Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genomic Evidence for Convergent Adaptation through Tandem Gene Duplications in Carnivorous Butterworts and Bladderworts

Victor Albert

vaalbert@buffalo.edu

University at Buffalo https://orcid.org/0000-0002-0262-826X

Steven Fleck

University at Buffalo https://orcid.org/0000-0001-5370-2258

Jonathan Kirshner

University of Buffalo https://orcid.org/0009-0008-5521-0351

Sitaram Rajaraman

Tianying Lan

Crystal Tomlin

Charlotte Lindqvist

University at Buffalo https://orcid.org/0000-0002-4190-727X

Richard Jobson

Enrique Ibarra-Laclette

Luis Herrera-Estrella

Texas Tech University

Martin Mata-Rosas

Article

Keywords:

Posted Date: September 26th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7552968/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Genomic Evidence for Convergent Adaptation through Tandem Gene Duplications in Carnivorous Butterworts and Bladderworts

Victor A. Albert[1,*,#], Steven J. Fleck[1,*], Jonathan Kirshner[1], Sitaram Rajaraman[2], Tianying Lan[1], Crystal Tomlin[3], Charlotte Lindqvist[1], Richard W. Jobson[5], Enrique Ibarra-Laclette[6], Luis Herrera-Estrella[4], Martín Mata-Rosas[7], Todd P. Michael[8]

1. Department of Biological Sciences, University at Buffalo, Buffalo, NY 14260, USA
2. Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland
3. Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, State University of New York, Buffalo, NY, USA.
4. Institute of Genomics for Crop Abiotic Stress Tolerance, Department of Plant and Soil Science, Texas Tech University, Lubbock, TX 79409-3135
5. Botanic Gardens of Sydney, Mrs Macquaries Rd, Sydney NSW 2000, Australia
6. Red de Estudios Moleculares Avanzados, Instituto de Ecología A.C., Carretera Antigua a Coatepec 351, El Haya, C.P. 91070, Xalapa, Ver, Mexico
7. Red Manejo Biotecnológico de Recursos, Instituto de Ecología, A.C., Carretera antigua a Coatepec 351, Xalapa, Veracruz, 91073, Mexico
8. Salk Institute for Biological Studies, La Jolla, CA, USA

[*]These authors contributed equally

[#]For correspondence: vaalbert@buffalo.edu (VAA)

Abstract

Cases of convergent adaptation, especially between close relatives within a lineage, provide insights into constraints underlying the mechanisms of evolution. We examined this in the carnivorous plant family Lentibulariaceae, with its highly divergent trap designs but shared need for prey digestion, by generating a chromosome-level genome for *Pinguicula gigantea*, the giant butterwort. Our work reveals a history of whole-genome duplication in the genus and provides strong phylogenomic evidence for a sister-group relationship between Lentibulariaceae and Acanthaceae. The genome also shows that a key digestive adaptation, the expansion of cysteine protease genes, arose through independent tandem duplications in *Pinguicula* and its close relative *Utricularia*. These parallel expansions arose on homologous regions in both genomes, consistent with a common ancestral gene occupying the same chromosomal context. This study provides genomic evidence for convergent evolution of digestive potential in a carnivorous plant lineage and illustrates how similar selective pressures can repeatedly shape genomes in analogous ways.

Introduction

The carnivorous syndrome, which encompasses prey attraction, capture, digestion, and nutrient absorption, is a classic example of convergent evolution in flowering plants. This complex trait has evolved independently at least eleven times, spanning thirteen families and six orders[1-4]. Carnivorous plants have adapted to nutrient-poor soils by evolving specialized mechanisms to obtain nutrients from animal prey[5].

The morphologically diverse Lentibulariaceae family, which includes *Pinguicula* (butterworts), *Genlisea* (corkscrew plants), and *Utricularia* (bladderworts), accounts for approximately 45% of all carnivorous plant species. Despite a single proposed origin of carnivory within the family, its members exhibit three highly divergent trapping mechanisms: *Utricularia* uses active suction bladder traps, *Genlisea* employs passive eel-trap structures, and *Pinguicula* utilizes passive adhesive leaves covered in mucilaginous glands[6].

Several genomic resources have been developed for Lentibulariaceae, including assemblies for *Genlisea aurea*[7], *Utricularia gibba*[8,9], *Utricularia reniformis*[10], *Pinguicula moranensis*, and *Pinguicula primuliflora*[11]. However, only *U. gibba* has been assembled to a chromosome-scale, highlighting the need for more contiguous reference genomes within this family. While the monophyly of the family is clear, the phylogenetic placement of Lentibulariaceae within Lamiales has been unstable, with proposed sister-group relationships to Stilbaceae[2], Schlegeliaceae[12], Pedaliaceae, and Acanthaceae[13-15]. Importantly, however, all of these results have been based on limited gene sets, a problem that should be solvable by comparing coding sequences at whole-genome scale.

Tandem gene duplications have been implicated as a key mechanism in the evolution of carnivory. The *U. gibba* genome contains tandem arrays of cysteine protease genes with trap-specific expression, enriched for functions related to prey digestion and nutrient transport[9]. Gene copy number of cysteine proteases was further potentiated by whole-genome duplications that occurred during *U. gibba*'s evolutionary history. Similarly, the pitcher plant *Nepenthes gracilis* shows subgenome dominance within its decaploid architecture, with many trap-specific and digestion-related tandem clusters located on recessive subgenomes[16]. Beyond digestion, tandem duplication has also been linked to trap morphology. In the Australian pitcher plant *Cephalotus follicularis*, tandem arrays of cytochrome P450 and wax ester synthase genes are expressed predominantly in the pitcher. These genes are implicated in the formation of the slippery waxy surface critical for prey capture[17]. In *Pinguicula*, proteomic studies of digestive secretions have identified several enzymes involved in prey breakdown, including proteases, peroxidases, esterases, nucleases, and amylases[18]. While some of the genetic basis of these functions may include small-scale gene duplications as a fundamental strategy, the architectural arrangement of such genes in butterwort genomes remains unknown.

To address these gaps, we generated a chromosome-level genome assembly for *Pinguicula gigantea* and two additional high-quality assemblies for *P. agnata* and *P. moctezumae*. A draft genome for a fourth species, *P. grandiflora*, was assembled to a level satisfactory for homology-based gene annotation purposes. Using these resources, we: 1) reconstructed the polyploid history of *Pinguicula*; 2) resolved the phylogenetic position of Lentibulariaceae; 3) identified tandem duplications of digestive enzyme genes; and 4) demonstrated independent expansions of cysteine protease gene clusters in *Pinguicula* and *Utricularia*.

Results

DNA Sequencing and Genome Assembly. We sequenced four *Pinguicula* species using both Oxford Nanopore Technologies (ONT) and Illumina platforms. ONT sequencing yielded between 28.4 and 55.5 Gb of data per species, with read N50 values ranging from 17.5 to 37.9 kb (Table S2). Illumina sequencing produced between 34.8 and 40.3 million 251-bp read pairs per species (Table S3). RNA-seq data were also generated for annotation purposes (Table S4).

The *P. gigantea* genome was assembled into 884 contigs, which were subsequently purged to 405 contigs to remove heterozygosity (Table S5). Using Dovetail Hi-C data (66.4 million read pairs), we scaffolded the assembly into 11 pseudomolecules with 10 additional small scaffolds (Figure 1). The final chromosome-scale assembly was 608.7 Mb, closely matching flow cytometry estimates (598 Mb[19]), and showed 96.8% complete BUSCOs[20] with 14.3% of these duplicated.

For *P. grandiflora*, the primary assembly (1,879 contigs) was scaffolded using Dovetail Omni-C data (62.4 million read pairs), resulting in 1,517 scaffolds. The final assembly size was 570.7 Mb, considerably larger than the flow cytometry estimate of 424 Mb[19]. Owing to this discrepancy and other quality concerns, the assembly was used only for homology-based annotation.

P. agnata and *P. moctezumae* genomes were first assembled using long reads, followed by purging to reduce heterozygosity (Table S6). The final assembly sizes were 676.0 Mb and 586.2 Mb, respectively, slightly larger than flow cytometry estimates (651 Mb and 572 Mb[19]). Both assemblies were scaffolded against the *P. gigantea* reference, with over 98% of bases placed into scaffolds (Table S7).

Genome Annotation and Repeat Content. *De novo* annotation of the *P. gigantea* Hi-C assembly predicted 35,940 genes, with 92.3% complete BUSCOs, 11.0% of these duplicated (Table S5). The *P. grandiflora* annotation contained 33,294 genes but showed higher duplicate gene content (33.8% duplicated BUSCOs), consistent with our assembly concerns. Annotations for *P. agnata* and *P. moctezumae* were generated using homology-based methods.

Repeat annotation revealed that Ty1/Copia LTR retrotransposons were the most abundant repeat class in *P. gigantea*, *P. agnata*, and *P. moctezumae* (Table S8). In contrast, *P. grandiflora* showed a distinct pattern, with Gypsy/DIRS1 elements comprising 15.31% of the genome and unclassified repeats accounting for 43.09%.

Phylogenomic Placement of Lentibulariaceae. To resolve the phylogenetic position of Lentibulariaceae, we constructed a species tree from 2,322 single-copy BUSCO proteins across 30 Lamiales and 2 Gentianales genomes. Our analysis robustly placed Lentibulariaceae as sister to Acanthaceae (Figure 2a), supporting some previous findings while using far more evidence[13-15].

Whole-Genome Duplication History. Synteny analyses revealed the complex polyploid history of *Pinguicula*. Self-alignments of the *P. gigantea* genome showed widespread internal synteny indicative of a recent whole-genome duplication event (WGD; Figure S2). Comparisons with *Vitis vinifera* (which has undergone no additional polyploidy events since the *gamma* hexaploidy event at the base of core eudicots) showed a 4:1 syntenic gene ratio (Figure S3), suggesting two WGDs in the *P. gigantea* lineage. The first WGD is shared across core Lamiales[21], while the second appears to be specific to *Pinguicula*.

This interpretation was further supported by comparisons with *Mimulus guttatus*, which has experienced only the core Lamiales WGD. The syntenic gene ratio between *P. gigantea* and *M. guttatus* was 4:2 (Figure S4), confirming an additional WGD in *Pinguicula* after their divergence.

The *P. grandiflora* assembly showed a 3:2 syntenic ratio against *P. gigantea* (Figure S5), deviating from the expected 4:2 ratio for a tetraploid[22], further justifying its exclusion from structural analyses.

Analyses of *P. agnata* and *P. moctezumae* initially showed aberrant syntenic ratios and synonymous substitution rate (Ks) distributions due to uncollapsed heterozygosity (Figures S6-S9). After purging to remove haplotigs, both assemblies showed the expected 2:2 ratio against *P. gigantea* (Figures S6c,d & S7c,d), consistent with their shared diploid chromosome number (2n=22[22]).

Comparisons with *Utricularia gibba* revealed syntenic depths of 2:3 to 2:6 (Figure S10), indicating that *U. gibba* has undergone more WGD events than *P. gigantea* since their divergence[8,9].

Tandem Duplications of Digestive Enzyme Genes. We identified a large tandem array of cysteine protease genes on chromosome 8 of *P. gigantea* (Figure 3). This cluster is syntenic with four cysteine protease arrays in *U. gibba* - on scaffolds (unitigs) 699, 85, 578, and 8 – which are WGD-derived homeologous blocks. Interestingly, trap-specific expression was reported for the arrays on *U. gibba* scaffolds 699 and 85 (Figures S11, S13; ref. [9]). As noted, *P. gigantea* harbors a

large cysteine protease tandem array on chromosome 8, and while its homeologous block on chromosome 5 derives from the same lineage-specific WGD, it retains only a single cysteine protease homolog. This pattern is consistent with tandem expansion confined to chromosome 8 after WGD, rather than parallel expansion and complete loss on chromosome 5. *P. agnata* and *P. moctezumae* also contain one *P. gigantea*-syntenic block with the tandem duplication and one without (Figure S12), indicating that whatever the evolutionary context of the tandem array may be, speciation occurred following any large-scale genome structural changes.

We identified 18 tandemly duplicated cysteine protease genes in *P. gigantea* (Table S9). Likely indicative of annotation incompleteness, domain analysis revealed that only three of these proteins contain the canonical cysteine active site (IPR000169), and none contain all three residues of the catalytic triad (Table S10). Phylogenetic analyses of CDS and protein sequences show taxon-specific clades for the tandem arrays (Figure S13), consistent with independent expansions from shared ancestral genes, possibly a single-copy locus.

Multiple WGD-derived *U. gibba* blocks (scaffolds 8, 85, 699, and 578) are syntenic to the *P. gigantea* cysteine protease blocks (chromosomes 8 and 5); for example, *U. gibba* scaffold 8 is syntenic to *P. gigantea* chromosome 8 (Figure S13). Given this positional homology but phylogenetic separation between the two species, we infer a single ancestral cysteine protease gene in the shared genomic block of the *Mimulus*/*Utricularia*/*Pinguicula* ancestor (retained as a single copy in *Mimulus*) from which the *Pinguicula* and *Utricularia* tandem arrays independently expanded.

We also identified smaller tandem duplications for other digestive enzymes, including aspartic proteases, peroxidases, and esterase/lipases (Figures S14-S17, Table S9). However, no tandem arrays were found for beta-xylosidase, desiccation-related protein, or alpha-amylase genes (Figures S18-S20).

Repeat analysis around the *P. gigantea* cysteine protease cluster revealed a predominance of unknown elements (63.9%), with LTR/Copia (10.4%) and simple repeats (10.1%) also abundant (Table S11). This profile differs from that of the *U. gibba* cluster regions, which are enriched for simple repeats (22.6%) and have lower LTR/Copia content (3.0%) (Table S12), suggesting different repeat elements may have facilitated the independent duplication events through unequal crossovers.

Discussion

We present the first chromosome-level genome assembly for the carnivorous plant *Pinguicula gigantea*, providing a high-quality resource for comparative genomics within Lentibulariaceae and across carnivorous plants. Our analyses reveal a complex history of polyploidy in *Pinguicula*, with at least two WGD events since the *gamma* hexaploidy event. The first is shared with other

core Lamiales[21], while the second is specific to the *Pinguicula* lineage. Our phylogenomic analysis robustly places Lentibulariaceae as sister to Acanthaceae, a position supported by several recent studies[13-15]. These findings resolve previous uncertainties about the polyploid history of the genus and provide context for interpreting its genomic evolution.

A key finding of our study is the independent evolution of tandem cysteine protease arrays in *Pinguicula* and *Utricularia*. While the genomic contexts of the arrays are syntenic across the genera (with only a single, likely transposed *U. gibba* gene on scaffold 32), several lines of evidence support convergent evolution of duplicate gene expansions: (1) phylogenetic trees show species-specific clustering of large arrays, indicating duplications occurred after species divergence; (2) tandem duplications occurred in the same homologous, WGD-derived blocks in each genus, indicating independent expansions from a single ancestral locus; and (3) repeat elements surrounding the clusters differ between genera, suggesting distinct, lineage-specific repeat environments that likely facilitated expansions through unequal crossover events. The single and syntenic cysteine protease gene of *M. guttatus* likely also represents the genomic state in the pre-carnivorous Lentibulariaceae ancestor. As such, our results represent a clear example of convergent evolution at the genomic level, where similar selective pressures (prey digestion) have led to similar genomic outcomes (tandem duplication of digestive enzyme genes) in closely related carnivorous plant lineages with radically different trapping strategies and prey species profiles.

The functional significance of the cysteine protease gene arrays in *Pinguicula* remains to be fully determined. While *U. gibba* shows clear trap-specific expression of its tandem arrays[9], similar expression analyses in *Pinguicula* are challenging due to its lack of dedicated non-trap vegetative tissues. Domain analyses suggest that not all duplicated genes retain full catalytic activity, which may reflect annotation incompleteness, or more interestingly, actual subfunctionalization or neofunctionalization within the arrays.

Beyond cysteine proteases, we found smaller tandem duplications for other digestive enzymes, suggesting that tandem duplication may be a general mechanism for evolving digestive capabilities in *Pinguicula*. The absence of tandem arrays for some enzymes (e.g., amylase and xylosidase) implies that other mechanisms, such as regulatory changes or gene co-option, may also contribute to the butterwort carnivorous syndrome.

Synthesis and Broader Implications. Our discovery of independently derived tandem duplications for digestion in Lentibulariaceae, when viewed alongside similar findings in other carnivorous lineages, suggests a broader genomic principle. In *Nepenthes*, tandemly duplicated clusters are linked to trap-specific functions like proteolysis, nutrient transport and chitin degradation[16]. Furthermore, in *Cephalotus*, tandem arrays of cytochrome P450 and wax ester synthase genes show pitcher-predominant expression and are hypothesized to be crucial for the

development of the trap's slippery, waxy surface[17]. The convergence on tandem duplication as a genomic solution to the challenges of carnivory, spanning digestion, nutrient absorption, and prey capture, across these distantly related orders (Lamiales, Caryophyllales, and Oxalidales) is significant. It suggests that the modular nature of tandem arrays provides an evolvable genetic architecture that may permit rapid gene-dosage amplification[9]. This structural specialization may also offer opportunities for neo- and subfunctionalization of duplicate genes, enabling the repeated and independent evolution of complex, coordinated traits like carnivory under selective pressure in nutrient-poor environments.

Conclusions

In conclusion, our study provides a key genomic resource in the form of a chromosome-scale assembly for *P. gigantea* that enables detailed comparative studies within carnivorous plants. It offers strong evidence for a sister-group relationship between Lentibulariaceae and Acanthaceae and clarifies the history of polyploidy in the *Pinguicula* lineage, revealing two whole-genome duplication events, one specific to the genus. Most significantly, we demonstrate that the expansion of digestive enzyme genes in *Pinguicula* and *Utricularia* occurred via independent tandem duplications, constituting a clear example of convergent evolution at the genomic level. By integrating our results with previous studies on *Nepenthes* and *Cephalotus*, we propose that tandem gene duplication is a major, recurrent genomic mechanism facilitating the evolution of plant carnivory across independent lineages. This genome resource establishes a foundation for future functional studies on the genetic mechanisms underlying carnivory, particularly the expression and regulation of these convergent tandem arrays, and for broader structural analyses across other carnivorous plants of the order Lamiales[2].

Methods

Plant Material and DNA Sequencing. Tissue cultured plant material for *Pinguicula agnata*, *P. gigantea*, *P. grandiflora*, and *P. moctezumae* was obtained from the Red Manejo Biotecnológico de Recursos, Instituto de Ecología, Mexico.

Oxford Nanopore Technologies (ONT) reads were generated at the J. Craig Venter Institute, La Jolla, California. Corrupted ONT reads for *P. agnata* were removed using fastqwiper v1.0.5-alpha (https://github.com/mazzalab/fastqwiper). Read quality was assessed with NanoPlot v1.42.0[23] (ONT) and FastQC[24] v0.11.9 (Illumina). RNA-seq was performed on *P. gigantea* and *P. grandiflora* tissues using Illumina 150 bp paired-end sequencing (Table S4).

Genome Assembly. *De novo* assemblies were generated using ONT long reads and Illumina short reads. For *P. gigantea*, we used Minimap2 v2.16 and Miniasm v0.3-r179[25,26]. Racon[27] v1.3.3 was run three times for consensus calling, followed by three rounds of Pilon[28] v1.23 error correction with Illumina data (Table S3). Genomes were indexed using SAMtools[29] version

0.1.19[29] and BWA version 0.7.17-r1188[30]. ONT long reads were mapped against the genome using Minimap2 version 2.16, Illumina short reads were mapped against the genome using bwa version 0.7.17-r1188, and the resulting BAM files were sorted, again using SAMtools version 0.1.19. The assembly was purged using Purge Haplotigs[31] v1.1.1.

A Dovetail Hi-C library was prepared for *P. gigantea* as described[32]. The assembly was scaffolded using HiRise[33] with 66.4 million read pairs.

P. grandiflora was assembled similarly, but with ONT reads <10 kb filtered using NanoFilt[23] v2.5.0 and without purging. A Dovetail Omni-C library was used for scaffolding with 62.4 million read pairs.

P. agnata and *P. moctezumae* were assembled using Minimap2 v2.24 and Miniasm v0.3-r179, with five rounds of Racon v1.5.0 and three rounds of Pilon v1.23. Purge Haplotigs v1.1.2 was used with repeat-masked BED files. The assemblies were scaffolded using RagTag[34] v2.1.0 with the *P. gigantea* reference.

All assemblies were assessed with QUAST[35] v5.0.2 and BUSCO v5.5.0 (embryophyta_odb10).

Genome Annotation. *P. gigantea* and *P. grandiflora* were annotated using EVidenceModeler[36,37] v1.1.1 to combine evidence from transcriptome assemblies and gene predictions. RNA-seq reads were trimmed with Trimmomatic[38] v0.39. Repeats were identified with RepeatModeler[39] v2.0.1 and masked with RepeatMasker[40] v4.1.0, including LTR discovery with LTRharvest[41] and LTR_retriever[42] v2.9.0.

Four transcriptome assemblies were generated: Trinity[36,43,44] v2.12.0; HISAT2[44] v2.2.1 + StringTie[45] v2.1.5; STAR[46] v2.7.9 + StringTie; and TransAbyss[47] v2.0.1. These were merged with EvidentialGene[48]. PASA[49] v2.5.1 was used with pblat[50] v2.5, GMAP[51], and Minimap2 v2.20 aligners.

Homology-based prediction used GeMoMa[52,53] v1.7.1 with *U. gibba*[9], *Mimulus guttatus*[54], *Tectona grandis*[55], *Sesamum indicum*[56], and *Callicarpa americana*[57] as references. *Ab initio* prediction used GeneMark-ES v4.68[58] and BRAKER[58] v2.1.6.

Annotations for *P. agnata* and *P. moctezumae* were generated with GeMoMa v1.9 using default settings, the *P. gigantea* and *P. grandiflora* annotations, and proteomes of five other species as references: *U. gibba*, *G. aurea*, *M. guttatus*, and *Arabidopsis thaliana*[59] as references. Furthermore, gene models were independently predicted against the masked genome assembly using GeneMark-ES version 4.68 and BRAKER version 2.1.6 (mapping done with STAR version 2.7.9 and sorted using SAMtools version 1.11). The PASA, BRAKER, GeneMark, and GeMoMa gff3 annotation files were combined and filtered using EVidenceModeler with segment size set

to 100,000 bp and overlap size set to 10,000 bp. Annotation completeness was measured by running BUSCO version 5.5.0 embryophyta_odb10.

Phylogenomic Analysis. BUSCO v5.4.4 (eudicots_odb10) was used to extract single-copy orthologs from 32 species (see Figure 2a). Proteins were aligned with MAFFT[60] v7.490, trimmed with trimAl v1.4.1 (-automated1), and used to build gene trees with IQ-TREE v1.6.12 (1000 SH-aLRT and UFBoot replicates, ModelFinder). A species tree was inferred from 2,322 gene trees using ASTRAL[61] v5.7.8.

Synteny and Ks Analysis. Select genome assemblies were analyzed on the CoGe platform[62]: *Coffea canephora* (CoGe id63467), *Catharanthus roseus* (CoGe id65259), *Gelsemium elegans* (CoGe id64491), *Gelsemium sempervirens* (CoGe id53941), *Gardenia jasminoides* (CoGe id66926), *Ophiorrhiza pumila* (CoGe id63710), *Prunus persica* (CoGe id36122), *Theobroma cacao* (CoGe id64859), and *Vitis vinifera* (CoGe id19990). For self-alignments, SynMap2[63] was used to generate syntenic dot plots and Ks histograms. For cross-species comparisons, SynMap2 with the Syntenic Path Assembly option was used to determine syntenic depth. Fractionation bias was assessed using SynMap's FractBias tool[64].

Ks values for syntenic paralogs were downloaded from CoGe and visualized in R using tidyverse[65], ggplot2[66], RColorBrewer[67], ggridges[68], ggpmisc[69], and data.table (DOI: 10.32614/CRAN.package.data.table).

Ksrates[64] v1.1.3 was used to position speciation events relative to WGDs. CDS files were extracted with AGAT v0.9.2. Paralogous Ks peaks were generated using the genome assembly for *Pinguicula gigantea*. Orthologous Ks peaks were generated using combinations of species pairs for each species pair. We justified the Ks value assigned to the gamma hexaploidy event by comparing Ks peaks from multiple other species that have not undergone additional WGD since that event (Figure S21). The phylogeny in Figure 2c was used as input for Ksrates.

Tandem Duplication Analysis. Digestive enzyme sequences from *Pinguicula × Tina*[18] were used as queries in CoGeBLAST against the *P. gigantea* assembly. Syntenic regions were visualized with GEvo. MCScan was used for synteny analysis, with annotations converted to gff3 using AGAT v0.9.2.

For phylogenetic trees of digestive enzymes, protein and CDS sequences were retrieved using BLAST+ v2.12.0. Alignments were made with MAFFT v7.490, trimmed with trimAl[70] v1.4.1 (-automated1), and trees built with IQ-TREE[71] v1.6.12 (using ModelFinder[72], 1000 SH-aLRT and UFBoot replicates). Sequences from multiple species (see Figure 2a) were included for context.

InterProScan[73] was used to identify protein domains in *P. gigantea* cysteine proteases. These domains were compared with those from *Arabidopsis*, *Drosera adelae*, and *U. gibba*. Notably, only three *P. gigantea* proteins contained IPR000169, the cysteine active site which is part of a

catalytic triad along with IPR025660 and IPR025661 (https://www.ebi.ac.uk/interpro/entry/InterPro/IPR000169/)[73]. The three *Pinguicula gigantea* proteins with this active domain were missing IPR025661 or both IPR025660 and IPR025661.

Data Visualization. The Circos[74] plot was generated by parsing RepeatModeler output and gene annotations to create density tracks for Gypsy, Copia, and genes. Synteny lines were derived from SynMap2[63] output. Circos software was used for visualization.

Acknowledgments

We thank Tanya Renner (Penn State University) for RNA-seq library preparation. This work was supported by NSF award 2030871 to VAA.

Author Contributions

VAA conceived and led the project. MM-R, LH-E, and EI-L provided plant material. CT, CL and TL prepared DNA. TPM sequenced DNA and produced initial assemblies. SJF, JK, and SR performed genome annotation and analysis. SJF conducted most genomic analyses. JK generated the Circos plot and BUSCO tree. RJ provided important input on Lentibulariaceae phylogeny. SJF and VAA wrote the manuscript with input from JK. All authors reviewed and approved the final manuscript.

Data Availability

Genome assemblies and annotations are available at CoGe (genomevolution.org) via the following link: https://genomevolution.org/coge/SearchResults.pl?s=pinguicula&p=genome. The data will also be made available at this Dryad link upon publication: https://doi.org/10.5061/dryad.b8gtht7p2. Raw sequencing data will be deposited in the NCBI SRA under BioProject PRJNA[XXXXXX] upon publication.

Main Text Figures

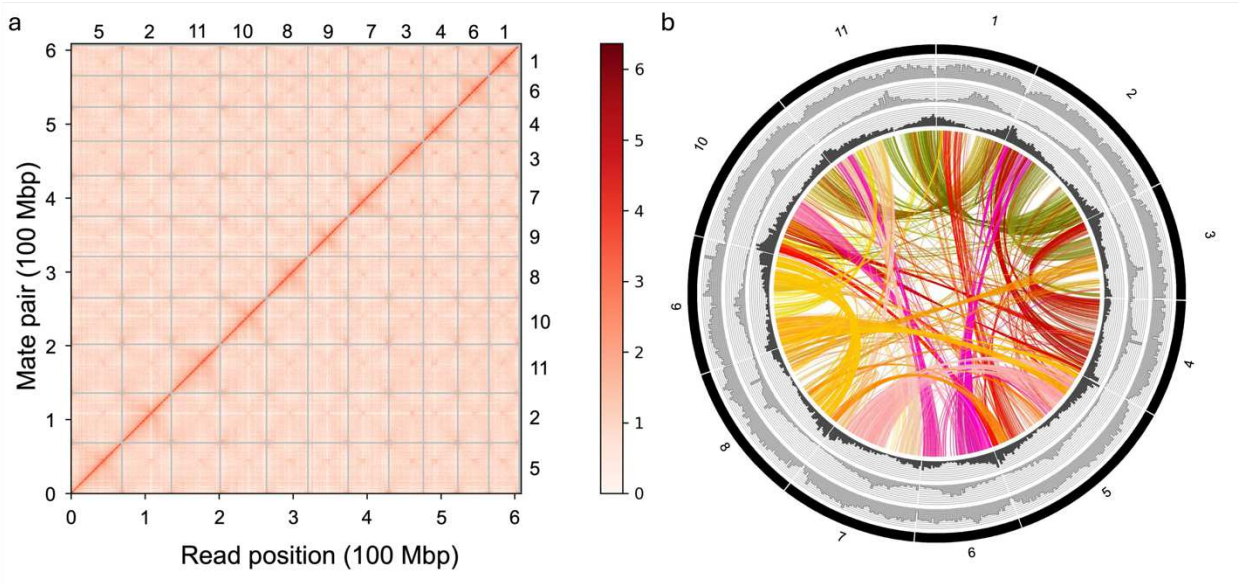


Figure 1: Chromosome-scale architecture of the *Pinguicula gigantea* genome. (a) (a) Darker red indicates a stronger signal. Axes show genomic coordinates (left/bottom). Chromosome numbers are on the right and top. **(b)** From the outer to the inner ring, the tracks represent: Chromosome number, LTR/Copia repeat density, LTR/Gypsy repeat density, and gene density. Putative centromeric regions are highlighted by peaks of increased transposable element (TE) density and decreased gene density. The plot also includes self-self syntenic connections, illustrating regions of genomic similarity.

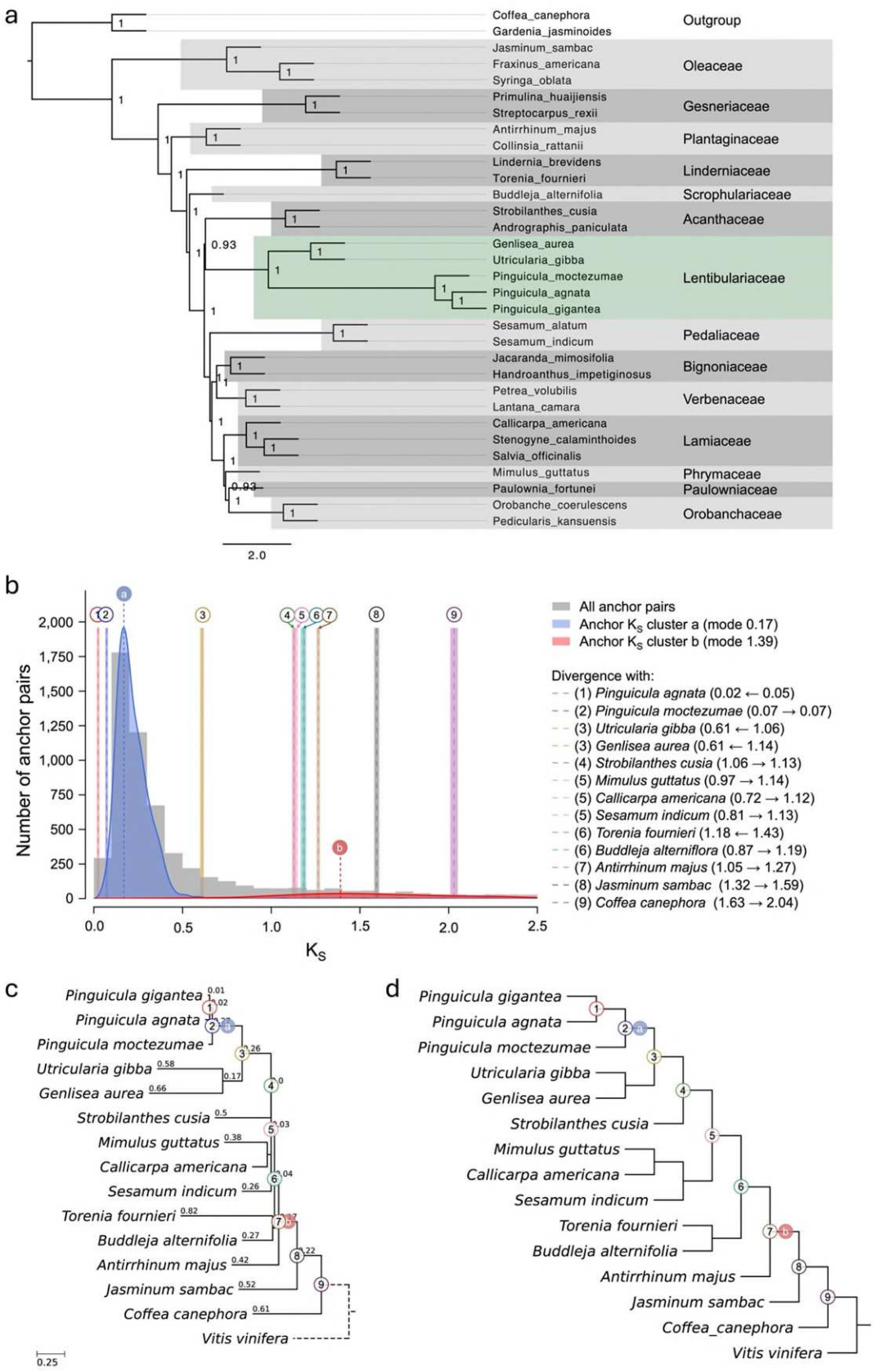

a
Coffea_canephora
Gardenia_jasminoides
Outgroup
Jasminum_sambac
Fraxinus_americana
Syringa_oblata
Oleaceae
Primulina_huaijiensis
Streptocarpus_rexii
Gesneriaceae
Antirrhinum_majus
Collinsia_rattanii
Plantaginaceae
Lindernia_brevidens
Torenia_fournieri
Linderniaceae
Buddleja_alternifolia
Scrophulariaceae
Strobilanthes_cusia
Andrographis_paniculata
Acanthaceae
Genlisea_aurea
Utricularia_gibba
Pinguicula_moctezumae
Pinguicula_agnata
Pinguicula_gigantea
Lentibulariaceae
Sesamum_alatum
Sesamum_indicum
Pedaliaceae
Jacaranda_mimosifolia
Handroanthus_impetiginosus
Bignoniaceae
Petrea_volubilis
Lantana_camara
Verbenaceae
Callicarpa_americana
Stenogyne_calaminthoides
Salvia_officinalis
Lamiaceae
Mimulus_guttatus
Phrymaceae
Paulownia_fortunei
Paulowniaceae
Orobanche_coerulescens
Pedicularis_kansuensis
Orobanchaceae
0.93
0.93
2.0
b
Number of anchor pairs
Ks
All anchor pairs
Anchor Ks cluster a (mode 0.17)
Anchor Ks cluster b (mode 1.39)
Divergence with:
(1) Pinguicula agnata (0.02 ← 0.05)
(2) Pinguicula moctezumae (0.07 → 0.07)
(3) Utricularia gibba (0.61 ← 1.06)
(3) Genlisea aurea (0.61 ← 1.14)
(4) Strobilanthes cusia (1.06 → 1.13)
(5) Mimulus guttatus (0.97 → 1.14)
(5) Callicarpa americana (0.72 → 1.12)
(5) Sesamum indicum (0.81 → 1.13)
(6) Torenia fournieri (1.18 ← 1.43)
(6) Buddleja alterniflora (0.87 → 1.19)
(7) Antirrhinum majus (1.05 → 1.27)
(8) Jasminum sambac (1.32 → 1.59)
(9) Coffea canephora (1.63 → 2.04)
c
Pinguicula gigantea
Pinguicula agnata
Pinguicula moctezumae
Utricularia gibba
Genlisea aurea
Strobilanthes cusia
Mimulus guttatus
Callicarpa americana
Sesamum indicum
Torenia fournieri
Buddleja alternifolia
Antirrhinum majus
Jasminum sambac
Coffea canephora
Vitis vinifera
0.25
d
Pinguicula gigantea
Pinguicula agnata
Pinguicula moctezumae
Utricularia gibba
Genlisea aurea
Strobilanthes cusia
Mimulus guttatus
Callicarpa americana
Sesamum indicum
Torenia fournieri
Buddleja alternifolia
Antirrhinum majus
Jasminum sambac
Coffea_canephora
Vitis vinifera

Figure 2: Phylogenomic placement and polyploid history of the Lentibulariaceae. (a) The Lentibulariaceae is highlighted in green. Other families of the Lamiales are highlighted in dark and light shades of grey. Two Rubiaceae species (order Gentianales) were used as the outgroup. Tree generated from 2,322 BUSCO protein sequences. Branch support is local posterior probability. **(b)** Positioning of divergence times in the Lamiales relative to polyploidy events in *Pinguicula gigantea*. Syntenic (anchor) Ks paralog pair distribution for *Pinguicula gigantea* is shown in gray. Lognormal mixture model clustering of the median Ks values of the syntenic paralog pairs indicate the core Lamiales WGD in red, and a *Pinguicula*-specific WGD in blue. Vertical dashed lines labeled a and b denote the modes of these curves. Vertical lines labeled with 1–9 are rate-adjusted modal estimates of one-to-one ortholog Ks distributions between *Pinguicula gigantea* and other species, representing divergence times. **(c)** Phylogenetic tree branch lengths represent Ks distances estimated from ortholog Ks distributions. Dashed branches have lengths that cannot be computed by the software. Manually superimposed labels match with the Ks plot and represent *Pinguicula gigantea* polyploidy events and divergence with other lineages on the tree. **(d)** Same as (c), but with meaningless branch lengths to help see tree topology.

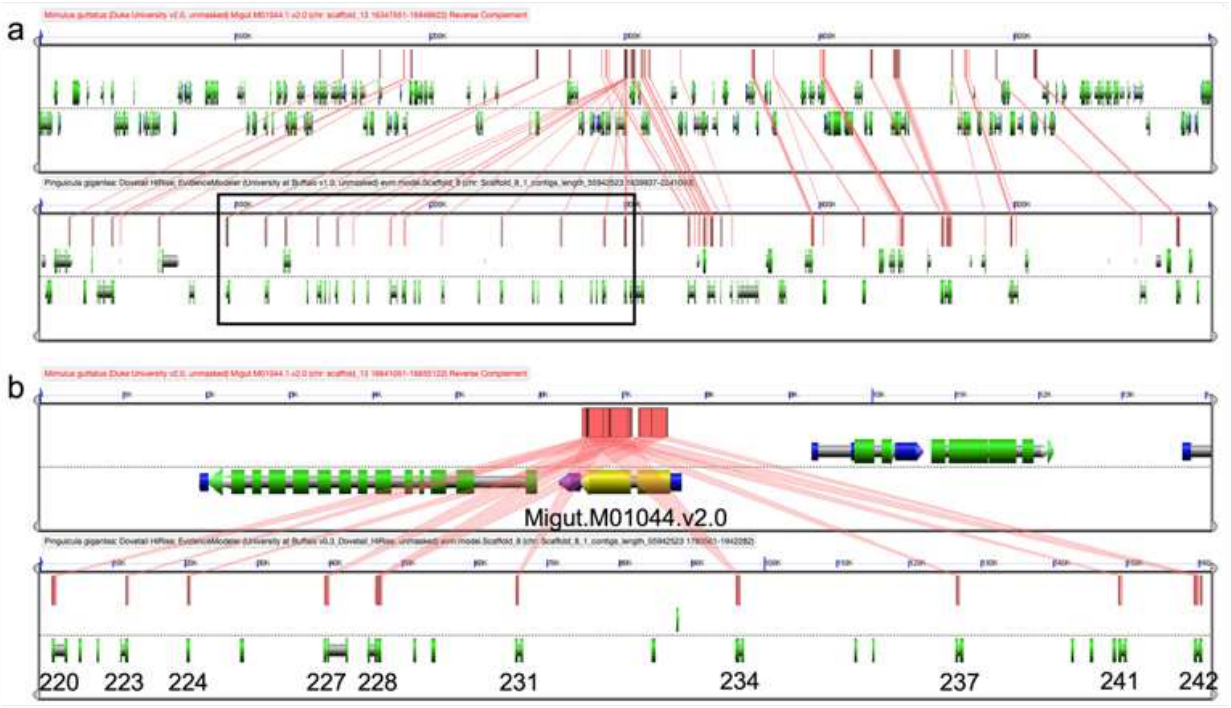


Figure 3: **A highly expanded cysteine protease tandem array in *Pinguicula gigantea* is syntenic to a single homolog in *Mimulus guttatus*. (a)** Syntenic blocks between *Mimulus guttatus* scaffold 13 and *Pinguicula gigantea* chromosome 8. The *P. gigantea* tandem array is marked with a black rectangle. **(b)** Zoom in of the tandem array, shown against the single *M. guttatus* syntenic homolog. Gene models are labeled with the full gene ID in *M. guttatus* (top block), but only with gene number in *P. gigantea* (bottom block). CDS are represented by green rectangles. Colored lines connect homologous gene pairs.

Supplementary Figures

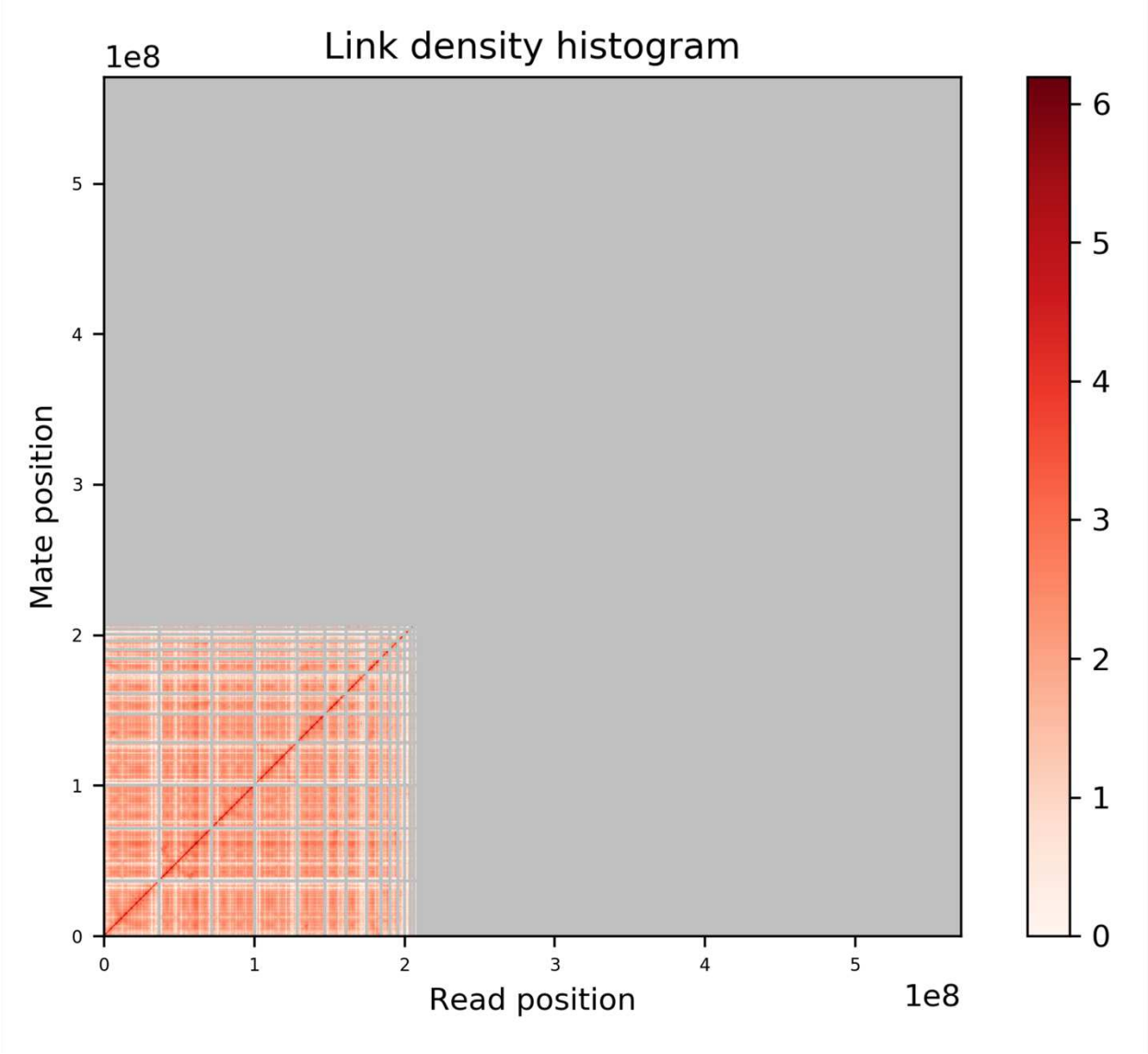


Figure S1: Hi-C contact map for the *Pinguicula grandiflora* assembly. Darker red indicates stronger interaction signals. The left and bottom axes show the genomic position along the assembly.

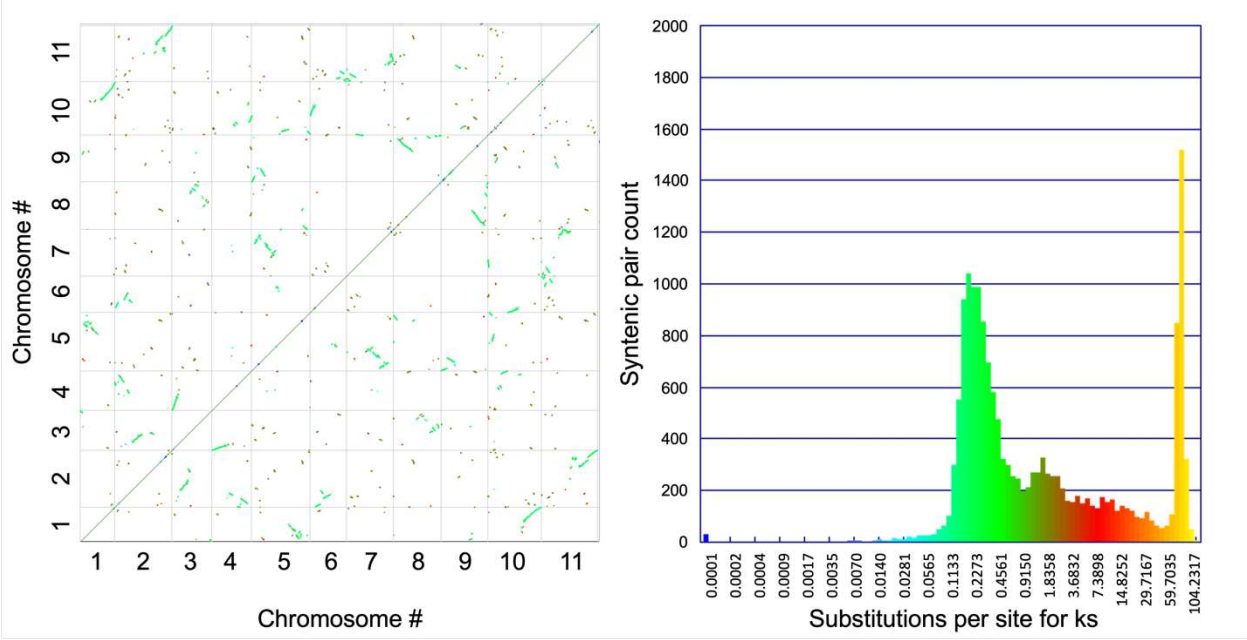


Figure S2: Intra-genomic synteny and synonymous substitution rates for the *Pinguicula gigantea* Hi-C assembly. (Left) Syntenic dot plot where each point represents a pair of paralogous genes; the genome is plotted on both the X and Y axes. (Right) Histogram of synonymous substitution rates (Ks) for these syntenic gene pairs. The color of each point in the dot plot corresponds to its Ks value on the histogram.

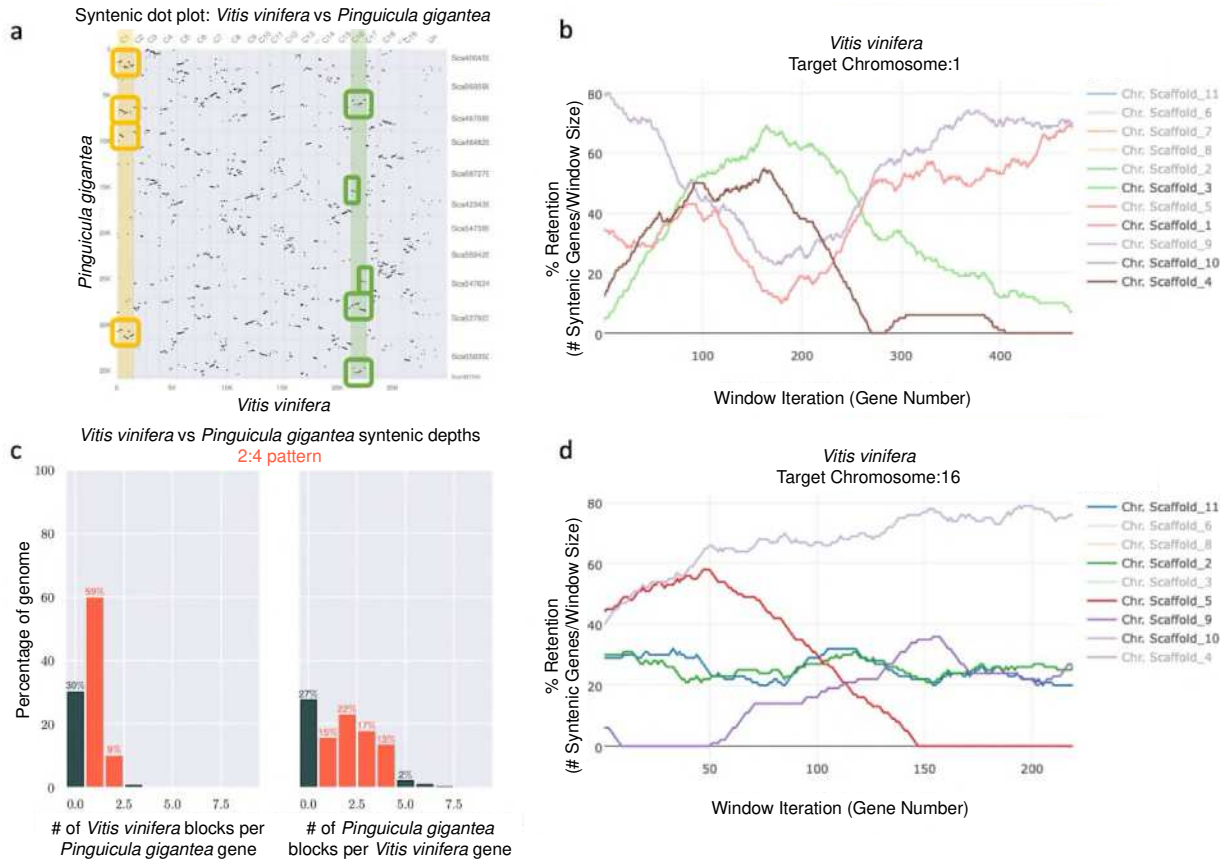


Figure S3: Comparative synteny and fractionation bias between *Pinguicula gigantea* and *Vitis vinifera*. (a) Syntenic dot plot of orthologous gene pairs. *Vitis vinifera* chromosomes 1 and 16 are highlighted in yellow and green, respectively. **(b)** Fractionation bias plot for *V. vinifera* chromosome 1, showing the percentage of retained genes (Y-axis) across the target chromosome (X-axis); lines represent different *P. gigantea* chromosomes. **(c)** Syntenic depth histograms showing the number of genomic blocks per gene for each species. **(d)** Fractionation bias plot against *V. vinifera* chromosome 16.

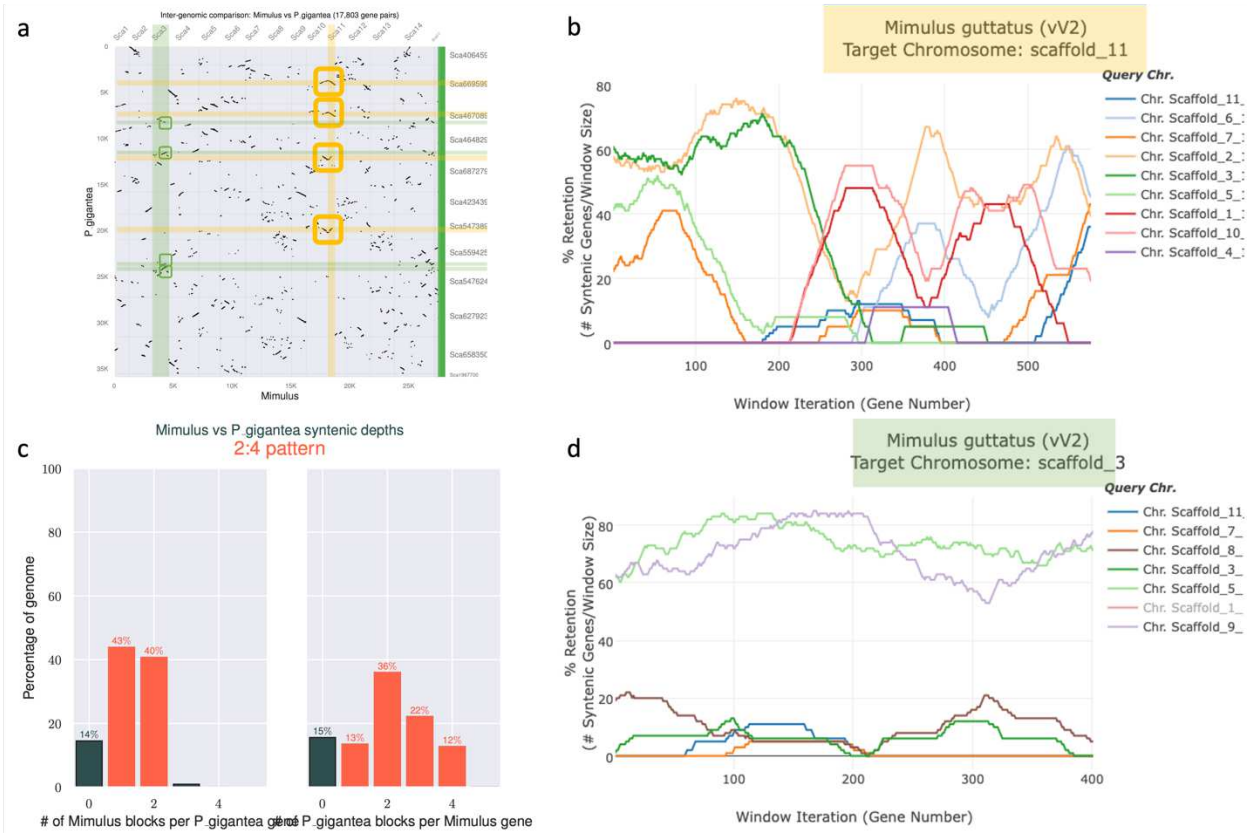


Figure S4: Comparative synteny and fractionation bias between *Pinguicula gigantea* and *Mimulus guttatus*. (a) Syntenic dot plot of orthologous gene pairs. *Mimulus guttatus* chromosomes 11 and 3 are highlighted in yellow and green, respectively. **(b)** Fractionation bias plot against *M. guttatus* chromosome 11. **(c)** Syntenic depth histograms. **(d)** Fractionation bias plot against *M. guttatus* chromosome 3.

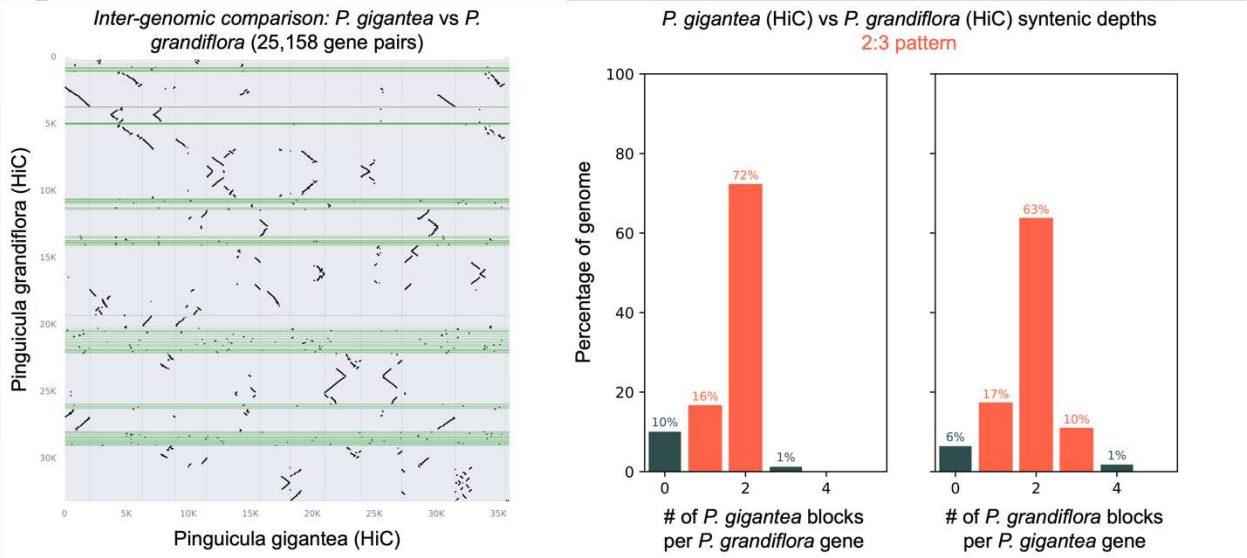


Figure S5: Comparative synteny between the *Pinguicula grandiflora* and *Pinguicula gigantea* Hi-C assemblies. (Left) Syntenic dot plot of orthologous gene pairs. (Right) Syntenic depth histograms.

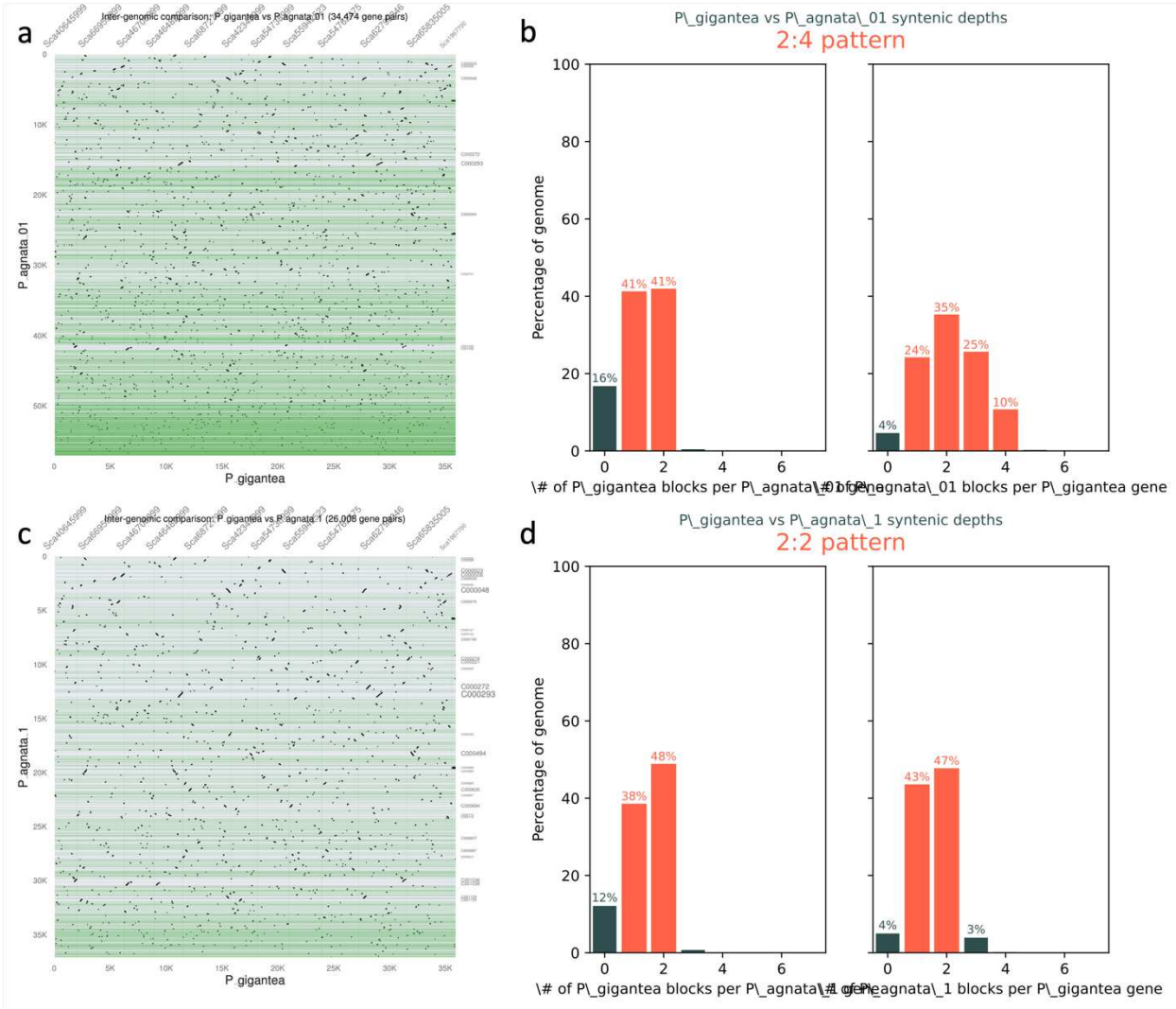


Figure S6: Syntenic comparison between *Pinguicula agnata* and *Pinguicula gigantea* assemblies. (a, b) Analysis of the primary *P. agnata* assembly before purging haplotigs. **(c, d)** Analysis after purging with PurgeHaplotigs. (Left) Syntenic dot plots of orthologous gene pairs (*P. agnata* on X-axis, *P. gigantea* on Y-axis). (Right) Corresponding syntenic depth histograms.

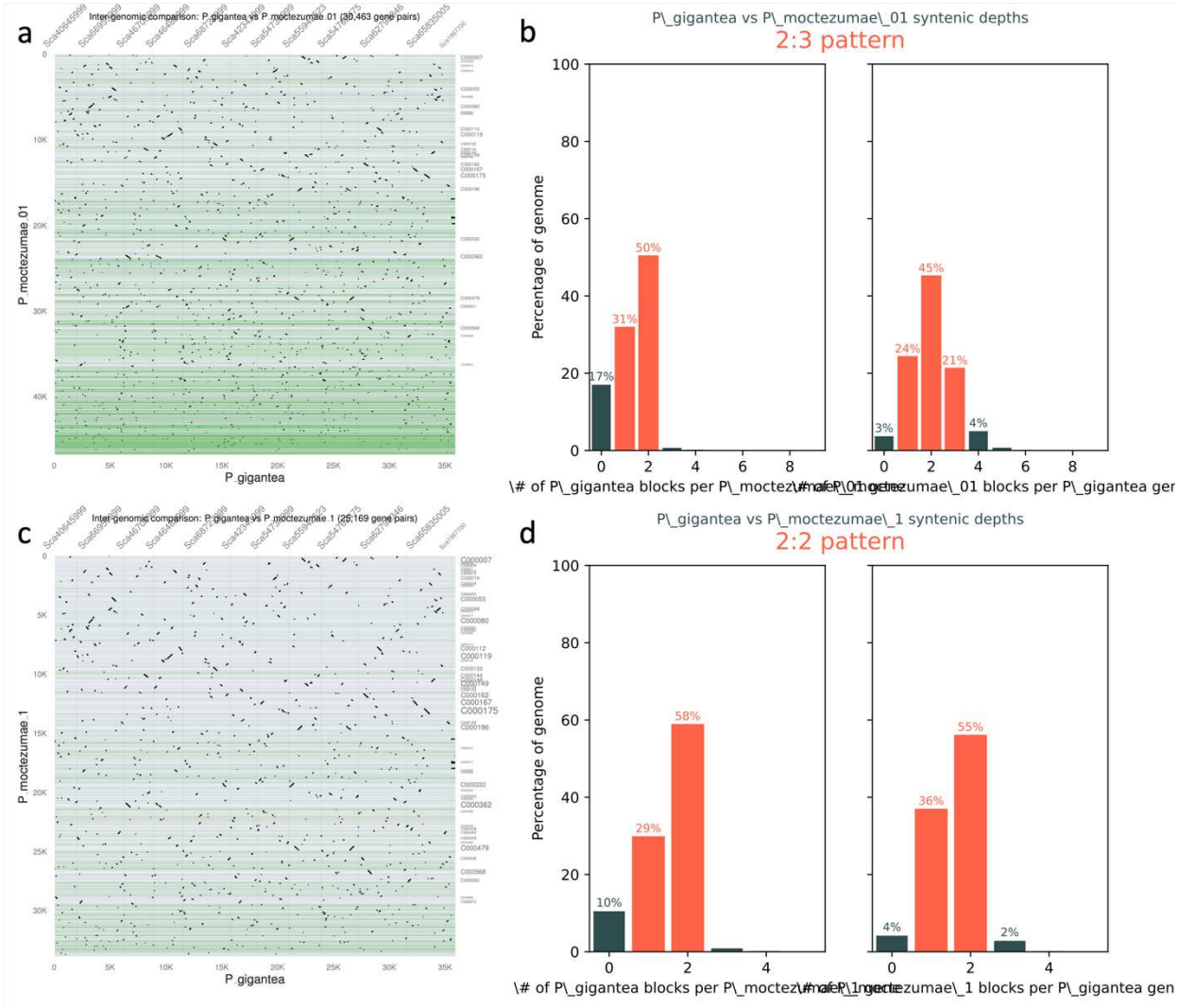


Figure S7: Syntenic comparison between *Pinguicula moctezumae* and *Pinguicula gigantea* assemblies. (a, b) Analysis of the primary *P. moctezumae* assembly before purging haplotigs. **(c, d)** Analysis after purging with PurgeHaplotigs. (Left) Syntenic dot plots of orthologous gene pairs (*P. moctezumae* on X-axis, *P. gigantea* on Y-axis). (Right) Corresponding syntenic depth histograms.

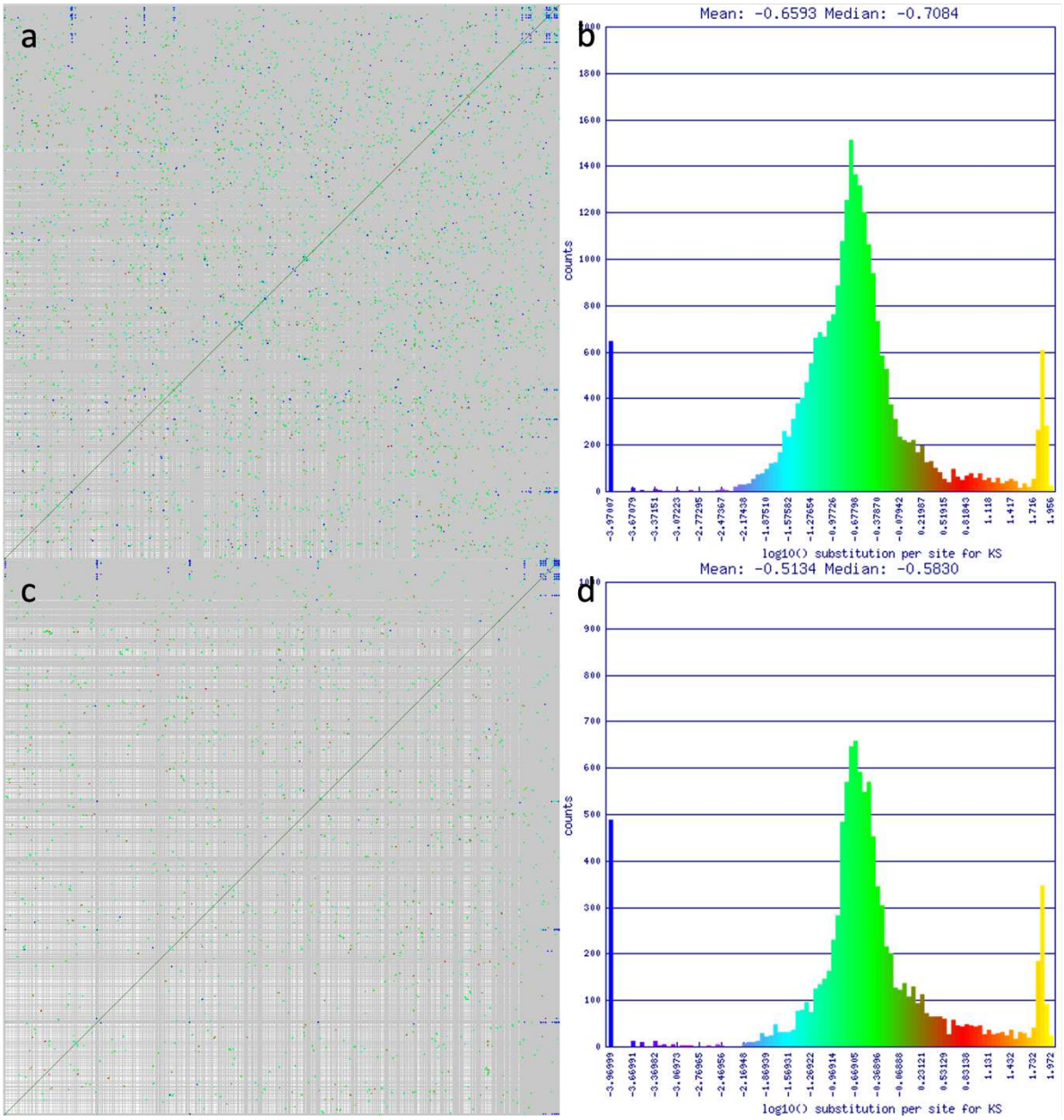


Figure S8: Intra-genomic synteny and Ks distribution for the *Pinguicula agnata* assembly. (a, b) Analysis of the primary assembly before purging. **(c, d)** Analysis after purging with PurgeHaplotigs. (Left) Syntenic dot plots of paralogous gene pairs. (Right) Histograms of synonymous substitution rates (Ks) for these pairs; point colors in the dot plots correspond to Ks values.

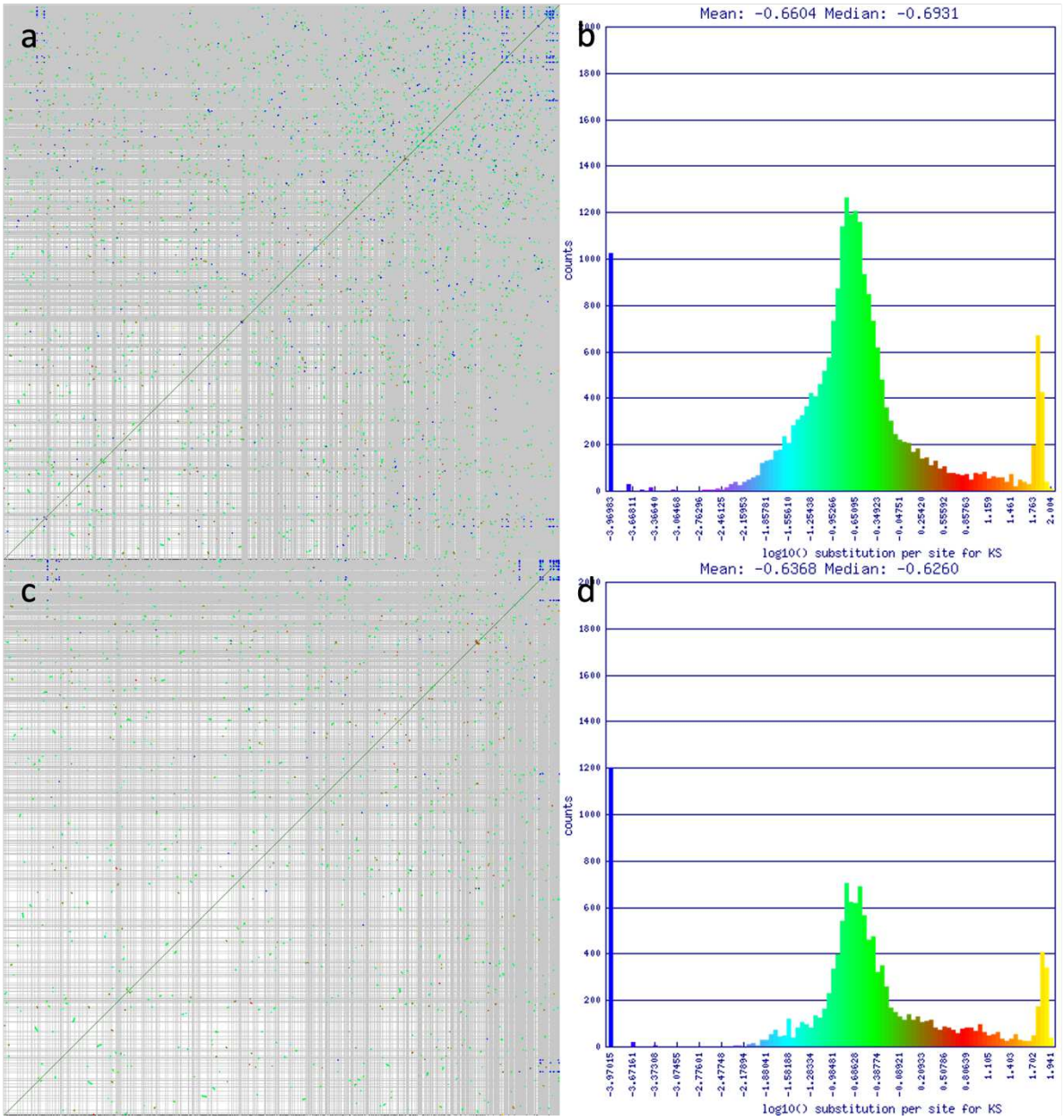


Figure S9: Intra-genomic synteny and Ks distribution for the *Pinguicula moctezumae* assembly. **(a, b)** Analysis of the primary assembly before purging. **(c, d)** Analysis after purging with PurgeHaplotigs. (Left) Syntenic dot plots of paralogous gene pairs. (Right) Histograms of synonymous substitution rates (Ks) for these pairs; point colors in the dot plots correspond to Ks values.

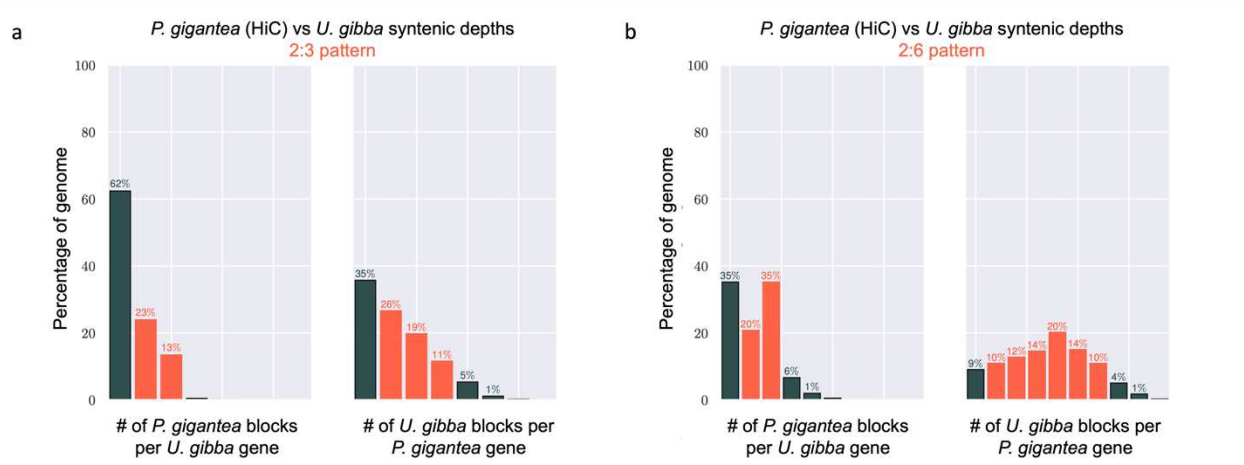


Figure S10: Comparative syntenic depth between *Pinguicula gigantea* and *Utricularia gibba*. Histograms show the number of genomic blocks per gene, inferred by searching for five syntenic genes within **(a)** twenty or **(b)** fifty nearby genes.

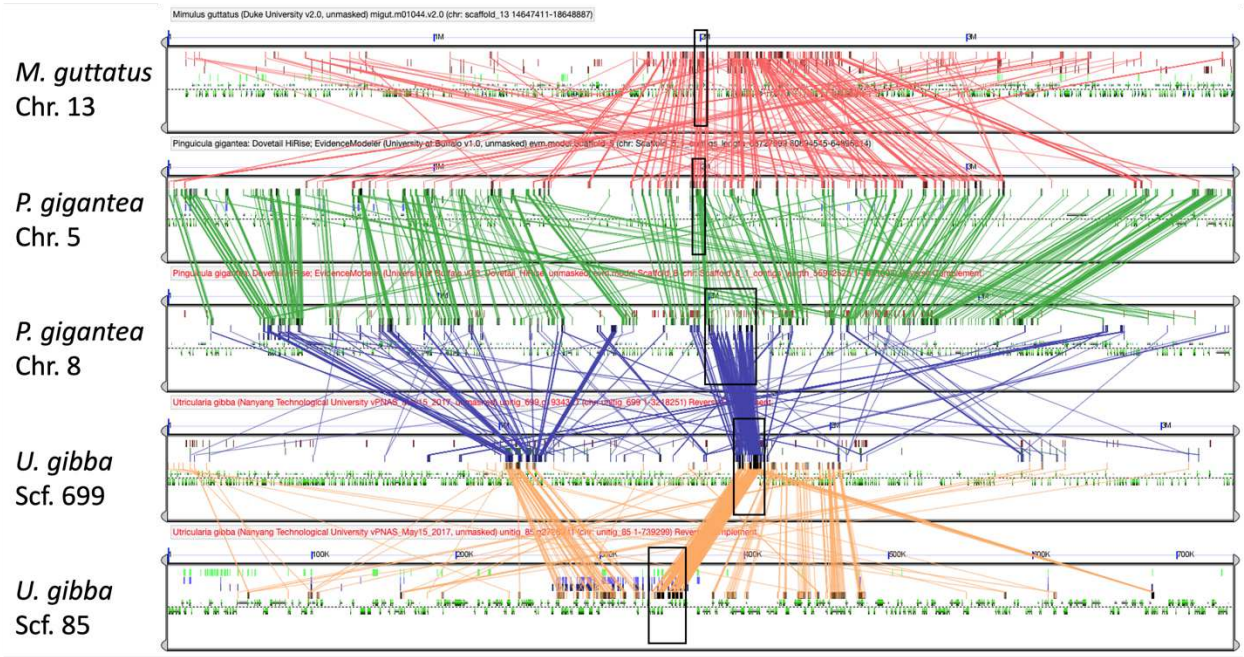


Figure S11: Microsynteny of cysteine protease genomic regions across species. Syntenic blocks are shown for *Mimulus guttatus* scaffold 13, *Pinguicula gigantea* chromosomes 5 and 8 (which are descendants of its lineage-specific WGD), and *U. gibba* scaffolds 699 and 85 (two of the four descendants of *U. gibba*'s two most-recent tetraploidy events; the other two are scaffolds 578 and 8). Tandem arrays are marked by black rectangles.

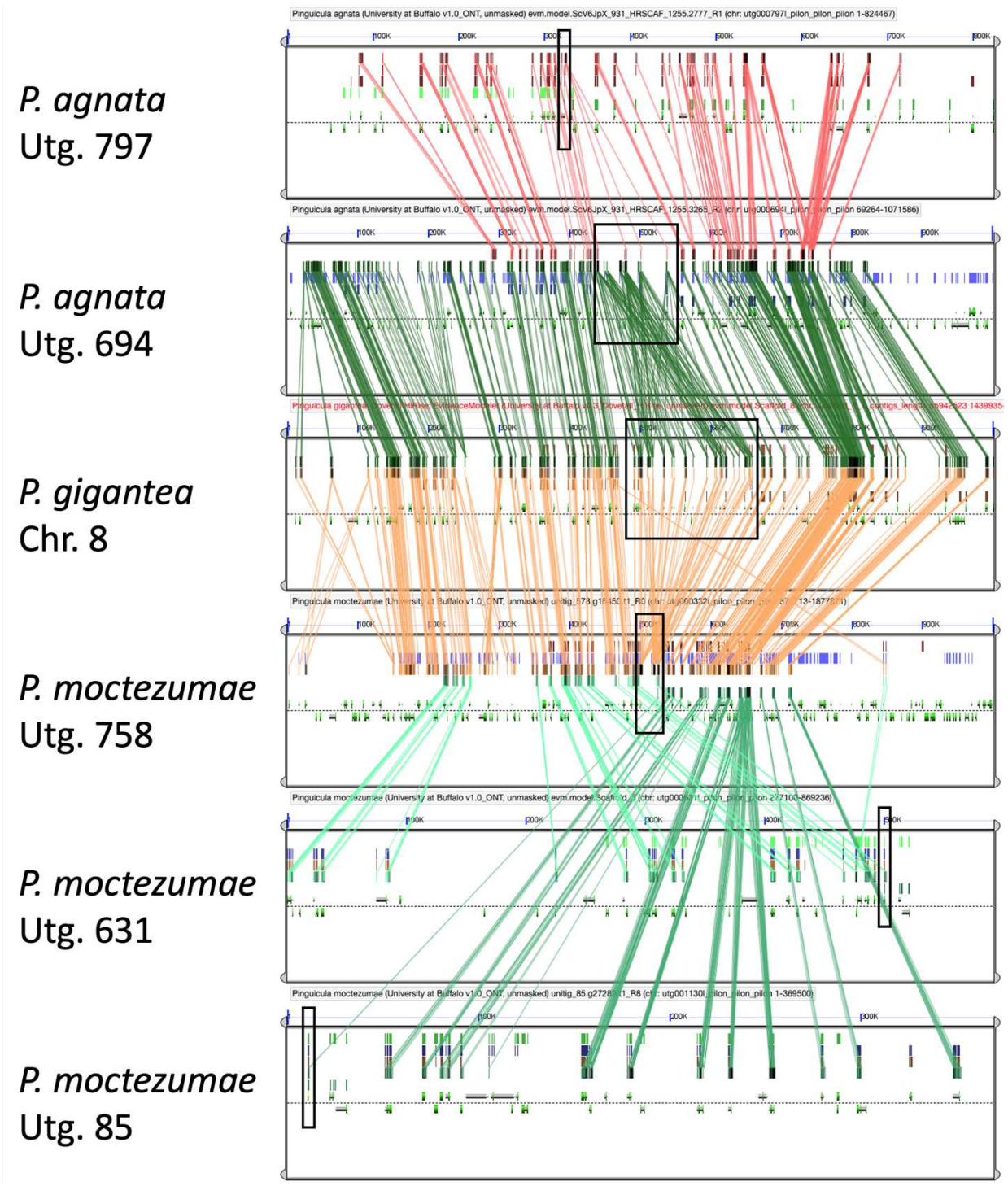
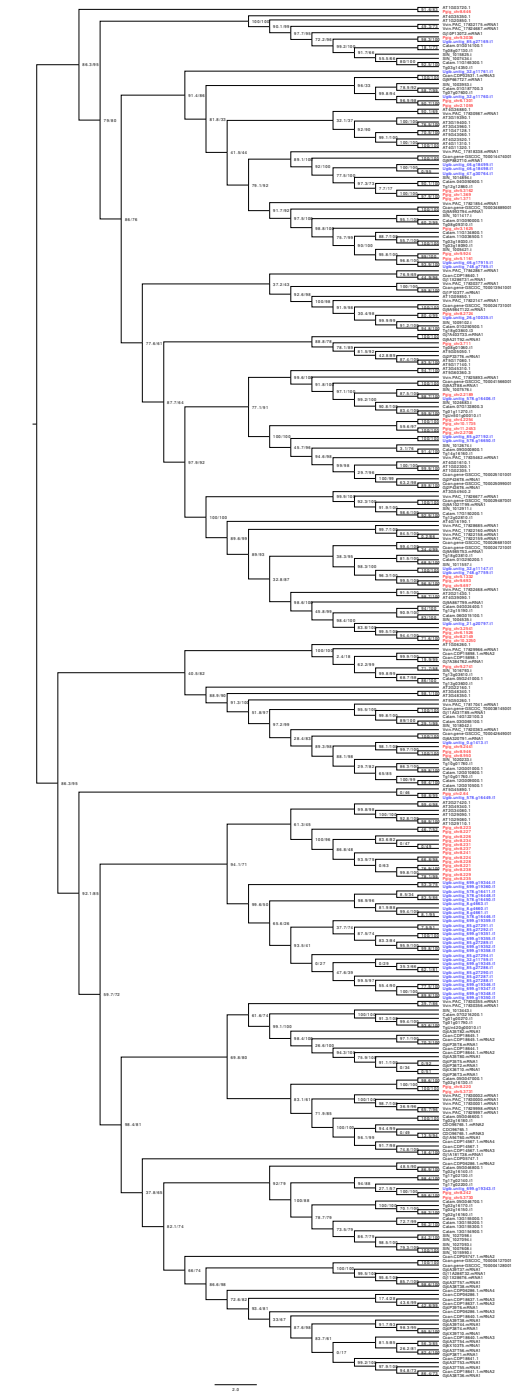
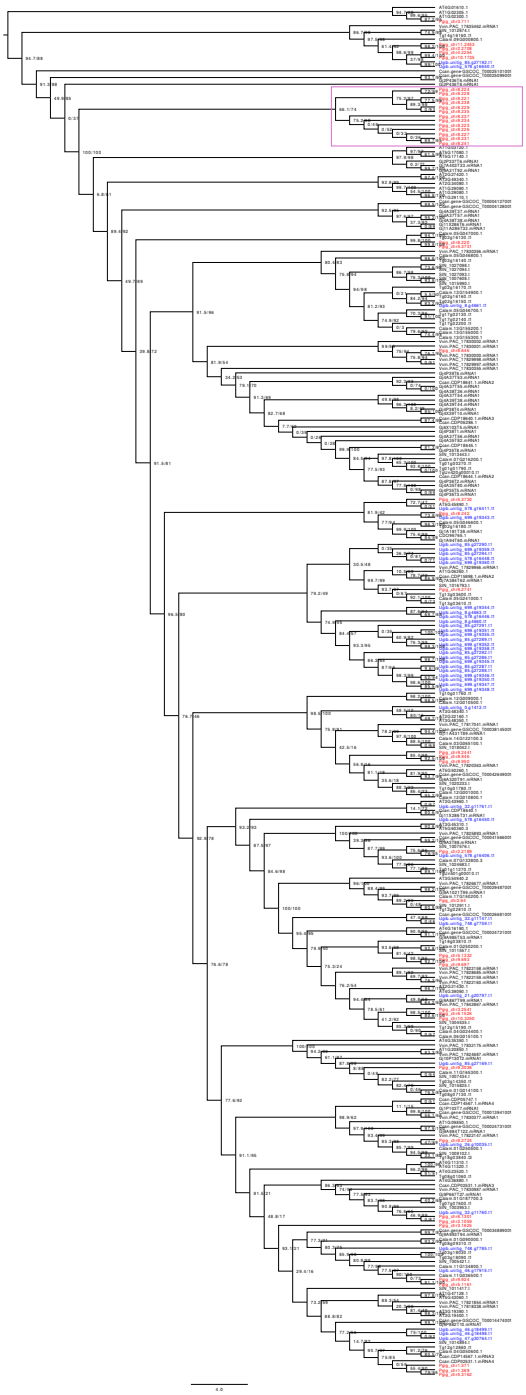


Figure S12: Macrosynteny of cysteine protease genomic regions across *Pinguicula* species. Syntenic blocks are shown for *Pinguicula agnata* unitigs 797 and 694, *Pinguicula gigantea* chromosome 8, and *Pinguicula moctezumae* unitigs 758, 631, and 85. Tandem arrays are marked by black rectangles

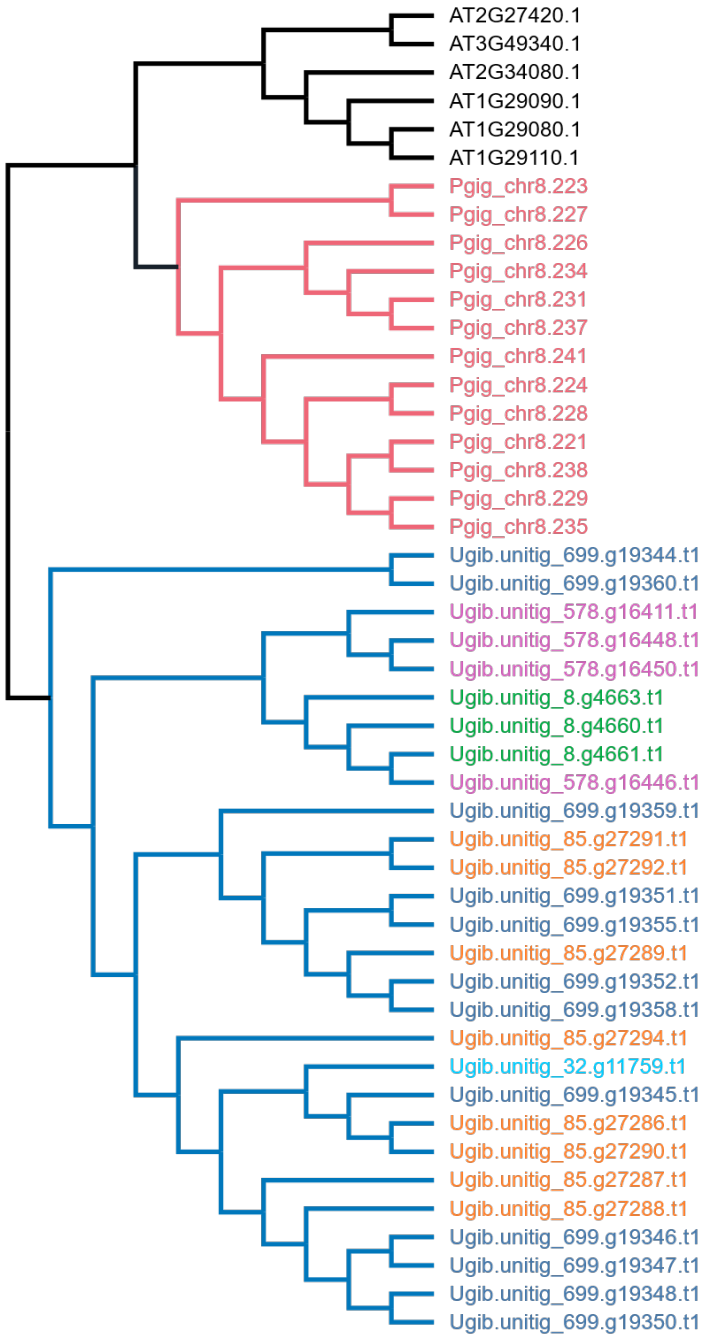
a



b



C


AT2G27420.1
AT3G49340.1
AT2G34080.1
AT1G29090.1
AT1G29080.1
AT1G29110.1
Pgig_chr8.223
Pgig_chr8.227
Pgig_chr8.226
Pgig_chr8.234
Pgig_chr8.231
Pgig_chr8.237
Pgig_chr8.241
Pgig_chr8.224
Pgig_chr8.228
Pgig_chr8.221
Pgig_chr8.238
Pgig_chr8.229
Pgig_chr8.235
Ugib.unitig_699.g19344.t1
Ugib.unitig_699.g19360.t1
Ugib.unitig_578.g16411.t1
Ugib.unitig_578.g16448.t1
Ugib.unitig_578.g16450.t1
Ugib.unitig_8.g4663.t1
Ugib.unitig_8.g4660.t1
Ugib.unitig_8.g4661.t1
Ugib.unitig_578.g16446.t1
Ugib.unitig_699.g19359.t1
Ugib.unitig_85.g27291.t1
Ugib.unitig_85.g27292.t1
Ugib.unitig_699.g19351.t1
Ugib.unitig_699.g19355.t1
Ugib.unitig_85.g27289.t1
Ugib.unitig_699.g19352.t1
Ugib.unitig_699.g19358.t1
Ugib.unitig_85.g27294.t1
Ugib.unitig_32.g11759.t1
Ugib.unitig_699.g19345.t1
Ugib.unitig_85.g27286.t1
Ugib.unitig_85.g27290.t1
Ugib.unitig_85.g27287.t1
Ugib.unitig_85.g27288.t1
Ugib.unitig_699.g19346.t1
Ugib.unitig_699.g19347.t1
Ugib.unitig_699.g19348.t1
Ugib.unitig_699.g19350.t1

d



Figure S13: Phylogeny and synteny between tandem arrays of cysteine protease genomes in *Pinguicula gigantea* and *Utricularia gibba*. Phylogenetic trees of cysteine protease **(a)** CDS, and **(b)** protein sequences show convergent tandem gene expansions in *U. gibba* and *P. gigantea*. Only CDS-based trees were further considered due to potential phylogeny-confounding functional convergence at the inferred amino acid level[17]. Gene locations are indicated by scaffold (unitig; *U. gibba*) or chromosome (*P. gigantea*), and gene numbers follow physical order. Branch support values are based on ultrafast bootstrap replicates. Since the gene copies from the two species do not form mixed clades in the tree, the tandem duplications occurred independently in each species. Genes from *U. gibba* are shown in blue, and *P. gigantea* in red. Purple boxes indicate clusters of three or more tandem duplicates. **(c)** Zoom-in of the boxed subtree in (a). Here, clades from *U. gibba* and *P. gigantea* are shown in blue and red, respectively, since they are taxon-specific, having tandemly duplicated following the *Utricularia-Pinguicula* species split. *U. gibba* arrays occur on four WGD-derived syntenic blocks (scaffolds 699, 85, 578, and 8); *P. gigantea* has a large array on chromosome 8. **(d)** A CoGe GEvo view of *U. gibba* scaffold 8 and *P. gigantea* chromosome 8; these are syntenic with *U. gibba* scaffolds 85, 699 and 578. Genes are colored blue, orange, purple and green, respectively, to signify location on those scaffolds. A single cysteine protease gene on *U. gibba* scaffold 32 (cyan) may be a transposed duplicate. A lineage of *Arabidopsis* cysteine protease genes (with prefix AT) is sister to the *P. gigantea* array. **(d)** A CoGe GEvo view of *U. gibba* scaffold 8 and *P. gigantea* chromosome 8; these are syntenic with *U. gibba* scaffolds 85, 699 and 578 (the first two scaffolds are shown in Figure S11). The tandem arrays of cysteine protease genes are highlighted in green shaded boxes.

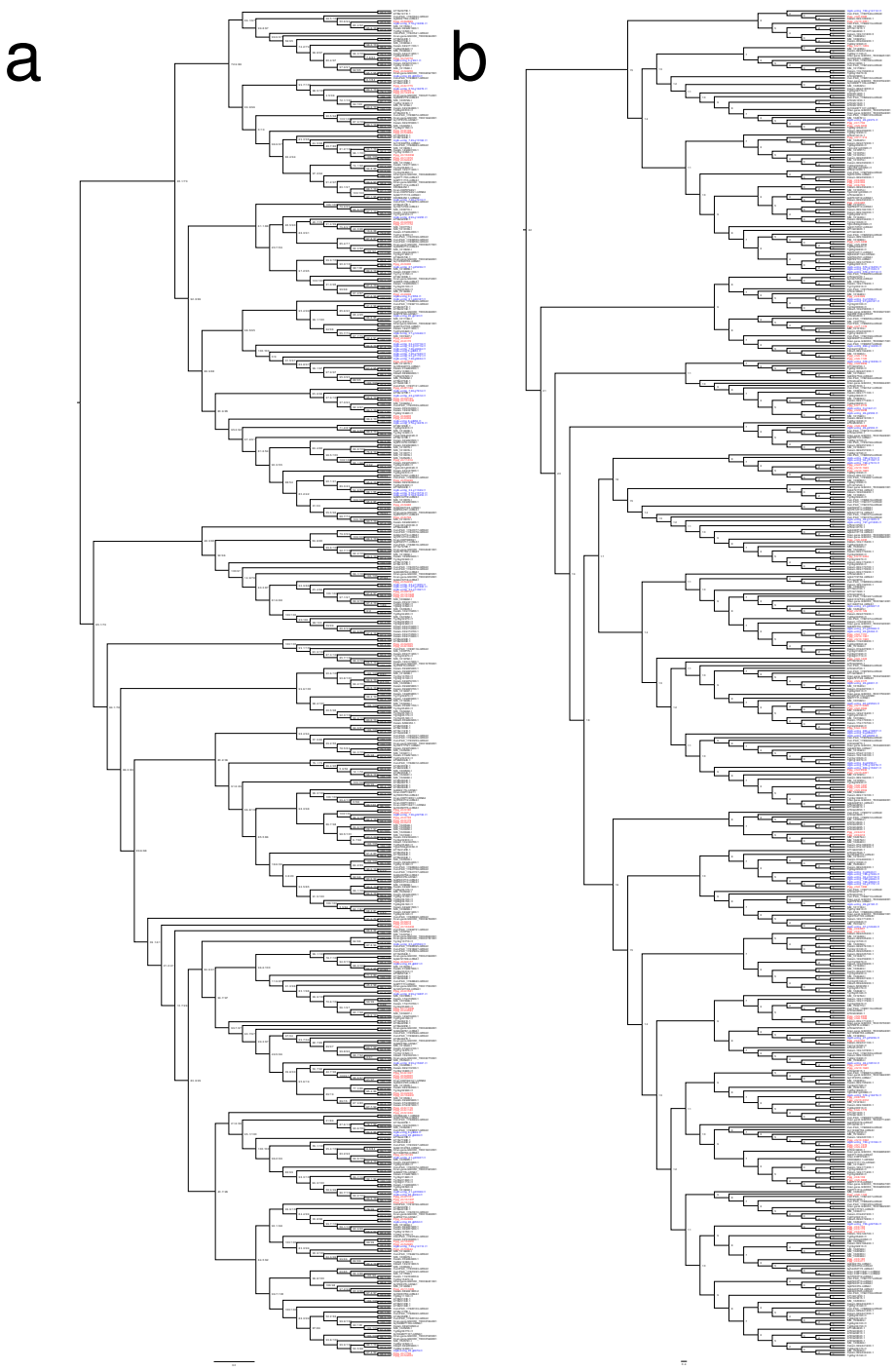


Figure S14: Phylogenetic trees of aspartic protease (a) CDS and (b) protein sequences. Branch support values are based on ultrafast bootstrap replicates. Sequences from *Utricularia gibba* are shown in blue, and *Pinguicula gigantea* in red. Purple boxes indicate clusters of three or more tandem duplicates.

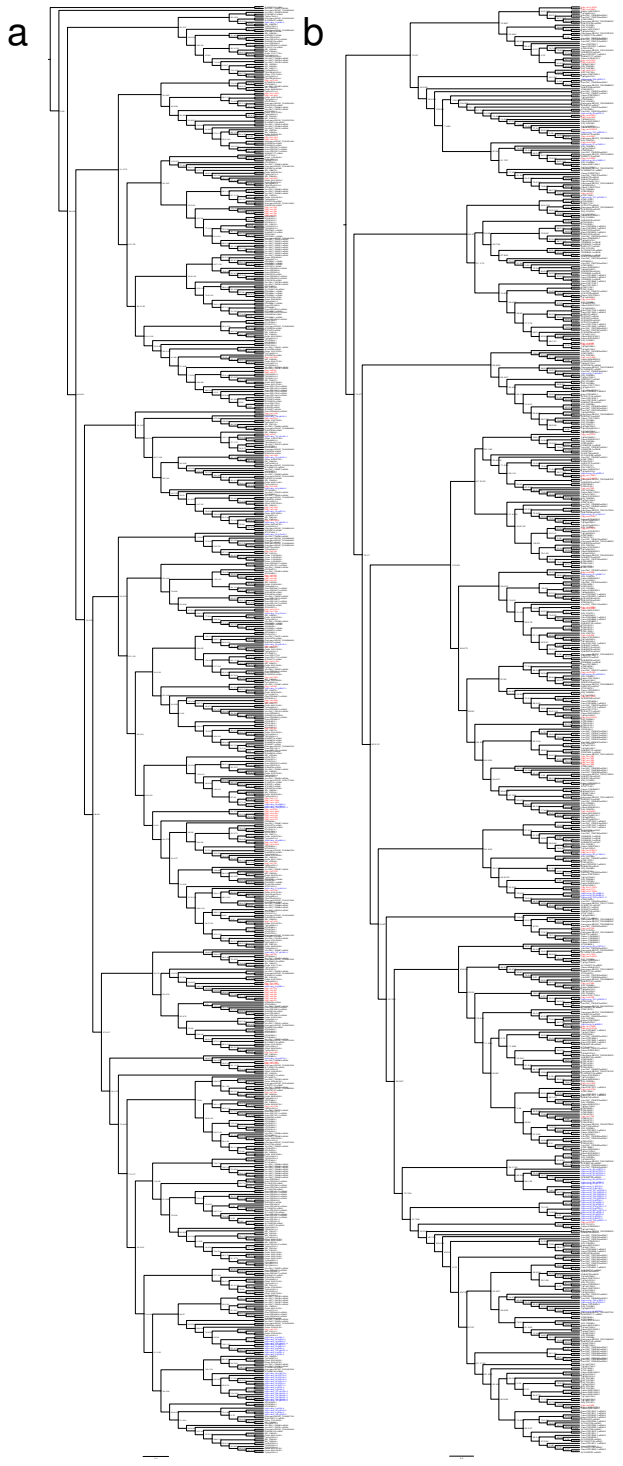


Figure S15: Phylogenetic trees of peroxidase 10-like (a) CDS and (b) protein sequences. Branch support values are based on ultrafast bootstrap replicates. Sequences from *Utricularia gibba* are shown in blue, and *Pinguicula gigantea* in red. Purple boxes indicate clusters of three or more tandem duplicates.

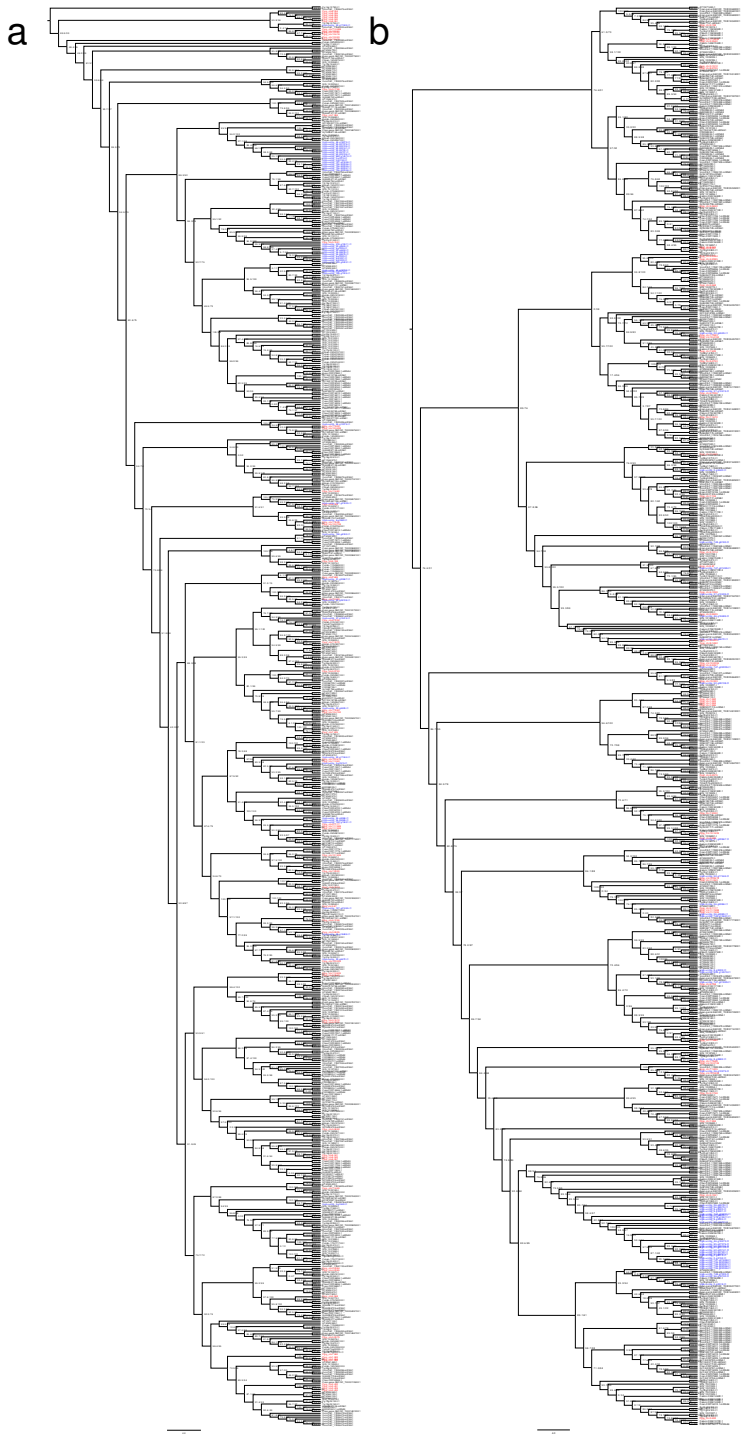


Figure S16: Phylogenetic trees of putative peroxidase (a) CDS and (b) protein sequences. Branch support values are based on ultrafast bootstrap replicates. Sequences from *Utricularia gibba* are shown in blue, and *Pinguicula gigantea* in red. Purple boxes indicate clusters of three or more tandem duplicates.

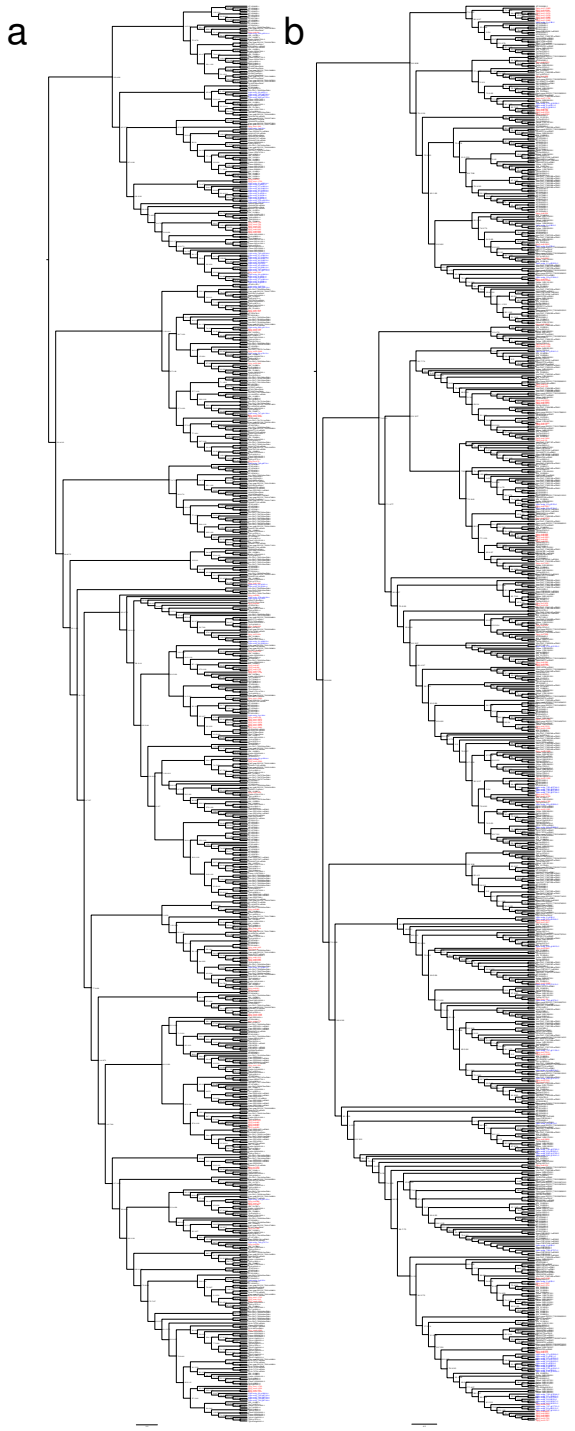


Figure S17: Phylogenetic trees of predicted esterase/lipase (a) CDS and (b) protein sequences. Branch support values are based on ultrafast bootstrap replicates. Sequences from *Utricularia gibba* are shown in blue, and *Pinguicula gigantea* in red. Purple boxes indicate clusters of three or more tandem duplicates.

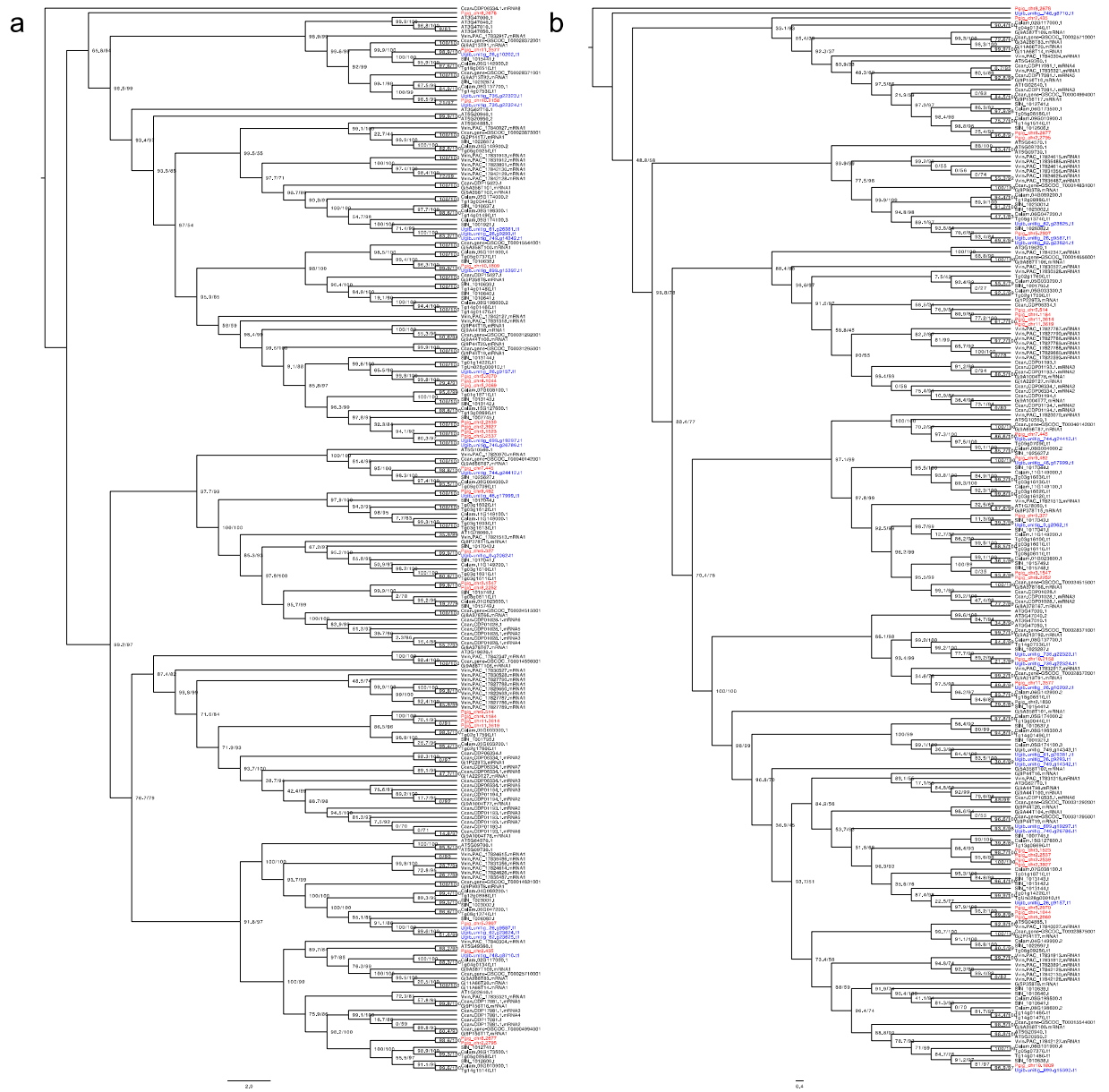


Figure S18: Phylogenetic trees of putative beta-xylosidase (a) CDS and (b) protein sequences. Branch support values are based on ultrafast bootstrap replicates. Sequences from *Utricularia gibba* are shown in blue, and *Pinguicula gigantea* in red. Purple boxes indicate clusters of three or more tandem duplicates.

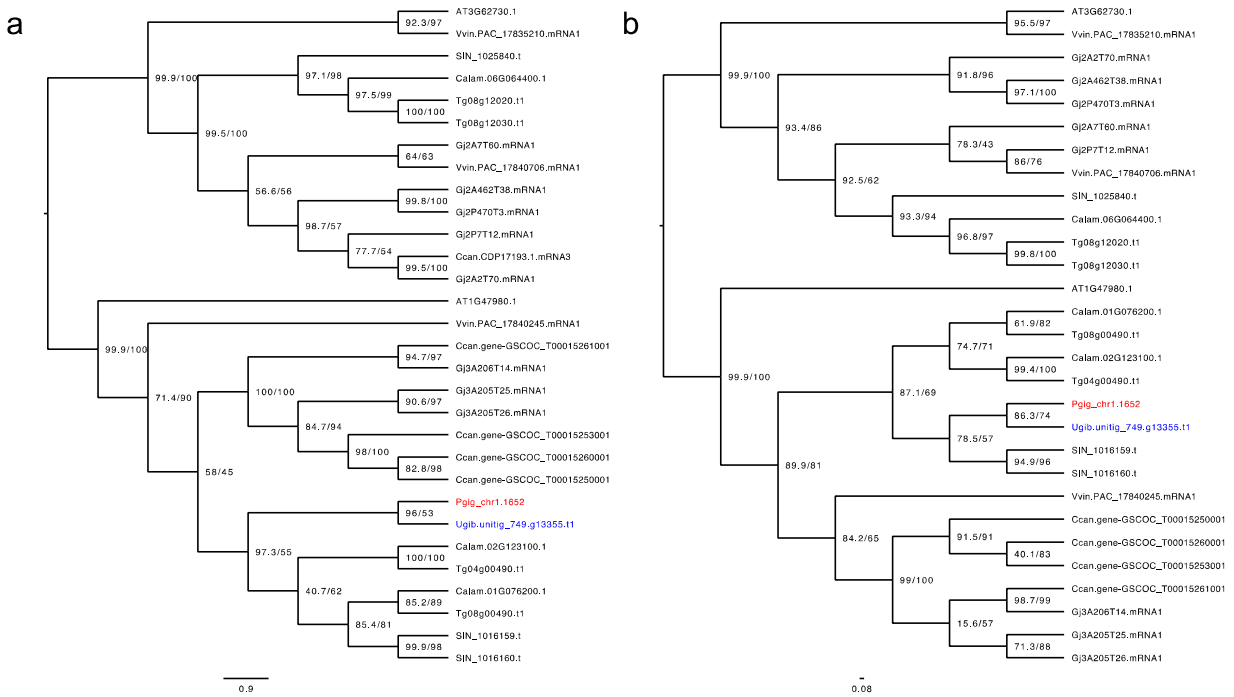


Figure S19: Phylogenetic trees of desiccation-related protein (a) CDS and (b) protein sequences. Branch support values are based on ultrafast bootstrap replicates. Sequences from *Utricularia gibba* are shown in blue, and *Pinguicula gigantea* in red. Purple boxes indicate clusters of three or more tandem duplicates.

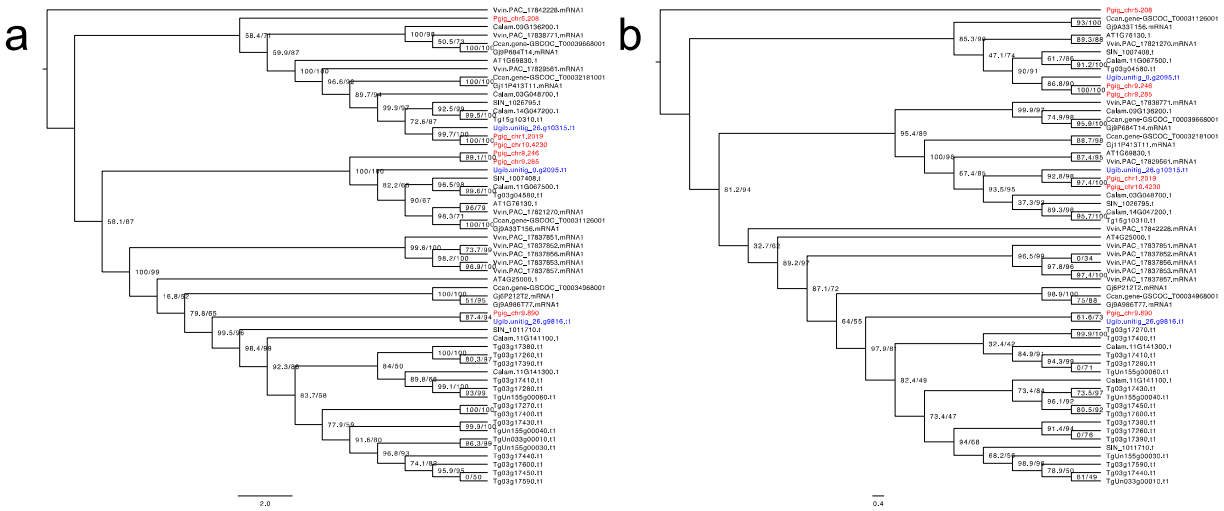


Figure S20: Phylogenetic trees of alpha-amylase (a) CDS and (b) protein sequences. Branch support values are based on ultrafast bootstrap replicates. Sequences from *Utricularia gibba* are shown in blue, and *Pinguicula gigantea* in red. Purple boxes indicate clusters of three or more tandem duplicates.

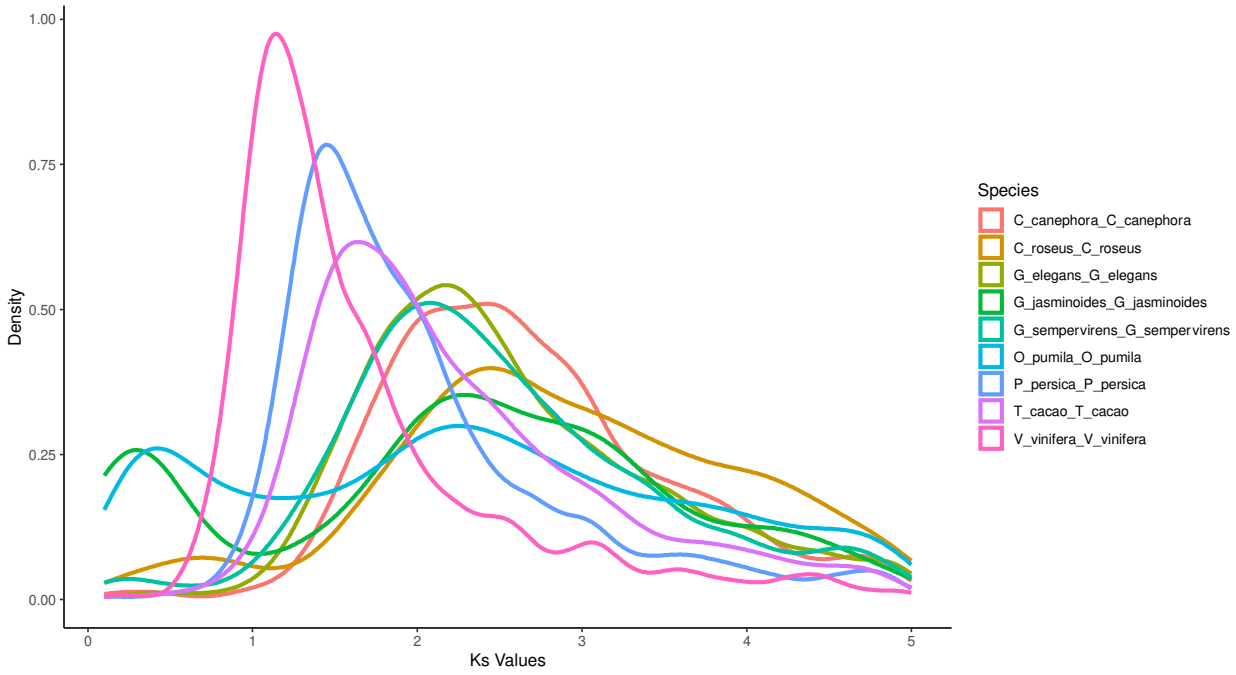


Figure S21: Density plots of synonymous substitution rate (Ks) peaks used to position the core eudicot *gamma* hexaploidy event. The color key for species is shown on the right. Analyzed species include: *Coffea canephora*, *Catharanthus roseus*, *Gelsemium elegans*, *Gardenia jasminoides*, *Gelsemium sempervirens*, *Ophiorrhiza pumila*, *Prunus persica*, *Theobroma cacao*, and *Vitis vinifera*.

Tables – See attached XLS

References Cited

1. Adamec, L., Matušíková, I. & Pavlovič, A. Recent ecophysiological, biochemical and evolutional insights into plant carnivory. *Annals of Botany* **128**, 241-259 (2021).
2. Fleck, S.J. & Jobson, R.W. Molecular Phylogenomics Reveals the Deep Evolutionary History of Carnivory across Land Plants. *Plants* **12**, 3356 (2023).
3. Fleischmann, A., Schlauer, J., Smith, S. & Givnish, T. Evolution of carnivory in angiosperms. in *Carnivorous Plants: Physiology, ecology, and evolution* (eds. AM, E. & L, A.) 22-41 (Oxford University Press, 2018).
4. Hedrich, R. & Fukushima, K. On the origin of carnivory: molecular physiology and evolution of plants on an animal diet. *Annual review of plant biology* **72**, 133-153 (2021).
5. Ellison, A.M. & Adamec, L. *Carnivorous Plants: physiology, ecology, and evolution*, (Oxford University Press, 2018).
6. Płachno, B. & Muravnik, L.E. Functional anatomy of carnivorous traps. (2018).
7. Leushkin, E.V. *et al.* The miniature genome of a carnivorous plant Genlisea aurea contains a low number of genes and short non-coding sequences. *BMC genomics* **14**, 1-11 (2013).
8. Ibarra-Laclette, E. *et al.* Architecture and evolution of a minute plant genome. *Nature* **498**, 94-98 (2013).
9. Lan, T. *et al.* Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. *Proceedings of the National Academy of Sciences* **114**, E4435-E4441 (2017).
10. Silva, S.R. *et al.* The terrestrial carnivorous plant Utricularia reniformis sheds light on environmental and life-form genome plasticity. *International journal of molecular sciences* **21**, 3 (2019).
11. Procko, C., Chory, J. & Pirro, S. The Genome Sequences of 17 Species of Carnivorous Plants. *Biodiversity genomes* (2023).
12. Refulio-Rodriguez, N.F. & Olmstead, R.G. Phylogeny of lamiidae. *American Journal of Botany* **101**, 287-299 (2014).
13. Initiative, O.T.P.T. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679-685 (2019).
14. Liu, B. *et al.* Phylogenetic relationships of Cyrtandromoea and Wightia revisited: A new tribe in Phrymaceae and a new family in Lamiales. *Journal of Systematics and Evolution* **58**, 1-17 (2020).
15. Xu, W.Q. *et al.* Comparative genomics of figworts (Scrophularia, Scrophulariaceae), with implications for the evolution of Scrophularia and Lamiales. *Journal of systematics and evolution* **57**, 55-65 (2019).
16. Saul, F. *et al.* Subgenome dominance shapes novel gene evolution in the decaploid pitcher plant Nepenthes gracilis. *bioRxiv*, 2023.06. 14.544965 (2023).
17. Fukushima, K. *et al.* Genome of the pitcher plant Cephalotus reveals genetic changes associated with carnivory. *Nature Ecology & Evolution* **1**, 0059 (2017).

18. Kocáb, O. *et al.* Jasmonate-independent regulation of digestive enzyme activity in the carnivorous butterwort Pinguicula× Tina. *Journal of Experimental Botany* **71**, 3749-3758 (2020).
19. Veleba, A. *et al.* Genome size and genomic GC content evolution in the miniature genome-sized family Lentibulariaceae. *New Phytologist* **203**, 22-28 (2014).
20. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
21. Zhang, C. *et al.* Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Molecular biology and evolution* **37**, 3188-3210 (2020).
22. Casper, S.J. & Stimper, R. Chromosome numbers in Pinguicula (Lentibulariaceae): survey, atlas, and taxonomic conclusions. *Plant Systematics and Evolution* **277**, 21-60 (2009).
23. De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666-2669 (2018).
24. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom, 2010).
25. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103-2110 (2016).
26. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
27. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* **27**, 737-746 (2017).
28. Walker, B.J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**, e112963 (2014).
29. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
30. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
31. Roach, M.J., Schmidt, S.A. & Borneman, A.R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC bioinformatics* **19**, 1-10 (2018).
32. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* **326**, 289-293 (2009).
33. Putnam, N.H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome research* **26**, 342-350 (2016).
34. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome biology* **23**, 1-19 (2022).
35. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075 (2013).

36. 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).
37. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, 1-22 (2008).
38. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
39. Smit, A. & Hubley, R. RepeatModeler Open-1.0. (2008-2015).
40. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013-2015).
41. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics* **9**, 1-14 (2008).
42. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant physiology* **176**, 1410-1422 (2018).
43. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644-652 (2011).
44. 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).
45. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290-295 (2015).
46. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
47. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nature methods* **7**, 909-912 (2010).
48. Gilbert, D. Gene-omes built from mRNA-seq not genome DNA. (2013).
49. Haas, B.J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**, 5654-5666 (2003).
50. Wang, M. & Kong, L. pblat: a multithread blat algorithm speeding up aligning sequences to genomes. *BMC bioinformatics* **20**, 1-4 (2019).
51. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875 (2005).
52. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. in *Gene prediction: Methods and protocols*, Vol. 1962 (ed. Kollmar, M.) 161-177 (Springer Science+Business Media, LLC, part of Springer Nature, 233 Spring Street, New York, NY 10013, U.S.A., 2019).
53. Keilwagen, J. *et al.* Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* **44**, e89 (2016).
54. Hellsten, U. *et al.* Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. *Proceedings of the National Academy of Sciences* **110**, 19478-19482 (2013).

55. Zhao, D. *et al.* A chromosomal-scale genome assembly of Tectona grandis reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. *Gigascience* **8**, giz005 (2019).
56. Wang, L. *et al.* Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome biology* **15**, 1-13 (2014).
57. Hamilton, J.P. *et al.* Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, Callicarpa americana. *GigaScience* **9**, giaa093 (2020).
58. Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR genomics and bioinformatics* **3**, lqaa108 (2021).
59. TAIR: The Arabidopsis Information Resource (TAIR). (www.arabidopsis.org, 2019).
60. Nakamura, T., Yamada, K.D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490-2492 (2018).
61. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).
62. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal* **53**, 661-673 (2008).
63. Haug-Baltzell, A., Stephens, S.A., Davey, S., Scheidegger, C.E. & Lyons, E. SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics* **33**, 2197-2198 (2017).
64. Joyce, B.L. *et al.* FractBias: a graphical tool for assessing fractionation bias following polyploidy. *Bioinformatics* **33**, 552-554 (2017).
65. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of open source software* **4**, 1686 (2019).
66. Hadley, W. *Ggplot2: Elegrant graphics for data analysis*, (Springer, 2016).
67. Neuwirth, E. & Neuwirth, M.E. Package 'RColorBrewer'. *ColorBrewer Palettes* (2014).
68. 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).
69. Aphalo, P. ggpmisc: Miscellaneous Extensions to "ggplot2"(R package version 0.3. 6). (2020).
70. Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
71. Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A. & Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32**, 268-274 (2015).
72. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., von Haeseler, A. & Jermiin, L.S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* **14**, 587 (2017).
73. Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic acids research* **51**, D418-D427 (2023).

74. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639-1645 (2009).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Pinguiculasupplementarytablessubmit..xlsx