Correction

EVOLUTION

Correction for "Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome," by Tianying Lan, Tanya Renner, Enrique Ibarra-Laclette, Kimberly M. Farr, Tien-Hao Chang, Sergio Alan Cervantes-Pérez, Chunfang Zheng, David Sankoff, Haibao Tang, Rikky W. Purbojati, Alexander Putra, Daniela I. Drautz-Moses, Stephan C. Schuster, Luis Herrera-Estrella, and Victor A. Albert, which appeared in issue 22, May 30, 2017, of *Proc Natl Acad Sci USA* (114:E4435–E4441; first published May 15, 2017; 10.1073/pnas.1702072114).

The authors note that on page E4436, right column, first full paragraph, lines 3–4, "30,689, 7.7% more than reported for our short-read assembly" should instead appear as "29,666, 4.1% more than reported for our short-ready assembly."

www.pnas.org/cgi/doi/10.1073/pnas.1709197114

Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome

Tianying Lan[a], Tanya Renner[b], Enrique Ibarra-Laclette[c], Kimberly M. Farr[a], Tien-Hao Chang[a], Sergio Alan Cervantes-Pérez[d], Chunfang Zheng[e], David Sankoff[e], Haibao Tang[f], Rikky W. Purbojati[g], Alexander Putra[g], Daniela I. Drautz-Moses[g], Stephan C. Schuster[g,1], Luis Herrera-Estrella[d,1], and Victor A. Albert[a,1]

[a]Department of Biological Sciences, University at Buffalo, Buffalo, NY 14260; [b]Department of Biology, San Diego State University, San Diego, CA 92182; [c]Red de Estudios Moleculares Avanzados, Instituto de Ecología A. C., C.P. 91070 Xalapa, México; [d]Laboratorio Nacional de Genómica para la Biodiversidad, Unidad de Genómica Avanzada del Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 36500 Guanajuato, México; [e]Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada K1N 6N5; [f]Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China; and [g]Singapore Centre on Environmental Life Sciences Engineering, Nanyang Technological University, Singapore 637551

Contributed by Luis Herrera-Estrella, April 7, 2017 (sent for review February 14, 2017; reviewed by Aaron Liston and Yves Van de Peer)

Utricularia gibba*, the humped bladderwort, is a carnivorous plant that retains a tiny nuclear genome despite at least two rounds of whole genome duplication (WGD) since common ancestry with grapevine and other species. We used a third-generation genome assembly with several complete chromosomes to reconstruct the two most recent lineage-specific ancestral genomes that led to the modern *U. gibba* genome structure. Patterns of subgenome dominance in the most recent WGD, both architectural and transcriptional, are suggestive of allopolyploidization, which may have generated genomic novelty and led to instantaneous speciation. Syntenic duplicates retained in polyploid blocks are enriched for transcription factor functions, whereas gene copies derived from ongoing tandem duplication events are enriched in metabolic functions potentially important for a carnivorous plant. Among these are tandem arrays of cysteine protease genes with trap-specific expression that evolved within a protein family known to be useful in the digestion of animal prey. Further enriched functions among tandem duplicates (also with trap-enhanced expression) include peptide transport (intercellular movement of broken-down prey proteins), ATPase activities (bladder-trap acidification and transmembrane nutrient transport), hydrolase and chitinase activities (breakdown of prey polysaccharides), and cell-wall dynamic components possibly associated with active bladder movements. Whereas independently polyploid *Arabidopsis* syntenic gene duplicates are similarly enriched for transcriptional regulatory activities, *Arabidopsis* tandems are distinct from those of *U. gibba*, while still metabolic and likely reflecting unique adaptations of that species. Taken together, these findings highlight the special importance of tandem duplications in the adaptive landscapes of a carnivorous plant genome.

plant genomics | evolution | polyploidy | carnivorous plant | *Utricularia*

The architectural evolution of flowering plant genomes includes a long history of gene duplication and diversification. Tandem gene duplication is an ongoing but nonglobal process that generates coding sequence diversity in eukaryotic genomes through subfunctionalization or neofunctionalization of gene copies on an individual basis (1). On the other hand, polyploidy events provide scores of genomically balanced duplicate genes all at once, on which divergent selection pressures can act to generate phenotypic diversity (2, 3). Evidence from available plant genomes supports the theory that modular, dosage-sensitive functions such as transcriptional regulation are enriched among duplicates surviving polyploidy events, whereas single-gene survivors of local duplication events have the opportunity to be enriched for dosage responsive functions, such as secondary metabolite production (e.g., refs. 4–7). Although it has been repeatedly noted that polyploidy events correlate with some major plant radiations (2, 8, 9), the specific roles that tandem duplicates play in species- or lineage-specific plant adaptation remain more poorly explored.

Utricularia gibba is an aquatic carnivorous plant with an unusually small but highly dynamic nuclear genome that experienced at least two whole-genome duplication (WGD) events during its evolutionary history since divergence from grapevine, tomato, and other species (10). Carnivorous plants are interesting model systems not only for understanding the molecular mechanisms underlying nutrient acquisition strategies, but also for discovering the regulatory underpinnings of their unique trapping morphologies. *U. gibba* is of particular interest given the previous publication of an ∼82-Mb short-read assembly (10), which revealed that its genome gained and deleted gene duplicates significantly faster than those of other genomes (11). Given that the *U. gibba* genome likely descended via considerable shrinkage from an ancestral genome up to 1.5 Gb in size (12), duplicates that survived deletion during its evolutionary history arguably evolved under greater purifying selection pressure compared with the more expansive genomes of most angiosperms. Therefore, we hypothesized that the deletion-prone genome of *U. gibba* could be particularly illustrative regarding

Significance

Carnivorous plants capture and digest animal prey for nutrition. In addition to being carnivorous, the humped bladderwort plant, *Utricularia gibba*, has the smallest reliably assembled flowering plant genome. We generated an updated genome assembly based on single-molecule sequencing to address questions regarding the bladderwort's genome adaptive landscape. Among encoded genes, we segregated those that could be confidently distinguished as having derived from small-scale versus whole-genome duplication processes and showed that conspicuous expansions of gene families useful for prey trapping and processing derived mainly from localized duplication events. Such small-scale, tandem duplicates are therefore revealed as essential elements in the bladderwort's carnivorous adaptation.

Author contributions: T.L., S.C.S., L.H.-E., and V.A.A. designed research; T.L., R.W.P., A.P., and D.I.D.-M. performed research; T.L., E.I.-L., S.A.C.-P., and L.H.-E. contributed new reagents/analytic tools; T.L., T.R., E.I.-L., K.M.F., T.-H.C., C.Z., D.S., H.T., R.W.P., and V.A.A. analyzed data; and T.L., T.R., D.S., R.W.P., D.I.D.-M., L.H.-E., and V.A.A. wrote the paper.

Reviewers: A.L., Oregon State University; and Y.V.d.P., Ghent University.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: This Whole Genome Shotgun project has been deposited at the DNA Data Bank of Japan/European Nucleotide Archive/GenBank (accession no. NEEC00000000). The version described in this paper is version NEEC01000000. The assembly and gene models are also available at https://genomevolution.org/coge/GenomeInfo.pl?gid=29027.

[1]To whom correspondence may be addressed. Email: scschuster@ntu.edu.sg, lherrerae@cinvestav.mx, or vaalbert@buffalo.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1702072114/-/DCSupplemental.

the adaptive legacy of differential duplicate survival following their two modes of generation, with tandems highlighting aspects of the carnivorous lifestyle and syntenic duplicates highlighting transcriptional functions.

To explore this possibility, we generated a highly contiguous nuclear genome assembly for *U. gibba* based on Pacific Biosciences (PacBio) Single Molecule, Real-Time (SMRT) technology. We used 10 SMRT cells and P6-C4 PacBio chemistry to produce 521,937 raw and 702,640 filtered subreads with N50 values of 21,825 and 15,244 bp, respectively. After assembly with HGAP.3 (13), we produced a genome of 581 contigs with an N50 of 3,424,836 bp and 101,949,210 total bases (*SI Appendix*, Fig. S2). Remarkably, base pair correction using either the PacBio data or Illumina MiSeq reads from our previous assembly (10) led to extremely minor improvements, only 0.071% and 0.01% of total bases, respectively (*SI Appendix*, section 1.5). Four contigs represented complete chromosomes marked on either end by telomeres, including the longest contig of the assembly at 8,502,017 bp (Fig. 1). Twenty additional contigs had telomere repeats on one end, the 14 largest being ≥1 Mb in size (Fig. 1). *Arabidopsis*-type telomeric repeats (TTTAGGG) were identified in these 24 contigs. Two variants, the *Chlamydomonas* type (14) (TTTTTAGGG) and TTCAGGG (similar to the variants TTCAGG and TTTCAGG known from the close carnivorous plant relative *Genlisea*) (15), were also found sporadically intermingled with the *Arabidopsis*-type telomeric repeats. Ten contigs were observed to have interstitial telomeric repeats, which were identified by searching for $(CCCTAAA)_3$ and $(TTTAGGG)_3$ within chromosomal arms (Fig. 1*A*). After filtration for bacterial and other contamination (*SI Appendix*, section 1.6), the assembled genome amounted to 100,688,548 bp (on 518 contigs), including a complete 172,489-bp plastid genome on a single contig and a 283,823-bp partial mitochondrial genome (*SI Appendix*, section 1.6.2). Therefore, our newly assembled nuclear genome gained 18,356,750 bp from the former assembly size of 81,875,486 bp.

Calculation of the genome space occupied by transposable elements (TEs) uncovered almost 9 Mb (~8.9%) complete TEs, with up to 59 Mb (~59%) of the nuclear genome possibly TE-derived (*SI Appendix*, Dataset S1); the latter amounted to ~16.6 Mb more TE-related genome space than was found in the previously published short-read assembly (*SI Appendix*, section 2.1). We found that ~2.9 Mb of the genome (on 115 contigs) was composed of ribosomal DNA repeats (*SI Appendix*, section 2.2). Indeed, a syntenic path alignment with the short-read assembly demonstrated that most of the DNA gained by PacBio sequencing contained repeated elements, particularly surrounding putative centromeres (Fig. 1*B* and *SI Appendix*, Figs. S4–S8).

To identify signature centromeric repeats in *U. gibba*, we selected tandem repeat clusters with average period size of 50–500 bp for identification as putative centromere repeats (*SI Appendix*, Fig. S5*B*), as described previously (16). The top 10 most abundant tandem repeat clusters were considered prime candidates for centromeric repeats, but these were not even preferentially located in our chromosome-sized contigs. We then manually checked the locations of the next 10 most abundant tandem repeat clusters in the genome, and found that none of these clusters showed unique localization in putative centromeric regions. Therefore, we conclude that *U. gibba* centromeres are devoid of high-copy tandem repeat arrays such as those known from *Arabidopsis* and maize (16). Similar findings also have been reported for the centromeres of several plant and animal species (17–19), including two closely related carnivorous plants, *Genlisea hispidula* and *Genlisea subglabra (15)*.

Although plant retrotransposon families generally are randomly dispersed, there are families distinctly concentrated in centromeric regions, such as the CRM centromeric chromoviruses. CRMs, a lineage of *Ty3/gypsy* retrotransposons, have been well characterized as centromeric retrotransposons in many species (20–25), including *G. hispidula* and *G. subglabra (15)*. Using phylogenetic analysis, we found that 55 *U. gibba* sequences are grouped within the subgroup A CRMs, which include the centromere-specific CRMs (*SI Appendix*, section 3.3.3). All but one of the *U. gibba* sequences form a single, monophyletic CRM subfamily. To investigate the chromosomal localization of the 55 *U. gibba* CRMs, we plotted them on the complete and near-complete chromosomes together with the TE and gene model tracks. As depicted in Fig. 1*A*, most *U. gibba* CRMs are located in the putative centromeric regions; however, not all putative centromeres have CRM elements. It has been proposed that CRMs may play an important role in stabilizing centromere structure and maintaining centromere function (26, 27), whereas an opposing hypothesis holds that they are merely parasitic and tend to accumulate in recombination-poor centromeric regions to escape negative selection against insertions in distal regions (28). Our finding that some putative centromeric regions in *U. gibba* lack CRMs or other high-copy centromeric tandem repeats suggests that neither CRMs nor tandem repeats are crucial for maintaining functional centromeres in the species.

Our highly contiguous genome assembly also permitted a much finer account of protein-coding gene number than previously available, which amounted to 30,689, 7.7% more than reported for our short-read assembly (10). Unlike the far shorter scaffolds from that assembly (10), our largely chromosome-sized contigs permitted us to conservatively distinguish the WGD-derived and tandem duplicate portions of *U. gibba*'s genome adaptive landscape. In both cases, we were concerned with duplicates that could still be discerned within their formative genome structural contexts, not with duplicates that might have migrated to other chromosomal positions after their generation via small-scale or WGD events, because such genes could be only indirectly assigned to one duplicative process versus the other.

Through syntenic analysis using CoGe (29, 30), we were able to identify 54 syntenic block pairs descending from the most recent *U. gibba* WGD event (*SI Appendix*, Fig. S11). We were then able to reconstruct the immediate, nine-chromosome pre-polyploid ancestor of the modern genome, following which numerous large-scale inversion events were required to account for modern gene order (*SI Appendix*, section 4.1). Further analysis permitted deconstruction of this ancestral genome into an earlier, six-chromosome pre-WGD ancestor that existed immediately before *U. gibba*'s second most recent polyploidy event (*SI Appendix*, Fig. S12); however, we could not reconstruct the third WGD event that was previously described based on visual inspection of syntenic dot plots and syntenic depth calculations (10). Nonetheless, microsynteny analyses did reveal many examples of eight (or more)-to-one syntenic block relationships with the *Vitis vinifera* genome (Fig. 2 and *SI Appendix*, section 4.5), some of which may include blocks dating to the gamma hexaploidy event at the base of all core eudicots (31).

We analyzed the duplicate block pairs from the most recent WGD event to assess the degree of fractionation (gene loss) experienced by each subgenome following polyploidization (*SI Appendix*, Fig. S13). This analysis yielded a clear pattern of deletion bias characteristic of subgenome dominance inherited through a polyploidy event (32, 33). Fractionation bias was matched by both subgenome expression dominance (34) and fewer single nucleotide polymorphisms on dominant blocks (35, 36) (*SI Appendix*, section 4.4, Figs. S13 and S14, and Datasets S3 and S4), indicating the influence of stronger purifying selection. Taken together, these data suggest that the most recent WGD in *U. gibba*'s past was an allopolyploidization event resulting from a broad cross (37), because autopolyploidies are not expected to show such strong biases; for example, unbiased fractionation has been discovered in the genomes of poplar, banana, and soybean (37, 38). Hybridization of two species accompanied by genome doubling can instantly generate a third species with novel and transcendent

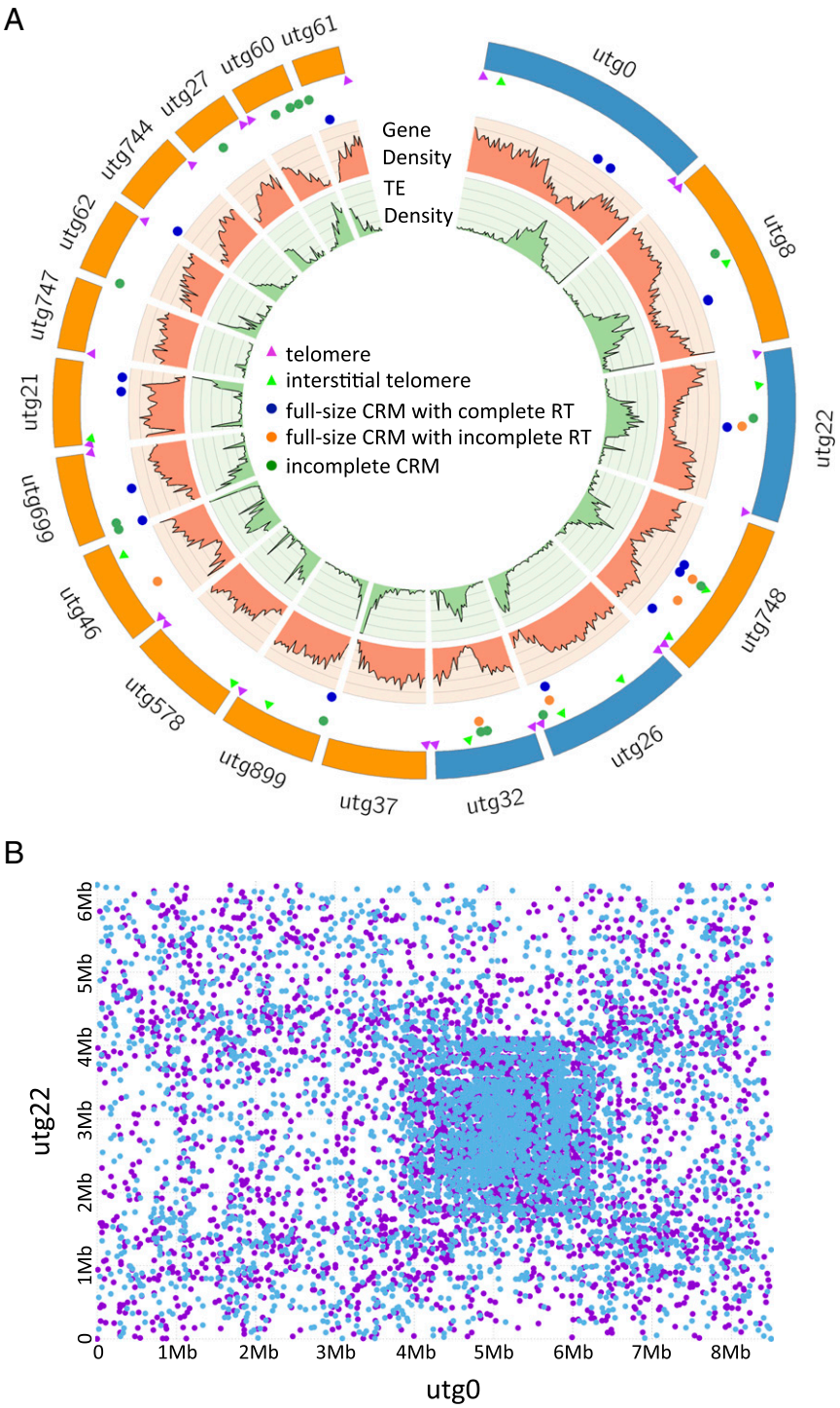


Fig. 1. A chromosome-scale view of the architecture of the *U. gibba* genome. (*A*) Gene density, TE density tracks, telomeres, and the locations of CRM centromeric retrotransposon sequences are shown for all *U. gibba* contigs >1 Mb in size. Four complete chromosomal contigs are shown in blue, and partial chromosomes that have at least one end with telomere sequence are shown in orange. Putative centromeric regions are visible as peaks of increased TE density and decreased gene density. Most CRMs are localized at putative centromeric regions. (*B*) MUMmer (82) pairwise dot-plot alignment of contigs 0 and 22, which represent complete chromosomes. Blue and purple dots indicate hits on each DNA strand, respectively. Putative centromeric regions of strong sequence similarity are apparent as a densely hit square.

phenotypic traits (39). Moreover, the modern *U. gibba* genome displays highly heterogeneous patterns of heterozygosity (*SI Appendix*, Dataset S4) that do not correlate with the structural limits of syntenic blocks, suggesting that outcrossing events subsequent to the most recent WGD were broad, but were not followed by ploidy changes. Given the highly clonal nature of aquatic *Utricularia* species (e.g., refs. 40, 41), this state could represent "frozen" heterozygosity in a particularly adaptive genotype, such as seen in unisexual hybrid vertebrates (42).

To examine polyploid adaptive genetic features in *U. gibba*, we evaluated gene ontology (GO) functional enrichments among

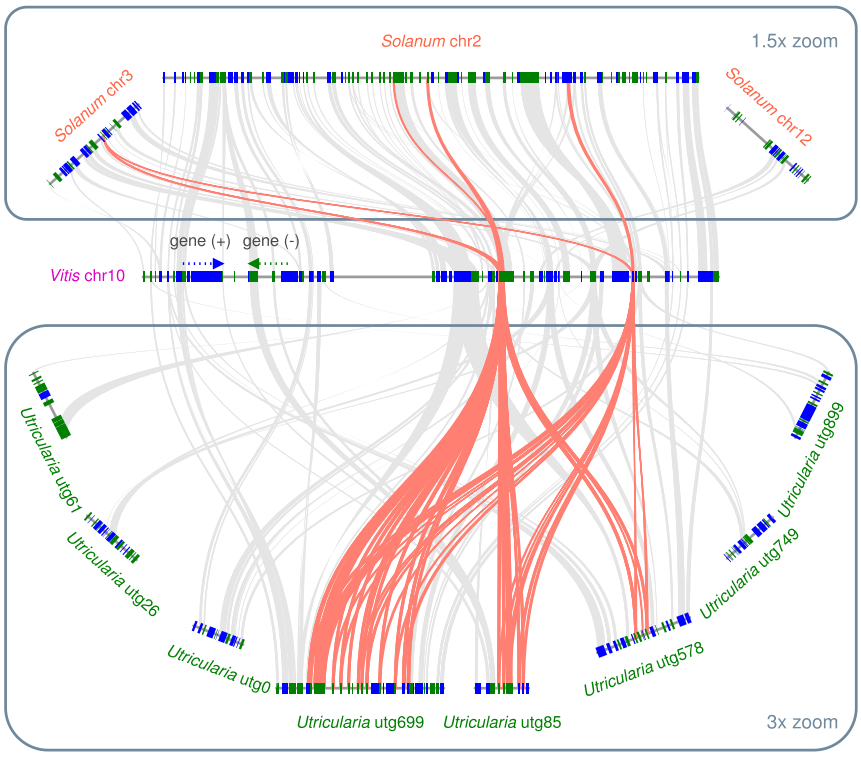


Fig. 2. Syntenic relationships among *V. vinifera*, *S. lycopersicum*, and *U. gibba* regions containing tandemly duplicated cysteine protease genes. Some parts of these tandem arrays clearly preexisted in *U. gibba*'s prepolyploid ancestral genomes, with further tandem duplications having occurred since those events, together increasing functional potential for *U. gibba*'s carnivory. A typical ancestral region in *Vitis* can be traced to up to three regions in *Solanum* (through the latter's genome triplication) and up to eight regions in *U. gibba* (where as many as three WGDs are possible). Red connecting lines highlight matching cysteine proteases in the selected regions; genes otherwise syntenic are shown in gray.

syntenically retained gene duplicates descending from *U. gibba*'s lineage-specific WGDs. Duplicates retained following WGD were mostly enriched for transcriptional regulatory functions (*SI Appendix*, Dataset S5). As expected based on earlier studies, very similar results were obtained for *Arabidopsis* WGD duplicates analyzed in the same manner (*SI Appendix*, Dataset S6) (4, 43, 44); however, comparing the 522 *U. gibba* WGD duplicates annotated with the GO "regulation of transcription, DNA-templated" with all *U. gibba* genes with this GO revealed no significant enrichment of any biological process category (*SI Appendix*, Dataset S14). Similar analysis of *Arabidopsis* WGD duplicates yielded only one significant biological process category, "response to jasmonic acid" (*SI Appendix*, Dataset S15), suggesting that in both species, transcriptional regulatory enrichment is functionally generic.

In contrast to functional enrichments of WGD duplicates, *U. gibba* genes filtered out by the blast_to_raw script in the QUOTA-ALIGN package [https://github.com/tanghaibao/quota-alignment (45), included in CoGe SynMap (29, 30)] as tandem duplicates in the modern genome (and thus ignored in syntenic dot plot comparison) were enriched for many secondary metabolic functions, including specific functions that could be anticipated for a carnivorous plant (*SI Appendix*, Datasets S7 and S8). *Arabidopsis* tandems discovered in the same manner were similarly enriched for secondary metabolic activities, as anticipated based on earlier results (5). However, in many cases the *Arabidopsis* activities were entirely different (*SI Appendix*, Dataset S9). Among the most significantly enriched categories in *U. gibba* was the category "oligopeptide transporter activity," assigned to 23 members of the OPT gene family (46). Importantly, oligopeptide transport was also among the most significantly enriched functional categories of genes specifically and strongly expressed in the bladder traps (47), with 13 genes showing 4- to 400-fold trap-enhanced transcription (*SI Appendix*, Dataset S8). Peptide transporters, which are involved in the plant nitrogen budget, have been identified as expressed in the trap fluid of the carnivorous pitcher plant *Nepenthes* (48, 49). The *Nepenthes* gene identified in that study is, however, a member of the PTR family, a group itself highlighted among *U. gibba* tandems by the significantly enriched term "dipeptide transporter activity," wherein there are 22 family members, including three homologs of the *Arabidopsis* nitrate transporter gene *NPF5.5* (50); *unitig_52.g17408.t1* and *unitig_26.g9035.t1* had >65-fold trap-enhanced expression (*SI Appendix*, Dataset S8). Carnivorous plants, bladderworts included, typically grow in nitrogen-poor habitats, where they compensate for deficiencies via prey capture and uptake of released nitrogen.

Another highly enriched functional category among tandem duplicates was "ATPase activity, coupled to transmembrane movement of substances," comprising 58 genes, mostly ABC transporters. Proteins encoded by such genes are known from *Nepenthes* traps, where they are hypothesized to be responsible for maintaining trap acidity and various molecular transport functions (51). Several of these genes show greater than ninefold trap-specific expression, including *unitig_85.g27344.t1*, *unitig_85.g27345.t1*, *unitig_750.g28500.t1*, and *unitig_750.g28501.t1* (*SI Appendix*, Dataset S8). Another enriched category was "transmembrane transport," which highlighted all of the foregoing genes and also included eight phosphate transporter genes homologous to *PHT1* (52). PHT1 family genes are induced during nutritional phosphate deficiency, a condition characteristic of the carnivorous plant lifestyle (53). Of these, *unitig_747.g21685.t1* and *unitig_747.g21690.t1* showed 2- to 24-fold trap-enhanced expression (*SI Appendix*, Dataset S8).

Another significantly enriched tandem duplicate functional category was "hydrolase activity, hydrolyzing O-glycosyl compounds." This GO category included a gene encoding a class III chitinase (*unitig_60.g25630.t1*, showing >20-fold trap-enhanced expression) (*SI Appendix*, Dataset S8), representing one of the chitinase families [glycoside hydrolase (GH) family 18] active within the digestive fluid of both open and closed traps of various carnivorous plant species. In *Nepenthes*, the GH family 18 enzyme is encoded by a single-copy gene that is up-regulated in response to prey in both the pitted glands and surrounding tissues (54). Galactosidases and xylosidases (55) are also among the genes with the hydrolase annotation, and enzymes encoding both have been identified in the *Nepenthes* trap fluid proteome (56, 57). *Nepenthes* and *Drosera* (carnivorous sundew plant) digestive mucilage contains galactose and xylose (58), which may require breakdown for peptide and other nutrient absorption in *U. gibba* traps as well (59). Three xylosidase genes—*unitig_62.g23624.t1*, *unitig_62.g23625.t1*, and *unitig_748.g7352.t1*—show 4- to 35-fold trap-enhanced expression (*SI Appendix*, Dataset S8).

The traps of *Utricularia* operate through an intricate triggering mechanism (60). High-speed snap-buckling movements (61, 62) occur following triggered release of negative internal trap pressure achieved by active pumping out of water (63). Prey is engulfed with the influx of liquid, after which the trap may reset itself with a new negative pressure potential. This repeating process likely demands highly dynamic cell-wall changes. Indeed, the tandems-enriched GO category "cell wall" annotated 17 genes encoding expansins (64) (none of which, however, showed uniformly trap-enriched expression) and 8 genes encoding xyloglucan endotransglycosylases (65) (of which *unitig_749.g14196.t1* and

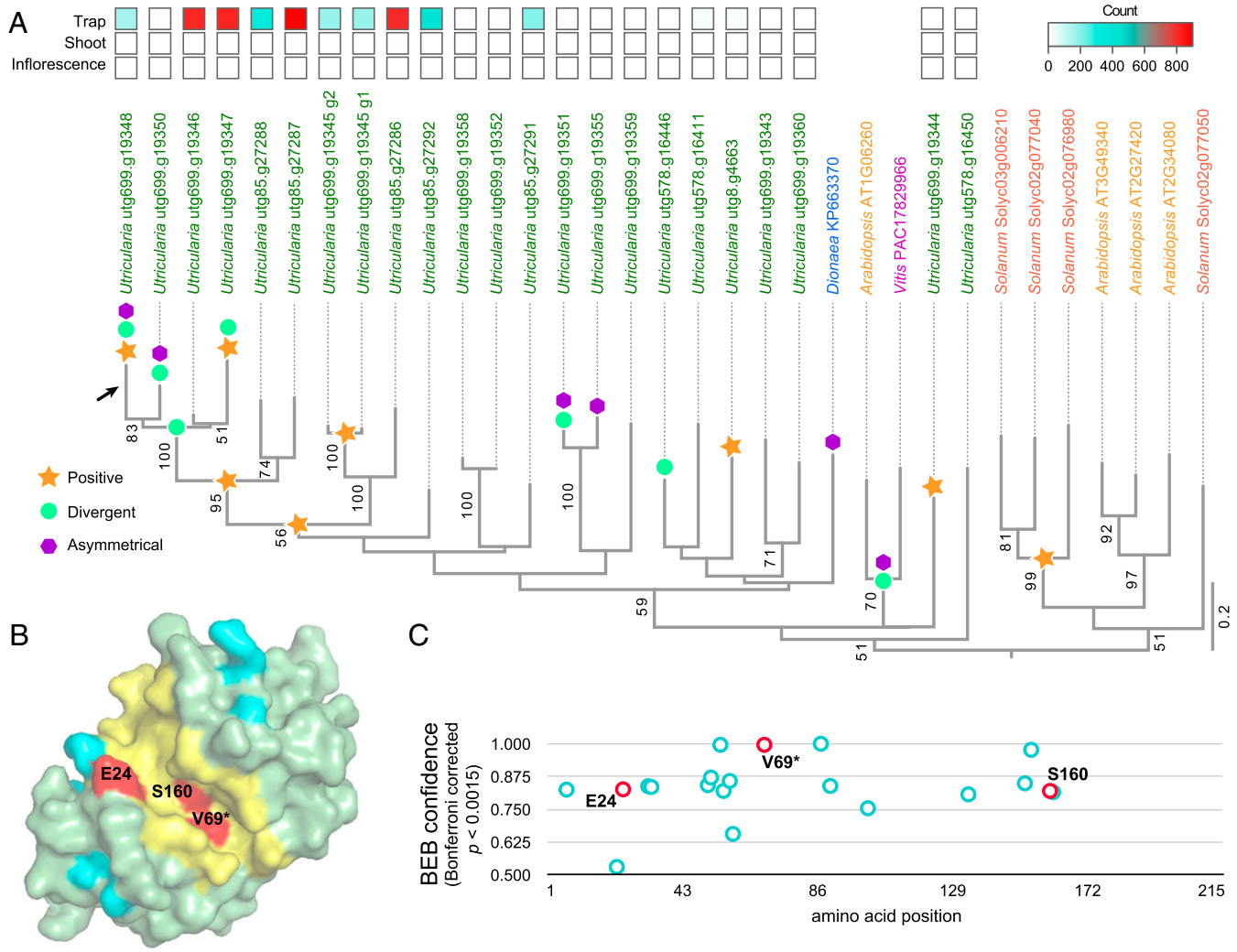


Fig. 3. Molecular and structural evolutionary analysis of *U. gibba* cysteine proteases suggests adaptive protein evolution accompanying WGD and tandem duplication events. (*A*) Best-scoring tree from maximum-likelihood based searches, with bootstrap support (BS) values ≥50 indicated at branches. Symbols on branches indicate significant evidence for positive selection (orange stars), divergent selection (green circles), or asymmetrical sequence evolution (purple hexagons) as determined using PAML (83) (*SI Appendix*, Dataset S10). The heatmap above the phylogeny shows trap-dominant expression of particular homologs in *U. gibba*, based on trap, shoot, and inflorescence transcriptome data (47) (*SI Appendix*, Dataset S2). Note that two tandem duplicates (g1 and g2) were repredicted at locus *utg699.g19345*. (*B*) The protein homology surface model for the catalytic domain of utg699.g19348 (encoded by the gene annotated by an arrow in *A*; based on the Venus flytrap [*D. muscipula*] enzyme structure (77)) shows that some residues under positive selection lie within or near the substrate-binding cleft. The cleft is depicted in yellow, and amino acid sites identified as under positive selection are indicated in red or cyan. Three (E24, V69, and S160) amino acid sites under positive selection (BEB confidence >0.82, Bonferroni-corrected $P < 0.0015$) are within five amino acids of known *D. muscipula* functional residues, where they line the substrate-binding cleft (red). (*C*) Plot of utg699.g19348 amino acid sites under positive selection, with colors corresponding to specific sites in the surface model (*SI Appendix*, Fig. S4*B*).

unitig_26.g9135.t1 showed greater than sixfold trap-enhanced expression) (*SI Appendix*, Dataset S8). Seventeen encoded peroxidases homologous to PRX52, which cross-link cell-wall strengthening extensins (*unitig_26.g8978.t1* and *unitig_22.g6605.t1* were >14-fold trap-enhanced), and 21 encoded polygalacturonases, which degrade cell-wall pectin (66) (*unitig_8.g3155.t1* and *unitig_8.g3156.t1* were >fourfold trap-enhanced) (*SI Appendix*, Dataset S8). Indeed, members of these protein families have been identified as candidates for involvement in plant mechanical stimulation or movements (62, 67, 68). Another cell-wall modification-related gene family under this GO term encoded a group of 19 pectin methylesterases and their inhibitors (69) (*unitig_899.g15179.t1* and *unitig_22.g5384.t1* were 2- to 32-fold trap-enhanced) (*SI Appendix*, Dataset S8). Interestingly, a second class of chitinases, the class IV enzymes, was also highlighted as an expanded gene family under the GO category "cell wall," but none of these five genes showed trap-enhanced expression. Class IV chitinases are defense response proteins that represent a second family of chitinase (GH family 19) involved in plant carnivory (70, 71). Finally, four genes encoding β-galactosidases (known from *Nepenthes* pitcher fluid) (57) appeared under the same GO category but did not have trap-enhanced expression in *U. gibba.* Another expanded GO category, "lipid catabolic process," comprised members of various lipase gene families, among them genes encoding patatin-like and GDSL lipases (*unitig_736.g22657.t1*, *unitig_37.g12702.t1*, *unitig_736.g22658.t1*, and *unitig_37.g12699.t1* showed 35- to 180-fold trap-enhanced expression) (*SI Appendix*, Dataset S8). A GDSL lipase likely related to carnivory was identified in the trap fluid of *Nepenthes* pitchers (57).

Strikingly, the most significantly enriched GO category among all tandemly duplicated genes, "senescence-associated vacuole," pointed to a specific expansion in one gene family encoding cysteine proteases that had nearly trap-specific expression patterns (*SI Appendix*, Datasets S2 and S8). Several other significantly enriched GOs are associated with this gene family. Cysteine proteases have been identified as major functional components of Venus flytrap (*Dionaea muscipula*) digestive fluid (72), reported in three *D. muscipula* transcriptomes (70, 73, 74), and structurally annotated for both Cape sundew (*Drosera capensis*) draft genome sequences (75, 76) and *D. muscipula* (77). We found tandem clusters of homologous protease-encoding genes in the *U. gibba* genome that had demonstrably undergone tandem duplication both before and after the most recent WGD event in *U. gibba*'s evolutionary history (Fig. 2). These tandem cysteine protease arrays are assignable to both dominant and recessive subgenomic blocks and are more preserved on the dominant block, where enhanced purifying selection on gene space is expected (*SI Appendix*, Fig. S13). Genome-wide BLAST search revealed that in general, *U. gibba* cysteine proteases have become nearly totally restricted to this single, specific subfamily, clearly indicating that diverse, related cysteine proteases known from various other species have become expendable during *U. gibba*'s genome evolution.

We further examined the cysteine proteases for molecular evolutionary features (*SI Appendix*, section 6.1), given that gene family members would have diversified in sequence and function to be retained by selection in the dynamically shrinking *U. gibba* genome. The alternative would be that the observed duplicates were extremely recent and functionally redundant; however, analyses of protein evolution showed this to not be the case, although tandem duplications did continue following the most recent WGD event that yielded arrays on contigs 85 and 699 (Fig. 3*A*). Instead, we detected evidence for positive selection acting on specific amino acid residues in a lineage leading to several of the *U. gibba* cysteine protease duplicates (Fig. 3*A*). When homology modeling these changes onto the *D. muscipula* cysteine protease structure (77) (Protein Data Bank ID code 5a24), we found some of these amino acids located within the substrate-binding cleft, near residues with known functions in protease activity (Fig. 3 *B* and *C*). These substitutions could affect polarity and charge within the cleft, as well as hydrogen bonding between residues essential for catalytic activity and the ligand.

SHORT VEGETATIVE PHASE (*SVP*) MADS box gene homologs and homologs of the cuticle biosynthesis gene *3-KETOACYL-COA SYNTHASE 6* (*KCS6*; highlighted by the significantly enriched GO category among tandems, "wax biosynthetic process") (*SI Appendix*, Dataset S8) are two additional cases of tandem duplicate arrays for which some members exhibit trap-enhanced gene expression. Both of these examples have been described previously, based on simple orthogroup clustering methods, as generic gene family expansions derived from unknown duplication mechanisms (11). However, only our highly contiguous PacBio genome provides the structural context necessary to discern that these duplicates are tandems. The *SVP*-like gene cluster may be involved with flowering phenology, and the *KCS6*-like genes may be involved in cuticle buttressing of the thin, two-celled trap wall (78–80). The *SVP*-like genes appear to have diversified anciently, whereas the *KCS6*-like array occurs in a region of the genome without internal synteny, so it is likely more recent than the last *U. gibba* WGD. Similar to the cysteine protease clusters, we discovered likely evidence of protein functional divergence in both of these array types (*SI Appendix*, Dataset S10). Also of note, both the cysteine protease and *KCS6*-like gene clusters occur within islands of mobile elements (*SI Appendix*, section 2.5) annotated as large retrotransposon derivatives (LARDs) (81). Serving as a good illustration of the repeat discovery power of PacBio sequencing, ∼47% of the total TE assembly space comprised LARDs, whereas these elements amounted to only ∼14.6% of TEs in the previous short-read assembly (*SI Appendix*, Dataset S1). We hypothesize that LARDs and other DNA repeats may have facilitated the tandem duplications that gave rise to metabolic gene arrays, as illustrated in the foregoing examples. Finally, we hypothesize that such tandem gene clusters could be coregulated to act in concert, perhaps at particular plant developmental stages or under particular environmental stimuli.

Taken together, our findings regarding the size-limited *U. gibba* genome highlight the important role that tandemly duplicated genes, under sufficiently substantial purifying selection to survive continual deletion pressure, may play in the individualized adaptive genomic architecture of a plant uniquely adapted for carnivorous morphology and physiology. Although WGD duplicates are not enriched for such niche-specific functions, polyploidy events clearly potentiated the evolutionary influence of preexisting tandem arrays.

Materials and Methods

U. gibba material was sourced from Umécuaro municipality, Michoacán, México, and grown in sterile tissue culture before nuclear DNA extraction. DNA was sequenced using PacBio SMRT technology and assembled using HGAP.3. Genome features were then annotated and analyzed using various bioinformatic tools. GO enrichments were analyzed within different gene pools. For selected gene families, molecular evolutionary pressures were evaluated using codon models and likelihood ratio tests. Detailed information is provided in *SI Appendix*.

ACKNOWLEDGMENTS. We thank Thomas J. Givnish for an insightful additional review. Funding for this work was provided by National Science Foundation Grants 0922742 and 1442190 (to V.A.A.).

1. Lynch M (2007) *The Origins of Genome Architecture* (Sinauer Associates, Sunderland, MA).
2. Soltis DE, et al. (2009) Polyploidy and angiosperm diversification. *Am J Bot* 96: 336–348.
3. Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. *Nat Rev Genet* 10:725–732.
4. Freeling M (2009) Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol* 60: 433–453.
5. Chae L, Kim T, Nilo-Poyanco R, Rhee SY (2014) Genomic signatures of specialized metabolism in plants. *Science* 344:510–513.

6. Myburg AA, et al. (2014) The genome of *Eucalyptus grandis*. *Nature* 510:356–362.
7. Sollars ES, et al. (2017) Genome sequence and genetic diversity of European ash trees. *Nature* 541:212–216.
8. Albert VA, et al.; Amborella Genome Project (2013) The *Amborella* genome and the evolution of flowering plants. *Science* 342:1241089.
9. Soltis PS, Soltis DE (2016) Ancient WGD events as drivers of key innovations in angiosperms. *Curr Opin Plant Biol* 30:159–165.
10. Ibarra-Laclette E, et al. (2013) Architecture and evolution of a minute plant genome. *Nature* 498:94–98.
11. Carretero-Paulet L, et al. (2015) High gene family turnover rates and gene space adaptation in the compact genome of the carnivorous plant *Utricularia gibba*. *Mol Biol Evol* 32:1284–1295.
12. Veleba A, et al. (2014) Genome size and genomic GC content evolution in the miniature genome-sized family Lentibulariaceae. *New Phytol* 203:22–28.
13. Chin C-S, et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10:563–569.
14. Fulnecková J, et al. (2013) A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. *Genome Biol Evol* 5:468–483.
15. Tran TD, et al. (2015) Centromere and telomere sequence alterations reflect the rapid genome evolution within the carnivorous plant genus *Genlisea*. *Plant J* 84:1087–1099.
16. Melters DP, et al. (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* 14:R10.
17. Wade CM, et al.; Broad Institute Genome Sequencing Platform; Broad Institute Whole Genome Assembly Team (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* 326:865–867.
18. Nasuda S, Hudakova S, Schubert I, Houben A, Endo TR (2005) Stable barley chromosomes without centromeric repeats. *Proc Natl Acad Sci USA* 102:9842–9847.
19. Locke DP, et al. (2011) Comparative and demographic analysis of orang-utan genomes. *Nature* 469:529–533.
20. Liu Z, et al. (2008) Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. *Chromosoma* 117:445–456.
21. Cheng Z, et al. (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* 14:1691–1704.
22. Nagaki K, et al. (2005) Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. *Mol Biol Evol* 22:845–855.
23. Zhong CX, et al. (2002) Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. *Plant Cell* 14:2825–2836.
24. Hudakova S, et al. (2001) Sequence organization of barley centromeres. *Nucleic Acids Res* 29:5029–5035.
25. Gorinšek B, Gubenšek F, Kordiš D (2004) Evolutionary genomics of chromoviruses in eukaryotes. *Mol Biol Evol* 21:781–798.
26. Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* 8:272–285.
27. Topp CN, Zhong CX, Dawe RK (2004) Centromere-encoded RNAs are integral components of the maize kinetochore. *Proc Natl Acad Sci USA* 101:15986–15991.
28. Gao X, Hou Y, Ebina H, Levin HL, Voytas DF (2008) Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res* 18:359–369.
29. Lyons E, et al. (2008) Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol* 148:1772–1781.
30. Tang H, et al. (2015) SynFind: Compiling syntenic regions across any set of genomes on demand. *Genome Biol Evol* 7:3286–3298.
31. Tang H, et al. (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* 18:1944–1954.
32. Sankoff D, Zheng C, Zhu Q (2010) The collapse of gene complement following whole genome duplication. *BMC Genomics* 11:313.
33. Thomas BC, Pedersen B, Freeling M (2006) Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* 16:934–946.
34. Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci USA* 108:4069–4074.
35. Schnable JC, Wang X, Pires JC, Freeling M (2012) Escape from preferential retention following repeated whole genome duplications in plants. *Front Plant Sci* 3:94.
36. Cheng F, et al. (2012) Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLoS One* 7:e36442.
37. Garsmeur O, et al. (2014) Two evolutionarily distinct classes of paleopolyploidy. *Mol Biol Evol* 31:448–454.
38. Joyce BL, et al. (2017) FractBias: A graphical tool for assessing fractionation bias following polyploidy. *Bioinformatics* 33:552–554.
39. Soltis PS (2013) Hybridization, speciation and novelty. *J Evol Biol* 26:291–293.
40. Kameyama Y, Toyama M, Ohara M (2005) Hybrid origins and F1 dominance in the free-floating, sterile bladderwort, *Utricularia australis* f. *australis* (Lentibulariaceae). *Am J Bot* 92:469–476.
41. Chormanski TA, Richards JH (2012) An architectural model for the bladderwort *Utricularia gibba* (Lentibulariaceae). *J Torrey Bot Soc* 139:137–148.
42. Lampert KP, Schartl M (2008) The origin and evolution of a unisexual hybrid: *Poecilia formosa*. *Philos Trans R Soc Lond B Biol Sci* 363:2901–2909.
43. Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16:1679–1691.
44. Maere S, et al. (2005) Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci USA* 102:5454–5459.
45. Tang H, et al. (2011) Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* 12:102.
46. Lubkowitz M (2006) The OPT family functions in long-distance peptide and metal transport in plants.. *Genetic Engineering: Principles and Methods*, ed Setlow JK (Springer Science and Business Media, Berlin), Vol 27, pp 35–55.
47. Ibarra-Laclette E, et al. (2011) Transcriptomics and molecular evolutionary rate analysis of the bladderwort (*Utricularia*), a carnivorous plant with a minimal genome. *BMC Plant Biol* 11:101.
48. Schulze W, Frommer WB, Ward JM (1999) Transporters for ammonium, amino acids and peptides are expressed in pitchers of the carnivorous plant *Nepenthes*. *Plant J* 17:637–646.
49. Adlassnig W, et al. (2012) Endocytotic uptake of nutrients in carnivorous plants. *Plant J* 71:303–313.
50. Léran S, et al. (2015) AtNPF5.5, a nitrate transporter affecting nitrogen accumulation in *Arabidopsis* embryo. *Sci Rep* 5:7962.
51. Brownlee C (2013) Carnivorous plants: Trapping, digesting and absorbing all in one. *Curr Biol* 23:R714–R716.
52. Stigter KA, Plaxton WC (2015) Molecular mechanisms of phosphorus metabolism and transport during leaf senescence. *Plants (Basel)* 4:773–798.
53. Nussaume L, et al. (2011) Phosphate import in plants: Focus on the PHT1 transporters. *Front Plant Sci* 2:83.
54. Rottloff S, et al. (2011) Functional characterization of a class III acid endochitinase from the traps of the carnivorous pitcher plant genus, *Nepenthes*. *J Exp Bot* 62:4639–4647.
55. Goujon T, et al. (2003) AtBXL1, a novel higher plant (*Arabidopsis thaliana*) putative beta-xylosidase gene, is involved in secondary cell wall metabolism and plant development. *Plant J* 33:677–690.
56. Hatano N, Hamada T (2008) Proteome analysis of pitcher fluid of the carnivorous plant *Nepenthes alata*. *J Proteome Res* 7:809–816.
57. Rottloff S, et al. (2016) Proteome analysis of digestive fluids in *Nepenthes* pitchers. *Ann Bot (Lond)* 117:479–495.
58. Erni P, Varagnat M, Clasen C, Crest J, McKinley GH (2011) Microrheometry of sub-nanolitre biopolymer samples: Non-Newtonian flow phenomena of carnivorous plant mucilage. *Soft Matter* 7(22):10889–10898.
59. Vintéjoux C, Shoar-Ghafari A (1997) Sécrétion de mucilages par une plante aquatique. *Acta Bot Gallica* 144(3):347–351.
60. Poppinga S, Weisskopf C, Westermeier AS, Masselter T, Speck T (2015) Fastest predators in the plant kingdom: Functional morphology and biomechanics of suction traps found in the largest genus of carnivorous plants. *AoB Plants* 8:plv140.
61. Skotheim JM, Mahadevan L (2005) Physical limits and design principles for plant and fungal movements. *Science* 308:1308–1310.
62. Forterre Y (2013) Slow, fast and furious: Understanding the physics of plant movements. *J Exp Bot* 64:4745–4760.
63. Llorens C, Argentina M, Bouret Y, Marmottant P, Vincent O (2012) A dynamical model for the *Utricularia* trap. *J R Soc Interface* 9:3129–3139.
64. Li Y, Jones L, McQueen-Mason S (2003) Expansins and cell growth. *Curr Opin Plant Biol* 6:603–610.
65. Campbell P, Braam J (1999) Xyloglucan endotransglycosylases: Diversity of genes, enzymes and potential wall-modifying functions. *Trends Plant Sci* 4:361–366.
66. Yadav S, Yadav PK, Yadav D, Yadav KDS (2009) Pectin lyase: A review. *Process Biochem* 44:1–10.
67. Humphrey TV, Bonetta DT, Goring DR (2007) Sentinels at the wall: Cell wall receptors and sensors. *New Phytol* 176:7–21.
68. Zonia L, Munnik T (2007) Life under pressure: Hydrostatic pressure in cell growth and function. *Trends Plant Sci* 12:90–97.
69. Micheli F (2001) Pectin methylesterases: Cell wall enzymes with important roles in plant physiology. *Trends Plant Sci* 6:414–419.
70. Schulze WX, et al. (2012) The protein composition of the digestive fluid from the Venus flytrap sheds light on prey digestion mechanisms. *Mol Cell Proteomics* 11: 1306–1319.
71. Renner T, Specht CD (2013) Inside the trap: Gland morphologies, digestive enzymes, and the evolution of plant carnivory in the Caryophyllales. *Curr Opin Plant Biol* 16:436–442.
72. Libiaková M, Floková K, Novák O, Slováková L, Pavlovič A (2014) Abundance of cysteine endopeptidase dionain in digestive fluid of Venus flytrap (*Dionaea muscipula* Ellis) is regulated by different stimuli from prey through jasmonates. *PLoS One* 9:e104424–e104424.
73. Jensen MK, et al. (2015) Transcriptome and genome size analysis of the Venus flytrap. *PLoS One* 10:e0123887.
74. Bemm F, et al. (2016) Venus flytrap carnivorous lifestyle builds on herbivore defense strategies. *Genome Res* 26:812–825.
75. Butts CT, Bierma JC, Martin RW (2016) Novel proteases from the genome of the carnivorous plant *Drosera capensis*: Structural prediction and comparative analysis. *Proteins* 84:1517–1533.
76. Butts CT, et al. (2016) Sequence comparison, molecular modeling, and network analysis predict structural diversity in cysteine proteases from the Cape sundew, *Drosera capensis*. *Comput Struct Biotechnol J* 14:271–282.
77. Risør MW, et al. (2015) Enzymatic and structural characterization of the major endopeptidase in the Venus flytrap digestion fluid. *J Biol Chem* 291:2271–87.
78. Mateos JL, et al. (2015) Combinatorial activities of SHORT VEGETATIVE PHASE and FLOWERING LOCUS C define distinct modes of flowering regulation in *Arabidopsis*. *Genome Biol* 16:31.
79. Gregis V, et al. (2013) Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in Arabidopsis. *Genome Biol* 14:R56.
80. Todd J, Post-Beittenmiller D, Jaworski JG (1999) KCS1 encodes a fatty acid elongase 3-ketoacyl-CoA synthase affecting wax biosynthesis in *Arabidopsis thaliana*. *Plant J* 17:119–130.
81. Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. *Genome Biol* 5:225.
82. Delcher AL, Salzberg SL, Phillippy AM (2003) Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics* Chapter 10: Unit 10.13.
83. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.

Supporting Information Appendix

Long-Read Sequencing Uncovers the Adaptive Topography of a Carnivorous Plant Genome

Tianying Lan, Tanya Renner, Enrique Ibarra-Laclette, Kimberly M. Farr, Tien-Hao Chang, Sergio Alan Cervantes-Pérez, Luis Herrera-Estrella, Chunfang Zheng, David Sankoff, Haibao Tang, Rikky W. Purbojati, Alexander Putra, Daniela I. Drautz-Moses, Stephan C. Schuster, Victor A. Albert

Table of Contents

1. PacBio Sequencing and Assembly of the *Utricularia gibba* genome

1.1. Plant Material

As for our previous short-read genome assembly, *U. gibba* material was sourced from Umécuaro municipality, Michoacán, México, whereafter plants were grown in sterile tissue culture prior to nuclear DNA extraction.

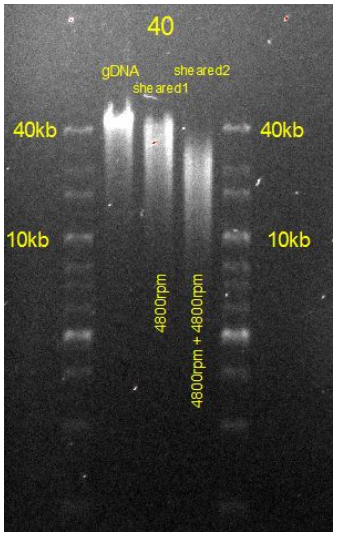
1.2. High Molecular Weight Nuclear DNA Preparation

In order to minimize chloroplast and mitochondrial DNA contamination, high molecular weight DNA was prepared from nuclei of *U. gibba* plants. Nuclear DNA was isolated according to the protocol described by the Mississippi Genome Exploration Laboratory (MGEL; http://www.mgel.msstate.edu/protocols.htm) based on Peterson et al. (1997) (1). The protocol was scaled-down to 10-15 g in order to decrease the amount of tissue required. In addition, isolated nuclei were collected from a 60% Percoll (Invitrogen) density gradient following low-speed centrifugation (4000g for 10 min at 4°C), after which high-quality megabase-sized DNA was isolated.

1.3. PacBio SMRT Sequencing

The quality of the *U. gibba* genomic DNA sample was first assessed by running 150 ng of DNA on a 0.6% pulsed-field agarose gel, stained with SYBR Safe (Invitrogen). Ten µg of DNA were then sheared to a size range of 10-40 kb using a Covaris g-TUBE. The fragment distribution of the sheared sample was validated by pulsed-field gel electrophoresis (Fig. S1A). The sheared DNA was then purified with 0.45X AMPure PB beads (Pacific Biosciences) according to the manufacturer's recommendations. Following the Pacific Biosciences (PacBio®) 20 kb SMRTbell Template Preparation Protocol, library preparation was subsequently performed using 5 µg of the sheared DNA as input. After library preparation, the library was assessed on an Agilent DNA 12000 bioanalyzer chip to determine the optimal cut-off for size selection (Fig. S1B). The libraries were then size-selected on a Sage Science BluePippin instrument using a dye-free 0.75% agarose cassette and 15 kb as the cut-off, followed by reanalysis with the bioanalyzer (Fig. S1C). For *U. gibba*, two libraries were prepared (internal IDs 40a and 40b). Library 40a was sequenced in two SMRTcells on a Pacific Biosciences RSII single-molecule sequencing platform at loading concentrations of 0.15nM and 0.2nM, respectively. Library 40b was sequenced in 8 SMRTcells at a loading concentration of 0.2nM.

A



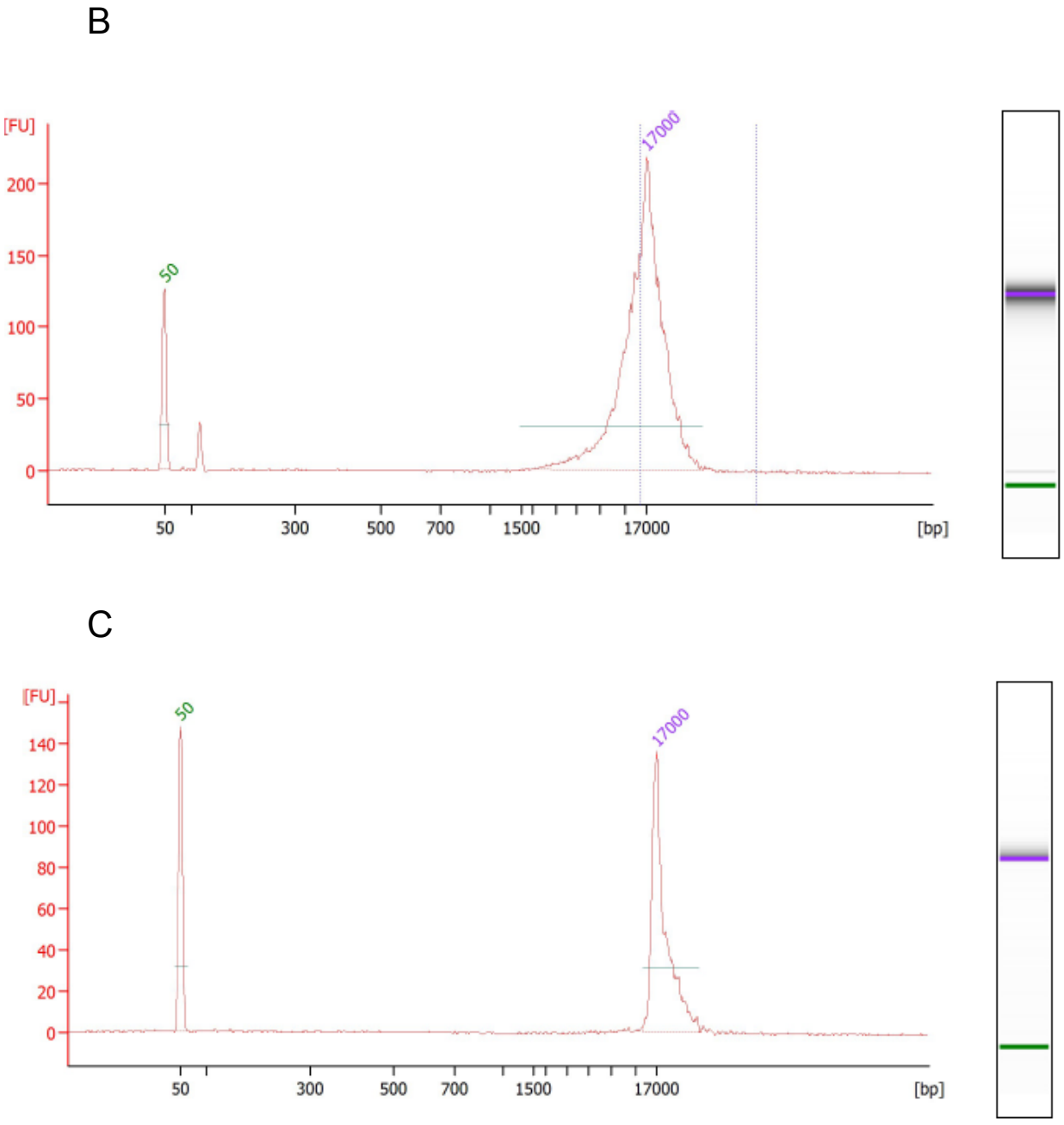


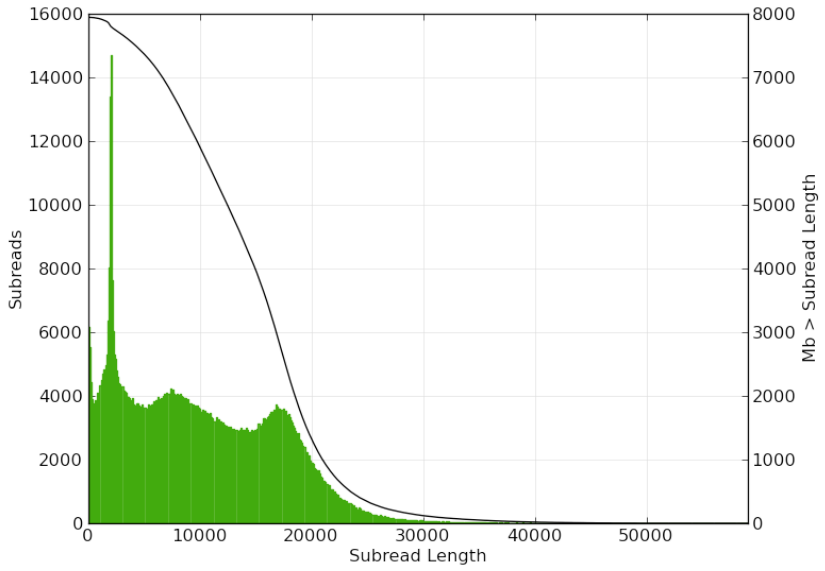
Fig. S1. (A), *U. gibba* gDNA before and after two rounds of shearing (sheared1 and sheared2) using a Covaris g-TUBE. Agilent bioanalyzer profiles for sample 40a, before and after Blue Pippin treatment, are shown in **(B)** and **(C)**, respectively. The spike-in control lies at 50 bp.

1.4. HGAP *De Novo* Genome Assembly

The HGAP3 workflow was used to assemble the raw h5 reads from 10 SMRT cells. The workflow used the software binaries from the SMRT analysis 2.3 package (http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/), and smrtmake (https://github.com/PacificBiosciences/smrtmake) was used as a wrapper for running the HGAP3

workflow. For the initial reads filtering, the "*MinReadScore=0.80, MinSRL=500, MinRL=100*" settings were used. Once the reads were filtered, the workflow attempted to detect overlap between reads. This step was done with the *"-bestn 10 -nCandidates 10 -noSplitSubreads -maxScore -1000 -maxLCPLength 16 -minMatch 14*" parameters. The result of this overlapping step was used to produce a set of corrected reads. Afterward, the corrected reads were used in the assembly process. The HGAP3 parameters used for this step were "*genomeSize = 90000000, xCoverge = 20, defaultFrgMinLen = 500, ovlErrorRate = 0.06, ovlMinLen = 40, merSize = 14*". The genome produced by HGAP3 consists of 581 contigs with N50 of 3,424,836 bp and 101,949,210 total bases (Fig. S2). We used this *de novo* assembly for all downstream analyses. A previous flow cytometry analysis estimated a genome size of 77Mb for *U. gibba*. However, the estimation was carried out using Golden Path *Arabidopsis* genome size (1C = 0.1605 pg or 135 Mb) as an internal standard calibration (2). In fact, *Arabidopsis* lines were proved to exhibit massive genome size variation, ranging from 161 to 184 Mb (3). Consequently, using this *Arabidopsis* genome size range as calibration, the flow cytometry estimated *U. gibba* genome size would be from 92.3 to 105.5 Mb, which is nicely consistent with our current PacBio genome size ~102 Mb.

A



B

	Raw Subreads (Post-Filter Polymerase Reads)	Filtered Subreads	Corrected Reads	Assembly
#Reads	521,937	702,640	155,051	-
#Contigs:	-	-	-	581
Average Len:	15,366	10,385	12,322	175,774.50
Total Bases:	8,020,277,606	7,297,249,484	1,910,594,344	101,949,210.00
Max Len:	59,061	59,061	39,438	8,502,017.00
N50:	21,825	15,244	17,834	3,424,836.00

Fig. S2. Reads and assembly statistics for the *de novo U. gibba* PacBio genome assembly. (A) Raw reads distribution plot. (B) Reads and assembly statistics.

1.5. Testing Genome Assembly Quality Using PacBio and MiSeq Read Data

To test for assembly quality at the nucleotide level, we also polished it using the Quiver workflow (4). The parameters used for this step were *"--forQuiver --seed=1 --minAccuracy=0.75 --minLength=50 --algorithmOptions="useQuality -minMatch 12 -bestn 10 -minPctIdentity 70.0" --hitPolicy=randombest*". At this step, 72,485 (0.071%) bases were corrected out of 101,810,468 bp total. In addition to the Quiver polishing, a set of Illumina MiSeq data (NCBI BioSample SAMN01940705, from (2)) was also used to provide a measure of assembly quality. A total of 3,120,016,344 base pairs reads (~39x coverage) were used with Pilon (5) to solve any discrepancies between the HGAP3 assembly and the MiSeq data. The settings that were used for this step were "*--fix all, --mindepth 0.1, --mingap 10, --K 47*". After this step, only 11,053 (0.01%) bases were corrected out of 101,203,230 total, which together with the Quiver statistics verifies the high quality of our PacBio assembly.

1.6. Identification of Contamination and Organellar Contigs

1.6.1. Identification of Environmental Sequence Contamination

Contigs with length < 1 Mb were blasted against the NCBI refseq non-redundant nucleotide database using NCBI blastn (6) v.2.2.30+ with an *E*-value threshold of 1E-5. Contigs with non-plant matches were determined to be environmental sequence contamination. In total, 65 contigs (1,260,662 bp) representing 0.01% of the genome were identified as bacterial contaminants. Among them, 62 contigs (1,171,157 bp) originated from *Methylobacterium*, a common bacterial contaminant of DNA extraction kit reagents (7).

1.6.2. Identification of Plastid and Mitochondrial Genome Contigs

Organelle DNA contigs were identified by: (1) blastn searches against the NCBI refseq non-redundant nucleotide database and, (2) using all available NCBI *U. gibba* organelle DNA from (2) as blastn queries against the new PacBio assembly with an *E*-value threshold of 1E-5. The chloroplast genome was assembled into a single 172,489 bp contig, which gained 20,376 bp compared to the previous 152,113 bp circular assembly (NC_021449.1; (2)). However, further investigation of the linear plastid contig showed that there are two identical terminal regions, while there is only one corresponding region in the previous plastid genome. Therefore, we speculate that the extra sequence in current chloroplast genome is likely misassembled and caused by using a linear *de novo* assembly strategy to assemble a circular genome. Our previous mitochondrial genome assembly constituted ten unique scaffolds/contigs with a combined total length of 222,145 bp (2), whereas the PacBio mitochondrial genome assembled into a single 283,823 bp contig. Due to the dynamic intramolecular recombination of mitochondrial DNA in plant cells, it is common to obtain multiple partially overlapping mitochondrial chromosomes in *de novo* assemblies (8-10).

2. Annotation

2.1. TE annotation and Repeat Masking

Prior to gene prediction, the REPET pipeline v2.2 (11, 12) was used for the *de novo* detection and annotation of transposable elements (TEs). The REPET TE annotation process was carried out in two phases: (i) *de novo* discovery and identification of TE families present in the genome studied, and (ii) the precise, comprehensive annotation of TE copies on contigs or scaffolds. TEs were predicted based on their typical features, and their quality was assessed by the extent to which full ancestral TE reference sequences were recovered andbased on their similarities to high-quality sequences available in public databases (Repbase (13) in this study). TE classes were classified based on Wicker's classification. (14) The annotation pipeline was performed for both the current PacBio assembly and our previous 454/Illumina/Sanger hybrid assembly (2). The number of annotated TEs, their lengths (range and average), the total bases contained, and the percentages that they represent of total genome space were calculated (Dataset S1). Considering the high recombination rates previously suggested to be linked with

DNA loss in *U. gibba* genomic regions with relaxed selective pressures (2), repeat masking was subsequently performed with RepeatMasker (http://www.repeatmasker.org/) using the library generated by REPET. Option -s was used (slow search; 0-5% more sensitive, 2-3 times slower than default), as was –cutoff. Cutoff score describes the overall quality of the alignments, with higher numbers corresponding to higher similarities for masking repeats with a custom library; a score of 250 was used in this study. This repeat masking strategy was also employed for both the current PacBio assembly and the previous 454/Illumina/Sanger hybrid assembly. Finally, the percentages of genome space assigned to TEs and TE-like regions in both genome assemblies were calculated, and the comparison is illustrated in Dataset S1 and Fig. S3. In summary, a total of 1,121 repetitive sequences were identified in our PacBio assembly, almost 3.5 times more TEs than identified in the published 454/Illumina/Sanger hybrid assembly (2). The percentage assigned to TEs plus TE-like regions in the current assembly is around 59%, which is about 16.5% more than that of the previous assembly. Complete TEs amounted to 8.9% and 2.3%, respectively. These figures can be compared with the 3% reported by Ibarra-Laclette et al. for the latter genome (2). One of the retrotransposon classes, the LARD class (Large Retrotransposon Derivatives), represents around 47% of total TE/TE-derived space identified in the current assembly. Interestingly, this class occupied only about 14.6% of TE/TE-derived space in the published short-read assembly. Increased identification of LARDs in the current assembly results from the capability of long read sequencing to obtain sequences from complex genomic regions with highly repetitive elements and condensed DNA structure (e.g., centromeres and telomeres).

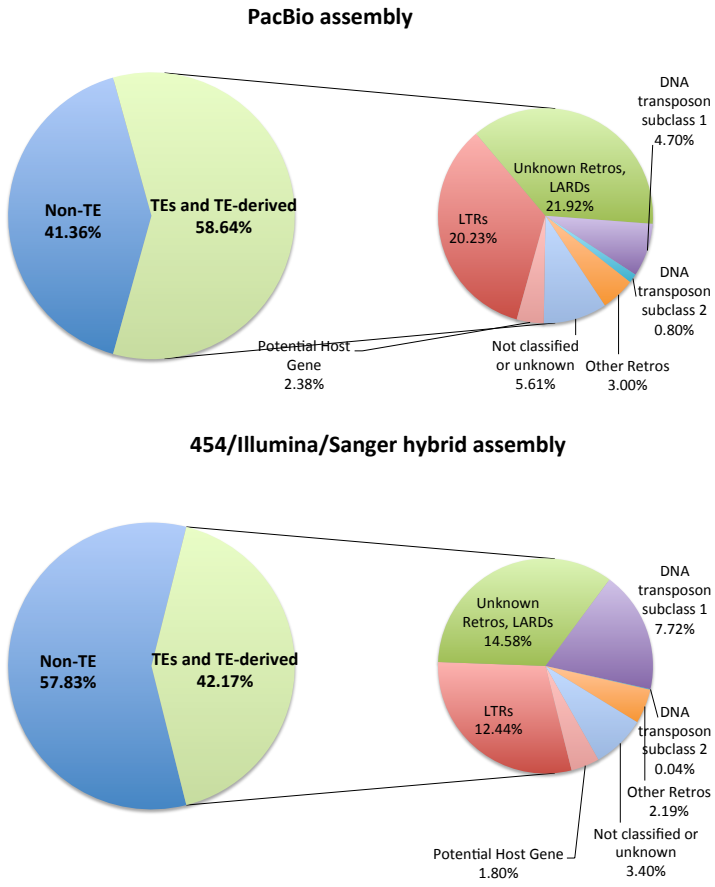


Fig. S3. Comparison of TE composition between the PacBio assembly and the 454/Illumina/Sanger hybrid assembly (2).

2.2. Non-coding RNA (ncRNA) Annotation

rRNA and tRNAs genes were identified using RNAmmer (15) and tRNAScan-SE (16), respectively. Other predicted ncRNA elements, including miRNAs, snRNAs and H/ACA box snoRNAs, were identified by the INFERNAL software (17) using the RFAM database (18, 19) records as models.

2.3. Identification of Protein-coding Genes

Gene models were predicted with an evidence-directed AUGUSTUS predictor (20). AUGUSTUS was trained for *U. gibba* gene parameters using full coding sequences (CDS) derived from gene models predicted for the previous version of *U. gibba* genome (2). This *U. gibba*-specific training set was generated through the Augustus training web interface (21) using only those CDS for which the gene model previously predicted represented "complete CDS", i.e., when proteins derived from the predicted genes represented at least 75% of the length of homologous proteins identified in related asterid species such as *Solanum lycopersicum* and *Mimulus guttatus*. AUGUSTUS was run for genomic PacBio contigs using default parameters, unigenes derived from transcriptome assemblies (2, 22), and those *U. gibba*-specific parameters calculated from the training set. In addition, gene models predicted by AUGUSTUS were complemented using the Maker-P pipeline (23). Inputs for Maker-P included the *de novo* draft genome assembly of *U. gibba*, *U. gibba* transcriptome assemblies (2, 22), a species-specific repeat library predicted using REPET, protein databases containing annotated proteins for *S. lycopersicum* and *M. guttatus* (downloaded from the CoGe OrganismView database; http://genomevolution.org/CoGe/OrganismView.pl), and gene models predicted by AUGUSTUS on the previously published *U. gibba* genome (2).

2.4. Gene Annotation

All unmasked gene models were first blasted against the Viridiplantae Repbase database 21.02 (24) using NCBI tblastx v.2.2.30+ with an *E*-value threshold of 1E-5. Gene models with at least one match were considered to be (e.g., TE) repeat-associated genes and therefore removed from the set. All filtered gene models were then annotated with the highest alignment score matches using tblastx versus the *Arabidopsis* coding sequences database v10.02 with an *E*-value cutoff of 1E-5.

2.5. Identification of Islands of Repeat Elements Surrounding Certain Tandemly Duplicated Gene Clusters

Some tandemly duplicated genes were difficult to predict with prior masking of repeat elements. Therefore, we performed gene model prediction without masking to discover a number of the tandem duplicate cases referred to below (e.g., sections 6.1 and 6.2, below). As part of our RepeatMasker survey (section 2.1), we were able to determine that incomplete recognition of tandem arrays were in some cases due to genes being surrounded by TEs annotated as LARDs (Large Retrotransposon Derivatives (25)), as visible in the browser plots below (Fig. S4) for cysteine protease and KCS gene clusters (sections 6.1 and 6.2, respectively). We hypothesize that such repetitive DNAs (green) might have facilitated the tandem duplications that generated the arrays (blue gene models).

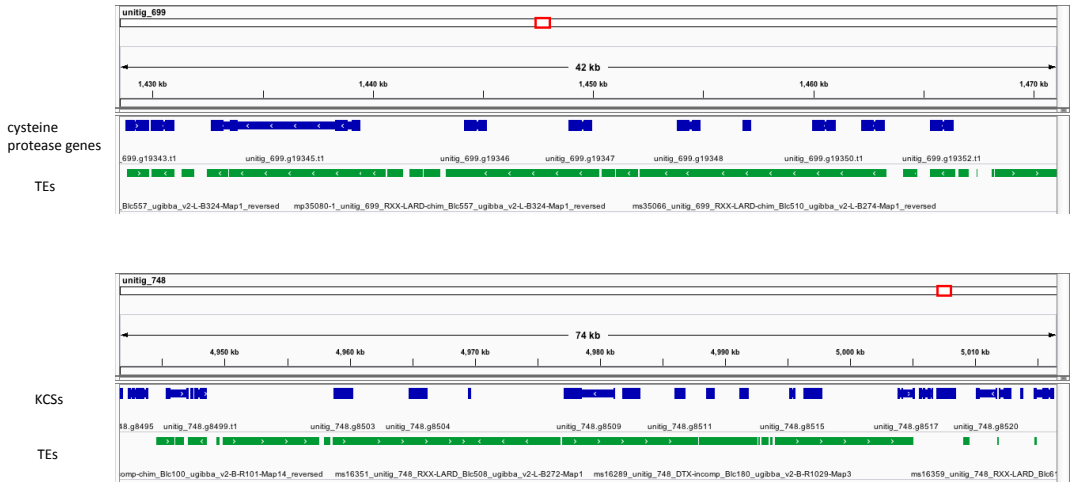


Fig. S4. IGV browser (26) plots showing cysteine protease and *KCS* gene clusters being surrounded by TEs annotated as LARDs.

2.6. Gene Expression Analysis

To investigate genes with trap-enhanced expression in *U. gibba*, 454 RNA-seq raw reads, which were sequenced from *U. gibba* shoot-like structures, traps and inflorescences as separate libraries in our previous study (27), were mapped to the PacBio genome using the subread-featureCounts pipeline (28, 29). The raw reads count table generated by featureCounts then served as input for the edgeR package (30) to calculate $\log_2$ fold change with model-based normalization, a minimum three reads filter, and the dispersion value 0.1 applied. The library size factors calculated by edgeR normalization were ~1 for all three libraries, which indicates that they have a similar library size. Comparisons were performed as traps versus inflorescences, and traps versus shoot-like structures, for which results are summarized in Dataset S2.

3. Identification of Centromeric and Telomeric Sequences

3.1. Identification of Tandem Repeats in the *U. gibba* Genome

We used Tandem Repeats Finder (31) (TRF) to search for tandem repeats in all contigs with alignment parameters 2, 7, and 7 for match, mismatch, and indels, respectively. Additionally, a minimum alignment score of 50 and a maximum period size of 2000 bp (https://tandem.bu.edu/trf/trf.html) were specified. The initial raw TRF output, which included TR arrays with overlapping genomic coordinates, was then processed with a redundancy elimination python script. The redundancy criterion was set to remove the repeat with a larger period size if two repeats shared 90% overlap. However, if the larger repeat had over a two-times greater alignment score, then the smaller repeat was removed. After redundancy elimination, 12,378,680 bp of the genome were annotated as tandem repeats. As shown in Fig. S5A, among repeats, size ranged from 0-600 bp, with the most abundant tandem repeats in the genome being the micro/minisatellite fraction, which has a period size smaller than 50 bp. The other abundant tandem repeats were distributed around 300-360 bp and 520-540 bp. We then performed an all versus all alignment using USEARCH global (32) with settings for 'global' alignment and 95% identity threshold. Using a clustering python script, 113 global clusters containing tandem repeats with >90% identity and near identical lengths were produced. Further analysis on tandem repeat clusters was performed for centromeric repeats screening (See section 3.3.2 for details).

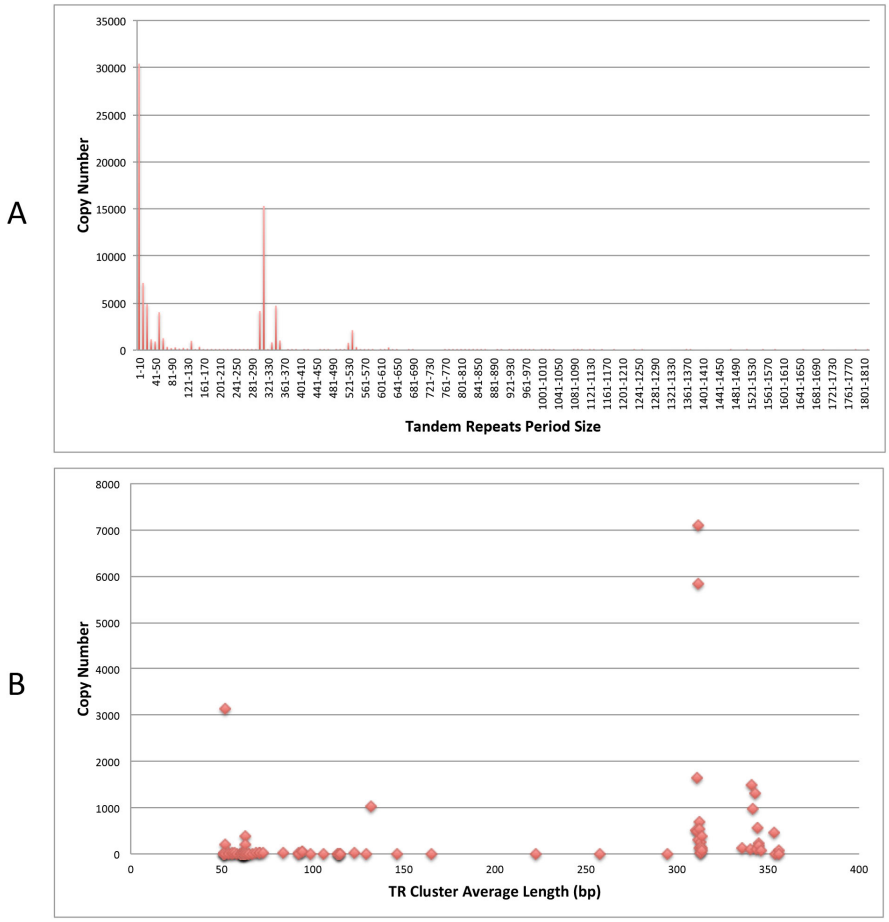


Fig. S5. Distribution plots of tandem repeats and repeat clusters in the *U. gibba* genome. **(A)** The period size and copy number of the tandem repeats identified using TRF. **(B)** The average length and copy number of tandem repeat clusters that had average length < 500 bp.

3.2. Identification of Telomeres

By searching the ends of contigs, high copy numbers of *Arabidopsis*-type telomeric repeats (TTTAGGG) were identified in 24 contigs.s. Two variants - the *Chlamydomonas*-type (33) (TTTTTAGGG) and and a novel type (TTCAGGG, similar to the variants TTCAGG and TTTCAGG known from the close carnivorous plant relative *Genlisea* (34)) - were also found sporadically intermingled with the *Arabidopsis*-type telomeric repeats. Four contigs, for which telomeric repeats were found on both ends, were identified as complete chromosomes. Ten contigs were observed to have internal telomeric repeats, which were identified by searching $(CCCTAAA)_3$ and $(TTTAGGG)_3$ in interstitial regions.

3.3. Identification of Putative Centromeres

The centromere is a complex chromosome component that is responsible for chromosome segregation during meiosis and mitosis. One of the distinctive properties of centromeres is that they are enriched in repetitive elements, including transposable elements (TEs) and tandem repeats (35, 36). Due to the presence and abundance of these identical or near-identical repeats, centromeric regions are a bioinformatic challenge for NGS-based *de novo* genome assembly, and therefore they often remain incomplete and largely uncharacterized even within extensively sequenced and studied genomes (37-39). To resolve this challenge, long reads that exceed TE and tandem array length are needed to obviate

misassembly and allow repeats to be unambiguously placed based on unique flanking sequences. PacBio SMRT sequencing, which can generate read lengths up to ~ 20 kb, permitted us to assemble four complete and several near-complete chromosomes, thereby permitting a rare view of the highly repetitive nature of plant centromeres (10).

3.3.1. Identification of Putative Centromeric Regions

For comparing our previous short-read assembly (2) with the new PacBio assembly, a syntenic map was generated using the SynMap tool (40) from the CoGe platform (https://genomevolution.org/CoGe/SynMap.pl). The short-read assembly showed clear gaps within the complete or near-complete chromosomes of the PacBio assembly (Fig. S6). These gaps were the putative repeat-rich centromeric regions that failed to be assembled in the short-read assembly. Note that the term "centromeric" used in this context refers to both the centromeric and pericentromeric regions, as they are difficult to distinguish from one another.

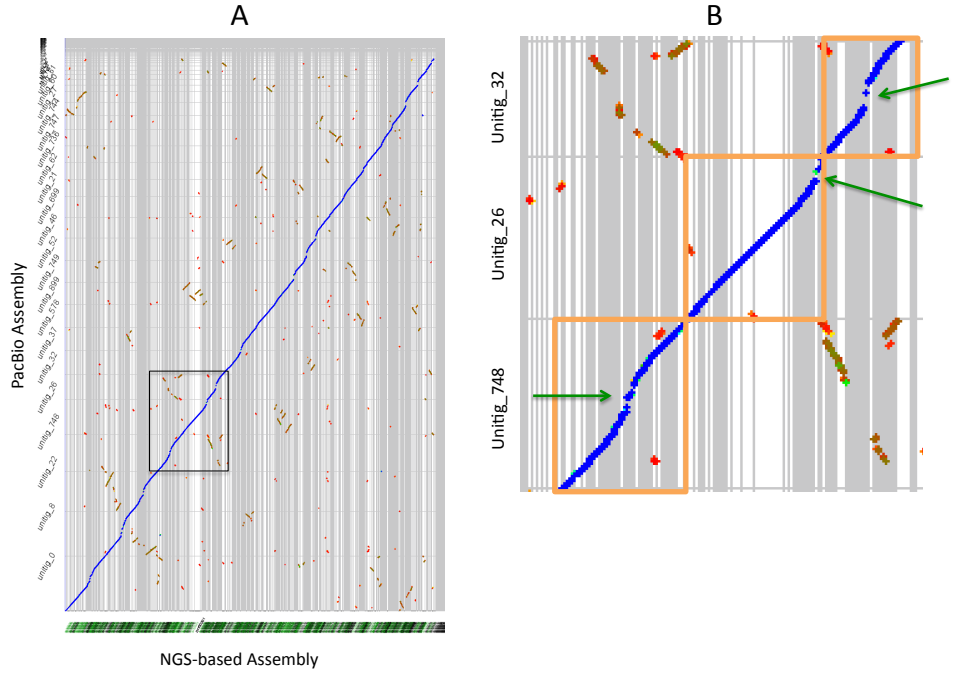


Fig. S6. Syntenic path alignment of the *U. gibba* short-read assembly (x-axis) against the PacBio assembly (y-axis), generated by SynMap in CoGe. Gray vertical and horizontal lines demarcate individual contigs/scaffolds from the short-read and PacBio assemblies, respectively. **(B)** The enlarged detail is encompassed within the black rectangle in **(A)**. Chromosomes are boxed in orange, and it is readily apparent that the x-axis dimensions of these boxes are narrower than their y-axis dimensions, indicating missing DNA in the short-read assembly. Putative centromeric regions lie at gaps in assembly cross-match (highlighted by green arrows), where short scaffolds from the short-read assembly are compacted. Furthermore, the "swooping" syntenic lines composed of homologous gene model matches (blue dots) are caused by the increasing absence of assembled repetitive DNA in the short-read genome toward the centromere (41). Green-red dots illustrate internally syntenic regions of both U. gibba genome assemblies (see section 4 for details).

To further investigate putative centromeric regions, we performed pairwise chromosomal alignments on the four complete chromosomes using MUMmer (42). TE families and other repetitive elements that accumulated in centromeric regions would show increased homology in the pairwise alignment. As expected, the putative centromeric regions were highlighted by increased dot density, whereupon the

boundaries of putative centromeres were estimated (Fig. S7). These regions were also confirmed as TE-rich and gene-poor regions through comparison to our repetitive element annotation and gene model prediction pipelines (Fig. S8).

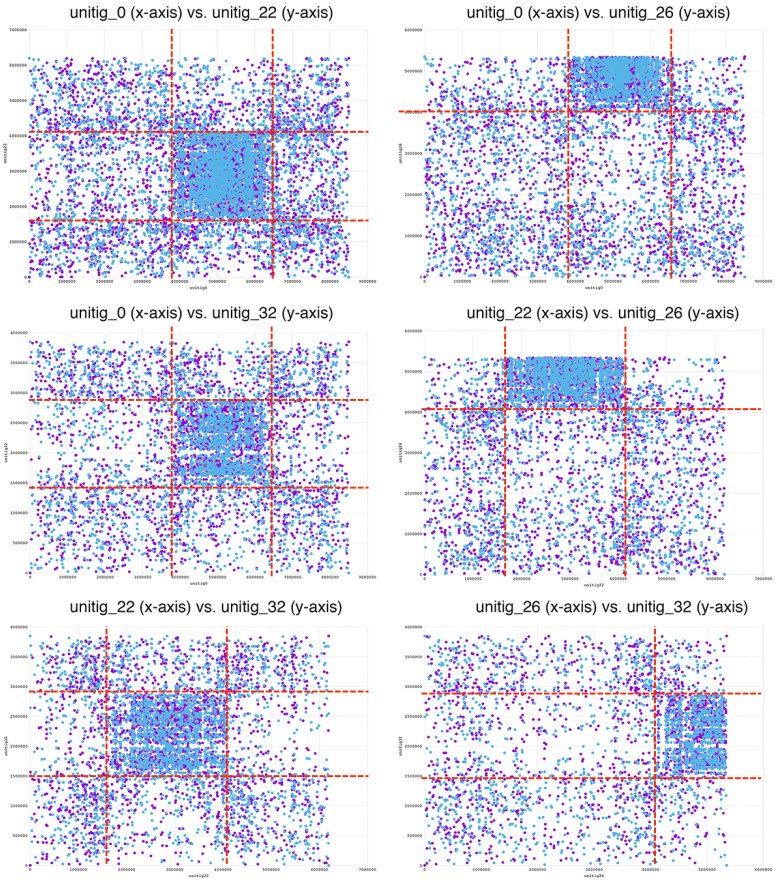


Fig. S7. Pairwise chromosomal alignment plots generated by MUMmer. Red dashed lines indicate the estimated boundaries of putative centromeric/pericentromeric regions.

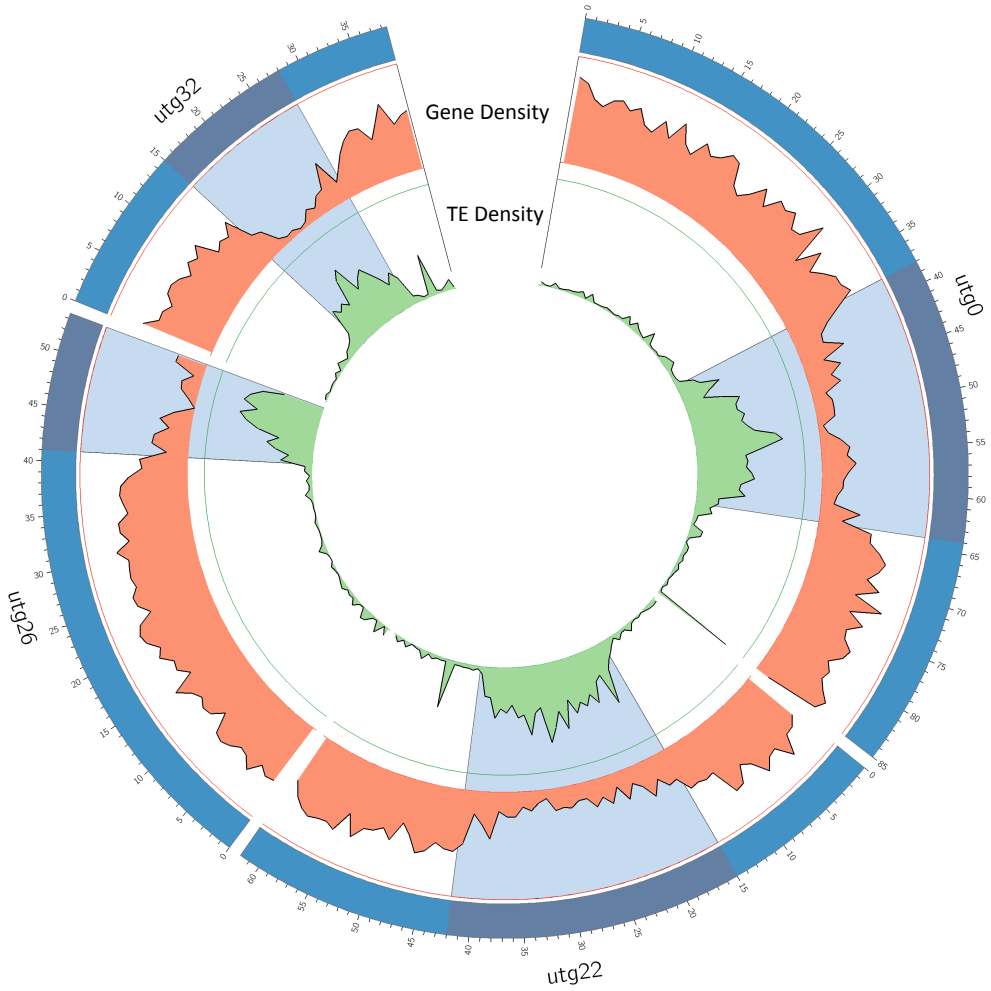


Fig. S8. Visualization of MUMmer-detected putative centromeric regions with TE and gene density tracks in a Circos (43) plot.

3.3.2. Centromeric Repeats Screening

Various arrays of simple tandem repeats are prevalent at the centromeric regions of plant and animal genomes (44-46), yet neither the tandem repeat monomer length nor the repeat sequences are conserved among species that diverged more than 50 MYA (36). In order to identify the signature centromeric repeats in *U. gibba*, tandem repeat clusters with average period size of 50-500 bp were selected for identification as putative centromere repeats (Fig. S5B) as in (36). The top 10 most abundant tandem repeat clusters were considered to be prime candidates for centromeric repeats, but these were not even preferentially located in our chromosome-sized contigs. We then manually checked the locations of the next 10 most abundant tandem repeat clusters in the genome; however, none of these clusters showed unique localization in putative centromeric regions. We thus speculated that *U. gibba* may be devoid of high-copy tandem repeat arrays in its centromeres. Similar findings have also been reported for the centromeres of several plant and animal species (47-49), including two carnivorous plant species, *Genlisea hispidula* and *G. subglabra* (34).

3.3.3. Identification of Putative Centromeric CRM Retrotransposons

While plant retrotransposon families are in general randomly dispersed, there are families distinctly concentrated in centromeric regions. Centromeric retrotransposons, CRMs, which locate preferentially in centromeric regions, are among the latter category. CRM chromoviruses, a lineage of Ty3/gypsy retrotransposons, have been well characterized as centromeric retrotransposons in many species (50-55), including *G. hispidula* and *G. subglabra* (34). CRM elements can be categorized into three subgroups, of which subgroup A and B are concentrated in centromeric regions (56). The RT domain is a reliable component to discriminate CRM elements from other LTR retrotransposons (57, 58).

Our search strategy for CRMs in *U. gibba* was carried out as follows. First, REPET-annotated TE families of *U. gibba* were queried with CRM sequences from 33 species (56) using blastn with an *E*-value threshold of 1E-5. The six resulting hits were processed with LTR_Finder (59), and only one among the six was identified as a full-sized LTR with an intact RT domain. We then searched the entire *U. gibba* genome using the ORF of the RT domain as a query with an *E*-value threshold of 1E-5. The resulting 55 hits distributed on 24 contigs were then extended by 15 kb both upstream and downstream to include the other portions of the LTRs, which were again processed with LTR_Finder for identification of full-sized LTRs and intact RT domains. As a result, 22 hits were identified as full-sized LTRs with an intact RT domain, while 11 hits were identified as full-sized LTRs with an incomplete RT domain, and 22 hits were identified as incomplete LTRs. The protein sequences of the RT ORFs from the 55 *U. gibba* hits and the 33 species were aligned using MUSCLE (60). Phylogenetic analysis based on the alignment was performed using RAxML-HPC BlackBox version 8.2.6 in CIPRES (61) with the GTR substitution model. A total of 1000 bootstrap relicates were conducted to evaluate branch support. In the maximum likelihood tree shown in Fig. S9, all 55 *U. gibba* sequences were grouped within the subgroup A CRMs, which include the centromere-specific CRMs. All but one of the *U. gibba* sequences together form a single, monophyletic CRM subfamily. To investigate the chromosomal localization of the 55 *U. gibba* CRMs, we plotted them on the complete and near-complete chromosomes together with the TE and gene model tracks. As depicted in Fig. 1 (main text), most *U. gibba* CRMs are located in the putative centromeric regions, however not all putative centromeres had CRM elements.

It has been proposed that CRMs may play an important role in stablizing centromere structure and maintaining centromere function (62, 63), while an opposing hypothesis holds that they are merely parasitic and tend to accumulate in recombination-poor centromeric regions to escape negative selection against insertions in distal regions (64). Our finding that the centromeric regions of several chromosomes lack CRMs, together with the finding that they also lack high-copy centromeric tandem repeats, suggests that neither CRMs nor tandem repeats are crucial for maintaining functional centromeres in *U. gibba*.

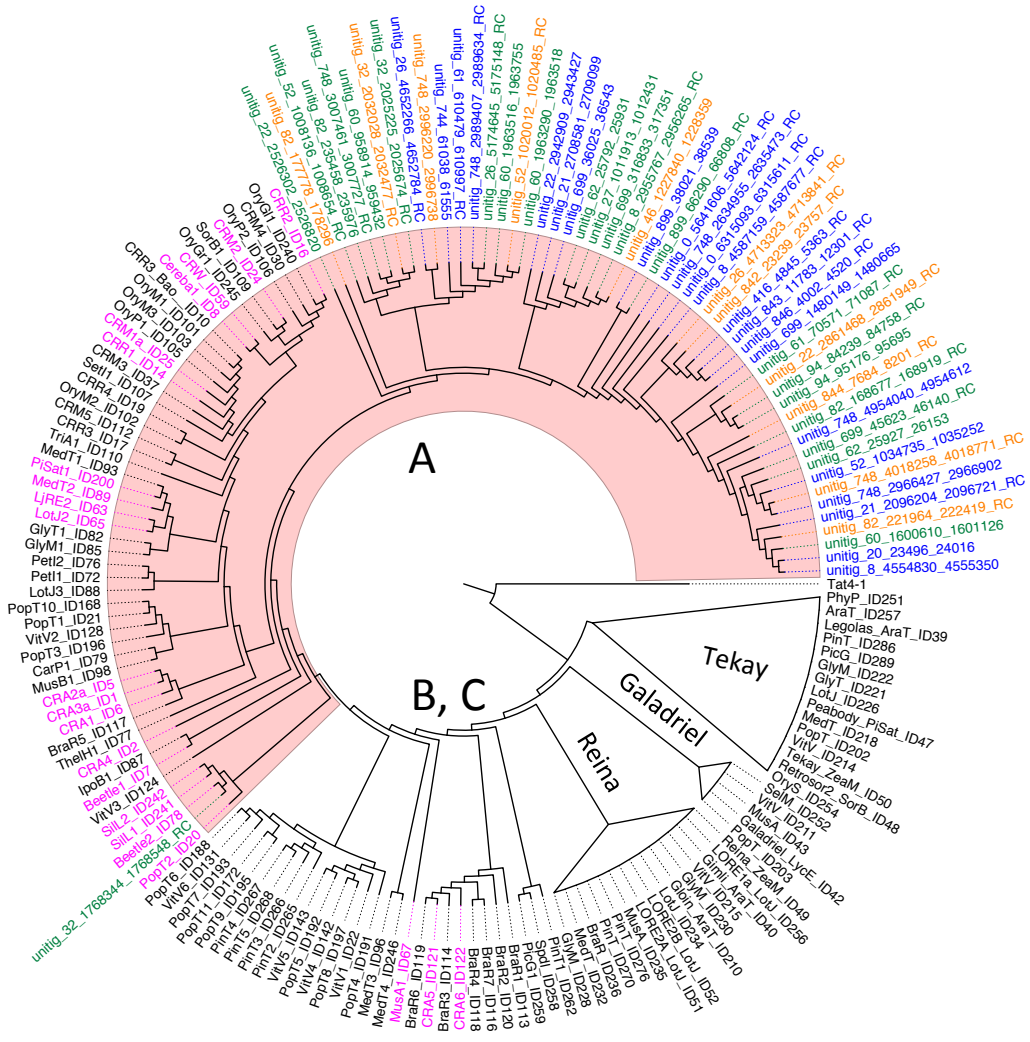


Fig. S9. Phylogenetic reconstruction based on the RAxML analysis of 55 *U. gibba* CRM-like blast hits and CRM reverse transcriptase (RT) domain sequences from (56). CRM subgroup A is highlighted in red. All 55 *U. gibba* sequences grouped with subgroup A CRMs. Full-sized *U. gibba* CRMs predicted by LTR_Finder to have intact RT domains are shown in blue, while the full-sized CRMs with incomplete RT domains and incomplete CRMs are shown in orange and green, respectively. CRM elements with previously confirmed centromeric localization are colored in purple. The collapsed Tekay, Reina and Galadriel clades were included as representatives of other non-CRM plant chromoviruses, and the non-chromovirus element Tat4-1 was used as an outgroup (56).

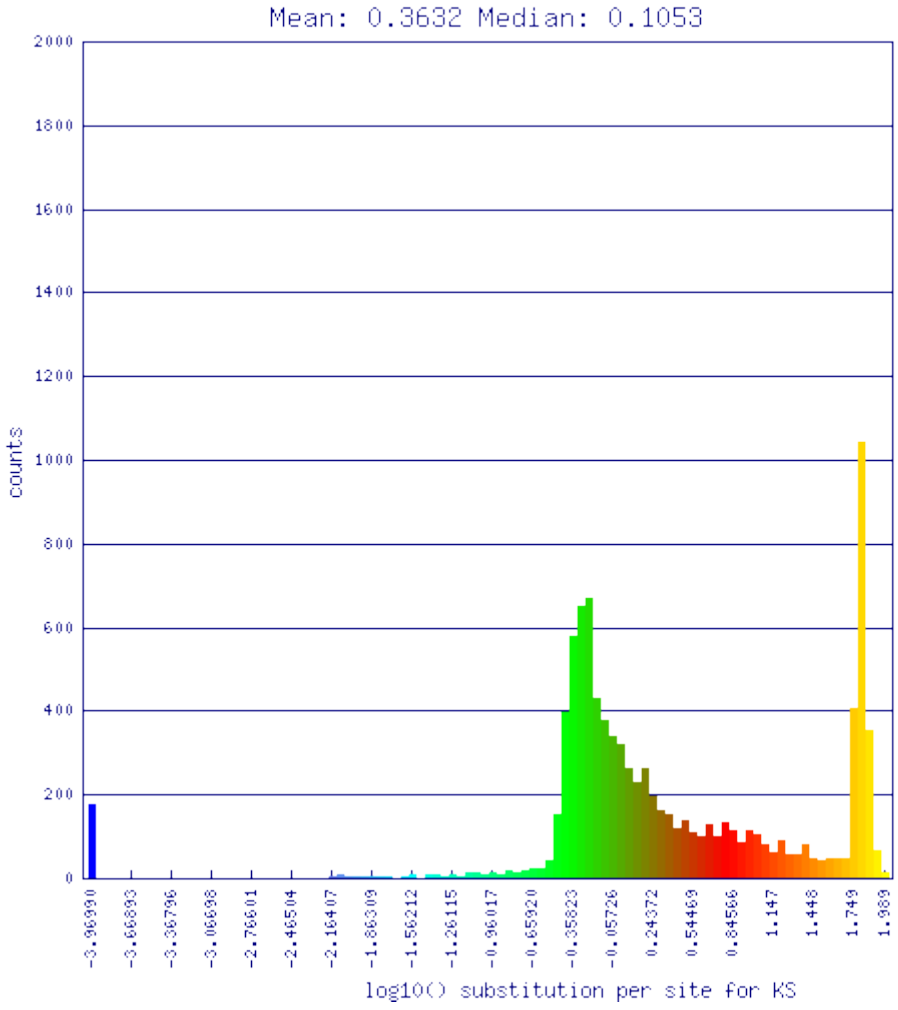
4. Ancestral Genome Reconstruction and Subgenome Dominance Analysis

4.1. Ancestral Genome Reconstruction

The analysis of the frequency distributions of duplicate gene similarities does not in the first instance involve syntenic considerations. Similarly, the detection of syntenic collinearity does not necessarily depend on similarity other than requiring identified gene pairs to be more similar than a given threshold. But there are many cases where combining the two kinds of data provides a powerful methodology. For example, in attempting to discern the evolution of a genome like *U. gibba* that has undergone at least 2

WGDs, the Ks distribution of similarities created by the more recent event tends to swamp and obscure that from an earlier event (Fig. S10), and the synteny of homeologous chromosomes is degraded, especially for the earlier event, by extensive fractionation and rearrangement.

A



B

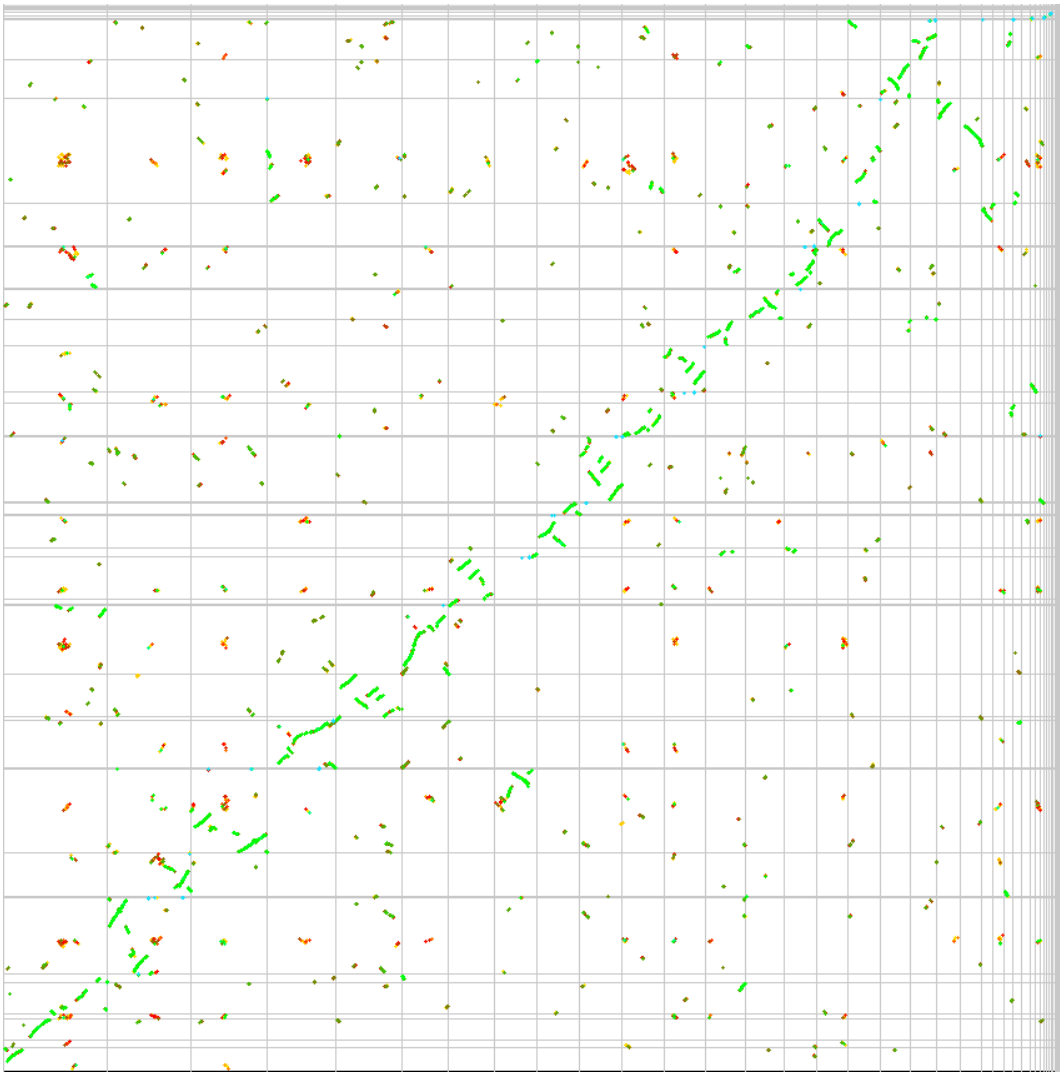


Fig. S10. (A) Distribution of Ks scores of paralogs in *U. gibba*, showing two overlapping distributions, resulting from a recent (light green) WGD event and an earlier (dark green - brown) WGD event. Red paralogous pairs at higher Ks largely represent regions that contain extensive tandemly duplicated genes. Given that Ks reasonably reflects the background (neutral) mutation rate, these red segments (i) may be older than the second-most recent WGD, (ii) they may lie in regions of enhanced mutation rate, or (iii) they may merely represent Ks saturation artifacts (65). Substitutional and insertional mutation rates are known to covary (66), so these regions could be hotspots for both nucleotide mutation and local (tandem) insertional duplication. The yellow peak represents irrational Ks values due to codon misalignments; these are commonly observed in CoGe SynMap results. **(B)** Syntenic dotplot from which **(A)** was derived; the *U. gibba* contigs are ordered with respect to each other (in a syntenic path alignment) such that the syntenic blocks from the most recent WGD event are forced as best possible to the diagonal.

Thus our analysis of WGD in *U. gibba* integrates both kinds of data. The first step is to describe and delimit the most recent event using synteny block construction provided by the SynMap package on the CoGe platform (40, 67), using default values for the parameters, and filtering out gene pairs with Ks

outside the interval [-0.6, 0.25]. Neighboring synteny blocks on the same two chromosomes were combined to produce a total of 54 pairs of (larger) blocks, containing 81% of the genes in the genome. These fell into six groupings, where each group contains from one to thirteen homeologous pairs of blocks. Five of these groups consisted essentially of two *U. gibba* chromosomes, either whole or fissioned into two pieces, or containing small translocations (Fig. S11).

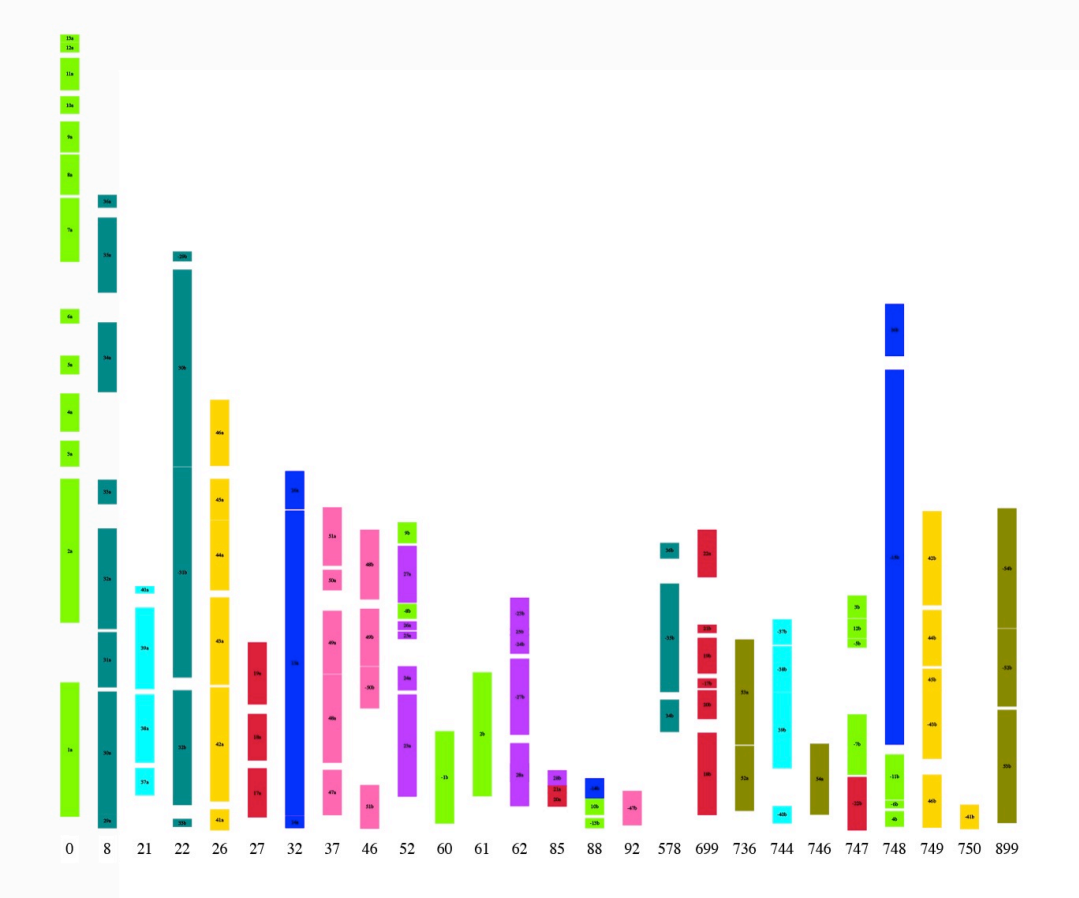


Fig. S11. Nine color-coded homeologous blocks in their current positions in the *U. gibba* karyotype.

In the sixth group, the largest *U. gibba* chromosome can be recognized as a fusion of four ancient chromosomes whose homeologs form all or part of various of the other extant chromosomes. A completely duplicated genome could then be recognized as underlying these nine groups, differing by a small number of readily identified rearrangements (Fig. S12).

Fig. S12. Reconstruction of *U. gibba* genome structural evolution through two rounds of WGD and subsequent diploidization. Internally syntenic blocks stemming from two ancient polyploidy events were identified in the highly contiguous genome assembly and tracked through multiple translocation, fusion, fission, and especially inversion events (diagramed at bottom) from two ancestral genomes, first an n = 6 chromosome pre-polyploid ancestor, and thereafter an n = 9 ancestor. Panels **(1)-(7)** show the structural rearrangement history of 54 syntenic blocks identified among the modern genome contigs **(7)**, with colors matching ancestral chromosomes. Numbers indicate block identities; "a" versus "b" represents subgenome pairs included in the most recent WGD event, which fractionation and expression and expression data suggest to have been an allopolyploidization. "-" indicates inverted orientation, and underscore between blocks in the n = 6 ancestor link blocks from the second WGD to those of the first.

In reconstructing the earlier WGD event, we found that the SynMap of the *U. gibba* genome against itself did not produce enough paralogous synteny blocks based on gene pairs with Ks levels around 0.25 or above. Since evolution after speciation retains WGD-generated orthologous pairs to a far greater extent than the paralogous pairs generated by fractionation after WGD, we used a conservative core eudicot genome – *V. vinifera* - to detect 113 sets of separate regions in *U. gibba* syntenic to the same region (or overlapping regions) in *Vitis*. Within each set of regions, for each pair of these regions, we then calculated the average Ks of all the paralogous gene pairs the two of them contained (if any). We then screened all the sets to find any that consisted of quadruples of regions, consisting of two pairs with average similarity clustered around a recent value, representing the recent WGD, where the four (at most) average similarity scores *across* the two pairs of regions were clustered around an earlier value of similarity. (Instability in calculations of average Ks for pairs of regions necessitated the use of an overall similarity measure.) We found 13 quadruples that verified that these conditions; the remaining non-retained sets either contained too many or too few regions, or an insufficient number of paralogous pairs to assess the similarity between all pairs of regions. The quadruples that emerged from this search should be suggestive of the chromosomes produced by the two rounds of WGD, although this may be obscured by rearrangement and fractionation. Using the intermediate ancestral genome as reconstructed in Fig. S12 to represent the more recent WGD, the pattern of common adjacencies in pairs of blocks containing members of different quadruples sufficed to determine a 12-chromosome karyotype resulting from the early WGD.

4.2. Syntenic Block Fractionation Rate Analysis

After reconstructing the amalgamated syntenic blocks and reconstructing their positions on the ancestral karyotype emerging from the recent WGD, we labeled each pair of reconstructed homeologous chromosomes a and b, in no particular order, and compared the fractionation patterns for each pair of homeologous blocks in the two chromosomes, according to the formulae:

$$\text{retention rate in homeolog a} = \frac{\#\text{ genes in a}}{\#\text{ genes in a only} + \#\text{ genes in b only} + \#\text{ genes in both}}$$

$$\text{retention rate in homeolog b} = \frac{\#\text{ genes in b}}{\#\text{ genes in a only} + \#\text{ genes in b only} + \#\text{ genes in both}}$$

The results for the 54 pairs of homeologous blocks, displayed in Fig. S13, show that for eight of the nine chromosome pairs, the blocks in one homeolog have consistently higher retention rates (Dataset S3). This suggests that one of the two subgenomes involved in the WGD was dominant, retaining more genes during fractionation than the other. Indeed, high versus low gene retention rates *within* chromosomes or chromosome-sized contigs (e.g., the light green unitig_0 in Fig. S13) may represent homeologous recombination between subgenomes after the most recent WGD.

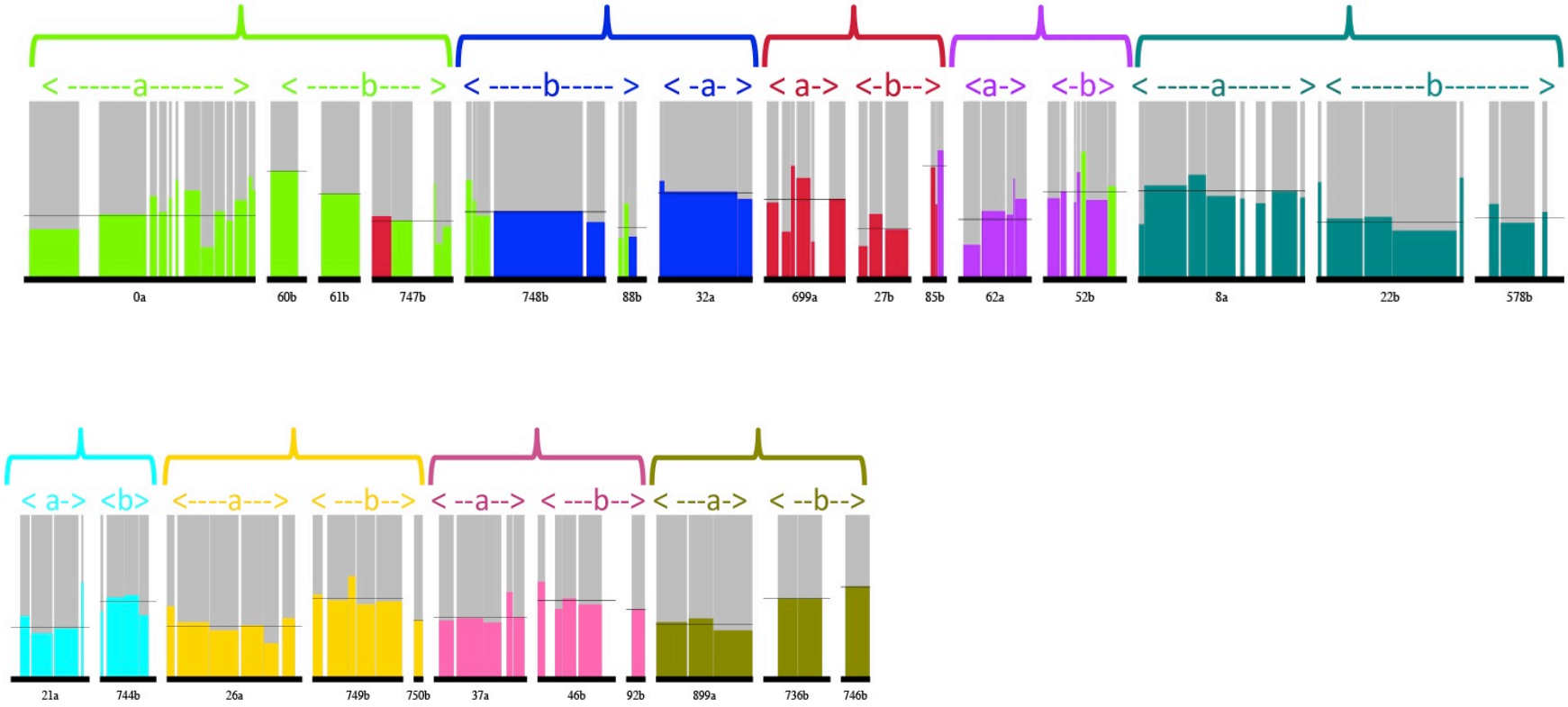


Fig. S13. Fractionation patterns in nine pairs of reconstructed chromosomes, colored according to chromosome, with a versus b subgenome pairs shown and sets of contributing contigs enveloped in brackets (with some interchromosomal rearrangements apparent). Y-axis indicates extent of fractionation, with contig-wise averages shown as dotted lines.

4.3. Subgenome Differential Expression Analysis

To investigate gene expression levels between the two subgenomes, we used raw read counts instead of normalized read counts, since all comparisons were performed within each library. Beside the shoot, trap, and inflorescence libraries, we also included the raw read counts from a stress condition library (27) and the raw read counts from the Ion Torrent RNA-seq reads sequenced from pooled tissues of the whole *U. gibba* plant (NCBI accession SRX247091, from (2)). Homeologs, which are the syntenic gene pairs obtained from SynMap, were assigned to the dominant (less fractionated) or recessive (more fractionated) subgenomes based on their locations within the syntenic blocks. Expression fold change was calculated for each homeolog using the ratio of the read counts in one subgenome versus the other. We considered homeologs to be differentially expressed if they had an expression fold change higher than the cutoff (2, 5, 10, 15, or 20-fold). Comparisons of total dominantly-expressed genes (as occurring either on the dominant or recessive subgenome) are summarized in Fig. S14. A consistent bias was observed (among most datasets and fold cutoffs) whereby the dominant subgenome showed a greater number of dominantly expressed genes. The *p*-values calculated using cumulative binomial distributions were significant at 2-fold expression difference in every transcriptome dataset except the one representing stress conditions. However, as the fold cutoff increased, the biased gene expression dominance in the dominant subgenome became less significant or non-significant. For example, when the fold cutoff was set to 20-fold, the recessive subgenome had a higher number of dominantly expressed genes in shoots.

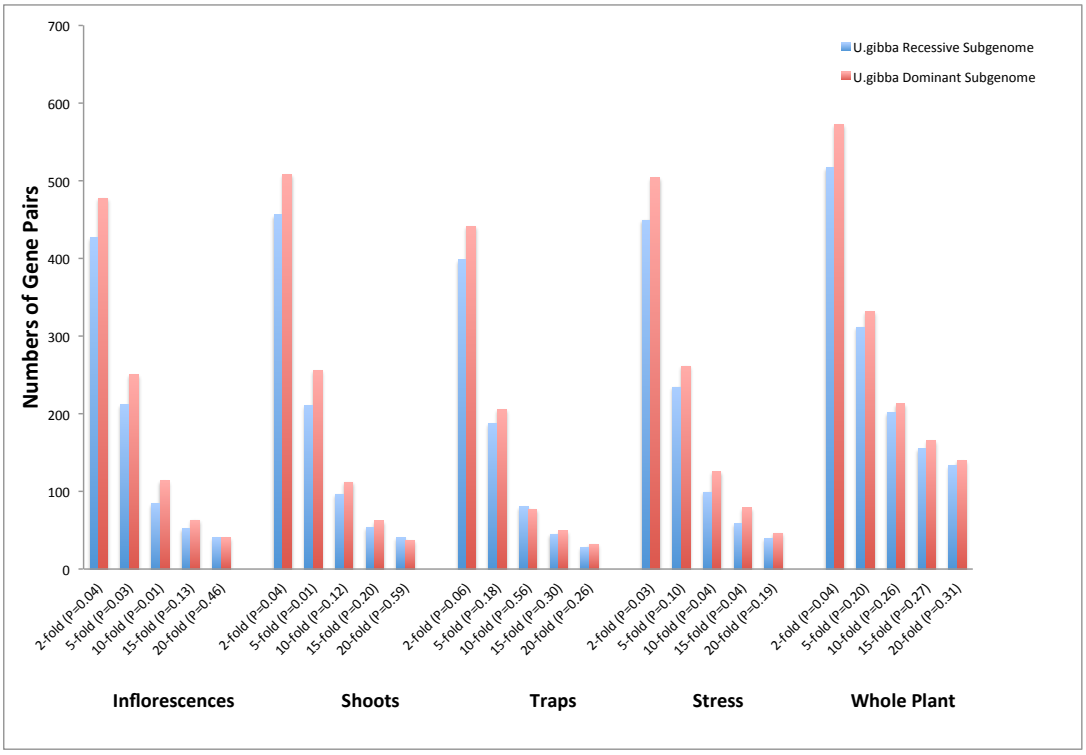


Fig. S14. Patterns of *U. gibba* homeolog expression in shoots, traps, inflorescences and the whole plant. All *p*-values were calculated using cumulative binomial distributions assuming an equal chance of gene copies for the dominant versus recessive subgenome to dominate total expression for the gene pair.

4.4. Variant Calling and Subgenome Heterozygosity Rate Analysis

Illumina and 454 raw reads from our previous short-read assembly (2) were aligned to the PacBio assembly using BWA mem version 0.7.7 (68) with default parameters. The resulting BAM files were filtered using SAMtools version 0.1.19 (69) with option "-q 30" to only keep reads that had mapping

quality larger than 30. Duplicated reads created by PCR amplification during library preparation were removed using the MarkDuplicates tool in the Picard software suite version 1.112 (http://broadinstitute.github.io/picard/) with lenient validation stringency. The Illumina- and 454-derived BAM files were then merged using SAMtools. Variant calling was implemented using the HaplotypeCaller tool in the GATK toolkit version 3.2 (70) with default settings. Heterozygous SNPs were then extracted from the resulting VCF file. Heterozygous SNPs were counted within each syntenic block and then added up for all dominant and recessive subgenome blocks, respectively. The heterozygosity rate was calculated as the total number of heterozygous SNPs divided by the total number of nucleotides in the blocks (Dataset S4). The bias ratio was calculated as the heterozygosity rate in the recessive subgenome divided by the rate in the dominant subgenome. The 1.5-times higher heterozygosity rate in the recessive subgenome implies stronger purifying selection acting on the dominant subgenome.

4.5. Whole Genome Duplication Analyses: Examples of Multiple *U. gibba* Blocks Syntenic to *Vitis*

Multiple *U. gibba* blocks in synteny with a single block of the *V. vinifera* genome could suggest either evidence for a third WGD (2) or retained synteny from the paleohexaploidy event that occurred at the base of core eudicots (71). We examined a number of cases of 8:1, or greater than 8:1 syntenic relationships compiled for *U. gibba* versus individual *Vitis* blocks using the SynFind tool (72). We illustrate 9 cases below of such multi-block relationships using CoGe's GEvo tool (Figs. S15-23). In each case, multiple *U. gibba* blocks show intercalated synteny (via block-specific, colored lines connecting BLAST HSPs) against a single *Vitis* block, as would be expected from a minimum of 3 WGD events. However, some of the multiple blocks are likely neighbors of each other instead of existing in complete overlap, since the rearranged structure of the *U. gibba* genome makes the latter status difficult to resolve for old, heavily fractionated duplicate regions. Furthermore, some of the multiple *U. gibba* blocks might instead be best matches to other, triplicated *Vitis* blocks that are homeologous with the query regions shown here; in other words, some of the multiply syntenic *U. gibba* blocks shown could simply date to the paleohexaploidy event. Preliminary analyses suggest that this might be the case for some comparisons, so we reserve judgment for the time being on the existence of a third lineage-specific WGD having occurred during *U. gibba* genome evolution.

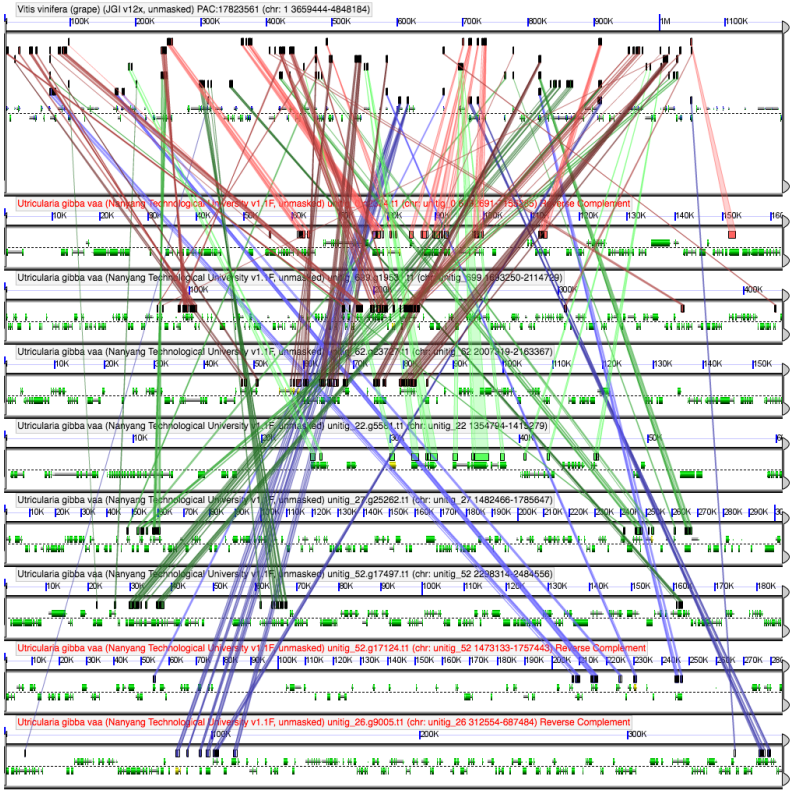


Fig. S15. GEvo plot for *Vitis* chr1. CoGe website link:
https://genomevolution.org/coge//GEvo.pl?prog=blastz;iw=1000;fh=10;padding=1;hsp_top=1;nt=0;cbc=0;spike_len=15;ca=1;skip_feat_overlap=1;skip_hsp_overlap=1;hs=0;bzW=8;bzK=3000;bzO=400;bzE=30;accn1=PAC%3A17823561;fid1=391283430;dsid1=80882;dsgid1=19990;chr1=1;dr1up=492308;dr1down=693716;ref1=1;mask1=non-cds;accn2=unitig_0.g2334.t1;fid2=826717306;dsid2=98983;dsgid2=28800;chr2=unitig_0;dr2up=80000;dr2down=80000;rev2=1;ref2=0;mask2=non-cds;accn3=unitig_699.g19531.t1;fid3=826753816;dsid3=98983;dsgid3=28800;chr3=unitig_699;dr3up=210000;dr3down=210000;ref3=0;mask3=non-cds;accn4=unitig_62.g23727.t1;fid4=826750848;dsid4=98983;dsgid4=28800;chr4=unitig_62;dr4up=55130;dr4down=96624;ref4=0;mask4=non-cds;accn5=unitig_22.g5581.t1;fid5=826725416;dsid5=98983;dsgid5=28800;chr5=unitig_22;dr5up=30000;dr5down=30000;ref5=0;mask5=non-cds;accn6=unitig_27.g25262.t1;fid6=826732448;dsid6=98983;dsgid6=28800;chr6=unitig_27;dr6up=250000;dr6down=50266;ref6=0;mask6=non-cds;accn7=unitig_52.g17497.t1;fid7=826743884;dsid7=98983;dsgid7=28800;chr7=unitig_52;dr7up=359615;dr7down=-175926;ref7=0;mask7=non-cds;accn8=unitig_52.g17124.t1;fid8=826743204;dsid8=98983;dsgid8=28800;chr8=unitig_52;dr8up=230000;dr8down=53025;rev8=1;ref8=0;mask8=non-cds;accn9=unitig_26.g9005.t1;fid9=826729650;dsid9=98983;dsgid9=28800;chr9=unitig_26;dr9up=82680;dr9down=290000;rev9=1;ref9=0;mask9=non-cds;num_seqs=9;hsp_overlap_limit=0;hsp_size_limit=0

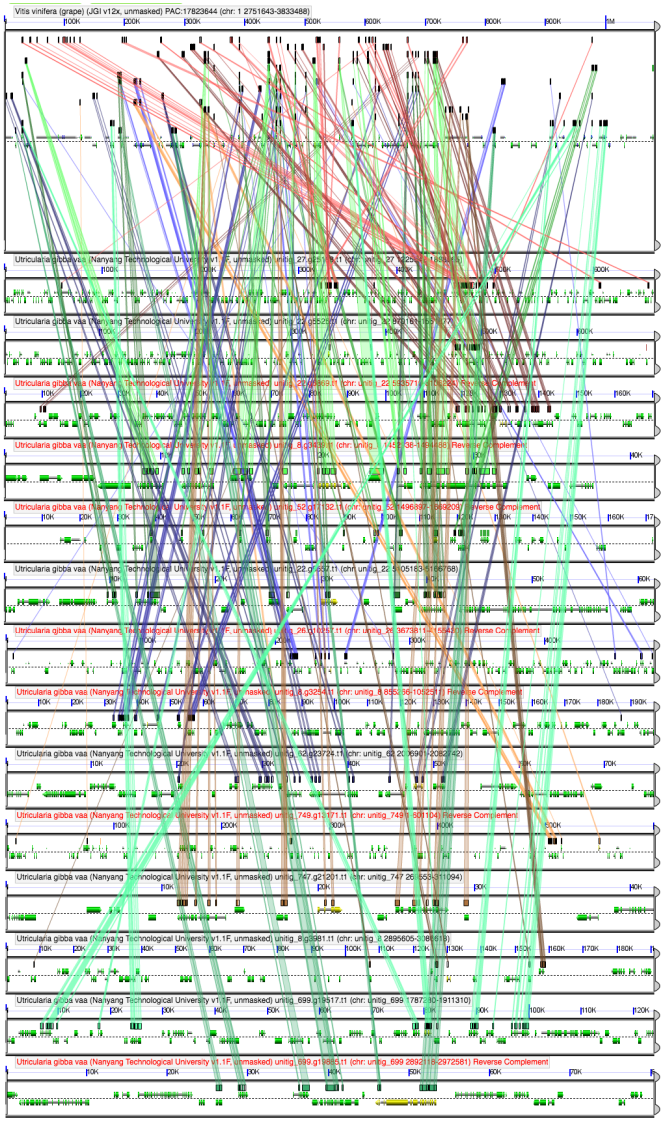


Fig. S16. A second GEvo plot for *Vitis* chr1. CoGe website link:
https://genomevolution.org/coge//GEvo.pl?prog=blastz;iw=1000;fh=10;padding=1;hsp_top=1;nt=0;cbc=0;spike_len=15;ca=1;skip_feat_overlap=1;skip_hsp_overlap=1;hs=0;bzW=8;bzK=3000;bzO=400;bzE=30;accn1=PAC%3A17823644;fid1=391283596;dsid1=80882;dsgid1=19990;chr1=1;dr1up=514180;dr1down=563491;ref1=1;mask1=non-cds;accn2=unitig_27.g25188.t1;fid2=826732302;dsid2=98983;dsgid2=28800;chr2=unitig_27;dr2up=330000;dr2down=330000;ref2=0;mask2=non-cds;accn3=unitig_22.g5525.t1;fid3=826725320;dsid3=98983;dsgid3=28800;chr3=unitig_22;dr3up=340000;dr3down=340000;ref3=0;mask3=non-cds;accn4=unitig_22.g6869.t1;fid4=826727502;dsid4=98983;dsgid4=28800;chr4=unitig_22;dr4up=130000;dr4down=38731;rev4=1;ref4=0;mask4=non-cds;accn5=unitig_8.g3439.t1;fid5=826767652;dsid5=98983;dsgid5=28800;chr5=unitig_8;dr5up=20000;dr5down=20000;rev5=1;ref5=0;mask5=non-cds;accn6=unitig_52.g17132.t1;fid6=826743220;dsid6=98983;dsgid6=28800;chr6=unitig_52;dr6up=113792;dr6down=57483;rev6=1;ref6=0;mask6=non-cds;accn7=unitig_22.g6557.t1;fid7=826726924;dsid7=98983;dsgid7=28800;chr7=unitig_22;dr7up=30000;dr7down=30000;ref7=0;mask7=non-cds;accn8=unitig_26.g10257.t1;fid8=826728790;dsid8=98983;dsgid8=28800;chr8=unitig_26;dr8up=240000;dr8down=240000;rev8=1;ref8=0;mask8=non-cds;accn9=unitig_8.g3254.t1;fid9=826767318;dsid9=98983;dsgid9=28800;chr9=unitig_8;dr9up=43178;dr9down=15000

0;rev9=1;ref9=0;mask9=non-cds;accn10=unitig_62.g23724.t1;fid10=826750844;dsid10=98983;dsgid10=28800;chr10=unitig_62;dr10up=47761;dr10down=26965;ref10=0;mask10=non-cds;accn11=unitig_749.g13171.t1;fid11=826763646;dsid11=98983;dsgid11=28800;chr11=unitig_749;dr11up=550000;dr11down=550000;rev11=1;ref11=0;mask11=non-cds;accn12=unitig_747.g21201.t1;fid12=826759016;dsid12=98983;dsgid12=28800;chr12=unitig_747;dr12up=20000;dr12down=20000;ref12=0;mask12=non-cds;accn13=unitig_8.g3981.t1;fid13=826768650;dsid13=98983;dsgid13=28800;chr13=unitig_8;dr13up=130000;dr13down=60016;ref13=0;mask13=non-cds;accn14=unitig_699.g19517.t1;fid14=826753788;dsid14=98983;dsgid14=28800;chr14=unitig_699;dr14up=80000;dr14down=42608;ref14=0;mask14=non-cds;accn15=unitig_699.g19885.t1;fid15=826754494;dsid15=98983;dsgid15=28800;chr15=unitig_699;dr15up=45909;dr15down=27029;rev15=1;ref15=0;mask15=non-cds;num_seqs=15;hsp_overlap_limit=0;hsp_size_limit=0

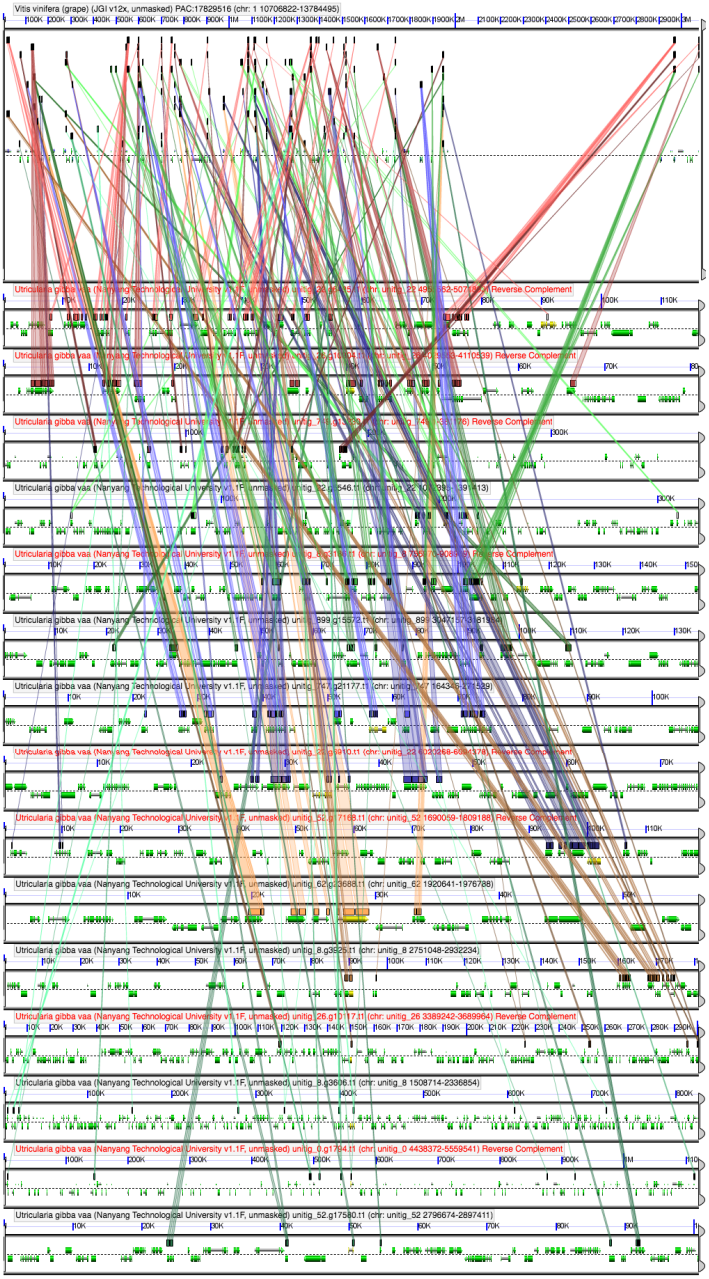


Fig. S17. A third GEvo plot for *Vitis* chr1. CoGe website link:
https://genomevolution.org/coge//GEvo.pl?prog=blastz;iw=1000;fh=10;padding=1;hsp_top=1;nt=0;cbc=0;spike_len=15;ca=1;skip_feat_overlap=1;skip_hsp_overlap=1;hs=0;bzW=8;bzK=3000;bzO=400;bzE=30;accn1=PAC%3A17829516;fid1=391284504;dsid1=80882;dsgid1=19990;chr1=1;dr1up=1050824;dr1down=2018316;ref1=1;mask1=non-cds;accn2=unitig_22.g6495.t1;fid2=826726810;dsid2=98983;dsgid2=28800;chr2=unitig_22;dr2up=90000;dr2down=23729;rev2=1;ref2=0;mask2=non-cds;accn3=unitig_26.g10304.t1;fid3=826728872;dsid3=98983;dsgid3=28800;chr3=unitig_26;dr3up=40000;dr3down=40000;rev3=1;ref3=0;mask3=non-cds;accn4=unitig_749.g13230.t1;fid4=826763712;dsid4=98983;dsgid4=28800;chr4=unitig_749;dr4up=125300;dr4down=390000;rev4=1;ref4=0;mask4=non-cds;accn5=unitig_22.g5546.t1;fid5=826725356;dsid5=98983;dsgid5=28800;chr5=unitig_22;dr5up=202210;dr5down=114874;ref5=0;mask5=non-cds;accn6=unitig_8.g3186.t1;fid6=826767192;dsid6=98983;dsgid6=28800;chr6=unitig_8;dr6up=113043;dr6down=37263;rev6=1;ref6=0;mask6=non-cds;accn7=unitig_899.g15572.t1;fid7=826775122;dsid7=98983;dsgid7=28800;chr7=unitig_899;dr7up=196672;dr7down=-62580;ref7=0;mask7=non-cds;accn8=unitig_747.g21177.t1;fid8=826758972;dsid8=98983;dsgid8=28800;chr8=unitig_747;dr8up=56312;dr8down=48122;ref8=0;mask8=non-cds;accn9=unitig_22.g6910.t1;fid9=826727576;dsid9=98983;dsgid9=28800;chr9=unitig_22;dr9up=32913;dr9down=39193;rev9=1;ref9=0;mask9=non-cds;accn10=unitig_52.g17168.t1;fid10=826743284;dsid10=98983;dsgid10=28800;chr10=unitig_52;dr10up=100000;dr10down=16785;rev10=1;ref10=0;mask10=non-cds;accn11=unitig_62.g23688.t1;fid11=826750774;dsid11=98983;dsgid11=28800;chr11=unitig_62;dr11up=26924;dr11down=26669;ref11=0;mask11=non-cds;accn12=unitig_8.g3925.t1;fid12=826768558;dsid12=98983;dsgid12=28800;chr12=unitig_8;dr12up=90000;dr12down=90000;ref12=0;mask12=non-cds;accn13=unitig_26.g10117.t1;fid13=826728526;dsid13=98983;dsgid13=28800;chr13=unitig_26;dr13up=150000;dr13down=150000;rev13=1;ref13=0;mask13=non-cds;accn14=unitig_8.g3606.t1;fid14=826767972;dsid14=98983;dsgid14=28800;chr14=unitig_8;dr14up=410000;dr14down=410000;ref14=0;mask14=non-cds;accn15=unitig_0.g1794.t1;fid15=826716354;dsid15=98983;dsgid15=28800;chr15=unitig_0;dr15up=560000;dr15down=560000;rev15=1;ref15=0;mask15=non-cds;accn16=unitig_52.g17580.t1;fid16=826744032;dsid16=98983;dsgid16=28800;chr16=unitig_52;dr16up=50000;dr16down=50000;ref16=0;mask16=non-cds;num_seqs=16;hsp_overlap_limit=0;hsp_size_limit=0

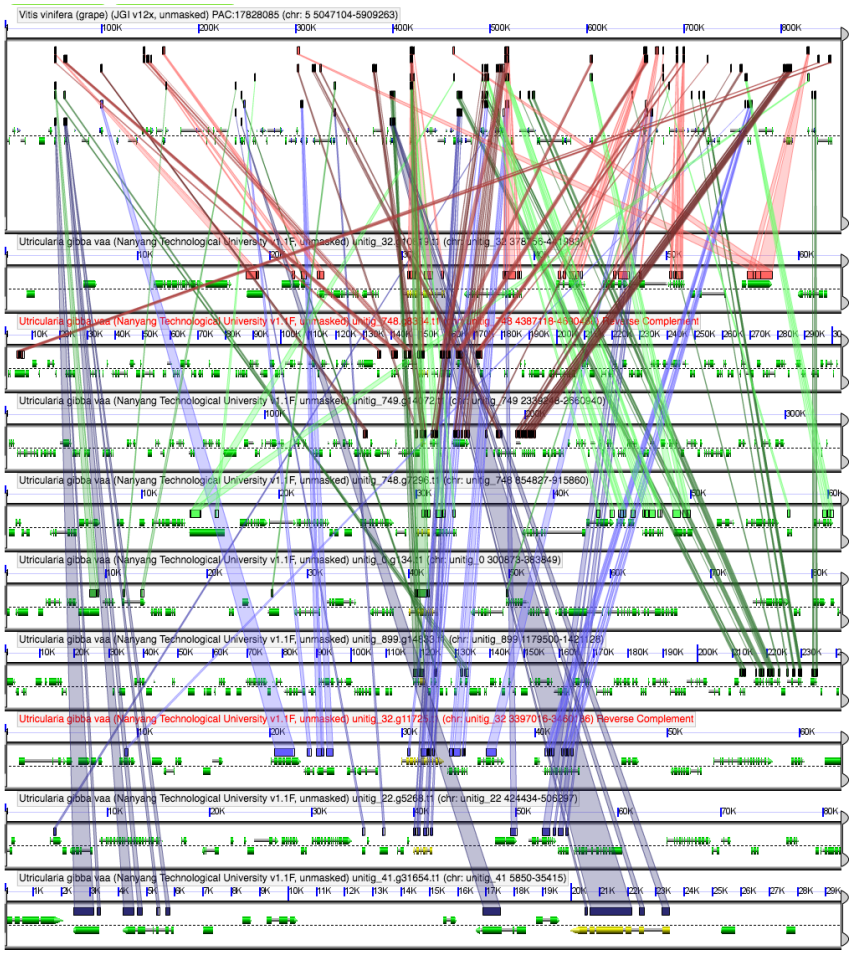


Fig. S18. GEvo plot for *Vitis* chr5. CoGe website link:
https://genomevolution.org/coge//GEvo.pl?prog=blastz;iw=1000;fh=10;padding=1;hsp_top=1;nt=0;cbc=0;spike_len=15;ca=1;skip_feat_overlap=1;skip_hsp_overlap=1;hs=0;bzW=8;bzK=3000;bzO=400;bzE=30;accn1=PAC%3A17828085;fid1=391335607;dsid1=80882;dsgid1=19990;chr1=5;dr1up=417852;dr1down=440698;ref1=1;mask1=non-cds;accn2=unitig_32.g10819.t1;fid2=826734252;dsid2=98983;dsgid2=28800;chr2=unitig_32;dr2up=30000;dr2down=30000;ref2=0;mask2=non-cds;accn3=unitig_748.g8334.t1;fid3=826762842;dsid3=98983;dsgid3=28800;chr3=unitig_748;dr3up=150000;dr3down=150000;rev3=1;ref3=0;mask3=non-cds;accn4=unitig_749.g14072.t1;fid4=826765200;dsid4=98983;dsgid4=28800;chr4=unitig_749;dr4up=160000;dr4down=160000;ref4=0;mask4=non-cds;accn5=unitig_748.g7296.t1;fid5=826761188;dsid5=98983;dsgid5=28800;chr5=unitig_748;dr5up=30000;dr5down=30000;ref5=0;mask5=non-cds;accn6=unitig_0.g134.t1;fid6=826715556;dsid6=98983;dsgid6=28800;chr6=unitig_0;dr6up=40000;dr6down=40000;ref6=0;mask6=non-cds;accn7=unitig_899.g14833.t1;fid7=826773746;dsid7=98983;dsgid7=28800;chr7=unitig_899;dr7up=120000;dr7down=120000;ref7=0;mask7=non-cds;accn8=unitig_32.g11725.t1;fid8=826735616;dsid8=98983;dsgid8=28800;chr8=unitig_32;dr8up=30000;dr8down=30000;rev8=1;ref8=0;mask8=non-cds;accn9=unitig_22.g5268.t1;fid9=826724862;dsid9=98983;dsgid9=28800;chr9=unitig_22;dr9up=40000;dr9down=40000;ref9=0;mask9=non-cds;accn10=unitig_41.g31654.t1;fid10=826738952;dsid10=98983;dsgid10=28800;chr10=unitig_41;dr10up=20000;dr10down=20000;ref10=0;mask10=non-cds;num_seqs=10;hsp_overlap_limit=0;hsp_size_limit=0

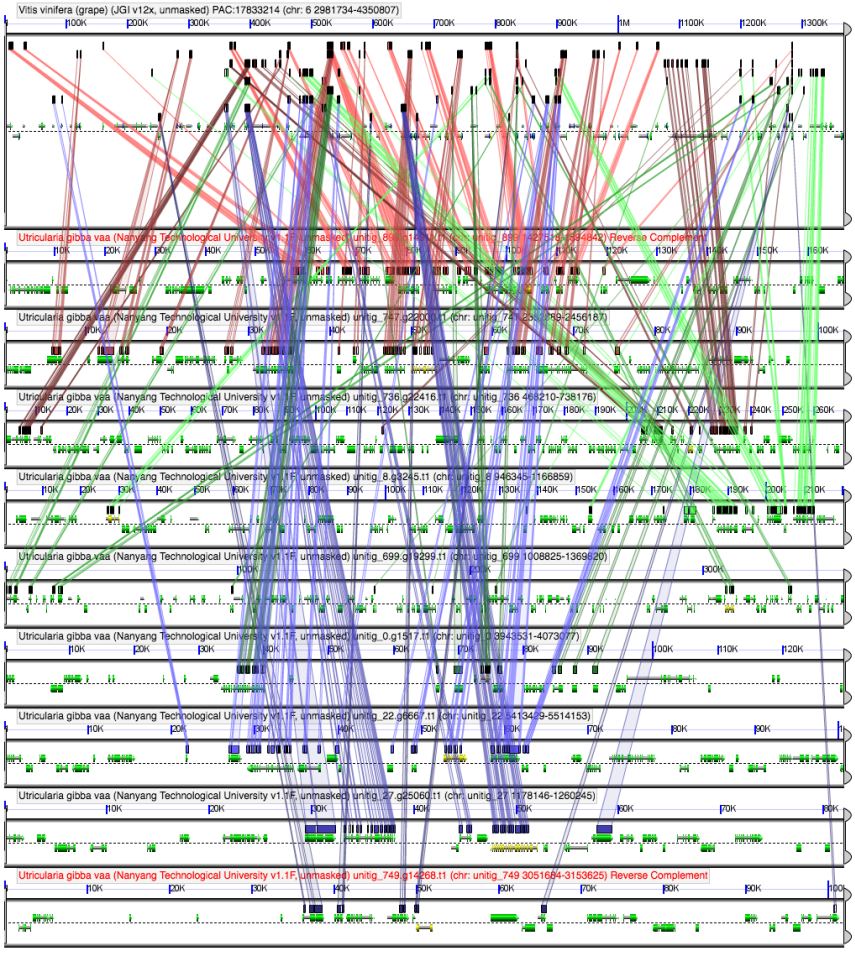


Fig. S19. GEvo plot for *Vitis* chr6. CoGe website link:
https://genomevolution.org/coge//GEvo.pl?prog=blastz;iw=1000;fh=10;padding=1;hsp_top=1;nt=0;cbc=0;spike_len=15;ca=1;skip_feat_overlap=1;skip_hsp_overlap=1;hs=0;bzW=8;bzK=3000;bzO=400;bzE=30;accn1=PAC%3A17833214;fid1=391339802;dsid1=80882;dsgid1=19990;chr1=6;dr1up=783997;dr1down=575518;ref1=1;mask1=non-cds;accn2=unitig_899.g14911.t1;fid2=826773892;dsid2=98983;dsgid2=28800;chr2=unitig_899;dr2up=101933;dr2down=62295;rev2=1;ref2=0;mask2=non-cds;accn3=unitig_747.g22000.t1;fid3=826760454;dsid3=98983;dsgid3=28800;chr3=unitig_747;dr3up=50000;dr3down=50000;ref3=0;mask3=non-cds;accn4=unitig_736.g22416.t1;fid4=826755688;dsid4=98983;dsgid4=28800;chr4=unitig_736;dr4up=220000;dr4down=48535;ref4=0;mask4=non-cds;accn5=unitig_8.g3245.t1;fid5=826767304;dsid5=98983;dsgid5=28800;chr5=unitig_8;dr5up=26856;dr5down=190000;ref5=0;mask5=non-cds;accn6=unitig_699.g19299.t1;fid6=826753386;dsid6=98983;dsgid6=28800;chr6=unitig_699;dr6up=310000;dr6down=47365;ref6=0;mask6=non-cds;accn7=unitig_0.g1517.t1;fid7=826715922;dsid7=98983;dsgid7=28800;chr7=unitig_0;dr7up=73562;dr7down=54442;ref7=0;mask7=non-cds;accn8=unitig_22.g6667.t1;fid8=826727126;dsid8=98983;dsgid8=28800;chr8=unitig_22;dr8up=52725;dr8down=45212;ref8=0;mask8=non-cds;accn9=unitig_27.g25060.t1;fid9=826732068;dsid9=98983;dsgid9=28800;chr9=unitig_27;dr9up=47565;dr9down=29997;ref9=0;mask9=non-cds;accn10=unitig_749.g14268.t1;fid10=826765562;dsid10=98983;dsgid10=28800;chr10=unitig_749;dr10up=50000;dr10down=50000;rev10=1;ref10=0;mask10=non-cds;num_seqs=10;hsp_overlap_limit=0;hsp_size_limit=0

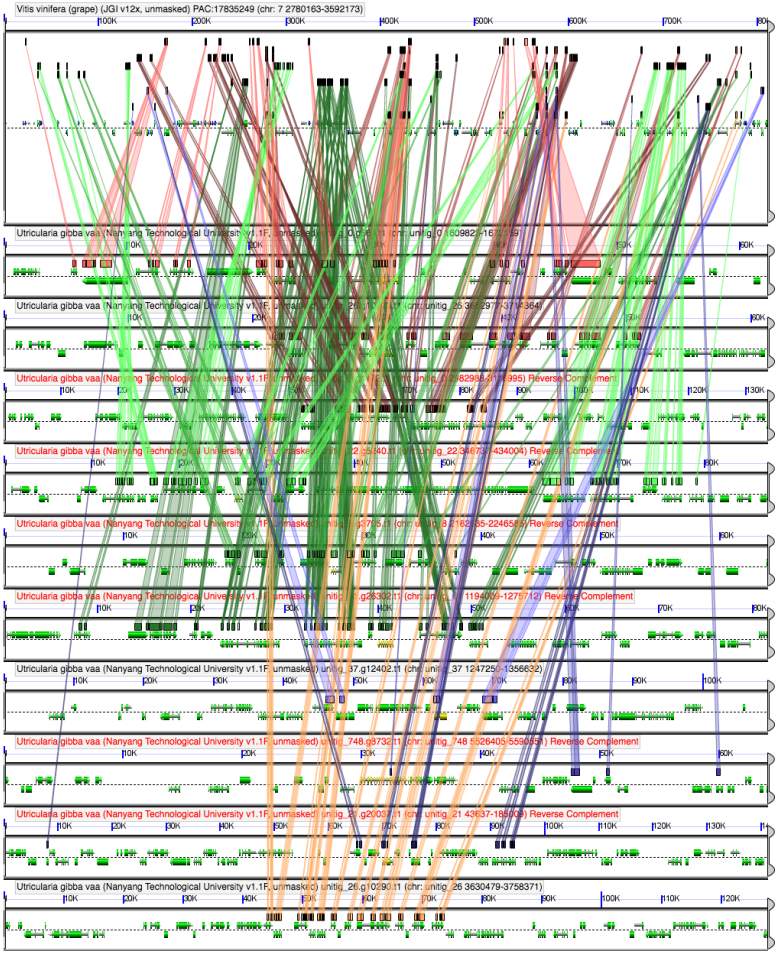


Fig. S20. GEvo plot for *Vitis* chr7. CoGe website link:
https://genomevolution.org/coge//GEvo.pl?prog=blastz;iw=1000;fh=10;padding=1;hsp_top=1;nt=0;cbc=0;spike_len=15;ca=1;skip_feat_overlap=1;skip_hsp_overlap=1;hs=0;bzW=8;bzK=3000;bzO=400;bzE=30;accn1=PAC%3A17835249;fid1=391344373;dsid1=80882;dsgid1=19990;chr1=7;dr1up=429973;dr1down=378851;ref1=1;mask1=non-cds;accn2=unitig_0.g664.t1;fid2=826719196;dsid2=98983;dsgid2=28800;chr2=unitig_0;dr2up=30000;dr2down=30000;ref2=0;mask2=non-cds;accn3=unitig_26.g10173.t1;fid3=826728634;dsid3=98983;dsgid3=28800;chr3=unitig_26;dr3up=30000;dr3down=30000;ref3=0;mask3=non-cds;accn4=unitig_0.g1172.t1;fid4=826715220;dsid4=98983;dsgid4=28800;chr4=unitig_0;dr4up=69448;dr4down=63269;rev4=1;ref4=0;mask4=non-cds;accn5=unitig_22.g5240.t1;fid5=826724808;dsid5=98983;dsgid5=28800;chr5=unitig_22;dr5up=26358;dr5down=60000;rev5=1;ref5=0;mask5=non-cds;accn6=unitig_8.g3705.t1;fid6=826768160;dsid6=98983;dsgid6=28800;chr6=unitig_8;dr6up=27467;dr6down=35647;rev6=1;ref6=0;mask6=non-cds;accn7=unitig_61.g26302.t1;fid7=826749334;dsid7=98983;dsgid7=28800;chr7=unitig_61;dr7up=40000;dr7down=40000;rev7=1;ref7=0;mask7=non-cds;accn8=unitig_37.g12402.t1;fid8=826737194;dsid8=98983;dsgid8=28800;chr8=unitig_37;dr8up=61633;dr8down=45978;ref8=0;mask8=non-cds;accn9=unitig_748.g8732.t1;fid9=826763546;dsid9=98983;dsgid9=28800;chr9=unitig_748;dr9up=30000;dr9down=30000;rev9=1;ref9=0;mask9=non-cds;accn10=unitig_21.g20037.t1;fid10=826722446;dsid10=98983;dsgid10=28800;chr10=unitig_21;dr10up=70000;dr10down=70000;rev10=1;ref10=0;mask10=non-cds;accn11=unitig_26.g10290.t1;fid11=826728848;dsid11=98983;dsgid11=28800;chr11=unitig_26;dr11up=401283;dr11down=-275819;ref11=0;mask11=non-cds;num_seqs=11;hsp_overlap_limit=0;hsp_size_limit=0

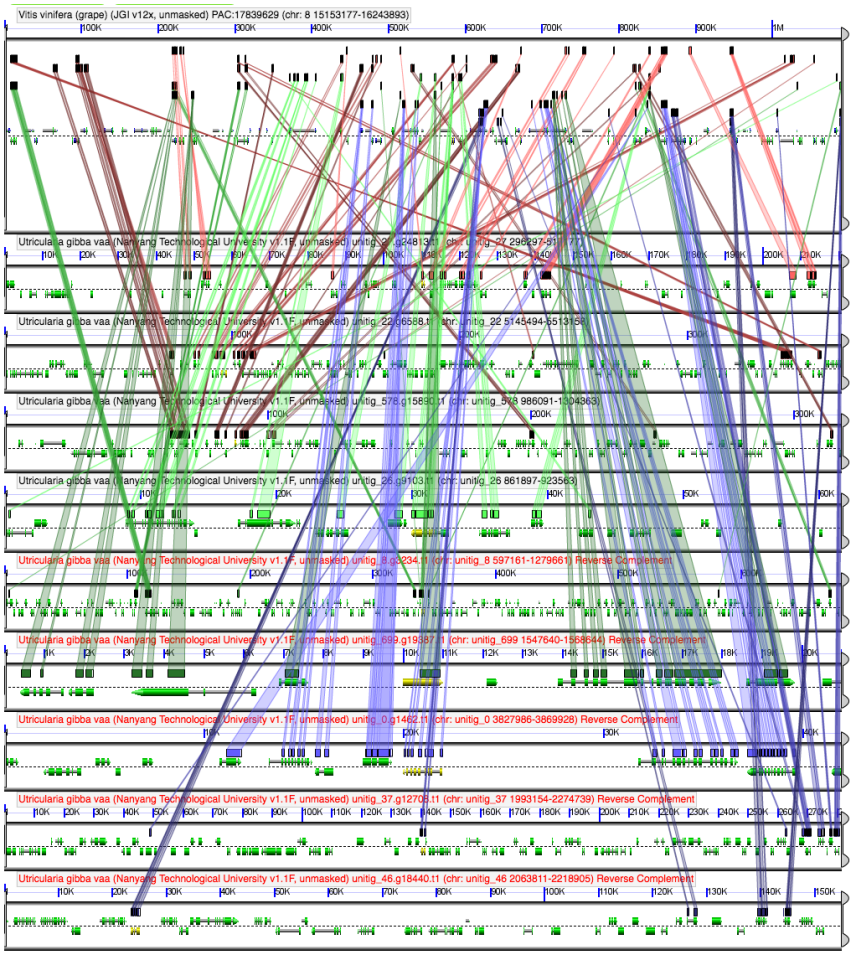


Fig. S21. GEvo plot for *Vitis* chr8. CoGe website link:
https://genomevolution.org/coge//GEvo.pl?prog=blastz;iw=1000;fh=10;padding=1;hsp_top=1;nt=0;cbc=0;spike_len=15;ca=1;skip_feat_overlap=1;skip_hsp_overlap=1;hs=0;bzW=8;bzK=3000;bzO=400;bzE=30;accn1=PAC%3A17839629;fid1=391349808;dsid1=80882;dsgid1=19990;chr1=8;dr1up=583528;dr1down=504151;ref1=1;mask1=non-cds;accn2=unitig_27.g24813.t1;fid2=826731654;dsid2=98983;dsgid2=28800;chr2=unitig_27;dr2up=110000;dr2down=110000;ref2=0;mask2=non-cds;accn3=unitig_22.g6588.t1;fid3=826726980;dsid3=98983;dsgid3=28800;chr3=unitig_22;dr3up=94971;dr3down=270000;ref3=0;mask3=non-cds;accn4=unitig_578.g15890.t1;fid4=826746028;dsid4=98983;dsgid4=28800;chr4=unitig_578;dr4up=87539;dr4down=230000;ref4=0;mask4=non-cds;accn5=unitig_26.g9103.t1;fid5=826729826;dsid5=98983;dsgid5=28800;chr5=unitig_26;dr5up=30000;dr5down=30000;ref5=0;mask5=non-cds;accn6=unitig_8.g3234.t1;fid6=826767284;dsid6=98983;dsgid6=28800;chr6=unitig_8;dr6up=340000;dr6down=340000;rev6=1;ref6=0;mask6=non-cds;accn7=unitig_699.g19387.t1;fid7=826753542;dsid7=98983;dsgid7=28800;chr7=unitig_699;dr7up=10000;dr7down=10000;rev7=1;ref7=0;mask7=non-cds;accn8=unitig_0.g1462.t1;fid8=826715810;dsid8=98983;dsgid8=28800;chr8=unitig_0;dr8up=20000;dr8down=20000;rev8=1;ref8=0;mask8=non-cds;accn9=unitig_37.g12708.t1;fid9=826737746;dsid9=98983;dsgid9=28800;chr9=unitig_37;dr9up=140000;dr9down=140000;rev9=1;ref9=0;mask9=non-cds;accn10=unitig_46.g18440.t1;fid10=826741104;dsid10=98983;dsgid10=28800;chr10=unitig_46;dr10up=23540;dr10down=130000;rev10=1;ref10=0;mask10=non-cds;num_seqs=10;hsp_overlap_limit=0;hsp_size_limit=0

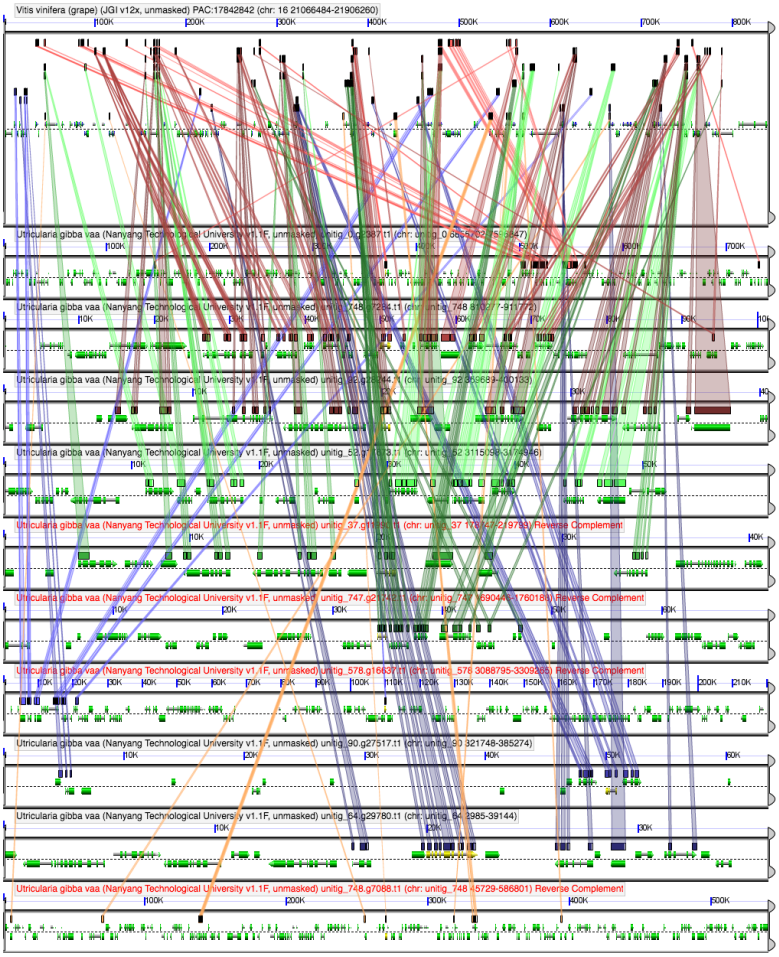


Fig. S22. GEvo plot for *Vitis* chr16. CoGe website link:
https://genomevolution.org/coge//GEvo.pl?prog=blastz;iw=1000;fh=10;padding=1;hsp_top=1;nt=0;cbc=0;spike_len=15;ca=1;skip_feat_overlap=1;skip_hsp_overlap=1;hs=0;bzW=8;bzK=3000;bzO=400;bzE=30;accn1=PAC%3A17842842;fid1=391309733;dsid1=80882;dsgid1=19990;chr1=16;dr1up=381041;dr1down=455942;ref1=1;mask1=non-cds;accn2=unitig_0.g2387.t1;fid2=826717414;dsid2=98983;dsgid2=28800;chr2=unitig_0;dr2up=370000;dr2down=370000;ref2=0;mask2=non-cds;accn3=unitig_748.g7284.t1;fid3=826761164;dsid3=98983;dsgid3=28800;chr3=unitig_748;dr3up=50000;dr3down=50000;ref3=0;mask3=non-cds;accn4=unitig_92.g28244.t1;fid4=826775990;dsid4=98983;dsgid4=28800;chr4=unitig_92;dr4up=20000;dr4down=20000;ref4=0;mask4=non-cds;accn5=unitig_52.g17673.t1;fid5=826744210;dsid5=98983;dsgid5=28800;chr5=unitig_52;dr5up=-54522;dr5down=110000;ref5=0;mask5=non-cds;accn6=unitig_37.g11990.t1;fid6=826736442;dsid6=98983;dsgid6=28800;chr6=unitig_37;dr6up=20000;dr6down=20000;rev6=1;ref6=0;mask6=non-cds;accn7=unitig_747.g21742.t1;fid7=826759988;dsid7=98983;dsgid7=28800;chr7=unitig_747;dr7up=34147;dr7down=34737;rev7=1;ref7=0;mask7=non-cds;accn8=unitig_578.g16637.t1;fid8=826747360;dsid8=98983;dsgid8=28800;chr8=unitig_578;dr8up=110000;dr8down=110000;rev8=1;ref8=0;mask8=non-cds;accn9=unitig_90.g27517.t1;fid9=826775418;dsid9=98983;dsgid9=28800;chr9=unitig_90;dr9up=50000;dr9down=12610;ref9=0;mask9=non-cds;accn10=unitig_64.g29780.t1;fid10=826751472;dsid10=98983;dsgid10=28800;chr10=unitig_64;dr10up=20000;dr10down=20000;ref10=0;mask10=non-cds;accn11=unitig_748.g7088.t1;fid11=826760818;dsid11=98983;dsgid11=28800;chr11=unitig_748;dr11up=270000;dr11down=270000;rev11=1;ref11=0;mask11=non-cds;num_seqs=11;hsp_overlap_limit=0;hsp_size_limit=0

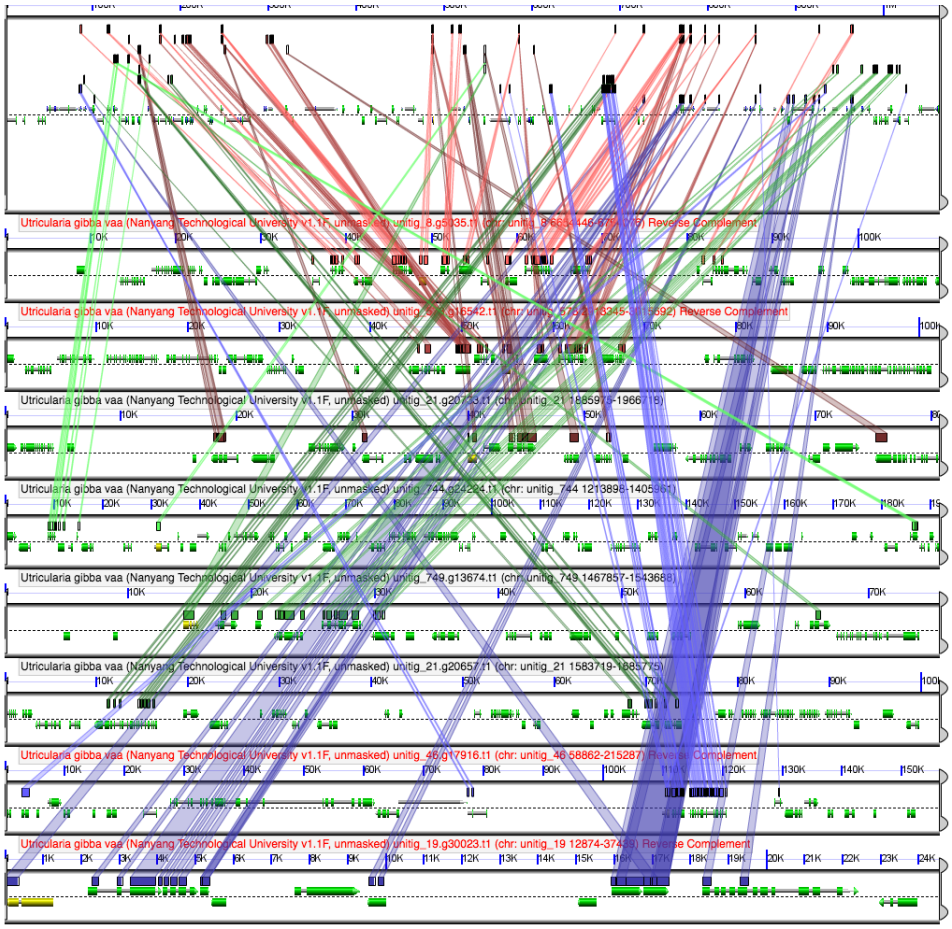


Fig. S23. GEvo plot for *Vitis* chr18. CoGe website link:
https://genomevolution.org/coge//GEvo.pl?prog=blastz;iw=1000;fh=10;padding=1;hsp_top=1;nt=0;cbc=0;spike_len=15;ca=1;skip_feat_overlap=1;skip_hsp_overlap=1;hs=0;bzW=8;bzK=3000;bzO=400;bzE=30;accn1=PAC%3A17821841;fid1=391316180;dsid1=80882;dsgid1=19990;chr1=18;dr1up=486751;dr1down=576416;ref1=1;mask1=non-cds;accn2=unitig_8.g5035.t1;fid2=826770172;dsid2=98983;dsgid2=28800;chr2=unitig_8;dr2up=48676;dr2down=60088;rev2=1;ref2=0;mask2=non-cds;accn3=unitig_578.g16542.t1;fid3=826747186;dsid3=98983;dsgid3=28800;chr3=unitig_578;dr3up=54524;dr3down=46945;rev3=1;ref3=0;mask3=non-cds;accn4=unitig_21.g20733.t1;fid4=826723670;dsid4=98983;dsgid4=28800;chr4=unitig_21;dr4up=40000;dr4down=40000;ref4=0;mask4=non-cds;accn5=unitig_744.g24224.t1;fid5=826757082;dsid5=98983;dsgid5=28800;chr5=unitig_744;dr5up=30838;dr5down=160000;ref5=0;mask5=non-cds;accn6=unitig_749.g13674.t1;fid6=826764462;dsid6=98983;dsgid6=28800;chr6=unitig_749;dr6up=14553;dr6down=60000;ref6=0;mask6=non-cds;accn7=unitig_21.g20657.t1;fid7=826723530;dsid7=98983;dsgid7=28800;chr7=unitig_21;dr7up=130000;dr7down=-28989;ref7=0;mask7=non-cds;accn8=unitig_46.g17916.t1;fid8=826740322;dsid8=98983;dsgid8=28800;chr8=unitig_46;dr8up=110000;dr8down=45200;rev8=1;ref8=0;mask8=non-cds;accn9=unitig_19.g30023.t1;fid9=826722082;dsid9=98983;dsgid9=28800;chr9=unitig_19;dr9up=30000;dr9down=23295;rev9=1;ref9=0;mask9=non-cds;num_seqs=9;hsp_overlap_limit=0;hsp_size_limit=0

5. Gene Ontology Enrichment Analyses

We obtained the generic gene ontology (GO) term annotations for *Arabidopsis* genes from TAIR and functionally annotated the RepBase-filtered *U. gibba* gene models by assigning the GO terms from their associated *Arabidopsis* gene annotation (see section 2.4, above). We then carried out GO term enrichment analyses of subsets of foreground genes versus all annotatable genes in the *U. gibba* genome as

background using Fisher's exact test in GOATOOLS (https://github.com/tanghaibao/goatools) to discover whether subsets of genes relate to specific biological functions or metabolic pathways. The *U. gibba* whole-genome background was custom-generated as the set of *U. gibba* genes annotatable against *Arabidopsis* genes at *E*-value cutoff of 1E-05, accepting the topmost hit as the match.

5.1. GO Enrichment Analysis of Syntenic Genes in *U. gibba* and *Arabidopsis*

To investigate GO enrichment among syntenic gene duplicates descending from *U. gibba* lineage-specific WGDs, a self-to-self SynMap was generated within CoGe using the QUOTA-ALIGN algorithm (73) with default parameters. Syntenic gene pairs was then downloaded from CoGe and used as the foreground subset in the GO enrichment analysis. As shown in Dataset S5, the topmost significantly enriched terms (Bonferroni-corrected p-values < 0.05) were mostly transcriptional regulatory functions. For comparison, the same pipeline was carried out on internally syntenic *Arabidopsis* genes descending from its own 2 lineage-specific WGDs, from which highly similar results were obtained (Dataset S6). The *Arabidopsis* background was all genes in the genome.

5.2. GO Enrichment Analysis of Tandem Duplicates in *U. gibba* and *Arabidopsis*

The blast_to_raw script in the QUOTA-ALIGN package (https://github.com/tanghaibao/quota-alignment), incorporated in CoGe's SynMap application, was used to filter out tandem duplicates before synteny plotting as in section 5.1. These genes calculated to be tandem duplicates in *U. gibba* were downloaded from a CoGe SynMap results link and used as a foreground subset for GO enrichment analyses. In contrast to functional enrichments of syntenic genes, the topmost significantly enriched terms for tandem duplicates were secondary metabolic functions, including specific functions that could be anticipated for a carnivorous plant (Dataset S7). Genes with significantly enriched GO terms assigned to them and their annotations are listed in Dataset S8. Although the specifically enriched terms were different for *Arabidopsis* tandem duplicates, they are also related mostly to secondary metabolic activities (Dataset S9). The *Arabidopsis* background used was all genes in its genome.

6. Molecular Evolution Analyses of Tandem Duplicated Genes

6.1. Cysteine Protease Genes

Cysteine protease genes identified within the collection of tandem duplicates derived in section 5.2 were used as queries for a NCBI local tblastx against *V. vinifera* (id 19990), *Arabidopsis* Col-0 (id 24424), *S. lycopersicum* (id 24769), and *U. gibba* (PacBio v1.1; id 28048) coding sequence databases downloaded from CoGe. The *Dionaea muscipula* cysteine protease (GenBank Accession KP663370) was also included in the dataset. Gene model repredictions were conducted using default settings of AUGUSTUS (21) with the genomic sequence of the previously predicted *U. gibba* gene models, plus 500-1000 bp of upstream and downstream genomic sequence. Two tandem duplicates (g1 and g2) were repredicted at locus utg699.g19345. Multiple sequence alignments were performed for CDS sequences using MAFFT E-INS-i (74). Regions corresponding to the variable signal peptide and propeptide were removed prior to phylogenetic analysis. Alignments were translated prior to phylogenetic analyses. Maximum-likelihood (ML) searches were used to reconstruct the cysteine protease phylogeny using RAxML v8.2.4 (75) on the CIPRES Science Gateway (http://www.phylo.org/index.php/) under the WAG+G model of evolution, as determined by the Akaike and Bayesian Information Criterion (AIC/BIC) in ProtTest v3.2 (76). Searches for the phylogenetic reconstruction with the highest likelihood score were performed simultaneously with rapid bootstrapping, allowing RAxML to automatically halt the analysis (at 552 bootstrap replicates). The resulting phylogeny was visualized using FigTree v1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/). The multiple sequence alignment and the resulting phylogeny used for subsequent molecular evolutionary analyses are provided in Dataset S11.

We estimated ω (dN/dS) values for the cysteine protease CDS alignment and RAxML phylogeny using the codeml part of the PAML v4.4 package (77). Gaps in the alignment were excluded by PAML. Two

types of models were implemented: "branch-specific" (ω ratio estimated for each branch in the tree (78)) and "branch-site" models (ω ratio varies in selected branches and across codons (79)).

Comparisons of two nested models were performed using a Likelihood Ratio Test (LRT) to test for the following: asymmetric sequence evolution (one-ratio model 0 ($\omega_0 = p_1$) versus two-ratio model 2 (ω_0, ω_1)), divergent selection (model 3 (discrete) versus clade model D ($K = 3$)), and positive selection (model A null ($\omega_2 = 1$) versus model A ($0 < \omega_0 < 1$). The chi square test was conducted using the log likelihood results of each branch and node of the phylogeny (Dataset S10; Cysteine Protease PAML Branches and Cysteine Protease PAML Nodes, Sheets 1 and 2). Sites listed as under positive selection in Dataset S10 correspond to amino acid residues in the multiple sequence alignment (Datasets S11-13) when gaps were removed by PAML. For subsequent homology modeling analyses of *U. gibba* cysteine protease, we matched sites identified by PAML as under positive selection in the un-gapped alignment to the original sites within contigs part of the alignment containing gaps (Datasets S10 and S11).

6.1.1. Cysteine Protease Homology Modeling

The protein structural model for the unitig699.g19348 catalytic domain was computed using the SWISS-MODEL server homology modeling pipeline (80) using PROMOD-II (81) and MODELLER (82). A crystal structure of a cysteine protease from *Dionaea muscipula* (PDB ID: 5a24) was identified as the top-ranking template in covalent complex with inhibitor E-64 (83). The program MacPyMOL v1.3 (Schrödinger LLC) was used to thread the 3D model of unitig_699.g19348 to 5a24 associated with E-64. Sites identified as evolving under positive selection pressure by the codeml branch-site model were mapped to PDB coordinates to detect substrate interacting regions and amino acids lining the substrate-binding cleft. Three (E24, V69, S160) of the unitig699.g19348 amino acid sites under positive selection (BEB confidence > 0.82, Bonferroni corrected $p < 0.0015$) are within five amino acids of the *D. muscipula* functional residues and line the substrate-binding cleft in the model (Dataset S10; Fig 3B and C in main text).

6.2. *KCS6*-like Genes

KCS6-like genes identified in the tandem duplicate analysis were used as a queries for NCBI local TBLASTX runs against *V. vinifera* (id 19990), *Arabidopsis* Col-0 (id 24424), *S. lycopersicum* (id 24769), and *U. gibba* (PacBio v1.1; id 28048) coding sequence databases downloaded from CoGe. Gene model reprediction was conducted as in section 6.1. Translated hits from the BLAST search were used to create an alignment in SeaView (84) using MUSCLE. Poorly aligned sequences were removed, the sequences were aligned again, and then the alignment was trimmed using Gblocks (85), with stringency parameters to allow smaller blocks, gap positions within the final blocks, and less strict flanking positions. Phylogenetic analysis was performed on back-translated nucleotide sequences using PhyML under default parameters in SeaView. As in section 6.1, ω values were estimated using the codeml program part of the PAML v4.4 package. The chi square test was conducted using the log-likelihood results of each branch and node of the phylogeny (Dataset S10; KCS PAML Branches and KCS PAML Nodes, Sheets 3 and 4). The multiple sequence alignment and resulting phylogeny for PAML analysis are available in Dataset S12.

6.3. *SVP*-like Genes

The *SVP*-like genes of *Arabidopsis*, tomato and grape were acquired from an ongoing MADS-box gene family analysis of 7 angiosperms being conducted by coauthors T.-H.C. and V.A.A. *U. gibba SVP*-like genes were identified by using the *Arabidopsis* and tomato SVP-like genes downloaded from TAIR (www.arabidopsis.org) to search against the *U. gibba* whole genome coding sequence dataset (PacBio v1.1; id 28048) by CoGeBlast with the TBLASTX algorithm with an *E*-value cutoff of 1E-10. Gene models were repredicted on the GeneWise website (86) for previously poorly predicted gene models. Genomic sequences of the target genes were acquired from CoGe and 5000 base pairs both upstream and downstream were added. Protein sequences serving as templates were selected based on the gene

subfamily phylogeny. Default parameters were applied to the gene model reprediction with the modeled split site setting. All *SVP*-like genes from four species were aligned using MUSCLE, and non-informative regions were removed using Gblocks (85), with stringency parameters to allow smaller blocks, gap positions within the final blocks, and less strict flanking positions. The phylogenetic analysis was performed on back-translated nucleotide sequences using PhyML under default parameters in SeaView (84). As in 6.1**,** ω values were estimated using the codeml part of the PAML v4.4 package. The chi square test was conducted using the log-likelihood results of each branch and node of the phylogeny (Dataset S10; SVP PAML Branches and SVP PAML Nodes, Sheets 5 and 6). The multiple sequence alignment and resulting phylogeny for PAML analysis are available in Dataset S13.

References

1. Peterson D, Boehm K, & Stack S (1997) Isolation of milligram quantities of nuclear DNA from tomato (*Lycopersicon esculentum*), A plant containing high levels of polyphenolic compounds. *Plant Mol. Biol. Rep.* 15(2):148-153.
2. Ibarra-Laclette E*, et al.* (2013) Architecture and evolution of a minute plant genome. *Nature* 498(7452):94-98.
3. Long Q*, et al.* (2013) Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature genetics* 45(8):884-890.
4. Chin CS*, et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10(6):563-569.
5. Walker BJ*, et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS One* 9(11):e112963.
6. Camacho C*, et al.* (2009) BLAST+: architecture and applications. *BMC bioinformatics* 10:421.
7. Salter S*, et al.* (2014) Reagent contamination can critically impact sequence-based microbiome analyses. *BMC Biology*.
8. Gualberto JM*, et al.* (2014) The plant mitochondrial genome: dynamics and maintenance. *Biochimie* 100:107-120.
9. Mao M, Austin AD, Johnson NF, & Dowton M (2013) Coexistence of minicircular and a highly rearranged mtDNA molecule suggests that recombination shapes mitochondrial genome organization. *Mol. Biol. Evol.*
10. VanBuren R*, et al.* (2015) Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature*.
11. Quesneville H*, et al.* (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS computational biology* 1(2):166-175.
12. Flutre T, Duprat E, Feuillet C, & Quesneville H (2011) Considering transposable element diversification in *de novo* annotation approaches. *PloS One* 6(1).
13. Kohany O, Gentles AJ, Hankus L, & Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC bioinformatics* 7:474.
14. Wicker T*, et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics* 8(12):973-982.
15. Lagesen K*, et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35(9):3100-3108.
16. Lowe T & Eddy S (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25(5):955-964.
17. Nawrocki E & Eddy S (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29(22):2933-2935.
18. Gardner PP*, et al.* (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.* 39(Database issue):D141-145.
19. Burge SW*, et al.* (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* 41(Database issue):D226-232.

20. Stanke M, Diekhans M, Baertsch R, & Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* 24(5):637-644.
21. Hoff KJ & Stanke M (2013) WebAUGUSTUS-a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Res.* 41(W1):W123-W128.
22. Ibarra-Laclette E*, et al.* (2011) Transcriptomics and molecular evolutionary rate analysis of the bladderwort (*Utricularia*), a carnivorous plant with a minimal genome. *BMC Plant Biol.* 11(1):101.
23. Campbell MS, Holt C, Moore B, & Yandell M (2014) Genome annotation and curation using MAKER and MAKER‐P. *Curr. Protoc. Bioinformatics*:4.11.11-14.11.39.
24. Jurka J*, et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* 110(1-4):462-467.
25. Kalendar R*, et al.* (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* 166(3):1437-1450.
26. Thorvaldsdottir H, Robinson JT, & Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 14.
27. Ibarra-Laclette E*, et al.* (2011) Transcriptomics and molecular evolutionary rate analysis of the bladderwort (*Utricularia*), a carnivorous plant with a minimal genome. *BMC Plant Biol.* 11(1):101.
28. Liao Y, Smyth GK, & Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 41(10):e108-e108.
29. Liao Y, Smyth GK, & Shi W (2013) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*:btt656.
30. Robinson MD, McCarthy DJ, & Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139-140.
31. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27(2):573-580.
32. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.
33. Fulnečková J*, et al.* (2013) A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. *Genome Biol. Evol.* 5(3):468-483.
34. Tran TD*, et al.* (2015) Centromere and telomere sequence alterations reflect the rapid genome evolution within the carnivorous plant genus *Genlisea*. *Plant J.* 84(6):1087-1099.
35. Bennetzen JL & Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Ann. Rev. Plant* 65(1):505-530.
36. Melters DP*, et al.* (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome biology* 14(1):R10.
37. Alkan C, Sajjadian S, & Eichler EE (2011) Limitations of next-generation genome sequence assembly. *Nat. Methods* 8(1):61-65.
38. McCoy RC*, et al.* (2014) Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* 9(9):e106689.
39. Hayden KE & Willard HF (2012) Composition and organization of active centromere sequences in complex genomes. *BMC Genomics* 13(1):1-13.
40. Lyons E*, et al.* (2008) Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids. *Plant Physiol.* 148:1772-1781.
41. Freeling M, Xu J, Woodhouse M, & Lisch D (2015) A solution to the C-value paradox and the function of junk DNA: the genome balance hypothesis. *Mol. Plant* 8(6):899-910.
42. Kurtz S*, et al.* (2004) Versatile and open software for comparing large genomes. *Genome biology* 5(2):R12.
43. Krzywinski M*, et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* 19.

44. Meraldi P, McAinsh AD, Rheinbay E, & Sorger PK (2006) Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. *Genome biology* 7(3):R23.
45. Birchler JA, Gao Z, & Han F (2009) A tale of two centromeres--diversity of structure but conservation of function in plants and animals. *Functional & integrative genomics* 9(1):7-13.
46. Wang G, Zhang X, & Jin W (2009) An overview of plant centromeres. *Journal of genetics and genomics = Yi chuan xue bao* 36(9):529-537.
47. Wade CM*, et al.* (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* 326(5954):865-867.
48. Nasuda S, Hudakova S, Schubert I, Houben A, & Endo TR (2005) Stable barley chromosomes without centromeric repeats. *Proceedings of the National Academy of Sciences of the United States of America* 102(28):9842-9847.
49. Locke DP*, et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature* 469(7331):529-533.
50. Liu Z*, et al.* (2008) Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. *Chromosoma* 117(5):445-456.
51. Cheng Z*, et al.* (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* 14.
52. Nagaki K*, et al.* (2005) Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. *Molecular Biol. Evol.* 22(4):845-855.
53. Zhong CX*, et al.* (2002) Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. *Plant Cell* 14(11):2825-2836.
54. Hudakova S*, et al.* (2001) Sequence organization of barley centromeres. *Nucleic Acids Res.* 29(24):5029-5035.
55. Gorin≈°ek B, Guben≈°ek F, & Kordi≈° Da (2004) Evolutionary genomics of chromoviruses in eukaryotes. *Mol. Biol. Evol.* 21(5):781-798.
56. Neumann P*, et al.* (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mobile DNA* 2(1):1.
57. Gorinsek B, Gubensek F, & Kordis D (2005) Phylogenomic analysis of chromoviruses. *Cytogenet Genome Res.* 110.
58. Gorinšek B, Gubenšek F, & Kordiš D (2004) Evolutionary genomics of chromoviruses in eukaryotes. *Mol. Biol. Evol.* 21(5):781-798.
59. Xu Z & Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35(suppl 2):W265-W268.
60. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32.
61. Miller MA, Pfeiffer W, & Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Gateway Computing Environments Workshop (GCE), 2010*, (IEEE), pp 1-8.
62. Slotkin RK & Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* 8(4):272-285.
63. Topp CN, Zhong CX, & Dawe RK (2004) Centromere-encoded RNAs are integral components of the maize kinetochore. *Pro. Natl. Acad. Sci. USA* 101(45):15986-15991.
64. Gao X, Hou Y, Ebina H, Levin HL, & Voytas DF (2008) Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* 18.
65. Vanneste K, Van de Peer Y, & Maere S (2013) Inference of genome duplications from age distributions revisited. *Molecular biology and evolution* 30(1):177-190.
66. Makova KD & Hardison RC (2015) The effects of chromatin organization on variation in mutation rates in the genome. *Nature Rev. Genet.* 16(4):213-223.
67. Lyons E & M F (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* 53(4):661-673.

68. Librado P & Rozas J (2013) Uncovering the functional constraints underlying the genomic organization of the odorant-binding protein genes. *Genome Biol. Evol.* 5:2096-2108.
69. Li H*, et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
70. McKenna A*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9):1297-1303.
71. Tang H*, et al.* (2008) Synteny and collinearity in plant genomes. *Science* 320(5875):486-488.
72. Tang H*, et al.* (2015) SynFind: Compiling Syntenic Regions across Any Set of Genomes on Demand. *Genome Biol. Evol.* 7(12):3286-3298.
73. Tang H*, et al.* (2011) Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC bioinformatics* 12(1):1.
74. Katoh K & Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30(4):772-780.
75. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312-1313.
76. Abascal F, Zardoya R, & Posada D (2005) ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics* 21:2104-2105.
77. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24(8):1586-1591.
78. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15(5):568-573.
79. Yang Z & Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19(6):908-917.
80. Biasini M*, et al.* (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 42(W1):W252-W258.
81. Guex N & Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18(15):2714-2723.
82. Sali A & Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234(3):779-815.
83. Risør MW*, et al.* (2015) Enzymatic and structural characterization of the major endopeptidase in the Venus flytrap digestion fluid. *J. Biol. Chem.*:jbc.M115.672550.
84. Gouy M, Guindon S, & Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution* 27(2):221-224.
85. Talavera G & Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56(4):564-577.
86. Birney E, Clamp M, & Durbin R (2004) GeneWise and genomewise. *Genome Res.* 14(5):988-995.