doi:10.1038/nature12132

CONTENTS

1. Sequencing and assembly of the *Utricularia gibba* genome

1.1. Plant Materials

U. gibba is a perennial, aquatic, photosynthetic herb that bears mats of reiterating vegetative structural units that lack roots[1]. The stems are very slender and up to 25 cm long. They may be floating, submerged or creeping along the bottom. The inflorescence has 1 to 4 yellow flowers 6-8 mm long at the end of a stalk less than 15 cm long. The leafy organs borne on stems are alternated, numerous, and 3-10 mm long; they are threadlike, have hairless margins, and may be undivided or generally 2-parted at the base and each part may be forked again. There are 1 or 2 valve-lidded bladders borne on the leaves that are less than 1-2 mm wide that trap small prey[2]. Plant investment in bladder number is inversely correlated with nutrient availability, reflective of the typical strategy of carnivorous plants[3-5]. It has recently been shown that traps of some aquatic species actually exude photosynthetically-derived carbon as a food source for associated bacterial assemblages that in turn supply vital nutrients[6]. Flowers of *Utricularia* species are monoecious, usually open, showy, and zygomorphic, typical of outcrossing plants serviced by insects[2,7]. However, *Utricularia* species are frequently characterised by considerable self-pollination or even predominant asexual phases[8-10]. Although the specific breeding system of *U. gibba* remains unstudied, the species likely exhibits different phases of outcrossing, inbreeding, and asexuality as do related *Utricularia* species[9,10]. For genomic DNA isolation, *U. gibba* was collected in the Umécuaro municipality, Michoacán, México, and grown outdoors in plastic containers (0.1 m^2, 10 L). Water depth was 15-20 cm, and was maintained by addition of soft tap water. At least 50% of the water used in the initial phase came from the dam in which these plants were collected.

1.2. Flow cytometric analysis

Independently, shoot-like structures and flowers were finely chopped with a razor blade in Petri dishes with 500µL of nuclei extraction buffer (Cystain ultraviolet Precise P Nuclei Extraction Buffer; Partec GmbH, Münster Germany). The solution was filtered using Partec Cell Trics disposable filters with a pore size of 50 µm to remove plant tissue debris. Nuclei were stained with 1.5 mL 4,6-diamidino-2 phenylindole Nuclei Extraction Buffer (Partec GmbH, Münster Germany) and incubated for 1 to 2 min at room temperature. A PARTEC CA II Cytometer (Partec GmbH, Münster Germany) was used to measure DAPI fluorescence (at least 3,000 nuclei) after UV excitation. *Arabidopsis thaliana* (1C = 0.1605 pg or 135 Mb, the approximate total chromosome length from the TAIR10 assembly; http://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/agicomplete.jsp) was used as an internal standard to calculate the *U. gibba* nuclear DNA content. The estimated genome size for *U. gibba* was 77.38 Mb (Supplementary Table 1; see also suppl. ref. 12).

1.3. Nuclear DNA preparation and sequencing

Nuclear and associated environmental DNA was isolated from tiny *U. gibba* shoot-like structures as described by Steinmüller and Apel[13], with minor modifications. After resuspending in isolation buffer, nuclear pellets from 50 g of fresh tissue were resuspended in 20 ml of Percoll (Sigma), and centrifuged at 4000g for 10 min at 4°C[14]. Floating nuclei were resuspended in 25 ml of isolation buffer, and then centrifuged at 800g for 15 min at 4°C. Next, nuclear DNA was purified as recommended by Steinmüller and Apel[13] and thereafter amplified by multiple displacement amplification using the GenomiPhi DNA amplification kit (Amersham Biosciences, Piscataway, NJ). Amplification was carried out according to the manufacturer's instructions. The DNA was sheared (Hydroshear) to obtain DNA fragments ranked according to the size required for sequencing libraries (1 Kb, 2 Kb, 2-4 Kb or 7-9 Kb). For whole genome sequencing, a total of eigth distinct libraries, one 3 Kb, three 8 Kb mate-pair libraries and four shotgun libraries, were constructed. Preparation, amplification and sequencing of these libraries were performed using GS FLX Titanium Sequencing Kits and Genome Sequencer FLX Instruments following the manufacturer's protocols (Roche Applied Science, Mannheim, Germany). One additional shotgun library was constructed and sequenced using the GS FLX XL+ Sequencing kit and corresponding platform. Additionally, one paired-end library of ~450 bp was prepared using Illumina's paired-end kit (Illumina, Sand Diego, CA). The DNA was sheared with a Covaris S2 ultrasonicator (Covaris Inc. Woburn, MA) and the library was sequenced (twice) as 2x250 bp on an Illumina MiSeq. Finally, conventional Sanger reads were generated with an ABI 3730xl sequencer (Applied Biosystems) using the Big Dye–terminator Cycle Sequencing kit. Recombinant clones (pJET1.2/blunt Cloning Vector; Fermentas) were used to transform DH10b cells to obtain two genomic libraries [(i) 43,968 clones, average insert size: 1.2 kb; and (ii) 55,680 clones, average insert size: 4 kb], and clones were sequenced both uni- and bidirectionally. In total ~5.2 Gb of sequence data was generated, consisting of 1.9 Gb of shotgun reads, 1.5 Gb of mate-pair reads, 1.5 Gb of paired-end reads and 119.5 Mb of Sanger reads (Supplementary Table 2).

1.4. *U. gibba de novo* assembly

The 454, Sanger and MiSeq reads were assembled using Newbler version 2.6 *de novo* genome assembler (with the -scaffold option). Vector and poor quality regions were masked in the Sanger reads using the LUCY2 software[15]. Natural and artificial duplicates in pyrosequencing reads were eliminated using the CD-HIT pipeline[16]. The MiSeq read pairs (2x250) were merged and adapter-trimmed with SeqPrep (https://github.com/jstjohn/SeqPrep) using default settings. Paired-end reads that did not overlap with at least 10 bases were subjected to stringent read filtering and trimming according to Minoche et al. 2011[17] prior to assembly. Reads were trimmed with a sliding window approach (window size 10 bases, shift 1 base). Bases were kept until the average

Illumina quality score Q of 10 adjacent bases was below Q=25. Reads were removed if they were smaller than 30 bases after trimming, had at least one uncalled base, contained the adapter sequence, or had less than two-thirds of the bases of the first half of the read with quality values of Q ≥ 30. In reads generated from pair-end libraries orphan reads were discarded in order to keep pairs only. Redundant read pairs that may originate from PCR artefacts were also removed by comparing the sequences of the read pairs. Out of 6,215,172 read pairs 28% could be merged and 60% passed the stringent filtering. The average length of the merged reads was 459 bp. The filtered MiSeq pairs were exclusively used for scaffolding by trimming them to 49 bases. We generated a total of 4.7 billion high-quality base pairs from 20.3 million high-quality reads. This represents 52.37-fold genome coverage, of which the Sanger reads provided 0.67-fold coverage, 454 reads provided 38.83-fold coverage and MiSeq reads provided 12.86-fold coverage (Supplementary Table 3). All high-quality reads were assembled into contigs containing 130 Mb and scaffolds spanning 130.09 Mb including embedded gaps (N50 = 28,028; Supplementary Table 4). The total length of the unfiltered assembly was about 40.05% higher than the genome size estimated by flow cytometry of isolated nuclei stained with DAPI (77.38 Mb; Supplementary Table 1, see also[12]).

1.5. Removal of organellar DNA and environmental sequence contamination

The 130.9 Mb, assembly comprised 57,732 sequences. Prior to analysis all low-complexity sequences were filtered out, especially artefacts and contaminating sequences that may have risen as a result of amplification. Our next generation sequence data shows an essentially unimodal distribution of local depth (coverage of each scaffold or contig estimated as total bases) when plotted against GC content (Supplementary Figure 1A). Since both GC-rich fragments and AT-rich fragments are always underrepresented in sequencing results, GC-content extremes around an extremely dominant mode can often be distinguished as contaminants or low-complexity sequences. The average GC content of the assembly was 40% and the local depth was ~35x in the majority of sequences (the major component). Scaffolds or contigs with significant differences in local depth (coverage > 50x or < 3x) also showed significant differences in GC content (Supplementary Figure 1B). Using a filtering strategy based on both GC content and assembly depth, we were able to cleanly classify misassigned *U. gibba* scaffolds and contigs (Supplementary Figure 1C). In total, 52,672 sequences spanning 49.03 Mb were identified as contaminants and removed from the assembly. The majority of these sequences were small contigs (with an average size ~850 bp) with extremely low coverage of ~3-4x and high GC content. These sequences were removed after confirming their likely environmental origin via significant match in BLAST comparisons to the NCBI refseq genomic database with plant genome sequences excluded. In the scaffolds or contigs with high coverage (≥ 60x), residual contamination was discovered to be from plant organellar DNA (see below, section 8). The high

proportion of contaminating sequences was expected since amplified DNA was used for constructing the sequencing libraries, and biases with respect to the distribution of amplified DNA are known[18]. We considered the remaining sequences after filtering (1,217 scaffolds and 3,843 contigs) to represent the *U. gibba* nuclear genome. All of these sequences showed significant matches against plant genomic sequences available in the refseq genomic database. This filtered assembly (at ~35x coverage) represented 81.87 Mb (N50 = 80,839; Supplementary Table 5), a total length 5.73% greater than the genome size estimated by flow cytometry (Supplementary Table 1).

1.6. Genome assembly validation

The assembly of *the U. gibba* genome was confirmed by single-pass primer walking re-sequencing of a ~100 Kb window (total) from two randomly selected scaffolds (Scf00089 and Scf00021; Supplementary Figure 2). A total of 211 sequences were generated with an estimated average size of 453.67 bp. Primers (described in Supplementary Table 6) were designed using Multiple Primer Design with Primer 3 (http://flypush.imgen.bcm.tmc.edu/primer/) with values set to produce primer pairs every 550 bp with an average and optimal length of 650 bp. The total overlap was 100 bp on average. Amplification was performed as follows: an initial step at 94°C for 10 min, followed by 35 cycles of 94°C for 20 s, 60°C for 30 s, and 72°C for 1 min, with the final step at 72°C was extended to 10 min. PCR products were sequenced after cleaning up with ExoSAP-IT (Affymetrix, USA). Additionally, using pCC1FOS™ vector (Epicentre) a fosmid library with ~1,000 clones was generated. Plasmid DNA was isolated using QIAprep Spin Miniprep Kits (QIAGEN, USA) and digested with *Not I* (New England BioLabs, USA). Insert size was then determined using CHEF gel electrophoresis. We sequenced 53 randomly selected clones (with insert size ranging from ~5-20 Kb, confirmed first as *U. gibba* by Sanger end-sequencing), using a Personal Genome Machine™ (PGM™) sequencer and a 3.18 semiconductor chips. A total of 4,973,037 reads (spanning 1.1 Gb) with an estimated average size of 229 bp were generated. The sequences were assembled using Newbler v2.6 (genomic option) with default parameters. A vector-trimming step was included in the assembly. The complete sequences of the 53 fosmids were obtained at an estimated coverage of ~250x (Supplemantary Data 1). The complete alignments of fosmid sequences to the *U. gibba* whole genome sequence revealed that we were able to generate a shotgun assembly with a low degree of misassembly (Supplementary Figure 3). Finally, the high coverage of the *U. gibba* nuclear genome was also confirmed using the Newbler Isotig sequences (see below section 2.3.1.). The genome assembly contains 99.45% of the 37,799 *U. gibba* Isotigs assembled from 4,687,343 sequenced ESTs (Supplementary Table 7).

2. Annotation

2.1. Identification of repetitive elements in the *U. gibba* genome

Transposable elements (TEs) in the *U. gibba* genome were identified both at the DNA and protein level. First, the REPET package[20] was used to search for TEs. TEs were classified according to Wicker's classification[21] (Supplementary Table 8). The classification takes into account the degree of completeness of the *de novo* TE consensus. For instance, if a consensus sequence has the required "structural features" — LTRs (long terminal repeats), TIRs (terminal inverted repeats) or a tail (poly-A or SSR-like [simple sequence repeats]) — and "coding features" — matches with known TEs in TBALSTX and BLASTX analyses — then it is considered "complete". If it has only one of these two types of features, it is classified as "incomplete". The coding sequence (CDS) and protein translation for each sequence was identified by comparison to available protein sequences (nr and Repbase databases) using the TransPipe pipeline[22]. Briefly, using BLASTX, best-hit proteins are paired with each gene at a minimum cut-off of 30% sequence similarity over at least 150 sites. To determine reading frame and generate estimated amino acid sequences, each gene was aligned against its best hit protein by Genewise 2.2.2[23]. Using the highest scoring Genewise DNA-protein alignments, custom Perl scripts were used to remove stop and 'N' containing codons and produce estimated amino acid sequences for each gene (Supplementary Table 9). A total of 532 TEs (both complete and incomplete) were identified, spanning a total of 2.5 Mb (3.1%) of *U. gibba* genome (Supplementary Table 8).

To confirm the degree of completeness of *U. gibba* LTR retrotransposons, characteristic elements (both 5'- and 3'-Long Terminal Repeats (LTRs), primer binding site (PBS), polypurine tract (PPT), conserved protein domains as IN (integrase), RT (reverse transcriptase) and RH (RNase H)) and their positions were identified using the LTR-Finder program[24] (Supplementary Figure 4 and Supplementary Data 2). LTR-Finder was used with default parameters. LTR TEs were considered only if they retained at least one of the LTR-retrotransposon characteristics such as a PBS, a PPT, or a conserved protein domain (IN, RT and/or RH) between both (5' and 3') LTRs. Using this approach for assessment of the intactness of the LTR retroelements, our data show a highly fragmented structure of LTR retrotransposon sequences. According to our analysis, only 15% of those retroelements present in the *U. gibba* genome are complete and therefore potentially capable of further retrotransposition. The high frequency of incomplete (or fragmented LTR TEs) associated with the deletions in *U. gibba* retroelements indicates that genome expansion through retrotransposon amplification can be counterbalanced by a gradual removal of the elements through illegitimate recombination[25,26]. Additionally, the LTRs of these elements were then used as query sequences in BLAST searches against the *U. gibba* genome with TEs masked. We identified many solo LTRs using this approach (Supplementary Figure 5),

but these were not characterised further because their highly fragmented structure made it difficult to determine the nature of specific rearrangements. The preponderance of solo LTRs suggests that unequal and illegitimate recombination is also a process that plays an important role in DNA loss in *U. gibba* genome. Illegitimate recombination is a process that has been seen as the driving force behind genome size decrease in *Arabidopsis thaliana* (Arabidopsis), removing at least fivefold more DNA than unequal homologous recombination[25].

2.2. Identification of noncoding RNA genes in the *U. gibba* genome

Non-coding RNAs (ncRNAs), including miRNA, small nuclear RNA, tRNA, ribosomal RNA and H/ACA-box small nucleolar RNA, were identified using INFERNAL software by searching against the Rfam database[27] (Supplementary Tables 10 and 11). The majority of them were also confirmed using software designed for specific types of RNA: tRNAscan-SE[28] for tRNAs, RNAMMER[29] for rRNA, snoscan[30] for snoRNAs, and SRPscan[31] for SRP RNA.

2.3. Identification of protein-coding genes in the *U. gibba* genome

2.3.1. Transcriptome sequencing

Total RNA was extracted from whole plants, shoot-like structures, inflorescences and traps using TRIZOL (Invitrogen) according to the manufacturer's instructions. To represent all *U. gibba* organs, 2 ug of RNA from each sample were pooled. cDNA synthesis was performed as described previously[32]. A total of 3,931,039 reads (with an estimated average size of 205 bases) were generated using a Personal Genome Machine™ (PGM™) sequencer and 3.18 semiconductor chips. These sequences were trimmed using SeqClean software (http://compbio.dfci.harvard.edu/tgi/software/) to eliminate sequence regions that would cause incorrect assembly (poly A/T tails, ends rich in undetermined bases, and low complexity sequences). To carry out the assembly process, 3,794,878 reads (96.5% of total reads, with an estimated average size of 185.29) were considered. In addition, we included in the assembly 817,792 pre-existing masked 454 reads generated in our laboratory (Accession number SRP005297[32]). These sequences were assembled with Newbler version 2.6 (using the -cdna option), producing a total of 37,799 Isotigs grouped in 21,775 Isogroups. Every Isotig, on average, was comprised of 112 reads and had a size of 868.29 bp.

2.3.2. Gene model prediction

The AUGUSTUS program[33] was trained on the *U. gibba* genome using the 37,799 Isotig sequences. First, using the $AUGUSTUS_{beta}$ web server training tool (http://bioinf.uni-greifswald.de/augustus-training-0.1/) and the *U. gibba* genome and transcriptome Isotigs, a data set with training gene structures (Supplementary Data 3) was generated. Using this training set, parameters required by AUGUSTUS were calculated. Gene models in the *U. gibba* genome

sequence were predicted, both *ab initio* and with hints, locally running AUGUSTUS with newly optimised parameters. A total of 28,494 gene models were predicted, with a mean coding sequence size of 1,023.92 bp and an average of 4.15 exons per gene (Supplementary Table 12). The *U. gibba* genome contains a similar number of genes than Arabidopsis, *Mimulus guttatus* (*Mimulus*), *Vitis vinifera* (grape) and *Carica papaya* (papaya) but a smaller number than *Solanum lycopersicum* (tomato) (Supplementary Table 12). 71.52% of genes were supported by transcriptional evidence, and 28.48% had an *ab initio* prediction. About 77.76% of the genes have homologues in the RefSeq plant or Arabidopsis protein databases, and 65.69% of the genes were assigned at least one protein domain using the protein families [Pfam;[34]] database (Supplementary Tables 13 and 14). A total of 41,034 protein domains with 4,297 distinct domain types were identified. The top 30 *U. gibba* Pfam domains are plotted in Supplementary Figure 6.

2.4. Construction of *U. gibba* Pfam domain families

Grouping genes according to similarities with known sequence signatures is a common approach for generating gene family classifications[35]. Classifying proteins based on their constituent domains is one of the most effective and efficient approaches to organise protein data both by structures and by evolutionary relationships[36]. In order to analyse the distribution of gene families over different plant species, we identified the Pfam domains present in gene models predicted in the Arabidopsis, tomato, grape, *Mimulus*, and papaya genomes (Supplementary Table 15). Gene models and their proteins were downloaded from the CoGe OrganismView database (http://genomevolution.org/CoGe/OrganismView.pl the same database (Pfam) and equal parameters to identify protein domains makes it possible to remove potential bias from comparisons of gene numbers within families.

To compare the abundance of domains in proteins of different plant species we used a modification of the method described by Stekel[37]. This method calculated a likelihood ratio (R) for comparing the abundance of a gene in any number of cDNA libraries. We used the method to compare the abundance of protein domains in the genes present over the six different plant genomes. Briefly, the likelihood ratio, denoted *Rj* for protein domain *j*, is given by the expression:

$$R_j = \sum_{i=1}^{m} x_{i,j} \log\left(\frac{x_{i,j}}{N_i f_j}\right)$$

where m represents the number of plant species, $x_{i,j}$ is the number of copies of domain j in the ith species and N_i is the total number of protein domains identified in the ith species. f_i is the frequency of copies of domain j in all of the species, given by the formula:

$$f_j = \frac{\sum_{i=1}^{m} x_{i,j}}{\sum_{i=1}^{m} N_i}$$

In a plant species in which there are no observed copies of the domain, that is, $x_{ij} = 0$, its contribution to *Rj* is zero. A total of 115 protein domain families with values of $Rj \geq 8$ showed significant differences among plant species (Supplementary Table 16).

Analysis of the distribution of protein domain families over different plant species reveals interesting insights into plant gene evolution, and identifies species-specific protein domains (e.g., PF06721; this family represents the C-terminus of a number of *Arabidopsis thaliana* hypothetical proteins of unknown function; family members contain a conserved DFD motif) and lineage-specific gene families (e.g., PF04776 and PF06746; proteins of unknown function, currently only identified in Brassicaceae), orphan genes (e.g., PF05617 and PF03478 proteins of unknown function present as single copy genes in Arabidopsis but not in other plant species), and conserved core genes across the green plant lineage (e.g., PF13650, with similar number of genes in *U. gibba* and *Mimulus*, but not in other plant species). In relation to other plant species, *U. gibba* shows fewer genes and/or domains in 40% (46 of 115) of the protein domain families identified with significant differences (R≥8) in number of members; however, this group represents less than 3% of total gene families grouped according to protein domains. In other words, 97% of gene families do not show significant differences among the plant species that we analysed. These data suggest a high proportion of genes lost after the *U. gibba* whole genome duplications (WGDs; see section 7, below); however, they also suggest a tendency to preserve a core set of genes distributed among the various gene families.

2.5. Expansions and contractions of *U. gibba* families

2.5.1 OrthoMCL analysis of protein family expansions and contractions

Clustering of orthologous (and close paralogous) genes in the *U. gibba*, Arabidopsis, tomato, grape and papaya genomes was performed using orthoMCL[38] on the translated protein sequences of all predicted genes. In our analysis we chose a stringent value for the e-value cut-off, $1E^{-10}$, in order to avoid false positive results (Supplementary Table 17). A total of 1,275 gene families are apparently absent in *U. gibba* genome (Supplementary Table 18). These families vary in size from 1-2 members to 25 members, and 57% of these are single-gene families. Additionally, a total of 1,804 gene families showed an increased number of genes in *U. gibba* (Supplementary Table 19).

2.5.1.1 Gene family annotations

All *U. gibba* gene models were processed through the Blast2GO program[39], which yields a set of GO annotations for each gene based on homology to proteins from other species as determined by BLAST. We used this software according to the default protocols and settings: BLAST searches were conducted for each protein (BLASTX, nr database, HSP cut-off length 33, report 20 hits, maximum e-value $1E^{-10}$), followed by mapping and annotation (e-value hit filter $1E^{-10}$, annotation cut-off 55, GO weight 5, HSP-hit coverage cut-off 20). We assigned 59,486 Gene Ontology (GO) terms to 16,699 or 58.6% of the 28,494 *U. gibba* genes (Supplementary Table 20). In order to establish a standard functional annotation process for different plant species, a similar approach was used to obtain the functional annotations of gene models predicted in the Arabidopsis, tomato, grape and papaya genomes (Supplementary Table 20).

2.5.1.2 Specific contractions

We surveyed 100 OrthoMCL clusters that contained genes from all genomes studied except for *U. gibba* (Supplementary Table 18). Based on their presence in both rosids and asterids, a number of interesting genes appear to have been lost from the *U. gibba* genome. *U. gibba* plants are noteworthy in their rootlessness, unusual embryogenesis (which frequently involves asymmetrical production of shoot apical organs and absence of true cotyledons), and frequent shoot-leaf indistinction.

Based on annotations of Arabidopsis orthologues, several of the genes missing in *U. gibba* were involved in aspects of root development and physiology: *WAK* (a cell wall-associated Ser/Thr kinase involved in cell elongation and lateral root development)[40], *NAXT1* (a nitrate efflux transporter mainly expressed in the cortex of adult roots)[41,42], *MYB48* and *MYB59* (nitrogen-responsive genes, involved in the regulation of cell cycle progression and root growth)[43,44] and *ANR1* (*ARABIDOPSIS NITRATE REGULATED 1* [*AGL44*], a root-specific MADS domain protein)[45]. Based on the absence of *ANR1* and the existence of other root-specific MADS box genes in Arabidopsis, we were motivated to perform a phylogenetic classification of the entire MADS box family (see below).

Other genes missing in *U. gibba* are specifically expressed or had function in embryos or cotyledons in other plants: AT1G68170 (a nodulin MtN21-like transporter, differentially expressed in mature and juvenile-phase shoots)[46], *PEI1* (an embryo-specific zinc finger transcription factor required for heart-stage embryo formation)[47], and *FD* and a paralogue (involved in flowering but also expressed in embryos and seed)[48].

Homologues of LOB (lateral organ boundaries) domain-containing protein 23 (LBD23) were also absent only in *U. gibba*.
In *U. gibba*, genes of the *CASPARIAN STRIP MEMBRANE DOMAIN PROTEIN* family, which mediate casparian strip formation in Arabidopsis roots[49], are reduced to single-copy, whereas 2-3 tomato, grape and papaya genes were present in the same orthogroup.

2.5.1.3 Specific expansions

We also surveyed genes from 100 OrthoMCL clusters with increased membership of *U. gibba* genes relative to other genomes studied (Supplementary Table 19). A number of additional orthogroups were only present in *U. gibba*, and some of these were also identified in our annotation process. Again, we focused on genes expressing in root, embryo, and lateral organs.

The TOP (TOPLESS) protein family, involved in transcriptional repression of root-promoting genes[50], had 7 members in *U. gibba*, compared to 2-6 in the other species. Interestingly, other root-functioning orthogroups were increased in membership (5 genes compared to 1-3), such as one containing *SHY2/IAA3*, which regulates multiple auxin responses in roots[51]. Another orthogroup (4 genes compared to 1-2) contained a multicopper oxidase that adjusts root meristem activity to Pi (inorganic phosphate) availability[52-54]. With rootlessness, *U. gibba* shoot or leaf organs must take over this function.

There were 6 homologues of RSM1, a small sub-family of single MYB transcription factors involved in embryo development[55], compared to 2-4 in other species.

A striking observation among the 100 *U. gibba*-increased orthogroups was 3 orthogroups representing members of different TCP (TEOSINTE BRANCHED1/CYCLOIDEA/PCF) transcription factor clades. These genes regulate multiple aspects of plant morphogenesis, including branching[56,57]. These findings motivated a phylogenetic classification of all *U. gibba* TCP genes to look at specific group expansions (see below).

Among the *U. gibba*-only orthogroups was a cluster containing 8 LOB homologues (of LBD41, LOB domain-containing protein 41) different from those specifically lost (above). Another comprised 5 *SPL* (squamosa-promoter binding protein-like) homologues, still other controllers of lateral organ development[58]. These findings suggest the possibility that new LOB and SPL functions related to the morphogenesis of *U. gibba*'s unusual lateral organs may be specific to its genome. Another *U. gibba*-specific cluster was related to *WOX1* (WUSCHEL-RELATED HOMEOBOX 1), the expression of which is confined to the initiating vascular primordium of the

cotyledons during heart and torpedo stages[59]. Still other *U. gibba*-specific orthogroups contained MADS box genes from different groups than those discussed above (see further analysis below).

2.5.2 Phylogenetic classifications of specific expanded and contracted transcription factor gene families

2.5.2.1 Analytical methods and basic results

We performed detailed phylogenetic classifications of 5 well-known transcription factor families to provide highly focused views of gene family expansion and contraction in *U. gibba* relative to Arabidopsis and tomato. Searches for MADS, TCP, GRAS, ARF, and AUX/IAA gene family sequences were performed throughout the whole proteomes of tomato (ITAG2.3 release) and *U. gibba* using HMMer v3.0[60]. Profile HMMs based on the alignment of Arabidopsis MADS[61] , TCP[62], and GRAS[63] protein domains or full length ARF[64] and AUX/IAA[65] proteins, respectively, were used as queries. Exon/intron location, distribution, and phases at the genomic sequences encoding for *U. gibba* MADSs and TCPs were predicted through comparisons with the predicted encoded protein using GENEWISE[23]. Phylogenetic analyses were performed on the basis of multiple alignments of amino acid sequences obtained using T-COFFEE[66] or MUSCLE[67]. Maximum Likelihood (ML) reconstructions were carried out using PhyML v3.0[68,69] and the best-fitting model selected by ProtTest v2.4 on the basis of the Akaike information criterion[70]; these were the LG (MADS, TCP), JTT+F (GRAS, ARF, AUX/IAA) models with a gamma-distribution with eight categories[71]. Tree topology searching was optimised using the subtree pruning and regrafting option. The statistical support of the retrieved topology was assessed using the Shimodaira-Hasegawa-like approximate likelihood ratio test[72]. Neighbour joining phylogenetic analyses were conducted in SeaView 4.3.3[73]. The evolutionary distances for neighbour joining phylogenetic reconstruction were computed using the Poisson correction method. To obtain statistical support on the resulting clades, a bootstrap analysis with 1000 replicates was performed. Resulting trees were represented and edited using FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

We identified a total of 82 MADS and 42 TCP sequences in *U. gibba*. Similar searches were performed in the asterid species tomato, resulting in the identification of 105 MADS and 36 TCP sequences respectively. The Arabidopsis reference has 108 MADS and 24 TCP genes. The MADS box gene family in *U. gibba* is therefore significantly reduced in size, while the TCP family is significantly larger. Likewise, the ARF and AUX/IAA families are largest in *U. gibba*, with 32 and 47 genes compared to 23 and 29 in Arabidopsis and 32 and 42 in tomato. The GRAS family, with 39 genes in *U. gibba*, is represented by 32 in Arabidopsis but is amplified to 47 in tomato (Supplementary Table 21). As such, there is no singular pattern of gene loss with decreasing genome size, but rather dynamic evolution of gene family size. Below, we detail gains

and losses in these particular transcription factor families that may have particular relevance in *U. gibba*.

2.5.2.2 MADS box genes

There are two main lineages of MADS-box genes, type I and type II, both of which are found in plants, yeast and animals[74]. Type I genes only share sequence similarity with type II genes in the MADS domain. Type II proteins in plants have three other domains, the K (keratin-like) domain, a less well conserved I (intervening) domain and the variable C-terminal region (C) and are therefore referred to as MIKC-type. These genes are best known for their roles in the specification of floral organ identity, in the regulation of flowering time and in other aspects of reproductive development[75,76]. However, MADS box genes are also widely expressed in vegetative tissues[61]. There is evidence that at least 50 MADS-box genes are expressed in Arabidopsis roots[61,77-79]. The AGL17-like type II clade is of particular note as all its members are expressed in roots and four (*AGL16*, *AGL17*, *AGL21* and *AGL44*) have been reported as root-specific, similarly to the type I genes *AGL26* and *AGL56*[45,61]. The type II *ANR1* (*AGL44*) and *XAL1* (*AGL12*) MADS-box genes are so far the only members of the family with characterised functions in roots. The *ANR1* gene has been identified as a component of a signalling pathway that regulates lateral root growth in response to changes in the external NO_3 supply[80] while *XAL1* is involved in root cell differentiation and flowering time[81]. It is interesting that *U. gibba*, which is rootless, has no genes grouping into these various root- expressed MADS-box gene clades (Supplementary Figure 7).

SOC1 (originally called *AGL20*), which has a well characterised role in the regulation of flowering time[82], is also expressed in shoots[45], and a possible role in a general response to nutrient stress has been suggested due to the gene's ability to respond to changes in phosphorus (P) and sulphur (but not nitrogen, N) supply. *U. gibba* has a considerably expanded SOC1-like clade in comparison with tomato and Arabidopsis. In *Utricularia vulgaris* it has been reported that investment in carnivory, calculated as the proportion of leaf biomass and leaf area comprising traps, is inversely proportional to the availability of P from non-carnivorous sources, whereas N showed no significant effect in the investment in carnivory[83]. The marked expansion in the *U. gibba* SOC1-like clade is consistent with the hypothesis that these genes are sensitive to P availability, and that P uptake from prey might be more important than that of N for *Utricularia* species.

2.5.2.3 TCP genes

Based on differences within their TCP domains, two main lineages of TCP proteins can be distinguished: class I (including the PCF subfamily) and class II (including the CIN and

CYC/TB1 subfamilies)[84]. Despite its smaller genome size, *U. gibba* shows a significant expansion in gene number in all three subfamilies (42 genes total) with respect to Arabidopsis and tomato (24 and 36 genes, respectively; Supplementary Table 21 and Supplementary Figure 8). One expanded clade comprised five *U. gibba TCP* genes grouping closely with Arabidopsis *PTF1* (*TCP13*) and its orthologue from tomato (Supplementary Figure 8). *PTF1* has been reported to express in cotyledons, particularly their vascular tissue[85]. Two other *U. gibba*-specific expansions of CIN-like genes have occurred in relatives of other Arabidopsis cotyledon-expressed genes[85], namely 4 genes grouping with Arabidopsis *TCP2/TCP24* (2 in tomato), and 5 genes clustered with *TCP10* (2 in tomato). Another expanded *U. gibba* clade was found in the CYC/TB1 subfamily, comprising 5 orthologues of the single Arabidopsis *BRC2* gene (2 in tomato). *BRC2* plays a key role in branching regulation by preventing bud outgrowth[86], being particularly associated with coordination of growth among branches in a phytochrome dependent manner[87]. It is tempting to speculate that these gene clade expansions may be related to the unusual cotyledonary structure of *U. gibba* (often asymmetrical, sometimes transformed into novel structures or even traps[88]), and its genus-wide diversity of branching patterns[2].

2.5.2.4 ARF and AUX/IAA genes

ARF and AUX/IAA transcription factors operate together in a number of auxin-dependent responses, including developmental processes in roots, shoots, embryos, cotyledons, and flowers[89]. Most previously defined subfamilies of these genes were represented in the *U. gibba* genome (Supplementary Table 21 and Supplementary Figure 9). The ARF-II clade, members of which (e.g., Arabidopsis *ETT*[90] and tomato *DR12*[91]) are involved in flower development, is significantly expanded (8 genes relative to 2 each in Arabidopsis and tomato). The ARF-V subfamily is also expanded, 6 genes relative to 3 and 4; the Arabidopsis members *ARF16* and *ARF10* are involved in root cap cell differentiation, although the *U. gibba* genes may not share this function (Supplementary Table 21 and Supplementary Figure 9A). Among the AUX/IAA-like genes, specific losses in *U. gibba* occur in small clades without known function (AUX/IAA-I and AUX/IAA-IV). In contrast, increased numbers of genes relative to Arabidopsis and tomato occur in 4 other lineages (AUX/IAA-II, VII, IX and XI)[92] containing genes mainly involved in root (*BDL*, *IAR2*) but also embryo, shoot and flower development[93,94] (Supplementary Table 21 and Supplementary Figure 9B). It will be interesting to investigate the roles for which these genes have been co-opted for in the evolution of a rootless species.

2.5.2.5 GRAS genes

GRAS transcription factors include the well-known root morphogenesis proteins SCARECROW (SCR)[95] and SHORTROOT (SHR)[96]. *U. gibba* genes were identified in the corresponding SCR and SHR subfamilies as well as in 8 others (Supplementary Table 21). Two subfamilies, SCL26

and TGRAS (tomato only), were absent from *U. gibba*. A considerable expansion, however, occurred in the HAM (HAIRY MERISTEM)[97] subfamily (7 genes in *U. gibba* compared to 4 in Arabidopsis and 3 in tomato), members of which are involved in shoot and root meristem indeterminacy (Supplementary Table 21 and Supplementary Figure 10).

2.6. *U. gibba* single-copy genes

It was recently reported that approximately 1,000 single-copy nuclear genes are shared among Arabidopsis, *Populus trichocarpa* (poplar), *Vitis vinifera* and *Oryza sativa* (rice)[98]. The majority of these genes are also present in the *Selaginella* and *Physcomitrella* genomes. There is evidence from Arabidopsis that genes that become single copy following WGD are more likely to return to single-copy status after subsequent genome duplications[99]. This suggests that there could be a small subset of single-copy nuclear genes that remain single copy throughout much of angiosperm diversity. Extensive loss of genes occurs after WGDs; however, assuming a random process, some duplicates may be retained, possibly followed by functional divergence. As a consequence, "single-copy" genes may in some cases become families that could exhibit variation in numbers of members. Using bidirectional best BLAST and synteny analysis (SynMap within CoGe), we discovered that 87.44% (824 of 948) of the previously reported single-copy genes were also present as single copy in the *U. gibba* genome. Three copies were identified from 3 single-copy genes (0.31%), two copies from 66 genes (6.96%), while 55 genes (5.82%) from this set were lost (Supplementary Table 22). Although these results suggest that paralogue gain:loss rates are close to 1:1, the 55 single-copy genes lost in *U. gibba* are apparently not essential because, with only three exceptions, insertion mutants have been reported for Arabidopsis orthologues (Supplementary Table 22). After similarly identifying orthologues in tomato, we discovered that there are a number of single-copy genes shared among Arabidopsis, poplar, grape and rice that were apparently lost in a lineage-specific manner. Except for rice (a monocot), the remaining species are rosid eudicots. *U. gibba* and tomato, which are asterids, have lost 8 single-copy genes otherwise shared among grape, Arabidopsis and poplar. Moreover, a total of 16 genes in *U. gibba* and tomato have increased their copy number to either 2 or 3. Furthermore, we identified a number of rice/grape/Arabidopsis/poplar genes (58) absent from tomato but present in *U. gibba* (these may be Lentibulariaceae-specific genes), while 46 genes were tomato-specific.

3. Promoter and untranslated region (UTR) analysis of *U. gibba*

In comparison with other angiosperms, *U. gibba* shows a smaller number of introns and also a smaller frequency of exons per gene. These results suggest that "non-essential" elements such as introns may be lost during the genome contraction process. Moreover, intergenic regions are substantially reduced in small genomes (like *U. gibba* and Arabidopsis; Supplementary Table

12). Lengths of both introns and intergenic regions are correlated with genome size (smaller genomes: shorter introns and intergenic regions[100]. The *U. gibba* and Arabidopsis genomes showed that this packing profile is an important contributor to the increase in gene density in these species.

3.1. Identification of UTRs in *U. gibba*

Intergenic regions encode essential regulatory elements such as promoters and terminators, which direct the accurate initiation and termination of transcription and prevent the expression of one gene from interfering with that of neighbouring genes. We estimated the average length of intergenic regions considering pairs of adjacent genes either convergent (→ ←), divergent (← →), or tandem (→ → or ← ←). *U. gibba*, like other plant species, showed the shortest intergenic region lengths between convergent gene pairs (Supplementary Table 23).

A total of 14 adjacent gene pairs (5 convergent, 4 divergent and 5 tandem) were selected to estimate UTR sizes in the *U. gibba* genome by amplification of cDNA ends (RACE-PCR). Using the Seaview program[73] and translated amino acid alignment to guide the alignment of nucleotide sequences, these *Utricularia* genes were compared against homologous Arabidopsis genes (Supplementary Data 4). Whole-plant total RNA from *U. gibba* was used for RACE-PCR as described in the GeneRacer™ Kit (Invitrogen, Life technologies). 2 µg of total RNA were used to carry out a 5'RACE-PCR reaction: 5' phosphate removal, RNA dephosphorylation and GeneRacer™ RNA Oligo (containing the priming sites for the GeneRacer™ 5'Primers) ligation, followed by reverse transcription. Reverse transcription for 3' RACE-PCR was carried out using 1 µg of original unligated total RNA. Both 5' and 3' transcriptions used GeneRacer™ Oligo dT Primer (containing the priming sites for the GeneRacer™ 3'Primers). Primary PCRs were carried out using 1µL of cDNA, gene-specific primers (Supplementary Table 24), hot start and touchdown PCR to minimise the background. HotStart-IT® FideliTaq™ Master Mix 2X (Affymetrix) was used with the following cycling parameters: 94°C for 2 min (1 cycle), 94°C for 30 sec, 72°C for 1 min (five cycles), 94°C for 30 sec, 70°C for 1 min (five cycles), 94°C for 30 sec, 65°C for 30 sec, 72C for 1 min (25 cycles), and 70°C for 10 min, in a 20ul reaction. Nested PCR was used to increase the specificity of RACE products for the 5' and 3' ends using 1µL of the original amplification reaction as a template, nested gene-specific primers (Supplementary Table 25) and Taq DNA Polymerase (Fermentas, Life Sciences). Cycling parameters used were: 94°C for 2 min (1 cycle), 94°C for 30 sec, 65°C for 30 sec, 72°C for 2 min (25 cycles) and 72°C for 10 min. Finally, 5µL of nested PCR reactions were analysed on a 1.2% agarose/ethidium bromide gel and the amplicons were sequenced unidirectionally using an ABI 3730xl sequencer (Applied Biosystems) (Supplementary Data 5). The average length of 3' UTRs was 269.69 bp, whereas for 5' UTRs the average was 149.45 bp (Supplementary Table 25).

We found that some adjacent convergent gene pairs overlapped in a portion of their 3' UTRs. The frequency of this phenomenon was 2 out of 5 convergent gene pairs tested. Although we only performed fine-scale analysis of intergenic regions between 195-427 bp (and the average size estimated was 1,039.90; see Supplementary Table 23) our data suggest that the *U. gibba* genome contains a high frequency of coding genes that overlap at their 3' ends. Additional evidence related to this phenomenon was found in our previously reported transcriptome assembly[32], in which 117 unique transcripts were identified that contained the CDSs of two neighbouring genes sharing a common polyadenylation region. The intergenic region size from these genes ranged from 59 to 925 bp, with an average length of 280.79 bp (Supplementary Table 26; Supplementary Data 6). In the *U. gibba* genome, 75% total of the convergent gene pairs have an intergenic region size ≤ 1000 bp, suggesting that a high proportion of convergent gene pairs may share a common polyadenylation region.

These sense and antisense poly(A) transcripts could participate in antisense-specific gene regulation, or could lead to the formation of dsRNA (natural antisense) substrates for RNA interference mechanisms that involve DICER-mediated cleavage and small RNA production[101,102]. In Arabidopsis (which also has a relatively small genome), similar sense-antisense transcripts have been reported[103]; however, alternative roles for these natural antisense transcripts have been suggested[104], or that they are simply targeted for degradation by the nonsense-mediated decay pathway.

3.2. Comparative analysis of the *rbcS* promoter

As in Arabidopsis, some extremely short intergenic regions (~150 bp) were detected in the *U. gibba* genome. These data suggest that some promoters have been contracted to minimal (or almost minimal) states. In the promoters of *rbcS* duplicates, a conserved family of genes contributing to the ribulose-1,5-bisphosphate carboxylase/oxygenase holoenzyme, a combination of at least two regulatory elements (the I- and G-boxes) is required to confer light responsiveness, although neither of these elements by themselves appears to be sufficient[105,106]. We analysed the upstream regions (400 bp) of selected *rbcS* genes from different plant species, including *U. gibba*. Using different programs (Weeder[107], Scope[108], rVISTA[109] and CoGE/GEvo), we identified the I- and G-boxes (and almost always, two other motifs) conserved in all species. Interestingly, the *U. gibba* rbcS promoter region in which these elements are contained is highly compacted toward the transcriptional start site (Supplementary Figure 11). These data suggest that some of the intergenic DNA contraction in *U. gibba* has been caused by microdeletions.

3.3. Transient expression assay

The functionality of some promoters in *U. gibba* was tested by transient expression assay. Specific primers (shown in Supplementary Table 27) for amplifying intergenic regions of each target gene were designed using the Primer3 version 4.0 website (http://frodo.wi.mit.edu/) with specific *U. gibba* contig sequence as the template. The PCR products were cloned into the pENTRTM TOPO vector (Invitrogen) and they were then transferred into the destination vector pKGWFS7 by recombination using a GATEWAY LR kit (Invitrogen) to generate transcriptional fusions and drive GFP-GUS expression. Transient gene expression was studied in Arabidopsis cell suspension culture. Cells were maintained at 25 ^{o}C with gentle agitation (125 rpm) in 50 ml of liquid growth medium supplemented with 2,4-dichlorphenoxyacetic acid, kinetin and sucrose (30 g/l). For bombardment, four days after transfer to fresh medium, Arabidopsis cells (0.343 g of fresh weight per 2 ml of medium) were loaded onto a 5 cm of filter paper (3MM Whatman) and placed on plant cell growth medium with 0.8% agar. The bombardment procedure was performed in a PDS/1000-He device (BIORAD, USA) essentially as described by Sanford[110,111]. 10 µg of each DNA was used for tungsten M10 particles. Following bombardment of cell suspensions, they were incubated in the dark for two days and then stained for *GUS* expression using GUS reaction buffer (0.5 mg/ml of 5-bromo-4-chloro-3-indolyl-b-Dglucuronide in 100 mM sodium phosphate, pH 7.0). Cell suspensions were incubated overnight at 37 °C. After the GUS reaction, they were observed in a LUMAR stereomicroscope (Zeiss). GUS expression was detected in five of the eight promoters tested (Supplementary Figure 12), including a 397 bp bidirectional promoter controlling a divergent gene pair.

4. RNA-mediated gene regulation pathways in *U. gibba*

We took a computational approach to gain insight into the different RNA-mediated gene regulatory pathways present in *U. gibba*. We used BLAST to look for genes similar to core components of the different small RNA mediated pathways[112], including microRNAs (miRNAs)[113,114] and short interfering RNAs (siRNAs)[115]. We found that essential genes involved in miRNA and siRNA biogenesis and function[112,113,115] are present in the *U. gibba* genome (Supplementary Table 28). miRNA prediction was performed by comparing all plant miRNAs (4,727 sequences) deposited in miRBase[116] against the *U. gibba* genome using the short read aligner bowtie[117] and a set of custom made PERL scripts. miRNA precursors were assayed with the UNAfold software[118]. We identified 75 miRNAs belonging to 19 families (Supplementary Table 29). All miRNA precursors fold into stable, minimum-free energy stem loop structures where the mature miRNA resides in the stem portion of the hairpin[119] (Supplementary Data 7). These results indicate that the general repertoire of RNA-mediated gene regulation mechanisms in plants is conserved in *U. gibba*. RNA-mediated gene regulation is essential for growth and

development in eukaryotic organisms[120-123], and is also responsible for the maintenance and reversal of epigenetic cellular memory, which records developmental and environmental cues[124]. Given the structural features and compact organisation of the *U. gibba* genome, it will be interesting to explore in the future how these diverse RNA-based mechanisms sense and respond to developmental and environmental cues and how the molecular processes are coordinated.

U. gibba contains only 379 retrotransposons, totalling ~2.6% of the genome (Supplementary Table 7). According to our analysis, only 15% of those retroelements present are complete and therefore potentially capable of further retrotransposition. Proliferation of retrotransposons (by a ‘‘copy and paste’’ mechanism) is involved in eukaryotic genome expansion, however, most retrotransposons are inactivated in plants by mechanisms involving DNA and histone modifications[125,126]. We found that homologues of all genes known to be involved in silencing of retrotransposons are present in the *U. gibba* genome (Supplementary Table 28). These data suggest that any influence of retrotransposon proliferation on *U. gibba* genome size must be countered by fractionation after WGDs (see section 7, below) and also by the silencing of these elements.

5. Genome compositional features of *U. gibba* compared to Arabidopsis

The small and highly compacted genome of *U. gibba* has a size of 82 Mb, whereas the Arabidopsis genome has a golden path 1.45 times longer (120 Mb). Transposable elements are largely responsible for the differences in genome size between these two species. *U. gibba* contains only 3.04% repetitive DNA whereas the Arabidopsis genome contains 12% (Supplementary Table 30). Although differences in gene space, ncRNAs and other repetitive sequences can also influence differences in genome size, basic genomic metrics reveal that intergenic regions size and TE numbers should be considered the principal contributors (Supplementary Table 30).

Differences in gene space can be attributed to fact that *U. gibba* shows fewer exons per gene than Arabidopsis, probably due to intron losses (see Supplementary Tables 30 and 12). Recent studies have shown that some eukaryotes have lost many introns, whereas others have gained many introns, and as consequence intron density in eukaryotic genomes varies considerably[127]. Currently, two main models are proposed for the mechanism of intron loss[128]: (1) deletion at the genome level[129]; and (2) homologous recombination between the genomic copy of a gene and the cDNA produced by the reverse transcription of its mature mRNA or partially spliced pre-mRNA[130]. Although the mechanism is poorly known, deletion under the first model can result in the exact removal of an intron region[131]. In order to evaluate intron loss from *U. gibba* genes, we first compared the number of introns in a total of 3,294 Arabidopsis and *U. gibba* orthologues

(Supplementary Table 31). As orthologues we consider those *U. gibba* and Arabidopsis genes grouped in the same othogroup (see Supplementary information section 2.5.1), provided these orthogroups contain a single member from each species (*U. gibba*, tomato, Arabidopsis, papaya and grape). The sum of the total number of introns identified in these gene models was smaller in *U. gibba* than in Arabidopsis while the CDS sizes were similar (±15% relative to Arabidopsis CDS, see Supplementary Table 31). Fewer introns were identified in 24.43% of 805 genes studied, the majority of which (82.83%) had 1-2 fewer introns (Supplementary Table 32 and Supplementary Data 8).

The apparent loss of introns might also reflect increased pesudogene number. We identified a total of 479 orthogroups (again, all of them containing genes from all species) that contained only one member from Arabidopsis and two from *U. gibba* (Supplementary Table 33). From these, we identified as putative pseudogenes only 23 candidates. An *U. gibba* gene that grouped with one or more *U. gibba* genes in the same orthogroup was considered a pseudogene if it met one of three criteria.

(i) A sequence was considered a pseudogene if its exon-intron structure was the same as that of its homologues, but CDSs were shorter; such pseudogenes may result from disruptive mutations such as frameshifts and premature stop codons.

(ii) We also considered as pseudogenes *U. gibba* sequences with 20% or more shorter CDS that also lacked one or more introns; such pseudogenes may result from incomplete copies of parental genes, or be the consequence of a mutation that disrupts the transcription and/or translation of the gene.

(iii) Finally, we documented retrotransposed pseudogenes, derived from intron-containing parental genes.

Although further analysis is warranted, these results suggest that approximately ~5% of *U. gibba* gene models could be considered pseudogenes. As such, in comparison with the Arabidopsis genome (which contains ~1,000 pseudogenes), *U. gibba* contains two times the number of pesudogenes, many of which probably result from the normal process of fractionation following whole-genome duplications (see below, section 7).

6. Population genomics of *U. gibba*

6.1 The Pairwise Sequentially Markovian Coalescent (PSMC) model

High-throughput genome sequencing provides an unprecedented opportunity for deciphering the population genetic information stored in single genomes. We applied the PSMC model, which was originally applied to human and other mammalian genomes, to study the history of *U. gibba* effective population size (Ne) over time. PSMC infers the local time to the most recent common ancestor of the present-day genome on the basis of the local density of heterozygotes by use of a

hidden Markov model in which the observation is a single diploid sequence[132]. PSMC utilises sequence reads as mapped to a reference genome to estimate historical fluctuations in Ne. Our use of the method assumes that the *U. gibba* genome is presently diploid despite its numerous WGDs. For scaling Ne, PSMC requires input of an estimated per-year mutation rate. Per-generation mutation rates have been shown to be generally related with genome size across a variety of organisms[133]. Taking this into account, we interpolated the mutation rates for several plants species, including *U. gibba*, using Lynch's published rate/genome-size relationship[133] (Supplementary Table 34). We mapped the *U. gibba* MiSeq genome reads using BWA, and then filtered them using SAMtools to obtain a mapping with approximately 10x coverage genome-wide. To scale PSMC results to real time, we assumed 3 years per *U. gibba* generation and a per-generation mutation rate (μ) of 3.2×10^{-9} (Supplementary Table 34). PSMC was otherwise conducted using default parameters. *U. gibba* N_e was estimated to be ~5,000 individuals from 10-25,000 years before present (BP), with the population represented by the modern genome coalescing ~600,000 years BP (Supplementary Figure 13A). Closer to the coalescent point, N_e was considerably larger, around 65,000, with a continuous decrease toward recent prehistory. Regardless, the magnitude of N_e over time is small and as such not conducive to augmenting global, weak selective forces that might favour genome size reduction[90]. Bootstrap values for 100 replicates frame the PSMC estimate.

Using a similar approach, but assuming 1 year per generation and a mutation rate (μ) of 4.1×10^{-9} per generation, we estimated the population size history of Arabidopsis (the raw reads from whole genome sequencing of *A. thaliana* Col-0 were used, as downloaded from GenBank accession number SRX158512). Arabidopsis coalesced more recently, approximately 25,000 years BP, with a N_e of ~15,000 (Supplementary Figure 13B). Unlike *U. gibba*, N_e increased toward recent prehistory, with the ~25,000 individuals at 10,000 years BP representing a small increase. Bootstrap analysis, also 100 replicates showed greater variation than in *U. gibba*.

In PSMC coalescent simulations, N_e is derived from heterozygosity of the sequenced genome (via $\theta = 4 N_e \mu$). For *U. gibba*, the average genome-wide θ calculated by PSMC was 1.54×10^{-3}. Expected heterozygosity (He) is closely correlated with θ when θ is small ($<< 1$), as here, since $He = \theta/(1 + \theta) \approx \theta$. For Arabidopsis, genome-wide θ was 0.99×10^{-3}, only slightly lower than *U. gibba* (although it should be noted that Arabidopsis neutral has been calculated as about 5 times greater using different methods[134]. As such, mutational diversity in the *U. gibba* genome is not appreciably enhanced over Arabidopsis, a finding that stands in contrast with earlier reports of enhanced molecular evolutionary rates based on selected gene alignments[32]. These earlier estimates were based on CDS alignments. To estimate θ values for *U. gibba* coding and non-coding regions separately, we mapped the MiSeq reads against concatenated CDSs predicted in

the *U. gibba* genome, and alternatively, against the genome assembly with these CDSs masked. Average coding and non-coding θ values were estimated to be 4.70×10^{-4} and 1.16×10^{-3}, respectively. As expected, the different heterozygosities of coding and noncoding regions suggest a lower mutation rate in coding sequences. Noncoding θ, although appropriately lower than genome-wide θ, is only slightly so. We attribute this unexpectedly small difference to the difficulty of mapping reads to the short intergenic and intronic regions apparent in the U. gibba genome; some coding sequence may have been inadvertently included. The magnitude of coding θ further undermines earlier interpretations of *U. gibba* molecular evolutionary rates based on limited CDSs (to be further addressed below). It is nonetheless possible that per-generation mutation rates in *U. gibba* might turn out to be higher than expected, and therefore we used PSMC to investigate N_e behaviour over time using an arbitrary rate value increased by 2x, i.e., 6.4×10^{-9}. It can be seen (Supplementary Figure 13C) that the overall behaviour of N_e is the same, although compressed on both the x- and y-axes to yield even smaller N_e estimates and shorter time to coalescence.

6.2 mlRho θ estimates

We also used the maximum-likelihood mlRho software[135,136] to evaluate genome-wide θ. Similarly to PSMC, the mlRho approach requires a diploid genome and a careful mapping of sequence reads to a genome assembly. We again used BWA to carefully mask out all the reads that map to multiple locations of the genome (i.e., gene duplicates, transposable elements, etc. The mlRho program generates joint maximum-likelihood estimates of heterozygosity $(\theta = 4N_e\mu)$ of the sequenced genome and sequencing error for a given sequencing project. For *U. gibba*, genome-wide θ was estimated to be 4.50×10^{-3}, somewhat larger than with PSMC (differing, however, by less than an order of magnitude), but much more similar to published estimates of Arabidopsis neutral θ [134]. As such, our point above regarding similar mutational diversity in *U. gibba* and Arabidopsis still holds.

To examine θ for different regions of our assembly, we performed window analyses of different numbers of nucleotides. We extracted non-overlapping windows from assembled scaffolds and analysed them similarly to the entire genome assembly. Window sizes used were 100Kb, 75Kb, 50Kb, and 25Kb across 101, 204, 482, and 1542 examples, respectively. A window size of 100Kb illustrated some θ heterogeneity across large stretches of the genome, with extremes at 5.4 and 1.4×10^{-03} (Supplementary Figure 14A). A moving average of 5 data points, however, revealed that most 100Kb blocks sampled varied only between 2.3 and 3.9×10^{-03} (a 1.7-fold difference). With the mlRho genome-wide average being 4.5×10^{-03}, we expected that smaller blocks of sequence would show greater θ heterogeneity. Indeed, for 75Kb windows, the minimum value was lower, 1.3×10^{-03}, and one extreme high was observed at $\sim 1.2 \times 10^{-02}$

(Supplementary Figure 14B). However, the 10-per moving average range was about 2.5-4.5 × 10^{-03} (a 1.8-fold difference), similar at the low end to 100Kb blocks, but trending higher toward the whole-genome average. Smaller window sizes revealed still more θ heterogeneity. For 50Kb windows the range was more extreme, 9.4 × 10^{-04} to 1.5 × 10^{-02}, with a 10-per moving average range of 2.2-5.8 × 10^{-03} (a 2.6-fold difference; Supplementary Figure 14C). Likewise, a window size of 25Kb showed similar extremes, from 7.0 × 10^{-04} to 2.3 × 10^{-02}, but still with 20-per moving average between 2.6-6.5 × 10^{-03} (a 2.5-fold difference; Supplementary Figure 14D). It is readily apparent from the 50Kb and 25Kb windows that θ outliers tend principally toward higher values. As such, we conclude that while most large (e.g., 25Kb) segments of the *U. gibba* genome (correspondingly, those capable of holding >5<10 genes) vary only as much as ~2-fold in heterozygosity, islands of considerably greater heterozygosity do exist. Since *Utricularia* species can have a mixed mating system with both selfing and outcrossing, strong variation in heterozygosity among chromosomal regions would be expected, since after even a single bout of selfing, in the next generation half of the chromosomal regions will be entirely homozygous while others that do not happen to experience shared inheritance will retain the heterozygosity of the parent. While there is always variation in levels of heterozygosity, even in randomly mating populations, this can become more extreme with partial inbreeding. In connection, it should be noted that since PSMC analysis assume random mating, the values obtained in Section 6.1 should be considered preliminary.

7. Polyploidy analyses

To examine WGD events we focused on comparing the genomes of *Utricularia gibba* (Ug), *Mimulus guttatus* (Mg), *Solanum lycopersicum* (Sl), and *Vitis vinifera* (Vv) using the comparative genomics system CoGe[137]. CoGe has several tools that were frequently used for identifying syntenic regions within and among genomes:

- SynMap[138]: SynMap was used for generating and visual whole-genome syntenic dotplots. The tool also includes a variety of options for modifying its visualisation scheme, identifying subsets of genes, and character large-scale evolutionary events such as WGDs or chromosome fusions. In addition, SynMap incorporates an additional algorithm, Quota Align[128], which can screen syntenic regions and select those giving a best user-defined ratio of coverage. Quota Align permits the rapid identification of orthologous syntenic gene sets between any two genomes. There are two major visualisation features that we employed within SynMap: (i) colouring syntenic gene pairs by synonymous substitution (Ks) values and (ii) ordering and orienting contigs based on synteny to a reference genome (also known as syntenic path assembly[139], SPA). Ks values, which are calculated using CodeML from the PAML package[140], may be used as a proxy for determining the relative age of genes. In SynMap, when syntenic gene pairs are coloured by Ks values,

syntenic regions derived from the same evolutionary event (e.g., polyploidy or divergence of lineages) tend to be coloured similarly[141]. By using the syntenic path assembly, evolutionarily or structurally related contigs that would otherwise be scattered across a dotplot will cluster, permitting the visualisation of evolutionary patterns such as polyploidy. Combined, SynMap's utilisation of visualisation and advanced comparative analytical tools permits the rapid characterisation of syntenic genes and genomic regions between any two genomes in CoGe's system. We used SynMap to characterise polyploidy events across entire genomes.

- GEvo[137]: GEvo is CoGe's tool for performing microsynteny analysis that permits comparison of multiple genomic regions with various algorithm and visualisation options. We used GEvo to validate synteny identified by SynMap across multiple genomic regions.
- SynFind: SynFind is CoGe's tool for identifying all regions across multiple genomes syntenic to a given gene, regardless of whether a homologous gene is present. SynFind was used extensively to find additional syntenic regions when comparing fragmented genomes such as Ug and Mg. In addition, SynFind will (i) generate master synteny tables where each gene in the reference genome has a list of all the identified syntenic genes/regions, which includes links to GEvo for validating the regions for microsynteny, and (ii) generate syntenic depth tables. Syntenic Depth measures the number of syntenic regions identified in genome A for a given gene in genome B. A syntenic depth of 0 means that no syntenic regions were identified; a syntenic depth of 1 means that one syntenic region was identified. We used SynFind to find potential syntenic regions for a given genomic region of interest by selecting a gene from the middle of that region.

Importantly, all of these tools permit on-the-fly analyses, let us manipulate parameters (e.g., higher or lower stringency), and are interconnected in order to characterise patterns of genome evolution, structure, and dynamics. A typical workflow would be to:

- Use SynMap with Ks colouration and syntenic path assembly to characterise whole genome polyploidy.
- Zoom-in on a pair of contigs/chromosomes that shows a pattern of polyploidy.
- Select a pair of genes from that region for microsynteny analysis with GEvo.
- Select a gene to fish out additional syntenic regions using SynFind.
- Validate all of the putatively syntenic regions using GEvo to ensure that each region covered the entire region of interest.

In addition, all of the tools in CoGe generate unique URLs that can be used to regenerate the previously run analysis. These URLs are included for all of our analyses. For recent reviews of

how to use these tools in CoGe for analysing plant genomes, please see Schnable and Lyons 2011[142] and Tang and Lyons 2012[141]

7.1. Summary of results

The results from the analyses detailed below are:

- Tomato's genome is a mix of singleton, duplicated and triplicated genome regions that arose after the eudicot paleohexaploidy, which is evidence for a more complicated genome evolutionary history than a straightforward whole genome triplication (WGT) followed by fractionation of homeologous gene content.
- *Mimulus guttatus* has had a WGD event subsequent to the eudicot paleohexaploidy event. This WGD is independent of tomato's most recent polyploidy event.
- *U. gibba* has had three sequential WGD events subsequent to the eudicot paleohexaploidy event. The most ancient of these WGDs may be shared with the most recent WGD of *M. guttatus*

7.1.1. *U. gibba* synteny analyses: evidence for at least two WGDs

The first step in character polyploidy is through intragenomic whole-genome analyses for synteny. Syntenic dotplots are one of the primary ways of visual the results of such an analysis. Supplementary Figure 15 shows a series of self-self syntenic dotplots for *U. gibba* required to unravel some of its polyploid history. Supplementary Figure 15A shows a self-self dotplot of *U. gibba* where contigs are ordered by size along each axis. While numerous small syntenic regions are identified as green dots, which are indicative of at least one polyploidy event in this lineage, this visualisation needs to be transformed into an easier form to interpret. Supplementary Figure 15B shows *U. gibba*'s contigs along the x-axis being arranged and ordered using the syntenic path assembly method (SPA). From this, it becomes clear that there is at least one round of polyploidy due to the syntenic signal along the 45-degree axis. However, there are several syntenic signals off this line, which may indicate a second, older polyploidy event. This can be further analysed by overlaying a colour scheme on the syntenic dots that corresponds to their relative age of divergence using Ks values. Supplementary Figure 15C shows this visual transformation using the Ks values show in the histogram in Supplementary Figure 15D. From this, it is apparent that the majority of genes comprising syntenic regions along the 45 degree line are from one age distribution (purple), and that there are numerous syntenic regions comprised of a different age class of gene pairs (cyan). The purple age class is younger than the cyan age class, indicative of at least two rounds of polyploidy in this lineage. However, it is not readily apparent from this view as to the nature of these polyploidy events (e.g. duplications or triplications). Supplementary Figure 15E shows a zoomed-in portion of the dotplot seen in Supplementary Figure 15C. Here, is obvious that for a given region of the *U. gibba* genome, there is one syntenic

region coloured purple, which is evidence for one WGD. For the cyan coloured regions, there are several cases where two occur for a given region of the genome. This is evidence for an older WGD event. This pattern of one recent syntenic region and two older syntenic regions is expected if there were two rounds of WGD in the lineage. These regions showing two older syntenic regions were analysed for microsynteny (Supplementary Figure 16A-D). In each of these analyses, there are a two pairs of regions showing a high degree of synteny, and the pairs of regions show, albeit more weakly, synteny between them. **This combination of macro- and microsynteny analyses provides strong evidence of at least two sequential WGDs in this lineage of *Utricularia***. It should be noted that given the highly reduced nature of this genome's size and the high degree of fractionation (homeologous gene loss) between syntenic regions derived from the second most recent WGD, identifying these cases is not trivial. In order to characterise this older WGD (and, as will be shown, an even older WGD) requires comparison to outgroup genomes that have not undergone all of these WGD events.

7.1.2. *Mimulus guttatus* synteny analyses: evidence for a WGD subsequent to the eudicot paleohexaploidy.

Mimulus guttatus (Mg) is an ideal comparator genome for *U. gibba* (Ug). However, before it can be used, its polyploidy history needs to be determined. Self-self synteny analysis shows that it has a relatively recent WGD superimposed on an older polyploidy event (Supplementary Figure 17A). The self-self syntenic dotplot shows that nearly the entire genome is covered by synteny from another part of the genome (Supplementary Figure 17B; purple regions), and microsynteny analysis of these regions shows the expected pattern of synteny with fractionated gene content (Supplementary Figure 17C). To determine whether the older syntenic regions (cyan) were derived from the eudicot paleohexaploidy event, the Mg genome was compared to the genome of *Vitis vinifera* (Vv). Vv has not had a WGD event since the eudicot paleohexplaoidy[143]. Whole genome syntenic dotplots of Mg versus Vv shows that there are two age classes of syntenic regions (Supplementary Figure 18A). Younger regions (purple) have a 2:1 syntenic relationship between Mg to Vv (Supplementary Figure 18C). This is the expected pattern if Mg has had a WGD subsequent to its divergence from the lineage of Vv. This pattern is confirmed by microsynteny analysis of one Vv region to two syntenic Mg regions (Supplementary Figure 18D). The Vv region contains nearly the entire gene content of the Mg regions combined. This pattern may be validated for nearly all regions of these genomes.

7.1.3. Syntenic analysis of *M. guttatus* and *U. gibba*: evidence that *U. gibba* has had three WGDs

Comparison of the genomes of *M. guttatus* (Mg) and *U. gibba* (Ug) are most revealing. The whole genome syntenic dotplot (Supplementary Figure 19) shows that nearly the entire genome of Ug is syntenic with at least one region of Mg. However, visually analysing the dotplot in more detail shows a pattern of many individual Mg regions being syntenic to four Ug regions (Supplementary Figure 20A). It can be noticed that the striking green colour of the syntenic lines seen in Supplementary Figure 19 are more difficult to discern in Supplementary Figure 20A. This is due to the order in which SynMap draws dots for gene pairs when Ks-value colours are used. SynMap will draw the younger dots (smaller Ks values) on top of the older dots (larger Ks values). This causes the lines to look mostly green in Supplementary Figure 19 when the dotplot is viewed at low resolution and this appears of all dots when the dotplot is viewed at high resolution (Supplementary Figure 20A). However, there is still a preponderance of green dots in Supplementary Figure 20A.

Seven of the 1xMg:4xUg regions identified in the syntenic dotplot (dashed boxes, Supplementary Figure 20A) were further characterised for microsynteny (Supplementary Figure 20B-H). Each of these analyses yields the expected pattern of fractionated gene content whereby nearly the entire gene content of the Mg region was contained in the four Ug regions (Supplementary Figure 21). This is indicative of Ug having undergone two sequential WGD events following its divergence from Mg, as well as evidence that the WGD in Mg is shared with Ug (Supplementary Figure 21). To further characterise this, we also expect that there will be an additional syntenic region within the Mg genome for each of identified Mg regions, and that region will be syntenic to an additional four regions of the Ug genome. This is due to Mg having had a WGD followed by two additional WGDs in the lineage of Ug. Together, this would create a syntenic set of regions that are comprised of 2xMg regions and 8xUg regions. Of the seven sets identified in Supplementary Figure 19A, we identified intragenomic syntenic regions for six of the Mg regions (Supplementary Figure 19B-G). In turn, five of the six newly identified Mg syntenic regions were syntenic to an additional set of four Ug regions (Supplementary Figure 19B, C, D, F, G), while one only identified three additional Ug syntenic regions (Supplementary Figure 19E). Note that since these are fragmented genome assemblies, there are places where multiple contigs were identified in order to provide full coverage of a syntenic region. Supplementary Figure 19E shows this for the case of Mg, where two contigs were used to represent full syntenic coverage to the other regions (Supplementary Figure 19G, where additional Ug regions were added). In total, 4.785 MB of the Mg genome and 1.854 Mb of the Ug were manually validated for microsynteny in these analyses. This represents 1.5% of the Mg genome and 2.3% of the Ug genome.

Next, we identified the syntenic region from *V. vinifera* (Vv) for two of the sets of validated regions shown above (Supplementary Figure 22). Here, we show that there is one Vv region that is syntenic to two Mg regions, which are in turn syntenic to eight Ug regions. This is the expected pattern of synteny if, following the divergence of these lineages, Vv underwent no subsequent polyploidy event, Mg had one WGD, and Ug had three WGDs. While our data suggests that the most ancient of the WGDs in Ug may be the same event as the most recent WGD seen in Mg, our current evidence is not conclusive. Overall, Ug regions appear to have more gene content retained compared to one of the Mg syntenic regions, and this pattern often fits well into quartets of Ug regions having more gene content in comment with one of the Mg regions. This would be expected if Ug shared Mg's WGD, and that Ug had two subsequent WGDs. However, close examination shows that gene content of many of the Ug regions is found split between the syntenic regions of Mg (albeit with more present on one region) (Supplementary Figure 23). A similar pattern is also seen with regard to the gene content of Mg as it is represented in Ug's syntenic regions. This could be explained by two mechanisms. The first is that all of Ug's WGD events are independent from the one in Mg, and the fractionation of homeologous gene content occurred independently in both lineages. The second is that they share a WGD event, followed by a small amount of fractionation, followed by the divergence of the lineages. The immediate ancestor to the divergence of the lineages would still have been very early in the diploidisation process, and the two lineages would have continued to undergo fractionation independently. This was further complicated by two subsequent WGD events in the lineage of Ug. While more genomes will be required to fully unravel the evolutionary relationships of these WGDs (especially the sequencing of a lineage that diverged between Ug and Mg as well as between the two most recent WGDs in Ug), both scenarios are remarkable for two reasons. The first is that Ug has had three WGD events despite its small genome size (and an additional whole genome triplication if the eudicot paleohexaploid event is included). The second is that Ug may have had all three WGD events subsequent to its divergence from Mg.

7.1.4 *U. gibba versus V. vinifera:* additional evidence of multiple rounds of polyploidy in the lineage of *Utricularia*

To further validate that *Utricularia gibba* (Ug) has had three WGD events, we compared its genome to that of *V. vinifera* (Vv). Their whole genome syntenic dotpot is shown in Supplementary Figure 24A. While Ug's contigs are ordered and arranged by SPA, this dotplot does not show any Ug contigs that do not show synteny to grape. A close up of a region of this dotplot (Supplementary Figure 24A, red dash box) shows that many regions of Vv's genome are each syntenic to multiple Ug contigs (Supplementary Figure 24C). This is expected if Ug has had three WGDs following the divergence of these lineages.

7.1.5. *Solanum lycopersicum* versus *V. vinifera*: character tomato's polyploidy

Another genome that is more closely related to *U. gibba* (Ug) than *V. vinifera* (Vv) is *S. lycopersicum* (tomato; Sl). However, before its genome can be compared to Ug's genome, its WGD event needs to be characterised. The published report on its genome states that it is underwent a whole genome triplication, but that much of its genome appears to have lost the signal of that event[144]. We characterised Sl's polyploidy through comparison to the genome of Vv. Supplementary Figure 25A shows a syntenic dotplot between the genomes of Vv and Sl that has been screened using Quota Align to show only the best three syntenic regions of Sl to grape. This syntenic screen helps cut down on older syntenic signals derived from the eudicot paleohexaploidy. In addition, the syntenic gene pairs are coloured according to their synonymous mutation values (Supplementary Figure 25B). This permits the differentiation of syntenic region that are orthologous (purple in Sup. Fig. 25) versus out-paralogous (cyan in Sup Fig L). We next visually annotated the dotplot for regions of the Vv genome that are orthologously represented once (green dash boxes), twice (blue dash boxes), and three times (red dash boxes) in the genome of Sl. Surprisingly, the majority of the Sl genome appears to be doubled, with the next major set being triplicated, and the final set being represented in one copy. While this is evidence that the genome evolution history of the Sl genome is more complicated than a single triploidy, we needed to determine which sets of Sl regions carry synteny from its most recent polyploidy event in order to best understand the polyploid nature of Ug. We examined duplicated and triplicated regions of Sl for microsynteny to Vv (Supplementary Figure 26). In both of these cases we saw the expected pattern of fractionated gene content across the Sl regions when compared to an unduplicated/untriplicated Vv region. In other words, the microsynteny analysis of the duplicated Sl regions (Supplementary Figure 26A) does not appear to be missing more genes than we saw from the microsynteny comparison to with the triplicated Sl regions (Supplementary Figure 26B). In addition, the pattern of some genomic regions of Sl being triplicated or duplicated (with the majority being duplicated) also shows in self-self syntenic dotplots of Sl (Supplementary Figure 27). For the purpose of comparing syntenic regions between S1 and Ug, some regions of S1 are treated as being duplicated and some triplicated.

7.1.6 *U. gibba* versus *S. lycopersicum*: additional evidence of multiple independent WGD events in the lineage of *Utricularia*.

A whole genome syntenic dotplot of *U. gibba* (Ug) versus *S. lycopersicum* (Sl) is shown in Supplementary Figure 28. Here, it is clear that, as previously seen for *Vitis vinifera*, Ug has multiple regions of its genome syntenic to a single region of Sl (Supplementary Figure 28C); such regions were analysed for microsynteny (Supplementary Figure 29). Since the genome of Sl may act as a functional tetraploid or functional hexaploid, we identified eight syntenic Ug regions

to either a pair (Supplementary Figure 29A) or a triplet set of Sl regions (Supplementary Figure 29B). Interestingly, the pattern of fractionation appears to be independent in the Ug regions when compared to the Sl regions, which is evidence that the polyploidy events in Ug are independent of the polyploidy event in Sl (Supplementary Figure 29A and B; coloured arrows). Each differentially fractionated syntenic gene between Sl and Ug was labelled by an arrow to signify which tomato region has lost it. Independence of polyploidy events is evidenced by each syntenic region of Ug having genes differentially lost among the Sl regions. If these lineages shared tomato's most recent polyploidy event, then we would expect an equal proportion of the Ug regions to be most similar to one region of tomato based on retention of gene content from their common ancestry. Instead, a given region of Sl appears to be dominant in terms of retaining gene content in Ug, but all tomato regions have their gene content represented among the combined *Utricularia* regions, not split to half of the Ug regions. This pattern is predicted by biased fractionation following polyploidy[145,146] in tomato. Since both Sl and *Mimulus guttatus* (Mg) have a polyploidy event in their lineage, but the Mg events appears to be a clean WGD while the polyploid status of Sl is a mix of duplicated and triplicated regions. We analysed syntenic regions to determine if their polyploidy events are shared or independent. Microsynteny analysis shows Sl and Mg regions show independent fractionation (Supplementary Figure 30), which is strong evidence that their polyploidy events are independent.

7.2. Randomised *U. gibba* genomes and the patterns of synteny

To test if the syntenic patterns observed when comparing the genome of *U. gibba* to the genomes of *Solanum lycopersicum* (Sl) and *Mimulus guttatus* (Mg) appear more often than random chance would predict, we generated 100 random permutations of the Ug genome and tested for significance of synteny. The random permutations of the Ug genome mimicked the quality of the wild-type (wt) genome by using the same number of contigs with the same number of genes per contig as observed in the wt genome. Our procedure for generating the randomised Ug genomes was:

1. Extract all genes from the Ug genome
2. Randomise the list of Ug genes
3. For each contig in the Ug genome
    a. Determine the number of genes the contig has
    b. Pick the same number of genes from the randomised list (without replacement)
    c. Use the random gene's CDS sequence to generate a new contig
    d. Add 200 nucleotides of "N" between each gene

These 100 randomised genomes were then added to CoGe and analysed for synteny using SynFind and SynMap. Overall, the randomised genomes showed a significant decrease in the

observed syntenic signal, nearly to the point of not identifying any syntenic regions. Syntenic dotplots showed nearly no syntenic regions when the randomised Ug genomes were compared to either Sl or Mg (Supplementary Figure 31). Statistical analysis of the number of genes in Mg or Sl at a particular depth all showed significant difference between the distribution of values obtained for the randomised genomes versus the wt genome (see below).

Syntenic depth refers to the number of times a genomic region (or genome) is syntenic to regions in another genome. In a typical case for two related organisms with no history of WGD, their syntenic depth is 1:1. This depth ratio changes when genomic regions are duplicated or deleted. For example, if one of the two genomes in the aforementioned example underwent WGD subsequent to the divergence of their lineages, then syntenic depth is 1:2. Estimates of syntenic depth using structural syntenic comparisons are complicated by two major factors: evolutionary time and completeness of genomic sequence. Genomes change over time, which obfuscates identifying syntenic regions, specifically when polyploidy is involved since the diploidisation process fractionates duplicated genes. Since many genome sequences are generated by NextGen shotgun sequencing, the resulting assemblies have many small chromosome fragments. Such small contigs often lack enough genes to infer synteny through either a colinear arrangement of genes[147] or through a local density of colinear genes[148], a problem that is exacerbated by genome evolution. SynFind permits a user to select one genome to which any number of additional genomes may be compared and screened for synteny. For each of these comparator genomes, SynFind identifies syntenic regions to the query genome using a Synteny-Score algorithm available from the TangTools package[149]. After identifying syntenic regions, SynFind generates a summary table of the number of syntenic regions identified for each gene in the query genome to each of the comparator genomes. These tables can provide evidence for the syntenic depth of the comparator genome to the query genome[150].

Our statistics used a two-tailed probability value of a z-test in order to assess the significance of the deviation of the value obtained by the wt versus the distribution of the randomised genomes (Supplementary Figure 32 and Tables 35 and 36). We tested syntenic depth with two different parameters sets, one stringent and one relaxed, for both Sl and Mg. In all cases, the number of genes at a particular syntenic depth was significantly different. Of note, the randomised genomes had fewer genes with synteny and showed none of the increased syntenic depths observed with the wt genome.

7.3. Syntenic depth tables

We compared tomato to *U. gibba* using two parameter sets that differ in the window size of genes used to define a minimum number of colinear genes allowing two regions to be called syntenic (Supplementary Tables 37 and 38). In both cases, a minimum number of four genes was required to seed the syntenic region. Due to the repeated number of WGDs inferred in these lineages since their divergence, one in tomato and three in *U. gibba*, and their shared whole genome triplication (basal to the eurosid-euastrid divergence), following by extensive fractionation in the *U. gibba* lineage, large gene window sizes are required to detect synteny. However, such large windows are prone to decreasing the signal to noise ratio by increasing the number of false-positive syntenic region calls. These tables provide evidence that *U. gibba* has undergone repeated WGD events since the divergence of these lineages, but, as stated above, they are not a strong sole source of evidence. When interpreting these tables, it is important to note that while there may be a given expectation of syntenic depth (in this case, a syntenic depth of eight *U. gibba* regions to one tomato region); post-polyploidy genome evolution can cause syntenic regions to become undetectable. In addition, since these genomes share a history of ancient whole-genome triplication, synteny from that event may further complicate the interpretation. However, given the strong microsynteny analyses showing eight regions of *U. gibba* being syntenic to one (or a pair) of tomato regions (above), these syntenic depth tables are in agreement with *U. gibba* having undergone three independent whole genome duplication events following the divergence with tomato. Supplemental Figure 33 summarises ploidy level findings for all genomes considered in this paper, both with respect to the pre- and post-hexaploidisation ancestor of core eudicots.

7.4. Fractionation depth

Fractionation depth refers to the number of syntenic genes that reduce to single-, double-, or n-copy over the course of *U. gibba*'s three independent WGDs since common ancestry with tomato (Supplementary Table 39). This table was generated using results from SynMap that generate a master table of all genes in tomato along with their matching syntenic regions in *U. gibba*. If a homologous *U. gibba* gene is present in an identified syntenic region, that gene was listed. If no gene was present, but the region was called syntenic due to neighbouring genes, the word "proxy" was listed. These results were parsed using a custom Perl program in order to tabulate the retention of *Utricularia* genes following *U. gibba*'s multiple WGD events. The looser parameter set from the tomato-*U. gibba* syntenic depth tables (see Supplementary Tables 37 and 38) was used for this analysis (gene window size of 160, at least four genes required to call a region syntenic) in order to capture as many syntenic regions as possible (i.e., to sacrifice a high false positive rate for increasing true positives and decreasing false negatives). As shown in Supplementary Table 39, the majority of *U. gibba* genes are retained as a single copy (62.39%).

22.10% were retained as two copies, and less than 9% were retained in three copies. Only 0.11% (11 genes) were retained in eight copies. This shows the strong effects of fractionation on the *U. gibba* genome following its series of 3 WGD events, which is expected given its small extant genome size.

7.5. Chromosome fusions

Suplementary Figure 34 shows syntenic mapping of one scaffold of the *U. gibba* genome to multiple genomic regions in tomato, providing evidence of multiple fusion events of ancestral chromosomes in the *U. gibba* lineage. Each pair of *U. gibba*-tomato regions has a series of colinear homologous gene pairs, which is evidence for synteny; each region of *U. gibba* matches approximately two or three regions of tomato, which is expected due to the independent duplications in that lineage. The single *U. gibba* region is syntenic to a series of regions of the tomato genome located on chromosomes 3, 6, 5, 4, which would be separated by tens of megabases if located on the same chromosomes. This provides evidence that during the multiple rounds of WGD and fractionation in the lineage of *U. gibba*, the genome underwent several chromosome fusion events. A tendency to fuse chromosomes is not unexpected for a genome undergoing a contraction process. Indeed, the genome of *Arabidopsis thaliana* (120 Mb golden path), compared to its close relative *Arabidopsis lyrata* (206 Mb genome), has had several putative chromosome fusion events that have reduced its chromosome count to five while *Arabidopsis lyrata* has retained eight. Since the lineages of these two *Arabidopsis* species diverged, neither has undergone any subsequent WGDs. Similarly, the genome of *Brachypodium distachyon* (272 Mb genome) has 5 chromosomes while it's more distant relative[151], *Oryza sativa* (374 Mb genome), has 12 chromosomes[152]. The chromosome fusions that have occurred in *Brachypodium*'s lineage are morphologically unique, showing repeated fusion events into centromeres of whole chromosomes, and whether a similar phenomenon has happened to the *U. gibba* genome is unclear. As more closely related genomes are sequenced and the *U. gibba* assembly improved, it will be interesting to try to determine the pattern of its chromosome fusion events.

8. Organelle genomes of *U. gibba*

Scaffolds or contigs consisting only of plastid or mitochondrial sequences were identified in the filtering process of *de novo* assembly (see section 1.5) through alignments to angiosperm chloroplast and mitochondrial genomes available in GenBank (232 and 43 genome sequences, respectively). Reference mapper software (Newbler v2.6) with default parameters was used. In total 48 sequences (38 from chloroplast and 10 from mitochondrial) spanning ~0.4 Mb were identified as organelle-like sequences and removed. The references showing the highest %

alignment were *Sesamum indicum*, with 66.98% coverage (chloroplast), and *Nicotiana tabacum* with 45.34% coverage (mitochondrial).

8.1. Plastid genome of *U. gibba*

Scaffolds/contigs originating from the chloroplast genome of *U. gibba* were merged into a single, unique sequence using the Megamerger program[153]. The chloroplast genome sequence of *Sesamum indicum* (NC_016433) was used as a guide for the alignment and determination of insertion (overlapping) points. The final, complete chloroplast genome sequence was initially annotated using the DOGMA program[154]. Annotations of protein-coding genes were refined through manual BlastX searches of the NCBI databases and annotations of tRNA genes were verified with the program tRNAscan-SE version 1.21[28,155].

The complete *U. gibba* plastid genome is 152,113 base pairs (bp) in length; it contains a pair of inverted repeats of 27,316 bp separated by two single-copy regions: the large single copy region is 81,819 bp long and the small single copy region is 15,662 bp. There are a total of 135 predicted coding regions, 95 of which are single copy (72 CDS and 23 tRNAs), and 20 of which are duplicated in the inverted repeats (9 CDS, 7 tRNAs, and 4 rRNAs; Supplementary Table 40 and Supplementary Figure 35). The *U. gibba* plastid genome is highly similar in size, gene content, and arrangement to the chloroplast genomes of other angiosperms, such as *Arabidopsis thaliana* and *Nicotiana tabacum* (Supplementary Figure 36). The most notable difference is that the *ndhF* gene is present in the inverted repeat region of our assembly, leading to the interruption of the ycf1 open reading frame that spans from the small single copy region into the inverted repeat region in other angiosperms. Eighteen genes contained at least one intron. Approximately 59% of the chloroplast genome is made up of coding DNA; introns represent 11%, and the remaining 30% of the genome is made up of non-coding, intergenic spacers, which is, again, similar to other angiosperms such as Arabidopsis and *Nicotiana* (Supplementary Table 41).

8.2. Mitochondrial genome of *U. gibba*

In comparison to non-plant unicellular and multicellular eukaryotes, plants have larger and more complex mitochondrial genomes[157]. All the features of plant mt genomes, including RNA editing, genomic recombination, trans-splicing, and insertions of "foreign" DNA from other genomes[158], make assembling mt genomes difficult. As recent studies have shown, genome sequences vary exceptionally in size, structure, and sequence content, especially among seed plants[159,160]. Ten unique scaffolds/contigs originating from the mitochondrial genome of *U. gibba* were identified during the process of *de novo* assembly using Newbler. These sequences ranged in size from 881 to 64,360 bp with an average length of approximately 22,215 bp. Their total, combined length is 222,145 bp, but this may not represent the size of the complete mitochondrial genome because

some regions may be duplicated or there may be some missing data. The sequences were initially annotated for common mitochondrial genes using the program Mitofy[159]. Annotations were refined using BlastX searches of NCBI databases, and the locations of tRNA genes were confirmed using tRNAscan-SE[28]. Mitochondrial copies of chloroplast genes were identified using Dogma[154]. Finally we used the Open Reading Frame Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) to identify any potentially coding regions that remained undiscovered.

Our *U. gibba* mitochondrial draft assembly includes a total of 55 genes, including 33 protein coding genes, 3 rRNAs, and 17 tRNAs (Supplementary Figure 37, Supplementary Table 42). The overall GC content was 46.03%. While incomplete, we found that most of the essential genes that are highly conserved in plant mt genomes, such as NADH dehydrogenase, succinate dehydrogenase, cytochrome *c* oxidase, and ATPsynthase, are present in our assembly. The only conserved gene that is apparently missing is *atp4*, although BLAST searches identified a partial sequence at the end of scaffold 01146 (Scf01146), indicating that it may in fact be present in the *U. gibba* mitochondrial genome. A total of 9 genes contained at least one intron. While most of the traditional mitochondrial genes were present in our assembly of the *U. gibba* mitochondrial genome, the order of these genes differs drastically from that of other plant species. Even though the chloroplast genomes of *U. gibba* and *Nicotiana tabacum* were highly syntenic, the mitochondrial genomes of these species are much less so (Supplementary Figure 38). Indeed, mitochondrial genome synteny is low even among *Arabidopsis thaliana*, *Brassica napus*, and *Carica papaya*, which all belong to the order Brassicales (Supplementary Figure 22). We identified 5 chloroplast-like regions in the mitochondrial contigs. This fact was not a surprise, because fragments of the chloroplast genome are frequently transferred to the mitochondrial genome[162]. In the *U. gibba* mitochondrial genome, these regions contained degenerate, pseudogene copies of *psbL*, *rps12*, *psbA*, *atpF*, *atpA*, and *rrn16*, as well as 5 tRNAs. In the complete genomes of *Nicotiana tabacum* and *Arabidopsis thaliana* intergenic regions are on average 1,818.61 bp and 2,053.03 bp, respectively. Similarly, the average intergenic region size in the current assembly of *U. gibba* mitochondria genome is 2,107.45 bp (Supplementary Table 43).

9. Molecular Dating Analyses

In order to investigate the divergence time of *U. gibba*, we obtained phylogenetic data sets for the family Lentibulariaceae from three regions of the chloroplast genome and one region of the mitochondrial genome. The nucleotide data sets were downloaded from the PhyLoTA Browser (http://phylota.net/) and included the *trnL* gene and *trnL-trnF* intergenic spacer region (1237 bp, 82 taxa), the *trnK* and *matk* genes (3048 bp, 101 taxa), the *rps16* intron (1371 bp, 131 taxa), and partial sequences of the *coxI* gene (708 bp, 34 taxa). Unfortunately no nuclear data sets with

appropriate taxonomic sampling were available. Species from the genera *Pinguicula* and *Genlisea* were used as outgroups.

We used the sequence alignments provided by the PhyLoTA browser for the *trnL* and *coxI* regions. We used the program MAFFT[163] to generate sequence alignments for the *rps16* data set, because no alignment was available, as well as for the matk data set, to which we added our own sequence from the *U. gibba* chloroplast genome. We used the Akaike Information Criterion (AIC) and the program jModelTest version 0.1.1[68,164] to investigate the nucleotide substitution model that best fit each data set. The GTR + G model was selected for the trnL and rps16 data sets, the GTR + I model was selected for the cox1 region, and the GTR + I + G model was selected for the matk region.

Divergence time estimates were obtained using the program BEAST version 1.7.1[165]. The appropriate nucleotide substitution model was selected, and where necessary, site rate heterogeneity was modelled by a gamma distribution with four discrete rate categories. Similar results were obtained from preliminary analyses using empirical and estimated base frequencies, therefore empirical base frequencies were utilised for final analyses. In all cases starting trees were randomly selected; however for the coxI data set, we enforced monophyly on the genera *Utricularia* and *Genlisea* because of positive selection on this gene[166]. We implemented the uncorrelated lognormal relaxed clock for divergence date estimates. The tree was calibrated by setting a lognormal prior on the estimated divergence time of the common ancestor of *Pinguicula* and *Utricularia/Genlisea* of 42 million years (31-54 million years;[167]). We used a Yule Process prior for the tree and default distributions for the remaining priors. Multiple independent runs were performed for between 1.5×10^7 and 1.0×10^8 generations, with the first 10% of the generations discarded as burn-in. Stationarity was investigated by examining plots of the –lnL across generations in Tracer version 1.5. For the vast majority of the parameters the effective sample size (ESS) was greater than 1000.

Overall, the divergence date estimates had wide 95% HPD (highest posterior density) intervals but median estimates were relatively consistent between different data sets despite differences in taxonomic sampling (Supplementary Figure 39). The wide HPD intervals likely reflect the uncertainty in the calibration point, as well as the fact that only one dating point was available. From the analyses conducted here it appears that *U. gibba* diverged from other *Utricularia* species approximately 13 million years ago (mya; Supplementary Table 44), although support for the branch leading to this split in the coxI data set had low support. The divergence date for the *Utricularia* crown group was about 28.5 mya (14.4 – 43.8 mya).

10. Evolutionary Rates

Results from previous work have suggested that members of the genus *Utricularia* have a higher mutation rate, and hence higher evolutionary rate, than other plants in the family Lentibulariaceae[32]. Therefore, we examined the estimates of evolutionary rates from the BEAST analyses. Overall, the relaxed clock analyses indicated that rates of evolution were variable across the tree. The 95% HPD interval for coefficient of variation did not overlap zero for any data set, indicating that the rates are not clock-like. The 95% HPD for the covariance did include zero, suggesting that evolutionary rates are uncorrelated in the tree. The mean substitution rates across the tree were similar for the three chloroplast regions were very similar ($3.16 \times 10^{-3} – 4.1 \times 10^{-3}$ substitutions per site per million years), but the substitution rate for the mitochondrial coxI gene was significantly lower (3.1×10^{-4} substitutions per site per million years), potentially due to selection on this region. Further investigation of the trees suggested that rate variation occurred primarily at the tips of the tree, and that rates were similar for deeper branches (Supplementary Figure 40). The rates for the branches leading to the crown groups of *Utricularia*, *Genlisea*, and *Pinguicula* were not significantly different (Supplementary Table 45), although 95% HPD intervals were wide. The discovery of additional fossils that can be used as calibration points would be helpful.

To further investigate differences in evolutionary rates, we conducted pairwise relative rate tests using the maximum likelihood program HyPhy[168]. For these tests we used *Byblis liniflora* as an outroup since sequences from this species were available for all four data sets. After the addition of the outgroup, sequences were re-aligned using MAFFT. We performed pairwise comparisons between *U. gibba* and all other species of Lentibulariaceae in the data set, using the Bonferroni correction for multiple tests.

Overall, for most data sets the relative rate tests from Hyphy suggest that there is little or no significant difference in rate between *U. gibba* and other members of its own genus, or between it and members of the other genera of Lentibulariaceae (Supplementary Table 46). However, significant differences were found in comparison of *Utricularia* to most or all species of *Pinguicula* for the *trnL* and *matk* regions. In these cases, the rate for *U. gibba* was nearly twice as high as for *Pinguicula* species. However, in general the BEAST and relative rate tests seem to indicate that *U. gibba* is not evolving at a faster rate than other species.

11. References

1 Chormanski, T. A. & Richards, J. H. An architectural model for the bladderwort *Utricularia gibba* (Lentibulariaceae). *The Journal of the Torrey Botanical Society* **139**, 137-148, (2012).

2 Taylor, P. G. *The genus Utricularia L. a taxonomical monograph*. (London: Her Majesty's Stationery Office, 1989).

3 Adamec, L. Mineral nutrition of carnivorous plants: A review. *The Botanical Review* **63**, 273-299, (1997).

4 Jobson, R. W., Morris, E. C. & Burgin, S. Carnivory and nitrogen supply affect the growth of the bladderwort *Utricularia uliginosa*. *Australian Journal of Botany* **48**, 549-560, (2000).

5 Adamec, L. r. Mineral nutrient relations in the aquatic carnivorous plant *Utricularia australis* and its investment in carnivory. *Fundamental and Applied Limnology* **171**, 175-183 (2008).

6 Sirová, D. *et al.* Ecological implications of organic carbon dynamics in the traps of aquatic carnivorous *Utricularia* plants. *Functional Plant Biology* **38**, 583-593, (2011).

7 Hobbhahn, N., Kuchmeister, H. & Porembski, S. Pollination biology of mass flowering terrestrial *Utricularia* species (lentibulariaceae) in the Indian Western Ghats. *Plant Biology (Stuttg)* **8**, 791-804, (2006).

8 Kondo, K. A Comparison of Variability in *Utricularia cornuta* and *Utricularia juncea*. *American Journal of Botany* **59**, 23-37, (1972).

9 Araki, S. & Kadono, Y. Restricted seed contribution and clonal dominance in a free-floating aquatic plant *Utricularia australis* R. Br. in southwestern Japan. *Ecological Research* **18**, 599-609, (2003).

10 Kameyama, Y. & Ohara, M. Predominance of clonal reproduction, but recombinant origins of new genotypes in the free-floating aquatic bladderwort *Utricularia australis* f. tenuicaulis (Lentibulariaceae). *Journal of Plant Research* **119**, 357-362, (2006).

11 Doležel, J., Bartoš, J., Voglmayr, H. & Greilhuber, J. Letter to the editor. *Cytometry Part A* **51A**, 127-128, (2003).

12 Greilhuber, J. *et al.* Smallest Angiosperm Genomes Found in Lentibulariaceae, with Chromosomes of Bacterial Size. *Plant Biology (Sttug)* **8**, 770-777, (2006).

13 Steinmüller, K. & Apel, K. A simple and efficient procedure for isolating plant chromatin which is suitable for studies of DNase I-sensitive domains and hypersensitive sites. *Plant Molecular Biology* **7**, 87-94, (1986).

14 Rabinowicz, P. D. *et al.* Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nature Genetics* **23**, 305-308 (1999).

15 Li, S. & Chou, H.-H. Lucy2: an interactive DNA sequence quality trimming and vector removal tool. *Bioinformatics* **20**, 2865-2866, (2004).

16 Niu, B., Fu, L., Sun, S. & Li, W. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* **11**, 187 (2010).

17 Minoche, A., Dohm, J. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology* **12**, R112 (2011).

18 Abbai, N. S., Govender, A., Shaik, R. & Pillay, B. Pyrosequence analysis of unamplified and whole genome amplified DNA from hydrocarbon-contaminated groundwater. *Molecular Biotechnology* **50**, 39-48, (2012).

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

20 Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, e16526, (2011).

21 Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**, 973-982 (2007).

22 Barker, M. S. *et al.* EvoPipes.net: Bioinformatic Tools for Ecological and Evolutionary Genomics. *Evolutionary Bioinformatics* **6**, 143, (2010).

23 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Research* **14**, 988-995, (2004).

24 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265-268, (2007).

25 Devos, K. M., Brown, J. K. M. & Bennetzen, J. L. Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis. *Genome Research* **12**, 1075-1079, (2002).

26 Ma, J., Devos, K. M. & Bennetzen, J. L. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Research* **14**, 860-869, (2004).

27 Gardner, P. P. *et al.* Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Research* **39**, D141-145, (2011).
28 Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research* **25**, 0955-0964, (1997).
29 Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**, 3100-3108, (2007).
30 Lowe, T. M. & Eddy, S. R. A Computational Screen for Methylation Guide snoRNAs in Yeast. *Science* **283**, 1168-1171, (1999).
31 Regalia, M., Rosenblad, M. A. & Samuelsson, T. Prediction of signal recognition particle RNA genes. *Nucleic Acids Research* **30**, 3368-3377, (2002).
32 Ibarra-Laclette, E. *et al.* Transcriptomics and molecular evolutionary rate analysis of the bladderwort (*Utricularia*), a carnivorous plant with a minimal genome. *BMC Plant Biology* **11**, 101 (2011).
33 Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
34 Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Research* **32**, D138-141, (2004).
35 Frech, C. & Chen, N. Genome-wide comparative gene family classification. *PLoS One* **5**, e13409, (2010).
36 Zhang, Y., Chandonia, J. M., Ding, C. & Holbrook, S. R. Comparative mapping of sequence-based and structure-based protein domains. *BMC Bioinformatics* **6**, 77, (2005).
37 Stekel, D. J., Git, Y. & Falciani, F. The comparison of gene expression from multiple cDNA libraries. *Genome Research* **10**, 2055-2061 (2000).
38 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**, 2178-2189, (2003).
39 Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676, (2005).
40 Lally, D., Ingmire, P., Tong, H. Y. & He, Z. H. Antisense expression of a cell wall-associated protein kinase, WAK4, inhibits cell elongation and alters morphology. *The Plant Cell* **13**, 1317-1331 (2001).
41 Segonzac, C. *et al.* Nitrate Efflux at the Root Plasma Membrane: Identification of an Arabidopsis Excretion Transporter. *The Plant Cell Online* **19**, 3760-3777, (2007).
42 Lin, S. H. *et al.* Mutation of the Arabidopsis NRT1.5 nitrate transporter causes defective root-to-shoot nitrate transport. *The Plant Cell* **20**, 2514-2528, (2008).
43 Bi, Y. M., Wang, R. L., Zhu, T. & Rothstein, S. J. Global transcription profiling reveals differential responses to chronic nitrogen stress and putative nitrogen regulatory components in Arabidopsis. *BMC Genomics* **8**, 281, (2007).
44 Mu, R. L. *et al.* An R2R3-type transcription factor gene AtMYB59 regulates root growth and cell cycle progression in Arabidopsis. *Cell Research* **19**, 1291-1304, (2009).
45 Gan, Y., Filleur, S., Rahman, A., Gotensparre, S. & Forde, B. Nutritional regulation of ANR1 and other root-expressed MADS-box genes in *Arabidopsis thaliana*. *Planta* **222**, 730-742, (2005).
46 Busov, V. B. *et al.* An auxin-inducible gene from loblolly pine (*Pinus taeda* L.) is differentially expressed in mature and juvenile-phase shoots and encodes a putative transmembrane protein. *Planta* **218**, 916-927, (2004).
47 Li, Z. & Thomas, T. L. PEI1, an embryo-specific zinc finger protein gene required for heart-stage embryo formation in Arabidopsis. *The Plant Cell* **10**, 383-398 (1998).
48 Huang, N. C., Jane, W. N., Chen, J. & Yu, T. S. *Arabidopsis thaliana* CENTRORADIALIS homologue (ATC) acts systemically to inhibit floral initiation in Arabidopsis. *The Plant Journal : for cell and molecular biology*, (2012).
49 Roppolo, D. *et al.* A novel protein family mediates Casparian strip formation in the endodermis. *Nature* **473**, 380-383, doi:10.1038/nature10070 (2011).
50 Pauwels, L. *et al.* NINJA connects the co-repressor TOPLESS to jasmonate signalling. *Nature* **464**, 788-791, (2010).
51 Goh, T., Kasahara, H., Mimura, T., Kamiya, Y. & Fukaki, H. Multiple AUX/IAA-ARF modules regulate lateral root formation: the role of Arabidopsis SHY2/IAA3-mediated auxin signalling. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **367**, 1461-1468, (2012).
52 Svistoonoff, S. *et al.* Root tip contact with low-phosphate media reprograms plant root architecture. *Nature Genetics* **39**, 792-796, (2007).
53 Ward, J. T., Lahner, B., Yakubova, E., Salt, D. E. & Raghothama, K. G. The effect of iron on the primary root elongation of Arabidopsis during phosphate deficiency. *Plant Physiology* **147**, 1181-1191, (2008).

54 Ticconi, C. A. *et al.* ER-resident proteins PDR2 and LPR1 mediate the developmental response of root meristems to phosphate availability. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 14174-14179, (2009).

55 Hamaguchi, A. *et al.* A small subfamily of Arabidopsis RADIALIS-LIKE SANT/MYB genes: a link to HOOKLESS1-mediated signal transduction during early morphogenesis. *Bioscience, Biotechnology, and Biochemistry* **72**, 2687-2696 (2008).

56 Kieffer, M., Master, V., Waites, R. & Davies, B. TCP14 and TCP15 affect internode length and leaf shape in Arabidopsis. *The Plant Journal : for cell and molecular biology* **68**, 147-158, (2011).

57 Steiner, E. *et al.* The Arabidopsis O-linked N-acetylglucosamine transferase SPINDLY interacts with class I TCPs to facilitate cytokinin responses in leaves and flowers. *The Plant Cell* **24**, 96-108, (2012).

58 Zhang, Y., Schwarz, S., Saedler, H. & Huijser, P. SPL8, a local regulator in a subset of gibberellin-mediated developmental processes in Arabidopsis. *Plant Molecular Biology* **63**, 429-439, (2007).

59 Nakata, M. *et al.* Roles of the middle domain-specific WUSCHEL-RELATED HOMEOBOX genes in early development of leaves in Arabidopsis. *The Plant Cell* **24**, 519-535, (2012).

60 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Computational Biology* **7**, e1002195, (2011).

61 Parenicova, L. *et al.* Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. *The Plant Cell* **15**, 1538-1551 (2003).

62 Cubas, P., Lauter, N., Doebley, J. & Coen, E. The TCP domain: a motif found in proteins regulating plant growth and development. *The Plant Journal : for cell and molecular biology* **18**, 215-222 (1999).

63 Bolle, C. The role of GRAS proteins in plant signal transduction and development. *Planta* **218**, 683-692, (2004).

64 Okushima, Y. *et al.* Functional genomic analysis of the AUXIN RESPONSE FACTOR gene family members in *Arabidopsis thaliana*: unique and overlapping functions of ARF7 and ARF19. *The Plant Cell* **17**, 444-463, (2005).

65 Overvoorde, P. J. *et al.* Functional genomic analysis of the AUXIN/INDOLE-3-ACETIC ACID gene family members in *Arabidopsis thaliana*. *The Plant Cell* **17**, 3282-3300, (2005).

66 Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**, 205-217, (2000).

67 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797, (2004).

68 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**, 696-704, (2003).

69 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**, 307-321, (2010).

70 Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-2105, (2005).

71 Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Molecular Biology and Evolution* **25**, 1307-1320, (2008).

72 Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology* **55**, 539-552, (2006).

73 Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* **27**, 221-224, (2010).

74 Alvarez-Buylla, E. R. *et al.* An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5328-5333 (2000).

75 Messenguy, F. & Dubois, E. Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development. *Gene* **316**, 1-21 (2003).

76 Jack, T. Molecular and genetic mechanisms of floral control. *The Plant Cell* **16 Suppl**, S1-17, (2004).

77 Rounsley, S. D., Ditta, G. S. & Yanofsky, M. F. Diverse roles for MADS box genes in Arabidopsis development. *The Plant Cell* **7**, 1259-1269, (1995).

78 Alvarez-Buylla, E. R. *et al.* MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. *The Plant Journal : for cell and molecular biology* **24**, 457-466 (2000).

79 Burgeff, C., Liljegren, S. J., Tapia-Lopez, R., Yanofsky, M. F. & Alvarez-Buylla, E. R. MADS-box gene expression in lateral primordia, meristems and differentiated tissues of *Arabidopsis thaliana* roots. *Planta* **214**, 365-372 (2002).

80 Zhang, H. & Forde, B. G. An Arabidopsis MADS Box Gene That Controls Nutrient-Induced Changes in Root Architecture. *Science* **279**, 407-409, (1998).

81 Tapia-López, R. *et al.* An AGAMOUS-Related MADS-Box Gene, XAL1 (AGL12), Regulates Root Meristem Cell Proliferation and Flowering Transition in Arabidopsis. *Plant Physiology* **146**, 1182-1192, (2008).

82 Komeda, Y. Genetic regulation of time to flower in *Arabidopsis thaliana*. *Annual Review of Plant Biology* **55**, 521-535, (2004).

83 Kibriya, S. & Iwan Jones, J. Nutrient availability and the carnivorous habit in *Utricularia vulgaris*. *Freshwater Biology* **52**, 500-509, (2007).

84 Martin-Trillo, M. & Cubas, P. TCP genes: a family snapshot ten years later. *Trends in Plant Science* **15**, 31-39, (2010).

85 Koyama, T., Furutani, M., Tasaka, M. & Ohme-Takagi, M. TCP transcription factors control the morphology of shoot lateral organs via negative regulation of the expression of boundary-specific genes in Arabidopsis. *The Plant Cell* **19**, 473-484, (2007).

86 Aguilar-Martinez, J. A., Poza-Carrion, C. & Cubas, P. Arabidopsis BRANCHED1 acts as an integrator of branching signals within axillary buds. *The Plant Cell* **19**, 458-472, (2007).

87 Finlayson, S. A., Krishnareddy, S. R., Kebrom, T. H. & Casal, J. J. Phytochrome regulation of branching in Arabidopsis. *Plant Physiology* **152**, 1914-1927, (2010).

88 Plachno, B. J. & Swiatek, P. Unusual embryo structure in viviparous *Utricularia nelumbifolia*, with remarks on embryo evolution in genus *Utricularia*. *Protoplasma* **239**, 69-80, (2010).

89 Liscum, E. & Reed, J. W. Genetics of Aux/IAA and ARF action in plant growth and development. *Plant Molecular Biology* **49**, 387-400 (2002).

90 Sessions, R. A. & Zambryski, P. C. Arabidopsis gynoecium structure in the wild and in ettin mutants. *Development* **121**, 1519-1532 (1995).

91 Jones, B. *et al.* Down-regulation of DR12, an auxin-response-factor homolog, in the tomato results in a pleiotropic phenotype including dark green and blotchy ripening fruit. *The Plant Journal : for cell and molecular biology* **32**, 603-613 (2002).

92 Kim, D. H. *et al.* A phytochrome-associated protein phosphatase 2A modulates light signals in flowering time control in Arabidopsis. *The Plant Cell* **14**, 3043-3056 (2002).

93 Hamann, T., Mayer, U. & Jurgens, G. The auxin-insensitive bodenlos mutation affects primary root formation and apical-basal patterning in the Arabidopsis embryo. *Development* **126**, 1387-1395 (1999).

94 Rogg, L. E., Lasswell, J. & Bartel, B. A gain-of-function mutation in IAA28 suppresses lateral root development. *The Plant Cell* **13**, 465-480 (2001).

95 Di Laurenzio, L. *et al.* The SCARECROW gene regulates an asymmetric cell division that is essential for generating the radial organization of the Arabidopsis root. *Cell* **86**, 423-433 (1996).

96 Helariutta, Y. *et al.* The SHORT-ROOT gene controls radial patterning of the Arabidopsis root through radial signaling. *Cell* **101**, 555-567 (2000).

97 Stuurman, J., Jäggi, F. & Kuhlemeier, C. Shoot meristem maintenance is controlled by a GRAS-gene mediated signal from differentiating cells. *Genes & Development* **16**, 2213-2218, (2002).

98 Duarte, J. *et al.* Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* **10**, 61 (2010).

99 Chapman, B. A., Bowers, J. E., Feltus, F. A. & Paterson, A. H. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 2730-2735, (2006).

100 Lynch, M. *The Origin of Genome Architecture*. (Sinauer Associates, Inc. Publishers, 2007).

101 Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R. & Zhu, J. K. Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell* **123**, 1279-1291, (2005).

102 Carlile, M., Nalbant, P., Preston-Fayers, K., McHaffie, G. S. & Werner, A. Processing of naturally occurring sense/antisense transcripts of the vertebrate Slc34a gene into short RNAs. *Physiol Genomics* **34**, 95-100, (2008).

103 Zubko, E., Kunova, A. & Meyer, P. Sense and antisense transcripts of convergent gene pairs in *Arabidopsis thaliana* can share a common polyadenylation region. *PLoS One* **6**, e16769, (2011).

104 Jen, C. H., Michalopoulos, I., Westhead, D. R. & Meyer, P. Natural antisense transcripts with coding capacity in Arabidopsis may have a regulatory role that is not linked to double-stranded RNA degradation. *Genome Biology* **6**, R51, (2005).

105 Martinez-Hernandez, A., Lopez-Ochoa, L., Arguello-Astorga, G. & Herrera-Estrella, L. Functional properties and regulatory complexity of a minimal RBCS light-responsive unit activated by phytochrome, cryptochrome, and plastid signals. *Plant Physiology* **128**, 1223-1233, (2002).

106 López-Ochoa, L., Acevedo-Hernández, G., Martínez-Hernández, A., Argüello-Astorga, G. & Herrera-Estrella, L. Structural relationships between diverse cis-acting elements are critical for the functional properties of a rbcS minimal light regulatory unit. *Journal of Experimental Botany* **58**, 4397-4406, (2007).

107 Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research* **32**, W199-203, (2004).

108 Carlson, J. M., Chakravarty, A., DeZiel, C. E. & Gross, R. H. SCOPE: a web server for practical de novo motif discovery. *Nucleic Acids Research* **35**, W259-264, (2007).

109 Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I. & Rubin, E. M. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Research* **12**, 832-839, (2002).

110 Sanford, J. C. *et al.* An improved, helium-driven biolistic device. *Technique* **3**, 3-16 (1991).

111 Tomes, D. T., Ross, M. C. & Songstad, D. D. in *Plant Cell, Tissue and Organ Culture* (eds O.L. Gamborg & G.C. Phillips) 197-213 (Springer-Verlag, 1995).

112 Matzke, M. A. & Birchler, J. A. RNAi-mediated pathways in the nucleus. *Nature Reviews Genetics* **6**, 24-35, (2005).

113 Voinnet, O. Origin, biogenesis, and activity of plant microRNAs. *Cell* **136**, 669-687, (2009).

114 Chen, X. Small RNAs and Their Roles in Plant Development. *Annual Review of Cell and Developmental Biology* **25**, 21-44, (2009).

115 Matzke, M., Kanno, T., Daxinger, L., Huettel, B. & Matzke, A. J. RNA-mediated chromatin-based silencing in plants. *Current Opinon in Cell Biology* **21**, 367-376, (2009).

116 Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* **33**, D121-D124, (2005).

117 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25, (2009).

118 Markham, N. R. & Zuker, M. UNAFold: software for nucleic acid folding and hybridization. *Methods in Molecular Biology* **453**, 3-31, (2008).

119 Meyers, B. C. *et al.* Criteria for annotation of plant MicroRNAs. *The Plant Cell* **20**, 3186-3190, (2008).

120 Henikoff, S. & Matzke, M. A. Exploring and explaining epigenetic effects. *Trends in Genetics* **13**, 293-295, (1997).

121 Henikoff, S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nature Reviews Genetics* **9**, 15-26, (2008).

122 Shilatifard, A. Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. *Annual Review of Biochemtry* **75**, 243-269 (2006).

123 Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics* **9**, 465-476, doi:10.1038/nrg2341 (2008).

124 Bonasio, R., Tu, S. & Reinberg, D. Molecular signals of epigenetic states. *Science* **330**, 612-616, (2010).

125 Zaratiegui, M., Irvine, D. V. & Martienssen, R. A. Noncoding RNAs and gene silencing. *Cell* **128**, 763-776, (2007).

126 Chan, S. W., Henderson, I. R. & Jacobsen, S. E. Gardening the genome: DNA methylation in Arabidopsis thaliana. *Nature Reviews Genetics* **6**, 351-360, (2005).

127 Jeffares, D. C., Mourier, T. & Penny, D. The biology of intron gain and loss. *Trends in Genetics* **22**, 16-22, doi:10.1016/j.tig.2005.10.006 (2006).

128 Roy, S. W. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics* **7**, 211-221, (2006).

129 Llopart, A., Comeron, J. M., Brunet, F. G., Lachaise, D. & Long, M. Intron presence-absence polymorphism in Drosophila driven by positive Darwinian selection. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 8121-8126, (2002).

130 Fink, G. R. Pseudogenes in yeast? *Cell* **49**, 5-6 (1987).

131 Kawaguchi, M. *et al.* Intron-loss evolution of hatching enzyme genes in Teleostei. *BMC Evolutionary Biology* **10**, 260, (2010).

132 Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496 (2011).

133 Lynch, M. Evolution of the mutation rate. *Trends in Genetics* **26**, 345-352, (2010).

134 Nordborg, M. *et al.* The Pattern of Polymorphism in *Arabidopsis thaliana*. *PLoS Biology* **3**, e196, (2005).

135 Haubold, B., Pfaffelhuber, P. & Lynch, M. mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Molecular Ecology* **19 Suppl 1**, 277-284, (2010).

136 Lynch, M. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Molecular Biology and Evolution* **25**, 2409-2419, doi:10.1093/molbev/msn185 (2008).

137 Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal* **53**, 661-673, (2008).

138 Lyons, E., Pedersen, B., Kane, J. & Freeling, M. The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. *Tropical Plant Biology* **1**, 181-190, (2008).

139 Lyons, E., Freeling, M., Kustu, S. & Inwood, W. Using Genomic Sequencing for Classical Genetics in *E. coli* K12. *PLoS ONE* **6**, e16717, (2011).

140 Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).

141 Tang, H. & Lyons, E. Unleashing the genome of brassica rapa. *Frontiers in Plant Science* **3**, 172, (2012).

142 Schnable, J. C. & Lyons, E. Comparative genomics with maize and other grasses: from genes to genomes! *Maydica* **56**, 183-200 (2011).

143 The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-467, (2007).

144 The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641, (2012).

145 Thomas, B. C., Pedersen, B. & Freeling, M. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research* **16**, 934-946, (2006).

146 Sankoff, D., Zheng, C. & Wang, B. A model for biased fractionation after whole genome duplication. *BMC Genomics* **13**, S8 (2012).

147 Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643-3646, (2004).

148 Tang, H. *et al.* Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* **12**, 102 (2011).

149 Tang, H. *quota-alignment*, <https://github.com/tanghaibao/quota-alignment/tree/3c92561ff0b86119fc84b9838a2c6f7c23167b8>

150 Lyons, E. *SynFind Syntenic Depth Examples*, <http://genomevolution.org/wiki/index.php/SynFind_Syntenic_Depth_Examples>

151 Hu, T. T. *et al.* The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature genetics* **43**, 476-481, (2011).

152 International Brachypodium, I. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763-768, (2010).

153 Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**, 276-277 (2000).

154 Wyman, S. K., Jansen, R. K. & Boore, J. L. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**, 3252-3255, (2004).

155 Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research* **33**, W686-W689, (2005).

156 Lohse, M., Drechsel, O. & Bock, R. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* **52**, 267-274, (2007).

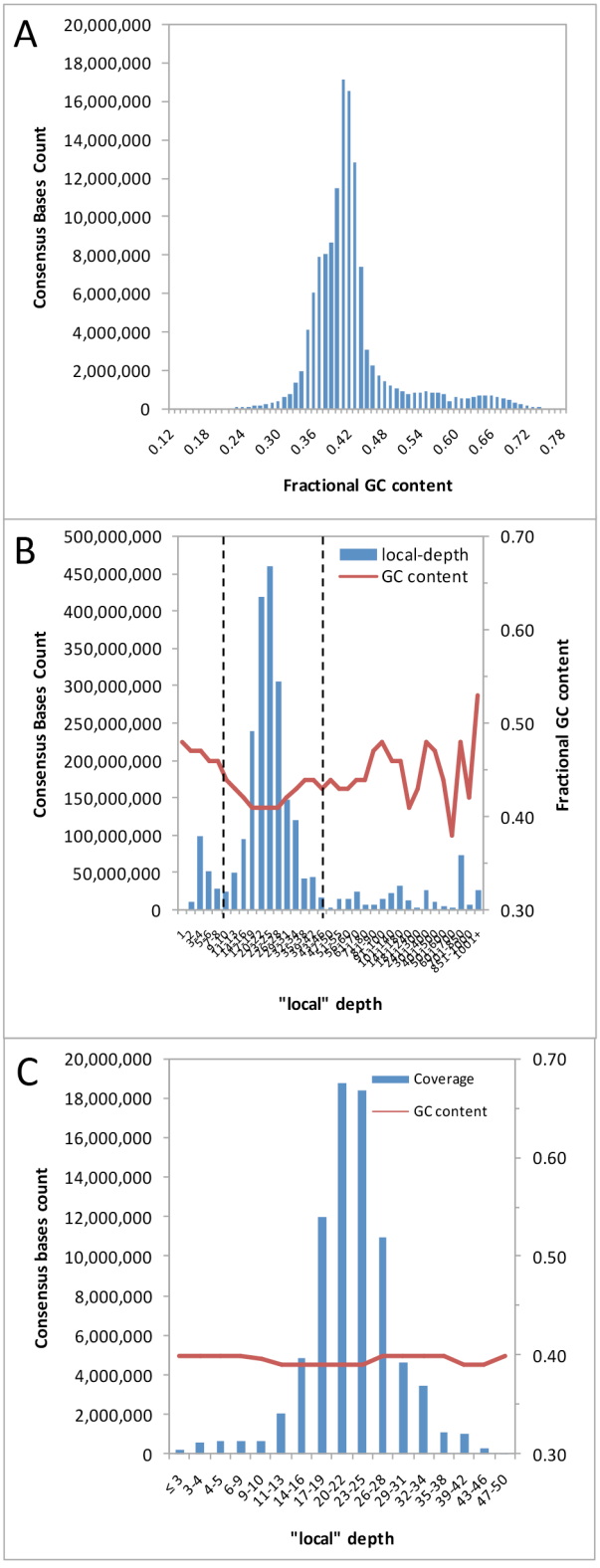
157 Xue, J. Y., Liu, Y., Li, L., Wang, B. & Qiu, Y. L. The complete mitochondrial genome sequence of the hornwort Phaeoceros laevis: retention of many ancient pseudogenes and conservative evolution of mitochondrial genomes in hornworts. *Current Genetics* **56**, 53-61, (2009).

158 Hecht, J., Grewe, F. & Knoop, V. Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. *Genome Biology and Evolution*, (2011).

159 Alverson, A. J. *et al.* Insights into the Evolution of Mitochondrial Genome Size from Complete Sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Molecular Biology and Evolution* **27**, 1436-1448, (2010).

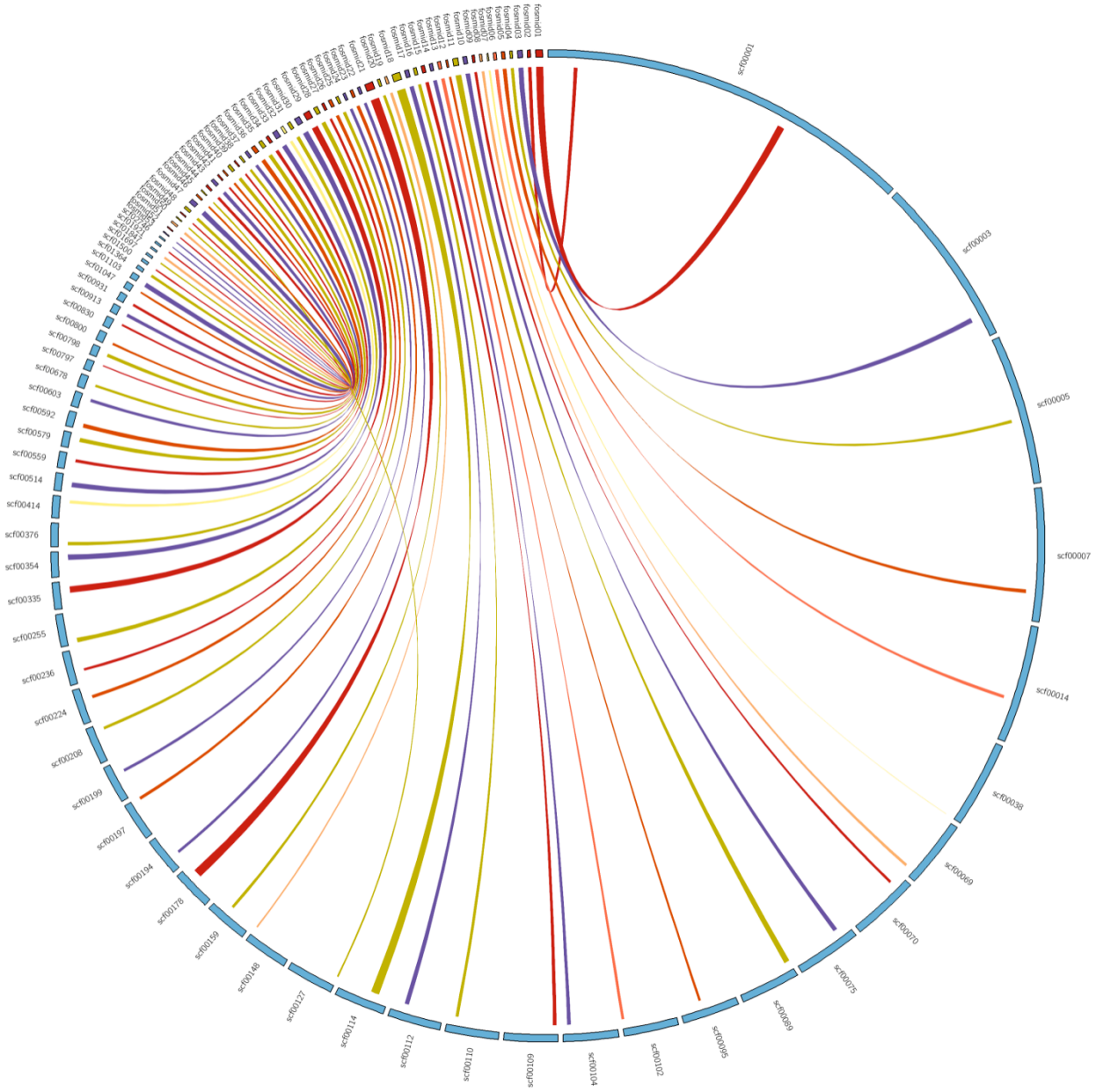
160 Alverson, A. J., Zhuo, S., Rice, D. W., Sloan, D. B. & Palmer, J. D. The Mitochondrial Genome of the Legume Vigna radiata and the Analysis of Recombination across Short Mitochondrial Repeats. *PLoS ONE* **6**, e16404, (2011).
161 Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* 25(1):119-120, (2009)
162 Wang, D. *et al.* Transfer of Chloroplast Genomic DNA to Mitochondrial Genome Occurred At Least 300 MYA. *Molecular Biology and Evolution* **24**, 2040-2048, (2007).
163 Katoh, K., Misawa, K., Kuma, K. i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066, (2002).
164 Posada, D. jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution* **25**, 1253-1256, (2008).
165 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**, 1969-1973, (2012).
166 Jobson, R. W., Nielsen, R., Laakkonen, L., Wikstrom, M. & Albert, V. A. Adaptive evolution of cytochrome c oxidase: Infrastructure for a carnivorous plant radiation. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 18064-18068, (2004).
167 Bell, C. D., Soltis, D. E. & Soltis, P. S. The age and diversification of the angiosperms re-revisited. *American Journal of Botany* **97**, 1296-1303, (2010).
168 Pond, S. L. K., Frost, S. D. W. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679, (2005).

12. Supplementary figures and legends

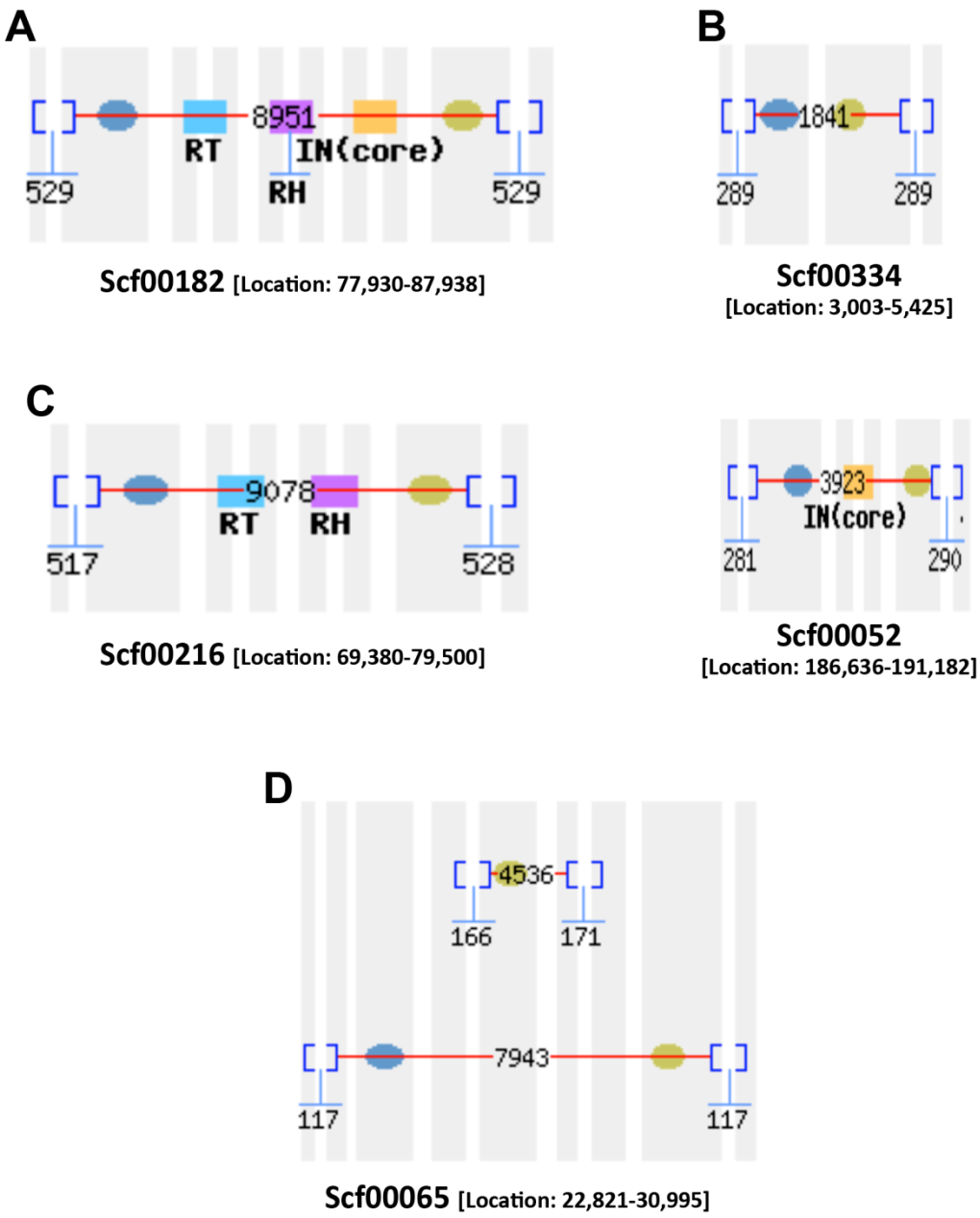


Supplementary figure 1: Filtering of contaminating environmental and organellar DNA from the *U. gibba* assembly. Plotting base counts of scaffolds and contigs against GC content reveals a basically unimodal GC distribution with a strong central mode and high vs. low GC tails (**A**). Base counts plotted against local depth of scaffolds and contigs similarly reveals a single major sequence depth mode (**B**). These results suggested the presence of a single major DNA sample (assumed to be the *U. gibba* nuclear genome) accompanied by contaminants from other sources. Low-coverage sequences to the left of the hatched box were identified as environmental DNA via BLAST (with plant sequences excluded) against the NCBI genome refseq database, and high-coverage sequences to the right of the box BLASTed as organellar DNA. The resulting filtered assembly (**C**) contained sequences showing BLAST hits to plant genomes in the refseq database.

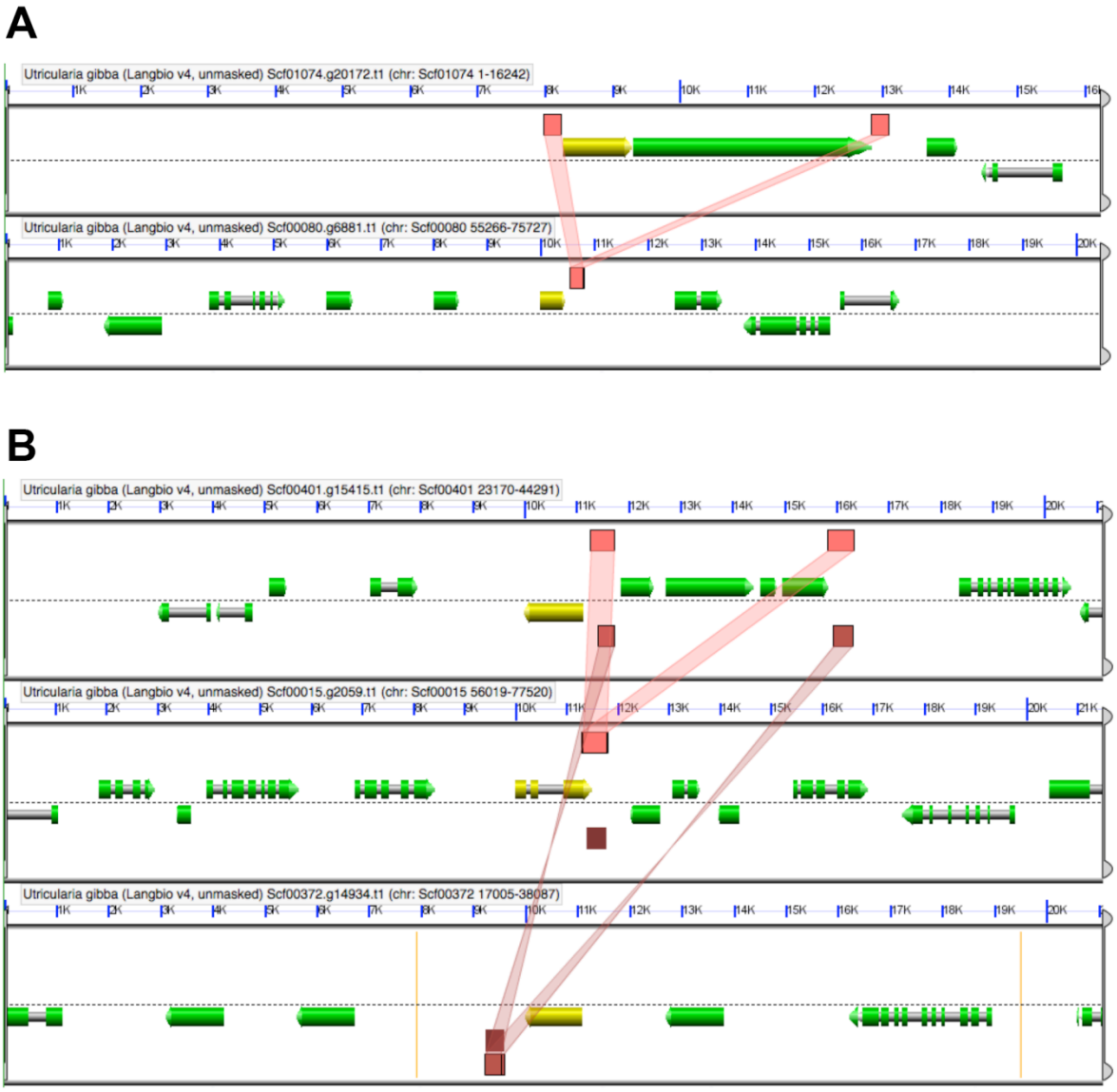
Supplementary Figure 2: Validation of the *U. gibba* genome assembly. Single pass primer walking of a 52,998 bp and 42,728 bp windows from Scaffold00089 (A) and Scaffold00021 (B), respectively. The windows represent 34.3 % and 15.5 % of total length of the sequence contained in the scaffolds.



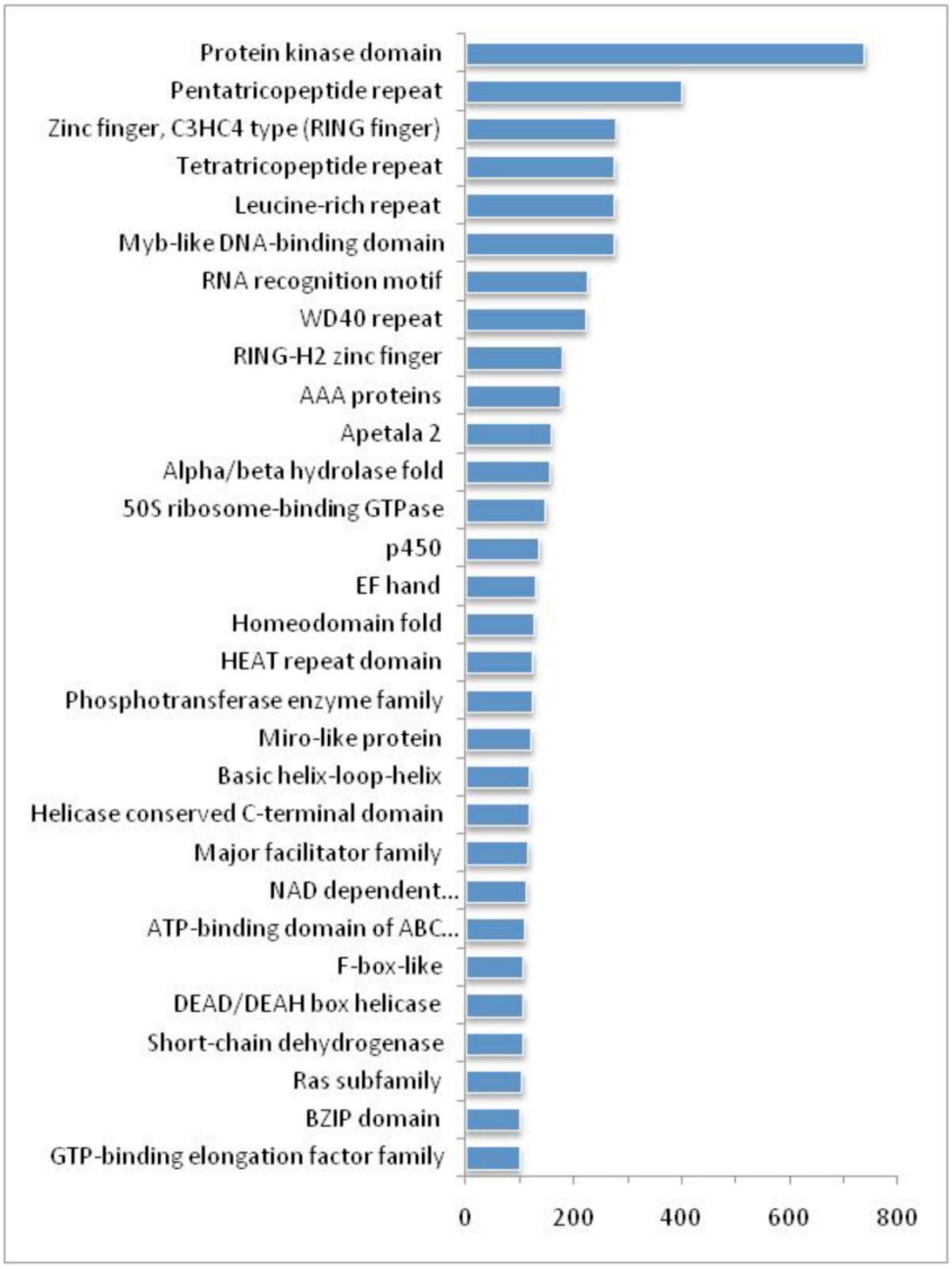
Supplementary Figure 3: Fosmids-scaffolds comparison. Circos map[19] depicting the alignment results. Each coloured segment represents a fosmid sequence while the scaffolds are coloured in blue. The fosmid positions were converted to approximate physical positions in *U. gibba* scaffolds. Coordinates were generated using Reference Mapper (Newbler) with default parameters.



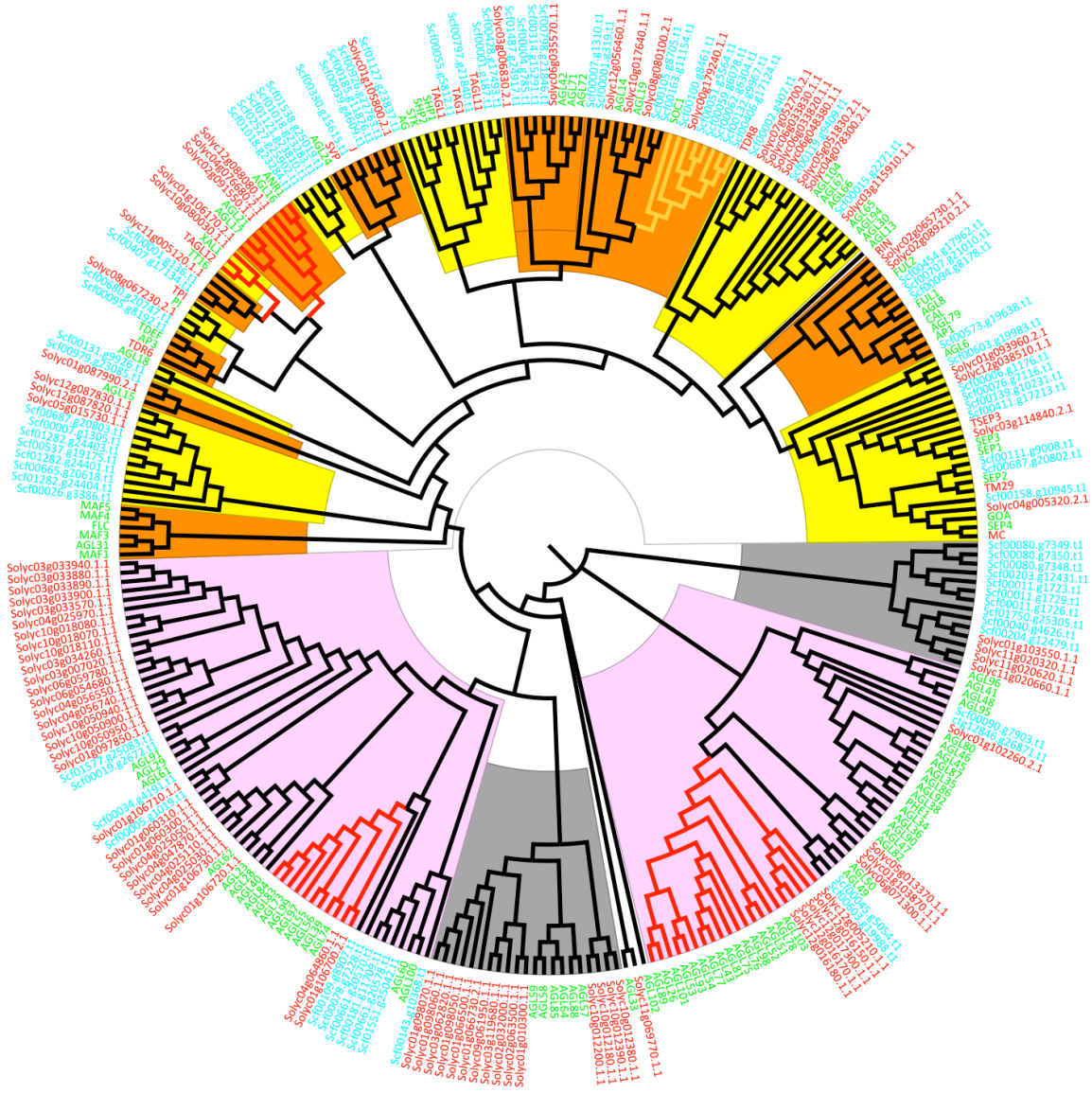
Supplementary Figure 4: *U. gibba* LTR retrotransposon structures. Structure of a complete element: long terminal repeats (LTRs; blue boxes), a primer-binding site (PBS; blue circules) and polypurine tract (PPT; green circules) needed for element replication, and encoded gag-pol gene products, the protein domains of which are labelled as IN (integrase; yellow), RT (reverse transcriptase; cyan) and RH (RNase H; purple) (**A**). Structure of incomplete LTR retrotransposons in which some elements are missed, such as complete gag-pol gene (**B**) or some protein domains (**C**). Additionally, evidence of ancient events of retrotransposition followed by DNA loss were also identified (**D**)



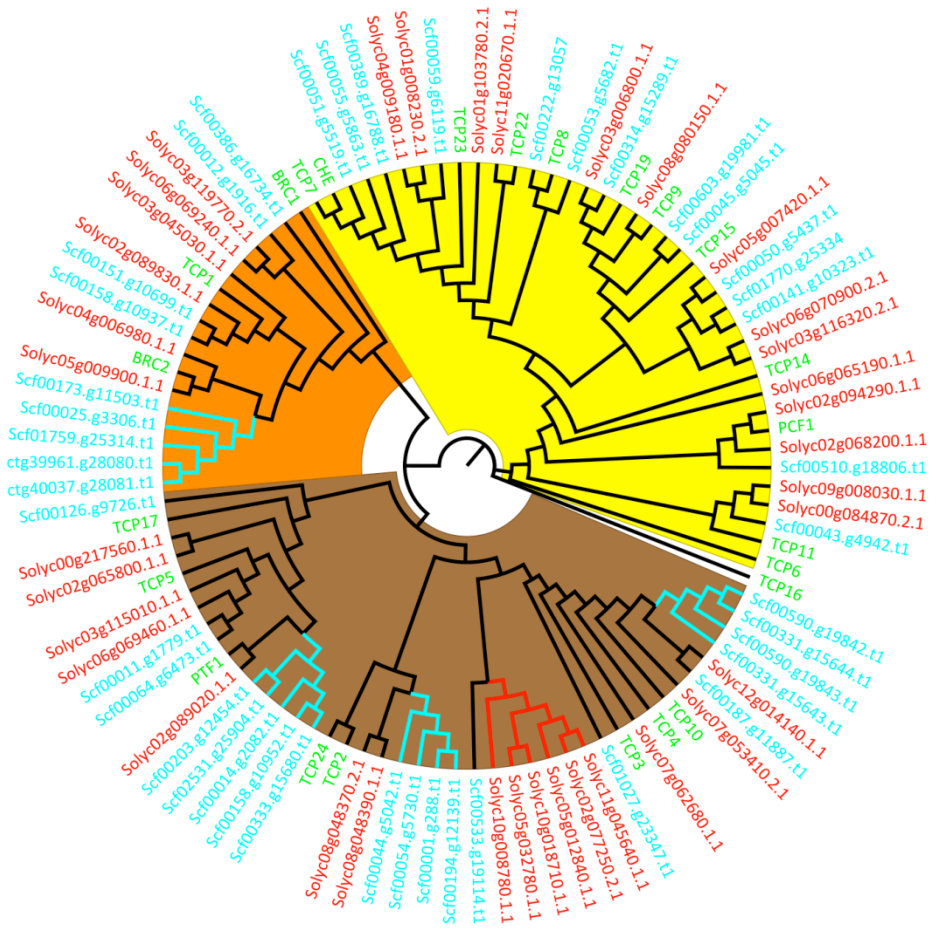
Supplementary Figure 5: Examples of "solo" LTRs identified in the *U. gibba* genome. Solo LTRs resulting from intra- or inter LTR element recombination[25]. These analyses may be regenerated at http://genomevolution.org/r/5f7l (**A**) and http://genomevolution.org/r/5f7z (**B**), respectively.



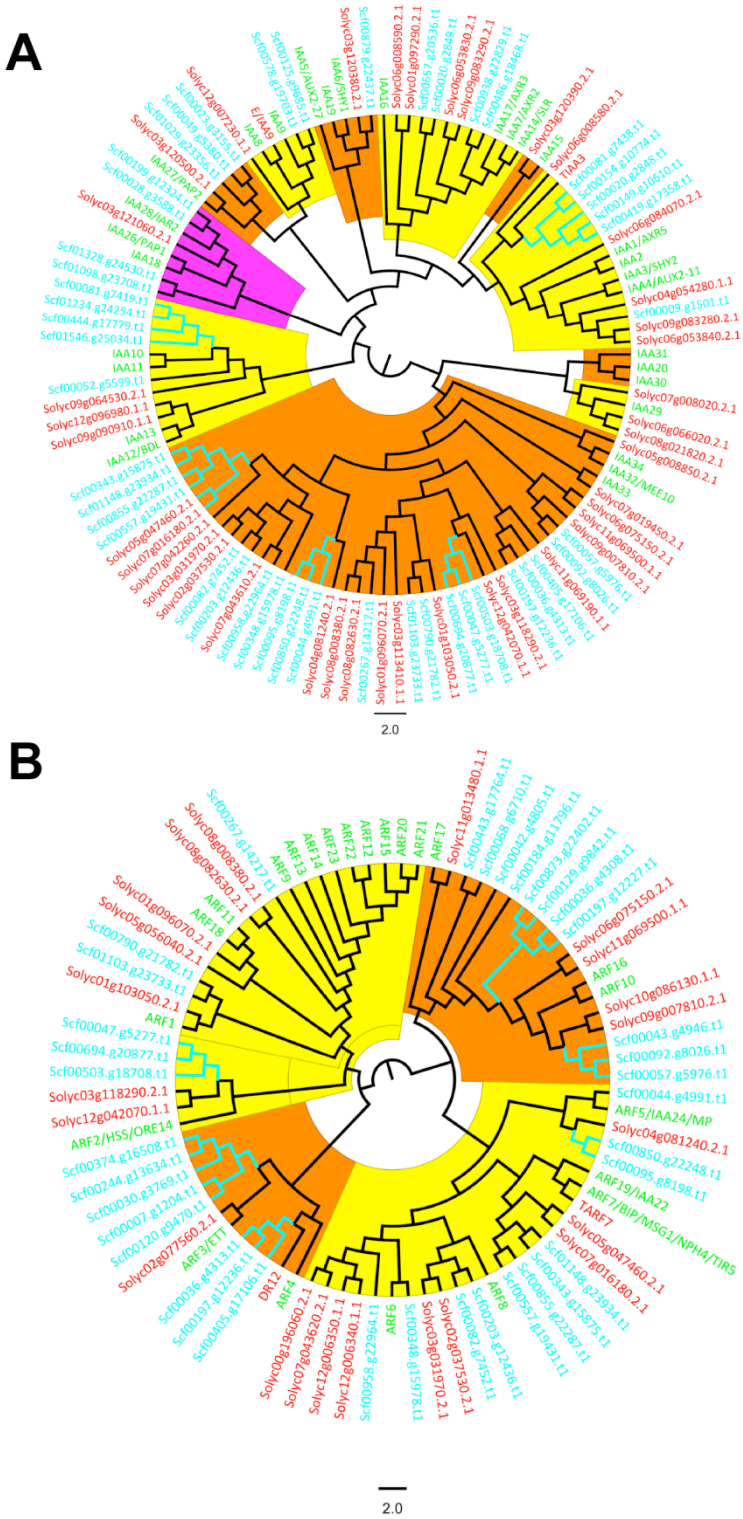
Supplementary Figure 6: The top 30 Pfam domains identified in *U. gibba* gene models.



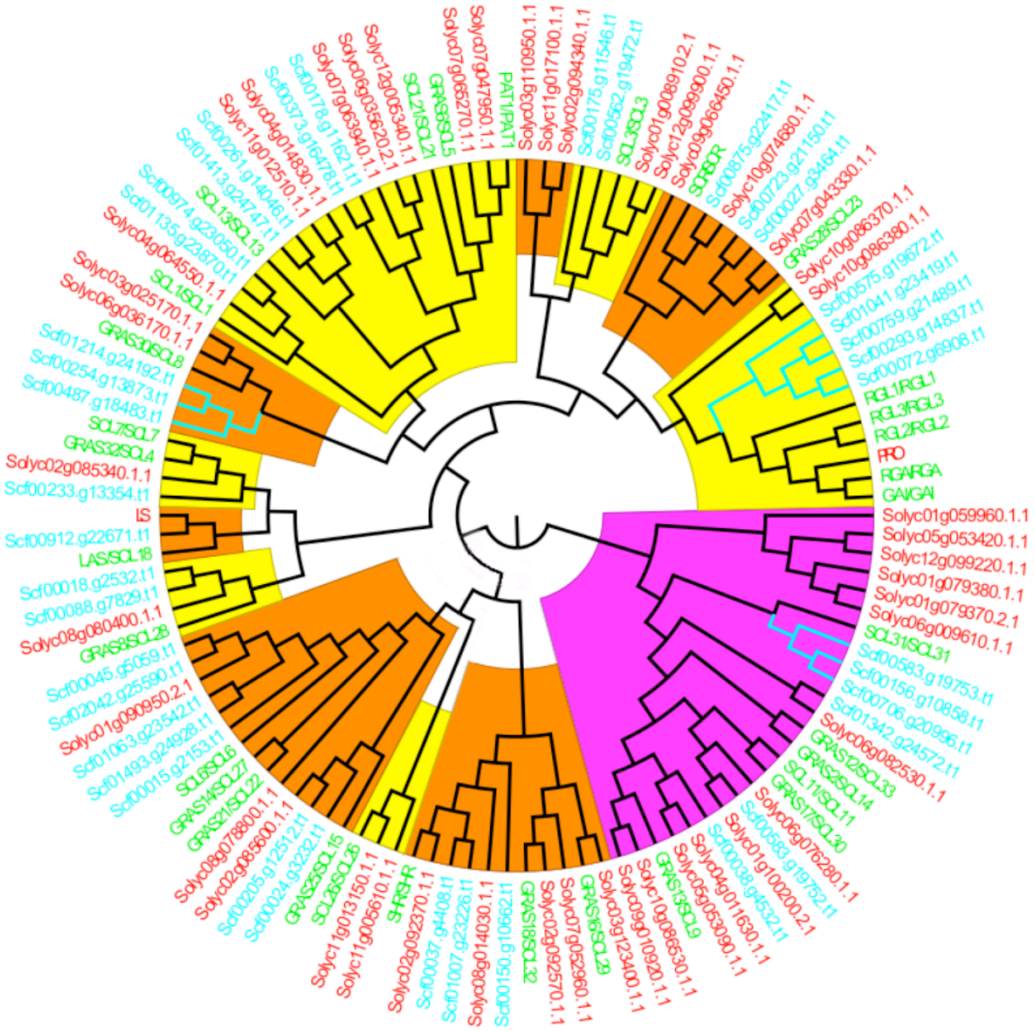
Supplementary Figure 7: Phylogenetic tree of the MADs-box transcription factor family. *U. gibba*, *S. lycopersicum* and *A. thaliana* gene names are in cyan, red and green, respectively. Clades representing specific subfamilies are colour-shadowed. Orange and yellow clades, alternating in colour for clarity, represent type II MADs-box genes, while type I MADs genes are shown alternating as pink and gray clades. Branch lines coloured in red indicate clades in which the *A. thaliana* genes are expressed in roots, while branch lines coloured in yellow represent homologs of *SOC1*, a gene expressed in shoots that is involved in a general response to nutrient stress. Gene subfamily classification and their members are listed in Supplementary Table 21.



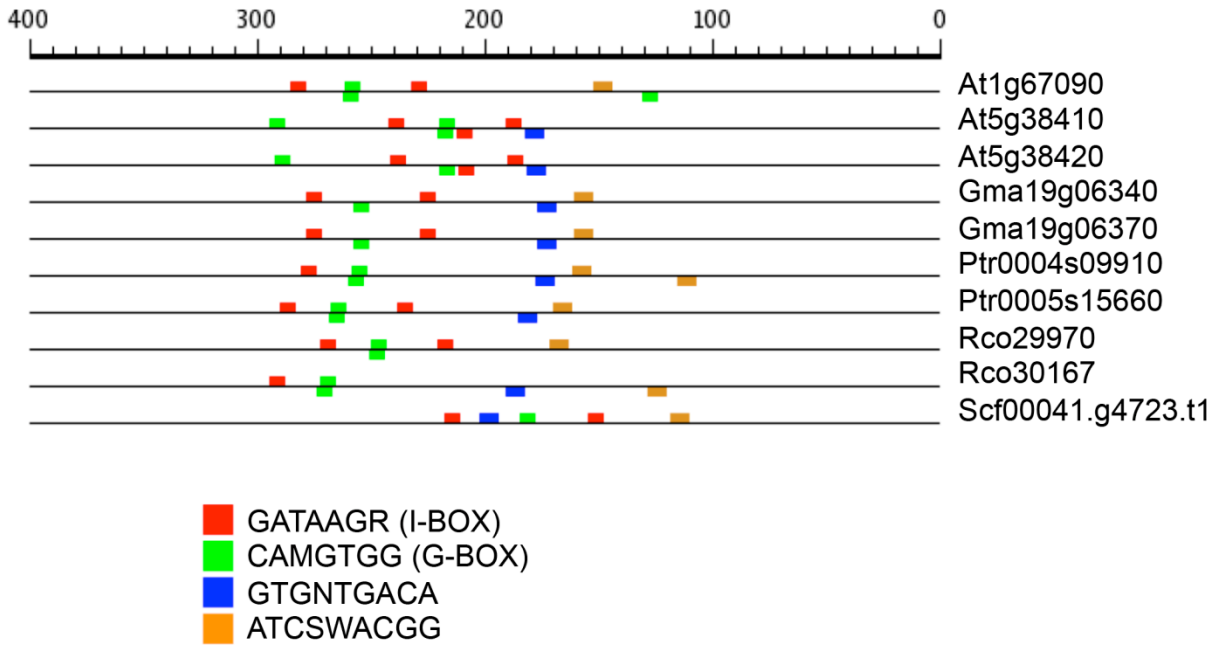
Supplementary Figure 8: Phylogenetic tree of the TCP transcription factor family. *U. gibba*, *S. lycopersicum* and *A. thaliana* gene names are in cyan, red and green, respectively. Clades representing specific subfamilies are shadowed in alternating colours for clarity. Specific clades show expanded families in *U. gibba* (cyan branches) or *S. lycopersicum* (red branches). Gene subfamily classification and their members are listed in Supplementary Table 21.



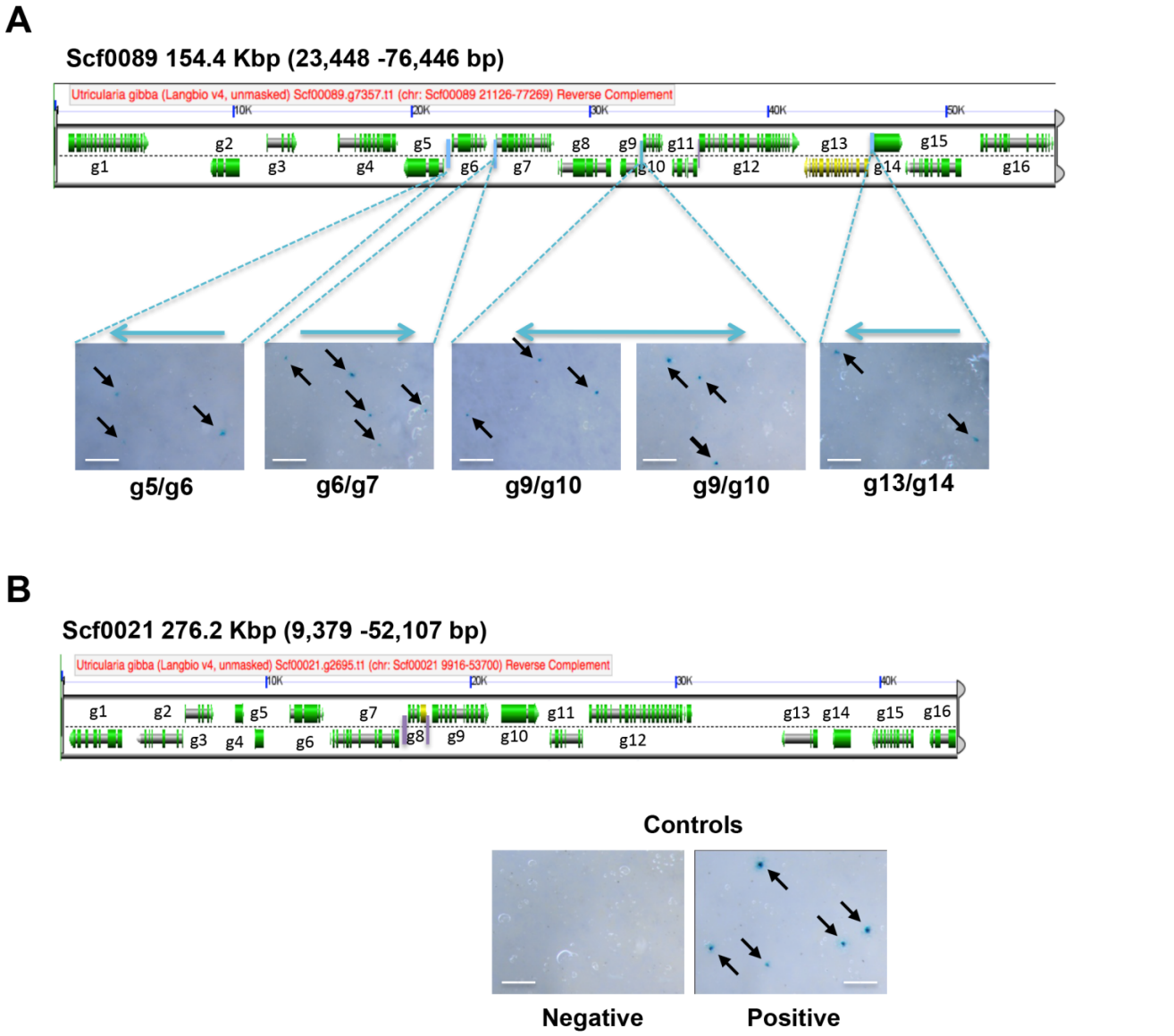
Supplementary Figure 9: Phylogenetic trees of AUX/IAA (**A**) and ARF (**B**) transcription factor families. *U. gibba*, *S. lycopersicum* and *A. thaliana* gene names are in cyan, red and green colour, respectively. Clades representing specific subfamilies are shadowed in alternating colours for clarity. Specific clades show expanded families in *U. gibba* (cyan branches). Gene subfamily classification and their members are listed in Supplementary Table 21.



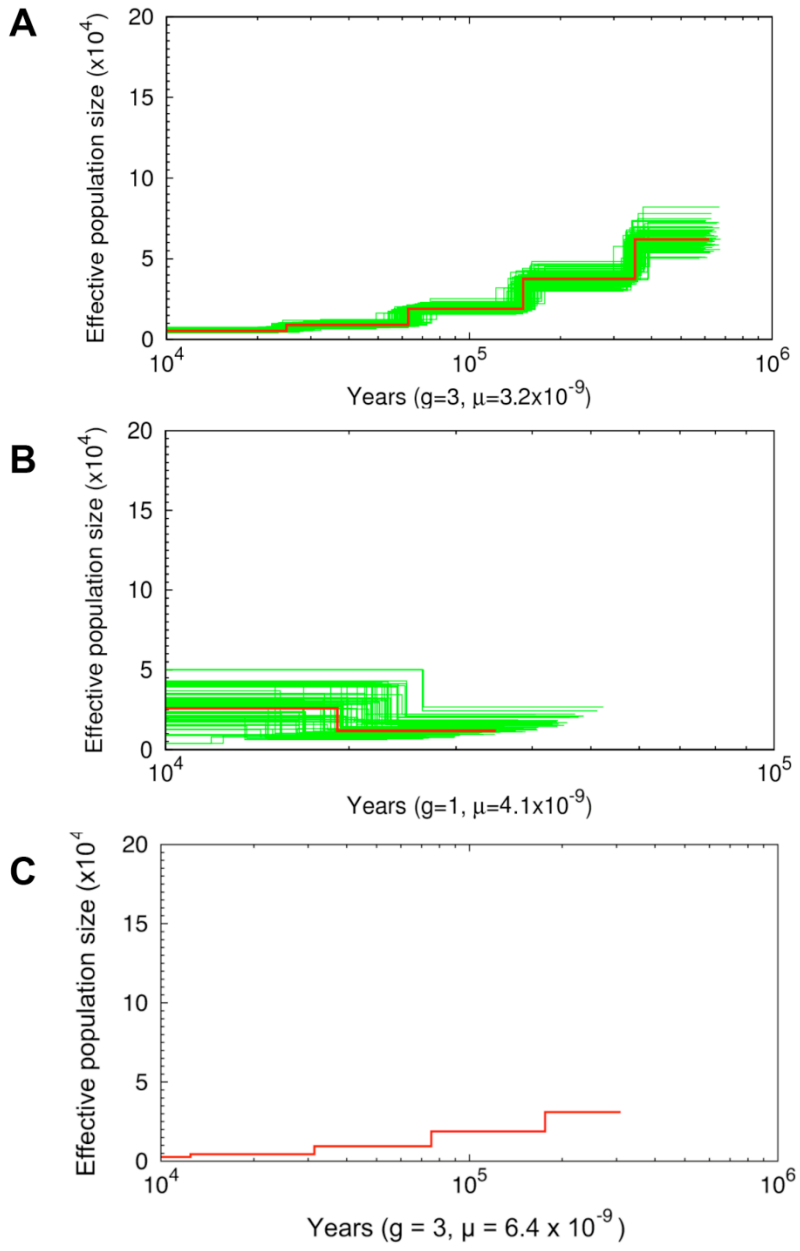
Supplementary Figure 10: Phylogenetic tree of the GRAS transcription factor family. *U. gibba*, *S. lycopersicum* and *A. thaliana* gene names are in cyan, red and green, respectively. Specific clades show expanded families in *U. gibba* (cyan branches). Gene subfamily classification and their members are listed in Supplementary Table 21.



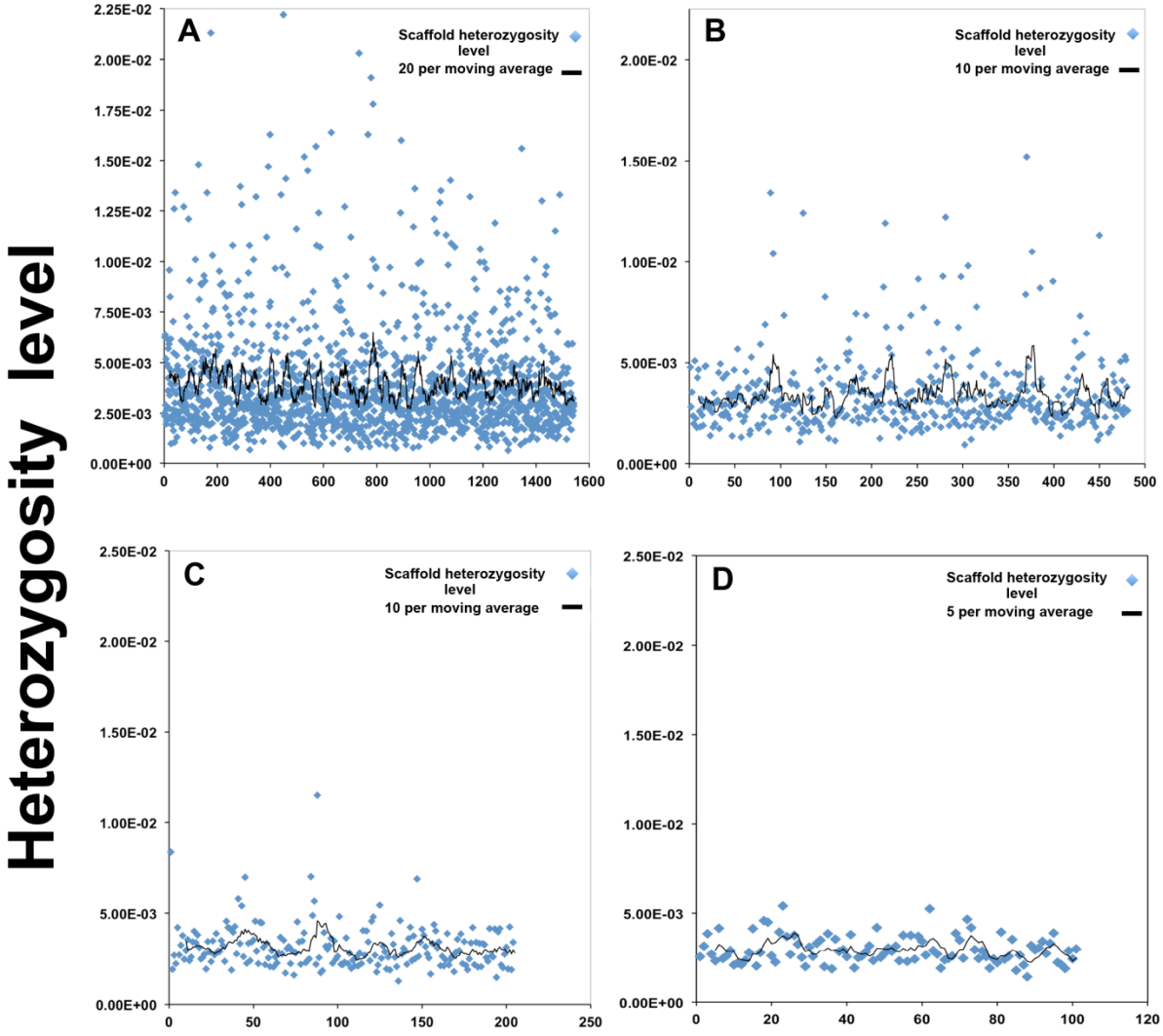
Supplementary Figure 11. Distribution of regulatory elements in the promoter region of *rbcS* genes of different plant species (*A. thaliana, Glycine max, Populus trichocarpa, Ricinus communis* and *U. gibba*).



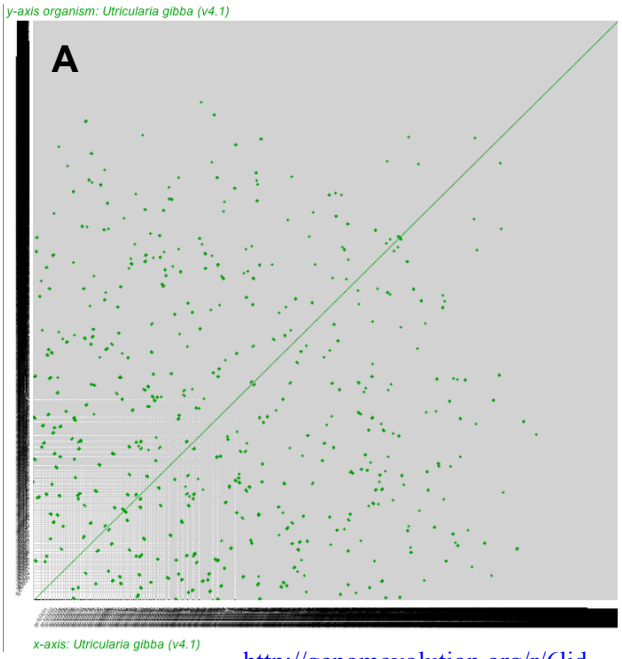
Supplementary Figure 12. GUS transient expression driven by *U. gibba* promoters. Intergenic regions between some gene pairs derived from two genomic scaffolds, previously validated by primer walking, were selected (scaffold00089 (**A**) and scaffold0021 (**B**) respectively). Promoters from *U. gibba* are labeled as g5/g6 (557 bp), g6/g7 (612 bp), g9/g10 (397 bp), g11/g12 (196 bp) and g13/g14 (441 bp), and on the other scaffold, g8/g9 (555 bp) and g9/g10 (397 bp). With only three exceptions (promoters represented as purple lines in the figure), GUS expression was detected in the promoters tested. These regions may be regenerated by CoGe at http://genomevolution.org/r/5l04 and http://genomevolution.org/r/5l0j, respectively.



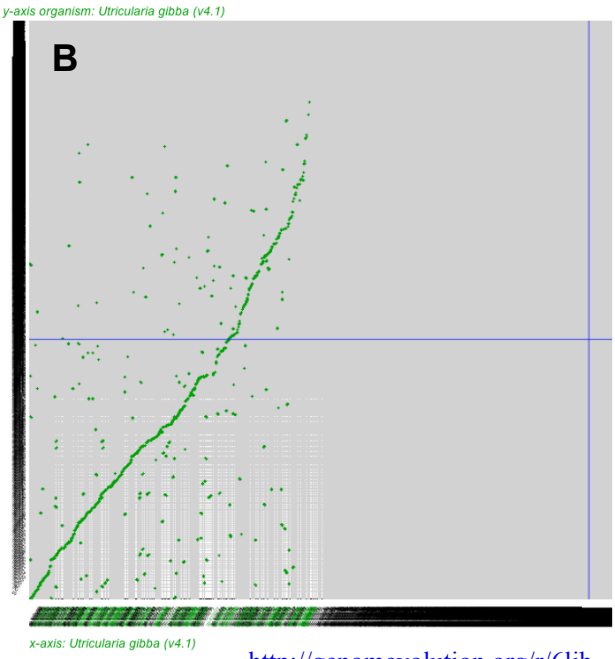
Supplementary Figure 13. Inferred change in effective population size (N_e) over time using the Pairwise Sequentially Markovian Coalescent (PSMC) model. To scale PSMC results to real time, we assumed 3 years per generation and a per-generation mutation rate (μ) of 3.2×10−9 for *U. gibba* (**A**) and 1 year per generation and a μ of 4.1 × 10−9 per generation for *Arabidopsis thaliana* (**B**). Changes in *U. gibba* N_e and coalescence time when using an arbitrary rate value increased by 2x (μ= 6.4 x 10-9) did not change the overall shape of the curve and still resulted in small N_e (**C**). See Supplementary information Section 6.1. for further discussion of A-C.



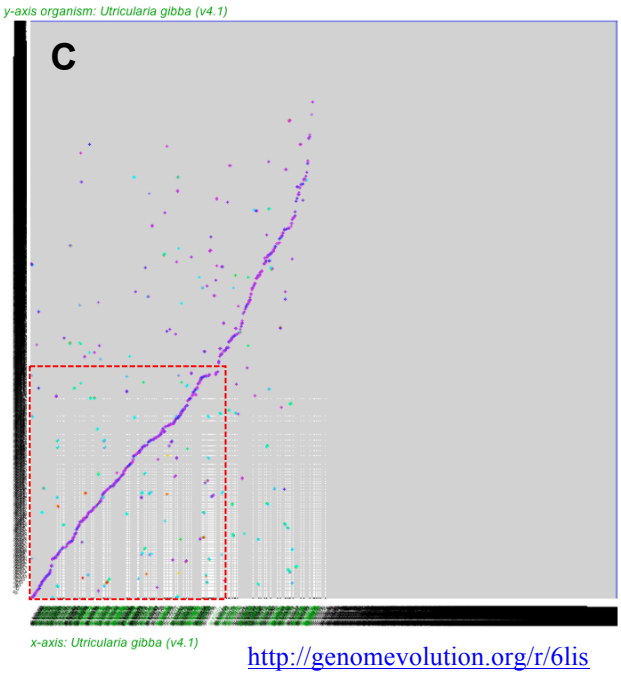
Supplementary Figure 14. Distribution of heterozygosity along genomic segments of *U. gibba*. Heterozygosity was estimated (as $\theta = 4Neu$, y-axis) in non-overlapping windows for intervals of 25 (A), 50 (B), 75(C) and 100 (D) Kb. Expected heterozygosity (He) is closely correlated with θ when θ is small ($<< 1$), as here, since $He = \theta/(1 + \theta) \approx \theta$. The x-axis shows total scaffolds. Moving averages are shown.



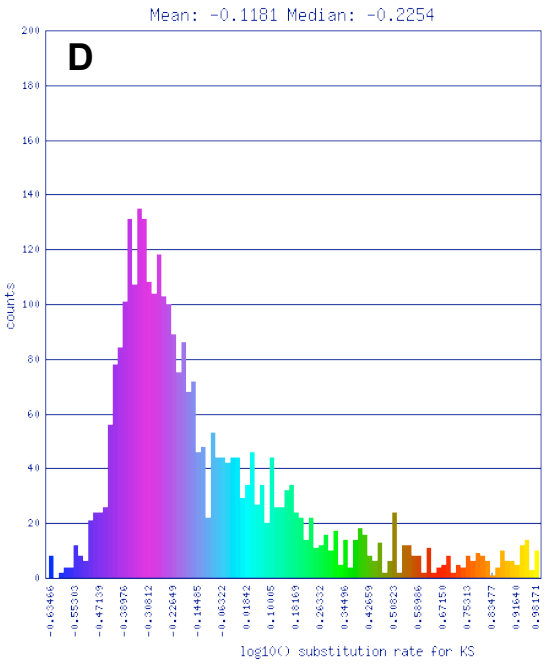
http://genomevolution.org/r/6lid

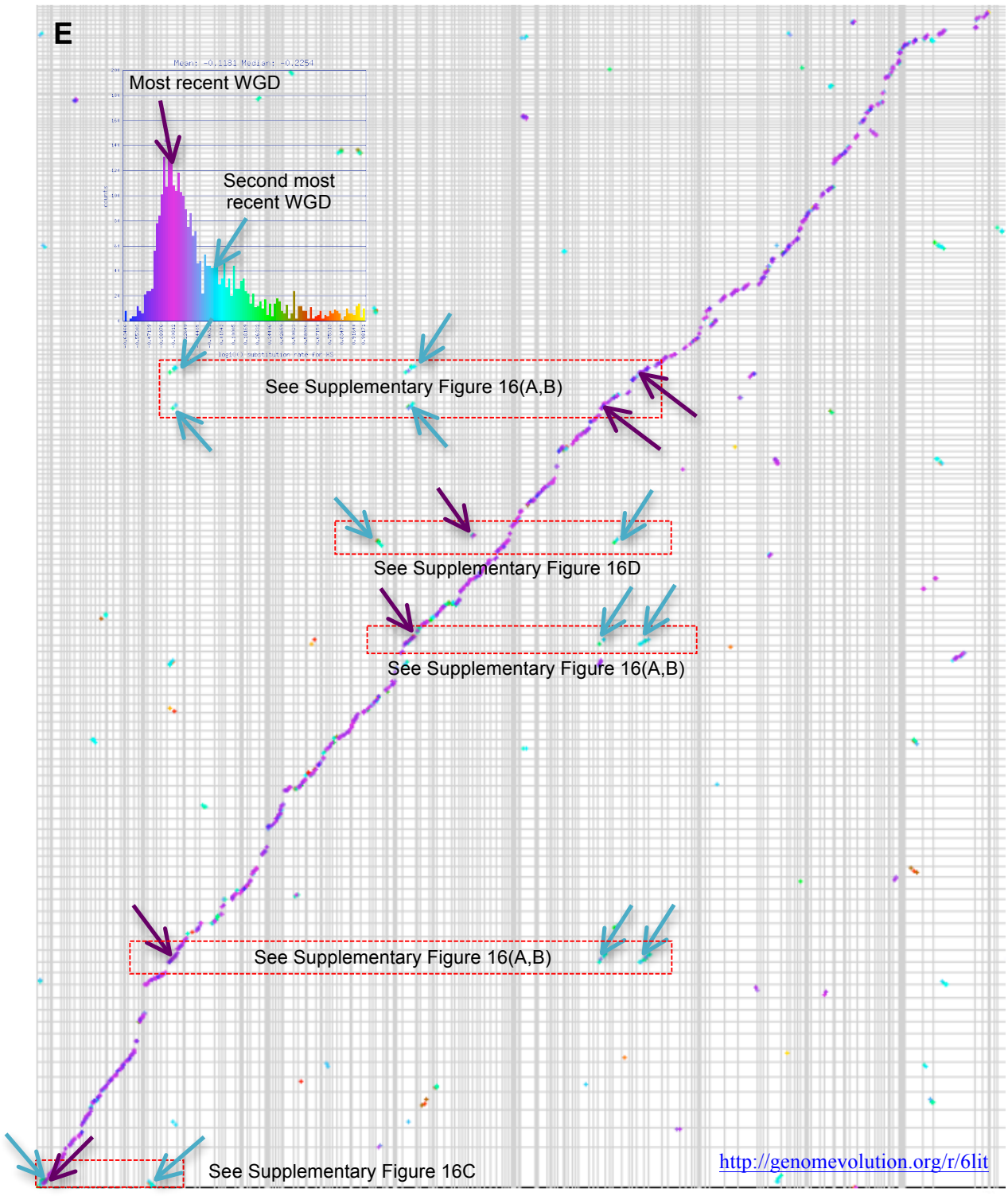


http://genomevolution.org/r/6lib



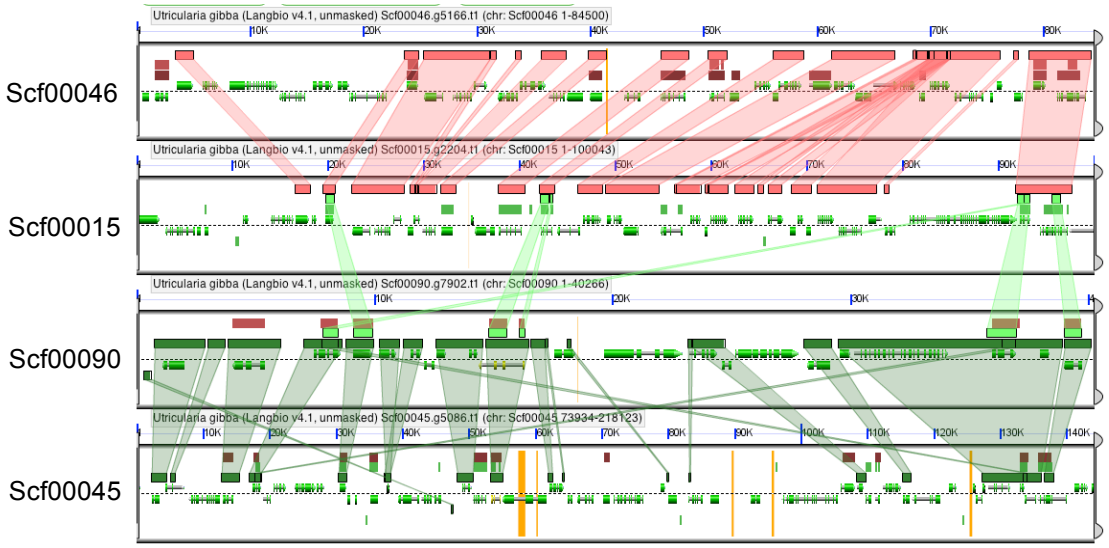
http://genomevolution.org/r/6lis





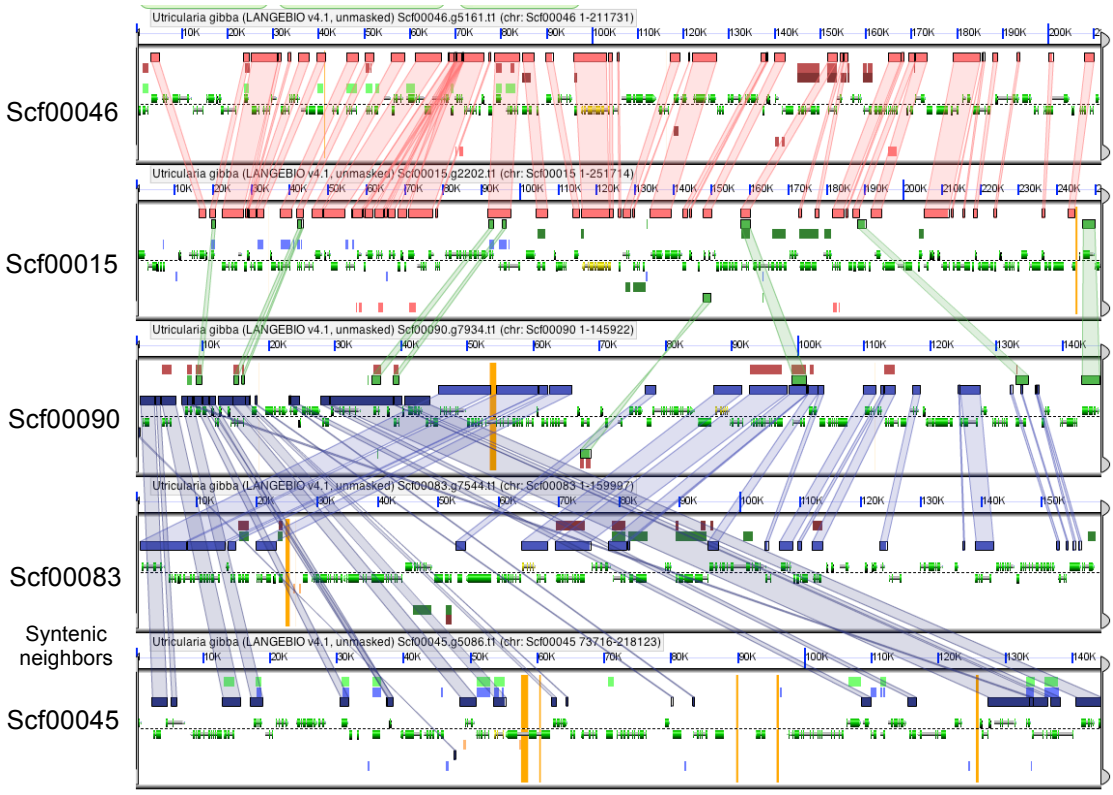
Supplementary Figure 15. Self-self syntenic dotplots of *U. gibba* showing evidence that *U. gibba* has had at least two WGDs. **(A)** Self-self dotplot; contigs ordered by size along each axis. **(B)** Self-self dotplot; contigs ordered by syntenic path assembly along x-axis. **(C)** Self-self dotplot; syntenic gene pair dots are coloured by synonymous mutation values. Red rectangle is shown zoomed in **(E)**. **(D)** Histogram of synonymous mutation values with colour scheme used for **(C)**. Note that values are $\log_{10}$ transformed and small values (younger syntenic gene pairs) are on the left of the histogram. **(E)** Zoomed-in portion of the syntenic dotplot. Purple lines are made up of genes derived from the most recent WGD, cyan from the second most recent WGD. Evidence for WGD is shown by coloured arrows and red boxes. Note that there is one purple and two cyan regions as would be expect from two sequential WGDs.

A

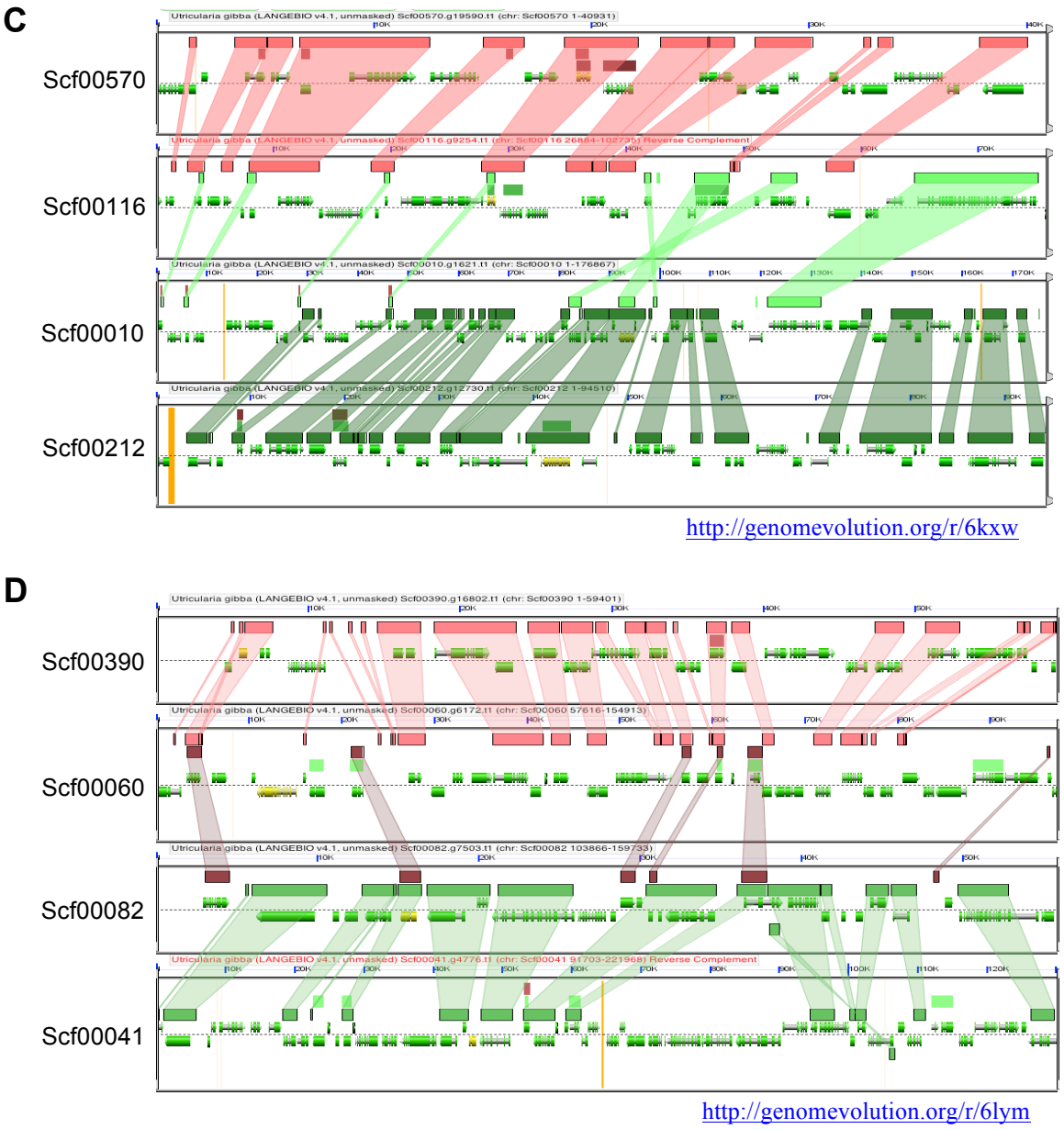


http://genomevolution.org/r/6g76

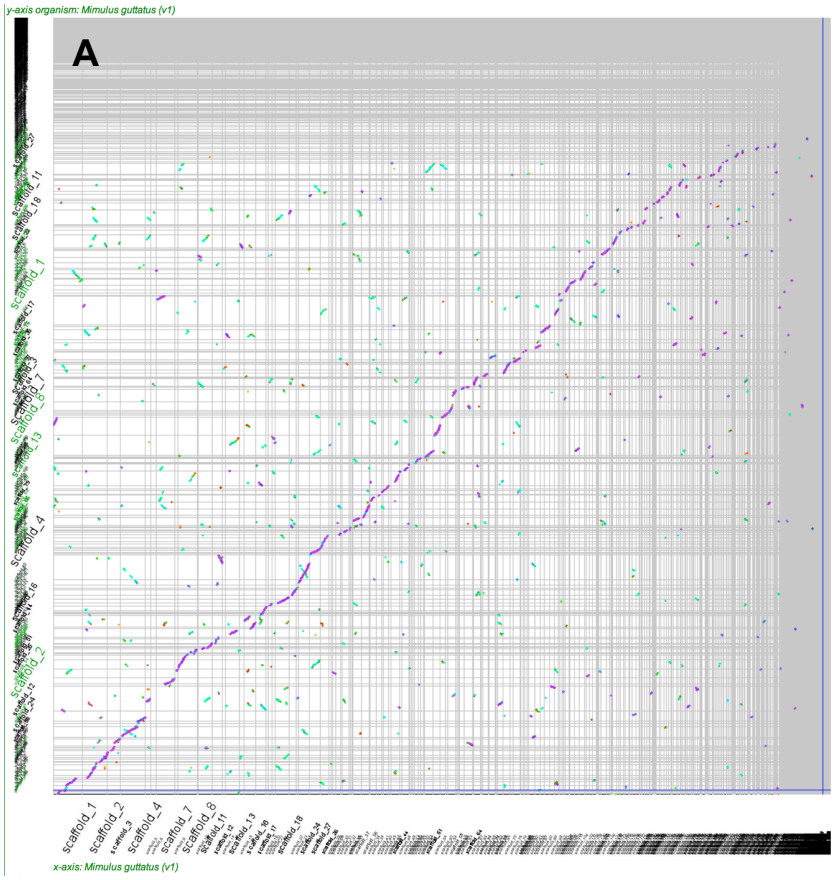
B

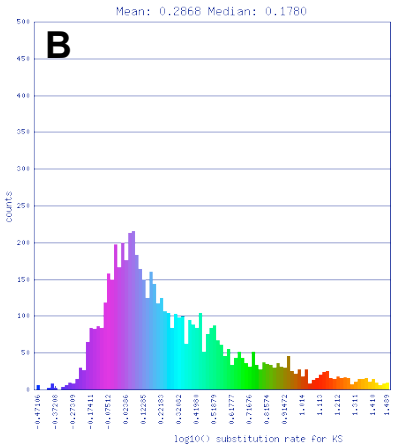


http://genomevolution.org/r/6kw5

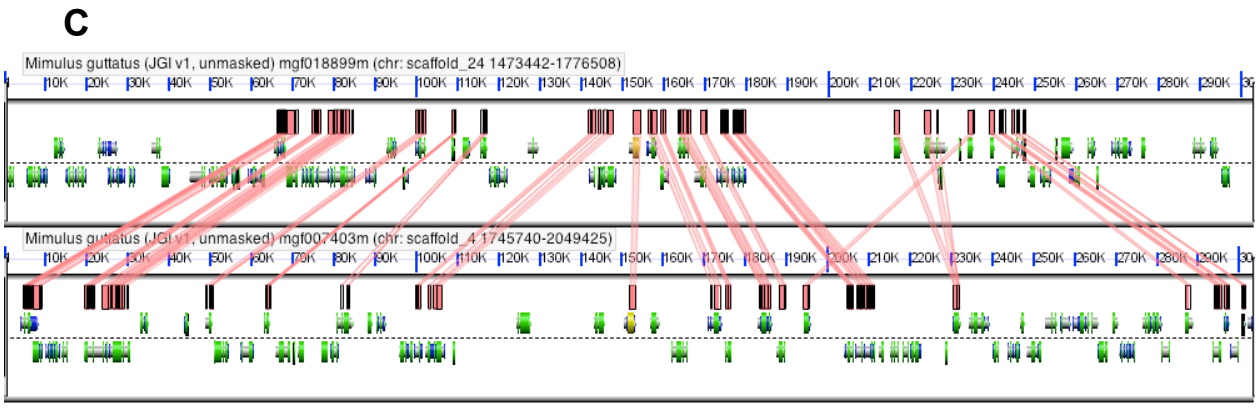


Supplementary Figure 16. Microsynteny analysis of set of purple and cyan syntenic regions identified in Supplementary Figure 15. Note that for each of these sets, the syntenic regions make up two pairs of regions (though A has two contigs making one half of one pair) with a relatively high degree of synteny. Each of these pairs is from the most recent WGD event in *U. gibba*. Synteny is also observed between pairs of regions, evidence that they are related from an older WGD event. **(A)** Two pairs of syntenic regions showing synteny across the pairs. **(B)** Same set of contigs shown in **(A)** with the addition of a fifth contig with synteny to Scf00090. Due to the large number of contigs assembled for this genome, it is a common occurence to have to use multiple contigs to see synteny. **(C)** Two pairs of syntenic regions showing synteny across the pairs. **(D)** Two pairs of syntenic regions showing synteny across the pairs.



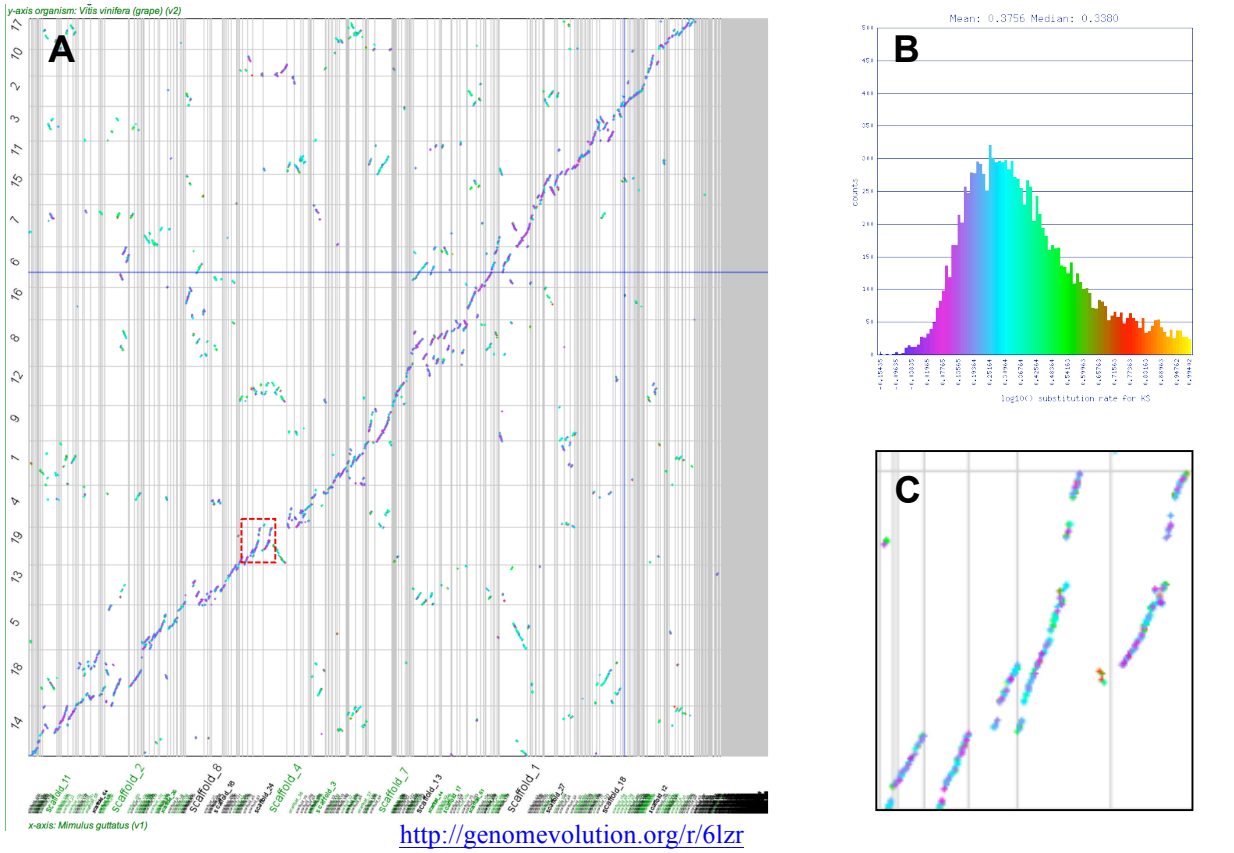


http://genomevolution.org/r/6eqq

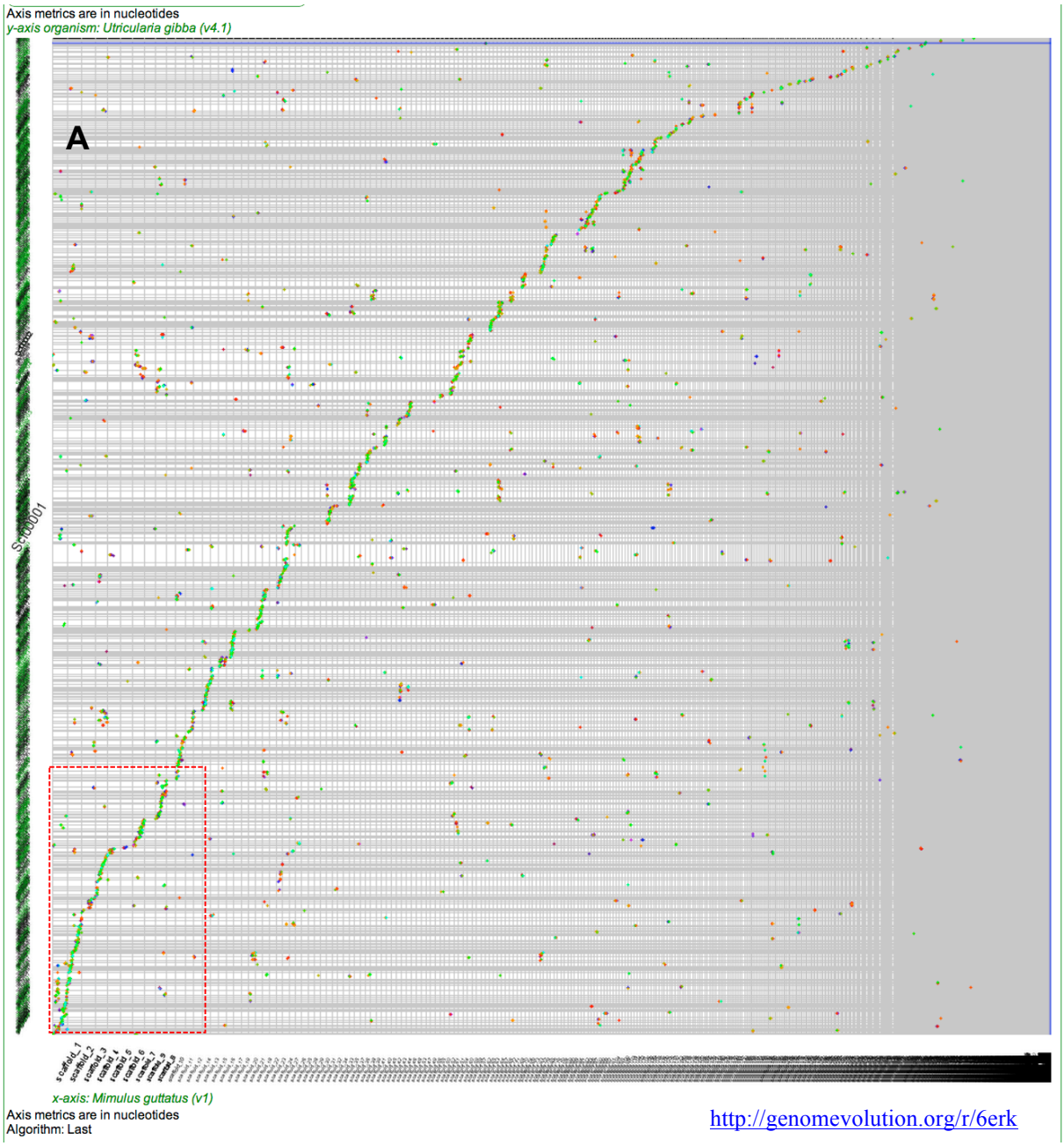


http://genomevolution.org/r/6lzp

Supplementary Figure 17. Synteny analysis of *Mimulus guttatus* against itself showing a recent WGD event. **(A)** Self-self syntenic dotplot where contigs on the y-axis have been ordered and oriented according to their syntenic path. Syntenic gene pair dots have been coloured by the Ks values. Note the two age distributions of syntenic regions. Purple are younger than cyan. **(B)** Histogram and colour scheme of Ks values derived from syntenic gene pairs in *M. guttatus*. **(C)** Microsynteny analysis of younger syntenic regions.

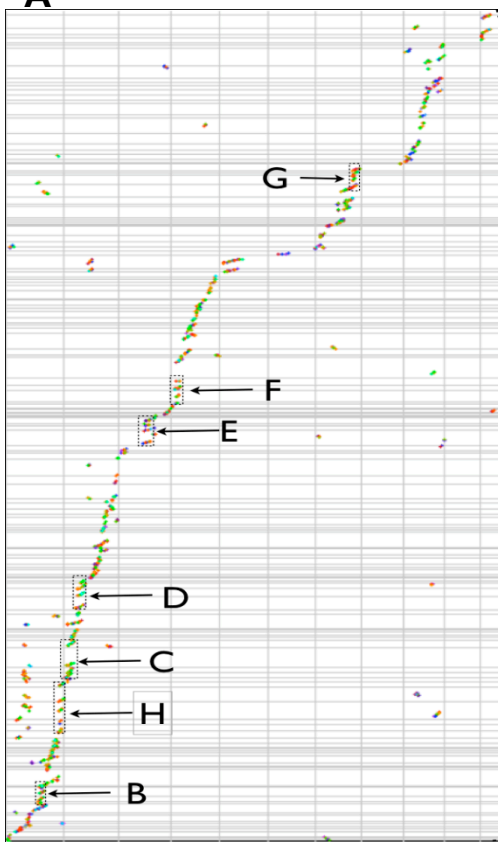


Supplementary Figure 18. Synteny analysis of *M. guttatus* (Mg) versus *V. vinifera* (Vv) shows evidence that Mg has had an independent WGD **(A)**. Syntenic dotplot where with Mg on the x-axis and Vv on the y-axis. Mg contigs have been ordered and oriented according to their syntenic path. Syntenic gene pair dots are coloured by the Ks values. Note the two age distributions of syntenic regions. Purple are younger than cyan. **(B)** Histogram and colour scheme of Ks values derived from syntenic gene pairs. **(C)** Close up of a region of the syntenic dotplot shown in **(A)**. Note that pairs of Mg syntenic regions to each grape region. This pattern is repeated across the entire genome, as is evidence that Mg has had a WGD. **(D)** Microsynteny analysis of one grape region to two Mg regions syntenic regions. Note that Vv has the entire gene content of the two Mg regions; the gene content of the Mg regions is fractionated.

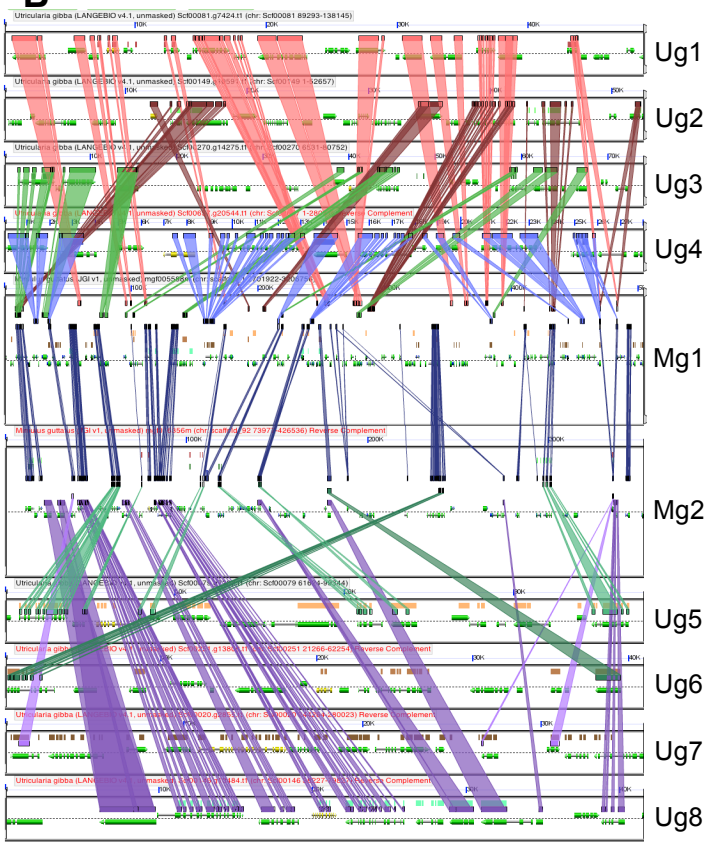


Supplementary Figure 19. (A) Syntenic dotplot of *Mimulus guttatus* (Mg; x-axis) versus *U. gibba* (Ug; y-axis) with Ug's contigs ordered and oriented by SPA. Syntenic gene pair dots are coloured by Ks values. **(B)** Histogram of Ks values derived from syntenic gene pairs. Note that red-dashed rectangle. This region of the dotplot is shown in higher resolution in Supplementary Figure 20 and shows evidence that Ug is 4x polyploid in relations to Mg.

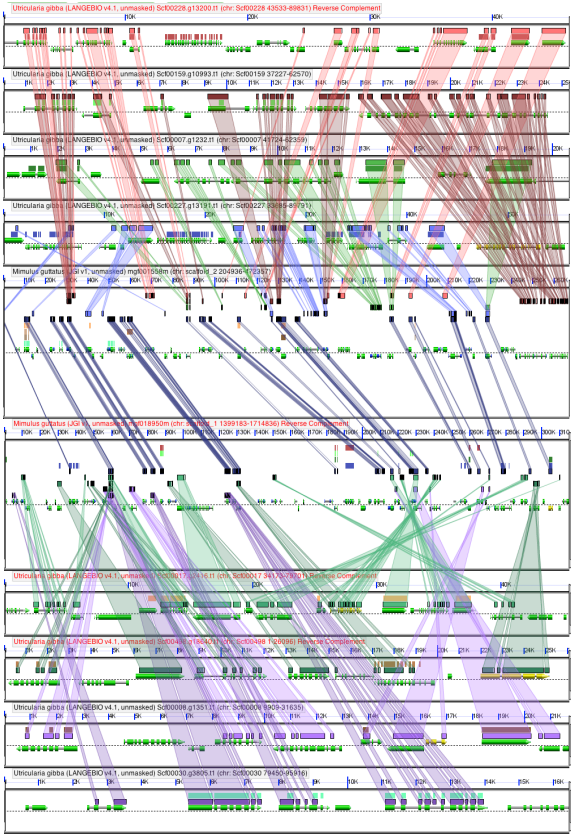
doi:10.1038/nature12132



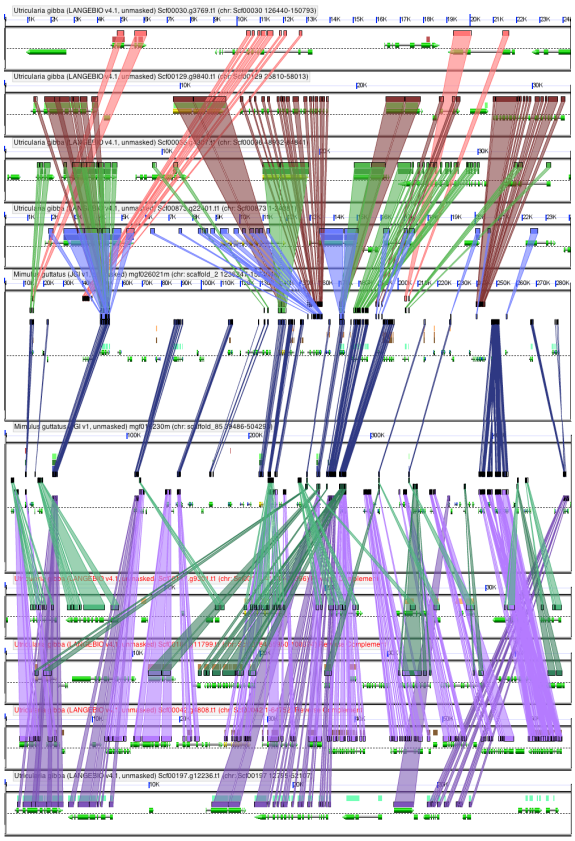
http://genomevolution.org/r/6l0z



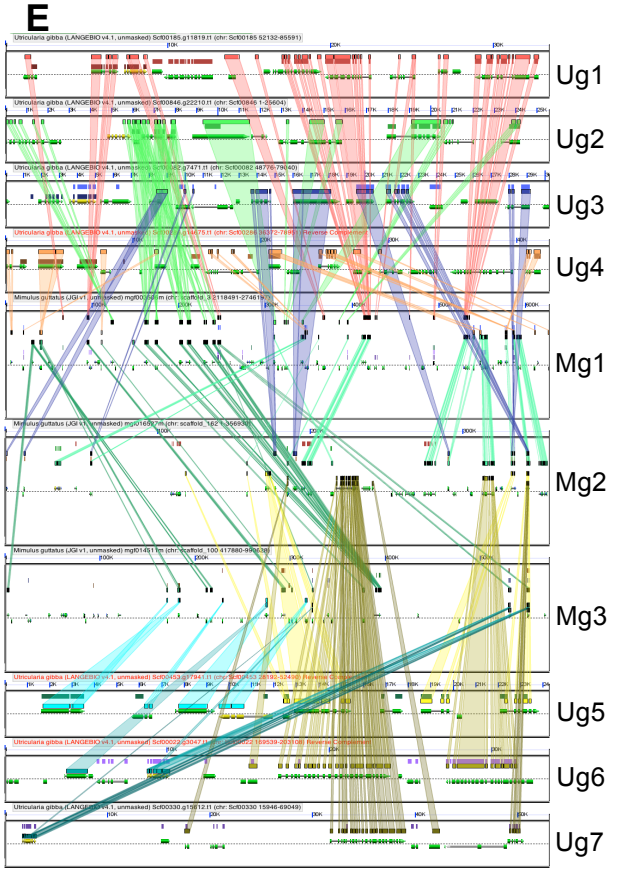
http://genomevolution.org/r/6l1r



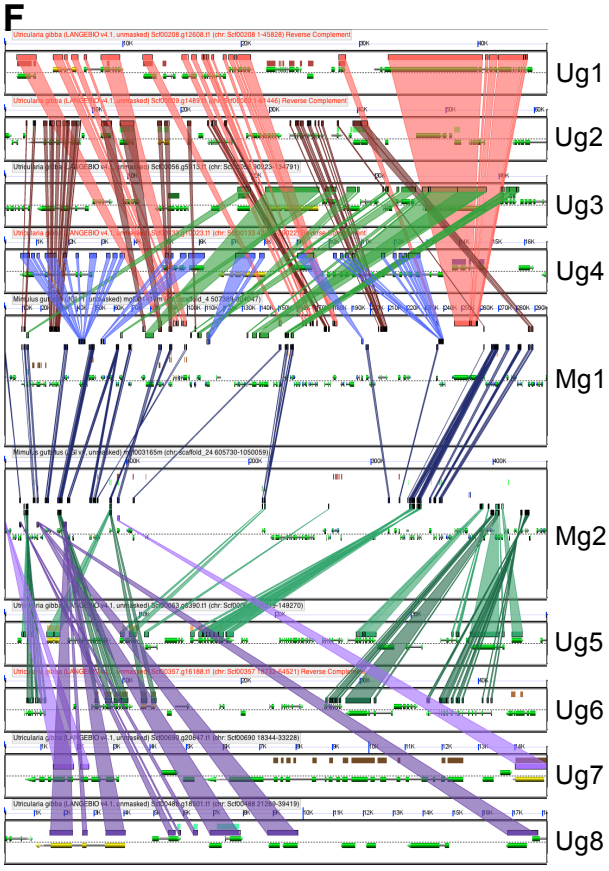
http://genomevolution.org/r/6m07



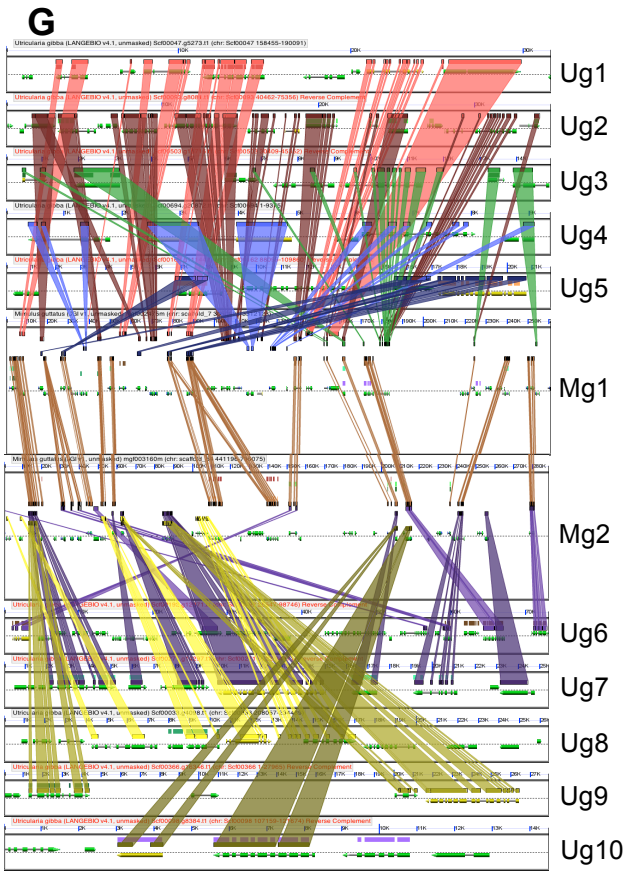
http://genomevolution.org/r/6m09



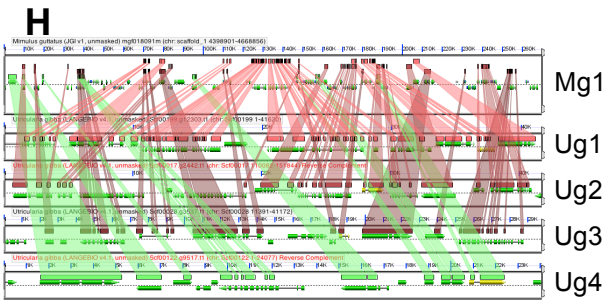
http://genomevolution.org/r/6m5c



http://genomevolution.org/r/6m5f

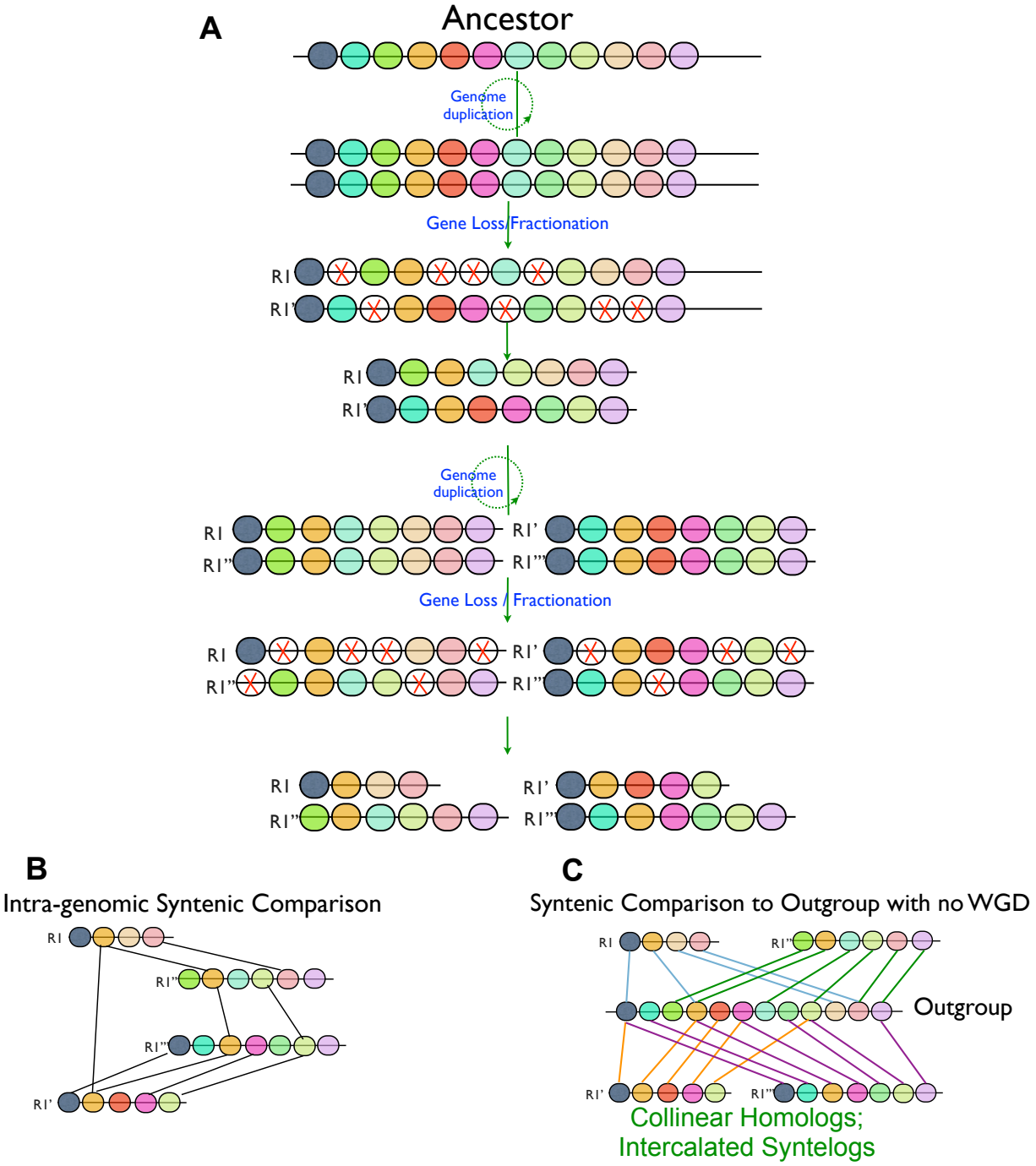


http://genomevolution.org/r/6m6v

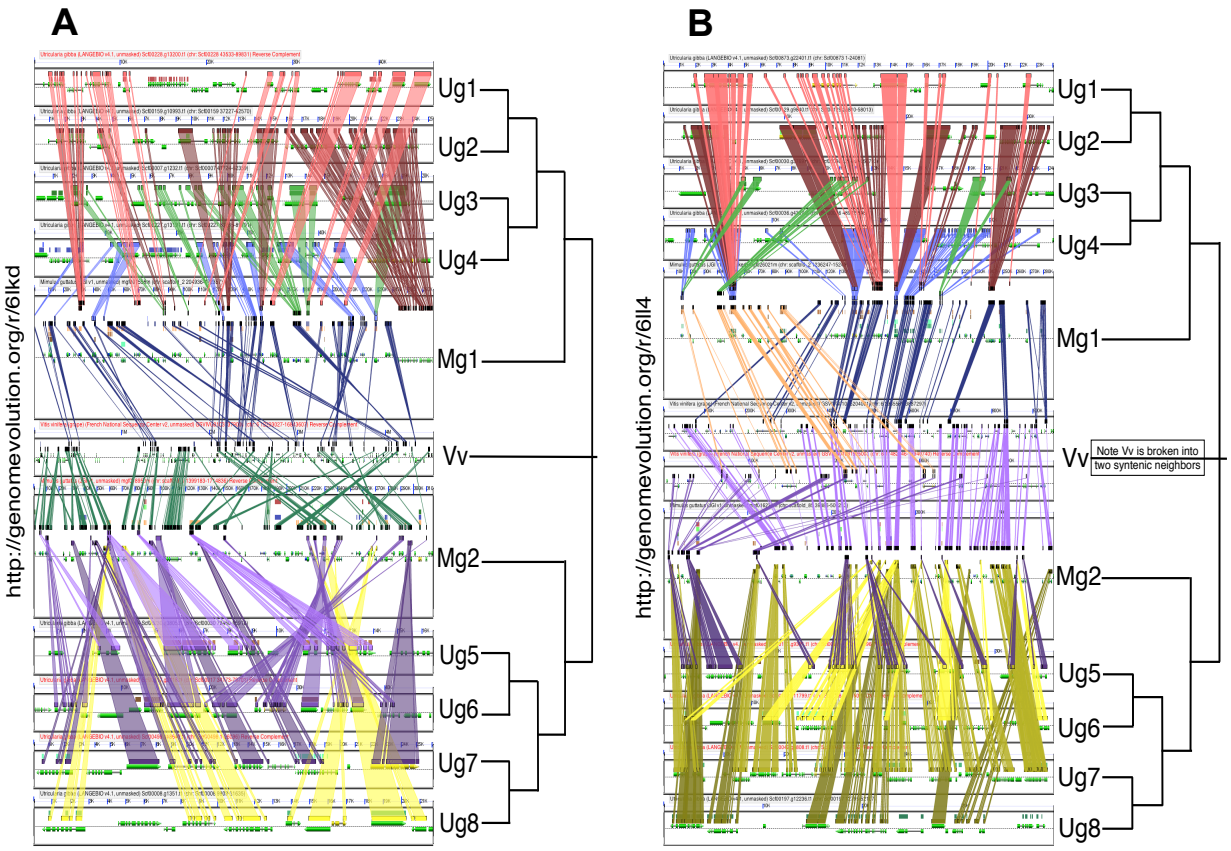


http://genomevolution.org/r/6127

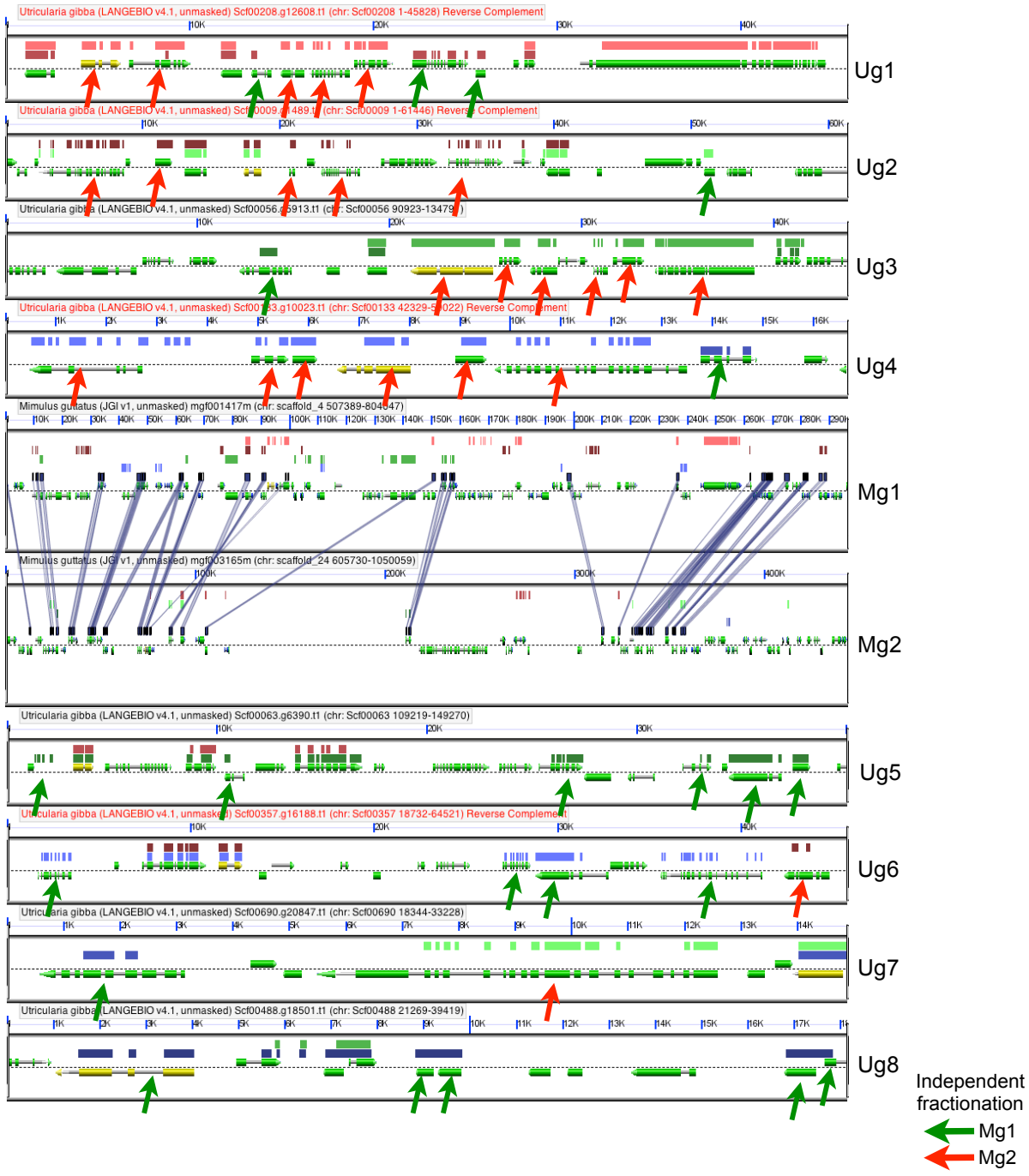
Supplementary Figure 20. Syntenic evidence that the genome of *U. gibba* (Ug) is 4x polyploid compared to the genome of *M. guttatus* (Mg). **(A)** Close up of a region of the syntenic dotplot shown in Sup. Fig. 19. Here, it is clearly shown that there are several regions of the Mg genome that are syntenic to 4 regions of Ug (square boxes). These regions are analysed for microsynteny and are shown in (B-H). Microsynteny analyses: For each set of 4xUg:1xMg syntenic regions identified, the syntenic region to Mg derived from its most recent WGD was identified and then used to identify additional Ug syntenic regions. For each of the microsynteny analyses, Mg regions are in the middle, with Ug regions above and below them. **(B)** 2xMg:8xUg. **(C)** 2xMg:8xUg. **(D)** 2xMg:8xUg. **(E)** 3xMg:7xUg. Note that two of the Mg regions are syntenic neighbors and syntenicly cover the other Mg region. Only 7 syntenic Ug regions were identified. The eighth may exist; Ug exists as too many small fragments to detect with our syntenic detection methods. **(F)** 2xMg:8xUg. **(G)** 2xMg:10xUg. Note that there are two pairs of Ug regions which are syntenic neighbours and together, this comes to 8xUg syntenic regions. **(H)** 1xMg:4xUg. No syntenic partner to the Mg region was identified.



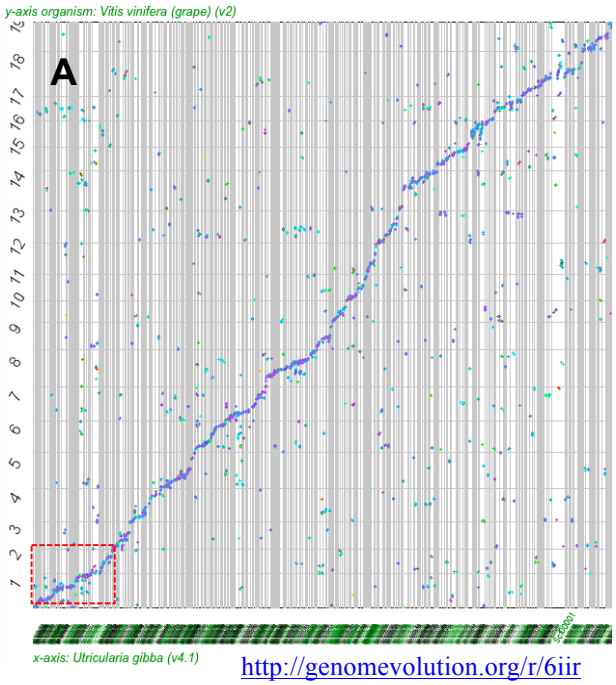
Supplementary Figure 21. Model for fractionation (homologous gene loss) following two rounds of WGD. **(A)** An ancestor genomic region is shown with coloured circles representing gene modes. Following a WGD, the genomic region and its underlying gene content is duplicated. Over time, homologous genes are lost (fractionated) from one genome region or the other. This pattern is repeated a second time resulting in four derived genomic regions. **(B)** Intragenomic comparison of the regions derived from (A). Note, there may be a very weak colinear signal amount syntenic region. **(C)** Comparison of the regions derived from (A) to an outgroup region without any WGD event. Note that each derived region has their entire gene content represented in the outgroup region and the colinear signal is much stronger. Also, there is a characteristic pattern of intercalated syntelogs when all four derived regions are compared to the outgroup region.

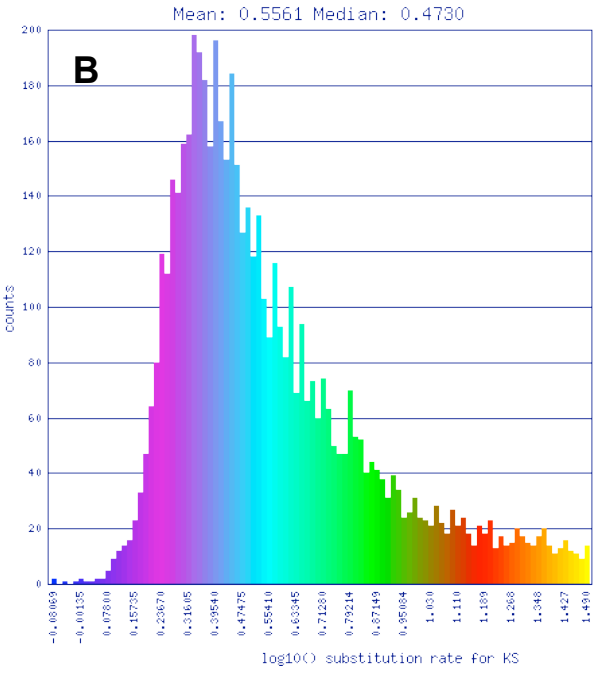


Supplementary Figure 22. Microsynteny analyses across *V. vinifera* (Vv), *M. guttatus* (Mg), and *U. gibba* (Ug). 1xVv:2xMgx8xUg. This shows the pattern of WGD following the divergence of the lineages. Vv has not undergone any ploypoidy event since their divergence; Mg has had one WGD; Ug has a three WGD events. **(A)** Derived from the same set of regions shown in Supplementary Figure 20C. **(B)** Derived from the same set of regions show in Supplementary Figure 20D. Note that in this analysis, two Vv regions were used to provide full syntenic coverage to the Mg regions.

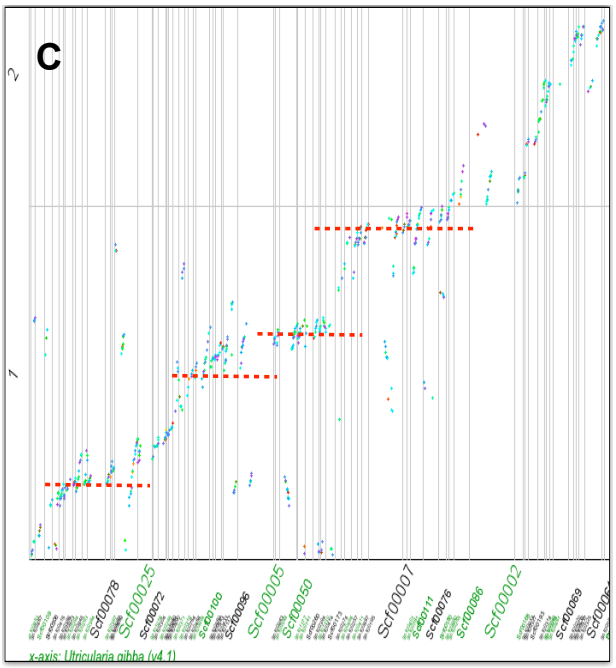


Supplementary Figure 23. Microsynteny analyses for independent fractionation across *Mimulus guttatus* (Mg) and *U. gibba* (Ug) showing a 1xVv:2xMgx8xUg. This region is the same as used for Supplementary Figure 20F. This shows the pattern of independent fractionation where genes in various Ug regions that are represented in either Mg region, but not necessarily both nor exclusively to one. Green arrows point to Ug genes that are fractionated (not present) from Mg1; red arrows point to Ug genes that are fractionated from Mg2. This pattern of independent fraction may be due to the Mg WGD being not shared with Ug.

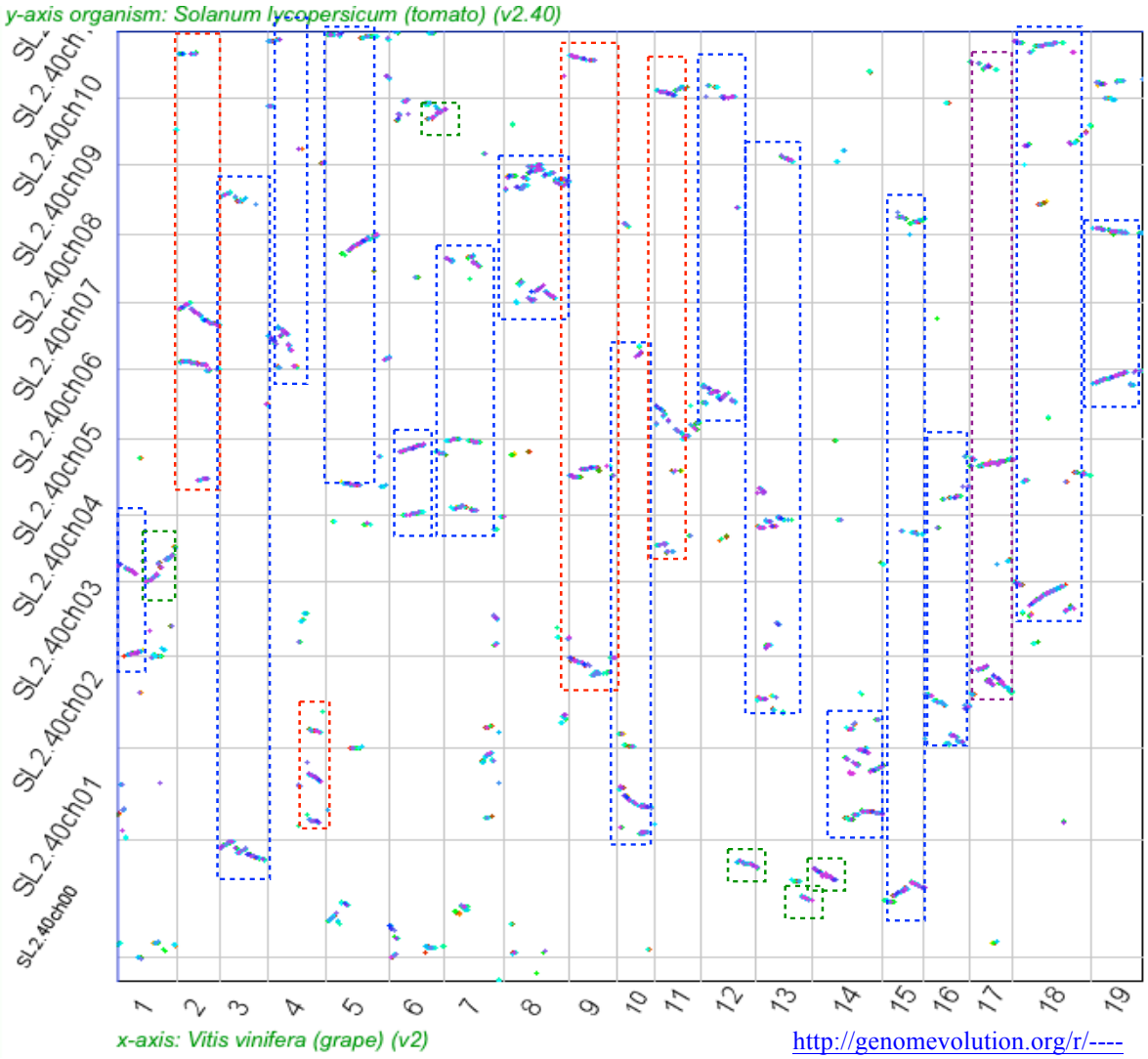




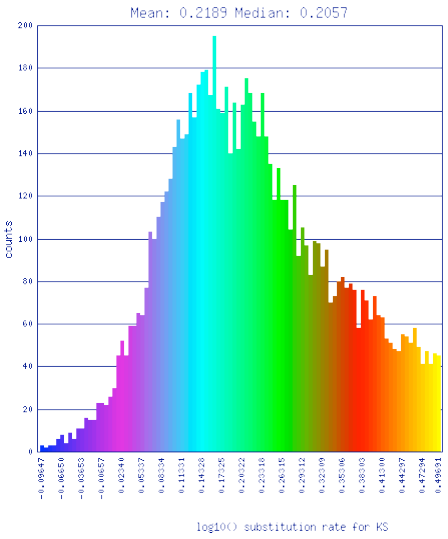
http://genomevolution.org/r/6iir



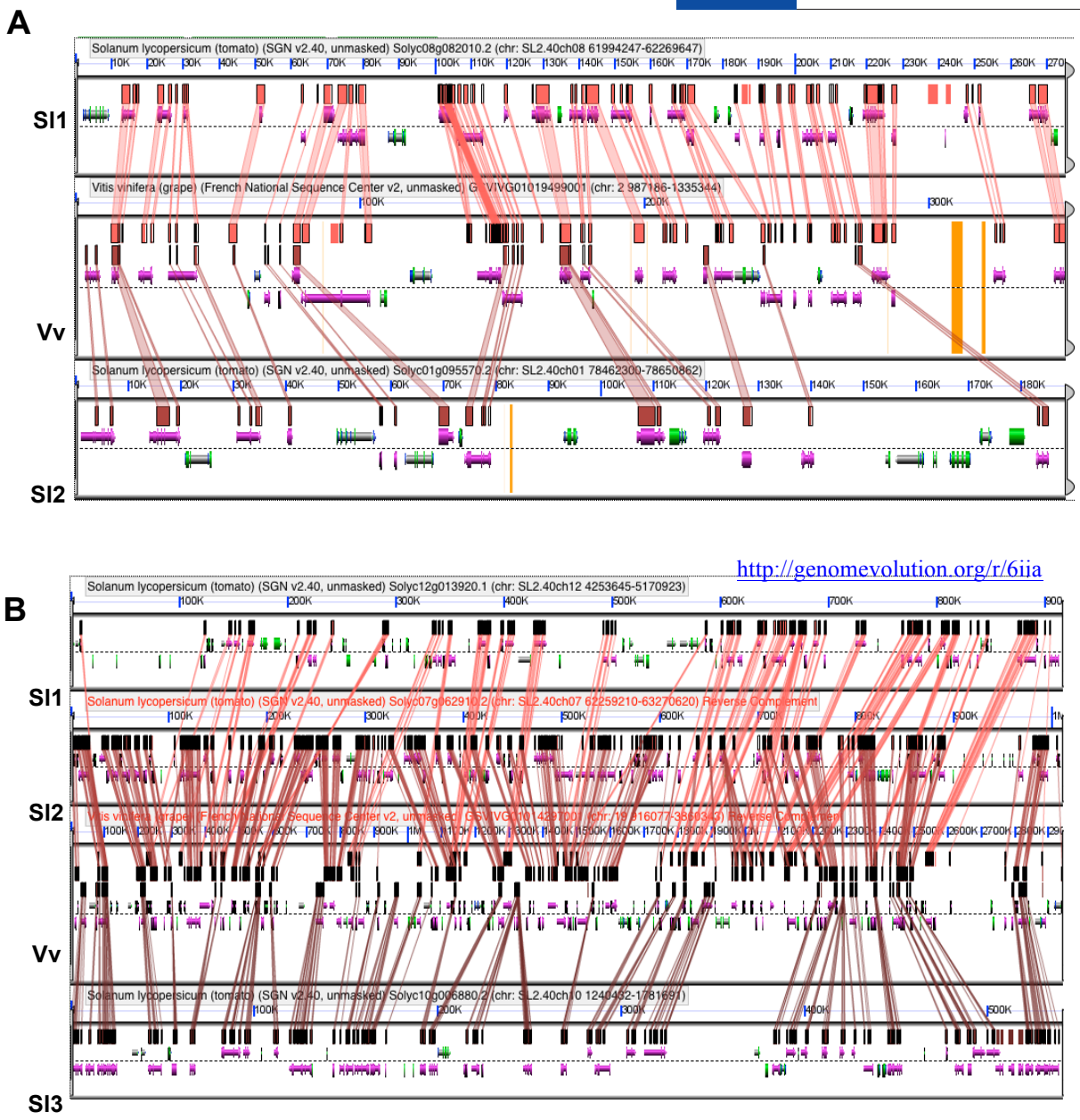
Supplementary Figure 24. (A) Syntenic dotplot of *U. gibba* (Ug; y-axis) versus *Vitis vinifera* (Vv; x-axis) with Ug's contigs ordered and oriented by SPA. Syntenic gene pair dots are coloured by Ks values. **(B)** Histogram of Ks values derived from syntenic gene pairs identified in (A). **(C)** High resolution of the region in (A) highlighted by the red dashed rectangle. Upon close inspection, there are multiple Ug regions syntenic to a single Vv region (red dashed lines). This pattern is expected if Ug has undergone three WGD events.



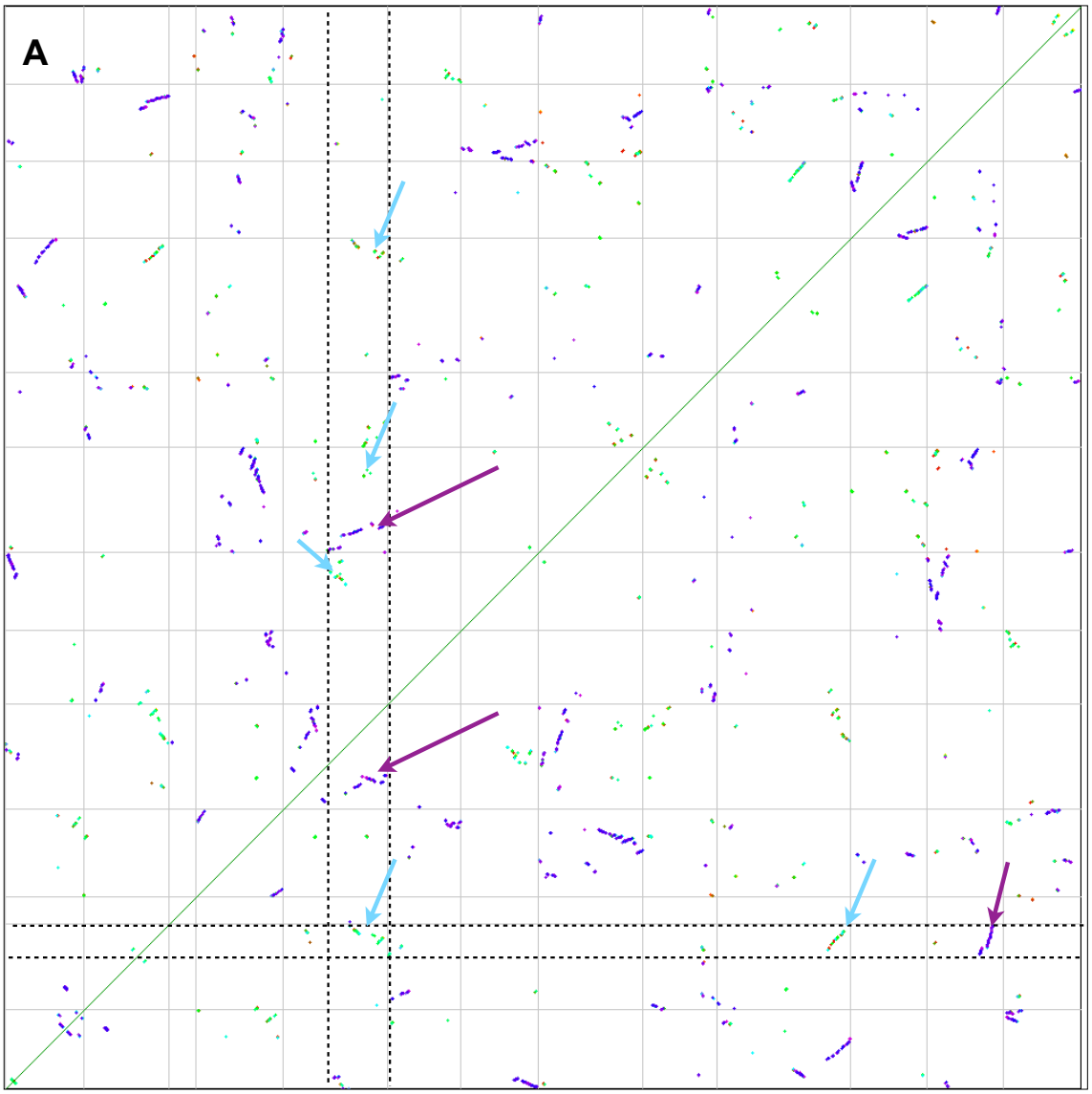
B



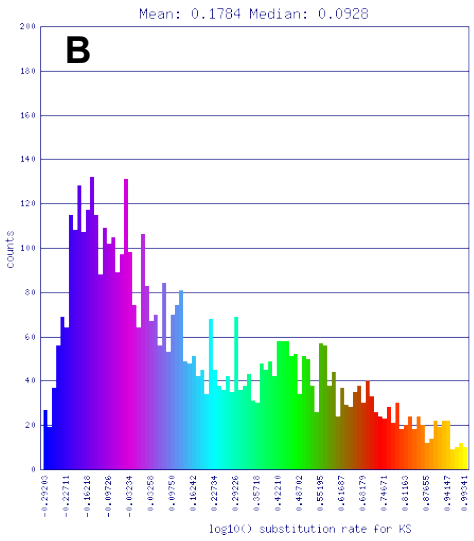
Supplementary Figure 25. (A) Syntenic dotplot of *Solanum lycopersicum* (Sl; y-axis) versus *Vitis vinifera* (Vv; x-axis). Syntenic gene pair dots are coloured by Ks values. Syntenic dotplot has been screen to show the best three regions of Sl to each region of Vv. Coloured dashed boxes correspond to Vv genomic regions that are represented as single (green), double (blue), or triple (red) copies in Vv. Note, orthologous regions are mostly purple, and those derived from the paleohexploidy are cyan. **(B)** Histogram of Ks values derived from syntenic gene pairs identified in (A).



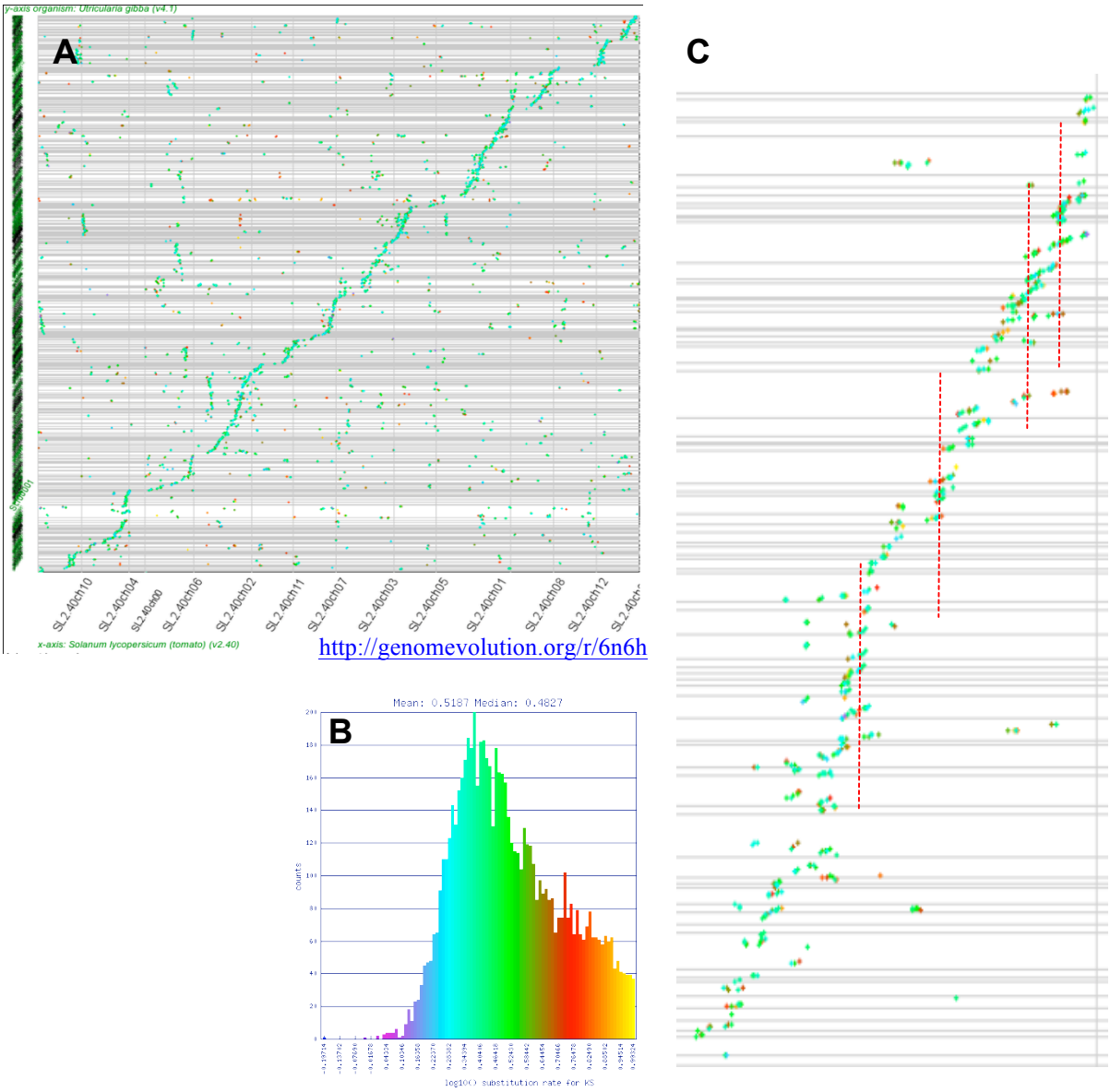
Supplementary Figure 26. Microsynteny analysis between *Solanum lycopersicum* (Sl) versus *Vitis vinifera* (Vv). Note that in these analyses, genes that are overlapped by regions of sequence similarity are coloured purple. **(A)** Microsynteny analysis between a region of Vv that shows synteny to two regions of Sl. Note that nearly the entire gene content of Vv is represented by the two Sl regions combined. (B) Microsynteny analysis between a region of Vv that shows synteny to three regions of Sl. Note that nearly the entire gene content of Vv is represented by the three Sl regions combined.



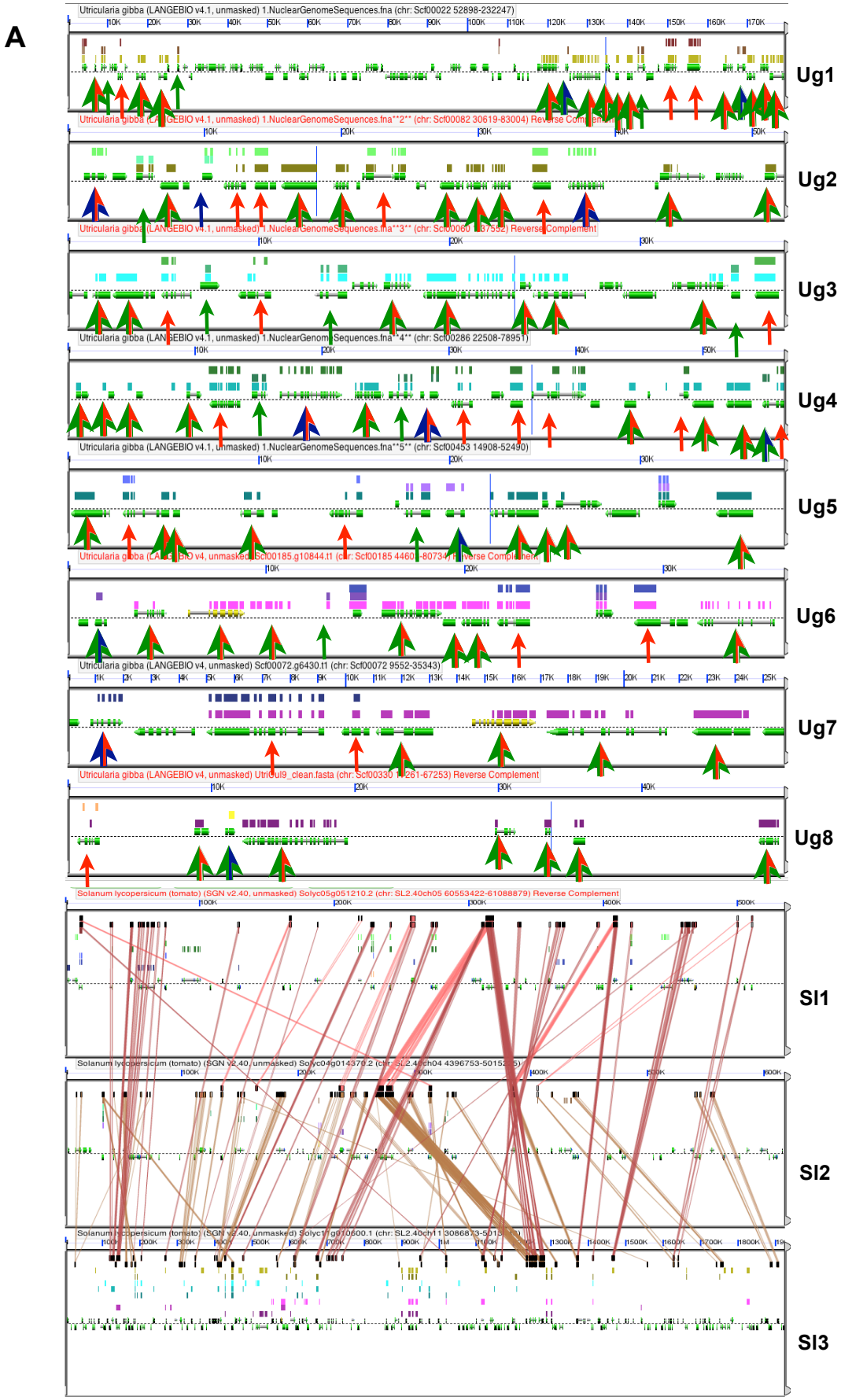
http://genomevolution.org/r/62tp



Supplementary Figure 27. (**A**) Self-self sytnenic dotplot of *Solanum lycopersicum* (Sl). Syntenic gene pairs are coloured by their Ks values. Purple regions are derived from the most recent polyploidy event; cyan from the eudicot paleohexaploidy event. Dashed line highlights regions that are duplicated (x-axis; one purple region) triplicated (y-axis; two purple regions. (**B**). Histogram of Ks values. Interestingly, there is a mixture of duplicated and triplicated regions in tomato instead of a pure triplication as reported during the sequencing of its genome[144]. A more in-depth analysis and discussion can be found at http://genomevolution.org/r/53g5.

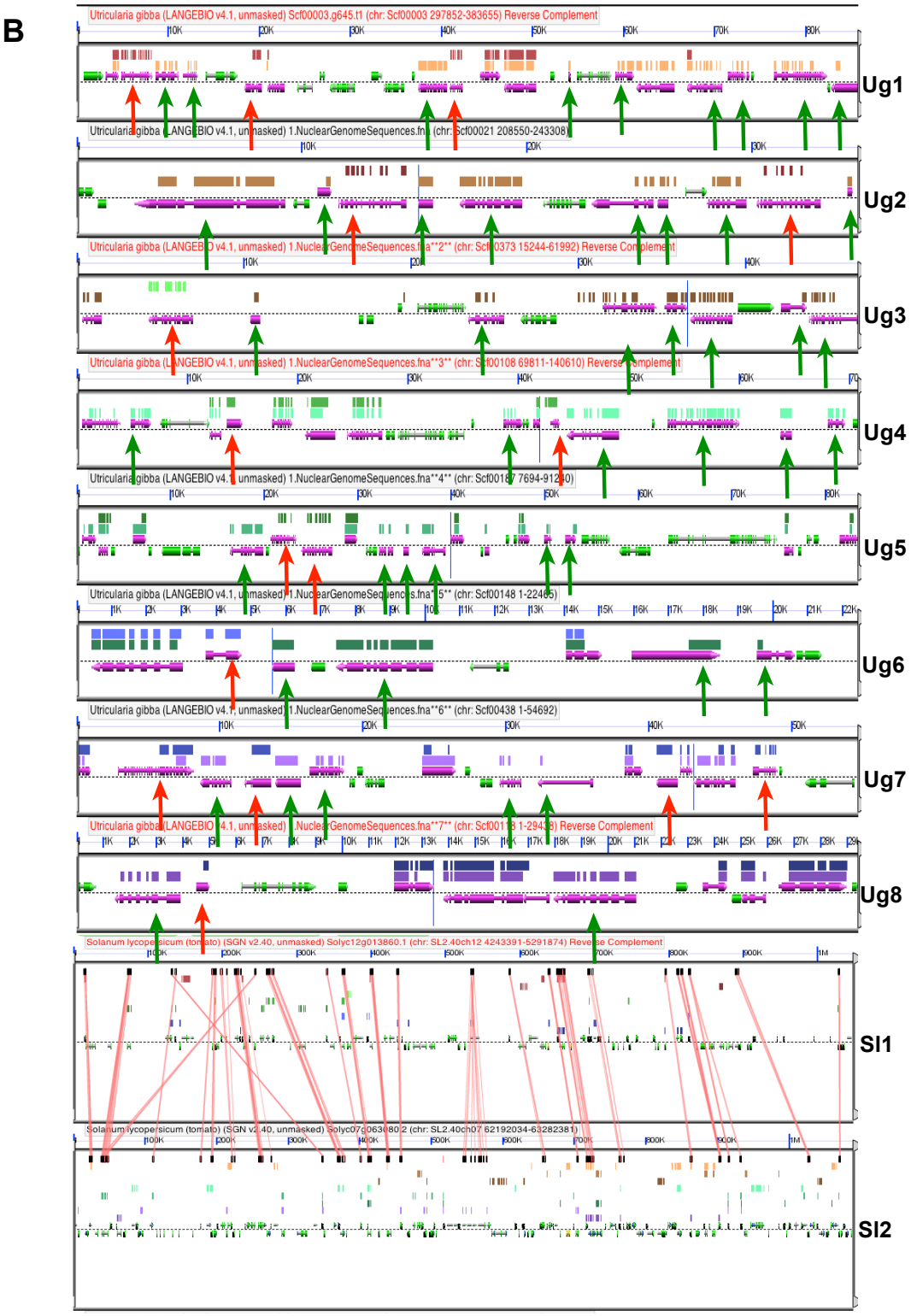


Supplementary Figure 28. Syntenic dotplot of the tomato genome (x-axis) vs. the *U. gibba* genome (y-axis) (**A**). Scaffolds/contigs of *U. gibba* have been ordered and oriented according to their syntenic relationship to the tomato genome (syntenic path assembly). Syntenic gene pairs are coloured by their $\log_{10}$ transformed synonymous mutation values (Ks) as visualised in the histogram shown in (**B**). Note that the large syntenic blocks are from the same portion of the Ks value distribution (green), including those matching multiple syntenic region in tomato derived from its most recent polyploidy event. This is evidence that *U. gibba*'s polyploidy events are independent from tomato's most recent polyploidy event. (**C**) Zooming in on a portion of the syntenic dotplot shows that many *U. gibba* scaffolds/contigs (y-axis) are syntenic to the same region of tomato (x-axis) and visualised by the red dashed lines, providing further evidence of multiple polyploidy events in the *U. gibba* lineage.



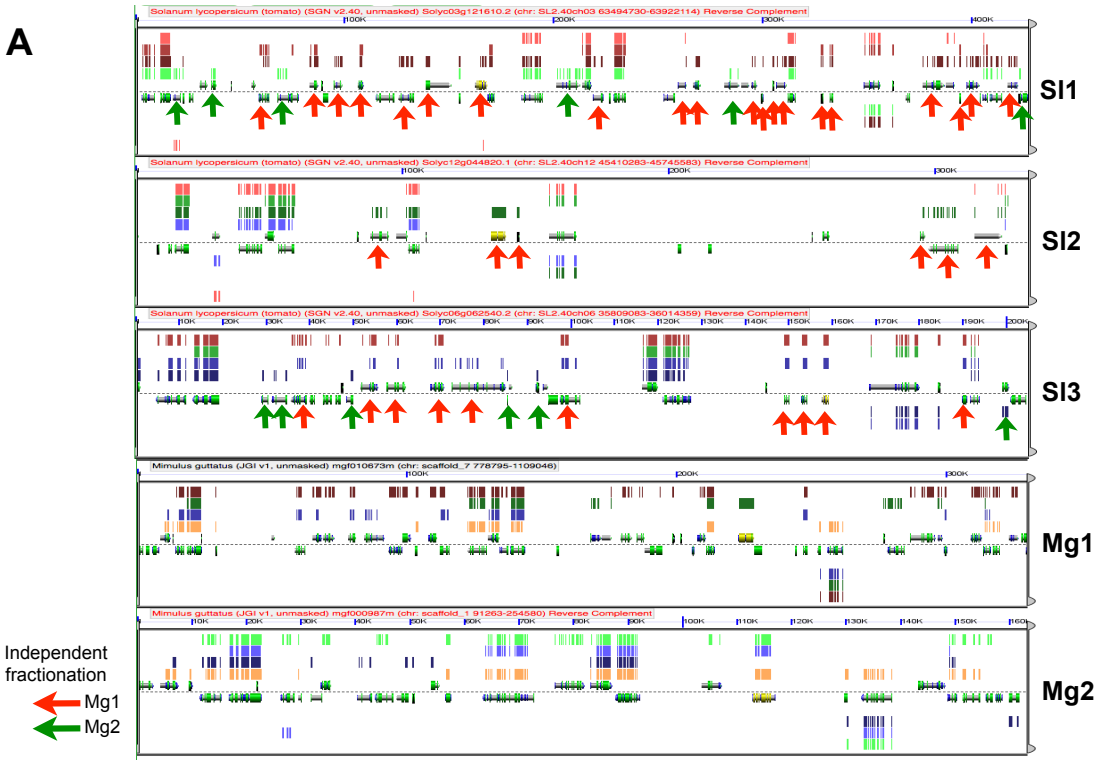
http://genomevolution.org/r/52cx

B

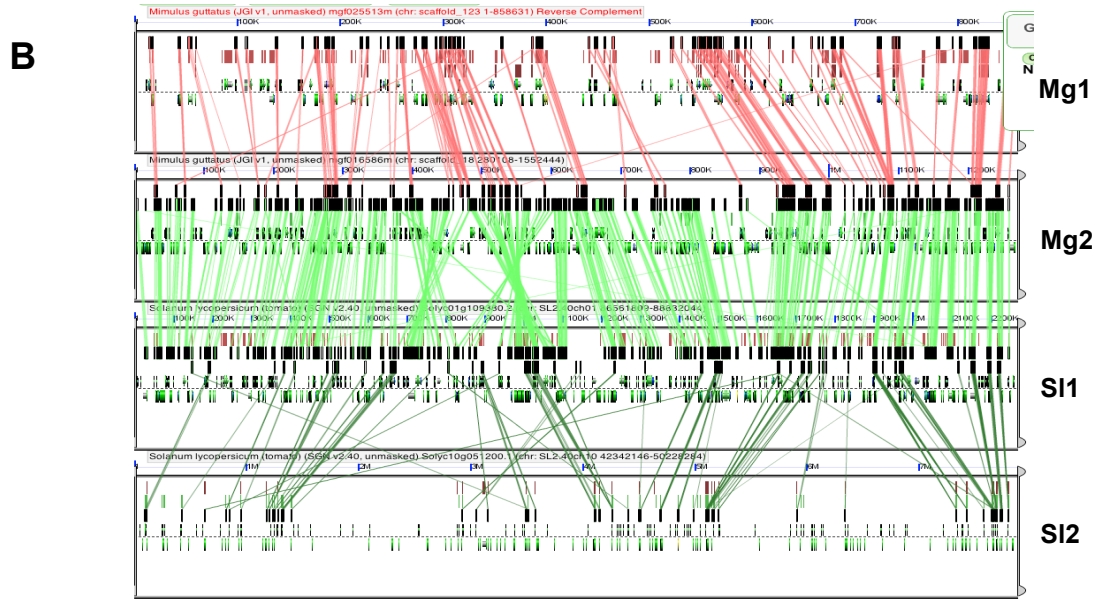


http://genomevolution.org/r/6jgm

Supplementary Figure 29. These 2 figures (**A** and **B**), show that the three WGD events in *U. gibba* are independent of tomato's mixed triplication(top)/ duplication(bottom) based on differential fractionation of gene content following polyploidy. Each panel represents a syntenic genomic region. The dashed line in each panel separates the top and bottom strands of DNA, gene models are composite green/yellow/grey arrows, and regions of sequence similarity are represented as additional tracks of coloured blocks. Panels from *U. gibba* are labeled Ug followed by their scaffold number, and those from tomato are labeled Sl followed by the chromosome number. Each differentially fractionated syntenic gene between tomato and *U. gibba* has been labeled by an arrow to signify which tomato region has lost it. Independence of polyploidy events is evidenced by each syntenic region of *U. gibba* having genes differentially lost among the tomato regions. If these lineages shared tomato's most recent polyploidy event, then we would expect an equal proportion of the *U. gibba* regions to be most similar to one region of tomato based on retention of gene content from their common ancestry. Instead, a given region of tomato appears to be dominant in terms of retaining gene content in *U. gibba*, but all tomato regions have their gene content represented among the combined *U. gibba* regions, not split to half of the *U. gibba* regions. This pattern is predicted by biased fractionation following polyploidy[145,146] in tomato.

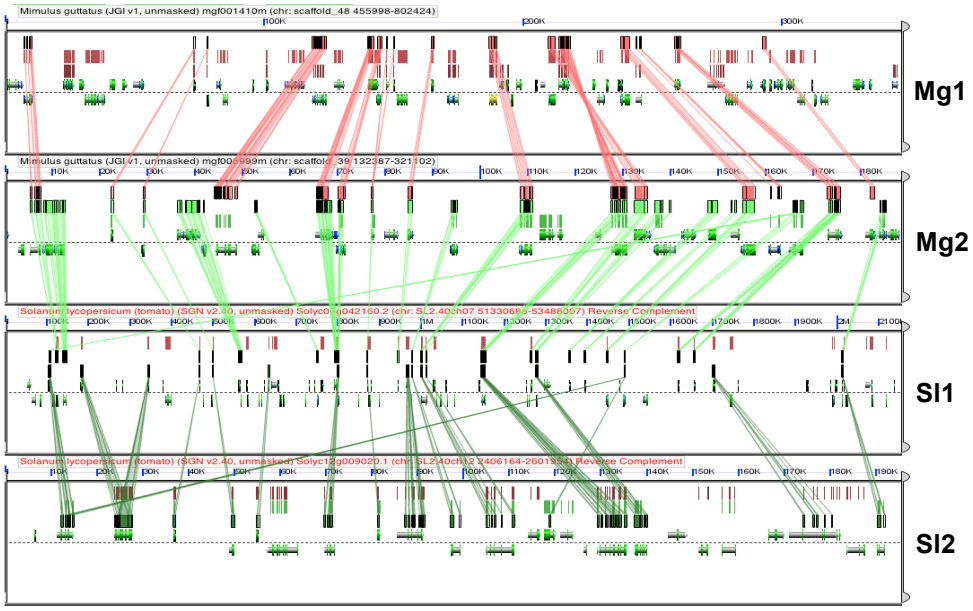


http://genomevolution.org/r/51mk



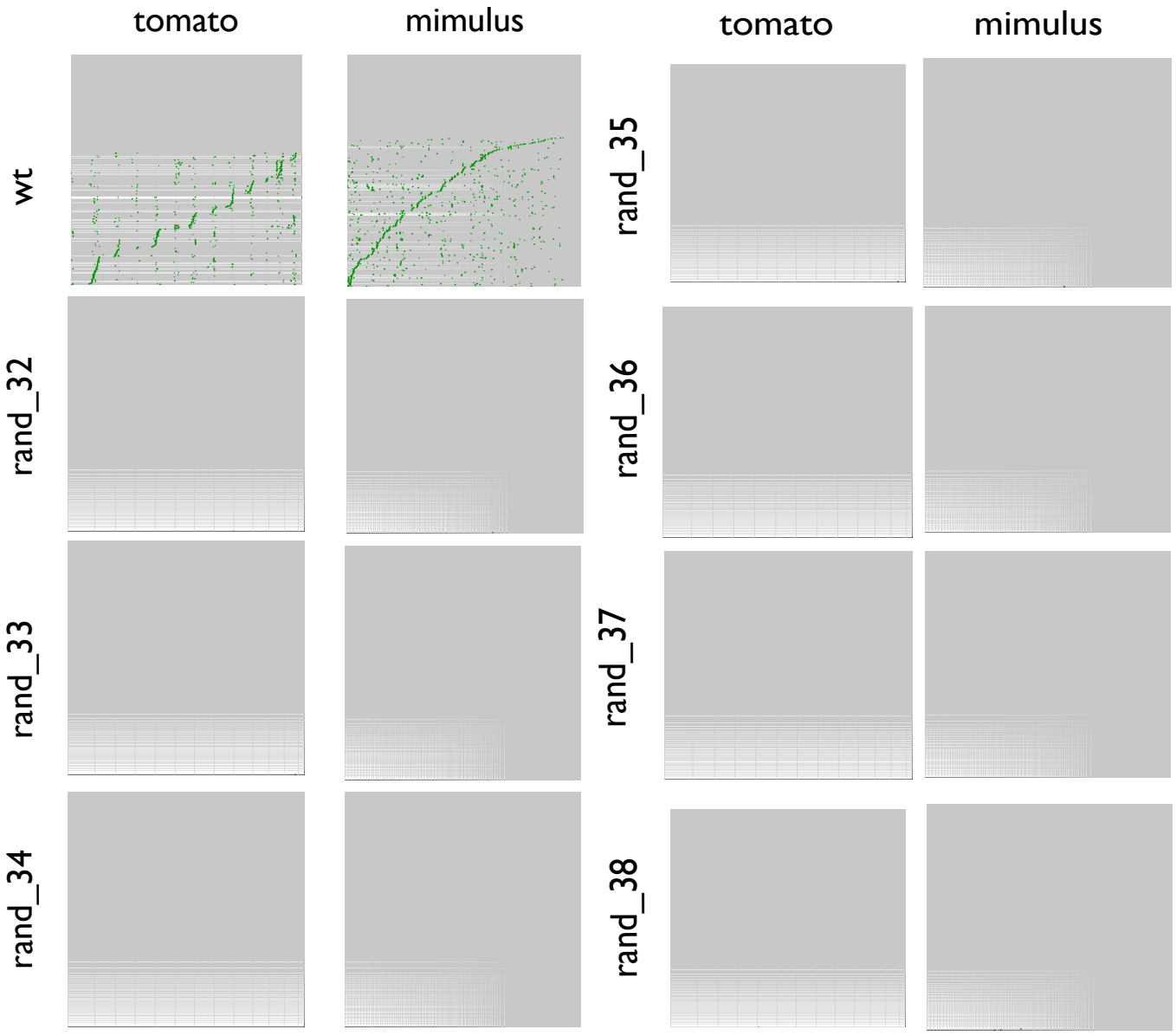
http://genomevolution.org/r/6kyt

C

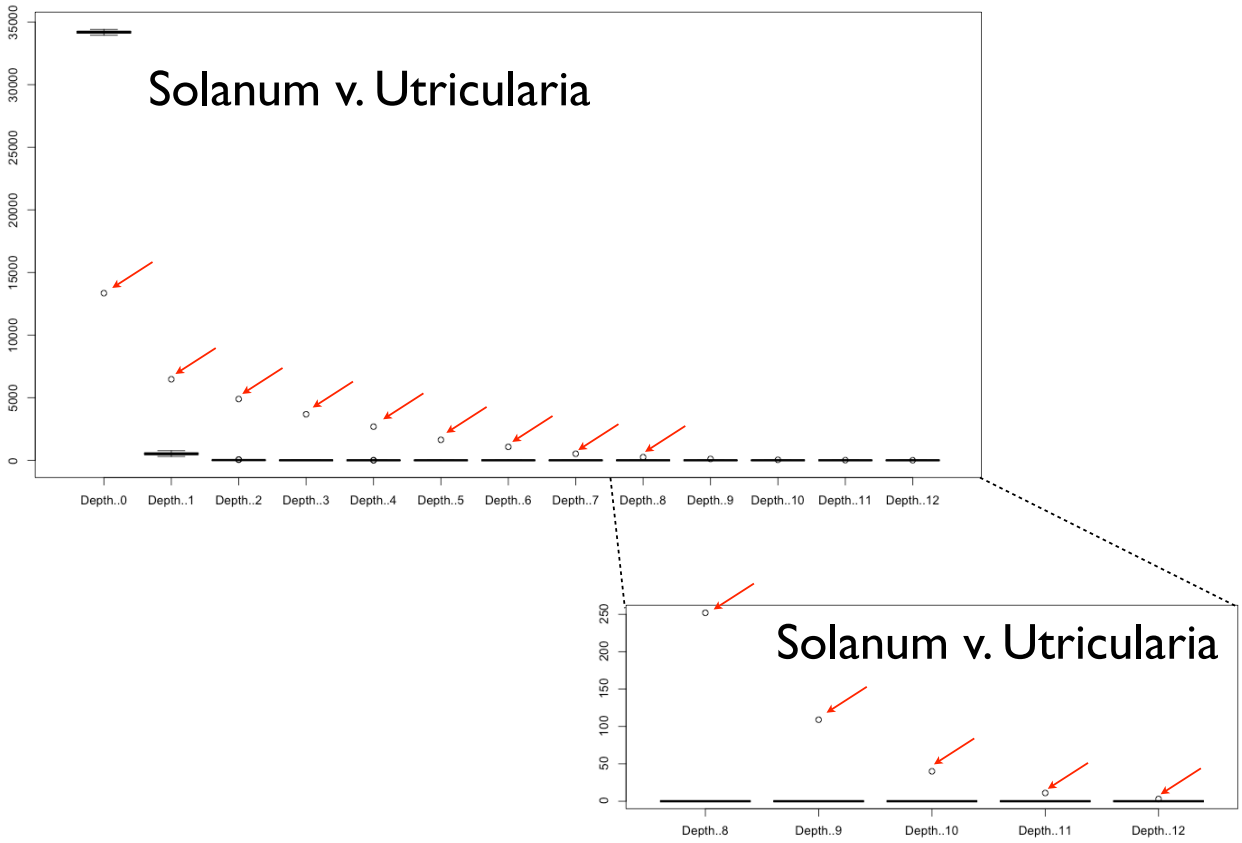


http://genomevolution.org/r/6kzl

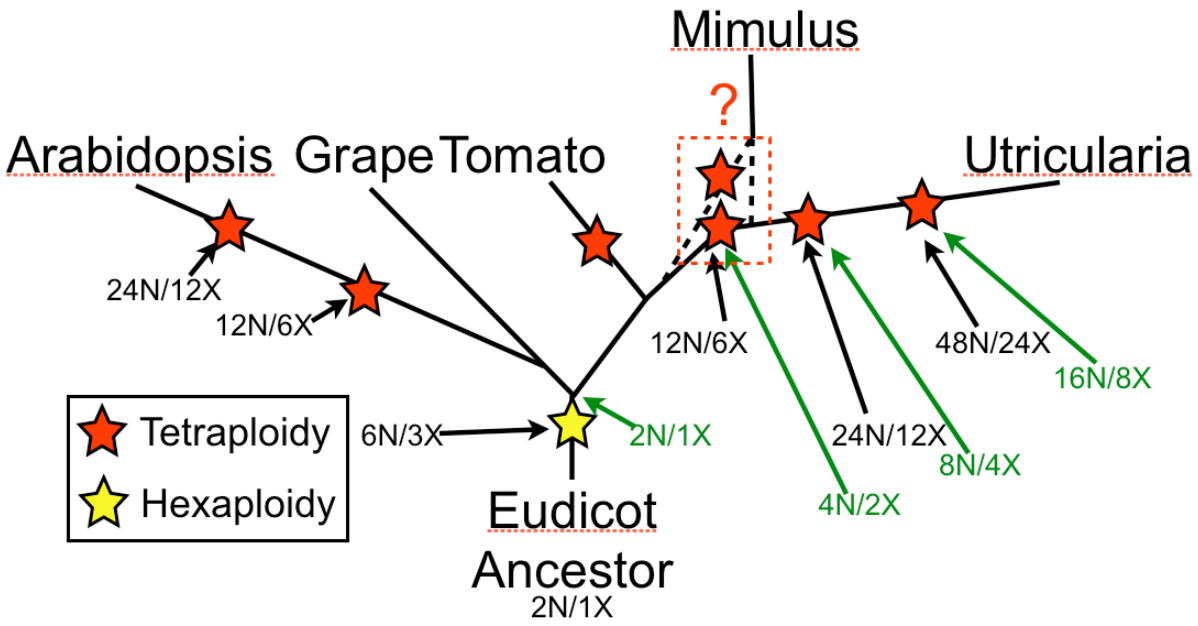
Supplementary Figure 30. Microsynteny analysis between syntenic regions of *Solanum lycopersicum* (Sl) and *Mimulus guttatus* (Mg) showing independent fractionation. **(A)** The Sl regions are differentially fractionation compared to the Mg regions, which is evidence that these regions are derived from independent polyploidy events in these lineages. Red arrows are genes lost in Mg1; green arrows are genes lost in Mg2. **(B)** and **(C)** differential fractionation can be seen in Sl1 where the red dashed box outlines regions of sequence similarity to Mg.



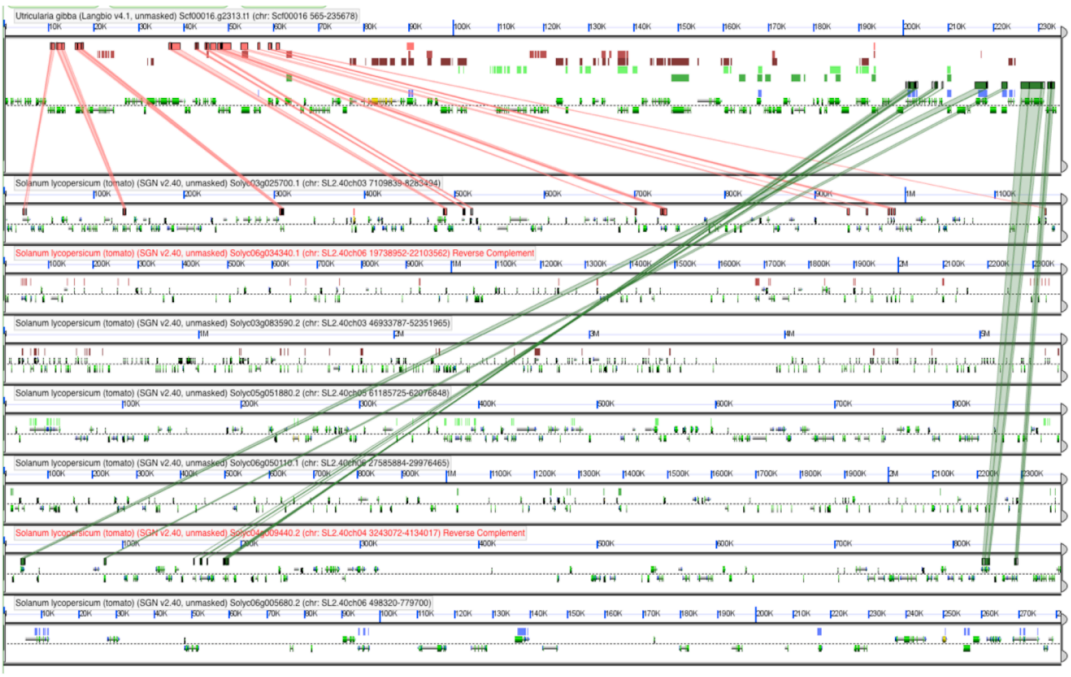
Supplementary Figure 31. Syntenic dotplots between the genomes of *Solanum lycopersicum* (tomato) or *Mimulus guttatus* (mimulus) and the wild-type (wt) or randomised (rand_XX) genomes of *U. gibba*. Randomised *U. gibba* genomes have same number of chromosomes and same gene density per chromosome, but randomised location of gene content. Eight of the 100 randomised genomes are shown. Note that the randomised genomes (rand_XX) have nearly no synteny using the same parameters used that detect extensive (nearly genome-die) synteny with the wild-type (WT) genome. This shows the near complete destruction of the syntenic signal in the randomised genomes.



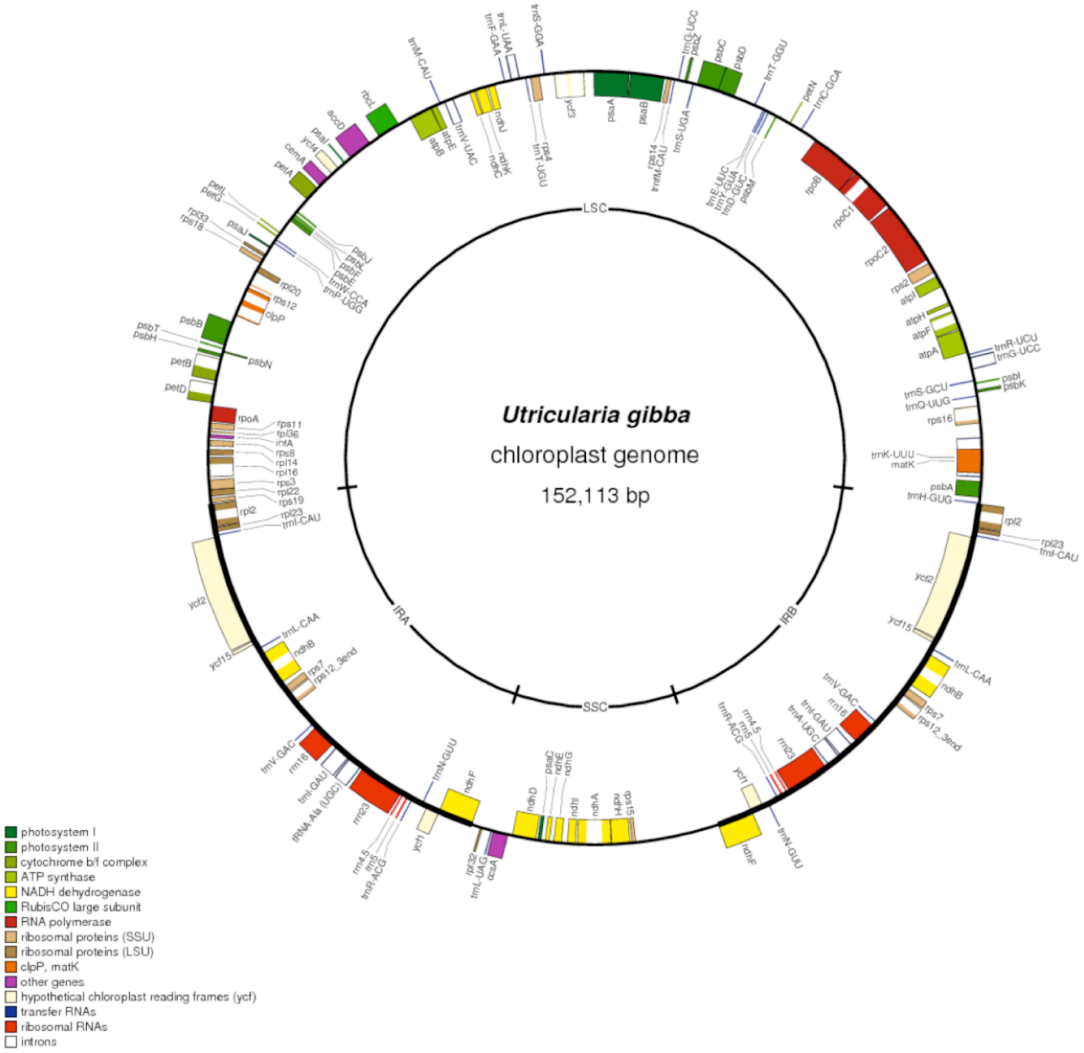
Supplementary Figure 32. Syntenic Depth analysis of randomised *U. gibba* (Ug) genomes versus *Solanum lycopersicum* (Sl)**.** Randomised Ug genomes have same number of chromosomes and same gene density per chromosome, but randomised location of gene content. Syntenic Depth measures the number of syntenic regions identified in genome A for a given gene in genome B. A syntenic depth of 0 means that no syntenic regions were identified; a syntenic depth of 1 means that one syntenic region was identified. These data have been pooled for the genes of Sl and averaged against 100 randomised genomes of Ug. Parameters are used to identify syntenic regions 4 collinear genes in a window of 80. The red arrows point to the syntenic depth of the wild-type Ug genome. All values are statistically different than mean of observed genes for a given syntenic depth for randomised genomes. This is further tested with Two-tailed probability-value of a z-test for both Sl and *Mimulus guttatus* under two sets of parameters in Supplementary Tables 35 and 36.



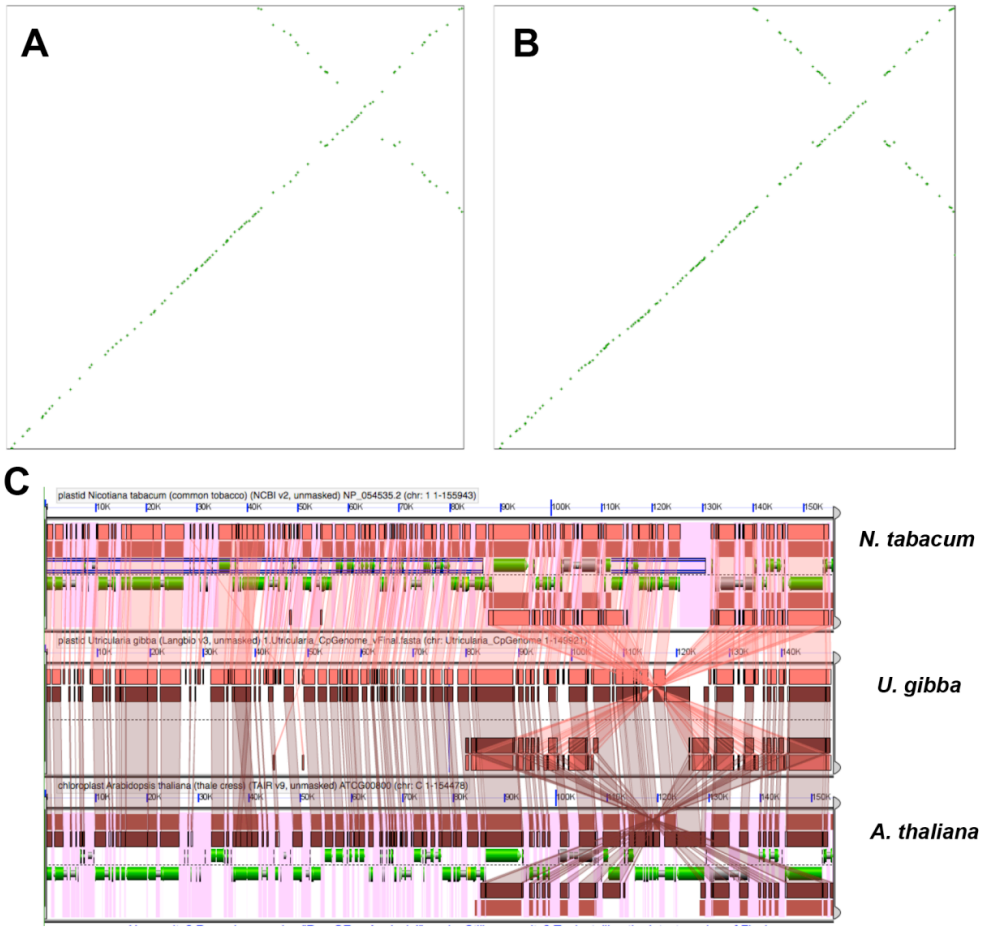
Supplementary Figure 33. Monoploid (x) and haploid (n) numbers assignable to the genomes under study here, with respect to the eudicot post-paleohexaploid ancestor (green) and the immediate pre-hexaploid ancestor (black). As such, *U. gibba* is 16n/8x with respect to the post-paleohexaploid ancestor, and 48n/24x with respect to the pre-hexaploid ancestor. While the *Mimulus* a WGD, it is uncertain if this event is shared with *U. gibba.*



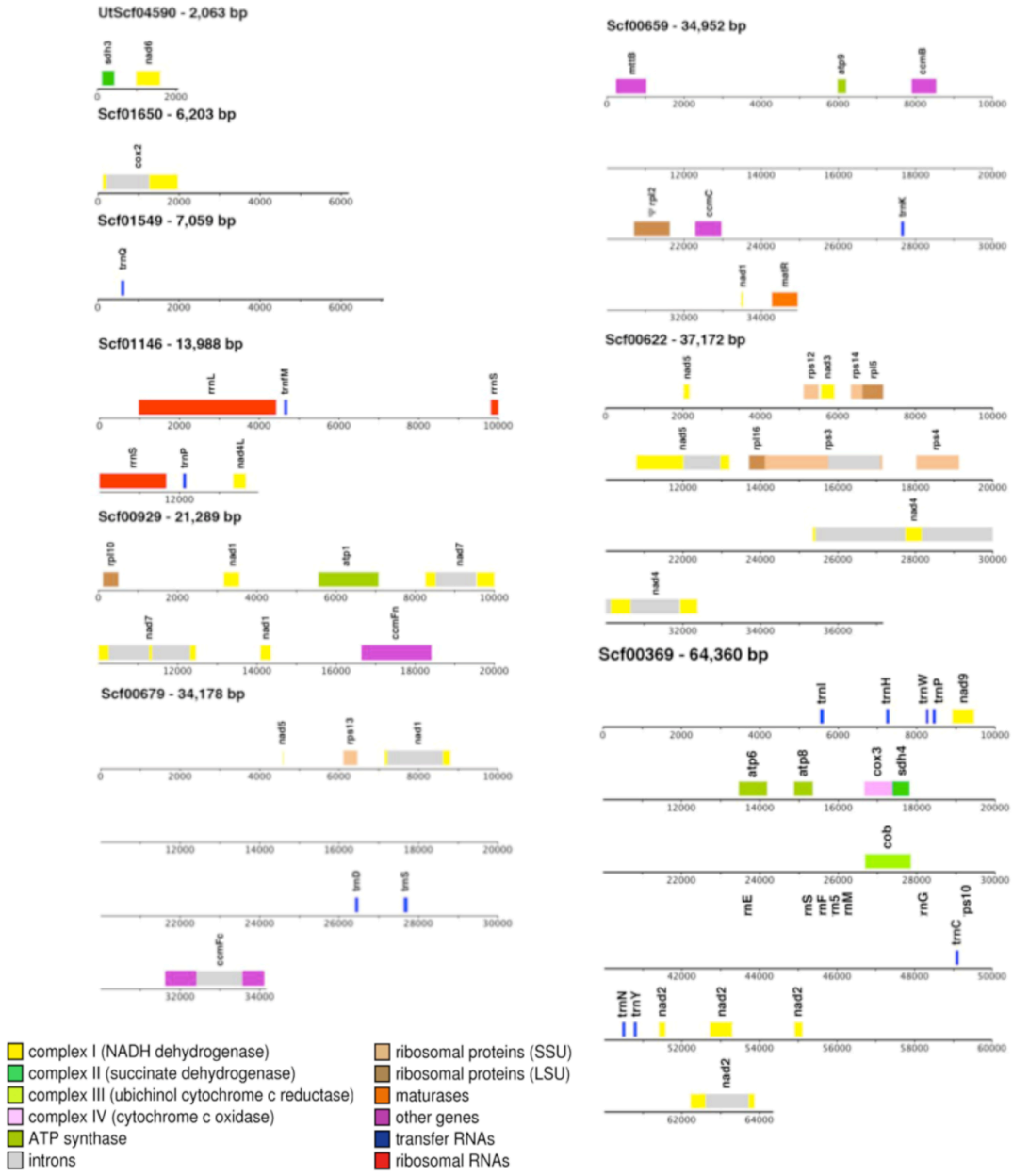
Supplementary Figure 34. Syntenic mapping of one scaffold of the *U. gibba* genome to multiple genomic regions in tomato, providing evidence of multiple fusion events of ancestral chromosomes in the lineage leading to *U. gibba*. The top panel shows a 80kb region of the genome of *U. gibba* from Scf00020. The dashed line in the panel separates the top and bottom strands of DNA and gene models are shown as composite coloured arrows. Above the gene models are tracks of coloured blocks with each track representing regions of sequence similarity to different tomato genomic regions. Each pair of *U. gibba*-tomato regions are syntenic based on the evidence of a collinear homologous gene pairs. Each region of *U. gibba* matches approximately two or three regions of tomato, which is expected due to the independent duplications in that lineage. The *U. gibba* region shows a synteny to a series of regions of the tomato genome located on chromosomes 3, 6, 5, 4, and separated by tens of megabases if located on the same chromosomes. This provides evidence that during the multiple rounds of WGD and fractionation in the lineage of *U. gibba*, the genome underwent several chromosome fusion events. The results displayed here may be regenerated at http://genomevolution.org/r/58e0.



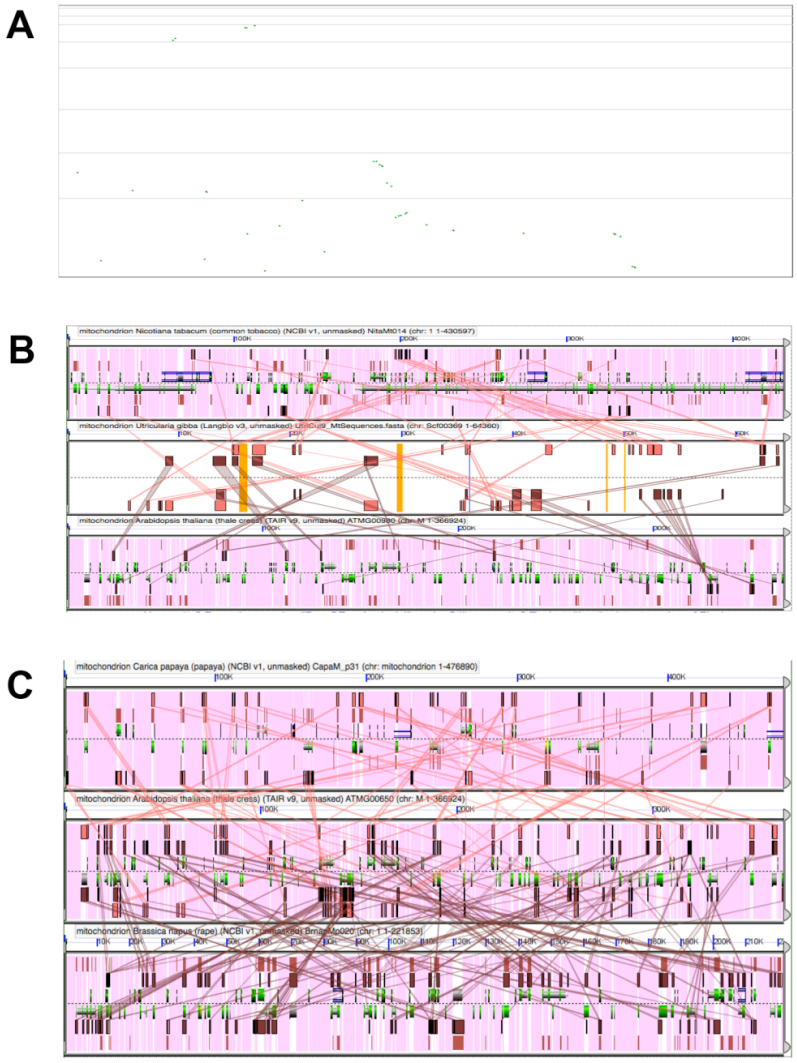
Supplementary Figure 35. Gene map of the chloroplast genome of *U. gibba*. This figure was made using the program OGDraw[156].



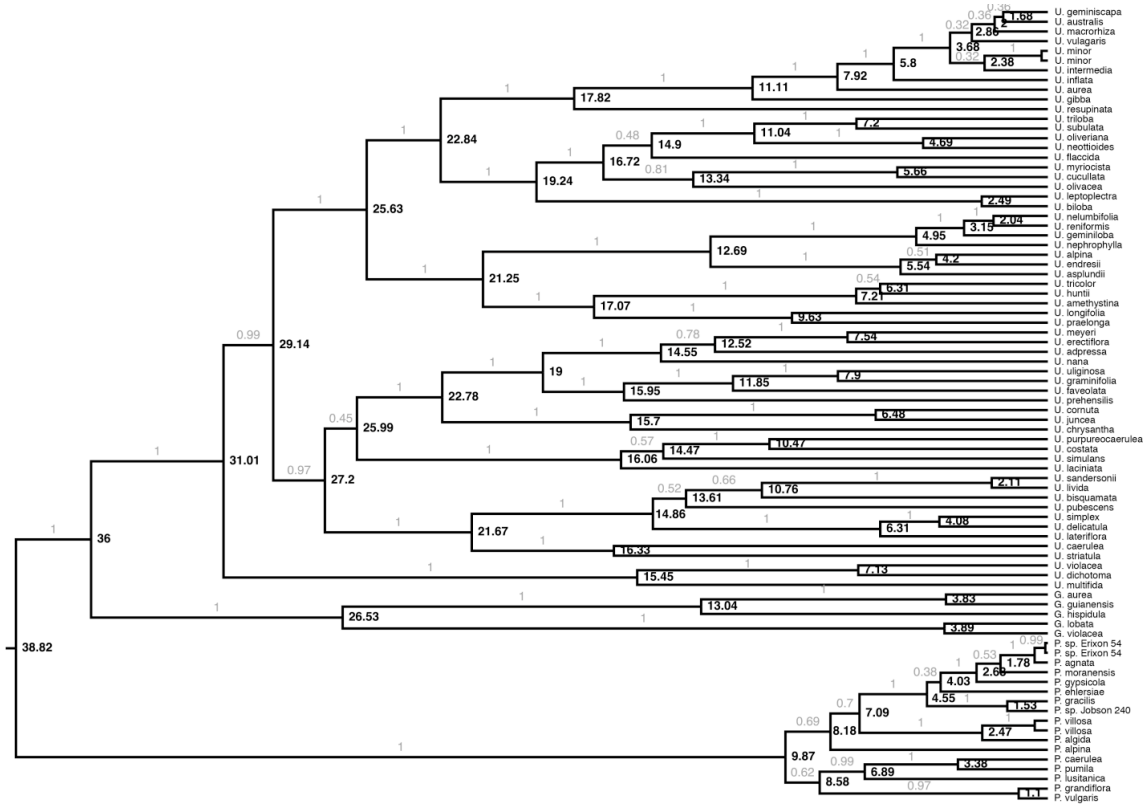
Supplementary Figure 36. Syntey between the *U. gibba* chloroplast genome (Cp) and other model plant species (*Nicotiana tabacum* and *Arabidopsis thaliana*). Syntenic dotplots of the *N. tabacum* Cp (x-axis) vs. the *U. gibba* Cp genome (y-axis) (**A**), and the *A. thaliana* Cp (x-axis) vs. the *U. gibba* Cp (y-axis) (**B**). GeVo analysis (C) showing synteny of the chloroplast genomes of *N. tabacum* and *U. gibba* (orange), and between *U. gibba* and *A. thaliana* (brown). Non-coding regions in *N. tabacum* and *A. thaliana* are masked in light purple. Inverted repeats are present on the right side of the figure.



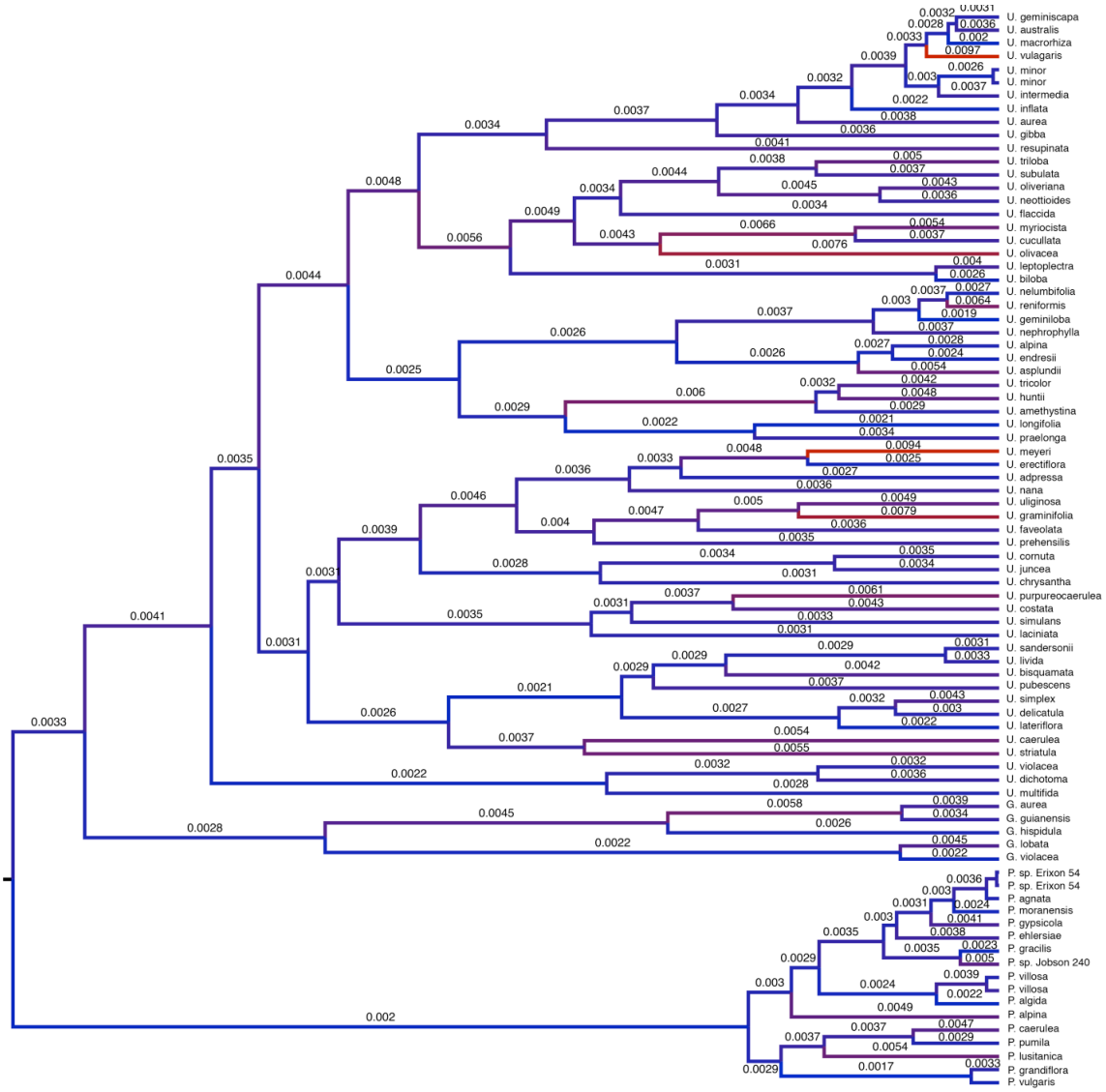
Supplementary Figure 37. Gene map of the scaffold/contigs from the mitochondrial genome of *U. gibba*. This figure was made using the program OGDraw[156].



Supplementary Figure 38. Analysis of synteny of the *U. gibba* mitochondrial (mt) genome. (**A**) Syntenic dotplot of the *U. gibba* mitochondrial (10 scaffolds/contigs) genome (x-axis) vs. the *Nicotiana tabacum* mitochondrial genome (y-axis). (**B**) GeVo analysis showing relatively low synteny of the mitochondrial genome of *N. tabacum* and the largest mitochondrial sequence of *U. gibba* (upper, pink lines), and between *U. gibba* and *Arabidopsis thaliana* (lower, dark red lines). Non-coding regions in *N. tabacum* and *A. thaliana* are masked in light purple. (**C**) Similar analysis showing low synteny in the mitochondrial genomes of plants from the same order (*A. thaliana*, *Brassica napus*, and *Carica papaya*; Brassicales) and family (*A. thaliana* and *B. napus*; Brassicaceae). *C. papaya* and *A. thaliana* are above, with pink lines, and *A. thaliana* and *Brassica napus* are below, with dark red lines. Non-coding regions are masked in light purple.



Supplementary Figure 39. Phylogenetic tree of Lentibulariaceae *trnL–trnF* chloroplast sequences from BEAST depicting estimated divergence dates. Numbers at selected nodes indicate the estimated divergence date, followed by the 95% HPD in brackets. Numbers along the branches represent the posterior probability.



Supplementary Figure 40. Phylogenetic tree of Lentibulariaceae *trnL-trnF* chloroplast sequences from BEAST depicting rate estimates. Branches are coloured according to rate, with a gradient between blue (relatively slow) and red (relatively fast). Numbers above the branch indicate their rate in substitutions per site per million years.