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.

Genome architecture partitions developmental stability and ecological specialization in the carnivorous pitcher plant Sarracenia purpurea

Victor Albert

vaalbert@buffalo.edu

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

Jonathan Kirshner

University at Buffalo

Christopher Page

University at Buffalo

Nicholas Pratt

University at Buffalo

Johanna Merkel

University at Buffalo

Matthias Freund

Julius-von-Sachs-Institute for Biosciences,

Kenji Fukushima

National Institute of Genetics

Charlotte Lindqvist

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

Michaela Richter

University at Buffalo https://orcid.org/0000-0001-7537-4902

Article

Keywords:

Posted Date: March 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8882670/v1

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

Additional Declarations: There is **NO** Competing Interest.

Genome architecture partitions developmental stability and ecological specialization in the carnivorous pitcher plant *Sarracenia purpurea*

Victor A. Albert[1,2,*], Jonathan Kirshner[1,#], Christopher Page[1,#], Nicholas Pratt[1], Johanna Merkel[1], Michaela Richter[1], Matthias Freund[3], Kenji Fukushima[4], Charlotte Lindqvist[1]

[1]Department of Biological Sciences, University at Buffalo, Buffalo, New York, United States 14260

[2]Institute of Life and Environmental Sciences, Sæmundargata 2, 101 Reykjavík, Iceland

[3]Julius-von-Sachs-Institute for Biosciences, Department of Molecular Plant Physiology and Biophysics, University of Wuerzburg, Julius-von-Sachs-Platz 2, 97082 Wuerzburg, Germany

[4]Center for Frontier Research, National Institute of Genetics, Mishima, 411- 4540, Japan

[*]Correspondence: vicsgenomics@gmail.com

[#]Equal contributions

Abstract

The repeated evolution of complex traits raises the question of how genomes generate ecological novelty while preserving developmental stability. Carnivorous pitcher plants modify the leaf into a prey-capturing organ, yet the genomic basis of their convergent evolution across flowering plants remains unresolved. We present a chromosome-scale genome assembly for the purple pitcher plant, *Sarracenia purpurea*, and analyze it alongside eight angiosperms spanning Ericales and independent carnivorous lineages. The genome retains extensive syntenic blocks from ancient polyploidy, organized into a mosaic of dominant and recessive subgenomic segments. Dominant regions preferentially retain dosage-sensitive regulators including AGO1, BRX, GATA11, ETC1, and RCD1, defining a conserved developmental scaffold associated with leaf morphogenesis. In contrast, tandem duplications concentrate in structurally dynamic regions and are enriched for detoxification, redox buffering, microbial-interaction, and cell-wall processes required to maintain digestive chemistry and trap physiology. Comparative analyses across four carnivorous taxa reveal convergent recruitment of oxidative, transport, and microbial-interaction pathways, with transcriptomic data confirming their activation in pitcher tissue. Pitcher evolution thus reflects genomic partitioning between conserved developmental control and rapidly evolving ecological effectors.

Main

Carnivorous plants have long attracted scientific attention because they display remarkable evolutionary innovations in leaf architecture, nutrient acquisition, and ecological strategy[1,2]. Among them, the North American pitcher plant *Sarracenia purpurea* is characterized by fluid-filled traps that collect, retain, and digest arthropod prey[3] (**Figure 1A-C**). These traps support a structured digestive ecosystem in which microbes, small invertebrates such as larval dipterans, and decomposing prey interact within a trophic network[4-12]. Nutrient release depends largely on microbial metabolism, with comparatively limited direct enzymatic contribution from the plant itself, especially relative to the Southeast Asian pitcher plant *Nepenthes*, which produce enzyme-rich secretions[13-15]. Although pitcher morphology is a classic example of convergent evolution among *Sarracenia*, *Cephalotus* (the Australian pitcher plant), and *Nepenthes*[16], the genomic mechanisms by which these lineages independently reprogram leaf development and implement distinct digestive strategies remain unclear.

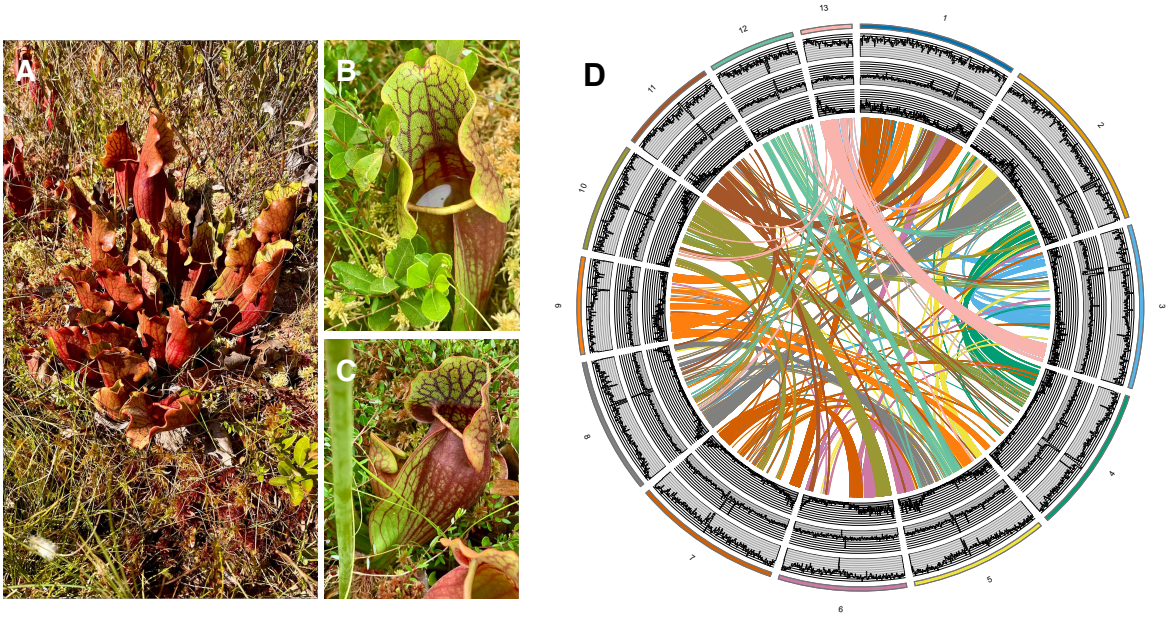


Figure 1. Plant habit and chromosome-wide genomic feature profiles in *Sarracenia purpurea*. (**A**) Whole plant forming a clump of pitchers, (**B**) a single pitcher showing the lid, peristome around the trap opening, and fluid-filled pitcher tube, and (**C**) a side view highlighting the conspicuous wing. (**D**) Genomic feature profiles in the *Sarracenia* genome assembly. Tracks across the 13 pseudochromosomes show – from the inside ring outward – gene density, Gypsy transposable element (TE) abundance, and Copia TE distribution. At the center, self-syntenic blocks inferred from duplicated anchor pairs (with log10(Ks) < 1; Ks = synonymous substitution rate) are shown as color-coded chords. Photos by V.A.A. – (**A**), Allenburg Bog Audubon Nature Reserve, New York; (**B**) and (**C**), Mendon Ponds Park, New York.

Sarracenia purpurea offers an opportunity to examine how a leaf is reconfigured to sustain a stable micro-ecosystem that performs much of the carnivorous digestive function. In the absence of a chromosome-scale genome, however, it has not been possible to assess how genome structure and gene duplication contribute to pitcher development, digestive ecology, or broader patterns of pitcher-plant genome evolution.

Despite considerable phenotypic and molecular developmental[17], environmental (e.g., [18]), and marker-based phylogenetic research (e.g.,[19,20]), the genomic basis of pitcher-leaf development

and digestive biology in *Sarracenia* have not been examined at chromosome scale. Previous studies of carnivorous plant genomes, including *Utricularia*, *Cephalotus*, and *Nepenthes*, were motivated by the evolution of genome reduction[21], convergent enzyme and amino acid usage[22], or sex determination and dominant-recessive subgenome functional partitioning[23]. In the absence of a resolved *Sarracenia* genome, how distinct duplication modes have shaped functional repertoires and duplicate retention–loss patterns across convergent pitcher plant lineages has remained unclear, limiting comparative insight. Although the incomplete *Cephalotus* assembly[22] has limited direct comparison with *Nepenthes*, a chromosome-scale *Sarracenia* assembly enables informative pairwise lineage inference.

Sarracenia belongs to the Ericales, a clade inferred to derive from an ancient allopolyploidy event[24] and encompassing a wide diversity of ecological strategies and well-known plants such as blueberry (*Vaccinium* species)[25]. Aligning a chromosome-scale *Sarracenia* genome with multiple Ericales relatives enables two related questions to be addressed: which duplicated blocks and structural features represent the ancestral framework within which carnivory evolved, and which rearrangements, fractionation patterns, and regional gene-family expansions are lineage-specific and potentially associated with pitcher evolution.

Addressing these questions requires integration of structural genomics, syntenic comparison, synonymous substitution rate (Ks)-based event dating interpreted using mixture components recovered for *Sarracenia* and other Ericales, and functional enrichment analyses. Synteny identifies conserved blocks inherited from the ancestral core eudicot and the *gamma* allohexaploidy[26,27] and reveals subsequent rearrangement. Ks distributions inform the relative timing of whole-genome and small-scale duplications through correspondence with background mutation rates. Comparisons with taxa such as *Vitis vinifera*, an early-branching rosid with conservative Ks behavior and modest structural evolutionary rate[28-30], provide temporal anchoring. Subgenome reconstruction clarifies how differential retention and loss shaped dominant and recessive compartments following polyploidy, a feature that may stabilize regulatory networks while permitting divergence of more flexible gene families[31-35]. Together, these approaches enable reconstruction of the genomic scaffolding underlying pitcher-leaf development and digestive function.

Repeated evolution of carnivorous traps provides a comparative framework [2]. By placing *Sarracenia* alongside *Nepenthes gracilis* (Caryophyllales), *Pinguicula gigantea* and *Utricularia gibba* (Lamiales), and by embedding these taxa within a broader phylogenetic context that includes non-carnivorous relatives, enables convergent solutions to be distinguished from lineage-specific innovations and ancestral traits. We used a nine-taxon occupancy-class framework based on exact presence–absence patterns across *Sarracenia*, *Nepenthes*, *Pinguicula*, *Utricularia*, *Vitis*, *Arabidopsis thaliana* (Brassicales, and our annotation reference), *Actinidia chinensis* (Ericales), *Mimulus guttatus* (Lamiales), and *Beta vulgaris* (Caryophyllales), to partition syntenically retained regulatory genes from tandem-duplicate–driven innovation and to trace functional categories across shared, convergent, and lineage-specific contexts. Because *Actinidia* is the only Ericales representative in this dataset, it serves as a filter for enrichments reflecting general Ericales biology rather than carnivory-associated change.

Here, we present a chromosome-level genome assembly of *Sarracenia purpurea*, detailed gene and repeat annotation, analysis of self-synteny revealing signatures of ancient whole-genome duplication and rearrangement, an Ericales anchoring framework situating *Sarracenia* within its clade, subgenome reconstruction using *Vitis* as outgroup, and evolutionary-functional analyses

integrating duplicate-gene classes with gene ontology enrichment. We further examine pitcher-specific gene expression for correlates of trap biology. These analyses clarify how distinct duplication modes partition developmental regulation from ecological function in a carnivorous leaf and provide a genomic perspective on repeated pitcher-plant evolution.

Results and Discussion

Chromosome-Scale Genome Assembly and Annotation

Genome Assembly of *Sarracenia purpurea*

The *Sarracenia purpurea* genome was assembled from high-coverage Oxford Nanopore Technology long reads[36] and reference-scaffolded against a publicly available *Sarracenia* genome assembly (https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_051027295.1/) into chromosome-scale sequences. The contig assembly spans 3,523,392,364 bp across 4,071 contigs with a contig N50 of 2,875,369 bp (**Table 1**). The scaffolded assembly totals 3,523,623,004 bp and is organized into 13 chromosome-length scaffolds (pseudochromosomes), consistent with the haploid chromosome number. Scaffold N50 is 241,630,572 bp. BUSCO[37] indicates 95.5% completeness for the genome and 84.1% completeness for the annotated gene set (**Table 1**; annotation below). Extended regions of reduced gene density mark presumptive peri- and centromeric compartments (**Figure 1D**), providing the chromosome-scale framework used for self-synteny, homoeolog reconstruction, and subgenome inference.

METRICS: GENOME	FILTERED CONTIG ASSEMBLY	RAGTAG SCAFFOLDED
COMPLETE BUSCOS (C)	2210 (95.0%)	2223 (95.5%)
COMPLETE AND SINGLE-COPY BUSCOS (S)	2007 (86.3%)	2036 (87.5%)
COMPLETE AND DUPLICATED BUSCOS (D)	203 (8.7%)	187 (8.0%)
FRAGMENTED BUSCOS (F)	35 (1.5%)	26 (1.1%)
MISSING BUSCOS (M)	81 (3.5%)	77 (3.4%)
TOTAL BUSCOS SEARCHED	2326	2326
# CONTIGS OR SCAFFOLDS	4071	1884
LARGEST CONTIGS OR SCAFFOLDS	45271496	321174296
TOTAL LENGTH	3523392364	3523623004
GC (%)	39.04	39.04
N50	2875369	241630572
N90	591864	7662836
L50	325	7
L90	1334	26
N'S PER 100 KBP	0	6.22
METRICS: ANNOTATION		
COMPLETE BUSCOS (C)		1956 (84.1%)
COMPLETE AND SINGLE-COPY BUSCOS (S)		1763 (75.8%)
COMPLETE AND DUPLICATED BUSCOS (D)		193 (8.3%)
FRAGMENTED BUSCOS (F)		173 (7.4%)
MISSING BUSCOS (M)		197 (8.5%)
TOTAL BUSCOS SEARCHED		2326

Table 1. Global assembly and gene space statistics for the *Sarracenia purpurea* reference genome. Columns include total assembled length, number of contigs and scaffolds, contig N50 and L50, scaffold N50 and L50, largest contig and scaffold sizes, GC content, and counts of contigs and scaffolds above specified size thresholds. BUSCO statistics are provided for the both contig- and scaffold-level assemblies, as well as for the genome annotation. These metrics confirm a high-quality contig assembly, the completeness of the gene space, and confirm chromosome-scale scaffolding for the 13 pseudochromosomes.

Annotation and Gene Space Characterization

The filtered, nonredundant annotation contains 28,835 protein-coding loci, each represented by a single transcript model. Transcript-supported models from trap, flower, and root tissues (see Methods, RNA-seq Processing and Expression Analyses) supported gene prediction and retained trap-associated expression contexts. Mean exon length is ~263 bp; genes average ~4–5 exons; ~25% are single-exon loci (**Table S1**).

Gene density is highly heterogeneous across chromosomes, with extended gene-poor regions interspersed with localized gene-rich segments (**Extended Data Fig. 1**). Genome-wide 1-Mb windows range from 0 to 42 genes per Mb and exhibit a strongly right-skewed distribution (**Extended Data Fig. 2A**).

RepeatMasker[38] estimates that ~88% of the genome is repetitive, with ~86.6% represented by interspersed repeats (**Table S2**). At the level of TE class composition, LTR (long terminal repeat) retrotransposons dominate, comprising ~71.8% of the genome, with Copia elements representing ~33.4% and Gypsy elements ~20.5%. DNA transposons account for an additional ~1.1%, while the remaining repeat fraction is distributed among unclassified elements (~12.4%) and other low-abundance families. Together, these proportions indicate that the *Sarracenia* genome has been shaped primarily by extensive LTR retrotransposon accumulation at the genome-wide scale.

Repeat content is uniformly high but varies modestly among pseudochromosomes: total TE coverage ranges from ~84.3-90.6% (Δ ~6.3 percentage points; coefficient of variation ~2%), and LTR-tagged coverage from ~38.9-45.3% (Δ ~6.3 percentage points; CV ~4%). This chromosome-scale variation, together with the strong window-level relationship between LTR coverage and low gene density (**Extended Data Figure 2B**), is consistent with megabase-scale compartmentalization of gene-rich and repeat-rich regions.

RepeatMasker divergence landscapes summarize relative repeat turnover (**Extended Data Figure 3**). Repeat fragments were binned by percent divergence from consensus (0–50%), and bp coverage was summed per bin to generate bp-weighted profiles. Genome-wide repeats show a mode near ~7% divergence, whereas LTR-tagged repeats peak near ~12%. The separation between genome-wide and LTR-only modes indicates that LTRs contribute disproportionately to structured turnover and complement chromosome-scale tracks (**Figure 1D, Extended Data Figure 1**) by showing temporally heterogeneous repeat accumulation.

Duplication history and subgenome architecture

Sarracenia Self-Synteny and Structural Genomic Landscape

Self-synteny analysis using SynMap in CoGe[39-41] shows that *Sarracenia purpurea* retains extensive chromosomal segments consistent with an ancient whole-genome duplication (WGD)

(**Figure 1D**; cf. [42]). Several chromosomes preserve long, near-collinear duplicated regions, whereas others contain shorter, fragmented blocks, consistent with gene loss, local rearrangement, and TE insertion. Ks distributions of syntenic gene pairs[43,44] resolve two dominant components with modal densities near ~0.65 and ~1.7 (**Figure 2**).

To contextualize duplication timing relative to repeat accumulation, Ks distributions were compared with RepeatMasker LTR divergence profiles using a calibrated time axis (where the *gamma* event was time-fixed and background mutation rate was standardized for both Ks and LTR divergence; **Figure 2**). The LTR divergence landscape is sharply peaked at low divergence, consistent with a recent burst of LTR activity (~10 Mya), whereas the Ks profile resolves a younger lineage-specific duplication component (~45–50 Mya) and an older component corresponding to the core eudicot *gamma* hexaploidy (~120–125 Mya; cf. [45-47]). The temporal separation between the LTR peak and both Ks components indicates that large-scale duplication and subsequent fractionation preceded the major episode of LTR proliferation. Thus, repeat expansion appears to have preferentially accumulated in already fractionated regions rather than driving the duplication event itself.

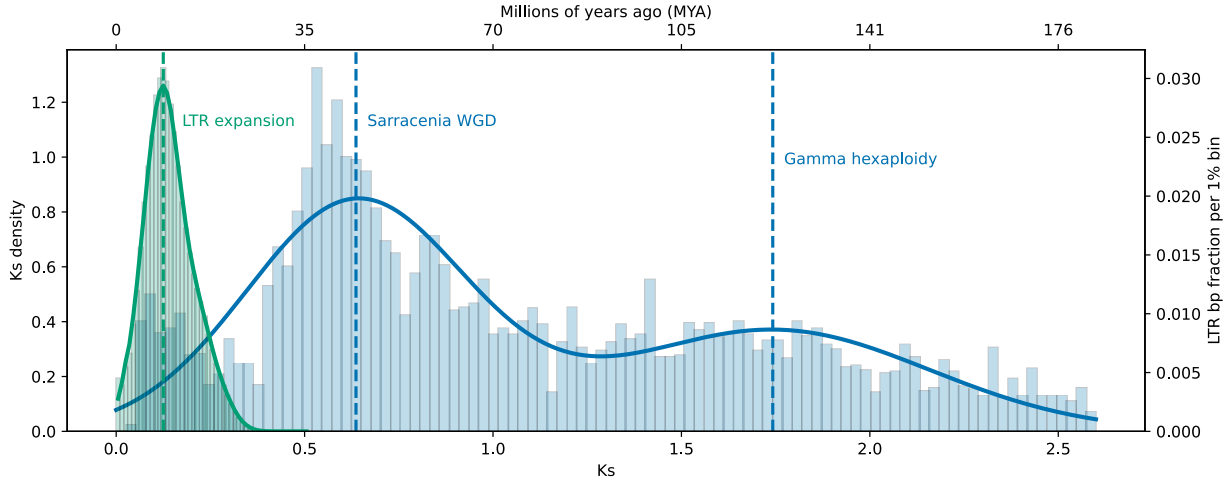


Figure 2. Temporal relationship between gene duplication and LTR expansion. Ks distributions of self-syntenic duplicate pairs compared with LTR divergence landscapes on a calibrated time axis. Dashed vertical lines indicate modal components corresponding to a lineage-specific WGD and the *gamma* paleohexaploidy. The LTR expansion peak is temporally distinct from both duplication components.

Multispecies Ericales Comparative Anchoring and Duplication Timing

To evaluate duplication timing in a broader context, Ks distributions were examined across Ericales taxa. All sampled species retain a broad, high-Ks component consistent with the core eudicot *gamma* hexaploidy. Some taxa, including *Actinidia* and *Vaccinium*, also exhibit prominent lower-Ks components (results not shown), likely reflecting more recent duplication events. Because these younger events obscure separation between *gamma*-derived and post-*gamma* signals that may be Ericales-general, *Actinidia* and *Vaccinium* were not used for cross-taxon comparison, although these genomes remain very structurally informative.

Focusing on taxa with simpler duplication landscapes, *Sarracenia*, *Rhododendron*, *Primula*, and *Impatiens* each display a distinct younger Ks mode in addition to *gamma*. Independent calibration using the *gamma* component within each species (set to 122.5 Mya) yields inferred ages of approximately from 45-64 Mya (**Extended Data Figure 4**).

Although the ages for post-*gamma* events detected in *Rhododendron* and *Impatiens* are particularly similar, differences in peak breadth, prominence, and modeling sensitivity among all taxa prevent confident assignment of homology among events. Therefore, we treat these Ks components as possibly lineage-specific, or potentially shared, duplications without asserting a single Ericales-wide polyploidy event.

Syntenic Architecture and Subgenome Structure Relative to *Vitis*

Comparative synteny across Ericales confirms that *Sarracenia purpurea* retains a conserved chromosomal backbone despite lineage-specific rearrangement and fractionation. Anchoring against *Actinidia*, *Vaccinium*, *Rhododendron*, and *Impatiens* reveals large blocks of conserved gene order tracing to shared ancestry. To place these relationships in a stable evolutionary coordinate system, we used *Vitis vinifera* as an outgroup reference.

Mapping *Sarracenia* chromosomes onto *Vitis* reveals extended syntenic correspondence, predominantly at an approximate 2:1 ratio, consistent with paleotetraploid ancestry and subsequent fractionation bias[40,48] (**Figure 3A**). Visualization of multiple Ericales taxa within the same *Vitis*-referenced framework emphasizes shared chromosomal correspondence while highlighting lineage-specific differences in syntenic anchor continuity (**Figure 3B**).

Within *Sarracenia*, syntenic anchor density alternates along chromosomes between regions of high and low retention, corresponding to dominant and recessive subgenomic segments (**Extended Data Figure 5**). Dominant segments preserve dense arrays of *Vitis*-derived anchors, whereas recessive segments show reduced anchor density and extended intervals of sparse signal. These recessive compartments frequently coincide with repeat-rich regions, consistent with the negative association between anchor density and transposable-element abundance observed genome-wide (Spearman's $\rho \approx -0.5$; $P \ll 0.001$).

Blockwise retention heatmaps quantify these patterns directly. Across *Vitis* chromosomes, *Sarracenia* syntenic anchor density forms a patchwork of high- and low-retention segments rather than chromosome-scale uniformity (**Extended Data Fig. 6A**). Adjacent bins often shift abruptly between conserved and fractionated states, showing that dominant and recessive segments are interleaved along chromosomes rather than segregated into intact domains (**Extended Data Fig. 6B**). This mosaic architecture reflects cumulative fractionation and localized rearrangement following polyploidy.

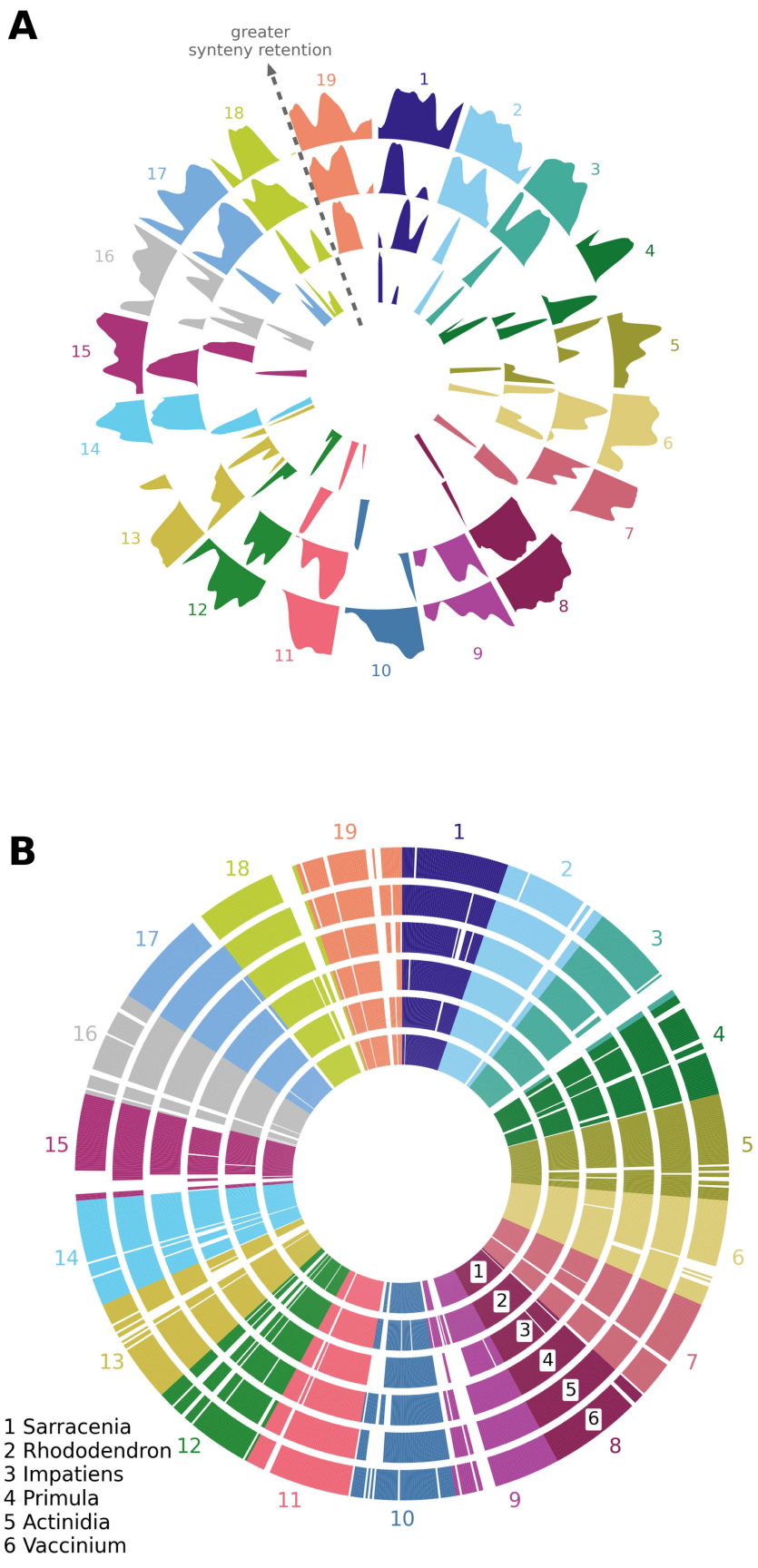

A
greater
synteny retention
19
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
B
1 Sarracenia
2 Rhododendron
3 Impatiens
4 Primula
5 Actinidia
6 Vaccinium

Figure 3. Radial representation of *Sarracenia–Vitis* fractionation bias and multispecies anchoring.
(**A**) Fractionation-bias profiles along *Vitis* chromosomes, showing alternating high- and low-retention regions consistent with dominant and recessive segments. (**B**) Multispecies anchor continuity across Ericales mapped onto the same *Vitis* coordinate system. In (**A**), the dashed guide indicates increasing synteny retention within a given chromosome-layer representation. For each *Vitis* chromosome, up to four concentric syntenically anchored layers were permitted for display, representing the four best-matching *Sarracenia* segments. The outer filled ribbon reflects the strongest syntenic match and the inner ribbons progressively weaker matches. In (**B**) concentric rings correspond to *Sarracenia*, *Rhododendron*, *Impatiens*, *Primula*, *Actinidia*, and *Vaccinium* (inner to outer), with each ring encoding per-bin syntenic anchor presence along *Vitis* chromosomes. (**B**) emphasizes the shared chromosomal coordinate framework across Ericales while illustrating lineage-specific variation in anchor presence and continuity, independent of *Sarracenia* fractionation-bias magnitude.

Comparative functional architecture of carnivory

Functional Architecture from Homolog Occupancy Classes Across Nine Taxa

To distinguish carnivorous innovation from background phylogenetic signal, we constructed a nine-taxon comparative framework including the carnivorous plants *Sarracenia*, *Nepenthes*, *Pinguicula*, and *Utricularia*, and their non-carnivorous relatives of reference species *Actinidia*, *Vitis*, *Arabidopsis*, *Mimulus*, and *Beta*. *Actinidia*, the closest non-carnivorous Ericales relative in our sample[49], provides an Ericales baseline. *Mimulus* and *Beta* serve as non-carnivorous controls for Lamiales and Caryophyllales carnivores[49]. *Vitis* anchors deeper structure, and *Arabidopsis thaliana* supplies standardized homolog identifiers and GO annotation.

We used two basic classification operations that are conceptually distinct. First, occupancy classes are defined by strict intersections of *Arabidopsis*-mapped homolog presence and absence across taxa, so that each class corresponds to an exact nine-taxon occupancy pattern. Second, functional enrichment operates on unions of occupancy classes, because convergence is expected at the level of biological processes even when underlying gene families differ.

All *Sarracenia* tandem and syntenic duplicates were mapped to their closest *Arabidopsis* homolog; equivalent mappings were generated for the other taxa. For each homolog detected in at least one species, a nine-element presence–absence vector was recorded, producing a fixed universe of intersection-defined occupancy classes. All *Sarracenia* duplicates were then assigned to these classes, establishing a phylogenetically explicit scaffold for functional comparison.

GO enrichment was performed on lineage-specific foregrounds constructed as unions of occupancy classes. For example, *Sarracenia*-specific novelty was defined by the union of classes containing *Sarracenia* and excluding all other taxa (**Extended Data Figure 7**). The pitcher-associated foreground combined enriched GO terms from *Sarracenia* and *Nepenthes*, reflecting convergence at the level of biological processes rather than strict gene identity. A similar union logic defined pan-carnivore foregrounds across the four carnivorous genera.

To remove background Ericales signal, GO categories enriched in both a *Sarracenia* foreground and *Actinidia* were excluded. Intersection-defined occupancy classes thus provide evolutionary structure; union-defined foregrounds capture functional innovation and convergence; *Actinidia* filtration removes ancestral Ericales bias.

Sarracenia-Specific Functional Innovations

Comparative foreground analyses identify detoxification, redox, antifungal, and cuticle-modifying activities associated with the *Sarracenia* lineage, including both *Sarracenia*-only and carnivore-union categories (**Table S3**). These functions recur across duplication modes and

include many tandem-derived loci. Named effectors include PEN3-like ABC transporters[50], *GSTF*-related enzymes[51], *CYP86*[52]- and *LACS2*[53]-associated cuticle-biosynthetic genes; additional defense-associated homologs in **Table S3** support an expanded antifungal and xenobiotic-handling repertoire in *Sarracenia* relative to comparator lineages. A *Sarracenia* homolog of *DML1*[54], implicated in maintaining stress-responsive gene accessibility by limiting inappropriate DNA methylation, also appears among tandem-derived loci, pointing to chromatin-level regulation within this innovation set.

Several enriched categories overlap with *Actinidia* and are removed under the strict *Actinidia* filter. After filtration, six biological-process GO categories remain enriched in *Sarracenia* but absent from *Actinidia*: response to oxidative stress, response to wounding, ethylene-activated signaling, cell differentiation, root-hair cell differentiation, and mitochondrial mRNA modification (**Table S3**). By contrast, detoxification, glutathione-based redox processes, antifungal functions, and cuticle-associated terms, although prominent in the broader *Sarracenia* landscape, are shared with *Actinidia* and therefore excluded by this filter.

Across the other carnivorous lineages, lineage-specific innovation sets emphasize secretory and vesicle-trafficking processes in *Nepenthes*, osmotic and desiccation-related pathways in *Pinguicula*, and rapid cell-wall remodeling and aquatic-habitat metabolic adjustments in *Utricularia*. Together, these analyses indicate a shared stress-, interaction-, and metabolism-associated repertoire across carnivores, with *Sarracenia* distinguished by the oxidative, wound-responsive, and ethylene-associated processes summarized in **Table S3**.

Sarracenia-Nepenthes **Shared Pitcher Toolkit**

The *Sarracenia–Nepenthes* comparison examines how two species with independently evolved pitfall traps deploy duplicate-gene repertoires. Because convergence is evaluated at the level of enriched biological processes rather than strict homolog identity, the foreground consists of the union of GO terms enriched in either lineage (**Table SB**). Under this union-based framework, both taxa are enriched for oxidative processes, cell-wall remodeling, and epidermal patterning functions, reflecting shared structural and ecological demands of constructing tubular traps that retain fluid and process arthropod prey.

Although *Nepenthes* and *Sarracenia* differ in digestive strategy, with the former secreting hydrolytic enzymes and the latter relying more heavily on microbial mediation, their overlapping GO categories indicate convergent recruitment of stress-responsive, cuticle-associated, and wall-modifying pathways. Convergence is therefore evident at the level of functional modules rather than specific regulators, consistent with independent recruitment of functionally analogous genes to achieve similar pitcher morphologies.

Deep Carnivorous Signatures Across Four Independent Lineages

Union-based functional enrichment across *Sarracenia*, *Nepenthes*, *Pinguicula*, and *Utricularia* (**Table SC**) identifies recurrent oxidative and biotic-interaction processes. These pan-carnivore foregrounds include *Arabidopsis* homologs such as AT1G15520, AT2G29430, and AT4G39840, annotated for defense response, interspecies interaction, and microbial-stimulus regulation. Their recurrence across four phylogenetically distant carnivorous lineages indicates repeated recruitment of stress- and microbe-associated molecular architectures during the evolution of leaf-derived traps.

Within this shared framework, lineage-specific emphases remain evident. *Sarracenia* shows stronger enrichment of detoxification and redox-buffering modules. *Nepenthes* emphasizes secretory and vesicle-trafficking processes. Lamialean carnivores, *Pinguicula* and *Utricularia*, highlight osmotic regulation and cell-wall adjustment (**Tables SD–SF**). Despite these differences, the overlapping foregrounds point to a common reliance on stress-responsive, interaction-mediated, and wall-modifying capacities. Collectively, these functions delineate a molecular repertoire repeatedly mobilized during the evolutionary transformation of leaves into specialized organs for prey acquisition and nutrient assimilation.

Distinguishing *Sarracenia* Innovations from Ericales Background

Comparative filtration using *Actinidia* narrows lineage-specific signatures in *Sarracenia* by excluding GO terms that reflect ancestral Ericales biology. Because *Actinidia* is a non-carnivorous Ericales relative lacking pitcher morphology and digestive physiology, GO categories enriched in both *Sarracenia* and *Actinidia* most parsimoniously represent background Ericales functions rather than carnivory-associated innovation.

After filtration, six biological-process GO categories remain uniquely enriched in *Sarracenia*: response to oxidative stress, response to wounding, ethylene-activated signaling, cell differentiation, root-hair cell differentiation, and mitochondrial mRNA modification (**Table SG**). These categories define a regulatory and stress-integration axis elaborated along the *Sarracenia* lineage and plausibly linked to pitcher development and physiology. By contrast, detoxification, glutathione-dependent redox activity, antifungal responses, cuticle modification, and broader transport-related functions do not survive the *Actinidia* filter. Although central to pitcher biology, their presence in *Actinidia* indicates retention from ancestral Ericales capacities rather than uniquely derived innovation.

Together, this filtration framework indicates that *Sarracenia* carnivory emerged through selective redeployment and elaboration of regulatory and stress-integration modules embedded within the Ericalean genomic background. After removal of shared Ericales signal, only a restricted set of enriched regulatory categories remains uniquely associated with the *Sarracenia* lineage (**Table S3**).

Regulatory deployment in pitcher tissue

Sarracenia gene expression analyses pursue the same conceptual objective as genomic comparisons but rely on regulatory rather than phylogenetic inputs. Differentially expressed genes (DEGs) were identified between flowers and pitcher leaves, mapped to *Arabidopsis* homologs using the same orthology framework, and subjected to GO enrichment. Because differential expression reflects regulatory deployment rather than taxonomic occupancy, presence–absence occupancy classes were not applied; instead, enriched GO terms from DEG sets were compared qualitatively with occupancy-class modules.

Transcriptome comparisons show that many genes underlying the nine-taxon occupancy class structure are also strongly expressed in traps. Within the full DEG set, multiple *Sarracenia* homologs of *PEN3*, *ZIP6*[55], and *DML1* display significant trap-biased expression (negative flower vs. leaf (FvL) $\log_2$FC), consistent with xenobiotic export, micronutrient transport, and redox-responsive chromatin regulation (**Tables SH and SI**). Several pan-carnivore homologs (AT1G01580, AT1G24140, AT2G16390) also show trap-biased expression for at least one paralog (**Tables SH, SI, and SC**). The heatmap and volcano plots (**Figure 4**; **Extended Data**

Figure 8) highlight the strongest FvL contrasts, while genes such as *PEN3*, *ZIP6*, and *DML1* exhibit consistent trap-biased expression across the broader DEG distribution. Flower-biased expression includes terpene synthases (*TPS03*[56], *TPS-CIN*[57]), peroxidases (*PER7*[58]), lipid-transfer proteins (*LTP4*[59]), cytochrome P450s (*CYP76G1*[60], *CYP71A25*[61]), and canonical floral organ identity regulators (*PI*, *SEP1*, *SEP4*, *AGL16*[62,63]) (**Tables SH and SI**). Their enrichment in flowers indicates that pitchers do not co-opt canonical floral identity networks, at least at the developmental stage sampled.

Trap-upregulated genes define the molecular architecture of pitcher specialization. Strongly trap-biased transcripts include *TDR1*[64], *KCS2*[65], *DWF1*[66], *HSP17.6C*[67], *MMP*[68], reflecting cuticle biosynthesis, brassinosteroid metabolism, stress buffering, osmotic regulation, and proteolytic remodeling. Additional regulators such as *CER1*[69], *CBF4*[70], *WRKY9*[71], *PCNA1*[72], *PRL*[73], and *DAO1*[74] support integration of surface modification, stress-response, hormonal, and redox modules.

Flower-enriched tandem-only homologs (AT1G02205, AT1G55260, AT1G62045) suggest broader defense capacity in reproductive tissues (**Tables SH and SI**). Distinct paralogs of *PEN3* and *ZIP6* occur in both trap- and flower-biased sets, consistent with tissue-level subfunctionalization.

Collectively, these transcriptional patterns reinforce the homolog occupancy class framework: lineage-specific and carnivore-level innovations identified genomically are actively deployed in traps, while other modules remain vegetative or floral. Pitchers represent modular redeployment of stress-response, developmental, metabolic, and structural programs within an Ericalean genomic context.

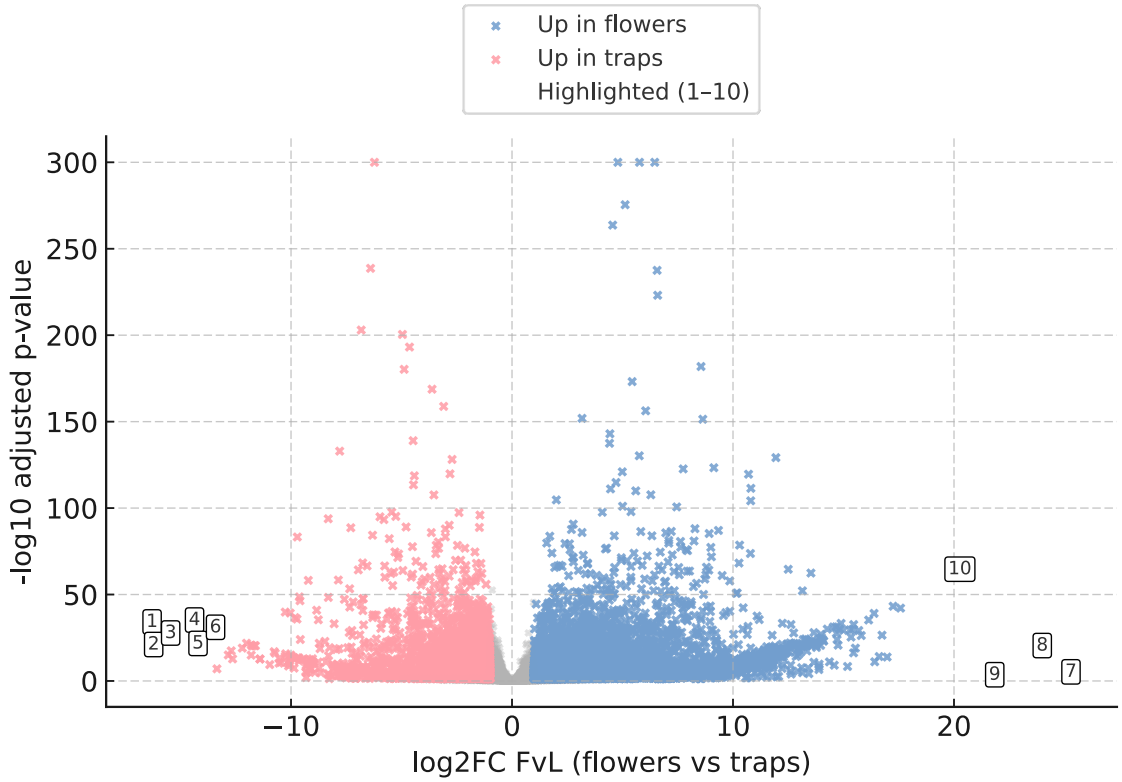


Figure 4. Differential expression volcano plot for flowers vs. pitcher leaves. Volcano plot of differential expression between flowers and pitcher leaves in *Sarracenia purpurea*. Positive $\log_2$FC indicates flower bias; negative indicates trap bias (**Tables SA and SG**). Blue points denote flower-up genes; pink points denote trap-up genes (FDR < 0.05, $|\log_2 FC| > 1$). Ten transcripts with the largest expression shifts are annotated by *Arabidopsis* best BLAST hit (**Table SI**):

1. AT3G23230 (*TDR1*) — trap-up ERF transcription factor
2. AT1G04220 (*KCS2*) — trap-biased 3-ketoacyl-CoA synthase
3. AT3G19820 (*DWF1*) — trap-up brassinosteroid biosynthesis
4. AT1G53540 (*HSP17.6C*) — trap-enriched heat-shock protein
5. AT1G70170 (*MMP*) — trap-up metalloprotease
6. AT3G09110 — trap-biased gene of unknown function
7. AT5G59310 (*LTP4*) — flower-biased lipid-transfer protein
8. AT1G30870 (*PER7*) — flower-up peroxidase
9. AT3G25830 (*TPS-CIN*),
10. and 22AT4G16740 (*TPS03*) — flower-enriched terpene synthases

Integrated Model for *Sarracenia* Pitcher Evolution

Combined syntenic, duplication-mode, and functional analyses support a two-tiered model of pitcher evolution. A conserved regulatory scaffold, preferentially retained on dominant subgenomes, anchors development. Core regulators including *AGO1*[75], *BRX*[76], *GATA11*[77], *ETC1*[78], and *RCD1*[79] persist as single or low-copy loci in high-retention syntenic blocks (**Table SI**), consistent with dosage sensitivity following ancient polyploidy. Their developmental roles in *Arabidopsis* support redeployment of conserved leaf regulatory pathways rather than invention of novel developmental circuits.

Superimposed on this scaffold are tandem expansions concentrated in structurally dynamic regions. These families are enriched for detoxification, glutathione-mediated redox buffering, antifungal defense, and cuticle and wall modification (**Tables SA and SG**), matching the biochemical demands of trap function.

Expression responsiveness differs significantly between duplication classes. Tandem-derived genes show greater responsiveness than syntenic genes (cf. [80]) (Mann–Whitney U test, $p = 6.94 \times 10^{-18}$; Cliff's $\delta = 0.179$; n = 934 tandem, 4,578 syntenic; **Figure 5A**). Responsiveness was quantified as $|\log_2 FC|$ between pitcher and flower tissues. Tandem genes display higher median responsiveness and broader variance, whereas syntenic genes show narrower distributions consistent with stronger regulatory constraint.

Functional breadth analysis using GO rarefaction analysis indicates that tandem-derived genes exceed syntenic genes in both term richness and effective functional diversity (D1) (**Figure 5B,C**), based on exponential Shannon entropy (cf. [81]). Rarefaction controls for gene-set size and is descriptive rather than inferential. These results support disproportionate functional diversification following tandem duplication.

Comparisons with *Nepenthes* show convergence at the level of functional categories, particularly oxidative and wall-modifying processes, under the union-based comparative framework. Convergence reflects analogous deployment in two pitcher-bearing species rather than gene-level identity.

Duplication timing indicates that the *Sarracenia* lineage-specific WGD predates major LTR retrotransposon expansion, placing large-scale gene duplication earlier than repeat proliferation.

TE-rich regions may provide structurally permissive environments for tandem expansion without destabilizing dosage-sensitive networks.

Together, these patterns indicate that *Sarracenia* pitcher evolution involved retention of a dosage-constrained regulatory backbone on dominant subgenomes and diversification of effector modules through tandem expansion in more flexible compartments. Dominant subgenomes preferentially retain regulatory loci, whereas recessive and boundary regions accumulate detoxification and interaction modules (**Tables SA, SG, and SI**). This partitioning permits rapid ecological specialization while maintaining developmental stability.

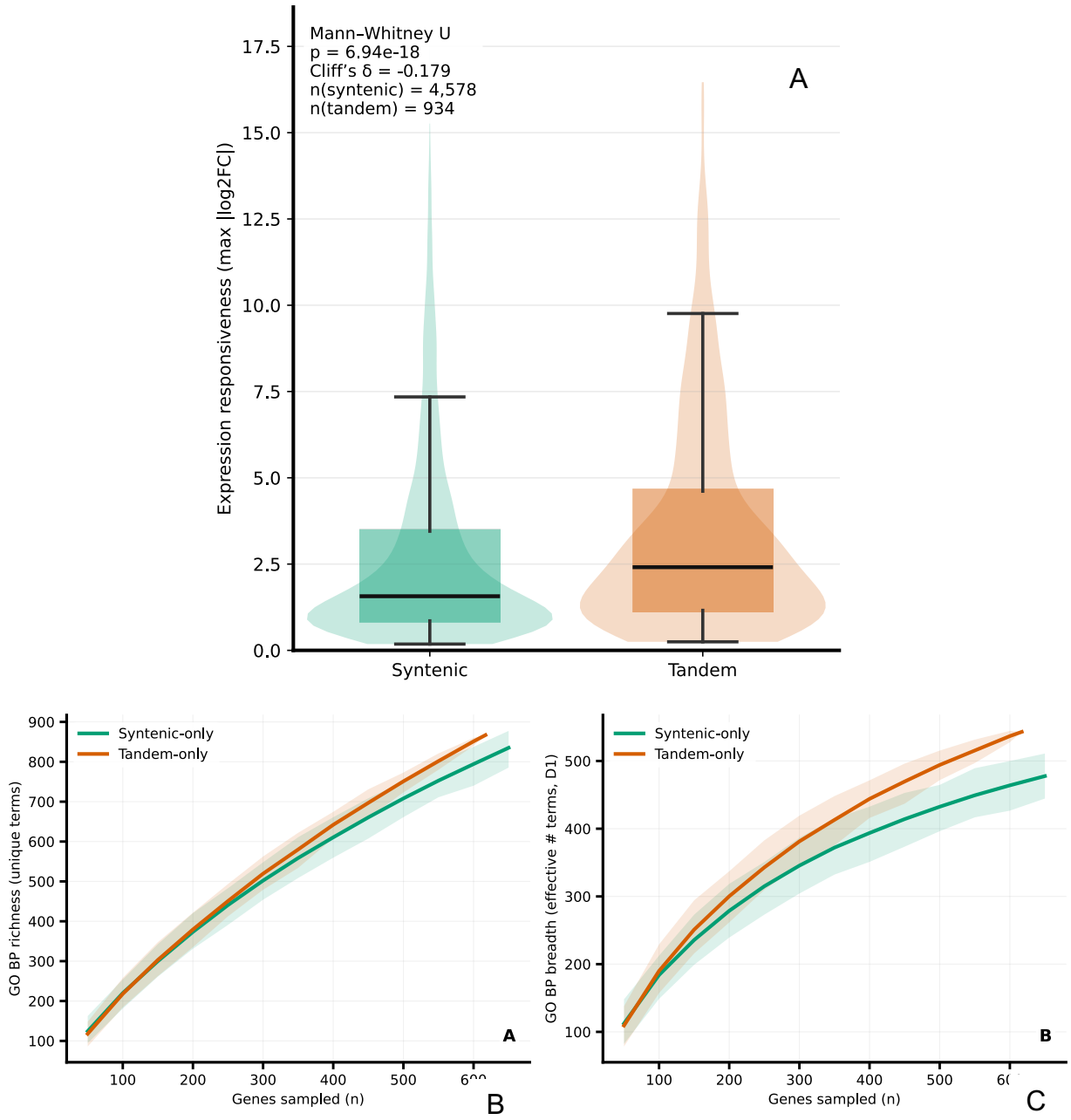


Figure 5. Expression responsiveness of syntenic and tandem gene sets and their functional breadth inferred from GO rarefaction. (**A**) Violin and box plots show expression responsiveness, defined as absolute $\log_2$ fold change between pitcher and flower tissues. Boxes indicate medians and interquartile ranges; violins depict full kernel density distributions. Statistical significance was assessed using a two-sided Mann–Whitney U test, with effect size reported as Cliff's δ; sample sizes are shown in the plot. Tandem-derived genes exhibit higher median responsiveness and greater variance than syntenic genes, consistent with relaxed regulatory constraint in tandem-expanded families (**Tables SA and SH**). Similarly, rarefaction curves depict greater functional breadth for tandem duplicates based on GO biological-process annotations. (**B**) Mean GO richness and (**C**) effective functional breadth (D1) are shown as functions of subsampled gene number, with shaded envelopes indicating subsampling variability. Functional breadth was quantified as the number of unique GO terms and the exponential of Shannon entropy (cf. [81]). Rarefaction controls for unequal gene-set size and visualizes contrasts in functional breadth without formal cross-curve hypothesis testing (**Tables SA and SC**).

Conclusions

The chromosome-scale genome assembly of *Sarracenia purpurea* provides an integrated structural and functional framework for the origin of pitcher leaves, clarifying how distinct duplication modes contributed to complementary evolutionary roles. The genome retains extensive Ericales-derived synteny and exhibits pronounced subgenome asymmetry. Dominant subgenomes are enriched for dosage-sensitive developmental and signaling regulators that form a conserved architectural scaffold, whereas recessive compartments harbor more flexible gene families. Tandem expansions in structurally dynamic regions contribute detoxification, redox, antifungal, and cuticle-associated functions that underpin the biochemical and ecological flexibility of a microbially mediated digestive system.

Expression data link this genomic partitioning to regulatory deployment in pitcher tissue. Pitchers upregulate homologs of *PEN3*, *ZIP6*, and *DML1*, together with broader suites of stress-associated, metal-handling, and wall-modifying loci. These genes correspond to enriched *Sarracenia*-specific and carnivore-union GO categories and support coordinated xenobiotic export, micronutrient balance, reactive-metabolite buffering, and surface remodeling. Distinct paralogs occupying different homolog occupancy classes indicate divergent regulatory trajectories among duplicates, suggesting subdivision and refinement of stress-integration modules during adaptation to a semi-open, microbially rich trap environment.

Comparative analyses across carnivorous taxa show that pitcher evolution repeatedly recruits shared developmental and stress-responsive capacities, while functional diversification reflects ecological context. Parallel enrichment of stress-, interaction-, and wall-modifying pathways across carnivorous lineages supports a convergent functional basis for trap formation, whereas lineage-specific effector composition distinguishes digestive strategies.

Actinidia-based filtration narrows uniquely derived innovation in *Sarracenia* to six GO biological-process categories: response to oxidative stress, response to wounding, ethylene-activated signaling, cell differentiation, root-hair cell differentiation, and mitochondrial mRNA modification. In contrast, detoxification, glutathione-based redox buffering, antifungal responses, cuticle-associated pathways, and transport-related functions likely represent elaborations of ancestral Ericales capacities rather than de novo invention. Carnivory thus emerges primarily through redeployment and amplification of pre-existing genomic programs.

Together, these results show how ancient polyploidy and tandem duplication partition regulatory stability from ecological flexibility. Conserved regulators retained on dominant subgenomes

maintain developmental robustness, whereas tandem expansions generate the metabolic specialization required for a fluid-filled, microbially mediated trap.

A future chromosome-scale genome for *Cephalotus* will enable direct tests of whether components of the shared pitcher toolkit represent true molecular convergence or parallel elaborations of distinct genomic substrates. The *Sarracenia* genome therefore anchors a unified model of carnivorous plant innovation integrating structural legacy, regulatory conservation, and effector diversification.

Methods

Genome Resources and Analytical Scope

All analyses were conducted on a chromosome-scale genome assembly of *Sarracenia purpurea* comprising 13 pseudochromosomes (publicly available at https://genomevolution.org/coge/GenomeInfo.pl?gid=69349). All genomic coordinates, gene models, and syntenic relationships refer to this assembly and its associated primary protein-coding gene annotation (also freely available at the above CoGe link). Each gene locus was represented by a single representative protein-coding model. Genome-wide transposable element annotations were available as genomic intervals and were used only as positional features in downstream analyses. Gene density was quantified by counting gene midpoints in non-overlapping 1-Mb windows across pseudochromosomes larger than 5 Mb, using coordinates from the final gene annotation GFF.

Genome Construction and Annotation Workflow

Plant Material, DNA Extraction, and Oxford Nanopore Sequencing

Plant material and voucher specimens for sequencing were collected from mature individuals of *Sarracenia purpurea* at Mendon Ponds Park, Rochester, New York (**Figure 1**). Species identification was performed in the field. Voucher specimens are currently deposited at the University at Buffalo. Leaf tissue was placed into individual vials and stored at −80 °C until DNA extraction.

Approximately 10 g of flash-frozen tissue was used for isolation of high-molecular-weight (HMW) genomic DNA. Nuclei isolation followed the BioNano NIBuffer protocol, in which frozen leaf tissue was homogenized in liquid nitrogen and subjected to a nuclei lysis step using IBTB buffer supplemented with spermine and spermidine and filtered immediately prior to use. IBTB buffer consisted of Isolation Buffer (IB; 15 mM Tris, 10 mM EDTA, 130 mM KCl, 20 mM NaCl, 8% (m/V) PVP-10, pH 9.4) containing 0.1% Triton X-100 and 7.5% (v/v) β-mercaptoethanol (BME), mixed and chilled on ice. The homogenized tissue-buffer mixture was strained to remove undissolved plant material. Triton X-100 was then added to a final concentration of 1% to lyse nuclei, followed by centrifugation at 2000 × g for 10 min to pellet the nuclei. Following nuclei isolation, DNA was extracted using a cetyltrimethylammonium bromide (CTAB) protocol modified for Oxford Nanopore sequencing. The quality and concentration of HMW genomic DNA were assessed using a NanoDrop spectrophotometer and agarose gel electrophoresis following standard protocols. Extracted genomic DNA was further purified using a Qiagen Genomic-Tip 500/G according to the manufacturer's instructions. Aliquots of purified

DNA were subjected to treatment with Short Read Eliminator (SRE; PacBio) to perform size selection and enrich for DNA fragments longer than 10 kb

High-molecular-weight genomic DNA was thereafter sequenced using Oxford Nanopore Technology (ONT) platforms. Library preparation and sequencing followed standard ONT protocols appropriate for high-molecular-weight plant genomic DNA. Sequencing generated long-read data without pre-filtering for read length or read quality. All genome assemblies and downstream analyses were performed using unfiltered ONT reads, unless explicitly stated otherwise, in order to preserve information from repeat-rich and structurally complex regions of the genome.

Genome Assembly

Unfiltered Oxford Nanopore reads were assembled using hifiasm v0.25.0 with the --ont flag enabled, which activates the assembler's long-read mode optimized for nanopore data. Assembly was performed to generate a collapsed primary (haploid) assembly, rather than phased haplotypes, reflecting the study's focus on genome structure, gene content, and duplication history rather than allelic variation.

Assembly completeness was evaluated using BUSCO v5.4.3 with the *eudicots_odb10* dataset, which assesses the presence and copy number of conserved single-copy orthologs expected in eudicot genomes. Assembly contiguity and size statistics were evaluated using QUAST v5.2.0.

To assess read support and facilitate downstream curation, unfiltered ONT reads were aligned back to the initial assembly using minimap2 v2.24, employing presets appropriate for long-read alignment. Resulting alignments were sorted using SAMtools v1.16.1.

Contamination Screening and Assembly Curation

The initial genome assembly was screened for potential contamination using BlobTools v1.1.1 in conjunction with BLAST+ v2.12.0. For each contig, taxonomic assignments were inferred from BLASTN similarity searches. Contigs classified as Streptophyta were retained, as were contigs lacking confident taxonomic assignment (no-hit or undefined). Contigs assigned to non-plant taxa were removed from further analyses.

Unfiltered ONT reads aligning to the retained contigs were extracted and reassembled using hifiasm with the same parameters as the initial assembly. This produced a curated assembly in which non-plant sequences were excluded while preserving read depth and long-range genomic structure across retained regions. Assembly quality and completeness were reassessed using BUSCO and QUAST as described above.

Homology-Based Scaffolding

The curated assembly was scaffolded using RagTag v2.1.0, employing a publicly available *Sarracenia purpurea* genome assembly (NCBI accession GCA_051027295.1) as a reference guide for contig ordering and orientation. RagTag was used strictly in reference-guided scaffolding mode; nucleotide sequences were not modified, and no gap filling or sequence correction was performed.

This scaffolding step produced a chromosome-scale assembly comprising 13 pseudochromosomes and 1,871 unplaced scaffolds. All subsequent analyses were conducted using this RagTag-scaffolded genome assembly.

Genome Annotation

Repetitive elements were identified using a combination of RepeatModeler v2.0.4 and EDTA v2.0.1, capturing both de novo repeat families and structurally defined transposable element classes. Repeat libraries generated by the two approaches were merged and collapsed into a non-redundant repeat library using CD-HIT v4.8.1, applying thresholds of 95% sequence identity and 90% alignment coverage.

Protein-coding gene prediction was performed using GeMoMa v1.9.0 with default parameters. Reference proteomes from five angiosperm species were used to guide prediction: *Arabidopsis thaliana*, *Camptotheca acuminata*, *Actinidia eriantha*, *Camellia lanceoleosa*, and *Vaccinium darrowii*. These species were selected to provide phylogenetic breadth within eudicots while remaining relevant to Ericales. RNA-seq data were incorporated as evidence for the prediction of splice sites.

This annotation was refined using EVidenceModeler (EVM) v2.1.0 to standardize naming and retain the best-supported primary gene model at each locus. The resulting EVM-integrated GFF3 annotation served as the final gene annotation used for all downstream analyses. Annotation completeness was assessed using BUSCO v5.5.0 on the predicted protein set.

Syntenic Anchor Processing and Coordinate Binning

Syntenic Anchor Preprocessing

Syntenic anchors were provided as gene-gene correspondences between *Sarracenia purpurea* and reference genomes. These runs were executed in SynMap[39-41] with FractBias[48] enabled to generate standardized syntenic anchor sets against *Vitis* as a joint reference. While FractBias summary outputs were inspected for consistency, all fractionation metrics, dominance assignments, and downstream visualizations reported here were recomputed independently from the raw syntenic anchor tables. SynMap settings for FractBias were:

- DAGChainer: relative gene order = nucleotide distance; maximum distance between two matches = 20 genes; minimum number of aligned pairs = 5 genes
- Skip random/unknown chromosomes
- Syntenic depth: Quota Align[82]
- Ratio of coverage depth: *Vitis* = 1 against all other taxa = 4
- Overlap distance: 40
- Window size: 100 genes; use all genes in target genome

Each anchor record obtained from SynMap (at the link: “Results with synonymous/non-synonymous rate values”) contained identifiers for the source and target chromosomes, genomic coordinates for both genes, and a Ks value. FractBias results were downloaded at the “Fractionation Bias synteny report” and “Fractionation Bias sliding window results” links.

Anchors were filtered as follows:

- Anchors lacking defined chromosomal coordinates were excluded.
- Anchors mapping to unplaced or unanchored scaffolds were excluded.
- Anchors lacking Ks values were excluded from Ks-based analyses but retained for purely structural analyses, they were used for reference-coordinate projection across multiple

Ericales taxa, where anchors were treated strictly as structural signals and were not used for dominance or fractionation-based inference.

Reference-Coordinate Binning of Syntenic Anchors

All reference-based analyses were conducted in a fixed bin coordinate system.

This reference-coordinate framework was applied uniformly wherever syntenic anchor positions were summarized along the *Vitis* genome, including both lineage-specific and multispecies analyses, where anchors were interpreted descriptively as indicators of syntenic presence rather than quantitative density.

For each reference chromosome c:

Let L_c be the maximum coordinate observed among all anchors on chromosome c.

Define a fixed number of bins B, held constant across all chromosomes for a given analysis.

Define the bin width:

$$\Delta_c = \frac{L_c}{B}$$

For each anchor with midpoint coordinate m on chromosome c, assign the anchor to bin:

$$b = \left\lfloor \frac{m}{\Delta_c} \right\rfloor$$

Clip b such that:

$$0 \leq b < B$$

This produces a discrete bin index for every anchor, enabling direct comparison across chromosomes and taxa.

Anchor Density Aggregation and Transformation

For each *Sarracenia* chromosome i and reference bin j, the raw anchor count is:

$$n_{ij}$$

To stabilize variance and reduce domination by anchor-rich bins, counts were transformed as:

$$A_{ij} = \log_{10}(n_{ij} + 1)$$

All subsequent analyses, including dominance, fractionation, path inference, and similarity matrices, operate exclusively on A_{ij}. In analyses where anchors were summarized across taxa rather than partitioned by subgenome state, transformed per-bin anchor signals were interpreted descriptively as measures of anchor presence and continuity along the reference coordinate system.

Syntenic Retention Metrics and Summary Statistics

Dominance Margin and Fractionation Bias

For each reference bin j, transformed anchor signals across chromosomes were ranked.

Let:

$$A_1 = \max_i A_{ij}$$
$$A_2 = \max_{i \neq i_1} A_{ij}$$

Where:

$$i_1 = \arg\max_i A_{ij}$$

then the dominance margin is:

$$DM_j = A_1 - A_2$$

Bins were classified as:

- dominant if $DM_j > 0$
- recessive if $DM_j \leq 0$

No magnitude cutoff beyond sign was applied.

Fractionation bias (see also[48]) was computed as:

$$FB_j = \frac{A_1 - \bar{A}_{\text{others}}}{A_1 + \bar{A}_{\text{others}}}$$

where $\bar{A}_{\text{others}}$ is the mean of all non-dominant transformed signals. Given $A_{ij} \geq 0$, then $0 \leq FB_j < 1$ whenever $A_1 \geq \bar{A}_{\text{others}}$.

Syntenic Retention Similarity Matrices

Each *Sarracenia* chromosome was represented as a vector:

$$\mathrm{v}_i = (A_{i1}, A_{i2}, \ldots, A_{iB})$$

Pairwise similarity was computed using correlation:

$$\rho_{ij} = \mathrm{corr}(\mathrm{v}_i, \mathrm{v}_j)$$

Bins with missing values were excluded pairwise. Similarities were assembled into square matrices and visualized as heatmaps.

Ks-Based Duplication Analyses

Ks Filtering and Mixture Decomposition

Ks values associated with syntenic anchors were filtered to retain:

$$0.05 \leq Ks \leq 2.0$$

Filtered Ks values were modeled using Gaussian mixture models[43,83] fitted by expectation-maximization.

Given Ks values $\{x_1, \ldots, x_n\}$, the model is:

$$p(x) = \sum_{k=1}^{K} \pi_k \ \mathcal{N}(x \mid \mu_k, \sigma_k^2),$$

where π_k denotes the mixture weight and $\mu_k$and σ_k represent the mean and standard deviation of component k. Mixture fitting was performed using expectation-maximization, and model complexity was chosen conservatively to capture interpretable duplication signal rather than maximize statistical fit.

Model selection was performed using Bayesian Information Criterion:

$$\mathrm{BIC} = -2\log L + p\log n$$

where p = number of estimated parameters (for 1D GMM with k components):

$$p = (k-1) + k + k = 3k - 1$$

Components were sorted by increasing $\mu_k$and treated strictly as relative age classes.

Event-Stratified Synteny Density

Each anchor was assigned to a Ks component.

For each component e, reference chromosome c, and bin b:

$$n_{e,c,b} \;=\; \sum_{r \in \mathcal{A}} 1\,(e_r = e \;\wedge\; c_r = c \;\wedge\; b_r = b)$$

where:

- r indexes syntenic anchor pairs
- $\mathcal{A}$ is the set of all anchors retained for analysis
- k_r is the Ks mixture component (event) assigned to anchor r
- c_r is the reference chromosome to which anchor $r$maps
- b_r is the reference-coordinate bin containing anchor r
- $1(\cdot)$ is the indicator function (1 if true, 0 otherwise)

The per-chromosome probability normalization, i.e. each chromosome's event-stratified density integrates to 1 across bins, is written as:

$$\tilde{n}_{e,c,b} = \frac{n_{e,c,b}}{\sum_{b'=1}^{B} n_{e,c,b'}}$$

where:

- $n_{e,c,b}$ is the raw event-stratified anchor count (as previously defined)
- $\tilde{n}_{e,c,b}$ is the normalized density
- B is the total number of bins on chromosome c
- the denominator sums only across bins on the same chromosome and event

As such, event-stratified anchor counts were normalized within each reference chromosome so that the summed density across bins equals one. These normalized densities were used for plotting age-stratified retention profiles.

Ks Analysis and Conservative Duplication Inference Across Ericales

Ks distributions were analyzed to characterize the temporal structure of gene duplication within *Sarracenia purpurea* and to place these patterns within a comparative Ericales context. Duplicate gene pairs were obtained from self:self CoGe SynMap analyses[39-41] for each species. Only gene pairs with defined chromosomal coordinates and valid Ks estimates were retained.

Ks values were filtered as above (see Ks Filtering and Mixture Decomposition**)** to reduce the influence of very recent noise at low divergence and extensive saturation at high divergence. Ks values outside this range were retained for visualization where informative but excluded from statistical modeling when dominated by multiple-hit effects. Filtered Ks values for each species were modeled using Gaussian mixture models, as in that same section above.

For all Ericales examined, the oldest and broadest Ks component was interpreted as retention of deeply diverged paralogs associated with the ancient core-eudicot *gamma* hexaploidy. This component was identified independently for each species and treated as a shared ancestral reference rather than evidence for recent duplication. Additional components at lower Ks were

considered candidate post-*gamma* duplication signals only when they were distinct from the *gamma* component and not dominated by saturation or diffuse background signal.

Several Ericales taxa, including *Actinidia* and *Vaccinium*, exhibit complex Ks profiles with prominent mid- to low-Ks peaks consistent with additional lineage-specific whole-genome duplications reported for these genera. The presence of these recent WGDs obscures separation between *gamma*-derived and post-*gamma* duplication signal and complicates direct cross-taxon comparison of younger Ks modes. Accordingly, *Actinidia* and *Vaccinium* were retained for syntenic anchoring and contextual visualization but excluded from comparative inference of post-*gamma* duplication timing.

Comparative timing analyses therefore focused on *Sarracenia purpurea*, *Rhododendron*, and *Impatiens*, which lack strong evidence for additional recent polyploidy and exhibit simpler Ks landscapes. For these taxa, a single younger Ks component was modeled in addition to the *gamma* component. Component means were treated strictly as relative temporal markers rather than discrete event boundaries.

To place Ks components on a comparative time axis, Ks values were converted to divergence time using:

$$t = \frac{Ks}{2\mu_{\text{eff}}},$$

where μ_{eff} is an effective long-term synonymous substitution rate. For each species, μ_{eff} was calibrated independently by constraining the mean of the *gamma* component to 120-125 million years ago, corresponding to widely inferred estimates for the timing of core-eudicot *gamma* hexaploidy[45-47]. Younger component ages were then scaled relative to this species-specific calibration.

Because Ks-based age estimates are sensitive to mixture model behavior, saturation effects, and the quality of genome assemblies and annotations, inferred duplication times are reported as approximate and are interpreted comparatively rather than as precise chronological estimates. No assumption of homology among younger duplication components across taxa was imposed.

Repeat Analyses

Repeat Divergence Landscape Analysis

Repeat divergence landscapes were generated from RepeatMasker annotations to provide a relative temporal summary of repeat accumulation in the *Sarracenia purpurea* genome. All RepeatMasker-annotated repeat fragments were treated as independent genomic intervals and analyzed without reference to gene models, syntenic structure, or chromosome identity.

For each annotated repeat fragment r, RepeatMasker reports a percent sequence divergence from the corresponding consensus element, reflecting accumulated substitutions since insertion relative to the library consensus. Let d_r denote this reported divergence value (in percent).

Repeat fragments were binned into integer divergence classes spanning 0-50% divergence. Formally, each fragment r was assigned to divergence bin:

$$b_r = \lfloor d_r \rfloor$$

such that:

$$0 \leq b_r \leq 50$$

For each divergence bin b, total base-pair coverage was computed as the sum of fragment lengths:

$$BP_b = \sum_r 1\,(\,b_r = b\,) \cdot \ell_r$$

where $\ell_r$is the genomic length of fragment r. These base-pair totals were then normalized by the total assembled genome size G to obtain bp-weighted divergence profiles:

$$F_b = \frac{BP_b}{G}$$

This procedure yields a genome-wide repeat divergence landscape in which each bin represents the fraction of the genome occupied by repeats with divergence values falling in that interval.

In addition to the genome-wide landscape, an LTR-restricted divergence landscape was constructed by applying the same binning and aggregation procedure to the subset of RepeatMasker fragments annotated as LTR-associated repeats. All normalization and bin definitions were identical between genome-wide and LTR-restricted analyses, enabling direct comparison of their divergence profiles.

For visualization, divergence landscape were plotted as histograms of F_b across divergence bins, with smoothed curves overlaid to highlight modal structure. The smoothed curves were used solely for visualization and did not alter the underlying bin counts. The modal divergence bin for each landscape was identified as the bin b with maximal F_b and indicated graphically as a reference.

Because RepeatMasker divergence values reflect sequence divergence from consensus rather than absolute insertion times, divergence landscapes were interpreted as relative summaries of repeat accumulation dynamics. Therefore, differences between genome-wide and LTR-restricted profiles were used to assess whether LTR retrotransposons contribute disproportionately to temporally structured components of repeat accumulation.

Comparative Analysis of Ks-Based Duplication Signals and LTR Divergence Landscapes

To examine the temporal relationship between gene duplication and transposable-element accumulation within the *Sarracenia purpurea* genome, we integrated Ks-based duplication age estimates with RepeatMasker-derived divergence landscapes for LTR retrotransposons. Ks values were obtained from self-syntenic duplicate gene pairs, and in this case, filtered to the range $0.05 \leq Ks \leq 2.0$ to minimize the influence of saturation at high divergence.

Filtered Ks values were modeled using a two-component Gaussian mixture model, representing a younger lineage-specific duplication signal and a broader, older component corresponding to deeply diverged paralogs. The latter component is interpreted as reflecting retention from the ancient core-eudicot *gamma* hexaploidy event, while the younger component captures post-*gamma* duplication retained within the *Sarracenia* lineage. Mixture fitting was performed as in section 4.5.1.

LTR retrotransposon divergence landscapes were constructed by parsing RepeatMasker annotations to extract all LTR-associated repeat fragments. For each fragment, the reported percent divergence from its consensus sequence was binned into integer divergence classes, and base-pair coverage was summed per bin to generate bp-weighted divergence profiles.

To enable direct visual comparison between duplication and repeat accumulation, Ks values and LTR divergence bins were transformed to a common time axis using the relationship:

Where:

$$d_{\text{frac}} = \frac{d_{\%}}{100}$$

and then:

$$t = \frac{d_{\text{frac}}}{2\mu_{\text{eff}}}$$

and d denotes Ks or fractional divergence (see also Ks Analysis and Conservative Duplication Inference Across Ericales, above). An effective long-term substitution rate μ_{eff} was calibrated by constraining the mean of the older Ks component to 120-125 million years ago, corresponding to the widely inferred timing of the core-eudicot *gamma* hexaploidy event. All time estimates are therefore calibration-dependent and interpreted comparatively. Ks density curves and LTR divergence histograms were plotted on aligned time axes with independent y-axis normalization to emphasize relative temporal structure rather than absolute magnitude.

Quantification of Transposable-Element Heterogeneity Across Pseudochromosomes

To quantify variation in transposable-element (TE) distribution across the *Sarracenia purpurea* genome, RepeatMasker annotations were summarized at the pseudochromosome scale using union base-pair coverage. All RepeatMasker-annotated intervals were parsed by pseudochromosome and merged prior to summation to avoid double-counting overlapping fragments. Total TE coverage was calculated as the number of base pairs covered by any RepeatMasker annotation (all TE classes combined) divided by the total length of each pseudochromosome.

LTR retrotransposon coverage was calculated analogously by restricting annotations to LTR-tagged elements, including Gypsy and Copia families, and computing union base-pair coverage relative to pseudochromosome length. Thus, heterogeneity was quantified separately for (i) overall TE occupancy and (ii) LTR-specific occupancy, enabling direct comparison between genome-wide repeat content and the dominant repeat class.

Inter-pseudochromosome variation was summarized using descriptive statistics across the 13 pseudochromosomes. For each metric, the range (Δ) was defined as the difference between the maximum and minimum coverage values, expressed in percentage points. The coefficient of variation (CV) was calculated as the standard deviation divided by the mean coverage and is reported as a percentage, providing a normalized measure of heterogeneity that accounts for differences in mean repeat abundance. These statistics were used to support statements regarding scaffold-level variation in repeat content and to contextualize the spatial distribution of LTR retrotransposons relative to overall TE composition.

Fragment counts per pseudochromosome were recorded as a measure of annotation density but were not used for coverage estimation. All reported percentages reflect union coverage and therefore represent conservative estimates of repeat occupancy.

Comparative Functional Genomics (Occupancy Classes, Foregrounds, GOs)

Homolog Occupancy Class Construction Using Strict Intersection Logic

Protein-coding genes were first mapped to *Arabidopsis thaliana* ortholog identifiers, providing a common reference framework for cross-taxon comparison. Homolog occupancy classes were then constructed using a strict intersection logic that encodes shared evolutionary presence patterns across a fixed set of taxa.

For a fixed set of taxa:

- For each *Arabidopsis* identifier, construct a binary presence-absence vector.
- Group identifiers with identical vectors.
- Each group defines a strict intersection occupancy class.

This procedure partitions the full gene set into equivalence classes defined solely by shared taxonomic presence patterns. These classes form the atomic units for all downstream functional analyses.

Properties:

- Each gene belongs to exactly one occupancy class.
- Occupancy classes are mutually exclusive.
- These classes are never modified downstream.

This binning step is intentionally conservative and is the only stage in the analysis where intersections across taxa are permitted. By fixing occupancy class membership at this stage, all subsequent analyses inherit a stable, reproducible partitioning of genes that is independent of downstream biological hypotheses.

This is the only stage where intersections are used.

Foreground Construction as Unions of Occupancy Classes

All downstream gene sets used for functional analysis were defined strictly as unions of the pre-defined intersection occupancy classes described above. No additional filtering, subdivision, or intersection operations were applied once these classes were established.

Explicitly:

- No genes were intersected across occupancy classes.
- No classes were intersected with other classes.
- Foregrounds are deterministic unions defined by occupancy class membership.

This union-based strategy ensures that foregrounds remain internally consistent, reproducible, and traceable back to the original occupancy class definitions. It also prevents inflation of statistical signal through repeated reuse or reshaping of overlapping gene sets.

This union-based foreground construction applies uniformly to:

- duplication-mode analyses
- subgenome compartment analyses
- lineage-specific gene sets

By enforcing this structure, functional comparisons across different biological questions (e.g., duplication mode versus lineage specificity) remain directly comparable and free from hidden dependencies.

GO Enrichment

Gene Ontology (GO) enrichment analyses were performed on foregrounds defined as unions of occupancy classes, as described above. Enrichment testing was conducted using a standard contingency-table framework, with careful separation of genome-derived and expression-derived analyses.

For each GO term:

- Count foreground genes annotated to the term.
- Count background genes annotated to the term.
- Perform Fisher's exact test.

The background consisted of all *Arabidopsis* identifiers eligible for occupancy class assignment, ensuring that enrichment tests were conditioned on the same orthology universe used to define classes.

Genome-derived enrichments used Bonferroni correction:

$$p_{\text{adj}} = \min\left(p \times N_{\text{tests}}, 1\right)$$

Expression-derived enrichments used Benjamini-Hochberg false discovery rate control, reflecting the higher degree of dependency expected among expression-based gene sets.

To avoid conflating broadly conserved Ericales functions with lineage-specific signal, GO terms enriched in *Actinidia*-containing foregrounds were identified separately and removed from *Sarracenia*-focused interpretations at the interpretation stage only. This filtering step was applied post hoc and did not alter the underlying enrichment calculations.

RNA-seq Processing and Expression Analyses

RNA Sequencing and Sample Design

RNA sequencing was performed on three tissue types: roots, trap leaves, and flowers. Three biological replicates were collected for each tissue type, yielding a total of nine RNA-seq samples. Samples were designated as follows: SPR1-SPR3 (roots), SPL1-SPL3 (trap leaves), and SPF1-SPF3 (flowers). Flower samples were harvested after petal abscission to capture post-floral developmental expression states.

Sarracenia purpurea



RNA-seq libraries were sequenced on an Illumina platform using paired-end reads. All RNA-seq data have been deposited under BioProject PRJDB15737. Detailed accession information for each sample is provided in section 4.9.

RNA-seq Samples and Accession Metadata

TISSUE	SAMPLE	BIOSAMPLE	EXPERIMENT	RUN	LIBRARY LAYOUT	READ FILES
FLOWER	SPF1	SAMD00754511	DRX522175	DRR538356	Paired-end	SPF1_1.fq.gz, SPF1_2.fq.gz
FLOWER	SPF2	SAMD00754512	DRX522176	DRR538357	Paired-end	SPF2_1.fq.gz, SPF2_2.fq.gz
FLOWER	SPF3	SAMD00754513	DRX522177	DRR538358	Paired-end	SPF4_1.fq.gz, SPF4_2.fq.gz
LEAF	SPL1	SAMD00754514	DRX522178	DRR538359	Paired-end	SPL1_1.fq.gz, SPL1_2.fq.gz
LEAF	SPL2	SAMD00754515	DRX522179	DRR538360	Paired-end	SPL2_1.fq.gz, SPL2_2.fq.gz
LEAF	SPL3	SAMD00754516	DRX522180	DRR538361	Paired-end	SPL3_1.fq.gz, SPL3_2.fq.gz
ROOT	SPR1	SAMD00754517	DRX522181	DRR538362	Paired-end	SPR1_1.fq.gz, SPR1_2.fq.gz
ROOT	SPR2	SAMD00754518	DRX522182	DRR538363	Paired-end	SPR2_1.fq.gz, SPR2_2.fq.gz
ROOT	SPR3	SAMD00754519	DRX522183	DRR538364	Paired-end	SPR3_1.fq.gz, SPR3_2.fq.gz

RNA-seq Read Mapping and Transcriptome Assembly

Genome indexing for RNA-seq read alignment was performed using HISAT2 v2.2.1, incorporating splice-site and exon information derived from the final EVM-integrated gene

annotation. Splice sites were extracted from the GFF3 annotation using *hisat2_extract_splice_sites.py*. Exon coordinates were obtained by converting the GFF3 file to GTF format using gffread v0.12.7, followed by *hisat2_extract_exons.py*.

The RagTag-scaffolded genome assembly was indexed using:

hisat2-build --ss Spur.ss --exon Spur.exons Spur.fasta --large-index -p 16 Spur_index

RNA-seq reads were aligned to the indexed genome using HISAT2 with parameters optimized for transcriptome assembly:

hisat2 --dta -p 16 -x Spur_index \

-1 SAMPLE_R1.fastq.gz -2 SAMPLE_R2.fastq.gz \

-S SAMPLE.sam

The --dta option was used to report alignments suitable for transcript assemblers. SAM files were converted to BAM format and sorted using SAMtools v1.16.1.

Transcript assembly and expression estimation were performed independently for each sample using StringTie v2.2.1:

stringtie SAMPLE.sorted.bam -G reference.gtf -e -B -o SAMPLE.gtf

The -e option restricted assembly to reference transcripts, and the -B option generated tabular output files required for downstream differential expression analyses.

Differential Expression Analysis

Gene- and transcript-level count matrices were generated using *prepDE.py* following the StringTie workflow. This procedure produced a transcript count matrix (*transcript_count_matrix.csv*) summarizing expression counts for each annotated gene across all samples. These matrices were used for downstream differential expression analyses described in the Results.

Expression Responsiveness of Syntenic Versus Tandem Duplicates

Expression responsiveness of duplicated genes in *Sarracenia purpurea* was compared between syntenically retained duplicates and tandem duplicates. Syntenic duplicates were defined as paralogous gene pairs retained within self-syntenic blocks identified through the Ericales anchoring framework, whereas tandem duplicates were defined as adjacent or near-adjacent paralogs identified using a CDS-based tandem detection pipeline. Genes classified as mixed or ambiguous with respect to duplication mode were excluded from this analysis.

Gene expression responsiveness was quantified using pitcher-associated transcriptomic datasets. For each gene, expression change was calculated as the absolute value of the log-transformed fold change across experimental contrasts, such that higher values reflect greater transcriptional responsiveness across pitcher-related conditions. This metric captures both up- and down-regulation while avoiding cancellation effects.

Distributions of expression responsiveness for syntenic and tandem gene sets were visualized using combined violin and box plots. Box plots indicate the median and interquartile range, while violins represent kernel density estimates of the full distribution. Differences between duplication classes were assessed using a two-sided Mann-Whitney U test, which does not

assume normality. Effect size was quantified using Cliff's delta (δ), which measures the probability that a randomly selected value from one group exceeds a randomly selected value from the other group, minus the reverse probability:

$$\delta = P(X > Y) - P(Y > X)$$

where X and Y denote observations drawn from the two duplication classes. Sample sizes for each group were retained as observed and are reported directly on **Figure 14**. All statistical tests were conducted using gene-level values without subsampling.

Functional Breadth Analysis Using GO Rarefaction

Functional breadth of syntenic and tandem gene sets was evaluated using Gene Ontology (GO) biological-process annotations under a rarefaction framework. GO annotations were derived from the curated genome annotation of *Sarracenia purpurea*, and genes lacking GO biological-process terms were excluded from this analysis.

For each duplication class, genes were repeatedly subsampled without replacement across a range of sampling depths. At each depth, two complementary measures of functional breadth were calculated. First, GO richness was defined as the number of unique GO biological-process terms represented in the subsample. Second, effective functional breadth ($D1$[84]) was calculated as the exponential of Shannon entropy (cf. [81]):

$$H = -\sum_{i=1}^{S} p_i \ln p_i$$

$$D_1 = e^H$$

where p_i is the proportional frequency of genes annotated to GO term i, and S is the total number of GO terms observed in the subsample. This metric captures both the diversity and evenness of functional annotations.

For each sampling depth, mean values and confidence envelopes were estimated from the distribution of subsampled values across iterations. Because rarefaction curves represent non-independent function-valued data, no formal hypothesis testing was applied across entire curves. Instead, rarefaction was used as a descriptive framework to visualize differences in functional breadth between duplication classes while controlling for gene set size.

Data Availability

The genome assembly and annotation of *Sarracenia purpurea* is publicly available at https://genomevolution.org/coge/GenomeInfo.pl?gid=69349. RNA-seq data are available at NCBI BioProject PRJDB15737. All other data can be requested from the corresponding author (Victor A. Albert; vicsgenomics@gmail.com) and will also be made available upon formal journal publication.

Acknowledgments

We acknowledge the following sources for funding: the U.S. National Science Foundation grants 2030871 and 2139311 (to VAA and CL, respectively), the Sofja Kovalevskaja Program of the Alexander von Humboldt Foundation (to KF), and JSPS KAKENHI 23K20050 (to KF).

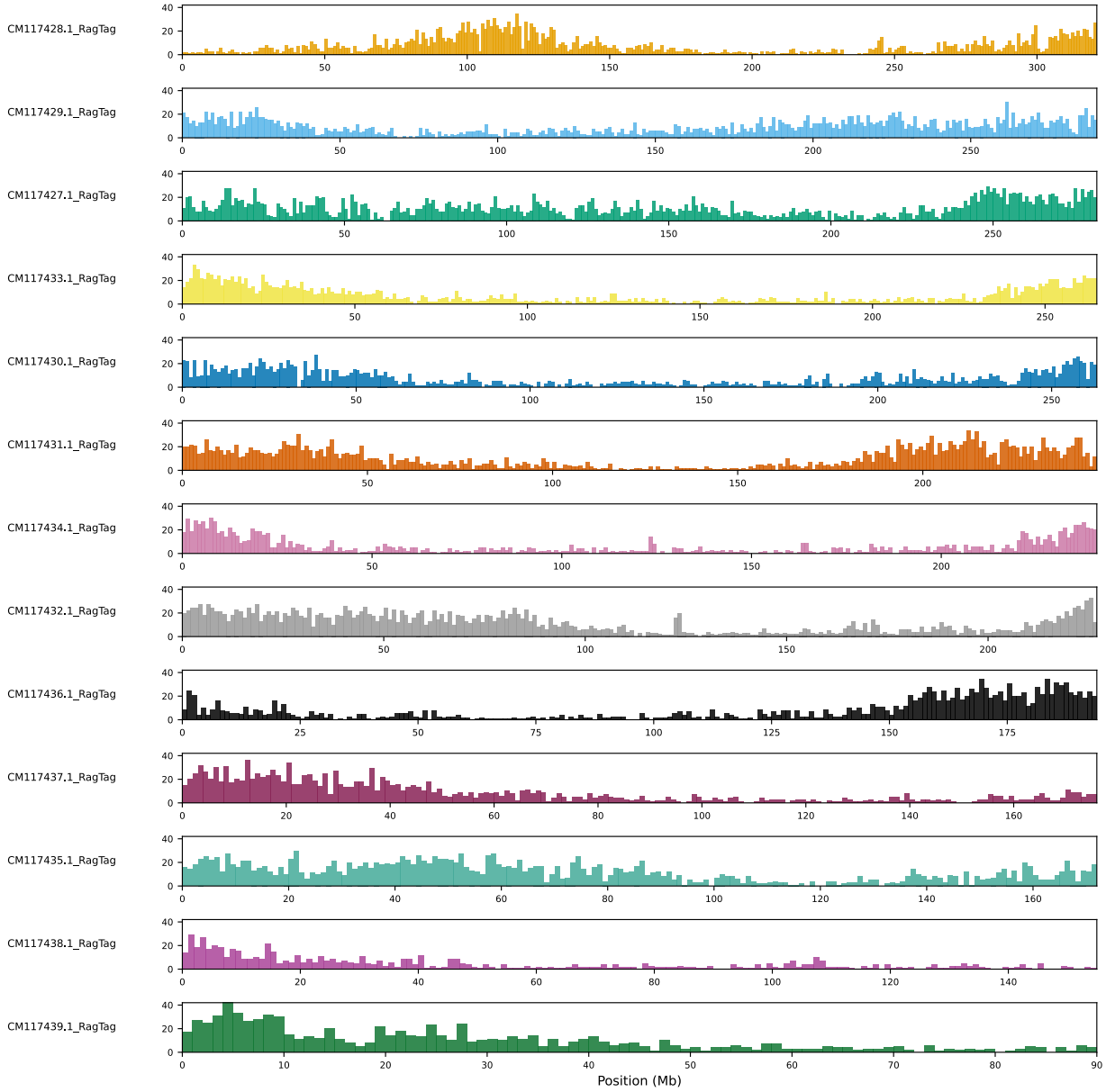
Author Contributions

VAA conceived the study; VAA led the research with assistance from JK; VAA, JK, MR, and CL provided materials; MR sequenced Oxford Nanopore reads; KF and MF provided RNA-seq data; CP, JK, and NP assembled and annotated the genome; CP decontaminated and reference-scaffolded the assembly; CP analyzed RNA-seq data; VAA, JK, CP, NP, and JM performed additional downstream bioinformatic analyses; VAA, CP, and JM produced graphics; VAA wrote the manuscript, and all authors approved the final version.

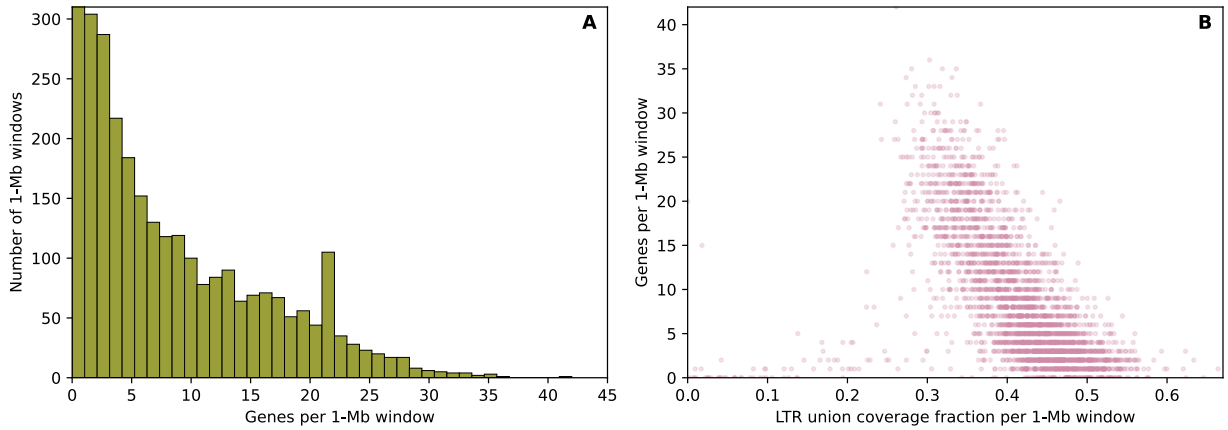
Disclosure

ChatGPT was used to help develop bioinformatics approaches, generate analysis scripts and graphics, and to vet and polish the text.

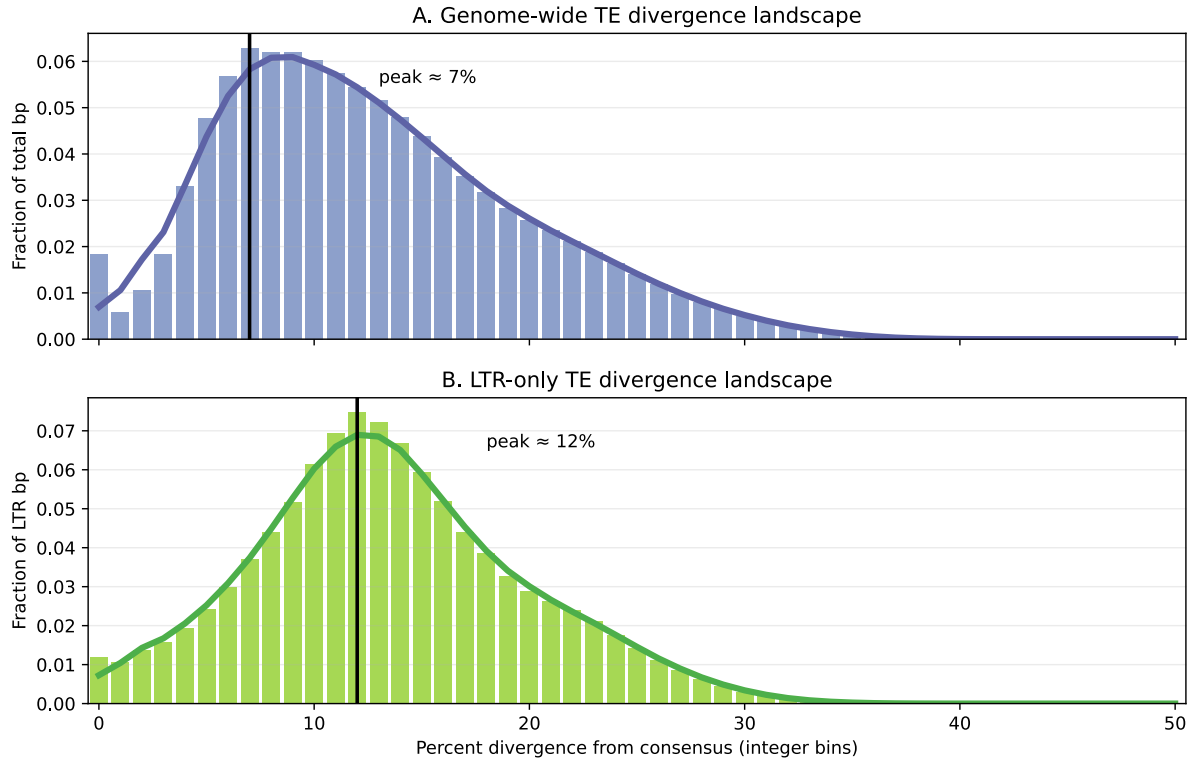
Extended Data Figures



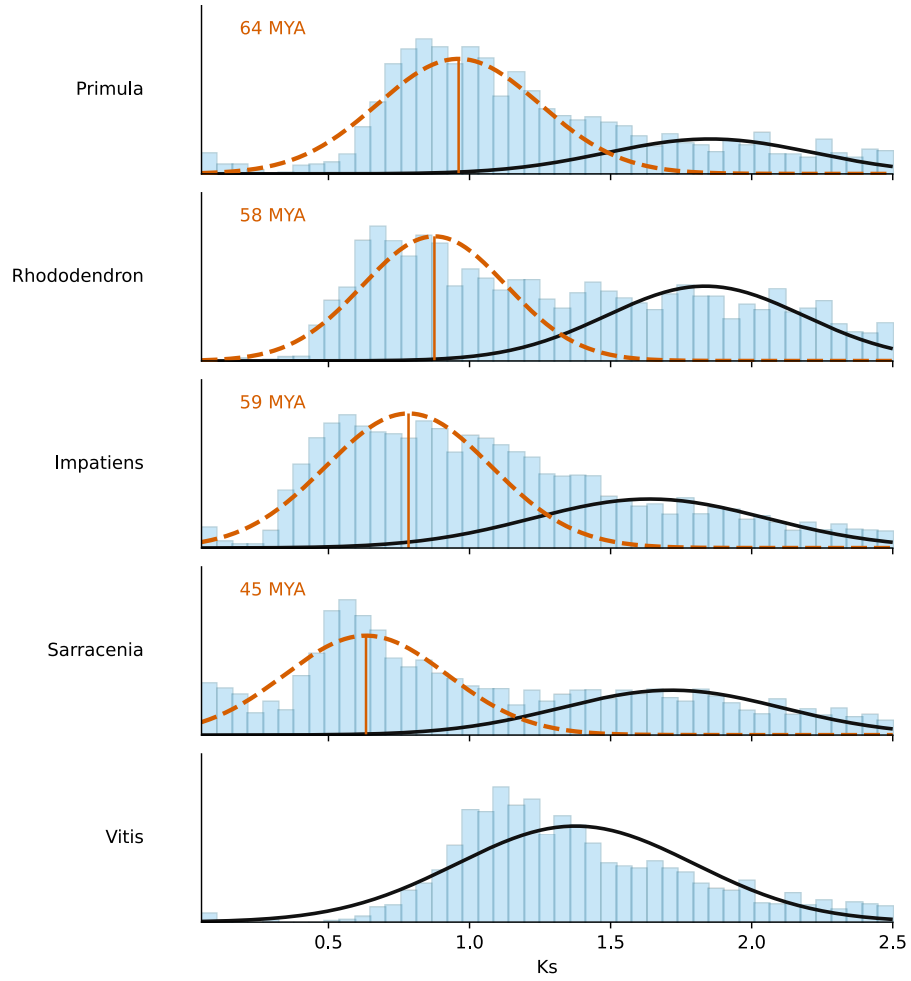
Extended Data Figure 1. Chromosome-scale gene-density variation. Gene counts per 1-Mb window across the 13 pseudochromosomes highlight extended gene-poor regions and localized peaks of elevated density.



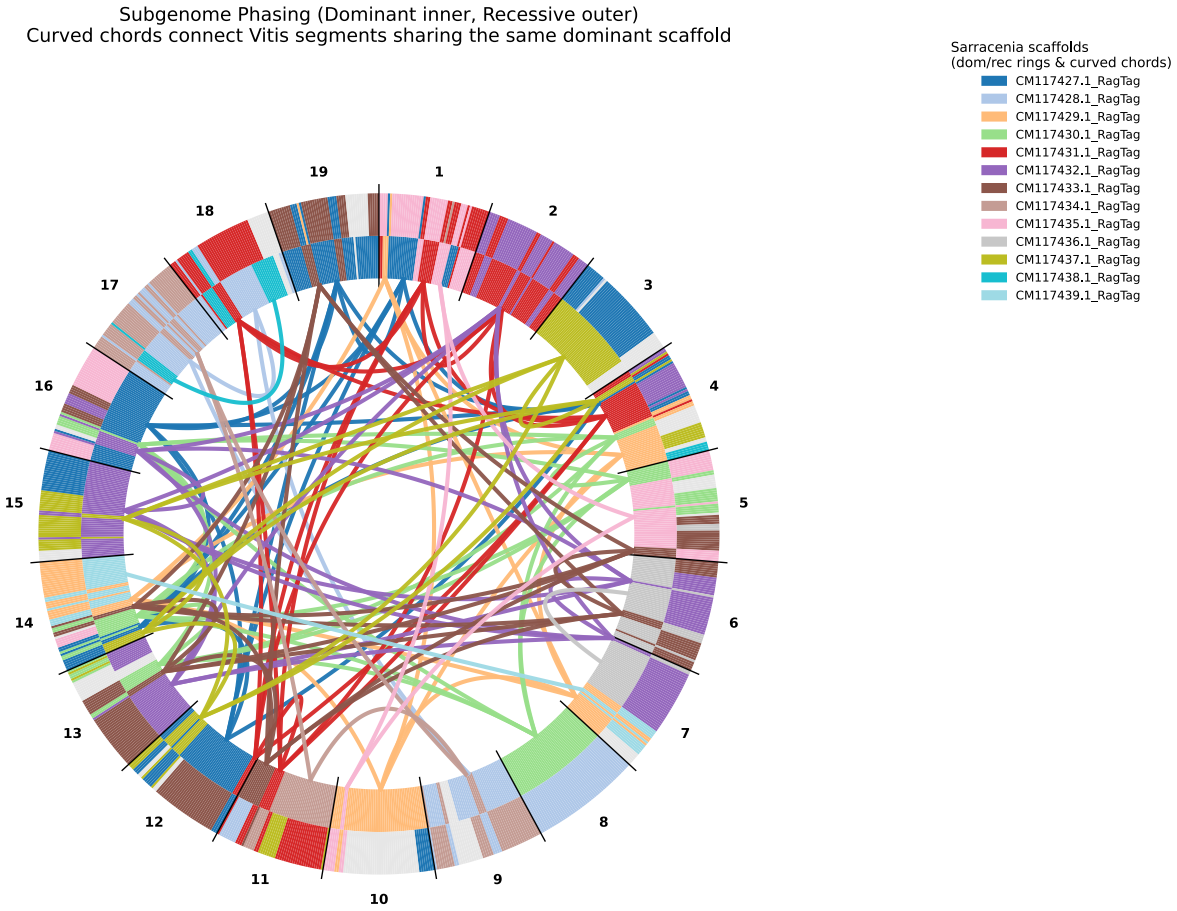
Extended Data Figure 2. Genome-wide gene density and LTR coverage. (**A**) Histogram of genes per 1-Mb window. (**B**) Gene density versus LTR union coverage per window, showing that LTR-rich windows tend to be gene-poor.



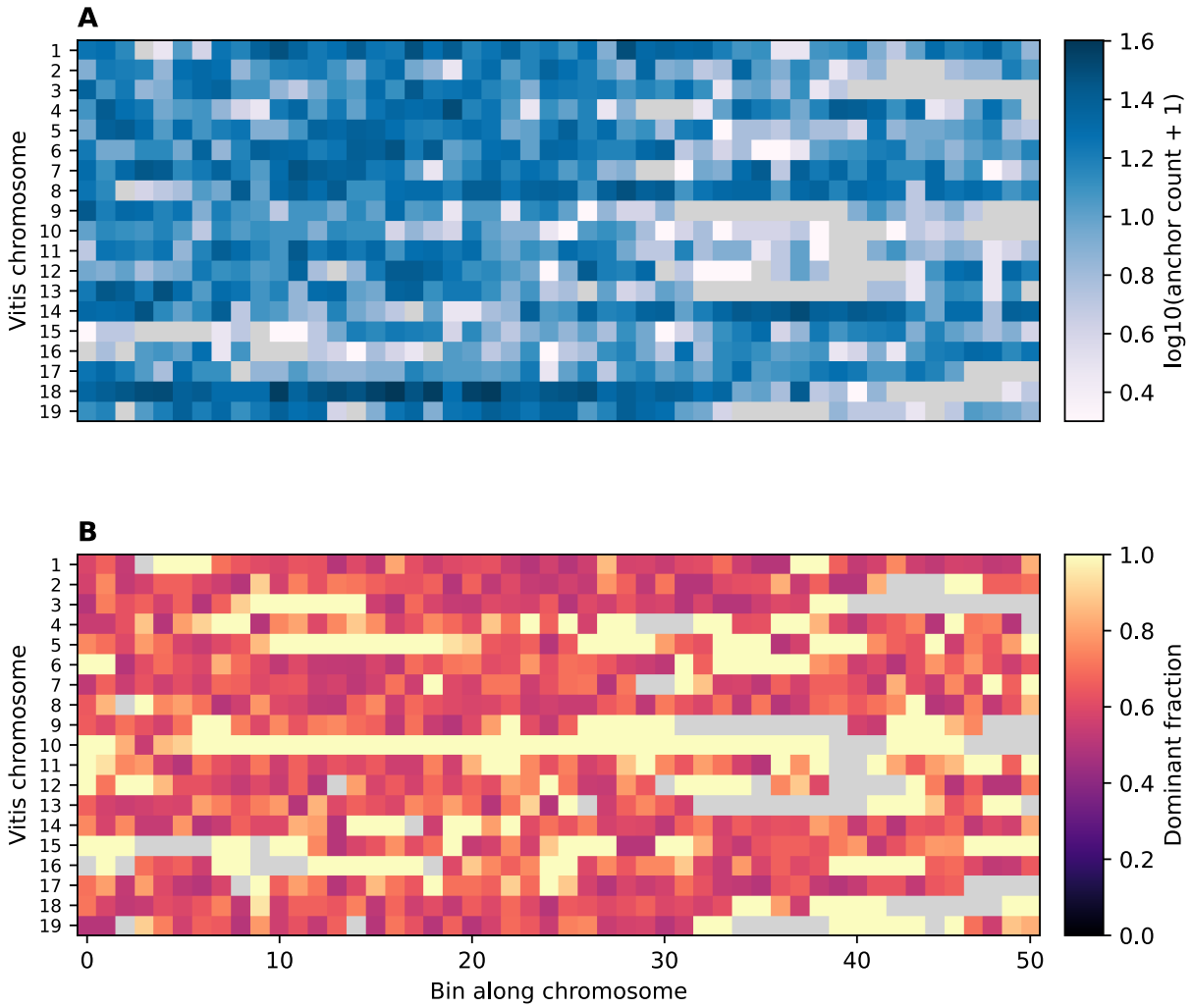
Extended Data Figure 3. Repeat divergence landscapes for *Sarracenia purpurea* repeats (genome-wide versus LTR-only). (**A**) Genome-wide and (**B**) LTR-only bp-weighted divergence distributions derived from RepeatMasker annotations; vertical lines mark modal divergence bins. For each RepeatMasker-annotated repeat fragment, the reported percent divergence from its consensus was binned into integer divergence bins (0-50%), and base-pair coverage was summed per bin to generate bp-weighted divergence profiles. Histograms show the fraction of total covered bp contributed by each divergence bin (panel-specific normalization), and the overlaid smoothed curves summarize the same distributions to highlight modal structure. Vertical reference lines mark the peak (modal) divergence bin for each panel.



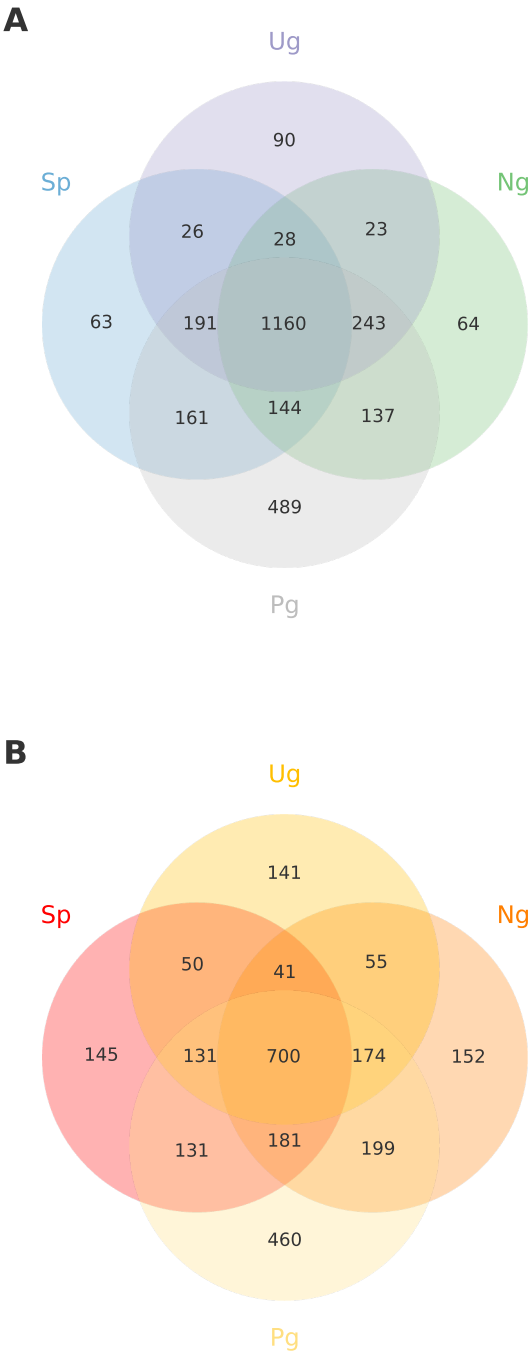
Extended Data Figure 4. Ks distributions across selected Ericales. Ks densities for self-syntenic duplicate pairs in *Sarracenia purpurea*, *Rhododendron*, *Impatiens*, *Primula*, and *Vitis vinifera* (0.05–2.5). Solid curves represent the *gamma*-associated component; dashed curves indicate younger inferred components. Modal Ks values of younger components are independently calibrated (and dated) relative to *gamma*.



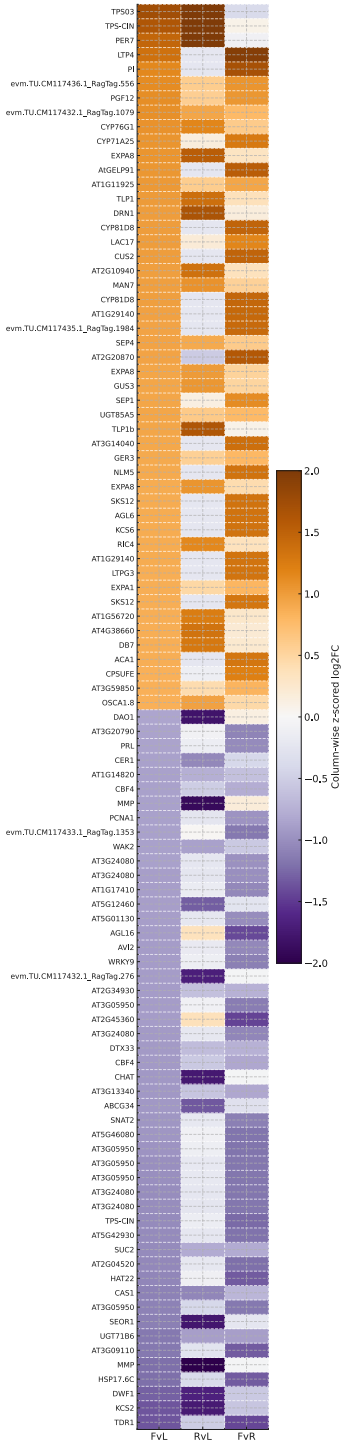
Extended Data Figure 5. Mosaic subgenome structure of *Sarracenia* relative to Vitis. Circular representation of dominant and recessive segments mapped onto *Vitis* chromosomes. Interleaved colored blocks illustrate non-one-to-one correspondence and rearranged relationships among ancestral segments. Dominant (inner ring) and recessive subgenome (outer ring) predictions based on fractionation patterns. Central chords connect *Vitis* segments sharing the same dominant *Sarracenia* scaffold.



Extended Data Figure 6. Blockwise retention heatmaps for *Sarracenia-Vitis* synteny. (**A**) Log_{10}-scaled syntenic anchor counts across 50 bins per *Vitis* chromosome. Alternating high- and low-retention domains reveal tightly interwoven conserved and fractionated regions. (**B**) Fraction of dominant subgenome anchors per bin across the same chromosomes. Warm hues denote dominant-trending bins and cooler hues recessive bins; interspersed color blocks demonstrate alternating dominant and recessive segments. Gray cells indicate absence of anchors.



Extended Data Figure 7. Enriched GO Biological Process categories across carnivorous lineages. Venn diagrams summarize significantly enriched GO BP terms derived from duplicate-gene foregrounds in *Sarracenia purpurea* (Sp), *Nepenthes gracilis* (Ng), *Utricularia gibba* (Ug), and *Pinguicula vulgaris* (Pg). (**A**) Enrichment among syntenically retained duplicates. (**B**) Enrichment among tandem duplicates. Counts indicate GO BP terms passing identical statistical thresholds within each lineage and their intersections.



Extended Data Figure 8. Differential expression heatmap for flowers versus pitcher leaves. Heatmap of $\log_2$ fold-change values (flowers vs. pitcher leaves, FvL) for top DEGs in *Sarracenia purpurea*. Positive values indicate flower-biased expression; negative values indicate trap-biased expression. Genes are ordered by FvL magnitude. For each gene, the best-supported *Arabidopsis* ortholog is shown by symbol or locus ID; *Sarracenia* IDs appear only when no homolog was assigned (**Table SI**). Floral regulators (*PI*, *SEP1*, *SEP4*), terpene synthases (*TPS03*, *TPS-CIN*), peroxidases (*PER7*), and lipid-transfer proteins (*LTP4*) dominate the flower-up spectrum. Trap-enriched genes include stress-response regulators, cell-wall modifiers, hormone-associated loci, and carnivore-associated homologs (**Tables SH and SC**).

Supplementary Tables

METRIC	VALUE
NUMBER OF GENE LOCI	28,835
NUMBER OF TRANSCRIPTS	28,835
TOTAL CDS FEATURES	132,001
MEAN EXON LENGTH (BP)	263.12
MEDIAN EXON LENGTH (BP)	139
MEAN GENE LENGTH (BP)	9,315.68
MEDIAN GENE LENGTH (BP)	3,112
MEAN TRANSCRIPT LENGTH (BP)	9,315.68
MEDIAN TRANSCRIPT LENGTH (BP)	3,112
MEAN EXON COUNT PER GENE	4.58
MEDIAN EXON COUNT PER GENE	3
NUMBER OF SINGLE-EXON GENES	7,179
MEAN INTRON LENGTH (BP)	2,267.08
MEDIAN INTRON LENGTH (BP)	558
GENOME-WIDE GENE DENSITY (GENES/MB)	8.18
TOTAL ASSEMBLY LENGTH (BP)	3,523,999,544

Table S1. Summary of gene and transcript structural features in the *Sarracenia purpurea* genome. Key annotation metrics for the primary gene set, including counts and length distributions for genes, transcripts, coding sequences, exons, and introns. Values summarize overall genome architecture, highlighting a moderately exon-rich gene space, substantial intron-length variation, and a genome-wide gene density of about 8 genes per megabase across the 3.52 Gb assembly. Each gene locus was represented by a single representative transcript model in the final annotation.

REPEAT CATEGORY	NUMBER OF ELEMENTS	BASES OCCUPIED (BP)	% OF GENOME
TOTAL MASKED SEQUENCE	-	3,099,995,857	87.97
TOTAL INTERSPERSED REPEATS	-	3,051,425,825	86.59
RETROELEMENTS (TOTAL)	3,018,215	2,578,482,566	73.17
SINES	8,757	9,815,940	0.28
LINES	63,405	38,409,777	1.09
LTR RETROTRANSPOSONS (TOTAL)	2,946,053	2,530,256,849	71.80
TY1/COPIA	1,143,498	1,176,556,747	33.39
GYPSY/DIRS1	555,662	721,463,009	20.47
RETROVIRAL	16,784	10,010,854	0.28
DNA TRANSPOSONS (TOTAL)	139,026	36,995,745	1.05
HOBO-ACTIVATOR	20,275	11,713,441	0.33
MULE-MUDR	9,274	6,885,918	0.20
OTHER DNA TRANSPOSONS	-	18,396,386	0.52
ROLLING-CIRCLE ELEMENTS	7,995	2,565,999	0.07
UNCLASSIFIED INTERSPERSED REPEATS	1,508,633	435,947,514	12.37
SMALL RNA	11,922	25,821,641	0.73
SIMPLE REPEATS	481,899	21,756,034	0.62
LOW-COMPLEXITY SEQUENCE	53,438	2,867,746	0.08
SATELLITES	264	79,125	<0.01

Table S2. Genome-wide distribution of transposable element classes and other masked repeat categories in the *Sarracenia purpurea* assembly. Summary of RepeatMasker annotations showing the proportional contributions of major TE groups, including LTR retrotransposons, LINEs, SINEs, DNA transposons, rolling-circle elements, and unclassified repeats. The table also reports ancillary repeat categories such as small RNAs, simple repeats, and low-complexity regions, providing an overview of the repeat landscape that shapes genome size, structure, and gene-density variation in *Sarracenia purpurea*.

FUNCTIONAL CATEGORY	REPRESENTATIVE GO BIOLOGICAL PROCESS	FOREGROUND CLASS	ACTINIDIA FILTER	*ARABIDOPSIS* GENE(S)	DUPLICATION MODE(S)
DETOXIFICATION / XENOBIOTIC TRANSPORT	Response to toxic substance (GO:0009636)	Carnivore-union	Shared with *Actinidia*	AT1G01580 (*PEN3*-like ABC transporter)	Tandem, syntenic
GLUTATHIONE-BASED REDOX REGULATION	Glutathione metabolic process (GO:0006749)	Carnivore-union	Shared with *Actinidia*	AT2G16390 (*GSTF* family)	Tandem
ANTIFUNGAL DEFENSE	Defense response to fungus (GO:0050832)	*Sarracenia*-only	Absent from *Actinidia*	AT2G26560	Tandem
WOUND / STRESS RESPONSE	Response to wounding (GO:0009611)	*Sarracenia*-only	Absent from *Actinidia*	AT1G24140	Tandem, syntenic
CUTICLE BIOSYNTHESIS / REMODELING	Cuticle development (GO:0042335)	Carnivore-union	Shared with *Actinidia*	*CYP86*- and *LACS2*-associated genes	Tandem
ETHYLENE-ACTIVATED SIGNALING	Ethylene-activated signaling pathway (GO:0009873)	*Sarracenia*-only	Absent from *Actinidia*	AT-annotated regulators	Mixed
CELL DIFFERENTIATION	Cell differentiation (GO:0030154)	*Sarracenia*-only	Absent from *Actinidia*	AT-annotated regulators	Mixed
ROOT-HAIR CELL DIFFERENTIATION	Root hair cell differentiation (GO:0048767)	*Sarracenia*-only	Absent from *Actinidia*	AT-annotated regulators	Mixed
MITOCHONDRIAL RNA MODIFICATION	Mitochondrial mRNA modification (GO:0042788)	*Sarracenia*-only	Absent from *Actinidia*	AT-annotated factors	Mixed
CHROMATIN-LEVEL REGULATION	DNA demethylation (GO:0080111)	*Sarracenia*-only	Absent from *Actinidia*	AT2G36490 (*DML1*)	Tandem

Table S3. *Sarracenia*-specific functional innovation identified by comparative foreground analyses. This table summarizes biological-process categories and representative genes associated with *Sarracenia*-specific and *Sarracenia*-enriched innovation foregrounds identified through comparative gene-set analyses across carnivorous plants and close Ericales relatives. Functional categories shown derive from gene sets present in *Sarracenia* but absent from, or explicitly filtered relative to, *Actinidia*, as well as from broader carnivore-union foregrounds. Representative *Arabidopsis* gene identifiers are provided to facilitate functional interpretation, with corresponding *Sarracenia* homologs listed where available. Duplication-mode annotations indicate whether genes occur among tandem-derived and/or syntenically retained loci; no enrichment, dominance, or causal relationship is implied. Rows reflect curated extractions from previously generated innovation and GO summary tables, and no additional statistical analyses were performed.

Tables SA-SI. See external Zip file: Tables_SA-SI.zip

References Cited

1. Darwin, C.R. *Insectivorous Plants*, (John Murray, London, 1875).
2. Ellison, A.M. & Adamec, L. *Carnivorous plants: physiology, ecology, and evolution*, (Oxford University Press, 2018).
3. Naczi, R.F. Systematics and evolution of Sarraceniaceae. *Carnivorous plants: Physiology, ecology, and evolution* **9**, 105-119 (2018).
4. Gray, S.M., Akob, D.M., Green, S.J. & Kostka, J.E. The bacterial composition within the Sarracenia purpurea model system: local scale differences and the relationship with the other members of the food web. *PLoS One* **7**, e50969 (2012).
5. Hale, R.E., Powell, E., Beikmohamadi, L. & Alexander, M.L. Effects of arthropod inquilines on growth and reproductive effort among metacommunities of the purple pitcher plant (Sarracenia purpurea var. montana). *Plos one* **15**, e0232835 (2020).
6. Grothjan, J.J. & Young, E.B. Diverse microbial communities hosted by the model carnivorous pitcher plant Sarracenia purpurea: analysis of both bacterial and eukaryotic composition across distinct host plant populations. *PeerJ* **7**, e6392 (2019).
7. Miner, J.A. & Taylor, P.D. Effects of peatland size and exposure on two species of Diptera inhabiting the pitcher plant Sarracenia purpurea L. *Ecoscience* **9**, 347-354 (2002).
8. Freedman, Z.B. *et al.* Environment–host–microbial interactions shape the Sarracenia purpurea microbiome at the continental scale. (Wiley Online Library, 2021).
9. Bittleston, L.S., Gralka, M., Leventhal, G.E., Mizrahi, I. & Cordero, O.X. Context-dependent dynamics lead to the assembly of functionally distinct microbial communities. *Nature Communications* **11**, 1440 (2020).
10. Grothjan, J.J. & Young, E.B. Bacterial Recruitment to Carnivorous Pitcher Plant Communities: Identifying Sources Influencing Plant Microbiome Composition and Function. *Front Microbiol* **13**, 791079 (2022).
11. Armitage, D.W. Linking the development and functioning of a carnivorous pitcher plant's microbial digestive community. *Isme j* **11**, 2439-2451 (2017).
12. Cai, J. *et al.* The microbiome and metatranscriptome of a panel from the Sarracenia mapping population reveal complex assembly and function involving host influence. *Front Plant Sci* **15**, 1445713 (2024).
13. Kocáb, O. *et al.* Enzyme activities in two sister-species of carnivorous pitcher plants (Nepenthes) with contrasting nutrient sequestration strategies. *Plant Physiology and Biochemistry* **161**, 113-121 (2021).
14. Ravee, R., Baharin, A., Cho, W.T., Ting, T.Y. & Goh, H.H. Protease activity is maintained in Nepenthes ampullaria digestive fluids depleted of endogenous proteins with compositional changes. *Physiologia Plantarum* **173**, 1967-1978 (2021).
15. Freund, M. *et al.* The digestive systems of carnivorous plants. *Plant physiology* **190**, 44-59 (2022).
16. Thorogood, C.J., Bauer, U. & Hiscock, S.J. Convergent and divergent evolution in carnivorous pitcher plant traps. *New Phytologist* **217**, 1035-1041 (2018).
17. Fukushima, K. *et al.* Oriented cell division shapes carnivorous pitcher leaves of Sarracenia purpurea. *Nature Communications* **6**, 6450 (2015).
18. Segala, M.C. & Horner, J.D. The effects of light availability, prey capture, and their interaction on pitcher plant morphology. *Plant Ecology* **224**, 539-548 (2023).
19. Baldwin, E., McNair, M. & Leebens-Mack, J. Rampant chloroplast capture in Sarracenia revealed by plastome phylogeny. *Frontiers in Plant Science* **14**, 1237749 (2023).

20. Stephens, J.D. *et al.* Resolving phylogenetic relationships of the recently radiated carnivorous plant genus Sarracenia using target enrichment. *Molecular Phylogenetics and Evolution* **85**, 76-87 (2015).
21. 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).
22. Fukushima, K. *et al.* Genome of the pitcher plant Cephalotus reveals genetic changes associated with carnivory. *Nature Ecology & Evolution* **1**, 0059 (2017).
23. Saul, F. *et al.* Subgenome dominance shapes novel gene evolution in the decaploid pitcher plant Nepenthes gracilis. *Nature plants* **9**, 2000-2015 (2023).
24. Nie, S. *et al.* Potential allopolyploid origin of Ericales revealed with gene-tree reconciliation. *Frontiers in Plant Science* **13**, 1006904 (2022).
25. Rose, J.P. *et al.* Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections. *Molecular Phylogenetics and Evolution* **122**, 59-79 (2018).
26. Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome research* **18**, 1944-1954 (2008).
27. Chanderbali, A.S. *et al.* Buxus and Tetracentron genomes help resolve eudicot genome history. *Nature communications* **13**, 643 (2022).
28. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-7 (2007).
29. Sankoff, D. *et al.* Towards improved reconstruction of ancestral gene order in angiosperm phylogeny. *Journal of Computational Biology* **16**, 1353-1367 (2009).
30. Wang, Y., Wang, J., Li, Y., Jin, Y. & Wang, X. Divergent evolutionary paces among eudicot plants revealed by simultaneously duplicated genes produced billions of years ago. *Frontiers in Plant Science* **16**, 1518981 (2025).
31. Alger, E.I. & Edger, P.P. One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Current opinion in plant biology* **54**, 108-113 (2020).
32. Twyford, A.D. *et al.* The polyploid continuum and the landscape of polyploid genomic variation. *American Journal of Botany*, e70121 (2025).
33. Wendel, J.F., Lisch, D., Hu, G. & Mason, A.S. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Current opinion in genetics & development* **49**, 1-7 (2018).
34. Wendel, J.F., Jackson, S.A., Meyers, B.C. & Wing, R.A. Evolution of plant genome architecture. *Genome Biol* **17**, 37 (2016).
35. Bird, K.A., VanBuren, R., Puzey, J.R. & Edger, P.P. The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytol* **220**, 87-93 (2018).
36. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K.F. Nanopore sequencing technology, bioinformatics and applications. *Nature biotechnology* **39**, 1348-1365 (2021).
37. 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).
38. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **5**, 4.10. 1-4.10. 14 (2004).

39. 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).
40. Albert, V.A. & Krabbenhoft, T.J. Navigating the CoGe Online Software Suite for Polyploidy Research. *Methods Mol Biol* **2545**, 19-45 (2023).
41. Lyons, E., Pedersen, B., Kane, J. & Freeling, M. The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. *Tropical Plant Biology* **1**, 181-190 (2008).
42. Srivastava, A., Rogers, W.L., Breton, C.M., Cai, L. & Malmberg, R.L. Transcriptome analysis of Sarracenia, an insectivorous plant. *DNA research* **18**, 253-261 (2011).
43. Cui, L. *et al.* Widespread genome duplications throughout the history of flowering plants. *Genome research* **16**, 738-749 (2006).
44. Lynch, M. & Conery, J.S. The evolutionary demography of duplicate genes. *Journal of structural and functional genomics* **3**, 35-44 (2003).
45. Jiao, Y. *et al.* A genome triplication associated with early diversification of the core eudicots. *Genome biology* **13**, R3 (2012).
46. Vekemans, D. *et al.* Gamma Paleohexaploidy in the Stem Lineage of Core Eudicots: Significance for MADS-Box Gene and Species Diversification. *Molecular Biology and Evolution* **29**, 3793-3806 (2012).
47. Wang, Y., Ficklin, S.P., Wang, X., Feltus, F.A. & Paterson, A.H. Large-Scale Gene Relocations following an Ancient Genome Triplication Associated with the Diversification of Core Eudicots. *PLoS One* **11**, e0155637 (2016).
48. Joyce, B.L. *et al.* FractBias: a graphical tool for assessing fractionation bias following polyploidy. *Bioinformatics* **33**, 552-554 (2017).
49. Fleck, S.J. & Jobson, R.W. Molecular Phylogenomics Reveals the Deep Evolutionary History of Carnivory across Land Plants. *Plants* **12**, 3356 (2023).
50. Stein, M. *et al.* Arabidopsis PEN3/PDR8, an ATP binding cassette transporter, contributes to nonhost resistance to inappropriate pathogens that enter by direct penetration. *Plant Cell* **18**, 731-46 (2006).
51. Ugalde, J.M., Nath, M., Wagner, S. & Meyer, A.J. Diversification of glutathione transferases in plants and their role in oxidative stress defense. *Biological Chemistry* **406**, 199-218 (2025).
52. Benveniste, I. *et al.* CYP86A1 from Arabidopsis thaliana encodes a cytochrome P450-dependent fatty acid omega-hydroxylase. *Biochem Biophys Res Commun* **243**, 688-93 (1998).
53. Ritpitakphong, U. *et al.* The microbiome of the leaf surface of Arabidopsis protects against a fungal pathogen. *New Phytologist* **210**, 1033-1043 (2016).
54. Du, X. *et al.* Molecular basis of the plant ROS1-mediated active DNA demethylation. *Nature Plants* **9**, 271-279 (2023).
55. Robe, K., Lefebvre-Legendre, L., Cleard, F. & Barberon, M. Four ZIPs contribute to Zn, Fe, Cu and Mn acquisition at the outer root domain. *PLoS Genet* **21**, e1011796 (2025).
56. Michael, R., Ranjan, A., Kumar, R.S., Pathak, P.K. & Trivedi, P.K. Light-regulated expression of terpene synthase gene, AtTPS03, is controlled by the bZIP transcription factor, HY5, in Arabidopsis thaliana. *Biochemical and Biophysical Research Communications* **529**, 437-443 (2020).

57. Roos, J., Bejai, S., Mozūraitis, R. & Dixelius, C. Susceptibility to Verticillium longisporum is linked to monoterpene production by TPS23/27 in Arabidopsis. *The Plant Journal* **81**, 572-585 (2015).
58. Vijayakumar, P., Datta, S. & Dolan, L. ROOT HAIR DEFECTIVE SIX-LIKE4 (RSL4) promotes root hair elongation by transcriptionally regulating the expression of genes required for cell growth. *New Phytologist* **212**, 944-953 (2016).
59. Brotman, Y. *et al.* Transcript and metabolite analysis of the Trichoderma-induced systemic resistance response to Pseudomonas syringae in Arabidopsis thaliana. *Microbiology* **158**, 139-146 (2012).
60. Kondou, Y. *et al.* RETARDED GROWTH OF EMBRYO1, a New Basic Helix-Loop-Helix Protein, Expresses in Endosperm to Control Embryo Growth *Plant Physiology* **147**, 1924-1935 (2008).
61. Wu, L. *et al.* Comparative Proteomics Identified Proteins in Mung Bean Sprouts Under Different Concentrations of Urea. *Molecules* **30**, 3176 (2025).
62. Nam, J., dePamphilis, C.W., Ma, H. & Nei, M. Antiquity and Evolution of the MADS-Box Gene Family Controlling Flower Development in Plants. *Molecular Biology and Evolution* **20**, 1435-1447 (2003).
63. Pařenicová, L. *et al.* Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in Arabidopsis : New Openings to the MADS World[W]. *The Plant Cell* **15**, 1538-1551 (2003).
64. Zhang, Z., Wang, J., Zhang, R. & Huang, R. The ethylene response factor AtERF98 enhances tolerance to salt through the transcriptional activation of ascorbic acid synthesis in Arabidopsis. *The Plant Journal* **71**, 273-287 (2012).
65. Lian, X.Y., Gao, H.N., Jiang, H., Liu, C. & Li, Y.Y. MdKCS2 increased plant drought resistance by regulating wax biosynthesis. *Plant Cell Rep* **40**, 2357-2368 (2021).
66. Best, N.B. *et al.* nana plant2 Encodes a Maize Ortholog of the Arabidopsis Brassinosteroid Biosynthesis Gene DWARF1, Identifying Developmental Interactions between Brassinosteroids and Gibberellins *Plant Physiology* **171**, 2633-2647 (2016).
67. Yamaguchi, N. *et al.* H3K27me3 demethylases alter HSP22 and HSP17.6C expression in response to recurring heat in Arabidopsis. *Nat Commun* **12**, 3480 (2021).
68. Wu, H. *et al.* Regulation of Arabidopsis Matrix Metalloproteinases by Mitogen-Activated Protein Kinases and Their Function in Leaf Senescence. *Front Plant Sci* **13**, 864986 (2022).
69. Yang, S.U., Kim, H., Kim, R.J., Kim, J. & Suh, M.C. AP2/DREB Transcription Factor RAP2.4 Activates Cuticular Wax Biosynthesis in Arabidopsis Leaves Under Drought. *Front Plant Sci* **11**, 895 (2020).
70. Guttikonda, S.K. *et al.* Overexpression of AtDREB1D transcription factor improves drought tolerance in soybean. *Mol Biol Rep* **41**, 7995-8008 (2014).
71. Krishnamurthy, P., Vishal, B., Bhal, A. & Kumar, P.P. WRKY9 transcription factor regulates cytochrome P450 genes CYP94B3 and CYP86B1, leading to increased root suberin and salt tolerance in Arabidopsis. *Physiol Plant* **172**, 1673-1687 (2021).
72. Strzalka, W.K. *et al.* Arabidopsis PCNAs form complexes with selected D-type cyclins. *Front Plant Sci* **6**, 516 (2015).
73. Holding, D.R. & Springer, P.S. The Arabidopsis gene PROLIFERA is required for proper cytokinesis during seed development. *Planta* **214**, 373-82 (2002).

74. Zhang, J. *et al.* DAO1 catalyzes temporal and tissue-specific oxidative inactivation of auxin in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **113**, 11010-5 (2016).
75. Bohmert, K. *et al.* AGO1 defines a novel locus of Arabidopsis controlling leaf development. *Embo j* **17**, 170-80 (1998).
76. Beuchat, J. *et al.* BRX promotes Arabidopsis shoot growth. *New Phytol* **188**, 23-9 (2010).
77. Manfield, I.W., Devlin, P.F., Jen, C.-H., Westhead, D.R. & Gilmartin, P.M. Conservation, Convergence, and Divergence of Light-Responsive, Circadian-Regulated, and Tissue-Specific Expression Patterns during Evolution of the Arabidopsis GATA Gene Family. *Plant Physiology* **143**, 941-958 (2007).
78. Tominaga-Wada, R., Kurata, T. & Wada, T. Localization of ENHANCER OF TRY AND CPC1 protein in Arabidopsis root epidermis. *J Plant Physiol* **214**, 48-52 (2017).
79. Shapiguzov, A. *et al.* Arabidopsis RCD1 coordinates chloroplast and mitochondrial functions through interaction with ANAC transcription factors. *Elife* **8**(2019).
80. Rendón-Anaya, M. *et al.* The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proceedings of the National Academy of Sciences* **116**, 17081-17089 (2019).
81. Chao, A. *et al.* Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* **84**, 45–67 (2014).
82. Tang, H. *et al.* Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* **12**, 102 (2011).
83. Schlueter, J.A. *et al.* Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**, 868-876 (2004).
84. Jost, L. Entropy and diversity. *Oikos* **113**, 363-375 (2006).

Supplementary Files

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

- TablesSASIf.zip