. M. Alonge *et al.*, Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145-161. e123 (2020).
32. N. Li *et al.*, Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nature Genetics* 10.1038/s41588-023-01340-y (2023).
33. C. Chiang *et al.*, SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods* **12**, 966-968 (2015).
34. R. Redon *et al.*, Global variation in copy number in the human genome. *nature* **444**, 444-454 (2006).
35. R. Blanchard-Gros *et al.*, Comparison of Drought and Heat Resistance Strategies among Six Populations of Solanum chilense and Two Cultivars of Solanum lycopersicum. *Plants* **10**, 1720 (2021).
36. A. Srikanth, M. Schmid, Regulation of flowering time: all roads lead to Rome. *Cellular and Molecular Life Sciences* **68**, 2013-2037 (2011).
37. A. Gaudinier, B. K. Blackman, Evolutionary processes from the perspective of flowering time diversity. *New Phytologist* **225**, 1883-1898 (2020).
38. C. Liu *et al.*, Direct interaction of AGL24 and SOC1 integrates flowering signals in Arabidopsis. *Development* **135**, 1481-1491 (2008).

39. J. Putterill, E. Varkonyi-Gasic, FT and florigen long-distance flowering control in plants. *Current Opinion in Plant Biology* **33**, 77-82 (2016).
40. Y. Makita *et al.*, Identification of a dual orange/far-red and blue light photoreceptor from an oceanic green picoplankton. *Nature communications* **12**, 3593 (2021).
41. V. r. Hecht *et al.*, Conservation of Arabidopsis Flowering Genes in Model Legumes *Plant Physiology* **137**, 1420-1434 (2005).
42. Rigoberto V. Pérez-Ruiz *et al.*, XAANTAL2 (AGL14) Is an Important Component of the Complex Gene Regulatory Network that Underlies Arabidopsis Shoot Apical Meristem Transitions. *Molecular Plant* **8**, 796-813 (2015).
43. J. Z. Cheng, Y. P. Zhou, T. X. Lv, C. P. Xie, C. E. Tian, Research progress on the autonomous flowering time pathway in Arabidopsis. *Physiol Mol Biol Plants* **23**, 477-485 (2017).
44. S. Bao, C. Hua, L. Shen, H. Yu, New insights into gibberellin signaling in regulating flowering in Arabidopsis. *Journal of Integrative Plant Biology* **62**, 118-131 (2020).
45. S. Xiao, L. Jiang, C. Wang, D. W. Ow, Arabidopsis OXS3 family proteins repress ABA signaling through interactions with AFP1 in the regulation of ABI4 expression. *Journal of Experimental Botany* **72**, 5721-5734 (2021).
46. Z. Liu *et al.*, Phytocytokine signalling reopens stomata in plant immunity and water loss. *Nature* **605**, 332-339 (2022).
47. X. Luo *et al.*, Abscisic acid inhibits primary root growth by impairing ABI4-mediated cell cycle and auxin biosynthesis. *Plant Physiology* **191**, 265-279 (2022).
48. M. Guo *et al.*, A single-nucleotide polymorphism in WRKY33 promoter is associated with the cold sensitivity in cultivated tomato. *New Phytologist* **236**, 989-1005 (2022).
49. E. Frichot, S. D. Schoville, G. Bouchard, O. François, Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution* **30**, 1687-1699 (2013).
50. K. Caye, B. Jumentier, J. Lepeule, O. François, LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. *Molecular Biology and Evolution* **36**, 852-860 (2019).
51. V. E. Tossi *et al.*, Beyond Arabidopsis: differential UV-B response mediated by UVR8 in diverse species. *Frontiers in Plant Science*, 780 (2019).
52. P. S. Thompson, D. Cortez, New insights into abasic site repair and tolerance. *DNA repair* **90**, 102866 (2020).
53. P. Castagnone-Sereno *et al.*, Gene copy number variations as signatures of adaptive evolution in the parthenogenetic, plant-parasitic nematode Meloidogyne incognita. *Molecular ecology* **28**, 2559-2572 (2019).
54. S. Lauer, D. Gresham, An evolving view of copy number variants. *Current genetics* **65**, 1287-1295 (2019).
55. C. Mérot, R. A. Oomen, A. Tigano, M. Wellenreuther, A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution* **35**, 561-572 (2020).
56. I. H. Cheeseman *et al.*, Population structure shapes copy number variation in malaria parasites. *Molecular biology and evolution* **33**, 603-620 (2016).
57. Y. Dorant *et al.*, Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species. *Molecular Ecology* **29**, 4765-4782 (2020).

58. R. Albalat, C. Cañestro, Evolution by gene loss. *Nature Reviews Genetics* **17**, 379-391 (2016).
59. J. Helsen *et al.*, Gene Loss Predictably Drives Evolutionary Adaptation. *Molecular Biology and Evolution* **37**, 2989-3002 (2020).
60. S. Gonzali, P. Perata, Fruit Colour and Novel Mechanisms of Genetic Regulation of Pigment Production in Tomato Fruits. *Horticulturae* **7**, 259 (2021).
61. S. Colanero, P. Perata, S. Gonzali, What's behind Purple Tomatoes? Insight into the Mechanisms of Anthocyanin Synthesis in Tomato Fruits[OPEN]. *Plant Physiology* **182**, 1841-1853 (2020).
62. J. G. Monroe, J. K. McKay, D. Weigel, P. J. Flood, The population genomics of adaptive loss of function. *Heredity* **126**, 383-395 (2021).
63. S. DeBolt, Copy Number Variation Shapes Genome Diversity in Arabidopsis Over Immediate Family Generational Scales. *Genome Biology and Evolution* **2**, 441-453 (2010).
64. S. De Mita *et al.*, Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology* **22**, 1383-1399 (2013).
65. S. Manel *et al.*, Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Molecular ecology* **25**, 170-184 (2016).
66. R. C. Iskow, O. Gokcumen, C. Lee, Exploring the role of copy number variants in human adaptation. *Trends in Genetics* **28**, 245-257 (2012).
67. R. M. Layer, C. Chiang, A. R. Quinlan, I. M. Hall, LUMPY: a probabilistic framework for structural variant discovery. *Genome biology* **15**, 1-19 (2014).
68. X. Chen *et al.*, Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222 (2016).
69. Z. N. Kronenberg *et al.*, Wham: identifying structural variants of biological consequence. *PLoS computational biology* **11**, e1004572 (2015).
70. T. Rausch *et al.*, DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).
71. D. C. Jeffares *et al.*, Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature communications* **8**, 14061 (2017).
72. W. Huang, L. Li, J. R. Myers, G. T. Marth, ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593-594 (2012).
73. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76-82 (2011).
74. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655-1664 (2009).
75. V. Boeva *et al.*, Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423-425 (2012).
76. B. S. Pedersen, A. R. Quinlan, Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867-868 (2018).
77. C. Camacho *et al.*, BLAST+: architecture and applications. *BMC bioinformatics* **10**, 1-9 (2009).
78. G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* **16**, 284-287 (2012).
79. J. Pickrell, J. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings*, 1-1 (2012).
80. L. J. Revell, phytools: an R package for phylogenetic comparative biology (and other things).

Methods in ecology and evolution **3**, 217-223 (2012).

81. M. V. Han, G. W. Thomas, J. Lugo-Martinez, M. W. Hahn, 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 (2013).
82. S. E. Fick, R. J. Hijmans, WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology* **37**, 4302-4315 (2017).
83. P. O. Title, J. B. Bemmels, ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography* **41**, 291-307 (2018).
84. T. Capblancq, B. R. Forester, Redundancy analysis: A Swiss Army Knife for landscape genomics. *Methods in Ecology and Evolution* **12**, 2298-2309 (2021).
85. B. R. Forester, J. R. Lasky, H. H. Wagner, D. L. Urban, Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology* **27**, 2215-2233 (2018).

Figures

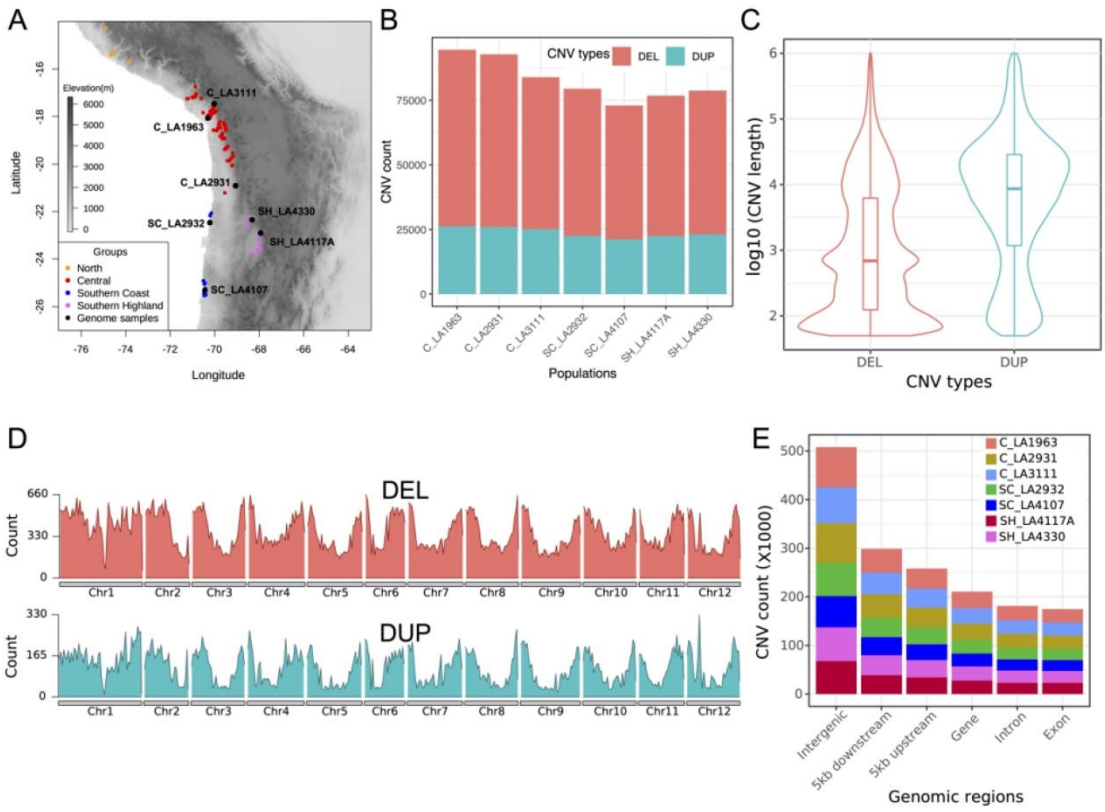


Fig. 1. The summary of the revealed CNVs in the genome of *S. chilense*. (A) Map with distribution of all *S. chilense* populations at the Tomato Genetics Resource Center (TGRC), the seven *S. chilense* populations in this study (black circles), the four population groups (circles with other colours).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

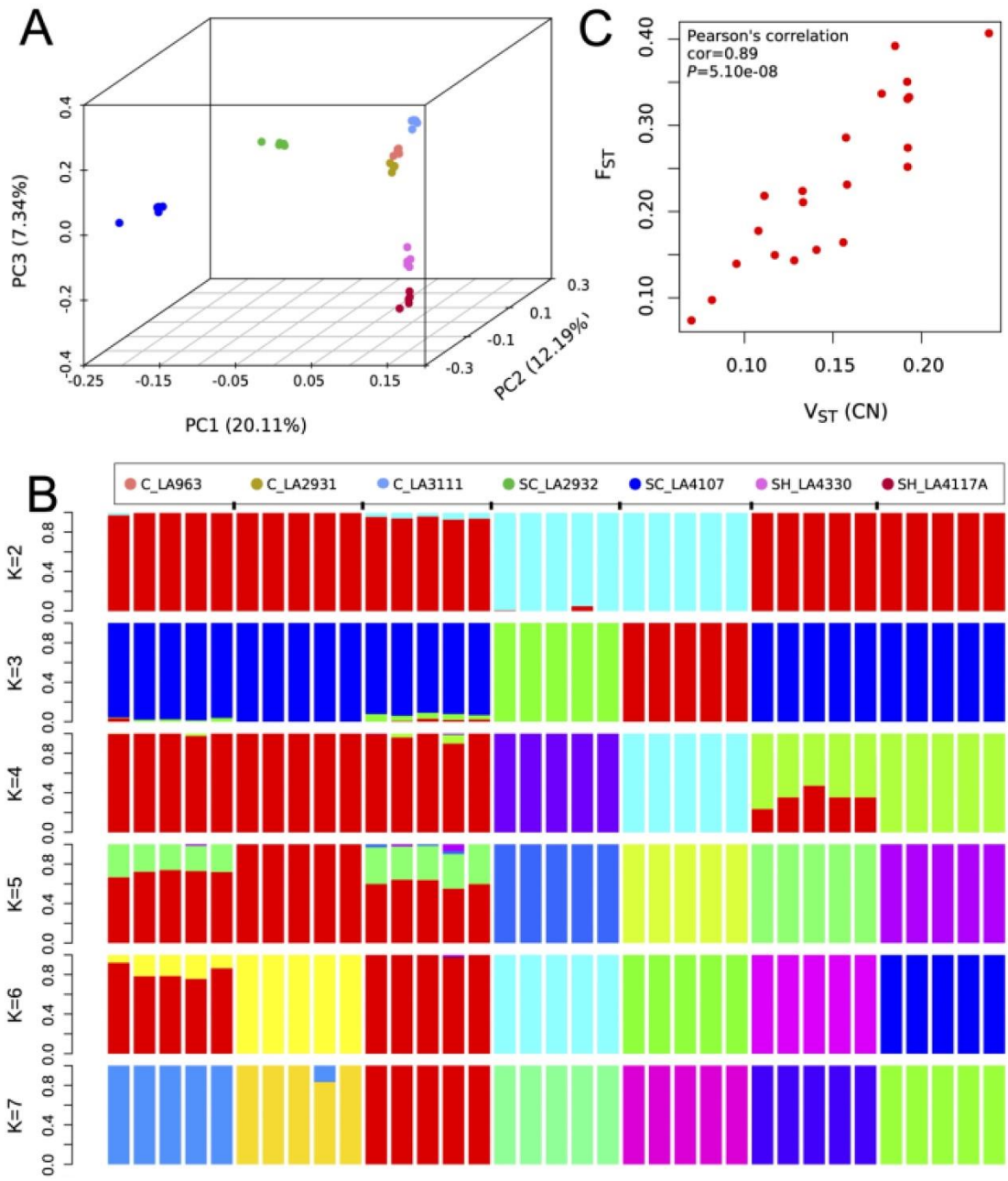


Fig. 2. Structure analysis based on genotyped CNVs. (A) Principal component analysis (PCA) based using genotyped CNVs from 35 S. chilense accessions. (B) Structure analysis based on genotyped CNVs and assuming *K* = [2, …, 7] subgroups (optimal *K* value is 4; Figure S2B). (C) The correlation between F_{ST} and V_{ST} indicates that CNVs inform that population differentiation.

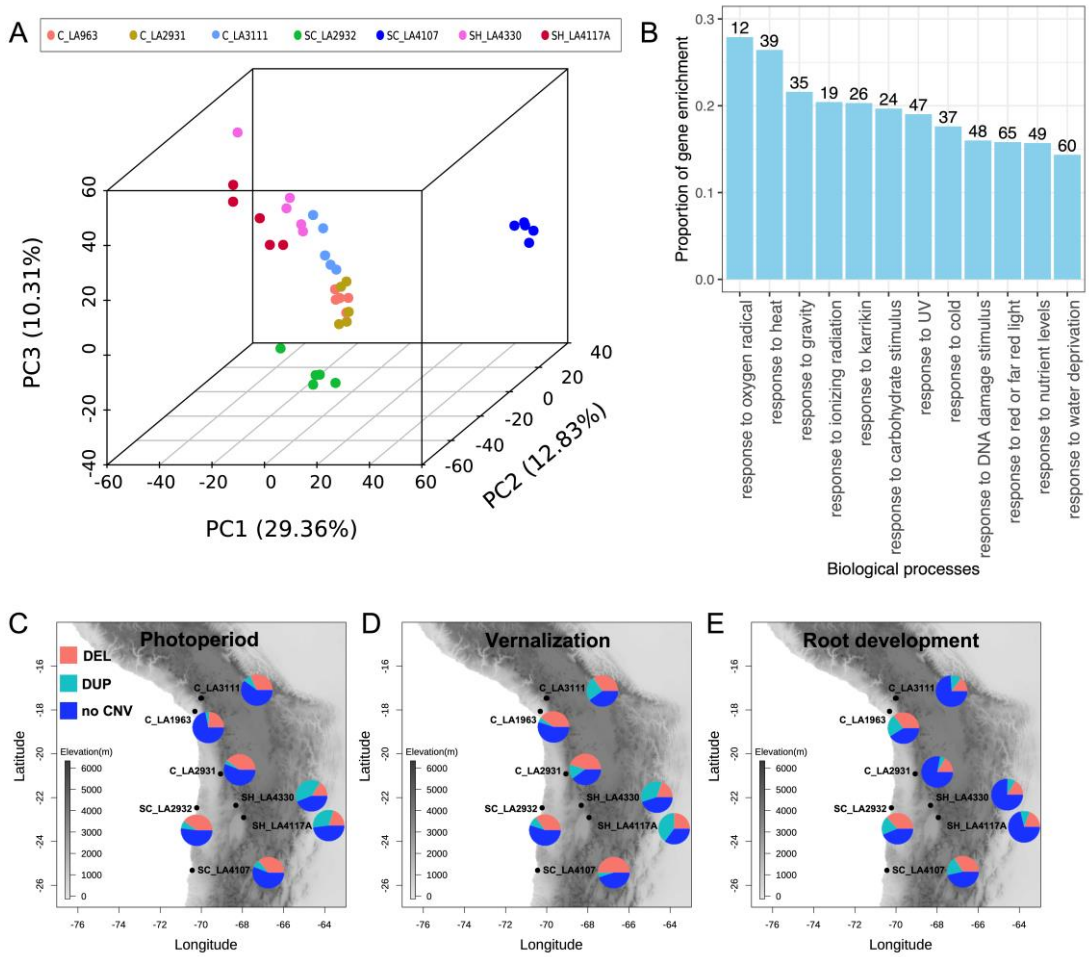
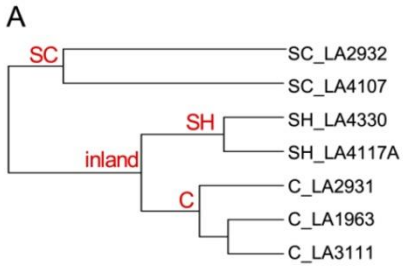


Fig. 3. Genes with differentiated CN profiles among seven populations are linked to response to multiple environmental stimulus. (A) PCA based on copy number (CN) of 3,539 differentiated genes. (B) The proportions of differentiated genes enriched in response to external stimulus/stresses (significantly enriched $P < 0.05$). The proportion 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), vernalization pathways to regulate flowering time (D), and root developmental process (E). The pie charts denote the proportions of CN differentiated genes with DEL, DUP or absence of CNV over all 3,539 genes (see also Table S8).



Nodes	Number of CN expanded genes	Number of CN contracted genes	Number of CN gained	Number of CN lost	Rate of average expansion/contraction	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)

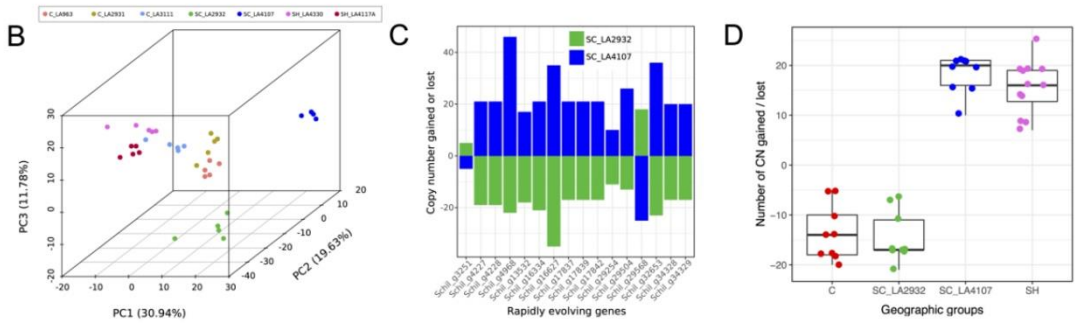


Fig. 4. The expansion and contraction of CN of differentiated genes in different populations. (A) The summary of CN expansion and contraction in different branches/populations in the phylogenetic and ultrametric tree. (B) PCA based on CN of 155 rapidly evolving genes with significant CN expansion or contraction (Viterbi $P < 0.05$). (C) The number of CN gained (positive values) or lost (negative values) for 16 rapidly evolving genes in two southern coast (SC) populations. (D) The number of CN gained or lost for rapidly evolving genes related to photosynthesis in different subgroups representing four different habitats.

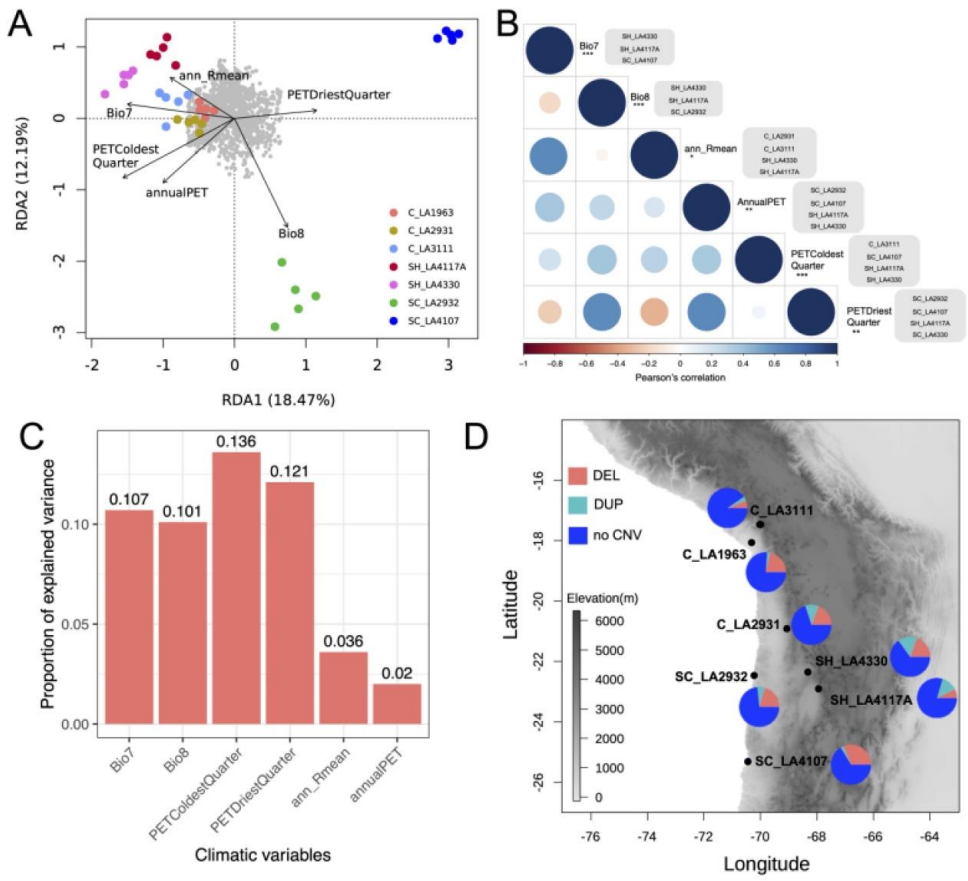


Fig. 5. Genome-environment association analysis reveals that CN differentiated genes adapt to different habitat environments. (A) Redundancy analysis (RDA) ordination biplots between the climatic variables (Dataset S7), populations and 3,539 differentiated gene CN. In the RDA, the loading of the climatic 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. (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 that are 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 DEL, DUP or absence of CNV (see also Table S9).