



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

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

SCHOLARONE™
Manuscripts

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

Kai Wei^{1*}, Remco Stam², Aurélien Tellier^{1*}, Gustavo A Silva-Arias^{1,3*}

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

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

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

*Corresponding authors: Aurélien Tellier: aurelien.tellier@tum.de

Kai Wei: kai.wei@tum.de

Gustavo A. Silva-Arias: gasilvaa@unal.edu.co

1 Abstract

2 Copy Number Variations (CNVs) are genomic structural changes constituting genetic diversity and
3 underpinning rapid ecological adaptation. The timing of, and the target genes involved in adaptation
4 through CNVs in the chamois and wild relative lineage still need to be explored at the population level.

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

flow
found?

1 2 3 19 Introduction 4 5

6 Copy number variation (CNV) is the primary type of structural variation (SV) caused by genomic
7 rearrangement, which mainly includes deletion (DEL) and duplication (DUP) events resulting from the loss
8 and gain of DNA segments (Feuk, et al. 2006; Źmieńko, et al. 2014). It is expected that CNV has a more
9 significant impact on gene function~~Y~~ because it covers more base-pairs (Shaikh, et al. 2009; Hämälä, et al.
10 2021) and has a higher per-locus mutation rate than point mutations (single nucleotide polymorphisms,
11 SNPs) (Lupsik 2007). CNV is recognised as an essential driver of genomic divergence and local
12 adaptation (Rinker, et al. 2019; Hämälä, et al. 2021; Marszalek-Zenczak, et al. 2023). Genome-wide
13 studies confirm the importance of CNVs as the basis of stress response and yield improvement in multiple
14 plants, such as maize (Springer, et al. 2009), rice (Fuentes, et al. 2019; Qin, et al. 2021), and *Arabidopsis*
15 *thaliana* (Zmienko, et al. 2020; Marszalek-Zenczak, et al. 2023). However, such studies have been so far
16 conducted in selfing species and/or crops characterised by small population size and domestication
17 bottlenecks (Alonso-Blanco, et al. 2016; Beissinger, et al. 2016; Brumlop, et al. 2019). Therefore, it is
18 difficult in such species to disentangle the effect of random evolutionary processes (genetic drift,
19 chromosomal rearrangements, and demographic history) generating fast and extensive CNVs between
20 populations from the impact of adaptive processes (here positive selection underpinning environmental
21 adaptation). In addition, the dynamics~~Y~~ of gene copy number also reflects population history and multiple
22 events, including selection, migration and recombination (Sudmant, et al. 2015; Zhou, et al. 2019; Otto,
23 et al. 2022; Antinucci, et al. 2023; Otto and Wiehe 2023). Indeed, the effective population size (N_e) of
24 populations determines the efficiency of positive and negative selection against genetic drift, as well as
25 the amount of genetic diversity (SNPs or CNV) available, thus being a major determinant of the genome
26 architecture (Lynch and Walsh 2007).

27 The tomato wild relative species *Solanum chilense* is proven to be an excellent model species to
28 study the genetic basis of adaptive evolution when colonising novel habitats (Böndel, et al. 2015; Stam,
29 et al. 2019b; Wei, et al. 2023b). The presence of outcrossing, gene flow, seed banks, and relatively mild
30 bottlenecks during the colonisation of new habitats results in high effective population sizes (N_e) as
31 reflected by high nucleotide diversity and high recombination rates, meaning that this species has a high
32 adaptive potential (Arunyawat, et al. 2007; Stam, et al. 2019b; Wei, et al. 2023b). *S. chilense* occurs in
33 southern Peru and northern Chile, from mesic to very arid habitats around the Atacama Desert, and
34 becomes the southern-most distributed species in the tomato clade (Nakazato, et al. 2010). Moreover,

1
2
3
4 49 within *S. chilense*, two groups of populations expanded southward during two independent colonisation
5 50 events (Böndel, et al. 2015; Stam, et al. 2019b; Wei, et al. 2023b): one towards the coastal part of northern
6 Chile (hereafter the southern coast group), and the other towards the high altitudes of the Chilean Andes
7 (southern highland group) (Fig. 1A). The populations currently occurring in the southern coast and
8 southern highland habitats have been shown to exhibit signatures of past positive selection for adaptation
9 to cold, drought, light (photoperiod), heat and biotic stresses (Xia, et al. 2010; Fischer, et al. 2011;
10 Nosenko, et al. 2016; Böndel, et al. 2018; Stam, et al. 2019b; Wei, et al. 2023b). These events of recent
11 positive selection at a cohesive set of genes and drought-responsive gene networks suggest the
12 occurrence of genetic underpinnings to the adaptation of novel habitats during the southward expansion
13 of *S. chilense* populations towards arid areas around the Atacama desert (Wei, et al. 2023a). Previous
14 population genomic studies revealed that these adaptive signatures are based on scans for positive
15 selection solely using single-nucleotide polymorphisms (SNPs). However, whether CNV can also
16 contribute to adaptation to novel habitats in *S. chilense* and the tomato clade is still unknown.

17
18
19
20
21
22
23
24
25
26
27
28
29 62 Reference genomes of several species of the tomato clade, including numerous cultivated tomato
30 varieties, are now sequenced and assembled (Ranjan, et al. 2012; Sato, et al. 2012; Bolger, et al. 2014;
31 Stam, et al. 2019a). Three tomato SV sets have been recently constructed based on a tomato-clade
32 pangenome analysis to investigate the impact of genome rearrangements on gene expression and
33 genomic diversity and provide new genomic resources for the improvement of tomato (Alonge, et al. 2020;
34 Zhou, et al. 2022; Li, et al. 2023). These three studies compare cultivated tomato genomes with that of
35 several wild tomato species, including PacBio and Illumina sequencing from an individual of the *S.*
36 *chilense* accession LA1969 (belonging to our central group; Fig. 1A). Interestingly, we note that *S.*
37 *chilense* exhibits the highest number of SV among all wild and cultivated tomato species (Li, et al. 2023).
38
39 71 This difference is even more striking when considering that the closely related species *S. peruvianum* and
40 *S. corneliomulleri* show half or fewer SVs than *S. chilense*. All these three species exhibit a similar recent
41 proliferation of transposable elements (Li, et al. 2023). As *S. chilense* has one of the largest genome sizes
42 of the tomato clade and has the highest number of annotated genes, it is crucial to study processes driving
43 gene-copy number variation and its relevance for speciation and intraspecific diversification. However, the
44 studies mentioned above focus on the pangenome across species level (wild and cultivated) and an
45 understanding of the role of CNVs in local (ecological) adaptation is still lacking, especially for the
46 adaptation to new arid/southern populations in *S. chilense* habitats.

local adaptation in *S. chil.*
but only studied in SNPs

mainly
arid/southern

1
2
3
4 We generate whole-genome copy number (CN) profiles for 35 *S. chilense* plants from seven
5 populations (five diploid individuals per population) representing three different geographic habitats: three
6 central (C) populations, two southern highland (SH) populations and two southern coast (SC) populations
7 with different habitats (Fig. 1A; Dataset S1). We first identify candidate genes with highly differentiated
8 CN profiles between populations that are likely candidates of recent positive selection. We then measure
9 the evolutionary trend of CN expansion and contraction across different populations. Finally, we associate
10 the dynamics of gene CN with climatic variables to provide evidence for environmental stresses driving
11 CNV dynamics across populations. Our results suggest that gene CNV contributes to population
12 adaptation to novel habitats in an outcrossing species with a large effective population size and genetic
13 diversity. We illustrate the importance of including an analysis of CN variants to complement genomic
14 scans of recent positive selection based on SNPs.

25 Results

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

27
28
29 We identify a total of 212,207 CNVs (160,926 deletions and 51,281 duplications) by aligning each of the
30 35 whole-genome sequencing datasets (Dataset S1) against a chromosome-level *S. chilense* reference
31 genome (Silva-Arias, et al. 2023) using the combination of four CNV callers (Fig. S1; Dataset S2). We find
32 73,014 up to 94,621 CNVs per population (Fig. 1B; Table S1) and 31,923 up to 46,579 CNVs per individual
33 (Fig. S1; Table S2). Although the number of deletions in all individuals and populations is much larger
34 than the number of duplications (Fig. 1B; Fig. S1; Table S1 and 2), the size of duplications is larger (39,140
35 bp +/- 104,577) than ~~V~~deletions and exhibits a skewed distribution (14,052 bp +/- 59,930) (Fig. 1C;
36 Kolmogorov-Smirnov test, $P=2.2e-16$). Deletions are smaller than duplications, as 56% of deletions
37 display a size between 50bp and 1,000bp, against 26% for duplications. We find 37% to 43% of the CNVs
38 to be identified in only one individual for three central populations (Fig. S2; Table S3), while only 12% to
39 14% of all CNVs are ~~V~~observed in all five individuals of a given population (i.e., CNVs being fixed).
40 Furthermore, the number of CNVs is not homogeneously distributed among populations as more than 20%
41 of CNVs are detected in all five individuals in the southern coast and southern highland populations,
42 especially in the two southern coast populations (25% in SC_LA2932 and 31% in SC_LA4107).
43 Deletions and duplications are enriched at both ends of the chromosomes (Fig. 1D), consistent
44 with previous studies (Alonge, et al. 2020; Hämälä, et al. 2021; Li, et al. 2023). Although most CNVs (76%

1
2
3
4 108 to 79% per population) cover intergenic regions (Fig. 1E; Table S4), about 35% of CNVs are located in
5 genes annotated in the *S. chilense* reference. In addition, 45% and 50% of CNVs across populations
6 overlap with putative regulatory elements 5 kb upstream and 5 kb downstream of genes, respectively.
7
8 111 CNVs are typically shaped predominately by transposable elements (Fuentes, et al. 2019; Alonge, et al.
9 2020), and the annotation reveals that 68% of deletions and 82% of duplications match at least one
10 112 transposable element annotated in the *S. chilense* genome.
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

114 To confirm the validity of our pipeline, which assembles CNV detection from four tools specialised
115 for short-read datasets, we simulated 1,000 deletions and 1,000 duplications with lengths ranging from
116 50 bp to 1 Mb based on 150 bp short-reads (see supplementary methods). We subsequently detected
117 approximately 90% of simulated CNVs using our pipeline, and the false-positive rate was much smaller
118 than using a single caller (Table S5). Our results, as well as previous claims, indicate that combined
119 multiple callers improve the detection of CNVs and are robust to short-read data (Kosugi, et al. 2019;
120 Mahmoud, et al. 2019; Coutelier, et al. 2022).

121 **CNVs effectively capture the population differentiation**

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

131 We further explore the differentiation of populations using the V_{ST} statistic, which is analogous to
132 the classically used F_{ST} for SNP data (Redon, et al. 2006). We first compute the V_{ST} values along the
133 whole genome in 10 kb windows of 1 kb step size using two CN quantitative measurements: Control-
134 FREEC ($V_{ST}(CN)$) and read depth ($V_{ST}(RD)$) (Table S6). We find a significantly high correlation between
135 these two measures (Pearson's test, $P=1.06e-07$; Fig. S4A). Based on the V_{ST} values, we find similar
136 structure patterns as in previous studies using SNPs (Böndel, et al. 2015; Stam, et al. 2019b; Raduski

1
2
3
4 137 and Igić 2021; Wei, et al. 2023b), namely the high differentiation between southern coast and inland
5 populations, especially between southern coast and southern highland populations (Table S6). As
6 expected, both V_{ST} estimates ($V_{ST}(CN)$ and $V_{ST}(RD)$) show a significantly high correlation with F_{ST} (based
7 on SNPs) (Fig. 2C; Fig. S4B; Table S6).

8
9
10
11 141 **Differentiation of gene CN profiles in different populations**

12
13
14 142 To explore the role of natural selection in shaping CNV frequencies and distribution across populations,
15 we use both V_{ST} measures ($V_{ST}(CN)$ and $V_{ST}(RD)$) across the 39,245 genes to capture candidate genes
16 under divergent selective pressures by identifying genes with strong CN differentiation across populations
17
18 144 (Fig. S5). We perform a permutation test (1,000 times) for each gene using the 35 samples of all seven
19 populations. The candidate genes are identified by considering those that surpass a high differentiation
20 threshold for both V_{ST} measures. The rationale is that high V_{ST} values indicate strong differentiation and
21 possibly adaptive divergence at some CNVs between populations. In total, we obtain 3,539 candidate
22 genes that present CN differentiation across the seven populations (i.e., V_{ST} greater than the maximum
23
24 149 95th percentile of the 1,000 permuted V_{ST} values; Fig. S5; Table S7; Dataset S3) and 2,192 strongly CN-
25 differentiated genes of these belong to the top 99th percentile of the 1,000 permuted V_{ST} values (Fig. S5;
26
27 151 Table S7; Dataset S3). In Fig. S6A, we show the distribution of deletions and duplications for these 3,539
28 candidate genes. Southern highland populations exhibit a pronounced increase in gene gains
29 (duplications) and a minimal reduction in gene loss (deletions), whereas SC populations show a
30
31 154 comparatively higher incidence of gene loss.

32
33 156 We perform four PCA analyses based on the Control-FREEC-based CN values of 1) all annotated
34 157 23,911 genes with CN values (Fig. S6B); 2) the 12,392 genes with $V_{ST}(CN)>0$ (Fig. S6C); 3) the 3,539
35
36 158 differentiated gene set (Fig. 3A); and 4) the 2,192 strongly differentiated gene set (Fig. S6D). In the PCA
37 based on 23,911 genes with CN values (Fig. S6B), all samples exhibit a cohesive grouping, except
38 159 SC_LA4107. The southern coast populations separate from the five inland populations (central and
39 southern highland populations) in the second PCA (with $V_{ST}(CN)>0$; Fig. S6C). This suggests a large
40
41 162 difference in the CN range and composition between southern coast and inland populations. Consistent
42 with the PCA based on the genotyped CNVs (and previously on SNP data), PC3 separates the southern
43
44 163 highland populations from the central populations when using the differentiated genes CN values (Fig.
45
46 164 3A; Fig. S6D). Note, however, that southern highland populations still show ca. 20% of admixed ancestry
47 coefficients with the central populations (Fig. 2B). These admixture signatures can be interpreted as either

1
2
3 167 gene flow post-colonization of the southern habitats between southern highland and central populations
4 168 or that the divergence time is very short. Consequently, similar polymorphisms in some parts of the
5 genome are maintained between these populations (Wei, et al. 2023b). These results indicate that the
6 past demographic history of habitat colonisation (and the resulting genetic drift) is an important
7 evolutionary process shaping SNP and CNV frequencies within and between populations of *S. chilense*.
8
9
10
11
12
13

14 172 **Copy number variation illuminates enriched abiotic stress response pathways in *S. chilense***

15 173 We perform functional enrichment analysis of the 3,539 CN-differentiated genes according to GO
16 biological process categories (Dataset S4). We classify these significantly enriched GO categories ($P <$
17 0.05) into nine groups (Fig. S7A) enriched for 82 (cell wall organisation) to 580 (cellular metabolic process)
18 genes. Interestingly, 400 (11.30%) CN-differentiated genes are enriched in response to stimulus/stress
19 that can be linked to multiple environmental factors, for example response to drought (water deprivation;
20 14.35% with 60 genes), cold (17.62% with 37 genes), heat (26.43% with 39 genes), red/far red light (15.82%
21 with 65 genes), or ultraviolet (UV; 19.03% with 47 genes) (Fig. 3B; Fig. S7A; Dataset S4). These
22 responsive pathways support multiple sources of evidence of adaptive processes at genes associated
23 with responses to arid conditions along a steep altitudinal gradient in *S. chilense* (Fischer, et al. 2011;
24 Nosenko, et al. 2016; Böndel, et al. 2018; Blanchard-Gros, et al. 2021; Wei, et al. 2023b). For instance,
25 multiple drought- (HSF and DREB3), cold- (FAD7), and light/cold-responsive genes (FT, GI, and FLD) for
26 flowering regulation (Dataset S5). This supports that selection pressure is not only linked to point
27 mutations but is also manifested as CNVs.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42

43 186 We find 227 genes associated with flowering (Fig. S7A; Fig. S7B), an important fitness trait
44 conditioning local adaptation in plant species (Srikanth and Schmid 2011). As a critical part of the transition
45 from vegetative to reproductive growth, flowering is influenced by several environmental conditions.
46 Therefore, divergent flowering times and adaptation along the ecological gradient may be related to
47 differential CN-differentiated genes (Fig. S7C). We find 31 and 36 CN differentiated genes in response to
48 light and cold and involved in flowering regulation (Fig. S7C), of which 25 and 20 genes are linked to
49 photoperiod and vernalisation pathways (Fig. S8). The latter represent two regulatory flowering time
50 pathways by the relative lengths of light-dark periods and low temperature, respectively (Srikanth and
51 Schmid 2011; Gaudinier and Blackman 2020). These genes are increasingly duplicated in southern
52 highland populations (Fig. 3A and B; Table S8; t-test, $P < 0.05$). These include the potential homologs of
53
54
55
56
57
58
59
60

1
2
3
4 196 floral integrator genes FT and FD (Liu, et al. 2008; Srikanth and Schmid 2011; Putterill and Varkonyi-
5 Gasic 2016), putative homologs of CRY2, GI, and ELF3 in the photoperiod pathway (Srikanth and Schmid
6 197 2011; Makita, et al. 2021), and a putative homolog of AGL14 in the vernalisation pathway (Hecht, et al.
7 198 2005; Pérez-Ruiz, et al. 2015). These candidate genes are well-known flowering time regulators in *A.*
8 199 200 *thaliana* (Dataset S5). Note that these potential candidate genes related to flowering regulation are
10 201 duplicated only in southern highland populations and either no CNV or only copy loss in central and
11 202 southern coast populations (Fig. 3A and B; Fig. S8; Table S8; t-test, $P < 0.05$). These findings indicate
12 203 that gene gains in CN may promote colonisation and adaptation in the southern highland habitats by
13 204 regulating flowering time via the photoperiod and vernalisation pathways (Wei, et al. 2023b). This genomic
14 205 finding is consistent with the phenology observed in glasshouse conditions, in which southern highland
15 206 individuals consistently flower 5-10 days earlier than those from central populations. In addition, other
16 207 potential flowering regulatory genes in the differentiated gene set are likely involved in flowering regulation
17 208 via different pathways, namely the putative homologs of the genes FY and FLD (Dataset S5) (Srikanth
18 209 and Schmid 2011; Cheng, et al. 2017; Bao, et al. 2020) (Srikanth and Schmid 2011; Cheng, et al. 2017;
19 210 Bao, et al. 2020). The FLD gene shows an increased copy number in all populations (Dataset S5; Fig.
20 211 S8).
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

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

224 Our previous SNP study links root development genes to likely local adaptation processes in
225 coastal populations of *S. chilense* (Wei, et al. 2023b). Accordingly, here we find 73 CN-differentiated

1
2
3 226 genes involved in root development, these showing more CNVs in low-altitude populations (C_LA1963,
4 SC_LA2932, SC_LA4107) than in high-altitude populations (C_LA2931, C_LA3111, SH_LA4117A,
5 SH_LA4330) (Fig. 3E; Table S8; t-test, $P < 0.05$).
6
7
8
9
10
11 229 **Gene expansion and contraction patterns show differences along altitudinal gradients**
12
13 230 We reveal that a large number of CN-differentiated genes are potentially involved in response to habitat
14 specialisation. To investigate the CN dynamics of these genes across populations, we perform an analysis
15 of gene CN expansion and contraction across populations based on the population *S. chilense* ultrametric
16 tree (Fig. 4A). The CN of the differentiated genes is expanded (CN gain) in the inland group with a high
17 expansion rate of 1.788. At the same time, it is contracted (CN loss) in the southern coast group with a
18 contraction rate of -0.818 (Table 1). Within the inland group, the southern highland group exhibits an
19 expansion of CN (expansion rate of 0.416). In contrast, the central group shows the number of CN losses
20 (contraction rate of -0.767) three times higher than CN gains (Table 1). This likely indicates that the high
21 rate of CN expansion in the inland group is mainly due to southern highland populations exhibiting high
22 CN gains (Table 1). The two southern highland populations show distinct CN expansion rates of 1.663
23 (SH_LA4117A) and 1.375 (SH_LA4330). In the central group, although the C_LA1963 and C_LA2931
24 display a trend of CN contraction, the C_LA3111 exhibits a similar rate of CN expansion (1.037) as the
25 southern highland populations (Table 1). We relate this to a high migration rate between the high-altitude
26 C_LA3111 and southern highland populations and/or the recent divergence of the southern highland
27 group from the central group (Wei, et al. 2023b). In addition, the similar highland habitat environments
28 (Fig. 1A) may also contribute to the same evolutionary trends of CN gain affecting a similar set of genes
29 for C_LA3111 and southern highland populations. Interestingly, the opposite results are observed
30 between the two southern coast populations.

31
32 248 Gene CN appears as contraction in SC_LA2932 (contraction rate of -0.935) while expansion
33 occurred in SC_LA4107 (expansion rate of 0.534; Table 1). This follows our previous observation that the
34 two southern coast populations show a high degree of differentiation, possibly resulting from a long time
35 of evolution in isolation. These results are also consistent with the population structure (Fig. 2) and may
36 reflect the old southernmost colonisation of the coastal habitats and the recent colonisation of the
37 highlands (Stam, et al. 2019b; Wei, et al. 2023b). Considering that the reference genome is assembled
38 from population C_LA3111, which probably does not represent the ancestral state of the species, we also
39 perform the same analysis using gene CN profiles calculated from the reference genome of *S. pennelli*,
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 256 a drought-adapted wild tomato species. Almost consistent results were observed, except for a decrease
4 257 in the rate of CN expansion in C_LA3111 (Table S9). This may also be a further hint that the dynamics of
5 258 gene CN may reflect the evolutionary history of populations. Overall, the copy numbers of these potentially
6 259 adaptively differentiated genes show an expansion (CN gain) in the two previously elucidated southward
7 260 colonisation events (Fig. 4B; Fig. S10A) (Stam, et al. 2019b; Wei, et al. 2023b).

13
14 261 We define **155 “rapidly evolving genes”** that exhibit significantly higher CN expansion or contraction
15 262 (Viterbi $P < 0.05$) across the different groups/populations using the reference genome of *S. chilense* (Table
16 263 1; Dataset S6). The CN profiles of these rapidly evolving genes also clearly support the population clusters
17 264 in the PCA, but C_LA3111 appears closer to SH populations than to the other central populations (Fig.
18 265 S10B and C). The highest number of such rapidly evolving genes are found in the southern highland
19 266 populations (91 genes), including 71 significant CN expanded genes mainly related to photosynthesis of
20 267 light reaction, long-day photoperiodism (flowering), response to UV light and cold, and 20 significant CN
21 268 contracted genes primarily associated with developmental and metabolic processes.
22 269 We also found 56 rapidly CN-evolving genes in the central populations (Table 1; Dataset S6).
23 270 Few rapidly evolving genes in C_LA3111 and C_LA2931 with high altitudes (above 2200 m) exhibit a significant trend of CN expansion
24 271 at genes involved in long-day photoperiodism. This confirms that inland populations at high altitudes may
25 272 exhibit similar CNV signatures of adaptation as highland populations. Among the 51 rapidly evolving
26 273 genes in the southern coast populations, 16 genes show exactly opposite CN profiles: a significant
27 274 contraction in SC_LA2932 versus an expansion in SC_LA4107 (Fig. 4C). These genes include few
28 275 homologs of photosystem subunits (i.e., *psbB* and *petD*) mainly involved in photosynthesis (Dataset S5)
29 276 and may underpin the high genetic differentiation at the CNV level between the two southern coast
30 277 populations. In addition, the same CN rapidly evolving genes enriched for photosynthesis (light reaction)
31 278 GO categories are also found in central and southern highland groups (Fig. 4D). These potentially
32 279 photosynthetic gene families appear to have been contracting (CN loss) in the central group and
33 280 SC_LA2932 but expanding (CN gain) in the southern highland group and SC_LA4107, suggesting that
34 281 changes in the photosynthetic pathway are also an important adaptive strategy across the different
35 282 habitats in *S. chilense*.

36
37 283 **CN-differentiated genes are associated with climatic variation along the altitudinal gradient**
38 284 To further support CNV as the genetic underpinning of adaptive response to abiotic factors, we conduct

1
2
3
4 285 two genome-environment associations (GEA) analyses between gene CN and 37 climate variables
5 286 (Dataset S7).

6
7 287 We first implement a redundancy analysis (RDA) to identify climate variables significantly
8 associated with CN-differentiated genes across the seven populations. Three climatic variables are
9 observed to correlate with CN changes in the RDA based on $V_{ST}(CN) > 0$. The first
10 three RDA axes (Permutation test, $P < 0.001$) retain 22.62% of the putative adaptive gene CN variances
11 and only weakly distinguish between inland and southern coast populations (Fig. S11B to D). The gene
12 CN differentiation of 52.11% can be explained by six climate variables (explanatory variables) from five
13 significant RDA axes (Permutation test, $P < 0.001$) based on the 3,539 CN differentiated genes (Fig. 5A;
14 Fig. S11E). These climatic variables are significantly correlated with the profiles of the CN-differentiated
15 genes (Mantel test, $P < 0.05$; Fig. 5B). In concordance with the PCA (Fig. 2A), the two main ordination
16 axes do cluster the seven populations into four groups corresponding to the main geographical habitats
17 (central, southern highland and two southern coast habitats). RDA1 is correlated with the annual
18 temperature range (Bio7) and potential evapotranspiration during the driest period (PETDriestQuarter).
19 This axis represents the differentiation between the southern coast and inland populations (Fig. 5A and
20 Fig. S11E). RDA2 reflects the differentiation between two southern coast populations by mean temperature of the
21 wettest quarter (Bio8). RDA2 also summarises a climatic gradient differentiating the low altitude
22 (C_LA1963) and highland populations, which is mainly driven by solar radiation (ann_Rmean) and
23 potential evapotranspiration (annualPET and PETColdestQuarter) (Fig. 5A and B). These six climatic
24 variables are primarily associated with the colonisation of southern highland and southern coast
25 populations (Fig. 5B). The proportions of gene CN differentiation explained by these six climatic variables
26 range from 0.02 (annualPET) to 0.136 (PETColdestQuarter) (Fig. 5C), in which PETColdestQuarter and
27 PETDriestQuarter (0.121) exhibit the highest importance and correlate with inland and southern coast
28 populations, respectively (Fig. 5A to C). Moreover, temperature changes (Bio7 and Bio8) also explain
29 about 20.8% of the gene CN differentiation (Fig. 5C). Solar radiation (ann_Rmean) is a specific variable
30 correlated with high altitude populations and explains 3.6% of gene CN differentiation (Fig. 5A to C). The
31 consistent RDA model is obtained using the 2,192 strongly CN-differentiated genes (Fig. S11F to H).
32 Finally, as a control for the test, we observe a lack of significant RDA model or associated climate variables
33 (Permutation test, $P > 0.001$) when implemented on the 20,372 genes that are not in the CN-differentiated
34 gene set (Fig. S11A).

LFMM

1
2
3 315 We subsequently search for candidate genes (among the 3,539 CN-differentiated genes) that may
4 316 be associated with the six overrepresented climate variables using latent factor mixed models (LFMM)
5 317 (Fig. S12A) Campo (Frichot et al. 2013; Cave(Frichot, et al. 2013; Cave, et al. 2019). We identify 312 CN-
6 318 differentiated genes significantly associated with the six climatic variables (z-test; calibrated $P < 0.01$; Fig.
7 319 S12A and B; Dataset S8). The PCA based on the CN of these 312 candidate genes displays consistent
8 320 population clustering in the RDA models (Fig. S13A; Fig. 5A), supporting that the six climate variables
9 321 reflect gene CN dynamic changes across the species distribution. Among these 312 candidates, we find
10 322 217 genes to be significantly associated with three PET climate variables (annualPET, PETDriestQuarter,
11 323 and PETColdestQuarter), of which 98 genes are shared between the three variables (Fig. S12B). Indeed,
12 324 PET is the primary variable reflecting the drought status of the habitat. We note that these PET-associated
13 325 CN-differentiated genes are mainly involved in metabolic and root development processes and are found
14 326 across all populations (Fig. S13B and C). These physiological processes (ABA signalling pathway, root
15 327 hair differentiation) are essential responses to drought stress using transcriptome and genome analysis
16 328 (Wei, et al. 2023a; Wei, et al. 2023b). This V confirms that drought tolerance is likely the main environmental
17 329 pressure driving CNV evolution across S. chilense distribution. Further, 69% (34/49) of genes associated
18 330 with Bio7 V are also observed to be correlated with ann_Rmean (Fig. S12B); these genes are mainly
19 331 duplicated in southern highland populations and lost in southern coast populations (Fig. 5D; Table S10).
20 332 This V likely reflects that cold and high solar radiation are challenging conditions in southern highland
21 333 populations (Dataset S7). Multiple duplicated genes associated with solar radiation (ann_Rmean) are
22 334 responsive to UV in high-altitude populations (Fig. 5D), such as (likely) homologs of UV-B receptor ARI12,
23 335 and DNA repair protein REV1 (Dataset S5) (Tossi, et al. 2019; Thompson and Cortez 2020). In addition,
24 336 we also find a few CN-differentiated genes, such as putative homologs of CPD (Dataset S5), which relate
25 337 to pigment (anthocyanins) accumulation and are statistically associated with solar radiation variables.
26 338 We finally observe that the number of duplicated genes associated with the six climatic variables
27 339 in the southern coast and especially southern highland populations is much higher than in the central
28 340 populations (Fig. S13B). These duplicated genes are involved in response to environments, including
29 341 light, drought, cold, UV, and carbohydrate (photosynthesis), such as likely homologs of the genes FT, FD,
30 342 and ABI4 and genes involved in the formation of photosystem subunits (Dataset S5). The number of
31 343 candidate genes found as deletions is similar in different populations (Fig. S13C). Most lost genes are
32 344 related to plant growth and development. The GEA analyses confirm the adaptive relevance of gene CN

1
2
3 345 expansion and contraction: (i) the CN-differentiated genes in the central group appear mainly as
4 contraction genes (deletions) while these appear at expansion (duplications) in the southern highland
5 populations; (ii) the adaptive gene CN changes reflect the colonisation of novel habitats at the southern
6 edge of the species distribution; and (iii) the expansion and contraction of gene CN in different populations
7 348 are the consequences of the response to the different habitat environments.
8
9
10
11
12
13
14
15 350 **Discussion**



18 351 A set of key genomic CNVs are found to be highly correlated with the species colonisation process and
19 environmental variables and thus are likely implicated in the adaptive differentiation between populations.
20
21 353 most likely because of their major impact on gene expression (Fuentes, et al. 2019; Rinker, et al. 2019;
22 Alonge, et al. 2020; Hämälä, et al. 2021; Li, et al. 2023). This confirms that CNVs have ubiquitous roles
23 in adaptive processes in ecology and evolution (Żmieleńko, et al. 2014; Castagnone-Sereno, et al. 2019;
24 355 Lauer and Gresham 2019; Mérot, et al. 2020). To better understand the genetic basis behind the fitness
25 effect of CNVs in natural populations, we analyse whole-genome data for 35 *S. chilense* individuals from
26 seven populations, allowing us to identify a genome-wide CNVs dataset. Our CNV calling pipeline
27 358 resolves hundreds of thousands of CNVs. The number of CNV for each population of *S. chilense* is similar
28 to numbers found in the previous tomato clade panSV-genome study that includes a single sample of *S.*
29 360 *chilense* (Li, et al. 2023). CNVs are abundant across all chromosomes and frequently reside within or in
30 close proximity to genes (Fig. 1). Widespread CNVs in the genome exhibit similar performance as SNPs
31 362 for the inference of population structure and differentiation between populations (Fig. 2; Fig. S3)
32 Cheeseman, et al. 2016; Fuentes, et al. 2019). Based on the past demographic model we developed
33 364 previously (Wei, et al. 2023b) as a neutral evolution baseline and the dynamics of CN profile in two
34 southward colonisation events, our results support that most CNV is likely shaped by neutral processes
35 366 (Silva-Arias, et al. 2023). However, this genome-wide assessment allows us to identify CNV likely related
36 to the adaptive divergence in recently colonised regions in response to abiotic stress.
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

369 We identified gene signatures putatively exhibiting footprints of adaptive divergence using CN
370 profiles, and these candidate genes are associated with adaptation to local environments, consistent with
371 genome scans based on SNPs (Wei, et al. 2023b). CN differences of these genes across different
372 populations reflect the neutral and divergent selection process between populations (Fig. 3A; Fig. S6),

1
2
3 373 demonstrating that CNVs must be considered to fully understand how selection shapes genomic structural
4 diversity and local adaptation. Overall, the evolutionary processes generating CNV diversity and
5 divergence follow the historical demography of *S. chilense*, namely two southward independent
6 colonisation events. Genes CN appear expanded in the southernmost SC_LA4107 and southern highland
7 populations, which underwent recent colonisation events and exhibit lower population sizes (Stam, et al.
8 2019b; Wei, et al. 2023b), while gene CN reveals a trend of contraction in the central and SC_LA2932
9 populations (close to the species ~~Y~~ centre of origin). Therefore, we conclude that CN expansion and
10 contraction are not only due to neutral evolutionary processes (past demographic events) but likely reflect
11 and underpin selective events during the two southward colonisation events. Conversely, early
12 established populations exhibit adaptive loss of gene and function processes (Albalat and Cañestro 2016;
13 Helsen, et al. 2020), especially in genes involved in plant growth and development in central populations,
14 or the loss of genes involved in photosynthesis in central and SC_LA2932 populations (Fig. 4D). Changes
15 at photosynthetic gene CN underpin population differentiation between SC_LA2932 (gene loss) and
16 SC_LA4107 (gene gain) representing two different habitats of the southern coast. CN differentiated genes
17 were also enriched in response to multiple abiotic stresses, such as red/far red light, cold, UV, or drought.
18 These response processes can directly affect plant reproduction and growth and regulate flowering
19 regulatory processes (Fig. S7). This further emphasises results based on our SNP study showing that the
20 reproductive cycle, primarily ~~V~~ regulating ~~V~~ flowering time, may play a key role in adaptation to abiotic stress
21 in *S. chilense* (Wei, et al. 2023b).

22
23 392 Flowering regulation involved in response to light (photoperiod) and cold (vernalisation) are key
24 adaptive pathways for *S. chilense* populations to colonise southern habitats based on genome-wide SNPs
25 (Wei, et al. 2023b). Here, we obtain further candidate genes with differentiated gene CN profiles involved
26 in flowering regulatory pathways for response to changes in photoperiod and cold. These genes (putative
27 FT, FD, FLD homologs, etc.) are duplicated in the southern highland populations (Fig. S8). In addition,
28 solar radiation is also a challenging condition for plants at high altitudes. Many CN-differentiated genes
29 are indeed enriched in response to UV light (Fig. 3B; Dataset S4), including homologs of genes involved
30 in anthocyanin accumulation. Indeed, the anthocyanin pathway is switched off in cultivated tomato by
31 mutations of splice sites in regulatory genes and anthocyanin-producing tomato varieties have been
32 created by genetic engineering to obtain anthocyanin-rich purple fruits (Gonzali, et al. 2009; Sun, et al.
33 2020; Gonzali and Perata 2021). These CN-differentiated genes related to the anthocyanin pathway still
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

1
2
3 403 provide a potential source of natural variation for breeding tomato with anthocyanin. More generally, the
4 404 large number of gene losses possibly in response to environmental stresses, may indicate that the
5 405 reduction of the genome size is a powerful evolutionary driver of adaptation (Albalat and Cañestro 2016;
6 406 Helsen, et al. 2020; Monroe, et al. 2021). Further functional validation will help understand the molecular
7 407 mechanisms through which CNV drives adaptive evolution in natural populations.
8
9
10
11
12
13

14 408 Genome-Environment Association (GEA) analysis ultimately links the dynamics of gene CN to six
15 409 climatic variables and reveals the population structure of CNVs in connection to four different habitat
16 410 environments (Fig. 5A and B). These overrepresented climate variables are almost uniformly associated 
17 411 with SNPs in an RDA analysis (Wei, et al. 2023b). These potential CNV-environmental interactions have
18 412 been observed in *Arabidopsis thaliana* (DeBolt 2010; Zmienko, et al. 2020), *Solanum lycopersicum*
19 413 (Alonge, et al. 2020), *Theobroma cacao* (Hämälä, et al. 2021), *Oryza sativa* (Fuentes, et al. 2019; Qin, et
20 414 al. 2021). We are further explicit that CNVs play an essential role in southward colonisation in *S. chilense*.
21
22 415 CNVs, especially duplications in southern highland populations exposed to typical high-altitude stresses,
23 416 show adaptations in genes with functions related to cold, change of photoperiod and solar radiation. The
24 417 CN profiles of differentiated genes in southern coast populations mainly correlate with drought stress,
25 418 such as root development, cell homeostasis, or cell wall maintenance. Interestingly, gene CN
26 419 differentiation related to photosynthesis provides evidence for the genetic underpinning of the adaptive
27 420 differentiation between SC_LA2932 and SC_LA4107, representing two different coastal habitats (Fig. 1A
28 421 and 4C). These differentiated genes reveal exactly opposite CN evolutionary trends between them
29 422 (Dataset S6). Indeed, we see different habitats as SC_LA2932 grows in dry ravines (quebrada) in Lomas
30 423 formations, whereas SC_LA4107 grows in extremely fine alluvial soil (with even some running water).
31
32 424 Moreover, these chloroplast genes are detected in the nuclear genome indicating a widespread event of
33 425 organellar gene transfer to the nuclear genome in tomato (Pesaresi, et al. 2014; Lichtenstein, et al. 2016;
34 426 Kim and Lee 2018). These adaptive signatures were not found in previous studies based on genome
35 427 scans of SNPs (Wei, et al. 2023b). The three central populations display mainly a trend towards gene loss
36 428 and low correlation with climatic variables (Fig. 5A and B). This is consistent with the fact that GEA
37 429 analyses based on current climatic data have limited statistical power to detect old adaptive selection
38 430 signals, whether based on SNPs or CNVs, due to the occurrence of multiple historical confounding events
39 431 such as genetic drift, migration, and recombination (De Mita, et al. 2013; Manel, et al. 2016). The two
40 432 central populations (C_LA2931 and C_LA3111) found at high altitudes exhibit few  adaptive duplications

1
2
3 433 signatures, but some as possible responses to cold and solar radiation, similar to those observed for the
4 434 SH populations (Stam, et al. 2019b; Wei, et al. 2023b).
5
6
7
8
9 435 We finally advise that our study likely underestimates the amount and importance of CNVs in *S.*
10 436 *chilense* as we do not possess long-read data for all accessions. First, our pipeline to recover CNVs based
11 437 on short-read data is tested by simulations and is likely conservative, meaning that we probably miss
12 438 some CNVs. Second, there may be some potential bias in finding footprints of selection when using
13 439 populations multiplied at the TGRC (UC Davis, USA) as we discussed previously (Wei, et al. 2023b).
14
15 440 Though we point out that the use of several selective sweep detection methods conservatively
16 441 underestimate the amount of (positive) selection signals. The availability of a new reference genome
17
18 442 (Silva-Arias, et al. 2023) and few accessions sequenced with long-read (Li, et al. 2023) do open the path
19
20 443 to sequence wild accessions with long-read sequencing and a complete assessment of the importance of
21 444 CNVs at abiotic stress genes in *S. chilense*. Furthermore, the new simulation method to study and infer
22
23 445 the neutral and selective processes driving gene duplication and deletion (Otto, et al. 2022; Otto and
24
25 446 Wiehe 2023) can be used in the future to refine our conclusions regarding the neutral rates of gene
26
27 447 duplication/deletion during the species southward expansion. Despite being conservative regarding the
28
29 448 importance of positive selection shaping the CNV diversity in *S. chilense*, our results reinforce the
30
31 449 observation that CNV is an important contributor to adaptation across different ecological habitats
32
33 450 (Żmieleńko, et al. 2014; Rinker, et al. 2019; Hämälä, et al. 2021; Monroe, et al. 2021). The strong selective
34
35 451 pressure imposed by the range expansion of *S. chilense* and the need to adapt to novel stressful habitats
36
37 452 has shaped the genetic diversity at SNPs and CNVs. In agreement with previous studies, we confirm that
38
39 453 natural selection acting through CNVs can reshape the population genome to underpin adaptation (Iskow,
40
41 454 et al. 2012; Żmieleńko, et al. 2014; Rinker, et al. 2019; Hämälä, et al. 2021).
42
43
44
45
46
47
48
49 455 Materials and Methods 
50
51
52 456 For complete materials and methods, see SI Appendix, Supplementary Information Text.
53
54
55
56 457 Sequence Read Processing
57
58
59 458 We used 35 whole-genome paired-end Illumina data from seven populations of *S. chilense* (five diploid
60 459 plants for each population) representing four different geographic groups (Fig. 1A). The data are available

1
2
3 460 on European Nucleotide Archive (ENA) BioProject PRJEB47577. We performed the same pipeline of
4 461 read processing procedure as in a previous study (Wei, et al. 2023b) including quality trimming, mapping
5 462 and SNP calling based on the new reference genome of *S. chilense* (Silva-Arias, et al. 2023).

6
7
8 463 **Identification and genotyping of CNVs**

9
10
11 464 We used LUMPY (Layer, et al. 2014), Manta (Chen, et al. 2016), Wham (Kronenberg, et al. 2015) and
12 465 DELLY (Rausch, et al. 2012) to identify CNVs in the 35 samples. The CNV sets from LUMPY, DELLY,
13 466 Manta and Wham were merged using SURVIVOR v1.0.7 (Jeffares, et al. 2017). The merged CNV set was
14 467 inputted to SVTyper v0.7.0 to call genotypes using a Bayesian algorithm (Chiang, et al. 2015).

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

How do the analyses relate?

30 468 To assess the sensitivity and accuracy of our pipeline for CNV calling, we simulated 1,000 duplication and
31 469 1,000 deletion regions with sizes ranging from 50bp to 1Mb using CNV-Sim v0.9.2 employing the
32 470 functionality of ART (Huang, et al. 2012), and these simulated reads (150 bp) were used as the input for
33 471 the same CNV analysis pipelines to identify CNVs.

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

472 **Population structure analysis**

35 473 The principal component analysis (PCA) was performed using GCTA v1.91.4 (Yang, et al. 2011). The
36 474 inference of population structure was performed using the program ADMIXTURE v1.3.0 (Alexander, et al.
37 475 2009) based on six scenarios (K values from 2 to 7) using SNPs or CNVs.

38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

476 **Quantification of gene copy number**

46 477 We employed two strategies to quantify gene copy number (CN). First, we used Control-FREEC v11.6 to
47 478 estimate CN by 10 kb windows with 1 kb step size across the entire genome (Boeva, et al. 2012). We
48 479 then obtained gene CN from the Control-FREEC outputs, and gene coordinates in the genome. We also
49 480 used Mosdepth v0.3.2 (Pedersen and Quinlan 2018) to calculate read depth by 1,000 bp sliding windows,
50 481 and gene read depth was calculated from gene coordinates. We then used median read-depth values of
51 482 all windows and genes as a normalising factor to obtain the final window and gene CN estimate,
52 483 respectively, with the formula: CN = (read-depth / median value) × 2.

53
54
55
56
57
58
59
60

484 **Identification of candidate genes associated with population differentiation**

1
2
3 485 The V_{ST} and F_{ST} statistics are applied to quantify population differentiation and are computed over 1,000,
4 486 bp windows. We calculated two V_{ST} estimates based on two different CN quantitative strategies: Control-
5 487 FREEC and Read Depth. We performed permutation tests (1,000 times) for each gene to extract
6 488 candidate genes. We then selected candidate genes with V_{ST} values above the 95th and 99th percentile of
7 489 the permuted V_{ST} distribution for each V_{ST} estimate.
8
9
10
11
12
13
14
15 490 Gene ontology (GO) analysis
16
17
18 491 We first performed a blast of our genes to the *A. thaliana* dataset TAIR10 (e-value cutoff was 10^{-6})
19
20 (Camacho, et al. 2009). We used the R package clusterProfiler to perform GO enrichment analysis using
21
22 the *A. thaliana* annotation database as the background (Yu, et al. 2012). The Benjamini-Hochberg method
23
24 was used to calibrate initial P values, and calibrated P values smaller than 0.05 were used as the cutoff
25
26 for a significant level to obtain final GO terms.
27
28
29 496 Expansion and contraction of gene copy number
30
31
32 497 We computed the expansion and contraction of the 3,359 genes with high CN differentiation between
33
34 populations. We first constructed a population-based phylogenetic tree using SNPs and TreeMix v1.13
35
36 (Pickrell and Pritchard 2012). The ultrametric tree (Fig. 4A) was generated based on force.ultrametric
37
38 function of phytools R package (Revell 2012). We then performed analyses of the expansion and
39
40 contraction of gene CN using CAFE v4.2.1 Campo (Han(Han, et al. 2013)). The branch-specific p-values
41
42 are obtained by the Viterbi method with the randomly generated likelihood distribution. We set a p-value
43
44 smaller than 0.05 to detect gene CN with a significant rate of evolution (expansion or contraction) in
45
46 different groups/populations.
47
48
49 505 Association analysis between gene copy number and climatic conditions
50
51
52 506 The environmental data include 37 climatic layers (Dataset S7) obtained from two public databases,
53
54 WorldClim2 (Fick and Hijmans 2017) and ENVIREM (Title and Bemmels 2018). To evaluate the relative
55
56 contribution of the abiotic environment to explaining patterns of genetic variation, we first used the
57
58 Redundancy Analysis (RDA) (Capblancq and Forester 2021) to associate CNs of the 3,539 differentiated
59
60 genes with climatic variables. RDA was performed using the rda function from the vegan package as

1
2
3 511 implemented in R (Forester, et al. 2018). LFMM (latent factor mixed models) is a univariate test (Frichot,
4 et al. 2013; Caye, et al. 2019), which means it builds a model for each gene or SNP and each predictor
5 variable. We then implemented LFMM2 to perform an association test between gene CN and six
6 representative climate variables obtained in RDA, respectively. Benjamini-Hochberg's method was used
7 to calibrate the *P* values with 0.01 as the significant threshold.
8
9
10
11
12
13

14 516 **Supplementary material**

15 517 Supplementary data are available online at Molecular Biology and Evolution.
16
17

18 518 **Data Availability**
19
20

21 519 Raw sequence data are available at the European Nucleotide Archive (ENA) BioProject PRJEB47577.
22 520 The resource of copy number variation identified in this study and custom scripts for conducting the
23 analyses are available at our Gitlab at the following link:
24 521
25 522 https://gitlab.lrz.de/population_genetics/s_chilense_cnv.
26
27
28
29
30
31
32

33 523 **Acknowledgements**
34
35

36 524 KW was funded by the Chinese Scholarship Council. GAS-A was funded by the Technical University of
37 Munich. AT acknowledges funding from DFG (Deutsche Forschungsgemeinschaft) Grant Number:
38 525 317616126 (TE809/7-1). We thank the Tomato Genetics Resource Center (TGRC) of the University of
39 526 California, Davis for generously providing us with the seeds of the accession included in this study.
40
41
42
43
44

45 528 **Competing interests**
46
47

48 529 The authors have no conflicts of interest to declare.
49
50

51 530 **Author contributions**
52
53

54 531 KW, GAS-A and AT planned and designed the study. RS and AT obtained the sequencing data. KW
55 532 performed data analyses. KW wrote the first draft of the manuscript, and RS, GAS-A, and AT edited and
56 533 improved the manuscript. All authors approved the final manuscript.
57
58
59
60

534

535 **References**

- 536 Albalat R, Cañestro C. 2016. Evolution by gene loss. *Nature Reviews Genetics* 17:379-391.
- 537 Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated
538 individuals. *Genome research* 19:1655-1664.
- 539 Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Cirene D.
540 2020. Major impacts of widespread structural variation on gene expression and crop improvement in
541 tomato. *Cell* 182:145-161. e123.
- 542 Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, Cao J, Chae E, Dezwaan TM,
543 Ding W, et al. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*.
544 *Cell* 166:481-491.
- 545 Antinucci M, Comas D, Calafell F. 2023. Population history modulates the fitness effects of Copy
546 Number Variation in the Roma. *Human Genetics*:1-17.
- 547 Arunyawat U, Stephan W, Städler T. 2007. Using multilocus sequence data to assess population
548 structure, natural selection, and linkage disequilibrium in wild tomatoes. *Molecular biology and*
549 *evolution* 24:2310-2322.
- 550 Bao S, Hua C, Shen L, Yu H. 2020. New insights into gibberellin signaling in regulating flowering in
551 *Arabidopsis*. *Journal of integrative plant biology* 62:118-131.
- 552 Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J. 2016. Recent demography
553 drives changes in linked selection across the maize genome. *Nature Plants* 2:16084.
- 554 Blanchard-Gros R, Bigot S, Martinez J-P, Lutts S, Guerriero G, Quinet M. 2021. Comparison of Drought
555 and Heat Resistance Strategies among Six Populations of *Solanum chilense* and Two Cultivars of
556 *Solanum lycopersicum*. *Plants* 10:1720.
- 557 Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O,
558 Barillot E. 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-
559 generation sequencing data. *Bioinformatics* 28:423-425.
- 560 Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sørensen I,
561 Lichtenstein G. 2014. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nature*
562 *genetics* 46:1034-1038.
- 563 Böndel KB, Lainer H, Nosenko T, Mboup M, Tellier A, Stephan W. 2015. North-south colonization
564 associated with local adaptation of the wild tomato species *Solanum chilense*. *Molecular biology and*
565 *evolution* 32:2932-2943.
- 566 Böndel KB, Nosenko T, Stephan W. 2018. Signatures of natural selection in abiotic stress-responsive
567 genes of *Solanum chilense*. *Royal Society open science* 5:171198-171198.
- 568 Brumlop S, Weedon O, Link W, Finckh M. 2019. Effective population size (N_e) of organically and
569 conventionally grown composite cross winter wheat populations depending on generation. *European*
570 *Journal of Agronomy* 109:125922.
- 571 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+:
572 architecture and applications. *BMC bioinformatics* 10:1-9.

- 1
2
3 573 Capblancq T, Forester BR. 2021. Redundancy analysis: A Swiss Army Knife for landscape genomics.
4 574 Methods in Ecology and Evolution 12:2298-2309.
- 5 575 Castagnone-Sereno P, Mulet K, Danchin EG, Koutsovoulos GD, Karaulic M, Da Rocha M, Bailly-Béchet
6 576 M, Pratx L, Perfus-Barbeoch L, Abad P. 2019. Gene copy number variations as signatures of adaptive
7 577 evolution in the parthenogenetic, plant-parasitic nematode *Meloidogyne incognita*. Molecular Ecology
8 578 28:2559-2572.
- 9 579 Caye K, Jumentier B, Lepeule J, François O. 2019. LFMM 2: Fast and Accurate Inference of Gene-
10 580 Environment Associations in Genome-Wide Studies. Molecular biology and evolution 36:852-860.
- 11 581 Cheeseman IH, Miller B, Tan JC, Tan A, Nair S, Nkhoma SC, De Donato M, Rodulfo H, Dondorp A, Branch
12 582 OH. 2016. Population structure shapes copy number variation in malaria parasites. Molecular biology
13 583 and evolution 33:603-620.
- 14 584 Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT.
15 585 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing
16 586 applications. Bioinformatics 32:1220-1222.
- 17 587 Cheng JZ, Zhou YP, Lv TX, Xie CP, Tian CE. 2017. Research progress on the autonomous flowering time
18 588 pathway in *Arabidopsis*. Physiol Mol Biol Plants 23:477-485.
- 19 589 Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015.
20 590 SpeedSeq: ultra-fast personal genome analysis and interpretation. Nature methods 12:966-968.
- 21 591 Coutelier M, Holtgrewe M, Jäger M, Flöttman R, Mensah MA, Spielmann M, Krawitz P, Horn D, Beule D,
22 592 Mundlos S. 2022. Combining callers improves the detection of copy number variants from whole-
23 593 genome sequencing. European Journal of Human Genetics 30:178-186.
- 24 594 De Mita S, ThUILlet A-C, Gay L, Ahmadi N, Manel S, Ronfort J, Vigouroux Y. 2013. Detecting selection
25 595 along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and
26 596 selfing populations. Molecular Ecology 22:1383-1399.
- 27 597 DeBolt S. 2010. Copy Number Variation Shapes Genome Diversity in *Arabidopsis* Over Immediate Family
28 598 Generational Scales. Genome biology and evolution 2:441-453.
- 29 599 Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nature Reviews
30 600 Genetics 7:85-97.
- 31 601 Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land
32 602 areas. International journal of climatology 37:4302-4315.
- 33 603 Fischer I, Camus-Kulandaivelu L, Allal F, Stephan W. 2011. Adaptation to drought in two wild tomato
34 604 species: the evolution of the Asr gene family. New Phytologist 190:1032-1044.
- 35 605 Forester BR, Lasky JR, Wagner HH, Urban DL. 2018. Comparing methods for detecting multilocus
36 606 adaptation with multivariate genotype–environment associations. Molecular Ecology 27:2215-2233.
- 37 607 Fritchot E, Schoville SD, Bouchard G, François O. 2013. Testing for Associations between Loci and
38 608 Environmental Gradients Using Latent Factor Mixed Models. Molecular biology and evolution 30:1687-
39 609 1699.
- 40 610 Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, Wing RA, McNally KL,

- 1
2
3 611 Tatarinova T, Grigoriev A. 2019. Structural variants in 3000 rice genomes. *Genome research* 29:870-
4 612 880.
5
6 613 Gaudinier A, Blackman BK. 2020. Evolutionary processes from the perspective of flowering time
7 614 diversity. *New Phytologist* 225:1883-1898.
8
9 615 Gonzali S, Mazzucato A, Perata P. 2009. Purple as a tomato: towards high anthocyanin tomatoes. *Trends
10 in plant science* 14:237-241.
11
12 617 Gonzali S, Perata P. 2021. Fruit Colour and Novel Mechanisms of Genetic Regulation of Pigment
13 Production in Tomato Fruits. *Horticulturae* 7:259.
14
15 619 Guo M, Yang F, Liu C, Zou J, Qi Z, Fotopoulos V, Lu G, Yu J, Zhou J. 2022. A single - nucleotide
16 polymorphism in WRKY33 promoter is associated with the cold sensitivity in cultivated tomato. *New
17 Phytologist* 236:989-1005.
18
19 622 Hämälä T, Wafula EK, Guiltinan MJ, Ralph PE, dePamphilis CW, Tiffin P. 2021. Genomic structural
20 variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate
21 tree. *Proceedings of the National Academy of Sciences* 118:e2102914118.
22
23 625 Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the
24 presence of error in genome assembly and annotation using CAFE 3. *Molecular biology and evolution*
25 30:1987-1997.
26
27 628 Hecht Vr, Foucher F, Ferrández C, Macknight R, Navarro C, Morin J, Vardy ME, Ellis N, Beltrán JPo,
28 Rameau C, et al. 2005. Conservation of *Arabidopsis* Flowering Genes in Model Legumes Plant
29 Physiology 137:1420-1434.
30
31 631 Helsen J, Voordeckers K, Vanderwaeren L, Santermans T, Tsontaki M, Verstrepen KJ, Jelier R. 2020. Gene
32 Loss Predictably Drives Evolutionary Adaptation. *Molecular biology and evolution* 37:2989-3002.
33
34 633 Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator.
35 Bioinformatics 28:593-594.
36
37 635 Iskow RC, Gokcumen O, Lee C. 2012. Exploring the role of copy number variants in human adaptation.
38 Trends in Genetics 28:245-257.
39
40 637 Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017.
41 638 Transient structural variations have strong effects on quantitative traits and reproductive isolation in
42 fission yeast. *Nature communications* 8:14061.
43
44 640 Kim HT, Lee JM. 2018. Organellar genome analysis reveals endosymbiotic gene transfers in tomato.
45 PLoS One 13:e0202279.
46
47 642 Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of
48 structural variation detection algorithms for whole genome sequencing. *Genome biology* 20:1-18.
49
50 644 Kronenberg ZN, Osborne EJ, Cone KR, Kennedy BJ, Domyan ET, Shapiro MD, Elde NC, Yandell M. 2015.
51 645 Wham: identifying structural variants of biological consequence. *PLoS computational biology*
52 11:e1004572.
53
54 647 Lauer S, Gresham D. 2019. An evolving view of copy number variants. *Current genetics* 65:1287-1295.

- 1
2
3 648 Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant
4 649 discovery. *Genome biology* 15:1-19.
5
6 650 Li N, He Q, Wang J, Wang B, Zhao J, Huang S, Yang T, Tang Y, Yang S, Aisimutuola P, et al. 2023. Super-
7 651 pangenome analyses highlight genomic diversity and structural variation across wild and cultivated
8 652 tomato species. *Nature genetics*.
9
10
11 653 Lichtenstein G, Conte M, Asis R, Carrari F. 2016. Chloroplast and mitochondrial genomes of tomato. The
12 654 Tomato Genome:111-137.
13
14 655 Liu C, Chen H, Er HL, Soo HM, Kumar PP, Han JH, Liou YC, Yu H. 2008. Direct interaction of AGL24 and
15 656 SOC1 integrates flowering signals in *Arabidopsis*. *Development* 135:1481-1491.
16
17 657 Liu Z, Hou S, Rodrigues O, Wang P, Luo D, Munemasa S, Lei J, Liu J, Ortiz-Moreira FA, Wang X, et al. 2022.
18 658 Phytohormone signalling reopens stomata in plant immunity and water loss. *Nature* 605:332-339.
19
20 659 Luo X, Xu J, Zheng C, Yang Y, Wang L, Zhang R, Ren X, Wei S, Aziz U, Du J, et al. 2022. Abscisic acid inhibits
21 660 primary root growth by impairing ABI4-mediated cell cycle and auxin biosynthesis. *Plant Physiology*
22 661 191:265-279.
23
24 662 Lupski JR. 2007. Genomic rearrangements and sporadic disease. *Nature genetics* 39:S43-S47.
25
26 663 Lynch M, Walsh B. 2007. The origins of genome architecture: Sinauer Associates Sunderland, MA.
27
28 664 Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant
29 665 calling: the long and the short of it. *Genome biology* 20:246.
30
31 666 Makita Y, Suzuki S, Fushimi K, Shimada S, Suehisa A, Hirata M, Kuriyama T, Kurihara Y, Hamasaki H,
32 667 Okubo-Kurihara E. 2021. Identification of a dual orange/far-red and blue light photoreceptor from an
33 668 oceanic green picoplankton. *Nature communications* 12:3593.
34
35 669 Manel S, Perrier C, Pratlong M, Abi - Rached L, Paganini J, Pontarotti P, Aurelle D. 2016. Genomic
36 670 resources and their influence on the detection of the signal of positive selection in genome scans.
37 671 *Molecular Ecology* 25:170-184.
38
39 672 Marszalek-Zenczak M, Satyr A, Wojciechowski P, Zenczak M, Sobieszczanska P, Brzezinski K, Iefimenko
40 673 T, Figlerowicz M, Zmienko A. 2023. Analysis of *Arabidopsis* non-reference accessions reveals high
41 674 diversity of metabolic gene clusters and discovers new candidate cluster members. *Front Plant Sci*
42 675 14:1104303.
43
44 676 Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A roadmap for understanding the evolutionary
45 677 significance of structural genomic variation. *Trends in Ecology & Evolution* 35:561-572.
46
47 678 Monroe JG, McKay JK, Weigel D, Flood PJ. 2021. The population genomics of adaptive loss of function.
48 679 *Heredity* 126:383-395.
49
50 680 Nakazato T, Warren DL, Moyle LC. 2010. Ecological and geographic modes of species divergence in wild
51 681 tomatoes. *American Journal of Botany* 97:680-693.
52
53 682 Nosenko T, Böndel KB, Kumpfmüller G, Stephan W. 2016. Adaptation to low temperatures in the wild
54 683 tomato species *Solanum chilense*. *Molecular Ecology* 25:2853-2869.
55
56 684 Otto M, Wiehe T. 2023. The structured coalescent in the context of gene copy number variation.

- 1
2
3 685 Theoretical Population Biology 154:67-78.
4
5 686 Otto M, Zheng Y, Wiehe T. 2022. Recombination, selection, and the evolution of tandem gene arrays.
6
7 687 Genetics 221:iyac052.
8
9 688 Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes.
10 689 Bioinformatics 34:867-868.
11
12 690 Pérez-Ruiz Rigoberto V, García-Ponce B, Marsch-Martínez N, Ugartechea-Chirino Y, Villajuana-Bonequi
13 M, de Folter S, Azpeitia E, Dávila-Velderrain J, Cruz-Sánchez D, Garay-Arroyo A, et al. 2015. XAANTAL2
14 (AGL14) Is an Important Component of the Complex Gene Regulatory Network that Underlies
15 Arabidopsis Shoot Apical Meristem Transitions. Molecular Plant 8:796-813.
16
17 694 Pesaresi P, Mizzotti C, Colombo M, Masiero S. 2014. Genetic regulation and structural changes during
18 tomato fruit development and ripening. Frontiers in plant science 5:124.
19
20 696 Pickrell J, Pritchard J. 2012. Inference of population splits and mixtures from genome-wide allele
21 frequency data. Nature Precedings:1-1.
22
23 698 Putterill J, Varkonyi-Gasic E. 2016. FT and florigen long-distance flowering control in plants. Current
24 opinion in plant biology 33:77-82.
25
26 700 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
27 of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell 184:3542-3558.e3516.
28
29 702 Raduski AR, Igić B. 2021. Biosystematic studies on the status of Solanum chilense. American Journal of
30 Botany 108:520-537.
31
32 704 Ranjan A, Ichihashi Y, Sinha NR. 2012. The tomato genome: implications for plant breeding, genomics
33 and evolution. Genome biology 13:1-8.
34
35 706 Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery
36 by integrated paired-end and split-read analysis. Bioinformatics 28:i333-i339.
37
38 708 Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things).
39 Methods in Ecology and Evolution 3:217-223.
40
41 710 Rinker DC, Specian NK, Zhao S, Gibbons JG. 2019. Polar bear evolution is marked by rapid changes in
42 gene copy number in response to dietary shift. Proceedings of the National Academy of Sciences
43 116:13446-13451.
44
45 713 Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki
46 K. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635-
47 641.
48
49 716 Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, O'Hara R, Casalunovo T, Conlin LK, D'arcy M.
50 2009. High-resolution mapping and analysis of copy number variations in the human genome: a data
51 resource for clinical and research applications. Genome research 19:1682-1690.
52
53 719 Silva-Arias GA, Gagnon E, Hembrom S, Fastner A, Khan MR, Stam R, Tellier A. 2023. Contrasting patterns
54 of presence-absence variation of NLRS within S. chilense are mainly shaped by past
55 demographic history. bioRxiv:2023.2010.2013.562278.
56
57
58
59
60

- 1
2
3 722 Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H. 2009. Maize
4 inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in
5 genome content. *PLoS genetics* 5:e1000734.
6
7 725 Srikanth A, Schmid M. 2011. Regulation of flowering time: all roads lead to Rome. *Cellular and*
8 *Molecular Life Sciences* 68:2013-2037.
9
10 727 Stam R, Nosenko T, Hörger AC, Stephan W, Seidel M, Kuhn JM, Haberer G, Tellier A. 2019a. The de novo
11 reference genome and transcriptome assemblies of the wild tomato species *Solanum chilense*
12 highlights birth and death of NLR genes between tomato species. *G3: Genes, Genomes, Genetics*
13 9:3933-3941.
14
15 731 Stam R, Silva-Arias GA, Tellier A. 2019b. Subsets of NLR genes show differential signatures of adaptation
16 during colonization of new habitats. *New Phytologist* 224:367-379.
17
18 733 Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt
19 S, Bamshad M. 2015. Global diversity, population stratification, and selection of human copy-number
20 variation. *Science* 349:aab3761.
21
22 736 Sun C, Deng L, Du M, Zhao J, Chen Q, Huang T, Jiang H, Li C-B, Li C. 2020. A transcriptional network
23 promotes anthocyanin biosynthesis in tomato flesh. *Molecular Plant* 13:42-58.
24
25 738 Title PO, Bemmels JB. 2018. ENVIREM: an expanded set of bioclimatic and topographic variables
26 increases flexibility and improves performance of ecological niche modeling. *Ecography* 41:291-307.
27
28 740 Wei K, Sharifova S, Zhao X, Sinha N, Nakayama H, Tellier A, Silva-Arias GA. 2023a. Evolution of two gene
29 networks underlying adaptation to drought stress in the wild tomato *Solanum chilense*.
30 bioRxiv:2023.2001.2018.524537.
31
32 743 Wei K, Silva-Arias GA, Tellier A. 2023b. Selective sweeps linked to the colonization of novel habitats and
33 climatic changes in a wild tomato species. *New Phytologist* 237:1908-1921.
34
35 745 Xia HUI, Camus-Kulandaivelu L, Stephan W, Tellier A, Zhang Z. 2010. Nucleotide diversity patterns of
36 local adaptation at drought-related candidate genes in wild tomatoes. *Molecular Ecology* 19:4144-4154.
37
38 747 Xiao S, Jiang L, Wang C, Ow DW. 2021. *Arabidopsis OXS3* family proteins repress ABA signaling through
39 interactions with AFP1 in the regulation of ABI4 expression. *Journal of Experimental Botany* 72:5721-
40 5734.
41
42 750 Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis.
43 The American Journal of Human Genetics 88:76-82.
44
45 752 Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: an R package for comparing biological themes
46 among gene clusters. *Omics: a journal of integrative biology* 16:284-287.
47
48 754 Zhou Y, Minio A, Massonet M, Solares E, Lv Y, Beridze T, Cantu D, Gaut BS. 2019. The population
49 genetics of structural variants in grapevine domestication. *Nature Plants* 5:965-979.
50
51 756 Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, Wu Y, Cheng L, Fang Y, Wu K. 2022. Graph pangenome captures
52 missing heritability and empowers tomato breeding. *Nature* 606:527-534.
53
54 758 Zmienko A, Marszalek-Zenczak M, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozlowski P,
55 Karlowski WM, Figlerowicz M. 2020. AthCNV: A Map of DNA Copy Number Variations in the *Arabidopsis*

- 1
2
3 760 Genome[OPEN]. *The Plant Cell* 32:1797-1819.
4
5 761 Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M. 2014. Copy number polymorphism in plant genomes.
6
7 762 Theoretical and applied genetics 127:1-18.
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

PDF Proof: Mol. Biol. Evol.

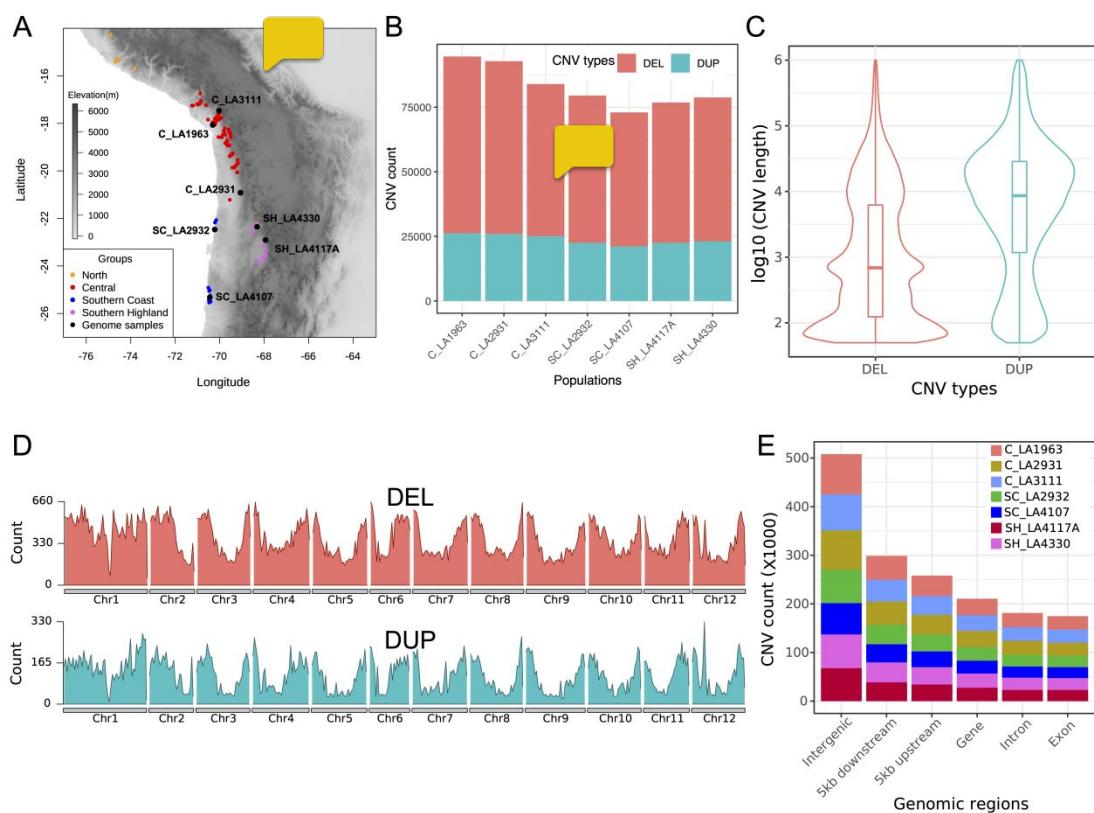
1
2
3 763 **Table1.** The summary of gene expansion and contraction in different groups/populations based on the
4 phylogenetic and ultrametric tree.
5
6

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

34
35 765 The table shows that the expansion and contraction of CN-differentiated genes in different groups /
36 populations based on an ultrametric tree (Fig. 4A). C: central; SH: southern highland; SC: southern coast.
37
38 767 ^aGroups and populations denote the branches in the phylogenetic and ultrametric tree (Fig. 4A).

39
40 768 ^bRate of average expansion / contraction = (Number of CN gained - Number of CN lost) / (Number of CN
41 expanded genes + Number of CN contracted genes). Positive values indicate CN expansion and negative
42 values indicate CN contraction.
43
44 771 ^cThe rapidly evolving genes indicate significant higher CN expansion or contraction (Viterbi $P < 0.05$)
45 across the different groups/populations. Values outside parentheses represent the total number of the
46 rapidly evolving genes. Positive values in parentheses denote the number of significantly expanded genes
47 and negative values denote the number of significantly contracted genes (see also Dataset S6).
48
49 775
50
51
52
53
54
55
56
57
58
59
60

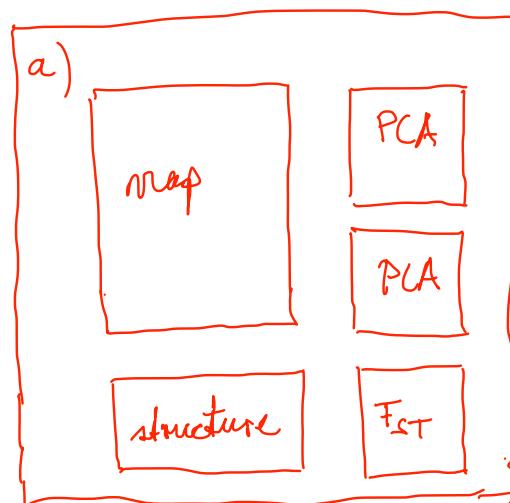
776 Figures



777

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

785



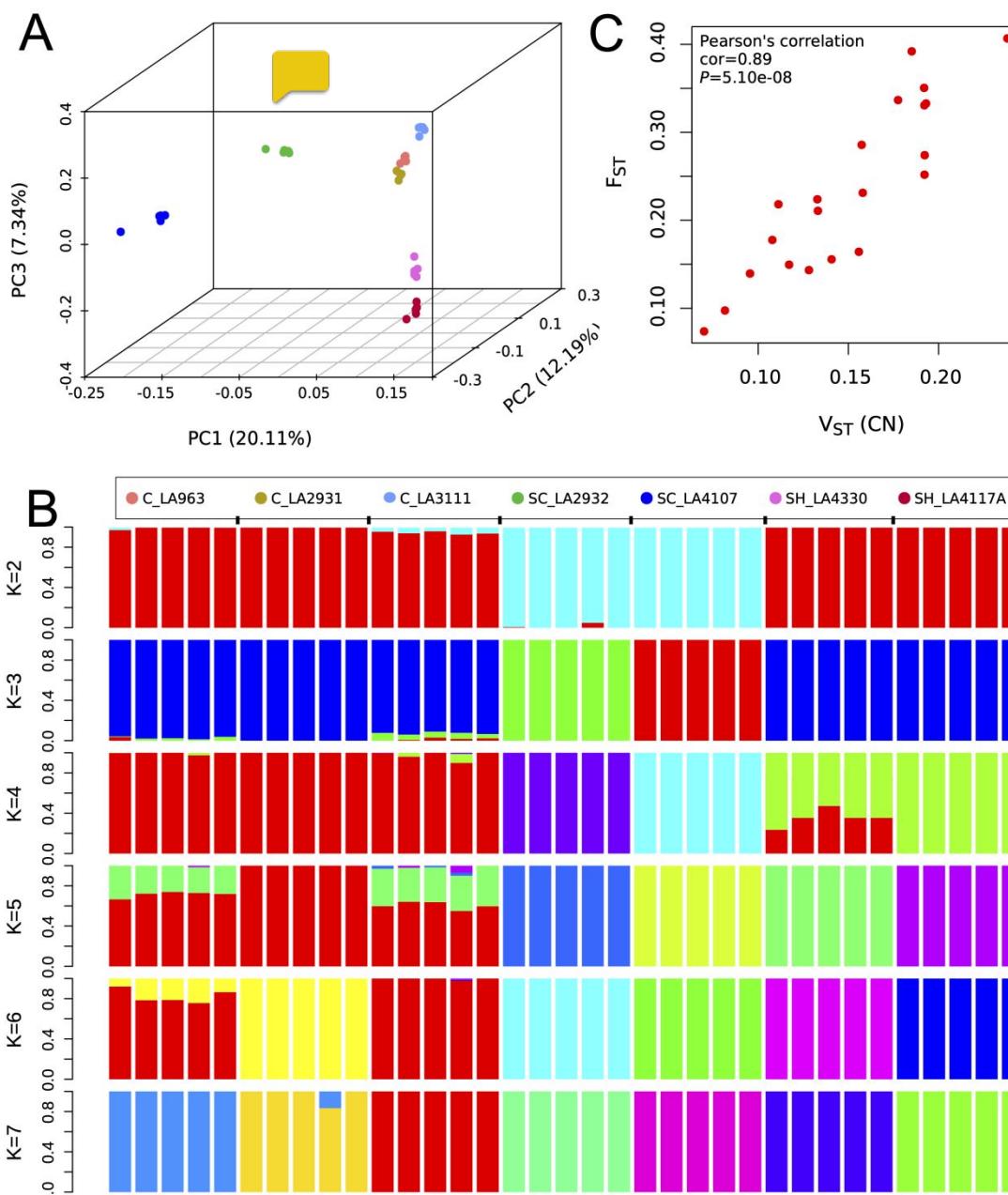
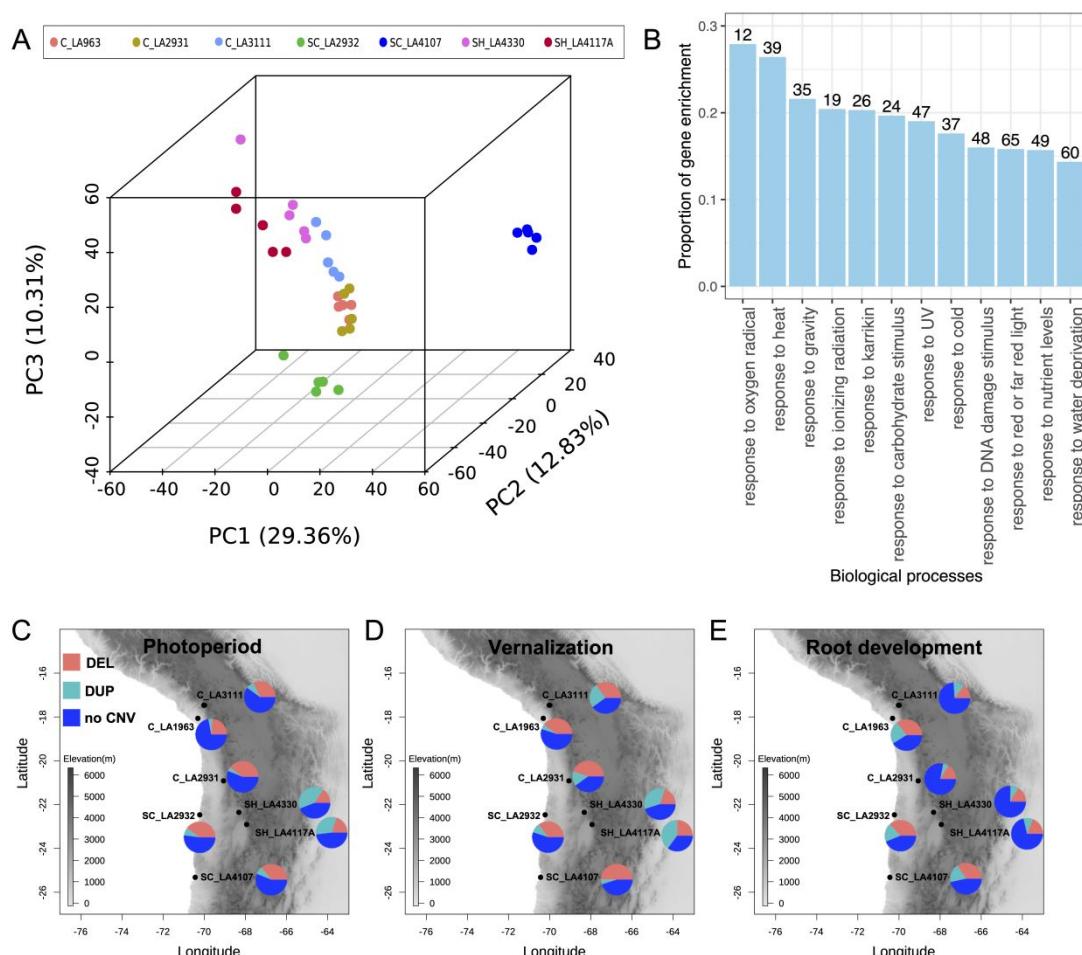


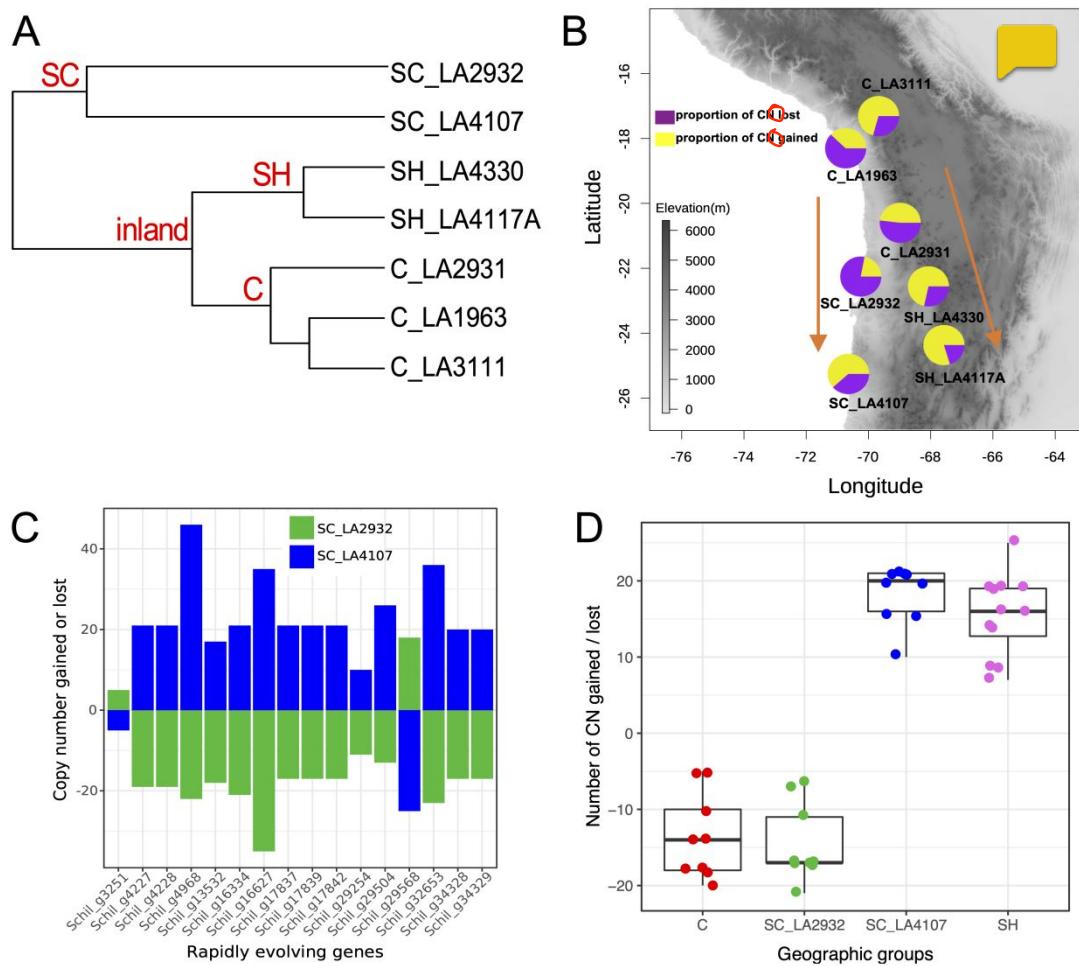
Fig. 2. Structure analysis based on genotyped CNVs. (A) Principal component analysis (PCA) based using genotyped CNVs from 35 *S. chilense* accessions. (B) Structure analysis based on genotyped CNVs and assuming $K = 2 - 7$ subgroups (optimal K value is 4; Fig. S3B). C: central; SH: southern highland;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



793

794 **Fig. 3.** Genes with differentiated CN profiles among seven populations are linked to response to multiple
 795 environmental stimuli. (A) PCA based on the copy number (CN) of 3,539 differentiated genes. C: central;
 796 SH: southern highland; SC: southern coast. (B) The proportions of CN-differentiated genes enriched in
 797 response to external stimulus/stresses (significantly enriched $P < 0.05$). The ratio of gene enrichment is
 798 equal to the number of genes enriched in one GO category divided by the number of background genes
 799 in this category. The number on each bar represents the number of genes enriched in that GO category.
 800 WThe CN-differentiated genes involved in photoperiod pathway to regulate flowering time (C), vernalisation
 801 pathways to regulate flowering time (D), and root developmental process (E). The pie charts denote the
 802 proportions of CN-differentiated genes with deletion (DEL), duplication (DUP) or absence of CNV over all
 803 3,539 genes (see also Table S8).

804



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

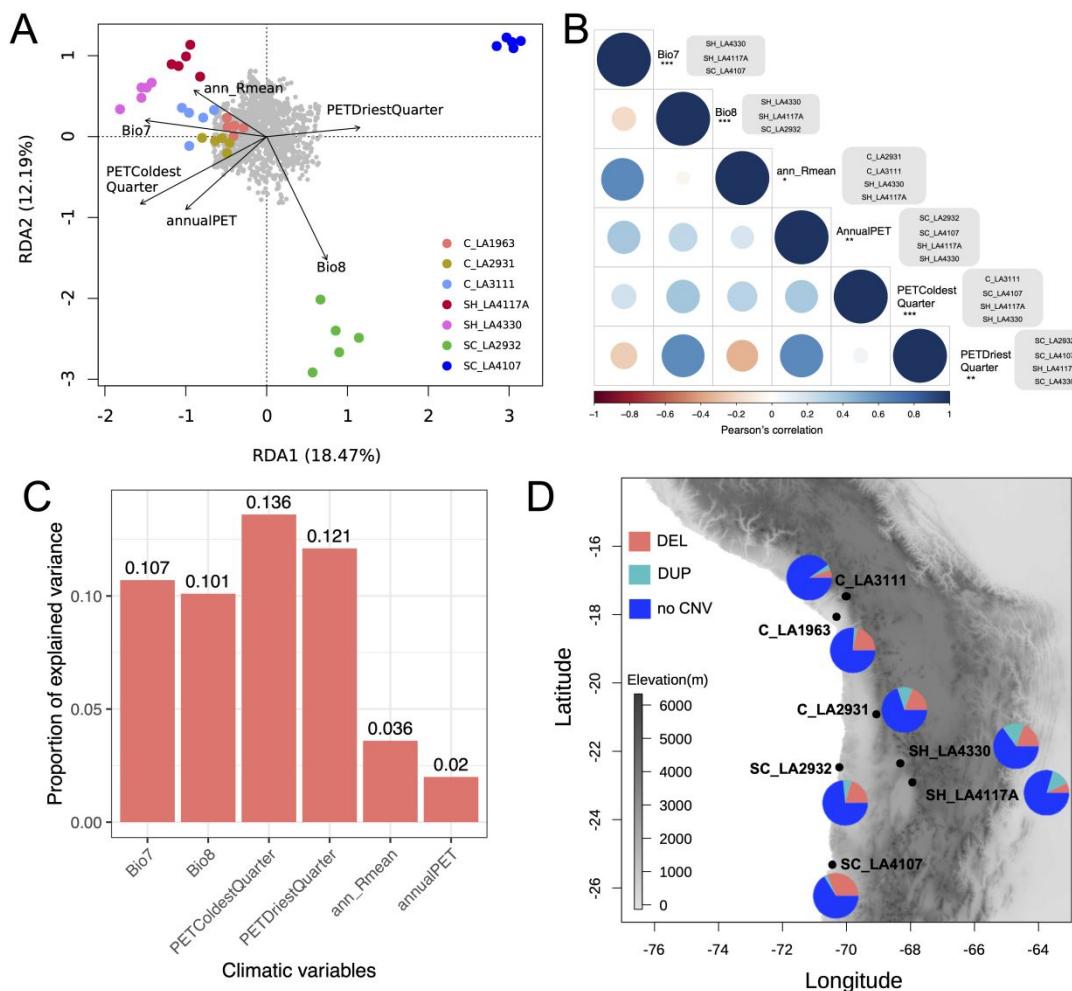


Fig. 5. Genome-Environment Association analysis reveals that CN differentiated genes adapt to different habitat environments. (A) Redundancy analysis (RDA) ordination biplots between the climatic variables (Dataset S7), populations, and 3,539 differentiated gene CN. In the RDA, the loading of the climatic variables or the length of the vector indicates the strength of the correlation with the ordination axis. The grey points denote the CN-differentiated genes. C: central; SH: southern highland; SC: southern coast.

(B) The correlations between six overrepresented climate variables and populations, respectively. The bubble chart shows correlations between six climate variables. The asterisks (*) indicate the levels of significance of the climate variables for the RDA model (Permutation test; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.0001$). The grey boxes to the right of the climatic variables show the populations significantly associated with that climatic variable (Mantel test, $P < 0.05$). (C) The proportion of explained variance for six overrepresented climate variables in the RDA model. (D) 34 CN-differentiated genes associated with both temperature annual range (Bio7) and solar radiation (ann_Rmean) in seven populations. The pie charts denote the proportions of CN-differentiated genes with deletion (DEL), duplication (DUP) or absence of CNV (see also Table S10).