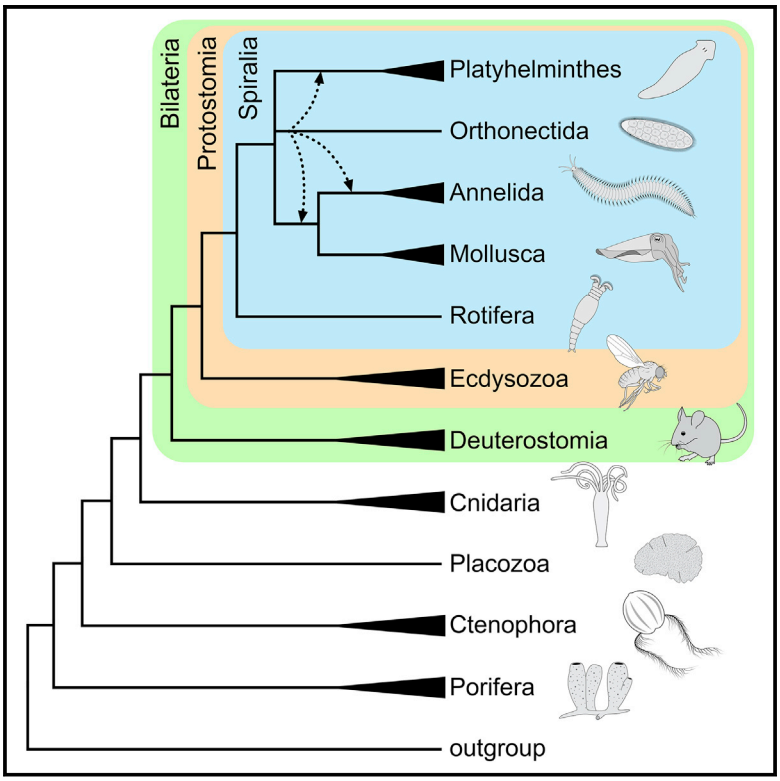
Report

Current Biology

The Genome of *Intoshia linei* Affirms Orthonectids as Highly Simplified Spiralians

Graphical Abstract



Authors

Kirill V. Mikhailov, Georgy S. Slyusarev, Mikhail A. Nikitin, Maria D. Logacheva, Aleksey A. Penin, Vladimir V. Aleoshin, Yuri V. Panchin

Correspondence

ypanchin@yahoo.com

In Brief

The genomic data presented by Mikhailov et.al show that orthonectids, a group of highly simplified parasitic animals, are true bilaterians related to Lophotrochozoa and possess a reduced complement of genes implicated in the metazoan development and nervous system activity.

Highlights

- The orthonectid *Intoshia linei* has a small genome with only around 9,000 genes
- The phylogenomic analysis affirms orthonectids as highly simplified spiralians
- The orthonectid simplification is associated with reduction of developmental genes
- They have a compact genetic toolkit for the nervous system development and activity

Mikhailov et al., 2016, Current Biology *26*, 1768–1774
July 11, 2016
http://dx.doi.org/10.1016/j.cub.2016.05.007

CellPress

Current Biology
Report

The Genome of *Intoshia linei* Affirms Orthonectids as Highly Simplified Spiralians

Kirill V. Mikhailov,[1,2] Georgy S. Slyusarev,[3] Mikhail A. Nikitin,[1] Maria D. Logacheva,[1,2] Aleksey A. Penin,[1,2] Vladimir V. Aleoshin,[1,2] and Yuri V. Panchin[1,2,*]

[1]Belozersky Institute for Physico-Chemical Biology, Lomonosov Moscow State University, Moscow 119991, Russian Federation
[2]Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russian Federation
[3]Saint Petersburg State University, St. Petersburg 199034, Russian Federation
*Correspondence: ypanchin@yahoo.com
http://dx.doi.org/10.1016/j.cub.2016.05.007

SUMMARY

Orthonectids are rare parasites of marine invertebrates [1] that are commonly treated in textbooks as a taxon of uncertain affinity [2]. Trophic forms of orthonectids reside in the tissues of their hosts as multinucleated plasmodia, generating short-lived, worm-like ciliated female and male organisms that exit into the environment for copulation [3]. These ephemeral males and females are composed of just several hundred somatic cells and are deprived of digestive, circulatory, or excretory systems. Since their discovery in the 19th century, the orthonectids were described as organisms with no differentiated cell types and considered as part of Mesozoa, a putative link between multicellular animals and their unicellular relatives. More recently, this view was challenged as the new data suggested that orthonectids are animals that became simplified due to their parasitic way of life [3, 4]. Here, we report the genomic sequence of *Intoshia linei*, one of about 20 known species of orthonectids. The genomic data confirm recent morphological analysis asserting that orthonectids are members of Spiralia and possess muscular and nervous systems [5]. The 43-Mbp genome of *I. linei* encodes about 9,000 genes and retains those essential for the development and activity of muscular and nervous systems. The simplification of orthonectid body plan is associated with considerable reduction of metazoan developmental genes, leaving what might be viewed as the minimal gene set necessary to retain critical bilaterian features.

RESULTS AND DISCUSSION

The Genome of *Intoshia linei* Is One of the Smallest among Metazoans

The genome of *Intoshia linei* was sequenced using the Illumina platform and assembled into a draft totaling 43.2 Mbp. The miniature by metazoan standards genome is predicted to encode about 9,000 genes—one of the lowest reported gene counts among metazoans, exceeding only the recently sequenced genomes of a myxozoan [6] and a plant-parasitic nematode [7]. The predicted genes are fairly intron rich and span an average of six exons, which in sum account for 23.1% of the genome. *I. linei* has some of the shortest introns seen in metazoans: the distribution of intron lengths peaks at 37 bp, and nearly a half of all introns are within a 30–50 bp size range (Figure S1). The genome compaction in *I. linei* is reflected in both lower gene count and higher average gene density in comparison with other metazoan genomes. However, it is not the most gene dense among metazoans: the average gene density in the genome of *I. linei* is around 200 genes per Mb, on par with that of *Caenorhabditis elegans* [8] but noticeably lower than the gene density in the highly compact genome of a parasitic nematode *Trichinella spiralis* [9] or a pelagic tunicate *Oikopleura dioica* [10, 11] (Figure S1). Despite its comparatively small size, the genome of *I. linei* carries a considerable amount of repetitive elements—more than a quarter of the total assembly size, with the largest contribution to the repetitive element repertoire provided by unclassified repeats (Table S1).

The Orthonectids Are Highly Simplified Spiralians

The orthonectid genes display an exceptionally high rate of sequence divergence, which is known to have a confounding effect on the phylogenetic inference [12, 13]. Previous studies using rRNA-based phylogenies placed Orthonectida within Bilateria but failed to determine their affiliation with any specific bilaterian taxon due to their highly divergent sequences [14, 15]. We investigated the phylogenetic position of *I. linei* using a 500-gene dataset and the tree inference methods of RAxML [16] and PhyloBayes [17]. To counteract the impact of systematic biases stemming from uneven evolutionary rates in the dataset, we performed tree reconstructions following the removal of homoplasy-prone fast-evolving sites and employing site-heterogeneous substitution model [18], which was shown to be more robust in dealing with reconstruction artifacts [19]. The obtained phylogenies agree on placing the orthonectid within Spiralia, a result supported by the morphological study [3], but nevertheless show discord in its exact phylogenetic placement. The maximum likelihood analysis or the Bayesian inference with a site-heterogeneous substitution model and a general time reversible exchange rate matrix (CAT-GTR) group *I. linei* with flatworms, while the analyses using the CAT model with flat exchange rates place it sister to annelids (Figures 1A, 1B, S2, and S3). To examine the phylogenetic relationship of *I. linei* within Spiralia in greater detail, we used the dataset of Struck et al.

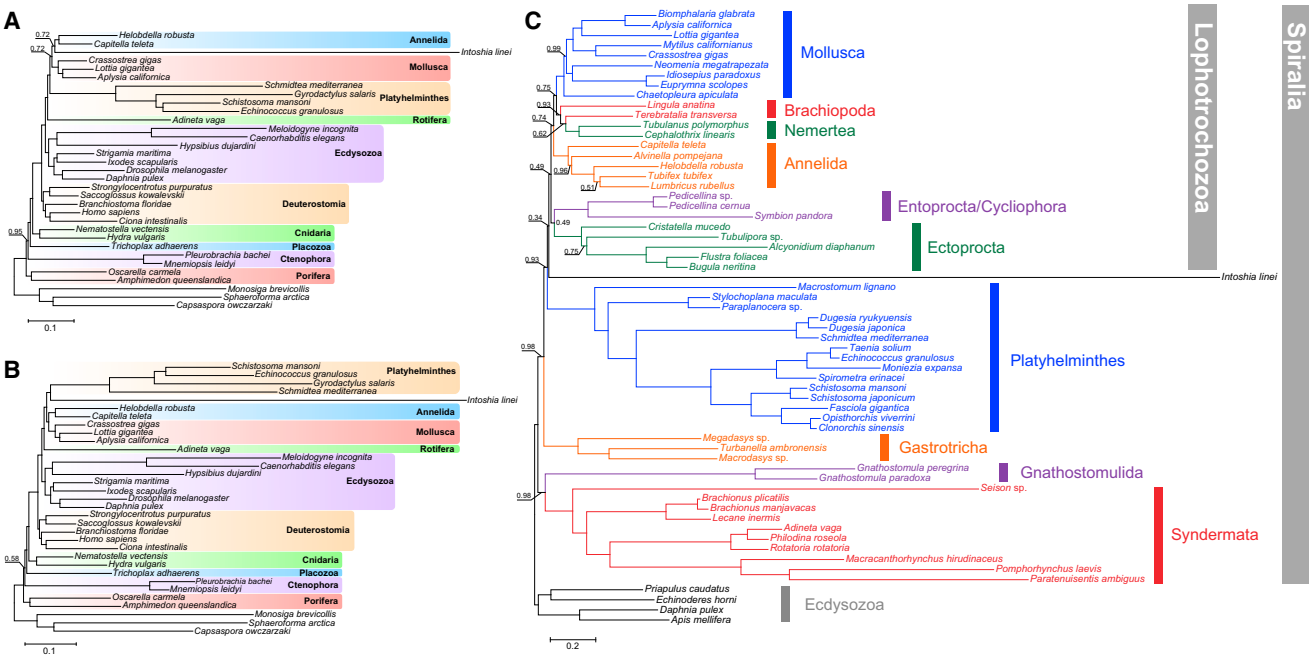


Figure 1. The Contentious Position of *Intoshi linei* in the Metazoan Phylogeny

(A and B) The trees in (A) and (B) were reconstructed by PhyloBayes with a dataset of slow-evolving sites (47,548 amino acid [aa] sites), assembled from 500 orthologous groups utilizing genomic data. The Bayesian inference was performed using the CAT + Γ4 model (A) and the CAT + GTR + Γ4 model (B).

(C) The spiralian phylogeny reconstructed on the basis of the dataset assembled by Struck et al. [20], relying primarily on transcriptomic data. The tree was reconstructed by PhyloBayes under the GTR + CAT + Γ4 model with a 22,909 aa site alignment that excludes positions with over 60% missing data. Support indexes for nodes with 1.00 posterior probability are omitted.

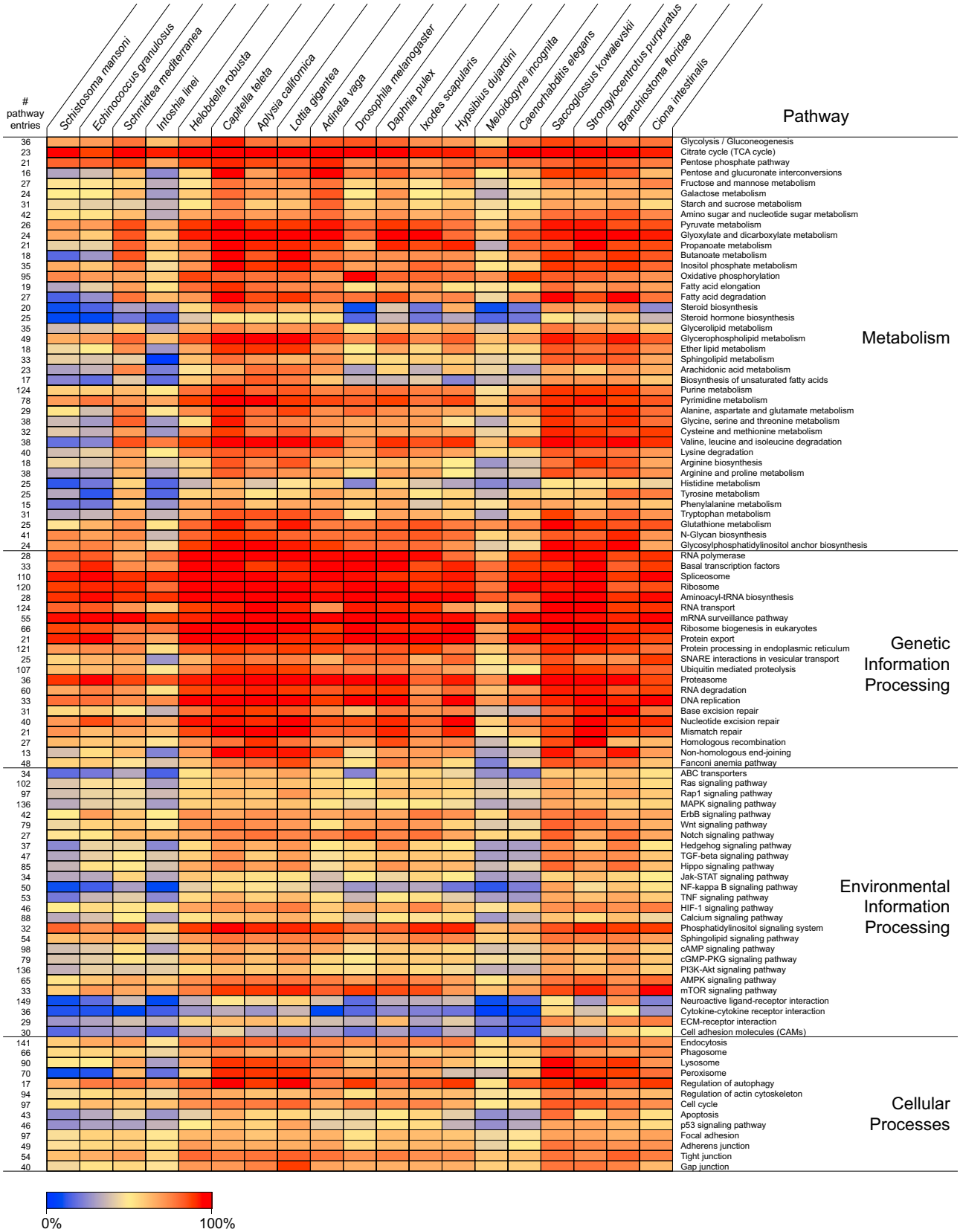
See also Figures S2 and S3.

[20] that focuses on taxonomic sampling of the spiralian taxa. The spiralian tree reconstructed by PhyloBayes (Figure 1C) is largely congruent with the previously reported multigene phylogenies [20, 21] but differs in some aspects, particularly in the position of gastrotrichs and the branching of the lophotrochozoan taxa. In the spiralian phylogeny, the orthonectid forms a lineage intercalating the branch that separates the classical Lophotrochozoa [22] from the rest of spiralians, including flatworms and rotifers. The latter result argues for an isolated position of orthonectids among the spiralian taxa, proposing a third alternative for the contested position within the group.

The spiralian ancestry of orthonectids implies a derived condition for the apparent simplicity of their organization and attributes meager gene complement of *I. linei* to extensive loss. Nearly 75% of orthologous groups inferred to be present in the spiralian ancestor are lost by the orthonectid (Figure S3D). The genome of *I. linei* retains only around 4,000 conserved ancestral orthologous groups—a thousand less than the corresponding number in the parasitic flatworms. A third of all predictions in the genome of *I. linei* find no hits in the NCBI's non-redundant database and are seen as orthonectid innovations. Among the domain families not associated with mobile genetic elements, the only family that appears to have expanded is a family of lecithin:cholesterol acyltransferase domains, which participate in lipid metabolism. Mapping the retained genes to molecular pathways shows that *I. linei* has a functional complement of components for glycolysis, pentose phosphate pathway, tricarboxylic acid cycle, and oxidative phosphorylation, but a range of metabolic pathways are likely impaired, including de novo synthesis of purine nucleotides and most amino acids (Figure 2). Almost all components of pathways for steroid biosynthesis and sphingolipid metabolism are missing. Similarly to the parasitic flatworms [23], *I. linei* lost most of the peroxisome components and may be devoid of the organelle itself (see Supplemental Information for details). The signaling pathways of NF-κB and STAT, implicated in immunity, growth, and development, also appear to be completely lost. The developmental signaling pathways of Wnt, Notch, and TGF-β seem to be intact, but the key elements of the Hedgehog pathway are missing (Table 1). It is notable that while most pathways that experience reduction in the parasitic flatworms also appear to be impaired in the orthonectid, the components of pathways for fatty acid metabolism and branched-chain amino acid degradation are preserved in *I. linei* but are absent from many of the parasitic flatworms [24].

The Orthonectid Simplification Is Associated with Reduction of Metazoan Developmental Genes

The repertoire of transcription factors in metazoan genomes is one of the properties that correlate with organismal complexity [25]. It is therefore unsurprising that the genome of *I. linei* has the minimal count of recognized transcription factors among bilaterian animals (Table 1). The families of C2H2 type zinc-finger proteins experience drastic contraction in *I. linei*, while the p53 homologs are absent from its genome altogether. The number of homeobox genes in the orthonectid is close to the number seen in parasitic flatworms—another group with extensively



(legend on next page)

Table 1. Transcription Factor Families and Signaling Pathway Ligands Implicated in the Metazoan Development

	I. linei	*E. granulosus*	*C. elegans*	*D. melanogaster*	*C. teleta*	*H. sapiens*
Homeobox	61	65	94	101	158	245
Forkhead box	20	15	15	18	42	50
High mobility group box	12	17	16	22	24	51
T-box	7	7	21	8	8	17
MADS box	2	4	2	2	2	5
Basic helix-loop-helix	24	25	38	53	82	105
Basic leucine zipper	8	19	22	16	26	41
C2H2 type zinc finger	75	153	186	296	443	803
ETS	3	9	10	8	13	29
Nuclear receptor	9	10	260	18	32	47
Rel/NF-κB	0	0	0	3	2	5
NFAT	0	1	0	1	1	5
SMAD	4	5	7	4	4	8
STAT	0	0	1	1	6	7
Wnt	3	6	5	7	12	19
TGF-β	4	3	5	7	14	37
Hedgehog	0	1	0	1	1	3
DSL ligand	1	3	10	2	6	5
Fibroblast growth factor	1	0	2	3	1	22

The gene counts are based on the analyses of protein domains, which were performed identically for the listed genomes. The majority of proteins were detected using the Pfam domain annotation with a gathering cutoff threshold; the C2H2 type zinc-finger proteins were detected using the InterPro domains (IPR007087 and IPR015880). Although the absence of the canonical Hedgehog signaling protein is not indicative of the whole pathway status, the genome of *I. linei* also lacks the Hint domain and orthologs of Patched, Smoothened, and Ci/GLI, which are integral to the Hedgehog pathway. *I. linei*, orthonectid; *E. granulosus*, flatworm; *C. elegans*, nematode; *D. melanogaster*, insect; *C. teleta*, annelid; *H. sapiens*, chordate. See also Figure S4.

reduced homeobox gene complement [23]. A closer look at their homeobox genes reveals that roughly a half of the retained families overlap between the orthonectid and parasitic flatworms (Figure S4). The genome of *I. linei* shows conservation of at least 37 homeobox gene families, which among others include *pax6*, *engrailed*, *sine oculis*, and *otx* orthologs. Notably, it encodes only three Hox type genes, which are known to play a pivotal role in regulating differentiation along the main body axis in bilaterians [26]. The Hox genes in *I. linei* represent the anterior and central Hox2, Hox4, and Hox6-8 families. The orthonectid Hox genes are located in different contigs and are neighbored by unrelated genes, which implies that unlike many of their bilaterian orthologs, they are not organized in a cluster. We found no posterior Hox genes in *I. linei*, which are usually conserved in bilaterians with one exception of a rotifer *A. vaga* [27]. A single ParaHox type posterior gene, *caudal/Cdx* ortholog, is present in *I. linei* genome. Aside from the transcription factors, an important part in regulation of gene expression in many metazoans is mediated by microRNAs [28]. The genome of *I. linei* encodes key elements of the microRNA pathway, including the Argonaute, Piwi, and Dicer orthologs, and components of the microprocessor complex, Drosha and Pasha orthologs. Presence of these genes might be an indication of a functional microRNA regulation system, although experimental data are needed to confirm its existence in the orthonectid.

Intoshia linei Retains a Compact Gene Set for the Nervous System Activity and Development

While the name Orthonectida suggests that they swim in a straightforward manner, in reality, *I. linei* exhibits complex movements, including spinning and bending (Figure 3; Movie S1). The presence of muscular and nervous systems, involved in coordinating these movements in Orthonectida, was only recently recognized [5, 29, 30]. The excitable tissues, however simply organized, imply the existence of action potential generating ion channels, neurotransmitter receptors, and electrical synapses. We searched the genome of *I. linei* to document the conserved genes involved in the development and functioning of these systems.

Most types of ion channels, including the voltage-gated channels necessary for the generation of action potential, are present in *I. linei*, yet their number is reduced in comparison to other metazoans [31]. There are 42 predicted proteins of tetrameric sodium, potassium, and calcium ion channels in *I. linei*.

Figure 2. Heatmap of Pathway Conservation in Bilaterians

The reference pathways were selected from the KEGG pathways collection. The number of pathway entries corresponds to the total amount of non-redundant pathway elements found in the surveyed bilaterian genomes. The color in each cell depicts the percentage of these non-redundant entries found in a given genome. See also Figure S3D.

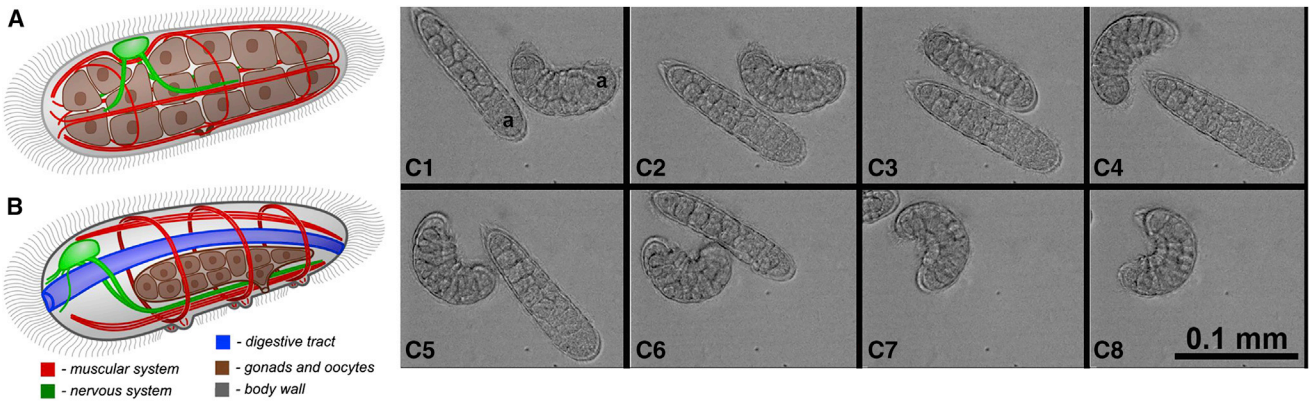


Figure 3. Orthonectid Body Plan and Locomotion

(A and B) Simplified scheme for the orthonectid body plan (A); the adult protostome bilaterian body plan (B). Orthonectids display typical bilaterian features such as the dorsal ganglion (brain) and have a muscular system and a single layer of ciliated epithelium cells (the only means for locomotion in Orthonectida). At the same time, they lack the digestive and excretory systems. The reproductive system of Orthonectida is also unusual: the germ cells are located in the body cavity without any gonad barrier.

(C1–C8) Two individuals of *I. linei* (adult females), which recently left their host, perform forward and reverse locomotion, propelled by means of numerous epithelial cilia. The organisms are seen turning and bending—the motions driven by coordinated muscle contractions. Time-lapse frames (1.25 s apart) are from Movie S1 (a, anterior part of the animal).

According to the Pfam analysis and reciprocal BLAST searches, one voltage-gated sodium ion channel and one voltage-gated calcium ion channel are present among the predicted proteins. *I. linei* has six predicted voltage-gated potassium channel proteins (one of them also had animal-specific KCNQ_channel signature and three proteins containing K+ channel tetramerisation domain BTB_2). One predicted *I. linei* protein is related both to the voltage-gated EAG K+ channels and to the cyclic nucleotide-gated cation channels. Three more proteins that encode cyclic nucleotide-binding domains and Ion_trans or Ion_trans_2 were predicted. One of them appears to have two copies of the cNMP_binding domain coexisting with the Ion_trans domains. The two-pore-domain inward rectifier potassium channels contribute to the resting potential and are known as ''leak channels.'' Three predicted proteins of this family were found in *I. linei*. Among other potassium channels, two calcium-activated BK potassium channel alpha subunit proteins and one calcium-activated SK potassium channel protein domains were identified. Notably, the SK protein family is present only in bilaterian animals. Two inward rectifier potassium channels (IRK) according to the reciprocal BLAST hits have the best similarity to the G protein-activated inward rectifier potassium channels that act in the seven transmembrane G protein-coupled receptor (GPCR) pathway.

The BLAST and Pfam searches detected 11 hits with the innexin/pannexin-specific Pfam domain (PF00876) in *I. linei*. Although gap junctions are found in many tissues, they often take part in the cell-to-cell communications and function as electrical synapses in excitable tissues.

An array of ionotropic and metabotropic receptor genes is present in the *I. linei*. We found 14 genes for neurotransmitter-gated ion-channels in *I. linei*, identified by the specific transmembrane region domain (PF02932) and the ligand binding domain (PF02931). Two of these receptors are predicted to be ionotropic glycine receptors, and the rest are identified as the nicotinic acetylcholine receptors (Table S2). The Pfam and BLAST searches and the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology assignments resulted in no hits for ionotropic glutamate receptors. Among the GPCRs, some hits indicate the presence of metabotropic glutamate receptors: five proteins with the domain of class 3 GPCRs (PF00003). Interestingly, the ionotropic neurotransmitter receptor proteins in the Orthonectida and Ctenophora, which are in some sense competing for the simplest nervous system in Metazoa, have very different profiles: while the orthonectid has lost the entire glutamate receptor family (iGluR), the ctenophores lack all ligand-gated Cys-loop receptors [32].

The locomotion in orthonectids relies on motile epithelial cilia, which are commonly activated by serotonin. Six of the *I. linei* neurons are also stained by the anti-serotonin antibodies [5]. Therefore, we expected to find elements of the serotonin (5-hydroxytryptamine [5-HT]) signaling in *I. linei*. We found no 5-HT-specific receptors of the ligand-gated ion channel family (PF02932), but two sequences for the 5-HT-specific GPCRs were detected and confirmed by bi-directional BLAST hits. Serotonergic neurons usually express serotonin symporter from the sodium:neurotransmitter symporter family (PF00209). This family of transporters is responsible for the synaptic recycling of neurotransmitters. The number of proteins from this family in *I. linei* is 27, which is comparable to the number seen in the genomes of animals with a developed nervous system. We were not able to reliably verify whether any of these transporters are specific for 5-HT. In summary, we found genes for ionotropic receptors to acetylcholine and glycine and genes for metabotropic receptors to acetylcholine, serotonin, histamine, dopamine, adrenalin, and glutamate, but no genes for receptors to GABA. The entire family of ionotropic glutamate receptors appears to be missing in the orthonectid.

The simplicity of *I. linei* nervous system is associated not only with the decrease in the receptor diversity but also with molecular mechanisms responsible for the nervous system development, axon guidance, and synapse formation. Semaphorins, important neuronal pathfinding signaling molecules, and their

receptors (plexins) are absent from the genome. This is also true for fasciclin domain involved in axonal guidance. At the same time, other players potentially involved in the nervous system development such as Netrin, Ephrins, Ephrin receptors, IgSF-CAMs, Cadherins, and Integrins are present.

While the core set of muscle proteins was already present before the emergence of animals, the troponin complex and titin appear to be an innovation specific to Bilateria and characteristic of bilaterian striated muscles [33]. Troponin is a complex of three proteins (troponin C, troponin I, and troponin T). These proteins are not detected in *I. linei* by BLAST search, and the troponin domain is not found by the Pfam search. The troponin complex in chordates is characteristic for skeletal and cardiac muscles, but not for smooth muscles. Morphological data suggest that *I. linei* muscles are similar to smooth muscles, so the troponin was likely lost in *I. linei*, and its absence is not a plesiomorphic trait. At the same time, another bilaterian hallmark, the myogenic regulatory factor, is present in the genome.

The Orthonectid Genome Retains Elements of the Metazoan Sensory Systems

Not much is known about the sensory systems in the orthonectids. The aquatic-stage *I. linei* females have a putative subepithelial receptor, which consists of three ciliated cells [3]. Additionally, ciliated and non-ciliated epithelial cells as well as neuronal processes may have a receptor function. We searched the genome of *I. linei* for clues on the putative sensory systems. Two predicted Piezo type proteins, ten Amiloride-sensitive sodium channels, two transient receptor potential (TRP) family proteins, and three TREK-1/TRAAK channel homologs could be potentially involved in mechanotransduction in *I. linei* by analogy to their use in *C. elegans*, *D. melanogaster*, and vertebrates. Some predicted *I. linei* proteins resemble TRPV and TRPM family members and may participate in temperature sensing. The presence or absence of photoreception in *I. linei* is not clearly confirmed by the genomic data. There are two distinct types of photoreceptive molecules in animals: 7-TM GPCRs class members (using retinal as chromophore) and flavoproteins cryptochromes. We found no proteins of the photolyase/cryptochrome family or flavin adenine dinucleotide (FAD) binding domain of DNA photolyase in *I. linei*, which excludes the common pathway for light sensing. The class of 7-TM GPCRs is extremely reduced in *I. linei* (comprising only 37 proteins), and photosensitive opsins are not easily recognized by bioinformatic approaches [34].

The morphological and genomic data clearly indicate that the simple organization of orthonectids is a derived trait associated with transition to obligate parasitism. Apart from the parasitic organisms comprising another enigmatic group, Rhombozoa, orthonectids represent an extreme case of simplification in Bilateria, which is reflected by the genome of *I. linei* in a remarkable extent of gene loss. The highly divergent sequences of *I. linei* remain a hindrance for phylogenetic inference, although it is worth noting that the affiliation of orthonectids with annelids obtained in some of our analyses was argued earlier on the basis of microvillar cuticle similarity and circular muscle metamery [3]. The orthonectid genome retains elements of the genetic toolkit for bilaterian development, which makes it a valuable object for evolutionary developmental biology as potentially the simplest model for the development of core bilaterian features, including muscular and nervous systems.

ACCESSION NUMBERS

The accession numbers for new data reported in this study are NCBI GenBank: Assembly GCA_001642005.1; NCBI BioProject: PRJNA316116; and NCBI BioSample: IDSAMN04576116.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, two tables, and one movie and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2016.05.007.

AUTHOR CONTRIBUTIONS

Conceptualization, K.V.M., V.V.A., and Y.V.P.; Methodology, K.V.M., V.V.A., and Y.V.P.; Investigation, K.V.M., V.V.A., M.A.N., and Y.V.P.; Writing – Original Draft, K.V.M, V.V.A., G.S.S., and Y.V.P.; Writing – Review & Editing, K.V.M, V.V.A., and Y.V.P.; Resources, G.S.S., M.A.N., M.D.L., and A.A.P.

ACKNOWLEDGMENTS

The authors thank A.C. Cherkasov, V.V. Starunov, and the staff of the Marine Biological Station of the Saint Petersburg State University for assistance in collecting material. The work was supported by Russian Scientific Foundation grant number 14-50-00150.

Received: April 1, 2016
Revised: May 2, 2016
Accepted: May 3, 2016
Published: June 30, 2016

REFERENCES

1. Kozloff, E.N. (1992). The genera of the phylum Orthonectida. Cah. Biol. Mar. *33*, 377–406.
2. Brusca, R.C., Moore, W., and Shuster, S.M. (2016). Invertebrates, Third Edition (Sinauer Associates).
3. Sliusarev, G.S. (2008). [Phylum Orthonectida: morphology, biology, and relationships to other multicellular animals]. Zh. Obshch. Biol. *69*, 403–427.
4. Ruppert, E.E., Fox, R.S., and Barnes, R.D. (2004). Invertebrate Zoology, A Functional Evolutionary Approach, Seventh Edition (Cengage Learning).
5. Slyusarev, G.S., and Starunov, V.V. (2016). The structure of the muscular and nervous systems of the female *Intoshia linei* (Orthonectida). Org. Divers. Evol. *16*, 65–71.
6. Chang, E.S., Neuhof, M., Rubinstein, N.D., Diamant, A., Philippe, H., Huchon, D., and Cartwright, P. (2015). Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proc. Natl. Acad. Sci. USA *112*, 14912–14917.
7. Burke, M., Scholl, E.H., Bird, D.M., Schaff, J.E., Colman, S.D., Crowell, R., Diener, S., Gordon, O., Graham, S., Wang, X., et al. (2015). The plant parasite *Pratylenchus coffeae* carries a minimal nematode genome. Nematology *17*, 621–637.
8. The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science *282*, 2012–2018.
9. Mitreva, M., Jasmer, D.P., Zarlenga, D.S., Wang, Z., Abubucker, S., Martin, J., Taylor, C.M., Yin, Y., Fulton, L., Minx, P., et al. (2011). The draft genome of the parasitic nematode *Trichinella spiralis*. Nat. Genet. *43*, 228–235.
10. Seo, H.C., Kube, M., Edvardsen, R.B., Jensen, M.F., Beck, A., Spriet, E., Gorsky, G., Thompson, E.M., Lehrach, H., Reinhardt, R., and Chourrout, D.

(2001). Miniature genome in the marine chordate *Oikopleura dioica*. Science *294*, 2506.

11. Danks, G., Campsteijn, C., Parida, M., Butcher, S., Doddapaneni, H., Fu, B., Petrin, R., Metpally, R., Lenhard, B., Wincker, P., et al. (2013). OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica*. Nucleic Acids Res. *41*, D845–D853.
12. Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. *27*, 401–410.
13. Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., and Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. *56*, 389–399.
14. Pawlowski, J., Montoya-Burgos, J.I., Fahrni, J.F., Wüest, J., and Zaninetti, L. (1996). Origin of the Mesozoa inferred from 18S rRNA gene sequences. Mol. Biol. Evol. *13*, 1128–1132.
15. Hanelt, B., Van Schyndel, D., Adema, C.M., Lewis, L.A., and Loker, E.S. (1996). The phylogenetic position of *Rhopalura ophiocomae* (Orthonectida) based on 18S ribosomal DNA sequence analysis. Mol. Biol. Evol. *13*, 1187–1191.
16. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.
17. Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. *62*, 611–615.
18. Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. *21*, 1095–1109.
19. Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. *7* (*Suppl 1*), S4.
20. Struck, T.H., Wey-Fabrizius, A.R., Golombek, A., Hering, L., Weigert, A., Bleidorn, C., Klebow, S., Iakovenko, N., Hausdorf, B., Petersen, M., et al. (2014). Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of spiralia. Mol. Biol. Evol. *31*, 1833–1849.
21. Laumer, C.E., Bekkouche, N., Kerbl, A., Goetz, F., Neves, R.C., Sørensen, M.V., Kristensen, R.M., Hejnol, A., Dunn, C.W., Giribet, G., and Worsaae, K. (2015). Spiralian phylogeny informs the evolution of microscopic lineages. Curr. Biol. *25*, 2000–2006.
22. Halanych, K.M., Bacheller, J.D., Aguinaldo, A.M., Liva, S.M., Hillis, D.M., and Lake, J.A. (1995). Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science *267*, 1641–1643.
23. Tsai, I.J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sanchez-Flores, A., Brooks, K.L., Tracey, A., Bobes, R.J., Fragoso, G., Sciutto, E., et al.; Taenia solium Genome Consortium (2013). The genomes of four tapeworm species reveal adaptations to parasitism. Nature *496*, 57–63.
24. Young, N.D., Nagarajan, N., Lin, S.J., Korhonen, P.K., Jex, A.R., Hall, R.S., Safavi-Hemami, H., Kaewkong, W., Bertrand, D., Gao, S., et al. (2014). The *Opisthorchis viverrini* genome provides insights into life in the bile duct. Nat. Commun. *5*, 4378.
25. Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature *424*, 147–151.
26. Lewis, E.B. (1978). A gene complex controlling segmentation in *Drosophila*. Nature *276*, 565–570.
27. Flot, J.F., Hespeels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E.G., Hejnol, A., Henrissat, B., Koszul, R., Aury, J.M., et al. (2013). Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. Nature *500*, 453–457.
28. Berezikov, E. (2011). Evolution of microRNA diversity and regulation in animals. Nat. Rev. Genet. *12*, 846–860.
29. Slyusarev, G.S., and Manylov, O.M. (2001). General morphology of the muscle system in the female orthonectid *Intoshia variabili* (Orthonectida). Cah. Biol. Mar. *42*, 239–242.
30. Slyusarev, G.S. (2003). The fine structure of the muscle system in the female of the orthonectid *Intoshia variabili* (Orthonectida). Acta Zoologica *84*, 107–111.
31. Moran, Y., Barzilai, M.G., Liebeskind, B.J., and Zakon, H.H. (2015). Evolution of voltage-gated ion channels at the emergence of Metazoa. J. Exp. Biol. *218*, 515–525.
32. Moroz, L.L., Kocot, K.M., Citarella, M.R., Dosung, S., Norekian, T.P., Povolotskaya, I.S., Grigorenko, A.P., Dailey, C., Berezikov, E., Buckley, K.M., et al. (2014). The ctenophore genome and the evolutionary origins of neural systems. Nature *510*, 109–114.
33. Steinmetz, P.R., Kraus, J.E., Larroux, C., Hammel, J.U., Amon-Hassenzahl, A., Houliston, E., Wörheide, G., Nickel, M., Degnan, B.M., and Technau, U. (2012). Independent evolution of striated muscles in cnidarians and bilaterians. Nature *487*, 231–234.
34. Liu, J., Ward, A., Gao, J., Dong, Y., Nishio, N., Inada, H., Kang, L., Yu, Y., Ma, D., Xu, T., et al. (2010). *C. elegans* phototransduction requires a G protein-dependent cGMP pathway and a taste receptor homolog. Nat. Neurosci. *13*, 715–722.

Current Biology, Volume *26*

Supplemental Information

The Genome of *Intoshia linei* Affirms Orthonectids as Highly Simplified Spiralians

Kirill V. Mikhailov, Georgy S. Slyusarev, Mikhail A. Nikitin, Maria D. Logacheva, Aleksey A. Penin, Vladimir V. Aleoshin, and Yuri V. Panchin

A

assembly size (Mbp)	43.2
assembly N50 (Kbp)	25.1
GC %	26.0
protein-coding region content %	23.1
repetitive content %	27.7
predicted protein-coding genes	8,728
gene density (genes per Mbp)	202
intron density (introns per gene)	5.07
mean intron length (bp)	342
mean exon length (bp)	187
mean length of intergenic region (bp)	1,390

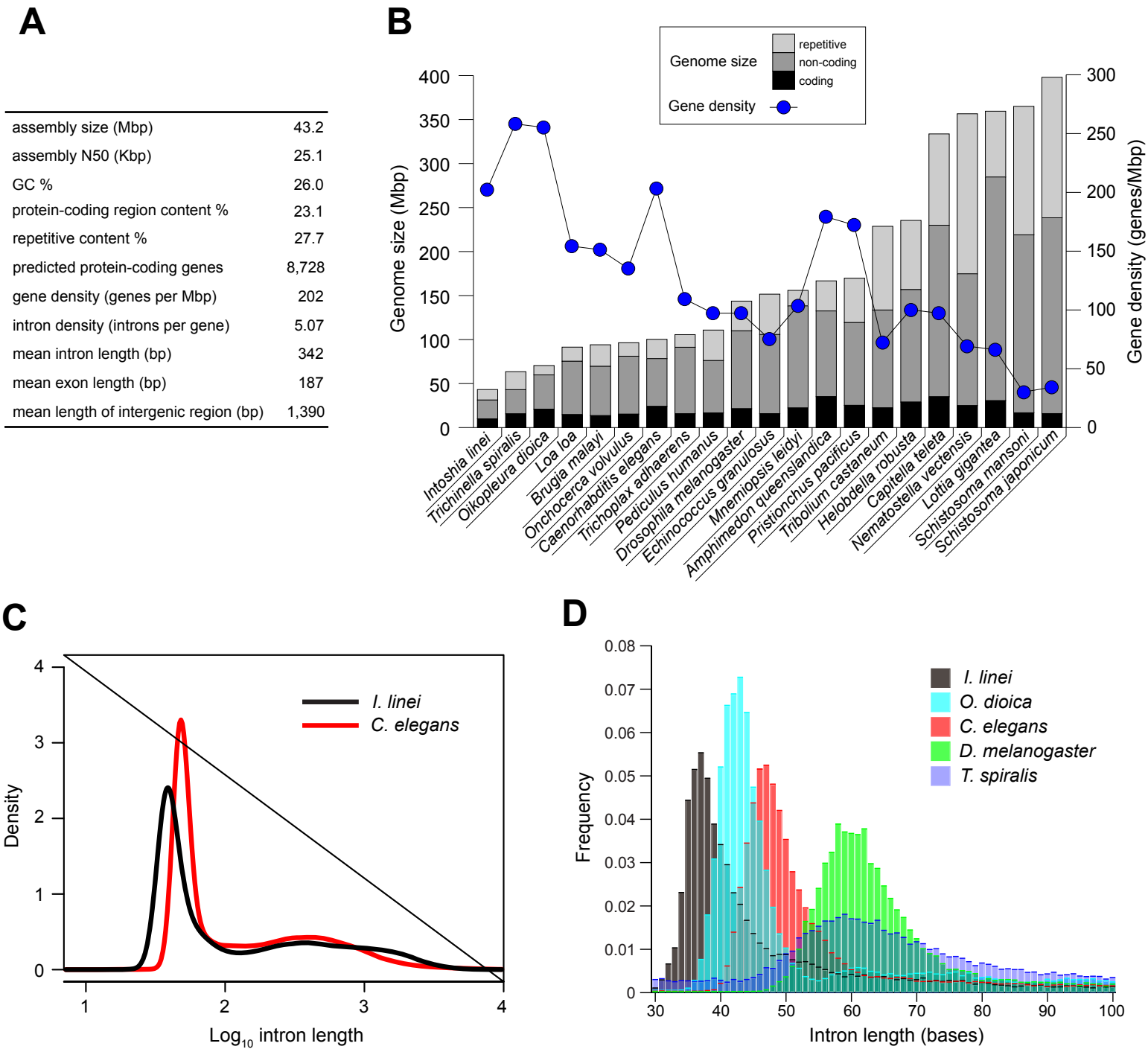
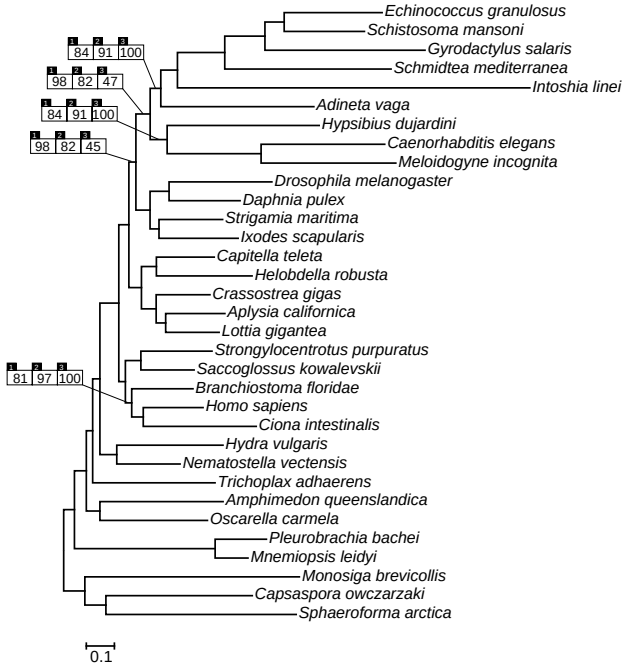


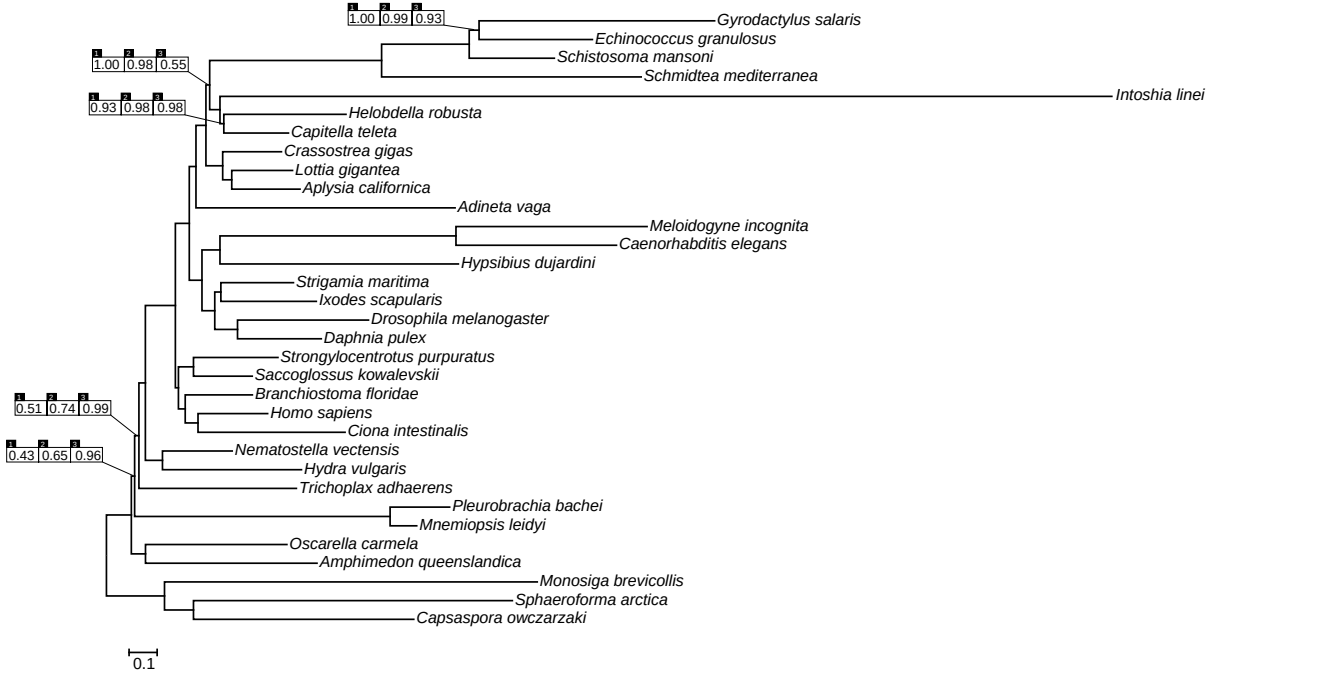
Figure S1 (Related to Figure 2). Annotation features of the *Intoshia linei* genome. **A**, genome assembly and annotation statistics. **B**, genome sizes and gene densities for selected metazoan genomes. Genome data for *H. robusta*, *C. teleta*, *L. gigantea*, *S. mansoni*, *S. japonicum* and *E. granulosus* were obtained from previous publications [S46, S47]. The data for *O. dioica* were obtained from the Genoscope *Oikopleura dioica* annotation v1.0; annotation files for other organisms were taken from release 26 of the Ensembl Genomes. The coding portion of the genome represents the cumulative size of protein-coding exon regions; the repetitive portion represents the sum of repeat-masked bases. **C**, intron size distribution in *I. linei* and *C. elegans*. The intron sizes are given in a logarithmic scale. The distributions in both genomes display a pronounced peak of short (<100 bp) introns. **D**, histogram of intron sizes depicting the intron frequencies in the 30-100 bp size range for 5 metazoan genomes, including those with the highest gene density. The mode of the distribution for *I. linei* is at 37 bp.

RAxML GTR+Γ4

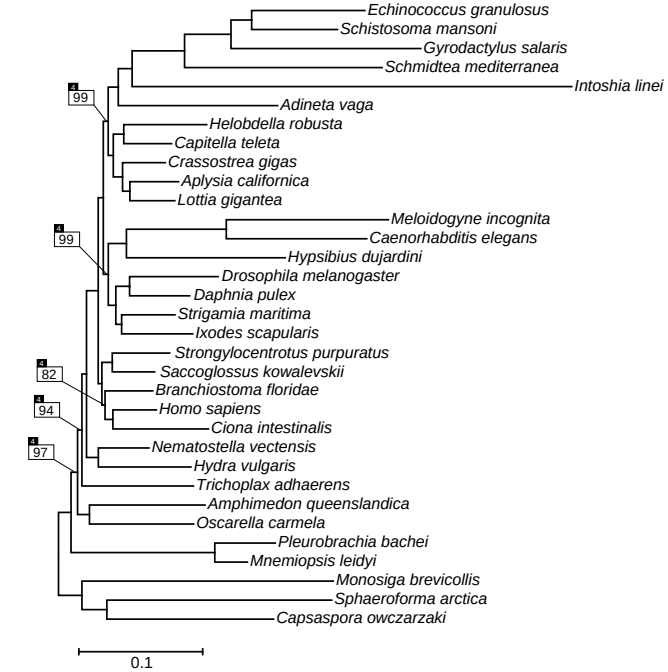
PhyloBayes CAT+Γ4

3 2 1 full dataset; – category 8; – categories 7,8





4 – categories 6,7,8



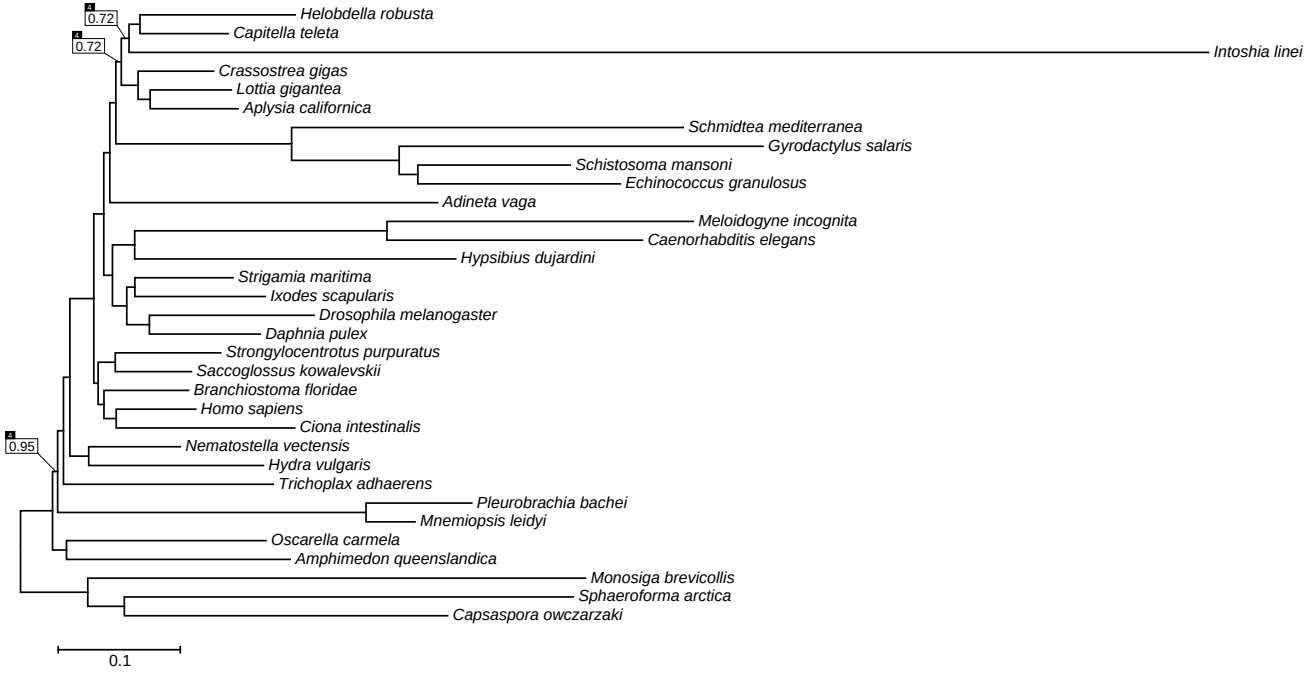
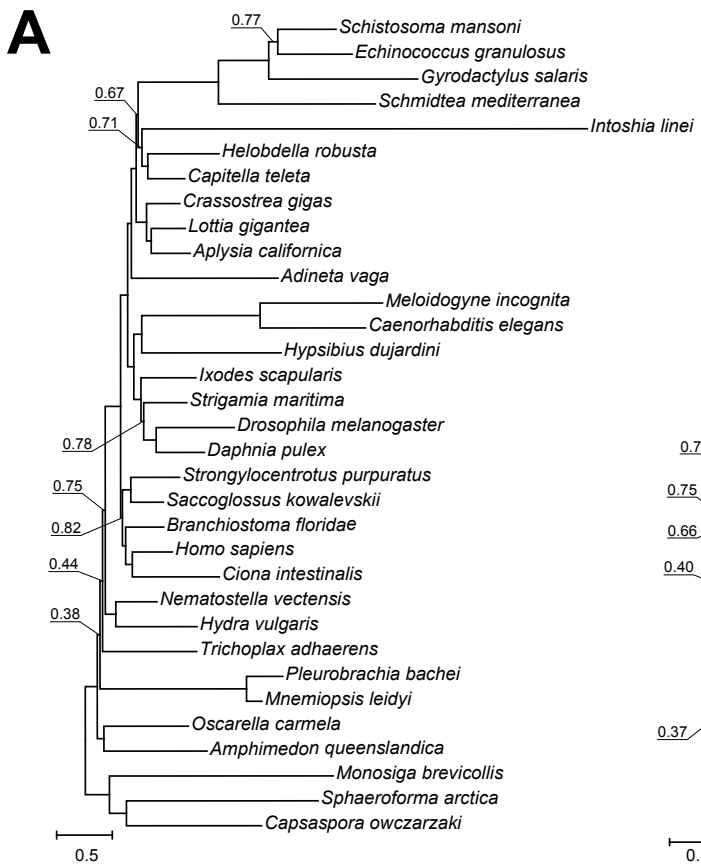
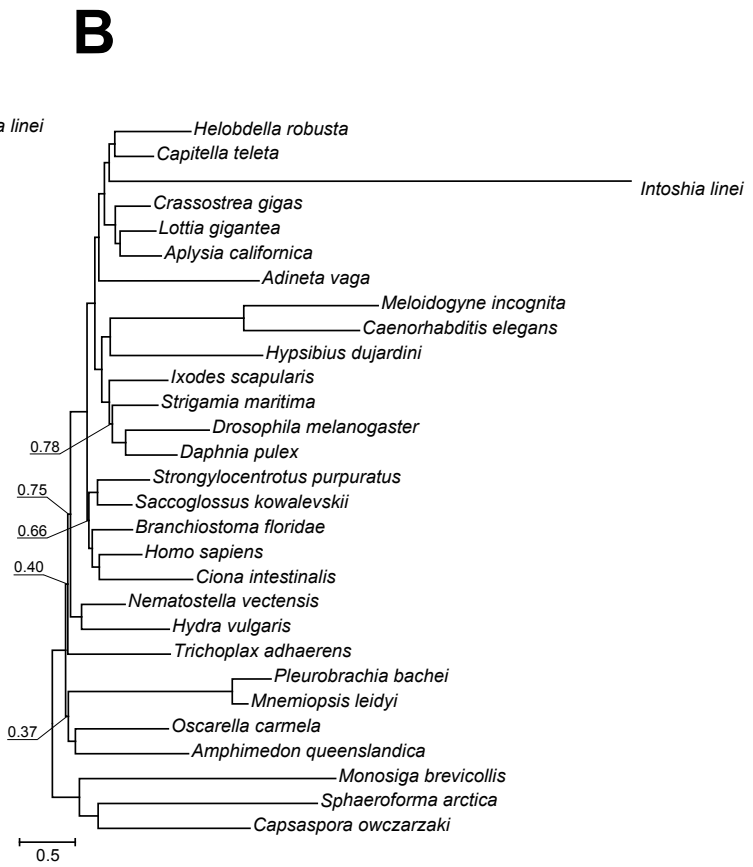


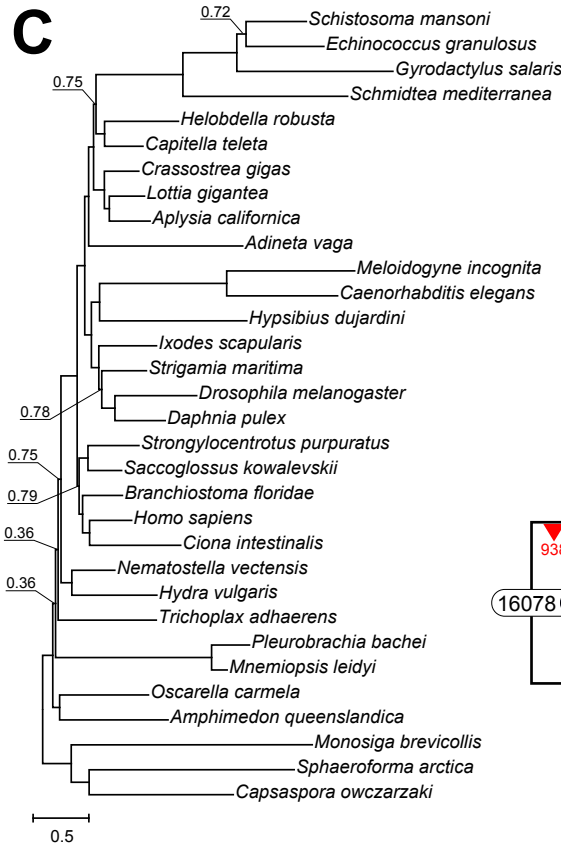
Figure S2 (Related to Figure 1). Phylogenetic trees reconstructed by RAxML and PhyloBayes with a 500-gene and derivative datasets, obtained by progressive elimination of fast-evolving sites. The trees reconstructed with the full dataset and datasets without sites of rate category 8 or categories 7 and 8 display no change in the topology and are represented by the varying support values superimposed on a tree reconstructed with the full dataset. The support indexes are given under a label specifying the corresponding dataset: (1) full dataset; (2) dataset without category 8 sites; (3) dataset without sites of categories 7 and 8; (4) dataset without sites of categories 6, 7 and 8. Support indexes for nodes with 100 bootstrap support or 1.00 posterior probability in all analyses are omitted. The PhyloBayes chain convergence statistics (largest difference across bipartitions, log likelihood discrepancy, log likelihood effective size) for the four datasets are: (1) 1.0, 0.33, 2115; (2) 1.0, 0.75, 100; (3) 1.0, 0.48, 500; (4) 0.9, 0.4, 133.

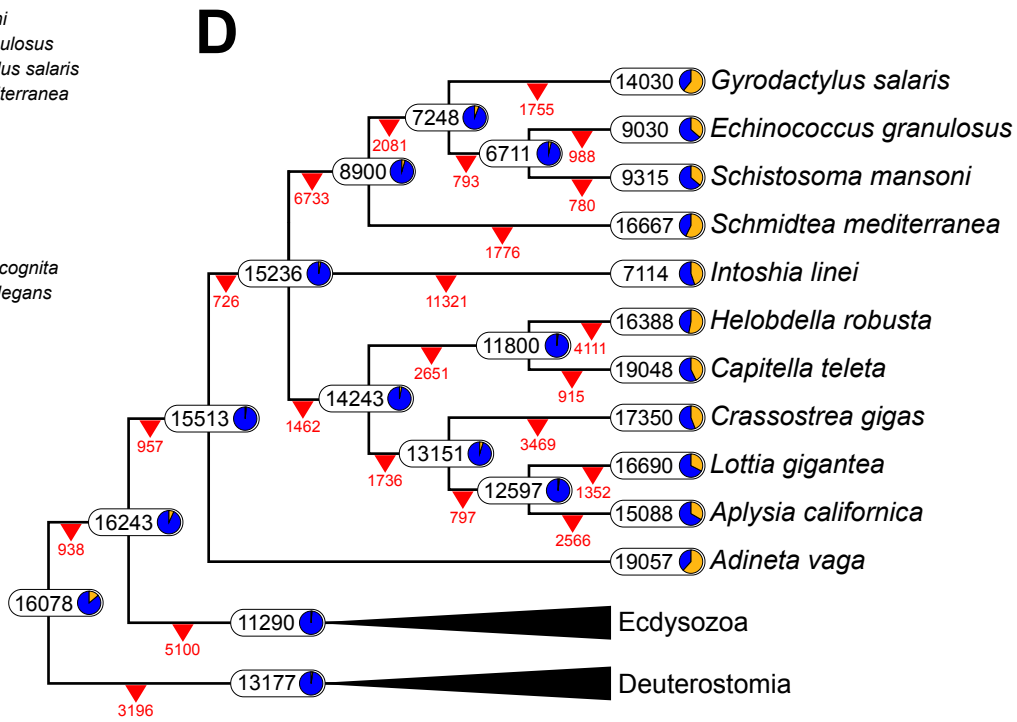




jackknife replicate	1	2	3	4	5	6	7	8	9	10
max. discrepancy across bipartitions	0.114	0.550	0.222	0.164	0.306	0.446	0.262	0.050	0.230	0.252
mean discrepancy across bipartitions	0.004	0.012	0.007	0.004	0.008	0.011	0.010	0.002	0.008	0.006
logLik discrepancy	0.780	0.725	0.045	1.097	0.520	0.900	0.261	0.521	1.255	0.265
logLik effective size	28	20	45	23	150	48	30	21	33	31

	1	2	3	4	5	6	7	8	9	10
	0.154	0.030	0.206	0.352	0.192	0.110	0.042	0.454	0.562	0.310
	0.006	0.001	0.007	0.007	0.009	0.002	0.001	0.013	0.013	0.012
	0.063	0.212	0.032	0.399	0.118	0.062	0.534	1.239	0.423	0.670
	20	18	21	39	22	150	21	16	10	50





jackknife replicate	1	2	3	4	5	6	7	8	9	10
max. discrepancy across bipartitions	0.262	0.040	0.390	0.018	0.188	0.054	0.838	0.042	0.540	0.320
mean discrepancy across bipartitions	0.008	0.001	0.012	0.000	0.004	0.002	0.032	0.002	0.014	0.012
logLik discrepancy	0.054	0.702	0.701	0.046	0.025	0.204	1.531	0.104	0.598	0.066
logLik effective size	16	15	9	74	69	34	18	36	29	45

Figure S3 (Related to Figure 1). Consensus trees of jackknife replicates inferred by PhyloBayes under the CAT-GTR model for the three datasets: **A**, full set of sequences; **B**, without sequences of flatworms; **C**, without sequences of *Intoshia linei*. The support values represent posterior probabilities of bipartitions evaluated by pooling chains from all jackknife replicates. Support indexes for nodes with over 0.95 jackknife-resampled posterior probability are omitted. A brief summary of PhyloBayes chain convergence statistics for each jackknife replicate is given below the trees for the three datasets; **D,** Gain and loss of orthologous groups in the evolution of Bilateria with focus on the spiralian taxa. The evolutionary history of OrthoMCL-generated groups of orthologs is superimposed on the prospective metazoan phylogeny using the Dollo parsimony principle. The uncertainty in the phylogenetic placement of *I. linei* is represented in the tree with a tripartition between the clade of Platyhelminthes, the Trochozoa (Annelida+Mollusca) and the orthonectid. The numbers at the nodes are the inferred counts of orthologous groups, and the pie charts represent the proportions of ancestral (blue portion) and novel (orange portion) groups at the node. The inferred number of losses at each branch is given in red (marked with a down-pointing triangle).

	I. linei	E. granulosus	S. mansoni	C. elegans	D. melanogaster	B. floridae	H. sapiens
ANTP-HOXL	**6**	**6**	**12**	**9**	**18**	**24**	**52**
Hox: Hox1			1	1	1	1	3
Hox: Hox2	1		1		1	1	2
Hox: Hox3					3	1	3
Hox: Hox4	1	1	1	1	1	1	4
Hox: Hox5					1	1	3
Hox: Hox6-8	1	1	2	2	4	3	8
Hox: Hox9-13(15)		2	2	2	1	7	16
ParaHox: Gsx					1	1	2
ParaHox: Pdx						1	1
ParaHox: Cdx	1			1	1	1	3
Evx	1	1	1	1	1	2	2
Gbx			1		1	1	2
Meox		1	1		1	1	2
Mnx	1		1	1	1	2	1
ANTP-NKL	**11**	**16**	**17**	**21**	**29**	**36**	**48**
Abox				1	1	1	
Barhl	1	2	2	2	2	1	2
Bari			1		1	1	
Barx						1	2
Bsx		1	1	1	1	1	1
Dbx	1				1	1	2
Dlx	1		1	1	1	1	6
Emx		1	1	2	2	3	2
En	1		1	1	2	1	2
Hhex				1	1	1	1
Hlx					1	1	1
Lbx	1	1	1		2	1	2
Msx	1		1	1	1	1	2
Msxlx		1	1		1	1	
Nedx					1	2	
Nk1	1	2	1	1	1	2	2
Nk2.1	1	1	1	2	1	1	2
Nk2.2		3	1	1	1	1	2
Nk3		1	1		1	1	2
Nk4		1	1	1	1	1	3
Nk5/Hmx	1	1	1	1	1	1	3
Nk6	1	1	1	1	1	1	3
Nk7				1	1	1	
Noto					1	1	1
Ro				1	1	1	
Tlx					1	1	3
Vax				2		1	2
Ventx						2	1
Unassigned	1		1				
POU	**2**	**5**	**5**	**4**	**5**	**7**	**16**
Hdx						1	1
Pou1						1	1
Pou2		1	1	2	2	1	3
Pou3	1	1	1	1	1	2	4
Pou4		1	1	1	1	1	3
Pou6	1	1	1		1	1	2
Unassigned		1	1				
SINE	**3**	**3**	**4**	**4**	**3**	**3**	**6**
Six1/2	1	1	1	2	1	1	2
Six3/6	1	1	1	1	1	1	2
Six4/5	1		1	1	1	1	2
Unassigned		1	1				

	I. linei	E. granulosus	S. mansoni	C. elegans	D. melanogaster	B. floridae	H. sapiens
PRD	**19**	**9**	**13**	**17**	**28**	**29**	**73**
Alx						1	3
Arx				1	2	1	1
Dmbx						1	1
Drgx					1	1	1
Gsc				1	1	1	2
Hbn			1		1		
Hopx						1	1
Isx						1	1
Otp	1	1	1		1	1	1
Otx	4	3	1	3	1	1	3
Pax2/5/8					1	1	3
Pax3/7			1	3	3	1	2
Pax4/6	1	1	2	1	4	1	2
Phox			1	2	1	1	2
Pitx	1	1	1	1	1	1	3
Prop	1	1	1	1	1	1	1
Prrx					1	1	2
Rax			1	1	1	1	2
Repo		1	1		1	1	
Shox					1	1	2
Uncx	1	1	2	1	2	3	1
Vsx	1			2	3	1	2
Unassigned	9						
LIM	**8**	**5**	**8**	**7**	**6**	**7**	**12**
Isl	1	1	1	1		1	2
Lhx1/5	1		1	2	1	1	2
Lhx2/9	1	1	2	1	1	2	2
Lhx3/4	1	1	1	1	1	1	2
Lhx6/8	1	1	1	1	1	1	2
Lmx	1	1	2	1	2	1	2
Unassigned	2						
HNF	**0**	**0**	**0**	**1**	**0**	**4**	**3**
Hmbox				1		2	1
Hnf1						1	2
TALE	**10**	**11**	**9**	**5**	**8**	**9**	**20**
Irx	1	1	2	1	3	3	6
Meis	1	3	2	1	1	1	3
Mkx		1			1	1	1
Pbx	1	1	1	3	1	1	4
Pknox		1	1			1	2
Tgif					2	1	4
Unassigned	7	4	3				
CUT	**1**	**3**	**2**	**6**	**3**	**4**	**7**
Cmp				1	1	1	
Cux		1		1	1	1	2
Onecut	1	2	2	4	1	1	3
PROS	**1**	**2**	**2**	**1**	**1**	**1**	**2**
Prox	1	2	2	1	1	1	2
ZF	**1**	**1**	**2**	**2**	**2**	**5**	**14**
Tshz						1	3
Zeb		1		1		1	2
Zfhx	1		2	1	2	1	3
Zhx/Homez						1	4
CERS	**0**	**1**	**1**	**0**	**1**	**1**	**5**
Cers		1	1		1	1	5
Other	**0**	**1**	**1**	**15**	**0**	**3**	**4**

Figure S4 (Related to Table 1). Homeobox genes in *Intoshia linei* and model bilaterians. The classification of homeobox gene classes and families follows the Homeobox Database [S32]. The data on the homeobox gene complement for the genomes of *Homo sapiens*, *Branchiostoma floridae*, *Drosophila melanogaster* and *Caenorhabditis elegans* were taken directly from the Homeobox Database. The data for *Schistosoma mansoni* and *Echinococcus granulosus* were obtained from a previous publication [S48]. Gene families that are unique to *H. sapiens* or *B. floridae* or *D. melanogaster* are excluded from the listing, but are accounted for in the total gene count for each class. The homeobox pseudogenes are omitted from the count.

Table S1 (Related to Figure 2). Repetitive content in the genome of *Intoshia linei*

Repetitive element class	count	bases	% genome
Retroelements	715	235,890	0.55
LTR/Gypsy	599	214,974	0.50
LINE/CR1	38	11,178	0.03
LTR/Pao	78	9,738	0.02
DNA transposons	14,522	2,995,690	6.94
hAT	9,670	1,886,890	4.37
TcMar	3,830	876,785	2.03
CMC-Chapaev-3	723	137,897	0.32
PiggyBac	76	21,676	0.05
Unclassified	35,539	7,610,437	17.62
Simple repeats	17,883	854,483	1.98
Low complexity	5,458	275,573	0.64
Total		11,972,073	27.72

Table S2 (Related to Figure 2 and Figure 3). The number of putative neurotransmitter receptors detected in *Intoshia linei*

	Metabotropic	Ionotropic
Acetylcholine	1	12
Glycine	0	2
Serotonin	2	0
Glutamate	5	0
Histamine	2	0
Dopamine	2	0
Octopamine	1	0
Adrenergic	1	0
Neuropeptide	16	0

Supplemental Experimental Procedures

Biological material collection

Individuals of the nemertean host *Lineus ruber* (Muller,1774) (Nemertea Enopla Heteronemertea) for the orthonectid *Intoshia linei* (Giard, 1877) were collected in 2001 and 2013 near the marine biological station Dalnie Zelentsi, Barents Sea (69°07' N, 36°05' E). The nemerteans were collected during a low tide and observed for infection under a stereomicroscope. The infected individuals were maintained in filtered sea water to allow the release of the aquatic stage of *I. linei*. For genomic sequencing the collected samples of *I. linei* were fixed using ethanol. Genomic DNA was extracted using the NucleoSpin Tissue kit (MACHEREY-NAGEL), and the DNA quantity assessments were performed with Qubit fluorometric quantification (Life Technologies). For transcriptomic sequencing the collected specimens of *I. linei* were stored in RNA*later* solution (Life Technologies), the nucleic acids were extracted using the TRIzol reagent and treated with a DNAase.

Genome sequencing and assembly

The genomic DNA libraries of *Intoshia linei* were constructed from two samples of reproductive stage organisms. The libraries were constructed using TruSeq library preparation protocol (Illumina) and sequenced on a HiSeq 2000 instrument. The libraries have an estimated insert length of 238 bp (67 bp standard deviation) for the first sample and 270 bp (100 bp standard deviation) for the second sample. We obtained 36M 100-bp paired-end reads for the first library and 13M 100-bp paired-end reads for the second library. The k-mer frequency analysis of the sequenced libraries, which was done using the Jellyfish 2.1.3 [S1] program, showed a noticeable difference between their k-mer frequency distributions, with the 270 bp insert library having a sharper peak of true k-mers and a higher count of low coverage k-mers. To assess how the difference between the libraries influenced the assembly quality we carried out assemblies using several combinations of reads from the two libraries. The assemblies were performed by the Velvet 1.2.10 [S2] assembler with k-mer lengths of 55 and 77 using: 1) all reads from both libraries, 2) equal amount of reads from both libraries, 3) only reads from the first library, 5) only reads from the second library. Prior to assembly reads were adapter trimmed with Trimmomatic v0.30 [S3] in Paired End mode, requiring a minimal length of 55 bp for each read to keep the pair; the assembly quality was evaluated by QUAST 2.3 [S4]. We found that despite having higher coverage, the assemblies which utilized reads from the 238 bp insert library had significantly poorer quality, presumably due to the greater age of the fixed sample used for preparation of this library. Therefore, we opted to use the assembly produced exclusively with the 270 bp insert library for all downstream analyses. The selected genome assembly was done by Velvet with k-mer length of 55, and has an average per nucleotide coverage of 35. The gaps in the assembly were closed iteratively by GapCloser v1.12 of the SOAPdenovo package [S5] and GapFiller v1.10 [S6]. Even though the samples of *I. linei* were collected upon non-traumatic release from the nemertean host [S7], any possible host contamination of the assembled genomic sequences was additionally ruled out by setting the appropriate k-mer coverage cutoff during the assembly (half of the median coverage depth), ensuring that no contigs with coverage below the k-mer frequency distribution corresponding to *I. linei* are produced by the assembler. Any haphazard contamination from prokaryotic or other sources was removed by performing a BLAST [S8] search against the 238 bp insert library assembly (obtained from an independent sample of *I. linei*) and flagging all sequences that fail to produce hits with over 95% identity. 2.1 Mbp of sequence in 4,359 contigs was discarded from the assembly as mostly prokaryotic contamination. The size of the cleaned assembly is 43.2 Mbp in 11,908 contigs with an N50 of 25 Kbp.

Transcriptome sequencing and assembly

The RNA-seq library was prepared using TruSeq library preparation protocol (Illumina) from a DNAase-treated nucleic acid extract of *Intoshia linei* reproductive stage organisms. The library was sequenced on a HiSeq2000 instrument, producing 16M 100-bp paired-end reads. The reads were adapter trimmed with Trimmomatic v0.30 [S3] and assembled into transcripts using the Cufflinks 2.2.1 pipeline [S9-S11] with the assembled genome as reference. To avoid excessive merging of transcripts during assembly the Cufflinks overlap-radius setting was lowered to 1. Using the Cufflinks pipeline 76.9% of reads from the sequenced RNA-seq library were mapped to the assembly, generating a total of 12,400 transcripts from 9,900 genomic loci.

Genome annotation

Prior to gene prediction the assembly was repeat masked using a custom library of *I. linei* specific repetitive elements constructed with RepeatModeler 1.0.8 [S12-S15]. Repetitive elements were classified using the 2014-01-31 repeatmasker version of the Repbase [S16]. The custom library of *Intoshia linei* repetitive elements created by RepeatModeler was used for masking repeats with RepeatMasker open-4.0.5 [S17]. A total of 27.7 % of the assembly was masked by RepeatMasker, with the most frequent repeat family accounting for 9.3% of the assembly. Gene predictions were carried out using the MAKER 2.31.6 pipeline [S18] with two programs for *ab initio* gene prediction: Augustus 3.0.3 [S19] and the self-training GeneMark-ES 2.3c [S20]. Cufflinks-generated transcripts were provided to MAKER as EST evidence and the Swiss-Prot database [S21] as homology based evidence. *I. linei* specific model for Augustus was constructed iteratively, first by using a set of predictions inferred with CEGMA 2.5 [S22], then by selecting the best gene models produced in the preliminary run of MAKER. Contigs shorter than 500 bp (3.7% of the assembly) were excluded from the annotation procedure. Annotation of the repeat-masked assembly returned 8,728 protein coding genes, which includes both evidence supported and unsupported predictions. This version of annotation was observed to have a number of artificial gene fusions, which in sum underestimates the actual gene count. Therefore, it represents an estimated lower bound on the number of genes encoded by the genome of *I. linei*, and most likely will be corrected upward with higher assembly quality and fuller RNA-seq data.

Protein orthology clustering

For comparative analyses the predicted proteins of *Intoshia linei* were clustered using the OrthoMCL v2.0.9 pipeline [S23, S24] with proteins from completely sequenced genomes of 29 metazoans and 3 unicellular relatives of Metazoa. The proteomes were obtained from the following resources: NCBI's GenBank database (http://www.ncbi.nlm.nih.gov/genbank), reference proteomes of the UniProtKB database (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes), JGI Genome Portal (http://genome.jgi.doe.gov/), Broad Institute genome annotation projects (http://www.broadinstitute.org/scientific-community/data), WormBase ParaSite online resource (http://parasite.wormbase.org/), Genoscope *Adineta vaga* Genome Browser (http://www.genoscope.cns.fr/adineta/), *Hypsibius dujardini* genome hosting site (http://badger.bio.ed.ac.uk/H_dujardini/), The *Gyrodactylus salaris* Genome Project site (http://invitro.titan.uio.no/gyrodactylus/), *Meloidogyne* genomic resources site (http://www6.inra.fr/meloidogyne_incognita), NHGRI *Mnemiopsis* Genome Project Portal (http://research.nhgri.nih.gov/mnemiopsis/), the comparative genomics platform for early branching metazoan animals (http://www.compagen.org/datasets.html), the comparative neurogenomics database (http://neurobase.rc.ufl.edu/). BLAST [S25] similarity search for OrthoMCL was carried out with an e-value cutoff of 1E-05, and the Markov clustering was performed with an inflation parameter of 1.5. The clustering returned 25,208 orthologous groups, where each group is represented by orthologs from at least 2 genomes. The predicted *I. linei* proteins were clustered into 7,114 orthologous groups, 3,915 of those groups are represented by at least one other genome and contain a total of 4,538 *I. linei* proteins or 52% of its whole proteome.

Comparative genomic analyses

The genes implicated in transcriptional regulation were detected using Pfam 27.0 [S26] and InterPro 48.0 [S27] database searches performed with HMMER 3.1b1 [S28] and InterProScan 5.6 [S29] programs. The domain architectures for specific families of multidomain transcription factors and protein ligands were verified using the NCBI's conserved domain database [S30]. The classification of *Intoshia linei* homeobox genes was done with the assistance of the CLANS clustering and visualization program [S31] using the homeodomain sequences of model bilaterian organisms, and verified using the BLAST search functionality of the Homeobox Database [S32]. For pathway assignments we used the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways collection [S33]. The KEGG orthology annotations and pathway mappings for the genomes of *I. linei* and other bilaterians were generated by the KEGG Automatic Annotation Server [S34] using best bi-directional hit method with a default bit score cutoff of 60. A total of 2,911 *I. linei* proteins were assigned a KEGG orthology. The dynamics of gene family gain and loss during the bilaterian evolution were modelled using the presence/absence profiles of the OrthoMCL-generated orthology groups for the selected set of metazoan genomes. The analysis was performed using the Dollo parsimony method implemented by the Count software [S35].

Phylogenetic analyses

The 500-gene dataset for phylogenetic analyses was constructed from a set of OrthoMCL-generated orthologous groups by requiring a group to be present in at least 31 sampled genomes (out of 33 total) and have exactly one ortholog in at least 27 of them. The alignments were created by MAFFT v7.215 [S36] and trimmed by trimAl 1.2rev57 [S37] with a gap threshold of 0.9 and a similarity threshold of 0.001 over a window of size 6 (-w 3). Trimmed alignments were inspected manually using BioEdit 7.2.5 [S38] to remove spurious sequences. The concatenated alignment was constructed by SCaFoS 1.25 [S39] using TREE-PUZZLE 5.2 [S40] evolutionary distance calculation for sequence selection per operational taxonomic unit. The concatenation resulted in a 108,321 position data matrix with 6.65% missing data; the number of constant positions in the alignment is 18,271 or 16.9% of the alignment size. Two tree inference methods were employed for the phylogenetic analyses of the concatenated alignment: maximum likelihood search implemented by RAxML 8.0.0 [S41] and Bayesian inference method of PhyloBayes MPI version 1.5a [S42]. The RAxML trees were reconstructed under a general time-reversible (GTR) model with Gamma-distributed rate variation across sites; node support for the ML trees was evaluated with 100 bootstrap replicates utilizing rapid search procedure (-f a). The PhyloBayes tree inference was performed in two variants: one using a site-specific profile mixture model (CAT) [S43] combined with uniform global substitution rates (-poisson), and the other using the CAT-GTR model, which combines site-specific profiles with a global empirically-estimated GTR substitution matrix. Gamma-distributed rate variation across sites with 4 categories was used for all PhyloBayes analyses. For each inference the tree was generated from 4 independent chains, which were run for 10,000 cycles each, with a burn-in of 2,000 cycles. Prior to PhyloBayes analyses all constant positions were removed from the alignment.

The trees reconstructed for the full dataset using RAxML and PhyloBayes are conflicting in regard to placement of *Intoshia linei* and the general topology of the Metazoan tree (Figure S2). The confounding influence of divergent sequences of *I. linei*, flatworms and nematodes, resulting in a long-branch attraction artefact, is evident in the RAxML tree obtained for the full dataset. The tree reconstructed by PhyloBayes using the site-heterogeneous substitution model (CAT) displays better reconciliation with the accepted Protostome phylogeny, recovering monophyletic Ecdysozoa. The position of *I. linei* in the PhyloBayes inference with the CAT model suggests orthonectid affiliation with the annelids. Problematically, the PhyloBayes tree reconstructed with the full dataset also places flatworms next to the annelid branch, disrupting the classical Trochozoa group, which unites annelids and mollusks.

With the reliability of obtained phylogenies in question, we sought to reduce the impact of homoplasy-prone fast-evolving sites by performing reconstructions on datasets subjected to progressive elimination of such sites. The estimation of evolutionary rates in the concatenated alignment was performed by TREE-PUZZLE under the LG substitution model with 8 categories of Gamma-distributed rates. The evolutionary rates of sites were evaluated independently under the constraints of two topologies obtained by RAxML and PhyloBayes on the full dataset, and the minimal of the two category values was assigned to each site. The fast-evolving sites were removed in three steps in accordance with their rate category assignment, generating three datasets: 1) alignment without category 8 sites (91% of the original alignment size or 89% of the alignment without constant sites); 2) alignment without sites of categories 7 and 8 (76% of the original alignment size or 73% of the alignment without constant sites); 3) alignment without sites of categories 6, 7 and 8 (61% of the original alignment size or 53% of the alignment without constant sites). The three derivative alignments were analysed by RAxML and PhyloBayes identically to the full alignment. Following the removal of fast-evolving sites, the RAxML and PhyloBayes trees showed more congruent topologies, with the exception for the positions of *Intoshia linei*, rotifer *A. vaga* and ctenophores (Figure S2). In particular, the contradiction between the placement of *I. linei* next to the flatworms in the RAxML analyses and its placement as a sister-group of annelids in the PhyloBayes analyses remained unresolved.

In addition to the fast-evolving site removal procedure, we examined the position of *Intoshia linei* using the more complex and computationally challenging CAT-GTR model of PhyloBayes. To alleviate the computational cost of running the CAT-GTR model for the full alignment, we applied a jackknife resampling procedure, generating 10 jackknife replicates of size 20,000 from the full alignment. Each replicate was analysed by PhyloBayes in two independent chains under the CAT-GTR model with 4 categories of Gamma-distributed rate variation. The chains were run for 10,000 cycles, following which a majority-rule consensus tree was obtained from all replicates sampled with a 50% burn-in. Chains from every replicate were summarized by the bpcomp program of PhyloBayes. To assess how the divergent sequences of *I. linei* and flatworms behaved independently of each other on the tree we repeated the computation with a variant of the dataset which excluded flatworms and a variant which excluded *I. linei*. These additional datasets were obtained from the original

10 jackknife replicates and were analyzed under the same conditions. The jackknife consensus tree largely mirrors the result obtained using the CAT model, placing *I. linei* with the annelids (Figure S3). Surprisingly, the analysis also recovers association of flatworms with annelids even when the divergent sequences of *I. linei* are excluded from the dataset. This result suggests that the association with the annelid lineage is not exclusive to the orthonectid, and may in fact be an artefact shared by the long branch taxa in the dataset (*I. linei* and flatworms).

The CAT-GTR analysis of the slow-evolving site dataset was performed by PhyloBayes in full without jackknife resampling. The analysis generates the least conflicting MCMC runs (largest difference across bipartitions = 0.17; log likelihood discrepancy = 0.04, log likelihood effective size = 1341), which converge on the topology respecting the monophyletic Trochozoa group and placing *Intoshia linei* sister to the flatworms (Figure 1B). To evaluate which of the two models, CAT-GTR or CAT, fit the dataset better we used a 5-fold cross-validation analysis with 10 replicates of the slow-evolving site dataset. Each replicate was run for 10,000 cycles under the two models and sampled with a 50% burn-in. The cross-validation score of the CAT-GTR model against the CAT model (2146.22 ± 71.3236) confirmed higher fit of the CAT-GTR model for our dataset.

For the analysis with the extended taxonomic sampling of the spiralian taxa we used a dataset assembled by Struck *et al.* [S44]. The dataset of 559 masked alignments prescreened for paralogy and contamination was used to extract the unmasked sequences from the initial dataset. To minimize the possibility of gathering paralogous sequences of *Intoshia linei* we established correspondence between the orthologous groups in the Struck *et al.* dataset and our OrthoMCL generated groups using the sequences of *Capitella teleta*, which were required to display identity between the two datasets. In total, 469 orthologous groups from the Struck *et al*. dataset were selected using this method, and the corresponding sequences of OrthoMCL-inferred *I. linei* orthologs were added to the dataset. The dataset was aligned, trimmed and concatenated similarly to the 500-gene dataset. Constant positions and positions with over 60% missing data were excluded from the concatenated alignment, which resulted in a 22,909 position data matrix. Sequences of *Dactylopodella baltica*, *Echinococcus multilocularis*, *Echinorhynchus truttae*, *Lepidodermella squamata*, and *Nematoplana coelogynoporoides* were excluded from the dataset due to being poorly represented or phylogenetically unstable. The number of operational taxonomic units in the final alignment is 61, and the proportion of missing data is 47%. The resulting PhyloBayes tree, which was constructed using the CAT-GTR model with 4 categories of Gamma-distributed rates, places *I. linei* at the base of the classical Lophotrochozoa [S45] (Figure 1C). The support for this association is low and does not lend itself to a definite conclusion. The chain convergence statistics in the analysis are poor (largest difference across bipartitions = 1.0; log likelihood discrepancy = 0.72, log likelihood effective size = 52), i.e. chains fail to converge, however, the independent position of the orthonectid is recovered in each of the four PhyloBayes chains, and the instability at the node is caused primarily by the shifting position of the Entoprocta/Cycliophora branch, which frequently finds itself sister to the flatworms in the analysis.

Peroxisome

Analyzing conserved genes lost in *Intoshia linei* we found that proteins and PFAM domains specific to peroxisomes organelles found in virtually all eukaryotic cells are absent in *I. linei* database. For instance the peroxisomal proteins: PEX3, PEX10, PEX12, and PEX19 mandatory for peroxisome are apparently missing. Failure to identify these markers unequivocally means absence of the organelle. Eight PFAM domains linked to peroxisome are present in GO database (PF01756, PF04088, PF04614, PF04882, PF05648, PF07163, PF09262, PF12634 http://geneontology.org/external2go/pfam2go). None of these domains were detected in *I. linei*. Further on we collected a set of 304 unique PFAM domains present in peroxisome related proteins from different species downloaded from http://www.peroxisomedb.org/ and compared them to the list PFAM domains detected in proteins predicted for *I. linei* . 124 of such domains were absent in *I. linei* .(compare to 36 PFAM domains absent in *Capitella teleta*, 60 in *Drosophila melanogaster*, 51 in *Caenorhabditis elegans* and 41 in *Homo sapiens*). This search confirms the hypothesis that peroxisome was lost in *I. linei* and suggests that the absence of 124 unique PFAM domains is associated with the loss of this organelle.

Supplemental References

S1. Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764-770.

S2. Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res *18*, 821-829.

S3. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

S4. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072-1075.

S5. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience *1*, 18.

S6. Boetzer, M., and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. Genome Biol *13*, R56.

S7. Slyusarev, G.S., and Cherkasov, A.S. (2001). Analisis of possible mechanisms of emission of the orthonectids from their hosts. Parazitologiya (St. Petersburg) *35*, 338-343.

S8. Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J Comput Biol *7*, 203-214.

S9. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology *28*, 511-515.

S10. Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

S11. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

S12. Smit, A.F.A., Hubley, R. (2008-2015). RepeatModeler Open-1.0. <http://www.repeatmasker.org>

S13. Bao, Z., and Eddy, S.R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res *12*, 1269-1276.

S14. Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21 Suppl 1*, i351-358.

S15. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res *27*, 573-580.

S16. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research *110*, 462-467.

S17. Smit, A.F.A., Hubley, R., Green, P. (2013-2015). RepeatMasker Open-4.0. <http://www.repeatmasker.org>

S18. Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics *12*, 491.

S19. Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics *19 Suppl 2*, ii215-225.

S20. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res *18*, 1979-1990.

S21. UniProt, C. (2015). UniProt: a hub for protein information. Nucleic Acids Res *43*, D204-212.

S22. Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061-1067.

S23. Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res *13*, 2178-2189.

S24. Dongen, S.v. (2000). Graph Clustering by Flow Simulation. (University of Utrecht).

S25. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

S26. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res *42*, D222-230.

S27. Mitchell, A., Chang, H.Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., et al. (2015). The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res *43*, D213-221.

S28. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol *7*, e1002195.

S29. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236-1240.
S30. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., et al. (2015). CDD: NCBI's conserved domain database. Nucleic Acids Res *43*, D222-226.
S31. Frickey, T., and Lupas, A. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics *20*, 3702-3704.
S32. Zhong, Y.F., and Holland, P.W. (2011). HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. Evol Dev *13*, 567-568.
S33. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28*, 27-30.
S34. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res *35*, W182-185.
S35. Csuros, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics *26*, 1910-1912.
S36. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol *30*, 772-780.
S37. Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972-1973.
S38. Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series *41*, 95-98.
S39. Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol Biol *7 Suppl 1*, S2.
S40. Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics *18*, 502-504.
S41. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312-1313.
S42. Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol *62*, 611-615.
S43. Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol *21*, 1095-1109.
S44. Struck, T.H., Wey-Fabrizius, A.R., Golombek, A., Hering, L., Weigert, A., Bleidorn, C., Klebow, S., Iakovenko, N., Hausdorf, B., Petersen, M., et al. (2014). Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of spiralia. Mol Biol Evol *31*, 1833-1849.
S45. Halanych, K.M., Bacheller, J.D., Aguinaldo, A.M., Liva, S.M., Hillis, D.M., and Lake, J.A. (1995). Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science *267*, 1641-1643.
S46. Simakov, O., Marletaz, F., Cho, S.J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.H., Larsson, T., Lv, J., Arendt, D., et al. (2013). Insights into bilaterian evolution from three spiralian genomes. Nature *493*, 526-531.
S47. Zheng, H., Zhang, W., Zhang, L., Zhang, Z., Li, J., Lu, G., Zhu, Y., Wang, Y., Huang, Y., Liu, J., et al. (2013). The genome of the hydatid tapeworm *Echinococcus granulosus*. Nat Genet *45*, 1168-1175.
S48. Tsai, I.J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sanchez-Flores, A., Brooks, K.L., Tracey, A., Bobes, R.J., Fragoso, G., Sciutto, E., et al. (2013). The genomes of four tapeworm species reveal adaptations to parasitism. Nature *496*, 57-63.