# Data, time and money: evaluating the best compromise for inferring molecular phylogenies of non-model animal taxa

Paul Zaharias, Eric Pante, Delphine Gey, Alexander E Fedosov, Nicolas Puillandre

# Data, Time and Money: Evaluating the Best Compromise for Inferring Molecular Phylogenies of Non-Model Animal Taxa

Zaharias Paul[1]*, Pante Eric[2], Gey Delphine[3], Fedosov Alexander[4], Puillandre Nicolas[1]

[1]*Institut Systématique Evolution Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, 43 rue Cuvier, CP 26, 75005 Paris, France.*

[2]*Littoral, Environnement et Sociétés (LIENSs), UMR 7266 CNRS - Université de La Rochelle, 2 rue Olympe de Gouges, 17042, La Rochelle, France.*

[3]*Acquisition et Analyses de Données pour l'histoire naturelle (2AD) UMS 2700, Muséum National d'Histoire Naturelle, Paris, France.*

[4]*A.N. Severtzov Institute of Ecology and Evolution, Russian Academy of Sciences, Leninski prospect 33, 119071Moscow, Russian Federation*

*Corresponding author: paul.zaharias@edu.mnhn.fr

*Abstract.* — For over a decade now, High Throughput sequencing (HTS) approaches have revolutionized phylogenetics, both in terms of data production and methodology. While transcriptomes and (reduced) genomes are increasingly used, generating and analyzing HTS datasets remains expensive, time consuming and complex for most non-model taxa. Indeed, a literature survey revealed that 74% of the molecular phylogenetics trees published in 2018 are based on data obtained through Sanger sequencing. In this context, our goal was to identify the strategy that would represent the best compromise among costs, time and robustness of the resulting tree. We sequenced and assembled 32 transcriptomes of the marine mollusk family Turridae, considered as a typical non-model animal taxon. From these data, we extracted the loci most commonly used in gastropod phylogenies (*cox1*, 12S, 16S, 28S, *h3* and 18S), full mitogenomes, and a reduced nuclear transcriptome representation. With each dataset, we reconstructed phylogenies and compared their robustness and accuracy. We discuss the impact of missing data and the use of statistical tests, tree metrics, and supertree and supermatrix methods to further improve the phylogenetic data acquisition pipelines. We evaluated the overall costs (time and money) in order to identify the best compromise for phylogenetic data sampling in non-model animal taxa. Although sequencing full mitogenomes seems to constitute the best compromise both in terms of costs and node support, they are known to induce biases in phylogenetic reconstructions. Rather, we recommend to systematically include loci commonly used for phylogenetics and taxonomy (i.e. DNA barcodes, rRNA genes, full mitogenomes, etc.) among the other loci when designing baits for capture.

[phylogenomics, transcriptomics, high throughput sequencing, Sanger sequencing, non-model taxa, Turridae]

## 1. Introduction

For over a decade now, high throughput sequencing (HTS) data has allowed not only the production of a substantial amount of DNA sequences relevant for phylogenetics, but also triggered many discussions on phylogenetic reconstruction methods (e.g. Edwards 2009; Lemmon & Lemmon 2013; Leaché et al. 2015a; Leaché & Oak 2017). Most authors concluded in the superiority of HTS approaches for reconstructing trees at all phylogenetic scales, especially in terms of robustness of the tree but also in the context of studying biological processes (e.g. introgression or horizontal transfer). Nevertheless, there is still a considerable amount of recent studies presenting trees obtained using first generation sequencing (chain-termination sequencing based on the incorporation of dideoxynucleotides, herein referred to as "Sanger sequencing" – e.g. Heather & Chain 2016). This technique has typically been used to sequence a few loci amplified by PCR. Although the first commercial HTS technology was introduced in 2004 (Mardis 2008), phylogenetic studies using this technology were not exceeding 12% of the total molecular phylogenetic studies up until 2016 (Fig. 1). In 2018, only 26% of molecular phylogenetic studies were based on HTS data. The simple, rapid, and stable standard protocols for producing and analyzing datasets based on Sanger sequencing data may explain why they are still primarily used in phylogenetic studies, while HTS-based studies remain costlier and more complex (both in terms of library preparation and data analysis). Thus, the sustained attractiveness of the Sanger sequencing approach to phylogenetics, indisputable in terms of number of published studies to date, contrasts with the premise that HTS data will allow us to "achieve phylogenomic Nirvana" (Faircloth 2013).

Multiple studies have explored tree robustness through the dissection of a particular HTS dataset – the recently defined practice of “phylogenomic subsampling” (reviewed in Edwards 2016). This practice mostly focuses on “the study of the information content of phylogenomic matrices of different sizes,” implying an *in silico* subsampling of loci (Edwards 2016). One way of extending the practice of phylogenomic subsampling beyond its quantitative aspects would be to take into account the nature of particular loci. This approach would evaluate the phylogenetic significance of the use of particular genome regions (e.g. coding vs. non-coding sequences; Chen et al. 2017). The development of orthology assessment pipelines (e.g. UPhO; Ballesteros & Hormiga 2016) also enabled their comparison with the same dataset, usually of transcriptomic nature (Washburn et al. 2017; Cuhna & Giribet 2019). Other studies also focused on the sequencing method used to extract a particular set of loci, mostly leading to studies comparing HTS vs. Sanger sequencing methods to recover phylogenetic datasets and draw conclusions on the superiority of one dataset type over another (e.g. Ruane et al. 2015; Lee et al. 2018).

In addition to data exploration alone, the computational time needed to analyze various datasets - including the phylogenetic reconstruction - can be calculated (e.g. Leaché et al. 2015b). The time for sample preparation and sequencing can also be estimated (e.g. Lemmon et al. 2012), but this information is more often reported in HTS method description articles, and rarely compared among methods of data acquisition. Even fewer studies evaluated monetary costs (reagents, library preparation and sequencing) for a particular phylogenetic dataset (e.g. Moreau & Wray 2017) or more generally for a taxonomic group (e.g. McKain et al. 2018). Finally, we only found two studies that evaluated both time and money in relation with the preparation of a specific

phylogenetic dataset (Lemmon et al. 2012; Cruaud et al. 2014). These studies led to the conclusion that HTS data will produce more data and more robust trees, justifying the higher costs.

However, such studies are often conducted on so-called “model taxa” (e.g. Primates in Collins & Hbrek 2018), for which genomic data is already abundant and the range of possible data acquisition methods is not limited by the absence of annotated genomes. However, more than 99% of the biodiversity is constituted on non-model taxa, i.e. taxa for which no or little genomic and/or transcriptomic data are available, and for which orthologous loci databases are information-poor. For those taxa, alternative strategies have been developed such as exon-capture (Bi et al. 2012), Anchored hybrid Enrichment (Lemmon et al., 2012) or Ultraconserved elements (McCormack et al., 2012), referred herein as “sequence capture”. These strategies allowed phylogenomicists to utilize very distant genomic resources for specific non-model groups (e.g. Haddad et al., 2017). Still, lacking whole genome data limit the possibilities to a handful of loci (the highly conserved ones), exclude non-coding material from potential markers, and complexify the task of sorting orthology from paralogy.

Thus, most available studies focus on model taxa only, are generally limited to the comparison of datasets of either different quantity or different nature of data, but rarely both, using criteria related to tree robustness or time and money needed, but rarely both. To provide arguments to choose one strategy over another in phylogenetic reconstruction in non-model taxa, we here provide a comparison of several sequencing and tree-reconstruction strategies in terms of robustness of resulting trees, and time and money needed to produce and analyze the datasets. To do so, we used the family

120 Turridae (Conoidea, Gastropoda), a group of marine molluscs, as an example. The
121 Turridae constitute a good example of non-model animal taxon because of the lack of
122 genomic resources (e.g. no assembled and annotated genome, less than ten nuclear
123 markers represented in public databases, no karyotypes available), even in closely
124 related groups. The closest reference would be the recently published genome of
125 *Pomacea canaliculata* (Liu et al. 2018) and its divergence time with the family Turridae
126 is estimated at 283 Ma (Zapata et al. 2014). The family comprises 216 species
127 (WoRMS, checked on May 2019) but this number is largely underestimated (Puillandre
128 et al. 2012). Most of the molecular phylogenies published for this group used the same
129 classical mitochondrial (*cox1*, 12S and 16S rRNA) and/or nuclear (28S rRNA) markers
130 (Heralde et al. 2007, 2010; Olivera et al. 2008; Puillandre et al. 2012; Fedosov et al.
131 2011; Todd & Rawlings 2014; Puillandre et al. 2017). Two venom-gland transcriptomes
132 are published (Gonzales and Saloma 2014) that were not used primarily for
133 phylogenetic purposes but for toxin research, because the Turridae are venomous and
134 constitute a group of interest for bioactive compound discovery (Puillandre & Holford
135 2010). Finally, only one phylogenomic (RAD-seq) study (Abdelkrim et al. 2018a) was
136 published for species-delimitation purposes on eight species in the *Xenuroturris*/*Iotyrris*
137 complex.
138 We analyzed 32 transcriptomes (29 Turridae + 3 outgroups), corresponding to 18
139 species, from different tissues (venom gland, salivary gland or foot), from which we
140 extracted *in silico* five datasets: (i) the barcode fragment (658bp) of the *cox1* gene, the
141 most frequently sequenced marker in gastropod systematics; (ii) a multilocus dataset
142 that is typically produced using Sanger sequencing to conduct phylogenetic studies in
143 gastropods, corresponding to fragments of the mitochondrial *cox1*, 16S rRNA and 12S

144 rRNAgenes, and the nuclear 28S rRNA, 18S rRNA and histone *h3* genes (e.g. Fedosov
145 et al. 2018; Johnson et al 2010); (iii) complete mitochondrial genomes (e.g. Uribe et al.
146 2018); (iv) a sequence capture approach, targeting a selection of nuclear loci (e. g.
147 Abdelkrim et al. 2018b); (v) an RNA-seq dataset (e.g. Cunha et Giribet 2019). Because
148 this dataset include only a limited number of Turridae lineages, the goal is not to resolve
149 the Turridae phylogeny, but to compare those five datasets. We empirically evaluated
150 the capacity of each dataset to resolve relationships among the 32 samples, within
151 which divergence ranges between 0 (intra-specimen divergence) and 79.4 Ma
152 (estimated age of origin of the family Turridae; Abdelkrim et al. 2018b). We also
153 evaluated the time necessary for sample preparation, sequencing and data analysis,
154 along with the monetary costs of each step to estimate the overall cost of producing
155 each dataset.
156

157 ## 2. Material and Methods

158 ### 2.1 Sampling

159 Twenty-eight specimens, representing six genera of Turridae and related outgroups
160 (Conidae and Mitridae) were collected during several field expeditions organized by the
161 Muséum national d'Histoire naturelle (MNHN; "KAVIENG" in Papua New Guinea,
162 "KANACONO" in South New Caledonia), by joined Russian-Vietnamese Tropical
163 Center (Vietnam), and by the University of Utah in collaboration with the University of
164 the Philippines (Philippines). Specimens were photographed and the shells were broken
165 to access the animal. For twenty-seven specimens, only one tissue type was sampled
166 (venom gland, salivary gland or foot) depending on the project they were associated
167 with; for one specimen, both venom gland and salivary gland tissue were sampled,

resulting in a total of 29 tissues (Supplementary Table 1). Remains of vouchers, when available, were kept and are deposited in the MNHN collections.

In addition, we used publicly available transcriptomes from three species: *Unedogemmula bisaya*, *Gemmula speciosa* (Turridae) and *Terebra subulata* from a closely related family Terebridae (NCBI Sequence Read Archive (SRA) accession no.'s SRR1574923, SRR1574907 and SRR2060989, respectively; Gonzales and Saloma 2014; Gorson et al. 2015).

### 2.2 RNA Extraction, Library Preparation and Sequencing

RNA was extracted using a Trizol protocol or the Qiagen RNeasy Micro kit, following the manufacturer's recommendations. Bioanalyzer traces were used to assess total RNA quality and determine suitability for sequencing. The cDNA libraries were prepared and sequenced either at the New York Genome Center or at the Evolutionary Genetics Lab at UC Berkeley (Supplementary Table 1). In New York, libraries were prepared using the automated polyA RNAseq library prep protocol and sequenced with Illumina HiSeq 4000 with 150-bp paired-end reads. In Berkeley, the KAPA Stranded mRNA-Seq kit was used to synthesize cDNA, ligate adapters using TruSeq HT adapters and barcode samples. Samples were then sequenced with Illumina HiSeq 2000 or 4000 (see Supplementary Table 1) with 100-bp paired-end.

### 2.3 Transcriptome Assembly and Quality Assessment

All the transcriptomes, including the ones downloaded from GenBank, were assembled following the same procedure. Trimmomatic v.0.36 (Bolger et al. 2014) was used to remove adapters and filter low quality reads (ILLUMINACLIP option enabled, seed

mismatch threshold = 2, palindrome clip threshold = 40, simple clip threshold of 15; SLIDING WINDOW option enabled, window size = 4, quality threshold = 20; MINLEN = 36; LEADING = 3; TRAILING = 3). Reads were merged using FLASH v1.2.8 (Magoc and Salzberg 2011) with a min. overlap parameter of 5, a maximum overlap parameter of 100 and a mismatch ratio of 0.05. FastQC (Andrews 2010) was used for raw reads quality control. Transcripts were assembled using Trinity v2.4 with default parameter (Grabherr et al. 2011). Cap3 (Huang and Madan 1999) with default parameters and cd-hit v4.6 (percent identity = 99%; Li and Godzik 2006) were finally applied to reduce redundancy in the assemblies.

BBMap (Bushnell 2014) was used to generate basic assembly statistics and BUSCO (Simão et al. 2015) to evaluate transcriptome completeness. Finally, bowtie2 v2.2.6 (Langmead and Salzberg 2012) and samtools v1.3 (Li et al. 2009) were used to evaluate read representation in each assembled transcriptome, as recommended in the Trinity manual.

**2.4 Transcriptome Orthology Inference**

Two approaches were used to assess orthology among transcripts, from here onwards referred to as “reference-based” approach and “graph-based” (without a reference genome) approach (Fig. 2).

For the reference-based approach, the *Pomacea canaliculata* genome (ASM307304v1; Liu et al. 2018) was used as a reference. Following the pipeline described in Phuong and Mahardika (2018) and Phuong et al. (2019), blastx was used to associate transcripts to peptide sequences of *P. canaliculata* and tblastn to associate peptides of *P.canaliculata* to transcripts from the BLAST + v2.2.31 suite (Altschul et al. 1990) with

an e-value threshold of $1e10^{-10}$ and a word size value of 11. For each sample, bowtie2 v2.3.4.1 was used with the very sensitive-local alignment option and not allowing for discordant pair mapping (unexpected paired read orientation during mapping) to map reads to the selected transcripts from the reciprocal blast step. Duplicates were marked using picard-tools v2.0.1 (http://broadinstitute.github.io/picard) using default parameters. All positions with a coverage < 5X were masked and the entire sequence was removed if >30% of the sequence was masked. To fix assembly errors, single nucleotide polymorphisms (SNPs) were called using samtools v1.3 (default parameters) and bcftools v1.3 (Li et al. 2009) using the call command. Transcripts for each locus were aligned as nucleotides using MAFFT v7.222 (Katoh et al. 2005) option -auto. To limit misalignments and paralogs inclusion, uncorrected pairwise distances were calculated at each locus for all possible pairwise comparisons and sequences were removed if the uncorrected pairwise distance was greater than the 90th percentile (threshold was set empirically) of pairwise distances across all loci for that pair of species.

For the graph-based approach, we used UPhO (Unrooted Phylogenetic Orthology; Ballesteros and Hormiga 2016), a method that uses the topology of individual gene trees to identify clades corresponding to orthologous groups. Following the workflow established by the authors, all transcripts in open reading frame (ORF) were extracted from the transcriptome assemblies with custom Python scripts, and all ORFs that were less than 100 amino-acid long were eliminated. An all-versus-all blastp search was then performed, using a relaxed expectation value threshold of e = 1 x 10-5.

To reduce missing data, only the clusters that contained the maximum number of samples (32) were selected. The gene-family amino-acid sequence clusters were aligned

and cleaned using mafft (option '-auto'), trimAL (option '-gappyout') and Al2phylo (-m 32 -t 300 -p 0.80). After alignments, the sequences were converted from amino acids back to nucleotides to increase the number of informative sites and improve the phylogenetic pipeline accuracy. Gene-family trees (GFTs) were estimated using IQ-tree (Nguyen et al. 2014). The best substitution model for each GFT was estimated with ModelFinder (Kalyaanamoorthy et al. 2017) following the BIC criterion. Subsequently, 1,000 ultrafast bootstraps (UFBoot) (Hoang et al. 2017) were performed on each GFT to obtain branch support. The branches representing putative orthogroups were finally extracted with UPhO (-m 4 -S 0.80). The orthogroup alignments obtained were cleaned and analyzed using MAFFT, trimAL, Al2phylo and IQ-tree with the same parameters as above (except for the -m parameter in Al2phylo, set to 4).

**2.5 Transcriptome Phylogeny**

Ten datasets were generated. For the reference-based approach three subsets were defined with a minimum of 4, 16 and 32 samples / locus. These subsets were analyzed using a supermatrix - concatenated alignment of all the loci - and a supertree approach, resulting in six datasets referred as follows: Ref-IQ4, Ref-IQ16, Ref-IQ32, Ref-AS4, Ref-AS16 and Ref-AS32 (IQ referring to IQ-tree and AS to ASTRAL – see below). Similar subsets were constructed for the graph-based approach with 16 and 4 samples / locus (the 32 sample/locus dataset was not analyzed here because only one locus was retrieved). The resulting four datasets are referred to as follows: Uph-IQ4, Uph-IQ16, Uph-AS4 and Uph-AS16.

Best substitution models were estimated for each partition (locus) in each concatenated dataset with ModelFinder following the BIC criterion. Supermatrix trees were

reconstructed using IQ-tree and 1,000 UFBoot were performed on each dataset. An individual tree for each locus was also generated with IQ-tree, using the associated best substitution model for datasets Ref-AS4, Ref-AS16, Ref-AS32, Uph-AS4 and Uph-AS16. The supertree approach implemented in the program ASTRAL-III (Zhang, Sayyari and Mirarab 2017) was then applied to combine the single-locus trees into a single supertree for each of these datasets.

### 2.6 Sequence Capture

We used the Ref-AS4 dataset and selected the 3,000 shortest loci (ranging from 96 to 839 bp) to simulate a sequence capture datasets (Bi et al. 2012; Jiang et al. 2017; Abdelkrim et al. 2018b). Three subsets were generated, with a minimum of 4, 16 and 32 samples / locus for which both supermatrix and supertree approaches were applied, as explained above. These datasets will be referred as follow: Cap-IQ4, Cap-IQ16, Cap-IQ32, Cap-AS4, Cap-AS16 and Cap-AS32.

### 2.7 Mitogenomes and Nuclear Markers

The *Pinguigemmula* sp. (Turridae) mitogenome (MH308408.1; Uribe et al. 2018) was used as a reference to extract partial (up to 20% missing data) to complete mitogenomes (including tRNAs) from the transcriptomes and create the dataset “MT.” Several sequences of 28S rRNA, 18S rRNA and histone 3 (*h3*) of Turridae from GenBank were used as references to extract the corresponding loci from the 32 transcriptomes by BLAST. Along with the mitochondrial *cox1*, 12S and 16S fragments, they constitute the Sanger multilocus dataset “SAN.” Finally, the *cox1* alone constitutes the Sanger barcode dataset “BC.” The same protocol as for the reference-based approach was

288 applied for mapping, filtering and alignment. For the MT, SAN and BC datasets, each
289 codon position of the protein coding genes was treated as an independent partition, as
290 well as each non-protein coding gene. The best substitution model was estimated for
291 each partition in each concatenated dataset with ModelFinder following the BIC
292 criterion and 1,000 UFBoot were performed on each dataset to obtain branch support
293 for the trees reconstructed with IQ-tree.
294
295 **2.8 Tree Topology Evaluation**
296 The Turridae trees published so far suffer from both incomplete sampling and lack of
297 resolution (e.g. Heralde et al. 2007; Puillandre et al. 2012). Thus, these published trees
298 can hardly be used as a reference tree for the Turridae. Consequently, two approaches
299 were used to evaluate tree topology decisiveness and informativeness.
300 For the matrix and supermatrix datasets (BC, SAN, MT, Cap-IQ, Ref-IQ, Uph-IQ), the
301 log-likelihood of multiple constrained tree searches for each dataset was compared and
302 the results were statistically tested with IQ-TREE using the Shimodaira-Hasegawa
303 (1999) (SH) test. The trees were constrained respectively following all the different
304 topologies retrieved with the different datasets, except for the intra-specific and
305 outgroup nodes, resulting in a total of eight unique constrained topologies (the same
306 topologies found for Cap-IQ32 and CapIQ16, Cap-IQ4 and Ref-IQ32, Ref-IQ16 and
307 Ref-IQ4).
308 For the supertree datasets (Cap-AS, Ref-AS and Uph-AS), tree metrics were used to
309 evaluate loci quality. The normalized quartet distance of each locus was calculated
310 using TreeCmp (Bogdanowicz et al. 2012) with reference to the corresponding supertree
311 with collapsed intraspecies nodes. Additionally, the quartet distance metric score

distribution of BUSCO (single-copy + fragmented) loci trees versus all other single-locus trees for Ref-AS16 and Ref-AS32 were compared to evaluate the quality of the reference-based approach. The quartet score (proportion of quartets satisfying the supertree) was also used to evaluate the overall support of supertree analysis using ASTRAL-III's log.

### 2.9 Data, Time and Money Evaluation

*Data* – The AMAS python program (Borowiec 2016) was used to calculate alignment statistics for each dataset, including the number of loci, the alignment length (in the case of ASTRAL-III, the median length of all loci), the total number of matrix cells and undetermined cells (to evaluate missing data) and the proportion of variable and parsimony-informative sites.

*Time and money* – Comparisons of costs (time and money) were measured respectively in number of days and euros but did not take into account specimen collection and salary costs, both varying too much depending respectively on the taxon, the country where research is carried out, or the academic level of the person employed (e.g. graduate or engineer). Costs were evaluated by the Service de Systématique Moléculaire (SSM) platform at the MNHN (UMS 2700). The time estimates were based on a realistic best-case scenario, meaning that each step of lab preparation and data analysis are supposed to work on the first try with the methods used at the SSM.

## 3. Results

### 3.1 Transcriptome Sequencing, Assembly and Quality Assessment

The total number of raw reads used for transcriptome assembly ranged from 42,770,212 to 138,181,918 and the number of assembled contigs ranged from 46,027 to 283,318. The mean value of N50 is 539. At least 80% of input reads mapped back to the transcriptome assemblies. The mean BUSCO completeness value is 49.1%, ranging from 36% to 83.7% (Supplementary Table 1). Pearson's r showed a strong correlation between assembly size and BUSCO completeness ($\rho$=0.78, p-value = 1.54E-07) but no correlation between the number of raw reads and BUSCO completeness ($\rho$=-0.01, p-value= 0.98) (Supplementary Table 2). Transcriptomes produced from foot tissue (*Gemmula* sp. and *M. mitra*) showed a greater BUSCO completeness than transcriptomes produced from venom or salivary glands, suggesting transcript abundance variation among tissues and/or overrepresentation of some transcripts in glands (e.g. highly expressed toxins – Dutertre et al. 2014). However, more transcriptomes assembled based on different tissues from the same specimen are needed to properly test this hypothesis.

### 3.2 Phylogenetic Results

The monophyly of the ingroup Turridae is always confirmed, except with two datasets (Uph-AS16 and Uph-AS4), where the outgroup *Terebra* is found in the ingroup (Supplementary Fig. 1). The genera *Gemmula* and *Turris* are systematically retrieved polyphyletic (Fig. 3), as shown in previous studies (Puillandre et al. 2012; Fedosov et al. 2011). The species represented by several specimens (*X. legitima*, *I. cingulifera*, *I. musivum and I. olangoensis*) are always recovered as monophyletic groups except for one dataset (Uph-IQ4), in which a specimen of *I. cingulifera* is placed as a sister group of the other members of *Iotyrris*. Apart from the Uph-IQ4 dataset, the relationships

359 inferred among *X. legitima* and all three *Iotyrris* species are always identical. The long
360 branches *Turris* and *Lophiotoma* are found as sister groups only in the “Ref” and “Cap”
361 datasets. Finally, the relationships among *Gemmula* sp., *T. nadaensis*, *Unedogemmula* –
362 the earliest offshoots in the ingroup – and the rest of the Turridae appear to be the most
363 problematical (Fig. 3). The phylogenetic results are globally congruent with previous
364 studies (e.g. Puillandre et al. 2012), despite the heterogeneity in the number of species
365 per lineage and several missing lineages. Overall, the graph-based approach (UPhO)
366 shows very low taxon occupancy (see also Fernandez et al., 2018) and fewer
367 parsimony-informative sites, and hence results in shortest branches and incongruent
368 results with the reference-based approach. An extreme case is the specimen of *I.*
369 *cingulifera* not retrieved within the *I. cingulifera* species node in the UPh-IQ4 dataset.
370 This specimen’s transcriptomes shows poorer results in terms of assembly size
371 (38,931,364 bp, compared to the 56,711,565 mean) and BUSCO completeness (23.8%
372 of complete single loci). Nevertheless, the reference-based reconstructions do not suffer
373 from this low-quality transcriptome.
374 Except for the BC and SAN datasets, support for specific to supra-specific nodes ranged
375 between 75% and 100% (Table 1), and shows no correlation with the dataset size.
376 Interestingly, in the mitogenome dataset (MT), bootstrap supports were similar or
377 superior to those of larger datasets, but those values were negatively affected by the
378 removal of some regions such as tRNAs (Supplementary Fig. 1).
379

380 **3.3 Topology Evaluation**

381 Except for UPh-IQ4, all the datasets had at least one alternative constrained topology
382 credible under the SH test (Table 2). The credible sets of trees for the smallest datasets

(BC, SAN and MT) contained more constrained trees than the credible sets of trees for the larger datasets (Cap, Ref & Uph).

Not a single-locus tree with 16 or more terminal entities fully matches its corresponding supertree (Figure 4). This is also true for the UPhO-AS16 single-locus tree distribution (Supplementary Fig. 2). The student's t test results of quartet distance metric score distribution of BUSCO (single-copy + fragmented) loci trees versus all other loci trees for Ref-AS16 showed a significant difference between the two distributions (p-value <2.2e-16; Fig. 4b). The quartet score decreases when reducing taxon occupancy: for Ref-AS32, Ref-AS16 and Ref-AS4 the normalized quartet scores were respectively 0.730, 0.709 and 0.707 (Supplementary Table 3).

### 3.4 Data, Time and Money

All Sanger markers were extracted from the transcriptomes except for *h3*, lacking in 25 of the transcriptomes. The largest dataset (DS5aIQ4) is a concatenated alignment of 14,586,607bp (71.7% of missing data), corresponding to 9,232 loci (DS5aAS4), of which all other datasets were constructed, except for the graph-based approach ones. The graph-based approach generated too few loci with no missing data (32 terminals/locus), therefore only four datasets were retained (Table 1). As shown on the Figure 5, the reference-based and graph-based approach used respectively 285,660 and 35,595 transcripts for each pipeline, but only 19,008 (6.8%) of the total transcripts are in common between the two pipelines.

Unsurprisingly, the larger datasets are also more costly (Table 1), ranging from an estimated 226€ for the CO1 dataset to 8,828€ for transcriptomes, for the production of a 32 terminal entity phylogeny (as for this study). But, while the Sanger datasets (BC,

SAN) costs increase proportionally with the number of specimens and number of loci targeted, the mitochondrial and sequence capture datasets costs will dramatically reduce when pooling a lot of specimens. This is particularly the case for the sequence capture dataset, especially when considering the price of custom baits. By pooling 100 post-capture libraries on a single sequencing lane (instead of the 32 in this study), the cost per specimen goes down from 196€ to 81€ (273€ to 105€ if including the transcriptome sequencing and the design of the probes). Finally, the transcriptomes dataset is the only HTS dataset not following the rule of decreasing costs when pooling more specimens, simply because there is a limit on the number of transcriptomes that can be sequenced on a single lane.

# 4. Discussion

## 4.1 Comparison of the Five Sequencing Strategies

In the present study, we compare datasets that are representative of the outputs of the pipelines used in most empirical phylogenetic studies in non-model animal taxa, and evaluate them in terms of costs (money and time) and robustness of the resulting tree. It should be noted that the conclusions on the cost evaluations rely on the assumption that the overall costs and timeframes of analyzed methodologies will be similar in other labs. Furthermore, another cost, the environmental cost (the impact of each pipeline on the environment), was not calculated due to the multicity of parameters to take in account. However, library preparations and the use of data centers (Jones, 2018) would surely represent a substantial environmental cost for HTS-based trees. If this cost is rarely considered, in the future scientists might be encouraged to lower their ecological footprint.

Our results show that traditional Sanger sequencing of one to six loci will retrieve trees with robust nodes for more than half of the clades, quickly and at very affordable costs. Indeed, the *cox1* barcode tree alone retrieved both monophyletic species and most nodes well supported. This particular result might partially explain why, despite 15 years of HTS development and democratization, the vast majority of articles is still presenting trees produced with such datasets (Fig. 1). Surely, the "Sanger era" has not yet arrived to its end, and many more phylogenies with such datasets will be published in the years to come.

Nevertheless, some nodes remain unsupported, in particular the deeper nodes. We found that the best compromise for retrieving a fully resolved and highly supported tree is the mitogenome dataset, for which all nodes have >80% bootstrap and the costs are less than half the price of a sequence capture. However, previous studies have already shown that mitogenomic trees are subject to artifacts, such as long-branch attraction generated because of the high rates of mutation of the mitochondrial genome, especially in the third codon positions (Bergsten 2005, Arabi et al. 2010). Moreover, a mitogenome can be considered as a single locus and thus cannot be subjected to congruence tests. The use of nuclear HTS data becomes even more indispensable when investigating biological processes such as introgression (e.g. Eaton et al. 2015; Zhang et al. 2015), where analysis of unlinked markers is necessary.

The sequence capture and RNA-seq datasets (based on a reference genome) yielded similar results in terms of phylogenetic reconstruction accuracy, number of credible sets of trees passing the SH test and single-loci tree metrics distribution. However, the costs of sequence capture are by far more affordable than costs of producing and analyzing transcriptomes. Furthermore, RNA-seq requires high-quality, fresh RNA samples, not

often available for a representative set of taxa. These considerations led to the conclusion that sequence capture might be the best method to produce a complete, high resolution tree for a non-model taxon, with a cost per specimen estimated at 80-100€ (if at least 100 specimens are sequenced on one lane) and a processing time of a few weeks to a few months (Supplementary Table 4). Nevertheless, transcriptomic data remains necessary to identify suitable markers that will be targeted by sequence-capture, especially when there is no available genome. Furthermore, transcriptomic data might be more suitable for backbone phylogenetic trees, including very deep relationships (i.e. several hundreds of millions of years; Cunha & Giribet 2019; Kocot et al. 2011). But very deep relationships also imply that it will be harder to distinguish orthology from paralogy. In summary, the Sanger approach still remains relevant to resolve phylogenetic relationships at a low price (both time and money), and can provide a preliminary outline of the taxon diversity, useful to select a subset of samples that can be analyzed with a more costly approach. However, some gene markers might not be as useful as thought, depending on the taxon (e.g. 18S, see Fig. 4), and 12S and 16S will generally only comfort the *cox1* results. We thus recommend starting with DNA barcoding but from there going directly to sequence capture (if there is a strong need to clarify the remaining challenging nodes). Mitogenomes indeed provide the best compromise between tree quality and costs, but are subject to potential biases. Finally, RNA-seq appears only appropriate for constructing phylogenies in the case of very deep relationships or simply to identify suitable markers for sequence capture.

Another class of HTS datasets that has not been explored is the reduced-representation approaches such as RAD-seq (e.g. Baird et al 2008). RAD-seq has already been

established as a suitable tool for phylogenetic inference (e.g., Cariou et al. 2013; Cruaud et al. 2014). In a recent *in silico* study (Collins & Hbrek 2018), the authors even found that RAD and sequence capture datasets gave highly congruent results. However, RAD-seq datasets are reduced-representation of genomes, and extracting an *in silico* RAD-seq dataset from our transcriptomes may have produced biased results, not equivalent to other RAD-seq datasets. Nonetheless, it could be argued that sequence capture methods are more promising for phylogenetic studies, because markers are not anonymous, and their sets can be tailored with more versatility according to the needs, samples with fragmented DNA can be sequenced more efficiently, information content per locus is higher (allowing the use of supertree approaches) and larger evolutionary time scales are covered (Harvey et al. 2016).

### 4.2 A Note on Topology Accuracy Assessment

As shown in Table 1, the majority of the concatenated datasets show >80% or even 100% bootstrap values for all nodes – the same applies for ASTRAL support values – even though the amount of data can vary by a factor of 100 between datasets. Despite high node support, several topologies are in conflict, especially for the earliest relationships of the Turridae (*Gemmula* sp., *Unedogemmula* and *T. nadaensis*). Even if the true tree is unknown, we know that, at best, only one of these topologies is correct. It has already been showed that the bootstrap support value can rapidly saturate when increasing the number of sites (especially invariant ones), proportion of missing data, or both (Simmons & Freudenstein 2011). Furthermore, when using supermatrix approaches, log-likelihood ratio tests have been used to statistically test if a given dataset can accommodate several topologies (e.g. McFadden et al. 2006). In our case, all

datasets (except the particular case of Uph-AS4) tolerated at least one, but not all, different (constrained) topologies, suggesting that the unconstrained topology is equal to or only slightly better than alternative one(s). The high-bootstrap values and non-conclusive log-likelihood ratio tests for each phylogenomic datasets called for alternative methods to measure tree robustness.

The normalized quartet score (Bayzid et al. 2015) is the proportion of quartets from the input single-locus trees that agree with the resulting supertree. We used it to measure the relevance of datasets with low taxon occupancy (e.g. Ref-AS4) when considering a supertree approach. Our results show that the normalized quartet scores for Cap-AS, Ref-AS and UPh-AS datasets are systematically lower with low taxon occupancy. Such results would imply that, as for supermatrix (Philippe et al. 2017), datasets with low taxon occupancy should be avoided (but see e.g. Kallal et al. 2018). Graphical representations of single-loci tree distribution, sometimes referred to tree space visualization in its extended version (Hillis et al. 2005) show promising results for understanding inconsistency among the datasets. The distribution of single-loci tree distance to a reference tree (Fig. 4a) has already been used to compare the quality of different datasets (Simmons 2017), but also within-dataset informativeness (e.g. intron vs. exon; Chen et al. 2017). In the case of non-model taxa, such distribution patterns can be used to compare loci with high reliability of orthologous relationships (e.g. BUSCO single-copies) versus shallow orthologous loci (e.g. from a reference-based or a graph-based approaches) and thus evaluate the quality of a pipeline (Fig. 4b). In our case, we show that a simple blast and downstream filtering approach against a reference genome, even a very distant one, gives satisfactory results, although not sufficient to obtain orthologous loci of similar confidence to BUSCO single-copy loci.

527

528 **4.3 Improving Sequence Capture: Challenges and Perspectives**

529 An important challenge of HTS in phylogenetic reconstruction is to *a priori* identify
530 loci that better reflect evolutionary relationships among taxa. Our reference-based and
531 graph-based approaches implemented herein correspond to the two alternative
532 strategies, widely used to infer orthologous loci from *de novo* assembled transcriptomes
533 (as reviewed in Laumer 2018). In our case, the graph-based approach with UPhO
534 yielded poor results in comparison to the reference-based approach, but more empirical
535 and *in silico* generated datasets need to be analyzed to properly compare them. The
536 UphO approach was especially sensitive to missing data (specimen of *I. cingulifera* not
537 found with other *I. cingulifera* specimens in DS Uph-IQ4) and the tree reconstruction
538 method (*Terebra* found in the ingroup for DS Uph-AS16 and Uph-AS4 ). One of the
539 reasons that *Terebra* was found in the ingroup for the Uph-AS16 and Uph-AS4 datasets
540 could be that the orthologs found with the graph-based approach were generally poorly
541 informative (~7% parsimony-informative sites on average), thus resulting in poorly
542 resolved single-locus trees. Conversely, the reference-based approach showed satisfying
543 results, both in terms of pipeline celerity (avoiding “all-vs-all” blast use), tree
544 robustness and congruency between subsamples. Furthermore, it retrieved far more loci
545 than the BUSCO database. However, single-loci tree evaluation (Fig. 4) showed that the
546 loci retrieved with our reference-based approach are not all informative and/or accurate,
547 and the loci selection could be improved. The use of other alignment statistics, such as
548 the proportion of parsimony-informative sites, could allow for a more precise *a priori*
549 selection of loci (e.g. HaMStR; Ebersberger et al. 2009). Nonetheless, in our dataset,
550 and quite paradoxically, there is a slight negative correlation between the number of

parsimony-informative sites in the single-locus alignments and the single-locus-tree distances to the supertree (e. g. for Ref-AS32 $\rho$=0.40, p-value = 9.93E-20 – Supplementary Table 5), suggesting that most of the phylogenetic signal retrieved in the single-locus trees would not be conveyed by the parsimony-informative sites. Finding true, orthologous, informative loci still needs development, especially when no close reference genome is available. This relies on finding a better combination of filtering thresholds, alignments statistics and tree metrics to reduce the costs and increase the tree robustness, generating a solid framework to test evolutionary hypotheses.

Finally, one particular advantage of the Sanger approach to reconstructing phylogenies is its routine application. A phylogenetic dataset can be completed regularly, by adding additional sequences on a day-to-day basis, with little doubt on the loci sequenced (but see Mutanen et al. 2016). This is less true for HTS based approaches, which usually provide a large amount of data requiring significant investment and staff trained in bioinformatics to eventually combine several datasets, produced in several batches and/or by different research teams. To combine the advantage of both approaches, i.e. a small set of well identified loci that can easily be incremented and a larger, more informative dataset, we propose the following strategy: together with the loci that will be identified as targets in the exon capture approach, the mitochondrial and nuclear loci traditionally used in Sanger sequencing (typically, the *cox1*, 16S, 12S, 28S, 18S and *h3* for the mollusks), and even full mitogenomes, could also be captured (e.g. Espeland et al. 2018with the *cox1* only). Hence, the backbone phylogeny obtained with a sequence capture dataset can further be completed with additional nuclear core markers or mitochondrial genomes, using a multilevel dataset approach.

## Fundings

This work was supported by the CONOTAX project funded by the French Agence Nationale pour la Recherche (ANR-13-JSV7-0013-01).

## Acknowledgments

Material was collected during several expeditions: the Kavieng Lagoon Biodiversity Survey in Papua New Guinea (June 2014, MoU with the University of Papua New Guinea, PIs: Philippe Bouchet, Jeff Kinch), as part of the Our Planet Reviewed expeditions organized jointly by Muséum national d'Histoire naturelle (MNHN), Pro-Natura International (PNI) and Institut de Recherche pour le Développement (IRD), with support from Papua New Guinea's National Fisheries Authority, the Total Foundation, the Laboratoire d'Excellence Diversités Biologiques et Culturelles (LabEx BCDiv, ANR-10-LABX-0003-BCDiv), the Programme Investissement d'Avenir (ANR-11-IDEX-0004-02), the Fonds Pacifique, and CNRS' Institut Ecologie et Environnement (INEE); the KANACONO expedition in New Caledonia (August 2016, convention MNHN-Province Sud, APA_NCPS_2016_012; PI N. Puillandre and S. Samadi), as part of the Our Planet Reviewed expeditions and the Tropical Deep-Sea Benthos program (expedition.mnhn.fr), with support from the LabEx BCDiv; the Nha-Trang expedition in Vietnam, supported by the Russian–Vietnamese Tropical Center, with support from the staff of the Tropical Center for assistance in organization of the field sampling and loan of some laboratory equipment; and a collection trip supported by the 'Conus-Turrid' project (principal investigator B. M. Olivera, University of Utah, USA). These expeditions were operated under the regulations then in force in the country in question and satisfy the conditions set by the Nagoya Protocol for access to

genetic resources. This project was partly supported by the Service de Systématique Moléculaire (UMS 2700 CNRS-MNHN). The authors also thank Laetitia Aznar-Cormano, Juliette Gorson and Mandë Holford for their help in the lab work, Yuri Kantor for his help in Turridae dissections, Barbara Buge for her help in curating the specimens, Mark Phuong, Lou Mary and Jérémie Bardin for their help with bioinformatics scripts.

## Declaration of interest

The authors declare no competing interests that could inappropriately influence (bias) their work.

## References

Abdelkrim, J., Aznar-Cormano, L., Fedosov, A. E., Kantor, Y. I., Lozouet, P., Phuong, M. A., Zaharias, P., & Puillandre, N. (2018a). Exon-Capture-Based Phylogeny and Diversification of the Venomous Gastropods (Neogastropoda, Conoidea). *Molecular biology and evolution*, 35(10), 2355-2374.

Abdelkrim, J., Aznar-Cormano, L., Buge, B., Fedosov, A., Kantor, Y., Zaharias, P., & Puillandre, N. (2018b). Delimiting species of marine gastropods (Turridae, Conoidea) using RAD sequencing in an integrative taxonomy framework. *Molecular ecology*, 27(22), 4591-4611.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

622 Arabi, J., Cruaud, C., Couloux, A., & Hassanin, A. (2010). Studying sources of
623 incongruence in arthropod molecular phylogenies: sea spiders (Pycnogonida) as a case
624 study. *Comptes rendus biologies*, 333(5), 438-453.

625 Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A.,
626 Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and
627 genetic mapping using sequenced RAD markers. *PloS one*, 3(10), e3376.

628 Ballesteros, J. A., & Hormiga, G. (2016). A new orthology assessment method for
629 phylogenomic data: unrooted phylogenetic orthology. *Molecular biology and evolution*,
630 33(8), 2117-2134.

631 Bayzid, M. S., Hunt, T., & Warnow, T. (2014). Disk covering methods improve
632 phylogenomic analyses. *BMC genomics*, 15(6), S7.

633 Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, 21(2), 163-193.

634 Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012).
635 Transcriptome-based exon capture enables highly cost-effective comparative genomic
636 data collection at moderate evolutionary scales. *BMC genomics*, 13(1), 403.

637 Bogdanowicz, D., Giaro, K., & Wróbel, B. (2012). TreeCmp: Comparison of trees in
638 polynomial time. *Evolutionary Bioinformatics*, 8, EBO-S9657.

639 Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for
640 Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120.

641 Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing
642 of summary statistics. *PeerJ*, 4, e1660.

643 Bushnell, B. (2014). *BBMap: a fast, accurate, splice-aware aligner* (No. LBNL-
644 7065E). Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

645 Cariou, M., Duret, L., & Charlat, S. (2013). Is RAD-seq suitable for phylogenetic
646 inference? An *in silico* assessment and optimization. *Ecology and evolution*, 3(4), 846-
647 852.

648 Chen, M. Y., Liang, D., & Zhang, P. (2017). Phylogenomic resolution of the phylogeny
649 of laurasiatherian mammals: Exploring phylogenetic signals within coding and
650 noncoding sequences. *Genome biology and evolution*, *9*(8), 1998-2012.

651 Collins, R. A., & Hrbek, T. (2018). An *in silico* comparison of protocols for dated
652 phylogenomics. *Systematic biology*, *67*(4), 633-650.

653 Cunha, T. J., & Giribet, G. (2019). A congruent topology for deep gastropod
654 relationships. *Proceedings of the Royal Society B*, *286*(1898), 20182776.

655 Cruaud, A., Gautier, M., Galan, M., Foucaud, J., Sauné, L., Genson, G., Dubois, E.,
656 Deuve, T., & Rasplus, J. Y. (2014). Empirical assessment of RAD sequencing for
657 interspecific phylogeny. *Molecular biology and evolution*, 31(5), 1272-1274.

658 Dutertre, S., Jin, A. H., Vetter, I., Hamilton, B., Sunagar, K., Lavergne, V., Dutertre, V.,
659 Fry, B. G., Antunes, A., Venter, D. J., Alewood, P. F., & Lewis, R. J. (2014). Evolution
660 of separate predation-and defence-evoked venoms in carnivorous cone snails. *Nature*
661 *communications*, 5, 3521.

662 Eaton, D. A., Hipp, A. L., González-Rodríguez, A., & Cavender-Bares, J. (2015).
663 Historical introgression among the American live oaks and the comparative nature of
664 tests for introgression. *Evolution*, 69(10), 2587-2601.

665 Ebersberger, I., Strauss, S., & von Haeseler, A. (2009). HaMStR: profile hidden markov
666 model based search for orthologs in ESTs. *BMC evolutionary biology*, 9(1), 157.

667 Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging?.
668 Evolution: *International Journal of Organic Evolution*, 63(1), 1-19.

669 Edwards, S. V. (2016). Phylogenomic subsampling: a brief review. *Zoologica Scripta*,
670 45, 63-74.

671 Espeland, M., Breinholt, J., Willmott, K. R., Warren, A. D., Vila, R., Toussaint, E. F.,
672 Maunsell, S. C., Aduse-Poku, K., Talavera, G., Eastwood, R., Jarzyna, M. A.,
673 Guralnick, R., Lohman, D. J., Pierce, N. E. & Kawahara, A. Y. (2018). A
674 comprehensive and dated phylogenomic analysis of butterflies. *Current Biology*, 28(5),
675 770-778.

676 Faircloth, 2013. *http://s3.ultraconserved.org/talks/faircloth-evolution-2013.pdf*

677 Fedosov, A., Watkins, M., Heralde III, F. M., Corneli, P. S., Concepcion, G. P., &
678 Olivera, B. M. (2011). Phylogeny of the genus Turris: Correlating molecular data with
679 radular anatomy and shell morphology. *Molecular phylogenetics and evolution*, 59(2),
680 263-270.

681 Fedosov, A., Puillandre, N., Herrmann, M., Kantor, Y., Oliverio, M., Dgebuadze, P.,
682 Modica, M. V. & Bouchet, P. (2018). The collapse of Mitra: molecular systematics and
683 morphology of the Mitridae (Gastropoda: Neogastropoda). *Zoological Journal of the*
684 *Linnean Society*, *183*(2), 253-337.

685 Gonzales, D. T. T., & Saloma, C. P. (2014). A bioinformatics survey for conotoxin-like
686 sequences in three turrid snail venom duct transcriptomes. *Toxicon*, 92, 66-74.

687 Gorson, J., Ramrattan, G., Verdes, A., Wright, E. M., Kantor, Y., Rajaram Srinivasan,
688 R., Musunuri, R., Packer, D., Albano, G., & Holford, M. (2015). Molecular diversity
689 and gene evolution of the venom arsenal of terebridae predatory marine snails. *Genome*
690 *biology and evolution*, 7(6), 1761-1778.

691 Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis
692 X., Fan L., Raychowdhury R., Zeng Q., Chen Z.,Mauceli E., Hacohen N., Gnirke A.,
693 Rhind N., Palma F.D., BirrenB.W., Nusbaum C., Lindblad-Toh K., Friedman N., &
694 Regev A. 2011.Full-length transcriptome assembly from RNA-Seq data without a
695 reference genome. *Nat. Biotechnol*. 29:644–652

696 Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Brumfield, R. T. (2016).
697 Sequence capture versus restriction site associated DNA sequencing for shallow
698 systematics. *Systematic biology*, 65(5), 910-924.

699 Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of
700 sequencing DNA. *Genomics*, *107*(1), 1-8.

701 Heralde III, F. M., Watkins, M., Ownby, J. P., Bandyopadhyay, P. K., Santos, A. D.,
702 Concepcion, G. P., & Olivera, B. M. (2007). Molecular phylogeny of some Indo-Pacific
703 genera in the subfamily Turrinae H. Adams and A. Adams, 1853 (1838)(Gastropoda:
704 Neogastropoda). *Nautilus*, *121*(3), 131-138.

705 Heralde FM, Kantor Y, Astilla MAQet al. (2010) The Indo-Pacific Gemmula species in
706 the subfamily Turrinae: aspects offield distribution, molecular phylogeny, radular
707 anatomy and feeding ecology. *Philippine Science Letters,* 3, 21–34

708 Hillis, D. M., Heath, T. A., & John, K. S. (2005). Analysis and visualization of tree
709 space. *Systematic biology*, 54(3), 471-482.

710 Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2017).
711 UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and*
712 *Evolution*, 35(2), 518-522.

713 Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome*
714 *research*, 9(9), 868-877.

715 Jiang, J., Yuan, H., Zheng, X., Wang, Q., Kuang, T., Li, J., Liu, J., Song, S., Wang, W.,
716 Cheng, F., Li, H., Huang, J. & Li, C. (2019). Gene markers for exon capture and
717 phylogenomics in ray-finned fishes. *Ecology and evolution*, 9(7), 3973-3983.

718 Johnson, S. B., Warén, A., Lee, R. W., Kano, Y., Kaim, A., Davis, A., Strong, E. E. &
719 Vrijenhoek, R. C. (2010). Rubyspira, new genus and two new species of bone-eating
720 deep-sea snails with ancient habits. *The Biological Bulletin*, 219(2), 166-177.

721 Jones, N. (2018). How to stop data centres from gobbling up the world's electricity.
722 *Nature*, 561(7722), 163.

723 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., & Jermiin, L. S.
724 (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature*
725 *methods*, 14(6), 587.

726 Katoh, K., Kuma, K. I., Toh, H., & Miyata, T. (2005). MAFFT version 5: improvement
727 in accuracy of multiple sequence alignment. *Nucleic acids research*, 33(2), 511-518.

728 Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2.
729 *Nature methods*, 9(4), 357.

730 Laumer, C. E. (2018). Inferring ancient relationships with genomic data: a commentary
731 on current practices. *Integrative and comparative biology*, 58(4), 623-639.

732 Leaché, A. D., Chavez, A. S., Jones, L. N., Grummer, J. A., Gottscho, A. D., & Linkem,
733 C. W. (2015a). Phylogenomics of phrynosomatid lizards: conflicting signals from
734 sequence capture versus restriction site associated DNA sequencing. *Genome biology*
735 *and evolution*, *7*(3), 706-719.

736 Leaché, A. D., Banbury, B. L., Felsenstein, J., De Oca, A. N. M., & Stamatakis, A.
737 (2015b). Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for
738 inferring SNP phylogenies. *Systematic biology*, 64(6), 1032-1047.

739 Leaché, A. D., & Oaks, J. R. (2017). The utility of single nucleotide polymorphism
740 (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *48*,
741 69-84.

742 Lee, K. M., Kivelä, S. M., Ivanov, V., Hausmann, A., Kaila, L., Wahlberg, N., &
743 Mutanen, M. (2018). Information Dropout Patterns in Restriction Site Associated DNA
744 Phylogenomics and a Comparison with Multilocus Sanger Data in a Species-Rich Moth
745 Genus. *Systematic biology*, 67(6), 925-939.

746 Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment
747 for massively high-throughput phylogenomics. *Systematic biology*, 61(5), 727-744.

748 Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in
749 systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*,
750 44, 99-121.

751 Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large
752 sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), 1658-1659.

753 Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., MarthG., Abecasis
754 G., Durbin R. 2009. The sequence alignment/mapformat and SAMtools. *Bioinformatics*
755 25:2078–9

756 Liu, C., Zhang, Y., Ren, Y., Wang, H., Li, S., Jiang, F., Yin, L., Qiao, X., Zhang, G.,
757 Qian, W., Liu, B., & Fan, W. (2018). The genome of the golden apple snail Pomacea
758 canaliculata provides insight into stress tolerance and invasive adaptation. *GigaScience*,
759 7(9), giy101.

760 Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics*
761 *Hum. Genet.*, *9*, 387-402

762 McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T.,
763 & Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that
764 resolve placental mammal phylogeny when combined with species-tree analysis.
765 *Genome research*, 22(4), 746-754..

766 McFadden, C. S., France, S. C., Sánchez, J. A., & Alderslade, P. (2006). A molecular
767 phylogenetic analysis of the Octocorallia (Cnidaria: Anthozoa) based on mitochondrial
768 protein-coding sequences. *Molecular phylogenetics and evolution*, 41(3), 513-527.

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., & Yang, Y. (2018). Practical considerations for plant phylogenomics. *Applications in plant sciences*, *6*(3), e1038.

Moreau, C. S., & Wray, B. D. (2017). An Empirical Test of Reduced-Representation Genomics to Infer Species-Level Phylogenies for Two Ant Groups. *Insect Systematics and Diversity*, *1*(2).

Mutanen, M., Kivelä, S.M., Vos, R.A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., Huemer, P., Dincă, V., Van Nieukerken, E. J., Lopez-Vaamonde, C., Vila, R., Aarvik. L., Decaëns, T., Efetov, K. A., Hebert, P. D. N., Johnsen, A., Karsholt, O., Pentinsaari, M., Rougerie, R., Segerer, A., Tarmann, G., Zahiri, R., & Godfray, H.C.J. (2016) Species-level para- and polyphyly in DNA barcode gene trees: Strong operational bias in European lepidoptera. *Systematic Biology* 65:1024–1040

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1), 268-274.

Olivera, B. M., Hillyard, D. R., & Watkins, M. (2008). A new species of Gemmula, Weinkauff 1875; Evidence of two clades of Philippine species in the genus Gemmula. *Philipp Sci Lett*, *11*, 11-5.

Philippe, H., Vienne, D. M. D., Ranwez, V., Roure, B., Baurain, D., & Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, 283, 1-25.

790 Phuong, M. A., & Mahardika, G. N. (2018). Targeted sequencing of venom genes from
791 cone snail genomes improves understanding of conotoxin molecular evolution.
792 *Molecular biology and evolution*, 35(5), 1210-1224.

793 Phuong, M. A., Alfaro, M. E., Mahardika, G. N., Marwoto, R. M., Prabowo, R. E., von
794 Rintelen, T., Vogt, P. W. H., Hendricks, J. R., & Puillandre, N. (2019). Lack of signal
795 for the impact of conotoxin gene diversity on speciation rates in cone snails. *Systematic*
796 *biology*.

797 Puillandre, N., Modica, M. V., Zhang, Y., Sirovich, L., Boisselier, M. C., Cruaud, C.,
798 Holford, M., & Samadi, S. (2012). Large-scale species delimitation method for
799 hyperdiverse groups. *Molecular ecology*, 21(11), 2671-2691.

800 Puillandre, N., & Holford, M. (2010). The Terebridae and teretoxins: Combining
801 phylogeny and anatomy for concerted discovery of bioactive compounds. *BMC*
802 *Chemical Biology*, 10(1), 7.

803 Puillandre, N., Fedosov, A. E., Zaharias, P., Aznar-Cormano, L., & Kantor, Y. I.
804 (2017). A quest for the lost types of Lophiotoma (Gastropoda: Conoidea: Turridae):
805 integrative taxonomy in a nomenclatural mess. *Zoological Journal of the Linnean*
806 *Society*, 181(2), 243-271.

807 Ruane, S., Raxworthy, C. J., Lemmon, A. R., Lemmon, E. M., & Burbrink, F. T.
808 (2015). Comparing species tree estimation with large anchored phylogenomic and small
809 Sanger-sequenced molecular datasets: an empirical study on Malagasy
810 pseudoxyrhophiine snakes. *BMC evolutionary biology*, 15(1), 221.

811 Shimodaira, H., & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with
812 applications to phylogenetic inference. Molecular biology and evolution, 16(8), 1114-
813 1114.

814 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M.
815 (2015). BUSCO: assessing genome assembly and annotation completeness with single-
816 copy orthologs. *Bioinformatics*, 31(19), 3210-3212.

817 Simmons, M. P., & Freudenstein, J. V. (2011). Spurious 99% bootstrap and jackknife
818 support for unsupported clades. *Molecular Phylogenetics and Evolution*, 61(1), 177-
819 191.

820 Simmons, M. P. (2017). Relative benefits of amino-acid, codon, degeneracy, DNA, and
821 purine-pyrimidine character coding for phylogenetic analyses of exons. *Journal of*
822 *systematics and evolution*, 55(2), 85-109.

823 Todd, J. A., & Rawlings, T. A. (2014). A review of the Polystira clade—the Neotropic's
824 largest marine gastropod radiation (Neogastropoda: Conoidea: Turridae sensu stricto).
825 *Zootaxa*, 3884(5), 445-491.

826 Uribe, J. E., Zardoya, R., & Puillandre, N. (2018). Phylogenetic relationships of the
827 conoidean snails (Gastropoda: Caenogastropoda) based on mitochondrial genomes.
828 *Molecular phylogenetics and evolution*, 127, 898-906.

829 Washburn, J. D., Schnable, J. C., Conant, G. C., Brutnell, T. P., Shao, Y., Zhang, Y.,
830 Ludwig, M., Davidse, G., & Pires, J. C. (2017). Genome-Guided Phylo-Transcriptomic
831 Methods and the Nuclear Phylogenetic Tree of the Paniceae Grasses. *Scientific reports*,
832 *7*(1), 13528.

833 Zapata, F., Wilson, N. G., Howison, M., Andrade, S. C., Jörger, K. M., Schrödl, M.,
834 Goetz, F. E., Giribet, G., & Dunn, C. W. (2014). Phylogenomic analyses of deep
835 gastropod relationships reject Orthogastropoda. *Proceedings of the Royal Society B:*
836 *Biological Sciences*, 281(1794), 20141739.

837 Zhang, W., Dasmahapatra, K. K., Mallet, J., Moreira, G. R., & Kronforst, M. R. (2016).
838 Genome-wide introgression among distantly related Heliconius butterfly species.
839 *Genome biology*, 17(1), 25.

840 Zhang, C., Sayyari, E., & Mirarab, S. (2017, October). ASTRAL-III: Increased
841 scalability and impacts of contracting low support branches. In *RECOMB International*
842 *Workshop on Comparative Genomics* (pp. 53-75). Springer, Cham.

## Figure legends

**Figure 1.** Proportion of articles over time that used Sanger sequencing (dark grey) or HTS (light grey) to reconstruct a phylogeny. Articles were extracted using the Web of Science "Basic Search", every two years from 2006 to 2018, and using the keyword "Phylogen*" in TITLE only. Only the first 50 articles of the list with newly produced genetic data with one of the two methods were screened and categorized as "Sanger" or "HTS".

**Figure 2.** Flowchart summarizing the in-silico approach used to generate all the datasets. Data are framed by parallelograms, tasks by rectangles and datasets by rounded rectangles.

**Figure 3.** From top to down and left to right: phylogenetic trees corresponding to the BC, SAN, MT, Cap-IQ16, Ref-IQ16 and UPh-IQ4 datasets. Outgroups are not shown. Bootstrap values for the fully supported and intraspecies nodes are not shown. Colors represent genera or genera-level groups. Scale: average number of substitutions per site.

**Figure 4.** Distribution of quartet distance of single-locus trees of the Ref-AS16 dataset against the Ref-AS16 supertree, for the BUSCO loci (dark grey) and the other loci (light grey). (a) Total number of counts, with indication of some specific loci (e.g. 28S) distance to supertree (arrows). (b) Scaled density plot, with dotted lines representing the mean values.

**Figure 5.** Number of unique and shared transcripts for the Ref, UPh and BUSCO sets of loci recovered after the first blast step of each pipeline. Total number of transcripts for all transcriptomes is 3,634,333 (supplementary Table 1).

# Tables

**Table 1.** Description of the datasets analyzed. Me = median loci length. For the robustness evaluation, only the nodes between the ingroup to the species nodes were taken into account. More details on time and money evaluation is available in Supplementary Table 4.

**Table 2.** Summary table of the Shimodaira-Hasegawa tests for each dataset constrained with each topology, with 1,000 resamplings using the RELL method. The topologies are on the top (as column headers) and the datasets on the side (as row headers). “+”: the corresponding topology is not rejected; “-”: vice-versa.

# Supplementary Material

**Supplementary Table 1.** Description of the specimens and transcriptomes.

**Supplementary Table 2.** Correlation table between different sequencing and assembly results

**Supplementary Table 3.** Quartet scores for ASTRAL-III datasets

**Supplementary Table 4.** Evaluation of the costs (time and money) for each dataset.

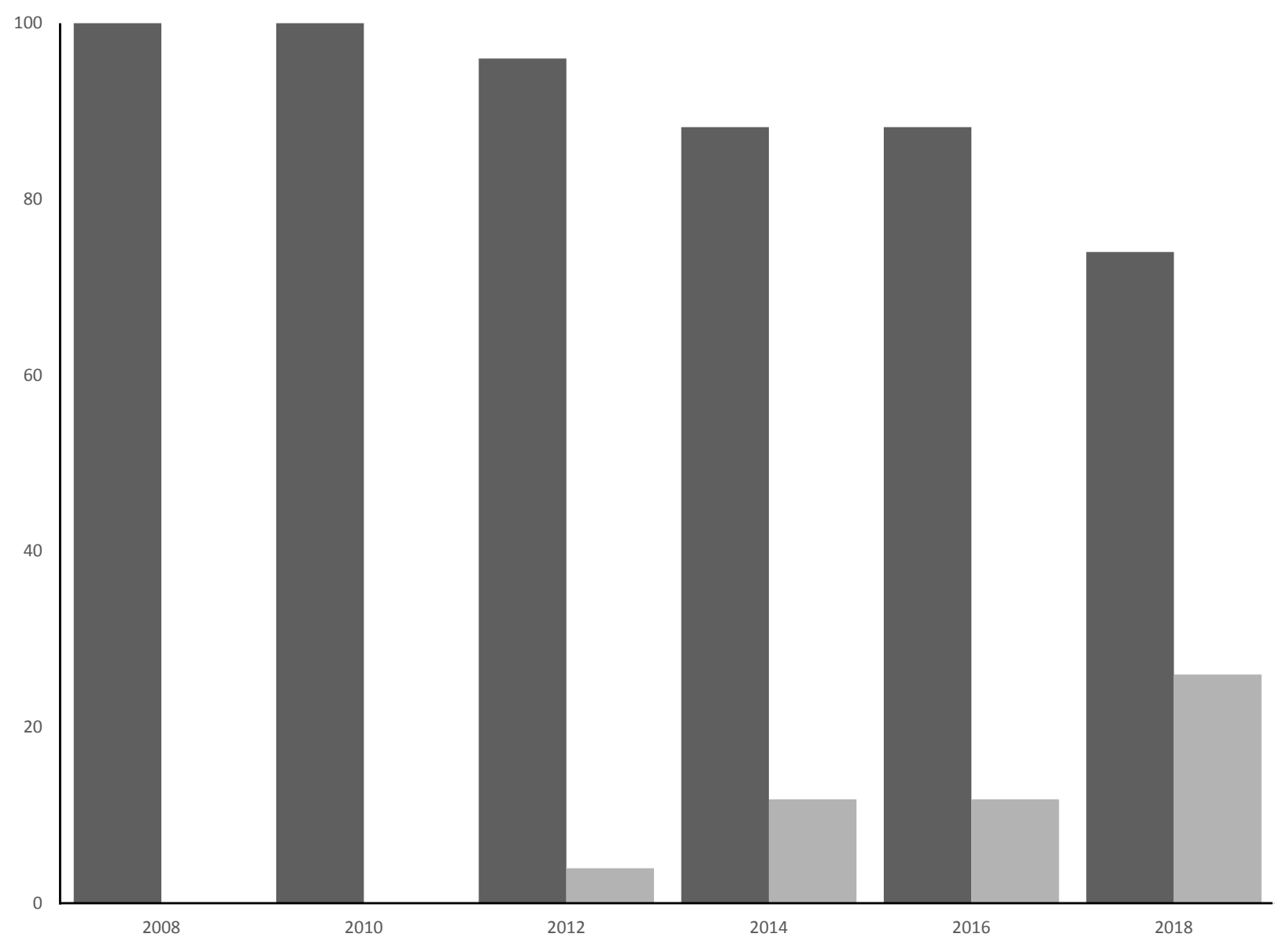
**Supplementary Table 5.** Correlation coefficient of single-loci's quartet distance against several alignment statistics.

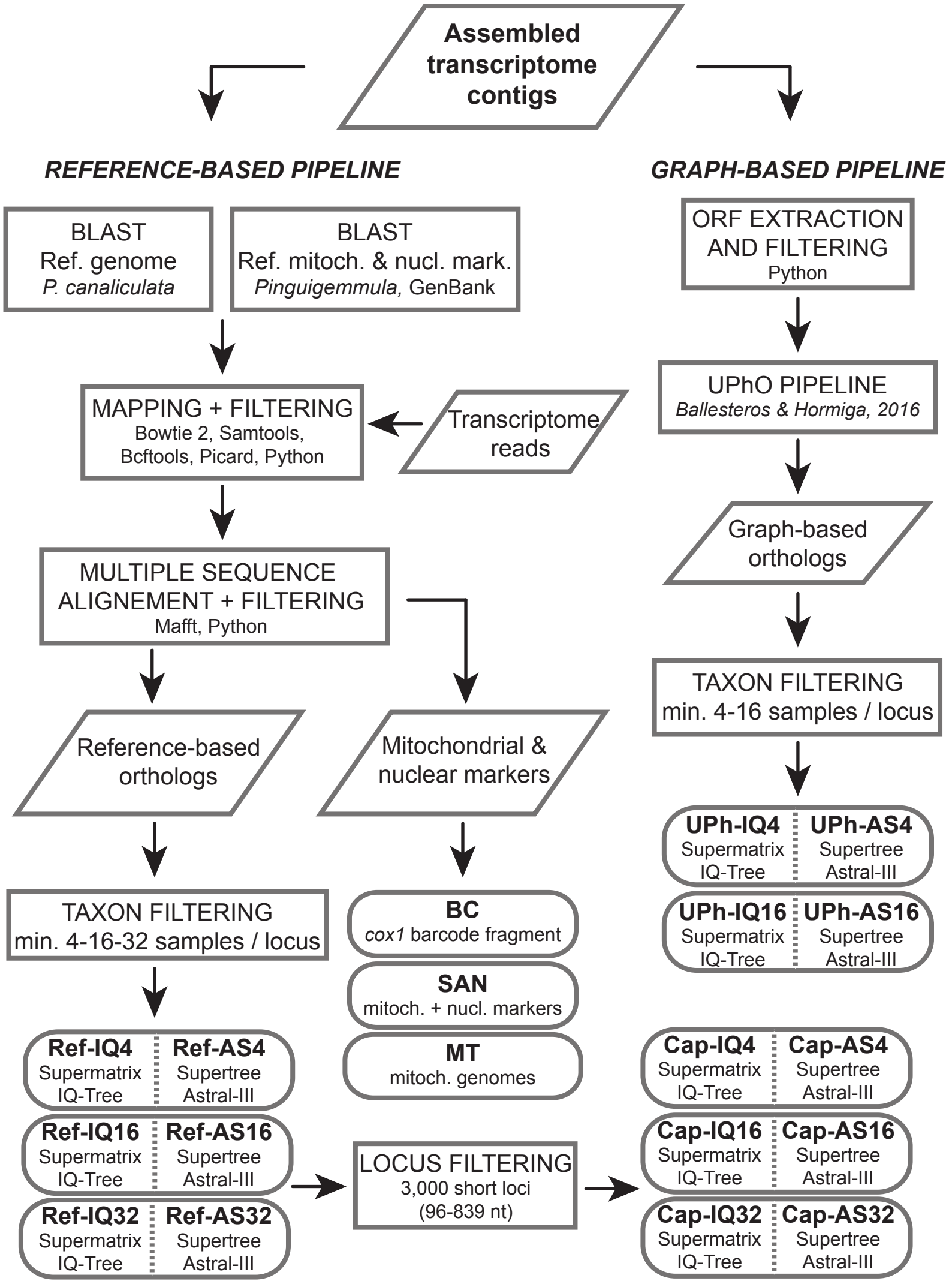
**Supplementary Figure 1.** 20 species tree produced for this study.

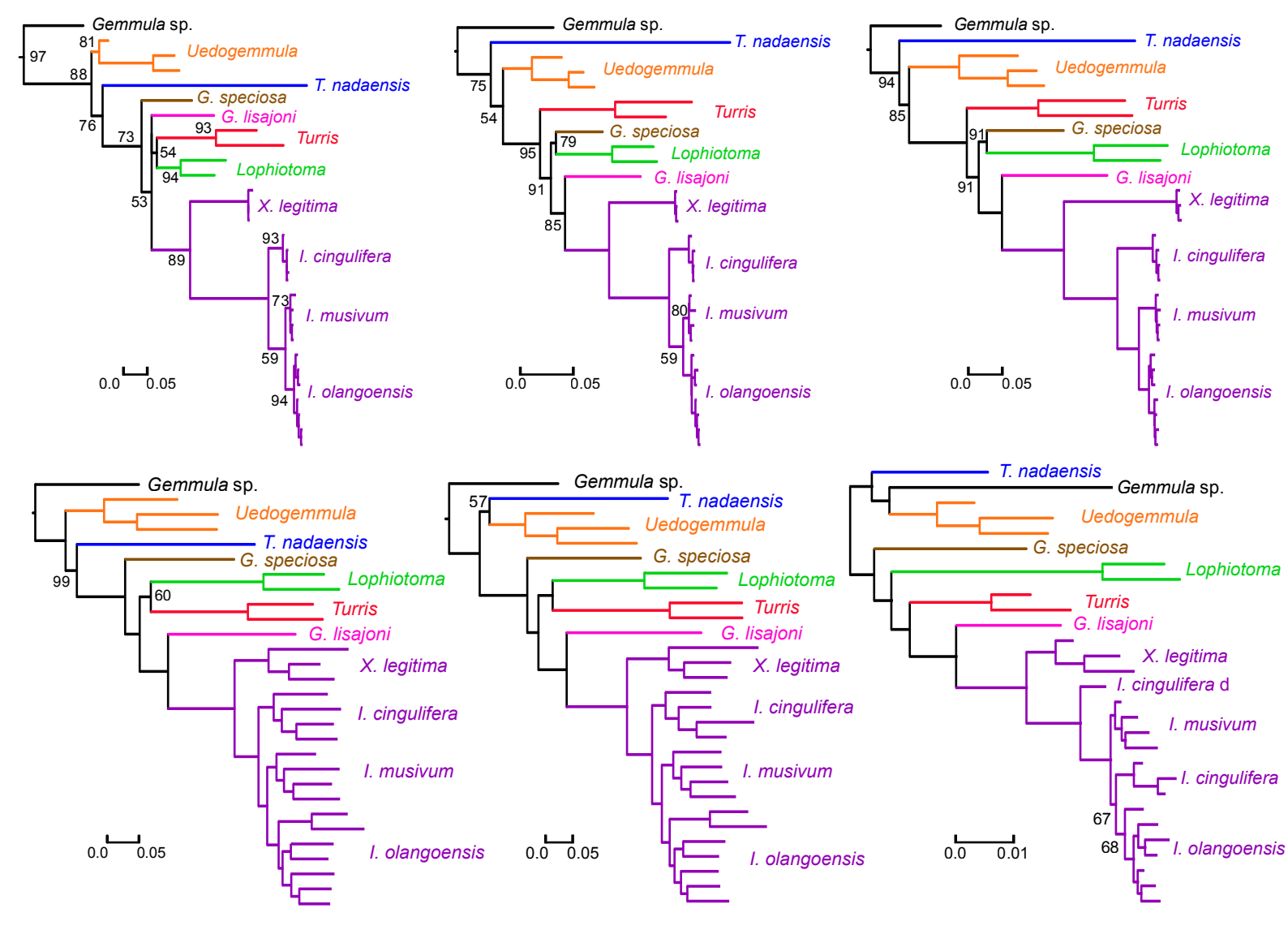
**Supplementary Figure 2.** Distribution of quartet distance of single-locus trees of the UPh-AS16 dataset against the UPh-AS16 supertree

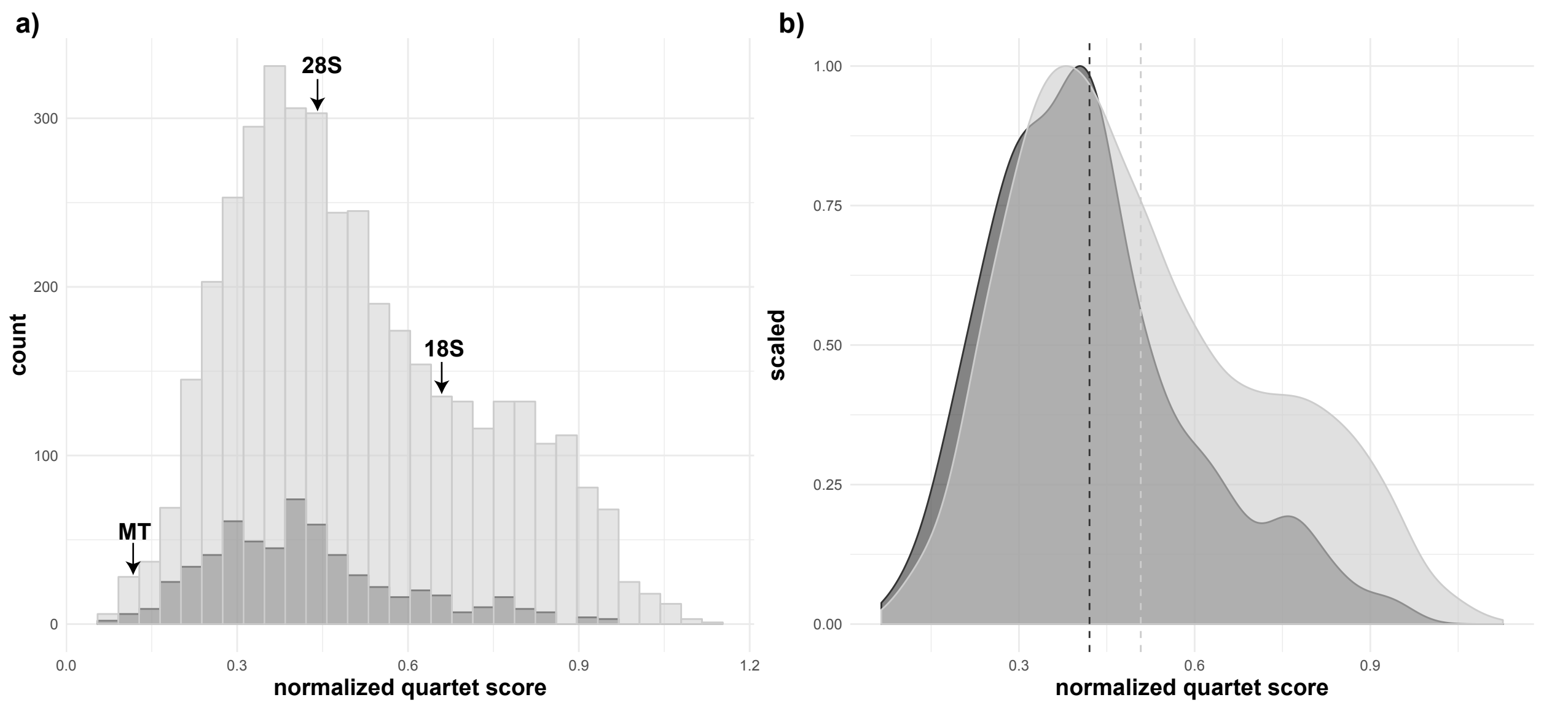
| | | | DATA | | | | | | TIME (days) | | MONEY (euros) | | | ROBUSTNESS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **dataset type** | **orthology assesment** | **Phylogenetic method** | **Dataset name** | **No of loci** | **Alignement length** | **Missing data** | **Variable sites** | **Parsimony informative sites** | **Lab work** | **Data analysis** | **Cost per specimen** | **Cost per base (per specimen)** | **Cost per variable site (per specimen)** | **% nodes > 80% BS or 95 PP*** | **% nodes = 1*** |
| Sanger - DNA barcoding gene | | ML (IQ-TREE) | BC | 1 | 658 | 76 (0.4%) | 258 (39.2%) | 165 (29%) | 2 | - | 7 | 0,011 | 0,027 | 61.1 | 16.6 |
| Sanger - multilocus | | ML (IQ-TREE) | SAN | 6 | 4,787 | 12,820 (8.4%) | 889 (18.6%) | 565 (11.8%) | 2 | - | 40 | 0,01 | 0,045 | 77.7 | 55.5 |
| mitogenome | | ML (IQ-TREE) | MT | 1 | 14,927 | 27,562 (5.8%) | 6,491 (43.5%) | 4,922 (33%) | 5 | 1 | 54 | 0,0036 | 0,0085 | 100 | 77.7 |
| sequence capture | Genome reference | ML (IQ-TREE) | Cap-IQ32 | 274 | 136,799 | 249,086 (5.7%) | 46,491 (34%) | 28,083 (20.5%) | 10 | 6 to 10 | 196 | 0,001432759 | 0,00445 | 100 | 83.3 |
| | | Supertree (ASTRAL-III) | Cap-AS32 | 274 | Me = 498 | Me = 402.5 (2.6%) | Me = 165.5 (32.6%) | Me = 95.5 (18.6%) | | | | | | 77.7 | 77.7 |
| | | ML (IQ-TREE) | Cap-IQ16 | 1373 | 743,778 | 8,009,019 (33.6%) | 266,325 (35.8%) | 148,171 (19.9%) | | | | 0,000263519 | 0,00078 | 100 | 88.8 |
| | | Supertree (ASTRAL-III) | Cap-AS16 | 1373 | Me = 548 | Me = 1,901 (15.2%) | Me = 182 (34%) | Me = 92 (17.5%) | | | | | | 94.4 | 83.3 |
| | | ML (IQ-TREE) | Cap-IQ4 | 3000 | 1,623,052 | 31,758,137 (61.1%) | 499,798 (30.8%) | 218,629 (13.5%) | | | | 0,00012076 | 0,00041 | 94.4 | 94.4 |
| | | Supertree (ASTRAL-III) | Cap-AS4 | 2999 | Me = 555 | Me = 1,491 (26.4%) | Me = 154 (29.9%) | Me = 56 (11.8%) | | | | | | 94.4 | 88.8 |
| transcriptomes | Genome reference | ML (IQ-TREE) | Ref-IQ32 | 473 | 480,293 | 2,533,447 (16.5%) | 158,798 (33.1%) | 91,619 (19.1%) | 8 | 20 to 40 | 275 | 0,000572567 | 0,00071 | 94.4 | 94.4 |
| | | Supertree (ASTRAL-III) | Ref-AS32 | 473 | Me = 698 | Me = 1,046 (4.2%) | Me = 239 (31.8%) | Me = 139 (17.8%) | | | | | | 88.8 | 88.8 |
| | | ML (IQ-TREE) | Ref-IQ16 | 4663 | 8,187,363 | 153,998,814 (58.8%) | 2,450,395 (29.9%) | 1,147,534 (14%) | | | | 0,00003359 | 0,000046 | 94.4 | 94.4 |
| | | Supertree (ASTRAL-III) | Ref-AS16 | 4663 | Me = 1,276 | Me = 9,438 (34.6%) | Me = 409 (31.2%) | Me = 183 (14.8%) | | | | | | 94.4 | 94.4 |
| | | ML (IQ-TREE) | Ref-IQ4 | 9232 | 14,586,607 | 334,525,406 (71.7%) | 3,832,278 (26,3%) | 1,465,372 (10%) | | | | 0,00001885 | 0,000029 | 94.4 | 94.4 |
| | | Supertree (ASTRAL-III) | Ref-AS4 | 9232 | Me = 1,173 | Me = 5,877.5 (42.9%) | Me = 314 (27.7%) | Me = 100 (9.6%) | | | | | | 94.4 | 94.4 |
| | UPhO | ML (IQ-TREE) | Uph-IQ16 | 347 | 245,095 | 2,812,587 (35.9%) | 43,022 (17.6%) | 20,211 (8.2%) | | 30 to 50 | | 0,001122014 | 0,0026 | 88.8 | 88.8 |
| | | Supertree (ASTRAL-III) | Uph-AS16 | 345 | Me = 618 | Me = 0 (0%) | Me = 88 (14.3%) | Me = 41 (6.7%) | | | | | | NA (84.2) | NA(78.9) |
| | | ML (IQ-TREE) | Uph-IQ4 | 7313 | 6,681,038 | 170,796,960 (79.9%) | 1,165,551 (17.4%) | 368,737 (5.5%) | | | | 0,000041161 | 0,000097 | 88.8 | 88.8 |
| | | Supertree (ASTRAL-III) | Uph-AS4 | 7058 | Me = 645 | Me = 2 (0%) | Me = 82 (11.6%) | Me = 16 (2.2%) | | | | | | NA (73.7) | NA (73.7) |

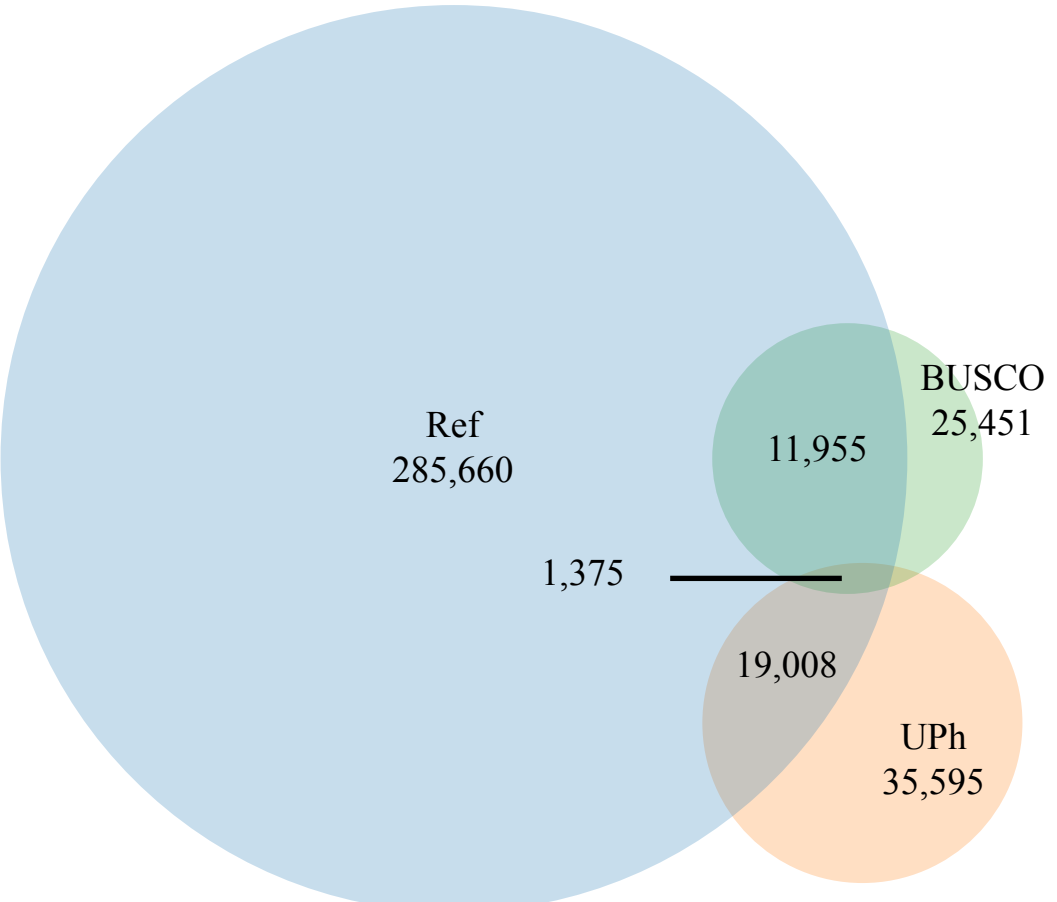
| **Dataset / Topology** | BC | SAN | MT | Cap-IQ32 | Cap-IQ16 | Cap-IQ4 | Ref-IQ32 | Ref-IQ16 | Ref-IQ4 | Uph-IQ16 | Uph-IQ4 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BC | | + | + | + | + | + | + | + | + | + | - |
| SAN | + | | + | + | + | + | + | + | + | + | - |
| MT | + | + | | + | + | + | + | + | + | + | - |
| Cap-IQ32 | - | - | - | | + | + | + | + | + | + | - |
| Cap-IQ16 | - | - | - | + | | + | + | + | + | + | - |
| Cap-IQ4 | - | - | - | + | + | | + | + | + | + | - |
| Ref-IQ32 | - | - | - | + | + | + | | - | - | + | - |
| Ref-IQ16 | - | - | - | + | + | + | + | | + | - | - |
| Ref-IQ4 | - | - | - | + | + | + | + | + | | - | - |
| Uph-IQ16 | - | + | + | + | + | + | + | + | + | | - |
| Uph-IQ4 | - | - | - | - | - | - | - | - | - | - | |


100
80
60
40
20
0
2008
2010
2012
2014
2016
2018


Assembled transcriptome contigs
REFERENCE-BASED PIPELINE
GRAPH-BASED PIPELINE
BLAST
Ref. genome
P. canaliculata
BLAST
Ref. mitoch. & nucl. mark.
Pinguigemmula, GenBank
ORF EXTRACTION AND FILTERING
Python
UPhO PIPELINE
Ballesteros & Hormiga, 2016
MAPPING + FILTERING
Bowtie 2, Samtools, Bcftools, Picard, Python
Transcriptome reads
Graph-based orthologs
MULTIPLE SEQUENCE ALIGNEMENT + FILTERING
Mafft, Python
TAXON FILTERING
min. 4-16 samples / locus
Reference-based orthologs
Mitochondrial & nuclear markers
UPh-IQ4
Supermatrix
IQ-Tree
UPh-AS4
Supertree
Astral-III
UPh-IQ16
Supermatrix
IQ-Tree
UPh-AS16
Supertree
Astral-III
TAXON FILTERING
min. 4-16-32 samples / locus
BC
cox1 barcode fragment
SAN
mitoch. + nucl. markers
MT
mitoch. genomes
Ref-IQ4
Supermatrix
IQ-Tree
Ref-AS4
Supertree
Astral-III
Ref-IQ16
Supermatrix
IQ-Tree
Ref-AS16
Supertree
Astral-III
Ref-IQ32
Supermatrix
IQ-Tree
Ref-AS32
Supertree
Astral-III
LOCUS FILTERING
3,000 short loci
(96-839 nt)
Cap-IQ4
Supermatrix
IQ-Tree
Cap-AS4
Supertree
Astral-III
Cap-IQ16
Supermatrix
IQ-Tree
Cap-AS16
Supertree
Astral-III
Cap-IQ32
Supermatrix
IQ-Tree
Cap-AS32
Supertree
Astral-III


Gemmula sp.
Uedogemmula
T. nadaensis
G. speciosa
G. lisajoni
Turris
Lophiotoma
X. legitima
I. cingulifera
I. musivum
I. olangoensis
97
81
88
76
73
93
54
94
53
89
73
59
94
75
54
95
79
91
85
80
59
94
85
91
91
57
99
60
0.0
0.05
0.01


a)
28S
18S
MT
count
300
200
100
0
0.0
0.3
0.6
0.9
1.2
normalized quartet score
b)
scaled
1.00
0.75
0.50
0.25
0.00
0.3
0.6
0.9
normalized quartet score


Ref
285,660
BUSCO
25,451
11,955
1,375
19,008
UPh
35,595