A genome database for a Japanese population of the larvacean Oikopleura dioica

doi: info:doi/10.1111/dgd.12689

Development, Growth & Differentiation
The Japanese Society of Developmental Biologists

A genome database for a Japanese population of the larvacean *Oikopleura dioica*

Journal:	*Development Growth and Differentiation*
Manuscript ID	DGD-00077-2019.R2
Manuscript Type:	Original Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Wang, Kai ; Osaka University, Biological Science Tomura, Ryo; Osaka University, Biological Science Chen, Wei ; National Institute of Genetics Kiyooka, Miho ; National Institute of Genetics Ishizaki, Hinako ; National Institute of Genetics Aizu, Tomoyuki ; National Institute of Genetics Minakuchi, Yohei ; National Institute of Genetics Seki, Masahide ; The University of Tokyo Graduate School of Frontier Sciences Suzuki, Yutaka ; The University of Tokyo Graduate School of Frontier Sciences Omotezako, Tatsuya ; Osaka University, Biological Science Suyama, Ritsuko ; Okinawa Institute of Science and Technology Graduate University Masunaga, Aki; Okinawa Institute of Science and Technology Graduate University Plessy, Charles ; Okinawa Institute of Science and Technology Graduate University Luscombe, Nicholas ; Okinawa Institute of Science and Technology Graduate University Dantec, Christelle; Centre de Recherche de Biochimie Macromoleculaire Lemaire, Patrick; Centre National de la Recherche Scientifique, CRBM Itoh, Takehiko ; Tokyo Institute of Technology Toyoda, Atsushi; National Institute of Genetics Nishida, Hiroki; Dept. Biol., Grad. Sch. Sci., Onuma, Takeshi; Osaka University, Biological Science
Key Words:	genome, transcriptome, gene model, larvacean, Y-chromosome

Development Growth and Differentiation (2^{nd} revision)

A genome database for a Japanese population of the larvacean *Oikopleura dioica*

Kai Wang[1], Ryo Tomura[1], Wei Chen[2], Miho Kiyooka[2], Hinako Ishizaki[2], Tomoyuki Aizu[2], Yohei Minakuchi[2], Masahide Seki[3], Yutaka Suzuki[3], Tatsuya Omotezako[1], Ritsuko Suyama[4], Aki Masunaga[4], Charles Plessy[4], Nicholas M. Luscombe[4], Christelle Dantec[5], Patrick Lemaire[5], Takehiko Itoh[6], Atsushi Toyoda[2], Hiroki Nishida[1] and Takeshi A. Onuma*[1]

[1] *Department of Biological Sciences, Graduate School of Science, Osaka University, 1-1 Machikaneyama-cho, Toyonaka, Osaka 560-0043, Japan*

[2] *Comparative Genomics Laboratory, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan*

[3] *Laboratory of Systems Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan*

[4] *Genomics and Regulatory Systems Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Kunigami-gun, Okinawa 904-0495, Japan*

[5] *Centre de Recherches de Biochimie Macromoleculaire (CRBM), UMR5237, CNRS-Universite de Montpellier, 1919 route de Mende, F-34090 Montpellier, France*

[6] *School of Life Science and Technology, Tokyo Institute of Technology, Meguro-ku, Tokyo, 152-8550, Japan.*

4 Tables, 2 Figures, 3 Supplemental Tables, and 2 Supplemental Materials.

Running title: Japanese *O. dioica* genome

*Correspondence

Tel.: +81-6-6850-5472

Fax: +81-6-6850-5472

E-mail: takeo@bio.sci.osaka-u.ac.jp

Abstract

The larvacean *Oikopleura dioica* is a planktonic chordate, and is tunicate that belongs to the closest relatives to vertebrates. Its simple and transparent body, invariant embryonic cell lineages, and short life cycle of five days make it a promising model organism for developmental biology research. The genome browser OikoBase was established in 2013 using Norwegian *O. dioica*. However, genome information for other populations is not available, even though many researchers have studied local populations. In the present study, we sequenced using Illumina and PacBio RSII technologies the genome of *O. dioica* from a southwestern Japanese population that was cultured in our laboratory for three years. The genome of Japanese *O. dioica* was assembled into 576 scaffold sequences with a total length and N50 length of 56.6 Mb and 1.5 Mb, respectively. A total of 18,743 gene models (transcript models) were predicted in the genome assembly, named as OSKA2016. In addition, 19,277 non-redundant transcripts were assembled using RNA-seq data. The OSKA2016 has global sequence similarity of only 86.5% when compared with the OikoBase, highlighting the sequence difference between the two far distant *O. dioica* populations on the globe. The genome assembly, transcript assembly, and transcript models were incorporated into ANISEED (https://www.aniseed.cnrs.fr/) for genome browsing and blast searches. Moreover, screening of the male-specific scaffolds revealed that over 2.6 Mb of sequence were included in the male-specific Y-region. The genome and transcriptome resources from two distinct populations will be useful datasets for developmental biology, evolutionary biology, and molecular ecology using this model organism.

Key words: genome, transcriptome, gene model, larvacean, Y-chromosome

1. Introduction

The larvacean (also known as appendicularian) *Oikopleura dioica* (*O. dioica*) is a planktonic chordate that retains the notochord and tadpole morphology throughout its entire life. Together with ascidians and thaliaceans, it belongs to tunicate, which is the closest invertebrate relative to vertebrates (Bourlat et al., 2006; Delsuc et al., 2006, 2018). This animal has a number of advantages as an experimental organism, such as a rapid developmental rate to complete morphogenesis in 10-h after fertilization, a small number of cells (approximately 4000 in a functional juvenile), and a transparent body (Nishida, 2008). Its embryonic cell lineage and fate map are well-characterized, and the fate map is comparable to those of ascidians and vertebrates (Fujii et al., 2008; Nishida, 2008; Stach et al., 2008; Nishida and Stach, 2014). The genome size of *O. dioica* has been estimated to be 70 Mb, the smallest genome described in non-parasitic metazoans to date (Seo et al., 2001; Denoeud et al., 2010; Naville et al., 2019). In addition, *O. dioica* is the sole dioecious species in tunicate, and the presence of a male-specific Y-chromosome has been reported (Denoeud et al., 2010; Navratilova et al., 2017).

The genome sequence of *O. dioica* was first obtained from a Norwegian population (Seo et al., 2001; Denoeud et al., 2010; Danks et al., 2013). It is available in a genome browser with predicted gene models, transcript models, expression sequence tags (ESTs), and microarray-based gene expression profiles (OikoBase, Danks et al. 2013). However, the genome sequences of other *O. dioica* populations are unavailable. Here, we determined the genome sequence of Japanese *O. dioica*. Using the same population, we have developed tools for molecular, cellular, and genetic techniques, including gene knockdown methods (Omotezako et al., 2013, 2015, 2017), fluorescent live imaging of larvae (Kishi et al., 2014, 2017), and maternal and zygotic transcriptomes of the eggs and larvae (Wang et al., 2015). These have been used to gain insights into cell migration (Kishi et al., 2014), patterning of epidermal cells (Kishi et al., 2017), left-right patterning (Onuma et al., 2020) and meiotic arrest of the unfertilized egg (Matsuo et al., 2020). Moreover, transcriptome-wide comparison has implied a high-level of sequence variation between the Japanese and Norwegian *O. dioica* (Wang et al., 2015): only 91.0 and 94.8% of exon sequences were conserved at nucleotide and amino acid levels on average, respectively. Having fully accessible genome information for two distinct geographical populations will contribute to the study of chordate development and to understand genome plasticity in this rapidly evolving tunicate.

In the present study, we generated the genome assembly of Japanese *O. dioica* using next-generation sequencing (NGS) and third-generation single molecule real-time (SMRT) sequencing. Accordingly, transcriptome information was also improved from the previous version (Wang et al., 2015). This genome assembly, named as OSKA2016, was compared with the OikoBase to estimate sequence variation between the two *O. dioica* populations. These resources have been incorporated into a publicly available genome database of tunicates, ANISEED (https://www.aniseed.cnrs.fr/) (Brozovic et al., 2018; Dardaillon et al. 2020), an open database for developmental, evolutionary, and comparative genome research involving this small chordate species. Using the genome scaffolds, we lastly tried to characterize male-specific scaffolds and gain insight into the male-specific Y-region of the Japanese *O. dioica*.

2. Materials and Methods

2.1 Genome materials, library construction, and sequencing

O. dioica were collected at Sakoshi Bay and Tossaki Port in Hyogo, Japan, and cultured for over 180 generations for more than three years in the inland laboratory, as previously described (Nishida, 2008; Omotezako et al., 2013). Animals were cultured at 20ºC. Under this condition, *O.dioica* hatch 3 h after fertilization (hpf) and complete organogenesis and form fully functional bodies by 10 hpf. Animals mature sexually and spawn on the fifth day.

Genomic DNA was extracted from sperm. Sperm was collected from a laboratory population derived from a single pair. Ordinarily, one female spawns 200-300 eggs. They develop into male and female adults at same ratio. The population was inbred for three generations to reduce allelic variation, and the cohorts were expanded for collection of specimens. More than 800 males were used to extract approximately 30 μg of genomic DNA (average fragment size was > 40 Kb). Three templates were generated for sequencing. First, genomic DNA was fragmented into an average size of 600 bp with the DNA Shearing System S220 (Covaris Inc., MA, USA). A paired-end library was constructed with a TruSeq DNA PCR-Free Library Prep kit (Illumina, CA, USA) and was size-selected on an agarose gel using a Zymoclean Large Fragment DNA Recovery Kit (Zymo Research, CA, USA). The final library was sequenced on the Illumina HiSeq 2500 sequencer with a read length of 250 bp. Next, a SMRTbell library was constructed using a SMRTbell Template Prep Kit v1.0 according to manufacturer's protocol (Pacific Bioscience, CA, USA). The sequencing library was size-selected using the BluePippin system (Saga Science, MA, USA) with a minimum fragment length cutoff of 15 kbp. Ten SMRT cell v3 were run on the PacBio RSII System with P6/C4 chemistry and 360 min movies. Lastly, construction and mate-pair library with inserts of approximately 5 Kb and Illumina sequencing were performed by Beijing Genome Institute (BGI Japan).

These sequence data have been deposited to the DNA databank of Japan (DDBJ) (accession number DRA010224).

2.2 Transcriptome materials and sequencing

Total RNA was collected from the same population used for the genome sequencing. Samples were collected in two ways. First, specimens were collected at 13 developmental stages (unfertilized egg, fertilized egg, embryos at the two-cell, four-cell, eight-cell, 16-cell, 32-cell, and 1.5 hours-post-fertilization (hpf) stages, and 2 hpf (tailbud embryo), 3 hpf (hatched), 5 hpf (early organogenesis), 8 hpf (late organogenesis), and 10 hpf (juvenile with functional body)) and from mature males or females. Embryos and larvae are reared at 20˚C. They were subjected to extraction of total RNA, and preparation of paired-end libraries with SureSelect Agilent Strand Specific RNA library prep kit. Libraries were sequenced using Illumina HiSeq 2500 with a read length of 100 bp. Second, animal and vegetal hemispheres of the eight-cell embryos were dissected and collected immediately upon dissection or after 12 hours at 11˚C (corresponds to 7 hpf at 20˚C). They were subjected to RNA-seq following an established protocol (Tang et al., 2010). Two to four replicates were prepared, and 0.7–9.0 million bases were sequenced from each library. These read sequences were used for transcriptome assembly described in Materials & Methods 2.3. Gene expression analysis using these reads will be published elsewhere.

2.3 Genome and transcript assemblies

For genome assembly, Illumina sequence data were assembled by Platanus v124b (PLATform for Assembling NUcleotide Sequences) (Kajitani et al., 2014). In addition, PacBio RS II sequence data were utilized to fill gaps between contigs using SMRT analysis v2.3.0 BridgeMapper. Other programs, including Falcon (Chin et al., 2013) and MuSuRCA (Zimin et al., 2013), were also used to consider assemblies obtained with different programs.

For transcriptome assembly, the aforementioned RNA-seq reads were pooled with previous RNA-seq reads (Wang et al., 2015) and genome-guided *de novo* assembly was carried out using Trinity (Haas et al., 2013) after adaptor trimming. Highly redundant sequences were removed. Hereafter, we call this assembled transcriptome "transcript assembly" to distinguish it from the "transcript model" (see below).

2.4 Gene prediction and annotation

Prediction of protein-coding genes was carried out using three different programs. SNAP v2006-07-28 (Korf, 2004) and Augustus v3.2.1 (Stanke and Morgenstern, 2005) were used to predict genes based on mapping results of the transcript assembly onto the genome assembly, and a self-training program, GeneMark v4.38 (Lukashin and Borodovsky, 1998), was also used. These three datasets were combined by MAKER v2.31.10 (Holt and Yandell, 2011), which was carried out following an opened pipeline (URL: https://reslp.github.io/blog/My-MAKER-Pipeline/). Hereafter, we call the predicted genes "transcript model" and "protein model".

2.5 Assessment of assembly using BUSCO

The genome assembly was assessed by Benchmarking Universal Single Copy Orthologs (BUSCO) (Simão et al., 2015). BUSCO searched for 303 genes, which are highly conserved in Eukaryotes (eukaryota_odb9), from the targets sequences, and evaluated the percentages that were covered by the assembly. BUSCO was carried out for all the assemblies generated by different assembly programs for comparison.

2.6 Comparison of the genome assembly with the OikoBase

The genome assembly was compared reciprocally (bi-directionally) with the scaffolds in the OikoBase (Danks et al., 2013) in three ways. First, we tested what percentage of scaffolds in the OikoBase could be recovered using the dnadiff tool from MUMmer v3.22 (Kurtz et al., 2004). Gap sequences (Ns) were removed before comparison. Second, percentages of sequence similarity and gap sequences of exons, introns and intergenic regions were estimated for three genes, *Brachury*, *Bmp.a*, and *Pax6*. These genes have been identified as plausible orthologous genes in both Japanese and Norwegian *O. dioica* (Omotezako et al., 2013; Danks et al., 2013; Onuma et al., 2020). Their gene IDs are summarized in Table 3. Third, BLAT searches (v35) were carried out to calculate what percentage of the transcript and protein models could be mapped to the genome sequences. Sequences that were regarded as the top hits were used for calculation.

2.7 Genome browser

Genome scaffolds, predicted genes, and transcriptome assembly were incorporated into ANISEED, a web-based and publicly available genome database (Brozovic et al., 2018; Dardaillon et al., 2020). It includes the genome information of 13 ascidian species as well as larvacean species (present data). This database enables us to visualize various features of these ascidian species, including genome scaffolds, gene features, annotation, and orthology.

2.8 Scaffolds corresponds to the male-specific Y-region

To characterize the male-specific Y-region, genomic DNA was extracted from mature males or females. Construction of pair-end libraries and sequencing were performed by BGI Japan. The Illumina HiSeq 4000 generated 39.1 Gb and 8.7 Gb of raw reads from the male and female libraries, respectively. The read size was 125 bp. These reads were mapped against the genome scaffolds for screening of male-enriched scaffolds. To confirm proximity of the male-enriched scaffolds, we mapped them onto a genome assembly produced from high-molecular weight DNA extracted from a single male individual and sequenced on a MinION instrument (Oxford Nanopore Technologies, ONT). This male is originated from the laboratory strain used to sequence the OSKA2016 genome, and was cultivated at 20 °C before sampling (Masunaga et al. 2020). Genomic DNA was extracted from the whole animal using a standard phenol-chloroform method, and was prepared for sequencing using a SQK-LSK109 kit (ONT). The raw signals were converted to sequence reads with the Guppy software version 3.3.0 using the dna_r9.4.1_450bps_hac model, and the reads were assembled using the Flye software (Kolmogorov et al., 2019) version 2.7 with a predicted genome size of 65 Mbp and a minimum overlap of 3000 bp between reads. An optimal set of alignments to the OSKA2016 genome was searched with the last-split software (Frith and Kawaguchi, 2015). These sequences, assemblies and alignments have been deposited in online repositories and will be published (Plessy et al., in preparation).

To detect the male-specific Y-region, genomic PCR was carried out. Mature males and females can be distinguished based on the morphology of the gonads (Nishida, 2008). The tail was manually dissected with a razor blade and subjected to analysis. Hatched larvae, which have unclear sexes based on appearance, were also used. For extraction of genomic DNA, samples were incubated with 20 µl of genome extraction buffer (10 mM EDTA, 10 mM Tris-EHCl, pH 8.0, and 0.2 mg of Protease K) at 50˚C for 3h, and then at 95°C for 10 min. 1 µl of the DNA was subjected to genomic PCR using Prime STAR^{GXL} (TAKARA). PCR was carried out with the following conditions: 98°C for 1 min, and 35 or 40 cycles of 98°C for 10 sec, 55°C for 15 sec, and 68°C for 30 sec. For long-range PCR, with product lengths longer than 10 Kb, the following condition was adopted: 98°C for 1 min, and 40 cycles of 98°C for 10 sec and 68°C for 30 min.

3. Results and Discussion

3.1 Genome assemblies of Japanese *O. dioica*

The genome assembly of Japanese *O. dioica* was generated through a hybrid program of Platanus (Illumina reads) and SMRT analysis BridgeMapper (PacBio RS II reads). Illumina HiSeq sequencing generated 217 million reads for the pair-end library (read size 250 bp, data size 54.2 Gb) and 49.8 million reads for the mate-pair library (read size 100 bp, insert length 5 Kb, data size 6.2 Gb). PacBio sequencing generated 1.2 million reads (data size 13.2 Gb). Therefore, the sequence coverage of the Japanese *O. dioica* genome was more than 1000×, considering the genome size (see below). This genome assembly is referred to as "OSKA2016".

The OSKA2016 is comprised of 576 scaffolds (length >= 1 Kb). The GC content was 41.4%. The N50 lengths of the scaffolds and contigs are 1.5 Mb and 0.62 Mb, respectively (Table 1 and Table S1). The longest scaffold was 6.8 Mb in length. The N50 length of the scaffolds and contigs are 3.8-

fold and 25-fold of those in the OikoBase, respectively (Table 1). The gap rate (Ns) was 0.4%. Numbers of scaffolds (576) and gaps (331,456 bp) in the present assembly are 46% and 8.4%, respectively, of the OikoBase assembly (build name: Oikopleura_reference_unmasked_v3.0) (http://oikoarrays.biology.uiowa.edu/Oiko/, Danks et al., 2013) (Table 1). These data indicate that OSKA2016 was better assembled than the OikoBase genome assembly in some regards.

The total length of the OSKA2016 genome assembly was estimated to be 56.6 Mb. It is smaller than that in the Oikobase (70 Mb) (Table 1). The sequence data in the OikoBase were collected by different techniques, i.e., shotgun sequencing and bacterial artificial chromosomes (Seo et al., 2001; Denoued et al., 2010; Danks et al., 2013). Also, highly repetitive sequences may affect the size of genome assembly (Naville et al., 2019). Our result is reminiscent of the genome size estimation in an ascidian *Ciona intestinalis* Type A (*C. robusta*) (Satou et al., 2019; Shoguchi et al., 2006). In this species, the genome size of the latest genome assembly had been determined as 123 Mb (Satou et al., 2019). This is smaller than the previous estimation (160 Mb) (Shoguchi et al., 2006).

3.2 Assessment of assembly and gene prediction

The quality of the OSKA2016 assembly was verified by making and comparing assemblies using different assembly programs/pipelines. Three representative results are shown in Table S1. Moreover, the completeness of the genome assemblies was assessed by BUSCO (Simão et al., 2015). BUSCO is a dataset of near-universal single copy orthologs, and can be used to assess completeness of newly sequenced genome. As shown in Table S2, BUSCO returned 87.2% completion for OSKA2016. This percentage is higher than other assemblies generated by different programs, namely Falcon or MaSuRCA (Table S2). We thus adopted the OSKA2016 as the best genome assembly for Japanese *O. dioica*. It is reasonable that less than 90% of BUSCO was recovered, because *O. dioica* is shown to have the smallest and most highly rearranged genome amongst non-parasitic metazoans (Seo et al., 2001; Denoeud et al., 2010), and often lacks evolutionarily-conserved genes, such as those for retinoic acid signal (Martí-Solans et al., 2016), Nodal signal (Onuma et al., 2020), and the non-homologous DNA-end joining pathway (Deng et al., 2018).

3.3 Latest version of transcriptome assembly

Our previous RNA-seq study identified 12,311 transcripts using samples collected at two developmental stages (unfertilized egg and late organogenesis larvae) (Wang et al., 2015). Search of similar sequences showed that these transcripts hit more than 95% of genes predicted in the OikoBase (Wang et al., 2015). However, the number, 12,311, is clearly less than that of protein-coding genes, which has been estimated to be 17,000-18,000 in the Oikobase (Seo et al., 2001; Denoeud et al., 2010; Danks et al., 2013), implying that there remains unassembled transcripts, probably due to transient expression or low expression levels. To improve this, we collected RNA-seq data more comprehensively from the following sources: 13 different developmental stages from egg to functional juveniles, mature males and females, and the animal/vegetal hemisphere of embryos. Genome-guided *de novo* assembly by Trinity and subsequent removal of redundant sequences finally yielded 19,277 transcripts with an average length of 1,375 bp ("Transcript assembly" in Table 1).

3.4 Gene prediction and generation of "transcript models"

Gene prediction was carried out using the MAKER2 (Holt and Yandell, 2011) pipeline (see Materials and Methods) based on the genome assembly. Mapping results of the transcript assembly onto the genome assembly were also used for a guide of gene prediction. A total of 18,743 gene models, which are regarded as "Transcript models"/"Protein models", were predicted in the OSKA2016 (Table 1). The average lengths of transcripts and amino acids were 1,309 bp and 393 aa, respectively. The number and average length of the transcript models are comparable with the "Transcript assembly" (Table 1).

3.5 Reciprocal searches between Japanese and Norwegian *O. dioica* genomes

To test the characteristics of the genome assembly, reciprocal searches were conducted between the OSKA2016 and OikoBase scaffolds using MUMmer v3.22. As shown in Table 2, 76.6% (441) and 91.2% (1,149) of scaffolds in the OSKA2016 and OikoBase were shared, respectively. Intriguingly, 44,685 distinct regions were aligned, with an overall sequence similarity of 86.5% (Table 2). Next, nucleotide sequences of three genes, *Brachury*, *Bmp.a* and *Pax6*, were compared manually (Table 3). Sequence similarities in the exon, intron and intergenic regions of *Brachury* were 92%, 82% and 82%, respectively. Moreover, rate of gaps was less in the exons (0.79%) than that in the introns (3.3%) and intergenic regions (7.2%). Likewise, sequence similarities of the exons were higher than those in the introns and the intergenic regions of the *Bmp.a* and *Pax6* (Table 3). These results suggest that there might be an abundance of insertions/deletions and/or more global differences between the genomes of the two *O. dioica* populations.

Next, we conducted BLAT searches of gene models against genome scaffolds. Of the 18,743 gene models in OSKA2016, 80.8% of the "transcript model" and 84.9% of the "protein model" coincided with the OikoBase scaffolds (Oikopleura_reference_unmasked_v3.0) when the coverage length percentage was set to more than 50% (Table 4). In comparison, 17,212 gene models in the OikoBase matched with the OSKA2016 scaffold at higher rates (87.2% and 91.0% of transcripts and proteins, respectively). These differences indicate that OSKA2016 predicts a larger number of transcript models because it incorporates more comprehensive RNA-seq data.

Based on these analyses, the predicted gene models, i.e., transcript models, are of sufficient quality for developmental, genetic, evolutionary, and ecological studies. They would cover most of the transcripts expressed during oogenesis, embryogenesis, and larval morphogenesis, and provide a valuable tool for developmental biology, such as making multiple *in situ* hybridization probes (Onuma et al., 2017) and designing knockdown experiments (Omotezako et al., 2013, 2015, 2017).

3.6 Genome browser construction on ANISEED

The OSKA2016 genome assembly, transcript models, and transcript assemblies were released as a genome browser in ANISEED (A tunicate database, https://www.aniseed.cnrs.fr/). Here, we briefly introduce several important links:

Home (Fig. 1a): Links to go to the BLAST and genome browser are indicated with blue and red rectangles, respectively.

Blast search (Fig. 1b): People can use a BLAST search to identify sequence information of interest. Blastn, tblastx, and tblastn are available as BLAST programs. The database lists the genome assembly and gene models (Transcript models). To get a link for the BLAST hits in the genome browser, the user must choose “genome assembly” as the database.

Genome browser (Fig. 1c): Choose the “Appendicularia” tab and select “OSKA2016 genome browser” (Fig. 1c). Click on the “Tracks” link, then “ANNOTATION tracks”, to show and compare the organizations of “Transcript models” and “Transcript assemblies”. To retrieve part of a genome sequence, click on the “Apps” link, then “More applications” and “Get sequence”, and input the ID of a scaffold, the start position you want to see, and the stop position. For instance, region 10000-20000 of the scaffold009 should be input as “S9:10000-20000”.

Gene Card (Fig. 1d): Click on the transcript model in the genome browser, then select the “Gene Card” link (yellow rectangle, Fig. 1c). Users can see annotated characteristics such as Gene ontology and InterPro annotations, if any. Click “Transcript Models and Sequences” at the top. Nucleotide and amino acid sequences can be retrieved in a FASTA format (Fig 1d). These data will be linked to other chordate resources.

3.7 Characterization of male-specific Y-scaffolds in the OSKA2016

Lastly, we used the OSKA2016 genome assembly for characterization of the male-specific chromosome. *O. dioica* is the sole dioecious species amongst tunicate species, and is reported to have a male-specific Y-region connected to a pseudo-autosomal region (Denoeud et al., 2010; Navratilova et al., 2017). The Y-specific region is reported to be a “gene desert” and contains seven giant male-specific genes (Denoeud et al., 2010). However, little information has been available on the male-specific genes. Characterization of the Y-specific scaffolds in another *O. dioica* population will be a useful resource for future studies on genetic and evolutionary studies of sex determination.

First, we isolated an EST clone encoding a putative Y-specific gene in the OikoBase (clone name: KT0AAA379YD22.CONTIG). It is 1,495 bp in length, but its encoding protein had no homology with known protein domains. Using this EST sequence, we isolated homologous partial cDNA clone that is 1,178 bp in length (Supplemental Material 1). A blast search using this cDNA as a query sequence indicated that this Y-specific gene is spread across scaffolds of OSKA2016 (Fig. 2a). Two scaffolds (S), S039 and S064, appeared to include the putative Y-specific exons and to be next to each other (Fig. 2a).

To test whether these scaffolds are male-specific, genomic PCR was carried out. Two different primer pairs targeting S039 and 064 detected PCR products from matured male, but not female, genomic DNA (Fig. 2b, d). Moreover, we randomly selected larvae, for which sexes are difficult to distinguish from appearance alone, and analyzed whether these scaffolds inherit randomly into cohorts. As shown in Fig. 2c, PCR signals were detected in about half of the larvae (5/10). By contrast, two primer pairs targeting different positions in the S009 amplified PCR products in both males and females (Fig. 2d).

Next, Y-specific scaffolds in the OSKA2016 were screened. Illumina sequencing was carried out using male- or female-specific genome libraries. Mapping of reads against the OSKA2016 assembly identified 20 male-enriched scaffolds, for which more than two-fold reads were mapped in the males (Fig. 2e). We noted that the “male to female ratio” of mapped reads did not show an “all-or-none” difference (Fig. 2e) because Y-specific region is enriched in repeated sequences such as transposable elements (Denoeud et al., 2010). Nonetheless, the 20 male-enriched scaffolds included S039 and S064 (Fig. 2e). Blast searches using S039 and S064 hit scaffold8 from the OikoBase. Scaffold8 is reported to be located in the terminus of the Y-specific region (Denoeud et al., 2010). To test whether these male-enriched scaffolds can be grouped into a single continuous sequence, they were mapped onto an ultra-long genome draft constructed using Oxford Nanopore sequencing reads from a single male individual. In this dataset, the Y-specific region is haploid, and could be assembled in a single sequence (Contig_4), on which 18 out of the 20 male-enriched scaffolds were grouped (Fig. 2f) The deduced order of the male-enriched scaffolds of OSKA2016 was summarized in Fig. 2g.

Nucleotide sequence of the Contig_4 was approximately 2.6 Mb in length (Supplemental Material 2). In comparison, rough calculation of the Y-specific region based on the previous report is around 4.9 Mb in the OikoBase (Denoeud et al., 2010). The reason of this discrepancy is not clear. Some chromosomal regions enriched in highly repetitive sequences that present outside of the Contig_4 could be overlooked in the present analysis. Alternatively, there could be global differences in the Y-specific region between Norwegian and Japanese populations. We also note that several male-enriched scaffolds, such as S026, S039, S048 and S055, were aligned with two regions in the Contig_4 (arrows in Fig, 2f), suggesting a possibility of variation between individuals and/or mis-assembly of the OSKA2016.

4. Summary and conclusion

Since the genome sequence of Norweigian *O. dioica* was read in 2001, most larvacean genome research has been dependent on OikoBase (Seo et al., 2001; Denoeud et al. 2010; Danks et al., 2013). The data in OikoBase are informative and useful for cloning and studying genes of interest. However, *O. dioica* is distributed in oceans around the world, and researchers use different geographical populations. Additionally, intra-species sequence variations have been evident (Wang et al., 2015 and present study), and might affect detailed analysis such as Cas9-mediated gene knockout (Deng et al., 2018). The present study enables researchers to use genome information of two distinct geographical populations. The OSKA2016 genome assembly, transcriptome assembly, and transcript models stored in ANISEED will help researchers using larvacean species conduct detailed analysis in the fields of developmental and evolutionary biology.

Acknowledgments

We thank Ms. Masae Suzuki, Ms. Keiko Kayahara, and several students at Osaka University for helping with the laboratory culture of *O. dioica*. We also appreciate the staff of the Comparative Genomics Laboratory at the National Institute of Genetics (NIG) and the Scientific Computing section of Research Support Division at OIST for supporting genome sequencing, Ms. Kiyomi Imamura and Ms. Terumi Horiuchi for their support in the collection of RNA-seq data, Dr. Shunsuke Yaguchi for his advice regarding making a genome browser, Dr. Garth R. Ilshley for critical reading of the manuscript, and Editage (www.editage.com) for English language editing. Computations were partially performed on the NIG supercomputer at the ROIS National Institute of Genetics.

Author contributions

TAO and HN designed the project overall. TAO collected genomic DNA and total RNA. WC, MK, HI, TA, TI, and AT conducted genome sequencing, *de novo* assembly, assessment of assemblies, and gene prediction. MS and YS collected staged RNA-seq data. OT, RS, and NML carried out embryo dissection and small RNA-seq. KW generated transcriptome assembly. KW, CD, and PL annotated genes and constructed the genome browser in ANISEED. TAO, KW, RT, CP and AM carried out analyses of the Y-chromosome. TAO and HN wrote the manuscript.

Funding information

This work was supported by JSPS KAKENHI Grant Number 16H06279, MEXT KAKENHI (No. 221S0002), Grants-in-Aid for Scientific Research from the JSPS to TAO (24870019, 26840079, 18K06256) and HN (26650079, 15H04377, 16K14735, 17KT0023, 19H03234), the Inamori Foundation (2015), the Kato Memorial Research Foundation (2015–2016), the Sumitomo Foundation (2016-2017), and the Japan Foundation for Applied Enzymology (2017, 2019) to TAO.

ORCID

Charles Plessy https://orcid.org/0000-0001-7410-6295

Takeshi A. Onuma https://orcid.org/0000-0002-9739-6333

Aki Masunaga https://orcid.org/0000-0002-6913-8417

References

Bourlat, S.J., Juliusdottir, T., Lowe, C.J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E.S., … Telford, M.J. (2006). Deuterostome phylogeny reveals monophyletic chordates and the new phylum *Xenoturbellida*. Nature, 444, 85–88.

Brozovic, M., Dantec, C., Dardaillon, J., Dauga, D., Faure, E., Gineste, M., Louis, A., … Lemaire, P. (2018). ANISEED 2017: Extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets. Nucleic Acids Research, 46, D718–D725.

Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., … Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods, 10, 563–569.

Danks, G., Campsteijn, C., Parida, M., Butcher, S., Doddapaneni, H., Fu, B., Petrin, R., … Manak, J.R. (2013). OikoBase: A genomics and developmental transcriptomics resource for the urochordate Oikopleura dioica. Nucleic Acids Research, 41, 1–9.

Dardaillon, J., Dauga, D., Simion, P., Faure, E., Onuma, T.A., DeBiasse, M., Louis, A. , … Lemaire, P. (2020). ANISEED 2019: 4D exploration of genetic data for an extended range of tunicates. Nucleic Acids Research, 48, D668-D675.

Delsuc, F., Brinkmann, H., Chourrout, D., Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature, 439, 965–968.

Delsuc, F., Philippe, H., Tsagkogeorga, G., Simion, P., Tilak, M.K., Turon, X., López-Legentil, S.,

… Douzery, E.J.P. (2018). A phylogenomic framework and timescale for comparative studies of tunicates. BMC Biology, 16, 1–14.

Deng, W., Henriet, S., Chourrout, D. (2018). Prevalence of mutation-prone microhomology-mediated end joining in a chordate lacking the c-NHEJ DNA repair pathway. Current Biology, 28, 3337-3341.e4.

Denoeud, F., Henriet, S., Mungpakdee, S., Aury, J.-M., Silva, C. Da., Brinkmann, H., Mikhaleva, J., et al. (2010). Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science, 330, 1381–1385.

Frith, M.C., Kawaguchi, R. (2015). Split-alignment of genomes finds orthologies more accurately. Genome Biology, 16, 106.

Fujii, S., Nishio, T., Nishida, H. (2008). Cleavage pattern, gastrulation, and neurulation in the appendicularian, *Oikopleura dioica*. Development Genes and Evolution, 218, 69–79.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., … Regev, A. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8, 1494–1512.

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

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., … Itoh, T (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Research, 24, 1384–1395.

Kishi, K., Hayashi, M., Onuma, T.A., Nishida, H. (2017). Patterning and morphogenesis of the intricate but stereotyped oikoplastic epidermis of the appendicularian, *Oikopleura dioica*. Developmental Biology, 428, 245–257.

Kishi, K., Onuma, T. A., Nishida, H. (2014). Long-distance cell migration during larval development in the appendicularian, *Oikopleura dioica*. Developmental Biology, 395, 299–306.

Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nature Biotechology, 37, 540–546.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics, 5, 1–9.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome biology, 5, 12.

Lukashin, A. V., Borodovsky, M. (1998). GeneMark.hmm: New solutions for gene finding. Nucleic Acids Research, 26, 1107–1115.

Martí-Solans, J., Belyaeva, O. V., Torres-Aguila, N.P., Kedishvili, N.Y., Albalat, R., Cañestro, C. (2016). Coelimination and Survival in Gene Network Evolution: Dismantling the RA-Signaling in a chordate. Molecular Biology and Evolution, 33, 2401–2416.

Masunaga, A., Liu, A.W., Tan, Y., Scott., A., Luscombe, A.W. (2020). Streamlined sampling and cultivation of the pelagic cosmopolitan larvacean, *Oikopleura dioica*. Journal of Visualized Experiments, in press.

Matsuo, M., Onuma, T., Omotezako, T., Nishida, H. (2020). Protein phosphatase 2A is essential to maintain meiotic arrest, and to prevent Ca^{2+} burst at spawning and eventual parthenogenesis in the larvacean *Oikopleura dioica*. Developmental Biology, 460, 155–163.

Naville, M., Henriet, S., Warren, I., Sumic, S., Reeve, M., Volff, J.N., Chourrout, D. (2019). Massive changes of genome size driven by expansions of non-autonomous transposable elements. Current Biology, 29, 1161-1168.e6.Navratilova, P., Danks, G.B., Long, A., Butcher, S., Manak, J.R., Thompson, E.M. (2017). Sex-specific chromatin landscapes in an ultra-compact chordate genome. Epigenetics and Chromatin, 10, 1–18.

Nishida, H. (2008). Development of the appendicularian *Oikopleura dioica*: Culture, genome, and cell lineages. Development Growth and Differentiation, 50, S239–S256.

Nishida, H., Stach, T. (2014). Cell Lineages and Fate Maps in Tunicates: Conservation and Modification. Zoological Science, 31, 645–652.

Omotezako, T., Matsuo, M., Onuma, T.A., Nishida, H. (2017). DNA interference-mediated screening of maternal factors in the chordate *Oikopleura dioica*. Scientific Reports, 7, 1–10.

Omotezako, T., Nishino, A., Onuma, T.A., Nishida, H. (2013). RNA interference in the appendicularian *Oikopleura dioica* reveals the function of the Brachyury gene. Development Genes and Evolution, 223, 261–267.

Omotezako, T., Onuma, T.A., Nishida, H. (2015). DNA interference: DNA-induced gene silencing in the appendicularian *Oikopleura dioica*. Proceedings of the Royal Society B: Biological Sciences, 282, 20150435–20150435.

Onuma, T.A., Matsuo, M., Nishida, H. (2017). Modified whole-mount in situ hybridisation and immunohistochemistry protocols without removal of the vitelline membrane in the appendicularian *Oikopleura dioica*. Development Genes and Evolution, 227, 367–374.

Onuma, T.A., Hayashi, M., Gyojya, F., Kishi, K., Wang, K., Nishida, H. (2020). A chordate species lacking *Nodal* utilizes calcium oscillation and *Bmp* for left-right patterning. Proceedings of the National Academy of Sciences of the United States of America, 117, 4188-4198.

Satou, Y., Nakamura, R., Yu, D., Yoshida, R., Hamada, M, Fujie, M., Hisata, K., …Satoh, N. (2019). Nearly complete genome of *Ciona intestinalis* Type A (*C. robusta*) reveals the contribution of inversion to chromosomal evolution in the genus *Ciona*. Genome Biology of Evolution, 11, 3144–3157

Seo, H.-C., Kube, M., Edvardsen, R.B., Jensen, M.F., Beck, A., Spriet, E., Gorsky, G., … Chourrout, D. (2001). Miniature genome in the marine chordate *Oikopleura dioica*. Science, 294, 2506.

Shoguchi, E., Kawashima, T., Satou, Y., Hamaguchi, M., Sin-I, T., Kohara, Y., Putnam, N., …Satoh, N. (2006) Chromosomal mapping of 170 BAC clones in the ascidian *Ciona intestinalis*. Genome Research, 16, 297–303.

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

Stach, T., Winter, J., Bouquet, J.-M., Chourrout, D., Schnabel, R. (2008). Embryology of a planktonic tunicate reveals traces of sessility. Proceedings of the National Academy of Sciences of the United States of America, 105, 7229–7234.

Stanke, M., Morgenstern, B. (2005). AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research, 33, 465–467.

Tang, F., Barbacioru, C., Nordman, E., Li, B., Xu, N., Bashkirov, V.I., Lao, K., Surani M.A. (2010). RNA-Seq analysis to capture the transcriptome landscape of a single cell. Nature Protocols, 5, 516–535.

Wang, K., Omotezako, T., Kishi, K., Nishida, H., Onuma, T.A. (2015). Maternal and zygotic transcriptomes in the appendicularian, *Oikopleura dioica*: novel protein-encoding genes, intra-species sequence variations, and trans-spliced RNA leader. Development Genes and Evolution, 225, 149–159.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A. (2013). The MaSuRCA genome assembler. Bioinformatics, 29, 2669–2677.

Figure legends

Figure 1 Screenshots of ANISEED (https://www.aniseed.cnrs.fr/).

(a) Home. (b) BLAST server. (c) Genome browser with transcript assembly and gene model (transcript model). (d) Gene Card with transcript and amino acid sequences.

Figure 2 Characterization of the male-specific Y-region.

(a) Schematic representation of scaffold (S) 064 and S039 encoding the male-specific Y-region. S064 is shown in the reverse orientation to the genome browser. The blue boxes represent exons of the Y-specific cDNA that is orthologous to the EST clone (KT0AAA379YD22.CONTIG) in the Oikobase. Green arrows and letters indicate positions of primers for genomic PCRs. (b-d) Genomic PCRs from mature males and females (b, d) or hatched larvae (c). Numbers depict larvae that were randomly collected and subjected to analysis. Green letters represent target regions of primers. N.C., Negative control with distilled water. Sequences of primers are summarized in Table S3. (e) Screening of male-enriched scaffolds. Illumina sequencing was carried out using pair-end libraries of male or female genomic DNA, and the obtained read sequences were mapped to the OSKA2016 scaffolds. The vertical axis represents the male to female ratio of mapped reads. Orange and gray bars indicate scaffolds for which mapped reads in the male were more than twice and 1.5 times that in the female, respectively. The two scaffolds in (a) are marked with blue rectangles. Scaffolds with red letter were grouped into a single sequence as shown in (f). (f) Alignment of the male-enriched scaffolds of OSKA2016 with Contig_4 of a draft Nanopore assembly from a single male individual. Scaffolds written in blue are shown in the reverse orientation to the genome browser. Scaffolds that were aligned with two regions of the Contig_4 were marked with arrows. (g) Deduced order of the male-enriched scaffolds of OSKA2016 in the Y-specific regions. Locations that were amplified by genomic PCRs in (b-d) were marked with blue rectangles.

Table 1 Summary of genome assembly, transcript assembly, and predicted gene models (transcript model and protein models).

Genome assembly	Total length (bp)	Number of scaffolds	Min. length (bp)	Max. length (bp)	Scaffold N50(bp)	Contig N50 (bp)	Gap (bp)	Resources
OSKA2016	56,625,162	576	1,000	6,807,127	1,506,282	623,729	331,456	Japanese population (present study)
Oikopleura_reference_unmasked_v3.0 (OikoBase)	70,471,451	1,260	2,002	3,167,015	395,387	24,917	3,938,492	Norwegian population (Danks et al., 2013)

Transcript model and transcript assembly	Total length (bp)	Number of transcripts	Min. length (bp)	Max. length (bp)	Average length (bp)	Resources
Transcript model	24,538,768	18,743	72	24,209	1,309	Japanese population (present study, OSKA2016 genome assembly)
Transcript assembly	26,524,329	19,277	300	24,145	1,375	Japanese population (present study, de novo assembly of RNA-seq data)

Protein model	Total length (aa)	Number of proteins	Min. length (aa)	Max. length (aa)	Average length (aa)	Resources
Protein model	7,382,255	18,743	9	7,980	393	Japanese population (present study, OSKA2016 genome assembly)

Table 2 Reciprocal searches of genome scaffolds between Japanese *O. dioica* (OSKA2016) and Norweigian *O. dioica* (Oikopleura_reference_unmasked_v3.0) using MUMmer v3.22.

Dataset	Number of scaffolds	Hit		No Hit	
		Number of scaffolds	(%)	Number of scaffolds	(%)
OSKA2016	576	441	76.6	135	23.4
Oikopleura_reference_unmasked_v3.0 (OikoBase)	1,260	1,149	91.2	111	8.8

Alignment criteria	Build	Number of aligned regions	Total length (bp)	Average length (bp)	Sequence similarity (%)
1-1 alignment	Oikopleura_reference_unmasked_v3.0	27,978	28,460,482	1,017	86.5
	OSKA2016	27,978	28,470,934	1,018	86.5
M-M alignment	Oikopleura_reference_unmasked_v3.0	44,685	33,633,403	753	86.5
	OSKA2016	44,685	33,640,518	753	86.5

Table 3 Sequence similarities of genes in the OSKA2016 against those in the OikoBase. Three genes were tested.

Brachury (Oidioi.g00005709 in the OSKA2016, GSOIDG00000279001 in the OikoBase)

Oidioi.g00005709	Number	Total length (bp)	Similarity with the GSOIDG00000279001 (OikoBase)	Gap (bp)	Gap (%)
introns	4	328	81.7%	11	3.3%
exons	5	1,642	92.4%	13	0.79%
intergenic regions	-	1,646	81.7%	127	7.2%

Bmp.a (Oidioi.g0000382 in the OSKA2016, GSOIDG00001216001 in the OikoBase)

Oidioi.g0000382	Number	Total length (bp)	Similarity with the GSOIDG00001216001 (OikoBase)	Gap (bp)	Gap (%)
introns	6	3,262	80.6%	438	17.2%
exons	7	1,099	90.0%	31	2.8%
intergenic regions	-	3,005	79.8%	223	7.2%

Pax6 (Oidioi.g00012737 in the OSKA2016, GSOIG00010489001 in the OikoBase)

Oidioi.g00012737	Number	Total length (bp)	Similarity with the OikoBase gene	Gap (bp)	Gap (%)
introns	11	2,740	80.6%	380	13.5%
exons	12	1,379	85.8%	105	7.3%
intergenic regions	-	6,744	69.6%	1295	17.4%

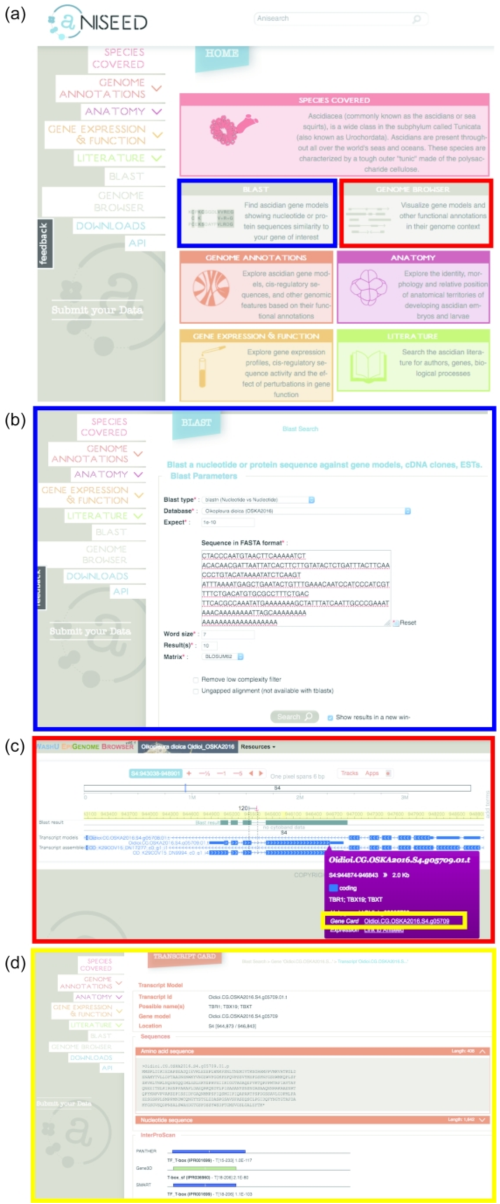
Table 4 Reciprocal BLAT searches of gene model (transcript or protein models) against genome scaffolds of Japanese O. dioica (OSKA2016) or the OikoBase (Oikopleura_reference_unmasked_v3.0).

Transcripts

Query (transcript models)	Number of transcript models used as query	Genome scaffolds	Coverage length percentage				
			>=90%	>=80%	>=70%	>=60%	>=50%
OSKA2016	18,743	OSKA2016	100.0%	100.0%	100.0%	100.0%	100.0%
OSKA2016	18,743	Oikopleura_reference_unmasked_v3.0 (OikoBase)	39.2%	59.1%	69.3%	76.0%	80.8%
Odioica_GeneModels_transcripts_reference_v1.0 (OikoBase)	17,212	OSKA2016	55.8%	72.8%	80.2%	84.3%	87.2%
Odioica_GeneModels_transcripts_reference_v1.0 (OikoBase)	17,212	Oikopleura_reference_unmasked_v3.0 (OikoBase)	99.0%	99.3%	99.4%	99.5%	99.5%

Amino acids

Query (protein models)	Number of protein models used as query	Genome scaffolds	Coverage length percentage				
			>=90%	>=80%	>=70%	>=60%	>=50%
OSKA2016	18,743	OSKA2016	99.8%	100.0%	100.0%	100.0%	100.0%
OSKA2016	18,743	Oikopleura_reference_unmasked_v3.0 (OikoBase)	56.0%	68.9%	75.7%	80.7%	84.9%
Odioica_GeneModels_peptides_reference_v1.0 (OikoBase)	17,212	OSKA2016	70.1%	80.9%	85.7%	88.8%	91.0%
Odioica_GeneModels_peptides_reference_v1.0 (OikoBase)	17,212	Oikopleura_reference_unmasked_v3.0 (OikoBase)	98.8%	99.2%	99.3%	99.5%	99.5%



SPECIES COVERED

Ascidiacea (commonly known as the ascidians or sea squirts), is a wide class in the subphylum called Tunicata (also known as Urochordata). Ascidians are present throughout all over the world's seas and oceans. These species are characterized by a tough outer "tunic" made of the polysaccharide cellulose.

BLAST

Find ascidian gene models showing nucleotide or protein sequences similarity to your gene of interest

GENOME BROWSER

Visualize gene models and other functional annotations in their genome context

GENOME ANNOTATIONS

Explore ascidian gene models, cis-regulatory sequences, and other genomic features based on their functional annotations

ANATOMY

Explore the identity, morphology and relative position of anatomical territories of developing ascidian embryos and larvae

GENE EXPRESSION & FUNCTION

Explore gene expression profiles, cis-regulatory sequence activity and the effect of perturbations in gene function

LITERATURE

Search the ascidian literature for authors, genes, biological processes

Blast a nucleotide or protein sequence against gene models, cDNA clones, ESTs.

Blast Parameters

Blast type* : blastn (Nucleotide vs Nucleotide)
Database* : Oikopleura dioica (OSKA2016)
Expect* : 1e-10

Sequence in FASTA format* :

```
CTACCCAATGTAACTTCAAAAATCT
ACACAACGATTAATTATCACTTCTTGTATACTCTGATTTACTTCAA
CCCTGTACATAAAATATCTCAAGT
ATTTAAAATGAGCTGAATACTGTTTGAAACAATCCATCCCATCGT
TTTCTGACATGTGCGCCTTTCTGAC
TTCACGCCAAATATGAAAAAAAGCTATTTATCAATTGCCCGAAAT
AAACAAAAAAAATTAGCAAAAAAAA
AAAAAAAAAAAAAAAAAA
```

Reset

Word size* : 7
Result(s)* : 10
Matrix* : BLOSUM62

□ Remove low complexity filter
□ Ungapped alignment (not available with tblastx)

☑ Show results in a new win-

One pixel spans 6 bp

Oidioi.CG.OSKA2016.S4.g05708.01.t
Oidioi.CG.OSKA2016.S4.g05709.01.t
OD_K29OOV15_DN17277_c0_g1_i1
OD_K29OOV15_DN9994_c0_g1_i4

Oidioi.CG.OSKA2016.S4.g05709.01.t
S4:944874-946843 » 2.0 Kb
🟦 coding
TBR1; TBX19; TBXT
Gene Card Oidioi.CG.OSKA2016.S4.g05709
Expression Link to Aniseed

Transcript Model

Transcript id	Oidioi.CG.OSKA2016.S4.g05709.01.t
Possible name(s)	TBR1; TBX19; TBXT
Gene model	Oidioi.CG.OSKA2016.S4.g05709
Location	S4 [944,873 / 946,843]

Sequences

Amino acid sequence Length: 408

>Oidioi.CG.OSKA2016.S4.g05709.01.p

Nucleotide sequence Length: 1,842

InterProScan

TF_T-box (IPR001699) - T[15-233] 1.0E-117
T-box_sf (IPR036960) - T[18-206] 2.1E-80
TF_T-box (IPR001699) - T[18-206] 1.1E-103

Figure 1

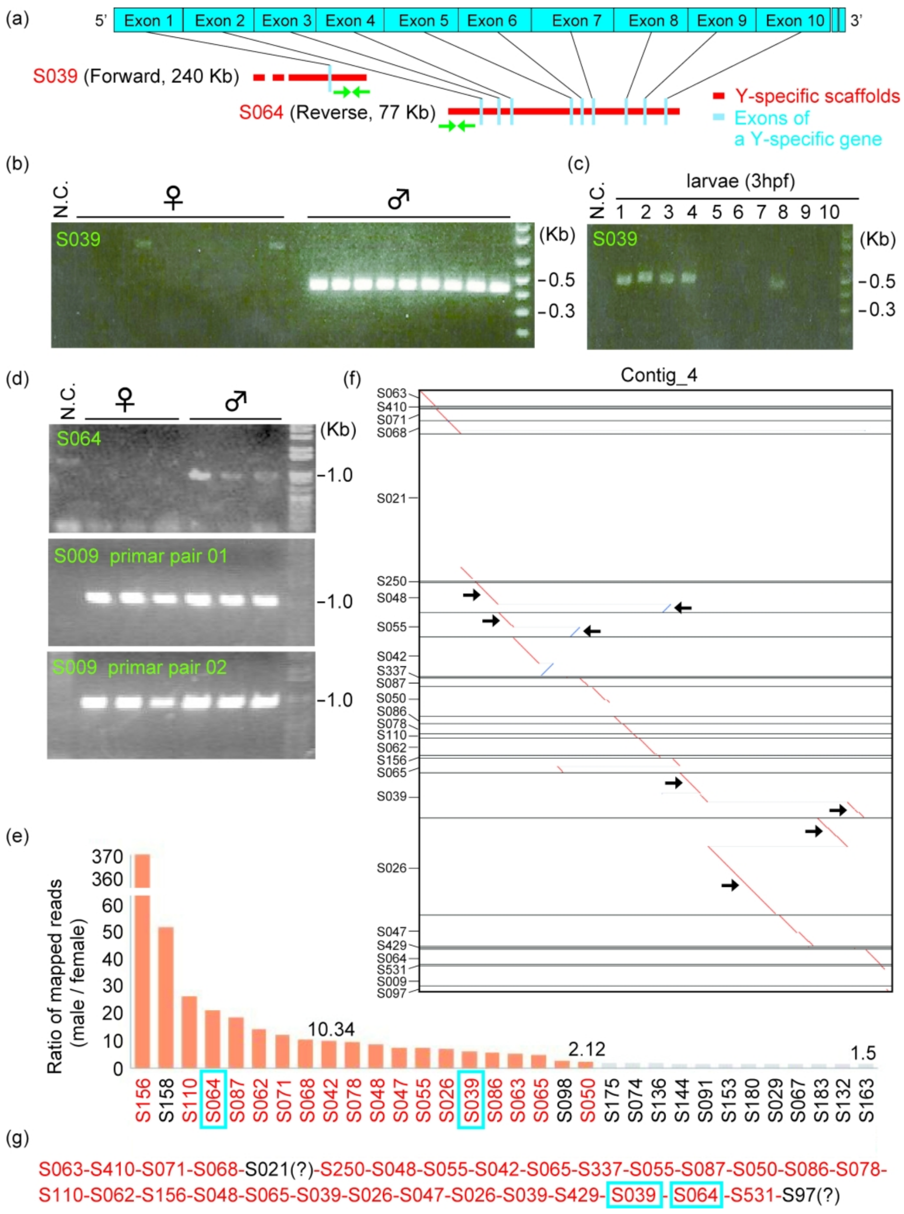


Figure 2

196x262mm (300 x 300 DPI)

Table S1 Summary of genome assembles that were generated by different programs. Three representative examples were shown.

Assembly program	Total length (bp)	GAP (bp)	Number of scaffold	Min. length (bp)	Max. length (bp)	Scaffold N50 (bp)	Contig N50 (bp)
Platanus_bridgemapper (OSKA2016)	56,625,162	331,456	576	1,000	6,807,127	1,506,282	623,729
Falcon_160511_whole	79,341,990	0	460	1,108	4,005,979	672,068	672,068
MaSuRCA (delete 50% Covered & Identity contigs)	65,729,327	0	124	1,141	7,530,094	1,691,822	1,691,822

Table S2 Assessment of genome assemblies by searching 303 BUSCOs. Note that the OSKA2016 was adopted as the genome assembly.

Build	% Complete single-copy BUSCOs	% Fragmented BUSCOs	% Completed single-copy & fragmented BUSCOs
Platanus_bridgemapper (OSKA2016)	73.3	13.9	87.2
Falcon_160511_whole	52.5	20.5	73.0
MaSuRCA(delete 50% Covered & Identity contigs)	70.6	11.2	81.8

Table S3 Sequences of primers used for genomic PCRs in Figure 2.

Target scaffold (S)	Forward primer (5'-3')	Reverse primer (5'-3')
S039	CCGCATTTTGGAACTCAGCA	CTTTAGAGGTGTGGATTTAATGGAAGATCC
S064	CGGAAAATAACGAACAGCTCCACTTCTACG	CATATCTCGTAACCCGAATAACCATGAGCC
S009 primer pair 01	TCTGGTGCAAAAACCAGCTTCCGAAGAATG	CCTTTGACGGGAAGCGAAACTAAAAGTGCG
S009 primer pair 02	TTGTTCCCGCCTGCCGAATAGCTTTATTGG	TTTGCAGAATCATCAAGGGAGTTCGCTGCC

Supplemental material 01 Nucleotide sequence of partial cDNA clone for the male-specific Y gene used in the Figure 2a.

```
CCGCATTTTGGAACTCAGCAACTACCTGGACACACAGAAACATCATTCAATCTTGATCTTGTTCTT
AAAAATTCAAATTATGAATCTATTTGTTATGATAACTCACATATTGACCACAATTCCGACGCTACA
ACTTATCCTTTAGCAATTGACATCGGATGTCCAGCAAGTCGCCATCTAGTACTCGACTATCAAAA
AATATTTGAATTGGATCAAGATAGAAGAAGAGGTAATTTTCTTTGCCAAATGGATATTGAGGAAA
AGAGACGATATGTTGTTGGGTATAAGTATCAACCTGAACGAAATGTTCTTTGTATAAGAAGTGGA
GCTGTTCGTTATTTTCCGCCAATGCTTACTGTAAAGGATAAATCGACGGGAAGAGAATGGCCTTA
TCATGAATCAGTTGTTTTTACAGTTACCAACGGGGAATGGCTAAAATTTAAAGACAATAAGCTTG
AAATACCGAACTTCAAATCAGAAAATGAATGTGCGATATTAAATAATAATCTTGTTGACAAAGATT
ATATCTTTACTTGTCTAGGTACGCATAGTTCAAATAATTCGACAACCGGTGATAATAAACCATCAG
TTCTAGAAGAGATATCAGTTTGTGGCTCAACGACTAGGCTGAAATTTGAATTAGACAAGAGTGAA
ACGGAAAAATCATATCCGAGTTTTTCAAATATGGAACTTCCTTTTGGTTTTTGGATGTCTGGATCT
TTGAAGACGGATGATTTTTTTCCTGCGGAATTTTACATGAATATTCGTGTATCTAACAGTATAGAT
GAGATTAATAAGCGTGATAAAAAACCATGTATAATTCCAGGAAATTGTTTTTTAGAAACATTTTTT
ACCATACGAATAGGAATGGTTATTTATGAAAGTGTTACTCAACTATTTTTCGTTATTATAGTTATA
GTCATTCTTACATTTGCTCTTATATTATATGTTCAATCAAAGAATCATGAAAATTTAATAGAGAAT
ATAAAAGACACTATACATGAGTGTTTGAATGCAAAAGCTATCGCCAAGGAGCGACACATGATAGA
GCTCTCTTATCAGAAT
```

Supplemental Material 1 (related to Fig. 2a-d). cDNAs and amino acid sequences from one of the male-specific gene of Japanese (present study) and Norwegian *O. dioica* (Oikobase). Vertical lines indicate conserved sequences. (A) Alignment of predicted amino acids. (B) Alignment of mRNA sequences.

(A)

```
Japan (partial)                             1 PHFGTQQLPGHTETSFNLDLVLKNS
                                              |||| ||||| |||||||| || ||
Norway (Oikobase)   1 MPIDIDIDSHKYFRNTTKSFILHIQPHFGRQQLPGYTETSFNLDVVLNNS

Japan (partial)    26 NYESICYDNSHIDHNSDATTYPLAIDIGCPASRHLVLDYQKIFELDQDRR
                      ||   | ||||||||||| |||| ||||||||||||||||||| ||||||
Norway (Oikobase)  51 NYSNTCFDNSHIDHNSDASSYPLGIDIGCPASRVLVLDYQKIFQLDQDRR

Japan (partial)    76 RGNFLCQMDIEEKRRYVVGYKYQPERNVLCIRSGAVRYFPPMLTVKDKST
                       |||||||||||||| ||||||||||||||||||||||||||||||||||
Norway (Oikobase) 101 KGNFLCQMDIEEKRRFVVGYKYQPERNVLCIRSGAVRYFPPMLTVKDKST

Japan (partial)   126 GREWPYHESVVFTVTNGEWLKFKDNKLEIPNFKSENECAILNNNLVDKDY
                      | |||||||||||||||||| |  ||||||||  | ||  ||||| | ||
Norway (Oikobase) 151 GKEWPYHESVVFTVTNGEWLRFRYNKLEIPNFRTEDECELLNNNLGDEDY

Japan (partial)   176 IFTCLGTHSSNNSTTGD--NKPSVLEEI-SVCGSTTRLKFELDKSETEKS
                      || ||| || ||||  |   | ||||||   | |   |||||||   |||
Norway (Oikobase) 201 IFICLGKHSPNNSTIVDYSGKTSVLEEIPRDCDSLPKLKFELDKNQIEKS

Japan (partial)   223 YPSFSNMELPFGFWMSGSLKTDDFFPAEFYMNIRVSNSIDEINKRDKKPC
                      |||||||||||||||||||||||||||||||||||||| ||||  ||| |
Norway (Oikobase) 251 YPSFSNMELPFGFWMSGSLKTDDFFPAEFYMNIRVSNSLDEINNQDKKQC

Japan (partial)   273 IIPGNCFLETFFTIRIGMVIYESVTQLFFVIIVIVILTFALILYVQSKNH
                      |.||||||||||||||||||||||||||||  ||      || |||||||
Norway (Oikobase) 301 ISPGNCFLETFFTIRIGMVIYESVTQLFFVFFVILLITIVLIIYVQSKNH

Japan (partial)   323 ENLIENIKDTIHECLNAKAIAKERHMIELSYQN 356-------------
                      || |||||||| ||||||||| || |||||||
Norway (Oikobase) 351 ENFIENIKDTIYECLNAKAIANERQMIELSYQSNLRAEYFKDSTLMTKRS

Japan (partial)       --------------------------------------------------

Norway (Oikobase) 401 RKQSSCQWNHSKISQHSSHDSQWSSPIYRQKVSSRFEHNKKSALVHQEII

Japan (partial)       --------------

Norway (Oikobase) 451 DKRRNTLRNLSFEV 464
```

(B)

```
Japan (partial)         --------------------------------------------------
Norway (Oikobase)     1 GTAGATACGACTTCTATAAGTTACACACTAATTAAACCAGATTTTATGCC

Japan (partial)         --------------------------------------------------
Norway (Oikobase)    51 GATTGATATTGATATCGATTCTCACAAATATTTCAGAAACACAACAAAGT

Japan (partial)         ------------------1 CCGCATTTTGGAACTCAGCAACTACCTGGA
                                            |||||||||||||  |||||||| ||||||
Norway (Oikobase)   101 CCTTTATACTACATATACAACCGCATTTTGGAAGACAGCAACTTCCTGGA

Japan (partial)      31 CACACAGAAACATCATTCAATCTTGATCTTGTTCTTAAAAATTCAAATTA
                         |||| || |||||||||||||||||| |||| || || |||||||||||
Norway (Oikobase)   151 TACACGGAGACATCATTCAATCTTGATGTTGTCCTAAATAATTCAAATTA

Japan (partial)      81 T--GAATCTATTTGTTATGATAACTCACATATTGACCACAATTCCGACGC
                        |  |||  || ||||| |||||| ||||||||||| || ||||| |||||
Norway (Oikobase)   201 TTCGAA--TACTTGTTTTGATAATTCACATATTGATCATAATTCGGACGC

Japan (partial)     129 TACAACTTATCCTTTAGCAATTGACATCGGATGTCCAGCAAGTCGCCATC
                        | || |||||||||| |  ||||||||||||||||||||||||||    |
Norway (Oikobase)   249 TTCATCTTATCCTTTGGGTATTGACATCGGATGTCCAGCAAGTCGAGTAC

Japan (partial)     179 TAGTACTCGACTATCAAAAAATATTTGAATTGGATCAAGATAGAAGAAGA
                        | || || || || |||||||||||| || | |||||||||||||||| |
Norway (Oikobase)   299 TTGTTCTTGATTACCAAAAAATATTTCAACTAGATCAAGATAGAAGAAAA

Japan (partial)     229 GGTAATTTTCTTTGCCAAATGGATATTGAGGAAAAGAGACGATATGTTGT
                        || || ||| | || |||||||||||||||||||||||||| | ||||||
Norway (Oikobase)   349 GGAAACTTTTTATGTCAAATGGATATTGAGGAAAAGAGACGTTTTGTTGT


Japan (partial)     279 TGGGTATAAGTATCAACCTGAACGAAATGTTCTTTGTATAAGAAGTGGAG
                        ||||||||| ||||||||||| || |||||||||||||||||||||||||
Norway (Oikobase)   399 TGGGTATAAATATCAACCTGAGCGTAATGTTCTTTGTATAAGAAGTGGAG

Japan (partial)     329 CTGTTCGTTATTTTCCGCCAATGCTTACTGTAAAGGATAAATCGACGGGA
                        ||||||| |||||||| || ||||| || || ||||||||||| ||||||
Norway (Oikobase)   449 CTGTTCGATATTTTCCTCCTATGCTCACAGTTAAGGATAAATCAACGGGA

Japan (partial)     379 AGAGAATGGCCTTATCATGAATCAGTTGTTTTTACAGTTACCAACGGGGA
                        | ||||||||||||||| ||||| || || ||||| || || ||||| ||
Norway (Oikobase)   499 AAAGAATGGCCTTATCACGAATCTGTAGTATTTACGGTAACGAACGGTGA

Japan (partial)     429 ATGGCTAAAATTTAAAGACAATAAGCTTGAAATACCGAACTTCAAATCAG
                        |||||| | ||||| | | ||||| |||||||||||||| |||| | |||
Norway (Oikobase)   549 ATGGCTTAGATTTAGATATAATAAACTTGAAATACCGAATTTCAGAACAG

Japan (partial)     479 AAAATGAATGTGCGATATTAAATAATAATCTTGTTGACAAAGATTATATC
                        |  ||||||||| | | |||||||||||||||| |||  ||||||||||
Norway (Oikobase)   599 AGGATGAATGTGAGTTGTTAAATAATAATCTTGGTGATGAAGATTATATT

Japan (partial)     529 TTTACTTGTCTAGGTACGCATAGTTCAAATAATTCGACAACCGGTGATAA
                        ||||  ||||||||||  |||||| ||||||||||||||| ||  ||  |
Norway (Oikobase)   649 TTTATATGTCTAGGTAAACATAGTCCAAATAATTCGACAATCGTAGACTA
```

```
Japan (partial)      579 T------AAACCATCAGTTCTAGAAGAGATATCA---GTTTGTG--GCTC
                         |      ||| | || ||||| ||||| ||| ||   |  ||||   |||
Norway (Oikobase)    699 TTCGGGAAAAACGTCTGTTCTTGAAGAAATACCACGTGACTGTGACTCTC

Japan (partial)      618 AACGACTAGGCTGAAATTTGAATTAGACAAGAGTGAAACGGAAAAATCAT
                          ||  | | ||| ||||||||||||||||||| | |||  ||||| ||||
Norway (Oikobase)    749 TAC--CAAAGCTAAAATTTGAATTAGACAAGAATCAAATCGAAAAGTCAT

Japan (partial)      668 ATCCGAGTTTTTCAAATATGGAACTTCCTTTTGGTTTTTGGATGTCTGGA
                         |||| |||||||||||||||||||| ||||||||||||||||||||||||
Norway (Oikobase)    797 ATCCAAGTTTTTCAAATATGGAACTCCCTTTTGGTTTTTGGATGTCTGGA

Japan (partial)      718 TCTTTGAAGACGGATGATTTTTTTCCTGCGGAATTTTACATGAATATTCG
                         || || || ||||||||||||||||| || |||||||||||||| ||  |
Norway (Oikobase)    847 TCATTAAAAACGGATGATTTTTTTCCGGCAGAATTTTACATGAACATAAG

Japan (partial)      768 TGTATCTAACAGTATAGATGAGATTAATAAGCGTGATAAAAAACCATGTA
                          |||||||||||| |||||||||| ||||| |  |||||||||| |||||
Norway (Oikobase)    897 AGTATCTAACAGTTTAGATGAGATAAATAACCAGGATAAAAAACAATGTA

Japan (partial)      818 TAATTCCAGGAAATTGTTTTTTAGAAACATTTTTTACCATACGAATAGGA
                         ||| |||||||||||||||||||||||| || || |||||||| ||||||
Norway (Oikobase)    947 TAAGTCCAGGAAATTGTTTTTTAGAAACCTTCTTCACCATACGTATAGGA

Japan (partial)      868 ATGGTTATTTATGAAAGTGTTACTCAACTATTTTTCGTTATTATAGTTAT
                         ||||| ||||||||||||||||| ||| ||||||||||| || | || ||
Norway (Oikobase)    997 ATGGTAATTTATGAAAGTGTTACACAATTATTTTTCGTTTTTTTCGTAAT

Japan (partial)      918 AGTCATTCTTACATTTGCTCTTATATTATATGTTCAATCAAAGAATCATG
                         | |  || ||||| | |  || ||  |||||||||||||||| |||||||
Norway (Oikobase)   1047 ATTACTTATTACAATCGTCCTGATTATATATGTTCAATCAAAAAATCATG

Japan (partial)      968 AAAATTTAATAGAGAATATAAAAGACACTATACATGAGTGTTTGAATGCA
                         ||||||| |||||||||||||||||||| ||  |||| ||| | |||||
Norway (Oikobase)   1097 AAAATTTTATAGAGAATATAAAAGACACGATTTATGAATGTCTAAATGCT

Japan (partial)     1018 AAAGCTATCGCCAAGGAGCGACACATGATAGAGCTCTCTTATCAG---AA
                         |||||||| || || ||  |||| ||||| ||||| |||||||||   ||
Norway (Oikobase)   1147 AAAGCTATTGCAAATGAAAGACAAATGATTGAGCTTTCTTATCAGTCTAA

Japan (partial)     1065 T-------------------------------------------------
                         |
Norway (Oikobase)   1197 TCTGAGGGCGGAATACTTTAAGGATTCAACACTCATGACAAAACGGTCAA

Japan (partial)          --------------------------------------------------

Norway (Oikobase)   1247 GAAAGCAATCTTCTTGTCAATGGAATCATTCGAAAATCTCTCAGCACAGT

Japan (partial)          --------------------------------------------------

Norway (Oikobase)   1297 TCCCATGACTCGCAATGGTCCTCGCCAATTTACAGACAAAAAGTTTCATC

Japan (partial)          --------------------------------------------------

Norway (Oikobase)   1347 TCGATTTGAACACAATAAAAAATCAGCTTTAGTTCATCAAGAAATAATAG
```

```
Japan (partial)        --------------------------------------------------

Norway (Oikobase) 1397 ACAAGAGAAGAAATACTCTGCGCAATCTCTCATTCGAGGTCTGAAATATT


Japan (partial)        -------------------------------------------------

Norway (Oikobase)  1447 TTTGACAACAAATGNTTTTTTAACCNTTACAAAATAAAAGTTATACAAA
                                                                       1499
```