

1 **Title**

2 Telomere-to-telomere assembly of the genome of an individual *Oikopleura*
3 *dioica* from Okinawa using Nanopore-based sequencing

4 **Authors**

5 Aleksandra Bliznina^{1*}, Aki Masunaga¹, Michael J. Mansfield¹, Yongkai Tan¹,
6 Andrew W. Liu¹, Charlotte West^{1,2}, Tanmay Rustagi¹, Hsiao-Chiao Chien¹, Saurabh
7 Kumar¹, Julien Pichon¹, Charles Plessy^{1*}, Nicholas M. Luscombe^{1,2,3}

8 ¹Okinawa Institute of Science and Technology Graduate University, Genomics and
9 Regulatory Systems Unit

10 ²Francis Crick Institute

11 ³UCL Genetics Institute, Department of Genetics, Evolution and Environment, University
12 College London

13 *Correspondence: aleksandra.bliznina2@oist.jp; charles.plessy@oist.jp

14

15 **Abstract**

16 **Background**

17 The larvacean *Oikopleura dioica* is an abundant tunicate plankton with the smallest (65-
18 70 Mbp) non-parasitic, non-extremophile animal genome identified to date. Currently,
19 there are two genomes available for the Bergen (OdB3) and Osaka (OSKA2016) *O. dioica*
20 laboratory strains. Both assemblies have full genome coverage and high sequence
21 accuracy. However, a chromosome-scale assembly has not yet been achieved.

22 **Results**

23 Here, we present a chromosome-scale genome assembly (OKI2018_I69) of the
24 Okinawan *O. dioica* produced using long-read Nanopore and short-read Illumina
25 sequencing data from a single male, combined with Hi-C chromosomal conformation
26 capture data for scaffolding. The OKI2018_I69 assembly has a total length of 64.3 Mbp
27 distributed among 19 scaffolds. 99% of the assembly is in five megabase-scale scaffolds.
28 We found telomeres on both ends of the two largest scaffolds, which represent
29 assemblies of two fully contiguous autosomal chromosomes. Each of the other three large
30 scaffolds have telomeres at one end only and we propose that they correspond to sex
31 chromosomes split into a pseudo-autosomal region and X-specific or Y-specific regions.
32 Indeed, these five scaffolds mostly correspond to equivalent linkage groups of OdB3,
33 suggesting overall agreement in chromosomal organization between the two populations.
34 At a more detailed level, the OKI2018_I69 assembly possesses similar genomic features
35 in gene content and repetitive elements reported for OdB3. The Hi-C map suggests few

36 reciprocal interactions between chromosome arms. At the sequence level, multiple
37 genomic features such as GC content and repetitive elements are distributed differently
38 along the short and long arms of the same chromosome.

39 **Conclusions**

40 We show that a hybrid approach of integrating multiple sequencing technologies with
41 chromosome conformation information results in an accurate *de novo* chromosome-scale
42 assembly of *O. dioica*'s highly polymorphic genome. This assembly will be a useful
43 resource for genome-wide comparative studies between *O. dioica* and other species, as
44 well as studies of chromosomal evolution in this lineage.

45

46 **Keywords**

47 *Oikopleura dioica*, Oxford Nanopore sequencing, telomere-to-telomere, chromosome-
48 scale assembly, single individual, Hi-C

49

50 **Background**

51 Larvaceans (synonym: appendicularians) are among the most abundant and ubiquitous
52 taxonomic groups within animal plankton communities (Allard, 1976; Hopcroft and
53 Roff, 1995). They live inside self-built "houses" which are used to trap food particles
54 (Sato *et al.*, 2001). The animals regularly replace houses as filters become damaged or
55 clogged and a proportion of discarded houses with trapped materials eventually sink to
56 the ocean floor. As such larvaceans play a significant role in global vertical carbon flux
57 (Allard, 2005).

58 *Oikopleura dioica* is the best documented species among larvaceans. It possesses
59 several invaluable features as an experimental model organism. It is abundant in coastal
60 waters and can be easily collected from the shore. Multigenerational culturing is possible
61 (Masunaga *et al.*, 2020). It has a short lifecycle of 4 days at 23°C and remains free-
62 swimming throughout its life (Feanux, 1998). As a member of the tunicates, a sister
63 taxonomic group to vertebrates, *O. dioica* offers insights into their evolution (Delsuc *et*
64 *al.*, 2006).

65 *O. dioica*'s genome size is 65-70 Mb (Seo *et al.*, 2001; Denoeud *et al.*, 2010), which is
66 one of the smallest among all sequenced animals. Interestingly, genome sequencing of
67 other larvacean species uncovered larger genome sizes, which correlated with the
68 expansion of repeat families (Naville *et al.*, 2019). *O. dioica* is distinguished from other
69 tunicates as it is the only reported dioecious species (Fredriksson and Olsson, 1991) and
70 its sex determination system uses an X/Y pair of chromosomes (Denoeud *et al.*, 2010).
71 The first published genome assembly of *O. dioica* (OdB3, B stands for Bergen) was
72 performed with Sanger sequencing which allowed for high sequence accuracy but limited

73 coverage (Denoeud *et al.*, 2010). The OdB3 assembly was scaffolded with a physical
74 map produced from BAC end sequences, which revealed two autosomal linkage groups
75 and a sex chromosome with a long pseudo-autosomal region (PAR; Denoeud *et*
76 *al.*, 2010). Recently, a genome assembly for a mainland Japanese population of *O. dioica*
77 (OSKA2016, OSKA denotes Osaka) was published, which displayed a high level of
78 coding sequence divergence compared with the OdB3 reference (Wang *et al.*, 2015;
79 Wang *et al.*, 2020). Although OSKA2016 was sequenced with single-molecule long reads
80 produced with the PacBio RSII technology, it does not have chromosomal resolution.

81 Historical attempts at karyotyping *O. dioica* by traditional histochemical stains arrived at
82 different chromosome counts, ranging between $n = 3$ (Körner, 1952) and $n = 8$
83 (Colombera and Fernaux, 1973). In preparation for this study, we karyotyped the
84 Okinawan *O. dioica* by staining centromeres with antibodies targeting phosphorylated
85 histone H3 serine 28 (Liu *et al.*, 2020), and concluded a count of $n = 3$. This is also in
86 agreement with the physical map of OdB3 (Denoeud *et al.*, 2010).

87 Currently, the method of choice for producing chromosome-scale sequences is to
88 assemble contigs using long reads (~10 kb or more) produced by either the Oxford
89 Nanopore or PacBio platforms, and to scaffold them using Hi-C contact maps (Lieberman-
90 Aiden *et al.*, 2009; Dudchenko *et al.*, 2017). To date, there have been no studies of
91 chromosome contacts in *Oikopleura* or any other larvaceans.

92 Here, we present a chromosome-length assembly of the Okinawan *O. dioica* genome
93 sequence generated with datasets stemming from multiple genomic technologies and
94 data types, namely long-read sequencing data from Oxford Nanopore, short-read
95 sequences from Illumina and Hi-C chromosomal contact maps (Fig. 1).

96

97 **Results**

98 **Genome sequencing and assembly**

99 *O. dioica*'s genome is small and highly polymorphic (Denoeud *et al.*, 2010), making
100 assembly of its complete sequence challenging. To reduce the level of variation, we
101 sequenced genomic DNA from a single *O. dioica* male. A low amount of extracted DNA
102 is an issue when working with small-size organisms like *O. dioica*. Therefore, we
103 optimized the extraction and sequencing protocols to allow for low-template input DNA
104 yields of around 200 ng and applied a hybrid sequencing approach using Oxford
105 Nanopore reads to span repeat-rich regions and Illumina reads to correct individual
106 nucleotide errors. The Nanopore run gave 8.2 million reads (221 \times coverage) with a
107 median length of 840 bp and maximum length of 166 kb (Fig. 2A). Based on k-mer
108 counting of the Illumina reads, the genome was estimated to contain ~50 Mbp (Fig. 2B)
109 – comparable in size to the OdB3 and OSKA2016 assemblies – and a relatively high
110 heterozygosity of ~3.6%. We used the Canu pipeline (Koren *et al.*, 2017) to correct, trim
111 and assemble Nanopore reads, yielding a draft assembly comprising 175 contigs with a

112 weighted median N50 length of 3.2 Mbp. We corrected sequencing errors and local
113 misassemblies of the draft contigs with Nanopore reads using Racon, and then with
114 Illumina reads using Pilon. The initial Okinawa *O. dioica* assembly length was 99.3 Mbp,
115 or ~1.5 times longer than the OdB3 genome at 70.4 Mbp. Merging haplotypes with
116 HaploMerger2 resulted in two sub-assemblies (reference and allelic) of 64.3 Mbp with an
117 N50 of 4.7 Mbp (I69-4). Repeating the procedure on a second individual from the same
118 culture showed overall agreement in assembly lengths, sequences and structures
119 (Fig. 2C).

120 To scaffold the genome, we sequenced Hi-C libraries from a pool of ~50 individuals from
121 the same culture. More than 99% of the Hi-C reads could be mapped to the contig
122 assembly. After removing duplicates, Hi-C contacts were passed to the 3D-DNA pipeline
123 to correct major misassemblies, as well as order and orient the contigs. The resulting
124 assembly named consisted of 8 megabase-scale scaffolds containing 99% of the total
125 sequence (Fig. 3A), and 14 smaller scaffolds that account for the remaining 663 Kbp
126 (lengths ranging from 2.9 to 131.6 Kbp). One of the small scaffolds is a draft assembly of
127 mitochondrial genome that we discuss below. Most of the other smaller scaffolds are
128 highly repetitive and might represent unplaced fragments of centromeric or telomeric
129 regions. We annotated telomeres by searching for the TTAGGG repeat sequence and
130 found that most of the megabase-scale scaffolds have single telomeric regions: therefore,
131 we reasoned that they represent chromosome arms. Indeed, pairwise genome alignment
132 to OdB3 identified two syntenic scaffolds for each autosomal linkage group, two for the
133 pseudo-autosomal region (PAR) and one for each sex-specific region. Since we had
134 previously inferred a karyotype of $n = 3$ by immunohistochemistry (Liu *et al.*, 2020), we
135 completed the assembly by pairing the megabase-scale scaffolds into chromosome arms
136 based on their synteny with the OdB3 physical map (Fig. 3B). The final assembly named
137 OKI2018_I69 (Table 1; Suppl. Table 1) comprises telomere-to-telomere assemblies of
138 the autosomal chromosomes 1 (chr 1) and 2 (chr 2). The sex chromosomes are split in
139 pseudo-autosomal region (PAR) and X-specific region (XSR) or Y-specific region (YSR;
140 Fig. 3). We assume that the sex-specific regions belong to the long arm of the PAR, as
141 the long arm does not comprise any telomeric repeats (Fig. 4A).

142 **Table 1:** Comparison of the OKI2018_I69 assembly with the previously published *O. dioica* genomes.

	OdB3	OSKA2016	OKI2018_I69
Geographical origin	Bergen, Norway (North Atlantic)	Osaka, Honshu, Japan (Western Pacific)	Okinawa, Japan (Ryukyu archipelago)
Assembly length (Mbp)	70.4	65.6	64.3
Number of scaffolds	1,260	576	19
Longest scaffold (Mbp)	3.2	6.8	17.1
Scaffold N50 (Mbp)	0.4	1.5	16.2
Number of contigs	5,917	746	42
Contig N50 (Mbp)	0.02	0.6	4.7
GC content (%)	39.77	41.34	41.06
Gap rate (%)	5.589	0.585	0.018
Complete BUSCOs (%)	70.8	71.7	73.6

144 The genome-wide contact matrix from the Hi-C data (Fig. 3C) shows bright, off-diagonal
145 spots that suggest spatial clustering of the telomeres and centromeres both within the
146 same and across different chromosomes (Dudchenko *et al.*, 2017). The three centromeric
147 regions are outside the sex-specific regions, dividing the PAR and both autosomes into
148 long and short arms. The two sex-specific regions have lower apparent contact
149 frequencies compared with the rest of the assembly which is consistent with their haploid
150 status in males. The chromosome arms themselves show few interactions between each
151 other, even when they are part of the same chromosome.

152 **Chromosome-level features**

153 The genome contains between 1.4 and 2.6 Mbp of tandem repeats (detected using the
154 tantan and ULTRA algorithms respectively with maximum period lengths of 100 and
155 2,000). Subtelomeric regions tend to contain retrotransposons or tandem repeats with
156 longer periods. We also found telomeric repeats in smaller scaffolds. A possible
157 explanation is that subtelomeric regions display high heterozygosity, leading to duplicated
158 regions that fail to assemble with the chromosomes. Alternatively, these scaffolds could
159 be peri-centromeric regions containing interstitial telomeric sequences. In some species,
160 high-copy tandem repeats can be utilized to discover the position of centromeric regions
161 (Melters *et al.*, 2013). However, we could not find such regions. Additional experimental
162 techniques such as chromatin immunoprecipitation and sequencing with centromeric
163 markers might be necessary to resolve the centromeres precisely. Therefore, the current
164 assembly skips over centromeric regions, represented as gaps of an arbitrary size of 500
165 bp in the chromosomal scaffolds.

166 We studied genome-scale features by visualizing them along whole chromosomes, from
167 the short to long arm, centered on their centromeric regions. Most strikingly, there is a
168 clear difference in sequence content between chromosome arms (Fig. 4; Supp. Table 3).
169 For each chromosome, small arms consistently display depleted GC content and elevated
170 repetitive content compared with the same chromosome's long arm. Although GC content
171 tends to be weakly negatively correlated with repeat content, it is difficult to ascertain
172 whether repetitive elements tend to drive changes in GC content or if changes in GC
173 content tend to drive the accumulation of repetitive sequence content. In either case, the
174 mechanism behind the marked difference in sequence content between short and large
175 chromosome arms remains unknown. It should be noted that the GC difference between
176 short and long arms also has an effect on the availability of DpnII restriction enzyme
177 recognition sites used for Hi-C library preparation, which recognizes the GC-rich motif
178 /GATC, although this bias is likely insufficient to explain the low degree of intra-
179 chromosomal interaction observed in the Hi-C contact maps.

180 **Quality assessment using BUSCO**

181 To assess the completeness of our assembly, we searched for 978 metazoan
182 Benchmarking Universal Single-Copy Orthologs (BUSCOs) provided with the BUSCO
183 tool (Simão *et al.*, 2015; Waterhouse *et al.*, 2017; Zdobnov *et al.*, 2017). To increase

184 sensitivity, we trained BUSCO's gene prediction tool, AUGUSTUS (Hoff and
185 Stanke, 2019), with transcript models generated from RNA-Seq data collected from the
186 same laboratory culture (see below). We detected 73.0% of BUSCOs (Table 1), which is
187 similar to OdB3 and OSKA2016 (Fig. 5A; Suppl. Table 4). All detected BUSCOs except
188 one reside on the chromosomal scaffolds. As the reported fraction of detected genes is
189 lower than for other tunicates such as *Ciona intestinalis* HT (94.6%; Satou *et al.*, 2019)
190 or *Botrylloides leachii* (89%; Blanchoud *et al.*, 2018), we searched for BUSCO genes in
191 the transcriptomic training data (83.0% present) and confirmed the presence of all but
192 one by aligning the transcript sequence to the genome. We then inspected the list of
193 BUSCO genes that were found neither in the genome nor in the transcriptome.
194 Bibliographic analysis confirmed that BUSCO genes related to the peroxisome were lost
195 (Žárský and Tachezy, 2015; Kienle *et al.*, 2016). There are two possible explanations for
196 the remaining missing genes: first is that protein sequence divergence (Berná *et al.*, 2012)
197 or length reduction (Berná and Alvarez-Valin, 2015) in *Oikopleura* complicate detection
198 by BUSCO, and second is gene loss. In line with the possibility of gene loss, most BUSCO
199 genes missing from our assembly are also undetectable in OdB3 and OSKA2016 (Fig. B;
200 Suppl. Table 5). To summarize, the Okinawa assembly achieved comparable detection
201 of universal single-copy conserved orthologs in comparison to previous *O. dioica*
202 assemblies, and consistently undetectable genes may be in fact missing or altered in
203 *Oikopleura*.

204 **Repeat annotation**

205 In order to identify repetitive elements in the OKI2018_I69 genome, we combined the
206 results of several *de novo* repeat detection algorithms and used this custom library as an
207 input to RepeatMasker to identify repeat sequences. Interspersed repeats make up
208 14.39% of the assembly (9.25 Mbp; Fig. 6), comparable to the 15% reported for OdB3
209 (Denoeud *et al.*, 2010). Of the annotated elements, the most abundant type is the long
210 terminal repeats (LTRs; ~4.6%) with Ty3/gypsy *Oikopleura* transposons (TORs)
211 dominating 2.97 Mbp of the sequence. Short interspersed nuclear elements (SINEs)
212 make up a smaller portion of the OKI2018_I69 sequence (<0.1%) compared with the
213 OdB3 (0.62%). It has been suggested that SINEs contribute significantly to genome size
214 variation in other oikopleurids (Naville *et al.*, 2019), but further analysis is required to
215 determine whether that is the case at shorter evolutionary distances. Non-LTR LINE/Odin
216 and Penelope-like elements are large components of most oikopleurid genomes
217 (Naville *et al.*, 2019), but they are almost absent in the OKI2018_I69 assembly. Indeed,
218 44% of the Okinawa *O. dioica* predicted repeats could not be classified through searches
219 against repeat databases and may either represent highly divergent relatives of known
220 repeat classes, or else potentially represent novel repeats specific to Okinawan *O. dioica*.

221 **Gene annotation**

222 We annotated the OKI2018_I69 assembly using *ab initio* and RNA-Seq-based gene
223 predictions. The different predictions were refined and merged with EVidenceModeler
224 using the Okinawan transcriptome and Bergen ESTs and proteins as additional support.

225 To predict alternative isoforms and update the models, we ran the PASA annotation
226 pipeline that yielded 18,485 transcript isoforms distributed among 16,936 protein-coding
227 genes. The number of predicted genes for the OKI2018_I69 is lower than what was
228 reported for OdB3 (18,020; Denoeud *et al.*, 2010) and OSKA2016 (18,743; Wang *et*
229 *al.*, 2020). The rest of the genes are either lost from the Okinawan *O. dioica* genome or
230 were not assembled and/or annotated with our pipeline. On the other side, higher number
231 of genes might be artifact of the OdB3 and OSKA2016 annotations. The completeness of
232 our annotation compares to the genome: BUSCO recovered 75.7% complete and 5.3%
233 fragmented metazoan genes (Fig. 5A). Like in the OdB3 assembly, gene density is very
234 high at one gene per 3.69 Kbp. Genes have very short introns (median length at 46 bp)
235 and intergenic spaces (average length 1,206 bp) (Table 2). Therefore, overall genomic
236 features seem to be conserved among *O. dioica* population despite large geographic
237 distance.

238 **Table 2:** Overall characterization of the OKI2018_I69 genome assembly

	OKI2018_I69
Repeats (%)	14.39
Number of genes	16,936
Number of isoforms	18,485
Median gene length (bp)	1,509
Median exon length (bp)	161
Median intron length	46

239
240 The ribosomal DNA gene encoding the precursor of the 18S, 5.8S and 28S rRNAs occurs
241 as long tandem repeats that form specific chromatin domains in the nucleolus. We
242 identified 4 full tandem copies of the rDNA gene at the tip of the PAR's short arm,
243 separated by 8,738 bp (median distance). As this region has excess coverage of raw
244 reads, and since assemblies of tandem repeats are limited by the read length (99% of
245 Nanopore reads in our data are shorter than 42,842 bp), we estimate that the real number
246 of the tandem rDNA copies could range between 5 (MiSeq) and 25 times (Nanopore)
247 larger. Between or flanking the rDNA genes, we also found short tandem repeats made
248 of 2 to 3 copies of a 96-bp sequence. This tandem repeat is unique to the rDNA genes
249 and to our reference and draft genomes, and was not found in the OdB3 reference nor in
250 other larvacean genomes. The 5S rRNA is transcribed from loci distinct to the rDNA gene
251 tandem arrays. In *Oikopleura*, it has the particularity of being frequently associated with
252 the spliced leader (SL) gene and to form inverted repeats present in more than 40 copies
253 (Ganot *et al.*, 2004). We found 27 copies of these genes on every chromosomal scaffold
254 except YSR, 22 of which were arranged in inverted tandem repeats. Altogether, we found
255 in our reference genome one rDNA gene repeat region assembled at the end of a
256 chromosome short arm. This sequence might provide useful markers for phylogenetic
257 studies in the future.

258

259

260 **Draft mitochondrial genome scaffold**

261 We identified a draft mitochondrial genome among the smaller scaffolds, chrUn_12, by
262 searching for mitochondrial sequences using the Cox1 protein sequence and the ascidian
263 mitochondrial genetic code (Pichon *et al.*, 2019). Automated annotation of this scaffold
264 using the MITOS2 server detected the coding genes *cob*, *cox1*, *nad1*, *cox3*, *nad4*, *cox2*,
265 and *atp6* (Fig. 7A), which are the same as in Denoeud *et al.*, 2010 except for the *nd5*
266 gene that is missing from our assembly. The open reading frames are often interrupted
267 by T-rich regions, in line with Denoeud *et al.* (2010). However, we cannot rule out the
268 possibility that these regions represent sequencing errors, as homopolymers are difficult
269 to resolve with the Nanopore technology available in 2019. The *cob* gene is interrupted
270 by a long non-coding region, but this might be a missassembly. Indeed, an independent
271 assembly using the *flye* software (Kolmogorov *et al.*, 2019) with the --meta option to
272 account for differential coverage also produced a draft mitochondrial genome, but its non-
273 coding region was ~2 kbp longer. Moreover, a wordmatch dotplot shows tandem repeats
274 in this region (Fig. 7B), and thus this region is prone to assembly errors, especially with
275 respect to the number of repeats. Altogether, the draft contig produced in our assembly
276 shows as a proof of principle that sequencing reads covering the mitochondrial genome
277 alongside the nuclear genome can be produced from a single individual, although it may
278 need supporting data such as targeted resequencing in order to be properly assembled.

279

280 **Discussion**

281 **OKI2018_I69 assembly quality**

282 Previously, different techniques have been used to sequence and assemble *O. dioica*
283 genomes which have produced assemblies of varying quality. The Sanger-based OdB3
284 sequence was published in 2010 (Denoeud *et al.*, 2010). Due to limitations in sequencing
285 technologies at the time, it is highly fragmented, comprising 1,260 scaffolds with an N50
286 of 0.4 Mbp. The recently released OSKA2016 assembly was generated from long-read
287 PacBio data and, therefore, has a larger N50 and fewer scaffolds (Table 1, Wang *et*
288 *al.* 2020). Both assemblies have high sequence quality and nearly full genome coverage,
289 but neither of them contains resolved chromosomes. However, Denoeud *et al.* (2010)
290 released a physical map calculated for OdB3 from BAC end sequences that comprises
291 five linkage groups (LGs): two autosomal LGs, one pseudo-autosomal region of sex
292 chromosomes, and two sex specific regions (X and Y).

293 The use of reference chromosome information from a closely related species to order
294 contigs or scaffolds into chromosome-length sequences is a common way to generate
295 final genome assemblies (*Drosophila* 12 Genomes Consortium, 2007). However, this
296 approach precludes discovery of structural variances. In our study, we first assembled
297 long Nanopore reads *de novo* into contigs that we ordered and joined into megabase-
298 scale scaffolds using long-range Hi-C data. The synteny-based approach with OdB3's

299 linkage groups as a reference was only required to guide final pairing of chromosome
300 arms into single scaffolds of chr 1, chr 2 and PAR, as we found that these scaffolds mostly
301 align to one of the autosomal LGs or PAR. Therefore, any potential assembly errors in
302 Odb3 would not be transferred to our assembly. Apart from these syntenic relationships,
303 our karyotyping results and the count of three centromeres on the Hi-C contact map
304 supports the presence of three pairs of chromosomes in the Okinawan *O. dioica*.
305 However, there is a possibility that chromosome arms might have been exchanged
306 between chromosomes in the Okinawan population. Additional experimental evidence is
307 needed to confirm the pairing of chromosome arms, such as data generated by the Omni-
308 C method which does not rely on restriction enzyme fragmentation.

309 Our synteny-based scaffolding assumes that animals collected from the Atlantic and
310 Pacific oceans are from the same species and conserve these chromosomal properties.
311 However, there are visible differences in gene number and repeat content compared with
312 the Odb3 and OSKA2016. *O. dioica* is distributed all over the world, and all the
313 populations are classified as a single species owing to the lack of obvious morphological
314 differences and limited understanding of population structure. However, the short life span
315 of *O. dioica* combined with limited mobility and high mutation rate contribute to an
316 accelerated genome evolution that might have led to multiple speciation events.
317 Sequence polymorphism was previously noted when comparing the Odb3 genome to
318 genomic libraries of a laboratory strain collected on the North American Pacific coast
319 (Denoeud *et al.*, 2010), and more recently when comparing Odb3 to OSKA2010 (Wang *et*
320 *al.*, 2015; Wang *et al.*, 2020). Further work will be needed to elucidate the relation of the
321 Okinawan population to the North Atlantic and North Pacific ones.

322 **Inter-arm contacts**

323 The sequence of *O. dioica*'s chromosomes and their contact map suggest that
324 chromosome arms may be the fundamental unit of synteny in larvaceans. Hi-C contact
325 matrices in vertebrates typically display greater intra-chromosomal than inter-
326 chromosomal interactions. A similar pattern was reported in the tunicate *Ciona robusta*
327 (also known as *intestinalis* type A; Satou *et al.*, 2019) and the lancelet *Branchiostoma*
328 *floridae* (Simakov, Marlétaz, *et al.*, 2020). By comparison, in flies and mosquitoes, the
329 degree of contacts between two arms of the same chromosome appear to be reduced
330 but nonetheless more frequent than between different chromosomes (Dudchenko *et*
331 *al.*, 2017). Indeed, in *Drosophila*, the chromosome arms – which are termed Muller
332 elements owing to studies with classical genetics (Schaeffer, 2018) – are frequently
333 exchanged between chromosomes across speciation events. *O. dioica*'s genome shares
334 with fruit flies its small size and small number of chromosomes. However, small
335 chromosome size is also seen in the tunicate *Ciona robusta*, which has 14 meta- or sub-
336 meta-centric pairs (Shoguchi *et al.*, 2005), with an average length of ~8 Mbp (Satou *et*
337 *al.*, 2019) that exhibit a more extensive degree of contacts, particularly for intra-
338 chromosomal interactions across the centromeres (Satou *et al.*, 2019). As we prepared
339 our Hi-C libraries from adult animals, where ploidy is high (Ganot and Thompson,

340 2002), we cannot rule out that it could be a possible cause of the low inter-arm interactions
341 in our contact matrix. Further studies such as investigations of other developmental
342 stages will be needed to elucidate the mechanism at work for the similarity between
343 *O. dioica* and insect's chromosome contact maps.

344 **Visualization and access**

345 We prepared a public view of our reference genome in the ZENBU browser (Severin *et*
346 *al.*, 2014), displaying tracks for our gene models, *in silico*-predicted features such as
347 repeats and non-coding RNAs, or syntenies with other *Oikopleura* genomes. To facilitate
348 the study of known genes, we screened the literature for published sequences
349 (Suppl. Table 6) and mapped them to the genome with a translated alignment. The
350 ZENBU track for these alignments is searchable by gene name, accession number and
351 PubMed identifier. Chromosome-level visualization of this track shows that the genes
352 studied so far are distributed evenly on each chromosome, except for the repeat-rich YSR
353 (Figure 8). In line with the observed loss of synteny in the Hox genes noted in *Oikopleura*
354 (Seo *et al.*, 2004), we did not see apparent clustering of genes by function or relatedness.
355 The view of the OKI2018_I69 genome assembly can be found here:

356 https://fantom.gsc.riken.jp/zenbu/gLyphs/#config=nPfav_juDdOmlG2t4LkJ1D;loc=OKI2018_I69_1.0::chr1:1..100000+.
357

358

359 **Conclusions**

360 We demonstrated that a combination of long- and short-read sequencing data from a
361 single animal, together with the long-range Hi-C data and the use of various bioinformatic
362 approaches can result in a high-quality *de novo* chromosome-scale assembly of
363 *O. dioica*'s highly polymorphic genome. However, further work is needed to properly
364 resolve the polymorphisms into separated haplotypes using a different approach, such as
365 trio-binning. We believe that the current version of the assembly will serve as an essential
366 resource for a broad range of biological studies, including genome-wide comparative
367 studies of *Oikopleura* and other species, and provides insights into chromosomal
368 evolution.

369

370 **Methods**

371 ***Oikopleura* sample and culture**

372 Wild live specimens were collected from Ishikawa Harbor (26°25'39.3"N 127°49'56.6"E)
373 by a hand-held plankton net and returned to the lab for culturing (Masunaga *et al.*, 2020).
374 A typical generation time from hatchling to fully mature adult is 4 days at 23°C for the
375 Okinawan *O. dioica*. Individuals I28 and I69 were collected at generation 44 and 47,
376 respectively.

377 **Isolation and sequencing of DNA**

378 Staged fully mature males were collected prior to spawning. These were each washed
379 with 5 ml filtered autoclaved seawater (FASW) for 10 min three times before resuspension
380 in 50 μ l 4 M guanidium isothiocyanate, 0.5% SDS, 50 mM sodium citrate and 0.05% v/v
381 2-mercaptoethanol. This was left on ice for 30 min before being precipitated with 2
382 volumes of ice-cold ethanol and centrifuged at 14,000 rpm 4°C for 20 min. The pellet was
383 washed with 1 ml of 70% cold-ethanol, centrifuged at 14,000 rpm 4°C for 5 min and air
384 dried briefly before resuspension in 200 μ l 100 mM NaCl, 25 mM EDTA, 0.5% SDS and
385 10 μ g/ml proteinase K. The lysates were incubated overnight at 50°C. The next morning,
386 the total nucleic acids were first extracted and then back-extracted once more with
387 chloroform:phenol (1:1). Organic and aqueous phases were resolved by centrifugation at
388 13,000 rpm for 5 min for each extraction; both first and back-extracted aqueous phases
389 were collected and pooled. The pooled aqueous phase was subjected to a final extraction
390 with chloroform and spun down as previously described. The aqueous fraction was then
391 removed and precipitated by centrifugation with two volumes of cold ethanol and 10 μ g/ml
392 glycogen; washed with 1 ml of cold 70% ethanol and centrifuged once more as previously
393 described. The resulting pellet was allowed to air-dry for 5 min and finally resuspended in
394 molecular biology grade H₂O for quantitation using a Qubit 3 Fluorometer (Thermo Fisher
395 Scientific, Q32850), and the integrity of the genomic DNA was validated using Agilent
396 4200 TapeStation (Agilent, 5067-5365).

397 Isolated genomic DNA used for long-reads on Nanopore MinION platform were
398 processed with the Ligation Sequencing Kit (Nanopore LSK109) according to
399 manufacturer's protocol, loading approximately 200 ng total sample per R9.4 flow-cell.
400 Raw signals were converted to sequence files with the Guppy proprietary software (model
401 "template_r9.4.1_450bps_large_flipflop", version 2.3.5). Approximately 5 ng was set
402 aside for whole genome amplification to perform sequencing on Illumina MiSeq platform,
403 using the TruePrime WGA Kit (Sygnis, 370025) according to manufacturer's protocol.
404 Magnetic bead purification (Promega, NG2001) was employed for all changes in buffer
405 conditions required for enzymatic reactions and for final buffer suitable for sequencing
406 system. Approximately 1 μ g of amplified DNA was sequenced by our core sequencing
407 facility with a 600-cycle MiSeq Reagent Kit v3 (Illumina, MS-102-3003) following the
408 manufacturer's instructions. These Illumina runs were used for polishing and error
409 checking of Nanopore runs.

410 **Hi-C library preparation**

411 50 fully matured males were rinsed 3 times for 10 min each by transferring from well to
412 well in a 6-well plate filled with 5 ml FASW. Rinsed animals were combined in a 1.5 ml
413 microcentrifuge tube. Tissues were pelleted for 10 min at 12,000 rpm and leftover FASW
414 was discarded. A Hi-C library was then prepared by following the manufacturer's protocol
415 (Dovetail, 21004). Briefly, tissues were cross-linked for 20 min by adding 1 ml 1 \times PBS
416 and 40.5 μ l 37% formaldehyde to the pellet. The tubes were kept rotating to avoid tissue
417 settle during incubation. Cross-linked DNA was then blunt-end digested with DpnII

418 (Dovetail) to prepare ends for ligation. After ligation, crosslinks were reversed, DNA was
419 purified by AMPure XP Beads (Beckman, A63880) and quantified by Qubit 3 Fluorometer
420 (Thermo Fisher Scientific, Q10210). The purified DNA was sheared to a size of 250-450
421 bp by sonication using a Covaris M220 instrument (Covaris, Woburn, MA) with peak
422 power 50 W, duty factor 20, and cycles/burst 200 times for 65 s. DNA end repair, adapter
423 ligation, PCR enrichment, and size selection were carried out by using reagents provided
424 with the kit (Dovetail, 21004). Finally, the library was checked for quality and quantity on
425 an Agilent 4200 TapeStation (Agilent, 5067-5584) and a Qubit 3 Fluorometer. The library
426 was sequenced on a MiSeq (Illumina, SY-410-1003) platform using a 300 cycles V2
427 sequencing kit (Illumina, MS-102-2002), yielding 20,832,357 read pairs.

428 **Genome size estimation**

429 Jellyfish (Marçais, Kingsford, 2011) was used to generate k-mer count profiles for various
430 values of k (17, 21, 25, 29, 33, 37, and 41) based on the genome-polishing Illumina MiSeq
431 reads, with a maximum k-mer count of 1000. These k-mer profiles were subsequently
432 used to estimate heterozygosity and genome size parameters using the GenomeScope
433 web server (<http://qb.cshl.edu/genomescope/>).

434 **Filtering of Illumina MiSeq raw reads**

435 Before using at different steps, all raw Illumina reads were quality-filtered (-q 30, -p 70)
436 and trimmed on both ends with the FASTX-Toolkit v0.0.14
437 (http://hannonlab.cshl.edu/fastx_toolkit/index.html). The quality of the reads before and
438 after filtering were checked with FASTQC v0.11.5 (Andrews *et al.*, 2010). Read pairs that
439 lacked one of the reads after the filtering were discarded in order to preserve paired-end
440 information.

441 **Genome assembly**

442 Genome assembly was conducted with the Canu pipeline v1.8 (Koren *et al.*, 2017) and
443 32.3 Gb (~221.69 \times) raw Nanopore reads (correctedErrorRate=0.105,
444 minReadLength=1000). The resulting contig assembly was polished three times with
445 Racon v1.2.1 (Vaser *et al.*, 2017) using Canu-filtered Nanopore reads. Nanopore-specific
446 errors were corrected with Pilon v1.22 (Walker *et al.*, 2014) using filtered 150-bp paired-
447 end Illumina reads (~99.7 \times). Illumina reads were aligned to the Canu contig assembly
448 with BWA v0.7.17 (Li *et al.*, 2013) and the corresponding alignments were provided as
449 input to Pilon. Next, one round of the Haplomerger2 processing pipeline (Huang *et*
450 *al.*, 2017) was applied to eliminate redundancy in contigs and to merge haplotypes.

451 Contigs were joined into scaffolds based on long-range Hi-C Dovetail™ data using Juicer
452 v1.6 (Durand *et al.*, 2016) and 3D de novo assembly (3D-DNA; Dudchenko *et al.*, 2017)
453 pipelines. The megabase-scale scaffolds were joined into pairs of chromosome arms
454 based on their synteny with the Odb3 physical map (see below). The candidate assembly
455 was visualized and reviewed with Juicebox Assembly Tools (JBAT) v1.11.08
456 (<https://github.com/aidenlab/Juicebox>; Durand and Robinson 2016).

457 Whole-genome alignment between OKI2018_I69 and OdB3 assemblies was performed
458 using LAST v1066 (Kiełbasa *et al.*, 2011). The sequence of OdB3 linkage groups were
459 reconstructed as defined in the Supplementary Figure 2 in Denoeud *et al.* 2010. The
460 resulting alignments were post-processed in R with a custom script
461 (<https://github.com/oist/oikGenomePaper>) and visualized using the R package
462 “networkD3” (“sankeyNetwork” function). The color scheme for chromosomes was
463 adopted from R Package RColourBrewer, “Set2”.

464 The final assembly was checked for contamination by BLAST searches against the NCBI
465 non-redundant sequence database. 12 smaller scaffolds were found to have strong
466 matches to bacterial DNA (Suppl. Table 2), as well as possessing significantly higher
467 Nanopore sequence coverage (>500×) than the rest of the assembly, and were therefore
468 removed from the final assembly.

469 The completeness and quality of the assembly were checked with QUAST v5.0.2
470 (Gurevich *et al.*, 2013) and by searching for the set of 978 highly conserved metazoan
471 genes (OrthoDB version 9.1; Zdobnov *et al.*, 2017) using BUSCO v3.0.2 (Simão *et*
472 *al.*, 2015; Waterhouse *et al.*, 2017). The --sp option was set to match custom AUGUSTUS
473 parameters (Hoff and Stanke, 2019) trained using the Trinity transcriptome assembly (see
474 below) split 50% / 50% for training and testing.

475 **Repeat masking and transposable elements**

476 A custom library of repetitive elements (RE) present in the genome assembly was built
477 with RepeatModeler v2.0.1 that uses three *-de novo* -repeat finding programs: RECON
478 v1.08, RepeatScout v1.0.6 and LtrHarvest/Ltr_retriever v2.8. In addition, MITE-Hunter
479 v11-2011 (Han and Wessler, 2010) and SINE_Finder (Wenke and Torsten, *et al.*, 2011)
480 were used to search for MITE and SINE elements, respectively. The three libraries were
481 pooled together as input to RepeatMasker v4.1.0 (Smit, Hubley and Green, 2015) to
482 annotate and soft-mask these repeats in the genomic sequence. Resulting sets of REs
483 were annotated by BLAST searches against RepeatMasker databases and sequences of
484 transposable elements published for different oikopleurids (Naville *et al.*, 2019).

485 Tandem repeats were detected using two different programs, tantan (Frith, 2011) and
486 ULTRA (Olson and Wheeler, 2018) using two different maximal period lengths (100 and
487 2000). Version 23 of tantan was used with the parameters -f4 (output repeats) and -w100
488 or 2000 (maximum period length). ULTRA version 0.99.17 was used with -mu 2 (minimum
489 number of repeats) -p 100 or 2000 (maximum period length) and -mi 5 -md 5 (maximum
490 consecutive insertions or deletions). ULTRA detected more tandem repeats than tantan,
491 but its predictions include more than 90% of tantan’s. Both tools detected *O. dioica*’s
492 telomeric tandem repeat sequence, which is TTAGGG as in other chordates
493 (Schulmeister *et al.*, 2007).

494

495 **Developmental staging, isolation and sequencing of mRNA, transcriptome**
496 **assembly**

497 Mixed stage embryos, immature adults (3 days after hatching) and adults (4 days after
498 hatching) were collected separately from our on-going laboratory culture for RNA-Seq
499 analysis. Eggs were washed three times for 10 min by moving eggs along with
500 micropipette from well to well in a 6-well dish each containing 5 ml of FASW and left in a
501 fresh well of 5 ml FASW in the same dish. These were stored at 17°C and set aside for
502 fertilization. Matured males, engorged with sperm, were also washed 3 times in FASW.
503 Still intact mature males were placed in 100 μ l of fresh FASW and allowed to spawn
504 naturally. Staged embryos were initiated by gently mixing 10 μ l of the spawned male
505 sperm to the awaiting eggs in FASW at 23°C. Generation 30 developing embryos at 1 h
506 and 3 h post-fertilization were visually verified by dissecting microscope and collected as
507 a pool for the mixed staged embryo time point. Immature adults at generation 31 and
508 sexually differentiated adults at generation 30 were used for the 2 adult staged time
509 points. All individuals for each time point were pooled and washed with FASW three times
510 for 10 min. Total RNA was extracted and isolated with RNeasy Micro Kit (Qiagen, 74004)
511 and quantitated using Qubit 3 Fluorometer (Thermo Fisher Scientific, Q10210). Additional
512 quality control and integrity of isolated total RNA was checked using Agilent 4200
513 TapeStation (Agilent, 5067-5576). Further processing for mRNA selection was performed
514 with Oligo-d(T)25 Magnetic Beads (NEB, E7490) and the integrity of the RNA was
515 validated once more with Agilent 4200 TapeStation (Agilent, 5067-5579). Adapters for the
516 creation of DNA libraries for the Illumina platform were added per manufacturer's
517 guidance (NEB, E7805) as were unique indexed oligonucleotides (NEB, E7600) to each
518 of the 3 staged samples. Each cDNA library was sequenced paired-end with a 300-cycle
519 MiSeq Reagent Kit v2 (Illumina, MS-102-2002) loaded at approximately 12 pM.

520 After quality assessment and data filtering (see Filtering of Illumina MiSeq raw reads),
521 Illumina RNA-Seq reads were pooled together and *de novo* assembled with Trinity v2.8.2
522 (Grabherr *et al.*, 2011). Redundancy in the transcriptome assembly was removed by CD-
523 HIT v4.8.1 (Li and Godzik, 2006) with a cut-off value of 95% identity. The quality and
524 completeness of the transcriptome assembly was verified with rnaQUAST v1.5.1
525 (Bushanova *et al.*, 2016) and BUSCO.

526 **Gene prediction and annotation**

527 Gene models were predicted using both AUGUSTUS v3.3 (Stanke *et al.* 2006) and the
528 MAKER pipeline v3.01.03 (Cantarel *et al.*, 2008). AUGUSTUS was trained following the
529 Hoff and Stanke protocol (2019) with the initial RNA-Seq reads and transcriptome
530 assembly used as intron and exon hints, correspondingly. Transcript models were
531 generated with the PASA pipeline v20140417 (Haas *et al.*, 2003) using BLAT v36 and
532 GMAP v2018-02-12 to align transcripts to the genome. RNA-Seq reads were mapped to
533 the genome with STAR v2.0.6a (Dobin *et al.*, 2013). Running AUGUSTUS in *ab initio*
534 mode resulted in 14,327 genes, whereas prediction with hints resulted in 17,277 genes.

535 ESTs and proteins from the Bergen *O. dioica* and our transcriptome assembly were used
536 as evidence to run the MAKER pipeline (<https://reslp.github.io/blog/My-MAKER-Pipeline/>)
537 that resulted in a set of 17,480 predicted genes. To finalize consensus gene structure,
538 three predictions (weight = 1) were combined and subsequently refined with
539 EvidenceModeler (EVM) v1.1.1 (Haas *et al.*, 2008) using PASA transcript assemblies
540 (weight = 5) and proteins from the Bergen *O. dioica* (weight = 1) aligned to the genome
541 with Exonerate v2.2.0, yielding 16,956 protein-coding genes. To predict UTRs and
542 alternatively spliced isoforms, the EVM models were updated using two rounds of the
543 PASA pipeline, which resulted in a final set of 18,485 transcript models distributed among
544 16,936 genes. Chromosomal coordinates were ported to our final assembly using the
545 Liftoff tool (Shumate and Salzberg, 2020). The quality of the predicted gene models was
546 assessed with BUSCO.

547 A draft annotation of the mitochondrial genome was obtained by submitting the
548 corresponding scaffold (chr_Un12) as input to the MITOS2 mitochondrial genome
549 annotation server (Bernt *et al.*, 2013; accessed May 28, 2020) with the ascidian
550 mitochondrial translation table specified (Denoeud *et al.*, 2010; Pichon *et al.*, 2019).

551 **Detection of coding RNAs**

552 A translated alignment was used to detect known *O. dioica* genes available from
553 GenBank using the TBLASTN software (Gertz *et al.*, 2006) with the options -ungapped
554 -comp_based_stats F to prevent *O. dioica*'s small introns from being incorporated as
555 alignment gaps, and -max_intron_length 100000 to reflect the compactness of
556 *O. dioica*'s genome. The best hits were converted to GFF3 format using BioPerl's
557 bp_search2gff program (Stajich *et al.*, 2002) before being uploaded to the ZENBU
558 genome browser (Severin *et al.*, 2014). For some closely related pairs of genes that gave
559 ambiguous results with that method, we searched for the protein sequence in our
560 transcriptome assembly with TBLASTN, located the genomic region where the best
561 transcript model hit was aligned, and selected the hit from the original TBLASTN search
562 that matched this region. We summarized our results in Suppl. Table 6. For both
563 searches, we used an *E*-value filter of 10^{-40} . Genes marked as not found in the table
564 might be present in the genome while failing to pass the filter.

565 **Detection of non-coding RNAs**

566 To validate the results of cmscan on rRNAs, genomic regions were screened with a
567 nucleotide BLAST search using the *O. dioica* isolate MT01413 18S ribosomal RNA gene,
568 partial sequence (GenBank:KJ193766.1). 200-kbp windows surrounding the hits where
569 then analysed with the RNAmer 1.2 web service (Lagesen *et al.*, 2007). RNAmer did
570 not detect the 5.8S RNA, but we could confirm its presence by a nucleotide BLAST search
571 using the AF158726.1 reference sequence. The loci containing the 5S rRNA (AJ628166)
572 and the spliced leader RNA (AJ628166) were detected with the exonerate 2.4 software
573 (Slater and Birney, 2005), with its affine:local model and a score threshold of 1000 using
574 the region chr1:8487589-8879731 as a query.

575 **Whole-genome alignments**

576 Pairs of genomes were mapped to each other with the LAST software (Kiełbasa *et*
577 *al.*, 2011) version 1066. When indexing the reference genome, we replaced the original
578 lowercase soft masks with ones for simple repeats (lastdb -R01) and we selected a
579 scoring scheme for near-identical matches (-uNEAR). Substitution and gap frequencies
580 were determined with last-train (Hamada *et al.*, 2017), with the alignment options -
581 E0.05 -C2 and forcing symmetry with the options --revsym --matsym --gapsym. An
582 optimal set of pairwise one-to-one alignments was then calculated using last-split
583 (Frith and Kawaguchi, 2015). For visualization of the results, we converted the alignments
584 to GFF3 format and collated the colinear "match_part" alignment blocks in "match"
585 regions using LAST's command maf-convert -J 200000. We then collated syntenic
586 region blocks (sequence ontology term SO:0005858) that map to the same sequence
587 landmark (chromosome, scaffold, contigs) on the query genome with a distance of less
588 than 500,000 bp with the custom script syntenic_regions.sh (supplementary Git
589 repository). In contrast to the "match" regions, the syntenic ones are not necessarily
590 colinear and can overlap with each other. The GFF3 file was then uploaded to the ZENBU
591 genome browser.

592 **Nanopore read realignments**

593 Nanopore reads were realigned to the genome with the LAST software as in the whole-
594 genome alignments above. FASTQ qualities were discarded with the option -Q0 of
595 lastal. Optimal split alignments were calculated with last-split. Alignment blocks
596 belonging to the same read were joined with maf-convert -J 1e6 and the custom script
597 syntenic_regions_stranded.sh. The resulting GFF3 files were loaded in the ZENBU
598 genome browser to visualize the alignments near gap regions in order to check for reads
599 spanning the gaps.

600 **Analysis of sequence properties across chromosome-scale scaffolds**

601 Each chromosome-scale scaffold was separated into windows of 50 Kbp and evaluated
602 for GC content, repeat content, sequencing depth, and the presence of DpnII restriction
603 sites. For chr 1, chr 2, and the PAR, windows corresponding to long and short
604 chromosome arms were separated based on their positioning relative to a central gap
605 region (chr 1 short arm: 1-5,191,657 bp, chr 1 long arm: 5,192,156-14,533,022 bp; chr 2
606 short arm: 1-5,707,009, chr 2 long arm: 5,707,508-16,158,756 bp; PAR short arm: 1-
607 6,029,625 bp, PAR long arm: 6,030,124-17,092,476). Since none of our assemblies or
608 sequencing reads spanned both the PAR and either sex-specific chromosome, the X and
609 Y chromosomes were excluded from this analysis. For each of GC content, sequencing
610 depth, repeat content, gene count, and DpnII restriction sites, the significance of the
611 differences between long and short arms was assessed with Welch's two-sided T test as
612 well as a nonparametric Mann-Whitney test implemented in R (Suppl. Table 3). The
613 results of the two tests were largely in agreement, but groups were only indicated as
614 significantly different if they both produced significance values below 0.05 ($p < 0.05$).

615 **Data access**

616 Sequence data was deposited to the ENA database (study ID PRJEB40135). Genome
617 assembly and annotation were deposited to the NCBI (Accession number pending) and
618 to Zenodo (DOI 10.5281/zenodo.4023777).

619 Custom scripts used in this study are available in GitHub
620 (<https://github.com/oist/oikGenomePaper>).

621

622 **Acknowledgements**

623 We would like to thank the DNA Sequencing Section and the Scientific Computing and
624 Data Analysis Section of the Research Support Division at OIST for their support, Danny
625 Miller for advices on Nanopore sequencing, Dan Rokhsar, Gene Myers, Ferdinand
626 Marlétaz and Konstantin Khalturin for critical comments, Takeshi Onuma and Hiroki
627 Nishida for sharing the OSKA2016 genome sequence prior publication, and Simon
628 Henriet for sharing a DNA extraction protocol. MJM received funding as an International
629 Research Fellow of the Japan Society for the Promotion of Science. This work was
630 supported by core funding from OIST.

631 **Author contributions**

632 Conceptualization: AB, CP, NML; data curation: AB, MJM, CP; formal analysis: AB, MJM,
633 CW, TR, HCC, SK, JP, CP; investigation: AB, AM, MJM, YKT, AWL, CP;
634 methodology: AB, CP; project administration: AB, CP; software: MJM, CW;
635 supervision: CP, NML; validation: AB; visualization: AB, AM, MJM, YKT; writing – original
636 draft: AB, AM, MJM, YKT, CP; writing – review & editing: AB, MJM, CP, NML.

637

638 **List of abbreviations**

639 chr 1: autosomal chromosome 1; chr 2: autosomal chromosome 2; Kbp: Kilobase pairs;
640 LGs: linkage groups; Mbp: Megabase pairs; PAR: pseudo-autosomal regions; XSR: X-
641 specific region; YSR: Y-specific region.

642

643 **Conflict of interests**

644 The authors declare that they have no competing interests.

645

646

647

648

649 **References**

- 650 1. Alldredge AL. Discarded appendicularian houses as sources of food, surface
651 habitats, and particulate organic matter in planktonic environments. *Limnology and
652 Oceanography*. 1976;21(1):14-24.
- 653 2. Hopcroft RR, Roff JC. Zooplankton growth rates: extraordinary production by the
654 larvacean *Oikopleura dioica* in tropical waters. *Journal of Plankton Research*.
655 1995;17(2):205-20.
- 656 3. Sato R, Tanaka Y, Ishimaru T. House production by *Oikopleura dioica* (Tunicata,
657 Appendicularia) under laboratory conditions. *Journal of plankton research*.
658 2001;23(4):415-23.
- 659 4. Alldredge, A. The contribution of discarded appendicularian houses to the flux of
660 particulate organic carbon from oceanic surface waters. In: Gorsky, G.,
661 Youngbluth, M. J., Deibel, D., editors. *Response of Marine Ecosystems to Global
662 Change: Ecological Impact of Appendicularians*. Contemporary Publishing
663 International; 2005. p.309-326.
- 664 5. Masunaga A, Liu AW, Tan Y, Scott A, Luscombe NM. Streamlined sampling and
665 cultivation of the pelagic cosmopolitan larvacean, *Oikopleura dioica*. *JoVE (Journal
666 of Visualized Experiments)*. 2020;16(160):e61279.
- 667 6. Fenaux, R. Anatomy and functional morphology of the Appendicularia. In: Bone,
668 Q., editor. *The biology of pelagic tunicates*. Oxford University Press; 1998. P.25-
669 34.
- 670 7. Delsuc F, Brinkmann H, Chourrout D, Philippe H. Tunicates and not
671 cephalochordates are the closest living relatives of vertebrates. *Nature*.
672 2006;439(7079):965-8.
- 673 8. Seo HC, Kube M, Edvardsen RB, Jensen MF, Beck A, Spiet E, Gorsky G,
674 Thompson EM, Lehrach H, Reinhardt R, Chourrout D. Miniature genome in the
675 marine chordate *Oikopleura dioica*. *Science*. 2001;294(5551):2506.
- 676 9. Denoeud F, Henriet S, Mungpakdee S, Aury JM, Da Silva C, Brinkmann H,
677 Mikhaleva J, Olsen LC, Jubin C, Cañestro C, Bouquet JM. Plasticity of animal
678 genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science*.
679 2010;330(6009):1381-5.
- 680 10. Naville M, Henriet S, Warren I, Sumic S, Reeve M, Wolff JN, Chourrout D. Massive
681 changes of genome size driven by expansions of non-autonomous transposable
682 elements. *Current Biology*. 2019;29(7):1161-8.
- 683 11. Fredriksson G, Olsson R. The subchordal cells of *Oikopleura dioica* and *O.
684 albicans* (Appendicularia, Chordata). *Acta Zoologica*. 1991;72(4):251-6.
- 685 12. Wang K, Omotezako T, Kishi K, Nishida H, Onuma TA. Maternal and zygotic
686 transcriptomes in the appendicularian, *Oikopleura dioica*: novel protein-encoding
687 genes, intra-species sequence variations, and trans-spliced RNA leader.
688 *Development Genes and Evolution*. 2015;225(3):149-59.
- 689 13. Wang K, Tomura R, Chen W, Kiyooka M, Ishizaki H, Aizu T, Minakuchi Y, Seki M,
690 Suzuki Y, Omotezako T, Suyama R. A genome database for a Japanese

691 population of the larvacean *Oikopleura dioica*. Development, Growth &
692 Differentiation. 2020;62(6):450-61.

693 14. Körner WF. Untersuchungen über die gehäusebildung bei appendicularien
694 (*Oikopleura dioica* fol). Zeitschrift für Morphologie und Ökologie der Tiere.
695 1952;41(1):1-53.

696 15. Colombera D, Fenaux R. Chromosome form and number in the Larvacea. Italian
697 Journal of Zoology. 1973;40(3-4):347-53.

698 16. Liu AW, Tan Y, Masunaga A, Plessy C, Luscombe NM. Centromere-specific
699 antibody-mediated karyotyping of Okinawan *Oikopleura dioica* suggests the
700 presence of three chromosomes. bioRxiv. 2020; doi:10.1101/2020.06.23.166173.

701 17. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling
702 A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R. Comprehensive
703 mapping of long-range interactions reveals folding principles of the human
704 genome. science. 2009;326(5950):289-93.

705 18. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim
706 MS, Machol I, Lander ES, Aiden AP, Aiden EL. De novo assembly of the *Aedes*
707 *aegypti* genome using Hi-C yields chromosome-length scaffolds. Science.
708 2017;356(6333):92-5.

709 19. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: a
710 scalable and accurate long-read assembly via adaptive k-mer weighting and
711 repeat separation. Genome research. 2017;27(5):722-36.

712 20. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso
713 P, Eid J, Rank D, Garcia JF. Comparative analysis of tandem repeats from
714 hundreds of species reveals unique insights into centromere evolution. Genome
715 biology. 2013;14(1):1-20.

716 21. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
717 assessing genome assembly and annotation completeness with single-copy
718 orthologs. Bioinformatics. 2015;31(19):3210-2.

719 22. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G,
720 Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to
721 gene prediction and phylogenomics. Molecular biology and evolution.
722 2018;35(3):543-8.

723 23. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simao FA, Ioannidis
724 P, Seppey M, Loetscher A, Kriventseva EV. OrthoDB v9. 1: cataloging evolutionary
725 and functional annotations for animal, fungal, plant, archaeal, bacterial and viral
726 orthologs. Nucleic acids research. 2017;45(D1):D744-9.

727 24. Hoff KJ, Stanke M. Predicting genes in single genomes with augustus. Current
728 protocols in bioinformatics. 2019;65(1):e57.

729 25. Satou Y, Nakamura R, Yu D, Yoshida R, Hamada M, Fujie M, Hisata K, Takeda
730 H, Satoh N. A nearly complete genome of *Ciona intestinalis* Type A (*C. robusta*)
731 reveals the contribution of inversion to chromosomal evolution in the genus *Ciona*.
732 Genome biology and evolution. 2019;11(11):3144-57.

733 26. Blanchoud S, Rutherford K, Zondag L, Gemmell NJ, Wilson MJ. De novo draft
734 assembly of the *Botrylloides leachii* genome provides further insight into tunicate
735 evolution. *Scientific reports*. 2018;8(1):1-8.

736 27. Žárský V, Tachezy J. Evolutionary loss of peroxisomes—not limited to parasites.
737 *Biology direct*. 2015;10(1):1-0.

738 28. Kienle N, Kloepper TH, Fasshauer D. Shedding light on the expansion and
739 diversification of the Cdc48 protein family during the rise of the eukaryotic cell.
740 *BMC evolutionary biology*. 2016; 16(1):215.

741 29. Berná L, D'Onofrio G, Alvarez-Valin F. Peculiar patterns of amino acid substitution
742 and conservation in the fast evolving tunicate *Oikopleura dioica*. *Molecular*
743 *phylogenetics and evolution*. 2012;62(2):708-17.

744 30. Berná L, Alvarez-Valin F. Evolutionary volatile Cysteines and protein disorder in
745 the fast evolving tunicate *Oikopleura dioica*. *Marine genomics*. 2015; 24:47-54.

746 31. Ganot P, Kallesøe T, Reinhardt R, Chourrout D, Thompson EM. Spliced-leader
747 RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome.
748 *Molecular and cellular biology*. 2004;24(17):7795-805.

749 32. Pichon J, Luscombe NM, Plessy C. Widespread use of the “ascidian” mitochondrial
750 genetic code in tunicates. *F1000Research*. 2019;8.

751 33. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads
752 using repeat graphs. *Nature biotechnology*. 2019;37(5):540-6.

753 34. *Drosophila* 12 Genomes Consortium. Evolution of genes and genomes on the
754 *Drosophila* phylogeny. *Nature*. 2007;450(7167):203.

755 35. Simakov O, Marlétaz F, Yue JX, O'Connell B, Jenkins J, Brandt A, Calef R, Tung
756 CH, Huang TK, Schmutz J, Satoh N. Deeply conserved synteny resolves early
757 events in vertebrate evolution. *Nature Ecology & Evolution*. 2020;20:1-11.

758 36. Schaeffer SW. Muller “Elements” in *Drosophila*: how the search for the genetic
759 basis for speciation led to the birth of comparative genomics. *Genetics*.
760 2018;210(1):3-13.

761 37. Shoguchi E, Kawashima T, Nishida-Umehara C, Matsuda Y, Satoh N. Molecular
762 cytogenetic characterization of *Ciona intestinalis* chromosomes. *Zoological*
763 *science*. 2005;22(5):511-6.

764 38. Ganot P, Thompson EM. Patterning through differential endoreduplication in
765 epithelial organogenesis of the chordate, *Oikopleura dioica*. *Developmental*
766 *biology*. 2002;252(1):59-71.

767 39. Severin J, Lizio M, Harshbarger J, Kawaji H, Daub CO, Hayashizaki Y, Bertin N,
768 Forrest AR. Interactive visualization and analysis of large-scale sequencing
769 datasets using ZENBU. *Nature biotechnology*. 2014;32(3):217-9.

770 40. Seo HC, Edvardsen RB, Maeland AD, Bjordal M, Jensen MF, Hansen A, Flaat M,
771 Weissenbach J, Lehrach H, Wincker P, Reinhardt R. Hox cluster disintegration
772 with persistent anteroposterior order of expression in *Oikopleura dioica*. *Nature*.
773 2004;431(7004):67-71.

774 41. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of
775 occurrences of k-mers. *Bioinformatics*. 2011;27(6):764-70.

776 42. Andrews S. FastQC: a quality control tool for high throughput sequence data.
777 2010.

778 43. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome
779 assembly from long uncorrected reads. *Genome research*. 2017;27(5):737-46.

780 44. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA,
781 Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for
782 comprehensive microbial variant detection and genome assembly improvement.
783 *Plos one*. 2014;9(11):e112963.

784 45. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-
785 MEM. *arXiv preprint arXiv:1303.3997*. 2013.

786 46. Huang S, Kang M, Xu A. HaploMerger2: rebuilding both haploid sub-assemblies
787 from high-heterozygosity diploid genome assembly. *Bioinformatics*.
788 2017;33(16):2577-9.

789 47. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL.
790 Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments.
791 *Cell systems*. 2016;3(1):95-8.

792 48. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden
793 EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited
794 zoom. *Cell systems*. 2016;3(1):99-101.

795 49. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic
796 sequence comparison. *Genome research*. 2011;21(3):487-93.

797 50. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for
798 genome assemblies. *Bioinformatics*. 2013;29(8):1072-5.

799 51. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-
800 repeat transposable elements from genomic sequences. *Nucleic acids research*.
801 2010;38(22):e199.

802 52. Wenke T, Döbel T, Sörensen TR, Junghans H, Weisshaar B, Schmidt T. Targeted
803 identification of short interspersed nuclear element families shows their
804 widespread existence and extreme heterogeneity in plant genomes. *The Plant
805 Cell*. 2011;23(9):3117-28.

806 53. Smit A.F.A., Hubley R. & Green P. RepeatMasker at <http://repeatmasker.org>

807 54. Frith MC. A new repeat-masking method enables specific detection of homologous
808 sequences. *Nucleic acids research*. 2011;39(4):e23.

809 55. Olson D, Wheeler T. ULTRA: A Model Based Tool to Detect Tandem Repeats.
810 InProceedings of the 2018 ACM International Conference on Bioinformatics,
811 Computational Biology, and Health Informatics 2018 (pp. 37-46).

812 56. Schulmeister A, Schmid M, Thompson EM. Phosphorylation of the histone H3. 3
813 variant in mitosis and meiosis of the urochordate *Oikopleura dioica*. *Chromosome
814 research*. 2007;15(2):189.

815 57. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X,
816 Fan L, Raychowdhury R, Zeng Q, Chen Z. Full-length transcriptome assembly
817 from RNA-Seq data without a reference genome. *Nature biotechnology*.
818 2011;29(7):644-52.

819 58. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of
820 protein or nucleotide sequences. *Bioinformatics*. 2006;22(13):1658-9.

821 59. Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD. rnaQUAST: a
822 quality assessment tool for de novo transcriptome assemblies. *Bioinformatics*.
823 2016;32(14):2210-2.

824 60. Stanke M, Schöfmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes
825 with a generalized hidden Markov model that uses hints from external sources.
826 *BMC bioinformatics*. 2006;7(1):62.

827 61. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS,
828 Yandell M. MAKER: an easy-to-use annotation pipeline designed for emerging
829 model organism genomes. *Genome research*. 2008;18(1):188-96.

830 62. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, Maiti R,
831 Ronning CM, Rusch DB, Town CD, Salzberg SL. Improving the *Arabidopsis*
832 genome annotation using maximal transcript alignment assemblies. *Nucleic acids
833 research*. 2003;31(19):5654-66.

834 63. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson
835 M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*.
836 2013;29(1):15-21.

837 64. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR,
838 Wortman JR. Automated eukaryotic gene structure annotation using
839 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome
840 biology*. 2008;9(1):R7.

841 65. Shumate A, Salzberg S. Liftoff: an accurate gene annotation mapping tool. *bioRxiv*.
842 2020; doi:10.1101/2020.06.24.169680

843 66. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J,
844 Middendorf M, Stadler PF. MITOS: improved de novo metazoan mitochondrial
845 genome annotation. *Molecular phylogenetics and evolution*. 2013;69(2):313-9.

846 67. Gertz EM, Yu YK, Agarwala R, Schäffer AA, Altschul SF. Composition-based
847 statistics and translated nucleotide searches: improving the TBLASTN module of
848 BLAST. *BMC biology*. 2006;4(1):1-4.

849 68. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G,
850 Gilbert JG, Korf I, Lapp H, Lehväslaiho H. The Bioperl toolkit: Perl modules for the
851 life sciences. *Genome research*. 2002;12(10):1611-8.

852 69. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW.
853 RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids
854 research*. 2007;35(9):3100-8.

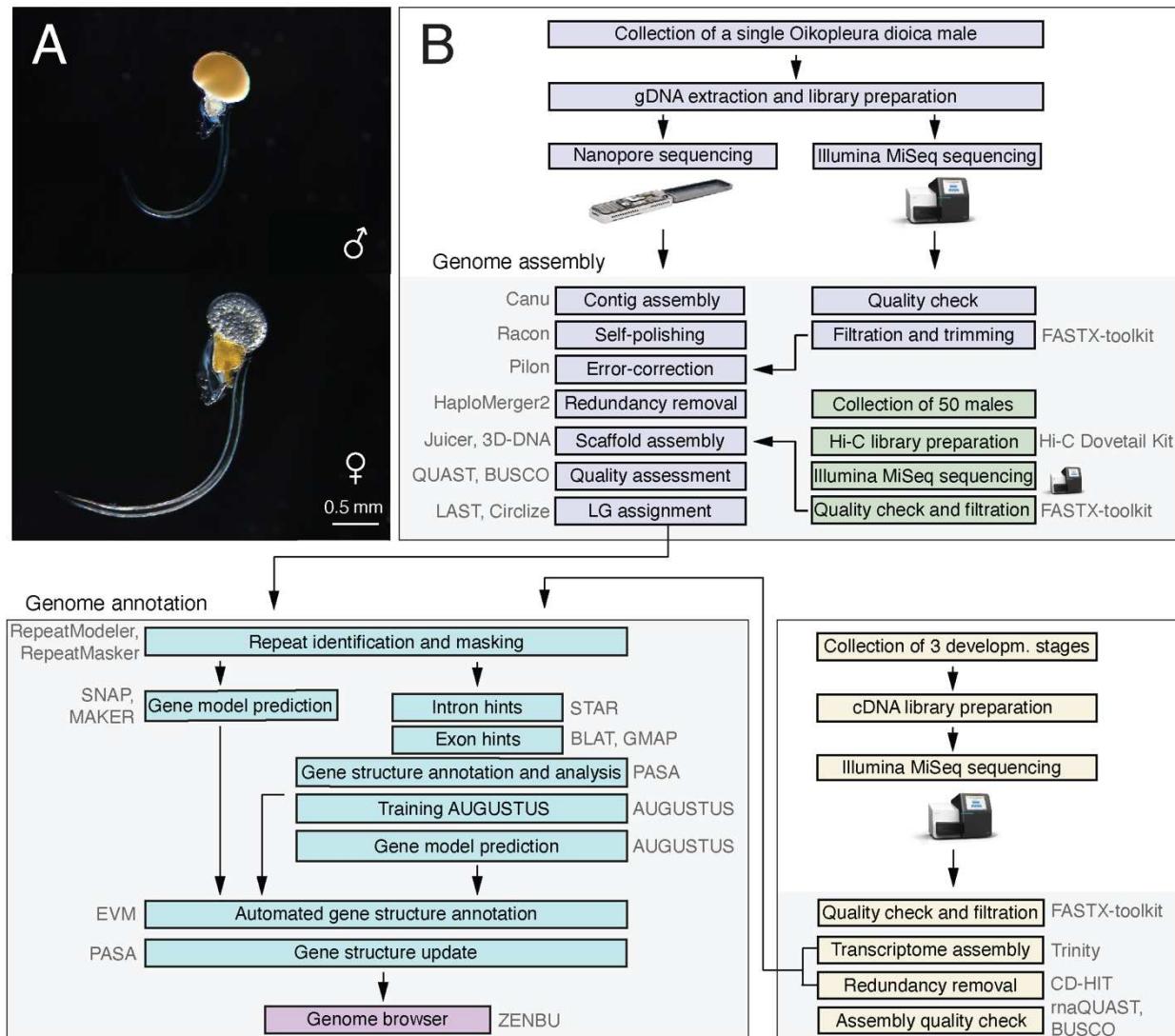
855 70. Slater GS, Birney E. Automated generation of heuristics for biological sequence
856 comparison. *BMC bioinformatics*. 2005;6(1):31.

857 71. Hamada M, Ono Y, Asai K, Frith MC. Training alignment parameters for arbitrary
858 sequencers with LAST-TRAIN. *Bioinformatics*. 2017;33(6):926-8.

859 72. Frith MC, Kawaguchi R. Split-alignment of genomes finds orthologies more
860 accurately. *Genome biology*. 2015;16(1):106.

861 **Figures**

862



863

864 **Figure 1:** (A) Life images of adult male (top) and female (bottom) *O. dioica*. (B) Genome assembly and
865 annotation workflow that was used to generate the OKI2018_I69 genome assembly.

866

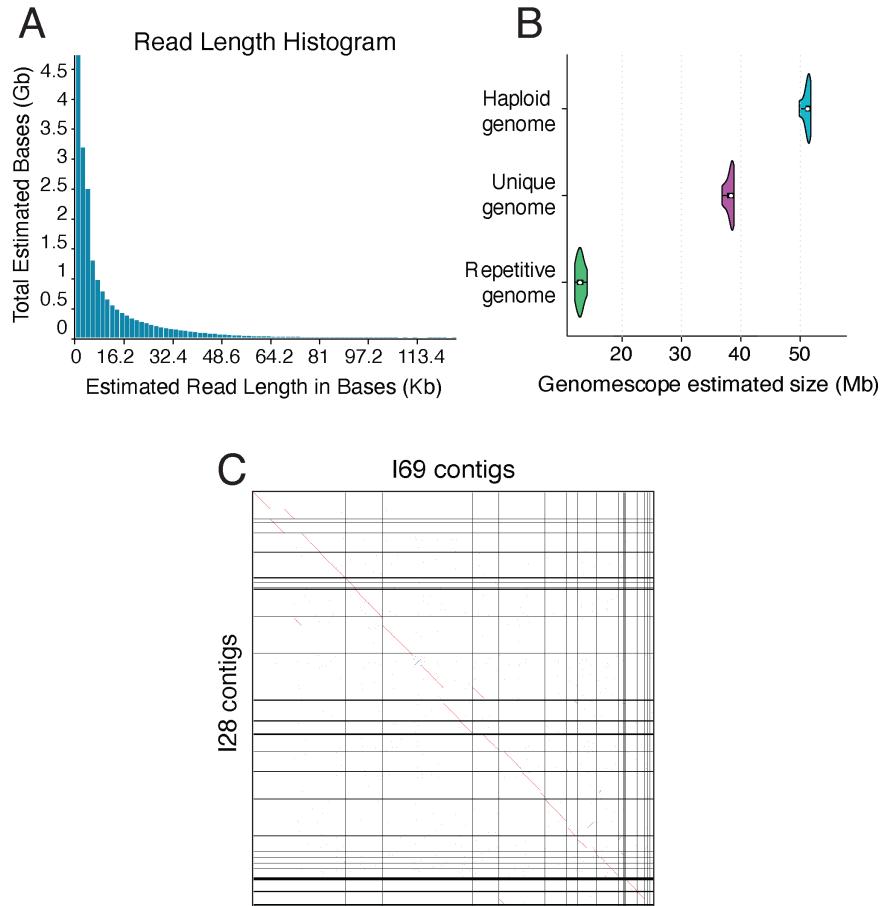
867

868

869

870

871



872

873 **Figure 2:** (A) Length distribution of raw Nanopore reads. (B) Estimated total and repetitive genome size
874 based on k -mer counting of the Illumina paired-end reads used for assembly polishing. (C) Pairwise
875 genome alignment of the contig assemblies of I69 and I28 *O. dioica* individuals.

876

877

878

879

880

881

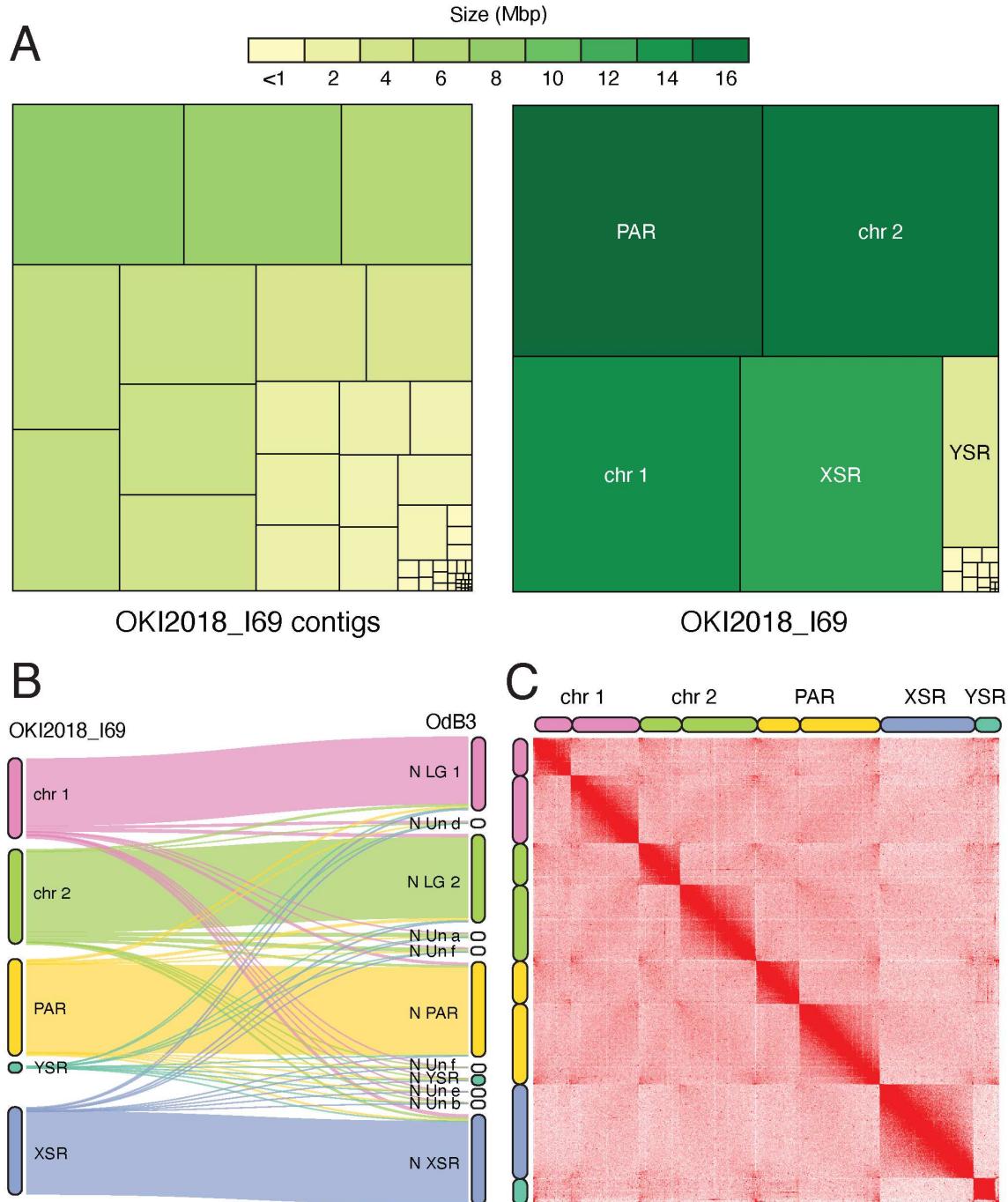
882

883

884

885

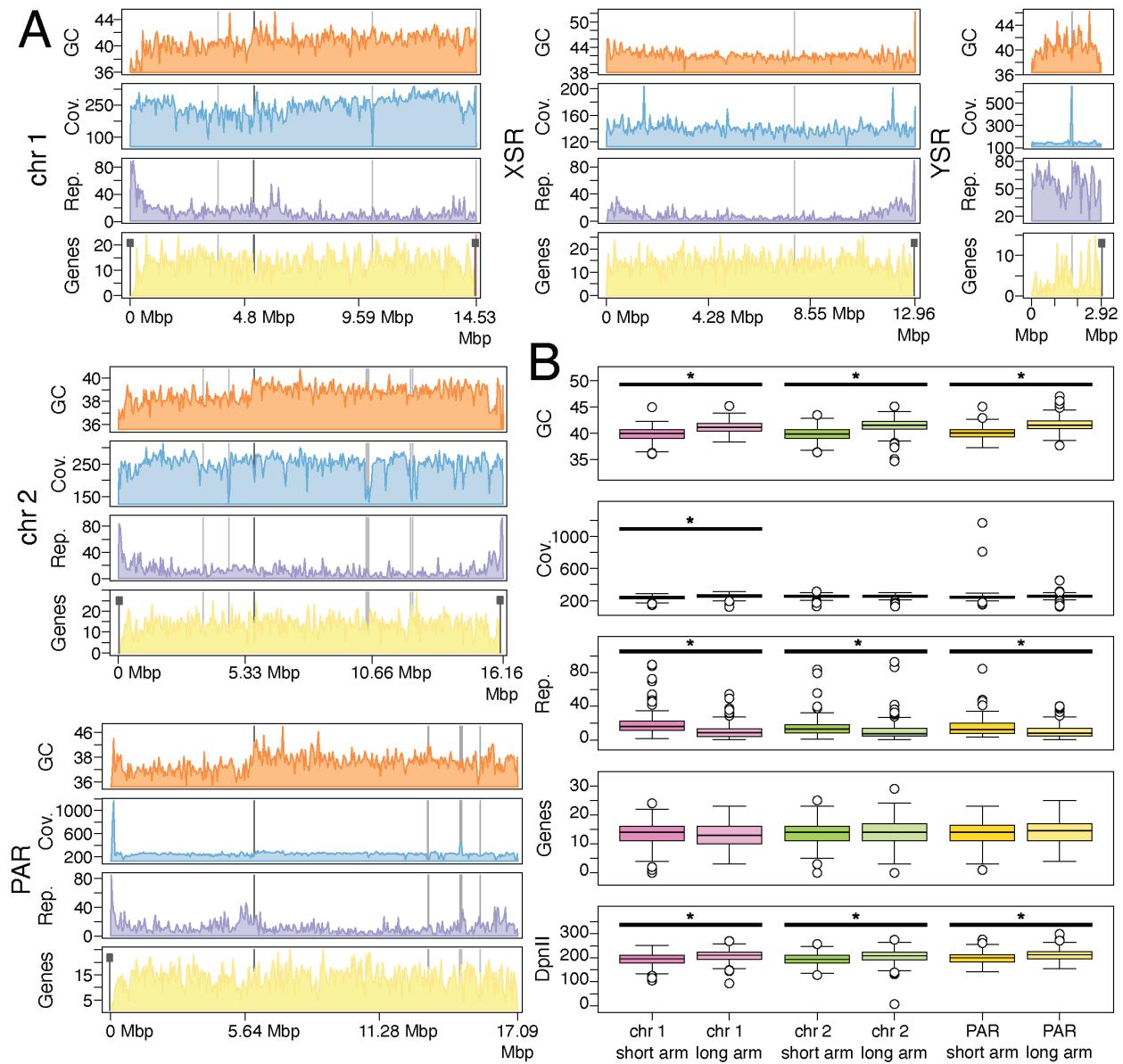
886



887

888 **Figure 3:** (A) Treemap comparison between the contig (left) and scaffold (right) assemblies of the *O. dioica*
889 genome. Each rectangle represents a contig or a scaffold in the assembly with the area proportional to its
890 length. (B) Comparison between the OKI2018_I69 (left) and OdB3 (right) linkage groups. The Sankey plot
891 shows what proportion of each chromosome in the OKI2018_I69 genome is aligned to the OdB3 linkage
892 groups. (C) Contact matrix generated by aligning Hi-C data set to the OKI2018_I69 assembly with Juicer
893 and 3D-DNA pipelines. Pixel intensity in the contact matrices indicates how often a pair of loci collocate in
894 the nucleus.

895

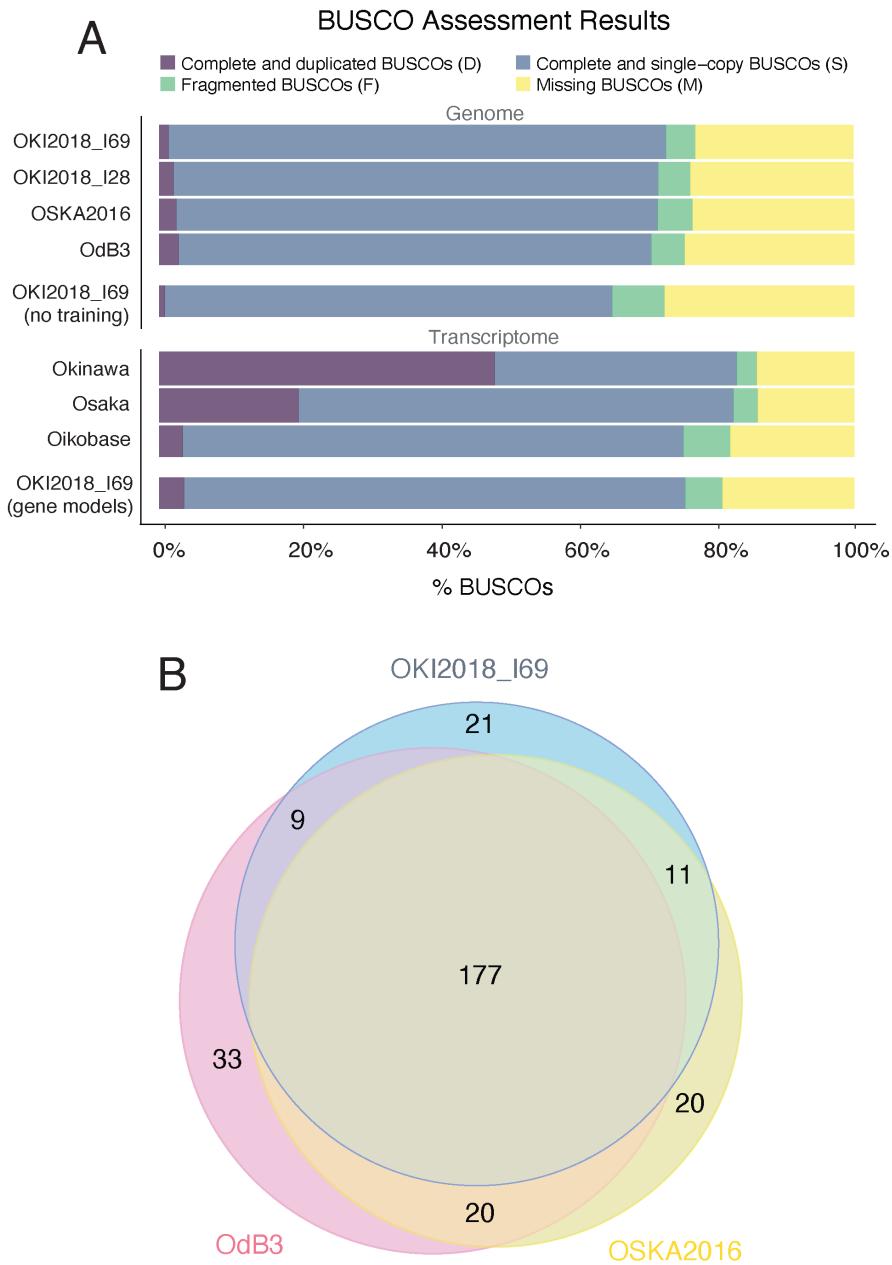


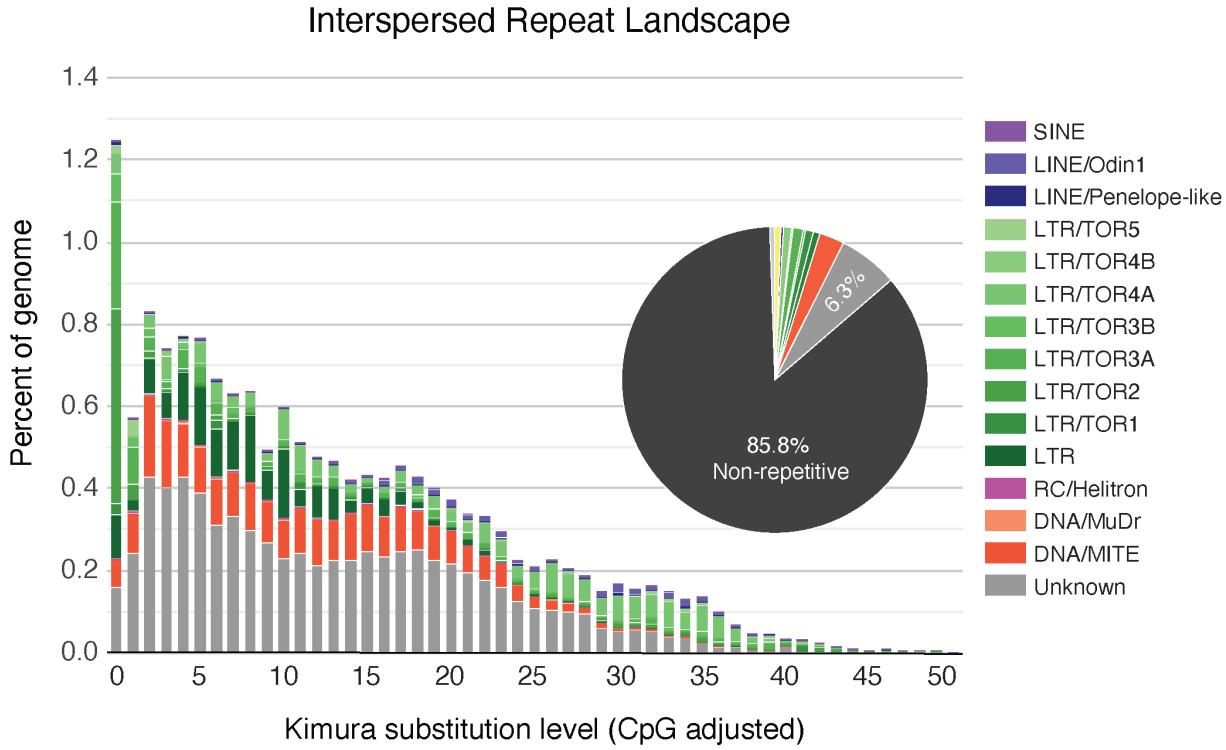
896

897 **Figure 4:** (A) Visualization of sequence properties across chromosomes in the OKI2018_169 assembly. For
 898 each chromosome, 50 Kbp windows of GC (orange), Nanopore sequence coverage (blue), the percent of
 899 nucleotides masked by RepeatMasker (purple), and the number of genes (yellow) are indicated. Differences
 900 in these sequence properties occur near predicted sites of centromeres and telomeres, as well as between
 901 the short and long arms of each non-sex-specific chromosome. Telomeres and gaps in the assembly are
 902 indicated with black and grey rectangles, respectively. B) Long and short chromosome arms exhibit
 903 significant differences sequence properties, including GC content, repetitive sequence content, and the
 904 number of restriction sites recognized by the Dpnl enzyme used to generate the Hi-C library.

905

906

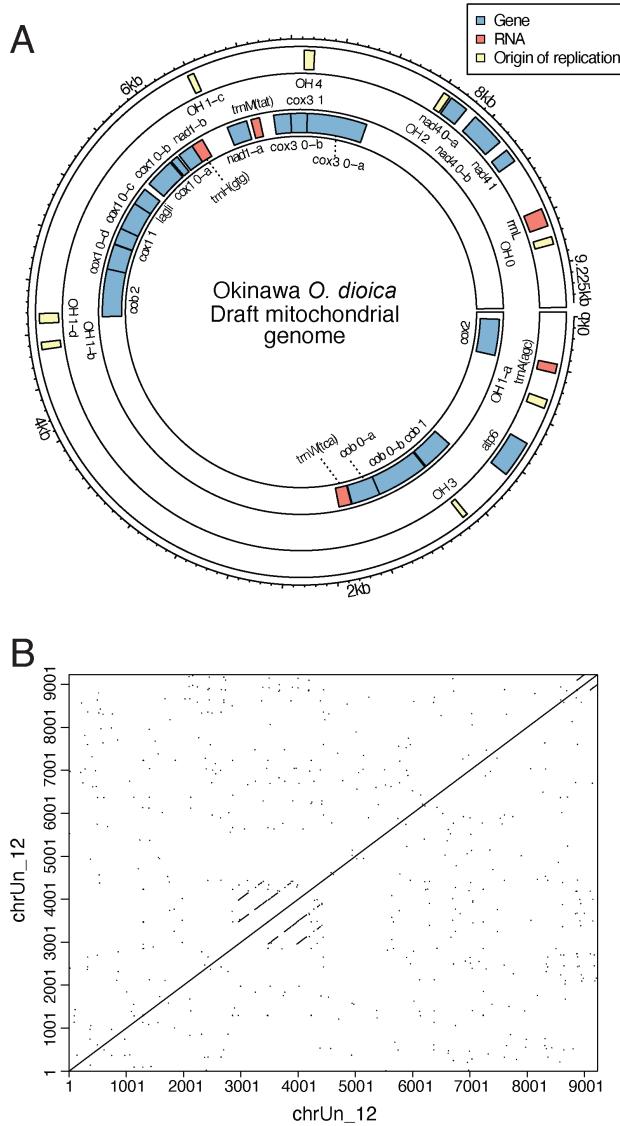


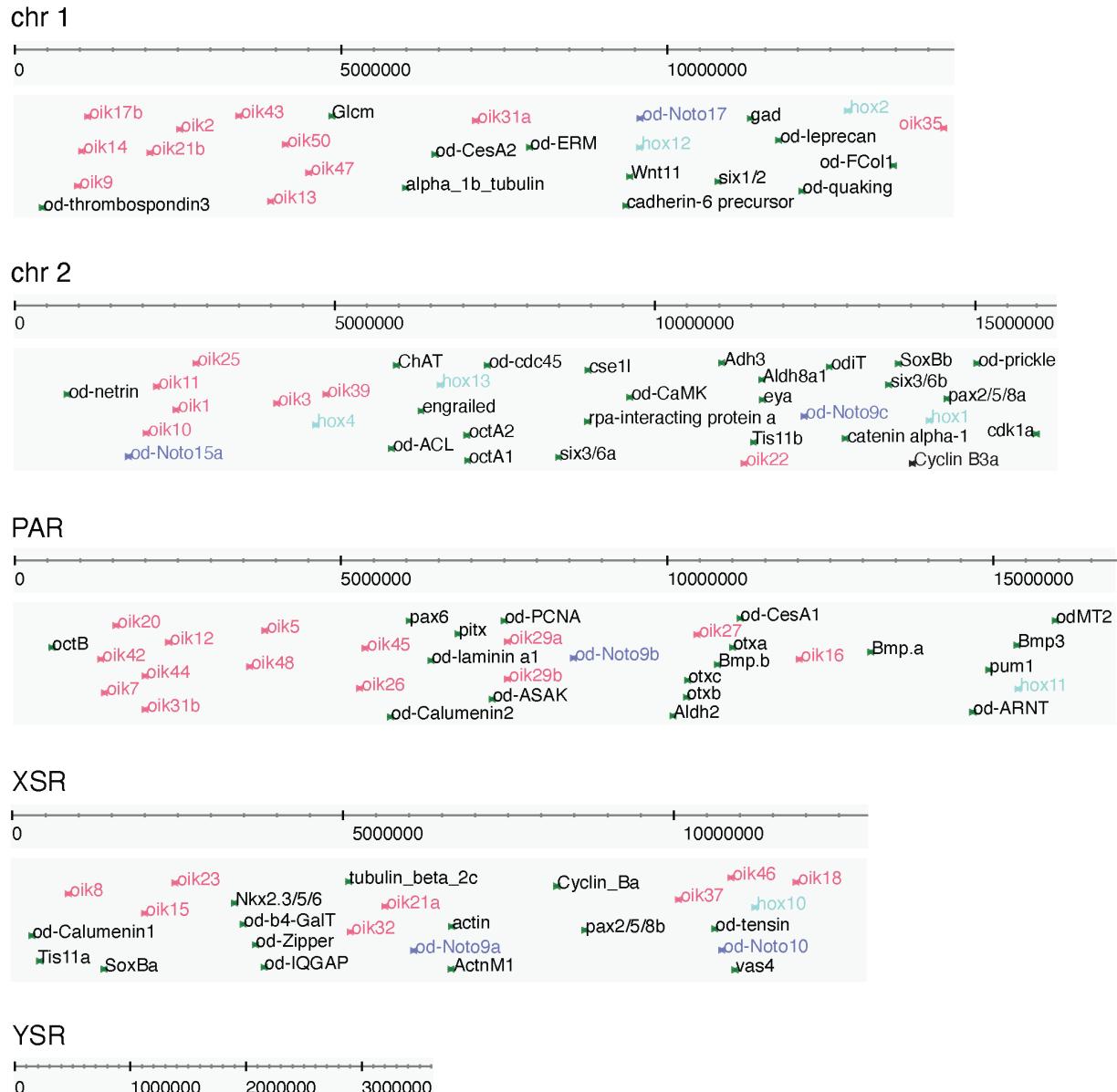


916

917 **Figure 6:** Analysis of repetitive elements. The repeat landscape and proportions of various repeat classes
 918 in the genome are indicated and color-coded according to the classes shown on the right side of the figure.
 919 The non-repetitive fraction of the genome is shown in black.

920





932

Figure 8: Genomic locations of various oikopleurid gene homologs searchable by name and PubMed identifiers in the ZENBU genome browser. Colours indicate genes from the same family.

935

936 **Supplementary materials**

937 **Supplementary Table 1:** Per-scaffold statistics

938 Supplementary Table 2: Contamination table

939 **Supplementary Table 3:** Statistics results for the analysis of sequence properties across chromosome-
940 scale scaffolds

941 **Supplementary Table 4: BUSCO scores**

942 **Supplementary Table 5:** List of missing BUSCO genes in OKI2018_I69, OdB3 and OSKA2016 genome
943 assemblies

944 **Supplementary Table 6:** Gene list uploaded to ZENBU

945

946