The whole genome sequence and mRNA transcriptome of the tropical cyclopoid copepod Apocyclops royi

Jørgensen, Tue Sparholt; Nielsen, Bolette Lykke Holm; Petersen, Bent; Browne, Patrick Denis; Hansen, Benni Winding; Hansen, Lars Hestbjerg

Published in:
G3: Genes, Genomes, Genetics

DOI:
10.1534/g3.119.400085

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Jørgensen, T. S., Nielsen, B. L. H., Petersen, B., Browne, P. D., Hansen, B. W., & Hansen, L. H. (2019). The whole genome sequence and mRNA transcriptome of the tropical cyclopoid copepod Apocyclops royi. *G3: Genes, Genomes, Genetics*, *9*(5), 1295-1302. https://doi.org/10.1534/g3.119.400085

G3 Genes | Genomes | Genetics

GENOME REPORT

The Whole Genome Sequence and mRNA Transcriptome of the Tropical Cyclopoid Copepod *Apocyclops royi*

Tue Sparholt Jørgensen,*,†,1 Bolette Lykke Holm Nielsen,* Bent Petersen,‡,§ Patrick Denis Browne,† Benni Winding Hansen,* and Lars Hestbjerg Hansen†

*Department of Science and Environment, Roskilde University, Roskilde, Denmark, 4000, †Department of Environmental Science - Environmental Microbiology and Biotechnology, Aarhus University, Roskilde, Denmark, 4000, ‡Natural History Museum of Denmark, University of Copenhagen, Denmark, 2100, and §Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia

ORCID IDs: 0000-0002-2437-7086 (T.S.J.); 0000-0002-8944-5879 (B.L.H.N.); 0000-0002-2472-8317 (B.P.); 0000-0001-8300-7758 (P.D.B.); 0000-0003-1145-561X (B.W.H.)

ABSTRACT Copepoda is one of the most ecologically important animal groups on Earth, yet very few genetic resources are available for this Subclass. Here, we present the first whole genome sequence (WGS, acc. UYDY01) and the first mRNA transcriptome assembly (TSA, Acc. GHAJ01) for the tropical cyclopoid copepod species *Apocyclops royi*. Until now, only the 18S small subunit of ribosomal RNA gene and the COI gene has been available from *A. royi*, and WGS resources was only available from one other cyclopoid copepod species. Overall, the provided resources are the 8th copepod species to have WGS resources available and the 19th copepod species with TSA information available. We analyze the length and GC content of the provided WGS scaffolds as well as the coverage and gene content of both the WGS and the TSA assembly. Finally, we place the resources within the copepod order Cyclopoida as a member of the *Apocyclops* genus. We estimate the total genome size of *A. royi* to 450 Mb, with 181 Mb assembled nonrepetitive sequence, 76 Mb assembled repeats and 193 Mb unassembled sequence. The TSA assembly consists of 29,737 genes and an additional 45,756 isoforms. In the WGS and TSA assemblies, >80% and >95% of core genes can be found, though many in fragmented versions. The provided resources will allow researchers to conduct physiological experiments on *A. royi*, and also increase the possibilities for copepod gene set analysis, as it adds substantially to the copepod datasets available.

KEYWORDS
Copepod genome assembly
Copepod transcriptome assembly
crustacean genomics
Apocyclops royi
Cyclopoida
Arthropoda

Copepods are among the most numerous animals on Earth, and the ecology, behavior, biotechnological and aquaculture potential of copepods has been scrutinized for decades. Yet very few molecular resources are available for the subclass Copepoda. *Apocyclops royi* is an omnivorous cyclopoid copepod found in estuaries, brackish-water aquaculture ponds and in freshwater areas in tropical regions (Chang and Lei 1993; Blanda *et al.* 2015; Blanda *et al.* 2017; Su *et al.* 2007). *A. royi* is a relatively small egg-carrying copepod with a prosome length of 0.5 mm (Figure 1A) (Chang and Lei 1993; Blanda *et al.* 2015). It has a life cycle of 7-8 days (Lee *et al.* 2013), and can tolerate temperatures of 15-35° (Yen Ju Pan *et al.* 2017; Blanda *et al.* 2015), and salinities of 0-35 psu (Y.-J. Pan *et al.* 2016). In a recent publication, we report the ability of *A. royi* to biosynthesize the polyunsaturated fatty acid Docosahexaenoic acid (DHA) from *alpha*-Linolenic acid (Y. J. Pan *et al.* 2018; Nielsen *et al.* 2019) which makes *A. royi* an interesting organism for copepod physiological studies.

Copepod genomes are infamously difficult to assemble (Bron *et al.* 2011; Rasch and Wyngaard 2006). This is likely caused by high repetitiveness, a low GC content of around 30% (Ross *et al.* 2013)

doi: https://doi.org/10.1534/g3.119.400085
Manuscript received February 12, 2019; accepted for publication March 8, 2019; published Early Online March 28, 2019.

Supplemental material available at Figshare: https://doi.org/10.25387/g3.7484492.
[1]Corresponding author Email: tuesparholt@gmail.com; lhha@envs.au.dk

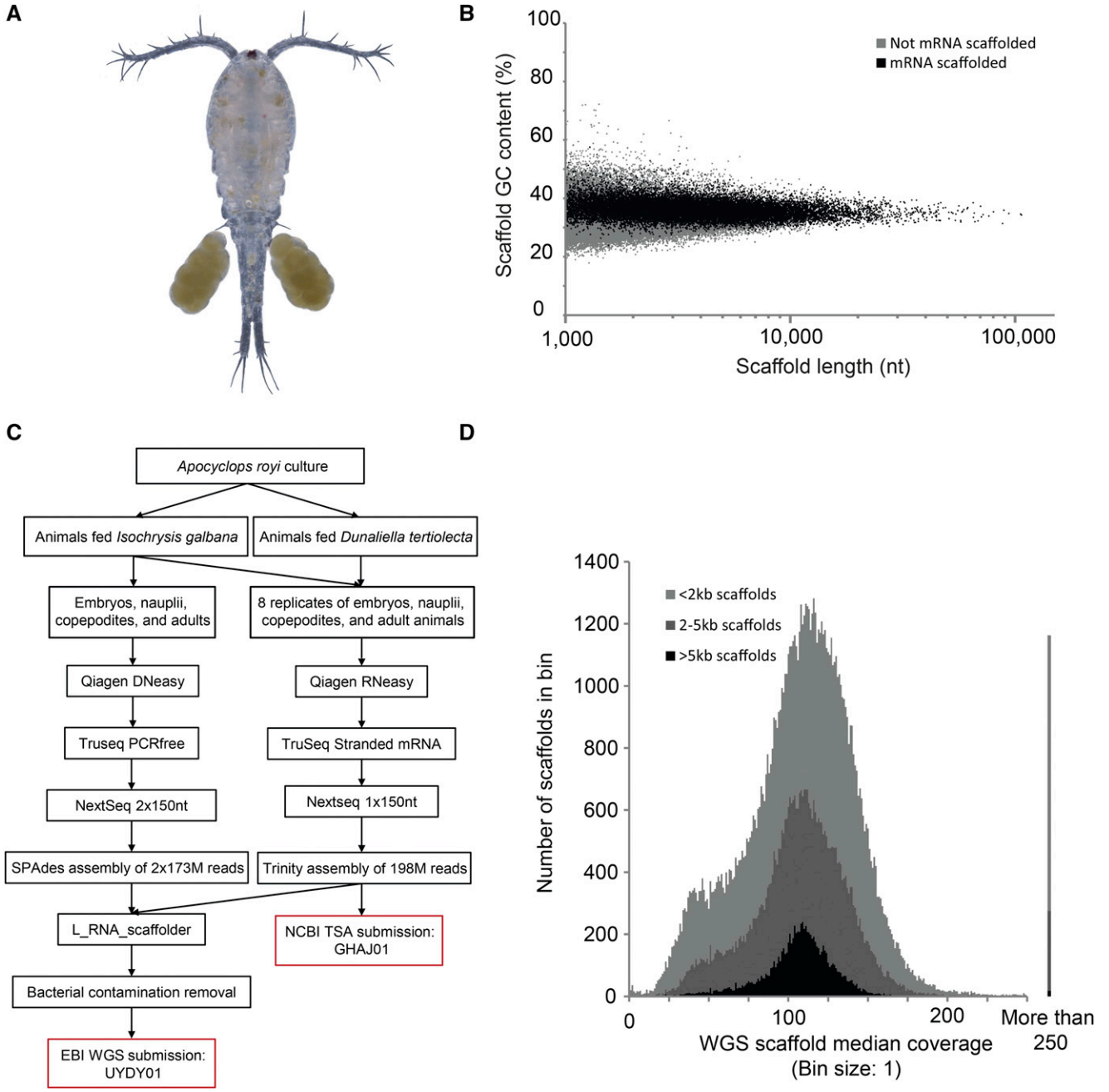

Figure 1 A, composite picture of a female *Apocyclops royi* with egg sacks from the culture used for experiments. B, WGS assembly GC% plotted against scaffold length for mRNA scaffolded sequences (black) and not mRNA scaffolded sequences (gray). Each dot represents one scaffold. C, workflow used in the present study from culture to data deposition. D, median coverage estimation for the WGS assembly in the three scaffold length subsets <2kb, 2-5kb, and >5kb. For each, a maximum number of scaffolds are seen in bins with coverage ca 110. Less than 1% of scaffolds have a coverage higher than 250 (illustrated on the right hand side of the plot). We chose to use median values to minimize the impact of highly covered regions, which are regularly seen in WGS datasets and which are likely owing to repetitive sequence.

and very variable genome sizes (Jørgensen *et al.* 2019), which means that it is difficult to assess the costs before undertaking whole genome sequencing (WGS). This is compounded by the often small physical size of the animals, which makes it necessary to use a collection of animals rather than a single individual for nucleic acid purification, adding to the complexity of genome assembly. Modern genome assembly pipelines and data generation workflows are optimized for mammalian genome assembly, and any deviation from mammalian like genomes are likely to result in lower quality assembly. Crucially, the total genome size often differs substantially from the assembly size, as repetitive DNA is collapsed or remains unassembled. Transcriptome assemblies, however, are significantly easier to obtain, as many of the clade-specific limitations of copepod WGS are overcome by focusing on mRNA. Here, the

highly repetitive regions are not transcribed or are removed post-transcriptionally and the assembly process is simpler as the remaining repetitive regions are dealt with simplistically (Grabherr *et al.* 2011). A recent paper presents a good example of a high quality transcriptome from a copepod where WGS information was not available (Lenz *et al.* 2014). A lot of information is however not captured by a transcriptome. For example, intron sizes and repeat structure can be derived from a genome assembly, but not from a transcriptome, which also fails to capture genes which are not highly expressed or genes which are expressed only in certain tissues.

For evolutionary analysis relying on existing DNA databases, it is imperative to have a diverse range of genomic information available. As of now, only one cyclopoid copepod genome is available, namely the high quality WGS assembly of *Oithona nana* (Madoui *et al.* 2017). Further, only eight copepod species have available WGS information, and only 19 copepod species have available TSA information, including the *A. royi* datasets. With the presented *A. royi* genome and transcriptome, we expand the possibilities for studies centered on *A. royi* physiology and improve the possibility for large scale phylogenetic and evolutionary studies. Further, our high-quality short read resources may prove pivotal in error correcting future genome projects which will utilize error prone single molecule DNA and RNA sequencing.

MATERIALS AND METHODS

Organism origin and derivation: An overview of the experimental and bioinformatical workflow of the genome assembly and transcriptome assembly can be seen in Figure 1C, and has been used first in a recent publication on the genome project on the calanoid copepod *Acartia tonsa* Dana (Jørgensen *et al.* 2019). Animal husbandry, sampling, RNA purification, RNA sequencing strategy and initial RNA data processing is also described in a companion paper where we used the mRNA dataset presented here to analyze the fatty acid metabolism genes and differential expression based on feeding regime (Nielsen *et al.* 2019). Briefly, an *Apocyclops royi* animal culture originating from Tungkang Biotechnology Research Center in Taiwan was split in two which were kept in 100 L tanks on separate microalga feeding regimes. One was fed *Isochrysis galbana* and the other was fed *Dunaliella tertiolecta*. The cultures were kept in the dark at 25° with aeration in 0.2um filtered seawater mixed to 20 psu with deionized water.

Sampling: Animals from each feeding regime were sampled as described in (Nielsen *et al.* 2019). Briefly, animals were starved for 2 h to empty their guts and a 53 μm filter was used to separate all the life stages (nauplii, copepodites and adults) from the sea water. Visual inspection showed that all life stages were present in the samples used for analysis. Four analytical replicates were made for each feeding regime, each consisting of hundreds to thousands of individuals. Animals were flushed with 0.2um filtered seawater (32ppt) until visual inspection showed very little particular contaminating matter. The remaining seawater was aspirated with a homemade small tip pasteur pipettor to ensure that animals were not removed during this step. A volume of 200 μl of RNAlater was added to the replicates of animals fed *I. galbana* and 500 μl of RNAlater was added to the replicates of animals fed *D. tertiolecta* to ensure a factor of at least 1:10 of animals in RNAlater. Samples were kept in a fridge for 24 H and frozen until use. These samples were used for both RNA and DNA extractions.

Sequencing methods and preparation details

Nucleic acid extraction: As described in (Nielsen *et al.* 2019), RNA was extracted with RNeasy (Qiagen, Venlo, Nederlands) according to protocol. Before extraction, all RNAlater was removed and the animal tissue was homogenized with a 1.5 mL RNase-Free Pellet Pestle (Kimble Chase, Vineland, New Jersey, USA) mounted on a Kontes Pellet Pestle motor (Kimble Chase, Vineland, New Jersey, USA) for 1 min on ice in 20 μl buffer RTL, before adding the remaining 320 μl Buffer RTL.

DNA was extracted from replicate 2 of animals fed *I. galbana* using the DNeasy blood and tissue kit from Qiagen according to protocol. Briefly, tissue from thousands of animals were homogenized manually with a 1.5 ml RNase-Free Pellet Pestle (Kimble Chase, Vineland, New Jersey, USA) to prevent unnecessary DNA shearing. The ground tissue was incubated for four hours at 56° in lysis buffer with Proteinase K and RNase A according to protocol, vortexing every 15-30 min. A Qubit 3.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) was used to determine the DNA and RNA concentrations.

Sequencing library construction: The RNA sequencing library strategy is described in (Nielsen *et al.* 2019). Briefly, an mRNA sequencing library was produced for each of the eight replicates with the Truseq stranded mRNA kit (Illumina, San Diego, California, United States) and SuperscriptII reverse transcriptase (Thermo Fisher Scientific, Waltham, MA, USA). 1 μg total RNA was used for each of the eight mRNA library preparations. DNase was not used to avoid breakdown of long transcripts and because the stranded protocol minimizes the influence of DNA contamination. The efficiency of the protocol was assessed using the directionality of reads. A PCR-free DNA sequencing library was produced using the Illumina TruSeq PCR-Free kit according to protocol. DNA was sheared in a Covaris E210 with the following settings: Intensity: 4, Duty cycle 10%, Cycles per burst: 200, Treatment time: 70 s intended to produce fragments of 350nt.

The library cluster forming molarity of all samples was evaluated using the KAPA qPCR system (Roche, Basel, Switzerland) and samples were run on a Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA) to evaluate the fragment length.

Sequencing: The eight mRNA libraries were pooled equimolarly and run on a single Illumina Nextseq 1x150nt mid output flowcell as described in (Nielsen *et al.* 2019). The PCR-free DNA library was run on a single, full Illumina Nextseq 2x150nt mid output flowcell. Data were demultiplexed using the blc2fastq tool provided by Illumina and all TruSeq indices.

Data processing methods: The data processing was performed similarly to a previous study on the genome of the calanoid copepod *A. tonsa* (Jørgensen *et al.* 2019). Initial data handling and basic statistics were carried out using Biopieces (Hansen, MA, maasha.github.io/biopieces/, unpublished). Raw illumina reads were trimmed using Adapterremoval v. 2.0 (Schubert, Lindgreen, and Orlando 2016) with the following parameters "–trimns–trimqualities" as well as standard parameters for paired end data for the PCR-free WGS reads. Trinity v. 2.5.1 (Grabherr *et al.* 2011) was used to assemble pooled mRNA reads from all eight sequenced replicates with the following parameters: "–SS_lib_type R–trimmomatic–single". Thus the transcriptomic data were run through adapter trimming twice. Transcripts shorter than 500nt were discarded and PhiX contigs removed by BLAST (Altschul *et al.* 1997) in CLCgenomics 11.0 (Qiagen, Venlo, Nederlands). The PreQC system

Table 1 Please provide Title

	WGS, Genome	TSA, Transcriptome
Assembly N50	3,257 nt	1,682 nt
Assembly size	257,5 Mb	100,7 Mb
Number of contigs	143,521	75,477
Number of scaffolds	97,072	—
Number of mRNA scaffolds	18,420	—
mRNA scaffolds N50	7,037 nt	—
Shortest sequence	1,001 nt	500 nt
Assembly GC%	34%	41%
Estimated genome size	450.6 Mb	—
Identified repetitive sequence in assembly	76.8 Mb	—
Sequencing technology	Illumina Nextseq500	Illumina Nextseq500
Assembler	SPAdes3.11	Trinity2.5.1
Scaffolding	L_RNA_scaffolder	—
Date of deposition	05/12/2018	13/11/2018
Accession	UYDY01	GHAJ01
SRA	ERR2811089	ERR2811715, ERR2811728-ERR2811734
Number of PE reads	173,365,491	—
Number of SE reads	—	203,548,224

from the SGA pipeline was used to estimate the total genome size based on read k-mer spectra (Simpson and Durbin 2012). SPAdes v. 3.11 (Bankevich *et al.* 2012) with the auto-selected k-mer sizes 21, 33, 55, and 77 was used to assemble the PCR-free WGS reads on the Computerome supercomputer on a 1TB RAM node. The SPAdes log can be found in Supplementary Material 1. The WGS assembly was scaffolded using the mRNA TSA assembly and the L_RNA_scaffolder program (Xue *et al.* 2013).

Contamination removal: Because whole animals were used for the WGS data generation, it is expected that bacterial symbionts also contributed DNA to the sequencing libraries. In order to remove any sequence of bacterial origin from the genome assembly, we first masked all scaffolds using Repeatmodeler and Repeatmasker (v. 4.0.7) (Smit and Hubley, n.d.; Smit, Hubley, and Green, n.d.). Repeats from RepeatModeler and the Arthropoda and ancestral (shared) repeats from repbase v. 22.05 (downloaded 2017-06-02) were used to mask scaffolds. The masked scaffolds were searched against the RefSeq database of representative prokaryotes (downloaded 2017-03-23) using the build-in BLAST in CLCgenomics 11.0. Scaffolds with BLAST hits longer than 500nt without mRNA proof were removed from the assembly. The output from a second round of Repeatmasker run on the assembly without contamination was used to estimate the assembled repetitive and non-repetitive fractions of the WGS assembly.

Additional analyses: The sequencing depth was estimated by mapping all reads on assemblies using Bowtie2 (Langmead and Salzberg 2012) (v. 2.3.4, switches:--local--no-unal) and extracting the median coverage of each transcript (TSA assembly) or scaffold (WGS assembly) using Samtools (Li *et al.* 2009) (samtools view |samtools sort -|samtools depth -aa -) and a custom python script which can be found in supplementary material 2. Both the mRNA TSA assembly and the WGS genome assembly were evaluated using the BUSCO Universal Single-Copy Ortholog v.2 (Simão *et al.* 2015) and the Arthropoda ODB9 dataset. In order to obtain a 18S rRNA gene sequence, paired reads from the WGS dataset was mapped on the partial Cyclopoida genes from the PopSet 442571920 (Wyngaard *et al.* 2011) using Bowtie2. The read pairs where at least one read mapped were then extracted and assembled using SPAdes v. 3.13 and the resulting 18S rRNA gene sequence was aligned to the reference sequences and trimmed using CLCgenomics 11.0. A neighbor-joining phylogram was constructed in CLC genomics 11.0 using 1000 bootstraps.

Data availablÌlity statement

All raw data (Acc. ERR2811089, ERR2811715, ERR2811728-ERR2811734), the TSA assembly (Acc. GHAJ01), and the WGS assembly (Acc. UYDY01) are available in the ENA/NCBI system under project accession number PRJEB28764. Supplemental material available at Figshare: https://doi.org/10.25387/g3.7484492.

RESULTS AND DISCUSSION

After quality and adapter trimming, the sequencing yielded 173,365,491 PCR-free WGS clusters (346,730,982 reads) and 203,548,224 mRNA derived SE reads constituting 52 gigabases (Gb) and 31 Gb of data, respectively (Table 1). In total, 99.9% and 97.1% of reads was left after quality and adapter trimming and filtering, respectively.

The TSA assembly yielded 100.7 Mb in 29,730 genes and additionally 45,747 alternative isoforms giving a total of 75,477 transcripts. The WGS assembly yielded 143,521 contigs in 97,072 scaffolds comprising a total length of 257.5 Mb, while 83.6% of sequencing reads mapped back to the assembly (data not shown). The size of the assembly is similar to other copepod WGS datasets, but three times larger than the *O. nana* assembly, which is the only other cyclopoid copepod WGS assembly available.

After bacterial contamination removal, the WGS assembly consists of scaffolds up to 116 Kb in length, with an average GC% of 33.5% (Figure 1B, Table 1), both of which are similar to other available copepod WGS assemblies, such as the *Acartia tonsa* resource (Jørgensen *et al.* 2019). Most scaffolds above 5kb are scaffolded with mRNA (Figure 1B, black), suggesting that the intron length is greater than the insert distance of the WGS sequencing library. The uniformity of the length and GC% in Figure 1B suggests that most contaminants are not present in the assembly, as bacteria and other contaminants would likely have a different pattern of distribution of scaffold length and GC%. For example, we removed several contigs in the size range 100 kb to 1 Mb, all with a GC content between 56% and 58% and highly similar to known bacterial sequences (Data now shown). In order to estimate the

genome size of *A. royi* including the unassembled and repetitive fraction, we used the preQC program which has previously been used for copepod genome size estimation (Jørgensen *et al.* 2019). The result shows that the expected complete genome size of *A. royi* is 450 Mb (supplementary material 3 preQC report). Of this, 181 Mb are assembled nonrepetitive sequence, 76 Mb are assembled repeats and 193Mb are unassembled sequence (Table 1, Repeatmodeler output can be found in Supplementary material 4). Much of the unassembled sequence can presumably be found in scaffolds smaller than the 1kb cutoff, though repeats also would be collapsed in these scaffolds. In a recent publication on the *Acartia tonsa* WGS assembly, the genome sizes of all copepod WGS projects was estimated and in all cases showed that less than half of the expected genome size was included in the WGS assembly (Jørgensen *et al.* 2019). The difference between the assembled and the actual size of the *A. royi* genome is thus expected, similar to the differences in other species, and hypothesized to be largely caused by unassembled repetitive/non-coding regions or collapsed scaffolds (Jørgensen *et al.* 2019). For example, if a repeat of 500 nt is found 1.001 times scattered throughout the genome, the sequence is unlikely to show up more than once in the assembly, which means that the assembly size is 500.000 nt smaller than the template genome. This repeat scaffold would then have 1.000 times higher coverage than the non-repetitive fraction of the genome assembly.

In Figure 1D, a histogram of the median WGS scaffold coverage (binsize 1) between 1 and 250 show that the largest amount of scaffolds in each of the three scaffold length fractions have a coverage of ca 110 (Supplementary Material 5). This result fits the simplistic coverage estimation: 52Gb of reads should give a coverage of ca 115 on a 450Mb genome. We chose to use median rather than mean values to minimize the impact of scaffold regions with extremely high coverage, which are often seen in copepod assemblies and potentially are the result of assembled repetitive sequence. In the smaller scaffold size fractions <2 kb and 2-5 kb, a distinct shoulder is observed at coverage ca 35. In Figure 1D, scaffold bins with a coverage between 0 and 250 are shown, but many scaffolds had a higher coverage than 250. These were collected in a separate bin (>250) which is displayed on the right hand side of Figure 1D, and likely constitute many of the repeated regions in the genome. In total, only 1.2% of scaffolds have coverages higher than 250. It is generally recommended to produce WGS assemblies from datasets with coverage of ca. 100, which the results in Figure 1D confirm was achieved. By mapping the mRNA derived reads to the transcripts of the TSA dataset, we similarly produced an overview of the median coverage of transcripts (Figure 2). Importantly, the coverage in transcriptomes are not similar to those in WGS assemblies in that differential expression of genes means that a uniform coverage is not expected. As a result of this, the range of transcript median coverage bins seen in Figure 2 had to accommodate a median coverage distribution from near-zero to more than 4,000,000 though >99% of transcripts had a median coverage of less than 1000 (Suplementary Material 5 mapping table).

In order to estimate the gene completeness of the WGS assembly, we used the BUSCO system of near-universal single-copy orthologous genes. We found 51% complete and single copy genes, 1% complete duplicated genes, 29% fragmented and 19% missing genes (Figure 3). These statistics are similar to some other copepod genome assemblies in the NCBI WGS database, and means that the large majority of conserved genes can be found in the assembly, though many are incomplete (Figure 3). The many fragmented genes could be explained by intron sizes up to 70 kb as recently reported in a

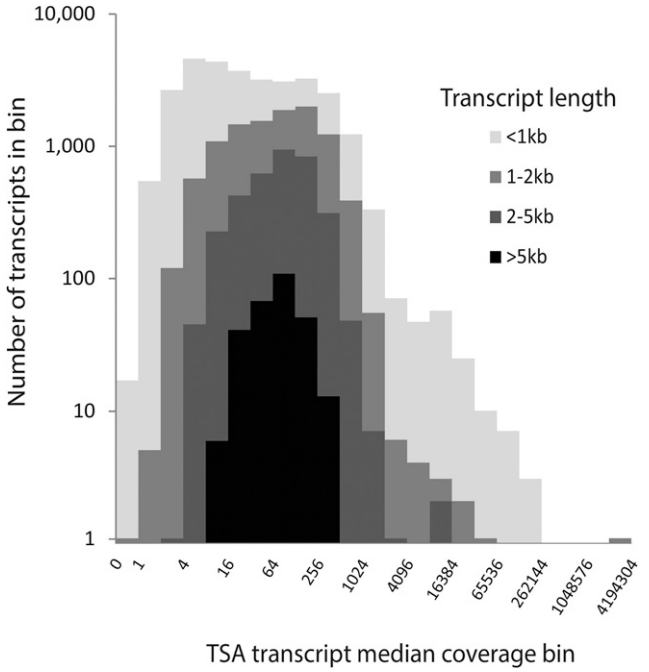

Figure 2 Median coverage of TSA isoform 1 of transcripts. While most transcripts have a median coverage of less than 200, several have median coverages of more than 100.000. The four length subset <1kb, 1-2kb, 2-5kb, and >5kb from the TSA are plotted, each with fewer transcripts, but each also with a similar distribution pattern. Because the coverage of TSA transcripts is related to the expression of genes, large variations of the coverage of transcripts are both expected and observed.

crustacean (Kao *et al.* 2016). For several practical applications, though, it is sufficient to have a gene fragment available to *e.g.*, design primers for qPCR as long as it can be annotated unequivocally. One example of the usefulness of incomplete database genes can be found in recent study by the same authors as this genome report, where fatty acid desaturase genes were found in fragmented versions and subsequently reconstructed to complete genes. The expression of the genes were found to be upregulated by starvation of polyunsaturated fatty acids in microalga feed (Nielsen *et al.* 2019). For the *A. royi* TSA dataset, 706 BUSCO genes are complete (66%), while another 311 BUSCO genes are fragmented (29%) and 49 missing (5%) (Figure 3). Because isoforms would show up as 'duplicate genes' in the BUSCO analysis, this category is not as problematic as it is for WGS resources. For comparison, we have included the BUSCO scores of all existing copepod WGS assemblies (*A. tonsa*, acc:OETC01, *Eurytemora affinis*, acc:AZAI02, *Caligus rogercresseyi*, acc:LBBV01, *Lepeophtheirus salmonis*, acc:LBBX01, *Tigriopus californicus*, not in NCBI databases, and *Tigriopus kingsejongensis*, not in NCBI databases) as well as the BUSCO scores of the aquatic arthropod species *Semibalanus balanoida* (Acorn Barnacle, acc:PHFM01), *Triops cancriformis* (acc:BAYF01), *Daphnea pulex* (water flea, acc:ACJG01), *Hyalella azteca* (acc:JQDR02), and *Limulus polyphemus* (Atlantic horseshoe crab, acc:AZTN01). The range of complete BUSCO genes from 21% (Acorn barnacle, *S. balenoides*) to 96% (*T. cancriformis*) show not only the status of genome sequencing of aquatic arthropods, but also the difficulty of genome assembly of non-model species. The complete BUSCO reports for the *A. royi* WGS and TSA assemblies are available in Supplementary Material

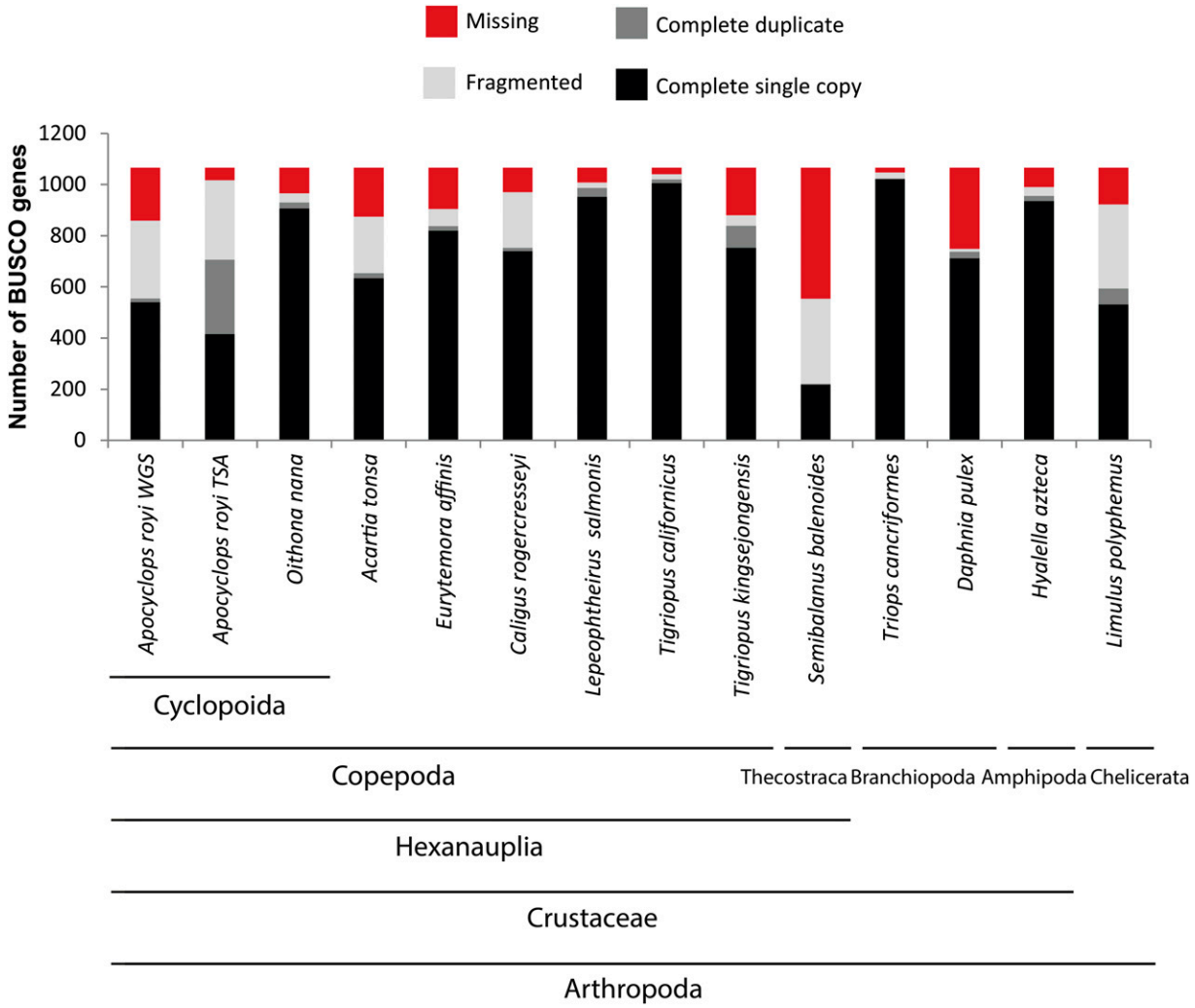

Figure 3 Universal single copy ortholog gene (BUSCO) scores for the the provided WGS and TSA assemblies of *A. royi* (UYDY01 and GHAJ01), the WGS assemblies of all available copepods (*A. tonsa* (OETC01), *E. affinis* (AZAI02), *C. rogercresseyi* (LBBV01), *L. salmonis* (LBBX01), *T. californicus* (Not in NCBI databases), and *T. kingsejongensis* (Not in NCBI databases) and WGS resources from the crustaceans *S. balanoides* (acorn barnacle, PHFM01), *T. cancriformis* (BAYF01), *D. pulex* (water flea, FLTH01), *H. Azteca* (JQDR02), and *L. polyphemus* (Atlantic horseshoe crab, AZTN01) for comparison. For the *A. royi* WGS assembly, very few genes are duplicated (2%), while only just over 50% are complete and single copy. Another 29% are fragmented while 19% of genes are missing in the WGS assembly. For the TSA dataset, 66% of genes are complete, and another 29% fragmented while only 5% are missing. For the TSA data, the categories complete single copy and complete duplicate are not as important as for the WGS dataset. Importantly, in the *A. royi* WGS and TSA datasets, only minimal fractions of the core BUSCO genes are not found in complete or fragmented versions. This will allow using the information, though some analyses might be impaired by the genes being fragmented.

6 and 7. Almost all mitochondrial genes can be found on scaffold_16888 where only the ND4L gene and the small subunit of the ribosomal RNA gene are missing: the remaining 13 genes are all present as well as all 22 tRNAs, as determined by MITOS2 (data not shown) (Bernt *et al.* 2013). In order to phylogenetically place the presented *A. royi* WGS within the order Cyclopoida, we aligned the identified 18S rRNA gene sequence to the partial 18S rRNA gene sequences from a publication on the family level phylogeny of cyclopoid copepods (Wyngaard *et al.* 2011). The nucleotide sequence of the 18S rRNA gene can be found in Supplementary Material 8. The identified *A. royi* WGS 18S rRNA gene sequence shared 598nt out of the 600nt fragment with a database sequence registred as *Apocyclops royi* (acc.: HQ008747.1, data not shown). We then created a neighbor-joining tree and found that the sequences registred as *Apocyclops* and the resources provided here form a clade with high support (bootstrap values: 93% and 100%, Figure 4). In general, readers are referred to (Wyngaard *et al.* 2011) for a thorough phylogeny of cyclopoids as many branchings in Figure 4 have little support. It does, however, thoroughly place the presented WGS assembly as *Apocyclops royi*.

In conclusion, we here present the WGS assembly (Acc. UYDY01) and an mRNA transcriptome assembly (Acc. GHAJ01) from the tropical cyclopoid copepod *Apocyclops royi*, along with the raw data used to produce them. We have shown that the provided datasets are sequenced to a sufficient depth, that any contamination in the raw reads has been removed from the WGS assembly, and that the phylogenetic placement within Cyclopoida matches our expectation for *Apocyclops royi*. Further, we have documented the completeness of core genes in

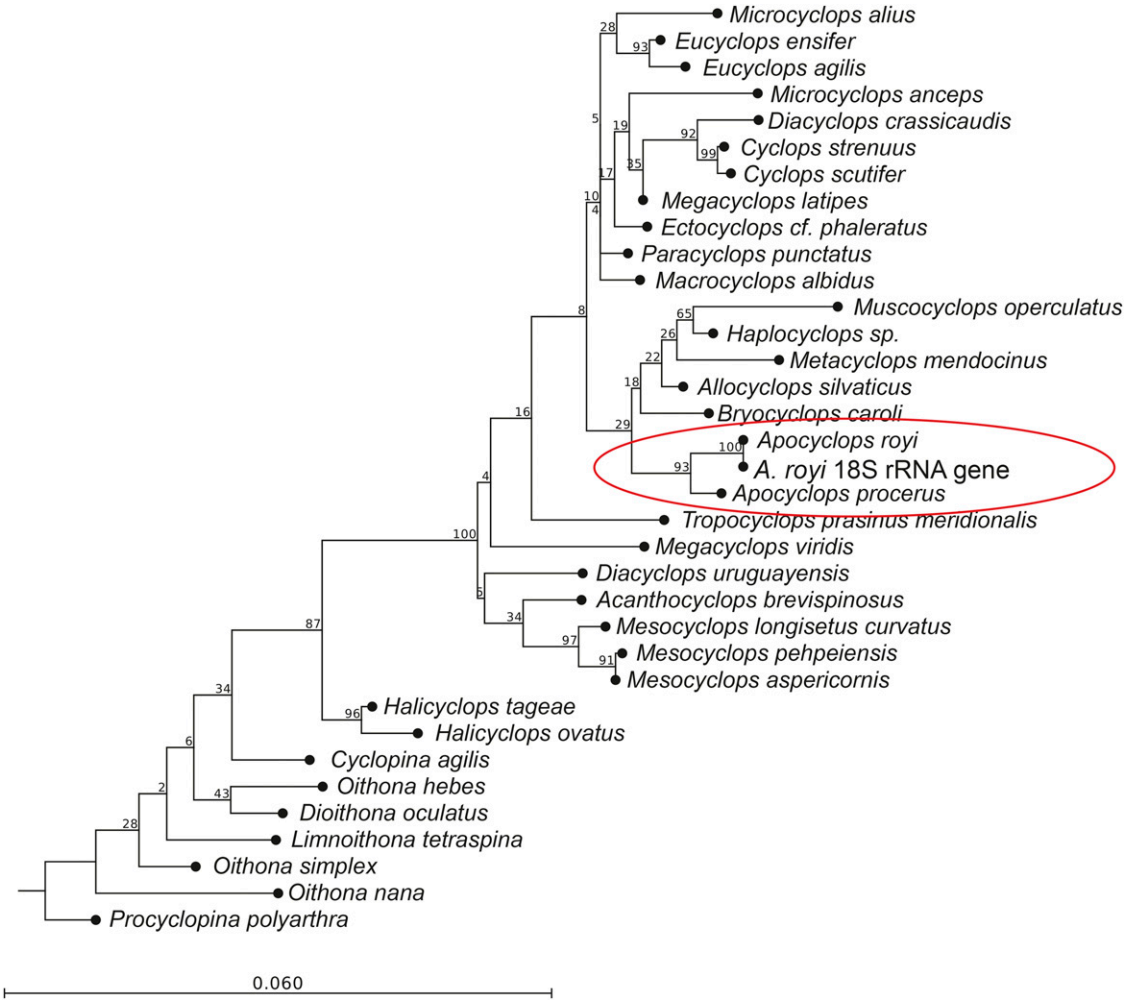

Figure 4 Neighbor-joining tree with the cyclopoid 18S rRNA gene dataset from Wyngaard *et al*. (2011) and the identified 18S rRNA gene sequence from the presented WGS. While many branches have very low support, the sequences from the genus *Apocyclops* group together with high support (Bootstrap values of 93% and 100%). Further, the *A. royi* 18S rRNA gene sequence from the presented dataset shares 598 nt out of 600 nt with the only *A. royi* sequence from the PopSet database used. This confirms the placement of the presented datasets as near identical to *A. royi*.

both the TSA and WGS dataset and found 95% and 80% of core genes, though many in fragmented versions.

ACKNOWLEDGMENTS

We thank Professor Sami Souissi (Université de Lille 1, Laboratoire d'Océanologie et de Géosciences) for sharing the *Apocyclops royi* culture on which this manuscript relies and the researchers at the Tungkang Biotechnology Research Center in Taiwan who originally isolated and cultured the strain. Further, We thank Marlene Danner Dalgaard (ORCID ID 0000-0002-4036-6408) for excellent technical support in fragmenting the DNA for WGS sequencing. This work was supported by the Villum Foundation; Project AMPHICOP No. 8960.

LITERATURE CITED

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin *et al.*, 2012 SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19: 455–477. https://doi.org/10.1089/cmb.2012.0021

Bernt, Matthias, Alexander Donath, Frank Jühling, Fabian Externbrink, Catherine Florentz, Guido Fritzsch, Joern Pütz, Martin Middendorf, and Peter F. Stadler. 2013. "MITOS: Improved de Novo Metazoan Mitochondrial Genome Annotation." *Molecular Phylogenetics and Evolution* 69 (2). Elsevier Inc.: 313–19. https://doi.org/10.1016/j.ympev.2012.08.023

Blanda, Elisa, Guillaume Drillet, Cheng-Chien Huang, Jiang-Shiou Hwang, Hans Henrik Jakobsen, Thomas Allan Rayner, Huei-Meei Su, Cheng-Han Wu, and Benni Winding Hansen. 2015. "Trophic Interactions and Productivity of Copepods as Live Feed from Tropical Taiwanese Outdoor Aquaculture Ponds." *Aquaculture* 445. Elsevier B.V.: 11–21. https://doi.org/10.1016/j.aquaculture.2015.04.003

Blanda, Elisa, Guillaume Drillet, Cheng Chien Huang, Jiang Shiou Hwang, Jacob Kring Højgaard, Hans Henrik Jakobsen, Thomas Allan Rayner, Huei Meei Su, and Benni Winding Hansen. 2017. "An Analysis of How to Improve Production of Copepods as Live Feed from Tropical Taiwanese Outdoor Aquaculture Ponds." *Aquaculture* 479 (December 2016). Elsevier: 432–41. https://doi.org/10.1016/j.aquaculture.2017.06.018

Bron, J. E., D. Frisch, E. Goetze, S. C. Johnson, C. E. Lee *et al.*, 2011 Observing Copepods through a Genomic Lens. Front. Zool. 8: 22. https://doi.org/10.1186/1742-9994-8-22

Chang, W.-B., and C.-H. Lei, 1993 Development and Energy Content of a Brackish-Water Copepod, *Apocyclops Royi* (Lindberg) Reared in a Laboratory. Bull. Inst. Zool. Acad. Sin. 32: 62–81.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome. Nat. Biotechnol. 29: 644–652. https://doi.org/10.1038/nbt.1883

Jørgensen, T. S., B. Petersen, H. C. B. Petersen, P. D. Browne *et al.*, 2019 "The genome and MRNA transcriptome of the cosmopolitan calanoid copepod acartia tonsa dana improve the understanding of copepod genome size evolution. Genome Biol. Evol.. https://doi.org/10.1093/gbe/evz067

Kao, D., A. G. Lai, E. Stamataki, S. Rosic, N. Konstantinides *et al.*, 2016 The genome of the crustacean *Parhyale Hawaiensis*, a model for animal development, regeneration, immunity and lignocellulose digestion. eLife 5: 1–45. https://doi.org/10.7554/eLife.20062

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923

Lee, K. W., H. U. Dahms, H. G. Park, and J. H. Kang, 2013 Population Growth and Productivity of the Cyclopoid Copepods *Paracyclopina Nana*, *Apocyclops Royi* and the Harpacticoid Copepod *Tigriopus Japonicus* in Mono and Polyculture Conditions: A Laboratory Study. Aquacult. Res. 44: 836–840. https://doi.org/10.1111/j.1365-2109.2011.03071.x

Lenz, P. H., V. Roncalli, R. P. Hassett, L. S. Wu, M. C. Cieslak *et al.*, 2014 De Novo Assembly of a Transcriptome for *Calanus Finmarchicus* (Crustacea, Copepoda) - The Dominant Zooplankter of the North Atlantic Ocean. PLoS One 9: e88589. https://doi.org/10.1371/journal.pone.0088589

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map Format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Madoui, M. A., J. Poulain, K. Sugier, M. Wessner, B. Noel *et al.*, 2017 New Insights into Global Biogeography, Population Structure and Natural Selection from the Genome of the Epipelagic Copepod Oithona. Mol. Ecol. 26: 4467–4482. https://doi.org/10.1111/mec.14214

Nielsen, Bolette LH, Louise Gøtterup, Tue Sparholt Jørgensen, Benni Winding Hansen, Lars Hestbjerg Hansen, John Mortensen, and Per Meyer Jepsen. 2019. "N-3 PUFA Biosynthesis by the Copepod *Apocyclops Royi* Documented Using Fatty Acid Profile Analysis and Gene Expression Analysis." *Biology Open* 8. https://doi.org/10.1242/bio.038331

Pan, Y. J., I. Sadovskaya, J. S. Hwang, and S. Souissi, 2018 Assessment of the Fecundity, Population Growth and Fatty Acid Composition of *Apocyclops Royi* (Cyclopoida, Copepoda) Fed on Different Microalgal Diets. Aquacult. Nutr. 24: 970–978. https://doi.org/10.1111/anu.12633

Pan, Yen-Ju, Anissa Souissi, Sami Souissi, and Jiang-Shiou Hwang. 2016. "Effects of Salinity on the Reproductive Performance of *Apocyclops Royi* (Copepoda, Cyclopoida)." *Journal of Experimental Marine Biology and Ecology* 475. Elsevier: 108–13. https://doi.org/10.1016/j.jembe.2015.11.011

Pan, Y. J., A. Souissi, I. Sadovskaya, B. W. Hansen, J. S. Hwang *et al.*, 2017 Effects of Cold Selective Breeding on the Body Length, Fatty Acid Content, and Productivity of the Tropical Copepod *Apocyclops Royi* (Cyclopoida, Copepoda). J. Plankton Res. 39: 994–1003. https://doi.org/10.1093/plankt/fbx041

Rasch, E. M., and G. A. Wyngaard, 2006 Genome Sizes of Cyclopoid Copepods (Crustacea): Evidence of Evolutionary Constraint. Biol. J. Linn. Soc. Lond. 87: 625–635. https://doi.org/10.1111/j.1095-8312.2006.00610.x

Ross, M. G., C. Russ, M. Costello, A. Hollinger, N. J. Lennon *et al.*, 2013 Characterizing and measuring bias in sequence data. Genome Biol. 14: R51. https://doi.org/10.1186/gb-2013-14-5-r51

Schubert, Mikkel, Stinus Lindgreen, and Ludovic Orlando. 2016. "AdapterRemoval v2: Rapid Adapter Trimming, Identification, and Read Merging." *BMC Research Notes* 9 (1). BioMed Central: 1–7. https://doi.org/10.1186/s13104-016-1900-2

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. Bioinformatics 31: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simpson, J. T., and R. Durbin, 2012 Efficient de Novo assembly of large genomes using compressed data structures. Genome Res. 22: 549–556. https://doi.org/10.1101/gr.126953.111

Smit, A. F. A., and R. Hubley, n.d. "RepeatModeler Open-1.0." http://www.repeatmasker.org.

Smit, A. F. A., R. Hubley, and P. Green, n.d. "RepeatMasker Open-4.0." http://www.repeatmasker.org.

Su, H.-M, S.-H Cheng, T.-I Chen, and M.-S Su. 2007. "Culture of Copepods and Applications to Marine Finfish Larval Rearing in Taiwan." *Copepods in Aquaculture*, November, 183–94. https://doi.org/10.1002/9780470277522.ch14

Wyngaard, Grace A, Carlos E F Rocha, and Almir Pepato. 2011. "Familial Level Phylogeny of Free-Living Cyclopoids (Copepoda), Inferred from Partial 18S Ribosomal DNA." *Studies on Freshwater Copepoda*, no. March 2016: 507–44.

Xue, W., J.-T. Li, Y.-P. Zhu, G.-Y. Hou, X.-F. Kong *et al.*, 2013 L_RNA_scaffolder: Scaffolding Genomes with Transcripts. BMC Genomics 14: 604. https://doi.org/10.1186/1471-2164-14-604

Communicating editor: B. Oliver