PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies

Dong Zhang[1,2], Fangluan Gao[3], Wen X. Li[1], Ivan Jakovlić[4], Hong Zou[1], Jin Zhang[4] and Gui T. Wang[1,*]

[1]Key Laboratory of Aquaculture Disease Control, Ministry of Agriculture, and State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, P. R. China,

[2]University of Chinese Academy of Sciences, Beijing, P. R. China,

[3]Institute of Plant Virology, Fujian Agriculture and Forestry University, Fuzhou 350002, Fujian, P. R. China,

[4]Bio-Transduction Lab, Wuhan Institute of Biotechnology, Wuhan, P. R. China

*** Corresponding author:** gtwang@ihb.ac.cn (GTW)

Abstract

Multi-gene and genomic datasets have become commonplace in the field of phylogenetics, but many of the existing tools are not designed for such datasets, which makes the analysis time-consuming and tedious. We therefore present PhyloSuite, a user-friendly workflow desktop platform dedicated to streamlining molecular sequence data management and evolutionary phylogenetics studies. It employs a plugin-based system that integrates a number of useful phylogenetic and

bioinformatic tools, thereby streamlining the entire procedure, from data acquisition to phylogenetic tree annotation, with the following features: (i) point-and-click and drag-and-drop graphical user interface, (ii) a workspace to manage and organize molecular sequence data and results of analyses, (iii) GenBank entries extraction and comparative statistics, (iv) a phylogenetic workflow with batch processing capability, (v) elaborate bioinformatic analysis for mitochondrial genomes. The aim of PhyloSuite is to enable researchers to spend more time playing with scientific questions, instead of wasting it on conducting standard analyses. The compiled binary of PhyloSuite is available under the LGPL license at https://github.com/dongzhang0725/PhyloSuite/releases, implemented in Python and runs on Windows, Mac OSX and Linux.

Introduction

Advancements in next-generation sequencing technologies (Metzker, 2009) have resulted in a huge increase in the amount of genetic data available through public databases. While this opens a multitude of research possibilities, retrieving and managing such large amounts of data may be difficult and time-consuming for researchers who are not computer-savvy. A standard analytical procedure for phylogenetic analysis is: selecting and downloading GenBank entries, extracting target genes (for multi-gene datasets, such as organelle genomes) and/or mining other data, sequence alignment, alignment optimization, concatenation of alignments (for multi-gene datasets), selection of best-fit partitioning schemes and evolutionary

models, phylogeny reconstruction, and finally visualization and annotation of the phylogram. This can be very time-consuming if different programs have to be employed for different steps, especially as they often have different input file format requirements, and sometimes even require manual file tweaking. Therefore, multifunctional, workflow-type software packages are becoming increasingly needed by a broad range of evolutionary biologists (Smith, 2015). Specifically, as single-gene datasets are rapidly being replaced by multi-gene or genomic datasets as a tool of choice for phylogenetic reconstruction (Degnan and Rosenberg, 2009; Rivera-Rivera and Montoya-Burgos, 2016), automated gene extraction from genomic data and batch manipulation in some of the above steps, like alignment, are becoming a necessity.

Although there are several tools in existence, designed to streamline this process by incorporating some or all of the steps mentioned above, none of these incorporate all of the above functions in a manner suitable for current trends in phylogenetic analyses (see detailed comparison in Supplementary data). Therefore, we present PhyloSuite, a versatile tool designed to incorporate all of the functions described above, including a series of different phylogenetic analysis algorithms, into a single workflow that does not require programming skills, has an intuitive graphical user interface (GUI), workspace, batch mode, extensive plugins support, inbuilt updating function, etc. (Fig. 1). This tool aims to be accessible to all scientists, streamline the phylogenetic analysis procedure, and allow scientists to focus on solving scientific questions rather than waste time on toying with different scientific software programs.

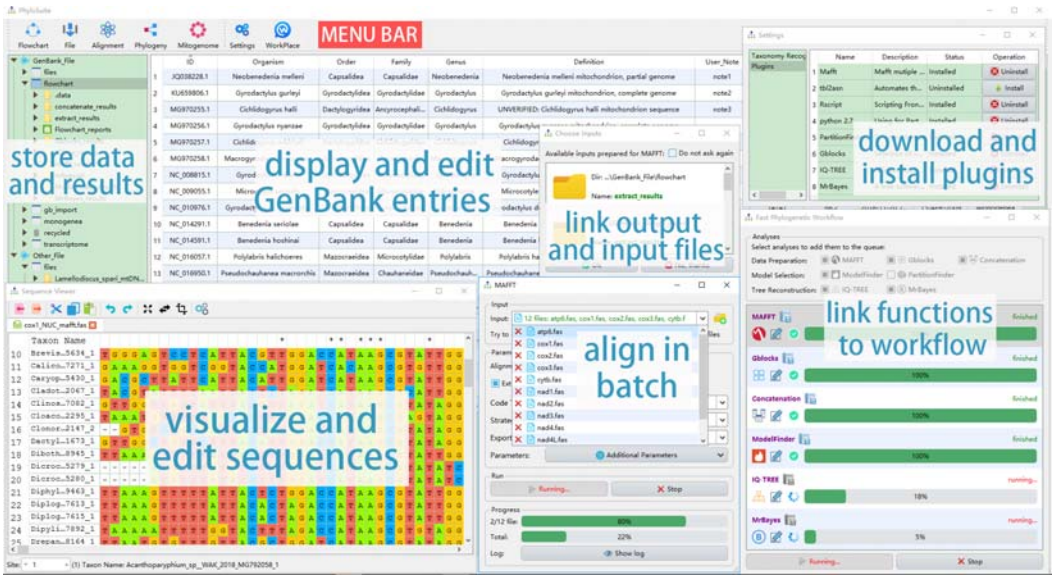

Fig. 1. The interface and the main functions of PhyloSuite

Implementation

PhyloSuite is a user-friendly stand-alone GUI-based software written in Python 3.6.7 and packaged and tested on Windows, Mac OSX and Linux. The functions are (Fig. 1, Supplementary data): (i) retrieving, extracting, organizing and managing molecular sequence data, including GenBank entries, nucleotide and amino acid sequences, and sequences annotated in Word documents; (ii) batch alignment of sequences with MAFFT (Katoh and Standley, 2013), for which we added a codon alignment (translation align) mode; (iii) batch optimization of ambiguously aligned regions using Gblocks (Talavera and Castresana, 2007); (iv) batch conversion of alignment formats (FASTA, PHYLIP, PAML, AXT and NEXUS); (v) concatenation of multiple alignments into a single dataset and preparation of a partition file for downstream analyses; (vi) selection of the best-fit evolutionary model and/or partitioning scheme using ModelFinder (Kalyaanamoorthy, et al., 2017) or PartitionFinder (Lanfear, et al., 2017); (vii) phylogeny reconstruction using IQ-TREE (maximum likelihood) (Nguyen,

et al., 2015) and/or MrBayes (Bayesian inference) (Ronquist, et al., 2012); (viii) linking the functions from (ii) to (vii) into a workflow; (ix) annotating phylogenetic trees in the iTOL webtool (Letunic and Bork, 2016) using datasets generated by the (i) function; (x) comprehensive bioinformatic analysis of mitochondrial genomes (mitogenomes); (xi) visualization and editing of sequences using a MEGA-like sequence viewer; (xii) storing, organizing and visualizing data and results of each analysis in the PhyloSuite workspace.

Genetic data management

PhyloSuite provides a flexible GenBank entries processing function (see Supplementary data). GenBank files can be imported either directly, or via a list of IDs, which PhyloSuite will automatically download from the GenBank. Almost all of the information in the annotation section of a GenBank record can be extracted and displayed in the GUI. Additionally, the information can be standardized in batch using a corresponding function or edited manually in the GUI, ambiguously annotated mitogenomic tRNA genes can be semi-automatically reannotated using ARWEN (Laslett and Canback, 2008), and taxonomic data can be automatically retrieved from WoRMS and NCBI Taxonomy databases. The ‘extract’ function allows users to extract genes in batches, as well as generate an assortment of statistics and dataset files (iTOL datasets). The extracted results can be used for downstream analyses without additional manipulation. The nucleotide and amino acid sequences can be visualized and edited in a MEGA-like explorer equipped with common functions (reverse complement, etc.). Importantly, PhyloSuite can parse the sequence

annotations recorded in a Word document via the inbuilt 'comment' function, and generate a GenBank file and an *.sqn file for direct submission to the GenBank. This function provides a novel and simple way to annotate genetic sequences, which shall benefit researchers who are not computer-savvy.

Phylogenetic analysis workflow

By allowing users to combine seven plugin programs/functions and execute them sequentially, PhyloSuite streamlines the evolutionary phylogenetics analysis (see Supplementary data). The standard execution order of these functions is: MAFFT, Gblocks, Concatenation, ModelFinder or PartitionFinder2, MrBayes and/or IQ-TREE. The results of upstream functions are directly prepared as the input for downstream functions, so only the first function of each workflow requires an input file(s). Functions can also be used in a non-standard order and/or separately, in which case PhyloSuite will automatically search for available input files (results of other tools) in the workspace. Before starting the workflow, PhyloSuite will summarize the parameters of each function, allowing the user to check and modify them, or autocorrect conflicting parameters, such as sequence types. Once a workflow is finished, PhyloSuite will describe the settings of each function as well as present the references for each plugin program in the GUI.

Bioinformatics analysis for mitogenomic data

PhyloSuite was originally designed for, and its major comparative advantages are in, the mitochondrial genomics analyses. There is a specialized configuration available for the extraction of mitogenomic features. In addition to gene extraction, PhyloSuite

will generate a dozen of statistics and dataset files useful for downstream analyses (see Supplementary data). The 'itol' dataset can be used to annotate the obtained phylogram (colorize lineages, map gene orders, etc.). The gene order file can be used to conduct gene order analysis with CREx (Bernt, et al., 2007) or treeREx (Bernt, et al., 2008). The tables generated include the list of mitogenomes and overall statistics, annotation, nucleotide composition and skewness, relative synonymous codon usage (RSCU) and amino acid usage. The RSCU figure (see Fig. 3 in Zhang, et al. (2017)) can be drawn using the RSCU table and 'Draw RSCU figure' function. The annotation table can be used to compare genomic annotations and calculate pairwise similarity of homologous genes with 'Compare table' function (see Table 1 in Zhang, et al. (2018)). In the future, PhyloSuite aims to gradually extend these analyses to other small genomes (organelles, viruses, etc.).

Discussion

PhyloSuite links the management of genetic sequence data and a series of phylogenetic analysis tools, thereby simplifying and speeding up multi-gene based phylogenetic analyses, from data acquisition to phylogram annotation. In summary, highlights of PhyloSuite include: (i) a user-friendly workspace to visualize, organize, manipulate and store sequence data and results; (ii) flexible GenBank entries processing (standardization, reannotation, etc.); (iii) batch data processing capability and workflow; (iv) a state of the art mitogenomic bioinformatics analysis. Although PhyloSuite is designed primarily to allow non-computer-savvy users to drag-and-drop and point-and-click their way through the phylogenetic analysis, experienced

scientists will also find it useful to streamline their research, store and manage results, and increase productivity. It will especially benefit evolutionary biologists who wish to test the effects of different datasets and analytical methods on the phylogenetic reconstruction.

Acknowledgements

The authors would like to thank Dr. Xiao-Qin Xia for modifying the manuscript and Mr. Cheng-En Zheng for technical assistance in the software development.

Funding

This work was supported by the National Natural Science Foundation of China [31872604]; the Earmarked Fund for China Agriculture Research System [CARS-45-15]; and the Major Scientific and Technological Innovation Project of Hubei Province [2015ABA045].

Conflict of Interest: none declared.

References

Bernt, M., Merkle, D. and Middendorf, M. (2008) An algorithm for inferring mitogenome rearrangements in a phylogenetic tree. In Nelson, C.E. and Vialette, S., (eds.), *Comparative Genomics, International Workshop, RECOMB-CG 2008, Proceedings of Lecture Notes in Bioinformatics*. Springer, Berlin, Vol. 5267, pp. 143-157.

Bernt, M., *et al.* (2007) CREx: inferring genomic rearrangements based on common

intervals. *Bioinformatics*, 23(21), 2957-2958.

Degnan, J.H. and Rosenberg, N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*, 24(6), 332-340.

Kalyaanamoorthy, S., *et al.* (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*, 14(6), 587-589.

Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4), 772-780.

Lanfear, R., *et al.* (2017) PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol*, 34(3), 772-773.

Laslett, D. and Canback, B. (2008) ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics*, 24(2), 172-175.

Letunic, I. and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*, 44(W1), W242-245.

Metzker, M.L. (2009) Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1), 31-46.

Nguyen, L.T., *et al.* (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*, 32(1), 268-274.

Rivera-Rivera, C.J. and Montoya-Burgos, J.I. (2016) LS(3): A Method for Improving Phylogenomic Inferences When Evolutionary Rates Are Heterogeneous among Taxa. *Mol Biol Evol*, 33(6), 1625-1634.

Ronquist, F., *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3), 539-542.

Smith, D.R. (2015) Buying in to bioinformatics: an introduction to commercial sequence analysis software. *Brief Bioinform*, 16(4), 700-709.

Talavera, G. and Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, 56(4), 564-577.

Zhang, D., *et al.* (2018) Three new Diplozoidae mitogenomes expose unusual compositional biases within the Monogenea class: implications for phylogenetic studies. *BMC Evol Biol*, 18(1), 133.

Zhang, D., *et al.* (2017) Sequencing of the complete mitochondrial genome of a fish-parasitic flatworm *Paratetraonchoides inermis* (Platyhelminthes: Monogenea): tRNA gene arrangement reshuffling and implications for phylogeny. *Parasites & Vectors*, 10(1), 462.

Supplementary Information

We merged our two older beta tools, MitoTool (https://github.com/dongzhang0725/MitoTool) and BioSuite (https://github.com/dongzhang0725/BioSuite), into a single plugin-based software program designed to streamline the multi-gene evolutionary phylogenetic analysis - PhyloSuite (Fig. S1).

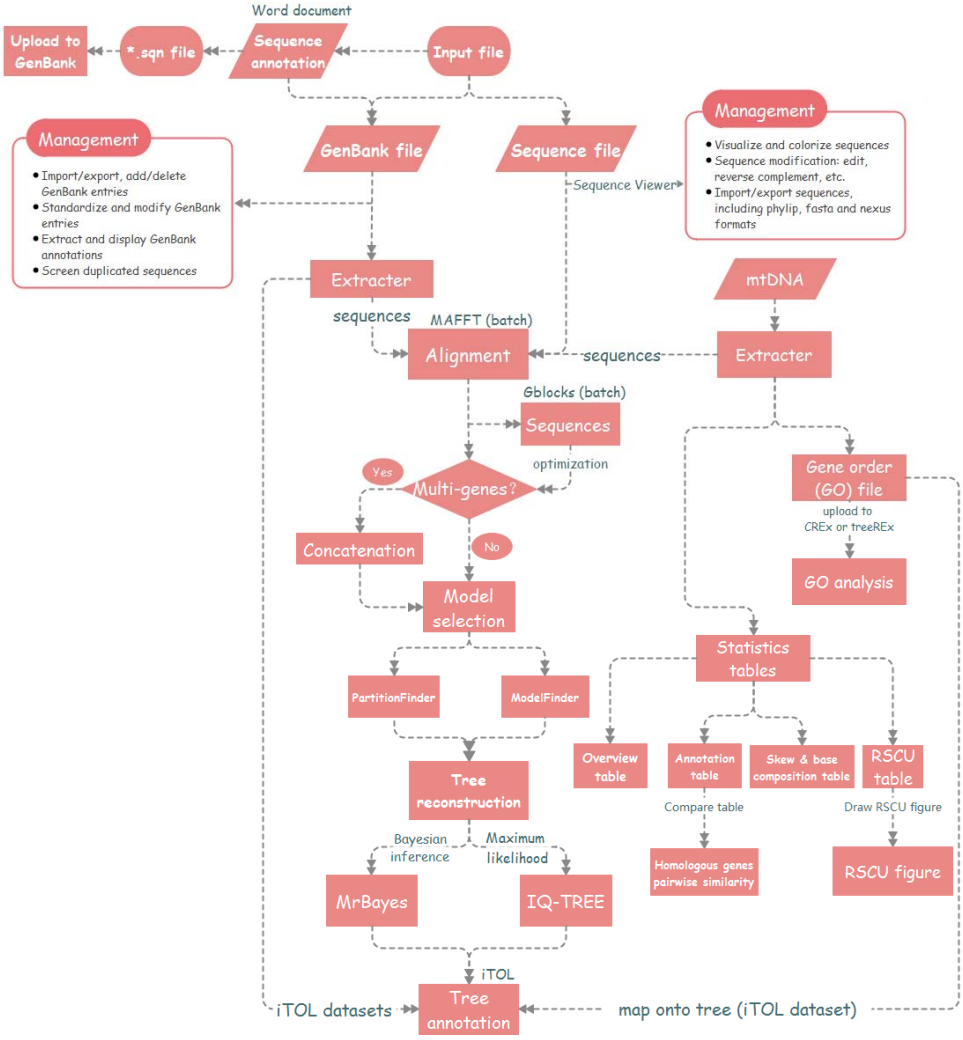

Fig. S1. The workflow diagram of PhyloSuite.

Comparison with extant software programs

Although the extant software programs possess some of the abilities of PhyloSuite, none of them incorporate all functions necessary for a streamlined multi-gene phylogenetic analysis, from data retrieval to the phylogenetic tree annotation (click here to access the attached Table S1). For example, FeatureExtract (Wernersson, 2005) and Geneious (Kearse, et al., 2012) can extract the annotations from GenBank files, but downstream analysis is not fully automated, so some manual data handling is required, especially for multi-gene datasets. Armadillo (Lord, et al., 2012), EPoS (Griebel, et al., 2008) and MEGA (Kumar, et al., 2016) do not possess the batch processing capability, which is indispensable for multi-gene datasets. Additionally, data partitioning and best-fit partitioning scheme estimation are also pivotal for multi-gene dataset-based phylogenetic analyses (Blair and Murphy, 2011; Lanfear, et al., 2012), but most other software programs lack this function, including Geneious, MEGA, Galaxy Workflow (Oakley, et al., 2014), etc. Although, MEGA and EPoS possess the ability to use the output of one tool directly as the input for another tool, they cannot link several functions into a single run (workflow). Probably the closest to meeting the described requirements is Geneious, but this is a commercial bioinformatics software, so it may not be an ideal option (i.e. too expensive) for all scientists, especially for students.

Functions and capabilities

Taking a recently published mitogenomic paper (Zhu, et al., 2018) as an example, using the ‘extract’ function user can quickly conduct most of the analyses reported in

that paper, and generate similar tables and figures: (i) mitogenome list and overall statistics table (Table 1 in Zhu et al.), (ii) annotation table (Table 2), (iii) nucleotide composition and skewness table (Table 3), (iv) relative synonymous codon usage (RSCU) table and figure (Fig. 2B), (v) amino acid usage statistics file used to draw Fig. 2A, and (vi) reconstruct and annotate (in iTOL) phylogenetic trees (Fig. 5) using the extracted genes. In comparison, most of the tables in that paper were made manually by the author, which is time-consuming, tedious and error-prone. Beyond these, several additional analyses are available: (i) gene order file is generated, which can be used to map gene orders of mitogenomes onto the phylograms in iTOL (Fig. S2, also see Fig. 6 in Zhang, et al. (2018)), (ii) statistics for individual genes, including size, start and terminal codons, base composition and skews (click here to access the attached Table S2), (iii) general statistics table, which can be used to draw skewness and base content figure (Fig. S3, also see Fig. 1 in Zhang, et al. (2018)), (iv) comparison of genomic annotations and pairwise similarity calculation for homologous genes, such as Table 1 in Zhang, et al. (2018).

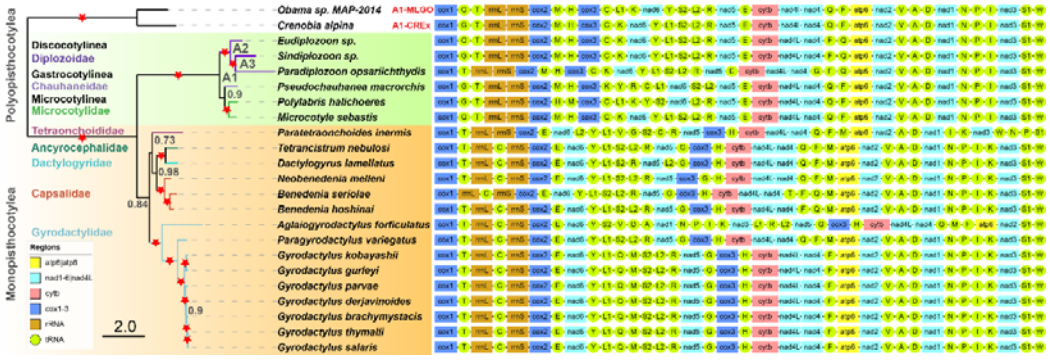

Fig. S2 Mapping gene orders of monogenean mitogenomes onto the phylogenetic tree.

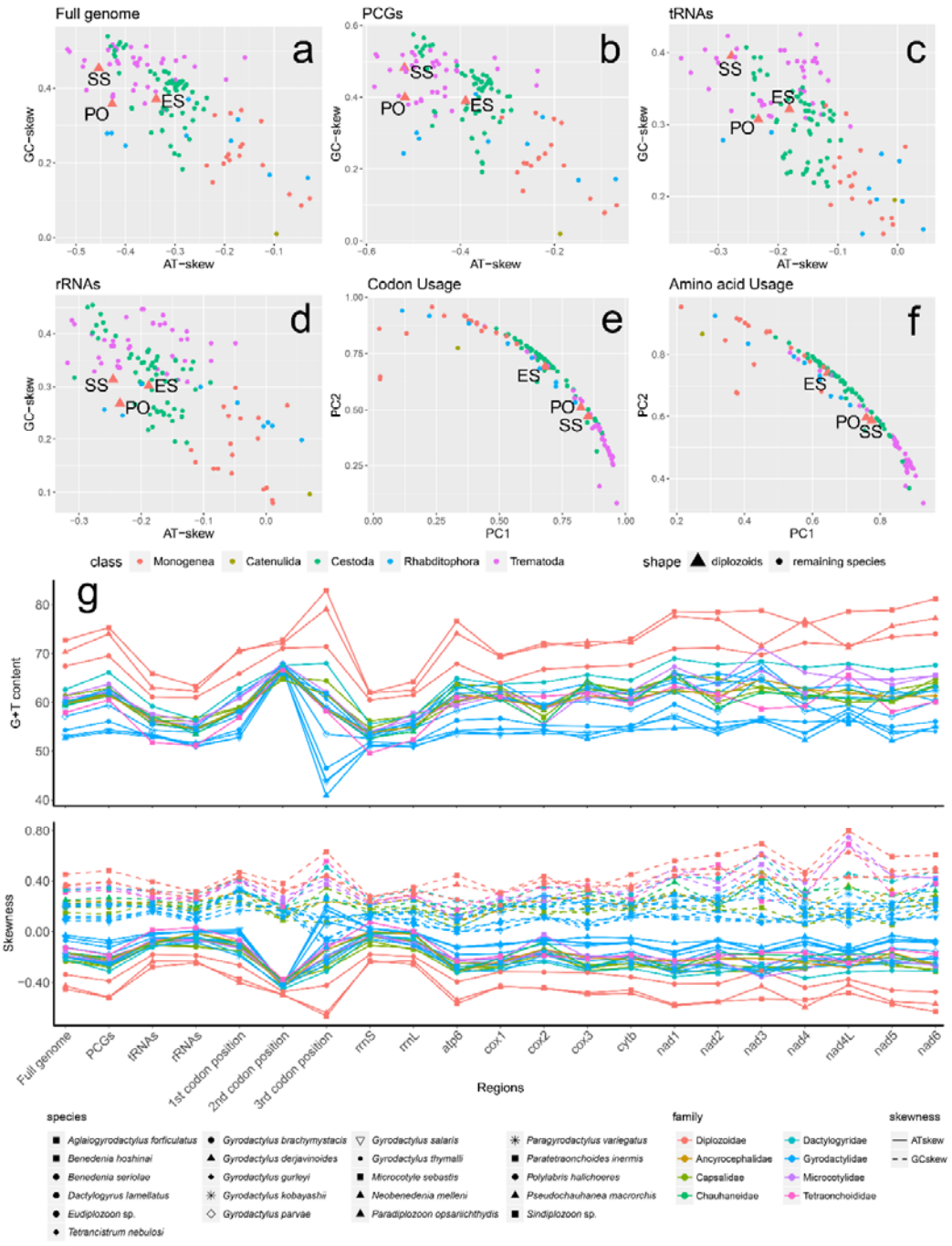

Fig. S3 Skewness and base content of some flatworm mitogenomes.

References

Blair, C. and Murphy, R.W. (2011) Recent trends in molecular phylogenetic analysis:

where to next? *J Hered*, 102(1), 130-138.

Griebel, T., Brinkmeyer, M. and Bocker, S. (2008) EPoS: a modular software framework for phylogenetic analysis. *Bioinformatics*, 24(20), 2399-2400.

Kearse, M., *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647-1649.

Kumar, S., Stecher, G. and Tamura, K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*, 33(7), 1870-1874.

Lanfear, R., *et al.* (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular biology and evolution*, 29(6), 1695-1701.

Lord, E., *et al.* (2012) Armadillo 1.1: an original workflow platform for designing and conducting phylogenetic analysis and simulations. *PLoS One*, 7(1), e29903.

Oakley, T.H., *et al.* (2014) Osiris: accessible and reproducible phylogenetic and phylogenomic analyses within the Galaxy workflow management system. *BMC Bioinformatics*, 15(1), 230.

Wernersson, R. (2005) FeatureExtract--extraction of sequence annotation made easy. *Nucleic Acids Res*, 33(Web Server issue), W567-569.

Zhang, D., *et al.* (2018) Three new Diplozoidae mitogenomes expose unusual compositional biases within the Monogenea class: implications for phylogenetic studies. *BMC Evol Biol*, 18(1), 133.

Zhu, H.F., *et al.* (2018) Complete mitochodrial genome of the crab spider Ebrechtella tricuspidata (Araneae: Thomisidae): A novel tRNA rearrangement and phylogenetic implications for Araneae. *Genomics*.