Integrative and Comparative Biology Advance Access published May 13, 2015

Integrative and Comparative Biology

Integrative and Comparative Biology, pp. 1–12
doi:10.1093/icb/icv037

Society for Integrative and Comparative Biology

SYMPOSIUM

Employing Phylogenomics to Resolve the Relationships among Cnidarians, Ctenophores, Sponges, Placozoans, and Bilaterians

Nathan V. Whelan,[1,*] Kevin M. Kocot[†] and Kenneth M. Halanych*

*Department of Biological Sciences, Molette Biology Laboratory for Environmental and Climate Change Studies, Auburn University, 101 Life Sciences Building, Auburn, AL 36849, USA; [†]School of Biological Sciences, The University of Queensland, 325 Goddard Building, St Lucia, QLD 4101, Australia

From the symposium ‘‘Origins of Neurons and Parallel Evolution of Nervous Systems: The Dawn of Neuronal Organization’’ presented at the annual meeting of the Society for Integrative and Comparative Biology, January 3–7, 2015 at West Palm Beach, Florida.

[1]E-mail: nwhelan@auburn.edu

Synopsis Despite an explosion in the amount of sequence data, phylogenomics has failed to settle controversy regarding some critical nodes on the animal tree of life. Understanding relationships among Bilateria, Ctenophora, Cnidaria, Placozoa, and Porifera is essential for studying how complex traits such as neurons, muscles, and gastrulation have evolved. Recent studies have cast doubt on the historical viewpoint that sponges are sister to all other animal lineages with recent studies recovering ctenophores as sister. However, the ctenophore–sister hypothesis has been criticized as unrealistic and caused by systematic error. We review past phylogenomic studies and potential causes of systematic error in an effort to identify areas that can be improved in future studies. Increased sampling of taxa, less missing data, and *a priori* removal of sequences and taxa that may cause systematic error in phylogenomic inference will likely be the most fruitful areas of focus when assembling future datasets. Ultimately, we foresee metazoan relationships being resolved with higher support in the near future, and we caution against dismissing novel hypotheses merely because they conflict with historical viewpoints of animal evolution.

Introduction

Phylogeny is the cornerstone of comparative biology, and interpretations of phenotypic evolution hinge on accurate hypotheses of organismal relationships (Felsenstein 1985). Transcriptomic and genomic sequences offer a nearly overwhelming source of information for inferring relationships, with some studies employing hundreds (Kocot et al. 2011; Bond et al. 2014; Cannon et al. 2014; Fernández et al. 2014; Moroz et al. 2014; Struck et al. 2014; Wickett et al. 2014) or thousands (Hejnol et al. 2009; Jarvis et al. 2014; Sharma et al. 2014) of genes. Despite great potential, phylogenomics has thus far failed to confidently resolve relationships of many animal groups (Dunn et al. 2014). Inferring relationships among major metazoan lineages (i.e., Bilateria, Ctenophora, Cnidaria, Placozoa, and Porifera) has been particularly difficult, with numerous recent studies recovering conflicting phylogenetic topologies (Fig. 1) (Dunn et al. 2008; Hejnol et al. 2009; Philippe et al. 2009, 2011; Pick et al. 2010; Nosenko et al. 2013; Ryan et al. 2013; Moroz et al. 2014; Borowiec et al. 2015; Whelan et al. 2015). This hinders our ability to study evolution of complex traits associated with the transition from unicellularity to multicellularity in early animals and their ancestor.

Of particular interest is evolution of neurons and complex neural systems. Communication between cells is integral for multicellular organisms (Kaiser 2001), and neurons provide a rapid communication network for most animals (Moroz 2009). Historically, the presence of neurons has been used as a morphological feature uniting ctenophores, cnidarians, and bilaterians (Ax 1996), but some phylogenomic studies have conflicted with this hypothesis and recovered ctenophores as the earliest branching lineage on the animal tree of life (Dunn et al. 2008; Hejnol et al. 2009; Nosenko et al. 2013; Ryan et al.

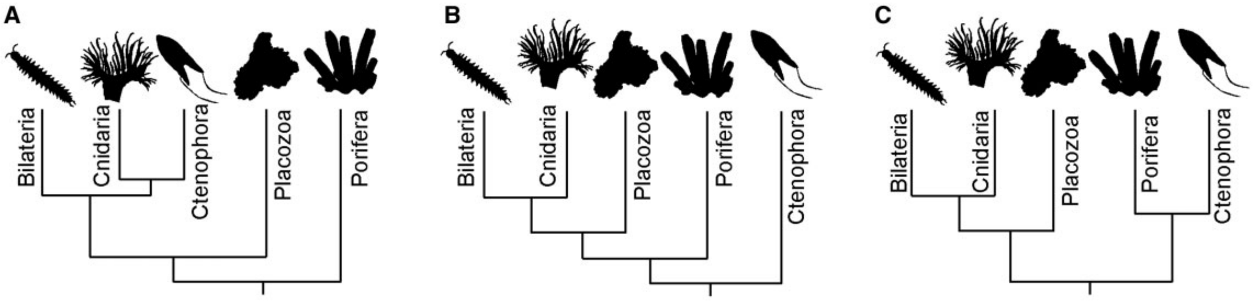

Fig. 1 Phylogenetic hypotheses of major metazoan lineages. (**A**) Traditional Porifera-sister hypothesis (Philippe et al. 2009, 2011; Pick et al. 2010; Nosenko et al. 2013), (**B**) Ctenophora-sister hypothesis (Dunn et al. 2008; Hejnol et al. 2009; Nosenko et al. 2013; Ryan et al. 2013; Moroz et al. 2014; Borowiec et al. 2015; Whelan et al. 2015), (**C**) Ctenophora + Porifera-sister hypothesis (Ryan et al. 2013).

2013; Moroz et al. 2014). Placement of ctenophores sister to all other animals would imply either parallel evolution of neural systems or extensive loss in sponges and placozoans (Ryan et al. 2013; Moroz et al. 2014). Pushback against the ctenophore–sister hypothesis has raised the possibility that systematic error caused ctenophores to be recovered sister to other animals in some of the aforementioned studies (Pick et al. 2010; Philippe et al. 2011; Nosenko et al. 2013). In order to move toward resolving metazoan relationships, pitfalls of current methods and ideal paths forward must be identified and addressed.

In recent years, systematists have faced many theoretical and methodological challenges associated with analyzing high-throughput sequencing data for phylogenetic inference (Nekrutenko and Taylor 2012; Chan and Ragan 2013; van Djik et al. 2014), and a major bottleneck for modern phylogenetic studies is the analysis of data, rather than the generation of sequences. Modern phylogenomics requires a new set of expertise and methodologies compared with phylogenetic studies with only one or a few genes. Analysis of sequences now requires multifaceted bioinformatic pipelines that piece together high-throughput assembly of sequences (El-Metwally et al. 2013; Nagarajan and Pop 2013), identification of orthologous sequences from incomplete transcriptomic or genomic data (Koonin 2005; Pearson and Sierk 2005; Dutilh et al. 2007), removal of potential causes of systematic error (Felsenstein 1978; Weisburg et al. 1989; Yang 1996), and, finally, phylogenetic inference (Lartillot et al. 2013; Stamatakis 2014). Determining which methods are the most appropriate to employ for these steps not only requires theoretical considerations, but also practical considerations given computational limitations.

Here, we examine data and methodologies that have led to support for different conflicting hypotheses of relationships among major metazoan lineages and consider future directions to resolve metazoan phylogeny. Analyses recovering non-traditional metazoan relationships have raised concerns that systematic error, as a result of poor quality of the data, or misapplied methods, has erroneously resulted in the inference of ctenophores sister to other metazoans. However, the assumed placement of sponges sister to other animals has not been as critically examined (Halanych 2015), and recent evidence suggests that even sponge choanocytes, a major morphological feature used to corroborate the sponge–sister hypothesis (Nielsen 2008), may not be homologous to choanoflagellates (Mah et al. 2014). Confident conclusions about early metazoan phylogeny and the pattern of the evolution of neurons cannot be made until a single, robust hypothesis of animal phylogeny is widely accepted. Biologists of all backgrounds should understand potential causes of systematic error and bioinformatic challenges associated with phylogenomic inference. Such understanding will promote an appreciation of how new and sometimes controversial hypotheses of early animal evolution have been generated and will be better resolved in the future.

Potential sources of systematic error

At least eight phylogenomic studies have recovered ctenophores as sister to other animals (Dunn et al. 2008; Hejnol et al. 2009; Nesnidal et al. 2013; Nosenko et al. 2013; Ryan et al. 2013; Moroz et al. 2014; Borowiec et al. 2015; Whelan et al. 2015). However, concerns have been raised that systematic error, usually long-branch attraction (LBA; Felsenstein 1978), a type of sequence-saturation that randomizes phylogenetic signals, caused ctenophores to be incorrectly placed (Pick et al. 2010; Philippe et al. 2011; Nosenko et al. 2013; Jékely et al. 2015). Other potential sources of systematic error include a limited sampling of taxa (Zwickl and Hillis 2002; Heath et al. 2008), too few characters to resolve relationships (Gatsey et al. 2007), misspecification of the model (Lartillot et al. 2007; Philippe et al. 2011; Straub et al. 2014), and misaligned sequences (Ogden and Rosenberg 2005). Sampling of taxa and characters are critical considerations for gathering data for any phylogenetic study. Misspecification of the model occurs when the probabilistic model of sequence evolution underlying phylogenetic inference poorly fits empirical sequence data. When sequences of a group of taxa are aligned for any given gene, aligned positions are statements of homology (Morrison and Ellis 1997) used in phylogenetic inference. Therefore, inaccurately aligned regions can introduce error and noise (i.e., homoplasy) into those inferences. Whereas sampling of taxa and characters are issues of experimental design, misspecification of the model, alignments of sequences, and saturated datasets usually are addressed by bioinformatics protocols.

Systematic error cannot be absolutely ruled out as the reason ctenophores have been resolved as sister to other animals even though both Ryan et al. (2013) and Moroz et al. (2014) tried to control for such error. Nonetheless, some have forcefully argued (Pick et al. 2010; Philippe et al. 2011; Nosenko et al. 2013) that systematic error *must* be the reason ctenophores were resolved as sister to all remaining animals in some studies despite weak support for sponges sister to all remaining animals in the above three studies. Nevertheless, future studies must take steps to limit the potential influence of error. The remainder of this article will focus on possible sources of systematic error and how they relate to recent and future analyses of relationships among Ctenophora, Cnidaria, Placozoa, Porifera, and Bilateria.

Sampling of taxa

Any phylogenetic study depends on adequate sampling of taxa. Collecting non-bilaterian metazoan species that represent sufficient phylogenetic diversity can be difficult as some are rare or difficult/expensive to collect. For example, most hexactinellid sponges are found in deep-sea or polar habitats (Janussen and Reiswig 2009). Tissue must also be appropriately preserved, which is not a trivial consideration. Most sequencing for phylogenomics currently utilizes RNA for transcriptomic sequencing, but RNA quickly degrades as an animal dies. Therefore, tissue must be flash-frozen or preserved with special fixatives like RNAlater (Life Technologies Inc.) as soon as possible after collection. Deep-sea animals, however, may not survive the trip to the surface when being collected, necessitating quick processing and appropriate preservation techniques. Field biologists must be aware of these considerations to maximize the utility of rarely collected animals for molecular work.

Further complicating the sampling of taxa is the fact that the majority of animal species are extinct (Simpson 1952; Raup 1986), and the common strategy of adding more taxa to an analysis to increase phylogenetic resolution will likely not work for some non-bilaterian metazoan lineages. Several phylogenomic studies have identified ctenophores as being a long-branched taxon with high mutation rates (Philippe et al. 2009, 2011; Osigus et al. 2013). However, ctenophores likely underwent a relatively recent radiation (Podar et al. 2001; Simion et al. 2015) and additional sampling will not considerably shorten the branch leading to the most recent common ancestor of extant ctenophores. A similar problem exists for Placozoans. *Trichoplax adhaerens* is the only nominal placozoan species, and other described genetic lineages are all closely related (Pearse and Voigt 2007; Eitel and Schierwater 2010; Eitel et al. 2013), which makes breaking up the long branch leading to modern-day Placozoans unlikely in future studies.

Most phylogenomic studies of non-bilaterian metazoans have used only one to four ctenophore species (Dunn et al. 2008; Hejnol et al. 2009; Philippe et al. 2009, 2011; Pick et al. 2010; Ryan et al. 2013) but Moroz et al. (2014) included 11 species in one analysis. Compared with their phylogenetic analyses involving only three ctenophore species, statistical support for the position of ctenophores was lowest when all 11 ctenophore species were included. However, the dataset with greater sampling of ctenophores also used stricter criteria for orthology than those applied to their other datasets, which confounds interpretations of how important increased sampling of ctenophore taxa was to phylogenetic inference. Nonetheless, decreased nodal support when more ctenophores were added raises the question of whether the inferred position of ctenophores is real, or instead caused by systematic error as a result of poor sampling of taxa. Although recent phylogenomic studies have focused on the placement of ctenophores, sampling of sponges and cnidarians is just as important for determining relationships among these taxa. Given the unstable placement of sponges (i.e., sister or not; Table 1) and the debate surrounding sponges' monophyly (Philippe et al. 2009; Nosenko et al. 2013; Ryan et al. 2013; Moroz et al. 2014; Riesgo et al. 2014) versus paraphyly

Table 1 Character-sampling schemes, substitution model employed in phylogenomic inference, and the hypothesized sister lineage to all other extant metazoans in past phylogenomics studies

Dataset	Genes	Gene occupancy (%)	Missing data including gaps (%)	Taxa	Total number of sites	Substitution model[a]	Inferred sister lineage
Dunn et al. (2008)	150	49	57	77	21,152	WAG & CAT	Ctenophora
Philippe et al. (2009)	128	81	27	55	30,257	CAT	Porifera
Hejnol et al. (2009) full dataset	1487	19	84	94	270,580	RTREV	Ctenophora
Hejnol 844 genes	844	25	80	94	153,925	RTREV	Ctenophora
Hejnol 330 genes	330	33	73	94	55,594	RTREV	Ctenophora
Hejnol 50 genes	50	50	56	94	7467	RTREV	Porifera[b]
Nosenko et al. (2013) full dataset[c]	122	85	28	71	23,799	CAT	Porifera
Nosenko non-ribosomal[c]	88	78	27	71	14,612	CAT	Ctenophora
Nosenko ribosomal[c]	35	75	30	71	9187	CAT	Porifera
Ryan et al. (2013) “genome”[c]	242	88	19	19	104,840	GTR & CAT	Ctenophora or Ctenophora + Porifera
Ryan et al. (2013) “EST”[c]	406	52	58	70	88,384	GTR & CAT	Ctenophora
Moroz et al. (2014) large dataset	586	51	45	44	170,871	WAG	Ctenophora
Moroz et al. (2014) more taxa	115	53	52	60	22,772	WAG	Ctenophora
Whelan et al. (2015)	89–251	71–82	35–44	60–76	23,680–81,008	Partitioned AA models & CAT	Ctenophora
Borowiec et al. (2015) “best108 matrix”	108	84	16	36	41,808	Partitioned AA models & CAT	Ctenophora

[a] Analyses using the CAT model were done with PhyloBayes. All others were done in RAxML.
[b] Support was low and sponges were paraphyletic.
[c] All outgroups included.

(Sperling et al. 2007; Nosenko et al. 2013; Osigus et al. 2013), increased sampling of poriferans may also be particularly fruitful for better resolution of non-bilaterian metazoan relationships.

Sampling of non-metazoan taxa must also be considered so trees can be accurately rooted. Outgroups allow characters to be polarized and the direction of evolution from one character-state to another to be determined through rooting the topology. Employing only distantly related or rapidly evolving outgroups can cause LBA artifacts (Philippe et al. 2005; Heath et al. 2008; Rota-Stabellia and Telford 2008). The choice of outgroup can be most problematic when the sister group is unknown, but choanoflagellates are widely accepted as the sister lineage to metazoans with Filasterea, Ichthyosporea, and Fungi also being closely related (King et al. 2008; Suga et al. 2013; Cavalier-Smith et al. 2014). Philippe et al. (2009), Ryan et al. (2013), and Moroz et al. (2014) explored how topologies were affected by different outgroup-sampling schemes, but this was primarily done to assess whether LBA was affecting phylogenetic inference rather than to determine an ideal set of outgroup taxa for best inferring animals' relationships. Notably, Philippe et al. (2009) recovered weaker support for a sister relationship between ctenophores and cnidarians with denser outgroup-sampling, but support was identical for sponges sister to all other animals under both schemes. Ryan et al. (2013) found that support for recovered relationships varied with different sampling of the outgroup, but inconsistent differences among methods of inference and of sampling characters make generalizing about effects of outgroup-sampling difficult. Despite challenges associated with the sampling of taxa, analyses that fully resolve early metazoan lineages will no doubt have more extensive sampling than in past studies.

Sampling of characters

Early phylogenomic studies on the earliest branches of the metazoan phylogeny relied heavily on expressed sequence tags via Sanger sequencing of cDNA libraries (Dunn et al. 2008; Hejnol et al. 2009; Philippe et al. 2009). Subsequent studies utilized 454 and Illumina high-throughput sequencing technologies (Nosenko et al. 2013; Ryan et al. 2013;

Moroz et al. 2014), which generate hundreds of thousands to millions of sequencing reads. Transcriptomic sequencing, particularly on Illumina platforms, is currently the most common method for generating large amounts of data for phylogenomic inference (Johnson et al. 2013; Bond et al. 2014; Fernández et al. 2014; Lemer et al. 2015), although whole-genome sequencing will likely become more common in coming years (e.g., Jarvis et al. 2014). Many past phylogenomic datasets used to infer metazoan relationships have had large amounts of missing data (Table 1), owing to the shotgun nature of sequencing technologies and differential expression of genes in different tissue used for RNA extractions. Missing data are known to affect phylogenetic inference (Lemmon et al. 2009; Roure et al. 2013), but phylogenomic data matrices have continuously seen a decrease in the amount of missing data due to high-throughput sequencing. Thus, future studies should benefit from fewer missing data as sequencing technologies continue to improve.

Of the most recent phylogenomic studies focusing on non-bilaterian metazoan relationships, that of Nosenko et al. (2013) had the smallest amount of data overall but also the smallest amount of missing data (Table 1). They analyzed datasets with the number of genes ranging from 35 to 122 and the number of amino-acid (aa) positions ranging from 9187 to 22,975. Nosenko et al. (2013) also analyzed non-ribosomal genes (35 genes, 9187 aa) and ribosomal genes (87 genes, 14,615 aa) separately, which appeared to have the greatest affect on phylogenetic inference as ctenophores were recovered sister to cnidarians in all analyses except when ribosomal protein genes were excluded. Ryan et al. (2013) employed two character-sampling schemes, one with 242 genes and 104,840 aa positions and one with 406 genes and 88,384 aa positions; the latter had many more missing data (Table 1). However, non-bilaterian relationships recovered by Ryan et al. (2013) appeared to be more influenced by substitution model and phylogenetic method (i.e., maximum likelihood [ML] with the site-homogenous general time reversible (GTR) model versus Bayesian inference [BI] with the site-heterogeneous CAT model) and by taxon-sampling rather than character-sampling. Moroz et al. (2014) had two primary datasets, one with 586 genes and 170,871 amino acids and one with 114 genes and 22,722 amino acids. Ryan et al. (2013) and Moroz et al. (2014) recovered ctenophores as sister to all other animals, but there was much lower support for critical nodes in their smaller datasets. In all three studies, differences in character-sampling were influenced by taxon-sampling (e.g., choice of outgroup) as fewer missing data were accompanied by a reduced sampling of taxa.

Saturation

Mutational saturation is also a concern for studies of deep evolutionary events. Sequence-saturation occurs when genes have undergone enough mutations between speciation events that observed genetic distances underestimate actual genetic distances, which can introduce homoplasy into sequence data. This process essentially randomizes sequences, obscuring phylogenetic signal. LBA, a special case of saturation, has been implicated as a reason ctenophores have been resolved as sister to all other animals (Philippe et al. 2011; Nosenko et al. 2013). LBA occurs when two rapidly evolving, or highly divergent, unrelated lineages are artificially drawn to each other during phylogenetic reconstruction (Felsenstein 1978; Hendy and Penny 1989). This is often manifested when an ingroup branch with a high substitution rate (i.e., a long branch) is drawn to the base of a tree toward outgroups (Bergsten 2005). Ctenophores tend to have long branches in molecular phylogenies so LBA seems like a plausible explanation for their recovery at, or near, the base of Metazoa. However, long branches toward the base of a tree may reflect real relationships.

Nosenko et al. (2013) noted higher saturation rates in their dataset of non-ribosomal genes compared with their ribosomal dataset and used this observation to justify placing more credibility upon the ribosomal dataset that recovered ctenophores as sister to cnidarians over the alternative finding of ctenophores as the sister lineage to other animals. Rather than relying too heavily on a single class of gene that happens to be less saturated than others, two alternative approaches would be to remove the most saturated genes regardless of their class or choose genes that produce individual gene trees with the highest average bootstrap support (see Salichos and Rokas 2013). For example, studies on arachnid (Sharma et al. 2014), arthropod (Regier et al. 2008), echinoderm (Telford et al. 2013), eukaryote (Hampl et al. 2009), and non-bilaterian animal (Borowiec et al. 2015; Whelan et al. 2015) relationships have ranked genes by evolutionary rate to remove the most saturated genes, and this may be an ideal method for removing the most saturated genes in future studies on metazoan relationships. No matter the approach, datasets with more signal and less noise will be essential for finally resolving the phylogenetic position of ctenophores.

Phylogenetics and conflicting tree inferences

One of the most troubling and difficult to explain conflicts in phylogenomic studies addressing the relative placement of cnidarians, ctenophores, placozoans, sponges, and bilaterians is the difference between hypotheses generated with ML and BI. At their core, both approaches are model-based algorithms, but ML is a frequentist approach (Felsenstein 1973, 1981), whereas BI is, of course, a Bayesian approach (Rannala and Yang 1996; Huelsenbeck et al. 2001). As a Bayesian statistical method, BI relies on prior distributions of probability, which have been identified as a potential bias (Pickett and Randle 2005; Yang and Rannala 2005; Ekman and Blaalid 2011) that should probably be more carefully considered than typically occurs (Brown et al. 2010; Rannala et al. 2012). Bayesian methods also employ Markov Chain Monte Carlo (MCMC) approaches (Gilks et al. 1996) to sample a posterior distribution of trees rather than making a direct estimation of the most likely tree as in ML. The ability to use more complex substitution models than those developed for ML is a major advantage to BI, but this comes with a tradeoff as BI is often prohibitively time consuming for phylogenomic datasets.

RAxML (Stamatakis 2014), a ML approach, and PhyloBayes (Lartillot et al. 2013), a BI approach, are by far the most commonly used algorithms for phylogenomic inference (Dunn et al. 2008; Hejnol et al. 2009; Philippe et al. 2009; Kocot et al. 2011; Struck et al. 2011; Nosenko et al. 2013; Ryan et al. 2013; Telford et al. 2013; Bond et al. 2014; Moroz et al. 2014; Struck et al. 2014). RAxML is utilized for its speed, whereas PhyloBayes is used because it is the only program that implements the CAT model, an infinite mixture model that does not assume site homogeneity (Lartillot and Philippe 2005). As a site heterogeneous model, the CAT model has many attractive theoretical properties, but several studies have found it to be too computationally intensive for some phylogenomic datasets (e.g., Nesnidal et al. 2010, Ryan et al. 2013; Moroz et al. 2014).

Generally, Bayesian analyses using the CAT model, which has been shown to suppress LBA artifacts (Latrillot et al. 2007), have recovered sponges as sister to all remaining animals (Philippe et al. 2009, 2011; Pick et al. 2010; Nosenko et al. 2013, but see Borowiec et al. 2015; Whelan et al. 2015). A sister relationship between cnidarians and ctenophores has only been recovered in analyses that used the CAT model on datasets dominated by ribosomal protein genes (Philippe et al. 2009; Nosenko et al. 2013), and Whelan et al. (2015) showed ribosomal protein genes, rather than the choice of model, was responsible for this inferred relationship. However, Ryan et al. (2013) also recovered support for ctenophores sister to sponges in some of their analyses that employed the CAT model. ML analyses have generally recovered ctenophores as sister to all other animals (Dunn et al. 2008; Hejnol et al. 2009; Ryan et al. 2013; Moroz et al. 2014; Boroweic et al. 2015; Whelan et al. 2015). Differences between topologies recovered with ML and BI may result from differences in the underlying substitution model. However, neither method has been consistent in their placement of ctenophores, which suggests that topological differences among phylogenies inferred with ML or BI are more complicated than model misspecification or LBA. In particular, how priors and MCMC chain-length affect phylogenetic inference using the CAT model has not been as thoroughly explored as has BI using different models (e.g., those implemented in MrBayes; Ronquist et al. 2012) and smaller datasets (Brown et al. 2010; Ekman and Blaalid 2011; Rannala et al. 2012).

Phylogenomic pipelines

Misapplied methods or poorly conceived bioinformatic pipelines can also introduce error into phylogenomic analyses. Bioinformatic pipelines for phylogenomics encompass procedures necessary for obtaining multi-gene alignments for phylogenetics from transcriptomic and genomic datasets. Failure to accurately identify orthologs, correctly align sequences, and/or remove problematic taxa or genes (e.g., those that may cause LBA) can introduce error into phylogenetic inference. Therefore, understanding steps and considerations necessary for bioinformatics in phylogenomics is essential for assessing the quality of inferred metazoan relationships.

Many bioinformatic procedures for phylogenomics are custom pipelines that piece together different tools to automate analyses that ultimately result in phylogenetic hypotheses, but broadly applicable pipelines have been published (e.g., AGALMA, Dunn et al. 2013; Osiris, Oakley et al. 2014; phylogenomic_dataset_construction, Yang and Smith 2014). Details vary from study to study, but almost every pipeline follows the same basic processes (Fig. 2). First raw data are either sequenced or retrieved from public databases and assembled (see reviews by Martin and Wang 2011; El-Metwally et al. 2013). Once assembled from raw nucleotide data,

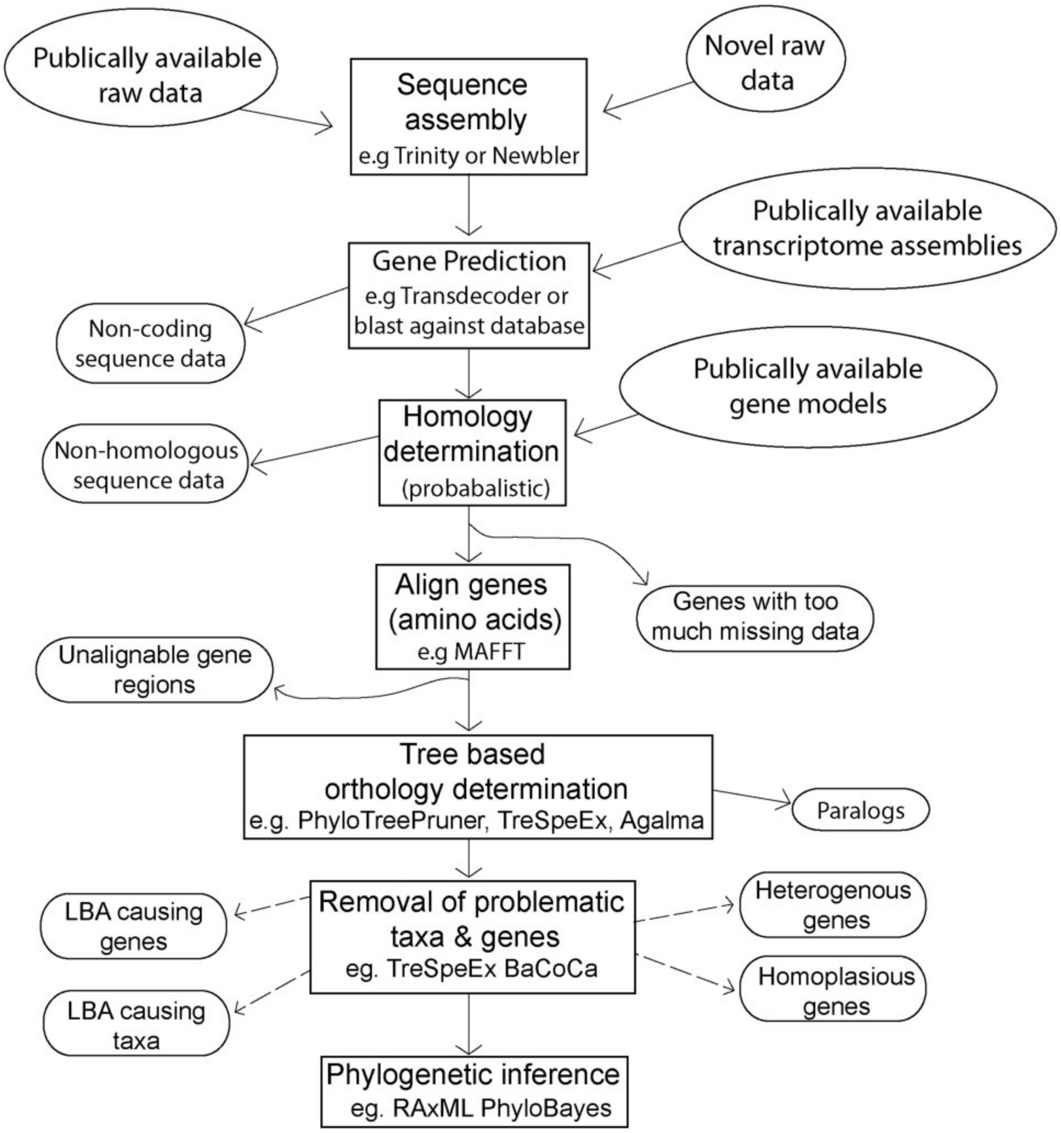
Publically available raw data
Sequence assembly
e.g Trinity or Newbler
Novel raw data
Gene Prediction
e.g Transdecoder or blast against database
Publically available transcriptome assemblies
Non-coding sequence data
Homology determination
(probabalistic)
Publically available gene models
Non-homologous sequence data
Align genes (amino acids)
e.g MAFFT
Genes with too much missing data
Unalignable gene regions
Tree based orthology determination
e.g. PhyloTreePruner, TreSpeEx, Agalma
Paralogs
Removal of problematic taxa & genes
eg. TreSpeEx BaCoCa
LBA causing genes
LBA causing taxa
Heterogenous genes
Homoplasious genes
Phylogenetic inference
eg. RAxML PhyloBayes

Fig. 2 Flow chart of a typical phylogenomic bioinformatics pipeline. Boxes represent procedures, ovals represent input data, curved boxes represent output data, and dotted lines represent optional steps that could be followed in future studies.

genes and open reading frames can be predicted with a variety of tools, including TransDecoder (transdecoder.github.io) or blast searches against well-curated databases (e.g., UniProt Consortium 2015) as in AGALMA (Dunn et al. 2013). For taxa with well-assembled genomes available, predictions of translated genes are often available from public databases.

Once genes are identified, sequences related by speciation (i.e., orthology) and not by duplication (i.e., paralogy) or lateral transfer of genes must be distinguished as the inclusion of even a limited number of paralogs can affect phylogenetic inference (Struck 2013). A number of non-tree-based methods exist for inferring which sequences from each taxon are homologous. A relatively simple, but time consuming, approach is all-by-all blast searches followed by Markov clustering (Enright et al. 2002), and a few software utilities are available for performing this type of clustering (Li et al. 2003; Lechner et al. 2011; Dunn et al. 2013; Yang and Smith 2014). As the number of taxa increase, however, all-versus-all blast searches can become computationally burdensome as every sequence must be blasted against every other sequence. An alternative approach is to use a set of ‘‘core-orthologs’’ and retrieve homologous sequences from each taxon with HaMSTR (Ebersberger et al. 2009). No matter the approach, the final result of initial determination of homology is a set of genes for phylogenetic analyses. Oftentimes, not every taxon will have a sequence for each gene because many assemblies of transcriptomes are incomplete. Therefore, genes with sequences from too few taxa are often discarded to minimize missing data. The cut-off for too few taxa per gene is often arbitrary, and many studies on metazoan relationships have generated two or more datasets with different amounts of missing data to explore how phylogenetic inferences may be influenced by missing data (e.g., Hejnol et al. 2009; Ryan et al. 2013; Moroz et al. 2014) (Table 1). The obvious trade-off to removing genes with too few taxa is that the overall amount of data is decreased (Roure et al. 2013).

After a set of genes has been assembled, each must be aligned. Two popular alignment algorithms are MAFFT (Katoh and Standley 2013; used by Ryan et al. 2013; Moroz et al. 2014; Whelan et al. 2015) and MUSCLE (Edgar 2004; used by Hejnol et al. 2009; Nosenko et al. 2013; Borowiec et al. 2015). Fast evolving sites, or sites with many indels, can be difficult or impossible to align accurately, especially at deep time-scales. Therefore, masking of alignments should be performed to remove these sites, which may introduce error, from phylogenetic inference. Programs such as ALISCORE (Misof and Misof 2009), TrimAl (Capella-Gutiérrez et al. 2009), REAP (Hartmann and Vision 2008), and GBLOCKS (Castresana 2000) are commonly used for trimming these unalignable regions (e.g., all studies in Table 1), which improves signal-to-noise ratios in empirical alignments (Kück et al. 2010).

The above methods of determining homology are not tree-based and likely contain paralogs. Therefore, additional sequence-filtering is required. Once genes are aligned, single gene trees can be generated and tree-based orthology-assignment and paralogy-pruning can be performed. Both FastTree-MP (Price et al. 2010) and RAxML (Stamatakis 2014) have been employed to generate single gene trees with FastTree-MP being computationally fast but less robust than RAxML. Gene trees are then used by programs such as PhyloTreePruner (Kocot et al. 2013), AGALMA (Dunn et al. 2013), or TreSpEx (Struck 2014) to identify and remove paralogs. Whereas PhyloTreePruner and AGALMA are purely tree-based, TreSpEx has an added blast search step to aid in identifying putative paralogs, which may offer more accuracy for identifying paralogs. This blast step confirms sequences as paralogous when different blast results are returned for the paralog and other sequences (i.e., orthologs). Notably, tree-based approaches can also help to remove exogenous contamination of sequences, which is a potentially important, but rarely considered, problem.

Once a set of orthologous genes has been identified, genes and taxa can be further screened to identify those that may cause systematic error. Programs such as TreSpEx (Struck 2014) and BaCoCa (Kück and Struck 2014) can be utilized to objectively identify genes and taxa that may cause LBA artifacts and compositionally heterogeneous genes that could violate assumptions of the model. Such *a priori* identification of problematic genes is rarely done (but see Boroweic et al. 2015; Golombek et al. 2015; Whelan et al. 2015), but some studies (Dunn et al. 2008; Hejnol et al. 2009; Moroz et al. 2014) have identified problematic taxa *a posteriori* by using leaf-stability indices (Thorley and Wilkinson 1999) and the program Phyutility (Smith and Dunn 2008). Furthermore, Philippe et al. (2011) and Nosenko et al. (2013) have emphasized using datasets with relatively low levels of saturation, and future studies may benefit from using a wide suite of genes with low mutational saturation. By focusing on using bioinformatics to identify orthologs and objectively remove different types of potential systematic error before phylogenetic inference (Fig. 2) future inference of metazoan relationships should more robustly resolve the phylogeny than have recent studies.

Conclusions

High-throughput sequencing has ushered in a new era of phylogenetics, but this has been accompanied with new challenges to accurate inference of trees. Many of these challenges are computational in nature, but traditional considerations such as taxon-sampling are still important. Instead of making blanket statements about systematic error influencing recent results, we argue that efforts should focus on assembling more complete datasets with greater sampling of taxa and characters, fewer missing data, and less potential causes of error (e.g., saturated genes or paralogs). Moreover, we must critically evaluate competing hypotheses with equal rigor and not show preferences to hypotheses just because they are traditional or commonly accepted. Overcoming the challenges to inferring the earliest splits in the animal tree of life appears within reach, and a consensus viewpoint of early animal evolution will likely materialize in the coming years.

Acknowledgments

The authors thank Leonid Moroz for organizing the symposium “Origins of Neurons and Parallel Evolution of Nervous Systems: The Dawn of Neuronal Organization” at the 2015 annual SICB meeting. This is Molette Biology Laboratory contribution 38 and Auburn University Marine Biology contribution 130.

Funding

This work was supported by the US National Aeronautics and Space Administration [grant number NASA-NNX13AJ31G].

References

Ax P. 1996. Multicellular animals: a new approach to the phylogenetic order in nature. Berlin: Springer Verlag.

Bergsten J. 2005. A review of long-branch attraction. Cladistics 21:163–93.

Bond JE, Garrison NL, Hamilton CA, Godwin RL, Hedin M, Agnarsson I. 2014. Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. Curr Biol 24:1765–71.

Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. 2015. Dissecting phylogenetic signal and accounting for bias in whole-genome data sets: a case study of the Metazoa. bioRxiv Published online January 16, 2015. http://dx.doi.org/10.1101/013946.

Brown JM, Hedtke SM, Lemmon AR, Lemmon EM. 2010. When trees grow too long: Investigating the causes of highly inaccurate Bayesian branch-length estimates. Syst Biol 59:145–61.

Cannon JT, Kocot KM, Waits DS, Weese DA, Swalla BJ, Santos SR, Halanych KM. 2014. Phylogenomic resolution of the hemichordate and echinoderm clade. Curr Biol 24:2827–32.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–3.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analyses. Mol Biol Evol 17:540–52.

Cavalier-Smith T, Fiore-Donno AM, Chao E, Kudryavtsev A, Berney C, Snell EA, Lewis R. 2014. Multigene phylogeny resolves deep branching of Amoebozoa. Mol Phylogenet Evol 81:71–85.

Chan CX, Ragan MA. 2013. Next-generation phylogenomics. Biol Direct 8:3.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–9.

Dunn CW, Giribet G, Edgecombe GD, Hejnol A. 2014. Animal phylogeny and its evolutionary implications. Annu Rev Ecol Evol Syst 45:371–95.

Dunn CW, Howison M, Zapata F. 2013. AGALMA: An automated phylogenomics workflow. BMC Bioinformatics 14:330.

Dutilh BE, van Noort V, van der Heijden RTJM, Boekhour T, Snel B, Huynen MA. 2007. Assessment of phylogenomic and orthology approaches for phylogenetic ineference. Bioinformatics 23:815–24.

Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: Profile hidden markov model based search for orthologs in ESTs. BMC Evol Biol 9:157.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–7.

Eitel M, Osigus H-J, deSalle R, Schierwater B. 2013. Global diversity of the placozoa. PLoS One 8:e57131.

Eitel M, Schierwater B. 2010. The phylogeography of the Placozoa suggests a taxon rich phylum in tropical and subtropical waters. Mol Ecol 19:2315–27.

Ekman S, Blaalid R. 2011. The devil in the details: Interactions between the branch-length prior and likelihood model affect node support and branch lengths in the phylogeny of the Psoraceae. Syst Biol 60:541–61.

El-Metwally S, Hamza T, Zakaria M, Helmy M. 2013. Next-generation sequence assembly: Four stages of data processing and computational challenges. PLoS Comp Biol 9:e1003345.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–84.

Felsenstein J. 1973. Maximum likelihood and minimum-steps for methods for estimating evolutioanry trees from data on discrete characters. Syst Zool 22:240–9.

Felsenstein J. 1978. Cases in which parsimony and compatability methods will be positively misleading. Syst Zool 27:401–10.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17:368–76.

Felsenstein J. 1985. Phylogenies and the comparative method. Am Zool 125:1–15.

Fernández R, Laumer CE, Vahtera V, Libro S, Kaluziak S, Sharma PP, Pérez-Porro AR, Edgecombe GD, Giribet G. 2014. Evaluating topological conflict in centipede phylogeny using transcriptome data sets. Mol Biol Evol 31:1500–13.

Gatsey J, DeSalle R, Wahlberg N. 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Syst Biol 56:355–63.

Gilks WR, Richardson S, Spiegelhalter DJ. 1996. Markov Chain Monte Carlo in practiceLondon: Chapman & Hall.

Golombek A, Tobergte S, Struck TH. 2015. Elucidating the phylogenetic position of Gnathostomulida and first mitochondrial genomes of Gnathostomulida, Gastrotricha and Polycladida (Platyhelminthes). Mol Phylogenet Evol 86:49–63.

Halanych KM. 2015. The ctenophore lineage is older than sponges? That cannot be right! Or can it? J Exp Biol 218:592–7.

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB, Roger AJ. 2009. Phylogenomic analyses support monophyly of Excavata and resolve relationships among eukaryotic ‘‘supergroups’’. Proc Natl Acad Sci USA 106:3859–64.

Hartmann S, Vision TJ. 2008. Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol Biol 8:95.

Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol 46:239–57.

Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic models. Proc R Soc Lond B Biol Sci 276:4261–70.

Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. Syst Zool 38:297–309.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–4.

Janussen D, Reiswig HM. 2009. Hexactinellida (Porifera) from the ANDEEP III expedition to the Weddel Sea, Antarctica. Zootaxa 2136:1–20.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–31.

Jékely G, Paps J, Nielsen C. 2015. The phylogenetic position of ctenophores and the origin(s) of nervous systems. EvoDevo 6:1.

Johnson BR, Borowiec ML, Chiu JC, Lee EK, Atallah J, Ward PS. 2013. Phylogenomics resolves evolutionary relationships among ants, bees, and wasps. Curr Biol 23:2565.

Kaiser D. 2001. Building a multicellular organism. Annu Rev Genet 35:103–23.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30:772–80.

King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, et al. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature 451:783–8.

Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B, et al. 2011. Phylogenomics reveals deep molluscan relationships. Nature 477:452–6.

Kocot KM, Citarella MR, Moroz LL, Halanych KM. 2013. PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. Evol Bioinform Online 9:429–35.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309–38.

Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Wägele JW, Misof B. 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Front Zool 7:10.

Kück P, Struck TH. 2014. BaCoCa—a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. Mol Phylogenet Evol 70:94–8.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artifacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 7:S4.

Lartillot N, Philippe H. 2005. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol 62:611–5.

Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: Detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics 12:124.

Lemer S, Kawauchi GY, Andrade SCS, González VL, Boyle MJ, Giribet G. 2015. Re-evaluating the phylogeny of Sipuncula through transcriptomics. Mol Phylogenet Evol 83:174–83.

Lemmon AR, Brown JM, Stanger-Hall K, Moriarty-Lemmon E. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst Biol 58:130–45.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–89.

Mah JL, Christensen-Dalsgaard KK, Leys SP. 2014. Choanoflagellate and choanocyte collar-flagellar systems and the assumption of homology. Evol Dev 16:25–37.

Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. Nat Rev Genet 12:671–82.

Misof B, Misof K. 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: A more objective means of data exclusion. Syst Biol 58:21–34.

Moroz LL. 2009. On the independent origins of complex brains and neurons. Brain Behav Evol 74:177–90.

Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov E, Buckley KM, et al. 2014. The ctenophore genome and the evolutionary origins of neural systems. Nature 510:109–14.

Morrison DA, Ellis JT. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa. Mol Biol Evol 14:428–41.

Nagarajan N, Pop M. 2013. Sequence assembly demystified. Nat Rev Genet 14:157–67.

Nekrutenko A, Taylor J. 2012. Next-generation sequencing data interpretation: Enhancing reproducibility and accesibility. Nat Rev Genet 13:667–72.

Nesnidal MP, Helmkampf M, Bruchhaus I, Hausdorf B. 2010. Compositional heterogeneity and phylogenomic inference of metazoan relationships. Mol Biol Evol 27:2095–104.

Nesnidal MP, Helmkampf M, Bruchhaus I, El-Matbouli M, Hausdorf B. 2013. Agent of whirling disease meets oprhan worm: Phylogenomic analyses firmly place Myxozoa in Cnidaria. PLOS One 8:e54576.

Nielsen C. 2008. Six major steps in animal evolution: Are we derived from sponge larvae? Evol Dev 10:241–57.

Nosenko T, Schreiber F, Adamska M, Adamski M, Eitel M, Hammel J, Maldonado M, Müller WEG, Nickel M, Schierwater B, et al. 2013. Deep metazoan phylogeny: When different genes tell different stories. Mol Phylogenet Evol 67:223–33.

Oakley TH, Alexandrou MA, Ngo R, Pankey MS, Churchill CKC, Chen W, Lopker KB. 2014. Osiris: Accessible and reproducible phylogenetic and phylogenomic analyses with the Galaxy workflow management system. BMC Bioinformatics 15:239.

Ogden TH, Rosenberg MS. 2005. Multiple sequence alignment accuracy and phylogenetic inference. Syst Biol 55:314–28.

Osigus H-J, Eitel M, Bernt M, Donath A, Schierwater B. 2013. Mitogenomics at the base of metazoa. Mol Phylogenet Evol 69:339–51.

Pearse VB, Voigt O. 2007. Field biology of placozoans (*Trichoplax*): Distribution, diversity, biotic interactions. Integr Comp Biol 47:677–92.

Pearson WR, Sierk ML. 2005. The limits of protein sequence comparision? Curr Opin Struct Biol 15:254–60.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide 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, Houliston E, Quéinnec E, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr Biol 19:706–12.

Philippe H, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroboratae the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol 22:1246–53.

Pick KS, Philippe H, Schreiber F, Erpenbeck D, Jackson DJ, Wrede P, Wiens M, Alié A, Morgenstern B, Manuel M, et al. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol Biol Evol 27:1983–7.

Pickett KM, Randle CP. 2005. Strange Bayes indeed: Uniform topological priors imply non-uniform clade priors. Mol Phylogenet Evol 34:203–11.

Podar M, Haddock SHD, Sogin ML, Harbison GR. 2001. A molecular phylogenetic framework for the phylum Ctenophora using 18S rRNA genes. Mol Phylogenet Evol 21:218–30.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—Approximately maximum-likelihood trees for large alignments. PLOS One 5:e9490.

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J Mol Evol 43:304–11.

Rannala B, Zhu T, Yang Z. 2012. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. Mol Biol Evol 29:325–35.

Raup DM. 1986. Biological extinction in earth history. Science 231:1528–33.

Regier JC, Shultz JW, Ganley ARD, Hussey A, Shi D, Ball B, Zwick A, Stajich JE, Cummings MP, Martin JW, et al. 2008. Resolving arthropod phylogeny: Exploring phylogenetic singal withn 41 kb of protein-coding nuclear gene sequence. Syst Biol 57:920–38.

Riesgo A, Farrar N, Windsor PJ, Giribet G, Leys SP. 2014. The analysis of eight transcriptomes from all Poriferan classes reveals surprising genetic complexity in sponges. Mol Biol Evol 31:1102–20.

Ronquist F, Teslenko M, van der Marl P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayes inference and model choice across a large model space. Syst Biol 61:539–42.

Rota-Stabellia O, Telford MJ. 2008. A multi criterion appraoch for the selection of optimal outgroups in phylogeny: Recovering some support for Mandibulata over Myriochelata using mitogenomics. Mol Phylogenet Evol 48:103–11.

Roure B, Baurain D, Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic datasets. Mol Biol Evol 30:197–214.

Ryan JF, Pang K, Schnitzler CE, Nguyen A-D, Moreland RT, Simmons DK, Koch BJ, Francis WR, Havlak P, Smith SA, et al. 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science 342:1242592.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–33.

Sharma PP, Kaluziak ST, Pérez-Porro AR, González VL, Hormiga G, Wheeler WC, Giribet G. 2014. Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal. Mol Biol Evol 31:2963–84.

Simion P, Bekkouche N, Jager M, Quéinnec E, Manuel M. 2015. Exploring the potential of small RNA subunit and ITS sequences for resolving the phylogenetic relationships within the phylum Ctenophora. Zoology 118:102–14.

Simpson GG. 1952. How many species? Evolution 6:342.

Sperling EA, Pisani D, Peterson KJ. 2007. Poriferan paraphyly and its implications for Precambrian paleobiology. Geol Soc Lond Spec Publ 286:355–68.

Smith SA, Dunn CW. 2008. Phyutility: A phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–6.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–3.

Straub SCK, Moore MJ, Soltis PS, Soltis DE, Liston A, Livshultz T. 2014. Phylogenetic signal detection from an ancient rapid radiation: Effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. Mol Phylogenet Evol 80:169–85.

Struck TH. 2013. The impact of paralogy on phylogenomic studies—A case study on Annelid Relationships. PLoS One 8:e62892.

Struck TH. 2014. TreSpEx—detection of misleading signal in phylogenetic reconstructions based on tree information. Evol Bioinform Online 10:51–67.

Struck TH, Paul C, Hill N, Hartmann S, Hosel C, Kube M, Lieb B, Meyer A, Tiedemann R, Purschke G, et al. 2011. Phylogenomic analyses unravel annelid evolution. Nature 471:95–8.

Struck TH, Wey-Fabrizius AR, Golombek A, Hering L, Weigert A, Bleidorn C, Klebow S, Iakovenko N, Hausdorf B, Petersen M, et al. 2014. Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestrsy of Spiralia. Mol Biol Evol 31:1833–49.

Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E, Carr M, Kerner P, Vervoort M, Sánchez-Pons N, et al. 2013. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. Nat Commun 4:2325.

Telford MJ, Lowe CJ, Cameron CB, Ortega-Martinex O, Aronowicz J, Oliveri P, Copley RR. 2013. Phylogenomic analysis of echinoderm class relationships supports Asterozoa. Proc R Soc Lond B Biol Sci 281:20140479.

Thorley J, Wilkinson M. 1999. Testing the phylogenetic stability of early tetrapods. J Theor Biol 200:343–4.

UniProt Consortium. 2015. UniProt: A hub for protein information. Nucleic Acids Res 43:D204–12.

van Djik EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. Trends Genet 30:418–26.

Weisburg WG, Giovannoni SJ, Woese CR. 1989. The Deinococcus-Thermus phylum and the effect of rRNA composition on phylogenetic tree construction. Syst Appl Microbiol 11:128–34.

Whelan NV, Kocot KM, Moroz LL, Halanych KM. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Scad Sci USA. doi:10.1073/pnas.1503453112.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci USA 111:E4859–68.

Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. Mol Biol Evol 31:3081–92.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–72.

Yang Z, Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst Biol 54:455–70.

Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51:588–98.