A novel phylogenomics pipeline reveals complex pattern of reticulate evolution in Cucurbitales

Running Title: **CAPTUS: a novel pipeline for phylogenomics**

Edgardo M. Ortiz[1*], Alina Höwener[1,2], Gentaro Shigita[1], Mustafa Raza[1], Olivier Maurin[3], Alexandre Zuntini[3], Félix Forest[3], William J. Baker[3] & Hanno Schaefer[1]

[1] Plant Biodiversity, Department Life Science Systems, Technical University of Munich (TUM), Emil-Ramann-Str. 2, D-85354 Freising, Germany

[2] current address: Systematics, Biodiversity & Evolution of Plants, Ludwig-Maximilian University Munich (LMU), Menzingerstr. 67, D-80638 Munich, Germany.

[3] Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, United Kingdom

[*] E-mail: e.ortiz.v@gmail.com

Abstract.--- A diverse range of high-throughput sequencing data, such as target capture, RNA-Seq, genome skimming, and high-depth whole genome sequencing, are amenable to phylogenomic analyses but the integration of such mixed data types into a single phylogenomic dataset requires a number of bioinformatic tools and significant computational resources. Here, we present a novel pipeline, CAPTUS, to analyze mixed data in a fast and efficient way. CAPTUS assembles these data types, allows searching of the assemblies for loci of interest, and finally produces alignments that have been filtered for paralogs. Compared to other software, CAPTUS allows the recovery of a greater number of more complete loci across a larger number of species. We apply CAPTUS to assemble a comprehensive mixed dataset, comprising the four types of sequencing data for the angiosperm order Cucurbitales, a clade of about 3,100 species in eight mainly tropical plant families, including begonias (Begoniaceae) and gourds (Cucurbitaceae). Our phylogenomic results support the currently accepted circumscription of Cucurbitales except for the position of the holoparasitic

Apodanthaceae. Within Cucurbitaceae, we confirm the monophyly of all currently accepted tribes. However, we also reveal deep reticulation patterns both in Cucurbitales and within Cucurbitaceae. We show that conflicting results of earlier phylogenetic studies in Cucurbitales can be reconciled when accounting for gene tree conflict.

Keywords.--- Angiosperms353, Cucurbitaceae, Cucurbitales, gene tree discordance, paralog filtering, genome skimming, phylotranscriptomics, nitrogen-fixing clade

Modern sequencing technology allows the rapid accumulation of large amounts of genomic data at low cost. The analysis of these raw data, however, is still a bottleneck and there is strong demand for skilled bioinformaticians who can handle such complex data in industry and in research institutions worldwide. Groups who cannot afford to pay for a bioinformatics expert rely on user-friendly analysis pipelines which are accessible without specialist training. For target capture data, HybPiper (Johnson et al. 2016) and SECAPR (Andermann et al. 2018) are the most widely used and of enormous importance for the entire field of phylogenomics. However, they are designed to process one sample at a time, only making use of multiple computer cores at certain stages [e.g., when assembling sequences *de novo* with SPADES (Bankevich et al. 2012)] .To process multiple samples simultaneously, the users must rely on external tools such as GNU parallel (Tange 2021) or additional containerization tools (Jackson et al. 2023). This lack of native parallelization capabilities can lead to suboptimal utilization of the computing resources in high-performance clusters or workstations and reduce accessibility for non-experts, potentially causing repeatability issues due to command inconsistencies across samples or human error. Also, since these pipelines were optimized for target capture data, their processing times for other data types can be exceptionally long. In contrast, for restriction site associated DNA (RAD) data, user-friendly specialized analysis pipelines have been published which efficiently analyze many samples in

parallel, e.g., IPYRAD (Eaton and Overcast 2020), Stacks (Rochette et al. 2019), and DDOCENT (Puritz et al. 2014). For example, IPYRAD (Eaton and Overcast 2020) uses a single or just a few consistent commands, that, combined with relatively short processing times, allow for rapid testing of alternative settings. IPYRAD allows users to summarize results of each processing step so that they can decide on the most appropriate settings needed for following steps.

With these features in mind, we developed CAPTUS, a pipeline written entirely in Python and aimed at building phylogenomic datasets from multiple types of high-throughput sequencing (HTS) data (target capture, RNA-Seq, genome skimming, and high-depth whole genome sequencing) by making extensive use of Python's native parallel computing to process many samples simultaneously in a consistent manner. Each step has a simple basic command syntax (although it can be customized with many options), provides complete reports in HTML to guide the settings of the next step, and is fully logged for repeatability. CAPTUS is able to extract nuclear, mitochondrial, and plastid proteins as well as any other type of DNA regions (such as ribosomal RNA, introns, spacers, RAD loci, etc.) in a single command, unlike HYBPIPER which would need as many runs as marker types or SECAPR which can only take a reference composed of individual exons in nucleotides. Additionally, as a feature unique to CAPTUS, it can also be used to search for novel conserved markers by clustering contigs that received no hits from the reference target loci across samples, thus making use of data that otherwise would be ignored. During the alignment stage, CAPTUS can also filter paralogs by taking advantage of reference target files that contain multiple sequences per locus such as Angiosperms353 (Johnson et al. 2019) and its more taxonomically comprehensive version Mega353 (McLay et al. 2021). The final outputs of CAPTUS are multiple sequence alignments (MSAs) for each reference locus found in the samples. The MSAs are organized by genomic compartment and format (amino acid, nucleotide, etc.), and multiple versions of each are provided (i.e., unfiltered, paralog-filtered,

untrimmed, and trimmed) so the user can select which type of alignment to analyze for phylogenetic inference.

We demonstrate the potential of CAPTUS in a test case focusing on the flowering plant order Cucurbitales, a clade of eight plant families, which have their diversity centers in the Tropics: Anisophylleaceae, Apodanthaceae, Begoniaceae, Coriariaceae, Corynocarpaceae, Cucurbitaceae, Datiscaceae, and Tetramelaceae (Zhang et al. 2006). Together, they include 110 genera with more than 3,100 species, about 2,000 of them in the mega-diverse genus *Begonia* in Begoniaceae (Goodall-Copestake et al. 2009) and 1,000 in Cucurbitaceae (Stevens 2001; Schaefer 2020). The genus *Begonia* is of great horticultural importance while Cucurbitaceae include some of the most important vegetable and fruit crops worldwide, like cucumber and watermelon (Schaefer and Renner 2011a). Morphologically, the taxa of Cucurbitales are rather diverse ranging from the holoparasitic Apodanthaceae (Bellot and Renner 2014), to annual and perennial herbs in Datiscaceae and Begoniaceae, to trees and shrubs in Anisophylleaceae, Corynocarpaceae, Coriariaceae, and Tetramelaceae and finally to woody or herbaceous climbers and creepers in Cucurbitaceae (Schaefer and Renner 2011a; Schaefer 2020).

A number of studies addressed phylogenetic problems in Cucurbitales in the past two decades. Zhang et al. (2006) produced the first comprehensive phylogeny estimate for the order based on nine plastid, nuclear and mitochondrial loci. Schaefer and Renner (2011b) inferred phylogenetic relationships in the order based on 14 DNA regions from all three genomes. Other studies targeted individual families: Zhang et al. (2007) provided the first comprehensive phylogeny estimate for Anisophylleaceae, Kocyan et al. (2007) and Schaefer et al. (2009) for Cucurbitaceae, Goodall-Copestake et al. (2009) for Begoniaceae, and Renner et al. (2020) for Coriariaceae. The results of Filipowicz and Renner (2010) suggested that the holoparasitic Apodanthaceae are best placed in Cucurbitales as sister lineage to all other taxa. In recent years, several phylogenomic studies contributed to an even better understanding of

the evolutionary relationships in Cucurbitaceae. Bellot et al. (2020) analyzed entire plastomes plus a set of 57 single-copy nuclear genes and the ITS region for 29 species from all but one tribe and detected four nodes with conflicting phylogenetic signal. With an impressive set of 136 transcriptomes and full genomes of Cucurbitaceae, representing 52 of the 97 genera, Guo et al. (2020) produced a well-supported (albeit incomplete) phylogeny estimate which conflicted with earlier studies in several positions (Kocyan et al. 2007; Schaefer et al. 2009; Schaefer and Renner 2011b). Finally, the angiosperm-wide genus level analysis (Zuntini et al. in review) in the framework of the Plant and Fungal Trees of Life (PAFTOL) project (Baker et al. 2022) with an almost complete representation of Cucurbitales places the holoparasitic Apodanthaceae outside Cucurbitales in Malpighiales and finds a sister group relationship between Cucurbitales and Rosales, challenging the results of earlier studies.

In this study, we demonstrate how our new analysis pipeline CAPTUS can be used not only to extract the Angiosperms353 genes (Johnson et al. 2019), but also to derive a new set of thousands of nuclear genes from transcriptomic data and to extract entire plastomes. The resulting nuclear phylogenomic datasets were analyzed with coalescent and concatenation methods to infer a complete genus-level phylogeny for the Cucurbitales. We chose to test the performance and efficiency of CAPTUS against HYBPIPER since, in contrast to the other available pipelines, these two are able to use amino acid sequences as reference targets and can handle multiple reference target sequences per locus. Overall, we demonstrate that CAPTUS is more efficient and user-friendly than currently available pipelines and thus, a good choice for most data types and users with different levels of informatics skills.

MATERIALS & METHODS

Sampling

We sequenced a total of 125 Cucurbitales samples, 118 samples for target capture, and seven for high-depth whole genome sequencing (Table S1). Additionally, we downloaded 327

samples available in NCBI's SRA for a total of 240 RNA-Seq samples, 48 genome skimming samples (<50M reads), 31 high-depth whole genome sequencing samples (>50M reads), and eight additional target capture samples (Table S2). The entire dataset comprises 342 samples of Cucurbitales (including Apodanthaceae), representing the 110 currently accepted genera and 249 species. In order to confirm the results of the recent angiosperm-wide analysis of Zuntini et al. (in review), who found Apodanthaceae placed outside Cucurbitales, we decided to use a rather broad outgroup selection. We included 110 additional samples covering taxa of the nitrogen-fixing clade (representing 25 species in eight families of Rosales, 11 species in six families of Fagales, and 28 species in three families of Fabales), as well as 31 species in 23 families of Malpighiales, six species in six families of Malvales, three species of Huaceae (Oxalidales), and three species of Vitaceae (Vitales).

DNA Extraction and Sample Preparation

DNA samples were taken from the Royal Botanic Gardens, Kew DNA bank and the DNA Bank of Biodiversity of Plants, Technical University of Munich. Tissue samples for further DNA extractions came from the herbaria of Royal Botanic Gardens, Kew (K), Muséum National d'Histoire Naturelle in Paris (P), Botanische Staatssammlung München (M), and Technical University of Munich (TUM).

Several methods were used for DNA extraction from herbarium material intended for target capture. These included: (i) a CTAB-chloroform-based protocol with ethanol washes and a caesium chloride/ethidium bromide density gradient cleaning and dialysis, (ii) a modified CTAB protocol (Doyle and Doyle 1987) followed by an Agencourt AMPure XP bead clean-up (Beckman Coulter, Indianapolis, Indiana, USA), (iii) an SDS-based protocol using Magen HiPure SF Plant DNA Kit (Angen Biotech Co., Ltd, Guangzhou, China) using magnetic beads for extraction and cleaning, and (iv) the manufacturers protocol of the CTAB-based NucleoSpin Plant II Extraction Kit (MACHEREY-NAGEL GmbH & Co. KG, Düren,

Germany) using silica columns for binding the DNA. Depending on the available herbarium material, we used 20-160 mg of material for extraction.

The seven samples intended for high-depth whole genome sequencing were extracted with the following methods: (i) the standard protocol of the CTAB-based NucleoSpin Plant II Extraction Kit, and (ii) the standard protocol of the MagBind® Plant DNA plus 96 Kit from Omega (https://www.omegabiotek.com/product/mag-bind-plant-dna-plus-96-kit/). These seven DNA extractions were sent for library preparation and sequencing to GENEWIZ Germany GmbH (Leipzig, Germany).

To measure DNA quality, we evaluated fragment size distribution using an Agilent Technologies 4200 TapeStation System with Genomic DNA ScreenTapes (Agilent Technologies, Santa Clara, California, USA) or by electrophoresis in 1% agarose gel. DNA was quantified using a Quantus™ Fluorometer (Promega Corporation, Madison, WI, USA) or with a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Invitrogen, Walthan, Massachusetts, USA).

Library Preparation

Shearing was only performed for samples with fragment sizes above 350 bp using a Covaris M220 Focused-ultrasonicator with Covaris microTUBES AFA Fiber Pre-Slit Snap-Cap (Covaris, Woburn, Massachusetts, USA). Libraries were prepared with the DNA NEBNext UltraTM II Library Prep Kit, including end-repair/end-prep, NEBNext Adapter ligation, size selection of preferred DNA fragments with a length of 300 to 400 bp using SPRI beads (Agencourt AMPure XP Bead Clean-up; Beckman Coulter, Indianapolis, IN, USA) and amplification with NEBNext Multiplex Oligos for Illumina (Dual Index Primer Set 1 and Set 2, New England BioLabs, Ipswich, Massachusetts, USA) as indices for sample identification.

Pooling, Hybridization, Target Enrichment, and Sequencing

Libraries were grouped and pooled according to quality and quantity. Hybridization of the pooled libraries was done using the Arbor Biosciences myBaits Target Capture Kit, "Angiosperms353 v1" (Catalog #3081XX), following the manufacturers manual Version 4.01 (myBaits® Kit Manual – Arbor Biosciences 2020). Hybridized libraries were enriched with the NEBNext Ultra II Q5 Master Mix (New England BioLabs, Ipswich, Massachusetts, USA). According to their respective quantity and quality, library pools were normalized and pooled for sequencing on a MiSeq platform (Illumina, San Diego, California, USA) at the Jodrell Laboratory, Royal Botanic Gardens, Kew or at Macrogen Europe B.V. in Amsterdam (Macrogen, Inc., Seoul, Korea) with a HiSeq system (Illumina, San Diego, California, USA).

Sequence Analysis Workflow

CAPTUS was implemented in Python 3, is freely available and maintained through GitHub (https://github.com/edgardomortiz/Captus) and fully documented (https://edgardomortiz.github.io/captus.docs/). For convenience, it can be installed directly from Bioconda (https://bioconda.github.io/recipes/captus/README.html). The workflow of CAPTUS consists of four steps controlled by their respective modules called `clean`, `assemble`, `extract`, and `align`, which are typically run in that order (Fig. 1).

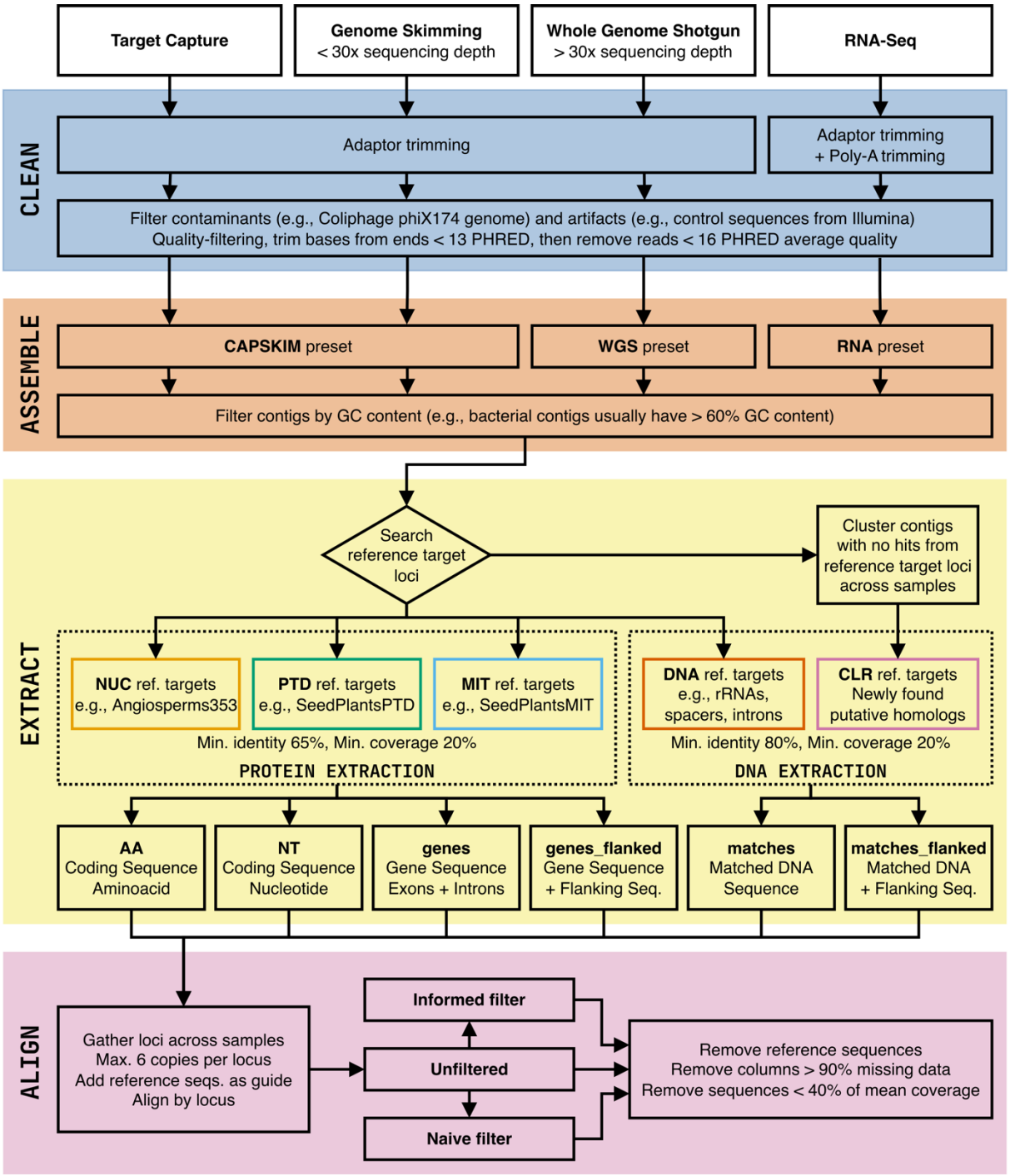


FIGURE 1. Workflow of the analysis pipeline CAPTUS. The default thresholds shown can be changed by the user via command options.

Alternatively, the analysis can be started at different points depending on the sample data provided, if raw reads are provided one should start with the `clean` module, if the reads have already been cleaned, they can be directly analyzed with the `assemble` module, and if

previously assembled data or reference genomes are provided the analysis can start using the `extract` module. CAPTUS provides flexibility in combining samples that enter the workflow at different stages. For example, a phylogenomic dataset could be composed of samples represented by target capture raw sequencing reads, samples with clean RNA-Seq reads, and genomic assemblies downloaded from GenBank in order to increase taxon sampling.

The `clean` *module.*--- CAPTUS uses BBDUK from BBTOOLS (Bushnell 2022) to clean raw reads. Cleaning is performed in two steps. First, adaptors and poly-A tails (when cleaning RNA-Seq reads) are trimmed in two consecutive rounds. Then, adaptor-free reads are quality-trimmed and filtered for common HTS contaminants. In this step, leading and trailing bases with PHRED scores <13 are trimmed and then trimmed reads with average PHRED score <16 are removed, both thresholds can be altered by the user.

When cleaning is completed, FASTQC (Andrews 2019) or FALCO (de Sena Brandine and Smith 2021) is run both on the raw and the cleaned files to evaluate and compile quality statistics. FALCO is a faster implementation of FASTQC that produces identical results (de Sena Brandine and Smith 2021). Finally, CAPTUS summarizes the cleaning statistics of all the samples before and after cleaning in a single HTML report that allows a quick overlook of all quality measurements.

For the Cucurbitales analysis, default settings were used for all samples in the `clean` module except RNA-Seq, for which the option `--rna` was added to trim poly-A tails.

The `assemble` *module.*--- The clean reads are passed to the assembly module which uses MEGAHIT (Li et al. 2015) for *de novo* assembly. MEGAHIT was designed to handle enormous metagenomic datasets and therefore its default settings are tuned to that purpose. We tested several combinations of MEGAHIT settings in order to optimize the assembly of other types of data, these are provided in CAPTUS under three presets: (i) `CAPSKIM` for target

sequence capture, genome skimming data, or a combination of both, (ii) `RNA` for RNA-Seq data, and (iii) `WGS` for high-depth whole genome sequencing data.

The `assemble` module also provides an option to subsample a fixed number of reads per sample prior to assembly, which is useful when the sequencing depth is too high for a particular sample or for speeding up assembly during trial runs.

Once assembly is completed, CAPTUS can also filter contigs based on their GC content. For example, a maximum of 60% GC would be appropriate to remove bacterial contamination in most eukaryotic genomes (Fierst and Murdock 2017). Finally, CAPTUS computes assembly statistics and produces a single HTML report for all processed samples.

Target capture and genome skimming samples of the Cucurbitales were run with the default assembly preset (`--preset CAPSKIM`), RNA-Seq samples were assembled using `--preset RNA`, and high-depth whole genome sequencing samples using `--preset WGS`. To remove potential bacterial contamination, we used the option `--max_contig_gc 60` for all assemblies.

The `extract` module.--- Once assemblies are completed, they can be searched for particular loci of interest (i.e., reference target loci). CAPTUS allows the simultaneous search and extraction of five types of markers (i) nuclear proteins (`NUC`), (ii) plastid proteins (`PTD`), (iii) mitochondrial proteins (`MIT`), (iv) miscellaneous DNA regions (`DNA`), and (v) new clustering-derived putative homologs (`CLR`).

Protein extractions (`NUC`, `PTD`, and `MIT` markers) use reference target loci provided as amino acid sequences or coding sequences in nucleotides (CDS). CAPTUS uses the program SCIPIO (Hatje et al. 2011) to perform protein extractions. SCIPIO is able to automatically correct frameshifts that could result from sequencing or assembly error and it can recover proteins that are spread across several contigs (Hatje et al. 2011), an important advantage

given that most assemblies resulting from target capture data are highly fragmented. CAPTUS comes bundled with four reference target sets, two `NUC` references (the Angiosperms353 (Johnson et al. 2019) target file and a curated version of the taxonomically expanded Mega353 (McLay et al. 2021) target file), as well as a plastid `PTD` reference (SeedPlantsPTD), and a mitochondrial `MIT` reference (SeedPlantsMIT).

Our curation of the Mega353 reference target file consisted in removing contaminated sequences (e.g., fungal, algal, organellar sequences) and then clustering at 78% identity to reduce sequence redundancy. Thus, CAPTUS still takes advantage of the expanded taxonomic representation of the Mega353 reference targets (McLay et al. 2021) while bypassing the additional step of filtering by taxon suggested by the instructions (https://github.com/chrisjackson-pellicle/NewTargets). The `PTD` and `MIT` references represent carefully curated sets of proteins found in all organellar genomes of seed plants originally downloaded from GenBank. Reference target sets for organism groups other than seed plants are being developed.

Miscellaneous DNA extractions are aimed at recovering other types of `DNA` regions (e.g., complete genes with introns, non-coding regions, ribosomal genes, individual exons, RAD markers, etc.). In this case, CAPTUS uses the program BLAT (Kent 2002) to search the assembly and the hits that are found across contigs are greedily assembled and concatenated using our own code.

Additionally, CAPTUS can be used to find new putative homologs by clustering across samples the contigs that had no hits from the reference target loci. In this case, the program MMSEQS2 (Steinegger and Söding 2018) is used, and several of its settings are available through CAPTUS. Once the sequence clusters have been found, they are filtered by the number of samples in the cluster and sequence length, and only the most represented sequences per cluster are selected as new reference targets to perform another extraction of the

283 miscellaneous `DNA` type in CAPTUS. The output from this kind of extraction is indicated by
284 the prefix `CLR`.

285 The output files from protein extractions (`NUC`, `PTD`, `MIT`) are provided in four
286 possible formats (Fig. S1a): the protein sequence in amino acids (`AA`), the coding sequence in
287 nucleotides or CDS (`NT`), the complete gene sequence including introns (`genes`), and the
288 complete gene sequence flanked by a fixed number of nucleotides (`genes_flanked`).
289 Similarly, the output from a miscellaneous DNA extraction and clustering-derived markers
290 (`DNA`, `CLR`), have two possible formats (Fig. S1b): the segment of sequence that was matched
291 to the reference (`matches`), and the matched segments flanked by a fixed number of
292 nucleotides (`matches_flanked`).

293 Once extractions are finished, CAPTUS collects extraction statistics (e.g., recovered
294 marker length, similarity to the reference sequence, number of paralogs, number of contigs
295 used in the gene assembly, etc.) from all the processed samples and produces a single HTML
296 report that provides a quick look at the marker recovery across markers and samples using a
297 dynamic heatmap.

298 For our Cucurbitales dataset, the minimum recovery percentage was set at 20%, the
299 default in CAPTUS. We used the option `-n Angiosperms353` to extract nuclear markers
300 using the original Angiosperms353 reference targets and `-n Mega353` to extract nuclear
301 markers using our curated version of the taxonomically expanded Mega353. Additionally,
302 organellar proteins were extracted with options `-p SeedPlantsPTD` and `-m`
303 `SeedPlantsMIT`. Finally, we created a custom miscellaneous DNA reference by
304 segmenting the Cucurbitales plastomes available in GenBank in 38 pieces ranging from ~3
305 kbp to ~5 kbp which we stored in a FASTA file (`Plastome38.fasta`, Appendix S1) and
306 extracted in CAPTUS using the option `-d`.

Additionally, we created a set of 5,435 of reference genes derived from publicly available transcriptomic data. We took the 240 transcriptomic assemblies from CAPTUS and removed the transcripts that had hits to organellar proteins, thereby retaining only putatively nuclear transcripts. We ran CODAN (Nachtigall et al. 2021) on the nuclear transcripts in order to find coding regions within them, searching both strands and using the `PLANTS_full` model which only emits a CDS when a complete protein, from start to stop codon, is identified within the transcript. Then we clustered the coding sequences across samples using the `extract` module of CAPTUS, with the following options: `-c --mmseqs2_method easy-cluster --cluster_mode 2 --cl_min_identity 70 --cl_seq_id_mode 1 --cl_min_coverage 66 --cl_rep_min_len 540 --cl_min_samples 144 --cl_max_copies 4`. These CAPTUS options retained 5,435 clusters that were at least 540 bp in length, grouped at least 144 transcriptomic samples (60%), and contained at most an average of four gene copies. After removing within-cluster redundant sequences, the resulting reference contained 13,492 sequences representing these 5,435 CDS clusters, this reference is called from this point onwards RNA5435 (`RNA5435.fasta`, Appendix S2). The new reference targets were used in the `extract` module of CAPTUS across all 454 samples using the option `-n RNA5435.fasta --nuc_min_score 0.15 --nuc_min_identity 70` to match the clustering identity threshold used for creating the reference.

The `align` *module.*--- The extraction output is then processed by the alignment module. Individual markers are collected across samples and organized in separate FASTA files per locus. By default, CAPTUS collects a maximum of five paralogs per sample and per marker to be aligned. The reference sequences are also added to each locus file to serve as alignment guides in case the sequences recovered from the samples are fragmentary. Then, the alignment is performed with MAFFT (Katoh and Standley 2013) or MUSCLE5 (Edgar 2022)

using default settings or by selecting one of their specific algorithms through CAPTUS. However, if protein sequence in amino acid (`AA`) and their corresponding coding sequence in nucleotides (`NT`) are aligned in the same run, CAPTUS aligns the `AA` with MAFFT and then uses the `AA` alignment as template for the `NT`, thus producing a codon-aware alignment for the CDS. CAPTUS also allows the user to provide the sample(s) that should be considered as outgroup, in this case the program will place those samples as the first sequence(s) in the alignments in the order provided. This feature (`--outgroup`) takes advantage of a common feature of many phylogenetic estimation programs (e.g., IQ-TREE, RAXML, MRBAYES) which arbitrarily draw the first sample in the alignment at the root of their output trees.

Once the FASTA files are aligned, paralogs are filtered using two alternative algorithms, `naive` and `informed`. In the `naive` filter, only the best hit is kept for each sample and no further filtering is performed. The `informed` filtering algorithm takes advantage of reference datasets that contain multiple sequences per locus, such as the Angiosperms353 (Johnson et al. 2019) or Mega353 (McLay et al. 2021) target files as well as the ones developed for CAPTUS (SeedPlantsPTD and SeedPlantsMIT). For example, the Angiosperms353 reference target set (as well as its expansion Mega353) was derived from the 1KP Project (Matasci et al. 2014; One Thousand Plant Transcriptomes Initiative 2019) data, where each gene could potentially be present in more than 1,000 species. To build a less redundant sequence collection while still covering the entire phylogenetic diversity of the angiosperms, each selected locus was clustered at ~70% identity, usually selecting only one representative sequence per cluster for the final reference dataset. Thanks to this feature, one could expect that when analyzing a single taxonomic group such as a family or genus, all samples would have a best match to only one sequence in the reference targets collection for a given locus (i.e., the closest relative present in the reference targets). However, a few samples could have better matches to a different reference sequence than the rest which can be possible when the locus in question has multiple copies (paralogs) in the studied group, and

the most common copy found in the group is absent from those few samples. In these cases, CAPTUS compares all recovered paralogs with the reference sequence that best matches most of the samples and keeps the copy that is most similar to that reference sequence across samples. Finally, it could also happen that a sample presents a single copy for a specific locus but with a sequence that is much more divergent than the average in the alignment (e.g., a remote paralog found in a contaminated contig). By default, CAPTUS will remove any sequence with average pairwise identity that is more than 4.0 standard deviations below the mean pairwise identity of the entire alignment, the number of standard deviations can be changed with the option `--tolerance`.

Recently developed coalescent methods, such as ASTRAL-PRO (Zhang et al. 2020; Zhang and Mirarab 2022a), are capable of analyzing trees of genes with paralogs (i.e., multi-copy genes), therefore CAPTUS also provides the alignments with paralogs as well as the file required by ASTRAL-PRO for mapping the paralog names to the names of the samples for species tree calculation. The reference target sequences are then removed from the alignments. Finally, all the produced alignments and their filtered versions are trimmed using CLIPKIT (Steenwyk et al. 2020), which can remove alignment columns based on criteria like informativeness or proportion of missing data. By default, CAPTUS removes columns with > 90% missing data and sequences with < 40% mean coverage. Alternatively, a minimum number of sites per column instead of a percentage can be specified with `--min_data_per_column`. As in previous modules, CAPTUS computes alignment statistics as well as sample occupancy statistics from all the FASTA files along all filtering and trimming stages and produces a comprehensive HTML report that can be used to determine outlier markers or samples that should be removed or curated more carefully before proceeding to phylogenetic analysis.

For the Cucurbitales data, all extracted loci were aligned in CAPTUS using MAFFT's most accurate algorithm E-INS-i (`--align_method mafft_genafpair`). For coding

markers we used the option `-f AA,NT` to take advantage of codon-aware alignments for CDS, and for Plastome38 we used `-f MA`. For trimming, we used `--min_data_per_column 6`, to keep alignment columns with six or more sequences. For the Plastome38 we also decreased the tolerance of the `informed` paralog filter to `--tolerance 2.0`.

Phylogenetic Analyses

We chose the trimmed, codon-aligned coding sequences (format `NT`) from the extractions performed with reference targets sets Angiosperms353, Mega353 and RNA5435. For the reference targets set Plastome38 (the plastome segments) we selected the format `matches` from the CAPTUS output. We analyzed the unfiltered alignments (i.e., including paralogs) as well as the alignments resulting from the `naive` and `informed` paralog filtering strategies for nuclear markers but only the alignments filtered by the `informed` algorithm for the plastome.

Gene tree estimation.--- For each individual nuclear gene alignment we inferred a phylogeny using IQ-TREE v. 2.2.2.6 (Minh et al. 2020b). During the run we used MODELFINDER (Kalyaanamoorthy et al. 2017) to determine the most appropriate nucleotide substitution model with `-m TEST`. Nodal support was inferred from 1,000 ultrafast bootstrap replicates with option `-bb 1000` (Hoang et al. 2018). For comparison purposes, we also estimated nuclear gene tree phylogenies using FASTTREE v. 2.1.11 (Price et al. 2010) with the manual recommended options to increase the accuracy of the search and using the GTR substitution model (`-pseudo -spr 6 -mlacc 3 -slownni -gtr`).

Species tree estimation.--- We estimated species trees using two methods, concatenation of alignments followed by maximum likelihood estimation in IQ-TREE v. 2.2.2.6 (Minh et al. 2020b), and coalescent estimation by summarization of quartet frequencies in gene trees using ASTRAL-PRO v. 1.15.1.3 (Zhang et al. 2020; Zhang and Mirarab 2022a). For concatenation, each individual locus alignment must contain a single sequence per sample (no paralogs allowed), therefore we can only apply this method to the alignments that were filtered for paralogs, while ASTRAL-PRO was applied to the filtered alignments as well as to the alignments containing multiple copies. The concatenation analyses in IQ-TREE were run with the same options as for the individual gene trees, and the loci alignments were provided as separate files in a single directory with the option `-p`, so IQ-TREE can automatically concatenate them into a supermatrix prior to analysis. The concatenation method was applied only to the set 353 nuclear genes (extracted using Angiosperm353 or Mega353) and to the plastome segments (Plastome38)

For the ASTRAL-PRO analysis, the individual gene trees calculated by IQ-TREE or FASTTREE were provided as well as the required file that maps the paralog names to their corresponding sample name produced by CAPTUS (`captus-assembly_align.astral-pro.tsv`). We also increased the number of placement and subsampling rounds to 16 (`-R`), as well as the proportion of taxa subsampled (`--proportion 0.75`) and calculated alternative quartet frequencies (`-u 3`).

Site concordance analyses.--- In order to measure concordance between alignment sites and the species tree, we performed a concordance factor analysis for each species tree (Minh et al. 2020a). The analysis was done in IQ-TREE v. 2.2.2.6 by supplying the species tree (`-t`), the folder containing the loci alignments (`-p`), and averaging the site concordance over 1,000 quartets (`--scfl 1000`).

Phylogenetic conflict analysis of selected nodes.--- Contentious relationships in the Cucurbitales are centered around the relationships among tribes in the Cucurbitaceae as well as the relationships among Cucurbitales families. In order to visualize the amount of conflict at selected nodes in the species tree, we used the branch quartets frequencies analysis in DISCOVISTA v. 1.0 (Sayyari et al. 2018), after annotating the species belonging to each Cucurbitaceae tribe, to each of the families in Cucurbitales, and to the clades in the outgroup.

Phylogenetic network estimation.--- We concatenated the Angiosperms353 alignments filtered by the `informed` method in CAPTUS keeping only the Cucurbitales families. This concatenated alignment was used as input for SPLITSTREE v. 4.18.2 (Huson and Bryant 2006) in order to estimate a phylogenetic network using the Neighbor Net algorithm on a matrix of uncorrected P distances.

Pipeline Comparison

To compare the locus recovery efficiency between CAPTUS and HYBPIPER (Johnson et al. 2016) on different data types, we extracted the Angiosperms353 loci (Johnson et al. 2019) from a selection of 80 samples (20 from each data type) representing all families of Cucurbitales and all tribes of Cucurbitaceae (Table S3). Since the HYBPIPER workflow does not include a read cleaning step, we used reads cleaned by CAPTUS as a common input for both pipelines. CAPTUS was run for the `assemble` and `extract` modules with the aforementioned presets optimized for each data type. HYBPIPER v2.1.2 was run using either BLASTX (Camacho et al. 2009) or DIAMOND (Buchfink et al. 2021), hereafter HYBPIPER-BLASTx and HYBPIPER-DIAMOND respectively, with an identity threshold of 65% (`--thresh 65`) to equalize with that of CAPTUS. A protein target file downloaded from https://github.com/mossmatters/Angiosperms353 was used as a reference for both pipelines. All commands were run on a MacOS X system equipped with a 2.7 GHz Intel Xeon E5

processor with 24 threads and 64 GB RAM, allocating 6 threads per sample. Running time, number of loci recovered, and total CDS length recovered for each sample were compared across pipelines, considering a locus to be "recovered" when at least 20% of the reference sequence length was retrieved.

To assess how the differences in locus recovery efficiency among pipelines affect phylogenetic inference, we estimated coalescent species trees from the CDS recovered by each pipeline and compared their topologies and nodal supports. For CAPTUS, trimmed CDS alignments with paralogs removed by the `informed` filtering were generated using the `align` module with default settings. For HYBPIPER-BLASTx and HYBPIPER-DIAMOND, trimmed CDS alignments were generated using MAFFT v7.520 (Katoh and Standley 2013) with the automatic strategy selection mode (`--auto`) and CLIPKIT v1.4.1 (Steenwyk et al. 2020) with default settings. A maximum likelihood tree was estimated for each locus using IQ-TREE v2.2.2.3 (Minh et al. 2020b) under the best-fit substitution model determined by MODELFINDER (Kalyaanamoorthy et al. 2017) with nodal support inferred from 1,000 ultrafast bootstrap replicates (Hoang et al. 2018). Coalescent species trees were estimated from the set of gene trees using WASTRAL-HYBRID v1.15.2.3 (Zhang and Mirarab 2022b) with 16 rounds each for placements and subsampling (`-R`), and then visualized using TOYTREE v2.0.5 (Eaton 2020).

RESULTS

Sequencing

From our 118 library preparations for target capture, 92 (78%) worked at the first attempt. Only eight of these successful libraries needed post-processing steps such as library concentration or increased number of PCR cycles to improve their quality prior to sequencing. For the remaining 26 libraries (22%) that had to be repeated, we used material of the same herbarium voucher for 14, while we had to select a different specimen for 12. Among the

repeated libraries, only two needed post-processing. Nonetheless, once the sequencing was completed, three samples had to be discarded due to insufficient data. Two of the discarded samples corresponded to replicated libraries of *Octomeles sumatrana* for which we had a replacement within the target capture batch. Only one of them, *Bambekea racemosa*, needed to be replaced by additional high-depth whole genome sequencing, which was performed for a total of seven additional samples. In the end, a total of 122 samples (115 target capture and 7 high-depth whole genome sequencing) yielded enough high-quality data to be used for the analyses (Table S4).

Read Cleaning and Assembly

RNA-Seq data had the largest average percentage of reads and percentage of base pairs removed with 4.7% and 9.2% respectively. The data type that had the largest average percentage of raw reads with adaptors was target capture with 14.6%, adaptors were fully removed from all data types after cleaning (Table 1: Cleaning). Average assembly size after removing contigs with GC content > 60% was similar for target capture and RNA-Seq data (c. 44 Mbp), followed by genome skimming (289.3 Mbp) and high-depth whole genome sequencing (591.1 Mbp). Despite their similar assembly sizes, target capture data produced more fragmented assemblies than RNA-Seq as indicated by their average number of contigs (88.3 k vs. 62.7 k respectively) and their average N50 (513.3 bp vs. 1084 bp respectively), while the largest average N50 corresponded to high-depth whole genome sequencing data with 5023.7 bp. The GC content across data types was similar and ranged between 34.8% to 42.6% (Table 1: Assembly).

The largest average percentage of contigs removed because of exceeding the threshold of 60% GC content belongs to target capture data with 4%. The average GC content of the filtered contigs ranged from 59.8% to 65.4% across data types (Table 1: Assembly).

513 TABLE 1. Summary statistics by data type.

	CAP N = 126		GSK N = 48		WGS N = 38		RNA N = 240	
	Mean	SD	**Mean**	SD	**Mean**	SD	**Mean**	SD
Raw reads (Millions)	**4.7**	2.6	**17.3**	12.1	**113.5**	99.4	**44.1**	33.2
Raw bases (Gbp)	**0.7**	0.4	**2.5**	1.8	**15.9**	12.2	**5.9**	4.4
Cleaning								
Removed reads (%)	**2.3**	2.7	**1.3**	2.3	**2.0**	3.7	**4.7**	10.3
Removed bases (%)	**8.4**	8.9	**4.2**	6.7	**4.3**	8.7	**9.2**	13.1
Raw reads with adaptors (%)	**14.6**	22.0	**5.6**	12.6	**1.8**	6.8	**8.5**	11.3
Clean reads with adaptors (%)	**0.0**	0.0	**0.0**	0.0	**0.0**	0.0	**0.0**	0.0
Assembly								
Contigs (Thousands)	**88.3**	61.1	**489.0**	261.4	**823.7**	654.3	**62.7**	37.9
Assembly size (Mbp)	**44.4**	31.7	**289.3**	162.2	**591.1**	256.3	**44.7**	23.8
N50 (bp)	**513.3**	116.1	**923.3**	964.8	**5023.7**	4384.0	**1084.0**	289.5
GC content (%)	**40.8**	3.6	**37.5**	2.7	**34.8**	1.5	**42.6**	2.1
Filtered contigs (Thousands)	**1.3**	3.1	**6.2**	20.1	**16.2**	49.1	**1.8**	6.1
Filtered contigs (%)	**4.0**	21.0	**1.8**	7.5	**1.5**	3.6	**1.9**	3.4
Filtered contigs GC content (%)	**59.8**	14.6	**64.1**	1.9	**65.4**	2.4	**62.9**	1.0
Extraction								
Angiosperms353								
Percentage of loci recov. (%)	**87.9**	19.3	**70.9**	30.7	**98.8**	4.1	**92.6**	14.0
Total CDS length recov. (kbp)	**183.8**	57.5	**137.2**	89.0	**255.9**	28.5	**229.2**	53.3
Mega353								
Percentage of loci recov. (%)	**88.0**	19.3	**71.3**	30.6	**98.9**	3.9	**92.7**	14.0
Total CDS length recov. (kbp)	**190.2**	60.6	**145.6**	95.1	**271.3**	30.1	**241.8**	57.1
RNA5435								
Percentage of loci recov. (%)	**20.8**	14.9	**61.2**	30.5	**97.4**	8.4	**90.5**	14.8
Total CDS length recov. (kbp)	**808.6**	713.6	**3964.6**	2757.1	**8207.7**	1136.5	**6968.8**	1832.1
SeedPlantsPTD								
Percentage of loci recov. (%)	**78.6**	24.9	**94.9**	5.3	**96.3**	0.7	**76.2**	20.6
Total CDS length recov. (kbp)	**45.9**	20.8	**64.8**	6.7	**67.0**	1.2	**39.9**	16.9
SeedPlantsMIT								
Percentage of loci recov. (%)	**73.9**	26.5	**92.4**	10.9	**93.7**	3.9	**63.8**	21.9
Total CDS length recov. (kbp)	**22.0**	10.6	**31.8**	4.1	**32.2**	0.9	**18.0**	8.2
Plastome38								
Percentage of loci recov. (%)	**88.6**	25.6	**99.0**	5.3	**100.0**	0.0	**81.3**	20.6
Total length recov. (kbp)	**145.3**	63.4	**208.3**	30.6	**211.7**	7.3	**106.6**	45.5

514

515

516

Extraction and Alignment

Since the Angiosperms353 and the Mega353 reference targets aim to recover the same set of 353 genes, gene recovery in terms of number of loci and total CDS length was very similar for both. The average percentage of genes recovered across data types varied from c. 71% for genome skimming to c. 99% for high-depth whole genome sequencing data while the data type with the longest total CDS lengths is high-depth whole genome sequencing with an average of 271.3 kbp and the shortest is genome skimming with an average of 145.6 kbp. However, the Mega353 reference targets recovered at most 0.4% more loci than Angiosperms353 for genome skimming and at most 0.1% more for other data types. Similarly, the Mega353 reference targets produced total CDS lengths only c. 10 kbp longer than the Angiosperms353 reference targets in average across data types.

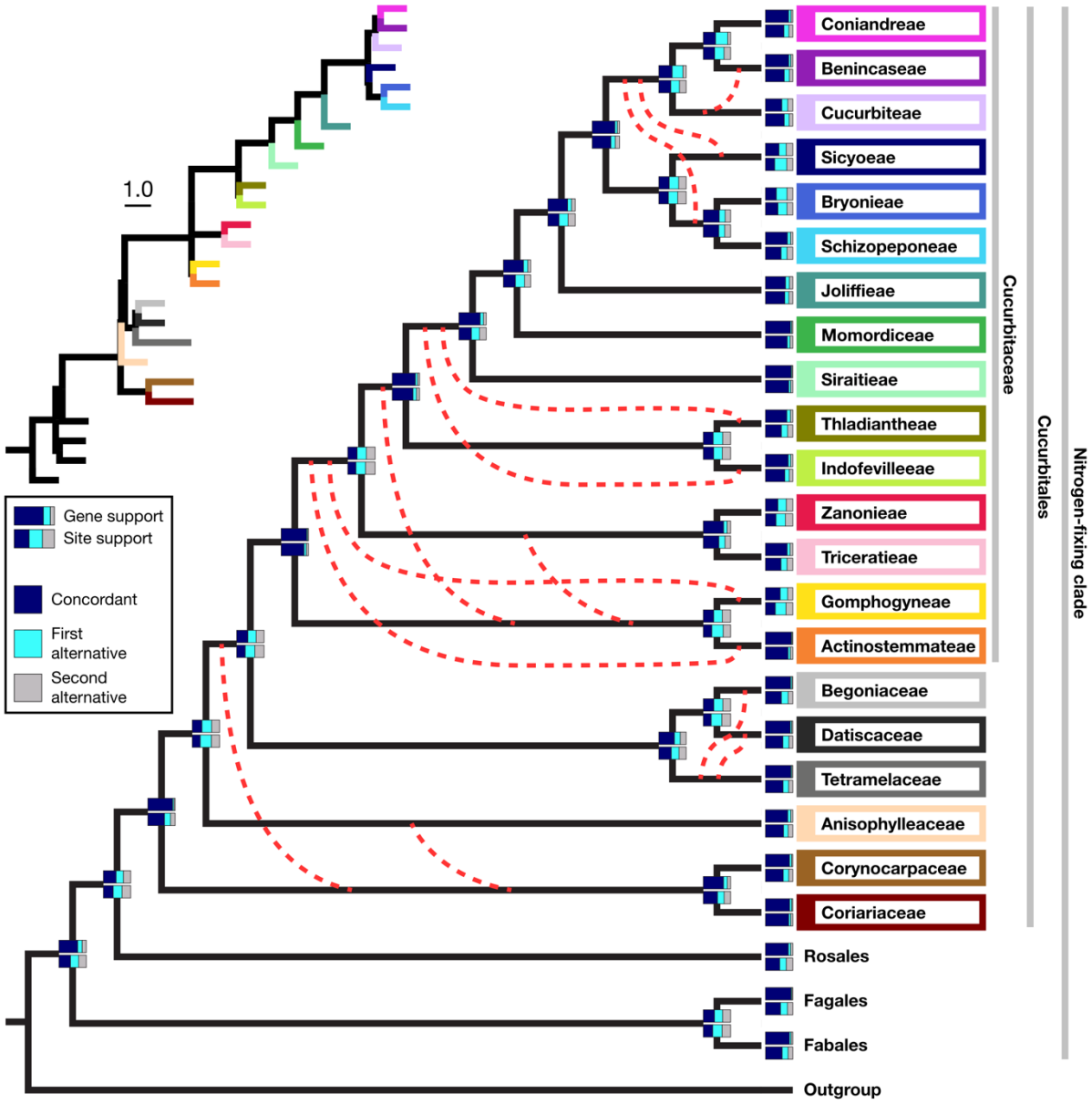
For the RNA5435 reference targets, target capture data had the lowest average percentage of recovered loci (20.8%, SD 14.9%) as well as the shortest average total CDS length (808.6 kbp, SD 713.6 kbp). For the rest of data types, the percentage of RNA5435 genes recovered ranges from 61.2% to 97.4% while the average total CDS length ranges from 3.96 Mbp to 8.2 Mbp. Recovery of organellar proteins was also high across data types, exceeding 73% except for mitochondrial proteins from RNA-Seq data where only 63.8% of the genes were recovered. The 38 plastome segments were successfully recovered across data types, where the minimum was 81.3% for RNA-Seq data (Table 1: Extraction). The total aligned ungapped length across samples was similar for Angiosperms353 (Fig. S2a) and Mega353 alignments (Fig. S2b) with c. 85% of samples with lengths between c. 200 kbps and c. 290 kbps and only c. 2.5% of samples showing less than 50 kbps aligned. As for RNA5435 (Fig. S2c) around 55% of samples had total ungapped aligned lengths between 6 and 8.2 Mbp, while around 20% of the samples had less than 1 Mbp aligned, corresponding mostly to target capture samples.

Based on a preliminary extraction report from CAPTUS which indicated an unusually high number of copies per DNA region in *Lemurosicyos variegata* (Cucurbitaceae), we were able to determine that this sample was contaminated with a plant species from another family. Given the evolutionary distance of the contaminant we could decontaminate it using CAPTUS and a custom reference specific for Cucurbitaceae (Supplementary Method). We also identified the Cucurbitaceae sample *Xerosicyos perrieri* as cross-contaminated with *Ibervillea sonorae*, another Cucurbitaceae. In this case, decontamination would have implied creating a *Xerosicyos*-specific reference, which was not possible, and therefore we decided to exclude the *Xerosicyos perrieri* data from further analyses. A preliminary plastome phylogeny revealed additional samples with chloroplast contamination or lack of sufficient data for correct placement (Table S5).

Phylogeny estimation

Coalescent species trees estimated from the set of 353 genes (using Angiosperms353 or Mega353 reference targets sets) or the RNA5435 gene trees, regardless if inferred from paralog-filtered or unfiltered alignments, or if estimated with FASTTREE or IQ-TREE, resulted in topologies with monophyletic and well supported families in Cucurbitales and monophyletic and well-supported tribes in Cucurbitaceae (Fig. 2, Fig. S3, Appendix S3) as well as in species trees calculated with the concatenation method for nuclear genes (Appendix S3) or for the plastome (Fig. S5). The holoparasitic Apodanthaceae are sister to Rafflesiaceae in Malpighiales in all inferred trees (Fig. S3, Appendix S3), so Cucurbitales are monophyletic only after exclusion of Apodanthaceae. For the remaining Cucurbitales families, the relationships vary, particularly around the nodes with ambiguous quartet support surrounding short branches (see inset in Fig. 2). However, the quartet frequency analyses from DISCOVISTA reveal that the apparently contradicting relationships among the major clades are all compatible with the reticulations depicted as dotted red lines in Fig. 2 because the different

568 species tree configurations observed coincide with the alternative quartet configurations
569 suggested for the nodes in conflict (Fig. S4). The best supported alternative shows a grade,
570 where the clade Coriariaceae + Corynocarpaceae is followed by Anisophylleaceae and by the
571 remaining families, but there is also considerable gene tree and site support for a clade in
572 which Anisophylleaceae is sister to a clade comprising Corynocarpaceae + Coriariaceae and
573 the remainder of the order (Fig. 2). The woody Tetramelaceae is resolved as sister to the two
574 herbaceous families Begoniaceae and Datiscaceae but again there is also considerable gene
575 support for the two alternative combinations within the triplet (Fig. 2). Within Cucurbitaceae,
576 there is also gene support for different relationships between the tribes, but the best supported
577 topology is a clade comprising Actinostemmateae and Gomphogyneae, found as sister to the
578 remaining tribes. Among the genera with multiple samples, only *Kedrostis* is found to be
579 polyphyletic (Fig. S3, Appendix S3, Fig. S5). The extinct Cambodian species *Khmeriosicyos*
580 *harmandii*, known only from the type collection, is confidently placed in Benincaseae as
581 sister to *Borneosicyos* from Mount Kinabalu in the nuclear species trees (Fig. S3, Appendix
582 S3) and in a clade with *Borneosicyos* and the southeast Asian *Solena* in the plastome tree
583 (Fig. S5). The network analysis revealed a deep reticulation within and between Benincaseae
584 and Cucurbiteae and betweeen Schizopeponeae and Sicyoeae. There is also evidence for more
585 recent reticulation within Coniandreae and very recent within Thladiantheae (Fig. 3). The
586 species trees resulting from the concatenated analysis of the complete plastomes (Fig. S5)
587 agree with the reticulated representation of the Cucurbitales topology (Fig. 2, Fig. 3).



588

589 FIGURE 2. Coalescent cladogram of the Cucurbitales inferred with ASTRAL-PRO based on the
590 RNA5435 nuclear gene trees (Fig. S4a) with families and Cucurbitaceae tribes collapsed into
591 a single branch each. Gene support calculated with ASTRAL-PRO and site support calculated
592 using IQ-TREE is indicated at each branch, the most important alternative topologies are
593 shown with red broken lines. Branch lengths in the inset phylogram are in coalescent units.

594

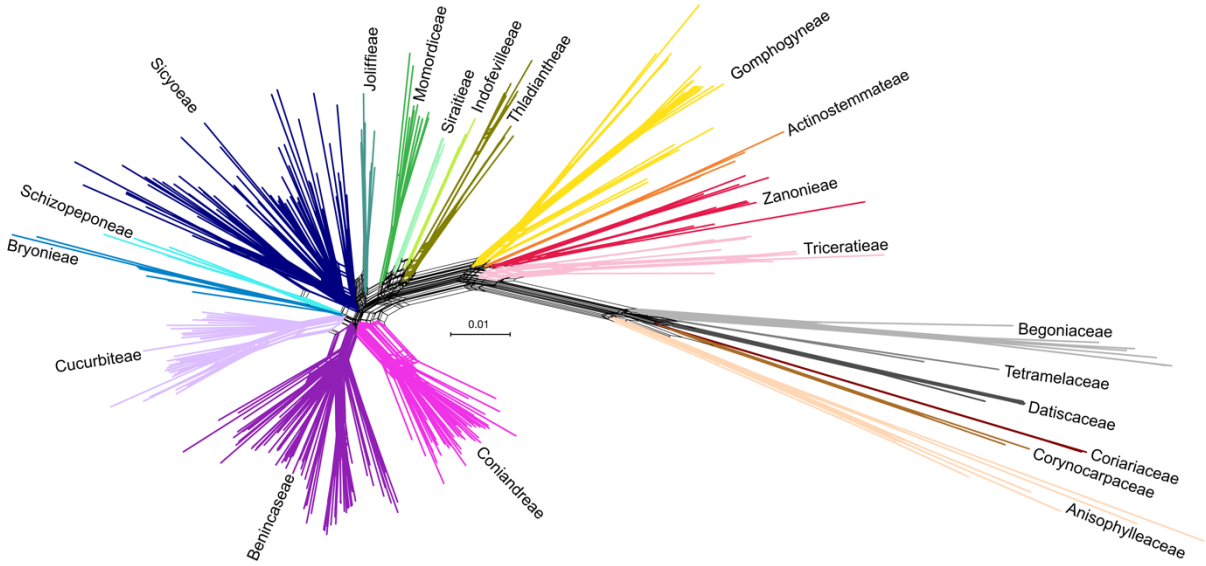


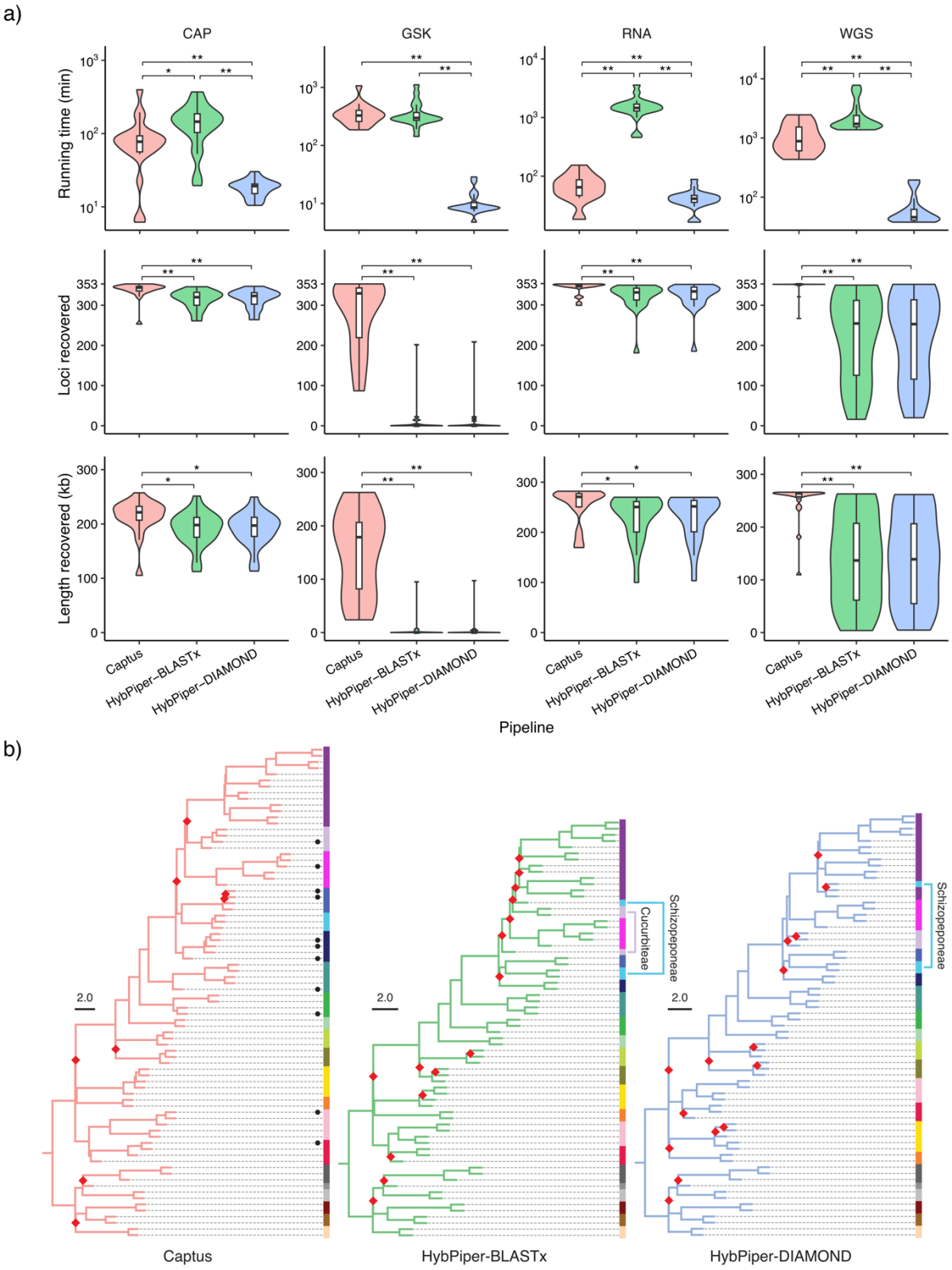
FIGURE 3. Phylogenetic network estimate for the Cucurbitales calculated with SPLITSTREE using the Neighbor Net algorithm on the concatenated alignment of the Angiosperms353 alignments which were filtered using the informed filter of CAPTUS.

Pipeline Comparison

In comparison to the two methods offered by HYBPIPER, CAPTUS consistently recovers a larger number of more complete genes regardless of data type, showing significant to highly significant differences in each case (Fig. 4a). Gene recovery is more similar for target capture data, however HYBPIPER could only recover a few gene fragments from genome skimming data. Regarding processing times, HYBPIPER-DIAMOND was faster than CAPTUS for every data type while keeping its gene recovery statistics essentially identical to HYBPIPER-BLASTx which was the slowest of the methods compared (Fig. 4a). HYBPIPER was not able to produce useful data for several genome skimming samples, which therefore could not be included in their respective species trees (Fig. 4b). The few genome skimming samples for which HYBPIPER produced data resulted in two polyphyletic groups, despite being consistently recovered as the monophyletic tribes Schizopeponeae and Cucurbiteae with CAPTUS for the reduced (Fig. 4b) and the full Cucurbitales species trees (Fig. 2, Fig. S3,

613 Appendix S3, Fig. S5). The number of nodes with low support is also greater in the

614 phylogenies derived from HybPiper data than in the Captus tree (Fig. 4b).



615

616 Figure 4. Benchmarking of Captus against HybPiper-BLASTx and HybPiper-DIAMOND

617 using a selection of 80 samples belonging to four different data types. **a)** Comparison of locus

recovery efficiency between pipelines for each data type, evaluating the running time required for the assembly and extraction of the Angiosperms353 loci (top), the number of loci recovered (middle), and the total length of coding sequences recovered (bottom). Statistical testing was performed using a Wilcoxon rank-sum test with Bonferroni correction: * $p < 0.05$, ** $p < 0.01$. **b)** Coalescent-based species trees resulting from each pipeline. Vertical bar along each tree color-codes the different tribes/families in the Cucurbitales. Red diamonds indicate conflicting or poorly supported nodes with a local posterior probability below 0.9. Samples recovered only by Captus are marked with a black dot. Two tribes inferred as polyphyletic groups in the HybPiper analyses are indicated with name and position. Tree files in Newick format are provided in Appendix S4.

DISCUSSION

The pipeline comparison shows that CAPTUS is slower than HYBPIPER-DIAMOND but consistently recovers a higher number of more complete genes across a larger number of species than either HYBPIPER-BLASTX or HYBPIPER-DIAMOND (Fig 4a). This can be explained by a combination of factors. First, the efficient and thorough *de novo* assembly by MEGAHIT allows the assembly of *all* reads in the sample in contrast to HYBPIPER which only assembles prefiltered reads that match the reference loci, which leads to a more restricted assembly potentially missing intronic regions and small exons. Second, SCIPIO's capacity to reconstruct gene models across several contigs is essential to deal with fragmentary assemblies such as the ones resulting from target capture data and particularly genome skimming (Hatje et al. 2011). Finally, SCIPIO also outperforms (Hatje et al. 2011) EXONERATE's (Slater and Birney 2005) in gene reconstruction, which is the method used by HYBPIPER.

The ability of CAPTUS to successfully combine different DNA data types is confirmed by the topological results where data available from diverse sources for the same species,

644 consistently formed well-supported and stable clades (Supplement Trees). Also, by
645 assembling the entire set of reads in a sample, CAPTUS uses off-target reads efficiently instead
646 of excluding them like HYBPIPER. Even though our target capture data was supposed to
647 comprise only 353 genes, CAPTUS was able to find almost three times as many nuclear genes
648 (c. 1130 genes on average) using our RNA5435 reference targets, as well as every organellar
649 protein and mostly complete plastomes. This indicates that CAPTUS takes better advantage of
650 the imperfect process of DNA hybridization that will usually carry over many other genomic
651 regions that would remain unused otherwise.

652 Regarding the phylogenetic estimate, our results show a well-supported nitrogen-
653 fixing clade where Cucurbitales + Rosales are sister to Fabales + Fagales (Fig. S3, Appendix
654 S3, Fig. S4). This topology has also been recovered by recent nuclear phylogenomic analyses
655 (Guo et al. 2021; Zuntini et al. in review) but differs from phylogenetic estimates mostly
656 based on chloroplast data (Soltis et al. 1995; The Angiosperm Phylogeny Group 2016; Li et
657 al. 2021) where the topology (Fabales (Rosales (Cucurbitales, Fagales) is recovered.
658 Assuming the gain of nitrogen-fixing capacity evolved in the common ancestor of this clade,
659 the net number of independent losses inside the clade (e.g., Griesmann et al. 2018) should not
660 be affected by this new topology, however future studies on the subject could benefit by also
661 interpreting gains and losses according to this new topology. Within the Cucurbitales, all the
662 currently accepted families are recovered as monophyletic groups in our analyses (Fig. 2, Fig.
663 S3, Appendix S3, Fig. S4, Fig. S5). The relationships among families agree well with
664 previous phylogenetic and phylogenomic studies of Cucurbitales, except for the position of
665 the holoparasitic Apodanthaceae, where we find strong support for a position outside
666 Cucurbitales, in contrast to the earlier results of Filipowicz and Renner (2010). In our dataset,
667 Apodanthaceae groups with Malpighiales in agreement with the recent result of Zuntini et al.
668 (in review) based on a 58% sampling of the 13,600 genera of angiosperms. However, our
669 analyses indicate a highly supported sister group relationship between the Apodanthaceae and

the Rafflesiaceae (Fig. S3, Appendix S3, Fig. S4) while Zuntini et al. (in review) recover the Rafflesiaceae well-nested within Malpighiales but with low nodal support. A deeper phylogenomic analysis of Malpighiales that takes advantage of the increased taxon sampling in Zuntini et al. (in review) and our increased gene sampling could prove useful to better resolve these relationships. For the autotrophic families we find most support for a grade where the clade Coriariaceae + Corynocarpaceae is followed by Anisophylleaceae, a clade with the triplet Tetramelaceae plus Begoniaceae and Datiscaceae, and by Cucurbitaceae (Fig. 2, Fig. S3a, Fig. S4). Earlier studies placed Anisophylleaceae as sister to all other except Apodanthaceae (Zhang et al. 2006; Schaefer and Renner 2011b; Zuntini et al. in review). Even though our analyses show the highest support for the clade Coriariaceae + Corynocarpaceae as sister to the rest of Cucurbitales, there is almost equal support among gene trees for Anisophylleaceae as sister to the rest (as in previous studies) as well as for a clade (Anisophylleaceae (Coriariaceae, Corynocarpaceae) as sister to the rest of families in the order (Fig. 2, Fig. S3a, Fig. S4). For Datiscaceae, Zhang et al. (2006) and Schaefer and Renner (2011b) also found a sister group relationship to Begoniaceae, albeit with low support. Here, we find that such a clade has the highest gene tree support but the alternatives Datiscaceae sister to Tetramelaceae [also recovered by Zuntini et al. (in review)] and Datiscaceae sister to Begoniaceae + Tetramelaceae also receive almost equal support (Fig. 2, Fig. S4). In the network (Fig. 3), the two dioecious families Datiscaceae + Tetramelaceae form a clade, which is sister to the monoecious Begoniaceae.

Tribal relationships within Cucurbitaceae match well the phylotranscriptomics results of Guo et al. (2020). The four conflicting nodes identified in Bellot et al. (2020) concerning the position of *Luffa*, *Hodgsonia*, *Bryonia*, and *Indofevillea* are stable when coalescent and concatenated phylogeny estimates are compared (Fig. S3, Appendix S3): *Indofevillea* is placed as sister to the Southeast Asian Thladiantheae; the sponge gourds (*Luffa*) are sister to all other Sicyoeae; the Asian *Hodgsonia* with the Neotropical Sicyoeae are sister to

696 *Trichosanthes*; and Bryonieae plus Schizopeponeae are sister to Sicyoeae. Looking at the
697 gene tree analyses, however, the considerable number of trees supporting alternative positions
698 of those lineages is evident (Fig. 2, indicated in red), indicating frequent hybridization events
699 in the evolutionary history of Cucurbitaceae.

700 CAPTUS allows the disentanglement of the complex pattern of deep reticulated
701 evolution in Cucurbitales, which seems to be prevalent across the angiosperms (Stull et al.
702 2023). The comparison of the gene tree frequencies obtained from the 353 captured regions
703 (Angiosperms353, or the taxonomically expanded Mega353) with the frequencies obtained
704 from the RNA5435 regions shows stable 1/3:1/3:1/3 ratio for nodes showing incomplete
705 lineage sorting (e.g., nodes 14, 24, 29, 32, and 41 in Fig. S4a) or 1/2:1/2:0 ratio for nodes of
706 hybrid origin (e.g. node 7 in Fig. S4a). This indicates that adding more genomic regions is
707 unlikely to change the family and tribal level relationships in Cucurbitales found in our study.
708 Future work should focus on the more recent evolutionary history of the clade and
709 phylogenetic patterns within Cucurbitaceae genera.

710 In conclusion, we show that our new pipeline can handle a complex phylogenomic
711 analysis in a very efficient way. The clustering capability of CAPTUS enables its application
712 not only to seed plants but to any taxonomic group, even those where a reference set of
713 orthologous loci has not yet been developed. Thus, CAPTUS can be used as a universal tool for
714 the assembly of phylogenomic datasets, even with mixed data of different origins, with
715 degraded and contaminated samples, and even in taxonomic groups with a very complex
716 evolutionary history.

717

718 ## ACKNOWLEDGEMENTS

719 We are grateful to the curators of the herbarium of Royal Botanic Gardens, Kew (K),
720 the herbarium of Muséum National d'Histoire Naturelle in Paris (P), and the herbarium of the
721 Botanische Staatssammlung München (M) for permission to extract DNA from selected

722 specimens. We also thank A. Tellier (TUM) for access to the Population Genetics HPC. The
723 study was funded by the German Science Foundation DFG - SCHA 1875/4-2 within SPP
724 1991 Taxon-Omics (to HS), and by grants from the Calleva Foundation to the Plant and
725 Fungal Trees of Life Project (PAFTOL) at the Royal Botanic Gardens, Kew.

726

727 LITERATURE CITED

728 Andermann T., Cano Á., Zizka A., Bacon C., Antonelli A. 2018. SECAPR—a bioinformatics
729 pipeline for the rapid and user-friendly processing of targeted enriched Illumina
730 sequences, from raw reads to alignments. PeerJ. 6:e5175.

731 Andrews S. 2019. FastQC: A quality control analysis tool for high throughput sequencing
732 data. Available from https://github.com/s-andrews/FastQC.

733 Baker W.J., Bailey P., Barber V., Barker A., Bellot S., Bishop D., Botigué L.R., Brewer G.,
734 Carruthers T., Clarkson J.J., Cook J., Cowan R.S., Dodsworth S., Epitawalage N.,
735 Françoso E., Gallego B., Johnson M.G., Kim J.T., Leempoel K., Maurin O., Mcginnie
736 C., Pokorny L., Roy S., Stone M., Toledo E., Wickett N.J., Zuntini A.R., Eiserhardt
737 W.L., Kersey P.J., Leitch I.J., Forest F. 2022. A Comprehensive Phylogenomic
738 Platform for Exploring the Angiosperm Tree of Life. Systematic Biology. 71:301–
739 319.

740 Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M.,
741 Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N.,
742 Tesler G., Alekseyev M.A., Pevzner P.A. 2012. SPAdes: A New Genome Assembly
743 Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational
744 Biology. 19:455–477.

745 Bellot S., Mitchell T.C., Schaefer H. 2020. Phylogenetic informativeness analyses to clarify
746 past diversification processes in Cucurbitaceae. Sci Rep. 10:488.

747 Bellot S., Renner S. 2014. The systematics of the worldwide endoparasite family
748 Apodanthaceae (Cucurbitales), with a key, a map, and color photos of most species.
749 PhytoKeys. 36:41–57.

750 Buchfink B., Reuter K., Drost H.-G. 2021. Sensitive protein alignments at tree-of-life scale
751 using DIAMOND. Nat Methods. 18:366–368.

752 Bushnell B. 2022. BBTools: A suite of fast, multithreaded bioinformatics tools designed for
753 analysis of DNA and RNA sequence data. Available from https://jgi.doe.gov/data-and-
754 tools/software-tools/bbtools/.

755 Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L.
756 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

757 Doyle J.J., Doyle J.L. 1987. A rapid DNA isolation procedure for small quantities of fresh
758 leaf tissue. Phytochemical Bulletin. 19:11–15.

759 Eaton D.A.R. 2020. Toytree: A minimalist tree visualization and manipulation library for
760 Python. Methods Ecol Evol. 11:187–191.

761 Eaton D.A.R., Overcast I. 2020. ipyrad: Interactive assembly and analysis of RADseq
762 datasets. Bioinformatics. 36:2592–2594.

763 Edgar R.C. 2022. Muscle5: High-accuracy alignment ensembles enable unbiased assessments
764 of sequence homology and phylogeny. Nat Commun. 13:6968.

765 Fierst J.L., Murdock D.A. 2017. Decontaminating eukaryotic genome assemblies with
766 machine learning. BMC Bioinformatics. 18:533.

767 Filipowicz N., Renner S.S. 2010. The worldwide holoparasitic Apodanthaceae confidently
768 placed in the Cucurbitales by nuclear and mitochondrial gene trees. BMC Evol Biol.
769 10:219.

770 Goodall-Copestake W.P., Harris D.J., Hollingsworth P.M. 2009. The origin of a mega-diverse
771 genus: dating *Begonia* (Begoniaceae) using alternative datasets, calibrations and
772 relaxed clock methods. Botanical Journal of the Linnean Society. 159:363–380.

773 Griesmann M., Chang Y., Liu X., Song Y., Haberer G., Crook M.B., Billault-Penneteau B.,
774 Lauressergues D., Keller J., Imanishi L., Roswanjaya Y.P., Kohlen W., Pujic P.,
775 Battenberg K., Alloisio N., Liang Y., Hilhorst H., Salgado M.G., Hocher V., Gherbi
776 H., Svistoonoff S., Doyle J.J., He S., Xu Y., Xu S., Qu J., Gao Q., Fang X., Fu Y.,
777 Normand P., Berry A.M., Wall L.G., Ané J.-M., Pawlowski K., Xu X., Yang H.,
778 Spannagl M., Mayer K.F.X., Wong G.K.-S., Parniske M., Delaux P.-M., Cheng S.
779 2018. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis.
780 Science. 361:eaat1743.

781 Guo J., Xu W., Hu Y., Huang J., Zhao Y., Zhang L., Huang C.-H., Ma H. 2020.
782 Phylotranscriptomics in Cucurbitaceae Reveal Multiple Whole-Genome Duplications
783 and Key Morphological and Molecular Innovations. Molecular Plant. 13:1117–1133.

784 Guo X., Fang D., Sahu S.K., Yang S., Guang X., Folk R., Smith S.A., Chanderbali A.S., Chen
785 S., Liu M., Yang T., Zhang S., Liu X., Xu X., Soltis P.S., Soltis D.E., Liu H. 2021.
786 Chloranthus genome provides insights into the early diversification of angiosperms.
787 Nat Commun. 12:6930.

788 Hatje K., Keller O., Hammesfahr B., Pillmann H., Waack S., Kollmar M. 2011. Cross-species
789 protein sequence and gene structure prediction with fine-tuned Webscipio 2.0 and
790 Scipio. BMC Res Notes. 4:265.

791 Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2018. UFBoot2:
792 Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution.
793 35:518–522.

794 Huson D.H., Bryant D. 2006. Application of Phylogenetic Networks in Evolutionary Studies.
795 Molecular Biology and Evolution. 23:254–267.

796 Jackson C., McLay T., Schmidt-Lebuhn A.N. 2023. hybpiper-nf and paragone-nf:
797 Containerization and additional options for target capture assembly and paralog
798 resolution. Appl Plant Sci. 11:e11532.

Johnson M.G., Gardner E.M., Liu Y., Medina R., Goffinet B., Shaw A.J., Zerega N.J.C., Wickett N.J. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Applications in Plant Sciences. 4:1600016.

Johnson M.G., Pokorny L., Dodsworth S., Botigué L.R., Cowan R.S., Devault A., Eiserhardt W.L., Epitawalage N., Forest F., Kim J.T., Leebens-Mack J.H., Leitch I.J., Maurin O., Soltis D.E., Soltis P.S., Wong G.K., Baker W.J., Wickett N.J. 2019. A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. Systematic Biology. 68:594–606.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 14:587–589.

Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 30:772–780.

Kent W.J. 2002. BLAT—The BLAST-Like Alignment Tool. Genome Res. 12:656–664.

Kocyan A., Zhang L.-B., Schaefer H., Renner S.S. 2007. A multi-locus chloroplast phylogeny for the Cucurbitaceae and its implications for character evolution and classification. Molecular Phylogenetics and Evolution. 44:553–577.

Li D., Liu C.-M., Luo R., Sadakane K., Lam T.-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 31:1674–1676.

Li H.-T., Luo Y., Gan L., Ma P.-F., Gao L.-M., Yang J.-B., Cai J., Gitzendanner M.A., Fritsch P.W., Zhang T., Jin J.-J., Zeng C.-X., Wang H., Yu W.-B., Zhang R., Van Der Bank M., Olmstead R.G., Hollingsworth P.M., Chase M.W., Soltis D.E., Soltis P.S., Yi T.-S., Li D.-Z. 2021. Plastid phylogenomic insights into relationships of all flowering plant families. BMC Biol. 19:232.

Matasci N., Hung L.-H., Yan Z., Carpenter E.J., Wickett N.J., Mirarab S., Nguyen N., Warnow T., Ayyampalayam S., Barker M., Burleigh J.G., Gitzendanner M.A., Wafula E., Der J.P., dePamphilis C.W., Roure B., Philippe H., Ruhfel B.R., Miles N.W., Graham S.W., Mathews S., Surek B., Melkonian M., Soltis D.E., Soltis P.S., Rothfels C., Pokorny L., Shaw J.A., DeGironimo L., Stevenson D.W., Villarreal J.C., Chen T., Kutchan T.M., Rolf M., Baucom R.S., Deyholos M.K., Samudrala R., Tian Z., Wu X., Sun X., Zhang Y., Wang J., Leebens-Mack J., Wong G.K.-S. 2014. Data access for the 1,000 Plants (1KP) project. GigaSci. 3:17.

McLay T.G.B., Birch J.L., Gunn B.F., Ning W., Tate J.A., Nauheimer L., Joyce E.M., Simpson L., Schmidt-Lebuhn A.N., Baker W.J., Forest F., Jackson C.J. 2021. New targets acquired: Improving locus recovery from the Angiosperms353 probe set. Appl Plant Sci. 9:aps3.11420.

Minh B.Q., Hahn M.W., Lanfear R. 2020a. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. Molecular Biology and Evolution. 37:2727–2733.

840 Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A.,
841 Lanfear R. 2020b. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
842 Inference in the Genomic Era. Molecular Biology and Evolution. 37:1530–1534.

843 Nachtigall P.G., Kashiwabara A.Y., Durham A.M. 2021. CodAn: predictive models for
844 precise identification of coding regions in eukaryotic transcripts. Briefings in
845 Bioinformatics. 22:bbaa045.

846 One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and
847 the phylogenomics of green plants. Nature. 574:679–685.

848 Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2 – Approximately Maximum-Likelihood
849 Trees for Large Alignments. PLoS ONE. 5:e9490.

850 Puritz J.B., Hollenbeck C.M., Gold J.R. 2014. *dDocent* : a RADseq, variant-calling pipeline
851 designed for population genomics of non-model organisms. PeerJ. 2:e431.

852 Renner S.S., Barreda V.D., Tellería M.C., Palazzesi L., Schuster T.M. 2020. Early evolution
853 of Coriariaceae (Cucurbitales) in light of a new early Campanian (ca. 82 Mya) pollen
854 record from Antarctica. TAXON. 69:87–99.

855 Rochette N.C., Rivera-Colón A.G., Catchen J.M. 2019. Stacks 2: Analytical methods for
856 paired-end sequencing improve RADseq-based population genomics. Mol Ecol.
857 28:4737–4754.

858 Sayyari E., Whitfield J.B., Mirarab S. 2018. DiscoVista: Interpretable visualizations of gene
859 tree discordance. Molecular Phylogenetics and Evolution. 122:110–115.

860 Schaefer H. 2020. Cucurbit Website. Version 1. Available from www.cucurbit.de.

861 Schaefer H., Heibl C., Renner S.S. 2009. Gourds afloat: a dated phylogeny reveals an Asian
862 origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events.
863 Proc. R. Soc. B. 276:843–851.

864 Schaefer H., Renner S.S. 2011a. Cucurbitaceae. In: Kubitzki K., editor. The Families and
865 Genera of Flowering Plants. X. Flowering Plants: Eudicots. Sapindales, Cucurbitales,
866 Myrtaceae. Springer, Berlin. p. 112–174.

867 Schaefer H., Renner S.S. 2011b. Phylogenetic relationships in the order Cucurbitales and a
868 new classification of the gourd family (Cucurbitaceae). Taxon. 60:122–138.

869 de Sena Brandine G., Smith A.D. 2021. Falco: high-speed FastQC emulation for quality
870 control of sequencing data. F1000Res. 8:1874.

871 Slater G., Birney E. 2005. Automated generation of heuristics for biological sequence
872 comparison. BMC Bioinformatics. 6:31.

873 Soltis D.E., Soltis P.S., Morgan D.R., Swensen S.M., Mullin B.C., Dowd J.M., Martin P.G.
874 1995. Chloroplast gene sequence data suggest a single origin of the predisposition for
875 symbiotic nitrogen fixation in angiosperms. Proc. Natl. Acad. Sci. U.S.A. 92:2647–
876 2651.

877 Steenwyk J.L., Buida T.J., Li Y., Shen X.-X., Rokas A. 2020. ClipKIT: A multiple sequence
878 alignment trimming software for accurate phylogenomic inference. PLoS Biol.
879 18:e3001007.

880 Steinegger M., Söding J. 2018. Clustering huge protein sequence sets in linear time. Nat
881 Commun. 9:2542.

882 Stevens P.F. 2001. Angiosperm Phylogeny Website. Version 14, July 2017 [and more or less
883 continuously updated since]. Available from
884 http://www.mobot.org/MOBOT/research/APweb/.

885 Stull G.W., Pham K.K., Soltis P.S., Soltis D.E. 2023. Deep reticulation: the long legacy of
886 hybridization in vascular plant evolution. The Plant Journal.

887 Tange O. 2021. GNU Parallel 20220422 ('Буча'). Zenodo.

888 The Angiosperm Phylogeny Group. 2016. An update of the Angiosperm Phylogeny Group
889 classification for the orders and families of flowering plants: APG IV. Bot. J. Linn.
890 Soc. 181:1–20.

891 Zhang C., Mirarab S. 2022a. ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-
892 copy gene family trees. Bioinformatics. 38:4949–4950.

893 Zhang C., Mirarab S. 2022b. Weighting by Gene Tree Uncertainty Improves Accuracy of
894 Quartet-based Species Trees. Molecular Biology and Evolution. 39:msac215.

895 Zhang C., Scornavacca C., Molloy E.K., Mirarab S. 2020. ASTRAL-Pro: Quartet-Based
896 Species-Tree Inference despite Paralogy. Molecular Biology and Evolution. 37:3292–
897 3307.

898 Zhang L.-B., Simmons M.P., Kocyan A., Renner S.S. 2006. Phylogeny of the Cucurbitales
899 based on DNA sequences of nine loci from three genomes: Implications for
900 morphological and sexual system evolution. Molecular Phylogenetics and Evolution.
901 39:305–322.

902 Zhang L.-B., Simmons M.P., Renner S.S. 2007. A phylogeny of Anisophylleaceae based on
903 six nuclear and plastid loci: Ancient disjunctions and recent dispersal between South
904 America, Africa, and Asia. Molecular Phylogenetics and Evolution. 44:1057–1067.

905 Zuntini A.R., Carruthers T., et al. in review. Phylogenomics and the rise of the angiosperms.
906 Manuscript submitted for publication.

907

908 **Supplementary Material**

909

910 TABLE S1. List of samples sequenced for this study.

911 TABLE S2. List of samples with public data used in this study.

912 FIGURE S1. The CAPTUS output formats, **a)** outputs available when a protein extraction (`NUC,`
913 `PTD, MIT`) is performed, **b)** outputs available when a miscellaneous DNA extraction
914 (`DNA, CLR`) is performed.

915 APPENDIX S1. Plastome segments used as reference targets provided as a file in FASTA
916 format `Plastome38.fasta`

917 APPENDIX S2. Newly found nuclear putative homologs used as reference targets provided as a
918 file in FASTA format `RNA5435.fasta`

919 TABLE S3. Per-sample comparison of running time, number of loci recovered, and total CDS
920 length recovered among CAPTUS, HYBPIPER-BLASTX, and HYBPIPER-DIAMOND.

921 TABLE S4. Summary statistics per sample at each analysis step.

922 FIGURE S2. Total aligned ungapped length per sample using different nuclear reference targets
923 sets, **a)** Angiosperms353, **b)** Mega353, **c)** RNA5435.

924 SUPPLEMENTARY METHOD. Decontamination process for target capture sample *Lemurosicyos*
925 *variegata*.

926 TABLE S5. Samples lacking sufficient chloroplast data for phylogenetic estimation or with
927 chloroplast contamination.

928 FIGURE S3. Coalescent species trees estimated with ASTRAL-PRO using gene trees estimated
929 by IQ-TREE on alignments derived from different reference targets sets (RNA5435,
930 Angiosperms353, Mega353) and paralog filtering strategies (`unfiltered`,
931 `informed`, `naive`).

932 APPENDIX S3. Estimated species trees in NEWICK format.

933 FIGURE S4. DISCOVISTA analyses of relative quartet frequencies imposed on the coalescent
934 ASTRAL-PRO topology shown in Fig. S3a using sets of gene trees derived from the
935 different reference target sets (RNA5435, Angiosperms353, Mega353) calculated by
936 different programs (IQ-TREE, FASTTREE) and filtered for paralogs (`informed`,
937 `naive`).
938 FIGURE S5. Species tree estimated by IQ-TREE on the concatenated alignments derived from
939 the Plastome38 reference targets set.
940 APPENDIX S4. Species trees in NEWICK format estimated for the pipeline comparison.
941 APPENDIX S5. Reference targets sets used in the decontamination of *Lemurosicyos variegata*
942 (Supplementary Method).
943