Comparative genomic studies on Dicyema japonicum: the phylogenetic position of dicyemids and the genomic adaptations to parasitic lifestyle

Okinawa Institute of Science and Technology

Graduate University

Thesis submitted for the degree

Doctor of Philosophy

Comparative genomic studies on *Dicyema japonicum*: the phylogenetic position of dicyemids and the genomic adaptations to parasitic lifestyle

by

Tsai-Ming Lu

Supervisor: Noriyuki Satoh

May 2018

Abstract

Parasitism has independently occurred more than 200 times across 15 animal phyla, yet remains a topic of debate how free-living ancestors evolved to parasitic organisms. Dicyemid mesozoans are microscopic endoparasites inhabiting the renal sacs of some cephalopods. They possess simplified body architecture without differentiated organs and have long fascinated biologists because of their incompletely known life cycles. Obtaining genomic data from enigmatic parasites would be essential to better comprehension of evolution of parasitism. Here I reexamined the phylogenetic position of dicyemids using genome-wide data. Dicyemids share a common ancestor with orthonectids endorsing the old Mesozoa clade, and mesozoans have a closer affinity to gastrotrichs and platyhelminthes, rather than mollusks and annelids. Secondly, I decoded the genome of *Dicyema japonicum* which is approximately 68 Mbp with substantially shortened introns. Comparisons of genomic data among bilaterians showed that *D. japonicum* retains fewer genes in most KEGG pathways, as in the case of four parasite species from different phyla that show a convergent gene number reduction in the metabolism pathways. In contrast, *D. japonicum* exhibits multi-copy gene clusters associated with endocytosis and membrane trafficking, perhaps reflecting its specialized nutrient-uptake strategy. Up-regulated transcripts at dispersal larvae stage indicate over-representation of gene ontology terms of motor activity and response to the stimulus. The occurrences of neurotransmitters and neuropeptides on apical cells of dispersal larva was evident. Taken together, dicyemids may utilize potential sensory functions to detect environmental cues in order to actively approach new hosts. In summary, the dicyemid genome provides a resource to uncover the mysterious life cycle of dicyemids, as well as for applying comparative genomic approaches to gain insights into the evolution of parasitism. Furthermore, genomes of parasites may adapt through eliminates of genes which are not necessary for parasitic lifestyle or through increasing gene copies corresponding to lineage-specific biological processes.

Acknowledgements

I greatly thank my supervisor, Prof. Noriyuki Satoh. He introduced me to the amazing organism, dicyemid mesozoan, and provided all the resources I needed for my thesis work. I appreciate his supervision and proofreading of this thesis as well. I also thank Prof. Hidetaka Furuya for kind hospitality during my sampling trips and the instructions on how to collect dicyemids.

I am grateful to have the support from all members of the Marine Genomics Unit. I especially thank Prof. Chuya Shinzato for the advice on genome assembly and for providing customized Perl scripts. I thank Dr. Jun Inoue for the advice on phylogenetic analyses. I also would like to thank Drs. Keisuke Nakashima, Eric Edsinger, and Yi-Jyun Luo for providing primary antibodies, CellMask, and DAPI reagents.

I would like to acknowledge assistance received from the OIST Imaging and Instrumental Analysis Section, the DNA Sequencing Section, and the Information Service Section for technical support. Additionally, I thank Dr. Miyuki Kanda for her assistance in library preparation and sequencing.

List of Abbreviations

BI	Bayesian inference
BLAST	basic local alignment search tool
BSA	bovine serum albumin
DAPI	4’,6-diamidino-2-phenylindole
DSL	Delta serrate ligand
DBH	dopamine beta-hydroxylase
FGF	fibroblast growth factor
GABA	gamma-aminobutyric acid
GC	guanine-cytosine
GO	gene ontology
GPCR	G protein-coupled receptor
HLH	helix-loop-helix
HMW	high-molecular-weight
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG ortholog
LBA	long-branch attraction
ML	maximum likelihood
mtSSB	mitochondrial single stranded binging protein
NPY	neuropeptide Y
OG	ortholog group
OT	oxytocin
PBS	phosphate-buffered saline
PFA	paraformaldehyde
RRM	RNA recognition motif
TMM	trimmed mean of M-values
TPM	transcripts per kilobase million
TPR	tetratricopeptide repeat
VP	vasopressin

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Spiralian phylogeny

Understanding the origin and evolutionary history of metazoans has been a biological research objective for more than a century, but phylogenetic relationships of many enigmatic taxa remain unsolved. Phylogenomic data from ambiguous taxa are essential for better comprehension of a total scope of animal evolution.

Referring to the ‘‘new animal phylogeny’’ inferred from morphological and molecular data, bilaterians comprise three clades: Deuterostomia, Ecdysozoa, and Spiralia (Giribet et al., 2015). The phyla within the Spiralia exhibit a wide range of variations in development and morphology (Nielsen 2012), and they are previously considered as two major groups, Lophotrochozoa and Platyzoa, mainly based on the morphological traits. The Lophotrochozoa comprises at least annelids, mollusks, and some other animals with a more complex morphology (Halanych 2004), while the Platyzoa includes small-size and simple appearing taxa such as platyhelminthes, gastrotriches, rotifers, etc. (Cavalier-Smith 1998). The earlier scenarios of spiralian evolution remain contentious, and the phylogenetic affiliations of many spiralian phyla remain unclear possibly due to low data coverage and limited taxa sampling (Giribet, 2008; Edgecombe et al., 2011). Even though recent studies use genome-wide data to examine the phylogenetic relationships within the Spiralia, the relationship of taxa involved in this clade is still debated (Nesnidal et al., 2013; Laumer et al., 2015; Luo et al., 2017). In addition, a paucity of morphological synapomorphies and problematic long branches of some taxa caused by unusually fast evolution potentially mislead the interpretation of phylogenetic relationships (Struck et al., 2014). The studies presenting inconsistent results on higher taxonomic grouping, e.g. the status of Platyzoa, Rouphozoa and Gnathifera, make the internal relationships of the Spiralia remain controversial and need to be elucidated further. Some analyses showed that microscopic spiralians, namely Platyhelminthes, Gastrotricha,

Gnathostomulida, and Syndermata (Rotifera + Acanthocephala), grouped into the Platyzoa (Edgecombe et al., 2011; Laumer et al., 2015). On the other hand, other phylogenomic studies support the monophyletic origin of Gnathifera (Syndermata + Gnathostomulida) (Witek et al., 2009) and have proposed that the Spiralia comprises three higher taxonomic units, i.e., the Gnathifera (Syndermata, Gnathostomulida, and Micrognathozoa), the Rouphozoa (Gastrotricha and Platyhelminthes) and the Lophotrochozoa (Mollusca, Annelida, Brachiopoda, Nemertea, etc.) (Laumer et al., 2015). It was further suggested that the Platyzoa would be a paraphyletic clade and the statement of its monophyly possibly represents an artifact due to long-branch attraction (Struck et al., 2014; Laumer et al., 2015). Adding a new taxon into the phylogenetic analysis could possibly lead to striking changes on the interpretation of the phylogenetic relationships between analyzed taxa, especially for the groups whose phylogenetic relationships are confounded by the long branch attraction problem. Crucial for clarifying the Spiralia internal phylogeny is the phylogenetic position of the enigmatic taxon Dicyemida, which has been regarded as microscopic spiralian with extremely simplified body architecture. If the phylogenetic position of the Dicyemida is basal to the remaining taxa of the Spiralia, it would echo the hypothesis of a noncoelomate common ancestor of the spiralians (Halanych, 2004; Hejnol et al., 2009).

Two hypothetical scenarios of spiralian evolutionary history have been proposed. One supports the traditional acoeloid-planuloid hypothesis of a noncoelomate common ancestor of the spiralians, whereas the other suggests that the common ancestor resembled an annelid-like organism with a segmented, coelomate body plan (Struck et al., 2014). As one of the issues, some microscopic lineages remain poorly studied, and thus their phylogenetic analyses may produce erroneous interpretations of spiralian evolution. Nowadays, taking advantage of new generation sequencing technology, phylogenomic approach becomes more popular and has been based on genome-wide information. This methodology provides better comprehension of

spiralian evolution. However, previous phylogenomic analyses have so far neglected a microscopic and parasitic acoelomate lineage, dicyemids, and this may mislead results of phylogenomic analyses of spiralians. Due to their highly simplified acoelomate body plan, the phylogenetic position of dicyemids remains enigmatic. The clarification of the precise position of dicyemids by phylogenomic approach is essential to fully comprehend spiralian evolution and to confirm whether or not dicyemids are morphologically simplified spiralians.

1.2 Evolution of parasitism

Parasitism is a type of symbiotic relationship in which one organism benefits while the other is harmed. However, the usage of terms, parasite and symbiont, might be occasionally ambiguous in some cases, because some symbionts may be incorrectly recognized as parasites due to lack of fundamental knowledge of organisms themselves. Parasites are present in 15 (43%) of the generally recognized 35 animal phyla, and parasitism has independently evolved at least 223 times (Weinstein & Kuris, 2016). This suggests that each parasitism event might reflect the interaction of each host–parasite pair, and the modes of adaptation to parasitic lifestyle might vary case by case. Parasites usually exhibit convergences such as reduction of external morphology, complicated life cycle, and host specificity, reflecting the common selection pressures during the parasitism process. A few molecular convergences have been suggested across parasite lineages; for example, gene expansion associated with the surface modification in relation to the host immune system and the loss of metabolism in flukes and tapeworms were reported (Wang et al., 2011; Tsai et al., 2013). By contrast, in some other cases, adaptations are divergent instead of convergent due to niche diversity. At genomic level, many unique adaptations have been observed in each evolution of flatworm parasitism to their specific niche (Zarowiechi & Berriman, 2014).

The disclosure of the impact of genomic innovation during the parasitism process is critical to understand the evolution of parasitism. The comparative genomic studies between parasites and close-relative non-parasitic species might provide an opportunity to test long-standing hypotheses of genome reduction and molecular adaptations to parasitic life style (Jackson, 2015). This could demonstrate how a new parasitic (or symbiotic) species diverges from a free-living ancestor and what genetic backgrounds are involved in morphological and physiological adaptations to their specific lifestyle.

Dicyemids have been recognized as parasites because of their habitat in the renal sac of host octopus with high population density, almost covering the whole surface of renal folds (Furuya & Tsuneki, 2003). Some species even insert into the tissue of renal folds. However, there is no physiological measure or morphological defect to show that the dense occupancy of dicyemids in the renal sac is really harmful to octopus hosts. In contrast, dicyemids may contribute to the establishment of the acidic condition in the renal sac, that probably facilitates the removal of ammonia from the less acidic blood (Lapan, 1975). Herein, I maintain the usage of the term "parasite" for dicyemids, simply because we still could not exclude the possibility that dicyemids are harmful to octopus hosts based on the evidence up to date. Decoding the *Dicyema japonicum* genome adds to a growing number of phylogenetically disparate genomes for parasitic (or symbiotic) organisms, and may illuminate the convergence of evolution of parasitism (or symbiosis more generally).

1.3 The dicyemid mesozoan

In the 19th century, owing to their simple body plan, dicyemids and orthonectids together were once placed in the old phylum Mesozoa, as intermediates between the Protozoa (unicellular animal-like eukaryotes) and the Metazoa (multicellular animals) (van Beneden et al., 1876). Both dicyemids and orthonectids lack differentiated tissues and consist of only two cell-layers.

In addition, they both possess an interior cavity, in which gametes are formed and embryos develop. They also share similar rootlet structure at the basal body of cilia, suggesting a parasitic stem species common to the Dicyemida and the Orthonectida (Ax, 1996). However, dicyemids inhabit renal sacs of cephalopods, mainly octopuses and cuttlefishes (Figure 1.1), while orthonectids are found in flatworms, polychaetes, molluscs, and echinoderms. Moreover, their differences in the sexual stages suggest that they are not closely related (Stunkard et al., 1954). It was proposed that they are two new phyla, Rhombozoa (Dicyemida) and Orthonectida, replacement of the phylum Mesozoa (Brusca & Brusca, 1990). Although dicyemids and orthonectids have many common features, a question whether Dicyemida and Orthonectida form a monophyletic group is still debated.

Figure 1.1 | Collecting dicyemids from the renal sac of octopus.
The urine of octopus is drawn from the renal sac. The dicyemid individuals at different life-cycle stages are intermixed with octopus blood cells in the urine, and therefore dicyemid individuals could be carefully isolated. N, nematogen; R, rhombogen; RS, renal sac; S, siphon; U, urine. Arrows, infusoriform larvae. Scale bars: 50 μm.

The occurrence of tight junctions between epidermal cells (Furuya et al., 1997) and the release of small polar bodies during meiosis (Furuya & Tsuneki, 2007) support the contention

that the dicyemids belong to the Metazoa. To date, 122 dicyemid species have been described worldwide (Catalano et al., 2014). Dicyemids are highly host-specific (Catalano, 2013), although more than one dicyemid species can inhabit one host individual (Furuya et al., 2003). Dicyemids have long fascinated biologists because of their highly simplified body organization and known but enigmatic life cycles (Katayama et al., 1995) (Figure 1.2). The simplified body architecture of an adult individual consists of approximately 30 cells, lacking digestive tract, coelom, circulatory system, and other differentiated tissues. Its body plan consists of three regions: collate, a central axial cell, and ciliated epidermal cells. This is probably the most extreme case of secondary reduction of body plan complexity in spiralian animals. In detail, the collate region (the most anterior 8 cells) is used to attach on the surface of the octopus renal tissues. The central axial cell is surrounded by an outer layer of ciliated epidermal cells. The central axial cell is the place mainly for reproduction, where vermiform or infusoriform embryos develop. Ciliated epidermal cells are supposed to absorb nutrients directly from the host urine via endocytosis (Ridley, 1968). The cilia appear to generate nutrient-containing water flow toward themselves (Furuya et al., 2003).

In contrast to the simple body plan, the life cycle of dicyemids is characterized by two reproduction modes (asexual and sexual) and there are respective larva and adult stages in each reproduction process (Figure 1.2h). After infecting an octopus host, the germinal cell develops into asexual reproductive adult (nematogen). Then, inside the central axial cell of nematogen, the agamete (axoblast) develops into asexual reproductive vermiform embryos (Furuya et al., 2001). While embryos mature, the vermiform larvae escape from the axial cell of nematogen and develop to a new nematogen to attach on renal tissue of the same host which increases the population density. Once the population density inside a host reaches certain threshold or nematogens receive certain chemical cue from the environment, the stage of sexual reproduction starts by transforming nematogen into sexually reproductive adult (rhombogen)

(Lapan & Morowitz, 1972). Inside the central axial cell of rhombogen, the hermaphroditic gonad (infusorigen) generates sperms and eggs, and gametes fertilize and develop to infusoriform larvae (Furuya et al., 1992). The matured, free-swimming infusoriform larvae break through the axial cell and trunk cells of the rhombogen (Figure 1.2g) and are released into the ocean through the excretion of octopus to infect new hosts.

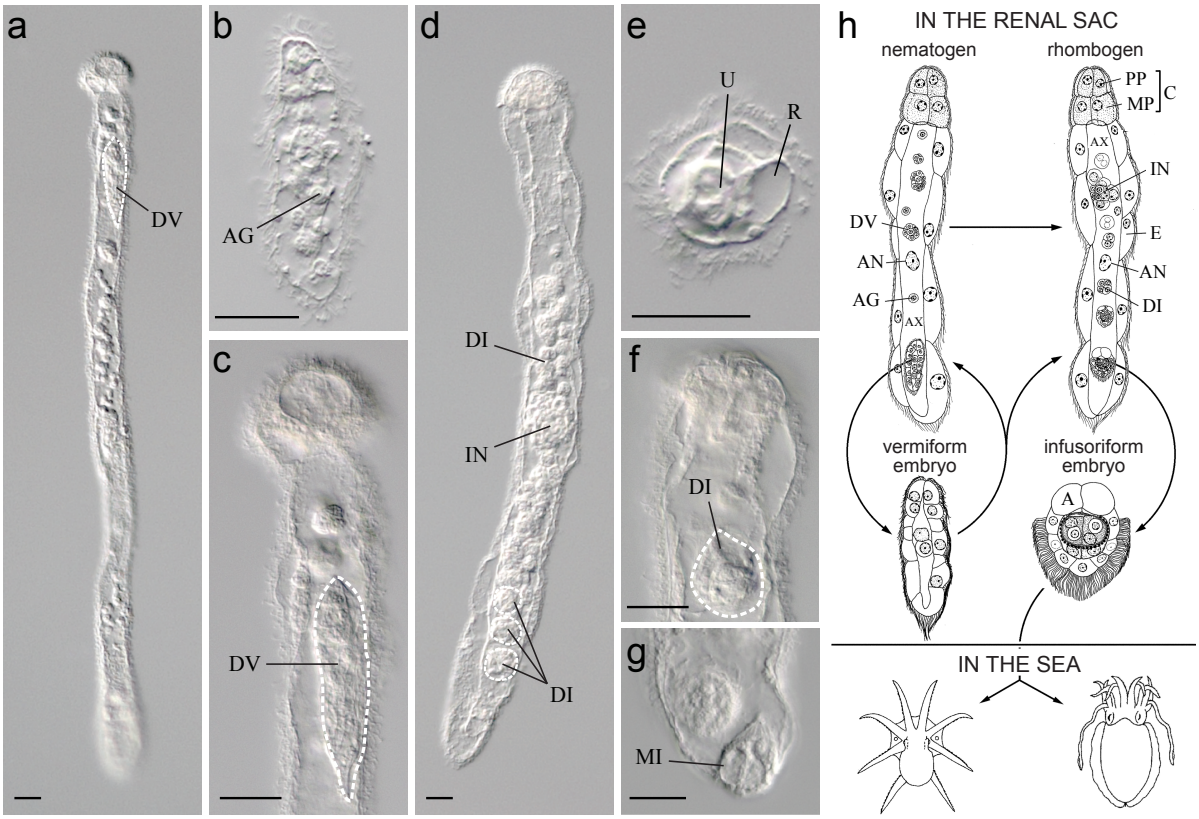

Figure 1.2 | Life-cycle stages of dicyemids.

(a) Nematogen; (b) Vermiform larva; (c) Vermiform embryo develops inside a nematogen; (d) Rhombogen; (e) Infusoriform larva; (f) Infusoriform embryo develops inside a rhombogen; (g) Mature infusoriform larva escapes from the parent rhombogen. (h) Life cycle of dicyemids. Adapted and modified from Furuya & Tsuneki (2003). AG, agamete; AN, axial cell nucleus; AX, axial cell; C, calotte; DI, developing infusoriform embryo; DV, developing vermiform embryo; E, epidermal cell; IN, infusorigen; MI, mature infusoriform larva; MP, metapolar cell; PP, propolar cell. R, refringent body; U, urn cell. Scale bars: 20 μm.

One infusoriform larva of *Dicyema japonicum* comprises 37 cells (Furuya et al., 1992). The infusoriform larva is the most complicated stage and its morphology is distinct from other three stages (Figure 1.2e). In the center of the infusoriform larva, four urn cells exist in a cup-shape cavity which is formed by two capsule cells, opening to the surface on the ventral side (Bresciani & Fenchel, 1967). Each urn cell contains one nucleus and one germinal cell, and germinal cells incorporating into urn cells is considered as an evolutionary foreshadowing of gastrulation (Lapan et al., 1975). Previous studies suggested that infusoriform larvae expel the urn cells after swimming for some time, and later the urn cells decompose releasing the contained germinal cells (McConnaughey et al., 1951). Inside the capsule cells, which are surrounded the urn cells, there are many granules, which are supposed to serve some kind of lytic functions associated with the liberation of urn cells (Ridley et al., 1969). The earliest stage of dicyemids found in the young cephalopod is a stem vermiform, which must derive directly or indirectly from the germinal cell released by infusoriform larva (Dodson et al., 1956; Lapan & Morowitz, 1972). Infusoriform larvae must find the next host and release germinal cells quickly, because they live for only a day or two (Stunkard et al., 1954). However, how an infusoriform larva searches for a new host and what mechanism triggers the liberating of germinal cells are still unclear.

1.4 Uncommon characteristics of dicyemid genome and mitochondrial DNA

Extrachromosomal, circular DNAs have been observed by scanning electron microscope in *D. japonicum* (Noto et al., 2003). They might derive from chromosomes through DNA elimination or selective replication during early embryogenesis. These circular DNAs are highly heterogeneous in sequence, not tandemly repetitive, and unlikely to encode proteins. Reflecting the circular DNA at different developmental stages, intense *in situ* hybridization signals appear in younger embryos but not in mature larvae. It suggests that these circular DNA

may originate during whole genome amplification; then unnecessary sequences are excised from chromosomal DNA and eliminated from the somatic genome as mini-circles (Noto et al., 2003). Another possibility is that selective replication of chromosomal elements that generate circles occurs during the early embryogenesis (Noto et al., 2003). The formation of small circular DNA during development has been described in many animals, e.g., fruitfly (Pont et al. 1987) and mouse (Yamagishi et al. 1983), but most of these small circular DNA are thought to be derived from tandemly repeated sequences, unlike to the case in dicyemids.

In the somatic genome of dicyemids, several hundred copies of various types of repetitive sequences are estimated, and they are expected to account for a relatively large part of the somatic genome (Awata et al., 2007). In addition, these repetitive sequences show different fate of single copy genes. Only the repetitive sequences are selectively amplified via endoreplication in the somatic nuclei, while the single copy genes stop replication after the early embryogenesis and are diluted via cell divisions into differentiated somatic cells (Awata et al., 2007). A survey of 39 dicyemid genes showed that most introns (97.6%, 205/210) are extremely short, with a mean length of 26 bp. This is nearly equal to that of the chlorarachniophyte, *Bigelowiella natans* (18–21 bp), which has the shortest introns of any known eukaryote (Ogina et al., 2010). Otherwise, the intron density (5.3 introns/gene) is similar to that of most invertebrates. This implies that dicyemids may hold a smaller genome size than other spiralians due to the short intron size. The shortening of introns has also been reported in the orthonectids, in which nearly a half of all introns are within a 30–50 bp size range, with a peak of 37 bp at the distribution of intron lengths, although the mean intron length is 342 bp (Mikhailov et al., 2016). This indicates that the shortening of introns could be a molecular convergence in the mesozoan lineages. Ogino et al. (2010) suggested that dicyemid genes show rather high evolutionary rates, despite the conserved exon/intron structure that may have

evolved under functional constraints. The dicyemid genome possesses many unusual properties that may reflect its parasitic life style. Other such attributes will undoubtedly be discovered.

In most metazoans, mitochondrial DNA (mtDNA) molecule is generally a small ring of 15-24 kbp, containing 13 protein coding genes, two ribosomal RNAs and 22 transfer RNAs, with variable gene orders. In dicyemids, three mitochondrial cytochrome oxidase genes (*cox1*, *cox2,* and *cox3*) of *D. misakiense* are encoded on three separate mini-circles (Watanabe et al., 1999). These mini-circles all have relatively long non-coding regions and carry only one gene per mini-circle. According to the results of *in situ* hybridization using probes of *cox1*and large subunit rRNA, mtDNA molecules are more abundant in early embryogenesis stage than stages of mature larva and adult. The copy number of mtDNA may decrease during development, and extremely low copy number of mtDNA remain (or completely absent) in the mature larva or adult cells (Awata et al., 2005). However, it has not been shown yet how mtDNA mini-circles form in dicyemids. The mini-circular mitochondrial genome has also been discovered in parasitic nematodes (Gibson et al., 2007), and the mtDNA circulation appears to arise multiple times within lice group (Shao et al., 2017). The study of mitochondrial replisome components in lice shows that mitochondrial single stranded binding protein (mtSSB) is absent in lice, which may cause the defect of mitochondrial replication leading to the favor of mini-circle formation (Cameron et al., 2011).

Except the mini-circle form of mtDNA, the existence of high-molecular-weight (HMW) mtDNA has been shown in *D. japonicum* by Southern blotting. Nonetheless, this type of mtDNA is probably restricted to germ cells and the minicircles are derived from the HMW mtDNA during the development process (Awata et al., 2005). Recently, eight mitochondrial transcripts of protein-coding genes (*cox1*, *cox2*, *cox3*, *nad1*, *nad3*, *nad4*, *nad5*, and *cob*) have been annotated in the transcriptome of *D. japonicum* (Robertson et al., 2018). However, the gene order and structure of entire dicyemid mitochondrial genome remain to be elucidated. This

information might provide valuable insights into the mechanism of possible HMW mtDNA degradation and mini-circle mtDNA formation.

Summarizing previous studies, dicyemids somehow can eliminate or selectively replicate chromosomes and mitochondrial DNA into small DNA circles during development. These phenomena may be beneficial for dicyemids' parasitic life style through reducing energy consumption during the mitotic DNA replication process. This also implied the situation of genome reduction. The genome-wide survey is required for future studies by comparing genome size, intron and exon size, gene number, and etc. Furthermore, understanding the genomic adaptations of parasitism is of fundamental biological interest and is essential to uncover the molecular mechanisms of parasite lifestyles.

1.5 Genes marker for distinct cell types and body organization

Although either adult or larva of dicyemids consists of less than forty cells without differentiated organs, the somatic cells may serve as diverged cell types corresponding to specific functions. The expression pattern of developmentally critical genes, e.g., transcription factor genes, could possibly distinguish cell types in the extremely simplified body plan of dicyemids. This could further provide molecular cues to speculate the possible evolutionary history of morphology reduction. *Pax6*, a tool-kit gene playing key roles in development of sensory organs, has been reported in dicyemids, implying that tool-kit genes are required even for the highly simplified bilaterians (Aruga et al., 2007). The expression of *brachyury* and *Otx* genes mark the ventral cells of the opening of the urn cavity and the vegetal blastomeres, respectively. The *Otx*-positive cells will invaginate and become germ cells in the infusoriform larva (Kobayashi et al., 2009). This raises a hypothesis that the invagination in the infusoriform larva that forms the germ cells is homologous to the gastrulation of other metazoans that forms a functional gut. The opening of urn cavity is regarded as a position homologous to the

trochophore stomodeum (Kobayashi et al., 2009). In the adult body plan, the anterior calotte region appears to have specific expression of chitinase-like protein (Ogino et al., 2007), while six genes associated with nutrient transportation and lysosomal proteolysis are expressed in the epidermal cells of trunk region (Ogino et al., 2011). This shows that the somatic cells of vermiform adult are physiologically differentiated between the calotte and trunk regions.

Hox genes are transcriptional factor genes and involved in the early development to specify segment identity along anterior-to-posterior axis of *Drosophila* and other animals (Mallo & Alonso, 2013). Hox genes are basically organized as a linked cluster in many animal genomes, although in some cases, it is accompanied with gene order rearrangement (Ikuta, 2011; Baughman et al., 2014). The loss of Hox genes has been reported in some taxa with less complex morphology such as Bryozoa, Rotifera, and orthonectids (Passamaneck & Halanych, 2004; Mikhailov et al., 2016; Frobius & Funch, 2017), which seems congruous with the simplification of body plan. In dicyemids, a Hox gene, *DoxC,* encodes the Lox5 peptide (also called the spiralian peptide), a key signature in spiralians, suggesting that dicyemids are secondarily simplified from higher protostome animals (Kobayashi et al., 1999). The dicyemid *DoxC* is expressed in the truck and tail of developing vermiform embryo, which exhibits a clear anterior boundary (Kobayashi et al., 2009). Since dicyemids have extremely reduced body architecture, it remains undiscovered that, in addition to the *DoxC* gene, which *Hox* genes exist in the dicyemid genome. Furthermore, comparison of the loss of Hox genes among taxa that exhibit simple body plans, would reveal the differences of genetic backgrounds for their simplified body organizations.

1.6 Potential sensory function

Although no nervous system has yet been reported in dicyemids, their behaviors appear to indicate that they utilize certain undescribed sensory machinery to receive signal cues and

respond to the surrounding environments. For example, dicyemids could sense the concentration of chemical signal related to the population density, which is a trigger for the transformation from asexual to sexual reproductive mode (Lapan & Morowitz, 1972). In this experiment, the asexually reproductive adults were incubated with a low population density in the culture medium which was adopted from a culture with a high population density for 48 hours. Within 24 hours, the uncrowded individuals started to produce infusoriform larva. This indicates that the asexually reproductive adults could sense the accumulated chemical factor in the culture medium to trigger the transformation from asexual to sexual reproduction (Lapan & Morowitz, 1972). The chemistry of the octopus urine and renal system has been examined (Lapan, 1975). No candidate chemical relevant to the transformation from asexual to sexual reproduction has been reported yet. Therefore, the molecular mechanism of chemoreception in dicyemids and the chemical which really triggers the switch between the reproductive modes remains to be discovered. In addition, when the infusoriform larvae are released to the open water, an unknown sensory mechanism may play a role in detecting new hosts in the ocean. The sensory function has been reported in non-nervous protists. They conduct sensory activities using chemoreceptors; for instance, *Tetrahymena* and *Paramecium* could utilize chemoreceptors to recognize many odorant compounds, similar to those conserved in higher organisms (Rodgers et al., 2008). In eukaryotes, G protein-coupled receptors (GPCRs) play an important role in sensory functions by receiving multiple stimuli from outside the cell and activating the internal signal transductions. GPCR superfamily consists of hundreds of receptor proteins, and they mostly contain a seven transmembrane domain (Pierce et al., 2002). Dicyemids have complex life cycles, and they may utilize different types of GPCRs to receive environmental signals to trigger the shift of their life-cycle stages. The annotation of GPCRs in dicyemids would be informative to uncover the unclear mechanism of how dicyemids interact with the surroundings.

Cilia are also considered as signal transduction components and participate in mechanoreception (Bloodgood, 2010). As the "avoiding reaction" shown in protist *Paramecium*, when hitting a solid object, a *Paramecium* shows a response of reversing the swimming direction (Jennings, 1906). By depolarizing mechanoreceptor potential on the plasma membrane, Ca^{2+} enter the cilia through voltage-sensitive ion channel, and the rise of Ca^{2+} concentration trigger reverse beating pattern (Ogura and Machemer, 1980). Dicyemids bear different cell types of cilia, and they contribute to various functions. The cilia on the adult individual could generate currents to circulate the urine in the renal sac and continuously bring nutrients forward to cell surface for endocytosis. On the other hand, the cilia on the posterior side of infusoriform larva contribute to the high efficient mobility. However, the possible function of a small patch of shorter cilia dorsal to the nucleus of apical cells of the infusoriform larva is still ambiguous (Short et al., 1966). Due to the short length, these cilia are unlikely to respond to the mobility, and probably participate in sensory function by mechanoreception.

Serotonin is well known as a conserved neurotransmitter. However, the function of serotonin is not restricted to the transmission of chemical signals by nerve cells (Czaker, 2006). Serotonin is also involved in early developmental process from cleavage to gastrulation and regulation of ciliary activity as well (Paparo,1986; Shmukler and Tosti, 2002). In dicyemids, the serotonin-like molecule is present either in the small vesicles of ciliated epidermal cells of adult individuals, or strikingly in the capsule cells and central internal cells of infusoriform larva (Czaker, 2006). Since dicyemids completely lack a nervous system, serotonin may support multiple functions such as regulation of developmental process, ciliary activity, and phagocytosis, as reported in mouse (Freire-Garabal et al., 2003). Considering the presence of serotonin in the capsule cells, I suspect that serotonin may also function as a hormone or intracellular regulator associated with the release of urn cells (germinal cells) from the infusoriform larva. Overall, the detail of sensory mechanisms retained in dicyemids leaves large

portions to be investigated, such as the potential functions of neurotransmitters and peptide hormones, and the presences of GPCRs.

Summarizing the conceptual background and motivation mentioned above, this PhD thesis aims to investigate (1) whether dicyemids are a group of morphologically reduced spiralians, and whether the Spiralia has a noncoelomate common ancestor; (2) whether dicyemid genome is reduced in association of their simplified morphology; (3) what are the genomic adaptations to their parasitic lifestyle, especially for the sensory function. I believe that, based on whole genome scale approach, I would be able to address these questions.

2 The phylogenetic position of dicyemid mesozoan

2.1 Introduction

In the 19th century, owing to their simple body plans, the name Mesozoa was proposed for dicyemids, as intermediates between the Protozoa (unicellular animal-like eukaryotes) and the Metazoa (multicellular animals) (van Beneden et al., 1876). However, developmental studies revealed that their embryos employ spiral cleavage, a characteristic feature of spiralians (Furuya et al., 1992). In addition, a "spiralian peptide," which is only found in spiralian lineages, is encoded by the dicyemid *DoxC* gene. Hence, thereafter, dicyemids have been regarded as degenerate triploblasts, and members of the Spiralia (Kobayashi et al., 1999). Several studies have examined the phylogenetic position of dicyemids, based on limited amounts of molecular data. Moreover, studies of tool-kit transcription factor genes, Pax6 and Zic, also suggested that dicyemids are highly simplified bilaterians (Aruga et al., 2007). Their morphology might have become simplified secondarily by virtue of their parasitic lifestyles (Awata et al., 2007). Inferred from 18S rRNA sequences, dicyemids were considered related to nematodes (Figure 2.1a) (Pawlowski et al., 1996), while analyses of 18S and 28S rRNA sequences suggested a close affinity of both dicyemids and orthonectids to annelids (Figure 2.1b) (Petrov et al., 2010). Another study using amino acid sequences of innexin suggested that dicyemids are a sister group to the clade consisting of annelids and mollusks (Figure 2.1c) (Suzuki et al., 2010). Take the advantage of next generation sequencing methodology, genome-wide molecular characters allow deep phylogenetic comparisons between taxa that lack clear relationships on morphology or ultrastructure. Recently, large quantities of transcriptomic data from microscopic metazoan lineages have been incorporated into phylogenomic analyses to improve our understanding of spiralian evolution (Struck et al., 2014; Laumer et al., 2015). However, dicyemids have been excluded from most recent phylogenomic studies of microscopic spiralian lineages. Here I used genome-wide data to reexamine the phylogenetic position of dicyemids using maximum

likelihood (ML) and Bayesian inference (BI) analyses. Because systematic biases may sometimes occur in phylogenomic studies, confounding analyses by creating artificial signals, I performed phylogenomic analyses using not only the complete dataset, but also sub-datasets from which rapidly evolving or compositionally heterogeneous sites had been eliminated. Namely, orthologs that could have caused long-branch attraction artifacts had been excluded.

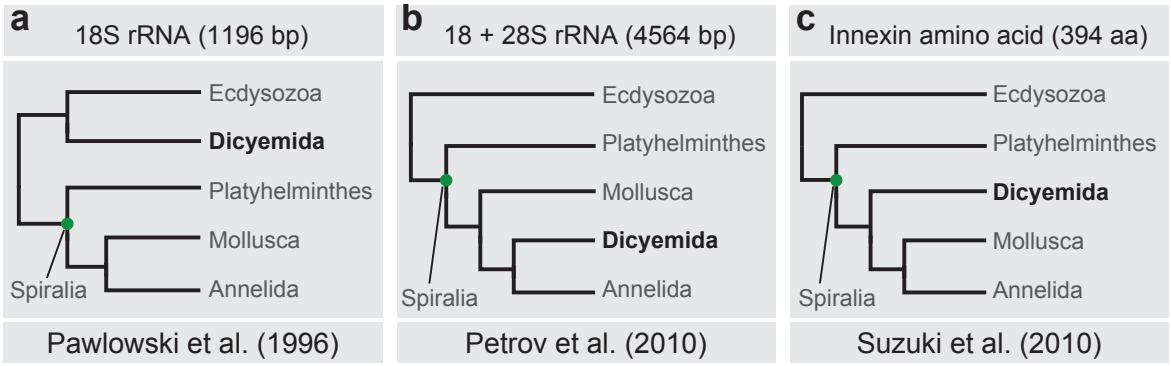

Figure 2.1 | The phylogenetic position of dicyemids remains controversial.
(a) Phylogenetic trees inferred from 18S rRNA sequences, (b) 18 + 28S rRNA sequences, and (c) innexin amino acid sequences.

2.2 Materials and methods

2.2.1 Sample collection, library preparation and Illumina sequencing

A mixed life-stage sample of *D. japonicum* was collected from urine in renal sacs of the host *Octopus sinensis*. Larger pieces of octopus tissue were discarded manually under the stereo microscope. After washing with artificial seawater for several times to remove contaminated octopus cells, the sample was fixed and homogenized in TRIzol Reagent (Ambion, #15596026). RNA from the mixed-stage dicyemid sample was extracted using the phenol-chloroform method, after which it was further purified with a QIAGEN RNeasy Micro Kit (QIAGEN, #74004). A stranded library of the mixed stage sample was prepared using a NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB, #E7420), and sequencing was performed

on an Illumina HiSeq2500. Newly sequenced *D. japonicum* transcriptomic data were deposited in the DNA Data Bank of Japan (DDBJ accession number: DRA004566).

In order to obtain the reference sequences of contamination, the gonad tissue from a host individual of *O. sinensis* was ground into fine powder in liquid nitrogen and lysed in the lysis buffer and proteinase K. After extracting the DNA-containing aqueous layer from phenol-chloroform method, RNase A was treated and then purified again by phenol-chloroform method. The octopus genomic DNA was precipitated in sodium acetate and ethanol solution and washed with 70% ethanol. One pair-end library with an insert size of 500 bp was prepared using an Illumina TruSeq DNA PCR-Free Library Prep Kits (Illumina, #20015962) and sequenced using a HiSeq Rapid run.

2.2.2 Available published data for phylogenomic analyses

Raw reads from transcriptome sequencing of *Limnognathia maerski* (NCBI accession number, SRR2131287), *Macracanthorhynchus hirudinaceus* (ERR454503), *Mesodasys laticaudatus* (SRR1797883), *Stenostomum sthenum* (SRR1801788) and *Brachionus koreanus* (SRR1658835) were downloaded from the National Center for Biotechnology Information (NCBI) database. Transcriptome assemblies of *Adineta vaga*, *Brachionus plicatilis*, *Gnathostomula paradoxa*, and *Macrodasys* sp. were adopted from the supplemental database of Struck et al. (2014). Transcriptome assemblies of *Stenostomum leucops*, *Microstomum lineare*, *Prostheceraeus vittatus*, *Geocentrophora applanata*, *Monocelis fusca*, and *Bothrioplana semperi* were adopted from Laumer et al. (2015). Protein sequences of *Tribolium castaneum*, *Drosophila melanogaster*, *Schistosoma mansoni*, and *Daphnia pulex* were downloaded from the Uniprot database. Protein sequences contributed by the *Intoshia linei* genome project (Mikhailov et al., 2016) were downloaded from NCBI BioProject: PRJNA316116. Gene models of *Octopus bimaculatus* were downloaded from the website of

the Molecular Genomics Unit at OIST (Albertin et al., 2015). Gene models of *Lottia gigantea*, *Capitella teleta*, and *Helobdella robusta* (Simakov et al., 2013) were downloaded from the JGI database (Project ID: 1091351, 16637, and 16078).

2.2.3 Transcriptome assembly and contamination control

Illumina raw reads from mixed-stage library were quality-trimmed with Trimmomatic (v0.33) (Bolger et al., 2014). Trimmomatic removed 3 bases from both ends of all reads and deleted them once the average quality within the window fell below a threshold of 20. If reads became shorter than 36 bases, they were discarded (LEADING:20, TRAILING:20, SLIDINGWINDOW:4:20, MINLEN:36). Afterward, quality-trimmed reads were assembled *de novo* using Trinity (v2.0.6) (Grabherr et al., 2011) with default settings. TransDecoder was used to extract likely coding regions within Trinity transcriptome assemblies and translate transcripts into amino acid sequences (Haas et al., 2013).

Although I washed dicyemid samples with artificial seawater several times and carefully collected individual dicyemids under a microscope, I still could not preclude the possibility of contamination with host octopus cells. In order to avoid the octopus contamination, I performed an assessment to confirm that our dicyemid transcriptome assembly was uncontaminated. I mapped 562 million raw reads from an Illumina pair-end library of the host *O. vulgaris* back to the dicyemid transcriptome assembly using Bowtie 2 (v2.2.3) (Langmead & Salzberg, 2012). Only 1% of dicyemid transcripts were mapped by octopus reads, and none of mapped transcripts were included in datasets for the present study.

2.2.4 Compilation of datasets and phylogenomic reconstructions

Translated transcripts of all 29 taxa were assigned into ortholog groups (OGs) using a hidden Markov model-based search with HaMStR (Ebersberger et al., 2009). In order to minimize

missing data, only OGs that contained orthologs from all 29 taxa were selected. Each OG was iterative refinement (-localpair -maxiterate 1000) (Katoh et al., 2013). Then alignments were trimmed using trimAl (v1.2) (Capella-Gutierrez et al., 2009) with a gap threshold of 0.9, a similarity threshold of 0.001, and a window size of 6. Trimmed alignments shorter than 30 amino acids were discarded. Ortholog alignments were concatenated into a supermatrix using FASconCAT-G (Kuck et al., 2014). For further filtering, the gene tree of each selected ortholog was reconstructed with PhyML (v3.1) (Guindon et al., 2010). A paralog screening function of TreSpEx (Struck et al., 2014) detected possible paralogs. I use TIGER v1.2 (Cummins et al., 2011) to rank sites into 20 bins based upon relative evolutionary rate of the applied dataset. Sums of branch lengths of each gene tree were determined using custom Perl scripts. An index of long-branch heterogeneity was calculated by TreSpEx (fun -e) for each gene tree. BMGE (v1.12) identified and removed the high-entropy regions and compositionally heterogeneous sites (Criscuolo et al., 2010). All aforementioned software was employed with default settings, unless otherwise specified. RAxML (v8.1.20) (Stamatakis et al., 2014) was employed to reconstruct phylogenetic trees using the maximum likelihood method under the GAMMA model of rate heterogeneity. Regarding to the bootstrap replicate, I tested the ML analysis with 100, 200, and 300 bootstrap replicates in the preliminary experiments. However, the tree topologies were consistent, and the bootstrap support values were nearly identical in the results of ML analyses using these three bootstrap replicate settings. Moreover, the ML analysis with 100 bootstrap replicates has been accepted in recent phylogenomic studies (Struck et al., 2014; Laumer et al., 2015). Therefore, in order to reduce the consumption of computational resources, I applied 100 bootstrap replications to all maximum likelihood analyses in this study. The partitioning scheme for each dataset was selected by PartitionFinderProtein v1.1.1 (Lanfear et al., 2012) using RAxML and relaxed clustering algorithm. For Bayesian inference analyses, a Bayesian Markov chain Monte Carlo (MCMC) sampler, PhyloBayes-MPI (v1.6j) (Lartillot et

al., 2013) was used. For each dataset, three independent chains were calculated using a CAT + GTR mixture model. I discarded at least the first 5,000 trees from each chain, and then sub-sampled every 10 trees to calculate a majority rule consensus tree of all remaining trees pooled across three chains. The PhyloBayes "bpcomp" command was used to calculate the largest discrepancy (maxdiff) observed across three independent chains. For all analyses, maxdiff values were lower than 0.15, indicating that all chains had converged. Calculated cycles and consumed time until three independent chains converged, depended on the datasets used. For instance, the analyses of Dataset 1 took 7 weeks to achieve convergence using 128 cores.

2.3 Results

Taking advantage of next-generation sequencing technology, phylogenomic methods utilize large molecular datasets to enable investigation of the phylogeny of animal taxa that have small body sizes and lack uniting synapomorphies. The *D. japonicum* was assembled using default settings for Trinity. The transcriptome assembly size is 14.2 Mbp, containing 19,498 contigs (40,707 transcripts), with a length-weighted median (N50) scaffold size of 1,586 bp. Newly sequenced dicyemid transcriptomic data were complemented with published transcriptomic data or predicted gene models from other spiralian, ecdysozoan, and deuterostome species. I developed a complete dataset for 29 taxa containing 348 orthologs, 58,124 amino acids, with only 6% missing data (Dataset 1). However, because large amounts of data may also amplify systematic biases, such as erroneously assigned orthologs, compositional heterogeneity, and long-branch attractions, I prepared two sub-datasets to assess the influences of potential bias sources.

Since orthologs were assigned by HaMStR, which sometimes groups paralogous sequences erroneously as sets of orthologous sequences, a paralog screening function of TreSpEx was used to detect paralogs. After removing 17 possible paralogs from the complete

dataset, two filtering criteria were used to generate Datasets 2 and 3, in which potential sources of systematic bias were removed. First, I obtained the sums of branch lengths of each gene tree using custom Perl scripts. If the branch length of dicyemids on a gene tree exceeded 30% of the sum of all branch lengths, this ortholog was excluded. Afterward, I ranked sites into 20 bins based upon relative evolutionary rate of the remaining alignment and removed bin20 (the most rapidly evolving sites). That resulted in Dataset 2, containing 321 orthologs and 45,359 amino acids. For Dataset 3, an index of long-branch heterogeneity was calculated with TreSpEx for each gene tree. After excluding orthologs with index values over 100, BMGE removed the high-entropy regions and compositionally heterogeneous sites, and generated Dataset 3, with 302 orthologs and 41,202 amino acids.

Maximum likelihood (ML) analyses were conducted using all datasets with partitioned analyses and 100 bootstrap replicates. ML trees from all three datasets showed identical topology (Figure 2.2), suggesting that our analyses were not affected by systematic bias, yet some basal splits within the Spiralia were supported with relatively mediocre bootstrap values. Bayesian inference with site-heterogeneous mixture models (CAT + GTR) is reportedly relatively resistant to long-branch attraction (LBA) artifacts (Lartillot et al., 2007), but it is computationally intensive. Therefore, I subjected only Datasets 1 and 2 to Bayesian inference (BI) analyses. BI tree topology for Dataset 1 (Figure 2.3a), differed from that of Dataset 2 only in the monophyly of the Gnathifera (Figure 2.3b). On BI analyses of Dataset 1, the posterior probability value was not significant for the node connecting the Mesozoa to the Rouphozoa (Figure 2.3a). However, the result of Dataset 2 offered solid support for this node (Figure 2.3b).

Both *D. japonicum* (Dicyemida) and *I. linei* (Orthonectida) exhibit long branches not seen in other animal taxa (Figures 2.2 and 2.3). In all ML and BI analyses, *D. japonicum* shows a close affinity to *I. linei* with strong statistical support, although I could not categorically exclude the influence of their long-branch lengths. However, parasitic organisms often have

short generation times and large population sizes, which may be associated with rapid evolution (Page et al., 1998; Zarowiecki et al., 2014). The long branch length probably reflects the accelerated evolutionary rate in dicyemids (Aruga et al., 2007). It may be that dicyemids and orthonectids both evolved rapidly after diverging from their common ancestor. In addition, both dicyemids and orthonectids possess simplified morphologies without obvious synapomorphies. Therefore, further comparative studies on the micro-structures or genomic features of these taxa may provide more evidence to assess affinities between them.

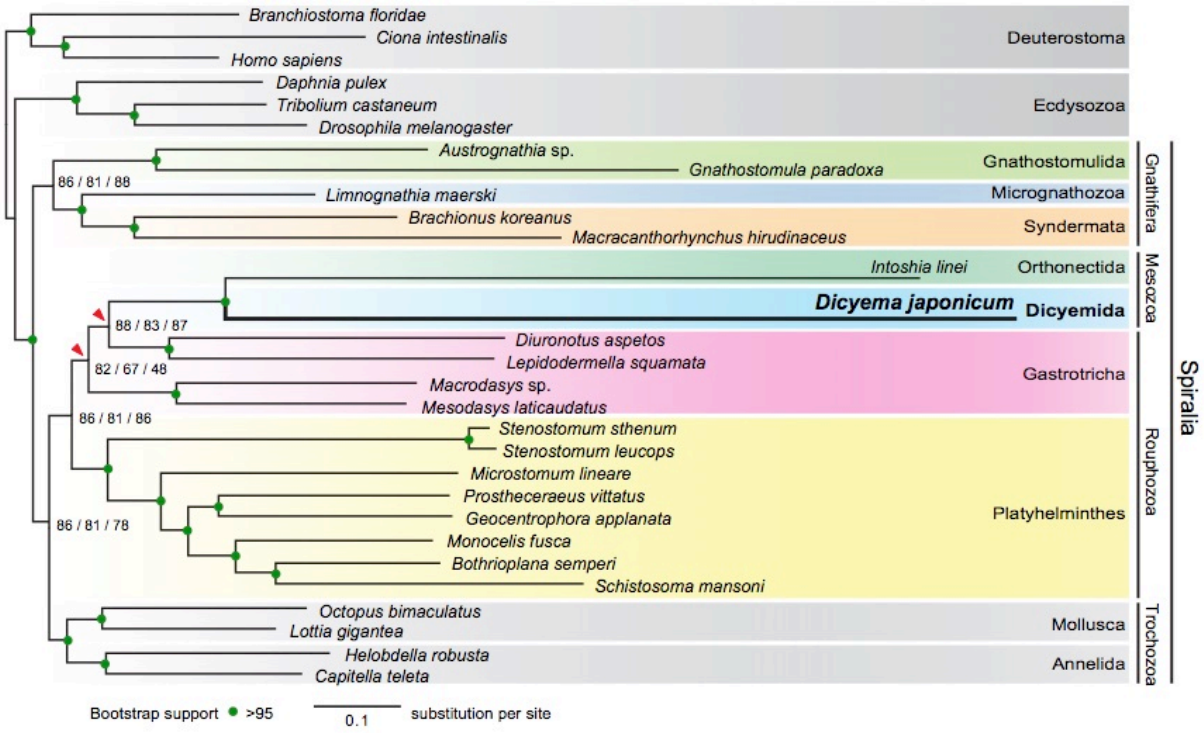

Figure 2.2 | The maximum likelihood tree inferred from 348 orthologs.

This tree topology is consistent with ML trees from analyses of two sub-datasets filtered to remove systematic biases. Analyses were executed under the GAMMA model of rate heterogeneity with 100 bootstrap replicates using RAxML. The Dicyemida displays close affinity to the Orthonectida, and both are nested within the Gastrotricha. Bootstrap values for three datasets (left to right): Datasets 1-3, respectively. Red triangles indicate different groupings from Bayesian analyses (Figure 2.3).

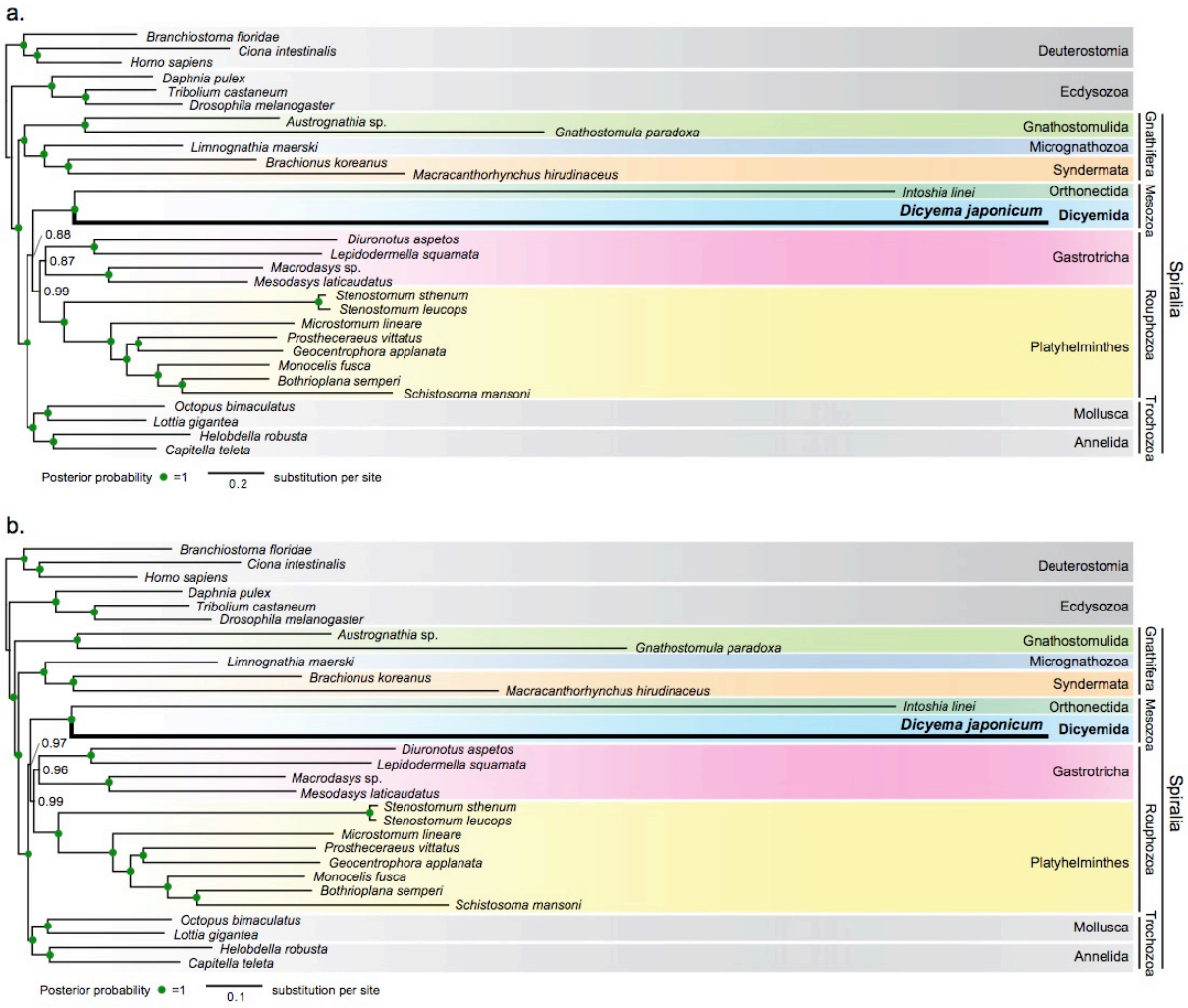

Figure 2.3 | Bayesian inference analyses using Datasets 1 and 2.

Bayesian inference analysis was performed on two datasets, and each dataset ran three independent trains under the CAT + GTR model using PhyloBayes-MPI. Convergence of three chains occurred with a maxdiff value of <0.15. The trees inferred from Dataset 1 (a) and Dataset 2 (b) show that the Dicyemida have a closer affinity to the Orthonectida, and that the Mesozoa diverged early as a sister group to the Rouphozoa. Filled green circles indicate a posterior probability of 1.

ML analyses placed the clade Mesozoa [Dicyemida + Orthonectida] as a sister group to one gastrotrich clade [*Diuronotus aspetos* + *Lepidodermella squamata*] (Figure 2.2), and the result of a taxon-exclusion experiment based on Dataset 3 showed *D. japonicum* as nested within the Gastrotricha as well (Figure 2.4). These indicate that the close relationship of *D. japonicum* with the Gastrotricha in ML analyses probably did not reflect the presence of the other long-branch taxon. Although the position of the Mesozoa does not alter across various ML analyses and is independent of potential systematic biases, the ML result for Dataset 3 (with long-branches and compositional heterogeneity removed) displayed poor bootstrap support for the root of the clade comprising dicyemids and gastrotrichs. In contrast, in BI analyses, the Dicyemida and Orthonectida formed a sister group to the Rouphozoa (Figure 2.3), and BI analysis of Dataset 2 (Figure 2.3b) demonstrated that the Mesozoa is a sister group to the Rouphozoa with significant statistical support.

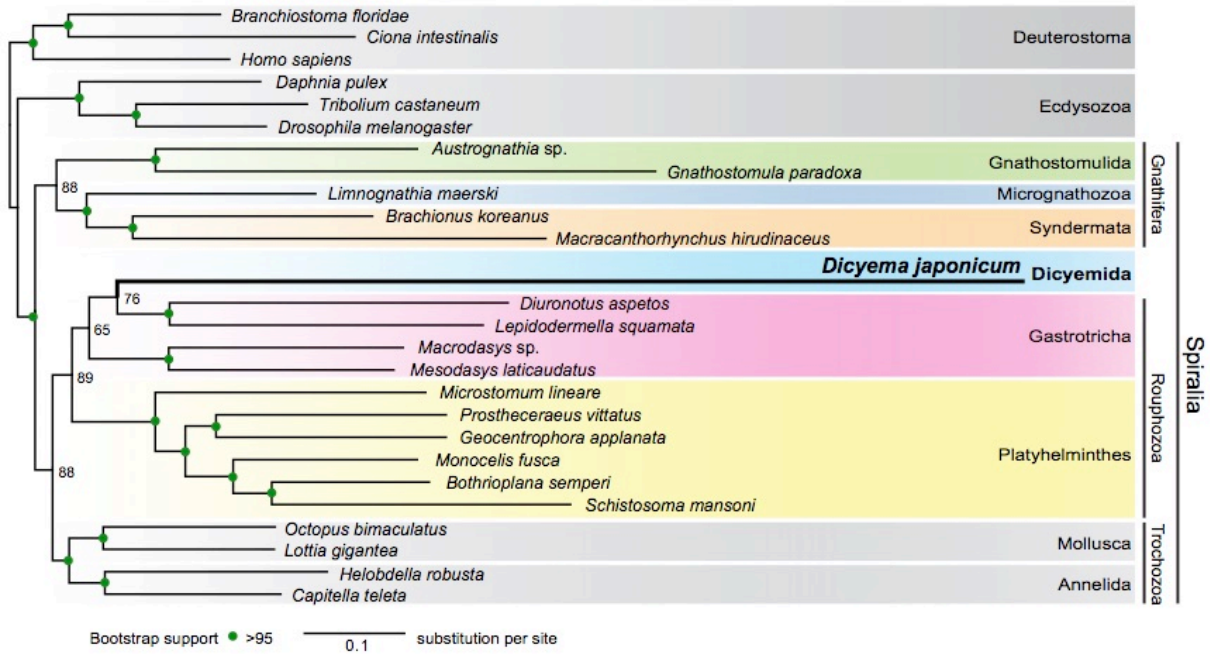

Figure 2.4 | The maximum likelihood tree inferred from the taxon-exclusion dataset. ML tree inferred from Dataset 3 covering 26 taxa indicates that the nesting of *D. japonicum* within the Gastrotricha probably does not reflect long-branch attraction artifacts. Filled green circles indicate >95% bootstrap support for all datasets.

2.4 Discussion

A previous developmental study showed that early cleavage of dicyemids exhibits stereotypical spiral cleavage, as in the case of spiralians such as annelids, mollusks, or platyhelminthes (Furuya et al., 1992). Furthermore, studies using different molecular markers have placed mesozoans as close relatives of annelids and mollusks (Pawlowski et al., 1996; Petrov et al., 2010). To identify orthologs for phylogenomic analysis, the so-called "genome-wide data" for phylogenetic analysis utilize either the transcripts which are assembled using transcriptome sequencing data from whole life-cycle stages or the predicted gene models which are annotated on the genome assembly with the support from transcriptome sequencing data. However, there are limited genome sequencing data in microscopic spiralians currently. In order to avoid the improper interpretation of phylogenetic relationships caused by low taxa coverage, I included representatives of microscopic spiralians whose transcriptomic data are currently available, even if they lack genome sequencing data to generate gene models. Hopefully, in the near future, the rapidly growing amount of genomic data would allow us to use more genomic regions to infer the phylogenetic relationships. One possibility for future study is that the conserved synteny regions on the genomes could be identified based on the conserved order of gene models to provide the information for phylogenetic analysis, once all analyzed taxa obtain good-quality genome assembly. Within the conserved synteny regions, taxa-shared non-coding sequences could be possible markers to provide phylogenetic information reflecting the relationships between analyzed taxa.

The present phylogenomic analyses, however, offer two additional possibilities. ML analyses suggest that the Mesozoa is nested within the Gastrotricha, implying that dicyemids may be degenerate gastrotrichs. However, taking morphological traits into account, the mono-ciliated epithelial cells of gastrotrichs has been considered a diagnostic trait (Cavalier-Smith, 1998), whereas dicyemids possess multi-ciliated epithelial cells. This morphological trait barely

supports the hypothesis of dicyemids nested within the Gastrotricha. Alternatively, BI trees insinuate that the Mesozoa diverged early from other rouphozoans. Even so, in all analyses, the Dicyemida, Orthonectida, Gastrotricha, and Platyhelminthes constitute a monophyletic group that is a sister group to the Trochozoa (Mollusca + Annelida). This indicates that dicyemids may share characters of a common acoelomate ancestor with gastrotrichs and platyhelminthes, rather than with mollusks and annelids. Nevertheless, morphological synapomorphies between dicyemids and gastrotrichs (or rouphozoans) remain to be discovered. Developing a firm grasp of spiralian evolution will still require additional developmental or genomic studies of a wide range of microscopic spiralian taxa.

Previous hypothetical scenarios of spiralian evolution remain controversial. One supports the traditional acoeloid-planuloid hypothesis of a noncoelomate common ancestor of the spiralians, whereas another suggests that the common ancestor resembled an annelid-like organism with a segmented, coelomate body plan (Struck et al., 2014). According to the present analyses of spiralian phylogeny, the small, non-coelomate Gnathifera branched off first, and forms a sister group to the Rouphozoa and Trochozoa, as reported in previous studies (Struck et al., 2014; Laumer et al., 2015). Moreover, within the Gnathifera and Rouphozoa, most species are small, with acoelomate or pseudocoelomate body plans, whereas animals with coeloms are only found in the Lophotrochozoa (Mollusca, Annelida, etc.). The foregoing analyses and observations support the conclusion that the last common ancestor of spiralians may have been a microscopic animal lacking a coelom. It may be that microscopic lineages either maintained their ancestral morphology or that they underwent regressive evolution secondarily simplifying their morphologies, while lophotrochozoan lineages evolved more complex morphologies with a coelomic cavity and larger body size.

The Chapter 2 has been published in a paper entitled “The phylogenetic position of dicyemid mesozoans offers insights into spiralian evolution” in the journal *Zoological Letters* (2017) 3: 6.

3 The genome of *Dicyema japonicum*

3.1 Introduction

Along with evolving from free-living organisms to parasites, dicyemids appear to develop characteristic genomic features. A survey of 39 dicyemid genes showed that most introns (97.6%, 205/210) are extremely short, with an average length of 26 bp (Ogino et al., 2010). This average intron size is nearly equal to that of the chlorarachniophyte, *Bigelowiella natans*, (18–21 bp), which has the shortest introns among any known eukaryotes. The shortening of introns may reflect the selective pressure during the course of adaptive simplification to reduce unnecessary expense such as long introns. This also implies that dicyemids may possess smaller genome than other non-parasitic spiralians. On the other hand, the intron density (5.3 introns/gene) is similar to that in other invertebrates examined so far. Ogino et al. (2010) mentioned that dicyemid genes show rather high evolutionary rates, despite the conserved exon/intron structure that may have evolved under functional constraints.

In dicyemids, chromatin elimination or selective replication of chromosomal elements into extrachromosomal DNA circles occurs during early embryogenesis (McConnaughey, 1951). These circular DNAs have been suggested to originate during whole genome amplification, in which unnecessary sequences are excised from chromosomal DNA and eliminated from the somatic genome as mini-circles during development (Noto et al., 2003). However, these extrachromosomal circular DNAs are highly heterogeneous and lack potential open reading frames, suggesting that they are unlikely to have protein-coding function (Noto et al., 2003). The similar phenomenon of chromosome elimination during development has also been reported in hagfish and lampreys (Kojima et al., 2010; Smith et al., 2010). Moreover, intracellular parasitic protists such as apicomplexans and microsporidia have dwarfed genomes, due to gene loss and elimination of noncoding DNA (Spano & Crisanti, 2000;

Cavalier-Smith, 2005). The downsizing of genomes seems a general strategy to reduce the cost of maintaining the original large genome, especially under limited nutrients (Gray et al., 1999).

Overall, the shortening of introns and chromatin elimination may probably be beneficial to reduce energy consumption during the DNA replication process, and also reflect an adaptation to parasitic lifestyle. However, except the aforementioned adaptations, the knowledge of dicyemid genome is limited. It is desired to know what kind of genomic feature changes occurred during the parasitism process and whether or not dicyemid genome retains some genomic convergences shared with genomes of other parasites. Especially, it should be investigated that the dicyemid lineage-specific genomic adaptations corresponding to the particular life cycle and any gene gain and loss along with the morphological simplification.

Here, I decoded, for the first time, the *Dicyema* genome and predicted the gene models. This provides the resources for present and future comparative studies. In addition to the shortening of introns, dicyemids might retain less genes than other parasitic spiralians, expecting a compact genome. Moreover, I also demonstrate examples of gene gain and loss, presumably corresponding to the particular lifestyle of dicyemids. The results show that gene number reduction of whole gene set, especially in the metabolism pathways, would be one of the molecular convergences among parasitic bilaterians. The partial loss of Hox cluster genes in parasitic organisms may echo the simplification of body organization. In addition, only six putative GPCR genes were found in dicyemids along with the reduction of nervous system, although they may play the roles in the retained sensory functions. In contrast to the gene loss, dicyemids possibly encountered gene expansions on nutrient-transportation-associated genes, which may explain how dicyemids acquire low molecular-weight nutrients from the urine of the host, although dicyemids do not have month and digestion system.

3.2 Materials and methods

3.2.1 Sample collection, library preparation and sequencing

For genome sequencings, mixed-stage samples were collected from one octopus individual. The gonad sample from a male octopus was also collected for sequencing as the contamination reference. The samples were frozen at -20 °C. The dicyemid genomic DNA was extracted from the samples using Promega Maxwell 16 Systems and Maxwell 16 Cell DNA purification Kit (Promega, #AS1020). One paired-end library with an insert size of 600 bp was prepared using TruSeq DNA PCR-Free Library Prep Kits (Illumina, #20015962). Four mate-pair libraries of insert lengths (1.6-7, 7-10, 10-12.5, and 12.5-20 Kbp) were prepared using the Nextera Mate Pair Sample Preparation Kit (Illumina, #FC-132-1001). Sequencing using the pair-end library was performed on an Illumina MiSeq, while mate-pair libraries were sequenced on an Illumina HiSeq 2500. In addition, PacBio extra-long reads were generated by single-molecule real-time (SMRT) sequencing method on PacBio RS II (Table 3.1).

Table 3.1 | Summary of genome sequencing libraries and data

Sequencing platform	Method	Library Length	Read Length	Raw read pairs	Raw bases
Dicyema japonicum genome					
Illumina MiSeq	Paired-end	600	2 x 300	206,959,118	113,349,153,168
Illumina HiSeq	Mate pair (Nextera)	1,600-7,000	2 x 150	4,433,810	982,878,355
Illumina HiSeq	Mate pair (Nextera)	7,000-10,000	2 x 150	8,683,880	1,919,579,646
Illumina HiSeq	Mate pair (Nextera)	10,000-12,500	2 x 150	12,629,929	2,790,166,864
Illumina HiSeq	Mate pair (Nextera)	12,500-20,000	2 x 150	9,097,714	2,003,644,168
PacBio RS II	SMRT	>7,000	7,068[a]	1,057,663[b]	6,797,850,486
Octopus sinensis genome					
Illumina HiSeq	Paired-end	600	2 x 250	312,232,334	181,116,326,462

[a] PacBio subread mean length; [b] subread number by PacBio

3.2.2 Genome size and heterozygosity estimation

I estimated the genome size using the *k*-mer coverage-based methods. The concept for the estimation method is intuitive that the average *k*-mer coverage is equal to total number of *k*-mers divided by genome size. I applied raw reads of two Illumina MiSeq runs on the pair-end library for this analysis. Since it was difficult to exclude octopus cell contamination completely in sample collection procedures, I first mapped raw reads to a draft *Octopus vulgaris* genome assembly and only the unmapped reads were used for estimations. Secondly, the unmapped reads were quality-trimmed by Trimmomatic (v0.33) to delete them once, the average quality within the window fell below a threshold of twenty. If reads became shorter than fifty bases, they were discarded (LEADING:20, TRAILING:20, SLIDINGWINDOW:4:20, MINLEN:50). Then, I performed the error-correction to trimmed reads using the Corrector_AR program in SOAPec (v2.01), which corrected the sequencing errors based on k-mer frequency spectrum (Luo et al., 2012).

Afterwards, Jellyfish (v.2.1.3) (Marçais & Kingsford, 2011) counted the *k*-mer occurrences of pair-end reads with the setting of $k = 17$. Plotting the histogram graph of *k*-mer distribution, the peak coverage was supposed to be the average *k*-mer coverage. The total number of *k*-mers was the sum of each coverage multiplied by its occurrence frequency. Then, I could calculate the estimated genome size. Alternatively, 17-mer profile was used to estimate the overall characteristics including genome size and heterozygosity by an open-source tool GenomeScope (Vurture et al., 2017), which applies a mixture model of four evenly spaced negative binomial distributions to the k-mer profile.

3.2.3 Genome assembly

I developed a pipeline to remove possibly contaminated sequences originated from the host octopus and performed two preliminary *de novo* assemblies using Illumina and PacBio data,

respectively. To achieve the most completeness, I merged two preliminary assemblies into the final assembly for further analyses. Sequencing raw reads from Illumina pair-end and mate-pair libraries were quality-trimmed using Trimmomatic (v0.33) with quality threshold of twenty and minimum length of fifty bases (SLIDINGWINDOW:4:20, LEADING:20, TRAILING:20, MINLEN:50). Prior to quality-trimming for mate-pair reads, non-adaptor read filtering and reverse-complement processes were performed. Biotin junction adaptors were added to circularize long DNA fragments before re-fragmented into small fragments by Nextera mate pair sample preparation kit. NextClip (Leggett et al., 2014) categorized raw reads from mate-pair libraries by checking the presence of adaptors on both reads. Reads without the junction adaptor likely from the pair-end sequences that have slipped through the biotin enrichment process were discarded.

Quality-trimmed reads from eight Illumina MiSeq runs on pair-end library were pooled together. I randomly selected one-third of reads (approximately 300X of the estimated genome size) and applied them to the assembler Platanus (Kajitani et al., 2014), which declared the ability to reconstruct genomic sequences of highly heterozygous diploids. The assembling was run with default setting and the initial *k*-mer coverage cutoff of twenty-five to avoid the low-coverage *k*-mers from host octopus. To eliminate possible octopus sequences from the assembly, output sequences of Platanus were mapped by 562 million octopus reads of Illumina paired-end library and were deleted if the average mapped base coverage larger than one. The remained Platanus assembly was proceeded to the scaffolding process conducted by SSPACE (Boetzer et al., 2011) and SSPACE-LongRead (Boetzer & Pirovano, 2014) incorporating sequences of four mate-pair libraries and PacBio long read sequences, respectively. Then, the gap closing process was performed using GapCloser (Luo et al., 2012) with whole pair-end library reads. Before removing redundant allelic scaffolds to obtain a haploid genome assembly

by HaploMerger (Huang et al., 2012), contamination filtering process by octopus reads mapping was performed again.

It was possible to perform *de novo* assembly of the *Dicyema* genome only based on PacBio sequencing data, because the total amount of PacBio data was more than fifty fold of estimated genome size. Since the reads of PacBio sequencing are longer than those of Illumina sequencing, the possible contamination of PacBio subreads could be filtered by back-mapping method of octopus reads prior to *de novo* assembly by Falcon (Chin et al., 2016). Afterwards, scaffolding was performed by SSPACE and SSPACE-LongRead using sequences of four mate-pair libraries and PacBio subreads, respectively. After GapCloser was used to fill gaps in scaffolds, contamination filtering process of octopus read mapping was performed again. Since PacBio reads were regarded with an error rate over 10%, scaffolds were corrected by GATK referring to Illumina pair-end reads after HaploMerger removal of redundant allelic scaffolds.

Dicyemid genome obtained using two pipelines in parallel and then merged together was expected to provide a genome assembly with a better completeness. Therefore, the preliminary haploid genome assemblies yielded from the Illumina and PacBio pipelines were merged by HaploMerger. The merged sequences were gone through scaffolding and gap-closing processes again to obtain the final genome assembly.

To assess the completeness of genome assembly, Assemblathon 2 (Bradnam et al., 2013) calculated overall statistic values of final assembly, while CEGMA (Parra et al., 2007) was used to analyze percentages of core eukaryotic genes present in the final assembly. In addition, GMAP mapped the longest isoform per gene of Trinity transcriptome assembly against the genome assembly with the criteria of minimum coverage and minimum identity cut-off of 0.6.

3.2.4 Gene prediction

Before conducting gene prediction, the gene predictor AUGUSTUS (Keller et al., 2011) was trained by a dicyemid training gene set to obtain dicyemid-specific prediction parameters. Then, gene models were predicted referring to the pre-trained parameters and the hints for introns, exons, and repeats. In order to create a gene-set for training the gene predictor AUGUSTUS, firstly PASA alignment assembly was generated by training gene structures according to the genome assembly, as well as the Trinity assembled cDNA sequences, which were removed the redundant sequences by CD-HIT (Fu et al., 2012) with 95% identity threshold in advance. The protein coding regions of PASA alignment assembly were extracted, and then filtered out redundant (more than 80% identical) and possibly error-causing genes. The training gene-set was randomly divided into two parts; a test-set of three hundred genes was used to evaluate the prediction accuracy of the trained parameters, and the rest of genes were applied for training. A script autoAugTrain.pl in AUGUSTUS package trained with the training gene set, and optimize_augustus.pl performed at most five rounds of optimizations to acquire the dicyemid-specific parameters for gene prediction.

As mentioned in the tutorial of AUGUSTUS, the repeated sequences should be masked prior mapping of transcriptomic data to generate hints, and later gene prediction should be run on the unmasked genome with the hints generated from repeat regions. RepeatScout (Price et al., 2005) discovered the repetitive DNA regions and counted the frequency of these regions into an index. Then, RepeatMasker (Smit et al., 2013) masked the repeat regions on the genome when the repeated sequences occurred more than thirty times. The output data of RepeatMasker could also be retrieved as the hints for repeats. The cDNA sequences of Trinity and PASA assemblies were aligned against the masked genome by BLAT (Kent, 2002) with at least 80% identity. The blat2hints.pl script converted the alignments of Trinity and PASA assemblies to the hints for cDNA sequences. Further, I incorporated quality-trimmed transcriptomic reads in

a two-step iterative mapping approach to generate hints for exons and introns. In the first step, spliced-alignments were performed by Tophat (Kim et al., 2013) to create intron hints, and these hints were applied to predict genes with AUGUSTUS. By concatenating the intron information in intron hints and predicted genes from the first step, I created the database of exon-exon junctions. The second step was to map quality-trimmed transcriptomic reads against the exon-exon junctions by Bowtie (Langmead & Salzberg, 2012) to increase the number of reads aligned to splice sites. Then, new hints generated from the merger of the second-round alignments could increase gene prediction accuracy.

3.2.5 Pfam domain and KEGG pathway analyses

The amino acid sequences of selected bilaterian species (Table 3.2) were applied to Pfam 30.0 domain search using HMMER v3.1b2, and the results with e-value larger than 1e-5 were filtered. A custom script was used to count how many genes contain specific domain each. For pathway analysis, I adopted the online service on KEGG Automatic Annotation Server to assign each predicted gene to KEGG ortholog using bi-directional best hit method and to map assigned orthologs to KEGG reference pathways. A custom Perl script was applied to count how many KEGG orthologs are involved in each pathway on each species studied.

Table 3.2 | The selected bilaterian species for comparative analyses

Short ID	Scientific name	Common name	Gene number	Download source
Spiralians				
dja	*Dicyema japonicum*	Dicyemid	5,012	This study
ili	*Intoshia linei*	Orthonectid	8,724	NCBI
sma	*Schistosoma mansoni*	Blood fluke	11,723	UniProt
emu	*Echinococcus multilocularis*	Tapeworm	10,656	WormBase
sme	*Schmidtea mediterranea*	Planarian	32,615	SmedGD
hro	*Helobdella robusta*	Leech	23,328	Uniprot
cte	*Capitella teleta*	Polychaete	31,236	Uniprot
obi	*Octopus bimaculoides*	Octopus	38,585	OIST
lgi	*Lottia gigantea*	Limpet	23,721	Uniprot
lan	*Lingula anatina*	Brachiopod	29,907	OIST
ava	*Adineta vaga*	Rotifer	49,300	Genoscope
Ecdysozoans				
tsp	*Trichinella spiralis*	Nematode (Clade I)	16,380	WormBase
bma	*Brugia malayi*	Nematode (Clade III)	13,436	WormBase
sst	*Strongyloides stercoralis*	Nematode (Clade IV)	13,098	WormBase
cel	*Caenorhabditis elegans*	Nematode (Clade V)	25,847	Uniprot
dme	*Drosophila melanogaster*	Fruit fly	19,684	Uniprot
Deuterostomes				
sko	*Saccoglossus kowalevskii*	Acorn worm	22,111	OIST
bfl	*Branchiostoma floridae*	Amphioxus	28,544	Uniprot
has	*Homo sapiens*	Human	20,257	Uniprot

3.2.6 Cluster orthologous groups

In order to explore any gene family expansion in the dicyemid genome, amino acid sequences of selected bilaterian species (Table 3.2) were applied to generate ortholog groups using OrthoMCL (Fischer et al., 2011). Firstly, low quality sequences of gene models from sixteen species were removed based on sequence length of OrthoMCL criteria. Afterwards, the sequences were applied to all-versus-all BLAST searches with e-value 1e-5 cutoff. The results proceeded through the internal algorithm of OrthoMCL to separate protein pairs into three relationship categories, namely orthologs, in-paralogs, and co-orthologs. Then MCL program (Enright et al., 2002) clustered the pairs into final ortholog groups and singletons which were not assigned into any ortholog group. The Venn diagram of shared ortholog groups between dicyemid, orthonectid, and other spiralians was plotted by jvenn online service (Bardou et al.,

2014). A custom Perl script was applied to count how many predicted genes involved in each ortholog group from each species studied.

3.2.7 Gene annotation

For preliminary annotation, predicted gene models were identified by reciprocal BLASTP search against Swiss-Prot database downloaded from UniProt (Boutet et al., 2016), in which the reference protein sequences were manually annotated and reviewed. The search outputs were filtered with the e-value threshold of 1e-5.

To search for Hox cluster genes, sixteen candidate gene models containing homeobox domain obtained by Pfam domain analysis were applied to reciprocal BLASTP search against Homeobox Database (Zhong et al., 2008) and Swiss-Prot database. For the candidate gene models which could not have bi-directional best hits, they were submitted to NCBI BLASTP search against nr database manually. Afterwards, the putative Hox genes from BLAST searches were identified by the phylogenetic analysis inferred from the homeobox domain. The homeobox domain sequences of *Lottia*, *Capitella*, *Drosophila*, and *Branchiostoma* were retrieved from the supplemental database of Simakov et al. (2014) and Homeobox Database (Zhong et al., 2008). The dataset was aligned using MAFFT (v7.220) (Katoh et al., 2013) and trimmed using trimAl (v1.2) (Capella-Gutierrez et al., 2009). The gene tree was reconstructed using RAxML (v8.1.20) (Stamatakis et al., 2014) based on the maximum likelihood method under the LG substitution model and the GAMMA model of rate heterogeneity with 100 bootstrap replications.

The protein sequences of G protein-coupled receptors (GPCRs) were downloaded from UniProt and built a database for BLASTP search. The gene models were applied to reciprocal BLAST search against GPCRs database. The bi-directional best hit gene models were confirmed again by preforming NCBI BLASTP search against nr database manually. As to the

gene models which were contain 7 transmembrane receptor (7tm) domain identified by Pfam domain analysis but not have bi-directional best hit against GPCRs database, they were also confirmed again by NCBI BLAST search against nr database.

3.2.8 Mitochondrial gene annotation

A total of 2,287 spiralian mitochondrial protein-coding gene sequences, including five *Dicyema* sequences, were downloaded from NCBI. They were used as queries to search for mitochondrial fragments in the transcriptome assembly, PacBio Read of Insert (referred to as PacBio read), and genome assembly scaffolds by TBLASTN search using invertebrate mitochondrial translation table (NCBI genetic code 5) with e-value threshold of 1e-10. Before TBLASTN search, the PacBio reads were corrected by referring to Illumina pair-end reads by GATK, because PacBio data usually come with high error rate. Since the use of *D. japonicum* mitochondrial sequences to search against *D. japonicum* genome assembly or PacBio reads may provide more accurate information, the transcripts with hit in TBLASTN search from the transcriptome assembly were used as queries to search again in the *Dicyema* genome scaffolds and PacBio reads by BLASTN.

3.3 Results

3.3.1 Assessing sequencing data and genome profile

Eight runs of MiSeq sequencing generated 207 million read pairs of paired-end library, and around 70% of them that passed the quality-trimming process were applied to genome assembling. In addition, the sequencing of four mate-pair libraries generated a total of 76.3 million read pairs. After NextClip-filtering and quality-trimming processes, 22% of the mate-pair reads were used for scaffolding. I also obtained sequences with length up to 65 Kbp from the PacBio platform, and 86% of them retained sequences over the contamination-removing process and were applied for *de novo* assembling and scaffolding.

The paired-end reads from two MiSeq runs were utilized to assess the genome profile. The quality-trimmed reads were applied to count the frequency of 17-mers, which generate 17-mer histogram. However, the GenomeScope analysis could not reach the converged state on model-fitting and failed to measure the relative abundances of heterozygous and homozygous sequences. The 17-mer histogram plot displayed a profile of relatively high frequency at low coverage *k*-mer without a clear coverage peak (Figure 3.1a). The low-coverage *k*-mers might mainly come from the sequencing errors, thus I preformed the error-correction process to the additional quality-trimmed reads prior to *k*-mer counting. After this treatment, I could observe the coverage peaks on the 17-mer histogram (Figure 3.1b). GenomeScope analysis overcame the obstruction to reach the converged state on model-fitting. The raw base error rate estimated by Corrector_AR program is 0.3%. The GenomeScope analysis indicated that the estimated genome size of *Dicyema japonicum* is about 65 Mbp with the heterozygosity rate of 1.24% (Figure 3.1c).

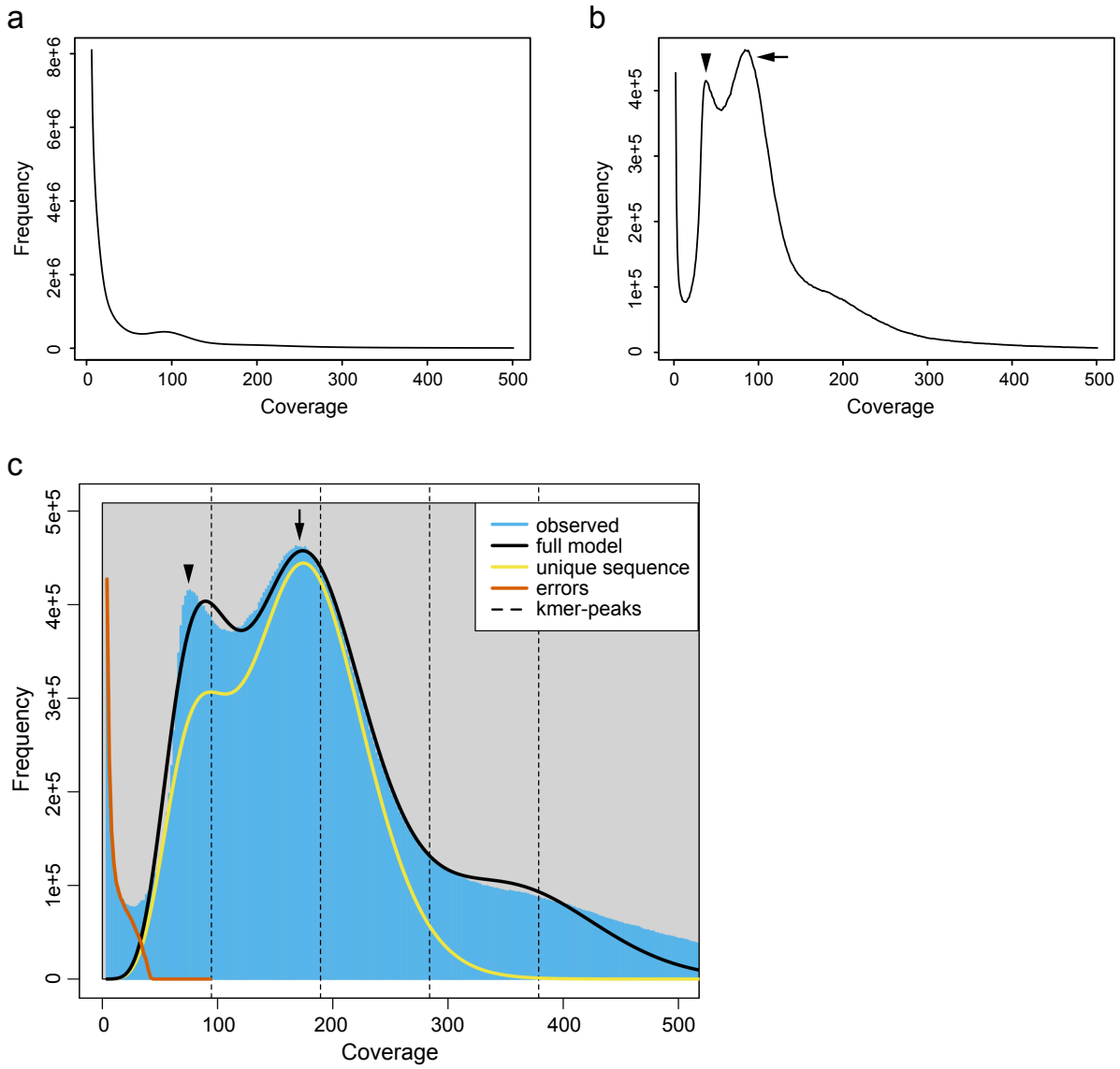

Figure 3.1 | Genome size estimation.

(a) Before error-correction, no peak was observed in the 17-mer frequency spectrum of Illumina paired-end reads. (b) The 17-mer frequency spectrum from the error-corrected Illumina paired-end reads shows the homozygous peak (arrow) and the heterozygous peak (arrowhead). (c) The GenomeScope profile shows the homozygous peak coverage (arrow) is twice as large as the heterozygous peak coverage (arrowhead), and the error rate is 0.012% after error correction. The estimated genome size of *D. japonicum* is about 65 Mbp with the heterozygosity rate of 1.24%.

3.3.2 Genome assembly

In order to obtain a non-contaminated and highly contiguous genome assembly, I developed a custom genome assembly pipeline (Figure 3.2) which includes the following two

characteristics: (1) repeatedly removing possibly contaminated sequences of host octopus, and (2) merging two preliminary *de novo* assemblies based on Illumina and PacBio data each into a consensus genome assembly.

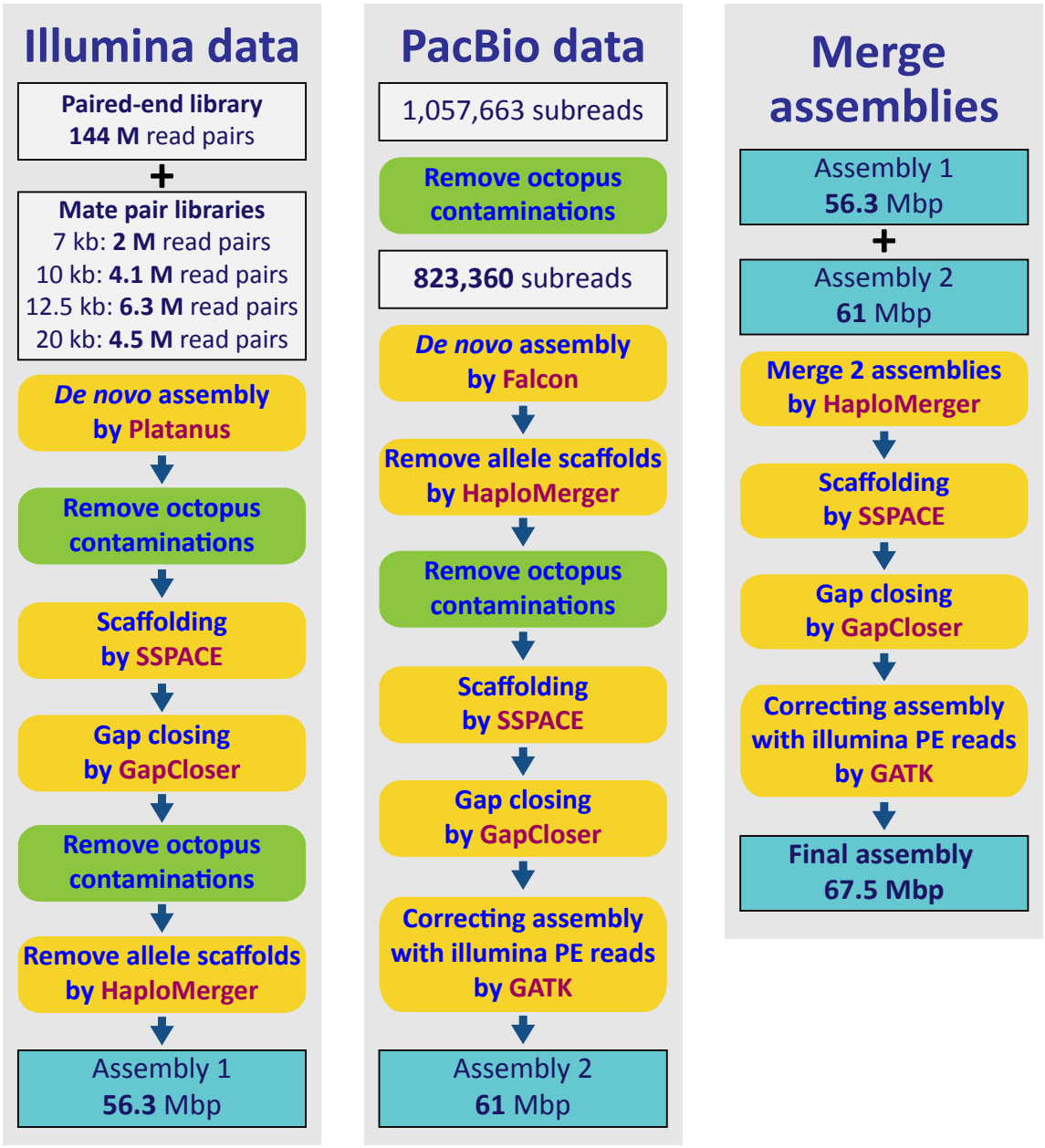

Figure 3.2 | The workflow of *de novo* assembly of *Dicyema japonicum* genome.
The sequencing data generated by Illumina and PacBio sequencing platforms were firstly assembled separately, and then the two data were merged into a final assembly for downstream analyses.

For a preliminary assembly of Illumina data, Platanus initially generated 11,729 scaffolds with a length-weighted median (N50) length of 85 Kbp and a gap rate of 14.7%. Then, this assembly was followed by the first-round contamination-removing step. This step removed 287 scaffolds (1% of the assembled length) which were mapped back by octopus reads with average base coverage larger than one. After improving the continuity and filling the gaps by SSPACE and GapCloser, the second-round contamination-removing step removed four scaffolds (about 0.5% of the assembled length). Afterwards, HaploMerger collapsed redundant allelic scaffolds to achieve a preliminary haploid assembly of 56.3 Mbp, including 462 scaffolds with an N50 length of 678 Kbp and a gap rate of 7.6%.

On the other hand, the first-round of contamination-removing step for the preliminary assembly of PacBio data was performed prior to *de novo* assembly, because the PacBio reads were with similar lengths to the contig assembly obtained using Illumina reads. Originally, Falcon generated an assembly with the size of 353 Mbp, including 18,404 contigs. HaploMerger removed redundant allelic contigs, so that the assembly size reduced to 63 Mbp, containing 864 contigs with an N50 length of 355 Kbp. Then, the second-round contamination-removing step was carried out to remove twelve contigs, decreasing 1.6% of the assembled length. Subsequently, the scaffolding, gap-closing, removing redundant allele, and error-correcting processes improved the continuity of assembly, which resulted in a 61 Mbp assembly, including 327 scaffolds with an N50 length of 834 Kbp and a gap rate of 1.4%.

The two preliminary assemblies, which could be considered as redundant haploid assemblies, were merged into a haploid assembly by HaploMerger. Additional round of scaffolding and gap-closing increased the continuity and reduced the gap rate of the merged assembly. Finally, additional error-correcting processes with Illumina paired-end reads yielded the final genome assembly of *Dicyema japonicum*. As summarized in Table 3.3, the final assembly size was 67.5 Mbp, which is close to the estimated genome size of 65 Mbp. This

genome was assembled by sequencing data of three next-generation sequencers with approximately 1000-fold genome coverage. It contained 377 scaffolds with an N50 length of 1 Mbp, 1,965 contigs with an N50 length 195.9 Kbp, and a gap rate of 4.1%. The GC content of *Dicyema japonicum* genome was 35.2%, which is similar to the other marine invertebrates. Repetitive elements occupy approximately 51.2% of the assembled genome (Table 3.3). To assess the assembly completeness, I carried out CEGMA test, which showed that 77% of 248 ultra-conserved core eukaryotic genes were partially covered and 58.5% were completely covered by the current genome assembly. In addition, 29,082 coding region detected Trinity transcripts were aligned to the genome assembly, and 98.4% of them could be aligned with the criteria of query coverage and identity more than 60%.

Table 3.3 | Summary of genome assembly

Genome size (Mbp)	67.5
GC content (%)	35.2
Gap rate (%)	4.1
Repeats (%)	51.2
Number of contigs	1965
Number of scaffolds	377
Contigs per scaffold	5.2
Contig N50 (Kbp)	195.9
Scaffold N50 (Kbp)	1000.2
Number of predicted genes	5012
Mean coding seq. size (bp)	1155.2
Introns per gene	6.2
Exons per gene	7.6
Mean intron size (bp)	38.2
Mean exon size (bp)	198.2

3-3-3. Gene model prediction

A set of 2,139 genes which were extracted from PASA alignment assembly and filtered by the selecting criteria was applied to train the gene predictor AUGUSTUS. Hints of repeat regions, exons, and introns were also prepared for gene predictions. The coordinate position information of 140,794 repeat regions on the genome was obtained from the output of RepeatMasker as the repeat region hints. By aligning transcripts from Trinity and PASA assemblies, the position information of exons and exon-exon junctions was acquired and organized as 850,785 exon hints and 1,811,535 intron hints. Utilizing the dicyemid-specific parameters and hints, AUGUSTUS predicted 5,012 gene models in the *D. japonicum* genome (Table 3.3), and 71% of transcripts were supported by hints. The average length of predicted protein-coding region was 1205.7 bp, and the average exon and intron number per gene were 7.6 and 6.2, respectively (Table 3.3). Notably, the average intron size was 38.2 bp, which was much smaller than other spiralians (Table 3.3). Comparisons of genomic characteristics such as genome size, intron size, and repeat contents demonstrated large variances among selected spiralians, especially in parasite species. Only the predicted gene number showed consistent reduction in parasite lineages (Figure 3.3).

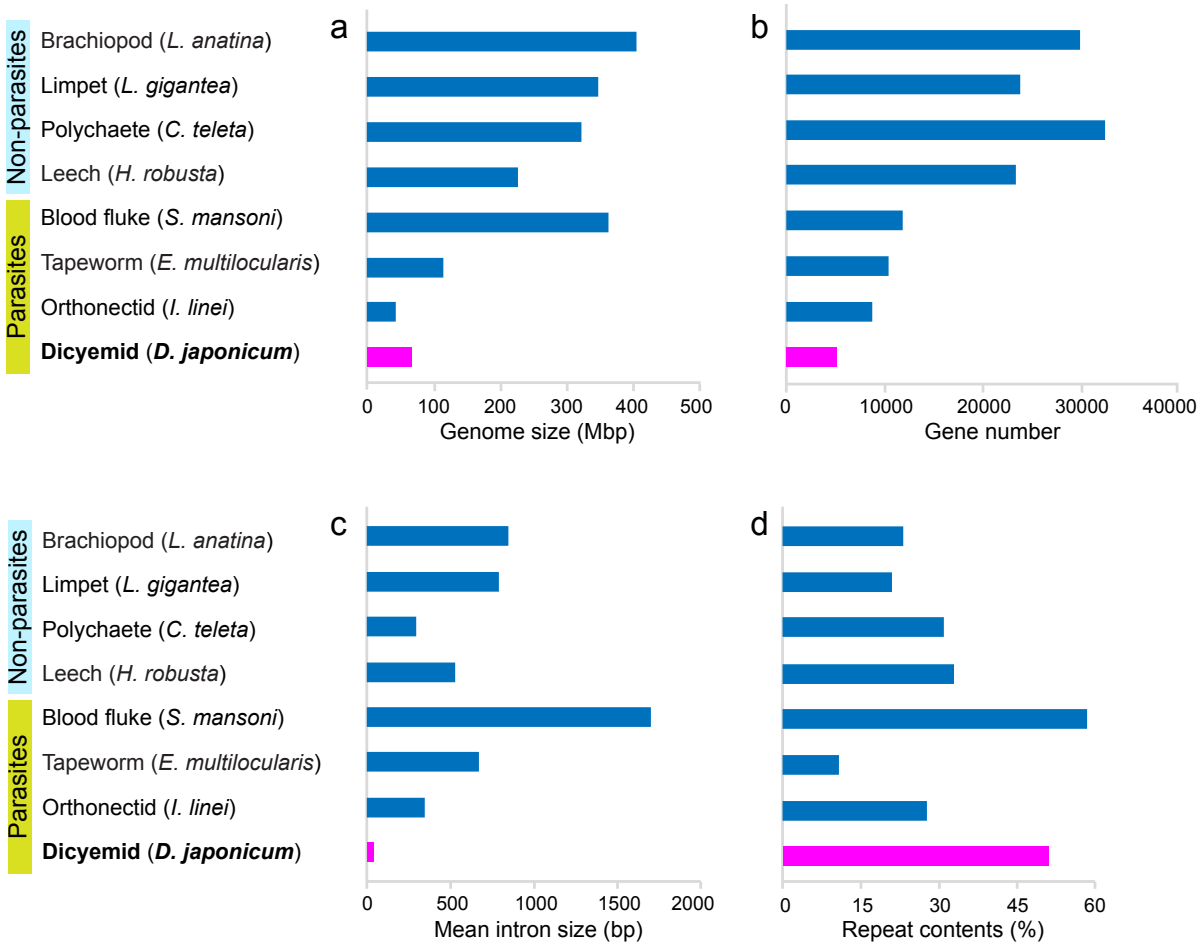

Figure 3.3 | The spiralians exhibit various genomic characteristics.

The predicted gene numbers of parasite lineages are less than half of the average gene number of non-parasite spiralians, suggesting that reduction of gene number could be a convergence of spiralian parasites. The genomic information of compared taxa are obtained from published sources as following: brachiopod (Luo et al., 2015); limpet, polychaete, and leech (Simakov et al., 2013); blood fluke (Berriman et al., 2009); tapeworm (Tsai et al., 2013), orthonectid (Mikhailov et al., 2016).

3.3.4 Pfam domain analysis

Domains contribute structural characteristics to particular proteins and also play roles in specific functions. To understand better the background knowledge of what constitutionally biological processes dicyemids retain, I examined the existence of functional domains which belong to some important transcription factors or signaling pathway genes. Comparing with other spiralians, dicyemids own less genes that contain functional domains in general (Tables

3.4 and 3.5). For example, only 16 homeobox domain-containing genes were found in dicyemids, the number of which was about quarter of those other parasitic spiralians possess. For Wnt pathway, only one Wnt family domain and one Dishevelled domain were found. In addition, some domains that are involved in development processes were not found in the dicyemid gene models such as fibroblast growth factor (FGF) domain, Hedgehog N-terminal signaling (HH_signal) domain, and CHRD domain in *Chordin* gene, which play a role in the dorsal-ventral patterning. Neuronal helix-loop-helix transcription factor (Neuro_bHLH) and Delta serrate ligand (DSL) domains which play roles in neuron cell fate determination were not detected as well. By contrast, the most abundant domains in dicyemids was protein kinase domain, which appeared in 142 genes. The next was AAA ATPase domain. Furthermore, WD40 domain, Ras domain, RNA recognition motif (RRM_1), tetratricopeptide repeat (TPR) domain, helicase (Helicase_C and DEAD) domains, and EF-hand domain were present in more than 50 genes (Table 3.6). However, even for these abundant domains, the numbers of genes containing these domains in *D. japonicum* were less than those in other spiralians (Tables 3.4, 3.5, and 3.6).

Table 3.4 | Number of genes with transcription factor domains in selected bilaterians

Pfam domain name	Pfam ID	Function	dja	ili	emu	sma	sme	lgi	obi	cte	hro	dme	cel	bfl	hsa
Homeobox	PF00046	Homeobox domain	16	65	77	77	133	141	96	183	242	164	125	127	244
zf-C2H2	PF00096	Zinc finger, C2H2 type	13	42	97	99	138	331	2086	325	234	338	157	986	709
HLH	PF00010	Helix-loop-helix DNA-binding domain	14	25	31	35	50	78	61	85	70	80	48	80	108
HMG_box	PF00505	HMG (high mobility group) box	11	13	20	20	47	29	71	25	66	37	23	45	56
Homeobox_KN	PF05920	Homeobox KN domain	10	19	34	27	49	50	27	66	115	42	41	40	79
zf-C4	PF00105	Zinc finger, C4 type (two domains)	8	12	15	21	32	36	33	38	50	47	325	29	46
Ets	PF00178	Ets-domain	7	3	9	9	22	10	13	13	22	15	17	13	28
bZIP_1	PF00170	bZIP transcription factor	5	6	14	13	31	38	27	33	28	39	39	38	50
Hormone_recep	PF00104	Ligand-binding domain of nuclear hormone receptor	4	10	13	14	15	32	35	38	32	42	320	29	48
ARID	PF01388	ARID/BRIGHT DNA binding domain	3	4	6	7	8	6	14	7	9	9	6	4	15
Pou	PF00157	Pou domain - N-terminal to homeobox domain	3	2	6	6	10	4	4	6	11	18	5	6	16
bZIP_Maf	PF03131	bZIP Maf transcription factor	2	1	8	8	9	13	11	13	16	22	19	15	35
Forkhead	PF00250	Fork head domain	2	21	16	16	33	31	19	47	31	27	32	31	50
GATA	PF00320	GATA zinc finger	2	3	5	6	8	7	10	16	15	11	27	8	19
RHD_DNA_bind	PF00554	Rel homology DNA-binding domain	2	0	1	0	4	4	5	3	4	12	0	2	10
CUT	PF02376	CUT domain	1	2	3	4	7	3	2	3	11	6	6	3	7
PAX	PF00292	Paired box' domain	1	5	4	4	14	8	6	9	10	20	12	5	9
Runt	PF00853	Runt domain	1	0	1	1	2	1	1	2	2	8	1	1	3
SRF-TF	PF00319	SRF-type transcription factor (DNA-binding and dimerisation domain)	1	2	3	4	5	3	3	2	6	3	2	4	5
Basic	PF01586	Myogenic Basic domain	0	1	1	1	1	1	1	1	1	2	0	4	4
DM	PF00751	DM DNA binding domain	0	1	2	4	6	4	2	5	3	4	13	10	7
GCM	PF03615	GCM motif protein	0	1	2	2	2	1	1	1	2	3	0	2	2
Hairy_orange	PF07527	Hairy Orange	0	1	0	1	1	18	5	11	2	12	0	13	11
Neuro_bHLH	PF12533	Neuronal helix-loop-helix transcription factor	0	0	1	1	1	1	1	2	1	0	1	1	4
OAR	PF03826	OAR domain	0	0	0	0	0	9	11	5	4	11	0	12	15
P53	PF00870	P53 DNA-binding domain	0	0	2	3	1	1	1	1	2	2	0	4	3
HPD	PF05044	Homeo-prospero domain	0	1	2	2	4	1	1	1	3	1	1	0	2
SCAN	PF02023	SCAN domain	0	0	0	0	0	2	3	0	2	0	0	0	60
T-box	PF00907	T-box	0	7	5	7	5	12	12	8	18	14	21	11	17
TF_AP-2	PF03299	Transcription factor AP-2	0	3	1	1	2	1	2	2	2	6	5	2	5
TF_Otx	PF03529	Otx1 transcription factor	0	0	0	0	0	0	0	0	0	0	0	1	3
zf-C2HC	PF01530	Zinc finger, C2HC type	0	0	2	2	5	3	6	5	3	7	2	2	7

dja, *Dicyema japonicum*; ili, *Intoshia linei*; emu, *Echinococcus multilocularis*; sma, *Schistosoma mansoni*; lgi, *Lottia gigantea*; obi, *Octopus bimaculoides*; cte, *Capitella teleta*; hro, *Hellobdella robusta*; dme, *Drosophila melanogaster*; cel, *Caenorhabditis elegans*; bfl, *Branchiostoma floridae*; hsa, *Homo sapiens*.

Table 3.5 | Number of genes with signaling pathway domains in selected bilaterians

Pfam domain name	Pfam ID	Function	dja	ili	emu	sma	sme	lgi	obi	cte	hro	dme	cel	bfl	hsa
G-alpha	PF00503	G-protein alpha subunit	14	25	26	32	66	36	33	42	35	36	50	31	49
RGS	PF00615	Regulator of G protein signaling domain	7	10	13	11	30	14	24	17	17	18	31	14	36
EGF	PF00008	EGF-like domain	4	25	27	25	38	109	169	225	102	87	74	541	127
DIX	PF00778	DIX domain	2	4	4	5	5	4	4	3	3	5	8	3	7
TGF_beta	PF00019	Transforming growth factor beta like domain	2	4	3	2	11	11	14	14	8	7	5	20	37
Notch	PF00066	LNR domain	1	0	2	1	3	2	3	6	3	3	2	3	6
Frizzled	PF01534	Frizzled/Smoothened family membrane region	1	2	6	7	12	5	6	8	6	10	7	7	12
STAT_bind	PF02864	STAT protein, DNA binding domain	1	0	0	0	2	2	2	6	3	1	5	4	7
STAT_int	PF02865	STAT protein, protein interaction domain	1	0	0	0	0	1	2	6	2	1	0	3	7
wnt	PF00110	wnt family	1	4	6	6	9	12	17	16	18	8	6	15	19
CHRD	PF07452	CHRD domain	0	0	0	0	0	1	1	0	0	6	0	1	1
Dishevelled	PF02377	Dishevelled specific domain	0	0	2	2	2	1	1	1	0	1	4	1	4
DSL	PF01414	Delta serrate ligand	0	1	12	7	36	11	9	8	5	10	15	11	4
FGF	PF00167	Fibroblast growth factor	0	1	0	0	1	2	4	1	1	4	2	9	25
Focal_AT	PF03623	Focal adhesion targeting region	0	0	1	1	1	1	2	1	1	3	3	2	2
G-gamma	PF00631	GGL domain	0	1	4	2	6	4	6	6	6	8	7	1	16
HH_signal	PF01085	Hedgehog amino-terminal signalling domain	0	0	1	1	1	3	3	1	1	1	0	3	3
MCPsignal	PF00015	Methyl-accepting chemotaxis protein (MCP) signalling domain	0	0	1	0	0	7	1	1	0	1	0	11	0
PDGF	PF00341	PDGF/VEGF domain	0	0	0	0	0	2	2	2	1	6	1	2	9
Phe_ZIP	PF08916	Phenylalanine zipper	0	0	0	0	0	0	0	0	0	1	0	0	3
PTN_MK_C	PF01091	PTN/MK heparin-binding protein family, C-terminal domain	0	0	0	0	0	0	2	1	0	4	0	2	2
PTN_MK_N	PF05196	PTN/MK heparin-binding protein family, N-terminal domain	0	0	0	0	0	1	1	0	1	0	0	1	2
Rabaptin	PF03528	Rabaptin	0	0	0	1	1	1	1	1	1	0	0	2	2
STAT_alpha	PF01017	STAT protein, all-alpha domain	0	0	0	0	0	1	0	5	1	1	1	3	7
TGFb_propeptide	PF00688	TGF-beta propeptide	0	0	1	0	3	10	11	8	5	7	1	15	24

dja, *Dicyema japonicum*; ili, *Intoshia linei*; emu, *Echinococcus multilocularis*; sma, *Schistosoma mansoni*; lgi, *Lottia gigantea*; obi, *Octopus bimaculoides*; cte, *Capitella teleta*; hro, *Hellobdella robusta*; dme, *Drosophila melanogaster*; cel, *Caenorhabditis elegans*; bfl, *Branchiostoma floridae*; hsa, *Homo sapiens*.

Table 3.6 | The most abundant domains in *Dicyema japonicum*

Pfam domain name	Pfam ID	Function	dja	ili	emu	sma	sme	lgi	obi	cte	hro	dme	cel	bfl	hsa
Pkinase	PF00069	Protein kinase domain	142	171	245	245	725	322	389	379	474	447	562	556	482
Pkinase_Tyr	PF07714	Protein tyrosine kinase	137	163	238	239	687	318	381	365	461	442	541	556	477
WD40	PF00400	WD domain, G-beta repeat	86	124	134	159	204	211	290	232	188	207	168	244	252
RRM_1	PF00076	RNA recognition motif	67	103	139	151	224	136	208	145	209	265	164	119	214
Helicase_C	PF00271	Helicase conserved C-terminal domain	62	56	82	83	133	83	118	94	91	99	107	96	107
ANAPC4_WD 40	PF12894	Anaphase-promoting complex subunit 4 WD40 domain	60	77	89	110	151	174	203	172	148	165	124	192	200
Roc	PF08477	Ras of Complex, Roc, domain of DAPkinase	56	73	66	83	200	136	131	142	118	133	97	206	185
Ras	PF00071	Ras family	55	68	65	78	180	137	134	134	119	127	93	191	178
DEAD	PF00270	DEAD/DEAH box helicase	55	54	67	68	123	70	90	86	81	81	89	92	90
TPR_2	PF07719	Tetratricopeptide repeat	54	49	59	81	107	106	140	119	86	102	73	396	142
Kinase-like	PF14531	Kinase-like	53	45	85	77	250	128	118	106	161	134	151	120	185
Arf	PF00025	ADP-ribosylation factor family	50	65	63	82	187	123	115	126	104	114	103	139	180
EF-hand_6	PF13405	EF-hand domain	49	68	87	100	212	202	176	175	146	151	108	248	173
EF-hand_7	PF13499	EF-hand domain pair	48	69	90	104	213	198	176	182	142	152	112	250	179
EF-hand_1	PF00036	EF hand	48	68	85	103	213	207	181	184	144	156	116	252	187
TPR_1	PF00515	Tetratricopeptide repeat	44	40	49	67	100	101	124	105	78	88	64	379	128
TPR_8	PF13181	Tetratricopeptide repeat	43	34	46	60	92	86	108	94	66	82	55	358	114
MMR_HSR1	PF01926	ribosome-binding GTPase	43	74	68	91	245	160	127	146	128	156	122	204	195
AAA_22	PF13401	AAA domain	42	51	61	66	121	109	118	134	87	130	104	147	120
AAA	PF00004	ATPase family associated with various cellular activities (AAA)	41	54	60	71	119	86	100	99	76	108	83	102	90
EF-hand_8	PF13833	EF-hand domain pair	40	56	67	83	154	143	121	142	117	123	83	194	140
ResIII	PF04851	Type III restriction enzyme, res subunit	40	47	47	53	107	74	71	73	76	75	69	73	88
TPR_12	PF13424	Tetratricopeptide repeat	39	37	45	63	88	98	109	97	69	70	60	392	111
LRR_4	PF12799.5	Leucine Rich repeats	34	50	60	60	130	163	186	434	120	169	93	964	251
AAA_5	PF07728	AAA domain (dynein-related subfamily)	34	45	50	52	86	71	82	77	62	82	65	78	65
TPR_19	PF14559	Tetratricopeptide repeat	33	28	44	61	73	74	92	77	47	68	59	229	98
AAA_16	PF13191.4	AAA ATPase domain	33	44	58	66	120	101	106	166	73	135	98	134	129
TPR_14	PF13428.4	Tetratricopeptide repeat	31	22	46	67	59	79	91	81	55	67	60	314	108
EF-hand_5	PF13202.4	EF hand	30	50	65	72	144	158	121	133	113	107	83	211	120

dja, *Dicyema japonicum*; ili, *Intoshia linei*; emu, *Echinococcus multilocularis*; sma, *Schistosoma mansoni*; lgi, *Lottia gigantea*; obi, *Octopus bimaculoides*; cte, *Capitella teleta*; hro, *Hellobdella robusta*; dme, *Drosophila melanogaster*; cel, *Caenorhabditis elegans*; bfl, *Branchiostoma floridae*; hsa, *Homo sapiens.*

3.3.5 Pathway analysis

Instead of regulation by a single gene, many biological processes are controlled by a group of genes on a function-specialized gene pathway. In order to understand how dicyemid gene models are conserved at the pathway level, I assigned predicted genes to KEGG orthologs (KO) and mapped to KEGG reference pathways. I also compared the KOs numbers in each pathway among eighteen bilaterians, including seven parasitic lineages and eleven non-parasitic lineages across the Spiralia, Ecdysozoa, and Deuterostomia. Within the dicyemid gene models, 2,305 KOs could be annotated. These KOs belonged to 345 KEGG pathways, which were about 5% less than the average number (362) of KEGG pathways among the compared species (Figure 3.4). This indicates that *D. japonicum* retains most of the pathways existing in other metazoans. However, it was noticed that *D. japonicum* has least KOs involved in most pathways, and spiralian parasite lineages generally have less KOs comparing with their close-related non-parasitic lineages (Figure 3.5). Particularly in metabolism pathways, both spiralian and ecdysozoan parasites possess less KOs than non-parasite lineages (Figure 3.6).

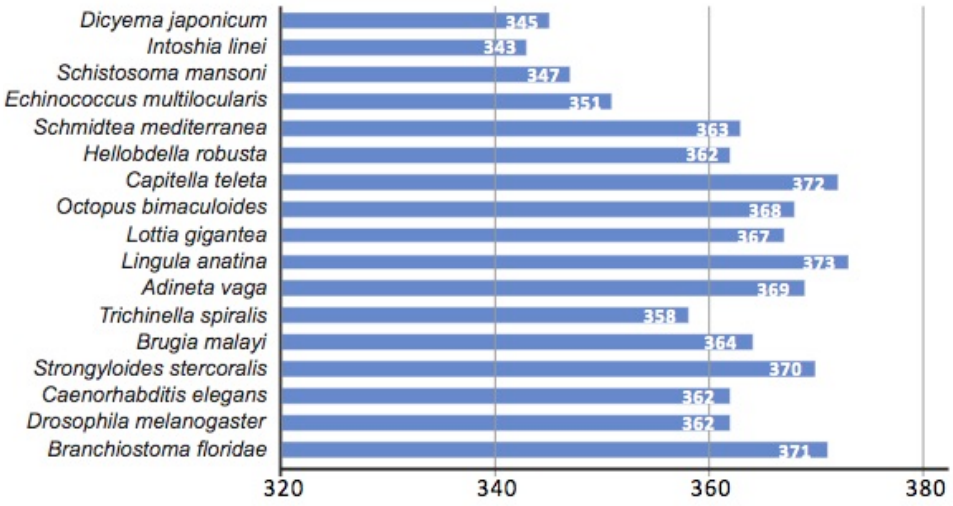

Figure 3.4 | Number of annotated KEGG pathways in the selected bilaterians.
Parasite lineages hold KEGG pathways 5% less than non-parasite lineages, although parasite lineages only possess half of predicted gene number of non-parasite species.

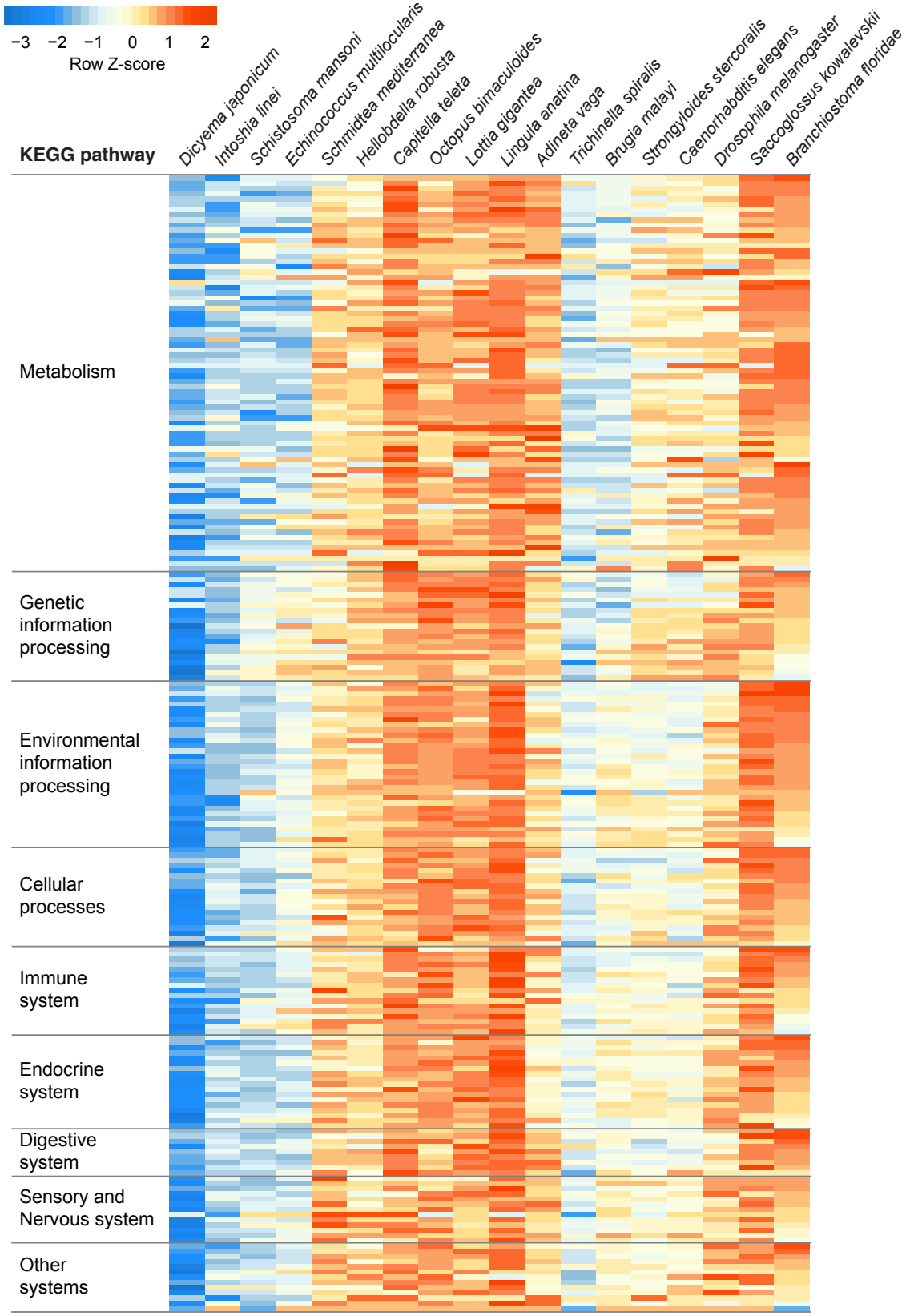

Figure 3.5 | Heatmap of gene number on KEGG pathways in bilaterians.

In general, parasites have less genes involved in most KEGG pathways than non-parasite species. Furthermore, *Dicyema* exhibits least genes among all parasites, reflecting the reduction of whole gene set and body organization.

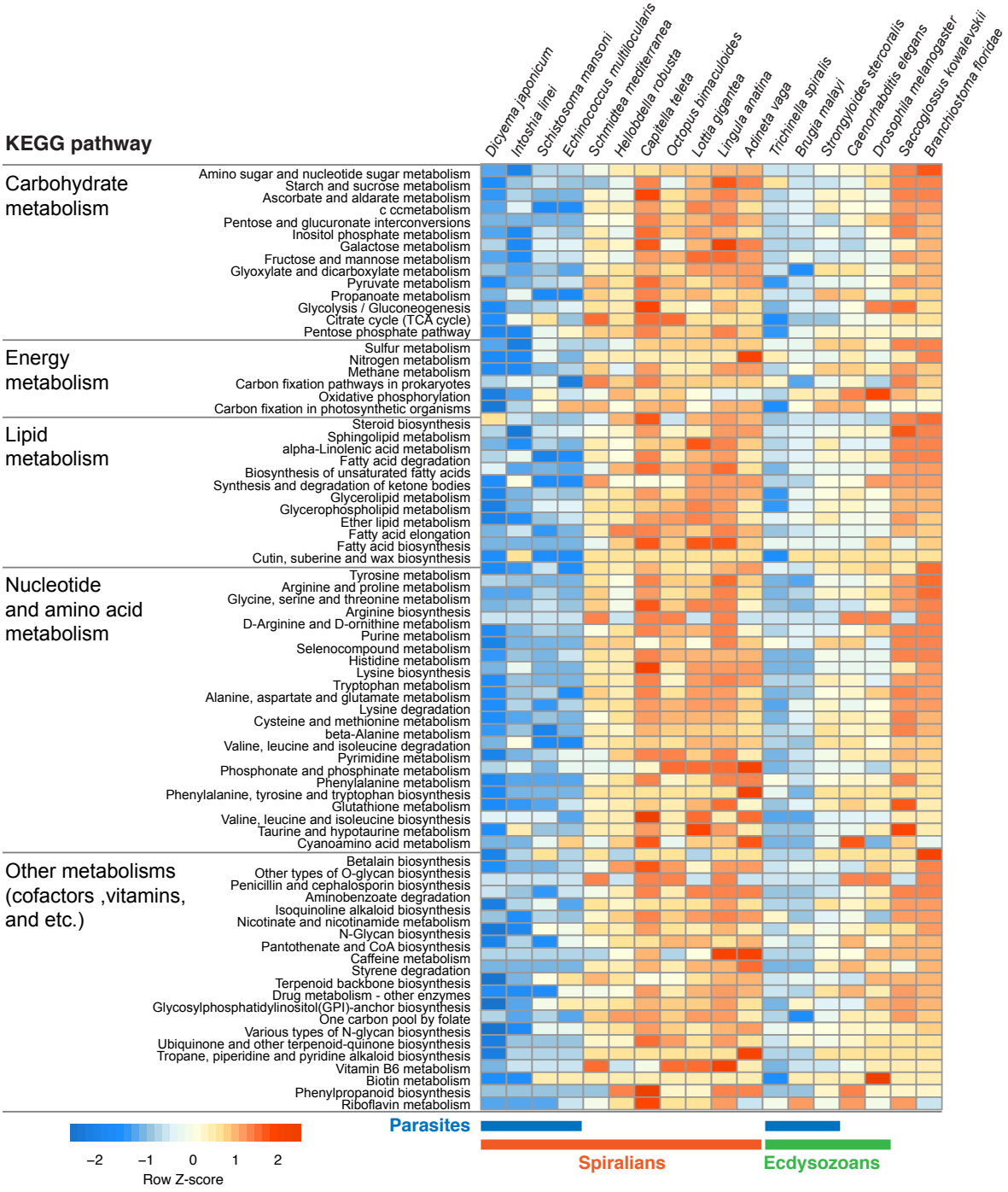

Figure 3.6 | Parasites possess less genes in KEGG metabolism pathways.

Not only in the spiralians but also ecdysozoans, except *Strongyloides stercoralis*, parasite species retain much less metabolism-associated genes than non-parasite species. This could be a convergence of parasites at least in spiralian lineages.

3.3.6 Orthologous groups clustering

The Pfam domain and KEGG pathway analyses showed that *D. japonicum* may lose many genes and retain approximately 5,000 genes. It raised questions how many orthologs are conserved to share with other spiralians and, on the other hand, what orthologous groups are specifically expanded in dicyemids. The gene models of 16 selected species were categorized into 12,297 clusters, which were identified by clustering orthologous groups by OrthoMCL and shared by at least three species. In *D. japonicum*, 3,966 proteins larger than 50 amino acids were categorized into 2,958 gene families, while 1,046 proteins were classified as singletons, those not clustered into any orthologous group. *D. japonicum* possessed 313 lineage-specific gene clusters and shared 2,645 gene clusters with other analyzed spiralian species (Figure 3.7). Dicyemids shared less orthologous groups with orthonectids than platyhelminthes and lophotrochozoans, even though dicyemids have a close affinity to the orthonectids (Figure 3.7).

To explore if any gene family has expanded, I counted the gene copy number of each gene clusters. In forty-six gene clusters, *D. japonicum* possessed more than four copies, although one-third of them were uncharacterized protein (Table 3.7). Within the spiralian-shared gene families, 1,811 gene clusters were shared by all spiralians analyzed (Figure 3.7). Some of these multiple copy gene clusters may be involved in cytoskeleton formation and cilia beating, e.g., dynein, actin, myosin, alpha-actinin, and beta-tubulin. In addition, three multiple copy gene clusters may be involved in nutrient-uptake via endocytosis (Figure 3.8). Macrophage mannose receptor and vacuolar protein sorting-associated protein may mediate the endocytosis and protein trafficking to lysosome.

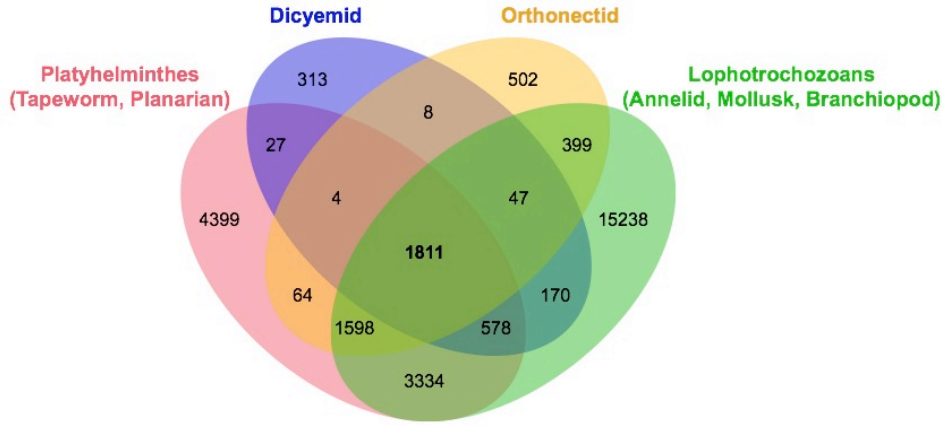

Figure 3.7 | Venn diagram of dicyemid orthologous groups shared with other spiralians. Dicyemids possess 313 lineage-specific gene clusters and share 2,645 gene clusters with other spiralian species analyzed.

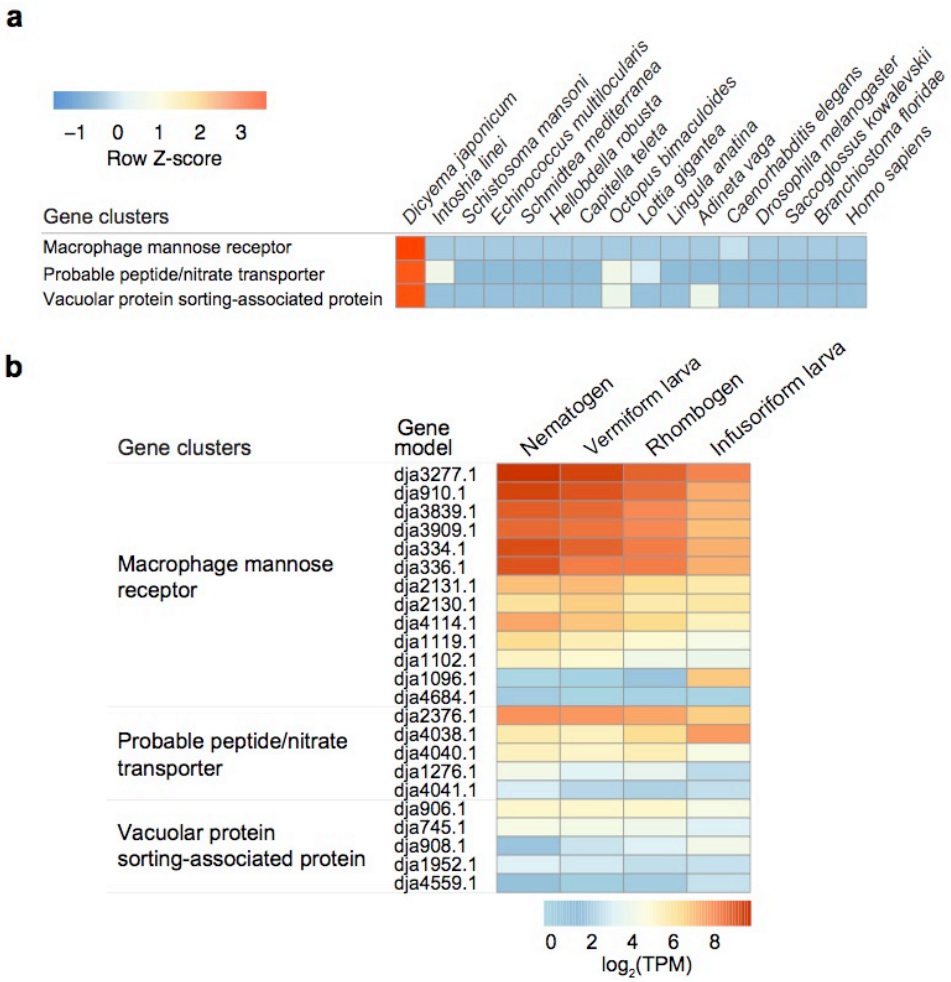

Figure 3.8 | Multiple copy genes could be associated with nutrient absorption. (a) *Dicyema* possesses multiple copy genes which might be involved in nutrient uptake. (b) The nutrient-uptake-related gene expression at four life-cycle stages. Some gene copies show stage-specific expression.

Table 3.7 | Multiple-copy gene clusters of *Dicyema japonicum*

Gene cluster	Copy number	Annotation
OG_2	16	Dynein heavy chain 1
OG_10592 *	15	Macrophage mannose receptor
OG_6596 *	10	Uncharacterized
OG_52	9	Actin
OG_1661 *	9	Probable serine carboxypeptidase
OG_8859 *	7	Probable peptide/nitrate transporter
OG_18681	7	Proton-coupled folate transporter
OG_2865 *	6	Malonyl-CoA-acyl carrier protein transacylase
OG_6419 *	6	Uncharacterized
OG_20785	6	Rho GTPase activator
OG_20788	6	Homeodomain-interacting protein kinase
OG_22151	6	Uncharacterized
OG_35	5	Myosin
OG_86	5	ATP-dependent RNA helicase
OG_603	5	Galactosyltransferase
OG_1951 *	5	Pyruvate dehydrogenase phosphatase
OG_3958 *	5	Vacuolar protein sorting-associated protein
OG_13754 *	5	Uncharacterized
OG_23486	5	Zinc finger protein
OG_23487	5	Uncharacterized
OG_23490	5	Uncharacterized
OG_23491	5	Aspartate/tyrosine/aromatic aminotransferase
OG_25447	5	Uncharacterized
OG_25448	5	Uncharacterized
OG_25449	5	Uncharacterized
OG_71	4	Tubulin beta
OG_157	4	Ferritin heavy chain
OG_520	4	Alpha-actinin
OG_584	4	DNA-directed RNA polymerase III subunit
OG_901	4	Protein polybromo-1
OG_1179 *	4	WD repeat domain phosphoinositide-interacting protein
OG_2280 *	4	RNA-binding protein MEX3B
OG_2515 *	4	Lysosomal Pro-X carboxypeptidase
OG_1973	4	Putative inorganic phosphate transporter
OG_11823	4	Uncharacterized
OG_15771 *	4	Ubiquitin-protein ligase
OG_17079 *	4	Cell division control protein 42
OG_18679 *	4	Uncharacterized
OG_23489	4	EF-hand calcium-binding protein
OG_27582	4	Bromodomain-domain-containing protein
OG_27583	4	IQ domain-containing protein
OG_27584	4	Serine hydroxymethyltransferase
OG_30590	4	Uncharacterized
OG_30591	4	Uncharacterized
OG_30593	4	Uncharacterized
OG_30594	4	Uncharacterized

* Z-score claculated among compared bilaterians is larger than 2

3.3.7 Identifications of Hox cluster genes

Hox cluster genes play key roles in regulating the body segmentation along the anterior-posterior axis, and the Hox cluster is often regarded as an indicator of genome assembly quality. The reciprocal BLAST searches of sixteen homeobox-containing gene models showed that *D. japonicum* retains four putative Hox genes (Table 3.8). They were further characterized by the phylogenetic analysis based on the amino acid sequences of homeobox domains (Figure 3.9). The phylogenetic relationships of analyzed Hox genes were consistent with previous reports (Simakov et al., 2013). Two putative Hox genes were likely *Hox1* and *Post*-like genes, and they could be representative for anterior and posterior Hox genes, respectively. The other two were *Lox5*-like genes, which could be representative for central Hox genes. One to them had the BLAST best hit to the previously published *DoxC* gene. However, these four genes were located on different scaffolds, thus the Hox cluster was not observed in *Dicyema japonicum.*

Table 3.8 | The annotation of homeobox domain genes in *Dicyema japonicum*

Gene ID	bbh vs HomeoboxDB	bbh vs UniProt database	BLASTP vs NCBI nr protein database
dja56.1	Frog | zfhx2	N/A	Zinc finger homeobox protein 4 | ELK27508.1
dja68.1	Frog | onecut1	P70512 | HNF6_RAT	HNF6a | AAW30419.1
dja131.1 *	Amphioxus | Hox6	O13074 | HXB4A_TAKRU	DoxC | BAA85262.1
dja291.1 *	N/A	Q90423 | HXB1B_DANRE	Hoxb1b protein | CAC34568.1
dja355.1	N/A	N/A	Homeobox protein SIX4 | KRZ84160.1
dja1156.1	N/A	N/A	SIX homeobox 2b | NP_001122206.1
dja1243.1	Frog | sial-4	N/A	Protein gooseberry-like | XP_022908761.1
dja1506.1	Frog | pou2f2	P70030 | P3F2B_XENLA	POU domain, transcription factor 2-like | XP_021467820.1
dja1633.1 *	N/A	N/A	Lox5, partial | ALQ28243.1
dja1943.1	Frog | mxtxl	Q9PVM0 | OTX5A_XENLA	Homeobox protein otx5 | CDW54751.1
dja2692.1	Honeybee | Hth	Q5U4X3 | MEI3A_XENLA	Homeobox protein meis3-A | NP_001081866.1
dja2996.1	Beetle | Lim1	P53411 | LHX1_CHICK	LIM/homeobox protein Lhx1 | ELK35898.1
dja3400.1	N/A	N/A	Unc-39 | PDM81100.1
dja3448.1 *	Zebrafish | hoxb10a	N/A	Homeobox D10, partial | AFJ72863.1
dja3851.1	Chicken | PBX1	P40425 | PBX2_HUMAN	PBX | AGH32779.1
dja4033.1	Frog | six6	A2D5H2 | SIX1_LAGLA	Homeobox protein SIX6 | OQV18562.1

* putative Hox gene

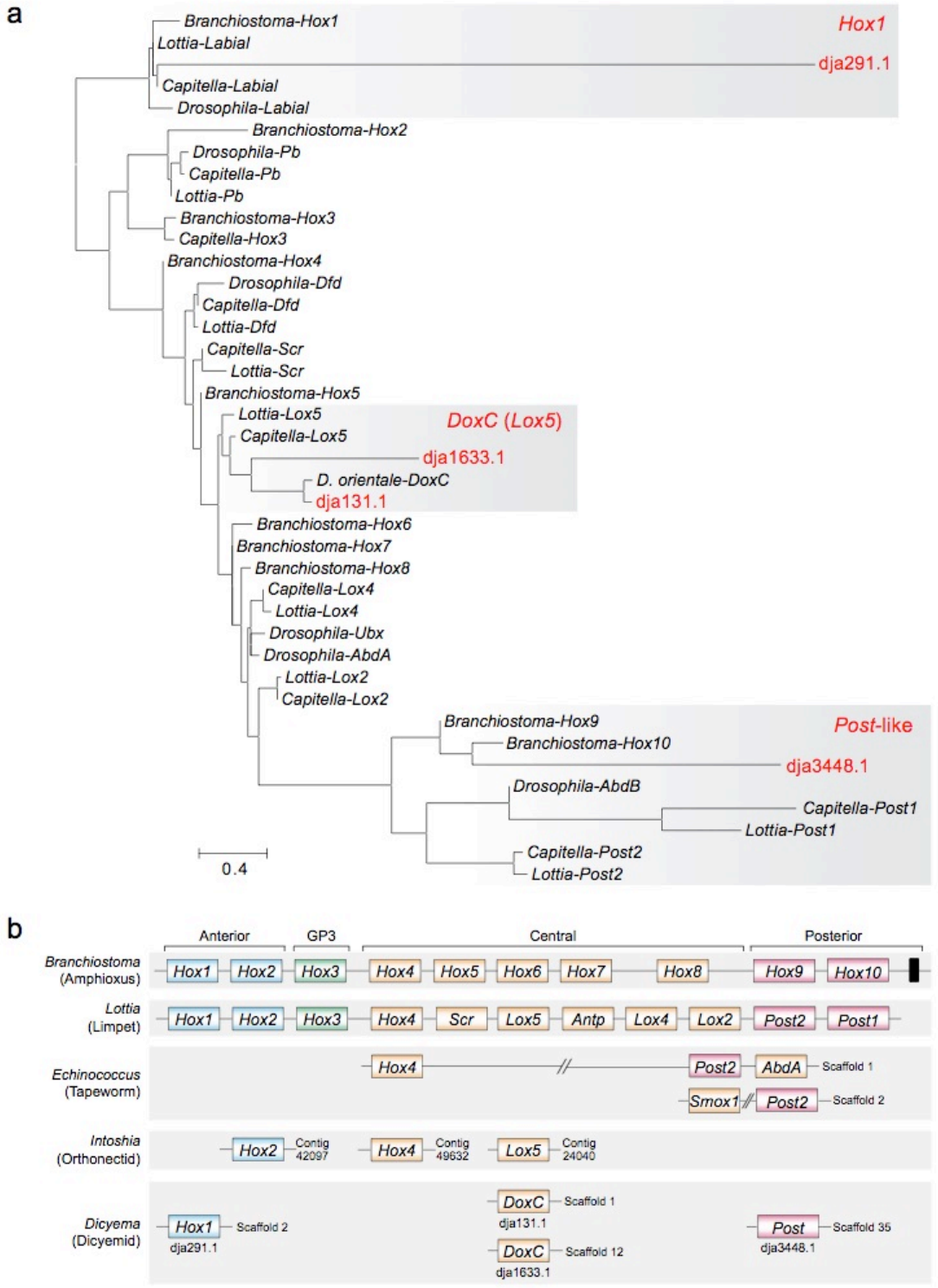

Figure 3.9 | *Dicyema* retains four putative Hox genes.

(a) Phylogenetic tree inferred from the homeobox domain using the maximum likelihood method. (b) Hox gene synteny in selected bilaterians. The scattered, non-clustered Hox gene structure commonly occurs in three parasitic spiralian lineages. Black block represents the rest of the posterior Hox genes in *Branchiostoma*. Double slashes signify non-continuous linkage between two genes.

3.3.8 Searching for G protein-coupled receptors

G protein-coupled receptors (GPCR) play crucial roles in sensory functions, receiving stimuli such as neurotransmitters and hormones outside the cell and activating down-stream signal transduction activities in the cell. GPCR family consists of hundreds of receptor proteins and could be further categorized into six classes according to the GPCRdb database (Pándy-Szekeres et al., 2017). Here, I found thirteen genes encoding GPCR proteins, which belonged to metabotropic glutamate receptor, rhodopsin-like, and secretin-like GPCR classes (Table 3.9). Six of them contained 7tm domains which are regarded as a common structure for GPCR family (Table 3.9). By the reciprocal BLASTP search between gene models and GPCRs database, nine gene models had bi-directional best hits and two of them contained 7tm domains. Moreover, four other 7tm domain-containing gene models were also confirmed as putative GPCRs by BLASTP searching against NCBI nr protein database manually, althoough they did not have bi-directional BLAST best hits against GPCRs database.

Table 3.9 | The putative G protein-coupled receptors in *Dicyema japonicum*

Gene ID	UniProt ID	Pfam domain			Annotation
		7tm_1	7tm_2	7tm_3	
dja442.1 *	Q5Y4N8				Adhesion G protein-coupled receptor E2
dja499.1	Q86Y34		✓		Adhesion G protein-coupled receptor G3
dja766.1 *	Q9R1M7				Glutamate receptor ionotropic, NMDA 3A
dja1033.1 *	G5ECX0		✓		Latrophilin-like protein LAT-2
dja1151.1 *	Q05586				Glutamate receptor ionotropic, NMDA 1
dja1543.1 *	P04755				Acetylcholine receptor subunit beta-like 1
dja2203.1 *	P09478				Acetylcholine receptor subunit alpha-like 1
dja3017.1	Q9R0M0		✓		Cadherin EGF LAG seven-pass G-type receptor 2
dja3130.1 *	P10824				Guanine nucleotide-binding protein G subunit alpha-1
dja3464.1 *	O35161				Cadherin EGF LAG seven-pass G-type receptor 1
dja4077.1 *	Q9VZW5	✓			FMRFamide receptor
dja4293.1	Q9Z0R8			✓	Taste receptor type 1 member 1
dja4301.1	P35400			✓	Metabotropic glutamate receptor 7

* Gene model vs UniProt BLASTP bi-direction best hit

3.3.9 *Dicyema japonicum* mitochondrial genome composition

Using amino acid sequences of spiralian mitochondrial genes, I searched for mitochondrial fragments in the *Dicyema* genome assembly, but I failed to find a scaffold containing the entire

mitochondrial genome. Only *cox1* of different spiralian species had TBLASTN search-hits to the present genome assembly. However, searching the *Dicyema* transcriptome assembly and PacBio long reads, there were 10 and 12 mitochondrial genes that have search-hits, respectively. I retrieved the sequences of transcripts with search-hits in the *Dicyema* transcriptome assembly as queries and performed BLASTN search against PacBio reads to obtain more position information on the PacBio reads.

No sequences resembling *atp8* could be found in all searches, but *atp6*, *nad2*, *nd4l* and *nd6* were newly identified. I found 25 PacBio reads containing at least two mitochondrial genes (Figure 3.10). By manual verification, I was able to determine the gene order of twelve mitochondrial genes and close the circular genome to obtain putative *Dicyema* mitochondrial genome composition (Figure 3.10), although *nd2*, *nd6*, and *cox2* were only supported by one PacBio reads. These twelve mitochondrial genes were found in two gene blocks of opposing transcriptional orientations. Six continuously linked genes were found on the forward (positive) strand (*nd1-nd6-cob-nd4l-nd4-nd5*) and the other six genes were on the reverse strand (Figure 3.10). In some PacBio reads, although I could find two to three *cox1* or *cox2* fragments on single PacBio read, these fragments did not display the tandem repeat structure.

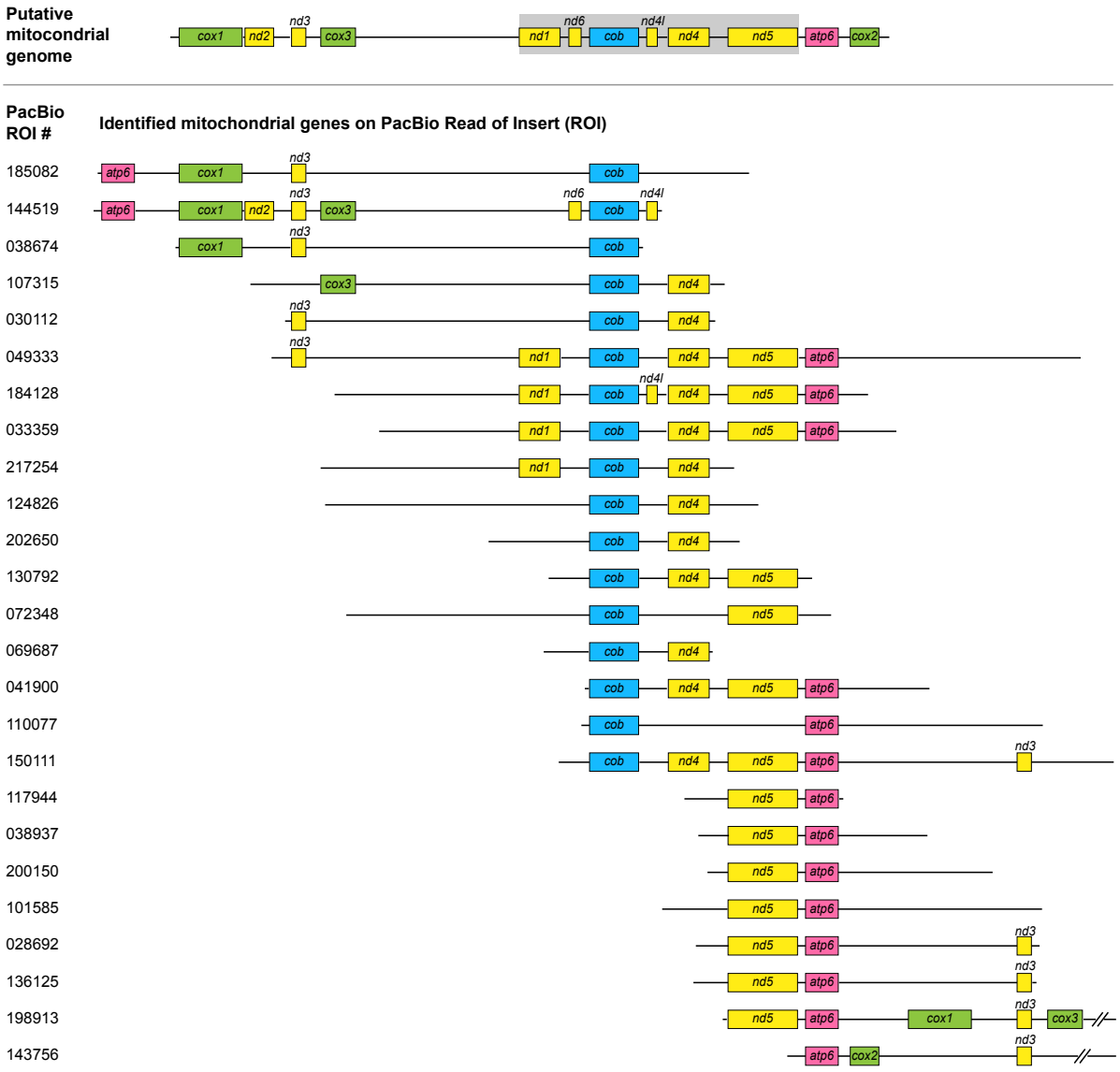

Figure 3.10 | Putative *Dicyema* mitochondrial genome composition.

The putative mitochondrial genome composition was reconstructed with the supports of 25 PacBio reads. Twelve mitochondrial genes are annotated in *Dicyama*, and display gene blocks of opposing transcriptional orientations. Gene orders were assumed that *cox1* reoriented as the first gene. Gray block shows that genes are annotated on the positive strand.

3.4 Discussion

Previous studies suggested that the dicyemid genome seems to acquire some unique characteristics during the process of parasitism (Awata et al., 2007; Ogino et al., 2010). Thus, decoding the dicyemid genome would provide better comprehension of the genomic evolution of parasitism, supplementary to the knowledge mainly on well-studied pathogenic parasites in human. After *de novo* assembling the genome and predicting the gene models of *Dicyema*, I implemented the comparative analyses on functional domains and pathways among species selected for representatives of major bilaterian taxa. I also explored gene family expansions and annotated genes which might be able to explain how dicyemids adapt to their lifestyle with the simplicity of their body plan.

Although the size of genome of *D. japonicum* is relatively small, the interferences of high heterozygosity and unpreventable contamination caused the *de novo* assembling more challenging. The customized pipeline successfully overcame the difficulties and assembled the genome of *D. japonicum* with a competent quality. The genome assembly was highly contiguous and covers over 98% transcripts included in the transcriptome assembly. Comparing with other spiralians, *D. japonicum* has the shortest average size of introns, which results in a compact genome. Conversely, the blood fluke possesses an average intron size over 1.5 Kbp, even much longer than that of non-parasitic lineages like mollusks and annelids, leading to the largest parasite genome size among spiralians (Figure 3.3c). Changes in intron length appear to reflect congruent changes in genome size as shown by comparison between dicyemid and blood fluke (Figure 3.3) and the previously reported cases in nematodes (Blaxter & Koutsovoulos, 2014). In addition, the genome size of bold fluke is approximately four-fold of that of dicyemid. However, the repeat contents occupy large portion of both bold fluke and dicyemid genomes. It suggests that the ratio of repeat contents alter the composition of each genome but independent of the genome size. Furthermore, the ratio of repeat contents of

dicyemid is 50% higher than that of orthonectid. In other words, less portion of the genome is for protein-coding genes in dicyemid than that of orthonectid. This probably could explain why the predicted gene number of dicyemid is 40% less than that of orthonectid, even dicyemid have bigger genome size and shorter mean intron size (Figure 3.3). Given the wide variance among compared eight spiralians, the small genome size and reduced intron size are incoherent with the simplified morphology as the convergence for spiralian parasites. These echo that the evolution of each parasitic lineage was driven by different forces of each evolutionary paths in parallel. Yet, *D. japonicum* and other three compared spiralian parasites hold predicted gene numbers less than half of that non-parasite spiralians. To save the energy costs for replicating redundant genetic materials, parasites may largely dismiss genes which are inessential for parasitic lifestyle during evolution to parasites. Therefore, the reduced gene number could be a possible convergence of spiralian parasites (Figure 3.3b).

Nonetheless, the reduced gene number may confuse the completeness assessment for the *de novo* assembling. This may raise an interpretation that some genes are missed owing to the inefficiency of *de novo* assembling, especially for the non-traditional model organisms with unique genome properties. CEGMA has been broadly used to assess the completeness of a genome assembly by examining the occurrence ratio of core eukaryotic genes (Parra et al., 2007). The occurrence rates of fully and partially covered genes are over 80% and 90%, respectively, in several high-quality spiralian genomes (Luo et al., 2018; Luo et al., 2015; Simakov et al., 2013; Tsai et al., 2013). However, the occurrence rates of fully and partially covered genes were around 55% and 80% in the genomes of *D. japonicum*. Similar occurrence rates are observed in blood fluke as well (Berriman et al., 2009). This may reflect the nature of these parasite genomes, in which predicted gene number are reduced and some core eukaryotic genes may be lost (Figure 3.3b). It is likely that overall consideration of CEGMA test and other measures, e.g., coverage ratio of transcripts from the transcriptome assembly, would provide

more reliable assessment for evaluating the genome assembly completeness, principally for the parasitic lineages.

Consistent with the reduced gene number and simplified morphology, most domains in the Pfam database showed much less occurrences in *D. japonicum*, even comparing to other spiralian parasites. Going through the genetical and morphological reductions that are involved in adaptations to the parasitic lifestyle, some domains with enzyme functions, such as protein kinase, ATPase, and helicase domains, were comparatively abundant in *D. japonicum*. This indicates that these domains may crucially support the basic physiological requirements for the survival of dicyemids. Although dicyemids have complex life cycles, their biological tasks may be mainly nutrition-uptake and reproduction. The domains (or genes) for other than these two primary tasks could be eliminated in such organisms with only around 30 constituting cells. From the pathway perspective, dicyemids retain most of the pathways existing in other metazoans. However, dicyemids possess least genes in most pathways among other parasites, suggesting that these pathways are still essential in most physiological processes but probably simplified and less genes are involved in dicyemids. In other words, dicyemids retain the most essential genes of each pathway, corresponding to the simplest body organization and the specificity of physiological tasks for each life-cycle stage. Both spiralian and ecdysozoan parasites possess less genes in most pathways than non-parasite lineages, particularly in metabolism pathways. This might show a convergence for parasitism evolution.

The Hox cluster is often used for assessing the completeness and contiguity of genome assembly (Shields et al., 2018). However, dicyemids retain only four Hox genes with disorganizing cluster structure. In contrast, most transcripts (98.4%) in the transcriptome assembly from a mixed-stage sample could be aligned to the current genome assembly with the criteria of query coverage and identity more than 60%, endorsing that the current genome assembly covers most expressed genes. In addition, the conspicuous reduction of Hox cluster

genes has been reported in other parasites such as orthonectids (Mikhailov et al., 2016) and tapeworms (Tsai et al., 2013), which perhaps reflects the convergent simplification of parasitic body organization. Moreover, *Paedocypris* fishes also show extensive *Hox* gene loss adapting to an extreme habitat (Malmstrøm et al., 2018). These studies raise the doubt whether Hox cluster is a proper indicator for completeness and contiguity of genome assembly, particularly for the parasitic lineages.

Comparing with the closest parasitic lineage, orthonectids, which retain one anterior (*Hox2*) and two central Hox genes (*Hox4* and *Lox5*), dicyemids keep one anterior (*Hox1*), two central (*DoxC* and *Lox5*-like*)*, and one posterior Hox genes (*Post*-like), respectively. Although the two organisms utilize different anterior Hox genes, the remaining gene set seems sufficient to provide the essential regulatory mechanism for basic anterior-posterior polarity, even they have no specialized segmentations. The phylogenetic analysis showed that one putative may be orthologous to the *DoxC* gene identified in *D. orientale* (Kobayashi et al., 1999). However, the other *Lox5*-like gene was not cloned and described in that report. In contrast to the other Hox genes, *DoxC* and *Lox5*-like genes were retained in dicyemids, suggesting that they may be responsible for crucial developmental processes or function at different stages of life cycle. The putative *Hox1* gene in the *Dicyema* genome encoded only 69 amino acids and whether it accomplishes similar function as its homologs in other organisms remains to be confirmed.

In association with receipt and transduction of a broad variety of chemical ligands, the expansion of GPCR family in metazoans has been reported (Srivastava et al., 2010). Due to the complexity of life cycles, I suspected that dicyemids have some specialized GPCRs to deal with their unusual living environment and lifestyle. However, instead of gene family expansion, dicyemids highly reduce GPCR gene family comparing with other bilaterians which usually possess more than dozens of GPCR genes (Cardoso et al., 2014; Simakov et al., 2013; Adema et al., 2017). Only six 7tm transmembrane domain-containing genes were identified in

D. japonicum. They are metabotropic glutamate receptors, adhesion G protein-coupled receptors, and neuropeptide Y (NPY) receptor (Table 3.7). Among them, cadherin EGF LAG seven-pass G-type receptors, a subclass of adhesion GPCRs, are speculated that they play a role in contact-mediated communication and are involved in many biological processes during embryonic development including neural cell differentiation and neural tube closure (Wang et al., 2014). On the other hand, NPY receptor is a multireceptor/multiligand system, in which it could be activated by three types of ligands with different potency (Pedragosa-Badia et al., 2013). Despite the lack of nervous system, these neuropeptide receptor and nervous system development-related genes still remain in dicyemids, indicating that they may provide functions to unspecified machinery of sensory transduction and signaling. Furthermore, even though dicyemids have complex life cycles, they may only sense few key chemicals to trigger certain critical biological processes for sustaining their survival owing to retaining and limited number of GPCR genes.

In contrast to gene reduction mentioned above, dicyemids display multiple copies on possible nutrient-uptake-associated gene families than other compared taxa. This provides genetic evidence to support previously proposed hypothesis that dicyemids may directly absorb low molecular-weight nutrients from the urine of host through endocytosis (Ridley, 1968). The present results indicate that parasites may reduce gene number and body organization, which are not required anymore for the parasitic lifestyle. Meanwhile, they may retain or expand essential gene families associated with particular biological process to compensate for the loss of tissues or organs.

A previous study reported that the canonical high-molecular-weight mitochondrial DNA is present in dicyemids (Awata et al., 2005). However, the mitochondrial gene composition of dicyemid remains unsolved, and the canonical high-molecular-weight mitochondrial circle may break during the development process into mini-circles with single-

gene tandem repeat structures. In the present study, I failed to find a long genome assembly scaffold or PacBio read that constitutes an entire mitochondrial genome, even though the mean length of scaffolds and PacBio reads are much longer than most metazoan circular mitochondrial genomes. I identified twelve dicyemid mitochondrial genes and manually verify the gene order with the support of 25 PacBio reads, although *atp8* was missing, similar to other marine creatures like most bivalves (Yu et al., 2008), brachiopod (Luo et al., 2015), and placozoan (Dellaporta et al., 2006). With inversion of the region from *cox3* to *atp6*, dicyemids possess very similar mitochondrial gene order to three other taxa, including freshwater snails (*Oncomelania*), ribbon worms, and terebratulide brachiopods, which have been reported to have exactly the same mitochondrial gene order (Luo et al., 2015). This suggests that this gene order might be conserved from the common ancestor of spiralians. Considering the size of putative mitochondrial genome of dicyemids, it is inconsistent with the reduction of their mitochondrial genome during the parasitism process as a general tendency for parasites (Gray et al., 1999; Saccone et al., 1999). However, based on the results of the present study, I still could not absolutely conclude that dicyemids have a canonical 15–20-kb circular mitochondrial genome or multiple mini-circular mitochondrial molecules similar to the cases in lice and parasitic nematode *Globodera* (Gibson et al., 2007; Herd et al, 2015).

For genome-sequencing, haploid gametes or sperm from one individual are generally considered as the better source for extracting genomic DNA, because high heterozygosity sometimes causes the difficulty for assembling. However, owing to the tiny individual size and limited cell number per individual, the dicyemid genomic DNA for preparing sequencing libraries was extracted from plenty of dicyemid individuals within one host octopus. This may lead to the high heterozygosity in the raw reads. The histogram plot of *k*-mer frequency distribution displayed this phenomenon, namely, there were lots of low frequency *k*-mers and no obvious k-mer frequency peak appeared before applying error-correction process. That is

probably because the *k*-mers from highly heterozygous region are usually with low frequency and easily mix up with those from sequencing errors. Despite of removing the low frequency *k*-mers, the GenomeScope estimated heterozygosity of 1.24%. This may be still underestimated since the error-correction process could discard the *k*-mers from highly heterozygous region. The high heterozygosity in raw reads raises a question whether the dicyemid individuals inside the same host are of one or multiple populations. The high heterozygosity could also explain why the size of assembly by Newbler assembler approach (Shinzato et al., 2011) continuously increased over 500 Mbp when incorporating more raw reads in the preliminary tests (Table 3.10). Newbler might recognize the variant sequences from the same highly heterozygous region as sequences from different regions and accumulated them repeatedly in the assembly. Therefore, the Platanus which declares the ability to reconstruct genomic sequences of highly heterozygous diploids was a suitable assembler for dicyemid genome.

Table 3.10 | The comparison between assemblies generated by Platanus and Newbler

Assembler	Newbler				Platanus			
Miseq data (runs)	1	2	3	4	1	2	3	4
Assembly length (Mbp)	111.7	197.3	443.2	522.5	19.6	32.3	40.6	42.8
Longest contig (bp)	22,988	32,307	32,408	41,493	19,041	34,958	41,376	31,345
N50 of contigs (bp)	955	962	992	969	1,720	2,096	2,050	1,968
Contig number	113,563	201,958	457,564	551,931	11,239	16,091	20,474	22,221

In addition to the high heterozygosity of genomic DNA sample, the unpreventable contamination from host cells was another concern for *de novo* assembly of dicyemid genome. At first, approximately 15% of the PacBio subreads were identified as the contamination sequences from the host octopus by mapping octopus short reads against the PacBio long reads of *D. japonicum*. Alternatively, only 2.4% of the scaffolds assembled by Platanus with *k*-mer coverage with cutoff of twenty-five were identified as contamination probably because the low-coverage octopus reads were not applied to the assembling process. In the present study,

combining these two computational strategies, back-mapping of contamination sequences and threshold cutoff of *k*-mer coverage, could eliminate the contaminating sequences efficiently. However, to obtain a high-quality genome assembly, the acquisition of a clean sample for genomic DNA extraction is the most essential procedure if it is possible. In contrast, fewer contamination sequences were found in the raw reads of transcriptome sequencing. Perhaps, because the RNA is comparatively unstable in dead cells, the contaminating octopus cells die and the octopus RNA degrade during the washing step of sample collection process, while the dicyemid individuals are still alive and contribute pure dicyemid RNA sample for sequencing.

For preliminary gene annotation, the reciprocal BLASTP searching against UniProt database is a common approach applying to predicted gene models. However, a few putative target genes could not be identified by this method; for example, four 7tm domain containing putative GPCR genes did not have bi-directional best hit in the reciprocal BLASTP search. Correspondingly, it has been reported that spiralians possess several atypical GPCRs with less similarity to both vertebrate and nematode GPCRs (Simakov et al., 2013). In addition, two of four Hox cluster genes found in this study could not be identified by reciprocal BLASTP search as well. The inefficiency of gene annotation by reciprocal BLASTP search may be due to that the similarity was insignificant between the presently studied sequences and UniProt contained reference sequences. Nevertheless, if studies adopt automatically annotated and not reviewed sequences, such as TrEMBL database from the UniProt, into the BLASTP database, the concern would be the accuracy of the reference sequences. In other words, the sequences outside the conserved regions of functional domains may be considerably divergent among organisms from a wide range of animal phyla, and the reviewed reference sequences in the UniProt database are adopted from relatively limited model organisms. These cause difficulty to cover whole sequence variations even within protein families. On the contrary, some putative genes identified by the reciprocal BLASTP search do not contain the characteristic

domain, which is conserved in such particular protein family. In the present case, seven of nine identified GPCR genes are without 7tm domain, which is believed as the functional and structural domain of the GPCR family. This suggests that other domains or sequences outside functional domain of some identified genes are relatively conserved and recognized as the sequences outside the characteristic domain of the target genes. Therefore, it is challenging to have both high efficiency and acceptable accuracy in one-step annotation. The reciprocal BLAST search against single database may be only suitable for preliminary annotation. Thus, consideration of evidences exclusively obtained from Pfam domain analysis and BLAST searches against different databases would be a better strategy for more precise annotation.

The dicyemid genome provides a resource to uncover the mysterious life cycles, as well as studying comparative genomics to gain insights into the parasitism evolution. *D. japonicum* possesses a compact genome, possibly accomplished by reducing intron size and gene number. Dicyemids retain least genes among compared species, reflecting their extremely simplified body organization. It also illustrates possible genomic convergences among spiralian (or bilaterian) parasites. Furthermore, the dicyemid-specific gene family expansions, which are particular adaptations to the lifestyle of dicyemids, may explain how dicyemids absorb the low molecular weight nutrients from the urine of their host. In general, parasites may adapt their genomes through reducing genes which are not necessary for parasitic lifestyle or increasing gene copies on lineage-specific biological processes.

4 Transcriptomic analyses: insights into the life cycle of dicyemids

4.1 Introduction

Dicyemids have a complicated life cycle and some portions of the life cycle have not been well studied yet, duo to practical difficulties. Nowadays, implement of new approach, such as transcriptome analysis, may provide additional interpretations for life cycle of dicyemids from a diverse aspect. A previous experiment showed that dicyemids could sense certain unknown soluble chemical which may trigger dicyemids shifting the phase from asexual to sexual reproduction (Lapan & Morowitz, 1972). Incubating asexually reproductive nematogens in the medium collected from a high-density culture for 24 hours, the axoblast inside the nematogen differentiated into the infusorigen, which is the hermaphroditic gonad in sexual reproductive rhombogen. This shift is an important step for dicyemids because the infusoriform larva from sexual reproductive rhombogen plays a crucial rule in infecting new hosts, sustaining the survival of dicyemid species. To date, the molecular mechanisms of how dicyemids that lack differentiated nervous system sense and respond to the chemical cues remains undiscovered. Moreover, how does a new infection establish? In other words, a question how infusoriform larvae search for a new host in the open water to close their life cycle still need to be investigated. Sensory function is essential for most animals which rely on their sensory organs to interact with the surrounding environments by receiving chemical or mechanical signals. A took-kit gene, *Pax6* has been reported in dicyemids. Since *Pax6* plays key roles in development of sensory organs, indicating that this tool-kit gene is still required and may perform conserved function even for highly simplified spiralians (Aruga et al., 2007). However, no nervous system or sensory cells have yet been reported in dicyemids, and it is believed that the sensory organs have secondarily reduced along with adapting to parasitic lifestyle. Therefore, the potential sensory function involved in closing the life cycle of dicyemids is worthy to investigate.

Here, I applied transcriptome analysis at four different life cycle stages. That could provide insights into the biological meaning and functions of each life cycle stage at molecular level. In addition, the immunostaining was performed to investigate the presences of neurotransmitter- and neuropeptide-like molecules, which may be involved in the signal processing for chemical cues. Such approach could be beneficial to further comprehend the mechanisms of potential sensory functions.

4.2 Materials and Methods

4.2.1 Sample collection

For sampling separated stages and for obtaining enough larva individuals, I combined samples from seven octopus individuals, and each dicyemid individual was manually picked up and identified into four stages (nematogen, asexual-reproductive adult; vermiform larva from asexual-reproduction; rhombogen, sexual-reproductive adult; infusoriform larva from sexual-reproduction) (Figure 1.2). Although the outward appearances of two adult stages look the same, they could be separated by careful inspection of the morphology of developing larva inside the adult individuals. In order to obtain enough amount of RNA, I collected approximately six to eight hundred individuals for each stage. However, I did not count the exact number of individuals used for each stage. RNA was extracted using Direct-zol RNA MicroPrep Kit (Zymo Research, #R2060). Instead of usage of the same number of individuals, I used the same amount of RNA (1.3 ng) for each stage to prepare the sequencing libraries. After reverse transcribe RNA to cDNA by SMART-Seq v4 Ultra Low Input RNA Kit (Clontech Laboratories, #634888), Nextera XT DNA Library Preparation Kit (Illumina, #FC-131-1024) was utilized for library preparation. Sequencing was performed on an Illumina HiSeq 4000.

4.2.2 Differential expression and gene ontology analyses on four stages

I pooled libraries of four life-cycle stages and performed sequencing on one HiSeq 4000 run to avoid the technical bias (Table 4.1). Moreover, in order to normalize the biological bias between individuals, I combined dicyemids collected from seven octopuses and picked up hundreds of individuals for each life-cycle stage to extract RNA. Transcripts of a mixed-stage transcriptome assembly were used as references, and quality-trimmed RNA-seq reads from each stage were applied to assess transcript abundance by kallisto (Bray et al., 2016), an alignment-free estimation method. The kallisto results were reported in estimated counts and TPM (Transcripts Per Kilobase Million) measures. Further, kallisto results of all stages were normalized to TMM (trimmed mean of M-values) measures cross samples. Differentially expressed transcripts were extracted and partitioned into clusters according to the expression pattern of four life-cycle stages using perl scripts in the Trinity package, analyze_diff_expr.pl and define_clusters_by_cutting_tree.pl, respectively. Gene ontology (GO) over-representation analyses were conducted using DAVID (Huang et al., 2009) and PANTHER (Mi et al., 2012).

Table 4.1 | Summary of RNA-seq library preparation methods and read numbers

Sample	Library preparation system	Raw read pairs	Both surviving (Q20)	Survival rate
Mixed stage	NEBNext Ultra Directional RNA Library Prep Kit.	59,459,075	38,579,433	64.9%
Nematogen	SMART-seq v4 Ultra Low Input RNA Kit	40,910,646	16,234,622	39.7%
Rhombogen	SMART-seq v4 Ultra Low Input RNA Kit	46,618,665	24,562,055	52.69%
Vermiform larva	SMART-seq v4 Ultra Low Input RNA Kit	44,080,942	21,368,427	48.5%
Infusoriform larva	SMART-seq v4 Ultra Low Input RNA Kit	42,611,014	18,638,608	43.7%

4.2.3 Immunostaining and imaging

Specimens for immunostaining were fixed in 4% paraformaldehyde (PFA) solution for 30 min, and then stored in 75% ethanol at -20°C. The immunostaining protocol was modified from a previous study (Lu et al., 2017). Samples were incubated for 1 hr in blocking solution (3% BSA and 0.1% Triton X-100 in PBS) and incubated in the primary antibody solution (Table 4.2), and acetylated tubulin mouse monoclonal antibody (Sigma, #T6793, 1:1000 diluted in blocking solution) at 4°C overnight. Fluorescent signals were detected after incubation in a secondary antibody solution of Alexa Fluor 594-conjugated, goat anti-mouse antibody. DAPI (Invitrogen, 1 μg/mL in PBST) was used for nuclear staining, and plasma membrane were stained with CellMask (Life technology, C10046). Fluorescent images were acquired using a Zeiss 780 confocal microscope with 20X and 100X objectives.

Table 4.2 | Antibodies of tubulin, neurotransmitters and neuropeptides

Antibody name	Host	Dilution	Catalog number
Acetylated-tubulin	Mouse	1:1000	Sigma T7451
Oxytocin	Rabbit	1:4000	Immunostar 20068
Vasopressin	Rabbit	1:2000	Immunostar 20069
FMRF-amide	Rabbit	1:1000	Abcam ab10352
FMRF-amide	Rabbit	1:1000	Immunostar 20091
Dopamine	Rabbit	1:1000	Abcam ab8888
Dopamine-beta-hydroxylase	Rabbit	1:2000	Immunostar 22806
Gamma-aminobutyric acid	Rabbit	1:1000	Sigma A2052
Gamma-aminobutyric acid	Rabbit	1:15000	Immunostar 20094

4.3 Results

4.3.1 Differential expression analysis and over-representation analysis

The quality-trimmed reads of four stages were separately aligned against 20,556 genes of the stages-mixed reference transcriptome assembly to estimate the abundance of each gene. Under the criteria of 4-fold changes and 1e-3 *p*-value cutoff for the false discovery rate, I identified

1,641 differentially expressed genes. They could be partitioned into eight gene clusters according to expression pattern (Figure 4.1). Most differentially expressed genes (81.6%) had higher expression in the infusoriform larvae (Figure 4.2a). The stage correlation matrix heatmap was congruous with the dicyemid life cycle, in which the vermiform larvae were more correlative to the nematogen than the rhombogen, and the infusoriform larvae was least correlative to all others (Figure 4.2b). GO over-representation analyses were performed for eight gene clusters. In gene cluster 3 in which infusoriform larvae displayed lower gene expressions than other three stages, GO terms involved in transmembrane transporter activity were over-represented, while I could not find any over-represented GO term for gene clusters 4, 6, and 8. For gene clusters with highest expressions in infusoriform larvae (gene cluster 1, 2, 5, and 7), over-represented GO terms were mainly associated with few biological processes, development, signal transduction, transmembrane transport, and sensory function and response (Figure 4.3).

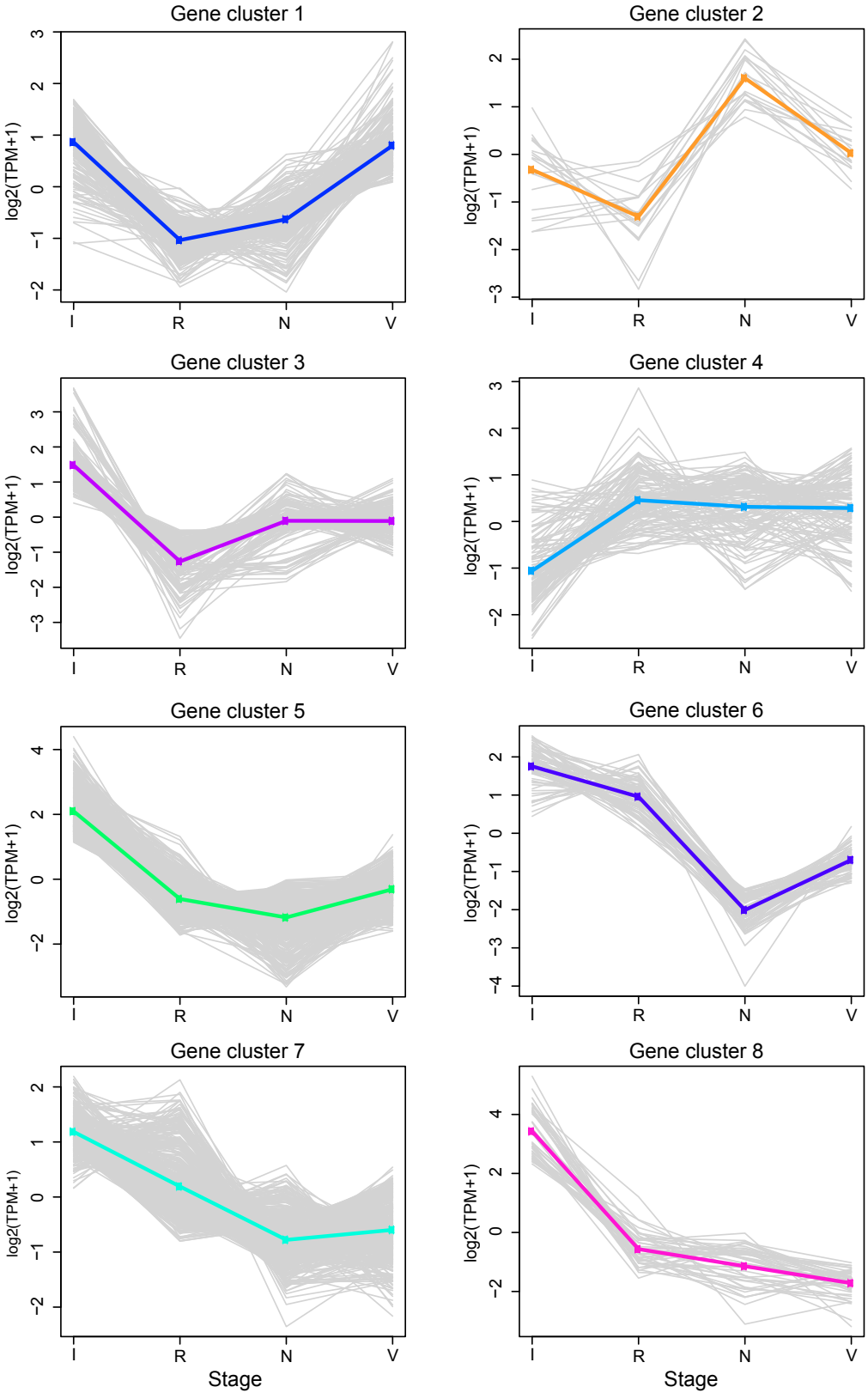

Figure 4.1| Eight gene clusters of differentially expressed genes.

According to the gene expression profile at each life-cycle stage, 1,641 differentially expressed genes were categorized into eight gene clusters. I, infusoriform larva; R, rhombogen; N, nematogen; V, vermiform larva.

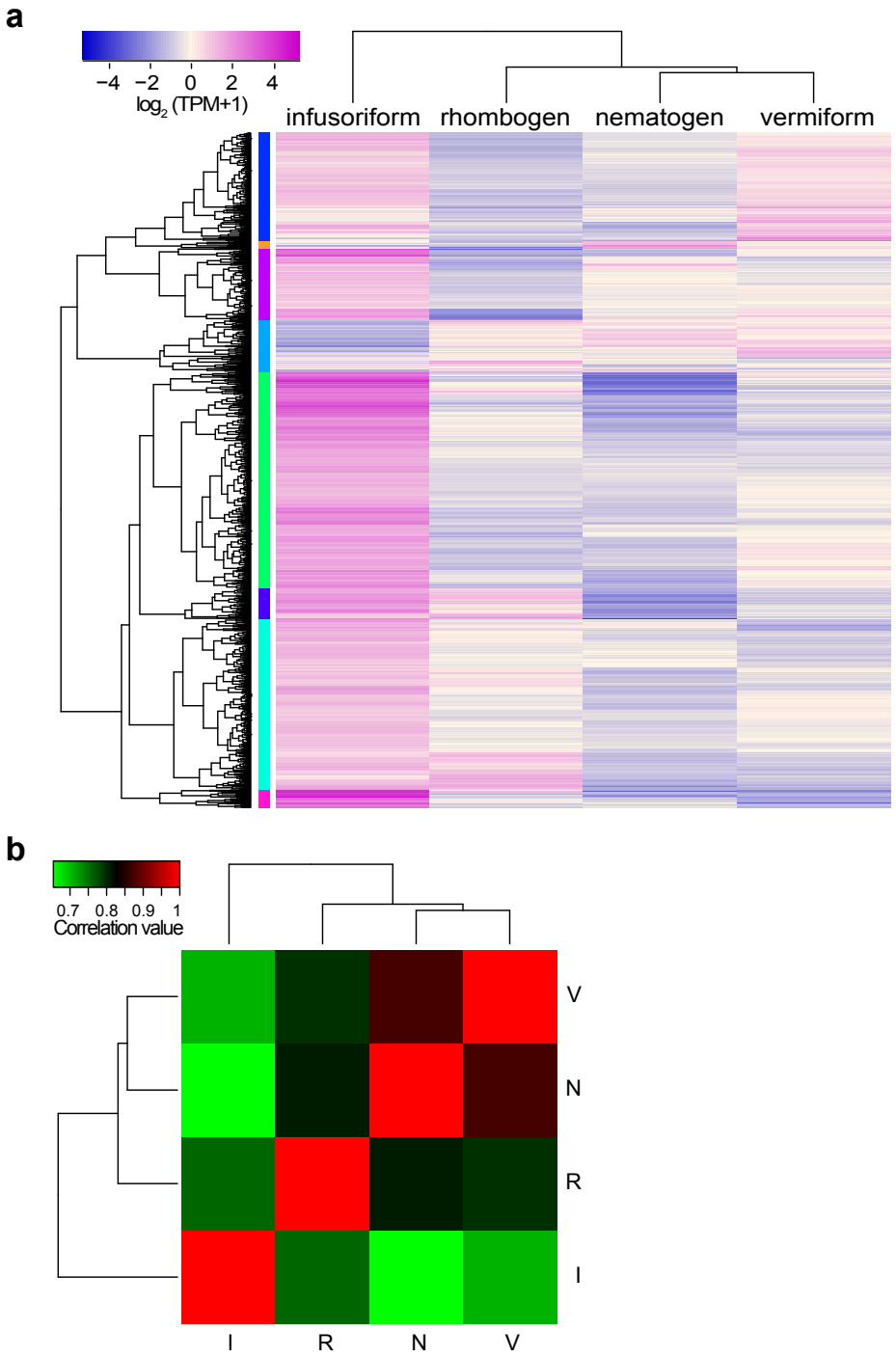

Figure 4.2 | Differentially expressed genes show correlations with life-cycle stages.

(a) Expression heatmap of differentially expressed genes. More than 80% of differentially expressed genes displayed higher expressions in the infusoriform larvae. Differentially expressed genes were grouped into eight gene clusters as indicated by color bars, also corresponding to the gene clusters in Figure 4.1. (b) Spearman's correlation relationship between four life-cycle stages of dicyemids. The gene expression profile of three stages inhabiting the renal sac exhibit higher correlation than that of dispersal infusoriform larva, reflecting that they utilize different genes to conduct divergent biological functions at different environments. I, infusoriform larva; R, rhombogen; N, nematogen; V, vermiform larva.

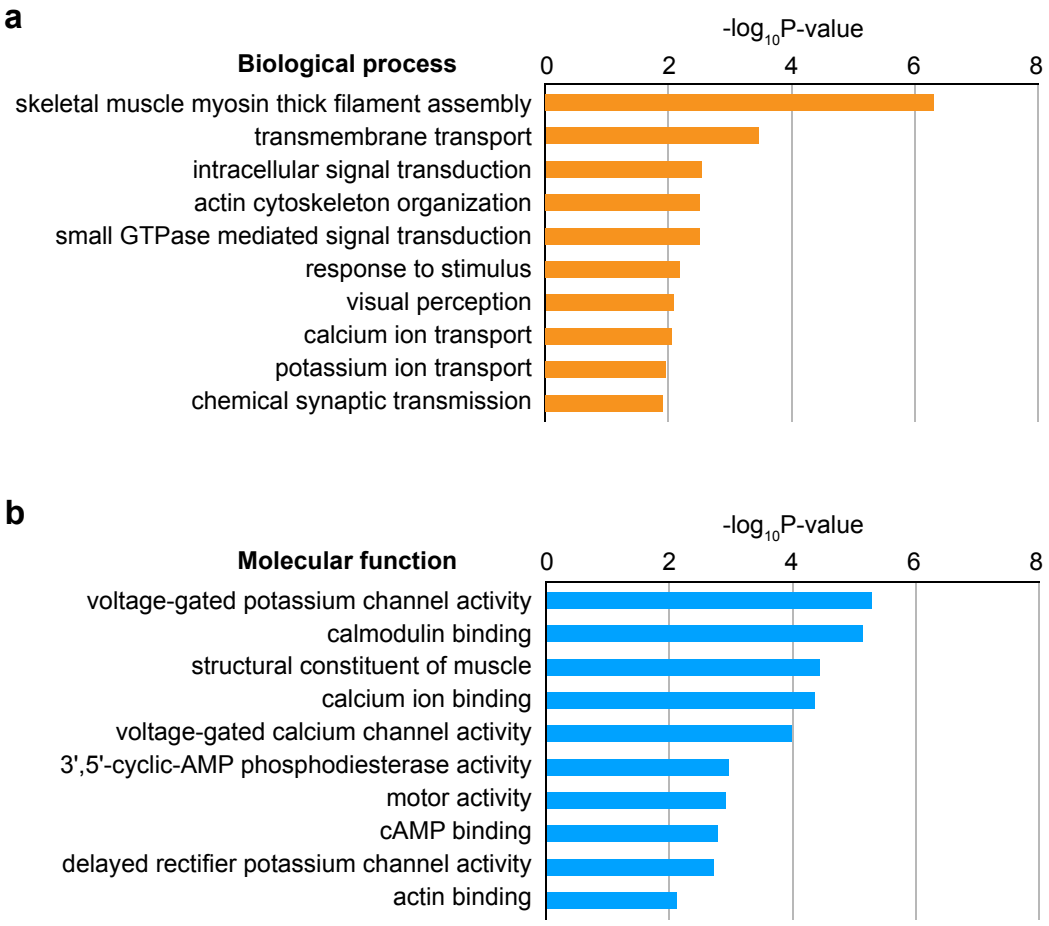

Figure 4.3 | Overrepresented GO terms on biological process and molecular function. The GO terms such as response to stimulus, chemical synaptic transmission, and ion channel activity suggest that highly expressed genes in infusoriform larva are probably associated with potential sensory function. On the other hand, skeletal muscle myosin thick filament assembly, and motor activity may link to the cilia beating of the infusoriform larva that contributes to the swimming ability.

4.3.2 Immunostaining of neuropeptides and neurotransmitters

To clarify if any neurotransmitter- and peptide hormone-like molecules are involved in certain biological function of dicyemids, especially in infusoriform larvae, I examined their protein expression by fluorescent immunostaining method. For FMRFamide, I could not detect any signal of expression at both adult and larvae stages. The immunostaining signals of two peptide hormones, vasopressin and oxytocin, and two neurotransmitters, dopamine and gamma-aminobutyric acid (GABA), specifically occurred on apical cells, while dopamine beta

hydroxylase (DBH) was expressed on both apical cells and germinal cells of infusoriform larvae (Figure 4.4). The cilia were also detected by acetylated alpha-tubulin antibody. The morphology of cilia was different among the collate and trunk regions of adult individual, the apical cells and other external cells of infusoriform larvae (Figure 4.5a-b). Notably, the bristles, a small patch of shorter cilia dorsal to the nucleus of apical cells which was described on *D. aegira* (Short et al., 1966), were also observed on the specimen studied herein (Figure 4.5c).

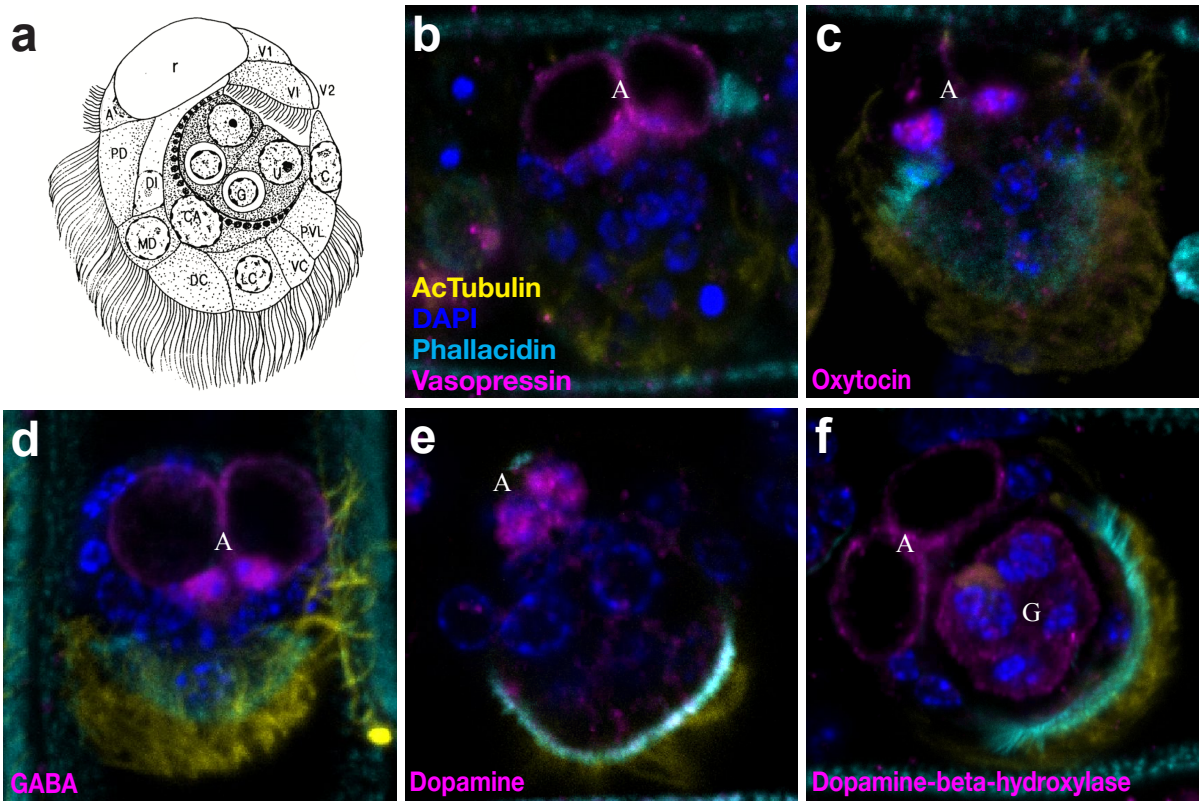

Figure 4.4 | Neurotransmitters and neuropeptides are expressed in the apical cells.
(a) Sagittal optical section of infusoriform larva. Adapted and modified from Furuya et al. (1992). The immunostaining signals of vasopressin (b), oxytocin (c), GABA (d), and dopamine (e) appear in the apical cells, while the signals of dopamine beta hydroxylase (DBH) exist in both apical cells and germinal cells (f). Apical cells seem to be a signal transduction center, playing a role in receiving the signals from the surroundings and coordinating the release of germinal cells. Nuclei were labeled with DAPI (blue), and acetylated tubulin antibodies labeled cilia (yellow). A, apical cell; G, germinal cell.

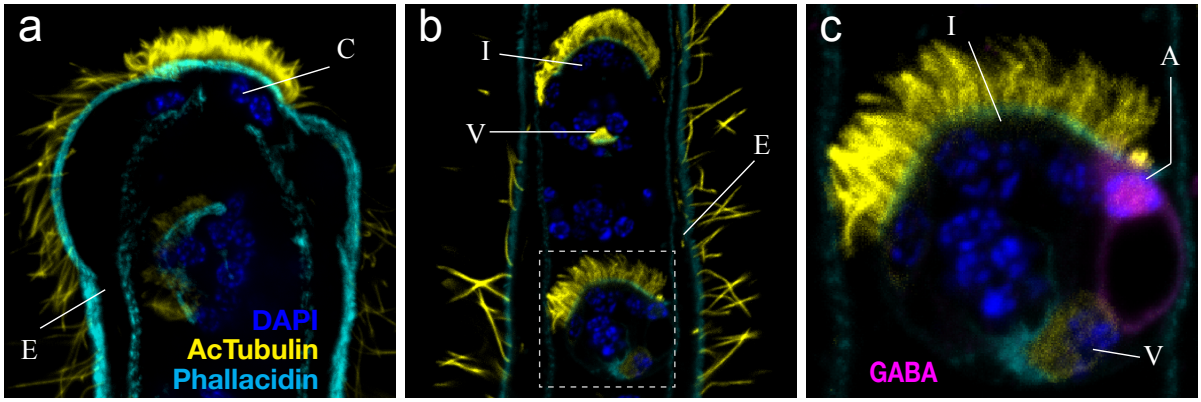

Figure 4.5 | Various cilium types in dicyemids.

(a) In the anterior end of rhombogen, cilia on the calotte cells are shorter and with higher density than those on the epithelial cells. (b) The density of cilia on the epithelial cells of infusoriform larva is higher than those on the epithelial cells of adults, reflecting that infusoriform larva have higher mobility than adults. (c) The infusoriform larvae, enlargement of the dashed square of (b). The length of bristles on apical cells is approximately one-third of the cilia length on the epithelial cells. A, apical cell; C, calotte cell; E, epithelial cell of adult; I, epithelial cell of infusoriform larvae; V, ventral internal cell of infusoriform larvae.

4.4 Discussion

Limited by the difficulties of tracing and observation owing to their microscopic individual size and living environment, the complicated life cycle of dicyemids remains incompletely known. The transcriptome data of four life-cycle stages would provide the information of what biological processes are carried out at each stage. This may help us to elucidate the gap in knowledge of life-cycle stages that how dicyemids search for and infect new hosts to close their life cycle. The up-regulated profile of differentially expressed genes on infusoriform larvae indicates the specialized functions for the dispersal larva, which escapes from the original hosts to search for new hosts in the open ocean. This contrasts with the other three stages keeping staying in the original hosts. Moreover, the over-represented GO terms such as visual perception, response to stimulus, and chemical synaptic transmission, bring up an intuition that infusoriform larvae utilize potential sensory function to detect new hosts and neighboring environments, although no sensory neurons or nervous system have been reported in the

simplified body organization of dicyemids. In addition, the over-represented GO terms such as signal transduction, ion channel activity, transmembrane transport and motor activity, suggest that infusoriform larvae could be able to process the received signals and turn on the downstream responses, like tracing the source of the signals.

Sensory receptor molecules contribute to receiving and transducing the sensory signals, that is the first step of perceiving an external stimulus. Unexpectedly, only six putative GPCRs are annotated in dicyemids as discussed in the previous chapter. It suggests that the retained sensory functions in dicyemids narrow down to detecting limited targets such as the hosts, due to the reduction of GPCR gene family. The infection to a new host could be the most crucial step in the entire life cycle of dicyemids to sustain the survival of this taxon, while the sensory functions for other than searching and infecting new hosts would be less essential to be lost during the evolution process.

The immunostaining of dopamine, GABA, DBH, and oxytocin- and vasopressin-like (OT/VP) peptide displayed expression signals in the apical cells, suggesting that apical cell may play an important role in signal processing for the surrounding stimuli, although their regulation and interaction within dicyemid individuals are still unclearly known. Dopamine presumably could excite the infusoriform larva to keep active momentum on swimming, which is essential for escaping the parent individual and looking for new hosts. GABA has been shown as an inhibitory neurotransmitter and its inhibition of dopaminergic activity (Garbutt & van Kammen, 1983), and it also affects the swimming behavior in *Paramecium* which do not have nervous system as well (Valentine et al., 2008). In the present data, GABA coexisting with dopamine in apical cells may carry out an antagonistic action to the effect of dopamine. DBH could modulate the amount of dopamine by catalyzing the chemical reaction of converting dopamine to norepinephrine. In addition to apical cells, DBH exhibits additional expressions in urn cells,

suggesting that DBH may generate norepinephrine for downstream regulation when urn cells receive the secreted dopamine from the neighboring apical cells.

Oxytocin and vasopressin are found in wide ranges of animals. However, invertebrates, except cephalopods, have only one oxytocin/vasopressin superfamily peptide homolog (Fieber, 2017), which plays roles in reproduction, learning, and osmoregulation in general. Members of this peptide family in molluscs, annelids, and nematodes share high sequence similarity (Gruber et al., 2014). Thus, the immunostaining signals of vasopressin and oxytocin antibodies both appeared on the apical cells (Figure 4.4b, c) should recognize the same OT/VP peptide in dicyemids. As in other animals, the OT/VP peptide may be associated with the reproduction-like action in infusoriform larva (if considering the release of the germinal cells similar to childbirth), and mediate the release of dopamine, as reported in vertebrates. Undoubtedly, further examinations on these assumed functions would be constructive to understand the behavior and function of infusoriform larva.

In this study, all antibodies against neuropeptides and neurotransmitters are commercially-purchased polyclonal antibodies, and they are produced against vertebrate antigens. The immunostaining results should be interpreted with caution when I use these antibodies in nonmammalian species due to its cross-reactivity to similar peptides. Even, the usages of dopamine and GABA antibodies used in this study have been published in nonmammalian species, e.g. GABA antibody shows the neuronal interactions in the insect antennal lobe (Husch et al., 2009), and dopamine antibody recognizes an insect-like mushroom body in a crustacean brain (Wolff et al., 2017). Although neuropeptides and neurotransmitters are often conserved throughout metazoans, the specificities of these antibodies in dicyemids need to be further validated. Despite the immunostaining signals shown in this study all appeared in the apical cells of dispersal larvae, I could not arbitrarily regard that these antibodies exactly recognized the target neuropeptides or neurotransmitters. In future, the

antibodies generated from dicyemid antigens would provide more reliable evidence for understanding the signal transduction of sensory function in dicyemids.

Dicyemids display variant cilia morphologies, suggesting that these cilia may play different roles. On the infusoriform larva, the external somatic cells bear long cilia, which are utilized for free moving in the ocean. Cilia on the ventral internal cells possibly circulate the fluid within the urn cavity and exchange the seawater providing nutrients and oxygen for both urn cells and germinal cells (Furuya et al., 2004). However, the actual function of bristles on apical cells is problematical. The length of bristles is approximately one third to that of the longer cilia on the external cells. Some short cilia labelled by the acetylated-tubulin antibody are seen on the apical cells and dorsal to the nucleus. This structure was first described by Short (1966) on *Dicyema aegira* that each apical cell bears a small patch of "bristles" (short cilia). The bristles on apical cells are obviously shorter than that on other epithelial cells (Figure 4.5c). It gives me an impression that these short cilia are not used for mobility, because there are too short. Taking the expressions of neurotransmitters and neuropeptides into account, it is an intriguing question whether it may be just a coincidence, or the apical cells may play a role in certain regulatory or sensory functions. During swimming, the weight of the refringent bodies may cause the infusoriform larvae to be oriented with the anterior end downward (Stunkard et al., 1954). Thus, the bristle side of infusoriform larvae may directly face the seabed and the open of the urn cavity is upward to the open water. Previous studies have proposed that abundant eosinophilic granules in the capsule cells may perform lytic function involved in the release of germinal cells containing urn cells (Ridley et al., 1969). However, the upstream signals that activate to lytic function in the capsule cells remain to be discovered. It could be possible that bristles may accomplish a function for mechanical or chemical senses on the seabed, which could presumably coordinate with the activation of lytic function for release of germinal cells.

The GO enrichment analyses and immunostaining results are summarized to indicate that the dispersal infusoriform larva may possess potential sensory function to receive chemical or mechanical cues. Even lacking the sensory cells and nervous system, certain dicyemid cells may retain sensory function, based on the fact that genes and receptors with conserved functions exist there. That may further lead to the secretion of peptide hormones and signaling transmitters in the apical cells. When new host is detected, these signals activate the lytic function of the capsule cells through intercellular signal transductions to release germinal cell containing urn cells. Coordinating with the mobility by long cilia, dicyemids probably could actively seek and approach new hosts, and then release the germinal cells under the siphon of the octopus for infection to close their life cycle. In the present study, I propose a hypothesis of how dicyemids look for new hosts and trigger the release of germinal cells for the unknown portion of dicyemids life cycle, yet how germinal cells migrate into the renal sac of octopus and develop into a nematogen adult still remain to be discovered.

Currently, since the closed and life-cycle-long cultivation system of octopus and dicyemids have not been well established, it hampers the infection experiment which probably could demonstrate how dicyemid dispersal larvae infect a new host. However, the potential *in vitro* experiments could be conducted in near future to test the proposed hypothesis. To test whether dicyemid dispersal larvae could be attracted by the octopus, the treatment of octopus extract could be applied to the experiment. We could observe whether dicyemid dispersal larvae gather to the source of the octopus extract when adding the concentrated octopus urine, octopus-cultured seawater, or extract from blended octopus tissues to the center of a 15 cm petri dish where dispersal larva are kept. The treatment experiments could also apply to test whether peptide hormones and signaling transmitters play a role in the release of germinal cells. We could record the responses of dicyemid dispersal larva when candidate peptide hormones or signaling transmitters added to the culture medium. These experiments could preliminarily test

the proposed hypothesis, once the method for efficiently collecting dicyemid dispersal larvae has been established.

5 Conclusions

5.1 Dicyemids are simplified spiralians and possess close affinity to the Rouphozoa

By incorporating new dicyemid transcriptomic data, this is the first dicyemid phylogenomic study to reflect reliability of evolutionary relationships. I excluded possible systematic biases and performed phylogenomic analyses. The present results clarify the phylogenetic position of dicyemids and present its new phylogenetic profile within the Spiralia. Analyses using the complete dataset and datasets with systematic biases removed, show concordant results that dicyemids share a common ancestor with orthonectids endorsing the old Mesozoa clade. Furthermore, mesozoans have a closer affinity to gastrotrichs and platyhelminthes, rather than mollusks and annelids. The long branch lengths of mesozoans likely reflect the nature of the rapid evolution of parasites, although I could not totally exclude the possible effects from compositional heterogeneity and long-branch attraction artifacts. Regarding spiralian phylogeny, my results support that the Spiralia comprises three monophyly clades, Gnathifera, Rouphozoa, and Lophotrochozoa. The small, non-coelomate Gnathifera branched off first and comprise a sister group to the Rouphozoa and Lophotrochozoa. This supports the previously proposed acoeloid-planuloid hypothesis of a nearly microscopic noncoelomate common ancestor of the spiralians.

5.2 Properties of *Dicyema* genome

In this study, I present the genome and predicted gene models of the dicyemid, *Dicyema japonicum*, for the first time. Investigating the genome and transcriptome gives us insights into their mysterious life cycle and genomic adaptations to the parasitic lifestyle. Comparing with other non-parasitic spiralians, the dicyemid genome is relatively compact and contains only approximately 5,000 predicted genes with extraordinarily shortened introns. In addition, the gene number reduction of whole gene set would be another cause leading to a small genome

size. Comparative study on KEGG pathways showed that *D. japonicum* retains less genes on most pathways than other bilaterians, instead of entirely eliminating functional pathways. It was also found that numbers of functional domains in *D. japonicum* are less than non-parasitic spiralians. The absence of important transcriptional domains for development processes, such as neuronal CHRD, FGF and DSL domains, may correspond to the loss of tissue specification and body organization. In contrast to the gene loss, dicyemids possibly encountered gene expansions on endocytosis-associated genes, which may explain how dicyemids acquire low molecular weight nutrients from the urine of the host, although they do not have month and digestion system.

5.3 Essential sensory function to close the life cycle

After releasing into the open water, how does 37-cell dispersed larva search for and infect a new host to close their life cycle? It is still a mystery. Although no nervous system has been reported in dicyemids, the sensory function and mobility associated genes are highly expressed in the dispersal larva. It indicates that these capabilities would be crucial for the survival of dicyemids, providing more efficiency to infect new hosts. However, only six putative GPCR genes are found in dicyemids, and they may play the roles in the remained sensory functions. Taken together with the immunostaining signals of neurotransmitters and neuropeptides that specifically occurred on apical cells of dispersal larva, I propose that dicyemid dispersal larva may have potential sensory ability to detect new hosts and actively approach them. Moreover, the apical cells may be responsible for processing the signals and coordinating the release of germinal cells near or inside the host individuals.

5.4 Molecular convergences of bilaterian parasites

Within spiralian species, the gene numbers of parasite species are less than half of the average gene number of non-parasite species. This indicates that the reduction of gene number could be a convergence among spiralian parasite species. More specifically, four spiralian and two ecdysozoan parasite species possess less gene number in KEGG metabolism pathways than those with compared non-parasite species. This also suggests that the reduction of metabolism associated genes could be a convergence among bilaterian parasites. Similar to tapeworm and orthonectid, only three Hox genes are identified in the *Dicyema* genome and they are scattered on different scaffolds. Therefore, the loss of partial Hox genes and Hox-cluster structure could be another convergence among spiralian parasite species, corresponding to their highly simplified body organization.

5.5 Concluding remarks

Decoding the *Dicyema* genome and transcriptome provides great resources to investigate the evolution of dicyemids and parasitism. It also allows us to speculate the mysterious life cycle of this microscopic endoparasite which was the technical barrier to observation. Here, I obtained high-quality reference genome and transcriptome assembly of *D. japonicum*, providing genome-wide datasets for future studies. Refer to genome-wide evidences, I clarified the phylogenetic position of dicyemids and the internal relationships with other microscopic lineages among the Spiralia. I obtained the genome structure information such as of intron, exon, and repetitive regions, and performed comprehensive gene model prediction and annotation. Further, the present study has advanced the understanding of genomic adaptions of parasitism (or symbiosis). I demonstrated the gene number reduction is a molecular convergence of spiralian parasites that would echo the simplification of morphological traits and the reduction of physiological functions in parasites. In contrast, I illustrated the dicyemid-

specific gene family expansion on endocytosis underlying the nutrient-uptake strategy of dicyemid. It suggests that each parasitic (or symbiotic) relationship evolved independently corresponding to its particular lifestyle and habitat environment. Overall, parasites may modify their genomes through reducing genes which are not necessary for parasitic lifestyle or increasing gene copies on lineage-specific biological processes. Finally, the differential gene expressions among different life-cycle stages and investigating possible sensory functions provide hints to uncover the mysterious life cycle of dicyemids and the biological functions of cells in the dispersal larva. Thus, I proposed a hypothesis that using potential sensory function, dicyemid dispersal larvae could actively detect new hosts and approach them for infection to close their life cycle. This study not only improves the knowledge of dicyemid biology but also provides sequence information for future genetic studies.

6 Reference list

Adema, C.M., Hillier, L.W., Jones, C.S., Loker, E.S., Knight, M., Minx, P., *et al*. (2017). Whole genome analysis of a schistosomiasis-transmitting freshwater snail. Nature Communications. *8*, 15451.

Albertin, C.B., Simakov, O., Mitros, T., Wang, Z.Y., Pungor, J.R., Edsinger-Gonzales, E., *et al*. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature. *524*, 220-224.

Aruga, J., Odaka, Y.S., Kamiya, A., and Furuya, H. (2007). Dicyema *Pax6* and *Zic*: tool-kit genes in a highly simplified bilaterian. BMC Evol Biol. *7*, 201.

Awata, H., Noto, T., and Endoh, H. (2005). Differentiation of somatic mitochondria and the structural changes in mtDNA during development of the dicyemid *Dicyema japonicum* (Mesozoa). Mol Genet Genomics. *273*, 441-449.

Awata, H., Noto, T., and Endoh, H. (2007). Peculiar behavior of distinct chromosomal DNA elements during and after development in the dicyemid mesozoan *Dicyema japonicum*. Chromosome Res. *14*, 817-830.

Ax, P. (1996). Multicellular animals: a new approach to the phylogenetic order in nature. Springer Verlag, Berlin.

Bardou, P., Mariette, J., Escudié, F., Djemiel, C., and Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. BMC Bioinformatics. *15*, 293.

Baughman, K.W., McDougall, C., Cummins, S.F., Hall, M., Degnan, B.M., Satoh, N., *et al*. (2014). Genomic organization of Hox and ParaHox clusters in the echinoderm, *Acanthaster planci*. Genesis. *52*, 952-958.

Berriman, M., Haas, B.J., LoVerde, P.T., Wilson, R.A., Dillon, G.P., Cerqueira, G.C., *et al*. (2009). The genome of the blood fluke *Schistosoma mansoni*. Nature. *460*, 352-358.

Blaxter, M., and Koutsovoulos, G. (2014). The evolution of parasitism in Nematoda. Parasitology. *142*, S26-S39.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. *27*, 578-579.

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics. *15*, 211.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. *30*, 2114-2120.

Boutet, E. Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A.J., *et al*. (2016). UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Methods in Molecular Biology, *1374*, 23-54.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., *et al*. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience. *2*, 10.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. *34*, 525-527.

Bresciani, J., and Fenchel, T. (1967). Studies on dicyemid Mesozoa II. The fine structure of the infusoriform larva. Ophelia.

Brusca, R.C., and Brusca, G.J. (1990). Invertebrates. Sinauer, Sunderland, Massachusetts.

Cameron, S.L., Yoshizawa, K., Mizukoshi, A., Whiting, M.F., and Johnson, K.P. (2011). Mitochondrial genome deletions and minicircles are common in lice (Insecta: Phthiraptera). BMC Genomics. *12*, 394.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. *25*, 1972-1973.

Cardoso, J.C.R., Félix, R.C., and Power, D.M. (2014). Nematode and arthropod genomes provide new insights into the evolution of class 2 B1 GPCRs. PLoS ONE. *9*, e92220.

Cavalier-Smith, T. (1998). A revised six-kingdom system of life. Biol Rev Camb Philos Soc. *73*, 203-266.

Cavalier-Smith, T. (2005). Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Annals of Botany. *95*, 147-175.

Catalano, S.R. (2013). Five new species of dicyemid mesozoans (Dicyemida: Dicyemidae) from two Australian cuttlefish species, with comments on dicyemid fauna composition. Syst Parasitol. *86*, 125-151.

Catalano, S.R., Whittington, I.D., Donnellan, S.C., and Gillanders, B.M. (2014). Dicyemid fauna composition and infection patterns in relation to cephalopod host biology and ecology. Folia Parasitologica. *61*, 301-310.

Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., *et al*. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat Meth. *13*, 1050-1054.

Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. *10*, 210.

Cummins, C.A., and McInerney, J.O. (2011). A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Systematic Biology. *60*, 833-844.

Czaker, R. (2006). Serotonin immunoreactivity in a highly enigmatic metazoan phylum, the pre-nervous Dicyemida. Cell Tissue Res. *326*, 843-850.

Dellaporta, S.L., Xu, A., Sagasser, S., Jakob, W., Moreno, M.A., Buss, L.W., *et al*. (2006). Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal lower metazoan phylum. Proc. Natl. Acad. Sci. U.S.A. *103*, 8751-8756.

Dodson, E.O. (1956). A note on the systematic position of the Mesozoa. Systematic Biology. *5*, 37-40.

Ebersberger, I., Strauss, S., and Haeseler, von A. (2009). HaMStR: Profile hidden markov model based search for orthologs in ESTs. BMC Evol Biol. *9*, 157.

Edgecombe G.D., Giribet G., Dunn C.W., Hejnol A., Kristensen R.M., Neves R.C., et al. (2011). Higher-level metazoan relationships: recent progress and remaining questions. Org Divers Evol. *11*, 151-172.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research. *30*, 1575-1584.

Fieber, L.A. (2017). Neurotransmitters and Neuropeptides of Invertebrates. Oxford Handbooks.

Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., *et al*. (2011). Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics. *6*, 12-19.

Freire-Garabal, M., Núñez, M.J., Balboa, J., López-Delgado, P., Gallego, R., García-Caballero, T., *et al*. (2003). Serotonin upregulates the activity of phagocytosis through 5-HT1A receptors. British Journal of Pharmacology. *139*, 457-463.

Fröbius, A.C., and Funch, P. (2017). Rotiferan Hox genes give new insights into the evolution of metazoan bodyplans. Nature Communications. *8*, 9.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. *28*, 3150-3152.

Furuya, H., Tsuneki, K., and Koshida, Y. (1992). Development of the infusoriform embryo of *Dicyema japonicum* (Mesozoa: Dicyemidae). Biol. Bull. MBL; *183*, 248-257.

Furuya, H., Tsuneki, K., and Koshida, Y. (1997). Fine structure of dicyemid mesozoans, with special reference to cell junctions. J. Morphol. *231*, 297-305.

Furuya, H., Hochberg, F.G., and Tsuneki, K. (2001). Developmental patterns and cell lineages of vermiform embryos in dicyemid mesozoans. Biol. Bull. *201*, 405-416.

Furuya, H., and Tsuneki, K. (2003). Biology of Dicyemid Mesozoans. Zool. Sci. *20*, 519-532.

Furuya, H., Hochberg, F.G., and Tsuneki, K. (2003). Calotte morphology in the phylum Dicyemida: niche separation and convergence. J Zoology. *259*, 361-373.

Furuya, H., Hochberg, F.G., and Tsuneki, K. (2004). Cell number and cellular composition in infusoriform larvae of dicyemid mesozoans (Phylum Dicyemida). Zool. Sci. *21*, 877-889.

Furuya, H., and Tsuneki, K. (2007). Developmental patterns of the hermaphroditic gonad in dicyemid mesozoans (Phylum Dicyemida). Invertebrate Biology. *126*, 295-306.

Garbutt, J.C. and van Kammen, D.P. (1983). The interaction between GABA and dopamine: implications for schizophrenia. Schizophrenia Bull, *9*, 336-353.

Gibson, T., Blok, V.C., and Dowton, M. (2007). Sequence and characterization of six mitochondrial subgenomes from *Globodera rostochiensis*: multipartite structure is conserved among close nematode relatives. J Mol Evol. *65*, 308-315.

Giribet G. (2008). Assembling the lophotrochozoan (=spiralian) tree of life. Philos Trans R Soc Lond B. *363*, 1513-1522.

Giribet, G. (2015). New animal phylogeny: future challenges for animal phylogeny in the age of phylogenomics. Org Divers Evol. *16*, 419-426.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., *et al*. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. *29*, 644-652.

Gray, M.W., Burger, G., and Lang, B.F. (1999). Mitochondrial evolution. Science. *283*, 1476-1481.

Gruber, C.W. (2014). Physiology of invertebrate oxytocin and vasopressin neuropeptides. Exp. Physiol. *99*, 55-61.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology. *59*, 307-321.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., *et al*. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. *8*, 1494-1512.

Halanych K.M. (2004). The new view of animal phylogeny. Annu Rev Ecol Evol Syst. *35*, 229-256.

Hejnol A., Obst M., Stamatakis A., Ott M., Rouse G.W., Edgecombe G.D., et al. (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci. *276*, 4261-4270.

Herd, K.E., Barker, S.C., and Shao, R. (2015). The mitochondrial genome of the chimpanzee louse, *Pediculus schaeffi*: insights into the process of mitochondrial genome fragmentation in the blood-sucking lice of great apes. BMC Genomics. *16*, 661.

Huang, D.W., Sherman, B.T., Zheng, X., Yang, J., Imamichi, T., Stephens, R., *et al*. (2009). Extracting biological meaning from large gene lists with DAVID. Curr Protoc Bioinformatics. *27*, 13.11.1-13.11.13.

Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., *et al*. (2012). HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Research. *22*, 1581-1588.

Husch A., Paehler M., Fusca D., Paeger L., Kloppenburg P. (2009). Calcium current diversity in physiologically different local interneuron types of the antennal lobe. J Neurosci. *29*, 716-726.

Ikuta, T. (2011). Evolution of invertebrate deuterostomes and Hox/ParaHox genes. Genomics Proteomics Bioinformatics. *9*, 77-96.

Jackson, A.P. (2015). The evolution of parasite genomes and the origins of parasitism. Parasitology. *142*, S1-S5.

Jennings, H.S. (1906). Behaviour of the lower lrganisms. London, Indiana University Press.

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

Katayama, T., Wada, H., Furuya, H., Satoh, N., and Yamamoto, M. (1995). Phylogenetic position of the dicyemid mesozoa inferred from 18S rDNA sequences. Biol. Bull. *189*, 81-90.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. *30*, 772-780.

Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. *27*, 757-763.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Research. *12*, 656-664.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kobayashi, M., Furuya, H., and Holland, P.W. (1999). Dicyemids are higher animals. Nature. *401*, 762-762.

Kobayashi, M., Furuya, H., and Wada, H. (2009). Molecular markers comparing the extremely simple body plan of dicyemids to that of lophotrochozoans: insight from the expression patterns of *Hox*, *Otx*, and *brachyury*. Evolution & Development. *11*, 582-589.

Kojima, N.F., Kojima, K.K., Kobayakawa, S., Higashide, N., Hamanaka, C., Nitta, A., *et al*. (2010). Whole chromosome elimination and chromosome terminus elimination both contribute to somatic differentiation in Taiwanese hagfish *Paramyxine sheni*. Chromosome Res. *18*, 383-400.

Kück, P., and Struck, T.H. (2014). BaCoCa – A heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. Molecular Phylogenetics and Evolution. *70*, 94-98.

Lanfear, R., Calcott, B., Ho, S.Y.W., and Guindon, S. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution. *29*, 1695-1701

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Mathods. *9*, 357-359.

Lapan, E.A., and Morowitz, H. (1972). The mesozoa. Sci. Am. *227*, 94-101.

Lapan, E.A., and Morowitz, H.J. (1975). The dicyemid Mesozoa as an integrated system for morphogenetic studies. I. Description, isolation and maintenance. J Exp Zool. *193*, 147-160.

Lapan E.A. (1975). Studies on the chemistry of the octopus renal system and an observation on the symbiotic relationship of the dicyemid Mesozoa. Comp Biochem Physiol. *52*, 651-657.

Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. *7*, S4.

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Systematic Biology. *62*, 611-615.

Laumer, C.E., Bekkouche, N., Kerbl, A., Goetz, F., Neves, R.C., Sørensen, M.V., *et al*. (2015). Spiralian phylogeny informs the evolution of microscopic lineages. Current Biology. *25*, 2000-2006.

Leggett, R.M., Clavijo, B.J., Clissold, L., Clark, M.D., and Caccamo, M. (2014). NextClip: an analysis and read preparation tool for Nextera long mate pair libraries. Bioinformatics. *30*, 566-568.

Lu, T.M., Kanda, M., Satoh, N., and Furuya, H. (2017). The phylogenetic position of dicyemid mesozoans offers insights into spiralian evolution. Zoological Letters; *3*, 6.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., *et al*. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience. *1*, 18.

Luo, Y.J., Takeuchi, T., Koyanagi, R., Yamada, L., Kanda, M., Khalturina, M., *et al*. (2015). The *Lingula* genome provides insights into brachiopod evolution and the origin of phosphate biomineralization. Nature Communications. *6*, 8301.

Luo, Y.J., Satoh, N., and Endo, K. (2015). Mitochondrial gene order variation in the brachiopod *Lingula anatina* and its implications for mitochondrial evolution in lophotrochozoans. Marine Genomics. *24*, 31-40.

Luo, Y.J., Kanda, M., Koyanagi, R., Hisata, K., Akiyama, T., Sakamoto, H., *et al*. (2018). Nemertean and phoronid genomes reveal lophotrochozoan evolution and the origin of bilaterian heads. Nat. ecol. evol. *2*, 141-151.

Mallo, M., and Alonso, C.R. (2013). The regulation of Hox gene expression during animal development. Development. *140*, 3951-3963.

Malmstrøm M., Britz R., Matschiner M., Tørresen O.K., Hadiaty R.K., Yaakob N., et al. (2018). The Most Developmentally Truncated Fishes Show Extensive Hox Gene Loss and Miniaturized Genomes. Genome Biol Evol. *10*, 1088-1103.

McConnaughey, B.H. (1951). The life cycle of the dicyemid Mesozoa. Univ. Calif. Publ. Zool. *55*, 295-336.

Mi, H., Muruganujan, A., and Thomas, P.D. (2012). PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Research. *41*, 377-386.

Mikhailov, K.V., Slyusarev, G.S., Nikitin, M.A., Logacheva, M.D., Penin, A.A., Aleoshin, V.V., *et al*. (2016). The Genome of Intoshia linei Affirms Orthonectids as Highly Simplified Spiralians. Current Biology. *26*, 1768-1774.

Nielsen C. (2012). Animal evolution - interrelationships of the living phyla. New York: Oxford University Press, Inc.

Nesnidal, M.P., Helmkampf, M., Meyer, A., Witek, A., Bruchhaus, I., Ebersberger, I., *et al*. (2013). New phylogenomic data support the monophyly of Lophophorata and an Ectoproct-Phoronid clade and indicate that Polyzoa and Kryptrochozoa are caused by systematic bias. BMC Evol Biol. *13*, 253.

Noto, T., Yazaki, K., and Endoh, H. (2003). Developmentally regulated extrachromosomal circular DNA formation in the mesozoan *Dicyema japonicum*. Chromosoma. *111*, 359-368.

Ogino, K., Tsuneki, K., and Furuya, H. (2007). Cloning of chitinase-like protein1 cDNA from dicyemid mesozoans (Phylum: Dicyemida). J. Parasitol. *93*, 1403-1415.

Ogino, K., Tsuneki, K., and Furuya, H. (2010). Unique genome of dicyemid mesozoan: highly shortened spliceosomal introns in conservative exon/intron structure. Gene. *449*, 70-76.

Ogura, A. and Machemer, H. (1980). Distribution of mechanoreceptor channels in the Paramecium surface membrane. J. Comp. Physiol. *135*, 233-242.

Page, R.D., Lee, P.L., Becher, S.A., Griffiths, R., and Clayton, D.H. (1998). A different tempo of mitochondrial DNA evolution in birds and their parasitic lice. Mol Phylogenet Evol. *9*, 276-293.

Pándy-Szekeres, G., Munk, C., Tsonkov, T.M., Mordalski, S., Harpsøe, K., Hauser, A.S., *et al*. (2017). GPCRdb in 2018: adding GPCR structure models and ligands. Nucleic Acids Research. *46*, 440-446.

Paparo, A.A. (1986). Average ciliary beat in the oyster: response to photoperiod, pentylenetetrazole, salyrgan, serotonin and dopamine. Mar. Behav. Physiol. *12*, 149-159.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. *23*, 1061-1067.

Passamaneck, Y.J., and Halanych, K.M. (2004). Evidence from Hox genes that bryozoans are lophotrochozoans. Evolution & Development. *6*, 275-281.

Pawlowski, J., Montoya-Burgos, J.I., Fahrni, J.F., Wüest, J., and Zaninetti, L. (1996). Origin of the Mesozoa inferred from 18S rRNA gene sequences. Molecular Biology and Evolution. *13*, 1128-1132.

Pedragosa-Badia, X., Stichel, J., and Beck-Sickinger, A.G. (2013). Neuropeptide Y receptors: how to get subtype selectivity. Front Endocrinol. *4*, 5.

Petrov, N.B., Aleshin, V.V., Pegova, A.N., Ofitserov, M.V., and Slyusarev, G.S. (2010). New insight into the phylogeny of Mesozoa: evidence from the 18S and 28S rRNA genes. Moscow University biological sciences bulletin. *65*, 167-169.

Pierce, K.L., Premont, R.T., and Lefkowitz, R.J. (2002). Seven-transmembrane receptors. Nat Rev Mol Cell Biol. *3*, 639-650.

Pont, G., Degroote, F., and Picard, G. (1987). Some extrachromosomal circular DNAs from Drosophila embryos are homologous to tandemly repeated genes. J Mol Biol. *195*, 447-451.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). *De novo* identification of repeat families in large genomes. Bioinformatics. *21*, 351-358.

Ridley, R.K. (1968). Electron microscopic studies on dicyemid Mesozoa. I. Vermiform stages. J. Parasitol. *54*, 975–998.

Ridley, R.K. (1969). Electron microscopic studies on dicyemid Mesozoa. II. Infusorigen and infusoriform stages. J. Parasitol. *55*, 779-793.

Robertson, H., Schiffer, P., and Telford, M. (2018). The mitochondrial genomes of the mesozoans *Intoshia linei*, *Dicyema* sp., and *Dicyema japonicum*. bioRxiv. doi: http://dx.doi.org/10.1101/282285.

Rodgers, L.F., Markle, K.L., and Hennessey, T.M. (2008). Responses of the ciliates tetrahymena and paramecium to vertebrate odorants and tastants. J. Eukaryot. Microbiol. *55*, 27-33.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., and Reyes, A. (1999). Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. Gene. *238*, 195-209

Shao, R., Li, H., Barker, S.C., and Song, S. (2017). The Mitochondrial Genome of the Guanaco Louse, Microthoracius praelongiceps: Insights into the ancestral mitochondrial karyotype of sucking lice (Anoplura, Insecta). Genome Biol Evol. *9*, 431-445.

Shields E.J., Sheng L., Weiner A.K., Garcia B.A., Bonasio R. (2018) High-Quality Genome Assemblies Reveal Long Non-coding RNAs Expressed in Ant Brains. Cell Rep. *23*, 3078-3090.

Shinzato, C., Shoguchi, E., Kawashima, T., Hamada, M., Hisata, K., Tanaka, M., *et al*. (2011). Using the *Acropora digitifera* genome to understand coral responses to environmental change. Nature. *476*, 320-323.

Shmukler, Yu.B., and Tosti, E. (2002). Serotonergic-induced ion currents in cleaving sea urchin embryo, Invertebr. Reprod. Devel. *42*, 43-49.

Short, R.B., and Damian, R.T. (1966). Morphology of the infusoriform larva of *Dicyema aegira* (Mesozoa: Dicyemidae). J. Parasitol. *52*, 746-751.

Simakov, O., Marletaz, F., Cho, S.J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., *et al*. (2013). Insights into bilaterian evolution from three spiralian genomes. Nature. *493*, 526-531.

Smit, A.F.A., Hubley, R. and Green, P. (2013). RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org.

Smith, J.J., Saha, N.R., and Amemiya, C.T. (2010). Genome biology of the cyclostomes and insights into the evolutionary biology of vertebrate genomes. Integrative and Comparative Biology. *50*, 130-137.

Spano, F., and Crisanti, A. (2000). *Cryptosporidium parvum*: the many secrets of a small genome. Int. J. Parasitol. *30*, 553-565.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E.A., Mitros, T., *et al*. (2010). The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature. *466*, 720-726.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. *30*, 1312-1313.

Struck, T.H., Wey-Fabrizius, A.R., Golombek, A., Hering, L., Weigert, A., Bleidorn, C., *et al*. (2014). Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of Spiralia. Molecular Biology and Evolution. *31*, 1833-1849

Stunkard, H.W. (1954). The life-history and systematic relations of the Mesozoa. Q Rev Biol. *29*, 230-244.

Suzuki, T.G., Ogino, K., Tsuneki, K., and Furuya, H. (2010). Phylogenetic analysis of dicyemid mesozoans (phylum Dicyemida) from innexin amino acid sequences: dicyemids are not related to Platyhelminthes. J. Parasitol. *96*, 614-625.

Tsai, I.J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sanchez-Flores, A., Brooks, K.L., *et al*. (2013). The genomes of four tapeworm species reveal adaptations to parasitism. Nature. *496*, 57-63.

Valentine, M., Yano, J., and Van Houten, J. (2010). Chemosensory transduction in Paramecium, Japanese Journal of Protozoology. *41*, 1-8.

van Beneden, E. (1876). Recherches sur les Dicykmides. Bull. Acad. Roy. Belg., *41*, 1160-1205.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., *et al*. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. *33*, 2202-2204.

Wang, X., Chen, W., Huang, Y., Sun, J., Men, J., Liu, H., *et al*. (2011). The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis*. Genome Biol. *12*, R107.

Wang, J., and Davis, R.E. (2014). Programmed DNA elimination in multicellular organisms. Curr. Opin. Genet. Dev. *27*, 26-34.

Watanabe, K.I., Bessho, Y., Kawasaki, M., and Hori, H. (1999). Mitochondrial genes are found on minicircle DNA molecules in the mesozoan animal Dicyema. J. Mol. Biol. *286*, 645-650.

Weinstein, S.B., and Kuris, A.M. (2016). Independent origins of parasitism in Animalia. Biol. Lett. *12*, 20160324.

Wolff G.H., Thoen H.H., Marshall J., Sayre M.E., Strausfeld N.J. (2017). An insect-like mushroom body in a crustacean brain. Elife. *6*, e29889.

Yamagishi, H., Tsuda, T., Fujimoto, S., Toda, M., Kato, K., Maekawa, Y., Umeno, M., and Anai, M. (1983). Purification of small polydisperse circular DNAs of eukaryotic cells by use of ATP-dependent deoxyribonuclease. Gene. *26*, 317-321.

Yu, Z., Wei, Z., Kong, X., and Shi, W. (2008). Complete mitochondrial DNA sequence of oyster *Crassostrea hongkongensis* - a case of "Tandem duplication-random loss" for genome rearrangement in Crassostrea? BMC Genomics. *9*, 477.

Zarowiecki, M., and Berriman, M. (2014). What helminth genomes have taught us about parasite evolution. Parasitology. *142*, S85-S97.

Zhong, Y.F., Butts, T., and Holland, P.W.H. (2008). HomeoDB: a database of homeobox gene diversity. Evolution & Development. *10*, 516-518.

Zhong, Y.F., and Holland, P.W. (2011). The dynamics of vertebrate homeobox gene evolution: gain and loss of genes in mouse. BMC Evol Biol. *11*, 169.