



**Copy number variation shapes structural genomic diversity  
associated with ecological adaptation in the wild tomato  
*Solanum chilense***

|                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Journal:                      | <i>Molecular Biology and Evolution</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| Manuscript ID                 | MBE-24-1101                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| Manuscript Type:              | Discoveries                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| Date Submitted by the Author: | 23-Oct-2024                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| Complete List of Authors:     | Wei, Kai; Xinjiang University College of Life Science and Technology, Xinjiang Key Laboratory of Biological Resources and Genetic Engineering; Technical University of Munich, Professorship for Population Genetics<br>Stam, Remco; Christian Albrechts University, Department of Phytopathology and Crop Protection<br>Tellier, Aurelien; Technical University of Munich, Department of Life Science Systems<br>Silva Arias, Gustavo; Universidad Nacional de Colombia - Sede Bogotá, Instituto de Ciencias Naturales; Technische Universität München, Life Science Systems |
| Keywords:                     | Flowering time, Population genomics, Abiotic stress response, Ecological adaptation, Gene duplication and deletion, Range expansion and colonization                                                                                                                                                                                                                                                                                                                                                                                                                          |
|                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |

SCHOLARONE™  
Manuscripts

**Copy number variation shapes structural genomic diversity associated with ecological adaptation  
in the wild tomato *Solanum chilense***

Kai Wei<sup>1,2\*</sup>, Remco Stam<sup>3</sup>, Aurélien Tellier<sup>2\*</sup>, Gustavo A Silva-Arias<sup>2,4\*</sup>

<sup>1</sup>Xinjiang Key Laboratory of Biological Resources and Genetic Engineering, College of Life Science and Technology, Xinjiang University, 830049, Urumqi, China

<sup>2</sup>Professorship for Population Genetics, Department of Life Science Systems, School of Life Sciences, Technical University of Munich, Liesel-Beckmann Strasse 2, 85354 Freising, Germany

<sup>3</sup>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

<sup>4</sup>Instituto de Ciencias Naturales, Facultad de Ciencias, Universidad Nacional de Colombia, Sede Bogotá, Av. Carrera 30 # 45-03, 111321, Bogotá, Colombia

\*Corresponding authors: Kai Wei: [kaiwei@xju.edu.cn](mailto:kaiwei@xju.edu.cn)  
Aurélien Tellier: [aurelien.tellier@tum.de](mailto:aurelien.tellier@tum.de)  
Gustavo A. Silva-Arias: [gasilvaa@unal.edu.co](mailto:gasilvaa@unal.edu.co)

## Abstract

Copy Number Variation (CNV) is a prevalent form of genetic variation affecting large genomic regions and contributes to both genetic diversity and ecological adaptation in plants. The target genes involved in adaptation through CNV in cultivated tomato and its wild relatives remain unexplored at the population level. Therefore, we characterized the CNV landscape of *Solanum chilense*, a wild tomato species adapted to dry habitats, using whole-genome short-read data of 35 individuals from seven populations. We identified 212,207 CNVs, including 160,926 deletions and 51,281 duplications. We found a higher number of CNVs in diverging populations occupying stressful habitats. CNVs and single nucleotide polymorphisms analyses concordantly revealed the known species' population structure, underscoring the impact of historical demographic and recent colonization events on the distribution of CNVs. Furthermore, we identified 3,539 candidate genes with highly divergent CNV profiles across populations. Interestingly, these genes are functionally associated with response to abiotic stress and linked to multiple pathways of flowering time regulation. Gene CNVs in *S. chilense* exhibit two evolutionary trends: gene loss in ancestral lineages distributed in central and southern coast populations and gene gain in the most recent diverged lineage from the southern highland region. Environmental association of the CNVs ultimately linked the dynamics of gene copy number to six climatic variables. It suggests that natural selection has likely shaped CNV patterns in response to the colonization of contrasting habitats during the southwards range expansion. Our findings provide insights into the role of CNV underlying adaptation in marginal populations.

20 **Introduction**

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

41 The tomato wild relative species *Solanum chilense* is an excellent model species to study the  
42 genetic basis of adaptive evolution when colonizing novel habitats (Böndel, et al. 2015; Stam, et al. 2019b;  
43 Wei, et al. 2023b). Features such as outcrossing, gene flow, seed banks, and relatively mild bottlenecks  
44 during the colonization of new habitats result in high  $N_e$ , as reflected by high nucleotide diversity and high  
45 recombination rates, meaning that this species has a high adaptive potential (Arunyawat, et al. 2007;  
46 Stam, et al. 2019b; Wei, et al. 2023b). *S. chilense* occurs in southern Peru and northern Chile, from mesic  
47 to very arid habitats around the Atacama Desert, and is the southernmost distributed species in the tomato  
48 clade (Nakazato, et al. 2010). Moreover, within *S. chilense*, two lineages expanded southward during two  
49 independent colonization events (Böndel, et al. 2015; Stam, et al. 2019b; Raduski and Igić 2021; Wei, et

al. 2023b): one, **early divergent** towards the coastal part of northern Chile (hereafter the southern coast group, **SC**), and the other **with a recent post-glacial divergence** towards the high altitudes of the Chilean Andes (**hereafter the** southern highland group, **SH**) (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 stress (Xia, et al. 2010; Fischer, et al. 2011; Nosenko, et al. 2016; Böndel, et al. 2018; Stam, et al. 2019b; Wei, et al. 2023b). This suggests a genetic basis for the adaptation to novel habitats during the southward expansion of *S. chilense* populations towards arid areas around the Atacama desert (Wei, et al. 2023a). **However, these studies revealed adaptive signatures based on scans for positive selection using solely SNP data: whether CNV can also contribute to adaptation to novel habitats in *S. chilense* is still unknown.**

Reference genomes of several species of the tomato clade, including numerous cultivated tomato varieties, **have been** sequenced and assembled (Ranjan, et al. 2012; Sato, et al. 2012; Bolger, et al. 2014; Stam, et al. 2019a). Three tomato SV sets have recently been 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 **compared** cultivated tomato genomes with that of several wild tomato species, including an individual of the *S. chilense* population LA1969 (belonging to our central group; Fig. 1A). Interestingly, **these studies showed** that *S. chilense* exhibited the highest number of SV among all wild and cultivated tomato species, **while** the closely related species *S. peruvianum* and *S. corneliomulleri* show only up to half of the number of SVs **found in *S. chilense*** (Li, et al. 2023). All these three species exhibit a similar recent proliferation of transposable elements (Li, et al. 2023). **As *S. chilense* occurs in a wide range of environments, this species is of key importance for the understanding of the role of CNV in speciation and intraspecific diversification processes in the tomato clade.** However, the studies mentioned above focused on the pangenome level across species (wild and cultivated), and an understanding of the role of CNV in local (ecological) adaptation is still lacking, especially for the adaptation to new arid habitats in southern populations in *S. chilense*.

**In this work,** we identified CNVs and **generated** whole-genome copy number (CN) profiles **based on genome-wide short-read sequencing data** for 35 *S. chilense* individuals 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 (Fig. 1A;

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Dataset S1). Based on these data, we first identified candidate genes with highly differentiated CN profiles between populations that are likely candidates associated with the inter-population differentiation and southward colonization in *S. chilense*. We then measured the evolutionary trend of CN expansion and contraction across different populations. Finally, we associated the dynamics of gene CN with climatic variables to provide evidence for environmental stresses driving CNV dynamics across populations. Our results suggest that CNV contributes to population adaptation to novel habitats in an outcrossing species with a large  $N_e$  and genetic diversity. We shed light on the importance of including an analysis of CNVs to complement genomic scans of recent positive selection based on SNPs.

**Results**

**Summary of CNVs in the genome of *S.chilense* and validation of the pipeline**

We identified a total of 212,207 CNVs (160,926 deletions and 51,281 duplications) using the combination of four CNV callers and the alignment of each of the 35 whole-genome sequencing datasets (Dataset S1) to the chromosome-level *S. chilense* reference genome (Silva-Arias, et al. 2024) (Fig. S1; Dataset S2). We found 73,014 to 94,621 CNVs per population (Fig. 1B; Table S1) and 31,923 to 46,579 CNVs per individual (Table S2). Although the number of deletions in all individuals and populations is much larger than the number of duplications (Fig. 1B; Fig. S1), the mean size of duplications (39,140 bp +/- 104,577) is larger than that of deletions (14,052 bp +/- 59,930) and exhibits a skewed distribution (Fig. 1C; Kolmogorov-Smirnov test,  $P=2.2e-16$ ). We found 37% to 43% of the CNVs to be private to one individual in the three central populations. In comparison, only 12% to 14% of all CNVs are fixed in each of the three central populations (Fig. S2), i.e., CNVs were observed in all five individuals of a given population. Southern populations (southern coast and southern highland) exhibited more fixed CNVs than the central populations, especially the two southern coast populations (25% in SC\_LA2932 and 31% in SC\_LA4107; Fig. S2).

Deletions and duplications were 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), about 35% to 38% of CNVs impacted coding sequences annotated in the *S. chilense* reference (some large CNVs were counted repeatedly due to covering multiple genes and intergenic regions). In addition, 45% and 50% of CNVs across populations overlapped with putative regulatory elements 5 kb upstream and 5 kb downstream of genes, respectively.

As expected, 68% of deletions and 82% of duplications matched at least one transposable element annotated in the *S. chilense* genome, supporting that CNVs are **predominately** shaped by transposable elements (Fuentes, et al. 2019; Alonge, et al. 2020).

To confirm the validity of our pipeline, which assembled CNV detection from four tools specialized 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 methods). **Our pipeline successfully** detected approximately 90% of the simulated CNVs, and the false-positive rate was much lower than based on a single caller (Table S3). Our results, as well as previous claims, indicated that combining multiple callers **can effectively improve the detection of CNVs based on short-read data** (Kosugi, et al. 2019; Mahmoud, et al. 2019; Coutelier, et al. 2022).

#### **CNVs effectively capture the known species population structure**

We compared the results of population structure analyses based on genome-wide SNPs and CNVs. The principal component analysis (PCA) based on the genotyped CNV dataset agreed with the clustering patterns from the genome-wide SNP dataset (Fig. 2A; Fig. S3A). **Both analyses suggested a division of our samples into four genetic clusters that aligned with the geographic structure of the populations.** The first principal component (PC1) separated the southern coast populations from inland (central and southern highland) populations, PC2 separated the southern coast subgroup into two **genetic** clusters (SC\_LA2932 and SC\_LA4107), and PC3 separated the inland populations into central and southern highland clusters (Fig. 2A; Fig. S3A). The ADMIXTURE analysis confirmed this result (Fig. 2B; Fig. S3B, with K=4 exhibiting the lowest cross-validation error) and was consistent with the results from the SNP dataset (Fig. S3C).

We further explored the population differentiation using the  $V_{ST}$  statistic. **This statistic is analogous to the classically used  $F_{ST}$  and  $D_{xy}$  statistics, but using CN values instead of allele frequencies (Redon, et al. 2006). The  $V_{ST}$  statistic ranges between 0 and 1, where 1 indicates that the populations are fully differentiated.** We first computed the  $V_{ST}$  values along the whole genome in 1 kb windows using two CN quantitative measurements: Control-FREEC ( $V_{ST}(CN)$ ) and read depth ( $V_{ST}(RD)$ ) (Table S4). We found a **highly significant positive** correlation between these two estimators of the  $V_{ST}$  statistic (Pearson's test,  $P=1.06e-07$ ; Fig. S4A). **In addition, all duplicated and lost fragments detected by Control-FREEC can be found in the CNV dataset obtained using the pipeline based on the four SV detection tools.** Based on the



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

138  $V_{ST}$  statistics, we found similar structure patterns as in previous studies **based on SNPs** (Böndel, et al.  
139 2015; Stam, et al. 2019b; Raduski and Igić 2021; Wei, et al. 2023b), namely the high differentiation  
140 between southern coast and inland populations, especially between southern coast and southern highland  
141 populations (Table S4). As expected, both  $V_{ST}$  statistics ( $V_{ST}(CN)$  and  $V_{ST}(RD)$ ) showed a highly  
142 significant positive correlation with  $F_{ST}$  and  $D_{xy}$  based on SNPs (Pearson's test,  $P$  values see Fig. 2C; Fig.  
143 S4B to D).

144 **Differentiation of gene CN profiles in different populations**

145 To explore the role of natural selection in shaping CNV frequencies and distribution across populations,  
146 we also calculated both  $V_{ST}$  statistics ( $V_{ST}(CN)$  and  $V_{ST}(RD)$ ) for each gene (39,245 genes in total). **We**  
147 **aimed to** capture candidate genes under divergent selective pressures by identifying genes with strong  
148 CN differentiation across populations (Fig. S5). In total, we identified 3,539 candidate genes that present  
149 outlier CN differentiation across the seven populations (*i.e.*, genes with  $V_{ST}$  greater than the top 95<sup>th</sup>  
150 percentile of the 1,000 permuted  $V_{ST}$  values; Fig. S5; Table S5; Dataset S3) and 2,192 strongly CN-  
151 differentiated genes of these belong to the top 99<sup>th</sup> percentile of the 1,000 permuted  $V_{ST}$  values (Fig. S5;  
152 Table S5; Dataset S3). In Fig. S6, we showed the distribution of deletions and duplications for these 3,539  
153 candidate genes. Southern highland populations **exhibited a comparatively large** increase in gene gains  
154 (duplications) and a small reduction in gene loss (deletions) **relative to the other populations**. In contrast,  
155 southern coast populations showed a comparatively **high number of deletions relative to the high-altitude**  
156 **populations**.

157 We performed four PCA analyses based on the Control-FREEC-based CN values of 1) all annotated  
158 23,911 genes **with the mapped reads** (Fig. S7A); 2) the 12,392 genes with  $V_{ST}(CN)>0$  (Fig. S7B); 3) the  
159 3,539 differentiated gene set (observed  $V_{ST}$  values > 95% confidence interval cutoff in both gene CN  
160 estimate methods; Fig. 3A); and 4) the 2,192 strongly differentiated gene set (observed  $V_{ST}$  values > 95%  
161 confidence interval cutoff; Fig. S7C). In the PCA based on the 23,911 genes (Fig. S7A), all samples  
162 exhibited a cohesive grouping, except **those from SC\_LA4107. In the PCA based on the 12,392 genes**  
163 **with  $V_{ST}(CN) > 0$**  (Fig. S7B), **two southern coast populations separated from the five inland populations**  
164 **(central and southern highland populations), suggesting** a large difference in the CN range and  
165 composition between southern coast and inland populations. In the PCA based on the differentiated gene  
166 set (Fig. 3A; Fig. S7C), PC3 separated the southern highland populations from the central populations,  
167 consistent with the PCA based on the genotyped CNVs and SNPs (Fig. 2A; Fig. S2A). Note, however,



that southern highland populations still showed ca. 20% of admixed ancestry coefficients with the central populations (Fig. 2B). These admixture signatures can reflect gene flow post-colonization of the southern habitats (between southern highland and central populations) or a very short divergence time. Consequently, similar polymorphisms in some parts of the genome were maintained between these populations (Wei, et al. 2023b). These results may indicate that the past demographic history of habitat colonization (and the resulting genetic drift) and gene flow are important evolutionary processes shaping both SNP and CNV frequencies within and between populations of *S. chilense*.

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

We performed functional enrichment analysis on the 3,539 CN-differentiated genes according to GO biological process categories (Dataset S4). We classified the significantly enriched GO categories ( $P < 0.05$ ) into nine groups (Fig. S8A) enriched for 82 genes (cell wall organization) up to 580 genes (cellular metabolic process). Interestingly, 400 (11.30%) CN-differentiated genes were enriched for a response to stimulus/stress that can be linked to multiple environmental factors (Fig. S8A), 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 light (UV; 19.03% with 47 genes) (Fig. 3B). The enrichment for these stress responses supported multiple sources of evidence for adaptation 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*) were found to be involved in flowering regulatory processes (Dataset S5). These findings suggested that selection pressures may occur at point mutations as well as at CNVs.

We found 227 CN-differentiated genes to be associated with flowering (Fig. S8A and B), an important fitness trait underlying 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 multiple environmental conditions. Therefore, divergent flowering times related to local adaptation processes along the ecological gradient may be driven by CN-differentiated genes (Fig. S8C). We found 31 and 36 CN-differentiated genes linked to response to light and cold among the genes involved in flowering regulation (Fig. S8C), of which 25 and 20 genes were linked to photoperiod and vernalization pathways (Fig. S8B). The latter represent two regulatory flowering time pathways sensitive to the relative lengths of light-dark

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

197 periods and low temperatures, respectively (Srikanth and Schmid 2011; Gaudinier and Blackman 2020).  
198 These genes showed a comparatively high overlap with duplications in southern highland populations (Fig.  
199 3C and D; Fig. S9; Table S6). These genes included the potential homologs of floral integrator genes *FT*  
200 and *FD* (Liu, et al. 2008; Srikanth and Schmid 2011; Putterill and Varkonyi-Gasic 2016), putative homologs  
201 of *CRY2*, *GI*, and *ELF3* in the photoperiod pathway (Srikanth and Schmid 2011; Makita, et al. 2021), and  
202 a putative homolog of *AGL14* in the vernalization pathway (Hecht, et al. 2005; Pérez-Ruiz, et al. 2015).  
203 These candidate genes are well-known flowering time regulators in *A. thaliana* (Dataset S5). Note that  
204 these potential candidate genes related to flowering regulation were duplicated only in southern highland  
205 populations and exhibited either no CNVs or copy loss in central and southern coast populations (Fig. 3C  
206 and D; Table S6; t-test,  $P < 0.05$ ). These findings indicate that genes with CN gains may promote  
207 colonization and adaptation in the southern highland habitats by regulating flowering time via the  
208 photoperiod and vernalization pathways (Wei, et al. 2023b). Remarkedly, this genomic finding was  
209 consistent with the phenology observed in glasshouse conditions, in which southern highland individuals  
210 consistently flower 5-10 days earlier than those from central populations. In addition, other potential  
211 flowering regulatory genes in the differentiated gene set were likely involved in flowering regulation via  
212 different pathways (Dataset S5), namely the putative homologs of the genes *FY* and *FLD* (Srikanth and  
213 Schmid 2011; Cheng, et al. 2017; Bao, et al. 2020). The *FLD* gene showed an increased copy number in  
214 all populations (Dataset S5).

215 We identified 60 drought-responsive CN-differentiated genes associated with direct responses to  
216 water deprivation (Fig. 3B), encompassing duplicated homologs of *ABI4* and *AFP1* in the abscisic acid  
217 (ABA) pathway, along with a putative *WRKY33* transcription factor homolog with varying CN across  
218 populations (Dataset S5). These genes were validated as drought stress-responsive in *A. thaliana* and  
219 crops (Xiao, et al. 2021; Liu, et al. 2022; Luo, et al. 2022), including *WRKY33*, which is linked to  
220 temperature stress in tomato (Guo, et al. 2022). Furthermore, eleven CN-differentiated genes also belong  
221 to the drought-response metabolism co-expression network we previously found to be over-expressed  
222 under drought compared to well-watered conditions (Fig. S10; t-test,  $P=2.68e-05$ ) (Wei, et al. 2023a),  
223 which corroborates their role in adaptive responses. Interestingly, we found similar numbers of deletion  
224 and duplication genes associated with water deprivation response across all populations (Fig. S8D; Table  
225 S6), suggesting a species-wide adaptation process in *S. chilense* through alterations in a metabolic gene  
226 network.

Our previous SNP study linked root development genes to putative local adaptation processes in coastal populations of *S. chilense* (Wei, et al. 2023b). Accordingly, we found 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 S6; t-test,  $P < 0.05$ ).

#### Gene expansion and contraction patterns show differences along altitudinal gradients

To this extent, our findings indicated that a considerable number of CN-differentiated genes may be implicated in the response to habitat specialization. To investigate the CN dynamics of these genes across populations, we performed an analysis of gene CN expansion and contraction across populations based on a phylogenetic tree derived from the inferred population genealogy (Fig. 4A). The CN of the differentiated genes was expanded (CN gain) in the inland group with an expansion rate of 1.788 (Table 1). On the other hand, we found a gene reduction (CN loss) in the southern coast group with a contraction rate of -0.818. Within the inland group, the southern highland group exhibited CN gain (expansion rate of 0.416). In contrast, the central group showed CN losses (contraction rate of -0.767) three times higher than CN gains (Table 1). This likely indicates that gene CN of inland populations presents different evolutionary trends along the two evolutionary lineages. The two southern highland populations showed 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 displayed a trend of CN contraction, the C\_LA3111 exhibited a similar rate of CN expansion (1.037) as the two southern highland populations (Table 1). The comparable CN expansion observed in the high-altitude populations (specifically, C\_LA3111, SH\_LA4330, and SH\_LA4117A) may be attributed to three factors: the recent divergence of the southern highland group from the central group, the recent (re-)colonization of highland habitats following the glacial maximum (Wei, et al. 2023b), and the ecological similarity of the habitats (Fig. 1A) which may also result in the duplication of a similar set of genes for C\_LA3111 and the southern highland populations.

Interestingly, the opposite results were observed between the two southern coast populations. Gene CN appeared to have contracted 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 showed a high degree of differentiation, possibly resulting from a long time of evolution in isolation and environmental differentiation. These results are also consistent with the population structure (Fig. 2) and may reflect the old southernmost colonization of the coastal habitats and

the recent colonization of the highlands (Stam, et al. 2019b; Wei, et al. 2023b).

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). Considering that the reference genome was assembled from population C\_LA3111, which probably does not represent the ancestral state of the species, we also performed the same analysis using gene CN profiles calculated from the reference genome of *S. pennellii*, a drought-adapted wild tomato species. We found consistent results, except for a slight decrease in the rate of CN expansion in C\_LA3111 (Table S7; Fig. S11).

We defined a set of 155 “rapidly evolving genes” that exhibited significantly higher CN expansion or contraction (Monte-Carlo re-sampling procedure;  $P < 0.05$ ) across the different groups/populations based on the reference genome of *S. chilense* (Table 1; Dataset S6). The CN profiles of the set of rapidly evolving genes also clearly supported the population clusters in the PCA (Fig. S12), but C\_LA3111 appeared closer to the southern highland populations than to the other central populations. The highest number of such rapidly evolving genes were found in the southern highland populations (91 genes), including 71 significant CN expanded genes with GO enriched for photosynthesis (light reaction), long-day photoperiodism (flowering), and response to UV light and cold. We also observed 20 significant CN-contracted genes primarily associated with developmental and metabolic processes. We also found 56 genes with rapidly evolving CN in the central populations (Table 1; Dataset S6), 75% of which exhibited a significant trend of CN contraction.

Among the 51 rapidly evolving genes in the southern coast populations, 16 genes showed opposite CN profiles: a significant contraction in SC\_LA2932 versus an expansion in SC\_LA4107 (Fig. 4C). These genes included few homologs of photosystem subunits (i.e., *psbB* and *petD*) 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 were also found in central and southern highland groups (Fig. 4D; Table S8). These potentially photosynthetic gene families appeared 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 (Fig. 4D; Table S8), suggesting that changes in the photosynthetic pathway may be also an important adaptive strategy across the different habitats in *S. chilense*.

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

287 To further explore CNV as the potential genetic basis of adaptive response to abiotic factors, we  
288 conducted two genome-environment associations (GEA) analyses between the gene CN profiles and 37  
289 climate variables (Dataset S7).

290 We first implemented a redundancy analysis (RDA) to identify climate variables significantly  
291 associated with CN-differentiated genes across the seven populations. Three climatic variables were  
292 observed to correlate with CN changes in the RDA based on 12,391 genes with  $V_{ST}(CN) > 0$  (Fig. 13A).  
293 The first three RDA axes retained only 22.62% of the putative adaptive gene CNV and only weakly  
294 distinguished between inland and southern coast populations (Permutation test,  $P < 0.001$ ; Fig. S13A to  
295 C). In the RDA based on the 3,539 CN-differentiated genes, 52.11% of the variance in CN can be  
296 explained by six climate variables (explanatory variables) from five significant RDA axes (Permutation  
297 test,  $P < 0.001$ ; Fig. 5A; Fig. S13D). These climatic variables were significantly correlated with the different  
298 populations (Mantel test,  $P < 0.05$ ; Fig. 5B). In concordance with the PCA (Fig. 2A), the two main  
299 ordination axes did cluster the seven populations into four groups corresponding to the main geographical  
300 habitats (central, southern highland and two southern coast habitats). RDA axis 1 (RDA1) was correlated  
301 with the annual temperature range (Bio7) and potential evapotranspiration during the driest period  
302 (PETDriestQuarter). This axis represented the differentiation between the southern coast and inland  
303 populations (Fig. 5A and B). RDA axis 2 (RDA2) reflected the differentiation between two southern coast  
304 populations by mean temperature of the wettest quarter (Bio8). RDA2 also summarized a climatic gradient  
305 differentiating the low altitude (C\_LA1963) and highland populations, which was mainly driven by solar  
306 radiation (ann\_Rmean) and potential evapotranspiration (annualPET and PETColdestQuarter) (Fig. 5A  
307 and B). These six climatic variables were primarily associated with the colonization of southern highland  
308 and southern coast populations (Fig. 5B). The proportions of gene CN differentiation explained by these  
309 six climatic variables ranged from 0.02 (annualPET) to 0.136 (PETColdestQuarter) (Fig. 5C), in which  
310 PETColdestQuarter and PETDriestQuarter (0.121) exhibited the highest importance and correlated with  
311 inland and southern coast populations, respectively (Fig. 5A to C). Moreover, temperature changes (Bio7  
312 and Bio8) also explained about 20.8% of the gene CN differentiation (Fig. 5C). Solar radiation  
313 (ann\_Rmean) was a specific variable correlated with high altitude populations and explained 3.6% of gene  
314 CN differentiation (Fig. 5A to C). A consistent RDA model was obtained using the 2,192 strongly CN-  
315 differentiated genes (Fig. S13E to G). Finally, as a comparison with the RDA based on the CN-

differentiated genes, we observed a null RDA model without significant RDA axis and associated climate variables (Permutation test,  $P < 0.001$ ) when implemented on the 20,372 genes that were not in the CN-differentiated gene set (Fig. S13H). This may further corroborate that the CN-differentiated genes respond to external environmental stimuli in *S. chilense*.

We subsequently searched for candidate genes that may be associated with the six overrepresented climate variables using latent factor mixed models (LFMM2) (Fig. S14A) (Frichot, et al. 2013; Caye, et al. 2019). Here, we performed an RDA association analysis between the climatic variables and 3,539 highly CN-differentiated genes (not all genes). We identified 312 CN-differentiated genes significantly associated with the six climatic variables (z-test; calibrated  $P < 0.01$ ; Fig. S14 B; Dataset S8). The PCA based on the CN of these 312 candidate genes displayed population clustering consistent with the one found in the RDA model (Fig. S15A; Fig. 5A), supporting that the six climate variables reflected gene CN changes across the species distribution. Among these 312 candidates, we found 217 genes to be significantly associated with three PET climate variables (annualPET, PETDriestQuarter, and PETColdestQuarter), of which 98 genes were shared between the three variables (Fig. S14B). Indeed, PET was the primary variable reflecting the drought status of the habitat. We noted that these PET-associated CN-differentiated genes were mainly GO-enriched in metabolic and root development processes and were found across all populations (Fig. S15B and C). The respective physiological processes (ABA signaling pathway, root hair differentiation) were essential responses to drought stress, as previously suggested by transcriptome and genome analysis (Wei, et al. 2023a; Wei, et al. 2023b). This result confirmed that drought tolerance is likely the main environmental pressure driving CN evolution across the population distribution of *S. chilense*. Furthermore, 69% (34/49) of the genes associated with Bio7 were also observed to be correlated with ann\_Rmean (Fig. S14B), because of the correlation between Bio7 and ann\_Rmean (Fig. 5B; Pearson's correlation = 0.50). These genes were mainly duplicated in the southern highland populations and lost in the southern coast populations (Fig. 5D; Table S9). This result 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) were enriched for response to UV light in high-altitude populations, such as (likely) homologs of UV-B receptor *ARI12*, and DNA repair gene *REV1* (Dataset S5) (Tossi, et al. 2019; Thompson and Cortez 2020). In addition, we also found a few CN-differentiated genes, such as putative homologs of *CPD*



(Dataset S5), which related to pigment (anthocyanins) accumulation and were statistically associated with solar radiation variables.

We finally observed that the number of duplicated genes associated with the six climatic variables in the southern coast and especially southern highland populations was much higher than in the central populations (Fig. S15B). The analysis of GO enrichment above showed that these duplicated genes are involved in response to environments, including light, drought, cold, UV, and photosynthesis, such as the 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 was similar in different populations (Fig. S15C). The analysis of GO enrichment showed that most lost genes are related to plant growth and development. The GEA analyses confirmed the adaptive relevance of gene CN expansion and contraction: (i) the CN-differentiated genes in the central group appeared mainly as contraction genes (deletions) while these appeared as expansion genes (duplications) in the southern highland populations; (ii) the gene CN changes were linked to the climatic variables and associated with colonization of novel habitats at the southern edge of the species distribution; and (iii) the expansion/contraction of gene CN in different populations and RDA model also matched the population structure.

## Discussion

In this study, we explored the role of genomic CNV in the ecological adaptations of *S. chilense*. A set of key genomic CNVs in *S. chilense* populations were found to be highly correlated with the species colonization process and environmental variables and thus were likely implicated in the adaptive differentiation between populations, probably 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 CNV has ubiquitous roles in adaptive processes in ecology and evolution (Żmienko, 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 CNV in natural populations, we analyzed whole-genome (short read) data for 35 *S. chilense* individuals from seven populations, which allowed us to identify genome-wide CNVs. Our CNV calling pipeline resolved hundreds of thousands of CNVs in *S. chilense*. The number of CNV for each population of *S. chilense* was similar to numbers found in the previous tomato clade CNV based on a pan-genome study that included a single sample of *S. chilense* (Li, et al. 2023). CNVs were abundant across all chromosomes and frequently resided within, or in close proximity to, genes in the *S.*



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

374 *chilense* genome (Fig. 1). Widespread CNVs in *S. chilense* genome exhibited similar performance as SNPs  
375 for the inference of population structure and differentiation between populations (Fig. 2; Fig. S3). Based  
376 on the demographic model we developed previously (Wei, et al. 2023b) as a neutral **null model** and the  
377 dynamics of CN **profiles** in two southward colonization events, our results supported that most CNV is  
378 likely shaped by neutral processes (Silva-Arias, et al. 2024). However, a genome-wide **perspective**  
379 allowed us to identify CNV likely related to the adaptive divergence in recently colonized regions in  
380 response to abiotic stress.

381 **We identified patterns of gene CN differentiation that are likely to represent footprints of adaptive**  
382 **divergence.** CN differences of these genes across different populations reflected the neutral and divergent  
383 selection process between populations, demonstrating that CNV must be considered to fully understand  
384 how selection shapes genomic structural diversity and local adaptation. Overall, the evolutionary  
385 processes generating CNV diversity and divergence **were dominated by the demographic history** of *S.*  
386 *chilense*, namely two southward independent colonization events. **Gene CN appears** expanded in the  
387 southernmost SC\_LA4107 and southern highland populations, which underwent recent colonization  
388 events and exhibited lower population sizes (Stam, et al. 2019b; Wei, et al. 2023b), while gene CN  
389 revealed a trend of contraction in the central and SC\_LA2932 populations (close to the species' center of  
390 origin). Therefore, we estimated that CN expansion and contraction likely reflect and underpin selective  
391 events during the two southward colonization events. **Conversely, some plants exhibit adaptive evolution**  
392 **by gene loss, for example, adaptive gene loss has been associated with changes in pollinators in *Petunia***  
393 ***axillaris* (Hoballah, et al. 2007), *Ipomoea quamoclit* (Zufall and Rausher 2004) and *A. thaliana* (Shimizu,**  
394 **et al. 2008). In this study, adaptive gene loss may also occur in genes involved in plant growth and**  
395 **development in central populations, and genes involved in photosynthesis in central and SC\_LA2932**  
396 **populations (Fig. 4D). This confirmed the critical role of gene loss in adaptive evolution. Changes in CN**  
397 **at photosynthetic genes** underpin population differentiation between SC\_LA2932 (gene loss) and  
398 SC\_LA4107 (gene gain), **two populations in** two different habitats **on** the southern coast. CN differentiated  
399 genes were also enriched in response to multiple abiotic stresses, such as red/far red light, cold, UV, or  
400 drought. These response processes can directly affect plant reproduction and growth and regulate  
401 flowering regulatory processes (Fig. S8). **These findings agreed with our** results based on SNPs showing  
402 that the reproductive cycle, namely the regulation **of** flowering time, may play a key role in adaptation to  
403 abiotic stress in *S. chilense* (Wei, et al. 2023b). **In addition to the ecological adaptation, we cannot ignore**

the shaping of gene CN patterns by neutral evolutionary events (genetic drift and mutation) especially during population expansions and bottlenecks.

The regulation of flowering time involved in response to light (photoperiod) and cold (vernalization) appear as key adaptive pathways for *S. chilense* populations to colonize southern habitats as suggested by the analysis of genome-wide SNPs (Wei, et al. 2023b). Here, we obtained further candidate genes based on differentiated gene CN profiles enriched for flowering regulatory pathways and response to changes in photoperiod and cold. These genes (putative *FT*, *FD*, *FLD* homologs) are duplicated in the southern highland populations (Fig. S9). Solar radiation is also a challenging condition for plants at high altitudes. Many CN-differentiated genes were enriched for a function in response to UV light (Fig. 3B; Dataset S4), including homologs of genes involved in anthocyanin accumulation in response to UV light. In plants, anthocyanin accumulation can improve the tolerance for drought, cold, salt and biotic stresses (Kaur, et al. 2023), especially anthocyanins act as potent antioxidants which help in eliminating Reactive Oxygen Species (ROS) molecules and protect the DNA damage under UV radiations (Catola, et al. 2017; Fang, et al. 2019). This may indicate that the gene CNVs in anthocyanin accumulation pathway are important for adaptation in high altitude populations of *S. chilense*. This follows a previous ecological niche study which suggested that *S. chilense* populations are expanding to the habitats of high altitude (Wei, et al. 2023b). More generally, the large number of gene losses 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 copy number variant drives adaptive evolution in natural populations.

RDA analysis ultimately linked the dynamics of gene CN across populations to six climatic variables (Fig. 5A and B), of which five climatic variables were consistent with previous RDA results based on SNPs (Wei, et al. 2023b). Similar CNV-environmental interactions have been observed in *A. thaliana* (DeBolt 2010; Zmienko, et al. 2020), *S. lycopersicum* (Alonge, et al. 2020), *Theobroma cacao* (Hämälä, et al. 2021), and *Oryza sativa* (Fuentes, et al. 2019; Qin, et al. 2021). Our results also highlight that CNV likely plays an essential role in response to the environments and in the southward colonization in *S. chilense*. CNVs, especially duplications in southern highland populations exposed to typical high-altitude stresses, were enriched in genes with functions related to cold, change of photoperiod and solar radiation.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

433 The CN profiles of differentiated genes in southern coast populations mainly correlated with drought stress,  
434 such as root development, cell homeostasis, or cell wall maintenance. Interestingly, gene CN  
435 differentiation related to photosynthesis provided evidence for the genetic underpinning of the adaptive  
436 differentiation between SC\_LA2932 and SC\_LA4107, representing two different coastal habitats (Fig. 1A  
437 and 4C). These differentiated genes revealed opposite CN evolutionary trends between the two southern  
438 coast populations. Indeed, we saw different habitats as SC\_LA2932 grows in dry ravines (quebrada) in  
439 Lomas formations, whereas SC\_LA4107 grows in extremely fine alluvial soil (with even some running  
440 water). Moreover, these chloroplast genes were detected in the nuclear genome, consistent with  
441 widespread events of organellar gene transfers to the nuclear genome in tomatoes (Pesaresi, et al. 2014;  
442 Lichtenstein, et al. 2016; Kim and Lee 2018). Since the chloroplast genome is much more conserved than  
443 the nuclear genome in plants, the transfer of chloroplast genes to the nuclear genome facilitates an  
444 increase in the genetic diversity of chloroplast genes (Daniell, et al. 2016). These adaptive signatures  
445 related to photosynthesis were not found in previous studies based on genome scans of SNPs (Wei, et  
446 al. 2023b). The three central populations displayed mainly a trend towards gene loss and low correlation  
447 with climatic variables (Fig. 5A and B). This is consistent with the fact that GEA analyses based on current  
448 climatic data have limited statistical power to detect old adaptive selection signals, whether based on  
449 SNPs or CNVs, due to the occurrence of multiple historical confounding events such as genetic drift,  
450 migration, and recombination (De Mita, et al. 2013; Manel, et al. 2016). The two central populations  
451 (C\_LA2931 and C\_LA3111) found at high altitudes exhibit few adaptive duplication signatures, but some  
452 as possible responses to cold and solar radiation, similar to those observed for the southern highland  
453 populations (Stam, et al. 2019b; Wei, et al. 2023b).

454 Finally, we would like to stress that our study likely underestimates the amount and importance of  
455 CNV in *S. chilense* as we do not possess long-read data for all populations. First, the tests with simulations  
456 based on the short-read data showed that our pipeline based on four tools to recover CNVs was likely  
457 conservative, meaning that we probably missed some CNVs. Second, there may be bias in finding  
458 footprints of selection when using seeds from accessions maintained and propagated at the Tomato  
459 Genetics Resource Center (TGRC; UC Davis, USA), as we discussed previously (Wei, et al. 2023b). We  
460 also point out that the detection of CN-differentiated genes by the  $V_{ST}$  statistic might be inflated because  
461 it is hard to correct for multiple testing (especially without a neutral demographic model of CNV evolution).

462 The availability of a new reference genome (Silva-Arias, et al. 2024) and a small number of populations

sequenced with long-read (Li, et al. 2023) do open the path to sequence wild populations with long-read sequencing and a complete assessment of the importance of CNV at abiotic stress genes in *S. chilense*. In addition, new simulation methods can be used to study and infer the neutral and selective processes driving gene duplication and deletion (Otto, et al. 2022; Otto and Wiehe 2023). These may be valuable options to quantify/infer the neutral rates of gene duplication/deletion during the species southward expansion, and, thereby, develop robust statistical selection tests for CNVs. Furthermore, instead of using the CN dataset of all genes to perform association analyses with climate variables, we used genes with high CN differentiation. The reason for this is that in the RDA analysis we did not obtain any associated climate variables when using CN dataset for all genes, indicating that the large number of genes with weak CN changes greatly reduced the resolution of the analysis.

Despite being conservative regarding the importance of positive selection shaping the CNV diversity in *S. chilense*, our results reinforced 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 suggest that natural selection acting on CNVs can reshape the genomic composition of populations and might form a basis for local adaptation (Iskow, et al. 2012; Żmieńko, et al. 2014; Rinker, et al. 2019; Hämälä, et al. 2021).

## Materials and Methods

### Sequence Read Processing

We retrieved whole-genome short-read sequencing data from 35 specimens from seven populations of *S. chilense* (five diploid plants for each population) representing four different geographic groups and environments (Fig. 1A). The data are available on the European Nucleotide Archive (ENA; BioProject accession no. PRJEB47577). We executed the same pipeline of read processing procedure as in our previous study (Wei, et al. 2023b), including quality trimming, mapping and SNP calling based on the reference genome of *S. chilense* (Silva-Arias, et al. 2024). The results of the sequencing and read mapping were documented in Dataset S1.

## 490 Identification and genotyping of CNVs

491 To obtain high-confidence CNVs including deletions and duplications, we chose four software tools for  
492 structural variation (SV) detection based on an evaluation of SV detection tools by Kosugi et al. (2019).  
493 This evaluation enumerated good tools for each SV type, among which LUMPY (Layer, et al. 2014), Manta  
494 (Chen, et al. 2016), Wham (Kronenberg, et al. 2015) and DELLY (Rausch, et al. 2012) were the better  
495 tools for the identification of deletions and/or duplications. These tools combined with multiple algorithms  
496 to detect CNVs using whole-genome sequencing data, including read depth, paired-end mapping, split  
497 read and *de novo* assembly approaches.

498 For Lumpy v0.3.1, we first extracted the discordant paired-end reads with abnormal insertion size  
499 from mapped results using 'view' function of Samtools v1.7 (Wysoker, et al. 2009), and the split-read  
500 alignments also were extracted using 'extractSplitReads\_BwaMem' script in Lumpy package. We used  
501 the 'sort' function of Samtools to sort the resulting BAM files. Next, we ran Lumpy using the mapped reads,  
502 discordant paired-end reads and split reads as inputs to detect CNVs. CNV calling used DELLY v0.7.6 by  
503 default parameters, and then we used bcftools v1.9 to convert bcf file outputted from DELLY into vcf file  
504 (Danecek, et al. 2011; Danecek, et al. 2021). Furthermore, we also ran Manta v1.6 and Wham v1.8 using  
505 default parameters. For each individual, we merged the CNV sets obtained with these four tools using  
506 SURVIVOR v1.0.7 (Jeffares, et al. 2017). We set the minimum CNV length as 50bp, the maximum CNV  
507 length as 1Mb, and CNV types and DNA strands must match. We retained CNVs that were called by at  
508 least two of the four tools.

509 We finally used the merged CNV set to SVTyper v0.7.0 to call genotypes (Chiang, et al. 2015).  
510 SVTyper performs breakpoint genotyping of structural variants using whole genome sequencing data. It  
511 assesses discordant and concordant reads from paired-end and split-read alignments to infer genotypes  
512 at each site. The script included CNV calling, merging and genotyping can be found on our Gitlab  
513 repository: [https://gitlab.lrz.de/population\\_genetics/s\\_chilense\\_cnv/-](https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-/blob/main/pipeline_of_CNV_calling_genotyping)  
514 [/blob/main/pipeline\\_of\\_CNV\\_calling\\_genotyping](https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-/blob/main/pipeline_of_CNV_calling_genotyping).

515 To assess the sensitivity and accuracy of our pipeline of CNV calling, we simulated short-read data  
516 using a python script 'CNV-Sim' obtained from <https://github.com/NabaviLab/CNV-Sim>. It extends the  
517 functionality of existing read simulators to introduce CNVs in the generated reads. We ran CNV-Sim v0.9.2

in the whole genome, which utilizes the functionality of ART (Huang, et al. 2012) to introduce CNVs in the genome. We simulated 1,000 duplication and 1,000 deletion regions ranging from 50 bp to 1 Mb based on 150 bp paired-reads. We then used our same pipeline to call CNVs based on this simulated short-read dataset (Table S5). The command lines for simulations can be found on: [https://gitlab.lrz.de/population\\_genetics/s\\_chilense\\_cnv/-/blob/main/CNVs\\_simulation](https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-/blob/main/CNVs_simulation).

### Population structure analysis

We inferred the population structure using the whole-genome SNPs and genotyped CNVs from SVTyper, respectively. We performed the principal component analysis (PCA) to seek a summary of the clustering pattern among sampled genomes using GCTA v1.91.4 (Yang, et al. 2011). We first converted vcf format to plink format using VCFtools v1.17 (Danecek, et al. 2011), then converted plink format to a binary format using PLINK v1.9 (Purcell, et al. 2007) with parameters '--noweb --make-bed' to generate input of GCTA. We next performed the analysis of admixture using the program ADMIXTURE v1.3.0 (Alexander, et al. 2009). We assessed six scenarios (ranging from  $K = 2$  to  $K = 7$ ) for genetic clustering using the same input as the PCA analysis.

### Quantification of gene copy number

We employed two strategies to quantify gene copy number (CN). First, we used the read-depth based method implemented in Control-FREEC v11.6 to estimate the CN in 1 kb sliding windows across the entire genome (Boeva, et al. 2012). We used the following parameters in Control-FREEC: ploidy=2, breakPointThreshold = 0.8, degree=3, minExpectedGC = 0.3, maxExpectedGC = 0.55, and telocentromeric=0. We then obtained gene CN from the Control-FREEC outputs and gene coordinates in the genome. However, some genes had more than one CN estimate. These events may be due to imperfect estimation of breakpoints using our window size. So, we calculated the average CN if one gene corresponds to multiple CN values.

We also employed an alternative strategy to calculate gene CN. We first extracted read depth using Mosdepth v0.3.2 (Pedersen and Quinlan 2018) in 1 kb sliding windows from BAM files, and then we calculated the read depth for each gene from gene coordinates. We used median read-depth values of all windows and genes as a normalizing factor to obtain the final window and gene CN estimate, respectively,



545 and the formula reads as:  $CN = (\text{read depth} / \text{median value}) \times 2$ . A factor of 2 stands for the species  
 546 diploidy (Rinker, et al. 2019).

#### 547 **Identification of CNV candidate genes associated with the population differentiation**

548 We calculated  $V_{ST}$  to estimate the population differentiation. The  $V_{ST}$  measurement, analogous to  $F_{ST}$ , is  
 549 applied to identify loci that differentiate by CN between populations (Redon, et al. 2006). Both  $V_{ST}$  and  $F_{ST}$   
 550 consider how genetic variation acts on the differentiation of populations or closely related species and  
 551 range from 0 (no differentiation) to 1 (complete differentiation). We first calculated pairwise  $F_{ST}$  and  $V_{ST}$  to  
 552 compare the efficacy to measure population differentiation by SNPs and CNVs using sliding window-  
 553 based approach. We calculated the  $F_{ST}$  statistics for each pair of populations using VCFtools over 1 kb  
 554 sliding windows. We calculated the  $V_{ST}$  statistics for each pair of populations using the method from Redon,  
 555 et al. (2006). We first calculated  $V_{ST}$  based on CN of 1kb sliding window across the reference genome. In  
 556 addition, we calculated two  $V_{ST}$  data sets based on the two different CN estimation strategies for each  
 557 pair of populations.

558 After assessing the strength of the effect of CN changes on population differentiation, we also  
 559 identified CN candidate genes related to population differentiation. Similar to the sliding window-based  
 560 method, we then independently calculated  $V_{ST}$  value of each gene based on the gene CN. An R script  
 561 with the pipeline of  $V_{ST}$  calculation and identification of candidate genes is found on:  
 562 [https://gitlab.lrz.de/population\\_genetics/s\\_chilense\\_cnv/-/blob/main/VST.R](https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-/blob/main/VST.R). We performed permutation  
 563 tests on the CN counts to identify which genes displayed the greatest degree of observed inter-population  
 564 CN differentiation while controlling for sampling bias. Here, we randomly permuted gene CN of each gene  
 565 for 35 individuals and calculated a new  $V_{ST}$  for every permutation and every gene, respectively. We  
 566 repeated 1,000 times the permutations to generate a random distribution of  $V_{ST}$  values for each gene. We  
 567 then selected candidate genes for which the observed  $V_{ST}$  fell above the 95<sup>th</sup> and 99<sup>th</sup> percentile of the  
 568 permuted  $V_{ST}$  distribution. These candidate genes displayed strong intra-population CN homogeneity and  
 569 high degrees of inter-population differentiation. Finally, genes were considered significant when observed  
 570  $V_{ST}$  values were above the maximum 95% (differentiated) or 99% (strongly differentiated) confidence  
 571 interval cutoff in both gene CN estimation methods (the  $V_{ST}$  cutoff see Table S5).

#### 572 **Gene ontology (GO) analysis**



We first performed BLAST (Camacho, et al. 2009) of our CN differentiated genes to the *A. thaliana* dataset TAIR10 (e-value cutoff was  $10^{-6}$ ). We selected the best matching entry (lowest e-value) as the target homologue for enrichment analysis. We performed GO enrichment analysis using *A. thaliana* annotation database as the background using the R package clusterProfiler (Yu, et al. 2012). We applied the Benjamini-Hochberg method as a false discovery rate (FDR) method 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

To gain insight into how CN of the CN differentiated genes vary across populations, we performed an analysis of gene CN expansion and contraction with the set of 3,359 differentiated genes. We first calculated the mean CN for each gene for each population. We then constructed a population-based phylogenetic tree using SNPs by TreeMix v1.13 (Pickrell and Pritchard 2012), and then the ultrametric tree (Figure 4A) was generated based on 'force.ultrametric' function of phytools R package (Revell 2012). Finally, we performed analysis of gene CN expansion and contraction in different groups using CAFE v4.2.1 (Han, et al. 2013). We first ran CAFE for genes with CN less than 100 to calculate an accurate *lambda* value ( $\lambda=0.00207$  in this study), because genes with large CN can lead to non-informative parameter estimates. We then ran CAFE for genes with CN larger than 100 using the same *lambda* value calculated from genes with CN less than 100. CAFE uses the Monte-Carlo re-sampling procedure to compute the *P* value for each gene to assess significant expansion or contraction along a specific branch. We set the threshold of *P* value smaller than 0.05 to extract genes with a significantly greater rate of evolution (expansion or contraction) in different groups/populations. The command lines of expansion and contraction analysis can be obtained on: [https://gitlab.lrz.de/population\\_genetics/s\\_chilense\\_cnv/-/blob/main/run\\_cafe.sh](https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-/blob/main/run_cafe.sh).

### Association analysis between gene copy number and climatic variables

We obtained the environmental data, including 37 climatic variables, from two public databases, WorldClim2 (Fick and Hijmans 2017) and ENVIREM (Title and Bemmels 2018) (Dataset S6). To evaluate the relative contribution of the abiotic environment to explaining patterns of genetic variation, we first performed a redundancy analysis (RDA) to associate CN of 3,539 differentiated genes with climatic variables. We performed RDA analysis using the *rda* function from the vegan package as implemented in

R (Forester, et al. 2018), modelling CN as a function of predictor variables and producing constrained axes and representative predictors (climatic variables). We assessed the multi-collinearity between representative predictors (climatic variables) using the variance inflation factor (VIF) and excluded all climatic variables that showed VIF > 10. We then calculated the significance of RDA ordination axes using the *anova.cca* function ( $P < 0.001$ ). The R script of RDA analysis, including all steps and parameters, can be obtained at [https://gitlab.lrz.de/population\\_genetics/s\\_chilense\\_cnv/-/blob/main/RDA.R](https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-/blob/main/RDA.R).

We identified six climatic variables correlated with the dynamics of gene CN across populations from RDA (Fig. 5A). To identify candidate genes associated with each climate variable, we used LFMM2 (latent factor mixed models) to build a model between each gene and climatic variable based on the univariate test (Caye, et al. 2019). We first performed *lfmm\_ridge* function implemented in the R library LFMM to obtain an object that contains the latent variable score matrix using a K value of four latent factors (as evaluated from analysis of population structure) based on the CN of 3,539 differentiated genes and six representative climate variables (as obtained from RDA), respectively. Then, we performed association testing using the *lfmm\_test* function. We finally used the method of Benjamini-Hochberg to calibrate  $P$  value and set 0.01 as the significance threshold to obtain candidate genes associated with the climatic variables. The R script of LFMM we used is available on our Gitlab repository [https://gitlab.lrz.de/population\\_genetics/s\\_chilense\\_cnv/-/blob/main/lfmm.R](https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-/blob/main/lfmm.R).

## Supplementary material

Supplementary data are available online at Molecular Biology and Evolution.

## Data Availability

Raw sequence data are available at the European Nucleotide Archive (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](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

627 Munich. KW acknowledges funding from Natural Science Foundation of Xinjiang Uygur Autonomous  
628 Region Grant Number: 2024D01C216. AT acknowledges funding from DFG  
629 (Deutsche Forschungsgemeinschaft) Grant Number: 317616126 (TE809/7-1). We thank the Tomato  
630 Genetics Resource Center (TGRC) of the University of California Davis for generously providing us with  
631 the seeds of the population included in this study, and the Greenhouses & Phytochambers Unit of the  
632 TUM Plant Technology Center in Dürnast for plant care.

633

#### 634 **Competing interests**

635 The authors have no conflicts of interest to declare.

#### 636 **Author contributions**

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

640

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

641     **References**

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

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

645     Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D.  
646     2020. Major impacts of widespread structural variation on gene expression and crop improvement in  
647     tomato. *Cell* 182:145-161. e123.

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

651     Antinucci M, Comas D, Calafell F. 2023. Population history modulates the fitness effects of Copy  
652     Number Variation in the Roma. *Human Genetics* 142:1327-1343.

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

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

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

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

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

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

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

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

675     Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+:  
676     architecture and applications. *BMC bioinformatics* 10:421.

677     Castagnone-Sereno P, Mulet K, Danchin EG, Koutsovoulos GD, Karaulic M, Da Rocha M, Bailly-Bechet  
678     M, Pratz L, Perfus-Barbeoch L, Abad P. 2019. Gene copy number variations as signatures of adaptive  
679     evolution in the parthenogenetic, plant-parasitic nematode *Meloidogyne incognita*. *Molecular Ecology*  
680     28:2559-2572.

681     Catola S, Castagna A, Santin M, Calvenzani V, Petroni K, Mazzucato A, Ranieri A. 2017. The dominant  
682     allele Aft induces a shift from flavonol to anthocyanin production in response to UV-B radiation in  
683     tomato fruit. *Planta* 246:263-275.

- 684 Caye K, Jumentier B, Lepeule J, François O. 2019. LFMM 2: Fast and Accurate Inference of Gene-  
685 Environment Associations in Genome-Wide Studies. *Molecular biology and evolution* 36:852-860.
- 686 Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT.  
687 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing  
688 applications. *Bioinformatics* 32:1220-1222.
- 689 Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015.  
690 SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods* 12:966-968.
- 691 Coutelier M, Holtgrewe M, Jäger M, Flöttman R, Mensah MA, Spielmann M, Krawitz P, Horn D, Beule D,  
692 Mundlos S. 2022. Combining callers improves the detection of copy number variants from whole-  
693 genome sequencing. *European Journal of Human Genetics* 30:178-186.
- 694 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT,  
695 Sherry ST. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156-2158.
- 696 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA,  
697 Davies RM. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10:giab008.
- 698 Daniell H, Lin C-S, Yu M, Chang W-J. 2016. Chloroplast genomes: diversity, evolution, and applications  
699 in genetic engineering. *Genome biology* 17:134.
- 700 De Mita S, Thuillet A-C, Gay L, Ahmadi N, Manel S, Ronfort J, Vigouroux Y. 2013. Detecting selection  
701 along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and  
702 selfing populations. *Molecular Ecology* 22:1383-1399.
- 703 DeBolt S. 2010. Copy Number Variation Shapes Genome Diversity in Arabidopsis Over Immediate Family  
704 Generational Scales. *Genome biology and evolution* 2:441-453.
- 705 Fang H, Dong Y, Yue X, Hu J, Jiang S, Xu H, Wang Y, Su M, Zhang J, Zhang Z, et al. 2019. The B-box zinc  
706 finger protein MdBBX20 integrates anthocyanin accumulation in response to ultraviolet radiation and  
707 low temperature. *Plant, cell & environment* 42:2090-2104.
- 708 Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nature Reviews*  
709 *Genetics* 7:85-97.
- 710 Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land  
711 areas. *International journal of climatology* 37:4302-4315.
- 712 Fischer I, Camus-Kulandaivelu L, Allal F, Stephan W. 2011. Adaptation to drought in two wild tomato  
713 species: the evolution of the Asr gene family. *New Phytologist* 190:1032-1044.
- 714 Forester BR, Lasky JR, Wagner HH, Urban DL. 2018. Comparing methods for detecting multilocus  
715 adaptation with multivariate genotype–environment associations. *Molecular Ecology* 27:2215-2233.
- 716 Frichot E, Schoville SD, Bouchard G, François O. 2013. Testing for Associations between Loci and  
717 Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution* 30:1687-  
718 1699.
- 719 Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, Wing RA, McNally KL,  
720 Tatarinova T, Grigoriev A. 2019. Structural variants in 3000 rice genomes. *Genome Research* 29:870-  
721 880.
- 722 Gaudinier A, Blackman BK. 2020. Evolutionary processes from the perspective of flowering time  
723 diversity. *New Phytologist* 225:1883-1898.
- 724 Guo M, Yang F, Liu C, Zou J, Qi Z, Fotopoulos V, Lu G, Yu J, Zhou J. 2022. A single - nucleotide  
725 polymorphism in WRKY33 promoter is associated with the cold sensitivity in cultivated tomato. *New*  
726 *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, Verstrepren KJ, Jelier R. 2020. Gene Loss Predictably Drives Evolutionary Adaptation. *Molecular Biology and Evolution* 37:2989-3002.
- Hoballah ME, Gübitz T, Stuurman J, Broger L, Barone M, Mandel T, Dell'Olivo A, Arnold M, Kuhlmeier C. 2007. Single gene-mediated shift in pollinator attraction in *Petunia*. *Plant Cell* 19:779-790.
- 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.
- Kaur S, Tiwari V, Kumari A, Chaudhary E, Sharma A, Ali U, Garg M. 2023. Protective and defensive role of anthocyanins under plant abiotic and biotic stresses: An emerging application in sustainable agriculture. *Journal of Biotechnology* 361:12-29.
- 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:R84.
- 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* 55:852-860.
- 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. Absciscic 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.
- Marszałek-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 JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8:e1002967.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* 81:559-575.
- Putterill J, Varkonyi-Gasic E. 2016. FT and florigen long-distance flowering control in plants. *Current Opinion in Plant Biology* 33:77-82.



1  
2  
3 813 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  
4 814 of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184:3542-3558.e3516.  
5  
6 815 Raduski AR, Igić B. 2021. Biosystematic studies on the status of *Solanum chilense*. *American Journal of*  
7 816 *Botany* 108:520-537.  
8  
9 817 Ranjan A, Ichihashi Y, Sinha NR. 2012. The tomato genome: implications for plant breeding, genomics  
10 818 and evolution. *Genome biology* 13:167.  
11  
12 819 Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery  
13 820 by integrated paired-end and split-read analysis. *Bioinformatics* 28:i333-i339.  
14  
15 821 Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen  
16 822 W. 2006. Global variation in copy number in the human genome. *Nature* 444:444-454.  
17  
18 823 Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things).  
19 824 *Methods in Ecology and Evolution* 3:217-223.  
20  
21 825 Rinker DC, Specian NK, Zhao S, Gibbons JG. 2019. Polar bear evolution is marked by rapid changes in  
22 826 gene copy number in response to dietary shift. *Proceedings of the National Academy of Sciences*  
23 827 116:13446-13451.  
24  
25 828 Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki  
26 829 K. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485:635-  
27 830 641.  
28  
29 831 Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, O'Hara R, Casalunovo T, Conlin LK, D'arcy M.  
30 832 2009. High-resolution mapping and analysis of copy number variations in the human genome: a data  
31 833 resource for clinical and research applications. *Genome Research* 19:1682-1690.  
32  
33 834 Shimizu KK, Shimizu-Inatsugi R, Tsuchimatsu T, Purugganan MD. 2008. Independent origins of self-  
34 835 compatibility in *Arabidopsis thaliana*. *Mol Ecol* 17:704-714.  
35  
36 836 Silva-Arias GA, Gagnon E, Hembrom S, Fastner A, Khan MR, Stam R, Tellier A. 2024. Patterns of  
37 837 presence-absence variation of NLRs across populations of *Solanum chilense* are clade-dependent and  
38 838 mainly shaped by past demographic history. *bioRxiv:2023.2010.2013.562278*.  
39  
40 839 Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H. 2009. Maize  
41 840 inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in  
42 841 genome content. *PLoS Genetics* 5:e1000734.  
43  
44 842 Srikanth A, Schmid M. 2011. Regulation of flowering time: all roads lead to Rome. *Cellular and*  
45 843 *Molecular Life Sciences* 68:2013-2037.  
46  
47 844 Stam R, Nosenko T, Hörger AC, Stephan W, Seidel M, Kuhn JM, Haberer G, Tellier A. 2019a. The de novo  
48 845 reference genome and transcriptome assemblies of the wild tomato species *Solanum chilense*  
49 846 highlights birth and death of NLR genes between tomato species. *G3: Genes, Genomes, Genetics*  
50 847 9:3933-3941.  
51  
52 848 Stam R, Silva-Arias GA, Tellier A. 2019b. Subsets of NLR genes show differential signatures of adaptation  
53 849 during colonization of new habitats. *New Phytologist* 224:367-379.  
54  
55 850 Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt  
56 851 S, Bamshad M. 2015. Global diversity, population stratification, and selection of human copy-number  
57 852 variation. *Science* 349:aab3761.  
58  
59 853 Title PO, Bemmels JB. 2018. ENVIREM: an expanded set of bioclimatic and topographic variables  
60 854 increases flexibility and improves performance of ecological niche modeling. *Ecography* 41:291-307.  
855 855 Wei K, Sharifova S, Zhao X, Sinha N, Nakayama H, Tellier A, Silva-Arias GA. 2023a. Evolution of two gene

1  
2  
3 856 networks underlying adaptation to drought stress in the wild tomato *Solanum chilense*.  
4 857 bioRxiv:2023.2001. 2018.524537.  
5  
6 858 Wei K, Silva-Arias GA, Tellier A. 2023b. Selective sweeps linked to the colonization of novel habitats and  
7 859 climatic changes in a wild tomato species. *New Phytologist* 237:1908-1921.  
8  
9 860 Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence  
10 861 alignment/map (SAM) format and SAMtools. *Bioinformatics* 25:2078-2079.  
11  
12 862 Xia HUI, Camus-Kulandaivelu L, Stephan W, Tellier A, Zhang Z. 2010. Nucleotide diversity patterns of  
13 863 local adaptation at drought-related candidate genes in wild tomatoes. *Molecular Ecology* 19:4144-4154.  
14  
15 864 Xiao S, Jiang L, Wang C, Ow DW. 2021. Arabidopsis OXS3 family proteins repress ABA signaling through  
16 865 interactions with AFP1 in the regulation of ABI4 expression. *Journal of Experimental Botany* 72:5721-  
17 866 5734.  
18  
19 867 Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis.  
20 868 *The American Journal of Human Genetics* 88:76-82.  
21  
22 869 Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: an R package for comparing biological themes  
23 870 among gene clusters. *Omics: a journal of integrative biology* 16:284-287.  
24  
25 871 Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, Cantu D, Gaut BS. 2019. The population  
26 872 genetics of structural variants in grapevine domestication. *Nature Plants* 5:965-979.  
27  
28 873 Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, Wu Y, Cheng L, Fang Y, Wu K. 2022. Graph pangenome captures  
29 874 missing heritability and empowers tomato breeding. *Nature* 606:527-534.  
30  
31 875 Zmienko A, Marszałek-Zenczak M, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozłowski P,  
32 876 Karłowski WM, Figlerowicz M. 2020. AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis  
33 877 Genome[OPEN]. *The Plant Cell* 32:1797-1819.  
34  
35 878 Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M. 2014. Copy number polymorphism in plant genomes.  
36 879 *Theoretical and applied genetics* 127:1-18.  
37  
38 880 Zufall RA, Rausher MD. 2004. Genetic changes associated with floral adaptation restrict future  
39 881 evolutionary potential. *Nature* 428:847-850.  
40 882  
41  
42 883  
43  
44 884  
45  
46 885  
47  
48 886  
49  
50 887  
51  
52 888  
53  
54 889  
55  
56 890  
57  
58 891  
59  
60 892  
893

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

894

895 **Table1.** The summary of gene expansion and contraction in different groups/populations based on an  
896 ultrametric tree.

| <sup>a</sup> Groups / Populations | Number of CN expanded genes | Number of CN contracted genes | Number of CN gained | Number of CN lost | <sup>b</sup> Rate of average expansion / contraction | <sup>c</sup> 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)                                  |

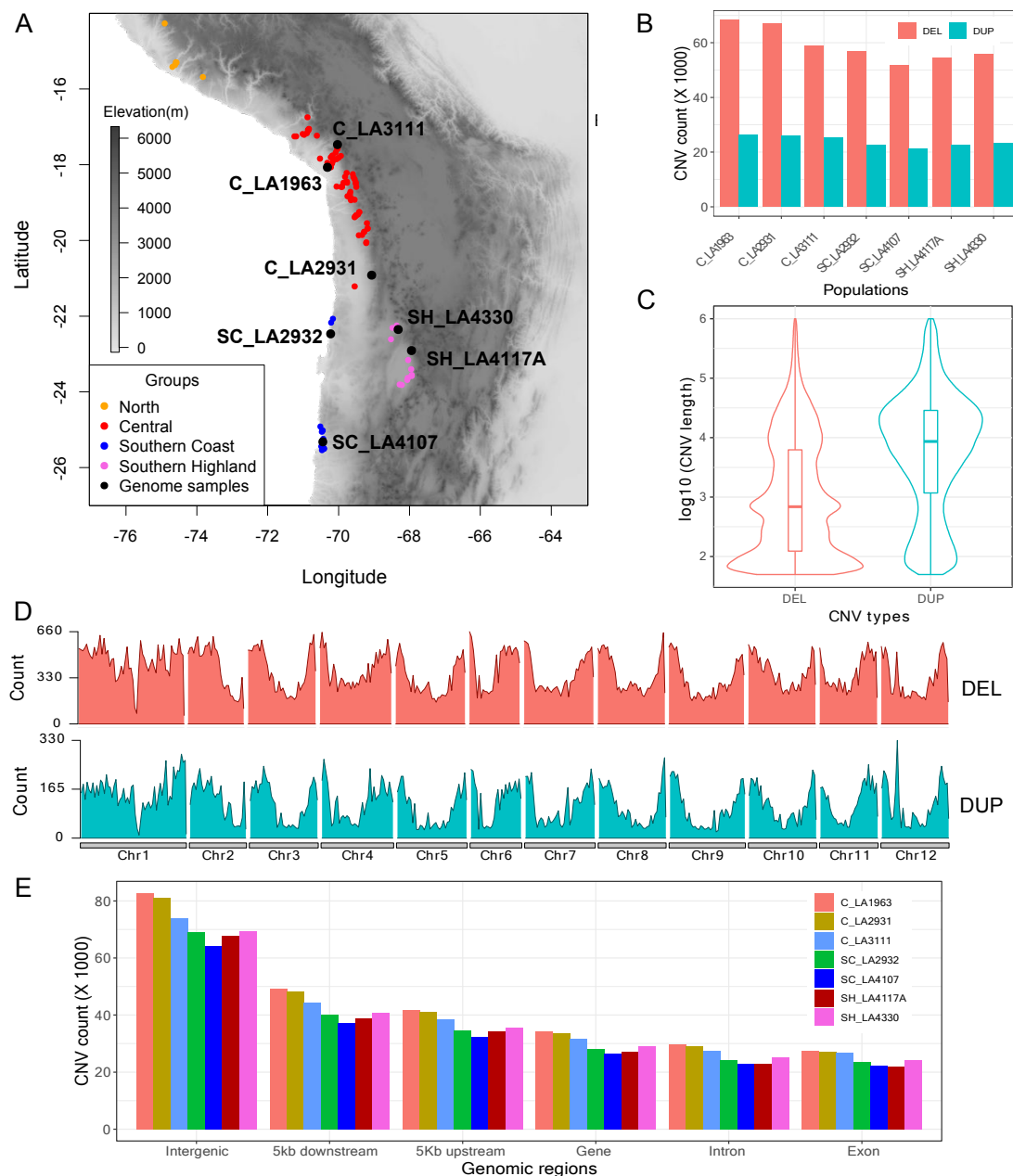
897 The table shows that the expansion and contraction of CN-differentiated genes in different groups /  
898 populations based on an ultrametric tree (Fig. 4A). C: central; SH: southern highland; SC: southern coast.

899 <sup>a</sup>Groups and populations denote the branches in the ultrametric tree (Fig. 4A).

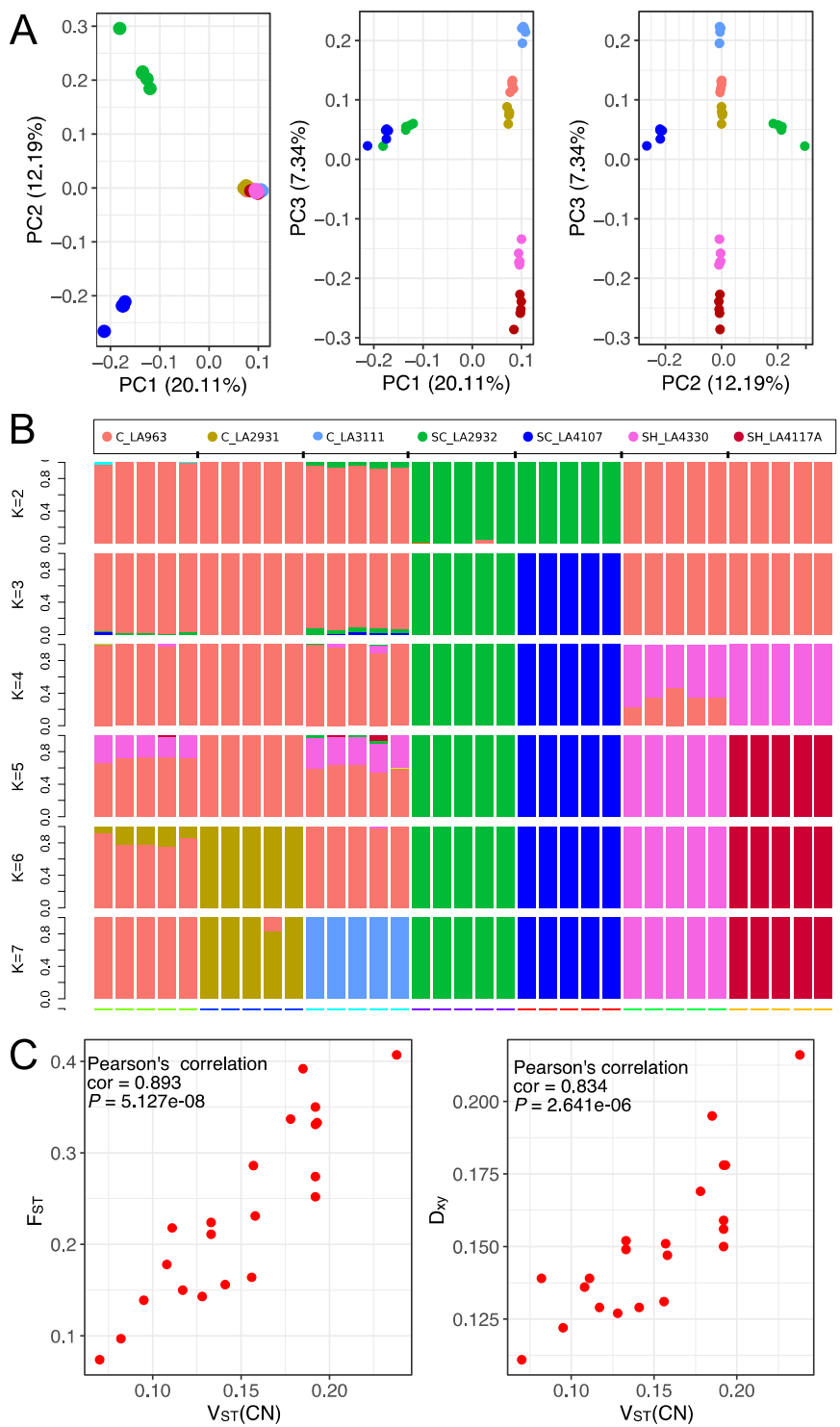
900 <sup>b</sup>Rate of average expansion / contraction = (Number of CN gained - Number of CN lost) / (Number of CN  
901 expanded genes + Number of CN contracted genes). Positive values indicate CN expansion and negative  
902 values indicate CN contraction.

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

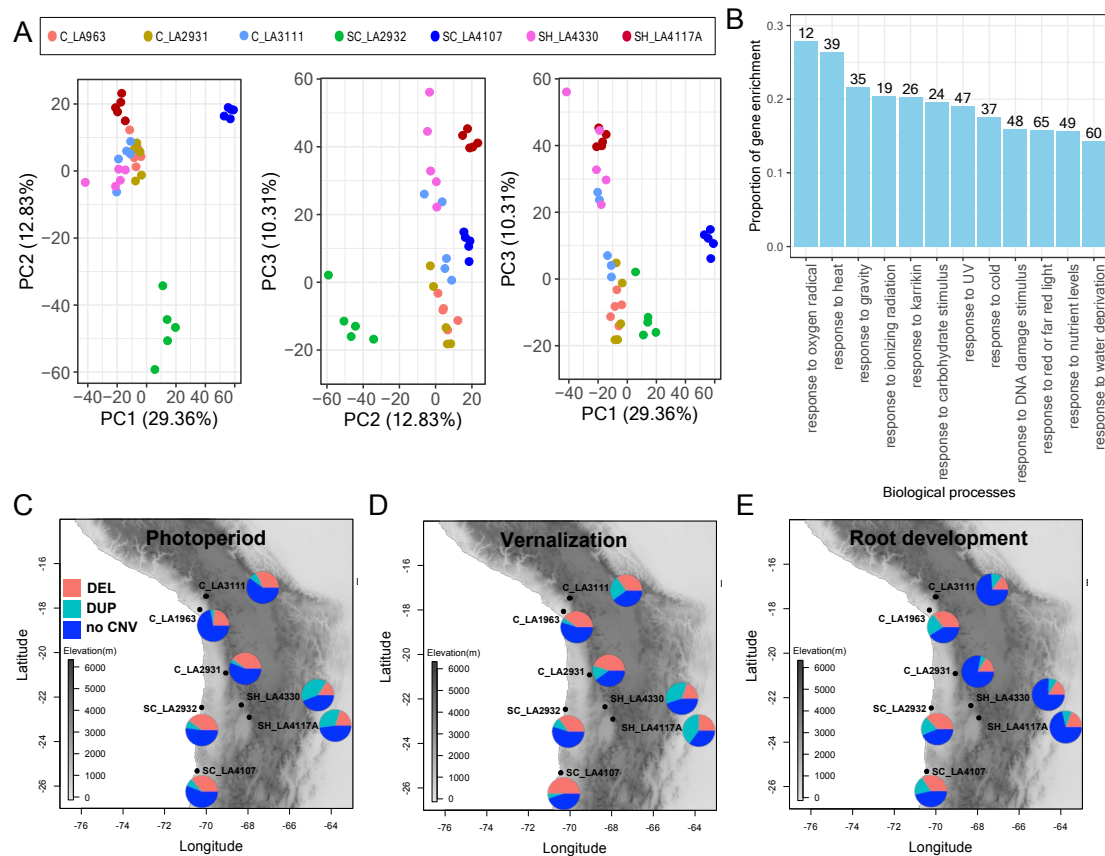
907



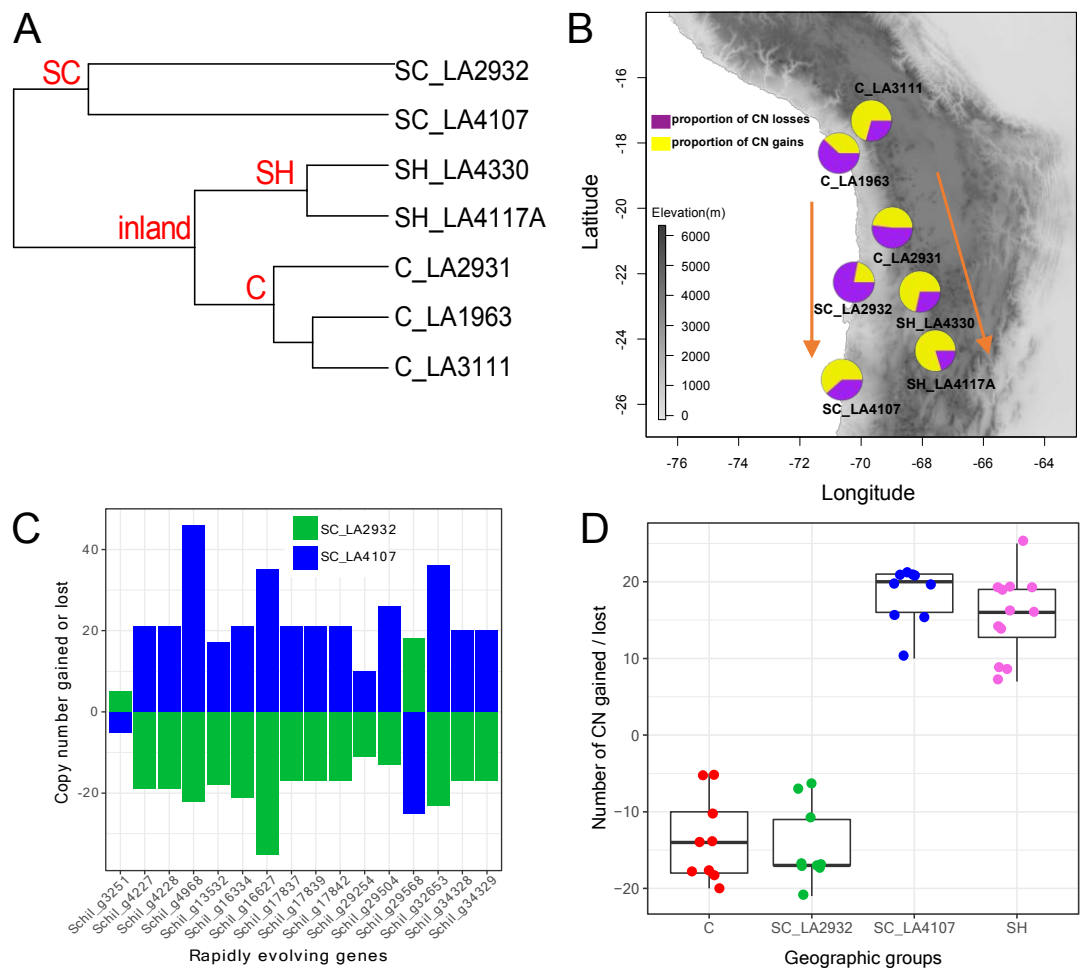
**Fig. 1.** The summary of the revealed CNVs in the genome of *S. chilense*. (A) The map with the distribution of all *S. chilense* populations at the Tomato Genetics Resource Center (TGRC), the *S. chilense* populations in this study (black circles), and the four population groups (circles with other colours). C: central; SH: southern highland; SC: southern coast. (B) The number of CNVs pooled across five individuals within each population. DEL: deletion; DUP: duplication. (C) The distribution of CNV size. (D) The CNV density along the genome is expressed as a count per 1Mb window. (E) The number of CNVs overlapping various genomic features for each population.



917  
918 **Fig. 2.** Population structure and differentiation analyses based on the genotyped CNVs. (A) Principal  
919 component analysis (PCA) based on the genotyped CNVs from 35 individuals from seven *S. chilense*  
920 populations. (B) Structure analysis based on genotyped CNVs and assuming  $K = 2 - 7$  subgroups (optimal  
921  $K$  value as determined by cross-validation error; Fig. S3B). C: central; SH: southern highland; SC:  
922 southern coast. (C) The correlation between  $F_{ST}/D_{xy}$  and  $V_{ST}$  indicates that CNVs support the known  
923 population differentiation.  
924

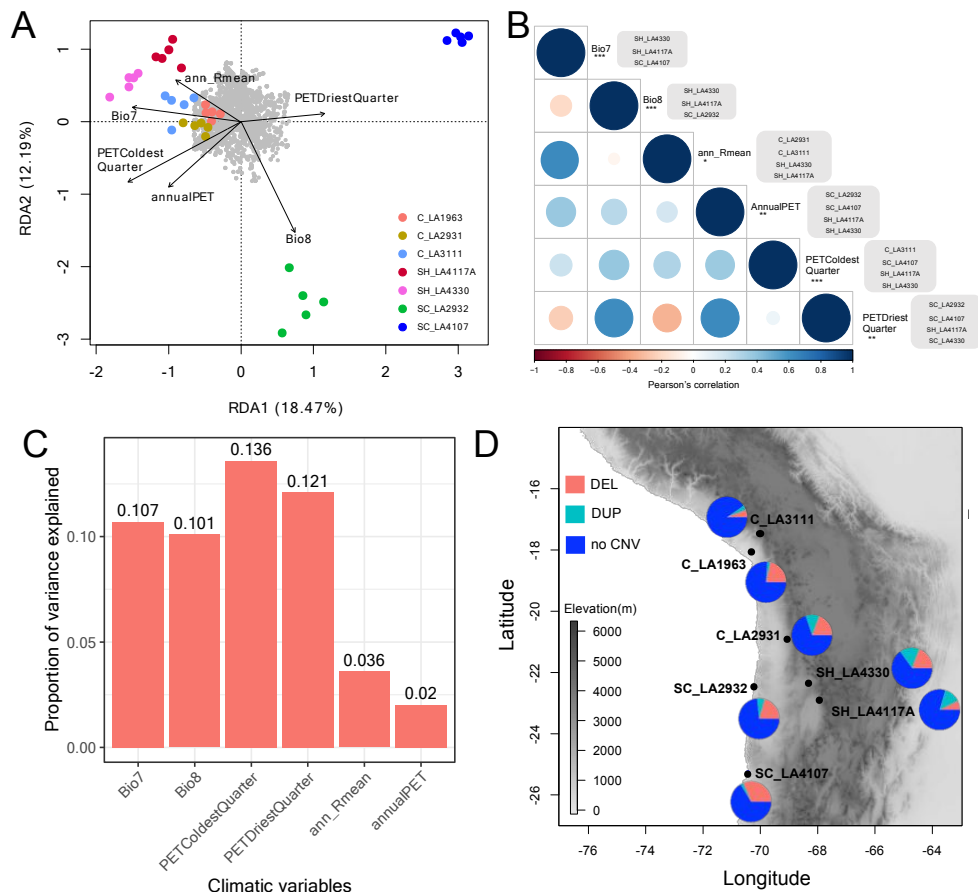


**Fig. 3.** Genes with differentiated copy number (CN) profiles among seven populations are linked to response to multiple environmental stimuli. (A) The PCA based on the copy number (CN) of 3,539 differentiated genes. C: central; SH: southern highland; SC: southern coast. (B) The proportions of CN-differentiated 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. (C-E) CN-differentiated genes involved in the photoperiod pathway to regulate flowering time (C), the vernalization pathway to regulate flowering time (D), and the root developmental process (E). The pie charts denote the proportions of CN-differentiated genes with deletion (DEL), duplication (DUP) or absence of CNV (see also Table S6).



**Fig. 4.** The expansion and contraction of CN-differentiated genes in different populations relative to the *S. chilense* reference genome. (A) The ultrametric phylogenetic 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 gene copy number (CN) loss and gain in the processes of two southward colonization events, first to the southern coast and second to the southern highland (orange arrows). Proportion of CN losses/gains = the number of CN losses or gains / (the number of CN losses + the number of CN gains). (C) The number of CN gains (positive values) or losses (negative values) for 16 rapidly evolving genes in two southern coast populations. (D) The number of CN gains and CN losses for rapidly evolving genes related to photosynthesis in different subgroups representing four different habitats (see also Table S8).





**Fig. 5. Genome-Environment Association (GEA) analysis between the gene copy number (CN) and the climatic data of the different habitats.** (A) Redundancy analysis (RDA) ordination biplot illustrates the association between the climatic variables (Dataset S7), individuals, and 3,539 differentiated gene CN. In the RDA, arrows indicate the direction of the climatic variables associated with the different populations, and the projection of arrows onto an ordination axis shows the correlation with that 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 variance explained by six overrepresented climate variable 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 S9).