Article

Current Biology

A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals

Highlights

- Largest and most internally consistent metazoan-scale superalignment to date
- Sponges (Porifera) are sister-group to all other multicellular animals
- Previous findings of trees with "Ctenophora-sister" were due to artifacts
- Analyses of multigene datasets should employ site-heterogeneous evolutionary models

Authors

Paul Simion, Hervé Philippe, Denis Baurain, ..., Nicole King, Gert Wörheide, Michaël Manuel

Correspondence

herve.philippe@sete.cnrs.fr (H.P.), michael.manuel@upmc.fr (M.M.)

In Brief

Simion et al. demonstrate that sponges (Porifera) are the earliest branching animal lineage, using a combination of 1,719 genes that outperforms in size and quality previous datasets used to address metazoan relationships. Previous findings of comb jellies sister to other animals were likely due to an artifact known as "long branch attraction."

Simion et al., 2017, Current Biology *27*, 958–967
April 3, 2017
http://dx.doi.org/10.1016/j.cub.2017.02.031

CellPress

CellPress

Current Biology
Article

A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals

Paul Simion,[1,17,18] Hervé Philippe,[2,3,17,*] Denis Baurain,[4] Muriel Jager,[1] Daniel J. Richter,[5,6,7] Arnaud Di Franco,[2] Béatrice Roure,[2,3] Nori Satoh,[8] Éric Quéinnec,[1] Alexander Ereskovsky,[9,10] Pascal Lapébie,[11] Erwan Corre,[12,13] Frédéric Delsuc,[14] Nicole King,[5] Gert Wörheide,[15,16] and Michaël Manuel[1,19,*]

[1]Sorbonne Universités, UPMC Univ Paris 06, CNRS, Evolution Paris-Seine UMR7138, Institut de Biologie Paris-Seine, Case 05, 7 quai St Bernard, 75005 Paris, France

[2]Centre de Théorisation et de Modélisation de la Biodiversité, Station d'Ecologie Théorique et Expérimentale, UMR CNRS 5321, Moulis 09200, France

[3]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, QC H3C 3J7, Canada

[4]InBios–Eukaryotic Phylogenomics, Department of Life Sciences and PhytoSYSTEMS, University of Liège, Bât. B22, Quartier Vallée 1, Chemin de la Vallée 4, 4000 Liège, Belgium

[5]Howard Hughes Medical Institute and Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720-3200, USA

[6]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[7]Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[8]Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan

[9]Aix Marseille Univ, Univ Avignon, CNRS, IRD, IMBE, Marseille, France, Chemin de la Batterie des Lions, 13007 Marseille, France

[10]Department of Embryology, Faculty of Biology, Saint-Petersburg State University, Universitetskaya nab. 7/9, Saint-Petersburg 199034, Russia

[11]Sorbonne Universités, UPMC Univ Paris 06, and CNRS, Laboratoire de Biologie du Développement de Villefranche-sur-mer, Observatoire Océanographique, 06230 Villefranche-sur-mer, France

[12]CNRS, FR2424, ABiMS, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[13]Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Paris 06, FR2424, ABiMS, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[14]Université de Montpellier (UM), Institut des Sciences de l'Évolution (ISEM), UMR 5554 CNRS-IRD-EPHE, Case Courier 64, Place Eugène Bataillon, 34095 Montpellier, France

[15]Department of Earth and Environmental Sciences & GeoBio-Center, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, 80333 München, Germany

[16]SNSB - Bayerische Staatssammlung für Paläontologie und Geologie, Richard-Wagner-Str. 10, 80333 München, Germany

[17]Co-first author

[18]Present address: Université de Montpellier (UM), Institut des Sciences de l'Évolution (ISEM), UMR 5554 CNRS-IRD-EPHE, Case Courier 64, Place Eugène Bataillon, 34095 Montpellier, France

[19]Lead Contact

*Correspondence: herve.philippe@sete.cnrs.fr (H.P.), michael.manuel@upmc.fr (M.M.)

http://dx.doi.org/10.1016/j.cub.2017.02.031

SUMMARY

Resolving the early diversification of animal lineages has proven difficult, even using genome-scale datasets. Several phylogenomic studies have supported the classical scenario in which sponges (Porifera) are the sister group to all other animals ("Porifera-sister" hypothesis), consistent with a single origin of the gut, nerve cells, and muscle cells in the stem lineage of eumetazoans (bilaterians + ctenophores + cnidarians). In contrast, several other studies have recovered an alternative topology in which ctenophores are the sister group to all other animals (including sponges). The "Ctenophora-sister" hypothesis implies that eumetazoan-specific traits, such as neurons and muscle cells, either evolved once along the metazoan stem lineage and were then lost in sponges and placozoans or evolved at least twice independently in Ctenophora and in Cnidaria + Bilateria. Here, we report on our reconstruction of deep metazoan relationships using a 1,719-gene dataset with dense taxonomic sampling of non-bilaterian animals that was assembled using a semi-automated procedure, designed to reduce known error sources. Our dataset outperforms previous metazoan gene superalignments in terms of data quality and quantity. Analyses with a best-fitting site-heterogeneous evolutionary model provide strong statistical support for placing sponges as the sister-group to all other metazoans, with ctenophores emerging as the second-earliest branching animal lineage. Only those methodological settings that exacerbated long-branch attraction artifacts yielded Ctenophora-sister. These results show that methodological issues must be carefully addressed to tackle difficult phylogenetic questions and pave the road to a better understanding of how fundamental features of animal body plans have emerged.

INTRODUCTION

The question of how animal-specific cell types and key animal body plan features first evolved cannot be answered without a clear understanding of the phylogenetic relationships among major animal lineages. Analyses of large-scale molecular superalignments assembled from genomic or transcriptomic data have failed thus far to provide a widely accepted consensus for the branching order among the five major animal lineages, i.e., sponges, placozoans, ctenophores, cnidarians, and bilaterians [1–12]. As a consequence, controversy prevails concerning early animal evolution.

Disagreement mainly hinges on the contradictory phylogenetic placements of ctenophores and sponges in different phylogenomic studies: either with sponges branching first ("Porifera-sister") and ctenophores grouping with cnidarians and bilaterians (consistent with the classical view of a single origin of neurons in the eumetazoan stem lineage) [1–4, 12], or instead with ctenophores as the first offshoot of the animal tree ("Ctenophora-sister") and sponges branching second [5–10]. The Ctenophora-sister hypothesis implies that the nerveless and morphologically much simpler sponges and placozoans are unexpectedly more closely related to cnidarians and bilaterians than are ctenophores, and has noticeably fuelled an intense debate around the possibility of two independent acquisitions for neurons and synapses [13–17]. Comparative analysis of gene content has also been proposed in support of Ctenophora-sister [7], but upon improvement of the inference method, the same data supported Porifera-sister instead [4].

Incongruence between phylogenies can arise from a number of sources, including the use of alignments that are flawed in some way (e.g., contaminated, poorly aligned), the inclusion of sequences that do not faithfully record the organismal phylogeny (e.g., because of lateral gene transfer or paralogy), and the use of inappropriate models of sequence evolution. Ctenophores have a high rate of molecular evolution, making them a priori difficult to place due to their potential for long branch attraction (LBA) artifacts [12]. LBA artifacts can result in the erroneous grouping of unrelated lineages due to their unusually high substitution rates, including the branching of a molecularly highly divergent lineage near the base of the tree due to artifactual attraction by distant outgroups. The conundrum of correctly placing ctenophores and sponges in the animal tree of life is stimulating not only due to its important implications for understanding early animal evolution, but also because it has prompted researchers to re-examine the sources of contradiction between phylogenomic studies.

We assembled an entirely new phylogenomic dataset designed to address relationships between early-diverging metazoan lineages. Our superalignment of 1,719 genes was constructed using a novel multi-step procedure devised to integrate knowledge accumulated in recent years about the various potential causes of artifacts and conflicts in phylogenomics. Analyses of this dataset using the site-heterogenous CAT model provide unambiguous support for the Porifera-sister hypothesis. Ctenophores emerge as the sister group to a clade containing placozoans + cnidarians + bilaterians. Ctenophora-sister was recovered only in analyses containing limited sampling of sponge classes and/or using suboptimal models of sequence evolution, strongly suggesting that it is an artifact of long branch attraction.

RESULTS AND DISCUSSION

A New Pan-metazoan Phylogenomic Dataset of Unprecedented Quality and Size

Phylogenomic datasets previously assembled to address non-bilaterian relationships have often contained a substantial amount of data error (both biological and in silico contamination, alignment errors, and false orthology, see below) and/or the sequences included did not contain enough phylogenetic signal to provide a statistically robust resolution to the problem (see [4, 12]). To tackle these limitations, we developed and implemented a new semi-automated pipeline (i.e., automated procedures supplemented with stringent manual controls of intermediate results) to comprehensively detect and eliminate as many data errors as possible (see Figure 1 for a graphical summary of our pipeline, Supplemental Experimental Procedures for details, and Figure S1 for examples of errors in published datasets constructed with less stringent automated procedures; an example showing trees for a given gene at all successive steps of our filtering procedure is provided at https://github.com/psimion/SuppData_Metazoa_2017). The goal was to simultaneously optimize taxonomic sampling, data quantity (gene number), and data quality. Viewed from this perspective, we therefore reconcile the two principal differing operational philosophies that have thus far competed in the field of phylogenomics: (1) reliance on a limited number of established gene alignments where potential sources of error can be manually curated (e.g., [1, 12]) and (2) entirely automated construction and limited quality control of hundreds of gene alignments (orthology groups) (e.g., [5–10]). The first (manual) approach is not scalable to thousands of genes while the second (automated) approach has not, until now, satisfactorily addressed all sources and types of error (see examples in Figures S1A–S1D).

The resulting dataset comprises 1,719 genes and 97 species (with 39.3% missing data), including 61 non-bilaterian metazoan species. For 21 of these 61 species, we produced new transcriptome assemblies as part of this study (see Supplemental Experimental Procedures for details of taxon sampling). This new dataset outperforms other previously published metazoan phylogenomic supermatrices both in terms of size (total number of amino acid residues) and quality (internal congruence, as estimated from the mean percentage of recovery of clades in single-gene phylogenies that are present in the species tree reconstructed from the supermatrix of concatenated genes) (Figure 2). It contains from two to ten times more information than other datasets and displays a level of congruence (60%) that exceeds that of a manually curated dataset supporting Porifera-sister [1] (57%) and of all automatically assembled datasets that have yielded trees in favor of Ctenophora-sister [5–10] (average of 39%). Our protocol for supermatrix construction therefore appears to be better suited for eliminating data errors than those that have been previously used.

Sponges Are Sister to All Other Metazoans

The supermatrix was analyzed in a Bayesian framework using the site-heterogeneous CAT model, which was originally conceived to minimize LBA artifacts by taking into account the observation that only a limited number of amino acids are functionally acceptable at a given position [18]. Cross-validation experiments, a

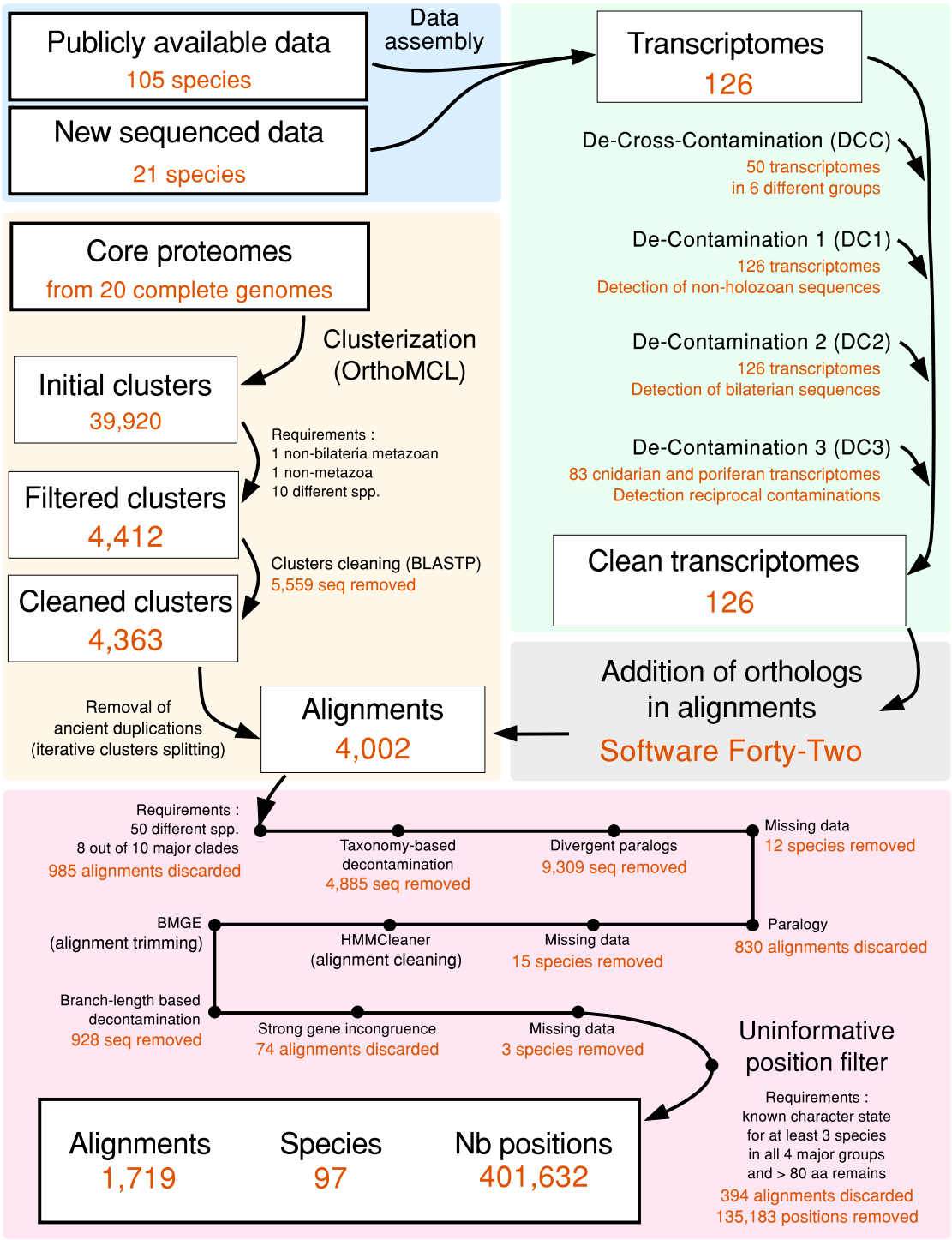

Figure 1. Graphical Summary of the Dataset Construction Protocol Used in This Study
See Experimental Procedures and Supplemental Experimental Procedures for the details of each step.

well-established statistical method to evaluate model fit [19], showed that the fit of the CAT model to the data used in this study is superior to that of site-homogeneous models (ΔlnL = 2,314 ± 164 compared to LG and 1,956 ± 154 compared to GTR), in agreement with previous studies (e.g., [3, 4]). A recent study using simulated data suggested that the CAT model might be less accurate than site-homogeneous models (e.g., LG) under some circumstances [20]. However, the biological relevance of these simulations has not yet been thoroughly explored. In particular, the CAT model appears to fit less well to these simulated data

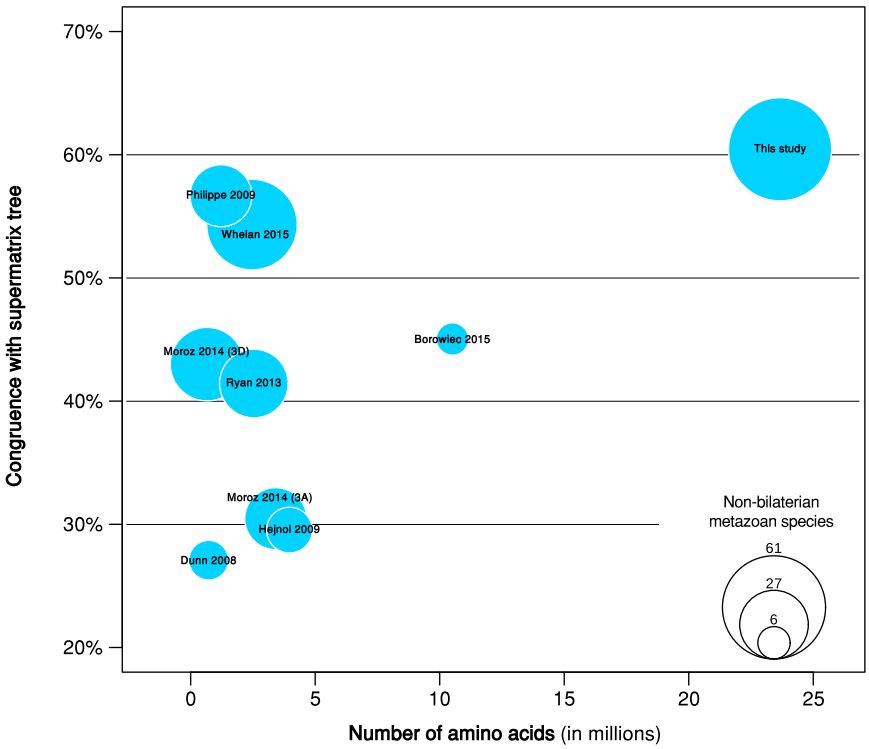

Figure 2. Comparison of Data Quantity and Quality between Nine Recent Phylogenomic Datasets

Quantity of molecular information (x axis) corresponds to the number of amino acids in the supermatrix, and data quality (y axis) corresponds to the percentage of bipartitions (internal tree branches) that are identical in both the single-gene trees and the supermatrix tree (internal congruence); only genes longer than 200 amino acids are represented (see https://github.com/psimion/SuppData_Metazoa_2017 for the corresponding plot with the remaining genes). Circled areas are proportional to the number of non-bilaterian metazoan species present in the datasets, indicated with abbreviated reference information. For each study included here, we analyzed the largest available dataset, except for Moroz et al. [8] where we analyzed both the largest one ("3A" dataset) as well as a smaller one enriched in ctenophores ("3D" dataset). Note that we also analyzed reduced datasets from Whelan et al. [10] (i.e., their datasets 6, 10, and 16, results not shown), which yielded congruence results almost identical to those of their dataset 1. See also Figure S1.

than do site-homogeneous models (H.P., unpublished results), while the opposite is true for most real datasets.

Because of the large size of our supermatrix, it was not computationally feasible to analyze the whole dataset using the CAT model. To circumvent this limitation, we used a gene jackknife strategy based on 100 separate analyses, each involving a random selection of roughly 25% of the genes in our dataset (i.e., about 100,000 amino acid positions per replicate). This strategy allowed us to combine the advantage of reduced computational burden with reliable estimates of the statistical robustness of each clade. Furthermore, jackknifing genes counteracts potentially strongly misleading signals that might be contained in a small number of genes. These analyses (Figure 3) revealed that monophyletic sponges (Jackknife Support [JS] of 100%) emerge as the sister group to all other metazoans (Porifera-sister) with ctenophores as the second diverging animal lineage (JS 95%), followed by placozoans, which are the sister group to a cnidarian + bilaterian clade (JS 100%). Since the inclusion of distant outgroups is known to amplify LBA artifacts [21–24], the tree of Figure 3 was rooted using only choanoflagellates, the closest living relatives of metazoans. In addition, an analysis including a more complete holozoan sampling also supports Porifera-sister; hence, this result does not depend on outgroup sampling (Figure S2A). Sponges were monophyletic in all analyses, and relationships within Porifera, Ctenophora, Cnidaria, and Bilateria were generally strongly supported and consistent with other molecular studies [25–27].

In previous studies, the Porifera-sister hypothesis has tended to be better supported by smaller, manually constructed and curated phylogenomic datasets [1–3], whereas datasets featuring many more genes and assembled using entirely automated procedures have tended to support Ctenophora-sister (e.g., [5–10]). This has fuelled the idea that increasing gene sampling and taxon quantity were the drivers of higher support for Ctenophora-sister [28]. Recent re-analyses of several of these datasets have cautioned, however, that this support vanishes once the effects of outgroup sampling and model choice (site-heterogeneous versus site-homogeneous) are simultaneously taken into account [4]. In this study, a multigene dataset generated using semi-automated procedures, including data quality controls of unparalleled stringency, and containing the largest amount of molecular information and taxonomic representation used to date in metazoan phylogenomics, yielded strong statistical support for sponges rather than ctenophores as the sister-group to all other animals. This result is consistent with previous propositions [4, 12] that Ctenophora-sister stems from an LBA artifact due to the use of poorly fitting evolutionary models that lead to statistical inconsistency (LBA being a form of systematic error) when analyzing large gene numbers.

Drastic Effects of Taxon Sampling and Site-Homogeneous versus Site-Heterogeneous Model Type on the Placement of Long Branches

To assess the impact of using a less well-fitting site-homogeneous model of sequence evolution (LG) versus a better-fitting site-heterogeneous model (CAT) on the placement of long branches in the phylogeny, we examined the behavior of the two metazoan lineages having the highest rate of substitution in our dataset, ctenophores and hexactinellid sponges, when other sponges (the closest relatives of hexactinellids) were either included or excluded from the dataset. With full taxonomic sampling, despite their high substitution rate, hexactinellids were correctly located as the sister group to demosponges [25] by both models (Figures 4A and 4D), probably because the branch that unites these two groups is sufficiently long to overcome

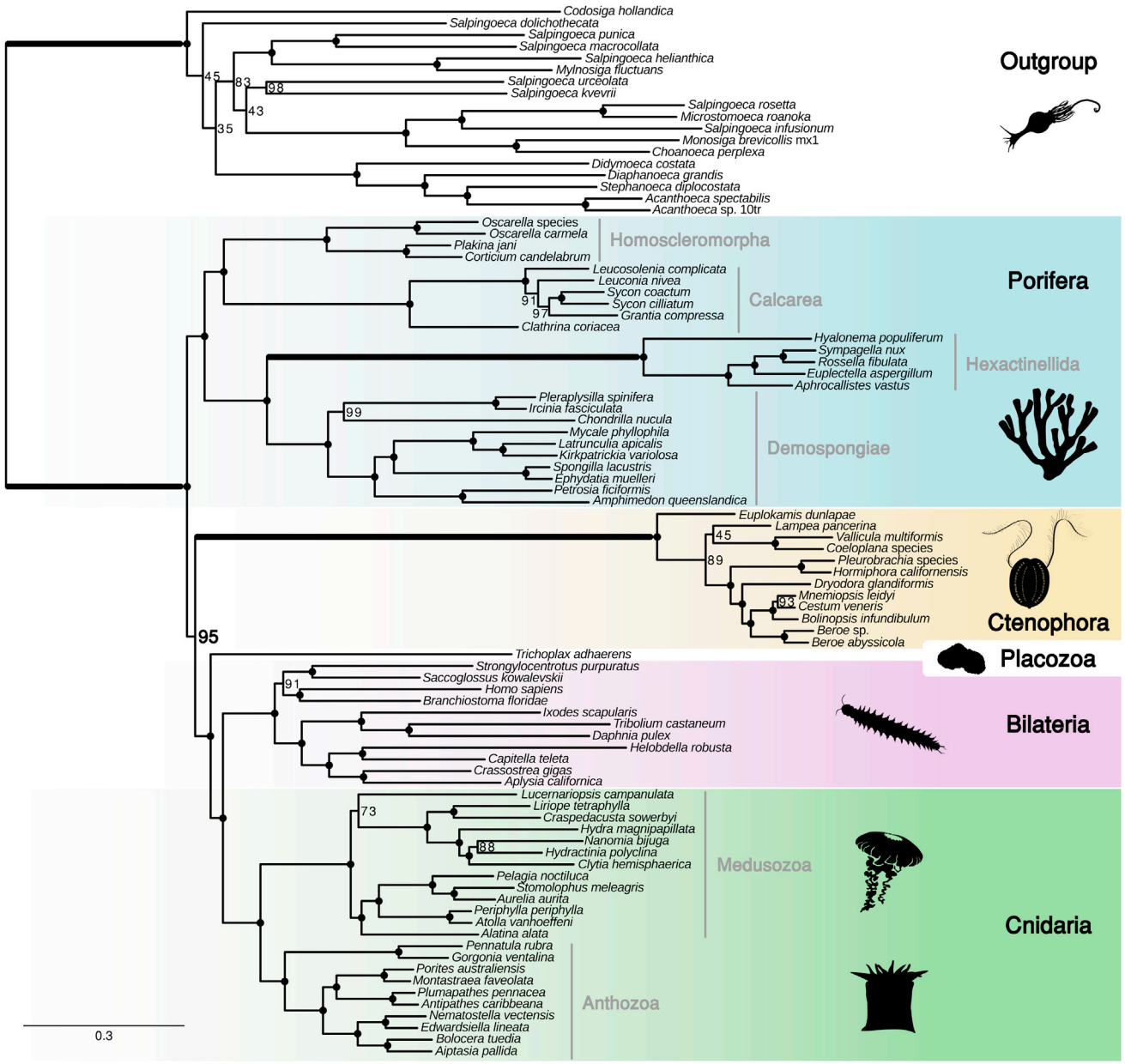

Figure 3. Metazoan Phylogenetic Relationships Inferred from a Supermatrix of 401,632 Amino Acid Positions for 90 Species

Every gene jackknife replicate was composed of ~100,000 positions (each replicate represents a larger dataset than those of any previous phylogenomic study [1, 5–8, 10] except [9]) and was analyzed using PhyloBayes_MPI 1.6j under the site-heterogeneous CAT+ Γ_4 model of sequence evolution. The tree shown here is the consensus of the 100 jackknife replicates and branch support values (JS %) represent the number of analyses in which each branch was recovered; black circles represent nodes with maximal support (100%). The three longest branches in terms of inferred substitutions are highlighted with thicker lines: the branch separating metazoans from outgroups and the terminal branches bearing hexactinellid sponges and ctenophores. The supermatrix had an overall percentage of missing data of 37.3%. Organism drawings were downloaded from the PhyloPic website. See also Figures S2 and S4.

any artifactual signal. We note that the site-homogeneous LG model (Figure 4D) recovered Ctenophora-sister, unlike the site-heterogeneous CAT model (Figure 4A), which recovered Porifera-sister. When demosponges were discarded, hexactinellids remained grouped with the other sponges when the site-heterogeneous CAT model was used (Figure 4B), but with the site-homogeneous LG model they formed a maximally supported clade with ctenophores (bootstrap support [BS] 100%), located at the base of metazoans (Figure 4E). Due to the removal of demosponges, the short internal branch linking hexactinellids to calcareous and homoscleromorph sponges, in combination with the use of a less well-fitting model, was insufficient to counteract the LBA artifact. This represents a quintessential LBA configuration, with the three longest branches (two internal and one external; highlighted with thicker lines in Figure 3) are clustered together. When all other sponges were removed, the

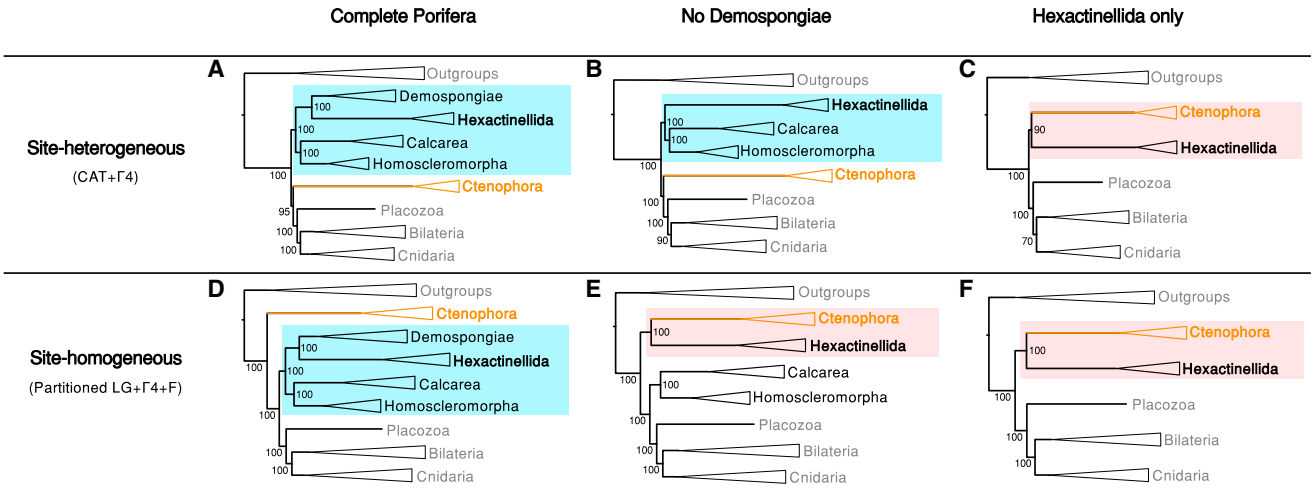

Figure 4. Comparison of the Behavior of Long Branches with Varying Taxon Sampling, under Site-Heterogeneous versus Site-Homogeneous Models

The multigene dataset of this study was analyzed with the CAT+ Γ_4 model (A, B, C) and the LG+ Γ_4+F model (D, E, F) with three different samplings of sponge taxa.
(A and D) Full taxonomic sampling, with all four sponge classes represented.
(B and E) Demosponges removed.
(C and F) Demosponges, calcareous sponges, and homoscleromorphs removed (sponges represented by hexactinellids only).
Node support for (A), (B), and (C) corresponds to jackknife support values (JS %) from 10 jackknife replicates on the complete supermatrix, while node support for (D), (E), and (F) corresponds to bootstrap support values (BS %). Analyses shown in (D) were run with complete and reduced outgroup samplings (see legend of Figure S2B). See also Figure S3 (same comparison performed on another phylogenomic dataset) and Table S1 (same comparison using likelihood computations with additional models of sequence evolution).

CAT analysis also yielded this erroneous placement of hexactinellids (Figure 4C), but with a slightly lower support than LG (JS 90% versus BS 100%, Figures 4C and 4F). This experiment confirms the higher sensitivity of site-homogeneous models to LBA [29] and indicates that the CAT model can also be impacted, albeit to a lesser extent. The same procedure applied to a previously published dataset [10] led to the same results, suggesting that the choice of the model of evolution, and not the dataset per se, is responsible for the effect observed (see Figure S3).

To test whether the CAT model and its Bayesian implementation are necessary to handle the heterogeneity of the evolutionary process across sites, we compared the likelihoods of the topologies of Figure 4 using various models, from purely site-homogeneous ones (LG, WAG) to site-heterogeneous models that are implemented in a maximum likelihood framework but are more restricted than CAT in the levels of pattern heterogeneity they can accommodate (see results in Table S1). Even with the simple removal of only demosponges, all these models yielded an erroneous clade of ctenophores and hexactinellids, although the likelihood difference (ΔlnL) in comparison to the correct topology (hexactinellids grouped with other sponges) was higher for site-homogeneous models and decreased as progressively more complex site-heterogeneous models were used (C20, C40, C60). These observations suggest that the levels of pattern heterogeneity present in empirical sequence alignments cannot be appropriately modeled by empirical finite mixtures and require instead the use of Dirichlet process models, such as CAT.

Implications for Body Plan Evolution

The principal outcome of this study is a rejection of the Ctenophora-sister hypothesis of animal evolution. With the artifactual basal long branch attraction of ctenophores now reduced, Porifera is strongly supported as the sister group of other metazoans. Therefore, the absence in sponges of features such as a gut, neurons, synapses, and muscles is more parsimoniously interpreted as ancestral in metazoans (i.e., plesiomorphic), in line with classical views.

The topology obtained in our analyses (Figure 3) does not fully clarify, however, the pattern of emergence of these features (shared by ctenophores, cnidarians, and bilaterians) within the non-sponge clade, because they are also lacking in placozoans, one of the most simply organized metazoan phyla. Thus, according to our phylogeny, either nervous systems and other advanced features originated independently in ctenophores and in the lineage leading to cnidarians + bilaterians, or placozoans represent the evolutionary loss of features that existed in a more complex ancestor, as has been previously suggested [14, 30], in consistence with the eumetazoan-like gene content of the placozoan genome [31]. Furthermore, "Ctenophora-second" as obtained here (Figure 3) could either represent the correct topology or reflect a trade-off between their attraction toward the base (attenuated but not suppressed with the CAT model) and a more internal true position. This is suggested by analyses in which most heteropecillous sites (i.e., violating CAT model assumptions [32]) were removed: in these trees, ctenophores are sister-group to other eumetazoans (Figure S4A) or to cnidarians (Figure S4B) [1], consistent with a single origin of eumetazoan features including nerve cells.

The anatomical features of ctenophores are of little help for understanding their relationships with other early-diverging animal lineages, because they include numerous unique traits, such as distinctive biradial symmetry, extraordinary macrocilia forming

swimming paddles or combs (arranged in eight rows around the body), a sophisticated aboral neuro-sensory complex without any counterpart in other phyla, and adhesive cells with unparalleled cytological features (colloblasts) (see [33, 34]). The problem of homology or convergence of their nervous system and muscles with those of cnidarians and bilaterians remains open to debate. Furthermore, regardless of its precise phylogenetic position, the ctenophore lineage has evolved a dramatic increase in anatomical complexity, which is paralleled only in bilaterians, and whose genomic and molecular developmental bases remain obscure.

Conclusions

The recent debate about deep metazoan relationships stems from methodological issues that call for improvements in terms of both data curation and inference. Our novel semi-automated protocol yielded a pan-metazoan phylogenomic dataset of greatly increased quality and size relative to other datasets used for reconstructing metazoan phylogenies. Using a stringent gene jackknifing strategy, we obtained strong support for placing sponges as the sister group to all other metazoans. Nonetheless, our observations with different sub-samplings of the sponge classes indicate that the site-heterogeneous CAT model, despite outcompeting any site-homogeneous model, is unsurprisingly not immune to reconstruction artifacts [23, 32].

These considerations call for further improvements in phylogenomics procedures, in order to better handle the tremendous level of data complexity that characterizes supermatrices composed of thousands of genes sampled across many different phyla. In-depth data curation in order to remove contaminants and paralogs is currently a necessity, as they are a major source of inter-gene incongruence, although a promising alternative approach could be to model them in a probabilistic framework (e.g., [35]). In parallel, the pervasive role of epistasis in protein evolution makes modeling of sequence evolution in a site-independent framework (currently a condition for models to remain computationally tractable) problematic. Therefore, determining which violations of model assumptions are the most detrimental to phylogenetic reconstruction accuracy is of primary importance [24]. These violations include heterotachy, heteropecilly, global compositional bias, and relationships among sites due to 3D structural features, among others. Integrating these various properties into models will significantly improve phylogenetic reconstruction and may resolve current controversies regarding deep metazoan relationships, in addition to those in other parts of the tree of life.

EXPERIMENTAL PROCEDURES

A graphical summary of the whole pipeline for dataset construction is provided in Figure 1. Supporting material such as gene alignments, programs, supermatrices and additional phylogenetic trees can be found at https://github.com/psimion/SuppData_Metazoa_2017.

Genome Sampling and Dataset Assembly

We used protein sequences predicted from the complete genomes of 20 selected "core species" (17 metazoans, 2 choanoflagellates, and 1 filasterean) (see Table S4) to create clusters of putative orthologous genes. For this, we used the programs USEARCH [36] (e-value = 1×10^{-5}; accel = 1.0) and OrthoMCL [37] with the default inflation parameter (I = 1.5), which resulted in the creation of a set of 39,920 clusters of putative orthologs. We then selected the 4,412 clusters that contained $\geq$10 different species, including at least one non-bilaterian metazoan and one non-metazoan species (outgroup). Because of the limitations resulting from the use of similarity scores and of single-linkage clustering, OrthoMCL clusters may contain non-homologous sequences. Thus, in a two-step BLAST procedure, we discarded sequences that did not match (e-value $\leq 1 \times 10^{-10}$): (1) $\geq$30% of other sequences in their cluster (2,990 sequences removed), (2) $\geq$50% other sequences in their cluster (2,520 sequences removed). Thinned clusters were aligned using MAFFT [38] (localpair, maxiterate = 5,000) then cleaned of non-homologous stretches via HMMCleaner [39] and of ambiguously aligned positions via BMGE [40]. The resulting alignments were analyzed with RAxML [41] (LG+Γ_4+F model) to yield single-gene trees. To determine whether clusters contained anciently duplicated genes, trees were split on branches (1) separating two subtrees with $\geq$10 different species in each of the subtrees and (2) within the top 10% longest branches of the tree. Clusters that could be split were iteratively reduced into smaller clusters. Finally, we applied the same taxonomic filter as above, which resulted in the generation of 4,002 core orthologous clusters.

RNA Preparation and Sequencing

Biological samples (see Table S2) were carefully cleaned to remove biological contaminants, then powdered in liquid nitrogen. RNA extraction was performed using either the QIAGEN RNeasy Kit (according to the manufacturer's instructions) or a TRIzol-based protocol. In the latter case, frozen sample powder was incubated for 5 min in TRIzol solution, before addition of chloroform for another 15 min of incubation. The solution was then centrifuged for 15 min at 4°C (12,000 × *g*) in order to retain the upper aqueous phase only, which was subsequently incubated for 10 min in isopropanol. Samples were then centrifuged for 10 min at 4°C (12,000 × *g*) and the supernatant eliminated. The pellet was vortexed and centrifuged for 5 min at 4°C (7,500 × *g*) in ethanol 75%. After supernatant elimination, the dried pellet was finally resuspended in RNase-free water. Construction of cDNA libraries and their sequencing using either 454 pyrosequencing or Illumina technology (see Table S2) was carried out at GATC Biotech. 6 of the 22 newly sequenced species were pooled for a single 454 run (group E in Table S3), while 14 others were pooled in two Illumina lanes of the same run (group F in Table S3).

Transcriptome Sampling, Assembly, and Decontamination

We used 126 non-bilaterian species for which sequence data were either publicly available or provided by Daniel Richter, Nicole King, and Nori Satoh (see third column in Table S3). 454 reads were assembled using MIRA alone [42] or a combination of MIRA and CAP3 [43], whereas Illumina reads were assembled using either Trinity [44] or SOAPdenovo-trans [45]. To reduce subsequent computational time, transcripts that did not match any of the 4,002 orthologous clusters (BLAST e-value $\leq 1 \times 10^{-10}$) were discarded, which reduced the total number of transcripts from 12,247,929 to 1,787,422. We then designed a new procedure to detect and remove cross-contaminating sequences between transcriptomic datasets obtained in the same lab, belonging to the same sequencing project or for which cross-contamination issues were observed in preliminary analyses (see details in Supplemental Experimental Procedures). Species transcriptomes were processed in six groups (see "group" column in Table S3), which allowed us to estimate the level of cross-contamination of each species, ranging from 0.16% (*Pleurobrachia pileus*, this study) to 70.64% (ctenophora sp3 A [8]) of the reads. Transcriptomes were further screened for additional contamination sources using a three-step procedure aiming at detecting contaminations at different taxonomical scales: by non-holozoans (DC1), by bilaterians (DC2), and reciprocal contamination between sponges and cnidarians (DC3). Details about decontamination procedures are given in Supplemental Experimental Procedures, and an illustration of their behavior, in the case of the highly expressed ribosomal protein rpl2 (an extreme case of contamination) is provided at https://github.com/psimion/SuppData_Metazoa_2017.

Transcriptomic Data Integration into Orthologous Clusters

Decontaminated transcriptomic data were then incorporated into the 4,002 previously assembled core orthologous clusters using a multiple Best Reciprocal Hit approach implemented in the newly designed Forty-Two software (see details in Supplemental Experimental Procedures). We then discarded

clusters with ≤50 species or ≤8 out of 10 major taxonomic groups (Bilateria, Anthozoa, Medusozoa, Ctenophora, Demospongiae, Hexactinellida, Calcarea, Homoscleromorpha, Placozoa, outgroup), thus retaining only 3,414 enriched orthologous clusters.

Paralogy Treatment and Removal of Contaminants

At this stage, despite considerable efforts to remove ancient paralogs and contaminants, some contaminating sequences or recent paralogs were still present in our alignments. That is why we applied several additional filters, based on BLAST similarity searches or on single-gene phylogenetic trees, to identify and remove them.

1) A genuine sequence from one of the ten major clades defined above should be more similar to other sequences of the same clade than to sequences of any other clade because of the long internal branch defining each of these clades. Each sequence was thus BLASTed against the other sequences of the same cluster (only if they were ≥90% complete on the overlapping part and after discarding positions containing ≤ 10% known character states) and sequences were removed when their best hit belonged to a clade other than the expected one. This step eliminated 4,885 sequences.
2) When multiple sequences from the same species are present in a given cluster, the one(s) that is(are) most similar to sequences from the other species is(are) more likely to be orthologs. Hence, for each species having multiple sequences, each sequence was BLASTed against the rest of the alignment and the best hit identified; a sequence was removed if it overlapped with the best hit sequence by ≥95% and if its BLAST score was below the best hit score by a given threshold. Using first a threshold of 25% and then a threshold of 10%, 21,444 and 4,668 sequences were removed, respectively. The resulting clusters were cleaned using HMMCleaner and the same process was repeated, this time removing 7,030 and then 2,279 additional sequences. Most of these sequences were variants of the same transcripts (due to sequencing errors or to in vivo transcript degradation), whereas the others corresponded to distant paralogs, and very few to previously undetected contaminants.
3) Based on a preliminary supermatrix tree built with RAxML using the LG+Γ_4+F model, 12 cnidarian and poriferan species that were incomplete and very closely related to more complete species were discarded, thereby reducing the number of species to 115. Subsequently, all alignments in which ctenophores were no longer represented were discarded. Finally, all alignments that did not contain ≥50 species in ≥8 out of the 10 major clades (see above) were discarded, leaving 3,176 clusters.
4) Paralogous genes in these 3,176 putatively orthologous clusters were discarded in two steps (see details in Supplemental Experimental Procedures): (1) only the 2,424 alignments with at most two of the previously defined major taxonomic groups affected by paralogy were conserved and (2) for each major taxonomic group affected by paralogy in these remaining alignments, we selected the largest set of species without out-paralogy.
5) We further eliminated 15 additional species that were incomplete and very closely related to more complete species, thereby reducing the number of species to 100. All clusters were re-aligned with MAFFT (same parameters as above) and applying the same taxonomic filter as above led us to retain 2,187 clusters.
6) Our last quality check was based on the rationale that non-orthologous sequences (being either a contaminant or a paralog) usually display very long branches when constrained on the species tree because they are misplaced. First, alignments were cleaned with HMMCleaner and BMGE and concatenated using SCaFoS [46]. The phylogeny inferred using RAxML from the supermatrix under the LG+Γ_4+F model was considered as a proxy of the species tree (note that ctenophores were sister to all other metazoans in this tree). Then, for each alignment, the reference topology was pruned of the species missing in that alignment, and branch lengths on this constrained topology were estimated using RAxML (LG+Γ_4+F model). This allowed us to compare terminal branch lengths observed in the single-gene tree to those observed in the pruned supermatrix tree and to remove sequences for which the branch-length ratio was >5, thereby eliminating 928 individual sequences. We repeated the same protocol, now computing the Pearson correlation coefficient R^2 between branch lengths in each single-gene tree and the corresponding pruned supermatrix tree. We obtained a mean R^2 of 0.797 and a standard deviation (SD) of 0.090, which led us to remove 74 clusters showing a R^2 outside the interval [0.6209, 0.9730] (i.e., the mean ± 1.96 SD). These included, for instance, a cluster in which a gene from a bacterium used for choanoflagellate culture was present in the transcriptomes of two closely related choanoflagellates.
7) Since missing data increases computational time and LBA artifacts [47], we removed three species that had >85% missing data. More importantly, to retain only genes that potentially bear phylogenetic information on the relative position of Ctenophora and Porifera, we now defined four major groups (Bilateria+Cnidaria+Placozoa, Ctenophora, Porifera, and outgroups) and removed positions that did not have a determined amino acid for ≥3 species in each of these four groups. Last, an alignment was discarded if its length was below 80 amino acid positions, leading to a final set of 1,719 orthologous gene clusters.

Phylogenetic Analyses

Supermatrix Construction

We used SCaFoS [46] to assemble the supermatrix, build chimeras of closely related species (Table S3), and retain only the slowest-evolving sequence when multiple copies were available for a given species (using Tree-Puzzle and the WAG+F model [48] to compute distances). This produced a supermatrix containing 401,632 amino acid positions for 97 species, with an overall amount of 39.3% missing data. A reduced sampling in which distant outgroups were removed resulted in a supermatrix with 90 species and an overall amount of 37.3% missing data.

Evaluation of Congruence

Nine phylogenomic datasets (our dataset and [1, 5–10]) were evaluated for their internal congruence. Single-gene trees were inferred with RAxML [49] under the LG+Γ_4 model, after discarding species with >50% missing data. The corresponding supermatrix trees were either retrieved from the original publications when possible or computed as for single-gene trees. For each gene, missing species were removed from the supermatrix tree, and the percent of bipartitions in agreement between each single-gene tree and its pruned supermatrix was computed. Lastly, we computed the mean percent of bipartition in agreement across single-gene trees for each dataset.

Model Testing

Bayesian cross-validation [50] implemented in PhyloBayes 3.3 [51] was used to compare the fit of the site-homogeneous LG and GTR models and of the site-heterogeneous CAT model. Ten replicates were considered, each one consisting of a random subsample of 10,000 sites for training the model and 2,000 sites for computing the cross-validation likelihood score.

Site-Homogeneous Model

Maximum likelihood analyses were run on the full dataset (i.e., 401,632 amino acid positions) using RAxML [49]. A partition was attributed to each gene and each partition was given an independent LG+Γ_4+F model. Such a partitioned analysis allows the alpha parameter of the gamma distribution and stationary frequencies of amino acids to vary across genes. See Supplemental Experimental Procedures for details on differences between site-homogeneous and site-heterogeneous models.

Site-Heterogeneous Model

Since the better fitting site-heterogeneous CAT model is very time consuming, we analyzed the dataset using a jackknifing strategy with 100 replicates. Each replicate was built by randomly sampling genes until >100,000 positions were obtained (equivalent to ~430 genes per replicate). Jackknife replicates (cleared of constant sites) were analyzed with PhyloBayes MPI 1.6j [52] under the CAT+Γ_4 model, until 6,000 cycles were obtained. Convergence of the parameters was assessed using criteria given in the PhyloBayes manual and a conservative burn-in of 3,000 cycles was used for all replicates.

Computation of Likelihoods for Various Models and Topologies

The relative fits of various available sequence evolution models, including several that aim to model site heterogeneity, were computed on four topologies (see Table S1) with iqtree [53]. This was done for three different taxon

samplings in order to observe the impact of progressive removal of poriferan clades on the position of hexactinellids and ctenophores. The complete taxon sampling corresponds to our 90 species supermatrix.

Removal of Heteropecillous Positions

Since heteropecillous positions (i.e., sites with a substitution process that is heterogeneous in time) violate the assumptions of the CAT model, they may lead to systematic error [32]. In order to account for this, we used the protocol of Roure and Philippe [32] to compute the level of heteropecilly of each position using five pre-defined clades (Choanoflagellatea, Porifera, Ctenophora, Cnidaria, and Bilateria). We then used the CAT+Γ_4 model to analyze the datasets obtained after removal of 60% and 70% of the most heteropecillous positions (136,618 and 102,464 remaining variable positions, respectively; Figure S4).

Testing the Impact of Compositional Bias

In order to reduce potential impact of saturation and compositional bias, we recoded our supermatrix using the Dayhoff 6-states alphabet corresponding to amino acid groups [54, 55], which we then analyzed with the CAT+Γ_4 and CAT+Γ_4+GTR models. This recoding did not affect in any way the deep metazoan relationships inferred in this study, as sponges were always recovered with maximal support (PP = 1) as sister group to all other metazoans, although slight incongruences within choanoflagellates hampered topological convergence of our replicates (data not shown).

Example Analyses of Data Errors in Previous Phylogenomic Datasets

Trees inferred with RAxML [49] under the LG+Γ_4 model from the single genes of ref [7–10] were manually scanned, and one for each study was arbitrarily selected to illustrate the occurrence of erroneous groupings. The original alignments were enriched with data from GenBank (nr) or transcriptomic datasets (retrieved from the NCBI portal) to improve taxonomic sampling and therefore reveal contaminations and/or paralogy. The same positions as in the original studies were selected and trees were inferred with RAxML [49] under the LG+Γ_4 model (Figure S1).

ACCESSION NUMBERS

The data are available under BioProject PRJNA316185 at SRA (accession number SRP072932).

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, four tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2017.02.031.

AUTHOR CONTRIBUTIONS

P.S., H.P., and M.M. designed the study. P.S., M.M., E.Q., A.E., P.L., D.J.R., N.S., and N.K. collected samples. P.S., M.J., and D.J.R. prepared RNA for sequencing. P.S., D.B., F.D., and D.J.R. assembled the transcriptomes. H.P. and P.S. conceived the protocol of data supermatrix assembly including data quality controls, with contribution from M.M.; D.B. wrote the Forty-Two software and H.P. and A.D. debugged it. H.P., P.S., B.R., and D.B. wrote the scripts for the various data quality controls; H.P. and P.S. built the supermatrix. H.P., P.S., G.W., and E.C. performed the phylogenetic analyses. P.S. made the figures and tables. M.M. drafted the manuscript main text and P.S. the Experimental Procedures; all other authors amended the manuscript and approved the final version.

ACKNOWLEDGMENTS

We thank Alain Goyeau's diving team for their help collecting antipatharians in Guadeloupe. We thank the UPMC biological stations of Banyuls-sur-Mer and Villefranche-sur-Mer, as well as Olivier Gros (Université des Antilles, Pointe-à-Pitre) for providing lab facilities. Funding was mainly from the Institut Universitaire de France (M.M. junior membership 2009–2014) and from the TULIP Laboratory of Excellence (ANR-10-LABX-41) to H.P. Computations were made on the supercomputers Mp2 and Ms2 from the Université de Sheerbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI), and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT). The Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities also provided some computational resources. D.J.R. was supported by a National Defense Science and Engineering Graduate fellowship from the United States Department of Defense, a National Science Foundation Central Europe Summer Research Institute Fellowship, a Chang-Lin Tien Fellowship in Environmental Sciences and Biodiversity, a postdoctoral fellowship from the Conseil Régional de Bretagne, and the French Government "Investissements d'Avenir" program OCEANOMICS (ANR-11-BTBR-0008). G.W. was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft [DFG]) and the Ludwig-Maximilians-Universität München LMUexcellent program (Project MODELSPONGE) through the German Excellence Initiative. This is publication ISEM 2017-043 of the Institut des Sciences de l'Evolution de Montpellier.

Received: December 2, 2016
Revised: February 7, 2017
Accepted: February 13, 2017
Published: March 16, 2017

REFERENCES

1. 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*, 706–712.
2. Pick, K.S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D.J., Wrede, P., Wiens, M., Alié, A., Morgenstern, B., Manuel, M., and Wörheide, G. (2010). Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. *27*, 1983–1987.
3. Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Müller, W.E., Nickel, M., Schierwater, B., et al. (2013). Deep metazoan phylogeny: when different genes tell different stories. Mol. Phylogenet. Evol. *67*, 223–233.
4. Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., Lartillot, N., and Wörheide, G. (2015). Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. USA *112*, 15402–15407.
5. Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., et al. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. Nature *452*, 745–749.
6. Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G.W., Edgecombe, G.D., Martinez, P., Baguñà, J., Bailly, X., Jondelius, U., et al. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. Biol. Sci. *276*, 4261–4270.
7. Ryan, J.F., Pang, K., Schnitzler, C.E., Nguyen, A.D., Moreland, R.T., Simmons, D.K., Koch, B.J., Francis, W.R., Havlak, P., Smith, S.A., et al.; NISC Comparative Sequencing Program (2013). The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science *342*, 1242592.
8. Moroz, L.L., Kocot, K.M., Citarella, M.R., Dosung, S., Norekian, T.P., Povolotskaya, I.S., Grigorenko, A.P., Dailey, C., Berezikov, E., Buckley, K.M., et al. (2014). The ctenophore genome and the evolutionary origins of neural systems. Nature *510*, 109–114.
9. Borowiec, M.L., Lee, E.K., Chiu, J.C., and Plachetzki, D.C. (2015). Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics *16*, 987.
10. Whelan, N.V., Kocot, K.M., Moroz, L.L., and Halanych, K.M. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. USA *112*, 5773–5778.
11. Chang, E.S., Neuhof, M., Rubinstein, N.D., Diamant, A., Philippe, H., Huchon, D., and Cartwright, P. (2015). Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proc. Natl. Acad. Sci. USA *112*, 14912–14917.

12. Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. *9*, e1000602.
13. Moroz, L.L. (2015). The genealogy of genealogy of neurons. Commun. Integr. Biol. *7*, e993269.
14. Ryan, J.F., and Chiodin, M. (2015). Where is my mind? How sponges and placozoans may have lost neural cell types. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 370.
15. Marlow, H., and Arendt, D. (2014). Evolution: ctenophore genomes and the origin of neurons. Curr. Biol. *24*, R757–R761.
16. Jékely, G., Paps, J., and Nielsen, C. (2015). The phylogenetic position of ctenophores and the origin(s) of nervous systems. Evodevo *6*, 1.
17. Leys, S.P. (2015). Elements of a 'nervous system' in sponges. J. Exp. Biol. *218*, 581–591.
18. Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. *21*, 1095–1109.
19. Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. Stat. Comput. *10*, 63–72.
20. Whelan, N.V., and Halanych, K.M. (2016). Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. Syst. Biol. Published online September 14, 2016. http://dx.doi.org/10.1093/sysbio/syw084.
21. Philippe, H., Lartillot, N., and Brinkmann, H. (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. *22*, 1246–1253.
22. Schneider, A., and Cannarozzi, G.M. (2009). Support patterns from different outgroups provide a strong phylogenetic signal. Mol. Biol. Evol. *26*, 1259–1272.
23. Gouy, R., Baurain, D., and Philippe, H. (2015). Rooting the tree of life: the phylogenetic jury is still out. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140329.
24. Philippe, H., and Roure, B. (2011). Difficult phylogenetic questions: more data, maybe; better methods, certainly. BMC Biol. *9*, 91.
25. Wörheide, G., Dohrmann, M., Erpenbeck, D., Larroux, C., Maldonado, M., Voigt, O., Borchiellini, C., and Lavrov, D.V. (2012). Deep phylogeny and evolution of sponges (phylum Porifera). Adv. Mar. Biol. *61*, 1–78.
26. Simion, P., Bekkouche, N., Jager, M., Quéinnec, E., and Manuel, M. (2015). Exploring the potential of small RNA subunit and ITS sequences for resolving phylogenetic relationships within the phylum Ctenophora. Zoology (Jena) *118*, 102–114.
27. Zapata, F., Goetz, F.E., Smith, S.A., Howison, M., Siebert, S., Church, S.H., Sanders, S.M., Ames, C.L., McFadden, C.S., France, S.C., et al. (2015). Phylogenomic analyses support traditional relationships within Cnidaria. PLoS ONE *10*, e0139068.
28. Halanych, K.M. (2015). The ctenophore lineage is older than sponges? That cannot be right! Or can it? J. Exp. Biol. *218*, 592–597.
29. Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. *7* (*Suppl 1*), S4.
30. Collins, A.G. (1998). Evaluating multiple alternative hypotheses for the origin of Bilateria: an analysis of 18S rRNA molecular evidence. Proc. Natl. Acad. Sci. USA *95*, 15458–15463.
31. Srivastava, M., Begovic, E., Chapman, J., Putnam, N.H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M.L., et al. (2008). The Trichoplax genome and the nature of placozoans. Nature *454*, 955–960.
32. Roure, B., and Philippe, H. (2011). Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol. Biol. *11*, 17.
33. Dunn, C.W., Leys, S.P., and Haddock, S.H. (2015). The hidden biology of sponges and ctenophores. Trends Ecol. Evol. *30*, 282–291.
34. Jager, M., and Manuel, M. (2016). Ctenophores: an evolutionary-developmental perspective. Curr. Opin. Genet. Dev. *39*, 85–92.
35. Boussau, B., Szöllosi, G.J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. Genome Res. *23*, 323–330.
36. Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.
37. Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189.
38. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.
39. Amemiya, C.T., Alföldi, J., Lee, A.P., Fan, S., Philippe, H., Maccallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., et al. (2013). The African coelacanth genome provides insights into tetrapod evolution. Nature *496*, 311–316.
40. Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. *10*, 210.
41. Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688–2690.
42. Chevreux, B. (2005). MIRA: An Automated Genome and EST Assembler (Ruprecht-Karls-University).
43. Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res. *9*, 868–877.
44. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. *8*, 1494–1512.
45. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., et al. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics *30*, 1660–1666.
46. Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol. Biol. *7* (*Suppl 1*), S2.
47. Roure, B., Baurain, D., and Philippe, H. (2013). Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. *30*, 197–214.
48. Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics *18*, 502–504.
49. Stamatakis, A., and 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.
50. Stone, M. (1974). Cross validatory choice and assessments of statistical predictions. J. R. Stat. Soc. B *36*, 111–117.
51. Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics *25*, 2286–2288.
52. Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. *62*, 611–615.
53. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.
54. Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., and Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. *56*, 389–399.
55. Hrdy, I., Hirt, R.P., Dolezal, P., Bardonová, L., Foster, P.G., Tachezy, J., and Embley, T.M. (2004). *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature *432*, 618–622.