bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, 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

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 GPL 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

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

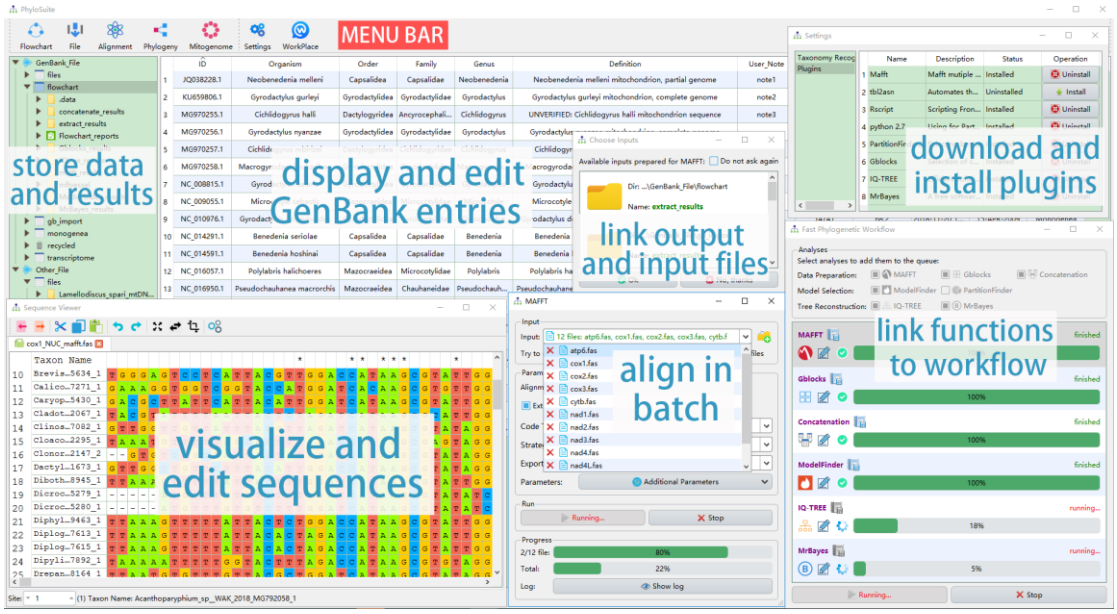

Fig. 1. The interface and the main functions of PhyloSuite

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,

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Information

Overview

PhyloSuite is designed to address the global trend towards multi-gene based phylogenetic analyses (Degnan and Rosenberg, 2009; Rivera-Rivera and Montoya-Burgos, 2016): this software program incorporates and streamlines all steps included in such analyses, from data acquisition to phylogenetic tree visualization and annotation (Fig. S1).

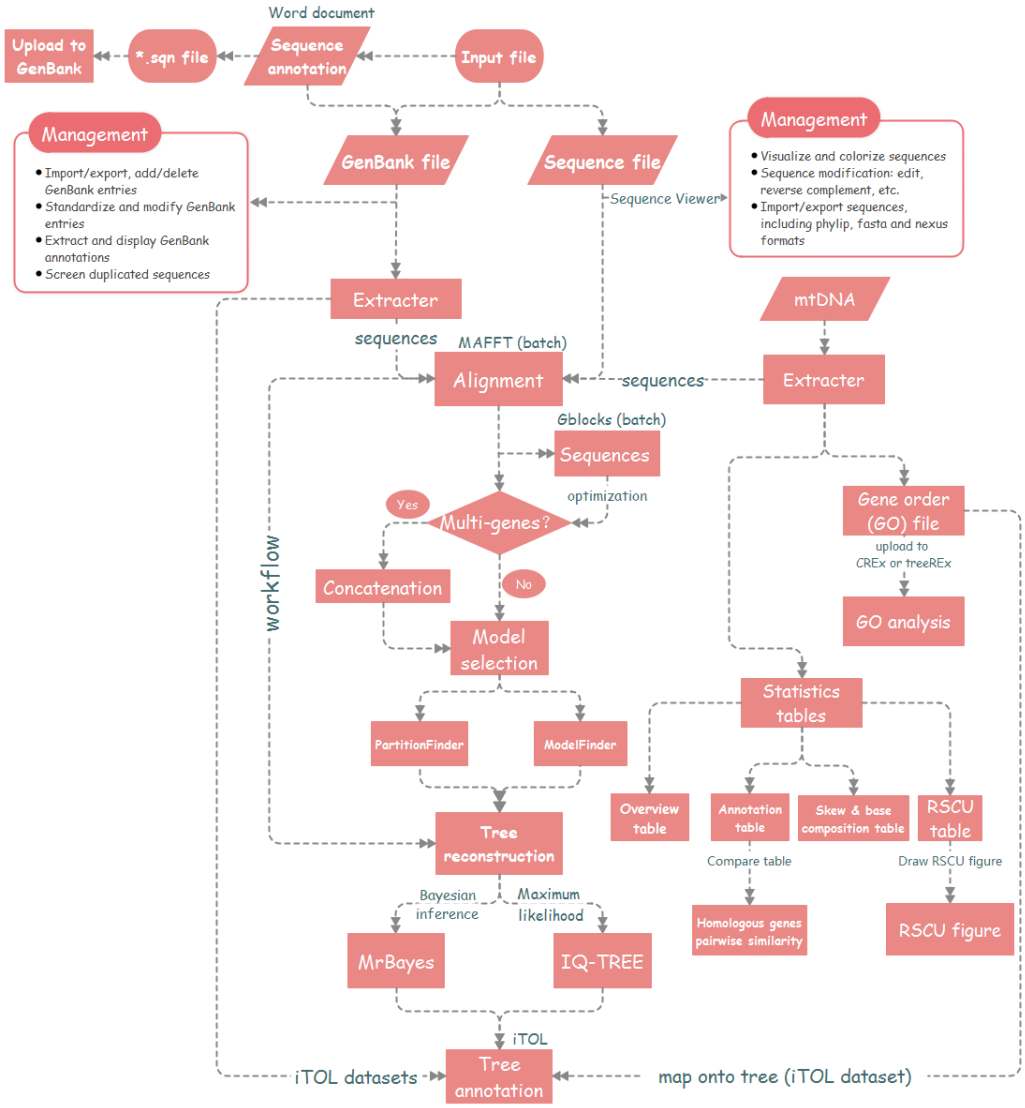

Fig. S1. The workflow diagram of PhyloSuite.

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Fig. S3). 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.

	PhyloSuite	MitoPhAST	HomBlocks	phylogeny.fr	Galaxy Workf	Armadillo	phyloGenerat	Geneious	MEGA	EPoS
Graphical interface/Webpage	√	×	×	√	√	√	×	√	√	√
Alignment	√	√	√	√	√	√	√	√	√	√
Codon alignment	√	√	×	×	×	×	×	√	√	×
Alignment optimization	√	√	√	√	√	×	√	×	×	√
Model selection	√	√	√	√	√	√	×	×	√	×
Maximum likelihood tree	√	√	×	√	√	√	√	√	√	√
Bayesian tree	√	×	×	√	√	√	√	√	×	√
Data partitioning	√	√	√	×	√	×	×	√	×	√
Best partitioning scheme estimate	√	√	√	×	×	×	×	×	×	×
Batch processing	√	√	√	×	√	×	×	√	×	×
Tree annotation	√	×	×	√	×	√	×	√	√	√
Workflow	√	√	√	√	√	√	√	√	√ [a]	√ [a]
Free	√	√	√	√	√	√	√	×	√	√

Fig. S3 Comparison of PhyloSuite with software programs with similar functions.

[a] semi-workflow.

Functions and capabilities

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Fig. S4, Table 1 in Zhu et al.), (ii) annotation table (Fig. S5, Table 2 in Zhu et al.), (iii) nucleotide composition and skewness table (Fig. S6, Table 3 in Zhu et al.), (iv) relative synonymous codon usage (RSCU) table (Fig. S7) and figure (Fig. S8, Fig. 2B in Zhu et al.), (v) amino acid usage statistics file (Fig. S9) used to draw Fig. 2A in Zhu et al., and (vi) reconstruct and annotate (in iTOL) phylogenetic trees (Fig. 5 in Zhu et al.) 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. S10, also see Fig. 6 in Zhang, et al. (2018)) and conduct gene order analysis with CREx (Bernt, et al., 2007) or treeREx (Bernt, et al., 2008), (ii) statistics for individual genes, including size, start and terminal codons, base composition and skews (Fig. S11), (iii) general statistics table, which can be used to draw skewness and base content figure (Fig. S12, also see Fig. 1 in Zhang, et al. (2018)), (iv) comparison of genomic annotations and pairwise similarity calculation for homologous genes (Fig. S13, Table 1 in Zhang, et al. (2018).

Taxon	Accession number	Whole Genome				PCGs			
		Size(bp)	AT%	AT-Skew	GC-Skew	Size(bp)	AT%	AT-Skew	GC-Skew
Araneae									
Arachnida									
Agelenidae									
N/A									
Agelena silvatica	NC_033971.1	14776	74.5	-0.163	0.302	10642	73.6	-0.171	0.103
Dipluridae									
Phyxioschema suthepium	NC_020322.1	13931	67.4	-0.04	0.472	10730	66.6	-0.232	0.153
Hypochilidae									
Hypochilus thorelli	NC_010777.1	13991	70.3	-0.14	0.266	10753	69	-0.192	0.064
Liphistiidae									
Liphistius erawan	NC_020323.1	14197	67.7	0.024	-0.361	10794	66.8	-0.165	-0.09
Songthela hangzhouensis	NC_005924.1	14215	72.2	-0.023	-0.235	10765	71.5	-0.16	-0.022
Nemesiidae									
Calisoga longitarsis	NC_010780.1	14070	64	-0.146	0.365	10738	63.1	-0.257	0.121
Pholcidae									
Pholcus phalangioides	NC_020324.1	14459	65.8	-0.191	0.371	10631	65.4	-0.175	0.078
Pholcus sp. HCP-2014	KJ782458.1	14279	65.8	-0.188	0.372	10631	64.9	-0.174	0.072
Salticidae									
Carrhotus xanthogramma	KP402247.1	14563	75.1	-0.089	0.26	10809	74.1	-0.147	0.057
Selenopidae									
Selenops bursarius	NC_024878.1	14272	74.4	-0.123	0.321	10756	73.8	-0.15	0.056
Theraphosidae									
Haplopelma schmidti	NC_005925.1	13874	69.8	-0.083	0.344	10724	69.8	-0.181	0.092
Thomisidae									
Ebrechtella tricuspidata	KU852748.1	14532	76.2	-0.097	0.221	5552	74.6	-0.143	-0.067
Oxytate striatipes	NC_025557.1	14407	78.2	-0.084	0.212	10784	77.7	-0.179	0.093

Fig. S4 Mitogenome list and overall statistics table.

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene	Position		Size	Intergenic nucleotide	Codon		Strand
	From	To			Start	Stop	
cox3	4	789	786	3	TTG	TAA	H
trnG	788	853	66	-2			H
nad3	837	1173	337	-17	ATT	T	H
trnL	1156	1218	63	-18			L
trnN	1214	1283	70	-5			H
trnA	1279	1335	57	-5			H
trnS1	1327	1381	55	-9			H
trnR	1377	1428	52	-5			H
trnE	1419	1479	61	-10			H
trnF	1460	1513	54	-20			L
nad5	1514	3142	1629		ATA	TAA	L
trnH	3147	3205	59	4			L
nad4	3207	4486	1280	1	ATA	TA	L
trnP	4755	4808	54	268			L
nad6	4817	5245	429	8	ATG	TAA	H
trnI	5247	5302	56	1			H

Fig. S5 Annotation table.

Regions	Size (bp)	T(U)	C	A	G	AT(%)	GC(%)	GT(%)	AT skew	GC skew
PCGs	9810	42.9	12.8	30.1	14.1	73	26.9	57	-0.175	0.048
1st codon position	3270	38.8	11.6	32.8	16.8	71.6	28.4	55.6	-0.084	0.183
2nd codon position	3270	44.1	15.8	23.8	16.3	67.9	32.1	60.4	-0.299	0.014
3rd codon position	3270	45.8	11.1	33.8	9.3	79.6	20.4	55.1	-0.151	-0.087
atp6	663	41	11.6	31.7	15.7	72.7	27.3	56.7	-0.129	0.149
atp8	159	36.5	9.4	41.5	12.6	78	22	49.1	0.065	0.143
cox1	1536	41.3	12.4	27.5	18.8	68.8	31.2	60.1	-0.2	0.205
cox3	786	42.1	10.7	28.4	18.8	70.5	29.5	60.9	-0.195	0.276
cytb	1134	43.5	12	28.2	16.3	71.7	28.3	59.8	-0.213	0.153
nad1	903	44.5	15.3	29.3	10.9	73.8	26.2	55.4	-0.205	-0.169
nad2	954	39.9	7.3	35.4	17.3	75.3	24.6	57.2	-0.06	0.404
nad3	337	49.3	11.9	28.8	10.1	78.1	22	59.4	-0.262	-0.081
nad4	1280	44.1	15.9	30.7	9.2	74.8	25.1	53.3	-0.18	-0.267
nad5	1629	44.9	17.4	28.5	9.1	73.4	26.5	54	-0.224	-0.312
nad6	429	40.3	5.1	36.4	18.2	76.7	23.3	58.5	-0.052	0.56
rrnL	1007	41.9	12.4	34.7	11	76.6	23.4	52.9	-0.095	-0.059
rrnS	698	41.4	10.9	34.4	13.3	75.8	24.2	54.7	-0.093	0.101
tRNAs	1222	38.3	9.7	38.6	13.4	76.9	23.1	51.7	0.004	0.163
rRNAs	1705	41.7	11.8	34.5	12	76.2	23.8	53.7	-0.094	0.007
Full genome	14344	36.6	10.5	37.1	15.8	73.7	26.3	52.4	0.007	0.205

Fig. S6 Nucleotide composition and skewness table.

Sequences used: Cyrtophora moluccensis											
Codon Table: 5											
Domain: Data											
Codon	Count	RSCU	Codon	Count	RSCU	Codon	Count	RSCU	Codon	Count	RSCU
UUU(F)	316	1.83	UCU(S)	112	2.47	UAU(Y)	113	1.65	UGU(C)	21	1.62
UUC(F)	30	0.17	UCC(S)	28	0.62	UAC(Y)	24	0.35	UGC(C)	5	0.38
UUA(L)	273	3.81	UCA(S)	88	1.94	UAA(*)	7	1.56	UGA(W)	75	1.76
UUG(L)	36	0.5	UCG(S)	6	0.13	UAG(*)	2	0.44	UGG(W)	10	0.24
CUU(L)	65	0.91	CCU(P)	75	2.54	CAU(H)	53	1.74	CGU(R)	5	0.45
CUC(L)	7	0.1	CCC(P)	13	0.44	CAC(H)	8	0.26	CGC(R)	4	0.36
CUA(L)	43	0.6	CCA(P)	24	0.81	CAA(Q)	42	1.68	CGA(R)	29	2.64
CUG(L)	6	0.08	CCG(P)	6	0.2	CAG(Q)	8	0.32	CGG(R)	6	0.55
AUU(I)	297	1.73	ACU(T)	54	1.86	AAU(N)	103	1.6	AGU(S)	13	0.29
AUC(I)	47	0.27	ACC(T)	11	0.38	AAC(N)	26	0.4	AGC(S)	9	0.2
AUA(M)	217	1.72	ACA(T)	45	1.55	AAA(K)	73	1.51	AGA(S)	86	1.9
AUG(M)	36	0.28	ACG(T)	6	0.21	AAG(K)	24	0.49	AGG(S)	21	0.46
GUU(V)	67	1.3	GCU(A)	96	2.65	GAU(D)	50	1.79	GGU(G)	41	0.89
GUC(V)	15	0.29	GCC(A)	9	0.25	GAC(D)	6	0.21	GGC(G)	11	0.24
GUA(V)	96	1.86	GCA(A)	38	1.05	GAA(E)	54	1.54	GGA(G)	87	1.89
GUG(V)	28	0.54	GCG(A)	2	0.06	GAG(E)	16	0.46	GGG(G)	45	0.98

Fig. S7 Relative synonymous codon usage (RSCU) table.

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

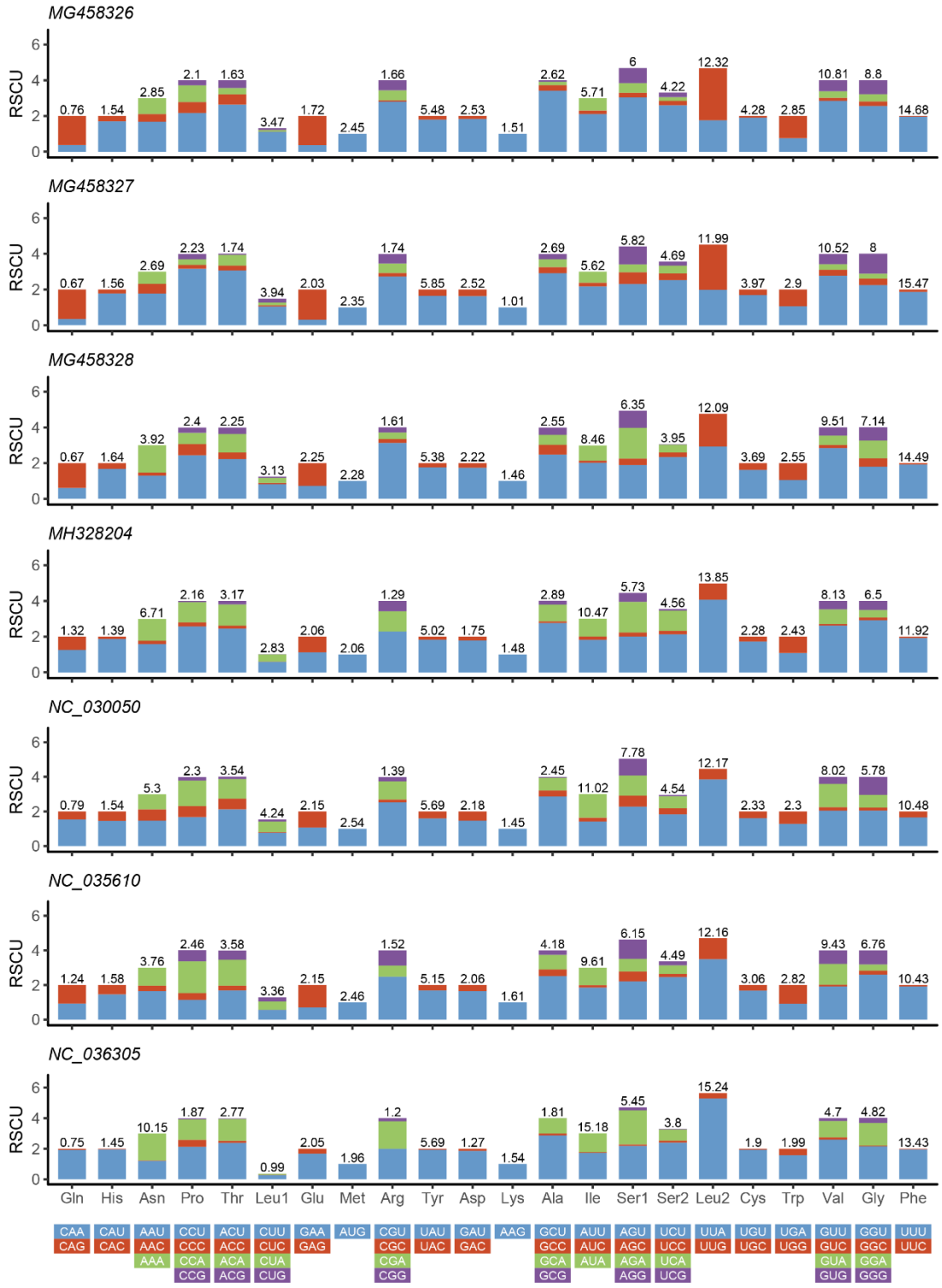

Fig. S8 Relative synonymous codon usage (RSCU) of seven species.

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AA	Count	%
Phe(F)	346	10.61
Leu2(L2)	309	9.48
Leu1(L1)	121	3.71
Ile(I)	344	10.55
Met(M)	253	7.76
Val(V)	206	6.32
Ser2(S2)	234	7.18
Pro(P)	118	3.62
Thr(T)	116	3.56
Ala(A)	145	4.45
Tyr(Y)	137	4.2
His(H)	61	1.87
Gln(Q)	50	1.53
Asn(N)	129	3.96
Lys(K)	97	2.98
Asp(D)	56	1.72
Glu(E)	70	2.15
Cys(C)	26	0.8
Trp(W)	85	2.61
Arg(R)	44	1.35
Ser1(S1)	129	3.96
Gly(G)	184	5.64
codon end in A or T	2758	84.6
codon end in G or T	1739	53.34

Fig. S9 Amino acid usage statistics file.

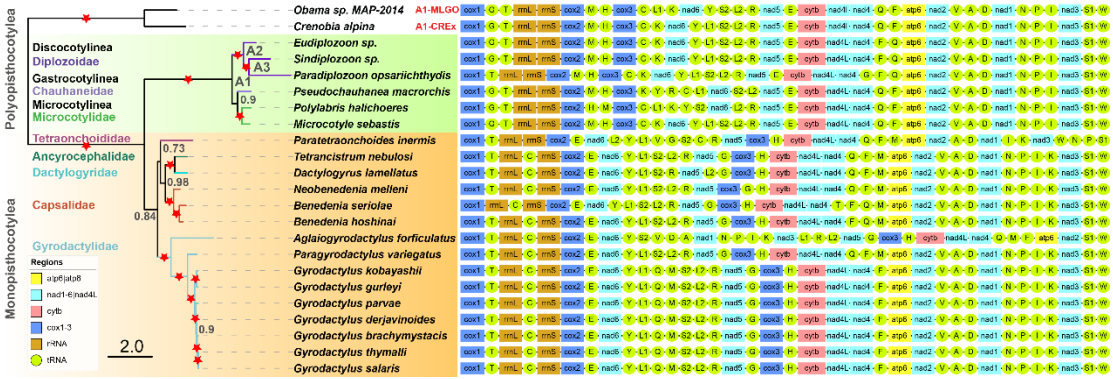

Fig. S10 Mapping gene orders of monogenean mitogenomes onto the phylogenetic tree. The figure was published in Zhang, et al. (2018).

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Species	P_sp	C_x	C_a	C_n	H_p	N_n	N_a	S_h	H_s	H_o
Length of PCGs (bp)										
atp6	663	663	663	663	663	663	663	661	669	660
atp8	150	153	153	162	156	159	159	150	153	153
cox1	1536	1539	1536	1533	1536	1536	1536	1533	1536	1542
cox2	640	664	666	666	666	666	666	667	669	666
Length of rRNA genes (bp)										
rrnL	1002	999	1026	1043	1028	1023	1022	1119	1048	1018
rrnS	691	693	680	696	696	699	700	698	666	691
Putative start codon										
atp6	ATG	ATA	ATA	ATA	ATA	ATA	ATA	ATG	ATG	ATT
atp8	ATT	ATT	ATT	ATT	ATA	ATT	ATT	ATA	ATG	ATT
cox1	ATA	TTA	TTA	CTG	TTA	TTA	TTA	TTG	TTA	ATT
cox2	GTG	TTG	TTG	TTG	TTG	TTG	TTG	CTG	ATG	TTG
Putative terminal codon										
atp6	TAA	TAA	TAA	TAA	TAA	TAA	TAA	T	TAG	TAA
atp8	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAG
cox1	TAG	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAG	TAA
cox2	T	T	TAA	TAA	TAA	TAA	TAA	T	TAG	TAG
AT skew										
atp6	-0.317	-0.195	-0.148	-0.171	-0.266	-0.173	-0.229	-0.122	-0.261	-0.21
atp8	-0.258	-0.023	0.083	0.015	-0.023	0.143	0.071	-0.135	0.053	-0.042
cox1	-0.316	-0.234	-0.205	-0.207	-0.222	-0.173	-0.207	-0.163	-0.255	-0.233
cox2	-0.233	-0.239	-0.117	-0.11	-0.157	-0.097	-0.149	-0.082	-0.127	-0.187
rRNAs	0.145	0.041	-0.023	0	0.04	-0.005	-0.015	0.034	0.022	0.087
rrnL	0.145	0.037	-0.027	0.002	0.005	-0.016	0.001	0.054	-0.021	0.076
rrnS	0.144	0.048	-0.018	-0.004	0.094	0.011	-0.039	0	0.089	0.103
tRNAs	0.006	0.009	0.04	0.007	0.036	0.01	0.003	-0.029	0.076	0.029
GC skew										
atp6	0.312	0.255	0.129	0.193	0.231	0.133	0.18	-0.284	0.295	0.264
atp8	0.472	0.4	0.212	0.308	-0.04	0.308	0.438	-0.59	0.487	0.515
cox1	0.293	0.206	0.165	0.202	0.245	0.193	0.216	-0.072	0.271	0.244
cox2	0.365	0.268	0.19	0.23	0.189	0.205	0.259	-0.196	0.384	0.347
rRNAs	0.145	0.041	-0.023	0	0.04	-0.005	-0.015	0.034	0.022	0.087
rrnL	-0.154	-0.039	-0.013	-0.005	-0.02	0	-0.041	0.26	-0.11	-0.161
rrnS	-0.101	0.015	0.126	0.076	0.032	0.047	0.018	0.175	-0.125	-0.09
tRNAs	0.006	0.009	0.04	0.007	0.036	0.01	0.003	-0.029	0.076	0.029
AT content										
atp6	66.7	75.7	74.4	75	76.5	78.4	73.1	73.3	71.1	73.6
atp8	64.7	86.9	78.4	84	83.9	83.6	79.9	74	74.5	77.8
cox1	61.5	69.4	69.6	68.4	69.7	72.1	68.6	67.5	66.4	69.7
cox2	63.6	72.5	70.7	73.9	73.7	75.8	71.6	67.9	68.5	70.6
rRNAs	71	80.1	76.2	80.6	79.6	80	77.7	74.3	70.3	78.2
rrnL	70.7	79.3	76.7	80.2	80.8	80.9	78.3	75.6	69.8	78.6
rrnS	71.3	81.3	75.5	81.1	77.7	78.7	76.8	72.2	71.1	77.4
tRNAs	66.8	74.4	76.6	77.2	76.9	78.6	75.9	73.3	70.1	73.4
GC content										
rRNAs	29	19.9	23.8	19.5	20.5	20	22.3	25.7	29.7	21.8
rrnL	29.3	20.6	23.3	19.9	19.2	19.2	21.7	24.4	30.3	21.3
rrnS	28.7	18.7	24.5	19	22.3	21.4	23.3	27.8	28.8	22.6
tRNAs	33.2	25.6	23.5	22.8	23.1	21.4	24.1	26.8	29.9	26.6

Fig. S11 Statistics for individual genes, including size, start and terminal codons, base composition and skews. Only 4 protein-coding genes (PCGs) are shown.

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

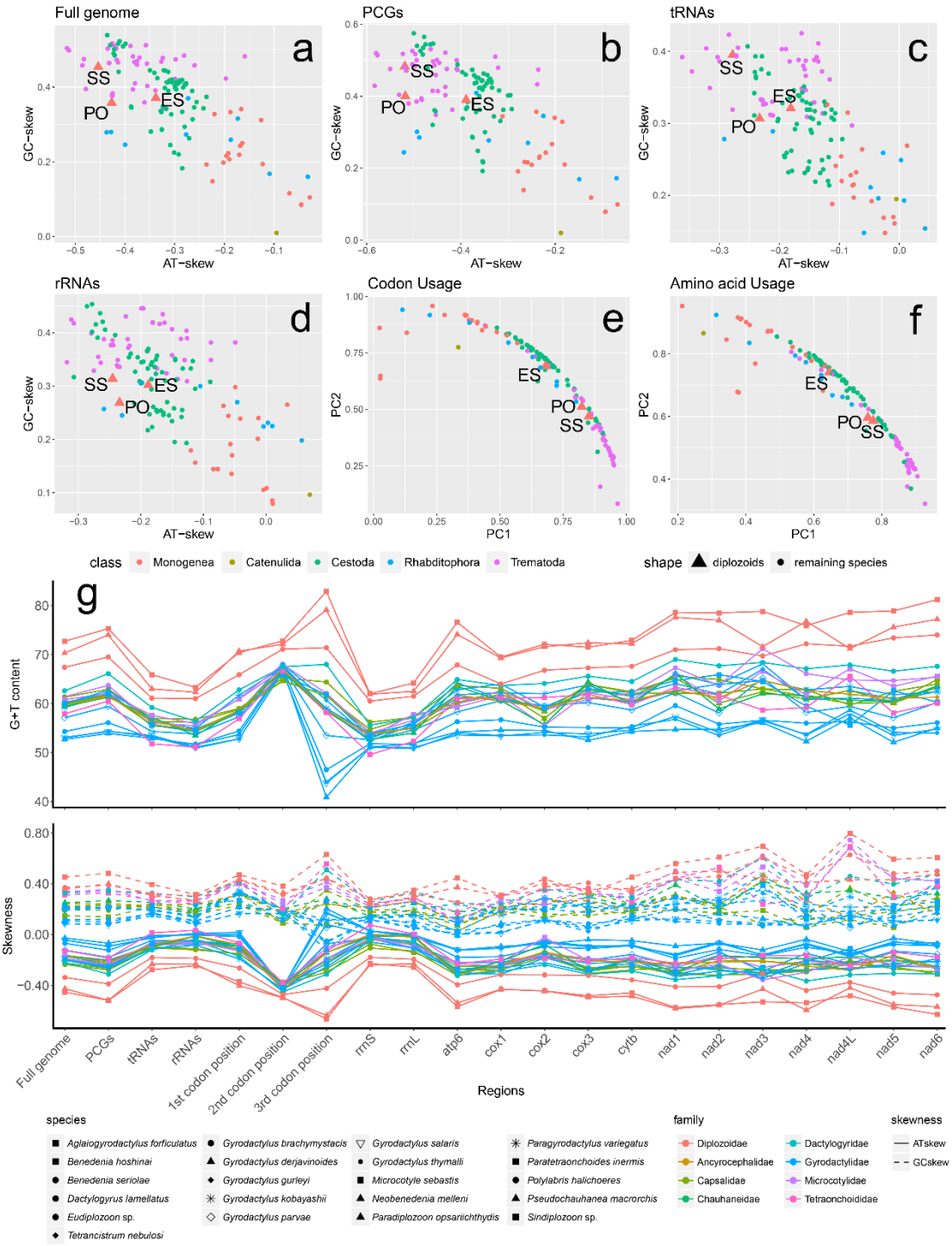

Fig. S12 Skewness and base content of some flatworm mitogenomes. The figure was published in Zhang, et al. (2018).

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene	Position		Size	Strand	Identity
	From	To			
NC_005924/NC_005925/KU852748					
cox1	1/1/1217	1533/1536/2752	1533/1536/1536	H/H/H	70.18/73.96/76.17/73.44
cox2	1537/1537/2756	2203/2205/3427	667/669/672	H/H/H	60.75/64.14/66.96/63.95
trnK	2204/2204/3428	2265/2263/3486	62/60/59	H/H/H	50.00/49.21/62.30/53.83
trnD	2264/2247/3472	2318/2301/3522	55/55/51	H/H/H	48.28/61.82/53.57/54.56
atp8	2319/2293/3532	2468/2445/3675	150/153/144	H/H/H	51.88/42.61/48.41/47.63
atp6	2462/2439/-	3122/3107/-	661/669/-	H/H/-	55.46/-/-/55.46
cox3	3124/3108/-	3897/3891/-	774/784/-	H/H/-	60.84/-/-/60.84
trnG	3909/3892/5126	3962/3945/5180	54/54/55	H/H/H	50.91/57.89/62.50/57.10
nad3	3969/3945/-	4298/4266/-	330/322/-	H/H/-	52.54/-/-/52.54
trnA	4298/4349/5608	4350/4405/5665	53/57/58	H/H/H	41.38/47.62/50.00/46.33
trnR	4357/4452/5718	4413/4501/5770	57/50/53	H/H/H	26.87/50.88/58.49/45.41
trnN	4413/4316/5562	4471/4363/5617	59/48/56	H/H/H	51.67/47.62/44.64/47.98
trnS1	4472/4396/5670	4523/4459/5721	52/64/52	H/H/H	42.19/44.23/51.56/45.99
trnE	4524/4493/5759	4583/4543/5811	60/51/53	H/H/H	49.18/58.33/56.60/54.71
trnF	4572/4531/5797	4626/4583/5855	55/53/59	L/L/L	58.18/46.97/54.24/53.13
nad5	4628/4584/5856	6263/6222/7479	1636/1639/1624	L/L/L	51.61/54.32/58.86/54.93
trnH	6262/6211/7492	6318/6268/7546	57/58/55	L/L/L	49.18/50.88/51.67/50.57
nad4	6319/6267/7547	7618/7544/8822	1300/1278/1276	L/L/L	53.40/55.84/58.40/55.88
nad4L	7618/7545/8791	7899/7821/9090	282/277/300	L/L/L	46.82/47.44/57.00/50.42
trnT	7898/9478/10812	7962/9540/10866	65/63/55	H/H/H	56.72/55.38/69.84/60.65
trnP	7959/7815/9083	8019/7867/9140	61/53/58	L/L/L	55.74/47.62/58.62/53.99
nad6	8031/7860/-	8462/8288/-	432/429/-	H/H/-	48.44/-/-/48.44
cytb	8469/8289/-	9590/9423/-	1122/1135/-	H/H/-	60.09/-/-/60.09
trnS2	9589/9425/10758	9651/9478/10811	63/54/54	H/H/H	34.33/53.97/48.33/45.54

Fig. S13 Comparison of genomic annotations and pairwise similarity calculation for homologous genes.

Usage examples

PhyloSuite has been used previously to conduct analyses in a number of published papers. You may refer to the following publications for examples of the use of PhyloSuite: (Hua, et al., 2018; Li, et al., 2018; Li, et al., 2017; Liu, et al., 2017; Liu, et al., 2018; Liu, et al., 2018; Wang, et al., 2017; Wen, et al., 2017; Xi, et al., 2018; Zhang, et al., 2018; Zhang, et al., 2018; Zhang, et al., 2018; Zhang, et al., 2017; Zhang, et al., 2017; Zou, et al., 2017; Zou, et al., 2018). Note that we have merged our two older beta tools, MitoTool (https://github.com/dongzhang0725/MitoTool) and BioSuite (https://github.com/dongzhang0725/BioSuite) into PhyloSuite, so some of our older published papers may refer these two tools instead.

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Blair, C. and Murphy, R.W. (2011) Recent trends in molecular phylogenetic analysis: where to next? *J. Hered.*, 102(1), 130-138.

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

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

Hua, C.J.*, et al.* (2018) Basal position of two new complete mitochondrial genomes of parasitic Cymothoida (Crustacea: Isopoda) challenges the monophyly of the suborder and phylogeny of the entire order. *Parasit. Vectors*, 11(1), 628.

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. *Mol. Biol. Evol.*, 29(6), 1695-1701.

Li, W.X.*, et al.* (2018) Comparative mitogenomics supports synonymy of the genera *Ligula* and *Digramma* (Cestoda: Diphyllobothriidae). *Parasit. Vectors*, 11(1), 324.

Li, W.X.*, et al.* (2017) The complete mitochondrial DNA of three monozoic tapeworms in the Caryophyllidea: a mitogenomic perspective on the phylogeny of eucestodes. *Parasit. Vectors*, 10(1), 314.

Liu, F.-F.*, et al.* (2017) Tandem duplication of two tRNA genes in the mitochondrial genome of *Tagiades vajuna* (Lepidoptera: Hesperiidae). *Eur. J. Entomol.*, 114, 407-415.

Liu, Q.*, et al.* (2018) Genetic diversity of glacier-inhabiting Cryobacterium bacteria in China and description of Cryobacterium zongtaii sp. nov. and Arthrobacter glacialis sp. nov. *Syst. Appl. Microbiol.*

Liu, Z.Q.*, et al.* (2018) Sequencing of complete mitochondrial genomes confirms synonymization of Hyalomma asiaticum asiaticum and kozlovi, and advances phylogenetic hypotheses for the Ixodidae. *PLoS One*, 13(5), e0197524.

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.

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.

Wang, J.G.*, et al.* (2017) Sequencing of the complete mitochondrial genomes of eight freshwater snail

bioRxiv preprint doi: https://doi.org/10.1101/489088; this version posted December 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

species exposes pervasive paraphyly within the Viviparidae family (Caenogastropoda). *PLoS One*, 12(7), e0181699.

Wen, H.B*., et al.* (2017) The complete maternally and paternally inherited mitochondrial genomes of a freshwater mussel Potamilus alatus (Bivalvia: Unionidae). *PloS one*, 12(1), e0169749.

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

Xi, B.W*., et al.* (2018) Characterization of the complete mitochondrial genome of Parabreviscolexniepini Xi et al., 2018 (Cestoda, Caryophyllidea). *Zookeys*, (783), 97-112.

Zhang, D*., et al.* (2018) Mitochondrial genomes of two diplectanids (Platyhelminthes: Monogenea) expose paraphyly of the order Dactylogyridea and extensive tRNA gene rearrangements. *Parasit. Vectors*, 11(1), 601.

Zhang, D*., et al.* (2018) Homoplasy or plesiomorphy? Reconstruction of the evolutionary history of mitochondrial gene order rearrangements in the subphylum Neodermata (under review). *Int. J. Parasitol.*

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, characterization and phylogenomics of the complete mitochondrial genome of *Dactylogyrus lamellatus* (Monogenea: Dactylogyridae). *J. Helminthol.*, 1-12.

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. *Parasit. Vectors*, 10(1), 462.

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*.

Zou, H*., et al.* (2017) The complete mitochondrial genome of parasitic nematode *Camallanus cotti*: extreme discontinuity in the rate of mitogenomic architecture evolution within the Chromadorea class. *BMC Genomics*, 18(1), 840.

Zou, H*., et al.* (2018) The complete mitochondrial genome of Cymothoa indica has a highly rearranged gene order and clusters at the very base of the Isopoda clade. *PLoS One*, 13(9), e0203089.