A comprehensive analysis of bilaterian mitochondrial genomes and phylogeny

Matthias Bernt[a]*, Christoph Bleidorn[b], Anke Braband[c], Johannes Dambach[d], Alexander Donath[d*], Guido Fritzsch[e], Anja Golombek[d], Heike Hadrys[f], Frank Jühling[e,g], Karen Meusemann[d], Martin Middendorf[a], Bernhard Misof[d], Marleen Perseke[h], Lars Podsiadlowski[i]*, Björn von Reumont[j], Bernd Schierwater[f], Martin Schlegel[b], Michael Schrödl[k], Sabrina Simon[l], Peter F. Stadler[e,m,n,o,p,q], Isabella Stöger[k], Torsten H. Struck[d]

(authors in alphabetical order)

Author affiliations

[a]*Parallel Computing and Complex Systems Group, Department of Computer Science, University of Leipzig, Augustusplatz 10, D-04109 Leipzig, Germany*

[b]*Molecular Evolution and Systematics of Animals, Institute of Biology, University of Leipzig, Talstraße 33, D-04103 Leipzig, Germany*

[c]*LGC Genomics GmbH, Ostendstr. 25, 12459 Berlin*

[d]*Centre for Molecular Biodiversity Research (ZMB), Zoologisches Forschungsmuseum Alexander Koenig (ZFMK), Adenauerallee 160, D-53113 Bonn, Germany*

[e]*Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*

[f]*ITZ, Ecology & Evolution, TiHo Hannover, Buenteweg 17d, 30559 Hannover, Germany*

[g]*Architecture et Réactivité de l'ARN, Université de Strasbourg, CNRS, IBMC; Strasbourg, France*

[h]*Laboratory of Marine Biology, South China Sea Institute of Oceanology, Chinese Academy of Science, 164 West Xingang Road, 510301 Guangzhou, PR China*

[i]*Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, D-53113 Bonn, Germany*

[j]*Department of Life Sciences, The Natural History Museum, Cromwell Road SW7 5BD London, United Kingdom*

[k]*Zoologische Staatssammlung München, Münchhausenstraße 21, 81247 München, Germany*

[l]*Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA*

[m]*Max-Planck-Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany*

[n]*Fraunhofer Institut fuer Zelltherapie und Immunologie, Perlickstrasse 1, D-04103 Leipzig,*

Germany

[o]*Department of Theoretical Chemistry, University of Vienna, Waehringerstrasse 17, A-1090 Wien, Austria*

[p]*Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegardsvej 3, DK-1870 Frederiksberg, Denmark*

[q]*Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

*these authors contributed equally to this work

corresponding authors:

Lars Podsiadlowski, Inst. Evolutionary Biology & Ecology, University of Bonn, An der Immenburg 1, D-53121 Bonn, Germany; fax +49228735129; email: lars@cgae.de

Matthias Bernt, Parallel Computing and Complex Systems Group, Department of Computer Science, University of Leipzig, Augustusplatz 10, D-04109 Leipzig, Germany; email: bernt@informatik.uni-leipzig.de

Abstract

About 2800 mitochondrial genomes of Metazoa are present in NCBI RefSeq today, two thirds belonging to vertebrates. Metazoan phylogeny was recently challenged by large scale EST approaches (phylogenomics), stabilizing classical nodes while simultaneously supporting new sister group hypotheses. The use of mitochondrial data in deep phylogeny analyses was often criticized because of high substitution rates on nucleotides, large differences in amino acid substitution rate between taxa, and biases in nucleotide frequencies. Nevertheless, mitochondrial genome data might still be promising as it allows for a larger taxon sampling, while presenting a smaller amount of sequence information. We present the most comprehensive analysis of bilaterian relationships based on mitochondrial genome data. The analyzed data set comprises more than 650 mitochondrial genomes that have been chosen to represent a profound sample of the phylogenetic as well as sequence diversity. The results are based on high quality amino acid alignments obtained from a complete reannotation of the mitogenomic sequences from NCBI RefSeq database. However, the results failed to give support for many otherwise undisputed high-ranking taxa, like Mollusca, Hexapoda, Arthropoda, and suffer from extreme long branches of Nematoda, Platyhelminthes, and some other taxa. In order to identify the sources of misleading phylogenetic signals, we discuss several problems associated with mitochondrial genome data sets, e.g. the nucleotide and amino acid landscapes and a strong correlation of gene rearrangements with long branches.

Key words: Mitochondrial genomes, animal phylogeny

1. Introduction

The suitability of molecular markers for phylogenetic analysis can be evaluated according to a set of criteria (Cruickshank, 2002). (1) The orthology criterion should be fulfilled, meaning that the changes between gene sequences are results of underlying speciation events and not of gene duplication events (as is the case when comparing paralogous genes). Orthology prediction is a non-trivial task and became an important part of phylogenomic approaches (Altenhoff and Dessimoz, 2012). (2) Marker genes should be present in all taxa under study.

Thus, “housekeeping genes”, responsible for basal cell functions and thus common to a wide array of organisms, were widely used in phylogenetics. Nevertheless, current phylogenomic studies often work with rather gappy data matrices, e.g. sets of genes derived from EST approaches which have a varying degree of incompleteness with respect to the whole matrix (Dunn et al., 2008; Pick et al., 2010). (3) Selection should only act as a stabilizing factor on marker genes. Otherwise phylogenetic signal may be obscured by positive or negative selection, e.g. by homoplasious changes in different taxa with similar selection pressure and by a strong difference of substitution rates depending on the strength of the selective forces. Again “housekeeping genes” seem to be a good choice, having the same functional role in basal cellular mechanisms of many organisms and being optimized for their functions long before the basal splits of the group under study (Butte et al., 2001). A recent study demonstrates that slowly evolving genes involved in the translation process provide best results in resolving basal metazoan relationships (Nosenko et al., 2013). (4) Ideal genetic markers exhibit constant character state frequencies (nucleotides or amino acids) and substitution rates in all studied lineages over time. However, these features are rarely met by real data sets. (5) Finally, a good mixture of conserved and variable parts must be present in the alignment. While conserved segments allow the construction of PCR primer sets suitable for many species and are important to obtain reliable sequence alignments, variable sites or segments provide a sufficient amount of phylogenetic signal.

At first glance animal mitochondrial genomes seem to fulfill most of these criteria. Gene duplications in mitochondrial genomes occur rarely. Therefore orthology prediction is apparently an easy task, especially for complete mitochondrial genomes. But the frequent detection of non-functional nuclear copies of mitochondrial sequences (numts) weakened this view. Identifying numts remains problematic, especially when complete nuclear genomes lack for comparison (Bensasson et al., 2001). Mitochondrial genomes are present in all Metazoa (with the single known exception of the Loricifera (Danovaro et al., 2010)) and contain an almost perfectly conserved complement of “housekeeping genes”. Their comparatively high mutation rate and mixture of conserved and variable sites facilitate the use of universal primer sets and provide sufficient phylogenetic signal (Moritz et al., 1987). In addition the lack of recombination and the strictly maternal mode of inheritance (for exceptions see Bernt et al., 2013a) make mitochondrial markers as well suitable to infer population structure (Avise, 2000).

Currently (October 1st, 2012) 2765 mitochondrial genomes of Metazoa were present in NCBI RefSeq database, covering 1829 (66%) vertebrate species. About one half (479) of the remaining 936 entries are from arthropod species. However, complete mitochondrial genomes are available for most animal phyla. In comparison to phylogenomic datasets mitochondrial genome data still allow a larger taxon sampling for most of the animal phyla. But they include a much smaller amount of sequence information. Moreover, working with complete mitochondrial genomes enables the additional analysis of features like gene content and gene order.

Thus, animal mitochondrial genome data have been widely used addressing phylogenetic questions ranging from population to phylum level (Avise, 2000). With an increasing number of studies the limits and problems of mitochondrial data became more evident and its value for phylogenetic analyses was criticized for specific points or even in general (Ballard and Whitlock, 2004; Ballard and Rand, 2005; Hurst and Jiggins, 2005; Galtier et al., 2009). Notable points are large divergence of substitution rates and base composition between taxa, the already mentioned presence of “numts”, change of inheritance pattern due to the presence of cytoplasmic bacteria, and frequent occurrence of mitochondrial introgression.

Nevertheless, mitochondrial genome data often proved its value in phylogenetic studies (Rubinoff and Holland, 2005).

State-of-the-art in animal phylogenetics is the analysis of large multilocus datasets, derived from whole genomes or large scale EST approaches (“phylogenomics”) (Hausdorf et al., 2007; Dunn et al., 2008; Philippe et al., 2009; Hejnol et al., 2009; Pick et al., 2010). These analyses largely confirmed the “new animal phylogeny” (Halanych, 1995; Aguinaldo et al., 1997; Adoutte et al., 2000; Halanych, 2004), with Bilateria subdivided into the major subtaxa Lophotrochozoa, Ecdysozoa, and Deuterostomia. However, a number of small phyla failed to be placed within this framework. The internal phylogeny of Lophotrochozoa and Ecdysozoa and the basal relationships between non-bilaterian taxa and Bilateria are far from being consistent between different published studies (e.g. Srivastava et al., 2008; Hejnol et al., 2009; Philippe et al., 2009; Pick et al., 2010; Philippe et al., 2011; Nosenko et al., 2013). To complement these approaches with a comparatively small set of genes, but larger taxon sampling, we exploit a comprehensive mitogenomic dataset for an analysis of metazoan phylogeny.

Former phylogenetic analyses of metazoan mitochondrial genomes with rather small taxon samplings frequently resulted in trees with problematic long branches (e.g. Nematoda, Platyhelminthes) and supported some barely reliable sister group relations (e.g. Hassanin et al., 2005; Steinauer et al., 2005; Yokobori et al., 2008; Jang and Hwang, 2009; Mwinyi et al., 2010). However, a broad comprehensive analysis was missing, which will clearly illustrate the prospects and limits of mitochondrial genome data in metazoan phylogenetics.

Here we present the most comprehensive analysis of bilaterian phylogeny based on mitochondrial genome data, involving most invertebrate species with a RefSeq entry for a complete mitochondrial genome and a selection of vertebrate species. Together with outgroup taxa from fungi and protists we analyzed a dataset comprising more than 650 mitochondrial genomes. A new optimized automated annotation pipeline was set up to overcome annotation errors known to be widespread in NCBI RefSeq entries of mitochondrial genomes (Bernt et al., 2013b). Alignments of protein-coding genes were subject to carefully modeled ML analyses. Inconsistencies between phylogenetic analyses of nuclear genes and our results, as well as an overview concerning mitochondrial gene orders, will be discussed in more detail in the taxon-specific reviews (other articles in this special issue). Here we focus on the general landscape of mitochondrial genome variation in Metazoa and the problems resulting from departures of the above mentioned criteria of ideal phylogenetic markers.

2. Materials and methods

2.1. Data

The analyses are based on all metazoan mitogenome sequences in RefSeq (Pruitt et al., 2007) release 41, excluding the sequence of *Anopheles funestus* (NC_008070), which consists of 27.5% non-standard bases. In addition, the mitochondrial genome sequences of four metazoan species which have been added to RefSeq recently plus a few new and so far unpublished mitochondrial sequences of metazoan species (see supplementary material) were added. We used the mitochondrial genome sequences of 20 fungi species from RefSeq release 41 and eight contributed other non-metazoan eukaryote species as outgroup representatives (see supplementary material).

The phylogenetic reconstruction is solely based on protein coding genes. In order to avoid potential inconsistencies or errors in the published annotations (e.g. Boore, 2006) we re-annotated all sequences using the protein prediction pipeline of MITOS (Bernt et al., 2013b). For each protein coding gene the MITOS prediction with the best quality value was used to extract the corresponding amino acid sequence. For the two species with a mitogenome consisting of two sequences, i.e. *Hydra magnipapillata* (NC 011221, NC 011220) and *Brachionus plicatilis* (NC 010472, NC 010484), the best prediction from the two sequences was taken. This affects only the three genes *cox1*, *nad4*, and *nad6* where a prediction was made by MITOS for both mitochondrial genome sequences. In each case the values for the quality scores of the best predictions for the two sequences differ by more than a magnitude. For each protein coding gene an alignment of the determined amino acid sequences has been created (see Section 2.2). The concatenated alignments for the different protein coding genes for a group of species have then been used for phylogenetic reconstruction (see Section 2.3). In addition to the complete dataset (denoted as METAZOA) subsets, partly complemented with additional data, were used in analyses presented in other articles of this special issue: ARTHROPODA - without neopteran insects - (Podsiadlowski et al., 2013), DEUTEROSTOMIA (PERSEKE ET AL., 2013); DIPLOBLASTS (Osigus et al., 2013); HEXAPODA (Simon and Hadrys, 2013), and MOLLUSCA (Stoeger and SchroEDL, 2013).

2.2. Creation and processing of alignments

Amino acid sequences were aligned separately for each protein coding gene with MAFFT version 6.716 (Katoh et al., 2002) using the default parameter values. The frayed ends of the aligned sequences were trimmed by employing a simple rule: Starting separately from both ends of an alignment, columns are removed until a column with less than 20% gaps is found or the total number of removed columns reaches 100. Homoplastic or random-like characters are removed by masking the trimmed alignments with the software noisy, rel. 1.5.9 (Dress et al., 2008), using a cutoff value of 0.8. The single protein alignments were concatenated in lexicographic order with respect to their names. In the few instances where an organism lacks a protein-coding gene the concatenated alignment is filled with gaps at the corresponding positions.

The phylogenetic analysis (Section 2.3) of the complete METAZOA data set is computationally extremely demanding. Therefore only a subset of species has been considered. The selection of such a subset has to regard the biases due to an over-representation of certain taxonomic groups. The reduction of the data set is carried out in such a way that the phylogenetic and sequence diversity within the data set is maintained. This is done with an automated approach as described in the following. A neighbor-joining tree of the concatenated alignments for the protein coding sequences has been calculated with QuickTree (Howe et al., 2002). Groups of very closely related sequences are identified as connected smallest subtrees with the property that the longest patristic distance between two leafs in the subtree is smaller than a cutoff value given as parameter. From such a group of sequences only two species having the sequence with the shortest and longest distances to the root node of the respective subtree are included in the data set. In order to prevent the exclusion of sequences belonging to species of high phylogenetic interest, all species from an expert curated list of 156 species (see supplement) are guaranteed to be included in METAZOA. The cutoff value is chosen to produce a data set of appropriate size (i.e. 684 species) such that a phylogenetic analysis is feasible in reasonable time.

In order to assess taxon sampling issues two smaller data sets have been analyzed. A data set containing 325 species (denoted as METAZOA-300) has been created by using a more restrictive threshold. Furthermore, a manually curated data set containing 114 species (denoted as METAZOA-100) has been analyzed (a detailed list of all taxa is provided in the supplement).

2.3. Phylogenetic reconstruction process

The Maximum Likelihood analysis was performed with RAxML version 7.2.8 (Stamatakis, 2006) by employing a protein mixed model, i.e. CAT+MTZOA+F (CAT+MTART+F for ARTHROPODA, and HEXAPODA, respectively) with GAMMA correction of the final tree. At least three batches of 100 rapid bootstrap trees were generated until all four convergence criteria provided by RAxML were met (Stamatakis et al., 2008; Pattengale et al., 2009). Additional batches of 100 rapid bootstraps were necessary for the data sets DIPLOBLASTS (400 in total), and HEXAPODA (400 in total). A best tree search for the best scoring ML tree was conducted. Except for the two large species sets METAZOA and DEUTEROSTOMIA 200 distinct starting trees were used. The run time requirements for the two larger datasets necessitated to select the best tree from separate runs with fewer starting trees, i.e. 10 times 10 and 50 times one starting tree for METAZOA and DEUTEROSTOMIA, respectively.

Bayesian Analysis was performed with PhyloBayes-MPI version 1.3b (Lartillot et al., 2009; Lartillot et al., 2013) on the smaller dataset (METAZOA-100) using the model CAT, MTZOA+Gamma. Six chains were run in parallel for at least 5500 iterations. The first 3000 samples were discarded as burn-in. From the remaining samples every tenth tree was used to compute a majority rule consensus tree and node support in form of Bayesian posterior probabilities. A PhyloBayes analysis was also started with the complete dataset (684 species), but the chains did not come to reasonable convergence and resolution of the consensus tree after comparatively long running time (1.8 CPU years).

2.4 Modeling amino acid substitution models

Two independent MAFFT (version 6.716, Katoh et al., 2002) alignments of amino acid sequences were obtained for light- and heavy strand encoded *nad5* genes from the METAZOA dataset. Best trees were calculated with RAxML version 7.2.8 (Stamatakis, 2006), model settings MTZOA+CAT+F for rapid bootstrapping and with MTZOA+GAMMA+F for the final tree. The resulting best trees were used for optimization of the model parameters under GTR+F model for amino acids. Substitution rates were obtained from the model parameters, amino acid frequencies were calculated directly from the alignments.

2.5 Nucleotide and amino acid statistics

AT and GC skew were determined for complete genomes (plus strand) according to the formula defined by Perna and Kocher (1995), AT skew = (A-T)/(A+T) and GC skew = (G-C)/(G+C), where the letters stand for the absolute number of the corresponding nucleotides in the sequences. We also analyzed the effect of AT content, GT and AC rich strands (measured by AT and GC skew) on the amino acid composition of mitochondrial protein coding genes. Considering the first two codon positions, which are crucial for coding, amino acids were grouped as follows: F, I, K, M, N, Y (encoded by AT-rich codons AAN, ATN, TAN and TTN) versus A, G, P, R (encoded by GC-rich codons GCN, CGN, CCN, GGN) and H, K, N, P, Q, T (encoded by CA-rich codons) versus C, F, G, V, W (encoded by GT-rich codons). For a species in an alignment the fraction of a set of amino acids denotes the fraction of these amino acids with respect to the total number of amino acids of the corresponding sequence (disregarding gaps). Leucine and serine are ignored since these amino acids are encoded by more than four codons (i.e. the first two codon positions must not be the same). If not stated otherwise, statistics of the complete genome are determined for the plus strand, i.e. the strand given in RefSeq; statistics for single genes always refer to the coding strand.

2.6 Gene order divergence

Gene orders were compared using the breakpoint distance (Blanchette et al., 1999). An adjacency of a gene order G is a pair of genes that are adjacent in G. A conserved adjacency of two gene orders G and F is an adjacency in both gene orders where the corresponding genes are either in the same or opposite order and orientation. A breakpoint in a gene order with respect to another gene order is a pair of adjacent genes that is not conserved, i.e. not adjacent in the other genome. The breakpoint distance is the average number of breakpoints for two gene orders with respect to each other.

We tested the correlation between gene order rearrangements and branch length of the corresponding taxon. For gene order we excluded the highly variable positions of tRNAs, thus in most cases 15 genes were considered. The branch length of each taxon from the base of Bilateria was determined as well as the minimal number of breakpoints needed to get from the taxons` gene order to one of three proposed ground patterns (corresponding to ground pattern hypotheses of Deuterostomia, Lophotrochozoa, and Ecdysozoa). As it is currently not possible to define a single most reasonable hypothesis for ground patterns of gene order for Metazoa, we used three different gene orders, defined as putative ground patterns for Ecdysozoa, Deuterostomia, and Lophotrochozoa, for an assessment of the derived nature of a given gene order. The deuterostome pattern is still realized in most of the deuterostome mitogenomes. The ecdysozoan pattern is the same as seen in most arthropods, an onychophoran species, and a tardigrade. The priapulid pattern is different from the ecdysozoan ground pattern by an inversion of half of the genome. Nematodes have a large variety of gene order patterns, not much resembling any of the presented three ground patterns. The lophotrochozoan pattern is one which is still realized in a brachiopod, some nemertean species, and in some molluscs. It is the only pattern realized in more than one phylum of Lophotrochozoa, and it is the lophotrochozoan pattern most similar to the ecdysozoan and deuterostome patterns.

2.7 Statistical analyses

Statistical analyses have been conducted with the R package (R Development Core Team, 2011). Pearson correlation coefficients have been computed with the function lm. The Wilcoxon signed-rank test (R function wilcox.test) was used to test statistical significance, using a p-value threshold of 0.01.

3. Results & Discussion

3.1 Phylogenetic trees obtained with mitochondrial genome data

Our most comprehensive dataset (METAZOA) includes almost all mitochondrial genome entries from invertebrate metazoans and a selection of vertebrate entries. A maximum likelihood analysis of this dataset using RAxML reveals an unbalanced tree with large differences in branch lengths and a lack of supported resolution for most basal nodes, clearly indicating some major problems in deep phylogeny reconstruction of Metazoa with mitochondrial genomes using up-to-date methods (Figure 1). This figure also displays that nucleotide frequencies and strand skews strongly vary among metazoan mt genomes.

At the base of this tree Cnidaria and Porifera appear polyphyletic, with Hydrozoa forming the sister group to Bilateria + Hexactinellida. The limited taxon sampling of mitochondrial genomes for several of these groups (e.g. Hydrozoa, Scyphozoa, Hexactinellida) clearly biases the analysis in this part of the tree. Only a few basal branches are well supported by bootstrap percentages (e.g. Bilateria, Bilateria+Hexactinellida). Mitochondrial genomics and the relationships of the basal metazoan splits are in focus of another article in this special issue (Osigus et al., 2013) and thus will not be discussed in detail here.

One remarkable feature of the tree presented in Figure 1 is the increase of branch lengths among bilaterian taxa in comparison to non-bilaterian taxa and outgroup members. Some unusual sister group relations found in the Bilateria part of the tree may be due to long-branch artifacts - most strikingly the assemblage of Nematoda, Platyhelminthes, Syndermata, and some long branching arthropod taxa like Acari and Phthiraptera. This group is nested within a likewise artificially assembled arthropod clade. Here only a small amount of “high-ranking” sister group relations show support values above 80% (Onychophora + Priapulida, Branchiura + Pentastomida), whereas several well-established monophyla fail to be supported by bootstrapping and even by the best tree topology, e.g. Hexapoda, Chelicerata, and Malacostraca.

The lophotrochozoan part of the tree shows bootstrap support for some of the traditional phyla, e.g. Brachiopoda, Nemertea, Annelida sensu lato (with Sipuncula and Echiura), Entoprocta, and Bryozoa. Mollusca are not supported as a monophylum, but instead are scattered between the other lophotrochozoan taxa. As well interrelationships between the lophotrochozoan phyla are not resolved by this dataset and are essentially disturbed by the scattered distribution of molluscan subtaxa between the other lophotrochozoan taxa.

The only part of the tree which is largely congruent with phylogenetic analyses obtained with nuclear genome datasets is the Deuterostomia clade, except for the position of tunicates. The basal splits of deuterostomes are reasonable and well supported by bootstrap values. Tunicates have much longer branches and do not end up with the other deuterostomes, but instead are found as sister to the Acoela.

For an evaluation of the effects of large versus small taxon sets we conducted further analyses with smaller taxon samplings (METAZOA-300 and METAZOA-100 containing approximately 300 and 100 taxa, respectively). Results from maximum likelihood analysis for the METAZOA-300 dataset are largely similar to results of the 684 taxon dataset (see supplementary material). In the METAZOA-100 taxa dataset we omitted the long-branching Nematoda and Platyhelminthes, as well as most of the molluscan taxa, to see if these had a shifting effect on the other long branches (Figure 2). Even in this strongly reduced taxon set the topology and bootstrap support of the RAxML analysis did not differ much in quality from the trees obtained from the two larger taxon sets. Again the arthropod assemblage seems arbitrarily arranged and includes the long branching Syndermata. In the lophotrochozoan part of the tree there is some resolution with moderate bootstrap support, probably due to the absence of many molluscan taxa. Brachiopods are sister to Annelida *sensu lato* and the remaining molluscs are combined in a clade with Nemertea and Phoronida. Thus a smaller taxon set results only in a slight improvement of phylogenetic support, especially when extreme long branching taxa are omitted.

Bayesian analysis of the METAZOA-100 dataset with (PhyloBayes-MPI) resulted in a different picture. Compared to the RAxML analysis long-branch phenomena did affect the outcome to a lesser extent, e.g in contrast to the RAxML tree Syndermata is found within a lophotrochozoan clade and some long-branching arthropods like Protura, Copepoda, and Branchiura are now found among Pancrustacea. Ecdysozoa and Lophotrochozoa found maximum support by Bayesian posterior probabilities and the Mollusca are found to be monophyletic. Nevertheless, some other well-defined taxa are still not supported, e.g. Chordata. Furthermore branching patterns among arthropods and lophotrochozoans are not resolved. Thus, the PhyloBayes approach seems to be promising in phylogenetic analysis of mitochondrial genome data with strong differences in branch lengths, but is far more expensive in computational time and not yet feasible for our biggest dataset.

The unsatisfying outcome of a phylogenetic analysis using mt genome data shows that a large taxon sampling cannot solve the problems that have been shown in many former analyses using more limited taxon samplings (Steinauer et al., 2005; Jang and Hwang, 2009; Mwinyi et al., 2010). Other studies omitted long-branching taxa from the analysis to circumvent these problems (Helfenbein et al., 2004; Yokobori et al., 2008). However, often the omitted taxa are of special interest concerning their phylogenetic position.

3.2 Nucleotide and amino acid frequencies of animal mitochondrial genomes

It is known that animal mitochondrial genomes vary significantly in nucleotide composition and almost all show a bias between the two strands of the genome (Perna and Kocher, 1995; Hassanin et al., 2005). This begs the question if the shifts in nucleotide composition affect amino acid alignments and subsequent phylogenetic analyses? The abundance of nucleotides demonstrates the AT-richness in animal mt genomes: A: 15.6%-48.7%, C: 4.4%-34.7%, G: 4.8%-31.3%, T: 21.0-54.9% (values from plus strand, due to NCBI RefSeq annotation). Almost balanced nucleotide frequencies (all four nucleotides around 25%) are found only in a few species, e.g. the placozoan *Trichoplax* species, the snail *Myosotella myositis,* and the anthozoan cnidarian *Savalia savaglia*. The lowest AT content is seen in *Balanoglossus* species (51.4% and 52.8%), as well as again *Savalia savaglia* (51.7%), and *Trichoplax adhaerens* (53%). Highest AT contents are found in insects, with an extreme value of 87.4% in the parasitic wasp *Diadegma semiclausum*. Several other species from Hymenoptera, Diptera, and Lepidoptera reach values higher than 80%, as well do some mites and nematodes.

Besides AT content variation, the strand bias is another factor yielding unbalanced nucleotide frequencies. Probably due to an asymmetry in the replication process of mitochondrial genomes, GC and AT skews characterize differences between the two strands of a mitochondria genome, with one strand favoring G/T over C/A (Perna and Kocher, 1995; Hassanin et al., 2005). Since G is by far the heaviest of the four nucleotides, the GT-rich strand corresponds to the “heavy strand”. It is important to note that this is completely different from the major/minor coding strand or plus/minus strand terminology. When most genes are coded on the same strand it is easy to define this one as the major coding strand, but not in the case where both strands show a similar amount of coding genes. The plus strand is mostly defined according to the orientation of the *cox1* gene, an arbitrary convention given that gene order (and relative orientation) is variable and replication and transcription origins are difficult to detect automatically from sequence information.

Additionally, the asymmetric replication process creates nucleotide skews differing along the mitogenome (Reyes et al., 1998), depending on the position and orientation of the replication origins. Comparing GC and AT skews in arthropods gave evidence for a number of independent reversals of nucleotide skews, some of them with little or no changes in gene order, e.g. in most spiders, the varroa mite, scorpions (Hassanin, 2006), some pycnogonids (Arabi et al., 2010), and isopods (Kilpert et al., 2012). Inversion of the replication origin was discussed as a putative mechanism for a reversal of the strand bias (Hassanin et al., 2005; Wei et al., 2010). Phylogenetic analyses in the above mentioned studies yielded longer branches (=more substitutions) for clades with a reversal of strand bias than for other clades.

There is a strong negative correlation between AT and GC skew, when all mt genomes (plus-strand) from our most comprehensive alignment (METAZOA) are compared (Figure 3). There are two clusters: one with positive GC skew and clearly negative AT skew, the other with predominantly negative GC skew and positive or moderately negative AT skew. Note that using the minus strand for one of the clusters would superimpose the clusters. It is unclear if the inversion of the skews is due to an inversion of the replication origin, which is not easily determinable. The long-branching taxa (red in Figure 3) have significantly larger GC and smaller AT skews. For instance, Nematoda, Platyhelminthes, and Tunicata have combinations of highly positive GC skew and a highly negative AT skew. The following alternative hypotheses are significantly supported (Wilcoxon test, p < 10^-16): a) AT skew for problematic taxa is less than the one for non-problematic taxa b) GC skew is larger than the one for non-problematic taxa.

Phylogenetic analyses on a high taxonomic level, like the study presented here, predominantly use amino acid sequences to overcome problems with aberrant nucleotide frequencies and nucleotide skews, which were assumed to have the strongest effect on synonymous substitutions. However, the variation in AT content and GC and AT skews obviously must lead to changes on the amino acid level, too (e.g. Foster et al., 1997; Min and Hickey, 2007). Figure 4 shows an example of amino acid frequency correlations for *nad5* across Metazoa (all other genes are presented in the supplement). Because the distribution of genes on the two strands differs among metazoan mitochondrial genomes, correlations of strand bias and amino acid composition can only be analyzed separately per gene. We chose *nad5* for two reasons: (1) it is the largest and among the least conserved protein coding genes in metazoan mitochondrial genomes, thereby providing most information and (2) in metazoan species *nad5* is well distributed on the plus and minus strand (approximately 2:1). Analysis of *nad5* shows a clear negative correlation of the fraction of amino acids coded by AT rich codons and the fraction of amino acids coded by GC rich codons (Figure 4A). The slope of the linear regression for the fractions of AT-rich and GC-rich codons is approximately -0.5, i.e. as for AT and GC content of the genome. Problematic taxa with long branches in Figure 1, as depicted by red dots, are slightly shifted from the main regression line to lower proportions for both AT and GC rich codons. The effects are less prominent in more conserved genes like *cox1-3* and *cytb*, but nevertheless visible, as well as in the complete alignment (see supplementary material). Thus, a strong dependence of AT / GC content and amino acid composition can be attested. This suggests homoplasious effects, at least when extreme AT / GC contents are reached, e.g. in the case of some hexapods.

The effects of strand bias (heavy strand is GT rich; light strand is CA rich) are shown in Figure 4B. The usage of amino acids encoded by GT rich and CA rich codons has a clear negative correlation. Here the formation of two clusters is noticeable, corresponding almost perfectly to heavy and light strand encoded *nad5* genes. Problematic long branched taxa (according to Figure 1) tend to accumulate high fractions of GT-rich codons and low fractions of CA codons, corresponding to genes located on the GT rich, i.e. heavy strand (red dots in Figure 4b).

The correlation of nucleotide composition and amino acid usage is as well reflected by the strong correlation of GC (resp. AT, GT, CA) content and the fraction of GC (resp. AT, GT, CA) rich codons (see figures for each protein coding gene in supplementary material). Thus, the amino acid composition of a gene strongly depends on whether it is located on the heavy or the light strand. This effect is visible also in the bimodal frequency distribution of several amino acids, e.g. those encoded by CA rich codons (Thr ACN; Gln CAA/C; His CAT/C), corresponding to the strand bias (Figure 5).

Optimized substitution model parameters for the two subsets of heavy strand and light strand encoded *nad5* genes have been determined (see Section 2.4). In accordance with the previous results, the two optimized models differ strongly (Figure 6). Thus, the usage of a unified substitution model (as generally applied in most analyses) barely fits to a dataset where model parameters strongly depend on the orientation of the gene. Hence, instead of a “one fits all” model “heavy” and “light” strand models should be used in turn depending on which strand the gene is encoded in the corresponding part of the tree.

Altogether our results suggest a strong relation of the strand bias and amino acid sequences and thus the danger of homoplasious substitutions in taxa that achieved a similar genome organization independently (at least when inversions are involved). In addition an accelerated substitution rate may occur each time a gene switches strands, hinting to a correlation of a high frequency of gene order changes with long branches in a phylogenetic tree (see next section).

3.3 Correlations of gene rearrangements and substitution rates

The structural genome variation of mitochondrial genomes is another source of phylogenetic information. Boore et al. (1995) were the first to demonstrate phylogenetic signal in mitochondrial gene order diversity. Mitochondrial gene order stayed relatively stable in vertebrates and insects, while highly variable patterns are found in e.g. Mollusca (Boore et al., 2004), Bryozoa (Waeschenbach et al., 2006; Jang and Hwang, 2009; Nesnidal et al., 2011), Tunicata (Gissi et al., 2010; Stach et al., 2010), and Acari (Shao et al., 2006). It was mentioned several times that a higher variation in gene order may correspond with higher substitution rates and therefore promotes long branches and problems in sequence-based analysis. Studies with arthropod examples show strong correlations between gene order and sequence distances (Shao et al., 2003; Xu et al., 2006). In the case of gene rearrangements involving strand switch of genes this could be explained with the strand bias of nucleotide frequencies (Hassanin et al., 2005), which also affects amino-acid frequencies in protein coding genes (Podsiadlowski and Braband, 2006; Min and Hickey, 2007).

In the absence of a coherent model for a ground pattern of mitochondrial gene order for all Metazoa or even Bilateria, we used three putative gene order ground patterns of protein coding and ribosomal RNA genes for Deuterostomia (Bourlat et al., 2009), Ecdysozoa (Webster et al., 2006), and Lophotrochozoa (Podsiadlowski et al., 2009) (Figure 7). Using data from our most comprehensive analysis, we determined for each taxon the branch length from the root and the breakpoint distance (Blanchette et al., 1999) of its gene order compared to the three putative ground patterns (Figure 8). We found a correlation of gene order change (quantified as the minimal number of breakpoints between the gene order under view and one of the three putative ground patterns) and amino acid substitution rate (here determined as root to leaf distance, i.e. the sum of the branch lengths in the phylogenetic tree from the root to the leaf). For up to seven breakpoints highly variable branch lengths were detected, but with more than seven breakpoints the number of taxa with short branches is in a minority. This is supported by the fact that the largest breakpoint distance where the null hypothesis (that the branch lengths are less or equal than those for equal gene orders) cannot be rejected is six. Thus, seven breakpoints lead to a significant increase in branch lengths. Complete shuffling of the mitochondrial genome is clearly correlated with long branches (=high substitution rate), while a moderate gene rearrangement (2-6 breakpoints) has virtually no effect. On the other hand extremely long branches (>5) are only found in genomes which are highly rearranged (eight or more breakpoints). Extreme values for both, branch lengths and breakpoint distance are found in Nematoda, Platyhelminthes, Tunicata, some Mollusca, and some Arthropoda (Acari, Copepoda). Nevertheless it should be mentioned that even taxa with the same gene order may have substantial variation in branch lengths, reaching mean substitution rates similar to those of taxa with highly rearranged genomes. Consequently, the gene order is only one of several factors related to substitution rate differences. An alternative explanation for this correlation would be that in some taxa unknown underlying features similarly affect both, substitution rates and rearrangement rates. These putative mechanisms may be relaxed repair mechanisms, high mutational stress in combination with lower importance of mitochondrial efficiency.

3.4 A more detailed discussion of problematic taxa

Several taxa were found in unexpected position within our tree (Figure 1), pointing to problems in constructing a reliable phylogenetic tree from a mitochondrial amino acid alignment. For an in depth discussion of selected phyla see the accompanying articles in this special issue. Here we will shortly re-examine some of the unexpected results from our phylogenetic analyses in the light of our results from nucleotide skew, amino acid frequencies, and gene order changes presented in Section 3.2 and 3.3.

Nematoda, Platyhelminthes, and Syndermata are most strongly affected by long-branch attraction, as seen in our phylogenetic tree (Figure 1). A clade composed of these groups was never supported by datasets obtained from nuclear genome sequences, where Platyhelminthes and Syndermata are part of Lophotrochozoa and Nematoda are part of Ecdysozoa (Dunn et al., 2008; Hejnol et al., 2009; Pick et al., 2010). Amino acid substitution rate (estimated from branch lengths in the trees) in syndermatans and nematodes is more than doubled, while in Platyhelminthes it appears to be more than four times higher than the average substitution rate in the remaining bilaterian taxa. It is apparent, that many of the long branches are comprised of species with endoparasitic lifestyle, but not all of them - rotifers and many non-parasitic nematodes are as well represented. In Syndermata the parasitic Acanthocephala have longer branches than free-living rotifers, but in nematodes no clear correlation between branch length and parasitic lifestyle is present. Anoxic conditions, a higher metabolic rate, a short generation time, and bottleneck effects, associated with low effective population size, were discussed to affect substitution rates of mitochondrial genomes (Martin, 1995; Min and Hickey, 2008). All of these effects are not restricted to an endoparasitic lifestyle, e.g. nematodes living in rotten plants, carcasses, or dung experience similar harsh conditions. Notable is that in phylogenomic datasets using nuclear genes Nematoda, Syndermata, and Platyhelminthes are among the longest branches as well (Dunn et al., 2008; Hejnol et al., 2009), suggesting an generally accelerated substitution rate in both, mitochondrial and nuclear genomes.

Tunicata show by far the longest branches among deuterostomes and never formed an exclusive clade with Vertebrata and *Branchiostoma* (compare also Perseke et al., 2013), as clearly supported by nuclear genes and morphological data (e.g. Delsuc et al., 2006). Their gene order is completely different from all other Deuterostomia and highly variable, i.e. even congeneric species differ in gene order (Iannelli et al., 2007; Gissi et al., 2010; Stach et al., 2010). Tunicates are usually not confronted with anoxic conditions, thus their high substitution rate and gene order variation may have other (unknown) reasons. Phylogenomic analysis of nuclear genes show extremely derived sequences for *Oikopleura dioca,* but average branch lengths for *Ciona* species (Denoeud et al., 2010).

Mollusca remain one of the most problematic taxa in mitochondrial genome based analyses. Interestingly, the taxa showing the least derived gene orders (the polyplacophoran *Katharina tunicata*, some gastropods, e.g. *Haliotis*, *Ilyanassa,* and the cephalopodes) have also the shortest branches in the tree. This indicates that the most probable reason for problems in phylogenetic analysis of Mollusca relates to their frequent gene order shuffling, promoting differences in strand bias - remark the extremely different GC skews between mollusk taxa in Figure 1 - which in turn affects the amino acid usage. As for tunicates, possible reasons for the comparatively unstable gene order in mollusks are unknown. More details for molluscan mitochondrial genomes are found in an accompanying article of this special issue (Stoeger and Schroedl, 2013).

The placement of several hexapod taxa in our tree seems to be influenced by accelerated substitution rates, most prominently in Thysanoptera, Phthiraptera, some Hymenoptera, Diptera, and Hemiptera. Besides long-branch attraction there must be some other reason for the lack of support for monophyletic Hexapoda. For instance the lack of inclusion of Collembola and Diplura into hexapods is a long known problem in phylogenetic analyses with mitochondrial datasets (Nardi et al., 2003; Pisani et al., 2004). In the case of Thysanoptera, Hemiptera, and Phthiraptera mitochondrial gene orders notably differ from the hexapod ground pattern (for more details see Simon and Hadrys, 2013). This is not the case for the dipteran and hymenopteran species placed outside of the main hexapod clade in Figure 1. Here the extreme values of AT content seem to contribute to the improper result of our phylogenetic analysis.

Conclusions

Nucleotide frequencies vary broadly among Metazoa. While slight differences may be overcome by the usage of amino acid alignments, stronger deviations are also reflected in shifts in amino acid frequencies. Amino acids can be grouped according to shared physical/chemical properties and are often interchangeable without changing the functional efficiency of the corresponding protein. Thus considerable differences in AT content, changes in strand bias or replication origins, and inversion of genes strongly affect amino acid substitution rates and the outcome of phylogenetic analysis based on amino acid alignments. The correlation between gene order distances and substitution rate fits well into this picture. However, it explains the exceptional high substitution rates in, e.g. Platyhelminthes, only to a certain degree. In our phylogenetic analysis of mitochondrial genome data from a broad taxon sampling of Bilateria sufficient resolution is lacking at the base of the tree as well as for ecdysozoan and lophotrochozoan interrelations. Several presuppositions for model-based phylogenetic analyses seem to be violated (frequency stationarity, even substitution rates, directed substitutions via strand bias or selection) in this dataset. On the other hand some parts of the tree show reasonable branching pattern and good bootstrap support even for deep splits, e.g the deuterostomes (compare also Perseke et al., 2013). This suggests that mitochondrial genomes may still have value in phylogenetic analyses, at least when gene order, nucleotide frequency, and strand bias does not vary extremely among the studied taxa.

To understand the dramatic differences of nucleotide abundance, strand bias, and mitochondrial substitution rates between metazoan taxa we are in need of a more thorough comparative analysis of mitochondrial functionality in cellular metabolism and the mitochondrial genetic machinery. Recent medical research revealed an unexpected complexity of the roles that mitochondria play for the maintenance of cellular functions (Zamzami et al., 1996; Szabadkai and Duchen, 2008; Dromparis and Michelakis, 2013), e.g., integrating energy metabolism, signaling pathways, and apoptotic processes - these functional roles may have varying degrees of importance among the different metazoan taxa.

Acknowledgements

We thank Omar Rota-Stabelli and Franz Lang for contributing amino acid sequences of additional taxa. This study was supported by the following grants from German Science foundation (DFG) priority programme “Deep Metazoan Phylogeny”, SPP 1174: BL 787/2-1,2 (CB); BA 1520/10-1,2 (LP); SCHO 442/8-1,2 (AB); PO 765/4-3 (LP); STA 850/3-2; STA 850/2 (MB,AD,GF,FJ,MM,PFS); STR-683/5-2, 6-1 (THS); SCHL 229/14-1 (MP,M.Schlegel); HA 1947/5 (HH); SCHR 667/9-1 (IS,M.Schroedl); WA530/34 (BvR); MI 649/7 (BM,KM); Schi-227/20, Schi-227/26-1 (BS). BvR received additional support from DFG grant RE-3454/1-1. FJ was funded by French-German PROCOPE program [DAAD D/0628236, EGIDE PHC 14770PJ]; French-German University (DFH-UFA, Cotutelle de thèse CT-08–10); a doctoral fellowship of the German Academic Exchange Service [DAAD D/10/43622]; and a bridge scholarship of the Collège Doctoral Européen (CDE), Université de Strasbourg.

References

Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B., de Rosa R., 2000. The new animal phylogeny: reliability and implications. Proc. Natl. Acad. Sci. USA 97, 4453-4456.

Aguinaldo, A.M., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., Lake, J.A., 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387, 489-493.

Altenhoff, A.M., Dessimoz, C., 2012. Inferring orthology and paralogy. Methods Mol. Biol. 855, 259-279.

Arabi, J., Cruaud, C., Couloux, A., Hassanin, A., 2010. Studying sources of incongruence in arthropod molecular phylogenies: sea spiders (Pycnogonida) as a case study. C. R. Biol. 333, 438-453.

Avise, J.C., 2000. Phylogeography - the history and formation of species. Harvard University Press, Cambridge, Mass.

Ballard, J.W.O., Rand, D.M., 2005. The population biology of mitochondrial DNA and its phylogenetic implications. Ann. Rev. Ecol. Evol. 36, 621-642.

Ballard, J.W.O., Whitlock, M.C., 2004. The incomplete natural history of mitochondria. Mol. Ecol. 13, 729-744.

Bensasson, D., Zhang, D., Hartl, D.L., Hewitt, G.M., 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol. Evol. 16, 314-321.

Bernt, M., Braband, A., Schierwater, B., Stadler, P.F., 2013a. Genetic aspects of mitochondrial genome evolution. Mol. Phylogenet. Evol. in press.

Bernt, M., Donath, A., Juhling, F., Externbrink, F., Florentz, C., Fritzsch, G., Putz, J., Middendorf, M., Stadler, P.F., 2013b. MITOS: Improved de novo metazoan mitochondrial genome annotation. Mol. Phylogenet. Evol. in press.

Blanchette, M., Kunisawa, T., Sankoff, D., 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. J. Mol. Evol. 49, 193-203.

Boore, J.L., 2006. Requirements and standards for organelle genome databases. OMICS. 10, 119-126.

Boore, J.L., Medina, M., Rosenberg, L.A., 2004. Complete Sequences of Two Highly Rearranged Molluscan Mitochondrial Genomes, Those of the Scaphopod *Graptacme eborea* and of the Bivalve *Mytilus edulis*. Mol. Biol. Evol. 21, 1492-1503.

Boore, J.L., Collins, T.M., Stanton, D., Daehler, L.L., Brown, W.M., 1995. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature 376, 163-165.

Bourlat, S.J., Rota-Stabelli, O., Lanfear, R., Telford, M.J., 2009. The mitochondrial genome structure of *Xenoturbella bocki* (phylum Xenoturbellida) is ancestral within the deuterostomes. BMC Evol. Biol. 9, 107.

Butte, A.J., Dzau, V.J., Glueck, S.B., 2001. Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". Physiol. Genomics 7, 95-96.

Cruickshank, R.H., 2002. Molecular markers for the phylogenetics of mites and ticks. Syst. Appl. Acarol. 7, 3-14.

Danovaro, R., Dell'anno, A., Pusceddu, A., Gambi, C., Heiner, I., Kristensen, R.M., 2010. The first Metazoa living in permanently anoxic conditions. BMC Biol. 8, 30.

Delsuc, F., Brinkmann, H., Chourrout, D., Philippe, H., 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439, 965-968.

Denoeud, F., Henriet, S., Mungpakdee, S., Aury, J.M., Da, S.C., Brinkmann, H., Mikhaleva, J., Olsen, L.C. and others, 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science 330, 1381-1385.

Dress, A.W., Flamm, C., Fritzsch, G., Grunewald, S., Kruspe, M., Prohaska, S.J., Stadler, P.F., 2008. Noisy: identification of problematic columns in multiple sequence alignments. Algorithms Mol. Biol. 3, 7.

Dromparis, P., Michelakis, E.D., 2013. Mitochondria in vascular health and disease. Ann. Rev. Physiol. 75, 21.1-21.32.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W. and others, 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745-749.

Foster, P.G., Jermiin, L.S., Hickey, D.A., 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J. Mol. Evol. 44, 282-288.

Galtier, N., Nabholz, B., Glemin, S., Hurst, G.D., 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol. Ecol. 18, 4541-4550.

Gissi, C., Pesole, G., Mastrototaro, F., Iannelli, F., Guida, V., Griggio, F., 2010. Hypervariability of Ascidian Mitochondrial Gene Order: Exposing the Myth of Deuterostome Organelle Genome Stability. Mol. Biol. Evol. 27, 211-215.

Halanych, K.M., 1995. The phylogenetic position of the pterobranch hemichordates based on 18S rDNA sequence data. Mol. Phylogenet. Evol. 4, 72-76.

Halanych, K.M., 2004. The new view of animal phylogeny. Ann. Rev. Ecol. Evol. 35, 229-256.

Hassanin, A., 2006. Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Mol. Phylogenet. Evol. 38, 100-116.

Hassanin, A., Leger, N., Deutsch, J., 2005. Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa, and consequences for phylogenetic inferences. Syst. Biol. 54, 277-298.

Hausdorf, B., Helmkampf, M., Meyer, A., Witek, A., Herlyn, H., Bruchhaus, I., Hankeln, T., Struck, T.H., Lieb, B., 2007. Spiralian phylogenomics supports the resurrection of Bryozoa comprising Ectoprocta and Entoprocta. Mol. Biol. Evol. 24, 2723-2729.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G.W., Edgecombe, G.D., Martinez, P., Baguna, J. and others, 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. Roy. Soc. B 276, 4261-4270.

Helfenbein, K.G., Fourcade, H.M., Vanjani, R.G., Boore, J.L., 2004. The mitochondrial genome of *Paraspadella gotoi i*s highly reduced and reveals that chaetognaths are a sister group to protostomes. Proc. Natl. Acad. Sci. USA 101, 10639-10643.

Howe, K., Bateman, A., Durbin, R., 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics 18, 1546-1547.

Hurst, G.D., Jiggins, F.M., 2005. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proc. Biol Sci. 272, 1525-1534.

Iannelli, F., Griggio, F., Pesole, G., Gissi, C., 2007. The mitochondrial genome of *Phallusia mammillata* and *Phallusia fumigata* (Tunicata, Ascidiacea): high genome plasticity at intra-genus level. BMC Evol. Biol. 7.

Jang, K.H., Hwang, U.W., 2009. Complete mitochondrial genome of *Bugula neritina* (Bryozoa, Gymnolaemata, Cheilostomata): phylogenetic position of Bryozoa and phylogeny of lophophorates within the Lophotrochozoa. BMC Genomics 10, 167.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059-3066.

Kilpert, F., Held, C., Podsiadlowski, L., 2012. Multiple rearrangements in mitochondrial genomes of Isopoda and phylogenetic implications. Mol. Phylogenet. Evol. 64, 106-117.

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286-2288.

Lartillot, N., Rodrigue, N., Stubbs, D., Richer, J., 2013. Phylobayes-MPI. A Bayesian software for phylogenetic reconstruction using mixed models, MPI version. Online ressource.

Martin, A.P., 1995. Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. Mol. Biol. Evol. 12, 1124-1131.

Min, X.J., Hickey, D.A., 2007. DNA asymmetric strand bias affects the amino acid composition of mitochondrial proteins. DNA Res. 14, 201-206.

Min, X.J., Hickey, D.A., 2008. An evolutionary footprint of age-related natural selection in mitochondrial DNA. J. Mol. Evol. 67, 412-417.

Moritz, C., Dowling, T.E., Brown, W.M., 1987. Evolution of Animal Mitochondrial-DNA - Relevance for Population Biology and Systematics. Ann. Rev. Ecol. Syst. 18, 269-292.

Mwinyi, A., Bailly, X., Bourlat, S.J., Jondelius, U., Littlewood, D.T.J., Podsiadlowski, L., 2010. The phylogenetic position of Acoela as revealed by the complete mitochondrial genome of *Symsagittifera roscoffensis*. BMC Evol. Biol. 10.

Nardi, F., Spinsanti, G., Boore, J.L., Carapelli, A., Dallai, R., Frati, F., 2003. Hexapod origins: Monophyletic or paraphyletic? Science 299, 1887-1889.

Nesnidal, M.P., Helmkampf, M., Bruchhaus, I., Hausdorf, B., 2011. The complete mitochondrial genome of *Flustra foliacea* (Ectoprocta, Cheilostomata) - compositional bias affects phylogenetic analyses of lophotrochozoan relationships. BMC Genomics 12, 572.

Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Muller, W.E. and others, 2013. Deep metazoan phylogeny: When different genes tell different stories. Mol. Phylogenet. Evol. 67, 223-233.

Osigus, H.-J., Eitel, M., Schierwater, B., 2013. Mitogenomics at the base of Metazoa. Mol. Phylogenet. Evol.

Pattengale, N., Alipour,M., Bininda-Emonds,O., Moret,B., Stamatakis,A., 2009. How many bootstrap replicates are necessary? In: Batzoglou,S. (Ed.), Research in Computational Molecular Biology (Lecture notes in Computat. Mol. Biol. 5541). Springer Berlin, pp. 184-200.

Perna, N.T., Kocher, T.D., 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J. Mol. Evol. 41, 353-358.

Perseke, M., Golombek, A., Schlegel, M., Struck, T.H., 2013. The impact of mitochondrial genome analyses on the understanding of deuterostome phylogeny. Mol. Phylogenet. Evol. 66, 898-905.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T., Manuel, M., Worheide, G., Baurain, D., 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 9, e1000602.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E. and others, 2009. Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19, 706-712.

Pick, K.S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D.J., Wrede, P., Wiens, M., Alie, A. and others, 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. 27, 1983-1987.

Pisani, D., Poling, L.L., Lyons-Weiler, M., Hedges, S.B., 2004. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. BMC. Biol. 2, 1.

Podsiadlowski, L., Braband, A., 2006. The complete mitochondrial genome of the sea spider *Nymphon gracile* (Arthropoda: Pycnogonida). BMC Genomics 7, 284.

Podsiadlowski, L., Braband, A., Struck, T.H., von Doehren J., Bartolomaeus, T., 2009. Phylogeny and mitochondrial gene order variation in Lophotrochozoa in the light of new mitogenomic data from Nemertea. BMC Genomics 10, 364.

Podsiadlowski, L., Meusemann, K., Reumont, B.v., Fahrein, K., DambachJ., Braband, A., Misof, B., 2013. Mitochondrial genome diversity in Ecdysozoa. Mol. Phylogenet. Evol.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucl. Acids Res. 35, D61-D65.

R Development Core Team, 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna , Austria www. r-project. org.

Reyes, A., Gissi, C., Pesole, G., Saccone, C., 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. 15, 957-966.

Rubinoff, D., Holland, B.S., 2005. Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. Syst. Biol. 54, 952-961.

Shao, R., Barker, S.C., Mitani, H., Takahashi, M., Fukunaga, M., 2006. Molecular mechanisms for the variation of mitochondrial gene content and gene arrangement among chigger mites of the genus *Leptotrombidium* (Acari: Acariformes). J. Mol. Evol. 63, 251-261.

Shao, R., Dowton, M., Murrell, A., Barker, S.C., 2003. Rates of gene rearrangement and nucleotide substitution are correlated in the mitochondrial genomes of insects. Mol. Biol. Evol. 20, 1612-1619.

Simon, S., Hadrys, H., 2013. A comparative analysis of complete mitochondrial genomes among Hexapoda . Mol. Phylogenet. Evol. in press.

Srivastava, M., Begovic, E., Chapman, J., Putnam, N.H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T. and others, 2008. The Trichoplax genome and the nature of placozoans. Nature 454, 955-960.

Stach, T., Braband, A., Podsiadlowski, L., 2010. Erosion of phylogenetic signal in tunicate mitochondrial genomes on different levels of analysis. Mol. Phylogenet. Evol. 55, 860-870.

Stamatakis, A., 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688-2690.

Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57, 758-771.

Steinauer, M.L., Nickol, B.B., Broughton, R., Orti, G., 2005. First sequenced mitochondrial genome from the phylum Acanthocephala (*Leptorhynchoides thecatus*) and its phylogenetic position within Metazoa. J. Mol. Evol. 60, 706-715.

Stoeger, I., Schroedl, M., 2013. Mitogenomics does not resolve deep molluscan relationships (yet?). Mol. Phylogenet. Evol. in press.

Szabadkai, G., Duchen, M.R., 2008. Mitochondria: the hub of cellular Ca2+ signaling. Physiology 23, 84-94.

Waeschenbach, A., Telford, M.J., Porter, J.S., Littlewood, D.T.J., 2006. The complete mitochondrial genome of *Flustrellidra hispida* and the phylogenetic position of Bryozoa among the Metazoa. Mol. Phylogenet. Evol. 40, 195-207.

Webster, B.L., Copley, R.R., Jenner, R.A., kenzie-Dodds, J.A., Bourlat, S.J., Rota-Stabelli, O., Littlewood, D.T.J., Telford, M.J., 2006. Mitogenomics and phylogenomics reveal priapulid worms as extant models of the ancestral Ecdysozoan. Evol. Dev. 8, 502-510.

Wei, S.J., Shi, M., Chen, X.X., Sharkey, M.J., van, A.C., Ye, G.Y., He, J.H., 2010. New views on strand asymmetry in insect mitochondrial genomes. PLoS One. 5, e12708.

Xu, W., Jameson, D., Tang, B., Higgs, P.G., 2006. The relationship between the rate of molecular evolution and the rate of genome rearrangement in animal mitochondrial genomes. J. Mol. Evol. 63, 375-392.

Yokobori, S., Iseto, T., Asakawa, S., Sasaki, T., Shimizu, N., Yamagishi, A., Oshima, T., Hirose, E., 2008. Complete nucleotide sequences of mitochondrial genomes of two solitary entoprocts, *Loxocorone allax* and *Loxosomella aloxiata*: Implications for lophotrochozoan phylogeny. Mol. Phylogenet. Evol. 47, 612-628.

Zamzami, N., Susin, S.A., Marchetti, P., Hirsch, T., Gomez-Monterrey, I., Castedo, M., Kroemer, G., 1996. Mitochondrial control of nuclear apoptosis. J Exp. Med. 183, 1533-1544.

Figure legends

Figure 1

Phylogenetic tree obtained from Maximum Likelihood analysis with amino acid alignments from mitochondrial protein coding genes. Best tree from RAxML analysis with bootstrap support from 100 pseudoreplicates. Branches with bootstrap support below 85% are shown in gray. Some major derivations of otherwise well supported phylogenetic hypothesis are highlighted (arrows show expected placement of Platyhelminthes, Syndermata, and Tunicata; polyphyletic hexapods are blue, molluscs pink). Next to the taxon names information about GC content (blue) and GC skew (green/red) of mt genomes is given (according to plus-strand sequence). Mean value of GC skew is shown in green if negative, and red if positive (if there is large variation, a span is given, e.g. Mollusca – Gastropoda). GC content is depicted as a left bound column with 50% at the right margin.

Figure 2

Phylogenetic tree obtained with an alignment of amino acid sequences from the dataset reduced to 100 taxa (methods for taxon selection see text). (A) Best tree from RAxML analysis with bootstrap percentages (>50%) beneath the branches. Differences to the tree shown in subfigure B are highlighted by arrows. (B) Consensus tree from six independent chains of PhyloBayes-MPI. Bayesian posterior probabilities are given when >0.95. In both trees the numbers in brackets after taxon names refer to the number of species representing this taxon in this reduced data set. Insufficiently supported parts of the tree are light Grey.

Figure 3

GC-skew versus AT-skew in complete mitochondrial genomes (plus-strand) from Metazoa. Red: 153 species with long branches and unusual phylogenetic position in the tree shown in Figure 1 (Nematodes, Platyhelminthes, Syndermata, Acari, Tunicata, some hexapods, and molluscs), black: remaining taxa with reasonable phylogenetic position.

Figure 4

Amino acid usage in the mitochondrial *nad5* gene. (A) The abundance of amino acids with AT-rich codons plotted against abundance of GC-rich codons. (B) The abundance of amino acids with CA-rich codons plotted against GT-rich codons. Data points corresponding to *nad5* genes located on the CA-rich strand are shown as triangles, those of nad5 genes on GT-rich strand as squares. Red data points correspond to long branched taxa as in Figure 1 (Nematodes, Platyhelminthes, Syndermata, Acari, Tunicata, some hexapods, and molluscs).

Figure 5

Density of amino acid frequencies in the mitochondrial *nad5* gene. Density of each amino acid is plotted against its frequency in the *nad5* gene. For each amino acid two density plots were computed independently for *nad5* genes from CA rich (Grey area) and GT rich strands (white area). Blue curves are from amino acids which should be affected by strand bias (GT and CA rich codons), also indicated by the term “true”.

Figure 6

Amino acid substitution models of *nad5* encoded on CA rich and GT rich strand. Between the two models the differences in amino acid frequencies of the two sets are shown (percent point difference of absolute proportion). Blue spheres indicate more than doubled substitution rate compared to the other model.

Figure 7

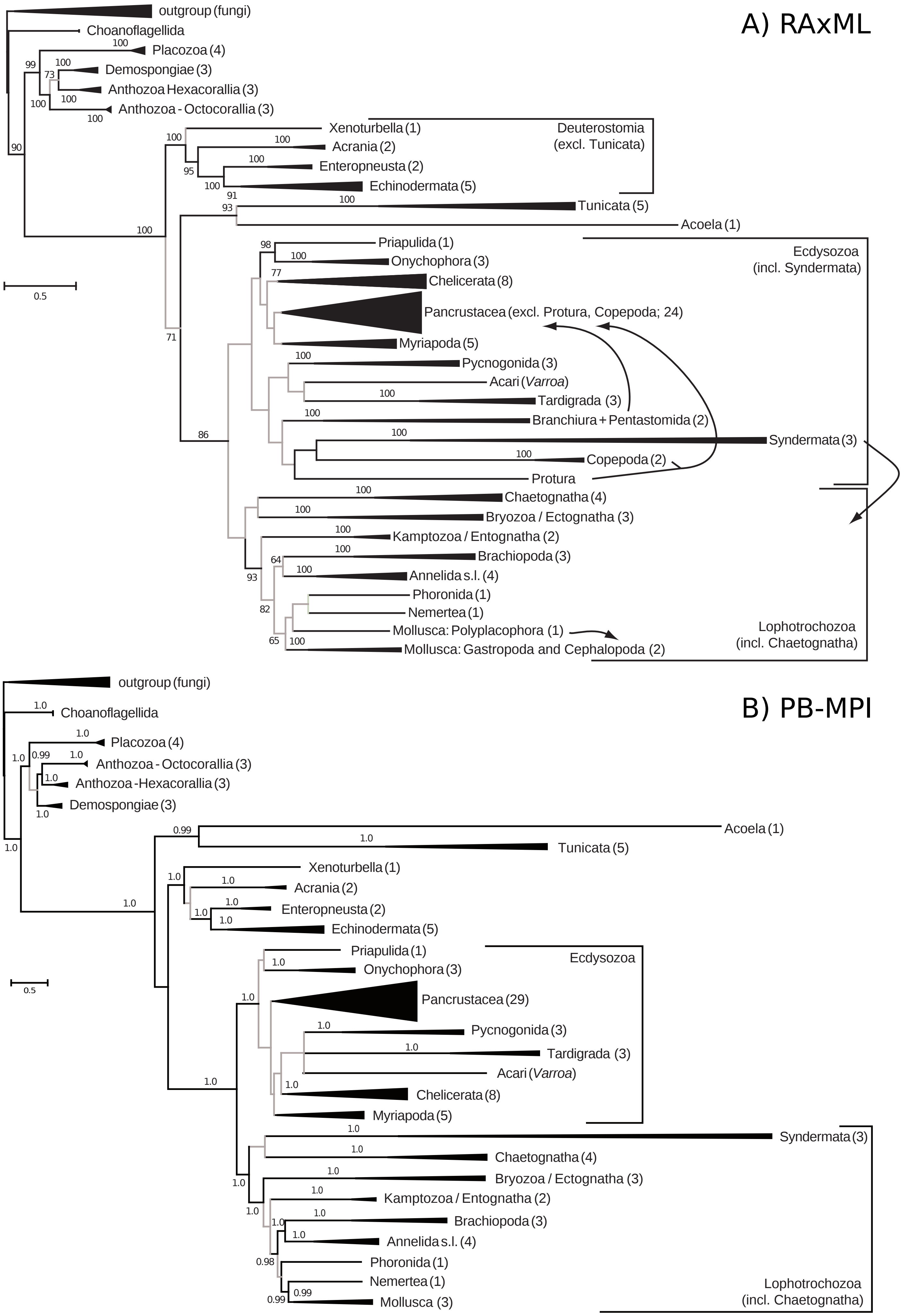
Gene order of Bilateria. Mainly protein coding and ribosomal genes are mentioned. Gene blocks found in at least two branches from the three groups Ecdysozoa, Lophotrochozoa and Deuterostomia were defined as conserved blocks. Putative ground patterns of Ecdysozoa, Lophotrochozoa, and Deuterostomia are defined for quantitative analysis of gene order change (see Figure 8).

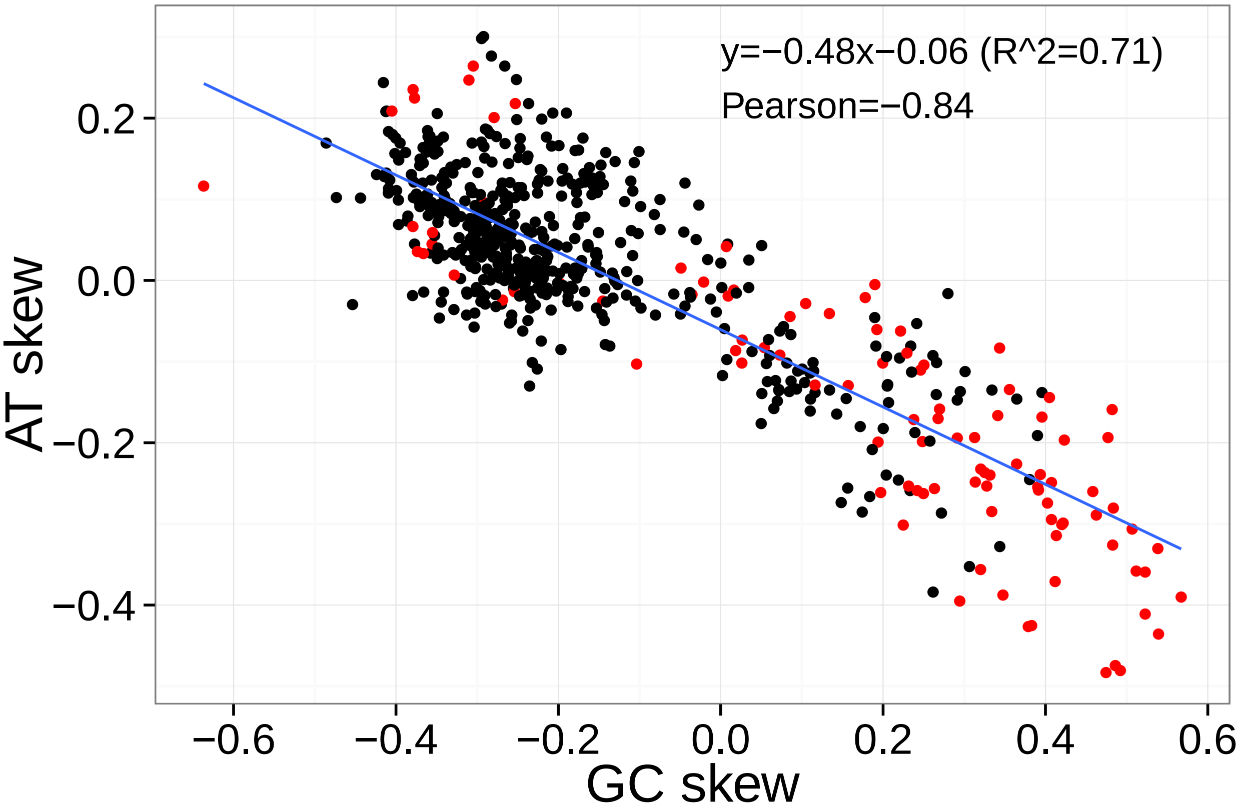
Figure 8

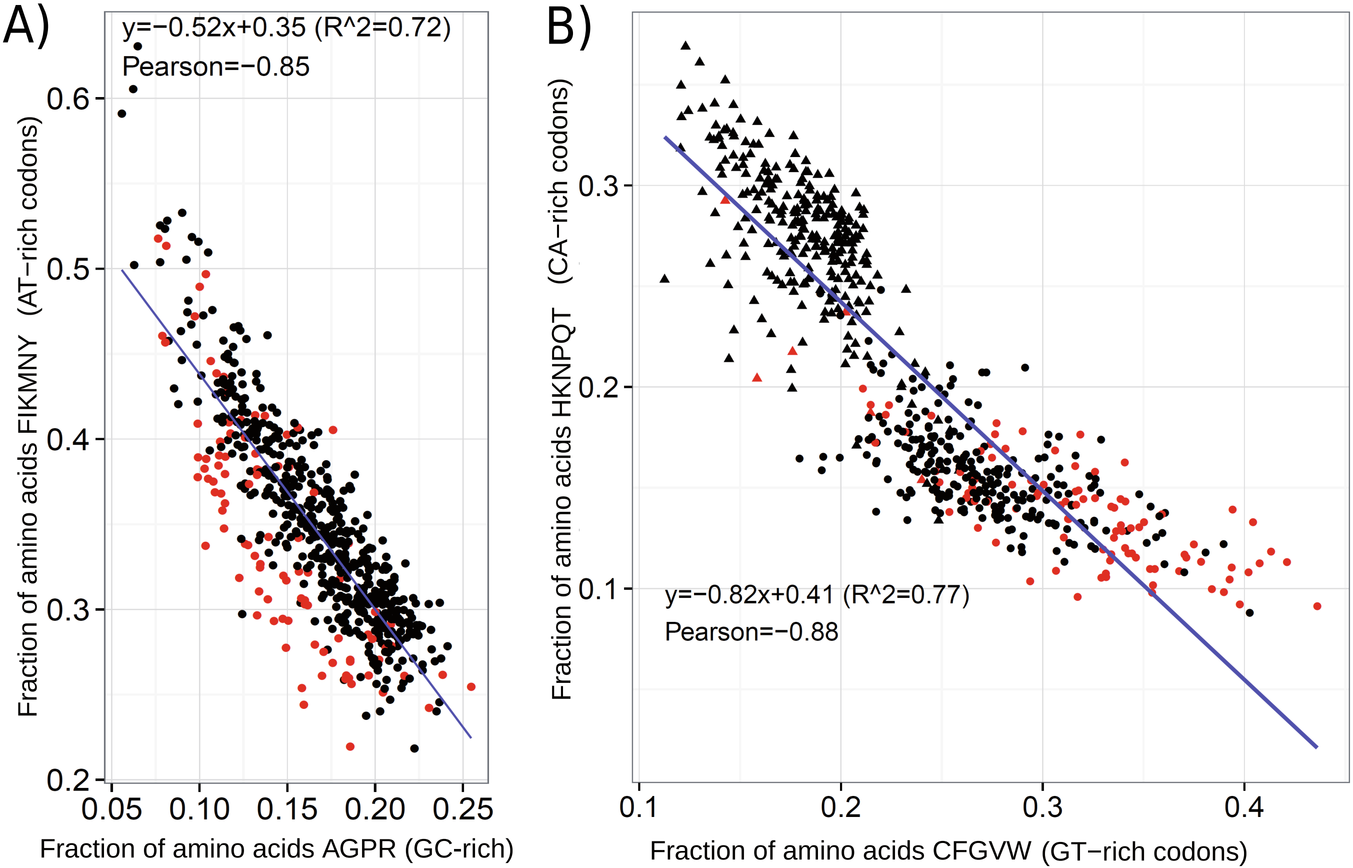
Correlation of gene order and branch length in phylogenetic analyses. Gene order changes are recorded as minimum breakpoint distance to one of three alternative ground pattern of Bilateria. Only protein coding and ribosomal RNA genes are used in this comparison. Breakpoint number is integer, data points are slightly scattered around values on the x-axis for better display of their quantity.

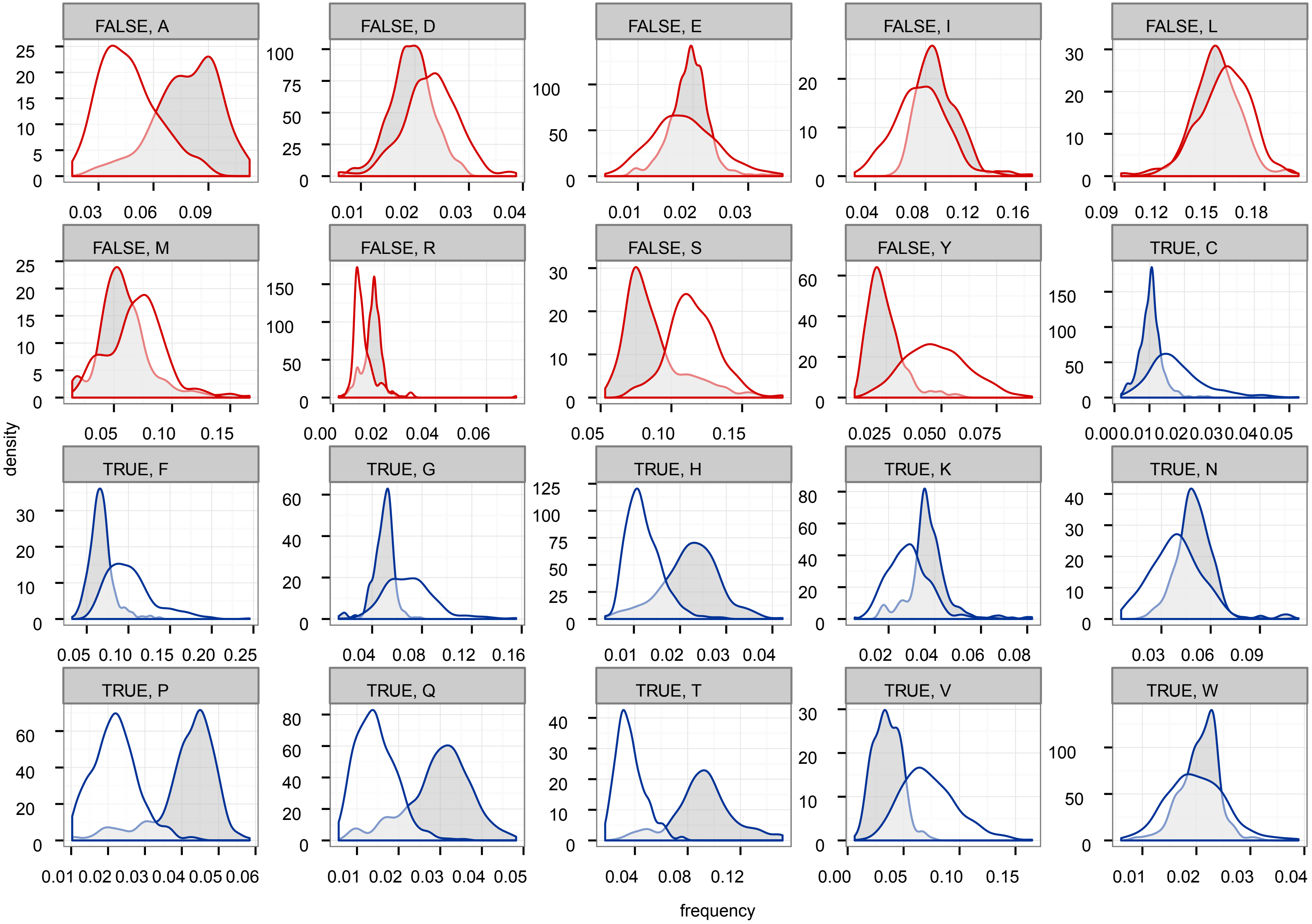
Fig-1

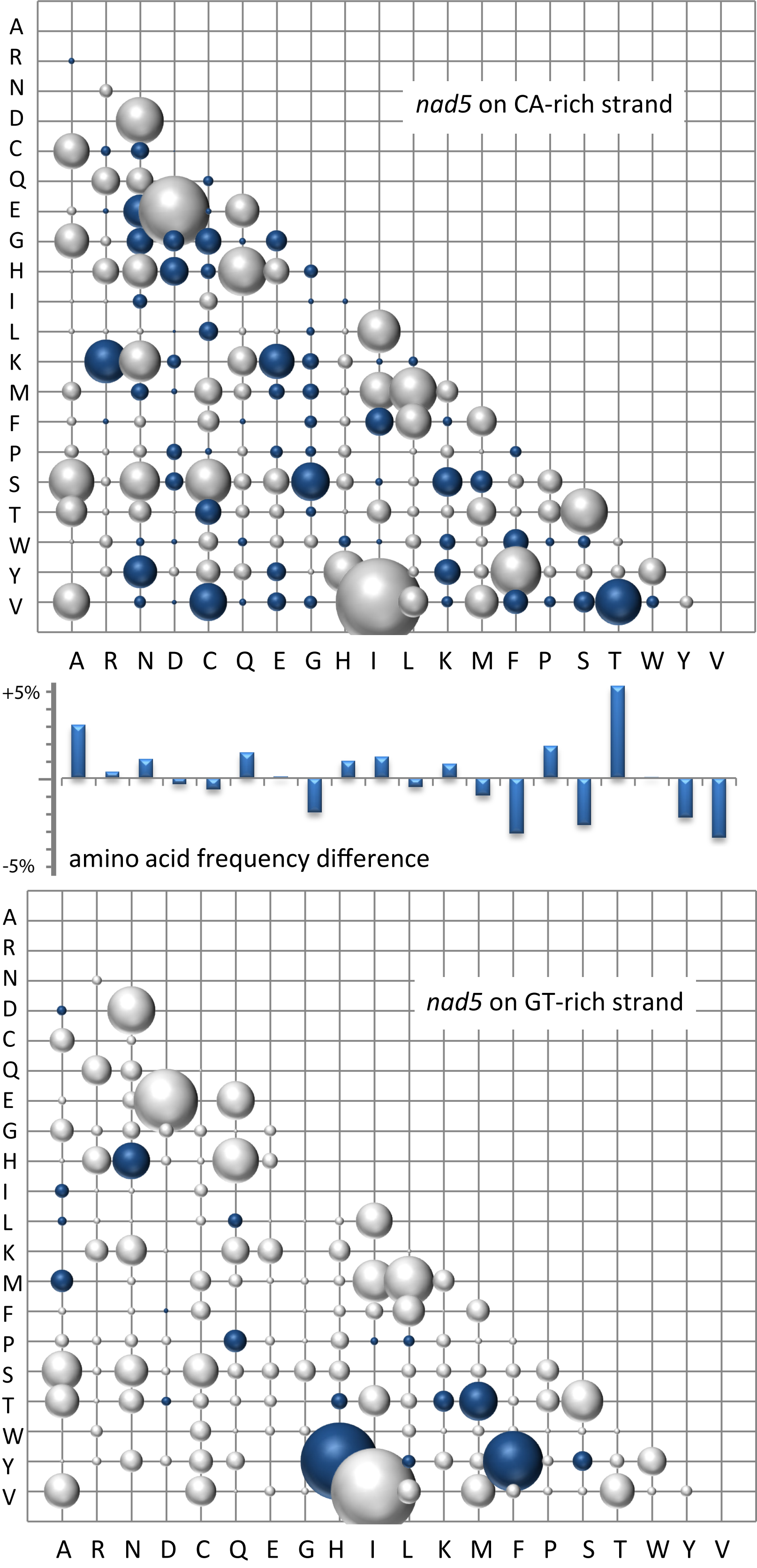
outgroup (fungi and protists)
100
Placozoa
85
Scyphozoa
100
Octocorallia
82
100
Hexacorallia
98
Homoscleromorpha
99
83
100
Demospongiae
Hydrozoa
100
Hexactinellida
Cnidaria
Porifera
100
Crustacea: Copepoda
Crustacea: Ostracoda
Crustacea: Cephalocarida
Ecdysozoa
93
Priapulida
100
Onychophora
52
Myriapoda
Chelicerata (main part)
63
Crustacea: Remipedia
100
Crustacea: Cirripedia
85
Crustacea: Branchiopoda
100
Crustacea: Isopoda
100
Hexapoda: Diplura (Campodea)
100
Hexapoda: Collembola
90
Hexapoda (main part)
and
Crustacea: Decapoda, Stomatopoda
(Hexapoda)
72
100
Crustacea: Amphipoda
100
Hexapoda: Diptera (part)
Hexapoda: Thysanoptera
96
Hexapoda: Hymenoptera
Hexapoda: Hemiptera (part)
100
Pycnogonida
100
82
Crustacea: Branchiura
100
Tardigrada
Hexapoda: Protura
88
Chelicerata: Acariformes
99
Hexapoda: Phthiraptera
82
Syndermata
65
100
54
Chelicerata: Acari (Tetranychus, Panonychus)
100
Platyhelminthes
87
Nematoda
100
Chaetognatha
Mollusca - Scaphopoda (Graptacme)
100
Mollusca - Gastropoda (part)
Mollusca - Scaphopoda (Siphonodentalium)
66
100
Bryozoa
(Mollusca)
Mollusca: Bivalvia
Mollusca - Aplacophora
100
Brachiopoda
70
100
100
Sipuncula
Annelida (incl. Echiura)
Mollusca - Polyplacophora
Phoronida
100
Nemertea
100
Mollusca - Gastropoda (part)
86
100
Mollusca - Cephalopoda
100
Entoprocta
Lophotrochozoa
Xenoturbellida
100
Acrania
Deuterostomia
100
96
100
100
Vertebrata
100
0.2
100
100
Enteropneusta
100
Echinodermata
81
Acoela
100
Tunicata
GC skew
GC content
-0.5
0
0.5
0%
25%
50%

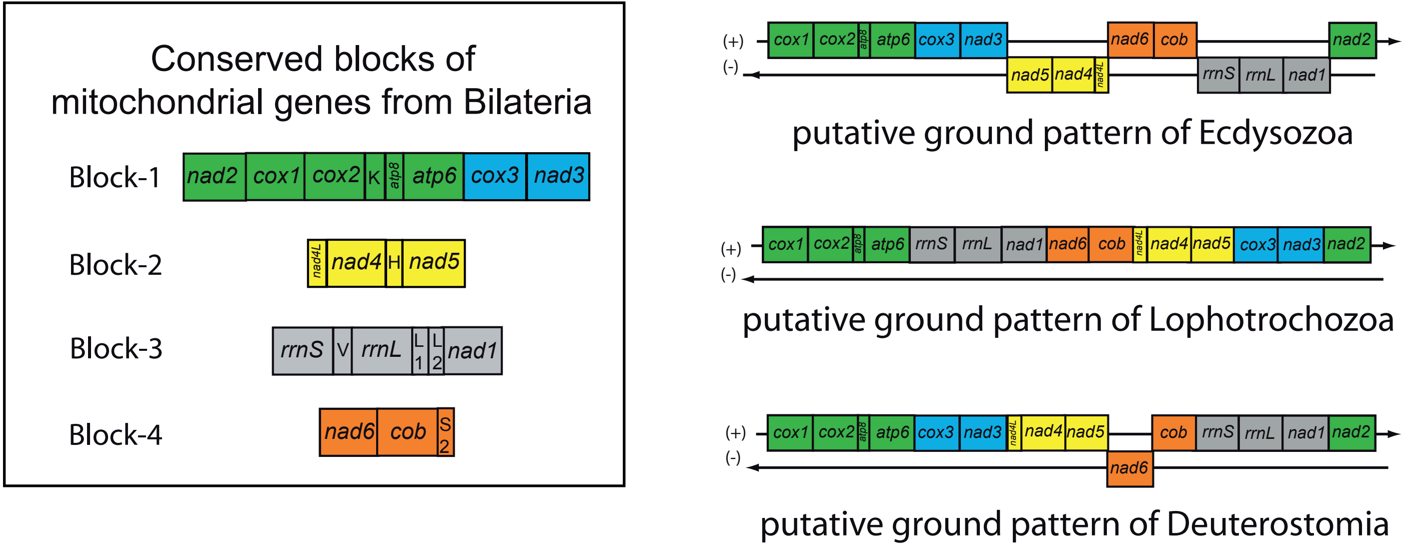

A) RAxML
outgroup (fungi)
Choanoflagellida
100
Placozoa (4)
99
100
Demospongiae (3)
73
Anthozoa Hexacorallia (3)
100
100
Anthozoa - Octocorallia (3)
100
90
Xenoturbella (1)
100
Deuterostomia
(excl. Tunicata)
100
Acrania (2)
95
100
Enteropneusta (2)
100
Echinodermata (5)
91
100
Tunicata (5)
93
Acoela (1)
100
0.5
Priapulida (1)
98
100
Onychophora (3)
Ecdysozoa
(incl. Syndermata)
77
Chelicerata (8)
Pancrustacea (excl. Protura, Copepoda; 24)
71
Myriapoda (5)
100
Pycnogonida (3)
Acari (Varroa)
100
Tardigrada (3)
100
Branchiura + Pentastomida (2)
86
100
Syndermata (3)
100
Copepoda (2)
Protura
100
Chaetognatha (4)
100
Bryozoa / Ectognatha (3)
100
Kamptozoa / Entognatha (2)
100
64
Brachiopoda (3)
100
93
Annelida s.l. (4)
Phoronida (1)
82
Nemertea (1)
Mollusca: Polyplacophora (1)
65
100
Mollusca: Gastropoda and Cephalopoda (2)
Lophotrochozoa
(incl. Chaetognatha)
B) PB-MPI
outgroup (fungi)
1.0
Choanoflagellida
1.0
Placozoa (4)
1.0
0.99
1.0
Anthozoa - Octocorallia (3)
1.0
Anthozoa -Hexacorallia (3)
Demospongiae (3)
1.0
1.0
0.99
Acoela (1)
1.0
Tunicata (5)
Xenoturbella (1)
1.0
1.0
Acrania (2)
1.0
1.0
Enteropneusta (2)
1.0
Echinodermata (5)
1.0
Priapulida (1)
Ecdysozoa
1.0
Onychophora (3)
0.5
1.0
Pancrustacea (29)
1.0
Pycnogonida (3)
1.0
Tardigrada (3)
Acari (Varroa)
1.0
1.0
Chelicerata (8)
Myriapoda (5)
1.0
Syndermata (3)
1.0
Chaetognatha (4)
1.0
Bryozoa / Ectognatha (3)
1.0
1.0
Kamptozoa / Entognatha (2)
1.0
1.0
1.0
Brachiopoda (3)
1.0
Annelida s.l. (4)
0.98
Phoronida (1)
Nemertea (1)
0.99
0.99
Mollusca (3)
Lophotrochozoa
(incl. Chaetognatha)

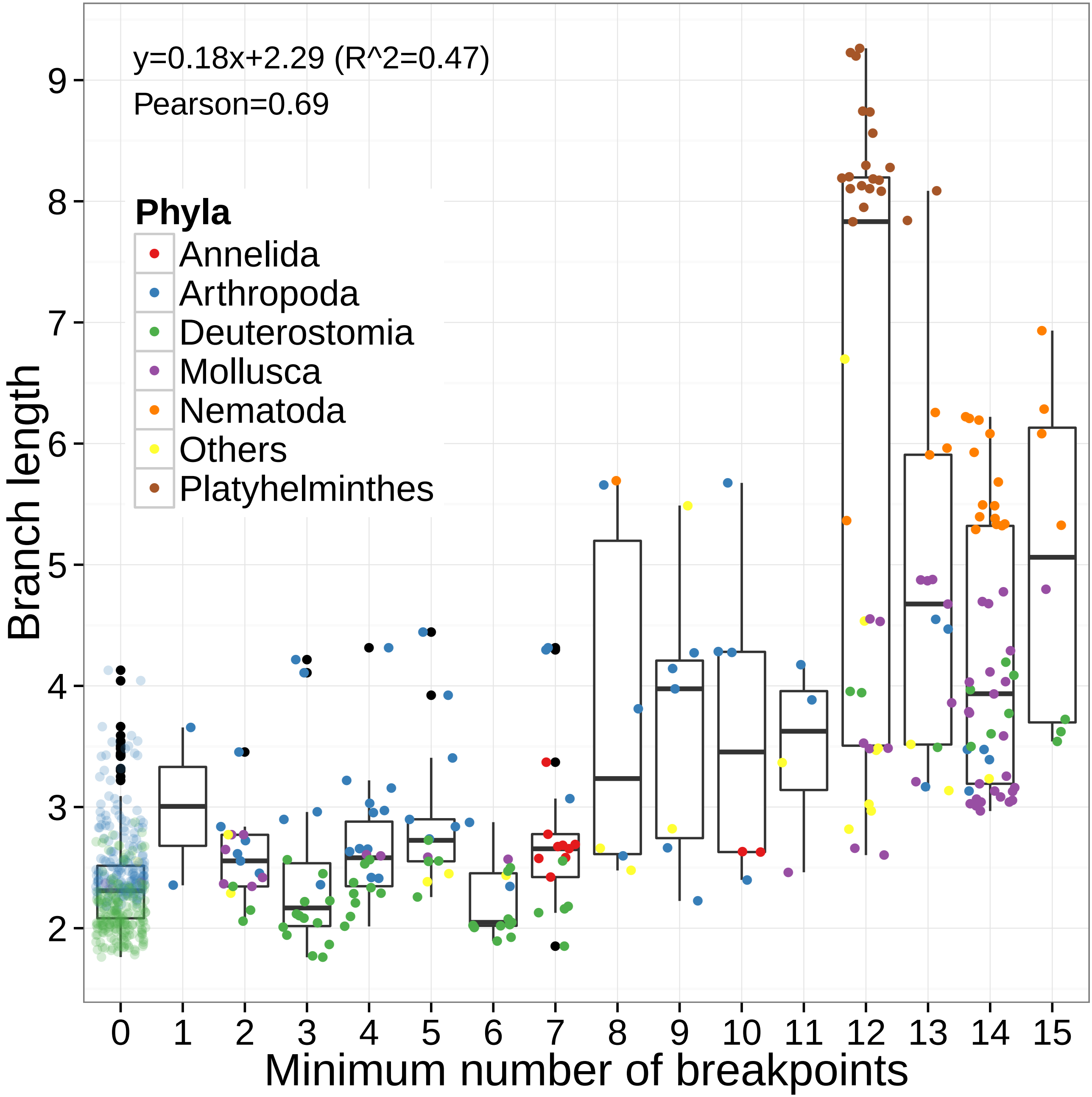

y=−0.48x−0.06 (R^2=0.71)
Pearson=−0.84
AT skew
0.2
0.0
−0.2
−0.4
GC skew
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


A)
y=−0.52x+0.35 (R^2=0.72)
Pearson=−0.85
Fraction of amino acids FIKMNY (AT−rich codons)
0.6
0.5
0.4
0.3
0.2
0.05
0.10
0.15
0.20
0.25
Fraction of amino acids AGPR (GC-rich)
B)
Fraction of amino acids HKNPQT (CA−rich codons)
0.3
0.2
0.1
y=−0.82x+0.41 (R^2=0.77)
Pearson=−0.88
0.1
0.2
0.3
0.4
Fraction of amino acids CFGVW (GT−rich codons)


FALSE, A
25
20
15
10
5
0
0.03 0.06 0.09
FALSE, D
100
75
50
25
0
0.01 0.02 0.03 0.04
FALSE, E
100
50
0
0.01 0.02 0.03
FALSE, I
20
10
0
0.04 0.08 0.12 0.16
FALSE, L
30
20
10
0
0.09 0.12 0.15 0.18
FALSE, M
25
20
15
10
5
0
0.05 0.10 0.15
FALSE, R
150
100
50
0
0.00 0.02 0.04 0.06
FALSE, S
30
20
10
0
0.05 0.10 0.15
FALSE, Y
60
40
20
0
0.025 0.050 0.075
TRUE, C
150
100
50
0
0.00 0.01 0.02 0.03 0.04 0.05
density
TRUE, F
30
20
10
0
0.05 0.10 0.15 0.20 0.25
TRUE, G
60
40
20
0
0.04 0.08 0.12 0.16
TRUE, H
125
100
75
50
25
0
0.01 0.02 0.03 0.04
TRUE, K
80
60
40
20
0
0.02 0.04 0.06 0.08
TRUE, N
40
30
20
10
0
0.03 0.06 0.09
TRUE, P
60
40
20
0
0.01 0.02 0.03 0.04 0.05 0.06
TRUE, Q
80
60
40
20
0
0.01 0.02 0.03 0.04 0.05
TRUE, T
40
30
20
10
0
0.04 0.08 0.12
TRUE, V
30
20
10
0
0.00 0.05 0.10 0.15
TRUE, W
100
50
0
0.01 0.02 0.03 0.04
frequency


A R N D C Q E G H I L K M F P S T W Y V
nad5 on CA-rich strand
A R N D C Q E G H I L K M F P S T W Y V
+5%
-5%
amino acid frequency difference
A R N D C Q E G H I L K M F P S T W Y V
nad5 on GT-rich strand
A R N D C Q E G H I L K M F P S T W Y V


Conserved blocks of mitochondrial genes from Bilateria
Block-1
nad2
cox1
cox2
K
atp8
atp6
cox3
nad3
Block-2
nad4L
nad4
H
nad5
Block-3
rrnS
V
rrnL
L 1
L 2
nad1
Block-4
nad6
cob
S 2
(+)
(-)
cox1
cox2
atp8
atp6
cox3
nad3
nad6
cob
nad2
nad5
nad4
nad4L
rrnS
rrnL
nad1
putative ground pattern of Ecdysozoa
(+)
(-)
cox1
cox2
atp8
atp6
rrnS
rrnL
nad1
nad6
cob
nad4L
nad4
nad5
cox3
nad3
nad2
putative ground pattern of Lophotrochozoa
(+)
(-)
cox1
cox2
atp8
atp6
cox3
nad3
nad4L
nad4
nad5
cob
rrnS
rrnL
nad1
nad2
nad6
putative ground pattern of Deuterostomia


y=0.18x+2.29 (R^2=0.47)
Pearson=0.69
Phyla
Annelida
Arthropoda
Deuterostomia
Mollusca
Nematoda
Others
Platyhelminthes
Branch length
2
3
4
5
6
7
8
9
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Minimum number of breakpoints