

Rearranged endogenized plant pararetroviruses as evidence of heritable RNA-based immunity

Adrian A. Valli^{1†}, Irene Gonzalo-Magro¹ and Diego H. Sanchez^{2†}

DHS ORCID iD: 0000-0002-4234-1124

[†] To whom correspondence shall be addressed: avalli@cnb.csic.es; diegosanchez@agro.uba.ar

¹ Centro Nacional de Biotecnología (CNB-CSIC). Calle Darwin 3, 28049 Madrid, Spain.

² IFEVA (CONICET-UBA), Facultad de Agronomía, Universidad de Buenos Aires. Av San Martín 4453, C1417DSE Buenos Aires, Argentina.

Adrian A. Valli: avalli@cnb.csic.es

Irene Gonzalo Magro: igonzalo@cnb.csic.es

Diego H. Sanchez: diegosanchez@agro.uba.ar

Supplemental Materials and Methods

Data sources

We retrieved from public repositories viral accessions representing the current eight genera of *Caulimoviridae* (<https://talk.ictvonline.org/> (Teycheney, et al. 2020)). These were *Banana streak virus* (NC 008018.1), *Blueberry red ringspot virus* (NC 003138.2), *Cacao swollen shoot virus* (NC 001574.1) *Carnation etched ring virus* (NC 003498.1), *Cassava vein mosaic virus* (NC 001648.1), *Cauliflower mosaic virus* (NC 001497.2), *Cestrum yellow leaf curling virus* (NC 004324.3), *Commelina yellow mottle virus* (NC 001343.1), *Horseradish latent virus* (NC 018858.1), *Petunia vein clearing virus* (NC 001839.2), *Rice tungro bacilliform virus* (NC 001914.1), *Rose yellow vein virus* (NC 020999.1), *Soybean chlorotic mottle virus* (NC 001739.2), *Sweet potato collusive virus* (NC 015328.1), *Sweet potato vein clearing virus* (MH 188860.1) and *Tobacco vein clearing virus* (NC 003378.1). Described *Florendovirus* were recovered from a specific report (Geering, et al. 2014). Scrutinized plant genomes were *S. lycopersicum* (ITAG4.0 (Hosmani, et al. 2019)), *S. pimpinellifolium* (LA2093 v1.5 (Wang, et al. 2020)), and PAS014479 and BGV006775 (Alonge, et al. 2020)), *S. pennellii* (Spenn v2.0 (Bolger, et al. 2014a)), *S. tuberosum* (DM v4.04 (Hardigan, et al. 2016)), *S. melongena* (Eggplant v3 (Barchi, et al. 2019)),

N. tabacum (Nitab v4.5 (Edwards, et al. 2017)), *N. benthamiana* (Niben v1.01 (Bombarely, et al. 2012)), *N. attenuata* (Niatt r2 (Xu, et al. 2017)), *G. max* (Gmax_508 v4.0 and ZH13 a1 (Schmutz, et al. 2010; Shen, et al. 2019)) and *G. soja* (PI483463 a1 and W05 a1 (Valliyodan, et al. 2019; Xie, et al. 2019)). *S. lycopersicum* genic features were recovered from ITAG4.1 annotation, whereas assemblies from other various accessions were recently reported (Brandywine, M82, Floradade, EA00371, LYC1410, EA00990, PI303721, PI169588, Fla.8924, BGV006865, BGV007989, and BGV007931; the last three *var. cerasiforme*) (Alonge, et al. 2020); all available at Solgenomics (www.solgenomics.net).

We obtained small-RNA libraries from publically available resources comprising data from different laboratories (Lunardon, et al. 2020) while *sldcl2ab* mutant libraries were from (Wang, et al. 2018), and further filtered for 18-25-nt sizes. *S. lycopersicum* RNA-seq tissue-specific libraries were reported previously (Sanchez, et al. 2019). PARE-seq raw data (Seo, et al. 2018) were analysed only for trimmed reads ≥ 14 bp. H3K9me2 and H3K9ac ChIP-seq and BS-seq raw data were publically available (Wang and Baulcombe 2020). Short-read DNA sequencing of different *S. lycopersicum* accessions were recovered from pan-genome and breeding history reports (Gao, et al. 2019; Lin, et al. 2014; Tomato Genome Sequencing, et al. 2014). In all cases, independent biological samples were merged to increase sequencing depth. [Supplemental Table S7](#) lists the different NGS datasets analyzed.

Bioinformatics analyses

De novo annotation of plant endogenous-pararetroviruses (EPRVs) was initiated with the search of apparent LTR retrotransposon sequences, taking advantage of EPRVs bearing terminal-repeats. The tool LTRharvest (Ellinghaus, et al. 2008), was used with parameters -v -mintsd 3 -maxtsd 6 -seed 30 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3 -minlenltr 100 -maxlenltr 7000 -mindistltr 1000 -maxdistltr 30000 -similar 97 -overlaps best -vic 60 -longoutput. The output was mined for sequences presenting an ORF > 100 amino-acids with significant similarity to Pfam PF01107.18 representing viral movement proteins (MP) (El-Gebali, et al. 2019); recognized applying HMMER3 (Eddy 2011) hmmscan function with parameter -T 40. Hits also presenting homologies to Pfam entries related to GAG and integrases from transposable

elements were filtered out. The outcome was used as input for the initial pararetroviral sequences search within *Solanum* clade, performed with BLASTN (Altschul, et al. 1990) at -
evaluate $1e^{-3}$, and then collapsing results from all genomes together with the sequences of 16
viral species from *Caulimoviridae* (www.ictvonline.org/) and different reported elements from
the *Florendovirus* genus (Geering, et al. 2014). Then, two rounds of such consecutive
Pfam+BLASTN searches were repeated. Subsequently, using a preliminary phylogenetic analysis
of reverse-transcriptase (RT) proteins of > 300 amino-acids (selected as a conservative
threshold to ensure robust confidence in alignment) inferred from listed pararetroviral
sequences, we recovered exemplary whole EPRVs from distinct phylogenetic groups in each
species. Such elements were assembled from marginally mutated sequences after careful
manual structural examination. Finally, another two consecutive rounds of Pfam+BLASTN
searches were conducted using as input the collapsed results from all *Solanum* genomes, but
now with a closing filtering step accepting only those sequences with at least 70% identity and
 ≥ 150 bp total alignment length to the above pararetrovirus species, exemplary assembled
EPRVs, or previously recognized complete *Florendovirus* elements from *Solanum* (Geering, et al.
2014). The identity threshold was placed conservatively above the *Caulimoviridae* genera call
(40-65% nucleotide identity (Sukal, et al. 2018)) but below the species call within genera (80%
nucleotide identity; <https://talk.ictvonline.org/>), making it very restrictive to this viral family. On
the other hand, the size filter avoided output inflation with small fragmented sequences that
might code for protein motifs shared with retrotransposons. A similar workflow was applied to
available *Glycine* and *Nicotiana* genomes, collapsing results from more than one species to
finally document the endogenized pararetroviral sequence space in *G. max* and *N. tabacum*.

Putative whole non-truncated EPRV candidates were recovered by aligning sequences
against pararetroviruses, some initially exemplary assembled EPRVs, or complete *Florendovirus*
elements from *Solanum*. First, the final pararetroviral sequence list was size-filtered in the
range between the smallest assembled EPRVs and the largest virus explored (between 6500-
9600 bp) ruling out those with ambiguous “N”. Then, those presenting at least 70% identity in
70% of their length using BLASTN were selected, subsequently performing a global alignment
with EMBOSS (Rice, et al. 2000) package needle function with parameters -gapopen 10 -

88 gapextend 0.5, accepting only hits above score 25000 with at least 70% similarity. As a mean to
89 assess the specificity/selectivity of our detection pipeline, the resulting 135 *S. lycopersicum*
90 whole non-truncated EPRVs were manually confirmed, and then compared with BLASTN against
91 very recent complete non-truncated insertions of *S. lycopersicum* LTR retrotransposons
92 representing both Copia (*Pseudoviridae*) and Gypsy (*Metaviridae*) superfamilies. These LTR
93 retrotransposons were called elements with extremely high LTR similarities (>99.5%) as
94 recognized by LTRharvest, and were annotated by the significant best blast hit (>80% identity
95 and alignment length of at least 500 bp) against Copia/Gypsy reported sequences from Repbase
96 (Bao, et al. 2015); finally accepting only those with translated in-between LTRs (with at least
97 100 amino-acids) showing compatibility to retrotransposon domains, as evidence by Pfam
98 recognition (but ruling out those with recognized MP domain). Importantly, the vast majority of
99 non-truncated EPRVs showed no significant similarities to any LTR retrotransposon listed at -
100 evalue 1e-3, save few irrelevant matches against fragments of extreme short length (29-31 bp);
101 however with one exception. This exception was revealing, presumably representing a
102 particular element called within coordinates SL4.0ch03:18539939-18553584, where the
103 alignments suspiciously matched only its extreme portions. It was later recognized that it was in
104 fact a composite chromosomal area, comprising pararetroviral-related sequences (recognized
105 by our pipeline: S lyc_Paraseq_325 to S lyc_Paraseq_328) genetically rearranged as similar
106 terminal-repeats around an historical remnant derived from a Gypsy element. We conclude
107 that such curious case defeated once our LTR retrotransposon discovery pipeline but did not
108 defeat our EPRV discovery pipeline, which showed correct detection of pararetroviral-related
109 portions while avoiding those originated in a LTR retrotransposon. As an exception that
110 confirmed the rule, this single hit highlighted that our recognized EPRVs presented no truly
111 relevant similarities to other potentially confounding intra-genomic retroelements. In addition,
112 *S. lycopersicum* whole non-truncated EPRVs were cross-compared to the REPET pipeline report
113 on repeated tomato sequences (Amselem, et al. 2019) (ITAG4.0_REPET_repeats_aggressive file,
114 available at www.solgenomics.net); where they were represented by 521 REPET fragments, but
115 only 60% of them were correctly distinguished as EPRVs. The rest presented non-declared
116 origins (16.5%), or were incorrectly annotated as retrotransposons (14.6% Gypsy, 2.1% Copia

and 0.6% LINE) or as different types of DNA transposable elements or simple sequence repeats. The analysis extrapolated to all our listed *S. lycopersicum* pararetroviral-related sequences resulted in a very similar figure, with only 55.9% of REPET fragments being called EPRV-related, suggesting a substantial number of database global misannotations. Taken together, this demonstrates that automatic annotations may represent a major constrain to independent call validation; underscoring the need of a dedicated analysis, such as the one we described above, in order to appropriately recognize EPRVs.

Flanking tandem-repeats were recognized with BLASTN through custom-made python scripts splitting elements in halves (considering only cases of no less than 100 bp aligning tandem-repeats above 85% identity, and occurring no further away than 20 bp from element's edge), whereas for inverted-repeats we initially used EMBOSS einverted with -gap 12 -threshold 200 -match 3 -mismatch -4 or -2 and -maxrepeat 30000, 5000 or 1000, with additional BLASTN and manual assessments.

Protein sequences were aligned with MAFFT (Katoh, et al. 2005) applying parameters --localpair --maxiterate 1000. Maximum likelihood phylogenetic analyses were performed in RAXML-NG (Kozlov, et al. 2019) with --model LG+G+F --tree pars(50),rand(50); no less than 1000 bootstraps were calculated till convergence at --bs-cutoff 0.02, which were mapped to the best reported tree and visualized in FigTree (<https://github.com/rambaut/figtree>).

Estimated boundaries of informative chromosomal areas were manually projected from siRNA signals visualized in genome-browsers, conservatively keeping as relevant those coordinates limited by inverted-repeats and/or pararetroviral sequences (depending on chromosomal context). Searches for *S. lycopersicum* TSAs in assemblies other than Heinz 1706 genome (ITAG4.0) –comprising nine *var. lycopersicum*, three *var. cerasiforme*, and the three *S. pimpinellifolium* LA2093, PAS014479 and BGV006775– were carried out by BLASTN, filtering for those with at least 90% similarity and 50% alignment length and further manually assessing syntenicity and uniqueness. The expected similarity between close homologous sequences was estimated from the presumed divergence time following the basic equation: $\text{time} = p \text{ genetic distance} / (2 * \text{substitution rate})$, using 1.3×10^{-8} mutations per site per year as inferred previously (Ma and Bennetzen 2004).

Estimates of Tajima's D summary statistic (Tajima 1989) were calculated genome-wide in overlapping sliding windows of 10000 bp with 1000 bp steps, from 124 *S. lycopersicum* accessions' private DNA-seq mapped data (Supplemental Table S7), using ANGSD software (Korneliussen, et al. 2014; Korneliussen, et al. 2013). The allele (site) frequency spectrum was estimated from allele frequency likelihoods, obtained with parameters `-uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -trim 0 -C 50 -baq 1 -minMapQ 20 -minQ 20 -GL 2 -doMajorMinor 1 -doCounts 0 -doSaf 1`. The ancestral state was inferred from the analyzed population's most common bases (i.e. majority-frequency allele for each SNP), with `-doFasta 2` and parameters `-uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -trim 0 -C 50 -baq 1 -minMapQ 20 -minQ 20 -basesPerLine 100 -explode 1 -seed 0 -doCounts 1`. For this, mappings were randomly subsampled beforehand toward the lowest available sequencing depth, to equalize mapping depth differences across samples. Fay and Wu's H summary test (Fay and Wu 2000) was calculated using Variscan 2.0 (Hutter, et al. 2006), with parameters `UseMuts = 1, UseLDSinglets = 1, CompleteDeletion = 1` and the maximum number of NumNuc; aligning with MAFFT `--globalpair --maxiterate 1000` those syntenic TSAs' nucleotide sequences from *S. lycopersicum* and *S. pimpinellifolium* accessions when reasonably complete.

Custom-made workflows for data explorations including Python scripts are available at <https://github.com/diegohernansanchez/>.

Next-generation sequencing and expression analyses

Next-generation sequencing reads were trimmed using Trimmomatic (Bolger, et al. 2014b) ILLUMINACLIP parameters `:2:10:5:1`, and further processed with open-source software such as BEDtools (Quinlan and Hall 2010), SAMtools (Li, et al. 2009) and Picard (<http://picard.sourceforge.net>). Small-RNA-seq, DNA-seq and ChIP-seq data were mapped with Bowtie2 (Langmead and Salzberg 2012), using parameters `--very-sensitive --non-deterministic`. For counting and size profiling of 'private' reads, these were filtered for only primary alignments with high MAPQ likelihood (SAMtools view parameters `-q 5 -F 256`), further applying BEDtools intersect with `-c` parameter. Counts per feature were then adjusted to the sum of total filtered counts per library (as counts-per-million, cpm) or per counts of EPRV-related

sequences (as fraction or percentage). For DNA-seq, only libraries presenting at least 40 million mapped private pair-end reads were explored (representing a minimum estimated primary alignment of ~x5 fold coverage). For PARE-seq and ChIP-seq, mapped data were collapsed with the bamCoverage function from deepTools2 suit (Ramirez, et al. 2016). BS-seq libraries were mapped, de-duplicated and methylation-called using Bismark (Krueger and Andrews 2011) with mapping parameters --bowtie2 -N 1 -L 20 -X 1000 -score_min L,0,-0.8 -R 3, while bismark_methylation_extractor was run as --comprehensive.

RNA-seq mapping was performed using STAR (Dobin, et al. 2013), with parameters --alignEndsType EndToEnd --twopassMode Basic --outReadsUnmapped None --outFilterMultimapNmax 10 --outMultimapperOrder Random. Reads were counted applying HT-seq count (Anders, et al. 2015) and normalized to the sum of HT-seq total counted library; present-call threshold for robust expression was set to >1 cpm in at least two independent samples under edgeR environment (Robinson, et al. 2010). Data manipulation workflows are available at <https://github.com/diegobernansanchez/>.

The expression and splicing of intronic/intergenic TSAs were validated by RT-PCR, performed with specific primers on cDNA template prepared from total DNA-free RNA from three-week-old *S. lycopersicum* leaves. Primers are available from [Supplemental Table S8](#).

RNA Interference

To build pRIC3.0-eGFP-TS and pRIC3.0-eGFP-NTS, a 22-bp potential target site for TSA10-derived siRNAs or its randomized sequence were introduced in the available unique XbaI restriction site within pRIC3.0-eGFP (Lamprecht, et al. 2016). Inserts were generated as small double-stranded DNAs obtained by *in vitro* annealing of specific oligos (TS/TSrevcomp and NTS/NTSrevcomp pairs, respectively; [Supplemental Table S8](#)).

S. lycopersicum cv M82 plants were grown in greenhouse with 16h/8h light/dark cycles at 20-24°C, with Supplemental light. *Agrobacterium tumefaciens* GV3101-pMP90RK (DSMZ) carrying pRIC3.0 (no reporter) or pRIC3.0-eGFP derivatives were infiltrated in three-week-old plant cotyledons, following classical reported protocols for agroinfiltration of *Nicotiana benthamiana* (Sparkes, et al. 2006). Portions of treated cotyledons were examined with a Leica

DMR epifluorescence microscope, using excitation and barrier filters at 450/490 nm and 500/550 nm, respectively, and photographed with an Olympus DP70 digital camera.

RLM-RACE was conducted as previously described (Llave, et al. 2011). Briefly, 1 µg of total DNA-free RNAs extracted with FavorPrep kit (Favorgene) from agroinfiltrated tissues were ligated to a 5' RACE adapter with T4 RNA ligase (NEB), and then reverse transcribed to cDNA with random primers using M-MuLV (NEB). Naturally cleaved products were amplified by two consecutive PCRs using 5' RACE forward/eGFP-3'UTR reverse and 5' RACE forward-nested /eGFP-3'UTR reverse-nested primers. RLM-RACE products were gel-purified, cloned into pCRII-TOPO (ThermoFisher) and Sanger sequenced (Macrogen Europe). The expression of *eGFP* reporter and *Actin* housekeeping gene were confirmed by RT-PCR and RT-qPCR from cDNA samples. Primers and oligos are available in [Supplemental Table S8](#).

References

- Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, Levy Y, Harel TH, Shalev-Schlosser G, Amsellem Z, Razifard H, Caicedo AL, Tieman DM, Klee H, Kirsche M, Aganezov S, Ranallo-Benavidez TR, Lemmon ZH, Kim J, Robitaille G, Kramer M, Goodwin S, McCombie WR, Hutton S, Van Eck J, Gillis J, Eshed Y, Sedlazeck FJ, van der Knaap E, Schatz MC, Lippman ZB 2020. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* 182: 145-161.e123. doi: 10.1016/j.cell.2020.05.021
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ 1990. Basic local alignment search tool. *J Mol Biol* 215: 403-410. doi: 10.1016/s0022-2836(05)80360-2
- Amselem J, Cornut G, Choisine N, Alaux M, Alfama-Depauw F, Jamilloux V, Maumus F, Letellier T, Luyten I, Pommier C, Adam-Blondon A-F, Quesneville H 2019. RepetDB: a unified resource for transposable element references. *Mob DNA* 10: 6-6. doi: 10.1186/s13100-019-0150-y
- Anders S, Pyl PT, Huber W 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31: 166-169. doi: 10.1093/bioinformatics/btu638
- Bao W, Kojima KK, Kohany O 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 6: 11. doi: 10.1186/s13100-015-0041-9
- Barchi L, Pietrella M, Venturini L, Minio A, Toppino L, Acquadro A, Andolfo G, Aprea G, Avanzato C, Bassolino L, Comino C, Molin AD, Ferrarini A, Maor LC, Portis E, Reyes-Chin-Wo S, Rinaldi R, Sala T, Scaglione D, Sonawane P, Tononi P, Almekias-Siegl E, Zago E, Ercolano MR, Aharoni A, Delledonne M,

Giuliano G, Lanteri S, Rotino GL 2019. A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. *Scientific Reports* 9: 11769. doi: 10.1038/s41598-019-47985-w

Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sørensen I, Lichtenstein G, Fich EA, Conte M, Keller H, Schneeberger K, Schwacke R, Ofner I, Vrebalov J, Xu Y, Osorio S, Aflitos SA, Schijlen E, Jiménez-Goméz JM, Ryngajlo M, Kimura S, Kumar R, Koenig D, Headland LR, Maloof JN, Sinha N, van Ham RCHJ, Lankhorst RK, Mao L, Vogel A, Arsova B, Panstruga R, Fei Z, Rose JKC, Zamir D, Carrari F, Giovannoni JJ, Weigel D, Usadel B, Fernie AR 2014a. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat Genet* 46: 1034. doi: 10.1038/ng.3046 <https://www.nature.com/articles/ng.3046#supplementary-information>

Bolger AM, Lohse M, Usadel B 2014b. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114-2120. doi: 10.1093/bioinformatics/btu170

Bombarely A, Rosli HG, Vrebalov J, Moffett P, Mueller LA, Martin GB 2012. A Draft Genome Sequence of *Nicotiana benthamiana* to Enhance Molecular Plant-Microbe Biology Research. *Molecular Plant-Microbe Interactions*® 25: 1523-1530. doi: 10.1094/MPMI-06-12-0148-TA

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15-21. doi: 10.1093/bioinformatics/bts635

Eddy SR 2011. Accelerated Profile HMM Searches. *PLoS Comput Biol* 7: e1002195. doi: 10.1371/journal.pcbi.1002195

Edwards KD, Fernandez-Pozo N, Drake-Stowe K, Humphry M, Evans AD, Bombarely A, Allen F, Hurst R, White B, Kernodle SP, Bromley JR, Sanchez-Tamburrino JP, Lewis RS, Mueller LA 2017. A reference genome for *Nicotiana tabacum* enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *BMC Genomics* 18: 448. doi: 10.1186/s12864-017-3791-6

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD 2019. The Pfam protein families database in 2019. *Nucleic Acids Res* 47: D427-d432. doi: 10.1093/nar/gky995

Ellinghaus D, Kurtz S, Willhoeft U 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9: 18. doi: 10.1186/1471-2105-9-18

Fay JC, Wu C-I 2000. Hitchhiking Under Positive Darwinian Selection. *Genetics* 155: 1405-1413. doi: 10.1093/genetics/155.3.1405

Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW, Foolad MR, Diez MJ, Blanca J, Canizares J, Xu Y, van der Knaap E, Huang S, Klee HJ, Giovannoni JJ, Fei Z 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet* 51: 1044-1051. doi: 10.1038/s41588-019-0410-2

Geering ADW, Maumus F, Copetti D, Choisne N, Zwickl DJ, Zytnicki M, McTaggart AR, Scalabrin S, Vezzulli S, Wing RA, Quesneville H, Teycheney P-Y 2014. Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat Commun* 5: 5269. doi: 10.1038/ncomms6269

Hardigan MA, Crisovan E, Hamilton JP, Kim J, Laimbeer P, Leisner CP, Manrique-Carpintero NC, Newton L, Pham GM, Vaillancourt B, Yang X, Zeng Z, Douches DS, Jiang J, Veilleux RE, Buell CR 2016. Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated *Solanum tuberosum*. *Plant Cell* 28: 388. doi: 10.1105/tpc.15.00538

Hosmani PS, Flores-Gonzalez M, van de Geest H, Maumus F, Bakker LV, Schijlen E, van Haarst J, Cordewener J, Sanchez-Perez G, Peters S, Fei Z, Giovannoni JJ, Mueller LA, Saha S 2019. An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *bioRxiv*: 767764. doi: 10.1101/767764

Hutter S, Vilella AJ, Rozas J 2006. Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics* 7: 409. doi: 10.1186/1471-2105-7-409

Katoh K, Kuma K-i, Toh H, Miyata T 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33: 511-518. doi: 10.1093/nar/gki198

Korneliussen TS, Albrechtsen A, Nielsen R 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15: 356. doi: 10.1186/s12859-014-0356-4

Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R 2013. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics* 14: 289. doi: 10.1186/1471-2105-14-289

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35: 4453-4455. doi: 10.1093/bioinformatics/btz305

Krueger F, Andrews SR 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27: 1571-1572. doi: 10.1093/bioinformatics/btr167

Lamprecht RL, Kennedy P, Huddy SM, Bethke S, Hendrikse M, Hitzeroth II, Rybicki EP 2016. Production of Human papillomavirus pseudovirions in plants and their use in pseudovirion-based neutralisation assays in mammalian cells. *Scientific Reports* 6: 20431. doi: 10.1038/srep20431

Langmead B, Salzberg SL 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357-359. doi: 10.1038/nmeth.1923

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079. doi: 10.1093/bioinformatics/btp352

Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, Huang Z, Li J, Zhang C, Wang T, Zhang Y, Wang A, Zhang Y, Lin K, Li C, Xiong G, Xue Y, Mazzucato A, Causse M, Fei Z, Giovannoni JJ, Chetelat RT, Zamir D, Städler T, Li J, Ye Z, Du Y, Huang S 2014. Genomic analyses provide insights into the history of tomato breeding. *Nat Genet* 46: 1220-1226. doi: 10.1038/ng.3117

- Llave C, Franco-Zorrilla JM, Solano R, Barajas D 2011. Target validation of plant microRNAs. *Methods Mol Biol* 732: 187-208. doi: 10.1007/978-1-61779-083-6_14
- Lunardon A, Johnson NR, Hagerott E, Phifer T, Polydore S, Coruh C, Axtell MJ 2020. Integrated annotations and analyses of small RNA-producing loci from 47 diverse plants. *Genome Res* 30: 497-513. doi: 10.1101/gr.256750.119
- Ma J, Bennetzen JL 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A* 101: 12404. doi: 10.1073/pnas.0403715101
- Quinlan AR, Hall IM 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841-842. doi: 10.1093/bioinformatics/btq033
- Ramirez F, Ryan DP, Gruning B 2016. deepTools2: a next generation web server for deep-sequencing data analysis. 44: W160-165. doi: 10.1093/nar/gkw257
- Rice P, Longden I, Bleasby A 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16: 276-277. doi: 10.1016/S0168-9525(00)02024-2
- Robinson MD, McCarthy DJ, Smyth GK 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139-140. doi: 10.1093/bioinformatics/btp616
- Sanchez DH, Gaubert H, Yang W 2019. Evidence of developmental escape from transcriptional gene silencing in MESSI retrotransposons. *New Phytol* 223: 950-964. doi: 10.1111/nph.15896
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X-C, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178-183. doi: 10.1038/nature08670
- Seo E, Kim T, Park JH, Yeom S-I, Kim S, Seo M-K, Shin C, Choi D 2018. Genome-wide comparative analysis in Solanaceous species reveals evolution of microRNAs targeting defense genes in *Capsicum* spp. *DNA Research* 25: 561-575. doi: 10.1093/dnares/dsy025
- Shen Y, Du H, Liu Y, Ni L, Wang Z, Liang C, Tian Z 2019. Update soybean Zhonghuang 13 genome to a golden reference. *Science China Life Sciences* 62: 1257-1260. doi: 10.1007/s11427-019-9822-2
- Sparkes IA, Runions J, Kearns A, Hawes C 2006. Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. *Nature Protocols* 1: 2019-2025. doi: 10.1038/nprot.2006.286
- Sukal AC, Kidanemariam DB, Dale JL, Harding RM, James AP 2018. Characterization of a novel member of the family Caulimoviridae infecting *Dioscorea nummularia* in the Pacific, which may represent a new genus of dsDNA plant viruses. *PLoS One* 13: e0203038. doi: 10.1371/journal.pone.0203038

Tajima F 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585-595. doi: 10.1093/genetics/123.3.585

Teycheney P-Y, Geering ADW, Dasgupta I, Hull R, Kreuze JF, Lockhart B, Muller E, Olszewski N, Pappu H, Pooggin MM, Richert-Pöggeler KR, Schoelz JE, Seal S, Stabolone L, Umber M, Report Consortium I 2020. ICTV Virus Taxonomy Profile: Caulimoviridae. *Journal of General Virology*. doi: <https://doi.org/10.1099/jgv.0.001497>

Tomato Genome Sequencing C, Aflitos S, Schijlen E, de Jong H, de Ridder D, Smit S, Finkers R, Wang J, Zhang G, Li N, Mao L, Bakker F, Dirks R, Breit T, Gravendeel B, Huits H, Struss D, Swanson-Wagner R, van Leeuwen H, van Ham RC, Fito L, Guignier L, Sevilla M, Ellul P, Ganko E, Kapur A, Reclus E, de Geus B, van de Geest H, Te Lintel Hekkert B, van Haarst J, Smits L, Koops A, Sanchez-Perez G, van Heusden AW, Visser R, Quan Z, Min J, Liao L, Wang X, Wang G, Yue Z, Yang X, Xu N, Schranz E, Smets E, Vos R, Rauwerda J, Ursem R, Schuit C, Kerns M, van den Berg J, Vriezen W, Janssen A, Datema E, Jahrman T, Moquet F, Bonnet J, Peters S 2014. Exploring genetic variation in the tomato (*Solanum section Lycopersicon*) clade by whole-genome sequencing. *Plant J* 80: 136-148. doi: 10.1111/tpj.12616

Valliyodan B, Cannon SB, Bayer PE, Shu S, Brown AV, Ren L, Jenkins J, Chung CYL, Chan T-F, Daum CG, Plott C, Hastie A, Baruch K, Barry KW, Huang W, Patil G, Varshney RK, Hu H, Batley J, Yuan Y, Song Q, Stupar RM, Goodstein DM, Stacey G, Lam H-M, Jackson SA, Schmutz J, Grimwood J, Edwards D, Nguyen HT 2019. Construction and comparison of three reference-quality genome assemblies for soybean. *The Plant Journal* 100: 1066-1082. doi: <https://doi.org/10.1111/tpj.14500>

Wang X, Gao L, Jiao C, Stravoravdis S, Hosmani PS, Saha S, Zhang J, Mainiero S, Strickler SR, Catala C, Martin GB, Mueller LA, Vrebalov J, Giovannoni JJ, Wu S, Fei Z 2020. Genome of *Solanum pimpinellifolium* provides insights into structural variants during tomato breeding. *Nat Commun* 11: 5817. doi: 10.1038/s41467-020-19682-0

Wang Z, Baulcombe DC 2020. Transposon age and non-CG methylation. *Nat Commun* 11: 1221. doi: 10.1038/s41467-020-14995-6

Wang Z, Hardcastle TJ, Canto Pastor A, Yip WH, Tang S, Baulcombe DC 2018. A novel DCL2-dependent miRNA pathway in tomato affects susceptibility to RNA viruses. *Genes & Development* 32: 1155-1160.

Xie M, Chung CY-L, Li M-W, Wong F-L, Wang X, Liu A, Wang Z, Leung AK-Y, Wong T-H, Tong S-W, Xiao Z, Fan K, Ng M-S, Qi X, Yang L, Deng T, He L, Chen L, Fu A, Ding Q, He J, Chung G, Isobe S, Tanabata T, Valliyodan B, Nguyen HT, Cannon SB, Foyer CH, Chan T-F, Lam H-M 2019. A reference-grade wild soybean genome. *Nat Commun* 10: 1216. doi: 10.1038/s41467-019-09142-9

Xu S, Brockmüller T, Navarro-Quezada A, Kuhl H, Gase K, Ling Z, Zhou W, Kreitzer C, Stanke M, Tang H, Lyons E, Pandey P, Pandey SP, Timmermann B, Gaquerel E, Baldwin IT 2017. Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proceedings of the National Academy of Sciences* 114: 6133. doi: 10.1073/pnas.1700073114