The Genome and mRNA Transcriptome of the Cosmopolitan Calanoid Copepod *Acartia tonsa* Dana Improve the Understanding of Copepod Genome Size Evolution

Tue Sparholt Jørgensen[1,2,*], Bent Petersen[3,4], H. Cecilie B. Petersen[1], Patrick Denis Browne[2], Stefan Prost[5,6,7], Jonathon H. Stillman[6,8], Lars Hestbjerg Hansen[2,*], and Benni Winding Hansen[1]

[1]Department of Science and Environment, Roskilde University, Denmark

[2]Department of Environmental Science – Environmental Microbiology and Biotechnology, Aarhus University, Roskilde, Denmark

[3]Natural History Museum of Denmark, University of Copenhagen, Denmark

[4]Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia

[5]Department of Integrative Biology and Evolution, Research Institute of Wildlife Ecology, University of Veterinary Medicine, Vienna, Austria

[6]Department of Integrative Biology, University of California, Berkeley

[7]National Zoological Garden, South African National Biodiversity Institute, Pretoria, South Africa

[8]Estuary and Ocean Science Center, San Francisco State University, Tiburon, California

*Corresponding authors: E-mails: lhha@envs.au.dk; tuesparholt@gmail.com.

Accepted: March 26, 2019

Data deposition: This project has been deposited at The European Nucleotide Archive under the accession PRJEB20069.

Abstract

Members of the crustacean subclass Copepoda are likely the most abundant metazoans worldwide. Pelagic marine species are critical in converting planktonic microalgae to animal biomass, supporting oceanic food webs. Despite their abundance and ecological importance, only six copepod genomes are publicly available, owing to a number of factors including large genome size, repetitiveness, GC-content, and small animal size. Here, we report the seventh representative copepod genome and the first genome and the first transcriptome from the calanoid copepod species *Acartia tonsa* Dana, which is among the most numerous mesozooplankton in boreal coastal and estuarine waters. The ecology, physiology, and behavior of *A. tonsa* have been studied extensively. The genetic resources contributed in this work will allow researchers to link experimental results to molecular mechanisms. From PCR-free whole genome sequence and mRNA Illumina data, we assemble the largest copepod genome to date. We estimate that *A. tonsa* has a total genome size of 2.5 Gb including repetitive elements we could not resolve. The nonrepetitive fraction of the genome assembly is estimated to be 566 Mb. Our DNA sequencing-based analyses suggest there is a 14-fold difference in genome size between the six members of Copepoda with available genomic information. This finding complements nucleus staining genome size estimations, where 100-fold difference has been reported within 70 species. We briefly analyze the repeat structure in the existing copepod whole genome sequence data sets. The information presented here confirms the evolution of genome size in Copepoda and expands the scope for evolutionary inferences in Copepoda by providing several levels of genetic information from a key planktonic crustacean species.

Key words: calanoid copepod genome, genome assembly, repetitive DNA, genome size evolution, invertebrate genomics, comparative genomics.

Introduction

Since the publication of the first version of the human genome sequence in 2001, >2,000 eukaryotic genomes have been collected in the reference sequence database under National Center for Biotechnology Information (NCBI) (Pruitt et al. 2007). The species with available genomic resources are predominately those which impact human health or are biomedically or agriculturally important. Genomic resources are

available to a far lesser extent in species with ecological significance. Arthropoda is the most species rich phylum on Earth, and in marine environments, Copepoda is the most species-rich subclass with >11,000 described species (Appeltans et al. 2012; Dunn and Ryan 2015), and the most abundant animal on Earth (Humes 1994). Yet, only six copepod genomes have hitherto been published. The six species are the calanoid *Eurytemor a affinis* (Evans et al. 2013), the cyclopoid *Oithona nana* (Madoui et al. 2017), the harpacticoid *Tigriopus californicus* (Barreto et al. 2018) and *Tigriopus kingsejongensis* (Han et al., 2016), and *Caligus rogercresseyi* and *Lepeophtheirus salmonis*, which both belong to the order Siphonostomatoida and are important pests in salmon aquaculture (Costello 2006).

Acartia tonsa is a marine, euryhaline calanoid copepod of about 1.5 mm in adult length with a cosmopolitan neritic distribution, and in many ecosystems, it is the most numerous mesozooplankton species (fig. 1*A*) (Albaina et al. 2016). It performs a vital function as it is a primary grazer on microalgae, and in turn is a main source of prey for the larvae of many fish species in estuarine, coastal and upwelling regions (Turner 2004). Further, *A. tonsa* is an emerging model organism, with research published in a diverse array of scientific fields such as ecology, physiology, ecotoxicology, and animal behavior (Støttrup et al. 1986; Drillet et al. 2006; Jepsen et al. 2015; Wendt et al. 2016; Hansen et al. 2017). *Acartia tonsa* is also an emerging live feed species in aquaculture, where it could trigger natural predation behavior and supply optimal nutrition for the larvae of fish species with economic importance or which are endangered in the wild (Støttrup et al. 1986; Broglio et al. 2003; Abate et al. 2015, 2016). Despite *A. tonsa*'s multifaceted importance, partial versions of the mitochondrial cytochrome oxidase subunit I (COI) and the ribosomal 18S rRNA genes have been the only available genetic resources for *A. tonsa* until now (Chen and Hare 2008; Drillet et al. 2008; Laakmann et al. 2013; Albaina et al. 2016).

Copepod genomes are particularly difficult to assemble as the genomes are often large, have a very low guanine–cytosine (GC)-content, around 30%, and because the animals are so small that a single animal rarely harbors a sufficient amount of genetic material for analysis (Gregory et al. 2000; Bron et al. 2011; Madoui et al. 2017). This is compounded by the medical and agriculture focus of modern genome assembly pipelines, where diploidy, small genome size, abundant genetic material, and a GC-content of about 50% is assumed, required, or favored (Miller et al. 2010).

Genome sizes of copepods species are highly variable. The genome assemblies available range 12-fold in size from 82 to 986 Mb (this study, table 1), whereas the haploid genome sizes available from Feulgen staining of nuclei or flow cytometry range 100-fold between 140 Mb and 14 Gb (fig. 1*D*, Gregory, TR, 2018, Animal Genome Size Database, http://www.genomesize.com; last accessed February 1, 2019). Within the order Calanoida, of which *A. tonsa* is a member, the reported haploid genome sizes vary 40-fold between 330 Mb and 14.4 Gb. For reference, the haploid human genome is about 3.2 Gb (Lander et al. 2001). The genome size of an animal is, however, not directly related to the assembly size, as the intronic and intergenic regions can be assembled to varying degrees which largely determine the assembly size (Francis and Wörheide 2017). Thus, we used all existing copepod whole genome sequence (WGS) resources and our contributed *A. tonsa* genome assembly to determine the total genome size of copepods from the four orders of Calanoida, Cyclopoida, Siphonostomatoida, and Harpacticoida using the *k*-mer frequency based preQC tool. We further characterized the contributed genome of *A. tonsa* Dana by analyzing the content of mitochondrial marker genes. Although few WGS data sets are available from Copepoda, transcriptome assemblies are much more common, with >20 data sets from 16 species available through the NCBI/EBI/DNA Data Bank of Japan, likely owing to a relative ease of obtaining good quality transcriptomes compared with genomes. Our aim with this genome project was to contribute to the knowledge base of genome evolution in Copepoda and for the first time provide the research community with sufficient genomic and transcriptomic resources to embark on evolutionary, ecological, and physiological studies involving the important copepod species *A. tonsa*.

Materials and Methods

Culture and Animal Husbandry

The *A. tonsa* culture strain DFU-ATI was used for all nucleic acid extractions. DFU-ATI has been in continuous culture without restocking. It was obtained off the coast of Helsingør in the Øresund strait in Denmark in 1981. Behavioral, ecological, physiological, and molecular aspects of the biology of *A. tonsa* strain DFU-ATI have been described in several publications (Støttrup et al. 1986; Tiselius et al. 1995; Drillet et al. 2006, 2011, 2015; Jepsen et al. 2015; Hansen et al. 2016, 2017). The continuous *A. tonsa* culture fed the microalga *Rhodomonas salina* in excess according to Berggreen et al. (1988) was kept in 70-l plastic buckets in a stable 17 °C environment in the dark. The culture was kept in 0.2-μm filtered water collected from the sea floor in Kattegat, near the site where the culture originated. The salinity was stable at 32 ± 1 ppt. Eggs and debris were collected from the bottom of the culture daily.

Animals were sorted by size by sequential filtering: adults were caught on a 250-μm filter, copepodites and nauplii were caught on a 125-μm filter and eggs were caught on a 70-μm filter. Animals were thoroughly rinsed with 0.2-μm filtered seawater which was removed prior to nucleic acid extraction. For the polymerase chain reaction (PCR)-free libraries, individual adult animals were picked with a Pasteur pipette and placed in sterile, 0.2-μm filtered seawater which was removed prior to nucleic acid extraction. Tissues for RNA

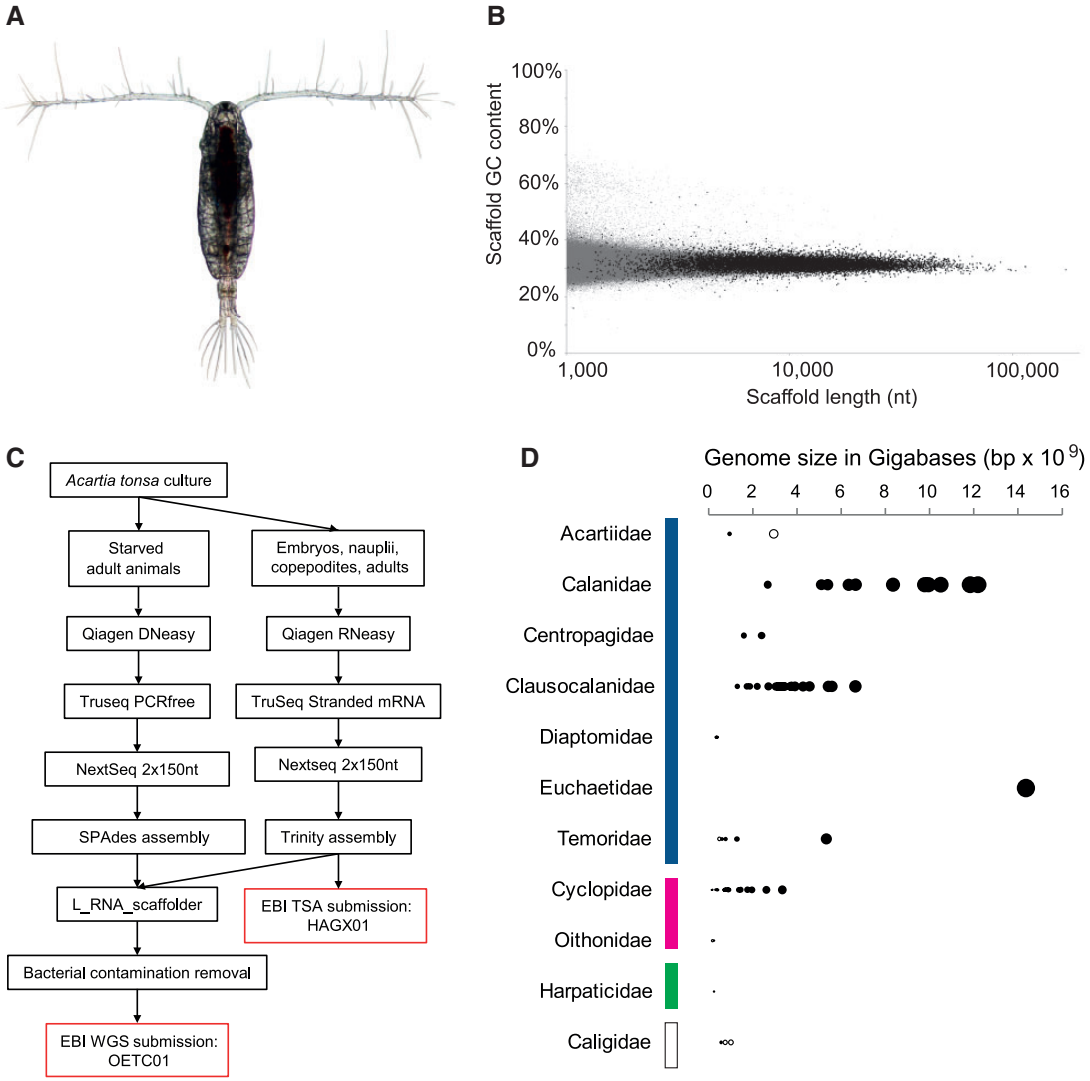

Fig. 1.—*Acartia tonsa* genome assembly. (*A*) Female specimen of the DFU-ATI strain of *A. tonsa* used in this study. Photo by Minh Vu Thi Thuy. (*B*) Length and GC-content of each scaffold in the Aton1.0 assembly. Black dots are scaffolds connected using mRNA information and gray dots are all other scaffolds. In total, 351,850 scaffolds are included in Aton1.0. The scaffolds are tightly distributed around 32% GC with lengths ranging from 1 to 174 kb. Most scaffolds of around or above 10 kb have been scaffolded using mRNA information (black dots). (*C*) Workflow for producing the Aton1.0 assembly from the DFU-ATI strain of *A. tonsa*. (*D*) Overview of reported genome sizes for the subclass Copepoda. The area of individual plot points is equal to the axis value. Black dots represent information from the Animal Genome Size Database based on nucleus staining (Gregory, TR, 2018, http://www.genomesize.com) and the five open circles represent the genome sizes estimated from WGS data in this study. Within Copepoda, a 100-fold difference in genome size from the smallest Cyclopoid (pink bar, 0.14 Gb) to the largest Calanoid (blue bar, 14 Gb) can be seen. Within the order Calanoida, the genome sizes vary >10-fold between the smallest Diaptomidae (0.95 Gb) and the largest Calanidae (12 Gb). Harpactidae species are marked with a green bar and Caligidae species with a white bar.

extraction were placed in at least five volumes of RNAlater 24 h prior to extraction.

Nucleic Acid Extraction, Library Construction, and Sequencing

DNA was extracted using the DNeasy mini Blood and Tissue kit from Qiagen Venlo, Netherlands according to the manufacturer's protocol with the following modifications: sample tissue was ground manually with a pestle in a 1.5-ml Eppendorf tube for at least 2 min and incubated with proteinase K and RNase A for 4 h with periodic mixing.

RNA was extracted using the RNeasy mini Blood and Tissue kit from Qiagen according to protocol with the following modifications: sample tissue was kept on ice and ground in a 1.5-ml Eppendorf tube using a pestle and electric motor for 2 min and incubated with proteinase K for 4 h with periodic mixing. The embryo RNA sample consisted of eggs from a few

Table 1
Overview of Existing Genomic Resources for Copepoda

Species	*Acartia tonsa*	*Eurytemora affinis*	*Oithona nana*	*Caligus rogercresseyi*	*Lepeophtheirus salmonis*	*Tigriopus californicus*
Species info	Common estuarine and coastal species	Common estuarine and coastal species	Common estuarine and coastal species	Sea louse, aquaculture pest	Salmon louse, aquaculture pest	Common Pacific intertidal species
Assembly size (Mb)	986	351	82	398	665	178
Assembly GC-content (%)	31	32	40	32	31	42
Sequencing effort (Gb)	116.3	39.3	5.1	37.8	29.8	42.4
Contig N50 (kb)	3.2	68	39	1.6	17	15
Number of scaffolds	351,850	6,171	4,626	288,616	83,165	2,385
Reported sequencing depth	50×	75×	50×	95×	45×	NA
Scaffolding	Paired end, mRNA	Paired end, Matepair	Paired end, Matepair, mRNA	Paired end	Paired end	Paired end, Matepair
Animal tissue	Pool of adult animals from culture	Pool of environmental eggs	Pool of environmental animals	Adult female animal	Adult female animal	Pools of animals from culture
Last updated	December 20, 2017	December 13, 2017	February 17, 2017	May 8, 2015	May 8, 2015	March 19, 2015
Assembly accession	OETC01	AZAI02	FTRT01	LBBV01	LBBX01	TCALIF_genome_v1.0
Reference	This study	The i5k Initiative	Madoui et al. (2017)	Unpublished, Leong et al.	Unpublished, Leong et al.	The i5k Initiative

and up to 50-h old to ensure that all stages in *A. tonsa* embryogenesis were present (Nilsson and Hansen 2018).

The PCR-free libraries were constructed using DNA from adult animals with the Truseq PCR free kit (Illumina, San Diego, CA) according to the manufacturer's protocol using 1-µg DNA as input. Shearing of DNA in the PCR-free protocol was done on a Covaris E210 (Woburn, MA) with miniTUBE with the following settings: intensity, 3; duty cycle, 5%; cycles/burst, 200; and treatment time, 80. The libraries were sequenced on an Illumina NextSeq 500 with 2× 150-bp PE high kits. The three transcriptome libraries covering all life stages of *A. tonsa* were built using the Illumina TruSeq stranded mRNA kit with half volumes according to Combs and Eisen (2015) immediately after RNA extraction. No DNase step was used as it would have negative impact on long fragments in the libraries. For each library, 1 µg of total RNA was used as input. RNA libraries were sequenced on an Illumina Nextseq 500 using a single 2× 150-bp PE high kit. All sequencing libraries were analyzed on a Bioanalyzer with DNA 7500 Assay chip (Santa Clara, CA) and the molarity of cluster forming fragments was analyzed using the KAPA Universal qPCR Master Mix (KK4824, KAPA Biosystems, Wilmington, MA). Nucleic acid concentration was measured using the Qubit system (Thermo Fisher Scientific, Waltham, MA).

An overview of the libraries was constructed, and more details on indexes, insert sizes, biological materials, and SRA accession numbers can be found in supplementary material 1, Supplementary Material online.

Data Handling, Assembly, Scaffolding, and Analysis

Basic statistics and data handling was done in a UNIX environment using Biopieces (Hansen, MA, www.biopieces.org, unpublished). mRNA data from eggs, nauplii, copepodites, and adults were pooled and assembled using the Trinity pipeline (v. 2.5.1) with default parameters and the built-in version of trimmomatic to quality trim the reads prior to assembly and to remove adapters (Grabherr et al. 2011; Bolger et al. 2014). Data from PCR-free libraries were pooled and used for assembly with SPAdes assembler (Bankevich et al. 2012) (v. 3.9.0, *k*-mer size 77), using paired end information for scaffolding. The SPAdes genome assembly was further scaffolded with the assembled mRNA transcriptome using L_RNA_scaffolder (Xue et al. 2013) with BLAT v. 36x2 (Kent 2002). Transcripts shorter than 500 nt were not used for scaffolding. Scaffolds smaller than 1,000 bp were discarded. After testing several bacterial contamination removal strategies, we decided on a BLAST-based method on scaffold level sequences as all sequences with obvious bacterial characteristics (high GC% and 100 kb–1 Mb in length) were removed and few likely copepod sequences were removed (data not shown). Briefly, scaffolds were masked using RepeatMasker with repeats from RepeatModeler and the Arthropoda and ancestral (shared) repeats from repbase v. 22.05 (downloaded June

2, 2017) (Smit and Hubley 2019; Smit et al. 2019). The masked scaffolds were BLAST-searched against the refseq database of representative prokaryotes (downloaded March 23, 2017) using the built-in BLAST in CLCgenomics 9.0 (*e*-value $\leq 10^{-6}$) (Altschul et al. 1997) and sequences with a longest hit longer than 500 bp were removed from the assembly. Raw reads and assemblies for the four published copepod reference genomes were downloaded from NCBI using the following accession numbers: *E. affinis* (assembly: AZAI02, raw reads: SRX387234-7), *O. nana* (assembly: FTRT01, raw reads ERX1858579-83), *C. rogercresseyi* (assembly: LBBV01, raw reads: SRX976492), *L. salmonis* (assembly: LBBX01, raw reads SRX976783), and *T. californicus* (assembly TCALIF_genome_v1.0, raw reads SRX469409 and SRX469410). Genome size was estimated using the preQC tool from the SGA assembler (Simpson and Durbin 2012) on reads cleaned using AdaptorRemoval (Lindgreen 2012) with the switches: –trimns –trimqualities. Repetitive sequence fractions in genome assemblies were identified using RepeatMasker on the Arthropoda and ancestral (shared) repeats from repbase v. 22.05 (downloaded June 2, 2017) merged with output from RepeatModeler run with standard parameters (Smit and Hubley 2019; Smit et al. 2019).

The nine complete copepod mitochondrial genomes used to find mitochondrial scaffolds in Aton1.0 were downloaded from the Organelles section of the NCBI genomes browser (https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/copepoda) in April 2018. Their accession numbers can be found in supplementary material 1, Supplementary Material online. BLAST search of the nine existing mitochondrial genomes against the Aton1.0 assembly was done in CLC Genomics workbench v. 10.1.1 using standard parameters and yielded three scaffolds with mitochondrial genes. The scaffolds carrying mitochondrial DNA were analyzed using the MITOS2 web interface with RefSeq63 Metazoa reference and table 5 invertebrate genetic code (Bernt et al. 2013).

COI genes from the genus *Acartia* along with 25 *Temora longicornis* COI genes were downloaded from the NCBI nucleotide collection (https://www.ncbi.nlm.nih.gov/nuccore/?term=Acartia%20COI) in April 2018. The accession numbers of the 544 + 25 sequences can be found in supplementary material 1, Supplementary Material online. Multiple alignment of Aton1.0 and database COI, trimming of sequence ends, realigning, and de novo Neighbor-Joining phylogenetic tree construction with 100 bootstraps were all performed in CLC genomics workbench version 10.1.1 using default parameters. Analysis of genome and transcriptome completeness was done using the Universal Single Copy Orthologs BUSCO (v2.0) with the arthropoda_odb9 lineage data set and ab initio gene prediction using Augustus (v.3.2.3), in all cases with the fly training set, and the switch "geno" for the genomes and "tran" for the transcriptome (Stanke et al. 2004; Simão et al. 2015).

To place Aton1.0 within Copepoda, the eight genes COX1, COX2, COX3, CYTB, ND1, ND3, ND4, and ND5 were extracted from the MITOS annotation of Aton1.0 scaffolds and from the nine complete copepod mitochondria downloaded from NCBI (supplementary material 1, Supplementary Material online). The genes were aligned individually using the MAFFT online platform (Katoh and Standley 2013), using the algorithm Q-INS-I iterative refinement method (Katoh and Toh 2008). Individual genes were then concatenated using Sequence Matrix (Vaidya et al. 2011).

The concatenated data set of the mitochondrial genes was analyzed using Bayesian method (BA). The analysis was performed using MrBayes version 3.2.6 (Ronquist and Huelsenbeck 2003) available on CIPRES Gateway. To identify the best substitution model for molecular evolution a Model Test was run on CLC Genomics Workbench 10.1.1 (https://www.qiagenbioinformatics.com/) on each individual gene prior to analyses, using Akaike information criterion for COX2, COX3, CYTB, ND1, and ND3 and corrected Akaike information criterion for COX1, ND4, and ND5. The models selected for each gene included a General time reversible (GTR) model of sequence evolution (Yang 1994) with gamma distribution and a proportion of invariable sites (GTR + I + Γ) for all genes. The data set was run with two independent analyses using four chains (three heated and one cold). Number of generations was set to 30 million, sampling every 1,000 generations. Burn-in was set to 10 million generations.

Results and Discussion

Sequencing and Assembly Metrics

The whole genome sequencing workflow for the *A. tonsa* genome (Aton1.0) and transcriptome (fig. 1*C*) yielded a total of 356,383,864 Illumina reads from PCR-free libraries and 112,558,144 Illumina reads from three stranded mRNA libraries covering all life stages from embryos over nauplii and copepodites to adults. Further, PacBio and Mate Pair data sets were produced, but not used in the assembly process because the coverage was insufficient to successfully scaffold contigs (data not shown but available under the study accession PRJEB20069). The decision to not use the distance information libraries closely resembles the conclusions in a recent article which reports that low coverage distance information does not improve assembly (Renaut et al. 2018).

In total, >145,000,000,000 sequenced bases were used for the assemblies of *A. tonsa*, which is 3–5 times more raw data than any other copepod WGS study to date. The assembly of mRNA data yielded 118,709,440 bases in 61,149 transcripts and an additional 56,257 isoforms which are available at ENA (https://www.ebi.ac.uk/) under the accession HAGX01 (fig. 1*C*). The SPAdes (Bankevich et al. 2012) assembly of PCR-free data was scaffolded with the transcriptome yielding a genome assembly of 989,163,677 bp distributed in 351,850

scaffolds (fig. 1*B*). The assembly is available at ENA (https://www.ebi.ac.uk/) under the name Aton1.0 and the accession OETC01. More than 20,000 contigs were joined using mRNA information, substantially adding to the contiguity of gene carrying scaffolds (fig. 1*B*).

The GC-content of the Aton1.0 assembly is 32% (fig. 1*B*), substantially lower than many model species such as human or mouse but similar to the available Copepod genomes (table 1). Because whole animals were used for nucleic acid extraction, bacterial contamination is expected to be present in the raw PCR-free data. The BLAST-based removal of scaffolds of bacterial origin eliminated 3,953 scaffolds, many of which had a substantially different sequence length and GC-content than other Aton1.0 scaffolds, further indicating bacterial origin (data not shown).

Assembly Completeness and Content

To estimate the completeness of the Aton1.0 assembly, we used the BUSCO system of orthologous single copy genes and the arthropods database on predicted genes (Simão et al. 2015). The Aton1.0 assembly carries 59.5% (634 of 1,066) complete single copy BUSCO genes, 1.9% complete but duplicated genes (20 of 1,066), 20.6% fragmented genes (220 of 1,066), and 18.0% missing genes (192 of 1,066) out of the 1,066 Arthropod gene models (fig. 2*C*). These numbers are not comparable with well-studied species such as *Drosophila melanogaster* (fig. 2*C*, 99.0% complete, single copy genes) but are close to those of the other published copepods, though the sequencing effort for Aton1.0 is unprecedented (fig. 2*C* and table 1). The low number of duplicate BUSCO genes suggests that the Aton1.0 assembly is not populated by many variants of the same core genes, even though the biological material was obtained from a large number of animals. For the *A. tonsa* transcriptome, 91.4% of BUSCO genes are complete, and a further 7.9% fragmented, suggesting that this resource is very useful for scaffolding, gene modeling, and gene functional annotation. Genes annotation was done using MAKER2 (Holt and Yandell 2011) and both the mRNA transcriptome, ab initio gene prediction using Augustus, and related species gene models, but because the resulting gene set had a substantially lower BUSCO score than the Aton assembly alone, we decided to not use it for further analysis (supplementary material 1, Supplementary Material online)

Placement of Aton1.0 in *Acartia* and Copepoda

Mitochondrial genes and genomes are widely used for phylogenetic analysis in Copepoda because they often can resolve specimens to species (Bernt et al. 2013). Because one of the few sequences available from the DFU-ATI strain of *A. tonsa* is the mitochondrial COI gene, we investigated the mitochondrial components in the Aton1.0 assembly. Three scaffolds were found to carry mitochondrial genes and they were annotated using MITOS2 (Bernt et al. 2013) (supplementary material 1, Supplementary Material online). Within these three scaffolds, 15 out of 22 expected tRNA genes are present as well as 11 of 15 expected protein coding genes (table 2). The Aton1.0 COI gene was aligned to all 541 COI entries for the genus *Acartia* and a de novo phylogenetic tree was constructed based on a region shared between all database versions using 25 *Temora longicornis* (Copepoda, Calanoida) COI genes as outgroup (fig. 2*A*). The Aton1.0 COI is 99.7% identical to many entries from the North Atlantic clade of *A. tonsa*, and most versions annotated as *A. tonsa* group together, confirming the placement of Aton1.0 within the most studied clade of *A. tonsa*, and the relatedness of database entries annotated as *A. tonsa*.

Because nine complete circular copepod mitochondrial genomes are available, the Aton1.0 assembly can be placed within Copepoda using a multigene strategy. We chose eight complete genes for analysis, as these eight genes aligned across the database and our *A. tonsa* resource. Extreme mitochondrial DNA divergence has previously been reported even within the copepod species *T. californicus*, why it is not surprising that not all copepod mitochondrial genes align across orders (Barreto et al. 2018). The genes used for phylogenetic placement of *A. tonsa* can be found in table 2. The result from the Bayesian phylogenetic analysis is presented in figure 2*B*. *A. tonsa* forms, together with *Calanus hyperboreus*, the clade of Calanoida (BPP: 1), as a sister group to the clade of the remaining orders of Cyclopoida, Harpacticoida, and Siphonostomatoida. *Paracyclopina nana*, *Lernaea cyprinacea*, and *Sinergasilus polycolpus* forms the monophyletic clade of Cyclopoida (BPP: 0.99). The paraphyletic clade of Harpacticoida and Siphonostomatoida is unsupported (BPP: 0.53). Siphonostomatoida is nested as a monophyletic sister clade (BPP: 1) to the clade of *Tigriopus japonicus* and *T. californicus* (BPP: 1). The split between Calanoida and the other Copepod orders correlates closely with recent phylogenetic work on copepod orders where the calanoid species form the superorder Gymnoplea, whereas the other copepods form the superorder Podoplea (Eyun 2017; Khodami et al. 2017). The placement of the orders Cyclopoida, Harpacticoida, and Siphonostomatoida, however, is inconsistent in the recent articles, both of which are also inconsistent with our results, which suggests that Siphonostomatoida is nested inside Harpacticoida, with Cyclopoida as an outgroup (fig. 2*B*). The cited studies use a larger number of species (Khodami et al. 2017) or genes (Eyun 2017) for the analysis than the present work. We do not intent to challenge the validity of either, yet our result adds to the uncertainty of the placement of the copepod orders.

Genome Sizes and Fractions

The total genome size of an animal including repetitive elements is not routinely deciphered from NGS data and reported along with the assembly. The preQC program

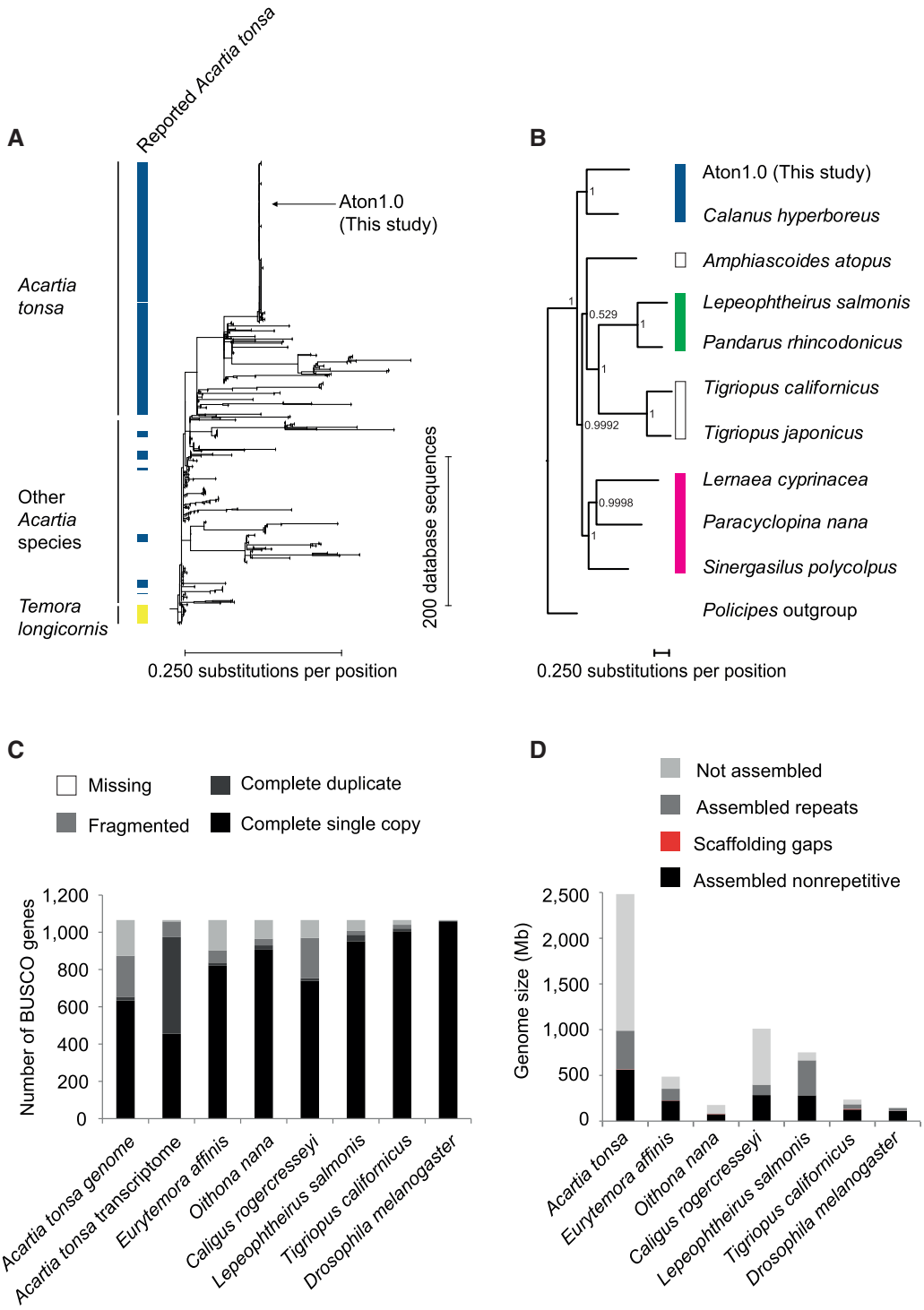

Fig. 2.—Placement and characterization of the Aton1.0 assembly. (*A*) Placement of Aton1.0 within the genus *Acartia*. The Aton1.0 COI gene groups within the most well-studied North Atlantic clade of *Acartia tonsa* which is in line with the origin of the culture. (*B*) Placement of Aton1.0 within the subclass Copepoda. Phylogenetic tree based on Bayesian analysis of a combined gene data set. Nodal support is displayed as Bayesian posterior probability at each branch. The colored bars represent the orders Calanoida (blue), Harpacticoida (outline), Cyclopoida (cyan), and Siphonostomatoida (green). The branch separating the calanoid copepods from the other orders closely resemble recent phylogenetic analyses based primarily on different genes (Eyun 2017; Khodami et al. 2017). (*C*) BUSCO core gene content of the genome assemblies of copepods and *Drosophila melanogaster*. Between 2.4% and 18% of BUSCO genes are missing from assemblies (outline), between 2.4% and 21% are fragmented (light gray), and between 1.2% and 3.2% exist in duplicate (dark gray). From 59% to 94% of BUSCO genes are complete and single copy in the assemblies. For all metrics, the *A. tonsa* genome assembly performs

Table 2

Overview of Aton1.0 Mitochondrial Resources. The identified genes are shown in black, and expected mitochondrial genes which were not identified are shown in red.

Aton1.0 mitochondrial genes and tRNAs	Used for phylogeny
ATP6	
ATP8	
COI	x
COI2	x
COI3	x
CYTB	x
ND1	x
ND2	
ND3	x
ND4	x
ND4L	
ND5	x
ND6	
rRNA lsu	
rRNA ssu	
trnA	
trnC	
trnD	
trnE	
trnF	
trnG	
trnH	
trnI	
trnK	
trnL1	
trnL2	
trnM	
trnN	
trnP	
trnQ	
trnR	
trnS	
trnS	
trnT	
trnV	
trnW	
trnY	

from the SGA genome assembly pipeline uses *k*-mer frequencies to predict the total genome size and can be used to evaluate WGS data before assembly (Simpson and Durbin 2012). We tested preQC on the *D. melanogaster* genome. The estimated genome size of *D. melanogaster* is, within 3%, the same than the high-quality assembly genome length (supplementary material 1, Supplementary Material online). This result permits us to use preQC on the five WGS genomes of copepods available at NCBI. For *A. tonsa*, the total genome size is estimated to be 2.48 Gb, slightly smaller than the size of the human genome (fig. 2*D*). The almost 2.5-Gb genome size estimate of *A. tonsa* differs substantially from the other copepod genomes which are estimated to be 0.49 Gb (*E. affinis*), 0.18 Gb (*O. nana*), 1.01 Gb (*C. rogercresseyi*), 0.75 Gb (*L. salmonis*), and 0.24 Gb (*T. californicus*). The complete preQC report for all species can be found in supplementary material 2, Supplementary Material online. A 100-fold range in genome size has been reported in Copepoda based on nucleic staining (fig. 1*D*), and the present study for the first time shows a large genome size range of 14-fold using NGS methods (fig. 2*D*). Because the quality and quantity of input data can influence the result of *k*-mer counting based analysis such as the preQC genome size analysis, it is important to be aware that the genome size results are estimates, and that they could change with more input data, or data with a different error profile. A recent study used flow cytometry to estimate the genome size of four species of calanoid copepods, three of which also has Feulgen staining genome size estimates available (Leinaas et al. 2016). The flow cytometry estimates were in all cases about half the size of the Feulgen staining estimates from the same species. This underlines the difficulty of copepod genome size, and makes comparisons across methods difficult. Of the species analyzed in this study using an NGS method, *E. affinis, L. salmonis*, and *T. californicus* also have genome size estimates from a different method. For all, our estimates (0.49, 0.75, and 0.24 Gb, respectively) are close to the Feulgen staining estimates (0.62, 0.57, and 0.25 Gb, respectively) (Rasch et al. 2004; Gregory TR, 2018, http://www.genomesize.com).

The large difference between the predicted genome sizes and the size of the genome assembly is hypothesized to be caused by both unassembled regions of the genome and the collapse of multiple repetitive regions to single scaffolds during assembly.

Because each assembly, scaffolding and gap filling approach yields different results, we determined the nonrepetitive fraction of each available copepod genome by modeling and masking out repeats and analyzing total genome size, assembled repetitive sequence size, and nonrepetitive sequence size. For *A. tonsa*, the nonrepetitive fraction is

Fig. 2.—Continued

worst, which is likely a result of the large genome size. The benchmarking species *D. melanogaster* has 99% complete single copy core genes. The mRNA transcriptome from all life stages of *A. tonsa* has 91% complete genes and additionally 8% fragmented genes, indicating that the resource is very powerful for identifying whole genes. (*D*) Total genome sizes for copepod WGS data sets and *D. melanogaster*. The unassembled genome fraction is depicted in light gray, the assembled repetitive genome fraction is in dark gray, the scaffolding gaps are in red and the nonrepetitive assembled fraction in black. The Aton1.0 assembly represents a genome that is estimated to be three to 20 times larger than the other copepods for which WGS resources are available through NCBI. The fraction of assembled nonrepetitive DNA is 22.7% (*A. tonsa*) to 53.8% (*Tigriopus californicus*) of the predicted total genome size, and only varies 7-fold from 75 Mb (*Oithona nana*) to 563 Mb (*A. tonsa*).

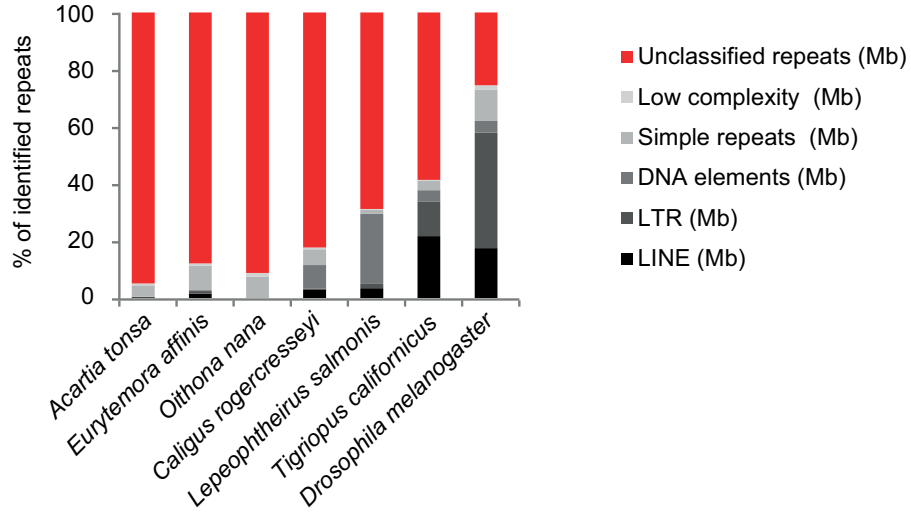

Fig. 3.—Classification of repeats in copepod WGS assemblies using RepeatModeler and RepeatMasker. Although >70% of identified repeats can be classified in the model species *Drosophila*, only between 5% and 20% of identified repeats from copepod genomes were classified. The unassembled genome fractions described in figure 2*D* and the large amount of unclassified repeats in copepods together illustrates how limited the current knowledge on this important animal group is.

566 Mb (fig. 2*D*). This means that only 22.7% of the *A. tonsa* genome is assembled and nonrepetitive. This figure is substantially lower than for the other copepod species which have assembled nonrepetitive fractions of 46.1% (224 Mb of 487 Mb), 42.4% (75 Mb of 177 Mb), 28.2% (285 Mb of 1011 Mb), 37.5% (282 Mb of 752 Mb), and 53.8% (127 Mb of 235 Mb) for *E. affinis*, *O. nana*, *C. rogercresseyi*, *L. salmonis*, and *T. californicus*, respectively. This difference is possibly caused by the large genome size of *A. tonsa*, as larger genome size can be associated with increased amounts of repetitive DNA, whereas the amount of exon DNA remains stable (Francis and Wörheide 2017). Figure 3 shows the amount of classified and unclassified repeats in the copepod genomes. Characteristically, the large majority of repeats in all copepod genomes cannot be classified by the RepeatMasker program (Smit et al. 2019) using RepeatModeler (Smit and Hubley 2019) output combined with the Repbase_arthropoda database (downloaded June 2, 2017). For *Drosophila*, most repeats are classified as long terminal repeats (40% of repeats) or long interspersed repeats (long interspersed nuclear element, 17% of repeats, fig. 3), whereas only 25% of identified repeats could not be classified. Likely, the *D. melanogaster* repeat classification is much better than the copepod repeat classification because *D. melanogaster* is a model species which specifically has been included in the RepBase repository. The WGS assembly of *T. californicus* is among the most contiguous copepod genome assemblies, and a larger fraction of repeats from this species can be classified by RepeatMasker than from the other copepod species (fig.3). Still, even for *T. californicus*, almost 60% of the repeats could not be classified. For *A. tonsa*, 95% of the identified repeats could not be classified, which is the highest rate of any of the analyzed copepods (fig. 3). The largest amount of classified repeats in the *A. tonsa* assembly is simple repeats, which make up 17 Mb or 4% of the identified repeats. This is equivalent to <1% of the total genome length. It is important to consider the large unassembled fraction of most copepod genomes when analyzing repeat structure, as the sequence absent from assemblies are very likely to be repetitive DNA and as the missing genome fraction constitute up to 60% of the total genome length.

Conclusions

Here, we present the first transcriptome and genome assembly of the ecologically important copepod species *A. tonsa* Dana. Eighty-two percent of the BUSCO core genes are present in the genome assembly, including 2% duplicated and 21% fragmented genes. In the transcriptome assembly, 99% of BUSCO genes could be found, including 8% fragmented genes. We further document the placement of the contributed genomic resources within Copepoda and the genus *Acartia* to the North Atlantic clade and estimate the genome size of *A. tonsa* to almost 2.5 Gb and compare with the other available copepod genomic resources where we find a 14-fold difference in estimated genome size. This is the first documentation of the range of genome size within Copepoda using DNA sequencing methods. Our resources are likely valuable to researchers in many scientific fields and can assist others to consider genome size when planning genome sequencing projects by elucidating the difference between the genome size and the assembly size of animal genomes.

Data Accessibility

Raw DNA sequencing data, the genome assembly, and the transcriptome assembly are available under the project PRJEB20069. The genome assembly prefix for Aton1.0 is OETC01 and the transcriptome prefix is HAGX01. All further data are available in supplementary material 1, Supplementary Material online, or upon request.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Anna la Cour for excellent technical assistance (ORCHID ID: 0000-0002-6990-9891), and Marlene Danner Dalgaard (ORCID ID: 0000-0002-4036-6408) for advice and assistance with DNA shearing. This study was supported by the Villum Foundation project AMPHICOP 8960. This work was supported by the Villum Foundation; Project AMPHICOP No. 8960.

Author Contributions

T.S.J., B.W.H., L.H.H., and J.H.S. designed research; T.S.J. performed research; B.P., S.P., J.H.S., P.D.B., and H.C.P. contributed analytical tools; T.S.J., B.P., S.P., J.H.S., and H.C.P. analyzed data; and T.S.J., B.P., S.P., J.H.S., H.C.P., P.D.B., L.H.H., and B.W.H. wrote the article.

Literature Cited

Abate TG, Nielsen R, Nielsen M, Jepsen PM, Hansen BW. 2016. A cost-effectiveness analysis of live feeds in juvenile turbot *Scophthalmus maximus* (Linnaeus, 1758) farming: copepods versus Artemia. Aquacult Nutr. 22(4):899–910.

Abate TG, et al. 2015. Economic feasibility of copepod production for commercial use: result from a prototype production facility. Aquaculture 436:72–79.

Albaina A, et al. 2016. Insights on the origin of invasive copepods colonizing Basque estuaries; a DNA barcoding approach. Mar Biodivers Rec. 9: 1–7.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.

Appeltans W, et al. 2012. The magnitude of global marine species diversity. Curr Biol. 22(23):2189–2202.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Barreto FS, et al. 2018. Genomic signatures of mitonuclear coevolution across populations of *Tigriopus californicus*. Nat Ecol Evol. 2(8):1250–1257.

Berggreen U, Hansen B, Kiørboe T. 1988. Food size spectra, ingestion and growth of the copepod *Acartia tonsa* during development: implications for determination of copepod production. Mar Biol. 99(3):341–352.

Bernt M, et al. 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 69(2):313–319.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Broglio E, Jónasdóttir SH, Calbet A, Jakobsen HH, Saiz E. 2003. Effect of heterotrophic versus autotrophic food on feeding and reproduction of the calanoid copepod *Acartia tonsa*: relationship with prey fatty acid composition. Aquat Microb Ecol. 31:267–278.

Bron JE, et al. 2011. Observing copepods through a genomic lens. Front Zool. 8(1):22.

Chen G, Hare MP. 2008. Cryptic ecological diversification of a planktonic estuarine copepod, *Acartia tonsa*. Mol Ecol. 17(6):1451–1468.

Combs PA, Eisen MB. 2015. Low-cost, low-input RNA-seq protocols perform nearly as well as high-input protocols. PeerJ. 3:e869.

Costello MJ. 2006. Ecology of sea lice parasitic on farmed and wild fish. Trends Parasitol. 22(10):475–483.

Drillet G, Goetze E, Jepsen PM, Højgaard JK, Hansen BW. 2008. Strain-specific vital rates in four *Acartia tonsa* cultures, I: strain origin, genetic differentiation and egg survivorship. Aquaculture 280(1–4):109–116.

Drillet G, Hansen BW, Kiørboe T. 2011. Resting egg production induced by food limitation in the calanoid copepod *Acartia tonsa*. Limnol Oceanogr. 56(6):2064–2070.

Drillet G, et al. 2006. Effect of cold storage upon eggs of a calanoid copepod, *Acartia tonsa* (Dana) and their offspring. Aquaculture 254(1-4):714–729.

Drillet G, et al. 2015. Total egg harvest by the calanoid copepod *Acartia tonsa* (Dana) in intensive culture—effects of high stocking densities on daily egg harvest and egg quality. Aquac Res. 46(12):3028–3039.

Dunn CW, Ryan JF. 2015. The evolution of animal genomes. Curr Opin Genet Dev. 35:25–32.

Evans JD, et al. 2013. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered. 104:595–600.

Eyun SI. 2017. Phylogenomic analysis of Copepoda (Arthropoda, Crustacea) reveals unexpected similarities with earlier proposed morphological phylogenies. BMC Evol Biol. 17:1–12.

Francis WR, Wörheide G. 2017. Similar ratios of introns to intergenic sequence across animal genomes. Genome Biol Evol. 9:1582–1598.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7):644–652.

Gregory TR, Hebert PDN, Kolasa J. 2000. Evolutionary implications of the relationship between genome size and body size in flatworms and copepods. Heredity (Edinb). 84(2):201–208.

Han J, Puthumana J, Lee M, Kim S, Lee J. 2016. Different susceptibilities of the Antarctic and temperate copepods Tigriopus kingsejongensis and T . japonicus to ultraviolet (UV) radiation. Mar. Ecol. Prog. Ser. 561:99–107.

Hansen BW, Buttino I, Cunha ME, Drillet G. 2016. Embryonic cold storage capability from seven strains of *Acartia* spp. isolated in different geographical areas. Aquaculture 457:131–139.

Hansen BW, Hansen PJ, Nielsen TG, Jepsen PM. 2017. Effects of elevated pH on marine copepods in mass cultivation systems: practical implications. J Plankton Res. 39(6):984–993.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491.

Humes AG. 1994. How many copepods? Hydrobiologia 293:1–7.

Jepsen PM, Andersen CVB, Schjelde J, Hansen BW. 2015. Tolerance of unionized ammonia in live feed cultures of the calanoid copepod *Acartia tonsa* Dana. Aquac Res. 46(2):420–431.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.

Katoh K, Toh H. 2008. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. BMC Bioinformatics 9: 1–13.
Kent WJ. 2002. BLAT—The BLAST-like alignment tool. Genome Res. 12(4):656–664.
Khodami S, McArthur JV, Blanco-Bercial L, Martinez Arbizu P. 2017. Molecular phylogeny and revision of copepod orders (Crustacea: Copepoda). Sci Rep. 7(1):9164.
Laakmann S, et al. 2013. Comparison of molecular species identification for North Sea calanoid copepods (Crustacea) using proteome fingerprints and DNA sequences. Mol Ecol Resour. 13(5):862–876.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860–921.
Leinaas HP, Jalal M, Gabrielsen TM, Hessen DO. 2016. Inter- and intraspecific variation in body- and genome size in calanoid copepods from temperate and arctic waters. Ecol Evol. 6(16):5585–5595.
Lindgreen S. 2012. AdapterRemoval: easy cleaning of next generation sequencing reads. BMC Res Notes 5(1):337.
Madoui MA, et al. 2017. New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. Mol Ecol. 26(17):4467–4482.
Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. Genomics. 95(6):315–327.
Nilsson B, Hansen BW. 2018. Timing of embryonic quiescence determines viability of embryos from the calanoid copepod, *Acartia tonsa* (Dana). PLoS One 13:1–16.
Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:61–65.
Rasch EM, Lee CE, Wyngaard GA. 2004. DNA–Feulgen cytophotometric determination of genome size for the freshwater-invading copepod *Eurytemora affinis*. Genome 47(3):559–564.
Renaut S, et al. 2018. Hybrid de novo assembly of the draft genome of the freshwater mussel *Venustaconcha ellipsiformis* (Bivalvia: Unionida). Genome Biol Evol. 10(7):1637–1646.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19(12):1572–1574.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.
Simpson J, Durbin R. 2012. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 22(3):549–556.
Smit A, Hubley R. 2019. RepeatModeler Open-1.0. Available from: http://www.repeatmasker.org.
Smit A, Hubley R, Green P. 2019. RepeatMasker Open-4.0. Available from: http://www.repeatmasker.org.
Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32:309–312.
Støttrup JG, Richardson K, Kirkegaard E, Pihl NJ. 1986. The cultivation of *Acartia tonsa* Dana for use as a live food source for marine fish larvae. Aquaculture 52(2):87–96.
Tiselius P, et al. 1995. Can we use laboratory-reared copepods for experiments? A comparison of feeding behaviour and reproduction between a field and a laboratory population of *Acartia tonsa*. ICES J Mar Sci. 52(3-4):369–376.
Turner JT. 2004. The importance of small pelagic planktonic copepods and their role in pelagic marine food webs. Zool Stud. 43:255–266.
Vaidya G, Lohman DJ, Meier R. 2011. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics 27(2):171–180.
Wendt I, Backhaus T, Blanck H, Arrhenius Å. 2016. The toxicity of the three antifouling biocides DCOIT, TPBP and medetomidine to the marine pelagic copepod *Acartia tonsa*. Ecotoxicology 25(5):871–879.
Xue W, et al. 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics. 14(1):604.
Yang Z. 1994. Estimating the pattern of nucleotide substitution. J Mol Evol. 91:105–111.

Associate editor: Liliana Milani