Supplementary Text: full materials and methods

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

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

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

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

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

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

Full Materials and Methods

Sequence Read Processing

We ~~used 35~~ whole-genome paired-end Illumina ~~data for~~ seven populations of *S. chilense* (five plants for each population) representing different geographic groups ~~available on~~ European Nucleotide Achieve (ENA) BioProject PRJEB47577. We ~~performed~~ the same pipeline of read processing as ~~the~~ previous study (Wei, et al. 2023) for quality trimming and mapping to reference genome of *S. chilense* (Silva-Arias, et al. 2023). SNPs ~~were also called and filtered~~ based on the *S. chilense* genome using the same pipeline in ~~a previous~~ study (Wei, et al. 2023).

Identification and genotyping of copy number ~~variations~~

To obtain high-confidence CNVs, including deletions (DELs) and duplications (DUPs), we chose four ~~CNV detection tools to perform CNV calling based on a comprehensive evaluation of structural variation detection algorithms (Kosugi, et al. 2019).~~ They enumerated potential good algorithms for each SV category, among which LUMPY (Layer, et al. 2014), Manta (Chen, et al. 2016), Wham (Kronenberg, et al. 2015) and DELLY (Rausch, et al. 2012) are better algorithms in deletion or duplication categories. These tools are combined with multiple algorithms to detect CNVs using whole-genome sequencing data, including read depth, paired-end mapping, split read and *de novo* assembly approaches.

In 35 samples, for Lumpy v0.3.1, we first extracted the discordant paired-end reads with abnormal insertion size from mapped results using the *view* function of Samtools v1.7 (Wysoker, et al. 2009), and the split-read alignments also were extracted using 'extractSplitReads_BwaMem' script in Lumpy package. The output BAM files were sorted using the *sort* function of Samtools. We then run Lumpy using the mapped reads, discordant paired-end reads, and split reads as inputs to detect CNVs. DELLY v0.7.6 was run using default parameters, and then outputted bcf file was converted into a vcf file using bcftools v1.9 (Danecek, et al. 2011; Danecek, et al. 2021). Furthermore, Manta v1.6 and Wham v1.8 were run using default parameters. For each accession, CNV call sets from LUMPY, DELLY, Manta and Wham were then merged with SURIVOR v1.0.7 (Jeffares, et al. 2017). We set minimum CNV length as 50bp, maximum CNV length as 1Mb, minimum distance of 1,000bp, and types must match. Only CNVs called by at least two of the four tools were retained. The merged CNV set was inputted to SVTyper v0.7.0 to call genotypes,

respectively, for population genetics analysis using a Bayesian algorithm (Chiang, et al. 2015). SVTyper performs breakpoint genotyping of structural variants using whole genome sequencing data. It assesses discordant and concordant reads from paired-end and split-read alignments to infer genotypes at each site. The pipeline included all command lines and parameters of CNV calling, merging, and genotyping, which can be found on our Gitlab repository: https://gitlab.lrz.de/population_genetics/s_chilense_cnv.

To verify the sensitivity and accuracy of our pipeline of CNV calling, we simulated NGS data using a Python script 'CNV-Sim' obtained on https://github.com/NabaviLab/CNV-Sim. It extends the functionality of existing NGS read simulators to introduce CNVs in the generated reads. We run CNV-Sim v0.9.2 in whole genome, which utilises the functionality of ART (Huang, et al. 2012) to introduce CNVs in the genome. We simulated 1,000 duplication and 1,000 deletion regions ranging from 50bp to 1Mb based on 150 bp short-reads. Then, these simulated reads were inputted into the same pipelines to identify CNVs. The command lines of simulation can be found at: https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-/blob/main/CNVs_simulation.

Population structure analysis

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

Quantification of gene copy number

We employed two strategies to quantify gene copy number (CN). First, we used the read-depth-based method implemented in Control-FREEC v11.6 to estimate copy numbers (CNs) by 10 kb windows with 1 kb step size across the entire genome (Boeva, et al. 2012). We used the following parameters in Control-

FREEC: ploidy=2, breakPointThreshold = 0.8, degree=3, minExpectedGC = 0.3, maxExpectedGC = 0.55, and telocentromeric=0. We then obtained gene CN from the Control-FREEC outputs, and gene coordinates in the genome. However, some genes were observed to have more than one CN estimate. These events may be due to imperfect estimation of breakpoints using our window size and sliding window. We calculated the average CN if one gene corresponds to multiple CNs.

Additionally, we also employed another strategy to calculate gene CN. For each sample, we estimated read depth using Mosdepth v0.3.2 (Pedersen and Quinlan 2018) by 1,000 bp sliding windows, and gene read depth was calculated from gene coordinates. We then used median read-depth values of all windows and genes as a normalising factor to obtain the final window and gene CN estimate, respectively, and the formula as CN = (read depth / median value) × 2.

Identification of candidate genes associated with population differentiation

We calculated V_{ST} to identify genes with divergent CN profiles among seven populations. The V_{ST} measurement, analogous to F_{ST}, is applied to identify loci that differentiate by CN between populations (Redon, et al. 2006). Both V_{ST} and F_{ST} consider how genetic variation acts on the differentiation of populations or closely related species and range from 0 (no differentiation) to 1 (complete differentiation). Using a sliding window-based approach, we first calculated pairwise FST and VST to compare the efficacy of population differentiation estimated by SNPs and CNVs. The F_{ST} was calculated for each pair of populations using VCFtools over a 1 kb non-overlapping sliding window. V_{ST} for each pair of populations was calculated by considering $(V_T-V_S)/V_T$, where V_T is the total variance among all individuals, and V_S is the average variance within each population, weighted for sample size (Redon, et al. 2006). V_{ST} was also calculated based on CNs of 1kb sliding windows across the reference genome. In addition, we calculated two V_{ST} data sets from two different CN quantitative strategies for each pair of populations, respectively.

After assessing the strength of the effect of copy number changes on population differentiation, we identified candidate genes related to population differentiation based on strategy from Rinker et al. 2019. Similar to the sliding window-based method, the V_{ST} value of each gene was independently calculated based on two CN quantifications. An R script shows the pipeline of V_{ST} calculation and identification of candidate genes (R script available at https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-

/blob/main/VST.R). We performed permutation tests on the CN counts to identify which genes displayed the most significant degree of observed inter-population CN variation that was likely not due to sampling bias. Here, we randomly permuted gene CNs of each gene for 35 individuals and calculated a new V_{ST} for every permutation and every gene, respectively. ~~This process (permutation test)~~ was repeated 1,000 times, creating a random distribution of V_{ST} values for each gene. We then selected candidate genes ~~that~~ observed V_{ST} fell above the 95th and 99th percentile of the permuted V_{ST} distribution. These candidate genes displayed substantial intra-population CN homogeneity and high degrees of inter-population differentiation. Finally, genes were considered significant when observed V_{ST} values were above the maximum 95% (differentiated) or 99% (extremely differentiated) confidence interval cutoff in both gene CN estimate methods.

Gene ontology (GO) analysis

We first performed a blast of our genes to the *A. thaliana* dataset TAIR10 (e-value cutoff was 10-6) (Camacho, et al. 2009; Berardini, et al. 2015). The most matching entry (lowest e-value) was selected as the target homologue for enrichment analysis. We used the R package clusterProfiler to perform GO enrichment analysis using the *A. thaliana* annotation database as the background (Yu, et al. 2012). The Benjamini-Hochberg method, a false discovery rate (FDR) method, was used to calibrate initial *P* values, and calibrated *P* values smaller than 0.05 were used as the cutoff for a significant level to obtain final GO terms.

Expansion and contraction of gene copy number

To gain insight into how copy numbers of these differentiated genes vary across populations, we analysed gene CNs expansion and contraction with 3,359 differentiated genes. We first calculated the mean CN for each gene for each population. We then constructed a population-based phylogenetic tree using SNPs by TreeMix v1.13 (Pickrell and Pritchard 2012), and finally the ultrametric tree was generated based on *force.ultrametric* function of phytools R package (Revell 2012). The ultrametric tree can be obtained at our GitLab repository https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-/blob/main/ultrametric_tree.nwk.

Finally, we analyse gene CN expansion and contraction in different groups using CAFE v4.2.1 (Han, et al. 2013) with the same *lambda* (the rate of change of evolution in a tree). We first run CAFE for genes with CN less than 100 to calculate an accurate *lambda* value (λ=0.00206736781311 in this study) because genes with large CNs can lead to non-informative parameter estimates. We then run CAFE for genes with CN larger than 100 using the same *lambda* value. The Viterbi method obtains the branch-specific P values with the randomly generated likelihood distribution. A low p-value indicates a rapidly evolving branch. Viterbi *P* values were computed for each significant gene to assess significant expansion or contraction along a specific branch. We set a p-value smaller ~~than~~ 0.05 to detect gene CN with a significantly greater rate of evolution (expansion or contraction) in different groups/populations.

Associated analysis between gene copy number and climatic conditions

The environmental data include 37 climatic ~~layers~~ obtained from two public databases, WorldClim2 (Fick and Hijmans 2017) and ENVIREM (Title and Bemmels 2018) (Dataset S6). To evaluate the relative contribution of the abiotic environment to explaining patterns of genetic variation, we first ~~used the~~ Redundancy Analysis (RDA) (Capblancq and Forester 2021) to associate CNs of 3,539 differentiated genes with climatic variables. RDA analyses were performed with an individual-based approach, using as input CNs for each differentiated gene for each sample. RDA was performed using the *rda* function from the vegan package as implemented in R (Forester, et al. 2018), modelling CNs as a function of predictor variables and producing constrained axes and representative predictors (climatic variables). All variables were centred and scaled before running the CN-environment association test. Multi-collinearity between representative predictors was assessed using the variance inflation factor (VIF), and since all predictor variables showed VIF < 10, none were excluded. The loadings of the CNs in the ordination space determined which genes were candidates for being under local adaptation. The CN loadings were stored as specified in the RDA object. The significance of RDA-constrained axes was assessed using the *anova.cca* function (*P* < 0.001).

Unlike RDA, which is a multivariate ordination technique that can analyse many loci and environmental predictors simultaneously, LFMM (latent factor mixed models) is a univariate test (Frichot, et al. 2013; Caye, et al. 2019), which means that it builds a model for each gene or SNP and each predictor variable. We first performed the lfmm_ridge function implemented in the R library LFMM to obtain an object that

contains the latent variable score matrix using a K value of four latent factors (as evaluated from analysis of population structure) based on CNs of 3,539 differentiated genes and six representative climate variables (as obtained from RDA). Then, we perform association testing using the *lfmm_test* function. The Benjamini-Hochberg method was used to calibrate the P-value and set 0.01 as the significance threshold to obtain candidate genes associated with climatic variables.

Reference

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

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. genesis 53:474-485.

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

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

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

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

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

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

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM. 2021. Twelve years of SAMtools and BCFtools. Gigascience 10:giab008.

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

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

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

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

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

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

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

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

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

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

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

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics 81:559-575.

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

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W. 2006. Global variation in copy number in the human genome. Nature 444:444-454.

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

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

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

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

Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078-2079.

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

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

216