Supplementary Figures and Tables

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

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

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

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

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

*Corresponding authors: Aurélien Tellier: aurelien.tellier@tum.de;
Kai Wei: kai.wei@tum.de

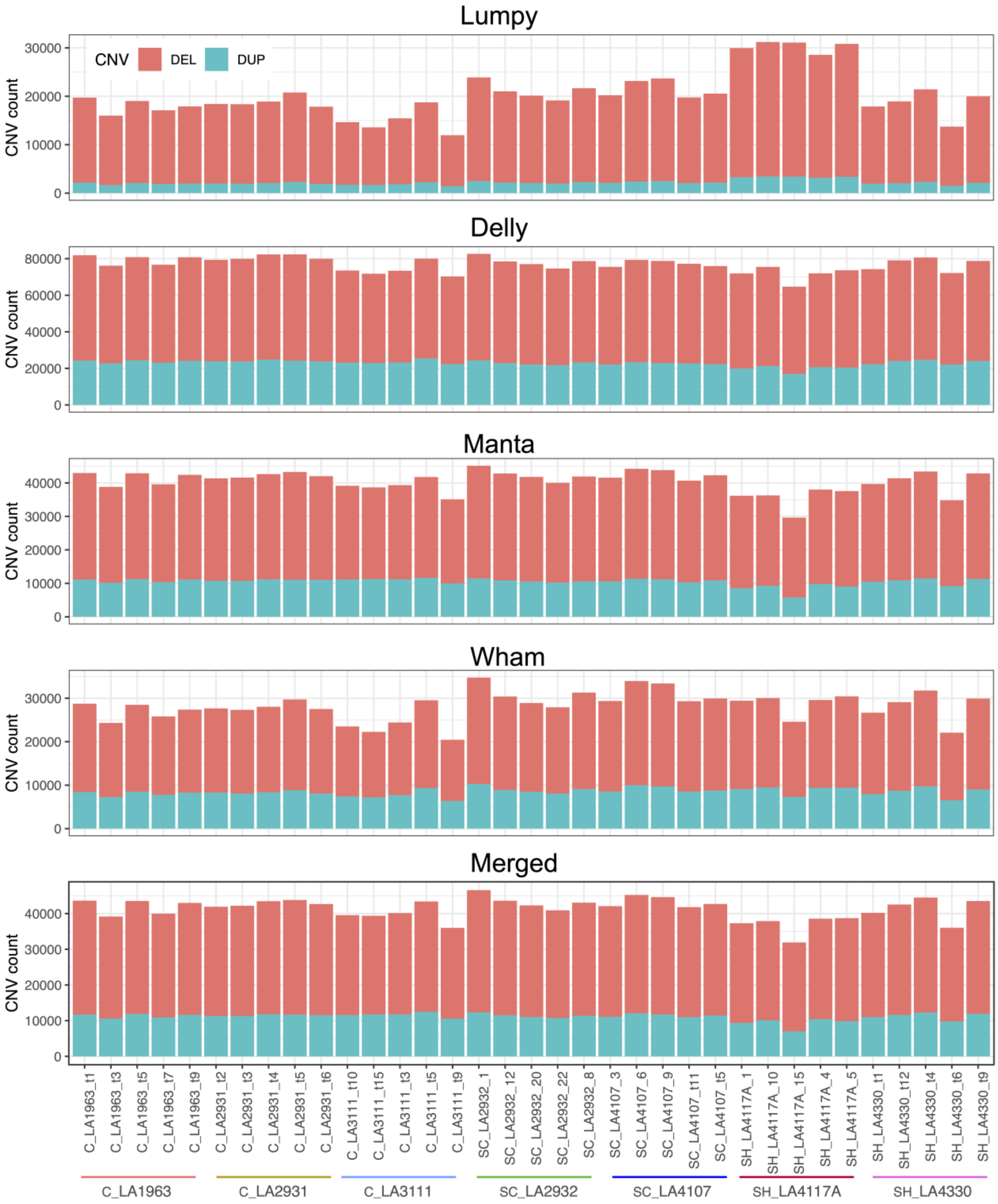

Figure S1. The summary of deletion (DEL) and duplication (DUP) using four CNV callers in each population, and merged result when CNVs need to be identified by at least two callers. See also Dataset S2 and Table S2.

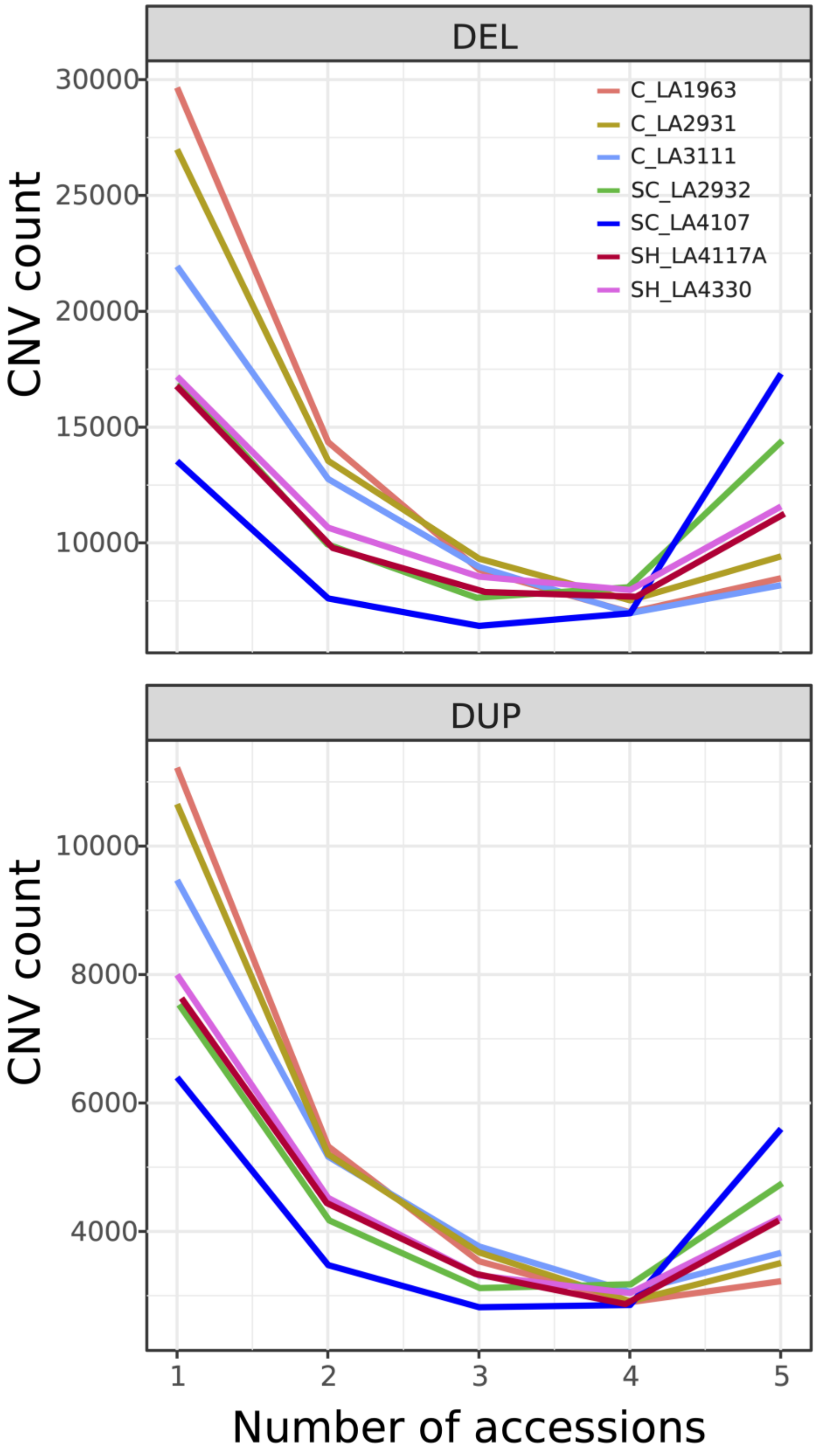

Figure S2. The number of CNVs identified in 1, 2, 3, 4 or 5 individuals in each population, respectively. DEL: deletion; DUP: duplication.

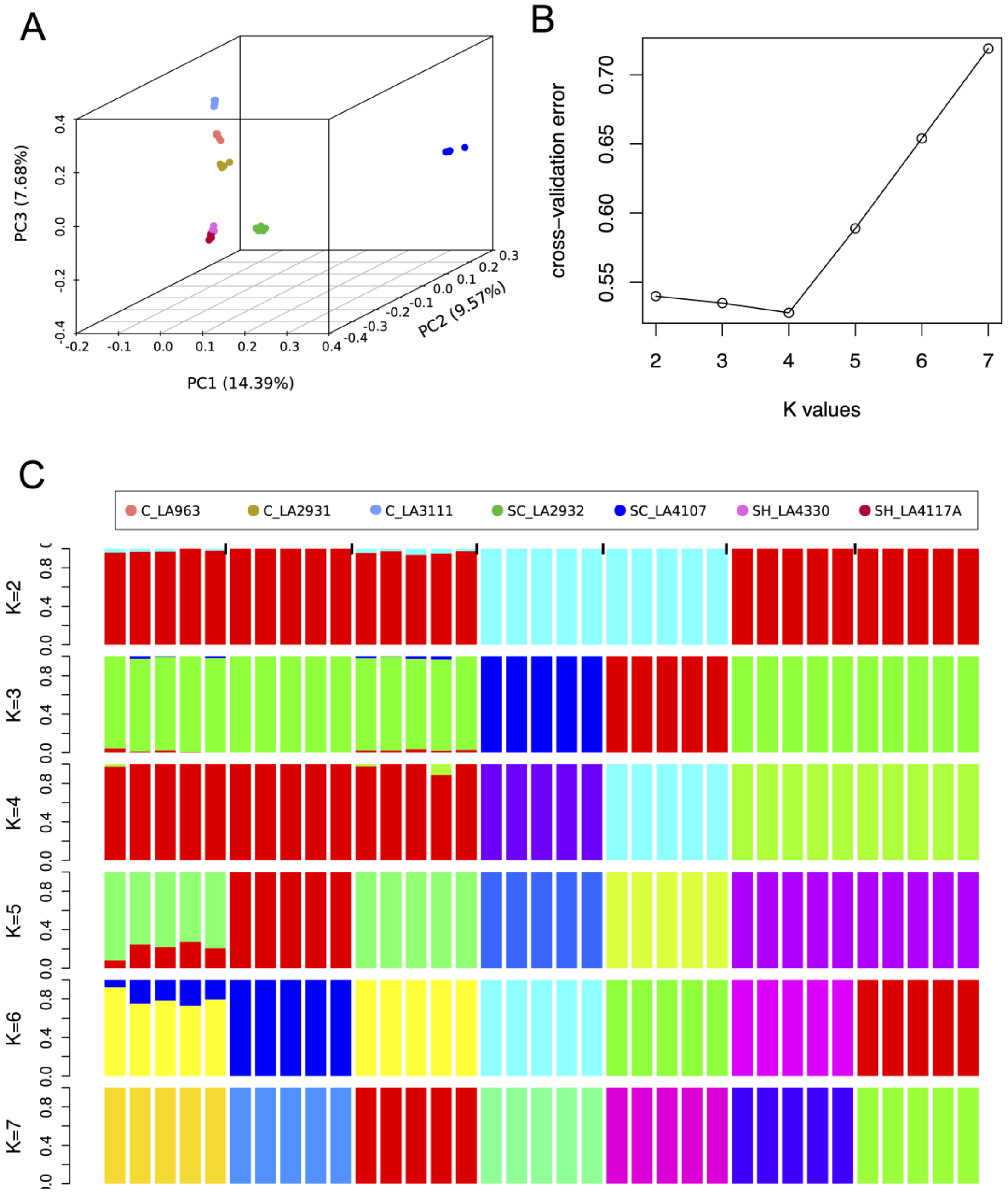

Figure S3. The population structure. (A) PCA based on genome-wide SNPs. (B) the cross-validation error based on different K values. (C) The admixture based on genome-wide SNPs.

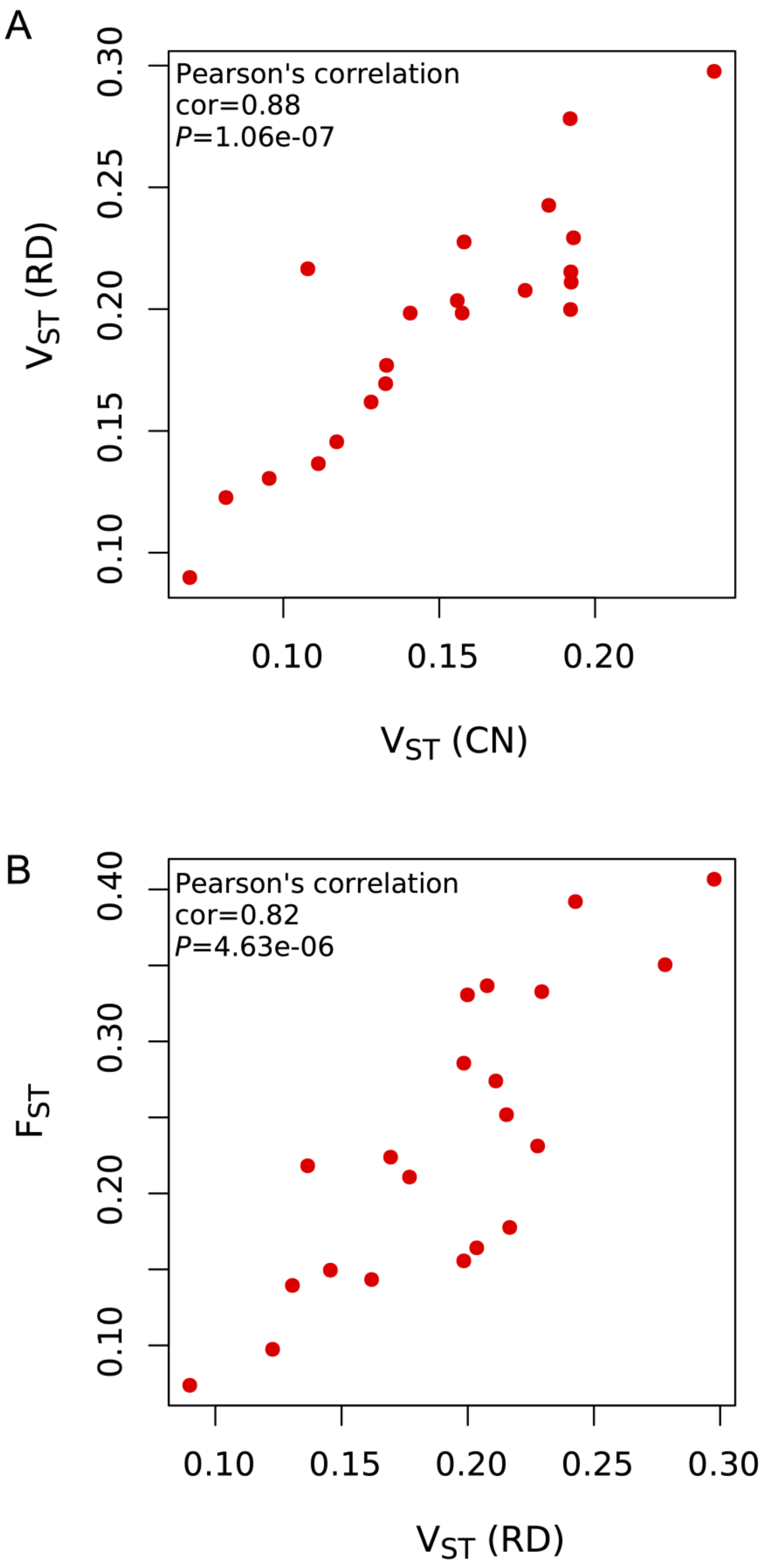

Figure S4. The correlations between V_{ST}(CN) and V_{ST}(RD) (A) and between V_{ST}(RD) and F_{ST} (B).

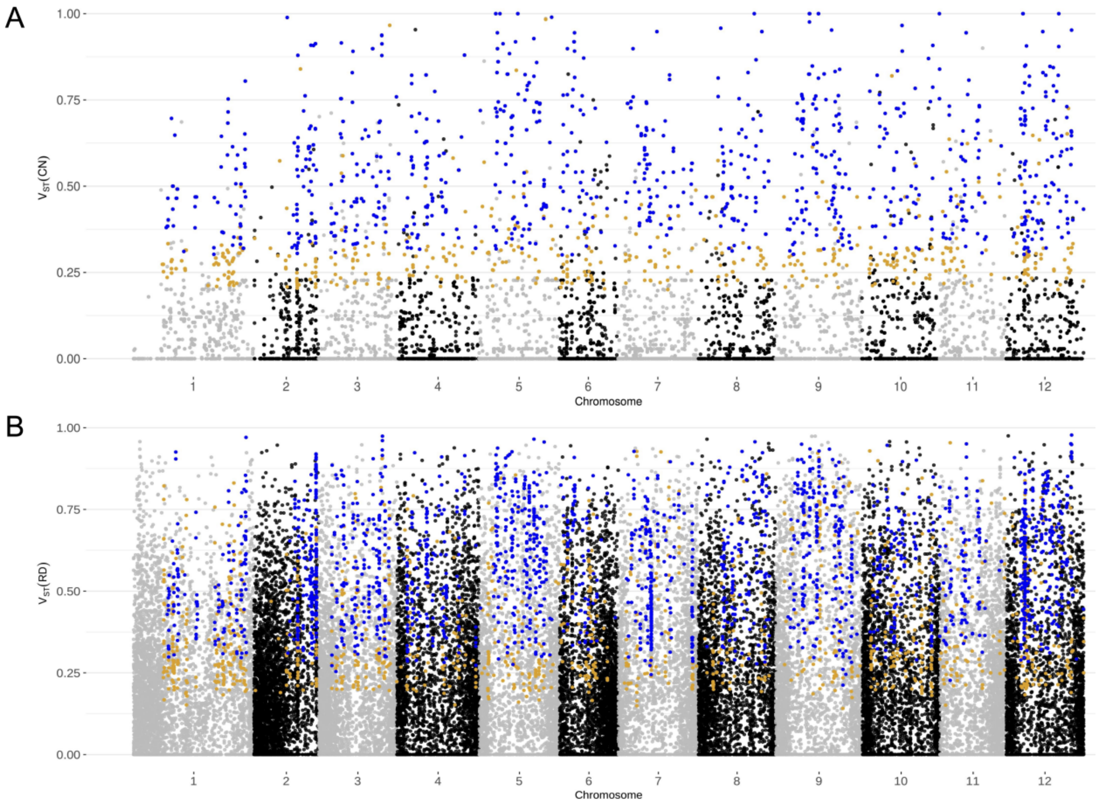

Figure S5. The identification of the differentiated genes based on VST(CN) (A) and VST(RD) (B) using gene copy number quantified by Control-FREEC and Read Depth, respectively. The orange dots denote the differentiated genes based 95th percentile using the permuted test (1,000 times), the blue dots denote the strongly differentiated genes based 99th percentile using the permuted test (1,000 times).

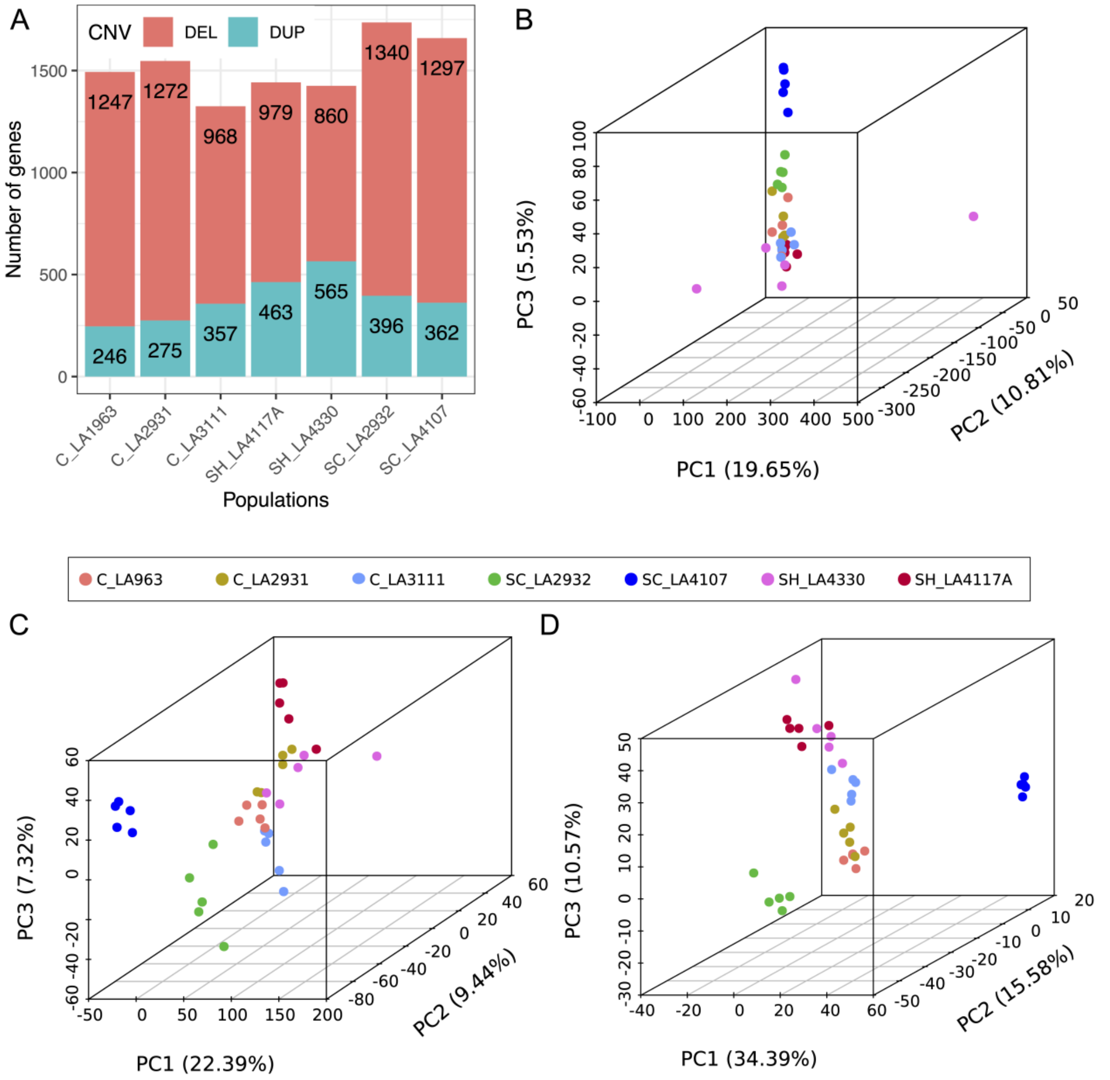

Figure S6. Genes with differentiated CN profiles among seven populations. (A) The distributions of 3,539 CN differentiated genes located at deletion (DEL) and duplication (DUP) regions in seven populations. (B) PCA based on CN values of 23,911 genes with CN values. (C) PCA based on CN values of 12,392 CN-variable genes with V_{ST}(CN) > 0. (D) PCA based on 2,192 strongly differentiated gene CN values.

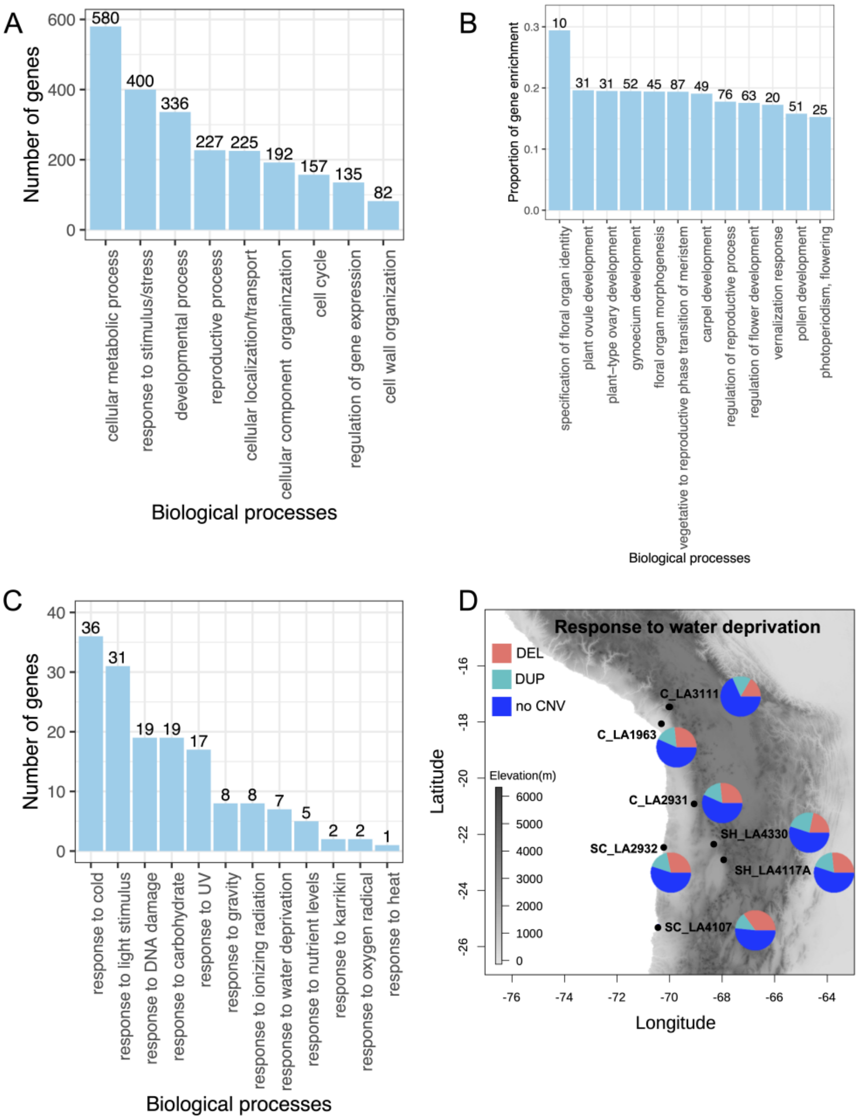

Figure S7. GO enrichment analysis of 3,539 CN differentiated genes. (A) The summary of GO enrichment. All significantly enriched GO terms (*P* < 0.05) were assigned into nine categories. (B) The proportions of differentiated genes enriched in different reproductive processes. The proportion of gene enrichment is equal to the number of genes enriched in one GO category divided by the number of background genes in this category. The number on top of each bar represents the number of genes enriched in that GO category. (C) The number of genes responded to external stimulus/stresses overlapping with genes enriched in reproductive process. (D) The 60 CN differentiated genes involved in response to water deprivation with deletion (DEL) and duplication (DUP) in seven populations, respectively. The pie charts denote number of CN differentiated genes with DEL, DUP or no CNV (see also Table S8).

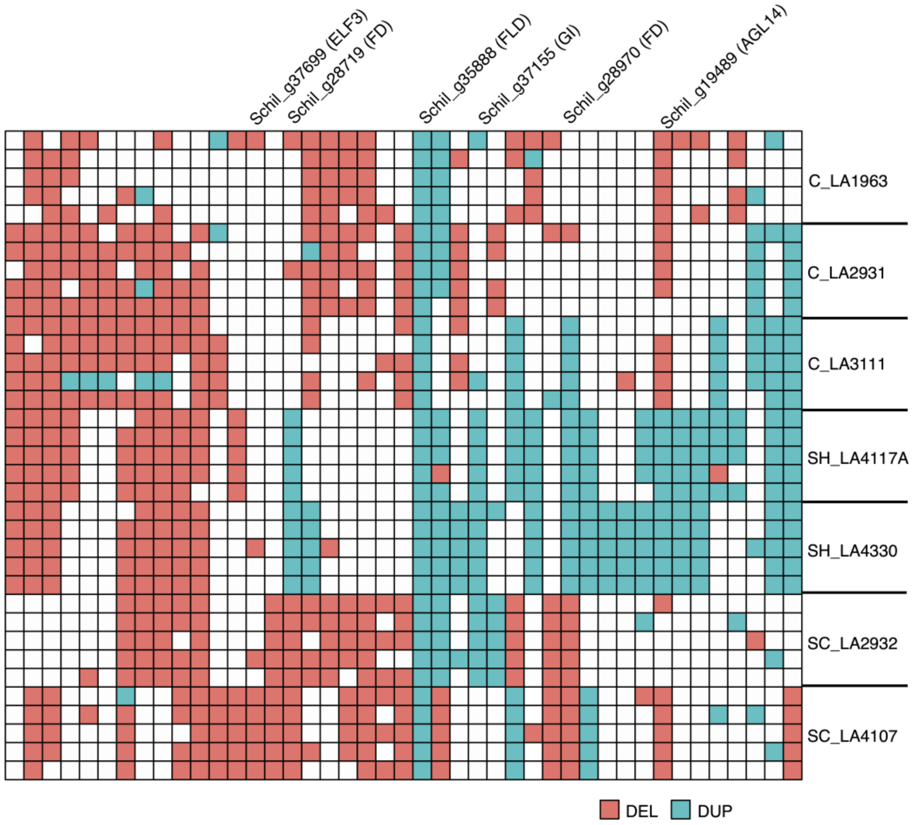

Figure S8. CN-differentiated genes enriched in photoperiod and vernalization pathways show the different CNV patterns across seven populations. Few putative homologs (see also Dataset S5) are validated as flowering regulatory genes in other plant species. White boxes indicate genes without CNV.

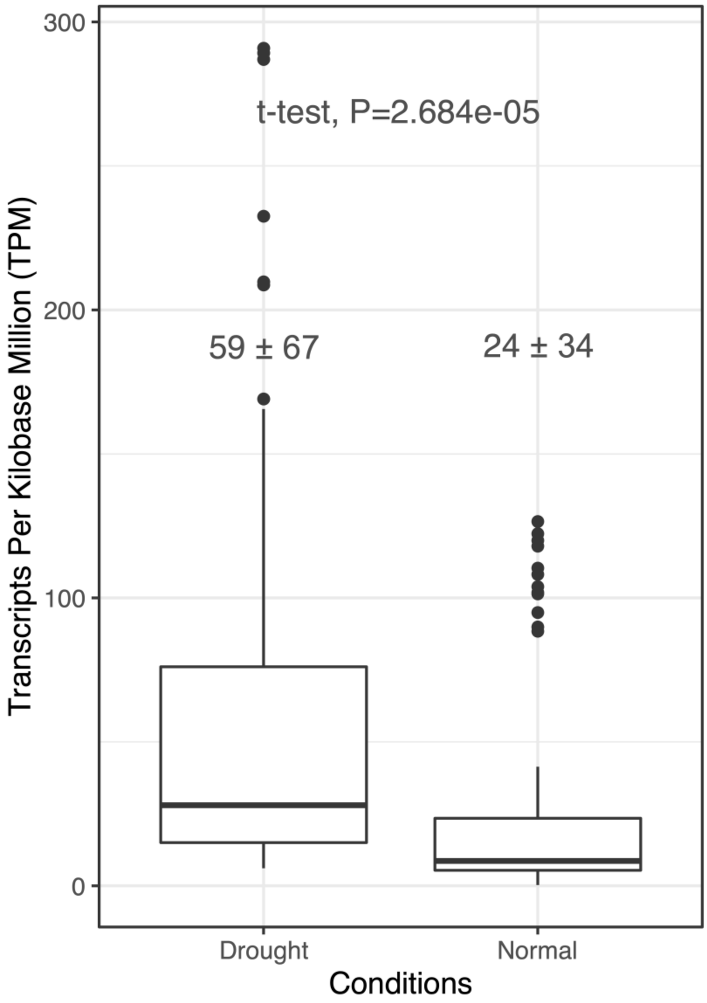

Figure S9. Gene expression of 11 genes responded to water deprivation overlapped with drought-responsive genes in metabolic network from the previous study of transcriptomes (Wei et al. 2023a).

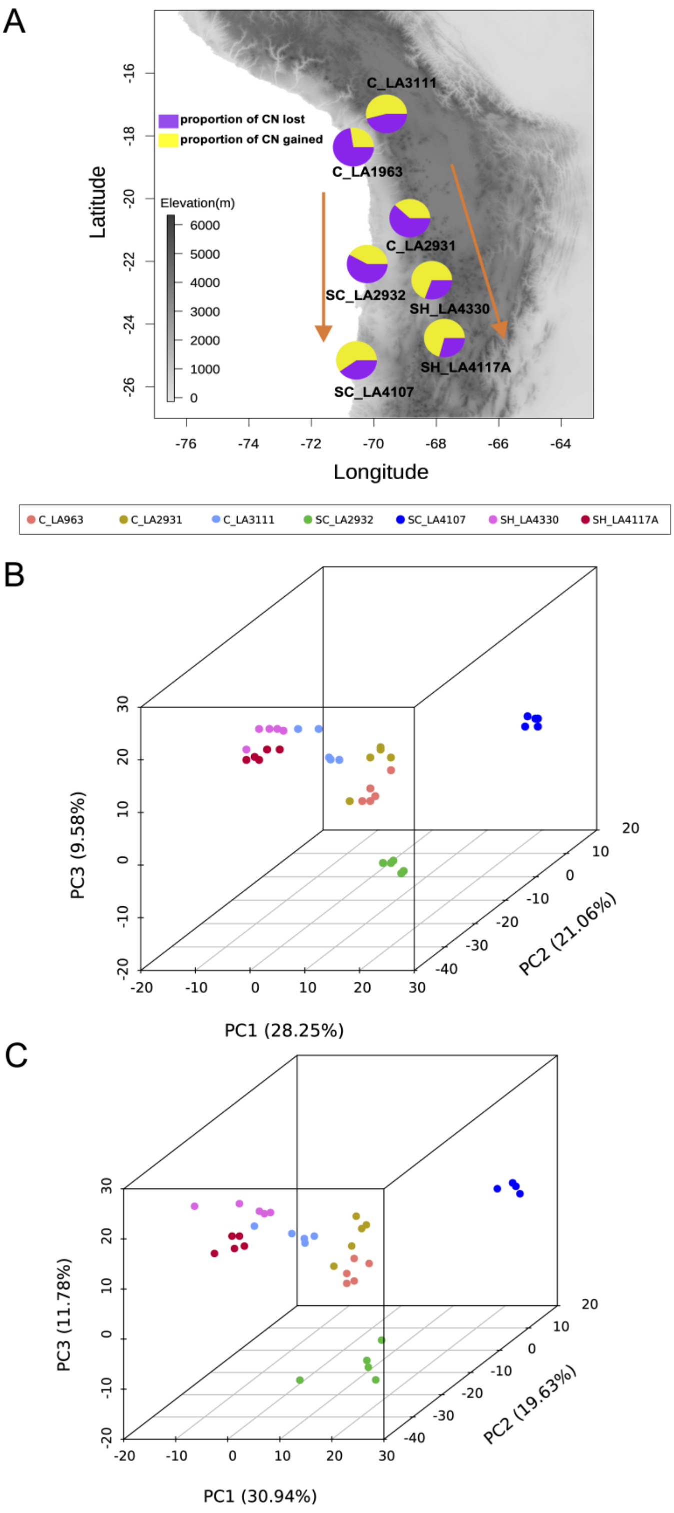

Figure S10. (A) The map and pie charts show that the dynamics of CN lost and gained in the processes of two southward colonization events, first to the southern coast (SC) and second to the southern highland (SH) (orange arrows) using the reference genome of *S. pennellii* (see also Table S9). (B) PCA based on CN of rapidly evolving genes with significant CN expansion or contraction (Viterbi P < 0.05) using the reference genome of *S. chilense*. (C) PCA based on CN of rapidly evolving genes with significant CN expansion or contraction (Viterbi P < 0.05) using the reference genome of *S. pennellii*.

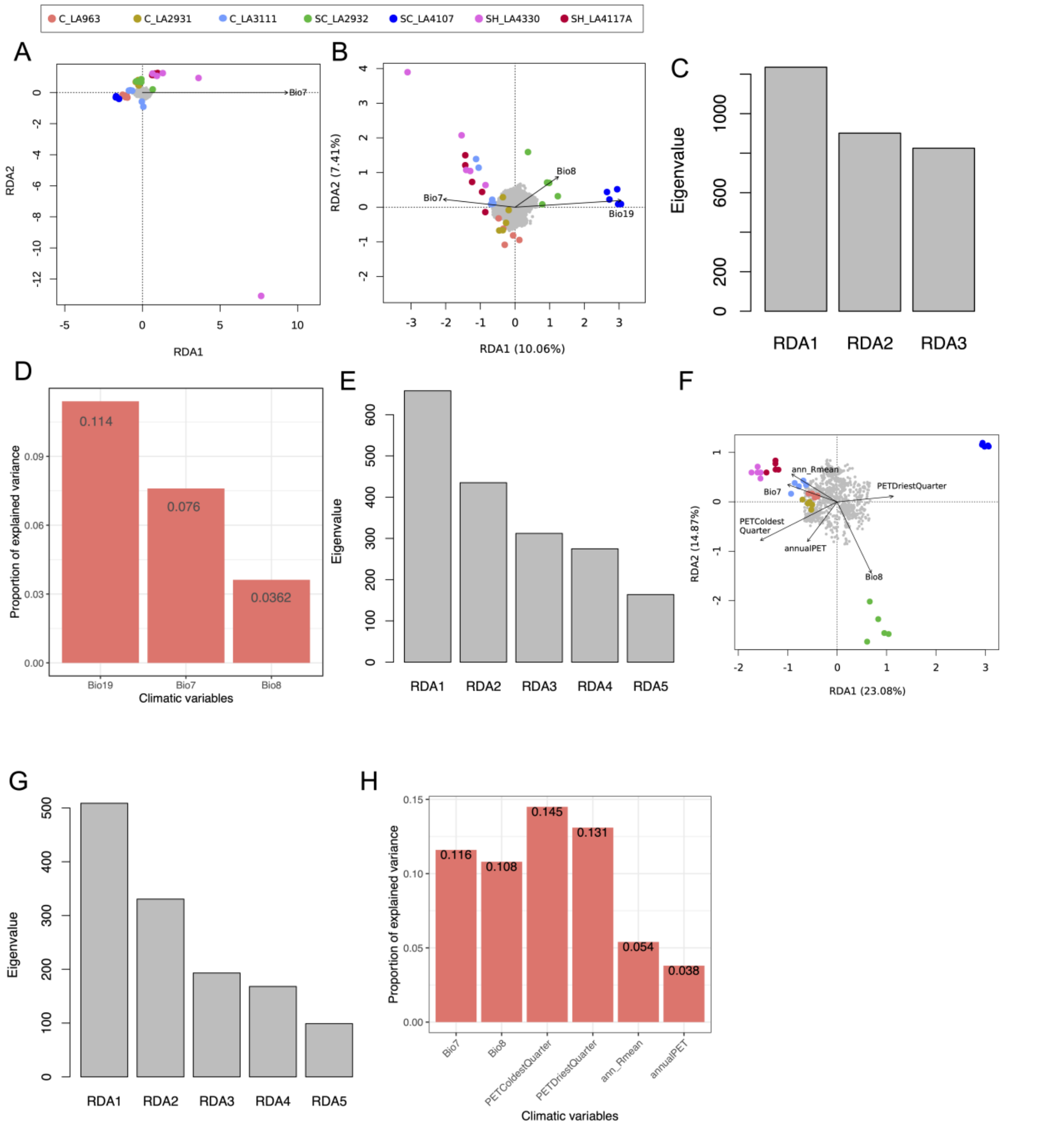

Figure S11. (A) RDA model shows no significant result based on CN of 23,911 genes. (B) RDA model based on CN of 12,391 genes with V_{ST} >0. (C) The eigenvalues of three significant ordination axes in RDA model based on CN of 12,391 genes with V_{ST} >0. (D) The proportion of explained variance of three overrepresented climate variables in RDA model based on CN of 12,391 genes with V_{ST} >0. (E) The eigenvalues of five significant ordination axes in RDA model based on CN of 3,539 differentiated genes. (F) RDA model based on CN of 2,192 strongly differentiated genes. (G) The eigenvalues of five significant ordination axes in RDA model based on CN of 2,192 strongly differentiated genes. (H) The proportion of explained variance of six overrepresented climate variables in RDA model based on CN of 2,192 strongly differentiated genes. In RDA models, the loading of the climatic variables or the length of the vector indicates the strength of the correlation with the ordination axis. Vectors of climate variables pointing in the same direction that populations indicate a high positive correlation, vector pointing at right angles indicate no correlation, and vectors pointing in opposite directions indicate high negative correlations. The grey dots represent genes. Colored dots represent different populations.

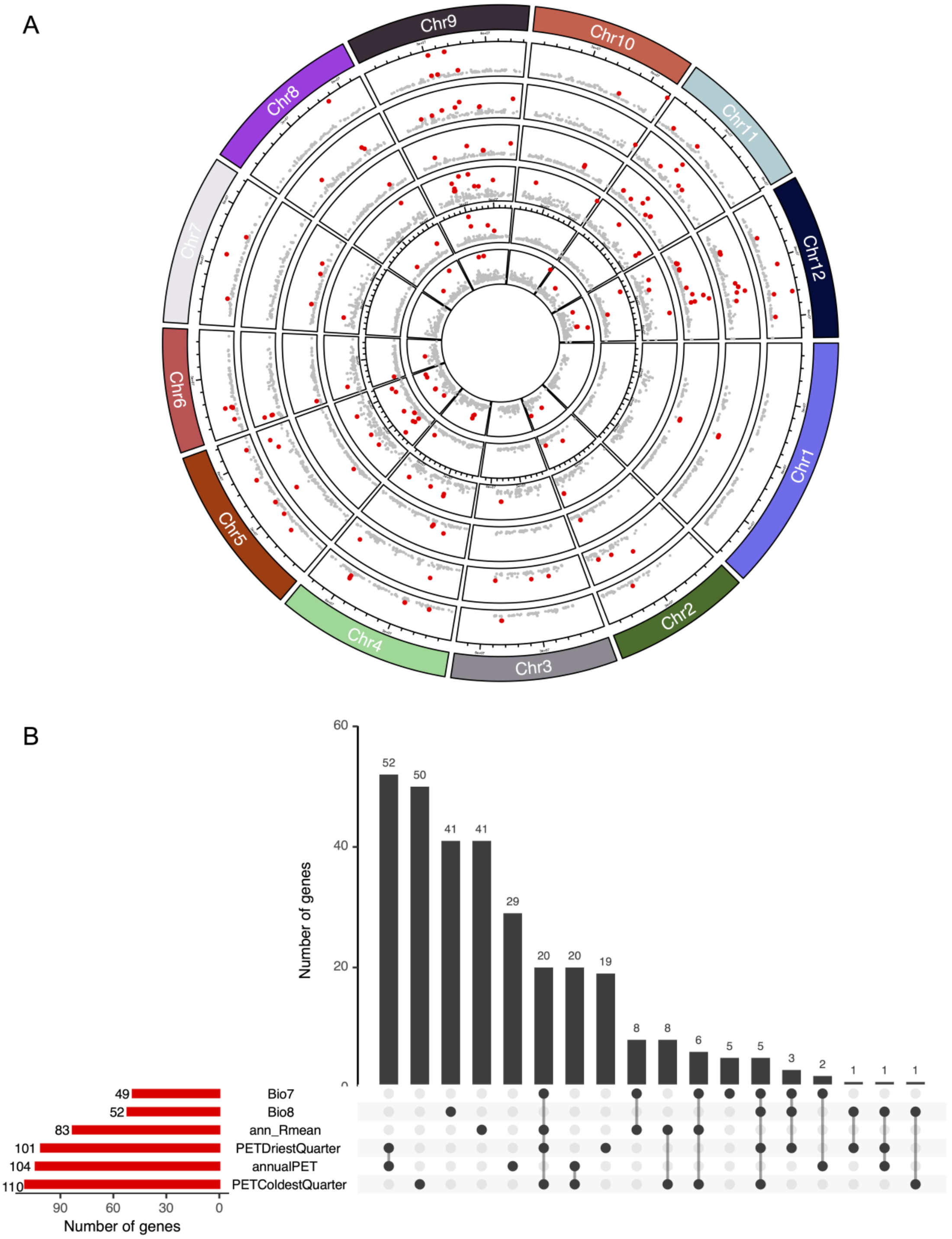

Figure S12. (A) The analyses of gene CN associated with six climatic variables using LFMM2. The circles from inside to outside are Bio7, Bio8, ann_Rmean, PETDriestQuarter, annualPET, PETColdestQuarter, respectively. (B) The number of candidate genes associated with six climatic variables (red bar), respectively, and shared candidate genes across climatic variables (black bar).

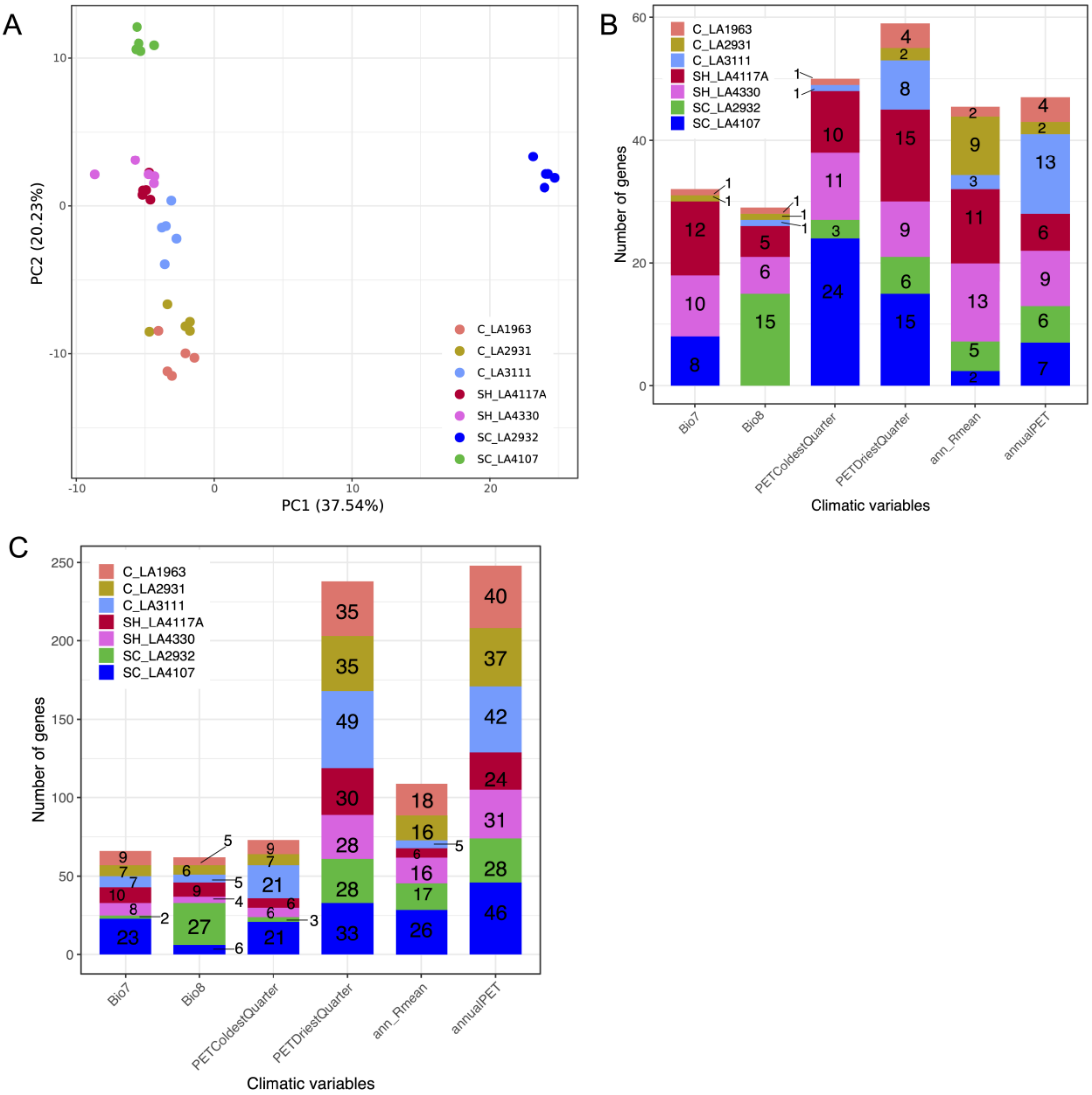

Figure S13. The GEA using LFMM2. (A) PCA based on CN of 312 candidate genes identified by LFMM2 correlated with six climatic variables. (B) The number of candidate genes identified by LFMM2 located at duplication (DUP) regions in seven populations. (C) The number of candidate genes identified by LFMM2 located at deletion (DEL) regions in seven populations.

Table S1. The number of deletion (DEL) and duplication (DUP) in each population

Populations	DEL	DUP
C_LA1963	68393	26228
C_LA2931	66859	25970
C_LA3111	58859	25133
SC_LA2932	56914	22576
SC_LA4107	51849	21165
SH_LA4117A	54314	22504
SH_LA4330	55661	23117

Table S2. The number of deletion (DEL) and duplication (DUP) in each accession

Groups	Populations	Accessions	DEL	DUP	Total
central	C_LA1963	C_LA1963_t1	31891	11736	43627
central	C_LA1963	C_LA1963_t2	28596	10605	39201
central	C_LA1963	C_LA1963_t5	31627	11908	43535
central	C_LA1963	C_LA1963_t7	29069	10913	39982
central	C_LA1963	C_LA1963_t9	31312	11679	42991
central	C_LA2931	C_LA2931_t2	30586	11342	41928
central	C_LA2931	C_LA2931_t3	30871	11334	42205
central	C_LA2931	C_LA2931_t4	31676	11809	43485
central	C_LA2931	C_LA2931_t5	32074	11730	43804
central	C_LA2931	C_LA2931_t6	31133	11556	42689
central	C_LA3111	C_LA3111_t3	28357	11789	40146
central	C_LA3111	C_LA3111_t5	30898	12507	43405
central	C_LA3111	C_LA3111_t9	25406	10597	36003
central	C_LA3111	C_LA3111_t10	27963	11594	39557
central	C_LA3111	C_LA3111_t15	27623	11765	39388
southern coast	SC_LA2932	SC_LA2932_1	34249	12330	46579
southern coast	SC_LA2932	SC_LA2932_8	31684	11379	43063
southern coast	SC_LA2932	SC_LA2932_t2	32097	11514	43611
southern coast	SC_LA2932	SC_LA2932_20	31230	11074	42304
southern coast	SC_LA2932	SC_LA2932_22	30155	10765	40920
southern coast	SC_LA4107	SC_LA4107_3	30991	11103	42094
southern coast	SC_LA4107	SC_LA4107_6	33141	12096	45237
southern coast	SC_LA4107	SC_LA4107_9	32849	11761	44610
southern coast	SC_LA4107	SC_LA4107_t5	31243	11456	42699
southern coast	SC_LA4107	SC_LA4107_t11	30818	11006	41824
southern highland	SH_LA4117A	SH_LA4117A_1	27913	9391	37304
southern highland	SH_LA4117A	SH_LA4117A_4	28128	10452	38580
southern highland	SH_LA4117A	SH_LA4117A_5	28858	9865	38723
southern highland	SH_LA4117A	SH_LA4117A_10	27772	10131	37903
southern highland	SH_LA4117A	SH_LA4117A_15	24905	7018	31923
southern highland	SH_LA4330	SH_LA4330_t1	29186	11019	40205
southern highland	SH_LA4330	SH_LA4330_t4	32217	12281	44498
southern highland	SH_LA4330	SH_LA4330_t6	26220	9823	36043
southern highland	SH_LA4330	SH_LA4330_t9	31562	11964	43526
southern highland	SH_LA4330	SH_LA4330_t12	30934	11610	42544

Table S3. The number of deletion (DEL) and duplication (DUP) identified in different numbers of accessions (1 to 5)

	DEL					DUP				
Populations	1	2	3	4	5	1	2	3	4	5
C_LA1963	29655	14361	8856	7025	8496	11222	5330	3541	2904	3231
C_LA2931	26987	13566	9337	7540	9429	10654	5204	3680	2918	3514
C_LA3111	21943	12756	8982	6990	8188	9471	5165	3767	3064	3666
SC_LA2932	16850	9546	7218	7852	14748	7469	4031	2987	3101	4788
SC_LA4107	13530	7604	6436	6971	17308	6399	3482	2825	2861	5598
SH_LA4117A	16031	9824	7891	6360	11208	7562	4242	3085	2585	3030
SH_LA4330	17009	10618	8547	7981	11506	7996	4528	3320	3047	4226

Table S4. The number of deletion (DEL) and duplication (DUP) overlapping different genomic features

	Gene		Intergenic		Exon		Intron		5kb upstream		5kb downstream	
Populations	DEL	DUP	DEL	DUP	DEL	DUP	DEL	DUP	DEL	DUP	DEL	DUP
C_LA1963	21672	12519	59522	23183	15857	11691	18869	10848	28199	13711	34051	15037
C_LA2931	21179	12465	58187	22915	15562	11681	18356	10774	27443	13653	33338	14991
C_LA3111	19582	12199	51632	22415	15096	11586	16899	10516	25201	13282	30024	14473
SC_LA2932	17478	10621	49333	19655	13448	10049	15018	9163	22927	11636	27365	12631
SC_LA4107	16319	10128	45631	18590	12712	9583	14034	8765	21296	11053	25220	12031
SH_LA4117A	16677	10590	47525	20331	12039	9970	14250	8795	22691	11555	26246	12527
SH_LA4330	17871	11132	48811	20468	13742	10565	15464	9627	23292	12165	27540	13197

Table S5. The validation of pipeline of CNV calling using 1,000 simulated deletions (DELs) and duplications (DUPs), respectively.

CNV caller	Number of DEL	Number of DUP
Lumpy	878 (21)	842(13)
Manta	795 (29)	774(36)
Wham	767 (33)	698 (19)
Delly	849 (40)	861 (22)
SURVIVOR merged	918(12)	879 (4)

Numbers in parentheses indicate the number of incorrect CNVs (false-positive).

Table S6. The measures of population differentiation based on copy number (V_{ST}(RD) and V_{ST}(CN)) and SNPs (F_{ST}).

Pairwise populations	V_{ST}(RD)	V_{ST}(CN)	F_{ST}
C_LA1963 vs C_LA2931	0.090 ±0.153	0.070±0.125	0.074±0.131
C_LA1963 vs C_LA3111	0.123±0.177	0.082±0.138	0.097±0.146
C_LA2931 vs C_LA3111	0.131±0.194	0.095±0.162	0.139±0.155
SH_LA4117A vs SH_LA4330	0.217±0.239	0.108±0.188	0.178±0.247
C_LA3111 vs SH_LA4330	0.137±0.212	0.111±0.177	0.218±0.227
C_LA2931 vs SH_LA4330	0.146±0.198	0.117±0.165	0.150±0.163
C_LA1963 vs SH_LA4330	0.162±0.203	0.128±0.190	0.143±0.147
C_LA2931 vs SC_LA2932	0.169±0.245	0.133±0.134	0.224±0.270
C_LA1963 vs SC_LA2932	0.177±0.248	0.133±0.174	0.211±0.196
C_LA1963 vs SH_LA4117A	0.198±0.244	0.141±0.189	0.156±0.149
C_LA2931 vs SH_LA4117A	0.204±0.248	0.156±0.206	0.164±0.172
SC_LA2932 vs C_LA3111	0.198±0.276	0.157±0.196	0.286±0.311
C_LA3111 vs SH_LA4117A	0.228±0.257	0.158±0.218	0.231±0.255
SC_LA2932 vs SH_LA4330	0.208±0.267	0.178±0.230	0.337±0.382
SC_LA4107 vs SH_LA4330	0.243±0.291	0.185±0.259	0.392±0.431
C_LA1963 vs SC_LA4107	0.215±0.272	0.192±0.236	0.252±0.277
C_LA2931 vs SC_LA4107	0.211±0.275	0.192±0.244	0.274±0.278
SC_LA2932 vs SC_LA4107	0.200±0.281	0.192±0.254	0.331±0.337
SC_LA2932 vs SH_LA4117A	0.278±0.300	0.192±0.259	0.350±0.414
C_LA3111 vs SC_LA4107	0.229±0.284	0.193±0.262	0.333±0.401
SC_LA4107 vs SH_LA4117A	0.298±0.315	0.238±0.287	0.407±0.448

Table S7. The number of candidate genes with differentiated gene CN across seven populations

	Differentiated (95th percentile)		Extremely differentiated (99th percentile)	
V_{ST}	Threshold	Number of genes	Threshold	Number of genes
V_{ST}(CN)	0.194	4,843	0.305	3,219
V_{ST}(RD)	0.157	16,655	0.244	12,228
Overlaps		3,539		2,192

The differentiated and extremely genes were identified using 95th and 99th percentile in 1,000 permutation tests, respectively.

Table S8. The number of CN differentiated genes involved in four GO terms in seven populations located in deletion (DEL) and duplication (DUP) regions.

Populations	photoperiod			vernalization			response to water deprivation			root development		
	DEL	DUP	no CNV	DEL	DUP	no CNV	DEL	DUP	no CNV	DEL	DUP	no CNV
C_LA1963	6	1	18	8	1	11	16	10	34	26	17	30
C_LA2931	10	1	14	9	3	8	16	10	34	12	4	57
C_LA3111	8	2	15	7	5	8	10	9	41	11	8	54
SC_LA2932	10	2	13	7	2	11	17	10	33	27	18	28
SC_LA4107	9	2	14	10	1	9	21	8	31	25	18	30
SH_LA4117A	5	8	12	5	8	7	16	11	33	14	7	52
SH_LA4330	4	10	11	4	7	9	13	14	33	12	6	55

Table S9. The summary of CN expansion and contraction in different branches/populations using the reference of *S. pennellii*.

Groups/Populations	Number of CN expanded genes	Number of CN contracted genes	Number of CN gained	number of CN contracted	[a]Rate of average expansion/ contraction	[b]Number of rapidly evolving genes
inland	63	48	344	86	2.324	28(+21/-7)
Central (C)	186	568	427	937	-0.676	35(+15/-20)
southern highland (SH)	522	384	1506	648	0.947	51(+38/-13)
southern coast (SC)	67	115	184	211	-0.1483	11(+4/-7)
C_LA1963	149	456	322	851	-0.874	27(+9/-18)
C_LA2931	116	292	360	587	-0.556	14(+3/-11)
C_LA3111	215	374	762	646	0.197	19(+11/-8)
SH_LA4117A	627	484	2007	824	1.065	54(+39/-15)
SH_LA4330	471	304	1340	588	0.970	43(+43/-0)
SC_LA2932	215	559	747	1035	-0.372	17(+5/-12)
SC_LA4107	384	306	1132	692	0.638	38(+26/-12)

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

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

Table S10. The number of CN differentiated genes associated with temperature annual range (Bio7) and solar radiation (ann_Rmean) in seven populations located in deletion (DEL) and duplication (DUP) regions.

Populations	DEL	DUP	no CNV
C_LA1963	7	1	26
C_LA2931	6	3	25
C_LA3111	2	1	31
SC_LA2932	7	2	25
SC_LA4107	10	1	23
SH_LA4117A	2	5	27
SH_LA4330	6	6	22