OPEN ACCESS Freely available online

Perspective

Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough

Hervé Philippe[1]*, Henner Brinkmann[1], Dennis V. Lavrov[2], D. Timothy J. Littlewood[3], Michael Manuel[4], Gert Wörheide[5,6], Denis Baurain[7]

1 Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, Québec, Canada, **2** Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, United States of America, **3** Department of Zoology, The Natural History Museum, London, United Kingdom, **4** Université Paris 6, UMR 7138 "Systématique, Adaptation, Evolution" UPMC CNRS IRD MHNH, Paris, France, **5** Department of Earth and Environmental Sciences, Ludwig-Maximilians-Universität München, München, Germany, **6** GeoBio-Center, Ludwig-Maximilians-Universität München, München, Germany, **7** Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège, Liège, Belgium

In the quest to reconstruct the Tree of Life, researchers have increasingly turned to phylogenomics, the inference of phylogenetic relationships using genome-scale data (Box 1). Mesmerized by the sustained increase in sequencing throughput, many phylogeneticists entertained the hope that the incongruence frequently observed in studies using single or a few genes [1] would come to an end with the generation of large multigene datasets. Yet, as so often happens, reality has turned out to be far more complex, as three recent large-scale analyses, one published in *PLoS Biology* [2–4], make clear. The studies, which deal with the early diversification of animals, produced highly incongruent (Box 2) findings despite the use of considerable sequence data (see Figure 1). Clearly, merely adding more sequences is not enough to resolve the inconsistencies.

Here, taking these three studies as a case in point, we discuss pitfalls that the simple addition of sequences cannot avoid, and show how the observed incongruence can be largely overcome and how improved bioinformatics methods can help reveal the full potential of phylogenomics.

Hurdles to Phylogenomics

Two factors contribute significantly to the difficulty of reconstructing the correct phylogenetic tree for a set of sequences. First, if speciation events are closely spaced in time, the amount of phylogenetic signal is often small, leading to short internal tree branches that are difficult to resolve [5,6]. Second, if the events of interest are ancient, terminal branches tend to be long and replete with multiple substitutions occurring at the same position (i.e., homoplasy). In the extreme case, insufficient signal may remain for very deep divergences to be resolved even when using very long gene sequences [7]—but this issue is outside the scope of the present contribution. Depending on the accuracy of the model of sequence evolution, multiple substitutions can go undetected or be wrongly inferred. In both situations spurious phylogenetic signals are generated; these constitute the major part of what we collectively term non-phylogenetic signal. The best known example of the misleading effect of non-phylogenetic signal is the long branch attraction (LBA) artifact [8]: when two (or more) lineages have much longer branches than the others, they tend to group together irrespective of their true relationships. Notably, the outgroup is a natural source of long branches that may attract fast-evolving (hence long branched) species of the ingroup. When this happens, attracted branches artifactually emerge too deeply in the tree [9].

Inferring phylogenies in difficult cases is akin to finding a needle (phylogenetic signal) in a haystack. Under the oversimplified assumption of an absence of non-phylogenetic signal, one can compute that the resolving power would increase from approximately 15 million years when using small subunit ribosomal RNA alone to less than 1 million years when using more than 50 genes [10]. At such levels of resolution, incomplete lineage sorting (i.e., the retention of ancestral polymorphisms over successive speciation events) should be taken into account as a potential source of phylogenetic error [11]. Nonetheless, even if conflicting gene genealogies were not an issue, throwing additional gene sequences at a difficult phylogenetic question does not necessarily solve the problem—the size of the needle is indeed increased, but so too is the size of the haystack. It follows that non-phylogenetic signal may become dominant and yield incongruent, yet statistically highly supported, phylogenomic trees [12].

How to Prevent Deleterious Effects of Non-Phylogenetic Signal

Non-phylogenetic signal has multiple and disparate sources [13]. When multiple genes are concatenated and analyzed with standard methods (but see [14]), non-phylogenetic signal is caused by the inclusion of sequences that deviate from the true species phylogeny or by the inability of our methods to correctly handle multiple substitutions. In practice, it mainly stems from (i) the incorrect identification of orthologs, (ii) erroneous alignments, or (iii) the incorrect reconstruction of multiple substitutions occurring at a given position, the last owing to model violations in probabilistic methods

The Perspective section provides experts with a forum to comment on topical or controversial issues of broad interest.

Citation: Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, et al. (2011) Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biol 9(3): e1000602. doi:10.1371/journal.pbio.1000602

Academic Editor: David Penny, Massey University, New Zealand

Published March 15, 2011

Funding: The work was funded by NSERC (www.nserc-crsng.gc.ca), CRC (www.chairs-chaires.gc.ca), Agence Nationale de la Recherche (http://www.agence-nationale-recherche.fr/), ARC Biomod (www.cfwb.be), and DFG (http://www.dfg.de/en/index.jsp). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: BS, bootstrap support; EST, expressed sequence tag; LBA, long branch attraction

* E-mail: herve.philippe@umontreal.ca

Box 1. From Phylogenetics to Phylogenomics

Phylogenetics, the determination of evolutionary relationships among organisms, is central to our understanding of the evolution of life. For instance, the three phylogenies of Figure 1 entail profoundly different interpretations about the complexity of the common ancestor of all animals. Important body plan characters (e.g., neurosensory and digestive systems and muscle cells) are found in cnidarians, ctenophores, and bilaterians but not in sponges and placozoans. According to the phylogenies of Schierwater et al. [4] and Dunn et al. [2], the taxonomic distribution of these characters implies either (i) that the ancestral metazoan already featured these traits and that sponges (and placozoans) have secondarily lost them or (ii) that these characters were acquired several times independently by convergence (e.g., in the cnidarian + ctenophore and in the bilaterian lineages, according to the tree of Figure 1A). In contrast, the phylogeny of Philippe et al. [3] is more congruent with morphological characters and compatible with a simple metazoan ancestor and a later emergence of these characters only once, in the lineage leading to the common ancestor of coelenterates (cnidarians+ctenophores) and bilaterians.

Phylogenies are generally depicted as trees (which are non-reticulated graphs, as in Figure 1) because vertical evolution is undisputedly the primary mechanism of inheritance for genetic material. However, the existence of horizontal transmission (e.g., hybridization of closely related taxa, organelle acquisition through endosymbiosis and horizontal gene transfer) makes phylogenetic trees only pragmatic approximations, which will probably be replaced by phylogenetic networks in the long term (particularly for unicellular organisms).

Recently, phylogenomics, the use of genomic data to infer evolutionary relationships, has emerged as a new domain of phylogenetics. The main strength of phylogenomics is the drastic reduction in random (or sampling) error brought by the use of large (multigene) datasets. Numerous approaches can be used to take advantage of genomic data (for review see [49]). Briefly, new methods based on oligonucleotide content, gene content, or intron positions look promising (as shown by their ability to yield reasonable trees) but require additional theoretical developments to achieve their full potential. That is why the two most popular phylogenomic approaches are simple extensions of the standard phylogenetics methods applied to single-gene datasets. The first, known as the "supermatrix" (or superalignment), consists in concatenating numerous orthologous genes into a single supergene, which is analyzed using standard methods (or slightly modified methods such as separate models allowing for multiple sets of branch lengths [50]). The second, "supertree," approach takes the opposite path by first inferring a tree for each gene in the dataset and then combining these individual trees into a single supertree. The supermatrix approach is the most commonly used, in agreement with the handful of studies suggesting that it offers greater accuracy than the supertree [13,51], though this remains to be formally demonstrated.

(i.e., Bayesian inference and maximum likelihood). Although all three aspects have received considerable attention from theoreticians, and despite the availability of numerous bioinformatics tools [15–17], there is still no magic bullet. That is why classic phylogenetics involves numerous refinements and controls, which are difficult, but not impossible, to apply at a phylogenomic scale.

Non-phylogenetic signal can be reduced by improving (i) the quality of primary alignments through selection of the orthologous genes that are least subject to saturation and (ii) the detection of multiple substitutions, which is best achieved by using both a large number of species and the most realistic model of sequence evolution. In the following, we show that both improvements are required at the same time to address the difficult question of the relationships among major animal groups, i.e., sponges, placozoans, ctenophores, cnidarians, and bilaterians. Reanalysis of the underlying data indicates that failure to apply one or more of the strategies intended to decrease non-phylogenetic signal is what caused the incongruent, though strongly supported, results that were recently observed [2–4].

Issues at the Level of Sequence Alignments

Selection of unambiguously orthologous genes [18] is usually achieved by targeting single-copy genes (e.g., mitochondrial genes) or pre-selected genes (e.g., ribosomal RNAs and proteins), or through automatic clustering methods. None of these options are without problems. Both manual and automatic methods [19–22] heavily rely on BLAST similarity scores, which are known to be a poor estimator of the true evolutionary distance [23]. Given the limitations of existing methods of orthology detection (Box 3), careful phylogenetic analysis of each alignment is important to achieve maximal accuracy. However, this manual step is difficult and subjective. That is why it is preferable to also verify orthology a posteriori. One possibility is to assess whether branches receiving high statistical support from every single gene tree are congruent with the species tree [18]. Though the latter is unknown, the phylogeny obtained by the concatenation of numerous genes constitutes a reasonable approximation. Hence, Philippe et al. [3] looked at every supported branch (bootstrap support [BS]$\geq$70%) from single-gene trees that were incongruent with the concatenated tree to assess the orthology of their preselected genes. Only 6.5% of the branches were incongruent, and almost all conflicts were best explained by reconstruction errors affecting single-gene trees [3]. According to this semi-automated approach, the 128 genes used in [3] can be provisionally considered as orthologous and suitable for phylogenetic analysis. In contrast, when applied to the datasets of Schierwater et al. [4] and Dunn et al. [2], the very same approach identifies several instances of incongruence between single-gene and concatenated trees (mainly apparent horizontal gene transfers that are in fact more likely due to contaminations, or deep unrecognized paralogy; see Text S1 and Figures S1, S2, S3, S4, S5, S6, S7, S8, S9).

This as well as the discovery of other important issues (see Table S1) prompted us to reassess and reanalyze the dataset of Schierwater et al. [4]. The revised phylogeny we generate (Figure 2B) differs from the original one (Figure 2A) in the deep animal relationships: the strong support for a sister-group relationship between Bilateria and a group composed of placozoans, sponges (Porifera), ctenophores, and cnidarians [4] has vanished, and sponges are now recovered as the sister group of all other Metazoa. Strikingly, this part of the revised tree (Figure 2B) suffers from a lack of statistical support (all BS$<$50% except for the monophyly of cnidarians). The simplest explanation for these results (Figure 2B) is that the genuine phylogenetic signal for non-bilaterian animal relationships is scarce, as reported in all previous studies (e.g., [24–28]). The possible inclusion of non-orthologous sequences (see Figures S1,

Box 2. Glossary

Homology/orthology/paralogy/xenology: Genes that derive from a common ancestor are termed homologs. Two homologous genes are orthologous if they diverged through a speciation event. In contrast, paralogs originate by duplication of a single gene within a given lineage, whereas xenologs result from the horizontal transfer of a gene from a donor species to a receiver species (which might eventually get its original copy replaced by the xenolog).

Homoplasy/convergence: Spurious similarity due to convergence or reversion and not to common ancestry is termed homoplasy. Convergence describes the independent acquisition by separate evolutionary lineages of the same nucleotide (or amino acid) at a given position. This is a direct consequence of multiple substitutions.

Incomplete lineage sorting: The transient retention of ancestral polymorphisms across speciation events. Speciations compressed in time and large reproductive populations both increase the likelihood of this phenomenon. Considering three lineages having rapidly diverged, by chance some sequence positions will be shared between one pair, while others will be shared between another pair, and yet others between the third possible pair, hence blurring the phylogenetic signal on the corresponding branches.

Incongruence: Two (or more) phylogenetic trees are said to be incongruent when they exhibit conflicting branching orders (i.e., topologies) and cannot be superimposed. This implies that at least one node (also known as a bipartition) present in one tree is not found in the other(s), where it is replaced by alternative groupings of taxa.

Model of sequence evolution: A statistical description of the process of substitution in nucleotide or amino acid sequences. Complex models better approximate the evolutionary process but at the expense of more parameters and computational time. As parameter-rich models require more data to behave properly, they have become really useful with the advent of phylogenomic datasets.

Monophyly: To be considered monophyletic, a taxonomic group must satisfy two conditions: (i) all its taxa must derive from a single ancestor and, reciprocally, (ii) all taxa deriving from this common ancestor must belong to the group.

Non-phylogenetic signal: The combination of different kinds of structured noise (e.g., undetected homoplasies) that compete with the genuine phylogenetic signal during tree reconstruction. Even if the non-phylogenetic content is partly a property of a multiple sequence alignment (notably related to its saturation level), the non-phylogenetic signal actually inferred heavily depends on the method and the model of evolution selected. In probabilistic methods, the non-phylogenetic signal mainly results from the data violating the model of sequence evolution. These violations arise because our models are inevitably oversimplified in comparison to the complexity of the natural evolutionary process. Eventually, the apparent signal analyzed will be a blend of phylogenetic and non-phylogenetic signal.

Outgroup/ingroup: Nearly all tree reconstruction methods produce unrooted trees, in which inferred relationships do not convey any information about the direction of time. To root a tree and turn it into a phylogeny, one has to include in the analysis a group of taxa that are known to be outside the group under study. This reference group is termed the outgroup, while the taxa of interest make the ingroup.

Patristic distance: The sum of the lengths of the branches that connect two nodes in a phylogenetic tree, where those nodes are typically terminal nodes representing extant taxa. It is thus an inferred distance (taking into account multiple substitutions) greater than the uncorrected distance directly computed from the number of differences observed between the two corresponding sequences in the alignment.

Phylogenetic signal/synapomorphy: The substitutions occurring along a given branch of the evolutionary tree. The strength of the phylogenetic signal is proportional to the number of substitutions occurring along the branch. In non-probabilistic methods, the signal is encoded in synapomorphies, i.e., shared residues (nucleotides or amino acids) at aligned positions that are specific to a set of sequences derived from a common ancestor. In probabilistic methods, the amount of phylogenetic signal actually extracted from a given dataset depends on the model and is expected to increase with the fit of the model to the data (i.e., the ability of the model to explain the data).

Phylogenetic tree: A (connected acyclic) graph describing the estimated evolutionary relationships among a group of species. In molecular trees, branch lengths are proportional to the genetic distances (and hence to some extent to time) inferred from the analysis of a multiple alignment of homologous sequences (nucleotide or amino acid sequences).

Probabilistic methods: A family of tree reconstruction methods from multiple sequence alignments that are grounded in statistical theory and make use of explicit models of sequence evolution. These include maximum likelihood and Bayesian inference approaches and are known to be the most accurate but also the most computationally demanding.

Saturation: When sequences in a multiple alignment have undergone so many multiple substitutions that apparent distances largely underestimate the real genetic distances, the alignment is said to be saturated. Phylogenetic inference works best with datasets that are only slightly saturated. Owing to their reduced state space (four possible bases), nucleotide sequences saturate more rapidly than protein sequences (20 possible amino acids).

Site-homogeneous/site-heterogeneous models: Most models of sequence evolution assume that the same evolutionary process takes place at every position (or site) of an alignment. With such models, only the evolutionary rate can be modeled as heterogeneous across sites, usually through a gamma distribution of rates. However, selective constraints are known to be quite heterogeneous across positions, hence seriously violating the hypotheses of site-homogeneous models. On the other hand, site-heterogeneous models assume that the evolutionary process varies widely across sites, in particular the set of acceptable amino acids (e.g., in the CAT model). A number of studies have demonstrated that site-heterogeneous models provide a better fit to phylogenomic datasets and tend to reduce the sensitivity to tree reconstruction artifacts (e.g., LBA).

S2, S3, S4, S5, S6, S7, S8, S9) might create a strong signal that could overcome the genuine but faint phylogenetic signal, and lead to the incorrect—but strongly supported—monophyly of "diploblasts" (sponges+placozoans+ctenophores+cnidarians) that was observed in the original study (Figure 2A). Otherwise, the topology we infer from the revised alignments is similar to the published tree [4], with only three nodes differing out of 21. This demonstrates that phylogenomics is relatively robust to the possible inclusion of non-orthologous sequences when the genuine phylogenetic signal is abundant (see also [29,30]), which can be explained by the randomness of most of the introduced errors preventing the appearance of a structured misleading signal.

On the other hand, phylogenomics is sensitive to the non-phylogenetic signal that stems from the incorrect inference of multiple substitutions. By devoting a large part of their dataset to mitochondrial genomes, which are fast-evolving in Bilateria (e.g., [24,31]), Schierwater et al.'s solution unwittingly favored the emergence of Bilateria between the outgroup and a group composed of all the non-bilaterian Metazoa, because of the LBA artifact. This artifact probably also affects the phylogeny of Dunn et al. [2]; in that case, the fast-evolving ctenophores are likely attracted by the distant outgroup (see Text S1). In the phylogeny inferred from an updated version of the alignments of Dunn et al. (purged of several sequencing errors and species misidentifications—see Table S2—and completed with new sequences, thereby reducing the amount of missing data from 55% to 35%), sponges are the sister group of all other Metazoa, with the fast-evolving Ctenophora representing the sister group of Cnidaria plus Bilateria (Figure S11; see also [32]).

In summary, analyzing the revised alignments from Schierwater et al. [4] and Dunn et al. [2] with their original taxon sampling and inference methods is sufficient to eliminate all significant incongruences among the three recent phylogenomic studies (Figure 1). The variability in robustness across the tree (e.g., Figure 2) underscores the importance of clean phylogenomic datasets: whereas large amounts of phylogenetic signal usually drown out any non-phylogenetic signal, for nodes characterized by a scarce phylogenetic signal, even small amounts of non-phylogenetic signal may dominate and eventually yield incorrect results [10].

Issues at the Level of Taxon Sampling

The lack of support observed in Figures 2 and S11 contrasts with the high bootstrap values obtained by Philippe

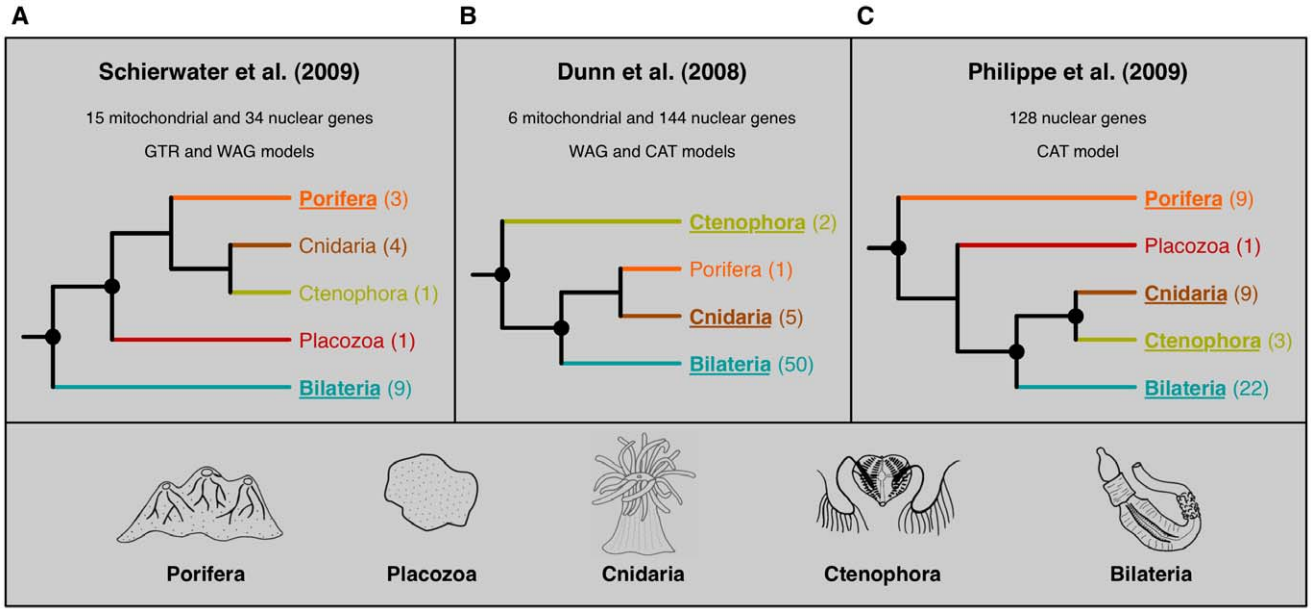

Figure 1. Simplified representation of the trees obtained in three recent phylogenomic analyses of early animal diversification. (A) Schierwater et al. [4] tree. (B) Dunn et al. [2] tree. (C) Philippe et al. [3] tree. Numbers in parentheses after taxon names indicate the number of species included in the dataset for the corresponding taxon. Bootstrap support values above 90% are indicated by a bullet (for nodes) or by underlining (for terminal taxa). It is worth mentioning that the monophyly of Porifera is not unequivocally accepted [28,46]; only the analysis of 30,000 positions with a rich taxon sampling and a complex model of evolution recovers it with significant statistical support [3]. Although such a sparse phylogenetic signal will require harnessing the full potential of phylogenomics to be confidently solved, this question is outside the scope of this study. Simplified drawings (redrawn from [74]) on the bottom illustrate the huge morphological disparity existing between the five terminal taxa. Porifera correspond to sponges; Cnidaria to sea anemones, jellyfishes, and allies; Ctenophora to comb jellies; and Bilateria to all other animals (characterized by their bilateral symmetry) except *Trichoplax* (Placozoa), which appears to be morphologically the most simply organized animal phylum.
doi:10.1371/journal.pbio.1000602.g001

Box 3. Quality Control of Phylogenomic Datasets

Despite the great progress in software development [19–22,52–54], our nine years of experience with large-scale multigene analyses [55] leads us to conclude that computer-assisted manual expertise is not yet dispensable. In particular when processing EST data, two issues are still challenging to handle by automation: (i) the non-homology of short sequence stretches due to frameshifts and point mutations and (ii) the non-orthology of one or more genes with similar sequence for some species, because of paralogy or xenology, along with taxonomic misidentifications and library contaminations (e.g., by parasites such as platyhelminthes). An important limitation of automated methods for checking single-gene alignments for orthology prior to concatenation is the limited amount of sequence information available in a single gene, which often makes current statistical analyses impractical. If the threshold used is stringent, almost every sequence will fail the test, whereas a loose threshold will lead to numerous false positives. Manual verification, through visual inspection of alignments and phylogenies, can to a large extent compensate for this lack of statistical power if a large number of species (much more than those eventually included in the final analysis) is taken into account. First, as conserved positions are clearly identified, both translational frameshifts (leading to stretches of amino acids highly different from the consensus, which are mostly found at EST extremities) and local sequencing errors (visible as unmatched amino acids at highly conserved positions) stand out. Based on manual analysis, we estimate that approximately 4,800 amino acids (0.66% of the complete alignment) were erroneous in the Dunn et al. dataset [2] because of frameshifts and local sequencing errors (including incorrect translation owing to a mistake in the specification of the genetic code for ambulacrarian mitochondria; see Table S2). Second, xenology, contaminations, and misidentification can be efficiently detected when individual alignments encompass a broad taxonomic diversity, as such diversity is much more likely to find a close relative of the donor species. For instance, in the Dunn et al. dataset [2], one acoel species, the marine flatworm *Neochildia fusca*, was contaminated by microsporidia (see Table S2). Since original alignments lacked microsporidial sequences, the contamination was overlooked and acoel sequences were simply considered as extremely divergent. Similarly, hidden paralogy is easier to detect with numerous species on hand (and with deeper sequencing of each of them), because they increase the chance of finding a species that has kept both copies. Interestingly, much more serious errors (including the use of paralogous, rather than orthologous, copies, and taxonomic misidentification; see Figures S1, S2, S3, S4, S5, S6, S7, S8, S9) were identified in the manually assembled Schierwater et al. dataset [4] than in the automatically assembled Dunn et al. dataset [2] (compare Tables S1 and S2). Manual assessment of the quality of primary data is particularly tedious and time-consuming, as well as error-prone. That is why automated approaches featuring refined statistics (e.g., hidden Markov models detecting frameshifts) are strongly needed to both speed up and improve the construction of phylogenomic datasets. Finally, it should be noted that missing data (i.e., incomplete sequences), which are on the rise in recent large-scale analyses (e.g., 55.5% of the characters in [2] and 81% in [46]), constitute an additional unpredictable issue, as they might further erode statistical power and sometimes enhance tree reconstruction artifacts [38,42] (see Text S1 and Figure S11).

et al. [3] for the monophyly of each of the Porifera (96%), Coelenterata (Cnidaria+Ctenophora, 93%), and Eumetazoa (all animals except Porifera and Placozoa, 90%) (Figure 3A). However, the number of non-bilaterian metazoan species used in [3] is larger, 22 versus 9 [2,4], which could account for the difference. Indeed, it is well known that including more species allows for a better detection of multiple substitutions [33], as it decreases the amount of non-phylogenetic signal while preserving phylogenetic signal [34]; this is why authors often mention that their results should be viewed as provisional until more taxa are considered (e.g., the position of Ctenophora in [2]). To test this hypothesis, we reduced the taxon sampling of [3] to match as closely as possible the sampling of Figure 2. Even though sequences and inference methods are exactly as in [3], the support for deep animal relationships decreases drastically (Figure 3B). While the monophyly of each of the Cnidaria (94%), Coelenterata (70%), and Demospongiae + Hexactinellida (86%) still receive some support, remaining relationships are unresolved (BS<60%); in particular, Porifera and Eumetazoa are not recovered. These results corroborate the hypothesis that the use of a limited number of species generates enough non-phylogenetic signal to swamp most of the faint genuine phylogenetic signal present in this part of the animal phylogeny (owing to short internal branches and heterogeneous rates among species).

However, taxon sampling is not simply a matter of number of species [35–37]. In particular, the inclusion of both slowly evolving species and closely related outgroups (e.g., choanoflagellates for animals; see [3] and Text S1) is often of prime importance. This point is well illustrated by a reanalysis of the original alignments of Schierwater et al. in which we eliminated the most distant outgroups. When rooting exclusively with choanoflagellates, the bootstrap support for a position of Porifera as the sister group to remaining animals rises to 80% (Figure S13). Although discarding very distant outgroups (e.g., Bacteria) undoubtedly improves accuracy, the effect of including moderately distant outgroups (e.g., Fungi) in addition to close outgroups (e.g., choanoflagellates) is more difficult to assess. Eventually, it will depend on the relative influence of introducing a very long branch (the distant outgroup) and breaking up an already existing long branch (the close outgroup). Even if further studies are needed to clarify this point, an effort to increase the taxon sampling of the close outgroup should help to resolve deep animal relationships.

Finally, phylogenomic datasets, especially when based on expressed sequence tag (EST) data, are frequently characterized by incomplete gene coverage for some taxa. Yet, there have been few attempts to determine whether missing data per se can cause errors in tree reconstruction [36,38–42] and how they may interfere with other aspects of phylogenetic inference. In particular, it is not known whether a smaller, but complete, alignment of targeted genes (e.g., selectively amplified by PCR) would yield a more accurate and robust tree than a large, but incomplete, alignment of highly expressed genes (obtained by EST sequencing). These questions can and should be better assessed in the near future.

Issues at the Level of Tree Reconstruction Methods

To further explore the idea that the paramount issue in phylogenomics pertains to the reduction of non-phylogenetic signal (more than the increase of phylogenetic signal with datasets containing more and more genes, especially in the short run), we

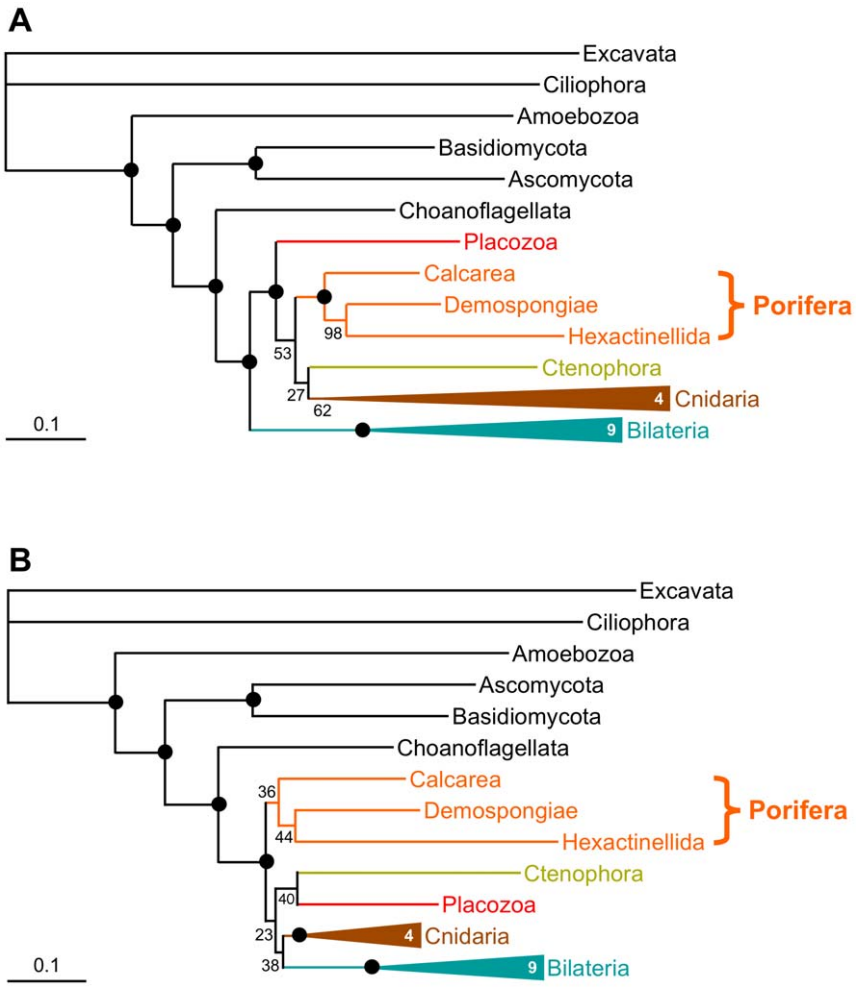

Figure 2. Analysis of the revised Schierwater et al. dataset. (A) Scheme of the original tree [4]. (B) Scheme of the tree obtained with the revised dataset. Both trees were inferred using exactly the same probabilistic method and model (i.e., using RAxML [75] with a GTR+Γ model for nucleotide sequences and a LG+F+Γ model for protein sequences). Numbers in the triangles indicate the number of species used for the corresponding clade. Bullets denote maximum bootstrap support values (BS = 100%); lower values are given. In the revised dataset, numerous discrepancies were corrected (Table S1), and a few genes were discarded because of dubious orthology; 14,112 unambiguously aligned positions were retained. Furthermore, the erroneous use of mitochondrial sequences of demosponge origin to represent both hexactinellids and calcareans (Figure S9) in the original study [4] drastically—yet probably artifactually—strengthened the support for the monophyly of sponges (BS = 100%; [A]), whereas it appeared much weaker in our reanalysis (BS = 36%; [B]), in line with previous studies [24,26–28] that failed to find significant support for or against sponge monophyly (but see [3]). See Figure S10 for the complete tree obtained with the revised dataset.
doi:10.1371/journal.pbio.1000602.g002

now turn to the selection of the model of sequence evolution. Since their origin [43], the main objective of these models has been to efficiently detect multiple substitutions (Box 4). We reanalyzed the dataset of [3] with a less accurate model, i.e., the site-homogeneous WAG+F+Γ model [44] used in [4] instead of the site-heterogeneous CAT+Γ model [45] used in the original study [3] (Figure 4A). In the WAG+F+Γ tree (Figure 4B), not only does resolution decrease (see BS of 43%, 45%, or 55%), but also the fast-evolving ctenophores now emerge at the base of all animals with strong support (BS = 98%), exactly as expected for a LBA artifact due to model mis-specifications. This indicates that when the less appropriate WAG+F+Γ model is used, multiple substitutions are so poorly inferred that branch lengths are miscalculated (i.e., non-phylogenetic signal has overwhelmed phylogenetic signal).

In summary, the incongruence at the base of the animal tree observed in recent phylogenomic studies [2–4] can be explained by (i) a limited amount of phylogenetic signal, reflected in the short internal branches, and (ii) a profusion of confounding non-phylogenetic signal in certain cases. Since genuine phylogenetic signal is similar in all three analyses (i.e., internal branch lengths are identical and datasets are of similar size), conflicts are due to variations in the level of non-phylogenetic signal—depending on the quantity of non-orthologous sequences included, the number of species considered, and the model of sequence evolution selected. Ultimately, the ratio of phylogenetic to non-phylogenetic signal will determine the outcome: (i) when the phylogenetic signal is strong (sufficiently long internal branches), phylogenomics is always able to recover the correct topology, as found in the three studies [2–4] for outgroup and bilaterian phylogenies; (ii) when both signals are weak, results are statistically non-significant, as is often observed for deep animal relationships; and (iii) when the phylogenetic signal is weak (short internal branches) and the non-phylogenetic signal is strong (e.g., scarce taxon sampling), an artifactual topology is robustly inferred, such as the monophyly of "diploblasts" [4] or the basal

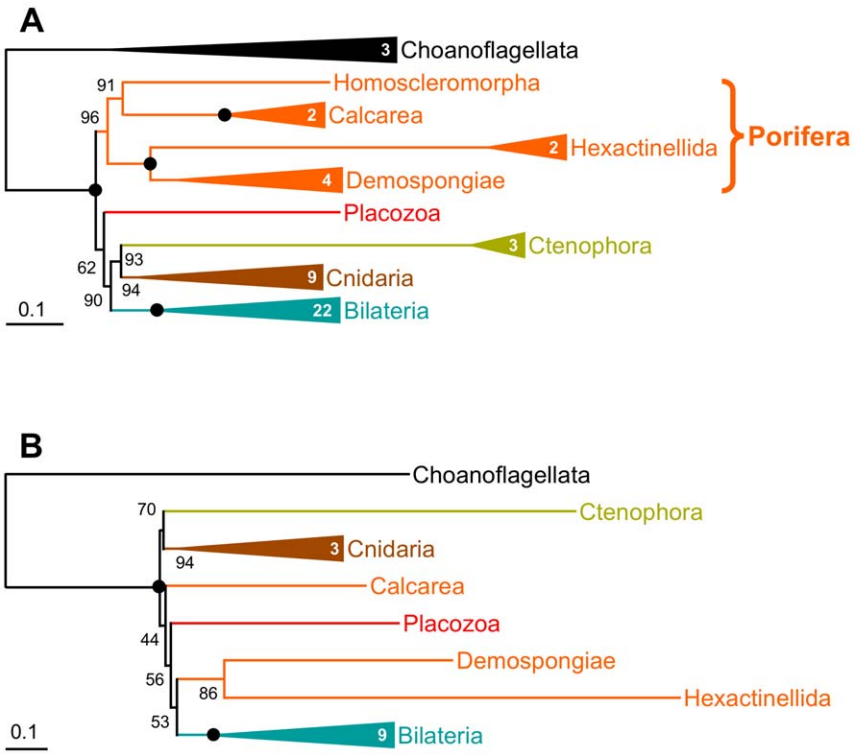

Figure 3. Reanalysis of the Philippe et al. dataset with a reduced taxon sampling. (A) Scheme of the original tree [3]. (B) Scheme of the tree obtained after reduction of the taxon sampling. Both trees were inferred using exactly the same probabilistic method and model (i.e., PhyloBayes using the CAT+Γ model [76]). Numbers in the triangles indicate the number of species used for the corresponding clade. Bullets denote maximum bootstrap support values (BS = 100%); lower values are given. See Figure S12 for the complete tree obtained after reduction of the taxon sampling.
doi:10.1371/journal.pbio.1000602.g003

Box 4. Improving Phylogenetic Inference Methods

There is broad consensus on the necessity of using probabilistic methods in phylogenetic inference. Development of more accurate models of sequence evolution is central to the improvement of these methods. This generally implies more complex models, which are expected to come with increased computational load. Hence, in-depth analyses of datasets that are rich in both genes and species with such models can become prohibitive [46]. Consequently, some promising approaches, e.g., accounting for three-dimensional structure of proteins [56,57] or performing joint alignment and phylogeny [58,59], will probably stay out of reach for years. Fortunately, numerous recent algorithmic developments [60–62] significantly speed up phylogenetic computations, thus paving the way for model improvements. One generally considers that models should be biologically sound. Although biological realism is particularly important for understanding molecular evolution, it is less central for phylogenetic inference, where improving detection of multiple substitutions should be the top priority. As a result, models that more accurately distinguish a synapomorphy from a convergence greatly improve phylogenetic accuracy. Briefly, major steps forward were the modeling of heterogeneity of rate across species [63], heterogeneity of rate across substitutions [64,65], heterogeneity of nucleotide/amino acid composition across species [66,67], heterogeneity of rate across sites [68], and heterogeneity of the substitution process across sites [45]. In contrast, some other improvements, e.g., to handle heterotachy (i.e., heterogeneity of rate over time), had limited effects on phylogenetic reconstruction [69]; heterogeneity of rates across genes, handled by separate models [50], also has limited impact ([70], but see [71]). Future progress is expected (i) from the combination of various existing models [72], (ii) from the handling of other complexities, such as the heterogeneity of the substitution process over time, and (iii) from the handling of incomplete lineage sorting [11,73].

emergence of ctenophores (Figure 4B) (see also [2,32,46]).

Issues at the Level of Gene Sampling

Last but not least, it should be noted that not all genes contain the same potential amount of non-phylogenetic signal. Depending on both functional constraints and evolutionary trajectory, different genes can include positions subject to different ranges of multiple substitutions, i.e., they may display variable levels of saturation. To estimate the saturation in the three datasets [2–4], we used the comparison of patristic and uncorrected distances [47]. As shown by the slope of the regression line (data without any saturation have slope = 1; see [12]), the three datasets (Figure 5) are different, with that of Schierwater et al. being the most saturated (slope = 0.38) and that of Philippe et al. the least affected by multiple substitutions (slope = 0.53). This uneven amount of non-phylogenetic signal explains in part the differences observed in the three studies, but is difficult to separate from other factors. The phylogeny of Figure 1C, with the monophyly of each of Coelenterata (cnidarians+ctenophores) and

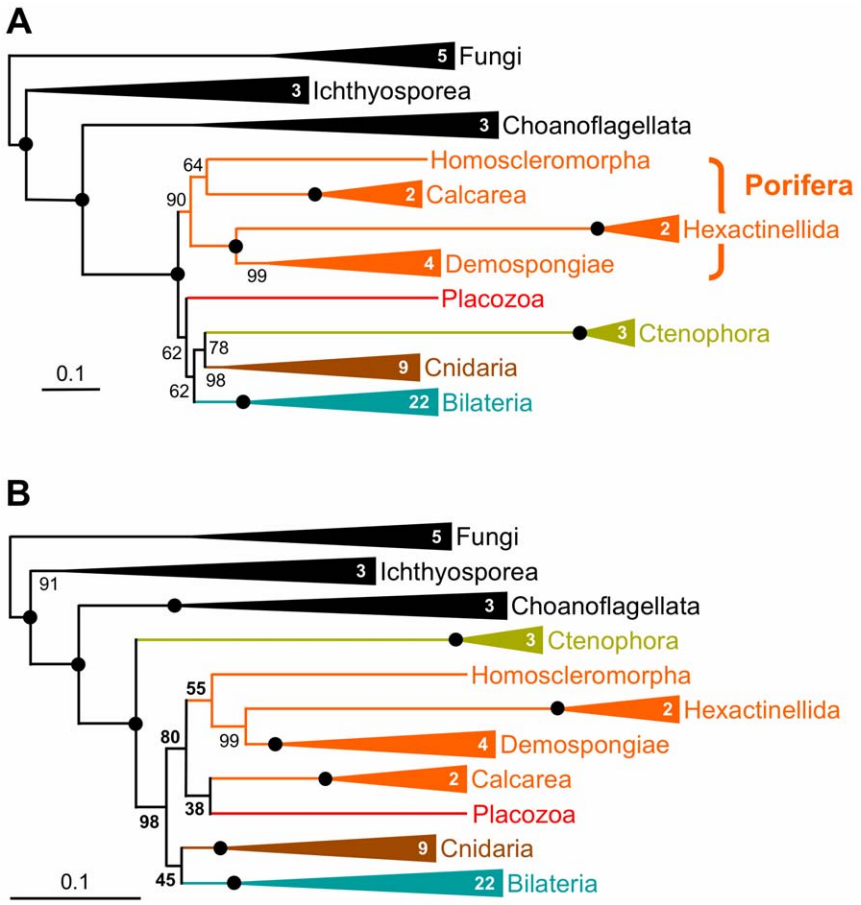

Figure 4. Reanalysis of the Philippe et al. dataset with a less complex model. (A) Scheme of the original tree [3] obtained with the CAT+Γ model. (B) Scheme of the tree obtained with the less complex WAG+F+Γ model. Both trees were inferred using exactly the same dataset. The WAG+F+Γ model has a less good fit to this alignment than the CAT+Γ model [3]. Numbers in the triangles indicate the number of species used for the corresponding clade. Bullets denote maximum bootstrap support values (BS = 100%); lower values are given. See Figure S14 for the complete tree obtained with the less complex WAG+F+Γ model.
doi:10.1371/journal.pbio.1000602.g004

Eumetazoa (all animals except sponges and placozoans), could be considered as the working hypothesis, because Philippe et al. [3] strived to minimize all three sources of non-phylogenetic signal (through the use of weakly saturated genes, a large number of species, and a complex model of sequence evolution). Nevertheless, the scarcity of phylogenetic signal shown here argues strongly for additional studies to confidently resolve the relationships among non-bilaterian animals.

Conclusion

Contrary to common belief, some degree of conflict has to be expected when applying phylogenomics to difficult phylogenetic questions, because of the preva-

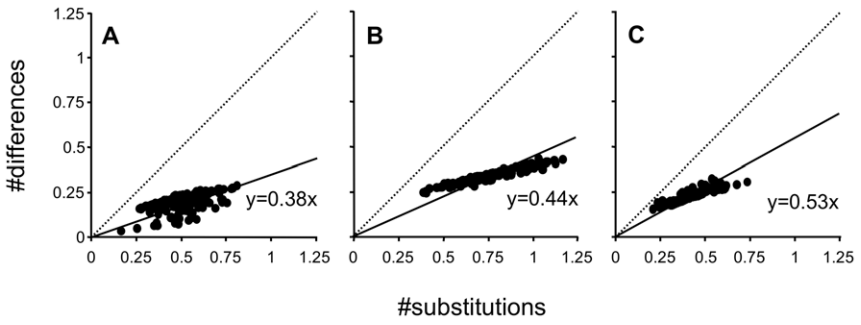

Figure 5. Saturation levels of datasets from Schierwater et al., Dunn et al., and Philippe et al. (A) Schierwater et al. [4] dataset. (B) Dunn et al. [2] dataset. (C) Philippe et al. [3] dataset. The revised alignments from Schierwater et al. and Dunn et al. were used (available as Datasets S1 and S2; see Text S1). The level of saturation was estimated for each dataset by computing the slope of the regression line of patristic distances (*y*-axis) versus uncorrected distances (*x*-axis), as previously described [12]. Patristic distances between two species were computed from branch lengths of the best maximum likelihood tree (using a GTR+Γ model for nucleotide sequences and a LG+F+Γ model for protein sequences).
doi:10.1371/journal.pbio.1000602.g005

lence of non-phylogenetic signal. Consequently, we stress the necessity of reducing its impact. Since taxon and gene sampling is being rapidly improved by the relentless progress in sequencing technology (even if obtaining well preserved and correctly identified specimens remains the limiting factor for several key taxa), full achievement of the ultimate goal of phylogenomics—i.e., accurate resolution of the Tree of Life—will primarily hinge on better procedures for the selection of orthologous and least saturated genes as well as on improved models of sequence evolution. In summary, while we certainly encourage the inclusion of neglected groups of organisms in large-scale sequencing studies (e.g., [2,3,46,48]), we consider at least as important that phylogeneticists engage in theoretical and bioinformatics developments that keep pace with sequencing technology to overcome these serious bottlenecks. This is essential to ensure that lessons learned from classical and molecular systematics are not forgotten in the phylogenomic era.

Supporting Information

Dataset S1 Updated alignment of the Schierwater et al. dataset under the Nexus format.
Found at: doi:10.1371/journal.pbio.1000602.s001 (0.10 MB ZIP)

Dataset S2 Updated alignment of the Dunn et al. dataset under the Nexus format.
Found at: doi:10.1371/journal.pbio.1000602.s002 (0.46 MB ZIP)

Figure S1 Phylogeny of the AT6 gene.
Found at: doi:10.1371/journal.pbio.1000602.s003 (0.06 MB PDF)

Figure S2 Phylogeny of the CDC gene.
Found at: doi:10.1371/journal.pbio.1000602.s004 (0.06 MB PDF)

Figure S3 Phylogeny of the RP3 gene.
Found at: doi:10.1371/journal.pbio.1000602.s005 (0.06 MB PDF)

Figure S4 Phylogeny of the EF1 gene.
Found at: doi:10.1371/journal.pbio.1000602.s006 (0.04 MB PDF)

Figure S5 Phylogeny of the H70 gene.
Found at: doi:10.1371/journal.pbio.1000602.s007 (0.04 MB PDF)

Figure S6 Phylogeny of the PAX gene.
Found at: doi:10.1371/journal.pbio.1000602.s008 (0.04 MB PDF)

Figure S7 Phylogeny of the RAS gene.
Found at: doi:10.1371/journal.pbio.1000602.s009 (0.06 MB PDF)

Figure S8 Phylogeny of the CO2 gene.
Found at: doi:10.1371/journal.pbio.1000602.s010 (0.06 MB PDF)

Figure S9 Taxonomic misidentification for mitochondrial proteins of sponges.
Found at: doi:10.1371/journal.pbio.1000602.s011 (0.06 MB PDF)

Figure S10 Analysis of the revised Schierwater et al. dataset.
Found at: doi:10.1371/journal.pbio.1000602.s012 (0.04 MB PDF)

Figure S11 Analysis of the updated Dunn et al. dataset.
Found at: doi:10.1371/journal.pbio.1000602.s013 (0.07 MB PDF)

Figure S12 Reanalysis of the Philippe et al. dataset with a reduced taxon sampling.
Found at: doi:10.1371/journal.pbio.1000602.s014 (0.04 MB PDF)

Figure S13 Reanalysis of the original Schierwater et al. alignment with only the closest outgroup (Choanoflagellata).
Found at: doi:10.1371/journal.pbio.1000602.s015 (0.04 MB PDF)

Figure S14 Reanalysis of the Philippe et al. dataset with a less complex model.
Found at: doi:10.1371/journal.pbio.1000602.s016 (0.06 MB PDF)

Table S1 List of errors detected in the dataset of Schierwater et al.
Found at: doi:10.1371/journal.pbio.1000602.s017 (0.05 MB PDF)

Table S2 List of errors detected in the dataset of Dunn et al.
Found at: doi:10.1371/journal.pbio.1000602.s018 (0.13 MB PDF)

Text S1 Methods and supporting information.
Found at: doi:10.1371/journal.pbio.1000602.s019 (0.16 MB DOC)

Acknowledgments

We thank Casey Dunn, Gavin Naylor, Davide Pisani, Scott Roy, Bernd Schierwater, and Mike Steel for critical comments on earlier versions of the manuscript. G. W. is supported by the Deutsche Forschungsgemeinschaft through the Priority Program SPP1174 "Deep Metazoan Phylogeny," D. B. by the Communauté Française de Belgique (ARC Biomod), M. M. by the Agence Nationale de la Recherche, and H. P. by the Canadian Research Chair program and Natural Sciences and Engineering Research Council. The Réseau Québécois de Calcul Haute Performance provided computational resources.

References

1. Gee H (2003) Evolution: ending incongruence. Nature 425: 782.
2. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452: 745–749.
3. Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, et al. (2009) Phylogenomics revives traditional views on deep animal relationships. Curr Biol 19: 706–712.
4. Schierwater B, Eitel M, Jakob W, Osigus HJ, Hadrys H, et al. (2009) Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol 7: e1000020. doi:10.1371/journal.pbio.1000020.
5. Philippe H, Chenuil A, Adoutte A (1994) Can the Cambrian explosion be inferred through molecular phylogeny? Development 120: S15–S25.
6. Saitou N, Nei M (1986) The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J Mol Evol 24: 189–204.
7. Mossel E, Steel M (2005) How much can evolved characters tell us about the tree that generated them? In: Gascuel O, ed. Mathematics of evolution and phylogeny. Oxford: Oxford University Press. pp 384–412.
8. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27: 401–410.
9. Philippe H, Laurent J (1998) How good are deep phylogenetic trees? Curr Opin Genet Dev 8: 616–623.
10. Baurain D, Philippe H (2010) Current approaches to phylogenomic reconstruction. In: Caetano-Anollés G, ed. Evolutionary genomics and systems biology. Hoboken (New Jersey): John Wiley and Sons. pp 17–41.
11. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24: 332–340.
12. Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? Trends Genet 22: 225–231.
13. Philippe H, Delsuc F, Brinkmann H, Lartillot N (2005) Phylogenomics. Annu Rev Ecol Evol Syst 36: 541–562.
14. Liu L, Pearl DK, Brumfield RT, Edwards SV (2008) Estimating species trees using multiple-allele DNA sequence data. Evolution 62: 2080–2091.
15. Kuzniar A, van Ham RC, Pongor S, Leunissen JA (2008) The quest for orthologs: finding the corresponding gene across genomes. Trends Genet 24: 539–551.
16. Felsenstein J (2004) Inferring phylogenies. Sunderland (Massachusetts): Sinauer Associates. 645 p.
17. Notredame C (2007) Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol 3: e123. doi:10.1371/journal.pcbi.0030123.
18. Fitch WM (2000) Homology: a personal view on some of the problems. Trends Genet 16: 227–231.
19. van Dongen S (2000) Graph clustering by flow simulation [PhD dissertation]. Utrecht, The Netherlands: University of Utrecht, Available: http://igitur-archive.library.uu.nl/dissertations/1895620/inhoud.htm. Accessed 9 February 2011.
20. Li L, Stoeckert CJ Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.

21. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22–28.
22. Schreiber F, Pick K, Erpenbeck D, Worheide G, Morgenstern B (2009) OrthoSelect: a protocol for selecting orthologous groups in phylogenomics. BMC Bioinformatics 10: 219.
23. Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. J Mol Evol 52: 540–542.
24. Haen KM, Lang BF, Pomponi SA, Lavrov DV (2007) Glass sponges and bilaterian animals share derived mitochondrial genomic features: a common ancestry or parallel evolution? Mol Biol Evol 24: 1518–1527.
25. Kobayashi M, Wada H, Satoh N (1996) Early evolution of the Metazoa and phylogenetic status of diploblasts as inferred from amino acid sequence of elongation factor-1 alpha. Mol Phylogenet Evol 5: 414–422.
26. Medina M, Collins AG, Silberman JD, Sogin ML (2001) Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. Proc Natl Acad Sci U S A 98: 9707–9712.
27. Rokas A, King N, Finnerty J, Carroll SB (2003) Conflicting phylogenetic signals at the base of the metazoan tree. Evol Dev 5: 346–359.
28. Sperling EA, Peterson KJ, Pisani D (2009) Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Mol Biol Evol 26: 2261–2274.
29. Galtier N (2007) A model of horizontal gene transfer and the bacterial phylogeny problem. Syst Biol 56: 633–642.
30. Brochier C, Bapteste E, Moreira D, Philippe H (2002) Eubacterial phylogeny based on translational apparatus proteins. Trends Genet 18: 1–5.
31. Dellaporta SL, Xu A, Sagasser S, Jakob W, Moreno MA, et al. (2006) Mitochondrial genome of Trichoplax adhaerens supports placozoa as the basal lower metazoan phylum. Proc Natl Acad Sci U S A 103: 8751–8756.
32. Pick KS, Philippe H, Schreiber F, Erpenbeck D, Jackson DJ, et al. (2010) Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol Biol Evol 27: 1983–1987.
33. Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst Zool 38: 297–309.
34. Baurain D, Brinkmann H, Philippe H (2007) Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Mol Biol Evol 24: 6–9.
35. Hillis DM (1998) Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst Biol 47: 3–8.
36. Wiens JJ (2005) Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst Biol 54: 731–742.
37. Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51: 588–598.
38. Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst Biol 58: 130–145.
39. 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.
40. Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, et al. (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21: 1740–1752.
41. Wiens JJ (2003) Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol 52: 528–538.
42. Wiens JJ, Moen DS (2008) Missing data and the accuracy of Bayesian phylogenetics. J Syst Evol 46: 307–314.
43. Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN, ed. Mammalian protein metabolism. New York: Academic Press. pp 21–132.
44. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–699.
45. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21: 1095–1109.
46. Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, et al. (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci 276: 4261–4270.
47. Philippe H, Sörhannus U, Baroin A, Perasso R, Gasse F, et al. (1994) Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. J Evol Biol 7: 247–265.
48. Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, et al. (2010) A phylogenomic approach to resolve the arthropod tree of life. Mol Biol Evol 27: 2541–2464.
49. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6: 361–375.
50. Yang Z (1996) Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42: 587–596.
51. Kupczok A, Schmidt HA, von Haeseler A (2010) Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Algorithms Mol Biol 5: 37.
52. Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, et al. (2009) Fast statistical alignment. PLoS Comput Biol 5: e1000392. doi:10.1371/journal.pcbi.1000392.
53. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17: 540–552.
54. Roure B, Rodriguez-Ezpeleta N, Philippe H (2007) SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol Biol 7(Suppl 1): S2.
55. Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, et al. (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc Natl Acad Sci U S A 99: 1414–1419.
56. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20: 1692–1704.
57. Rodrigue N, Philippe H, Lartillot N (2006) Assessing site-interdependent phylogenetic models of sequence evolution. Mol Biol Evol 23: 1762–1775.
58. Redelings BD, Suchard MA (2005) Joint Bayesian estimation of alignment and phylogeny. Syst Biol 54: 401–418.
59. Lunter G, Miklos I, Drummond A, Jensen JL, Hein J (2005) Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics 6: 83.
60. Lartillot N (2006) Conjugate Gibbs sampling for Bayesian phylogenetic models. J Comput Biol 13: 43–63.
61. Stamatakis A, Ott M (2008) Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. Philos Trans R Soc Lond B Biol Sci 363: 3977–3984.
62. de Koning AP, Gu W, Pollock DD (2010) Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories. Mol Biol Evol 27: 249–265.
63. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17: 368–376.
64. Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20: 86–93.
65. Dayhoff MO, Eck RV, Park CM (1972) A model of evolutionary change in proteins. In: Dayhoff MO, ed. Atlas of protein sequence and structure. Washington (District of Columbia): National Biomedical Research Foundation. pp 89–99.
66. Galtier N, Gouy M (1995) Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci U S A 92: 11317–11321.
67. Yang Z, Roberts D (1995) On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol Biol Evol 12: 451–458.
68. Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11: 367–370.
69. Kolaczkowski B, Thornton JW (2008) A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol Biol Evol 25: 1054–1066.
70. Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, et al. (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol 56: 389–399.
71. Nishihara H, Okada N, Hasegawa M (2007) Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol 8: R199.
72. Blanquart S, Lartillot N (2008) A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol 25: 842–858.
73. Than C, Nakhleh L (2009) Species tree inference by minimizing deep coalescences. PLoS Comput Biol 5: e1000501. doi:10.1371/journal.pcbi.1000501.
74. Houliston E, Momose T, Manuel M (2010) Clytia hemisphaerica: a jellyfish cousin joins the laboratory. Trends Genet 26: 159–167.
75. Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456–463.
76. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25: 2286–2288.