Supplementary Appendix A for ***A new story of the four Hexapoda classes: Protura as the sister group to all other hexapods***

—— Description of filtering strategy selection and matrix generation

Previous studies have shown that some gene properties are strongly correlated with phylogenetic signal (1, 2). Filtering genes based on their properties is a crucial step in phylogenomic studies (2). In the last few years, much effort has been devoted to uncovering and understanding the roles of gene properties in phylogenomic reconstruction (1–3). A total of ten properties were examined for each gene in this study, which we classified into two categories — four sequence-based (TAPER (4), nRCFV (5), SRH (6), API (7)), and six tree-based properties (PSH (8), ABS values (9), DVMC (10), treeness (11), treeness over rcv (11), and saturation (12)) — computed using custom scripts (13). Their full descriptions are provided in Table S1). We tested the effect of these gene properties on three proxies for phylogenetic signal: Phylogenetic signal was evaluated based on ABS values, Robinson-Foulds (RF) distances (14), and wASTRAL topologies.

First, filtering was performed with TAPER and TreeShrink (8), using default parameters (PSH analysis). Then, we focused on nRCFV and SRH that quantify compositional heterogeneity of alignments. We applied four thresholds (0.002, 0.003, 0.004, and 0.005) to nRCFV and examined the values for all genes, as shown in Figure S1a. Similarly, for assessing compositional homogeneity, we used two thresholds (0.05 and 0.1) for SRH and presented the values for all loci in Figure S1b. To evaluate the impact of these gene properties on phylogenetic reconstruction, we constructed wASTRAL trees using the matrices created from the sites retained by each threshold for nRCFV and SRH. Notably, all resulting trees strongly supported H1, as depicted in Figures S2 and S3. For detailed information on the number of loci, sites, and average locus length retained by each threshold, refer to Table S2. Additionally, the ABS values and RF distances of loci corresponding to different thresholds are displayed in Figures S4a, S4b, S5a, and S5b. To retain the maximum number of loci, we selected the threshold values of 0.004 for nRCFV (the “turning point” in Figure S1a) and 0.5 (the default value) for SRH. By combining these threshold values, we created the ‘symtest-matrix’ which included 949 loci, which served as the basis for subsequent analyses.

We then proceeded to investigate the six remaining properties of each analysed gene (API, ABS, DVMC, treeness, treeness over rcv, and saturation). Different thresholds were applied to filter the gene properties; the threshold and the resultant values are depicted in Figures S1c to S1h. For the API property, which was used to calculate the values of average pairwise identity (lower values indicating slowly evolving genes) three thresholds were used: 0.6, 0.7, and 0.8. The wASTRAL trees constructed using the matrices created by the sites retained at the threshold lower than 0.6 strongly supported H3, while the other thresholds supported H2 (Figure S6). ABS filtering employed five thresholds: 55, 60, 65, 70, and 75. Loci of higher ABS values are thought to harbor more phylogenetic signal. The topology of the wASTRAL tree constructed using the matrices created by the sites retained at the threshold greater than 55 supported H1,

while the remaining thresholds supported H2 (Figure S7). We further quantified the genic deviation from the assumptions of a molecular clock (i.e., DVMC), under five thresholds: 0.5, 1, 1.5, 2, and 4 (Figure S1e). The wASTRAL trees strongly supported H1 (Figure S8). Treeness describes the signal-to-noise ratio in a phylogeny, whereby higher values of treeness are thought to be desirable (11) for phylogenetic inference. We set five thresholds: 0.1, 0.15, 0.17, 0.18, and 0.19 (Figure S1f). The wASTRAL trees constructed using the matrices created by the sites retained greater than each treeness threshold also strongly supported H1 (Figure S9). For treeness over rcv, three thresholds were used: 0.5, 0.8, and 1.1. When the threshold value greater than 1.1 was chosen, the topology of the wASTRAL tree constructed using this matrix created by the sites retained supported H2, whereas the other thresholds supported the topology [Collembola + Diplura] + [Protura + Insecta] (Figure S10). In the saturation analysis, two thresholds were considered: 0.1 and 0.4. Data with no saturation will have higher values. When the threshold > 0.1 was chosen, the topology of the wASTRAL tree constructed using the matrices created by the retained sites supported H1, while the threshold greater than 0.4 supported the topology of [Collembola + Diplura] + [Protura + Insecta] (Figure S11). All thresholds ensured that the number of retained loci was not less than 200, as specified in Table S2, which provides information on the number of loci, number of sites, and average locus length retained for each threshold. Additionally, ABS values and RF distances of the loci were calculated for each matrix (Figure S4, S5). Notably, significant differences were observed in ABS values (Figure S4d, S5d) and treeness (Figure S4f, S5f). However, for the API property, only ABS values showed significant differences (Figure S4c). Therefore, the ABS values and RF distance of loci do not differ significantly for different gene properties, even under different threshold (apart from API, ABS, and treeness).

In addition to the thresholds discussed above, the influence of missing data, i.e., the absence of some sites for some taxa, was considered. Five different thresholds (50, 60, 70, 80, and 90) were set, corresponding to the proportion of missing data in each dataset. Table S2 provides basic information on the thresholds, including the number of loci, number of sites, and average locus length retained for each threshold. The wASTRAL trees were constructed using the matrices created by the sites retained at each missing data threshold, and all results supported H1 (Figure S12). Notably, the ABS values and RF distances of loci did not differ significantly for the missing data gene properties, even with different thresholds (Figure S4h, S5h).

In summary, topological changes in the wASTRAL trees were observed for API and ABS at thresholds of 0.6 and 70, respectively. We concluded that ABS and API have an impact on phylogenetic inference in our dataset, but DVMC, and treeness, treeness over rcv, missing data, and saturation were not strongly correlated with phylogenetic signal. Two new matrices, namely ‘API-matrix’ and ‘ABS-matrix’, were created based on these thresholds. Importantly, we followed the rule of thumb that when selecting a threshold, the number of loci retained should not be less than half of the original dataset, which corresponded to retaining at least 475 loci in our case.

Table S1. Information on the ten gene properties (filtering strategies) used in this study.

Property	Name	Description
Sequence-based	TAPER	Two-dimensional Algorithm for Pinpointing ERrors
	nRCFV	normalised Relative Compositional Frequency Variation
	API	Average Pairwise Identity
	SRH	Stationary, Reversible and Homogeneous
Tree-based	ABS	Average Bipartition Support
	DVMC	Degree of Violation of the Molecular Clock
	Treeness	Proportion of sum of internal branch lengths over sum of all branch lengths across the maximum likelihood tree of a given alignment
	Treeness over RCV	Treeness divided by RCV
	PSH	Potentially Spurious Homologs
	Saturation	The sequences in multiple sequence alignments that have undergone numerous substitutions such that the distances between taxa are underestimated

Table S2. Summary of the number of loci, number of sites, and average locus length for the matrices created using the sites retained at each threshold of nine distinct gene properties.

Gene property	threshold	Number of loci	Number of sites	Average locus length	wASTRAL topology
nRCFV	0.002	589	290,160	492.63	H1
	0.003	917	622,964	679.35	H1
	0.004	994	742,413	746.89	H1
	0.005	1,011	789,799	781.20	H1
SRH	0.05	949	685,769	722.62	H1
	0.1	915	653,885	714.62	H1
API	0.6	475	374,376	788.16	H3
	0.7	763	577,673	757.10	H2
	0.8	904	664,304	734.84	H2
ABS	55	935	681,300	728.66	H1
	60	885	664,515	750.86	H2
	65	763	611,047	800.84	H2
	70	554	499,622	901.84	H2
	75	280	294,924	1,053.30	H2
DVMC	0.5	349	212,482	608.83	H1
	1	749	509,478	680.21	H1
	1.5	862	602,252	698.66	H1
	2	900	637,395	708.21	H1
	4	939	677,449	721.45	H1
treeness	0.1	947	684,669	722.98	H1
	0.15	680	488,193	717.93	H1
	0.17	452	329,502	728.98	H1
	0.18	361	264,806	733.53	H1
	0.19	270	200,448	742.40	H1
treeness over rcv	0.5	883	604,438	684.52	H1
	0.8	564	362,770	643.20	H1
	1.1	249	151,690	609.19	(C+D)+(P+I)
missing data	50	948	685,373	722.97	H1
	60	939	679,407	723.54	H1
	70	900	635,031	705.59	H1
	80	787	539,939	686.07	H1
	90	470	333,303	709.16	H1
saturation	0.1	861	577,914	671.21	H1
	0.4	371	195,529	527.03	(C+D)+(P+I)

Note: H1, Collembola + (Protura + (Diplura + Insecta)); H2, (Collembola + Protura) + (Diplura + Insecta); H3, Protura + ((Collembola + Diplura) + Insecta).

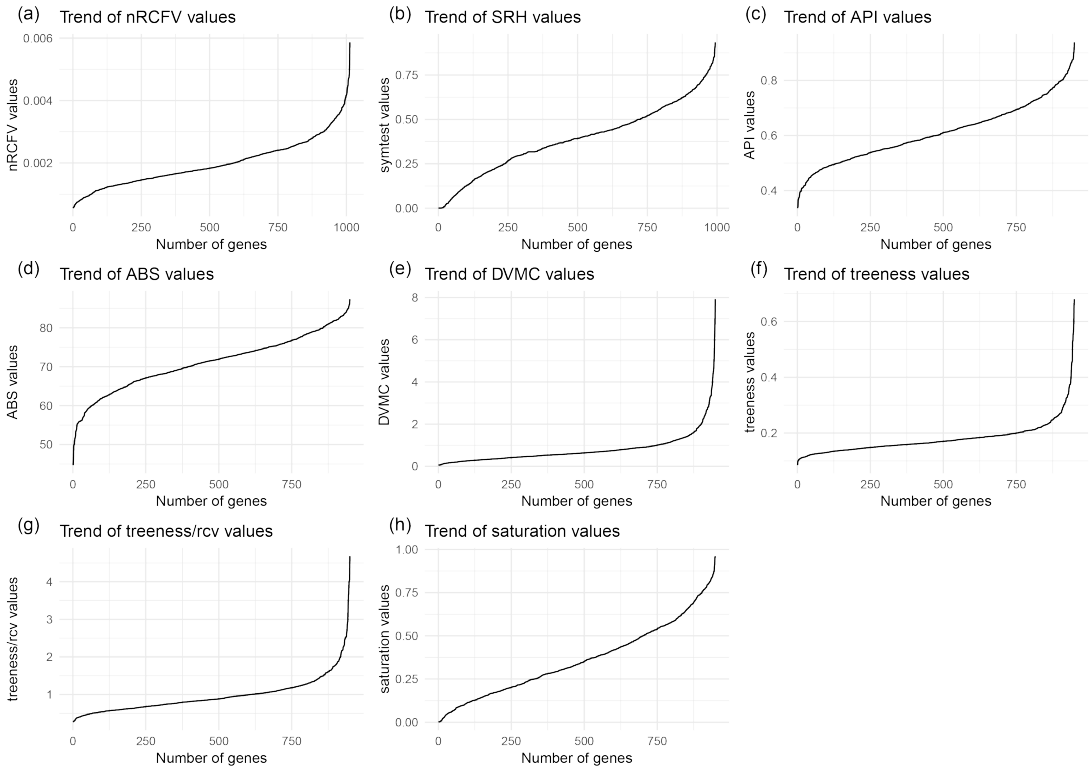

Figure S1. The values of different gene properties across different genes. (a) nRCFV (normalised Relative Compositional Frequency Variation); (b) SRH (Stationary, Reversible and Homogeneous); (c) API (Average Pairwise Identity); (d) ABS (Average Bipartition Support value); (e) DVMC (Degree of Violation of the Molecular Clock); (f) treeness; (g) treeness/rcv (treeness over rcv); (h) saturation.

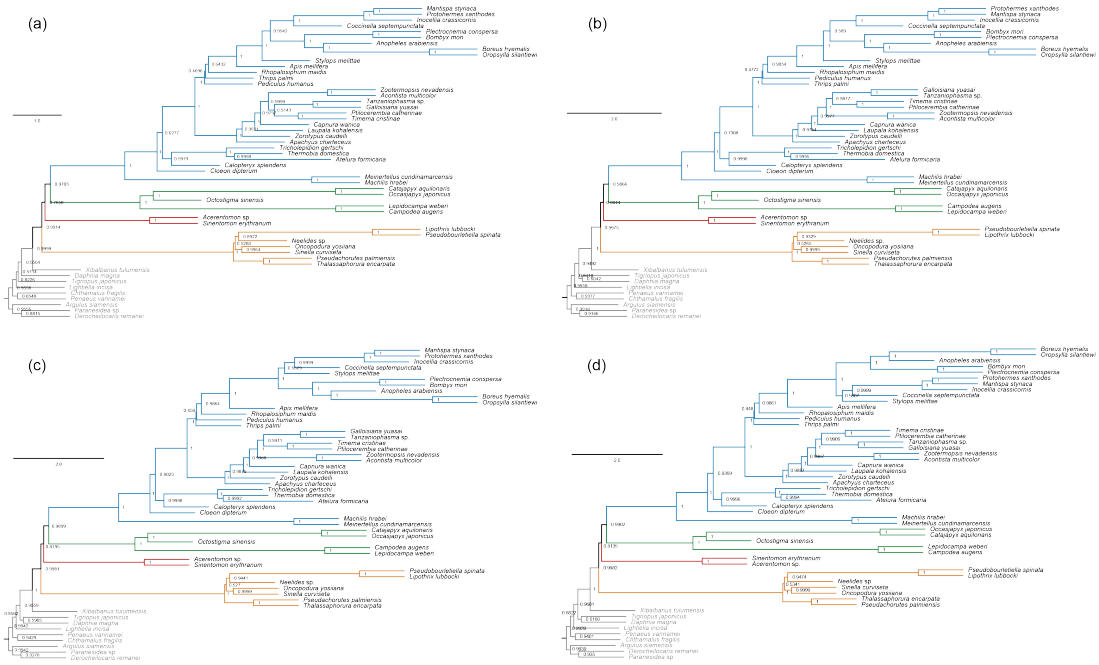

Figure S2. The phylogenetic trees constructed using wASTRAL with the matrices created from the sites retained under four different thresholds of nRCFV. (a) the threshold of 0.002; (b) the threshold of 0.003; (c) the threshold of 0.004; (d) the threshold of 0.005.

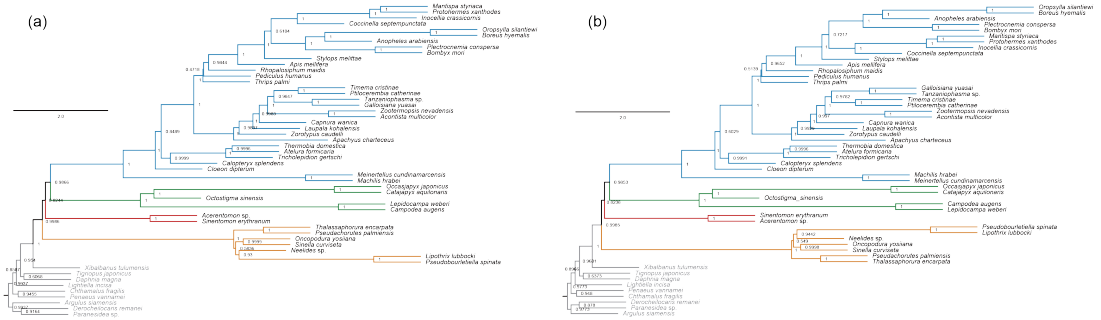

Figure S3. The phylogenetic trees constructed using wASTRAL with the matrices created from the sites retained under two different thresholds of SRH. (a) the threshold of 0.05; (b) the threshold of 0.1.

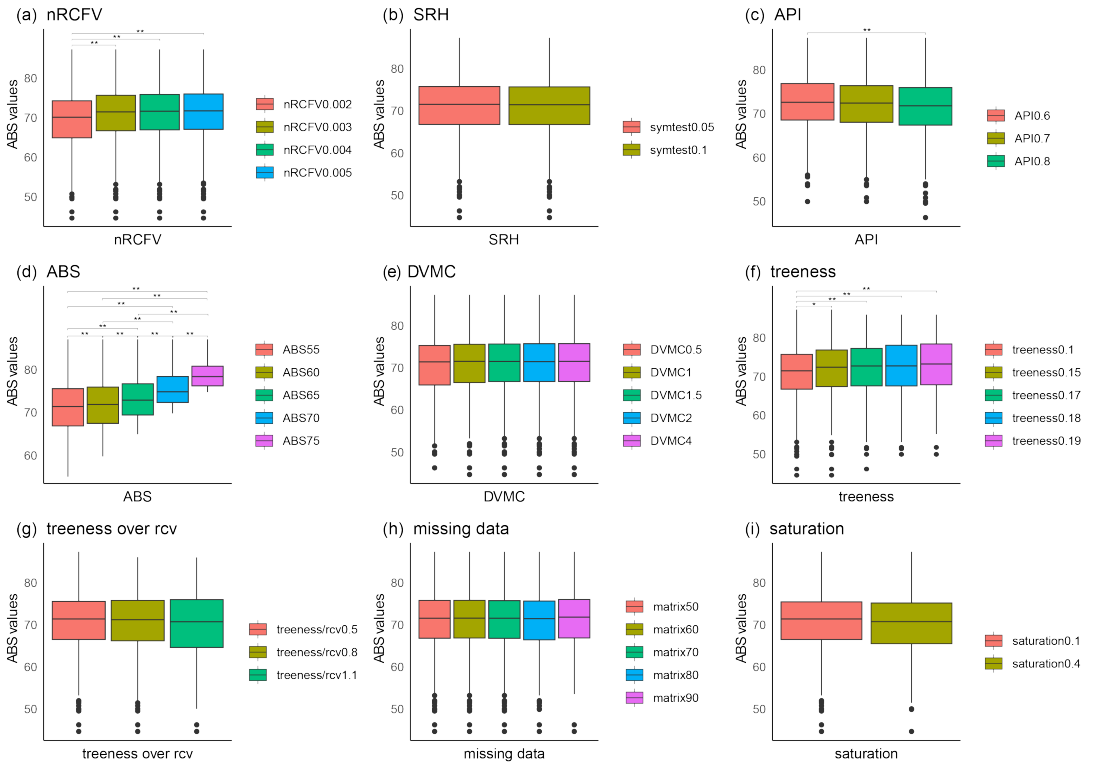

Figure S4. ABS values of matrix statistics created by loci under different threshold retention of (a) nRCFV, (b) SRH, (c) API, (d) ABS, (e) DVMC, (f) treeness, (g) treeness over rcv, (h) missing data, and (i) saturation. The top value corresponds to the p-value (* $p \leq 0.05$; ** $p \leq 0.01$) show that there is significant difference between two matrices. No significant difference between two matrices is not shown.

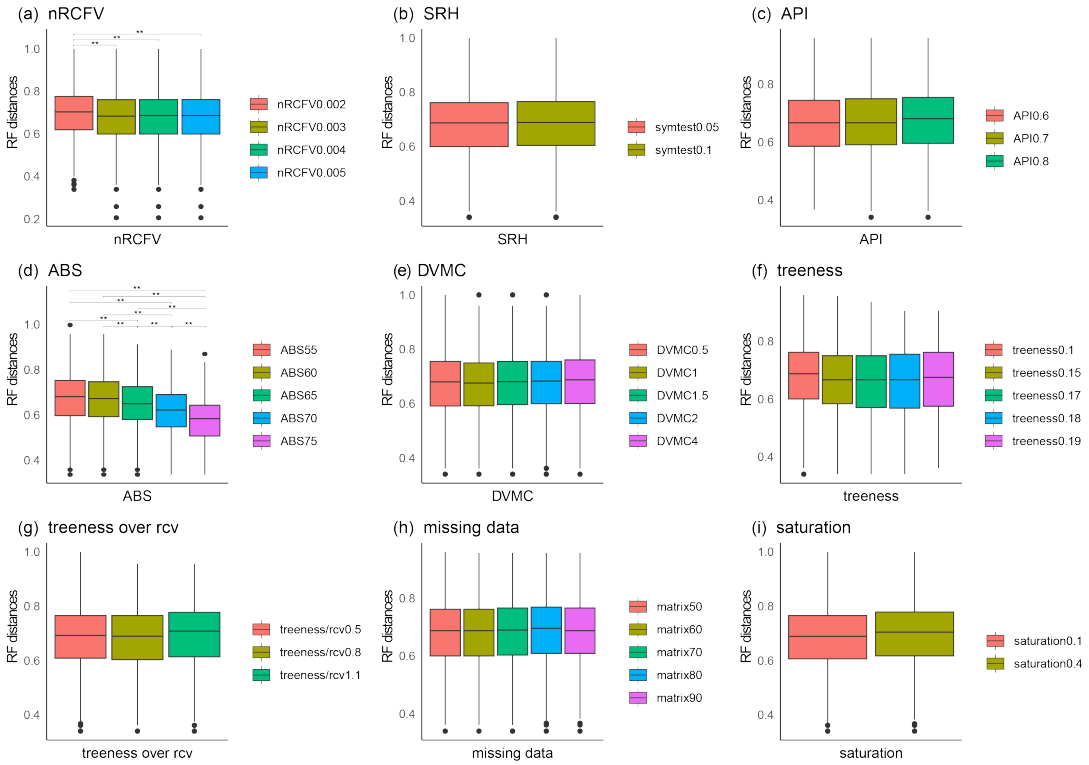

Figure S5. RF distances of matrix statistics created by loci under different threshold retention of (a) nRCFV, (b) SRH, (c) API, (d) ABS, (e) DVMC, (f) treeness, (g) treeness over rcv, (h) missing data, and (i) saturation. The top value corresponds to the p-value (* $p \leq 0.05$; ** $p \leq 0.01$) show that there is significant difference between two matrices. No significant difference between two matrices is not shown.

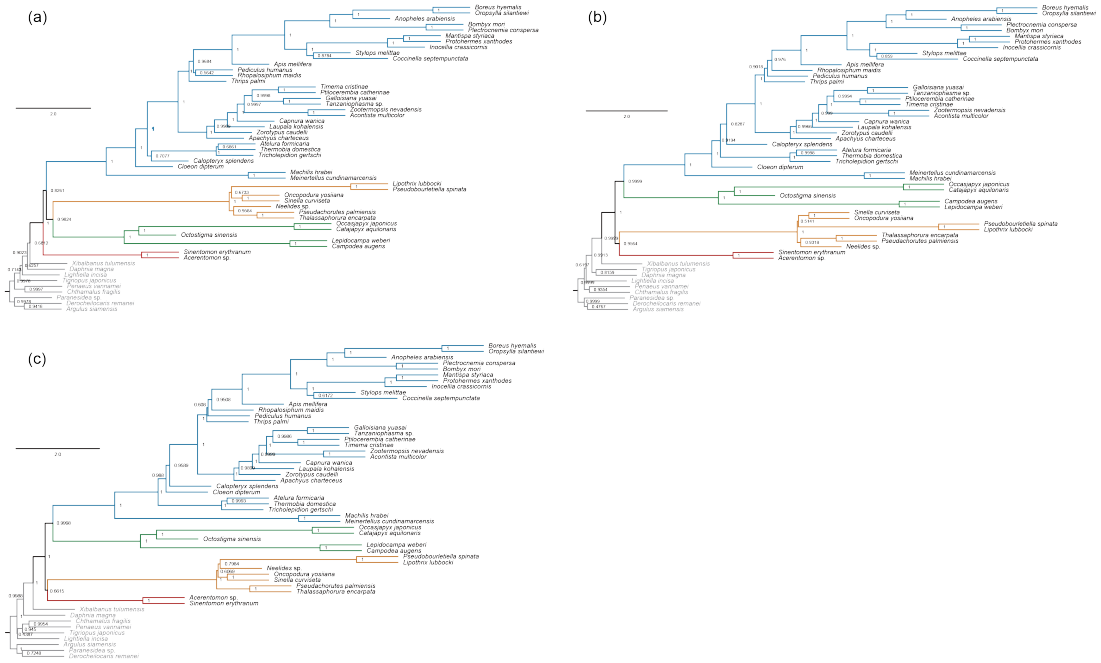

Figure S6. The phylogenetic trees constructed using wASTRAL with the matrices created from the sites retained under three different thresholds of API. (a) the threshold of 0.6; (b) the threshold of 0.7; (c) the threshold of 0.8.

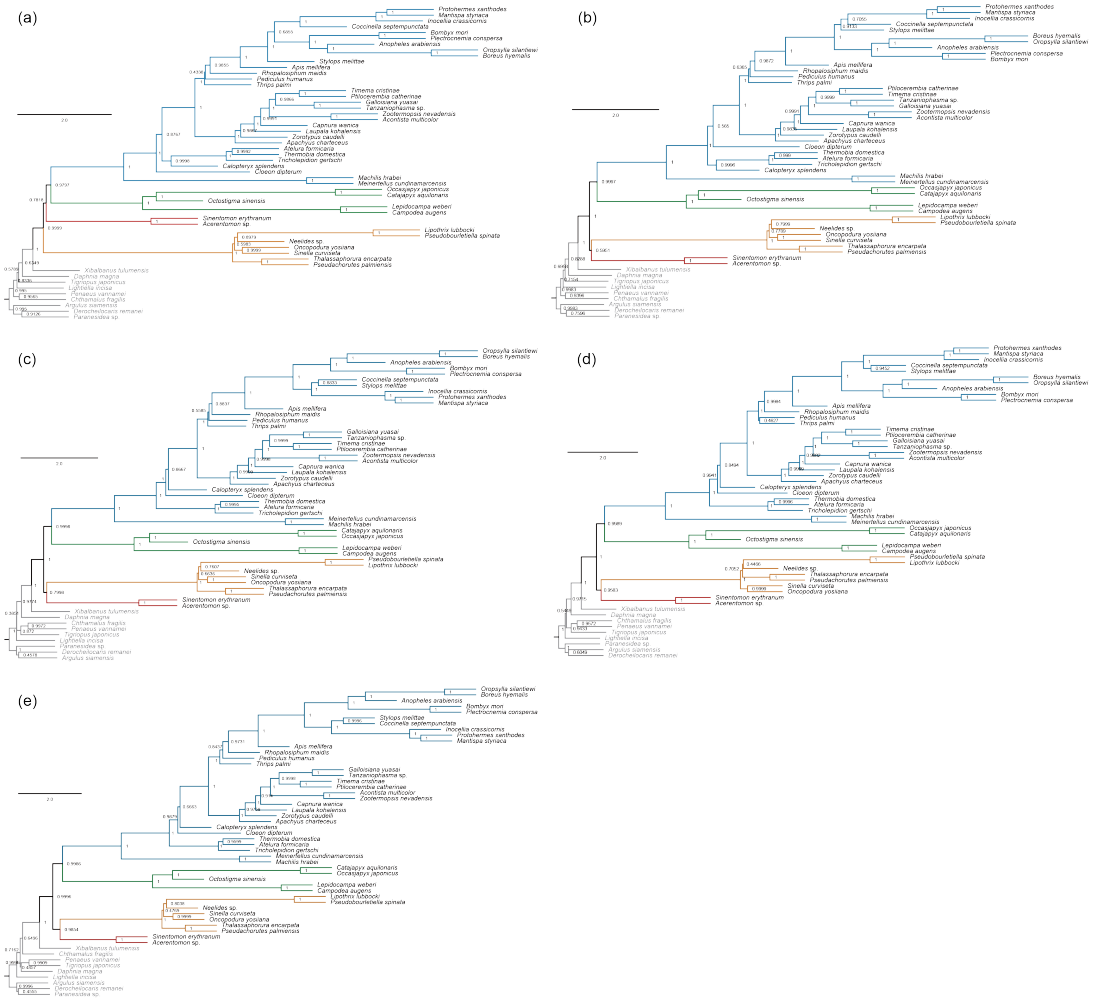

Figure S7. The phylogenetic trees constructed using wASTRAL with the matrices created from the sites retained under five different thresholds of ABS. (a) the threshold of 55; (b) the threshold of 60; (c) the threshold of 65; (d) the threshold of 70; (e) the threshold of 75.

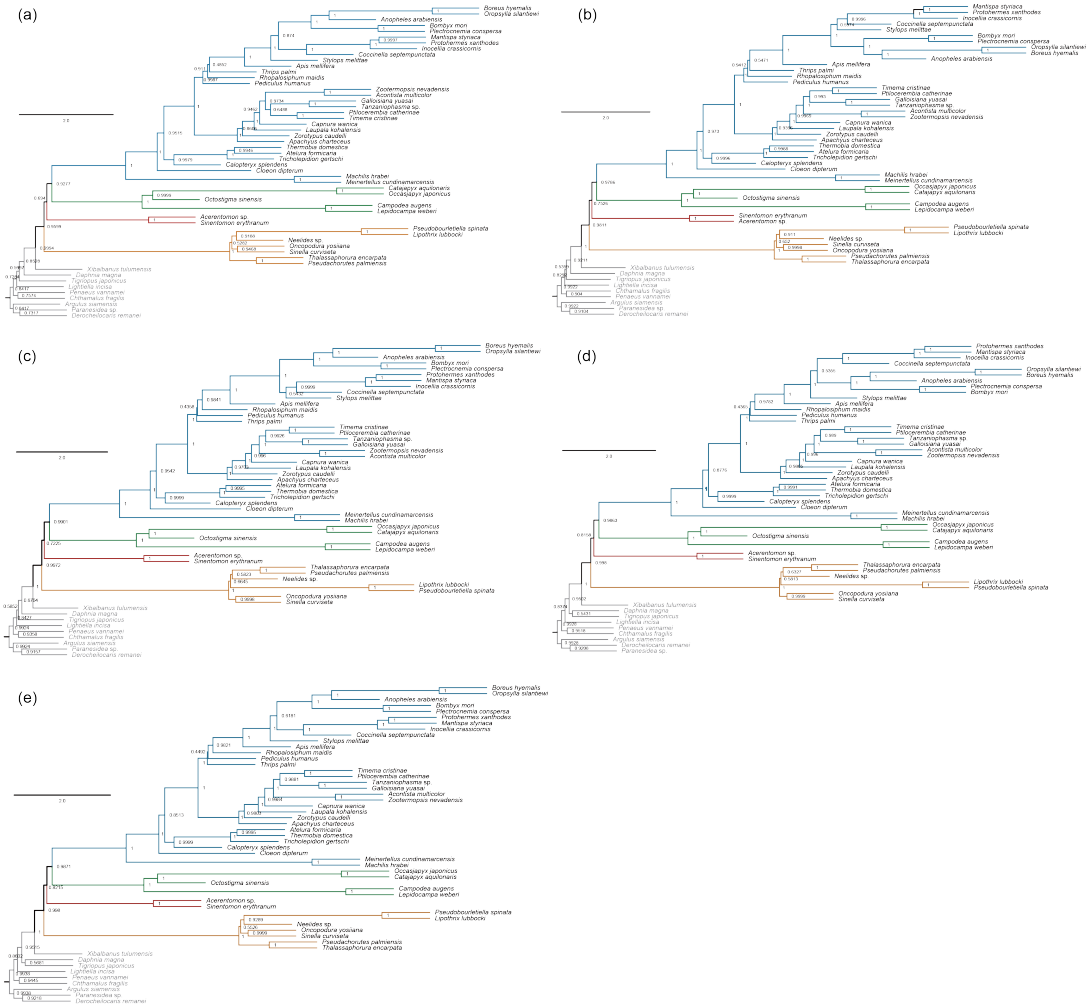

Figure S8. The phylogenetic trees constructed using wASTRAL with the matrices created from the sites retained under five different thresholds of DVMC. (a) the threshold of 0.5; (b) the threshold of 1; (c) the threshold of 1.5; (d) the threshold of 2; (e) the threshold of 4.

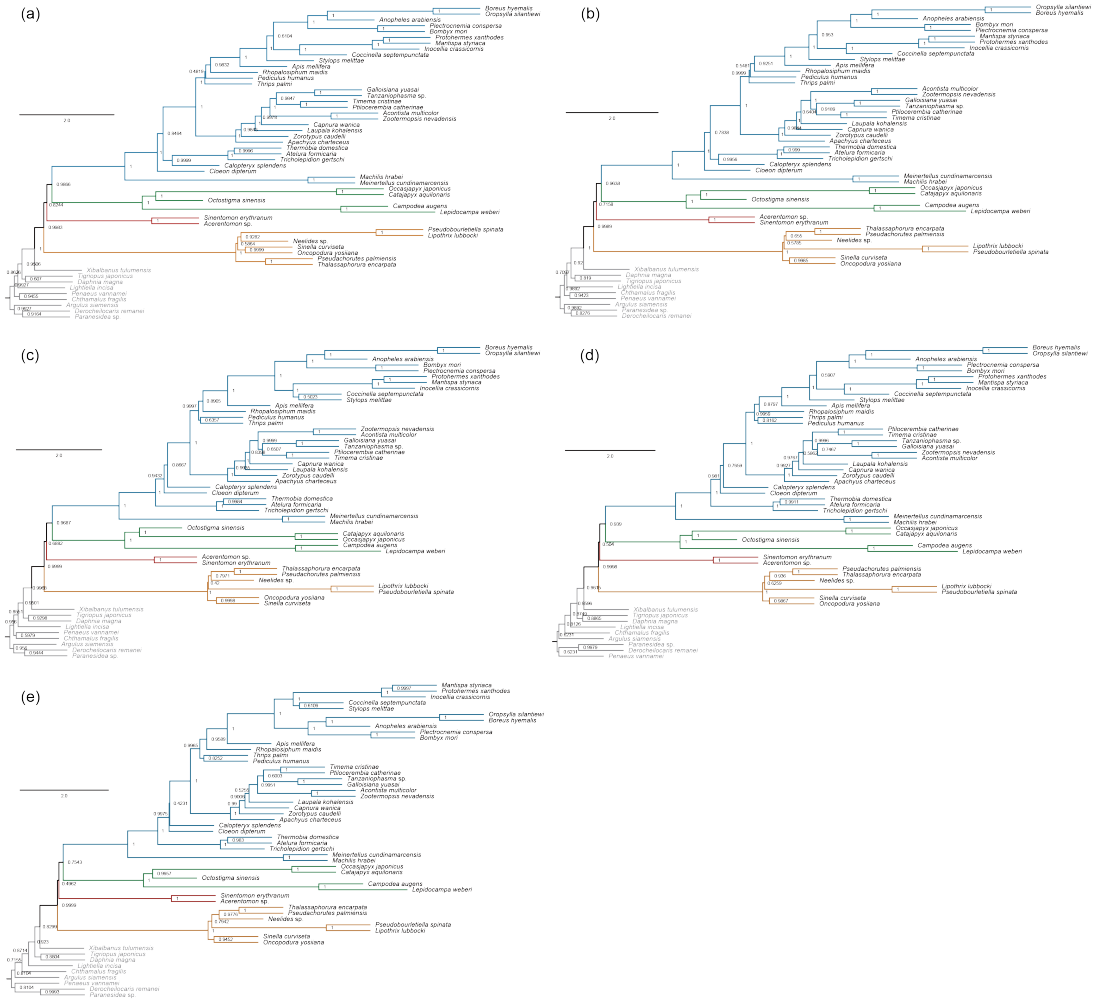

Figure S9. The phylogenetic trees constructed using wASTRAL with the matrices created from the sites retained under five different thresholds of treeness. (a) the threshold of 0.1; (b) the threshold of 0.15; (c) the threshold of 0.17; (d) the threshold of 0.18; (e) the threshold of 0.19.

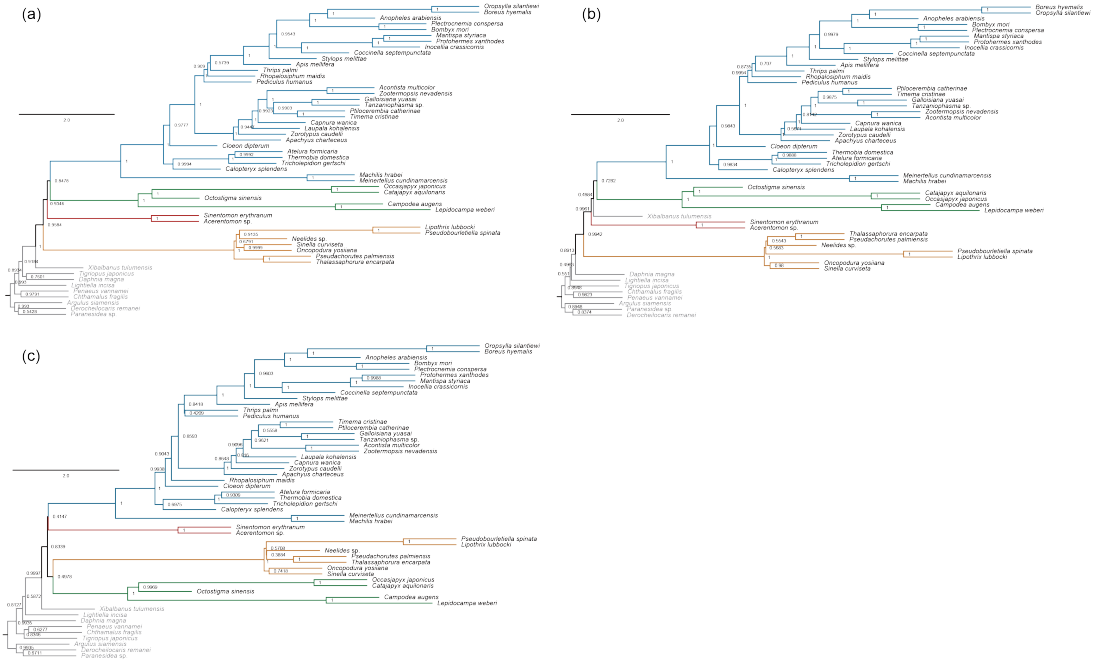

Figure S10. The phylogenetic trees constructed using wASTRAL with the matrices created from the sites retained under three different thresholds of treeness over rcv. (a) the threshold of 0.5; (b) the threshold of 0.8; (c) the threshold of 1.1.

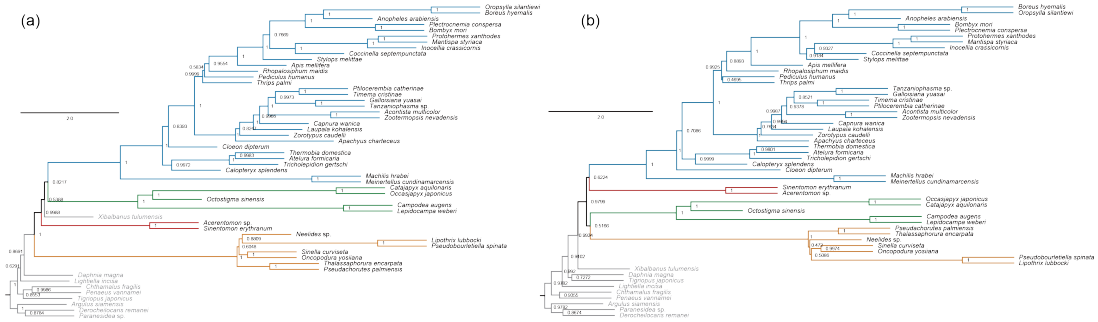

Figure S11. The phylogenetic trees constructed using wASTRAL with the matrices created from the sites retained under two different thresholds of saturation. (a) the threshold of 0.1; (b) the threshold of 0.4.

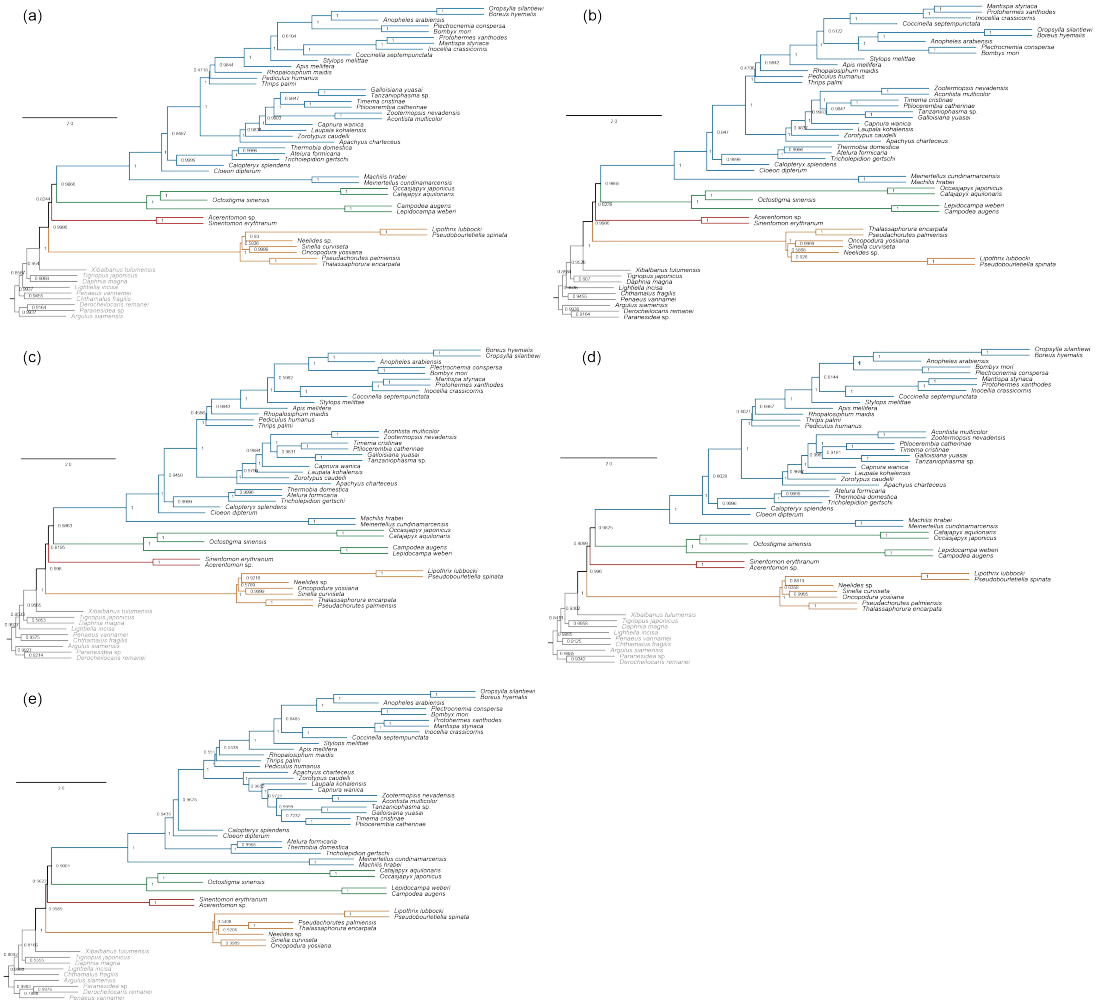

Figure S12. The phylogenetic trees constructed using wASTRAL with the matrices created from the sites retained under five different thresholds of missing data. (a) the threshold of 50; (b) the threshold of 60; (c) the threshold of 70; (d) the threshold of 80; (e) the threshold of 90.

References

1. P. Kapli, Z. Yang, M. J. Telford, Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* 21, 428–444 (2020).
2. A. D. Young, J. P. Gillung, Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. *Syst. Entomol.* 45, 225–247 (2020).
3. X.-X. Shen, J. L. Steenwyk, A. Rokas, Dissecting incongruence between concatenation- and quartet-based approaches in phylogenomic data. *Syst. Biol.* 70, 997–1014 (2021).
4. C. Zhang, Y. Zhao, E. L. Braun, S. Mirarab, TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution. *Methods Ecol. Evol.* 12, 2145–2158 (2021).
5. J. F. Fleming, T. H. Struck, nRCFV: a new, dataset-size-independent metric to quantify compositional heterogeneity in nucleotide and amino acid datasets. *BMC Bioinform.* 24, 145 (2023).
6. B. Q. Minh, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534 (2020).
7. M.-Y. Chen, D. Liang, P. Zhang, Phylogenomic resolution of the phylogeny of laurasiatherian mammals: exploring phylogenetic signals within coding and noncoding sequences. *Genome Biol. Evol.* 9, 1998–2012 (2017).
8. U. Mai, S. Mirarab, TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* 19, 272 (2018).
9. L. Salichos, A. Rokas, A. Rokas, Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497, 327–331 (2013).
10. L. Liu, et al. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. *Proc. Natl. Acad. Sci.* 114, E7282–E7290 (2017).
11. M. J. Phillips, D. Penny, The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol. Phylogenet. Evol.* 28, 171–185 (2003).
12. H. Philippe, et al. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* 9, e1000602 (2011).
13. S. Du, et al. Construction of a phylogenetic matrix: scripts and guidelines for phylogenomics. *Zool. Syst.* 48, 107–116 (2023).
14. D. F. Robinson, L. R. Foulds, Comparison of phylogenetic trees. *Mat. Biosci.* 53, 131–147 (1981).