Supplementary Material

1. New Transcriptome Data and Retrieval of Published Data

New transcriptome data (ESTs) were generated for *Megajapyx* sp. (Diplura: Japygidae) and *Thermobia domestica* (Packard 1837) (Zygentoma: Lepismatidae). For extraction of total and messenger RNA, we used 17 adult specimens of *Megajapyx* sp. collected from two localities in East Crete, (locality 1: 35,05619 N - 26,035108 E; locality 2: 35,04927 N - 26,020925 E, *legit* Emiliano Dell'Ampio, Günther Pass in April 2008), and 20 adult specimens of *T. domestica* obtained from a commercial labstock (Fressnapf, Bonn, Germany).

Following steps were carried out at the Max Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany. Total RNA was extracted with the Absolutely RNA Microprep Kit (Stratagene/Agilent, Germany). Complementary DNA (cDNA) was synthesized using the SMART approach (Mint-Universal cDNA synthesis kit, Evrogen, Russia), and subsequently normalized using duplex-specific nuclease (Trimmer kit, Evrogen, Russia) according to manufacturer's instructions, and directionally ligated to self-made 454 adaptors with MIDs following Roche's technical bulletin TCB 09004 introducing SfiI-sites. Obtained 454 libraries were immobilized on beads and clonally amplified using the GS FLX Titanium LV emPCR Kit. Libraries were sequenced using the GS FLX Titanium Sequencing Kit XLR70 and GS FLX Titanium PicoTiterPlate Kit. All kits used were purchased from Roche and followed the manufacturer's protocol.

Sequencing results of raw reads:

	Megajapyx sp.	*Thermobia domestica*
Total number of reads	84,841	82,154
Average read length (bp)	307	330
Modal read length (bp)	214	370
Total number of bases	260,964,700	271,878,820
Average GC content	50.0%	36.6%

Raw sequence reads were processed at the Center for Integrative Bioinformatics (CIBIV), Vienna (Austria) as described in von Reumont et al. (2012): Lowercase nucleotides were clipped with custom made perl modules. Vector sequences and poly-A-tails were removed using UNIVEC (build 5.2, December 2009) with Crossmatch (Green 1993-1996, version 0.990329). Masking of sequence repeats was performed with RepeatMasker (Smit et al. 1996-2010, open-3.1.6), using the database RepBase (version 20061006). Subsequently, de novo assembly of the sequence reads into contigs was performed with MIRA (Chevreux et al. 2004), version 3.0.3, with following settings: −job=denovo, est, accurate, 454 − highlyrepetitive; −CO: asir =yes; 454_SETTINGS−CO: fnicpst=yes; −FN: fai=.fasta : fqui=. Qual; s−LR: mxti=no; −CL: qc=yes : cpat=no −OUT: ssip=p=yes.

Published transcriptome and genome data were obtained from various public sources (see supplementary table 1). Transcriptome assemblies were obtained from the Deep Metazoan Phylogeny project database (www.deep-phylogeny.org), see also von Reumont et al. (2012) and Simon et al. (2012).

2. Orthology Assignment

For orthology assignment of assembled sequences we used HaMStR (Ebersberger et al. 2009), version 4 (http://www.deep-phylogeny.org/hamstr/download/archive/hamstrsearch_local-hmmer3.v4.tar.gz). We used *insecta-hmmer3-1* as ortholog reference set, including 1,886 orthologous genes (OGs) (available from www.deep-phylogeny.org/hamstr/download/datasets/hmmer3/insecta_hmmer3-1.tar.gz). In this set, five species are used to build the reference ortholog set (these species are called core or reference taxa): *Capitella* sp. (capsp 2853), *Ixodes scapularis* (ixosc 1169), *Daphnia pulex* (dappu 1391), *Apis mellifera* (apime 353), and *Aedes aegypti* (aedae 350). The assignment of orthologous genes (OGs) among these taxa was compiled using the InParanoid-TC approach (O'Brien et al. 2005; Ebersberger et al. 2009; Östlund et al. 2010). Multiple sequence alignments of the reference taxa for each OG served as input to build Hidden Markow Model profiles (pHMM) with the HMMER 3.0 software package (Eddy 2011). The algorithm, as implemented in HaMStR, uses pHMMs to search the transcriptome data (chosen *e*-value: 1e-05) for putative orthologous hits by mapping transcript sequences to the OGs of the reference ortholog set. For each hit received from the pHMM search, orthology was assured using reciprocal BLAST against the official gene sets of the refernce taxa. Bidirectional best hits (BBH) were accepted as 1:1 orthologs if the reciprocal BLAST delivered hits against official gene sets (amino acid level) of four selected reference taxa: *Daphnia pulex*, *Ixodes scapularis*, *Apis mellifera*, and *Capitella* sp. (option *-strict*). We chose the option *-representative* for concatenation of single, non-overlapping transcript sequences that had been assigned to the same OG. For orthology assignment of selected species which were not included in the reference set, and for which an official gene set on protein level was available, orthology was assigned using the option *-protein* (see supplementary table 2).

3. Rogue Taxa Analysis

To identify rogue taxa, we used the single taxon algorithm (Aberer and Stamatakis 2011). It iteratively assesses how the support in a majority rule consensus tree changes by alternately pruning each single taxon. Reasons for rogues can be various, *e.g.*, rogues may show long branches or may have low data coverage within a dataset (like, *e.g.*, *Orchesella cincta*, or *Coptotermes formosanus*, supplementary table 1).

The impact of a rogue taxon is described as the accumulated bootstrap support (AS, listed in column ImprovCons, supplementary table 7a) that is added to the consensus tree, if the rogue taxon is pruned from the bootstrap trees. For

example, a value of 100% AS indicates that the support in a consensus tree after pruning the respective taxon increased by an equivalent of one bipartition with 100% bootstrap support (*e.g.*, two bipartitions with 50% bootstrap support each will yield 100% AS improvement). As a cut-off, taxa are excluded that, if pruned, add more than the equivalent of a bipartition with 30% bootstrap support to the consensus tree.

A total of five rogues were identified in dataset *M_Ento* and in the three data subsets *M_Nono*, *M_Elli*, and *M_DiCo* (supplementary table 7a): *Coptotermes formosanus* (all four datasets), *Orchesella cincta* (set *M_Elli*), *Baetis* sp. (sets *M_Elli* and *M_DiCo*), *Pollicipes pollicipes* and *Pediculus humanus* (set *M_DiCo*).

In all four datasets, *C. formosanus* acts as a particularly strong rogue taxon (between 150% and 213% AS, if pruned). Only *O. cincta* in set *M_Elli* had a similarly detrimental effect on the support in the consensus tree (~211% AS). Moreover, the algorithm detected moderate rogue taxon effects for *Baetis* sp. in set *M_Elli* (72.1% AS) and *M_DiCo* (43.7% AS), and for *P. pollicipes* (63.4% AS) and *P. humanus* (61.6% AS) in set *M_DiCo*. We used a cut-off of 30% AS because elimination of very weak rogues did not result in an increase of the bootstrap support of bipartitions in the ML trees. In contrast, pruning the above mentioned rogue taxa, with exception of *P. humanus*, increased the support for bipartitions in the ML trees.

Pruning rogues resulted in increasing bootstrap support for various relationships among enthognathous lineages: After pruning *O. cincta* from the dataset *M_Elli* the support in the inferred ML tree became maximal for Ellipura and increased for Diplura + Ectognatha. In the datasets *M_Ento* and *M_Nono* the elimination of rogues resulted in an increased support for Nonoculata + Ectognatha (supplementary table 7b).

4. Annotation of Cellular Functional Catogeries for OGs included in dataset *M_Ento*

As proposed by Simon et al. (2009) we annotated all 117 OGs included in the decisive dataset *M_Ento* with functional KOG gene categories using *KOGnitor* as available at the COG database (Tatusov et al. 2003, see supplementary table 5). As query for the annotation of the OGs we used unmasked amino acid sequences of *Aedes aegypti*, which for this purpose was included as a reference species in our ortholog reference set. We assessed whether differences in inferred relationships among datasets *M_Nono*, *M_Elli*, and *M_DiCo* (supplementary figure 4, supplementary table 4) correlate with gene function by comparing the percentage of genes pertaining to functional categories among these datasets. We found no obvious correlation (supplementary figure 5).

5. Partitioned ML Analyses

We repeated ML tree reconstructions based on partitioned analyses for the datasets *M_Ento*, *M_Nono*, *M_Elli* and *M_DiCo* (rogue taxa pruned). This was done to detect possible differences in topology or support between results from

unpartitioned versus partitioned analyses that were possibly caused by model misspecifications in the unpartitioned analyses.

We used the Bayesian information criterion (BIC, Schwarz 1978) to select the best-fit model for each OG included in the respective datasets. Selected models were then used to apply a partitioned ML analysis for dataset *M_Ento, M_Nono, M_Elli,* and *M_DiCo*. This approach also allows unlinked branch length estimates per gene (the so-called full partition model approach) which is the most complex model accounting for among-gene differences in substitution processes and rate heterogeneity. The selected models for each OG are listed in supplementary table 6. Clade support was assessed using the ultrafast bootstrap algorithm (UFBoot; Minh et al. 2013) with 5,000 bootstrap replicates. The bootstrap resampling was done on a within-gene basis (*i.e.*, sites were resampled within each OG separately). Tree reconstructions and UFBoot analyses were carried out with the software IQ-TREE (Minh et al. 2013), v. 0.9.5, available from http://www.cibiv.at/software/iqtree/. IQ-TREE iteratively applies important quartet puzzling (Vinh and von Haeseler 2004) and nearest neighbor interchange heuristics to sample the tree space.

The inferred topologies from partitioned analyses did not differ from the results of the unpartitioned analyses with respect to the phylogenetic relationships under study, *i.e.*, relationships among entognathous primarily wingless insects. However, there were minor changes in bootstrap support particularly addressing entognathous hexapods comparing unpartitioned and partitioned analyses (see table 2 of the main text). For changes in other clades, please compare figure 1, supplementary figure 2, and supplementary figure 6, or supplementary figure 4 and supplementary figures 7-9.

6. Split analyses

Phylogenetic networks (Huson and Bryant 2006) were calculated from masked alignments for the dataset *M_Ento*, and each of its data subsets *M_Nono*, *M_Elli* and *M_DiCo* (supplementary figures 10-13) using SplitsTree 4.13. We calculated NeighborNetworks based on the neighbornet algorithm (Bryant and Moulton 2004) with uncorrected p-distances. The NeighborNetworks gives indications of signal-like patterns and conflicts in alignments. In each dataset, the topology that is predominantly supported in the ML tree reconstructions and the FcLM is reflected also in the network, as is conflicting signal. Comparisons of other parts of the tree, that are not related to the “Entognatha-problem”, should be taken with caution, since our data matrix and its submatrices are optimized for a specific question, the “Entognatha-problem”. Thus, differences in other parts of the tree may just reflect differences in coverage of genes.

Additional references

Bryant D, Moulton V. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 21:255–265.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14:1147–1159.

Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comp. Biol. 7: e1002195.

Green P. 1993-1996. Crossmatch. [cited 2011 Nov 14]. Available from: http://www.incogen.com/public_documents/vibe/details/crossmatch.html.

KOGnitor. Available from: http://www.ncbi.nlm.nih.gov/COG/grace/kognitor.html.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

O'Brien KP, Remm M, Sonnhammer ELL. 2005. Inparanoid: A comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 33:D476–D480.

Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38:D196–203.

Schwarz G. 1978. Estimating dimension of a model. Ann Stat. 6:461–464.

Simon S, Strauss S, von Haeseler A, Hadrys H. 2009. A phylogenomic approach to resolve the basal pterygote divergence. Mol Biol Evol. 26:2719–2730.

Smit AFA, Hubley R, Green P. 1996-2010. RepeatMasker Open 3.0. Available from: http://www.repeatmasker.org.

Tatusov RL, Fedorova ND, Jackson JD, et al. (17 co-authors). 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4:41. (*Please note that 17 includes all authors of the article.*)

Vinh LS, von Haeseler A (2004) IQPNNI: moving fast through tree space and stopping in time. Mol Biol Evol. 21:1565-1571.**Supplementary Material**

Tables

Supplementary table 1. Taxon sampling with number of orthologous genes (OGs) per species in the original supermatrix and in the three data subsets.

Supplementary table 2. Annotation of orthologous genes (OGs) used in this study.

Supplementary table 3. Treelikeness of orthologous genes (OGs) included in the dataset SOS.

Supplementary table 4. Results of the *FcLM* for each OG included in the decisive dataset *M_Ento.*

Supplementary table 5. Functional KOG categories of orthologous genes (OGs) included in dataset *M_Ento*.

Supplementary table 6. List of the best substitution model per orthologous gene (OG).

Supplementary table 7. Results from rogue taxa analyses.

Figures

Supplementary Fig. 1. Unreduced datamatrix displaying presence/absence of genes along with treelikeness value for each gene.

Unreduced datamatrix (73 taxa,1,866 OGs) analyzed with *mare* 0.12-rc. X-axis: Orthologous genes (OG ID). Y-axis: Taxa (ixosc_1169: *Ixodes scapularis*; dappu_1391: *Daphnia pulex;*capsp_2853: *Capitella* sp.; apime_353: *Apis mellifera*; aedae_350: *Aedes aegypti*; DROME_4: *Drosophila melanogaster*; PEDHU: *Pediculus humanus*; TRICA 1449: *Tribolium castaneum*; ACYPI_254: *Acyrthosiphon pisum*; BOMMO_1242: *Bombyx mori*; NASVI_363: *Nasonia vitripennis; Acerentomon franzi* was afterwards corrected and assigned as *Acerentomon* sp. Legend: Treelikeness values represent the information content for each gene analyzed with the extended geometry quartet mapping approach (for more details refer to the *mare* manual). Overall information content of the matrix: 0.103, matrix coverage in terms of presence of OGs: 36%.

Supplementary Fig. 2. ML phylograms and datamatrix from dataset *M_Ento* (117 OGs).

*: taxon present in the ortholog reference set. Best Maximum likelihood trees (RAxML v.7.2.8, PROTCAT, LG + GAMMA), bootstrap support is derived from 1,000 bootstrap replicates. Trees were rooted with *Capitella* sp.; nodes below 50% bootstrap support are collapsed.

(a) Tree after pruning rogues (72 taxa, 32,883 alignment positions, overall information content of the matrix 0.281, matrix coverage (presence/absence of OGs): 65.9%, pruned rogue: *Coptotermes formosanus*.

(b) Tree including rogues (73 taxa, 32,883 amino acid alignment positions, overall information content of the matrix 0.284, matrix coverage in terms of presence of OGs: 66.6%; **: the real branch length of *Coptotermes formosanus* is three times longer than displayed.

(c) Datamatrix of *M_Ento* including rogues, analyzed with *mare* 0.12-rc. X-axis, Y-axis, and legend see supplementary fig. 1.

Supplementary Fig. 3: ML phylograms and datamatrices from datasets SOS (a) and SOS_{ω} (b).

(a) Top: Best Maximum likelihood tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA) based on 253 orthologous genes (OGs), 79 of which are covered by Protura, Diplura and Collembola (SOS). Bootstrap support is derived from 1,000

bootstrap replicates. *: taxon present in the ortholog reference set. The tree was rooted with *Capitella* sp.
Bottom: Corresponding datamatrix (SOS) generated with *mare* 0.12-rc (62 taxa, 253 OGs, 55,429 amino acid alignment positions, overall information content of the matrix 0.337, matrix coverage in terms of presence of OGs: 66.5%). For extraction of this SOS we choose a taxon weighting (*-t* 1.5). X-axis, Y-axis, and legend see supplementary fig. 1.
(b) Top: Best Maximum likelihood tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA) based on 174 OGs, none of which is covered by Protura, Diplura and Collembola (SOS_{ω}). Bootstraop support is derived from 1,000 bootstrap replicates. *: taxon was present in the ortholog reference set. The tree was rooted with *Capitella* sp.
Bottom: Corresponding datamatrix (SOS_{ω}) analyzed with *mare* 0.12-rc (62 taxa, 174 OGs, 37,152 amino acid alignment positions, overall information content of the matrix 0.313, matrix coverage in terms of presence of OGs: 60.4%. X-axis, Y-axis, and legend see supplementary fig. 1.

Supplementary Fig. 4. ML phylograms and datamatrices from datasets *M_Nono* (a), *M_Elli* (b) and *M_DiCo (c)*
*: taxon was present in the ortholog reference set. Trees were rooted with *Capitella* sp. Bootstraop support is derived from 1,000 bootstrap replicates.
(a) Top left: Best Maximum likelihood tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA) of dataset *M_Nono* (51 OGs, 72 taxa, 12,548 amino acid alignment positions, after pruning the rogue taxon *Coptotermes formosanus*).
(a) Top right: Corresponding datamatrix (*M_Nono*) generated with *mare* 0.12-rc. Overall information content of the matrix 0.279, matrix coverage in terms of presence of OGs 65.8%. X-axis, Y-axis, and legend see supplementary fig. 1.
(a) Bottom left: Best Maximum likelihood tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA) of dataset *M_Nono* (51 OGs, 73 taxa, 12,548 amino acid alignment positions, including the rogue taxon *Coptotermes formosanus*). **: the real branch length of *Coptotermes formosanus* is three times longer than displayed.
(a) Bottom right: Corresponding datamatrix (*M_Nono*) generated with *mare* 0.12-rc. Overall information content of the matrix 0.275, matrix coverage in terms of presence of OGs 65%. X-axis, Y-axis, and legend see supplementary figure 1.
(b) Top left: Best Maximum likelihood tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA) of dataset *M_Elli* (35 OGs, 70 taxa, 11,789 amino acid alignment positions, after pruning rogues *Coptotermes formosanus, Orchesella cincta, Baetis* sp.).
(b) Top right: Corresponding datamatrix (*M_Elli*) generated with *mare* 0.12-rc. Overall information content of the matrix 0.286, matrix coverage in terms of presence of OGs 68.1%. X-axis, Y-axis, and legend see supplementary fig. 1.
(b) Bottom left: Best Maximum likelihood tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA) of dataset *M_Elli* (35 OGs, 73 taxa, 11,789 amino acid alignment positions, including rogue taxa).
(b) Bottom right: Corresponding datamatrix (*M_Elli*) generated with *mare* 0.12-rc. Overall information content of the matrix 0.275, matrix coverage in terms of presence of OGs 66%. X-axis, Y-axis, and legend see supplementary fig. 1.

(c) Top left: Best Maximum likelihood tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA) of dataset *M_DiCo* (31 OGs, 69 taxa, 8,546 amino acid alignment positions, after pruning rogues *Coptotermes formosanus, Pollicipes pollicipes, Pediculus humanus, Baetis* sp.).
(c) Top right: Corresponding datamatrix (*M_DiCo*) generated with *mare* 0.12-rc. Overall information content of the matrix 0.306, matrix coverage in terms of presence of OGs 68.2%. X-axis, Y-axis, and legend see supplementary fig. 1.
(c) Bottom left: Best Maximum likelihood tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA) of dataset *M_DiCo* (31 OGs, 73 taxa, 8,546 amino acid alignment positions, including rogue taxa).
(c) Bottom right: Corresponding datamatrix (*M_DiCo*) generated with *mare* 0.12-rc. Overall information content of the matrix 0.296, matrix coverage in terms of presence of OGs 67.1%. X-axis, Y-axis, and legend see supplementary fig. 1.

Supplementary Fig. 5. Orthologous genes (OGs) in data subsets *M_Nono, M_Elli and M_Dico* sorted by functional KOG categories.
Each chart shows the absolute number of OGs in different functional KOG categories (A, C, E, F. G, H, I, J, K, O, P, R, S, T, U, Z) according to the COG database (Tatusov et al. 2003). Top: data subset *M_Nono*; middle: data subset *M_Elli*; bottom: data subset *M_DiCo,* refer also to supplementary table 5.

Supplementary Fig. 6. ML phylogram (partitioned analyses with IQ-TREE) from dataset *M_Ento*.
Best Maximum likelihood tree inferred with IQ-TREE v. 0.9.5 from dataset *M_Ento* (117 OGs, 72 taxa, 32,883 amino acid alignment positions, after pruning the rogue taxon *Coptotermes formosanus*). *: taxon was present in the ortholog reference set. The tree was rooted with *Capitella* sp. and *Ixodes scapularis*. Bootstrap support is derived from 5,000 bootstrap replicates using the ultrafast bootstrap algorithm. For selected models per OG see supplementary table 6.

Supplementary Fig. 7. ML phylogram (partitioned analyses with IQ-TREE) from dataset *M_Nono*.
Best Maximum likelihood tree inferred with IQ-TREE v. 0.9.5 from dataset *M_Nono* (51 OGs, 72 taxa, 12,548 amino acid alignment positions, after pruning the rogue taxon *Coptotermes formosanus*). Legend see supplementary figure 6.

Supplementary Fig. 8. ML phylogram (partitioned analyses with IQ-TREE) from dataset *M_Elli*.
Best Maximum likelihood tree inferred with IQ-TREE v. 0.9.5 from dataset *M_Elli* (35 OGs, 70 taxa, 11,789 amino acid alignment positions, after pruning rogues *Coptotermes formosanus, Orchesella cincta, Baetis* sp.). Legend see supplementary figure 6.

Supplementary Fig. 9. ML phylogram (partitioned analyses with IQ-TREE) from dataset *M_DiCo*.

Best Maximum likelihood tree inferred with IQ-TREE v. 0.9.5 from dataset *M_DiCo* (31 OGs, 69 taxa, 8,546 amino acid alignment positions, after pruning rogues *Coptotermes formosanus, Pollicipes pollicipes, Pediculus humanus, Baetis* sp.). Legend see supplementary figure 6.

Supplementary Figs 10-13. NeighborNetworks.

Neighbor SplitNetworks from dataset *M_Ento* (supplementary figure 10), *M_Nono* (supplementary figure 11), *M_Elli* (supplementary figure 12), and *M_DiCo* (supplementary figure 13) (rogue taxa pruned) were calculated with SplitsTree 4.13 using uncorrected p-distances.

Supplementary Fig. 14. *FcLM* results for dataset *M_Ento* with all quartets.

(a) Histogram of *FcLM* results.

Each bar refers to an OG (for OG-IDs, see supplementary table 2). Y-axis: amount of quartets (in %), that predominantly support either T_1 [Protura+Diplura] – [Collembola+remaining taxa] (blue), T_2 [Protura+Collembola] – [Diplura+remaining taxa] (red), T_3 [Diplura+Collembola] – [Protura+remaining taxa] (yellow), and the amount of quartets that show ambiguous support (grey) (see fig. 5).

(b) Proportion of genes predominantly supporting T_1, T_2 or T_3 or showing ambiguous support.

The chart shows the proportion of quartets (summed up for the OGs included) that show predominant support for T_1, T_2, and T_3 (see above). The proportion of quartets that show ambiguous support (see fig. 5) is colored in grey.