The slowly evolving genome of the xenacoelomorph worm *Xenoturbella bocki*

Philipp H. Schiffer[1,2,*], Paschalis Natsidis[1], Daniel J. Leite[1,3], Helen E. Robertson[1], François Lapraz[1,4], Ferdinand Marlétaz[1], Bastian Fromm[5], Liam Baudry[6], Fraser Simpson[1], Eirik Høye[7,8], Anne-C. Zakrzewski[1,9], Paschalia Kapli[1], Katharina J. Hoff[10,11], Steven Mueller[1,12], Martial Marbouty[13], Heather Marlow[14], Richard R. Copley[15], Romain Koszul[13], Peter Sarkies[16], Maximilian J. Telford[1,*]

*Corresponding authors: p.schiffer@uni-koeln.de, m.telford@ucl.ac.uk

1 Center for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK
2 Institute for Zoology, University of Cologne, 50674 Cologne, Germany
3 Department of Biosciences, Durham University, Durham DH1 3LE, UK
4 Université Côte D'Azur, CNRS, Inserm, iBV, Nice, France
5 The Arctic University Museum of Norway, UiT – The Arctic University of Norway, Tromsø, Norway
6 Collège Doctoral, Sorbonne Université, F-75005 Paris, France
7 Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway
8 Institute of Clinical Medicine, Medical Faculty, University of Oslo, Oslo, Norway
9 Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Invalidenstr. 43, 10115 Berlin, Germany
10 University of Greifswald, Institute for Mathematics and Computer Science, Greifswald, Germany
11 University of Greifswald, Center for Functional Genomics of Microbes, Greifswald, Germany
12 Royal Brompton Hospital, Guy's and St Thomas' NHS Foundation Trust
13 Institut Pasteur, Université de Paris, CNRS UMR3525, Unité Régulation Spatiale des Génomes, F-75015 Paris, France
14 The University of Chicago, Division of Biological Sciences, Chicago, IL 60637, USA
15 Laboratoire de Biologie du Développement de Villefranche-sur-mer (LBDV), Sorbonne Université, CNRS, 06230 Villefranche-sur-mer, France
16 Department of Biochemistry, University of Oxford, Oxford, UK

Abstract

The evolutionary origins of Bilateria remain enigmatic. One of the more enduring proposals highlights similarities between a cnidarian-like planula larva and simple acoel-like flatworms. This idea is based in part on the view of the Xenacoelomorpha as an outgroup to all other bilaterians which are themselves designated the Nephrozoa (protostomes and deuterostomes). Genome data can help to elucidate phylogenetic relationships and provide important comparative data. Here we assemble and analyse the genome of the simple, marine xenacoelomorph *Xenoturbella bocki*, a key species for our understanding of early bilaterian and deuterostome evolution. Our highly contiguous genome assembly of *X.*

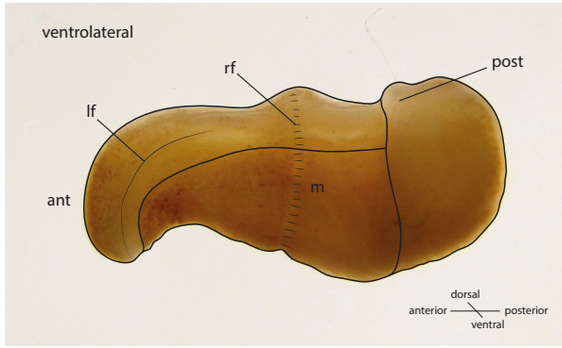
bocki has a size of ~110 Mbp in 18 chromosome like scaffolds, with repeat content and intron, exon and intergenic space comparable to other bilaterian invertebrates. We find *X. bocki* to have a similar number of genes to other bilaterians and to have retained ancestral metazoan synteny. Key bilaterian signalling pathways are also largely complete and most bilaterian miRNAs are present. We conclude that *X. bocki* has a complex genome typical of bilaterians, in contrast to the apparent simplicity of its body plan. Overall, our data do not provide evidence supporting the idea that Xenacoelomorpha are a primitively simple outgroup to other bilaterians and gene presence/absence data support a relationship with Ambulacraria.

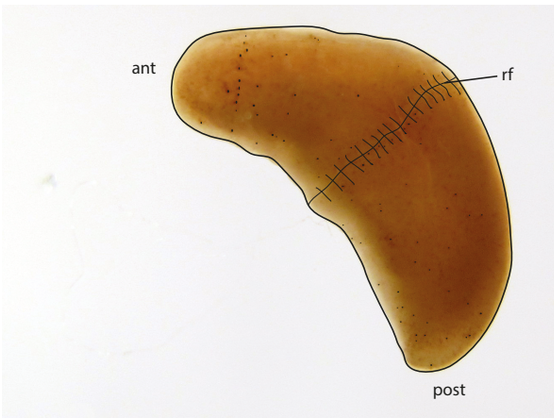
Introduction

Xenoturbella bocki (Fig 1) is a morphologically simple marine worm first described from specimens collected from muddy sediments in the Gullmarsfjord on the West coast of Sweden. There are now 6 described species of *Xenoturbella* - the only genus in the higher-level taxon of Xenoturbellida[1]. *X. bocki* was initially included as a species within the Platyhelminthes[2], but molecular phylogenetic studies have shown that Xenoturbellida is the sister group of the Acoelomorpha, a second clade of morphologically simple worms also originally considered Platyhelminthes. Analyses of phylogenomic data sets have shown that Xenoturbellida and Acoelomorpha constitute their own phylum, the Xenacoelomorpha[3,4]. The monophyly of Xenacoelomorpha is also convincingly supported by their sharing unique amino acid signatures in their Caudal genes[3] and Hox4/5/6 gene[5]. In the present work we analyse our data in this phylogenetic framework of a monophyletic taxon Xenacoelomorpha.

The simplicity of xenacoelomorph species compared to other bilaterians is a central feature of discussions over their evolution. While Xenacoelomorpha are clearly monophyletic, their phylogenetic position within the Metazoa has been controversial for a quarter of a century. There are two broadly discussed scenarios: a majority of studies have supported a position for Xenacoelomorpha as the sister group of all other Bilateria (the Protostomia and Deuterostomia, collectively named Nephrozoa)[4,6–8]; work we have contributed to[1,3,9,10], has instead placed Xenacoelomorpha within the Bilateria as the sister group of the Ambulacraria (Hemichordata and Echinodermata) to form a clade called the Xenambulacraria[9].

1a





1b

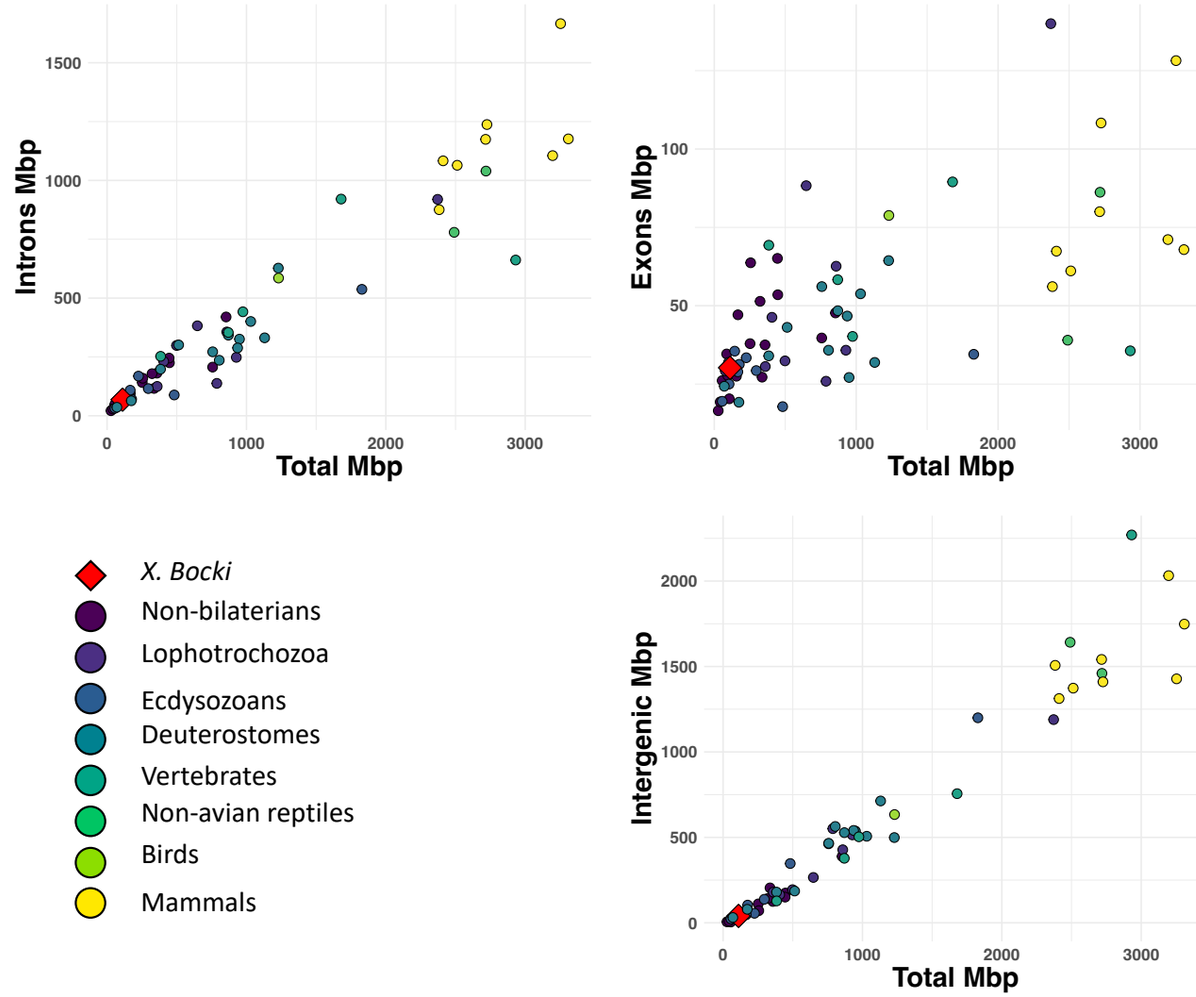


Figure 1: (a) Schematic drawings of *Xenoturbella bocki* showing the simple body organisation of the marine vermiform animal. Abbreviations: ant - anterior, post - posterior, lf - lateral furrow, rf - ring furrow, m - mouth opening. (b) A comparison of total length of exons, introns, and intergeneic space in the *X. bocki* genome with other metazoans (data from ref 20). *X. bocki* does not appear to be an outlier in any of theses comparisons.

75 *Xenoturbella bocki* has neither organized gonads nor a centralized nervous
76 system. It has a blind gut, no body cavities and lacks nephrocytes[11]. If
77 Xenacoelomorpha is the sister group to Nephrozoa these character absences can be
78 parsimoniously interpreted as representing the primitive state of the Bilateria.
79 According to advocates of the Nephrozoa hypothesis, these and other characters
80 absent in Xenacoelomorpha must have evolved in the lineage leading to Nephrozoa
81 after the divergence of Xenacoelomorpha. More generally there has been a tendency
82 to interpret Xenacoelomorpha (especially Acoelomorpha) as living approximations of
83 Urbilateria[12].
84 An alternative explanation for the simple body plan of xenaceolomorphs is
85 that it is derived from that of more complex urbilaterian ancestors through loss of
86 morphological characters. The loss or remodelling of morphological complexity is a
87 common feature of evolution in many animal groups and is typically associated with
88 unusualmodes of living[13,14] – in particular the adoption of a sessile (sea squirts,
89 barnacles) or parasitic (neodermatan flatworms, orthonectids) lifestyle, extreme
90 miniaturization (e.g. tardigrades, orthonectids), or even neoteny (e.g. flightless
91 hexapods).
92 In the past some genomic features gleaned from analysis of various
93 Xenacoelomorpha have been used to test these evolutionary hypotheses. For
94 example, the common ancestor of the protostomes and deuterostomes has been
95 reconstructed with approximately 8 Hox genes but only 4 have been found in the
96 Acoelomorpha (Nemertoderma) and 5 In *Xenoturbella*. This has been interpreted as
97 a primary absence with the full complement of 8 appearing subsequent to the
98 divergence of Xenacoelomorpha and Nephrozoa. Similarly, analysis of the microRNAs
99 (miRNAs) of an acoelomorph, *Symsagittifera roscoffensis*, found that many bilaterian
100 miRNAs were absent from its genome[15]. Some of the missing bilaterian miRNAs,
101 however, were subsequently observed in *Xenoturbella*[9].
102 The few xenacoelomorph genomes available to date are from the acoel
103 *Hofstenia miamia*[16] – like other Acoelomorpha it shows accelerated sequence
104 evolution relative to *Xenoturbella*[3] – and from two closely related species
105 *Praesagittifera naikaiensis*[17] and *Symsagittifera roscoffensis*[18]. The analyses of gene
106 content of *Hofstenia* showed similar numbers of genes and gene families to other

107 bilaterians[16], while an analysis of the neuropeptide content concluded that most
108 bilaterian neuropeptides were present in Xenacoelomorpha[19].
109 In order to infer the characteristics of the ancestral xenacoelomorph genome,
110 and to complement the data from the Acoelomorpha, we describe a highly-scaffolded
111 genome of the slowly evolving xenacoelomorph *Xenoturbella bocki*. This allows us to
112 contribute knowledge of Xenacoelomorpha and *Xenoturbella* in particular with
113 genomic traits, such as gene content and genome-structure and to help reconstruct
114 the genome structure and composition of the ancestral xenacoelomorph.
115

116 # Results

117 **Assembly of a draft genome of *Xenoturbella bocki*.**

118 We collected *Xenoturbella bocki* specimens (Fig. 1) from the bottom of the
119 Gullmarsfjord close to the biological field station in Kristineberg (Sweden). These adult
120 specimens were starved for several days in tubes with artificial sea water, and then
121 sacrificed in lysis buffer. We extracted high molecular weight (HMW) DNA from single
122 individuals for each of the different sequencing steps below.
123 We assembled a high-quality draft genome of *Xenoturbella bocki* using one short
124 read Illumina library and one TruSeq Synthetic Long Reads (TSLR) Illumina library.
125 We used a workflow based on a primary assembly with SPAdes (Methods; [20]). The
126 primary assembly had an N50 of 8,500kb over 37,880 contigs with a maximum length
127 of 206,709bp. After using the redundans pipeline[21] this increased to an N50 of ~62kb
128 over 23,094 contigs and scaffolds spanning ~121Mb, and a longest scaffold of
129 960,978kb (supplementary Table 1).
130 The final genome was obtained with Hi-C scaffolding using the program
131 instaGRAAL (Methods, see supplementary for contact map;[22]). The scaffolded
132 genome has a span of 111 Mbp (117 Mbp including small fragments unincorporated
133 into the HiC assembly) and an N50 of 2.7 Mbp (for contigs >500bp). The assembly
134 contains 18 megabase-scale scaffolds encompassing 72 Mbp (62%) of the genomic
135 sequence, with 43% GC content. The original assembly indicated a repeat content of
136 about 25% after a RepeatModeller based RepeatMasker annotation (Methods). As
137 often seen in non-model organisms, about 2/3 of the repeats are not classified.
138 We used BRAKER1[23,24] with extensive RNA-Seq data, and additional single-cell

139 UTR enriched transcriptome sequencing data to predict 15,154 gene models. 9,575
140 gene models (63%) are found on the 18 large scaffolds (which represent 62% of the
141 total sequence). 13,298 of our predicted genes (88%) have RNA-Seq support.
142 Although this is proportion is at the low end of bilaterian gene counts, we note that our
143 RNA-seq libraries were all taken from adult animals and thus may not represent the
144 true complexity of the gene complement. We consider our predicted gene number to
145 be a lower bound estimate for the true gene content.
146 The predicted *X. bocki* genes have a median coding length of 873 nt and a mean
147 length of 1330 nt. Median exon length is 132 nt (mean 212 nt) and median intron length
148 is 131 nt (mean 394 nt). Genes have a median of 4 exons and a mean of 8.5 exons.
149 2,532 genes have a single exon and, of these, 1,381 are supported as having a single
150 exon by RNA-Seq (TPM>1). A comparison of the exon, intron, and intergenic
151 sequence content in *Xenoturbella* with descriptions of other animal genomes[25] show
152 that *X. bocki* falls within the range of other similarly sized metazoan genomes (Fig. 1b)
153 for all these measures.
154
155 The genome of a co-sequenced *Chlamydia* species
156 We recovered the genome of a marine *Chlamydia* species from Illumina data obtained
157 from one *X. bocki* specimen and from Oxford Nanopore data from a second specimen
158 supporting previous microscopic analyses and single gene PCRs suggesting that *X.*
159 *bocki* is host to a species in the bacterial genus *Chlamydia*. The bacterial genome was
160 found as 5 contigs spanning 1,906,303 bp (N50 of 1,237,287 bp) which were
161 assembled into 2 large scaffolds. Using PROKKA[26], we predicted 1,738 genes in this
162 bacterial genome, with 3 ribosomal RNAs, 35 transfer RNAs, and 1 transfer-
163 messenger RNA. The genome is 97.5% complete for bacterial BUSCO[27] genes,
164 missing only one of the 40 core genes.
165 Marine chlamydiae are not closely related to the group of human pathogens[28]
166 and we were not able to align the genome of the *Chlamydia*-related symbioint from *X.*
167 *bocki* to the reference strain *Chlamydia trachomatis* F/SW4, nor to *Chlamydophila*
168 *pneumoniae* TW-183. To investigate the phylogenetic position of the species co-
169 occurring with *Xenoturbella*, we aligned the 16S rRNA gene from the *X. bocki*-hosted
170 *Chlamydia* with orthologs from related species including sequences of genes amplified

171 from DNA/RNA extracted from deep sea sediments. The *X. bocki*-hosted *Chlamydia*
172 belong to a group designated as Simkaniaceae in[28], with the sister taxon in our
173 phylogenetic tree being the *Chlamydia* species previously found in *X. westbladi* (*X.*
174 *westbladi* is almost certainly a synonym of *X. bocki*)[7] (Fig. 2a).
175 To investigate whether the *X. bocki*-hosted *Chlamydia* might contribute to the
176 metabolic pathways of its host, we compared the completeness of metabolic pathways
177 in KEGG for the *X. bocki* genome alone and for the *X. bocki* genome in combination
178 with the bacteria. We found only slightly higher completeness in a small number of
179 pathways involved in carbohydrate metabolism, carbon fixation, and amino acid
180 metabolism (see supplementary material) suggesting that the relationship is likely to
181 be commensal or parasitic rather than a true symbiosis.
182 A second large fraction of bacterial reads, annotated as Gammaproteobacteria,
183 were identified and filtered out during the data processing steps. These bacteria were
184 also previously reported as potential symbionts of *X. bocki*[29]. However, these
185 sequences were not sufficiently well covered to reconstruct a genome and we did not
186 investigate them further.
187
188 **A phylogenetic gene presence/absence matrix supports Xenambulacraria**
189 The general completeness of the *X. bocki* gene set allowed us to use the presence
190 and absence of genes identified in our genomes as a source of information to find the
191 best supported phylogenetic position of the Xenacoelomorpha. We conducted two
192 separate phylogenetic analyses of gene presence/absence data: one including the
193 fast-evolving Acoelomorpha and one without. In both analyses the best tree grouped
194 *Xenoturbella* with the Ambulacraria (Fig. 2b). The analysis including acoels, however,
195 placed the acoels as the sister-group to Nephrozoa separate from *Xenoturbella* (Fig.
196 2c). Because other data have shown the monophyly of Xenacoelomorpha to be robust,
197 we interpret this result as being the result of systematic error caused by a high rate of
198 gee loss or by orthologs being incorrectly scored as missing due to higher rates of
199 sequence evolution in acoelomorphs[30].
200
201 **The *X. bocki* molecular toolkit is typical of bilaterians.**
202 One of our principal aims was to ask whether the *Xenoturbella* genome lacks
203 characteristics otherwise present in the Bilateria. We found that for the Metazoa gene

2a

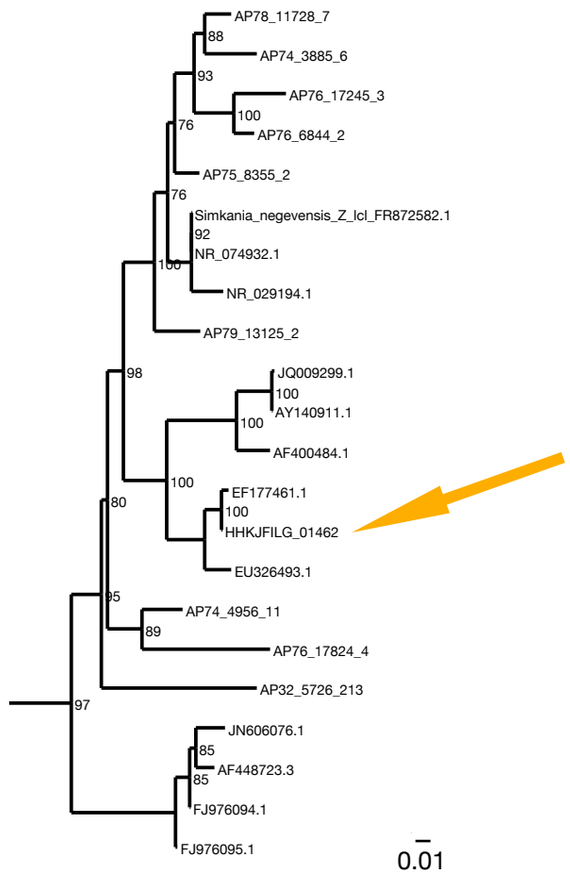
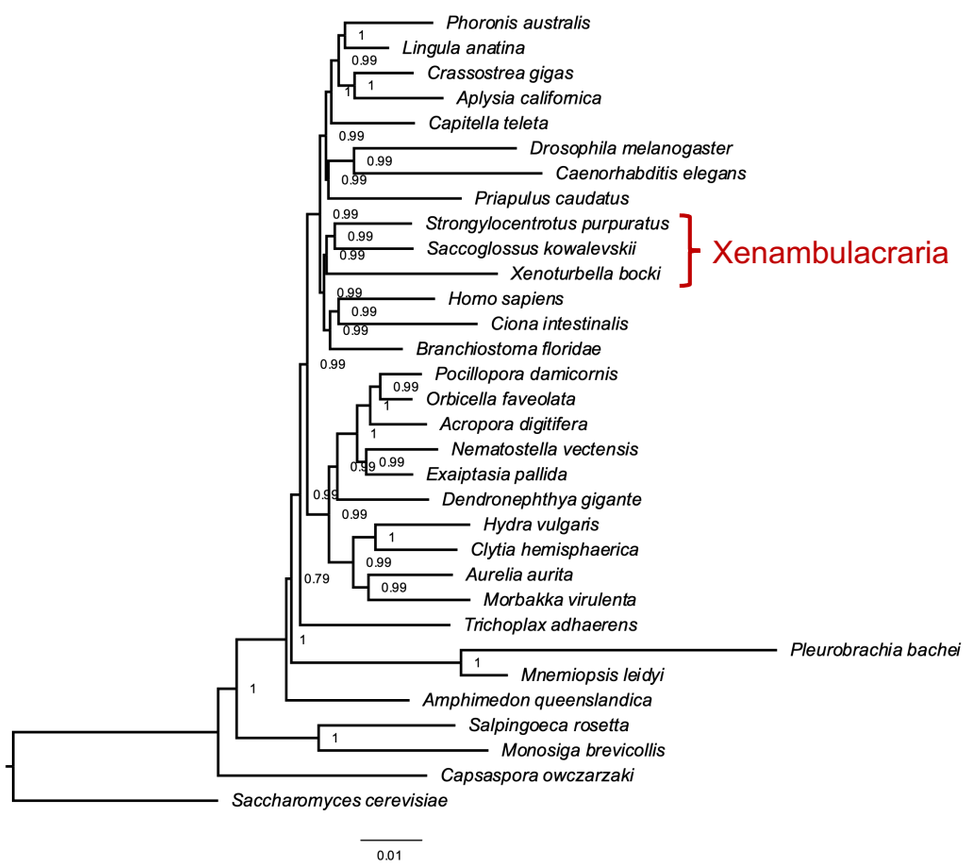
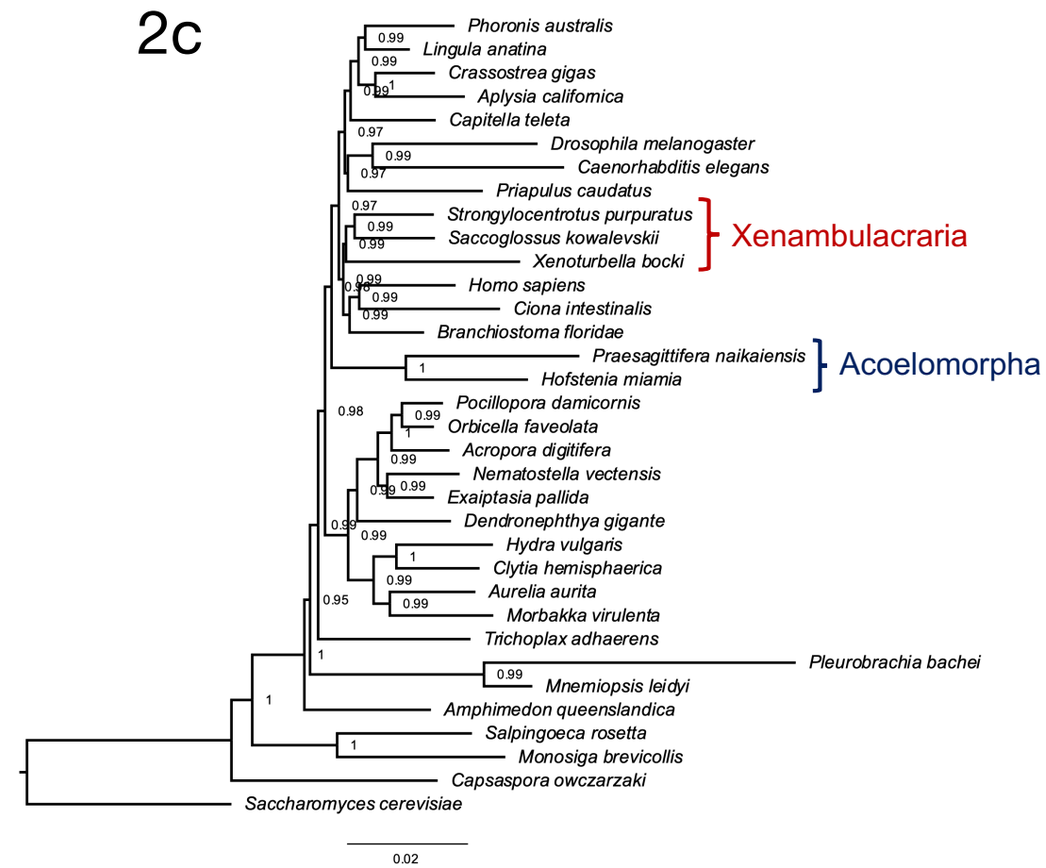


Figure 2a: *Xenoturbella bocki* harbours a marine Chlamydiae species as potential symbiont. In the phylogenetic analysis of 16S rDNA (ML: GTR+F+R7; bootstrap values included) the bacteria in our *X. bocki* isolate (arrow) are sister to a previous isolate from *X. westbaldi*. *X. westbaldi* is most likely a mis-identification of *X. bocki.* (b/c) A phylogeny based on presence and absence of genes calculated with OMA. Both analysis (b) and (c) confirm Xenambularcraria, i.e. Xenoturbellida in a group with Echinoderms and Hemichordates. Inclusion of the acoel flat worms places these as sister to all other Bilateria (b). This placement appears an artefact due to the very fast evolution in this taxon, in particular as good evidence exists for uniting Xenoturbellida and Acoela refs 5 and 6.

2b

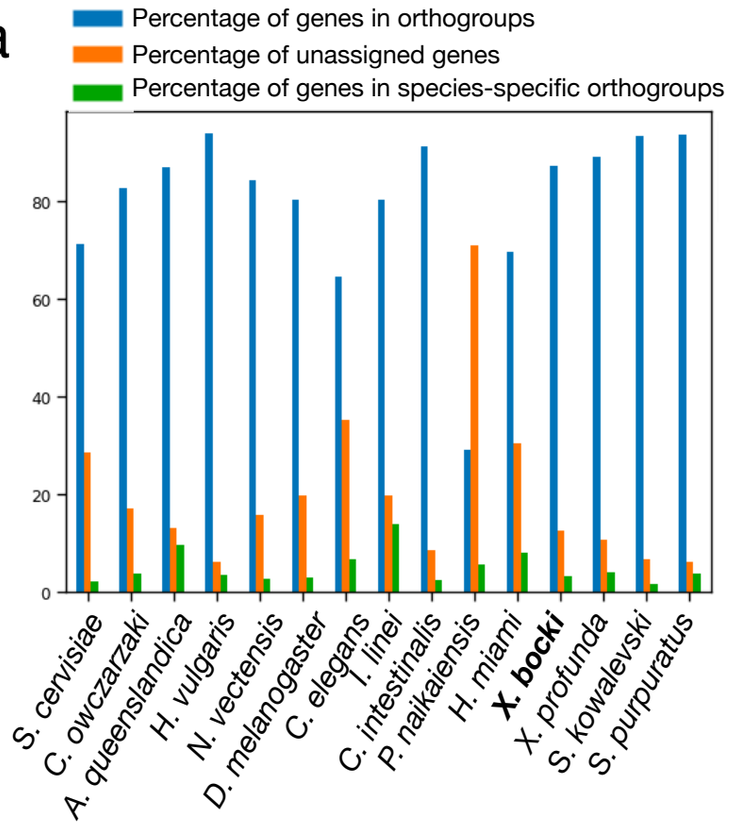


2c

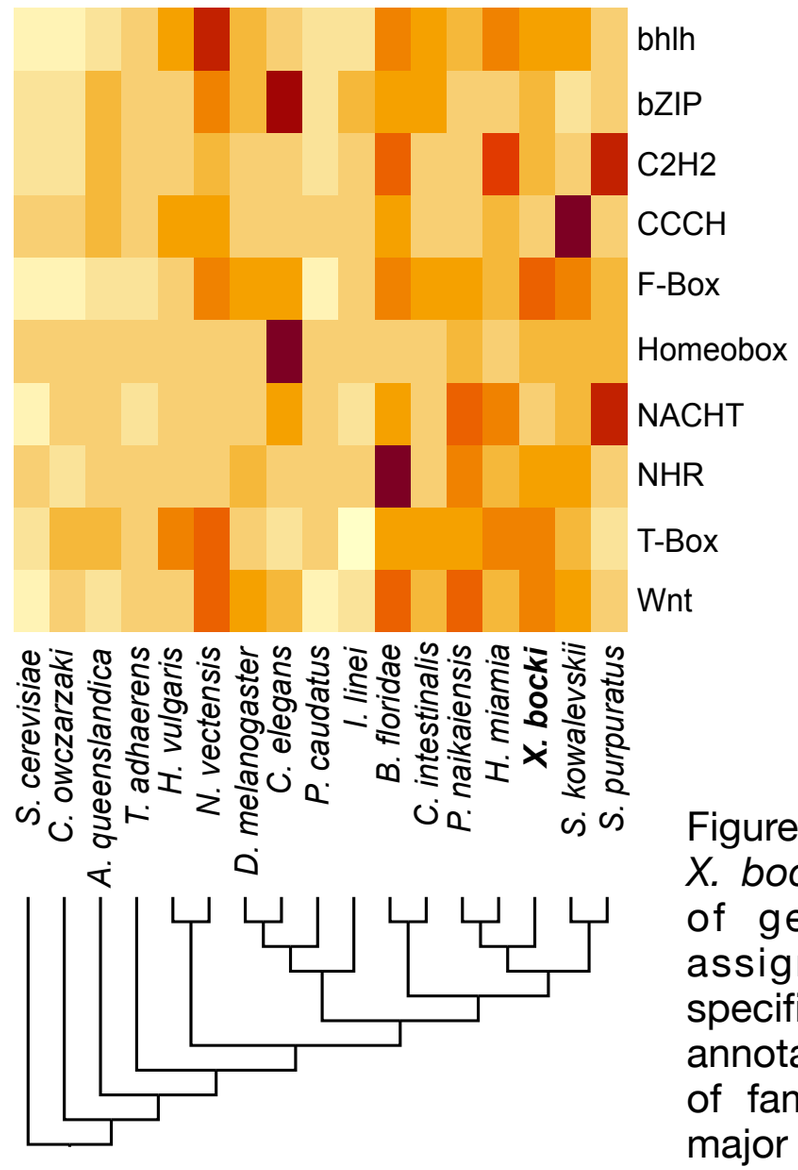


204 set in BUSCO (v5) the *X. bocki* proteome translated from our gene predictions is
205 82.5% complete and ~90% complete when partial hits are included. This estimate is
206 even higher in the acoel *Hofstenia miamia*, which was originally reported to be 90%[16],
207 but in our re-analysis was 95.71%. In comparison, the morphologically highly
208 simplified and fast evolving annelid *Intoshia linei*[31] has a genome of fewer than 10,000
209 genes[32] and in our analysis is only ~64% complete for the BUSCO (v5) Metazoa set.
210 The model nematode *Caenorhabditis elegans* is ~81% complete for the same set.
211 Despite the morphological simplicity of both *Xenoturbella*, and *Hofstenia,* these
212 Xenacoelomorpha are missing few core genes compared to other bilaterian lineages
213 that we perceive to have undergone a high degree of morphological evolutionary
214 change (associated with miniaturisation, parasitism, sessility etc).
215 Using our phylogenomic matrix of gene presence/absence (see above) we identified
216 all orthologs present in any bilaterian and any non-bilaterian; although a minority of
217 these may be explained by horizontal gene transfer or other rare events, most of these
218 must have existed in the bilaterian ancestor. All individual bilaterian genomes were
219 missing many of these orthologs but Xenacoelomorphs and some other bilaterians
220 lacked more of these than did other taxa. The average numbers of these genes
221 present in bilaterians = 7577; *Xenoturbella* = 5459; *Hofstenia* = 5438; *Praesagittifera*
222 = 4280; *Drosophila* = 4844; *Caenorhabditis* = 4323.
223 To better profile the *Xenoturbella* and xenacoelomorph molecular toolkit, we
224 used OrthoFinder to conduct orthology searches in a comparison of 155 metazoan
225 and outgroup species, including the transcriptomes of the sister species *X. profunda*
226 and an early draft genome of the acoel *Paratomella rubra* we had available*,* as well as
227 the *Hofstenia* and *Praesagittifera* proteomes (Supplementary online material). For
228 each species, we counted, in each of the three Xenacoelomorphs, the number of
229 orthogroups for which a gene was present. The proportion of orthogroups containing
230 an *X. bocki* and *X. profunda* protein (87.4% and 89.2%) are broadly similar to the
231 proportions seen in other well characterised genomes, for example *S. purpuratus*
232 proteins (93.8%) or *N. vectensis* proteins (84.3%) (Fig 3a). In this analysis, the fast-
233 evolving nematode *Caenorhabditis elegans* appears as an outlier, with only ~64% of
234 its proteins in orthogroups and ~35% unassigned. Both *Xenoturbella* species have an
235 intermediate number of unassigned genes of ~11-12%. Similarly, the proportion of

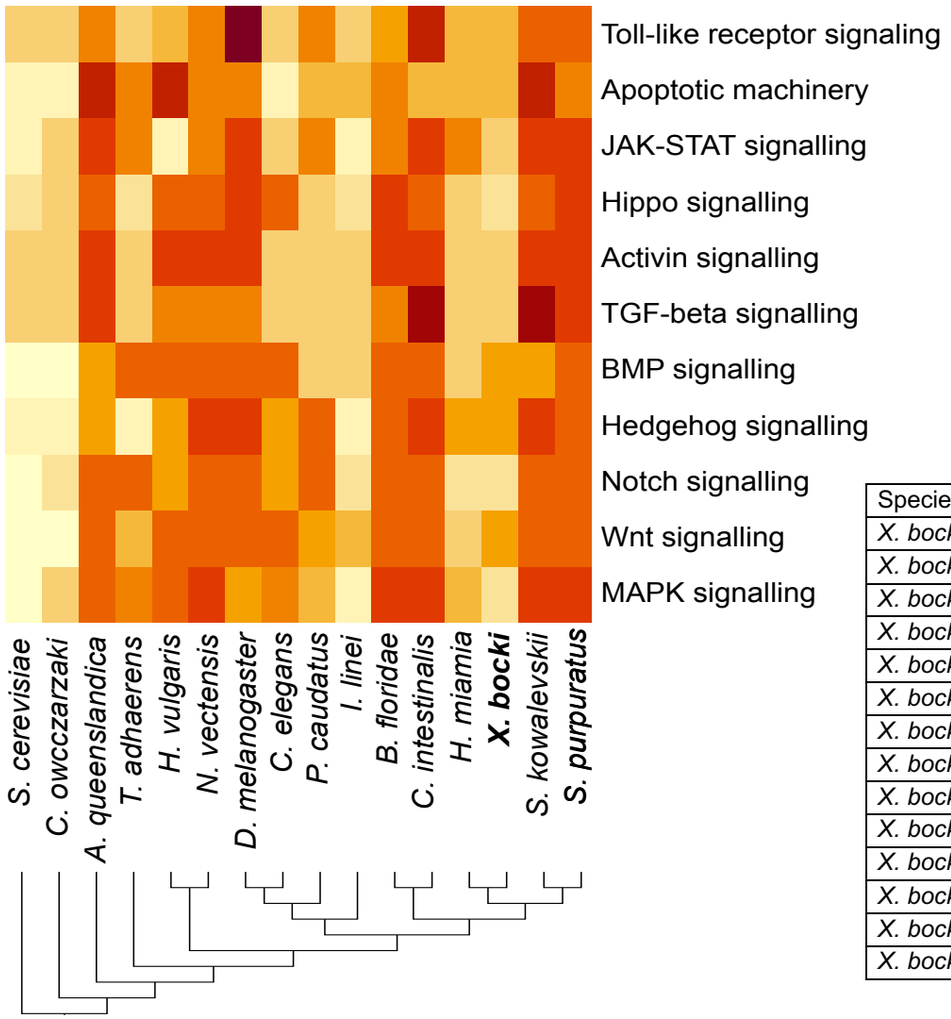
3a



3b



3c



t-test

Species 1	Species 2	p-value
X. bocki	*H. miamia*	0.9448
X. bocki	*S. kowalevskii*	**0.0001974**
X. bocki	*S. purpuratus*	**0.0004118**
X. bocki	*C. intestinalis*	**0.0003404**
X. bocki	*B. floridae*	**0.004928**
X. bocki	*C. elegans*	0.3893
X. bocki	*D. melanogaster*	**0.0004194**
X. bocki	*I. linei*	0.2469
X. bocki	*N. vectensis*	**0.00277**
X. bocki	*H. vulgaris*	0.06593
X. bocki	*T. adhaerens*	0.4552
X. bocki	*A. queenslandica*	**0.001896**
X. bocki	*C. owczarzaki*	0.1184
X. bocki	*S. cerevisiae*	**0.007309**

Figure 3: (a) In our orthology screen *X. bocki* shows similar percentages of genes in orthogroups, unassigned genes, and species-specific orthogroups as other well-annotated genomes. (b) The number of family members per species in major gene families (based on Pfam domains), like transcription factors, fluctuates in evolution. The *X. bocki* genome does not appear to contain particularly less or more genes in any of the analysed families. (c) Cell signalling pathways in *X. bocki* are functionally complete, but in comparison to other species contain less genes. The overall completeness is not significantly different to, for example, the nematode *C. elegans* (inset, t-test). Schematic cladograms in b/c drawn by the authors.

236 species-specific genes (~14% of all genes) corresponds closely to what is seen in
237 most other species (with the exception of the parasitic annelid *I. linei*, Fig. 3a).
238
239 Idiosyncrasies of *Xenoturbella*
240 In order to identify sets of orthologs specific to the two *Xenoturbella* species we used
241 the kinfin software[33] and found 867 such groups in the OrthoFinder clustering. We
242 profiled these genes based on Pfam domains and GO terms derived from
243 InterProScan. While these *Xenoturbella* specific proteins fall into diverse classes, we
244 did see a considerable number of C-type lectin, Immunoglobulin-like, PAN, and Kringle
245 domain containing Pfam annotations. Along with the Cysteine-rich secretory protein
246 family and the G-protein coupled receptor activity GO terms, these genes and families
247 of genes may be interesting for future studies into the biology of *Xenoturbella* in its
248 native environment.
249
250 Gene families and signaling pathways are retained in *X. bocki*
251 In our orthology clustering we did not see an inflation of *Xenoturbella*-specific groups
252 in comparison to other taxa, but also no conspicuous absence of major gene families
253 (Fig. 3b). Family numbers of transcription factors like Zinc-fingers or homeobox-
254 containing genes, as well as, for example, NACHT-domain encoding genes seem to
255 be neither drastically inflated nor contracted in comparison to other species in our
256 InterProScan based analysis.
257 To catalogue the completeness of cell signalling pathways we screened the *X.*
258 *bocki* proteome against KEGG pathway maps using GenomeMaple[34]. The *X. bocki*
259 gene set is largely complete in regard to the core proteins of these pathways, while an
260 array of effector proteins is absent (Fig. 3c). In comparison to other metazoan species,
261 as well as a unicellular choanoflagellate and a yeast, the *X. bocki* molecular toolkit has
262 significantly lower KEGG completeness than morphologically complex animals such
263 as the sea urchin and amphioxus (t-test; Fig. 3c). *Xenoturbella* is, however, not
264 significantly less complete when compared to other bilaterians considered to have low
265 morphological complexity and which have been shown to have reduced gene content,
266 such as *C. elegans*, the annelid parasite *Intoshia linei*, or the acoel *Hofstenia miamia*
267 (Fig. 3c).

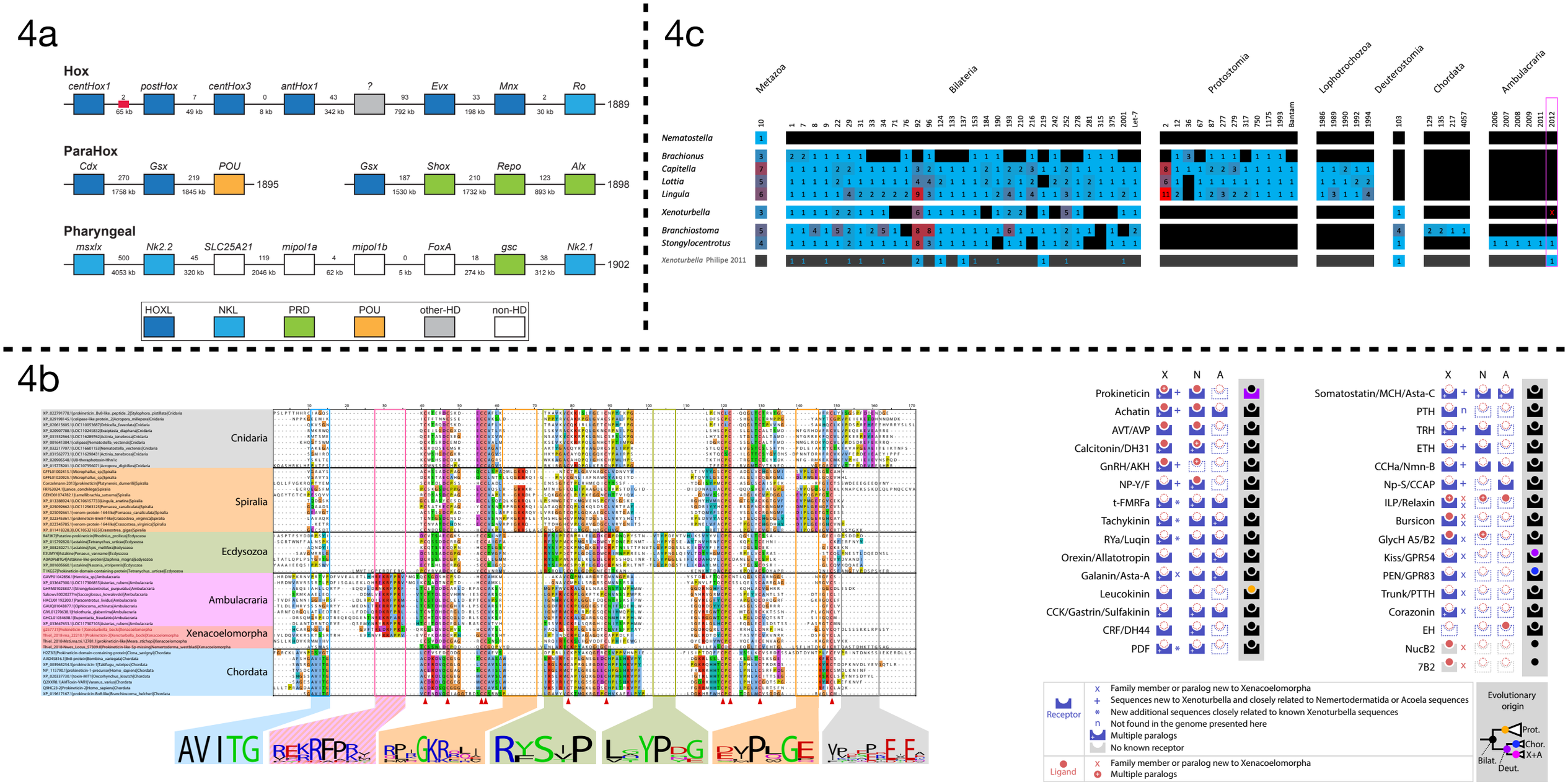


Figure 4: (a) *X. bocki* has 5 HOX genes, which are located in relatively close proximity on one of our chromosome size scaffolds. Similar clusters exist for the ParaHox and "pharyngeal" genes. Numbers between genes are distance (below) and number of genes between (below). Colours indicate gene families. Red box marks the position of a partial Hox gene. The "?" gene has an unresolved homeodomain identity. (b) We found a specific prokineticin ligand signature sequence in *X. bocki*, which was previously reported for Ecdysozoa and Chordata, as well as a "K/R-RFP-K/R", sequence shared only by ambulacrarians and *X. bocki*. Signature previously reported for Ecdysozoa and Chordata, as well as new signatures we found in Spiralia and Cnidaria is absent from ambulacrarian and *X. bocki* prokineticin ligand sequences. The inset cladogram depicts the evolutionary origin of sequences in accordance with our analysis: **Bilat**erian, **Prot**ostomia, **Chor**date, **X**enacoelomorpha + **A**mbulacraria last common ancestor respectively. (c) The revised microRNA complement of *X. bocki* has a near complete set of metazoan, bilaterian and deuterostome families and genes. Presence (color) and absence (black) of microRNA families (column), paralogue numbers (values & heatmap coloring) organized in node-specific blocks in a range of representative protostome and deuterostome species compared with *Xenoturbella* (species from MirGeneDB 2.1 - Fromm et al 2021). The bottom row depicts 2011 complement by Philippe et al 2011 (blue numbers on black depict detected miRNA reads, but lack of genomic evidence). Red "x" in pink box highlights the lack of evidence for an ambulacraria-specific microRNA in *X. bocki*.

268 Clustered homeobox genes in the *X. bocki* genome

269 Acoelomorph flatworms possess three unlinked HOX genes, orthologs of anterior
270 (Hox1), central (Hox4/5 or Hox5) and posterior Hox (HoxP). In contrast, previous
271 analysis of *X. bocki* transcriptomes identified one anterior, three central and one
272 posterior Hox genes. We identified clear evidence of a syntenic Hox cluster with four
273 Hox genes (centHox1, postHox, centHox3, and antHox1) in the *X. bocki* genome (Fig.
274 4). There was also evidence of a fragmented annotation of centHox2, split between
275 the 4 gene Hox cluster and a separate scaffold (Fig. 4). In summary, this suggests that
276 all five Hox genes form a Hox cluster in the *X. bocki* genome, but that there are
277 possible unresolved assembly errors disrupting the current annotation. We also
278 identified other homeobox genes on the Hox cluster scaffold, including Evx (Fig. 4a).

279 Along with the Hox genes, we also surveyed other homeobox genes that are
280 typically clustered in Bilateria. The canonical bilaterian paraHox cluster contains three
281 genes Cdx, Xlox (=Pdx) and Gsx. We identified Cdx and a new Gsx annotation on the
282 same scaffold, as well as a previously reported Gsx paralog on a separate scaffold.
283 This indicates partial retention of the paraHox cluster in *X. bocki* along with a
284 duplication of Gsx. On both of these paraHox containing scaffolds we observed other
285 homeobox genes.

286 Hemichordates and chordates have a conserved cluster of genes involved in
287 patterning their pharyngeal pores - the so-called 'pharyngeal cluster'. The homeobox
288 genes of this cluster (Msxlx, Nk2-1/2/4/8) were present on a single *X. bocki* scaffold.
289 Another pharyngeal cluster transcription factor, the Forkhead containing Foxa, and
290 'bystander' genes from that cluster including Egln, Mipol1 and Slc25a21 are found in
291 the same genomic region. Different sub-parts of the cluster are found in non-bilaterians
292 and protostomes and the cluster may well be plesiomorphic for the Bilateria rather
293 than a deuterostome synapomorphy[35].

294

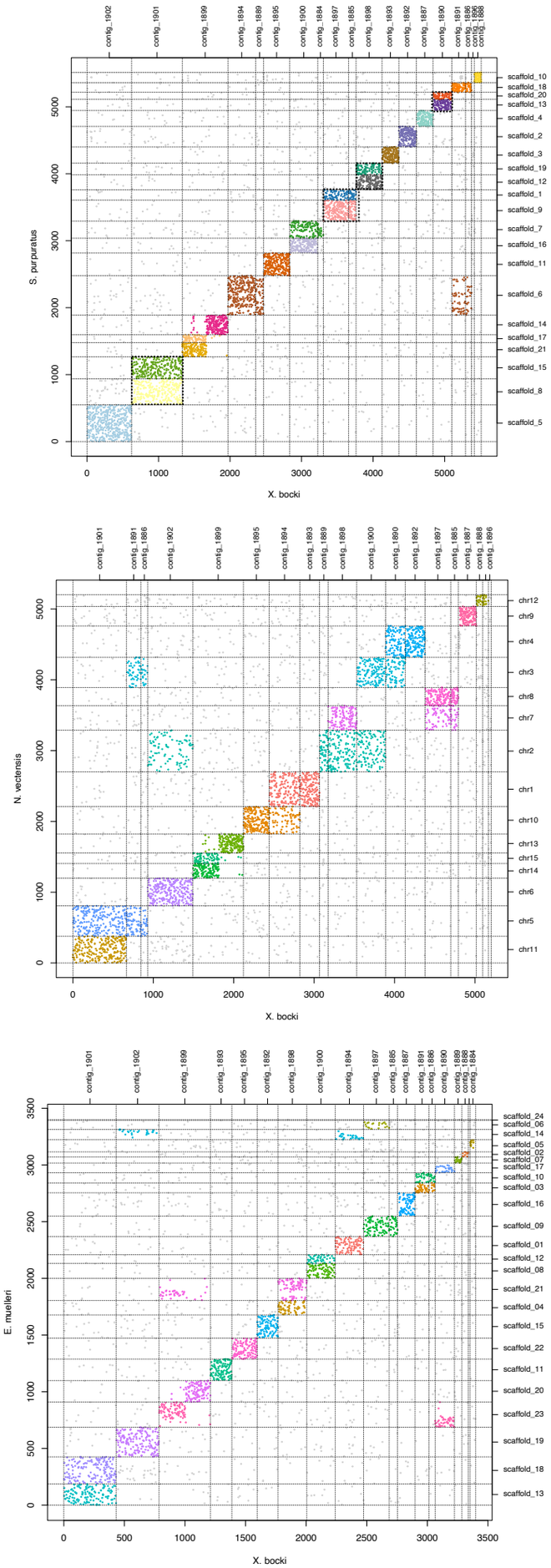
295 The *X. bocki* neuropeptide complement is larger than previously thought

296 A catalogue of acoelomorph neuropeptides was previously described using
297 transcriptome data[36]. We have discovered 12 additional neuropeptide genes and 39
298 new neuropeptide receptors in *X.bocki* adding 6 bilaterian peptidergic systems to the
299 *Xenoturbella* catalogue (NPY-F ; MCH/Asta-C ; TRH ; ETH ; CCHa/Nmn-B ; Np-

300 S/CCAP), and 6 additional bilaterian systems to the Xenacoelomorpha catalogue
301 (Corazonin ; Kiss/GPR54 ; GPR83 ; 7B2 ; Trunk/PTTH ; NUCB2) making a total of 31
302 peptidergic systems (Fig. 4, Supplementary).
303 Among the ligand genes, we identified 6 new repeat-containing sequences. One
304 of these, the LRIGamide-peptide, had been identified in Nemertodermatida and
305 Acoela and its loss in *Xenoturbella* had been proposed[36]. We also identified the first
306 7B2 neuropeptide and NucB2/Nesfatin genes in Xenacoelomorpha. Finally, we
307 identified 3 new *X.bocki* insulin-like peptides, one of them showing sequence similarity
308 and an atypical cysteine pattern shared with the Ambulacrarian octinsulin, constituting
309 a potential synapomorphy of Xenambulacraria (see Supplementary).
310 Our searches also revealed the presence of components of the arthropod
311 moulting pathway components (PTTH/trunk, NP-S/CCAP and Bursicon receptors),
312 which have recently been shown to be of ancient origin (de Oliveira et al., 2019). We
313 further identified multiple paralogs for, e.g the Tachykinin, Rya/Luquin, tFMRFa,
314 Corazonin, Achatin, CCK, and Prokineticin receptor families. Two complete *X. bocki*
315 Prokineticin ligands were also found in our survey (see Supplementary).
316 Chordate Prokineticin ligands possess a conserved N-terminal "AVIT" sequence
317 required for the receptor activation[37]. This sequence is absent in arthropod Astakine,
318 which instead possess two signature sequences within their Prokineticin domain [38].
319 To investigate Prokineticin ligands in Xenacoelomorpha we compared the sequences
320 of their prokineticin ligands with those of other bilaterians (Fig. 4b, Supplementary).
321 Our alignment reveals clade specific signatures already reported in Ecdysozoa and
322 Chordata sequences, but also two new signatures specific to Lophotrochozoa and
323 Cnidaria sequences, as well as a very specific "K/R-RFP-K/R" signature shared only
324 by ambulacrarian and *Xenoturbella bocki* sequences. The shared
325 Ambulacrarian/Xenacoelomorpha signature is found at the same position as the
326 Chordate sequence involved in receptor activation - adjacent to the N-terminus of the
327 Prokineticin domain (Fig. 4b).
328
329 **The *X. bocki* genome contains most bilaterian miRNAs reported missing from**
330 **acoels.**
331 microRNAs have previously been used to investigate the phylogenetic position of the

332 acoels and *Xenoturbella.* The acoel *Symsagittifera roscoffensis* lacks protostome and
333 bilaterian miRNAs and this lack was interpreted as supporting the position of acoels
334 as sister-group to the Nephrozoa. Based on shallow 454 microRNA sequencing (and
335 sparse genomic traces) of *Xenoturbella*, some of the bilaterian miRNAs missing from
336 acoels were found - 16 of the 32 expected metazoan (1 miRNA) and bilaterian (31
337 miRNAs) microRNA families – of which 6 could be identified in genome traces[9].
338 By deep sequencing two independent small RNA samples, we have now
339 identified the majority of the missing metazoan and bilaterian microRNAs and
340 identified them in the genome assembly (Fig. 4c). Altogether, we found 23 out of 31
341 bilaterian microRNA families (35 genes including duplicates); the single known
342 Metazoan microRNA family (MIR-10) in 2 copies; the Deuterostome-specific MIR-103;
343 and 7 *Xenoturbella*-specific microRNAs giving a total of 46 microRNA genes. None of
344 the protostome-specific miRNAs were found. We could not confirm in the RNA
345 sequences or new assembly a previously identified, and supposedly
346 xenambulacrarian-specific MIR-2012 ortholog.
347
348 **The *X. bocki* genome retains ancestral metazoan linkage groups.**
349 The availability of chromosome-scale genomes has made it possible to reconstruct 24
350 ancestral linkage units broadly preserved in bilaterians[39]. In fast-evolving genomes,
351 such as those of nematodes, tunicates or platyhelminths, these ancestral linkage
352 groups (ALGs) are often dispersed and/or extensively fused (Supplementary). We
353 were interested to test if the general conservation of the gene content in *X. bocki* is
354 reflected in its genome structure.
355 We compared the genome of *Xenoturbella* to several other metazoan genomes
356 and found that it has retained most of these ancestral bilaterian units: 12
357 chromosomes in the *X. bocki* genome derive from a single ALG, five chromosomes
358 are made of the fusion of two ALGs, and one *Xenoturbella* chromosome is a fusion of
359 three ALGs, as highlighted with the comparison of ortholog content with amphioxus,
360 the sea urchin and the sea scallop (Fig. 5 and Supplementary).
361 One ancestral linkage group that has been lost in chordates but not in
362 ambulacrarians nor in molluscs (ALG R in sea urchin and sea scallop) is detectable in
363 *X. bocki* (Fig. 5), while *X. bocki* does not show the fusions that are characteristic of
364 lophotrochozoans.

5a



5b

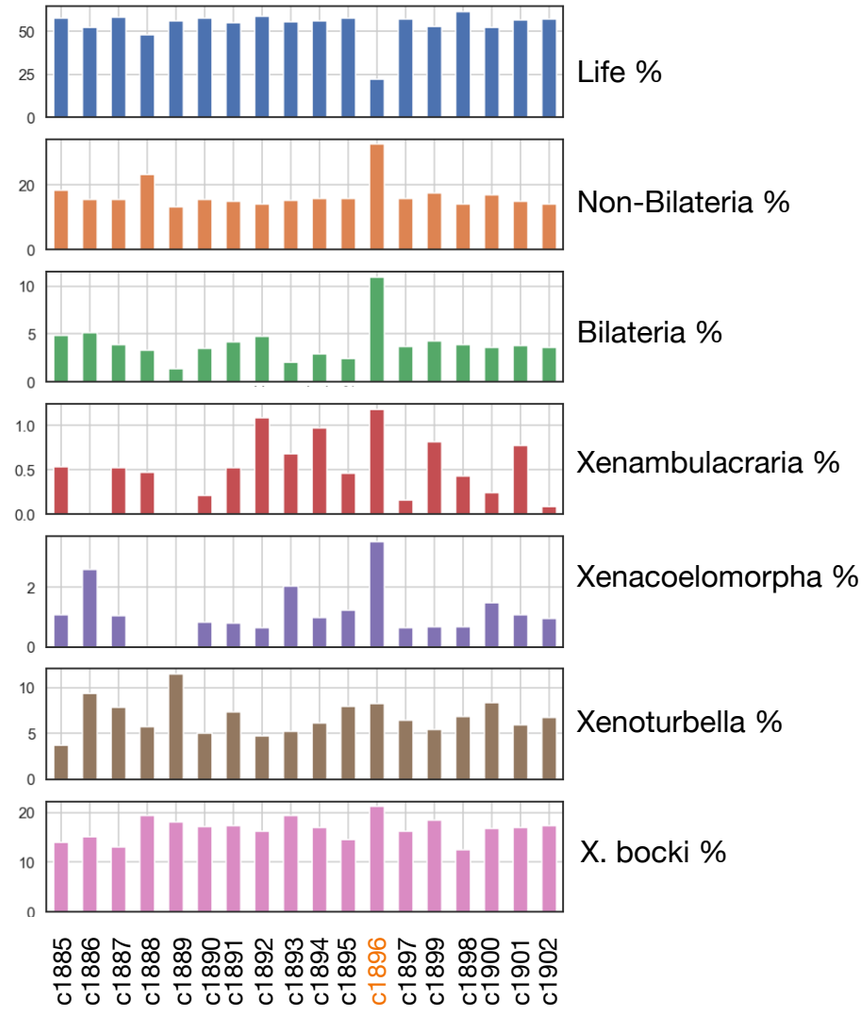


Figure 5: (a) A comparison of scaffolds in the *X. bocki* genome with other Metazoa. 17 of the 18 large scaffolds in the *X. bocki* genome are linked via synteny to distinct chromosomal scaffolds in these species. (b) Phylostratigraphic age distribution of genes on all major scaffolds in the *X. bocki* genome. One scaffold (c1896), which showed no synteny to a distinct chromosomal scaffold in the other metazoan species also had a divergent gene age structure in comparison to other *X. bocki* scaffolds.

365 We also attempted to detect some pre-bilaterian arrangement of ancestral
366 linkage: for instance, ref [40] predicted that several pre-bilaterian linkage groups
367 successively fused in the bilaterian lineage to give ALGs A1, Q and E. These are all
368 represented as single units in *X. bocki* in common with other Bilateria. None of the
369 inferred pre-bilaterian chromosomal arrangements that could have provided support
370 for the Nephrozoa hypothesis were found *in X. bocki* although of course this does not
371 rule out Nephrozoa.
372
373 <u>One *X. bocki* chromosomal fragment appears aberrant</u>
374 The smallest of the 18 large scaffolds in the *X. bocki* genome did not show strong 1:1
375 clustering with any scaffold/chromosome of the bilaterian species we compared it to.
376 To exclude potential contamination in the assembly as a source for this contig we
377 examined the orthogroups to which the genes from this scaffold belong. We found that
378 *Xenoturbella profunda*[41], for which a transcriptome is available, was the species that
379 most often occurred in the same orthogroup with genes from this scaffold (41 shared
380 orthogroups), suggesting the scaffold is not a contaminant.
381 We did observe links between the aberrant scaffold and several scaffolds from
382 the genome of the sponge *E. muelleri* in regard to synteny, but could not detect distinct
383 synteny relationships to a single scaffold in another species. In line with this, genes
384 on the scaffold show a different age structure compared to other scaffolds, with both
385 more older genes (pre bilaterian) and more *Xenoturbella* specific genes (Fig. 5b;
386 supported by Ks statistics, Supplementary). This aberrant scaffold also had
387 significantly lower levels of methylation than the rest of the genome (Supplementary).
388
389 **DISCUSSION**
390 The phylogenetic positions of *Xenoturbella* and the Acoelomorpha have been
391 controversial since the first molecular data from these species appeared over
392 twentyfive years ago. Today we understand that they constitute a monophyletic group
393 of morphologically simple worms[1,9,42], but there remains a disagreement over whether
394 they represent a secondarily simplified sister group of the Ambulacraria or a primitively
395 simple sister group to all other Bilateria.
396 Previous analyses of the content of genomes, especially of Acoela, have been
397 used to bolster the latter view, with the small number of Hox genes and of microRNAs

398 of acoels being interpreted as representing an intermediate stage on the path to the
399 ~8 Hox genes and 30 odd microRNAs of the Nephrozoa. A strong version of the
400 Nephrozoa idea would go further than these examples and anticipate, for example, a
401 genome-wide paucity of bilaterian genes, GRNs and biochemical pathways and/or an
402 arrangement of chromosomal segments intermediate between those of the
403 Eumetazoa and the Nephrozoa.

404 One criticism of the results from analyses of acoel genomes is that the
405 Acoelomorpha have evolved rapidly (their long branches in phylogenetic trees
406 showing high rates of sequence change). This rapid evolution might plausibly be
407 expected to correlate with other aspects of rapid genome evolution such as higher
408 rates of gene loss and chromosomal rearrangements leading to significant differences
409 from other Bilateria. The more normal rates of sequence evolution observed in
410 *Xenoturbella* therefore recommend it as a more appropriate xenacoelomorph to study
411 with fewer apomorphic characters expected.

412 We have sequenced, assembled, and analysed a draft genome of *Xenoturbella*
413 *bocki*. To help with annotation of the genome we have also sequenced miRNAs and
414 small RNAs as well as using bisulphite sequencing, Hi-C and Oxford nanopore. We
415 compared the gene content of the *Xenoturbella* genome to species across the
416 Metazoa and its genome structure to several other high-quality draft animal genomes.

417 We found the *X. bocki* genome to be fairly compact, but not unusually reduced
418 in size compared to many other bilaterians. It appears to contain a similar number of
419 genes (~15,000) as other animals, for example from the model organisms *D.*
420 *melanogaster* (>14,000) and *C. elegans (*~20,000). The BUSCO completeness, as
421 well as a high level of representation of *X. bocki* proteins in the orthogroups of our 155
422 species orthology screen indicates that we have annotated a near complete gene set.
423 Surprisingly, there are fewer genes than in the acoel *Hofstenia* (>22,000; BUSCO_v5
424 score ~95%). This said, of the genes found in Urbilateria (orthogroups in our
425 presence/absence analysis containing a member from both a bilaterian and an
426 outgroup) *Xenoturbella* and *Hofstenia* have very similar numbers (5459 and 5438
427 respectively). Gene, intron and exon lengths all also fall within the range seen in many
428 other invertebrate species[25]. It thus seems that basic genomic features in *Xenoturbella*
429 are not anomalous among Bilateria. Unlike some extremely simplified animals, such

430 as orthonectids, we observe no extreme reduction in gene content.

431 All classes of homeodomain transcription factors have previously been reported
432 to exist in Xenacoelomorpha[43]. We have identified 5 HOX-genes in *X. bocki* and at
433 least four, and probably all five of these are on one chromosomal scaffold within 187
434 Kbp. *X.bocki* also has the parahox genes Gsx and Cdx; while Xlox/pdx is not found, it
435 is present in Cnidarians and must therefore have been lost[44]. If block duplication
436 models of Hox and Parahox evolutionary relationships are correct, the presence of a
437 complete set of parahox genes implies the existence of their Hox paralogs in the
438 ancestor of Xenacoelomorphs suggesting the xenacoelomorph ancestor also
439 possessed a Hox 3 ortholog. If anthozoans also have an ortholog of bilaterian Hox 2[45],
440 this must also have been lost from Xenacoelomorphs. The minimal number of Hox
441 genes in the xenacoelomorph stem lineage was therefore probably 7 (AntHox1, lost
442 Hox2, lost Hox 3, CentHox 1, CentHox 2, CentHox 3 and postHoxP).

443 Based on early sequencing technology and without a reference genome
444 available, it was thought that Acoelomorpha lack many bilaterian microRNAs. Using
445 deep sequencing of small RNAs and our high-quality genome, we have shown that
446 *Xenoturbella* shows a near-complete bilaterian set of miRNAs including the single
447 deuterostome-specific miRNA family (MIR-103) (Figure X). The low number of
448 differential family losses of *Xenoturbella* (8 of 31 bilaterian miRNA families) inferred is
449 the same as the number lostin the flatworm *Schmidtea,* and substantially lower than
450 the number lost in the rotifer *Brachionus* (which has lost 14 bilaterian families). It is
451 worth mentioning that *X. bocki* shares the absence of one miRNA family (MIR-216)
452 with all Ambulacrarians although if Deuterostomia are paraphyletic this could be
453 interpretable as a primitive state[35].

454 The last decade has seen a re-evaluation of our understanding of the evolution
455 of the neuropeptide signaling genes[46,47]. The peptidergic systems are thought to have
456 undergone a diversification that produced approximately 30 systems in the bilaterian
457 common ancestor[46,47]. Our study identified 31 neuropeptide systems in *X.bocki* and
458 for all of these either the ligand, receptor, or both are also present in both protostomes
459 and deuterostomes indicating conservation across Bilateria. It is likely that more
460 ligands (which are short and variable) remain to be found with better detection
461 methods. It appears that the *Xenoturbella* genome contains a nearly complete

462 bilaterian neuropeptide complement with no signs of simplification but rather signs of
463 expansions of certain gene families. Our analyses also reveal a potential
464 synapomorphy linking Xenacoelomorpha with Ambulacraria (Fig 4 and
465 Supplementary).
466 We have used the predicted presence and absence of genes across a selection
467 of metazoan genomes as characters for phylogenetic analyses. Our trees re-confirm
468 the findings of recent phylogenomic gene alignment studies in linking *Xenoturbella* to
469 the Ambulacraria. We also used these data to test different bilaterians for their
470 propensity to lose otherwise conserved genes (or for our inability to identify
471 orthologs[30]). While the degree of gene loss appears similar between *Xenoturbella* and
472 acoels, the phylogenetic analysis shows longer branches leading to the acoels, most
473 likely due to faster evolution and some degree of gene loss in the branch leading to
474 the Acoelomorpha. Recent work has shown the tendency of rapidly evolving genes
475 (such as those belonging to rapidly evolving species) to be missed by orthology
476 detection software[48,49].
477
478 This pattern of conservation of evolutionarily old parts of the Metazoan genome
479 is further reinforced by the retention in *Xenoturbella* of linkage groups present from
480 sponges to vertebrates. It is interesting to note that*X. bocki* does not follow the pattern
481 of seen in other morphologically simplified animals such as nematodes and
482 platyhelminths, which have lost and/or fused these ancestral linkage groups. We
483 interpret this to be a signal of comparably slower genomic evolution in *Xenoturbella* in
484 comparison to some other bilaterian lineages. The fragmented genome sequence of
485 *Hofstenia* prevents us from asking whether the ancient linkage groups have also been
486 preserved in the Acoelomorpha.
487 One of the chromosome-scale scaffolds in our assembly showed a different
488 methylation and age signal, with younger genes, and no clear relationship to metazoan
489 linkage groups. By analyzing orthogroups of genes on this scaffold for their
490 phylogenetic signal and finding *X. bocki* genes to cluster with those of *X. profunda* we
491 concluded that the scaffold most likely does not represent a contamination. It remains
492 unclear whether this scaffold is a fast-evolving chromosome, or a chromosomal
493 fragment or arm. Very fast evolution on a chromosomal arm has for example been

494 shown in the zebrafish[50].

495 Apart from DNA from *X. bocki* we also obtained a highly contiguous genome of
496 a species related to marine *Chlamydia* species (known from microscopy to exist in *X.*
497 *bocki)*; a symbiotic relationship between *Xenoturbella* and the bacteria has been
498 thought possible[51]. The large gene number and the completeness of genetic pathways
499 we found in the chlamydial genome do not support an endosymbiotic relationship.

500 Overall, we have shown that, while *Xenoturbella* has lost some genes - in
501 addition to the reduced number of Hox genes previously noted, we observe a reduction
502 of some signaling pathways to the core components - in general, , the *X. bocki* genome
503 is not strikingly simpler than many other bilaterian genomes. We do not find
504 support for a strong version of the Nephrozoa hypothesis which would predict many
505 missing bilaterian genes. Bilaterian Hox and microRNA absent from Acoelomorpha
506 are found in *Xenoturbella* reducing the impact of two character types that were
507 previously cited in support of Nephrozoa. The *Xenoturbella* genome has also largely
508 retained the ancestral linkage groups found in other bilaterians and does not represent
509 a structure intermediate between Eumetazoan and bilaterian ground states. Overall,
510 while we can rule out a strong version of the Nephrozoa hypothesis with many
511 Bilaterian characteristics missing in xenacoelomorphs, our analysis of the
512 *Xenoturbella* genome cannot distinguish between a weak version of Nephrozoa and
513 the Xenambulacraria topology; nevertheless, our phylogenetic analysis of gene
514 presence and absence supports the latter.

515

516 # Methods

517 **Genome Sequencing, Assembly, and Scaffolding**

518 We extracted DNA from individual *Xenoturbella* specimens with a standard and
519 additionally worked with a Phenol-Chloroform protocol specifically developed to
520 extract HMW DNA (dx.doi.org/10.17504/protocols.io.mrxc57n). The extracted DNA
521 was quality controlled with a Nanodrop instrument in our laboratory and subsequently
522 a TapeStation at the sequencing center. Worms were first starved and kept in
523 repeatedly replaced salt water, reducing the likelihood of food or other contaminants
524 in the DNA extractions. First, we sequenced Illumina short paired-end reads and mate
525 pair libraries (see ref [3] for details). As the initial paired-read datasets were of low

526 complexity and coverage, we later complemented these data with an Illumina HiSeq
527 2000/2500 series paired-end dataset with ~700 bp insert size and 250bp read lengths,
528 yielding ~354 Million reads. Additionally, we generated ~40 Million
529 Illumina TruSeq Synthetic Long Reads (TSLR) for high confidence primary
530 scaffolding.

531 After read cleaning with Trimmomatic v.0.38[52] we conducted initial test assemblies
532 using the clc assembly cell v.5 and ran the blobtools pipeline[53] to screen for
533 contamination (Supplementary). Not detecting any significant numbers of reads from
534 suspicious sources in the HiSeq dataset we used SPAdes v. 3.9.0[20] to correct and
535 assemble a first draft genome. We also tried to use dipSPAdes but found the runtime
536 to exceed several weeks without finishing. We submitted the SPAdes assembly to the
537 redundans pipeline to eliminate duplicate contigs and to scaffold with all available mate
538 pair libraries. The resulting assembly was then further scaffolded with the aid of
539 assembled transcripts (see below) in the BADGER pipeline[54]. In this way we were able
540 to obtain a draft genome with ~60kb N50 that could be scaffolded to chromosome
541 scale super-scaffolds with the use of 3C data.

542 We also used two remaining specimens to extract HMW DNA for Oxford Nanopore
543 PromethION sequencing in collaboration with the Loman laboratory in Birmingham.
544 Unfortunately, the extraction failed for one individual with the DNA appearing to be
545 contaminated with a dark coloured residue. We were able to prepare a ligation and a
546 PCR library for DNA from the second specimen and obtain some genomic data.
547 However, due to pore blockage on both flow cells the combined data amounted to only
548 about 0.5-fold coverage of the genome and was thus not useful in scaffolding. We
549 suspect that the dark colouration of the DNA indicates a natural modification to be
550 present in *X. bocki* DNA that inhibits sequencing with the Oxford Nanopore method.

551 Library preparation for genome-wide bisulfite sequencing was performed as
552 previously described[55]. The resulting sequencing data were aligned to the *X. bocki*
553 draft genome using Bismark in non-directional mode to identify the percentage
554 methylation at each cytosine genome-wide. Only sites with >10 reads mapping were
555 considered for further analysis.

556

557

558 Preparation of the Hi-C libraries

559 The Hi-C protocol was adapted at the time from (Lieberman-Aiden et al., 2009; Sexton
560 et al., 2012 and Marie-Nelly et al., 2014). Briefly, an animal was chemically cross-
561 linked for one hour at room temperature in 30 mL of PBS 1X added with 3%
562 formaldehyde (Sigma – F8775 - 4x25 mL). Formaldehyde was quenched for 20 min
563 at RT by adding 10 ml of 2.5 M glycine. The fixed animal was recovered through
564 centrifugation and stored at -80°C until use. To prepare the proximity ligation library,
565 the animal was transferred to a VK05 Precellys tubes in 1X DpnII buffer (New England
566 Biolabs; 0.5mL) and the tissues were disrupted using the Precellys Evolution
567 homogenizer (Bertin-Instrument). SDS was added (0.3% final) to the lysate and the
568 tubes were incubated at 65°C for 20 minutes followed by an incubation at 37°C for 30
569 minutes and an incubation of 30 minutes after adding 50 µL of 20% triton-X100. 150
570 units of the DpnII restriction enzyme were then added and the tubes were incubated
571 overnight at 37°C. The endonuclease was inactivated 20 min at 65°C and the tubes
572 were then centrifuged at 16,000 x g during 20 minutes, supernatant was discarded
573 and pellets were re-suspended in 200 µl NE2 1X buffer and pooled. DNA ends were
574 labeled using 50 µl NE2 10X buffer, 37.5 µl 0.4 mM dCTP-14-biotin, 4.5 µl 10mM
575 dATP-dGTP-dTTP mix, 10 µl klenow 5 U/µL and incubation at 37°C for 45 minutes.
576 The labeling mix was then transferred to ligation reaction tubes (1.6 ml ligation buffer;
577 160 µl ATP 100 mM; 160 µl BSA 10 mg/mL; 50 µl T4 DNA ligase (New England
578 Biolabs, 5U/µl); 13.8 ml H2O) and incubated at 16°C for 4 hours. A proteinase K mix
579 was added to each tube and incubated overnight at 65°C. DNA was then extracted,
580 purified and processed for sequencing as previously described[22]. Hi-C libraries were
581 sequenced on a NextSeq 500 (2 × 75 bp, paired-end using custom made
582 oligonucleotides as in Marie-Nelly et al., 2014). Libraries were prepared separately on
583 two individuals in this way but eventually merged. Note that more recent version of the
584 HI-C protocol than the one used here have been described elsewhere[56].

585

586 InstaGRAAL assembly pre-processing

587 The primary Illumina assembly contains a number of very short contigs, which are
588 disruptive when computing the contact distribution needed for the instaGRAAL
589 proximity ligation scaffolding (pre-release version, see[57] and[22] for details). Testing

590 several Nx metrics we found a relative length threshold, that depends on the scaffolds'
591 length distribution, to be a good compromise between the need for a low-noise contact
592 distribution and the aim of connecting most of the genome. We found N90 a suitable
593 threshold and excluded contigs below 1,308 bp. This also ensured no scaffolds shorter
594 than three times the average length of a DpnII restriction fragment (RF) were in the
595 assembly. In this way every contig contained enough RFs for binning and were
596 included in the scaffolding step.

597 Reads from both libraries were aligned with bowtie2 (v. 2.2.5)[58] against the
598 DpnII RFs of the reference assembly using the hicstuff pipeline
599 (https://github.com/koszullab/hicstuff) and in paired-end mode (with the options: -fg-
600 maxins 5 -fg-very-sensitive-local), with a mapping quality >30. The pre-processed
601 genome was reassembled using instaGRAAL. Briefly, the program uses a Markov
602 Chain Monte Carlo (MCMC) method that samples DNA segments (or bins) of the
603 assembly for their best relative 1D positions with respect to each other. The quality of
604 the positions is assessed by fitting the contact data first on a simple polymer model,
605 then on the plot of contact frequency according to the genomic distance law computed
606 from the data. The best relative position of a DNA segment with respect to one of its
607 most likely neighbours consists in operations such as flips, swaps, merges or a split
608 of contigs. Each operation is either accepted or rejected based on the computed
609 likelihood, resulting in an iterative progression toward the 1D structure that best fits
610 the contact data. Once the entire set of DNA segments is sampled for position (i.e. a
611 cycle), the process starts over. The scaffolder was run independently for 50 cycles,
612 long enough for the chromosome structure to converge. The corresponding genome
613 is then considered stable and suitable for further analyses. The scaffolded assemblies
614 were then refined using instaGRAAL's instaPolish module, to correct small artefactual
615 inversions that are sometimes a byproduct of instaGRAAL's processing.

616

617 **Genome Annotation**

618 Transcriptome Sequencing

619 We extracted total RNA from a single *X. bocki* individual and sequenced a strand
620 specific Illumina paired end library. Extraction of total RNA was performed using a
621 modified Trizol & RNeasy hybrid protocol for which tissue had to be stored in RNAlater.

622 cDNA transcription reaction/cDNA synthesis was done using the RETROscript kit
623 (Ambion) using both Oligo(dT) and Random Decamer primers. Detailed extraction and
624 transcription protocols are available from the corresponding authors. The resulting
625 transcriptomic reads were assembled with the Trinity pipeline[59,60] into 103,056
626 sequences (N50: 705; BUSCO_v5 Eukaryota scores: C:65.1%, [S:34.1%, D:31.0%],
627 F:22.0%, M:12.9%) for initial control and then supplied to the genome annotation
628 pipeline (below).
629
630 Repeat annotation
631 In the absence of a repeat library for Xenoturbellida we first used RepeatModeller
632 v. 1.73 to establish a library *de novo*. We then used RepeatMasker v. 4.1.0
633 (https://www.repeatmasker.org) and the Dfam library[61,62] to soft-mask the genome.
634 We mapped the repeats to the instaGRAAL scaffolded genome with RepeatMasker.
635
636 Gene prediction and annotation
637 We predicted genes using Augustus[63] implemented into the BRAKER (v.2.1.0)
638 pipeline[23,24] to incorporate the RNA-Seq data. BRAKER uses spliced aligned RNA-
639 Seq reads to improve training accuracy of the gene finder GeneMark-ET[64].
640 Subsequently, a highly reliable gene set predicted by GeneMark-ET in *ab initio* mode
641 was selected to train the gene finder AUGUSTUS, which in a final step predicted
642 genes with evidence from spliced aligned RNA-Seq reads. To make use of additional
643 single cell transcriptome data allowing for a more precise prediction of 3'-UTRs we
644 employed a production version of BRAKER (August 2018 snapshot). We had
645 previously mapped the RNA-Seq data to the genome with gmap-gsnap v. 2018-07-
646 04[65] and used samtools[66] and bamtools[67] to create the necessary input files. This
647 process was repeated in an iterative way, visually validating gene structures and
648 comparing with mappings loci inferred from a set of single-cell RNA-Seq data
649 (published elsewhere, see: [68]) in particular regarding fused genes. Completeness of
650 the gene predictions was independently assessed with BUSCO_v5[27] setting the
651 eukaryote dataset as reference on gVolante[69]. We used InterProScan v. 5.27-66.0
652 standalone[70,71] on the UCL cluster to annotate the predicted *X. bocki* proteins with
653 Pfam, SUPERFAM, PANTHER, and Gene3D information.

654

655 Horizontal Gene Transfer

656 To detect horizontally acquired genes in the *X. bocki* genome we used a pipeline
657 available from (https://github.com/reubwn/hgt). Briefly, this uses blasts against the
658 NCBI database, alignments with MAFFT[72], and phylogenetic inferences with
659 IQTree[73,74] to infer most likely horizontally acquired genes, while trying to discard
660 contamination (e.g. from co-sequenced gut microbiota).

661

662 Orthology inference

663 We included 155 metazoan species and outgroups into our orthology analysis. We
664 either downloaded available proteomes or sourced RNA-Seq reads from online
665 repositories to then use Trinity v 2.8.5 and Trinnotate v. 3.2.0 to predict protein sets.
666 In the latter case we implemented diamond v. 2.0.0 blast[75,76] searches against UniProt
667 and Pfam[77] hmm screens against the Pfam-A dataset into the prediction process. We
668 had initially acquired 185 datasets, but excluded some based on inferior BUSCO
669 completeness, while at the same time aimed to span as many phyla as possible.
670 Orthology was then inferred using Orthofinder v. 2.2.7[78,79], again with diamond as the
671 blast engine.

672 Using InterProScan v. 5.27-66.0 standalone on all proteomes we added
673 functional annotation and then employed kinfin[33] to summarise and analyse the
674 orthology tables. For the kinfin analysis, we tested different query systems in regard
675 to phylogenetic groupings (Supplementary).

676 To screen for inflation and contraction of gene families we first employed
677 CAFE5[80], but found the analysis to suffer from long branches and sparse taxon
678 sampling in Xenambulacraria. We thus chose to query individual gene families (e.g.
679 transcription factors) by looking up Pfam annotations in the InterProScan tables of
680 high-quality genomes in our analysis.

681 Through the GenomeMaple online platform we calculated completeness of
682 signaling pathways within the KEGG database using GhostX as the search engine.

683

684 Presence/absence phylogenetics

685 We used a database of metazoan proteins, updated from ref [81], as the basis for

686 an OMA analysis to calculate orthologous groups, performing two separate runs, one
687 including *Xenoturbella* and acoels, and one with only *Xenoturbella*. We converted
688 OMA gene OrthologousMatrix.txt files into binary gene presence absence matrices in
689 Nexus format with datatype = restriction. We calculated phylogenetic trees on these
690 matrices using RevBayes (see https://github.com/willpett/metazoa-gene-content for
691 RevBayes script), as described in ref 74 with corrections for no absent sites
692 and no singleton presence, using the reversible, not the Dollo model,
693 as it is more likely to be able to correct for noise related to
694 prediction errors [82,78]. For each matrix, two runs were performed and compared and
695 consensus trees generated with bpcomp from Phylobayes[83].
696
697 *Hox and ParaHox gene cluster identification and characterisation*
698 Previous work has already used transcriptomic data and phylogenetic inference
699 to identify the homeobox repertoire in *Xenoturbella bocki*. These annotations were
700 used to identify genomic positions and gene annotations that correspond to Hox and
701 ParaHox clusters in *X. bocki*. Protein sequences of homeodomains (Evx, Cdx, Gsx,
702 antHox1, centHox1, centHox2, cent3 and postHoxP) were used as TBLASTN queries
703 to identify putative scaffolds associated with Hox and ParaHox clusters. Gene models
704 from these scaffolds were compared to the full length annotated homeobox transcripts
705 from[84] using BLASTP, using hits over 95% identity for homeobox classification. There
706 were some possible homeodomain containing genes on the scaffolds that were not
707 previously characterised and were therefore not given an annotation.
708 There were issues concerning the assignment of postHoxP and Evx to gene
709 models. To ascertain possible CDS regions for these genes, RNA-Seq reads were
710 mapped with HISAT2 to the scaffold and to previous annotation[84],were assembled
711 with Trinity and these were combined with BRAKER annotations.
712 Some issues were also observed with homeodomain queries matching genomic
713 sequences that were identical, suggesting artefactual duplications. To investigate
714 contiguity around genes the ONT reads were aligned with Minimap2 to capture long
715 reads over regions and coverage.
716
717

718 Small RNA Sequencing and Analysis

719 Two samples of starved worms were subjected to 5' monophosphate dependent
720 sequencing of RNAs between 15 and 36 nucleotides in length, according to previously
721 described methods[85]. Using miRTrace[86] 3.3, 18.6 million high-quality reads were
722 extracted and merged with the 27 635 high quality 454 sequencing reads from Philippe
723 et al. The genome sequence was screened for conserved miRNA precursors using
724 MirMachine[87] followed by a MirMiner run that used predicted precursors and
725 processed and merged reads on the genome[88]. Outputs of MirMachine and MirMiner
726 were manually curated using a uniform system for the annotation of miRNA genes[89]
727 and by comparing to MirGeneDB[90].

728

729 Neuropeptide prediction and screen

730 Neuropeptide prediction was conducted on the full set of *X.bocki* predicted
731 proteins using two strategies to detect neuropeptide sequence signatures. First, using
732 a custom script detecting the occurrence of repeated sequence patterns:
733 RRx(3,36)RRx(3,36)RRx(3,36)RR,RRx(2,35)ZRRx(2,35)ZRR,
734 RRx(2,35)GRRx(2,35)GRR, RRx(1,34)ZGRRx(1,34)ZGRR where R=K or R; x=any
735 amino acid; Z=any amino acid but repeated within the pattern. Second, using
736 HMMER3.1[91] (hmmer.org), and a combination of neuropeptide HMM models obtained
737 from the PFAM database (pfam.xfam.org) as well as a set of custom HMM models
738 derived from alignment of curated sets of neuropeptide sequences[46,47,92]. Sequences
739 retrieved using both methods and comprising fewer than 600 amino acids were further
740 validated. First, by blast analysis: sequences with E-Value ratio "best blast hit versus
741 ncbi nr database/best blast hit versus curated neuropeptide dataset" < 1e-40 were
742 discarded. Second by reciprocal best blast hit clustering using Clans[93]
743 (eb.tuebingen.mpg.de/protein-evolution/software/clans/) with a set of curated
744 neuropeptide sequences[46]. SignalP-5.0[94] (cbs.dtu.dk/services/SignalP/) was used to
745 detect the presence of a signal peptide in the curated list of predicted neuropeptide
746 sequences while Neuropred[95] (stagbeetle.animal.uiuc.edu/cgi-bin/neuropred.py) was
747 used to detect cleavage sites and post-translational modifications. Sequence
748 homology of the predicted sequence with known groups was analysed using a
749 combination of (i) blast sequence similarity with known bilaterian neuropeptide

750 sequences, (ii) reciprocal best blast hit clustering using Clans and sets of curated
751 neuropeptide sequences, (iii) phylogeny using MAFFT
752 (mafft.cbrc.jp/alignment/server/), TrimAl[96] (trimal.cgenomics.org/) and IQ-TREE[97]
753 webserver for alignment, trimming and phylogeny inference respectively. Bilaterian
754 prokineticin-like sequences were searched in ncbi nucleotide, EST and SRA
755 databases as well as in the *Saccoglossus kowalevskii* genome assembly[74,98]
756 (groups.oist.jp/molgenu) using various bilaterian prokineticin-related protein
757 sequences as query. Sequences used for alignments shown in figures were collected
758 from ncbi nucleotide and protein databases as well as from the following publications:
759 7B2[46]; NucB2[92]; Insulin[99]; Prokineticin[37,38,100]. Alignments for figures were created with
760 Jalview (jalview.org).
761
762 Neuropeptide receptor search
763 §Neuropeptide Receptor sequences for Rhodopsin type GPCR, Secretin type GPCR
764 and tyrosine and serine/threonine kinase receptors were searched by running
765 HMMER3.1 on the full set of *X.bocki* predicted proteins using the 7tm_1 (PF00001),
766 7tm_2 (PF00002) and PK_Tyr_Ser-Thr (PF07714) HMM models respectively which
767 were obtained from the PFAM database (pfam.xfam.org). Sequences above the
768 significance threshold were then aligned with sequences from the curated dataset,
769 trimmed and phylogeny inference was conducted using same method as for the
770 neuropeptide. A second alignment and phylogeny inference was conducted after
771 removal of all *X.bocki* sequences having no statistical support for grouping with any of
772 the known neuropeptide receptors from the curated dataset. Curated datasets were
773 collected from the following publications: Rhodopsin type GPCR beta and gamma and
774 Secretin type GPCR[100]; Rhodopsin type GPCR delta (Leucine-rich repeat-containing
775 G-protein coupled Receptors)[101]; Tyrosine kinase receptors[102,103]; and were
776 complemented with sequences from NCBI protein database.
777
778 Synteny
779 Ancestral linkage analyses rely on mutual-best-hits computed using Mmseqs2[104]
780 between pairs of species in which chromosomal assignments to ancestral linkage
781 groups (ALG) was previously performed, such as *Branchiostoma floridae* or *Pecten*

782 *maximus*[39]. Oxford dotplots were computed by plotting reciprocal positions of indexed
783 pairwise orthologs between two species as performed previously[39,40]. The significance
784 of ortholog enrichment in pairs of chromosomes was assessed using a fisher test.
785 We also used a Python implementation of MCscanX[105] (Haibao Tang and available
786 on https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)) to compare *X.*
787 *bocki* to *Euphydtia muelleri*, *Trichoplax adhearens*, *Branchiostoma floridae*,
788 *Saccoglossus kowalevskii*, *Ciona intestinalis*, *Nematostella vectensis*, *Asteria rubens,*
789 *Pecten maximus*, *Nemopilema nomurai, Carcinoscorpius rotundicauda* (see
790 Supplementary). Briefly, the pipeline uses high quality genomes and their annotations
791 to infer syntenic blocks based on proximity. For this an all vs. all blastp is performed
792 and synteny extended from anchors identified in this way. Corresponding heatmaps
793 (see Supplementary) were plotted with Python in a Jupyter notebooks instance.
794
795 *Chlamydia* assembly and annotation
796 We identified a highly contiguous *Chlamydia* genome in the *X. bocki* genome
797 assembly using blast. We then used our Oxford Nanopore derived long-reads to
798 scaffold the *Chlamyida* genome with LINKS[106] and annotated it with the automated
799 PROKKA pipeline. To place the genome on the *Chlamydia* tree we extracted the 16S
800 ribosomal RNA gene sequence, aligned it with set of *Chlamydia* 16S rRNA sequences
801 from[28] using MAFFT, and reconstructed the phylogeny using IQ-TREE 2[73] We
802 visualized the resulting tree with Figtree (http://tree.bio.ed.ac.uk/).
803
804
805 **Acknowledgements**
806 We thank Josh Quick and Nick Loman for help with the generation of ONT long-read
807 data. Analyses were conducted mainly on the UCL Cluster, with some computations
808 also run on the CHEOPS cluster at the University of Cologne. We are grateful to Kevin
809 J. Peterson for his comments on the manuscript, the miRNA section in particular. We
810 thank the Kristineberg Center for Marine Research and Innovation for their essential
811 support in sampling *Xenoturbella*.
812
813 **Conflict of interest**
814 The authors declare no conflict of interest.
815
816

Data availability

All read sets (RNA and DNA derived) used in this study will be made available with the publication of this manuscript on the SRA database under the BioProject ID PRJNA864813. Hi-C reads are deposited under SAMN30224387, RNA-Seq under SAMN35083895. The genome assemblies of *X. bocki* (ERS12565994, ERA16814408) and the *Chlamydia* sp. (ERS12566084, ERA16814775) are deposited under PRJEB55230 at ENA.

Funding

PHS was funded by an ERC grant (ERC-2012-AdG 322790) to MJT, which also supported HR, ACZ, SM. PHS was also funded through an Emmy-Noether grant (434028868) to himself. Part of this work was funded by BBSRC grant BB/R016240/1 (M.J.T./P.K.), by a Leverhulme Trust Research Project Grant RPG-2018-302 (M.J.T./D.J.L.), and by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no 764840 IGNITE (M.J.T./P.N.).

References

1. Telford, M. J. Xenoturbellida: the fourth deuterostome phylum and the diet of worms. *Genesis (New York, N.Y. : 2000)* 46, 580–586 (2008).

2. Westblad, E. *Xenoturbella bocki* n. g., n. sp., a peculiar, primitive Turbellarian type. *Arkiv för Zoologi* 3–29 (1949).

3. Philippe, H. *et al.* Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria. *Current Biology* 29, 1818-1826.e6 (2019).

4. Cannon, J. T. *et al.* Xenacoelomorpha is the sister group to Nephrozoa. *Nature* 530, 89–93 (2016).

5. Ueki, T., Arimoto, A., Tagawa, K. & Satoh, N. Xenacoelomorph-Specific Hox Peptides: Insights into the Phylogeny of Acoels, Nemertodermatids, and Xenoturbellids. *Zool Sci* 36, 395–401 (2019).

6. Hejnol, A. *et al.* Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings. Biological sciences / The Royal Society* 276, 4261–4270 (2009).

7. Rouse, G. W., Wilson, N. G., Carvajal, J. I. & Vrijenhoek, R. C. New deep-sea species of *Xenoturbella* and the position of Xenacoelomorpha. *Nature* 530, 94–97 (2016).

8. Srivastava, M., Mazza-Curll, K. L., Wolfswinkel, J. C. van & Reddien, P. W. Whole-body acoel regeneration is controlled by Wnt and Bmp-Admp signaling. *Current Biology* 24, 1107–1113 (2014).

9. Philippe, H. *et al.* Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature* 470, 255–258 (2011).

10. Bourlat, S. J. *et al.* Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* 444, 85–88 (2006).

11. Nakano, H. What is *Xenoturbella*? *Zoological Letters* 1, 1 (2015).

12. Hejnol, A. & Martindale, M. Q. Acoel development supports a simple planula-like urbilaterian. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 363, 1493–1501 (2008).

13. Martynov, A. *et al.* Multiple paedomorphic lineages of soft-substrate burrowing invertebrates: parallels in the origin of *Xenocratena* and *Xenoturbella*. *Plos One* 15, e0227173 (2020).

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

867 15. Sempere, L. F., Cole, C. N., McPeek, M. A. & Peterson, K. J. The phylogenetic
868 distribution of metazoan microRNAs: insights into evolutionary complexity and constraint.
869 306, 575–588 (2006).

870 16. Gehrke, A. R. *et al.* Acoel genome reveals the regulatory landscape of whole-body
871 regeneration. *Science* 363, 1–9 (2019).

872 17. Arimoto, A. *et al.* A draft nuclear-genome assembly of the acoel flatworm *Praesagittifera*
873 *naikaiensis*. *Gigascience* 8, giz023 (2019).

874 18. Martinez, P. *et al.* Genome assembly of the acoel flatworm *Symsagittifera roscoffensis*, a
875 model for research on body plan evolution and photosymbiosis. *G3 Genes Genomes Genetics*
876 13, jkac336 (2022).

877 19. Moroz, L. L., Romanova, D. Y. & Kohn, A. B. Neural versus alternative integrative
878 systems: molecular insights into origins of neurotransmitters. *Philosophical Transactions*
879 *Royal Soc Lond Ser B Biological Sci* 376, 20190762 (2021).

880 20. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to
881 single-cell sequencing. *www.liebertpub.com* 19, 455–477 (2012).

882 21. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous
883 genomes. *Nucleic Acids Research* 44, e113–e113 (2016).

884 22. Baudry, L. *et al.* instaGRAAL: chromosome-level quality scaffolding of genomes using a
885 proximity ligation-based scaffolder. *Genome Biol* 21, 148 (2020).

886 23. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome annotation with
887 BRAKER. *Methods Mol Biology Clifton N J* 1962, 65–95 (2019).

888 24. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1:
889 Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS.
890 *Bioinformatics* 32, 767–769 (2016).

891 25. Francis, W. R. & Wörheide, G. Similar ratios of introns to intergenic sequence across
892 animal genomes. *Genome Biol Evol* 9, evx103- (2017).

893 26. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinform Oxf Engl* 30, 2068–
894 9 (2014).

895 27. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M.
896 BUSCO: assessing genome assembly and annotation completeness with single-copy
897 orthologs. 31, 3210–3212 (2015).

898 28. Dharamshi, J. E. *et al.* Marine sediments illuminate Chlamydiae diversity and evolution.
899 *Curr Biol* 30, 1032-1048.e7 (2020).

900 29. Kjeldsen, K. U., Obst, M., Nakano, H., Funch, P. & Schramm, A. Two types of
901 endosymbiotic bacteria in the enigmatic marine worm *Xenoturbella bocki*. 76, 2657–2662
902 (2010).

903 30. Natsidis, P., Kapli, P., Schiffer, P. H. & Telford, M. J. Systematic errors in orthology
904 inference and their effects on evolutionary analyses. *Iscience* 102110 (2021).

905 31. Schiffer, P. H., Robertson, H. E. & Telford, M. J. Orthonectids are highly degenerate
906 annelid worms. *Current Biology* 1–9 (2018).

907 32. Mikhailov, K. V. *et al.* The genome of *Intoshia linei* affirms orthonectids as highly
908 simplified spiralians. *Current Biology* 26, 1768–1774 (2016).

909 33. Laetsch, D. R., Laetsch, D. R., Blaxter, M. L. & Blaxter, M. L. KinFin: Software for
910 taxon-aware analysis of clustered protein sequences. *G3 (Bethesda, Md.)* 7, 3349–3357
911 (2017).

912 34. Takami, H. *et al.* An automated system for evaluation of the potential functionome:
913 MAPLE version 2.1.0. *Dna Res* 23, 467–475 (2016).

914 35. Kapli, P. *et al.* Lack of support for Deuterostomia prompts reinterpretation of the first
915 Bilateria. *Sci Adv* 7, eabe2741 (2021).

916 36. Thiel, D., Franz-Wachtel, M., Aguilera, F. & Hejnol, A. Xenacoelomorph
917 neuropeptidomes reveal a major expansion of neuropeptide systems during early bilaterian
918 evolution. *Molecular Biology And Evolution* 35, 2528–2543 (2018).

919 37. Negri, L. & Ferrara, N. The Prokineticins: neuromodulators and mediators of
920 inflammation and myeloid cell-dependent angiogenesis. *Physiol Rev* 98, 1055–1082 (2018).

921 38. Ericsson, L. & Söderhäll, I. Astakines in arthropods–phylogeny and gene structure. *Dev*
922 *Comp Immunol* 81, 141–151 (2018).

923 39. Simakov, O. *et al.* Deeply conserved synteny resolves early events in vertebrate
924 evolution. *Nat Ecol Evol* 1–11 (2020).

925 40. Simakov, O. *et al.* Deeply conserved synteny and the evolution of metazoan
926 chromosomes. *Sci Adv* 8, eabi5884 (2022).

927 41. Rouse, G. W., Wilson, N. G., Carvajal, J. I. & Vrijenhoek, R. C. New deep-sea species of
928 *Xenoturbella* and the position of Xenacoelomorpha. *Nature* 530, 94–97 (2016).

929 42. Hejnol, A. Acoelomorpha and Xenoturbellida. in 203–214 (Springer Vienna, 2015).

930 43. Brauchle, M. *et al.* Xenacoelomorpha survey reveals that all 11 animal homeobox gene
931 classes were present in the first bilaterians. *Genome Biol Evol* 10, 2205–2217 (2018).

932 44. Jimenez-Guri, E., Paps, J., Garcia-Fernandez, J. & Salo, E. Hox and ParaHox genes in
933 Nemertodermatida, a basal bilaterian clade. *Int J Dev Biology* 50, 675–679 (2006).

934 45. Ryan, J. F. *et al.* The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes:
935 evidence from the starlet sea anemone, *Nematostella vectensis*. *Genome Biol* 7, R64–R64
936 (2006).

937 46. Jekely, G. Global view of the evolution and diversity of metazoan neuropeptide signaling.
938 *Proc National Acad Sci* 110, 8702–8707 (2013).

939 47. Mirabeau, O. & Joly, J.-S. Molecular evolution of peptidergic signaling systems in
940 bilaterians. *Proc National Acad Sci* 110, E2028–E2037 (2013).

941 48. Natsidis, P., Kapli, P., Schiffer, P. H. & Telford, M. J. Systematic errors in orthology
942 inference and their effects on evolutionary analyses. *iScience* 24, 102110 (2021).

943 49. Weisman, C. M., Murray, A. W. & Eddy, S. R. Many, but not all, lineage-specific genes
944 can be explained by homology detection failure. *Plos Biol* 18, e3000862 (2020).

945 50. Howe, K. *et al.* Structure and evolutionary history of a large family of NLR proteins in
946 the zebrafish. *Open Biology* 6, 160009 (2016).

947 51. Pillonel, T., Bertelli, C. & Greub, G. Environmental metagenomic assemblies reveal
948 seven new highly divergent chlamydial lineages and hallmarks of a conserved intracellular
949 lifestyle. *Front Microbiol* 9, 79 (2018).

950 52. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
951 sequence data. *Bioinformatics* 30, 2114–2120 (2014).

952 53. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies.
953 *F1000research* 6, 1287 (2017).

954 54. Elsworth, B., Jones, M. & Blaxter, M. Badger--an accessible genome exploration
955 environment. *Bioinformatics* 29, 2788–2789 (2013).

956 55. Lewis, S. H. *et al.* Widespread conservation and lineage-specific diversification of
957 genome-wide DNA methylation patterns across arthropods. *Plos Genet* 16, e1008864 (2020).

958 56. Lafontaine, D. L., Yang, L., Dekker, J. & Gibcus, J. H. Hi-C 3.0: improved protocol for
959 genome-wide chromosome conformation capture. *Curr Protoc* 1, e198 (2021).

960 57. Marie-Nelly, H. *et al.* High-quality genome (re)assembly using chromosomal contact
961 data. *Nat Commun* 5, 5695 (2014).

962 58. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature*
963 *Methods* 9, 357–359 (2012).

964 59. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-Seq: reference
965 generation and analysis with Trinity. 8, 1494–1512 (2013).

966 60. *RNA-Seq De novo Assembly Using Trinity*. 1–7 (2015).

967 61. Wheeler, T. J. *et al.* Dfam: a database of repetitive DNA based on profile hidden Markov
968 models. *Nucleic Acids Res* 41, D70–D82 (2013).

969 62. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res* 44,
970 D81–D89 (2016).

971 63. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron
972 submodel. *Bioinformatics* 19 Suppl 2, ii215-25 (2003).

973 64. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads
974 into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* 42, e119–
975 e119 (2014).

976 65. Wu, T. D., Reeder, J., Lawrence, M., Becker, G. & Brauer, M. J. GMAP and GSNAP for
977 Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality.
978 *Methods Mol Biology Clifton N J* 1418, 283–334 (2016).

979 66. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25,
980 2078–2079 (2009).

981 67. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T.
982 BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* 27,
983 1691–1692 (2011).

984 68. Robertson, H. E. *et al.* Single cell atlas of *Xenoturbella bocki* highlights the limited cell-
985 type complexity of a non-vertebrate deuterostome lineage. *Biorxiv* 2022.08.18.504214 (2022)
986 doi:10.1101/2022.08.18.504214.

987 69. Nishimura, O., Hara, Y. & Kuraku, S. gVolante for standardizing completeness
988 assessment of genome and transcriptome assemblies. *Bioinformatics* 33, 3635–3637 (2017).

989 70. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.
990 *Bioinformatics* 30, 1236–1240 (2014).

991 71. Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence
992 classification and comparison. *Methods in molecular biology (Clifton, N.J.)* 396, 59–70
993 (2007).

994 72. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7:
995 Improvements in Performance and Usability. *Mol Biol Evol* 30, 772–780 (2013).

996 73. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic
997 inference in the genomic era. *Mol Biol Evol* 37, 1530–1534 (2020).

998 74. Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. IQ-TREE: A fast and
999 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*
1000 32, 268–274 (2015).

1001 75. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
1002 DIAMOND. *Nature Methods* 12, 59–60 (2014).

1003 76. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale
1004 using DIAMOND. *Nat Methods* 18, 366–368 (2021).

1005 77. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future.
1006 *Nucleic Acids Research* 44, D279-85 (2016).

1007 78. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
1008 genomics. *Genome Biol* 20, 238 (2019).

1009 79. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
1010 comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16, E9-
1011 13 (2015).

1012 80. Han, M. V., Thomas, G. W. C., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain
1013 and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol*
1014 *Biol Evol* 30, 1987–1997 (2013).

1015 81. Leclère, L. *et al.* The genome of the jellyfish *Clytia hemisphaerica* and the evolution of
1016 the cnidarian life-cycle. *Nat. Ecol. Evol.* 1–41 (2019).

1017 82. Pett, W. *et al*. The Role of Homology and Orthology in the Phylogenomic Analysis of
1018 Metazoan Gene Content. *Mol Biol Evol,* 1–7 (2019).

1019 83. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for
1020 phylogenetic reconstruction and molecular dating. 25, 2286–2288 (2009).

1021 84. Brauchle, M. *et al.* Xenacoelomorpha survey reveals that all 11 animal homeobox gene
1022 classes were present in the first bilaterians. *Genome biology and evolution* (2018)
1023 doi:10.1093/gbe/evy170.

1024 85. Sarkies, P. *et al.* Ancient and novel small RNA pathways compensate for the loss of
1025 piRNAs in multiple independent nematode lineages. *Plos Biol* 13, e1002061 (2015).

1026 86. Kang, W. *et al.* miRTrace reveals the organismal origins of microRNA sequencing data.
1027 *Genome Biol* 19, 213 (2018).

1028 87. Umu, S. U. *et al.* Accurate microRNA annotation of animal genomes using trained
1029 covariance models of curated microRNA complements in MirMachine. *Biorxiv*
1030 2022.11.23.517654 (2023).

1031 88. Wheeler, B. M. *et al.* The deep evolution of metazoan microRNAs. *Evol Dev* 11, 50–68
1032 (2009).

1033 89. Fromm, B. *et al.* A uniform system for the annotation of vertebrate microRNA genes and
1034 the evolution of the human microRNAome. *Annu Rev Genet* 49, 213–242 (2015).

1035 90. Fromm, B. *et al.* MirGeneDB 2.1: toward a complete sampling of all major animal phyla.
1036 *Nucleic Acids Res* 50, D204–D210 (2022).

1037 91. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and
1038 iterative HMM search procedure. *Bmc Bioinformatics* 11, 431–431 (2010).

1039 92. Zandawala, M. *et al.* Discovery of novel representatives of bilaterian neuropeptide
1040 families and reconstruction of neuropeptide precursor evolution in ophiuroid echinoderms.
1041 *Open Biol* 7, 170129 (2017).

1042 93. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein families based
1043 on pairwise similarity. *Bioinformatics* 20, 3702–3704 (2004).

1044 94. Armenteros, J. J. A. *et al.* SignalP 5.0 improves signal peptide predictions using deep
1045 neural networks. *Nat Biotechnol* 37, 420–423 (2019).

1046 95. Southey, B. R., Rodriguez-Zas, S. L. & Sweedler, J. V. Prediction of neuropeptide
1047 prohormone cleavages with application to RFamides. *Peptides* 27, 1087–1098 (2006).

1048 96. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldon, T. trimAl: a tool for automated
1049 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973
1050 (2009).

1051 97. Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. IQ-TREE: A fast and
1052 effective stochastic algorithm for estimating Maximum-Likelihood phylogenies. *Mol Biol*
1053 *Evol* 32, 268–274 (2015).

1054 98. Simakov, O. *et al.* Hemichordate genomes and deuterostome origins. *Nature* 527, 459–
1055 465 (2015).

1056 99. Cherif--Feildel, M., Berthelin, C. H., Rivière, G., Favrel, P. & Kellner, K. Data for
1057 evolutive analysis of insulin related peptides in bilaterian species. *Data Brief* 22, 546–550
1058 (2019).

1059 100. Thiel, D., Franz-Wachtel, M., Aguilera, F. & Hejnol, A. Changes in the neuropeptide
1060 complement correlate with nervous system architectures in xenacoelomorphs. 1–57 (2018)
1061 doi:10.1101/265579.

1062 101. Roch, G. J. & Sherwood, N. M. Glycoprotein hormones and their receptors emerged at
1063 the origin of metazoans. *Genome Biol Evol* 6, 1466–79 (2014).

1064 102. Oliveira, A. L. de, Calcino, A. & Wanninger, A. Ancient origins of arthropod moulting
1065 pathway components. *elife* 8, e46113 (2019).

1066 103. Smýkal, V. *et al.* Complex evolution of insect insulin receptors and homologous decoy
1067 receptors, and functional significance of their multiplicity. *Mol Biol Evol* 37, 1775–1789
1068 (2020).

1069 104. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for
1070 the analysis of massive data sets. *Nat Biotechnol* 35, 1026–1028 (2017).

1071 105. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene
1072 synteny and collinearity. *Nucleic Acids Research* 40, e49–e49 (2012).

1073 106. Warren, R. L. *et al.* LINKS: Scalable, alignment-free scaffolding of draft genomes with
1074 long reads. *Gigascience* 4, 35 (2015).