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.*

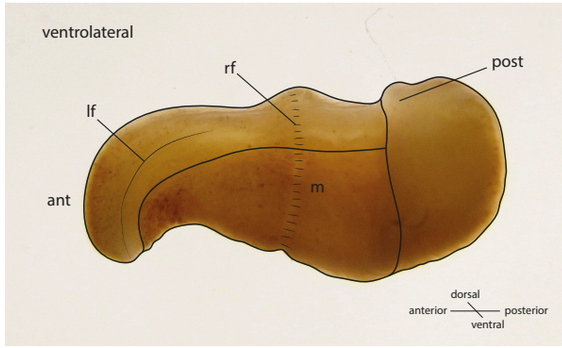
bocki has a size of ~111 Mbp in 18 chromosome like scaffolds, with repeat content and intron, exon and intergenic space comparable to other bilaterian invertebrates. We find *X. bocki* to have a similar number of genes to other bilaterians and to have retained ancestral metazoan synteny. Key bilaterian 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.

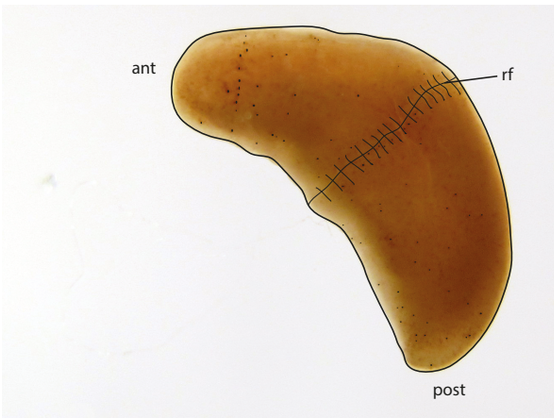
Introduction

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: Xenoturbellida and Acoelomorpha constitute their own phylum, the Xenacoelomorpha[3,4]. The monophyly of Xenacoelomorpha is 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

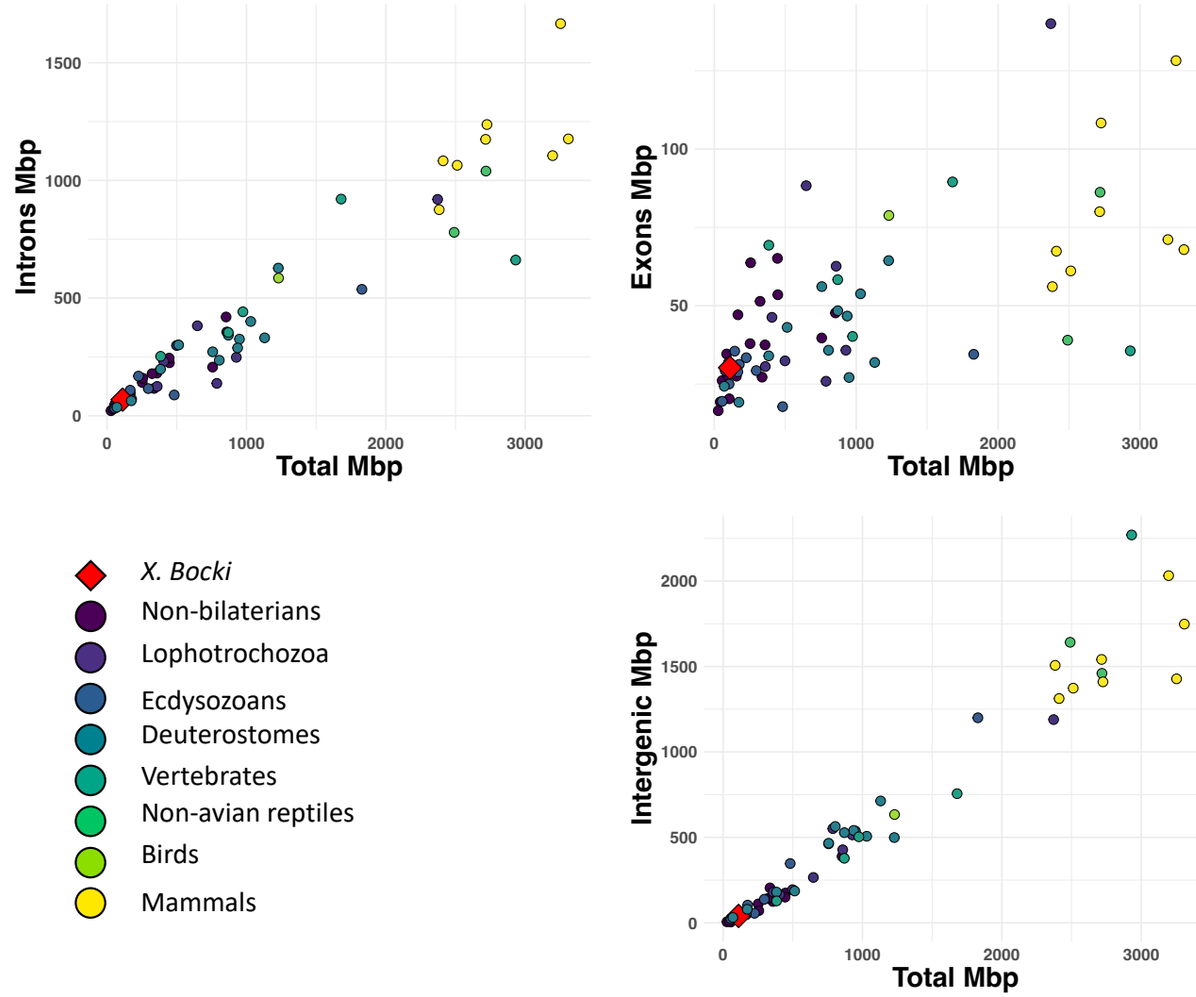


Figure 1: (a) Schematic drawings of *Xenoturbella bocki* showing the simple body organisation of the marine vermiform animal. Abbreviations: ant - anterior, post - posterior, lf - lateral furrow, rf - ring furrow, m - mouth opening. (b) A comparison of total length of exons, introns, and intergeneic 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 unusual modes 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 of genomic
113 traits, such as gene content and genome-structure and to help reconstruct the genome
114 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 8.5kb over 37,880 contigs with a maximum length of
127 206,709bp. 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 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 those described in other animal genomes[25]
152 show that *X. bocki* falls within the range of other similarly sized metazoan genomes
153 (Fig. 1b) for all these measures.
154
155 <u>The genome of a co-sequenced *Chlamydia* species</u>
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 gene 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

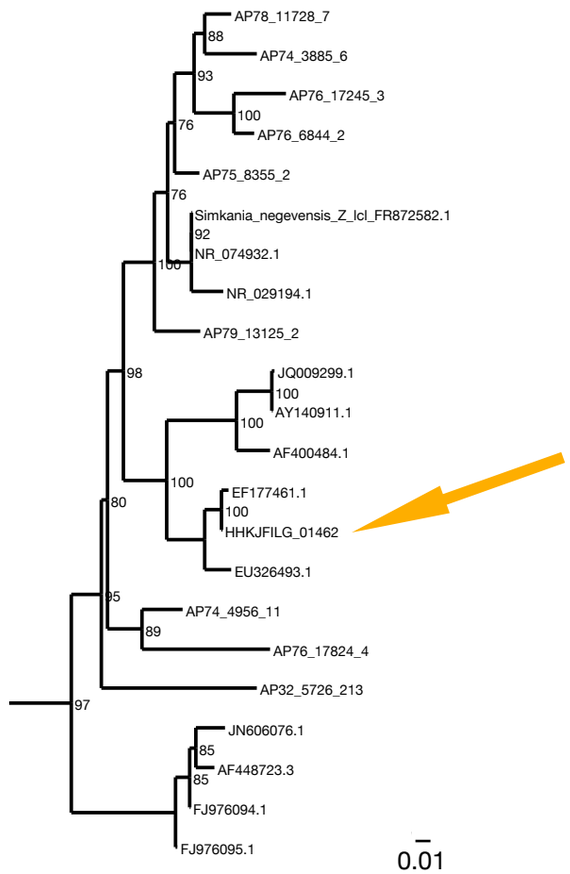
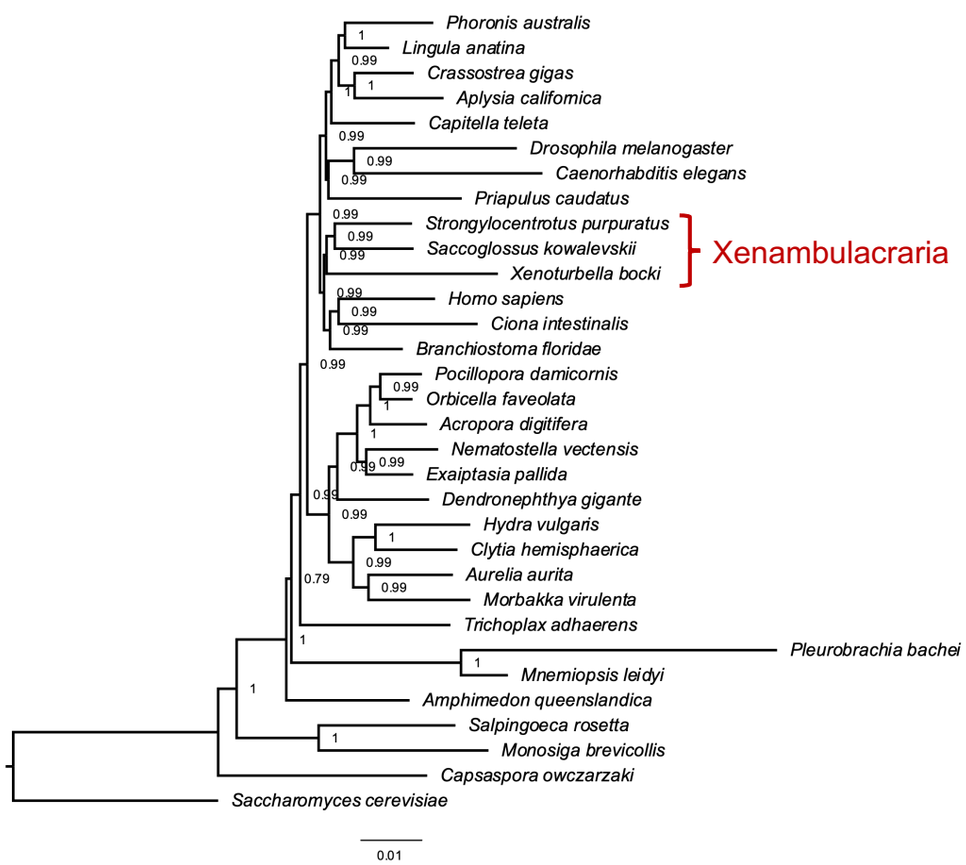
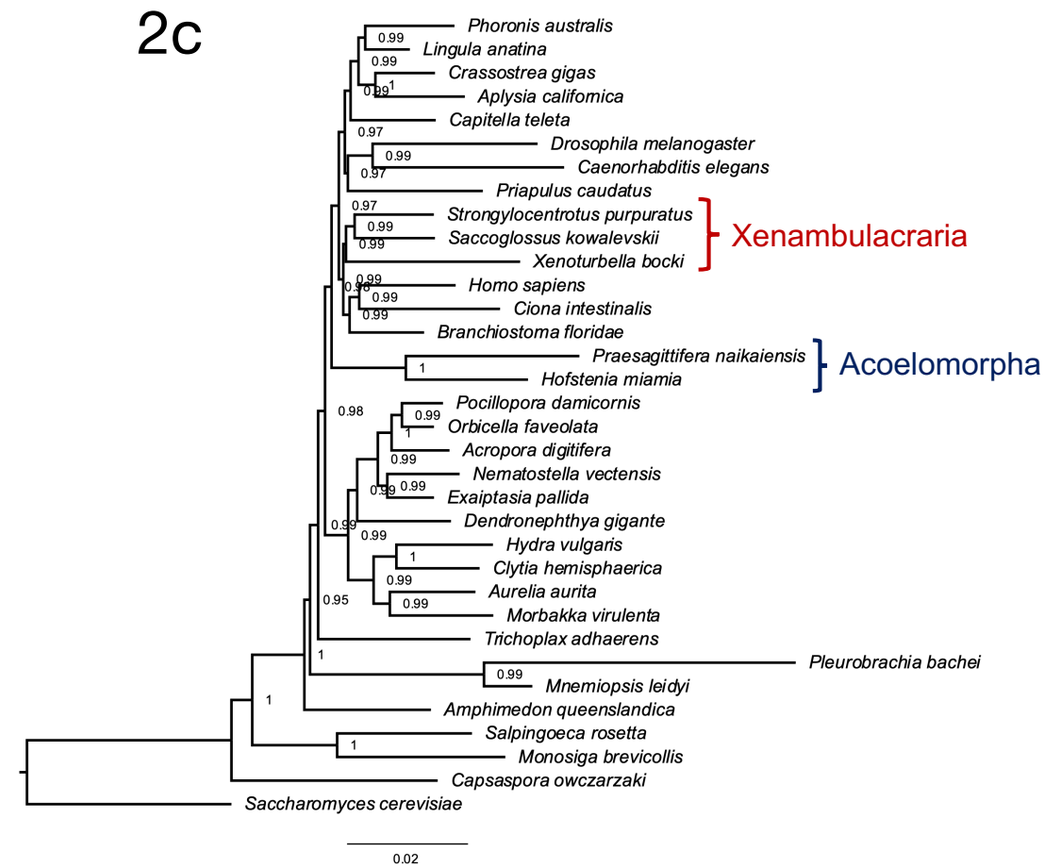


Figure 2a: *Xenoturbella bocki* harbours a marine Chlamydiae species as potential symbiont. In the phylogenetic analysis of 16S rDNA (ML: GTR+F+R7; bootstrap values included) the bacteria in our *X. bocki* isolate (arrow) are sister to a previous isolate from *X. westbaldi*. *X. westbaldi* is most likely 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

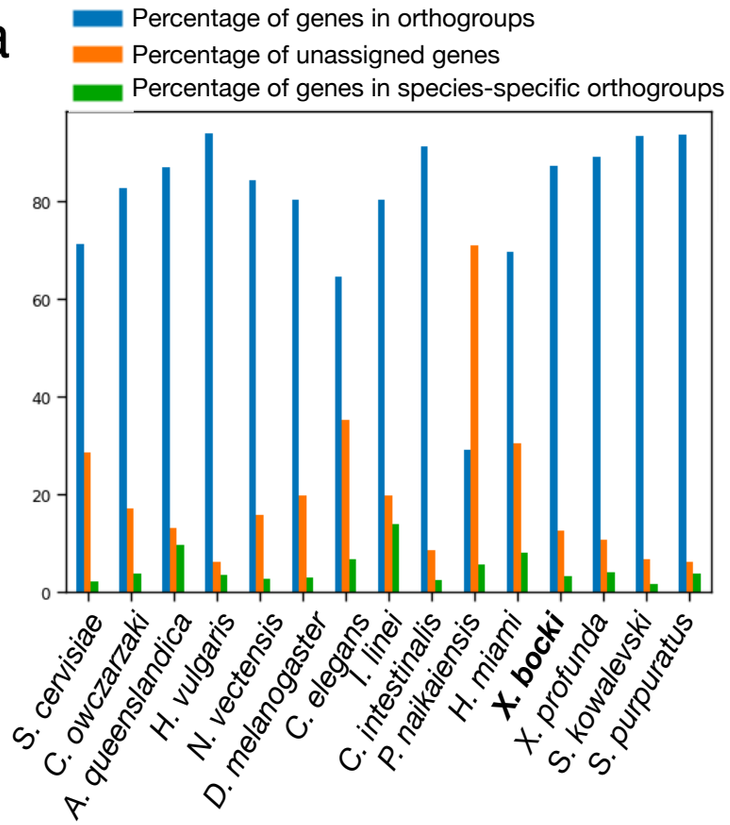


2c

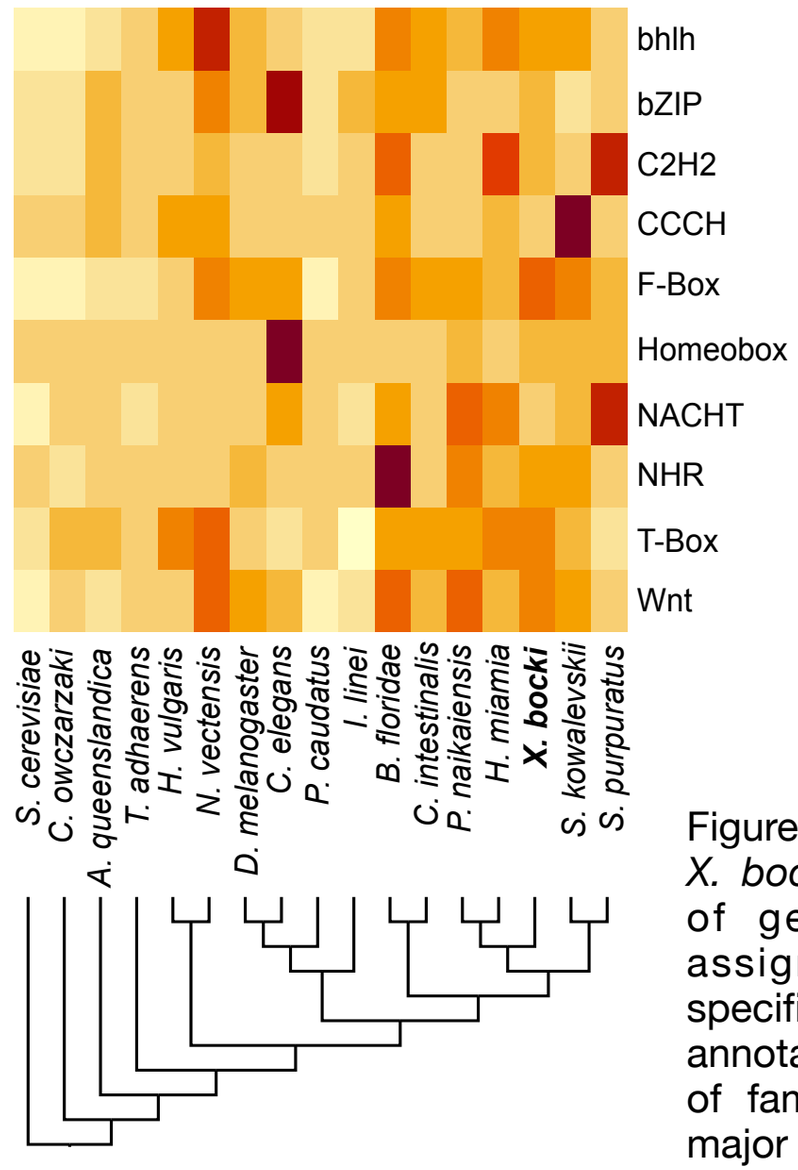


204 set in BUSCO (v5) the *X. bocki* proteome translated from our gene predictions is
205 82.5% complete and ~90% complete when partial hits are included (82% and 93%
206 respectively for the Eukaryote gene set). This estimate is even higher in the acoel
207 *Hofstenia miamia*, which was originally reported to be 90%[16], but in our re-analysis
208 was 95.71%. In comparison, the morphologically highly simplified and fast evolving
209 annelid *Intoshia linei*[31] has a genome of fewer than 10,000 genes[32] and in our analysis
210 is only ~64% complete for the BUSCO (v5) Metazoa set. The model nematode
211 *Caenorhabditis elegans* is ~79% complete for the same set. Despite the morphological
212 simplicity of both *Xenoturbella*, and *Hofstenia,* these Xenacoelomorpha are missing
213 few core genes compared to other bilaterian lineages that we perceive to have
214 undergone a high degree of morphological evolutionary change (such as the evolution
215 of miniaturisation, parasitism, sessility etc).
216 Using our phylogenomic matrix of gene presence/absence (see above) we identified
217 all orthologs that could be detected both in Bilateria (in any bilaterian lineage) and in
218 any non-bilaterian; ignoring horizontal gene transfer and other rare events, these
219 genes must have existed in Urbilateria (and, of less interest to us, in Urmetazoa). The
220 absence of any of these bilaterian genes in any lineage of Bilateria must therefore be
221 explained by loss of the gene. All individual bilaterian genomes were missing many
222 of these orthologs but Xenacoelomorphs and some other bilaterians lacked more of
223 these than did other taxa. The average numbers of these genes present in bilaterians
224 = 7577; *Xenoturbella* = 5459; *Hofstenia* = 5438; *Praesagittifera* = 4280; *Drosophila* =
225 4844; *Caenorhabditis* = 4323.
226 To better profile the *Xenoturbella* and xenacoelomorph molecular toolkit, we
227 used OrthoFinder to conduct orthology searches in a comparison of 155 metazoan
228 and outgroup species, including the transcriptomes of the sister species *X. profunda*
229 and an early draft genome of the acoel *Paratomella rubra* we had available, as well as
230 the *Hofstenia* and *Praesagittifera* proteomes (Supplementary online material). For
231 each species we counted, in each of the three Xenacoelomorphs, the number of
232 orthogroups for which a gene was present. The proportion of orthogroups containing
233 an *X. bocki* and *X. profunda* protein (87.4% and 89.2%) are broadly similar to the
234 proportions seen in other well characterised genomes, for example *S. purpuratus*
235 proteins (93.8%) or *N. vectensis* proteins (84.3%) (Fig 3a). In this analysis, the fast-

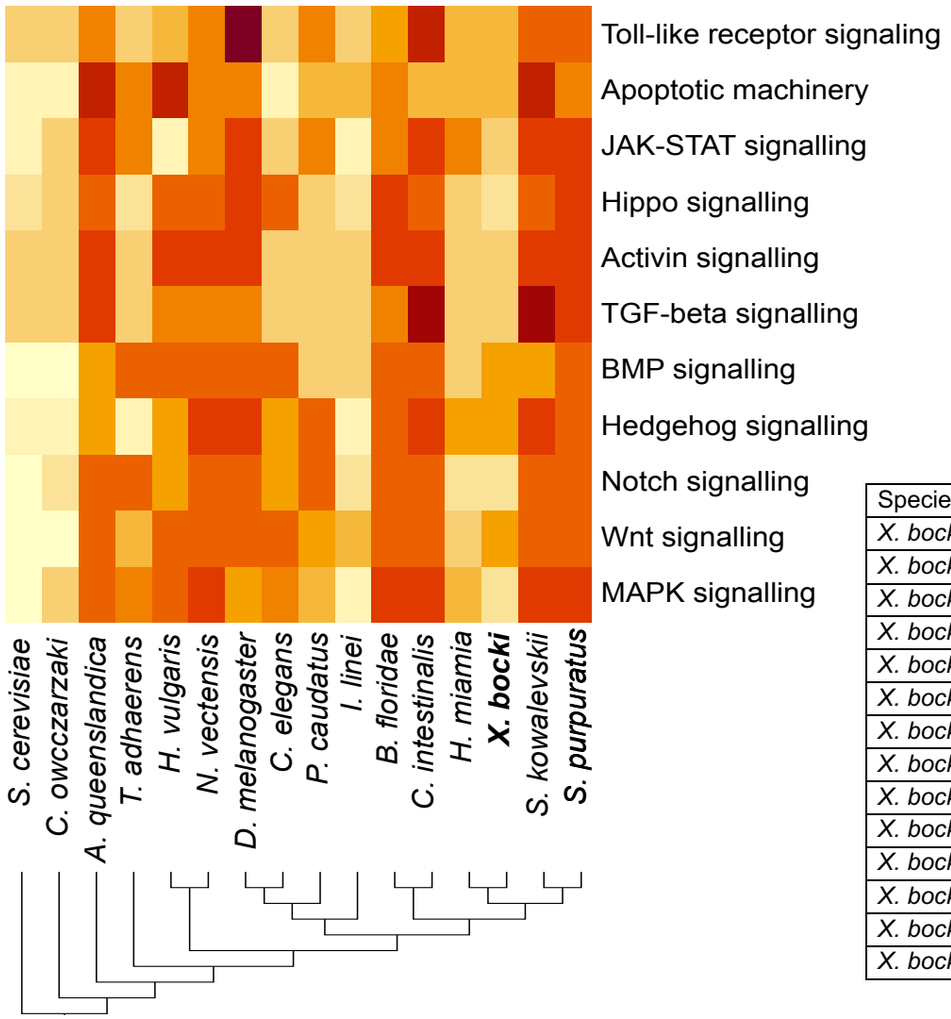
3a



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 evolving nematode *Caenorhabditis elegans* appears as an outlier, with only ~64% of
237 its proteins in orthogroups and ~35% unassigned. Both *Xenoturbella* species have an
238 intermediate number of unassigned genes of ~11-12%. Similarly, the proportion of
239 species-specific genes (~14% of all genes) corresponds closely to what is seen in
240 most other species (with the exception of the parasitic annelid *I. linei*, Fig. 3a).
241
242 Idiosyncrasies of *Xenoturbella*
243 In order to identify sets of orthologs specific to the two *Xenoturbella* species we used
244 the kinfin software[33] and found 867 such groups in the OrthoFinder clustering. We
245 profiled these genes based on Pfam domains and GO terms derived from
246 InterProScan. While these *Xenoturbella* specific proteins fall into diverse classes, we
247 did see a considerable number of C-type lectin, Immunoglobulin-like, PAN, and Kringle
248 domain containing Pfam annotations. Along with the Cysteine-rich secretory protein
249 family and the G-protein coupled receptor activity GO terms, these genes and families
250 of genes may be interesting for future studies into the biology of *Xenoturbella* in its
251 native environment.
252
253 Gene families and signaling pathways are retained in *X. bocki*
254 In our orthology clustering we did not see an inflation of *Xenoturbella*-specific groups
255 in comparison to other taxa, but also no conspicuous absence of major gene families
256 (Fig. 3b). Family numbers of transcription factors like Zinc-fingers or homeobox-
257 containing genes, as well as, for example, NACHT-domain encoding genes seem to
258 be neither drastically inflated nor contracted in comparison to other species in our
259 InterProScan based analysis.
260 To catalogue the completeness of cell signalling pathways we screened the *X.*
261 *bocki* proteome against KEGG pathway maps using GenomeMaple[34]. The *X. bocki*
262 gene set is largely complete in regard to the core proteins of these pathways, while an
263 array of effector proteins is absent (Fig. 3c). In comparison to other metazoan species,
264 as well as a unicellular choanoflagellate and a yeast, the *X. bocki* molecular toolkit has
265 significantly lower KEGG completeness than morphologically complex animals such
266 as the sea urchin and amphioxus (t-test; Fig. 3c). *Xenoturbella* is, however, not
267 significantly less complete when compared to other bilaterians considered to have low

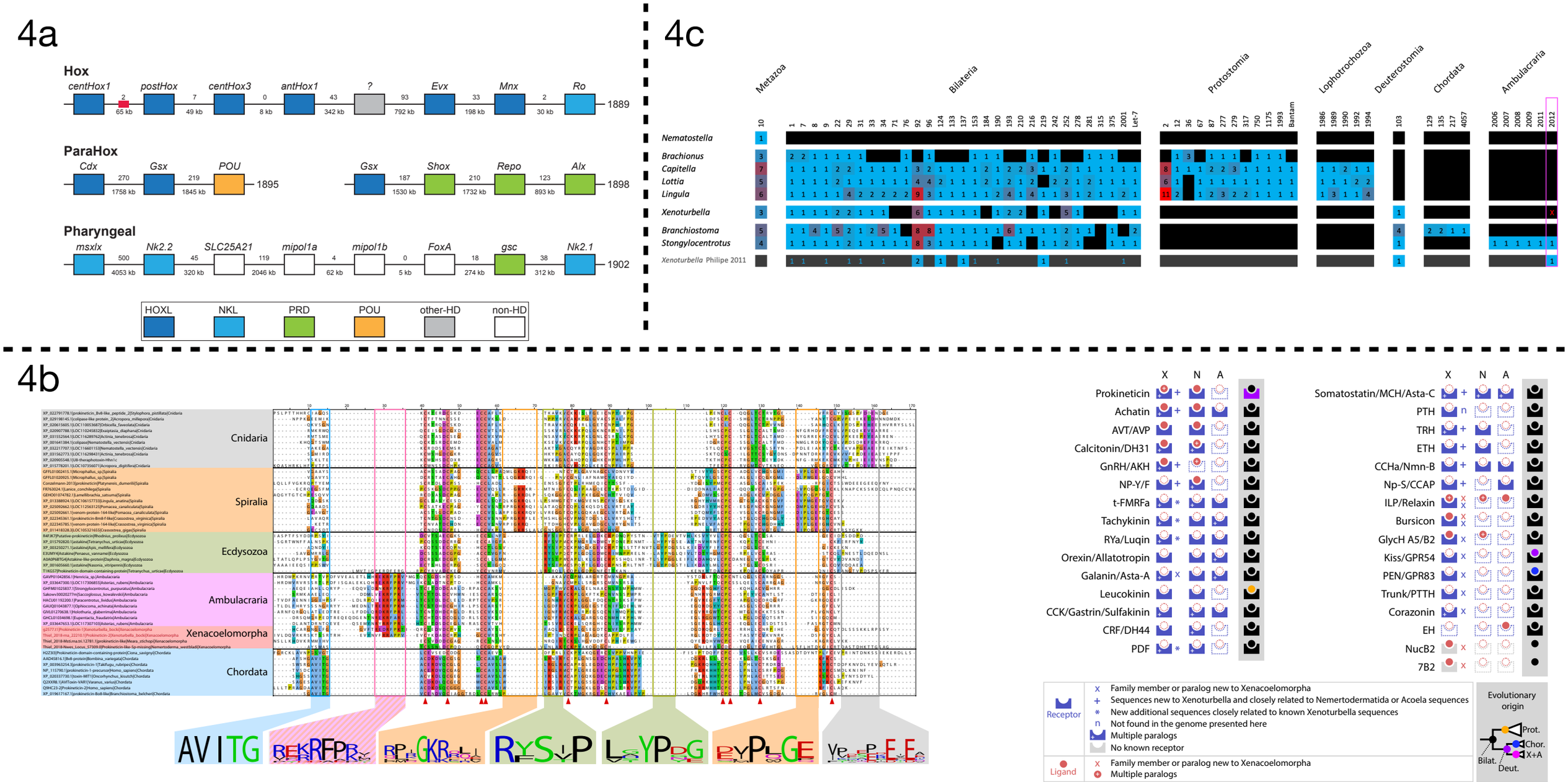


Figure 4: (a) *X. bocki* has 5 HOX genes, which are located in relatively close proximity on one of our chromosome size scaffolds. Similar clusters exist for the ParaHox and "pharyngeal" genes. Numbers between genes are distance (below) and number of genes between (below). Colours indicate gene families. 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 morphological complexity and which have been shown to have reduced gene content,
269 such as *C. elegans*, the annelid parasite *Intoshia linei*, or the acoel *Hofstenia miamia*
270 (Fig. 3c).

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

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

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

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

297

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

299 A catalogue of acoelomorph neuropeptides was previously described using

300 transcriptome data[36]. We have discovered 12 additional neuropeptide genes and 39
301 new neuropeptide receptors in *X. bocki* adding 6 bilaterian peptidergic systems to the
302 *Xenoturbella* catalogue (NPY-F ; MCH/Asta-C ; TRH ; ETH ; CCHa/Nmn-B ; Np-
303 S/CCAP), and 6 additional bilaterian systems to the Xenacoelomorpha catalogue
304 (Corazonin ; Kiss/GPR54 ; GPR83 ; 7B2 ; Trunk/PTTH ; NUCB2) making a total of 31
305 peptidergic systems (Fig. 4, Supplementary).

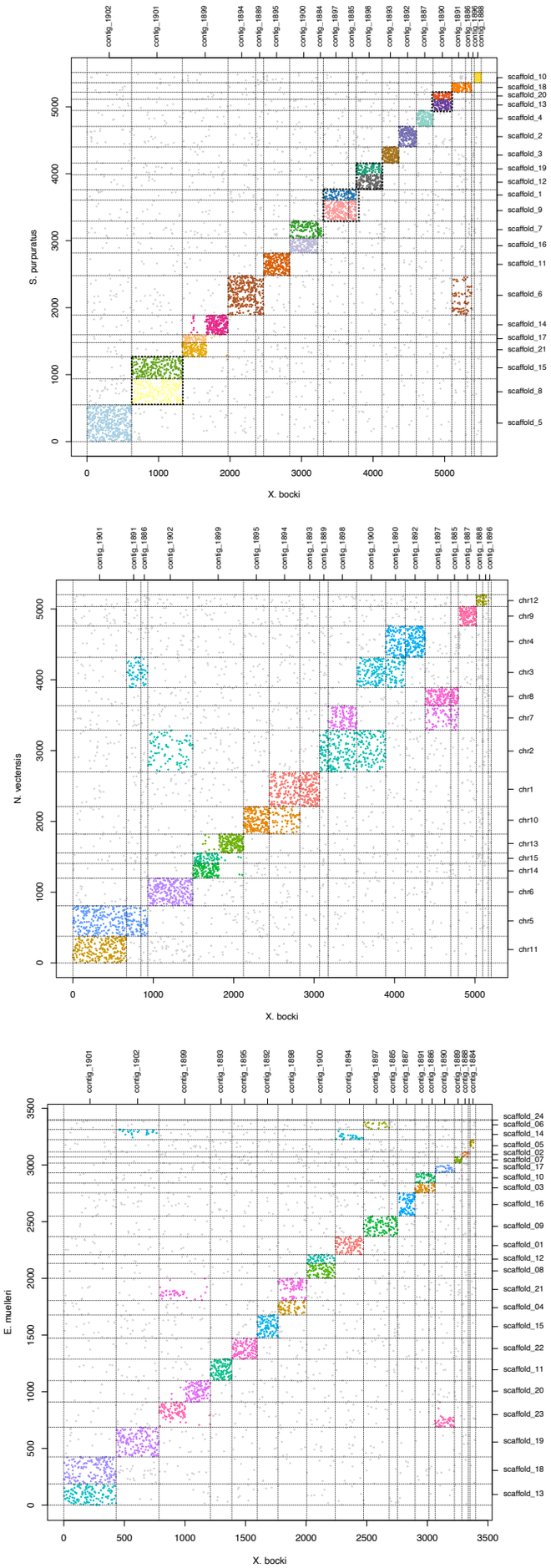
306 Among the ligand genes, we identified 6 new repeat-containing sequences. One
307 of these, the LRIGamide-peptide, had been identified in Nemertodermatida and
308 Acoela and its loss in *Xenoturbella* had been proposed[36]. We also identified the first
309 7B2 neuropeptide and NucB2/Nesfatin genes in Xenacoelomorpha. Finally, we
310 identified 3 new *X. bocki* insulin-like peptides, one of them sharing sequence similarity
311 and an atypical cysteine pattern with the Ambulacrarian octinsulin, constituting a
312 potential synapomorphy of Xenambulacraria (see Supplementary).

313 Our searches also revealed the presence of components of the arthropod
314 moulting pathway components (PTTH/trunk, NP-S/CCAP and Bursicon receptors),
315 which have recently been shown to be of ancient origin (de Oliveira et al., 2019). We
316 further identified multiple paralogs for, e.g the Tachykinin, Rya/Luquin, tFMRFa,
317 Corazonin, Achatin, CCK, and Prokineticin receptor families. Two complete *X. bocki*
318 Prokineticin ligands were also found in our survey (see Supplementary).

319 Chordate Prokineticin ligands possess a conserved N-terminal “AVIT” sequence
320 required for the receptor activation[37]. This sequence is absent in arthropod Astakine,
321 which instead possess two signature sequences within their Prokineticin domain [38].
322 To investigate Prokineticin ligands in Xenacoelomorpha we compared the sequences
323 of their prokineticin ligands with those of other bilaterians (Fig. 4b, Supplementary).
324 Our alignment reveals clade specific signatures already reported in Ecdysozoa and
325 Chordata sequences, but also two new signatures specific to Lophotrochozoa and
326 Cnidaria sequences, as well as a very specific “K/R-RFP-K/R” signature shared only
327 by ambulacrarian and *Xenoturbella bocki* sequences. The shared
328 Ambulacrarian/Xenacoelomorpha signature is found at the same position as the
329 Chordate sequence involved in receptor activation - adjacent to the N-terminus of the
330 Prokineticin domain (Fig. 4b).
331

332 **The *X. bocki* genome contains most bilaterian miRNAs reported missing from**
333 **acoels.**
334 microRNAs have previously been used to investigate the phylogenetic position of the
335 acoels and *Xenoturbella.* The acoel *Symsagittifera roscoffensis* lacks protostome and
336 bilaterian miRNAs and this lack was interpreted as supporting the position of acoels
337 as sister-group to the Nephrozoa. Based on shallow 454 microRNA sequencing (and
338 sparse genomic traces) of *Xenoturbella*, some of the bilaterian miRNAs missing from
339 acoels were found - 16 of the 32 expected metazoan (1 miRNA) and bilaterian (31
340 miRNAs) microRNA families – of which 6 could be identified in genome traces[9].
341 By deep sequencing two independent small RNA samples, we have now
342 identified the majority of the missing metazoan and bilaterian microRNAs and
343 identified them in the genome assembly (Fig. 4c). Altogether, we found 23 out of 31
344 bilaterian microRNA families (35 genes including duplicates); the single known
345 Metazoan microRNA family (MIR-10) in 2 copies; the Deuterostome-specific MIR-103;
346 and 7 *Xenoturbella*-specific microRNAs giving a total of 46 microRNA genes. None of
347 the protostome-specific miRNAs were found. We could not confirm in the RNA
348 sequences or new assembly a previously identified, and supposedly
349 xenambulacrarian-specific MIR-2012 ortholog.
350
351 **The *X. bocki* genome retains ancestral metazoan linkage groups.**
352 The availability of chromosome-scale genomes has made it possible to reconstruct 24
353 ancestral linkage units broadly preserved in bilaterians[39]. In fast-evolving genomes,
354 such as those of nematodes, tunicates or platyhelminths, these ancestral linkage
355 groups (ALGs) are often dispersed and/or extensively fused (Supplementary). We
356 were interested to test if the general conservation of the gene content in *X. bocki* is
357 reflected in its genome structure.
358 We compared the genome of *Xenoturbella* to several other metazoan genomes
359 and found that it has retained most of these ancestral bilaterian units: 12
360 chromosomes in the *X. bocki* genome derive from a single ALG, five chromosomes
361 are made of the fusion of two ALGs, and one *Xenoturbella* chromosome is a fusion of
362 three ALGs, as highlighted with the comparison of ortholog content with amphioxus,
363 the sea urchin and the sea scallop (Fig. 5 and Supplementary).
364 One ancestral linkage group that has been lost in chordates but not in

5a



5b

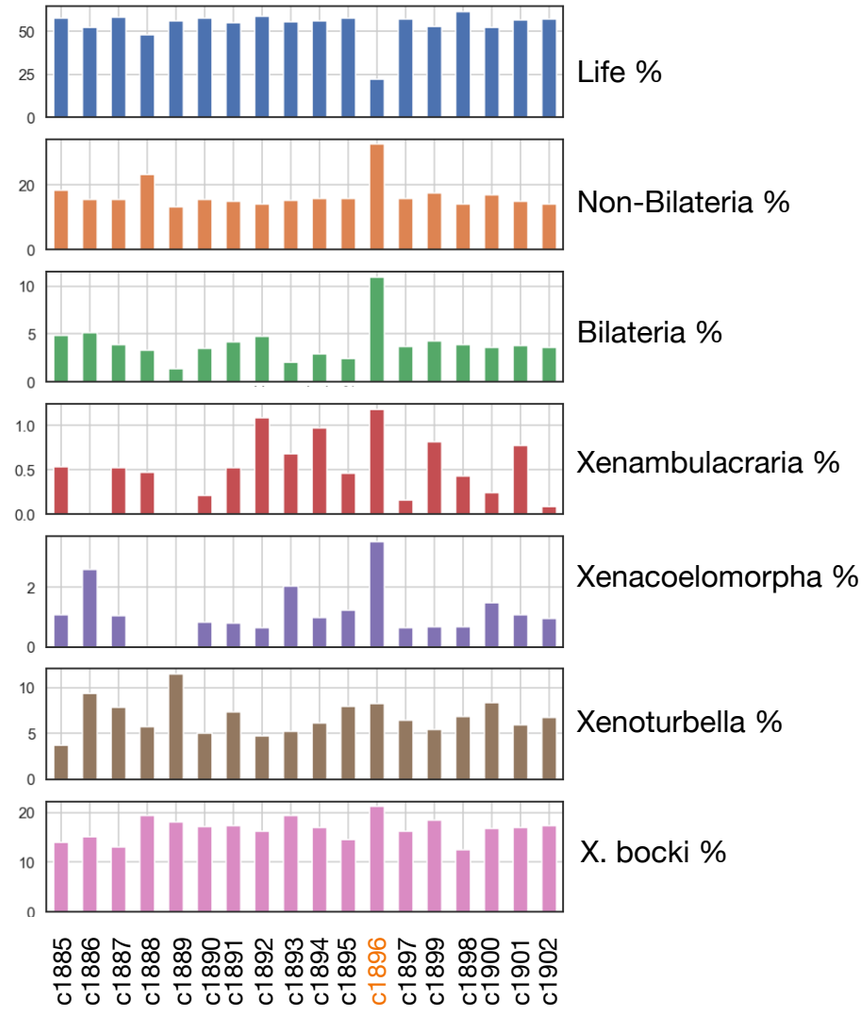


Figure 5: (a) A comparison of scaffolds in the *X. bocki* genome with other Metazoa. 17 of the 18 large scaffolds in the *X. bocki* genome are linked via synteny to distinct chromosomal scaffolds in these species. (b) Phylostratigraphic age distribution of genes on all major scaffolds in the *X. bocki* 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 ambulacrarians nor in molluscs (ALG R in sea urchin and sea scallop) is detectable in
366 *X. bocki* (Fig. 5), while *X. bocki* does not show the fusions that are characteristic of
367 lophotrochozoans.

368 We also attempted to detect some pre-bilaterian arrangement of ancestral
369 linkage: for instance, ref [40] predicted that several pre-bilaterian linkage groups
370 successively fused in the bilaterian lineage to give ALGs A1, Q and E. These ALGs
371 are all represented as single units in *X. bocki* in common with other Bilateria. None of
372 the inferred pre-bilaterian chromosomal arrangements that could have provided
373 support for the Nephrozoa hypothesis were found *in X. bocki* although of course this
374 does not rule out Nephrozoa.

375

376 <u>One *X. bocki* chromosomal fragment appears aberrant</u>

377 The smallest of the 18 large scaffolds in the *X. bocki* genome did not show strong 1:1
378 clustering with any scaffold/chromosome of the bilaterian species we compared it to.
379 To exclude potential contamination in the assembly as a source for this contig we
380 examined the orthogroups to which the genes from this scaffold belong. We found that
381 *Xenoturbella profunda*[41], for which a transcriptome is available, was the species that
382 most often occurred in the same orthogroup with genes from this scaffold (41 shared
383 orthogroups), suggesting the scaffold is not a contaminant.

384 We did observe links between the aberrant scaffold and several scaffolds from
385 the genome of the sponge *E. muelleri* in regard to synteny, but could not detect distinct
386 synteny relationships to a single scaffold in another species. In line with this, genes
387 on the scaffold show a different age structure compared to other scaffolds, with both
388 more older genes (pre bilaterian) and more *Xenoturbella* specific genes (Fig. 5b;
389 supported by Ks statistics, Supplementary). This aberrant scaffold also had
390 significantly lower levels of methylation than the rest of the genome (Supplementary).

391

392 **DISCUSSION**

393 The phylogenetic positions of *Xenoturbella* and the Acoelomorpha have been
394 controversial since the first molecular data from these species appeared over twenty
395 five years ago. Today we understand that they constitute a monophyletic group of
396 morphologically simple worms[1,9,42], but there remains a disagreement over whether
397 they represent a secondarily simplified sister group of the Ambulacraria or a primitively

398 simple sister group to all other Bilateria.

399 Previous analyses of the content of genomes, especially of Acoela, have been
400 used to bolster the latter view, with the small number of Hox genes and of microRNAs
401 of acoels being interpreted as representing an intermediate stage on the path to the
402 ~8 Hox genes and 30 odd microRNAs of the Nephrozoa. A strong version of the
403 Nephrozoa idea would go further than these examples and anticipate, for example, a
404 genome-wide paucity of bilaterian genes, GRNs and biochemical pathways and/or an
405 arrangement of chromosomal segments intermediate between those of the
406 Eumetazoa and the Nephrozoa.

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

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

420 We found the *X. bocki* genome to be fairly compact, but not unusually reduced
421 in size compared to many other bilaterians. It appears to contain a similar number of
422 genes (~15,000) as other animals, for example from the model organisms *D.*
423 *melanogaster* (>14,000) and *C. elegans (*~20,000). The BUSCO completeness, as
424 well as a high level of representation of *X. bocki* proteins in the orthogroups of our 155
425 species orthology screen indicates that we have annotated a near complete gene set.
426 Surprisingly, there are fewer genes than in the acoel *Hofstenia* (>22,000; BUSCO_v5
427 score ~95%). This said, of the genes found in Urbilateria (orthogroups in our
428 presence/absence analysis containing a member from both a bilaterian and an
429 outgroup) *Xenoturbella* and *Hofstenia* have very similar numbers (5459 and 5438

430 respectively). Gene, intron and exon lengths all also fall within the range seen in many
431 other invertebrate species[25]. It thus seems that basic genomic features in *Xenoturbella*
432 are not anomalous among Bilateria. Unlike some extremely simplified animals, such
433 as orthonectids, we observe no extreme reduction in gene content.

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

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

457 The last decade has seen a re-evaluation of our understanding of the evolution
458 of the neuropeptide signaling genes[46,47]. The peptidergic systems are thought to have
459 undergone a diversification that produced approximately 30 systems in the bilaterian
460 common ancestor[46,47]. Our study identified 31 neuropeptide systems in *X.bocki* and
461 for all of these either the ligand, receptor, or both are also present in both protostomes

462 and deuterostomes indicating conservation across Bilateria. It is likely that more
463 ligands (which are short and variable) remain to be found with better detection
464 methods. It appears that the *Xenoturbella* genome contains a nearly complete
465 bilaterian neuropeptide complement with no signs of simplification but rather signs of
466 expansions of certain gene families. Our analyses also reveal a potential
467 synapomorphy linking Xenacoelomorpha with Ambulacraria (Fig 4 and
468 Supplementary).

469 We have used the predicted presence and absence of genes across a selection
470 of metazoan genomes as characters for phylogenetic analyses. Our trees re-confirm
471 the findings of recent phylogenomic gene alignment studies in linking *Xenoturbella* to
472 the Ambulacraria. We also used these data to test different bilaterians for their
473 propensity to lose otherwise conserved genes (or for our inability to identify
474 orthologs[30]). While the degree of gene loss appears similar between *Xenoturbella* and
475 acoels, the phylogenetic analysis shows longer branches leading to the acoels, most
476 likely due to faster evolution, gain of lineages specific genes, and some degree of
477 gene loss in the branch leading to the Acoelomorpha. Recent work has shown the
478 tendency of rapidly evolving genes (such as those belonging to rapidly evolving
479 species) to be missed by orthology detection software[48,49].

480

481 This pattern of conservation of evolutionarily old parts of the Metazoan genome
482 is further reinforced by the retention in *Xenoturbella* of linkage groups present from
483 sponges to vertebrates. It is interesting to note that *X. bocki* does not follow the pattern
484 seen in other morphologically simplified animals such as nematodes and
485 platyhelminths, which have lost and/or fused these ancestral linkage groups. We
486 interpret this to be a signal of comparably slower genomic evolution in *Xenoturbella* in
487 comparison to some other bilaterian lineages. The fragmented genome sequence of
488 *Hofstenia* prevents us from asking whether the ancient linkage groups have also been
489 preserved in the Acoelomorpha.

490 One of the chromosome-scale scaffolds in our assembly showed a different
491 methylation and age signal, with both older and younger genes, and no clear
492 relationship to metazoan linkage groups. By analyzing orthogroups of genes on this
493 scaffold for their phylogenetic signal and finding *X. bocki* genes to cluster with those

494 of *X. profunda* we concluded that the scaffold most likely does not represent a
495 contamination. It remains unclear whether this scaffold is a fast-evolving chromosome,
496 or a chromosomal fragment or arm. Very fast evolution on a chromosomal arm has for
497 example been shown in the zebrafish[50].

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

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

518

519 # Methods

520 **Genome Sequencing, Assembly, and Scaffolding**

521 We extracted DNA from individual *Xenoturbella* specimens with a standard and
522 additionally worked with a Phenol-Chloroform protocol specifically developed to
523 extract HMW DNA (dx.doi.org/10.17504/protocols.io.mrxc57n). The extracted DNA
524 was quality controlled with a Nanodrop instrument in our laboratory and subsequently
525 a TapeStation at the sequencing center. Worms were first starved and kept in

526 repeatedly replaced salt water, reducing the likelihood of food or other contaminants
527 in the DNA extractions. First, we sequenced Illumina short paired-end reads and mate
528 pair libraries (see ref [3] for details). As the initial paired-read datasets were of low
529 complexity and coverage, we later complemented these data with an Illumina HiSeq
530 2000/2500 series paired-end dataset with ~700 bp insert size and 250bp read lengths,
531 yielding ~354 Million reads. Additionally, we generated ~40 Million
532 Illumina TruSeq Synthetic Long Reads (TSLR) for high confidence primary
533 scaffolding.

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

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

554 Library preparation for genome-wide bisulfite sequencing was performed as
555 previously described[55]. The resulting sequencing data were aligned to the *X. bocki*
556 draft genome using Bismark in non-directional mode to identify the percentage

557 methylation at each cytosine genome-wide. Only sites with >10 reads mapping were
558 considered for further analysis.
559
560
561 Preparation of the Hi-C libraries
562 The Hi-C protocol was adapted at the time from (Lieberman-Aiden et al., 2009; Sexton
563 et al., 2012 and Marie-Nelly et al., 2014). Briefly, an animal was chemically cross-
564 linked for one hour at room temperature in 30 mL of PBS 1X added with 3%
565 formaldehyde (Sigma – F8775 - 4x25 mL). Formaldehyde was quenched for 20 min
566 at RT by adding 10 ml of 2.5 M glycine. The fixed animal was recovered through
567 centrifugation and stored at -80°C until use. To prepare the proximity ligation library,
568 the animal was transferred to a VK05 Precellys tubes in 1X DpnII buffer (New England
569 Biolabs; 0.5mL) and the tissues were disrupted using the Precellys Evolution
570 homogenizer (Bertin-Instrument). SDS was added (0.3% final) to the lysate and the
571 tubes were incubated at 65°C for 20 minutes followed by an incubation at 37°C for 30
572 minutes and an incubation of 30 minutes after adding 50 µL of 20% triton-X100. 150
573 units of the DpnII restriction enzyme were then added and the tubes were incubated
574 overnight at 37°C. The endonuclease was inactivated 20 min at 65°C and the tubes
575 were then centrifuged at 16,000 x g during 20 minutes, supernatant was discarded
576 and pellets were re-suspended in 200 µl NE2 1X buffer and pooled. DNA ends were
577 labeled using 50 µl NE2 10X buffer, 37.5 µl 0.4 mM dCTP-14-biotin, 4.5 µl 10mM
578 dATP-dGTP-dTTP mix, 10 µl klenow 5 U/µL and incubation at 37°C for 45 minutes.
579 The labeling mix was then transferred to ligation reaction tubes (1.6 ml ligation buffer;
580 160 µl ATP 100 mM; 160 µl BSA 10 mg/mL; 50 µl T4 DNA ligase (New England
581 Biolabs, 5U/µl); 13.8 ml H2O) and incubated at 16°C for 4 hours. A proteinase K mix
582 was added to each tube and incubated overnight at 65°C. DNA was then extracted,
583 purified and processed for sequencing as previously described[22]. Hi-C libraries were
584 sequenced on a NextSeq 500 (2 × 75 bp, paired-end using custom made
585 oligonucleotides as in Marie-Nelly et al., 2014). Libraries were prepared separately on
586 two individuals in this way but eventually merged. Note that more recent version of the
587 HI-C protocol than the one used here have been described elsewhere[56].
588

589 InstaGRAAL assembly pre-processing

590 The primary Illumina assembly contains a number of very short contigs, which are
591 disruptive when computing the contact distribution needed for the instaGRAAL
592 proximity ligation scaffolding (pre-release version, see[57] and[22] for details). Testing
593 several Nx metrics we found a relative length threshold, that depends on the scaffolds'
594 length distribution, to be a good compromise between the need for a low-noise contact
595 distribution and the aim of connecting most of the genome. We found N90 a suitable
596 threshold and excluded contigs below 1,308 bp. This also ensured no scaffolds shorter
597 than three times the average length of a DpnII restriction fragment (RF) were in the
598 assembly. In this way every contig contained enough RFs for binning and were
599 included in the scaffolding step.

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

619

620 **Genome Annotation**

621 Transcriptome Sequencing

622 We extracted total RNA from a single *X. bocki* individual and sequenced a strand
623 specific Illumina paired end library. Extraction of total RNA was performed using a
624 modified Trizol & RNeasy hybrid protocol for which tissue had to be stored in RNAlater.
625 cDNA transcription reaction/cDNA synthesis was done using the RETROscript kit
626 (Ambion) using both Oligo(dT) and Random Decamer primers. Detailed extraction and
627 transcription protocols are available from the corresponding authors. The resulting
628 transcriptomic reads (deposited under SRX20415651) were assembled with the Trinity
629 pipeline[59,60] into 103,056 sequences (N50: 705; BUSCO_v5 Eukaryota scores:
630 C:65.1%, [S:34.1%, D:31.0%], F:22.0%, M:12.9%) for initial control and then supplied
631 to the genome annotation pipeline (below).

632

633 Repeat annotation

634 In the absence of a repeat library for Xenoturbellida we first used RepeatModeller
635 v. 1.73 to establish a library *de novo*. We then used RepeatMasker v. 4.1.0
636 (https://www.repeatmasker.org) and the Dfam library[61,62] to soft-mask the genome.
637 We mapped the repeats to the instaGRAAL scaffolded genome with RepeatMasker.

638

639 Gene prediction and annotation

640 We predicted genes using Augustus[63] implemented into the BRAKER (v.2.1.0)
641 pipeline[23,24] to incorporate the RNA-Seq data. BRAKER uses spliced aligned RNA-
642 Seq reads to improve training accuracy of the gene finder GeneMark-ET[64].
643 Subsequently, a highly reliable gene set predicted by GeneMark-ET in *ab initio* mode
644 was selected to train the gene finder AUGUSTUS, which in a final step predicted
645 genes with evidence from spliced aligned RNA-Seq reads. To make use of additional
646 single cell transcriptome data allowing for a more precise prediction of 3'-UTRs we
647 employed a production version of BRAKER (August 2018 snapshot). We had
648 previously mapped the RNA-Seq data to the genome with gmap-gsnap v. 2018-07-
649 04[65] and used samtools[66] and bamtools[67] to create the necessary input files. This
650 process was repeated in an iterative way, visually validating gene structures and
651 comparing with mappings loci inferred from a set of single-cell RNA-Seq data
652 (published elsewhere, see: [68]) in particular regarding fused genes. Completeness of

653 the gene predictions was independently assessed with BUSCO_v5[27] setting the
654 metazoan and the eukaryote datasets as reference respectively on gVolante[69]. We
655 used InterProScan v. 5.27-66.0 standalone[70,71] on the UCL cluster to annotate the
656 predicted *X. bocki* proteins with Pfam, SUPERFAM, PANTHER, and Gene3D
657 information.
658
659 Horizontal Gene Transfer
660 To detect horizontally acquired genes in the *X. bocki* genome we used a pipeline
661 available from (https://github.com/reubwn/hgt). Briefly, this uses blasts against the
662 NCBI database, alignments with MAFFT[72], and phylogenetic inferences with
663 IQTree[73,74] to infer most likely horizontally acquired genes, while trying to discard
664 contamination (e.g. from co-sequenced gut microbiota).
665
666 Orthology inference
667 We included 155 metazoan species and outgroups into our orthology analysis. We
668 either downloaded available proteomes or sourced RNA-Seq reads from online
669 repositories to then use Trinity v 2.8.5 and Trinnotate v. 3.2.0 to predict protein sets.
670 In the latter case we implemented diamond v. 2.0.0 blast[75,76] searches against UniProt
671 and Pfam[77] hmm screens against the Pfam-A dataset into the prediction process. We
672 had initially acquired 185 datasets, but excluded some based on inferior BUSCO
673 completeness, while at the same time aimed to span as many phyla as possible.
674 Orthology was then inferred using Orthofinder v. 2.2.7[78,79], again with diamond as the
675 blast engine.
676 Using InterProScan v. 5.27-66.0 standalone on all proteomes we added
677 functional annotation and then employed kinfin[33] to summarise and analyse the
678 orthology tables. For the kinfin analysis, we tested different query systems in regard
679 to phylogenetic groupings (Supplementary).
680 To screen for inflation and contraction of gene families we first employed
681 CAFE5[80], but found the analysis to suffer from long branches and sparse taxon
682 sampling in Xenambulacraria. We thus chose to query individual gene families (e.g.
683 transcription factors) by looking up Pfam annotations in the InterProScan tables of
684 high-quality genomes in our analysis.

685 Through the GenomeMaple online platform we calculated completeness of
686 signaling pathways within the KEGG database using GhostX as the search engine.
687
688 Presence/absence phylogenetics
689 We used a database of metazoan proteins, updated from ref [81], as the basis for
690 an OMA analysis to calculate orthologous groups, performing two separate runs, one
691 including *Xenoturbella* and acoels, and one with only *Xenoturbella*. We converted
692 OMA gene OrthologousMatrix.txt files into binary gene presence absence matrices in
693 Nexus format with datatype = restriction. We calculated phylogenetic trees on these
694 matrices using RevBayes (see https://github.com/willpett/metazoa-gene-content for
695 RevBayes script), as described in ref 74 with corrections for no absent sites
696 and no singleton presence, using the reversible, not the Dollo model,
697 as it is more likely to be able to correct for noise related to
698 prediction errors [82,78]. For each matrix, two runs were performed and compared and
699 consensus trees generated with bpcomp from Phylobayes[83].
700
701 *Hox and ParaHox gene cluster identification and characterisation*
702 Previous work has already used transcriptomic data and phylogenetic inference
703 to identify the homeobox repertoire in *Xenoturbella bocki*. These annotations were
704 used to identify genomic positions and gene annotations that correspond to Hox and
705 ParaHox clusters in *X. bocki*. Protein sequences of homeodomains (Evx, Cdx, Gsx,
706 antHox1, centHox1, centHox2, cent3 and postHoxP) were used as TBLASTN queries
707 to identify putative scaffolds associated with Hox and ParaHox clusters. Gene models
708 from these scaffolds were compared to the full length annotated homeobox transcripts
709 from[84] using BLASTP, using hits over 95% identity for homeobox classification. There
710 were some possible homeodomain containing genes on the scaffolds that were not
711 previously characterised and were therefore not given an annotation.
712 There were issues concerning the assignment of postHoxP and Evx to gene
713 models. To ascertain possible CDS regions for these genes, RNA-Seq reads were
714 mapped with HISAT2 to the scaffold and to previous annotation[84],were assembled
715 with Trinity and these were combined with BRAKER annotations.

716 Some issues were also observed with homeodomain queries matching genomic
717 sequences that were identical, suggesting artefactual duplications. To investigate
718 contiguity around genes the ONT reads were aligned with Minimap2 to capture long
719 reads over regions and coverage.
720
721
722 Small RNA Sequencing and Analysis
723 Two samples of starved worms were subjected to 5' monophosphate dependent
724 sequencing of RNAs between 15 and 36 nucleotides in length, according to previously
725 described methods[85]. Using miRTrace[86] 3.3, 18.6 million high-quality reads were
726 extracted and merged with the 27 635 high quality 454 sequencing reads from Philippe
727 et al. The genome sequence was screened for conserved miRNA precursors using
728 MirMachine[87] followed by a MirMiner run that used predicted precursors and
729 processed and merged reads on the genome[88]. Outputs of MirMachine and MirMiner
730 were manually curated using a uniform system for the annotation of miRNA genes[89]
731 and by comparing to MirGeneDB[90].
732
733 Neuropeptide prediction and screen
734 Neuropeptide prediction was conducted on the full set of *X.bocki* predicted
735 proteins using two strategies to detect neuropeptide sequence signatures. First, using
736 a custom script detecting the occurrence of repeated sequence patterns:
737 RRx(3,36)RRx(3,36)RRx(3,36)RR,RRx(2,35)ZRRx(2,35)ZRR,
738 RRx(2,35)GRRx(2,35)GRR, RRx(1,34)ZGRRx(1,34)ZGRR where R=K or R; x=any
739 amino acid; Z=any amino acid but repeated within the pattern. Second, using
740 HMMER3.1[91] (hmmer.org), and a combination of neuropeptide HMM models obtained
741 from the PFAM database (pfam.xfam.org) as well as a set of custom HMM models
742 derived from alignment of curated sets of neuropeptide sequences[46,47,92]. Sequences
743 retrieved using both methods and comprising fewer than 600 amino acids were further
744 validated. First, by blast analysis: sequences with E-Value ratio "best blast hit versus
745 ncbi nr database/best blast hit versus curated neuropeptide dataset" < 1e-40 were
746 discarded. Second by reciprocal best blast hit clustering using Clans[93]
747 (eb.tuebingen.mpg.de/protein-evolution/software/clans/) with a set of curated

748 neuropeptide sequences[46]. SignalP-5.0[94] (cbs.dtu.dk/services/SignalP/) was used to
749 detect the presence of a signal peptide in the curated list of predicted neuropeptide
750 sequences while Neuropred[95] (stagbeetle.animal.uiuc.edu/cgi-bin/neuropred.py) was
751 used to detect cleavage sites and post-translational modifications. Sequence
752 homology of the predicted sequence with known groups was analysed using a
753 combination of (i) blast sequence similarity with known bilaterian neuropeptide
754 sequences, (ii) reciprocal best blast hit clustering using Clans and sets of curated
755 neuropeptide sequences, (iii) phylogeny using MAFFT
756 (mafft.cbrc.jp/alignment/server/), TrimAl[96] (trimal.cgenomics.org/) and IQ-TREE[97]
757 webserver for alignment, trimming and phylogeny inference respectively. Bilaterian
758 prokineticin-like sequences were searched in ncbi nucleotide, EST and SRA
759 databases as well as in the *Saccoglossus kowalevskii* genome assembly[74,98]
760 (groups.oist.jp/molgenu) using various bilaterian prokineticin-related protein
761 sequences as query. Sequences used for alignments shown in figures were collected
762 from ncbi nucleotide and protein databases as well as from the following publications:
763 7B2[46]; NucB2[92]; Insulin[99]; Prokineticin[37,38,100]. Alignments for figures were created with
764 Jalview (jalview.org).
765
766 Neuropeptide receptor search
767 §Neuropeptide Receptor sequences for Rhodopsin type GPCR, Secretin type GPCR
768 and tyrosine and serine/threonine kinase receptors were searched by running
769 HMMER3.1 on the full set of *X.bocki* predicted proteins using the 7tm_1 (PF00001),
770 7tm_2 (PF00002) and PK_Tyr_Ser-Thr (PF07714) HMM models respectively which
771 were obtained from the PFAM database (pfam.xfam.org). Sequences above the
772 significance threshold were then aligned with sequences from the curated dataset,
773 trimmed and phylogeny inference was conducted using same method as for the
774 neuropeptide. A second alignment and phylogeny inference was conducted after
775 removal of all *X.bocki* sequences having no statistical support for grouping with any of
776 the known neuropeptide receptors from the curated dataset. Curated datasets were
777 collected from the following publications: Rhodopsin type GPCR beta and gamma and
778 Secretin type GPCR[100]; Rhodopsin type GPCR delta (Leucine-rich repeat-containing
779 G-protein coupled Receptors)[101]; Tyrosine kinase receptors[102,103]; and were

780 complemented with sequences from NCBI protein database.

781

782 Synteny

783 Ancestral linkage analyses rely on mutual-best-hits computed using Mmseqs2[104]
784 between pairs of species in which chromosomal assignments to ancestral linkage
785 groups (ALG) was previously performed, such as *Branchiostoma floridae* or *Pecten*
786 *maximus*[39]. Oxford dotplots were computed by plotting reciprocal positions of indexed
787 pairwise orthologs between two species as performed previously[39,40]. The significance
788 of ortholog enrichment in pairs of chromosomes was assessed using a fisher test.
789 We also used a Python implementation of MCscanX[105] (Haibao Tang and available
790 on https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)) to compare *X.*
791 *bocki* to *Euphydtia muelleri*, *Trichoplax adhearens*, *Branchiostoma floridae*,
792 *Saccoglossus kowalevskii*, *Ciona intestinalis*, *Nematostella vectensis*, *Asteria rubens,*
793 *Pecten maximus*, *Nemopilema nomurai, Carcinoscorpius rotundicauda* (see
794 Supplementary). Briefly, the pipeline uses high quality genomes and their annotations
795 to infer syntenic blocks based on proximity. For this an all vs. all blastp is performed
796 and synteny extended from anchors identified in this way. Corresponding heatmaps
797 (see Supplementary) were plotted with Python in a Jupyter notebooks instance.

798

799 *Chlamydia* assembly and annotation

800 We identified a highly contiguous *Chlamydia* genome in the *X. bocki* genome
801 assembly using blast. We then used our Oxford Nanopore derived long-reads to
802 scaffold the *Chlamyida* genome with LINKS[106] and annotated it with the automated
803 PROKKA pipeline. To place the genome on the *Chlamydia* tree we extracted the 16S
804 ribosomal RNA gene sequence, aligned it with set of *Chlamydia* 16S rRNA sequences
805 from[28] using MAFFT, and reconstructed the phylogeny using IQ-TREE 2[73] We
806 visualized the resulting tree with Figtree (http://tree.bio.ed.ac.uk/).

807

808

809 **Acknowledgements**

810 We thank Josh Quick and Nick Loman for help with the generation of ONT long-read
811 data. Analyses were conducted mainly on the UCL Cluster, with some computations
812 also run on the CHEOPS cluster at the University of Cologne. We are grateful to Kevin
813 J. Peterson for his comments on the manuscript, the miRNA section in particular. We

thank the Kristineberg Center for Marine Research and Innovation for their essential support in sampling *Xenoturbella*.

Conflict of interest

The authors declare no conflict of interest.

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).

871 15. Sempere, L. F., Cole, C. N., McPeek, M. A. & Peterson, K. J. The phylogenetic
872 distribution of metazoan microRNAs: insights into evolutionary complexity and constraint.
873 306, 575–588 (2006).

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

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

878 18. Martinez, P. *et al.* Genome assembly of the acoel flatworm *Symsagittifera roscoffensis*, a
879 model for research on body plan evolution and photosymbiosis. *G3 Genes Genomes Genetics*
880 13, jkac336 (2022).

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

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

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

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

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

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

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

897 26. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinform Oxf Engl* 30, 2068–
898 9 (2014).

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

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

904 29. Kjeldsen, K. U., Obst, M., Nakano, H., Funch, P. & Schramm, A. Two types of
905 endosymbiotic bacteria in the enigmatic marine worm *Xenoturbella bocki*. 76, 2657–2662
906 (2010).

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

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

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

913 33. Laetsch, D. R., Laetsch, D. R., Blaxter, M. L. & Blaxter, M. L. KinFin: Software for
914 taxon-aware analysis of clustered protein sequences. *G3 (Bethesda, Md.)* 7, 3349–3357
915 (2017).

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

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

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

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

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

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

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

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

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

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

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

938 45. Ryan, J. F. *et al.* The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes:
939 evidence from the starlet sea anemone, *Nematostella vectensis*. *Genome Biol* 7, R64–R64
940 (2006).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

983 66. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25,
984 2078–2079 (2009).

985 67. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T.
986 BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* 27,
987 1691–1692 (2011).

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

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

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

995 71. Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence
996 classification and comparison. *Methods in molecular biology (Clifton, N.J.)* 396, 59–70
997 (2007).

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

1000 73. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic
1001 inference in the genomic era. *Mol Biol Evol* 37, 1530–1534 (2020).

1002 74. Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. IQ-TREE: A fast and
1003 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*
1004 32, 268–274 (2015).

1005 75. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
1006 DIAMOND. *Nature Methods* 12, 59–60 (2014).

1007 76. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale
1008 using DIAMOND. *Nat Methods* 18, 366–368 (2021).

1009 77. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future.
1010 *Nucleic Acids Research* 44, D279-85 (2016).

1011 78. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
1012 genomics. *Genome Biol* 20, 238 (2019).

1013 79. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
1014 comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16, E9-
1015 13 (2015).

1016 80. Han, M. V., Thomas, G. W. C., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain
1017 and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol*
1018 *Biol Evol* 30, 1987–1997 (2013).

1019 81. Leclère, L. *et al.* The genome of the jellyfish *Clytia hemisphaerica* and the evolution of
1020 the cnidarian life-cycle. *Nat. Ecol. Evol.* 1–41 (2019).

1021 82. Pett, W. *et al*. The Role of Homology and Orthology in the Phylogenomic Analysis of
1022 Metazoan Gene Content. *Mol Biol Evol,* 1–7 (2019).

1023 83. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for
1024 phylogenetic reconstruction and molecular dating. 25, 2286–2288 (2009).

1025 84. Brauchle, M. *et al.* Xenacoelomorpha survey reveals that all 11 animal homeobox gene
1026 classes were present in the first bilaterians. *Genome biology and evolution* (2018)
1027 doi:10.1093/gbe/evy170.

1028 85. Sarkies, P. *et al.* Ancient and novel small RNA pathways compensate for the loss of
1029 piRNAs in multiple independent nematode lineages. *Plos Biol* 13, e1002061 (2015).

1030 86. Kang, W. *et al.* miRTrace reveals the organismal origins of microRNA sequencing data.
1031 *Genome Biol* 19, 213 (2018).

1032 87. Umu, S. U. *et al.* Accurate microRNA annotation of animal genomes using trained
1033 covariance models of curated microRNA complements in MirMachine. *Biorxiv*
1034 2022.11.23.517654 (2023).

1035 88. Wheeler, B. M. *et al.* The deep evolution of metazoan microRNAs. *Evol Dev* 11, 50–68
1036 (2009).

1037 89. Fromm, B. *et al.* A uniform system for the annotation of vertebrate microRNA genes and
1038 the evolution of the human microRNAome. *Annu Rev Genet* 49, 213–242 (2015).

1039 90. Fromm, B. *et al.* MirGeneDB 2.1: toward a complete sampling of all major animal phyla.
1040 *Nucleic Acids Res* 50, D204–D210 (2022).

1041 91. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and
1042 iterative HMM search procedure. *Bmc Bioinformatics* 11, 431–431 (2010).

1043 92. Zandawala, M. *et al.* Discovery of novel representatives of bilaterian neuropeptide
1044 families and reconstruction of neuropeptide precursor evolution in ophiuroid echinoderms.
1045 *Open Biol* 7, 170129 (2017).

1046 93. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein families based
1047 on pairwise similarity. *Bioinformatics* 20, 3702–3704 (2004).

1048 94. Armenteros, J. J. A. *et al.* SignalP 5.0 improves signal peptide predictions using deep
1049 neural networks. *Nat Biotechnol* 37, 420–423 (2019).

1050 95. Southey, B. R., Rodriguez-Zas, S. L. & Sweedler, J. V. Prediction of neuropeptide
1051 prohormone cleavages with application to RFamides. *Peptides* 27, 1087–1098 (2006).

1052 96. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldon, T. trimAl: a tool for automated
1053 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973
1054 (2009).

1055 97. Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. IQ-TREE: A fast and
1056 effective stochastic algorithm for estimating Maximum-Likelihood phylogenies. *Mol Biol*
1057 *Evol* 32, 268–274 (2015).

1058 98. Simakov, O. *et al.* Hemichordate genomes and deuterostome origins. *Nature* 527, 459–
1059 465 (2015).

1060 99. Cherif--Feildel, M., Berthelin, C. H., Rivière, G., Favrel, P. & Kellner, K. Data for
1061 evolutive analysis of insulin related peptides in bilaterian species. *Data Brief* 22, 546–550
1062 (2019).

1063 100. Thiel, D., Franz-Wachtel, M., Aguilera, F. & Hejnol, A. Changes in the neuropeptide
1064 complement correlate with nervous system architectures in xenacoelomorphs. 1–57 (2018)
1065 doi:10.1101/265579.

1066 101. Roch, G. J. & Sherwood, N. M. Glycoprotein hormones and their receptors emerged at
1067 the origin of metazoans. *Genome Biol Evol* 6, 1466–79 (2014).

1068 102. Oliveira, A. L. de, Calcino, A. & Wanninger, A. Ancient origins of arthropod moulting
1069 pathway components. *elife* 8, e46113 (2019).

1070 103. Smýkal, V. *et al.* Complex evolution of insect insulin receptors and homologous decoy
1071 receptors, and functional significance of their multiplicity. *Mol Biol Evol* 37, 1775–1789
1072 (2020).

1073 104. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for
1074 the analysis of massive data sets. *Nat Biotechnol* 35, 1026–1028 (2017).

1075 105. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene
1076 synteny and collinearity. *Nucleic Acids Research* 40, e49–e49 (2012).

1077 106. Warren, R. L. *et al.* LINKS: Scalable, alignment-free scaffolding of draft genomes with
1078 long reads. *Gigascience* 4, 35 (2015).