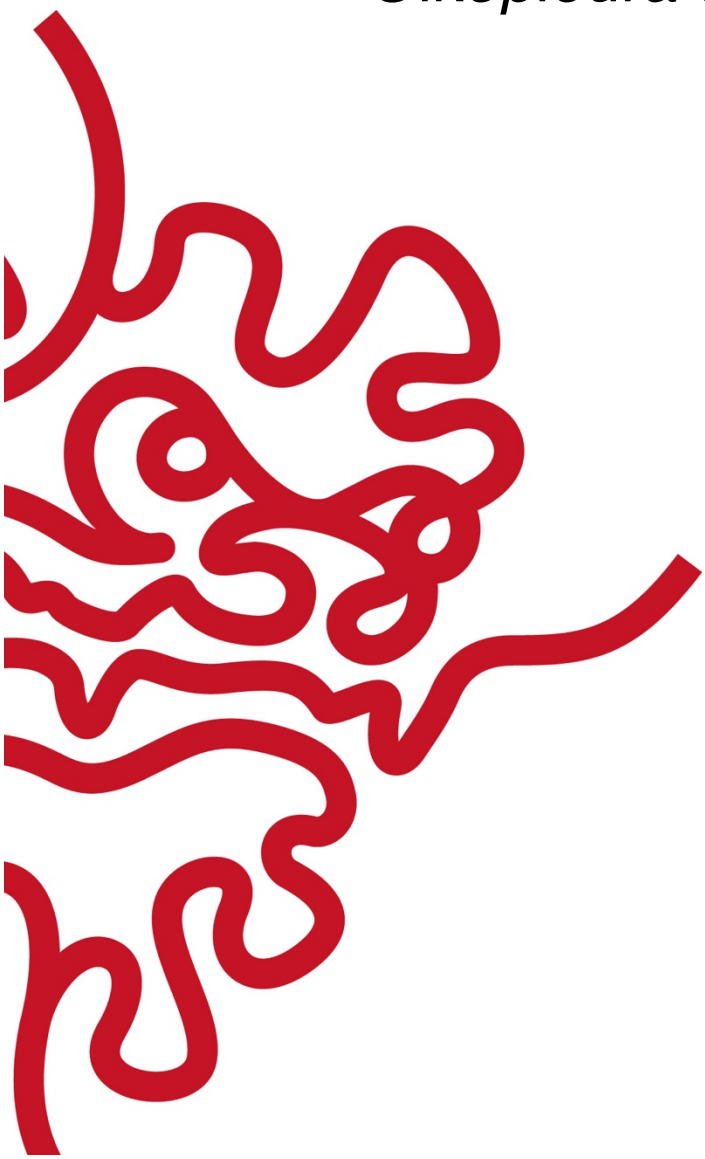
Okinawa Institute of Science and Technology Graduate University

Thesis submitted for the degree
Doctor of Philosophy

Cross-genome Comparison of Global *Oikopleura dioica* Populations



by Aleksandra Bliznina

Supervisor: Nicholas M. Luscombe

20 December, 2022

Declaration of Original and Sole Authorship

I, Aleksandra Bliznina, declare that this thesis entitled "Cross-genome comparison of global *Oikopleura dioica* populations" and the data presented in it are original and my own work.

I confirm that:

- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly acknowledged. Quotations from the work of others have been clearly indicated, and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.
- None of this work has been previously published elsewhere, with the exception of the following:

  Bliznina A., Masunaga A., Mansfield M.J., Tan Y., Liu A.W., West C., Rustagi T., Chien H.C., Kumar S., Pichon J., Plessy C., Luscombe N.M. (2021). Telomere-to-telomere assembly of the genome of an individual *Oikopleura dioica* from Okinawa using Nanopore-based sequencing. *BMC genomics, 22*(1): 1-18. doi:10.1186/s12864-021-07512-6

Date: 20 December, 2022

Signature: 

Abstract

Larvaceans represent the second most abundant zooplankton in all the world's oceans, with key roles in marine food chains and global carbon flux. *Oikopleura dioica* is a free-swimming planktonic tunicate from the group and possesses the smallest animal genome with extremely dynamic organization: multiple genomic features such as transposon diversity, intron repertoire, gene content and order are altered in *Oikopleura* compared with other metazoans. Intriguingly, such genome reorganization has not affected the preservation of their ancestral morphology, since *O. dioica* maintains a chordate-like body plan throughout its life. *O. dioica* can be easily distinguished from other larvaceans mainly based on separate sexes and the presence of two subchordal cells on its tail. My research is focused on the cross-genome comparison of three *O. dioica* populations sampled from the Northern hemisphere: one from North Atlantic (Barcelona/Bergen) and two from Pacific (Osaka/Aomori and Okinawa/Kume) Oceans. For each population, I generated high-quality genome assemblies using a combination of short- and long-read sequencing technologies, as well as chromatin conformation data, confirming preservation of three chromosome pairs. A pairwise comparison of populations revealed a striking degree of genome reshuffling that involves a vast number of synteny breaks and rearrangements. My research also shows that rearrangements mostly happen within individual chromosomes and generally preserve protein-coding features, such as genes and their constituent exons, although the gene order has been effectively randomized. *O. dioica* populations exhibit differences in repeats and gene content that affect even evolutionary conserved clusters, such as Hox genes. Consistent with an increased evolutionary rate, the accumulation of rearrangements in *O. dioica* appears to have happened much faster than in other animals and resulted in the divergence of multiple lineages of dioecious *Oikopleura*. The fact that their morphology stayed virtually identical makes *O. dioica* a perfect model to study genotype-phenotype correlation and the possible existence of unknown regulatory mechanisms. Overall, my thesis contributes new insights into the evolution of chordate genomes and, thus, may be interesting beyond the field of *Oikopleura* research.

Acknowledgements

First of all, I would like to thank past and present members of the OIST Genomics and Regulatory Systems Unit, who each played an essential role in completing this thesis by creating a comfortable and collaborative work environment, engaging in scientific discussions, providing data crucial for this thesis, reading and editing manuscript drafts, and much more.

I would like to express my sincere gratitude to my primary supervisor, Nicholas Luscombe, who gave me an opportunity to work in his lab, allowed me to make independent research decisions and guided me throughout the past five years. I also owe so much to Charles Plessy for his invaluable contribution and continued mentorship. I am forever grateful for the enormous time Charles spent answering my questions, discussing the project, troubleshooting problems with data analysis, and proofreading manuscripts.

I wish to extend my special thanks to Aki Masunaga, Yongai Tan and Andrew Liu for doing incredible work of troubleshooting and maintaining the *Oikopleura dioica* culture, collecting samples around Japan and generating endless sequencing data for the whole lab and this thesis in particular. My sincere thanks to Michael Mansfield for always being ready to help with data analysis and providing valuable scientific, career and life advice.

I would like to thank our collaborators, especially Cristian Cañestro from the University of Barcelona, Takeshi Onuma and Hiroki Nishida from the University of Osaka for providing samples of *O. dioica* crucial for this thesis. Thank you to my thesis committee members, Evan Economo and Akihiro Kusumi, for keeping an eye on my progress as a PhD student. I would also like to extend my special thanks to Konstantin Khalturin for helpful scientific discussions and tips in data analysis and life.

I wish to acknowledge the help of the OIST DNA Sequencing Section for their support in sequencing. Thank you to the OIST Scientific Computing and Data Analysis section, especially Jan Moren, who provided support in running data analysis on the cluster. Special thanks to all the Graduate School staff for constant help with academic and nonacademic-related things.

Most importantly, I would like to thank my family and friends for their unwavering love and encouragement. My mom, Olga Bliznina, helped me to choose my career path and taught me important values I still use to guide my way through life. The warmest thanks to my sister and best friend, Maria, for her unconditional support with anything I do and for creating a safe space to share not only my biggest dreams but also my worst fears.

I wish to thank Georg Fischer who was always the first to proofread this thesis and provide crucial feedback. He gave me confidence by believing in me like nobody else and providing the greatest support and encouragement through good and bad times, even when we were living thousand kilometers apart.

I wish to thank my friends from Moscow, Anya and Alina, for the emotional support that cheered me up even in the most challenging times. My friends from Okinawa – especially Kamila, Aki, Tanya, Khelil, Ainash and Aisulu – for making this beautiful island feel like home and for always being up for any kind of adventure. Last but not least, I would like to acknowledge my dog Kona, the sweetest, most intelligent and loving girl, for constantly providing reasons to smile and making sure that I get my daily dose of walks and sunshine.

Abbreviations

a-NHEJ	alternative non-homologous end-joining
Aom	Aomori
Bar	Barcelona
Ber	Bergen
bp	base pairs
BUSCOs	Benchmarking Universal Single-Copy Orthologs
chr1	Autosomal chromosome 1
chr2	Autosomal chromosome 2
CNEs	conserved non-coding elements
c-NHEJ	classical non-homologous end-joining
DSBs	DNA double strand breaks
GO	Gene Ontology
Hox	Homeobox
IOs	Isolated orthologs
kbp	Kilobase pairs
Kum	Kume
LGs	Linkage groups
LINEs	Long interspersed nuclear elements
LTRs	Long terminal repeats
Mbp	Megabase pairs
MH	micro-homologous
MITEs	Miniature inverted-repeat transposable elements
Mya	Million years ago
Odin	*Oikopleura dioica* non-LTR
Oki	Okinawa
ORF	Open reading frame
Osa	Osaka
PAR	Pseudo-autosomal regions
SBs	Synteny blocks
SINEs	Short interspersed nuclear elements
TEs	Transposable elements
TOR	*Ty3/Gypsy*-related *Oikopleura*-specific LTR retrotransposon
XSR	X-specific region
YSR	Y-specific region

Table of Contents

List of Figures

List of Tables

Chapter One

Introduction

1.1 Tunicates, the sister group to vertebrates

Tunicates, also called urochordates, are a group of worldwide marine invertebrates comprising the subphylum Tunicata. Together with cephalochordates (such as lancets) and vertebrates, tunicates constitute the chordate phylum. Animals are classified as chordates based on the presence of a notochord and a dorsal hollow neural tube.

It is established that tunicates are the closest relatives to vertebrates (Fig. 1.1; Delsuc et al., 2018). However, the phylogenetic relationships inside the chordate phylum have only recently been resolved. Previously, tunicates were placed at the most basal position of the chordate phylogeny. Therefore, both cephalochordates and vertebrates have been thought to evolve from a tunicate-like ancestor. This view was mainly based on the overall morphological similarities between cephalochordates and vertebrates relative to tunicates. However, phylogenomic analyses have provided new evidence that supports strong phylogenetic affinity between tunicates and vertebrates. Tunicates and vertebrates were grouped into a sister clade called Olfactores (Bourlat et al., 2006; Delsuc et al., 2006; Dunn et al., 2008; Putnam et al., 2008). The reordering of the chordate phylogeny suggested that the tunicate body plan is evolutionarily derived from a more complex chordate ancestor (Delsuc et al., 2006).

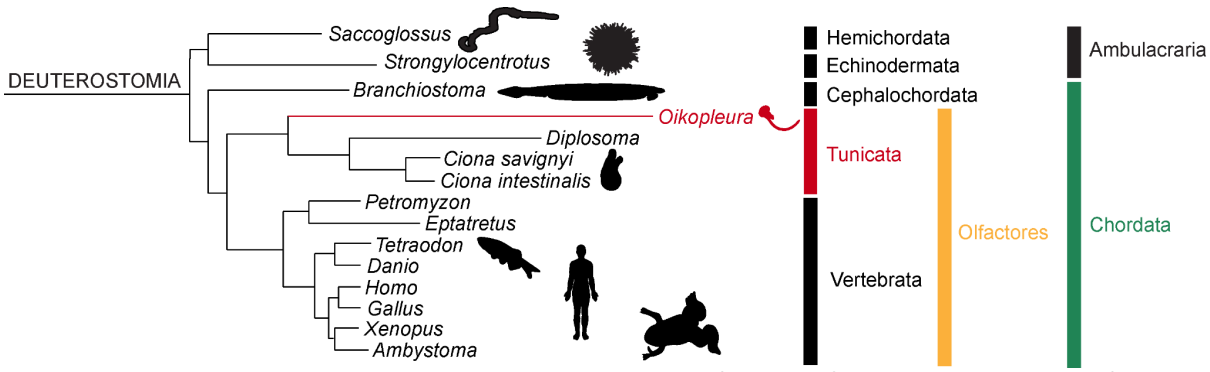


Figure 1.1: Phylogenetic relationships within deuterostomes (adapted from Delsuc et al., 2018).

Tunicates are extremely diverse with approximately 3,000 extant species (Appeltans et al., 2012) that occupy most marine habitats - from shallow waters to deep seas. They exhibit a wide range of life cycles, developmental strategies, reproductive methods (asexual and sexual reproduction; the majority of species are hermaphrodites) and regenerative abilities (regeneration of the whole body or only specific organs; Lemaire et al., 2008). Indeed, genome sequences have revealed that tunicates have undergone a rapid evolution compared with other deuterostomes. As a result, the tunicate phylogeny has been notoriously difficult to resolve.

Traditionally, the subphylum Tunicata is divided into three major classes - Ascidiacea (phlebobranchs, aplousobranchs, and stolidobranchs), Thaliacea (salps, doliolids, and pyrosomes) and Appendicularia (larvaceans). Previous phylogenetic studies relying on 18S rRNA (Swalla et al., 2000; Zeng and Swalla, 2005; Tsagkogeorga et al., 2009) and mitochondrial protein-coding genes (Singh et al., 2009; Rubinstein et al., 2013; Shenkar et al., 2016) have suggested the paraphyly of ascidians and proposed three additional clades: Appendicularia, Phlebobranchia + Thaliacea + Applosobranchia and Stolidobranchia. Indeed, two recent analyses of genomic (Kocot et al., 2018) and transcriptomic (Delsuc et al., 2018) datasets have supported the major splits within Tunicata and placed Appendicularia as a sister group to all other tunicates (Fig. 1.2).

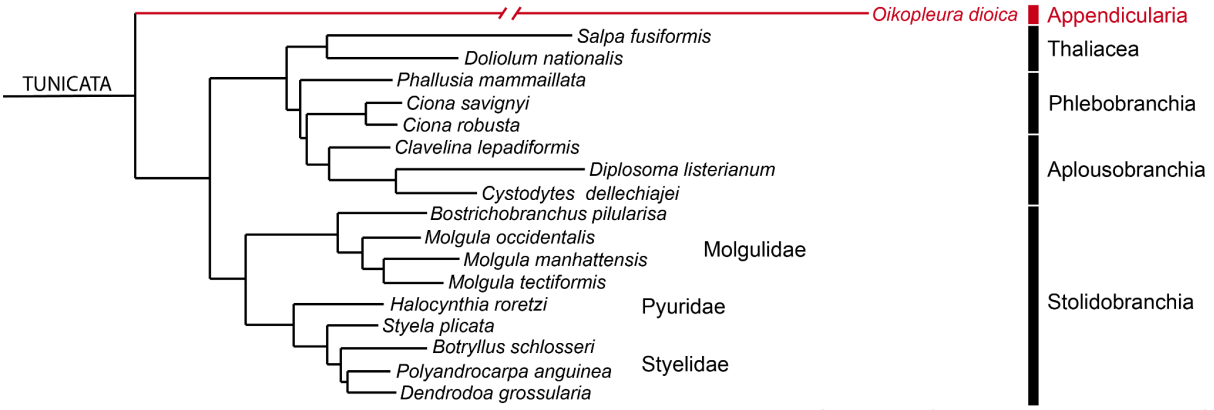


Figure 1.2: Phylogenetic relationships within the subphylum Tunicata (adapted from Delsuc et al., 2018).

All tunicate clades are united by a suite of common morphological features. The entire body of an adult tunicate is encased in a thick covering - tunic, also known as test (Hirose et al., 1999). The major constituent of the tunic is tunicin - a specific type of cellulose that tunicates synthesize directly. Cellulose production is normally confined to bacteria and plants, making tunicates unique among animals. It is proposed that the last common tunicate ancestor acquired the ability to synthesize cellulose through horizontal gene transfer of a prokaryotic gene (Nakashima et al., 2004; Sagane et al., 2010). Most tunicates have a tadpole-like free-swimming larva that exhibits simplified chordate morphology. Among tunicates, only larvaceans (appendicularians) retain this tadpole shape into adult life; all others resorb the tail through metamorphosis after a brief larva phase. Despite diverse adult body plans, all clades exhibit anatomic features that are closely related to those of vertebrates, including a heart, vascular system, notochord, and an endostyle which assists an animal to filter feed (Millar, 1971).

1.2 Larvaceans

The group Larvaceans represent the second most abundant zooplankton in all the world's oceans (Alldredge, 1976). Larvaceans serve as an important component of marine food chains and make up around 10% of zooplanktonic biomass (Gorsky and Fenaux, 1998). Despite their abundance and ecological importance, larvacean diversity remains relatively low with only ~70 described species worldwide (Tokioka 1960; Hopcroft 2005; Castellani and Edwards, 2017), belonging to two main families: Oikopleuridae and Fritillariidae (Fig. 1.3; Fenaux et al., 1998).

The families are distinct in anatomical structures, body sizes and house complexity (Alldredge, 1976; Flood and Deibel, 1998).

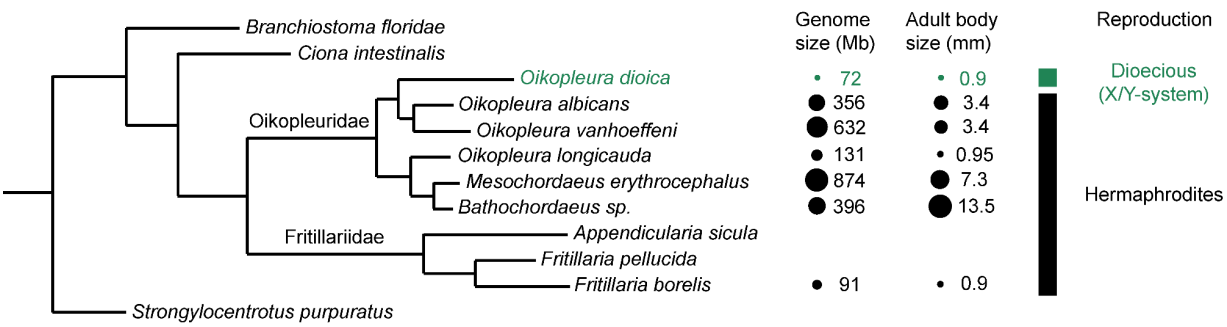


Figure 1.3: Phylogenetic relationships within larvaceans (adapted from Naville et al., 2019).

Larvaceans have a simplified but conserved chordate body plan that they retain throughout life. The tadpole-shaped morphology is very distinctive with a trunk and muscular tail. The size of adult animals varies among species: from ~1-2 mm for *Oikopleura dioica* up to ~8-9 cm among the *Bathochordaeus* species (Fenaux, 1998; Sherlock et al., 2017). All larvaceans are hermophrodites with the only exception, *O. dioica*, described so far (Fig. 1.3).

Larvaceans are enclosed inside a complex cellulose-based structure, the so-called "house", that serves as a filtration device to enable feeding on a concentrated suspension of bacteria and algae extracted from seawater (Bone, 1998). The water currents through the house are generated by the rapid movements of the tail. The size of the house varies from 4 mm (*O. dioica*) to 1 m (giant larvacean from the subfamily Bathochordaeinae; Hamner and Robison, 1992). Larvaceans resynthesize their houses five to eight times a day; the expansion of new houses usually happens within a few minutes. The clogged and discarded houses gradually collapse to the ocean floor taking with them remnants of food and other particles. This way houses make up a significant portion of marine snow - the constant rain of organic matter falling from the upper water layers to the depths (Gorsky and Fenaux 1998). Therefore, larvaceans are considered as major contributors to carbon circulation.

1.3 Morphology of *Oikopleura dioica*

Oikopleura dioica (the Oikopleuridae family) is easily distinguished from other larvaceans as the only reported species with separate sexes (Fig 1.3). Mature males and females are distinguished based on the presence of helmet-like gonads that they carry on top of their trunks (Fig. 1.4a,b). Sex in *O. dioica* is genetically determined: the presence of a male-specific Y-chromosome has been reported (Denoeud et al., 2010; Navratilova et al., 2017). Apart from that, *O. dioica* has two visible subchordal cells in the distal half of the tail (Fig. 1.6a). The number of these cells is considered a differentiating feature of *Oikopleura* species and varies from zero (*O. longicauda*) to one (*O. rufescens*) and many (*O. albicans*; Fredriksson and Olsson, 1991).

The tadpole-like *O. dioica* adults do not grow bigger than ~1-2 mm - the smallest body size reported for a larvacean (Fenaux, 1998; Sherlock et al., 2017). The life span is only five days at 20°C; the animals die soon after releasing sperm and eggs into the water for fertilization. The fecundity of *O. dioica* is high: one female produces more than 300 oocytes on average.

Karyotypic analysis using H3S28p (phospho-histone 3) antibodies to detect centromeres showed that *O. dioica* has a small karyotype of only 3 chromosomes: two autosomes and one sex chromosome (Liu et al., 2020; Denoeud et al., 2010). It can be easily collected from the shore and cultured in laboratories for hundreds of generations (Nishida, 2008; Bouquet et al., 2009; Masunaga et al., 2020). All features combined make *O. dioica* a promising non-classical model species for a wide range of biological studies.

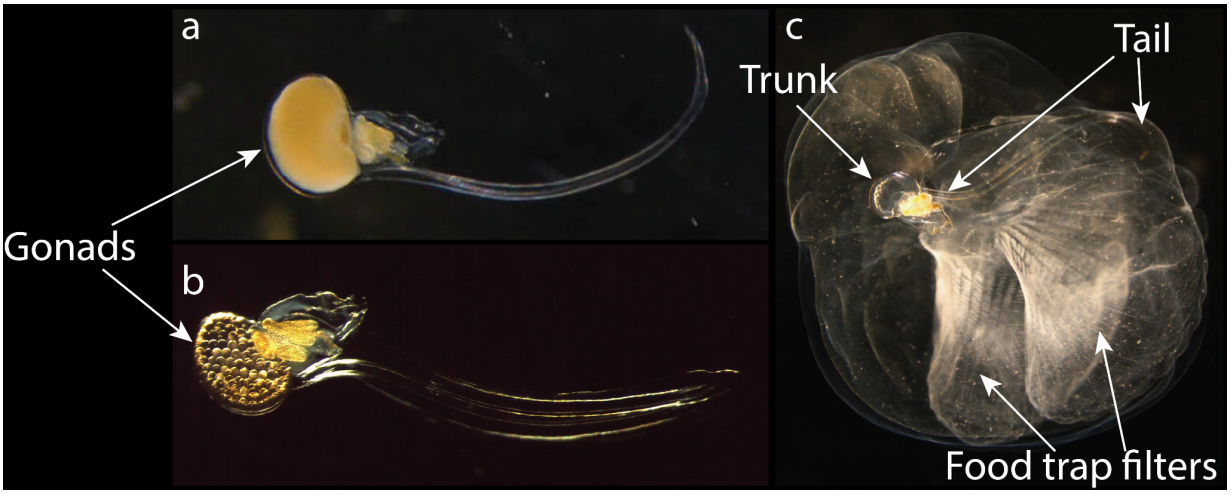


Figure 1.4: Morphology of *Oikopleura dioica*. The adult (a) male and (b) female *O. dioica* with helmet-like gonads (photo credit: Aki Masunaga). (c) *O. dioica* in its house (photo credit: Sars International Center: https://www.uib.no/en/sarssenteret).

O. dioica develops very quickly: organ formation is complete within 10 h (at 20°C) after fertilization. At that point *O. dioica* juveniles are fully functional except for the reproductive system, they are capable of creating their first houses and filter-feeding. The subsequent maturation to males and females with full-grown gonads is complete within 5 days (Nishida 2008; Fig. 1.5). Both embryos and adults are transparent and consist of a small number of cells (10-h juvenile has approximately 4000 cells). The *O. dioica* anatomy and development have been extensively studied and reviewed in detail (Fenaux, 1998; Nishida, 2008; Onuma and Nishida, 2021). Briefly, the tadpole body is divided in two parts, a trunk with endostyle and gill openings (ventral side) and a muscular tail with a notochord and dorsal neural tube. In addition, *O. dioica* has a well described and complex nervous system (Olsson et al., 1990; Cañestro et al., 2005; Glover and Fritzsch, 2009), a vascular system with an easily recognizable heart (Nishida, 2008), and a digestive tract with some extent of organ specialization (Burighel et al., 2001; Fig. 1.6a).

O. dioica builds and lives inside a house, a filter-feeding device made of cellulose, glycopolysaccharides and mucopolysaccharides (Fig. 1.3c; Hosp et al., 2012). It is secreted by the trunk epidermis that is represented by a single layer of the oikoplastic epithelium (Fenaux, 1998; Flood and Deibel, 1998; Thompson et al., 2001). The oikoplastic epithelium is highly patterned, it is composed of a fixed number of cells (~2000) that are grouped into “territories” according to the size and shape of nuclei, extent of polyploidization and gene expression patterns related to the formation of certain house structures (Fig. 1.6c; Thompson et al., 2001; Nishida, 2008; Ganot and Thompson, 2002). The pattern is essential for a quick expansion of a new house and is often used as a species-defining feature (Spada et al., 2001, Flood, 2005).

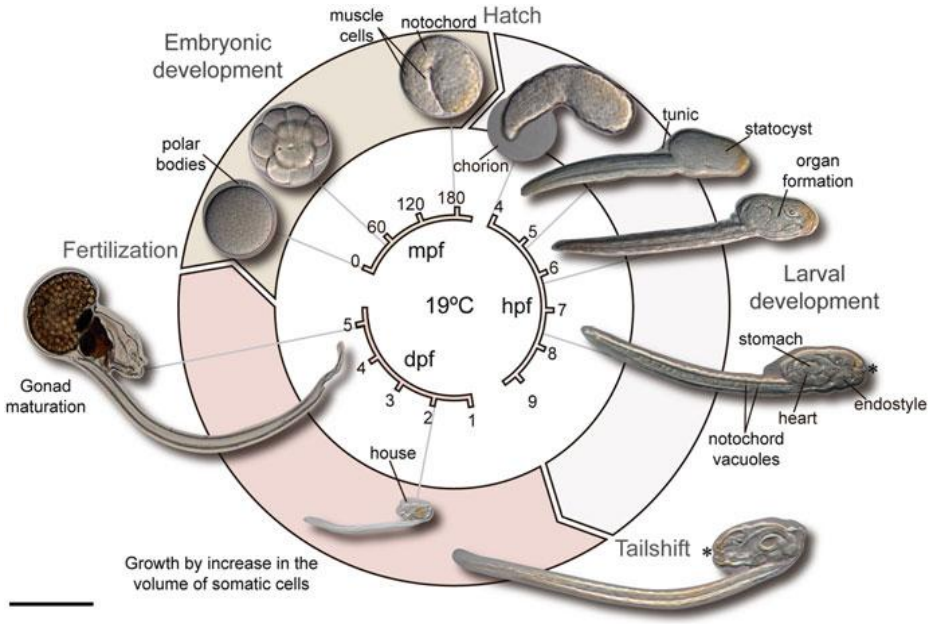


Figure 1.5: Life cycle of *Oikoplura dioica* (from Ferrández-Roldán et al., 2019).

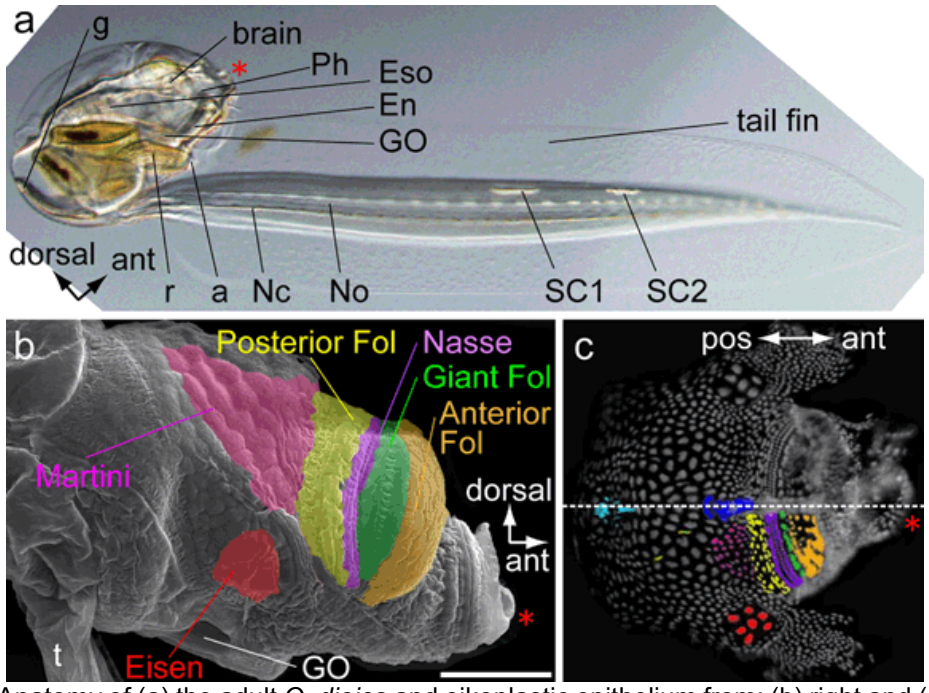


Figure 1.6: Anatomy of (a) the adult *O. dioica* and oikoplastic epithelium from: (b) right and (c) top views. *a* anus, *En* endostyle, *g* gonad, *GO* gill opening, *Nc* nerve cord, *No* notochord, *Ph* pharynx, *r* rectum, *SC1* first subchordal cell, *SC2* second subchordal cell, *t* tail. Asterisk represents the mouth (adapted from: Onuma et al., 2017).

1.4 Genome of *Oikopleura dioica*

O. dioica has a 55-70 Mbp genome, one of the smallest animal genomes identified to date (Seo et al., 2001; Denoeud et al., 2010). By comparison, the genome of the ascidian *Ciona intestinalis* 'type A' (also known as *C. robusta*; Brunetti et al., 2015) is 123 Mbp (Satou et al., 2019), the lancet *Branchiostoma floridae* genome is 520 Mb and the human genome is 3,300 Mbp. The size of other sequenced larvacean genomes varies from 131 Mbp (*Oikopleura longicauda*) to 834 Mbp (*Mesochordaeus erythrocephalus*; Naville et al. 2019).

The sequence of the *O. dioica* genome (here referred to as OdB3, where B stands for Bergen) was first assembled by Denoeud et al. (2010) from the shotgun sequencing of sperm DNA extracted from approximately 200 partially inbred males (11 successive full-sib matings). The males were taken from a laboratory culture derived from a wild population sampled in the North Atlantic Ocean (Bergen coastline, Norway). The total length of the OdB3 assembly is 70.4 Mbp – this falls within the range of the expected genome size (72±13 Mbp) predicted earlier using flow cytometry (Seo et al., 2001). Despite inbreeding, two distinct haplotypes were preserved. Denoeud et al. (2010) released a physical map for OdB3 that was calculated from BAC (bacterial artificial chromosome) end sequences and comprises five linkage groups (LGs): two autosomal LGs, pseudo-autosomal region (PAR) of sex chromosomes, and X- and Y-specific regions (XSR and YSR), suggesting the presence of three pairs of chromosomes.

At 70 Mbp, the OdB3 genome encodes for 18,020 genes with a density of one gene per 3.9 kbp, the highest gene density reported for a chordate genome (Denoeud et al., 2010). In comparison, the human genome has around 20,600 genes, while the lancet genome consists of around 28,500 genes. Within tunicates, the gene density in the *C. intestinalis* genome is three times lower: ~16,500 genes with one gene per 10.5 kbp (Berná and Alvarez-Valin, 2015). The OdB3 assembly is available at the genome browser, OikoBase (http://oikoarrays.biology.uiowa.edu/Oiko/), along with gene and transcript models, functional annotation, expression sequence tags (ESTs) and microarray-based gene expression data (Danks et al., 2013). After its release, the OdB3 assembly has provided key insights into organization and evolution of the *O. dioica* genome and it is still used as the reference genome sequence for many *O. dioica* populations worldwide.

Recently, another sequence assembly was published for a mainland Japanese *O. dioica* (Wang et al., 2020). Here, this assembly is referred to as OSKA2016, where OSKA stands for the Osaka *O. dioica* population. The Osaka population was sampled from Hyogo prefecture and, at the time of sequencing, was already cultured in the laboratory for several years. In contrast with the OdB3 genome, the Osaka *O. dioica* was sequenced with Illumina and PacBio RS II technologies. This assembly is available through the Aniseed database (https://www.aniseed.cnrs.fr/). Total length of OSKA2016 assembly is 56.6 Mbp with 18,743 predicted protein-coding genes. More comparative statistics on *O. dioica* genome assemblies can be found in Table 2.3 (Chapter two). Transcriptomic data for unfertilized eggs and larvae of the Osaka *O. dioica* population is also available (Wang et al., 2015).

The *O. dioica* genome exhibits an unusual level of plasticity (Denoeud et al., 2010). It has undergone an extreme level of compaction accompanied by extensive loss of genes and most ancient animal transposable elements (TEs). The most notable examples are genes that are known to be involved in immune system, epigenetic machinery, apoptotic system, xenobiotic defense systems and developmental mechanisms (Denoeud et al., 2010; Albalat et al., 2012; Weill et al., 2005; Yadetie et al., 2012; Deng et al., 2018; Seo et al., 2004; Edvardsen et

al., 2005). Among the key developmental genes, *O. dioica* has lost the main components of the retinoic acid signaling pathway (Cañestro et al., 2006), and more than 30% of the homeobox (Hox) genes (Edvardsen et al., 2005). In metazoan genomes, Hox genes are organized into an evolutionary conserved cluster that is crucial for anterior-posterior axial patterning during animal development (Carroll, 1995; Pearson et al., 2005). However, that is not the case with the *O. dioica* genome, the remaining Hox genes are scattered across 9 different loci (Seo et al., 2004). Despite that, *O. dioica* still preserves the canonical chordate body plan and developmental trajectories (Denoeud et al., 2010).

Interestingly, the loss of genes and TEs is only partially responsible for the observed compaction of the *O. dioica* genome. Smaller genome size was also achieved through reduction of intergenic and intronic regions, and by packing genes into operons (Fig. 1.7; Denoeud et al., 2010).

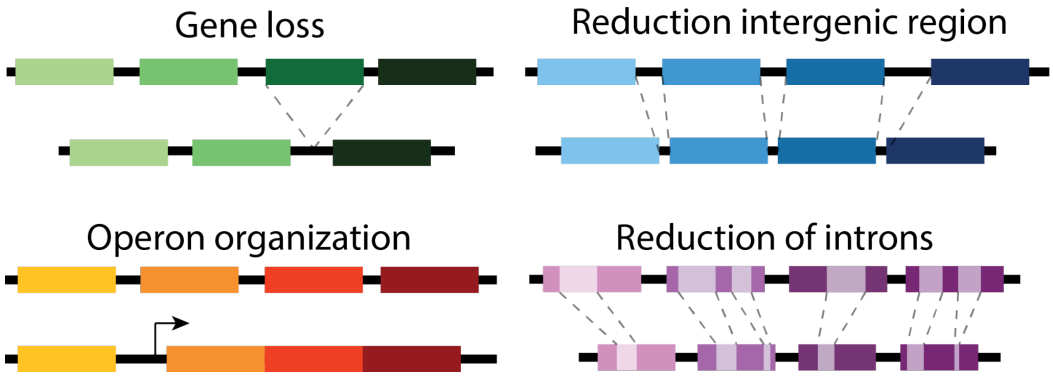


Figure 1.7: Mechanisms of genome reduction in *O. dioica* (adapted from Bern and Alvarez-Valin, 2014).

Operons are polycistronic transcription units mostly found in bacteria and their viruses. So far in metazoans, operons have been identified for nematodes, flatworms and tunicates (Zorio et al., 1994; Blumenthal et al., 2002; Davis and Hodgson, 1997; Satou et al., 2006). *O. dioica* has about 1,800 operons that unite 27% of the genes (Denoeud et al., 2010). Genes in operons are densely-packed and collinear with small intergenic regions that can encode core promoters and binding sites for transcription factors. In *O. dioica*, the highest (74%) fraction of operons are bicistronic with only two genes. The rest of the operons have three or more genes with up to a maximum of 11 genes. Polycistronic pre-mRNA co-transcribed from genes in operons undergo the maturation process by the addition of a *trans*-spliced leader RNA to facilitate translation. This process has been called *trans*-splicing (Ganot et al., 2004; Denoeud et al., 2010). Unlike bacterial operons where genes tend to be functionally related, functions of co-transcribed genes in *Ciona* and *Oikopleura* operons are more loosely related, with a trend towards house-keeping, cell cycle, translation and germline functions (Zeller, 2010; Danks et al., 2015; Wang et al., 2015). In contrast, genes that are involved in development processes, such as morphogenesis and organogenesis, tend to stay outside operons and have relatively large introns and intergenic regions due to the abundance of adjacent regulatory elements (Denoeud et al., 2010).

The genes outside operons are also densely packed, with 53% of intergenic regions being less than 1 kb (Denoeud et al., 2010). Also, the genes in *O. dioica* have 4.1 introns on average. The introns are short with a peak length of 47 bp; only 2.4% of the introns are longer than 1 kbp (Denoeud et al., 2010). Indeed, introns in *O. dioica* have been subjected to a high turnover with a massive loss of old ones (Denoeud et al., 2010). Specifically, only 17% of 5,589 mapped

introns were at ancestral positions (old introns); 76% of introns have been identified as newly acquired (found at genetic loci specific to *O. dioica*); 7% of introns remained unclassified. Indeed, there is a tendency that the largest introns are more often old and canonical with a GT/AG splice-site. Canonical introns are very conservative and represent the majority of introns in most eukaryotic genomes, and the *O. dioica* genome is no exception. However, the frequency of non-canonical introns in *O. dioica* is unusually high – G(non-T)/AG-introns make up around 12% of all annotated introns; the non-canonical splice sites are often hosted by the newly acquired introns. The intron gain in *O. dioica* might have happened through two distinct mechanisms: insertion of transposable elements and reverse splicing (Denoeud et al., 2010). Since *O. dioica* lacks the minor U12 spliceosome, Denoeud et al. (2010) proposed that a single and permissive spliceosome is used instead, with U1snRNP and U2AF being able to recognize both canonical and non-canonical splice sites.

Henriet et al. (2019) showed that non-GT/AG introns are also frequent in other larvacean genomes. In *Fritillaria borealis*, the non-canonical introns are exceptionally diverged (AG/AC and AG/AT are the most frequent), and many (or even most) of them originated from DNA transposons, in particular, miniature inverted-repeat transposable elements (MITEs). Unlike in *O. dioica*, the non-canonical introns in *F. borealis* are correctly processed by the evolutionary conserved U2 spliceosome that evolved a new level of selectivity in this species (Henriet et al., 2019).

The genome size in larvaceans strongly correlates with the animal body size (Fig. 1.3; Naville et al., 2019). Six recently sequenced genomes are bigger (91-874 Mbp) than the one of *O. dioica*, but show no signs of whole-genome duplication events. Moreover, they exhibit similar diversity of TEs, suggesting that the last common ancestor of larvaceans might have also had a small and compact genome. Thus, larger genomes could have occurred through the species-specific bursts in transposon activities, in particular, short interspersed nuclear elements (SINEs), that lack significant homology across species. Their abundance explains more than 83% of genome size variations in larvaceans. In *O. dioica*, SINEs occupy only 0.6% of the genome (Naville et al., 2019).

1.5 DNA repair in the *Oikopleura dioica* genome

A classical (or canonical) non-homologous end-joining (c-NHEJ) is a fundamental pathway that repairs double strand DNA breaks (DSBs). The pathway is almost universal in eukaryotes – seven c-NHEJ proteins (Ku70, Ku80, DNA-PKcs, Lig4, XRCC4, XLF and Artemis) are evolutionary conserved from yeasts to humans. However, genes that encode these proteins could not be identified in the genomes of *O. dioica* and six other larvaceans (Denoeud et al., 2010; Deng et al., 2018). Therefore, Deng et al. (2018) suggested that the c-NHEJ might have been lost in one of the common larvacean ancestors.

Instead of the c-NHEJ, *O. dioica* exploits micro-homologous (MH) sequences, mostly 4-bp long, to join DNA ends after a DSB (Deng et al., 2018). The molecular mechanism underlying MH-dependent repair in *O. dioica* has not yet been described. However, it shows strong similarity with the alternative NHEJ (a-NHEJ), or microhomology-mediated end joining (MMEJ), pathway that is found in other animals and is used as a back-up mechanism to repair DSBs when the c-NHEJ is inhibited. The main candidate genes of the a-NHEJ pathway - Mre11, Parp1, Xrcc1, RAD50, Ligase 1 - were detected in the *O. dioica* genome (Denoeud et al., 2010). In fact, almost all of them are highly expressed in ovaries and oocytes (Danks et al., 2013).

However, that does not exclude a possibility that another, so far undescribed, pathway may be involved (Deng et al., 2018).

It was shown that the presence of c-NHEJ is strongly required for genomic stability (Ferguson et al., 2000; Simsek and Jasin, 2010; Villarreal et al., 2012). However, it is not understood yet whether or not loss of this pathway in *O. dioica* affected its genome evolution. Larger deletions and other rearrangements that often tend to occur when the MH-mediated mechanism is activated would have directly contributed to genome compaction. However, that is not the case in other genomes. For example, *Ciona intestinalis* retains and mostly depends on the c-NHEJ (Deng et la., 2018), but still exhibits a compact genome.

1.6 Genomic diversity of tunicate species

Comparative genomic analyses revealed that tunicates have experienced particularly rapid evolution compared with other deuterostomes (Delsuc et al., 2006; Singh et al., 2009; Denoeud et al., 2010; Tsagkogeorga et al., 2010). The organization of their genomes is highly dynamic, and considerable diversity has been observed within a class (for instance, between the ascidians *Ciona intestinalis* and *Halocynthia roretzi*; Oda-Ishii et al., 2005), genus (such as among species of the genus *Ciona*; Hill et al., 2008; Johnson et al., 2004; Satou et al., 2019), or a single species (for example, *C. intestinalis*; Tsagkogeorga et al., 2012).

Ciona species evolve 50% faster than vertebrates on average (Berná et al., 2009). Comparative genomic studies revealed that two model ascidians, *C. intestinalis* and *C. savignyi*, have higher genomic divergence than humans and chickens (*Gallus gallus*, red jungle fowl; Johnson et al., 2004; Berná et al., 2009). Hill et al. (2008) suggested that extensive genomic rearrangements occurred in two Ciona species (Hill et al., 2008), possibly after their split approximately 122 ± 33 million years ago (Mya; Delsuc et al., 2018). This observation was confirmed by Satou et al. (2019) that found many chromosomal inversions in the genomes of the two *Ciona* species, suggesting that such rearrangements occurred frequently and might have contributed to chromosomal evolution in the *Ciona* genus.

Moreover, multiple phylogenetic and population studies observed a high genomic diversity within the species of *C. intestinalis*. *Ciona intestinalis* is a popular model organism for evo-devo studies and its complete genome sequence has been characterized (Dehal et al., 2002). Wide range of evidence from the analyses of mitochondrial data, microsatellites, five nuclear genes and crossing experiments proved that morphologically indistinguishable strains of *C. intestinalis* ‘type A’ (Northeast Pacific/Mediterranean) and ‘type B’ (Northwest Atlantic) represent two cryptic species (Iannelli et al., 2007; Nydam and Harrison, 2010; Caputi et al., 2007). *Ciona intestinalis* ‘type A; is now also known as *C. robusta* (Brunetti et al. 2015). Indeed, Tsagkogeorga et al. (2012) used a transcriptome-based framework to show that there is a significant diversity among eight wild individuals of the *C. intestinalis* ‘type B’.

O. dioica exhibits an even higher evolutionary rate compared with *Ciona* species: 95% of *O. dioica* genes evolve faster than those in *Ciona* (Berná et al., 2012). A relative rate test across *O. dioica*, *C. intestinalis*, vertebrates and cephalochordates revealed that the average distance between *O. dioica* and an outgroup (vertebrates or cephalochordates) is always higher than those between *C. intestinalis* and the same outgroup. Several independent studies identified *O. dioica* as possibly the fastest-evolving metazoan sequenced so far. In fact, it is always represented by a very long branch in any phylogenetic tree (Delsuc et al., 2006, 2008, 2018; Putnam et al., 2008; Denoeud et al., 2010). *O. dioica* is mainly distinguished from other

larvaceans based on separate sexes and the presence of two subchordal cells on its tail. However, remarkable sequence variations were observed between *O. dioica* collected from Norway and northern Japan on a nucleotide and amino acid levels (Denoeud et al., 2010; Wang et al., 2015, 2020), despite the low level of phenotypic disparity. Therefore, we believe that *O. dioica* exhibit higher within-species diversity that has been suspected before and use of chromosome-scale assemblies may shed light on that question.

1.7 A hybrid approach for genome assembly: from raw reads to complete chromosomes

The development of single-molecule sequencing technologies, such as those from Oxford Nanopore (ONT) and Pacific Biosciences (PacBio), continue to revolutionize the field of genomics. The ONT and PacBio platforms are able to produce reads that are ten kilobases to over a megabase long (Grohme et al., 2018; Tyson et al., 2018; Payne et al., 2019). At such lengths, reads span genomic regions that used to be difficult to reconstruct with only short-read technologies. Therefore, the assemblies of even highly repetitive and complex genomes are now less fragmented and can be further upgraded to complete chromosome sequences with additional data.

Both ONT and PacBio technologies are under constant development and have improved significantly in terms of read length and accuracy over the past years. However, they still exhibit a relatively higher error rate than the short-read NGS (next generation sequencing) technologies, such as Illumina (Laehnemann et al., 2015; Bowden et al., 2019). Therefore, the current standard for *de novo* genome assemblies is to apply a hybrid approach, in which contigs are first assembled with only ONT or PacBio reads and further polished with more accurate Illumina reads generated from the same DNA. As a next step, the polished contigs are joined into longer scaffolds with either Hi-C or Omni-C data. Both Hi-C and Omni-C are chromatin conformation capture methods that produce sequencing data to cover genomic regions in close spatial proximity but are distant linearly (Putnam et al., 2016; https://dovetailgenomics.com/). Compared with the restriction enzyme-based Hi-C technology, Omni-C uses a sequence-independent endonuclease for chromatin digestion, providing higher resolution, especially in genomic regions with a low density of restriction enzyme sites. Use of Hi-C or Omni-C data together with a high-quality contig assembly has a remarkable capability for scaffolding, allowing to resolve near complete telomere-to-telomere genomes.

It is preferred that the DNA for sequencing is extracted from a single individual to reduce heterozygosity levels in genomic data that often result in various assembly errors, such as region duplications. However, a challenge in working with small organisms like *O. dioica* is the amount of DNA that can be extracted. That is why in our laboratory we use the Oxford Nanopore MinION sequencer as it requires a minimum of only 400 ng compared with several µg for PacBio. Moreover, our technical staff have optimized the sequencing protocol to work with even lower DNA input – less than 100 ng (Masunaga et al., 2022).

Despite all the achievements in this field, there are still no definitive guidelines established to verify the correctness and completeness of the assembly. To assess these parameters, one may check if the number of final sequences corresponds to the count of haploid chromosomes determined for the species. As this information is not available for all organisms, especially non-model ones, the total assembly size can be compared to the expected genome

size estimated with other methods (flow cytometry, k-mer counting of short reads). The final assembly may also be examined for the presence of "core" genes, for example, BUSCOs (Benchmarking Universal Single-Copy Orthologs), a set of genes that are highly conserved in species evolutionary close to those that are being sequenced.

Also, one has to keep in mind that the minimum standards to assess the quality of the final assembly can vary depending on what studies the genome is going to be used for. As an example, the VGP (Vertebrate Genome Project) consortium requires more than 95% of the genome, that is used to study chromosomal evolution, to be haplotype phased and assigned to chromosomes. The 50% (N50 length) of the genome length should be in contigs or scaffolds longer than 1 Mb or 10 Mb, respectively, and more than 90% of gene structures should be complete (Rhie et al., 2021). In contrast, an assembly for phylogenomics or population-scale SNP studies may have a low continuity but high base accuracy. Of course, these standards should be taken with caution, especially when working with non-model species. The correct choice of criteria to assess the quality of the final assembly can provide confidence in downstream biological insights.

After confirming the quality of the assembly, it is ready for gene annotation. But before that, the genome sequence has to be properly masked for repeat sequences. Here, the term repeat defines two classes of sequences: low-complexity DNA regions (homopolymeric runs of nucleotides) and mobile elements (retrotransposons and DNA transposons). These sequences can be identified and masked in the genome by homology search with RepeatMasker (Smit et al. at http://repeatmasker.org) using repeat sequences from closely related species available from databases, such as Repbase (Bao et al., 2015) or Dfam (Storer et al., 2021). On the other hand, genome-specific repeat libraries can be generated with *de novo* repeat prediction software specifically trained to identify family-specific structural features, for example, LTR (long terminal repeat) sequences found in LTR retrotransposons or TIRs (terminal inverted repeats) of DNA transposons. Good repeat masking is crucial for the accurate annotation of genes, given that most transposable elements possess protein-coding ORFs and, hence, can be falsely identified as genes, influencing downstream analysis.

After the repeat masking, genes can be identified using either *ab initio* or evidence-based approaches (Yandell and Ence, 2012). The great advantage of the *ab initio* approach is that it requires no external evidence, such as EST or protein sequences, to identify a gene. The tools predict genes only based on the organism-specific genomic traits, such as codon-frequencies and intron/exon distributions, to distinguish gene structures from intergenic regions. Most gene predictors, such as AUGUSTUS (Stanke et al. 2006), already come with files containing pre-calculated parameters. However, such information is often available for genomes of only a few model organisms, like *Drosophila melanogaster, Caenorhabditis elegans* and *Mus musculus*. Therefore, if one works with a non-model organism, the gene predictor needs to be trained on the genome of interest using organism-specific data. For example, AUGUSTUS can be trained with a test dataset containing alignments of ESTs, RNA-Seq, protein and many other sequences. However, performing this step can be challenging given that training AUGUSTUS is a supervised procedure and requires basic programming skills. Fortunately, Hoff and Stanke (2019) released a comprehensive step-by-step guide explaining how to train AUGUSTUS for the annotation of individual genomes. After the training, gene predictor can be run on the genome with the pre-calculated species parameters using an evidence-driven approach, where transcriptome, RNA-Seq and/or protein sequence alignments are treated as hints for exon/intron structures. Evidence-driven gene prediction is computationally more demanding but has greater

potential to provide accurate gene models than *ab initio* approach, and, thus, is considered more standard practice for non-model eukaryotic genomes.

1.8 Summary and thesis outline

The tiny chordate, *O. dioica*, contributes to a wide range of biological research areas, including developmental biology, evolution and ecology (Nishida, 2008; Lombard et al., 2009; Troedsson et al., 2013; Deng et al., 2018; Onuma and Nishida 2021; Ferrández-Roldán et al., 2021). Moreover, much of our knowledge of larvacean biology comes from studies on *O. dioica*, as it can be easily cultured in the laboratory for many generations (Bouquet et al., 2009; Martí-Solans et al., 2015; Masunaga et al., 2020). *O. dioica* maintains a classical chordate-like morphology throughout its life, despite having little synteny preserved with the ancestral chordate linkage groups. The organization of its genome is highly dynamic: it has undergone an extreme level of compaction, resulting in the smallest and fastest-evolving non-parasitic animal genome sequenced so far (Seo et al., 2001; Denoeud et al., 2010). This process has been accompanied by the loss of genes and most pan-animal transposable elements. Even gene clusters that are conserved throughout metazoan genomes, such as Hox genes, are entirely dispersed in *O. dioica* (Seo et al., 2004; Edvardsen et al., 2005; Blanchoud et al., 2018). It is believed that DNA repair, which is microhomology-dependent in *O. dioica*, has contributed to genome compaction and rearrangement (Deng et al., 2018).

Two features that clearly distinguish *O. dioica* from other larvacean species are separate sexes and the presence of two subchordal cells in its tail. *O. dioica* is characterized by its ubiquitous distribution, given that all populations of dioecious oikopleurids around the globe exhibit a low level of phenotypic disparity. Owing to that and the limited availability of genomic sequencing data for this species, the population structure and genomic diversity of *O. dioica* are still unclear. To fill this gap, we decided to perform a cross-comparison of three *O. dioica* populations on the level of whole genomes, the results of which I present in this thesis.

For this research project, chromosome-scale genome sequences are required in order to better refine the gene order and rule out potential assembly artifacts. The genome of *O. dioica* is highly polymorphic (Denoeud et al., 2010), making the assembly of its complete sequence challenging. Chapter two takes a hybrid approach of integrating multiple sequencing technologies with the Hi-C confirmation results to generate a first *de novo* chromosome-scale assembly of an *O. dioica* individual from Okinawa (OKI2018_I69). In this chapter, we discuss the quality of the final genome assembly and annotation in comparison to the previously published genomes for Bergen (OdB3; Denoeud et al., 2010) and Osaka *O. dioica* (OSKA2016; Wang et al., 2020). Given the chromosomal resolution of the final OKI2018_I69 assembly, we investigate whether any genomic features are distributed differently along chromosome arms.

Chapter three presents results from cross-genome comparisons of three *O. dioica* populations from globally distributed locations: one from North Atlantic (Barcelona/Bergen) and two from Pacific (Osaka/Aomori and Okinawa/Kume) Oceans. Through collaborations with the Cañestro laboratory at the University of Barcelona and the Nishida laboratory at Osaka University, we received samples of the Barcelona and Osaka *O. dioica* individuals. The samples of *O. dioica* from Kume and Aomori were collected by our technical staff, Aki Masunaga, Yongai Tan and Andrew Liu. Chapter three investigates the preservation of genome synteny across the populations on nucleotide and protein levels, providing first evidence that *O. dioica* exhibits higher genetic diversity than has been suspected before.

Chapter four gives a final glance at the updated annotations of genes and repeats in the *O. dioica* genomes and an outlook of further research that could be built onto the work and data presented in this thesis.

Finally, chapter five provides an overview of the thesis results and conclusions, discussing the possibility of the existence of multiple lineages of dioecious *Oikopleura* around the globe and how and when they might have diverged.

Chapter Two

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

This chapter is published as:

Bliznina A., Masunaga A., Mansfield M.J., Tan Y., Liu A.W., West C., Rustagi T., Chien H.C., Kumar S., Pichon J., Plessy C., Luscombe N.M. (2021). Telomere-to-telomere assembly of the genome of an individual *Oikopleura dioica* from Okinawa using Nanopore-based sequencing. *BMC genomics, 22*(1): 1-18. doi:10.1186/s12864-021-07512-6

Members of the Genomics and Regulatory Systems Unit at OIST contributed to this project as following: I led the project, performed genome assembly, annotation and analysis; Aki Masunaga, Yongkai Tan, and Andrew Liu generated the sequencing data; Hsiao-Chiao Chien assisted in generating the contig assembly; Charlotte West generated the Sankey plot; Charles Plessy, Tanmay Rustagi, Saurabh Kumar and Julien Pichon performed analysis of the mitochondrial genome assembly and BUSCO genes; Michael Mansfield studied the distribution of various genomic features across chromosome arms.

Abstract

Background

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

Results

Here, we present a chromosome-scale genome assembly (OKI2018_I69) of the Okinawan *O. dioica* produced using long-read Nanopore and short-read Illumina sequencing data from a single male, combined with Hi-C chromosomal conformation capture data for scaffolding. The OKI2018_I69 assembly has a total length of 64.3 Mbp distributed among 19 scaffolds. 99% of the assembly is contained within five megabase-scale scaffolds. We found telomeres on both ends of the two largest scaffolds, which represent assemblies of two fully contiguous autosomal chromosomes. Each of the other three large scaffolds have telomeres at one end only and we propose that they correspond to sex chromosomes split into a pseudo-autosomal region and X-specific or Y-specific regions. Indeed, these five scaffolds mostly correspond to equivalent

linkage groups in OdB3, suggesting overall agreement in chromosomal organization between the two populations. At a more detailed level, the OKI2018_I69 assembly possesses similar genomic features in gene content and repetitive elements reported for OdB3. The Hi-C map suggests few reciprocal interactions between chromosome arms. At the sequence level, multiple genomic features such as GC content and repetitive elements are distributed differently along the short and long arms of the same chromosome.

Conclusions

We show that a hybrid approach of integrating multiple sequencing technologies with chromosome conformation information results in an accurate de novo chromosome-scale assembly of *O. dioica*'s highly polymorphic genome. This genome assembly opens up the possibility of cross-genome comparison between *O. dioica* populations, as well as of studies of chromosomal evolution in this lineage.

2.1 Background

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

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

O. dioica's genome size is 55–70 Mbp (Seo et al., 2001; Denoeud et al., 2010), making it one of the smallest among all sequenced animals. Interestingly, genome-sequencing of other larvacean species uncovered large variations in genome sizes, which correlated with the expansion of repeat families (Naville et al., 2019). *O. dioica* is distinguished from other larvaceans as the only reported dioecious species (Fredriksson and Olsson, 1991) with sex determination system using an X/Y pair of chromosomes (Denoeud et al., 2010). The first published genome assembly of *O. dioica* (OdB3, B stands for Bergen) was performed with Sanger sequencing which allowed for high sequence accuracy but limited coverage (Denoeud et al., 2010). The OdB3 assembly was scaffolded with a physical map produced from BAC-end sequences, which revealed two autosomal linkage groups and a sex chromosome with a long pseudo-autosomal region (PAR; Denoeud et al., 2010). Recently, a genome assembly for a mainland Japanese population of *O. dioica* (OSKA2016, OSKA denotes Osaka) was published, which displayed a high level of coding sequence divergence compared with the OdB3 reference (Wang et al., 2015; Wang et al., 2020). Although OSKA2016 was sequenced with single-molecule long reads produced with the PacBio RSII technology, it does not have chromosomal resolution.

Historical attempts at karyotyping *O. dioica* by traditional histochemical stains arrived at different chromosome counts, ranging between n = 3 (Körner, 1952) and n = 8 (Colombera

and Fenaux, 1973). In preparation for this study, we karyotyped the Okinawan *O. dioica* by staining centromeres with antibodies targeting phosphorylated histone H3 serine 28 (Liu et al., 2020), and determined a count of n = 3. This is also in agreement with the physical map of OdB3 (Denoeud et al., 2010).

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

Here, we present a chromosome-length assembly of the Okinawan *O. dioica* genome sequence generated with datasets stemming from multiple genomic technologies and data types, namely long-read sequencing data from Oxford Nanopore, short-read sequences from Illumina and Hi-C chromosomal contact maps (Fig. 2.1).

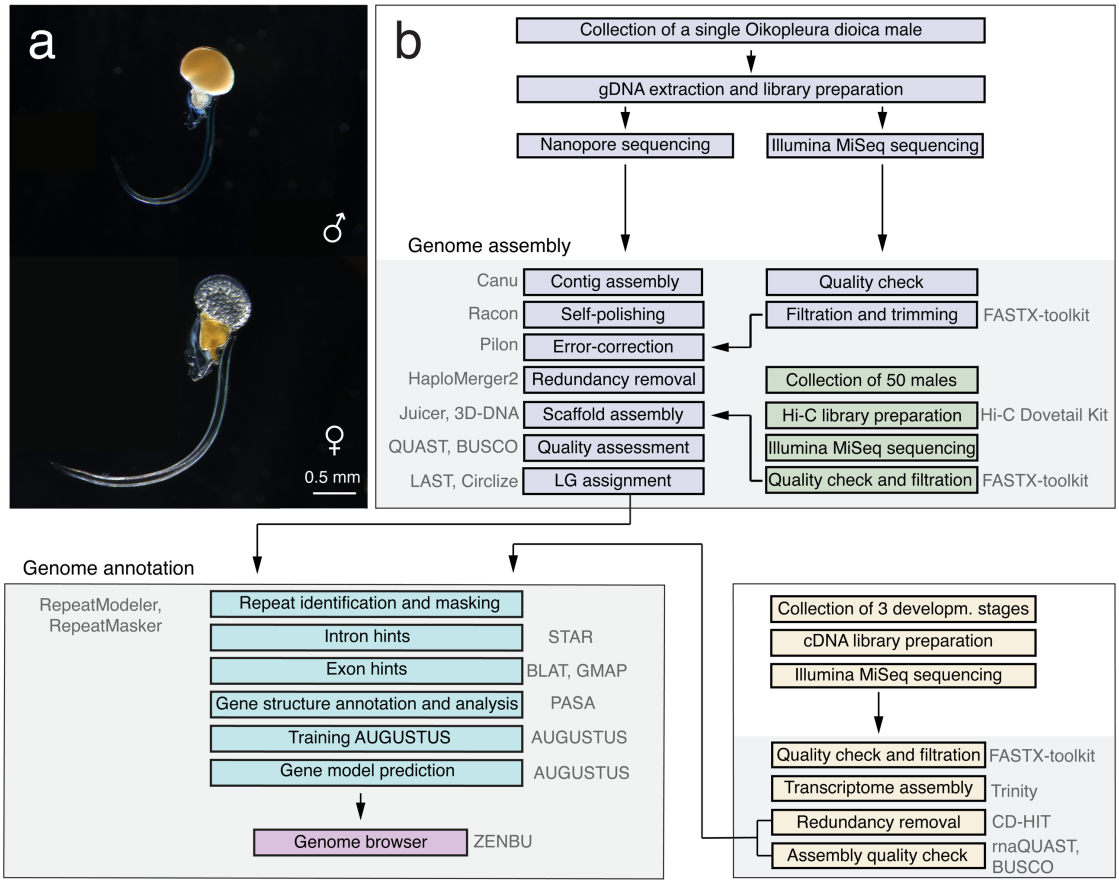


Figure 2.1: Genome assembly and annotation workflow used to generate the OKI2018_I69 genome assembly. (a) Life images of adult male (top) and female (bottom) *O. dioica*. (b) The assembly was generated using Nanopore and Illumina data, followed by scaffolding using Hi-C chromosomal capture information data.

2.2 Methods

2.2.1 *Oikopleura* sample and culture

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

2.2.2 Isolation and sequencing of DNA

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

Isolated genomic DNA used for long-reads on Nanopore MinION platform were processed with the Ligation Sequencing Kit (Nanopore LSK109) according to manufacturer's protocol, loading approximately 200 ng total sample per R9.4 flow-cell. Raw signals were converted to sequence files with the Guppy proprietary software (model "template_r9.4.1_450bps_large_flipflop", version 2.3.5). Approximately 5 ng was set aside for whole genome amplification to perform sequencing on Illumina MiSeq platform, using the TruePrime WGA Kit (Sygnis, 370,025) according to manufacturer's protocol. Magnetic bead purification (Promega, NG2001) was employed for all changes in buffer conditions required for enzymatic reactions and for final buffer suitable for sequencing system. Approximately 1 µg of amplified DNA was sequenced by our core sequencing facility with a 600-cycle MiSeq Reagent Kit v3 (Illumina, MS-102-3003) following the manufacturer's instructions. These Illumina runs were used for polishing and error checking of Nanopore runs.

2.2.3 Hi-C library preparation

50 fully matured males were rinsed three times for 10 min each by transferring from well to well in a 6-well plate filled with 5 ml FASW. Rinsed animals were combined in a 1.5 ml

microcentrifuge tube. Tissues were pelleted for 10 min at 12,000 rpm and leftover FASW was discarded. A Hi-C library was then prepared by following the manufacturer's protocol (Dovetail, 21,004). Briefly, tissues were cross-linked for 20 min by adding 1 ml 1× PBS and 40.5 µl 37% formaldehyde to the pellet. The tubes were kept rotating to avoid tissue settle during incubation. Cross-linked DNA was then blunt-end digested with DpnII (Dovetail) to prepare ends for ligation. After ligation, crosslinks were reversed, DNA was purified by AMPure XP Beads (Beckman, A63880) and quantified by Qubit 3 Fluorometer (Thermo Fisher Scientific, Q10210). The purified DNA was sheared to a size of 250–450 bp by sonication using a Covaris M220 instrument (Covaris, Woburn, MA) with peak power 50 W, duty factor 20, and cycles/burst 200 times for 65 s. DNA end repair, adapter ligation, PCR enrichment, and size selection were carried out by using reagents provided with the kit (Dovetail, 21,004). Finally, the library was checked for quality and quantity on an Agilent 4200 TapeStation (Agilent, 5067–5584) and a Qubit 3 Fluorometer. The library was sequenced on a MiSeq (Illumina, SY-410-1003) platform using a 300 cycles V2 sequencing kit (Illumina, MS-102-2002), yielding 20,832,357 read pairs.

2.2.4 Genome size estimation

Jellyfish (Marçais and Kingsford, 2011) was used to generate k-mer count profiles for various values of *k* (17, 21, 25, 29, 33, 37, and 41) based on the genome-polishing Illumina MiSeq reads, with a maximum k-mer count of 1000. These k-mer profiles were subsequently used to estimate heterozygosity and genome size parameters using the GenomeScope web server (Vurture et al., 2017).

2.2.5 Filtering of Illumina MiSeq raw reads

Before using at different steps, all raw Illumina reads were quality-filtered (−q 30, −p 70) and trimmed on both ends with the FASTX-Toolkit v0.0.14 (Gordon and Hannon, 2010). The quality of the reads before and after filtering were checked with FASTQC v0.11.5 (Andrews, 2010). Read pairs that lacked one of the reads after the filtering were discarded in order to preserve paired-end information.

2.2.6 Genome assembly

Genome assembly was conducted with the Canu pipeline v1.8 (Koren et al., 2017) and 32.3 Gb (~ 221.69×) raw Nanopore reads (correctedErrorRate = 0.105, minReadLength = 1000). The resulting contig assembly was polished three times with Racon v1.2.1 (Vaser et al., 2017) using Canu-filtered Nanopore reads. Nanopore-specific errors were corrected with Pilon v1.22 (Walker et al., 2014) using filtered 150-bp paired-end Illumina reads (~ 99.7×). Illumina reads were aligned to the Canu contig assembly with BWA v0.7.17 (Li, 2013) and the corresponding alignments were provided as input to Pilon. Next, one round of the HaploMerger2 processing pipeline (Huang et al., 2017) was applied to eliminate redundancy in contigs and to merge haplotypes.

Contigs were joined into scaffolds based on long-range Hi-C Dovetail™ data using Juicer v1.6 (Durand, Shamim et al., 2016) and 3D de novo assembly (3D-DNA; Dudchenko et al., 2017) pipelines. The megabase-scale scaffolds were joined into pairs of chromosome arms based on the assumption of conserved synteny with the OdB3 physical map. The candidate assembly was visualized and reviewed with Juicebox Assembly Tools v1.11.08 (JBAT; Durand, Robinson et al., 2016).

Whole-genome alignment between OKI2018_I69 and OdB3 assemblies was performed using LAST v1066 (Kiełbasa et al., 2011). The sequence of OdB3 linkage groups were reconstructed as defined in the Supplementary Fig. 2 ("Draft chromosome scale assembly based on scaffolds of the reference genome sequence") in Denoeud et al. (2010). The resulting alignments were post-processed in R with a custom script (https://github.com/oist/oikGenomePaper) and visualized using the R package "networkD3" ("sankeyNetwork" function). The color scheme for chromosomes was adopted from R Package RColourBrewer, "Set2".

The final assembly was checked for contamination by BLAST searches against the NCBI non-redundant sequence database. 12 smaller scaffolds were found to have strong matches to bacterial DNA (Table 2.1), as well as possessing significantly higher Nanopore sequence coverage (> 500×) than the rest of the assembly, and were therefore removed from the final assembly.

The completeness and quality of the assembly were checked with QUAST v5.0.2 (Gurevich et al., 2013) and by searching for the set of 978 highly conserved metazoan genes (OrthoDB version 9.1; Zdobnov et al., 2017) using BUSCO v3.0.2 (Simão et al., 2015; Waterhouse et al., 2018). The --sp option was set to match custom AUGUSTUS parameters (Hoff and Stanke, 2019) trained using the Trinity transcriptome assembly (see below) split 50% / 50% for training and testing.

Table 2.1: Contaminations found in smaller scaffolds of the OKI2018_I69 assembly.

Contig	Length	Depth of coverage (Nanopore reads)	Depth of coverage (Illumina MiSeq reads)	Top BLAST hits
HiC_scaffold_15	6809	17127.64	0	*Escherichia coli* chromosome/cloning vector (50% coverage over 3447bp, 99% identity)
HiC_scaffold_16	6224	17304.25	0	*Escherichia coli* chromosome/cloning vector (97% coverage over 3525bp, 99% identity)
HiC_scaffold_17	5528	18306.63	1.99	*Escherichia coli* chromosome/cloning vector (62% coverage over 3447bp, 99% identity)
HiC_scaffold_18	4785	2161.17	0.10	*Escherichia coli* chromosome (66% coverage over 1734bp, 99% identity)
HiC_scaffold_19	4434	17283.97	1.95	*Escherichia coli* chromosome (75% coverage over 1909bp, 99% identity)
HiC_scaffold_20	4093	1398.96	1.39	*Escherichia coli* chromosome (73% coverage over 1523bp, 99% identity)
HiC_scaffold_21	4008	1396.25	0.75	*Escherichia coli* chromosome (72% coverage over 1718bp, 99% identity)
HiC_scaffold_22	3462	613.98	0.35	*Escherichia coli* chromosome (69% coverage over 1232bp, 99.11% identity)
HiC_scaffold_23	3220	7421.60	2.11	*Bacteriophage sp.* isolate 181; Escherichia cloning vector (82% coverage over 2165bp, 99.45% identity)
HiC_scaffold_24	2963	7002.27	4.87	*Escherichia coli* chromosome (80% coverage over 2405bp, 99% identity)
HiC_scaffold_25	2452	1144.96	0.05	*Escherichia coli* chromosomes (54% coverage over 1341bp, 99% identity)
HiC_scaffold_26	2142	4.99	0	*Pseudomonas spp.*, chromosome/plasmid (99% coverage 2134bp, 99.48% identity)

2.2.7 Repeat masking and transposable elements

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

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

2.2.8 Developmental staging, isolation and sequencing of mRNA, transcriptome assembly

Mixed stage embryos, immature adults (3 days after hatching) and adults (4 days after hatching) were collected separately from our on-going laboratory culture for RNA-Seq analysis. Eggs were washed three times for 10 min by moving eggs along with micropipette from well to well in a 6-well dish each containing 5 ml of FASW and left in a fresh well of 5 ml FASW in the same dish. These were stored at 17 °C and set aside for fertilization. Matured males, engorged with sperm, were also washed 3 times in FASW. Still intact mature males were placed in 100 µl of fresh FASW and allowed to spawn naturally. Staged embryos were initiated by gently mixing 10 µl of the spawned male sperm to the awaiting eggs in FASW at 23 °C. Generation 30 developing embryos at 1 h and 3 h post-fertilization were visually verified by dissecting microscope and collected as a pool for the mixed staged embryo time point. Immature adults at generation 31 and sexually differentiated adults at generation 30 were used for the two adult staged time points. All individuals for each time point were pooled and washed with FASW three times for 10 min. Total RNA was extracted and isolated with RNeasy Micro Kit (Qiagen, 74,004) and quantitated using Qubit 3 Fluorometer (Thermo Fisher Scientific, Q10210). Additional quality control and integrity of isolated total RNA was checked using Agilent 4200 TapeStation (Agilent, 5067–5576). Further processing for mRNA selection was performed with Oligo-d(T)25 Magnetic Beads (NEB, E7490) and the integrity of the RNA was validated once more with Agilent 4200 TapeStation (Agilent, 5067–5579). Adapters for the creation of DNA libraries for the Illumina platform were added per manufacturer's guidance (NEB, E7805) as were unique indexed oligonucleotides (NEB, E7600) to each of the three staged samples. Each cDNA library was sequenced paired-end with a 300-cycle MiSeq Reagent Kit v2 (Illumina, MS-102-2002) loaded at approximately 12 pM.

After quality assessment and data filtering (see Filtering of Illumina MiSeq raw reads), Illumina RNA-Seq reads were pooled together and *de novo* assembled with Trinity v2.8.2 (Grabherr et al., 2011). Redundancy in the transcriptome assembly was removed by cd-hit

v4.8.1 (Li and Godzik, 2006) with a cut-off value of 95% identity. The quality and completeness of the transcriptome assembly was verified with rnaQUAST v1.5.1 (Bushmanova et al., 2016) and BUSCO.

2.2.9 Gene prediction and annotation

Gene models were predicted using AUGUSTUS v3.3 (Stanke et al., 2006). AUGUSTUS was trained following the Hoff and Stanke protocol (Hoff and Stanke, 2019) with the initial RNA-Seq reads and transcriptome assembly used as intron and exon hints, correspondingly. Transcript models were generated with the PASA pipeline v20140417 (Haas et al., 2003) using BLAT v36 and GMAP v2018-02-12 to align transcripts to the genome. RNA-Seq reads were mapped to the genome with STAR v2.0.6a (Dobin et al., 2013). Running AUGUSTUS using hints resulted in a set of 17,277 protein-coding genes and 18,811 transcript models. Chromosomal coordinates were ported to our final assembly using the Liftoff tool (Shumate and Salzberg, 2021) filtering out 17 genes and corresponding transcripts. The quality of the predicted gene models was assessed with BUSCO.

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

2.2.10 Detection of coding RNAs

A translated alignment was used to detect known *O. dioica* genes available from GenBank using the TBLASTN software (Gertz et al., 2006) with the options "-ungapped -comp_based_stats F" to prevent *O. dioica*'s small introns from being incorporated as alignment gaps, and -max_intron_length 100,000 to reflect the compactness of *O. dioica*'s genome. The best hits were converted to GFF3 format using BioPerl's bp_search2gff program (Stajich et al., 2002) before being uploaded to the ZENBU genome browser (Severin et al., 2014). For some closely related pairs of genes that gave ambiguous results with that method, we searched for the protein sequence in our transcriptome assembly with TBLASTN, located the genomic region where the best transcript model hit was aligned, and selected the hit from the original TBLASTN search that matched this region. We summarized our results in Appendix 1. For both searches, we used an E-value filter of 10^{-40}. Genes marked as not found in the table might be present in the genome while failing to pass the filter.

2.2.11 Detection of non-coding RNAs

To validate the results of cmscan on rRNAs, genomic regions were screened with a nucleotide BLAST search using the *O. dioica* isolate MT01413 18S ribosomal RNA gene, partial sequence (GenBank:KJ193766.1). 200-kbp windows surrounding the hits were then analyzed with the RNAmmer 1.2 web service (Lagesen et al., 2007). RNAmmer did not detect the 5.8S RNA, but we could confirm its presence by a nucleotide BLAST search using the AF158726.1 reference sequence. The loci containing the 5S rRNA (AJ628166) and the spliced leader RNA (AJ628166) were detected with the exonerate 2.4 software (Slater and Birney, 2005), with its affine:local model and a score threshold of 1000 using the region chr1:8487589–8,879,731 as a query.

2.2.12 Whole-genome alignments

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

2.2.13 Nanopore read realignments

Nanopore reads were realigned to the genome with the LAST software (Kiełbasa et al., 2011) as in the whole-genome alignments above. FASTQ qualities were discarded with the option –Q0 of lastal. Optimal split alignments were calculated with last-split. Alignment blocks belonging to the same read were joined with maf-convert -J 1e6 and the custom script syntenic_regions_stranded.sh. The resulting GFF3 files were loaded in the ZENBU genome browser to visualize the alignments near gap regions in order to check for reads spanning the gaps.

2.2.14 Analysis of sequence properties across chromosome-scale scaffolds

Each chromosome-scale scaffold was separated into windows of 50 kbp and evaluated for GC content, repeat content, sequencing depth, and the presence of DpnII restriction sites. For chr1, chr2, and the PAR, windows corresponding to long and short chromosome arms were separated based on their positioning relative to a central gap region (chr1 short arm: 1–5,191,657 bp, chr1 long arm: 5,192,156-14,533,022 bp; chr2 short arm: 1–5,707,009, chr2 long arm: 5,707,508-16,158,756 bp; PAR short arm: 1–6,029,625 bp, PAR long arm: 6,030,124-17,092,476). Since none of our assemblies or sequencing reads spanned both the PAR and either sex-specific chromosome, the X and Y chromosomes were excluded from this analysis. For each of GC content, sequencing depth, repeat content, gene count, and DpnII restriction sites, the significance of the differences between long and short arms was assessed with Welch's two-sided T test as well as a nonparametric Mann-Whitney test implemented in R (Table 2.2). The results of the two tests were largely in agreement, but groups were only indicated as significantly different if they both produced significance values below 0.05 ($p < 0.05$).

Table 2.2: Statistics results for the analysis of sequence properties across chromosome-scale scaffolds in the OKI2018_I69 genome assembly.

Chromosome	Variable	Comparison	Welch's two-sided T test			Wilcoxon rank sum test with continuity correction		
			statistics	degrees of freedom	p_value	statistics	degrees of freedom	p_value
chr1	GC	short arm vs. long arm	8.249560417	160.676	5.41539E-14	15138	289	1.89957E-15
chr1	Depth	short arm vs. long arm	6.148269875	201.945	4.11958E-09	14013.5	289	2.80065E-10
chr1	Repetitive regions	short arm vs. long arm	-5.652255201	134.7954	9.05379E-08	4660.5	289	2.58083E-13
chr1	Gene count	short arm vs. long arm	0.158067145	181.2778	0.874579829	9359.5	289	0.638076149
chr1	DpnII sites	short arm vs. long arm	4.267339627	187.535	3.13608E-05	12759.5	289	7.35493E-06
chr2	GC	short arm vs. long arm	9.420654022	242.0464	3.73638E-18	19291.5	322	9.71857E-20
chr2	Depth	short arm vs. long arm	-0.490563818	269.5921	0.62413402	12467.5	322	0.537086311
chr2	Repetitive regions	short arm vs. long arm	-3.489588876	207.9979	0.000590457	7625	322	6.83519E-08
chr2	Gene count	short arm vs. long arm	0.679643529	238.0144	0.497390647	12642	322	0.403027205
chr2	DpnII sites	short arm vs. long arm	4.100284161	284.7456	5.38876E-05	15535	322	9.54534E-06
PAR	GC	short arm vs. long arm	11.11845982	254.6067	1.13965E-23	22101.5	340	8.05645E-24
PAR	Depth	short arm vs. long arm	-0.047655828	134.4092	0.962061276	17335	340	4.21408E-06
PAR	Repetitive regions	short arm vs. long arm	-4.210464487	179.5016	4.02491E-05	9000.5	340	7.44088E-07
PAR	Gene count	short arm vs. long arm	1.277661482	245.9621	0.2025734	14376.5	340	0.225039821
PAR	DpnII sites	short arm vs. long arm	4.672020687	236.7953	5.00372E-06	17333.5	340	4.23669E-06

2.3 Results

2.3.1 Genome sequencing and assembly

O. dioica's genome is highly polymorphic (Denoeud et al., 2010), making assembly of its complete sequence challenging. To reduce the level of variation, we sequenced genomic DNA from a single *O. dioica* male. The low amount of extracted DNA is an issue when working with small-size organisms like *O. dioica*. Therefore, we optimized the extraction and sequencing protocols to allow for low-template input DNA yields of around 200 ng and applied a hybrid sequencing approach using Oxford Nanopore reads to span repeat-rich regions and Illumina reads to correct individual nucleotide errors. The Nanopore run gave 8.2 million reads (221× coverage) with a median length of 840 bp and maximum length of 166 kb (Fig. 2.2a). Based on k-mer counting of the Illumina reads, the genome was estimated to contain ~ 50 Mbp (Fig. 2.2b) – comparable in size to the OdB3 and OSKA2016 assemblies – and a relatively high heterozygosity of ~ 3.6%. We used the Canu pipeline (Koren et al., 2017) to correct, trim and assemble Nanopore reads, yielding a draft assembly comprising 175 contigs with a weighted median N50 length of 3.2 Mbp. We corrected sequencing errors and local misassemblies of the draft contigs with Nanopore reads using Racon, and then with Illumina reads using Pilon. The initial Okinawa *O. dioica* assembly length was 99.3 Mbp, or ~ 1.5 times longer than the OdB3 genome at 70.4 Mbp. Merging haplotypes with HaploMerger2 resulted in two sub-assemblies (reference and allelic) of 64.3 Mbp with an N50 of 4.7 Mbp. Repeating the procedure on a second individual from the same culture showed overall agreement in assembly lengths, sequences and structures (Fig. 2.2c).

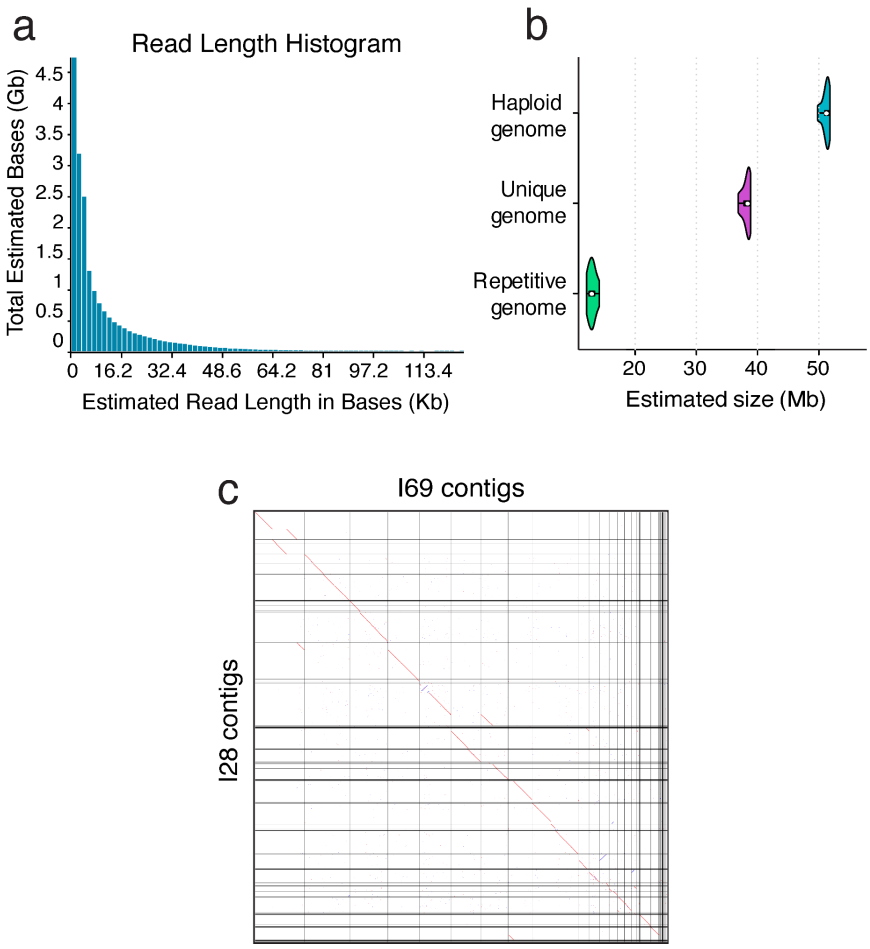


Figure 2.2: Quality control checks implemented on different steps of genome sequencing and assembly. (a) Graph showing length distribution of raw Nanopore reads used to generate the OKI2018_I69 assembly. (b) Estimated total and repetitive genome size based on *k*-mer counting of the Illumina paired-end reads used for polishing the OKI2018_I69 assembly. (c) Pairwise genome alignment of the contig assemblies of I69 and I28 *O. dioica* individuals.

To scaffold the genome, we sequenced Hi-C libraries from a pool of ~ 50 individuals from the same culture. More than 99% of the Hi-C reads could be mapped to the contig assembly. After removing duplicates, Hi-C contacts were passed to the 3D-DNA pipeline to correct major misassemblies, as well as order and orient the contigs. The resulting assembly consisted of 8 megabase-scale scaffolds containing 99% of the total sequence (Fig. 2.3a), and 14 smaller scaffolds that account for the remaining 663 kbp (lengths ranging from 2.9 to 131.6 kbp). One of the small scaffolds is a draft assembly of the mitochondrial genome that we discuss below. Most of the other smaller scaffolds are highly repetitive and might represent unplaced fragments of centromeric or telomeric regions. We annotated telomeres by searching for the TTAGGG repeat sequence and found that most of the megabase-scale scaffolds have single telomeric regions: therefore, we reasoned that they represent chromosome arms. Indeed, pairwise genome alignment to OdB3 identified two syntenic scaffolds for each autosomal linkage group, two for the pseudo-autosomal region (PAR) and one for each sex-specific region. Since we had previously inferred a karyotype of n = 3 by immunohistochemistry (Liu et al., 2020), we completed the assembly by pairing the megabase-scale scaffolds into chromosome

arms based on the assumption of conserved synteny with the OdB3 physical map (Fig. 2.3b). The final assembly named OKI2018_I69 comprises telomere-to-telomere assemblies of the autosomal chromosomes 1 (chr1) and 2 (chr2). The sex chromosomes are split into pseudo-autosomal region (PAR) and X-specific region (XSR) or Y-specific region (YSR; Table 2.3, 2.4; Fig. 2.3). We assume that the sex-specific regions belong to the long arm of the PAR, as the long arm does not contain any telomeric repeats (Fig. 2.4a). Alignment of the Illumina polishing reads to the OKI2018_I69 assembly estimated an error rate of 1.3% showing high sequence accuracy.

The genome-wide contact matrix from the Hi-C data (Fig. 2.3c) shows bright, off-diagonal spots that suggest spatial clustering of the telomeres and centromeres both within the same and across different chromosomes (Dudchenko et al., 2017). The three centromeric regions are outside the sex-specific regions, dividing the PAR and both autosomes into long and short arms. The two sex-specific regions have lower apparent contact frequencies compared with the rest of the assembly which is consistent with their haploid status in males. The chromosome arms themselves show few interactions between each other, even when they are part of the same chromosome.

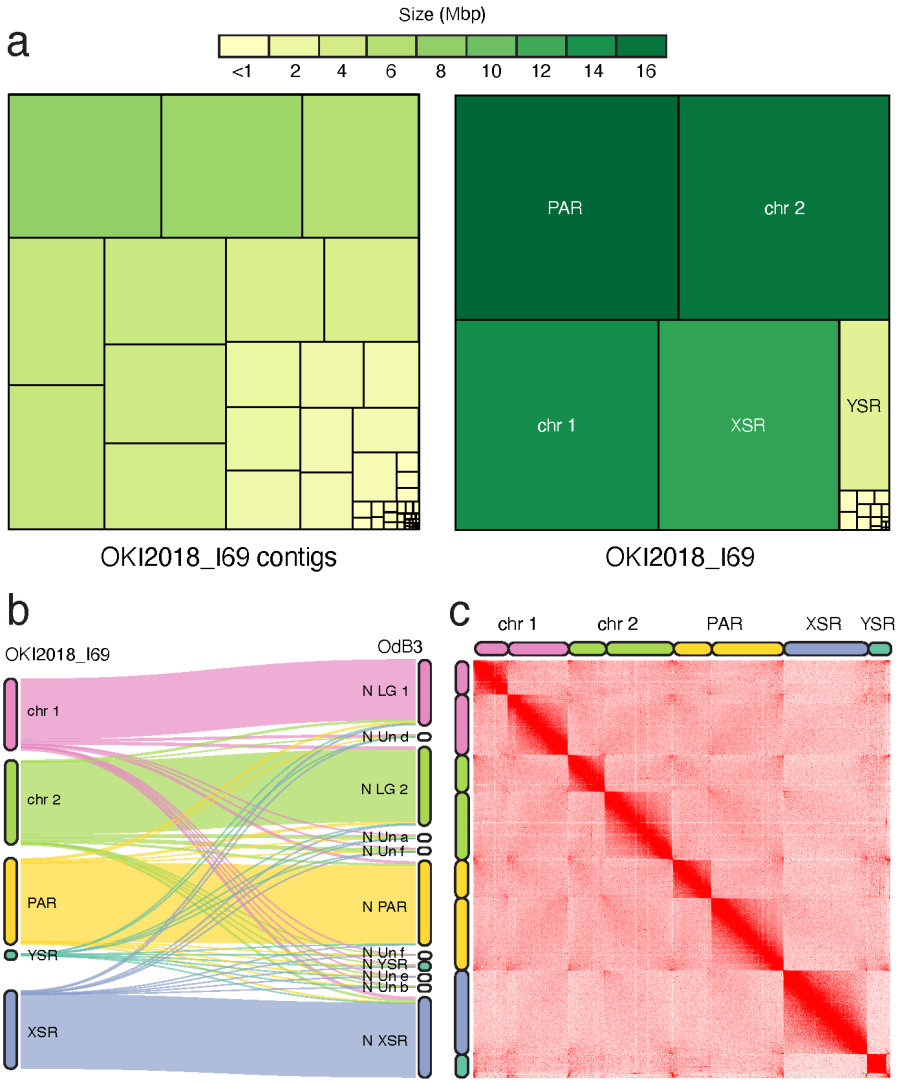


Figure 2.3: OKI2018_I69 assembly of the Okinawan *O. dioica*. (a) Treemap comparison between the contig (left) and scaffold (right) assemblies of the *O. dioica* genome. Each rectangle represents a contig or a scaffold in the assembly with the area proportional to its length. (b) Comparison between the OKI2018_I69 (left) and OdB3 (right) linkage groups. The Sankey plot shows what proportion of each chromosome in the OKI2018_I69 genome is aligned to the OdB3 linkage groups. (c) Contact matrix generated by aligning Hi-C data set to the OKI2018_I69 assembly with Juicer and 3D-DNA pipelines. Pixel intensity in the contact matrices indicates how often a pair of loci collocate in the nucleus.

Table 2.3: Comparison of the OKI2018_I69 assembly with the previously published *O. dioica* genomes.

	OdB3	OSKA2016	OKI2018_I69
Geographical origin	Bergen, Norway (North Atlantic)	Hyogo, Japan (North Pacific)	Okinawa, Japan (Ryukyu archipelago)
Assembly length (Mbp)	70.4	56.6	64.3
Number of scaffolds	1,260	576	19
Longest scaffold (Mbp)	3.2	6.8	17.1
Scaffold N50 (Mbp)	0.4	1.5	16.2
Number of contigs	5,917	746	42
Contig N50 (Mbp)	0.02	0.6	4.7
GC content (%)	39.77	41.34	41.06
Gap rate (%)	5.589	0.585	0.034
Complete BUSCOs (%)	70.8	71.7	73.01

Table 2.4: Per-scaffold statistics of the OKI2018_I69 genome assembly.

Scaffold	Length (Mbp)	Number of contigs	Number of protein-coding genes	Repetitive sequences (%)	GC content (%)	Gaps (%)	Depth of coverage (Nanopore reads)	Median coverage (Nanopore reads)
chr1	14.533	6	3943	13.86	40.68	0.018%	247.18	258
chr2	16.159	10	4499	12.67	40.86	0.028%	257.24	261
PAR	17.092	8	4797	11.94	41.09	0.059%	255.07	261
XSR	12.959	2	3798	8.83	42.08	0.004%	140.32	137
YSR	2.916	2	170	54.9	40.44	0.135%	163.1	133
chrUn_1	0.011	1	4	0.88	42.04	0	130.51	136
chrUn_2	0.058	1	20	3.42	41.79	0	136.26	137
chrUn_3	0.132	1	1	80.66	31.44	0	114.11	108
chrUn_4	0.111	1	5	76.77	38.31	0	128.12	125
chrUn_5	0.080	1	0	83.68	35.18	0	99.72	102
chrUn_6	0.068	1	0	91.2	29.47	0	121.67	96
chrUn_7	0.057	1	20	11.68	42.71	0	126.52	127
chrUn_8	0.035	1	0	86.93	33.44	0	260.98	238
chrUn_9	0.035	1	0	78.01	33.81	0	134.53	114
chrUn_10	0.004	1	3	2.18	41.42	0	140.09	141
chrUn_11	0.014	1	0	98.62	45.97	0	244.69	138
chrUn_12	0.009	1	0	2.43	27.51	0	126.73	130
chrUn_13	0.006	1	0	81.41	15.89	0	294.69	279
chrUn_14	0.003	1	0	61.48	36.16	0	129.76	132
Sum	64.28	42	17260	–	–	–	–	–

2.3.2 Chromosome-level features

The genome contains between 1.4 and 2.6 Mbp of tandem repeats (detected using the tantan and ULTRA algorithms respectively with maximum period lengths of 100 and 2000). Subtelomeric regions tend to contain retrotransposons or tandem repeats with longer periods. We also found telomeric repeats in smaller scaffolds. A possible explanation is that subtelomeric regions display high heterozygosity, leading to duplicated regions that fail to assemble with the chromosomes. Alternatively, these scaffolds could be peri-centromeric regions containing interstitial telomeric sequences. In some species, high-copy tandem repeats can be utilized to discover the position of centromeric regions (Melters et al., 2013); however, we could not find such regions. Additional experimental techniques such as chromatin immunoprecipitation and sequencing with centromeric markers might be necessary to resolve the centromeres precisely. Therefore, the current assembly skips over centromeric regions, represented as gaps of arbitrary size of 500 bp in the chromosomal scaffolds.

We studied genome-scale features by visualizing them along whole chromosomes, from the short to long arm, centered on their centromeric regions. Most strikingly, there is a clear difference in sequence content between chromosome arms (Fig. 2.4; Table 2.2). The short arms consistently display depleted GC content and elevated repetitive content compared with the corresponding long arms. Although GC content tends to be weakly negatively correlated with repeat content, it is not currently possible to ascertain causality and the mechanism behind the marked difference in sequence content between the short and long chromosome arms remains unknown. It should be noted that the differences in GC contents affects the density of the GATC DpnII restriction enzyme recognition sites used for Hi-C library preparation; however, this bias is insufficient to explain the low degree of intra-chromosomal interaction observed in the Hi-C contact maps.

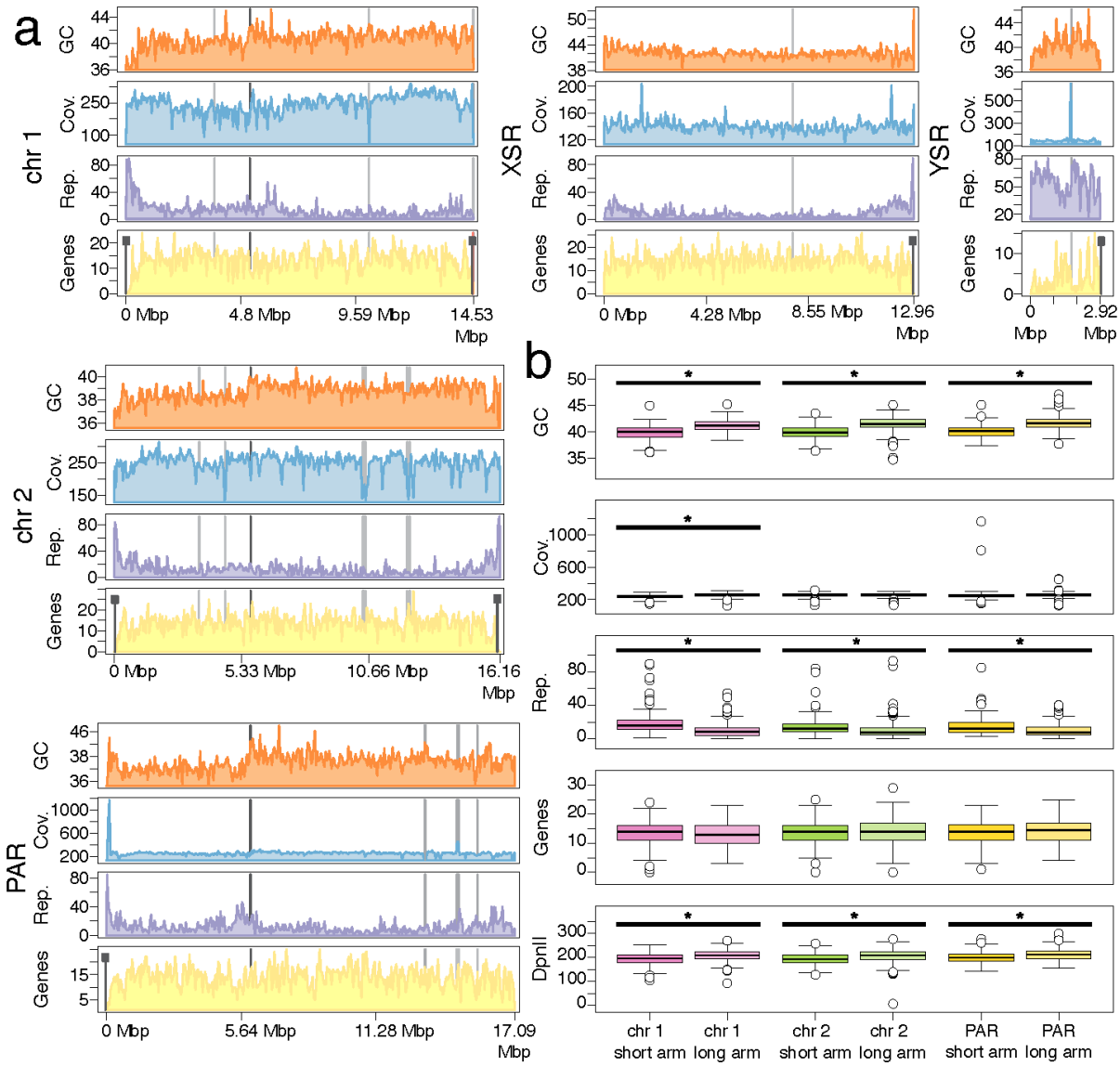


Figure 2.4: Chromosome-level features of the Okinawan *O. dioica* genome. (a) Visualization of sequence properties across chromosomes in the OKI2018_I69 assembly. For each chromosome, 50 kbp windows of GC (orange), Nanopore sequence coverage (blue), the percent of nucleotides masked by RepeatMasker (purple), and the number of genes (yellow) are indicated. Differences in these sequence properties occur near predicted sites of centromeres and telomeres, as well as between the short and long arms of each non-sex-specific chromosome. Telomeres and gaps in the assembly are indicated with black and grey rectangles, respectively. (b) Long and short chromosome arms exhibit significant differences in sequence properties, including GC content, repetitive sequence content, and the number of restriction sites recognized by the DpnII enzyme used to generate the Hi-C library.

2.3.3 Quality assessment using BUSCO

To assess the completeness of our assembly, we searched for 978 metazoan Benchmarking Universal Single-Copy Orthologs (BUSCOs) provided with the BUSCO tool (Simão et al., 2015; Waterhouse et al., 2018; Zdobnov et al., 2017). To increase sensitivity, we trained BUSCO's gene prediction tool, AUGUSTUS (Hoff and Stanke, 2019), with transcript models generated from RNA-Seq data collected from the same laboratory culture (see below). We detected 73.0% of BUSCOs, which is similar to OdB3 and OSKA2016 (Fig. 2.5a; Table 2.5). All detected BUSCOs except one reside on the chromosomal scaffolds. As the reported fraction of detected genes is lower than for other tunicates such as *Ciona intestinalis* HT (94.6%; Satou et al., 2019) or *Botrylloides leachii* (89%; Blanchoud et al., 2018), we searched for BUSCO genes in the transcriptomic training data (83.0% present) and confirmed the presence of all but one by aligning the transcript sequence to the genome. We then inspected the list of BUSCO genes that were found neither in the genome nor in the transcriptome. Bibliographic analysis confirmed that BUSCO genes related to the peroxisome were lost from *O. dioica* (Žárský and Tachezy, 2015; Kienle et al., 2016). There are two possible explanations for the remaining missing genes: first is that protein sequence divergence (Berná et al., 2012) or length reduction (Berná and Alvarez-Valin, 2015) in *Oikopleura* complicate detection by BUSCO, and second is gene loss. In line with the possibility of gene loss, most BUSCO genes missing from our assembly are also undetectable in OdB3 and OSKA2016 (Fig. 2.5b; Appendix 2). To summarize, the Okinawa assembly achieved comparable detection of universal single-copy conserved orthologs compared with previous *O. dioica* assemblies, and consistently undetectable genes may have been lost or diverged extensively in *Oikopleura*.

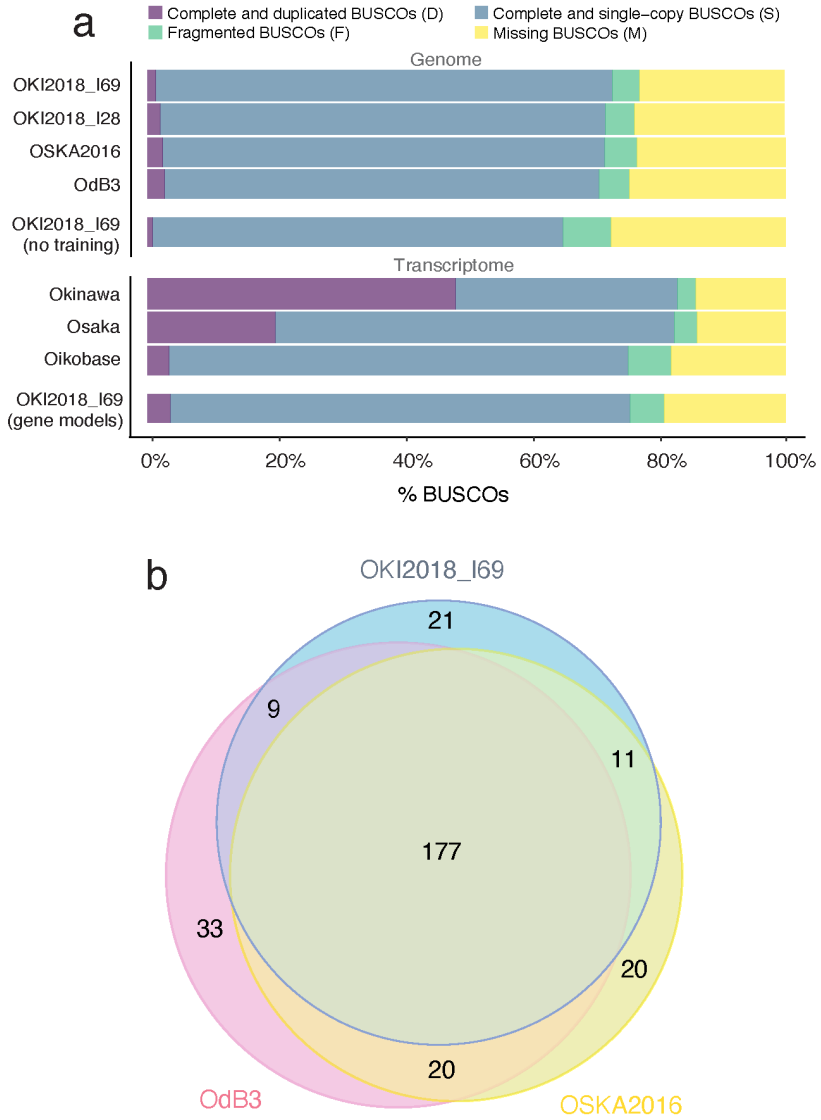


Figure 2.5: Quality assessment of the OKI2018_I69 genome assembly. (a) Proportion of BUSCO genes detected or missed in *Oikopleura* genomes and transcriptomes. The search on the OKI2018_I69 assembly was repeated with default parameters ("no training") to display the effect of AUGUSTUS training. (b) Venn diagram showing the number of BUSCO genes missing in OKI2018_I69, OdB3 and/or OSKA2016 genomes.

Table 2.5: BUSCO scores for genome and transcriptome assemblies.

		Complete (C)	Complete and single-copy (S)	Complete and duplicated (D)	Fragmented (F)	Missing (M)	Total BUSCO genes searched
Genomes	OKI2018_I69	714	701	13	41	223	978
		73%	71.7%	1.33%	4.2%	22.8%	—
	OKI2018_I28	703	683	20	45	230	978
		71.9%	69.8%	2%	4.6%	23.5%	—
	OSKA2016	701	677	24	49	228	978
		71.7%	69.2%	2.5%	5%	23.3%	—
	OdB3	692	665	27	47	239	978
		70.8%	68%	2.8%	4.8%	24.4%	—
	OKI2018_I69 (no training)	637	629	8	74	267	978
		65.1%	64.3%	0.8%	7.6%	27.3%	—
Transcriptomes	Okinawa	812	340	472	29	137	978
		83%	34.8%	48.3%	3%	14%	—
	Osaka	808	612	196	34	136	978
		82.6%	62.6%	20%	3.5%	13.9%	—
	OikoBase (Bergen)	737	704	33	66	175	978
		75.4%	72%	3.4%	6.7%	17.9%	—
	OKI2018_I69 (gene models)	736	662	74	47	195	978
		75.3%	67.7%	7.6%	4.8%	19.9%	—

2.3.4 Repeat annotation

In order to identify repetitive elements in the OKI2018_I69 genome, we combined the results of several de novo repeat detection algorithms and used this custom library as an input to RepeatMasker to identify repeat sequences. Interspersed repeats make up 14.4% of the assembly (9.25 Mbp; Fig. 2.6), comparable to the 15% reported for OdB3 (Denoeud et al., 2010). Of the annotated elements, the most abundant type is the long terminal repeat (LTRs; ~4.6%) with Ty3/gypsy *Oikopleura* retrotransposons (TORs) dominating 2.97 Mbp of the sequence. Short interspersed nuclear elements (SINEs) make up a smaller portion of the OKI2018_I69 sequence (< 0.1%) compared with the OdB3 (0.6%). It has been suggested that SINEs contribute significantly to genome size variation in other oikopleurids (Naville et al., 2019), but further analysis is required to determine whether that is the case at shorter evolutionary distances. Non-LTR LINE/Odin and Penelope-like elements are large components of most oikopleurid genomes (Naville et al., 2019), but they are almost absent from the OKI2018_I69 assembly. Indeed, 44% of the predicted repeats in the Okinawan *O. dioica* could not be classified through searches against repeat databases and may either represent highly divergent relatives of known repeat classes, or novel repeats specific to Okinawan *O. dioica*.

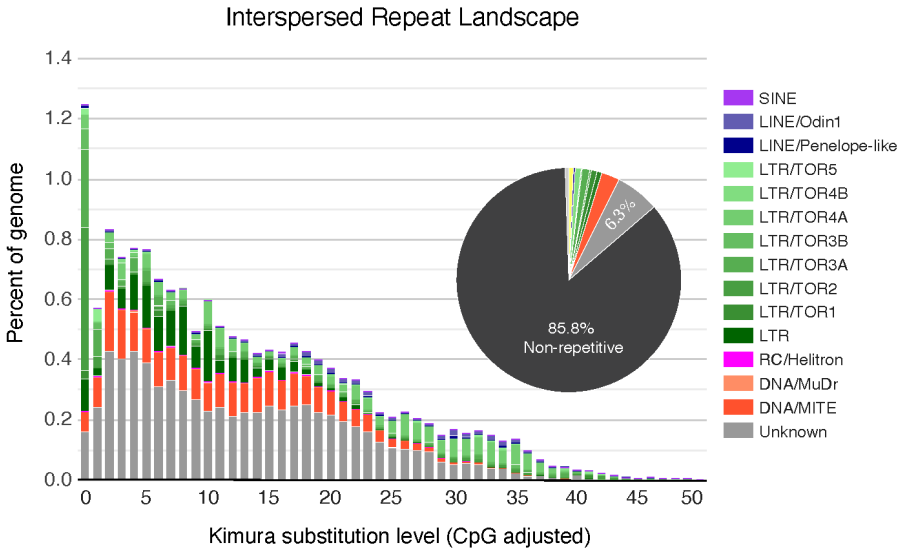


Figure 2.6: Analysis of repetitive elements. The repeat landscape and proportions of various repeat classes in the genome are indicated and color-coded according to the classes shown on the right side of the figure. The non-repetitive fraction of the genome is shown in black.

2.3.5 Gene annotation

We annotated the OKI2018_I69 assembly using RNA-Seq-based gene prediction. RNA-Seq reads mapped to the assembly showed 99.14% agreement between the genome and transcriptome indicating high sequence accuracy. Annotation of the genome yielded 18,794 transcript isoforms distributed among 17,260 protein-coding genes. The number of predicted genes for the OKI2018_I69 is slightly lower than what was reported for OdB3 (18,020; Denoeud et al., 2010) and OSKA2016 (18,743; Wang et al., 2020; Table 2.6). The rest of the genes are either lost from the Okinawan *O. dioica* genome or were not assembled and/or annotated with our pipeline. On the other hand, the higher number of genes might be artifacts of the OdB3 and OSKA2016 annotations. The completeness of the annotation compares to the genome: BUSCO

recovered 75.3% complete and 4.8% fragmented metazoan genes (Fig. 2.5a). Like the OdB3 assembly, gene density is very high at one gene per 3.7 kbp. OKI2018_I69 has similar gene and exon length distributions, and very short introns with a median length of only 49 bp (Table 2.6). Indeed, we found a high frequency of the non-canonical (non-GT/AG) introns in the OKI2018_I69 (11%). Previously, Denoeud et al. (2010) reported that 12% of the introns were non-canonical in the OdB3. Some of those non-canonical introns were found in the same genes as in the OdB3. However, more close examination is required to understand if it is the case for the rest of the genes. Therefore, overall genomic features seem to be conserved among *O. dioica* populations despite the large geographic distance.

Table 2.6: Comparison of the annotations of the three *O. dioica* genome assemblies.

	OdB3	OSKA2016	OKI2018_I69
Masked sequence (%)	15.0	–	14.4
Number of genes	18,020	18,743	17,260
Median gene length (bp)	1,488	1,483	1,505
Median exon length (bp)	159	155	152
Median intron length	48	51	49

The ribosomal DNA gene encoding the precursor of the 18S, 5.8S and 28S rRNAs occurs as long tandem repeats that form specific chromatin domains in the nucleolus. We identified 4 full tandem copies of the rDNA gene at the tip of the PAR's short arm, separated by 8738 bp (median distance). As this region has excess coverage of raw reads, and assemblies of tandem repeats are limited by the read length (99% of Nanopore reads in our data are shorter than 42,842 bp), we estimate that the real number of the tandem rDNA copies could range between 20 (MiSeq) and 100 (Nanopore) copies. Between or flanking the rDNA genes, we also found short tandem repeats made of two to three copies of a 96-bp sequence. This tandem repeat is unique to the rDNA genes and to our reference and draft genomes, and was not found in the OdB3 reference nor in other larvacean genomes. The 5S rRNA is transcribed from loci distinct to the rDNA gene tandem arrays. In *Oikopleura*, they have the particularity of being frequently associated with the spliced leader (SL) gene and to form inverted repeats present in more than 40 copies (Ganot et al., 2004). We found 27 copies of these genes on every chromosomal scaffold except YSR, 22 of which were arranged in inverted tandem repeats. Altogether, we found in our reference genome one rDNA gene repeat region assembled at the end of a chromosome short arm. This sequence might provide useful markers for phylogenetic studies in the future.

2.3.6 Draft mitochondrial genome scaffold

We identified a draft mitochondrial genome among the smaller scaffolds, chrUn_12, by searching for mitochondrial sequences using the Cox1 protein sequence and the ascidian mitochondrial genetic code (Pichon et al., 2019). Automated annotation of this scaffold using the MITOS2 server detected the coding genes *cob*, *cox1*, *nad1*, *cox3*, *nad4*, *cox2*, and *atp6* (Fig. 2.7a), which are the same as in Denoeud et al. (2010) except for the *nd5* gene that is missing from our assembly. The open reading frames are often interrupted by T-rich regions, in line with

Denoeud et al. (2010). However, we cannot rule out the possibility that these regions represent sequencing errors, as homopolymers are difficult to resolve with the Nanopore technology available in 2019. The cob gene is interrupted by a long non-coding region, but this might be a missassembly. Indeed, an independent assembly using the flye software (Kolmogorov et al., 2019) with the --meta option to account for differential coverage also produced a draft mitochondrial genome, but its non-coding region was ~ 2 kbp longer. Moreover, a wordmatch dotplot shows tandem repeats in this region (Fig. 2.7b), and thus this region is prone to assembly errors, especially with respect to the number of repeats. Altogether, the draft contig produced in our assembly shows as a proof of principle that sequencing reads covering the mitochondrial genome alongside the nuclear genome can be produced from a single individual, although it may need supporting data such as targeted resequencing in order to be properly assembled.

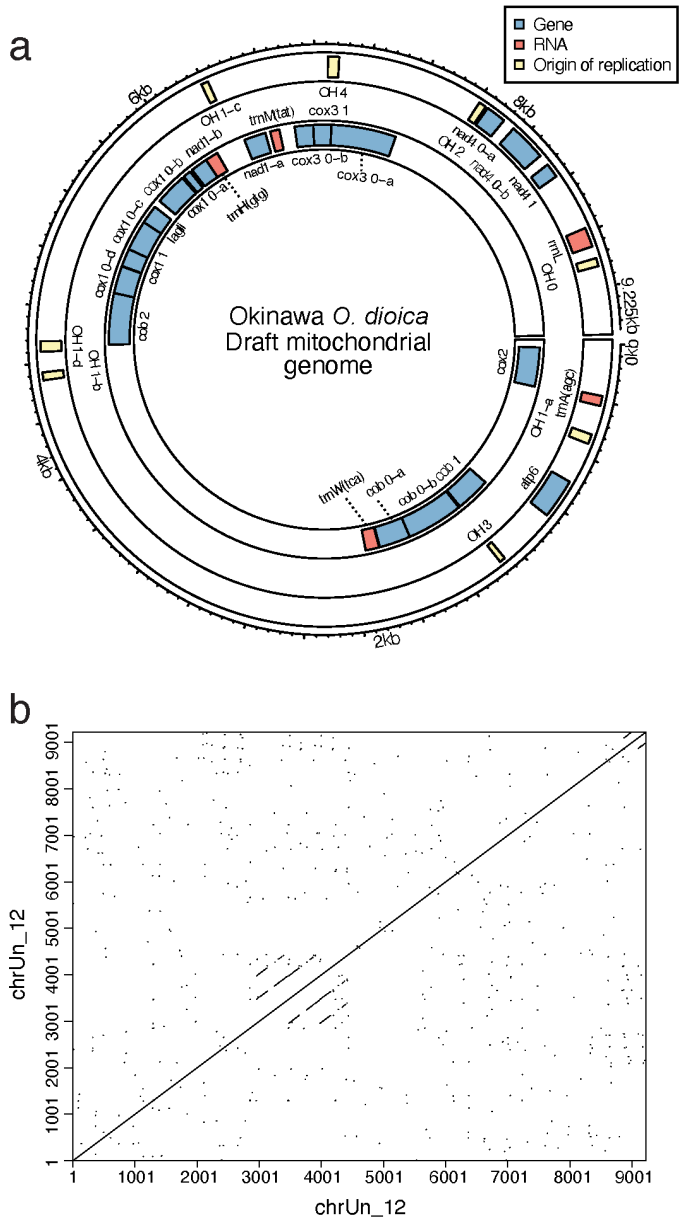


Figure 2.7: Draft scaffold of the mitochondrial genome in the OKI2018_I69 assembly. (a) Predicted gene annotation of the draft mitochondrial genome sequence. (b) Self-similarity plot of the draft mitochondrial genome sequence. A tandem repeat can be seen, which complicates the complete assembly of the mitochondrial genome from whole-genome sequencing data.

2.4 Discussion

2.4.1 OKI2018_I69 assembly quality

Previously, different techniques have been used to sequence and assemble *O. dioica* genomes which have produced assemblies of varying quality. The Sanger-based OdB3 sequence was published in 2010 (Denoeud et al., 2010). Due to limitations in sequencing technologies at the time, it is highly fragmented, comprising 1260 scaffolds with an N50 of 0.4 Mbp. The recently released OSKA2016 assembly was generated from long-read PacBio data and, therefore, has a larger N50 and fewer scaffolds (Table 2.4; Wang et al., 2020). Both assemblies have high sequence quality and nearly full genome coverage, but neither of them contains resolved chromosomes. However, Denoeud et al. (2010) released a physical map calculated for OdB3 from BAC end sequences that comprises five linkage groups (LGs): two autosomal LGs, one pseudo-autosomal region of sex chromosomes, and two sex specific regions (X and Y).

The use of reference chromosome information from a closely related species to order contigs or scaffolds into chromosome-length sequences is a common way to generate final genome assemblies (Drosophila 12 Genomes Consortium et al., 2007). However, this approach precludes discovery of structural variants. In our study, we first assembled long Nanopore reads de novo into contigs that we ordered and joined into megabase-scale scaffolds using long-range Hi-C data. The synteny-based approach with OdB3's linkage groups as a reference was only required to guide final pairing of chromosome arms into single scaffolds of chr1, chr2 and PAR, as we found that these scaffolds mostly align to one of the autosomal LGs or PAR. Therefore, any potential assembly errors in OdB3 would not be transferred to our assembly. Apart from these syntenic relationships, our karyotyping results and the count of three centromeres on the Hi-C contact map supports the presence of three pairs of chromosomes in the Okinawan *O. dioica*. However, there is a possibility that chromosome arms might have been exchanged between chromosomes in the Okinawan population. Additional experimental evidence is needed to confirm the pairing of chromosome arms, such as data generated by the Omni-C method which does not rely on restriction enzyme fragmentation.

Our synteny-based scaffolding is based on the simplest definition of synteny meaning "on the same chromosome". It does not make assumptions on gene order, which is why we report our results with a position-independent Sankey plot in Fig. 2.3b. We initially assumed that animals collected from the Atlantic and Pacific oceans are from the same species and conserve these chromosomal properties. However, there are visible differences in gene number, gene order and repeat content compared with the OdB3 and OSKA2016. *O. dioica* is distributed all over the world, and all the populations are classified as a single species owing to the lack of obvious morphological differences and limited understanding of population structure. However, the short life span of *O. dioica* combined with limited mobility and high mutation rate contribute to an accelerated genome evolution that might have led to multiple speciation events. Sequence polymorphism was previously noted when comparing the OdB3 genome to genomic libraries of a laboratory strain collected on the North American Pacific coast (Denoeud et al., 2010), and more recently when comparing OdB3 to OSKA2016 (Wang et al., 2015; Wang et al., 2020). The chromosome-scale OKI2018_I69 assembly opens up the possibility for further work on cross-comparison among *O. dioica* populations that will elucidate the relation of the Okinawan populations to the North Atlantic and North Pacific ones.

2.4.2 Inter-arm contacts

The sequence of *O. dioica*'s chromosomes and their contact map suggest that chromosome arms may be the fundamental unit of synteny in larvaceans. Hi-C contact matrices in vertebrates typically display greater intra-chromosomal than inter-chromosomal interactions. A similar pattern was reported in the tunicate *Ciona robusta* (also known as *intestinalis* type A; Satou et al., 2019) and the lancelet *Branchiostoma floridae* (Simakov et al., 2020). By comparison, in flies and mosquitoes, the degree of contacts between two arms of the same chromosome appears to be reduced but nonetheless more frequent than between different chromosomes (Dudchenko et al., 2017). Indeed, in *Drosophila*, the chromosome arms – which are termed Muller elements owing to studies with classical genetics (Schaeffer, 2018) – are frequently exchanged between chromosomes across speciation events. *O. dioica*'s genome shares with fruit flies its small size and small number of chromosomes. However, small chromosome size is also seen in the tunicate *Ciona robusta*, which has 14 meta- or sub-meta-centric pairs (Shoguchi et al., 2005), with an average length of ~ 8 Mbp (Satou et al., 2019) that exhibit a more extensive degree of contacts, particularly for intra-chromosomal interactions across the centromeres (Satou et al., 2019). As we prepared our Hi-C libraries from adult animals, where polyploidy is high (Ganot and Thompson, 2002), we cannot rule out that it could be a possible cause of the low inter-arm interactions in our contact matrix. Further studies such as investigations of other developmental stages will be needed to elucidate the mechanism at work for the similarity between *O. dioica* and insect's chromosome contact maps.

2.4.3 Visualization and access

We prepared a public view of our reference genome in the ZENBU browser (Severin et al., 2014), displaying tracks for our gene models, in silico-predicted features such as repeats and non-coding RNAs, or syntenic regions with other *Oikopleura* genomes. To facilitate the study of known genes, we screened the literature for published sequences (Appendix 1) and mapped them to the genome with a translated alignment. The ZENBU track for these alignments is searchable by gene name, accession number and PubMed identifier. Chromosome-level visualization of this track shows that the genes studied so far are distributed evenly on each chromosome, except for the repeat-rich YSR (Fig. 2.8). In line with the observed loss of synteny in the Hox genes noted in *Oikopleura* (Seo et al., 2004), we did not see apparent clustering of genes by function or relatedness. The view of the OKI2018_I69 genome assembly can be found here:
https://fantom.gsc.riken.jp/zenbu/gLyphs/#config=0tPT7vwSO1Vm5QV9iKqfAC;loc=OKI2018_I69_1.0::chr1:677717..880998+ (ZENBU view "OKI2018_I69_1.0 view with tracks (updated)").

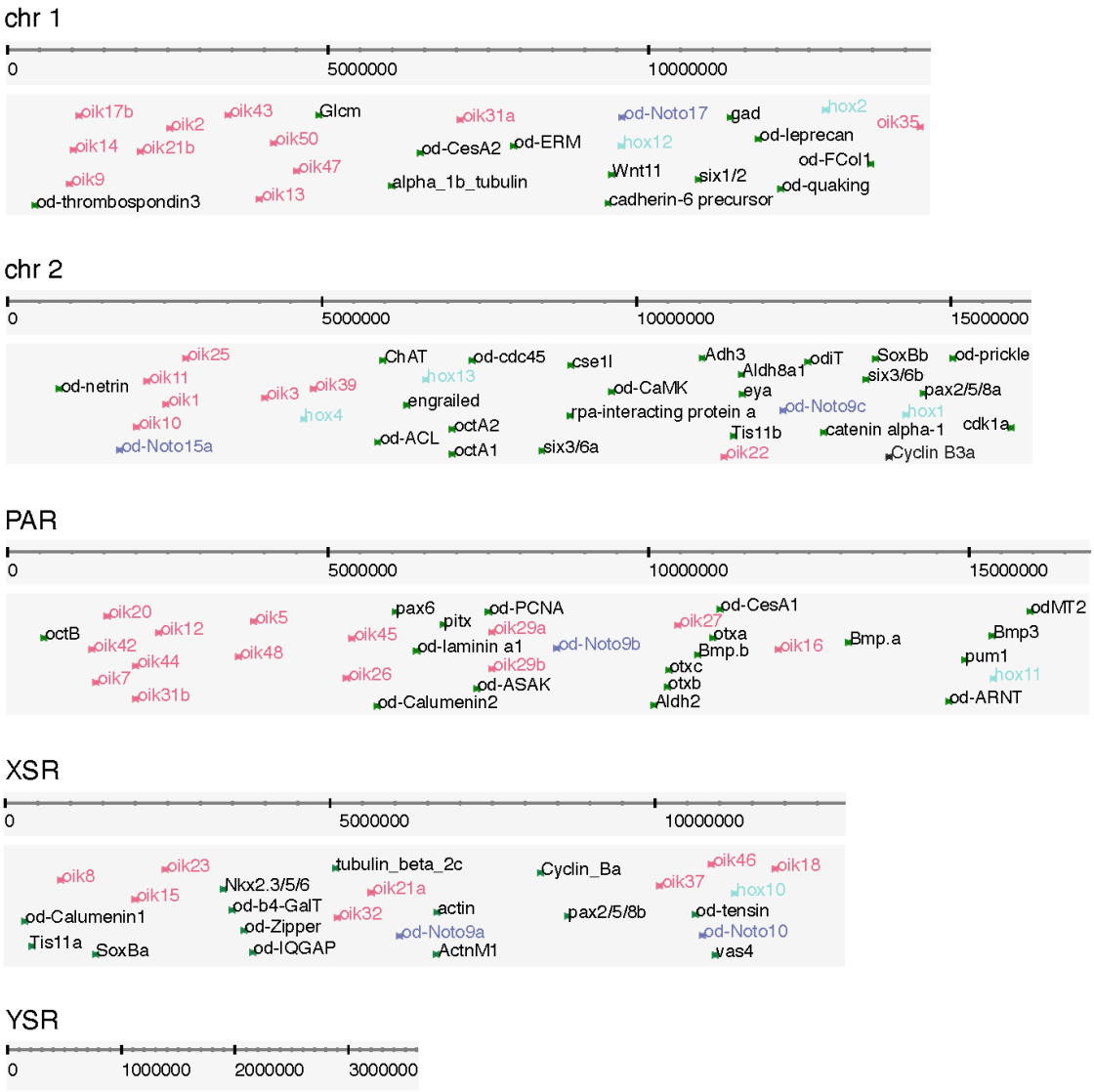


Figure 2.8: Genomic locations of various oikopleurid gene homologs in the OKI2018_I69. The genes are searchable by name and PubMed identifiers in the ZENBU genome browser. Colors indicate genes from the same family.

2.5 Conclusions

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

2.6 Availability of data and materials

All sequence data presented here, the final OKI2018_I69_1.0 genome assembly and annotation were deposited to the ENA database under BioProject ID PRJEB40135 and Zenodo (DOI https://doi.org/10.5281/zenodo.4604144). Custom scripts used in this study are available in GitHub (https://github.com/oist/oikGenomePaper)

Chapter Three

Extensive genomic rearrangements in phenotypically similar populations of *Oikopleura dioica*

The work and research in this chapter represent a collaborative project with the contribution of all members of the Genomics and Regulatory Systems Unit. In particular, Aki Masunaga, Yongkai Tan, and Andrew Liu generated the sequencing data; Charles Plessy and I assembled the genomes; Charles Plessy computed pairwise genome alignments and analyzed them; I annotated the repeats and genes in the genomes, performed the functional annotation, reconstructed gene orthology, computed and analyzed syntenic blocks; Michael Mansfield calculated dN/dS values and the molecular clock.

3.1 Introduction

It is well-known that many organisms, from sponges to humans, show a certain conservation of global genome architecture. In particular, hundreds of conserved gene blocks were found throughout different metazoan genomes (Simakov et al., 2013), including the Hox cluster that can be traced back to the origin of bilaterian animals more than 500 Mya (Balavoine et al., 2002). However, *Oikopleura dioica*, a tiny planktonic chordate, does not seem to follow the same patterns (Denoeud et al., 2010).

Sequencing of the *O. dioica* genome indicated that little synteny has been preserved with the ancestral chordate linkage groups. The genome organization is extremely compact and highly dynamic: multiple genomic features such as transposon diversity, intron repertoire, gene content and order are scattered in *Oikopleura* (Denoeud et al., 2010). The Hox cluster has been entirely dispersed and more than 30% of Hox genes are missing (Seo et al., 2004; Denoeud et al., 2010; Blanchoud et al., 2018). The genome compaction coincided with low repeat content (~15%) and gene loss, including the c-NHEJ genes (Denoeud et al., 2010; Deng et al., 2018). There are other unusual features in the genome, such as co-expression of genes within operons, *trans*-splicing, and high intron turnover (Denoeud et al., 2010; Ganot et al., 2004). These dramatic genomic features have not affected the preservation of ancestral morphology of the species: *O. dioica* possesses a chordate-like body plan and early development (Denoeud et al., 2010).

Considering such scientific importance, it is surprising to learn that so little is known about the within-species diversity of *O. dioica*. Establishing any major morphological differences between *O. dioica* from diverse geographical locations proved to be difficult (Masunaga et al., 2022). The karyotype of three chromosome pairs seems to be preserved (Denoeud et al., 2010; Liu et al., 2020; Chapter two). Currently, all dioecious *Oikopleura* around the globe are considered to represent a single species. However, sequence variation has been observed across populations at single nucleotide and amino acid levels (Denoeud et al., 2010; Wang et al., 2015; Wang et al., 2020).

Here, we examine synteny conservation and variation in *O. dioica* by comparing chromosome-scale genome assemblies from three populations from the Northern hemisphere: one from North Atlantic (Bergen/Barcelona) and two from Pacific (Okinawa/Kume and Osaka/Aomori) Oceans (Fig. 3.1). Despite their broadly conserved morphology and karyotype, these populations exhibit extreme levels of genomic rearrangements. These rearrangements appear to preserve protein-coding elements, with genes and exons being more conserved than operon structures. At the macro scale, arms within individual chromosomes seem to exhibit different evolutionary rates, with fewer synteny blocks and more breakpoints observed in the short arms. Moreover, consistent with the fast evolutionary rate in the species, these genetic events appear to have accumulated in *O. dioica* much faster than in other animals and may have resulted in multiple speciation events.

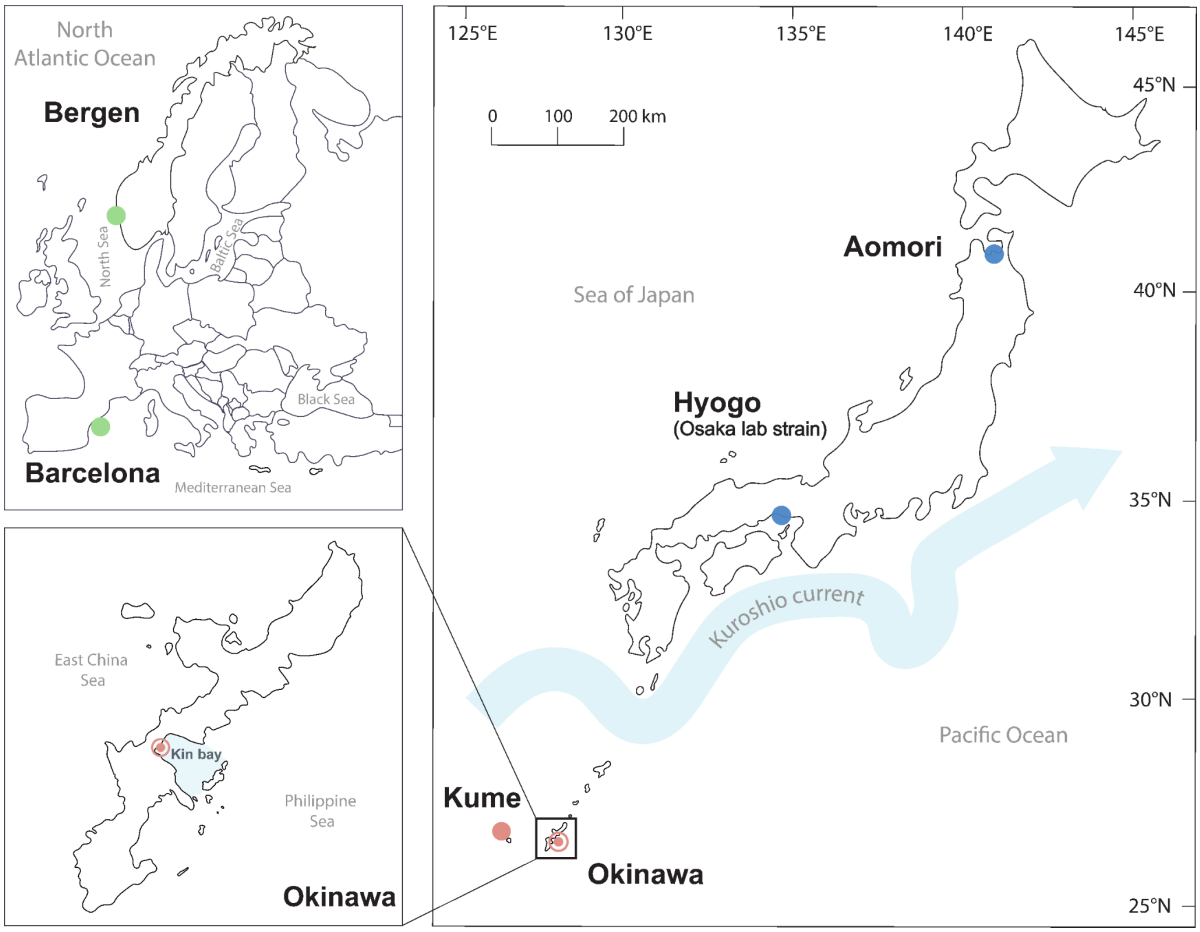


Figure 3.1: Sampling locations of dioecious *Oikopleura* (adapted from Masunaga et al., 2022): three populations from the Ryukyu archipelago (Okinawa and Kume), the North Atlantic (Bergen and Barcelona) and North Pacific (Osaka and Aomori) Oceans.

3.2 Materials and Methods

3.2.1 Genome sequencing and assembly

For the four genomes presented in this chapter, the DNA from male *O. dioica* individuals was sequenced to simplify the assembly of the sex-specific regions, which are single-copy in that case. For the Barcelona genome, high-molecular-weight DNA was extracted from the "Bar2" individual of the Barcelona laboratory strain (Martí-Solans et al., 2015) using a modified salting-out protocol (Masunaga et al., 2022) and sequenced on MinION sequencer Mk1B (Oxford Nanopore Technologies) with a SQK-LSK109 kit (ONT) following the manufacturer's instructions. Basecalling was performed with the Guppy software (ONT) v4.4.2 using the Rerio model res_dna_r941_min_crf_v031(https://github.com/nanoporetech/rerio). The shortest reads were discarded using the filtlong software (https://github.com/rrwick/Filtlong), resulting in a N50 read length higher than 30,000 nt. The reads were assembled into contigs using the Flye software v.2.8.2-b1689 with the --min-overlap 10000 parameter (Kolmogorov et al., 2019). The alternative haplotype sequences were removed from the assemblies using the purge_dups tool (Guan et al., 2020). Since a single run of the purge_dups tool was not efficient enough, an alternative approach was introduced: in one iteration, the haplotigs were first flagged with purge_dups, the Nanopore reads were then mapped to assembled contigs with LAST and last-split, and reads mapping to purged haplotigs were removed prior to restarting of the assembly process. Iterations were stopped when the purge_dups stopped discovering alternative haplotypes. The contig assembly that provided the best tradeoff between contiguity and low number of duplicated single-copy orthologs (BUSCOs) was selected for further analysis. The contigs were polished with Pilon v1.22 (Walker et al., 2014) using 150-bp paired-end Illumina reads generated from the DNA of the same individual to remove Nanopore-specific errors, and joined into scaffolds with Hi-C data from the Bergen *O. dioica* line at tailbud stage (SRR14470734) using Juicer v1.6 (Durand et al., 2016) and 3D-DNA (Dudchenko et al., 2017) pipelines. The resulting assembly was called "Bar2_p4". For the Osaka genome (OSKA2016v1.9), the original OSKA2016 assembly (Wang et al., 2020) was re-scaffolded manually by merging scaffolds overlapped by long contigs from the Nanopore-based genome assembly drafts generated for the single individuals from the same laboratory strain (SAMEA6864573). The genomes of Kume (KUM-M3-7f) and Aomori (AOM-5-5f) *O. dioica* were sequenced using single animals isolated from wild populations (Masunaga et al., 2022) and assembled with the same method as for the Barcelona genome, except for the polishing and scaffolding steps that were not performed.

To ensure the completeness of the assemblies, metazoan near-universal single-copy genes were counted using the BUSCO tool v5.2.1 (Manni et al. 2021) and an AUGUSTUS model trained for the Okinawa *O. dioica* (see Materials and Methods in chapter two). Unfortunately, this version of BUSCO appears to have a lower detection baseline compared to the v3.0.2 (Simão et al., 2015) that was used in chapter two to assess the completeness of the OKI2018_I69, OSKA2016 (Wang et al., 2015) and OdB3 (Denoeud et al., 2010) genome assemblies. For example, only 64% of complete BUSCOs were predicted for the OKI2018_I69 with the new version of the software. We already discussed the high accuracy of the OKI2018_I69 genome assembly in the previous chapter. Therefore, we can assume the

completeness of the other five assemblies presented in this chapter based on the similarity of their scores with the OKI2018_I69.

3.2.2 Annotation of the genomes

The repeat and gene annotations were performed using a procedure similar to that for OKI2018_I69 (see Materials and Methods in chapter two). For each genome, a custom library of repetitive elements was merged from the outputs of three different software: RepeatModeler v2.0.1 (Flynn et al., 2020), MITE-Hunter v11–2011 (Han and Wessler, 2010), and SINE_Finder (Wenke et al., 2011), and used as input for a RepeatMasker search against the genome (v4.1.0; Smit, Hubley and Green at http://repeatmasker.org). The identified repeats were soft-masked to keep the genome sequence information.

Gene models were predicted using AUGUSTUS v3.3.3 (Stanke et al., 2006) based on the species model trained for the Okinawa population (see Materials and Methods in chapter two). In order to produce more accurate annotations, transcripts aligned to genomes with BLAT v36 were used as "hints". For specimens where an assembled transcriptome was not available, data from related individuals was used. For instance, the transcriptome assembly generated by Wang et al. (2015) for the Osaka laboratory strain was used for predicting genes in both OSKA2016v1.9 and AOM-5-5f, whereas the Okinawan transcriptome from chapter two (see Materials and Methods) was used for reannotation of OKI2018_I69 and annotation of KUM-M3-7f. For the Bar2_p4 genome, we used an assembled transcriptome shared by Professor Cristian Cañestro (University of Barcelona). The parameter "--allow_hinted_splicesites" was used with AUGUSTUS to allow the prediction of non-canonical splice sites. Operons were identified as a set of colinear genes in the same orientation separated by 500 bp at most using the "bedtools merge" function ("-s" parameter to force strandedness). The threshold of 500 bp was chosen to recover the operon from Ganot et al. (2004). Further, the distribution of operon sizes in all three genomes matches the ones shown in Supplementary Figure S8 in Denoeud et al. (2010; Figure 3.11c).

All *O. dioica* translated gene sequences were subjected to InterProScan v5.22-86.0 (Jones et al., 2014) for functional annotation. InterProScan was run with parameters of "-appl Pfam -iprlookup -goterms -pa -f tsv" to annotate Pfam protein families and Gene Ontology (GO) terms for each translated protein sequence. The GO terms show which biological processes (BPs), molecular functions (MFs) and cellular components (CCc) the gene is involved in. In total, around 60% of transcripts in all genomes had a functional annotation with Pfam IDs, and around 40% with GO terms. The Bioconductor GOstats package in R was used to identify GO terms enriched in a given set of genes (Falcon and Gentleman, 2007).

3.2.3 Pairwise genome alignment and comparison

To align genomes to each other, we developed a reproducible and standardized workflow using Nextflow (pipeline system; Di Tommaso et al., 2017) and a local alignment method called LAST. The LAST method is especially good at finding structural rearrangements and recombinations, and, thus, is well-suited for whole-genome alignments (Kiełbasa et al., 2011; Frith and Kawaguchi, 2015; Mitsuhashi et al., 2020). The pipeline is available from https://github.com/oist/plessy_pairwiseGenomeComparison/tree/v5.1.0. In the pipeline, a *target* genome is indexed with the YASS seeding scheme (Noé and Kucherov, 2005) to allow searching for "long-and-weak similarities". A *query* genome is aligned to the *target* genome with parameters and a scoring matrix determined by the LAST-TRAIN software (Hamada et al., 2017). The resulting sets of many-to-many alignments were filtered with the last-split tool (Frith

and Kawaguchi, 2015) in order to find the optimal set of one-to-one alignments. In addition, alignments comprising soft-masked repeat sequences were removed from the dataset with the "last-postmask" tool. To load the alignment coordinates in the R environment for further statistical analysis, we developed an R package called "GenomicBreaks" (https://oist.github.io/GenomicBreaks/) using core Bioconductor libraries (Lawrence et al., 2013). For each pair of genomes, the "strand_randomisation_index()" function from the GenomicBreaks package was used to calculate the strand randomization index which indicates that either all alignments are on the same strand (a value of 1) or overall orientation is random (a value of zero). The breakpoints and bridge regions were computed based on the strictest definition of colinearity, where it is interrupted by inversions or translocations of any length. The minimum length of the bridge region is ~200 bp. Smaller bridge regions may be missed as they are represented as a gap within an alignment region by the aligner.

Pairwise comparison between the *O. dioica* genomes (this work) and the *O. vanhoeffeni*, *O. longicauda* and *O. albicans* genomes (Naville et al., 2019) were loaded in the CNEr package (Tan et al., 2019) to define conserved non-coding elements with a window size of 50 and an identity threshold of 48.

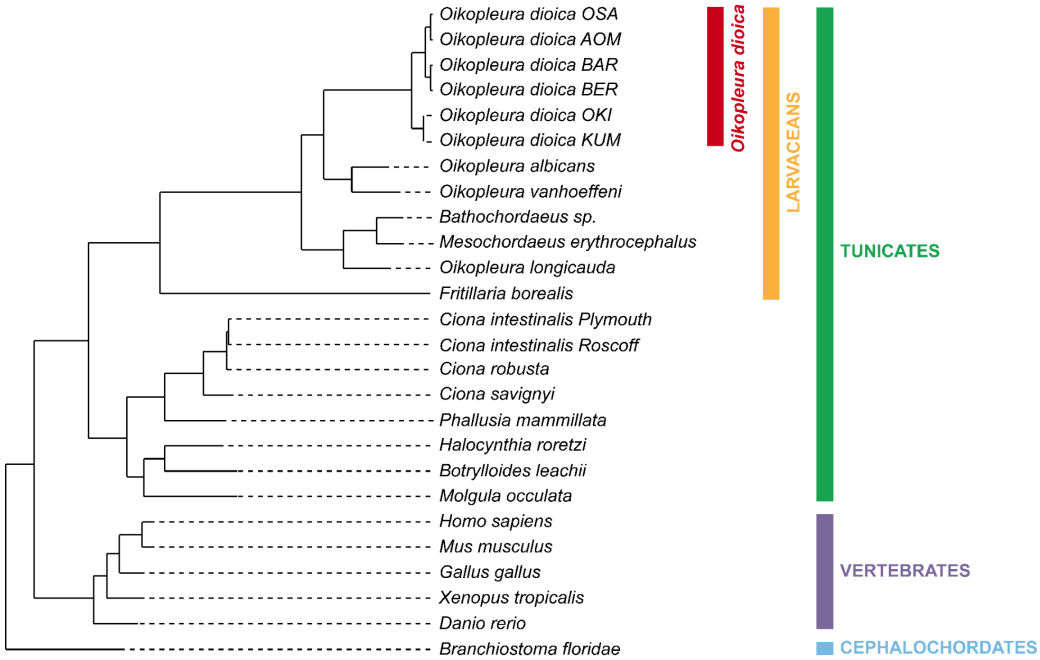
3.2.4 Identification of orthologous genes and gene synteny analysis

To ensure a good orthology assignment between *O. dioica* genomes, six oikopleurid species from Naville et al. (2019) were used as recommended in the OrthoFinder tutorials (https://davidemms.github.io/). Since the oikopleurid genomes lack publicly available annotations, we masked them using repeat libraries generated by Naville et al. (2019) and annotated *de novo* with AUGUSTUS v3.3.3 based on either Okinawan *O. dioica* (Chapter two) or *Ciona* models. Unfortunately, predicted gene models did not yield good BUSCO scores owing to the high fragmentation of the genomes and the absence of transcriptomic data. Also, the AUGUSTUS models optimized for *O. dioica* and *Ciona* may not be appropriate for gene predictions in other larvaceans, given the existence of operons and non-canonical splice-sites in this clade that complicates the generation of accurate gene models. Nevertheless, that was not a problem for this analysis since assigning orthologous genes within *O. dioica* is the main purpose of this step. In addition, two *Ciona* genomes, *C. intestinalis* P (Plymouth) and *C. intestinalis* R (Roscoff) (Satou et al., 2021), were annotated using the same pipeline based on the *Ciona* AUGUSTUS model ("--species=ciona"). Results of the gene predictions can be found in Table 3.1.

Table 3.1: Annotation results of the six oikopleurid genomes (from Naville et al., 2019) and two *Ciona intestinalis* (from Satou et al., 2021).

Species	Total assembly size (Mbp)	Repeat content (%)	AUGUSTUS model	# of predicted genes	BUSCO score
Oikopleura albicans	365	26.58	*Ciona*	24,943	C:33.7%, F:18.3%
Oikopleura vanhoeffeni	643.6	40.45	*Ciona*	18,686	C:25.4%, F:18.7%
Oikopleura longicauda	308.7	20.75	*Oikopleura*	27,566	C:54.9%, F:12.2%
Mesochordaeus erythrocephalus	874	51.37	*Ciona*	48,793	C:11.7%, F:20.2%
Bathochordaeus sp.	396.5	38.65	*Ciona*	25,023	C:12.7%, F:17.2%
Fritillaria borealis	143.1	11.25	*Oikopleura*	17,861	C:28.6%, F:11.5%
Ciona intestinalis P (Plymouth)	175.3	39.11	*Ciona*	16,492	C:90.1%, F:3.2%
Ciona intestinalis R (Roscoff)	275.8	43.5	*Ciona*	23,131	C:91.7%, F:3.1%

Several other species spanning three Chordata subphyla were added to the dataset: *Branchiostoma floridae* (Cephalochordata), two *Ciona* species (*C. intestinalis* "type A" also known as "*robusta*" and *C. savignyi*; Tunicata), four other tunicate species (*Botrylloides leachii*, *Halocynthia roretzi*, *Molgula oculata*, *Phallusia mammillata*) and five vertebrates (*Danio rerio*, *Xenopus tropicalis*, *Gallus gallus*, *Mus musculus*, *Homo sapiens*). See Table 3.2 for a full list of species used in this analysis and the source information. To remove redundancy in the dataset, protein sequences were clustered at 100% identity using cd-hit version 4.8.1 (Li and Godzik, 2006; Table 3.2). Also, genes from alternative haplotypes were removed from the Bergen *O. dioica* and only the longest isoforms per genes in other *O. dioica* were used for the analysis. Finally, OrthoFinder v2.5.4 (Emms and Kelly, 2015; Emms and Kelly, 2019) was run on protein sets from 26 organisms with the parameters "-M msa -T raxml-ng" for higher sensitivity. A fixed species tree was used to avoid long branch artifacts and to ensure that *O. dioica* sequences fall within the larvacean clade. This tree was generated with a pre-run of OrthoFinder and edited manually based on the phylogeny presented in Delsuc et al. (2018) and Naville et al. (2019):



Later, the accuracy of this tree was confirmed by an independent phylogenomic analysis performed by Michael Mansfield (see Discussion).

Table 3.2: Per species statistics of orthologous genes assignment performed with OrthoFinder.

Species	Phylogenetic group	Source	Number of genes	Number of genes after clustering	Number of genes in orthogroups	Number of unassigned genes	Percentage of genes in orthogroups	Percentage of unassigned genes	Number of orthogroups containing species	Percentage of orthogroups containing species	Number of species-specific orthogroups	Number of genes in species-specific orthogroups	Percentage of genes in species-specific orthogroups
Branchistoma floridae (Amphioxus)	Cephalochordata	Uniprot (UP000001554)	28542	28438	26567	1871	93.4	6.6	8377	24.2	863	4828	17
Okinawa (OKI2018_I69)	Tunicata	This thesis	17291	17109	16854	255	98.5	1.5	10378	29.9	44	132	0.8
Kume (KUM-M3-7f)	Tunicata	This thesis	16852	16711	16514	197	98.8	1.2	10220	29.5	31	104	0.6
Osaka (OSAK2016v1.9)	Tunicata	This thesis	15720	15662	15480	182	98.8	1.2	9698	28	27	71	0.5
Aomori (AOM-5-5f)	Tunicata	This thesis	15224	15160	15047	113	99.3	0.7	9595	27.7	14	39	0.3
Barcelona (Bar2_p4)	Tunicata	This thesis	14272	14169	14020	149	98.9	1.1	8980	25.9	21	49	0.3
Bergen (OdB3)	Tunicata	Denoeud et al., 2010	18020	16899	16105	794	95.3	4.7	9122	26.3	54	178	1.1
Oikopleura albicans	Tunicata	Naville et al., 2019 and this thesis	24943	23830	20976	2854	88	12	7814	22.5	1155	5547	23.3
Oikopleura vanhoeffeni	Tunicata	Naville et al., 2019 and this thesis	18686	17973	15451	2522	86	14	6270	18.1	462	3799	21.1

Table 3.2: Per species statistics of orthologous genes assignment performed with OrthoFinder (continued).

Species	Phylogenetic group	Source	Number of genes	Number of genes after clustering	Number of genes in orthogroups	Number of unassigned genes	Percentage of genes in orthogroups	Percentage of unassigned genes	Number of orthogroups containing species	Percentage of orthogroups containing species	Number of species-specific orthogroups	Number of genes in species-specific orthogroups	Percentage of genes in species-specific orthogroups
Oikopleura longicauda	Tunicata	Naville et al., 2019 and this thesis	27566	27264	24820	2444	91	9	8314	24	1094	5840	21.4
Mesochordaeus erythrocephalus	Tunicata	Naville et al., 2019 and this thesis	48793	48717	41303	7414	84.8	15.2	9028	26	1300	11764	24.1
Bathochordaeus sp.	Tunicata	Naville et al., 2019 and this thesis	25023	24979	21146	3833	84.7	15.3	7511	21.7	627	7308	29.3
Fritillaria borealis	Tunicata	Naville et al., 2019 and this thesis	17861	17770	13262	4508	74.6	25.4	4714	13.6	1221	5814	32.7
Ciona intestinalis (*robusta*)	Tunicata	Uniprot (UP000008144)	16678	16644	14541	2103	87.4	12.6	7802	22.5	58	200	1.2
Ciona intestinalis P (Plymouth)	Tunicata	Satou et al., 2021 and this thesis	16492	16284	16044	240	98.5	1.5	8220	23.7	65	252	1.5
Ciona intestinalis R (Roscoff)	Tunicata	Satou et al., 2021 and this thesis	23131	22412	21951	461	97.9	2.1	8408	24.2	107	423	1.9
Ciona savignyi	Tunicata	Uniprot (UP000007875)	11592	11570	10895	675	94.2	5.8	6445	18.6	64	170	1.5

Table 3.2: Per species statistics of orthologous genes assignment performed with OrthoFinder (continued).

Species	Phylogenetic group	Source	Number of genes	Number of genes after clustering	Number of genes in orthogroups	Number of unassigned genes	Percentage of genes in orthogroups	Percentage of unassigned genes	Number of orthogroups containing species	Percentage of orthogroups containing species	Number of species-specific orthogroups	Number of genes in species-specific orthogroups	Percentage of genes in species-specific orthogroups
Botrylloides leachii	Tunicata	Aniseed	15839	15782	14796	986	93.8	6.2	7147	20.6	298	1250	7.9
Halocynthia roretzi	Tunicata	Aniseed	16404	13909	13017	892	93.6	6.4	7359	21.2	85	258	1.9
Molgula oculata	Tunicata	Aniseed	16616	15301	14076	1225	92	8	6886	19.9	251	1009	6.6
Phallusia mammillata	Tunicata	Aniseed	23828	19370	18404	966	95	5	7787	22.5	353	1707	8.8
Danio rerio (Zebrafish)	Vertebrata	Uniprot (UP000000437)	25706	25616	24829	787	96.9	3.1	8773	25.3	244	1323	5.2
Xenopus tropicalis (Frog)	Vertebrata	Uniprot (UP000008143)	22514	22369	22052	317	98.6	1.4	8960	25.8	125	667	3
Gallus gallus (Chicken)	Vertebrata	Uniprot (UP000000539)	18113	17980	17708	272	98.5	1.5	8279	23.9	45	508	2.8
Mus musculus (Mouse)	Vertebrata	Uniprot (UP000000589)	22001	21944	21513	431	98	2	9829	28.3	176	1434	6.5
Homo sapiens (Human)	Vertebrata	Uniprot (UP000005640)	20600	20501	19874	627	96.9	3.1	9772	28.2	82	418	2

The identified orthologs with one-to-one relationships were loaded in the R environment and visualized with Oxford (macro-synteny) plots with the "makeOxfordPlots()" function from the GenomicBreaks package. For each pair of genomes, gene synteny blocks were predicted based on the same orthologous set using the "coalesce_contigs()" function, defining a synteny block as a set of colinear genes that appear in the same order independent of orientation.

3.2.5 Identification of ancestral gene clusters

To study the preservation of ancestral gene clusters, protein sequences of the Bergen *O. dioica* were used as queries to search against protein and genome sequences of other *O. dioica* with BLAST v2.10.1. Genes were considered orthologous if their protein sequences shared more than 75-80% identity over 80% of their lengths, and belonged to the same orthogroup computed by OrthoFinder. The ortholog IDs and their genomic locations in each genome are shown in Table 3.3. To plot orthologs on the same figure, gene coordinates were transposed to the OKI2018_I69 genome: initial coordinates were first divided to the corresponding chromosome length and multiplied by the length of the same chromosome in the OKI2018_I69.

3.2.6 dN/dS estimation

The dN/dS values were estimated for single-copy orthologs common to all six genomes of *O. dioica*. Each orthologous protein was aligned using PRANK and trimmed for unreliable sites with the GUIDANCE2 algorithm (v2.02; Sela et al., 2015). The resulting protein alignments were converted to codon alignments using PAL2NAL (v14.1; Suyama et al., 2006). Phylogenetic trees were estimated for each orthologue using RAxML with the PROTCATAUTO model and 100 rapid bootstraps. The dN/dS values were estimated using the CODEML program of the PAML package (version 4.9j; Yang, 1997; Yang, 2007).

Table 3.3a: Genome locations and ids of the Hox cluster gene orthologous in the three different populations of *O. dioica*.

Gene name	Orthogroup ID	OdB3 orthologs	OKI2018_I69		OSKA2016v1.9		Bar2_p4	
			ID	Genome location	ID	Genome location	ID	Genome location
Hox1	HOG0000193	GSOIDT00017529001	g8043	chr2:14331701-14339427	g6520	Chr2:11876406-11881984	g4409	Chr2:6198535-6201003
Hox2	HOG0000191	GSOIDT00016901001	g1772	chr1:6753050-6760732	g1327	Chr1:5476819-5481313	g2590	Chr1:9825924-9831120
Hox4	HOG0000192	GSOIDT00012820001	g5193	chr2:4686018-4687168	g3346	Chr2:1156287-1157412	g3778, g3781	Chr2:3941566-3942121, Chr2:3946172-3947335
Hox9	HOG0000198	GSOIDT00013300001, AAS21428.1	g7783	chr2:13443950-13444626	g5216	Chr2:7763982-7777636	g4658	Chr2:7123494-7124409
Hox10	HOG0000200	GSOIDT00003106001	g16661	XSR:11241991-11244701	g12075	XSR:2068460-2071007	g12969	XSR:10969494-10972064
Hox11	HOG0000195	GSOIDT00007256001	g12896	PAR :15661221-15664812	g11069	PAR:13798474-13807133	not annotated	PAR:5221095-5223352
Hox12	HOG0000199	GSOIDT00011323001	g2652	chr1:9941100-9945705	g2511	Chr1:9788291-9791413	g1071	Chr1:4266047-4269168
Hox13	HOG0000196	GSOIDT00000159001	g5822	chr2:6661083-6665189	g4769	Chr2:6220372-6223917	g5228	Chr2:9288283-9291948

Table 3.3b: Genome locations and ids of the pharyngeal cluster gene orthologous in the three different populations of *O. dioica*.

Gene name	Orthogroup ID	OdB3 orthologs	OKI2018_I69		OSKA2016v1.9		Bar2_p4	
			ID	Genome location	ID	Genome location	ID	Genome location
FoxA1	HOG0000722	GSOIDT00003756001	g2744	chr1:10296365-10297679	g2788	Chr1:10740371-10741672	g1404	Chr1:5369941-5371263
FoxA2	HOG0000724	GSOIDT00005965001	g1126	chr1:4222672-4229527	g454	Chr1:2009845-2014213	g151	Chr1:773797-779314
FoxA3	HOG0000723	GSOIDT00004124001	g4703	chr2:3019820-3024869	g3878	Chr2:3017153-3020706	g2871	Chr2:813796-817047
Nkx2.1	HOG0013656	GSOIDT00010368001	g13695	XSR:1528628-1534042	g14990	XSR:11774258-11778796	g10634	XSR:2692305-2696783
Nkx2.2	HOG0003442	GSOIDT00011992001	g16546	XSR:10803779-10806625	g13003	XSR:5104824-5107578	g12090	XSR:7750435-7753186
Pax3/7	HOG0002231	GSOIDT00008979001	g1551	chr1:5849112-5850200	g2263	Chr1:8869732-8870820	g1780	Chr1:6764631-6765722
slc25A21	HOG0009967	GSOIDT00009358001	g1803	chr1:6863227-6865895	g1131	Chr1:4692595-4694226	g2073	Chr1:7829690-7831386

Table 3.3c: Genome locations and ids of the NK cluster gene orthologous in the three different populations of *O. dioica*.

Gene name	Orthogroup ID	OdB3 orthologs	OKI2018_I69		OSKA2016v1.9		Bar2_p4	
			ID	Genome location	ID	Genome location	ID	Genome location
Lbx	HOG0000157	GSOIDT00013323001	g5882	chr2:6848505-6849242	g5203	Chr2:7719517-7720288	g4674	Chr2:7170058-7170833
Msx	HOG0000165	GSOIDT00001108001	g12369	PAR:13708956-13724166	g8875	PAR:7310423-7311666	g8147	PAR:7368876-7370126
NKx4	HOG0003444	GSOIDT00003812001	g14210	XSR:3368067-3377209	g14032	XSR:8365594-8374800	g10540	XSR:2340027-2350208
NKx5	HOG0000183	GSOIDT00008102001	g3572	chr1:13040081-13043521	g1462	Chr1:5968003-5974206	g2254	Chr1:8513178-8515459

3.3 Results

3.3.1 A pan-genomic comparison of *Oikopleura dioica*

To study the extent of chromosomal rearrangements in *O. dioica*, we prepared a set of six genomes of dioecious *Oikopleura* from globally distributed locations: the Ryukyu archipelago (Okinawa and Kume), North Pacific (Osaka and Aomori) and North Atlantic (Barcelona and Bergen) Oceans (Fig. 3.1; Table 3.4). To the genome assembly of the Okinawa *O. dioica* (OKI2018_I69) presented in chapter two, we added two more chromosomal genomes, Bar2_p4 and OSKA2016v1.9, of individuals from the Barcelona and Osaka *O. dioica* populations, correspondingly (Table 3.4). The Osaka assembly was constructed by scaffolding the previously published OSKA2016 genome (Wang et al., 2020) with the long Nanopore reads of single individuals from the same laboratory strain. The Barcelona genome was assembled following a workflow similar to the Okinawa one, including chromosome conformation data (Hi-C libraries) to aid scaffolding. All populations possess the same karyotype of three haploid chromosomes: two autosomes (chr1 and chr2) and one sex chromosome, containing a sex-specific X or Y region and a pseudo-autosomal region (PAR) shared by males and females (Fig. 3.2). The Hi-C contact map of the Barcelona genome showed that the chromosome arms and sex-specific regions had little interactions with each other (Fig. 3.3). Moreover, the assembly graph connected the sex-specific regions to the PAR's long arm through ribosomal DNA repeats, as observed previously for the Okinawa genome (see chapter two).

To support the Osaka and Okinawa chromosomal assemblies, we prepared two contig-level haplotype-purged genome sequences of *O. dioica* individuals from Aomori prefecture (northern Japan) and Kume island (west of mainland Okinawa; Fig. 3.2). As a validation genome to the Barcelona assembly, we used the original Bergen genome sequence (OdB3) from Denoeud et al. (2010). Therefore, each chromosomal assembly in our dataset is supported by an additional contig-level genome sequence of an *O. dioica* individual from the same population, which we refer to as "sister genomes" (Table 3.4). Using the workflow presented in chapter two, we updated gene and repeat annotation for all genomes, except for the Bergen one to keep the original functional annotation from Denoeud et al. (2010).

We summarized the main characteristics of the final genome assemblies in Table 3.4. The genomes show a large variability in length across populations, although all fall mostly around expected genome size (72 ± 13 Mbp; Seo et al., 2001). There are some visible differences in the GC and repeat content, as well as in the number of predicted protein-coding genes. However, these differences might be population-specific, given that genomes from the same population exhibit comparable characteristics. The overall completeness of all genomes is around 60-65%, judging based on the presence of universal single-copy orthologous (BUSCOs). The Okinawa, Osaka and Barcelona genomes have five main scaffolds that make up more than 99% of the total assembly lengths (Fig. 3.2), and correspond to two autosomes, pseudo-autosomal and two sex-specific regions. Given this, we conclude that by taking similar steps in assembly and annotation, we were able to produce high-quality genomes suitable for direct comparison to each other.

Figure 3.2: Treemap comparisons between Osaka, Barcelona, Aomori and Kume *O. dioica* genomes presented in the chapter. Each rectangle represents a contig or a scaffold in the assembly with the area proportional to its length.

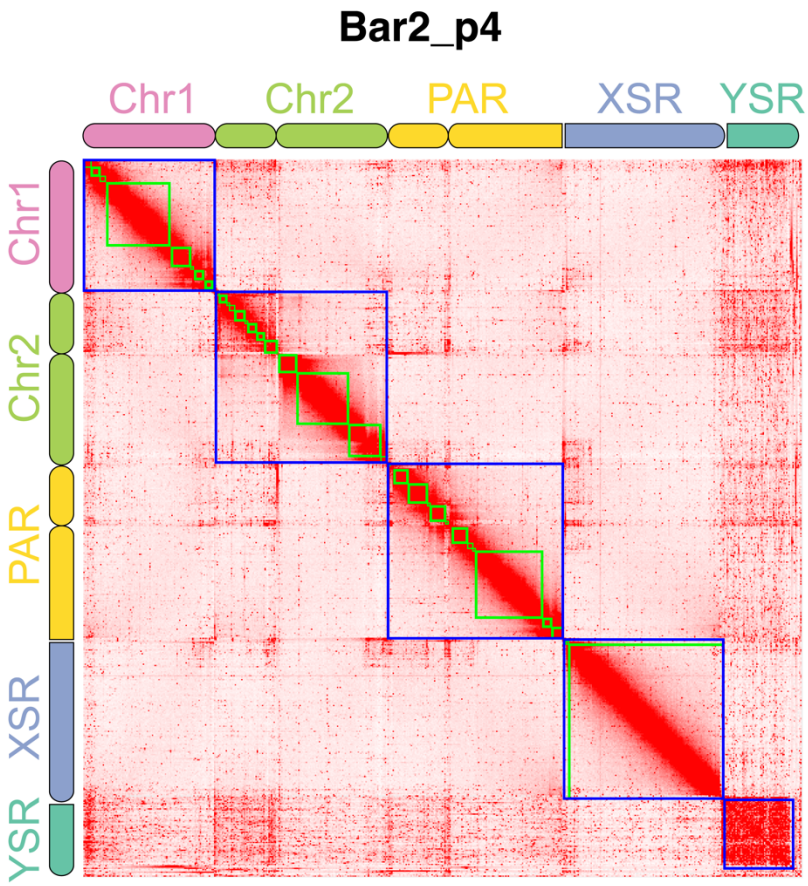


Figure 3.3: Contact matrix generated by aligning the Hi-C data set to the Bar2_p4 assembly with Juicer and 3D-DNA pipelines. Pixel intensity in the contact matrices indicates how often a pair of loci collocate in the nucleus.

Table 3.4: Statistics for the *Oikopleura dioica* genome assemblies.

	Okinawa	Kume	Osaka	Aomori	Barcelona	Bergen
Genome id	OKI2018_I69	KUM-M3-7f	OSKA2016v1.9	AOM-5-5f	Bar2_p4	OdB3
Location	Okinawa island, Ryukyu archipelago	Kume island, Ryukyu archipelago	Hyogo, Japan inland sea	Honshu, Japan N.E. Pacific coast	Mediterranean Sea, Spain	North Atlantic coast, Norway
Group	Ryukyu	Ryukyu	North Pacific	North Pacific	North Atlantic	North Atlantic
Length (Mbp)	64.3	64.7	56.6	56.8	55.8	70.4
N50 (Mbp)	16.2	4.7	13.4	6.4	12.5	0.4
GC richness	41%	41%	41.4%	41.5%	40%	40%
Repeats (%)	14.4%	13.7%	13.2%	14.1%	15%	13.5%
Genes	17,291	16,852	15,720	15,224	14,272	18,020
Transcripts	18,906	18,321	17,199	16,606	15,741	18,020
Completeness	64%	66%	63%	65%	64%	60%
Technology	Nanopore+ Illumina+HiC	Nanopore	PacBio+ Illumina	Nanopore	Nanopore+ Illumina+HiC	Sanger
Reference	Chapter two	This chapter	Wang et al., 2020 and this chapter	This chapter	This chapter	Denoeud et al., 2010

3.3.2 Pairwise alignment of *Oikopleura dioica* genomes

Given the chromosome-scale assemblies (Okinawa, Osaka, Barcelona), we compared chromosomal structures in *O. diocia* by aligning the genomes in a pairwise matter. We created reproducible Nextflow pipeline (Di Tommaso et al., 2017) and an alignment method called LAST (Frith and Kawaguchi, 2015; Mitsuhashi et al., 2020) to identify regions with one-to-one correspondence (mapping) between pairs of genomes. Using this method we computed all-by-all alignments, revealing an unprecedented genome rearrangement in *O. dioica* (Fig. 3.4).

Strikingly, the rearrangements are primarily restricted to homologous chromosomes: interchromosomal rearrangements are rare and represented by only 6% of the alignments (compared to 94% of intrachromosomal ones). Within chromosomes, rearrangements tend to occur within arms or sex-specific regions (~99%). Thus, chromosome arms and sex-specific regions might represent the primary unit of synteny in *O. dioica*. However, within the arms the genome is scattered to the extent that the position of a genomic segment in one population has little predictive power for the same segment's position and/or orientation in another population – a phenomenon that we refer to as "scrambling".

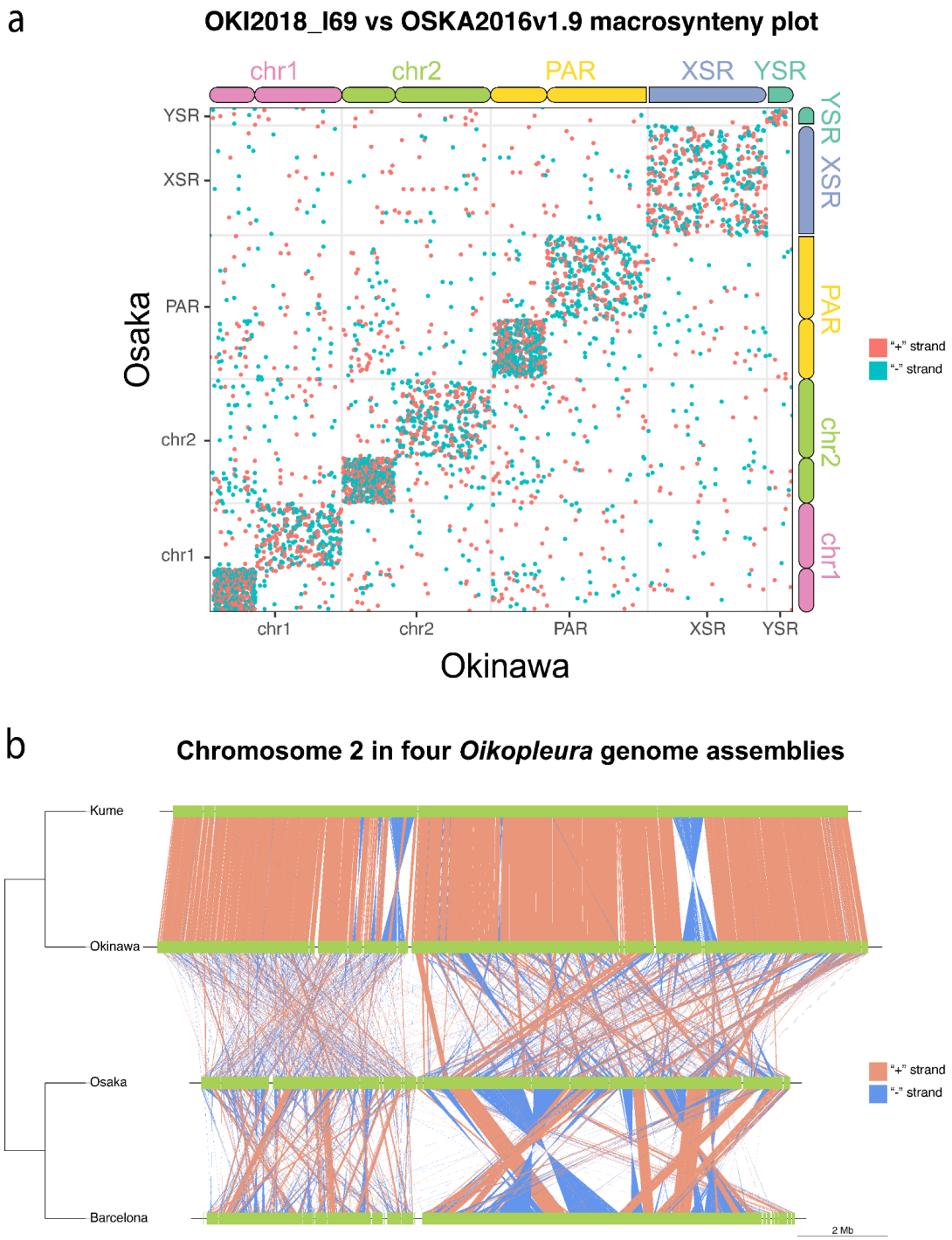


Figure 3.4: Extensive genomic rearrangements observed between *O. dioica* populations. (a) A dot plot representation of a pairwise whole-genome alignment between Osaka and Okinawa *O. dioica* populations. Blocks of strong synteny between genomic regions always appear on a dot plot as a diagonal line. Here, we see significant genomic rearrangements across the length of all chromosomes. (b) Pairwise comparison of chromosome 2 between *Oikopleura* populations.

Visual comparison of genome pairs revealed different levels of scrambling between populations (Fig. 3.4b; Fig. 3.5). Charles Plessy created a strand randomization index to measure the loss of collinearity between aligned regions in a pair of genomes (see Section 3.2.3 in Materials and Methods). Computation of the index over all pairs revealed three classes of scores: within one population (highest scores; e.g., Okinawa-Kume), between the North Atlantic and North Pacific (e.g., Osaka-Barcelona), and between the Ryukyu population and the rest (lowest scores; e.g., Okinawa-Osaka; Table 3.5). Later, the result was confirmed by an independent phylogenetic analysis performed by Masunaga et al. (2022) and an estimation of molecular divergence time performed by Michael Mansfield (see Discussion). All pairs of sister genomes show high within-population scores (0.5-0.7), ruling out the presence of potential technological bias that might have occurred due to the use of different sequencing platforms. Thus, the contig assembly set (Kume, Aomori, Bergen) provides a validation of the results with independent pairs of individuals representing the same comparison of populations (for instance, Osaka-Bergen and Aomori-Barcelona). In total, the *O. dioica* genome appears to be scrambled over separate populations. Among them, the Ryukyu population possesses the most different genome, representing an outgroup to the other two populations.

Table 3.5: Strand randomization index for pairs of *O. dioica* genomes: same-population (green), North Atlantic – North Pacific (yellow), and Ryukyu–non-Ryukyu (red) populations.

	Okinawa	**Kume**	**Osaka**	**Aomori**	**Barcelona**	**Bergen**
Okinawa	1	0.675	0.045	0.08	0.1	0.18
Osaka	0.045	0.06	1	0.545	0.23	0.285
Barcelona	0.1	0.115	0.23	0.22	1	0.5

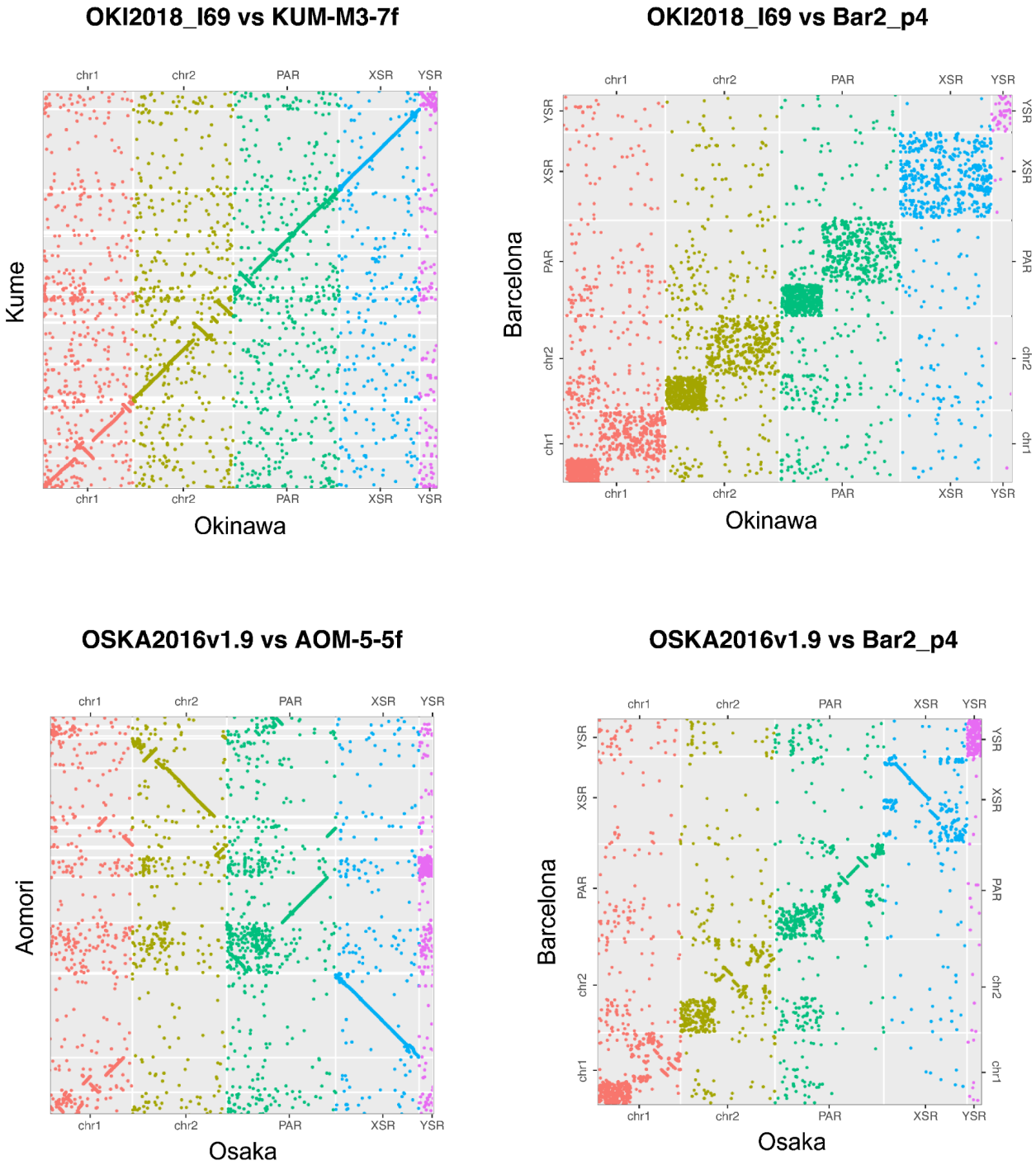


Figure 3.5: Dot plot representations of pairwise whole-genome alignments between *O. dioica* genomes.

3.3.3 Characterization of genomic breaks

Given that the most scrambling is observed between Ryukyu and the rest, we used the Okinawa-Osaka genome pair to understand how the phenomenon of scrambling relates to the various genomic features. In order to do that, we segmented the genome into five categories: colinear alignments, bridge regions, colinear regions, isolated alignments and breakpoint regions (Fig. 3.6a). We first defined colinear alignments as one-to-one mapped segments in the same orientation adjacent to each other in a pair of genomes. After that, we mapped regions flanked by colinear alignments to their counterparts and named them "bridge regions". The successions of colinear alignments and bridge regions make larger syntenic blocks – "colinear regions". By definition, the aligned genomic regions that are not colinear to anything else represent "isolated alignments". Therefore, we called the remaining unmapped regions "breakpoint regions" since they consistently interrupt colinearity.

As a next step, we calculated the frequencies of multiple genomic features at the edges of the segments (Fig. 3.6c). Isolated alignments are short in length (0.48 ± 1.9 kbp) and represent less than 5% of the genome. Their boundaries coincide with the exon start positions and with the intron ends to a lesser degree. At the same time, colinear regions (alignments + bridge regions) are large (1.7 ± 20 kbp) and cover ~70% of the genome. These regions are strongly enriched for gene and exon structures, with a local depletion of introns and repeats at their edges. Also, they overlap with the operon structures more frequently than the isolated alignments. Of non-coding features, the repeat sequences are depleted, but not completely absent, in both isolated alignments and colinear regions, while the conserved non-coding elements (CNEs) show the peak of frequency distant from the start of aligned segments. The short bridge regions (0.32 ± 5.1 kbp) mainly intersect with genes, showing strong enrichment for exon/intron boundaries at their edges. Also, there is an increase in repeat frequency further from the edges of the bridge regions, suggesting the presence of intronic repetitive elements. Strikingly, the length distribution of bridge regions has two main peaks, which may reflect bimodal intron lengths (Denoeud et al., 2010). Lastly, the breakpoint regions vary in size but are generally short (0.32 ± 5.1 kbp) and the least likely to be within a gene. These regions also cover a considerable fraction of the genome (~23.5%), suggesting that some of the scrambling events must have happened at enough time that the diverged sequences lost their ability to align or that the mechanism involves the loss of DNA fragments. Altogether, these results indicate that the most changes at the segment boundaries are related to the frequencies of protein-coding features, with the exception of operons that showed little changes in frequencies at the edges of isolated alignments and bridge regions.

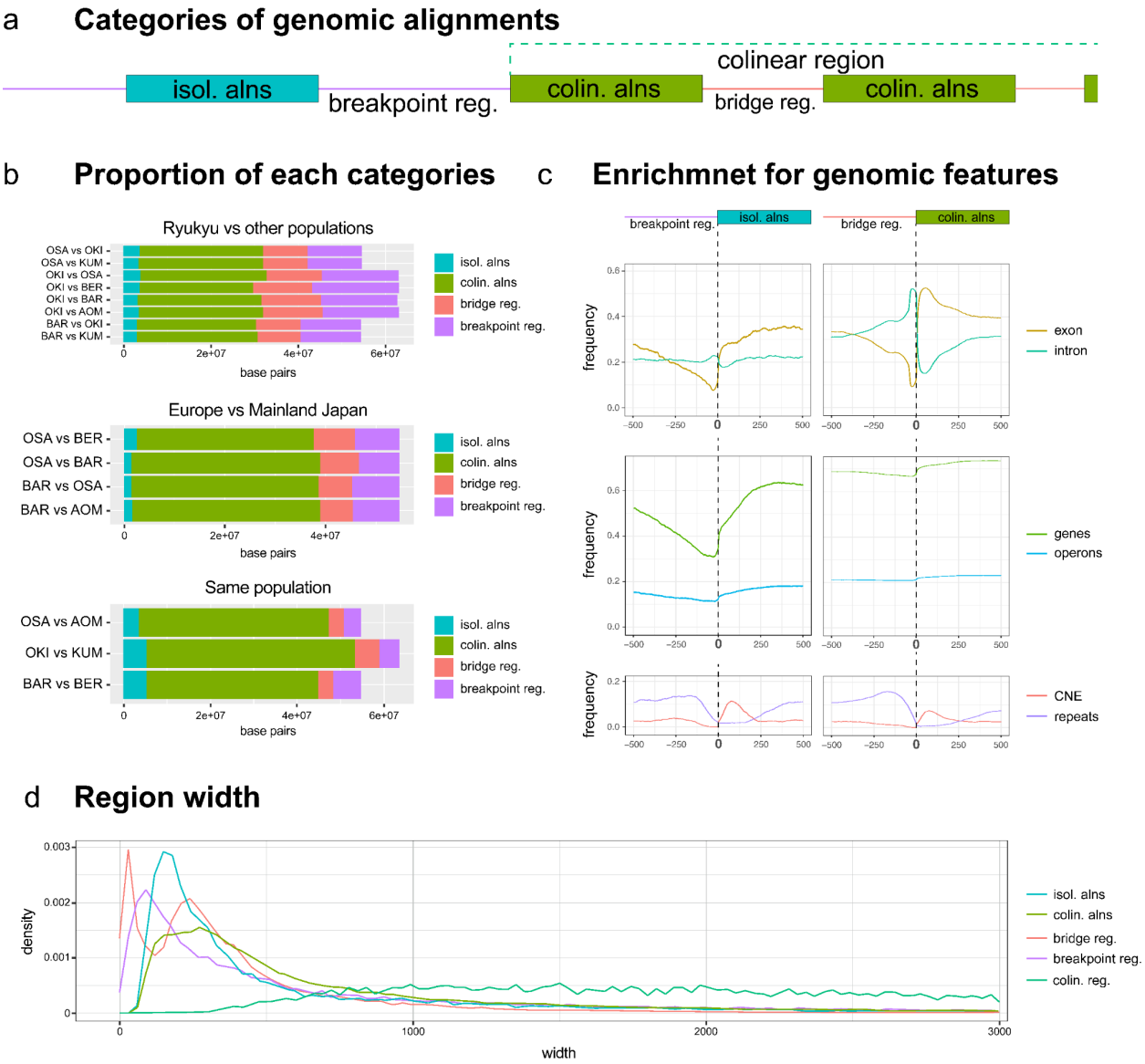


Figure 3.6: Properties of genomic alignments: (a) Categories of genomic alignments according to their participation into colinear regions. Colinear regions are defined by the uninterrupted succession of alignments on the same chromosome strand and in the same order in both genomes; (b) proportion of the categories in different alignment pairs, grouped by evolutionary distance; (c) enrichment of genomic features at the boundary between breakpoint or bridge regions and aligned regions; (d) length distribution of the alignment categories and of the colinear regions in the Okinawa-Osaka genome pair; isol. alns – isolated alignment, breakpoint reg. – breakpoint region, colin. alns – colinear alignments, bridge reg. – bridge region, CNE – conserved non-coding elements.

3.3.4 Synteny analysis on a protein level confirms the scrambling

To understand how the gene order is preserved in the context of scrambling, we reconstructed gene orthology using a set of protein sequences from 26 organisms, spanning three chordate lineages. We included multiple species of larvaceans and other tunicates, to ensure high identification of orthologous genes between *O. dioica* populations (Emms and Kelly, 2015; Emms and Kelly, 2019). As a result, all *O. dioica* individuals have more than 95% of genes in orthogroups within the given set of species (Table 3.2). Depending on the evolutionary distance, the fraction of genes in orthogroups for a pair of *O. dioica* genomes ranges from 70% (e.g., Okinawa-Barcelona) to 90% (e.g., Okinawa-Kume), suggesting a variation in gene content between populations (Table 3.6). The rest of the genes might have diverged to the point that there is not enough sequence similarity left to identify them as orthologs. Strikingly, the following vertebrate genomes have a similar amount of genes in orthogroups as the two *O. dioica* pairs mentioned above: zebrafish (*Danio rerio*) and mouse (*Mus musculus*) share around 70% of orthologous genes, while mouse and human (*Homo sapiens*) share about 90% (see Appendix 3,4).

Table 3.6: Proportions of genes in orthogroups between pairs of *O. dioica* genomes.

	Number of genes	Barcelona	Bergen	Aomori	Osaka	Okinawa	Kume
Barcelona	14168		11470 (~81%)	11719 (~83%)	11805 (~83%)	11749 (~83%)	11750 (~83%)
Bergen	16899	12044 (~71%)		12529 (~74%)	12377 (~73%)	12752 (~76%)	12529 (~74%)
Aomori	15160	11806 (~78%)	11692 (~77%)		13327 (~88%)	12495 (~82%)	12444 (~82%)
Osaka	15662	11903 (~76%)	11787 (~75%)	13404 (~86%)		12687 (~81%)	12546 (~80%)
Okinawa	17109	12065 (~71%)	12268 (~72%)	12719 (~74%)	12729 (~74%)		15137 (~89%)
Kume	16711	11951 (~72%)	12135 (~73%)	12535 (~75%)	12499 (~75%)	14998 (~90%)	

Within the orthologous gene set, a significant fraction of genes have one-to-one relationships across all pairs of *O. dioica*, with the highest number reported for the sister genomes (for example, Okinawa-Kume have 13,095 single-copy orthologs; Fig. 3.7a and Table 3.7). The only exception is the Bergen genome which has a generally low number of single-copy orthologs with all *O. dioica*. However, it still shares the highest number of one-to-ones with the Barcelona genome, but not vice versa. Apart from that, the Bergen genome has the highest number of genes not found in other *O. dioica* (~450 non-orthologous genes; Fig. 3.7b). For orthology reconstruction, we used the gene models generated by Denoeud et al. (2010) with different software to keep the original functional annotation. Therefore, such variety in gene content likely reflects annotation bias, since mixing gene annotation methods inflates the apparent number of lineage-specific genes (Weisman et al., 2022).

Chromosomal assemblies share a similar number of one-to-one orthologs, although Okinawa always has fewer genes in common with either Osaka or Barcelona than the last two share together (Fig. 3.7a and Table 3.7). Interestingly, the number of one-to-ones shared by

these three genomes is significantly lower (7,717), suggesting that the set of single-copy orthologous genes varies across populations.

Table 3.7: Proportions of genes with one-to-one orthologous relationships between pairs of *O. dioica* genomes.

	Number of genes	Barcelona	Bergen	Aomori	Osaka	Okinawa	Kume
Barcelona	14168		9742 (~69%)	10075 (~71%)	10038 (~71%)	9649 (~68%)	9580 (~68%)
Bergen	16899	9742 (~58%)		9463 (~56%)	9436 (~56%)	9254 (~55%)	9202 (~55%)
Aomori	15160	10075 (~67%)	9463 (~62%)		11825 (~78%)	9991 (~66%)	9894 (~65%)
Osaka	15662	10038 (~64%)	9436 (~60%)	11825 (~76%)		9898 (~63%)	9855 (~63%)
Okinawa	17109	9649 (~57%)	9254 (~54%)	9991 (~58%)	9898 (~58%)		13095 (~77%)
Kume	16711	9580 (~57%)	9202 (~55%)	9894 (~59%)	9855 (~59%)	13095 (~78%)	

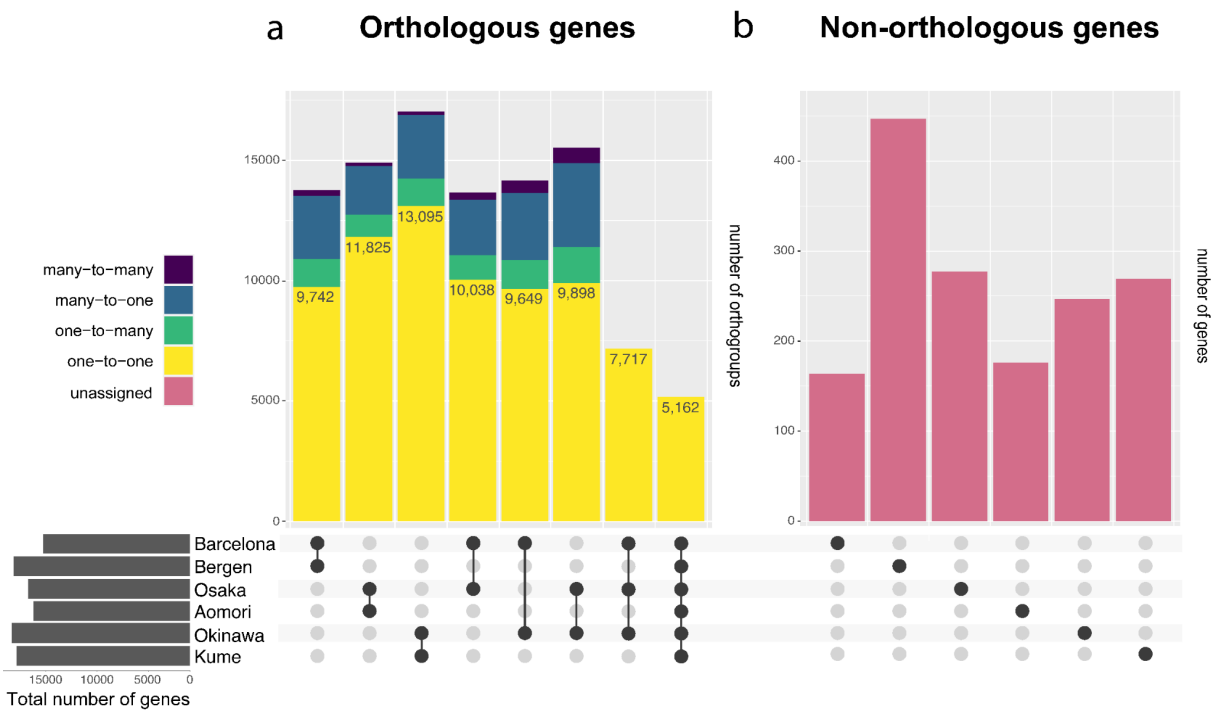
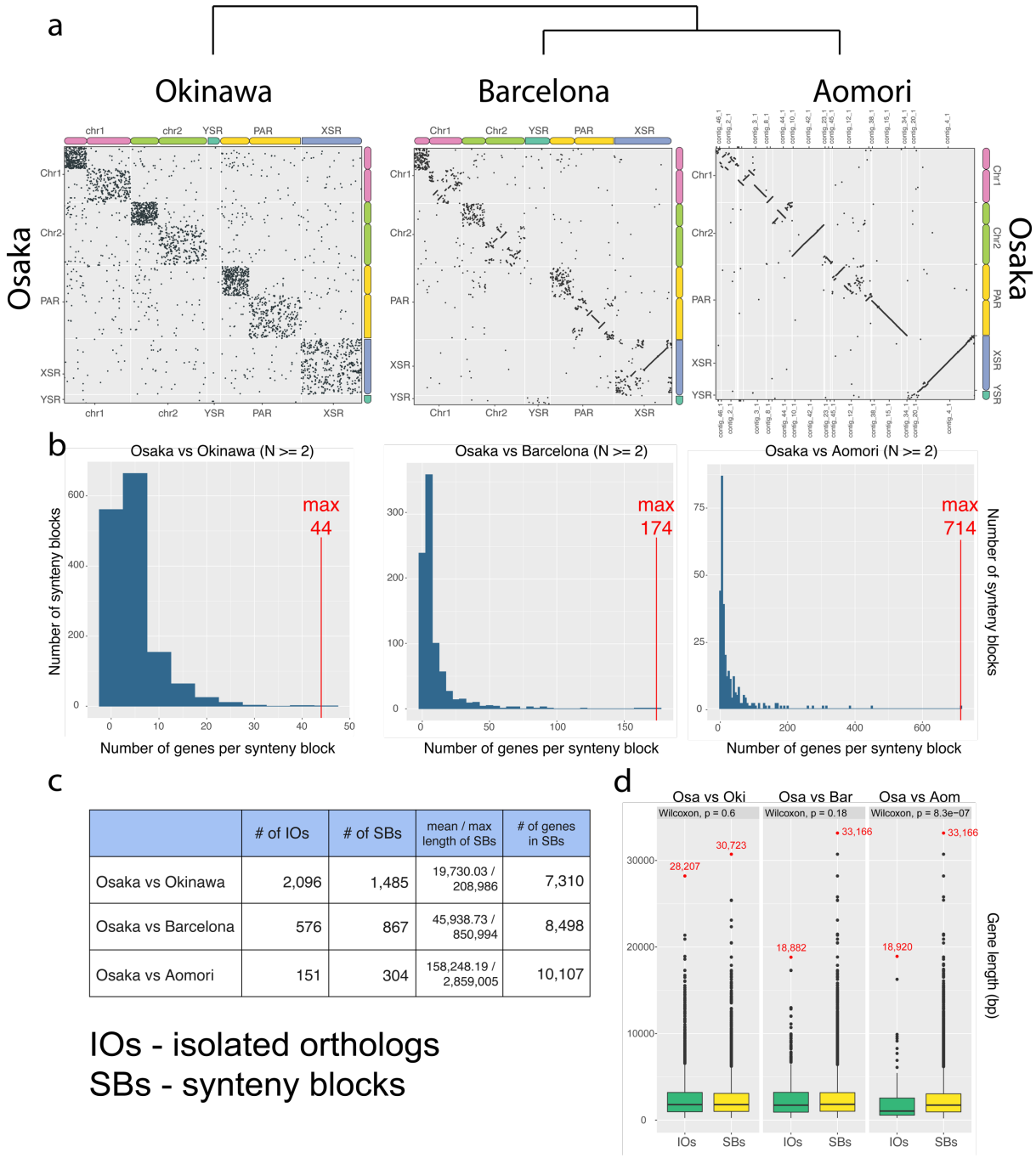


Figure 3.7: Statistics of orthologous gene assignment across six *O. dioica* genomes performed with OrthoFinder: (a) number of orthogroups shared between genomes, (b) number of genes with no orthologs in other *O. dioica* genomes.

Apart from one-to-one relationships, a significant fraction of genes in all genomes has multiple orthologs in the other genomes. These genes get assigned in "one-to-many", "many-to-one" or "many-to-many" orthogroup categories (Fig. 3.7a). We found that some of these cases represent actual population-specific gene duplications, as shown for the *Hox4* gene that has one copy in Okinawa and Osaka, but two copies in Barcelona (see below; Fig. 3.13a). Genes with many-to-many relationships in two genomes are often hard to resolve due to the presence of highly-conserved domains. At the same time, we could also identify artifacts of annotations when a single gene in one genome gets split into two genes in the other genome due to long intron or transposon insertion. A situation when one gene has too many orthologs with another genome often indicates a problem related to protein-coding transposable elements. Given that the tools used for *de novo* repeat prediction are primarily trained on repeats available in databases, species-specific transposons with unknown structures often get overlooked and remain unmasked and, thus, predicted as genes in following steps. One example is the *Oikopleura*-specific retrotransposon with an incomplete structure called *Odin,* whose copies often get annotated as genes in *O. dioica* genomes. Therefore, given identified artifacts of the annotations, we focused our next analysis on the set of only single-copy resolved orthologs.

To compare gene order across populations, we visualized single-copy orthologs with macro-synteny plots to show positions of the same gene in a pair of genomes (Fig. 3.8a, Appendix 7). We used Osaka as a reference genome for the figure to show how the gene order is preserved at all evolutionary distances. Similar to the whole-genome alignments, gene order rearrangements are mostly restricted to arms and sex-specific regions of the homologous chromosomes, with interchromosomal gene translocations rarely observed. Plus, there are various levels of scrambling at different evolutionary distances, with the gene order in the Ryukyu genomes being more shuffled compared to the rest. At the same time, sister genomes have longer stretches of synteny, with more genes along the diagonal line (e.g., Osaka-Aomori pair). Therefore, we can conclude that the protein orthologous set fully supports our finding that genomic sequences in *O. dioica* have been rearranged to an extent where even the gene order is scrambled.



	# of IOs	# of SBs	mean / max length of SBs	# of genes in SBs
Osaka vs Okinawa	2,096	1,485	19,730.03 / 208,986	7,310
Osaka vs Barcelona	576	867	45,938.73 / 850,994	8,498
Osaka vs Aomori	151	304	158,248.19 / 2,859,005	10,107

Figure 3.8: Comparison of synteny blocks at different evolutionary distances: (a) macro-synteny plots built based on one-to-one orthologs; (b) distribution of genes across synteny blocks; (c) number of isolated orthologs, synteny blocks, and genes in synteny blocks; (d) size distribution of isolated orthologs and genes in synteny blocks. N – number of genes, p – p-value, bp - base pairs. Osa – Osaka, Oki – Okinawa, Bar – Barcelona, Aom – Aomori.

Using the coalescence algorithm implemented in the GenomicBreaks R package, we predicted synteny blocks for each pair of *O. dioica* genomes. We defined synteny breaks as two or more orthologous genes that occurred in the same order (colinear) in both genomes despite the orientation. Thus, the one-to-one orthologs not colinear to any other genes were marked as isolated orthologs. In agreement with whole-genome alignments, sister genomes are more colinear, having fewer but longer syntenic regions in contrast to genomes from different populations (Fig. 3.8b,c; Table 3.8). The largest synteny block found between Osaka and Aomori is 2.86 Mbp long, comprising 714 genes. In contrast, synteny blocks between Osaka and Okinawa unite no more than 44 genes covering less than 0.2 Mbp of the genomic sequence. There is no significant difference in length between isolated orthologs and genes in synteny blocks (Fig. 3.8d), although Denoeud et al. (2010) suggested that larger genes tend to stay more isolated owing to an abundance of regulatory elements. Visual comparisons with parallel plots revealed that the synteny blocks' size gradually reduces with evolutionary distance: the smallest synteny blocks observed between Okinawa and Barcelona fully intersect the larger Barcelona-Osaka and Osaka-Aomori syntenic regions (Fig. 3.9). The ZENBU view of the Okinawa genome shows that these genomic regions contain individual genes and multiple operons, with some comprising up to seven genes (operon_2321). Therefore, although scrambling shuffles the gene order, some of the operon structures seem to be preserved even at the largest evolutionary distance (Ryukyu–non-Ryukyu).

Table 3.8: Comparison of synteny blocks at different evolutionary distances. IOs – isolated orthologs, SBs – synteny blocks.

Comparison	# of IOs	mean / max length of IOs	# of SBs	mean / max length of SBs	# of genes in SBs	mean / max length of genes in SBs
Okinawa to Kume	76	1856.83 / 15360	240	240598.73 / 1951479	11247	2407.46 / 42121
Okinawa to Osaka	2096	2702.56 / 26112	1485	21619.43 / 204106	7310	2456.76 / 32391
Okinawa to Aomori	2061	2772.98 / 42121	1472	21949.35 / 219589	7377	2469.00 / 32391
Okinawa to Barcelona	1963	2781.33 / 42121	1412	22577.06 / 243899	7179	2529.42 / 32391
Okinawa to Bergen	2093	2738.78 / 24003	1421	19658.76 / 211799	6559	2427.72 / 20579
Osaka to Okinawa	2096	2614.00 / 28207	1485	19730.03 / 208986	7310	2438.88 / 30723
Osaka to Kume	2068	2609.89 / 28207	1465	19951.02 / 164547	7283	2463.35 / 33166
Osaka to Aomori	151	2032.89 / 18920	304	158248.19 / 2859005	10107	2412.49 / 33166
Osaka to Barcelona	576	2553.02 / 18822	867	45938.73 / 850994	8498	2515.74 / 33166
Osaka to Bergen	778	2708.30 / 28207	1028	34481.76 / 620155	7534	2465.98 / 23103
Barcelona to Okinawa	1963	2626.14 / 27584	1410	20172.64 / 182247	7178	2446.65 / 29169
Barcelona to Kume	1940	2647.65 / 27584	1393	20559.1 / 182247	7164	2470.59 / 38340
Barcelona to Aomori	596	2513.66 / 27584	822	47242.45 / 756789	8546	2449.74 / 38340
Barcelona to Osaka	576	2377.41 / 22262	867	44475.38 / 840174	8498	2472.45 / 38340
Barcelona to Bergen	319	2573.41 / 16013	588	67750.02 / 978787	8064	2444.64 / 25277

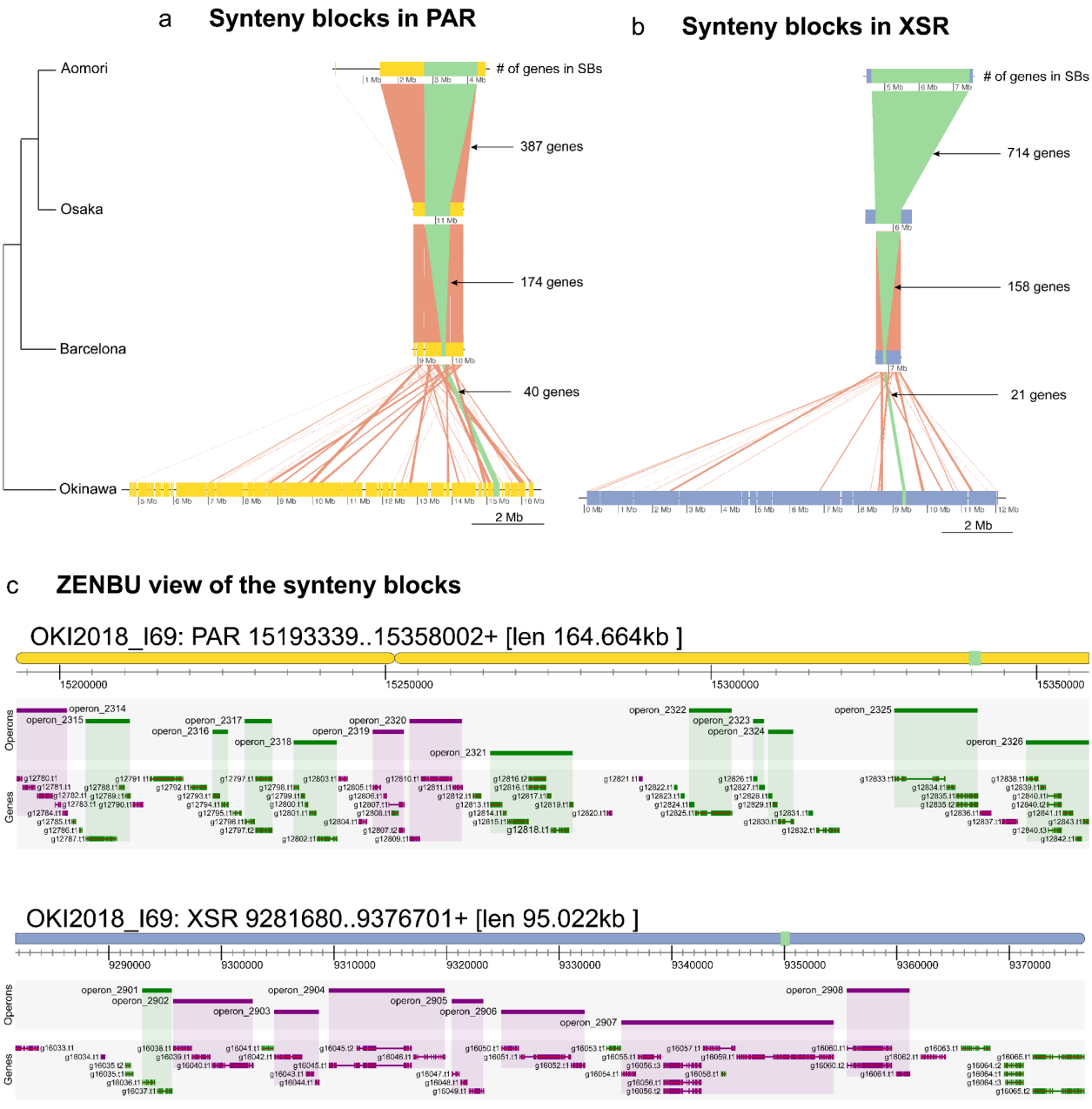


Figure 3.9: Examples of synteny block conservation: in the (a) PAR and (b) XSR, and (c) ZENBU views of these regions in the Okinawa genome. Chromosomes are color-coded: yellow – PAR, blue – XSR.

3.3.5 Chromosome arms evolve at different rates

To understand how the scrambling phenomenon affects the genome at a macro scale, we compared the distribution of synteny blocks between short and long chromosome arms. The analysis revealed that short arms tend to have fewer and significantly shorter synteny blocks compared to long arms or the XSR (Fig. 3.10a). Also, this difference is observed at all evolutionary distances, meaning that short arms are generally less conserved in *O. dioica*. Indeed, short arms are more rich in repeats and have reduced GC content (see chapter two), fewer genes and operon structures (Fig. 3.10b). Moreover, genes encoded on short arms more frequently overlap breakpoint regions than those on long arms (~50% vs. ~20%, p ~0.01), while single-copy orthologs exhibit higher dN/dS values (Fig. 3.10b). Therefore, chromosome regions seem to evolve at different rates, with a higher pace observed for the short chromosome arms. In our observations, the XSR shows the same pattern as the long arms, while the YSR represents a more unique case: it is rich in repeats and has only a few single-copy orthologs shared by the genomes, with the highest number between the YSR in the Osaka and Barcelona *O. dioica* – 12 genes. Interestingly, some of these genes belong to orthogroups that are specific to *O. dioica* and, therefore, represent good candidates for sex-determination genes in this species.

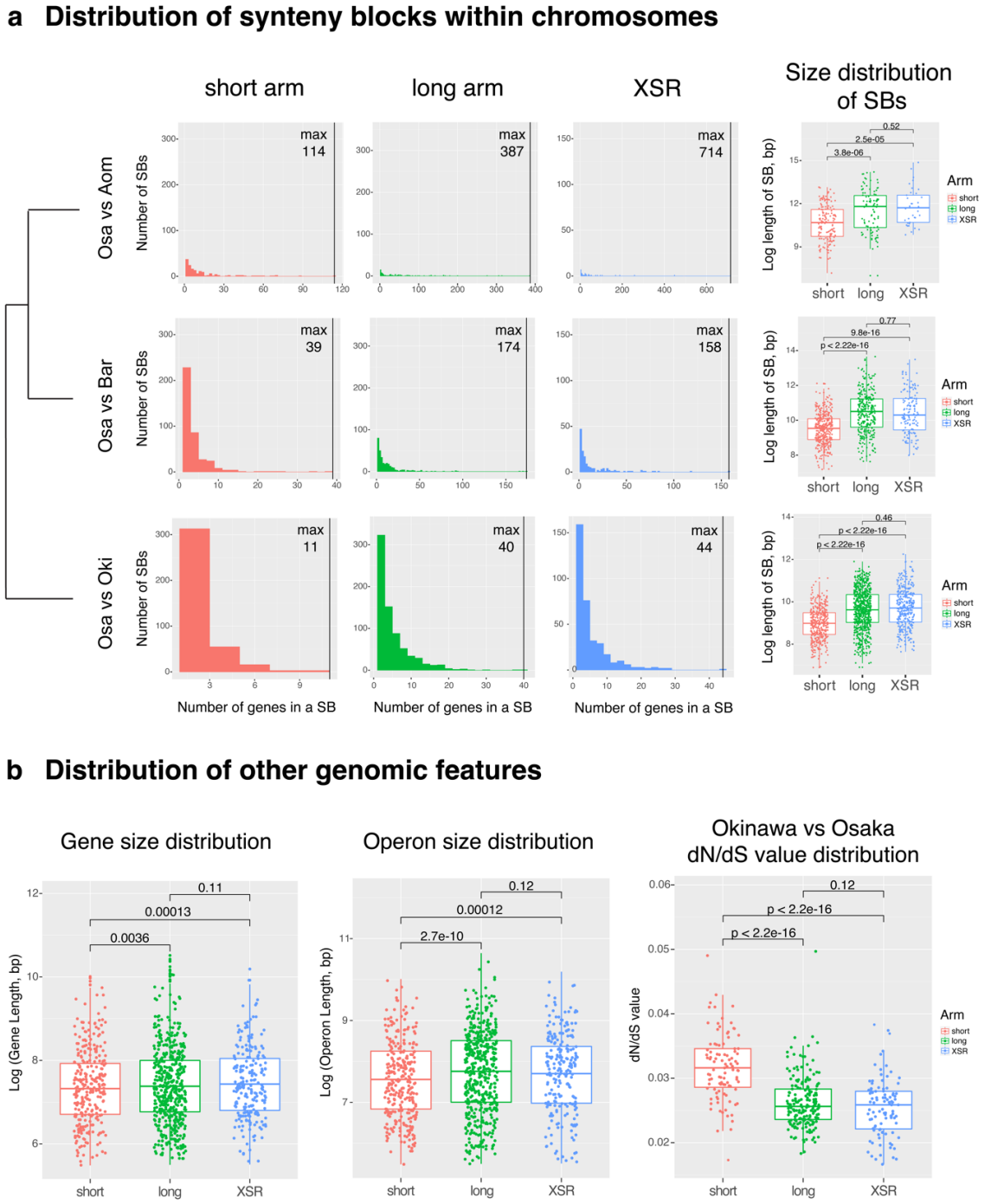


Figure 3.10: Comparison of chromosome arm preservation. (a) Distribution of synteny blocks across chromosome arms in the Osaka genome. In this case, we took the Osaka genome to show conservation of synteny blocks within the arms at all possible evolutionary distances. (b) Distribution of gene and operon sizes, and dN/dS values of orthologous genes between Okinawa and Osaka across chromosome arms in the Okinawa genome.

3.3.6 Scrambling does not preserve operon structures

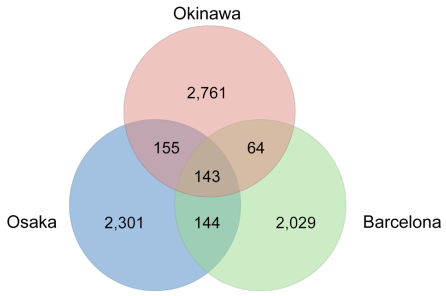
Given that a significant proportion of genes in *O. dioica* are co-expressed within operons, we investigated the degree to which operon structures are conserved between *O. dioica* populations. Here, we defined an operon as a set of co-oriented genes separated by 500 bp at most. At the chosen threshold, the Okinawa genome possesses more operon structures than the Osaka or Barcelona ones, although the proportion of genes in operons is relatively similar (~50%), as well as the distribution of operon sizes between the three populations (Fig. 3.11a,c). However, most of the operons appear to be population-specific, as determined by the exact matching of single-copy resolved orthologs: with the current analyses, only 143 full-length operons are shared between three genomes at full length (Fig. 3.11b). Functional annotation with GO terms shows that the operon gene sets in three populations are significantly enriched for genes involved in metabolic processes and housekeeping functions, such as RNA processing, translation, gene expression, and ribosome biogenesis (Fig. 3.12a; Appendix 5). On the other hand, genes involved in regulatory activities (biological regulation, regulation of biosynthetic process and gene expression) and signaling are over-represented in the non-operon gene sets (Fig. 3.12b; Appendix 6). At the same time, genes involved in processes, such as transport, can be found both out of and in operons, suggesting some level of flexibility.

In the context of scrambling, operons are less preserved than other protein-coding features since only ~50% of operons completely intersect the synteny blocks, in contrast to genes (~70%) and exons (~80%; Fig. 3.11d). The apparent interchromosomal translocation of a *PAC3* gene between operons encoded on the chr1 (Okinawa) and XSR (Osaka and Barcelona) is one example of the operon rearrangement (Fig. 3.13a). Operon switching is also exemplified in Fig. 3.13b, where a single gene has moved between operons encoded on the XSR (Osaka and Barcelona) and PAR (Okinawa). At the same time, some of the operons are nonetheless conserved, including an operon of nine genes that is common to the three genomes (Fig. 3.13c). In some cases, even an interchromosomal translocation of the whole operon without breaking can be observed, such as a five-gene operon that was translocated between the chr1 (Okinawa) and PAR (Osaka) (Fig. 3.13d). However, the same operon appears to be only partially preserved and duplicated in Barcelona, with a following translocation of one of the copies to the chr2. To sum up, operons appear to exhibit little evidence of conservation between *O. dioica* from Okinawa and Osaka and, unlike genes and their intron/exon structures, might operate under different selective constraints. On the other hand, there might be a limit to which a gene can be co-expressed within an operon structure in *O. dioica* with a trend towards genes with metabolic and house-keeping functions.

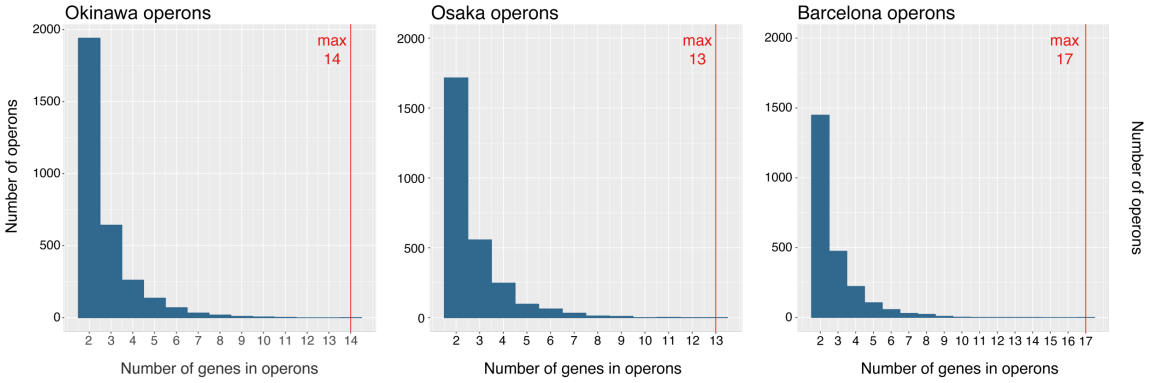
a Operon structures

	Okinawa	Osaka	Barcelona
# of operons	3,123	2,743	2,380
# of genes in operons	8,535	7,451	6,659
# of genes out of op.	8,756	8,269	7,613

b Operons common between genomes



c Operon size distribution



d Protein-coding features in synteny blocks

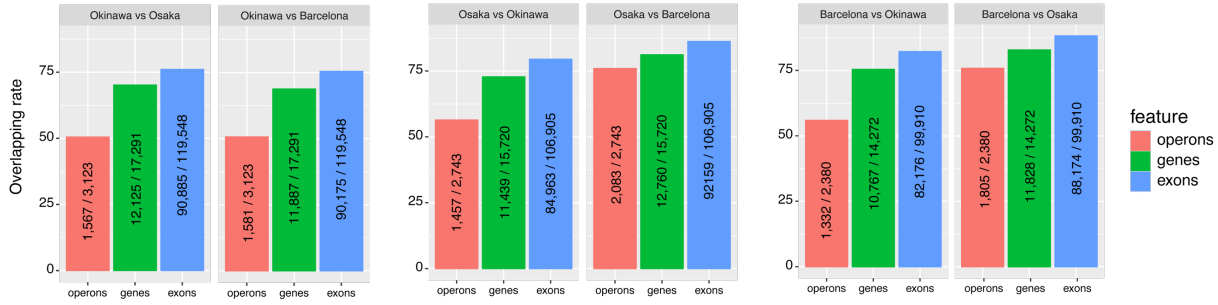


Figure 3.11: Comparison of operon structures in the Okinawa, Osaka and Barcelona genomes: (a) numbers of predicted operons; (b) the proportion of the same operons shared between three populations; (c) operon size distribution; (d) the proportion of genomic features in synteny blocks between genome pairs.

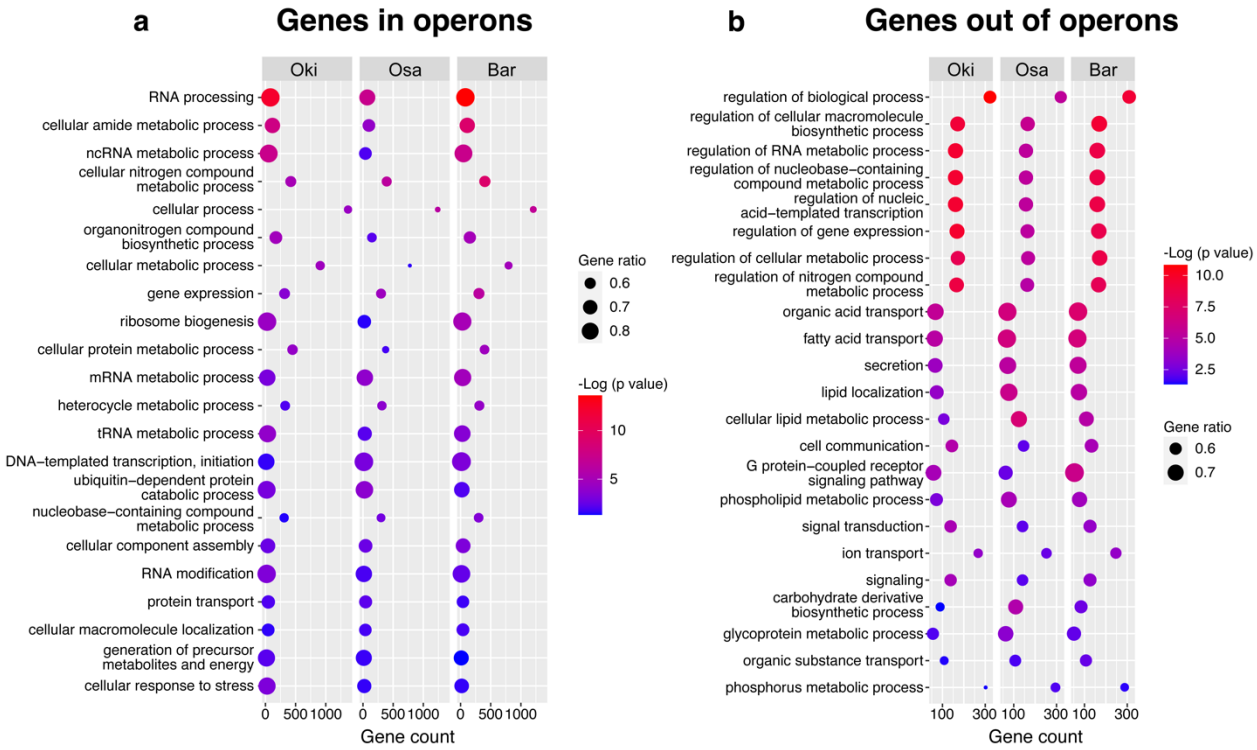


Figure 3.12: Enrichment analysis of genes in and out of operons in the Okinawa, Osaka and Barcelona genomes: (a) 25 common GO terms enriched in the operon gene sets; (b) 25 common GO terms enriched in the non-operon gene sets. “Gene ratio” was calculated by dividing the number of observed genes (“Count”) by the number of expected genes (“Size”; see Appendix 5,6).

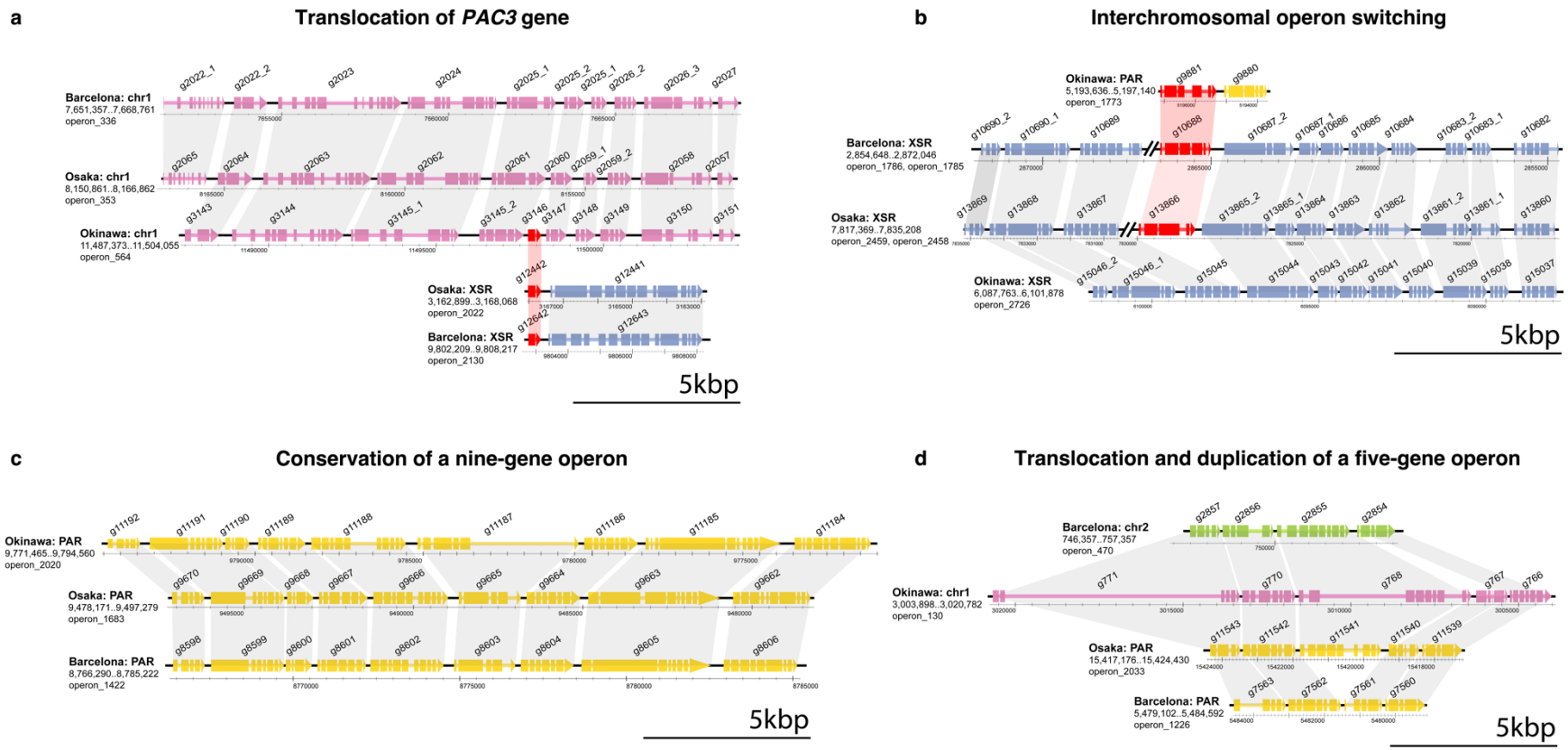


Figure 3.13: Example of the operon preservation between the Okinawa, Osaka and Barcelona *O. dioica* genomes: (a) the *PAC3* gene translocation between two operons in the chr1 and XSR; (b) an operon rearrangement in the XSR in the Osaka and Barcelona genomes involving a single gene insertion from the PAR; (c) conservation of a nine-gene operon between three populations; (d) translocation of a 5-gene operon between the PAR and chr1, and its duplication and translocation of another copy to the chr2 in the Barcelona genome. Chromosomes are color-coded: pink – chr1, green – chr2, yellow – PAR, blue – XSR.

3.3.7 Scrambling does not preserve ancestral gene clusters

To understand how scrambling affects individual genes, we looked at the ancestral clusters, which in most animals represent highly conserved sets of genes typically adjacent to each other within genomes because of the shared regulatory mechanisms. The Hox gene cluster, required for anterior-posterior patterning during embryonic development, represents one of the most noted examples of gene synteny in metazoans. The cluster has been atomized in the Bergen *O. dioica* genome, which has only eight genes instead of 13, with apparent duplication of the posterior *Hox9* gene. Moreover, the remaining Hox genes do not cluster together but have been wholly dispersed throughout six scaffolds in the Bergen genome (Seo et al., 2004; Edvardsen et al., 2005; Blanchoud et al., 2018). Consistent with this pattern, the locations of Hox genes appear to also change between three populations of *O. dioica*, although the genes stay within the same chromosome arm or sex-specific region (Fig. 3.14a and Table 3.3a). Unlike the Bergen *O. dioica*, all three genomes seem to have only one copy of the *Hox9* gene. We also observed a unique duplication of the *Hox4* gene in the Barcelona genome that may represent a pseudogene.

Other examples of evolutionary conserved clusters, that have been broken between populations of *O. dioica,* are the deuterostome-specific pharyngeal cluster (Simakov et al., 2015; Fig. 3.14b and Table 3.3b) and the NK cluster that was present in the last common ancestor of bilaterians (Luke et al., 2003; Fig. 3.14c and Table 3.3c). Therefore, ancestral gene clusters appear to have been shuffled in the *O. dioica* populations not only relative to other species, but also relative to one another.

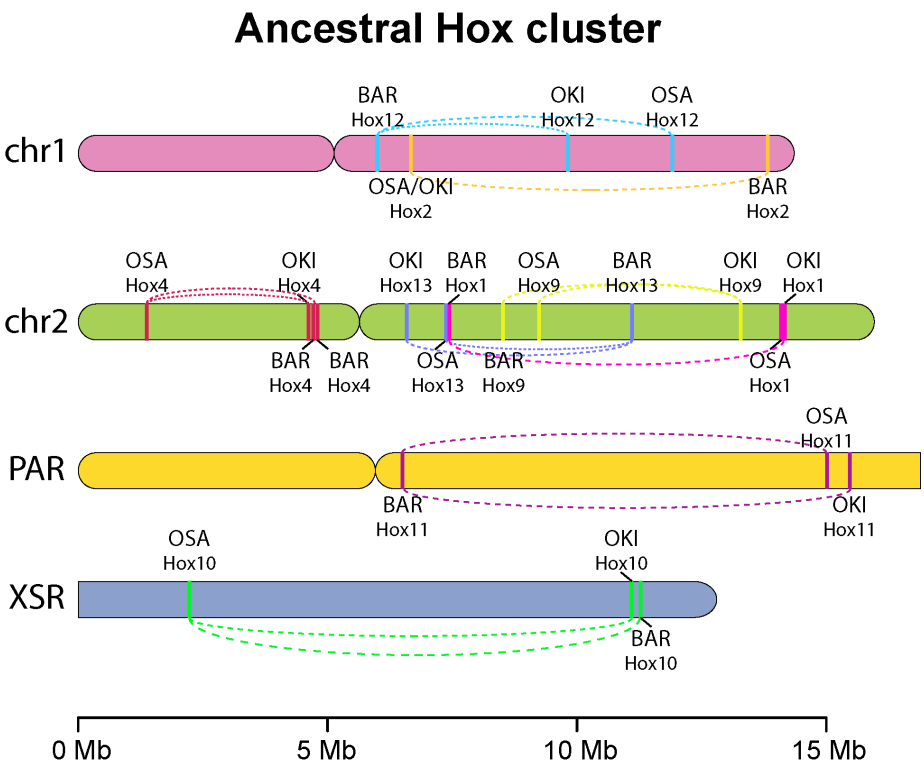


Figure 3.14a: Chromosomal locations of the Hox cluster gene orthologs in the three different populations of *O. dioica*. Positions are transposed upon coordinates of the Okinawa genome.

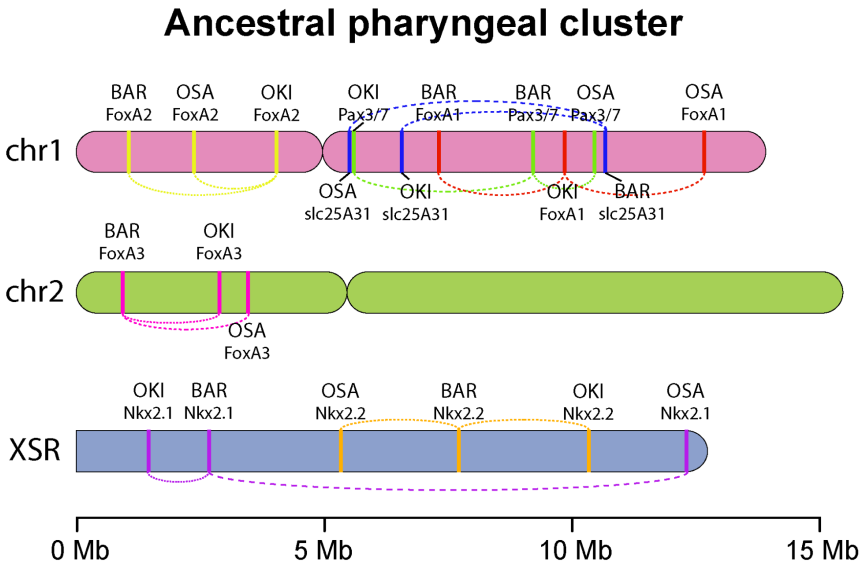


Figure 3.14b: Chromosomal locations of the pharyngeal cluster gene orthologous in the three different populations of *O. dioica*. Positions are transposed upon coordinates of the Okinawa genome.

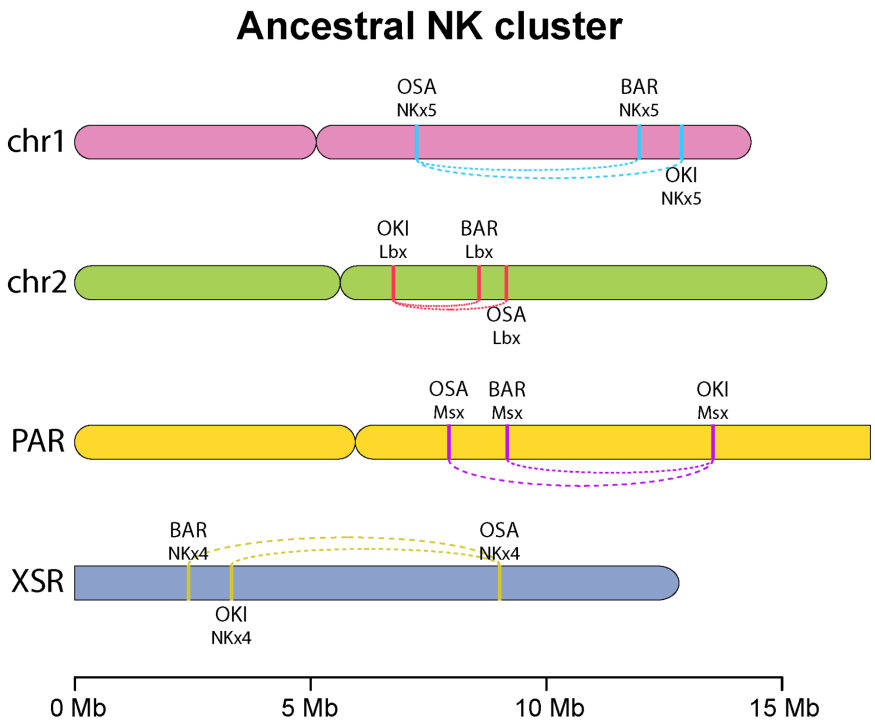


Figure 3.14c: Chromosomal locations of the NK cluster gene orthologous in the three different populations of *O. dioica*. Positions are transposed upon coordinates of the Okinawa genome.

3.4 Discussion

3.4.1 Scrambling of *Oikopleura* genome

In this chapter, we reported extensive scrambling between the genomes of three independent *O. dioica* populations from the Northern hemisphere. Currently, these animals are classified as a single species owing to the presence of separate sexes and no apparent differences in morphology. The observed scrambling is corroborated by the fact that the *O. dioica* genomes have no synteny left with other animals, such as *Ciona intestinalis* and amphioxus (Denoeud et al., 2010), which one may use as a proxy to the ancestral chordate linkage groups (Simakov et al., 2020). However, the speed at which the synteny has been lost remains unknown, given that no chromosome-scale assemblies are available for other larvaceans.

Our study design using six individual genomes of *O. dioica* ensures that we are able to support our findings with at least two inter-population comparisons. This design, combined with the molecular clock estimation (see below), allowed us to examine the scrambling at a closer evolutionary distance compared to previous studies on other organisms, for example, in *Drosophila* (Drosophila 12 Genomes Consortium et al., 2007) or cephalopods (Albertin et al., 2022). In cephalopods, this phenomenon has been associated with emergence of new regulatory mechanisms (Schmidbaur et al., 2022). We believe that the extent of rearrangements in *O. dioica* has been difficult to estimate until now due to the lack of multiple chromosome-scale assemblies for the species, although previous comparisons have shown sequence variation between populations (Denoeud et al., 2010; Wang et al.,2015; Wang et al., 2020). The chromosomal resolution of our genomes allowed us to better refine gene order and assess the scrambling phenomenon at multiple levels, revealing a consistent pattern of heterogeneity in evolutionary rates between chromosome regions in *Oikopleura*. Thus, ongoing efforts to obtain longer contiguous high-quality genome assemblies for other species may prove helpful in assessing the prevalence of this phenomenon in different clades.

3.4.2 The unit of scrambling

Genes and their constituent exons might be the smallest units of synteny preserved between populations. Although, even their structures are not wholly protected from the mutational processes that produce scrambled genomes, since some genes still overlap breakpoint regions. In fact, we already found some rearrangements in ancestral gene clusters between chromosomal genomes, especially in the Hox cluster. We assume that more differences in gene content are yet to be discovered given that the orthologous gene set seems to vary between populations. However, the current gene models are not entirely free from artifacts. Therefore, updating annotations is required before analyzing lineage-specific gene variations.

Operon structures are the least conserved among protein-coding features. Most rearrangements within operons, such as a single gene inversion or translocation, would likely decouple the gene from its primary regulatory elements and result in abnormalities. Given that, it is puzzling to learn that such scrambling has occurred in an organism where the expression of a significant gene fraction relies upon the existence of operons. That suggests that the primary function of the operons may not be related to co-expression. Indeed, the expression level of genes within operons is not always correlated as well as the functional categories of genes (Danks et al., 2015). On the other hand, operon structures could decrease gene reliance on

keeping their own promoters by allowing a gene to be inserted into an operon with the existing transcriptional machinery. This way, the presence of operon-like structures makes gene movements easier and helps to preserve gene expression in the context of scrambling. However, gene switching between operons might be limited to the extent that not all categories of genes can be co-expressed within operons, given that larger genes involved in regulatory processes are found only in the non-operon gene set and, therefore, may rely on their own promoters.

3.4.3 Mechanism behind the scrambling

The mechanism behind the genome scrambling in *O. dioica* remains unknown since the breakpoint regions are too large at all evolutionary distances to identify the exact spot where the DNA was initially broken. Although it is tempting to speculate that the loss of the c-NHEJ pathway in *O. dioica* created synergies that promote scrambling since the a-NHEJ mechanism used instead (Deng et al., 2018) is slower and might allow for chromatin movements before the repair of a DNA double-strand break. At the same time, using the alternative pathway, "cut-and-paste" DNA transposons, such as MITE, may act as a source of microhomologies facilitating repair by the a-NHEJ. Therefore, the low repeat content in *O. dioica* might reflect genome instability that also causes scrambling. Although the genome is repeat-sparse, a small fraction of interspersed repeats with a high identity of their copies is enough to facilitate rearrangements via other mechanisms, such as homologous recombination. Also, the active transposons can directly induce rearrangements by inserting into and spreading within the genome. Given that the extent of scrambling strongly correlates with the evolutionary distance and, hence, time since divergence, we assume that the mechanism behind scrambling is more likely to involve a gradual accumulation of rearrangements without significant variations in gene content. However, closer distant comparisons, for instance, haplotypes within one individual or multiple generations of *O. dioica* from the same laboratory strain, are required to get more precise insights into the molecular mechanism. The overall result of genome scrambling is that genes are being put into new locations in the genome, creating opportunities for the evolution of novel traits and regulatory mechanisms.

3.4.4 Molecular clock estimation

The various levels of scrambling observed across *O. dioica* populations let us speculate about the existence of at least three lineages of dioecious *Oikopleura*. However, more may be discovered with extended sampling of the animals from the Southern hemisphere. When computing the alignments, we ensured that no genome plays a special role by using paired comparisons that removed us from the need to resolve phylogenetic relationships prior to analysis. Also, later our observations were confirmed by comparisons of multiple phylogenetic markers (nuclear 18S, ITS, and mitochondrial COI) performed by Aki Masunaga (Masunaga et al., 2022) and the molecular clock calculated by Michael Mansfield based on the set of single-copy resolved orthologs presented in this chapter. The analysis showed that the three different *O. dioica* populations had shared an ancestor as recently as ~20-30 Mya. In contrast, the North Atlantic and North Pacific populations have diverged less than 10 Mya (Fig. 3.15a). To compare with, species within another tunicate genus, such as *Ciona*, represent cases of either recent (*C. intestinalis* vs. *C. robusta*, ~11 Mya; 35 breakpoint regions per megabase) or more ancient (*C. savignyi* vs. other *Ciona* at more than 100 Mya, ~33 breakpoints per megabase) speciation events (Fig. 3.15b). One has to keep in mind that these numbers are approximate owing to a lack of suitable fossil taxa that would allow us to calibrate the internal clock nodes of the

appendicularian clade. Therefore, we must limit ourselves on broad conclusions that stay true even if these numbers would vary by almost an order of magnitude (manuscript in preparation).

Also, *O. dioica* competes with *Drosophila* species regarding the number of breakpoints per megabase accumulated within a time frame. Even the morphologically similar and closely related species *D. melanogaster* and *D. mauritiana* that diverged ~3.5 Mya have fewer breakpoints (~19) per megabase than any pair of *Oikopleura* genomes (73 and 140 breakpoint regions per megabase pairs aligned; Fig. 3.15c). Scaling the number of breakpoints to the estimated divergence times reaffirms that *O. dioica* genomes accumulate breakpoints faster than other animal species, with more than five breakpoints per megabase aligned per million years.

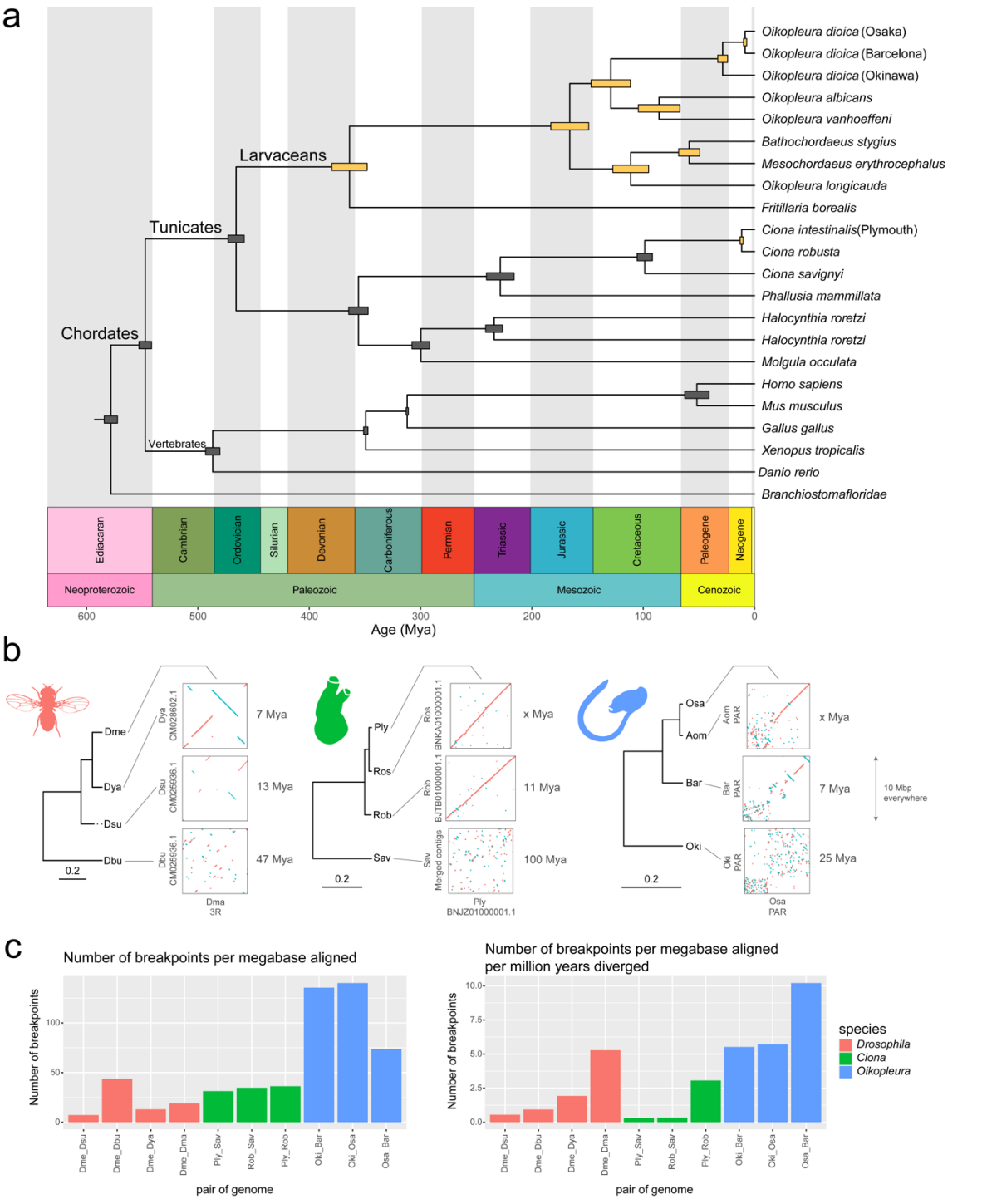


Figure 3.15: Divergence time estimates and number of breakpoints per million years for various chordate species: (a) Time-scaled phylogenetic tree including several larvaceans, tunicates, and vertebrates; (b) 10-Mbp genomic regions in *Drosophila*, *Ciona* species and *O. dioica* lineages at different divergence time; (c) The number of breakpoint regions for the three clades, scaled according to the number of aligned megabases (left) and number of aligned megabases per million years diverged (right).

3.5 Conclusions

We report that the genome of the planktonic tunicate *O. dioica* has been reconstructed multiple times over the past ~20-30 million years. At the same time, the organism's morphology and ecology stayed unchanged, allowing globally distributed populations to be identified as the same species. The genome scrambling was observed on multiple levels, including operon and gene structures, with the latter being more preserved. Also, a different pace of rearrangements was reported for chromosome arms, with the short arms evolving at faster rates. Further work is required to understand the role of mobile elements, gene operons, and the loss of c-NHEJ in promoting scrambling. To our knowledge, *O. dioica* possesses the fastest rearranging animal genome described so far, showing much more significant changes in gene order when compared to other species at similar alignment distances. Overall, such genome revolution might have led to the emergence of at least three independent lineages of dioecious *Oikopleura*, and more sampling is required to study populations from the Southern hemisphere. Altogether that makes *O. dioica* a perfect model to study genotype-phenotype correlation and the possible existence of new regulatory mechanisms.

Chapter Four

Updated repeat and gene predictions in *Oikopleura dioica* using cross-genome protein and transcript alignments

4.1 Background

The annotation of the *Oikopleura dioica* genomes showed a significant variation in gene number between populations (see chapter two and three). However, reconstruction of the gene orthology in chapter three revealed that some of these differences could be explained by artifacts of the annotation. One of the problems was that some mobile elements, for example, the *Oikopleura*-specific family of retrotransposons called *Odin*, were not identified and properly masked before the gene prediction step, resulting in their annotations as potential protein-coding genes. That created a need for more comprehensive prediction of both repeats and genes in the *O. dioica* genomes.

Currently, two main ways exist to identify transposable elements (TEs) in a genome. The first approach is homology-based, where repeat sequences from closely-related species available in public databases, such as RepBase (Bao et al., 2015) and Dfam (Storer et al., 2021), are used as queries to search against the genome. The second way is to predict TEs *de novo* using software trained to search for the family-specific structural features, for example, long terminal repeat (LTR) sequences in LTR retrotransposons or terminal inverted repeats (TIRs) of DNA transposons. Repeat sequences in public databases are often manually curated; therefore, the first approach identifies real TEs but may overlook organism-specific ones. Moreover, we have to keep in mind the high genomic divergence observed between *O. dioica* populations (see chapter three) and the fact that all *O. dioica* TE sequences in databases have been obtained from the Bergen (OdB3) *O. dioica* (Volff et al., 2004; Naville et al., 2019). Thus, use of only the first approach may result in undermasking of the genome. The second approach has higher potential to predict novel repeat sequences, but these sequences have to be appropriately studied in order to validate that they are coming from actual TE-containing loci and to classify them into families.

This chapter provides a more exhaustive masking of the six *O. dioica* genomes using population-specific repeat libraries generated with two approaches. Annotation of newly masked genomes resulted in fewer protein-coding genes compared to chapters two and three, but still showed variation in gene number, which, at this point, we consider to be population-specific. In the last section, I discuss further research projects that we envision in the OIST Genomics and Regulatory Systems Unit to further improve the work presented in this thesis.

4.2 Methods

4.2.1 Annotation of transposable elements

First, a de novo approach was used to find potential TEs within genomes based on the family-specific structural features (Fig. 4.1). Here, EDTA v1.9.6 (Extensive de novo TE Annotator) was used to identify and filter candidates of three TE subclasses: LTRs, TIRs, and Helitrons (Ou et al., 2019). The EDTA pipeline incorporates LTR_Finder and a parallel version of LTRharvest together with LTR_retriever for identification of LTR elements, HelitronScanner – for helitrons and GRF and TIR-Learner – for TIR transposons. In addition to that, miniature inverted-repeat TE (MITE) candidates were identified with MITE-Hunter v11-2011 (Han and Wessler, 2010), whereas short interspersed nuclear element (SINE) loci were predicted with SINE_Finder using standard parameters (Wenke et al., 2011). Finally, RepeatModeler v2.0.2a (Flynn et al., 2020) was used to locate other potential repeat families that might not have been identified with other software. All putative TE loci identified *de novo* were joined together into one file, merging the overlapping regions with the "bedtools merge" (v2.29.2) function.

For homology-based identification of TE-containing loci, protein sequences of TEs from *Ciona* and *Oikopleura* species were collected from the RepBase v25.10 and searched against the *O. dioica* genomes using LAST "DNA-versus-proteins" pipeline v1238 with the parameters "-D1e9 --codon -X1 -m100 -p" (Yao and Frith, 2021). The same pipeline was used to align ORF protein sequences of LTR elements extracted from the Gypsy Database (https://gydb.org; Llorens et al., 2010). Moreover, nucleotide sequences of TEs identified in the OdB3 *O. dioica* genome by Naville et al. (2019) were used as queries for BLASTn and tBLASTx searches against the genomes. Hits of TE loci obtained with LAST, BLASTn, and tBLASTx were joined together into one file. The overlapping hits were merged with "bedtools merge" and resolved manually.

Next, homology-based and *de novo* TE-containing regions that overlapped for at least 80% of the sequence were merged together. In order to remove redundancy, genomic sequences of each locus were extracted with "bedtools getfasta" and clustered together with cd-hit (parameters "-d 0 -aS 0.8 -c 0.8 -G 0 -g 1 -b 500"), following the "80-80-80" rule: "any two sequences longer than 80nt that share more than 80% identity over 80% of their sequences belong to the same family" (Wicker et al., 2007). Finally, TEs that were less than 500 bp and did not group with any other sequences were removed from further analysis. The resulting repeat library was provided as input for a RepeatMasker (v4.1.0; Smit, Hubley and Green at http://repeatmasker.org) search against the genome.

This workflow was used to generate population-specific repeat libraries for the OKI2018_I69, OSKA2016v1.9, and Bar2_p4 genomes and to mask them. In addition, repeat sequences from Okinawa and Osaka were used to mask the KUM-M3-7f and AOM-5-5f assemblies, correspondingly, whereas the Bergen *O. dioica* genome was masked using the repeat library generated by Naville et al. (2019).

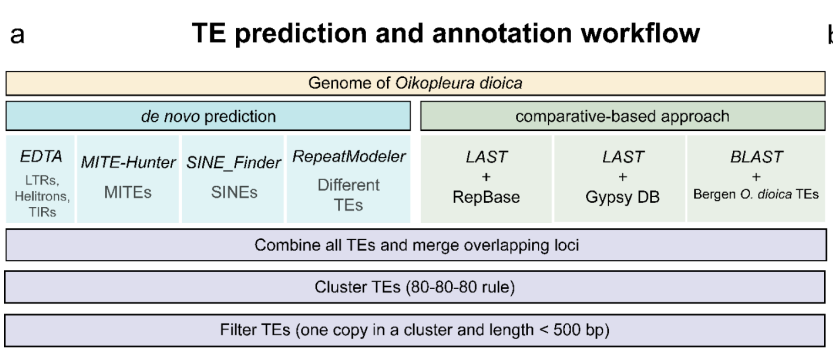


Figure 4.1: Repeat identification in the *O. dioica* genomes: (a) TE prediction and annotation workflow. (b) Number of putative TEs in the three *O. dioica* genomes generated after each step of the annotation.

4.2.2 Transcriptome assembly, gene prediction and gene orthology reconstruction.

To update the gene annotation, new transcriptome assemblies for the Okinawa and Osaka *O. dioica* were generated, using RNA-Seq data from six developmental time points: eggs, 32-cells, tailbud, hatchling, heartbeat and tailshift. After quality assessment and data filtering, Illumina RNA-Seq reads were pooled together and *de novo* assembled with Trinity v2.14.0 (Grabherr et al., 2011). Quality assessment with BUSCO v3.0.2 showed that the assembled transcriptomes exhibit higher completeness scores compared to the OikoBase (Danks et al. 2013) and the Barcelona assembly shared earlier by Prof. Cristian Cañestro (unpublished; Table 4.1). However, a high proportion of transcripts are duplicated (~74%), suggesting that there are genes that appear to express on multiple or even all the stages of *O. dioica* development.

Table 4.1: Comparison of the transcriptome assemblies of *O. dioica* from Okinawa, Osaka, Barcelona and Bergen. C – complete genes, S – complete and single copy, D – complete and duplicated, F – fragmented, M – missing.

	Okinawa	Osaka	Barcelona	Bergen (OikoBase)
Number of transcripts	1,245,387	1,224,601	192,268	17,212
Transcripts longer than 1000 bp	97,778	86,171	19,078	7,543
N50 length (bp)	602	574	128	1,587
BUSCO score (metazoa_odb9)	C:84.4% [S:10.4%, D:74.0%], F:4.5%, M:11.1%	C:85.3% [S:12.2%, D:73.1%], F:5.9%, M:8.8%	C:79.0% [S:41.5%, D:37.5%], F:5.5%, M:15.5%	C:75.4% [S:72.0%, D:3.4%], F:6.7%, M:17.9%
Source	This chapter	This chapter	Shared by the Cañestro laboratory (unpublished)	Danks et al., 2013

The transcriptome assemblies from Table 4.1 were aligned to genomes using BLAT v36 and used as "hints" to predict genes with AUGUSTUS v3.3.3 (Stanke et al., 2006). More specifically, the Okinawa transcriptome was used for gene prediction in OKI2018_I69 and KUM-M3-7f, whereas the Osaka one was used for annotating the OSKA2016v1.9 and AOM-5-5f genome assemblies. The Barcelona transcriptome was used to predict genes in the Bar2_p4 assembly. Gene structures in the Bergen (OdB3) genome were annotated with the aligned transcripts from the OikoBase (Danks et al., 2013). Moreover, protein sequences from other populations were aligned to genomes with Exonerate v2.2.0 (Slater and Birney, 2005) and used as additional evidence for gene prediction; for example, the Barcelona and Osaka *O. dioica* protein sequences were used for the OKI2018_I69 and KUM-M3-7f genome assemblies. For all genomes, gene prediction was performed based on the species model trained for the Okinawa *O. dioica* in chapter two. The resulting protein sequences were searched against the population-specific repeat libraries with the LAST "DNA-versus-proteins" pipeline v1238 (Yao and Frith, 2021), filtering out TEs that were still annotated as protein-coding genes. The quality of the final gene annotations was checked with BUSCO v3.0.2 (Simão et al., 2015). Finally, the gene orthology for *O. dioica* was reconstructed with OrthoFinder, using the workflow and the species list from chapter three (see section 3.2.4 in Material and Methods).

4.3 Results

I used a combination of comparative and *de novo* approaches to produce more comprehensive libraries of *O. dioica* repeats, which comprise up to 19% of the genomes. Further analysis of repeat structures is required to validate the sequences and classify them into families and subfamilies. Annotation of the newly masked genomes based on the population-specific transcript and cross-population protein alignments yielded fewer genes than the results presented in chapters two and three (see Table 2.6, Table 3.4). At the same time, the proportion of the "complete" BUSCO genes remained unchanged (~75% for all genomes; see Table 2.5), meaning that more comprehensive masking of repeat elements has likely not affected the core gene models. Nevertheless, the number of protein-coding genes varies depending on the population: from ~14,000-15,500 genes in the North Atlantic and North Pacific *O. dioica* and up to ~16,500 genes in the Ryukyu ones. Despite such a difference, the gene structures exhibit similar properties in terms of intron/exon and gene lengths and the fraction of non-canonical splice sites (non-GT/AG; 10-12%), which strongly agrees with the results presented by Denoeud et al. (2010) for the original annotation of the Bergen genome (also see Table 2.6). Similar to the results in chapter three, gene orthology reconstruction shows that a pair of *O. dioica* genomes shares from ~70% to ~90% of orthologous genes (Table 4.3). Between all of them, the highest proportion of single-copy resolved orthologs is reported for genomes from the same population (e.g., ~80% for the Okinawa-Kume pair), while the lowest is for the Ryukyu *O. dioica* and the rest (e.g., ~60% for the Okinawa-Osaka pair; Table 4.4). We can also see that updating the Bergen genome annotation significantly increased the number of one-to-one orthologs it shares with other *O. dioica*, especially with the Barcelona one. Unfortunately, there are still a lot of orthologous genes (~10%) with one-to-many or many-to-many relationships. Further analysis is required to understand whether there are real lineage-specific duplications and/or gene family expansions. However, some of the cases may still represent the annotation artifacts such as when genes in one genome get split or fused due to the variability of intergenic regions. I believe that the main challenge of getting the precise gene boundaries is related to the unique organization

of the *O. dioica* genome, which exhibits one of the highest densities of coding features reported to date: genes in *O. dioica* can be separated by as few as ~200 bases and have intronic sequences as short as ~30 nucleotides. Although I specifically trained AUGUSTUS on the *O. dioica* genome using *O. dioica*-specific transcriptome data, the sensitivity and accuracy of gene prediction is still not as high as for other organisms with more studied genome organization, such as fruit flies, humans or mice. Also, we have to keep in mind that ~50% of genes in *O. dioica* are co-transcribed as polycistronic pre-mRNAs, and, thus, may contribute to the same transcript assemblies, resulting in fused gene structures. Therefore, further analysis and more data is required in order to finalize annotation of genes in *O. dioica*. In sum, I can conclude that the current set of gene models is improved compared to the previous versions (see chapters two and three), given that artifacts such as TEs predicted as genes have been removed from the annotation. However, it clearly still leaves room for further improvement.

Table 4.2: Updated annotation of gene models in the six *O. dioica* genomes.

	Okinawa	Kume	Osaka	Aomori	Barcelona	Bergen
Genome id	OKI2018_I69	KUM-M3-7f	OSKA2016v1.9	AOM-5-5f	Bar2_p4	OdB3
Masked sequence (%)	19.27	18.69	18.95	19.09	18.95	15.76
Number of genes	16,460	16,513	15,301	14,848	14,111	16,430
Number of alternative transcripts	18,275	18,435	16,964	16,336	16,267	18,378
Median gene length (bp)	1,525	1,526	1,582	1,542	1,593	1,565
Median exon length (bp)	148	149	152	149	155	160
Median intron length (bp)	48	48	49	49	48	48
Non-canonical introns (%)	12.65	12.81	11.47	11.33	11.13	10.38
BUSCO scores (metazoa_odb9)	Complete: 76.9%, Fragmented: 4.3%	Complete: 77.8%, Fragmented: 4.1%	Complete: 72.7%, Fragmented: 6.1%	Complete: 76.0%, Fragmented: 4.7%	Complete: 77.5%, Fragmented: 4.6%	Complete: 75.7%, Fragmented: 4.9%

Table 4.3: Proportions of genes in orthogroups between pairs of *O. dioica* genomes.

	Number of genes	Barcelona	Bergen	Aomori	Osaka	Okinawa	Kume
Barcelona	14111		12214 (~87%)	11721 (~83%)	11850 (~84%)	11801 (~84%)	11844 (~84%)
Bergen	16430	13008 (~79%)		12730 (~77%)	12794 (~78%)	12747 (~78%)	12716 (~77%)
Aomori	14848	11829 (~80%)	11994 (~81%)		13066 (~88%)	12178 (~82%)	12200 (~82%)
Osaka	15301	12039 (~79%)	12202 (~80%)	13101 (~86%)		12363 (~81%)	12346 (~81%)
Okinawa	16460	12040 (~73%)	12165 (~74%)	12224 (~74%)	12271 (~75%)		14665 (~89%)
Kume	16513	12110 (~73%)	12165 (~74%)	12277 (~74%)	12339 (~75%)	14739 (~89%)	

Table 4.4: Proportions of genes with one-to-one orthologous relationships between pairs of *O. dioica* genomes.

	Number of genes	Barcelona	Bergen	Aomori	Osaka	Okinawa	Kume
Barcelona	14111		10719 (~76%)	10320 (~73%)	10334 (~73%)	9878 (~70%)	9942 (~70%)
Bergen	16430	10719 (~65%)		9813 (~60%)	9886 (~60%)	9383 (~57%)	9392 (~57%)
Aomori	14848	10320 (~70%)	9813 (~66%)		11615 (~78%)	9894 (~67%)	9917 (~67%)
Osaka	15301	10334 (~68%)	9886 (~65%)	11615 (~76%)		9833 (~64%)	9812 (~64%)
Okinawa	16460	9878 (~60%)	9383 (~57%)	9894 (~60%)	9833 (~60%)		12998 (~79%)
Kume	16513	9942 (~60%)	9392 (~57%)	9917 (~60%)	9812 (~59%)	12998 (~79%)	

4.4 Discussion and future work

The organization of the *O. dioica* genomes is unique compared to other chordates, and thus is quite challenging to annotate. Here, I improved the previous versions of gene models by applying a more exhaustive masking of the repeats. Also, I used cross-genome protein alignments as hints for gene prediction in order to generate more even annotations across populations. Unfortunately, a high density of coding sequences is still a problem for gene predictors, even after exhaustive training with the organism-specific data. Therefore, our next step is to apply CAGE (Cap Analysis of Gene Expression) data for accurate annotation of transcription start sites (TSS) and operon structures, and long Nanopore RNA sequencing that has the potential to provide more accurate detection of organism-specific isoforms and splice junctions, by sequencing transcript from start to end (Workman et al., 2019). Unfortunately, automatic gene annotation has limitations, especially when working with a genome of a non-model organism like *O. dioica*, which has short intergenic and intronic regions, operon structures, and non-canonical exon-intron boundaries. Therefore, manual curation of gene models in genome browser using evidence of transcript and protein alignments and coverage of RNA-Seq data is clearly warranted for further validation of the models.

Nevertheless, at this point our results clearly suggest that the number of protein-coding genes in *O. dioica* is population-specific and, thus, may explain the difference in genome size between populations that we observed in chapter three. Our next step is to analyze gene variations and family expansions specific to populations, in the context of their native marine environments. Moreover, in chapter three we showed that the genome of *O. dioica* is highly scrambled over separate populations, although morphology and developmental features of the animals remained virtually unchanged. We plan to combine the updated annotations, together with the RNA-seq data and open chromatin information (ATAC-seq) from different populations, to evaluate gene expression on multiple stages of the *O. dioica* development. We will also investigate the degree to which sequences of proximal regulatory elements (PRE) upstream of orthologous genes are conserved between populations. Altogether, these analyses may help to explain how gene expression is preserved in the context of genomic rearrangement and what mechanisms allow *O. dioica* to maintain its basic body structure and functions.

In this chapter, I also presented the more comprehensive annotation of repetitive elements in the *O. dioica* genomes, although further classification of them into families and subfamilies is still needed. I believe that TEs play an important role in the evolution of the *O. dioica* genome. By nature, TEs can act as a source of genomic variability, by moving through the genome and causing various genomic rearrangements that often lead to the diversification of species and populations (Warren et al., 2015). Moreover, mobile elements contribute to genome size variations in chordates (Chalopin et al., 2015), including larvaceans where accumulation of non-autonomous SINEs has driven multiple independent genome expansions (Naville et al. 2019). Thus, I am curious to see if some of the genome size variation between *O. dioica* populations could be explained by differences in diversity and distribution of various TE families. Moreover, given the high level of genome reshuffling that we observed between the *O. dioica* populations in chapter three, it is important to investigate how much of that is caused by the activity of TEs. Although proportion of interspersed repeats in the *O. dioica* genome is sparse compared to other animals, some transposons still show a low level of sequence corruption that suggest a rather recent activity of the elements in the genome (Volff et

al., 2004; Denoeud et al., 2010). Also, I observed that even despite masking the genome, some of the transposons still got identified as protein-coding genes, which is possible when a transposon is expressed and contributes to the transcriptomic data. Therefore, by looking at the transcriptome alignment and sequence identity of TE copies, I can identify transposons that are potentially active in the genome. With the complete assemblies of *O. dioica*'s chromosomes presented in this thesis, I can also study the distribution of various elements across them in order to understand if any repeat families played a specific role in the divergence of sex chromosomes. For example, a good candidate would be the *Oikopleura*-specific *Tor* elements whose insertions are 9× enriched in chromosome Y compared to the rest of the genome (Naville et al., 2019; Denoeud et al., 2010). Altogether, this research will contribute further insights into understanding the organization of the *O. dioica* genome and to which extent its features vary across geographically distant populations.

Chapter Five

Thesis conclusions

This thesis aimed to assess the diversity of *O. dioica* populations using a comparative genomics approach. In general, the genome sequence of *O. dioica* is highly polymorphic (Denoeud et al., 2010), making the assembly of its complete sequence challenging. Several attempts to sequence the *O. dioica* genome have been reported before, including a Sanger assembly for an *O. dioica* laboratory strain from the SARS Institute in Bergen (OdB3; Denoeud et al., 2010) and a more recent assembly with PacBio and Illumina reads for mainland Japanese (Osaka area) laboratory strain (OSKA2016; Wang et al., 2020). Both assemblies have full genome coverage and high sequence accuracy, but lack chromosomal resolution.

Chapter two took a hybrid approach to produce a first chromosome-scale genome assembly of an *O. dioica* individual from Okinawa (OKI2018_I69). To reduce variation in the data, we extracted high-molecular-weight DNA from a single *O. dioica* male, which we further sequenced with Oxford Nanopore and Illumina MiSeq technologies. We used long-range chromatin conformation data to enable chromosomal resolution. 99% of the resulting assembly is contained within five megabase-scale scaffolds, that represent telomere-to-telomere sequences of two autosomes and sex chromosomes split into pseudo-autosomal region (PAR) and X-specific or Y-specific regions. The assembled chromosomes mostly align to corresponding linkage groups predicted for the Bergen *O. dioica* (Denoeud et al., 2010), suggesting the conservation of three chromosome pairs between populations. We confirmed this result by karyotyping the Okinawa population using antibody staining (Liu et al., 2020).

The chromosomal resolution of our genome assembly allowed us to make several observations that will be of interest beyond the field of *Oikopleura* research. First, the Hi-C contact matrix shows that arms within individual chromosomes have only a few reciprocal interactions, similar to the "type-I" genome architecture reported by Hoencamp et al. (2021). Second, there are arm-specific differences in repeat and GC content, protein-coding features and dN/dS values, suggesting that chromosome arms in *O. dioica* may evolve at different rates. To our knowledge, this has not been observed before for other chordate genomes.

Chapter three gives an overview of the genomic diversity of three *O. dioica* populations from globally distributed locations: one from North Atlantic (Barcelona/Bergen) and two from Pacific (Osaka/Aomori and Okinawa/Kume) Oceans. Each population in our analysis is represented by one chromosome-level (Barcelona, Osaka, Okinawa) and one contig level (Bergen, Aomori, Kume) assembly, making up a dataset of six *O. dioica* genomes.

Whole-genome alignments of *O. dioica* populations revealed a striking degree of sequence rearrangements that is by far unprecedented in metazoa. These rearrangements are mostly restricted to homologous chromosome arms and sex-specific regions, which appear to represent the primary unit of synteny in *O. dioica*. The smallest units of synteny seem to be genes and their constituent exons. However, even their structures are not entirely protected from the mutational processes that produce scrambled genomes, given that some genes and exons still overlap breakpoint regions. Further, gene orthology reconstruction revealed significant variation in the number of orthologous genes shared between populations. The order of genes has been effectively randomized, affecting even evolutionary conserved clusters.

Operon structures are less conserved than genes, since only half of them overlap syntenic regions between Okinawa and Osaka genomes. Indeed, most of the operons in Okinawa and Osaka are population-specific. We also found that gene movements between operons seem possible in *O. dioica* – something that has not been reported before for other animals with operon structures. We believe that the existence of operons in *O. dioica* may help to retain gene expression in the context of scrambling by allowing a gene to be inserted into an operon with the already existing transcriptional machinery. However, there might be an extent to which a gene can be co-expressed within operons, with a trend towards genes with house-keeping and metabolic functions. Additional data such as CAGE (Cap Analysis of Gene Expression) and long RNA-Seq is required to validate the results and confirm the functions of operons in *O. dioica*.

Our research shows that the level of scrambling in *O. dioica* varies across populations on both nucleotide and gene level, with the lowest difference between genomes from the same populations (e.g., Osaka-Aomori) and the highest between the Ryukyu genomes and the rest (e.g. Okinawa-Osaka). Therefore, the Ryukyu population seems to possess the most scrambled genome, representing an outgroup to the North Pacific and North Atlantic populations. We compared multiple phylogenetic markers, such as nuclear 18S and ITS (internal transcribed spacers), and mitochondrial COI (cytochrome oxidase I), proving that the Ryukyu, North Pacific and North Atlantic specimens comprise three genetically distinct lineages that are currently considered conspecific (Masunaga et al., 2022). In addition to that, Aki Masunaga investigated the fertilization success of Okinawa and Osaka lab strains, showing the lack of interbreeding between the two – one of the most common criteria for defining species. On the other hand, the overall morphology of *O. dioica* specimens is indistinguishable between Okinawa, Osaka and Barcelona. All lineages have separate sexes and two subchordal cells on their tails – two features that clearly distinguish *O. dioica* from other larvaceans. Close examination of the samples revealed only minor differences in the trunk-tail ratios, egg diameter and nuclei shape of oikoplastic epithelium. Altogether, results suggest that the current taxonomic *O. dioica* hides multiple cryptic species (Masunaga et al., 2022). Here, cryptic species are defined as "two or more distinct species that are erroneously classified (and hidden) under one species name" because of their superficially indistinguishable morphology (Bickford et al., 2007). Cryptic speciation is a common phenomenon in many cosmopolitan marine taxa (Borges et al., 2022). Further studies are required to establish markers that can reliably distinguish lineages of dioecious *Oikopleura* in the field. Approaches targeting environmental DNA (eDNA) using ITS primers may provide enough molecular data to draw clear boundaries in geographical distribution of the lineages.

Michael Mansfield estimated divergence time of *O. dioica* populations using single-copy resolved orthologous genes from chapter two. The analysis showed that the three *O. dioica* populations shared a common ancestor as recently as ~20-30 Mya, with the Ryukyu one first to diverge. The North Atlantic and North Pacific populations have split less than 10 Mya. How these lineages diverged and why in this order remain unknown. It is possible that the extreme level of genome rearrangements in *O. dioica* could promote sympatric speciation by creating reproductively incompatible subpopulations within marine environments that lack clear geographical boundaries. Thus, further studies of genetic diversity within the lineages are needed to validate this hypothesis.

We can also explain the existence of reproductive isolation between Okinawa and Osaka *Oikopleura* by the Kuroshio – a fast and strong ocean current that flows north from Taipei towards mainland Japan on the West side of the Pacific through the Ryukyu archipelago

(Fig. 3.1). Kuroshio transports many tropical and subtropical species (Saito, 2019). However, it can also act as a potential geographical barrier and limit gene flow between the Ryukyu archipelago and mainland Japan, promoting lineage diversification in marine organisms (Kojima et al., 2000). More sampling from both sides of the Kuroshio current is required to confirm whether this may be the case in *O. dioica*. Further, there is a significant difference in the average annual temperature of the sea surface near the sampling locations of Okinawa (~25 ℃; Kin bay) and Osaka *O. dioica* (~19 ℃; Sakoshi bay; Masunaga et al., 2022). Depending on the temperature, the generation time of *O. dioica* can be as short as one day at 28 °C or 16 days at 10 °C (Uye and Ichino, 1995). Thus, the shorter life cycle of Okinawan *O. dioica* may limit long-distance dispersal and accelerate the accumulation of rearrangements in the population, resulting in the most scrambled genome.

Overall, this thesis contributes new insights into the evolution of basal chordate genomes and provides first evidence that *O. dioica* might be hiding more genetic diversity than has been suspected before. Follow-up comparisons, especially using gene annotations, should further assess genomic differences in these lineages and uncover lineage-specific gene variations that could shed light on possible adaptations to certain marine environments. If the high genomic divergence demonstrated here leads to the recognition of *O. dioica* as multiple separate species by the scientific community, the chromosomal sequences with complete gene and repeat annotations presented in this thesis may serve as references for each of the three species. Moreover, we expect that the pool of natural *Oikopleura* diversity has not been exhaustively studied yet as there is no genomic data of dioecious oikopleurids from the Southern hemisphere.

Our study design allowed us to examine the genome scrambling phenomenon at a closer evolutionary distance than previous studies on other organisms (Drosophila 12 Genomes Consortium et al., 2007; Albertin et al., 2022; Hane et al., 2011). To our knowledge, *O. dioica* possesses the fastest-rearranging genome described so far. Further studies are needed to understand the way the loss of the c-NHEJ DNA repair machinery has impacted the evolution of these organisms. A functional genomic approach using the chromatin status and transcriptome data from different lineages may uncover unknown regulatory mechanisms that allow *O. dioica* to maintain classical chordate morphology and developmental features despite pronounced genomic divergence.

Considering *O. dioica*'s chordate nature and uniqueness of its genome structure, along with the high level of within-species diversity demonstrated in this thesis, further study of this organism is clearly warranted. *O. dioica* appears as an animal with considerable promise in cross-disciplinary research ranging from basic evolutionary developmental (evo-devo) studies to ecology and biomedicine. A privileged phylogenetic position of tunicates as a sister clade to vertebrates makes *O. dioica* more closely related to humans than worms or fruit flies, but more genetically traceable than mice or zebrafish. It can be easier and much faster cultured in the laboratory for many generations (Bouquet et al., 2009; Martí-Solans et al., 2015; Masunaga et al., 2020) and used for genetic manipulations, such as different knockdown approaches for altering gene expression (Sagane et al., 2010; Omotezako et al., 2013, 2015, 2017; Mikhaleva et al., 2015), functional screening (Onuma and Nishida, 2021) and genome editing based on CRISPR-Cas9 (Deng et al., 2018). *O. dioica* possess organs, tissues, and structures that are unequivocally homologous to those in vertebrates (Cañestro et al. 2005, 2008; Nishida, 2008; Ferrández-Roldán et al., 2021) and, thus, can be used in biomedical research as a good proxy for understanding the genetic basis of various disorders. Moreover, we showed that the miniature genome of *O. dioica* can be affordably sequenced in many individuals, facilitating

population genomics studies for understanding the effect evolutionary forces have on natural populations. Like other zooplankton, *O. dioica* is highly sensitive to changes in the ecosystem, responding quickly to seasonal variations in water temperature, nutrient balance, and ocean currents. Therefore, tracking *O. dioica* populations has been proposed as valuable means for following and assessing the ecological health and integrity of marine systems, and understanding the impact of climate change on marine food webs, nutrient cycles, and ocean production (Troedsson et al., 2013; Bouquet et al., 2018; Torres-Águila et al., 2018). Moreover, larvaceans are potential key links in the biomagnification of industrial pollutants: a recent study revealed how giant larvaceans could serve as efficient vectors of microplastics using their filter-feeding systems (Katija et al., 2017). All combined, we believe that *O. dioica* research already has a fascinating present, but an even more exciting future that will only benefit with the increase of genomic data from local populations around the globe.

References

Albalat R., Martí-Solans J., Cañestro C. (2012). DNA methylation in amphioxus: from ancestral functions to new roles in vertebrates. *Briefings in Functional Genomics, 11*(2), 142–155. https://doi.org/10.1093/bfgp/els009

Albertin C.B., Medina-Ruiz S., Mitros T., Schmidbaur H., Sanchez G., Wang Z.Y., Grimwood J., Rosenthal J.J. C., Ragsdale C.W., Simakov O., Rokhsar D.S. (2022). Genome and transcriptome mechanisms driving cephalopod evolution. *Nature Communications, 13*(1), 2427. https://doi.org/10.1038/s41467-022-29748-w

Alldredge A. (1976). Appendicularians. *Scientific American, 235*(1), 94–105. http://www.jstor.org/stable/24950396

Alldredge A.L. (1976) Discarded appendicularian houses as sources of food, surface habitats, and particulate organic matter in planktonic environments. *Limnol Oceanogr., 21*(1):14–24. https://doi.org/10.4319/lo.1976.21.1.0014

Alldredge A. (2005). The contribution of discarded appendicularian houses to the flux of particulate organic carbon from oceanic surface waters. In: Gorsky G, Youngbluth M.J, Deibel D, editors. Response of Marine Ecosystems to Global Change: Ecological Impact of Appendicularians: Contemporaty Publishing International. p. 309–26.

Appeltans W., Ahyong S.T., Anderson G., Angel M.V., Artois T., Bailly N., Bamber R., Barber A., Bartsch I., Berta A., Błażewicz-Paszkowycz M., Bock P., Boxshall G., Boyko C.B., Brandão S.N., Bray R.A., Bruce N.L., Cairns S.D., Chan T.-Y., … Costello M.J. (2012). The magnitude of global marine species diversity. *Current Biology, 22*(23), 2189–2202. https://doi.org/10.1016/j.cub.2012.09.036

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data.

Balavoine G., de Rosa R., Adoutte A. (2002). Hox clusters and bilaterian phylogeny. *Molecular Phylogenetics and Evolution, 24*(3), 366–373. https://doi.org/10.1016/s1055-7903(02)00237-3

Bao W., Kojima K.K., Kohany O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA, 6*, 11. https://doi.org/10.1186/s13100-015-0041-9

Berná L., Alvarez-Valin F. (2014). Evolutionary genomics of fast evolving tunicates. *Genome Biology and Evolution, 6*(7), 1724–1738. https://doi.org/10.1093/gbe/evu122

Berná L., Alvarez-Valin F. (2015). Evolutionary volatile Cysteines and protein disorder in the fast evolving tunicate *Oikopleura dioica*. *Marine Genomics, 24*, 47–54. https://doi.org/10.1016/j.margen.2015.07.007

Berná L., Alvarez-Valin F., D'Onofrio G. (2009). How fast is the sessile ciona? *Comparative and Functional Genomics,* 875901. https://doi.org/10.1155/2009/875901

Berná L., D'Onofrio G., Alvarez-Valin F. (2012). Peculiar patterns of amino acid substitution and conservation in the fast evolving tunicate *Oikopleura dioica*. *Molecular Phylogenetics and Evolution, 62*(2), 708–717. https://doi.org/10.1016/j.ympev.2011.11.013

Bernt M., Donath A., Jühling F., Externbrink F., Florentz C., Fritzsch G., Pütz J., Middendorf M., Stadler P.F. (2013). MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution, 69*(2), 313–319. https://doi.org/10.1016/j.ympev.2012.08.023

Bickford D., Lohman D.J., Sodhi N.S., Ng P.K., Meier R., Winker K., Ingram K.K. and Das I. (2007). Cryptic species as a window on diversity and conservation. *Trends in ecology & evolution, 22*(3), pp.148-155. https://doi.org/10.1016/j.tree.2006.11.004

Blanchoud S., Rutherford K., Zondag L., Gemmell N.J., & Wilson M.J. (2018). De novo draft assembly of the *Botrylloides leachii* genome provides further insight into tunicate evolution. *Scientific Reports*, *8*(1), 5518. https://doi.org/10.1038/s41598-018-23749-w

Bliznina A., Masunaga A., Mansfield M.J., Tan Y., Liu A.W., West C., Rustagi T., Chien H.C., Kumar S., Pichon J., Plessy C., Luscombe N.M. (2021). Telomere-to-telomere assembly of the genome of an individual *Oikopleura dioica* from Okinawa using Nanopore-based sequencing. *BMC genomics, 22*(1): 1-18. doi:10.1186/s12864-021-07512-6

Blumenthal T., Evans D., Link C.D., Guffanti A., Lawson D., Thierry-Mieg J., Thierry-Mieg D., Chiu W.L., Duke K., Kiraly M., Kim S.K. (2002). A global analysis of *Caenorhabditis elegans* operons. *Nature, 417*(6891), 851–854. https://doi.org/10.1038/nature00831

Borges L.M., Treneman N.C., Haga T., Shipway J.R., Raupach M.J., Altermark B. and Carlton J.T. (2022). Out of taxonomic crypsis: A new trans-arctic cryptic species pair corroborated by phylogenetics and molecular evidence. *Molecular Phylogenetics and Evolution,* 166, 107312. https://doi.org/10.1016/j.ympev.2021.107312

Bouquet J.-M., Spriet E., Troedsson C., Otterå H., Chourrout D., Thompson E.M. (2009). Culture optimization for the emergent zooplanktonic model organism *Oikopleura dioica*. *Journal of Plankton Research, 31*(4), 359–370. https://doi.org/10.1093/plankt/fbn132

Bouquet J.-M., Troedsson C., Novac A., Reeve M., Lechtenbörger A.K., Massart W., Skaar K.S., Aasjord A., Dupont S. and Thompson E.M., 2018. Increased fitness of a key appendicularian zooplankton species under warmer, acidified seawater conditions. *PLoS One, 13*(1), p.e0190625. https://doi.org/10.1371/journal.pone.0190625

Bourlat S.J., Juliusdottir T., Lowe C.J., Freeman R., Aronowicz J., Kirschner M., Lander E.S., Thorndyke M., Nakano H., Kohn A.B., Heyland A., Moroz L.L., Copley R.R., Telford M.J. (2006). Deuterostome phylogeny reveals monophyletic chordates and the new phylum *Xenoturbellida. Nature, 444*(7115), 85–88. https://doi.org/10.1038/nature05241

Bowden R., Davies R.W., Heger A., Pagnamenta A.T., de Cesare M., Oikkonen L.E., Parkes D., Freeman C., Dhalla F., Patel S.Y., Popitsch N., Ip C.L.C., Roberts H.E., Salatino S., Lockstone H., Lunter G., Taylor J.C., Buck D., Simpson M.A., Donnelly P. (2019). Sequencing of human genomes with nanopore technology. *Nature Communications, 10*(1), 1869. https://doi.org/10.1038/s41467-019-09637-5

Brunetti R., Gissi C., Pennati R., Caicci F., Gasparini F., Manni L. (2015). Morphological evidence that the molecularly determined Ciona intestinalis type A and type B are different species: *Ciona robusta* and *Ciona intestinalis*. *Journal of Zoological Systematics and Evolutionary Research, 53*(3), 186–193. https://doi.org/10.1111/jzs.12101

Burighel P., Brena C., Martinucci G.B., Cima F. (2001). Gut ultrastructure of the appendicularian *Oikopleura dioica* (Tunicata). *Invertebrate Biology: A Quarterly Journal of the American Microscopical Society and the Division of Invertebrate Zoology/ASZ, 120*(3), 278–293. https://doi.org/10.1111/j.1744-7410.2001.tb00038.x

Bushmanova E., Antipov D., Lapidus A., Suvorov V. (2016). rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. *Bioinformatics, 32*(14):2210–2. https://doi.org/10.1093/bioinformatics/btw218.

Cañestro C., Bassham S., Postlethwait J. (2005). Development of the central nervous system in the larvacean *Oikopleura dioica* and the evolution of the chordate brain. *Developmental Biology, 285*(2), 298–315. https://doi.org/10.1016/j.ydbio.2005.06.039

Cañestro C., Bassham S. and Postlethwait J.H. (2008). Evolution of the thyroid: anterior-posterior regionalization of the Oikopleura endostyle revealed by Otx, Pax2/5/8, and Hox1 expression. *Developmental dynamics: an official publication of the American Association of Anatomists, 237*(5), 1490–1499. doi:10.1002/dvdy.21525.

Cañestro C., Postlethwait J.H., Gonzàlez-Duarte R., Albalat R. (2006). Is retinoic acid genetic machinery a chordate innovation? *Evolution & Development, 8*(5), 394–406. https://doi.org/10.1111/j.1525-142X.2006.00113.x

Caputi L., Andreakis N., Mastrototaro F., Cirino P., Vassillo M., Sordino P. (2007). Cryptic speciation in a model invertebrate chordate. *Proceedings of the National Academy of Sciences of the United States of America, 104*(22), 9364–9369. https://doi.org/10.1073/pnas.0610158104

Carroll S.B. (1995). Homeotic genes and the evolution of arthropods and chordates. *Nature, 376*(6540), 479–485. https://doi.org/10.1038/376479a0

Castellani C., Edwards M. (2017). Marine Plankton: A practical guide to ecology, methodology, and taxonomy. *Oxford University Press*. https://play.google.com/store/books/details?id=l3QzDwAAQBAJ

Chalopin D., Naville M., Plard F., Galiana D., Volff J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution, 7*(2), 567–580. https://doi.org/10.1093/gbe/evv005

Colombera D., Fenaux R. (1973). Chromosome form and number in the Larvacea. *Ital J Zool.*, *40*(3-4), 347–353. https://doi.org/10.1080/11250007309429248

Danks G.B., Raasholm M., Campsteijn C., Long A.M., Manak J.R., Lenhard B., Thompson E.M. (2015). Trans-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Molecular Biology and Evolution, 32*(3), 585–599. https://doi.org/10.1093/molbev/msu336

Danks G., Campsteijn C., Parida M., Butcher S., Doddapaneni H., Fu B., Petrin R., Metpally R., Lenhard B., Wincker P., Chourrout D., Thompson E.M., Manak J.R. (2013). OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica. Nucleic Acids Research, 41*(Database issue), D845–D853. https://doi.org/10.1093/nar/gks1159

Davis R.E., Hodgson S. (1997). Gene linkage and steady state RNAs suggest trans-splicing may be associated with a polycistronic transcript in *Schistosoma mansoni*. *Molecular and Biochemical Parasitology, 89*(1), 25–39. https://doi.org/10.1016/s0166-6851(97)00097-2

Dehal P., Satou Y., Campbell R.K., Chapman J., Degnan B., De Tomaso A., Davidson B., Di Gregorio A., Gelpke M., Goodstein D.M., Harafuji N., Hastings K.E.M., Ho I., Hotta K., Huang W., Kawashima T., Lemaire P., Martinez D., Meinertzhagen I.A., … Rokhsar D.S. (2002). The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science, 298*(5601), 2157–2167. https://doi.org/10.1126/science.1080049

Delsuc F., Brinkmann H., Chourrout D., Philippe H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, *439*(7079), 965–968. https://doi.org/10.1038/nature04336

Delsuc F., Philippe H., Tsagkogeorga G., Simion P., Tilak M.-K., Turon X., López-Legentil S., Piette J., Lemaire P., Douzery, E.J.P. (2018). A phylogenomic framework and timescale for comparative studies of tunicates. *BMC Biology, 16*(1), 39. https://doi.org/10.1186/s12915-018-0499-2

Delsuc F., Tsagkogeorga G., Lartillot N., Philippe H. (2008). Additional molecular support for the new chordate phylogeny. *Genesis, 46*(11), 592–604. https://doi.org/10.1002/dvg.20450

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*(20), 3337–3341.e4. https://doi.org/10.1016/j.cub.2018.08.048

Denoeud F., Henriet S., Mungpakdee S., Aury J.-M., Da Silva, C., Brinkmann H., Mikhaleva J., Olsen L.C., Jubin C., Cañestro C., Bouquet J.-M., Danks G., Poulain J., Campsteijn C., Adamski M., Cross I., Yadetie F., Muffato M., Louis A., … Chourrout D. (2010). Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science*, *330*(6009), 1381–1385. https://doi.org/10.1126/science.1194167

Di Tommaso P., Chatzou M., Floden E.W., Barja P.P., Palumbo E., Notredame C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology, 35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Dobin A., Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Drosophila 12 Genomes Consortium, Clark A.G., Eisen M.B., Smith D.R., Bergman C.M., Oliver B., Markow T.A., Kaufman T.C., Kellis M., Gelbart W., Iyer V.N., Pollard D.A., Sackton T.B., Larracuente A.M., Singh N.D., Abad J.P., Abt D.N., Adryan B., Aguade M., … MacCallum I. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, *450*(7167), 203–218. https://doi.org/10.1038/nature06341

Dudchenko O., Batra S.S., Omer A.D., Nyquist S.K., Hoeger M., Durand N.C., Shamim M.S., Machol I., Lander E.S., Aiden A.P., Aiden E.L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95. https://doi.org/10.1126/science.aal3327

Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature, 452*(7188), 745–749. https://doi.org/10.1038/nature06614

Durand N.C., Robinson J.T., Shamim M.S., Machol I., Mesirov J.P., Lander E.S., Aiden E.L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems*, *3*(1), 99–101. https://doi.org/10.1016/j.cels.2015.07.012

Durand N.C., Shamim M.S., Machol I., Rao S.S.P., Huntley M.H., Lander E.S., Aiden E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems*, *3*(1), 95–98. https://doi.org/10.1016/j.cels.2016.07.002

Edvardsen R.B., Seo H.-C., Jensen M.F., Mialon A., Mikhaleva J., Bjordal M., Cartry J., Reinhardt R., Weissenbach J., Wincker P., Chourrout D. (2005). Remodelling of the homeobox gene complement in the tunicate *Oikopleura dioica*. *Current Biology, 15*(1), R12–R13. https://doi.org/10.1016/j.cub.2004.12.010

Emms D.M., Kelly S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology, 16*, 157. https://doi.org/10.1186/s13059-015-0721-2

Emms D.M., Kelly S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology, 20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Falcon S., Gentleman R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics, 23*(2), 257–258. https://doi.org/10.1093/bioinformatics/btl567

Fenaux R. (1998) Anatomy and functional morphology of the Appendicularia. In: Bone Q, editor. The biology of pelagic tunicates: Oxford University Press. p. 25–34.

Fenaux R., Bone Q., Deibel D. (1998). Appendicularian distribution and zoogeography. In *The biology of pelagic tunicates* (ed. Q. Bone), pp.251-264. Oxford University Press, New York.

Fenaux R. (1986). The house of *Oikopleura dioica* (Tunicata, Appendicularia): structure and functions. *Zoomorphology 106*(4), pp.224–31.

Ferguson D.O., Sekiguchi J.M., Chang S., Frank K.M., Gao Y., DePinho R.A., Alt F.W. (2000). The nonhomologous end-joining pathway of DNA repair is required for genomic stability and the suppression of translocations. *Proceedings of the National Academy of Sciences of the United States of America, 97*(12), 6630–6633. https://doi.org/10.1073/pnas.110152897

Ferrández-Roldán A., Fabregà-Torrus M., Sánchez-Serna G., Duran-Bello E., Joaquín-Lluís M., Bujosa P., Plana-Carmona M., Garcia-Fernàndez J., Albalat R., Cañestro C. (2021). Cardiopharyngeal deconstruction and ancestral tunicate sessility. *Nature, 599*(7885), 431–435. https://doi.org/10.1038/s41586-021-04041-w

Ferrández-Roldán A., Martí-Solans J., Cañestro C., Albalat R. (2019). *Oikopleura dioica*: An Emergent Chordate Model to Study the Impact of Gene Loss on the Evolution of the Mechanisms of Development. In W. Tworzydlo & S. M. Bilinski (Eds.), *Evo-Devo: Non-model Species in Cell and Developmental Biology* (pp. 63–105). Springer International Publishing. https://doi.org/10.1007/978-3-030-23459-1_4

Flynn J.M., Hubley R., Goubert C., Rosen J., Clark A.G., Feschotte C., Smit A.F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America, 117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Flood P. (2005). Toward a photographic atlas on special taxonomic characters of oikopleurid Appendicularia (Tunicata). *Response of marine ecosystems to global change: ecological impact of appendicularians. Paris: Contemporary Publishing International*, 59-85

Flood P., Deibel D. (1998). The appendicularian house. In "The Biology of Pelagic Tunicates" (Q. Bone, Ed.), 105-124.

Fredriksson G., Olsson R. (1991). The subchordal cells of *Oikopleura dioica* and *O. albicans* (Appendicularia, Chordata). *Acta Zoologica*, *72*(4), 251–256. https://doi.org/10.1111/j.1463-6395.1991.tb01203.x

Frith M.C. (2011). A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Research*, *39*(4), e23. https://doi.org/10.1093/nar/gkq1212

Frith M.C., Kawaguchi R. (2015). Split-alignment of genomes finds orthologies more accurately. *Genome Biology*, *16*, 106. https://doi.org/10.1186/s13059-015-0670-9

Ganot P., Kallesøe T., Reinhardt R., Chourrout D., Thompson E.M. (2004). Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. *Molecular and Cellular Biology*, *24*(17), 7795–7805. https://doi.org/10.1128/MCB.24.17.7795-7805.2004

Ganot P., Thompson E.M. (2002). Patterning through differential endoreduplication in epithelial organogenesis of the chordate, *Oikopleura dioica*. *Developmental Biology*, *252*(1), 59–71. https://doi.org/10.1006/dbio.2002.0834

Gertz E.M., Yu Y.-K., Agarwala R., Schäffer A.A., Altschul S.F. (2006). Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biology*, *4*, 41. https://doi.org/10.1186/1741-7007-4-41

Glover J.C., Fritzsch B. (2009). Brains of primitive chordates. *Encyclopedia of Neurosciences*, 439-448.

Gordon A., Hannon G.J. (2010) Fastx-toolkit. FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab.cshl.edu/fastx_toolkit/.

Gorsky G., Fenaux, R. (1998). The role of Appendicularia in marine food webs. In The biology of pelagic tunicates (ed. Q. Bone), pp.161-170. Oxford University Press, New York.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., … Regev A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Grohme M.A., Schloissnig S., Rozanski A., Pippel M., Young G.R., Winkler S., Brandl H., Henry I., Dahl A., Powell S., Hiller M., Myers E., Rink J.C. (2018). The genome of *Schmidtea mediterranea* and the evolution of core cellular mechanisms. Nature, 554(7690), 56–61. https://doi.org/10.1038/nature25473

Guan D., McCarthy S.A., Wood J., Howe K., Wang Y., Durbin R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics, 36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

Gurevich A., Saveliev V., Vyahhi N., Tesler G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Haas B.J., Delcher A.L., Mount S.M., Wortman J.R., Smith R.K., Jr, Hannick L.I., Maiti R., Ronning C.M., Rusch D.B., Town C.D., Salzberg S.L., White O. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*(19), 5654–5666. https://doi.org/10.1093/nar/gkg770

Hamada M., Ono Y., Asai K., Frith M.C. (2017). Training alignment parameters for arbitrary sequencers with LAST-TRAIN. *Bioinformatics*, *33*(6), 926–928. https://doi.org/10.1093/bioinformatics/btw742

Hamner W.M., Robison B.H. (1992). In situ observations of giant appendicularians in Monterey Bay. *Deep-Sea Research. Part A, Oceanographic Research Papers, 39*(7), 1299–1313. https://doi.org/10.1016/0198-0149(92)90070-A

Hane J.K., Rouxel T., Howlett B.J., Kema G.H.J., Goodwin S.B., Oliver R.P. (2011). A novel mode of chromosomal evolution peculiar to filamentous *Ascomycete fungi*. *Genome Biology, 12*(5), R45. https://doi.org/10.1186/gb-2011-12-5-r45

Han Y., Wessler S.R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, *38*(22), e199. https://doi.org/10.1093/nar/gkq862

Henriet S., Colom Sanmartí B., Sumic S., Chourrout D. (2019). Evolution of the U2 spliceosome for processing numerous and highly diverse non-canonical introns in the chordate *Fritillaria borealis*. *Current Biology, 29*(19), 3193–3199.e4. https://doi.org/10.1016/j.cub.2019.07.092

Hill M.M., Broman K.W., Stupka E., Smith W.C., Jiang D., Sidow A. (2008). The *C. savignyi* genetic map and its integration with the reference sequence facilitates insights into chordate genome evolution. *Genome Research, 18*(8), 1369–1379. https://doi.org/10.1101/gr.078576.108

Hirose E., Kimura S., Itoh T., Nishikawa J. (1999). Tunic morphology and cellulosic components of pyrosomas, doliolids, and salps (thaliacea, urochordata). *The Biological Bulletin, 196*(1), 113–120. https://doi.org/10.2307/1543173

Hoencamp C., Dudchenko O., Elbatsh A.M.O., Brahmachari S., Raaijmakers J.A., van Schaik T., Sedeño Cacciatore Á., Contessoto V.G., van Heesbeen R.G.H.P., van den Broek B., Mhaskar A.N., Teunissen H., St Hilaire B.G., Weisz D., Omer A.D., Pham M., Colaric Z., Yang Z., Rao S.S.P., … Rowland B.D. (2021). 3D genomics across the tree of life reveals condensin

II as a determinant of architecture type. *Science, 372*(6545), 984–989. https://doi.org/10.1126/science.abe2218

Hoff K.J, Stanke M. (2019) Predicting genes in single genomes with augustus. *Curr Protoc Bioinformatics*, *65*(1), e57. https://doi.org/10.1002/cpbi.57

Hopcroft R.R., Clarke C., Nelson R.J., Raskoff K.A. (2005). Zooplankton communities of the Arctic's Canada Basin: the contribution by smaller taxa. *Polar Biology, 28*(3), 198–206. https://doi.org/10.1007/s00300-004-0680-7

Hopcroft R.R., Roff J.C. (1995). Zooplankton growth rates: extraordinary production by the larvacean *Oikopleura dioica* in tropical waters. *Journal of Plankton Research*, *17*(2), 205–220. https://doi.org/10.1093/plankt/17.2.205

Hosp J., Sagane Y., Danks G., Thompson E.M. (2012). The evolving proteome of a complex extracellular matrix, the Oikopleura house. *PloS One, 7*(7), e40172. https://doi.org/10.1371/journal.pone.0040172

Huang S., Kang M., Xu A. (2017). HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics*, *33*(16), 2577–2579. https://doi.org/10.1093/bioinformatics/btx220

Iannelli F., Pesole G., Sordino P., Gissi C. (2007). Mitogenomics reveals two cryptic species in *Ciona intestinalis*. Trends in Genetics: TIG, 23(9), 419–422. https://doi.org/10.1016/j.tig.2007.07.001

Johnson D.S., Davidson B., Brown C.D., Smith W.C., Sidow A. (2004). Noncoding regulatory sequences of *Ciona* exhibit strong correspondence between evolutionary constraint and functional importance. *Genome Research, 14*(12), 2448–2456. https://doi.org/10.1101/gr.2964504

Jones P., Binns D., Chang H.-Y., Fraser M., Li W., McAnulla C., McWilliam H., Maslen J., Mitchell A., Nuka G., Pesseat S., Quinn A.F., Sangrador-Vegas A., Scheremetjew M., Yong S.-Y., Lopez R., Hunter S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Katija K., Choy C.A., Sherlock R.E., Sherman A.D., Robison B.H. (2017). From the surface to the seafloor: How giant larvaceans transport microplastics into the deep sea. *Science Advances, 3*(8), e1700715. https://doi.org/10.1126/sciadv.1700715

Kiełbasa S.M., Wan R., Sato K., Horton P., Frith M.C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493. https://doi.org/10.1101/gr.113985.110

Kienle N., Kloepper T.H., Fasshauer D. (2016). Shedding light on the expansion and diversification of the Cdc48 protein family during the rise of the eukaryotic cell. *BMC Evolutionary Biology*, *16*(1), 215. https://doi.org/10.1186/s12862-016-0790-1

Kocot K.M., Tassia M.G., Halanych K.M., Swalla B.J. (2018). Phylogenomics offers resolution of major tunicate relationships. *Molecular Phylogenetics and Evolution, 121*, 166–173. https://doi.org/10.1016/j.ympev.2018.01.005

Kojima S., Segawa R., Hayashi I. (2000). Stability of the courses of the warm coastal currents along the Kyushu island suggested by the population structure of the Japanese turban shell, Turbo (Batillus) Cornutus. *Journal of Oceanography, 56*(5), 601–604. https://doi.org/10.1023/A:1011113430343

Kolmogorov M., Yuan J., Lin Y., Pevzner P.A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8

Koren S., Walenz B.P., Berlin K., Miller J.R., Bergman N.H., Phillippy A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Körner W.F. (1952) Untersuchungen über die gehäusebildung bei appendicularien (*Oikopleura dioica* fol). *Z Morphol Okol Tiere, 41*(1), 1–53. https://doi.org/10.1007/BF00407623

Laehnemann D., Borkhardt A., McHardy A.C. (2015). Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Briefings in Bioinformatics, 17*(1), 154–179. https://doi.org/10.1093/bib/bbv029

Lagesen K., Hallin P., Rødland E.A., Staerfeldt H.-H., Rognes T., Ussery D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108. https://doi.org/10.1093/nar/gkm160

Lawrence M., Huber W., Pagès H., Aboyoun P., Carlson M., Gentleman R., Morgan M.T., Carey V.J. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology, 9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Lemaire P., Smith W.C., Nishida H. (2008). Ascidians and the plasticity of the chordate developmental program. *Current Biology, 18*(14), R620–R631. https://doi.org/10.1016/j.cub.2008.05.039

Lieberman-Aiden E., van Berkum N. L., Williams L., Imakaev M., Ragoczy T., Telling A., Amit I., Lajoie B.R., Sabo P.J., Dorschner M.O., Sandstrom R., Bernstein B., Bender M.A., Groudine M., Gnirke A., Stamatoyannopoulos J., Mirny L.A., Lander E.S., Dekker J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. preprint arXiv, 1303:3997.

Liu A.W., Tan Y., Masunaga A., Bliznina A., West C., Plessy C., Luscombe N.M. (2020). H3S28P antibody staining of Okinawan *Oikopleura dioica* suggests the presence of three chromosomes. *F1000Research*, *9*, 780. https://doi.org/10.12688/f1000research.25019.2

Li W., Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Llorens C., Futami R., Covelli L., Domínguez-Escribá L., Viu J. M., Tamarit D., Aguilar-Rodríguez J., Vicente-Ripolles M., Fuster G., Bernet G.P., Maumus F., Munoz-Pomer A., Sempere J.M., Latorre A., Moya A. (2011). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Research, 39*(Database issue), D70–D74. https://doi.org/10.1093/nar/gkq1061

Lombard F., Renaud F., Sainsbury C., Sciandra A., Gorsky G. (2009). Appendicularian ecophysiology I: Food concentration dependent clearance rate, assimilation efficiency, growth and reproduction of *Oikopleura dioica*. *Journal of Marine Systems, 78*(4), 606–616. https://doi.org/10.1016/j.jmarsys.2009.01.004

Luke G.N., Castro L.F.C., McLay K., Bird C., Coulson A., Holland P.W.H. (2003). Dispersal of NK homeobox gene clusters in amphioxus and humans. *Proceedings of the National Academy of Sciences of the United States of America, 100*(9), 5292–5295. https://doi.org/10.1073/pnas.0836141100

Manni M., Berkeley M.R., Seppey M., Simão F.A., Zdobnov E.M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution, 38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Marçais G., Kingsford C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

Martí-Solans J., Ferrández-Roldán A., Godoy-Marín H., Badia-Ramentol J., Torres-Aguila N.P., Rodríguez-Marí A., Bouquet J.M., Chourrout D., Thompson E.M., Albalat R., Cañestro C. (2015). *Oikopleura dioica* culturing made easy: a low-cost facility for an emerging animal model in EvoDevo. *Genesis, 53*(1), 183–193. https://doi.org/10.1002/dvg.22800

Masunaga A., Liu A.W., Tan Y., Scott A., Luscombe N.M. (2020). Streamlined sampling and cultivation of the pelagic cosmopolitan larvacean, *Oikopleura dioica*. *Journal of Visualized Experiments: JoVE*, *160*. https://doi.org/10.3791/61279

Masunaga A., Mansfield M.J., Tan Y., Liu A.W., Bliznina A., Barzaghi P., Hodgetts T.L., Ferrández-Roldán A., Cañestro C., Onuma T., Plessy C., Luscombe N.M. (2022). The cosmopolitan appendicularian *Oikopleura dioica* reveals hidden genetic diversity around the globe. *Marine Biology,* 169(12), 1–17. https://doi.org/10.1007/s00227-022-04145-5

Melters D.P., Bradnam K.R., Young H.A., Telis N., May M.R., Ruby J.G., Sebra R., Peluso P., Eid J., Rank D., Garcia J.F., DeRisi J.L., Smith T., Tobias C., Ross-Ibarra J., Korf I., Chan S.W.L. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, *14*(1), R10. https://doi.org/10.1186/gb-2013-14-1-r10

Mikhaleva Y., Kreneisz O., Olsen L.C., Glover J.C., Chourrout D. (2015). Modification of the larval swimming behavior in *Oikopleura dioica*, a chordate with a miniaturized central nervous system by dsRNA injection into fertilized eggs. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution, 324*(2), 114–127. https://doi.org/10.1002/jez.b.22607

Millar R.H. (1971). The biology of ascidians. In F.S. Russell and M.Yonge (Eds.), *Advances in Marine Biology* (Vol. 9, pp. 1–100). Academic Press. https://doi.org/10.1016/S0065-2881(08)60341-7

Mitsuhashi S., Ohori S., Katoh K., Frith M.C., Matsumoto N. (2020). A pipeline for complete characterization of complex germline rearrangements from long DNA reads. *Genome Medicine, 12*(1), 67. https://doi.org/10.1186/s13073-020-00762-1

Nakashima K., Yamada L., Satou Y., Azuma J.-I., Satoh N. (2004). The evolutionary origin of animal cellulose synthase. *Development Genes and Evolution, 214*(2), 81–88. https://doi.org/10.1007/s00427-003-0379-8

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*(7), 1161–1168.e6. https://doi.org/10.1016/j.cub.2019.01.080

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 & Chromatin, 10*, 3. https://doi.org/10.1186/s13072-016-0110-4

Nishida H. (2008). Development of the appendicularian *Oikopleura dioica*: culture, genome, and cell lineages. *Development, Growth & Differentiation, 50* Suppl 1, S239–S256. https://doi.org/10.1111/j.1440-169X.2008.01035.x

Noé L., Kucherov G. (2005). YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Research, 33*(Web Server issue), W540–W543. https://doi.org/10.1093/nar/gki478

Nydam M.L., Harrison R.G. (2010). Polymorphism and divergence within the ascidian genus *Ciona*. In *Molecular Phylogenetics and Evolution* (Vol. 56, Issue 2, pp. 718–726). https://doi.org/10.1016/j.ympev.2010.03.042

Oda-Ishii I., Bertrand V., Matsuo I., Lemaire P., Saiga H. (2005). Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians *Halocynthia roretzi* and *Ciona intestinalis*. *Development, 132*(7), 1663–1674. https://doi.org/10.1242/dev.01707

Olson D., Wheeler T. (2018). ULTRA: a model based tool to detect tandem repeats. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; p. 37–46. https://doi.org/10.1145/3233547.3233604

Olsson R., Holmberg K., Lilliemarck Y. (1990). Fine structure of the brain and brain nerves of *Oikopleura dioica* (Urochordata, Appendicularia). *Zoomorphology, 110*(1), 1–7. https://doi.org/10.1007/BF01632806

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*, 44226. https://doi.org/10.1038/srep44226

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*(4), 261–267. https://doi.org/10.1007/s00427-013-0438-8

Omotezako T., Onuma T.A., Nishida H. (2015). DNA interference: DNA-induced gene silencing in the appendicularian *Oikopleura dioica*. *Proceedings. Biological Sciences / The Royal Society, 282*(1807), 20150435. https://doi.org/10.1098/rspb.2015.0435

Onuma T.A., Isobe M., Nishida H. (2017). Internal and external morphology of adults of the appendicularian, *Oikopleura dioica*: an SEM study. *Cell and Tissue Research, 367*(2), 213–227. https://doi.org/10.1007/s00441-016-2524-5

Onuma T.A., Nishida H. (2021). Developmental biology of the larvacean *Oikopleura dioica*: Genome resources, functional screening, and imaging. *Development, Growth & Differentiation.* https://doi.org/10.1111/dgd.12769

Ou S., Su W., Liao Y., Chougule K., Agda J.R.A., Hellinga A.J., Lugo C.S.B., Elliott T.A., Ware D., Peterson T., Jiang N., Hirsch C.N., Hufford M.B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology, 20*(1), 275. https://doi.org/10.1186/s13059-019-1905-y

Payne A., Holmes N., Rakyan V., Loose M. (2019). BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics, 35*(13), 2193–2198. https://doi.org/10.1093/bioinformatics/bty841

Pearson J.C., Lemons D., McGinnis W. (2005). Modulating Hox gene functions during animal body patterning. *Nature Reviews. Genetics, 6*(12), 893–904. https://doi.org/10.1038/nrg1726

Pichon J., Luscombe N.M., Plessy C. (2019). Widespread use of the "ascidian" mitochondrial genetic code in tunicates. *F1000Research*, *8*, 2072. https://doi.org/10.12688/f1000research.21551.2

Putnam N.H., Butts T., Ferrier D.E.K., Furlong R.F., Hellsten U., Kawashima T., Robinson-Rechavi M., Shoguchi E., Terry A., Yu J.-K., Benito-Gutiérrez E.L., Dubchak I., Garcia-Fernàndez J., Gibson-Brown J.J., Grigoriev I.V., Horton A.C., de Jong P.J., Jurka J., Kapitonov V.V., … Rokhsar D.S. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature, 453*(7198), 1064–1071. https://doi.org/10.1038/nature06967

Putnam N.H., O'Connell B.L., Stites J.C., Rice B.J., Blanchette M., Calef R., Troll C.J., Fields A., Hartley P.D., Sugnet C.W., Haussler D., Rokhsar D.S., Green R.E. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research, 26*(3), 342–350. https://doi.org/10.1101/gr.193474.115

Rubinstein N.D., Feldstein T., Shenkar N., Botero-Castro F., Griggio F., Mastrototaro F., Delsuc F., Douzery E.J.P., Gissi C., Huchon D. (2013). Deep sequencing of mixed total DNA without barcodes allows efficient assembly of highly plastic ascidian mitochondrial genomes. *Genome Biology and Evolution, 5*(6), 1185–1199. https://doi.org/10.1093/gbe/evt081

Sagane Y., Zech K., Bouquet J.-M., Schmid M., Bal U., Thompson E.M. (2010). Functional specialization of cellulose synthase genes of prokaryotic origin in chordate larvaceans. *Development, 137*(9), 1483–1492. https://doi.org/10.1242/dev.044503

Saito H. (2019). The kuroshio. In *Kuroshio Current* (pp. 1–11). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119428428.ch1

Sato R., Tanaka Y., Ishimaru T. (2001). House production by *Oikopleura dioica* (Tunicata, Appendicularia) under laboratory conditions. *Journal of Plankton Research*, *23*(4), 415–423. https://doi.org/10.1093/plankt/23.4.415

Satou Y., Hamaguchi M., Takeuchi K., Hastings K.E.M., Satoh, N. (2006). Genomic overview of mRNA 5′-leader trans-splicing in the ascidian *Ciona intestinalis*. Nucleic Acids Research, 34(11), 3378–3388. https://doi.org/10.1093/nar/gkl418

Satou Y., Nakamura R., Yu D., Yoshida R., Hamada M., Fujie M., Hisata K., Takeda H., Satoh N. (2019). A nearly complete genome of *Ciona intestinalis* Type A (*C. robusta*) reveals the contribution of inversion to chromosomal evolution in the genus *Ciona*. *Genome Biology and Evolution*, *11*(11), 3144–3157. https://doi.org/10.1093/gbe/evz228

Satou Y., Sato A., Yasuo H., Mihirogi Y., Bishop J., Fujie M., Kawamitsu M., Hisata K., Satoh N. (2021). Chromosomal inversion polymorphisms in two sympatric ascidian lineages. *Genome Biology and Evolution, 13*(6). https://doi.org/10.1093/gbe/evab068

Schaeffer S.W. (2018). Muller "elements" in Drosophila: how the search for the genetic basis for speciation led to the birth of comparative genomics. *Genetics*, *210*(1), 3–13. https://doi.org/10.1534/genetics.118.301084

Schulmeister A., Schmid M., Thompson E.M. (2007). Phosphorylation of the histone H3.3 variant in mitosis and meiosis of the urochordate *Oikopleura dioica*. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *15*(2), 189–201. https://doi.org/10.1007/s10577-006-1112-z

Sela I., Ashkenazy H., Katoh K., Pupko T. (2015). GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Research, 43*(W1), W7–W14. https://doi.org/10.1093/nar/gkv318

Seo H.-C., Edvardsen R.B., Maeland A.D., Bjordal M., Jensen M.F., Hansen A., Flaat M., Weissenbach J., Lehrach H., Wincker P., Reinhardt R., Chourrout D. (2004). Hox cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*. *Nature*, *431*(7004), 67–71. https://doi.org/10.1038/nature02709

Seo H.-C., Kube M., Edvardsen R.B., Jensen M.F., Beck A., Spriet E., Gorsky G., Thompson E.M., Lehrach H., Reinhardt R., Chourrout D. (2001). Miniature genome in the marine chordate *Oikopleura dioica*. *Science*, *294*(5551), 2506. https://doi.org/10.1126/science.294.5551.2506

Severin J., Lizio M., Harshbarger J., Kawaji H., Daub C.O., Hayashizaki Y., FANTOM Consortium, Bertin N., Forrest A.R.R. (2014). Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nature Biotechnology*, *32*(3), 217–219. https://doi.org/10.1038/nbt.2840

Schmidbaur H., Kawaguchi A., Clarence T., Fu X., Hoang O.P., Zimmermann B., Ritschard E.A., Weissenbacher A., Foster J.S., Nyholm S.V. and Bates P.A. (2022). Emergence of novel

cephalopod gene regulation and expression through large-scale genome reorganization. *Nature communications, 13*(1), pp.1-11.

Shenkar N., Koplovitz G., Dray L., Gissi C., Huchon D. (2016). Back to solitude: Solving the phylogenetic position of the *Diazonidae* using molecular and developmental characters. *Molecular Phylogenetics and Evolution, 100*, 51–56. https://doi.org/10.1016/j.ympev.2016.04.001

Sherlock R.E., Walz K.R., Schlining K.L., Robison B.H. (2017). Morphology, ecology, and molecular biology of a new species of giant larvacean in the eastern North Pacific: *Bathochordaeus mcnutti* sp. nov. *Marine Biology, 164*(1), 20. https://doi.org/10.1007/s00227-016-3046-0

Shoguchi E., Kawashima T., Nishida-Umehara C., Matsuda Y., Satoh N. (2005). Molecular cytogenetic characterization of *Ciona intestinalis* chromosomes. *Zoological Science*, *22*(5), 511–516. https://doi.org/10.2108/zsj.22.511

Shumate A., Salzberg S.L. (2020). Liftoff: accurate mapping of gene annotations. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa1016

Simakov O., Marletaz F., Cho S.-J., Edsinger-Gonzales E., Havlak P., Hellsten U., Kuo D.-H., Larsson T., Lv J., Arendt D., Savage R., Osoegawa K., de Jong P., Grimwood J., Chapman J.A., Shapiro H., Aerts A., Otillar R.P., Terry A.Y., … Rokhsar D.S. (2013). Insights into bilaterian evolution from three spiralian genomes. *Nature, 493*(7433), 526–531. https://doi.org/10.1038/nature11696

Simakov O., Kawashima T., Marlétaz F., Jenkins J., Koyanagi R., Mitros T., Hisata K., Bredeson J., Shoguchi E., Gyoja F., Yue J.-X., Chen Y.-C., Freeman R. M., Jr, Sasaki A., Hikosaka-Katayama T., Sato A., Fujie M., Baughman K.W., Levine J., … Gerhart J. (2015). Hemichordate genomes and deuterostome origins. *Nature, 527*(7579), 459–465. https://doi.org/10.1038/nature16150

Simakov O., Marlétaz F., Yue J.-X., O'Connell B., Jenkins J., Brandt A., Calef R., Tung C.-H., Huang T.-K., Schmutz J., Satoh N., Yu J.-K., Putnam N. H., Green R. E., Rokhsar D.S. (2020). Deeply conserved synteny resolves early events in vertebrate evolution. *Nature Ecology & Evolution*, *4*(6), 820–830. https://doi.org/10.1038/s41559-020-1156-z

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*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simsek D., Jasin M. (2010). Alternative end-joining is suppressed by the canonical NHEJ component Xrcc4–ligase IV during chromosomal translocation formation. *Nature Structural & Molecular Biology, 17*(4), 410–416. https://doi.org/10.1038/nsmb.1773

Singh T.R., Tsagkogeorga G., Delsuc F., Blanquart S., Shenkar N., Loya Y., Douzery E.J., Huchon D. (2009). Tunicate mitogenomics and phylogenetics: peculiarities of the Herdmania momus mitochondrial genome and support for the new chordate phylogeny. BMC Genomics, 10, 534. https://doi.org/10.1186/1471-2164-10-534

Slater G.S.C., Birney E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31. https://doi.org/10.1186/1471-2105-6-31

Smit A.F.A., Hubley R., Green P. RepeatMasker at http://repeatmasker.org

Spada F., Steen H., Troedsson C., Kallesoe T., Spriet E., Mann M., Thompson E.M. (2001). Molecular patterning of the oikoplastic epithelium of the larvacean tunicate *Oikopleura dioica*. *The Journal of Biological Chemistry, 276*(23), 20624–20632. https://doi.org/10.1074/jbc.M100438200

Stajich J.E., Block D., Boulez K., Brenner S.E., Chervitz S.A., Dagdigian C., Fuellen G., Gilbert J.G.R., Korf I., Lapp H., Lehväslaiho H., Matsalla C., Mungall C.J., Osborne B.I., Pocock M.R., Schattner P., Senger M., Stein L.D., Stupka E., … Birney E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, *12*(10), 1611–1618. https://doi.org/10.1101/gr.361602

Stanke M., Schöffmann O., Morgenstern B., Waack S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, *7*, 62. https://doi.org/10.1186/1471-2105-7-62

Storer J., Hubley R., Rosen J., Wheeler T.J., Smit A.F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA, 12*(1), 2. https://doi.org/10.1186/s13100-020-00230-y

Suyama M., Torrents D., Bork P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research, 34*(Web Server issue), W609–W612. https://doi.org/10.1093/nar/gkl315

Swalla B.J., Cameron C.B., Corley L.S., Garey J.R. (2000). Urochordates are monophyletic within the deuterostomes. *Systematic Biology, 49*(1), 52–64. https://doi.org/10.1080/10635150050207384

Tan G., Polychronopoulos D. and Lenhard B. (2019). CNEr: A toolkit for exploring extreme noncoding conservation. *PLoS computational biology, 15*(8), p.e1006940. https://doi.org/10.1371/journal.pcbi.1006940

Thompson E.M., Kallesøe T., Spada F. (2001). Diverse genes expressed in distinct regions of the trunk epithelium define a monolayer cellular template for construction of the oikopleurid house. *Developmental Biology, 238*(2), 260–273. https://doi.org/10.1006/dbio.2001.0414

Tokioka T. (1960). Studies on the distribution of appendicularians and some thaliaceans of the north pacific, with some morphological notes. *Publications of the Seto Marine Biological Laboratory, 8*(2), 351–443. https://doi.org/10.5134/174644

Torres-Águila N.P., Martí-Solans J., Ferrández-Roldán A., Almazán A., Roncalli V., D'Aniello S., Romano G., Palumbo A., Albalat R., Cañestro C. (2018). Diatom bloom-derived biotoxins cause aberrant development and gene expression in the appendicularian chordate *Oikopleura dioica*. *Communications Biology, 1*, 121. https://doi.org/10.1038/s42003-018-0127-2

Troedsson C., Bouquet J.-M., Lobon C.M., Novac A., Nejstgaard J.C., Dupont S., Bosak S., Jakobsen H.H., Romanova N., Pankoke L.M., Isla A., Dutz J., Sazhin A.F., Thompson E.M. (2013). Effects of ocean acidification, temperature and nutrient regimes on the appendicularian *Oikopleura dioica*: a mesocosm study. Marine Biology, 160(8), 2175–2187. https://doi.org/10.1007/s00227-012-2137-9

Tsagkogeorga G., Cahais V., Galtier N. (2012). The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. Genome Biology and Evolution, 4(8), 740–749. https://doi.org/10.1093/gbe/evs054

Tsagkogeorga G., Turon X., Galtier N., Douzery E.J.P., Delsuc F. (2010). Accelerated evolutionary rate of housekeeping genes in tunicates. *Journal of Molecular Evolution, 71*(2), 153–167. https://doi.org/10.1007/s00239-010-9372-9

Tsagkogeorga G., Turon X., Hopcroft R.R., Tilak M.-K., Feldstein T., Shenkar N., Loya Y., Huchon D., Douzery E.J.P., Delsuc F. (2009). An updated 18S rRNA phylogeny of tunicates based on mixture and secondary structure models. *BMC Evolutionary Biology, 9*, 187. https://doi.org/10.1186/1471-2148-9-187

Tyson J.R., O'Neil N.J., Jain M., Olsen H.E., Hieter P., Snutch T.P. (2018). MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. *Genome Research, 28*(2), 266–274. https://doi.org/10.1101/gr.221184.117

Uye S.-I., Ichino S. (1995). Seasonal variations in abundance, size composition, biomass and production rate of *Oikopleura dioica* (Fol) (Tunicata: Appendicularia) in a temperate eutrophic inlet. *Journal of Experimental Marine Biology and Ecology, 189*(1), 1–11. https://doi.org/10.1016/0022-0981(95)00004-B

Vaser R., Sović I., Nagarajan N., Šikić M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/gr.214270.116

Villarreal D.D., Lee K., Deem A., Shim E.Y., Malkova A., Lee S.E. (2012). Microhomology directs diverse DNA break repair pathways and chromosomal translocations. *PLoS Genetics, 8*(11), e1003026. https://doi.org/10.1371/journal.pgen.1003026

Volff J.-N., Lehrach H., Reinhardt R., Chourrout D. (2004). Retroelement dynamics and a novel type of chordate retrovirus-like element in the miniature genome of the tunicate *Oikopleura dioica*. *Molecular Biology and Evolution, 21*(11), 2022–2033. https://doi.org/10.1093/molbev/msh207

Vurture G.W., Sedlazeck F.J., Nattestad M., Underwood C.J., Fang H., Gurtowski J., Schatz M.C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Walker B.J., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar S., Cuomo C.A., Zeng Q., Wortman J., Young S.K., Earl A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

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*(3), 149–159. https://doi.org/10.1007/s00427-015-0502-7

Wang K., Tomura R., Chen W., Kiyooka M., Ishizaki H., Aizu T., Minakuchi Y., Seki M., Suzuki Y., Omotezako T., Suyama R., Masunaga A., Plessy C., Luscombe N.M., Dantec C., Lemaire P., Itoh T., Toyoda A., Nishida H., Onuma T.A. (2020). A genome database for a Japanese population of the larvacean *Oikopleura dioica*. *Development, Growth & Differentiation*, *62*(6), 450–461. https://doi.org/10.1111/dgd.12689

Warren I.A., Naville M., Chalopin D., Levin P., Berger C.S., Galiana D., Volff J.-N. (2015). Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology, 23*(3), 505–531. https://doi.org/10.1007/s10577-015-9493-5

Waterhouse R.M., Seppey M., Simão F.A., Manni M., Ioannidis P., Klioutchnikov G., Kriventseva E.V., Zdobnov E.M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548. https://doi.org/10.1093/molbev/msx319

Weisman C.M., Murray A.W., Eddy S.R. (2022). Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Current Biology, 32*(12), 2632–2639.e2. https://doi.org/10.1016/j.cub.2022.04.085

Weill M., Philips A., Chourrout D., Fort P. (2005). The caspase family in urochordates: distinct evolutionary fates in ascidians and larvaceans. *Biology of the Cell / under the Auspices of the European Cell Biology Organization, 97*(11), 857–866. https://doi.org/10.1042/BC20050018

Wenke T., Döbel T., Sörensen T. R., Junghans H., Weisshaar B., Schmidt T. (2011). Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *The Plant Cell*, *23*(9), 3117–3128. https://doi.org/10.1105/tpc.111.088682

Wicker T., Sabot F., Hua-Van A., Bennetzen J.L., Capy P., Chalhoub B., Flavell A., Leroy P., Morgante M., Panaud O., Paux E., SanMiguel P., Schulman A.H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews. Genetics, 8*(12), 973–982. https://doi.org/10.1038/nrg2165

Workman R.E., Tang A.D., Tang P.S., Jain M., Tyson J.R., Razaghi R., Zuzarte P.C., Gilpatrick T., Payne A., Quick J., Sadowski N., Holmes N., de Jesus J.G., Jones K.L., Soulette C.M., Snutch T.P., Loman N., Paten B., Loose M., … Timp W. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods, 16*(12), 1297–1305. https://doi.org/10.1038/s41592-019-0617-2

Yadetie F., Butcher S., Førde H.E., Campsteijn C., Bouquet J.-M., Karlsen O.A., Denoeud F., Metpally R., Thompson E.M., Manak J.R., Goksøyr A., Chourrout D. (2012). Conservation and divergence of chemical defense system in the tunicate *Oikopleura dioica* revealed by genome wide response to two xenobiotics. *BMC Genomics, 13*, 55. https://doi.org/10.1186/1471-2164-13-55

Yandell M., Ence D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews. Genetics, 13*(5), 329–342. https://doi.org/10.1038/nrg3174

Yang Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences: *CABIOS, 13*(5), 555–556. https://doi.org/10.1093/bioinformatics/13.5.555

Yang Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution, 24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yao Y., Frith M.C. (2021). Improved DNA-versus-Protein homology search for protein fossils. *In bioRxiv* (p. 2021.01.25.428050). https://doi.org/10.1101/2021.01.25.428050

Žárský V., Tachezy J. (2015). Evolutionary loss of peroxisomes--not limited to parasites. *Biology Direct*, *10*(1), 74. https://doi.org/10.1186/s13062-015-0101-6

Zdobnov E.M., Tegenfeldt F., Kuznetsov D., Waterhouse R.M., Simão F.A., Ioannidis P., Seppey M., Loetscher A., Kriventseva E.V. (2017). OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research*, *45*(D1), D744–D749. https://doi.org/10.1093/nar/gkw1119

Zeller R.W. (2010). Computational analysis of *Ciona intestinalis* operons. Integrative and Comparative Biology, 50(1), 75–85. https://doi.org/10.1093/icb/icq040

Zeng L., Swalla B.J. (2005). Molecular phylogeny of the protochordates: chordate evolution. *Canadian Journal of Zoology, 83*(1), 24–33. https://doi.org/10.1139/z05-010

Zorio D.A., Cheng N.N., Blumenthal T., Spieth J. (1994). Operons as a common form of chromosomal organization in *C. elegans*. *Nature, 372*(6503), 270–272. https://doi.org/10.1038/372270a0

Appendices

Appendix 1: List of oikopleurid gene homologs in the OKI2018_I69 genome uploaded to ZENBU.

Gene name/symbol	Gene name/symbol description	Gene accession number	Coordinates in OKI2018_I69 assembly	PMIDs
OdiT	Brachyury	AF204208	chr2:12776999-12778322	10753519
hox1	homeobox 1	AAS21474	chr2:14331704-14334515	15343333
hox2	homeobox 2	CBY42438	chr1:6753504-6754675	
hox4	homeobox 4	AAT47829	chr2:4686021-4687168	
hox10	homeobox 10	AAT47861	XSR:11243346-11244698	
hox11	homeobox 11	AAS21413	PAR:15661224-15664812	
hox12	homeobox 12	AAS21383	chr1:9941250-9945471	
hox13	homeobox 13	CBY20174	chr2:6661193-6665186	
gad	glutamic acid dehydroxylase	GSOIDT00008657001	chr1:8244080-8253873	16041716
otx a	orthodenticle homeobox a	AY886542	PAR:11286911-11287300	
otx b	orthodenticle homeobox b	AY897556	PAR:10581308-10582329	
otx c	orthodenticle homeobox c	AY897557	PAR:10598978-10601491	
pax6	paired box 6	GSOIDT00010489001	PAR:6330739-6336615	
pax2/5/8a (pax2-5-8a)	paired box 2/5/8a	DQ020279	chr2:14617328-14620230	
pax2/5/8b(pax2-5-8b)	paired box 2/5/8b	AY870649	XSR:8667124-8668521	
engrailed	engrailed gene	AY870647	chr2:6350277-6351157	
eya	eyes absent	DQ011272	chr2:11719030-11722577	16120641
pitx	pituitary homeobox transcript	DQ011274	PAR:7083214-7084966	
six1/2(six1.2)	six(Sineoculis homeobox homolog 1) homeobox1/2	DQ011279	chr1:8738535-8739860	
six3/6a(six3.6a)	six (Sineoculis homeobox homolog 1) homeobox 3/6a	DQ011281	chr2:8503938-8504892	
six3/6b	six (Sineoculis homeobox homolog 1) homeobox 3/6b	DQ011283	chr2:13702728-13703597	

Appendix 1: List of oikopleurid gene homologs in the OKI2018_I69 genome uploaded to ZENBU (continued).

Gene name/symbol	Gene name/symbol description	Gene accession number	Coordinates in OKI2018_I69 assembly	PMIDs
Od-prickle	Od-prickle	GSOIDT00017601001	chr2:15081611-15085582	
Od-quaking	Od-quaking	GSOIDT00018697001	chr1:7452574-7456213	
Od- thrombospondin 3	Od- thrombospondin 3	GSOIDT00001381001	chr1:567892-569445	
Od-FCol1	Od-Fibrillar collagen 1	GSOIDT00025286001	chr1:6055785-6059763	
Od-b4-GalT	Od- b1,4-Galactosyltransferase	GSOIDT00003166001	XSR:3486652-3488744	
Od- Calumenin1	Od- Calumenin1	GSOIDT00007932001	XSR:305232-306472	
Od-Calumenin2	Od-Calumenin2	GSOIDT00008481001	PAR:6046647-6047779	
Od-ERM	Od-Ezrin/radixin/moesin	GSOIDT00009195001	chr1:11624438-11625687	
Od-IQGAP	Od-IQ motif-containing GTPase-activating protein	GSOIDT00008004001	XSR:3814962-3818503	
Od-leprecan	Od-leprecan	GSOIDT00020891001	chr1:7808710-7810912	
Od-Zipper	Od-Zipper	GSOIDT00003226001	XSR:3684218-3691118	
Od-netrin	Od-netrin	GSOIDT00023202001	chr2:853071-856681	
Od-laminin a1	Od-laminin a1	GSOIDT00030725001	PAR:6663470-6672829	
Od- cdc45	Od-cell division cycle	GSOIDT00000575001	chr2:7387928-7389648	21251251
Od-Noto9a	Od-notochord homeobox 9a	GSOIDT00027089001	XSR:6076441-6078772	
Od-Noto9b	Od-notochord homeobox 9b	GSOIDT00012942001	PAR:8852530-8854062	
Od-Noto9c	Od-notochord homeobox 9c	GSOIDT00017670001	chr2:12381322-12384526	
Od-Noto10	Od-notochord homeobox 10	GSOIDT00007691001	XSR:10740653-10741915	
Od-Noto15a	Od-notochord homeobox 15a	GSOIDT00012625001	chr2:1715120-1717880	
Od-Noto17	Od-notochord homeobox 17	GSOIDT00025490001	chr1:9934843-9936189	
Od-ARNT	Od-aryl hydrocarbon receptor nuclear translocator	GSOIDT00027679001	PAR:14983557-14984855	
Od-PCNA	Od-proliferating cell nuclear antigen	GSOIDT00012375001	PAR:7601325-7604703	
Od-ACL	Od- tensin, Od-ATP citrate lyase	GSOIDT0013472001	chr2:5885019-5891083	
Od-ASAK	ATP sulfurylase/ adenosine 5'-phosphosulfate (APS) kinase	GSOIDT00007183001	PAR:7601325-7604703	
Od- CaMK	Od-calmodulin-dependent protein kinase	GSOIDT00012261001	chr2:9619068-9621630	
Od- tensin	Od-tensin	GSOIDT00008372001	XSR:10631881-10633946	
ChAT	choline acetyltransferase	GSOIDT00013449001	chr2:5954949-5960580	25676192

Appendix 1: List of oikopleurid gene homologs in the OKI2018_I69 genome uploaded to ZENBU (continued).

Gene name/symbol	Gene name/symbol description	Gene accession number	Coordinates in OKI2018_I69 assembly	PMIDs
cadherin-6 precursor	cadherin-6 precursor	GSOIDT00011474001	chr1:10141257-10146632	28281645
catenin alpha-1	catenin alpha-1	GSOIDT00013235001	chr2:13019153-13021938	
rpa-interacting protein a	replication protein A (rpa) interacting protein a	GSOIDT00013373001	chr2:8959846-8960515	
CSE1L	chromosome segregation 1-like	GSOIDT00013374001	chr2:8955007-8959732	
alpha 1b tubulin	tubulin alpha 1b	GSOIDT00000916001	chr1:13529476-13531209	
tubulin beta 2c	tubulin beta 2c	GSOIDT00006170001	XSR:5093488-5094942	
Actin	Actin	GSOIDG00000756001	XSR:6650011-6650459	30272001
Tis11a	Tis11a zinc finger protein	GSOIDG00015222001	XSR:434576-438400	
SoxBa	thiosulfohydrolase SoxB a	GSOIDG00010386001	XSR:1417659-1419459	
SoxBb	thiosulfohydrolase SoxB b	GSOIDG00013526001	chr2:13856787-13858894	
Wnt11	Wnt family member 11	GSOIDG00011688001	chr1:10092471-10096192	
Nkx2.3/5/6	NK2 homeobox 3/5/6	GSOIDG00003812001	XSR:3368088-3369109	
Tis11b	Tis11b zinc finger protein	GSOIDG00017080001	chr2:11614083-11614962	
Glcm	glutamate-cysteine ligase, modifier subunit	GSOIDG00006303001	chr1:5045440-5046968	
Aldh3	Alcohol dehydrogenase 3	GSOIDG00000110001	chr2:11096657-11098257	
Aldh2	Alcohol dehydrogenase 2	GSOIDG00002220001	PAR:10375200-10376853	
Aldh8a1	Alcohol dehydrogenase 8a1	GSOIDG00021101001	chr2:11735877-11737715	
pum1	pumilio	CBY19123.1	PAR:15230053-15232590	29486709
vas4	vas4	CBY13641.1	XSR:10942618-10944231	
CDK1a	Cyclin-Dependant Kinase 1a	FR822377	chr2:16006560-16010370	29969934
CDK1d	Cyclin-Dependant Kinase 1d	FR822380	hard to resolve	
Cyclin Ba	Cyclin Ba	FR821604	XSR:8244212-8245406	
Cyclin B3a	Cyclin B3a	FR821607	chr2:14062690-14064201	
Bmp3	bone morphogenetic protein 3	GSOIDT00007253001	PAR:15677313-15677706	32029598
Bmp.a	bone morphogenetic protein a	GSOIDT00001216001	PAR:13401259-13409969	
Bmp.b	bone morphogenetic protein b	GSOIDT00001809001	PAR:11058990-11061552	
Od-CesA1	*Oikopleura dioica Cellulose synthase A1*	AB543594	PAR:11400951-11404576	20972815
Od-CesA2	*Oikopleura dioica Cellulose synthase A2*	AB543593	chr1:13074593-13080860	20335363
OdMT1	*Oikopleura dioica Metallothioneins 1*	MH577047	no hits found	30284576

Appendix 1: List of oikopleurid gene homologs in the OKI2018_I69 genome uploaded to ZENBU (continued).

Gene name/symbol	Gene name/symbol description	Gene accession number	Coordinates in OKI2018_I69 assembly	PMIDs
OdMT2	*Oikopleura dioica Metallothioneins 2*	MH577046	PAR:16247483-16248470	30284576
ActnM1	muscle actin 1	GSOIDT00000756001	XSR:6650011-6650459	30217598
ActnM2	muscle actin 2	GSOIDT00011141001	hard to resolve	
ActnM3	muscle actin 3	GSOIDT00012400001	no hits found	
ActnM4	muscle actin 4	GSOIDT00016817001	hard to resolve	
ActnC1	cytoplasmic actins 1	GSOIDT00000372001	hard to resolve	
ActnC2	cytoplasmic actins 2	GSOIDT00013012001	hard to resolve	
ActnC3	cytoplasmic actins 3	GSOIDT00013080001	hard to resolve	
OctA1	octamer A1 POU(pit, oct, unc) domain containing transcription factors	DQ328331	chr2:7057188-7062411	16989962
OctA2	octamer A2 POU(pit, oct, unc) domain containing transcription factors	DQ328332	chr2:7057323-7062411	
OctB	octamer B POU(pit, oct, unc) domain containing transcription factors	DQ328333	PAR:711236-712452	
oik1	Oikosin(Oikopleura house proteins)1	AJ308491	chr2:2447948-2455328	22792236
oik2	Oikosin(Oikopleura house proteins)2	AJ308492	chr1:2603562-2604812	
oik3	Oikosin(Oikopleura house proteins)3	AJ308495	chr2:4029333-4030738	
oik4	Oikosin(Oikopleura house proteins)4	AJ310624	no hits found	
oik5	Oikosin(Oikopleura house proteins)5	AJ310627	PAR:4067945-4068922	
oik6	Oikosin(Oikopleura house proteins)6	AJ310629	no hits found	
oik7	Oikosin(Oikopleura house proteins)7	AJ310634	PAR:1552601-1553164	
oik8	Oikosin(Oikopleura house proteins)8	FN806849	XSR:854966-855978	
oik9	Oikosin(Oikopleura house proteins)9	HE663406	chr1:1041205-1043279	
oik10	Oikosin(Oikopleura house proteins)10	HE663407	chr2:1991876-1993444	
oik11	Oikosin(Oikopleura house proteins)11	HE663408	chr2:2156021-2157043	
oik12	Oikosin(Oikopleura house proteins)12	HE663409	PAR:2593491-2594228	
oik13	Oikosin(Oikopleura house proteins)13	HE663410	chr1:4110443-4113040	
oik14	Oikosin(Oikopleura house proteins)14	HE663411	chr1:1087968-1101578	
oik15	Oikosin(Oikopleura house proteins)15	HE663412	XSR:1996460-1997122	
oik16	Oikosin(Oikopleura house proteins)16	HE774605	PAR:12310822-12311469	
oik17a	Oikosin(Oikopleura house proteins)17a	HE774606	hard to resolve	
oik17b	Oikosin(Oikopleura house proteins)17b	HE774607	chr1:1192466-1195503	

Appendix 1: List of oikopleurid gene homologs in the OKI2018_I69 genome uploaded to ZENBU (continued).

Gene name/symbol	Gene name/symbol description	Gene accession number	Coordinates in OKI2018_I69 assembly	PMIDs
oik18	Oikosin(Oikopleura house proteins)18	FN806850	XSR:11866211-11867008	
oik19	Oikosin(Oikopleura house proteins)19	HE774608	hard to resolve	
oik20	Oikosin(Oikopleura house proteins)20	HE774609	PAR:1790119-1791291	
oik21a	Oikosin(Oikopleura house proteins)21a	HE774610	XSR:5640811-5643564	
oik21b	Oikosin(Oikopleura house proteins)21b	HE774611	chr1:2145866-2148510	
oik22	Oikosin(Oikopleura house proteins)22	HE774612	chr2:11445269-11446392	
oik23	Oikosin(Oikopleura house proteins)23	HE774613	XSR:2469609-2475836	
oik24a	Oikosin(Oikopleura house proteins)24a	HE774614	hard to resolve	
oik24b	Oikosin(Oikopleura house proteins)24b	HE774615	hard to resolve	
oik24c	Oikosin(Oikopleura house proteins)24c	HE774616	hard to resolve	
oik24d	Oikosin(Oikopleura house proteins)24d	HE774617	hard to resolve	
oik24e	Oikosin(Oikopleura house proteins)24e	HE774618	hard to resolve	
oik24f	Oikosin(Oikopleura house proteins)24f	HE774619	hard to resolve	
oik24g	Oikosin(Oikopleura house proteins)24g	HE774620	hard to resolve	
oik24h	Oikosin(Oikopleura house proteins)24f	HE774621	hard to resolve	22792236
oik25	Oikosin(Oikopleura house proteins)25	HE774622	chr2:2769109-2770644	
oik26	Oikosin(Oikopleura house proteins)26	HE774623	PAR:5523438-5523912	
oik27	Oikosin(Oikopleura house proteins)27	HE774624	PAR:10741877-10742325	
oik28a	Oikosin(Oikopleura house proteins)28a	HE774625	hard to resolve	
oik28b	Oikosin(Oikopleura house proteins)28b	HE774626	hard to resolve	
oik29a	Oikosin(Oikopleura house proteins)29a	HE774627	PAR:7782026-7783046	
oik29b	Oikosin(Oikopleura house proteins)29b	HE774628	PAR:7782053-7782829	
oik30a	Oikosin(Oikopleura house proteins)30a	HE774629	hard to resolve	
oik30b	Oikosin(Oikopleura house proteins)30b	HE774630	hard to resolve	
oik30c	Oikosin(Oikopleura house proteins)30c	HE774631	hard to resolve	
oik30d	Oikosin(Oikopleura house proteins)30d	HE774632	hard to resolve	
oik30e	Oikosin(Oikopleura house proteins)30e	HE774633	hard to resolve	
oik31a	Oikosin(Oikopleura house proteins)31a	HE774634	chr1:12452800-12453551	
oik31b	Oikosin(Oikopleura house proteins)31b	HE774635	PAR:2222582-2223325	

Appendix 1: List of oikopleurid gene homologs in the OKI2018_I69 genome uploaded to ZENBU (continued).

Gene name/symbol	Gene name/symbol description	Gene accession number	Coordinates in OKI2018_I69 assembly	PMIDs
oik32	Oikosin(Oikopleura house proteins)32	FN806851	XSR:5121369-5122993	
oik33a	Oikosin(Oikopleura house proteins)33a	HE774636	hard to resolve	
oik33b	Oikosin(Oikopleura house proteins)33b	HE774637	hard to resolve	
oik34a	Oikosin(Oikopleura house proteins)34a	HE774638	hard to resolve	
oik34b	Oikosin(Oikopleura house proteins)34b	HE774639	hard to resolve	
oik35	Oikosin(Oikopleura house proteins)35	HE774640	chr1:5277102-5279044	
oik36a	Oikosin(Oikopleura house proteins)36a	HE774641	chr2:2308496-2309463	
oik36b	Oikosin(Oikopleura house proteins)36b	HE774642	chr2:2308496-2309463	
oik37	Oikosin(Oikopleura house proteins)37	HE774643	XSR:10085876-10087427	
oik38	Oikosin(Oikopleura house proteins)38	HE774644	no hits found	
oik39	Oikosin(Oikopleura house proteins)39	HE774645	chr2:4845138-4846194	
oik40a	Oikosin(Oikopleura house proteins)40a	HE774646	no hits found	
oik40b	Oikosin(Oikopleura house proteins)40b	HE774647	no hits found	
oik41a	Oikosin(Oikopleura house proteins)41a	HE774648	no hits found	
oik41b	Oikosin(Oikopleura house proteins)41b	HE774649	no hits found	22792236
oik42	Oikosin(Oikopleura house proteins)42	HE774650	PAR:1612422-1615266	
oik43	Oikosin(Oikopleura house proteins)43	HE774651	chr1:3611589-3617755	
oik44	Oikosin(Oikopleura house proteins)44	HE774652	PAR:2232451-2233274	
oik45	Oikosin(Oikopleura house proteins)45	HE774653	PAR:5608964-5609715	
oik46	Oikosin(Oikopleura house proteins)46	HE774654	XSR:10878688-10879967	
oik47	Oikosin(Oikopleura house proteins)47	HE774655	chr1:4693619-4698444	
oik48	Oikosin(Oikopleura house proteins)48	HE774656	PAR:3835513-3835878	
oik49a	Oikosin(Oikopleura house proteins)49a	HE774657	hard to resolve	
oik49b	Oikosin(Oikopleura house proteins)49b	HE774658	hard to resolve	
oik50	Oikosin(Oikopleura house proteins)50	HE774659	chr1:4340494-4341624	
oik51a	Oikosin(Oikopleura house proteins)51a	HE774660	hard to resolve	
oik51b	Oikosin(Oikopleura house proteins)51b	HE774661	hard to resolve	
oik51c	Oikosin(Oikopleura house proteins)51c	HE774662	hard to resolve	
oik51d	Oikosin(Oikopleura house proteins)51d	HE774663	hard to resolve	

Appendix 2: List of missing BUSCO genes in OKI2018_I69, OdB3 and OSKA2016 genome assemblies.

Genome	Missing BUSCO genes
OKI2018_I69	EOG091G00MI, EOG091G01MK, EOG091G020E, EOG091G02E4, EOG091G02NO, EOG091G02OE, EOG091G036C, EOG091G03H4, EOG091G03JF, EOG091G03JG, EOG091G048B, EOG091G0495, EOG091G04JK, EOG091G04MA, EOG091G05D7, EOG091G05IH, EOG091G05K7, EOG091G066W, EOG091G074J, EOG091G0769, EOG091G07PH, EOG091G07VQ, EOG091G08DN, EOG091G08IJ, EOG091G08JG, EOG091G08PW, EOG091G090D, EOG091G099D, EOG091G09IW, EOG091G09J7, EOG091G09QO, EOG091G09R0, EOG091G09T3, EOG091G09UA, EOG091G0AGM, EOG091G0ARL, EOG091G0AVW, EOG091G0AX1, EOG091G0AZ7, EOG091G0BDT, EOG091G0BED, EOG091G0BKX, EOG091G0BN4, EOG091G0BNI, EOG091G0BPX, EOG091G0C2Z, EOG091G0CLR, EOG091G0CXE, EOG091G0D0D, EOG091G0DQS, EOG091G0DYU, EOG091G0E11, EOG091G0E5P, EOG091G0EA2, EOG091G0EBA, EOG091G0EHF, EOG091G0EHY, EOG091G0EM6, EOG091G0EO4, EOG091G0EZ8, EOG091G0F93, EOG091G0FH3, EOG091G0FOR, EOG091G0FSK, EOG091G0FYA, EOG091G0FZN, EOG091G0G1G, EOG091G0G58, EOG091G0G8S, EOG091G0GA7, EOG091G0GA8, EOG091G0GD0, EOG091G0GKX, EOG091G0GMC, EOG091G0GQO, EOG091G0H6J, EOG091G0HBC, EOG091G0HDF, EOG091G0HMT, EOG091G0HP9, EOG091G0HQ3, EOG091G0HT4, EOG091G0I1O, EOG091G0I5H, EOG091G0I7M, EOG091G0IDS, EOG091G0IHL, EOG091G0IMB, EOG091G0IXY, EOG091G0IZD, EOG091G0J09, EOG091G0J69, EOG091G0J98, EOG091G0JBZ, EOG091G0JC7, EOG091G0JHN, EOG091G0JIA, EOG091G0JTR, EOG091G0JWC, EOG091G0JWK, EOG091G0JZV, EOG091G0K0H, EOG091G0K1H, EOG091G0K6O, EOG091G0K82, EOG091G0KNT, EOG091G0L2T, EOG091G0LBN, EOG091G0LCT, EOG091G0LKE, EOG091G0LMQ, EOG091G0LMX, EOG091G0LPW, EOG091G0LT7, EOG091G0LTE, EOG091G0M09, EOG091G0M0J, EOG091G0M0T, EOG091G0M24, EOG091G0M4Q, EOG091G0MBG, EOG091G0MBX, EOG091G0MCM, EOG091G0MCZ, EOG091G0MG7, EOG091G0MS7, EOG091G0MSR, EOG091G0N0U, EOG091G0N9D, EOG091G0NKD, EOG091G0NM1, EOG091G0NP6, EOG091G0NP7, EOG091G0NRN, EOG091G0NTZ, EOG091G0O4W, EOG091G0O5A, EOG091G0O7O, EOG091G0O97, EOG091G0OFC, EOG091G0OPI, EOG091G0OWC, EOG091G0OY0, EOG091G0PVT, EOG091G0Q05, EOG091G0QE1, EOG091G0QS6, EOG091G0QZ2, EOG091G0R3S, EOG091G0R9X, EOG091G0RI9, EOG091G0RRA, EOG091G0RRT, EOG091G0RTI, EOG091G0RWI, EOG091G0S2R, EOG091G0S5L, EOG091G0SAU, EOG091G0SB2, EOG091G0SCV, EOG091G0SGH, EOG091G0SGT, EOG091G0SRJ, EOG091G0T3D, EOG091G0T3X, EOG091G0T5O, EOG091G0TAC, EOG091G0TAU, EOG091G0TZU, EOG091G0U2U, EOG091G0UE6, EOG091G0UOM, EOG091G0UQ0, EOG091G0UTL, EOG091G0UUT, EOG091G0UZ0, EOG091G0V3C, EOG091G0VIF, EOG091G0VMU, EOG091G0VQK, EOG091G0W0U, EOG091G0W26, EOG091G0W86, EOG091G0WCB, EOG091G0WPV, EOG091G0WUF, EOG091G0X1V, EOG091G0XKP, EOG091G0XN1, EOG091G0XRQ, EOG091G0Y05, EOG091G0Y09, EOG091G0Y35, EOG091G0Y6M, EOG091G0Y96, EOG091G0YBP, EOG091G0YUO, EOG091G0YY3, EOG091G0YYA, EOG091G0Z43, EOG091G0Z6N, EOG091G0Z7J, EOG091G0ZEJ, EOG091G0ZNH, EOG091G0ZTA, EOG091G10IX, EOG091G10MO, EOG091G10SJ, EOG091G126R, EOG091G12A5, EOG091G141O, EOG091G15HV, EOG091G15KV, EOG091G15XS, EOG091G15XZ, EOG091G1757, EOG091G17I3, EOG091G17K3, EOG091G184V, EOG091G18B1, EOG091G18BK, EOG091G18Z5, EOG091G1A3H

Appendix 2: List of missing BUSCO genes in OKI2018_I69, OdB3 and OSKA2016 genome assemblies (continued).

Genome	Missing BUSCO genes
OdB3	EOG091G01MK, EOG091G020E, EOG091G02E4, EOG091G02NO, EOG091G02OE, EOG091G036C, EOG091G03H4, EOG091G03JG, EOG091G048B, EOG091G0495, EOG091G04JK, EOG091G04MA, EOG091G05D7, EOG091G05K7, EOG091G062X, EOG091G066W, EOG091G06CB, EOG091G06DH, EOG091G074J, EOG091G07PH, EOG091G07VQ, EOG091G087E, EOG091G08DN, EOG091G08IJ, EOG091G08JG, EOG091G090D, EOG091G099D, EOG091G09IW, EOG091G09J7, EOG091G09KX, EOG091G09QO, EOG091G09R0, EOG091G0A79, EOG091G0AGM, EOG091G0ARL, EOG091G0AVW, EOG091G0AX1, EOG091G0AZ7, EOG091G0BCO, EOG091G0BDT, EOG091G0BED, EOG091G0BKX, EOG091G0BN4, EOG091G0BNI, EOG091G0BPX, EOG091G0BX3, EOG091G0C2Z, EOG091G0CEG, EOG091G0CEW, EOG091G0CLR, EOG091G0CXE, EOG091G0D0D, EOG091G0D8N, EOG091G0DQS, EOG091G0DYU, EOG091G0E5P, EOG091G0EA2, EOG091G0EBA, EOG091G0EHF, EOG091G0EHY, EOG091G0EM6, EOG091G0EO4, EOG091G0F93, EOG091G0FH3, EOG091G0FOR, EOG091G0FSK, EOG091G0FYA, EOG091G0FZN, EOG091G0G1G, EOG091G0G58, EOG091G0G8S, EOG091G0GA7, EOG091G0GA8, EOG091G0GD0, EOG091G0GI9, EOG091G0GKX, EOG091G0H6J, EOG091G0HBC, EOG091G0HDF, EOG091G0HMT, EOG091G0HP9, EOG091G0HQ3, EOG091G0HSB, EOG091G0HT4, EOG091G0I1O, EOG091G0I5H, EOG091G0I7M, EOG091G0IDS, EOG091G0IHL, EOG091G0ILG, EOG091G0IMB, EOG091G0IMD, EOG091G0IS6, EOG091G0IXY, EOG091G0J14, EOG091G0J56, EOG091G0J61, EOG091G0JBM, EOG091G0JBZ, EOG091G0JIA, EOG091G0JTR, EOG091G0JUY, EOG091G0JWC, EOG091G0JZV, EOG091G0K0H, EOG091G0K1H, EOG091G0K98, EOG091G0KNT, EOG091G0L2T, EOG091G0L3L, EOG091G0L5X, EOG091G0LB8, EOG091G0LBN, EOG091G0LCT, EOG091G0LKE, EOG091G0LMQ, EOG091G0LMX, EOG091G0LPW, EOG091G0LT7, EOG091G0LTE, EOG091G0LZ0, EOG091G0M09, EOG091G0M0J, EOG091G0M24, EOG091G0M4Q, EOG091G0MBG, EOG091G0MBX, EOG091G0MCM, EOG091G0MCZ, EOG091G0MG7, EOG091G0MS7, EOG091G0MSR, EOG091G0MVR, EOG091G0MZC, EOG091G0NIU, EOG091G0NK3, EOG091G0NKD, EOG091G0NM1, EOG091G0NP6, EOG091G0NRN, EOG091G0NTZ, EOG091G0O4W, EOG091G0O5A, EOG091G0O7O, EOG091G0O97, EOG091G0OFC, EOG091G0OPI, EOG091G0OY0, EOG091G0PD3, EOG091G0PUZ, EOG091G0PVT, EOG091G0Q05, EOG091G0QBN, EOG091G0QE1, EOG091G0QIP, EOG091G0QS6, EOG091G0QZ2, EOG091G0R3S, EOG091G0RA2, EOG091G0RI9, EOG091G0RM8, EOG091G0RRA, EOG091G0RRT, EOG091G0RTI, EOG091G0RVD, EOG091G0RWI, EOG091G0S2R, EOG091G0S5L, EOG091G0SAU, EOG091G0SB2, EOG091G0SCV, EOG091G0SGT, EOG091G0SRJ, EOG091G0T3D, EOG091G0T3X, EOG091G0T5O, EOG091G0TAC, EOG091G0TAU, EOG091G0TCJ, EOG091G0TZU, EOG091G0U3W, EOG091G0UE6, EOG091G0UOM, EOG091G0UQ0, EOG091G0UTL, EOG091G0UUT, EOG091G0UZ0, EOG091G0V3C, EOG091G0VIF, EOG091G0VMU, EOG091G0VSN, EOG091G0VVZ, EOG091G0W0U, EOG091G0W26, EOG091G0W86, EOG091G0WCB, EOG091G0WUF, EOG091G0X1V, EOG091G0X3B, EOG091G0XKP, EOG091G0XN1, EOG091G0XQ3, EOG091G0XRQ, EOG091G0XXM, EOG091G0Y05, EOG091G0Y35, EOG091G0Y6M, EOG091G0Y96, EOG091G0YBP, EOG091G0YUO, EOG091G0YY3, EOG091G0YYA, EOG091G0Z43, EOG091G0Z6N, EOG091G0ZEJ, EOG091G0ZNH, EOG091G0ZTA, EOG091G1081, EOG091G10IX, EOG091G10MO, EOG091G10SJ, EOG091G116R, EOG091G11PB, EOG091G126L, EOG091G141O, EOG091G146E, EOG091G14DH, EOG091G1549, EOG091G15HV, EOG091G15KV, EOG091G15XS, EOG091G1757, EOG091G17I3, EOG091G17K3, EOG091G184V, EOG091G18B1, EOG091G18BK, EOG091G18Z5, EOG091G1A3H

Appendix 2: List of missing BUSCO genes in OKI2018_I69, OdB3 and OSKA2016 genome assemblies (continued).

Genome	Missing BUSCO genes
OSKA2016	EOG091G01MK, EOG091G020E, EOG091G02E4, EOG091G02NO, EOG091G02OE, EOG091G02WU, EOG091G02ZI, EOG091G03H4, EOG091G03JF, EOG091G03JG, EOG091G048B, EOG091G0495, EOG091G04JK, EOG091G04MA, EOG091G05D7, EOG091G05IH, EOG091G05K7, EOG091G05RA, EOG091G060D, EOG091G066W, EOG091G074J, EOG091G07PH, EOG091G07VQ, EOG091G087E, EOG091G08DN, EOG091G08IJ, EOG091G08JG, EOG091G090D, EOG091G095O, EOG091G099D, EOG091G09IW, EOG091G09J7, EOG091G09QO, EOG091G09R0, EOG091G09SS, EOG091G0ARL, EOG091G0AVW, EOG091G0AX1, EOG091G0AZ7, EOG091G0BDT, EOG091G0BED, EOG091G0BKX, EOG091G0BN4, EOG091G0BNI, EOG091G0BPX, EOG091G0BX3, EOG091G0C2Z, EOG091G0CLR, EOG091G0CXE, EOG091G0D0D, EOG091G0DQS, EOG091G0DVQ, EOG091G0DYU, EOG091G0E11, EOG091G0E5P, EOG091G0EA2, EOG091G0EBA, EOG091G0EHF, EOG091G0EHY, EOG091G0EM6, EOG091G0EO4, EOG091G0EZ8, EOG091G0F93, EOG091G0FH3, EOG091G0FOR, EOG091G0FSK, EOG091G0FYA, EOG091G0FZN, EOG091G0G1G, EOG091G0G58, EOG091G0G8S, EOG091G0GA7, EOG091G0GA8, EOG091G0GD0, EOG091G0GKX, EOG091G0H6J, EOG091G0HBC, EOG091G0HDF, EOG091G0HMT, EOG091G0HP9, EOG091G0HQ3, EOG091G0HS8, EOG091G0HSB, EOG091G0I1O, EOG091G0I5H, EOG091G0I7M, EOG091G0IDS, EOG091G0IHL, EOG091G0IMB, EOG091G0IMI, EOG091G0IS6, EOG091G0IXY, EOG091G0IZD, EOG091G0J61, EOG091G0JBZ, EOG091G0JC7, EOG091G0JIA, EOG091G0JTR, EOG091G0JWC, EOG091G0JZV, EOG091G0K0H, EOG091G0K1H, EOG091G0KNT, EOG091G0L2T, EOG091G0L3L, EOG091G0L5X, EOG091G0LBN, EOG091G0LCT, EOG091G0LKE, EOG091G0LMQ, EOG091G0LMX, EOG091G0LPW, EOG091G0LT7, EOG091G0LTE, EOG091G0M09, EOG091G0M0J, EOG091G0M24, EOG091G0M4Q, EOG091G0MBG, EOG091G0MBX, EOG091G0MCM, EOG091G0MCZ, EOG091G0MG7, EOG091G0MS7, EOG091G0MSR, EOG091G0NIU, EOG091G0NKD, EOG091G0NM1, EOG091G0NP6, EOG091G0NP7, EOG091G0NRN, EOG091G0NTZ, EOG091G0O4W, EOG091G0O5A, EOG091G0O7O, EOG091G0O97, EOG091G0OFC, EOG091G0OPI, EOG091G0OWC, EOG091G0OY0, EOG091G0PAG, EOG091G0PD3, EOG091G0PIU, EOG091G0PVT, EOG091G0Q05, EOG091G0QCX, EOG091G0QE1, EOG091G0QS6, EOG091G0QZ2, EOG091G0R3S, EOG091G0R9X, EOG091G0RHX, EOG091G0RI9, EOG091G0RRA, EOG091G0RRT, EOG091G0RTI, EOG091G0RWI, EOG091G0S2R, EOG091G0S5L, EOG091G0SAU, EOG091G0SB2, EOG091G0SCV, EOG091G0SGT, EOG091G0SRJ, EOG091G0T3D, EOG091G0T3X, EOG091G0T5O, EOG091G0TAU, EOG091G0TCJ, EOG091G0TM2, EOG091G0TN4, EOG091G0TZU, EOG091G0U2U, EOG091G0U3W, EOG091G0UOM, EOG091G0UQ0, EOG091G0UTL, EOG091G0UUT, EOG091G0UZ0, EOG091G0V3C, EOG091G0VIF, EOG091G0VSN, EOG091G0VVZ, EOG091G0W0U, EOG091G0W86, EOG091G0WCB, EOG091G0WPV, EOG091G0WUF, EOG091G0X1V, EOG091G0X3B, EOG091G0XKP, EOG091G0XN1, EOG091G0XRQ, EOG091G0Y05, EOG091G0Y35, EOG091G0Y6M, EOG091G0Y96, EOG091G0YBP, EOG091G0YUO, EOG091G0YY3, EOG091G0YYA, EOG091G0Z43, EOG091G0Z7J, EOG091G0ZEJ, EOG091G0ZNH, EOG091G0ZTA, EOG091G10IX, EOG091G10MO, EOG091G10SJ, EOG091G116R, EOG091G11PB, EOG091G11QL, EOG091G126L, EOG091G126R, EOG091G1285, EOG091G141O, EOG091G14DH, EOG091G1549, EOG091G15HV, EOG091G15KV, EOG091G15XS, EOG091G1757, EOG091G17I3, EOG091G184V, EOG091G18B1, EOG091G18BK, EOG091G1A3H, EOG091G1AP6

Appendix 3: Proportions of genes in orthogroups between pairs of vertebrate species.

	Number of genes	*Danio rerio*	*Xenopus tropicalis*	*Gallus gallus*	*Mus musculus*	*Homo sapiens*
Danio rerio	25616		19988 (~78%)	17908 (~70%)	19187 (~75%)	19071 (~75%)
Xenopus tropicalis	22369	17279 (~77%)		15078 (~67%)	16652 (~74%)	16750 (~75%)
Gallus gallus	17980	15062 (~84%)	15119 (~84%)		15234 (~85%)	15383 (~86%)
Mus musculus	21944	15760 (~72%)	16227 (~74%)	15134 (~69%)		19426 (~89%)
Homo sapiens	20501	15553 (~76%)	15880 (~78%)	14798 (~72%)	18666 (~91%)	

Appendix 4: Proportions of genes with one-to-one orthologous relationships between pairs of vertebrate species.

	Number of genes	*Danio rerio*	*Xenopus tropicalis*	*Gallus gallus*	*Mus musculus*	*Homo sapiens*
Danio rerio	25616		9209 (~36%)	8263 (~32%)	9546 (~37%)	9555 (~37%)
Xenopus tropicalis	22369	9209 (~41%)		10647 (~48%)	12404 (~56%)	12412 (~56%)
Gallus gallus	17980	8263 (~46%)	10647 (~60%)		11580 (~64%)	11617 (~65%)
Mus musculus	21944	9546 (~44%)	12404 (~57%)	11580 (~53%)		16183 (~74%)
Homo sapiens	20501	9555 (~47%)	12412 (~61%)	11617 (~57%)	16183 (~79%)	

Appendix 5a: Genes in operons in the Okinawa genome, top 50 terms in Gene Ontology Biological Process.

	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
1	GO:0006396	5.44E-13	6.12	50.03	83	96	RNA processing
2	GO:0006412	2.29E-10	3.55	66.18	100	127	translation
3	GO:0043604	3.38E-10	3.46	67.22	101	129	amide biosynthetic process
4	GO:0043043	3.38E-10	3.46	67.22	101	129	peptide biosynthetic process
5	GO:0043603	1.64E-08	2.65	82.34	116	158	cellular amide metabolic process
6	GO:0006518	3.41E-08	2.60	81.30	114	156	peptide metabolic process
7	GO:0034660	4.73E-08	5.09	33.35	54	64	ncRNA metabolic process
8	GO:0071840	1.64E-07	2.33	91.20	124	175	cellular component organization or biogenesis
9	GO:0044085	1.12E-06	3.16	45.34	67	87	cellular component biogenesis
10	GO:0034470	2.13E-06	6.55	20.84	35	40	ncRNA processing
11	GO:0034641	8.07E-06	1.45	368.95	420	708	cellular nitrogen compound metabolic process
12	GO:0030163	1.03E-05	5.98	19.28	32	37	protein catabolic process
13	GO:1901566	1.67E-05	1.73	140.18	173	269	organonitrogen compound biosynthetic process
14	GO:0044237	2.21E-05	1.34	849.43	908	1630	cellular metabolic process
15	GO:0022613	2.91E-05	5.60	18.24	30	35	ribonucleoprotein complex biogenesis
16	GO:0009987	3.86E-05	1.41	1322.09	1369	2537	cellular process
17	GO:0032259	4.05E-05	17.65	10.42	19	20	methylation
18	GO:0042254	4.12E-05	6.29	16.15	27	31	ribosome biogenesis
19	GO:0044257	4.85E-05	5.41	17.72	29	34	cellular protein catabolic process
20	GO:0051603	4.85E-05	5.41	17.72	29	34	proteolysis involved in cellular protein catabolic process
21	GO:0044267	6.67E-05	1.38	402.31	449	772	cellular protein metabolic process
22	GO:0006399	9.30E-05	3.97	21.89	34	42	tRNA metabolic process
23	GO:0016043	0.000107341	1.95	75.04	97	144	cellular component organization
24	GO:0016072	0.000122716	9.29	11.46	20	22	rRNA metabolic process
25	GO:0006364	0.000122716	9.29	11.46	20	22	rRNA processing

Appendix 5a: Genes in operons in the Okinawa genome, top 50 terms in Gene Ontology Biological Process (continued).

	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
26	GO:0044265	0.000146758	3.85	21.37	33	41	cellular macromolecule catabolic process
27	GO:0051649	0.000187967	2.43	41.17	57	79	establishment of localization in cell
28	GO:0010467	0.000254717	1.40	279.84	317	537	gene expression
29	GO:0006397	0.000368315	4.19	17.20	27	33	mRNA processing
30	GO:0046907	0.000373425	2.34	40.13	55	77	intracellular transport
31	GO:0009451	0.000478425	6.19	11.99	20	23	RNA modification
32	GO:0006974	0.000577465	4.46	15.11	24	29	cellular response to DNA damage stimulus
33	GO:0033554	0.000587267	4.03	16.68	26	32	cellular response to stress
34	GO:0009057	0.000617984	2.87	25.53	37	49	macromolecule catabolic process
35	GO:0051641	0.000667498	2.12	45.86	61	88	cellular localization
36	GO:0016071	0.000803886	3.21	20.84	31	40	mRNA metabolic process
37	GO:0070647	0.000858973	3.37	19.28	29	37	protein modification by small protein conjugation or removal
38	GO:0019941	0.000894451	4.87	13.03	21	25	modification-dependent protein catabolic process
39	GO:0006281	0.000894451	4.87	13.03	21	25	DNA repair
40	GO:0043632	0.000894451	4.87	13.03	21	25	modification-dependent macromolecule catabolic process
41	GO:0006511	0.000894451	4.87	13.03	21	25	ubiquitin-dependent protein catabolic process
42	GO:0043414	0.00148234	12.03	7.30	13	14	macromolecule methylation
43	GO:1901565	0.001951912	2.64	23.97	34	46	organonitrogen compound catabolic process
44	GO:0022607	0.002356956	2.33	29.18	40	56	cellular component assembly
45	GO:0044248	0.002478494	2.23	31.79	43	61	cellular catabolic process
46	GO:0032446	0.002664275	11.10	6.77	12	13	protein modification by small protein conjugation
47	GO:0001510	0.002805539	Inf	4.69	9	9	RNA methylation
48	GO:0043933	0.003301694	2.27	28.66	39	55	protein-containing complex subunit organization
49	GO:0006260	0.003301694	2.27	28.66	39	55	DNA replication
50	GO:0044281	0.00341827	1.64	72.96	89	140	small molecule metabolic process

Appendix 5b: Genes in operons in the Osaka genome, top 50 terms in Gene Ontology Biological Process.

	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
1	GO:0006996	1.18E-09	4.28	43.52	71	89	organelle organization
2	GO:0006396	5.02E-08	3.39	46.45	72	95	RNA processing
3	GO:0071840	9.31E-08	2.32	87.04	121	178	cellular component organization or biogenesis
4	GO:0044248	3.07E-07	4.21	31.30	51	64	cellular catabolic process
5	GO:0009056	3.25E-07	3.39	40.59	63	83	catabolic process
6	GO:0016192	3.34E-07	4.38	29.83	49	61	vesicle-mediated transport
7	GO:0034641	4.78E-07	1.54	335.94	393	687	cellular nitrogen compound metabolic process
8	GO:0009987	6.69E-07	1.54	1183.35	1239	2420	cellular process
9	GO:0016043	7.50E-07	2.36	71.39	100	146	cellular component organization
10	GO:0006412	2.06E-06	2.45	60.63	86	124	translation
11	GO:0043604	2.41E-06	2.41	61.61	87	126	amide biosynthetic process
12	GO:0043043	2.41E-06	2.41	61.61	87	126	peptide biosynthetic process
13	GO:1901575	7.51E-06	3.00	37.16	56	76	organic substance catabolic process
14	GO:0051276	7.71E-06	3.98	25.43	41	52	chromosome organization
15	GO:0030163	1.78E-05	4.86	19.07	32	39	protein catabolic process
16	GO:0010467	2.48E-05	1.48	256.23	299	524	gene expression
17	GO:0044265	2.60E-05	4.14	21.52	35	44	cellular macromolecule catabolic process
18	GO:0048193	4.33E-05	Inf	6.85	14	14	Golgi vesicle transport
19	GO:0009057	4.67E-05	3.46	24.94	39	51	macromolecule catabolic process
20	GO:0044257	5.15E-05	4.55	18.09	30	37	cellular protein catabolic process
21	GO:0051603	5.15E-05	4.55	18.09	30	37	proteolysis involved in cellular protein catabolic process
22	GO:1901565	5.42E-05	3.58	23.47	37	48	organonitrogen compound catabolic process
23	GO:0006518	7.20E-05	1.94	73.84	97	151	peptide metabolic process
24	GO:0043603	7.82E-05	1.92	74.82	98	153	cellular amide metabolic process
25	GO:1901360	9.06E-05	1.42	277.26	318	567	organic cyclic compound metabolic process

Appendix 5b: Genes in operons in the Osaka genome, top 50 GO terms in Gene Ontology Biological Process (continued).

	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
26	GO:0034622	9.29E-05	4.77	16.14	27	33	cellular protein-containing complex assembly
27	GO:0016071	0.000115552	3.77	20.05	32	41	mRNA metabolic process
28	GO:0046483	0.000152286	1.41	274.81	314	562	heterocycle metabolic process
29	GO:0043632	0.000166272	5.08	14.18	24	29	modification-dependent macromolecule catabolic process
30	GO:0019941	0.000166272	5.08	14.18	24	29	modification-dependent protein catabolic process
31	GO:0006511	0.000166272	5.08	14.18	24	29	ubiquitin-dependent protein catabolic process
32	GO:0006725	0.000181183	1.40	274.32	313	561	cellular aromatic compound metabolic process
33	GO:0051641	0.000218505	2.35	39.12	55	80	cellular localization
34	GO:0006397	0.000241704	4.09	16.63	27	34	mRNA processing
35	GO:0051649	0.00024418	2.49	34.23	49	70	establishment of localization in cell
36	GO:0044085	0.000272398	2.14	46.94	64	96	cellular component biogenesis
37	GO:0046907	0.000357255	2.43	33.74	48	69	intracellular transport
38	GO:0006352	0.000850499	7.91	8.31	15	17	DNA-templated transcription, initiation
39	GO:0006139	0.001253392	1.34	266.50	299	545	nucleobase-containing compound metabolic process
40	GO:0008380	0.001553166	7.37	7.82	14	16	RNA splicing
41	GO:0006366	0.001717609	3.47	14.67	23	30	transcription by RNA polymerase II
42	GO:0065003	0.001797489	2.37	26.89	38	55	protein-containing complex assembly
43	GO:0045184	0.001972057	2.12	33.74	46	69	establishment of protein localization
44	GO:0022607	0.002181593	2.17	31.30	43	64	cellular component assembly
45	GO:0043933	0.002393772	2.23	28.85	40	59	protein-containing complex subunit organization
46	GO:0032259	0.002790174	5.27	8.80	15	18	methylation
47	GO:0030029	0.002814868	6.84	7.33	13	15	actin filament-based process
48	GO:0030036	0.002814868	6.84	7.33	13	15	actin cytoskeleton organization
49	GO:0006888	0.003237963	Inf	3.91	8	8	endoplasmic reticulum to Golgi vesicle-mediated transport
50	GO:0071103	0.004359101	4.21	9.78	16	20	DNA conformation change

Appendix 5c: Genes in operons in the Barcelona genome, top 50 GO terms in Gene Ontology Biological Process.

	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
1	GO:0006396	2.48E-14	7.08	4.57E+01	80	92	RNA processing
2	GO:0043043	1.52E-11	3.87	6.21E+01	98	125	peptide biosynthetic process
3	GO:0043604	1.52E-11	3.87	6.21E+01	98	125	amide biosynthetic process
4	GO:0006412	3.99E-11	3.78	6.11E+01	96	123	translation
5	GO:0006518	6.36E-10	3.01	7.40E+01	110	149	peptide metabolic process
6	GO:0034641	7.68E-10	1.71	3.34E+02	403	672	cellular nitrogen compound metabolic process
7	GO:0043603	8.18E-10	2.96	7.50E+01	111	151	cellular amide metabolic process
8	GO:0071840	9.83E-10	2.70	8.75E+01	126	176	cellular component organization or biogenesis
9	GO:0034660	4.57E-08	5.55	2.83E+01	48	57	ncRNA metabolic process
10	GO:0044085	8.32E-08	3.38	4.62E+01	71	93	cellular component biogenesis
11	GO:0009987	1.73E-07	1.60	1.15E+03	1206	2313	cellular process
12	GO:0010467	5.22E-07	1.62	2.54E+02	305	512	gene expression
13	GO:0022613	1.56E-06	6.81	1.89E+01	33	38	ribonucleoprotein complex biogenesis
14	GO:0034470	2.34E-06	7.73	1.69E+01	30	34	ncRNA processing
15	GO:0016043	3.29E-06	2.27	7.06E+01	97	142	cellular component organization
16	GO:0042254	8.59E-06	6.18	1.74E+01	30	35	ribosome biogenesis
17	GO:1901566	1.03E-05	1.80	1.22E+02	154	245	organonitrogen compound biosynthetic process
18	GO:0006996	1.28E-05	2.66	4.42E+01	64	89	organelle organization
19	GO:0044237	1.44E-05	1.37	7.39E+02	796	1487	cellular metabolic process
20	GO:0016071	1.56E-05	4.51	2.14E+01	35	43	mRNA metabolic process
21	GO:0006364	1.70E-05	19.48	9.94E+00	19	20	rRNA processing
22	GO:0016072	1.70E-05	19.48	9.94E+00	19	20	rRNA metabolic process
23	GO:0044267	2.24E-05	1.43	3.50E+02	398	705	cellular protein metabolic process
24	GO:0006397	2.71E-05	5.15	1.79E+01	30	36	mRNA processing
25	GO:1901360	4.76E-05	1.46	2.73E+02	315	550	organic cyclic compound metabolic process

Appendix 5c: Genes in operons in the Barcelona genome, top 50 GO terms in Gene Ontology Biological Process (continued).

	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
26	GO:0006725	5.51E-05	1.45	2.71E+02	312	545	cellular aromatic compound metabolic process
27	GO:0046483	6.76E-05	1.45	2.71E+02	312	546	heterocycle metabolic process
28	GO:0016192	1.43E-04	2.84	2.98E+01	44	60	vesicle-mediated transport
29	GO:0032446	2.30E-04	15.33	7.95E+00	15	16	protein modification by small protein conjugation
30	GO:0009057	2.37E-04	3.01	2.53E+01	38	51	macromolecule catabolic process
31	GO:0044248	2.69E-04	2.67	3.03E+01	44	61	cellular catabolic process
32	GO:0030163	2.71E-04	3.54	1.99E+01	31	40	protein catabolic process
33	GO:0006399	3.05E-04	3.72	1.84E+01	29	37	tRNA metabolic process
34	GO:0051276	3.62E-04	2.93	2.48E+01	37	50	chromosome organization
35	GO:0006139	3.75E-04	1.39	2.64E+02	300	532	nucleobase-containing compound metabolic process
36	GO:0018193	5.03E-04	3.08	2.19E+01	33	44	peptidyl-amino acid modification
37	GO:0044265	5.03E-04	3.08	2.19E+01	33	44	cellular macromolecule catabolic process
38	GO:0051641	5.20E-04	2.23	3.93E+01	54	79	cellular localization
39	GO:0065003	5.51E-04	2.85	2.43E+01	36	49	protein-containing complex assembly
40	GO:0043933	5.78E-04	2.68	2.68E+01	39	54	protein-containing complex subunit organization
41	GO:0006352	5.82E-04	8.18	8.94E+00	16	18	DNA-templated transcription, initiation
42	GO:0022607	5.88E-04	2.55	2.93E+01	42	59	cellular component assembly
43	GO:0018208	6.68E-04	5.12	1.19E+01	20	24	peptidyl-proline modification
44	GO:0000413	6.68E-04	5.12	1.19E+01	20	24	protein peptidyl-prolyl isomerization
45	GO:0051603	7.81E-04	3.46	1.74E+01	27	35	proteolysis involved in cellular protein catabolic process
46	GO:0044257	7.81E-04	3.46	1.74E+01	27	35	cellular protein catabolic process
47	GO:0009056	8.38E-04	2.15	3.98E+01	54	80	catabolic process
48	GO:0034622	8.86E-04	3.66	1.59E+01	25	32	cellular protein-containing complex assembly
49	GO:0006367	9.04E-04	Inf	4.97E+00	10	10	transcription initiation from RNA polymerase II promoter
50	GO:0070647	9.11E-04	3.08	1.99E+01	30	40	protein modification by small protein conjugation or removal

Appendix 6a: Genes out of operons in the Okinawa genome, top 50 terms in Gene Ontology Biological Process.

	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
1	GO:0050789	1.46E-11	1.91	251.89	322	526	regulation of biological process
2	GO:0050794	2.26E-11	1.90	247.10	316	516	regulation of cellular process
3	GO:0065007	8.90E-11	1.83	265.30	334	554	biological regulation
4	GO:0010468	1.91E-10	2.35	120.20	168	251	regulation of gene expression
5	GO:1903506	2.24E-10	2.39	114.45	161	239	regulation of nucleic acid-templated transcription
6	GO:0051252	2.24E-10	2.39	114.45	161	239	regulation of RNA metabolic process
7	GO:2001141	2.24E-10	2.39	114.45	161	239	regulation of RNA biosynthetic process
8	GO:0006355	2.24E-10	2.39	114.45	161	239	regulation of transcription, DNA-templated
9	GO:0019219	2.24E-10	2.39	114.45	161	239	regulation of nucleobase-containing compound metabolic process
10	GO:0009889	4.32E-10	2.28	123.55	171	258	regulation of biosynthetic process
11	GO:2000112	4.32E-10	2.28	123.55	171	258	regulation of cellular macromolecule biosynthetic process
12	GO:0031326	4.32E-10	2.28	123.55	171	258	regulation of cellular biosynthetic process
13	GO:0010556	4.32E-10	2.28	123.55	171	258	regulation of macromolecule biosynthetic process
14	GO:0080090	1.04E-09	2.26	120.20	166	251	regulation of primary metabolic process
15	GO:0051171	1.04E-09	2.26	120.20	166	251	regulation of nitrogen compound metabolic process
16	GO:0031323	4.52E-09	2.14	126.90	172	265	regulation of cellular metabolic process
17	GO:0060255	1.34E-08	2.06	130.73	175	273	regulation of macromolecule metabolic process
18	GO:0019222	1.34E-08	2.06	130.73	175	273	regulation of metabolic process
19	GO:0006814	1.91E-06	5.54	20.11	35	42	sodium ion transport
20	GO:0015718	2.67E-06	2.78	45.01	67	94	monocarboxylic acid transport
21	GO:0046942	2.67E-06	2.78	45.01	67	94	carboxylic acid transport
22	GO:0015849	2.67E-06	2.78	45.01	67	94	organic acid transport
23	GO:0015711	4.88E-06	2.68	45.49	67	95	organic anion transport
24	GO:0050482	9.29E-06	2.65	43.58	64	91	arachidonic acid secretion
25	GO:0032309	9.29E-06	2.65	43.58	64	91	icosanoid secretion

Appendix 6a: Genes out of operons in the Okinawa genome, top 50 terms in Gene Ontology Biological Process (continued).

	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
26	GO:1903963	9.29E-06	2.65	43.58	64	91	arachidonate transport
27	GO:0071715	9.29E-06	2.65	43.58	64	91	icosanoid transport
28	GO:0015908	9.29E-06	2.65	43.58	64	91	fatty acid transport
29	GO:0015909	9.29E-06	2.65	43.58	64	91	long-chain fatty acid transport
30	GO:0007154	1.79E-05	1.78	113.02	144	236	cell communication
31	GO:0007165	4.83E-05	1.73	109.18	138	228	signal transduction
32	GO:0007186	5.83E-05	2.48	40.23	58	84	G protein-coupled receptor signaling pathway
33	GO:0023052	6.54E-05	1.71	109.66	138	229	signaling
34	GO:0097659	8.86E-05	1.59	142.71	174	298	nucleic acid-templated transcription
35	GO:0006351	8.86E-05	1.59	142.71	174	298	transcription, DNA-templated
36	GO:0046903	0.000159619	2.16	47.89	66	100	secretion
37	GO:0006869	0.000184537	2.06	53.16	72	111	lipid transport
38	GO:0032774	0.000190958	1.55	145.10	175	303	RNA biosynthetic process
39	GO:0010876	0.000301878	1.98	54.59	73	114	lipid localization
40	GO:0006811	0.000369572	1.40	232.26	267	485	ion transport
41	GO:0006810	0.000787624	1.30	369.21	408	771	transport
42	GO:0051234	0.001188506	1.29	371.61	409	776	establishment of localization
43	GO:0006644	0.001740985	1.78	56.03	72	117	phospholipid metabolic process
44	GO:0051179	0.001749616	1.27	376.88	413	787	localization
45	GO:0044255	0.001963372	1.58	86.68	106	181	cellular lipid metabolic process
46	GO:0006629	0.002081714	1.50	108.71	130	227	lipid metabolic process
47	GO:0034654	0.002737537	1.39	161.38	186	337	nucleobase-containing compound biosynthetic process
48	GO:0051716	0.006926142	1.39	123.55	143	258	cellular response to stimulus
49	GO:0055085	0.006981211	1.31	184.85	208	386	transmembrane transport
50	GO:0050896	0.007170213	1.38	127.38	147	266	response to stimulus

Appendix 6b: Genes out of operons in the Osaka genome, top 50 terms in Gene Ontology Biological Process.

	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
1	GO:0044255	1.05E-07	2.34	90.45	124	177	cellular lipid metabolic process
2	GO:0015711	1.65E-07	3.34	47.01	71	92	organic anion transport
3	GO:0006629	2.09E-07	2.13	108.33	144	212	lipid metabolic process
4	GO:0046942	2.55E-07	3.29	46.50	70	91	carboxylic acid transport
5	GO:0015718	2.55E-07	3.29	46.50	70	91	monocarboxylic acid transport
6	GO:0015849	2.55E-07	3.29	46.50	70	91	organic acid transport
7	GO:0032309	2.79E-07	3.35	44.97	68	88	icosanoid secretion
8	GO:0050482	2.79E-07	3.35	44.97	68	88	arachidonic acid secretion
9	GO:1903963	2.79E-07	3.35	44.97	68	88	arachidonate transport
10	GO:0071715	2.79E-07	3.35	44.97	68	88	icosanoid transport
11	GO:0015908	2.79E-07	3.35	44.97	68	88	fatty acid transport
12	GO:0015909	2.79E-07	3.35	44.97	68	88	long-chain fatty acid transport
13	GO:0010876	9.93E-07	2.82	53.15	77	104	lipid localization
14	GO:0006869	1.47E-06	2.78	52.63	76	103	lipid transport
15	GO:0010556	1.60E-06	1.89	129.29	165	253	regulation of macromolecule biosynthetic process
16	GO:0031326	1.60E-06	1.89	129.29	165	253	regulation of cellular biosynthetic process
17	GO:0009889	1.60E-06	1.89	129.29	165	253	regulation of biosynthetic process
18	GO:2000112	1.60E-06	1.89	129.29	165	253	regulation of cellular macromolecule biosynthetic process
19	GO:0065007	2.73E-06	1.54	284.12	333	556	biological regulation
20	GO:0050794	3.46E-06	1.55	267.77	315	524	regulation of cellular process
21	GO:0050789	4.27E-06	1.54	272.88	320	534	regulation of biological process
22	GO:0019219	4.70E-06	1.85	123.66	157	242	regulation of nucleobase-containing compound metabolic process
23	GO:2001141	4.70E-06	1.85	123.66	157	242	regulation of RNA biosynthetic process
24	GO:1903506	4.70E-06	1.85	123.66	157	242	regulation of nucleic acid-templated transcription
25	GO:0051252	4.70E-06	1.85	123.66	157	242	regulation of RNA metabolic process

Appendix 6b: Genes out of operons in the Osaka genome, top 50 terms in Gene Ontology Biological Process (continued).

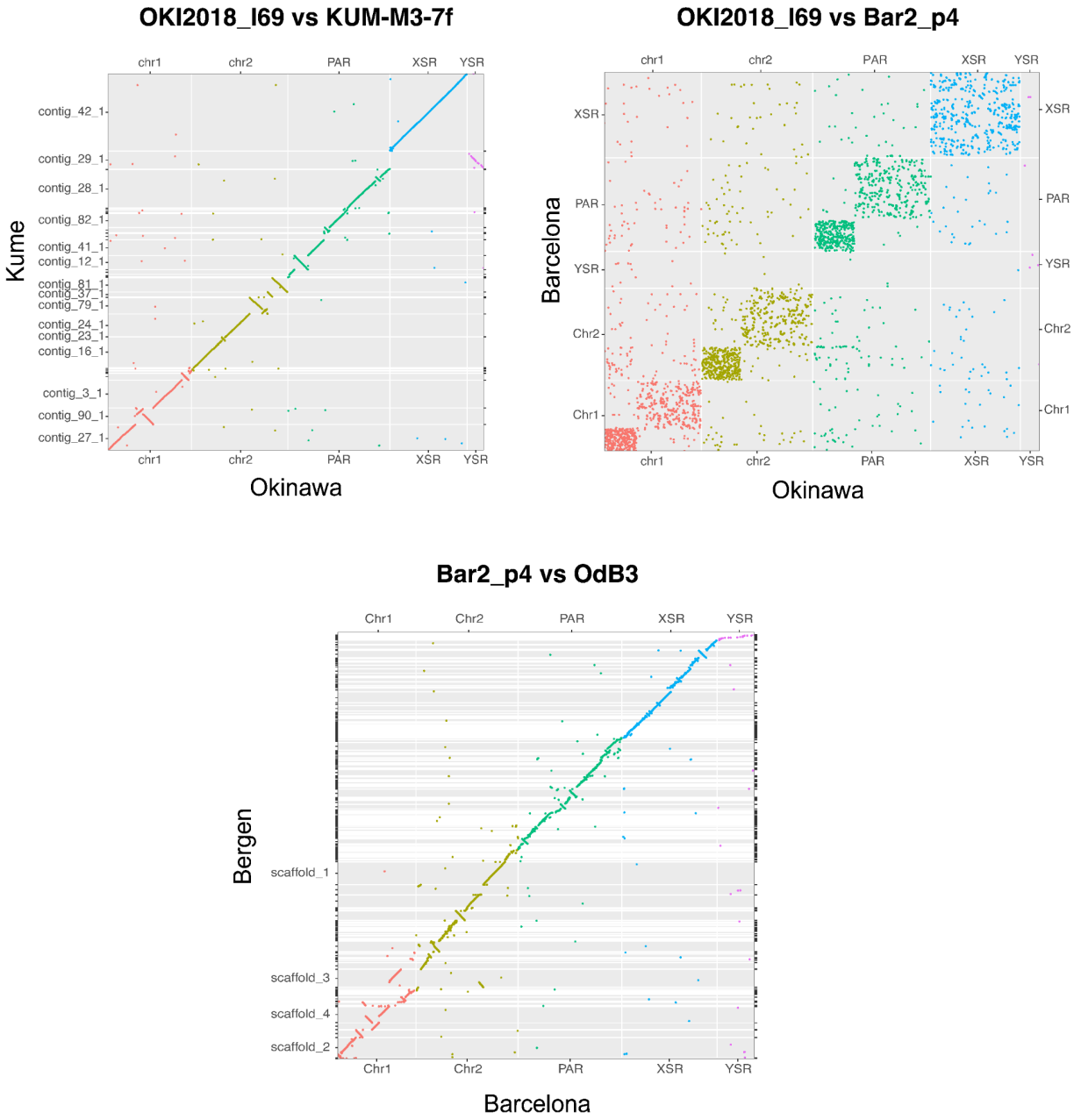
	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
26	GO:0006355	4.70E-06	1.85	123.66	157	242	regulation of transcription, DNA-templated
27	GO:0031323	5.67E-06	1.80	132.86	167	260	regulation of cellular metabolic process
28	GO:0010468	6.19E-06	1.81	130.31	164	255	regulation of gene expression
29	GO:0046903	6.84E-06	2.63	50.59	72	99	secretion
30	GO:0080090	1.13E-05	1.78	130.31	163	255	regulation of primary metabolic process
31	GO:0051171	1.13E-05	1.78	130.31	163	255	regulation of nitrogen compound metabolic process
32	GO:0019222	1.41E-05	1.74	136.95	170	268	regulation of metabolic process
33	GO:0060255	1.41E-05	1.74	136.95	170	268	regulation of macromolecule metabolic process
34	GO:1901137	2.10E-05	2.00	83.29	109	163	carbohydrate derivative biosynthetic process
35	GO:0006644	5.38E-05	2.23	56.72	77	111	phospholipid metabolic process
36	GO:1901135	0.000106815	1.77	99.65	125	195	carbohydrate derivative metabolic process
37	GO:0006643	0.000327037	2.68	30.66	44	60	membrane lipid metabolic process
38	GO:0070085	0.000393355	2.16	45.99	62	90	glycosylation
39	GO:0006486	0.000393355	2.16	45.99	62	90	protein glycosylation
40	GO:0043413	0.000393355	2.16	45.99	62	90	macromolecule glycosylation
41	GO:0009101	0.000393355	2.16	45.99	62	90	glycoprotein biosynthetic process
42	GO:0006664	0.000478859	2.62	30.15	43	59	glycolipid metabolic process
43	GO:0009247	0.000478859	2.62	30.15	43	59	glycolipid biosynthetic process
44	GO:1903509	0.000478859	2.62	30.15	43	59	liposaccharide metabolic process
45	GO:0046467	0.000478859	2.62	30.15	43	59	membrane lipid biosynthetic process
46	GO:0009100	0.000627242	2.09	46.50	62	91	glycoprotein metabolic process
47	GO:0006814	0.002096242	3.15	17.37	26	34	sodium ion transport
48	GO:0008610	0.002503555	1.95	42.92	56	84	lipid biosynthetic process
49	GO:0007186	0.004696902	1.78	49.06	62	96	G protein-coupled receptor signaling pathway
50	GO:0006811	0.004991203	1.31	227.40	253	445	ion transport

Appendix 6c: Genes out of operons in the Barcelona genome, top 50 terms in Gene Ontology Biological Process.

	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
1	GO:0050794	2.43E-10	1.89	2.41E+02	304	480	regulation of cellular process
2	GO:0009889	3.00E-10	2.38	1.23E+02	169	244	regulation of biosynthetic process
3	GO:0010556	3.00E-10	2.38	1.23E+02	169	244	regulation of macromolecule biosynthetic process
4	GO:0031326	3.00E-10	2.38	1.23E+02	169	244	regulation of cellular biosynthetic process
5	GO:2000112	3.00E-10	2.38	1.23E+02	169	244	regulation of cellular macromolecule biosynthetic process
6	GO:0050789	3.46E-10	1.87	2.47E+02	309	490	regulation of biological process
7	GO:1903506	1.59E-09	2.34	1.17E+02	160	232	regulation of nucleic acid-templated transcription
8	GO:0019219	1.59E-09	2.34	1.17E+02	160	232	reg. of nucleobase-containing compound metabolic process
9	GO:0051252	1.59E-09	2.34	1.17E+02	160	232	regulation of RNA metabolic process
10	GO:0006355	1.59E-09	2.34	1.17E+02	160	232	regulation of transcription, DNA-templated
11	GO:2001141	1.59E-09	2.34	1.17E+02	160	232	regulation of RNA biosynthetic process
12	GO:0031323	1.82E-09	2.26	1.26E+02	171	251	regulation of cellular metabolic process
13	GO:0065007	1.82E-09	1.80	2.58E+02	319	513	biological regulation
14	GO:0010468	2.73E-09	2.26	1.23E+02	167	245	regulation of gene expression
15	GO:0080090	6.17E-09	2.21	1.23E+02	166	245	regulation of primary metabolic process
16	GO:0051171	6.17E-09	2.21	1.23E+02	166	245	regulation of nitrogen compound metabolic process
17	GO:0060255	6.92E-09	2.16	1.30E+02	174	259	regulation of macromolecule metabolic process
18	GO:0019222	6.92E-09	2.16	1.30E+02	174	259	regulation of metabolic process
19	GO:0015711	3.18E-08	3.63	4.58E+01	71	91	organic anion transport
20	GO:0015718	5.05E-08	3.58	4.53E+01	70	90	monocarboxylic acid transport
21	GO:0015849	5.05E-08	3.58	4.53E+01	70	90	organic acid transport
22	GO:0046942	5.05E-08	3.58	4.53E+01	70	90	carboxylic acid transport
23	GO:0032309	1.98E-07	3.42	4.38E+01	67	87	icosanoid secretion
24	GO:0050482	1.98E-07	3.42	4.38E+01	67	87	arachidonic acid secretion
25	GO:0015908	1.98E-07	3.42	4.38E+01	67	87	fatty acid transport

Appendix 6c: Genes out of operons in the Barcelona genome, top 50 terms in Gene Ontology Biological Process (continued).

	Gene Ontology ID	p-value	OddsRatio	ExpCount	Count	Size	Term
26	GO:0015909	1.98E-07	3.42	4.38E+01	67	87	long-chain fatty acid transport
27	GO:0071715	1.98E-07	3.42	4.38E+01	67	87	icosanoid transport
28	GO:1903963	1.98E-07	3.42	4.38E+01	67	87	arachidonate transport
29	GO:0007186	8.73E-07	3.84	3.37E+01	53	67	G protein-coupled receptor signaling pathway
30	GO:0046903	3.80E-06	2.74	4.83E+01	70	96	secretion
31	GO:0006869	5.00E-06	2.60	5.18E+01	74	103	lipid transport
32	GO:0010876	8.94E-06	2.51	5.23E+01	74	104	lipid localization
33	GO:0006629	1.23E-05	1.91	9.91E+01	128	197	lipid metabolic process
34	GO:0044255	1.26E-05	2.03	8.25E+01	109	164	cellular lipid metabolic process
35	GO:0007154	5.50E-05	1.78	1.06E+02	133	210	cell communication
36	GO:0006644	1.11E-04	2.11	5.74E+01	77	114	phospholipid metabolic process
37	GO:0006814	1.72E-04	5.01	1.51E+01	25	30	sodium ion transport
38	GO:0098609	2.74E-04	14.96	8.05E+00	15	16	cell-cell adhesion
39	GO:0098742	2.74E-04	14.96	8.05E+00	15	16	cell-cell adhesion via plasma-membrane adhesion molecules
40	GO:0007156	2.74E-04	14.96	8.05E+00	15	16	homophilic cell adhesion via plasma membr. adhesion mol.
41	GO:0006811	2.94E-04	1.45	2.14E+02	247	425	ion transport
42	GO:0007165	3.18E-04	1.67	1.02E+02	126	203	signal transduction
43	GO:0023052	4.28E-04	1.65	1.03E+02	126	204	signaling
44	GO:0006810	5.91E-04	1.34	3.40E+02	377	675	transport
45	GO:0051234	1.11E-03	1.31	3.42E+02	377	679	establishment of localization
46	GO:0051179	1.81E-03	1.29	3.47E+02	381	690	localization
47	GO:0097659	2.25E-03	1.44	1.42E+02	165	282	nucleic acid-templated transcription
48	GO:0006351	2.25E-03	1.44	1.42E+02	165	282	transcription, DNA-templated
49	GO:0055085	2.86E-03	1.39	1.72E+02	196	341	transmembrane transport
50	GO:0006508	3.35E-03	1.35	1.95E+02	220	387	proteolysis



Appendix 7: The rest of the macro-synteny plots built based on one-to-one orthologs between *O. dioica* genomes.