

1 **Insights into early animal evolution form the genome of the**  
2 **xenacoelomorph worm *Xenoturbella bocki***

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



8  
9 \*Corresponding authors: [p.schiffer@uni-koeln.de](mailto:p.schiffer@uni-koeln.de), [m.telford@ucl.ac.uk](mailto:m.telford@ucl.ac.uk)

10  
11 1 Center for Life's Origins and Evolution, Department of Genetics, Evolution and Environment,  
12 University College London, London WC1E 6BT, UK

13 2 worm-lab, Institute of Zoology, University of Cologne, 50674 Cologne, Germany

14 3 Department of Biosciences, Durham University, Durham DH1 3LE, UK

15 4 Université Côte D'Azur, CNRS, Inserm, iBV, Nice, France

16 5 The Arctic University Museum of Norway, UiT – The Arctic University of Norway, Tromsø, Norway

17 6 Collège Doctoral, Sorbonne Université, F-75005 Paris, France

18 7 Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo  
19 University Hospital, Oslo, Norway

20 8 Institute of Clinical Medicine, Medical Faculty, University of Oslo, Oslo, Norway

21 9 Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Invalidenstr. 43,  
22 10115 Berlin, Germany

23 10 University of Greifswald, Institute for Mathematics and Computer Science, Greifswald, Germany

24 11 University of Greifswald, Center for Functional Genomics of Microbes, Greifswald, Germany

25 12 Royal Brompton Hospital, Guy's and St Thomas' NHS Foundation Trust

26 13 Institut Pasteur, Université de Paris, CNRS UMR3525, Unité Régulation Spatiale des Génomes, F-  
27 75015 Paris, France

28 14 The University of Chicago, Division of Biological Sciences, Chicago, IL 60637, USA

29 15 Laboratoire de Biologie du Développement de Villefranche-sur-mer (LBDV), Sorbonne Université,  
30 CNRS, 06230 Villefranche-sur-mer, France

31 16 Department of Biochemistry, University of Oxford, Oxford, UK

32  
33 **Abstract**

34 The evolutionary origins of Bilateria remain enigmatic. One of the  
35 more enduring proposals highlights similarities between a  
36 cnidarian-like planula larva and simple acoel-like flatworms. This idea is based in  
37 part on the view of the Xenacoelomorpha as an outgroup to all other bilaterians  
38 which are themselves designated the Nephrozoa (protostomes and deuterostomes).  
39 Genome data can provide important comparative data and help to understand the  
40 evolution and biology of enigmatic species better. Here we assemble and analyse  
41 the genome of the simple, marine xenacoelomorph *Xenoturbella bocki*, a key  
42 species for our understanding of early bilaterian evolution. Our highly contiguous  
43 genome assembly of *X. bocki* has a size of ~111 Mbp in 18 chromosome like  
44 scaffolds, with repeat content and intron, exon and intergenic space comparable to

45 other bilaterian invertebrates. We find *X. bocki* to have a similar number of genes to  
46 other bilaterians and to have retained ancestral metazoan synteny. Key bilaterian  
47 signalling pathways are also largely complete and most bilaterian miRNAs are  
48 present. Overall, we conclude that *X. bocki* has a complex genome typical of  
49 bilaterians, which does not reflect the apparent simplicity of its body plan that has  
50 been so important to proposals that the Xenacoelomorpha are the simple sister  
51 group of the rest of the Bilateria.

52

## 53 **Introduction**

54 *Xenoturbella bocki* (Fig. 1) is a morphologically simple marine worm first described  
55 from specimens collected from muddy sediments in the Gullmarsfjord on the West  
56 coast of Sweden. There are now 6 described species of *Xenoturbella* - the only  
57 genus in the higher-level taxon of Xenoturbellida<sup>1</sup>. *X. bocki* was initially included as a  
58 species within the Platyhelminthes<sup>2</sup>, but molecular phylogenetic studies have shown  
59 that Xenoturbellida is the sister group of the Acoelomorpha, a second clade of  
60 morphologically simple worms also originally considered Platyhelminthes:  
61 Xenoturbellida and Acoelomorpha constitute their own phylum, the  
62 Xenacoelomorpha<sup>3,4</sup>. In addition to multiple phylogenetic studies that support the  
63 monophyly of the phylum, Xenacoelomorpha is convincingly supported by classical  
64 analysis in the field of evolution of development, for example their sharing unique  
65 amino acid signatures in their Caudal genes<sup>3</sup> and Hox4/5/6 gene<sup>5</sup>. Here we analyse  
66 our data in this phylogenetic framework of a monophyletic taxon.

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

76 *Xenoturbella bocki* has neither organized gonads nor a centralized nervous  
77 system. It has a blind gut, no body cavities and lacks nephrocytes<sup>11</sup>. If

78 Xenacoelomorpha is the sister group to Nephrozoa these character absences can be  
79 parsimoniously interpreted as representing the primitive state of the Bilateria.  
80 According to advocates of the Nephrozoa hypothesis, these and other characters  
81 absent in Xenacoelomorpha must have evolved in the lineage leading to Nephrozoa  
82 after the divergence of Xenacoelomorpha. More generally there has been a  
83 tendency to interpret Xenacoelomorpha (especially Acoelomorpha) as living  
84 approximations of Urbilateria<sup>12</sup>.

85 An alternative explanation for the simple body plan of xenacoelomorphs is  
86 that it is derived from that of more complex urbilaterian ancestors through loss of  
87 morphological characters. The loss or remodelling of morphological complexity is a  
88 common feature of evolution in many animal groups and is typically associated with  
89 unusual modes of living<sup>13,14</sup> – in particular the adoption of a sessile (sea squirts,  
90 barnacles) or parasitic (neodermatan flatworms, orthonectids) lifestyle, extreme  
91 miniaturization (e.g. tardigrades, orthonectids), or even neoteny (e.g. flightless  
92 hexapods).

93 The biology of *Xenoturbella* is difficult to study in vivo - they are hard to collect  
94 and mostly inactive in culture: knowledge of their embryology is restricted to one  
95 descriptive paper of a handful of embryos for example<sup>15</sup>. One route to better  
96 understanding the biology of this key taxon in the phylogeny of the animals is to read  
97 and study their genome.

98 In the past some genomic features gleaned from analysis of various  
99 Xenacoelomorpha have been used to test these evolutionary hypotheses. For  
100 example, the common ancestor of the protostomes and deuterostomes has been  
101 reconstructed with approximately 8 Hox genes but only 4 have been found in the  
102 Acoelomorpha (Nemertoderma) and 5 in *Xenoturbella*. This has been interpreted as  
103 a primary absence with the full complement of 8 appearing subsequent to the  
104 divergence of Xenacoelomorpha and Nephrozoa. Similarly, analysis of the  
105 microRNAs (miRNAs) of an acoelomorph, *Symsagittifera roscoffensis*, found that  
106 many bilaterian miRNAs were absent from its genome<sup>16</sup>. Some of the missing  
107 bilaterian miRNAs, however, were subsequently observed in *Xenoturbella*<sup>9</sup>.

108 The few xenacoelomorph genomes available to date are from the acoel  
109 *Hofstenia miamia*<sup>17</sup> – like other Acoelomorpha it shows accelerated sequence  
110 evolution relative to *Xenoturbella*<sup>3</sup> – and from two closely related species  
111 *Praesagittifera naikaiensis*<sup>18</sup> and *Symsagittifera roscoffensis*<sup>19</sup>. The analyses of gene

112 content of *Hofstenia* showed similar numbers of genes and gene families to other  
113 bilaterians<sup>17</sup>, while an analysis of the neuropeptide content concluded that most  
114 bilaterian neuropeptides were present in Xenacoelomorpha<sup>20</sup>.

115 In order to infer the characteristics of the ancestral xenacoelomorph genome,  
116 and to complement the data from the Acoelomorpha, we describe a highly-scaffolded  
117 genome of the slowly evolving xenacoelomorph *Xenoturbella bocki*. Our data allow  
118 us to contribute knowledge of Xenacoelomorpha and *Xenoturbella* in particular of  
119 genomic traits, such as gene content and genome-structure and to help reconstruct  
120 the genome structure and composition of the ancestral xenacoelomorph. Our data  
121 suggest that, while *Xenoturbella* is generally described as having a very simple body  
122 (interpreted by many as primitively simple), that its genome is of a similar complexity  
123 to many other bilaterians perhaps lending support to the idea that the simplicity of *X.*  
124 *bocki* is derived. animal.

125

## 126 **Results**

### 127 **Assembly of a draft genome of *Xenoturbella bocki*.**

128 We collected *Xenoturbella bocki* specimens (Fig. 1) from the bottom of the  
129 Gullmarsfjord close to the biological field station in Kristineberg (Sweden). These  
130 adult specimens were starved for several days in tubes with artificial sea water, and  
131 then sacrificed in lysis buffer. We extracted high molecular weight (HMW) DNA from  
132 single individuals for each of the different sequencing steps below.

133 We assembled a high-quality draft genome of *Xenoturbella bocki* using one  
134 short read Illumina library and one TruSeq Synthetic Long Reads (TSLR) Illumina  
135 library. We used a workflow based on a primary assembly with SPAdes (Methods;  
136 <sup>21</sup>). The primary assembly had an N50 of 8.5kb over 37,880 contigs with a maximum  
137 length of 206,709bp. After using the redundans pipeline<sup>22</sup> this increased to an N50 of  
138 ~62kb over 23,094 contigs and scaffolds spanning ~121Mb, and a longest scaffold of  
139 960,978kb (Table 1).

140 The final genome was obtained with Hi-C scaffolding using the program  
141 instaGRAAL<sup>23</sup>. The scaffolded genome has a span of 111 Mbp (117 Mbp including  
142 small fragments unincorporated into the HiC assembly) and an N50 of 2.7 Mbp (for  
143 contigs >500bp). The assembly contains 18 megabase-scale scaffolds  
144 encompassing 72 Mbp (62%) of the genomic sequence, with 43% GC content. The

145 original assembly indicated a repeat content of about 25% after a RepeatModeller  
146 based RepeatMasker annotation (Methods). As often seen in non-model organisms,  
147 about 2/3 of the repeats are not classified.

148

149 **Table 1.** *Improvement of assembly and scaffolding metrics.*

Assembly step	# seqs	# reals	# Ns	Max length	N50
redundans contigs	37,880	113,212,556	38,3327	206,709	8,544
redundans scaffolds	24,538	117,405,089	3,021,351	952,321	52,073
pre instaGRAAL	23,094	117,396,873	3,534,582	960,978	61,989
final scaffolds	27,939	107,712,917	3,328,069	8,757,424	2,730,651

150 Assessed with the jvci toolbox: <https://github.com/tanghaibao/jvci.git>.

151

152 We used BRAKER1<sup>24,25</sup> with extensive RNA-Seq data, and additional single-  
153 cell UTR enriched transcriptome sequencing data to predict 15,154 gene models.  
154 9,575 gene models (63%) are found on the 18 large scaffolds (which represent 62%  
155 of the total sequence). 13,298 of our predicted genes (88%) have RNA-Seq support.  
156 Although this proportion is at the low end of bilaterian gene counts, we note that our  
157 RNA-seq libraries were all taken from adult animals and thus may not represent the  
158 true complexity of the gene complement. We consider our predicted gene number to  
159 be a lower bound estimate for the true gene content.

160 The predicted *X. bocki* genes have a median coding length of 873 nt and a  
161 mean length of 1330 nt. Median exon length is 132 nt (mean 212 nt) and median  
162 intron length is 131 nt (mean 394 nt). Genes have a median of 4 exons and a mean  
163 of 8.5 exons. 2,532 genes have a single exon and, of these, 1,381 are supported as  
164 having a single exon by RNA-Seq (TPM>1). A comparison of the exon, intron, and  
165 intergenic sequence content in *Xenoturbella* with those described in other animal  
166 genomes<sup>26</sup> show that *X. bocki* falls within the range of other similarly sized metazoan  
167 genomes (Fig. 2) for all these measures.

168

### 169 The genome of a co-sequenced *Chlamydia* species

170 We recovered the genome of a marine *Chlamydia* species from Illumina data  
171 obtained from one *X. bocki* specimen and from Oxford Nanopore data from a second  
172 specimen supporting previous microscopic analyses and single gene PCRs  
173 suggesting that *X. bocki* is host to a species in the bacterial genus *Chlamydia*. The  
174 bacterial genome was found as 5 contigs spanning 1,906,303 bp (N50 of 1,237,287  
175 bp) which were assembled into 2 large scaffolds. Using PROKKA<sup>27</sup>, we predicted  
176 1,738 genes in this bacterial genome, with 3 ribosomal RNAs, 35 transfer RNAs, and

177 1 transfer-messenger RNA. The genome is 97.5% complete for bacterial BUSCO<sup>28</sup>  
178 genes, missing only one of the 40 core genes.

179 Marine chlamydiae are not closely related to the group of human pathogens<sup>29</sup>  
180 and we were not able to align the genome of the *Chlamydia*-related symbiont from  
181 *X. bocki* to the reference strain *Chlamydia trachomatis* F/SW4, nor to *Chlamydophila*  
182 *pneumoniae* TW-183. To investigate the phylogenetic position of the species co-  
183 occurring with *Xenoturbella*, we aligned the 16S rRNA gene from the *X. bocki*-hosted  
184 *Chlamydia* with orthologs from related species including sequences of genes  
185 amplified from DNA/RNA extracted from deep sea sediments. The *X. bocki*-hosted  
186 *Chlamydia* belong to a group designated as Simkaniaceae in<sup>29</sup>, with the sister taxon  
187 in our phylogenetic tree being the *Chlamydia* species previously found in *X.*  
188 *westbladi* (*X. westbladi* is almost certainly a synonym of *X. bocki*)<sup>7</sup> (Fig. 3).

189 To investigate whether the *X. bocki*-hosted *Chlamydia* might contribute to the  
190 metabolic pathways of its host, we compared the completeness of metabolic  
191 pathways in KEGG for the *X. bocki* genome alone and for the *X. bocki* genome in  
192 combination with the bacteria. We found only slightly higher completeness in a small  
193 number of pathways involved in carbohydrate metabolism, carbon fixation, and  
194 amino acid metabolism (see supplementary material) suggesting that the relationship  
195 is likely to be commensal or parasitic rather than a true symbiosis.

196 A second large fraction of bacterial reads, annotated as Gammaproteobacteria,  
197 were identified and filtered out during the data processing steps. These bacteria  
198 were also previously reported as potential symbionts of *X. bocki*<sup>30</sup>. However, these  
199 sequences were not sufficiently well covered to reconstruct a genome and we did not  
200 investigate them further.

201

#### 202 HGT into the *X. bocki* genome is low

203 Given the close association with bacteria we were curious to see whether the *X.*  
204 *bocki* genome contains an elevated number of horizontally acquired genes. We did  
205 not find this to be the case. We were able to detect 56 potential horizontal gene  
206 transfer (HGT) events. Phylogenies generated using closest blast hits for each HGT  
207 candidate unveiled one of the 56 genes to be of chlamydial origin and thus likely  
208 originating from a bacterial contig. A number of the HGT candidates appear to be of  
209 Proteobacteria origin, coding for a functionally diverse set of proteins. In summary,

210 0.35% of the *X. bocki* genes we have identified might be horizontally acquired. See  
211 supplementary online material for alignments and gene trees.

212

### 213 **A phylogenetic gene presence/absence matrix supports Xenambulacraria**

214 The general completeness of the *X. bocki* gene set allowed us to use the presence  
215 and absence of genes identified in our genomes as a source of information to find  
216 the best supported phylogenetic position of the Xenacoelomorpha. We conducted  
217 two separate phylogenetic analyses of gene presence/absence data: one including  
218 the fast-evolving Acoelomorpha and one without. In both analyses the best tree  
219 grouped *Xenoturbella* with the Ambulacraria (Fig. 4a). The analysis including acoels,  
220 however, placed the acoels as the sister-group to Nephrozoa separate from  
221 *Xenoturbella* (Fig. 4b). There are two explanations for this finding. The first would be  
222 that the Xenacoelomorphs are paraphyletic; that *Xenoturbella* is the sister group of  
223 the Ambulacraria and Acoelomorpha the sister group of Nephrozoa. Because many  
224 other studies have shown the monophyly of Xenacoelomorpha to be robust<sup>3,4,7-10,31</sup>,  
225 we do not think this a plausible explanation. The second explanation of this  
226 observation is that it is the result of systematic error caused by a high rate of gene  
227 loss or by orthologs being incorrectly scored as missing due to higher rates of  
228 sequence evolution in acoelomorphs<sup>32</sup>. Under this second scenario, we consider it  
229 more likely that, of the two clades, it is the Acoelomorpha not *Xenoturbella* that are  
230 wrongly placed and that the position of *Xenoturbella* represents the more likely  
231 position of the entire phylum of Xenacoelomorpha. We note that under both  
232 scenarios, the focus of our work, *Xenoturbella*, is the sistergroup of the Ambulacraria  
233 though the implied error suggests that using gene presence/absence may not be the  
234 ideal way to solve difficult phylogenetic problems.

235

### 236 **The *X. bocki* molecular toolkit is typical of bilaterians.**

237 One of our principal aims was to ask whether the *Xenoturbella* genome lacks  
238 characteristics otherwise present in the Bilateria. We found that for the Metazoa  
239 gene set in BUSCO (v5) the *X. bocki* proteome translated from our gene predictions  
240 is 82.5% complete and ~90% complete when partial hits are included (82% and 93%  
241 respectively for the Eukaryote gene set). This estimate is even higher in the acoel  
242 *Hofstenia miamia*, which was originally reported to be 90%<sup>17</sup>, but in our re-analysis  
243 was 95.71%. In comparison, the morphologically highly simplified and fast evolving

244 annelid *Intoshia linei*<sup>33</sup> has a genome of fewer than 10,000 genes<sup>34</sup> and in our  
245 analysis is only ~64% complete for the BUSCO (v5) Metazoa set. The model  
246 nematode *Caenorhabditis elegans* is ~79% complete for the same set. Despite the  
247 morphological simplicity of both *Xenoturbella*, and *Hofstenia*, these  
248 Xenacoelomorpha are missing few core genes compared to other bilaterian lineages  
249 that we perceive to have undergone a high degree of morphological evolutionary  
250 change (such as the evolution of miniaturisation, parasitism, sessility etc).

251 Using our phylogenomic matrix of gene presence/absence (see above) we identified  
252 all orthologs that could be detected both in Bilateria (in any bilaterian lineage) and in  
253 any non-bilaterian; ignoring horizontal gene transfer and other rare events, these  
254 genes must have existed in Urbilateria (and, of less interest to us, in Urmetazoa).  
255 The absence of any of these bilaterian genes in any lineage of Bilateria must  
256 therefore be explained by loss of the gene. All individual bilaterian genomes were  
257 missing many of these orthologs but Xenacoelomorphs and some other bilaterians  
258 lacked more of these than did other taxa. The average numbers of these genes  
259 present in bilaterians = 7577; *Xenoturbella* = 5459; *Hofstenia* = 5438; *Praesagittifera*  
260 = 4280; *Drosophila* = 4844; *Caenorhabditis* = 4323.

261 To better profile the *Xenoturbella* and xenacoelomorph molecular toolkit, we  
262 used OrthoFinder to conduct orthology searches in a comparison of 155 metazoan  
263 and outgroup species, including the transcriptomes of the sister species *X. profunda*  
264 and an early draft genome of the acoel *Paratomella rubra* we had available, as well  
265 as the *Hofstenia* and *Praesagittifera* proteomes (Supplementary File 1). For each  
266 species we counted, in each of the three Xenacoelomorphs, the number of  
267 orthogroups for which a gene was present. The proportion of orthogroups containing  
268 an *X. bocki* and *X. profunda* protein (87.4% and 89.2%) are broadly similar to the  
269 proportions seen in other well characterised genomes, for example *S. purpuratus*  
270 proteins (93.8%) or *N. vectensis* proteins (84.3%) (Fig. 5). In this analysis, the fast-  
271 evolving nematode *Caenorhabditis elegans* appears as an outlier, with only ~64% of  
272 its proteins in orthogroups and ~35% unassigned. Both *Xenoturbella* species have  
273 an intermediate number of unassigned genes of ~11-12%. Similarly, the proportion  
274 of species-specific genes (~14% of all genes) corresponds closely to what is seen in  
275 most other species (with the exception of the parasitic annelid *I. linei*, Fig. 5).

276

277 Idiosyncrasies of *Xenoturbella*



278 In order to identify sets of orthologs specific to the two *Xenoturbella* species we used  
279 the kinfin software<sup>35</sup> and found 867 such groups in the OrthoFinder clustering. We  
280 profiled these genes based on Pfam domains and GO terms derived from  
281 InterProScan. While these *Xenoturbella* specific proteins fall into diverse classes, we  
282 did see a considerable number of C-type lectin, Immunoglobulin-like, PAN, and  
283 Kringle domain containing Pfam annotations. Along with the Cysteine-rich secretory  
284 protein family and the G-protein coupled receptor activity GO terms, these genes  
285 and families of genes may be interesting for future studies into the biology of  
286 *Xenoturbella* in its native environment.

287

#### 288 Gene families and signaling pathways are retained in *X. bocki*

289 In our orthology clustering we did not see an inflation of *Xenoturbella*-specific groups  
290 in comparison to other taxa, but also no conspicuous absence of major gene families  
291 (Fig. 6). Family numbers of transcription factors like Zinc-fingers or homeobox-  
292 containing genes, as well as, for example, NACHT-domain encoding genes seem to  
293 be neither drastically inflated nor contracted in comparison to other species in our  
294 InterProScan based analysis.

295 To catalogue the completeness of cell signalling pathways we screened the *X.*  
296 *bocki* proteome against KEGG pathway maps using GenomeMaple<sup>36</sup>. The *X. bocki*  
297 gene set is largely complete in regard to the core proteins of these pathways, while  
298 an array of effector proteins is absent (Fig. 6). In comparison to other metazoan  
299 species, as well as a unicellular choanoflagellate and a yeast, the *X. bocki* molecular  
300 toolkit has significantly lower KEGG completeness than morphologically complex  
301 animals such as the sea urchin and amphioxus (t-test; Fig. 6). *Xenoturbella* is,  
302 however, not significantly less complete when compared to other bilaterians  
303 considered to have low morphological complexity and which have been shown to  
304 have reduced gene content, such as *C. elegans*, the annelid parasite *Intoshia linei*,  
305 or the acoel *Hofstenia miamia* (Fig. 6).

#### 306 Clustered homeobox genes in the *X. bocki* genome

307 Acoelomorph flatworms possess three unlinked HOX genes, orthologs of anterior  
308 (Hox1), central (Hox4/5 or Hox5) and posterior Hox (HoxP). In contrast, previous  
309 analysis of *X. bocki* transcriptomes identified one anterior, three central and one  
310 posterior Hox genes. We identified clear evidence of a syntenic Hox cluster with four  
311 Hox genes (centHox1, postHox, centHox3, and antHox1) in the *X. bocki* genome

312 (Fig. 7). There was also evidence of a fragmented annotation of centHox2, split  
313 between the 4 gene Hox cluster and a separate scaffold (Fig. 7). In summary, this  
314 suggests that all five Hox genes form a Hox cluster in the *X. bocki* genome, but that  
315 there are possible unresolved assembly errors disrupting the current annotation. We  
316 also identified other homeobox genes on the Hox cluster scaffold, including Evx (Fig.  
317 7).

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

325 Hemichordates and chordates have a conserved cluster of genes involved in  
326 patterning their pharyngeal pores - the so-called 'pharyngeal cluster'. The homeobox  
327 genes of this cluster (Msx1x, Nk2-1/2/4/8) were present on a single *X. bocki* scaffold.  
328 Another pharyngeal cluster transcription factor, the Forkhead containing Foxa, and  
329 'bystander' genes from that cluster including EglN, Mipol1 and Slc25a21 are found in  
330 the same genomic region. Different sub-parts of the cluster are found in non-  
331 bilaterians and protostomes and the cluster may well be plesiomorphic for the  
332 Bilateria rather than a deuterostome synapomorphy<sup>37</sup>.

333

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

335 A catalogue of acoelomorph neuropeptides was previously described using  
336 transcriptome data<sup>38</sup>. We have discovered 12 additional neuropeptide genes and 39  
337 new neuropeptide receptors in *X. bocki* adding 6 bilaterian peptidergic systems to  
338 the *Xenoturbella* catalogue (NPY-F ; MCH/Asta-C ; TRH ; ETH ; CCHa/Nmn-B ; Np-  
339 S/CCAP), and 6 additional bilaterian systems to the Xenacoelomorpha catalogue  
340 (Corazonin ; Kiss/GPR54 ; GPR83 ; 7B2 ; Trunk/PTTH ; NUCB2) making a total of  
341 31 peptidergic systems (Fig. 8).

342 Among the ligand genes, we identified 6 new repeat-containing sequences.  
343 One of these, the LRIGamide-peptide, had been identified in Nemertodermatida and  
344 Acoela and its loss in *Xenoturbella* had been proposed<sup>38</sup>. We also identified the first  
345 7B2 neuropeptide and NucB2/Nesfatin genes in Xenacoelomorpha. Finally, we

346 identified 3 new *X. bocki* insulin-like peptides, one of them sharing sequence  
347 similarity and an atypical cysteine pattern with the Ambulacrarian octinsulin,  
348 constituting a potential synapomorphy of Xenambulacraria (see Supplementary).

349 Our searches also revealed the presence of components of the arthropod  
350 moulting pathway components (PTTH/trunk, NP-S/CCAP and Bursicon receptors),  
351 which have recently been shown to be of ancient origin (de Oliveira et al., 2019). We  
352 further identified multiple paralogs for, e.g the Tachykinin, Rya/Luquin, tFMRFa,  
353 Corazonin, Achatin, CCK, and Prokineticin receptor families. Two complete *X. bocki*  
354 Prokineticin ligands were also found in our survey (Fig. 8).

355 Chordate Prokineticin ligands possess a conserved N-terminal “AVIT”  
356 sequence required for the receptor activation<sup>39</sup>. This sequence is absent in  
357 arthropod Astakine, which instead possess two signature sequences within their  
358 Prokineticin domain<sup>40</sup>. To investigate Prokineticin ligands in Xenacoelomorpha we  
359 compared the sequences of their prokineticin ligands with those of other bilaterians  
360 (Fig. 8). Our alignment reveals clade specific signatures already reported in  
361 Ecdysozoa and Chordata sequences, but also two new signatures specific to  
362 Lophotrochozoa and Cnidaria sequences, as well as a very specific “K/R-RFP-K/R”  
363 signature shared only by ambulacrarian and *Xenoturbella bocki* sequences. The  
364 shared Ambulacrarian/Xenacoelomorpha signature is found at the same position as  
365 the Chordate sequence involved in receptor activation - adjacent to the N-terminus of  
366 the Prokineticin domain (Fig. 8).

367

368 **The *X. bocki* genome contains most bilaterian miRNAs reported missing from**  
369 **acoels.**

370 microRNAs have previously been used to investigate the phylogenetic position of the  
371 acoels and *Xenoturbella*. The acoel *Symsagittifera roscoffensis* lacks protostome  
372 and bilaterian miRNAs and this lack was interpreted as supporting the position of  
373 acoels as sister-group to the Nephrozoa. Based on shallow 454 microRNA  
374 sequencing (and sparse genomic traces) of *Xenoturbella*, some of the bilaterian  
375 miRNAs missing from acoels were found - 16 of the 32 expected metazoan (1  
376 miRNA) and bilaterian (31 miRNAs) microRNA families – of which 6 could be  
377 identified in genome traces<sup>9</sup>.

378 By deep sequencing two independent small RNA samples, we have now  
379 identified the majority of the missing metazoan and bilaterian microRNAs and

380 identified them in the genome assembly (Fig. 9). Altogether, we found 23 out of 31  
381 bilaterian microRNA families (35 genes including duplicates); the single known  
382 Metazoan microRNA family (MIR-10) in 2 copies; the Deuterostome-specific MIR-  
383 103; and 7 *Xenoturbella*-specific microRNAs giving a total of 46 microRNA genes.  
384 None of the protostome-specific miRNAs were found. We could not confirm in the  
385 RNA sequences or new assembly a previously identified, and supposedly  
386 xenambulacrarian-specific MIR-2012 ortholog.

387

### 388 **The *X. bocki* genome retains ancestral metazoan linkage groups.**

389 The availability of chromosome-scale genomes has made it possible to reconstruct  
390 24 ancestral linkage units broadly preserved in bilaterians<sup>41</sup>. In fast-evolving  
391 genomes, such as those of nematodes, tunicates or platyhelminths, these ancestral  
392 linkage groups (ALGs) are often dispersed and/or extensively fused  
393 (Supplementary). We were interested to test if the general conservation of the gene  
394 content in *X. bocki* is reflected in its genome structure.

395 We compared the genome of *Xenoturbella* to several other metazoan  
396 genomes and found that it has retained most of these ancestral bilaterian units: 12  
397 chromosomes in the *X. bocki* genome derive from a single ALG, five chromosomes  
398 are made of the fusion of two ALGs, and one *Xenoturbella* chromosome is a fusion  
399 of three ALGs, as highlighted with the comparison of ortholog content with  
400 amphioxus, the sea urchin and the sea scallop (Fig. 10 and Supplementary).

401 One ancestral linkage group that has been lost in chordates but not in  
402 ambulacrarians nor in molluscs (ALG R in sea urchin and sea scallop) is detectable  
403 in *X. bocki* (Fig. 10), while *X. bocki* does not show the fusions that are characteristic  
404 of lophotrochozoans.

405 We also attempted to detect some pre-bilaterian arrangement of ancestral  
406 linkage: for instance, ref <sup>42</sup> predicted that several pre-bilaterian linkage groups  
407 successively fused in the bilaterian lineage to give ALGs A1, Q and E. These ALGs  
408 are all represented as single units in *X. bocki* in common with other Bilateria. None  
409 of the inferred pre-bilaterian chromosomal arrangements that could have provided  
410 support for the Nephrozoa hypothesis were found *in X. bocki* although of course this  
411 does not rule out Nephrozoa.

412

413 One *X. bocki* chromosomal fragment appears aberrant

414 The smallest of the 18 large scaffolds in the *X. bocki* genome did not show strong  
415 1:1 clustering with any scaffold/chromosome of the bilaterian species we compared it  
416 to. To exclude potential contamination in the assembly as a source for this contig we  
417 examined the orthogroups to which the genes from this scaffold belong. We found  
418 that *Xenoturbella profunda*<sup>43</sup>, for which a transcriptome is available, was the species  
419 that most often occurred in the same orthogroup with genes from this scaffold (41  
420 shared orthogroups), suggesting the scaffold is not a contaminant.

421 We did observe links between the aberrant scaffold and several scaffolds  
422 from the genome of the sponge *E. muelleri* in regard to synteny, but could not detect  
423 distinct synteny relationships to a single scaffold in another species. In line with this,  
424 genes on the scaffold show a different age structure compared to other scaffolds,  
425 with both more older genes (pre bilaterian) and more *Xenoturbella* specific genes  
426 (Fig. 11; supported by Ks statistics, Supplementary). This aberrant scaffold also had  
427 significantly lower levels of methylation than the rest of the genome.

428

## 429 **DISCUSSION**

430 The phylogenetic positions of *Xenoturbella* and the Acoelomorpha have been  
431 controversial since the first molecular data from these species appeared over twenty  
432 five years ago. Today we understand that they constitute a monophyletic group of  
433 morphologically simple worms<sup>1,9,44</sup>, but there remains a disagreement over whether  
434 they represent a secondarily simplified sister group of the Ambulacraria or a  
435 primitively simple sister group to all other Bilateria. Here we wanted to analyse the  
436 genome of *Xenoturbella* to glean insights into their biology from a new perspective.

437 Previous analyses of the content of genomes, especially of Acoela, have found  
438 a small number of Hox genes and of microRNAs of acoels and this has been  
439 interpreted as representing an intermediate stage on the path to the ~8 Hox genes  
440 and 30 odd microRNAs of the Nephrozoa. A strong version of the Nephrozoa idea  
441 would go further than these examples and anticipate, for example, a genome-wide  
442 paucity of bilaterian genes, GRNs and biochemical pathways and/or an arrangement  
443 of chromosomal segments intermediate between those of the Eumetazoa and the  
444 Nephrozoa.

445 One criticism of the results from analyses of acoel genomes is that the  
446 Acoelomorpha have evolved rapidly (their long branches in phylogenetic trees  
447 showing high rates of sequence change). This rapid evolution might plausibly be

448 expected to correlate with other aspects of rapid genome evolution such as higher  
449 rates of gene loss and chromosomal rearrangements leading to significant  
450 differences from other Bilateria. The more normal rates of sequence evolution  
451 observed in *Xenoturbella* therefore recommend it as a more appropriate  
452 xenacoelomorph to study with fewer apomorphic characters expected.

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

459 We found the *X. bocki* genome to be fairly compact, but not unusually reduced  
460 in size compared to many other bilaterians. It appears to contain a similar number of  
461 genes (~15,000) as other animals, for example from the model organisms *D.*  
462 *melanogaster* (>14,000) and *C. elegans* (~20,000). The BUSCO completeness, as  
463 well as a high level of representation of *X. bocki* proteins in the orthogroups of our  
464 155 species orthology screen indicates that we have annotated a near complete  
465 gene set. Surprisingly, there are fewer genes than in the acoel *Hofstenia* (>22,000;  
466 BUSCO\_v5 score ~95%). This said, of the genes found in Urbilateria (orthogroups in  
467 our presence/absence analysis containing a member from both a bilaterian and an  
468 outgroup) *Xenoturbella* and *Hofstenia* have very similar numbers (5459 and 5438  
469 respectively). Gene, intron and exon lengths all also fall within the range seen in  
470 many other invertebrate species<sup>26</sup>. It thus seems that basic genomic features in  
471 *Xenoturbella* are not anomalous among Bilateria. Unlike some extremely simplified  
472 animals, such as orthonectids, we observe no extreme reduction in gene content.

473 All classes of homeodomain transcription factors have previously been reported  
474 to exist in Xenacoelomorpha<sup>45</sup>. We have identified 5 HOX-genes in *X. bocki* and at  
475 least four, and probably all five of these are on one chromosomal scaffold within 187  
476 Kbp. *X. bocki* also has the parahox genes *Gsx* and *Cdx*; while *Xlox/pdx* is not found,  
477 it is present in Cnidarians and must therefore have been lost<sup>46</sup>. If block duplication  
478 models of Hox and Parahox evolutionary relationships are correct, the presence of a  
479 complete set of parahox genes implies the existence of their Hox paralogs in the  
480 ancestor of Xenacoelomorphs suggesting the xenacoelomorph ancestor also  
481 possessed a Hox 3 ortholog. If anthozoans also have an ortholog of bilaterian Hox

482 2<sup>47</sup>, this must also have been lost from Xenacoelomorphs. The minimal number of  
483 Hox genes in the xenacoelomorph stem lineage was therefore probably 7 (AntHox1,  
484 lost Hox2, lost Hox 3, CentHox 1, CentHox 2, CentHox 3 and postHoxP).

485 Based on early sequencing technology and without a reference genome  
486 available, it was thought that Acoelomorpha lack many bilaterian microRNAs. Using  
487 deep sequencing of small RNAs and our high-quality genome, we have shown that  
488 *Xenoturbella* shows a near-complete bilaterian set of miRNAs including the single  
489 deuterostome-specific miRNA family (MIR-103) (Fig. 9). The low number of  
490 differential family losses of *Xenoturbella* (8 of 31 bilaterian miRNA families) inferred  
491 is the same as the number lost in the flatworm *Schmidtea*, and substantially lower  
492 than the number lost in the rotifer *Brachionus* (which has lost 14 bilaterian families).  
493 It is worth mentioning that *X. bocki* shares the absence of one miRNA family (MIR-  
494 216) with all Ambulacrarians although if Deuterostomia are paraphyletic this could be  
495 interpretable as a primitive state<sup>37</sup>.

496 The last decade has seen a re-evaluation of our understanding of the evolution  
497 of the neuropeptide signaling genes<sup>48,49</sup>. The peptidergic systems are thought to  
498 have undergone a diversification that produced approximately 30 systems in the  
499 bilaterian common ancestor<sup>48,49</sup>. Our study identified 31 neuropeptide systems in  
500 *X.bocki* and for all of these either the ligand, receptor, or both are also present in  
501 both protostomes and deuterostomes indicating conservation across Bilateria. It is  
502 likely that more ligands (which are short and variable) remain to be found with better  
503 detection methods. It appears that the *Xenoturbella* genome contains a nearly  
504 complete bilaterian neuropeptide complement with no signs of simplification but  
505 rather signs of expansions of certain gene families. Our analyses also reveal a  
506 potential synapomorphy linking Xenacoelomorpha with Ambulacraria (Fig. 8 and  
507 Supplementary).

508 We have used the predicted presence and absence of genes across a selection  
509 of metazoan genomes as characters for phylogenetic analyses. Our trees re-confirm  
510 the findings of recent phylogenomic gene alignment studies in linking *Xenoturbella* to  
511 the Ambulacraria. We also used these data to test different bilaterians for their  
512 propensity to lose otherwise conserved genes (or for our inability to identify  
513 orthologs<sup>32</sup>). While the degree of gene loss appears similar between *Xenoturbella*  
514 and acoels, the phylogenetic analysis shows longer branches leading to the acoels,  
515 most likely due to faster evolution, gain of lineages specific genes, and some degree

516 of gene loss in the branch leading to the Acoelomorpha. Recent work has shown the  
517 tendency of rapidly evolving genes (in particular those belonging to rapidly evolving  
518 species) to be missed by orthology detection software<sup>50,51</sup>.

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

528 One of the chromosome-scale scaffolds in our assembly showed a different  
529 methylation and age signal, with both older and younger genes, and no clear  
530 relationship to metazoan linkage groups. By analyzing orthogroups of genes on this  
531 scaffold for their phylogenetic signal and finding *X. bocki* genes to cluster with those  
532 of *X. profunda* we concluded that the scaffold most likely does not represent a  
533 contamination. It remains unclear whether this scaffold is a fast-evolving  
534 chromosome, or a chromosomal fragment or arm. Very fast evolution on a  
535 chromosomal arm has for example been shown in the zebrafish<sup>52</sup>.

536 Apart from DNA from *X. bocki* we also obtained a highly contiguous genome  
537 of a species related to marine *Chlamydia* species (known from microscopy to exist in  
538 *X. bocki*); a symbiotic relationship between *Xenoturbella* and the bacteria has been  
539 thought possible<sup>53,54</sup>. The large gene number and the completeness of genetic  
540 pathways we found in the chlamydial genome do not support an endosymbiotic  
541 relationship.

542 Overall, we have shown that, while *Xenoturbella* has lost some genes - in  
543 addition to the reduced number of Hox genes previously noted, we observe a  
544 reduction of some signaling pathways to the core components - in general, the *X.*  
545 *bocki* genome is not strikingly simpler than many other bilaterian genomes. We do  
546 not find support for a strong version of the Nephrozoa hypothesis which would  
547 predict many missing bilaterian genes. Bilaterian Hox and microRNA absent from  
548 Acoelomorpha are found in *Xenoturbella* eliminating the impact of two character  
549 types that were previously cited in support of Nephrozoa. The *Xenoturbella* genome



550 has also largely retained the ancestral linkage groups found in other bilaterians and  
551 does not represent a structure intermediate between Eumetazoan and bilaterian  
552 ground states. Overall, while we can rule out a strong version of the Nephrozoa  
553 hypothesis with many Bilaterian characteristics missing in xenacoelomorphs, our  
554 analysis of the *Xenoturbella* genome cannot distinguish between a weak version of  
555 Nephrozoa and the Xenambulacraria topology.

556

## 557 **Methods**

### 558 **Genome Sequencing, Assembly, and Scaffolding**

559 We extracted DNA from individual *Xenoturbella* specimens with a standard and  
560 additionally worked with a Phenol-Chloroform protocol specifically developed to  
561 extract HMW DNA ([dx.doi.org/10.17504/protocols.io.mrxc57n](https://doi.org/10.17504/protocols.io.mrxc57n)). The extracted DNA  
562 was quality controlled with a Nanodrop instrument in our laboratory and  
563 subsequently a TapeStation at the sequencing center. Worms were first starved and  
564 kept in repeatedly replaced salt water, reducing the likelihood of food or other  
565 contaminants in the DNA extractions. First, we sequenced Illumina short paired-end  
566 reads and mate pair libraries (see ref <sup>3</sup> for details). As the initial paired-read datasets  
567 were of low complexity and coverage, we later complemented these data with an  
568 Illumina HiSeq 2000/2500 series paired-end dataset with ~700 bp insert size and  
569 250bp read lengths, yielding ~354 Million reads. Additionally, we generated ~40  
570 Million Illumina TruSeq Synthetic Long Reads (TSLR) for high confidence primary  
571 scaffolding.

572 After read cleaning with Trimmomatic v.0.38<sup>55</sup> we conducted initial test  
573 assemblies using the clc assembly cell v.5 and ran the blobtools pipeline<sup>56</sup> to screen  
574 for contamination (Figure 12). Not detecting any significant numbers of reads from  
575 suspicious sources in the HiSeq dataset we used SPAdes v. 3.9.0<sup>21</sup> to correct and  
576 assemble a first draft genome. We also tried to use dipSPAdes but found the runtime  
577 to exceed several weeks without finishing. We submitted the SPAdes assembly to  
578 the redundans pipeline to eliminate duplicate contigs and to scaffold with all available  
579 mate pair libraries. The resulting assembly was then further scaffolded with the aid of  
580 assembled transcripts (see below) in the BADGER pipeline<sup>57</sup>. In this way we were  
581 able to obtain a draft genome with ~60kb N50 that could be scaffolded to  
582 chromosome scale super-scaffolds with the use of 3C data.

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

593 Library preparation for genome-wide bisulfite sequencing was performed as  
594 previously described<sup>58</sup>. The resulting sequencing data were aligned to the *X. bocki*  
595 draft genome using Bismark in non-directional mode to identify the percentage  
596 methylation at each cytosine genome-wide. Only sites with >10 reads mapping were  
597 considered for further analysis.

598

#### 599 Preparation of the Hi-C libraries

600 The Hi-C protocol was adapted at the time from (Lieberman-Aiden et al., 2009;  
601 Sexton et al., 2012 and Marie-Nelly et al., 2014). Briefly, an animal was chemically  
602 cross-linked for one hour at room temperature in 30 mL of PBS 1X added with 3%  
603 formaldehyde (Sigma – F8775 - 4x25 mL). Formaldehyde was quenched for 20 min  
604 at RT by adding 10 ml of 2.5 M glycine. The fixed animal was recovered through  
605 centrifugation and stored at -80°C until use. To prepare the proximity ligation library,  
606 the animal was transferred to a VK05 Precellys tubes in 1X DpnII buffer (New  
607 England Biolabs; 0.5mL) and the tissues were disrupted using the Precellys  
608 Evolution homogenizer (Bertin-Instrument). SDS was added (0.3% final) to the lysate  
609 and the tubes were incubated at 65°C for 20 minutes followed by an incubation at  
610 37°C for 30 minutes and an incubation of 30 minutes after adding 50 µL of 20%  
611 triton-X100. 150 units of the DpnII restriction enzyme were then added and the tubes  
612 were incubated overnight at 37°C. The endonuclease was inactivated 20 min at  
613 65°C and the tubes were then centrifuged at 16,000 x g during 20 minutes,  
614 supernatant was discarded and pellets were re-suspended in 200 µl NE2 1X buffer  
615 and pooled. DNA ends were labeled using 50 µl NE2 10X buffer, 37.5 µl 0.4 mM  
616 dCTP-14-biotin, 4.5 µl 10mM dATP-dGTP-dTTP mix, 10 µl klenow 5 U/µL and

617 incubation at 37°C for 45 minutes. The labeling mix was then transferred to ligation  
618 reaction tubes (1.6 ml ligation buffer; 160 µl ATP 100 mM; 160 µl BSA 10 mg/mL; 50  
619 µl T4 DNA ligase (New England Biolabs, 5U/µl); 13.8 ml H<sub>2</sub>O) and incubated at 16°C  
620 for 4 hours. A proteinase K mix was added to each tube and incubated overnight at  
621 65°C. DNA was then extracted, purified and processed for sequencing as previously  
622 described<sup>23</sup>. Hi-C libraries were sequenced on a NextSeq 500 (2 × 75 bp, paired-end  
623 using custom made oligonucleotides as in Marie-Nelly et al., 2014). Libraries were  
624 prepared separately on two individuals in this way but eventually merged. Note that  
625 more recent version of the Hi-C protocol than the one used here have been  
626 described elsewhere<sup>59</sup>.

627

### 628 InstaGRAAL assembly pre-processing

629 The primary Illumina assembly contains a number of very short contigs, which are  
630 disruptive when computing the contact distribution needed for the instaGRAAL  
631 proximity ligation scaffolding (pre-release version, see<sup>60</sup> and<sup>23</sup> for details). Testing  
632 several Nx metrics we found a relative length threshold, that depends on the  
633 scaffolds' length distribution, to be a good compromise between the need for a low-  
634 noise contact distribution and the aim of connecting most of the genome. We found  
635 N90 a suitable threshold and excluded contigs below 1,308 bp. This also ensured no  
636 scaffolds shorter than three times the average length of a DpnII restriction fragment  
637 (RF) were in the assembly. In this way every contig contained enough RFs for  
638 binning and were included in the scaffolding step.

639 Reads from both libraries were aligned with bowtie2 (v. 2.2.5)<sup>61</sup> against the  
640 DpnII RFs of the reference assembly using the hicstuff pipeline  
641 (<https://github.com/koszullab/hicstuff>) and in paired-end mode (with the options: -fg-  
642 maxins 5 -fg-very-sensitive-local), with a mapping quality >30. The pre-processed  
643 genome was reassembled using instaGRAAL. Briefly, the program uses a Markov  
644 Chain Monte Carlo (MCMC) method that samples DNA segments (or bins) of the  
645 assembly for their best relative 1D positions with respect to each other. The quality  
646 of the positions is assessed by fitting the contact data first on a simple polymer  
647 model, then on the plot of contact frequency according to the genomic distance law  
648 computed from the data. The best relative position of a DNA segment with respect to  
649 one of its most likely neighbours consists in operations such as flips, swaps, merges  
650 or a split of contigs. Each operation is either accepted or rejected based on the

651 computed likelihood, resulting in an iterative progression toward the 1D structure that  
652 best fits the contact data. Once the entire set of DNA segments is sampled for  
653 position (i.e. a cycle), the process starts over. The scaffolder was run independently  
654 for 50 cycles, long enough for the chromosome structure to converge. The  
655 corresponding genome is then considered stable and suitable for further analyses  
656 (Figure 13). The scaffolded assemblies were then refined using instaGRAAL's  
657 instaPolish module, to correct small artefactual inversions that are sometimes a  
658 byproduct of instaGRAAL's processing.

659

## 660 **Genome Annotation**

### 661 Transcriptome Sequencing

662 We extracted total RNA from a single *X. bocki* individual and sequenced a strand  
663 specific Illumina paired end library. Extraction of total RNA was performed using a  
664 modified Trizol & RNeasy hybrid protocol for which tissue had to be stored in  
665 RNAlater. cDNA transcription reaction/cDNA synthesis was done using the  
666 RETROscript kit (Ambion) using both Oligo(dT) and Random Decamer primers.  
667 Detailed extraction and transcription protocols are available from the corresponding  
668 authors. The resulting transcriptomic reads (deposited under [SRX20415651](#)) were  
669 assembled with the Trinity pipeline<sup>62,63</sup> into 103,056 sequences (N50: 705;  
670 BUSCO\_v5 Eukaryota scores: C:65.1%, [S:34.1%, D:31.0%], F:22.0%, M:12.9%) for  
671 initial control and then supplied to the genome annotation pipeline (below).

672

### 673 Repeat annotation

674 In the absence of a repeat library for *Xenoturbellida* we first used RepeatModeller v.  
675 1.73 to establish a library *de novo*. We then used RepeatMasker v. 4.1.0  
676 (<https://www.repeatmasker.org>) and the Dfam library<sup>64,65</sup> to soft-mask the genome.  
677 We mapped the repeats to the instaGRAAL scaffolded genome with RepeatMasker.

678

### 679 Gene prediction and annotation

680 We predicted genes using Augustus<sup>66</sup> implemented into the BRAKER (v.2.1.0)  
681 pipeline<sup>24,25</sup> to incorporate the RNA-Seq data. BRAKER uses spliced aligned RNA-  
682 Seq reads to improve training accuracy of the gene finder GeneMark-ET<sup>67</sup>.  
683 Subsequently, a highly reliable gene set predicted by GeneMark-ET in *ab initio* mode  
684 was selected to train the gene finder AUGUSTUS, which in a final step predicted

685 genes with evidence from spliced aligned RNA-Seq reads. To make use of additional  
686 single cell transcriptome data allowing for a more precise prediction of 3'-UTRs we  
687 employed a production version of BRAKER (August 2018 snapshot). We had  
688 previously mapped the RNA-Seq data to the genome with gmap-gsnap v. 2018-07-  
689 04<sup>68</sup> and used samtools<sup>69</sup> and bamtools<sup>70</sup> to create the necessary input files. This  
690 process was repeated in an iterative way, visually validating gene structures and  
691 comparing with mappings loci inferred from a set of single-cell RNA-Seq data  
692 (published elsewhere, see: <sup>71</sup>) in particular regarding fused genes. Completeness of  
693 the gene predictions was independently assessed with BUSCO\_v5<sup>28</sup> setting the  
694 metazoan and the eukaryote datasets as reference respectively on gVolante<sup>72</sup>. We  
695 used InterProScan v. 5.27-66.0 standalone<sup>73,74</sup> on the UCL cluster to annotate the  
696 predicted *X. bocki* proteins with Pfam, SUPERFAM, PANTHER, and Gene3D  
697 information.

698

#### 699 Horizontal Gene Transfer

700 To detect horizontally acquired genes in the *X. bocki* genome we used a pipeline  
701 available from (<https://github.com/reubwn/hgt>). Briefly, this uses blasts against the  
702 NCBI database, alignments with MAFFT<sup>75</sup>, and phylogenetic inferences with  
703 IQTree<sup>76,77</sup> to infer most likely horizontally acquired genes, while trying to discard  
704 contamination (e.g. from co-sequenced gut microbiota).

705

#### 706 Orthology inference

707 We included 155 metazoan species and outgroups into our orthology analysis. We  
708 either downloaded available proteomes or sourced RNA-Seq reads from online  
709 repositories to then use Trinity v 2.8.5 and Trinotate v. 3.2.0 to predict protein sets.  
710 In the latter case we implemented diamond v. 2.0.0 blast<sup>78,79</sup> searches against  
711 UniProt and Pfam<sup>80</sup> hmm screens against the Pfam-A dataset into the prediction  
712 process. We had initially acquired 185 datasets, but excluded some based on inferior  
713 BUSCO completeness, while at the same time aimed to span as many phyla as  
714 possible. Orthology was then inferred using Orthofinder v. 2.2.7<sup>81,82</sup>, again with  
715 diamond as the blast engine.

716 Using InterProScan v. 5.27-66.0 standalone on all proteomes we added  
717 functional annotation and then employed kinfin<sup>35</sup> to summarise and analyse the  
718 orthology tables. For the kinfin analysis, we tested different query systems in regard

719 to phylogenetic groupings (Supplementary).

720 To screen for inflation and contraction of gene families we first employed  
721 CAFE5<sup>83</sup>, but found the analysis to suffer from long branches and sparse taxon  
722 sampling in *Xenambulacraria*. We thus chose to query individual gene families (e.g.  
723 transcription factors) by looking up Pfam annotations in the InterProScan tables of  
724 high-quality genomes in our analysis.

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

727

#### 728 Presence/absence phylogenetics

729 We used a database of metazoan proteins, updated from ref <sup>84</sup>, as the basis for an  
730 OMA analysis to calculate orthologous groups, performing two separate runs, one  
731 including *Xenoturbella* and acoels, and one with only *Xenoturbella*. We converted  
732 OMA gene OrthologousMatrix.txt files into binary gene presence absence matrices in  
733 Nexus format with datatype = restriction. We calculated phylogenetic trees on these  
734 matrices using RevBayes (see <https://github.com/willpett/metazoa-gene-content> for  
735 RevBayes script), as described in ref 74 with corrections for no absent sites  
736 and no singleton presence, using the reversible, not the Dollo model,  
737 as it is more likely to be able to correct for noise related to  
738 prediction errors <sup>85,78</sup>. For each matrix, two runs were performed and compared and  
739 consensus trees generated with bpcomp from Phylobayes<sup>86</sup>.

740

#### 741 Hox and ParaHox gene cluster identification and characterisation

742 Previous work has already used transcriptomic data and phylogenetic inference to  
743 identify the homeobox repertoire in *Xenoturbella bocki*. These annotations were used  
744 to identify genomic positions and gene annotations that correspond to Hox and  
745 ParaHox clusters in *X. bocki*. Protein sequences of homeodomains (Evx, Cdx, Gsx,  
746 antHox1, centHox1, centHox2, cent3 and postHoxP) were used as TBLASTN  
747 queries to identify putative scaffolds associated with Hox and ParaHox clusters.  
748 Gene models from these scaffolds were compared to the full length annotated  
749 homeobox transcripts from<sup>87</sup> using BLASTP, using hits over 95% identity for  
750 homeobox classification. There were some possible homeodomain containing genes  
751 on the scaffolds that were not previously characterised and were therefore not given  
752 an annotation.

753 There were issues concerning the assignment of postHoxP and Evx to gene  
754 models. To ascertain possible CDS regions for these genes, RNA-Seq reads were  
755 mapped with HISAT2 to the scaffold and to previous annotation<sup>87</sup>, were assembled  
756 with Trinity and these were combined with BRAKER annotations.

757 Some issues were also observed with homeodomain queries matching  
758 genomic sequences that were identical, suggesting artefactual duplications. To  
759 investigate contiguity around genes the ONT reads were aligned with Minimap2 to  
760 capture long reads over regions and coverage.

761

### 762 Small RNA Sequencing and Analysis

763 Two samples of starved worms were subjected to 5' monophosphate dependent  
764 sequencing of RNAs between 15 and 36 nucleotides in length, according to  
765 previously described methods<sup>88</sup>. Using miRTrace<sup>89</sup> 3.3, 18.6 million high-quality  
766 reads were extracted and merged with the 27 635 high quality 454 sequencing reads  
767 from Philippe et al. The genome sequence was screened for conserved miRNA  
768 precursors using MirMachine<sup>90</sup> followed by a MirMiner run that used predicted  
769 precursors and processed and merged reads on the genome<sup>91</sup>. Outputs of  
770 MirMachine and MirMiner were manually curated using a uniform system for the  
771 annotation of miRNA genes<sup>92</sup> and by comparing to MirGeneDB<sup>93</sup>.

772

### 773 Neuropeptide prediction and screen

774 Neuropeptide prediction was conducted on the full set of *X.bocki* predicted proteins  
775 using two strategies to detect neuropeptide sequence signatures. First, using a  
776 custom script detecting the occurrence of repeated sequence patterns:  
777 RRx(3,36)RRx(3,36)RRx(3,36)RR,RRx(2,35)ZRRx(2,35)ZRR,  
778 RRx(2,35)GRRx(2,35)GRR, RRx(1,34)ZGRRx(1,34)ZGRR where R=K or R; x=any  
779 amino acid; Z=any amino acid but repeated within the pattern. Second, using  
780 HMMER3.1<sup>94</sup> (hmmer.org), and a combination of neuropeptide HMM models  
781 obtained from the PFAM database (pfam.xfam.org) as well as a set of custom HMM  
782 models derived from alignment of curated sets of neuropeptide sequences<sup>48,49,95</sup>.  
783 Sequences retrieved using both methods and comprising fewer than 600 amino  
784 acids were further validated. First, by blast analysis: sequences with E-Value ratio  
785 "best blast hit versus ncbi nr database/best blast hit versus curated neuropeptide  
786 dataset" < 1e-40 were discarded. Second by reciprocal best blast hit clustering using

787 Clans<sup>96</sup> (eb.tuebingen.mpg.de/protein-evolution/software/clans/) with a set of curated  
788 neuropeptide sequences<sup>48</sup>. SignalP-5.0<sup>97</sup> (cbs.dtu.dk/services/SignalP/) was used to  
789 detect the presence of a signal peptide in the curated list of predicted neuropeptide  
790 sequences while Neuropred<sup>98</sup> (stagbeetle.animal.uiuc.edu/cgi-bin/neuropred.py) was  
791 used to detect cleavage sites and post-translational modifications. Sequence  
792 homology of the predicted sequence with known groups was analysed using a  
793 combination of (i) blast sequence similarity with known bilaterian neuropeptide  
794 sequences, (ii) reciprocal best blast hit clustering using Clans and sets of curated  
795 neuropeptide sequences, (iii) phylogeny using MAFFT  
796 (mafft.cbrc.jp/alignment/server/), TrimAl<sup>99</sup> (trimal.cgenomics.org/) and IQ-TREE<sup>100</sup>  
797 webserver for alignment, trimming and phylogeny inference respectively. Bilaterian  
798 prokineticin-like sequences were searched in ncbi nucleotide, EST and SRA  
799 databases as well as in the *Saccoglossus kowalevskii* genome assembly<sup>77,101</sup>  
800 ([groups.oist.jp/molgenu](http://groups.oist.jp/molgenu)) using various bilaterian prokineticin-related protein  
801 sequences as query. Sequences used for alignments shown in figures were  
802 collected from ncbi nucleotide and protein databases as well as from the following  
803 publications: 7B2<sup>48</sup>; NucB2<sup>95</sup>; Insulin<sup>102</sup>; Prokineticin<sup>39,40,103</sup>. Alignments for figures  
804 were created with Jalview (jalview.org).

805

#### 806 Neuropeptide receptor search

807 §Neuropeptide Receptor sequences for Rhodopsin type GPCR, Secretin type GPCR  
808 and tyrosine and serine/threonine kinase receptors were searched by running  
809 HMMER3.1 on the full set of *X.bocki* predicted proteins using the 7tm\_1 (PF00001),  
810 7tm\_2 (PF00002) and PK\_Tyr\_Ser-Thr (PF07714) HMM models respectively which  
811 were obtained from the PFAM database (pfam.xfam.org). Sequences above the  
812 significance threshold were then aligned with sequences from the curated dataset,  
813 trimmed and phylogeny inference was conducted using same method as for the  
814 neuropeptide. A second alignment and phylogeny inference was conducted after  
815 removal of all *X.bocki* sequences having no statistical support for grouping with any  
816 of the known neuropeptide receptors from the curated dataset. Curated datasets  
817 were collected from the following publications: Rhodopsin type GPCR beta and  
818 gamma and Secretin type GPCR<sup>103</sup>; Rhodopsin type GPCR delta (Leucine-rich  
819 repeat-containing G-protein coupled Receptors)<sup>104</sup>; Tyrosine kinase receptors<sup>105,106</sup>;  
820 and were complemented with sequences from NCBI protein database.



821

## 822 Synteny

823 Ancestral linkage analyses rely on mutual-best-hits computed using Mmseqs2<sup>107</sup>  
824 between pairs of species in which chromosomal assignments to ancestral linkage  
825 groups (ALG) was previously performed, such as *Branchiostoma floridae* or *Pecten*  
826 *maximus*<sup>41</sup>. Oxford dotplots were computed by plotting reciprocal positions of  
827 indexed pairwise orthologs between two species as performed previously<sup>41,42</sup>. The  
828 significance of ortholog enrichment in pairs of chromosomes was assessed using a  
829 fisher test.

830 We also used a Python implementation of MCscanX<sup>108</sup> (Haibao Tang and available  
831 on [https://github.com/tanghaibao/jcvi/wiki/MCscan-\(Python-version\)](https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version))) to compare *X.*  
832 *bocki* to *Euphydtia muelleri*, *Trichoplax adhaerens*, *Branchiostoma floridae*,  
833 *Saccoglossus kowalevskii*, *Ciona intestinalis*, *Nematostella vectensis*, *Asteria*  
834 *rubens*, *Pecten maximus*, *Nemopilema nomurai*, *Carcinoscorpius rotundicauda*. Briefly,  
835 the pipeline uses high quality genomes and their annotations to infer syntenic blocks  
836 based on proximity. For this an all vs. all blastp is performed and synteny extended  
837 from anchors identified in this way. Corresponding heatmaps were plotted with  
838 Python in a Jupyter notebooks instance.

839

## 840 *Chlamydia* assembly and annotation

841 We identified a highly contiguous *Chlamydia* genome in the *X. bocki* genome  
842 assembly using blast. We then used our Oxford Nanopore derived long-reads to  
843 scaffold the *Chlamydia* genome with LINKS<sup>109</sup> and annotated it with the automated  
844 PROKKA pipeline. To place the genome on the *Chlamydia* tree we extracted the 16S  
845 ribosomal RNA gene sequence, aligned it with set of *Chlamydia* 16S rRNA  
846 sequences from<sup>29</sup> using MAFFT, and reconstructed the phylogeny using IQ-TREE  
847 2<sup>76</sup> We visualized the resulting tree with Figtree (<http://tree.bio.ed.ac.uk/>).

848

849

## 850 **Acknowledgements**

851 We thank Josh Quick and Nick Loman for help with the generation of ONT long-read  
852 data. Analyses were conducted mainly on the UCL Cluster, with some computations  
853 also run on the CHEOPS cluster at the University of Cologne. We are grateful to  
854 Kevin J. Peterson for his comments on the manuscript, the miRNA section in  
855 particular. We thank the Kristineberg Center for Marine Research and Innovation for

856 their essential support in sampling *Xenoturbella*.

857

### 858 **Conflict of interest**

859 The authors declare no conflict of interest.

860

### 861 **Data availability**

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

868

### 869 **Funding**

870 PHS was funded by an ERC grant (ERC-2012-AdG 322790) to MJT, which also  
871 supported HR, ACZ, SM. PHS was also funded through an Emmy-Noether grant  
872 (434028868) to himself. Part of this work was funded by BBSRC grant  
873 BB/R016240/1 (M.J.T./P.K.), by a Leverhulme Trust Research Project Grant RPG-  
874 2018-302 (M.J.T./D.J.L.), and by the European Union's Horizon 2020 research and  
875 innovation program under the Marie Skłodowska-Curie grant agreement no 764840  
876 IGNITE (M.J.T./P.N.).

877 **References**

878

- 879 1. Telford, M. J. Xenoturbellida: the fourth deuterostome phylum and the diet of worms.  
880 *Genesis (New York, N.Y.: 2000)* **46**, 580–586 (2008).
- 881 2. Westblad, E. Xenoturbella bocki n. g., n. sp., a peculiar, primitive Turbellarian type. *Arkiv*  
882 *för Zoologi* 3–29 (1949).
- 883 3. Philippe, H. *et al.* Mitigating Anticipated Effects of Systematic Errors Supports Sister-  
884 Group Relationship between Xenacoelomorpha and Ambulacraria. *Current Biology* **29**, 1818-  
885 1826.e6 (2019).
- 886 4. Cannon, J. T. *et al.* Xenacoelomorpha is the sister group to Nephrozoa. *Nature* **530**, 89–93  
887 (2016).
- 888 5. Ueki, T., Arimoto, A., Tagawa, K. & Satoh, N. Xenacoelomorph-Specific Hox Peptides:  
889 Insights into the Phylogeny of Acoels, Nemertodermatids, and Xenoturbellids. *Zool Sci* **36**,  
890 395–401 (2019).
- 891 6. Hejnol, A. *et al.* Assessing the root of bilaterian animals with scalable phylogenomic  
892 methods. *Proceedings. Biological sciences / The Royal Society* **276**, 4261–4270 (2009).
- 893 7. Rouse, G. W., Wilson, N. G., Carvajal, J. I. & Vrijenhoek, R. C. New deep-sea species of  
894 *Xenoturbella* and the position of Xenacoelomorpha. *Nature* **530**, 94–97 (2016).
- 895 8. Srivastava, M., Mazza-Curll, K. L., Wolfswinkel, J. C. van & Reddien, P. W. Whole-Body  
896 Acoel Regeneration Is Controlled by Wnt and Bmp-Admp Signaling. *Current Biology* **24**,  
897 1107–1113 (2014).
- 898 9. Philippe, H. *et al.* Acoelomorph flatworms are deuterostomes related to *Xenoturbella*.  
899 *Nature* **470**, 255–258 (2011).
- 900 10. Bourlat, S. J. *et al.* Deuterostome phylogeny reveals monophyletic chordates and the new  
901 phylum Xenoturbellida. *Nature* **444**, 85–88 (2006).
- 902 11. Nakano, H. What is *Xenoturbella*? *Zoological Letters* **1**, 1 (2015).
- 903 12. Hejnol, A. & Martindale, M. Q. Acoel development supports a simple planula-like  
904 urbilaterian. *Philosophical transactions of the Royal Society of London. Series B, Biological*  
905 *sciences* **363**, 1493–1501 (2008).
- 906 13. Martynov, A. *et al.* Multiple paedomorphic lineages of soft-substrate burrowing  
907 invertebrates: parallels in the origin of *Xenocratena* and *Xenoturbella*. *Plos One* **15**,  
908 e0227173 (2020).
- 909 14. Westheide, W. Progenesis as a principle in meiofauna evolution. *J Nat Hist* **21**, 843–854  
910 (1987).

- 911 15. Nakano, H. *et al.* *Xenoturbella bocki* exhibits direct development with similarities to  
912 Acoelomorpha. *Nature Communications* **4**, 1537 (2013).
- 913 16. Sempere, L. F., Cole, C. N., McPeck, M. A. & Peterson, K. J. The phylogenetic  
914 distribution of metazoan microRNAs: insights into evolutionary complexity and constraint.  
915 **306**, 575–588 (2006).
- 916 17. Gehrke, A. R. *et al.* Acoel genome reveals the regulatory landscape of whole-body  
917 regeneration. *Science* **363**, 1–9 (2019).
- 918 18. Arimoto, A. *et al.* A draft nuclear-genome assembly of the acoel flatworm *Praesagittifera*  
919 *naikaiensis*. *Gigascience* **8**, giz023 (2019).
- 920 19. Martinez, P. *et al.* Genome assembly of the acoel flatworm *Symsagittifera roscoffensis*, a  
921 model for research on body plan evolution and photosymbiosis. *G3 Genes Genomes Genetics*  
922 **13**, jkac336 (2022).
- 923 20. Moroz, L. L., Romanova, D. Y. & Kohn, A. B. Neural versus alternative integrative  
924 systems: molecular insights into origins of neurotransmitters. *Philosophical Transactions*  
925 *Royal Soc Lond Ser B Biological Sci* **376**, 20190762 (2021).
- 926 21. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to  
927 single-cell sequencing. *www.liebertpub.com* **19**, 455–477 (2012).
- 928 22. Prysycz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous  
929 genomes. *Nucleic Acids Research* **44**, e113–e113 (2016).
- 930 23. Baudry, L. *et al.* instaGRAAL: chromosome-level quality scaffolding of genomes using a  
931 proximity ligation-based scaffold. *Genome Biol* **21**, 148 (2020).
- 932 24. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-Genome Annotation  
933 with BRAKER. *Methods Mol Biology Clifton N J* **1962**, 65–95 (2019).
- 934 25. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1:  
935 Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS.  
936 *Bioinformatics* **32**, 767–769 (2016).
- 937 26. Francis, W. R. & Wörheide, G. Similar ratios of introns to intergenic sequence across  
938 animal genomes. *Genome Biol Evol* **9**, evx103- (2017).
- 939 27. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinform Oxf Engl* **30**, 2068–  
940 9 (2014).
- 941 28. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M.  
942 BUSCO: assessing genome assembly and annotation completeness with single-copy  
943 orthologs. **31**, 3210–3212 (2015).
- 944 29. Dharamshi, J. E. *et al.* Marine Sediments Illuminate Chlamydiae Diversity and Evolution.  
945 *Curr Biol* **30**, 1032-1048.e7 (2020).

- 946 30. Kjeldsen, K. U., Obst, M., Nakano, H., Funch, P. & Schramm, A. Two Types of  
947 Endosymbiotic Bacteria in the Enigmatic Marine Worm *Xenoturbella bocki*. **76**, 2657–2662  
948 (2010).
- 949 31. Ueki, T., Arimoto, A., Tagawa, K. & Satoh, N. Xenacoelomorph-Specific Hox Peptides:  
950 Insights into the Phylogeny of Acoels, Nemertodermatids, and Xenoturbellids. *Zoöl. Sci.* **36**,  
951 395–401 (2019).
- 952 32. Natsidis, P., Kapli, P., Schiffer, P. H. & Telford, M. J. Systematic errors in orthology  
953 inference and their effects on evolutionary analyses. *Iscience* 102110 (2021)  
954 doi:10.1016/j.isci.2021.102110.
- 955 33. Schiffer, P. H., Robertson, H. E. & Telford, M. J. Orthonectids Are Highly Degenerate  
956 Annelid Worms. *Current Biology* 1–9 (2018) doi:10.1016/j.cub.2018.04.088.
- 957 34. Mikhailov, K. V. *et al.* The genome of *Intoshia linei* affirms orthonectids as highly  
958 simplified spiralian. *Current Biology* **26**, 1768–1774 (2016).
- 959 35. Laetsch, D. R., Laetsch, D. R., Blaxter, M. L. & Blaxter, M. L. KinFin: Software for  
960 Taxon-Aware Analysis of Clustered Protein Sequences. *G3 (Bethesda, Md.)* **7**, 3349–3357  
961 (2017).
- 962 36. Takami, H. *et al.* An automated system for evaluation of the potential functionome:  
963 MAPLE version 2.1.0. *Dna Res* **23**, 467–475 (2016).
- 964 37. Kapli, P. *et al.* Lack of support for Deuterostomia prompts reinterpretation of the first  
965 Bilateria. *Sci Adv* **7**, eabe2741 (2021).
- 966 38. Thiel, D., Franz-Wachtel, M., Aguilera, F. & Hejnol, A. Xenacoelomorph  
967 Neuropeptidomes Reveal a Major Expansion of Neuropeptide Systems during Early  
968 Bilaterian Evolution. *Molecular Biology And Evolution* **35**, 2528–2543 (2018).
- 969 39. Negri, L. & Ferrara, N. The Prokineticins: Neuromodulators and Mediators of  
970 Inflammation and Myeloid Cell-Dependent Angiogenesis. *Physiol Rev* **98**, 1055–1082 (2018).
- 971 40. Ericsson, L. & Söderhäll, I. Astakines in arthropods—phylogeny and gene structure. *Dev*  
972 *Comp Immunol* **81**, 141–151 (2018).
- 973 41. Simakov, O. *et al.* Deeply conserved synteny resolves early events in vertebrate evolution.  
974 *Nat Ecol Evol* 1–11 (2020) doi:10.1038/s41559-020-1156-z.
- 975 42. Simakov, O. *et al.* Deeply conserved synteny and the evolution of metazoan  
976 chromosomes. *Sci Adv* **8**, eabi5884 (2022).
- 977 43. Rouse, G. W., Wilson, N. G., Carvajal, J. I. & Vrijenhoek, R. C. New deep-sea species of  
978 *Xenoturbella* and the position of Xenacoelomorpha. *Nature* **530**, 94–97 (2016).
- 979 44. Hejnol, A. Acoelomorpha and Xenoturbellida. in 203–214 (Springer Vienna, 2015).  
980 doi:10.1007/978-3-7091-1862-7\_9.

- 981 45. Brauchle, M. *et al.* Xenacoelomorpha Survey Reveals That All 11 Animal Homeobox  
982 Gene Classes Were Present in the First Bilaterians. *Genome Biol Evol* **10**, 2205–2217 (2018).
- 983 46. Jimenez-Guri, E., Paps, J., Garcia-Fernandez, J. & Salo, E. Hox and ParaHox genes in  
984 Nemertodermatida, a basal bilaterian clade. *Int J Dev Biology* **50**, 675–679 (2006).
- 985 47. Ryan, J. F. *et al.* The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes:  
986 evidence from the starlet sea anemone, *Nematostella vectensis*. *Genome Biol* **7**, R64–R64  
987 (2006).
- 988 48. Jekely, G. Global view of the evolution and diversity of metazoan neuropeptide signaling.  
989 *Proc National Acad Sci* **110**, 8702–8707 (2013).
- 990 49. Mirabeau, O. & Joly, J.-S. Molecular evolution of peptidergic signaling systems in  
991 bilaterians. *Proc National Acad Sci* **110**, E2028–E2037 (2013).
- 992 50. Natsidis, P., Kapli, P., Schiffer, P. H. & Telford, M. J. Systematic errors in orthology  
993 inference and their effects on evolutionary analyses. *iScience* **24**, 102110 (2021).
- 994 51. Weisman, C. M., Murray, A. W. & Eddy, S. R. Many, but not all, lineage-specific genes  
995 can be explained by homology detection failure. *Plos Biol* **18**, e3000862 (2020).
- 996 52. Howe, K. *et al.* Structure and evolutionary history of a large family of NLR proteins in  
997 the zebrafish. *Open Biology* **6**, 160009 (2016).
- 998 53. Pillonel, T., Bertelli, C. & Greub, G. Environmental Metagenomic Assemblies Reveal  
999 Seven New Highly Divergent Chlamydial Lineages and Hallmarks of a Conserved  
1000 Intracellular Lifestyle. *Front Microbiol* **9**, 79 (2018).
- 1001 54. Robertson, H. E. *et al.* Single cell atlas of *Xenoturbella bocki* highlights limited cell-type  
1002 complexity. *Nat. Commun.* **15**, 2469 (2024).
- 1003 55. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina  
1004 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- 1005 56. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies.  
1006 *F1000research* **6**, 1287 (2017).
- 1007 57. Elsworth, B., Jones, M. & Blaxter, M. Badger--an accessible genome exploration  
1008 environment. *Bioinformatics* **29**, 2788–2789 (2013).
- 1009 58. Lewis, S. H. *et al.* Widespread conservation and lineage-specific diversification of  
1010 genome-wide DNA methylation patterns across arthropods. *Plos Genet* **16**, e1008864 (2020).
- 1011 59. Lafontaine, D. L., Yang, L., Dekker, J. & Gibcus, J. H. Hi-C 3.0: Improved Protocol for  
1012 Genome-Wide Chromosome Conformation Capture. *Curr Protoc* **1**, e198 (2021).
- 1013 60. Marie-Nelly, H. *et al.* High-quality genome (re)assembly using chromosomal contact data.  
1014 *Nat Commun* **5**, 5695 (2014).

- 1015 61. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature*  
1016 *Methods* **9**, 357–359 (2012).
- 1017 62. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-Seq: reference  
1018 generation and analysis with Trinity. **8**, 1494–1512 (2013).
- 1019 63. *RNA-Seq De Novo Assembly Using Trinity*. 1–7 (2015).
- 1020 64. Wheeler, T. J. *et al.* Dfam: a database of repetitive DNA based on profile hidden Markov  
1021 models. *Nucleic Acids Res* **41**, D70–D82 (2013).
- 1022 65. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**,  
1023 D81–D89 (2016).
- 1024 66. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron  
1025 submodel. *Bioinformatics* **19 Suppl 2**, ii215–25 (2003).
- 1026 67. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads  
1027 into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**, e119–  
1028 e119 (2014).
- 1029 68. Wu, T. D., Reeder, J., Lawrence, M., Becker, G. & Brauer, M. J. GMAP and GSNAP for  
1030 Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality.  
1031 *Methods Mol Biology Clifton N J* **1418**, 283–334 (2016).
- 1032 69. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,  
1033 2078–2079 (2009).
- 1034 70. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T.  
1035 BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**,  
1036 1691–1692 (2011).
- 1037 71. Robertson, H. E. *et al.* Single cell atlas of *Xenoturbella bocki* highlights the limited cell-  
1038 type complexity of a non-vertebrate deuterostome lineage. *Biorxiv* 2022.08.18.504214 (2022)  
1039 doi:10.1101/2022.08.18.504214.
- 1040 72. Nishimura, O., Hara, Y. & Kuraku, S. gVolante for standardizing completeness  
1041 assessment of genome and transcriptome assemblies. *Bioinformatics* **33**, 3635–3637 (2017).
- 1042 73. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.  
1043 *Bioinformatics* **30**, 1236–1240 (2014).
- 1044 74. Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence  
1045 classification and comparison. *Methods in molecular biology (Clifton, N.J.)* **396**, 59–70  
1046 (2007).
- 1047 75. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7:  
1048 Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772–780 (2013).

- 1049 76. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic  
1050 inference in the genomic era. *Mol Biol Evol* **37**, 1530–1534 (2020).
- 1051 77. 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 78. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using  
1055 DIAMOND. *Nature Methods* **12**, 59–60 (2014).
- 1056 79. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale  
1057 using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
- 1058 80. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future.  
1059 *Nucleic Acids Research* **44**, D279-85 (2016).
- 1060 81. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative  
1061 genomics. *Genome Biol* **20**, 238 (2019).
- 1062 82. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome  
1063 comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**, E9-  
1064 13 (2015).
- 1065 83. Han, M. V., Thomas, G. W. C., Lugo-Martinez, J. & Hahn, M. W. Estimating Gene Gain  
1066 and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3.  
1067 *Mol Biol Evol* **30**, 1987–1997 (2013).
- 1068 84. Leclère, L. *et al.* The genome of the jellyfish *Clytia hemisphaerica* and the evolution of  
1069 the cnidarian life-cycle. 1–41 (2018) doi:10.1101/369959.
- 1070 85. The Role of Homology and Orthology in the Phylogenomic Analysis of Metazoan Gene  
1071 Content. 1–7 (2019) doi:10.1093/molbev/msz013.
- 1072 86. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for  
1073 phylogenetic reconstruction and molecular dating. **25**, 2286–2288 (2009).
- 1074 87. Brauchle, M. *et al.* Xenacoelomorpha survey reveals that all 11 animal homeobox gene  
1075 classes were present in the first bilaterians. *Genome biology and evolution* (2018)  
1076 doi:10.1093/gbe/evy170.
- 1077 88. Sarkies, P. *et al.* Ancient and Novel Small RNA Pathways Compensate for the Loss of  
1078 piRNAs in Multiple Independent Nematode Lineages. *Plos Biol* **13**, e1002061 (2015).
- 1079 89. Kang, W. *et al.* miRTrace reveals the organismal origins of microRNA sequencing data.  
1080 *Genome Biol* **19**, 213 (2018).
- 1081 90. Umu, S. U. *et al.* Accurate microRNA annotation of animal genomes using trained  
1082 covariance models of curated microRNA complements in MirMachine. *Biorxiv*  
1083 2022.11.23.517654 (2023) doi:10.1101/2022.11.23.517654.



- 1084 91. Wheeler, B. M. *et al.* The deep evolution of metazoan microRNAs. *Evol Dev* **11**, 50–68  
1085 (2009).
- 1086 92. Fromm, B. *et al.* A Uniform System for the Annotation of Vertebrate microRNA Genes  
1087 and the Evolution of the Human microRNAome. *Annu Rev Genet* **49**, 213–242 (2015).
- 1088 93. Fromm, B. *et al.* MirGeneDB 2.1: toward a complete sampling of all major animal phyla.  
1089 *Nucleic Acids Res* **50**, D204–D210 (2022).
- 1090 94. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and  
1091 iterative HMM search procedure. *Bmc Bioinformatics* **11**, 431–431 (2010).
- 1092 95. Zandawala, M. *et al.* Discovery of novel representatives of bilaterian neuropeptide  
1093 families and reconstruction of neuropeptide precursor evolution in ophiuroid echinoderms.  
1094 *Open Biol* **7**, 170129 (2017).
- 1095 96. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein families based  
1096 on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).
- 1097 97. Armenteros, J. J. A. *et al.* SignalP 5.0 improves signal peptide predictions using deep  
1098 neural networks. *Nat Biotechnol* **37**, 420–423 (2019).
- 1099 98. Southey, B. R., Rodriguez-Zas, S. L. & Sweedler, J. V. Prediction of neuropeptide  
1100 prohormone cleavages with application to RFamides. *Peptides* **27**, 1087–1098 (2006).
- 1101 99. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldon, T. trimAl: a tool for automated  
1102 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973  
1103 (2009).
- 1104 100. Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. IQ-TREE: A Fast and  
1105 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol*  
1106 *Evol* **32**, 268–274 (2015).
- 1107 101. Simakov, O. *et al.* Hemichordate genomes and deuterostome origins. *Nature* **527**, 459–  
1108 465 (2015).
- 1109 102. Cherif--Feildel, M., Berthelin, C. H., Rivière, G., Favrel, P. & Kellner, K. Data for  
1110 evolutive analysis of insulin related peptides in bilaterian species. *Data Brief* **22**, 546–550  
1111 (2019).
- 1112 103. Thiel, D., Franz-Wachtel, M., Aguilera, F. & Hejnol, A. Changes in the neuropeptide  
1113 complement correlate with nervous system architectures in xenacoelomorphs. 1–57 (2018)  
1114 doi:10.1101/265579.
- 1115 104. Roch, G. J. & Sherwood, N. M. Glycoprotein hormones and their receptors emerged at  
1116 the origin of metazoans. *Genome Biol Evol* **6**, 1466–79 (2014).
- 1117 105. Oliveira, A. L. de, Calcino, A. & Wanninger, A. Ancient origins of arthropod moulting  
1118 pathway components. *Elife* **8**, e46113 (2019).

- 1119 106. Smýkal, V. *et al.* Complex Evolution of Insect Insulin Receptors and Homologous  
1120 Decoy Receptors, and Functional Significance of Their Multiplicity. *Mol Biol Evol* **37**, 1775–  
1121 1789 (2020).
- 1122 107. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for  
1123 the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
- 1124 108. Wang, Y. *et al.* MCSscanX: a toolkit for detection and evolutionary analysis of gene  
1125 synteny and collinearity. *Nucleic Acids Research* **40**, e49–e49 (2012).
- 1126 109. Warren, R. L. *et al.* LINKS: Scalable, alignment-free scaffolding of draft genomes with  
1127 long reads. *Gigascience* **4**, 35 (2015).

1128

1129

### 1130 **Figure Supplement legends**

1131 **Figure 8 – supplement 1.** Radial tree representation of the phylogenetic analysis of  
1132 bilaterian Glycoprotein hormone and Bursicon. Colored dots indicate support (UFB, 1000  
1133 ultrafast bootstrap replicates; SHL, 1000 SH-aLRT replicates) and follow the color code in  
1134 the left inset. Scale bar unit for branch length is the number of substitutions per site.  
1135 Branches are colored according to the phylogenetic position of the organism from which  
1136 the sequence originates and follow the color code in the left inset. Abbreviation: Nbl1,  
1137 neuroblastoma suppressor of tumorigenicity 1 — Sequences, alignment and IQTREE tree  
1138 files are available as supplementary online material.

1139

1140 **Figure 8 – supplement 2.** Radial tree representation of the sequence similarities analysis  
1141 of bilaterian insulin related peptides. Tree is calculated from concatenated alignment of A  
1142 and B chains. Scale bar unit for branch length is the number of substitutions per site.  
1143 Branches are colored according to the phylogenetic position of the organism from which  
1144 the sequence originates and follow the color code in the bottom inset. Abbreviations : dILP,  
1145 drosophila Insulin-like peptide ; GSS, gonad stimulating substance ; ILP, Insulin-like peptide  
1146 ; IGF, Insulin-like growth factor. Sequences, alignment and IQTREE tree files are available  
1147 as supplementary online material.

1148

1149 **Figure 8 – supplement 3.** Full tree representation of the sequence similarities analysis of  
1150 bilaterian insulin related peptides. Tree is calculated from concatenated alignment of A and  
1151 B chains. Numbers represent support for nodes calculated using 1000 ultrafast bootstrap  
1152 replications and 1000 SH-aLRT replicates respectively. Scale bar unit for branch length is  
1153 the number of substitutions per site. Branches are colored according to the phylogenetic  
1154 position of the organism from which the sequence originates : red, *Xenoturbella* ; pink,  
1155 Ambulacraria ; blue, Chordata ; orange, Ecdysozoa ; green, Ecdysozoa ; gray, Cnidaria.  
1156 Abbreviations : dILP, drosophila Insulin-like peptide ; GSS, gonad stimulating substance ;

1157 ILP, Insulin-like peptide ; IGF, Insulin-like growth factor. Sequences, alignment and IQTREE  
1158 tree files are available as supplementary online material.

1159

1160 **Figure 8 – supplement 4.** Full tree representation of the phylogenetic analysis of  
1161 bilaterian Leucine-rich repeat-containing G-protein coupled Receptors (Rhodopsin type G-  
1162 protein coupled Receptors delta). Numbers represent support for nodes calculated using  
1163 1000 Ultrafast bootstrap replications and 1000 SH-aLRT replicates respectively. Scale bar  
1164 unit for branch length is the number of substitutions per site. Branches are colored  
1165 according to the phylogenetic position of the organism from which the sequence originates  
1166 : red, *Xenoturbella* ; pink, Ambulacraria ; blue, Chordata ; orange, Ecdysozoa ; green,  
1167 Ecdysozoa ; gray, Cnidaria. Collapsed group colored in red indicate that they contain at  
1168 least one *Xenoturbella bocki* sequence. Abbreviations : GPA<sub>2</sub>, Glycoprotein Hormone  
1169 alpha<sub>5</sub> ; GPB<sub>5</sub>, Glycoprotein Hormone beta<sub>2</sub> ; GPCR, G Protein-Coupled Receptor ; GRL-  
1170 101, G-protein coupled receptor GRL101. Sequences, alignment and IQTREE tree files are  
1171 available as supplementary online material.

1172

1173 **Figure 8 – supplement 5.** Circular tree representation of the phylogenetic analysis of  
1174 bilaterian Rhodopsin type G-protein coupled Receptors beta and gamma. Colored dots  
1175 indicate support (UFB, 1000 ultrafast bootstrap replicates; SHL, 1000 SH-aLRT replicates)  
1176 for main nodes and follow the color code in the bottom inset. Scale bar unit for branch  
1177 length is the number of substitutions per site. Branches are colored according to the  
1178 phylogenetic position of the organism from which the sequence originates and follow the  
1179 color code in the bottom inset. Circular gray bars highlights names of groups of annotated  
1180 sequences. Circular red bars indicate position of groups of Xenacoelomorpha sequences  
1181 and associated number the number of *Xenoturbella bocki* sequence(s) within these groups.  
1182 Abbreviations : AKH, adipokinetic hormone ; Asta-A, Allatostatin-A ; Asta-C, Allatostatin-C  
1183 ; CAPA, Cardio acceleratory peptide ; CCAP, crustacean cardioactive peptide ; CCHa,  
1184 CCHamide peptide ; CCK, cholecystokinin ; CRZ, Corazonin ; eFMRF, ecdysozoan-  
1185 FMRFamide peptide ; GGN-EP, GGN excitatory peptide ; ETH, ecdysis triggering hormone ;  
1186 GnRH, Gonadotropin Releasing Hormone ; GPR<sub>150</sub>, G Protein-Coupled Receptor 150 ;  
1187 GPR<sub>54</sub>, G Protein-Coupled Receptor 54 ; GPR<sub>83</sub>, G Protein-Coupled Receptor 83 ; MCH,  
1188 melanin concentrating hormone ; Myomod, Myomodulin ; NK-2, Neurokinin 2 ; Np-B/W,  
1189 Neuropeptide B/W ; Np-FF, Neuropeptide FF ; Np-F, Neuropeptide F ; Np-S, Neuropeptide  
1190 S ; Np-Y, Neuropeptide Y ; PBAN, pheromone biosynthesis activation neuropeptide ; PEN,  
1191 neuroendocrine peptide PEN ; PRP, Prolactin releasing peptide ; QRFP, Neuropeptide  
1192 QRFP ; RYa, RYamide peptide ; SIFa, SIFamide peptide ; SPR, Sex peptide receptor ;  
1193 tFMRFa, trochozoan-FMRFamide peptide ; TRH, thyrotrophin-releasing hormone.  
1194 Sequences, alignment and IQTREE tree files are available as supplementary online  
1195 material.

1196

1197 **Figure 8 – supplement 6.** Circular tree representation of the phylogenetic analysis of  
1198 bilaterian Tyrosine kinase Receptors. Colored dots indicate support (UFB, 1000 ultrafast  
1199 bootstrap replicates; SHL, 1000 SH-aLRT replicates) and follow the color code in the  
1200 bottom inset. Scale bar unit for branch length is the number of substitutions per site.  
1201 Branches are colored according to the phylogenetic position of the organism from which  
1202 the sequence originates and follow the color code in the bottom inset. Collapsed group

1203 colored in red indicate that they contain at least one *Xenoturbella bocki* sequence.  
1204 Abbreviations : EGF, Epidermal Growth Factor ; Discoidin cont. R, discoidin domain-  
1205 containing receptor ; Orphan Tyr. Kinase Ror2, receptor tyrosine kinase-  
1206 like orphan receptor 2 ; VKR, Venus kinase Receptor ; ILP, Insulin-like peptide ; PDGF,  
1207 Platelet-derived growth factor ; VEGF, Vascular endothelial growth factor ; GDNF, Glial cell  
1208 line-derived neurotrophic factor; FGF, fibroblast growth factor ; PTTH, Prothoracicotropic  
1209 hormone. Sequences, alignment and IQTREE tree files are available as supplementary  
1210 online material.

1211

1212 **Figure 8 – supplement 7.** Full tree representation of the phylogenetic analysis of  
1213 bilaterian Tyrosine kinase Receptors. Numbers represent support for nodes calculated  
1214 using 1000 ultrafast bootstrap replications and 1000 SH-aLRT replicates respectively. Scale  
1215 bar unit for branch length is the number of substitutions per site. Branches are colored  
1216 according to the phylogenetic position of the organism from which the sequence originates  
1217 : red, *Xenoturbella* ; pink, Ambulacraria ; blue, Chordata ; orange, Ecdysozoa ; green,  
1218 Ecdysozoa ; gray, Cnidaria. Collapsed group colored in red indicate that they contain at  
1219 least one *Xenoturbella bocki* sequence. Abbreviations : EGF, Epidermal Growth Factor ;  
1220 Discoidin cont. R, discoidin domain-containing receptor ; Orphan Tyr. Kinase Ror2,  
1221 receptor tyrosine kinase-like orphan receptor 2 ; VKR, Venus kinase Receptor ; ILP, Insulin-  
1222 like peptide ; PDGF, Platelet-derived growth factor ; VEGF, Vascular endothelial growth  
1223 factor ; GDNF, Glial cell line-derived neurotrophic factor ; FGF, fibroblast growth factor ;  
1224 PTTH, Prothoracicotropic. Sequences, alignment and IQTREE tree files are available as  
1225 supplementary online material.

1226

1227 **Figure 8 – supplement 8.** Circular tree representation of the phylogenetic analysis of  
1228 bilaterian Secretin type G-protein coupled Receptors. Colored dots indicate support (UFB,  
1229 1000 ultrafast bootstrap replicates; SHL, 1000 SH-aLRT replicates) for main nodes and  
1230 follow the color code in the bottom inset. Scale bar unit for branch length is the number of  
1231 substitutions per site. Branches are colored according to the phylogenetic position of the  
1232 organism from which the sequence originates and follow the color code in the bottom  
1233 inset. Circular gray bars highlights names of groups of annotated sequences. Circular red  
1234 bars indicate position of groups of Xenacoelomorpha sequences and associated number  
1235 the number of *Xenoturbella bocki* sequence(s) within these groups. Abbreviations : DH<sub>31</sub>,  
1236 diuretic hormone 31 ; Np-R B<sub>1</sub>, Neuropeptide receptor B<sub>3</sub> ; Np-R B<sub>4</sub>, Neuropeptide receptor  
1237 B<sub>1</sub> ; PDF, Pigment-dispersing factor; CRF, Corticotropin-releasing factor; DH-<sub>44</sub>, diuretic  
1238 hormone 44; PTH<sub>2/3</sub>-R, Parathyroid hormone receptor<sub>2/3</sub>; GIP, Gastric inhibitory  
1239 polypeptide; PACAP, Pituitary adenylate cyclase-activating polypeptide ; VIP-R, Vasoactive  
1240 intestinal polypeptide receptor ; GHRH, Growth hormone-releasing hormone; PTH,  
1241 Parathyroid hormone receptor ; SCTR, Secretin Receptor. Sequences, alignment and  
1242 IQTREE tree files are available as supplementary online material.

1243

1244 **Figure 10 – supplement 1.** Conservation of metazoan synteny and Methylation in *X.*  
1245 *bocki*. (a) A summary plot of synteny between major scaffolds in the *X. bocki* genome  
1246 assembly and early branching highly contiguous metazoan genome assemblies: *Euphydtia*

1247 *muelleri*, *Trichoplax adhaerens*, *Branchiostoma floridae*, *Saccoglossus kowalevskii*, *Ciona*  
1248 *intestinalis*, *Nematostella vectensis*, *Asteria rubens*, *Pecten maximus*, *Nemopilema nomurai*,  
1249 *Carcinoscorpius rotundicauda*. All but one of the chromosome sized scaffolds in our  
1250 assembly have at least one syntenic match in the each of the other species (see main text  
1251 for one-to-one plots with key species and a description of the aberrant scaffold). We  
1252 performed the same analysis with *Amphioxus* as the focal species as a proof of principle  
1253 (inset). (b) Analysis of methylation on the largest scaffold in the *X. bocki* genome assembly.  
1254 One scaffold with a deviant gene age and synteny structure (see main text) also stands out  
1255 in terms of methylation. A detailed analysis of methylation patterns across the genome and  
1256 classes of genes will be published separately.

1257

1258 **Figure 10 – supplement 2.** Intergenomic comparison of *X. bocki* and *E. muelleri*  
1259 highlighting synteny connections between the aberrant scaffold c18g6 and scaffolds across  
1260 the sponge genome.

1261

1262 **Figure 13 – figure supplement 1.** Kmer profile of the *X. bocki* Illumina WGS reads obtained  
1263 with GenomeScope2 (Ranallo-Benavidez et al. 2020). Linear plot and transformed linear  
1264 plots are shown. Per description on (<http://qb.cshl.edu/genomescope/>) we used 21mers  
1265 counted with jellyfish (Marçais and Kingsford 2011). GenomeScope genome property  
1266 estimates and measures were len: 222,242,800bp, uniq: 31.7%, aa: 99.1%, ab: 0.929%,  
1267 kcov: 8.72, err: 0.665%, dup: 0.527, k: 21, p:2, model fit min: 34.6%, model fit max: 96.3.

1268

1269 **Supplementary online material on Zenodo (doi:10.5281/zenodo.6962271):**

1270 S. File 1: Orthofinder Orthology, Xbocki\_genome\_Orthogroups.csv.gz;

1271 S. File 2: Comparative annotation of KEGG pathway completeness,

1272 Xbocki\_genome.KEGG\_module\_completeness.xls;

1273 S. File 3: Neuropeptide screen, alignments and treefiles;

1274 S. File 4: 16S rRNA tree for Chlamydia species, Chlamydia\_16S\_rRNA.tree;

1275 S. File 5: Alignments and trees from HGT screen, HGT\_screen\_aln-tree.tgz.

1276 S. File 6: Excel table with data sources for OrthoFinder analysis.

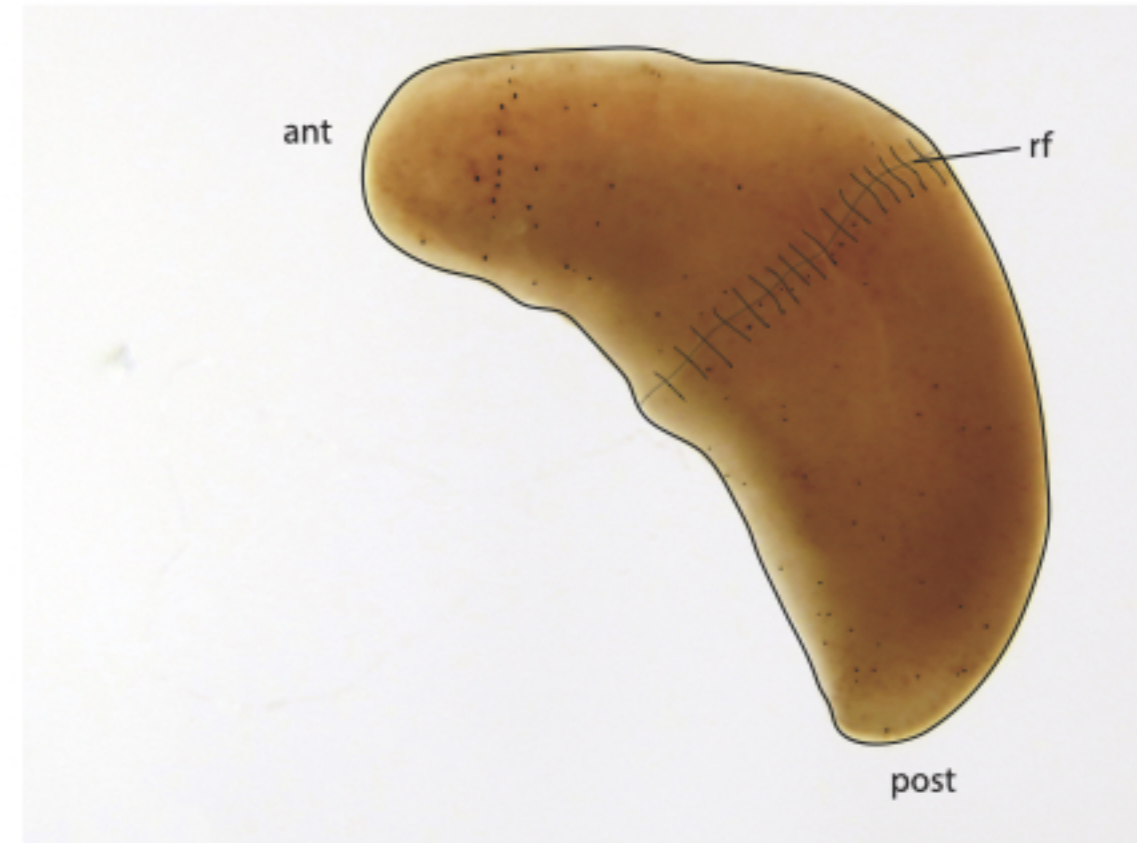
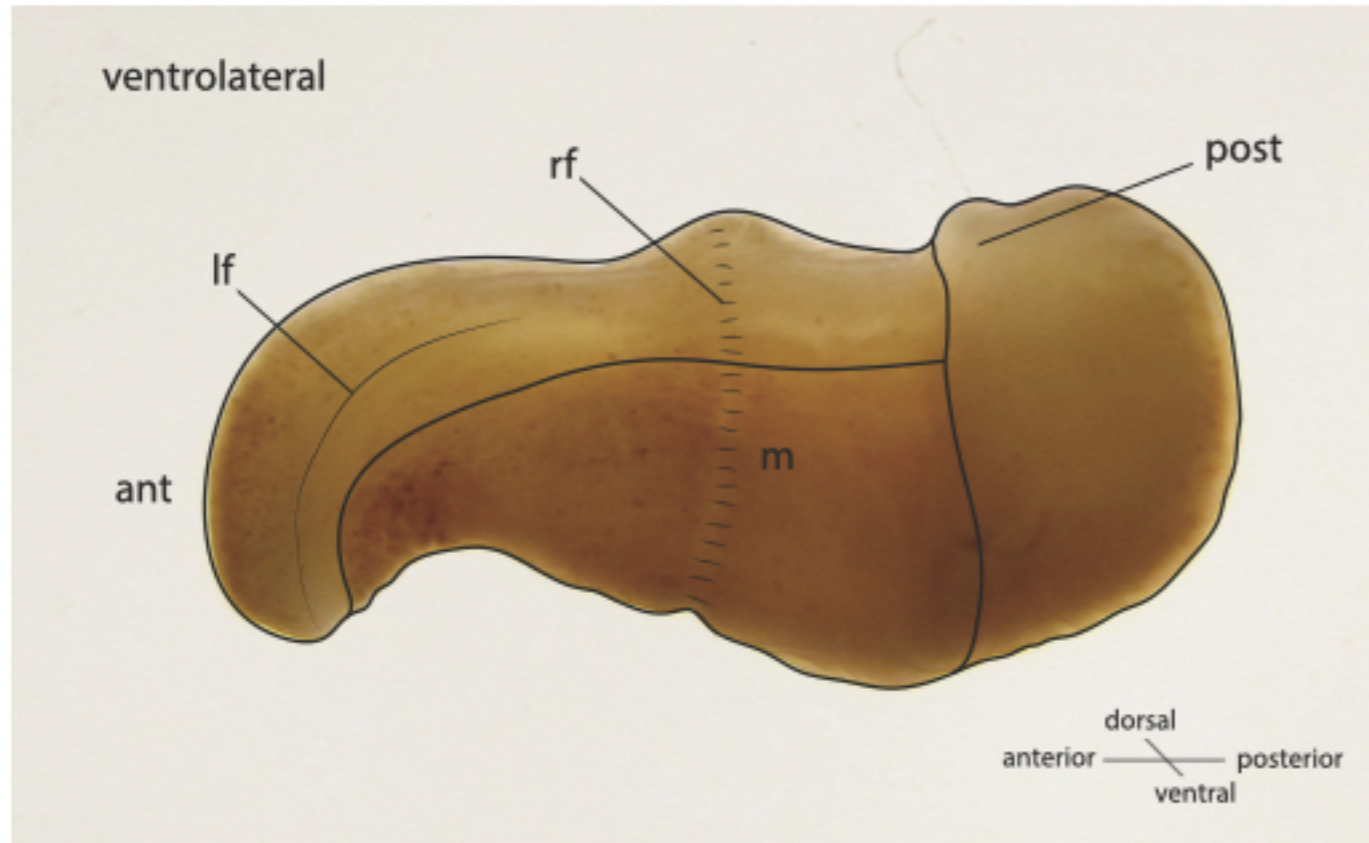


Figure 1: Schematic drawings of *Xenoturbella bocki* showing the simple body organisation of the marine vermiform animal. Abbreviations: ant - anterior, post - posterior, lf - lateral furrow, rf - ring furrow, m - mouth opening.

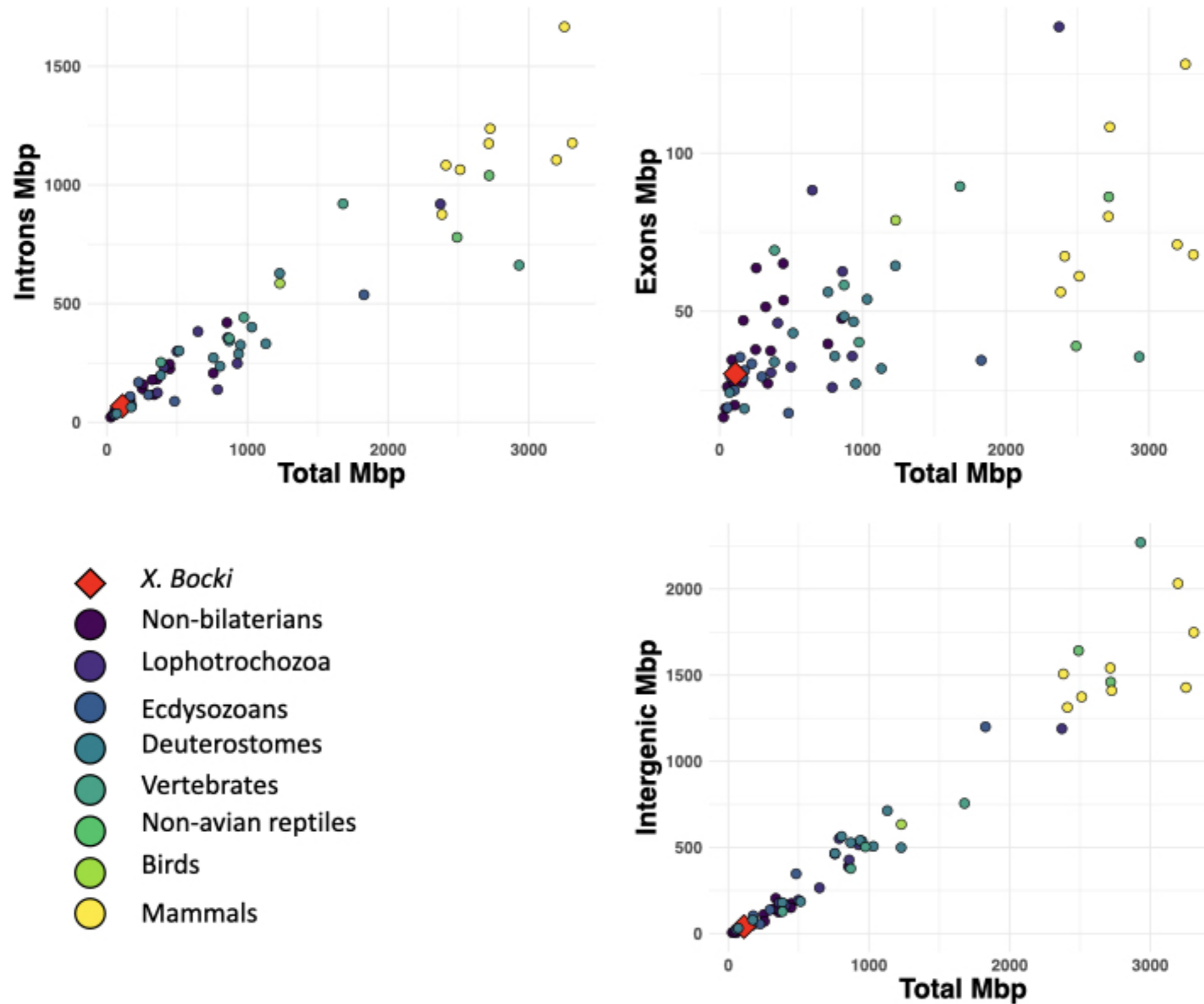


Figure 2: A comparison of total length of exons, introns, and intergenic space in the *X. bocki* genome with other metazoans (data from ref 26). *X. bocki* does not appear to be an outlier in any of these comparisons.



Figure 3: *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*.



4a

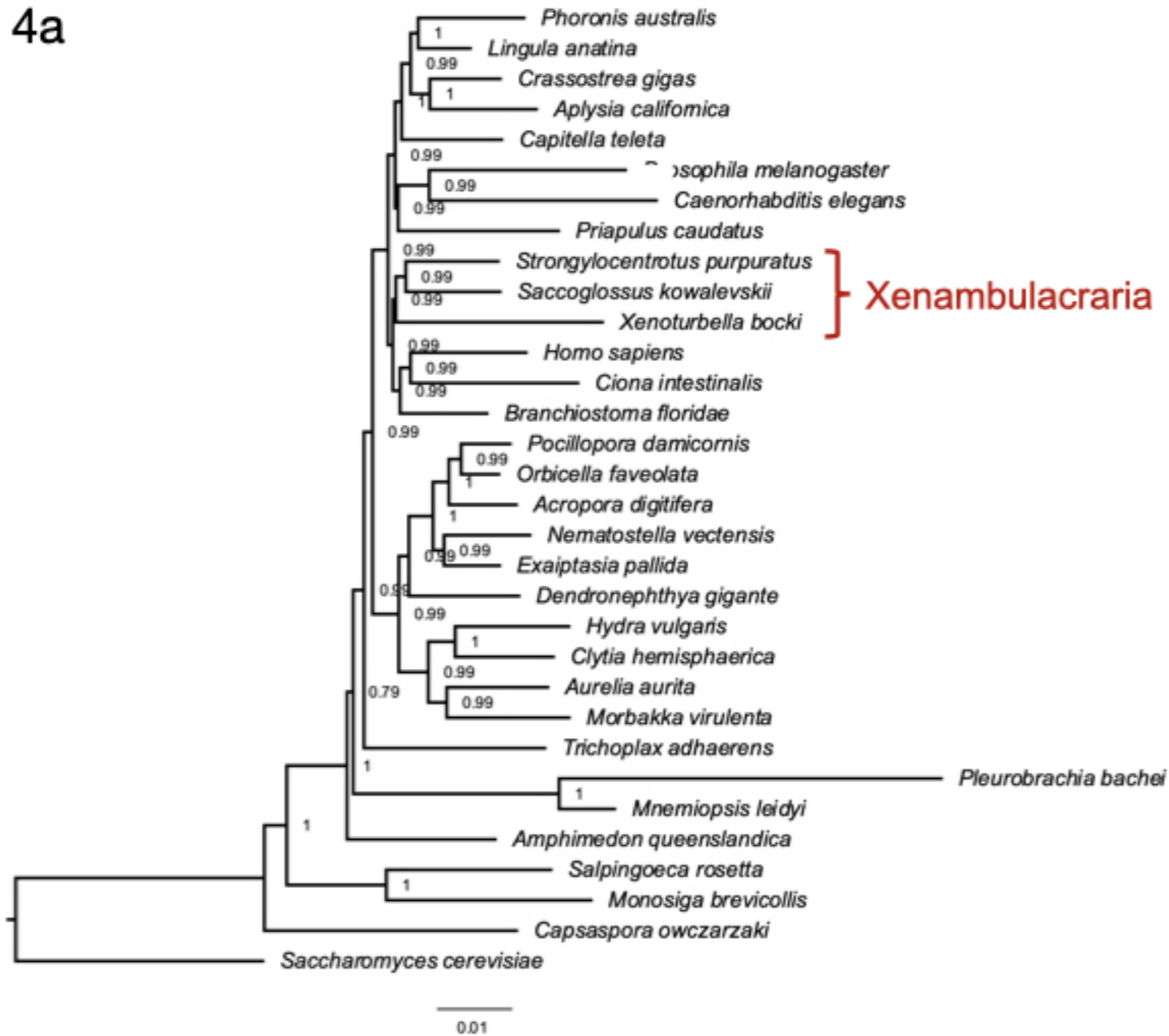
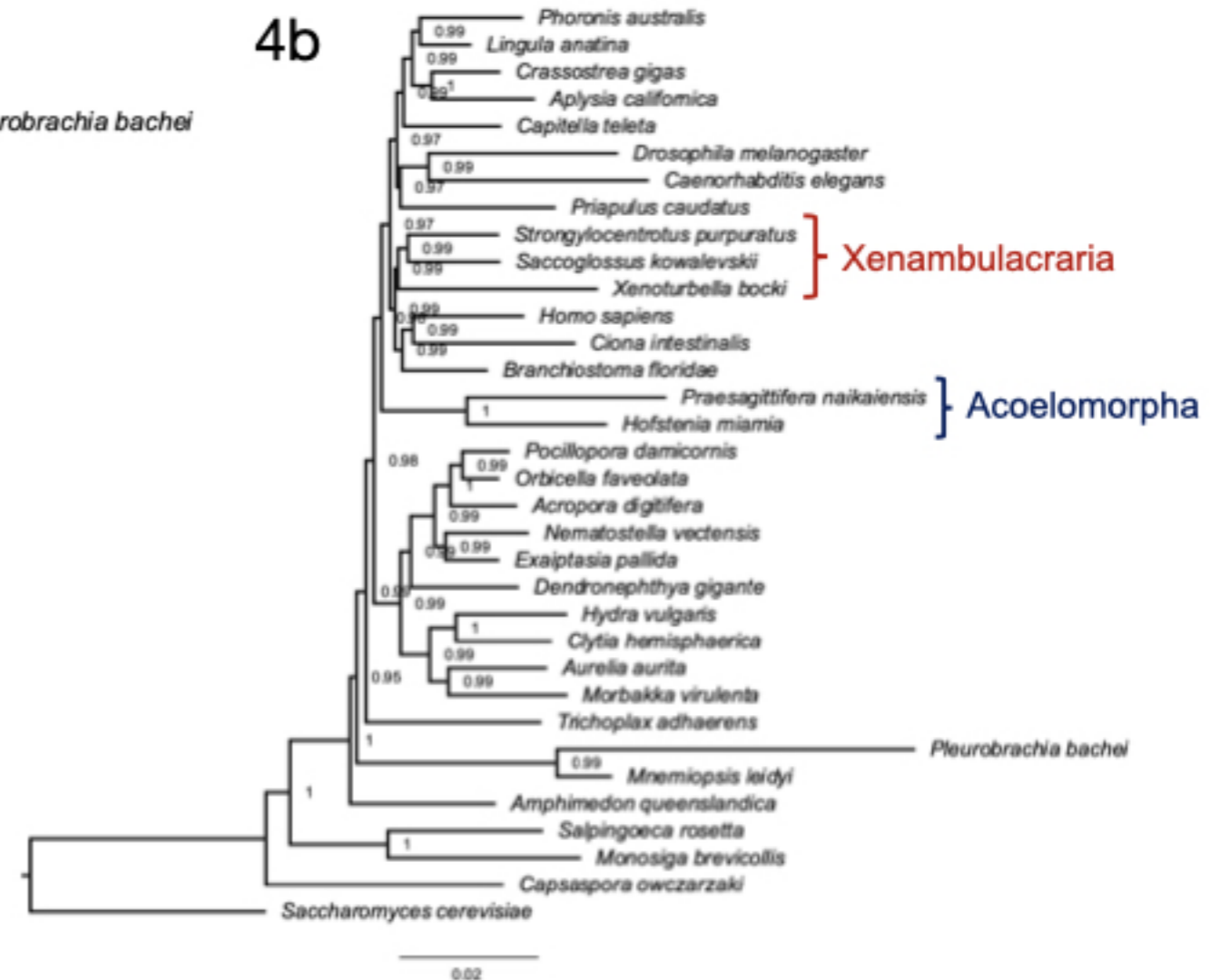


Figure 4: A phylogeny based on presence and absence of genes calculated with OMA. Both analysis (a) and (b) confirm Xenambulacraria, i.e. Xenoturbellida in a group with Echinoderms and Hemichordates. Inclusion of the acoel flat worms places these as sister to all other Bilateria (b). This placement appears an artefact due to the very fast evolution in this taxon, in particular as good evidence exists for uniting Xenoturbellida and Acoela refs 3,4,7–10,31.

4b



5

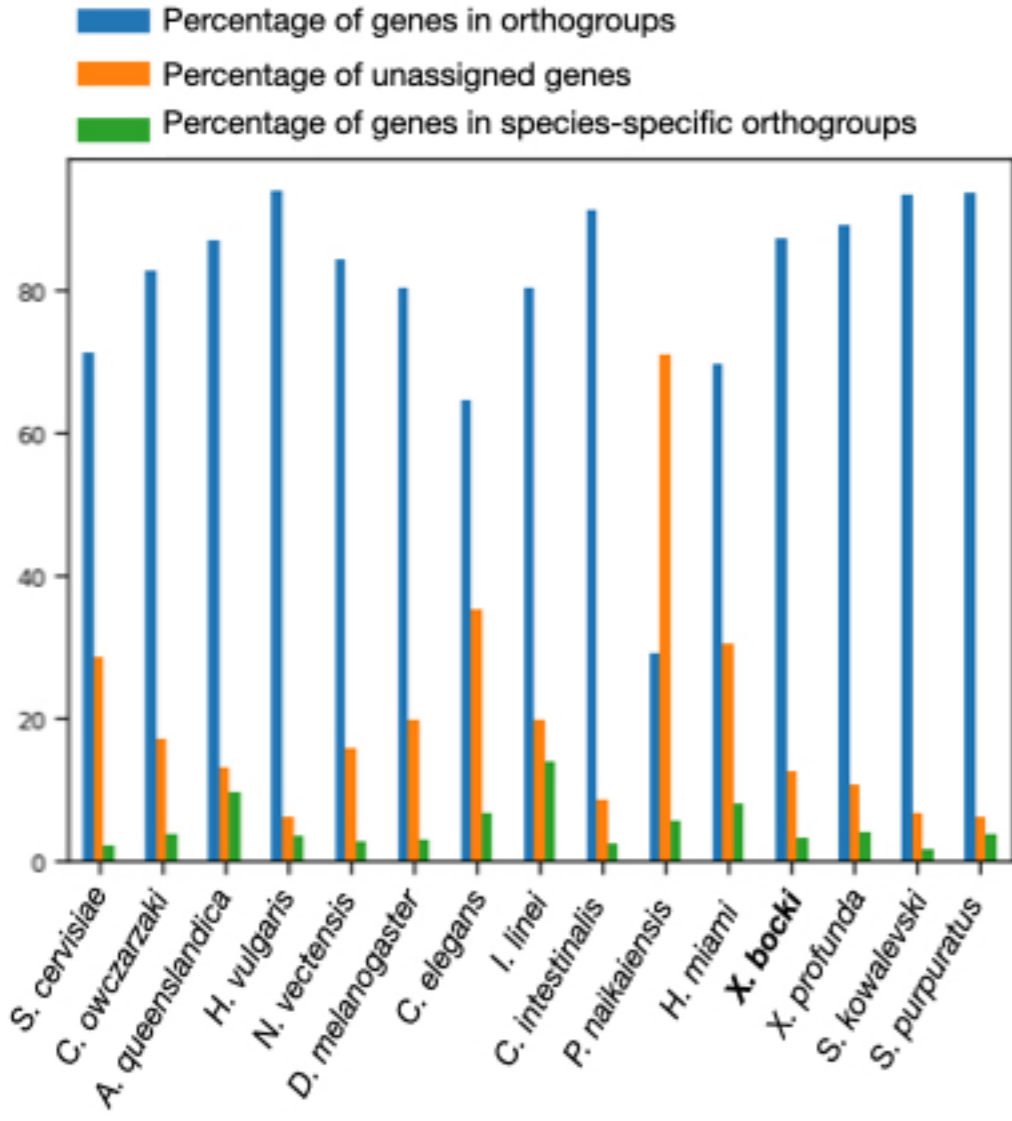
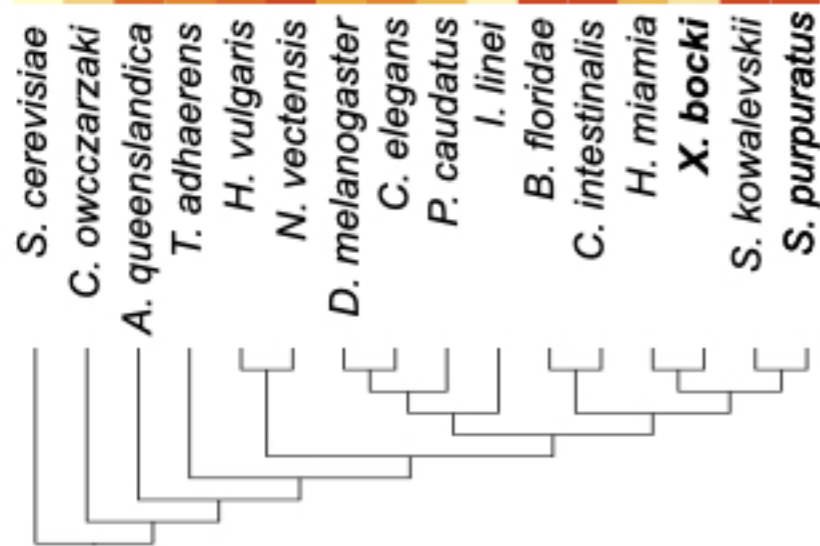
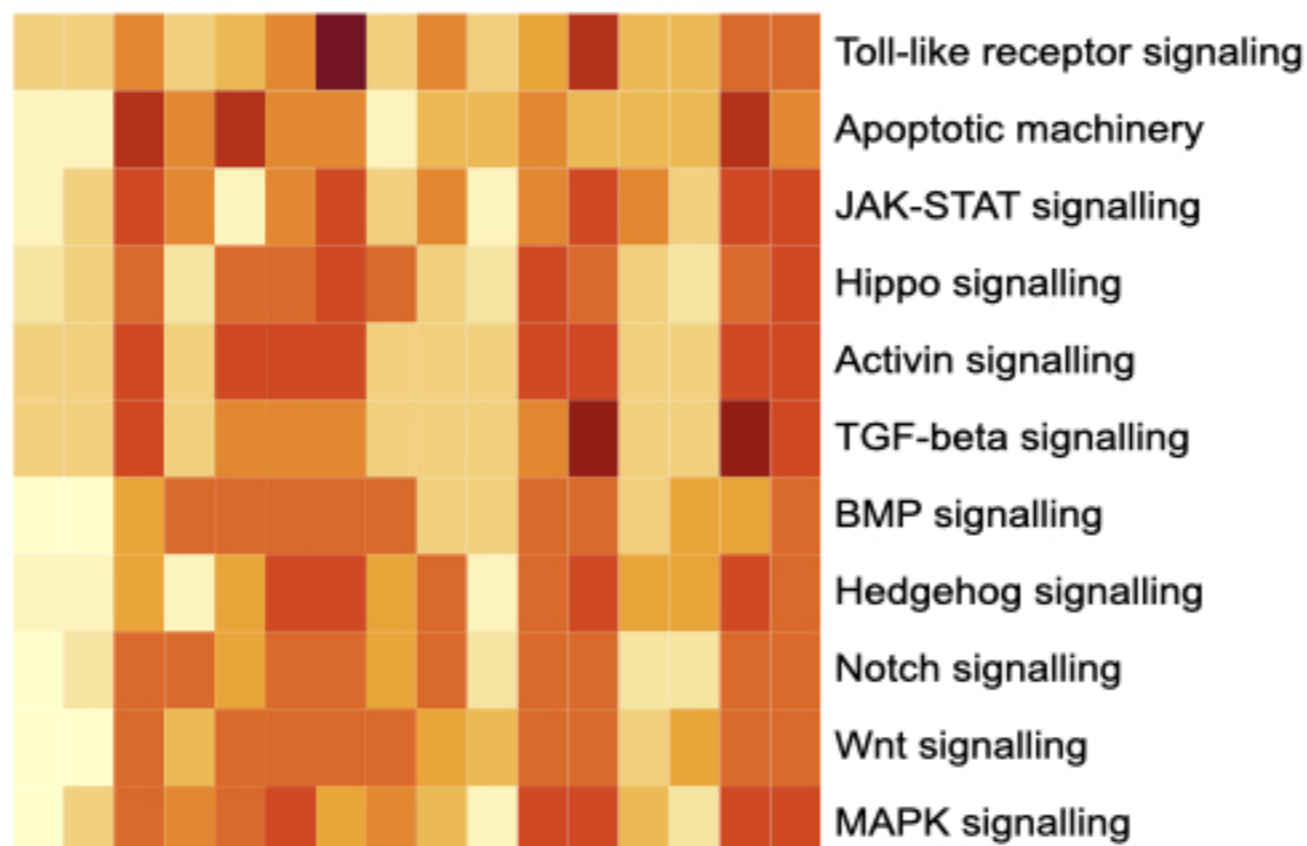


Figure 5: In our orthology screen *X. bocki* shows similar percentages of genes in orthogroups, unassigned genes, and species-specific orthogroups as other well-annotated genomes.

6a



### t-test

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

6b

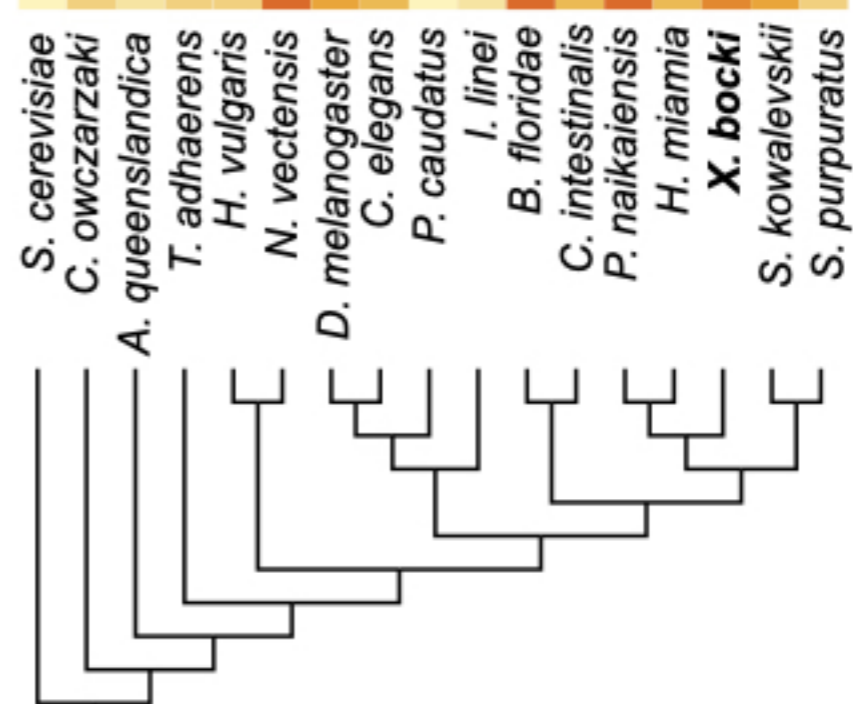
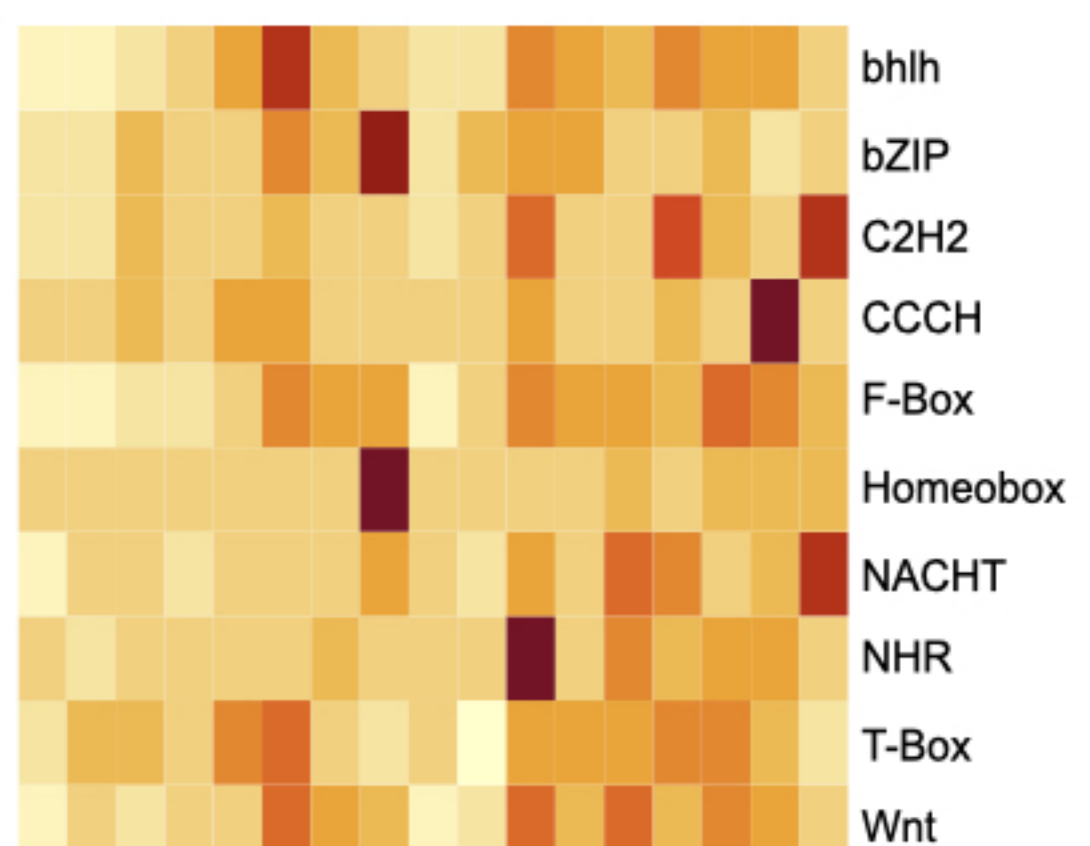


Figure 6: The heatmaps show a comparative measure of relative completeness of signalling pathways based on KEGG and assessed with Genomagle or abundance of genes in a given gene-family based on InterProScan annotations. (a) The number of family members per species in major gene families (based on Pfam 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. (b) 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).

Due to the comparative nature of the assay no “true” scale can be given: darker colours indicate higher comparative completeness. Schematic cladograms in b/c drawn by the authors.

7a

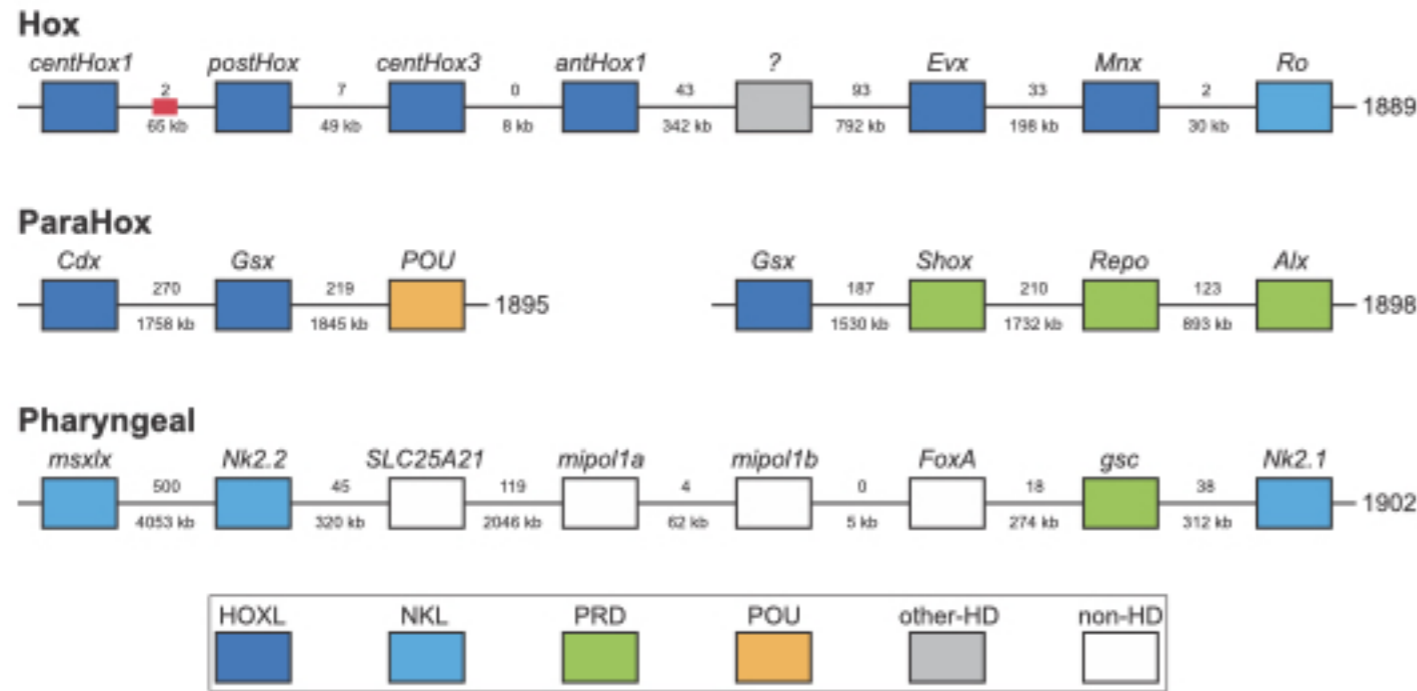


Figure 7: *X. bocki* has 5 HOX genes, which are located in relatively close proximity on one of our chromosome size scaffolds. Similar clusters exist for the ParaHox and “pharyngeal” genes. Numbers between genes are distance (below) and number of genes between (below). Colours indicate gene families. Red box marks the position of a partial Hox gene. The “?” gene has an unresolved homeodomain identity.

8a



8b

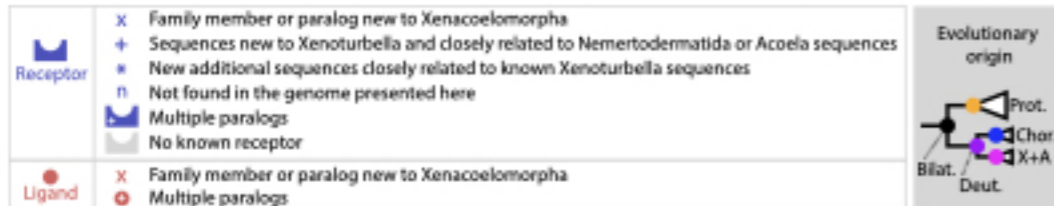
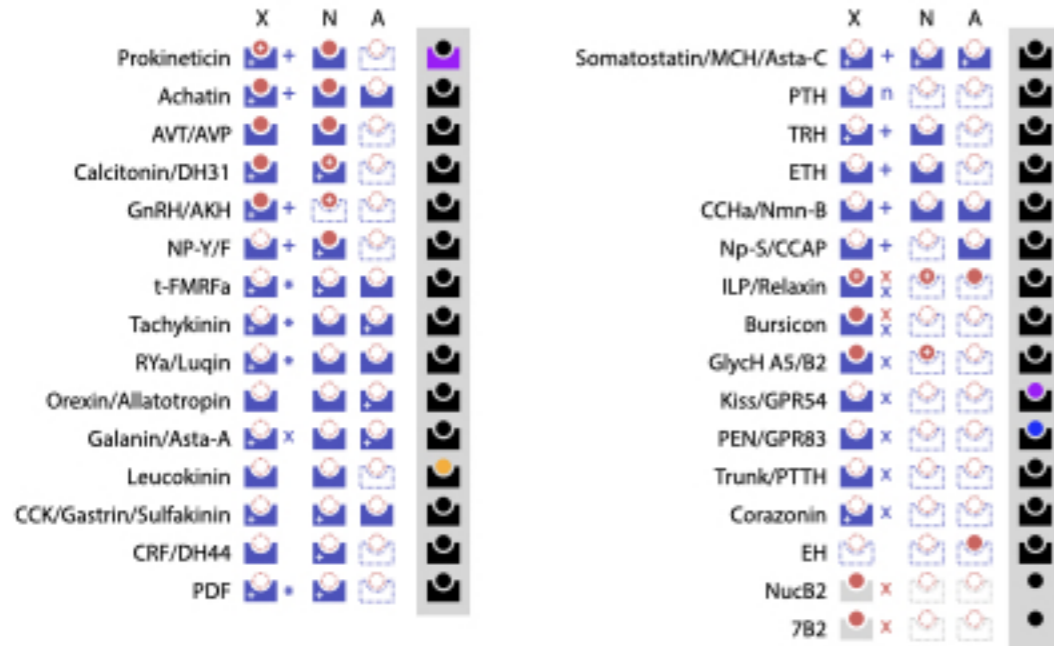


Figure 8: (a/b) We found a specific prokineticin ligand signature sequence in *X. bocki*, which was previously reported for Ecdysozoa and Chordata, as well as a “K/R-RFP-K/R”, sequence shared only by ambulacrarians and *X. bocki*. The signature previously reported for Ecdysozoa and Chordata, as well as new signatures we found in Spiralia and Cnidaria is absent from ambulacrarian and *X. bocki* prokineticin ligand sequences. The inset cladogram in (b) depicts the evolutionary origin of sequences in accordance with our analysis: **Bilateral**, **Protostomia**, **Chordate**, **Xenacoelomorpha + Ambulacraria** last common ancestor respectively.

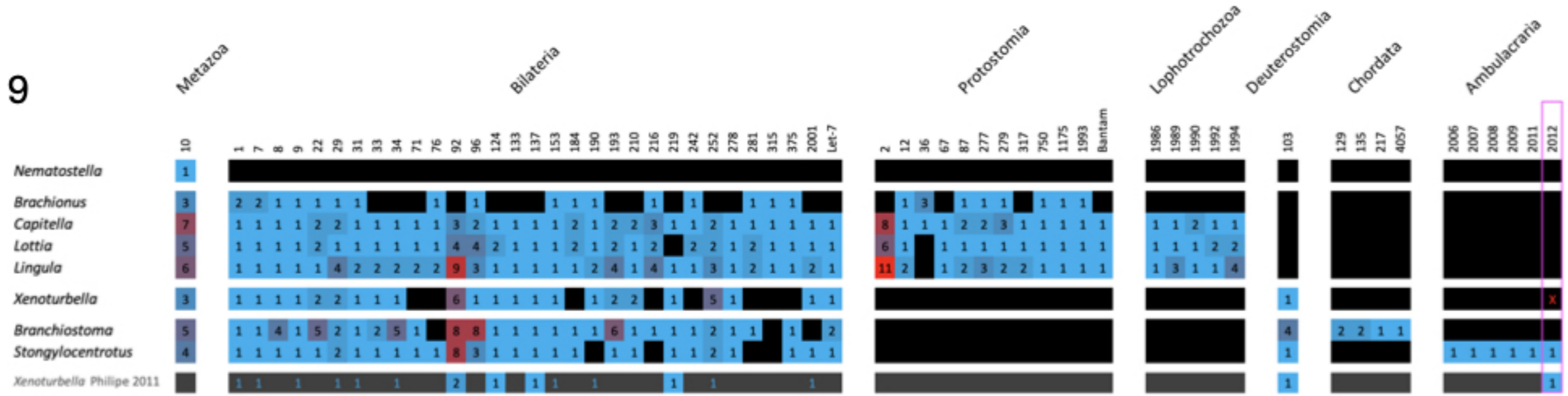


Figure 9: The revised microRNA complement of *X. bocki* has a near complete set of metazoan, bilaterian and deuterostome families and genes. Presence (color) and absence (black) of microRNA families (column), paralogue numbers (values & heatmap coloring) organized in node-specific blocks in a range of 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*.

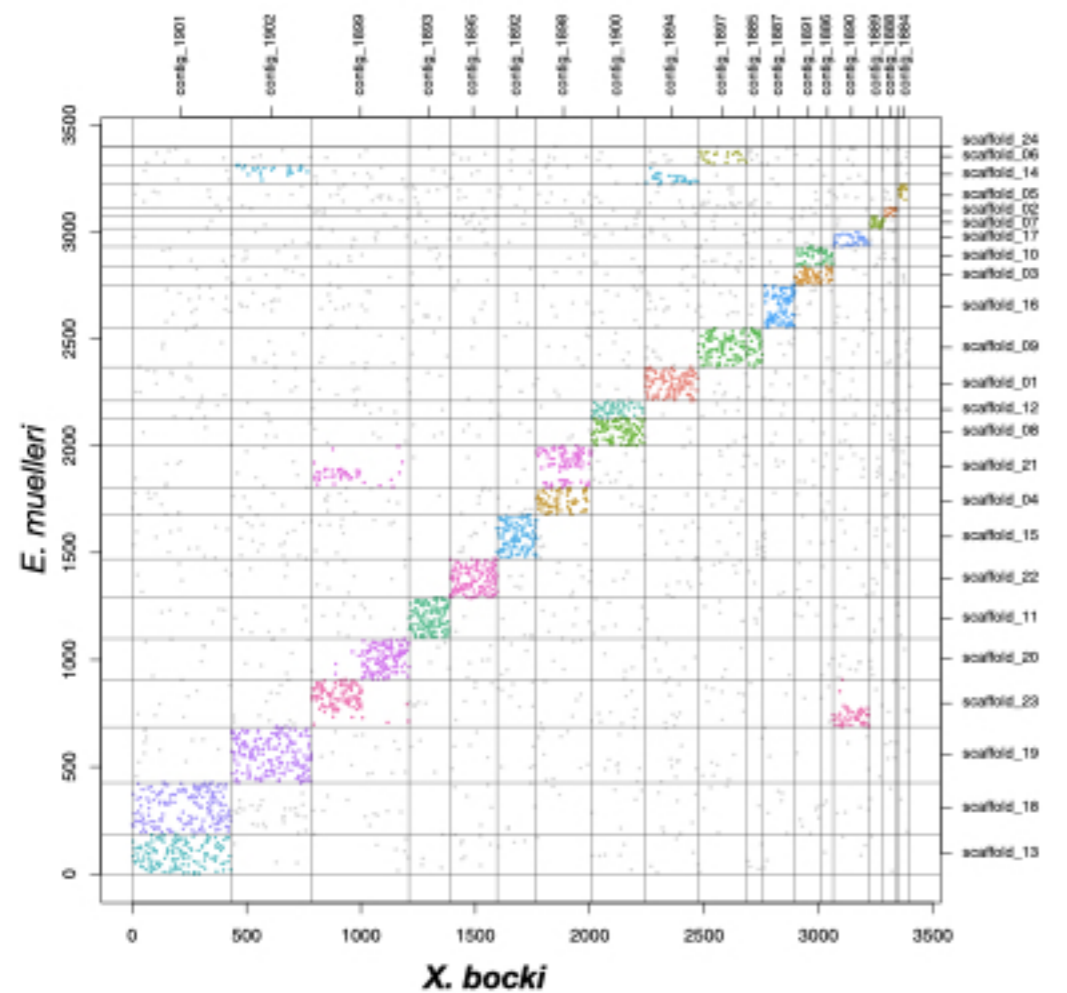
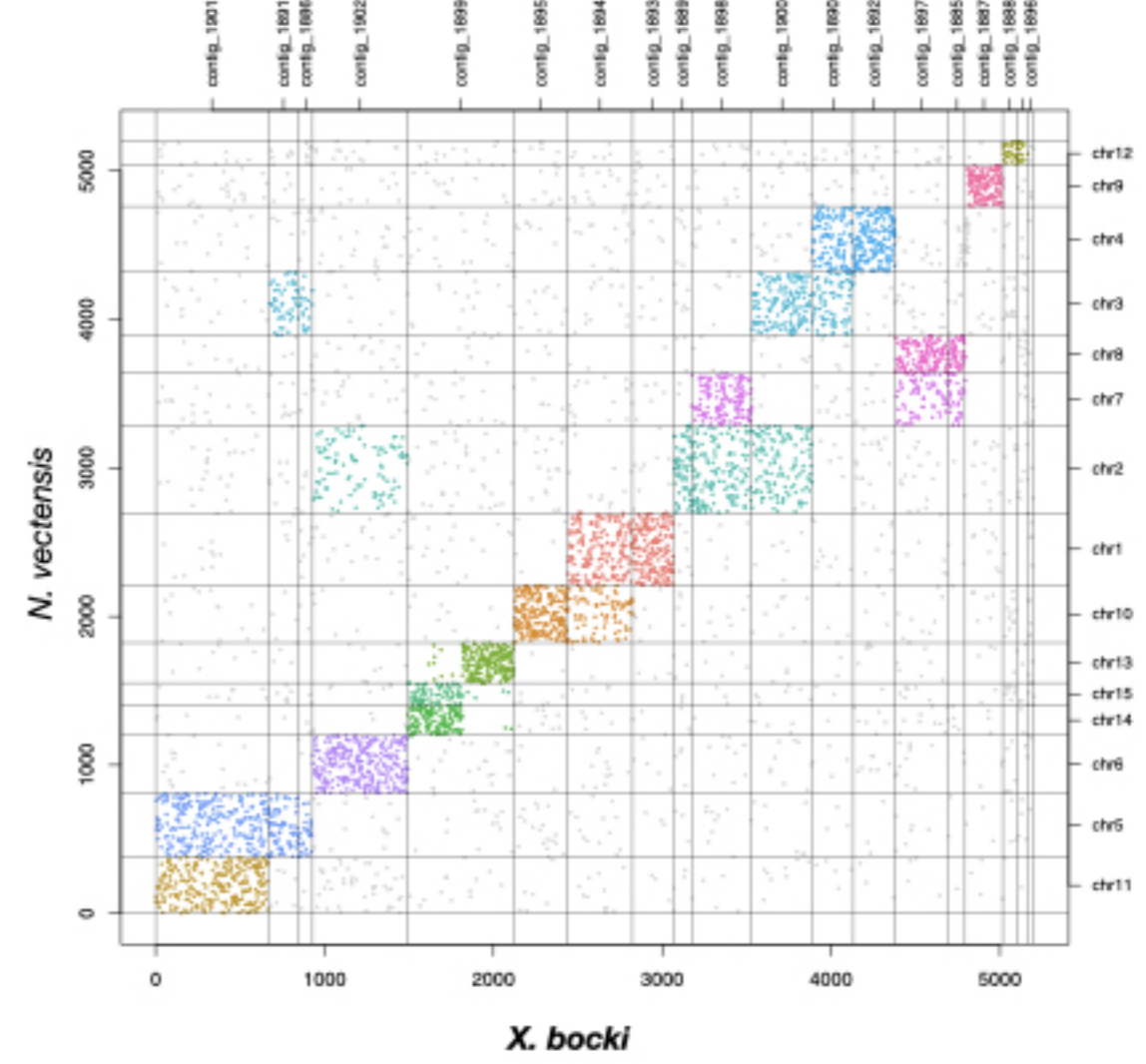
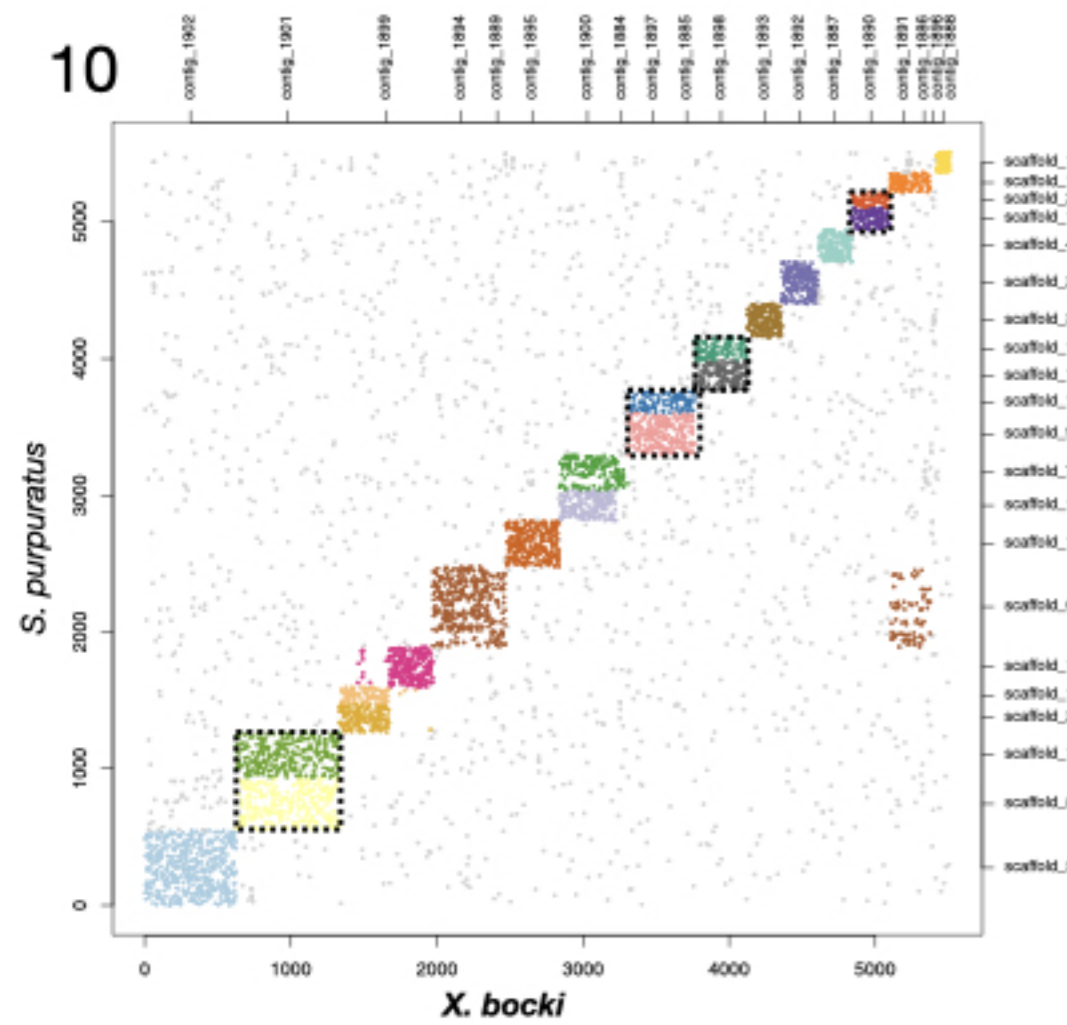


Figure 10: A comparison of scaffolds in the *X. bocki* genome with other Metazoa. 17 of the 18 large scaffolds in the *X. bocki* genome are linked via synteny to distinct chromosomal scaffolds in these species.

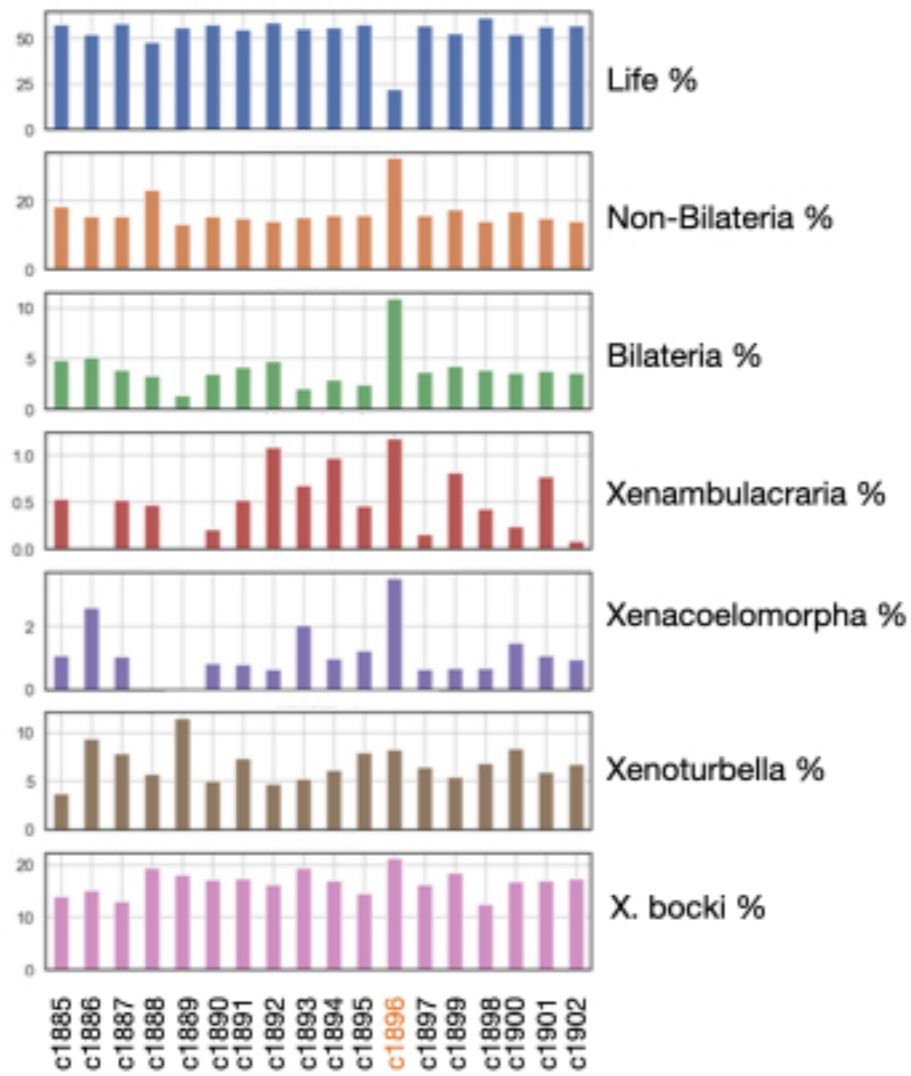


Figure 11: Phylostratigraphic age distribution of genes on all major scaffolds in the *X. bocki* genome. One scaffold (c1896), which showed no synteny to a distinct chromosomal scaffold in the other metazoan species also had a divergent gene age structure in comparison to other *X. bocki* scaffolds.



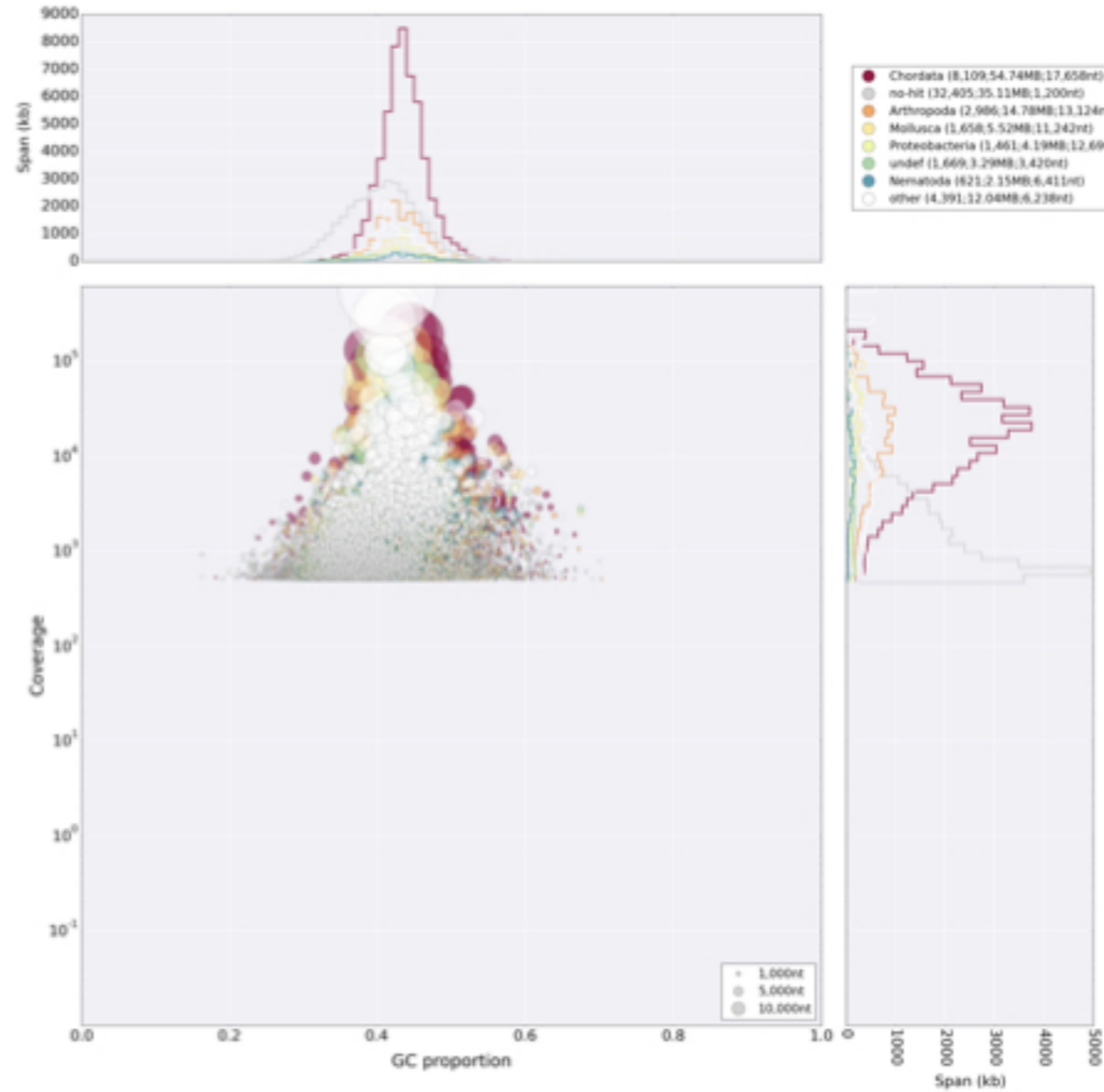
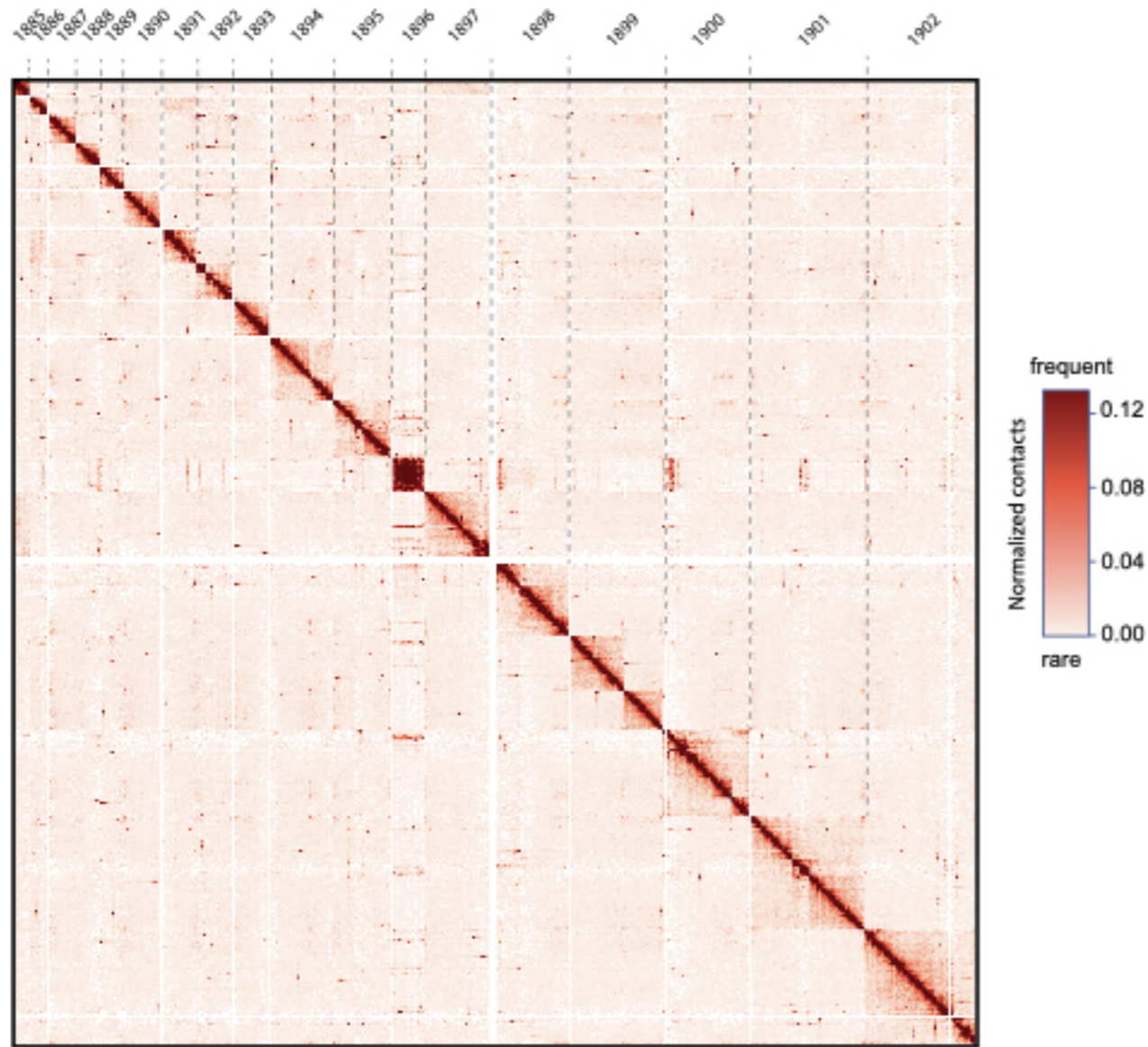


Figure 12: Blobplot analysis of the primary Illumina genome assembly. The assembly shows no major microorganismal contamination, apart from the Chlamydia and Gammaproteobacteria described in the main text. The diamond tool was used to blast against the UniProt database for this analysis.

13a



13b

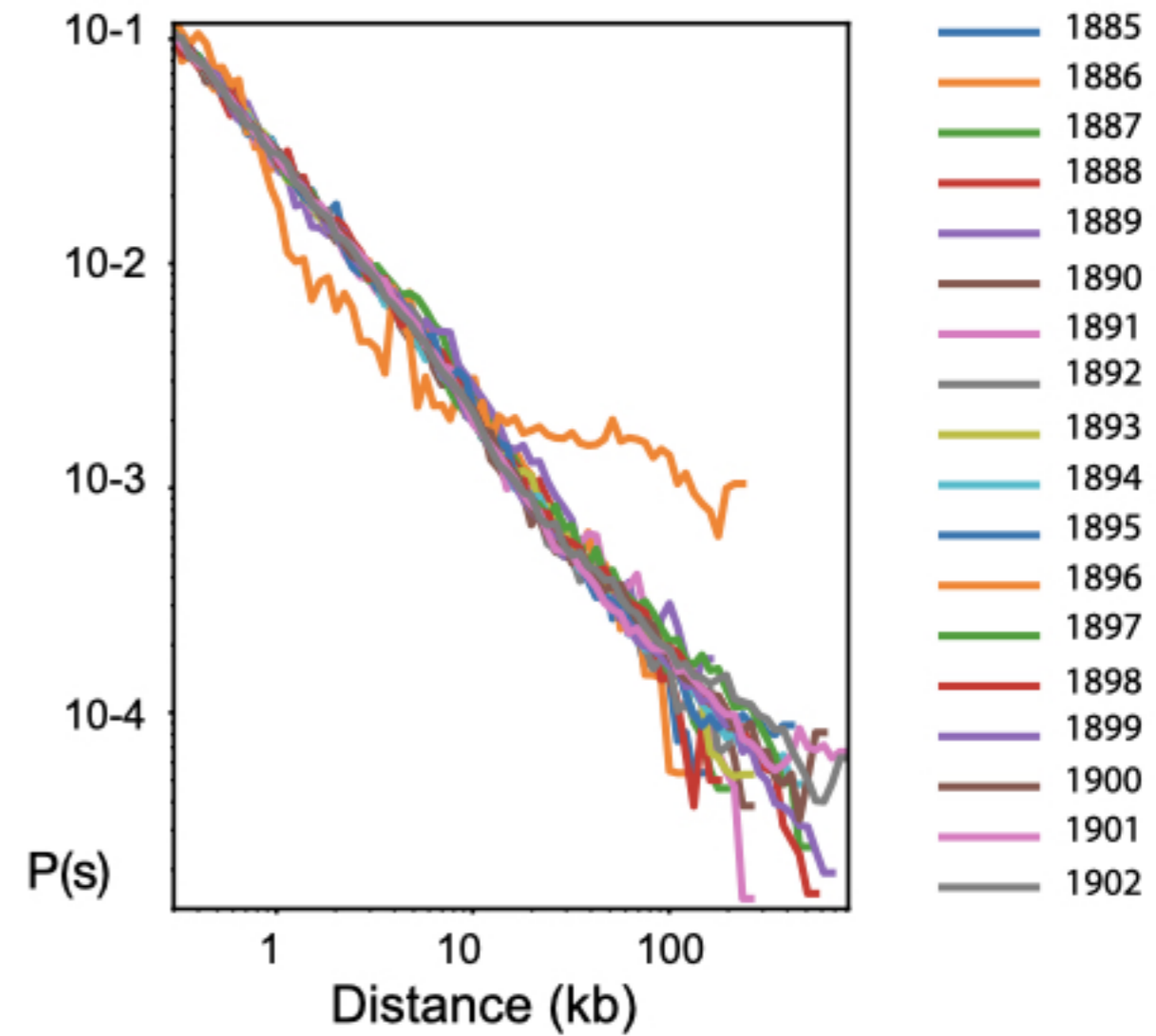


Figure 13: (a) Contact frequency map of the largest 18 scaffolds and (b) distribution of contact frequency as function of distance (distance law).

Figure 8 – supplement 1.

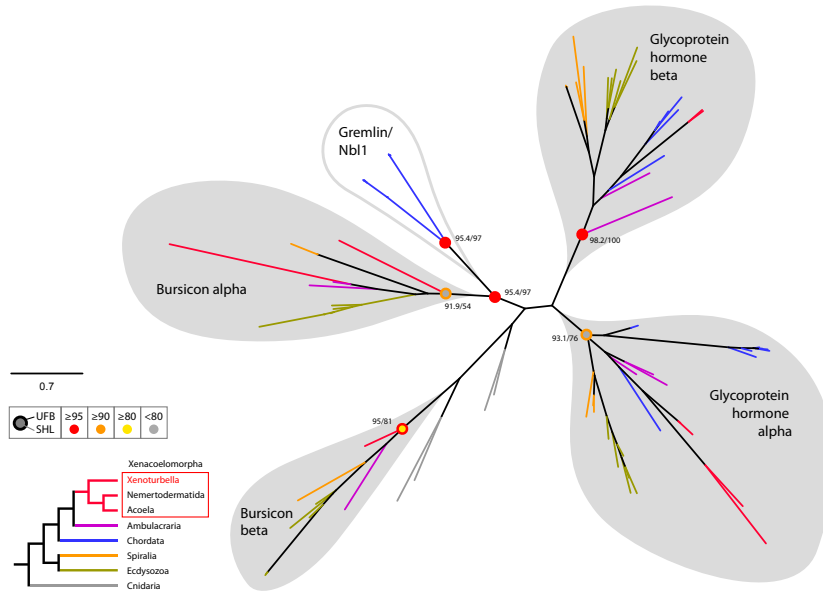


Figure 8 – supplement 2.

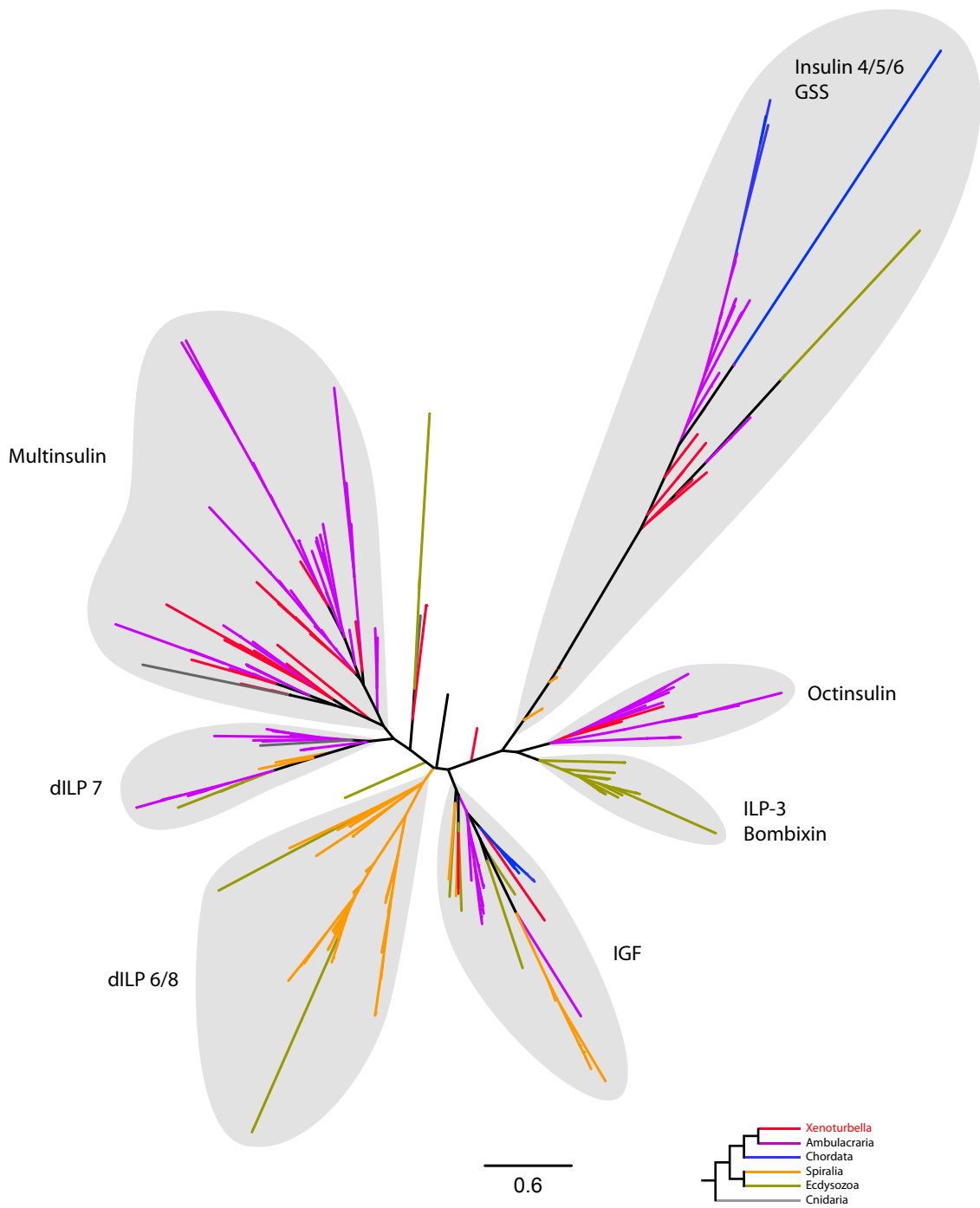






Figure 8 – supplement 5.

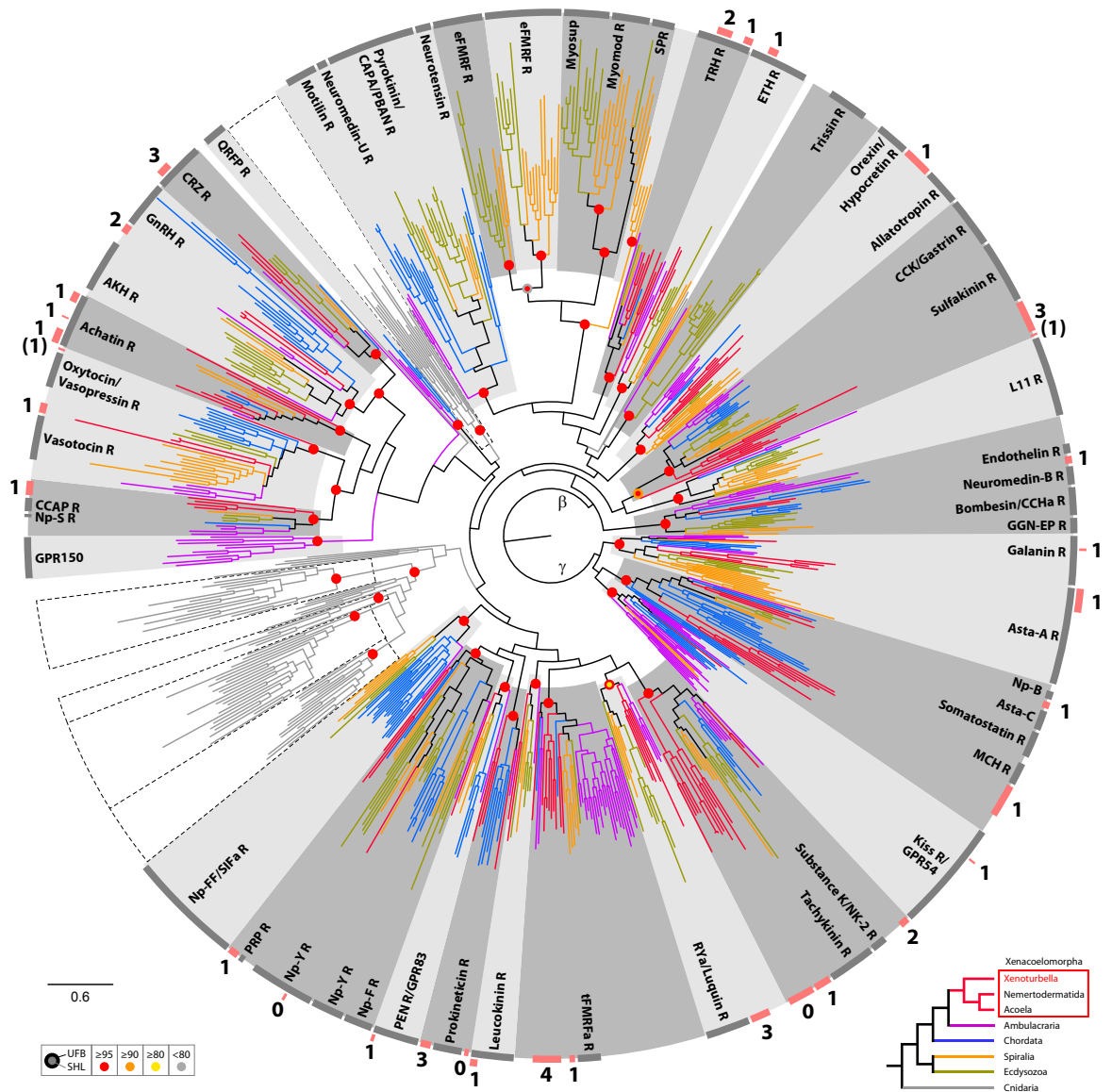
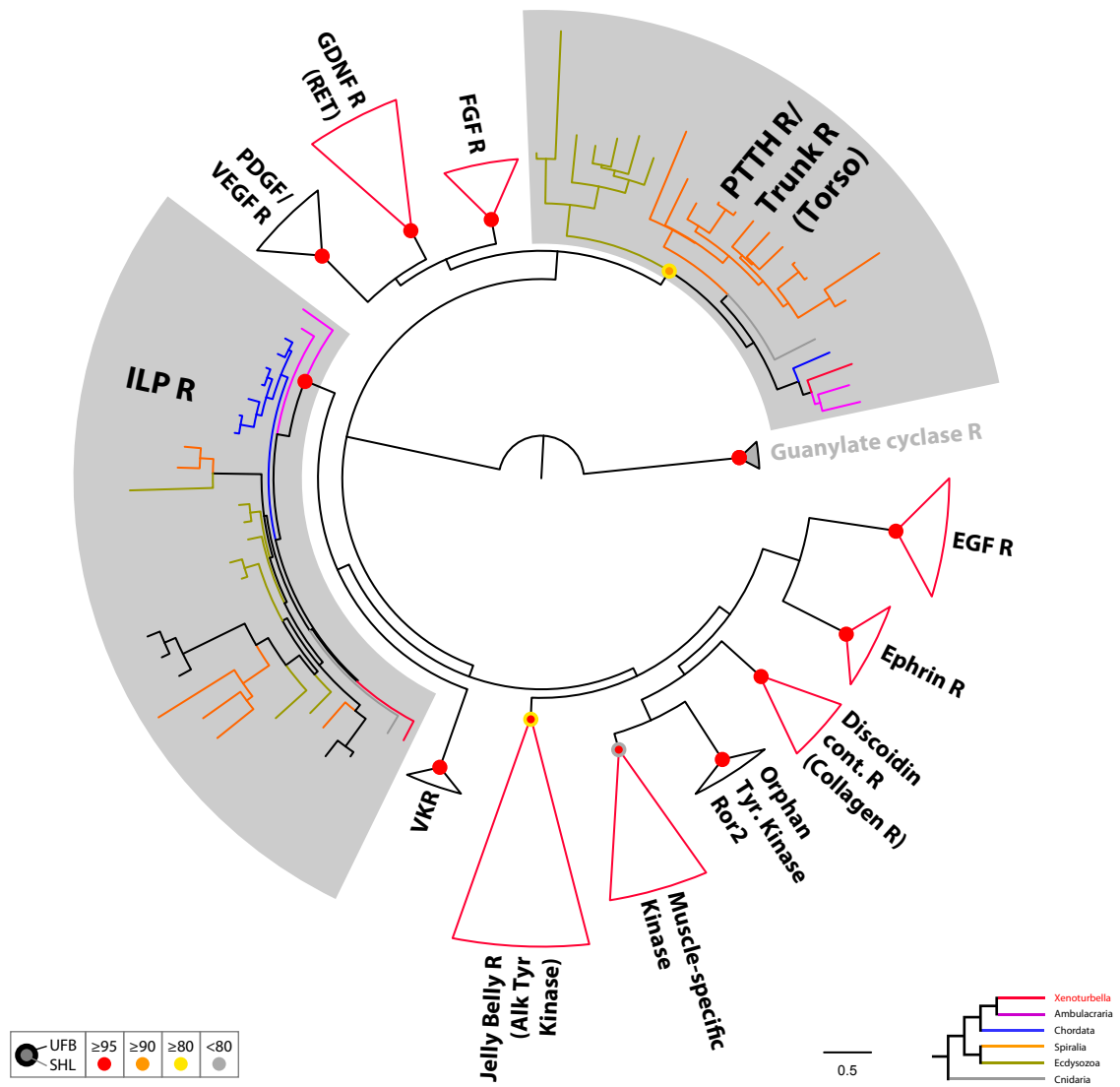


Figure 8 – supplement 6.





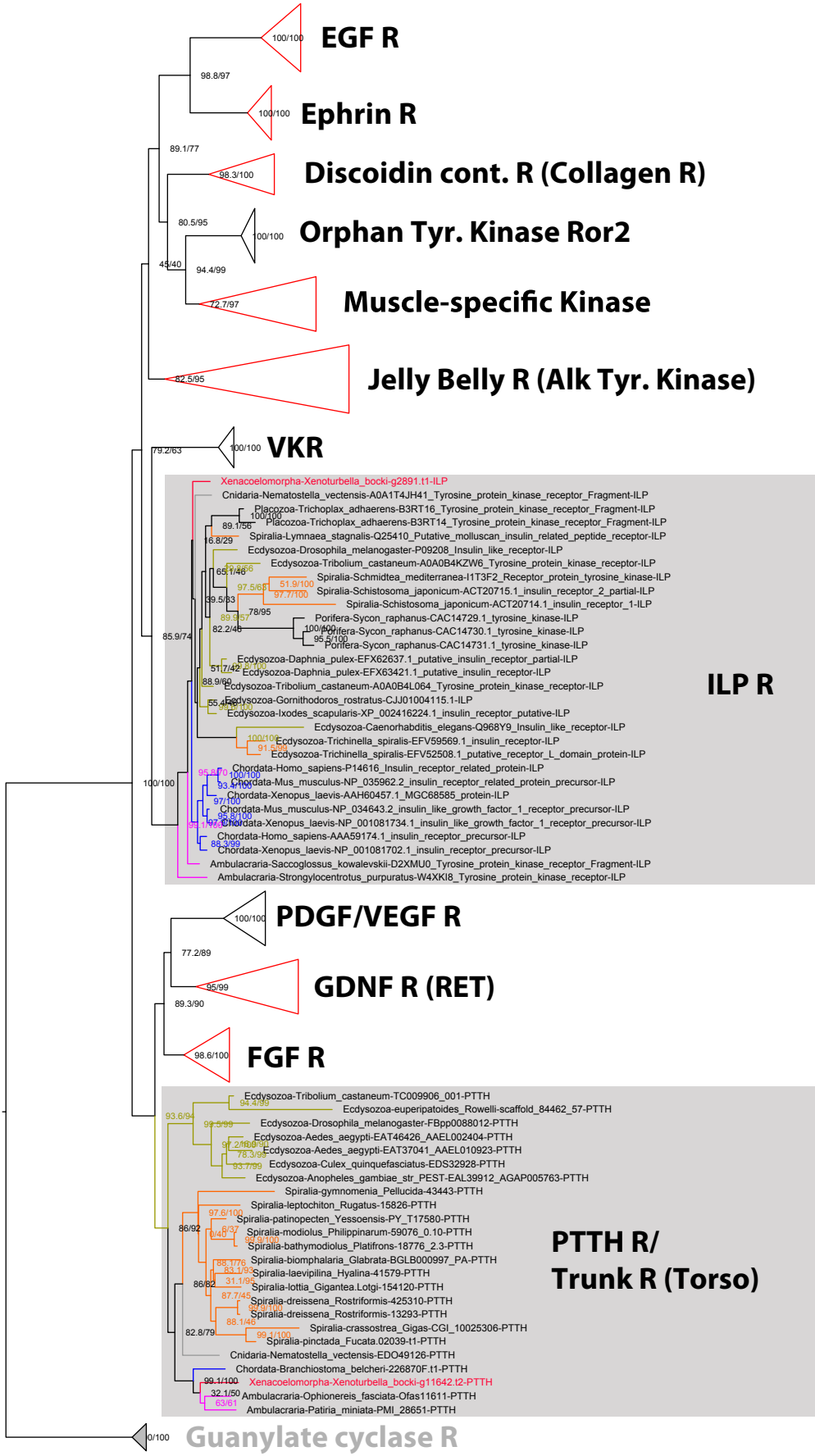


Figure 8 – supplement 8.

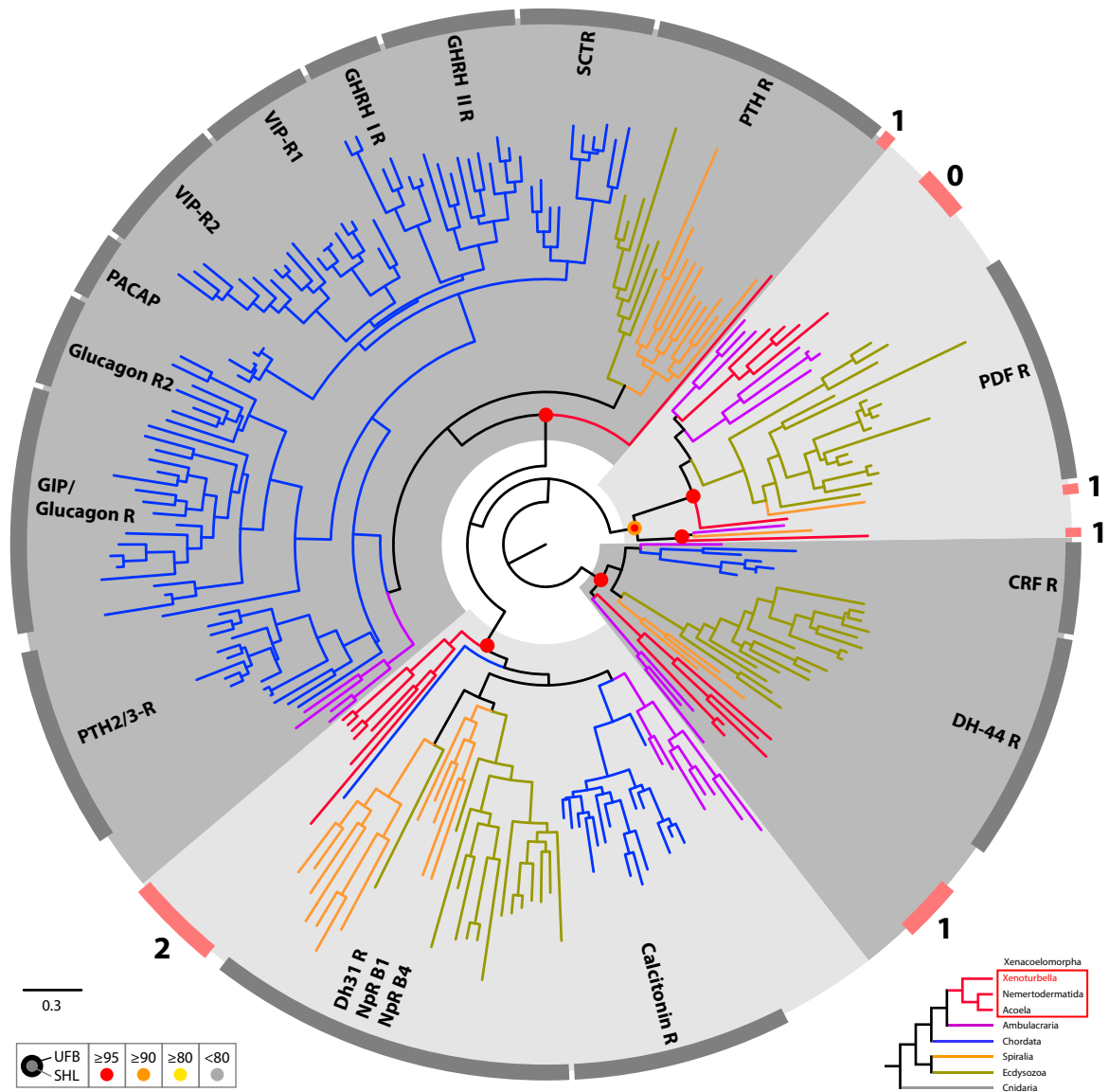


Figure 10 – supplement 1.

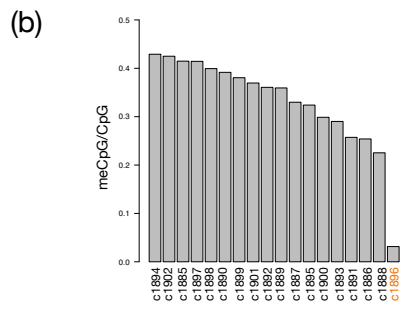
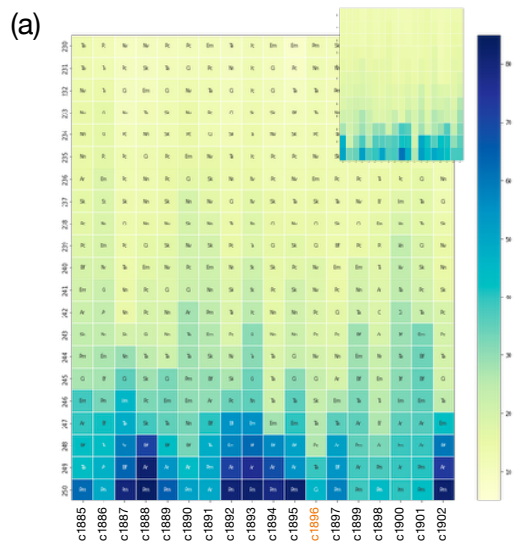


Figure 10 – supplement 2.

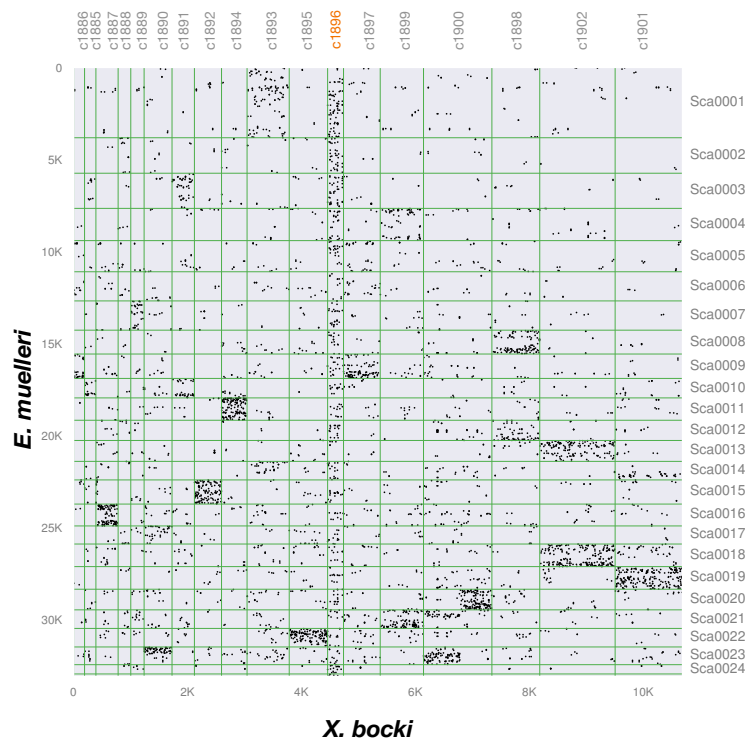


Figure 13 – figure supplement 1

