

## Supplementary Text

### Phylogenomic analyses reveal paraphyly across Pill-Millipede families and subfamilies (Diplopoda: Glomerida)

Oeyen et al.

#### 1. Sample collection and preparation

With the exception of *Geoglomeris subterranea*, all samples were collected by hand. Samples of *G. subterranea* were collected via Berlese extractions of soil and leaf-litter. For details on collecting sites and voucher specimens deposited please see Table 1 and S1. Samples used for RNA extractions were ground with a plastic pestle, conserved in RNAlater (Qiagen, Hilden, Germany), and stored at -80 °C.

#### 2. Extraction, library preparation, and sequencing.

The samples of *Adenomeris gibbosa*, *Eupeyerimhoffia archimedis*, *Geoglomeris subterranea*, *Onomeris underwoodi*, *Rhopalomeris carnifex*, *Trachysphaera* sp., and *Typhloglomeris martensi* were extracted and sequenced by StarSEQ GmbH (Mainz, Germany; [www.starseq.com](http://www.starseq.com)) following their protocols. The tissue was homogenized using Precellys grinding balls in a Bertin Minilys Homogenizier, Total RNA was isolated using the Total RNA Kit peqGOLD (VWR International GmbH, Darmstadt, Germany), and paired-end cDNA libraries were constructed using the Illumina TruSeq RNA stranded HT kit. The libraries were sequenced on an Illumina NextSeq 500 sequencer with a read length of 150 bp, obtaining ~50 million reads for each sample. The samples of *Hyleoglomeris* sp. Japan and *Protoglomeris vasconica* were extracted and sequenced according to the protocols described by Misof et al. (2014) with the modifications described in Peters et al. (2017). All raw reads obtained for this study were submitted to the NCBI SRA database, please see Table 2 for accession numbers.

#### 3. Data processing and assembly

The reads of all samples, except *Hyleoglomeris* sp. Japan and *Protoglomeris vasconica* but including those downloaded from NCBI, were screened for adapters and trimmed using Cutadapt (version 1.18; Martin 2011) as implemented in Trim Galore (version 0.5.0; [github.com/FelixKrueger/TrimGalore](https://github.com/FelixKrueger/TrimGalore)). Reads were trimmed requiring a minimum Phred score of 25 and retaining only reads that have a minimum length of 75 bp after trimming. In the case of *Glomeridesmus* sp. and *Brachygybe* sp., the downloaded reads were only 50 bp, and the minimum read length after trimming was therefore set to 45 bp. The cleaned and trimmed reads were then de-novo assembled using Trinity (version 2.8.3; XXX) under default settings. The default settings include an *in-silico* normalization of the reads to remove highly redundant sequences before the assembly process. The reads of *Hyleoglomeris* sp. Japan, *Progolomeris vasconica* were processed and assembled according to previously established and published protocols (Misof et al. 2014, Peters et al. 2017) Finally, all transcripts under 200 bp in length were discarded from each assembly before contamination screening.

The assembled transcriptome of *Glomeris pustulata* was downloaded from the TSA (accession number: GAKW000000000.1) and those of *Haploglomeris multistriata* and *Glomeridella minima* were kindly shared with us by Szucsich et al. (in review) prior to their publication.

#### **4. Screening for contaminants**

The assembled transcripts were screened against the Univec database (version 10.0; [www.ncbi.nlm.nih.gov/tools/vecscreen/univec/](http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/)) using a local installation of the Vecscreen pipeline ([www.ncbi.nlm.nih.gov/tools/vecscreen/](http://www.ncbi.nlm.nih.gov/tools/vecscreen/)). The potential contamination identified were removed both from terminal and internal regions of transcripts, the latter resulting in split contigs. Following Peters et al. (2017), an additional cross-contamination check was conducted for the transcriptomes of *Hyleoglomeris* sp. Japan, *Progolomeris vasconica*, *Haploglomeris multistriata*, and *Glomeridella minima* against all other transcriptomes that were sequenced on the same lane. No such check was

possible for the remaining transcriptomes as they were sequenced separately or downloaded from NCBI and no sequence data was available from samples that were sequenced on the same lane. Finally, the newly sequenced transcriptomes were uploaded to the NCBI Transcriptome Shotgun Assembly Sequence Database (TSA), which provides additional filtering for possible contamination. A detailed overview of all transcriptomes, filtering steps, and final assembly statistics can be found in Table 1.

## **5. Set of orthologous genes**

In order to identify single copy genes in the assembled transcriptomes, we generated a reference set of orthologous genes. As there is currently only a single high-quality myriapod genome available (*Strigamia maritima*; Chipman et al. 2014), we used the BUSCO v2 Arthropoda ortholog set (Simao et al. 2015) as a basis. The set comprises of 1,066 clusters of orthologous genes (COGs) which are single-copy and present in most species across Arthropoda. We selected a subset of 16 species, representing the diversity of the group (including *Strigamia maritima* as a Myriapod representative) for which the official gene sets are considered to be of high-quality and that are readily available (Table S2). Following the approach outlined by Vasilikopoulos et al. (2018), we downloaded the official gene sets (OGSs) of all 16 species from their respective original databases (Table S2). All OGSs were the same versions as used to construct the original BUSCO v2 ortholog set and were downloaded as both amino acid and nucleotide sequences. The OGSs were subsequently modified according to the input requirements of Orthograph (Petersen et al. 2017) using custom PERL scripts.

## **6. Orthology assignment**

The assembled and cleaned transcriptome libraries were searched for single copy orthologous genes using Orthograph (version 0.6.3; Petersen et al. 2017) and the ortholog set described in section 5. Orthograph was ran using the following settings: max-blast-searches = 50, blast-max-hits = 50, orf-

overlap-minimum = 0.5, extend-orf = 1, minimum-transcript-length = 30, substitute-u-with = X. A detailed description of all settings can be found in Petersen et al. (2017) as well as on the online documentation of Orthograph (<https://github.com/mptrsen/Orthograph>). The results for each species were summarized using the `summarize_orthograph_results.pl`, which is included in the Orthograph package. The results were summarized using the `-t -u -s` options to remove terminal stop codons not encoded by the nucleotide sequence and substituting internal stop codons as well as selenocysteins (U) with 'X' and 'NNN' in the amino acid and nucleotide sequences, respectively. The transcriptomes of *Glomeris pustulata*, *Cyliosoma* sp., and *Narceus americanus* were discarded at this point due to the overall low number of hits (62-84% of COGs) and total number of amino acids identified (Table 3) in order to prevent them from negatively impacting the alignment process as well as the completeness of the final datamatrix.

## 7. Outlier check

We identified outliers (potentially miss-identified or erroneously aligned sequences) in each of the amino acid multiple sequence alignments using the outlier check scripts (`checker_complete_1.3.1.2` and `checker_complete_1.3.1.2`; Misof et al. (2014)). Sequences are considered as outliers if their BLOSUM62 distance to their closest reference taxon, as identified by Orthograph, is above the cut-off value. The cut-off is defined as 2.25 times the distance of the upper quartile to the mean of the BLOSUM62 distances among the reference species. All sequences identified as outliers, 35 in total, were removed from the amino acid alignments and the corresponding unaligned nucleotide sequences.

## 8. Analysis of confounding factors: FcML and data permutations

To test for the presence of confounding factors in our dataset, we tested two hypotheses (Figure S7) derived from the maximum likelihood and multi-species coalescence analyses of the amino acid supermatrix (without BMGE-filtering) using the four-cluster likelihood mapping approach (FcLM;

Strimmer and Von Haeseler 1997). Following the analyses of the original dataset, we performed three permutations of the dataset, as described by Misof et al. (2014), before reanalyzing using FcLM to assess whether the analyses were affected by confounding factors such as compositional biases, non-stationary processes, or non-random distribution of missing data. The first permutation removes any phylogenetic signal in the data, retaining only the amino acid composition and distribution of missing or ambiguous characters within each partition. The second permutation replaces the amino acids in the original matrix using amino acid frequencies according to the LG substitution matrix, thereby retaining only the distribution of missing data within each partition. The third permutation is similar to the second, but additionally randomizes the distribution of missing data in the data set. Any phylogenetic signal detected in the FcML analyses of the permuted data thus indicates that the signal is derived from the confounding factors which have not been eliminated in that specific permutation. Please see the supplementary material of Misof et al. (2014) for additional details. To mitigate any stochastic effects of the permutation scheme, each permutation of the original supermatrix was repeated 100 times, each of the permuted matrixes were analyzed separately, and the results were summarized as medians. The results for each individual analysis are given in File S2.

The explicit testing of hypothesis 1 using FcLM revealed strong support (83.3 %) for the placement of *A. gibbosa* and *G. subterranea* as the earliest branching member of Glomerida (Hypothesis 1; Figure S7). The signal is still present, although much weaker (Median = 38.89 %), after eliminating the phylogenetic signal in the datamatrix (Permutation 1) and is completely removed after eliminating any compositional bias (Permutation 2) (Figure S8; File S2). This indicates that the phylogenetic signal was stronger than the signal of the bias and that the tree reconstructions should not have been affected. Testing Hypothesis 2 revealed equal support for and against (35.7 %) a sister taxon relationship between *O. underwoodi* and *E. archimedis* (Hypothesis 2; Figure S7). Here, no signal remained in the dataset after the elimination of the phylogenetic signal (Permutation 1)

(Figure S8; File S2), indicating that there were no confounding factors affecting the analyses. We did not detect confounding factors that affected the branches in question, however the small taxon sampling might have negatively impacted the analyses. FcLM relies on drawing quartets from predefined groups and the small number of taxa means only a small number of quartets could be drawn. Therefore, potential confounding factors cannot be completely ruled out.

### Figure captions

**Figure S1:** Heat maps showing pairwise p-values for Bowker's test of symmetry, computed using SymTest. **A:** Codon pos. 1+2+3. **B:** Codon pos. 1+2+3 + BMGE-filtering. **C:** Codon pos. 1+2. **D:** Codon pos. 1+2 + BMGE-filtering. **E:** Codon pos. 1. **F:** Codon pos. 1 + BMGE filtering. **G:** Codon pos. 3. **H:** Codon pos. 3 + BMGE-filtering.

**Figure S2:** Heat maps showing pairwise completeness scores computed using AliStat. **A:** Codon pos. 1+2+3. **B:** Codon pos. 1+2+3 + BMGE-filtering. **C:** Codon pos. 1+2. **D:** Codon pos. 1+2 + BMGE-filtering. **E:** Codon pos. 1. **F:** Codon pos. 1 + BMGE-filtering. **G:** Codon pos. 3. **H:** Codon pos. 3 + BMGE-filtering.

**Figure S3:** Phylogenetic relationships of Glomerida derived from analyzing the amino acid dataset. **A:** Tree inferred from the supermatrix using metapartitions in IQTREE, branch support estimated from 500 non-parametric bootstrap replicates. **B:** Species tree inferred from individual gene trees using ASTRAL, branch support values are posterior probabilities and quartet support values (quartet-based frequencies of the given topology and the two alternative topologies).

**Figure S4:** Phylogenetic relationships of Glomerida derived from analyzing the BMGE-filtered amino acid dataset. **A:** Tree inferred from the supermatrix using metapartitions in IQTREE, branch support estimated from 500 non-parametric bootstrap replicates. **B:** Species tree summarized from

individual gene trees using ASTRAL, branch support values are posterior probabilities and quartet support values (quartet-based frequencies of the given topology and the two alternative topologies). **C:** Tree inferred from the supermatrix using PhyloBayes, branch support values are posterior probabilities.

**Figure S5:** Phylogenetic relationships of Glomerida derived from analyzing the nucleotide dataset of the 2<sup>nd</sup> codon position. **A:** Tree inferred from the supermatrix using metapartitions in IQTREE, branch support estimated from 500 non-parametric bootstrap replicates. **B:** Tree inferred from the supermatrix using IQTREE and the ghost-model, branch support estimated from 500 non-parametric bootstrap replicates. **C:** Species tree inferred from individual gene trees using ASTRAL, branch support values are posterior probabilities and quartet support values (quartet-based frequencies of the given topology and the two alternative topologies).

**Figure S6:** Phylogenetic relationships of Glomerida inferred from analyzing the BMGE-filtered nucleotide dataset of the 2<sup>nd</sup> codon position. **A:** Tree inferred from analyzing the supermatrix using metapartitions in IQTREE, branch support estimated from 500 non-parametric bootstrap replicates. **B:** Tree inferred from the supermatrix using IQTREE and the ghost-model, branch support estimated from 500 non-parametric bootstrap replicates. **C:** Species tree inferred from individual gene trees using ASTRAL, branch support values are posterior probabilities and quartet support values (quartet-based frequencies of the given topology and the two alternative topologies).

**Figure S7:** **A:** Comparison of topologies derived from the maximum likelihood and ASTRAL analyses of the amino acid dataset without BMGE-filtering. **B:** Results of the FcLM analysis on the amino acid dataset without BMGE-filtering for the identification of the earliest branching Glomerida (hypothesis 1). **C:** Results of the FcLM analysis on the amino acid dataset without BMGE-filtering for the placement of *E. archimedis* and *O. underwoodi* (hypothesis 2).

**Figure S8:** Results of the FcLM analyses of the amino acid dataset without BMGE-filtering for hypothesis 1 and 2 after permutation. Values represent medians of 100 permutations.

## REFERENCE

- Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, et al. (2014) The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede *Strigamia maritima*. PLoS Biology. 12: e1002005.
- Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet Journal. 17: 10.
- Misof B, Liu S, Meusemann K, Peters RS, Donath A, et al. (2014) Phylogenomics resolves the timing and pattern of insect evolution. Science. 346: 763–767.
- Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, et al. (2017) Evolutionary History of the Hymenoptera. Current Biology. 27: 1013–1018.
- Petersen M, Meusemann K, Donath A, Dowling D, Liu S, et al. (2017) Orthograph: A versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinformatics. 18: 1–10.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31: 3210–3212.
- Strimmer K, von Haeseler A (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. proceedings of the national academz of sciences of the U.S.A. 94: 6815–6819.
- Szucsich N, Bartel D, Blanke A, Böhm A, Donath A, et al. (in review) Four myriapod relatives – but who are sisters? No end to debates on relationships among the four major myriapod subgroups. BMC Evolutionary Biology.



Vasilikopoulos A, Balke M, Beutel RG, Donath A, Podsiadlowski L, et al. (2019) Phylogenomics of the superfamily Dytiscoidea (Coleoptera: Adephaga) with an evaluation of phylogenetic conflict and systematic error. *Molecular Phylogenetics and Evolution*. 135: 270–285.