LETTER

doi:10.1038/nature09864

Phylogenomic analyses unravel annelid evolution

Torsten H. Struck[1], Christiane Paul[2], Natascha Hill[3], Stefanie Hartmann[3], Christoph Hösel[1], Michael Kube[4], Bernhard Lieb[5], Achim Meyer[5], Ralph Tiedemann[2], Günter Purschke[1] & Christoph Bleidorn[2,6]

Annelida, the ringed worms, is a highly diverse animal phylum that includes more than 15,000 described species and constitutes the dominant benthic macrofauna from the intertidal zone down to the deep sea. A robust annelid phylogeny would shape our understanding of animal body-plan evolution and shed light on the bilaterian ground pattern. Traditionally, Annelida has been split into two major groups: Clitellata (earthworms and leeches) and polychaetes (bristle worms), but recent evidence suggests that other taxa that were once considered to be separate phyla (Sipuncula, Echiura and Siboglinidae (also known as Pogonophora)) should be included in Annelida[1–4]. However, the deep-level evolutionary relationships of Annelida are still poorly understood, and a robust reconstruction of annelid evolutionary history is needed. Here we show that phylogenomic analyses of 34 annelid taxa, using 47,953 amino acid positions, recovered a well-supported phylogeny with strong support for major splits. Our results recover chaetopterids, myzostomids and sipunculids in the basal part of the tree, although the position of Myzostomida remains uncertain owing to its long branch. The remaining taxa are split into two clades: Errantia (which includes the model annelid *Platynereis*), and Sedentaria (which includes Clitellata). Ancestral character trait reconstructions indicate that these clades show adaptation to either an errant or a sedentary lifestyle, with alteration of accompanying morphological traits such as peristaltic movement, parapodia and sensory perception. Finally, life history characters in Annelida seem to be phylogenetically informative.

Annelids are found throughout the world's terrestrial, aquatic and marine habitats. They represent one of three major animal groups with segmentation, so understanding annelid body-plan evolution is crucial for elucidating aspects of the evolution of Bilateria[5–7]. Several annelid taxa have recently emerged as model organisms in various biological disciplines[8]. Surprisingly, the evolution of Annelida is still poorly understood, and it is uncertain how well these model organisms represent the ancestral character traits in Annelida. To rectify this situation, multigene data sets are needed to evaluate the diversity and the relationships of major annelid clades.

Annelida traditionally included Polychaeta and Clitellata. Morphological and molecular data corroborate clitellate monophyly and provide robust phylogenetic hypotheses within this taxon[9]. Polychaetes are classified into approximately 80 family-level taxa that are generally supported as monophyletic; however, arrangement of these taxa into well-supported, more-inclusive nodes is problematic[2,10]. Historically, polychaetes were classified as either Sedentaria or Errantia on the basis of their morphology and mode of life[11–13]. This systematization was dismissed in the 1970s as being arbitrary groupings useful only for practical purposes[14]. About 15 years ago, on the basis of morphological cladistic analyses, a monophyletic Polychaeta consisting of two major clades, Scolecida and Palpata, was proposed, with the latter clade divided into Canalipalpata and Aciculata[15]. There is increasing molecular evidence, however, that places Clitellata, as well as the non-segmented taxa Echiura and Sipuncula, within polychaetes and thus renders Polychaeta paraphyletic[1–4]. So far, molecular work based on only a few genes has not supported the proposed monophyly[15] of most major polychaete clades. Yet, support for basal nodes in these studies is less than 50 or 0.50 for bootstrap support (BS) or posterior probability (PP), respectively, resulting in a lack of support for alternative hypotheses[2,3].

To address these major outstanding issues of annelid phylogeny, we used a phylogenomic approach, generating expressed sequence tag (EST) libraries for 17 annelid taxa, which are in addition to the publicly available EST or genomic data from annelids. We reconstructed relationships of major annelid taxa using 47,953 amino acid positions derived from 231 gene fragments that span 20 traditional polychaete 'families', Siboglinidae, Myzostomida, Echiura, Clitellata, Sipuncula and five outgroup taxa. This is the largest phylogenomic data set explored so far in annelid phylogeny and has a mean data coverage of 41.7% per taxon.

Sensitivity analyses of our data (Supplementary Tables 4 and 6) showed that increasing the number of positions and mean leaf stability had a positive impact on BS, whereas increasing the data coverage by removing either genes or taxa with low coverage had no such impact (Supplementary Fig. 1). Therefore, the largest data set (47,953 positions), with either all taxa (denoted ALL) or excluding the five annelid taxa that showed leaf stabilities below 0.925 (denoted EX) was used in maximum likelihood and Bayesian inference analyses. These analyses retrieved a clade (called clade 1) comprising all annelids with the exception of Chaetopteridae, Sipuncula and Myzostomida. This clade received significant branch support: ALL, PP = 0.98 (Bayesian inference), BS = 88 (maximum likelihood); EX, PP = 0.99 (Bayesian inference), BS = 100 (maximum likelihood) (Fig. 1 and Supplementary Figs 2, 3, 6 and 8). Reconstructing ancestral morphological traits for clade 1 and Annelida revealed that they were similar, except for some larval characters (Fig. 2a and Supplementary Table 5).

On the basis of this reconstruction, the ancestral annelid had a pair of anterior appendages (that is, grooved palps), which functioned in food gathering and sensory perception. Other head or pygidial appendages were absent. Eyes and nuchal organs were present as sensory organs. Of the different chaetal types, only internalized supporting chaetae and simple chaetae were part of the ancestral pattern. Reconstructions of most other parapodial characters were uncertain, except for the possession of prominent notopodial lobes. Although the fossil record of early annelids from the Cambrian period is sparse, it nonetheless reveals that, congruent with our reconstructions, the early annelids had palps, simple chaetae and internalized supporting chaetae but did not have other chaetae or appendages such as tentacular, parapodial or pygidial cirri[16,17].

In agreement with previous molecular studies[1–3,18], Chaetopteridae, which have three distinct body regions, are found in the basal part of the annelid tree. Thus, the evolution of segmentation—with predominantly homonomous segmentation in clade 1 and Myzostomida, heteronomous segmentation in Chaetopteridae and complete reduction in Sipuncula—is already highly variable at basal nodes in the

[1]University of Osnabrück, FB05 Biology/Chemistry, AG Zoology, Barbarastrasse 11, 49069 Osnabrück, Germany. [2]University of Potsdam, Institute of Biochemistry and Biology, Unit of Evolutionary Biology/Systematic Zoology, Karl-Liebknecht-Strasse 24-25, Haus 26, 14476 Potsdam, Germany. [3]University of Potsdam, Institute of Biochemistry and Biology, Unit of Bioinformatics, Karl-Liebknecht-Strasse 24-25, Haus 14, 14476 Potsdam, Germany. [4]Max Planck Institute for Molecular Genetics, Ihnestrasse 63–73, 14195 Berlin, Germany. [5]Johannes Gutenberg University, Institute of Zoology, Müllerweg 6, 55099 Mainz, Germany. [6]University of Leipzig, Institute for Biology II, Molecular Evolution and Systematics of Animals, Talstrasse 33, 04103 Leipzig, Germany.

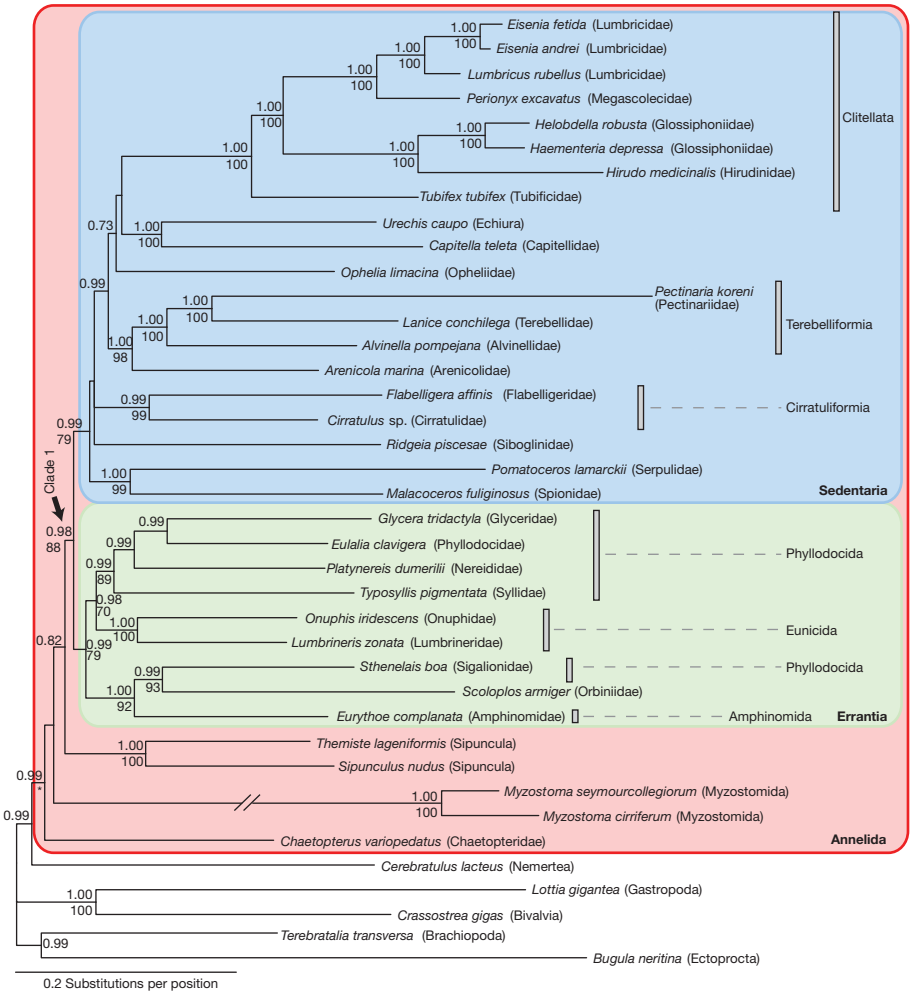

Figure 1 | Reconstruction of the Annelida phylogenetic tree. Majority rule consensus trees of the Bayesian inference analysis using the site-heterogeneous CAT model of the data set with 39 taxa and 47,953 amino acid positions. Only PP (top of branch or alone) and BS (bottom) values ≥ 0.70 or 70, respectively, are shown. The branch leading to Myzostomida is reduced by 75%. Annelida are highlighted in red, with Sedentaria in blue and Errantia in green. Grey bars indicate additional annelid groups. *, BS value for the monophyly of Annelida without Myzostomida in the maximum likelihood analysis is 99.

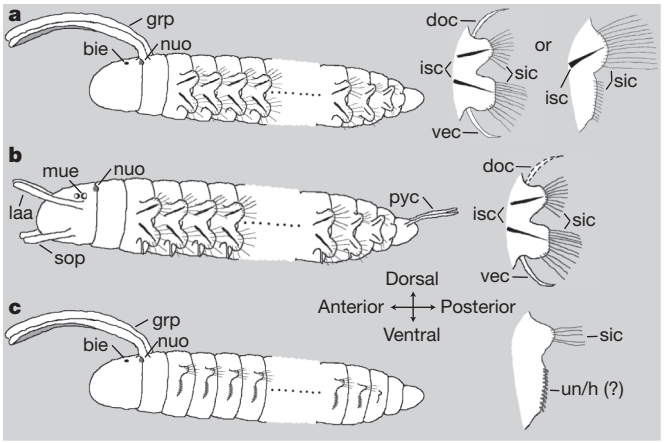

Figure 2 | Ancestral reconstructions of body and parapodial characters. **a**, Annelida and clade 1. **b**, Errantia. **c**, Sedentaria. Body characters (left) and parapodial characters (right) are depicted. The state of several parapodial characters in Annelida and clade 1 is uncertain, so we depict the two most extreme possibilities. Dashed lines or question marks indicate that the state of the character is uncertain. bie, bicellular eyes; doc, dorsal cirrus; grp, grooved palps; isc, internalized supporting chaetae; laa, lateral antenna; mue, multicellular eyes; nuo, nuchal organ; pyc, pygidial cirrus; sic, simple chaetae; sop, solid palps; un/h, uncini/hooks; vec, ventral cirrus.

annelid phylogeny[1]. Moreover, we acknowledge that, in addition to Chaetopteridae, Myzostomida and Sipuncula, other taxa such as Oweniidae, Dinophilidae or Protodrilida, which were not covered here because of a lack of data, might also be placed in the basal part of the annelid tree[2].

The major difference between the maximum likelihood and Bayesian inference analyses is the placement of Myzostomida. Myzostomids are either ectocommensals or endoparasites of echinoderms, and the systematic placement of this aberrant taxon has proved to be problematic[19,20]. Bayesian inference analysis places Myzostomida within Annelida (PP = 0.99 for both data sets (ALL and EX); Fig. 1 and Supplementary Fig. 6). By contrast, by maximum likelihood analyses, long-branched Myzostomida are grouped with Ectoprocta, the outgroup taxon with the longest branch (Supplementary Figs 2 and 3). There is conclusive support from mitochondrial gene order and morphological data that Myzostomida are part of the annelid radiation[19,20], and it has been shown that their derived sequences can be affected by long-branch attraction (LBA)[18]. The CAT model of Bayesian inference analyses is known to be less affected by LBA than other models, and this model proved to be better suited for our data set than was the LG model of maximum likelihood analyses (Supplementary Information). Notwithstanding the different position of Myzostomida (possibly owing to LBA), both maximum likelihood analyses support the monophyly of Annelida: ALL, BS = 99; EX, BS = 100 (Supplementary Figs 2 and 3). Moreover, the exclusion of Myzostomida did not substantially

alter the phylogenetic reconstruction of annelid ingroup relationships and BS values (Supplementary Fig. 7). Finally, the different placement of Myzostomida in the Bayesian inference and maximum likelihood analyses did not affect the reconstructions of ancestral morphological traits (Supplementary Table 5).

Clade 1 split into two well-supported clades: Errantia, which comprised Phyllodocida, Eunicida, Amphinomida and Orbiniidae; and Sedentaria, which comprised Clitellata and Echiura, as well as most other Scolecida (Capitellidae, Opheliidae and Arenicolidae) and Canalipalpata (Terebelliformia, Cirratuliformia, Siboglinidae, Serpulidae and Spionidae). Both clades were significantly supported: ALL, PP = 0.99 (Bayesian inference), BS = 79 (maximum likelihood); EX, PP = 0.99 (Bayesian inference), BS = 100 (maximum likelihood) (Fig. 1 and Supplementary Figs 2, 3, 6 and 8). The placement of Clitellata indicated a closer relationship to Terebelliformia/Arenicolidae, Opheliidae and Capitellidae/Echiura. Moreover, analyses of branch attachment frequencies based on the data set comprising all taxa showed that each of the five removed annelid taxa is nested in either Sedentaria (*Ridgeia*, *Ophelia*, *Pomatoceros* and *Malacoceros*) or Errantia (*Eurythoe*), and none is moving between clades (Supplementary Figs 4 and 5).

In an influential publication in the 1990s, the two main competing hypotheses of annelid evolution were discussed[21]: one, starting with a ground pattern that resembles an errant, epibenthic organism; and, the other, starting with an infaunal burrowing form. Interestingly, we found both trends to be realized within annelids. The ground pattern of Errantia reveals some important changes with respect to sensory perception and motility. On the basis of our reconstructions, the last common ancestor of Errantia had lateral antennae, palps (which are solid and restricted to sensory perception), a pair of pygidial cirri, nuchal commissures and two pairs of multicellular eyes facing in different directions[22] (Fig. 2b and Supplementary Table 5). The parapodia had prominent notopodial and neuropodial lobes supported by internalized chaetae, as well as ventral cirri. Overall, this pattern can be regarded as adaptations to a more active and mobile lifestyle, which requires increased perception of the environment, as well as motility by undulation. Prominent parapodial lobes are advantageous for rapid movements based on undulation, which is mainly achieved by the well-developed longitudinal musculature arranged in at least four separate bundles. For example, in sexually mature (that is, epigamous) nereidids, adopting a temporary pelagic reproductive stage, parapodial lobes are even further enlarged and paddle-like than in immature stages[17]. Most taxa of Phyllodocida, Eunicida and Amphinomida show such an errant, often predatory, mode of life and hence were traditionally named Errantia[11]. The position of Orbiniidae, which were previously grouped with Scolecida[15], might be surprising; however, placing them within or close to the errant forms had previously been debated on the basis of morphological and molecular evidence[2,3,12]. Therefore, we named this clade Errantia, as it is characterized by adaptation to a more errant life.

The evolution of parapodia in Sedentaria shows the opposite trend. Neuropodial and notopodial lobes are generally smaller than in Errantia and lack internalized supporting chaetae (Fig. 2c). In general, chaetae are in close proximity to the stiff body wall, an arrangement that facilitates a better anchorage in tubes and burrows. Moreover, antennae are absent, and palps have been lost independently in several taxa. The taxa of this clade are commonly characterized by a sedentary life, as more or less sessile organisms that live below stones, tube builders, or burrowers by means of peristalsis such as earthworms[17]. Sedentaria are generally microphagous. Taxa without appendages such as those formerly grouped as Scolecida[15] are deposit feeders, often ingesting sediment. By contrast, taxa with sometimes elaborate head appendages such as terebellids or serpulids are surface deposit feeders or filter feeders, respectively[23]. The deposit feeding lifestyle also generally applies to most Clitellata. Therefore, we named this clade Sedentaria[12] (now including Clitellata), and it is characterized by adaptations to a more sedentary lifestyle by, for example, the reduction of parapodia and loss of internalized supporting chaetae. A key feature is that the chaetae are in closer proximity to the stiff body wall, rather than being embedded in parapodial lobes (which are more flexible) as is typical for errant annelids. Interestingly, errant polychaetes with sedentary life strategies such as Lumbrineridae or Onuphidae have adapted to such a lifestyle by using different solutions[17].

Hence, within Annelida, there are two major clades, Errantia and Sedentaria, whose evolution was driven by the adaptation to two different modes of life. Errantia show a more mobile and active life strategy than Sedentaria, and this is correlated to increased sensory perception and motility. Sedentaria are more sessile, with accompanying reductions of head and body appendages and the position of the chaetae being in closer proximity to the body wall than in Errantia. Annelids have been successfully established as models in evolutionary developmental studies to deduce the characteristics of the last common bilaterian ancestor[24]. Of the recent model organisms, *Platynereis*, with its well-developed head and parapodial appendages, is a good representative of Errantia. By contrast, *Capitella* (as a burrower with reduced appendages), *Helobdella* (as a clitellate) and *Hydroides* (as a filter feeder using its radiolar crown) represent different microphagous feeder types in Sedentaria.

METHODS SUMMARY

EST libraries of 1,370 clones, on average, were prepared for 17 annelid species (Supplementary Table 1). All original sequence data have been deposited in the NCBI Expressed Sequence Tag database (dbEST). EST or genomic data from 17 additional annelid species and 5 outgroup species were obtained from public archives (Supplementary Table 1). These raw EST data were further processed as described previously[25]. Sets of orthologous genes were determined using the program HaMStR in combination with the InParanoid database (without ribosomal proteins)[26], and were translated into amino acid sequences using the program ESTwise[27]. In parallel, we retrieved all ribosomal proteins from databases as described previously[25] (Supplementary Table 2). Each orthologous gene set was aligned using MAFFT software[28] and masked using the program REAP[29]. Only genes that had taxon coverage of at least 33.3% were included in the final super-matrix.

Phylogenetic trees were inferred from this data set of 39 taxa by using a Bayesian inference approach (using the site-heterogeneous CAT model) and a maximum likelihood approach (using the LG model). Stabilities of taxa were assessed using the leaf stability index as calculated by Phyutility software[30] (Supplementary Table 3). The five annelid taxa with an index below 0.925 were removed in the second data set, and the Bayesian inference analysis was repeated. Branch attachment frequencies of these unstable annelid taxa were assessed using the lineage movement option in Phyutility[30] based on the data set with all taxa.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 2 September 2010; accepted 18 January 2011.

1. Dordel, J., Fisse, F., Purschke, G. & Struck, T. H. 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 (2010).
2. Struck, T. H., Nesnidal, M. P., Purschke, G. & Halanych, K. M. Detecting possibly saturated positions in 18S and 28S sequences and their influence on phylogenetic reconstruction of Annelida (Lophotrochozoa). *Mol. Phylogenet. Evol.* **48,** 628–645 (2008).
3. Struck, T. H. *et al.* Annelida phylogeny and the status of Sipuncula and Echiura. *BMC Evol. Biol.* **7,** 57 (2007).
4. McHugh, D. Molecular evidence that echiurans and pogonophorans are derived annelids. *Proc. Natl Acad. Sci. USA* **94,** 8006–8009 (1997).
5. Raible, F. *et al.* Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science* **310,** 1325–1326 (2005).
6. Tessmar-Raible, K. & Arendt, D. Emerging systems: between vertebrates and arthropods, the Lophotrochozoa. *Curr. Opin. Genet. Dev.* **13,** 331–340 (2003).
7. Rivera, A. & Weisblat, D. And Lophotrochozoa makes three: *Notch/Hes* signaling in annelid segmentation. *Dev. Genes Evol.* **219,** 37–43 (2009).
8. Shain, D. H. *Annelids in Modern Biology* (Wiley, 2009).
9. Erséus, C. Phylogeny of oligochaetous Clitellata. *Hydrobiologia* **535–536,** 357–372 (2005).
10. McHugh, D. Molecular systematics of polychaetes (Annelida). *Hydrobiologia* **535–536,** 309–318 (2005).
11. Fauvel, P. Polychètes errantes. *Faune de France* **5,** 1–488 (1923).
12. Fauvel, P. Polychètes sédentaires. *Faune de France* **16,** 1–494 (1927).
13. de Quatrefages, A. M. *Histoire Naturelle des Annelides, Marine et d'Eau Douce. Annelides et Gephyriens* Vol. 1 (Librairie Encyclopédique de Roret, 1866).

14. Day, J. H. *A Monograph on the Polychaeta of Southern Africa. Part 1. Errantia* (British Museum (Natural History), 1967).
15. Rouse, G. W. & Fauchald, K. Cladistics and polychaetes. *Zool. Scr.* **26,** 139–204 (1997).
16. Eibye-Jacobsen, D. A reevaluation of *Wiwaxia* and the polychaetes of the Burgess Shale. *Lethaia* **37,** 317–335 (2004).
17. Rouse, G. W. & Pleijel, F. *Polychaetes* (Oxford Univ. Press, 2001).
18. Bleidorn, C. *et al.* On the phylogenetic position of Myzostomida: can 77 genes get it wrong? *BMC Evol. Biol.* **9,** 150 (2009).
19. Bleidorn, C. *et al.* Mitochondrial genome and nuclear sequence data support Myzostomida as part of the annelid radiation. *Mol. Biol. Evol.* **24,** 1690–1701 (2007).
20. Eeckhaut, I., Fievez, L. & Müller, M. C. M. Larval development of *Myzostoma cirriferum* (Myzostomida). *J. Morphol.* **258,** 269–283 (2003).
21. Westheide, W. The direction of evolution within the Polychaeta. *J. Nat. Hist.* **31,** 1–15 (1997).
22. Suschenko, D. & Purschke, G. Ultrastructure of pigmented adult eyes in errant polychaetes (Annelida): implications for annelid evolution. *Zoomorphology* **128,** 75–96 (2009).
23. Fauchald, K. & Jumars, P. A. The diet of worms: a study of polychaete feedings guilds. *Oceanogr. Mar. Biol. Annu. Rev.* **17,** 193–284 (1979).
24. Christodoulou, F. *et al.* Ancient animal microRNAs and the evolution of tissue identity. *Nature* **463,** 1084–1088 (2010).
25. Hausdorf, B. *et al.* Spiralian phylogenomics supports the resurrection of Bryozoa comprising Ectoprocta and Entoprocta. *Mol. Biol. Evol.* **24,** 2723–2729 (2007).
26. Ebersberger, I., Strauss, S. & von Haeseler, A. HaMStR: profile Hidden Markov Model based search for orthologs in ESTs. *BMC Evol. Biol.* **9,** 157 (2009).
27. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).
28. Katoh, K., Kuma, K.-i., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33,** 511–518 (2005).
29. Hartmann, S. & Vision, T. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evol. Biol.* **8,** 95 (2008).
30. Smith, S. A. & Dunn, C. W. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24,** 715–716 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to I. Ebersberger, S. Strauss and A. von Haeseler for the processing of our raw EST libraries. We also thank M. Aguado for species identification of *Typosyllis pigmentata*, as well as K. M. Halanych and P. A. Ramey-Balci for suggestions. T.H.S. and C.B. acknowledge support from the marine biological stations in Bamfield, Helgoland, List and Roscoff for collection of annelids. This work was funded by the priority programme 'Deep Metazoan Phylogeny' of the Deutsche Forschungsgemeinschaft by grants DFG-STR 683/5-2 (T.H.S.), DFG-Pu-84/5-1 (G.P. and T.H.S.), DFG-Tl-349/4-1 (R.T.), DFG Li 998/3-1 (B.L.), DFG-HA 5744/1-1 (S.H.) and DFG-BL-787/2-1 (C.B.).

Author Contributions T.H.S., G.P., R.T. and C.B. conceived this study. T.H.S. took the lead on data collection of sedentary polychaetes, and writing. T.H.S. and S.H. performed phylogenomic analyses. C.H. aided in the data collection of Sedentaria. C.B. and C.P. took the lead on data collection of errant polychaetes, and C.B., S.H. and N.H. on compilation of the data sets from the EST libraries. A.M. and B.L. generated the EST library of *Sipunculus nudus*, and M.K. was responsible for the sequencing of the EST libraries. T.H.S., G.P., R.T. and C.B. were the main contributors to the writing of the manuscript.

Author Information Sequence data have been deposited in the NCBI Expressed Sequence Tag database (dbEST) under accession numbers FN424437–FN428571, FR754554–FR771822, HQ729923–HQ729975. The largest aligned data set has been deposited at http://www.treebase.org. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.H.S. (struck@biologie.uni-osnabrueck.de) or C.B. (bleidorn@uni-leipzig.de).

METHODS

General outline. EST libraries were prepared for 17 annelid species, and they were used in combination with EST or genomic data from 17 additional annelid species and 5 outgroup species from public archives, and further processed as described previously[25]. Sets of orthologous genes were determined using the program HaMStR in combination with the InParanoid database (without ribosomal proteins)[26], and were translated into amino acid sequences using the program ESTwise[27]. In parallel, we retrieved all ribosomal proteins from databases as described previously[25]. Each orthologous gene set was aligned using MAFFT software[28] and masked using the program REAP[29]. Phylogenetic trees were constructed by using a Bayesian inference approach and a maximum likelihood approach. Stabilities of taxa were assessed using the leaf stability index as calculated by Phyutility software[30]. The five annelid taxa with an index below 0.925 were removed, and the Bayesian inference analysis was repeated. Branch attachment frequencies of these unstable annelid taxa were assessed using the lineage movement option in Phyutility[30].

Data assembly. Supplementary Table 1 lists taxa (34 annelids and 5 outgroup taxa) used in this study. On collection, samples were frozen at −80 °C. Total RNA was isolated using an RNeasy Plant Mini Kit (Qiagen) and then reverse transcribed to double-stranded cDNA with the Mint- Universal cDNA synthesis kit (Evrogen) to produce amplified cDNA libraries. The cDNA was size fractioned using CHROMA SPIN-1000 (Clontech). SfiI-digested cDNA allowed directional cloning into pDNR-lib. On average, 1,370 clones—ranging from 368 in *Ophelia limacina* to 4,135 in *Myzostoma cirriferum*—were successfully 5′-end sequenced from recombinant plasmids (at the htpt group of R. Reinhardt) by using Sanger-based sequencing technology. For *Glycera tridactyla*, sequences were generated with 454 technology by LGC Genomics. All original sequence data have been deposited in NCBI dbEST[31].

Recent studies successfully used ribosomal proteins obtained from EST databases to resolve deep metazoan phylogeny[1,25,32,33]. Therefore, ribosomal protein sequences were extracted from these EST data (Supplementary Table 2) using the human ribosomal proteome (retrieved from the Ribosomal Protein Gene Database[34]) as a search template during local BLAST searches (tblastn algorithm and an e-value $<e^{-10}$ as a match criterion). To substantially increase the amount of data, we also determined sets of orthologous genes using the program HaMStR[26], which derives a set of primer taxa from the InParanoid database[35], generating a set of core orthologous genes to build, train and calibrate a profile Hidden Markov Model. This model is then used to search for orthologues in the EST data. Further confirmation of the orthology of determined EST sequences was achieved in a final step of a reciprocal BLAST search against the proteome of one of the primer taxa, ideally the closest relative of the primer taxa to the query taxon. Orthology was accepted only if the same gene was retrieved as the best hit as in the set of core orthologous genes. We used the following set of primer taxa: *Capitella teleta*, *Helobdella robusta*, *Lottia gigantea*, *Schistosoma mansoni*, *Daphnia pulex*, *Apis mellifera* and *Caenorhabditis elegans*. The nucleotide sequence was translated into amino acids using ESTwise[27], and each set of orthologous genes was individually aligned using MAFFT[28] with default settings. Questionably aligned positions were eliminated with the alignment masker REAP[29] for each individual partition using default parameters.

For the sensitivity analyses, we generated three super-matrices based on taxon coverage per gene. The first matrix consisted of genes that were present in at least one-third of the taxa. In the second, the genes were present in at least one-half of the taxa, and in the third, the genes were present in at least two-thirds of the taxa. Thus, matrix coverage increased from the first to the third super-matrix, but the number of positions decreased. Custom Perl scripts were written for all of these steps. The data set consisting of genes that were present in at least one-third of the taxa was deposited at http://www.treebase.org and can be accessed at http://purl.org/phylo/treebase/phylows/study/TB2:S10986. Together with the information provided in Supplementary Table 6 and the Supplementary Information, all data sets used in the course of our analyses can be generated from this data set.

Phylogenetic analyses. The most appropriate substitution model for these three matrices was LG + I + Γ as determined based on the Akaike information criterion using ProtTest[36]. Before the time-consuming Bayesian inference analyses, we conducted a series of maximum likelihood analyses to assess the influence of the number of positions, the percentage of missing data and leaf stability on BS. Therefore, taxa that had less than 15%, 17.5% or 20% of the total positions in the largest super-matrix were excluded (Supplementary Table 2). Similarly, taxa with a leaf stability index of less than 0.875, 0.900 or 0.925 were excluded from the three super-matrices (Supplementary Table 3). We did not exclude annelid taxa with an index less than 0.950 because this was above the mean leaf stability of 0.943. Moreover, we also prepared one data set excluding only Myzostomida from the largest data set with 47,953 positions. Finally, we partitioned this data set based on our two strategies to assemble the data set. The first data set comprised only the ribosomal proteins; the second, the genes that were identified by HaMStR, without any ribosomal proteins; and the third, all HaMStR-identified genes, including the ribosomal genes (which were also detected by HaMStR). Thus, we had a total of 25 data sets (Supplementary Table 4). Maximum likelihood analyses were conducted with RAxML version 2.7.6 (ref. 37), using 100 replicate searches starting from randomized maximum parsimony trees. Confidence values for the edges of the maximum likelihood trees were determined based on bootstrap replicates. We used the automatic bootstopping option[38] (-# autoMRE) in RAxML to a maximum of 1,000 replicates (Supplementary Table 4). Leaf stability indices, as well as lineage movements of the unstable taxa, were determined using Phyutility[30] and the bootstrap trees of the analyses comprising all taxa.

On the basis of the results of the sensitivity analyses, we conducted two Bayesian inference analyses using PhyloBayes v3.2d[39] and the site-heterogeneous CAT model (which is not available for RAxML), as it has been shown that this model is more robust against LBA artefacts and thus less prone to systematic errors in phylogenetic data sets[40]. One data set comprised all 39 taxa and 47,953 positions, and in the other all annelids showing a leaf stability index below 0.925 were excluded (34 taxa, 47,953 positions). Each analysis ran eight chains in parallel for 29,525 cycles on average (ranging from 28,894 to 29,808) for the data set with 39 taxa and for 34,693 on average (ranging from 33,560 to 34,944) for the one with 34 taxa. To conduct these analyses, we used Mac OS X v10.6.4 with 2 × 2.93 GHz Quad-core Intel Xeon processors and 16 GB, 1,066 MHz DDR3 RAM. Using all eight processors in parallel, the two PhyloBayes analyses ran for 37 days, which is equivalent to nearly 10,000 h of CPU time. Stable convergence of likelihood values, alpha parameter and tree length of the eight chains was assessed using Tracer v1.4.1 (ref. 41), and if we had sampled nearly two times more trees than would be discarded as burn-in, this was taken as a stopping point. The first 10,000 cycles (trees) of each chain were discarded as burn-in, and the majority rule consensus tree containing the PPs was calculated from the remaining trees of the eight chains of each Bayesian inference analysis, sampling each second tree. Thus, the consensus trees are based on a total of 78,099 or 98,771 trees, respectively. We also tested *a posteriori* whether the CAT model was superior to the LG model in the PhyloBayes analyses using the cross-validation test[42] implemented in PhyloBayes. On the basis of the data set comprising 39 taxa and 47,953 positions, this test was conducted using ten replicates with a tenfold cross-validation. This means that the learning alignment consisted of 90% of the positions of the original alignment, and the test alignment consisted of the remaining 10%. The tests were run using the tree shown in Fig. 1 as a fixed topology, as suggested by the manual, and 1,100 cycles with a burn-in of 100.

Ancestral state reconstruction. For the ancestral reconstructions, we used a morphological data matrix reported previously[43], which is largely based on previously published data matrices[15,44]. We slightly modified this matrix (Supplementary Information) by updating/changing the coding of characters related to "shape of parapodia", "pygidial cirri", "uncini", "hooks" and "presence of eyes" according to the literature[17,22,43,45–47]. Instead of the character "aciculae", we coded the presence of internalized supporting chaetae[17].

Ancestral reconstructions were done for the last common ancestor of Annelida, clade 1, Sedentaria and Errantia based on the Bayesian inference as well as the maximum likelihood tree of the 39-taxa data set with 47,953 positions, using Mesquite v2.72 (ref. 48). We used the parsimony reconstruction option, and all characters were regarded as unordered. Sipuncula and Echiura have lost nearly all of their morphological annelid characters. However, it is well known that severe secondary losses of characters can strongly hamper reconstructions based on morphological data because they cannot easily be differentiated from primary absence[49–55]. Therefore, we did not consider these two taxa in the ancestral reconstructions.

To visualize the results of the ancestral reconstructions (Supplementary Table 5), we drew graphical depictions of body and parapodial characters using a basic schematic approach (Fig. 2). We refrained from using a representative approach for two reasons. First, no family of recent polychaetes shows only all of the characters of any of the four ancestral reconstructions. Second, a representative approach using, for example a recent polychaete family, has the potential to mislead in that this family might be taken to fully represent basal Annelida, Sedentaria or Errantia. However, each recent taxon is a patchwork of plesiomorphic and apomorphic characters and is as closely or distantly related to an ancestor in evolutionary times as any other recent descendant of that ancestor is.

For each of the four clades, we used a schematic representation of a homonomously segmented worm with parapodia. For Annelida and clade 1, this worm also had grooved palps, nuchal organs and bicellular eyes (Fig. 2a). The reconstruction of parapodial features was uncertain except for the composition of chaetae, because the last common ancestors of both clades had only simple chaetae and internalized supporting chaetae. Therefore, we depicted the two extreme possibilities: we included either all features that were eventually present in the ground pattern (such as large dorsal notopodial and ventral neuropodial lobes and dorsal and ventral cirri (sensory parapodial appendages)) or only those features that, on the basis of the reconstruction, were definitely present (such as a large

notopodial and a small neuropodial lobe). For Sedentaria, the parapodia were reduced, with a small neuropodial lobe and the absence of internalized supporting chaetae. The uncertain presence of uncini/hooks in the ground pattern of Sedentaria is indicated by a question mark (Fig. 2c). For Errantia, we also inferred one pair of antennae and pygidial cirri (sensory appendages at the body end), multicellular eyes (instead of bicellular eyes) and solid, sensory palps, which moved from a dorsal to a ventral position as is typical for such palps (Fig. 2b). The parapodia consisted of large notopodial and neuropodial lobes and ventral cirri. The presence of dorsal cirri in the ground pattern of Errantia was uncertain, so we have shown them with dashed lines only.

31. *NCBI dbEST (Expressed Sequence Tags Database)* ⟨http://www.ncbi.nlm.nih.gov/projects/dbEST/⟩ (2010).
32. Helmkampf, M., Bruchhaus, I. & Hausdorf, B. Phylogenomic analyses of lophophorates (brachiopods, phoronids and bryozoans) confirm the Lophotrochozoa concept. *Proc. R. Soc. Lond. B* **275,** 1927–1933 (2008).
33. Struck, T. H. & Fisse, F. Phylogenetic position of Nemertea derived from phylogenomic data. *Mol. Biol. Evol.* **25,** 728–736 (2008).
34. *Ribosomal Protein Gene Database* ⟨http://ribosome.med.miyazaki-u.ac.jp⟩ (2010).
35. *InParanoid: Eukaryotic Ortholog Groups (100 organisms: 1687023 sequences)* ⟨http://inparanoid.sbc.su.se/cgi-bin/index.cgi⟩ (2010).
36. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21,** 2104–2105 (2005).
37. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
38. Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R. P., Moret, B. M. E. & Stamatakis, A. in *RECOMB 2009, LNCS 5541* (ed. Batzoglou, S.) 184–200 (Springer, 2009).
39. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21,** 1095–1109 (2004).
40. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7,** S4 (2007).
41. Rambaut, A. & Drummond, A. J. *Tracer v1. 4* ⟨http://beast.bio.ed.ac.uk/Tracer⟩ (2007).
42. Zhou, Y., Rodrigue, N., Lartillot, N. & Philippe, H. Evaluation of the models handling heterotachy in phylogenetic inference. *BMC Evol. Biol.* **7,** 206 (2007).
43. Zrzavý, J., Riha, P., Pialek, L. & Janouskovec, J. Phylogeny of Annelida (Lophotrochozoa): total-evidence analysis of morphology and six genes. *BMC Evol. Biol.* **9,** 189 (2009).
44. Rouse, G. W. Trochophore concepts: ciliary bands and the evolution of larvae in spiralian Metazoa. *Biol. J. Linn. Soc.* **66,** 411–464 (1999).
45. Hartmann-Schröder, G. *Teil 58. Annelida, Borstenwürmer, Polychaeta* 2nd edn (Gustav Fischer, 1996).
46. Westheide, W. & Rieger, R. M. *Spezielle Zoologie. Erster Teil: Einzeller und Wirbellose Tiere* (Gustav Fischer, 1996).
47. Purschke, G., Arendt, D., Hausen, H. & Müller, M. C. M. Photoreceptor cells and eyes in Annelida. *Arthropod Struct. Dev.* **35,** 211–230 (2006).
48. Maddison, W. P. & Maddison, D. R. Mesquite: a modular system for evolutionary analysis. Version 2.71. *Mesquite Project* ⟨http://mesquiteproject.org⟩ (2009).
49. Purschke, G., Hessling, R. & Westheide, W. The phylogenetic position of the Clitellata and the Echiura — on the problematic assessment of absent characters. *J. Zool. Syst. Evol. Res.* **38,** 165–173 (2000).
50. Wiens, J. J. Does adding characters with missing data increase or decrease phylogenetic accuracy? *Syst. Biol.* **47,** 625–640 (1998).
51. Wiens, J. J., Bonett, R. M. & Chippindale, P. T. Ontogeny discombobulates phylogeny: paedomorphosis and higher-level salamander relationships. *Syst. Biol.* **54,** 91–110 (2005).
52. Bleidorn, C. The role of character loss in phylogenetic reconstruction as exemplified for the Annelida. *J. Zool. Syst. Evol. Res.* **45,** 299–307 (2007).
53. Bleidorn, C., Hill, N., Erséus, C. & Tiedemann, R. On the role of character loss in orbiniid phylogeny (Annelida): molecules vs. morphology. *Mol. Phylogenet. Evol.* **52,** 57–69 (2009).
54. Struck, T. H. Progenetic species in polychaetes (Annelida) and problems assessing their phylogenetic affiliation. *Integr. Comp. Biol.* **46,** 558–568 (2006).
55. Struck, T. H. Data congruence, paedomorphosis and salamanders. *Front. Zool.* **4,** 22 (2007).