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

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

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

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

Abstract

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

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

Introduction

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

The simplicity of xenacoelomorph species compared to other bilaterians is a central feature of discussions over their evolution. While Xenacoelomorpha are clearly monophyletic, their phylogenetic position within the Metazoa has been controversial for a quarter of a century. There are two broadly discussed scenarios: a majority of studies have supported a position for Xenacoelomorpha as the sister group of all other Bilateria (the Protostomia and Deuterostomia, collectively named Nephrozoa)[4]; work we have contributed to[1,3,6,7], has instead placed Xenacoelomorpha within the Bilateria as the sister group of the Ambulacraria (Hemichordata and Echinodermata) to form a

clade called the Xenambulacraria[6].

Xenoturbella bocki has neither organized gonads nor a centralized nervous system. It has a blind gut, no body cavities and lacks nephrocytes[8]. If Xenacoelomorpha are the sister group to Nephrozoa these character absences can be interpreted as representing the primitive state of the Bilateria. According to advocates of the Nephrozoa hypothesis, these and other characters absent in Xenacoelomorpha must then have evolved in the lineage leading to Nephrozoa after the divergence of Xenacoelomorpha. More generally there has been a tendency to interpret Xenacoelomorpha (especially Acoelomorpha) as living approximations of Urbilateria[9].

An alternative explanation for the simple body plan of xenaceolomorphs is that it is derived from that of more complex urbilaterian ancestors through a major loss of morphological characters found in other bilaterians. The loss of morphological complexity is a common feature of evolution in many animal groups and is typically associated with new modes of living[10,11] – in particular the adoption of a sessile (sea squirts, barnacles) or parasitic (neodermatan flatworms, orthonectids) lifestyle, extreme miniaturization (e.g. tardigrades, orthonectids), or even neoteny (e.g. flightless hexapods).

In the past some genomic features gleaned from analysis of various Xenacoelomorpha have been used to test these evolutionary hypotheses. For example, the common ancestor of the protostomes and deuterostomes has been reconstructed with approximately 8 Hox genes but only 4 have been found in the Acoelomorpha (Nemertoderma) and 5 In *Xenoturbella*. This has been interpreted as a primary absence with the full complement of 8 appearing subsequent to the divergence of Xenacoelomorpha and Nephrozoa. Similarly, analysis of the microRNAs (miRNAs) of an acoelomorph, *Symsagittifera roscoffensis*, found that many bilaterian miRNAs were absent from its genome[12]. Some of the missing bilaterian miRNAs, however, were subsequently observed in *Xenoturbella*[6].

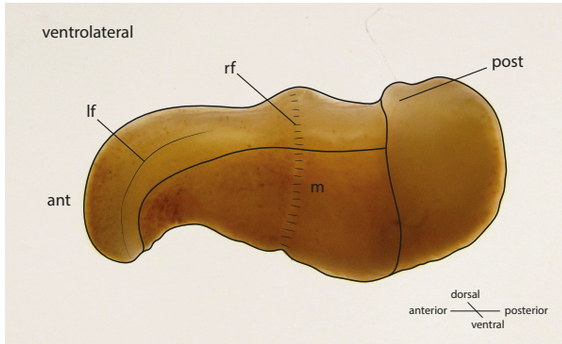
The only Xenacoelomorpha genomes available to date are from the acoel *Hofstenia miamia*[13] – like other Acoelomorpha it shows accelerated sequence evolution relative to *Xenoturbella*[3] – and from *Praesagittifera naikaiensis*[14]. The analyses of gene content of *Hofstenia* showed similar numbers of genes and gene

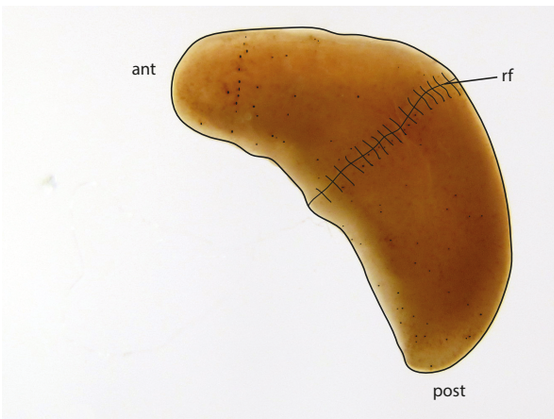
107 families to other bilaterians[13], while an analysis of the neuropeptide content concluded
108 that most bilaterian neuropeptides were present in Xenacoelomorpha[15].
109 In order to infer the characteristics of the ancestral xenacoelomorph genome,
110 and to complement the data from the Acoelomorpha, here we describe a high-quality
111 genome of the slowly evolving xenacoelomorph *Xenoturbella bocki*. This allows us to
112 contribute knowledge of Xenacoelomorpha and *Xenoturbella* in particular with
113 genomic traits, such as gene content and genome-structure and helping to reconstruct
114 the genome structure and composition of the ancestral xenacoelomorph.
115

116 Results

117 **Assembly of a draft genome of *Xenoturbella bocki*.**
118 We collected *Xenoturbella bocki* specimens (Fig. 1) from the bottom of the fjord close
119 to the biological field station in Kristineberg (Sweden). These adult specimens were
120 starved for several days in tubes with artificial sea water, and then sacrificed in lysis
121 buffer. We extracted high molecular weight (HMW) DNA from single individuals for
122 each of the different sequencing steps below.
123 We assembled a high-quality draft genome of *Xenoturbella bocki* using one short
124 read Illumina library and one TruSeq Synthetic Long Reads (TSLR) Illumina library.
125 We used a workflow based on a primary assembly with SPAdes (Methods; [16]). The
126 primary assembly had an N50 of ~62kb over 23,094 contigs and scaffolds spanning
127 ~121Mb. The longest scaffold was 960,978kb.
128 The final genome was obtained using the redundans pipeline, and Hi-C
129 scaffolding using the program instaGRAAL (Methods;[17]). The scaffolded genome has
130 a span of 111 Mbp (117 Mbp including small fragments unincorporated into the HiC
131 assembly) and an N50 of 2.7 Mbp (for contigs >500bp). The assembly contains 18
132 megabase-scale scaffolds encompassing 72 Mbp (62%) of the genomic sequence,
133 with 43% GC content. The original assembly indicated a repeat content of about 25%
134 after a RepeatModeller based RepeatMasker annotation (Methods). As often seen in
135 non-model organisms, about 2/3 of the repeats are not classified.
136 We used BRAKER1[18,19] with extensive RNA-Seq data, and additional single-cell
137 UTR enriched transcriptome sequencing data to predict 15,154 gene models. 9,575
138 gene models (63%) are found on the 18 large scaffolds (which represent 62% of the

1a





1b

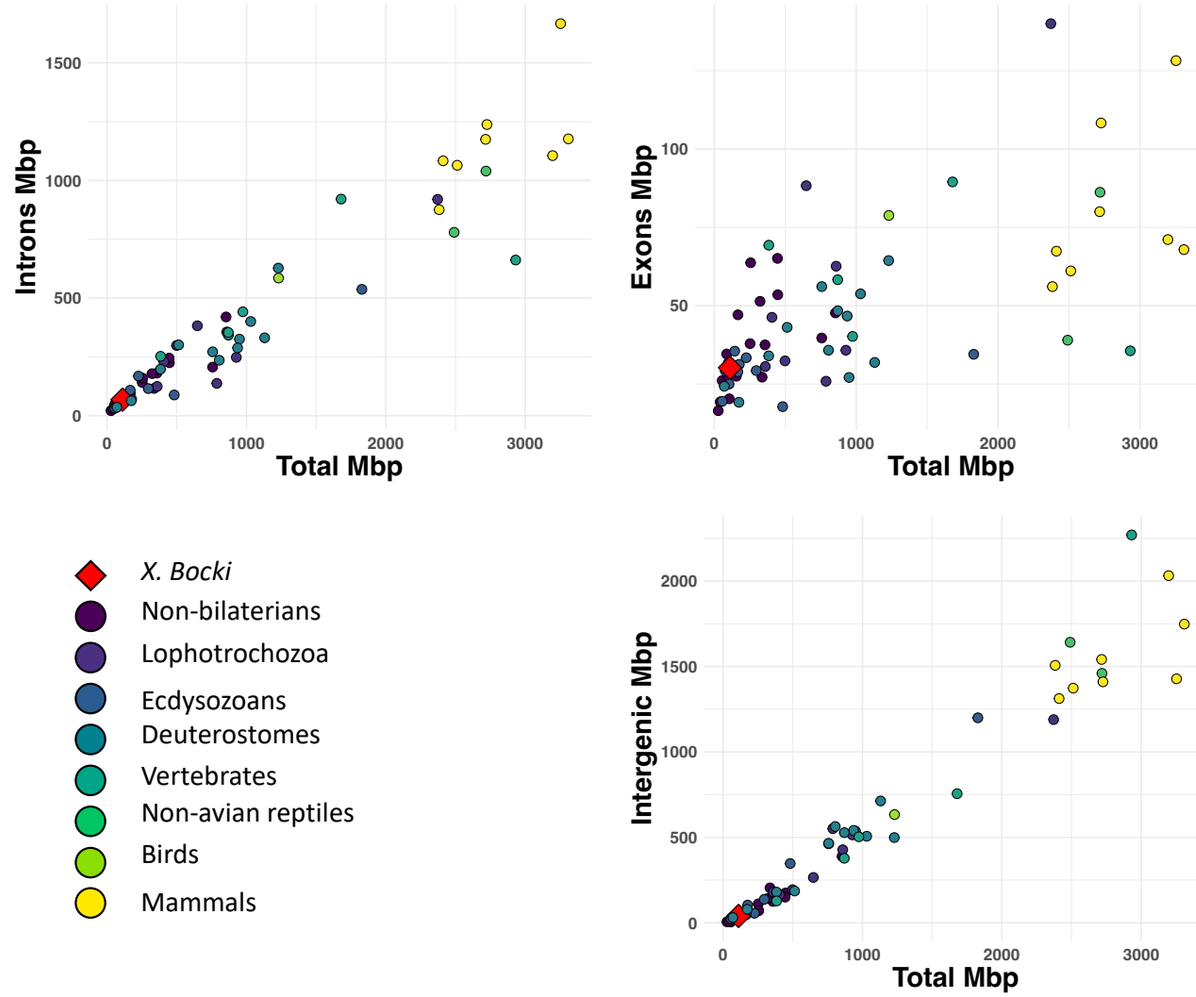


Figure 1: (a) Schematic drawings of *Xenoturbella bocki* showing the simple body organisation of the marine vermiform animal. (b) A comparison of total length of exons, introns, and intergeneic space in the *X. bocki* genome with other metazoans (data from ref 20). *X. bocki* does not appear to be an outlier in any of theses comparisons.

139 total sequence). 13,298 of our predicted genes (88%) have RNA-Seq support.
140 Although at the low end of bilaterian gene counts, we note that our RNA-seq libraries
141 were all taken from presumably adult animals and thus may not represent the true
142 complexity of the gene complement. We consider our predicted gene number to be a
143 lower bound estimate for the true gene content.

144 The predicted *X. bocki* genes have a median coding length of 873 nt and a mean
145 length of 1330 nt. Median exon length is 132 nt (mean 212 nt) and median intron length
146 is 131 nt (mean 394 nt). Genes have a median of 4 exons and a mean of 8.5 exons.
147 2,532 genes have a single exon and, of these, 1,381 are supported as having a single
148 exon by RNA-Seq (TPM>1). A comparison of the exon, intron, and intergenic
149 sequence content in *Xenoturbella* with descriptions of other animal genomes[20] show
150 that *X. bocki* falls within the range of other similarly sized metazoan genomes (Fig. 1b)
151 for all these measures.

152

153 <u>The genome of a co-sequenced *Chlamydia* species</u>

154 We recovered the genome of a marine *Chlamydia* species from Illumina data obtained
155 from one *X. bocki* specimen and from Oxford Nanopore data from a second specimen
156 supporting previous microscopic analyses and single gene PCRs suggesting that *X.*
157 *bocki* is host to a species in the bacterial genus *Chlamydia*. The bacterial genome was
158 found as 5 contigs spanning 1,906,303 bp (N50 of 1,237,287 bp) which were
159 assembled into 2 large scaffolds. Using PROKKA[21] we predicted 1,738 genes in this
160 bacterial genome, with 3 rRNAs, 35 tRNAs, and 1 tmRNA. The genome is 97.5%
161 complete for bacterial BUSCO[22] genes, missing only one of the 40 core genes.

162 Marine chlamydiae are not closely related to the group of human pathogens[23]
163 and we were not able to align the genome of the Chlamydia-related symbioint from *X.*
164 *bocki* to the reference strain *Chlamydia trachomatis* F/SW4, nor to *Chlamydophila*
165 *pneumoniae* TW-183. To investigate the phylogenetic position of the species co-
166 occurring with *Xenoturbella*, we aligned the 16S rRNA gene from the *X. bocki*-hosted
167 *Chlamydia* with orthologs from related species including sequences of genes amplified
168 from DNA/RNA extracted from deep sea sediments. The *X. bocki*-hosted *Chlamydia*
169 belong to a group designated as Simkaniaceae in[23], with the sister taxon in our

2a

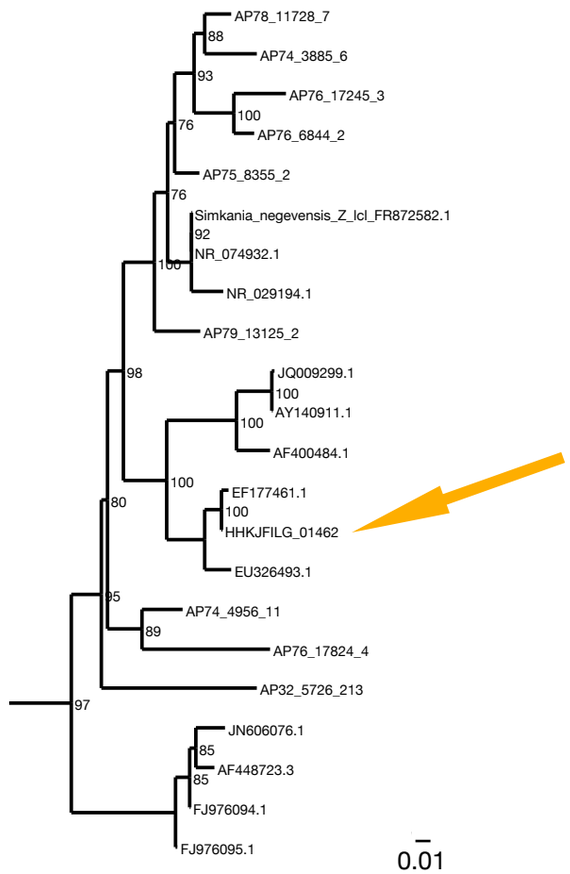
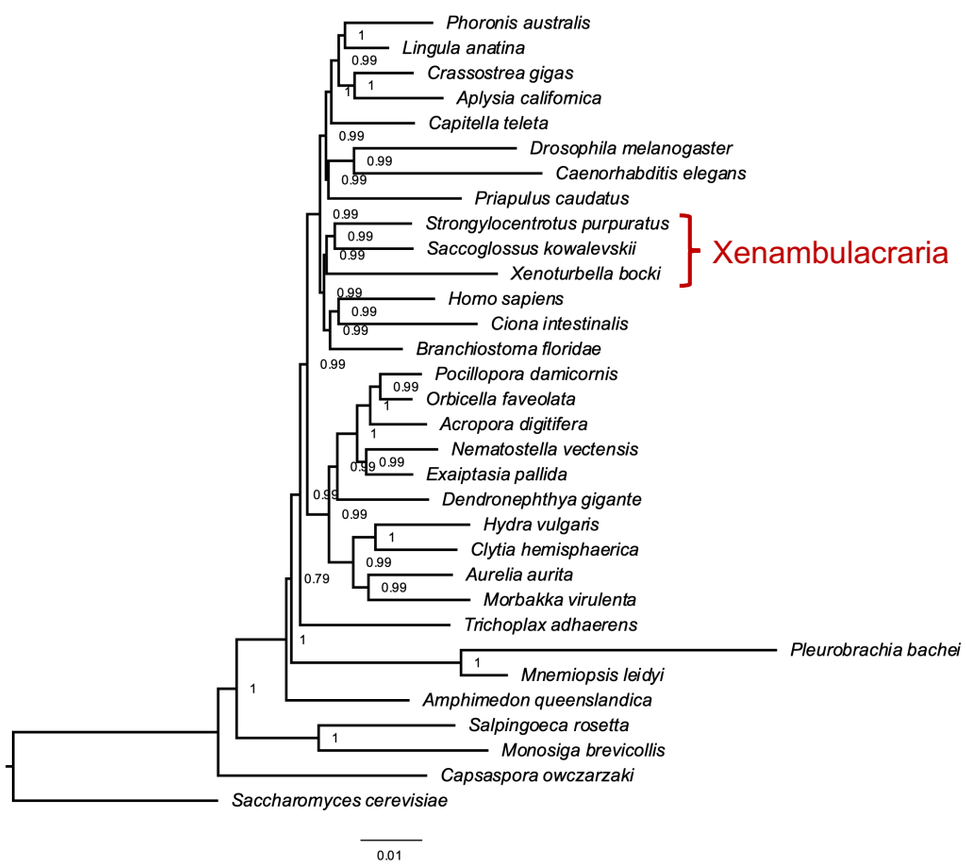
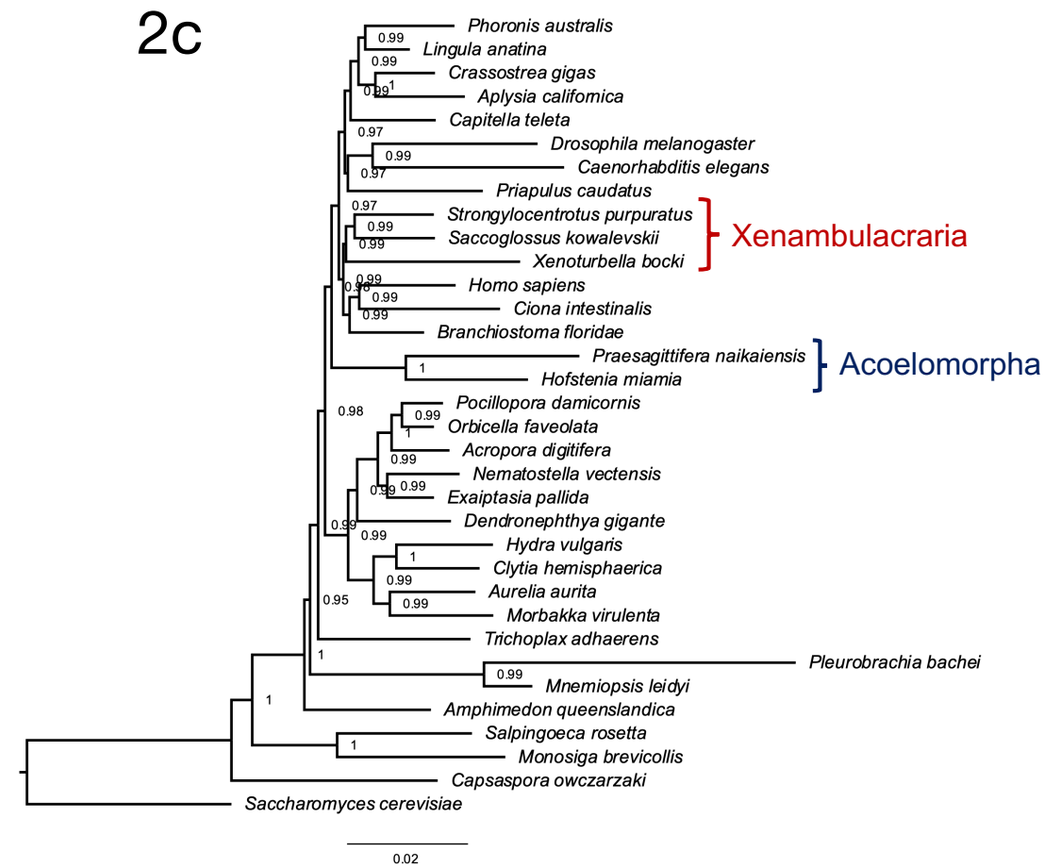


Figure 2a: *Xenoturbella bocki* harbours a marine Chlamydiae species as potential symbiont. In the phylogenetic analysis of 16S rDNA (ML: GTR+F+R7; bootstrap values included) the bacteria in our *X. bocki* isolate (arrow) are sister to a previous isolate from *X. westbaldi*. *X. westbaldi* is most likely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

2b



2c



170 phylogenetic tree being the *Chlamydia* species previously found in *X. westbladi* (*X.*
171 *westbladi* is almost certainly a synonym of *X. bocki*)[24] (Fig. 2a).
172 To investigate whether the *X. bocki*-hosted Chlamydia might contribute to the
173 metabolic pathways of its host, we compared the completeness of metabolic pathways
174 in KEGG for the *X. bocki* genome alone and for the *X. bocki* genome in combination
175 with the bacteria. We found only slightly higher completeness in a small number of
176 pathways involved in carbohydrate metabolism, carbon fixation, and amino acid
177 metabolism (Supplementary) suggesting that the relationship is likely to be
178 commensal or parasitic rather than a true symbiosis.
179 A second large fraction of bacterial reads, annotated as Gammaproteobacteria,
180 were identified and filtered out during the data processing steps. These bacteria were
181 also previously reported as potential symbionts of *X. bocki*[25]. However, these
182 sequences were not sufficiently well covered to reconstruct a genome and we did not
183 investigate them further.
184
185 **The *X. bocki* molecular toolkit is typical of bilaterians.**
186 The general completeness of the *X. bocki* gene set allowed us to use the presence
187 and absence of genes identified in our genomes as a source of information to find the
188 best supported phylogenetic position of the Xenacoelomorpha. We conducted two
189 separate phylogenetic analyses of gene presence/absence data: one including the
190 fast evolving Acoelomorpha and one without. In both analyses the best tree grouped
191 *Xenoturbella* with the Ambulacraria (Fig. 2b). The analysis including acoels, however,
192 placed this taxon separate from *Xenoturbella* as the sister-group to Nephrozoa (Fig.
193 2c). Because other data have shown the monophyly of Xenacoelomorpha to be robust,
194 we interpret this result as being the result of systematic error caused by a high rate of
195 gene loss or by orthologs being incorrectly scored as missing due to higher rates of
196 sequence evolution in acoelomorphs[26].
197 One of our principal aims was to ask whether the *Xenoturbella* genome lacks
198 characteristics otherwise present in the Bilateria. We found the set of *X. bocki* proteins
199 translated from our gene predictions to be 90% complete for the Metazoa gene set in
200 BUSCO (v5). This estimate is similar to the acoel *Hofstenia miamia*, which was
201 originally reported to be 90%[13], but in our re-analysis was 95.71%. In comparison, the
202 morphologically highly simplified and fast evolving annelid *Intoshia linei*[27] has a

203 genome of fewer than 10,000 genes[28] and in our analysis is only ~64% complete for
204 the BUSCO (v5) Metazoa set. The model nematode *Caenorhabditis elegans* is ~81%
205 complete for the same set. Despite the morphological simplicity of both *Xenoturbella*,
206 and even more so the acoel *Hofstenia,* the Xenacoelomorpha show few absences of
207 core genes compared to other fast evolving bilaterian lineages.

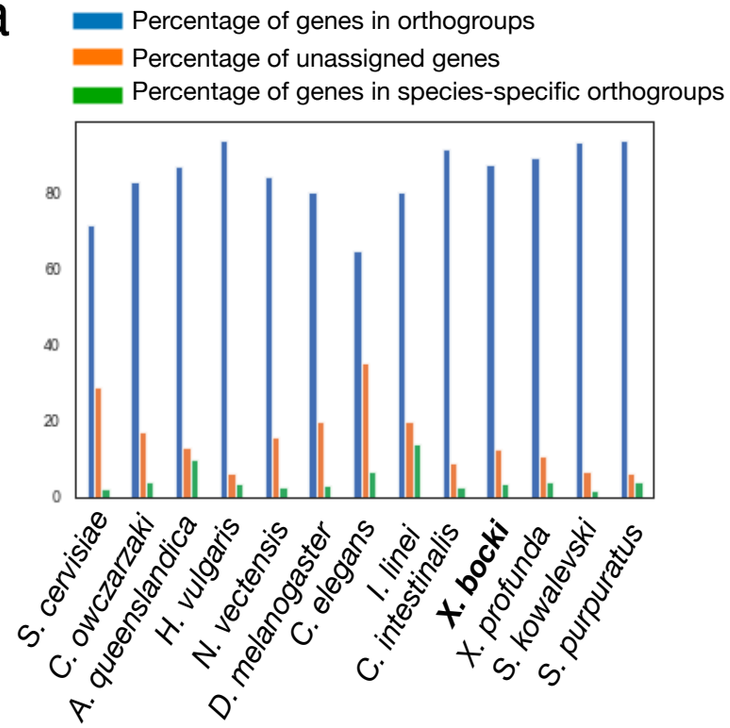
208 Using our phylogenomic matrix of gene presence/absence (see above) we
209 identified all orthologs present in any bilaterian and any non-bilaterian; these must
210 have existed in the bilaterian ancestor. All individual bilaterian genomes were missing
211 many of these orthologs but Xenacoelomorphs and some other derived bilaterians
212 lacked more of these than most other taxa. Average number of these genes present
213 in bilaterians = 7577; *Xenoturbella* = 5459; *Hofstenia* = 5438; *Praesagittifera* = 4280;
214 *Drosophila* = 4844; *Caenorhabditis* = 4323.

215 To better profile the *Xenoturbella* and xenacoelomorph molecular toolkit, we
216 used OrthoFinder to conduct orthology searches in a comparison of 155 metazoan
217 and outgroup species, including the transcriptomes of the sister species *X. profunda*
218 and an early draft genome of the acoel *Paratomella rubra* we had available*,* as well as
219 the *Hofstenia* and *Praesagittifera* proteomes (Supplementary online material). For
220 each species, we counted, in each of the three Xenacoelomorphs, the number of
221 orthogroups for which a gene was present. The proportion of orthogroups containing
222 an *X. bocki* and *X. profunda* protein (87.4% and 89.2%) are broadly similar to the
223 proportions seen in other well characterised genomes, for example *S. purpuratus*
224 proteins (93.8%) or *N. vectensis* proteins (84.3%) (Fig 3a). In this analysis, the fast-
225 evolving nematode *Caenorhabditis elegans* appears as an outlier, with only ~64% of
226 its proteins in orthogroups and ~35% unassigned. Both *Xenoturbella* species have an
227 intermediate number of unassigned genes of ~11-12%. Similarly, the proportion of
228 species-specific genes (~14% of all genes) corresponds closely to what is seen in
229 most other species (with the exception of the parasitic annelid *I. linei*, Fig. 3a).
230

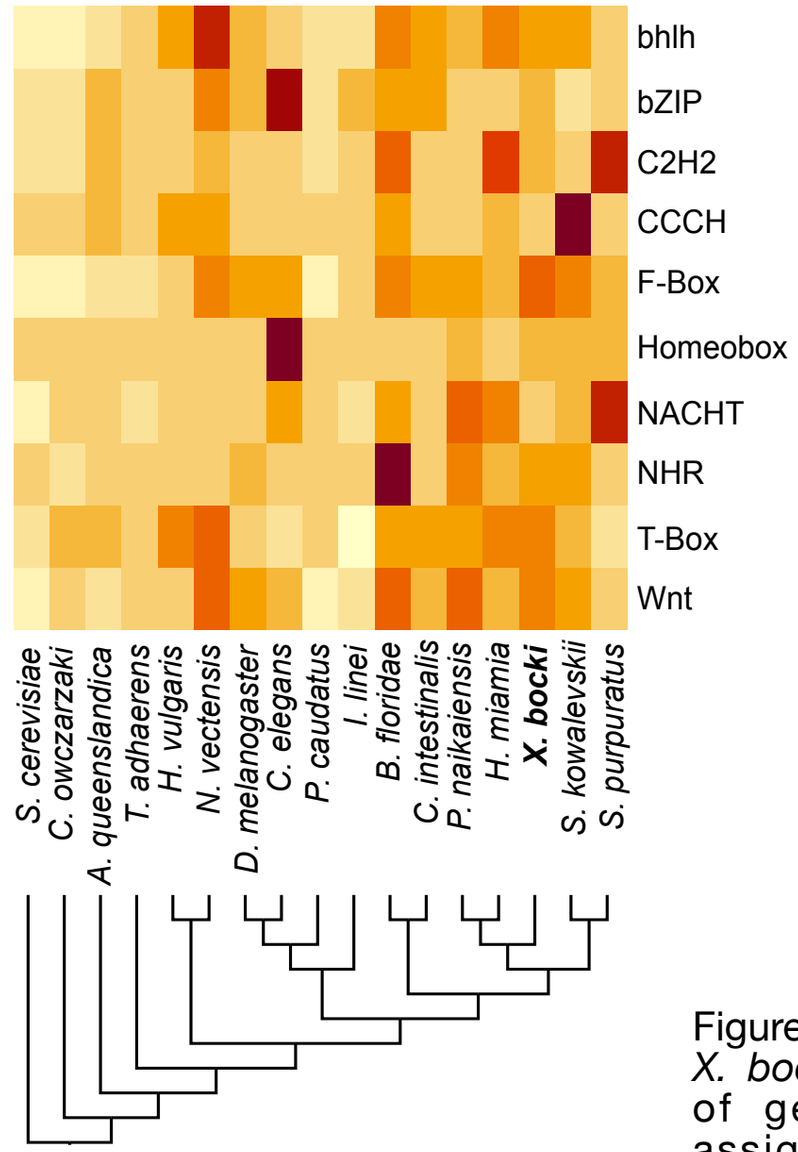
231 <u>Idiosyncrasies of *Xenoturbella*</u>

232 In order to identify sets of orthologs specific to the two *Xenoturbella* species we used
233 the kinfin software[29] and found 867 such groups in the OrthoFinder clustering. We
234 profiled these genes based on Pfam domains and GO terms derived from

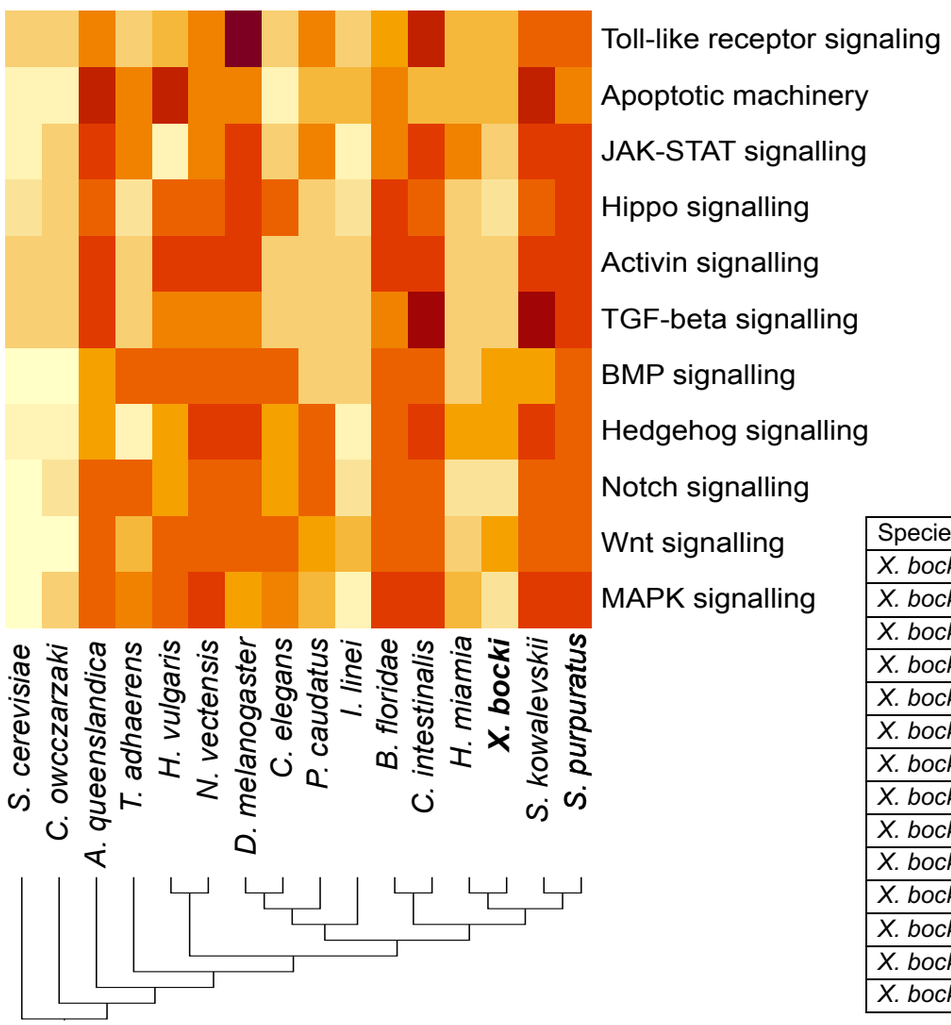
3a



3b



3c



t-test

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

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

235 InterProScan. While these *Xenoturbella* specific proteins fall into diverse classes, we
236 did see a considerable number of C-type lectin, Immunoglobulin-like, PAN, and Kringle
237 domain containing Pfam annotations. Along with the Cysteine-rich secretory protein
238 family and the G-protein coupled receptor activity GO terms, these genes and families
239 of genes will be important for future studies into *Xenoturbella* biology in its native
240 environment.
241
242 Gene families and signaling pathways are retained in *X. bocki*
243 In our orthology clustering we did not see an inflation of *Xenoturbella*-specific groups
244 in comparison to other taxa, but also no conspicuous absence of major gene families
245 (Fig. 3b). Family numbers of transcription factors like Zinc-fingers or homeobox-
246 containing genes, as well as, for example, NACHT-domain encoding genes seem to
247 be neither drastically inflated nor contracted in comparison to other species in our
248 InterProScan based analysis.
249 To catalogue the completeness of cell signalling pathways we screened the *X.*
250 *bocki* proteome against KEGG pathway maps using GenomeMaple[30]. The *X. bocki*
251 gene set is largely complete in regard to core proteins of these pathways, while an
252 array of effector proteins is absent (Fig. 3c). In comparison to other metazoan species,
253 as well as a unicellular choanoflagellate and a yeast, the *X. bocki* molecular toolkit has
254 significantly lower KEGG completeness than morphologically complex animals such
255 as the sea urchin and amphioxus (t-test; Fig. 3c). *Xenoturbella* is, however, not
256 significantly less complete when compared to other bilaterians considered to have low
257 morphological complexity and which have been shown to have reduced gene content,
258 such as *C. elegans*, the annelid parasite *Intoshia linei*, or the acoel *Hofstenia miamia*
259 (Fig. 3c).
260
261 Clustered homeobox genes in the *X. bocki* genome
262 Acoelomorph flatworms possess three dispersed HOX genes, orthologs of anterior
263 (Hox1), central (Hox4/5 or Hox5) and posterior Hox (HoxP). In contrast, previous
264 analysis of *X. bocki* transcriptomes identified one anterior, three central and one
265 posterior Hox genes. We identified in *Xenoturbella* clear evidence of a syntenic Hox
266 cluster with four Hox genes (antHox1, centHox1, centHox3 and postHox) in the *X.*

267 *bocki* genome (Fig. 4). There was also evidence of a fragmented annotation of
268 centHox2, split between the Hox cluster and a separate scaffold (Fig. 4). In summary,
269 this suggests that all five Hox genes form a Hox cluster in the *X. bocki* genome, but
270 that there are possible unresolved assembly errors disrupting the current annotation.
271 We also identified other homeobox genes on the Hox cluster scaffold, including Evx
272 (Fig. 4a).

273 Along with the Hox genes, we also surveyed other homeobox genes that are
274 often clustered. The canonical bilaterian paraHox cluster contains three genes Cdx,
275 Xlox (=Pdx) and Gsx. We identified Cdx and a new Gsx annotation on the same
276 scaffold, as well as a previously reported Gsx paralog on a separate scaffold. This
277 indicates partial retention of the paraHox cluster in *X. bocki* along with a further
278 duplication of Gsx. On both of these paraHox containing scaffolds we observed other
279 homeobox genes.

280 Hemichordates and chordates have a conserved cluster of genes involved in
281 patterning their pharyngeal pores - the so-called 'pharyngeal cluster'. The homeobox
282 genes of this cluster (Msxlx, Nk2-1/2/4/8) were present on a single *X. bocki* scaffold.
283 Another pharyngeal cluster transcription factor, the Forkhead containing Foxa, and
284 'bystander' genes from that cluster including Egln, Mipol1 and Slc25a21 are found in
285 the same genomic region. Different sub-parts of the cluster are found in non-bilaterians
286 and protostomes and the cluster may well be plesiomorphic for the Bilateria rather
287 than a deuterostome synapomorphy[31].

288

289 <u>The *X. bocki* neuropeptide complement is larger than previously thought</u>

290 A catalogue of acoelomorph neuropeptides was previously described using
291 transcriptome data[32]. We have discovered 12 additional neuropeptide genes and 39
292 new neuropeptide receptors in *X.bocki* adding 6 bilaterian peptidergic systems to the
293 *Xenoturbella* catalogue (NPY-F ; MCH/Asta-C ; TRH ; ETH ; CCHa/Nmn-B ; Np-
294 S/CCAP), and 6 additional bilaterian systems to the Xenacoelomorpha catalogue
295 (Corazonin ; Kiss/GPR54 ; GPR83 ; 7B2 ; Trunk/PTTH ; NUCB2) making a total of 31
296 peptidergic systems (Fig. 4, Supplementary).

297 Among the ligand genes, we identify 6 new repeat-containing sequences. One
298 of these, the LRIGamide-peptide, had been identified in Nemertodermatida and

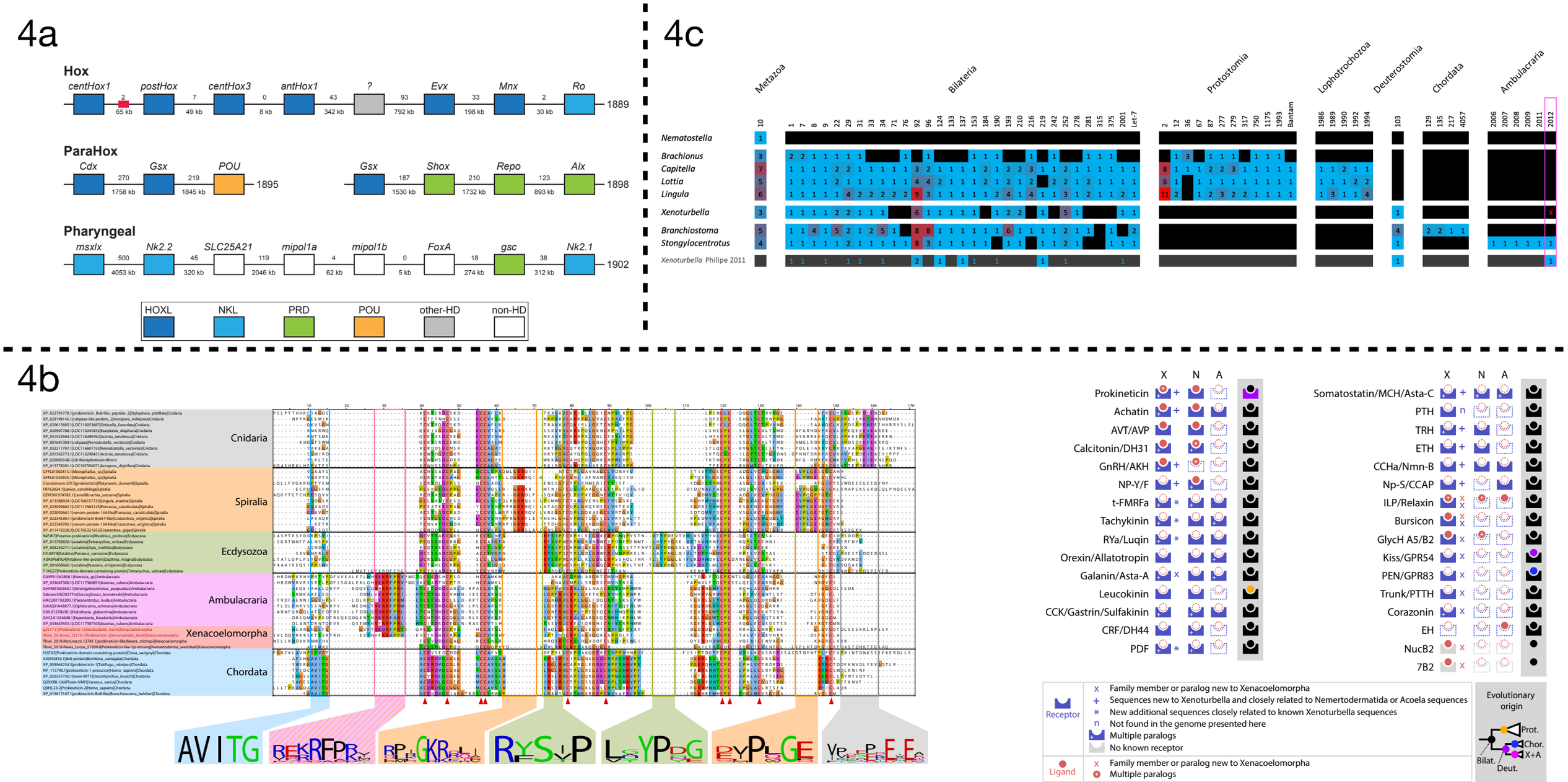


Figure 4: (a) *X. bocki* has 5 HOX genes, which are located in relatively close proximity on one of our chromosome size scaffolds. Similar clusters exist for the ParaHox and "pharyngeal" genes. Numbers between genes are distance (below) and number of genes between (below). Colours indicate gene families.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

299 Acoela and its loss in *Xenoturbella* was proposed[32]. We also identify the first 7B2
300 neuropeptide and NucB2/Nesfatin genes in Xenacoelomorpha. Finally, we identified 3
301 new *X.bocki* insulin-like peptides, one of them showing sequence similarity and an
302 atypical cysteine pattern shared with the Ambulacrarian octinsulin, constituting a
303 potential synapomorphy of Xenambulacraria (see Supplementary).
304 Our searches also revealed the presence of components of the arthropod
305 moulting pathway components (PTTH/trunk, NP-S/CCAP and Bursicon receptors),
306 which have recently been shown to be of ancient origin (de Oliveira et al., 2019). We
307 further identified multiple paralogs for, e.g the Tachykinin, Rya/Luquin, tFMRFa,
308 Corazonin, Achatin, CCK, and Prokineticin receptor families. Two complete *X. bocki*
309 Prokineticin ligands were also found in our survey (see Supplementary).
310 Chordate Prokineticin ligands possess a conserved N-terminal "AVIT" sequence
311 required for the receptor activation[33]. This sequence is absent in arthropod Astakine,
312 which instead possess two signature sequences within their Prokineticin domain [34].
313 To investigate Prokineticin ligands in Xenacoelomorpha we compared the sequences
314 of their prokineticin ligands with those of other bilaterians (Fig. 4b, Supplementary).
315 Our alignment reveals clade specific signatures already reported in Ecdysozoa and
316 Chordata sequences, but also two new signatures specific to Lophotrochozoa and
317 Cnidaria sequences, as well as a very specific "K/R-RFP-K/R" signature shared only
318 by ambulacrarian and *Xenoturbella bocki* sequences. The shared
319 Ambulacrarian/Xenacoelomorpha signature is found at the same position as the
320 Chordate sequence involved in receptor activation - adjacent to the N-terminus of the
321 Prokineticin domain (Fig. 4b).
322
323 **The *X. bocki* genome contains most bilaterian miRNAs reported missing from**
324 **acoels.**
325 microRNAs have previously been used to investigate the phylogenetic position of the
326 acoels and *Xenoturbella.* The acoel *Symsagittifera roscoffensis* lacks protostome and
327 bilaterian miRNAs and this lack was interpreted as supporting the position of acoels
328 as sister-group to the Nephrozoa. Based on shallow 454 microRNA sequencing (and
329 sparse genomic traces) of *Xenoturbella*, some of the bilaterian miRNAs missing from

330 acoels were found - 16 of the 32 expected metazoan (1 miRNA) and bilaterian (31
331 miRNAs) microRNA families – of which 6 could be identified in genome traces[6].
332 By deep sequencing two independent small RNA samples, we have now
333 identified the majority of the missing metazoan and bilaterian microRNAs and
334 identified them in the genome assembly (Fig. 4c). Altogether, we found 23 out of 31
335 bilaterian microRNA families (35 genes including duplicates); the single known
336 Metazoan microRNA family (MIR-10) in 2 copies; the Deuterostome-specific MIR-103;
337 and 7 *Xenoturbella*-specific microRNAs giving a total of 46 microRNA genes. None of
338 the protostome-specific miRNAs were found and we could not confirm in the RNA
339 sequences or new assembly a previously identified, and supposedly
340 xenambulacrarian-specific MIR-2012 ortholog.
341
342 **The *X. bocki* genome retains ancestral metazoan linkage groups.**
343 The availability of chromosome-scale genomes has made it possible to reconstruct 24
344 ancestral linkage units broadly preserved in bilaterians[35]. In fast-evolving genomes,
345 such as those of nematodes, tunicates or platyhelminths, these ancestral linkage
346 groups (ALGs) are often dispersed and/or extensively fused (Supplementary). We
347 were interested to test if the general conservation of the gene content in *X. bocki* is
348 reflected in its genome structure.
349 We compared the genome of *Xenoturbella* to several other metazoan genomes
350 and found that it has retained most of these ancestral bilaterian units: 12
351 chromosomes in the *X. bocki* genome derive from a single ALG, five chromosomes
352 are made of the fusion of two ALGs, and one *Xenoturbella* chromosome is a fusion of
353 three ALGs, as highlighted with the comparison of ortholog content with amphioxus,
354 the sea urchin and the sea scallop (Fig. 5 and Supplementary).
355 One ancestral linkage group that has been lost in chordates but not in
356 ambulacrarians (such as sea urchin, ALG R) is detectable in *X. bocki* (Fig. 5), while *X.*
357 *bocki* does not show the fusions that are characteristic of lophotrochozoans.
358 We also attempted to detect some pre-bilaterian arrangement of ancestral
359 linkage: for instance, ref [36] predicted that several pre-bilaterian linkage groups
360 successively fused in the bilaterian lineage to give ALGs A1, Q and E. These are all
361 represented as a single unit in *X. bocki* in common with other Bilateria and ultimately,
362 we found none of the inferred pre-bilaterian chromosomal arrangements in *X. bocki*

363 that could have provided support for the Nephrozoa hypothesis.

364

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

366 The smallest of the 18 large scaffolds in the *X. bocki* genome did not show strong 1:1
367 clustering with any scaffold/chromosome of the bilaterian species we compared it to.
368 To exclude potential outside contamination in the assembly as a source for this contig
369 we examined the orthogroups to which the genes from this scaffold belong. We found
370 that *Xenoturbella profunda*[37], for which a transcriptome is available, was the species
371 that most often occurred in the same orthogroup with genes from this scaffold (41
372 shared orthogroups), suggesting the scaffold is not a contaminant.

373 We did observe links between the aberrant scaffold and several scaffolds from
374 the genome of the sponge *E. muelleri* in regard to synteny (Supplementary). In line
375 with this, genes on the scaffold show a different age structure compared to other
376 scaffolds, with both more older genes (pre bilaterian) and more *Xenoturbella* specific
377 genes (Fig. 5b; supported by Ks statistics, Supplementary). The scaffold also had
378 significantly lower levels of methylation than the rest of the genome (Supplementary).

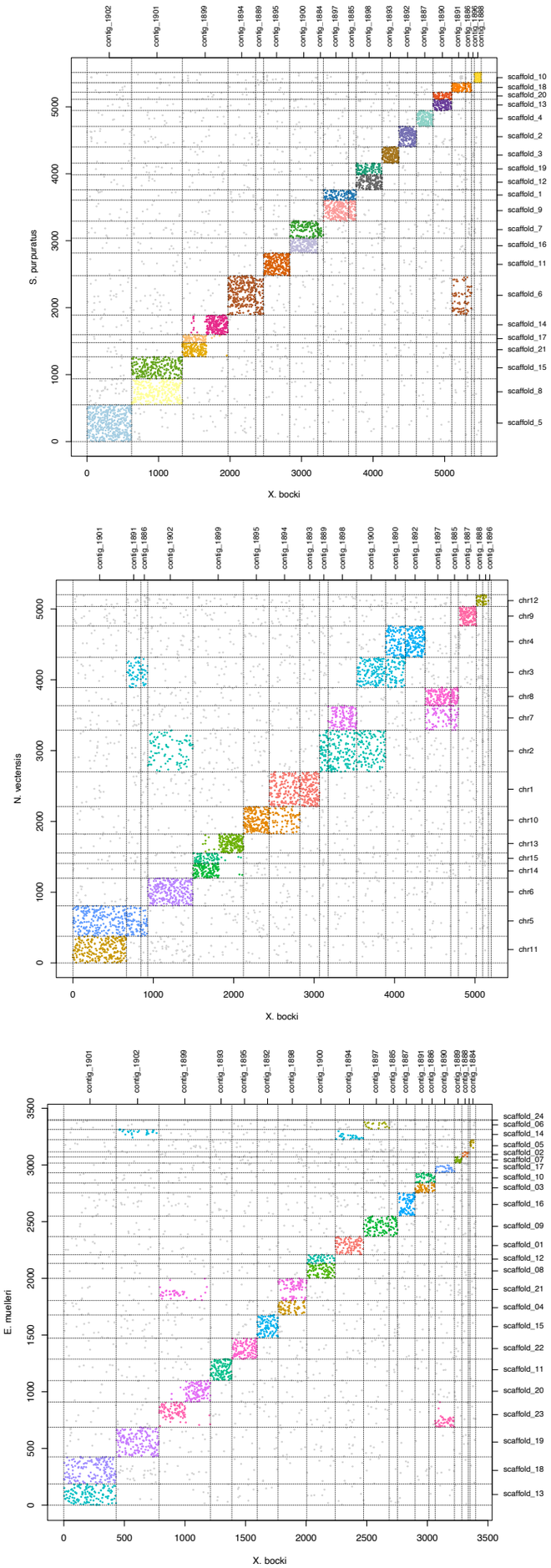
379

380 **DISCUSSION**

381 The phylogenetic positions of *Xenoturbella* and the Acoelomorpha have been
382 controversial since the first molecular data appeared over twenty five years ago. Today
383 we understand that they constitute a monophyletic group of morphologically simple
384 worms[1,6,38], but there remains a disagreement over whether they represent a
385 secondarily simplified sister group of the Ambulacraria or a primitively simple sister
386 group to all other Bilateria (and can be interpreted as evolutionary intermediates
387 between the eumetazoan and nephrozoan common ancestors).

388 Previous analyses of the content of genomes, especially of Acoela, have been
389 used to bolster the latter view, with the small number of Hox genes and of microRNAs
390 of acoels interpreted as representing an intermediate stage on the path to the ~8 Hox
391 genes and 30 odd microRNAs of the Nephrozoa. A strong version of the Nephrozoa
392 idea would go further than these examples and anticipate, for example, a genome
393 wide paucity of bilaterian genes, GRNs and biochemical pathways and/or an
394 arrangement of chromosomal segments intermediate between that of the Eumetazoa
395 and Nephrozoa.

5a



5b

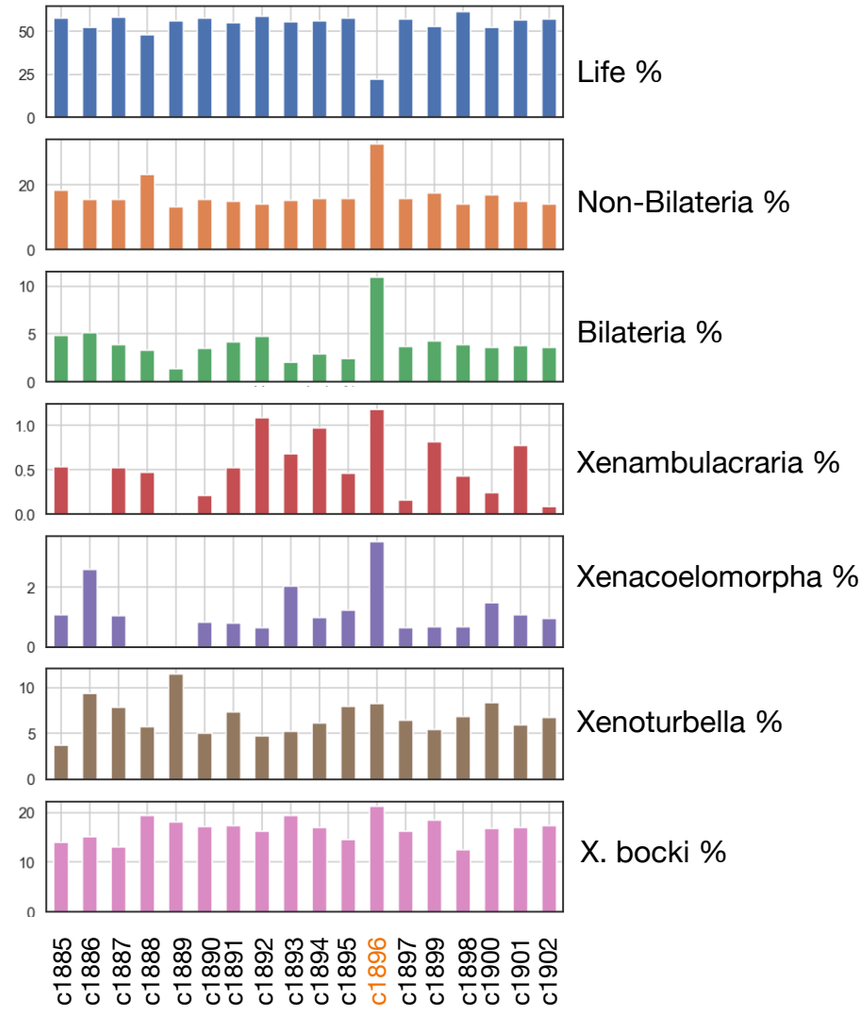


Figure 5: (a) A comparison of scaffolds in the *X. bocki* genome with other Metazoa. 17 of the 18 large scaffolds in the *X. bocki* genome are linked via synteny to distinct chromosomal scaffolds in these species. (b) Phylostratigraphic age distribution of genes on all major scaffolds in the *X. bocki* genome. One scaffold, 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.

396 One criticism of the results from analyses of acoel genomes is that the
397 Acoelomorpha have evolved rapidly (their long branches in phylogenetic trees
398 showing high rates of sequence change). This rapid evolution might plausibly be
399 expected to correlate with other aspects of rapid genome evolution such as higher
400 rates of gene loss and chromosomal rearrangements leading to significant differences
401 from other Bilateria. The much more normal rates of sequence evolution observed in
402 *Xenoturbella* therefore recommend it as a more appropriate xenacoelomorph to study
403 with fewer apomorphic characters expected.
404 We have sequenced, assembled, and analysed a draft genome of *Xenoturbella*
405 *bocki*. To help with annotation of the genome we have also sequenced miRNAs and
406 small RNAs as well as using bisulphite sequencing, Hi-C and Oxford nanopore. We
407 compared the gene content of the *Xenoturbella* genome to species across the
408 Metazoa and its genome structure to several other high-quality draft animal genomes.
409 We found the *X. bocki* genome to be fairly compact, but not unusually reduced
410 in size compared to many other bilaterians. It appears to contain a similar number of
411 genes (~15,000) as other animals, for example from the model organisms *D.*
412 *melanogaster* (>14,000) and *C. elegans (*~20,000). The BUSCO completeness, as
413 well as a high level of representation of *X. bocki* proteins in the orthogroups of our 155
414 species orthology screen (discussed in detail below) indicates that we have annotated
415 a near complete gene set. Surprisingly, there are fewer genes than in the acoel
416 *Hofstenia* (>22,000; BUSCO score ~95%). This said, of the genes found in Urbilateria
417 (orthogroups in our presence/absence analysis containing a member from both a
418 bilaterian and an outgroup) *Xenoturbella* and *Hofstenia* have very similar numbers
419 (5459 and 5438 respectively). Gene, intron and exon lengths also fall in a range seen
420 in many other invertebrate species[20]. It thus seems that basic genomic features in
421 *Xenoturbella* are not anomalous among Bilateria. Unlike some extremely simplified
422 animals, such as orthonectids, we observe no extreme reduction in gene content.
423 All classes of homeodomain transcription factors have previously been reported
424 to exist in Xenacoelomorpha[39]. We have identified 5 HOX-genes in *X. bocki* at least
425 four, and probably all five of these are on one chromosomal scaffold within 187 Kbp.
426 *X.bocki* also has the parahox genes Gsx and Cdx; while Xlox/pdx is not found, it is
427 present in Cnidarians and must therefore have been lost[40]. If block duplication models

428 of Hox and Parahox evolutionary relationships are correct, the presence of a complete
429 set of parahox genes implies the existence of their Hox paralogs in the ancestor of
430 Xenacoelomorphs suggesting the xenacoelomorph ancestor also possessed a Hox 3
431 ortholog. If anthozoans also have an ortholog of bilaterian Hox 2[41], this must also have
432 been lost from Xenacoelomorphs. The minimal number of Hox genes in the
433 xenacoelomorph stem lineage was therefore probably 7 (AntHox1, lost Hox2, lost Hox
434 3, CentHox 1, CentHox 2, CentHox 3 and postHoxP).

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

445 The last decade has seen a re-evaluation of our understanding of the evolution
446 of the neuropeptide signaling genes[42,43]. The peptidergic systems are thought to have
447 undergone a diversification that produced approximately 30 systems in the bilaterian
448 common ancestor[42,43]. Our study identified 31 neuropeptide systems in *X.bocki* and
449 for all of these either the ligand, receptor, or both are also present in both protostomes
450 and deuterostomes indicating conservation across Bilateria. It is likely that more
451 ligands (which are short and variable) remain to be found with enhanced detection
452 methods. It appears that the *Xenoturbella* genome contains a nearly complete
453 bilaterian neuropeptide complement with no signs of simplification but rather signs of
454 expansions of certain gene families and reveals a new potential synapomorphy with
455 Ambulacraria (Fig 4 and Supplementary).

456 We have used the predicted presence and absence of genes across a selection
457 of metazoan genomes as characters for phylogenetic analyses. Our trees re-confirm
458 the findings of recent phylogenomic gene alignment studies in linking *Xenoturbella* to
459 the Ambulacraria. We also used these data to test different bilaterians for their

460 propensity to lose otherwise conserved genes (or for our inability to identify
461 orthologs[26]). While the degree of gene loss appears similar between *Xenoturbella* and
462 acoels, the phylogenetic analysis shows longer branches leading to the acoels most
463 likely due to faster evolution and some degree of gene loss in the branch leading to
464 the Acoelomorpha.

465 This pattern of conservation of evolutionarily old parts of the Metazoan genome
466 is further reinforced by the retention in *Xenoturbella* of linkage groups present from
467 sponges to vertebrates. It is interesting to note that as a morphologically simplified
468 organism *X. bocki* would not follow the pattern of for example nematodes and
469 Platyhelminthes, which have lost or fused the ancestral units of conservation. We
470 interpret this to be a signal of comparably slower genomic evolution in *Xenoturbella* in
471 comparison to some derived protostome lineages. The fragmented genome sequence
472 of *Hofstenia* prevents us from asking whether the ancient linkage groups have also
473 been preserved in the Acoelomorpha.

474 One of the chromosome-scale scaffolds in our assembly showed a different
475 methylation and age signal, with more younger genes, and also no clear relationship
476 to metazoan linkage groups. By analyzing orthogroups of genes on this scaffold for
477 their phylogenetic signal and finding *X. bocki* genes to cluster with those of *X. profunda*
478 we concluded that the scaffold most likely does not represent a contamination. It
479 remains unclear whether this scaffold is a fast-evolving chromosome, or a
480 chromosomal fragment or arm. Very fast evolution on a chromosomal arm has for
481 example been shown in the zebrafish[44].

482 Apart from DNA from *X. bocki* we also obtained a highly contiguous genome of
483 a species related to marine *Chlamydia* species (known from microscopy to exist in *X.*
484 *bocki)*; a symbiotic relationship between the species and the bacteria has been
485 thought possible[45] The large gene number and the completeness of genetic pathways
486 we found in the chlamydial genome do not support an endosymbiotic relationship.

487 Overall, we have shown that, while *Xenoturbella* has lost some genes - in
488 addition to the reduced number of Hox genes previously noted, we observe a reduction
489 of some signaling pathways to the core components - in general, the *X. bocki* genome
490 is not strikingly simpler than many other bilaterian genomes. Some degree of
491 secondary absence is in fact expected as our analysis of the gene presence/absence

492 matrix shows Xenacoelomorpha have lost a greater number of genes we know to have
493 existed in Urbilateria than have most other bilaterians.
494 We do not find support for a strong version of the Nephrozoa hypothesis which
495 would predict many missing bilaterian genes. Bilaterian Hox and microRNA absent
496 from Aceolomorpha are found in *Xenoturbella* removing two character types that were
497 previously used in support of Nephrozoa. The *Xenoturbella* genome has also largely
498 retained the ancestral linkage groups found in other bilaterians and does not represent
499 a structure intermediate between Eumetazoan and bilaterian ground states. Overall,
500 while we can rule out a strong version of the Nephrozoa hypothesis, our analysis of
501 the *Xenoturbella* genome cannot distinguish between a weaker version of Nephrozoa
502 (with character absences being typically primary) and the Xenambulacraria topology
503 (with character absences being mostly secondary); nevertheless, our phylogenetic
504 analysis of gene presence and absence supports the latter.
505

506 # Methods

507 **Genome Sequencing, Assembly, and Scaffolding**

508 We extracted DNA from individual *Xenoturbella* specimens with Phenol-
509 Chloroform protocols and Qiagen kits. Worms were first starved and kept in repeatedly
510 replaced salt water, reducing the likelihood of food or other contaminants in the DNA
511 extractions. Initially, we sequenced Illumina short paired-end reads and mate pair
512 libraries. As the initial paired-read datasets were of low complexity and coverage we
513 later complement this data with an Illumina HiSeq 4000 paired-end dataset with ~700
514 bp insert size and 250bp read lengths, yielding ~354M reads. Additionally, we
515 generated ~40M Illumina TruSeq Synthetic Long Reads (TSLR) for high confidence
516 primary scaffolding.
517 After read cleaning with Trimmomatic v.0.38[46] we conducted initial test assemblies
518 using the clc assembly cell v.5 and ran the blobtools pipeline[47] to screen for
519 contamination. Not detecting any significant numbers of reads from suspicious
520 sources in the HiSeq 4000 dataset we used SPAdes v. 3.9.0[16] to correct and assemble
521 a first draft genome. We also tried to use dipSPAdes but found the runtime to exceed
522 several weeks without finishing. Thus, we submitted SPAdes assembly to the
523 redundans pipeline to eliminate duplicate contigs and to scaffold with all available mate

524 pair libraries. The resulting assembly was then further scaffolded with the aid of
525 assembled transcripts (see below) in the BADGER pipeline[48]. In this way we were able
526 to obtain a draft genome with ~60kb N50 that could be scaffolded to chromosome
527 scale super-scaffolds with the use of 3C data.
528
529 Preparation of the Hi-C libraries
530 The Hi-C protocol was adapted from refs[49] and[50]. Briefly, an animal was chemically
531 cross-linked for one hour at room temperature. Formaldehyde was quenched for 20
532 min at RT by adding 10 ml of 2.5 M glycine. The fixed animal was recovered through
533 centrifugation and stored at -80°C until use. For library preparation, the animal was
534 transferred to a VK05 Precellys tubes in 1X DpnII buffer (NEB) and the tissues were
535 disrupted using the Precellys Evolution homogenizer (Bertin-Instrument). SDS was
536 added to the lysate and the tubes were incubated at 65°C for 20 minutes followed by
537 an incubation at 37°C for 30 minutes and an incubation of 30 minutes after adding 50
538 μL of 20% triton-X100. 150 units of the DpnII restriction enzyme were then added and
539 the tubes were incubated overnight at 37°C. The tubes were then centrifuged and
540 pellets were re-suspended in 200 μl NE2 1X buffer and pooled. DNA ends were
541 labeled using 50 μl NE2 10X buffer, 37.5 μl 0.4 mM dCTP-14-biotin, 4.5 μl 10mM
542 dATP-dGTP-dTTP mix, 10 μl klenow 5 U/μL and incubation at 37°C for 45 minutes.
543 The labeling mix was then transferred to ligation reaction tubes (1.6 ml ligation buffer;
544 160 μl ATP 100 mM; 160 μl BSA 10 mg/mL; 50 μl ligase 5U/μl; 13.8 ml H2O) and
545 incubated at 16°C for 4 hours. A proteinase K mix was added to each tube and
546 incubated overnight at 65°C. DNA was then extracted, purified and processed for
547 sequencing. Hi-C libraries were sequenced on a NextSeq 500 (2 × 75 bp, paired-end
548 using custom made oligonucleotides[50]). Libraries were prepared separately on two
549 individuals in this way but eventually merged.
550
551 InstaGRAAL assembly pre-processing
552 The primary Illumina assembly contains a number of very short contigs, which are
553 disruptive when computing the contact distribution needed for the instaGRAAL
554 proximity ligation scaffolding (see[50] and[17] for details). Testing several Nx metrics we
555 found a relative length threshold, that depends on the scaffolds' length distribution, to

556 be a good compromise between the need for a low-noise contact distribution and the
557 aim of connecting most of the genome. We found N90 a suitable threshold and
558 excluded contigs below 1,308 bp. This also ensured no scaffolds shorter than three
559 times the average length of a DpnII restriction fragment (RF) were in the assembly. In
560 this way every contig contained enough RFs for binned and were included in the
561 scaffolding step.

562 Reads from both libraries were aligned with bowtie2 (v. 2.2.5)[51] against the
563 DpnII RFs of the reference assembly using the hicstuff pipeline
564 (https://github.com/koszullab/hicstuff) and in paired-end mode (with the options: -fg-
565 maxins 5 -fg-very-sensitive-local), with a mapping quality >30. The pre-processed
566 genome was reassembled using instaGRAAL. Briefly, the program uses a Markov
567 Chain Monte Carlo (MCMC) method that samples DNA segments (or bins) of the
568 assembly for their best relative 1D positions with respect to each other. The quality of
569 the positions is assessed by fitting the contact data first on a simple polymer model,
570 then on the plot of contact frequency according to genomic distance law computed
571 from the data. The best relative position of a DNA segment with respect to one of its
572 most likely neighbors consists in operations such as flips, swaps, merges or split of
573 contigs. Each operation is either accepted or rejected based on the computed
574 likelihood, resulting in an iterative progression toward the 1D structure that best fits
575 the contact data. Once the entire set of DNA segments is sampled for position (i.e. a
576 cycle), the process starts over. The scaffolder was run independently for 50 cycles,
577 long enough for the chromosome structure to converge. The corresponding genome
578 is then considered stable and suitable for further analyses. The scaffolded assemblies
579 were then refined using instaGRAAL's instaPolish module, to correct small artefactual
580 inversions that are sometimes a byproduct of instaGRAAL's processing.

581

582 **Genome Annotation**

583 Transcriptome Sequencing

584 We extracted total RNA from a single *X. bocki* individual and sequenced a strand
585 specific Illumina paired end library. The resulting transcriptomic reads were assembled
586 with the Trinity pipeline[52,53] for initial control and then supplied to the genome
587 annotation pipeline (below).

588 Repeat annotation

589 In the absence of a repeat library for Xenoturbellida we first used RepeatModeller
590 v. 1.73 to establish a library *de novo*. We then used RepeatMasker v. 4.1.0
591 (https://www.repeatmasker.org) and the Dfam library[54,55] to soft-mask the genome.
592 We mapped the repeats to the instaGRAAL scaffolded genome with RepeatMasker.
593

594 Gene prediction and annotation

595 We predicted genes using Augustus[56] implemented into the BRAKER (v.2.1.0)
596 pipeline[18,19] to incorporate the RNA-Seq data. BRAKER uses spliced aligned RNA-
597 Seq reads to improve training accuracy of the gene finder GeneMark-ET[57].
598 Subsequently, a highly reliable gene set predicted by GeneMark-ET in *ab initio* mode
599 was selected to train the gene finder AUGUSTUS, which in a final step predicted
600 genes with evidence from spliced aligned RNA-Seq reads. To make use of additional
601 single cell transcriptome data allowing for a more precise prediction of 3'-UTRs we
602 employed a production version of BRAKER (August 2018 snapshot). We had
603 previously mapped the RNA-Seq data to the genome with gmap-gsnap v. 2018-07-
604 04[58] and used samtools[59] and bamtools[60] to create the necessary input files. This
605 process was repeated in an iterative way, visually validating gene structures and
606 comparing with mappings loci inferred from the single cell data, in particular in regard
607 to fused genes. Completeness of the gene predictions was independently assessed
608 with BUSCO[22] setting the eukaryote dataset as reference on gVolante[61]. We used
609 InterProScan v. 5.27-66.0 standalone[62,63] on the UCL cluster to annotate the predicted
610 *X. bocki* proteins with Pfam, SUPERFAM, PANTHER, and Gene3D information.
611

612 Horizontal Gene Transfer

613 To detect horizontally acquired genes in the *X. bocki* genome we made use of a
614 pipeline available from (https://github.com/reubwn/hgt). Briefly, this uses blast against
615 the NCBI database, alignments with MAFFT[64], and phylogenetic inferences with
616 IQTree[65,66] to infer most likely horizontally acquired genes, while trying to discard
617 contamination (e.g. from co-sequenced gut microbiota).
618
619

620 Orthology inference

621 We included 155 metazoan species and outgroups into our orthology analysis. We
622 either downloaded available proteomes or sourced RNA-Seq reads from online
623 repositories to then use Trinity v 2.8.5 and Trinnotate v. 3.2.0 to predict protein sets.
624 In the latter case we implemented diamond v. 2.0.0 blast[67,68] searches against UniProt
625 and Pfam[69] hmm screens against the Pfam-A dataset into the prediction process. We
626 had initially acquired 185 datasets, but excluded some based on inferior BUSCO
627 completeness, while at the same time aimed to span as many phyla as possible.
628 Orthology was then inferred using Orthofinder v. 2.2.7[70,71], again with diamond as the
629 blast engine.

630 Using InterProScan v. 5.27-66.0 standalone on all proteomes we added
631 functional annotation and then employed kinfin[29] to summarise and analyse the
632 orthology tables. For the kinfin analysis, we tested different query systems in regard
633 to phylogenetic groupings (Supplementary).

634 To screen for inflation and contraction of gene families we first employed
635 CAFE5[72], but found the analysis to suffer from long branches and sparse taxon
636 sampling in Xenambulacraria. We thus chose to query individual gene families (e.g.
637 transcription factors) by looking up Pfam annotations in the InterProScan tables of
638 high-quality genomes in our analysis.

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

641

642 Presence/absence phylogenetics

643 We used a database of metazoan proteins, updated from ref [73], as the basis for
644 an OMA analysis to calculate orthologous groups, performing two separate runs, one
645 including *Xenoturbella* and acoels, and one with only *Xenoturbella*. We converted
646 OMA gene OrthologousMatrix.txt files into binary gene presence absence matrices in
647 Nexus format with datatype = restriction. We calculated phylogenetic trees on these
648 matrices using RevBayes (see https://github.com/willpett/metazoa-gene-content for
649 RevBayes script), as described in ref [74], with corrections for no absent sites and no
650 singleton presence. For each matrix, two runs were performed and compared and
651 consensus trees generated with bpcomp from Phylobayes[75].

652 *Hox and ParaHox gene cluster identification and characterisation*

653 Previous work has already used transcriptomic data and phylogenetic inference
654 to identify the homeobox repertoire in *Xenoturbella bocki*. These annotations were
655 used to identify genomic positions and gene annotations that correspond to Hox and
656 ParaHox clusters in *X. bocki*. Protein sequences of homeodomains (Evx, Cdx, Gsx,
657 antHox1, centHox1, centHox2, cent3 and postHoxP) were used as TBLASTN queries
658 to initially identify putative scaffolds associated with Hox and ParaHox clusters. Gene
659 models from these scaffolds were compared to the full length annotated homeobox
660 transcripts from[76] using BLASTP, using hits over 95% identity for homeobox
661 classification. There were some possible homeodomain containing genes on the
662 scaffolds that were not previously characterised and where therefore not given an
663 annotation.

664 There were issues concerning the assignment of postHoxP and Evx to gene
665 models. To ascertain possible CDS regions for these genes, RNA-Seq reads were
666 mapped with HISAT2 to the scaffold and to previous annotation[76],were assembled
667 with Trinity and these were combined with BRAKER annotations.

668 Some issues were also observed with homeodomain queries matching genomic
669 sequences that were identical, suggesting artifactual duplications. To investigate
670 contiguity around genes the ONT reads were aligned with Minimap2 to capture long
671 reads over regions and coverage.

672

673 Small RNA Sequencing and Analysis

674 Two samples of starved worms were subjected to 5' monophosphate dependent
675 sequencing of RNAs between 15 and 36 nucleotides in length, according to previously
676 described methods[77]. Using miRTrace[78] 3.3 18.6 million high-quality reads were
677 extracted and merged with the 27 635 high quality 454 sequencing reads from Philippe
678 et al. The genome sequence was screened for conserved miRNA precursors using
679 MirMachine (Umu et al in prep; https://github.com/sinanugur/MirMachine) followed by
680 a MirMiner run that used predicted precursors and processed and merged reads on
681 the genome[79]. Outputs of MirMachine and MirMiner were manually curated using a
682 uniform system for the annotation of miRNA genes[80] and by comparing to
683 MirGeneDB[81].

684 Neuropeptide prediction and screen

685 Neuropeptide prediction was conducted on the full set of *X.bocki* predicted proteins
686 using two strategies to detect neuropeptide sequence signatures. First, using a
687 custom script detecting the occurrence of repeated sequence patterns:
688 RRx(3,36)RRx(3,36)RRx(3,36)RR,RRx(2,35)ZRRx(2,35)ZRR,
689 RRx(2,35)GRRx(2,35)GRR, RRx(1,34)ZGRRx(1,34)ZGRR where R=K or R ; x=any
690 amino acid ; Z=any amino acid but repeated within the pattern. Second, using
691 HMMER3.1[82] (hmmer.org), and a combination of neuropeptide HMM models obtained
692 from the PFAM database (pfam.xfam.org) as well as a set of custom HMM models
693 derived from alignment of curated sets of neuropeptide sequences[42,43,83]. Sequences
694 retrieved using both methods and comprising fewer than 600 amino acids were further
695 validated. First, by blast analysis: sequences with E-Value ratio "best blast hit versus
696 ncbi nr database/best blast hit versus curated neuropeptide dataset" < 1e-40 were
697 discarded. Second by reciprocal best blast hit clustering using Clans[84]
698 (eb.tuebingen.mpg.de/protein-evolution/software/clans/) with a set of curated
699 neuropeptide sequences[42]. SignalP-5.0[85] (cbs.dtu.dk/services/SignalP/) was used to
700 detect the presence of a signal peptide in the curated list of predicted neuropeptide
701 sequences while Neuropred[86] (stagbeetle.animal.uiuc.edu/cgi-bin/neuropred.py) was
702 used to detect cleavage sites and post-translational modifications. Sequence
703 homology of the predicted sequence with known groups was analysed using a
704 combination of (i) blast sequence similarity with known bilaterian neuropeptide
705 sequences, (ii) reciprocal best blast hit clustering using Clans and sets of curated
706 neuropeptide sequences, (iii) phylogeny using MAFFT
707 (mafft.cbrc.jp/alignment/server/), TrimAl[87] (trimal.cgenomics.org/) and IQ-TREE[88]
708 webserver for alignment, trimming and phylogeny inference respectively. Bilaterian
709 prokineticin-like sequences were searched in ncbi nucleotide, EST and SRA
710 databases as well as in the *Saccoglossus kowalevskii* genome assembly[66,89]
711 (groups.oist.jp/molgenu) using various bilaterian prokineticin-related protein
712 sequences as query. Sequences used for alignments shown in figures were collected
713 from ncbi nucleotide and protein databases as well as from the following publications:
714 7B2[42]; NucB2[83]; Insulin[90]; Prokineticin[33,34,91]. Alignments for figures were created with
715 Jalview (jalview.org).

716

717 Neuropeptide receptor search

718 Neuropeptide Receptor sequences for Rhodopsin type GPCR, Secretin type GPCR
719 and tyrosine and serine/threonine kinase receptors were searched by running
720 HMMER3.1 on the full set of *X.bocki* predicted proteins using the 7tm_1 (PF00001),
721 7tm_2 (PF00002) and PK_Tyr_Ser-Thr (PF07714) HMM models respectively which
722 were obtained from the PFAM database (pfam.xfam.org). Sequences above
723 significance threshold were then aligned with sequences from curated dataset,
724 trimmed and phylogeny inference was conducted using same method as for the
725 neuropeptide. A second alignment and phylogeny inference was conducted after
726 removal of all *X.bocki* sequences having no statistical support for grouping with any of
727 the known neuropeptide receptor from curated dataset. Curated datasets were
728 collected from the following publications: Rhodopsin type GPCR beta and gamma and
729 Secretin type GPCR[91]; Rhodopsin type GPCR delta (Leucine-rich repeat-containing
730 G-protein coupled Receptors)[92]; Tyrosine kinase receptors[93,94]; and were
731 complemented with sequences from NCBI protein database.

732

733 Synteny

734 Ancestral linkage analyses rely on mutual-best-hits computed using Mmseqs2[95]
735 between pairs of species in which chromosomal assignment to ancestral linkage
736 groups (ALG) was previously performed, such as *Branchiostoma floridae* or *Pecten*
737 *maximus*[35]. Oxford dotplots were computed by plotting reciprocal positions of indexed
738 pairwise orthologs between two species as performed previously[35,36]. The significance
739 of ortholog enrichment in pairs of chromosomes was assessed using a fisher test.
740 We also used a Python implementation of MCscanX[96] (Haibao Tang and available on
741 https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)) to compare *X. bocki*
742 to *Euphydtia muelleri*, *Trichoplax adhearens*, *Branchiostoma floridae*, *Saccoglossus*
743 *kowalevskii*, *Ciona intestinalis*, *Nematostella vectensis*, *Asteria rubens, Pecten*
744 *maximus*, *Nemopilema nomurai,* *Carcinoscorpius rotundicauda* (see Supplementary).
745 Briefly the pipeline uses high quality genomes and their annotations to infer syntenic
746 blocks based on proximity. For this an all vs. all blastp is performed and synteny
747 extended from anchors identified in this way. Corresponding heatmaps (see

748 Supplementary) were plotted with Python in a Jupyter notebooks instance.

749

750 *Chlamydia* assembly and annotation

751 We extracted a highly contiguous *Chlamydia* genome from the *X. bocki* genome
752 assembly. We then use our Oxford Nanopore derived long-reads to polish the
753 *Chlamyida* genome with LINKS[97] and annotated it with the automated PROKKA
754 pipeline. To place the genome on the *Chlamydia* tree we extracted the 16S gene
755 sequence, aligned it with set of *Chlamydia* 16S sequences from[23] using MAFFT, and
756 reconstructed the phylogeny using IQ-TREE 2[65] We visualized the resulting tree with
757 Figree (http://tree.bio.ed.ac.uk/).

758

759

760 **Acknowledgements**
761 We thank Josh Quick and Nick Loman for help with the generation of ONT long-read
762 data. Analyses were conducted mainly on the UCL Cluster, with some computations
763 also run on the CHEOPS cluster at the University of Cologne. We are grateful to Kevin
764 J. Peterson for his comments on the manuscript, the miRNA section in particular. We
765 thank the Kristineberg research station to support us with sampling worms.

766

767 **Conflict of interest**
768 The authors declare no conflict of interest.

769

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

776

777 **Funding**
778 PHS was funded by an ERC grant (ERC-2012-AdG 322790) to MJT, which also
779 supported HR, ACZ, SM. PHS was also funded through an Emmy-Noether grant
780 (434028868) to himself. Part of this work was funded by BBSRC grant
781 BB/R016240/1 (M.J.T./P.K.), by a Leverhulme Trust Research Project Grant RPG-
782 2018-302 (M.J.T./D.J.L.), and by the European Union's Horizon 2020 research and
783 innovation program under the Marie Skłodowska-Curie grant agreement no 764840
784 IGNITE (M.J.T./P.N.).

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

825 18. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-Genome Annotation
826 with BRAKER. *Methods Mol Biology Clifton N J* 1962, 65–95 (2019).

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

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

832 21. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinform Oxf Engl* 30, 2068–
833 9 (2014).

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

837 23. Dharamshi, J. E. *et al.* Marine Sediments Illuminate Chlamydiae Diversity and Evolution.
838 *Curr Biol* 30, 1032-1048.e7 (2020).

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

841 25. Kjeldsen, K. U., Obst, M., Nakano, H., Funch, P. & Schramm, A. Two Types of
842 Endosymbiotic Bacteria in the Enigmatic Marine Worm *Xenoturbella bocki*. 76, 2657–2662
843 (2010).

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

846 27. Schiffer, P. H., Robertson, H. E. & Telford, M. J. Orthonectids Are Highly Degenerate
847 Annelid Worms. *Current Biology* 1–9 (2018).

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

850 29. Laetsch, D. R., Laetsch, D. R., Blaxter, M. L. & Blaxter, M. L. KinFin: Software for
851 Taxon-Aware Analysis of Clustered Protein Sequences. *G3 (Bethesda, Md.)* 7, 3349–3357
852 (2017).

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

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

857 32. Thiel, D., Franz-Wachtel, M., Aguilera, F. & Hejnol, A. Xenacoelomorph
858 Neuropeptidomes Reveal a Major Expansion of Neuropeptide Systems during Early
859 Bilaterian Evolution. *Molecular Biology and Evolution* 35, 2528–2543 (2018).

860 33. Negri, L. & Ferrara, N. The Prokineticins: Neuromodulators and Mediators of
861 Inflammation and Myeloid Cell-Dependent Angiogenesis. *Physiol Rev* 98, 1055–1082
862 (2018).

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

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

867 36. Simakov, O. *et al.* Deeply conserved synteny and the evolution of metazoan
868 chromosomes. *Sci Adv* 8, eabi5884 (2022).

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

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

872 39. Brauchle, M. *et al.* Xenacoelomorpha Survey Reveals That All 11 Animal Homeobox
873 Gene Classes Were Present in the First Bilaterians. *Genome Biol Evol* 10, 2205–2217 (2018).

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

876 41. Ryan, J. F. *et al.* The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes:
877 evidence from the starlet sea anemone, *Nematostella vectensis*. *Genome Biol* 7, R64–R64
878 (2006).

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

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

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

885 45. Pillonel, T., Bertelli, C. & Greub, G. Environmental Metagenomic Assemblies Reveal
886 Seven New Highly Divergent Chlamydial Lineages and Hallmarks of a Conserved
887 Intracellular Lifestyle. *Front Microbiol* 9, 79 (2018).

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

890 47. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies.
891 *F1000research* 6, 1287 (2017).

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

894 49. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals
895 folding principles of the human genome. *Sci New York N Y* 326, 289–93 (2009).

896 50. Marie-Nelly, H. *et al.* High-quality genome (re)assembly using chromosomal contact
897 data. *Nat Comm* 5, 5695 (2014).

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

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

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

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

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

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

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

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

915 59. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25,
916 2078–2079 (2009).

917 60. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T.
918 BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* 27,
919 1691–1692 (2011).

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

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

924 63. Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence
925 classification and comparison. *Methods in molecular biology (Clifton, N.J.)* 396, 59–70
926 (2007).

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

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

931 66. Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. IQ-TREE: A Fast and
932 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol*
933 *Evol* 32, 268–274 (2015).

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

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

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

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

942 71. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
943 comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16, E9-
944 13 (2015).

945 72. Han, M. V., Thomas, G. W. C., Lugo-Martinez, J. & Hahn, M. W. Estimating Gene Gain
946 and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE
947 3. *Mol Biol Evol* 30, 1987–1997 (2013).

948 73. Leclère, L. *et al.* The genome of the jellyfish Clytia hemisphaericaand the evolution of
949 the cnidarian life-cycle. 1–41 (2018).

950 74. The Role of Homology and Orthology in the Phylogenomic Analysis of Metazoan Gene
951 Content. 1–7 (2019).

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

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

957 77. Sarkies, P. *et al.* Ancient and Novel Small RNA Pathways Compensate for the Loss of
958 piRNAs in Multiple Independent Nematode Lineages. *Plos Biol* 13, e1002061 (2015).

959 78. Kang, W. *et al.* miRTrace reveals the organismal origins of microRNA sequencing data.
960 *Genome Biol* 19, 213 (2018).

961 79. Wheeler, B. M. *et al.* The deep evolution of metazoan microRNAs. *Evol Dev* 11, 50–68
962 (2009).

963 80. Fromm, B. *et al.* A Uniform System for the Annotation of Vertebrate microRNA Genes
964 and the Evolution of the Human microRNAome. *Annu Rev Genet* 49, 213–242 (2015).

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

967 82. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and
968 iterative HMM search procedure. *BMC Bioinformatics* 11, 431–431 (2010).

969 83. Zandawala, M. *et al.* Discovery of novel representatives of bilaterian neuropeptide
970 families and reconstruction of neuropeptide precursor evolution in ophiuroid echinoderms.
971 *Open Biol* 7, 170129 (2017).

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

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

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

978 87. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldon, T. trimAl: a tool for automated
979 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973
980 (2009).

981 88. Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. IQ-TREE: A Fast and
982 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol*
983 *Evol* 32, 268–274 (2015).

984 89. Simakov, O. *et al.* Hemichordate genomes and deuterostome origins. *Nature* 527, 459–
985 465 (2015).

986 90. Cherif--Feildel, M., Berthelin, C. H., Rivière, G., Favrel, P. & Kellner, K. Data for
987 evolutive analysis of insulin related peptides in bilaterian species. *Data Brief* 22, 546–550
988 (2019).

989 91. Thiel, D., Franz-Wachtel, M., Aguilera, F. & Hejnol, A. Changes in the neuropeptide
990 complement correlate with nervous system architectures in xenacoelomorphs. 1–57 (2018).

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

993 93. Oliveira, A. L. de, Calcino, A. & Wanninger, A. Ancient origins of arthropod moulting
994 pathway components. *eLife* 8, e46113 (2019).

995 94. Smýkal, V. *et al.* Complex Evolution of Insect Insulin Receptors and Homologous Decoy
996 Receptors, and Functional Significance of Their Multiplicity. *Mol Biol Evol* 37, 1775–1789
997 (2020).

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

1000 96. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene
1001 synteny and collinearity. *Nucleic Acids Research* 40, e49–e49 (2012).

1002 97. Warren, R. L. *et al.* LINKS: Scalable, alignment-free scaffolding of draft genomes with
1003 long reads. *Gigascience* 4, 35 (2015).