Zoological Systematics, 48(2): 107–116 (April 2023), DOI: 10.11865/zs.2023201

COMMUNICATION

Construction of a phylogenetic matrix: Scripts and guidelines for phylogenomics

Shiyu Du[1], Yinhuan Ding[2], Hu Li[3], Aibing Zhang[4], Arong Luo[5], Chaodong Zhu[4, 5], Feng Zhang[1, 5 *]

[1]*Department of Entomology, College of Plant Protection, Nanjing Agricultural University, Nanjing 210095, China*
[2]*Department of Agronomy and Horticulture, Jiangsu Vocational College of Agriculture and Forestry, Nanjing 212400, China*
[3]*Department of Entomology and MOA Key Lab of Pest Monitoring and Green Management, College of Plant Protection, China Agricultural University, Beijing 100193, China*
[4]*College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*
[5]*Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China*
[*]*Corresponding author, E-mail: fzhang@njau.edu.cn*

Abstract Phylogenomics is a new field that infers evolutionary relationships of taxa at the genome-scale level. The increment of molecular data may raise the potential bias as the limiting factor in phylogenomics. It is particularly important to explore these factors in phylogenomic analyses by simple, convenient, time-saving and (relatively) robust means. Here, we construct a set of custom scripts for USCO (universal single-copy orthologs) loci extraction, multiple sequence alignment, trimming poorly aligned regions, loci filtering and creating a concatenation matrix, prior to reconstructing the phylogenetic trees, to simplify analytical pipelines and improve the accuracy of tree estimation. These scripts employed a series of computationally efficient bioinformatic tools, and were used with a universal 'BASH' shell or visual interface by Windows-like 'drag and drop' operations in LINUX systems. Most steps in these scripts are parallelized to accelerate analyses. These new custom scripts provide a convenient analytical solution for phylogenomics data preparation, data quality control, and detection of potential analytical errors. Details and scripts usage are provided at https://github.com/xtmtd/Phylogenomics/tree/main/scripts. The virtual mirror file (.vmdk) integrates the operating system and required environment. All tools and scripts can be downloaded from https://dx.doi.org/10.6084/m9.figshare.21283026. Besides, the video introduction and "walk-through" for each script are provided at https://space.bilibili.com/319699648/channel/seriesdetail?sid=2682055.

Key words Analytical bias, sequence aligning and trimming, loci filtering, phylogenetic inference.

1 Introduction

Phylogenomics, putting forward in 1998 for the first time, was used to predict gene function at the genome-scale level (Eisen, 1998) and was applied to the phylogenetic inference in the ensuing year (O'Brien & Stanyon, 1999). In recent years, phylogenomics has played an essential role in understanding the phylogenetic relationships and evolution between taxa (Young & Gillung, 2020). It is no longer challenging to obtain whole-genome data with advances in high-throughput sequencing technology (next-generation sequencing, NGS; Metzker, 2010). Abundant universal molecular marker set design or extraction processes, such as USCO (universal single-copy orthologues; Simão *et al.*, 2015), UCE (ultraconserved elements; Faircloth *et al.*, 2012), AHE (anchored hybrid enrichment; Lemmon *et al.*, 2012), and SNP (single nucleotide

Received 8 November 2022, accepted 30 January 2023
Executive editor: Fuqiang Chen

polymorphism; Sherry *et al.*, 1999), *etc.*, have been very popular (*e.g.* Zhang *et al.*, 2019; Sun *et al.*, 2020). Resolving some tricky nodes in the tree of animals based on hundreds or thousands of markers has been a common practice, for example, in mammals (*e.g.* Chen *et al.*, 2019; Jebb *et al.*, 2020; Johnson *et al.*, 2022), birds (*e.g.* Jarvis *et al.*, 2014; Manthey *et al.*, 2016), especially insects (*e.g.* Misof *et al.*, 2014; Johnson *et al.*, 2018; McKenna *et al.*, 2019; Wipfler *et al.*, 2019; Allio *et al.*, 2020; de Moya *et al.*, 2021; Tihelka *et al.*, 2021). Apparently, big-data-based phylogenetics, *i.e.*, phylogenomics, provides new thinking in studying evolutionary relationships among most biological groups.

However, simply increasing the number of molecular markers will not effectively resolve problematic nodes in the tree of life (Young & Gillung, 2020). Because the low-quality phylogenetic matrix suffered from the analytical errors (biological and methodological; Bouckaert & Lockhart, 2015). Major confounding factors detection is therefore crucial in phylogenomics (Kapli *et al.*, 2020). Most common analytical errors include four types (Young & Gillung, 2020). Type 1, missing data, which means loci are not sampled for the taxa in an alignment or part of the sites are missing in a locus (Hosner *et al.*, 2016; Kocot *et al.*, 2017; Roure *et al.*, 2013; Smith *et al.*, 2018), can bias phylogenetic relationships (*e.g.* Lemmon *et al.*, 2009; Simmons, 2014). Missing data is a common but an overlooked confounding factor in phylogenomics. It is often an excellent choice to design a more minor (*i.e.*, with fewer loci) but a more completed matrix (the premise is that enough sites/loci are still retained) to overcome impacts of missing data (Young & Gillung, 2020). Type 2 is compositional heterogeneity states sequence differences of individual bases or amino acids between groups (Jeffroy *et al.*, 2006; Philippe *et al.*, 2011; Duchêne, *et al.*, 2017; Borowiec *et al.*, 2019). It can deviate estimates of topology and branch lengths regardless and/or overthinking sequence differences (*e.g.* Jermiin *et al.*, 2004; Nesnidal *et al.*, 2010; Nabholz *et al.*, 2011). Using proper models and removing the loci or groups exhibiting deviation in composition, can reduce such heterogeneity (Young & Gillung, 2020). Type 3, rate heterogeneity, *i.e.*, long-branch heterogeneity or long-branch attraction (LBA), a major obstacle in phylogenomics (Qu *et al.*, 2017), negatively affect the accuracy of tree reconstruction (Nosenko *et al.*, 2013; Struck, 2014; Kück & Wägele, 2015). Increasing the taxa sampling (*e.g.* Bergsten, 2005; Pick *et al.*, 2010; Zhong *et al.*, 2010), removing the long-branch (*e.g.* Bergsten, 2005; Hampl *et al.*, 2009), using amino acids and site-heterogeneous models (*e.g.* Lartillot *et al.*, 2007; Talavera & Vila, 2011), *etc.*, to reconstruct phylogenetic trees, can alleviate the impact of LBA. Type 4, gene tree heterogeneity, which is variously termed incomplete lineage sorting (ILS) (*e.g.* Betancur-R *et al.*, 2013; Copetti *et al.*, 2017; Richards *et al.*, 2018; Kapli *et al.*, 2020). It can pose challenges to species tree inference (Edwards, 2009). Using coalescence approaches, such as multi-species coalescent (MSC) model (Rannala & Yang, 2003; Degnan & Rosenberg, 2009), can reduce this type of error appropriately (Young & Gillung, 2020). Hence, detection confounding factors can reduce most analytical errors and reconstruct reliable phylogenetic trees using the high-quality matrix in phylogenomics.

To date, some widely applicable tools for aligning, trimming, filtering, and matrix generation for phylogenomics have been proposed (Table 1). Five major categories of multiple sequence alignment software have been classified (Kapli *et al.*, 2020; Smirnov & Warnow, 2021): (1) the progressive approach (most commonly used), including MUSCLE (Edgar, 2004), CLUSTAL O (Sievers *et al.*, 2011), and MAFFT (Katoh & Standley, 2013); (2) the consistency-based methods, such as ProbCons (Do *et al.*, 2005), Probalign (Roshan & Livesay, 2006) and T-Coffee (Notredame *et al.*, 2003); (3) the statistical or evolution-based methods, for example, PRANK (Löytynoja & Goldman, 2008) and StatAlign (Novák *et al.*, 2008); (4) the divide-and-conquer strategy, for instance, PASTA (Mirarab *et al.*, 2015) and SATé-II (Liu *et al.*, 2012); (5) the graph clustering approach, such as POA (Lee *et al.*, 2002) and MAGUS (is similar to MAFFT; Smirnov & Warnow, 2021). ProbCons, T-Coffee, Probalign, MAFFT, PASTA, SATé-II, and MAGUS are more accurate than other programs mentioned above (Pais *et al.*, 2014; Kapli *et al.*, 2020; Smirnov & Warnow, 2021). POA is the fastest program (Kapli *et al.*, 2020). MAGUS and MAFFT are the two programs that have more and more been used in alignment in recent years because of their efficiency and accuracy. All the alignment trimming software are focus on the removal of highly divergent sites and/or remain of the parsimony-informative sites, for instance, Gblocks (Talavera & Castresana, 2007), Guidance2 (Sela *et al.*, 2015), Noisy (Dress *et al.*, 2008), trimAl (Capella-Gutiérrez *et al.*, 2009), BMGE (Criscuolo & Gribaldo, 2010), and ClipKIT (Steenwyk *et al.*, 2020b), *etc.* Gblocks, Guidance2, and Noisy have lower tree certainty values than trimAl, BMGE, and ClipKIT (Steenwyk *et al.*, 2020b). For filtering, a toolkit, PhyKIT (Steenwyk *et al.*, 2020a), aimed at detecting unreliable alignments or trees have been published. It has more than 30 alignment- and tree-based functions. For example, calculating the number of parsimony-informative sites in an alignment (Shen *et al.*, 2016; Steenwyk *et al.*, 2020a), the RCV (relative composition variability) for an alignment (Phillips & Penny, 2003), the GC content (Guanine-Cytosine content) of a fasta file (Shen *et al.*, 2016), the average pairwise identity among sequences (Chen *et al.*, 2017) and the DVMC (degree of violation of the molecular clock) in a phylogeny (Liu *et al.*, 2017), identifying the potentially spurious homologs (Shen *et al.*, 2018), and creating a concatenation matrix, *etc.* Moreover, TreeShrink (Mai & Mirarab, 2018) also can detect outlier long branches in phylogenetic trees quickly and accurately. Although these tools provided valuable source of data processing, they can merely identify bias in an individual locus. Therefore, using various tools to perform analyses is inconvenient,

Table 1. Features of different sequence alignment, trimming and filtering programs.

	Software	Method	Advantage	Shortcoming	Link	Refences
Alignment	MUSCLE	Progressive	Time- and memory-saving	Unstable	http://www.drive5.com/muscle/	Edgar, 2004
	CLUSTAL O	Progressive	Time-saving	Inaccurate, memory-consuming	http://www.clustal.org/	Sievers *et al.*, 2011
	MAFFT	Progressive	Time-saving, accurate, scalable	Memory-consuming	https://mafft.cbrc.jp/alignment/software/	Katoh & Standley, 2013
	ProbCons	Consistency	Accurate	Time- and memory-consuming (not apply to nucleotides)	http://probcons.stanford.edu/	Do *et al.*, 2005
	Probalign	Consistency	Accurate	Time-consuming	http://www.cs.njit.edu/usman/probalign	Roshan & Livesay, 2006
	T-Coffee	Consistency	Accurate	Time- and memory-consuming	https://tcoffee.crg.eu/	Notredame, Higgins & Heringa, 2003
	PRANK	Statistical or evolution-based	Methodologically sound	Time- and memory-consuming	https://www.ebi.ac.uk/goldman-srv/webprank/	Löytynoja & Goldman, 2008
	StatAlign	Statistical or evolution-based	Methodologically sound	Time- and memory-consuming	https://statalign.github.io/	Novák *et al.*, 2008
	PASTA	Divide-and-conquer	Accurate	Time-consuming	https://github.com/smirarab/pasta	Mirarab *et al.*, 2015
	SATé-II	Divide-and-conquer	Accurate	Time-consuming	https://phylo.bio.ku.edu/software/sate/sate.html	Liu *et al.*, 2012
	POA	Graphs	Time- and memory-saving	Inaccurate	https://sourceforge.net/projects/poamsa/	Lee, Grasso & Sharlow, 2002
	MAGUS	Graphs	Accurate, memory-saving, scalable	Time-consuming	https://github.com/vlasmirnov/MAGUS	Smirnov & Warnow, 2021
Trimming	Gblocks	Remove gap rich and highly variable sites	Time-saving	Limited, inaccurate	https://gensoft.pasteur.fr/docs/gblocks/0.91b/	Talavera & Castresana, 2007
	Guidance2	Remove uncertain sites	Interfacial, with alignment function	Inaccurate	http://guidance.tau.ac.il/ver2/	Sela *et al.*, 2015
	Noisy	Remove homoplastic sites	Reliable	Conservative	http://www.bioinf.uni-leipzig.de/Software/noisy/	Dress *et al.*, 2008
	trimAl	Remove highly gappy and/or variable sites	Time-saving, reliable, accurate	-	http://trimal.cgenomics.org/	Capella-Gutiérrez, Silla-Martínez & Gabaldón, 2009
	BMGE	Remove high entropy sites	Time-saving	Strict	https://gitlab.pasteur.fr/GIPhy/BMGE	Criscuolo & Gribaldo, 2010
	ClipKIT	Keep parsimony-informative and/or constant sites Remove highly gappy sites	Reliable, accurate, multi-strategies	Time-consuming	https://jlsteenwyk.com/ClipKIT/	Steenwyk *et al.*, 2020b
Filtering	PhyKIT	Process and analyze alignments and phylogenies	More than 30 functions	-	https://jlsteenwyk.com/PhyKIT/	Steenwyk *et al.*, 2020a
	TreeShrink	Detect the outlier long branches	Time-saving, accurate	Conservative	https://github.com/uym2/TreeShrink	Mai & Mirarab, 2018

from aligning sequences to creating a matrix using hundreds or thousands of loci. Analyzing phylogenomic data is challenging due to the lack of a unified script or pipeline, especially the commands of batch processing in LINUX. Constructing some efficient scripts or pipelines, which integrate these tools, is of great significance for phylogenetic analyses.

We propose a set of custom scripts (available from GitHub: https://github.com/xtmtd/Phylogenomics/tree/main/scripts) for constructing the high-quality matrix. This study aims to introduce the scripts for phylogenomics by integrating the more accessible, more convenient, and low-consumption bioinformatic tools, including extracting USCO loci, aligning multiple sequences, trimming poorly aligned regions, filtering loci, and creating concatenation matrix. The virtual mirror file (.vmdk) with CentOS 7.3 system is also prepared at https://dx.doi.org/10.6084/m9.figshare.21283026, to facilitate the use of these scripts and simplify the installation process of various packages. Besides, the video introductions for all scripts have been uploaded to https://space.bilibili.com/319699648/channel/seriesdetail?sid=2682055. This work improves our understanding and detection of analytical error, generates the matrix for downstream analyses and provides an important solution for phylogenomics.

2 Method overview

The widely used sequence format in phylogenomics — *fasta* file format (.fasta, .fas, or .fa) was used as the input and output. The files or folders needed in each script/step can 'drag and drop' as prompted, like the Windows interface. It is worth noting that optionally changing the names of files (loci) or folders in the process were not allowed. Our scripts and processes apply to different molecular markers, such as USCO, UCE, AHE, *etc*. The process of constructing a high-quality matrix consists of five steps (Fig. 1) with step independence. Users can combine the appropriate analyses as needed:

(1) For all species, their BUSCO results folders (*run_$SPECIES* folder generated from BUSCO v3) are extracted from BUSCO (https://busco.ezlab.org/; Waterhouse *et al.*, 2018) and deposited in the same folder (*e.g. BUSCOs* folder; also, can name it whatever you want). BUSCOs are the significant source of USCO (*i.e.*, provided the USCO reference sequences for almost all organisms of OrthoDB database (www.orthodb.org)), which are the most commonly used molecular marker. This step aims to extract the amino acid and nucleotide sequences, *i.e.*, modifying the head name of the *fasta* files for each locus and merging sequences of the same locus into the *fasta* files, then filtering loci having too few taxa (less than three). The tool TransDecoder (https://github.com/TransDecoder/TransDecoder) and script *BUSCO_extraction.sh* will be used in this step. Moreover, the UCEs extracted using the scripts4 from Zhang *et al.* (2019) are also applied for the downstream analyses.

(2) Multiple sequence alignment (MSA) plays an important role in identifying homologous regions of biological sequences. Accurate alignment is fundamental in the inference of evolutionary relationships (Kapli *et al.*, 2020). The two most efficient and accurate tools of alignment, MAFFT, and MAGUS, are employed in our script (*i.e.*, *align_MAFFT.sh*). A variety of strategies can be chosen to align the sequences conveniently and accurately, as prompted in this step.

(3) Erroneously inferring site homology or saturation of multiple substitutions in MSAs is thought to negatively impact phylogenetic inference (Talavera & Castresana, 2007). Several trimming tools (including trimAl, BMGE, and ClipKIT) for identifying and removing the poorly aligned regions/highly divergent sites, or retaining parsimony-informative/constant sites in MSAs are used in the script of *trimming_alignments.sh*, prior to phylogenetic inference. The appropriate trimming tools and parameters should be selected according to the research targets.

(4) The properties of genes (sequence- and tree-based) can strongly influence phylogenomic inference (Shen, Salichos & Rokas, 2016). Filtering — removing these confusing genes — has been promoted as a way to increase the phylogenetic signals (Talavera & Castresana, 2007; Tan *et al.*, 2015). Estimated sequence-based properties include alignment length or the number of parsimony-informative sites (Shen *et al.*, 2016), GC content (Shen *et al.*, 2016), RCV (Phillips & Penny, 2003), SRH (stationary, reversible and homogeneous; Naser-Khdour *et al.*, 2021), average pairwise identity among sequences (*i.e.*, the proxy of evolutionary rate; Chen *et al.*, 2017), and likelihood-mapping (Strimmer & von Haeseler, 1997). Tree-based approaches refer to the identification of potentially spurious homologs (Mai & Mirarab, 2018; Shen *et al.*, 2018), ABS calculation (average bootstraps support; Salichos & Rokas, 2013), DVMC (Liu *et al.*, 2017), treeness (*i.e.*, the proportion of the tree distance found on internal branches; Phillips & Penny, 2003), and 'inconsistent' genes identification (*i.e.*, the topological conflict or incongruence between concatenation- and coalescent-based approaches; Shen, Steenwyk & Rokas, 2021). The major aim is to obtain the alignments with the strong phylogenetic signal. The scripts '*loci_filtering_alignment-based.sh*' and '*loci_filtering_tree-based.sh*' combined the tools PhyKIT, TreeShrink, ASTRAL-III v5.6.1 (Zhang *et al.*, 2018), and IQ-TREE. Selecting the suitable strategies (*i.e.*, genic properties) and thresholds are needed.

(5) Construction of an appropriate matrix, the requisite step before reconstructing the trees. The percentage value for taxa occupancy (*i.e.*, completeness, which represents the lowest ratio of taxa for all partitions, usually ranging from 50% to

100%) can be input in the script of *matrix_generation.sh* to create the supermatrix (concatenated matrix) and partition file by using tools PhyKIT and FASconCAT-g v1.04 (Kück & Longo, 2014) with amino acid or nucleotide sequences, for phylogenetic inference. Furthermore, we have added an extra function in this script to move the outgroup species to the first one in the supermatrix to make it more intuitive in the final tree file.

Details and script tutorials are provided on GitHub. Meanwhile, a simulated example is also given in the virtual mirror (BUSCO results folder: /home/zf/Desktop/materials/datasets_examples/BUSCOs/; BUSCO loci extraction, multiple sequence alignment, trimming, filtering and matrix generation folders: /home/zf/Desktop/materials/phylogenomics_examples/). Users can use our examples or their datasets to illustrate the use and effectiveness of these scripts.

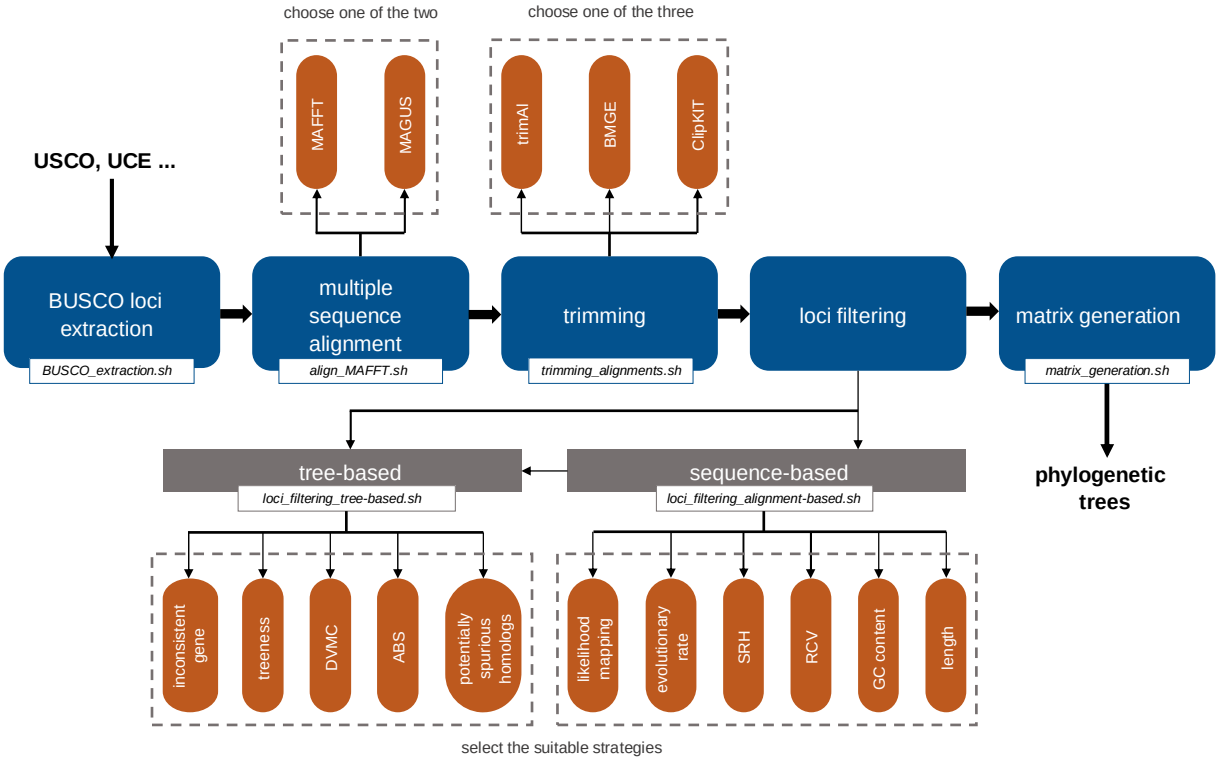

Figure 1. Flowchart of constructing a phylogenetic matrix for phylogenomics. The custom scripts used in each step are marked as italic. Dashed boxes indicate that these strategies of each step choose only one or more suitable strategies.

3 Conclusions

A set of custom scripts and analysis processes were presented in this study to construct a suitable matrix for phylogeny. The use and effectiveness of these scripts are successfully verified in the hexapod groups, Collembola (Yu *et al.*, 2022), Hemiptera (Hu *et al.*, 2023; Song & Zhang, 2022), and Hymenoptera (Zhang *et al.*, 2022), for example. This study highlights the impact of high-quality matrix on phylogenetic inference and promotes the progress of big-data-based phylogenetics. To further optimize and enhance the analysis process of phylogenomics, we will invest future efforts in reconstructing the phylogenetic trees with multiple strategies (approaches and models). Furthermore, future studies would also benefit from the abundant high-quality genomes, which would permit further studying phylogenomics by mining more potential genomic information (*e.g.* duplication, rearrangement, and syntenic block, *etc.*). Finally, we believe our scripts, by integrating the efficient tools of extracting, aligning, trimming, filtering, and matrix generation will be useful in improving the accuracy of phylogenomic trees.

Funding This research was supported by the National Natural Science Foundation of China (31970434, 32270470) to FZ, and the Youth support project of Jiangsu Vocational College of Agriculture and Forestry (2022kj27) to YHD.

Acknowledgments We thank the executive editor Dr. Fuqiang Chen and the anonymous reviewers for their thoughtful comments. We thank Zhihong Zhan (Nanjing Agricultural University) for his comments and suggests on earlier versions of the manuscript. The analyses in this article supported by the high-performance computing platform of Bioinformatics Center, Nanjing Agricultural University.

Data Availability Statement The scripts of phylogenomics are available at GitHub (https://github.com/xtmtd/Phylogenomics/tree/main/scripts, accessed on 1 June 2022). Details and script usage are provided also on the same webpage. The virtual mirror (.vmdk) contains all the bioinformatic tools and custom scripts ready in the CentOS 7.3 system are available at figshare (https://dx.doi.org/10.6084/m9.figshare.21283026, accessed on 6 October 2022). It can be directly opened by VMware or VirtualBox, which are often installed on Windows. The video introduction and "walk-through" for each script are provided at https://space.bilibili.com/319699648/channel/seriesdetail?sid=2682055, accessed on 10 October 2022).

References

Allio, R., Scornavacca, C., Benoit, N., Clamens, A-L., Sperling, F.A.H., Condamine, F.L. 2019. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Systematic Biology*, 69: 38–60. doi: 10.1093/sysbio/syz030

Bergsten, J. 2005. A review of long-branch attraction. *Cladistics*, 21: 163–193. doi: 10.1111/j.1096-0031.2005.00059.x

Betancur-R, R., Li, C., Munroe, T.A., Ballesteros, J.A., Ortí, G. 2013. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). *Systematic Biology*, 62: 763–785. doi: 10.1093/sysbio/ syt039

Borowiec, M.L., Rabeling, C., Brady, S.G., Fisher, B.L., Schultz, T.R., Ward, P.S. 2019. Compositional heterogeneity and outgroup choice influence the internal phylogeny of the ants. *Molecular Phylogenetics and Evolution*, 134: 111–121. doi: 10.1016/j.ympev.2019.01.024

Bouckaert, R., Lockhart, P. 2015. Capturing heterotachy through multi-gamma site models. bioRxiv. doi: 10.1101/018101

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25: 1972–1973. doi: 10.1093/bioinformatics/btp348

Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., Bibi, F., Yang, Y., Wang, J., Nie, W., Su, W., Liu, G., Li, Q., Fu, W., Pan, X., Liu, C., Yang, J., Zhang, C., Yin, Y., Wang, Y., Zhao, Y., Zhang, C., Wang, Z., Qin, Y., Liu, W., Wang, B., Ren, Y., Zhang, R., Zeng, Y., da Fonseca, R.R., Wei, B., Li, R., Wan, W., Zhao, R., Zhu, W., Wang, Y., Duan, S., Gao, Y., Zhang, Y.E., Chen, C., Hvilsom, C., Epps, C.W., Chemnick, L.G., Dong, Y., Mirarab, S., Siegismund, H.R., Ryder, O.A., Gilbert, M.T.P., Lewin, H.A., Zhang, G., Heller, R., Wang, W. 2019. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science*, 364(6446): eaav6202. doi: 10.1126/science.aav6202

Chen, M.Y., Liang, D., Zhang, P. 2017. Phylogenomic resolution of the phylogeny of laurasiatherian mammals: Exploring phylogenetic signals within coding and noncoding sequences. *Genome Biology and Evolution*, 9(8): 1998–2012. doi: 10.1093/gbe/evx147

Copetti, D., Búrquez, A., Bustamante, E., Charboneau, J.L.M., Childs, K.L., Eguiarte, L.E., Lee, S., Liu, T.L., McMahon, M.M., Whiteman, N.K., Wing, R.A., Wojciechowski, M.F., Sanderson, M.J. 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. *Proceedings of the National Academy of Sciences*, 114(45): 12003–12008. doi: 10.1073/pnas.1706367114

Criscuolo, A., Gribaldo, S. 2010. BMGE (Block Mapping and Gathering with Entropy): selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*, 10: 210. doi: 10.1186/1471-2148-10-210

Crotty, S.M., Minh, B.Q., Bean, N.G., Holland, B.R., Tuke, J., Jermiin, L.S., von Haeseler, A. 2019. GHOST: Recovering historical signal from heterotachously evolved sequence alignments. *Systematic Biology*, 69(2): 249–264. doi: 10.1093/sysbio/syz051

de Moya, R.S., Yoshizawa, K., Walden, K.K.O., Sweet, A.D., Dietrich, C.H., Johnson, K.P. 2021. Phylogenomics of parasitic and nonparasitic lice (Insecta: Psocodea): Combining sequence data and exploring compositional bias solutions in next generation data sets. *Systematic Biology*, 70(4): 719–738. doi: 10.1093/sysbio/syaa075

Degnan, J.H., Rosenberg, N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6): 332–340. doi: 10.1016/j.tree.2009.01.009

Do, C.B., Mahabhashyam, M.S., Brudno, M., Batzoglou, S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2): 330–340. doi: 10.1101/gr.2821705

Dress, A.W., Flamm, C., Fritzsch, G., Grünewald, S., Kruspe, M., Prohaska, S.J., Stadler, P.F. 2008. Noisy: Identification of problematic columns in multiple sequence alignments. *Algorithms for Molecular Biology*, 3: 7. doi: 10.1186/1748-7188-3-7

Duchêne, D.A., Duchêne, S., Ho, S.Y.W. 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. *Molecular Biology and Evolution*, 34(6): 1529–1534. doi: 10.1093/molbev/msx092

Edgar, R.C. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5: 113. doi: 10.1186/1471-2105-5-113

Edwards, S.V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution*, 63: 1–19. doi: 10.1111/j.1558-5646.2008.00549.x

Eisen, J.A. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, 8(3): 163–167. doi: 10.1101/gr.8.3.163

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, 61: 717–726. doi: 10.1093/sysbio/sys004

Hampl, V., Hug, L., Leigh, J.W., Dacks, J.B., Lang, B.F., Simpson, A.G.B., Roger, A.J. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proceedings of the National Academy of Sciences*, 106(10): 3859–3864. doi: 10.1073/pnas.0807880106

Hosner, P.A., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T. 2016. Avoiding missing data biases in phylogenomic inference: an empirical study in the Landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, 33(4): 1110–1125. doi: 10.1093/molbev/msv347

Hu, Y., Dietrich, C.H., Skinner, R.K., Zhang, Y. 2023. Phylogeny of Membracoidea (Hemiptera: Auchenorrhyncha) based on transcriptome data. *Systematic Entomology*, 48(1): 97–110. doi: 10.1111/syen.12563

Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y.W., Faircloth, B.C., Nabholz, B., Howard, J.T., Suh, A., Weber, C.C., da Fonseca, R.R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M.S., Zavidovych, V., Subramanian, S., Gabaldon, T., Capella-Gutierrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W.C., Ray, D., Green, R.E., Bruford, M.W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E.P., Bertelsen, M.F., Sheldon, F.H., Brumfield, R.T., Mello, C.V., Lovell, P.V., Wirthlin, M., Schneider, M.P.C., Prosdocimi, F., Samaniego, J.A., Velazquez, A.M.V., Alfaro-Nunez, A., Campos, P.F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D.M., Zhou, Q., Perelman, P., Driskell, A.C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F.E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F.K., Jonsson, K.A., Johnson, W., Koepfli, K.-P., O'Brien, S., Haussler, D., Ryder, O.A., Rahbek, C., Willerslev, E., Graves, G.R., Glenn, T.C., McCormack, J., Burt, D., Ellegren, H., Alstrom, P., Edwards, S.V., Stamatakis, A., Mindell, D.P., Cracraft, J., Braun, E.L., Warnow, T., Jun, W., Gilbert, M.T.P., Zhang, G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215): 1320–1331. doi: 10.1126/science.1253451

Jebb, D., Huang, Z., Pippel, M., Hughes, G.M., Lavrichenko, K., Devanna, P., Winkler, S., Jermiin, L.S., Skirmuntt, E.C., Katzourakis, A., Burkitt-Gray, L., Ray, D.A., Sullivan, K.A.M., Roscito, J.G., Kirilenko, B.M., DÃ¡valos, L.M., Corthals, A.P., Power, M.L., Jones, G., Ransome, R.D., Dechmann, D.K.N., Locatelli, A.G., Puechmaille, S.J., Fedrigo, O., Jarvis, E.D., Hiller, M., Vernes, S.C., Myers, E.W., Teeling, E.C. 2020. Six reference-quality genomes reveal evolution of bat adaptations. *Nature*, 583(7817): 578–584. doi: 10.1038/s41586-020-2486-3

Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H. 2006. Phylogenomics: the beginning of incongruence? *Trends in Genetics*, 22(4): 225–231. doi: 10.1016/j.tig.2006.02.00

Jermiin, L., Ho, S.Y.W., Ababneh, F., Robinson, J., Larkum, A.W. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic Biology*, 53: 638–643. doi: 10.1080/10635150490468648

Johnson, K.P., Dietrich, C.H., Friedrich, F., Beutel, R.G., Wipfler, B., Peters, R.S., Allen, J.M., Petersen, M., Donath, A., Walden, K.K.O., Kozlov, A.M., Podsiadlowski, L., Mayer, C., Meusemann, K., Vasilikopoulos, A., Waterhouse, R.M., Cameron, S.L., Weirauch, C., Swanson, D.R., Percy, D.M., Hardy, N.B., Terry, I., Liu, S., Zhou, X., Misof, B., Robertson, H.M., Yoshizawa, K. 2018. Phylogenomics and the evolution of hemipteroid insects. *Proceedings of the National Academy of Sciences*, 115(50): 12775–12780. doi: 10.1073/pnas.1815820115

Johnson, K.P., Matthee, C., Doña, J. 2022. Phylogenomics reveals the origin of mammal lice out of Afrotheria. *Nature Ecology & Evolution*, 6(8): 1205–1210. doi: 10.1038/s41559-022-01803-1

Kapli, P., Yang, Z., Telford, M.J. 2020. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21: 428–444. doi: 10.1038/s41576-020-0233-0

Katoh, K., Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4): 772–780. doi: 10.1093/molbev/mst010

Kocot, K.M., Struck, T.H., Merkel, J., Waits, D.S., Todt, C., Brannock, P.M., Weese, D.A., Cannon, J.T., Moroz, L.L., Lieb, B., Halanych, K.M. 2017. Phylogenomics of Lophotrochozoa with consideration of systematic error. *Systematic Biology*, 66(2): 256–282. doi: 10.1093/sysbio/syw079

Kück, P., Longo, G.C. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology*, 11: 81. doi: 10.1186/s12983-014-0081-x

Kück, P., Wägele, W.J. 2015. Plesiomorphic character states cause systematic errors in molecular phylogenetic analyses: a simulation study. *Cladistics*, 32: 461–478. doi: 10.1111/cla.12132

Lartillot, N., Brinkmann, H., Philippe, H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, 7: S4. doi: 10.1186/1471-2148-7-S1-S4

Lee, C., Grasso, C., Sharlow, M.F. 2002. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3): 452–464. doi: 10.1093/bioinformatics/18.3.452

Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. *Systematic Biology*, 58(1): 130–145. doi: 10.1093/sysbio/syp017

Lemmon, A.R., Emme, S.A., Lemmon, E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, 61(5): 727–744. doi: 10.1093/sysbio/sys049

Liu, K., Warnow, T.J., Holder, M.T., Nelesen, S.M., Yu, J., Stamatakis, A.P., Linder, C.R. 2012. SATé-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology*, 61(1): 90–106. doi: 10.1093/sysbio/syr095

Liu, L., Zhang, J., Rheindt, F.E., Lei, F., Qu, Y., Wang, Y., Zhang, Y., Sullivan, C., Nie, W., Wang, J., Yang, F., Chen, J., Edwards, S.V., Meng, J., Wu, S. 2017. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. *Proceedings of the National Academy of Sciences*, 114(35): E7282–E7290. doi: 10.1073/pnas.1616744114

Löytynoja, A., Goldman, N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320(5883): 1632–1635. doi: 10.1126/science.1158395

Mai, U., Mirarab, S. 2018. TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, 19(Suppl 5): 272. doi: 10.1186/s12864-018-4620-2

Manthey, J.D., Campillo, L.C., Burns, K.J., Moyle, R.G. 2016. Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: A test in *Cardinalid tanagers* (Aves, Genus: Piranga). *Systematic Biology*, 65(4): 640–650. doi: 10.1093/sysbio/syw005

McKenna, D.D., Shin, S., Ahrens, D., Balke, M., Beza-Beza, C., Clarke, D.J., Donath, A., Escalona, H.E., Friedrich, F., Letsch, H., Liu, S., Maddison, D., Mayer, C., Misof, B., Murin, P.J., Niehuis, O., Peters, R.S., Podsiadlowski, L., Pohl, H., Scully, E.D., Yan, E.V., Zhou, X., Ślipiński, A., Beutel, R.G. 2019. The evolution and genomic basis of beetle diversity. *Proceedings of the National Academy of Sciences*, 116(49): 24729–24737. doi: 10.5281/zenodo.3522944

Metzker, M.L. 2010. Sequencing technologies – the next generation. *Nature Reviews Genetics*, 11: 31–46. doi:10.1038/nrg2626.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R., Teeling, E. 2020. IQ-TREE 2: New Models and efficient methods for phylogenetic inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5): 1530–1534. doi: 10.1093/molbev/msaa015

Mirarab, S., Nguyen, N., Guo, S., Wang, L-S., Kim, J., Warnow, T. 2015. PASTA: Ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology*, 22(5): 377–386. doi: 10.1089/cmb.2014.0156

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A.J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T.B., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L.S., Kawahara, A.Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D.D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J.L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B.M., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N.U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M.G., Wiegmann, B.M., Wilbrandt, J., Wipfler, B., Wong, T.K.F., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D.K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao., J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K.M., Zhou, X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210): 763–767. doi: 10.1126/science.1254426

Nabholz, B., Künstner, A., Wang, R., Jarvis, E.D., Ellegren, H. 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. *Molecular Biology and Evolution*, 28(8): 2197–2210. doi: 10.1093/molbev/msr047

Naser-Khdour, S., Minh, B.Q., Lanfear, R. 2021. Assessing confidence in root placement on phylogenies: An empirical study using nonreversible models for mammals. *Systematic Biology*, 71(4): 959–972. doi: 10.1093/sysbio/syab067

Nesnidal, M.P., Helmkampf, M., Bruchhaus, I., Hausdorf, B. 2010. Compositional heterogeneity and phylogenomic inference of metazoan relationships. *Molecular Biology and Evolution*, 27(9): 2095–2104. doi: 10.1093/molbev/msq097

Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Müller, W.E.G., Nickel, M., Schierwater, B., Vacelet, J., Wiens, M., Wörheide, G. 2013. Deep metazoan phylogeny: when different genes tell different stories. *Molecular Phylogenetics and Evolution*, 67(1): 223–233. doi: 10.1016/j.ympev.2013.01.010

Notredame, C., Higgins, D.G., Heringa, J. 2003. T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1): 205–217. doi: 10.1006/jmbi.2000.4042

Novák, Á., Miklós, I., Lyngsø, R., Hein, J. 2008. StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics*, 24: 2403–2404. doi: 10.1093/bioinformatics/btn457

O'Brien, S.J., Stanyon, R. 1999. Phylogenomics: ancestral primate viewed. *Nature*, 402(6760): 365–366. doi: 10.1038/46450

Pais, F.S.M., Ruy, P.C., Oliveira, G., Coimbra, R.S. 2014. Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology*, 9(1): 4. doi: 10.1186/1748-7188-9-4

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology*, 9(3): e1000602. doi: 10.1371/journal.pbio.1000602

Phillips, M.J., Penny, D. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. *Molecular Phylogenetics and Evolution*, 28(2): 171–185. doi: 10.1016/s1055-7903(03)00057-5

Pick, K.S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D.J., Wrede, P., Wiens, M., Alié, A., Morgenstern, B., Manuel, M., Wörheide, G. 2010. Improved phylogenomic taxon sampling noticeably affects Nonbilaterian relationships. *Molecular Biology and Evolution*, 27(9): 1983–1987. doi: 10.1093/molbev/msq089

Qu, X.J., Jin, J.J., Chaw, S.M., Li, D.Z., Yib, T.S. 2017. Multiple measures could alleviate long-branch attraction in phylogenomic reconstruction of Cupressoideae (Cupressaceae). *Scientific Reports*, 7: 41005. doi: 10.1038/srep41005

Rannala, B., Yang, Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4): 1645–1656. doi: 10.1093/genetics/164.4.1645

Richards, E.J., Brown, J.M., Barley, A.J., Chong, R.A., Thomson, R.C. 2018. Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological? *Systematic Biology*, 67(5): 847–860. doi: 10.1093/sysbio/syy013

Roshan, U., Livesay, D.R. 2006. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, 22(22): 2715–2721. doi: 10.1093/bioinformatics/btl472

Roure, B., Baurain, D., Philippe, H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular Biology and Evolution*, 30(1): 197–214. doi: 10.1093/molbev/mss208

Salichos, L., Rokas, A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, *497*: 327–331. doi: 10.1038/nature12130

Sela, I., Ashkenazy, H., Katoh, K., Pupko, T. 2015. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Research*, 43: W7–W14. doi: 10.1093/nar/gkq443

Shen, X-X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K.V., Haase, M.A.B., Wisecaver, J.H., Wang, M., Doering, D.T., Boudouris, J.T., Schneider, R. M., Langdon, Q.K., Ohkuma, M., Endoh, R., Takashima, M., Manabe, R. ichiroh, Čadež, N., Libkind, D., Rosa, C.A., DeVirgilio, J., Hulfachor, A.B., Groenewald, M., Kurtzman, C.P., Hittinger, C.T., Rokas, A. 2018. Tempo and mode of genome evolution in the budding yeast subphylum. *Cell*, 175(6): 1533–1545. doi: 10.1016/j.cell.2018.10.023

Shen, X-X., Salichos, L., Rokas, A. 2016. A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. *Genome Biology and Evolution*, *8*(8): 2565–2580. doi: 10.1093/gbe/evw179

Shen, X-X., Steenwyk, J.L., Rokas, A. 2021. Dissecting incongruence between concatenation- and quartet-based approaches in phylogenomic data. *Systematic Biology*, 70(5): 997–1014. doi: 10.1093/sysbio/syab011

Sherry, S.T., Ward, M., Sirotkin, K. 1999. dbSNP — database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Research*, 9(8): 677–679. doi: 10.1101/gr.9.8.677

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7: 539. doi: 10.1038/msb.2011.75

Simmons, M.P. 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. *Molecular Phylogenetics and Evolution*, 80: 267–280. doi: 10.1016/j.ympev.2014.08.021

Smirnov, V., Warnow, T. 2021. MAGUS: Multiple sequence Alignment using Graph clUStering. *Bioinformatics*, 37: 1666–1672. doi: 10.1093/bioinformatics/btaa992

Smith, B.T., Mauck, W.M., Benz, B.W., Andersen, M.J. 2018. Uneven missing data skew phylogenomic relationships within the Lories and Lorikeets. *Genome Biology and Evolution*, 12(7): 1131–1147. doi: 10.1093/gbe/evaa113

Song, N., Zhang, H. 2022. A comprehensive analysis of higher-level phylogenetic relationships of Hemiptera based on transcriptome data. *Journal of Systematics and Evolution*. 1–15. doi: 10.1111/jse.12855

Steenwyk, J.L., Buida, T.J., Labella, A.L., Li, Y., Shen, X-X., Rokas, A. 2020a. PhyKIT: A UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics*, 37(16): 2325–2331. doi: 10.1093/bioinformatics/btab096

Steenwyk, J.L., Buida, T.J., Li, Y., Shen, X-X., Rokas, A. 2020b. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biology*, 18(12): e3001007. doi: 10.1371/journal.pbio.3001007

Strimmer, K., von Haeseler, A. 1997. Likelihood-mapping: A simple method to visualize phylogenetic. *Proceedings of the National Academy of Sciences*, 94(13): 6815–6819. doi: 10.1073/pnas.94.13.6815

Struck, T.H. 2014. TreSpEx-detection of misleading signal in phylogenetic reconstructions based on tree information. *Evolutionary Bioinformatics*, 10: 51–67. doi: 10.4137/EBO.S14239

Suchard, M.A., Redelings, B.D. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, 22(16): 2047–2048. doi: 10.1093/bioinformatics/btl175

Sun, X., Ding, Y., Orr, M.C., Zhang, F. 2020. Streamlining universal single-copy orthologue and ultraconserved element design: A case study in Collembola. *Molecular Ecology Resources*, 20: 706–717. doi: 10.1111/1755-0998.13146

Talavera, G., Castresana, J. 2007. Improvement of Phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, 56(4): 564–577. doi: 10.1080/10635150701472164

Talavera, G., Vila, R. 2011. What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny. *BMC Evolutionary Biology*, 11: 315. doi: 10.1186/1471-2148-11-315

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., Dessimoz, C. 2015. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Systematic Biology*, 64(5): 778–791. doi: 10.1093/sysbio/syv033

Tihelka, E., Cai, C., Giacomelli, M., Lozano-Fernandez, J., Rota-Stabelli, O., Huang, D., Engel, M.S., Donoghue, P.C.J., Pisani, D. 2021. The evolution of insect biodiversity. *Current Biology*, 31(19): 1299–1311. doi: 10.1016/j.cub.2021.08.057

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., Zdobnov, E.M. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, 35: 543–548. doi: 10.1093/molbev/msx319

Wipfler, B., Letsch, H., Frandsen, P.B., Kapli, P., Mayer, C., Bartel, D., Buckley, T.R., Donath, A., Edgerly-Rooks, J.S., Fujita, M., Liu, S., Machida, R., Mashimo, Y., Misof, B., Niehuis, O., Peters, R.S., Petersen, M., Podsiadlowski, L., Schütte, K., Shimizu, S., Uchifune, T., Wilbrandt, J., Yan, E., Zhou, X., Simon, S. 2019. Evolutionary history of Polyneoptera and its implications for our understanding of early winged insects. *Proceedings of the National Academy of Sciences*, 116(8): 3024–3029. doi: 10.1073/pnas.1817794116

Young, A.D., Gillung, J.P. 2020. Phylogenomics – principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology*, 45(2): 225–247. doi: 10.1111/syen.12406

Yu, D., Ding, Y., Tihelka, E., Cai, C., Hu, F., Liu, M., Zhang, F. 2022. Phylogenomics of elongate-bodied springtails reveals independent transitions from aboveground to belowground habitats in deep time. *Systematic Biology*, 71(5): 1023–1031. doi: 10.1093/sysbio/syac024

Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(6): 153. doi: 10.1186/s12859-018-2129-y

Zhang, D., Niu, Z., Luo, A., Orr, M.C., Ferrari, R.R., Jin, J., Wu, Q., Zhang, F., Zhu, C-D. 2022. Testing the systematic status of *Homalictus* and *Rostrohalictus* with weakened cross-veingroups within Halictini (Hymenoptera: Halictidae) using low-coverage whole-genome sequencing. *Insect Science*, 29(6): 1819–1833. doi: 10.1111/1744-7917.13034

Zhang, F., Ding, Y., Zhu, C-D., Zhou, X., Orr, M.C., Scheu, S., Luan, Y-X. 2019. Phylogenomics from low-coverage whole-genome sequencing. *Methods in Ecology and Evolution*, 10: 507–517. doi: 10.1111/2041-210X.13145

Zhong, B.J., Yonezawa, T., Zhong, Y., Hasegawa, M. 2010. The position of Gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. *Molecular Biology and Evolution*, 27(12): 2855–2863. doi: 10.1093/molbev/msq170