Cladistics 29 (2013) 435–448

Cladistics

10.1111/cla.12015

Phylogenomics of Annelida revisited: a cladistic approach using genome-wide expressed sequence tag data mining and examining the effects of missing data

Sebastian Kvist[a,b,*,†] and Mark E. Siddall[b,c]

[a]*Richard Gilder Graduate School, American Museum of Natural History, Central Park West at 79th Street, New York, NY, 10024, USA;* [b]*Division of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, New York, NY, 10024, USA;* [c]*Sackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West at 79th Street, New York, NY, 10024, USA;* [†]*Present address: Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA, 02138, USA*

Accepted 31 December 2012

Abstract

We present phylogenomic analyses of the most comprehensive molecular character set compiled for Annelida and its constituent taxa, including over 347 000 aligned nucleotide sites for 39 taxa. The nucleotide data set was recovered using a pre-existing amino acid data set of almost 48 000 aligned sites as a backbone for tBLASTn searches against NCBI. In addition, orthology determinations of the loci in the original amino acid data set were scrutinized using an All vs All Reciprocal Best Hit approach, employing BLASTp, and examining for statistical interdependency among the loci. This approach revealed considerable sequence redundancy among the loci in the original data set and a new data set was compiled, with the redundancy removed. The newly compiled nucleotide data set, the original amino acid data set, and the new reduced amino acid data set were subjected to parsimony analyses and two forms of bootstrap resampling. The last-named data set also was analysed using a maximum-likelihood approach. There were two main objectives to these analyses: (i) to examine the general topology, including support, resulting from the analyses of the new data sets and (ii) to assess the consistency of the branching patterns across optimality criteria by comparison with previous probabilistic approaches. The phylogenetic hypotheses resulting from analyses of the three data sets are largely unsupported, reflecting the continued difficulty of finding numerous, reliable, and suitable loci for a group as ancient as Annelida. Resulting parsimonious hypotheses disagree, in some respects, with the previous probabilistic approaches; Sedentaria and, in most cases, Errantia are not supported as monophyletic groups but Pleistoannelida is recovered as a (unsupported) monophyletic group in one of the three parsimony analyses as well as the likelihood analysis. In addition, we performed missing data titration studies to estimate the impact of missing data on overall support and support for specific clades.

Notwithstanding recent efforts to resolve the evolutionary history and phylogenetic relationships of Annelida (e.g. Erséus, 2005; Struck et al., 2007, 2008, 2011; Rousset et al., 2007; Zrzavý et al., 2009), the debates relating to this are still contentious and the hypotheses concerning certain groups are unstable. While the class Clitellata (Hirudinida, oligochaetous clitellates, Branchiobdellida, and Acanthobdellida) is frequently recovered as monophyletic, Hirudinida commonly nests within oligochaetous clitellates, rendering the latter paraphyletic (see also Brinkhurst and Nemec, 1987; Erséus, 1987; Siddall et al., 2001). Moreover, the class Polychaeta, including the vast majority of the 17 210 annelid species currently recognized (Zhang, 2011), is regularly rendered paraphyletic with respect to Clitellata, as well as the "phylum" Sipuncula (see also Struck et al., 2007; Dunn et al., 2008; Dordel

*Corresponding author:
E-mail address: skvist@fas.harvard.edu

et al., 2010), and several of the long-held clades among polychaetous annelids are themselves paraphyletic (see also Fauchald and Rouse, 1997). Equally disconcerting is the fact that the traditionally held order Sedentaria is seldom recovered as monophyletic (Day, 1967; Westheide et al., 1999) and its counterpart, Errantia, has only been recovered as monophyletic in one study, in which only morphological characters were used (Rouse and Fauchald, 1997). Sedentaria has historically referred to annelid worms with a more sessile or semi-sessile lifestyle (but including Clitellata) and with weakly developed, more or less absent parapodia, while Errantia includes worms with a more vagile lifestyle and more well-developed parapodia and chaetae (Perrier, 1897; Westheide et al., 1999; Bartolomaeus et al., 2005). Whereas Errantia is upheld by a number of synapomorphic morphological characters (Bartolomaeus et al., 2005), Sedentaria includes such morphologically disparate taxa that establishing homologies is often difficult, if not impossible. Until Fauchald's (1977) treatment of the two orders, there had also been doubt as to whether these different modes of life accurately reflected common ancestry (e.g. Day, 1967). The taxonomy behind these groups was probably a matter of convenience as opposed to their being reflective of true evolutionary relationships and, as a result, Fauchald (1977) eliminated the Errantia and Sedentaria dichotomy, whilst erecting 17 new, taxonomically equivalent groups, which were later expanded to include more orders (Rouse and Fauchald, 1997; Rouse and Pleijel, 2001; Bartolomaeus et al., 2005).

Until recently, the phylogenetic analyses of molecular characters that underlie some of the taxonomic changes mentioned above have been based on only a small number of different loci (e.g. McHugh, 1997, 2000; Bleidorn et al., 2003; Bely and Wray, 2004; Erséus and Källersjö, 2004; Rousset et al., 2007), which may have contributed to inconsistency in the monophyletic groups recovered. To address this and other issues, Struck et al. (2011) compiled a large data set of expressed sequence tags (ESTs), including 231 loci for 39 taxa across Annelida. Their phylogenetic analyses, based on Bayesian inference and maximum-likelihood (ML) estimations of almost 48 000 aligned amino acid sites, recovered both Errantia and Sedentaria as monophyletic groups (albeit somewhat taxonomically modified by Struck et al., 2011). While their rediscovered monophyly may be reason enough to re-erect these old groupings, we believe that some reconsiderations of the analyses are pertinent before restoring a previous classification of such a large group of organisms. Chiefly, we investigate whether a parsimony analysis of the nucleotides coding for the amino acids used by Struck et al. (2011) results in the same hypothesis. In addition, we examine whether locus interdependencies could lead to artificial clades or at least artificially inflated support values for clades. The large data set compiled by Struck et al. (2011) also lends itself well to a timely discussion on orthology statements and a desire for consistency of results across methods.

To this end, we here recovered orthologous nucleotide sequences for each locus and taxon using the amino acids as a source for targeted searches, and analyse this larger data set (as well as the amino acid data set independently) under a parsimony, as opposed to probabilistic, framework.

Material and methods

Data set reconstruction

The amino acid alignment compiled by Struck et al. (2011) (47 953 aligned sites for 39 taxa) was parsed and transferred to 231 separate files, each representing a single locus as defined in Supplementary Table 6 in Struck et al. (2011). The nucleotide data set then was compiled using the individual loci as queries for independent BLAST searches as described below.

Using BLAST client 3 (NCBI), the separate loci were compared remotely against both the EST database and the non-redundant (nr) sequence database at NCBI using a tBLASTn protocol (searching translated nucleotide databases using a protein query) and employing a cutoff *e*-value of $1E^{-5}$, retaining the best hit for each query, insofar as the best hit in a database should represent the sequence used by Struck et al. (2011). Data were obtained in this way for each of the following taxa—Polychaeta: *Alvinella pompejana* (Alvinellidae), *Arenicola marina* (Arenicolidae), *Cirratulus* sp. (Cirratulidae), *Eulalia clavigera* (Phyllodocidae), *Eurythoe complanata* (Amphinomidae), *Flabelligera affinis* (Flabelligeridae), *Glycera tridactyla* (Glyceridae), *Lanice conchilega* (Terebellidae), *Lumbrineris zonata* (Lumbrineridae), *Malacoceros fuliginosus* (Spionidae), *Onuphis iridescens* (Onuphidae), *Ophelia limacina* (Opheliidae), *Pectinaria koreni* (Pectinariidae), *Platynereis dumerilii* (Nereididae), *Pomatoceros lamarckii* (Serpulidae), *Ridgeia piscesae* (Siboglinidae), *Scoloplos armiger* (Orbiniidae), *Sthenelais boa* (Sigalionidae), and *Typosyllis pigmentata* (Syllidae); Clitellata: *Eisenia andrei* (Lumbricidae), *Eisenia fetida* (Lumbricidae), *Haementeria depressa* (Glossiphoniidae), *Hirudo medicinalis* (Hirudinidae), *Lumbricus rubellus* (Lumbricidae), *Perionyx excavatus* (Megascolecidae), and *Tubifex tubifex* (Naididae); Bivalvia: *Crassostrea gigas* (Ostreidae); Myzostomida: *Myzostoma cirriferum* (Myzostomidae); and Sipuncula: *Sipunculus nudus* (Sipunculidae). In cases where the best hit among the targets did not match the taxonomic identities of the queries, a new tBLASTn search was performed on the NCBI website using the Entrez option to confine the search to the respective taxon.

Complementary to the foregoing, the amino acid sequences for each of the following taxa were compared against the annotated genomes of the respective taxon using a local tBLASTn search—Polychaeta: *Capitella teleta* (Capitellidae); Clitellata: *Helobdella robusta* (Glossiphoniidae); and Gastropoda: *Lottia gigantea* (Lottiidae). Again, the search employed a cutoff *e*-value of $1E^{-5}$ and the best matching hit was retrieved. The nucleotide sequences from the genomic hits were extracted and added to the data set acquired from the EST and nr databases.

In addition, tBLASTn searches ($1E^{-5}$ cutoff) were performed against assembled trace archive data for each of—Polychaeta: *Chaetopterus variopedatus* (Chaetopteridae) and *Urechis caupo* (Urechidae); Myzostomida: *Myzostoma seymourcollegiorum* (Myzostomidae); Ectoprocta: *Bugula neritina* (Bugulidae); Nemertea: *Cerebratulus lacteus* (Cerebratulidae); Brachiopoda: *Terebratalia transversa* (Laqueidae); and Sipuncula: *Themiste lageniformes* (Themistidae). To ensure maximum correspondence, the exact same trace archive assemblies as used by Struck et al. (2011) were used as targets. Nucleotide sequences from the best hit for each locus and taxon were extracted and added to the total data set now consisting of data from all four sources. All of these data, including the entire nucleotide data set as well as hit descriptions for each of the tBLASTn searches, are available as Supplementary material.

Repeat masking

All of the data retrieved from the tBLASTn searches were imported to and indexed in a database created in FileMaker Pro ver. 5 (FileMaker Inc., Santa Clara, CA, USA). Polyadenosine tails were identified and, in cases where they were longer than four bases, each was truncated at the upstream polyadenylation signal (AATAA; Zaret and Sherman, 1982). In addition, homopolymer leading and trailing sequences were excised where they occurred in multiples of two or more (largely in *Platynereis dumerilii*). Other dinucleotide, trinucleotide, and tetranucleotide repeat patterns also were identified with FileMaker Pro and excised using the RepBase repeats library for the nematode *Caenorhabditis elegans* (Rhabditidae) as implemented in the software EGassembler (Masoudi-Nejad et al., 2006) employing the "slow" option with a default cutoff score of 225.

Alignment and phylogenetic analyses

Sequences for each locus in the nucleotide data set were aligned independently using MAFFT (Katoh et al., 2005) on the European Bioinformatics Institute website employing a gap-opening cost of three and default settings for all other parameters. Each of (i) the joined nucleotide data set, (ii) the amino acid data set from Struck et al. (2011), and (iii) the dataset following from the removal of some redundant loci (retaining the representative locus with the highest taxonomic coverage; see below) then were subjected to parsimony analyses using TNT (Goloboff et al., 2008). New Technology searches were conducted employing sectorial searches, with the tree fusing and ratcheting algorithms turned on. Trees were retrieved by a driven search using 100 initial addition sequences and requiring that the minimum length tree be found at least five times. All characters were equally weighted and nonadditive, and gaps were treated as missing data. The results of the New Technology searches were subsequently resubmitted to TNT for TBR branch swapping. Support values for nodes also were estimated in TNT through both standard bootstrap resampling and partition (i.e. locus) bootstrapping (Siddall, 2009). Both bootstrap analyses employed 100 iterations, each subjected to five iterations of ratcheting and three rounds of tree fusing after an initial five rounds of Wagner tree building.

To elucidate whether the differences in results are contingent on the optimality criteria used or differences in the signal of the data, ML analyses were carried out using RAxML ver. 7.3.1. (Stamatakis, 2006) on the data set with redundant loci removed. A heuristic search was performed using the PROTCAT model of amino acid evolution and employing the JTTF transition matrix. Runs were performed for 1000 iterations with an initial 25 CAT rate categories and final optimization using four gamma shape categories. Support values were estimated from 1000 pseudoreplicates with a different starting tree for each replicate. All trees presented here were rooted at *Bugula neritina* following Struck et al. (2011).

Examination of interdependent loci

An "all vs. all" reciprocal BLAST or bi-directional best hit search (Ge et al., 2005; Fang et al., 2010) using a BLASTp protocol (searching protein databases using a protein query) with a cutoff *e*-value of $1E^{-20}$ was performed on the amino acid loci to determine the level of similarity and coverage between them (hits that were less similar than $1E^{-20}$ were also retained separately to be used in the next step). Using the loci that showed hits at $1E^{-20}$ as a guide, all data (i.e. even hits less similar than $1E^{-20}$ but excluding self-hits) both for intra-locus (within a particular locus) and for inter-locus (between loci) similarity values were compiled. In turn, averages, standard deviations, and range intervals, as well as 95% confidence intervals were calculated both for intra-locus and for inter-locus values. The putative overlap between the 95% confidence interval of intra-locus and inter-locus range values was assessed assuming that an overlap between 95% confidence intervals of e-value ranges indicated orthologous gene families or redundant use of sequences in more than one locus.

Missing data

To assess the level at which the taxa grouped together based on mere presence or absence of data (and the support values related to this), rather than true phylogenetic signal, all positions in the original amino acid data set were changed to an "A" such that the only information present in the data set consisted of patterns of presence or absence of data. The parsimony analysis then was re-run using the same parameters as mentioned above but now treating gaps as a fifth state and supported clades were crosschecked against clades present in the Bayesian tree found by Struck et al. (2011).

Complementary to this, missing data titrations were performed on the data set that had redundant loci already removed by considering the relative representation of characters for each taxon or the relative representation of taxa for each character and then evaluating clade support with a resampling scheme. That is, for characters, parsimony bootstrap support values were calculated (using the same parameters as mentioned above) for various main clades when characters were missing for a given percentage of the taxa (e.g. Oakley et al., 2013). This was performed in increments of 5% such that when characters were missing for 100, 95, 90 […] 25% of the taxa, they were removed; resampling was performed at each increment. A similar strategy was then applied for the exclusion of taxa for which a certain number of characters were absent (i.e. a taxon was completely excluded from the analysis if it was lacking 100, 85, 80, […] 55% of the characters). Again, parsimony bootstrap resampling was applied at each increment and the resulting values were investigated. Scripts for implementation of our titration scheme (excludemiss and deactmiss) are available on the TNT Wiki (http://tnt.insectmuseum.org/index.php/Scripts).

Results

Data independency

In the results of the BLASTp search, loci were deemed orthologous if the 95% confidence interval of the e-value range for hits between sequences from at least two different loci overlapped with that of the range found within a locus. A total of 23 loci (9.96%) satisfied this criterion. Some of the redundancy was conjoined in triple-locus hits (i.e. if locus A = locus B and locus B = locus C then, transitively, locus A = locus C). A total of 13 loci (5.63%) were found to be redundant in that they are already represented by ten other orthologues. Figure 1 illustrates this phenomenon using real similarity values from the data set generated by Struck et al. (2011). Specifically, non-redundant loci 96 and 143 each demonstrate intra-locus similarity values that do not significantly overlap with values obtained from comparisons between loci 96 and 143 (Fig. 1a). In contrast, the inter-locus similarities for loci 224 and 225 are indistinguishable from

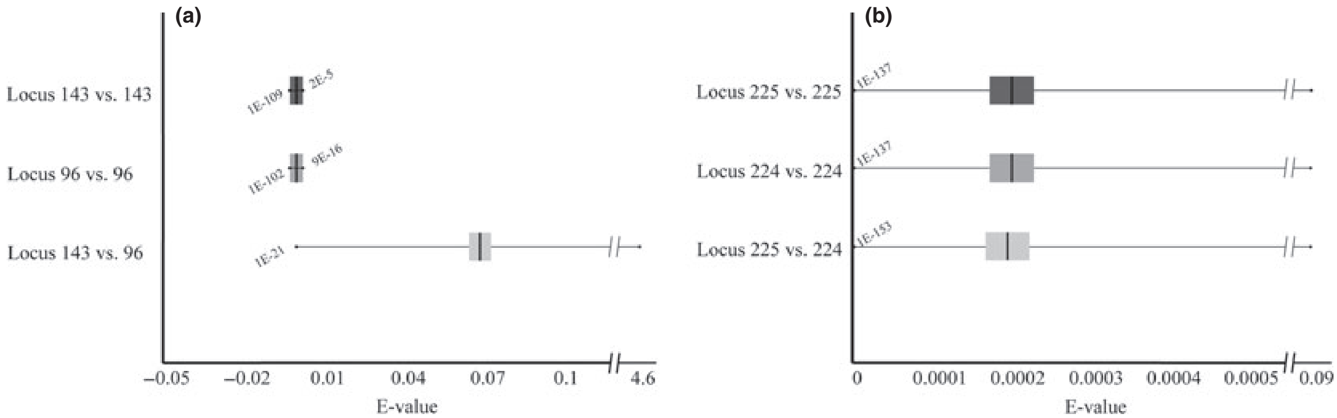

Fig. 1. Two examples of the calculations of overlap between ranges of e-values within a 95% confidence interval: (a) an instance in which the 95% confidence interval of similarity values between two loci does not overlap with that of similarity values within each of those loci and (b) a case in which these values do overlap. Black horizontal lines bind the minimum and maximum e-values (i.e. ranges); shaded areas indicate the 95% confidence interval; and vertical lines within the shading indicate the average e-values for comparisons within and between the loci. Broken lines denote a compression of the actual ranges in order to fit the values within the ranges of the *x*-axes. Lower bound values of the e-value ranges, and in some cases also upper bound values, are printed as they are very close to 0. All e-values were recovered using a BLASTp protocol and the remaining comparisons for other loci are presented in the text.

intra-locus values (Fig. 1b), illustrating their mutual redundancy. The following loci (with the single redundant locus retained for analysis appearing in parentheses) were removed from subsequent phylogenetic analyses: locus 14 (locus 7); locus 16 and locus 136 (locus 5); locus 32 and locus 109 (locus 35); locus 64 (locus 120); locus 125 and locus 137 (locus 134); locus 126 (locus 50); locus 129 (locus 74); locus 141 (locus 59); locus 212 (locus 53); locus 224 (locus 225).

Phylogenetic analyses

While a lenient e-value of $1E^{-5}$ was used as a cutoff for the tBLASTn searches, in almost all cases, resulting *e*-values were much lower (i.e. better matching) than $1E^{-20}$. With a rather high incidence, however, the best BLAST-hits for the amino acid sequences represented a different taxon than the query (see Supplementary material), requiring an additional tBLASTn search using the Entrez-option to confine the search to the same taxon. All of the individual BLAST searches for *Onuphis iridescens* matched *Lumbrineris zonata*, and vice versa, at a lower e-value due to a confusion of these taxa (T. Struck pers. comm.) in the final data set used by Struck et al. (2011). That is, the NCBI submissions for these taxa were correct, whereas there is an error in the final data set used by Struck et al. (2011); the TreeBase submission for this data set now includes a disclaimer noting this confusion. In the present study, this was remedied by extracting the best hit for the opposite taxon in every case. As an aside, Supplementary Table 6 in Struck et al. (2011) states that data for *Myzostoma seymourcollegiorum*, *Platynereis dumerilii*, and *Glycera tridactyla* are available in the NCBI EST database but, rather, data for *M. seymorcollegiorum* are found in NCBI trace archives and for the latter two in the NCBI nr database. Hits equal to or lower than $1E^{-5}$ were not found for *Glycera tridactyla* at loci 163, 164, 171, 175, 187 and 189 or for *Sthenelais boa* at locus 228. Either no sequence (or an erroneous one) was deposited in NCBI for those taxa at those loci or the translation from nucleotides to amino acids by Struck et al. (2011) was not equivalent. For information on data coverage for each taxon, see Supplementary material in Struck et al. (2011).

The final nucleotide data set consisted of 347 298 aligned sites, 118 163 of which were parsimony informative. The parsimony analysis of this data set returned a single most-parsimonious tree with 932 748 steps (Fig. 2). Support values were low across the entire topology of the tree, with the exception of support values relating to Clitellata, which were relatively high. Both of the molluscs (*Lottia gigantea* and *Crassostrea gigas*) nest within Annelida (standard bootstrap support; BS < 50%, partition bootstrap support; PBS < 50%), rendering the phylum non-monophyletic. Neither Errantia nor Sedentaria is monophyletic by virtue of taxa from each group nesting within the other. Clitellata is recovered as a monophyletic group (BS <50%, PBS 78%) as sister to an unsupported clade containing *Alvinella pompejana*, *Crassostrea gigas,* and *Lottia gigantea*. Neither Canalipalpata (including Terebelliformia, Cirratuliformia, Siboglinidae, Serpulidae, and Spionidae) nor Scolecida (including Capitellidae, Ophellidae, and Arenicolidae) were recovered as monophyletic. The echiuran *Urechis caupo* nests within Annelida as sister to the clade containing *Platynereis*/*Capitella***,** *Pomatoceros*/Myzostomida, *Alvinella*/Mollusca, and Clitellata.

The original amino acid data set used by Struck et al. (2011) consisted of 47 953 aligned sites, 18 628 of which were parsimony informative. Analysis of that data set returned a single most-parsimonious tree with 110 516 steps (Fig. 3). Again, support values were low across most of the topology, but show relatively high values for clitellate clades. In the tree, both species of *Myzostoma* [these were included in Annelida by Struck et al. (2011), but see their discussion] nest among the outgroup taxa (BS <50%, PBS 86%). Sedentaria is rendered paraphyletic (BS <50%, PBS <50%) with respect to both the monophyletic Errantia (BS <50%, PBS <50%) and another clade containing both the sipunculans (placed together with BS 98%, PBS 99%) and *Chaetopterus variopedatus* (BS <50%, PBS <50%). Clitellata is monophyletic (BS 100%, PBS 100%) but placed as sister to Opheliidae. Within Sedentaria, neither Canalipalpata nor Scolecida is monophyletic. *Urechis caupo* nests well within Annelida, and its position as sister to *Capitella teleta* is supported by BS of 67% and PBS of 96%; this relationship has also been consistently recovered with high support in previous studies (Dunn et al., 2008; Hejnol et al., 2009).

After removing the 13 redundant loci (retaining ten representatives), the data set comprised 44 265 aligned amino acid sites (92.30% of the total data set), 17 555 of which were parsimony informative. The parsimony analysis of these resulted in two equally parsimonious trees, the strict consensus of which is shown in Fig. 4. The tree shows largely the same topology as the analysis of the original amino acid data set that included the redundant loci. Nevertheless, differences exist relative to the tree from the original data set as well as the tree found by Struck et al. (2011). Significantly, Errantia is not recovered as monophyletic: *Ridgeia piscesae* nests within the group. Additionally, *Flabelligera affinis* and *Cirratulus* sp. appear in different parts of the tree when redundant loci are excised (Fig. 4). However, Sedentaria + Errantia [named clade 1 by Struck et al. (2011), see Discussion below] is recovered as monophyletic. Neither the standard bootstrap values nor the partition bootstrap values associated with the tree vary markedly when excluding redundant loci but

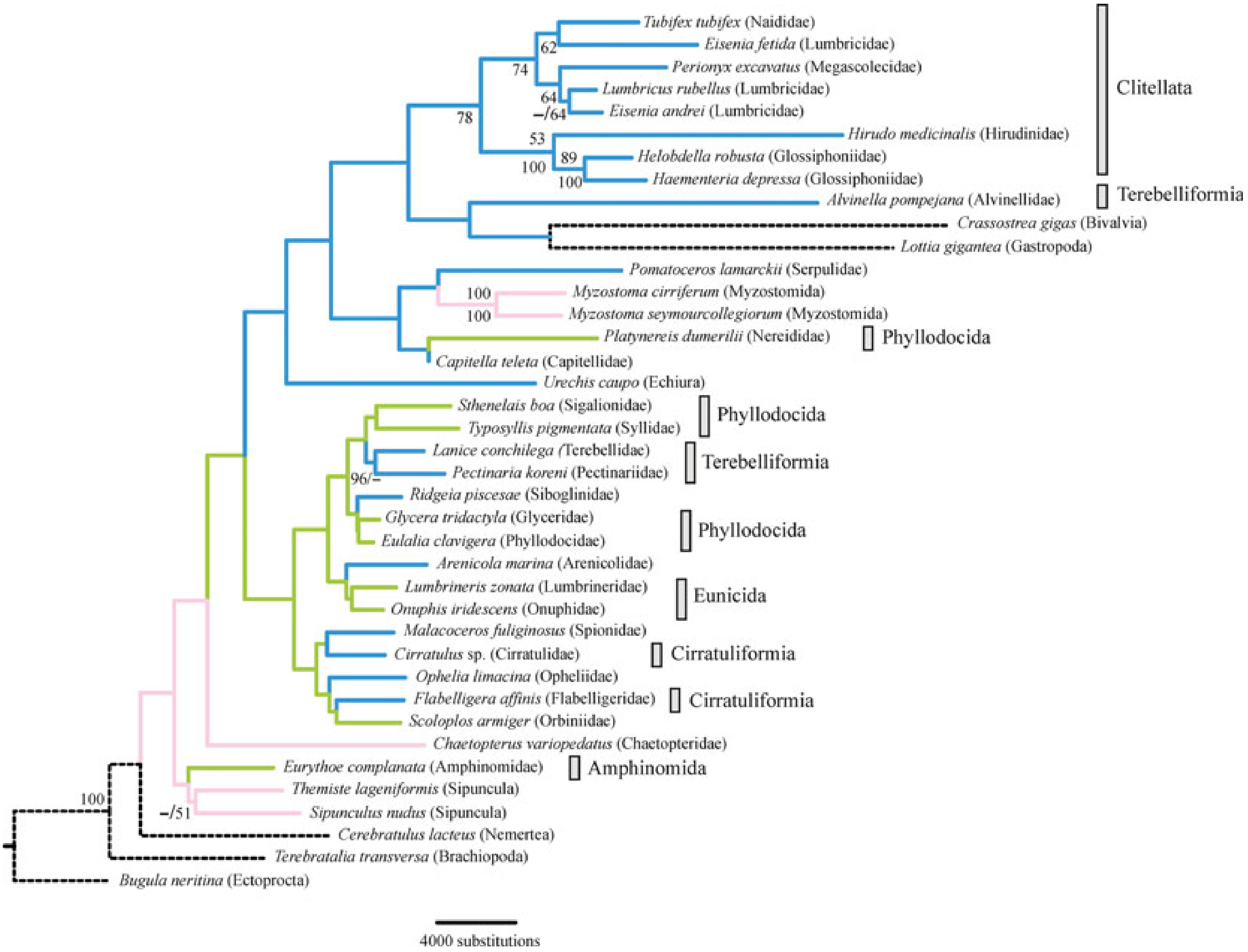
Tubifex tubifex (Naididae)
Eisenia fetida (Lumbricidae)
Perionyx excavatus (Megascolecidae)
Lumbricus rubellus (Lumbricidae)
Eisenia andrei (Lumbricidae)
Hirudo medicinalis (Hirudinidae)
Helobdella robusta (Glossiphoniidae)
Haementeria depressa (Glossiphoniidae)
Alvinella pompejana (Alvinellidae)
Crassostrea gigas (Bivalvia)
Lottia gigantea (Gastropoda)
Pomatoceros lamarckii (Serpulidae)
Myzostoma cirriferum (Myzostomida)
Myzostoma seymourcollegiorum (Myzostomida)
Platynereis dumerilii (Nereididae)
Capitella teleta (Capitellidae)
Urechis caupo (Echiura)
Sthenelais boa (Sigalionidae)
Typosyllis pigmentata (Syllidae)
Lanice conchilega (Terebellidae)
Pectinaria koreni (Pectinariidae)
Ridgeia piscesae (Siboglinidae)
Glycera tridactyla (Glyceridae)
Eulalia clavigera (Phyllodocidae)
Arenicola marina (Arenicolidae)
Lumbrineris zonata (Lumbrineridae)
Onuphis iridescens (Onuphidae)
Malacoceros fuliginosus (Spionidae)
Cirratulus sp. (Cirratulidae)
Ophelia limacina (Opheliidae)
Flabelligera affinis (Flabelligeridae)
Scoloplos armiger (Orbiniidae)
Chaetopterus variopedatus (Chaetopteridae)
Eurythoe complanata (Amphinomidae)
Themiste lageniformis (Sipuncula)
Sipunculus nudus (Sipuncula)
Cerebratulus lacteus (Nemertea)
Terebratalia transversa (Brachiopoda)
Bugula neritina (Ectoprocta)
Clitellata
Terebelliformia
Phyllodocida
Phyllodocida
Terebelliformia
Phyllodocida
Eunicida
Cirratuliformia
Cirratuliformia
Amphinomida
62
74
64
–/64
78
53
89
100
100
100
100
96/–
100
–/51
4000 substitutions

Fig. 2. Single most-parsimonious tree recovered from the analysis of the nucleotide data set (L = 932 748; CI = 0.487; RI = 0.236). Standard bootstrap values ≥ 50% are indicated above each node and partition bootstrap values ≥ 50% below each node. Note the low support for most nodes, which is discussed further in the text. Branches are coloured following taxonomic affiliations in Struck et al. (2011): blue = Sedentaria, green = Errantia, red = Annelida but not Errantia or Sedentaria. Dashed lines denote non-annelid taxa and grey bars denote additional taxonomic information. Branch lengths are drawn proportional to change.

it can be noted that the standard bootstrap support for *Ophelia limacina* as sister to Clitellata is decreased to below 50%.

The resulting tree from the likelihood analysis of the redundant data set is shown in Fig. 5. By and large, the tree agrees well with the Bayesian and ML trees shown by Struck et al. (2011), in that both Errantia and Sedentaria are each recovered as monophyletic groups (BS 74 and 71%, respectively). Yet, several of the sister-group relationships change when using the reduced data set. These include *Ophelia limacina* as sister to Clitellata (BS <50%), *Ridgeia piscesae* as sister to the cirratuliform species (BS <50%; this node was unresolved in Struck et al., 2011), *Malacoceros fuliginosus* and *Pomatoceros lamarckii* as sister to the cirratuliforms and *Ridgeia* piscesae (BS <50%), and the myzostomids as sister to the remaining ingroup taxa (BS 76%). By extension, the positioning of *Urechis caupo* and *Capitella teleta* as sister to Clitellata, as recovered by Struck et al. (2011), changes when using the reduced data set. The likelihood bootstrap values are consistently higher than the parsimony bootstrap values, although 11 of the 37 nodes receive likelihood bootstrap values below 75%.

Missing data titration

The topology of the tree based on missing data only (not shown) is completely incongruent with the trees shown here, as well as the Bayesian tree recovered by Struck et al. (2011), affirming that the taxa are not grouping based on mere presence/absence of data.

The parsimony bootstrap results from the titration of characters and that of taxa are presented in

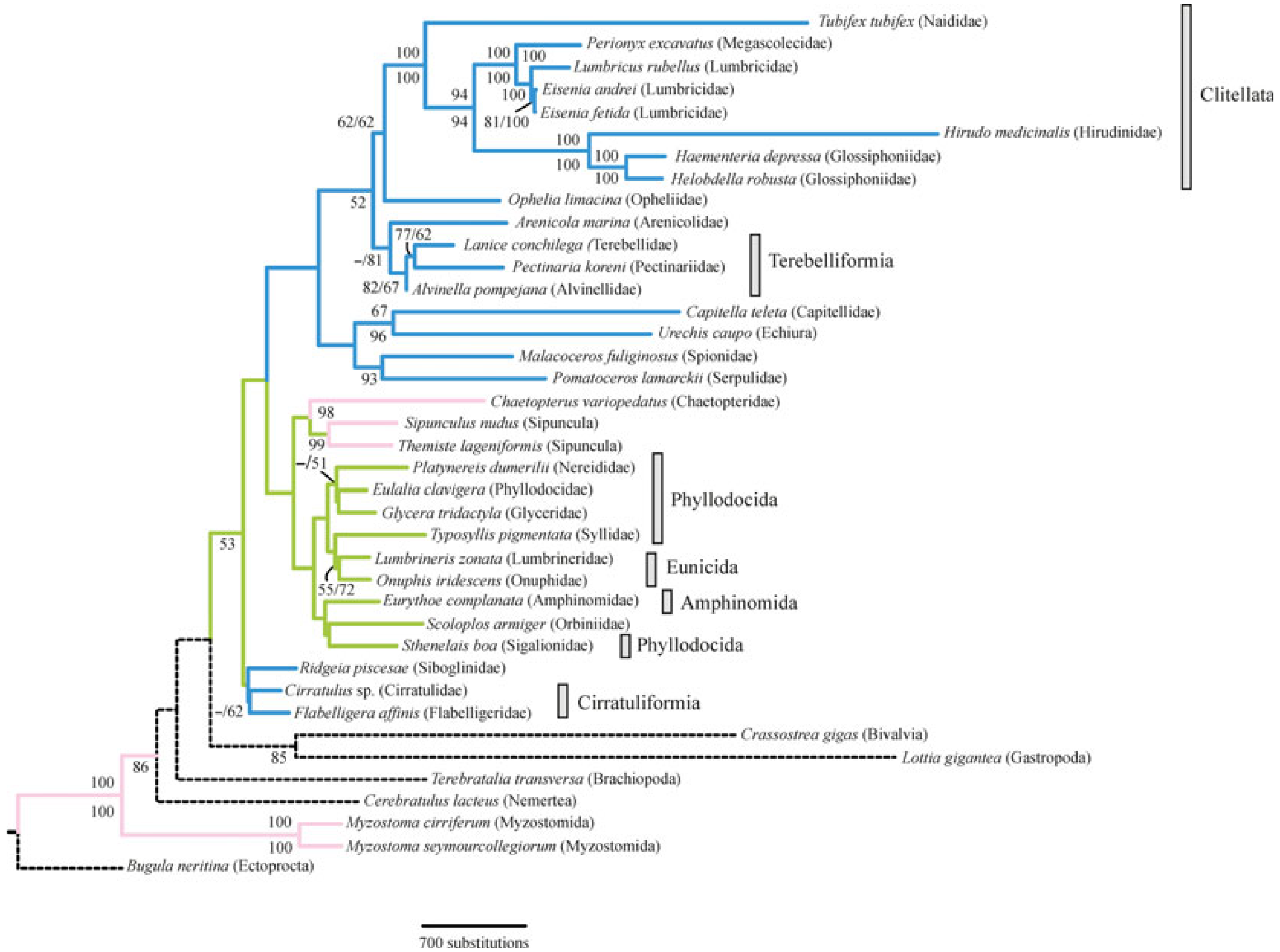
Tubifex tubifex (Naididae)
Perionyx excavatus (Megascolecidae)
Lumbricus rubellus (Lumbricidae)
Eisenia andrei (Lumbricidae)
Eisenia fetida (Lumbricidae)
Hirudo medicinalis (Hirudinidae)
Haementeria depressa (Glossiphoniidae)
Helobdella robusta (Glossiphoniidae)
Ophelia limacina (Opheliidae)
Arenicola marina (Arenicolidae)
Lanice conchilega (Terebellidae)
Pectinaria koreni (Pectinariidae)
Alvinella pompejana (Alvinellidae)
Capitella teleta (Capitellidae)
Urechis caupo (Echiura)
Malacoceros fuliginosus (Spionidae)
Pomatoceros lamarckii (Serpulidae)
Chaetopterus variopedatus (Chaetopteridae)
Sipunculus nudus (Sipuncula)
Themiste lageniformis (Sipuncula)
Platynereis dumerilii (Nereididae)
Eulalia clavigera (Phyllodocidae)
Glycera tridactyla (Glyceridae)
Typosyllis pigmentata (Syllidae)
Lumbrineris zonata (Lumbrineridae)
Onuphis iridescens (Onuphidae)
Eurythoe complanata (Amphinomidae)
Scoloplos armiger (Orbiniidae)
Sthenelais boa (Sigalionidae)
Ridgeia piscesae (Siboglinidae)
Cirratulus sp. (Cirratulidae)
Flabelligera affinis (Flabelligeridae)
Crassostrea gigas (Bivalvia)
Lottia gigantea (Gastropoda)
Terebratalia transversa (Brachiopoda)
Cerebratulus lacteus (Nemertea)
Myzostoma cirriferum (Myzostomida)
Myzostoma seymourcollegiorum (Myzostomida)
Bugula neritina (Ectoprocta)
Clitellata
Terebelliformia
Phyllodocida
Eunicida
Amphinomida
Phyllodocida
Cirratuliformia
100
100
100
100
100
94
94
100
100
81/100
62/62
100
100
100
100
52
77/62
–/81
82/67
67
96
93
98
99
–/51
53
55/72
–/62
85
86
100
100
100
100
700 substitutions

Fig. 3. Single most-parsimonious tree recovered from the analysis of the amino acid data set (L = 110 516; CI = 0.619; RI = 0.347). Standard bootstrap values ≥ 50% are indicated above each node and partition bootstrap values ≥ 50% below each node. Note the low support for most nodes, which is discussed further in the text. Branches are coloured following taxonomic affiliations in Struck et al. (2011): blue = Sedentaria, green = Errantia, red = Annelida but not Errantia or Sedentaria. Dashed lines denote non-annelid taxa and grey bars denote additional taxonomic information. Branch lengths are drawn proportional to change.

Tables 1 and 2, respectively. These titration analyses were carried out using the amino acid data set that already had the redundant loci removed. For the titration of characters (Table 1), the support for Annelida (including Myzostomida) reaches levels above 75% only when characters that are missing for 55–75% of the taxa are deleted. Sedentaria is never recovered as monophyletic, regardless of the amount of data involved. Most often, this is due to the "sedentary" *Ridgeia piscesae* nesting within Errantia instead. Similarly, the support values associated with Errantia are low across the entire character-titration scheme, averaging 2 ± 4.84% in the 13 simulations, the highest support (14%) occurring when those characters are deleted that are missing for 70% of the taxa. Errantia is also not monophyletic in nine of the 13 simulations. The non-monophyletic nature of the group is due to either *Ridgeia piscesae* nesting within the group or, when *R. piscesae* jumps out of the group, its additional exclusion of the "errant" *Scoloplos armiger*. Moreover, the phyllodocid species are rarely recovered as a monophylum and when they are (when those characters are deleted that are missing for 60–70% of the taxa), the maximum bootstrap value recovered is 30%. Interestingly, *Ophelia limacina* is recovered as sister to Clitellata across the titration scheme save for the most stringent iteration, in which those characters are deleted that are missing from 25% of the taxa. However, bootstrap values associated with this clade rarely become relevant, the maximum value being 73% when only 4.4% of the characters remain.

In terms of titrating out taxa missing large amounts of data (Table 2), the average parsimony bootstrap support value remained fairly constant throughout

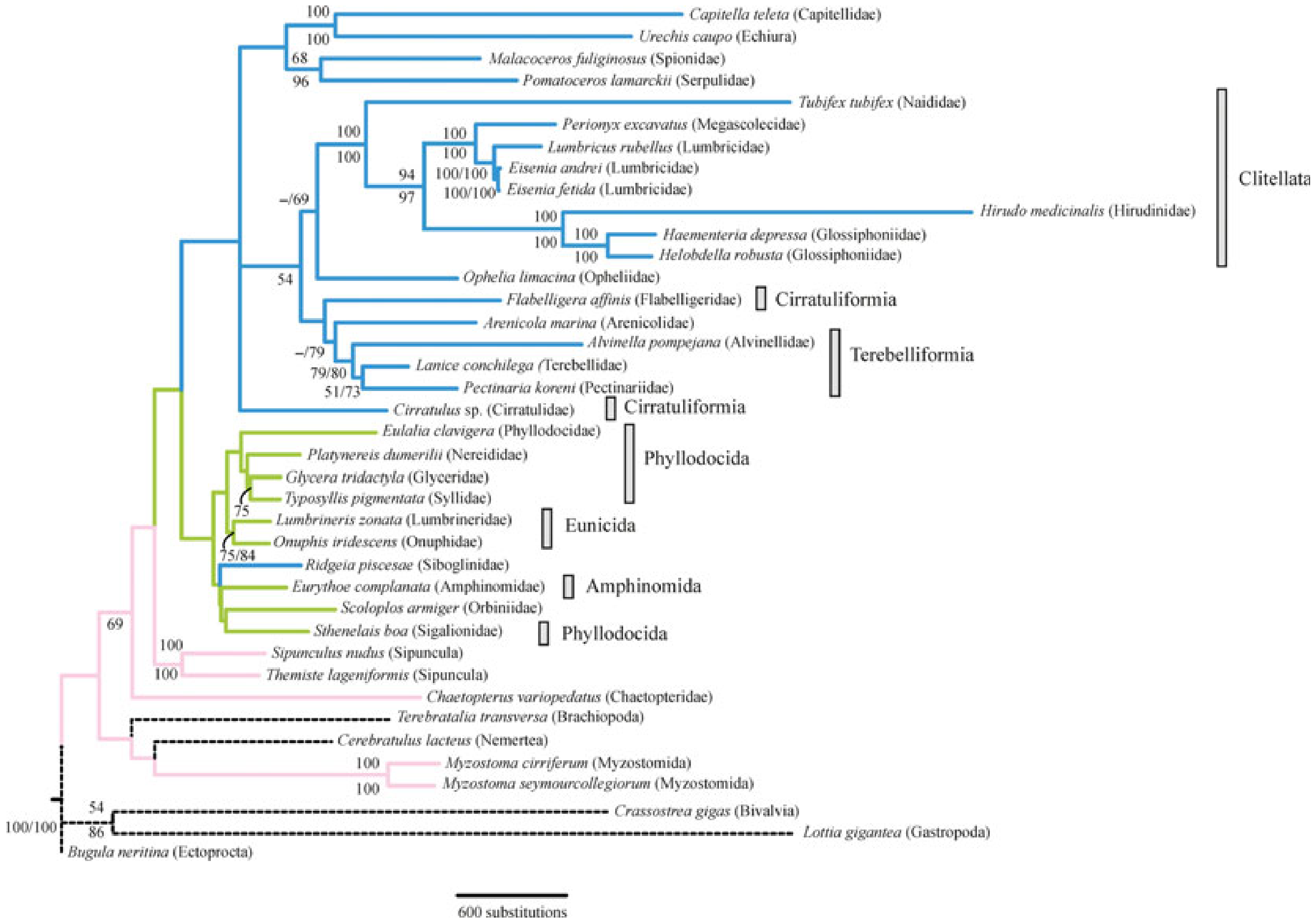
100
100
Capitella teleta (Capitellidae)
Urechis caupo (Echiura)
68
96
Malacoceros fuliginosus (Spionidae)
Pomatoceros lamarckii (Serpulidae)
Tubifex tubifex (Naididae)
100
100
100
100
Perionyx excavatus (Megascolecidae)
Lumbricus rubellus (Lumbricidae)
94
97
100/100
100/100
Eisenia andrei (Lumbricidae)
Eisenia fetida (Lumbricidae)
–/69
100
100
Hirudo medicinalis (Hirudinidae)
100
100
Haementeria depressa (Glossiphoniidae)
Helobdella robusta (Glossiphoniidae)
Clitellata
54
Ophelia limacina (Opheliidae)
Flabelligera affinis (Flabelligeridae)
Cirratuliformia
Arenicola marina (Arenicolidae)
–/79
Alvinella pompejana (Alvinellidae)
79/80
Lanice conchilega (Terebellidae)
51/73
Pectinaria koreni (Pectinariidae)
Terebelliformia
Cirratulus sp. (Cirratulidae)
Cirratuliformia
Eulalia clavigera (Phyllodocidae)
Platynereis dumerilii (Nereididae)
Glycera tridactyla (Glyceridae)
Typosyllis pigmentata (Syllidae)
Phyllodocida
75
Lumbrineris zonata (Lumbrineridae)
Onuphis iridescens (Onuphidae)
Eunicida
75/84
Ridgeia piscesae (Siboglinidae)
Eurythoe complanata (Amphinomidae)
Amphinomida
Scoloplos armiger (Orbiniidae)
Sthenelais boa (Sigalionidae)
Phyllodocida
69
100
100
Sipunculus nudus (Sipuncula)
Themiste lageniformis (Sipuncula)
Chaetopterus variopedatus (Chaetopteridae)
Terebratalia transversa (Brachiopoda)
Cerebratulus lacteus (Nemertea)
100
100
Myzostoma cirriferum (Myzostomida)
Myzostoma seymourcollegiorum (Myzostomida)
54
86
100/100
Crassostrea gigas (Bivalvia)
Lottia gigantea (Gastropoda)
Bugula neritina (Ectoprocta)
600 substitutions

Fig. 4. Strict consensus of two most-parsimonious trees recovered from the amino acid data set with 13 redundant loci removed (L = 104 470; CI = 0.615; RI = 0.347). Standard bootstrap values ≥ 50% are indicated above each node and partition bootstrap values ≥ 50% below each node. Note the low support for most nodes, which is discussed further in the text. Branches are coloured following taxonomic affiliations in Struck et al. (2011): blue = Sedentaria, green = Errantia, red = Annelida but not Errantia or Sedentaria. Dashed lines denote non-annelid taxa and grey bars denote additional taxonomic information. Branch lengths are drawn proportional to change.

with an average of 59 ± 7.97% (note that no taxa were missing more than 86% of the characters and that deleting all taxa missing at least 55% of the total characters eliminates the outgroup and prevents meaningful analysis). In total, 38% of the taxa fulfil the requirement of lacking no more than 55% of the characters from the original data set. The data in Table 2 illustrate a substantial change in the number of taxa that fulfil the requirement of lacking no more than 75% of the characters (72% of the taxa remaining) as opposed to 70% of the characters (56% of the taxa remaining). Thus, 44% of the taxa in the data set are lacking 70% or more of the characters.

Taxon titration was also performed on the character-titrated data set with the highest average parsimony bootstrap score (the 65% data set; Table 1), in an attempt to tease out a stable tree topology. Interestingly, in the data set that already had 50% of the characters removed because they were not represented in at least 65% of the taxa, we get an added improvement if we then also eliminate about 12% of the taxa because they are still missing at least 70% of the characters (Table 3).

Discussion

Using the amino acid data set from Struck et al. (2011) as well as a newly compiled nucleotide representation of it, we show that parsimony analyses produce trees with topologies that are at odds with the Bayesian and ML trees recovered by that study and in this study. However, the resulting parsimony bootstrap values were found wanting for almost all clades, indicating that the topologies shown in Figs 2–4 are no more reliable than topologies of previous studies (e.g. Rousset et al., 2007). That is, our study shows that even when using this large EST data set, cladistic

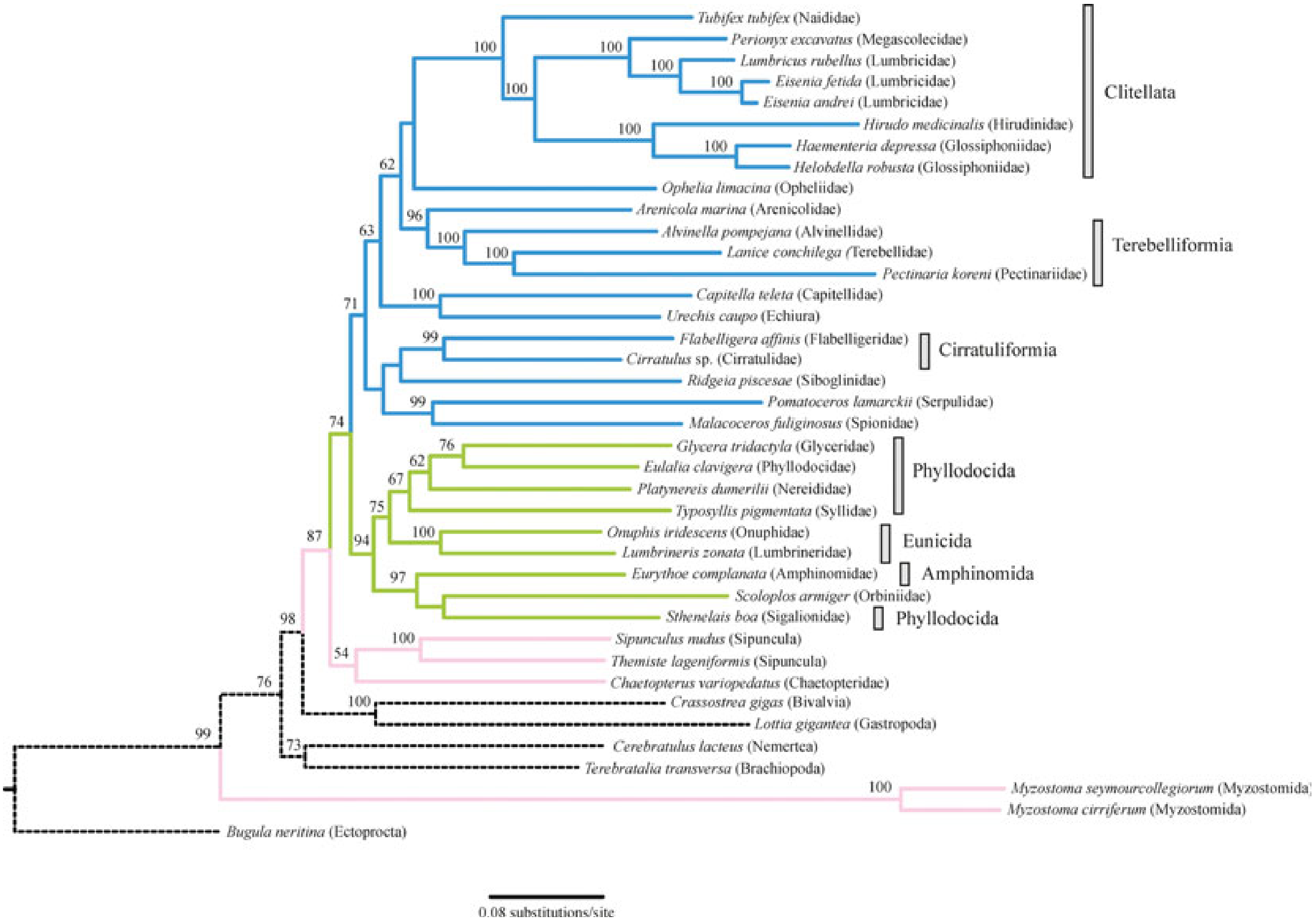
Tubifex tubifex (Naididae)
Perionyx excavatus (Megascolecidae)
Lumbricus rubellus (Lumbricidae)
Eisenia fetida (Lumbricidae)
Eisenia andrei (Lumbricidae)
Hirudo medicinalis (Hirudinidae)
Haementeria depressa (Glossiphoniidae)
Helobdella robusta (Glossiphoniidae)
Ophelia limacina (Opheliidae)
Arenicola marina (Arenicolidae)
Alvinella pompejana (Alvinellidae)
Lanice conchilega (Terebellidae)
Pectinaria koreni (Pectinariidae)
Capitella teleta (Capitellidae)
Urechis caupo (Echiura)
Flabelligera affinis (Flabelligeridae)
Cirratulus sp. (Cirratulidae)
Ridgeia piscesae (Siboglinidae)
Pomatoceros lamarckii (Serpulidae)
Malacoceros fuliginosus (Spionidae)
Glycera tridactyla (Glyceridae)
Eulalia clavigera (Phyllodocidae)
Platynereis dumerilii (Nereididae)
Typosyllis pigmentata (Syllidae)
Onuphis iridescens (Onuphidae)
Lumbrineris zonata (Lumbrineridae)
Eurythoe complanata (Amphinomidae)
Scoloplos armiger (Orbiniidae)
Sthenelais boa (Sigalionidae)
Sipunculus nudus (Sipuncula)
Themiste lageniformis (Sipuncula)
Chaetopterus variopedatus (Chaetopteridae)
Crassostrea gigas (Bivalvia)
Lottia gigantea (Gastropoda)
Cerebratulus lacteus (Nemertea)
Terebratalia transversa (Brachiopoda)
Myzostoma seymourcollegiorum (Myzostomida)
Myzostoma cirriferum (Myzostomida)
Bugula neritina (Ectoprocta)
Clitellata
Terebelliformia
Cirratuliformia
Phyllodocida
Eunicida
Amphinomida
Phyllodocida
100
100
100
100
100
100
100
62
96
100
100
63
100
71
99
99
74
76
62
67
75
100
94
97
87
98
100
54
76
100
73
99
100
0.08 substitutions/site

Fig. 5. ML tree using the data sets with 13 redundant loci removed (ln $L = -651928.083620$). Standard likelihood bootstrap values are shown above each node. Branches are coloured following taxonomic affiliations in Struck et al. (2011): blue = Sedentaria, green = Errantia, red = Annelida but not Errantia or Sedentaria. Dashed lines denote non-annelid taxa and grey bars denote additional taxonomic information.

Table 1

Results from the character titration analysis with associated bootstrap support values for some major groupings. Characters were incrementally removed from the data set if they were lacking for a certain percentage of the taxa and bootstrap resampling was performed at each increment to identify poorly supported groups and the effects of missing data on the analyses. CM cutoff indicates the level at which characters needed to be missing from taxa in order to be excluded from the data set (i.e. at 100, a character needed to be absent from 100% of the taxa to be excluded);% Characters indicates the percentage of the original character data set that was still included after each titration level.

CM cutoff	Average support	Annelida	Errantia	Arenicola + Terrebelliformia	Phyllodocida	Clitellata + Ophelia	Cirratuliformia	% Characters
100	39.8	0	0	0	0	41	0	100
85	38.4	0	0	4	0	31	0	97.32
80	48.2	47	0	0	4	43	0	87.33
75	58.9	83	8	81	0	42	20	76.73
70	60.8	96	14	83	21	67	48	63.04
65	62.3	95	0	80	30	63	87	50.44
60	57.9	88	10	71	10	64	77	42.15
55	57.4	75	0	79	0	70	87	36.46
50	59.8	9	0	82	0	60	90	33.83
45	56.9	10	0	88	0	73	100	29.08
40	56.7	33	0	97	0	57	99	24.30
35	53.3	25	0	70	0	13	95	15.42
25	39.1	12	0	0	0	0	91	4.37

Table 2
Results from the taxon titration scheme. Taxa were incrementally excluded from the data set if they were lacking a certain percentage of the characters and bootstrap resampling was performed at each increment. TM cutoff indicates the level of missing characters at which taxa were excluded (i.e. at 100, taxa had to lack 100% of the characters in order to be removed);% Taxa indicates the percentage of the taxa from the original data set that were remaining after each titration level; average support is across the entire topology; the last column indicates the product of the number of supported groups and the average support (i.e. the sum of all bootstrap values across the tree)

TM cutoff	% Taxa	Average support	Average support * no. of groups
100	100	57.4	2123.8
85	94.87	51.5	1802.5
80	82.05	53	1590
75	71.79	57.6	1497.6
70	56.41	59.4	1188
65	48.72	64.7	1099.9
60	41.03	54.5	763
55	38.46	76.2	990.6

Table 3
Results from the taxon titration on the data set with highest overall clade support in the character titration scheme (exclusion of characters that were missing for 65% of the taxa; see Table 1). Taxa were incrementally excluded from the data set if they were lacking a certain percentage of the characters and bootstrap resampling was performed at each increment. TM cutoff indicates the level of missing characters at which taxa were excluded (i.e. at 100, taxa had to lack 100% of the characters in order to be removed);% Taxa indicates the percentage of the taxa from the original data set that were remaining after each titration level; and average support is across the entire topology. See text for further details

TM cutoff	% Taxa	Average support
100	100	62.3
80	97.44	55.81
75	92.31	65.29
70	84.62	69.26
65	79.49	68.17
60	74.36	63.74
55	82.76	78.91
50	56.41	76.25

analysis does not support the probabilistic hypotheses of phylogenetic relationships of Annelida, such as the monophyly of Sedentaria and, in most cases, Errantia. Regardless of one's preferred optimality criterion, these findings reflect the continued difficulty of finding numerous reliable and suitable loci for a group as ancient as Annelida (Struck et al., 2007). The fact that the Bayesian tree shown by Struck et al. (2011) receives dramatically higher support values than the parsimony trees shown here may not be surprising. Previous studies have shown that Bayesian posterior probabilities are often considerably higher when compared with parsimony or ML bootstrap values (Suzuki et al., 2002; Alfaro et al., 2003; Cummings et al., 2003; Douady et al., 2003; Lewis et al., 2005), and might be treated with appropriate caution. Note here, however, that Struck et al. (2011) base their conclusions only on nodes that also received high ML bootstrap support.

With the exception of a few unsupported clades, the ML tree shown here mirrors the Bayesian tree shown by Struck et al. (2011), exposing the fact that the optimality criterion of choice will dictate the tree topology. This is unfortunate but solidifies a historical contingency: finding numerous reliable loci that will consistently lead to the same phylogenetic hypothesis across optimality criteria is a very problematic issue in Annelida. In addition to recovering Errantia and Sedentaria as monophyletic, our ML analysis recovered a tree with relatively higher support than the parsimony trees. That said, there are still several clades, chiefly within Sedentaria (see Fig. 5), that consistently remain unsupported by both parsimony and ML.

Stochastic models of evolution may be prone to difficulties in parameter estimation when the number of parameters is greatly increased (e.g. amino acid substitution matrices) or when a large number of characters are added in conjunction with the use of certain models of evolution (Felsenstein, 1982, 1983, 2004; Lartillot and Phillippe, 2004). Because (site-specific) mutations occur at the nucleotide level, and because nucleotides entail fewer parameters, we were curious to see how well a tree recovered from a nucleotide representation would compare with that of the amino acid data set used by Struck et al. (2011). Nucleotide data sets potentially hold an advantage over those of amino acids in that they consider potentially informative synonymous substitutions, a measure that is lost after translation into amino acids. If the rate of third-position nucleotide substitutions greatly exceeds "normal" evolutionary rates, this may in some cases lead to less resolved phylogeny reconstructions (Cunningham, 1997). However, in other cases, inclusion of such sites increases the resolution and support of the trees (e.g. Källersjö et al., 1999; Rydin et al., 2002; Kim et al., 2004). Bearing all of this in mind, our nucleotide data set produced a tree with a substantially different topology (Fig. 2) as compared with both the parsimony tree recovered from the original (Struck et al., 2011) amino acid data set (Fig. 3), as well as the trees recovered from the reduced amino acid data set used here (Figs 4 and 5), much as it differs from the Bayesian tree shown by Struck et al. (2011). None of Annelida, Errantia, and Sedentaria was monophyletic in the resulting nucleotide hypothesis, based on over 347 000 aligned sites. Nor was there much support for any but the already obviously well-supported groups (Rousset et al., 2007). It is also notable that the support values related to our analysis of nucleotide data agree, in most respects, to those estimated by the parsimony analysis performed by Zrzavý et al. (2009) on the basis of only six nucleotide loci, suggesting that a 38-fold increase in sequence information is no more or less helpful in resolving annelid relationships. However, this seems to be true only under the parsimony criterion as likelihood bootstrap values seem to increase greatly with an increased number of characters (see Struck et al., 2008; Struck, 2011). In comparing the parsimony analyses performed here with previous studies of annelid phylogenetics, it seems that using amino acids to reconstruct the trees is more propitious than using nucleotides; this was also noted by Dordel et al. (2010) for probabilistic analysis.

Redundancy of loci and orthology determination

All phylogenetic methods assume independence of characters (Farris, 1983), a prerequisite that is transitive to loci or other sets of characters. Whether automated orthology determinations are sensitive to this requirement is poorly evaluated in phylogenomics. Using a straightforward BLAST to establish orthology will not itself ensure locus independence, especially if coverage in a database is patchy (Koski and Golding, 2001). Our simple double-check was to perform an "all vs. all" BLAST search and contrast the ranges (rather than a single match) of the e-values *among* the suspected overlapping loci against the ranges of e-values *within* each putative locus. When applied to the amino acid data set used here, this simple method showed that 23 loci belonged to ten distinct, redundant sets of loci. Exemplary of this, sequences for *Helobdella robusta*, *Lottia gigantea*, *Capitella teleta*, and *Crassostrea gigas* at locus 59 and locus 141 all display perfect e-values (0), even for inter-locus comparisons. Non-independence of (e.g.) loci inflates the number of ad-hoc hypotheses needed to describe the data on any tree (Farris, 1983). For example, if two taxa are supported as a group, any homoplasy supporting that group would be counted as many times as there are redundant non-independent sites (Farris, 1983; Farris and Kluge, 1985). Clearly, the use of non-independent data generates an artificial inflation of support values for clades that are themselves supported by the redundant loci involved. As the size of the data set increases, as it does when many orthologous loci are used more than once in a data set, the support values will also artificially increase (see de Queiroz et al., 1995).

A corollary problem, and one not fully explored here, concerns orthology statements that belie multiple loci in one. Several of the loci show abnormally high e-values for intra-locus comparisons, suggesting that the sequences have been forced to represent a single locus, regardless of sequence similarity or coverage. For example, within locus 48, *Cirratulus* sp. and *Platynereis dumerilii* return an e-value of $5E^0$ when compared against each other, much like *Sipunculus nudus* vs. *Arenicola marina* within locus 95 (see Supplementary material). Of the 107 376 individual intra-locus BLAST hits, 2320 (2.16%) show e-values higher than $1E^{-5}$ and 1679 (1.56%) of them show e-values of 0.001 or higher (see Supplementary material). This could of course be exacerbated by the often fragmentary nature of EST data leading to very little overlap between taxa on any given locus. Regardless, if orthology statements for the loci involved in these hits were based on e-values alone, many would not accept that these loci should be considered orthologues. Indeed, BLAST-based orthology prediction is becoming more stringent (Kharchenko et al., 2006; Chen et al., 2007) and there is some precedence for using an e-value cut-off of $1E^{-20}$ for unequivocal orthology determination (e.g. Putta et al., 2004; Pel et al., 2007). If these stringent e-values were to be applied to the current amino acid data set, fully 12 481 sequences (11.62%) would have to be removed from the partition to which they currently belong.

Errantia, Sedentaria, and Pleistoannelida

Until the 1970s, Polychaeta was widely accepted to consist of Archiannelida, Errantia, and Sedentaria (Bartolomaeus et al., 2005), although some authors had expressed doubt about the reliability of these taxa (e.g. Dales, 1962; Day, 1967). In the last quarter of the twentieth century, each of these large taxa was eliminated on the basis of both phylogenetic analysis and morphological examinations. While Hermans (1969) argued that archiannelid taxa are more closely related to each other than any other group based on morphological phylogenetics, Westheide (1985, 1987) based his arguments on ontogeny and a more comprehensive phylogeny, which suggested that Archiannelida is a paraphyletic assemblage. Errantia and Sedentaria also were eliminated during this period of rapid advancement of annelid systematics; Fauchald (1977) replaced them with 17 orders on the basis of morphological examination. Struck et al.'s (2011) analyses imply a resurrection of Errantia and Sedentaria, and Struck (2011) also erects a new taxon, Pleistoannelida (defined as the group consisting of Errantia and Sedentaria, and exclusive of Sipuncula, Myzostomida, and *Chaetopterus*).

In the most comprehensive analyses of Annelida prior to Struck et al. (2011) none of Errantia, Sedentaria or Pleistoannelida was reliably recovered as monophyletic groups (Zrzavý et al., 2009). With the most taxonomically broad analysis to date, Rousset et al. (2007), with four loci for 217 taxa, also failed to recover the monophyly of Errantia, Sedentaria, or Pleistoannelida. Our analysis of nucleotides and reanalyses of amino acids with redundant sequences removed continue to suggest a lack of compelling evidence for the monophyly of Errantia or Sedentaria, much like both the nucleotide data set and the original amino acid data set used by Struck et al. (2011) do not reliably find the monophyly of Pleistoannelida under a parsimony framework. Interestingly, however, when removing the redundant loci from the amino acid data set, Pleistoannelida is recovered as monophyletic but without support. If the root of the tree resulting from our analysis of the original amino acid data set was applied at the node leading to *Chaetopterus*, however, our tree would result in monophyly of both Errantia and Sedentaria. Increased taxon sampling may be of even more importance than increased genetic coverage, and could potentially accurately resolve the relationships at the base of the tree of Annelida, which would have a large effect on the overall topology and groupings found (see Phillippe et al., 2011). Indeed, some of the unconventional relationships found in the present study, such as the sister grouping of the amphinomid *Eurythoe complanata* and the phyllodocids *Scoloplos armiger* and *Sthenelais boa*, or the paraphyly of Cirratuliformia in Figs 2 and 4, can probably be explained by the low taxon sampling in the data set. Other atypical relationships, such as the placement of *Ridgeia piscesae* within Errantia, may be an artefact of missing data (see above and Tables 1 and 2). Adding to the controversies surrounding the phylogeny of Annelida, our analyses corroborate previous studies in suggesting that the hypotheses of relationships both within Annelida, and between the phylum and its constituent taxa are not consistent nor reasonably supported across optimality criteria, even when large amounts of data are employed by the analyses.

Transcriptomics and missing data

Differential gene expression is common in both vertebrates and invertebrates (e.g. Okubo et al., 1992; Evans and Wheeler, 1999; Carleton and Kocher, 2001; Franzellitti and Fabbri, 2005) and may explain the inability to gather comparable transcriptomic data from organisms with diverging expression patterns. Moreover, data recovered from EST libraries are often fragmentary with respect to the original transcript, and alignments of such sequences will overlap incompletely, leaving ragged ends (Hartmann and Vision, 2008). These issues are geared towards transcriptomic data (as opposed to PCR-directed studies) and may lead to either a highly mosaic data set with high levels of missing data or a data set with erroneously inferred orthology between the loci involved. The effect of missing data on large-scale, empirical EST studies is largely unknown (but see Phillippe et al., 2004; Siddall, 2009) but several simulation studies have found a large impact of both random and non-random missing data on the resulting topology (e.g. Wiens, 2003, 2006). Contrary to the results conveyed by Struck et al. (2011), here we show that the resulting average clade stability across a topology is reduced when including taxa with high levels of missing data. Thus, the lack of support for most clades in the parsimony hypotheses presented here is mostly attributable to the lack of data for several taxa in the original matrix. Unfortunately, this issue may be hard to resolve as a single transcriptome necessarily changes in nucleotide composition throughout the lifespan of any given organism, a concern that can only be partially mediated by exceedingly deep next-generation cDNA sequencing. The data set presented by Struck et al. (2011) will probably prove to be a cornerstone for future studies of annelid systematics and it is important that this data set now be amended such that the amount of missing data is reduced.

Acknowledgements

We are greatly indebted to Torsten Struck (University of Osnabrück) for providing the trace archive assemblies, for his valuable input on the process, and for his thoughtful comments that greatly improved this contribution. We also thank Alejandro Oceguera, Anna Phillips, Liz Wright and two anonymous reviewers for their insightful comments on an early draft of this manuscript, and Apurva Narechania for technical assistance. We thank the Willi Hennig Society for making TNT freely available. This contribution was supported the Swedish-American Foundation, the Wenner-Gren Foundations, the Olle Engkvist Byggmästare Foundation, the Sixten Gemzéus Foundation, and Helge Ax:son Johnson's Foundation.

References

Alfaro, M.E., Zoller, S., Lutzoni, F., 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov Chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20, 255–266.

Bartolomaeus, T., Purschke, G., Hausen, H., 2005. Polychaete phylogeny based on morphological data—a comparison of current attempts. Hydrobiologia 535/536, 341–356.

Bely, A.E., Wray, G.A., 2004. Molecular phylogeny of naidid worms (Annelida: Clitellata) based on cytochrome oxidase I. Mol. Phylogenet. Evol. 30, 50–63.

Bleidorn, C., Vogt, L., Bartolomaeus, T., 2003. New insights into polychaete phylogeny (Annelida) inferred from 18S rDNA sequences. Mol. Phylogenet. Evol. 29, 279–288.

Brinkhurst, R.O., Nemec, A.F.L., 1987. A comparison of phenetic and phylogenetic methods applied to the systematics of Oligochaeta. Hydrobiologia 155, 65–74.

Carleton, K.L., Kocher, T.D., 2001. Cone opsin genes of African cichlid fishes: tuning spectral sensitivity by differential gene expression. Mol. Biol. Evol. 18, 1540–1550.

Chen, F., Mackey, A.J., Vermunt, J.K., Roos, D.S., 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2, e383.

Cummings, M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A., Winka, K., 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52, 477–487.

Cunningham, C.W., 1997. Can three incongruence tests predict when data should be combined? Mol. Biol. Evol. 14, 733–740.

Dales, R.P., 1962. The polychaete stomatodeum and the interrelationships of the families of Polychaeta. Proc. Zool. Soc. London 139, 389–428.

Day, J.H. 1967. A monograph on the Polychaeta of Southern Africa. Vol. British Museum (Natural History) Publication 656 (Part I. Errantia, Part II. Sedentaria). British Museum (Natural History), London, UK.

Dordel, J., Fisse, F., Purschke, G., Struck, T.H., 2010. Phylogenetic position of Sipuncula derived from multi-gene and phylogenomic data and its implication for the evolution of segmentation. J. Zool. Syst. Evol. Res. 48, 197–207.

Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., Douzery, E.J.P., 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20, 248–254.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., et al., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.

Erséus, C., 1987. Phylogenetic analysis of the aquatic Oligochaeta under the principle of parsimony. Hydrobiologia 155, 75–89.

Erséus, C., 2005. Phylogeny of oligochaetous Clitellata. Hydrobiologia 535/536, 357–372.

Erséus, C., Källersjö, M., 2004. 18S rDNA phylogeny of Clitellata (Annelida). Zool. Scr. 33, 187–196.

Evans, J.D., Wheeler, D.E., 1999. Differential gene expression between developing queens and workers in the honey bee, *Apismellifera*. Proc. Natl Acad. Sci. USA 96, 5575–5580.

Fang, G., Bhardwaj, N., Robilotto, R., Gerstein, M.B., 2010. Getting started in gene orthology and functional analysis. PLoS Comput. Biol. 6, e1000703.

Farris, J.S. 1983. The logical basis of phylogenetic analysis. In: Platnick, N. I., Funk, V. A. (Eds.), Advances in Cladistics 2. Columbia University Press, New York, 7–36.

Farris, J.S., Kluge, A.G., 1985. Parsimony, synapomorphy, and explanatory power: a reply to Duncan. Taxon 34, 130–135.

Fauchald, K., 1977. The Polychaete Worms: Definitions and Keys to the Orders, Families and Genera. Natural History Museum of Los Angeles County, Science Series 28, Los Angeles, CA.

Fauchald, K., Rouse, G.W., 1997. Polychaete systematics: past and present. Zool. Scr. 26, 71–138.

Felsenstein, J., 1982. How can we infer geography and history from gene frequencies? J. Theoret. Biol. 1, 9–20.

Felsenstein, J., 1983. Parsimony in systematics: biological and statistical issues. Ann. Rev. Ecol. Syst. 14, 313–333.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Franzellitti, S., Fabbri, E., 2005. Differential HSP70 gene expression in the Mediterranean mussel exposed to various stressors. Biochem. Biophys. Res. Commun. 4, 1157–1163.

Ge, F., Wang, L-S., Kim, J., 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3, 1709–1718.

Goloboff, P.A., Farris, J.S., Nixon, K.C., 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786.

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

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G.W., Edgecombe, G.D., Martinez, P., Baguñà, J., Bailly, X., Jondelius, U., Wiens, M., Müller, W.E.G., Seaver, E., Wheeler, W.C., Martindale, M.Q., Giribet, G., Dunn, C.W., 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. R. Soc. B. 276, 4261–4270.

Hermans, C.O., 1969. The systematic position of Archiannelida. Syst. Biol. 18, 85–102.

Källersjö, M., Albert, V.A., Farris, J.S., 1999. Homoplasy *increases* phylogenetic structure. Cladistics 15, 91–93.

Katoh, K., Kuma, K-I, Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518.

Kharchenko, P., Chen, L., Freund, Y., Vitkup, D., Church, G. M., 2006. Identifying metabolic enzymes with multiple types of association evidence. BMC Evol. Biol. 7, 177.

Kim, S., Soltis, D.E., Soltis, P.S., Suh, Y., 2004. DNA sequences from Miocene fossils: an *ndhF* sequence of *Magnolia lathahensis* (Magnoliaceae) and an *rbcL* sequence of *Persea pseudocarolinensis* (Lauraceae). Am. J. Bot. 91, 615–620.

Koski, L.B., Golding, G.B., 2001. The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52, 540–542.

Lartillot, N., Phillippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109.

Lewis, P.O., Holder, M.T., Holsinger, K.E., 2005. Polytomies and Bayesian inference. Syst. Biol. 54, 241–253.

Masoudi-Nejad, A., Tonomura, K., Kawashima, S., Moriya, Y., Suzuki, M., Itoh, M., Kanehisa, M., Endo, T., Goto, S., 2006. EGassembler: online bioinformatics service for large-scale processing, clustering and assembling EST's and genomic DNA fragments. Nucleic Acids Res. 34, W459–W462.

McHugh, D., 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. Proc. Natl Acad. Sci. USA 94, 8006–8009.

McHugh, D., 2000. Molecular phylogeny of the Annelida. Can. J. Zool. 78, 1873–1884.

Oakley, T.W., Wolfe, J.M., Lindgren, A.R., Zaharoff, A.K. 2013. Phylotranscriptomics to bring the understudied into the fold: monophyletic Ostracoda, fossil placement and pancrustacean phylogeny. Mol. Biol. Evol. 30, 215–233.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., Matsubara, K., 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat. Genet. 2, 173–179.

Pel, H.J., de Winde, J.H., Archer, D.B., Dyer, P.S., Hofmann, G., et al., 2007. Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. Nat. Biotechnol. 25, 221–231.

Perrier, E., 1897. Traité de Zoologie. , Fascicule IV. Vers. Molusques, Tuniciers, Masson et Cie, Paris.

Phillippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W.H., Casane, D., 2004. Phylogenomics of Eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21, 1740–1752.

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

Putta, S., Smith, J.J., Walker, J.A., Rondet, M., Weisrock, D.W., Monaghan, J., Samuels, A.K., Kump, K., King, D.C., Maness, N.J., Habermann, B., Tanaka, E., Bryant, S.V., Gardiner, D.M., Parichy, D.M., Voss, S.R., 2004. From biomedicine to natural history research: EST resources for ambystomatid salamanders. BMC Genomics 5, 54.

de Queiroz, A., Donoghue, M.J., Kim, J., 1995. Separate versus combined analysis of phylogenetic evidence. Ann. Rev. Ecol. Syst. 26, 657–681.

Rouse, G.W., Fauchald, K., 1997. Cladistics and polychaetes. Zool. Scr. 26, 139–204.

Rouse, G.W., Pleijel, F., 2001. Polychaetes. Oxford University Press, New York.

Rousset, V., Pleijel, F., Rouse, G.W., Erséus, C., Siddall, M.E., 2007. A molecular phylogeny of annelids. Cladistics 23, 41–63.

Rydin, C., Källersjö, M., Friis, E.M., 2002. Seed plant relationships and the systematic position of Gnetales based on nuclear and chloroplast DNA: conflicting data, rooting problems, and the monophyly of conifers. Int. J. Plant Sci. 163, 197–214.

Siddall, M.E., 2009. Unringing a bell: metazoan phylogenomics and the partition bootstrap. Cladistics 26, 444–452.

Siddall, M.E., Apakupakul, K., Burreson, E. M., Coates, K. A., Erséus, C., Gelder, S. R., Källersjö, M., Trapido-Rosenthal, H., 2001. Validating Livanow: molecular data agree that leeches, branchiobdellidans, and *Acanthobdella peledina* form a monophyletic group of oligochaetes. Mol. Phylogenet. Evol. 21, 346–351.

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

Struck, T.H., 2011. Direction of evolution within Annelida and the definition of Pleistoannelida. J. Zool. Syst. Evol. Res. 49, 340–345.

Struck, T.H., Schult, N., Kusen, T., Hickman, E., Bleidorn, C., McHugh, D., Halanych, K.M., 2007. Annelid phylogeny and the status of *Sipuncula* and *Echiura*. BMC Evol. Biol. 7, 57.

Struck, T.H., Nesnidal, M.P., Purschke, G., Halanych, K.M., 2008. Detecting possibly saturated positions in 18S and 28S sequences and their influence on phylogenetic reconstruction of Annelida (Lophotrochozoa). Mol. Phylogenet. Evol. 48, 628–645.

Struck, T.H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., Meyer, A., Tiedemann, R., Purschke, G., Bleidorn, C., 2011. Phylogenomic analyses unravel annelid evolution. Nature 471, 95–98.

Suzuki, Y., Glazko, G. V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl Acad. Sci. USA 99, 16138–16143.

Westheide, W. 1985. The systematic position of Dinophilidae and the archiannelid problem. In: Conway Morris, S., George, J. D., Gibson, R., Platt, H. M. (Eds.), The Origins and Relationships of Lower Invertebrates. Oxford University Press, Oxford, pp. 310–326.

Westheide, W., 1987. Progenesis as a principle in meiofauna evolution. J. Nat. Hist. 21, 843–854.

Westheide, W., McHugh, D., Purschke, G., Rouse, G., 1999. Systematization of the Annelida: different approaches. Hydrobiologia 402, 291–307.

Wiens, JJ., 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52, 528–538.

Wiens, JJ., 2006. Missing data and the design of phylogenetic analyses. J. Biomed. Inform. 39, 34–42.

Zaret, K.S., Sherman, F., 1982. DNA sequence required for efficient transcription termination in yeast. Cell 28, 563–573.

Zhang, Z-Q., 2011. Animal biodiversity: an introduction to higher-level classification and taxonomic richness. Zootaxa 3148, 7–12.

Zrzavý, J., Říha, P., Piálek, L., Janouškovec, J., 2009. Phylogeny of Annelida (Lophotrochozoa): total-evidence analysis of morphology and six genes. BMC Evol. Biol. 9, 189.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Full nucleotide data set in TNT format.

Data S2. Full amino acid data set, in TNT format, after removing redundant data.

Data S3. Results from the tBLASTn search using the original amino acids as queries against GenBank EST in .xls file.