Supplementary Materials

Centromere divergence and allopolyploidy reshape carnivorous sundew genomes

Laura Ávila Robledillo[1]†, Steven J. Fleck[2]‡, Jonathan Kirshner[2]‡, Dirk Becker[3], Aaryan Bhatia[1], Gerhard Bringmann[4], Jordan R. Brock[5], Daniela Drautz-Moses[6], Matthias Freund[3], Rainer Hedrich[7], Luis Herrera-Estrella[8], Enrique Ibarra-Laclette[9], Ines Kreuzer[3], Tianying Lan[2], Sachiko Masuda[10], Martín Mata-Rosas[11], Todd P. Michael[12,13,14,15], Héctor Montero[16], Sitaram Rajaraman[17,18], Michaela Richter[2], David Sankoff[19], Stephan C. Schuster[6,17], Ken Shirasu[10,20], Sonja Trebing[3], Yves Van de Peer[21,22,23,24], Gerd Vogg[25], Tan Qiao Wen[17], Yue Zhang[19], Chunfang Zheng[19], Kenji Fukushima[16]*, Jarkko Salojärvi[17,18]*, André Marques[1,27]*, Victor A. Albert[2]*

Corresponding authors. Email: vaalbert@buffalo.edu (V.A.A.); amarques@mpipz.mpg.de (A.M.); jarkko@ntu.edu.sg (J.S.); kenji.fukushima@nig.ac.jp (K.F.)

This file includes:

Materials and Methods

Figs. S1 to S9

Tables S1 to S2

Materials and methods

Plant materials and genome sequencing. In vitro cultures of *D. regia* and *D. capensis* were initiated from seeds derived from plants maintained in the carnivorous plant collection at the Red Manejo Biotecnológico de Recursos, Instituto de Ecología, A.C., Xalapa, Veracruz, Mexico. For superficial disinfection, seeds were placed in filter paper envelopes (Whatman No. 1, 110 mm diameter) and submerged in sterile distilled water for 30 minutes. They were then soaked in a 10% (v/v) commercial bleach solution (1.8% NaOCl) containing two drops of Tween-80 per 100 mL (Sigma, St. Louis, MO) for 10 minutes. This was followed by four rinses with sterile distilled water under aseptic conditions. The disinfected seeds were sown in 125-mL baby food jars containing 25 mL of half-strength MS medium (Murashige and Skoog, 1962), supplemented with 30 $g \cdot L^{-1}$ sucrose. The pH was adjusted to 5.0 ± 0.1 using 0.5 N NaOH and 0.5 N HCl prior to the addition of 7.5 $g \cdot L^{-1}$ Agar, Plant TC (Caisson A111), followed by autoclaving at 1.2 $kg \cdot cm^{-2}$ and 120 °C for 15 minutes. Cultures were incubated in a growth chamber at 25 ± 1 °C under a 16-hour photoperiod provided by LED lamps (50 $\mu mol \cdot m^{-2} \cdot s^{-1}$). Plants obtained from germinated seeds were subcultured every 3-4 months on the same medium to promote growth and multiplication.

High-molecular-weight DNA was extracted from nuclei isolated from the young leaves of *Drosera* species, following the protocol by Steinmüller and Apel, 1986[1]. To reduce contamination from chloroplast and mitochondrial DNA, nuclei were first collected from a 60% Percoll (Invitrogen) density gradient after centrifugation at 4000g for 10 minutes at 4°C. Next, high-quality megabase-sized DNA was obtained using the MagAttract HMW DNA Kit (Qiagen). Before library preparation for sequencing, the integrity of the high molecular weight (HMW) DNA was confirmed using pulsed-field gel electrophoresis (CHEF-DRIII system, Bio-Rad), as described elsewhere[2]. For library preparation, 10 µg of DNA were sheared to a fragment size range of 10-40 kb using a Covaris g-TUBE. The resulting fragment distribution was verified by pulsed-field gel electrophoresis. The sheared DNA was purified using 0.45× AMPure PB beads (Pacific Biosciences) following the manufacturer's protocol.

Library preparation for the PacBio RS II instrument was carried out according to the PacBio 20 kb SMRTbell Template Preparation Protocol, using 5 µg of the sheared DNA as input material. After preparation, the library size distribution was analyzed on an Agilent DNA 12000 Bioanalyzer chip to determine the appropriate size-selection cut-off. Libraries were size-selected with a Sage Science BluePippin system, employing a dye-free 0.75% agarose cassette and a 15 kb cut-off. The selected libraries were reanalyzed on the Bioanalyzer to confirm size distribution. Two libraries per species were prepared for *D. capensis* and *D. regia*. The *D. capensis* library was sequenced on two SMRT cells of the PacBio RSII single-molecule sequencing platform at loading concentrations of 0.15 nM and 0.2 nM, respectively. The *D. regia* library was sequenced on eight SMRT cells at a loading concentration of 0.2 nM. For Illumina sequencing of *D. capensis*, *D. regia* and the ten additional *Drosera* species, the Illumina HiSeq 2500 (rapid run, 2x250bp; https://www.illumina.com/documents/products/datasheets/datasheet_hiseq2500.pdf) was employed.

Genome assembly and Hi-C scaffolding. The 2x250bp Illumina reads were first filtered for adapter sequences using Trimmomatic v0.36[3]. A hybrid assembly including the filtered reads and PacBio RS II reads was then carried out using MaSuRCA v3.2.1[4]. The contig-level assemblies were evaluated for completeness with BUSCO v3.0.2[5] using the embryophyta odb9 database, and for contiguity, using QUAST v4.3[6].

Dovetail Hi-C reads were first mapped to the contig file obtained from the MaSuRCA assembler using BWA[7] following the hic-pipeline (https://github.com/esrice/hic-pipeline). Hi-C scaffolding was performed using 3D-DNA pipeline[8,9] with default parameters using '*GATC, GAATC, GATTC, GAGTC, GACTC*' as restriction sites. After testing several minimum mapping quality values of bam alignments, the final scaffolding was performed with MAPQ10. Following the automated scaffolding by 3D-DNA, several rounds of visual assembly correction guided by Hi-C heatmaps were performed. When regions showed multiple contact patterns, manual re-organization of the scaffolds was performed with Juicebox and 3D-DNA assembly pipeline[8] to correct position/orientation and to obtain the pseudomolecules.

Transcriptome sequencing and genome annotation. Total RNA was extracted from *D. capensis* leaf tissues and petioles. Sample preparation employed a single stranded mRNA library kit, and libraries were subsequently sequenced using an Illumina HiSeq instrument. We obtained a total of 105,784,845 read pairs. To generate our transcriptome assembly, we first merged the RNA-Seq data from the two tissues. We then assembled one de novo transcriptome using transAbyss v2.0.1[10]. In transAbyss, we used multiple *k*-mers (33-75) in steps of 2 to generate multiple assemblies before merging them all into a single large set. We then assembled a second de novo transcriptome using Trinity v2.6.6[11] using the default *k*-mer size of 25. We generated one reference-guided transcriptome by first mapping the RNA-Seq reads against the reference *D. capensis* genome using HISAT2 v.2.1.0[12] and then assembling the transcriptome using StringTie v.1.3.4c[13]. We then passed the three transcriptome assemblies to EvidentialGene v2017.12.21[14] which produced a final high confidence transcriptome assembly.

For repeat masking, we first generated a de novo repeat library for the *D. capensis* genome using RepeatModeler v1.0.9[15] and then masked the genome using RepeatMasker v4.0.7[16]. For gene model prediction, the transcriptome assembly was first splice-aligned against the unmasked reference *D. capensis* genome using PASA v2.2.0[17] to generate ORFs. Secondly, ab initio gene model prediction was carried out using BRAKER v.2.0.3[18], which internally used the reference aligned RNA-Seq data and Genemark-ET to train AUGUSTUS v3.3[18,19] for its final prediction. Additionally, the self-training Genemark-ES v.4.33[20] was run independently to generate a second set of predictions. Finally the 2 prediction tracks and two spliced-alignment tracks were passed to the combiner tool Evidence Modeler v1.1.1[17] with highest weights to transcriptome evidence and lowest weights to Genemark-ES to generate a final high confidence gene prediction set. This final prediction set was re-run through PASA to update the gene models, add UTRs and identify and generate alternate spliced models.

For *D. regia*, the *D. capensis* transcriptome assembly was used as evidence for training purposes. First, the assembly was splice-aligned against the *D. regia* genome using PASA v.2.2.0[17] to generate ORFs. Additionally, *Arabidopsis thaliana* gene models were aligned against the genome using exonerate v2.2.0[21]. Genemark-ES v4.33[20] was then used to generate one set of predictions. BRAKER v2.0.3[18] was used in the protein mode where the *D. capensis* gene models were aligned against the *D. regia* genome using GenomeThreader v1.7.1 (https://genomethreader.org/), and that set was used to train BRAKER v2.0.3 to generate a second set of gene models. AUGUSTUS v3.3[18,19] was run independently using *D.*

capensis parameters to generate a third set of predictions for *D. regia*. All of these gene predictions were then passed to Evidence Modeler v1.1.1[17] to generate a single high confidence gene prediction set. This set was once again passed through PASA to update the gene models with UTR regions and also to generate the splice variants.

Synteny analysis. Synteny analyses were performed with GENESPACE (https://github.com/jtlovell/GENESPACE)[22]. To identify shared chromosomal rearrangements, we first identified all end-to-end fusions of non-homoeologous chromosomes within the *D. regia* genome. For each such fusion event, we then extracted corresponding genomic regions from *D. capensis* that involved the same ancestral *Nepenthes*-like chromosomes. Shared rearrangements were inferred when the breakpoint boundary regions in *D. regia* precisely matched the syntenic block locations in *D. capensis* that represented these fusions.

Further synteny analyses were performed using the CoGe SynMap platform (https://genomevolution.org/coge/SynMap.pl)[23]. Synteny plots were obtained using the following steps: (1) using the Last tool, (2) synteny analysis was performed using DAGChainer, using 20 genes as the maximum distance between two matches (-D) and 5 genes as the minimum number of aligned pairs (-A). Then (3) either default (no syntenic depth) or Quota Align (with syntenic depth) was used with overlap distance of 40 genes, and (4) orthologous and paralogous blocks were differentiated according to the synonymous substitution rate (Ks) using CoGe-integrated CodeML, and represented with different colors in the dot plot. FractBias[24] was run using Quota Align window size of 100 for all genes in the target genome, with a syntenic depth ratio of 12:12 with maximum query and target chromosome numbers of 64 each for the concatenated *D. regia*-*N. gracilis* genome assembly.

For the characterization of regions involved in fusions, we followed Hofstatter et al, 2022[25]. The syntenic alignment obtained in GENESPACE between the *D. regia* genome and the *N. gracilis* dominant subgenome allowed us to pinpoint regions around the borders of proposed homologous fusion events. To evaluate homology, we loaded and compared annotation features for genes, TEs, and tandem repeats along the syntenic alignments using Geneious (https://www.geneious.com).

Subgenome-aware phasing of* D. regia *and* Dionaea muscipula *genomes. We used SubPhaser[26] (default parameters) to phase and partition the subgenomes of *D. regia* and *Dionaea muscipula* genomes simultaneously by assigning chromosomes to subgenomes based on differential repetitive *k*-mers. Additionally, Ks distributions of homeologous duplicate gene pairs were extracted from CoGe SynMap calculations (above) to generate density plots for each triplet of *D. regia* ancestral chromosomes. Density plots were generated in R using the tidyverse[27], ggplot2[28], RColorBrewer[29], ggridges[30], and ggpmisc[31] packages.

ksrates *analysis.* *ksrates* version 1.1.4[32] was used to position species splits relative to polyploidy events. Coding sequence (CDS) fasta files were extracted using AGAT version 1.4.0[33]. Paralogous Ks peaks were generated for the following focal species: *Ancistrocladus abbreviatus* (this study), *Dionaea muscipula* (this study), *D. capensis* (this study), *D. regia* (this study), *Nepenthes gracilis*[34], and *Triphyophyllum peltatum* (this study). *Beta vulgaris* (GCA_026745355.1), *Coffea canephora*[35], *Gelsemium elegans* (CoGe id64491), and

Spinacia oleracea (GCA_020520425.1) were also included in the analysis, but not as focal species. Orthologous Ks peaks were generated for each required species pair.

For the tree topology used in *ksrates*, OrthoFinder v2.5.5[36] was run with default settings to generate a tree for all species listed above. Each annotation was reduced to the longest isoform and proteins were extracted using AGAT version 1.4.0[33]. The species tree inference was performed with STAG[37], which uses the proportion of species trees derived from single-locus gene trees supporting each bipartition as its measure of support, and rooted using STRIDE[38].

Repeat characterization. DANTE and DANTE-LTR retrotransposon identification (Galaxy Version 3.5.1.1) pipelines[39] were used to identify full-length LTR retrotransposons in the assembled genomes of *D. capensis* and *D. regia*, using a set of protein domains from REXdb[40]. All complete LTR-RTs contain GAG, PROT, RT, RH and INT domains, including some lineages encoding additional domains, such as chromodomains (CHD and CHDCR) from chromoviruses or ancestral RNase H (aRH) from Tat elements. DANTE_LTR retrotransposon filtering (Galaxy Version 3.5.1.1) was used to search for high quality retrotransposons, those with no cross-similarity between distinct lineages. This tool produced a GFF3 output file with detailed annotations of the LTR-RTs identified in the genome and a summary table with the numbers of the identified elements. Overall repeat composition was calculated excluding clusters of organelle DNA (chloroplast and mitochondrial DNA).

Tandem repeat sequences were identified using RepeatExplorer2 (https://repeatexplorer-elixir.cerit-sc.cz/) and further verified using the TideCluster pipeline (https://github.com/kavonrtep/TideCluster)[41]. All putative tandem sequences were compared for homology using DOTTER. These tandem sequences were individually mapped to the genome by BLAST[42], with 95% similarity in Geneious (https://www.geneious.com). The mapped sequence files were converted to BED and used as an input track for a genome-wide overview with ShinyCircos[43] using a 100kb window**.**

Comparative repeatome analysis of *Drosera* genomes was made using Illumina reads from the species listed in **Supplementary Materials Table S2**. First, reads were filtered by quality with 95% of bases equal to or above the quality cut-off value of 10 using the RepeatExplorer2 pipeline (https://repeatexplorer-elixir.cerit-sc.cz/)[41]. The clustering was performed using the default settings of 90% similarity over 55% of the read length. For the comparative analyses, we performed an all-to-all similarity comparison across all species following the same approach. Because genome sizes are unknown for some analyzed species, each set of reads was down-sampled to 1,000,000 for each species. The automated annotations of repeat clusters obtained by RepeatExplorer2 were manually inspected and reviewed, followed by recalculation of the genomic proportion of each repeat type when appropriate.

Immunostaining and FISH. Flower buds of *D. capensis and D. regia* at various developmental stages were harvested and immediately fixed in freshly prepared 4% formaldehyde in Tris buffer (10 mM Tris-HCl (pH 7.5), 10 mM EDTA, and 0.5% Triton X-100), when immunostaining was intended, or ethanol-acetic acid (3:1 v/v) when only FISH was performed.

For immunostaining, fixation was carried out under vacuum infiltration for at least 15 minutes at room temperature, followed by an additional 45 minutes of incubation without vacuum. Fixed tissues were washed twice in 1× phosphate-buffered saline (1x PBS) at 4 °C until further processing. For enzymatic digestion, individual fixed buds were incubated in a solution containing 2% (w/v) cellulase (Onozuka R-10) and 2% (w/v) pectinase (Sigma) prepared in 1× PBS. Digestion was performed at 37 °C for 1 hour to facilitate cell wall degradation and release of nuclei. Following digestion, the softened tissue was gently macerated on a clean slide and squashed under the coverslip. Chromosome spreads were washed in 1× PBS for 5 minutes, followed by incubation in PBS-T1 buffer (1× PBS, 0.5% Triton X-100, pH 7.4) for 25 minutes. After two additional 5-minute washes in PBS, slides were incubated in PBS-T2 buffer (1× PBS, 0.1% Tween 20, pH 7.4) for 30 minutes. Primary antibody incubation was performed overnight at 4 °C using rabbit anti-CENH3 (specific for each species; **Supplementary Materials Fig. 2**), rabbit anti-KNL1[44] (GenScript, NJ, USA) and mouse anti α-tubulin (Sigma-Aldrich, St. Louis, MO; catalog number T6199) diluted 1:1,000 in blocking buffer (3% BSA, 1x PBS, 0.1% Tween 20, pH 7.4). Slides were then washed twice in 1x PBS (5 minutes each) and once in PBS-T2 (5 minutes), followed by incubation with secondary antibodies for at least 1 hour at room temperature. As the secondary antibody, goat anti-rabbit IgG antibody conjugated with Alexa Fluor 488 (Invitrogen; catalog number A27034), goat anti-rabbit conjugated with Rhodamine Red X (Jackson ImmunoResearch, catalog number: 111-295-144) or goat anti-mouse conjugated with Alexa Fluor 488 (Jackson ImmunoResearch; catalog number 115-545-166) were used in a 1:500 dilution. Final washes included two rounds in PBS and one in PBS-T2, each for 5 minutes. Slides were then mounted for fluorescence microscopy. Microscopic images were recorded using a Zeiss Axiovert 200M microscope equipped with a Zeiss AxioCam CCD. Images of at least 5 cells were analyzed using the ZEN software (Carl Zeiss GmbH).

For FISH experiments, material fixation was performed for 2 days at 4 °C, washed in ice-cold water, and digested in a solution of 4% cellulase (Onozuka R10, Serva Electrophoresis, Heidelberg, Germany), 2% pectinase, and 0.4% pectolyase Y23 (both MP Biomedicals, Santa Ana, CA) in 0.01 M citrate buffer (pH 4.5) for 60 min at 37 °C. The digested material was transferred to a drop of 45% acetic acid, macerated and squashed under a coverslip. FISH was performed using species-specific oligonucleotide probes that were 5′-labeled with Cy3 during their synthesis:

*D. regia*__Regi90 (Cy3)-
CAAGTATTTCAATGGAAATGGTGAAATAACATGTTTTACACCTATTTCC;

*D. capensis*__Cape71 (Cy3)-
CCCTTTAAATGAGCTTAAAACACTCAAAACCCCTTGAAAAGGCTAAAAAC

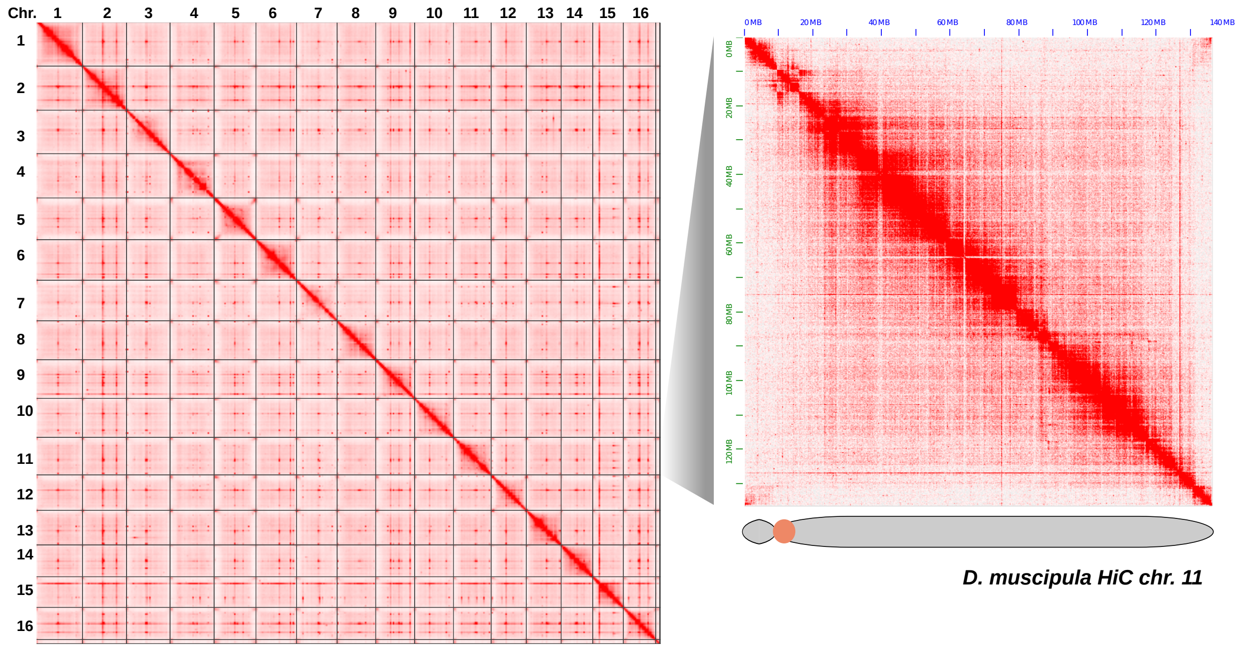
FISH was performed as described in Macas, et al. 2007[45], with hybridization and washing temperatures adjusted to account for AT/GC content and hybridization stringency allowing for 10-20% mismatches. The slides were counterstained with 2 µg/mL DAPI in Vectashield (Vector) mounting medium. The images of at least 10 cells were captured as described above.

Satellite DNA phylogeny. Centromere tracks were generated using the consensus sequences of the main satellite repeats identified using TideCluster, and the coordinates of *Cape71* and *Regi90* satDNA polymers were used to guide sequence extraction in Geneious Prime (https://www.geneious.com). In total, 5,371 *Cape71* and 18,178 *Regi90* were extracted from *D. capensis* and *D. regia*, respectively. The collected centromeric repeat sequences were aligned using MAFFT v7.490[46]. Phylogenetic trees were inferred using FastTree v2.1.11[47] under the generalized time-reversible (GTR) model. Resulting trees were visualized and annotated using the Interactive Tree of Life (iTOL) web server[48], with colors corresponding to their chromosome of origin.

The phylogenetic trees were imported into R using the ape package[49]. Tip labels included chromosome ID and genomic coordinates. Genomic positions of sampled tips were plotted using ggplot2[28] as normalized positions (relative to chromosome length) in order to show the spatial distribution of tips along individual chromosomes. Tips were plotted using ggplot2, with color representing the order of appearance in the tree (NodeOrder), allowing assessment of how phylogenetic relationships correlate with genomic location. All analyses and plotting were performed in R (version 4.1.2) using the packages ape[49], dplyr[50], ggplot2[28], and viridis (https://sjmgarnier.github.io/viridis/).

Species tree inference with BUSCO genes. To extract BUSCO genes, we collected coding sequence sets from sequenced genomes as well as previously assembled transcriptomes[51] (**Supplementary Materials Table S1**). We identified eudicot-conserved single-copy genes with BUSCO v5.3.2 (https://gitlab.com/ezlab/busco). Genes classified as single-copy (S) or fragmented (F) were retained, whereas those classified as duplicated (D) or missing (M) were considered absent, as described previously. Protein sequences were aligned with MAFFT v7.508 (https://mafft.cbrc.jp/alignment/server/index.html), trimmed with ClipKIT v2.1.1 (https://github.com/JLSteenwyk/ClipKIT), and back-translated to codons with CDSKIT v0.10.10 (https://github.com/kfuku52/cdskit) to produce in-frame nucleotide alignments. For each gene, nucleotide and protein maximum-likelihood trees were inferred with IQ-TREE v2.2.5 using GTR+R4 and LG+R4 models, respectively. Individual gene trees were then used for coalescent species-tree inference with ASTRAL v5.7.8[52]. Concatenated nucleotide and protein alignments, generated by catfasta2phyml v2018-09-28 (https://github.com/nylander/catfasta2phyml), served as additional input to IQ-TREE[53] with the same substitution models. *Amborella trichopoda* was specified as the outgroup.

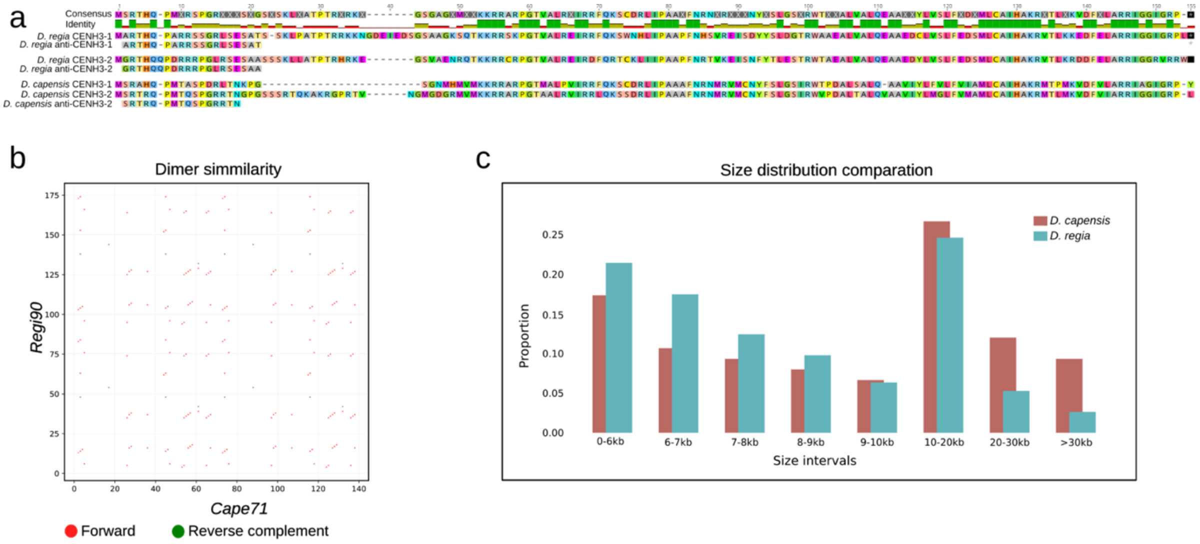
Supplementary Materials: Figures



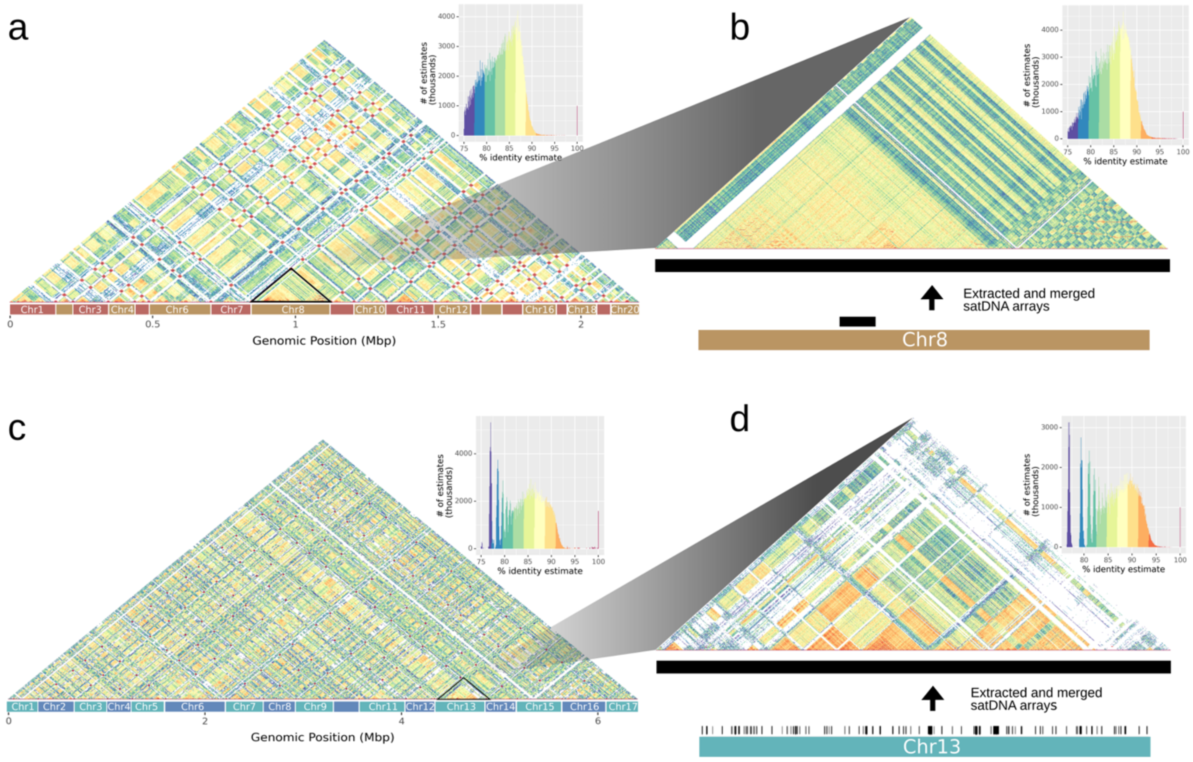
D. muscipula HiC chr. 11

***D. muscipula* HiC**

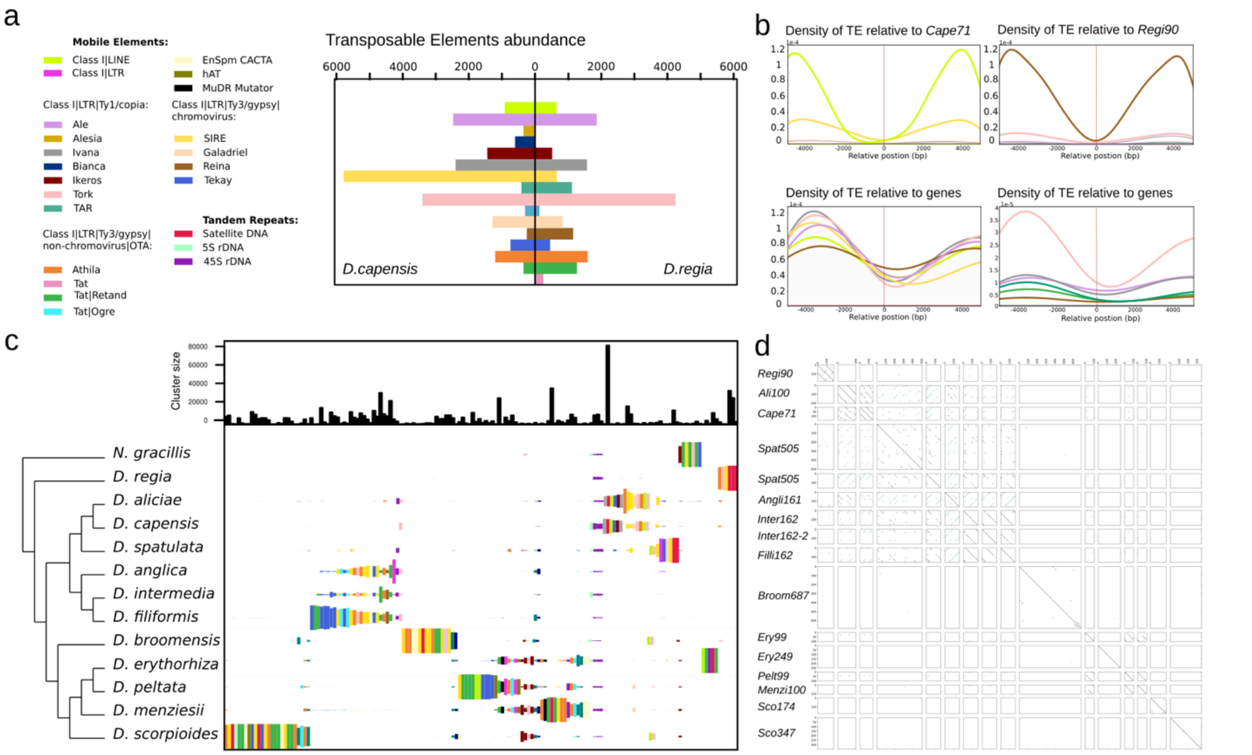
Supplementary Materials Fig. S1. *Dionaea muscipula* Hi-C contact map. Hi-C contact map showing every *Dionaea muscipula* chromosome, plus a zoom-in to chromosome 11.



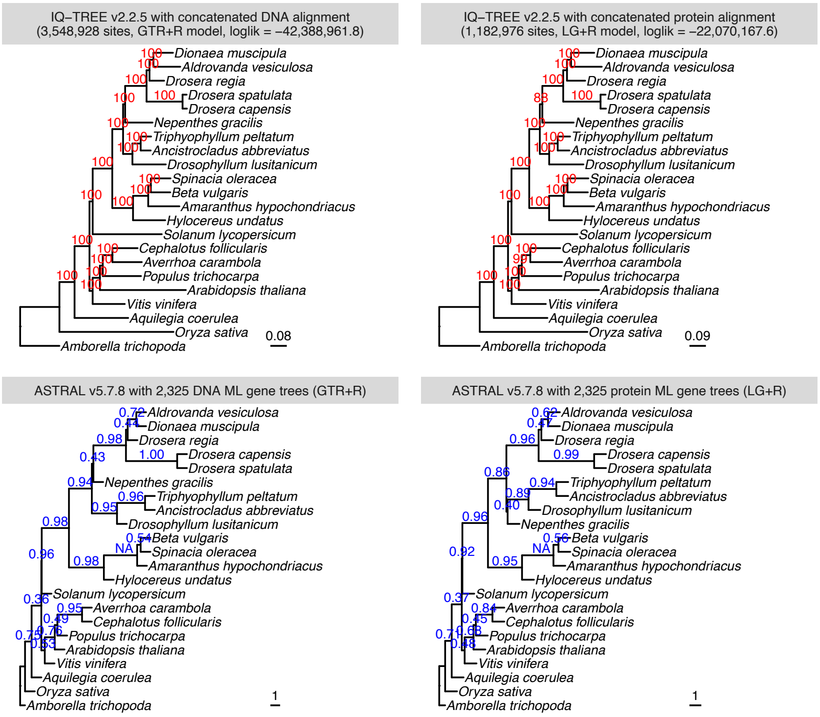
Supplementary Materials Fig. S2: Characterization of CENH3 sequence and centromeric repeats in *D. regia* and *D. capensis*. (a) Sequence alignment of inferred CENH3 proteins found in *D. regia* and *D. capensis*. The antibody against CENH3 in each species is shown below each CENH3 sequence. **(b)** Dot plot highlighting short stretches of identity between the two satDNA families *Cape71* in *D. capensis* and *Regi90* in *D. regia*. **(c)** Size distribution comparison of *Cape71* and *Regi90*. Notice the prevalent array size of 10-20kbp for both species.



Supplementary Materials Fig. S3: Comparison of centromeric sequence homogenization and the abundance of different transposable elements present in the *D. capensis* and *D. regia* genomes. **(a)** Genome-wide ModDot plot histograms of concatenated isolated centromeric arrays of all *D. capensis* chromosomes. **(b)** Close-up view of satellite array homogenization in chromosome 8 centromeric arrays. **(c)** Genome-wide ModDot plot histograms of concatenated isolated centromeric arrays of all *D. regia* chromosomes. **(d)** Close-up view of satellite array homogenization in chromosome 13 centromeric arrays.



Supplementary Materials Fig. S4: Comparison of the most abundant transposable-element (TE) families across *Drosera capensis* and *D. regia* genomes, and comparison of total repeats among 12 *Drosera* species plus *Nepenthes gracilis*. **(a)** Abundance differences of major TE families between *D. capensis* and *D. regia*. **(b)** Distribution of densities of TEs in relation to their position (+- 5000 bp) to nearby satDNA (upper panels), and genes (lower panels) **(c)** Relative contribution of each repetitive family to the total repetitive fraction in 12 *Drosera* species and *N. gracilis*. Black bars above the panels denote the total cluster size for each TE family (columns), while colored bars within each species row indicate the relative abundance of that family in that genome. **(d)** Dot plot comparison of the main satDNA families found in the genome of the 12 *Drosera* species analyzed.

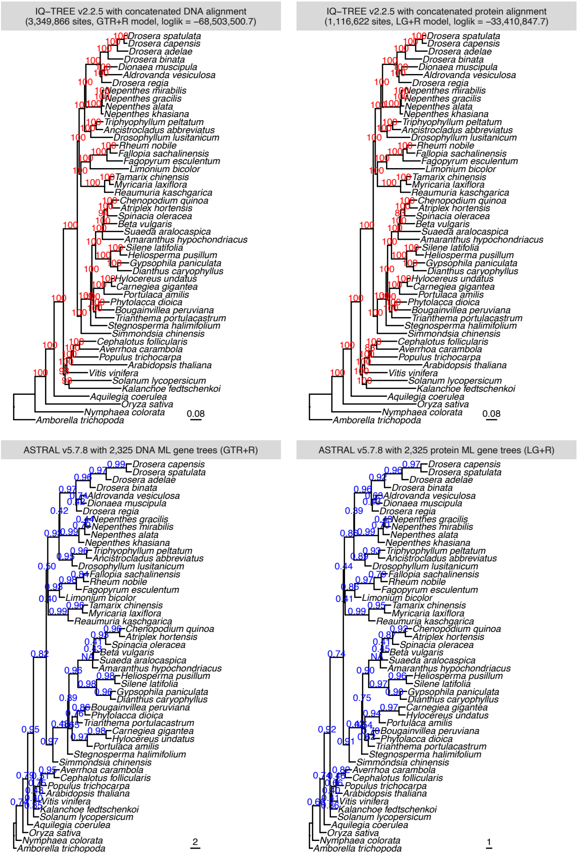


IQ–TREE v2.2.5 with concatenated DNA alignment (3,548,928 sites, GTR+R model, loglik = –42,388,961.8)

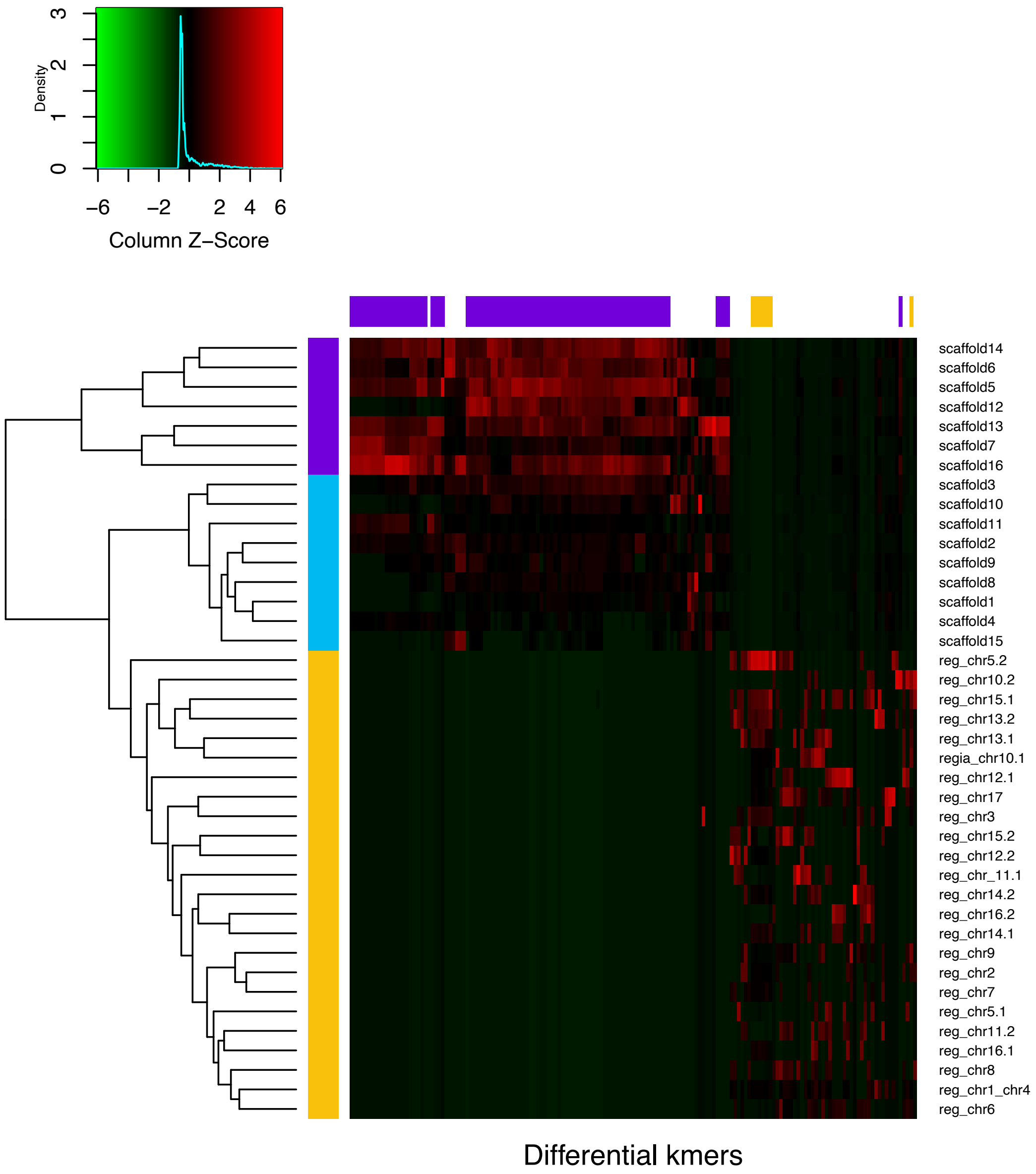
IQ–TREE v2.2.5 with concatenated protein alignment (1,182,976 sites, LG+R model, loglik = –22,070,167.6)

ASTRAL v5.7.8 with 2,325 DNA ML gene trees (GTR+R)

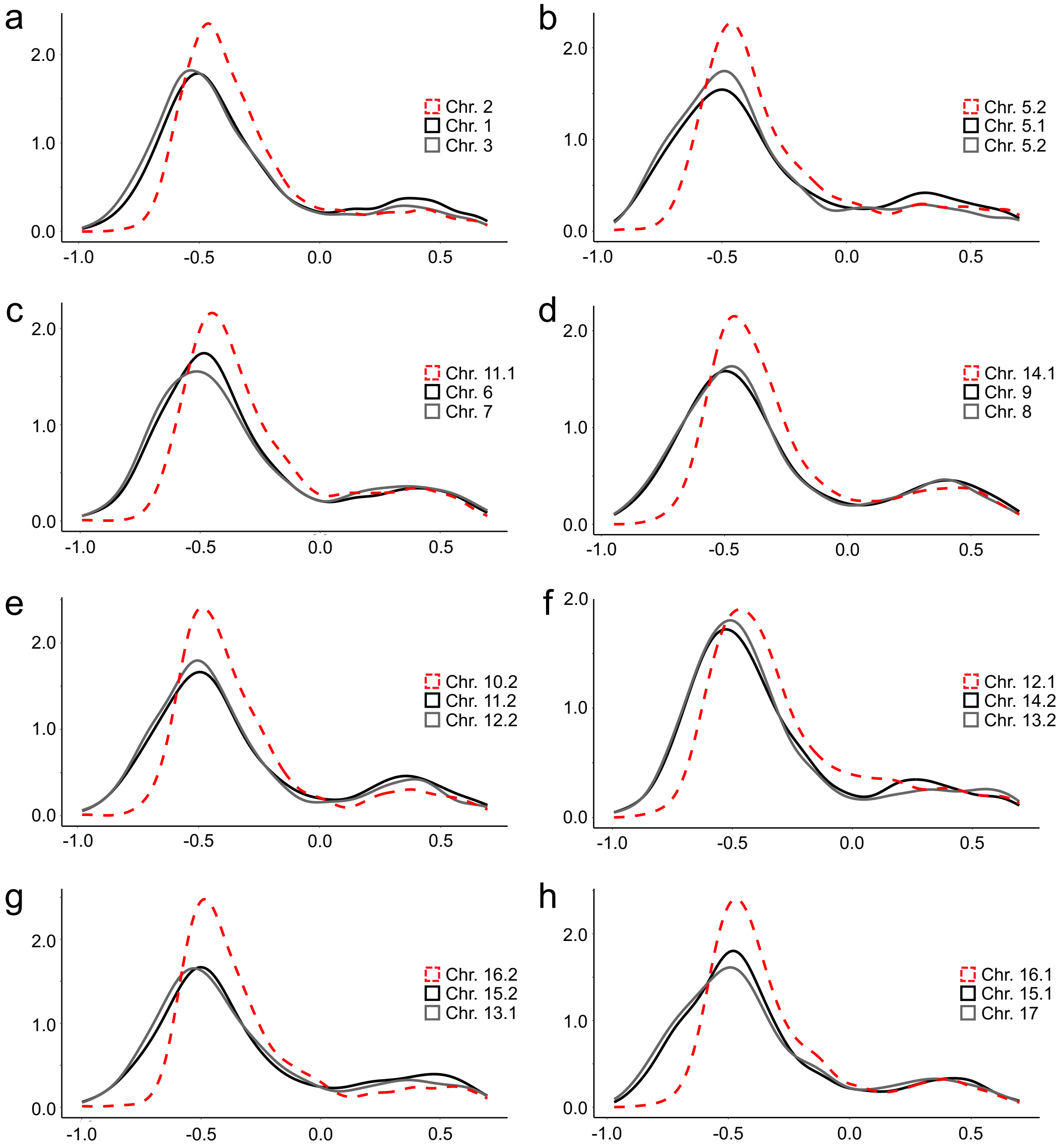
ASTRAL v5.7.8 with 2,325 protein ML gene trees (LG+R)


IQ–TREE v2.2.5 with concatenated DNA alignment
(3,349,866 sites, GTR+R model, loglik = –68,503,500.7)
IQ–TREE v2.2.5 with concatenated protein alignment
(1,116,622 sites, LG+R model, loglik = –33,410,847.7)
0.08
0.08
ASTRAL v5.7.8 with 2,325 DNA ML gene trees (GTR+R)
ASTRAL v5.7.8 with 2,325 protein ML gene trees (LG+R)
2
1

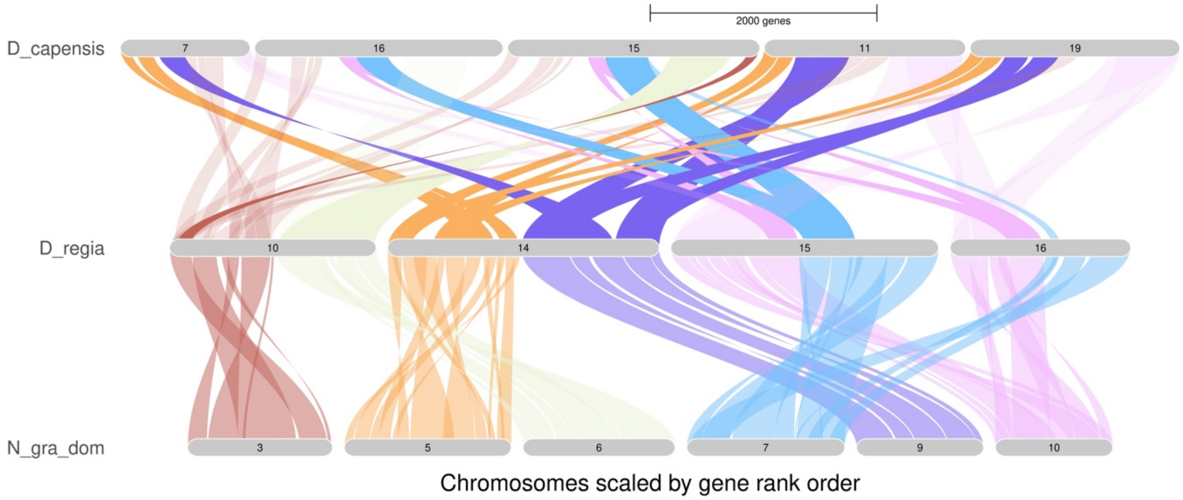
Supplementary Materials Fig. S5. BUSCO gene and species trees. Two datasets were examined: nucleotide coding sequences or inferred amino acids from 22 species or 50 species. Phylogenetic relationships were reconstructed using two different approaches: (1) maximum likelihood (ML) analysis based on concatenated BUSCO gene alignments (**top**), and (2) ASTRAL species tree inference summarizing gene trees from individual BUSCOs (**bottom**). Both methods consistently placed *D. regia* as the sister group to *Aldrovanda* plus *Dionaea*, rather than to other *Drosera* species. *Nepenthes* was generally recovered as sister to *Drosera* and the snap-trapping lineages, while the clade containing *Drosophyllum* and *Triphyophyllum/Ancistrocladus* formed their sister group.



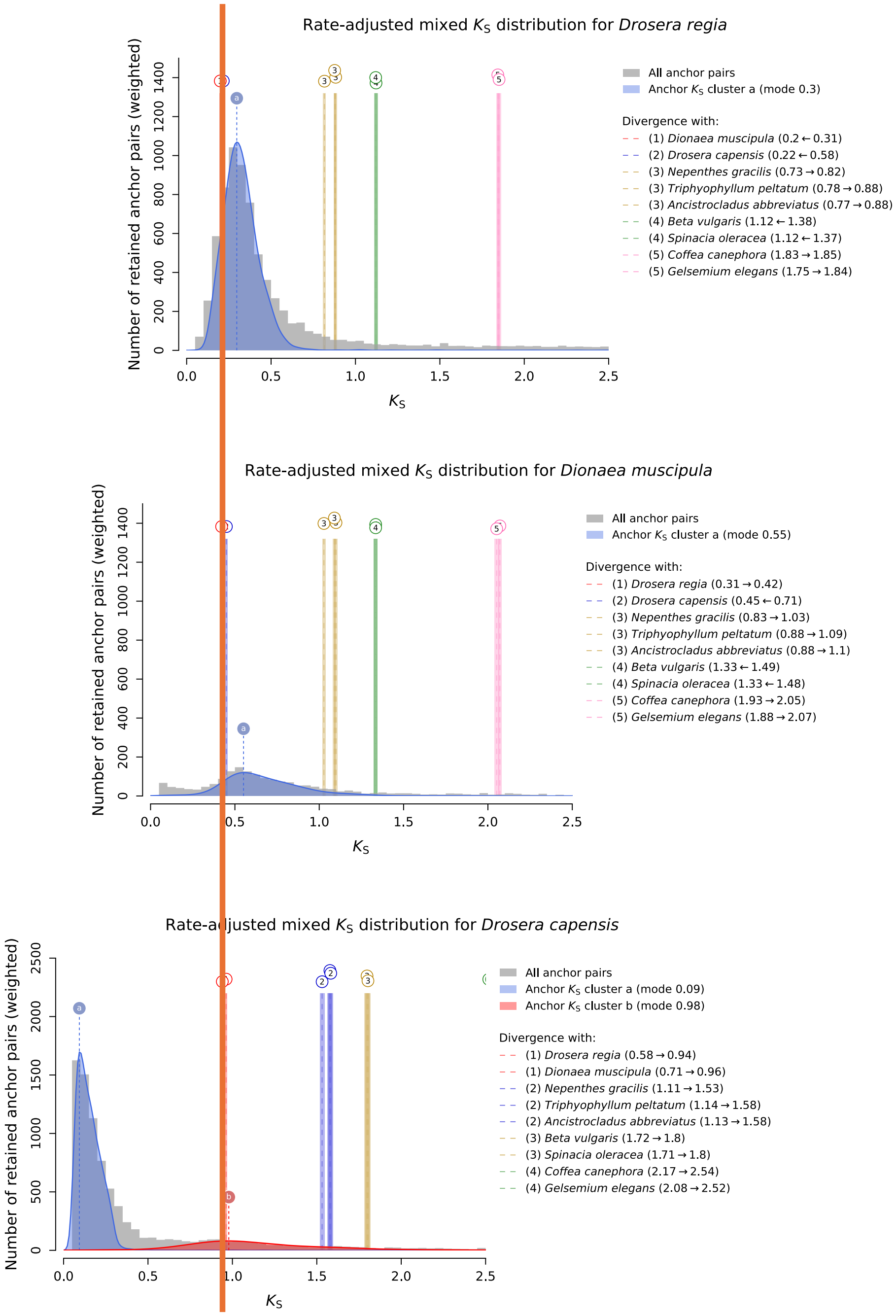
Supplementary Materials Fig. S6: Unsupervised hierarchical clustering of ancestral chromosomes of *D. regia* and modern chromosomes of Dionaea. The horizontal color bar at the top (x-axis) indicates to which subgenome the *k*-mer is specific; the vertical color bar on the left (y-axis) indicates the subgenome to which the chromosome is assigned. The heatmap indicates the Z-scale relative abundance of *k*-mers. The larger the Z-score is, the greater the relative abundance of a *k*-mer. *Dionaea* subgenomes are represented in purple and blue colors, while *D. regia* (yellow) did not show any particular subgenome differentiation based on this approach.



Supplementary Materials Fig. S7: Ks density plots of homeologous gene pairs in *D. regia* reveals a clear triplicate subgenomic structure that exhibits a 2:1 configuration. Two of three chromosomes show similar distributions (gray), while a third (dashed red lines) stands out as an older subgenome, characterized by a higher Ks value (x-axis = $\log_{10}$ Ks; y-axis = density).



Supplementary Materials Fig. S8: Shared chromosomal rearrangements and subgenome relationships across *Drosera* species, as compared to *Nepenthes*. GENESPACE riparian plot illustrating shared chromosomal rearrangements among selected chromosomes from the *N. gracilis* dominant subgenome, *D. regia* and *D. capensis*. This subset of chromosomes highlights key patterns supporting shared ancestral genomic restructuring. Note the orange:purple fusion in *D. regia* chromosome 14, and the three *D. capensis* chromosomes in this view (7, 11 and 19) that appear to show the same fusion. Likewise, note the pink:sky-blue fusion shared by *D. regia* chromosomes 15 and 16, which appears similarly reflected in *D. capensis* chromosomes 15 and 16.


Rate-adjusted mixed K_S distribution for *Drosera regia*
Number of retained anchor pairs (weighted)
0 200 400 600 800 1000 1200 1400
0.0 0.5 1.0 1.5 2.0 2.5
K_S
All anchor pairs
Anchor K_S cluster a (mode 0.3)
Divergence with:
(1) *Dionaea muscipula* (0.2 ← 0.31)
(2) *Drosera capensis* (0.22 ← 0.58)
(3) *Nepenthes gracilis* (0.73 → 0.82)
(3) *Triphyophyllum peltatum* (0.78 → 0.88)
(3) *Ancistrocladus abbreviatus* (0.77 → 0.88)
(4) *Beta vulgaris* (1.12 ← 1.38)
(4) *Spinacia oleracea* (1.12 ← 1.37)
(5) *Coffea canephora* (1.83 → 1.85)
(5) *Gelsemium elegans* (1.75 → 1.84)
Rate-adjusted mixed K_S distribution for *Dionaea muscipula*
Number of retained anchor pairs (weighted)
0 200 400 600 800 1000 1200 1400
0.0 0.5 1.0 1.5 2.0 2.5
K_S
All anchor pairs
Anchor K_S cluster a (mode 0.55)
Divergence with:
(1) *Drosera regia* (0.31 → 0.42)
(2) *Drosera capensis* (0.45 ← 0.71)
(3) *Nepenthes gracilis* (0.83 → 1.03)
(3) *Triphyophyllum peltatum* (0.88 → 1.09)
(3) *Ancistrocladus abbreviatus* (0.88 → 1.1)
(4) *Beta vulgaris* (1.33 ← 1.49)
(4) *Spinacia oleracea* (1.33 ← 1.48)
(5) *Coffea canephora* (1.93 → 2.05)
(5) *Gelsemium elegans* (1.88 → 2.07)
Rate-adjusted mixed K_S distribution for *Drosera capensis*
Number of retained anchor pairs (weighted)
0 500 1000 1500 2000 2500
0.0 0.5 1.0 1.5 2.0 2.5
K_S
All anchor pairs
Anchor K_S cluster a (mode 0.09)
Anchor K_S cluster b (mode 0.98)
Divergence with:
(1) *Drosera regia* (0.58 → 0.94)
(1) *Dionaea muscipula* (0.71 → 0.96)
(2) *Nepenthes gracilis* (1.11 → 1.53)
(2) *Triphyophyllum peltatum* (1.14 → 1.58)
(2) *Ancistrocladus abbreviatus* (1.13 → 1.58)
(3) *Beta vulgaris* (1.72 → 1.8)
(3) *Spinacia oleracea* (1.71 → 1.8)
(4) *Coffea canephora* (2.17 → 2.54)
(4) *Gelsemium elegans* (2.08 → 2.52)
a
b
1 2 3 4 5

Supplementary Materials Fig. S9. *ksrates* synonymous substitutions calibration clarifies split times and polyploidy events for *D. regia*, *D. capensis*, and *Dionaea muscipula*. Here, focal species are (top to bottom) *D. regia*, *Dionaea*, and *D. capensis*. Colorations for events are the same for *D. regia* and *Dionaea*; they share the blue tetraploidy, which is the allotetraploidy still visible in *Dionaea*. For *D. regia*, one third of its hexaploid genome matches *Dionaea*; the hexaploidy was likely time-coincident with *Dionaea*'s allotetraploidy, with *D. regia*'s third subgenome possibly being one of the two original *Dionaea* parents. The species splits for *D. regia*, *Dionaea*, and *D. capensis* (aligned at the orange vertical line) are also time-coincident, lying after the allotetraploidy and allohexaploidy events. While the *D. capensis*-focused analysis at the bottom appears different at a glance, it is not. Due to how the *ksrates* software handles plotting, the most recent tetraploidy is blue, and it occurred after *D. capensis* split from *D. regia* and *Dionaea* (orange line). With color coding changed, the older event shared with *D. regia* and *Dionaea* is red and is barely visible, with its remains in the modern *D. capensis* genome having been overlaid by an extremely recent tetraploidy event.

Supplementary Materials: Tables

	BUSCO Library: eudicot	*Drosera_capensis*	*Drosera_regia*	*Triphyophyllum_peltatum*	*Ancistrocladus_abbreviatus*	*Dionaea_muscipula*
BUSCO	Complete BUSCOs (C)	2,086 (89.7%)	2,179 (93.7%)	2,255 (96.9%)	2,240 (96.3%)	2,128 (91.5%)
	Complete and single-copy BUSCOs (S	1,471 (63.2%)	1,859 (79.9%)	2,089 (89.8%)	2,033 (87.4%)	1,995 (85.8%)
	Complete and duplicated BUSCOs (D)	615 (26.4%)	320 (13.8%)	166 (7.1%)	207 (8.9%)	133 (5.7%)
	Fragmented BUSCOs (F)	37 (1.6%)	27 (1.2%)	18 (0.8%)	29 (1.2%)	37 (1.6%)
	Missing BUSCOs (M)	203 (8.7%)	120 (5.2%)	53 (2.3%)	57 (2.5%)	161 (6.9%)
	Total BUSCO groups searched	2,326 (100.0%)	2,326 (100.0%)	2,326 (100.0%)	2,326 (100.0%)	2,326 (100.0%)
	Number of scaffolds	1,089	1,338	429	131	3,706
	Number of contigs	1,431	3,275	468	653	11,425
	Total length	284,038,097	282,006,236	552,408,181	1,187,329,241	2,551,530,229
	Percent gaps	0	0	0	0	0
	Scaffold N50	12 MB	15 MB	26 MB	64 MB	144 MB
	Contigs N50	1 MB	287 KB	14 MB	4 MB	432 KB

	BUSCO Library: embryophyta	*Drosera_capensis*	*Drosera_regia*	*Triphyophyllum_peltatum*	*Ancistrocladus_abbreviatus*	*Dionaea_muscipula*
BUSCO	Complete BUSCOs (C)	1,530 (94.8%)	1,561 (96.7%)	1,578 (97.8%)	1,572 (97.4%)	1,540 (95.4%)
	Complete and single-copy BUSCOs (S	1,115 (69.1%)	1,390 (86.1%)	1,507 (93.4%)	1,459 (90.4%)	1,466 (90.8%)
	Complete and duplicated BUSCOs (D)	415 (25.7%)	171 (10.6%)	71 (4.4%)	113 (7.0%)	74 (4.6%)
	Fragmented BUSCOs (F)	23 (1.4%)	11 (0.7%)	15 (0.9%)	18 (1.1%)	25 (1.5%)
	Missing BUSCOs (M)	61 (3.8%)	42 (2.6%)	21 (1.3%)	24 (1.5%)	49 (3.0%)
	Total BUSCO groups searched	1,614 (100.0%)	1,614 (100.0%)	1,614 (100.0%)	1,614 (100.0%)	1,614 (100.0%)
	Number of scaffolds	1,089	1,338	429	131	3,706
	Number of contigs	1,431	3,275	468	653	11,425
	Total length	284,038,097	282,006,236	552,408,181	1,187,329,241	255,153,029
	Percent gaps	0	0	0	0	0
	Scaffold N50	12 MB	15 MB	26 MB	64 MB	144 MB
	Contigs N50	1 MB	287 KB	14 MB	4 MB	432 KB

QUAST	# contigs	1,089	1,338	429	131	3,706
	Largest contig	17,145,115	20,363,681	31,991,434	85,057,085	169,349,142
	Total length	284,038,097	282,006,236	552,408,181	1,187,329,241	2,551,530,229
	GC (%)	37	37	40	36	44
	N50	12,650,043	15,761,547	26,241,775	64,216,489	144,508,999
	N75	10,814,855	12,243,869	24,909,253	58,859,448	126,729,089
	L50	10	8	10	9	9
	L75	16	13	16	14	13
	# N's per 100 kbp	34	120	1	4	30

Supplementary Materials Table S1. Quality assessment of the genomes.

Species	Total reads analyzed	Read length	Reads: repetitive Elements	% reads repetitive elements
N. gracilis	236096	150	78331	33.18
D. regia	235232	251	161825	68.79
D. aliciae	236124	251	156822	66.42
D. capensis	235410	251	178862	75.98
D. tokaiensis	235912	145	172730	73.22
D. anglica	235654	101	136961	58.12
D. intermedia	235694	251	203669	86.41
D. filiformis	234720	251	203766	86.81
D. broomensis	235806	251	158050	67.03
D. erythrorhiza	235186	151	113059	48.07
D. peltata	235214	251	195065	82.93
D. menziesii	235850	251	184643	78.29
D. scorpioides	234890	251	172536	73.45

Supplementary Materials Table S2. Short reads analyzed in the repeatome comparative analysis.

Supplementary References

1. Steinmüller, K. & Apel, K. A simple and efficient procedure for isolating plant chromatin which is suitable for studies of DNase I-sensitive domains and hypersensitive sites. *Plant molecular biology* **7**, 87-94 (1986).
2. Hatano, S., Yamaguchi, J. & Hirai, A. The preparation of high-molecular-weight DNA from rice and its analysis by pulsed-field gel electrophoresis. *Plant Science* **83**, 55-64 (1992).
3. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
4. Zimin, A.V. *et al.* Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome research* **27**, 787-792 (2017).
5. Manni, M., Berkeley, M.R., Seppey, M. & Zdobnov, E.M. BUSCO: assessing genomic data quality and beyond. *Current Protocols* **1**, e323 (2021).
6. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075 (2013).
7. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754-1760 (2009).

8. Durand, N.C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems* **3**, 95-98 (2016).
9. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).
10. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nature methods* **7**, 909-912 (2010).
11. Haas, B.J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494-1512 (2013).
12. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907-915 (2019).
13. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290-295 (2015).
14. Gilbert, D. Gene-omes built from mRNA-seq not genome DNA. (2013).
15. Flynn, J.M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451-9457 (2020).
16. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **5**, 4.10. 1-4.10. 14 (2004).
17. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, R7 (2008).
18. Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767-769 (2016).
19. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **34**, W435-W439 (2006).
20. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research* **33**, 6494-6506 (2005).
21. Slater, G.S.C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics* **6**, 31 (2005).
22. Lovell, J.T. *et al.* GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *elife* **11**, e78526 (2022).
23. Albert, V.A. & Krabbenhoft, T.J. Navigating the CoGe Online Software Suite for Polyploidy Research. in *Polyploidy: Methods and Protocols* (ed. Van de Peer, Y.) 19-45 (Springer US, New York, NY, 2023).
24. Joyce, B.L. *et al.* FractBias: a graphical tool for assessing fractionation bias following polyploidy. *Bioinformatics* **33**, 552-554 (2016).
25. Hofstatter, P.G. *et al.* Repeat-based holocentromeres influence genome architecture and karyotype evolution. *Cell* (2022).
26. Jia, K.H. *et al.* SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. *New Phytologist* **235**, 801-809 (2022).
27. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of open source software* **4**, 1686 (2019).
28. Hadley, W. *Ggplot2: Elegrant graphics for data analysis*, (Springer, 2016).

29. Neuwirth, E. & Neuwirth, M.E. Package 'RColorBrewer'. *ColorBrewer Palettes* (2014).
30. Wilke, C.O. Ridgeline Plots in 'ggplot2'[R Package Ggridges Version 0.5. 3]. *January. https://cran. r-project. org/web/packages/ggridges/index. html* (2021).
31. Aphalo, P. ggpmisc: Miscellaneous Extensions to "ggplot2"(R package version 0.3. 6). (2020).
32. Sensalari, C., Maere, S. & Lohaus, R. ksrates: positioning whole-genome duplications relative to speciation events in KS distributions. *Bioinformatics* **38**, 530-532 (2022).
33. Dainat, J. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format (version 1.4.0). (Zenodo, 2022).
34. Saul, F. *et al.* Subgenome dominance shapes novel gene evolution in the decaploid pitcher plant Nepenthes gracilis. *Nature plants* **9**, 2000-2015 (2023).
35. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *science* **345**, 1181-1184 (2014).
36. Emms, D.M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology* **20**, 1-14 (2019).
37. Emms, D.M. & Kelly, S. STAG: Species tree inference from all genes. Preprint at https://doi.org/10.1101/267914. (2018).
38. Emms, D.M. & Kelly, S. STRIDE: species tree root inference from gene duplication events. *Molecular biology and evolution* **34**, 3267-3278 (2017).
39. Novák, P., Hoštáková, N., Neumann, P. & Macas, J. DANTE and DANTE_LTR: lineage-centric annotation pipelines for long terminal repeat retrotransposons in plant genomes. *NAR Genomics and Bioinformatics* **6**(2024).
40. Neumann, P., Novák, P., Hoštáková, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA* **10**, 1 (2019).
41. Novák, P., Neumann, P. & Macas, J. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nature Protocols* **15**, 3745-3776 (2020).
42. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410 (1990).
43. Yu, Y., Ouyang, Y. & Yao, W. shinyCircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics* **34**, 1229-1231 (2017).
44. Oliveira, L. *et al.* KNL1 and NDC80 represent new universal markers for the detection of functional centromeres in plants. *Chromosome Research* **32**, 3 (2024).
45. Macas, J., Neumann, P. & Navrátilová, A. Repetitive DNA in the pea (Pisum sativum L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. *BMC Genomics* **8**, 427 (2007).
46. Katoh, K. & Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772-780 (2013).
47. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490 (2010).
48. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic acids research* **52**, W78-W82 (2024).
49. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2018).

50. Wickham, H. dplyr: A grammar of data manipulation. *R package version 04.* **3**, p156 (2015).
51. Montero, H., Freund, M. & Fukushima, K. Convergent losses of arbuscular mycorrhizal symbiosis in carnivorous plants. *bioRxiv*, 2025.04. 03.646726 (2025).
52. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541-i548 (2014).
53. Wong, T.K. *et al.* IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models. (2025).